Phénomène Big Data en entreprise : processus projet, génération de valeur et Médiation Homme-Données

Anna Nesvijevskaia

le cnam

ÉCOLE DOCTORALE Abbé Grégoire

Laboratoire Dispositifs d'Information et de la Communication à l'Ere Numérique

THÈSE présentée par

Anna NESVIJEVSKAIA

soutenue le 18 octobre 2019

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline : Sciences de l'Information et de la Communication

Phénomène Big Data en entreprise : processus projet, génération de valeur et Médiation Homme-Données.

THÈSE dirigée par :

Madame CHARTRON Ghislaine	Professeur, CNAM Paris, Sciences de l'information et de la communication

RAPPORTEURS :

Monsieur BOURRET Christian	Professeur, Université Paris-Est Marne-la-Vallée, Sciences de l'information et de la communication
Monsieur MOINET Nicolas	Professeur, IAE de Poitiers, Sciences de l'information et de la communication

---

JURY :

Madame DUDEZERT Aurélie	Professeur, Université Paris Sud, Management
Monsieur GAREL Gilles	Professeur, CNAM Paris, Gestion de l'innovation
Madame PINEDE Nathalie	Maître de conférences HDR, Université Bordeaux Montaigne, Sciences de l'Information et de la Communication

A ma famille

Remerciements

Je tiens à remercier en tout premier lieu le Professeur Ghislaine Chartron, ma directrice de thèse, qui m'a accordé sa confiance tout le long de ces six années de recherche, pour ses conseils et son partage d'expérience, pour ses intuitions et ses propositions d'angles de recherche captivants, et pour sa constance, sa délicatesse et sa bienveillance dans la façon de me faire garder le cap.

Je tiens à exprimer toute ma reconnaissance à Messieurs Christian Bourret, Professeur des Universités en Sciences de l'Information et de la Communication, et Nicolas Moinet, Professeur des universités à l'IAE de Poitiers, membres de mon Comité de suivi de thèse pour leurs encouragements, pour m'avoir fait assumer la transdisciplinarité de ces travaux et pour l'honneur qu'ils m'ont fait en acceptant d'être les rapporteurs de cette thèse. Je remercie également les Professeurs Aurélie Dudézert et Gilles Garel ainsi que Madame Nathalie Pinède, Maître de Conférences HDR, qui ont bien voulu être examinateurs.

Je remercie Madame Maria Mercanti Guérin, pour avoir partagé avec moi les spécificités des méthodes des Sciences de Gestion, et pour ses riches retours, francs et pragmatiques.

Je tiens à remercier les Professeurs Jean Charles Clément et Pierre Yves Gomez pour m'avoir non seulement initiée aux Sciences de Gestion lorsque j'étais de leurs élèves, mais aussi pour avoir pris le temps de me guider dans mes premières réflexions balbutiantes sur un troisième cycle dont cette thèse est l'accomplissement. Merci également au Professeur Michel Bera ainsi qu'au Professeur Gilbert Saporta pour m'avoir donné leurs avis précieux sur la mise en perspective du phénomène Big Data et pour les premières références structurantes lorsque je m'attaquais à la problématique.

J'adresse tous mes remerciements au Professeur Manuel Zacklad pour m'avoir accueillie dans le laboratoire DICEN IDF qu'il dirige. Je remercie également l'équipe de l'INTD, et en particulier Mesdames Béa Arruabarrena et Evelyne Broudoux avec qui j'ai gouté aux plaisirs et aux exigences des publications et des conférences scientifiques, sans oublier Mesdames Adriana Lopez Uroz et Catherine de Laitre pour leur aide documentaire salutaire.

Je souhaite remercier les équipes de KPMG pour leur enthousiasme et leur soutien lors de mon lancement dans cette aventure.

Je remercie également les équipes IMA, et plus particulièrement Monsieur Antoine Trarieux et le plateau Assistance Télématique pour m'avoir offert mes premiers entretiens terrain et, à travers eux, les axes principaux sur lesquels j'ai pu orienter ces travaux de recherche.

J'adresse tous mes remerciements à Madame Alaoui et à Messieurs Alexandre Templier et Guillaume Bourdon pour m'avoir ouvert, en toute liberté et confiance, l'accès à une matière première aussi riche au sein des équipes de Quinten. Merci à Messieurs Lucas Davy et Alexandre Civet pour avoir pris le temps de m'exposer leurs projets en entretien, avant le démarrage de notre collaboration, puis à tous les membres de l'équipe qui, sans le savoir, m'inspiraient et alimentaient mes observations au quotidien. Qu'ils soient remerciés pour ces années de partage professionnel et humain.

Mes remerciements vont aussi aux entreprises qui ont mis en place les projets auxquels j'ai eu la chance de participer, et en particulier aux sponsors et membres des équipes projet. Je remercie également mes interlocuteurs professionnels et académiques ainsi que les élèves pour leur curiosité et pour les échanges qui ont fait murir mes réflexions.

J'exprime ma gratitude au Professeur François Ewald pour l'intérêt qu'il a manifesté pour mes travaux, ainsi que pour son aide inestimable et les encouragements qu'il m'a inlassablement prodigués pour mes travaux et au-delà.

Je félicite par ailleurs tous les « membres de l'équipe médicale et logistique » qui ont réussi l'exploit remarquable de me maintenir debout et administrativement équipée tout le long de ce marathon, et plus particulièrement dans la dernière ligne droite.

Merci enfin à mes proches et à ma famille pour leur enthousiasme infatigable, leur écoute en période de doute et leur patience en période d'absorption, ainsi que leur joyeux et indispensable support, même de loin, lors de la soutenance. Face à toutes les épreuves qui ont jalonné, inévitablement, ce travail de longue haleine, leur soutien infaillible m'a permis de ne jamais dévier de l'objectif.

Résumé

Le Big Data, phénomène sociotechnique porteur de mythes, se traduit dans les entreprises par la mise en place de premiers projets, plus particulièrement des projets de Data Science. Cependant, ils ne semblent pas générer la valeur espérée. La recherche-action menée au cours de 3 ans sur le terrain, à travers une étude qualitative approfondie de cas multiples, pointe des facteurs clés qui limitent cette génération de valeur, et notamment des modèles de processus projet trop autocentrés. Le résultat est (1) un modèle ajusté de dispositif projet data (Brizo_DS), ouvert et orienté sur les usages, dont la capitalisation de connaissances, destiné à réduire les incertitudes propres à ces projets exploratoires, et transposable à l'échelle d'une gestion de portefeuille de projets data en entreprise. Il est complété par (2) un outil de documentation de la qualité des données traitées, le Databook, et par (3) un dispositif de Médiation Homme-Données, qui garantissent l'alignement des acteurs vers un résultat optimal.

Mots clés :

Big Data, Data Science, Intelligence Artificielle, Qualité des données, Médiation Homme-Donnée, Stratégie d'entreprise, Capitalisation de connaissances, Projet Data, Indicateurs de valeur, Cas d'Usage Métier.

Résumé en anglais

Big Data, a sociotechnical phenomenon carrying myths, is reflected in companies by the implementation of first projects, especially Data Science projects. However, they do not seem to generate the expected value. The action-research carried out over the course of 3 years in the field, through an in-depth qualitative study of multiple cases, points to key factors that limit this generation of value, including overly self-contained project process models. The result is (1) an open data project model (Brizo_DS), orientated on the usage, including knowledge capitalization, intended to reduce the uncertainties inherent in these exploratory projects, and transferable to the scale of portfolio management of corporate data projects. It is completed with (2) a tool for documenting the quality of the processed data, the Databook, and (3) a Human-Data Mediation device, which guarantee the alignment of the actors towards an optimal result.

Keywords:

Data Science, Artificial Intelligence, Data Quality, Human-Data Mediation, Business Strategy, Knowledge Capitalization, Data Project, Value Metrics, Business Use Case.

Sommaire

Liste des figures

Liste des annexes

Préambule

Le Big Data alimente un discours de promesse de création de valeur sans précédent grâce à une meilleure exploitation des données à l'ère de l'information. La Data Science, processus de transformation de données en connaissances utiles grâce à des algorithmes de pointe, rendus opérationnels sur les technologies de nouvelle génération, est annoncée comme disruptive pour la construction de connaissances métier inédites, et pour l'optimisation, voire l'automatisation, de la prise de décision. Pourtant, les entreprises n'ont pas attendu le phénomène pour mobiliser des outils et des méthodes analytiques afin de mieux décider et capitaliser des connaissances. Le buzz éveille alors l'émerveillement, la crainte ou la perplexité chez les acteurs historiques, et les premiers projets mis en œuvre dans les entreprises tardent à tenir la promesse de génération de bénéfices tangibles et significatifs. La révolution incontestable assurée semble remise en cause. Dans ce contexte, ces travaux de recherche sont guidés par un désir de faire la part des choses entre le mythe et la réalité du Big Data, et par l'intuition d'une relation, indissociable mais immature, entre la valeur des nouveaux usages et le dispositif qui permet de les construire. Ils ont alors une visée double : dresser un état des lieux du phénomène en enquêtant sur ses éventuelles nouveautés, et comprendre sous quelles conditions les nouveaux modes d'exploitation de la donnée par l'homme seraient plus bénéfiques, en plongeant au cœur des dispositifs de projets Data Science dans les entreprises françaises entre 2014 et 2017.

Face au manque de recul et de définitions partagées et académiques sur ce phénomène récent, ces travaux de recherche font un détour introductif par une mise en perspective historique du Big Data et des enjeux soulevés (Introduction du contexte) avant d'énoncer la problématique et poser le plan de thèse et le cadre conceptuel (Première Partie). La définition du terrain et des méthodes (Deuxième Partie) suivra cet énoncé avant de proposer les résultats (Troisième Partie) et les conclusions.

Introduction du contexte

1 L'homme et la donnée : un historique multidisciplinaire

L'exploitation de la donnée au service de la prise de décision par l'homme s'enracine dans les pratiques des Etats, ayant guidé les développements mathématiques et technologiques bien avant les derniers progrès en informatique, les algorithmes à la mode, la naissance du terme « Big Data » et sa propagation dans le monde des entreprises sous la forme d'un phénomène sociotechnique, porteur de promesses (voir illustration de l'ensemble de ce chapitre en Annexe 1 – Courte histoire du Big Data et des algorithmes).

1.1 La donnée, une affaire d'Etat millénaire

La collecte des informations à des fins de création de connaissance et de prise de décision n'est pas nouvelle. Les recensements démographiques, opération statistique visant à dénombrer et qualifier une population à des fins militaires, fiscales, économiques, comparatives ou autres, existent déjà dans l'Egypte des pharaons, en Grèce Antique, ou encore en Chine sous la dynastie Han. Le recensement de population se distingue du sondage, basé sur un échantillon de la population, bien que les deux soient utilisés dans la recherche, le marketing, ou bien la politique. La démographie, discipline couvrant ces méthodes, s'intéresse ainsi à la taille, à la distribution, aux caractéristiques de la structure et à la dynamique d'une population, allant parfois de pair avec la cartographie et d'autres disciplines anciennes. Elle définit notamment un cadre méthodologique élaboré et en transition, afin de répondre aux enjeux et aux problèmes de collecte de données (représentativité de l'échantillon, dénombrement...), de classification, de qualité de l'information, de méthodes d'extrapolation ou d'analyse des données.

Les statistiques modernes, s'appuyant sur les premiers travaux des statisticiens français et fondées au début du 19ème siècle sur une formalisation mathématique rigoureuse, s'ancrent dans cette discipline millénaire, comme en témoigne encore le vocabulaire de base (*population*, *effectif, individu*...), ou simplement leur nom. En effet, *statista*, homme d'Etat en italien, connaisseur de son pays, est à l'origine du nom donné à ce champ en 1749 par l'allemand Gottfried Achenwal, en le définissant comme la science politique de plusieurs pays. La même année, le premier institut de statistique officiel, la Tabellverket (Bureau des Tables), est créé

par Pehr Wilhelm Wargentin pour les recensements suédois : au-delà d'un simple dénombrement d'hommes mobilisables, l'étude porte sur la natalité, la mortalité, la pratique du luthéranisme, puis l'emploi ou encore les mouvements de population au cours de la Révolution Industrielle. Les statistiques ne sont élargies que plus tard à l'ensemble des terrains de collecte et d'analyse des données, élevées au rang de science (Quetelet, 1849), et enrichies de concepts majeurs comme la régression (Galton, 1886), la corrélation (Pearson, 1900), ou encore le design expérimental (Fisher, 1937).

Au-delà de l'évolution des fondements méthodologiques et mathématiques, la démographie fait face à un autre problème : une insuffisance des supports pour la collecte des données. Lorsqu'en juin 1880 les Etats Unis entament le 10ème recensement démographique, la collecte d'informations, sur seulement 5 paramètres, dure près de 7 ans, un temps estimé trop long. Le Bureau de Recensement finit alors par signer un contrat avec Herman Hollerith pour le développement d'une tabulatrice pour le référencement de 1890. De cette impulsion naît la mécanographie, c'est-à-dire l'usage des techniques mécaniques et électromécaniques au service de l'exécution rapide d'un algorithme pour en afficher le résultat. Les brevets donneront lieu aux développements de machines de Computing-Tabulating-Recording Company, renommée plus tard en 1924 par Thomas J. Watson en International Business Machines, ou IBM. L'usage des cartes perforées se diffuse en Europe et en Russie au-delà des statistiques générales : le transport, la téléphonie, l'éducation, l'administration des salaires sont autant de domaines dans lesquels la technologie a été mise en place au cours des deux premières décennies du XXème siècle. Rapidement, les industriels pionniers, comme Renault ou Michelin en France, introduisent la technologie au service de l'optimisation du processus industriel et de la performance d'entreprise. Si les tabulatrices ont, depuis, fait place aux ordinateurs, la mécanographie reste utilisée pour certaines de leurs composantes, et sera remise au goût du jour grâce à la nano-mécanique, autrement dit les nanotechnologies.

La croissance de la population aux Etats-Unis au début du XXème siècle, et notamment la nécessité de répondre aux enjeux multiculturels liés aux vagues migratoires, s'accompagne d'une croissance des données de recherche et de sécurité. Cette croissance impacte l'organisation et la classification des informations dans les administrations et les librairies, et alimente le développement des sciences de l'information à travers l'extension de techniques de la gestion documentaire. Malgré ses évolutions, bibliothécaires et chercheurs tirent la sonnette d'alarme. Fremont Rider démontre en 1944 que les bibliothèques de recherche américaines

doublent en taille tous les 16 ans. Derek Price pointe en 1961 l'apparition exponentielle de nouveaux journaux et revues scientifiques, doublant tous les 15 ans, expliquée par le fait que chaque avancée génère une nouvelle série d'avancées à un taux d'apparition relativement constant, de sorte que le nombre d'apparitions est strictement proportionnel à la taille de la population de découvertes à un moment donné. Le concept d'explosion d'information fait son apparition en 1964 dans les titres du New Statesman et du New York Times, poursuivi en 1970 par la notion de surcharge informationnelle (Toffler, 1984), ou infobésité, c'est-à-dire l'excès d'informations reçues par un système, dépassant ses capacités à les traiter. Ce défi, comme conséquence de l'explosion informationnelle, est adressé par un ensemble de recherches scientifiques, allant de l'informatique à la théorie d'économie d'attention en sciences sociales, au cours du dernier demi-siècle.

Les solutions face à l'explosion des données intègrent alors des avancées théoriques et pratiques récentes, s'appuyant notamment sur les progrès tractés par des intérêts militaires au cours des guerres mondiales : Enigma et la Machine de Turing en sont des exemples bien connus. Tout d'abord, la formulation de la théorie de l'information en 1948 par Claude Shannon, pour le modèle télégraphique, ouvre la voie au codage de l'information, à la compression ou encore à la cryptographie. Ses fondements probabilistes visent à qualifier et quantifier la notion de contenu d'un ensemble de messages en information, concept physique, mesurable, bien que non observable. La théorie de l'information de Shannon (dite « Communication Theory ») porte notamment sur le processus de transmission d'informations entre hommes ou entre machines, en tenant compte de la notion de bruit et d'entropie, c'est-à-dire de l'incertitude, et s'applique jusqu'à ce jour en informatique ou en télécommunication.

Elle est reprise au cours du développement de la cybernétique, qui bat son plein aux Etats-Unis dès 1948 sous l'impulsion de Norbert Wiener, en tant que science interdisciplinaire à l'intersection de l'automatique, de l'électronique et de la théorie mathématique de l'information formulée par Shannon. La cybernétique, du grec *kubernân*, « gouverner, piloter », comme « théorie entière de la commande et de la communication, aussi bien chez l'animal que dans la machine » pose le concept de boite noire et de rétroaction, et s'applique en robotique, en sciences cognitives ou en intelligence artificielle, ainsi qu'en Sciences de Gestion. A l'est, dans la Russie des années 60', en pleine déstalinisation et conquête de l'espace, la théorie de l'information et la cybernétique trouvent une résonnance forte. Emerge alors, sous l'impulsion de Kolmogorov, Solomonov et Chaitin, la théorie algorithmique de l'information, reprenant le

concept de l'algorithme de la machine de Turing. Bien que compatible avec la première, elle apporte la notion de calculabilité face à un ensemble statistique des données. Ses applications sont nombreuses, là aussi, notamment en physique et en biologie, et elle constitue un domaine de l'informatique.

Les deux écoles, l'une portant sur la modélisation stochastique, c'est-à-dire aléatoire, des données, et l'autre sur la modélisation algorithmique, plus déterministe, poursuivent leur évolution de façon assez indépendante (Béra, 2014) en pleine guerre froide, ce qui présente des limites. La première se heurte à la pratique de la collecte des données et au manque de pragmatisme, et la seconde manque de fondements statistiques, s'étant développée essentiellement de façon appliquée, en dehors de la discipline. Lorsque, dans le milieu des années 80', sous l'impulsion de la disponibilité de deux nouveaux algorithmes (arbres de décision et réseaux de neurones), se met en place une nouvelle communauté de recherche, visant à faire converger les deux écoles, les résultats sont phénoménaux, comme l'illustrent les apports théoriques de Vladimir Vapkin, contrant l'enjeu de la malédiction des grands nombres (Béra, 2011; Pajot, 2016), et aboutissant aux techniques de machines à vecteurs de support (SVM).

1.2 Les progrès technologiques et informatiques

Cependant les avancées en statistiques seules ne permettent pas de répondre aux enjeux liés à l'explosion des données, qui nécessitent des solutions technologiques appropriées au stockage et au traitement de l'information. Alors que les cartes perforées montrent leurs limites évidentes et que l'ordinateur théorique est déjà imaginé par l'inventeur Charles Babbage, IBM s'attaque en 1937 au projet ASCC, dit Harvard Mark I. Conçu par Howard Hathaway Aiken, ce calculateur électromécanique n'est pas doté de la possibilité de programmation, et fonctionne avec des cartes perforées qu'il est nécessaire de remettre en entrée manuellement en cas de « boucle » conditionnelle. En parallèle, le Z3, première machine programmable automatique, est créé entre 1938 et 1941 en Allemagne par Konrad Zuse. La conception des deux premiers ordinateurs entièrement électroniques débute en 1943. Les Etats-Unis lancent le développement de l'ENIAC, et la Grande Bretagne celui du Colossus Mark I. Ce dernier est conçu pour déchiffrer le code Lorenz, utilisé par les Allemands, tout comme le code Enigma, mais seulement pour de rares communications entre hauts dirigeants allemands. Le développement de ces ordinateurs s'est largement appuyé sur les travaux de Turing et de Shannon, et intègrent le calcul binaire. Le développement futur des ordinateurs à travers l'introduction de transistors dans les années 50' s'appuie sur ce système binaire. D'autres solutions technologiques, comme

la compression (Marron & de Maine, 1967) ou VLSI, technologie de circuit intégré permettant la fabrication de puces comprenant des millions de transistors, commencent à faire leur apparition. Lorsque sont conçues les puces électroniques, Moore, directeur de recherche et de développement chez Fairchild Semiconductor et futur co-fondateur de Intel en 1968, formule la conjecture, dite la loi de Moore (Moore, 1965), qui anticipe un doublement de la capacité des composants électroniques tous les 18 mois. Sa conjoncture se vérifie (Fanet, 2008; Schaller, 1997) depuis sa formulation, en traversant la période de développement des circuits intégrés lancés par Intel : tous les 18 mois, le nombre de transistors qui peuvent être installés sur une puce double, avec une baisse des coûts des microprocesseurs, liée aux progrès de la miniaturisation, au traitement collectif de silicium et au parallélisme. Cependant, le manque de fondements de cette conjecture est largement souligné (Kish, 2002; Meindl, 2003), que ce soit en termes théoriques, pratiques ou physiques, en particulier en se basant sur ses limites de consommation énergétique ou de taille de support en silicone, voué à être remplacé grâce au développement des nanotechnologies. En attendant, la baisse des coûts de production des ordinateurs et autres dispositifs développés, notamment la téléphonie mobile, et plus généralement les objets communicants (réseaux de capteurs, mouvement Quantified Self…), permet la multiplication des usages auprès de la recherche, du grand public, des institutions publiques ou des entreprises.

Au-delà de l'aspect technique et industriel, la programmation informatique évolue (Knuth, 1969) et les langages développés donnent la possibilité d'implémenter des algorithmes de plus en plus sophistiqués, appliqués par des programmes. En 1970, Xerox PARC (Palo Alto Research Center) est fondé en Californie par Jacob Goldman et Robert Taylor. Le premier est physicien, travaillant pour Xerox, qui fabrique à l'époque des imprimantes et craint la concurrence japonaise. Le second, Robert Taylor, est directeur du Bureau des techniques de traitement de l'information de l'ARPA au Pentagone, responsable du projet ARPAnet (Taylor & Licklider, 1968), lancé dans le cadre du retard pris sur les soviétiques dans l'aérospatial et précurseur d'Internet. Xerox PARC a alors pour objectif d'accélérer l'innovation pour Xerox, et donne lieu, au cours de la décennie, à la mise en place de standards informatiques, comme l'invention de l'imprimante laser, la souris, la programmation orientée objet, la conceptualisation de l'ordinateur personnel (PC), l'interface graphique utilisateur ou encore Ethernet et le calcul distribué. Ces avancées, couplées avec le développement de l'ergonomie et des sciences cognitives, contribuent à la propagation des ordinateurs auprès d'un public moins expert. A ces progrès s'ajoute la mise à disposition dans le domaine public par le CERN,

en 1993, du logiciel World Wide Web, créé quatre années plus tôt par le scientifique britannique Tim Berners-Lee. En s'appuyant sur l'effet réseau (Bomsel, 2007), commun dans les télécommunications, Internet atteint rapidement les ordinateurs personnels, et permet l'explosion de l'univers des objets connectés (IOT, Internet Of Things). Il participe ainsi à l'augmentation du flux informationnel, effaçant (Béra & Méchoulan, 1999) les spécificités des notions de « donnée », « information » ou « communication », et nourrissant l'avènement de la société de l'information (Duff, 2000; Gillies & Cailliau, 2000).

L'un des paradigmes nouveaux en informatique, mentionné plus haut, est le calcul distribué. Il fonde une branche de recherche des sciences mathématiques et informatiques. De nombreux projets permettent des avancées dans ce sens, notamment des projets utilisant la bande passante inutilisée des ordinateurs personnels connectés. A Berkeley, Seti@home, projet utilisant des ordinateurs reliés à internet pour la recherche d'une intelligence extraterrestre en analysant une quantité de signaux impossible à traiter jusqu'alors, est rendu public en 1999 et prouve, entre autres et pour commencer, la fiabilité du calcul distribué. Le calcul distribué est utilisé en mathématiques à partir de 1996 pour divers projets scientifiques marqués par une limite infinie de solutions comme GIMPS (recherche de nombres premiers) ou distributed.net (recherche sur le chiffrement et les règles de Gollomb), ou bien dans la branche cryptologie. En informatique, la recherche sur le parallélisme se développe avec l'étude de langages comme le π-calcul de Milner, pour fusionner avec le domaine de calcul distribué grâce au déploiement d'Internet. Ces technologies sont utilisées en science pour les supercalculateurs comme Roadrunner avec une application à des domaines très variés, et s'imposent en paradigme dominant dès le début du millénaire grâce à la décroissance du coût de matériel permettant la construction des systèmes à multiprocesseurs, mais aussi grâce aux progrès dans l'intégration à très grande échelle et l'augmentation de la vitesse de traitement des ordinateurs.

D'autres progrès technologiques marquent ce début du siècle, dont certains sont régulièrement cités comme associés au phénomène Big Data (Varian, 2014). Tout d'abord, le modèle de programmation MapReduce distribué, développé en 2004 par Google (Dean & Ghemawat, 2004), puis breveté en 2010 (Dean & Ghemawat, 2010), permet d'accéder et de manipuler les données dans des structures de données volumineuses, comme BigTable, table de données présente dans le système GoogleFS. Sur ce parton d'architecture s'appuient des frameworks comme Hadoop, créé en 2009 par Doug Cutting et récupéré par Yahoo avec son créateur. Le modèle MapReduce est jugé prometteur (Ranger et al., 2007), la technologie Hadoop, en open

source, est largement reprise dans les logiciels comme Oracle, Microsoft, IBM ou EMC, et donne lieu à de nouveaux développements, tels que Spark qui connait un succès croissant en 2015. Ensuite, le Cloud, comme moyen d'exploitation de la puissance de calcul ou de stockage à distance à travers internet, fait son apparition progressivement (ASP, mails, CRM) et le terme se popularise dès 2006 grâce à l'introduction de l'Elastic Compute Cloud par Amazon.com. Par ailleurs, les outils de Business Intelligence, qui comprennent généralement des bases de données relationnelles extraites par des ETL pour être chargées dans un Data Warehouse structuré, puis requêtées pour des besoins de reporting et de prise de décision, s'adaptent à l'apparition des systèmes de stockage qui utilisent des bases de données orientées objet, grâce à l'évolution des langages de traitement de l'information comme le NoSQL (Not Only SQL). En réaction, les fournisseurs de bases de données relationnelles s'ajustent avec des structures horizontales utilisant les langages NewSQL, comme MySQL. Ces progrès guident la constitution du concept de Data Lake, outil de stockage de données dans leur format natif, qui, contrairement à un Data Warehouse classique, s'affranchit du besoin de structuration amont (voir Annexe 2 - Data Lakes et Informatique Décisionnelle). Enfin, les outils de Data Visualisation évoluent (Tableau, Target...) pour permettre la représentation de données plus volumineuses de façon ergonomique. Ces évolutions constituent le socle technologique des opportunités liées au phénomène Big Data en 2015.

Il semble complexe de trancher entre les tenants du principe que la technologie ait été développée en conséquence des besoins militaires, gouvernementaux, scientifiques ou individuels (Edmunds & Morris, 2000), et les défenseurs de l'idée que les possibilités offertes par le progrès technologique aient crée un vide et tiré l'explosion des données (Tjomsland, 1980). De même, le débat entre l'explosion d'information comme continuité des progrès passés (Barnes, 2013) ou comme révolution (Gillies & Cailliau, 2000; Mayer-Schönberger & Cukier, 2013; McAfee & Brynjolfsson, 2012) reste ouvert. Les efforts de standardisation sur la mesure de l'information (Coffman & Odlyzko, 1998; Dienes, 1994; Pool, 1984; Varian & Lyman, 2003) butent sur l'absence de consensus autour de la définition de celle-ci (Hilbert, 2012). Pourtant, les points de vue s'accordent sur les difficultés à absorber cette information, et la nécessité d'avoir recours à des machines pour préserver la possibilité de découvrir de nouvelles connaissances dans les données (Denning, 1990; Lesk, 1997) ou bien simplement pour visualiser (Cox & Ellsworth, 1997a) de façon digeste l'information existante. Les fruits de la convergence des écoles en statistiques et les derniers progrès en informatique peuvent alors apporter des réponses à l'enjeu de l'explosion des données.

1.3 Le « Big Data » : des origines du terme au phénomène sociotechnique

Le terme « Big Data » semble éclore dans la littérature scientifique en informatique, en mai 1997 dans la publication « Managing big data for scientific visualization » (Cox & Ellsworth, 1997b). Les deux auteurs de l'article sont alors chercheurs en infographie à NASA Ames Research Center, centre de recherche à vocation militaire et civile pour l'aviation américaine, et leurs publications précédentes ciblent le rendu par la programmation parallèle (parallel rendering, ou distributed rendering). L'article en question soulève le concept de Big Data en tant que problème en cours de résolution par les applications commerciales classiques, comme les systèmes de réservation des compagnies aériennes, et par les applications plus récentes, comme le stockage et la fédération de bases de données. Ce problème étant moins appréhendé par l'ingénierie et la visualisation scientifique, l'article fournit des explications en gestion de données, à partir notamment de l'étude sur la visualisation d'écoulement de fluides. Les auteurs continueront à aborder le sujet Big Data sous l'angle de la visualisation des données (Bryson et al., 1999), mais l'apport majeur pour la définition de « Big Data » de ce premier article, qui s'inscrit dans un débat à la NASA sur l'intérêt de l'automatisation ou de l'interaction (Kenwright, 1999) face au Big Data, est la première utilisation du terme associé à une définition qui se précise. Il s'agit de l'accumulation de deux problèmes distincts : « Big Data Collections » et « Big Data Objects ». Le premier correspond à l'agrégation d'un grand nombre de bases de données en provenance de plusieurs sources, souvent pluridisciplinaires, et généralement distribués sur des sites physiques et types de référentiels différents. Le second indique un ensemble de données trop volumineux pour être traité par des algorithmes et logiciels standards sur le matériel disponible. « Big data objects » sont particulièrement problématiques lorsqu'ils sont générés par la méthode de simulation de phénomène physique dans divers domaines scientifiques, comprenant la dynamique des fluides, l'analyse structurelle, la modélisation météo ou l'astrophysique. La combinaison de ces deux problèmes est alors de plus en plus répandue, notamment avec l'approche scientifique double alliant l'expérimentation et la simulation. Big Data, en tant que méthode combinatoire de recherche utilisant l'informatique, est par ailleurs présentée dans des publications scientifiques dans d'autres domaines, comme la génomique (Lenski, 2002).

La seconde apparition du terme « Big Data », sans lien traçable avec la première, a lieu en statistiques dans la préface du livre « Predictive Data Mining » (Weiss & Indurkhya, 1998) de

Weiss et Indurkhya en 1998 lorsqu'ils décrivent le volume de données accumulées dans des entrepôts centralisés de stockage de données. Cette masse de données représente alors une opportunité théorique avec un renforcement des conclusions, mais aussi une difficulté pratique pour ses applications de Data Mining, les techniques duquel sont traitées dans le livre. Il s'agit de l'extraction, de la transformation et de l'organisation de données brutes en vue d'effectuer des recherches multidimensionnelles pour des solutions prédictives. Big Data n'est pas traité en tant que concept, mais comme la première caractéristique du Data Mining, la seconde étant le nombre de dimensions, sujet principal du livre. Plusieurs publications en statistiques se réfèrent à cette définition, comme en 2001 le livre « Data Mining for design and Manufacturing : methods and applications » (Braha, 2001), où, dans un chapitre rédigé avec deux ingénieurs industriels, Dan Braha précise les caractéristiques du Big Data (p.236 : « many variables, many values, and many records »), composante du Data Mining englobant et le nombre d'observations, et leur richesse en termes de dimensions. Cette publication met en évidence le changement dans l'approche scientifique : les techniques de Data Mining dans un contexte Big Data constituent alors une clé permettant de passer d'un modèle de recherche classique guidé par les hypothèses à une approche nouvelle basée sur la donnée.

La définition de Big Data comme accumulation de données stockées décrite par Weiss et Indurkhya est par ailleurs reprise en science informatique après l'intervention de John Mashey en conférence annuelle à USENIX en 1999. Il présente la notion d'Infrastress, due au Big Data (accroissement accélérée du stockage) ainsi qu'à la croissance de l'attente des utilisateurs du Net en termes de type de données différentes et complexes. Notons que ces deux facteurs sont bien distincts, mais leur représentation confondante conduit le lecteur à envisager le Big Data comme un déluge d'information tout court auquel fait face le progrès technique actuellement. La recherche dans la base de données ACM (Association for Computing Machinery digital library) du terme « Big Data » indique par ailleurs que les premières utilisations du terme par d'autres chercheurs dans le domaine SI ont lieu suite à cette conférence de 1999, en particulier dans la recherche sur le World Wide Web (Gschwind & Hauswirth, 1999) ou en gestion de systèmes de fichiers (Randolph Y. Wang et al., 1999).

En 2000, le terme « Big Data » apparait accompagné d'une définition nouvelle dans la publication en statistiques et économétrie de Francis X. Diebold « Big Data Dynamic Factor Models (DFM) for Macroeconomic Measurement and Forecasting » (Diebold, 2012) et désigne un phénomène se référant à « l'explosion en quantité (et parfois qualité) de données disponibles

et potentiellement pertinentes ». L'objectif du terme utilisé consiste alors à marquer le contraste entre l'ancien et le nouvel environnements économétriques DFM décrits par Reichlin et Watson, mais aussi mettre en évidence cette nouvelle caractéristique commune à un ensemble de domaines de recherche plus large que l'économétrie. En effet, le Big Data est un phénomène auquel est confrontée et dont bénéficie la recherche scientifique en physique, biologie et sciences sociales. Ce phénomène nouveau est induit par des avancées technologiques significatives, touchant d'une part à la création, et d'autre part au stockage de données.

Enfin, le terme « Big Data » fait son apparition dans le marketing en 2005 (Ratner, 2004), où la nouveauté est lié à l'entrée dans le domaine, jusqu'alors basé sur la statistique classique, de l'architecture orientée évènement, ou EDA (Event-Driven Architecture). Celle-ci est permise pour un grand volume de données par la possession d'ordinateurs personnels et inverse la relation fournisseur-client classique grâce à l'émission par un service d'un évènement auquel le client doit répondre. Ratner précise qu'il reprend le concept de « Big Data » de Wiess et Indurkhya tout en appuyant d'autres caractéristiques, à savoir son opposition avec « Small Data », représentable sous forme de table de lignes (observations ou individus) et de colonnes (variables ou paramètres) pour un échantillon atteignant rarement 200 lignes pour une poignée de colonnes, caractérisée par sa « propreté » et sa complétude. L'ajout de la notion d'évènement extérieur, la multiplication des tables et l'existence de données secondaires (captées en parallèle de celles qui devaient servir un objectif prédéfini) poussent à revoir les techniques de l'échantillonnage et les méthodes de calcul de répartition, ce qui constitue selon lui le phénomène Big Data.

Le terme se structure progressivement autour d'un ensemble de concepts issus des paradigmes informatiques et statistiques pour être transposés en économétrie et en marketing, et globalement dans un jargon plus commun au service du Data Management moderne. En particulier Doug Laney rédige en 2001 une note de recherche de Gartner, à l'époque META Group (Laney, 2001) où il associe au Big Data trois dimensions : « 3D Data Management : Controlling Data Volume, Velocity, and Variety ». Aucune limite quantitative n'est alors précisée pour que les dimensions soient propres au Big Data, cependant Laney annonce la nécessité pour les entreprises, de e-commerce notamment, de contrôler ces facteurs dont l'ampleur croit rapidement, et donne des solutions concrètes afin de les maîtriser. Gartner affiche alors cette définition du terme Big Data : « Big data is high-volume, high-velocity and

high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making ».

Ces dimensions sont rapidement reprises sous l'abréviation des « 3V », et enrichies sans modération par des « V » supplémentaires, comme la « Valeur », la « Vulnérabilité », mais aussi la « Validité », la « Versatilité », la « Visibilité » ou la « Véracité », voire d'autres dimensions encore moins inspirées, comme la « Complexité ». Les grands acteurs du marché de l'informatique, comme IBM, Microsoft, SAS et leurs observateurs donnent chacun leur propre définition de Big Data, mettant généralement en valeur l'opportunité que le Big Data représente pour un client. La communauté de professionnels Opentracker recense ainsi plus d'une trentaine de définitions du terme[1]. Pour résumer, il s'agit soit d'outils et de processus permettant à une organisation de capturer, créer, traiter et gérer une masse de données importante, soit d'un ensemble de données qui devient tellement volumineux qu'il en devient difficile à travailler avec des outils classiques de gestion de bases de données ou de gestion de l'information. Malgré l'absence de consensus apparent sur l'objet pointé par le terme, l'intention commune dans l'arène commerciale publique est de désigner une évolution technologique actuelle, basculant les entreprises vers des solutions d'une nouvelle génération.

Ce sens est limité par son manque de spécificité et de limites de l'extension du concept. En effet, une définition similaire était d'ores et déjà attribuée (Senge, 1990) à l'information en entreprise dans les années 90, et rien dans la définition ne fait référence à la nature des outils classiques, ni ne quantifie les seuils. Enfin, le concept même d'« ensemble des données » est discutable. Pourtant, cette définition des « 3V » reste actuellement courante, et sera saluée par Francis X. Diebold lorsqu'il reviendra sur son terme en 2012 (Diebold, 2012). Il s'affranchit alors de toute définition quantitative, propose de considérer le Big Data comme un phénomène évolutif, et l'élève au niveau d'une discipline émergeante, intégrant des concepts tels que le Cloud Computing ou les algorithmes massivement parallèles. Ces concepts ne sont pas, selon lui, couverts par les domaines de science existants seuls, comme la statistique ou l'informatique. Diebold s'attribue à cette occasion le mérite de la première définition du terme en citant un ensemble d'apparitions du terme dans des publications académiques et non académiques.

[1] OpenTracker est la version Open Source des outils de gestion de fichiers et de web analytics, animée depuis 2001 par des webmasters et des professionnels de marketing qui les utilise.
https://www.opentracker.net/article/definitions-big-data/

L’un des efforts de formulation les plus intéressants au sein de la communauté scientifique (Boyd & Crawford, 2012) consiste à définir le Big Data comme « un phénomène culturel, technologique et scientifique qui repose sur l’interaction entre :

- **La technologie** : maximisation de la puissance de calcul et de la précision algorithmique dans le recueil, l’analyse, la liaison et la comparaison des grands ensembles de données

- **L’analyse** : représentation à partir de grands ensembles de données pour identifier des tendances (« patterns ») afin de réaliser des déclarations économiques, sociales, techniques et juridiques

- **La mythologie** : croyance largement répandue que de grands ensembles de données offrent une forme supérieure de l'intelligence et des connaissances qui peuvent générer des idées (« insights ») qui étaient auparavant impossibles, avec une aura de vérité, d'objectivité et d'exactitude »

L’intérêt de cette définition est de mettre en valeur non pas les caractéristiques des données (les « 3V », notions relatives et difficiles à cerner), mais bien les processus associés au traitement des données, c'est-à-dire l’optimisation permise par une technologie et les capacités de représentation facilitant une analyse pour une prise de décision. Ces deux caractéristiques ne semblent pas pourtant pointer une nouveauté particulière en dehors des progrès au niveau des outils et processus associés. Cependant, grâce à la troisième caractéristique, cette définition embrasse la prise de conscience et le buzz autour du phénomène, tout en les mettant en perspective de façon critique. Dans la suite de cette thèse, le Big Data portera le sens de cette définition, en considérant le Big Data comme un phénomène sociotechnique dont les avantages réels devraient être critiqués et examinés avec attention.

2 Les enjeux Big Data pour les communautés d'acteurs

Les progrès multidisciplinaires ont généré des attentes et des craintes différentes selon les communautés d'acteurs impliqués. En effet, il est, à ce stade, utile de faire une distinction entre deux familles d'acteurs : les communautés dont l'enjeu principal est de contribuer au progrès et d'en tirer des bénéfices directs (Ecosystème Big Data et recherche, financés notamment par les pouvoirs publics), et les entreprises historiques et les citoyens qui subissent un discours médiatique agressif en faveur du phénomène et qui ont du mal à concrétiser son impact sur leur fonctionnement historique.

2.1 Les pouvoirs publics au service de la recherche et de l'Ecosystème Big Data

Bien que les dispositifs de réflexion et de communication intra-disciplinaires sur les enjeux liés aux données massives ne datent pas de ce siècle, le phénomène semble prendre une ampleur conséquente, et les sujets abordés sont de plus en plus ambitieux et transversaux. Face aux progrès pouvant apporter des réponses à l'enjeu de l'explosion des données, la NSA encourage et supporte la réflexion dans le monde de la recherche et des entreprises au cours d'ateliers de 1995 organisés par CATS en partenariat avec les laboratoires AT&T (Kettenring, 2001; National Research Council, 1996). Ces ateliers regroupent alors des chercheurs issus de grandes universités américaines et quelques acteurs privés comme le laboratoire partenaire ou McKinsey, représentant des disciplines très diversifiées (mathématiques, physique, aérospatial, santé, criminologie, sciences sociales, marketing…), autour des problématiques majeures qui peuvent être résolues grâce à ces progrès en statistiques et en informatique pour traiter les « données massives ».

En 2010, *Communication and Society Programm* de Aspen Institute accueille une communauté de dirigeants et experts sur le sujet, et donne lieu à l'une des publications (Bollier, 2010) des plus complètes résumant les enjeux liés au phénomène du point de vue business, gouvernemental, démocratique et culturel. Il expose notamment les inquiétudes face aux abus potentiels au niveau des données touchant à la vie privée, aux libertés civiles et aux libertés de consommateur. De nombreux domaines de recherche intègrent le phénomène Big Data dans leurs réflexions. La génomique, l'épidémiologie, l'astronomie (Sloan Digital Sky Survey,

analyse de l'imagerie spatiale…), mais aussi l'éducation ou l'histoire utilisent des technologies Big Data. L'utilisation de données massives, notamment en sciences sociales et sciences humaines, est considérée comme pouvant fondamentalement bouleverser la recherche, à condition de donner les outils appropriés aux chercheurs en sciences humaines (Manovich, 2012). Des chercheurs appellent aux investissements (Bertino et al., 2011) et à une création d'interconnexions croissantes (Guo, 2013) dans la communauté R&D afin d'augmenter la qualité, la disponibilité et la traçabilité de l'information scientifique. Les exemples d'applications et de résultats concrets sont nombreux.

En 2012, la maison blanche lance « Big Data Initiative », finançant à hauteur de 200 M$ des projets de recherche et de développement afin d'améliorer les outils et techniques nécessaires à l'accès et à l'organisation de découvertes à partir d'énormes volumes de données digitales. Les communautés visées sont diverses : recherche scientifique, environnement, biomédical, éducation et sécurité nationale. L'un des objectifs de l'initiative est d'établir un état de l'art dans ce domaine très étendu. De nombreux projets de recherche sont lancés par des universités, des acteurs privés ou des institutions comme la Commission Européenne ou l'Etat Français, qui injecte 11,5 millions d'euros dans 7 projets formant le volet Big Data des Investissements d'avenir, pour des constructeurs, des laboratoires, des éditeurs et des intégrateurs. Au-delà de l'ensemble des programmes de développement initiés par les Etats, les institutions gouvernementales constituent aussi un secteur d'activité privilégié pour l'écosystème Big Data. En effet, le développement des applications au service de l'urbanisation (« Smart Cities »), de la surveillance ou bien de la gestion des foules en situation de crise, illustrent les applications possibles pour les institutions publiques.

Globalement, le début des années 2010 est marqué par des démarches larges et coercitives entre les disciplines scientifiques et portées par les Etats. Ces derniers, chacun pour son compte, deviennent des clients du phénomène, mais aussi des fournisseurs de ressources pour le progrès à long terme, notamment en cofinançant le développement de l'Ecosystème Big Data (Turck, 2016), constitué d'une communauté d'organisations privées (start-ups, grands groupes industriels…) et publiques (plateformes open source…) ayant pour objet de fournir aux entreprises des prestations et des moyens technologiques liés au phénomène Big Data (voir Annexe 3 - Ecosystème Big Data). Par ailleurs, les institutions étatiques constituent des garde-fous face à l'expression de craintes des dérives du phénomène, que ce soit à travers des instances

de régulation habituellement actives dans les différents secteurs ou des dispositifs plus récents, comme celui de la Régulation de la Protection des Données Personnelles.

2.2 Le buzz face au grand public et aux entreprises

Jusqu'en 2011, Big Data apparait dans les publications non académiques selon une approche principalement basée sur sa caractéristique « Volume », appréhendée du point de vue de l'augmentation de sources de données différentes, comme en témoignent, en 2008, l'article « The Petabyte Age » publié dans le magazine Wired, ou bien, en 2010, le rapport dirigé par Joe Hellerstein publié dans The Economist « Data, data everywhere ». De même, IDC attire l'attention sur le volume croissant de données : cette société d'études et de conseil démontre en 2007 l'écart entre une croissance rapide de l'univers digital et une croissance lente des ressources humaines et financières pour le gérer. Le ton de ces publications est partagé : d'un côté le Big Data pourrait être une source de revenus (non estimé par les auteurs), mais, d'un autre côté, ce déluge informationnel provoquerait des problématiques liées à la sécurité des données et aux ressources (informatiques, applicatives et humaines).

Cependant ces réflexions semblent impacter peu l'intérêt public pour le phénomène Big Data, d'après l'observation des courbes de tendances représentant la recherche de mots clés sur Google (voir Figure 1). Alors que l'intérêt pour le « Data Mining » décroit progressivement, les prémices du buzz émergent avec l'apparition du « Cloud » et du framework « Hadoop », posant le socle technologique qui permet de résoudre les problématiques informatiques. Il faudra attendre la traduction des enjeux en termes d'opportunités concrètes et de menaces pour que le terme « Big Data » se propage auprès des utilisateurs de Google, avant de se stabiliser depuis 2014 (Turck, 2016) au profit de l'apparition de termes plus précis, comme le Machine Learning, le Deep Learning ou encore la Data Science (voir Annexe 4 - Data Science et algorithmes). Ces concepts font référence plus précisément aux moyens humains et aux méthodes de conception algorithmique. Dernier terme en date, l'Intelligence Artificielle, revient au galop dans les discours médiatiques à partir de 2015-2016 pour englober les concepts algorithmiques et les opposer au socle technologique qui les rend praticables : le barycentre du terme « Big Data » se décale alors sur ces concepts plus techniques (nous continuons ici à l'utiliser au sens plus large de phénomène global).

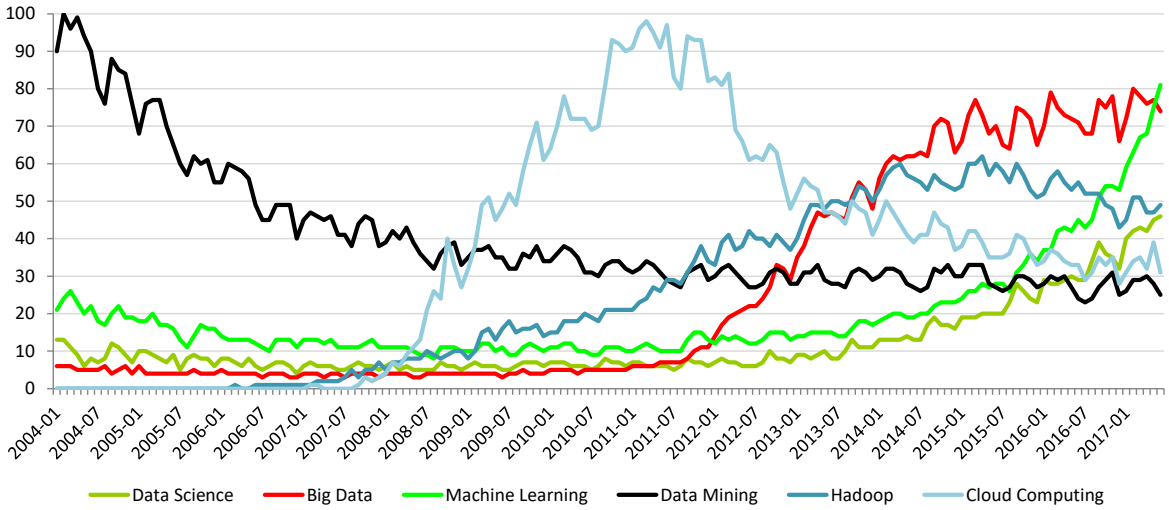


Figure 1 – Evolution de l'intérêt pour les termes recherchés sur Google – *Construit avec Google trends, 25/06/2017*

L'analyse de l'intérêt pour le terme « Big Data » conduit à un constat clair : le terme connait un grand succès à partir de l'été 2011. Le premier article grand public, paru dans le Wall Street Journal, renvoie au lancement d'un fond d'investissement de 100 millions de dollars destiné à financer la croissance des start-ups Big Data, par Accel. Cette société de capital-risque basée à Palo Alto opère aux Etats Unis, au Royaume Uni, en Inde et en Chine. Elle est connue pour être le deuxième actionnaire de Facebook après Mark Zuckerberg, ayant investi dans l'entreprise dès 2005 à l'époque où elle s'appelait encore « Thefacebook », mais aussi pour ses participations dans les entreprises comme Rovio Entertainment (Angry Birds), Dropbox ou Cloudera, qui commercialise Hadoop. Puis, de plus en plus de publications commerciales, conférences et livres blancs des fournisseurs de logiciels (EMC, SAS...) vantent les opportunités business des analyses permises par les outils de dernière génération, dits Big Data Analytics. Les premières expériences se propagent dans les médias, parfois assez alarmantes, comme c'est le cas de Target, acteur de grande distribution américain qui détecte qu'une cliente est enceinte avant sa famille, en analysant sa consommation pour lui proposer des produits dédiés.

Les médias et les sites de référence pour l'audience de masse s'emparent du sujet, qui fait régulièrement les unes, les numéros spéciaux, ou des articles de choc dans la presse grand public à l'international. L'intérêt pour le Big Data croit, et les blogs et magazines spécialisés fusent, y compris en France (Decideo.fr, Bigdatafrance.wordpress.com, Blog.businessdecision.com...). Cependant, le discours reste assez ambigu : d'un côté, l'analyse massive des données est présentée comme une opportunité d'innovation sans précédent et d'évolution du confort des

citoyens et des consommateurs, et d'un autre côté un phénomène anxiogène de « Big Brother » lui est rapidement associé, plus particulièrement lorsqu'il s'agit de traiter les données personnelles. Parmi les acteurs pointés du doigt, les gouvernements et les GAFA (Google, Amazon, Facebook, Apple), mais aussi les entreprises plus traditionnelles qui pourront utiliser les données issues des objets connectés, des réseaux sociaux ou encore les consommations afin de manipuler les citoyens.

Au-delà du grand public, la publication en 2011 du rapport de recherche du cabinet McKinsey (Manyika et al., 2011) produit un effet significatif sur la communauté des entreprises : dépassant les considérations de volume, McKinsey Global Institute, largement soutenu par des recherches scientifiques, s'intéresse en profondeur aux enjeux économiques du phénomène et aux conséquences sur le management et la gouvernance. Balayant tous les secteurs (voir Figure 2), le cabinet identifie les différences de potentiels pour chacun et des leviers de création de valeur. Les leviers les plus approfondis sont alors la santé et la distribution aux Etats Unis, les services publics en Europe, l'industrie en général et les services basés sur la localisation individuelle. La gestion du Big Data est alors annoncée comme potentiellement un avantage concurrentiel significatif, plus ou moins accessible d'un secteur à l'autre, notamment en fonction des barrières existantes à l'heure actuelle, comme la maturité et les investissements dans les systèmes d'information ou encore la culture du secteur. Le rapport chiffre par ailleurs une manne potentielle de 140 000 à 190 000 postes marqués par une profonde connaissance analytique : le métier de Data Scientist sera nommé comme le plus sexy du siècle un an plus tard par Harvard Business Review (Davenport & Patil, 2012). Le cabinet poursuit une activité de publication et de communication sur le phénomène Big Data, alimentant l'intérêt des entreprises, notamment à travers des articles complémentaires du MGI (Brown et al., 2011).

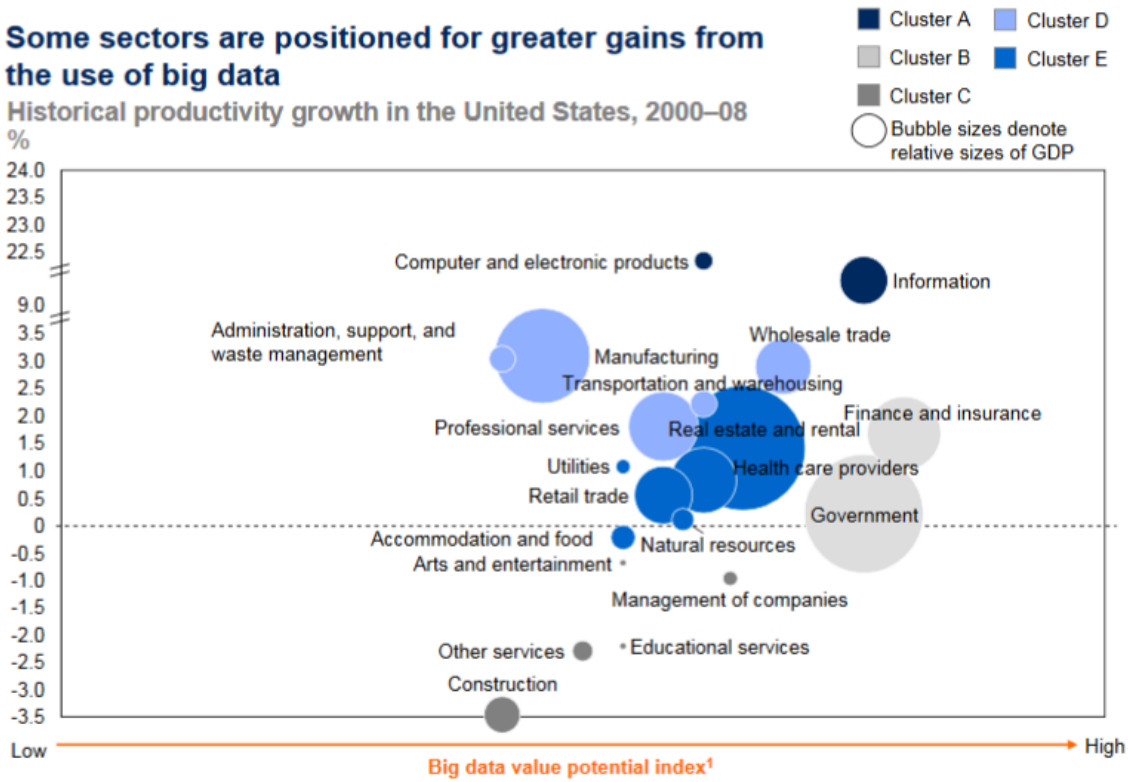


Figure 2 – Les bénéfices de l'usage du Big Data attendus par secteur – *Source : Big Data: The next Fronteer for innovation, competition and productivity, McKinsey & Company, 2011*

Les opportunités pour les entreprises semblent innombrables (Bertino et al., 2011), et l'intérêt des dirigeants annoncé pour le Big Data est vif aux Etats-Unis. Marketsansmarkets estime, en avril 2016, que le marché du Big Data va croitre de 29 Milliards de dollars en 2016 à 67 Milliards en 2021. Le marché français à lui seul est estimé à 1.9 Milliards en 2015[2], avec une croissance de 12% entre 2016 et 2018, décollant à son tour dès 2012 selon les sondages d'IDC[3]. L'index de maturité Big Data a augmenté de 66% entre 2012 et 2014, et les initiatives ont été multipliées par 6 : 43% des Directions Informatiques ont un projet Big Data en 2014, contre 7% en 2012. La nouveauté de ce sondage consiste à montrer l'intérêt des acteurs métier, dans ce cas-là du Marketing, pour le phénomène. Un tiers des dirigeants marketing sondés jugent qu'il n'est pas nécessaire d'impliquer les DSI dans ces projets, avant tout analytiques. Les

[2] Source : MARKESS, 15 mars 2016, *« Analytique, big data & gestion des données : un marché de 1,9 milliard d'euros en France »*, http://blog.markess.fr/2016/03/analytique-big-data-et-gestion-des-donnees-un-marche-de-pres-de-2-mds.html

[3] Source : ICD, 10 novembre 2014, « Le Big Data Index EMC/IDC: Le Big Data décolle enfin en France! », http://france.emc.com/about/news/press/2014/20141010-01.htm

experts IT et marketing montrent des intérêts divergents face au phénomène, même si les deux s'accordent sur un potentiel significatif de contribution à la performance de l'entreprise.

Dès le début du buzz en 2011, le discours sur la création de valeur grâce au Big Data pointe l'opportunité dégagée par la prise de décision « data-driven », à condition d'adapter le style managérial et investir dans les ressources humaines et technologiques capables d'analyser des données et les traduire en information économique utile pour la génération de connaissances et la prise de décision. L'efficience de la transformation des données constitue ainsi un avantage concurrentiel, que ce soit pour la conception et la mise en œuvre de nouveau business modèles de l'entreprise (nouveaux produits et services), ou bien pour la recherche de performance sur les activités existantes. L'impact du Big Data sur la performance des entreprises semble d'ailleurs se vérifier, comme l'illustre, entre 2006 et 2012, une croissance supérieure de la productivité (4-5%) et des ventes (2%) pour les entreprises qui ont investi dans l'embauche d'ingénieurs spécialisés dans le Big Data (McAfee & Brynjolfsson, 2012; Tambe, 2012). Un autre sondage (Brynjolfsson et al., 2011) auprès de dirigeants de 330 compagnies Américaines, recoupé avec les résultats financiers et opérationnels extraits des rapports annuels et des sources indépendantes, montre que plus les entreprises se définissent comme dirigées par les données, meilleures sont leurs performances. Elles ont alors des gains de productivité de 5 à 6%, et une amélioration de l'utilisation des actifs, du ROE et de la valeur de marché, et ce sans effet de causalité inverse. Plus généralement, de nombreux articles prônent l'efficacité du Data-Driven Decision Management (DDDM), c'est-à-dire la pratique managériale basée sur l'analyse des données et non pas sur l'intuition du décideur.

Cependant, ces sondages ne démontrent ni l'aspect inédit du sujet, ni le lien causal entre les résultats des projets en Data Science et la performance des entreprises. Les entreprises restent réticentes à fournir les résultats de leurs projets et initiatives, contrairement aux institutions de recherche. En effet, les cas concrets sont distillés dans une démarche de communication, et non pas de retour d'expérience ni de pédagogie. La publication des études de cas présentant des résultats quantifiés est rare dans les premières années du buzz, en dehors de cas (Bartram, 2013; P. Simon, 2013) en Web Marketing ou des estimations de potentiels en Santé sur l'usage des données en temps réel. Seuls quelques auteurs dans le monde du conseil proposent des recueils plus pédagogiques de cas d'usages dans différents secteurs (Biernat & Lutz, 2015; Ezratty, 2017), mais beaucoup plus tard et sans confirmer dans la majeure partie des cas la génération de la valeur. Ces ouvrages, publiés de façon autonome, ont alors l'avantage de démystifier

certains aspects opérationnels des projets data, mais peuvent difficilement servir de référence. Au-delà des effets de communication sur le caractère innovant des usages embarquant l'Intelligence Artificielle, peu de publications abordent la nouveauté induite concrètement par le Big Data sur la valeur générée pour les secteurs, plus précisément son impact sur les business modèles ou la performance de l'entreprise, et ce d'autant plus que si l'on s'intéresse aux sociétés françaises. Cette absence de retours d'expérience concrets, plusieurs années après le début du buzz Big Data, instille un doute quant au caractère révolutionnaire du phénomène Big Data et à la pertinence des usages potentiels en entreprise, conformément au discours médiatique tenu et à l'ampleur des investissements.

3 Une prise de position des SIC au cœur du phénomène Big Data

Deux communautés d'acteurs aux enjeux différents se font ainsi face dans le phénomène Big Data. Cette dichotomie complexe, marquée par des intérêts contradictoires, mixtes et parfois opportunistes (obtenir des financements, générer des bénéfices, imposer une image d'innovation par la communication institutionnelle, protéger son pouvoir de décision et son marché face aux concurrents et nouveaux entrants…), se heurte à un débat de fond sur la pertinence de l'invasion du champ de la prise de décision humaine par les solutions embarquant l'intelligence artificielle. Peu d'acteurs sont aujourd'hui légitimes et assez multidisciplinaires pour contribuer à apaiser le débat de façon pédagogique et contribuer à établir le lien entre les hommes et les données. Les Sciences de l'Information et de la Communication apparaissent alors comme un terrain de médiation doté d'un cadre assez transversal, souple, et conceptuellement armé pour assumer ce rôle et permettre la génération de sens et de valeur dans les projets data en entreprise.

3.1 Débat épistémologique

Le Big Data comme phénomène social (Boyd & Crawford, 2012) comprend la notion de « mythologie » à travers la propagation de l'idée que le traitement de grands volumes de données permet de construire des connaissances plus vraies, plus pertinentes et plus objectives. Cette dernière caractéristique est illustrée par un débat épistémologique (Kitchin, 2014; Noyer & Carmes, 2014) qui dépasse les milieux scientifiques et attise les désirs d'intégrer ce phénomène au cœur de la prise de décision, voire l'automatiser. En effet, deux visions radicales s'opposent.

La première, assimilable à un « style de pensée » (Fleck, 1935) en faveur du Big Data prêche que la donnée est le moteur principal de la découverte sans hypothèse préalable (H. J. Miller, 2010), que le traitement de la totalité des données est nécessaire pour une analyse fine, voire individuelle, des éléments, et que le besoin actuel cible la prise de décision qui ne nécessite qu'une analyse inductive par corrélation sans recherche de causalité. Cette vision est portée par les adeptes de la e-science (Bohle, 2013; Hey et al., 2009; Hey & Trefethen, 2005), démarche

appliquant systématiquement les avancées technologiques à l'ensemble de la science, commencée dans les années 90, mais aussi par des journalistes influents comme Chris Anderson du magazine Wired (Anderson, 2008). Elle est élevée au grade de 4ème paradigme formulé par Microsoft (Hey et al., 2009) en 2009 et basé sur le travail de Jim Gray. Pour 4 domaines scientifiques (terre et environnement, santé et bien-être, infrastructure de recherche et communication scientifique), la méthode basée sur la capture, le traitement et l'analyse de données massives illustre les avancées réalisées et possibles. Dépassant les méthodes scientifiques existantes (méthodes empiriques par l'expérience, méthode systémique par l'analyse, et la méthode informatique par simulation), le livre exprime une invitation à utiliser l'informatique dans la recherche scientifique afin de répondre aux problématiques impliquant un traitement massif de données. La transformation de la pratique de recherche par cette méthode implique notamment un croisement croissant des domaines de recherches pour accélérer la découverte de nouvelles connections. Jim Gray prédit un intérêt sans précédent de la part de la science pour cette méthode (et donc pour ces technologies), plus intense que dans le domaine industriel et commercial, étant donné que ce dernier ne nécessite qu'un ensemble de données (valeurs, noms, métriques...) très limité par rapport aux domaines scientifique comme la génomique ou l'aérospatial.

La seconde vision (Frické, 2014; Graham, 2012), conservatrice, considère que la découverte est avant tout issue d'hypothèses bâties sur des théories, que les méthodes statistiques d'échantillonnage actuelles sont éprouvées et suffisantes, et que toute connaissance doit être fondée sur une logique causale exclusivement. Cette controverse sur la légitimité des méthodes scientifiques se cristallise autour de certains caractères dichotomiques[4] : corrélation et causalité, data-driven et hypothesis-driven, totalité et échantillon, décision et connaissance, ou encore absence et nécessité de théorie. La vision conservatrice se radicalise jusqu'à la dénonciation d'une véritable idéologie nouvelle qui consiste à opposer à la construction scientifique de connaissances l'objectivité totale, dénuée de toute idéologie face au réel (Ouellet et al., 2014).

Le débat porte sur la génération et la gestion de connaissances scientifiques (Rheinberger, 2014), mais aussi la sphère politique et sociale ainsi que l'économie de l'information (Porat, 1977). En effet, le rejet de l'algorithme, en tant qu'entité quasi personnifiée au fonctionnement opaque (dit « boite noire ») et à impact allant à l'encontre des libertés démocratiques (Cardon,

[4] Nesvijevskaia, Anna. 28 mai 2014. « Epistémologie - Big Data ». Présentation au séminaire « Big Data, fouille de données dans les domaines scientifiques », Conservatoire National des Arts et Métiers.

2015), marque fortement les disciplines sociales en France. Les algorithmes sont perçus comme porteurs de projets politiques qui tendent à individualiser les sociétés, et les concepteurs des algorithmes (peu identifiables en dehors des GAFA et des Etats, et aux responsabilités peu définies) sont opposés à leurs utilisateurs. Paradoxalement, ces derniers se voient enfermés dans des modèles comportementaux portés par les algorithmes de profilage sous prétexte d'une personnalisation (d'offre de crédit, de contenu à lire, de trajet à suivre...) : par corrélation à des comportements similaires, l'algorithme s'affranchit des normes sociales issues d'un travail de conventions et d'une vision par la moyenne, pour générer de nouvelles « normes » plus ciblées, et mouvantes dans le cadre d'utilisation d'algorithmes auto-apprenants (Rouvroy & Berns, 2013). Les chercheurs craignent alors avoir à faire à une « colonisation de l'espace public par une sphère privée hypertrophiée […] à l'ère de la gouvernementalité algorithmique », c'est-à-dire d'un « certain type de rationalité (a)normative ou (a)politique reposant sur la récolte, l'agrégation et l'analyse automatisée de données en quantité massive de manière à modéliser, anticiper et affecter par avance les comportements possibles ». Le débat touche ainsi non seulement les GAFA, mais aussi les entreprises existantes pour qui le Big Data représente des opportunités spécifiques et se traduit par la mise en œuvre de projets, contribuant à répondre à leurs enjeux stratégiques particuliers, notamment dans un contexte d'accélération de la prise de décision.

Le renouvellement des Sciences de l'Information et de la Communication par le Big Data a d'ores et déjà été mis en perspective (Noyer & Carmes, 2014) à travers l'impact du Data Mining sur les pratiques théoriques, notamment suite à ce débat épistémologique. La mobilisation des concepts tels que l'approche sociologique de la théorie acteur–réseau (Akrich et al., 2006) ou l'agencement collectif d'énonciation (Deleuze & Guattari, 1972), permet alors de s'émanciper des principes dogmatiques des disciplines scientifiques et des pratiques. Il vise à mettre en perspective les enjeux complexes, ouverts et évolutifs liés à la production de savoirs à travers l'enrôlement produit par des acteurs humains et non-humains (objets et discours) qui constituent une méta-organisation, un réseau hétérogène où se traduit et se stabilise un ensemble de positions et de savoirs. Cette approche inscrit les disciplines transversales au cœur de la génération de valeur au sein des organisations hétérogènes à travers un processus de fabrication de faits, étroitement liés à la robustesse du dispositif qui le garantit. Ce processus de fabrication est alors constitué d'un ensemble de controverses qui permettent d'élaborer les faits à travers la recherche dialectique de consensus.

3.2 Spécificités de l'angle de vue

Il ne s'agit plus d'établir des principes techniques ou quantitatifs en termes de génération de valeur pour une organisation, mais bien de se pencher sur la nature et le dispositif, indissociables, de production de cette valeur. En effet, si les travaux de recherche sur les progrès techniques et analytiques attribués au phénomène sont riches dans les sciences dures et les Sciences de Gestion, les interactions entre ces disciplines ainsi qu'entre le monde des entreprises et l'Ecosystème Big Data, sont peu étudiées. L'explosion du phénomène crée l'urgence d'une prise en main de ce sujet par les Sciences de l'Information et de la Communication, par nature transdisciplinaire et d'ores et déjà ancrée dans les organisations, notamment à travers l'Intelligence Economique ou le Knowledge Management. Identifiés comme leviers de transformation digitale (Dudezert, 2018; Chastenet de Géry, 2018), ils semblent pourtant à la traine en termes d'implication dans le phénomène Big Data au sein des entreprises, derrière les DSI ou le Marketing. En effet, face au nombre de publications de recherche techno-centrées sur le Big Data, à la promesse de gain financiers et performatifs pour les entreprises, à un nombre réduit de travaux anthropocentrés, à la transdisciplinarité du phénomène et à la complexité de l'écosystème en pleine croissance, les Sciences de l'Information et de la Communication peuvent permettre d'aborder de façon pertinente la génération de sens, directement ou indirectement valorisable, sous l'angle du dispositif complet.

La notion de génération de sens est confrontée au débat épistémologique lié au phénomène Big Data sous plusieurs angles.

D'une part, la multitude des acteurs impliqués, qu'ils soient issus de l'écosystème Big Data, hétérogène et instable, ou alors des organisations existantes, fait appel à la notion d'intersubjectivité, à une construction de sens à travers la communication entre individus et entre collectivités avec reconfiguration technologique et sociale. Selon la théorie de l'interactionnisme, cette construction de sens s'accompagne d'une transformation de l'individu, en donnant ainsi une place privilégiée à l'acquis face à l'inné. Or, le mythe autour du phénomène Big Data comporte la recherche d'objectivité, que ce soit en termes de nature des données ou de la construction et de la diffusion de résultats des algorithmes, ce qui empêche *a priori* cette construction des individus. Ce point est éclairci par la mise en évidence d'un processus collectif d'objectivation, comportant des mécanismes de gouvernance non neutres (Bonenfant et al., 2014).

D’autre part, la génération de sens comporte la notion de causalité, difficile à prouver à ce jour dans le cadre de l’élaboration d’indicateurs de pilotage en gestion classique, et encore plus difficile dans le cadre de la Data Science (Provost & Fawcett, 2013a; Templ, 2012), marquée par des risques de sur-apprentissage, des facteurs de confusion ou par la complexité de certains algorithmes. Cette problématique conforte la nécessité de mettre en lumière les mécanismes humains et non humains de choix des données, des indicateurs et des usages, et plus généralement un nouveau rapport entre la génération d’hypothèses, de connaissances et de leviers d’action.

Enfin, le phénomène Big Data renouvèle les défis d’ores et déjà abordés dans la discipline, comme la surcharge informationnelle et l’économie d’attention, la structuration et la sélection de l’information, ou encore la transformation des interfaces homme-machine. Ces défis s’inscrivent dans la transformation numérique des organisations (Boustany et al., 2014), que ce soit à travers le processus, les compétences, ou les artefacts technologiques de transformation des données en éléments utiles à l’homme. L’algorithme semble constituer un nouveau moyen de médiation entre l’homme et la donnée, perturbant les modèles d’ores et déjà observés.

Dans ces conditions, les Sciences de l’Information et de la Communication doivent pouvoir faire face à la perturbation induite par le phénomène sur les informations, outils et savoirs utilisés dans le cadre de la construction de sens en amont de la prise de décision. Et, inversement, aborder le sujet sous un angle fondamentalement anthropocentré semble constituer une nouveauté à ce stade de développement du phénomène Big Data.

***

Ainsi, le phénomène Big Data s’enracine dans des besoins politiques, à travers l’évolution des méthodes statistiques et des outils de recensement, ainsi que dans les progrès des technologies militaires puis civiles, et s’inscrit dans une mutation longue des technologies et des approches analytiques. Il est marqué dès 2011 par un buzz médiatique qui fait naître des attentes fortes de la part des entreprises, et ce en absence d’un terrain de médiation, pouvant être joué par les SIC, propice à une meilleure appropriation du potentiel de valeur. Cette appropriation de valeur est clé dans la problématisation de ces travaux de recherche.

Première partie :
Problématique et cadre conceptuel

1 Problématique

« Pour un esprit scientifique, toute connaissance est une réponse à une question. S'il n'y a pas eu de question, il ne peut y avoir connaissance scientifique.

Rien ne va de soi. Rien n'est donné. Tout est construit ».

***Gaston Bachelard**, Philosophe*

Le Big Data, phénomène sociotechnique à définition instable, plonge ainsi ses racines dans une évolution millénaire des approches de la donnée et de progrès technologiques. L'étude du contexte de son développement conduit à deux constats principaux. D'une part, le terme naît dans le milieu scientifique et connait une propagation interdisciplinaire, en particulier dans la recherche en informatique, en statistique, en économétrie et en marketing. Les usages auxquels il répond sont alors ancrés dans des projets de recherche fondamentale et appliquée et de gouvernance publique, qui ont besoin de s'affranchir des limites des méthodes de traitement de données existantes. D'autre part, il est sujet à un « buzz » à partir de 2011, poussé par la dynamique de croissance de l'Ecosystème Big Data qui draine des investissements étatiques et privés et tire le système éducatif pour la mise sur le marché de nouvelles compétences. Les acteurs hétérogènes de l'écosystème orientent alors leur discours sur les opportunités pour un plus grand nombre d'organisations, non demandeuses mais intéressées, et plus particulièrement des entreprises. Pour ces dernières, le phénomène semble atteindre dès 2014 un palier, où l'utopie « Big Data » se confronte aux résultats, discrets si ce n'est décevants, des projets sur le terrain.

Dans l'hypothèse initiale où les projets data constituent des dispositifs d'interaction caractérisables et pertinents, c'est-à-dire pouvant contribuer à une modulation des usages métier générant de la valeur, deux pistes simples s'ouvrent pour la compréhension de ces échecs. La première est une induction en erreur, par le buzz, sur la nature ou l'ampleur de la valeur potentielle pour les entreprises. Dans ce cas, il faut mettre en évidence l'écart entre les résultats attendus et les résultats atteignables grâce à la Data Science, ainsi que les modalités de convergence sur des usages réalistes. La seconde, cumulable avec la première, est l'inefficacité du dispositif projet : il est alors nécessaire de pointer les difficultés de cette convergence et les

résistances à la construction des résultats. Dans les deux cas, il est indispensable de se pencher sur les dispositifs projet data, sur leur processus de génération de valeur et ses caractéristiques, ainsi que sur les médiations entre acteurs hétérogènes impliqués. Que peut-on espérer plus largement de cette forme d'interaction et quels facteurs favoriseraient la satisfaction des attentes des entreprises ?

1.1 Les processus propres aux dispositifs projet data sont-ils efficients ?

Le dispositif projet data fait d'ores et déjà l'objet de nombreuses études, et donne lieu à l'établissement de modèles de référence qui tiennent compte des spécificités de ces projets. Plus particulièrement, le modèle de processus CRISP_DM (Chapman, 1999; Shearer, 2000), décrivant les processus et les cycles de vie intrinsèques des projets data, met en évidence l'ensemble des activités, tâches et résultats produits au cours d'un projet. Il distingue la phase de conception de l'algorithme, qui inclut la définition de l'objectif métier, et la préparation du déploiement d'usages, sous forme de solutions algorithmiques ou de connaissances utiles, évaluées comme opérationnelles pour l'entreprise. Ces modèles, qui semblent autocentrés sur le travail de production algorithmique par les Data Scientists et hors sol vis-à-vis de la pratique métier, sont complétés par des concepts plus orientés sur la valeur et la gestion de projet issus des Sciences de Gestion (Camiciotti & Racca, 2015; Marr, 2015). Cependant, la génération de sens et la dimension relationnelle, notamment au cours de l'appropriation des résultats par les acteurs métier dont la prise de décision est potentiellement impactée par la mise en œuvre des résultats de ces projets, restent peu étudiées. Le dispositif semble pourtant partager des caractéristiques communes avec d'autres projets liés à la transformation numérique (Boustany et al., 2014; Noyer & Carmes, 2014) où l'interaction et la capitalisation de connaissances sont clés. Quels modèles de référence sont appliqués sur le terrain ? Sont-ils considérés comme assez robustes par les Data Scientists et les experts métier pour générer des résultats utiles pour la pratique métier ? Peut-on capter, à travers l'étude de projets jugés comme réussis, des pistes de remise en cause ou d'enrichissement de ces dispositifs qui favoriseraient leur efficience ?

1.2 Quelle est la valeur générée par ces projets ?

En pratique, l'étude de ce lien indissociable entre le dispositif et la valeur qu'il génère se heurte à l'importance de la contextualisation du projet au sein d'une entreprise : les usages possibles mobilisant l'Intelligence Artificielle semblent innombrables, or chaque usage peut avoir une

utilité très variable selon l'entreprise qui le met en œuvre. Cela induit une difficulté à établir des indicateurs de valeur génériques, moins arides et indirects que l'impact comptable sur les bénéfices de l'entreprise. Face à la richesse des modèles de mesure et de pilotage de la valeur quantitative et qualitative, souvent mixtes sur le terrain et dépendants des fonctions de l'entreprise et de la subjectivité de la stratégie voulue par les décideurs, une simplification s'impose. En effet, les indicateurs clés employés en entreprise se déclinent à l'échelle de systèmes plus réduits, composés d'acteurs réunis autour d'un ensemble d'activités communes et homogènes. Ces sous-systèmes constituent des modèles logiques, inscrits dans un processus de causalité (Lebas, 1995) décrivant la production de produits ou de connaissances à partir de ressources. Cette mise à plat systémique permet l'étude de l'action située (Laville, 2000) en se basant sur l'activité des agents dans un environnement pouvant être facilement représenté et jugé. Dans ce cadre d'analyse, en quoi le modèle logique d'un sous-système traitant est-il différent entre son état initial (avant le projet data) et final (après le projet data) ? Autrement dit, peut-on détecter des impacts directs d'un projet data sur l'entreprise à travers un changement des pratiques métier au sein de la fonction responsable de la mise en œuvre de ses résultats ?

Bien que ce cadre d'analyse soit approprié pour confirmer ou infirmer l'existence d'effets immédiats d'un projet data, son application brute, quantitative, le réduirait à une approche financière par le retour sur investissement. Or, l'effet des projets data semble plus subtil dès lors qu'on s'intéresse à l'écart entre la valeur espérée et la valeur potentielle (c'est-à-dire le niveau de maturité des acteurs impliqués pour définir et anticiper la valeur du projet), mais aussi aux externalités positives des projets, à la génération de valeur à plus long terme, plus indirecte. Le discours prometteur sur la création de connaissances à travers la Data Science dans des domaines professionnels variés, la capacité d'innovation induite par ces projets, ou encore l'urgence d'internalisation de nouvelles compétences spécifiques à l'exploration des données ne permettent pas d'exclure *a priori* ces axes de création de valeur. Peut-on observer sur le terrain des éléments qualitatifs permettant d'appréhender cette génération de valeur indirecte et complexe à mesurer ? Enfin, la donnée, matière première de ces projets et, potentiellement, des usages qui en découlent, est d'ores et déjà identifiée, notamment par la CIGREF[5], comme l'actif incorporel principal des entreprises. Sa valeur dépasse dès lors les usages circonscrits dans des pratiques isolées de certains métiers. Or, la valorisation de cet actif s'appuie sur une meilleure

[5] « Enjeux business des données ». 2014. CIGREF. https://www.cigref.fr/rapport-cigref-enjeux-business-des-donnees.

gestion de la qualité des données. Les projets data constituent un vecteur de transformation des données de l'entreprise, mais contribuent-ils à la valorisation des données au-delà de l'usage visé ?

Ces travaux de recherche visent ainsi à identifier les différents paramètres, quantitatifs mais aussi qualitatifs, qui permettent de pointer des indicateurs de valeur, qu'elle soit directement liée à l'usage immédiat des résultats du projet ou s'inscrive dans des modalités de mesure plus complexes et plus long terme.

1.3 Quelle est la nature de la médiation humaine dans les projets de Data Science ?

Un modèle impacté par un projet data implique des enjeux cognitifs nouveaux dans la mesure où les praticiens, possédant des compétences métier, ont besoin de s'approprier des paramètres inédits impactant soit leurs ressources, soit leurs activités, soit leurs connaissances et produits, en particulier informationnels. Ces différents paramètres reflètent le résultat d'un processus de choix stratégique subjectif (Atamer et Calori, 2003) et de priorisation. Une analyse de la relation entre différents acteurs basée sur la sociologie de l'innovation et sur la théorie de l'acteur-réseau se prête à la lecture des logiques d'actions dans le système de relations entre entités hétérogènes composé d'une «méta-organisation d'humains et de non humains » liées les unes aux autres et agissant comme intermédiaires (Akrich et al., 2006). Cette approche théorique permet de concevoir l'étude des projets sous l'angle du dispositif foucaldien incluant des pratiques, des discours sur des pratiques, des objets techniques, et des relations qui se concrétisent par les interactions entre acteurs. Il s'agit de pointer « les conventions, les négociations, les compromis préalables » (Desrosières, 2008), et ce plus particulièrement au sein des projets data qui n'ont pas encore été décrits sous cet angle. Ce processus d'appropriation est-il assez efficient et facilité par le dispositif de projet data ? Le dispositif nécessite-t-il d'être enrichi avec des facteurs favorisant la médiation entre l'homme et la donnée au service d'une meilleure appropriation des résultats et d'une génération de valeur supérieure ? L'absence de l'appropriation défavoriserait-elle l'engagement des acteurs métier historiques au profit des acteurs concurrents de l'Ecosystème, dans le cas où celui-ci s'avèrerait véritablement disruptif ?

Les enjeux liés à l'appropriation de ces paramètres sont multiples pour une exploitation finale des résultats de projets data qui conditionne la génération de valeur. Le changement des ressources de l'activité (volume et nature de données d'entrée, technologies disponibles,

nouvelles compétences sur le marché comme les Data Scientists…), implique une montée en compétences pour maîtriser leur utilisation et appréhender les biais pour justifier la prise de décision. Les limites des pratiques liées au manque, à la qualité et à l'éthique de l'usage des données accessibles (Bonenfant et al., 2014; Boyd & Crawford, 2012) ainsi que leur conditionnement à un changement culturel et organisationnel profond des métiers en entreprises sont d'ores et déjà soulevées (Bollier, 2010; Manovich, 2012; Manyika et al., 2011), ce qui justifie un questionnement plus approfondi dans ce sens. Par ailleurs, la modification des activités (nature, ordre, méthode, indicateurs de mesure…) et des processus de formation du savoir peut impacter l'organisation du système de décision, et, plus spécifiquement, l'introduction d'un algorithme, acteur « non-humain » (Latour, 1994 ; McMaster et Wastell, 2005) supposé objectif dans les modèles de décision repose les questions de reconfiguration technologique et sociale, ainsi que de responsabilité. Inversement, l'observation des corrélations sur des données passées peut conduire à l'application d'une logique inductive qui amplifierait l'inertie du système en privilégiant la normalisation de la prise de décision ou la production au détriment de l'innovation. Comment les acteurs métier impliqués dans les projets Big Data intègrent-ils ces évolutions de leurs activités et maîtrisent les risques et les incertitudes induites ? Ces changements risquent-ils de perturber le processus de création des savoirs et d'entrer en contradiction avec des exigences théoriques et règlementaires de causalité ? Existe-t-il, derrière ces projets perçus comme innovants, un risque de compromission de la capacité d'innovation métier ? Quelle place est laissée à la découverte, à la capitalisation et à la valorisation des connaissances métier ? Enfin, comment les entreprises préservent-elles ce lien nécessaire de causalité pour justifier une décision et en porter la responsabilité ? L'appropriation des nouveaux paramètres dans ces organisations possède-t-elle des caractéristiques d'une rupture épistémologique dans le sens de « polémique contre l'immédiat » (Bachelard, 1938), voire de changement de paradigme (Gray, 2009 ; Kuhn, 1996) ?

***

Pour répondre à ces questions, marquées par ces dépendances fortes et multiples, ces travaux de recherche s'attachent à décomposer l'impact du phénomène Big Data en entreprise selon plusieurs axes : l'étude des processus projet data, la génération de valeur qui en découle, dont l'impact sur la qualité des données, et enfin la médiation entre l'homme et la donnée au cours de ce processus. Chacun de ces axes et les concepts clés associés font l'objet d'une mise en perspective théorique initiale avant d'être confrontées à la pratique. Cette confrontation a lieu

sous la forme de recherche-action et aboutit à une étude de cas multiples. Adoptant alors une approche **anthropocentrée**, la démarche de recherche favorise l'étude des interactions humaines sur des approches financière, analytique ou technologique seules, afin d'établir des liens interdisciplinaires sur cette base commune et de proposer un modèle global, à la fois théorique et opérationnel, qui articulerait les axes de l'analyse et les facteurs favorisant l'efficience du dispositif projet data en entreprise.

2 Plan de thèse

Une fois posée l'introduction du contexte, pointant les enjeux naissants et les concepts satellites, historiques et plus récents, qui gravitent autour du phénomène Big Data, la problématique appelle des réponses concrètes et robustes pour faire la part des choses entre le mythe et la réalité, en se détachant des discours médiatiques et technico-mathématiques pour se concentrer sur la génération de valeur et de sens pour les entreprises. Pour ce faire, ce travail de recherche présente une structure en trois parties.

La suite de la **Première Partie** dressera le cadre conceptuel de ces travaux de recherche. Tout d'abord, il est en effet nécessaire de comprendre ce qu'est un projet data, dispositif privilégié d'observation de l'interaction immédiate entre l'écosystème Big Data et l'entreprise. L'objectif est de révéler les processus de référence sur lesquels il s'appuie, et les limites théoriques de ces processus de référence. Ensuite, il faut s'approprier les dimensions clés qui manquent à ces processus, et mettre en perspective les concepts clés qui pourraient constituer des leviers activables favorisant la convergence vers une valeur optimale au cours des projets data. Ces concepts clés seront les indicateurs de valeur dans le cadre de l'utilisation des données en entreprise, la qualité des données, et enfin la Médiation Homme-Donnée, comprenant la complexité des interactions entre acteurs mobilisés dans un projet data.

La **Deuxième Partie** présentera le terrain et les méthodes qui permettront de confronter les éléments théoriques établis en première partie avec le terrain. En effet, cette partie s'attardera tout d'abord sur le choix du terrain, c'est-à-dire l'intégration progressive et délibérée des projets Data Science au sein d'un cabinet spécialiste, Quinten, puis exposera l'approche méthodologique. Il s'agira d'une recherche-action, mode de recherche qualitative et abductive marquée par une posture constructiviste, qui s'appuiera sur une étude de cas multiples, selon un protocole à la fois construit et contraint. Cette méthode aura pour objectif d'aboutir à un recueil d'observations terrain et à la modélisation itérative des résultats.

Enfin, la **Troisième Partie** couvrira d'une part la restitution de dix études de cas, comprenant le recueil détaillé et la synthèse des observations clés, riches en termes de dynamique d'interaction entre acteurs mobilisés, et d'autre part résultat de la modélisation à proprement parler. Le modèle proposé sera constitué d'un ajustement du modèle de processus projet data

de référence, identifié en Première Partie, puis d'un modèle de dispositif de projet data plus large, orienté sur la valeur et encapsulant ce modèle de processus de référence ajusté. Ce nouveau modèle sera alors articulé avec les dimensions clés, émergeant comme manquantes au cours de la confrontation entre la théorie et le travail terrain. Chacune des dimensions manquantes, c'est-à-dire la qualité des données et le dispositif de Médiation Homme-Donnée, s'appuiera sur des propositions de supports inédits et opérationnels, pouvant favoriser l'efficience du dispositif projet data en entreprise. Enfin, cette partie discutera les limites de ces travaux de recherche avant de passer aux conclusions et aux perspectives de recherche.

3 Cadre conceptuel

L'interdisciplinarité du sujet implique une mise en perspective de l'objet de la recherche avec des concepts issus de plusieurs domaines.

Le premier emprunt est la notion de modèle de processus d'un projet data. Au-delà de ces aspects théoriques, sa mise en perspective permet de choisir, de juger et d'envisager un enrichissement du modèle de référence à confronter au terrain. En effet, le projet data constitue la matière première de ces travaux de recherche.

Le second emprunt est la notion d'indicateur de valeur dans le cadre d'une entreprise. Issu d'abord des Sciences de Gestion, cet emprunt permet de situer l'entreprise en tant que système doté d'une finalité, d'une stratégie au sein de son environnement, ainsi que de définir les notions de mesure de la performance et de la prise de décision. Cette conception, malgré son aridité, a l'avantage de procurer un cadre simplifié à l'analyse de l'impact des projets data, c'est-à-dire la comparaison entre les usages et les bénéfices qu'ils génèrent avant et après les projets data étudiés. Par ailleurs, elle permet une mise en perspective de la valeur de l'information avec ses définitions en Sciences de l'Information et de la Communication. Enfin, elle est conforme aux attentes des entreprises à l'heure du buzz.

En distinguant les usages, dont la diversité est un facteur de complexité de ces travaux, du processus d'un projet data, une attention particulière est portée aux différentes dimensions de ce processus, voulues génériques et extrapolables à l'ensemble des projets data.

La première dimension, celle de la valeur, fait partie des hypothèses de travail initiales : elle est couverte par le second emprunt évoqué. Les deux suivantes sont la qualité des données, sous l'angle des différentes disciplines qui portent sa déclinaison opérationnelle, et enfin la médiation entre l'homme et la donnée, pivot dans la prise de décision et de création de connaissance qui se complexifie avec l'arrivée de ces nouvelles technologies et compétences. Les trois dimensions et leur articulation font l'objet, tout le long de ces travaux de recherche, d'une confrontation itérative entre des canons académiques transdisciplinaires et les résultats pratiques de leur application, afin de dégager une zone de convergence à la fois opérationnelle et scientifiquement légitime (voir Figure 3). L'adoption de la vision anthropocentrée, et non pas

techno-centrée, comme dans le cas de la majorité des travaux sur la Data Science, place le rôle et l'interaction entre acteurs au cœur de ces travaux.

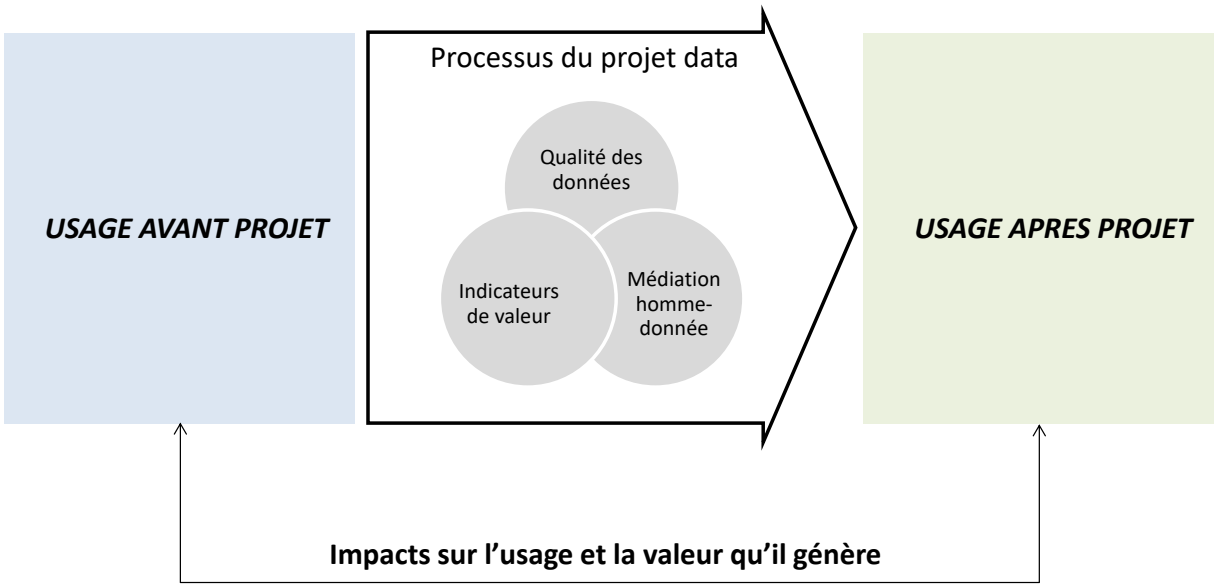


Figure 3 – Cadre conceptuel du projet data

3.1 Processus du projet data

La Data Science est réduite dans ces travaux à son exercice au cours d'un projet, c'est-à-dire « un processus unique qui consiste en un ensemble d'activités coordonnées et maîtrisées, comportant des dates de début et de fin, entrepris dans le but d'atteindre un objectif conforme à des exigences spécifiques, incluant des contraintes de délais, de coûts et de ressources » [6]. Le processus d'exécution de ces activités est réalisé par les acteurs impliqués dans le projet, soit l'équipe projet. Il s'agit d'une vision managériale de la notion de « projet », par opposition à une activité de type « opération » (Declerck et al., 1983; Garel et al., 2003) : c'est une activité non répétitive, dotée d'incertitudes fortes et aux décisions irréversibles. Un projet est une activité réalisée pour la première fois et tend vers un résultat identifiable et inédit, et comporte un ensemble de difficultés traitées notamment par les Sciences de Gestion (complexité de la définition initiale des objectifs, pression externe, gouvernance et management de l'équipe projet, spécificités des critères d'avancement utiles pour le pilotage...). Cette définition du « projet » est un cadre qui nécessite une mise en perspective dans la Data Science.

[6] Organisation Mondiale de Normalisation, la norme ISO 10006 (version 2003) reprise par l'AFNOR sous la norme X50-105

3.1.1 Projet, gestion de projet et processus

Le concept de « projet » commence à s'établir à la Renaissance en architecture (Boutinet, 2012) sous une définition différente : « En dissociant le projet de son exécution, Brunelleschi, en même temps qu'il organise une division technique et sociale du travail, spécifie le projet comme le premier acte caractéristique de toute création architecturale ». L'architecte, précurseur des méthodes de conduite de projet actuelles (Aïm, 2011), conceptualise ainsi le travail préparatoire et l'anticipation du résultat à travers le projet. Un autre architecte florentin de la même époque, Alberti, distingue la volonté qui « fournit le pouvoir moteur qui permet à l'homme de réaliser ce qu'il désire réaliser » de la raison qui lui « permet de connaître exactement ce qu'il désire obtenir comme ce qu'il doit éviter ». Il pose les fondements théoriques de la distinction entre la « finalité du projet » et la « gestion de projet », ce qui met en lumière les notions de coût et de temps de projet, mais aussi du dialogue, dont l'importance est soulignée à la vue du rôle de consolidateur que l'architecte doit jouer au-delà de son statut d'artiste.

La pratique de projet, notamment en entreprise, continuera à irriguer la recherche dans les différentes disciplines telles que les Sciences de Gestion, l'informatique ou les Sciences de l'Information et de la Communication. Inversement, la théorisation du « projet » et de la « gestion de projet », ainsi que la modélisation des processus mis en œuvre au cours d'un projet, donnent lieu à un ensemble de méthodes pratiques.

Le projet est caractérisé par son objet, sa place économique dans l'entreprise et son client (Garel et al., 2003). L'objet permet de distinguer les projets de production unitaire (marqués par le triptyque « maîtrise d'ouvrage, maîtrise d'œuvre et responsable de lots de travaux »), les projets de conception de nouveaux produits (marqués par l'émergence progressive du rôle clé de chef puis de directeur de projet comme coordinateur, et par une absence de spécifications initiales nécessitant un management des connaissances produites et des trajectoires d'innovation) et les opérations exceptionnelles réalisées en mode projet. La place économique du projet dans l'entreprise détermine son impact sur celle-ci (un projet vital pour une entreprise, un projet clé réunissant plusieurs entreprises, un projet parmi d'autres ou un projet-entreprise, spécifique aux start-ups). Enfin, le client du projet détermine ses contraintes et la capacité à renégocier les ressources dédiées : on distingue alors les projets à coûts contrôlés (contrats négociés au forfait ou en régie) et les projets à rentabilité contrôlée (souvent pour un lancement de produit, avec un besoin de préciser les bénéfices potentiels au fur et à mesure). Dans ce cadre, la place des projets Data Science est loin d'être évidente : toutes les combinaisons de caractéristiques

semblent probables étant donné la diversité des usages possibles et le manque de stabilité de l'offre et des modèles économiques des acteurs de l'écosystème Big Data.

Or, les différents outils et dispositifs de pilotage, l'organisation des ressources humaines et les modèles standards ont une efficacité totalement dépendante de la typologie du projet. Le choix de ces dispositifs et modèles est l'un des enjeux phares de la « gestion de projet », qui « pose le double problème de la conception d'une réalisation à venir, puis du passage à l'acte au travers de la réalisation elle-même » (Garel, 2003). La gestion de projet implique la mise en place et le respect d'un « contrat », un engagement dont fait l'objet le projet, y compris en cas de projet interne à l'entreprise. La gestion de projet est un concept qui a connu une évolution terrain millénaire, notamment dans l'artisanat, l'architecture, la marine ou l'entreprenariat, pour faire émerger des démarches plus rationnalisées, puis des modèles standards à partir des années 1960, progressivement institutionnalisés et formalisés. Elle est devenue l'un des sujets clés des Sciences de Gestion depuis les années 80, et présente une grande diversité d'écoles de pensée et de pratique. Elles comprennent les démarches classiques (démarche séquentielle, cycle en V, jalonnement...), mais aussi un ensemble de démarches plus récentes. Notamment, le Lean Management et le modèle d'ingénierie concourante (Garel, 2003), issu de l'industrie automobile, partagent avec la Data Science le besoin de mobiliser des acteurs humains aux compétences hétérogènes pour la conception d'un produit final unique (un véhicule ou un algorithme). La pratique de la Data Science use par ailleurs des méthodes de gestion de projet issues du domaine du développement des applicatifs web ou des méthodes dite « agiles » dans le cadre d'une conception de solution métier.

En effet, si un projet peut être caractérisé par un objet, une place économique et un client spécifiques, il s'agit aussi d'un dispositif, c'est-à-dire d'un agencement, temporaire et pilotable, d'acteurs humains et non humains. Cette notion d'agencement renvoie non seulement aux relations physiques et communicationnelles entre les acteurs, mais aussi aux liens temporels qui relient la dynamique d'activités d'acteurs constituant un processus au sein d'un projet. Or, les processus peuvent être dotés de caractéristiques spécifiques, comme une recherche de conformité (réalisation d'une tâche anticipé ou normée), une volonté de changement (et notamment d'apprentissage), ou encore comme une favorisation de la créativité et de l'innovation (Caliste & Bourret, 2013). Les projets data semblent être des révélateurs des trois dynamiques à la fois, ce qui nécessite de se pencher sur les caractéristiques des processus projet data.

La diversité des cas d'usages et des finalités des projets data résulte ainsi en une diversité de modes de gestion de projet, ce qui empêche en théorie un rattachement des projets Data Science à des courants spécifiques de gestion de projet, bien que la propagation des méthodes agiles dans les dispositifs d'innovation dans les entreprises soit assez concomitante avec le Big Data. Il est donc nécessaire de mettre en lumière en premier lieu les spécificités propres à ces « projets data », comme les modèles de processus de transformation des données en informations utiles, portés par des cadres de référence dédiés, ou encore les ressources mobilisées (données, compétences, technologies…).

3.1.2 Modélisation de processus en Data Science

La modélisation de processus est l'une des composantes principales de la gestion de projet. Elle est définie (Marban et al., 2009; Pressman, 2005) comme un ensemble de taches réalisées pour développer un élément particulier, « l'output », à partir de ressources disponibles, « les inputs ». L'objectif de la modélisation est de rendre le processus reproductible, mesurable et pilotable. La modélisation s'appuie sur une méthodologie, c'est-à-dire une instance de la modélisation de processus qui définit les tâches, les inputs et les outputs, et spécifie comment les tâches doivent être réalisées (techniques et outils pouvant rendre les tâches plus performantes). L'ordre dans lequel l'activité doit être réalisée correspond à la notion de cycle de vie.

Les pratiques propres aux projets data sont largement formalisées au moment du développement du Data Mining et du Knowledge Discovery, et font l'objet d'une convergence entre la pratique et la théorie : un ensemble de modélisations est proposé par les praticiens intervenant sur le terrain au sein des entreprises au cœur du phénomène, ou par les chercheurs, qui théorisent ou post-rationnalisent ces modèles.

L'une des premières modélisations du processus de projet data a lieu au début des années 1990 avec le Knowledge Discovery in Databases Process, ou KDD (Fayyad et al., 1996; Piatetsky-Shapiro, 1994). Ce modèle place le traitement de la donnée au cœur du projet et met en évidence sa transformation en connaissance. Il comprend le Data Mining comme l'une des étapes, et pose ainsi des bases d'un projet plus global. Il se veut applicable à la découverte de connaissances nouvelles, non triviales, valides, et potentiellement utiles. Les connaissances sont dites compréhensibles, soit immédiatement à l'issue du processus, soit à l'issue d'un post-processing. Le processus KDD met en évidence un ensemble d'itérations qui ont lieu entre les étapes (voir Figure 4), et souligne l'importance de l'interactivité entre les concepteurs et les utilisateurs,

dont les décisions impactent la construction des résultats. Cette première modélisation constitue l'une des références pour les travaux qui ciblent le traitement des données massives et à dimensions multiples, le surapprentissage et l'évaluation des modèles algorithmiques, la mise à disposition d'interfaces homme-machine interactive pour le déploiement d'outils de découverte de connaissances assistée par ordinateur, ou encore la prise en compte de la diversification des sources et des formats de données. Elle soumet par ailleurs des perspectives d'optimisation sur les axes tels que l'interaction avec l'utilisateur tout le long du processus, l'évaluation de ses connaissances préalables, mais aussi la prise en compte de la qualité des données, et notamment les données manquantes. Si le dernier sujet est assez largement couvert par la recherche en mathématiques et en informatique, les premières perspectives s'inscrivent au cœur des problématiques des SIC et sont peu abordées.

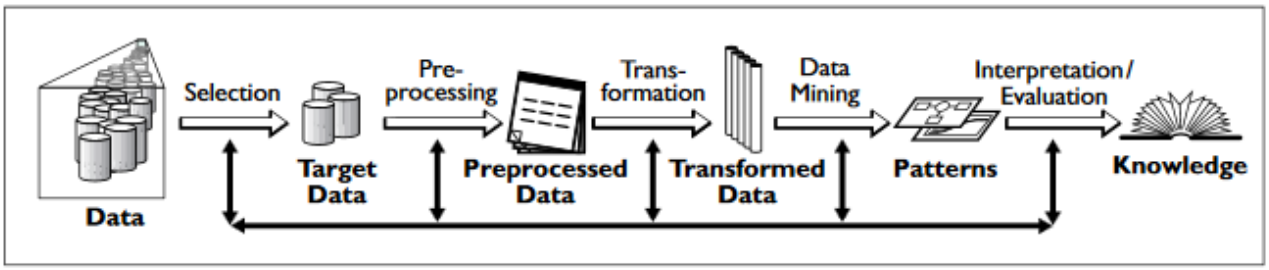


Figure 4 – Synthèse des étapes constituant le processus KDD – *Source : Fayyad, Piatetsky-Shapiro et Smyth, 1996.*

Les praticiens de divers horizons poursuivent le travail de modélisation de processus data, que ce soit pour proposer des modèles inédits, ou pour faire évoluer le modèle KDD : la recherche d'un standard est enclenchée pour garantir l'efficacité et l'efficience de la gestion de projets data. La méthode SEMMA (Sample, Explore, Modify, Model, and Assess) est proposée par SAS Institute Inc : cette méthode en cascade correspond à l'organisation fonctionnelle de l'un des outils de l'éditeur, SAS Entreprise Miner. Cette méthode est critiquable dans la mesure où elle n'a pas de visée neutre, étant dépendante de la technologie sous-jacente. Par ailleurs, elle laisse de côté les aspects métier dans le processus de création de connaissances. KDD Process et SEMMA restent alors toutes les deux pauvres en termes de description de la phase métier amont, mais aussi de la phase de déploiement en aval : en effet, aucune des deux ne précise les modes de transformation des connaissances générées en leviers opérationnels générant de la valeur.

En parallèle, la méthode Cross Industry Standard Process for Data Mining, dite CRISP_DM (Shearer, 2000; Wirth & Hipp, 2000), se stabilise. Initiée par European Strategic Program on Research in Information Technology en 1996 et par un consortium regroupant 5 acteurs de l'écosystème Big Data et d'entreprises, cette méthode est basée sur l'expérience terrain pragmatique, et se veut neutre. L'apport principal de cette méthode consiste à préciser en détail la première phase du projet, c'est-à-dire la compréhension business, qui manquait aux méthodes précédentes évoquées (voir Figure 5). Elle a par ailleurs l'avantage d'avoir été assez documentée, que ce soit en termes de description du modèle de référence (étapes, tâches, livrables...) ou bien de description de la manière de faire (Chapman, 1999), ce qui la rend opérationnelle et enseignable (voir Figure 6). Enfin, cette méthode pose une structure stable grâce à une hiérarchisation entre 4 niveaux d'abstraction : les phases, les tâches, les tâches spécifiques, et les instances. La stabilité est assurée par les deux premiers niveaux, génériques, alors que le troisième assure la possibilité d'adapter le modèle aux spécificités des problématiques traitées. Le dernier niveau correspond à l'enregistrement de l'ensemble des actions, décisions et résultats propres à chaque problématique. Contrairement au KDD Process qui modélise l'interactivité comme possible entre n'importe quelle phase du processus, celle-ci est plus spécifique entre les phases dans le modèle CRISP_DM, tout en gardant le principe général d'itération. Les enchainements restent toutefois libres entre les tâches d'une même phase : la modélisation exhaustive du cycle complexe est jugée peu utile.

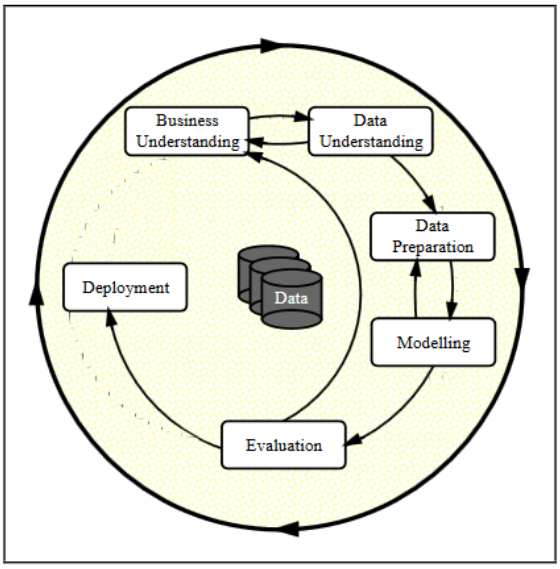


Figure 5 – Les phases du modèle de processus CRISP_DM – *Source : Colin Shearer, The CRISP_DM Model, Continued, 2000*

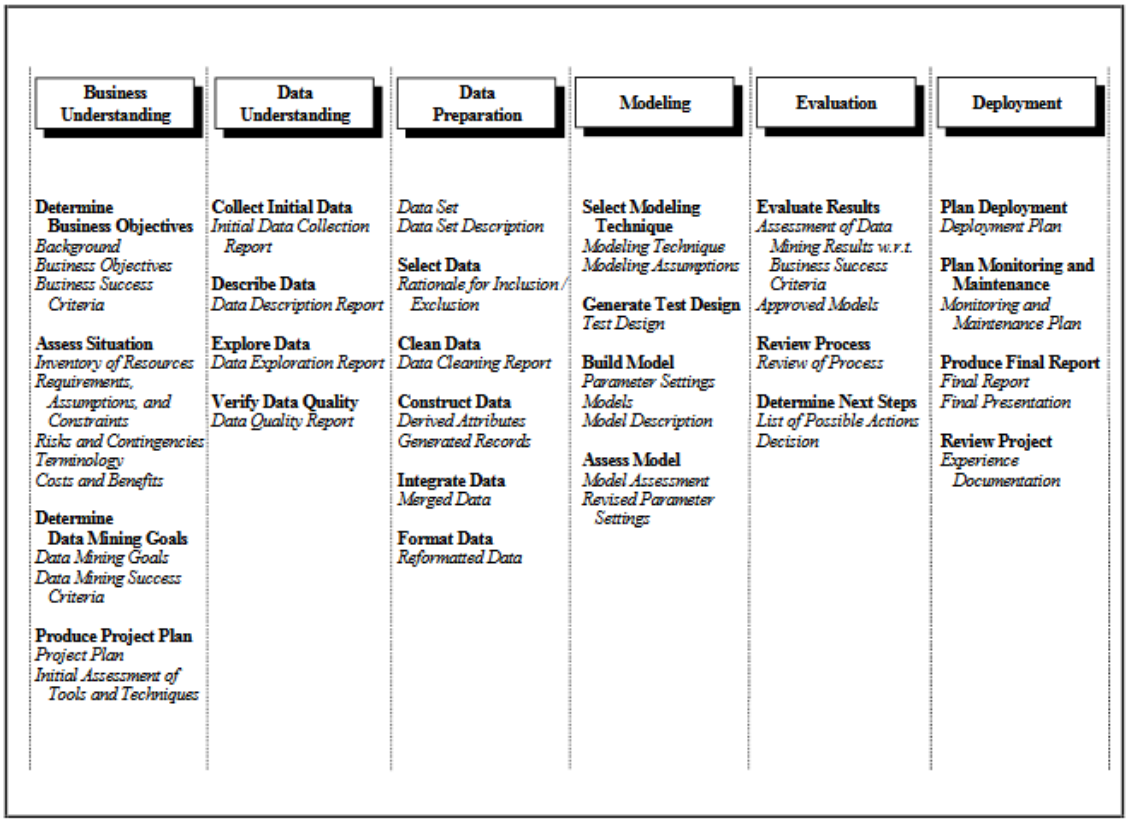


Figure 6 – Synthèse des tâches et des livrables du modèle CRISP_DM – *Source : Colin Shearer, The CRISP_DM Model, Continued, 2000*

Le modèle CRISP_DM est souvent cité par les praticiens comme un processus systématique pour l'extraction des connaissances utiles à partir des données, pour résoudre des problèmes métier (Provost & Fawcett, 2013a). Cependant, il cesse d'évoluer à la disparition du consortium, et sa mise à jour CRISP_DM 2.0 reste inachevée. Les pistes d'optimisation prévues, comme les retours d'expérience des membres du consortium et les études de cas, des modèles de livrables, le mapping de tâches génériques et spécifiques, et autres, sont alors restées en suspens. Par ailleurs, le modèle reste limité par certains éléments comme une description du déploiement réduite aux étapes de sa phase préparatoire, une spécification imprécise des rôles des acteurs impliqués, de l'impact du projet sur la génération de valeur directe et sur la génération de connaissances, indirectement transformables en usages opérationnels. Le modèle est **autocentré** et n'établit que peu de liens opérationnels avec les enjeux de l'entreprise non liés directement au projet en question, bien que leur existence soit évoquée sous la forme de prise en compte du contexte business.

Ces trois méthodes, considérées comme traditionnelles, font l'objet d'un ensemble d'analyses comparatives (Azevedo & Santos, 2008; Kurgan & Musilek, 2006; Marban et al., 2009) qui visent la standardisation du processus, sans pour autant proposer de processus unifié qui prenne racine dans la pratique. En effet, les sondages réalisés par KDNuggets, blog de référence en Data Mining crée par le fondateur de la méthode KDD, illustrent (voir Figure 7) la dominance de la méthode CRISP_DM entre 2002 et 2014, stabilisée depuis 2004 autour de 42% de projets sondés. Par ailleurs, ils montrent un usage significatif de méthodes individuelles, soit 28% des projets en 2014, ce qui confirme les difficultés à converger vers une méthode standardisée.

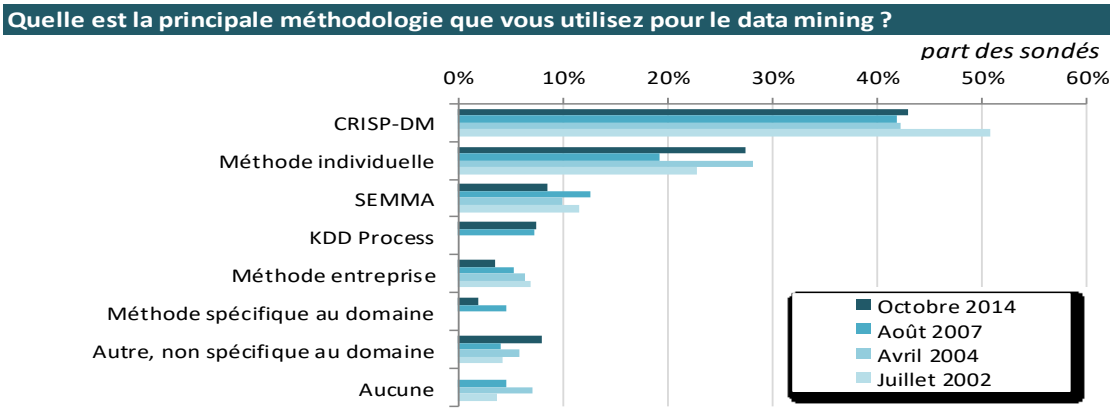


Figure 7 – Utilisation des méthodes projet Data Mining et leur évolution : synthèse de 4 sondages réalisés par KDNuggets entre 2002 et 2014

En Irlande, l'une des modélisations des plus complètes sur l'aspect RH, conçue grâce à l'analyse comparative approfondie de 8 modèles d'autorité, dont CRISP_DM comme seul modèle non issu du monde académique, donne lieu au modèle générique DMCL (Data Mining Life Cycle) (Hofmann & Tierney, 2009). Ce modèle de cycle de vie constitue une évolution du CRISP_DM et son enrichissement sur l'axe des ressources humaines grâce à l'identification et à la description détaillée des rôles de 8 groupes de compétences, impliqués sur un projet data (voir Figure 8). Le modèle est constitué de 3 phases, chacune composée de 3 étapes. Comparé au modèle CRISP_DM, il regroupe les phases de compréhension business et compréhension data dans une seule phase initiale, en la complétant avec la préparation des hypothèses et des objectifs. Il décompose en 3 étapes la phase de préparation des données et regroupe les phases de modélisation, d'évaluation et de déploiement en une seule phase, dite de découverte et de validation. Enfin, le modèle s'appuie sur une conception de Datawarehouse ou de Datamart et d'un entrepôt de données et de connaissances (IKR) résultant du projet. Cette modélisation remet à plat la notion de cycle de vie (ensemble de processus jalonnant le projet) et de

méthodologie (ensemble d'actions et de tâches composant un processus), et aligne les modèles précédents sur un cycle de vie simplifié et cohérent. L'effort significatif de convergence, orienté sur les processus, les données (source et destination), et les personnes, est toutefois limité par trois éléments. Premièrement, il reste imprécis sur l'interaction entre les 8 acteurs impliqués sur le projet, notamment en termes de transmission de connaissances et de convergence sur les objectifs opérationnels avec les experts métier. Deuxièmement, il ne précise pas la nature de l'IKR, le stockage de connaissances et leur utilisation pour créer et mettre à jour une stratégie business. Enfin, ce modèle, comme aucun autre, ne couvre l'impact des processus sur les indicateurs, et donc sur la création de valeur du projet data. Le modèle DLMC semble rester au stade théorique, sans être mis à l'épreuve et confronté au terrain afin de démontrer le gain de création de valeur qu'il clame générer.

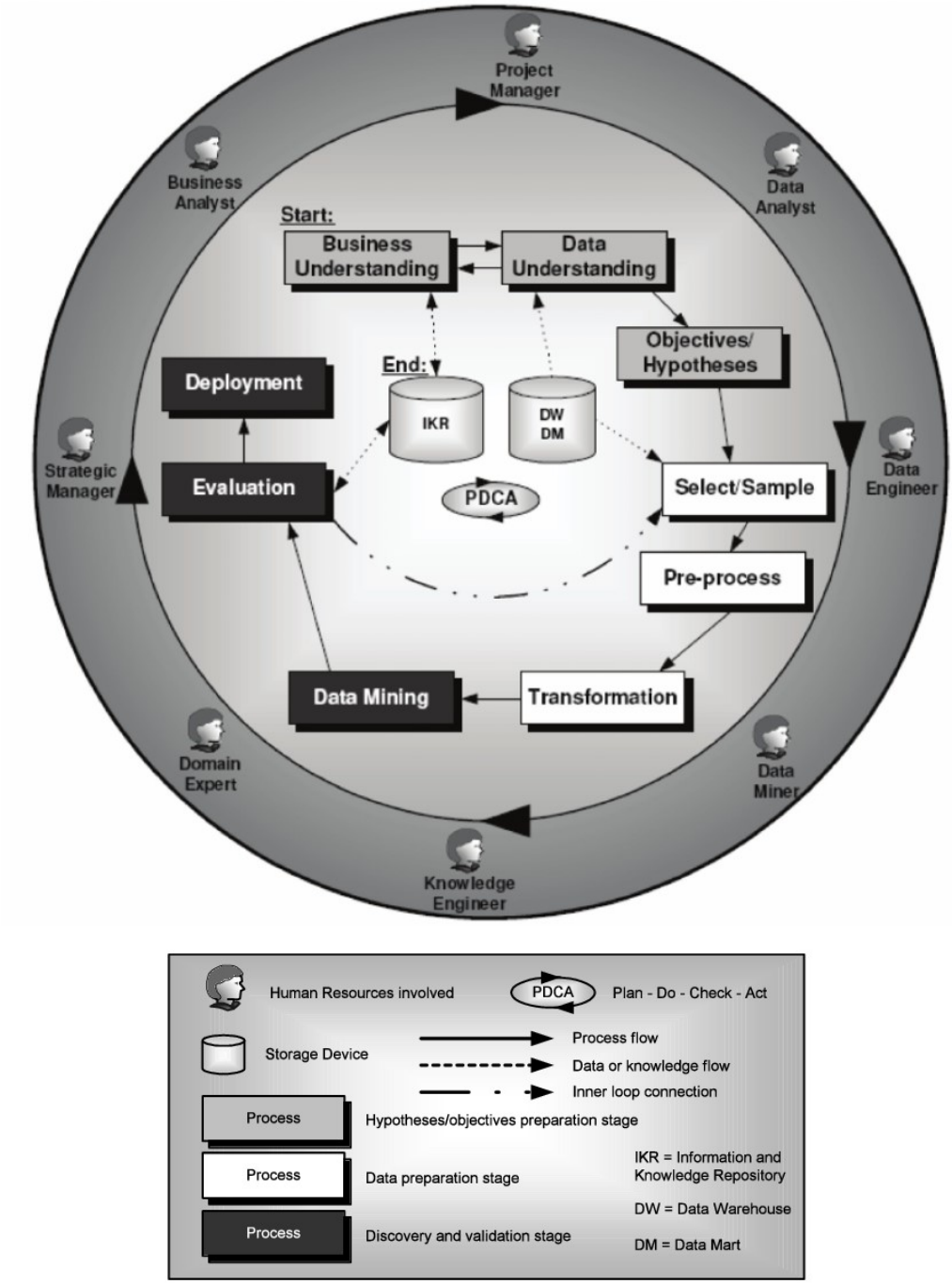


Figure 8 – Modèle de cycle de vie Data Mining – *Source : Hofmann et Tierney, Data Miing Life Cycle (DMLC), 2009*

Les modèles de processus proposés depuis CRISP_DM partagent et approfondissent l'un des apports principaux de celui-ci : l'identification en amont du projet de la phase de compréhension métier. Cette dernière fait l'objet d'un ensemble de travaux de recherche dans le champ des Sciences de Gestion, qui privilégient les enjeux métier tout en établissant des relations avec le processus analytique. L'un de ces modèles est particulièrement représentatif :

il s'agit du modèle SMART (Marr, 2015). Celui-ci pose un ancrage simple du processus dans la génération de valeur en entreprise grâce à l'identification des phases de définition de la stratégie, de mesure des métriques métier et data, suivis de l'analyse proprement dite et de la restitution des résultats avant le déploiement de l'usage à travers la transformation du métier et de la prise de décision. La prise en compte des besoins stratégiques vise avant tout l'établissement des données réellement utiles à l'exploration dans le cadre d'une problématique donnée. Cette phase constitue alors une réduction du périmètre d'analyse aux seules données pré-qualifiées comme utiles. Contrairement aux autres modèles évoquée, le modèle SMART présente une démarche séquentielle plus proche des projets incluant des analyses métier classiques (voir Figure 9) : le résultat des analyses est, en effet, une connaissance qu'il s'agira de transformer en levier, et non pas un algorithme à déployer. Chaque phase du modèle SMART s'appuie sur un socle technologique, plus ou moins mobilisé selon les phases.

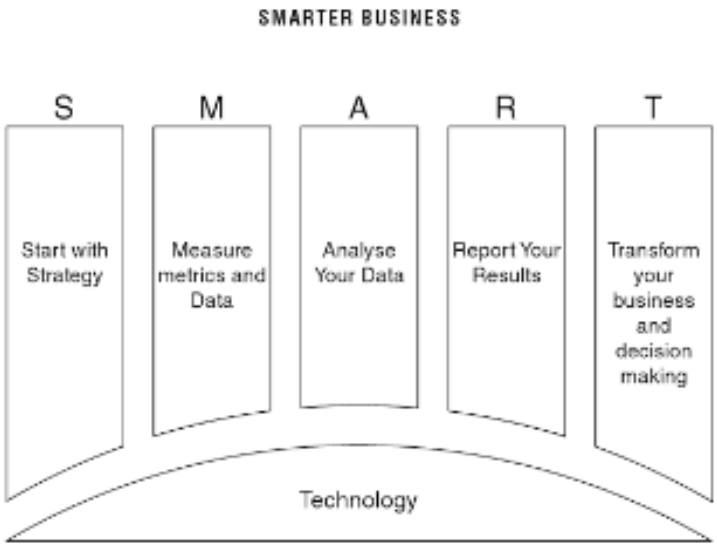


Figure 9 – Modèle SMART – *Source : Marr, 2015*

3.1.3 Méthodes de prise en compte de la technologie dans le processus data

Le support technique mobilisé tout le long du processus data fait l'objet de nombreuses publications, et semble nécessiter une remise en ordre des grandes familles d'outils disponibles pour ce type de projet. D'une part, se développe un ensemble d'outils capables de capter, stocker et mettre à disposition des données brutes : il s'agit traditionnellement de Data Warehouses, Data Marts, et, sous l'impulsion du développement des technologies Big Data, des Data Lakes. D'autre part, la propagation de la Data Science s'accompagne du

développement de plateformes de Data Science qui permettent de traiter les données recueillies, et notamment réaliser les étapes de structuration et de modélisation : ces plateformes incluent des algorithmes et leurs outils d'évaluation. La mise à disposition des résultats issus de cette exploration s'appuie, quant à elle, sur les interfaces de restitution, notamment les outils de Data Visualisation. Enfin, de nouveaux outils dédiés voient le jour dans le cadre de l'industrialisation de solutions métier embarquant des modèles algorithmiques automatisés : ces outils sont interfacés avec les données source en amont et alimentent, complètent ou remplacent les outils de prise en décision en aval. L'évocation de la phase de déploiement dans le processus d'exécution de projet data semble couvrir des éléments hétérogènes selon l'angle choisi par les chercheurs, voire être propre à l'offre de l'éditeur lorsqu'il s'agit de modèles développés par les entreprises comme IBM[7]. Ces offres technologiques commerciales sont complétées par l'Open Source qui nourrit largement les projets data, ce qui complexifie le choix des solutions technologiques et crée un manque de visibilité sur la place de la technologie dans le modèle projet global. Nous distinguerons plus particulièrement les solutions destinées à l'équipe projet pour l'exécution du projet data (apprentissages exploratoires), et les solutions applicatives opérationnelles (solution résultant du projet data destinée à soutenir un usage métier).

Deux écoles de développement de solutions applicatives opérationnelles se profilent : la première s'appuie sur les modèles en cascade, issus de l'ingénierie (Rohanizadeh & Moghadam, 2009), avec une anticipation des solutions en amont du projet et leur implémentation en aval, et la deuxième issue de la méthode agile (Abdel & El Sheikh, 2011). Lancées sous l'influence de la méthode RAD (Rapid Application Development) (Martin, 1991) et regroupées sous l'impulsion du manifeste pour le développement Agile des logiciels[8] de 2001, les méthodes agiles mettent au cœur de la création d'applications le mode itératif, incrémental et adaptatif. Ces méthodes permettent de contrer l'une des limites principales du modèle en cascade, c'est-à-dire son absence de réactivité face aux incertitudes, liée à la rigidité de l'expression initiale des besoins. En revanche, ils présentent un risque fort en termes de qualité des résultats et de manque de documentation, qui ne constitue pas le mode de transmission privilégié : les échanges humains tiennent en effet une place prépondérante, avec partage de connaissances tacites ou explicites. Ces deux approches peuvent toutefois être mixées, ce qui donne des

[7] L'entreprise a en effet utilisé la méthode CRISP-DM, l'ayant incorporé dans son produit SPSS Modeler, avant de proposer une nouvelle méthodologie en 2015 sous la forme de ASUM-DM (Analytics Solutions Unified Method for Data Mining/Predictive Analytics), en développant d'avantage l'aspect infrastructure pour le déploiement de la solution IBM.

[8] http://agilemanifesto.org/

approches Data Science comme ASD-DM (Adaptive Software Development) (Alnoukari et al., 2008) ou Agile Knowledge Discovery in Databases Process Model (Nascimento & Oliveira, 2012). Ce dernier constitue plus précisément un processus de développement de solution applicative, inspiré des modèles de projet data traditionnels et l'OpenUP dont il emprunte les « disciplines » comme la conduite de changement et la gestion de projet. Cet emprunt ne semble pas constituer une innovation en soi, mais apporte les premiers éléments de convergence entre la modélisation de projets data, la méthode de développement d'applicatifs et les modèles de projet classiques. Le modèle accentue l'importance du rôle des utilisateurs sans donner de précisions sur les modalités de son exercice.

Si les méthodes agiles semblent faire le buzz outre-Atlantique, en particulier la méthode Scrum où un Product Owner représente l'intérêt des utilisateurs d'un applicatif à développer par une équipe menée par un « maître de mêlée » facilitateur des interactions, elles restent critiquées en France pour leur incomplétude et leur manque d'adéquation avec l'organisation complexe des entreprises (Khalil, 2011). Prônant l'auto-organisation des équipes, elles s'appliquent essentiellement à des équipes d'informaticiens, réduites en nombre et rapprochées dans un espace commun, et ne remplacent pas le besoin de méthodes classiques de gestion de projet. Par ailleurs, elles ne répondent pas suffisamment au besoin de mobiliser une équipe multi-compétente sur un projet data, qui n'est pas réduit au seul codage de la solution applicative.

Plusieurs tentatives récentes de rapprochement entre les processus projet classiques, les processus de développement itératif de solutions applicatives et les processus projets data, en particulier selon le modèle CRISP_DM, ont permis de générer des modèles assez complets et opérationnels. Par exemple, le modèle dit « Data Ring », crée par deux praticiens italiens et repris par des acteurs comme IFC (Caire et al., 2017; Camiciotti & Racca, 2015), établit une check liste opérationnelle des éléments à prendre en compte de façon itérative au cours d'un projet data, comprenant les objectifs, les outils, les compétences, les processus et la valeur. Le canevas du modèle a l'avantage d'être assez complet (bien que complexe à utiliser), et propose des pistes de réflexion sur la nature du livrable du projet data sous forme de 5 choix :

- Un processus automatisé (transformation de données d'entrée fiables en résultat)

- Un MVP (Minimum Viable Product, c'est-à-dire un concept de produit et de processus dont le résultat témoigne d'une valeur essentielle)

- Un Prototype (concept de produit avec un déploiement, une facilité d'utilisation et une fiabilité basiques)

- Un Produit (concept éprouvé, déployé de façon fiable, et ayant démontré la proposition de valeur)

- Une Production (produit systématiquement déployé et livré aux utilisateurs)

Le Data Ring (voir Figure 10) s'appuie sur une hypothèse de base qui consiste à admettre que les objectifs du projet ne soient pas toujours clairs, et que l'usage ne soit pas complètement anticipé. L'incomplétude potentielle de l'usage est alors contournée grâce à la recommandation de produire un MVP, puis de procéder de façon itérative pour le transformer en prototype. Toutefois, la méthode n'exclut pas non plus la découverte, à condition que celle-ci soit générée de façon structurée à travers un test d'hypothèses. Cette proposition de modèle a par ailleurs l'avantage de regrouper un spectre large d'améliorations apportées au modèle de référence, CRISP_DM, en donnant un outil de pilotage transversal qui rapproche le modèle de la gestion de projet classique. Le modèle omet cependant la génération de nouvelles connaissances et d'usages indirects, et n'évoque pas de capitalisation de savoirs particulière, ce qui maintient l'inconvénient d'un modèle autocentré.

Figure 10 – Modèle Data Ring Canevas – *Source : IFC, 2017, d'après Camiciotti et Racca, 2015*

3.1.4 Synthèse des limites des modèles actuels et pistes de recherche

La revue des modèles de projet permet d'identifier un ensemble de similitudes. Tout d'abord, les modèles pointent l'importance de l'interaction entre les utilisateurs et l'équipe projet data. Cette interaction est définie comme clé dans la phase amont du projet, sous forme de compréhension métier ou de définition de la stratégie et des objectifs du projet. Cependant, cette interaction semble présente à chaque étape du projet, et notamment au cours de l'évaluation des résultats construits de façon itérative. Les modalités de ces interactions ne sont que peu décrites, voire sont contradictoires selon les approches de la gestion de projet, notamment entre les modèles en cascade et le mode agile. Ensuite, aucun modèle ne semble relier la notion de découverte à un capital de connaissances préexistant ou généré au cours du projet, en particulier lorsque la génération de connaissances est exclue des modèles au profit de la construction de

solutions applicatives. Enfin, les spécificités de la notion de valeur semblent floues : bien que certains modèles la mettent en perspective avec les méthodes d'évaluation classiques de gestion de projet, aucun de pointe de particularités d'évaluation de valeur propres aux projets data. Ces similitudes semblent liées à la difficulté à établir un cadre commun de génération de valeur à travers l'usage, dont les définitions divergent entre la génération de connaissances et le déploiement de solutions applicatives.

Face à cet état de l'art du concept de la modélisation de l'exécution de projets data, le modèle CRISP_DM, stable et le plus fréquemment utilisé comme modèle de référence à la fois pour la génération de connaissances (Provost & Fawcett, 2013a) et pour la conception de solutions métier (Caire et al., 2017; Camiciotti & Racca, 2015), est identifiée comme le plus approprié pour servir de référence à l'analyse des études de cas. Le vocabulaire mis à part, ce modèle CRISP_DM regroupe à la fois un socle pour une gestion de projet en cascade, et un cycle de vie itératif plus propre aux méthodes agiles, et ouvre l'éventail des possibilités à un ensemble de méthodes intermédiaires. Il permet alors de s'affranchir du choix des méthodes de gestion de projet pour mieux couvrir la richesse des finalités potentielles de ces projets et de se concentrer sur l'usage.

Au-delà de cette définition vaste d'usage, la prise en compte de la qualité des données semble mise à l'écart. Seul le modèle SMART inclut la préqualification des données dans la phase initiale, les autres modèles isolent cette tâche dans la phase de préparation des données pour l'alimentation des modèles. Or, la qualité des données semble non seulement relative, mais aussi dépendante de l'activité de l'entreprise en amont des projets, et de l'activité réalisée au cours du projet, ce qui nécessite une fois de plus d'envisager une capitalisation sur cet axe. Ces éléments pointent le manque d'ancrage des modèles de projets data, autocentrés, dans l'activité l'entreprise et dans son évolution à travers la transformation du capital de connaissances.

Etant donné ces limites des modèles actuels, et plus particulièrement du modèle de référence (CRISP_DM), une piste de recherche s'ouvre sur la construction d'un modèle global plus évolué, s'inscrivant dans une temporalité observable sur le terrain et faisant l'objet d'un enrichissement sur les 3 dimensions identifiées comme prioritaires à l'issue de la revue de la littérature : il s'agit des indicateurs de valeur, de la qualité des données, et de la médiation entre acteurs impliqués et données.

3.2 Indicateurs de valeur

Les indicateurs stratégiques constituent l'un des angles d'observation privilégiés des études de cas, étant donné qu'ils constituent l'une des finalités principales et directes des projets data au sens global : en effet, ces projets visent généralement une création de valeur quantifiable et l'amélioration de la performance de l'entreprise, or celles-ci sont reflétées par un ensemble d'indicateurs propres à chaque entreprise. Ayant écarté du terrain de recherche l'implémentation de nouveaux business modèles grâce au Big Data, les résultats des projets data ne se positionnent pas en remplacement des produits ou services existants : ils contribuent uniquement à perfectionner les avantages concurrentiels pour répondre à un enjeu stratégique. Cet avantage peut être traduit par des indicateurs de performance classiques, inscrivant les projets data dans une stratégie globale comme vecteur de performance causal, ce qui ne les distingue pas des arbitrages et projets habituels, et ne les place pas en rupture malgré l'innovation apportée. En revanche, de nouveaux paramètres apparaissent localement sous forme d'indicateurs de performance d'activité d'un métier restreint, de données d'entrée nouvelles et opérationnelles, de produits informationnels inédits, ou encore d'éléments de mesure de l'activité métier au service de sa performance. Si la complexité et les spécificités des systèmes d'évaluation de la performance en entreprise ne permettent pas d'établir des indicateurs précis et exhaustifs visés par ces projets, il reste utile et opérationnel d'identifier les catégories d'éléments impactés. L'hypothèse de création d'éléments de valeur est alors mise à l'épreuve à travers les études de cas pour tendre vers un modèle générique de génération de valeur. Cette analyse est réalisée à travers la mobilisation de l'approche systémique des organisations et la catégorisation des natures de valeur générée par les projets data, permettant une mise à plat de l'impact de projets data sur le fonctionnement d'une entreprise et visant la simplification de l'appropriation de la valeur.

3.2.1 Entreprise, stratégie, savoir-faire et usages

Une entreprise est une organisation qui désigne « l'ensemble de moyens structurés, constituant une unité de coordination, ayant des frontières identifiables, fonctionnant en continu, en vue d'atteindre un ensemble d'objectifs partagés par l'ensemble de ses membres » (Robbins & Judge, 2011). Basée sur l'approche dite biologique (Von Bertalanffy, 2012) de la théorie des organisations, la notion d'organisation fait l'objet de nombreux travaux de recherche depuis 1950. Autre notion issue de la recherche biologique, le holisme, défini à l'origine comme « la tendance dans la nature de former des ensembles (des « wholes ») qui sont plus grands que la

somme de leurs parties au travers de l'évolution créative » (Smuts Hon J. C, 1927), est un élément important pour ces travaux : en effet, l'étude d'un sous-système, comme par exemple un centre fonctionnel isolé (marketing, contrôle de gestion…) ne parait pas suffisante pour décrire l'évolution globale d'une entreprise.

La notion d'organisation, abordée selon une analyse systémique, tient par ailleurs de l'approche cybernétique développée au cours de la seconde moitié du XX^{ème} siècle (Wiener, 1948). Un système peut alors être défini comme un « ensemble d'éléments en interaction dynamique organisé en fonction d'un but » (Rosnay, 1991), et l'organisation est alors considérée comme un système ouvert en interaction avec son environnement. Si l'analyse classique de l'organisation, cartésienne, cherchant à démontrer et à obtenir des certitudes, reste compatible et complémentaire avec l'analyse systémique (Guerra, 2007), cette dernière, visant la compréhension et la maîtrise sans chercher de certitudes, a deux avantages dans le cadre de ces travaux de recherche. D'une part, elle permet une simplification face à la complexité croissante des organisations actuelles, dans un contexte de mondialisation et d'intensification des échanges internes et externes en termes de flux transactionnels, énergétiques, physiques et informationnels. Et d'autre part, elle offre la possibilité de dégager des caractéristiques du phénomène étudié à travers le prisme de son impact sur des organisations délimitées dans un environnement, sans pour autant se restreindre à une analyse cartésienne de cet impact.

« La pensée systémique est une discipline […] qui permet d'étudier les interrelations plutôt que les éléments individuels, d'observer les processus de changements. Ce mode de raisonnement devient plus nécessaire que jamais car nous sommes dépassés par la complexité. Pour la première fois dans l'histoire de l'humanité, l'homme est capable de créer des quantités d'information plus grandes que ce qu'il peut absorber, de concevoir des relations d'interdépendances plus complexes que ce qu'il est capable de gérer, et d'accélérer le changement à un rythme que personne n'est capable de suivre » (Senge, 1990, p. 95). Cette définition du cadre conceptuel de l'approche systémique, formulée par Peter M. Senge en 1990, semble faire un clin d'œil à la définition populaire du Big Data à l'heure du buzz, c'est-à-dire « l'ensemble de données qui devient tellement volumineux qu'il en devient difficile à travailler avec des outils classiques de gestion de bases de données ou de gestion de l'information »[9].

[9] Définition tirée de Wikipedia en 2014, toujours présente en 2019.

Un système est défini par une frontière, une finalité, une évolution dans le temps et une organisation (Guerra, 2007). La première caractéristique implique une séparation entre le système et son environnement : le Big Data en tant que phénomène fait partie de l'environnement des systèmes étudiés, c'est-à-dire des entreprises, l'Ecosystème Big Data étant considéré comme un système en soi. Ces systèmes étudiés sont nécessairement ouverts, c'est-à-dire qu'il existe une interaction entre le système et les composantes de son environnement (flux financiers, technologiques, culturels…) (voir Figure 11). La deuxième caractéristique suppose l'intention d'atteindre un objectif fixé : il s'agit de la stratégie, et par conséquent des mesures de contrôle de la trajectoire stratégique définie, les métriques étant des outils de mesure de l'objectif et d'atteinte de cet objectif. La troisième caractéristique permet de cibler les systèmes étudiés grâce à la notion de temporalité : l'analyse de l'évolution implique l'existence du système dans le passé et d'une dynamique. Enfin, la dernière caractéristique renvoie à la notion d'organisation détaillée, soit la structure et les processus du système analysé.

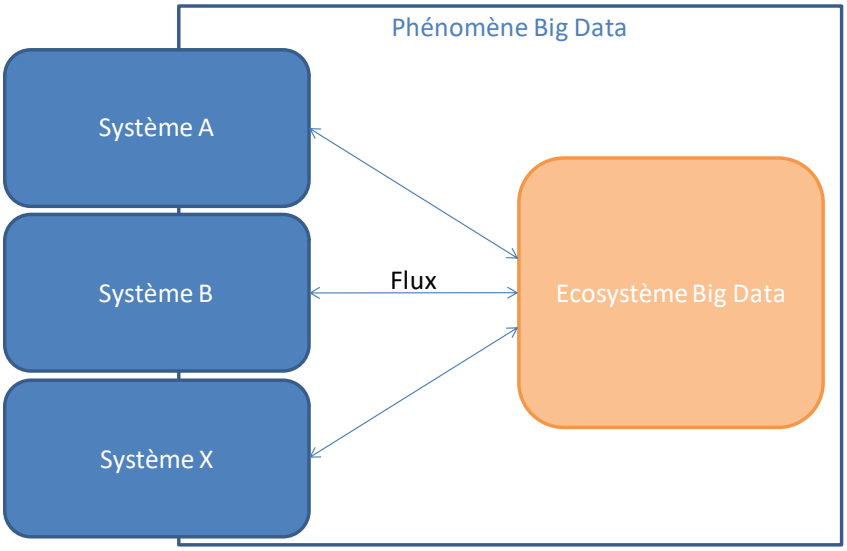


Figure 11 – Vision systémique du phénomène Big Data face aux entreprises

L'entreprise peut se définir par ailleurs comme une somme de processus, c'est-à-dire une chaîne de valeur (Porter, 1998) d'« un ensemble d'activités reliées entre elles par des flux d'information ou de matière significatifs, et qui se combinent pour fournir un produit matériel ou immatériel important et bien défini » (Lorino, 1997). Autrement dit, l'entreprise est un continuum d'activités qui lui sont propres et qui concourent à la création d'un produit ou d'un service. Les activités sont ainsi organisées selon la logique d'objectifs et de résultats, ce que l'organisation hiérarchique (Fayol, 1916) ou fonctionnelle taylorienne ne permet pas d'aborder,

au-delà de la dynamique et de l'interactivité. Les processus sont des activités récurrentes et peuvent être étudiés, comparés et mesurés. Mise en perspective avec l'analyse systémique, qui comprend le concept de rétroaction, la mesure de la performance des processus est clé dans la performance globale de l'entreprise face à une finalité stratégique donnée.

La stratégie est une pensée qui existe depuis plusieurs millénaires. Elle fut tout d'abord une pensée à application militaire avec des grands théoriciens comme Sun Tzu ou Clausewitz. La stratégie s'oppose à la tactique : une décision stratégique est une décision difficilement réversible, à fort enjeu et à effet de système (Atamer & Calori, 2003). Cette pensée s'est structurée après la seconde guerre mondiale avec les apports théoriques de Porter, du Boston Consulting Group et de McKinsey, et constitue un champ de recherche riche dans les Sciences de Gestion. Toute entreprise a un objectif autour duquel elle va construire une stratégie, c'est-à-dire anticiper et orienter l'action sur le long terme, exprimer une volonté de mouvement de changement. Cet objectif peut être décomposé (Atamer & Calori, 2003), à un moment donné de la vie d'une entreprise, comme une pondération non exclusive de 3 orientations :

- La volonté de croissance, c'est-à-dire l'augmentation des volumes et des parts de marché de l'entreprise

- La volonté de rentabilité, c'est-à-dire l'augmentation de la marge et de la profitabilité de l'entreprise

- La volonté de plus-value sociale, c'est-à-dire le développement harmonieux avec l'ensemble des parties prenantes

Une entreprise est insérée dans un environnement dans lequel évoluent les concurrents, les clients, et les autres parties prenantes (salariés, banquiers, actionnaires, collectivité). Un rapport de force s'établit entre ces acteurs, et est représenté par la gouvernance de l'entreprise (Gomez, 1996; Pérez, 2010). Ce rapport de force est momentané, donc susceptible d'évoluer et propre à chaque entreprise. Les parties prenantes sont en lutte permanente pour obtenir un partage de la valeur ajoutée qui leur soit le plus bénéfique. Chacun attend une rémunération en fonction de son apport, et la mesure selon ses propres critères et son utilité. Il y a donc un enjeu dans la répartition de la valeur ajoutée. Le premier principe de l'analyse stratégique consiste alors à définir qui exerce la plus forte pression pour obtenir la valeur ajoutée (de Margerie, 2008). Une fois que l'entreprise a déterminé pour qui elle doit créer de la valeur elle peut commencer à

construire sa stratégie. Se pose alors la question du comment. Dans une économie concurrentielle les clients sont les plus volatils, ils exercent donc une forte pression sur la stratégie. Or le client perçoit deux sources de valeur : la qualité (produit, services, image) et les prix. Le client effectue une comparaison permanente entre les produits de l'entreprise et ceux des concurrents sur ces deux critères. La valeur obtenue par le client est donc relative à ce que proposent les autres entreprises. Partant de ce principe, Michael E. Porter (Porter, 1998) définit en 1985 deux sources d'avantages concurrentiels pour les entreprises : l'avantage concurrentiel par les coûts et l'avantage concurrentiel par la différenciation (qualité). Le stratège doit trouver les sources d'avantages concurrentiels et orienter les ressources vers l'optimisation de la création de valeur. Pour ce faire, les entreprises disposent d'un ou de plusieurs savoir-faire :

- Le savoir-faire technologique, qui procure un avantage en différenciation et/ou en coûts
- Le savoir-faire marketing, qui permet à l'entreprise d'obtenir un avantage en différenciation
- Le savoir-faire en termes de coûts unitaires des facteurs de production, qui permet à l'entreprise d'obtenir un avantage en coûts
- Le savoir-faire de productivité qui procure un avantage en coûts également

Ces savoir-faire sont déclinés à l'échelle opérationnelle sous la forme d'usages. Dans cette approche systémique de l'entreprise dotée d'une finalité stratégique, un usage est porté par un sous-système traitant doté d'un processus causal. Le système traitant est constitué d'une communauté d'acteurs, pouvant représenter une fonction ou une direction, mais aussi une communauté d'utilisateurs ou de contributeurs à une activité. L'usage correspond alors à un modèle logique, c'est-à-dire la production, à partir de ressources, au cours d'activités déterminées et mesurables, de produits ou de connaissances (Curry, Flett, et Hollingsworth, 2006) pour atteindre des objectifs de performance stratégiques à l'échelle du macro-système. Il peut être représenté à travers un modèle Input-Process-Output (Samsonowa et al., 2009), mettant à plat la liste des ressources (Inputs) nécessaires aux différentes activités du processus de production réalisées par le système traitant pour aboutir à des résultats (Outputs). Ces résultats sont constitués de produits, de services, de communications, de connaissances, de processus et autres éléments qui peuvent à leur tour servir d'autres systèmes, dits de réception.

L'enchainement des différentes activités de production en entreprise assoit les avantages compétitifs qui génèrent des bénéfices (Outcomes) sous la forme de contribution à la croissance de l'entreprise, à sa rentabilité, ou bien à sa plus-value sociale. Les bénéfices (Outcomes) se distinguent des résultats des activités (Outputs) par leur caractère transversal : cette distinction se retrouve entre les indicateurs de pilotage stratégique et leur déclinaison en indicateurs de pilotage opérationnels. La représentation de ces différents éléments (voir Figure 12) est utilisée dans ces travaux pour figurer non pas les usages en soi, mais le différentiel entre usages avant et après le projet, en précisant la nature du différentiel et en établissant des éléments de mesure des activités (efficacité des processus), des outputs (résultats opérationnels) et des outcomes (bénéfices stratégiques) espérés.

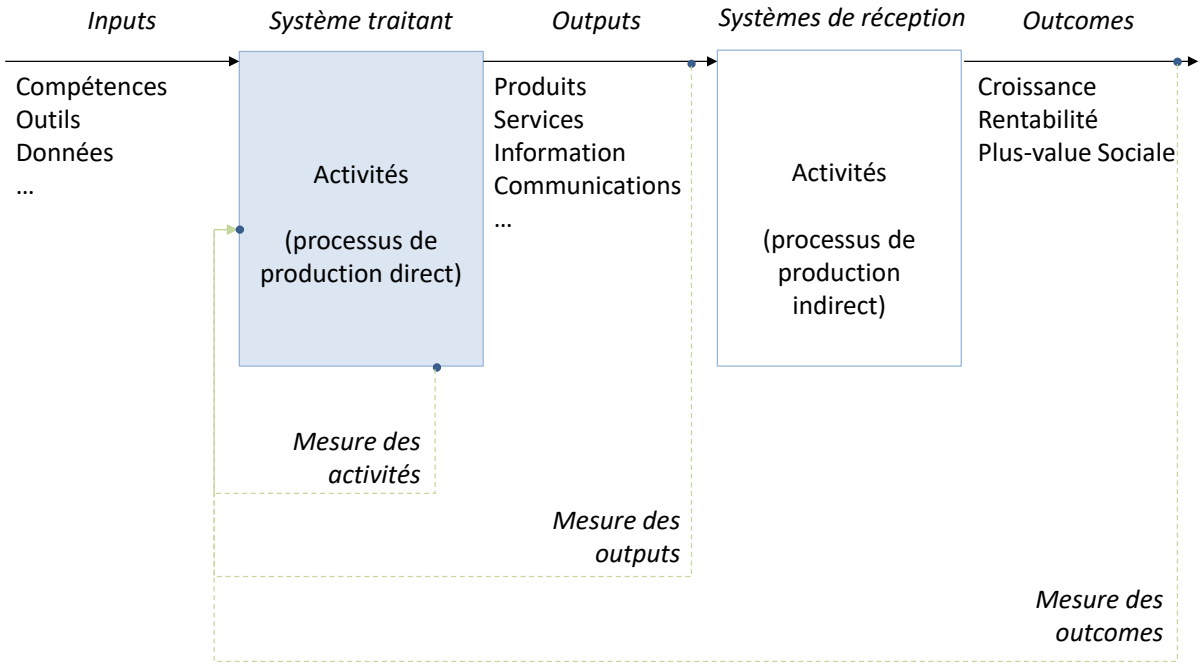


Figure 12 – Modèle Input – Process – Output. *Inspiré de Curry, Flett, et Hollingsworth, 2006, en rendant un modèle R&D plus générique et ajusté selon Atamer et Calori, 2003*

Cette mise en perspective de la notion d'usage ne fait pas l'objet d'un consensus dans les différentes disciplines : en effet, les Sciences de l'Information et de la Communication distinguent habituellement la notion de « pratique » et d'« usage » (Jeanneret, 2007; Jihjah, 2015). L'usage place l'homme dans un rôle d'utilisateur d'un dispositif technologique à sa disposition pour réaliser son action de production : il s'impose alors indépendamment de son contexte. La pratique, quant à elle, est orientée sur l'homme qui met en œuvre un savoir-faire : elle est contextualisée, c'est-à-dire marquée par une insertion sociale et temporelle. Dès

l'époque de la démocratisation de l'informatique, l'étude de la pratique « permet de saisir, de façon concrète, les modes d'interaction entre l'informatique et la société », et de sortir ainsi du schéma du stimulus-réponse (Jouët, 1990) à sens unique pour révéler des dynamiques d'apprentissage, voire de contournement de l'usage initialement prévu. Cette différence implique une prise en compte de déviation possible d'usage dans la pratique terrain, qui devient une « actualisation de l'usage ». Dans ce sens, le savoir-faire, en tant qu'activité de production portée par des acteurs humains dans le cadre d'un dispositif donné, fait référence en Sciences de l'Information et de la Communication à la notion de pratique métier, puisque le savoir-faire est issu d'un apprentissage, d'une appropriation contextualisée du dispositif technologique. La subtilité de cette distinction entre les termes « usage » et « pratique » ne semble pas exister en Sciences de Gestion, qui s'intéressent en priorité, depuis les mouvements de standardisation fordistes, à la fonction productive du savoir-faire et aborde le sujet sous un terme plus générique d'usage métier, ou de « cas d'utilisation », c'est-à-dire d'une manière d'utiliser un outil qui détermine en génie logiciel ses exigences fonctionnelles.

Ces savoir-faire et leurs dispositifs d'application, le rapport qualité/coût, mais aussi la maîtrise des délais, les barrières à l'entrée et bien évidemment la capacité financière sont autant d'avantages concurrentiels à conjuguer pour un stratège, qui doit répondre aux attentes des parties prenantes en trouvant et développant les sources de la valeur dans un environnement de compétition. Il existe deux niveaux de formulation de la stratégie d'une entreprise : le niveau de domaines d'activités homogènes (ou business strategy) et le niveau d'ensemble (ou corporate strategy) (Porter, 1989). Comme il y a autant de stratégies d'activité que de domaines d'activité, la stratégie d'ensemble assure l'intégration nécessaire pour assurer la cohérence. Elle oriente ainsi les ressources vers les domaines d'activité qui créent ou sont susceptibles de créer le plus de valeur, et planifie l'atteinte d'objectifs qu'il sera nécessaire de piloter par domaine. S'il existe un grand nombre de stratégies de pilotage, la plupart partagent des concepts de base : la notion de contrôle (qui fait référence à la rétroaction cybernétique), la mesure de la performance, et l'informatique décisionnelle en appui de ces deux premiers.

3.2.2 Mesure de la performance et prise de décision

L'évolution des objectifs et de l'environnement des entreprises repousse le modèle taylorien-fordiste visant la production de masse, et montre les limites de modèles de contrôle de la productivité et de la rentabilité financière. La performance peut s'exprimer à ce jour en termes plus variés (délais, qualité, conformité…), et provient de la nécessité d'apprendre des erreurs

passées. L'analyse de l'évolution de la notion de performance en entreprise conduit à la définition générale (ou plutôt un ensemble de caractéristiques) suivante (Lebas, 1995) : « la performance n'est pas une simple constatation, elle se construit. Elle est le résultat d'un processus de causalité. Elle est une indication d'un potentiel de résultats futurs. Elle se définit par un vecteur de paramètres reflétant le modèle de causalité dans l'espace et dans le temps. Elle n'a de sens que par rapport à une prise de décision. Elle est relative à un contexte choisi en fonction de la stratégie. Elle est spécifique à un utilisateur et à une stratégie. Elle correspond à un domaine d'action et à un horizon de temps. Elle résulte de la définition d'un champ de responsabilité et le définit en retour ».

La mesure de la performance consiste alors dans un premier temps à définir les objets de mesure, c'est-à-dire des points stratégiques, à leur associer des indicateurs de performance (KPI en anglais, Key Performance Indicator), puis à restituer l'évolution de ces indicateurs pour permettre la prise de décision. Les KPIs doivent être fidèles aux objectifs, rapidement chiffrables en cours d'exercice, additifs pour remonter dans l'organisation. Historiquement, ils peuvent être de 4 types : des quantités à produire ou à vendre, des recettes à réaliser, des dépenses correspondant aux moyens à consommer et un niveau de qualité à respecter. La mesure de la performance classique a été bouleversée par la diffusion des tableaux de bord prospectifs (Kaplan & Norton, 1996) (« balanced scorecard »), qui permet de synthétiser les KPIs à un instant donné, de mettre en relation les indicateurs financiers et non financiers, internes et externes. Les indicateurs financiers objectifs et facilement quantifiables et le résultat des efforts passés sont en effet doublés d'indicateurs (« enablers », ou « drivers ») des performances futures, subjectifs, ce qui relie la stratégie aux facteurs opérationnels. 4 axes d'indicateurs sont généralement représentés de façon « équilibrée » : finance, client, processus et apprentissage opérationnel, ce qui en fait l'un des outils théorisés des plus complets pour l'entreprise.

La conception des indicateurs est fondamentalement top-down, traduisant les orientations stratégiques en commençant par les résultats financiers. Or, des objectifs différents supposent que l'entreprise ne mesurera pas les mêmes processus selon la stratégie qu'elle s'est fixée. Ainsi, le tableau de bord prospectif permet de mettre le doigt sur des processus, donc des indicateurs, qui auraient été invisibles selon la démarche classique. Cette personnalisation des indicateurs ne passe pas à côté de la recherche de causalité qui relie les indicateurs, bien que la causalité entre certains facteurs opérationnels et les indicateurs stratégiques s'avère parfois

difficile à prouver (Lippman & Rumelt, 1982), construisant un réseau de relations de cause à effets appelé « la carte stratégique ». Chaque indicateur résulte d'un modèle de calcul, défini *a priori* en fonction des objectifs de mesure, et basé sur un ensemble de données qui représentent une perception de la réalité terrain. Par ailleurs, en partant du principe qu'« un objet est déterminé par la marge d'erreur qui le sépare à un moment donné de l'objectif qu'il cherche à atteindre » (Rosenblueth et al., 1943), selon les objectifs et les priorités définis par la stratégie d'une entreprise, chaque indicateur n'aura pas le même impact sur la prise de décision.

Mais l'établissement des indicateurs pertinents n'est pas suffisant : il est nécessaire de fournir aux décideurs des outils de visualisation de ces indicateurs en tant qu'aide à la prise de décision. L'ergonomie de la présentation dite « Management Cockpit » (George, 2002), permettant une appropriation plus adaptée aux compétences cognitives des managers, l'intégration d'indicateurs non financiers au sein d'un tableau de bord prospectif (Kaplan & Norton, 1996), l'orientation Business Intelligence de la méthode GIMSI (Fernandez, 2000) ou la déclinaison d'indicateurs top-down par périmètre de responsabilité selon la méthode OVAR (Fiol et al., 2004) sont des méthodes qui aboutissent à la conception et à la mise en œuvre d'un ensemble d'indicateurs de performance traduisant la stratégie d'une entreprise en leviers de pilotage. Ces méthodes, principalement basées sur une déclinaison descendante de la stratégie, s'opposent aux constructions valorisant le capital humain, plus bottom-up, de type Navigator Skandia (Wegmann, 2008). De façon plus réduite, un acteur de l'entreprise peut avoir accès à ses propres outils de visualisation d'informations utiles à son périmètre de prise de décision défini dans le cadre de la stratégie de l'organisation. Ces outils peuvent alors descendre à la maille la plus fine des indicateurs, y compris aux données brutes si la nature de cette donnée brute est interprétée comme un driver de performance.

En théorie, les outils évoqués (Business Intelligence, tableaux de bord, reportings, applicatifs métier dédiés…) sont opérationnels et comportent des Data Visualisations, destinées à fournir aux responsables les informations lisibles et confortablement utiles en termes de qualité, de temporalité et de pertinence afin d'actionner des leviers à leur disposition. En pratique, les entreprises ne sont pas à ce jour toutes équipées de ces outils, ou sont en attente de progrès informatiques sur ce sujet. En effet, une meilleure identification des facteurs de performance et une restitution efficiente des mesures de la performance permettrait d'apprendre plus rapidement, ce qui constitue un avantage concurrentiel. Or, l'un des problèmes majeurs des entreprises, qui ont tendance à complexifier leur organisation et leur système de production, est

celui de la surabondance (Ackoff, 1967; Edmunds & Morris, 2000) de données. Une transformation des données en informations intelligibles et de confiance, utiles à la prise de décision managériale est alors indispensable, ce qui a lieu lors de la mise en place de ces outils et de leur utilisation récurrente. Leur conception comprend la sélection des informations en ligne avec les objectifs stratégiques, la sélection et le nettoyage des données, l'automatisation de la transformation des données, le contrôle de la cohérence des résultats, et la restitution ergonomique des informations permettant une interprétation aisée. Les progrès réalisés sur ces technologies visent essentiellement à équiper de plus en plus les différents preneurs de décision, et à dégager progressivement de plus en plus de temps aux décideurs pour le travail analytique des indicateurs restitués au lieu de les produire.

La limite de ces méthodes de définition de la mesure de la performance est un manque d'espace laissé pour l'évolution et l'émergence de nouveaux indicateurs dans la phase de déploiement du pilotage stratégique. D'autres approches sont proposées pour faire face au besoin d'intégrer les incertitudes liées à l'environnement de l'entreprise, comme l'adoption des systèmes en Open Source (Chau & Tam, 1997) pour répondre aux besoin d'innovation, ou bien des méthodes de création de programmes de KPIs efficients (Kaskinen, 2007). Ces méthodes prônent notamment plus de flexibilité en termes d'analyse descriptive et de reporting, ainsi que la mise en place de processus d'amélioration continue. L'évolution des tableaux de bord grâce à l'intégration de briques de type « Business Analytics » dans les outils de Business Intelligence classique (Fernandez, 2013) fait partie des progrès attendus dans ce sens. Au-delà des outils de Data Visualisation de nouvelle génération, ces outils permettent de faciliter l'accès à l'information et l'investigation humaine grâce à l'analyse des données passées, mais aussi l'identification de tendances futures, ou la génération de recommandations de décisions optimales, voire automatiques. Communément, trois familles d'analyses (Raiffa et al., 1988) sont identifiées : les analyses descriptives (comment et pourquoi a lieu un phénomène ?), normatives (que va-t-il se passer dans des conditions idéales et identiques), et prescriptives (que faire pour maximiser le phénomène, en conditions réelles). Ces dernières visent une évaluation de la valeur pragmatique, c'est-à-dire la capacité des analyses à aider les décideurs, ou alors à les remplacer dans la prise de décision. Cette catégorisation est reprise par Gartner en 2014 pour illustrer les apports potentiels des différentes approches analytiques pour la prise de décision, en remplaçant notamment le terme « normatif » par « prédictif », plus proche du champs lexique du phénomène Big Data (voir Figure 13).

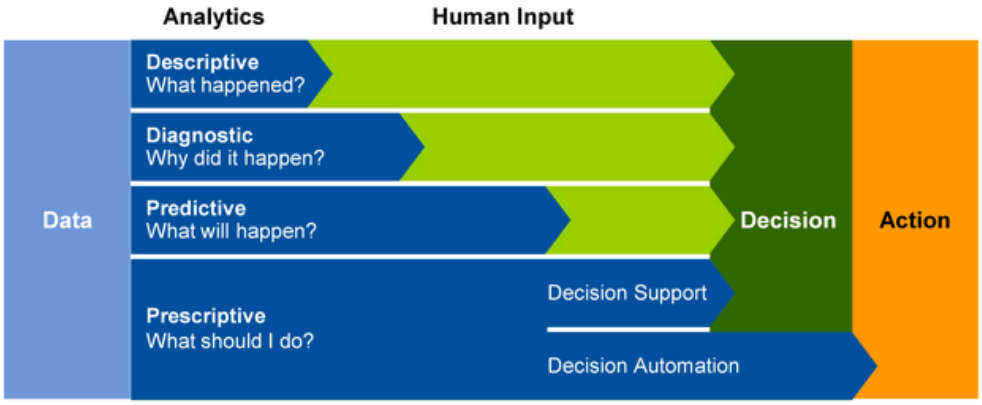


Figure 13 – Chaîne de transformation des données en actions, et 4 approches analytiques possibles – *Source : Four Types of Analytics Capability, Gartner, 2014*

La propagation d'algorithmes prédictifs ou prescriptifs dans la sphère de l'entreprise permet d'enrichir les informations utiles en isolant de façon automatisée l'information du bruit, et s'inscrit plus globalement dans l'évolution des concepts issus de la théorie de la décision (Kast, 1993; Luce & Raiffa, 1989; Von Neumann & Morgenstern, 1945). Elle permet par conséquent d'imaginer des indicateurs inédits, nouveaux par méthode de construction. Ils complètent les réponses des analyses descriptives à des objectifs business existants, et accélèrent le temps humain dédié à l'élaboration de la décision[10], voire remplacent la prise de décision humaine, comme c'est le cas dans certaines applications militaires (A. C. Miller et al., 1976). Ces dernières sont issues de la combinaison de la recherche opérationnelle et de la théorie statistique de la décision, et mettent en lumière 3 caractéristiques : la nature de l'environnement de la décision, les préférences et ressources des décideurs, et le processus d'interaction entre les individus pour atteindre une décision. Le modèle décisionnel est alors défini comme « une abstraction quantitative ou logique de la réalité, construite et analysée dans le but d'aider la prise de décision ». Un tel modèle est caractérisé par les variables du système, représentant l'environnement de la prise de décision, les hypothèses sur le séquençage ou la temporalité de l'arrivée de nouvelles informations sur l'environnement, les hypothèses de liens structurels ou interdépendances entre les variables du système, et les spécificités des préférences du décideur parmi les différents états à l'issue de la décision. L'importance de ces caractéristiques est alors déterminante pour le choix du modèle décisionnel.

[10] *Gartner Says Advanced Analytics Is a Top Business Priority, 2014,* http://www.gartner.com/newsroom/id/2881218

En définissant une métrique comme un indicateur de performance (ou de risque), clé pour la prise de décision, les décideurs pourraient s'appuyer sur des apprentissages algorithmiques supervisés, c'est-à-dire des modèles mettant en lumière cette métrique comme phénomène d'intérêt (ou phénomène à éviter, respectivement). Il s'agit alors d'expliquer ou de prédire le phénomène d'intérêt afin de prendre la décision la plus appropriée. Contrairement à la modélisation classique où le décideur doit établir un modèle *a priori*, en s'appuyant sur les données représentant sa perception de la causalité, les modèles nouveaux fonctionnent en partie sans *a priori*, c'est-à-dire qu'ils dégagent des corrélations entre le phénomène d'intérêt et les données sans que ceux-ci ne soient spécifiés en amont. Ces algorithmes intéressent les décideurs d'autant plus que certains sont auto-apprenants, c'est-à-dire qu'ils s'ajustent sans intervention humaine aux variations de l'environnement : en effet, les modèles traditionnels nécessitent un paramétrage amont, et donc une revue du paramétrage dès la détection d'un changement d'environnement externe ou interne. Les algorithmes de nouvelle génération permettraient de prendre en charge de façon native la détection de certains changements de tendances dans les données qui les alimentent et qui représentent une partie de la réalité du terrain, et d'ajuster automatiquement les paramètres de calcul et la restitution des indicateurs utiles à la prise de décision.

Ainsi, les nouvelles possibilités analytiques offertes par le phénomène Big Data pourraient permettre de mesurer, de prévoir et d'optimiser la performance de l'entreprise, et d'améliorer la prise de décision dans le cadre d'une stratégie subjective propre à l'entreprise. Le gain de performance implique alors deux conditions nécessaires : il faut que la décision basée sur le traitement des données soit effectivement transformée en action (il s'agit de l'exploitation du résultat sous forme d'usage), et que la valeur générée par cette action soit plus grande qu'en absence du traitement des données. Rétroactivement, ce gain de performance semble ainsi pouvoir se propager au résultat du traitement des données (informations utiles à la prise de décision), et donc aux données source elles-mêmes, ce qui pousse à s'interroger sur la valeur de ces objets. Il est ainsi nécessaire de confronter cette vision de la valeur comme déclinaison de la mesure de la performance, descendante et assez aride, à d'autres visions de la valeur de la donnée et de l'information.

3.2.3 Valeur de l'information et de la donnée

La valorisation de la « donnée », complexe du point de vue économique, passe par une décomposition de sa chaîne de transformation et par une prise en compte de l'utilité directe et

indirecte de l'information construite en bout de chaine dans le cadre d'usages, pouvant dépasser la simple prise de décision en contexte incertain.

3.2.3.1 Le paradoxe de la valeur économique de l'information

La valeur économique, au sens classique d'Adam Smith, distingue la valeur d'échange et la valeur d'usage. La valeur d'échange permet de définir un prix objectif résultant des conditions de production d'un bien ou d'un service : il s'agit d'un calcul des coûts des ressources mobilisées, comme la matière première, le travail ou encore le coût du capital, considérés comme plus ou moins prépondérants selon les courants de pensée. Elle sert de socle pour l'estimation d'une valeur de marché, c'est-à-dire la valeur matérielle d'un bien impactée par des facteurs d'offre et de demande entre agents informés sur la nature du bien. La valeur d'usage, quant à elle, consiste à estimer les avantages économiques futurs attendus de l'utilisation d'un actif. Elle est par définition dépendante des préférences d'un agent donné, dans un contexte précis. Ancienne et intuitive, elle a longtemps été considérée comme une valeur psychologique, et non pas économique, issue d'un processus d'évaluation subjectif du prix que l'agent serait prêt à payer, avant d'être remise au gout du jour par le courant marginaliste. L'utilité marginale est alors liée à la consommation d'une unité complémentaire de bien ou de service. Généralement, l'utilité marginale décroit avec le nombre de biens ou de services consommés, sauf dans certains cas comme des biens addictifs ou dans le cadre de l'effet réseau (Bomsel, 2007), cher au déploiement des télécoms, puis d'internet et des réseaux numériques.

Cependant, la valeur économique est difficilement applicable à l'information en tant que bien ou service pour lequel des acheteurs souhaitent payer un prix. Tout d'abord, il existe un déséquilibre entre son coût de production élevé (fixe) et son coût de reproduction quasi nul (marginal), ne diminuant pas sa valeur (Shapiro & Varian, 1998). Ensuite, l'appréciation de la valeur de l'information est postérieure à sa consommation, ce qui empêche un alignement entre le prix, au sens de valeur de marché, et la valeur de l'information. Enfin, l'utilisation de l'information ne la « consomme » pas, elle est dans ce sens inépuisable, tout en étant très volatile, en particulier dans le cadre d'une prise de décision. En effet, les délais entre un évènement, l'acquisition d'information sur l'évènement et la prise de décision face à cet évènement, elle-même dépendante du temps d'activation des leviers possibles, peuvent faire varier drastiquement l'utilité de l'information en question.

A ces difficultés d'évaluation économique de l'information destinée à être vendue, s'ajoute, en entreprise, une absence de cadre homogène d'évaluation de son utilité interne pour la prise de décision et la performance des dispositifs de sa création. Les entreprises ont globalement une utilisation mixte de deux types d'informations : les informations possédées de façon unique (il s'agit alors d'un actif à potentiel d'avantage concurrentiel lié à un produit, un service ou une méthode de production ou de vente) et non unique (dans ce cas, c'est un avantage informationnel au sens polémique, permettant d'adapter la tactique d'interaction à l'environnement externe). La variété des approches de la valorisation de l'information en entreprise est liée notamment à des courants divergentes de l'intelligence économique (Bulinge & Moinet, 2013) comme la guerre, la sécurité, la compétitivité ou encore la diplomatie. Ces approches rendent peu pertinente la conception d'un cadre commun d'évaluation économique de l'information.

La valeur de l'information fait ainsi l'objet d'un paradoxe : l'information est un concept trop large et diversifié pour avoir une valeur économique interne transverse aux entreprises et objective, « ne dispose pas des caractéristiques adéquates pour être une marchandise compte tenu des axiomes de base de la cybernétique », et pourtant « les phénomènes observables tendent à montrer que l'information peut être une marchandise, et des industries se sont développées pour sa production et sa distribution malgré les démentis théoriques » (Mayère, 1990). Ce paradoxe provient notamment d'une vision trop plate de l'« information-donnée » : l'information est alors considérée comme « donnée », au sens mathématique, étymologique et technique, lié à une accessibilité directe. La résolution de ce paradoxe passe par une distinction de l'objet à valoriser entre la « donnée », matière première de l'information, l'« information » en tant que résultat du traitement de la donnée, et enfin l'utilité de cette information dans son contexte spécifique d'application.

3.2.3.2 Chaine de valeur de la donnée

Contrairement à la vision précédente de l'« information-donnée », la distinction entre donnée, information, et utilité de l'information permet de se pencher sur la chaîne de valeur de la donnée, c'est-à-dire le processus de transformation d'une donnée brute en information utile à la prise de décision, que ce soit selon ses aspects informatiques, cognitifs, ou performatifs.

Du point de vue informatique, le phénomène Big Data impacte de plein fouet le coût de production de l'information en accélérant et en enrichissant les possibilités liées au traitement

des données sur toute leur chaîne de valeur (H. G. Miller & Mork, 2013). La multiplication des supports technologiques, y compris mobiles, l'Open Data, et autres évolutions récentes peuvent impacter l'acquisition des données, matière première de l'information. Les progrès en termes de stockage, d'agrégation, de filtrage, peuvent accélérer la transformation initiale des données à partir de sources et formats hétérogènes, et permettre une augmentation du volume et de la richesse des informations détenues. La mise à disposition de nouveaux algorithmes et des compétences pour concevoir des stratégies de modélisation au sein des entreprises peuvent conduire à optimiser et enrichir les usages existants. La connectivité et l'évolution des outils de référencement, de recherche d'information, peuvent mener à une dissémination plus efficiente auprès des utilisateurs. Les innovations en termes d'ergonomie de la visualisation des données peuvent permettre une présentation des résultats plus adaptée à un contexte d'économie de l'attention (H. A. Simon, 1994). Ces changements sont considérés comme pouvant produire des décisions plus pertinentes (automatisées ou non), et à leur mise en œuvre sous forme d'actions appropriées.

Du point de vue cognitif, le processus de transformation de la « donnée » en « information » comporte une dimension d'appropriation, permettant de juger le niveau d'incertitudes contenu dans l'information (Mayère, 1990). Ce point est clé pour la prise de décision et pour la construction de sens, cher à l'approche causale de la prise de décision. Cette appropriation s'appuie notamment sur une démarche de recherche progressive : l'information initiale est alors enrichie par de nouvelles « informations utiles » afin de constituer l'information recherchée. Ces informations utiles peuvent concerner les caractéristiques des sources ou des différentes étapes de retraitement des données, ou bien constituer une description sémantique des données sources, des traitements et de l'information finale : il s'agit des « métadonnées », identifiées en Sciences de l'Information et de la Communication comme l'un des défis majeurs du traitement des données à l'ère du Big Data. Sous cet angle, l'hétérogénéité et le volume des données ont un impact sensible sur la chaîne de valeur de la donnée. Par ailleurs, la possibilité de mobiliser de nouveaux algorithmes pour la réduction des incertitudes contenues dans l'information, par exemple sous la forme de moteurs de recommandation (Kembellec et al., 2014), présente de nouvelles opportunités pour accélérer le processus de construction cognitif.

Enfin, l'aspect performatif, directement inspiré des Sciences de Gestion et des travaux de Porter sur la chaîne de valeur de l'entreprise, découle des deux précédents et se traduit par l'efficacité interne du processus de transformation de la donnée en information utile, c'est-à-dire son coût

de traitement et d'appropriation, ainsi que par l'efficacité externe de l'action qui en résulte. Les dimensions de génération de valeur, spécifiques au Big Data sont présentées par les acteurs de l'Ecosystème essentiellement sous leur forme commerciale, et portent une confusion entre ces deux sources d'aspect performatif. L'émergence de cadres d'analyse de la valeur plus théoriques ne semble pas se profiler, bien que plusieurs tentatives intéressantes aient bien lieu. Purdue University publie un livre blanc (Bertino et al., 2011) regroupant des apports transdisciplinaires de représentants des sciences de l'informatique, des mathématiques et statistiques, de l'ingénierie, de la médecine, de l'éducation et autres. Dans ce livre blanc, l'impact du Big Data est décomposé selon 2 dimensions : les étapes de la chaîne de valeur de données en général, et les spécificités du Big Data impactant chaque étape (hétérogénéité, échelle, temporalité, caractère privé et collaboration humaine). Chaque croisement est discuté et mis en perspective selon l'avancée de la recherche et des technologies, et converge avec les approches précédentes sur les deux aspects, informatique et cognitif (par exemple, sur l'importance de la sémantisation des données pour la pratique métier). L'aspect performatif interne est alors abordé sous la forme de nouveaux défis à relever, et l'externe sous la forme d'actions optimisées visées par la prise de décision (dite « data-driven ») à laquelle aboutit la chaîne de valeur de la donnée.

La comparaison des visions de la chaîne de valeur des données à l'ère du Big Data, selon les différentes disciplines et dans le temps, permet d'entrevoir une évolution possible des modèles en Sciences de l'Information et de la Communication (Delecroix, 2005) vers des modèles plus riches. En effet, plusieurs pistes d'enrichissement apparaissent à travers cette comparaison : tout d'abord, la chaîne de valeur s'étoffe sur la phase de structuration et de modélisation des données, ce qui génère des opportunités cognitives, mais aussi des difficultés au cours de la phase d'interprétation. Ensuite, la multiplication de sources de données hétérogènes est porteuse de défis nouveaux pour la qualification sémantique des données et la gestion des métadonnées en général. Enfin, l'IT retarde dans la chaîne de valeur la représentation des informations pour ne proposer la visualisation qu'à la fin de la chaîne, or le processus cognitif nécessite d'avoir des représentations et une dissémination des informations dès la phase d'intégration. La valeur ajoutée de cette tâche de représentation intermédiaire mérite alors une attention particulière dans la suite de ces travaux. Les trois chaînes de valeur comparées aboutissent dans tous les cas à une prise de décision considérée comme plus pertinente, et donc une action plus efficace, c'est-à-dire plus génératrice de valeur. Les synergies interdisciplinaires sont ainsi mises en évidence pour profiter de nouvelles opportunités et relever les défis

communs à chaque étape du processus de transformation des données, sous l'angle à la fois informatique et cognitif (voir Figure 14).

Comparaison des chaînes de valeur de la donnée entre disciplines et dans le temps				
SIC (Delecroix, **2005**)	Transdisciplinaire (Bertino et al., **2011**)	IT (Miller & Mork, **2012**)	Opportunités et défis **informatiques**	Opportunités et défis **cognitifs**
Acquisition des informations	**Acquisition / Enregistrement**	**Collecte et annotation**	Multiplication des sources de données (web, supports technologiques de collecte, Open Data,…), et des critères de qualification techniques des données	Capture du contexte de collecte, sélection des données utiles, génération des métadonnées sur la source
		Préparation (source)		
Transformation initiale	**Extraction / Nettoyage / Annotation**	**Organisation des données**	Formatage et contrôle de qualité des données hétérogènes	Vérification des la qualité des données et des métadonnées, y compris sémantiques
Dissémination	**Intégration / Agrégation / Représentation**	**Intégration**	Structuration accélérée de formats hétérogènes et de plus grands volumes (progrès de stockage, d'agrégation, de filtrage…)	Richesse des informations et dissémination plus efficiente auprès les utilisateurs (connectivité, évolution des outils de référencement, de recherche d'information)
Modélisation et présentation	**Analyses / Modélisation**	**Analyse**	Capacité de requête sur données hétérogènes non fiables Nouveaux algorithmes, notamment plus appropriés pour traiter des données volumineuses et hétérogènes Puissance de calcul des infrastructures informatiques	Evolution des langages des requêtes sur les données hétérogènes et non fiables (nouvelles compétences) Nouveaux algorithmes pouvant accélérer le processus cognitif (moteurs de recommandation…)
	Interprétation	**Visualisation**	Ergonomie de la visualisation des données	Présentation plus adaptée dans un contexte d'économie d'attention, mais interprétabilité difficile de l'ensemble de la chaine de construction des résultats par les décideurs
Décisions / Actions	**Prise de décision**	**Prise de décision**	Mise en œuvre de décisions plus pertinentes (automatisées ou non)	

Figure 14 – Synthèse des similitudes des chaînes de valeur de la donnée en SIC (Delecroix, 2005), dans des travaux transdisciplinaires (Bertino et al., 2011) et dans l'IT (H. G. Miller & Mork, 2013), illustrant les synergies des opportunités et défis informatiques et cognitifs.

Les propositions de chercheurs issus du monde des données (Provost & Fawcett, 2013a) confirment que la Data Science, comme processus d'extraction de l'information et de la connaissance à partir de la donnée brute, ne génère pas de valeur en soi. Ils avertissent en effet que l'emploi de technologies nouvelles par les entreprises ne sera pas bénéfique (voire aura un impact négatif) tant que l'information issue n'est pas incorporée dans le processus de prise de décision (McAfee & Brynjolfsson, 2012). Cela confirme *a priori* la valeur d'usage comme seule piste possible d'évaluation, mais passe à côté des usages secondaires liés au gain de performance interne de la chaîne de valeur. Provost et Fawcett précisent par ailleurs deux autres points sur la mesure de la valeur générée par les projets data : l'évaluation des résultats de la Data Science nécessite un examen attentif du contexte dans lequel ils seront utilisés, et la relation entre le problème de l'entreprise et la solution d'analyse peut souvent être décomposée en sous-questions via le cadre de l'analyse de la valeur attendue. Cela revient à décomposer la notion de valeur générée non pas à l'échelle des usages, mais potentiellement à celle de questions métier plus fines, transposables en problèmes analytiques. D'une part, cette précision

évacue d'autant plus le mirage d'attribution d'une valeur directement à la donnée ou à l'information pour l'ancrer dans les modèles d'évaluation propres aux contextes des pratiques métier et des incertitudes qui les caractérisent. D'autre part, cette ouverture renvoie aux aspects cognitifs de la transformation de données en informations utiles, évoqués plus haut : « l'information acquiert une signification, devient "informationnelle" dans ce processus qui lie étroitement un traitement et son résultat » (Mayère, 1990).

Il faut, à ce stade, envisager la chaîne de valeur de la donnée comme une suite de transformations informatiques et cognitives, chaque étape étant dotée d'un coût (optimisé par les progrès informatiques) et génératrice de valeur ajoutée à la donnée brute comme matière première. Cette valeur ajoutée ne se révèle alors que lorsque l'information finale générée est activée par la prise de décision dans le cadre d'un usage métier.

3.2.3.3 Valeur des usages issus des progrès sur la chaine des données

Si la valeur d'usage semble généralement la plus mise en avant[11] pour aborder la création de valeur en bout de chaîne de traitement des données, la génération d'une valeur économique plus complexe, au-delà de l'effet restreint sur la prise de la décision, n'est pas absente des débats. Salaün propose, par exemple, de distinguer l'information (contenu) et le « document » (contenu, forme et relation) qui l'englobe (Salaün et al., 2011). En effet, « la valeur [de l'information] est la perception et le jugement qu'un acteur donné a en tête au moment d'un choix à faire (achat, investissement...). Elle le conduit à décider d'acheter ou non, ou encore de préférer telle solution à telle autre. La valeur, construction mentale de l'acteur-décideur, est contextuelle, conjecturelle (spéculation sur les avantages et les inconvénients) et surtout subjective (propre à l'acteur sujet décideur). Cette valeur « décisionnelle » est une mise en relation entre un certain nombre d'avantages (services rendus, impacts espérés...) et des efforts à faire, de l'argent ou du temps à dépenser. ». Le document, quant à lui, s'inscrit dans une logique de production et de marché. Salaün explore l'articulation entre la valeur d'usage de l'information et la valeur économique classique du document à travers l'évolution des modèles économiques de l'industrie de l'information et de la confrontation entre le droit d'auteur européen et le droit à l'information anglo-saxon. Il propose des dimensions de valeur au-delà du simple accès à l'information sous forme de pistes, comme une mise à disposition plus efficiente des documents (synthèses, interfaces plus intuitives, implication plus directe

[11] « Enjeux business des données ». 2014. CIGREF. https://www.cigref.fr/rapport-cigref-enjeux-business-des-donnees.

d'acteurs multiples et leur formation…), une augmentation de la qualité des contenus (richesse, critique des sources, éditorialisation plus intelligente…), une personnalisation de l'usage de l'information (intégration dans des projets personnels, géolocalisation…), ou encore la dynamisation informationnelle. Toutes ces pistes impactent potentiellement les business modèles liés à la production et au marché du document. Les pistes de génération de valeur évoquées à travers l'évolution de la chaîne de valeur de la donnée renvoient ainsi bien à valeur du « document », qui comprend non seulement le contenu, mais aussi sa forme et sa relation, c'est-à-dire ses modalités anthropologiques, intellectuelles et sociales (Salaün, 2007).

Une autre version de proposition de valeur, ambitieuse et annoncée comme révolutionnaire, est offerte par Cukier et Mayer-Schönberger en 2013. Il s'agit d'une catégorisation de « valeurs des données » : si le terme est ici un abus de langage qui télescope la chaîne de traitement de la donnée, il fait essentiellement référence à la façon de mesurer la valeur à travers des usages directs, mais surtout indirects (Mayer-Schönberger & Cukier, 2013) :

- La valeur d'option : création d'un véritable marché de la connaissance autour d'usages secondaires, telle la face immergée de l'iceberg, à forts enjeux de standardisation dont se sont emparées les sociétés les plus innovantes dans l'Ecosystème Big Data (par exemple, collecte et traitement prédictif de données sur les véhicules connectés de Honda par IBM pour prévoir les meilleures périodes de recharge et les lieux pour construire les stations de recharge)

- La réutilisation des données : génération de produits et de services à partir des données initialement collectées à d'autres fins (par exemple, utilisation des données de visionnage de contenu et de vente d'AOL pour faire du ciblage par Amazon, qui hébergeait sa plateforme)

- La recombinaison des données : génération de connaissances inédites à partir de sources non confrontées auparavant (par exemple, la recombinaison des données des opérateurs de téléphone mobile danois, du registre des patients atteints du cancer et du registre national des niveaux d'études et des revenus des citoyens Danois montre une absence de corrélation entre le cancer et l'usage de téléphone mobile[12])

[12] https://www.rtbf.be/info/societe/detail_un-lien-entre-usage-du-gsm-et-cancer-du-cerveau-les-avis-divergent?id=6955583

- L'extension des données : collecte de données complémentaires à coût marginal (par exemple, collecte par Google, dans le cadre de Street View, des photos des lieux, mais aussi des éléments GPS, des contrôles de la cartographie, voire des noms des réseaux wifi)

- La dépréciation de la valeur : modélisation de valeur de données pour distinguer les vitesses et modes de dépréciation, hétérogènes selon leur nature, et optimiser les coûts de stockage

- La valeur des traces numériques : valorisation de rebuts d'information à des fins secondaires (par exemple, utilisation des erreurs de frappe sur Google pour automatiser l'ajustement de son correcteur orthographique ou la saisie automatique, ou encore utilisation des traces de comportement sur les sites pour améliorer le classement du moteur de recherche)

- La valeur des données ouvertes : divulgation des données publiques à des fins commerciales et civiques

- La valeur non mesurable : évaluation de « ce qui n'a pas de prix », comme la valorisation initiale de Facebook à 104 milliards de dollars alors que ses actifs valaient 6,3 milliards, soit une valorisation à environ 100 dollars par compte utilisateur comme actif incorporel, ce qui questionne la valorisation comptable des données sous la forme d'écart d'acquisition

L'avantage de cette catégorisation est son caractère économique plus large : une séparation claire est énoncée entre la valeur immédiate et sa valeur latente (usage indirect, impactant les activités futures en capitalisant sur les activités passées, voire les business modèles de l'entreprise au sens plus large). Cette vision de valeur est assez orientée sur l'Ecosystème naissant : elle incite les entreprises à se renouveler en générant une crainte de se faire dépasser, et non pas à optimiser leur métier ou innover leur modèle librement. Cette vision permet de concevoir le document, vu précédemment, non pas comme un objet plus riche que l'information qu'il contient, mais comme un artefact issu d'un secteur d'activité économique qui peut, lui aussi, bénéficier des apports possibles du phénomène, et craindre l'entrée sur le marché de ces nouveaux entrants.

De façon moins anxiogène et assez coercitive, McKinsey (Manyika et al., 2011) partage cette vision plus ambitieuse et donne cinq leviers de création de valeur par l'usage des données : en créant de la transparence, en permettant l'expérimentation afin de déterminer des besoins et améliorer la performance, en segmentant la population afin de personnaliser les approches, en remplaçant ou aidant les prises de décisions humaines avec les algorithmes automatisés, et en

innovant les business modèles, produits et services. Ces pistes mettent en évidence, une fois de plus, les différents types de résultats directs pouvant être attendus des projets data, tout en ouvrant les portes aux usages indirects, et laissent une place aux objectifs opérationnels individuels des organisations, dotées de stratégies propres. Elles élargissent ainsi définitivement le spectre des usages de la simple prise de décision dans le cadre du pilotage des entreprises.

Bien que les pistes d'usages proposées à travers ses trois visions soient, en soi, différentes, elles pourraient se nourrir mutuellement dès lors qu'on accepte l'inscription de chaque usage dans son propre contexte économique. Par exemple, la géolocalisation de l'information et la personnalisation du produit en marketing ne semblent pas dénoués de points communs, ce qui permet d'envisager des partages de pratiques intéressantes, au-delà du fait que les deux pistes s'appuient sur des progrès de structuration et de modélisation des données. De même, la création de la transparence et d'interfaces intuitives pour le partage du document semble assez transposable. Cela confirme qu'un décloisonnement des pratiques entre secteurs économiques peut avoir lieu du point de vue informatique et cognitif sur la chaîne de valeur de la donnée, à condition de maintenir une contextualisation des modalités de mesure de la valeur générée par les usages. Enfin, les trois visions (Salaün, Manyika, Cukier et Mayer-Schönberger) partagent cette opportunité d'évolution des business modèles, qui constitue un usage indirect assez radical. Ce dernier est exclu de ces travaux de recherche qui se concentrent sur l'optimisation des business modèles existants en entreprise.

Ainsi, un consensus se dégage sur l'usage comme seul levier de génération de valeur. Cependant, l'usage peut être direct (prise de décision contextualisée comme finalité de la chaîne de valeur de la donnée, et objet principal visé par l'approche analytique en Data Science) ou indirect (performance interne du processus de transformation informatique et cognitif de la donnée, capitalisation de connaissance, expérimentation, et potentiel d'activation d'usages latents, voire de nouveaux business modèles qui en découlent). La difficulté initiale de l'appréhension de la valeur économique des données et de l'information, provenant d'une confusion sur l'objet valorisé, semble alors résolue. Les données apparaissent comme une ressource pouvant être transformée en informations utiles, voire en artefacts plus riches (documents, solutions métier…), nécessaires à des usages directs et indirects. Ces usages sont dotés d'un certain niveau d'incertitude que le processus analytique et cognitif de transformation des données permet de réduire. Ils sont les seuls à pouvoir générer une quelconque valeur.

Le projet data est dans ce cadre un processus de réduction d'incertitudes correspondant à la conception de l'usage. Toute ressource, dont la donnée, est alors un investissement dans la construction analytique et cognitive de l'information utile, et l'objectif du projet est de valider cette utilité, c'est-à-dire de confirmer si son exploitation est bien bénéfique. La donnée faisant l'objet d'un investissement projet est à distinguer de la donnée comme ressource nécessaire à l'exploitation même de l'usage : par exemple, dans le cadre de déploiement de moteurs de recommandations, les données alimentent continuellement un algorithme d'ores et déjà déployé (conçu en amont au cours d'un dispositif projet) pour réduire les incertitudes au grès des questionnements des utilisateurs (Kembellec et al., 2014). Dans ce cas, la donnée est une dépense qui diminue les recettes de l'usage. L'aspect performatif de la chaîne de la valeur de la donnée, à la fois informatique et cognitif, permet ainsi de diminuer l'investissement dans le cadre d'un projet data, plus abordable, ou bien, dans un second temps, d'augmenter les bénéfices à l'issue du projet lorsque l'usage est activé. Cette vision, on ne peut plus économique, de la donnée reste à mettre en perspective avec les pratiques habituelles de gestion de cet actif, et plus particulièrement de sa notion de qualité.

3.3 Qualité des données

La qualité des données est étudiée ici comme l'un des éléments de la chaine de transformation des données perturbé par le phénomène selon ses aspects techniques, économiques et cognitifs, ce qui renouvelle les enjeux de gouvernance des données, notamment à travers l'arrivée d'un nouvel objet : les algorithmes.

3.3.1 Des approches opérationnelles différents selon les disciplines

La « qualité » est un concept qui a évolué à travers le temps et les cultures, en suivant notamment les mouvements industriels, économiques et sociaux. La qualité des données en particulier s'inscrit au cœur de création de valeur par l'entreprise à travers la gestion des connaissances et de l'amélioration de la prise de décision, et, dans certains cas, comme un service en soi proposé par l'entreprise à ses clients. Dans ce sens, les Sciences de Gestion posent un socle commun pour la convergence des disciplines sur la notion d'utilisateur. En effet, pour les sciences informatiques « une donnée de qualité est une donnée qui convient aux usages auxquels les utilisateurs sont en droit de s'attendre » (Richard Y. Wang, 1998). De même, pour les Sciences de l'Information et de la Communication, la qualité des données garantit un niveau suffisant des attributs de l'information, notamment sémantiques, pour satisfaire les besoins,

plus ou moins clairement exprimés (Cottin & Nesme, 2017), des clients-décideurs sur leur périmètre de responsabilité d'action. Cependant les conséquences théoriques et opérationnelles divergent rapidement entre ces trois disciplines qui projettent leurs propres définitions sur le concept de la qualité des données, et un consensus ne semble pas se dégager.

En effet, les Sciences de Gestion se restreignent à une vision fonctionnelle d'utilité dans le cadre de la prise de décision, comme pour la valeur des données et de l'information, les sciences de l'informatique proposent des définitions plus technico-mathématiques, et les Sciences de l'Information et de la Communication abordent le sujet en termes de génération de sens. Or, le phénomène Big Data semble perturber de front les trois approches disciplinaires, plus particulièrement en favorisant l'intégration des algorithmes dans la palette des données à gérer en entreprise. Une mise en perspective est alors nécessaire pour aborder ce sujet sous l'angle des projets data, à la fois consommateurs et générateurs de données.

3.3.1.1 Le besoin de qualité pour les Sciences de Gestion

En Sciences de Gestion, la conceptualisation de la notion de qualité s'inscrit dans le cadre de la diversification et de complexification des produits et des services, en particulier pour les organisations qui nécessitent un management distancié, une maîtrise des coûts pour un niveau de qualité acceptable, et une mitigation des risques à chaque étape de production. La qualité est généralement définie soit comme l'aptitude d'un ensemble de caractéristiques d'un produit ou d'un service à satisfaire l'usage attendu et les besoins, exprimés ou implicites, soit comme l'inverse de non-qualité, ou l'absence de défauts nécessitant un retraitement (Juran & Godfrey, 1999). Le niveau de qualité se traduit en termes de coûts (effort à fournir, contrainte à subir, investissement pour la mise en qualité...) selon un bénéfice généré (satisfaction, performance, différentiation, diminution de pertes...) au cours de l'usage en question. Il s'agit du rapport qualité-prix, déterminant pour la comparaison de produits ou services sur un marché concurrentiel. L'application opérationnelle de ce rapport qualité-prix nécessite la distinction (Doucet, 2010) entre la qualité théorique (spécifications générales d'un produit ou service), la qualité réelle (conformité d'un produit ou service donné aux spécifications générales), la qualité souhaitée (attentes subjectives dépendantes de chaque client, ou d'un profil de clients), et la qualité perçue (par le client, ou alors par le fournisseur).

Au concept de qualité s'ajoute alors le concept de gestion de la qualité, ou « qualitique », regroupant un ensemble de « méthodes, normes et guides qui ont pour but d'aider à réaliser

cette qualité » (Doucet, 2010). La gestion de la qualité permet d'enrichir la définition de la qualité sous l'angle de l'efficacité optimale de l'action de production, c'est-à-dire la qualité « interne », par opposition à la qualité « externe » correspondant à la satisfaction des objectifs de l'usage. Elle constitue un moyen pour réduire le risque de non-qualité à travers un triple processus de planification, de contrôle et d'amélioration, marquée par un certain niveau de formalisme des procédures ou d'indicateurs de qualité. S'il n'existe pas de méthode commune de gestion de qualité, plusieurs écoles sont identifiables, comme le contrôle qualité (marqué par la dominance de la métrologie et des statistiques), les méthodologies organisationnelles (gestion de projet, gestion de risque), les normes ISO 9000 (systèmes qualité supervisant l'activité de l'organisation de façon procédurale), l'éthique Lean 6 Sigma (mesure et analyse des défauts), les approches RH (implication et motivation des individus et des groupes de travail), ainsi qu'un grand nombre de pratiques de niche. Ces différentes méthodes sont cumulables et plus ou moins adaptées à un contexte ou à une culture, plus ou moins génériques ou propres à un secteur d'activité ou à une fonction.

La gestion de la qualité des données est un alors processus de gestion de la qualité comme un autre, propre à chaque entreprise, voire à chaque usage. Il vise une priorisation des critères (choix du modèle de qualité des données) et une amélioration des valeurs des critères priorisés. Cette amélioration peut avoir lieu sous forme d'impact technologique (automatisation de la mise en qualité des données, des métadonnées…), d'impact processus (simplification, standardisation…) ou bien d'impact humain (alignement sur les processus). Encore faut-il pouvoir prioriser les champs d'actionnement de ces leviers, et s'appuyer sur les disciplines qui les portent.

3.3.1.2 La réponse de l'Informatique : des indicateurs de qualité génériques

Lorsque la question de la qualité des données est abordée dans les sciences dures de la donnée, elle se réfère essentiellement à la quantification et la qualification du contenu en information d'un ensemble de données, en mobilisant la théorie de la calculabilité ou encore la notion de machine universelle de Turing. Cette approche, selon la théorie algorithmique de l'information par Kolmogorov, Solomonoff et Chaitin vient de pair avec la théorie de l'information de Shannon, probabiliste, permettant de quantifier le contenu moyen en information d'un ensemble de messages, dont le codage informatique satisfait une distribution statistique précise (Shannon, 1948). On cherche alors à optimiser un ratio, fournir le moins de données possibles

pour maximiser les informations qu'elles contiennent. Au-delà de la beauté théorique de l'approche, le but est d'économiser les ressources nécessaires au traitement de l'information.

Dans les sciences informatiques moins théoriques, la recherche sur la qualité des données a donné lieu à plusieurs propositions de modélisation (Berti, 1999; Jarke et al., 1997; Richard Y. Wang, 1998). Si la nature et le niveau de qualité des données ne font pas l'objet d'un consensus, il existe trois approches pour identifier des critères de qualité (Harrathi & Calabretto, 2006; Naumann & Rolker, 2000) : l'approche sémantique (critères de qualité explicites), l'approche par processus (critères s'appliquant à chaque étape de la chaîne de traitement de la donnée) et l'approche par objectifs (critères priorisés et mesurés selon des cibles définies au préalable). Ces modèles visent essentiellement la conception d'indicateurs de mesure des critères de qualité des données. En France, l'association ExQI[13] (Berti-Equille, 2012) a animé un groupe de travail sur le sujet, et présenté dans un ouvrage collectif les résultats de ces travaux, afin de dégager des bases théoriques communes et les meilleures pratiques du marché. S'il reste complexe d'établir des définitions partagées face à la richesse du sujet, les indicateurs génériques émergeant sur la qualité des données semblent faire plus ou moins consensus (voir Figure 15).

[13] Excellence Qualité Information, http://exqi.asso.fr/

Familles d'indicateurs	Description
Pertinence	capacité des données à répondre aux besoins actuels et futurs des utilisateurs », incluant la prise en compte de la diversité des processus métier traitées à partir de la même donnée, la restriction des données utilisées actuellement aux seules données utiles, et l'adaptabilité des données pour les besoins futurs.
Exactitude et justesse	conformité des données à la réalité
Complétude	exhaustivité de toutes les entités dans le modèle de données, de tous les attributs nécessaires au métier dans les entités du modèle des données, des relations entre les entités du modèle de données, et des occurrences d'une entité
Consistance	pour chaque occurrence d'une entité recopiée et maintenue en plusieurs exemplaires, toutes les valeurs des attributs sont identiques entre les bases
Précision temporelle	exactitude des données par rapport à l'instant qu'elles sont censées représenter
Accessibilité	facilité de localisation et d'accès aux données par l'utilisateur
Facilité d'interprétation	facilité de compréhension des données, de leur analyse et de leur usage », que ce soit à travers la documentation, l'existence de nomenclatures partagées et de métadonnées, ou le choix des termes, proches du vocabulaire métier
Unicité	une entité du monde réel correspond à un seul et unique enregistrement (absence de doublons)
Cohérence	absence d'informations conflictuelles au sein d'un même objet ou entre objets différents
Conformité	respect par un ensemble de données de certaines contraintes, relatives à un standard, un format ou une convention de nommage

Figure 15 – Familles d'indicateurs de qualité des données - *Source : Berti-Equille 2012.*

La description des indicateurs génériques de qualité des données renvoie aux objets plus précis qu'ils permettent d'évaluer : les enregistrements unitaires, les entités, les attributs et plus généralement les modèles de données qui les structurent. Ces objets, dits « données », correspondent à deux composantes principales : les « valeurs », ou enregistrements, et les « modèles de données » (Fox et al., 1994), c'est-à-dire les entités, les attributs et les relations entre les eux que l'entreprise utilise pour structurer sont environnement. Par ailleurs, un autre objet de qualité est mentionné dans cette famille d'indicateurs : il s'agit de métadonnées. Les « métadonnées » sont des données servant à décrire une autre donnée pour l'utiliser ou faciliter son interprétation. Un indicateur de qualité peut lui aussi, dans ce cadre, être considéré comme une métadonnée décrivant un enregistrement, une entité, une relation ou un modèle, c'est-à-dire les « données ». Si la qualité des données est prise dans son sens plus large orienté usage, il s'avère indispensable de prendre en compte d'autres objets à qualifier, tels que les interfaces utilisateur, leur capacité à appréhender le sens de l'information, ou encore la déclinaison des choix stratégiques qui permettent de prioriser les critères de qualité. Les solutions technologiques qui soutiennent les processus de traitement sont elles aussi dotées de critères de qualité comme la sécurité, la fiabilité, l'accessibilité, la disponibilité, la maintenabilité,

l'interopérabilité ou encore la confidentialité[14]. Cet élargissement, non exhaustif, permet la prise en compte de trois facteurs de qualité, généralement invoqués dans les travaux sur les modèles de qualité : l'utilisateur, la source, et le système porteur du processus d'accès à l'information.

Cependant, la diversité et l'aridité de ces modèles limitent leur mise en perspective avec les notions plus anthropologiques de génération de connaissances ou de perception de la qualité par l'utilisateur. Dans ce cadre, Rami Harrathi et Sylvie Calabretto (Harrathi & Calabretto, 2006) exposent un modèle de qualité qui varie non pas en fonction de la définition de l'ensemble des facteurs de qualité, mais de la structure, de la représentation de la notion de qualité en général. Ils expriment l'hypothèse que la valeur de l'information est étroitement liée à sa qualité : elle est donc relative à son utilisateur et à son objectif, c'est-à-dire à la satisfaction de ses besoins en termes de choix et d'appréciation des facteurs de qualité. Ainsi, une seule et même information sera jugée de bonne qualité dans un contexte d'usage, et de mauvaise qualité dans un autre contexte d'usage, car les facteurs de qualité seront appréciés de façon différente. Le processus de choix d'indicateurs (de pilotage ou, dans ce cas, de qualité des données) n'est pourtant pas nouveau, il est d'ores et déjà mis en lumière comme une déclinaison de la stratégie de l'entreprise sur des pratiques métier plus restreintes, plus opérationnelles. Ici, les chercheurs en informatique vont un cran plus loin en déclinant cette spécificité à la maille de chaque utilisateur, ou familles homogènes d'utilisateurs, ce qui permet de prendre en compte les spécificités des usages et de converger avec les Sciences de Gestion. Mais cette approche ne permet toujours pas d'aborder la richesse de la dimension anthropologique et de l'exploration d'un contenu porteur de sens.

3.3.1.3 Les SIC : une approche de la qualité des données orientée sur le sens

Face à cette approche informatique, qui cherche à objectiver la qualité des données et les modèles de qualité tout en essayant de les ajuster en fonction des priorités de l'entreprise, les Sciences de l'Information et de la Communication adoptent une posture plus anthropologique favorisant la génération de sens pour les utilisateurs. Cette génération de sens passe par l'étude et l'amélioration des éléments descriptifs permettant de faciliter l'exploration des données à la recherche d'informations utiles. L'approche passe, par exemple, par la constitution de thésaurus de documents, mis à l'épreuve à l'ère du numérique autant sur la phase d'indexation que de

[14] Critères de qualité des systèmes data issus du cours MDM – Béa Arruabarrena, 2018-2019.

recherche (Dalbin, 2007). Elle s'intéresse aux ontologies, aux lexiques, aux dictionnaires de données, ou plus récemment aux folksonomies ou aux indexations collectives : ces différents concepts pointent une différence profonde entre la définition de la notion de « métadonnée » au sens informatique, technique, et celle des Sciences de l'Information et de la Communication, qui visent en priorité la facilitation de l'appréhension du sens des données par l'utilisateur. Ainsi une donnée de qualité est une donnée documentée, c'est-à-dire comportant des métadonnées utiles à la compréhension du sens. La qualité des métadonnées elles-mêmes dans ce cadre est alors une notion clé.

Or, l'assouplissement de l'accès à l'information par l'utilisateur, l'explosion de documents numériques et numérisés en parallèle d'un maintien des documents papier, ou encore le poids croissant des éditeurs ont soulevé de nouveaux défis pour cette discipline marquée par un renouvellement des pratiques professionnelles sur la sélection, la qualification et la mise à disposition des documents (Battisti et al., 2010; Dacos & Mounier, 2010). Ces préoccupations constantes n'attendent pas le buzz Big Data pour naître, et se poursuivent tout le long du phénomène, comme le montre notamment le Programme National de Numérisation et de Valorisation des Contenus Culturels, dans le cadre duquel le ministère de la culture sensibilise les porteurs et les évaluateurs de projets aux enjeux liés aux métadonnées et à la standardisation[15]. Les métadonnées sont alors considérées comme un vecteur majeur de réduction de « bruit et de silence ». Dans le cadre des ressources informationnelles du web, les métadonnées visent la « redocumentarisation » (Broudoux & Scopsi, 2011), c'est-à-dire la recomposition des médiations documentaires (contexte, processus complexe et implication d'acteurs multiples liés à la génération du document à la disposition du lecteur). Les métadonnées peuvent alors être interprétatives (ajout par le praticien) ou intrinsèques (extraction automatique), et sont marquées par un manque de standards accessibles. Dans le milieu du livre numérique, les métadonnées font l'objet d'un développement d'une économie nouvelle (Odeh & Chartron, 2016) avec une reconfiguration possible de la valeur ajoutée des bibliothèques, des producteurs et des intermédiaires, dans la mesure où la visibilité et l'accessibilité des livres dépendent de leurs métadonnées. Cette nouvelle économie des métadonnées pointe des perspectives d'orientation de ses métadonnées sur l'œuvre, sur l'autorité, ou encore sur la recommandation.

[15] http://www.culture.gouv.fr/Media/Thematiques/Innovation-numerique/Folder/Livrables-GT-Numerisation/Les-enjeux-des-metadonnees-et-des-standards

En entreprises, la qualité des données fait l'objet d'une approche similaire grâce à une orientation sur le sens des données, et l'arrivée massive de données hétérogènes à travers la mobilisation de nouvelles technologies perturbe leur sémantisation (mise en place de Data Lakes sans qualification préalable du sens des données, ajout de données brutes issues des objets connectés, multiplication de données statistiques...). Or, les Sciences de l'Information et de la Communication n'abordent que peu l'impact de ces nouvelles technologies et méthodes analytiques dans les entreprises, notamment dans le cadre de projets data. Pourtant, les concepts traités dans cette discipline inscrivent la notion de qualité des données au cœur de ces projets, dispositifs de production d'informations utiles caractérisés par un aspect non seulement informatique, mais aussi cognitif. Ils renvoient la gestion de la qualité des données dans ses dimensions applicatives, psycho-sociales, éthiques et contextuelles (Arruabarrena et al., 2019) autant que techniques.

3.3.2 Enjeux de gouvernance à l'échelle de l'entreprise

Les enjeux liés à la qualité des données, des informations et des documents s'inscrivent déjà dans des réflexions à la frontière entre les Sciences de Gestion et les Sciences de l'Information et de la Communication, tout en s'appuyant sur des avancées techniques et analytiques. En effet, son impact est conséquent sur l'entreprise. « [Le] document représente [] des enjeux stratégiques (assurer la continuité de l'activité, la confidentialité, la capitalisation de l'information), des enjeux juridiques (répondre aux obligations légales, apporter des preuves lors de litiges), des enjeux économiques (pallier le coût d'une perte ou de la non fiabilité d'un document, d'un stockage inutile, d'une recherche trop longue). » (Battisti, 2017). Les enjeux s'élargissent de plus en plus du document à l'ensemble des données de l'entreprise.

A l'heure où la donnée est perçue de plus en plus comme un actif incontestable d'une entreprise, voire l'un de ses principaux actifs immatériels, comme l'annonce le rapport CIGREF de 2014[16], la gestion de la qualité des données fait l'objet de nombreux travaux de recherche théorique et opérationnelle. Inspirées des Sciences de Gestion, ils adaptent l'approche aux spécificités du sujet et à son évolution à l'ère de la transformation numérique et du phénomène Big Data. Trois enjeux sont plus particulièrement développés : la nécessité d'une gouvernance propre à la dimension patrimoniale des données, la prise en compte des spécificités liées au cycle de vie des données et des documents, et enfin les méthodes de valorisation des données, circonscrites

[16] « Enjeux business des données ». 2014. CIGREF. https://www.cigref.fr/rapport-cigref-enjeux-business-des-donnees.

dans un cadre éthique. Ces enjeux, identifiés comme une opportunité historique pour les professionnels de l'information dès 2013 dans le dossier spécial de la revue Documentaliste (Jules & Lebigre, 2013), guident la gestion de la qualité des données dans un processus d'amélioration continue, contrairement aux projets data qui ne s'intéressent que ponctuellement à la qualité des données, dans le cadre d'un usage donné. Ainsi, restreindre l'étude de la qualité des données à l'impact des projets data sur ces usages limités peut empêcher une génération de valeur plus globale pour l'entreprise.

Appliqué plus spécifiquement aux données clés de l'entreprise, c'est-à-dire essentielles à sa performance (comme ses données client dans les CRM, ou encore les données nécessaires à la prise de décision générées dans les outils de Business Intelligence), le processus de qualité des données est communément géré par la fonction Master Data Management, ou MDM (voir Figure 16). Cette fonction « incorpore les applications métier, les méthodes de gestion de l'information, et les outils de gestion des données pour implémenter les politiques, les procédures, et les infrastructures qui supportent la capture, l'intégration, et l'usage partagé qui en résulte des données maître exactes, opportunes dans le temps, cohérentes et complètes » (Loshin, 2010). Son objectif est d'« augmenter la performance de l'entreprise (en ajustant la valeur des données) et diminuer les coûts liés au traitement et à la gestion des données maîtres » (Mariko, 2016).

Si, pour les sciences de l'informatique, le MDM correspond essentiellement à un modèle physique de données optimal, sa dimension sémantique reste négligée et aucun standard n'est établi aujourd'hui en termes de description des données. Cependant, le MDM fait l'objet de nombreux travaux de recherche et de publications de praticiens, mettant en évidence ses facteurs d'efficacité, les plus récents (Vilminko-Heikkinen & Pekkola, 2017) incluant bien la compréhension mutuelle des domaines de données maîtres, mais aussi les challenges guidés par la législation, ou encore le niveau de granularité.

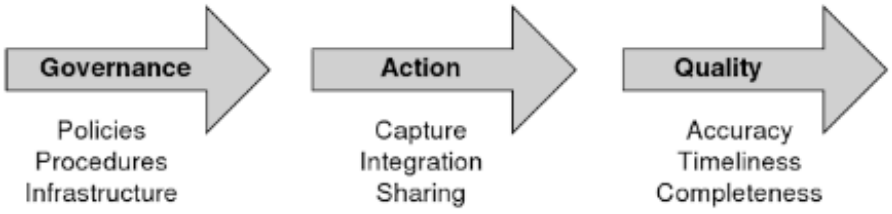


Figure 16 – Eléments essentiels du Master Data Management - *Source : Loshin, 2010.*

La complexité du MDM freine d'ores et déjà son application dans les entreprises sur les données de référence. Cela devient d'autant plus problématique à l'ère du Big Data, notamment à travers l'accès facilité à d'autres sources de données internes et externes et les difficultés de sémantisation. Cette infobésité touche plus largement la gestion de la qualité dans le domaine de la gestion info-documentaire, qui tend en France vers des standards internationaux et transversaux de gestion de la qualité, comme en témoigne la publication thématique I2D (Nesme & Cottin, 2017a) reprenant les normes ISO 9000:2015 et invoquant 7 principes de gestion de qualité, mis à l'épreuve à l'ère digitale.

1. **La génération de valeur pour le client -** Les organisations dépendent de leurs clients et doivent donc comprendre leurs besoins actuels et futurs, y répondre, et s'efforcer de dépasser leurs attentes. Or, le client voit son rôle d'évaluateur de la qualité augmenter au détriment de celui de l'intermédiaire (professionnel de Knowledge Management, de veille, de gestion d'archives...), ce qui réoriente la notion de qualité des critères classiques (cycle de vie des documents, sécurité...) vers des critères de perception de qualité (facilité d'accès, ergonomie des outils collaboratifs...). Ces nouveaux critères génèrent eux aussi des bénéfices directs ou indirects (Nesme & Cottin, 2017b) pour l'organisation, et doivent être intégrés dans la gestion de la qualité par le métier sans toutefois entraver les critères de qualité traditionnels.

2. **Le leadership -** Les dirigeants établissent une orientation stratégique partagée pour l'organisation. Ils doivent créer et maintenir l'environnement interne dans lequel les collaborateurs peuvent participer pleinement à la réalisation des objectifs de l'organisation. La nature transversale de l'information au sein d'une organisation nécessite une gouvernance capable de mobiliser les acteurs aux intérêts différents autour d'une stratégie de qualité commune, répondant aux enjeux et à la culture de l'organisation. Au-delà du rôle de décideur, ancré dans un environnement collaboratif complexe, le leader a la responsabilité « d'influencer et fédérer des acteurs autour d'un but commun dans une relation de confiance mutuelle » (Jules, 2017). Cette génération de confiance nécessite des compétences pédagogiques plus spécifiques à la transformation digitale.

3. **L'engagement des collaborateurs -** Les personnes à tous les niveaux sont l'essence d'une organisation et leur implication complète permet à leurs capacités d'être utilisées au service des bénéfices de l'organisation. Dans ce contexte, la direction définit une stratégie de qualité et les rôles et les responsabilités qui la soutiennent, et l'encadrement anime la mise en œuvre

de la stratégie de qualité sur son périmètre. Quant aux collaborateurs, ils sont impliqués de deux manières : soit dans l'amélioration de la qualité des informations dont ils sont responsables, soit dans la définition des critères de satisfaction du service de mise à disposition d'informations qu'ils utilisent dans le cadre de leur activité (Nesme, 2017).

4. **L'approche par le processus -** Un résultat souhaité est obtenu plus efficacement lorsque les activités et les ressources relatives à ces activités sont gérées en tant que processus. Etape normalisée, inscrite ou non dans le cadre d'une démarche de qualité, l'analyse des processus permet d'identifier et de caractériser les documents qui prouvent le déroulé d'une activité. La gestion des métadonnées des documents produits au cours des activités, réalisée selon les méthodes de *records management*, est alors de résultat d'un travail collaboratif de compréhension de l'activité terrain (séquences, fonctions…) et de sa confrontation avec les exigences contractuelles, réglementaires ou financières (Brahim, 2017).

5. **L'amélioration** - L'amélioration de la performance globale de l'organisation doit être un objectif permanent de l'organisation. Les actions d'amélioration découlent du principe précédent qui permet d'affecter les ressources aux processus, d'établir des critères d'évaluation de la valeur générée, et de mettre en place des moyens de détection et de pilotage de l'amélioration.

6. **La prise de décision basée sur des preuves -** Les décisions efficaces sont basées sur l'analyse des données et des informations. Qu'il s'agisse de fiabiliser ou de légitimer la prise de décision, la notion de preuve renvoie à la véracité démontrable d'une information, « fondée sur des faits obtenus par observation, mesurage, essai ou autres moyens »[17]. Ce principe renvoie à la notion de traçabilité qui s'inscrit dans le cadre d'audits divers, propres à un métier, à une fonction, ou bien à un type de données, comme c'est le cas des données à caractère personnel dans le cadre de réglementations nationales ou européennes[18]. Il commence à se heurter à la volonté des décideurs d'accéder directement à l'information source au lieu d'être restreint à l'accès à un résultat de travail séquentiel de synthèse (Cottin, 2017).

[17] http://www.qualiteonline.com/definition-193-preuve-tangible.html
[18] https://www.cnil.fr/fr/reglement-europeen-sur-la-protection-des-donnees-ce-qui-change-pour-les-professionnels

7. **La gestion de la relation avec les prestataires -** Une organisation et ses fournisseurs externes sont interdépendants et une relation mutuellement avantageuse améliore la capacité à créer de la valeur. Le choix du prestataire, qu'il s'agisse d'une prestation de production ou de conseil, est étroitement liée à la définition du besoin et du périmètre d'intervention attendu, définissant la nature du contrat, les livrables et les modalités d'allocation du budget.

Les 7 principes sont regroupés dans ces travaux sous les notions de gouvernance (décideurs, contributeurs et prestataires), d'efficience interne (cohérence et amélioration des processus), et d'efficacité externe (usage et prise de décision légitime). Ils dessinent les grandes lignes des enjeux actuels de la qualité des données et sont à mettre en perspective avec le Big Data.

En effet, d'un côté ce phénomène présente des opportunités de renouvellement, à travers de nouveaux usages et outils de traitement et de restitution des données, notamment pour contribuer à la qualité perçue. Et d'un autre côté, il semble être porteur d'un certain nombre de risques : catalyse d'infobésité, complexification du leadership et de la contribution des collaborateurs par une maturité insuffisance pour générer la confiance transversale, incompréhension des métiers des prestataires de l'Ecosystème Big Data… Par ailleurs, ces risques, associés aux lacunes d'entendement des modèles algorithmiques, nuisent à la perception de la qualité par les clients ou par les utilisateurs internes. Ils sont progressivement soumis à de nouvelles règlementations, plus particulièrement sur les données personnelles dans le cadre du RGPD[19], doté d'un dispositif punitif assez dissuasif. Ce règlement européen assoit les principes d'un traitement licite, loyal et transparent au regard de la personne physique concernée, et ce sur l'ensemble du cycle de vie des données personnelles. En découlent les règles de consentement, de portabilité des données, le droit à l'effacement, les limitations du profilage automatique, la mise en place de guichet unique ou encore le droit à l'information en cas de piratage.

Ainsi, les enjeux de la gestion de la qualité des données s'inscrivent dans un système à triple contrainte : la contrainte de la valeur de l'usage, la contrainte de coût de la qualité, et les contraintes externes. Dans le contexte d'évolution des enjeux, deux facteurs en particulier

[19] Règlement Général sur la Protection des Données, https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32016R0679

attirent l'attention au cours de la découverte des premiers cas terrain : la qualité des algorithmes, et notamment leur intelligibilité, et l'adoption des nouvelles technologies Big Data.

3.3.3 L'algorithme, un nouveau type de modèle de données à qualifier

L'« algorithme » est un modèle, c'est-à-dire une suite d'instructions, finie et non ambiguë, qui prend en entrée des éléments issus d'un ensemble spécifié, et donne en sortie des éléments ayant une relation spécifiée avec les entrées. L'objectif d'un algorithme est de fournir la solution à un problème sous la forme d'un résultat. Dans ce cadre, une formule de parfum, un tarif d'assurance, un rapport financier, un plan de vol ou une simple addition sur calculette sont des algorithmes. La CNIL évoque à ce titre la métaphore de la recette de cuisine pour décrire un algorithme : en effet, il s'agit d'utiliser une liste d'ingrédients selon des étapes de préparation pour obtenir un plat. Dans le cas d'un algorithme informatique, il s'agit d'un modèle de données qui se présente sous la forme d'un code informatique, c'est-à-dire une séquence d'instructions dans un langage programmatique à appliquer à un ensemble de données spécifié pour obtenir un résultat analytique.

Les algorithmes sont utilisés par les entreprises dès la mise en place des systèmes d'information. Alimentés par des ensembles de données, leurs résultats peuvent à la fois être des relations (sous la forme d'une formule, c'est-à-dire d'une association entre des éléments entrants et sortants), ou un nouvel élément en sortie (une valeur, un attribut, une entité…). Cette définition de l'algorithme comme modèle de données le rend parfaitement compatible avec une gestion de qualité des données classique dans un système d'information. Cependant, un glissement de la définition commune des algorithmes s'opère au cours des dernières années au profit des algorithmes issus du phénomène Big Data (application sur des données de type image ou voix, Deep Learning, algorithmes auto-apprenants…) (voir Annexe 4 - Data Science et Algorithmes). Ce décalage génère un questionnement méthodologique et des craintes alimentées par des exemples d'applications algorithmiques sur des données personnelles à des fins commerciales (O'Neil, 2018). Face à ces craintes, une étude est demandée au Conseil Général de l'Economie : cette étude vise plus particulièrement les algorithmes mis à disposition sur les plateformes fournies par Google (moteur de recherche), Amazon (moteur de recommandation), et plus généralement les nombreuses plateformes émergeantes dans l'Ecosystème Big Data ou hébergeant les algorithmes au sein des entreprises. En effet, le rapport affirme que « ces algorithmes sont inséparables des données qu'ils traitent et des plateformes qui les utilisent pour proposer un service » (Pavel & Serris, 2016). La responsabilité des éditeurs de ces plateformes,

au sens plus large, est alors évoquée. Cette question pointe la possibilité d'une déviation de certains usages, pouvant mener à un dysfonctionnement de la concurrence, notamment sur les marchés financiers, à une protection amoindrie des consommateurs, au risque de discrimination, ou encore à la perte de confiance dans l'économie numérique. Elle soulève le besoin d'une réflexion éthique et sociétale.

Ces craintes conduisent à une recherche de position française et européenne à ce jour inaboutie, et à l'affirmation de principes fondateurs comme la loyauté et la vigilance par la CNIL dès 2017 (Falque-Pierrotin et al., 2017). Ces principes trouvent un aboutissement opérationnel à travers un ensemble de recommandations, comme la formation des acteurs impliqués sur toute la chaîne à l'éthique, la médiation entre les utilisateurs pour rendre les algorithmes plus compréhensibles, l'asservissement des algorithmes à la liberté humaine et à l'intérêt général dès la phase de conception, ou encore la constitution d'une plateforme nationale d'audit des algorithmes. La capitalisation sur les atouts français, comme un ensemble de valeurs culturelles orientées sur l'homme et l'éthique ou encore la qualité des formations en sciences de l'ingénieur et en mathématiques, est d'ores et déjà en cours de mobilisation, comme le montre le rapport annuel de l'INRIA et ses publications en 2017[20].

La transparence des algorithmes dans ce cadre présente des enjeux clés pour les années à venir, et donne lieu au lancement de projets nationaux, comme TransAlgo[21] qui vise à centraliser les travaux de recherche transdisciplinaires sur l'« auditabilité des algorithmes et le développement de nouvelles générations d'algorithmes "transparents par construction" qui facilitent la mesure de leur transparence, leur explication et la traçabilité de leur raisonnement ». Le lien avec le monde professionnel commence à donner ces fruits dans les médias et la gestion des contenus, et des pratiques communes émergent au service de la transparence des algorithmes, c'est-à-dire la « façon dont les acteurs à la fois à internes et externes au journalisme ont la possibilité de surveiller, vérifier, critiquer et même intervenir dans le processus journalistique. » (Diakopoulos & Koliska, 2017). Les travaux sur ce terrain démontrent clairement un lien indissociable entre les humains et les algorithmes au cours de leur phase de construction, mais aussi de leur application continue, et mettent en évidence un ensemble de facteurs de la transparence pointant sur les données comme matière première, le modèle en soi, les inférences

[20] https://www.inria.fr/centre/saclay/actualites/marc-schoenauer-epaule-cedric-villani-pour-definir-une-strategie-ia-pour-la-france
[21] https://www.inria.fr/actualite/actualites-inria/transalgo

et les interfaces. L'implication humaine tout le long fait partie des facteurs contribuant à la transparence, que ce soit à travers le choix des données pertinentes, la définition des méthodes, la documentation des données et des sources, les hypothèses de collecte, la documentation du nom et du type de modèle algorithmique, les arbitrages tout le long du projet, ou encore la justification du niveau de confiance dans les résultats.

Si le sujet de la transparence des algorithmes s'inscrit dans un rapport de force qui dépasse largement les projets data sur le terrain, il nécessite tout de même une montée en maturité dans laquelle ces projets peuvent représenter un vecteur concret pour les entreprises. Or, aucun cadre n'est disponible début 2019 pour les entreprises afin d'avancer dans ce sens, et ce encore moins au cours de la période étudiée ici, entre 2015 et 2017. Les algorithmes sont alors abordés de façon beaucoup plus terre à terre : d'ailleurs, si les facteurs de transparence exprimés en 2017 renvoient à une documentation dite propre aux algorithmes, ils ne semblent pas apporter de nouveautés par rapport à la documentation plus habituelle de la qualité des données (sémantisation des données collectées et exploitées, métadonnées techniques, documentation du processus de conception et d'exploitation…).

Pour faire le point sur ce sujet, il faut distinguer quatre phases de vie pour un algorithme :

- La première est sa naissance, à issue d'un travail de recherche théorique de création d'algorithmes en tant qu'objets mathématiques nouveaux. Historiquement, les algorithmes sont conçus dans des laboratoires, académiques ou privés, de recherche mathématique ou informatique. Dans le cadre de ces travaux, le processus de création d'un algorithme mathématiquement et conceptuellement nouveau est écarté.

- La seconde est sa mise à disposition dans l'éventail des algorithmes mobilisables. A ce stade, un algorithme est doté d'un ensemble de propriétés : il est supervisé ou non supervisé, paramétrique ou non, vise un résultat discret ou continu... Il est associé aux problématiques mathématiques qu'il sert à résoudre à partir d'un certain type de matière première (Brownlee, 2013). Les algorithmes mobilisables sont catégorisables, comme le montrent certaines cartes heuristiques (voir Figure 17), et donc qualifiables *a priori*, en fonction des catégories de structure de leur source, des structures de contrôle, ou encore du résultat attendu. Ces catégories peuvent être complétées par des indicateurs de qualité théoriques, comme les notions de correction, de complétude, de terminaison, ou encore de complexité, qui régit, entre autres, sa

scalabilité. Les algorithmes et leurs propriétés sont de plus en plus stockés sous la forme de librairies, souvent en Open Source. A ce stade, ils ne sont pas alimentés par des données.

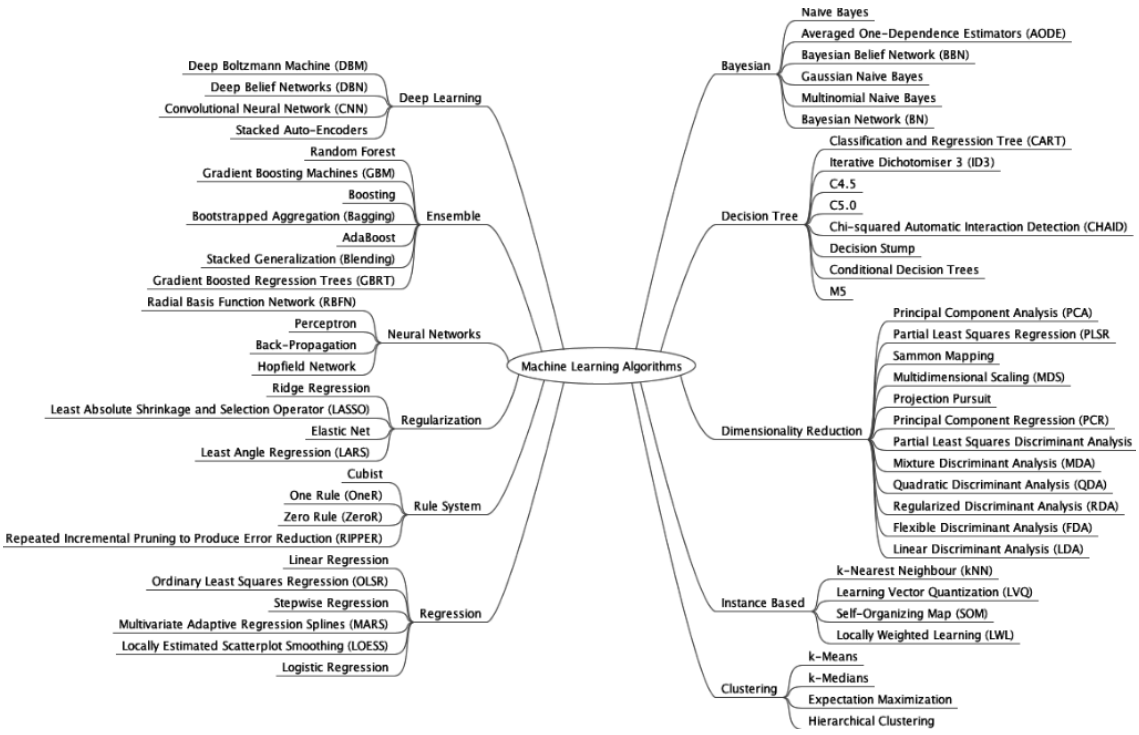


Figure 17 – Carte heuristique des algorithmes de machine learning - *Source : Brownlee 2013*

- La troisième phase de la vie d'un algorithme est l'apprentissage sur les données, c'est-à-dire son alimentation par un ensemble de données et son adaptation pour la résolution d'un problème concret. Les propriétés des algorithmes mobilisables permettent alors de se diriger vers des familles d'algorithmes pertinents selon le contexte précis du problème à résoudre, mais ne sont pas suffisantes pour identifier l'algorithme optimal. Ce choix est effectué au cours de la phase d'apprentissage par benchmark comparatif entre algorithmes et par ajustement des données et des instructions. Cette phase aboutit à l'algorithme proprement dit, c'est-à-dire un modèle de données, et il peut être qualifié en termes de 3 éléments qui le composent : la nature informatique et sémantique des données en entrée, la nature des instructions et des paramètres de ces instructions, et la nature et les critères d'évaluation (statistiques et métier) des résultats qu'il génère dans le cadre de la résolution de la problématique en question. Il s'agit d'une qualification comparative, *a posteriori*. Si le résultat est jugé de qualité suffisante, l'algorithme est bien une solution algorithmique au problème.

- Enfin, les algorithmes ainsi « entraînés » sur des ensembles de données et évalués peuvent être stockés et utilisés : ils produisent alors des résultats à chaque fois qu'ils sont appliqués à

de nouveaux ensembles de données dotées de la même nature et structure que celles utilisées pour l'apprentissage. Dans le cas où l'algorithme n'a été utilisé qu'une seule fois (par exemple pour générer des connaissances à partir d'un historique figé), cette quatrième phase de vie n'a pas lieu d'être. Dans le cas où l'algorithme est auto-apprenant, les paramètres peuvent évoluer au fur et à mesure qu'il s'alimente de nouvelles données (ajustement automatique en fonction de nouvelles tendances détectées à partir des données mises à jour).

Un algorithme cumule ainsi tout le long de sa vie des caractéristiques complémentaires : d'abord ses propriétés intrinsèques, puis des paramètres résultants de sa phase d'apprentissage, et enfin ses paramètres finaux lorsqu'il est appliqué à de nouveaux ensembles de données. Cette accumulation de critères doit alors faire l'objet d'une documentation qui expliciterait les choix humains réalisés à chaque étape. Il ne s'agit donc pas uniquement des caractéristiques techniques, mais bien des traces du travail d'interaction entre l'homme et la donnée. Tout comme pour les données au sens plus classique, ces caractéristiques peuvent se présenter sous la forme de « métadonnées », liées à leurs propriétés *a priori*, à l'évaluation de leur qualité *a posteriori*, et au processus de transformation (informatique et cognitif). La gestion des algorithmes semble alors s'inscrire au cœur de la gestion de la qualité des données, voire dans le processus continu du Master Data Management si les usages qu'ils soutiennent sont clés pour l'entreprise.

Peu de travaux couvrent les enjeux liés à cet élargissement de la gestion de la qualité des données, alors que certaines perturbations semblent exister, au-delà de l'aspect hétérogène et volumineux des données qui les alimentent. Tout d'abord, la méconnaissance de ces algorithmes, et plus particulièrement des critères d'évaluation statistiques de leur qualité, semble poser un problème pour la validation de la pertinence d'une solution algorithmique au sein d'un usage. Les critères statistiques seuls paraissent en effet extrêmement arides. Ensuite, la situation se complique dans le cadre des algorithmes dits « auto-apprenants », dotés d'une fonction de générateur de formule optimale au cours d'un apprentissage automatisé. Dans ce cadre, les métadonnées générées au cours de l'utilisation évoluent de façon autonome et nécessitent un suivi. Ce sujet est d'autant plus sensible qu'il s'applique à des données personnelles, et notamment les données client dans les entreprises en BtoC, soumises aux exigences de transparence (voir Annexe 5 – Transparence des algorithmes). L'algorithme en tant qu'objet questionne alors les pratiques de gestion de la qualité des données, et notamment leur sémantisation, et nécessite une prise en compte plus particulière.

3.4 Médiation Homme-Données et co-construction de sens

La notion d'usage s'inscrit à l'intersection d'enjeux stratégiques et opérationnels, et s'appuie sur un socle de pratiques métier historiques pour mieux s'adapter à l'environnement changeant de l'entreprise. Cette adaptation passe habituellement par certaines fonctions spécifiques en entreprise, comme le marketing et l'innovation, mais aussi l'intelligence économique ou encore le Knowledge Management, dont l'implication dans le phénomène Big Data semble secondaire, si ce n'est inexistante, à l'aube du phénomène. L'usage est par ailleurs porteur d'enjeux plus vastes : éthiques, juridiques, financiers, économiques, techniques, organisationnels, communicationnels, qualitatifs et autres, ce qui induit la mobilisation d'acteurs aux compétences hétérogènes pour la confirmation de son potentiel. De même, les projets data sont marqués par la variété et l'hétérogénéité des sources de données, la nouveauté de la démarche et de la nature complexe des résultats, impliquant une mobilisation de compétences plus spécialisées afin d'évaluer la faisabilité de l'usage et de le mettre en œuvre. Enfin, les acteurs non-humains, comme les interfaces entre l'homme et la donnée ou les algorithmes, se multiplient et s'ajoutent aux acteurs humains engagés dans ces projets. Ces éléments dessinent une complexité de mise en musique, et constituent des facteurs de risque de mauvaise circulation de connaissances, ce qui peut mener vers un usage impossible (pas de transformation opérationnelle des résultats), inadéquat (génération de valeur insuffisante), voire dangereux (dégradation de la valeur ou prise de décision mettant en péril l'entreprise). Il est alors indispensable de mettre en lumière les nouvelles compétences mobilisées et les modalités de circulation des connaissances pour la co-construction de sens au cours de ces projets.

3.4.1 Acteurs et cadre de compétences mobilisées

Trois types d'acteurs clés sont présentés ici : les fonctions historiques de médiation (Intelligence Economique et Knowledge Management), les médiateurs humains et techniques et enfin les nouveaux acteurs intervenant dans les projets data : les Data Scientists.

3.4.1.1 Intelligence Economique et Knowledge Management, en retrait

L'histoire de la donnée présentée dans l'introduction du contexte de ces travaux pointait un phénomène large et un lien indissociable entre la politique et le développement des disciplines et des technologies analytiques. L'exploitation de la donnée devait alors fournir des informations utiles pour la prise de décision des hommes d'état, que ce soit en temps de paix ou en temps de guerre. Dans la première partie de l'état de l'art sur la valeur et la qualité des

données, plus orienté sur les entreprises, un consensus se dégageait sur le fait que l'information n'avait de sens que lorsqu'elle permettait de guider la prise de décision dans le cadre d'un usage. En entreprise ou en politique, l'information utile, transformable en action, procure un avantage. Lorsque, en 2013, l'Etat français souhaite favoriser la compétitivité et promouvoir une coordination entre les acteurs publics et privés, il s'intéresse plus particulièrement à l'internalisation par les entreprises de l'Intelligence Economique. Celle-ci est issue du paradigme du « renseignement », pratique empirique plutôt qu'objet théorique dont les contours sont encore mal définis et pluridisciplinaires (Bulinge & Boutin, 2015). Défini par le PIA 02-200 comme le « résultat d'un processus d'exploitation de données et d'informations dont la collecte a été orientée et décidée pour répondre à un besoin décisionnel », il correspond, dans la chaîne de valeur de la donnée décrite dans les chapitres précédents, à l'information utile, c'est-à-dire évaluée et exploitée. Il s'agit d'une connaissance produite par un dispositif sociotechnique au sein d'une organisation.

Le rapport Martre, établi dans ce contexte en France, définit l'information utile comme « celle dont ont besoin les différents niveaux de décision de l'entreprise ou de la collectivité, pour élaborer et mettre en œuvre de façon cohérente la stratégie et les tactiques nécessaires à l'atteinte des objectifs définis par l'entreprise dans le but d'améliorer sa position dans son environnement concurrentiel » (Martre et al., 1994). Il s'agit de mettre à disposition les informations pertinentes aux preneurs de décisions au moment de cette prise de décision : ce processus fait l'objet de l'intelligence économique, qui remplit les fonctions de veille, de protection des informations et d'influence. Résolument tournée vers l'environnement externe de l'entreprise et la construction de l'avantage concurrentiel grâce à l'établissement de liens avec les données internes, notamment à travers les outils décisionnels (Business Intelligence), l'intelligence économique se distingue de la Gestion des Connaissances (Knowledge Management), orienté vers les connaissances existantes et circulantes au sein même de l'entreprise. Cette démarche stratégique vise à atteindre un objectif fixé grâce à une exploitation optimale des connaissances de l'entreprise. Comme le métier du renseignement, les deux fonctions sont pluridisciplinaires et multidimensionnelles (Bulinge & Boutin, 2015).

Vu la mise en perspective des projets data, ceux-ci impacteraient à la fois l'Intelligence Economique grâce à de nouvelles informations utiles à la prise de décision, et la Gestion des Connaissances, grâce à la capitalisation de connaissances qui rend possibles des usages futurs et la valorisation des informations internes. Par ailleurs, les nouvelles possibilités d'analyse

combinatoire d'informations internes et externes à l'entreprise présagent un effacement de frontière entre ces deux fonctions. En effet, l'intelligence économique, « avatar du renseignement » en entreprise (Bulinge & Moinet, 2016) et la gestion des connaissances partagent la notion d'utilité d'une information et sa distinction de l'information brute, constituée de données prises en dehors de leur contexte. Or, à l'heure où les directions Marketing semblent s'être approprié le terrain des projets data en entreprise, les deux fonctions ne sont que très peu intégrées dans les différents modèles projet, alors qu'elles garantissent cette construction de sens, c'est-à-dire l'aspect cognitif du travail réalisé tout le long de la transformation de la donnée brute en informations utiles. Des dispositifs d'intelligence économique et de gestion des connaissances assurent habituellement un rôle de médiateur entre la donnée et leurs utilisateurs : l'hypothèse de ces travaux de recherche consiste à croire que cette médiation existe bel et bien dans les projets data sous une forme renouvelée, et joue un rôle de guide pour une convergence vers l'usage.

3.4.1.2 Médiateurs humains et techniques

Le processus d'interprétation au service de l'action fait l'objet d'un ensemble de recherches dans les Sciences de l'Information et de la Communication qui définissent et développent le concept de médiation. Il ne s'agit pas d'une simple transmission d'information, mais bien de l'accompagnement dans la convergence, parfois brutale, de deux mondes (Olivesi, 2014), c'est-à-dire du monde des données, doté d'une culture basée sur un ensemble de mythes, d'évolutions technologiques et analytiques et de pratiques sociales, confronté avec le monde de l'entreprise. A la différence d'une interaction au sens commun, la médiation exclut le fait que les objets soient constitués en amont de l'établissement de lien communicationnel, et permet d'admettre la construction interactive de l'objet. La médiation est alors une condition à la convergence vers un usage à travers l'exploration.

Historiquement, la médiation documentaire fait partie des médiations des savoirs, c'est-à-dire l'intervention d'un tiers qui concilie l'information et la communication pour l'accompagnement de l'usager dans l'accès à l'information et son usage. Le tiers agit comme interface ou met à disposition une interface technique, ce qui inclut dans la médiation des savoirs la médiation humaine et la médiation technique. Fondée sur la maîtrise des techniques et outil documentaires, la médiation des savoirs est un processus support à la création de valeur, un service (Espaignet et al., 2003) par nature dépendant des compétences relationnelles du professionnel de l'information face à l'utilisateur de l'information. Le médiateur se positionne

comme interface entre les « données recherchées par un demandeur » et « les connaissances contenues dans les documents » (Sutter, 2005). L'objection possible, qui consiste à dire que dans un projet data le niveau de maturité du demandeur ne lui donne pas la capacité à formuler une demande précise, est levée dans le champ de la documentation à travers la distinction entre une demande et un besoin : la demande dissimule en effet un ensemble d'intentions, d'objectifs, de contextes, de maturité des connaissances. Le rôle du médiateur consiste alors à questionner le demandeur afin de formuler le besoin, grâce à la mobilisation de ses compétences de communiquant et de conseiller. Par ailleurs, Sutter met en évidence la médiation comme un dispositif qui « exploite des sources formelles et ouvertes d'information, généralement externes, et rediffuse des informations « brutes » même si des traitements à valeur ajoutée sont effectués en matière de sélection, de mise en forme ». Ce recul par rapport à la nature « brute » de l'information s'inscrit tout à fait dans la problématique évoquée sur les projets data, sans toutefois notifier les activités de génération de données nouvelles liées aux algorithmes. Enfin, le médiateur, en tant que garant de la communication optimale de l'information, se fait attribuer un rôle d'évaluateur de la valeur créée par l'accès à l'information.

L'implication des professionnels de l'information dans le management global de l'information a démontré un besoin de repenser le positionnement des acteurs de médiation, en particulier à la suite du mouvement de désintermédiation opérée au cours des années 1980-90. Dans le cadre d'un centre documentaire, ce repositionnement modifie l'activité documentaire, la médiation et les compétences des documentalistes (Galaup, 2017; Panissier, 2007) avec l'ajout dans l'activité de la veille informationnelle et au développement des services d'accès à l'information, notamment à distance. En effet, l'ouverture de l'usage des métadonnées des documents auprès des agents, et non pas seulement auprès des documentalistes, a nécessité de mettre en place des moyens de contextualisation, c'est-à-dire une description du document « en fonction du contexte dans lequel le document sera appréhendé » (Lamouroux & Ferchaud, 2008). Ce changement s'inscrit dans l'introduction de la complexité, la transversalité des tâches (à la fois unification des pratiques d'archivage, de gestion de l'information, de veille, de gestion des connaissances, et spécialisation selon la nature des documents), et l'exigence de qualité de la part des utilisateurs. Les projets data semblent *a priori* suivre un mouvement similaire de désintermédiation, à travers l'absence des professionnels de l'information dans les équipes projet composées d'acteurs métier et des Data Scientists. En effet, aucun modèle de référence ne mentionne les métiers de la connaissance, sauf le modèle DMLC (Data Mining Life Cycle), seul qui inclut le profil de Knowledge Engineer dans une équipe projet data. Il faut rappeler que

ce modèle est aujourd'hui théorique (Hofmann & Tierney, 2003, 2009). Par ailleurs, l'évolution des métiers du document montre une décentralisation progressive grâce à l'introduction des outils informatiques, un décalage depuis les réservoirs informationnels vers des métiers d'accompagnement méthodologique et technique de l'utilisateur (Prévot-Hubert, 2004), alors que l'exercice de la Data Science semble aujourd'hui se recentrer autour des DataLabs et des DataFabs en entreprises, avoir une fonction productive, et garder la main sur les méthodes et les nouvelles technologies. Cela questionne la présence de nouveaux dispositifs de médiation, internes aux projets data et la pérennité de cette organisation centralisée.

Au-delà du rôle de médiateur des acteurs humains, la présence d'interfaces techniques n'est pas une nouveauté dans les Sciences de l'Information et de la Communication. Leur maîtrise constitue l'un des fondements du travail des documentalistes, et les bibliothèques et les musées ont été les premiers à mettre en œuvre les nouveautés technologiques comme les outils du Web 2.0, intégrant les médias sociaux dans le dispositif de médiation (Besset, 2011). La dimension médiatique des moteurs de recherche web fait l'objet d'analyses approfondies qui pointent son impact sur la rationalité des pratiques de construction de sens, d'appropriation de l'information par les agents, et sur le paradigme « acteur » (Simonnot, 2012) qui s'inscrit dans les théories de l'action située et distribuée. Enfin, la numérisation croissante des documents modifie les modalités d'interaction avec le document-objet, générant le besoin de créer des espaces d'interaction complémentaires, comme des environnements de découverte et de manipulation des documents, et transfert ainsi le document de son statut d'objet vers un nouveau de rôle de médiateur à part entière : il s'agit alors d'une médiation numérique, c'est-à-dire le fait de « mettre en relation des ressources et des usagers » (Boustany et al., 2014). Dans ce cadre, un dispositif de médiation, comme par exemple un jeu multimédia, constitue un moyen d'appropriation du contenu, un facilitateur de création de sens, particulièrement utile dans le cadre de l'évolution des comportements des usagers, de plus en plus nomades, exigeants, pressés et habitués à la gratuité des supports numériques. Il en va de même pour les nouvelles solutions de visualisation des données. Bien que la Data Visualisation ne soit pas nouvelle et que ses vertus interactives et exploratoires soient connues, l'explosion des données remet le concept au goût du jour et l'ancre dans les dispositifs de médiation numérique comme intermédiaire entre les données et la connaissance (Arruabarrena, 2015). Elle est alors considérée comme un prolongement des cartographies multidimensionnelles, dont l'intérêt est tout à fait accentué dans le cadre du Big Data (Amato, 2015). Si les technologies à l'ère du Big Data sont bien identifiées comme des éléments clés pour la génération finale de connaissances

et d'usages, elles ne sont que rarement citées comme contributrices à l'aspect cognitif du processus de transformation des données : une fois de plus, seul le modèle de référence DMLC évoque la nécessité de mettre en place un entrepôt de données et de connaissances, sans toutefois préciser ni la nature des interfaces, ni leur spécificité et leur aspect innovant, ni les modalités de transition d'information entre acteurs utilisant ce socle technologique commun au cours du projet.

Afin de clarifier la nature de la médiation visée dans ces travaux, il faut distinguer la médiation comme service opérationnel et récurrent fourni aux demandeurs (par exemple, dans le cadre d'un centre de service de recherche documentaire, ou d'une utilisation régulière d'un modèle de prise de décision d'ores et déjà déployé), et la médiation entre les hommes et les données qui vise l'acquisition de connaissances pour guider les arbitrages réalisés au cours d'un projet data lui-même. La première est issue d'une coopération de deux démarches, inductive et déductive. Cette double démarche est mise en évidence à travers les modèles sociocognitifs (Ellis, 1989; Simonnot, 2012) qui pointent les réactions des utilisateurs de moteurs de recherche face à la découverte d'informations qui peuvent entrer en conflit entre elles ou avec leur capital de connaissances. Elle guide les procédés d'élimination, de compromis, d'acceptation et de confusion, identifiés comme des réactions typiques des utilisateurs. La seconde n'est pas décrite aujourd'hui dans la littérature, mais semble pouvoir s'appuyer sur des mécanismes similaires, notamment en termes d'arbitrages successifs sur l'utilité directe, l'utilité indirecte, voire l'inutilité des informations générées au cours d'un projet data. Cette similitude de mécanismes est d'autant plus intéressante à mettre en évidence que, sous l'impulsion du phénomène Big Data, les projets data se multiplient et peuvent être gérés sous la forme d'un portefeuille par une entreprise, ce qui peut faire basculer la médiation au cours d'un projet data unique vers une démarche opérationnelle plus récurrente et capitalisable.

La Médiation Homme-Données partage ainsi avec la médiation documentaire de nombreuses caractéristiques (voir Figure 18), modes opératoires et rôles, mais se distingue par le lieu de son exercice en entreprise et les métiers associés, ainsi que par un nouvel objet dans le dispositif : l'algorithme, que ce soit en tant qu'acteur non humain du dispositif de médiation (fonction de transformation d'une donnée brute en résultat intelligible) ou qu'élément construit au cours du projet. Cette Médiation Homme-Donnée fait l'objet dans ces travaux d'une attention particulière, notamment à travers son observation sur le terrain et sa mise à l'épreuve au cours des projets data réalisés.

	Médiation documentaire	Médiation Homme – Donnée
Lieu commun d'exercice	Entreprise : Fonction Business Intelligence ou Knowledge Management Hors entreprise : Centre Documentaire, Bibliothèque, Musée,… Tendance globale à la **décentralisation**	Entreprise : Dispositif Projet Data (DataLabs, DataFabs, DSI, Marketing, fonction métier bénéficiaire,…) Hors entreprise : écosystème Big Data (instable) Tendance globale à la **centralisation**
Acteurs du dispositif	Demandeurs métier Objets techniques, Interfaces	
	Professionnel de l'information (Knowledge Engineer, archiviste, documentaliste,…)	Algorithmes Data Scientists (métier aux frontières et parcours encore instables)
Rôle de médiateur	Accompagne l'utilisateur (méthode et technique) Garantit une communication optimale Evalue la valeur des documents ou des données face à un besoin Génère et gère les métadonnées	
	---	Construit de nouveaux modèles de données
Objet d'appropriation	Métadonnées	
	Documents	Données (valeurs et modèles, dont algorithmes)

Légende :

	Différences
	Similitudes

Figure 18 – Synthèse comparative entre la Médiation Documentaire et la Médiation Homme-Données

Si le médiateur est considéré habituellement comme un acteur (humain ou technique) tiers entre deux mondes en convergence, son rôle apparait plus ambigu dans la Médiation Homme-Donnée : en effet, le projet data mobilise des Data Scientists, qui semblent porter à la fois la responsabilité d'un professionnel de l'information, catalyseur de la dynamique de convergence, et celle d'un constructeur d'algorithmes à partir des données. Son cadre de compétences, encore flou, est précisé dans le chapitre qui suit.

3.4.1.3 Data Scientists : un nouvel éventail de compétences encore instable

Les compétences constituent un « ensemble des savoirs, savoir-faire et savoir-être mobilisés dans l'exercice d'un métier ou d'une activité professionnelle »[22]. Or, bien que le « Data Scientist » ait été identifié comme le métier le plus sexy du siècle (Davenport & Patil, 2012), les contours de ce métier ne sont pas stabilisés. Selon la nomenclature RH des métiers par

[22] Définition FAFIEC : https://www.fafiec.fr/80-l-observatoire-opiiec/etudes/metiers-du-numerique/315-formations-et-competences-big-data-et-cloud-computing-en-france.html

CIGREF[23], un Data Scientist réalise la mission suivante : « positionné auprès des Métiers, il exploite, analyse et évalue la richesse de données structurées ou non, appartenant à l'entreprise ou non, pour établir des scénarios permettant de comprendre et d'anticiper de futurs leviers métier ou opérationnels pour l'entreprise ». Dans ce cadre, il possède des compétences liées à la veille technologique, à l'innovation, à la gestion de l'information et de la connaissance, à l'identification des besoins, au marketing numérique, au développement prévisionnel et à l'amélioration des processus. Il s'agit d'un acteur qui s'inscrit côté métier dans l'organisation et la gestion des évolutions du système d'information, contrairement à un Data Analyst qui appartient aux fonctions support (méthode, qualité et sécurité). Ce dernier est alors positionné à la DSI, et doit « organiser, synthétiser et traduire efficacement » les données. Il ne conçoit pas de leviers métier, mais exécute leur développement (conception de l'architecture et de l'application, développement technique, tests, mise en production et documentation) et leur gestion (risques, processus, sécurité de l'information).

Cependant, cette définition des compétences du Data Scientist est loin d'être partagée. L'OPIIEC présente un référentiel des métiers sous forme de fiches[24] où « Data Scientist » est synonyme avec « Data Analyst » et « Data Miner ». Métier positionné dans l'amélioration continue, il n'est qu'une simple évolution des deux métiers précédents en intégrant les savoir-faire techniques liés à l'exploitation des données massives, structurées et non structurées. Cette vision se confirme généralement par les fiches de poste publiées par les entreprises, qui mettent en avant les savoir-faire techniques nécessaires au poste lors des recrutements des Data Scientists, comme la maîtrise des outils analytiques (R, SAS…), des langages de programmation (Python,..) ou des plateformes (Hadoop, Spark…). Cette vision est difficilement compatible avec une approche plus large, où un Data Scientist doit posséder toutes les qualités d'un hacker, analyste, communiquant et conseiller de confiance (Milleker, 2014), être capable de réaliser un projet data de bout en bout (y compris en termes de création d'applicatif), ou réaliser l'ensemble de la mise en perspective narrative du projet, c'est-à-dire le storytelling (Bladt & Filbin, 2013). Enfin, les compétences en Machine Learning font elles aussi l'objet d'une intégration ambiguë dans l'éventail des compétences du Data Scientist : soit il s'agit d'un domaine dont le Data Scientist doit avoir de simples notions pour mobiliser les

[23] CIGREF, « Nomenclature RH », *Les métiers des systèmes d'information dans les grandes entreprises*, Octobre 2015, https://www.cigref.fr/wp/wp-content/uploads/2015/12/CIGREF-Nomenclature-RH-Metiers-Competences-2015.pdf

[24] OPIIEC, Référentiel métiers de la branche numérique, de l'ingénierie, des études et du conseil et de l'événement, http://referentiels-metiers.opiiec.fr/fiche-metier/113-data-scientist

algorithmes les plus appropriés dans une situation donnée[25], soit le Data Scientist est justement un Machine Learner : cette compétence constitue alors son cœur de métier, complété avec des compétences périphériques. Dans ce cas, le Data Scientist tient plus du chercheur, doit avoir une maîtrise pointue de l'Intelligence Artificielle, voire être capable de concevoir des algorithmes théoriquement nouveaux.

Cette absence de consensus sur les compétences, y compris minimales, peut être expliquée par des besoins différents en fonction de la nature de l'entreprise qui emploie les Data Scientists[26]. En effet, pour une entreprise qui a l'habitude de mobiliser des Data Miners ou des Data Analysts pour des rôles support, le Data Scientist peut être une évolution marginale de ces métiers historiques, mis au goût du jour et enrichis avec des compétences techniques liées à l'évolution des outils basiques. Pour des entreprises qui se lancent dans la production ou la consommation des données volumineuses et hétérogènes sans bouleversement particulier des usages, un Data Scientist est essentiellement orienté sur les requêtes, le nettoyage et la structuration des données avec les solutions technologiques Big Data : il s'agit d'un profil d'Ingénieur Data, assez proche de celle d'un ingénieur de la connaissance maîtrisant, une fois de plus, de nouveaux outils. Pour les entreprises dont la donnée constitue le cœur de métier (comme les GAFA) et n'a pas besoin d'être nettoyée, l'éventail de compétences est beaucoup plus large et comprend une maîtrise solide du Machine Learning et des modèles complexes, et des capacités créatives d'un chercheur. Enfin, pour une entreprise qui souhaite valoriser les données et créer de nouveaux usages sur leur cœur d'activité historique, le Data Scientist est beaucoup plus orienté sur le métier et la communication, tout en possédant des compétences en Machine Learning et un état d'esprit de chercheur, sans toutefois exceller dans les modèles les plus complexes. Ce profil est le plus complet en termes de compétences et présente un intérêt particulier pour ces travaux de recherche orientés sur les entreprises traditionnelles. Il permet de concevoir le Data Scientist comme la jonction nécessaire entre le système de l'entreprise et l'écosystème Big Data. Dans ce cadre, une entreprise d'assurance, de parfum, ou encore un acteur de l'économie du document sont autant d'entités dites « métier » qui doivent intégrer ce nouvel acteur, le « Data Scientist », pour établir le lien avec l'écosystème.

[25] Blog « Le Big Data », *Voici les 13 compétences nécessaires pour devenir Data Scientist*, 20 juillet 2017, https://www.lebigdata.fr/13-competences-necessaires-devenir-data-scientist

[26] Distinction inspirée de la présentation des compétences de Data Scientists par Udacity, centre de formation continue, https://blog.udacity.com/2014/11/data-science-job-skills.html

Cette mise en perspective permet, en absence de consensus, de s'affranchir d'un établissement exhaustif des compétences d'un Data Scientist travaillant dans une entreprise qui souhaite créer de nouvelles connaissances et améliorer sa prise de décision à travers les projets Data Science, et de dresser simplement ses 3 domaines de compétences principaux (Conway, 2010; Milleker, 2014), symétriques avec la définition du phénomène Big Data (Boyd & Crawford, 2012). Il s'agit :

- de la connaissance mathématique et statistique (dimension analytique du phénomène)
- d'un goût prononcé pour l'informatique afin d'intégrer dans ce processus créatif les outils naissants (dimension technologique du phénomène)
- d'une expertise métier qui permet d'imaginer et de mettre en œuvre une stratégie de valorisation des données et une démarche d'extraction de connaissances utiles (dimension mythique du phénomène, nécessitant une traduction en usage concret dans le contexte de l'entreprise en question)

Le rôle du Data Scientist consiste alors à utiliser les données disponibles, des approches analytiques et des technologies nouvelles et son expérience de projets data pour construire des produits et services afin d'aider les organisations à formuler et atteindre des objectifs (Segaran & Hammerbacher, 2009). Ce rôle semble isoler ensemble des métiers historiques, exerçant des missions récurrentes de production ou de support dans l'entreprise, mais il n'est pas exclu que des métiers historiques puissent évoluer vers le rôle de Data Scientist grâce à une meilleure maîtrise des nouvelles technologies (par exemple, en se formant à un nouveau langage de programmation). Le Data Scientist apparaît dans tous les cas comme un acteur clé indispensable à un projet data, révélateur dans un temps limité de l'interaction entre l'entreprise et l'écosystème Big Data.

Or, cette définition des familles de compétences est large, et comprend la connaissance métier, ce qui fait du Data Scientist un profil extrêmement rare, d'autant plus que le système de formation français n'en est qu'à sa mise en route pour former ces profils. Le profil semble presque utopique, et le Data Scientist est en effet présenté comme un mouton à cinq pattes, comme en témoignent de nombreux articles grand public ou les synthèses des salons Big Data

de Paris en 2014 et 2016 par les journalistes du site Clubic[27]. Une réponse simple à cette problématique est proposée par le modèle de projet data DMLC (Data Mining Life Cycle, présentant une version théorique construite sur le modèle CRISP_DM et enrichie notamment sur les compétences humaines mobilisées tout le long du projet) qui décrit non pas un, mais des profils mobilisés dans la dynamique d'un projet data. Pour rappel, ce modèle introduit la mobilisation, au cours d'un projet data, des acteurs suivants : Project Manager, Data Analyst, Data Engineer, Data Miner, Knowledge Engineer, Domain Expert, Strategic Manager, et Business Analyst (le « Data Scientist » en soi est absent de ce modèle, proposé avant le buzz, mais le rôle du Data Miner comprend bien la génération d'un modèle algorithmique). Il élargit ainsi les métiers impliqués aux interlocuteurs ne présentant pas de compétences spécifiques en Data Science, et identifiés pourtant comme fondamentaux dans les projets data (Hofmann & Tierney, 2009). Cette vision est cohérente et basée sur les travaux de recherche antérieurs, et semble réaliste face à la diversité des métiers, y compris historiques, potentiellement impliqués dans la construction d'un usage en mode projet. A ces compétences clés semble s'ajouter progressivement un acteur appelé Data Steward, issu d'un enjeu fort à lier les projets data avec l'organisation de la gouvernance des données plus générale en entreprise.

Pour mieux comprendre le rôle d'un profil donné dans ce dispositif, on peut considérer par exemple l'un des profils contributeurs : le Knowledge Engineer. Dans le cadre de ce modèle, il ne s'agit pas d'envisager cet acteur comme un rôle récurrent, mais bien comme un contributeur temporaire à un projet data précis. La contribution de l'ingénieur de la connaissance est de « garantir que la connaissance est non seulement obtenue, mais également représentée et structurée pour une utilisation et une réutilisation optimales ». Il est notamment responsable du stockage, de l'extraction et de la maintenance des connaissances dans un entrepôt de connaissances, appelé IKR dans le modèle DMLC. Il sert le projet grâce à la mobilisation d'un capital de connaissances préalables permettant de mieux converger sur un usage, sans toutefois formuler directement l'objectif du projet. Il mobilise ainsi ses compétences historiques, largement documentées dans les métiers du Knowledge Management (Chastenet de Géry, 2018), dans les différentes phases du projet : dans un premier temps, il aide à orienter la convergence en s'appuyant sur la psychologie cognitive pour reproduire ce qui réussit et ne pas reproduire ce qui échoue, et, dans un second temps, il construit la capitalisation optimale de

[27] Voir articles Clubic : https://www.clubic.com/pro/emploi-informatique/actualite-798752-data-scientist-metier-voie-decomposition.html et https://www.clubic.com/pro/it-business/actualite-693592-data-scientist-mouton-5-pattes-coeur-donnees.html

l'ensemble des nouvelles connaissances métier générées au cours du projet. L'opportunité de mobilisation d'un ingénieur de connaissances dans un projet data a d'ores et déjà commencé à être identifiée pour le Knowledge Management, qui devrait y trouver une occasion pour s'orienter davantage sur les métiers, sortir d'une activité autocentrée, capitaliser et mieux collaborer pour transformer les processus métier. Cette réflexion peut être menée de façon indépendante pour chacun des 8 profils décrits dans le modèle DMLC, mobilisés ponctuellement dans le cadre d'un projet data : chaque acteur semble devoir mobiliser, le temps d'un projet, ses compétences habituelles pour contribuer à la convergence vers un usage cohérent.

Le modèle a toutefois l'inconvénient d'établir une correspondance stricte entre les 8 groupes de compétences et les individus, rendant la mobilisation humaine d'un projet data potentiellement lourde et peu subtile. Cette limite peut être levée grâce à une autre vision du « Data Scientist », complémentaire avec le cadre défini ci-dessus, qui consiste à le considérer non pas comme un individu doté de compétences, mais comme une équipe d'individus possédant toutes les compétences nécessaires à l'ensemble de la chaîne de valeur du projet (Bououchma, 2016). Cette vision présente l'avantage de regrouper les compétences dans des ensembles cohérents au service de l'exécution d'une activité, ce qui semble convenir plus particulièrement dans la modélisation de processus. Elle rompt la bijection entre les individus et les métiers, pour aborder le sujet sous l'angle de rôles nécessaires à la réalisation d'un projet, plus pragmatique et plus malléable face aux choix de méthode projet. En effet, la variété de projets data semble nécessiter des dispositifs ajustés en termes de mix de compétences, de niveaux de compétences et d'expérience sur chaque rôle, de fonction et d'hiérarchie au sein de l'organisation d'un projet data ou de l'entreprise, et enfin d'implication lors des différentes étapes de la chaîne de transformation des données. L'ensemble du projet devient ainsi proche du concept de « bundle », c'est-à-dire une recombinaison de ressources et de compétences, mobilisé de façon hétérogène dans le temps. On peut poursuivre l'exemple sur le Knowledge Engineer, mais dans le cas particulier où le projet vise un usage en Knowledge Management (par exemple, la mise en place d'un moteur de recherche d'informations internes dont la fonction KM est le bénéficiaire direct) : le contributeur issu de la fonction KM peut se retrouver à la fois comme contributeur en ingénierie de connaissances, comme un expert du domaine, voire comme un manager de projet, en tant que porteur du budget dédié. S'il possède des compétences analytiques et techniques nécessaire, il peut exercer d'autres rôles plus spécifiques au projet data. Cette vision permet d'écarter une approche trop individuelle, et d'isoler les

interactions entre acteurs humains liées à l'apprentissage des savoirs, savoir-faire et savoir-être au cours d'un projet.

En effet, l'apprentissage correspond, dans ce cas, à l'acquisition d'expérience par l'application des savoirs dans un contexte d'action. Il permet ainsi de mieux anticiper les risques, et d'imaginer des stratégies de réduction des incertitudes. Bien au-delà du simple apprentissage des savoir-faire « techniques », il s'inscrit dans la finalité générative du savoir, dans le sens capital de connaissances théoriques et pratiques, comprenant le vocabulaire, les normes, les règles de pratique, la description des ressources mobilisables, mais aussi la connaissance des modes opératoires et des processus en soi. Il fait l'objet d'un investissement initial sous forme d'acquisition de l'expertise ou d'allocation de temps d'apprentissage sur des dispositifs similaires. Chaque acteur humain mobilise ainsi, en intégrant un projet data, un capital de connaissance et une expérience propre à son domaine de pratiques, et « apprend » au cours du projet, à travers l'application de ses savoirs, pour enrichir ce capital initial. Mais, si l'évolution du capital de connaissances global est clé pour la compréhension de la dynamique d'un projet data, les processus et modalités d'apprentissage individuel sont écartées de ces travaux de recherche afin de s'affranchir des unicités des compétences individuelles, passées et apprises, pour se concentrer sur l'apport de leur mobilisation sur le résultat du projet.

Pour résumer, un projet data s'appuie sur un dispositif comportant nécessairement les trois domaines de compétences en Data Science, à savoir la maîtrise des mathématiques et statistiques, des technologies informatiques, et l'expertise métier (ces trois compétences constituent sa spécificité), mais aussi des compétences historiques mobilisées ponctuellement dans le cadre du projet afin de converger sur un usage tout le long du processus projet. De cette diversité des compétences, fluctuantes et hétérogènes, naît un risque lié à la difficulté d'une convergence entres les acteurs impliqués au cours des interactions, et une hypothèse forte d'existence d'une médiation entre l'homme et la donnée permettant de sécuriser cette convergence grâce à une co-construction de sens par les individus impliqués.

3.4.2 Interactions au sein d'un projet data

Si la notion d'interaction est bel et bien identifiée comme clé dans les projets data, essentiellement en termes de compétences de communiquant et d'orchestrateur attribuées aux Data Scientists, l'immatérialité de cet échange au sein du processus pourrait expliquer son absence des travaux de recherche à ce jour, et sa confusion avec la notion, peu précisée, de

« coopération » suggérée dans le modèle DMLC. Par ailleurs, le caractère exploratoire de ces projets ne permet pas à ce jour de s'appuyer sur les standards matériels de transmission de savoirs au sein d'un projet, comme des expressions des besoins ou des spécifications techniques. Or, le savoir constitue *a priori* une ressource, un input nécessaire à la réalisation d'une activité, mais aussi un output, sous sa finalité fonctionnelle (usage) ou générative (capitalisation de connaissances). Ces considérations poussent à préciser la nature des savoirs, connaissances et informations échangées entre les acteurs au cours des différentes étapes du projet, et les facteurs favorisant l'efficience de cette transmission pour converger sur un résultat optimal.

La construction d'informations utiles est dotée d'un aspect à la fois informatique et cognitif sur toute la chaîne de valeur de la donnée, ce qui a d'ores et déjà été mis en évidence. Dans ce cadre, les échanges d'informations permettraient aux acteurs de s'aligner et seraient indispensables à la production du résultat. Ils se focaliseraient sur le mouvement de va-et-vient entre les données (observations, métadonnées, paramètres, indicateurs...) et le sens qu'elles portent, et tendraient à faire converger les acteurs vers la génération d'une connaissance utile à la progression du projet. Elles viseraient un compromis négocié et s'inscrivent dans un mouvement de volonté de quantifier des indicateurs de performance de l'entreprise et de découvrir des connaissances utiles pour l'optimisation de cette performance. « Ce verbe *quantifier,* dans sa forme active (*faire* du nombre), implique qu'il existe une série de conventions préalables, de négociations, de compromis, de traductions, d'inscriptions, de codages et de calculs conduisant à la mise en nombre. La quantification se décompose en deux moments : *convenir* et *mesurer*. » (Desrosières & Kott, 2005). La mesure est généralement précédée par cette phase interactive de convention, ce qui priorise et structure la représentation de la réalité afin de la rendre convenable pour la compréhension ou pour l'action. Or, cette définition contredit la mythologie Big Data, selon laquelle les données sont objectives et permettent une mesure plus précise et opérationnelle de la réalité : elles génèreraient à elles-mêmes un nouveau cadre de mesure, une « auto-convention ». Il est nécessaire de mettre en lumière ce paradoxe, en partant d'une hypothèse forte d'existence d'activité de convention au sein des projets data, contrairement au mythe.

Cet écart entre le mythe et la réalité est en effet progressivement abordé, notamment sous l'impulsion de managers en entreprise qui cherchent à naviguer entre les opportunités réelles et les promesses du Big Data, en dehors des approches algorithmiques évoluées, par exemple dans

le cadre du traitement des flux d'information en temps réel (Pigni et al., 2016). La convention, au cours d'interactions d'alignement, au sens cognitif, est alors décrite comme une action de contextualisation qui aboutit à la construction d'informations, par opposition aux données « brutes ». La contextualisation comprend un processus de construction de sens à partir de données, validées, confrontées et caractérisées par des réponses aux questions clés (Quand ? Où ? Qui ? Quoi ? Comment ? Pourquoi ?). Ce processus peut être modélisé comme une convergence entre l'action de conversion des données en information, et l'action de requête, c'est-à-dire la mise en œuvre de la capacité à formuler une question, à problématiser, cette capacité étant dépendante de l'ensemble des savoirs existants. Une fois que le processus de contextualisation est abouti, l'information qui en résulte fait l'objet d'une interprétation en vue de prise de décision, c'est-à-dire une transformation en information utile pour l'action : il s'agit alors de connaissance nouvelle.

Cette recherche de compromis intermédiaires pointe sur un besoin d'informations circulantes et de traçabilité des décisions prises au cours du projet afin de converger vers le résultat final, mais aussi de rejeter certaines connaissances, considérées comme « non utiles » dans le cadre de la production du résultat, et pourtant « utiles » à l'échelle de l'organisation. Ce mouvement distingue alors les connaissances mobilisables au service de la conception de l'usage direct, et celles qui permettront de générer de nouvelles pistes de contextualisation.

Ce mouvement s'inscrit plus globalement dans la notion de capital de connaissances (Ermine, 2003), ou patrimoine de connaissances, désignant un stock de connaissances considéré dans le Knowledge Management comme la notion de savoir. L'étude des conséquences de la montée des investissements en capital informationnel et du développement des Systèmes d'Information a démontré qu'ils ont non seulement contribué à l'amélioration du processus technique de production, mais aussi, plus largement, révélé un système de connaissances, à mémoire, qui rend les hommes capables de réutiliser l'information et de développer des savoir-faire de façon non délibérée. Le capital de connaissances, par définition propre à une entreprise, voire à des individus dits « sachants », est immatériel et considéré comme valorisable (Edvinsson & Malone, 1999; Martory & Pierrat, 1996) malgré l'absence de consensus sur les méthodes de valorisation.

La subjectivité d'information, découlant de cette mise en relation, n'est pas partagée en sciences humaines et sociales. En sciences de l'éducation (Monteil & Truchot, 1986), l'information est extérieure au sujet : on peut la stocker, elle est transmissive et placée « sous le primat de

l'objectivité ». Elle s'oppose alors à la connaissance, résultat d'une expérience personnelle, et donc subjective. Lorsque la connaissance est dépouillée de sa subjectivité afin de devenir un « produit communicable », il s'agit de savoir, c'est-à-dire de l'information appropriée par un sujet : le savoir se situe alors à l'intersection des concepts d'information et de la connaissance. Cette mise en perspective est reprise pour opposer le savoir comme simple accumulation de contenus intellectuels à la notion de « rapport au savoir » (Charlot, 2005), où « le savoir est construit dans une histoire collective qui est celle de l'esprit humain et des activités de l'homme, et il est soumis à des processus collectifs de validation, de capitalisation, de transmission ». Il fait alors référence à l'apprentissage, c'est-à-dire la capacité à établir un rapport au monde, à soi et à l'autre, et au capital social de savoirs.

Ce rapport au savoir s'assujettit à un ensemble des représentations sociales du savoir, c'est-à-dire des artefacts qui incluent « croyances, valeurs, attitudes, opinions, images » (Jodelet, 2003) qui facilitent la diffusion des informations. « Le concept de représentation sociale désigne une forme de connaissance spécifique, le savoir de sens commun, dont les contenus manifestent l'opération de processus génératifs et fonctionnels socialement marqués. Plus largement, il désigne une forme de pensée sociale. Les représentations sociales sont des modalités de pensée pratique orientées sur la communication, la compréhension et la maîtrise de l'environnement social, matériel et idéal. En tant que telles, elles présentent des caractères spécifiques sur le plan de l'organisation des contenus, des opérations mentales et de la logique. Le marquage social des contenus ou des processus de représentation est à référer aux conditions et aux contextes dans lesquels émergent les représentations, aux communications par lesquelles elles circulent, aux fonctions qu'elles servent dans l'interaction avec le monde et les autres ».

La définition du concept de représentations sociales partage avec la gestion des connaissances la vision de deux finalités des savoirs : il s'agit de la finalité générative et de la finalité fonctionnelle. La première fait référence à la capitalisation progressive des connaissances en vue de l'augmentation du savoir (ou du patrimoine informationnel) : le savoir est alors un accélérateur de son propre processus de génération. La seconde fait référence à la mise en application des connaissances à travers la prise de décision au cours d'usages : il s'agit de l'application opérationnelle des savoirs dans le cadre d'une stratégie de l'entreprise.

Il est alors nécessaire de distinguer les représentations sociales préalables à la conception d'un nouvel usage (patrimoine informationnel), qui garantissent l'interprétation et l'appropriation des connaissances générées, et les représentations sociales générées dans le contexte de

conception de cet usage. Or, lorsque cet usage est créé dans le cadre d'un projet data, de nouvelles compétences analytiques et technologiques seraient intégrées dans un dispositif et confrontées avec des représentations sociales historiques. Au-delà des acteurs humains (Data Scientists), aux rôles à clarifier, impliqués sur ces projets, les ressources mobilisées contiendraient un ensemble d'objets non humains, dotés de leurs propres marqueurs sociaux et dont le rôle ne semble pas négligeable dans le dispositif étudié.

Tout d'abord, il s'agit des données. Les projets data s'appuient en effet sur l'existence de données préalables, générées dans le cadre d'un usage indépendant. Ce sont des données « brutes », dans le sens où elles n'ont pas fait au préalable l'objet de contextualisation dans le cadre de l'usage nouveau visé par le projet data en question. Le savoir mobilisé dans le cadre de la génération de ces données a pu donc avoir une finalité fonctionnelle différente. Ensuite, les nouvelles technologies de traitement des données donnent lieu à la mobilisation potentielle de systèmes de stockage, comme les Data Lakes, de systèmes de traitement analytique, dotés d'une interface dédiée à l'analyse, et de fonctionnalités de restitution des résultats Data Visualisation. Ces éléments technologiques constituent des voies d'accès aux données sous différents angles, non exempts de représentations sociales du savoir qui lui seraient propres, distinctes du paradigme disciplinaire et professionnel de l'usage visé. Enfin, l'introduction des algorithmes, avec toute leur technicité et imaginaire, perturberaient d'autant plus les représentations sociales habituelles.

Dans l'hypothèse où ces acteurs, humains et non humains, induisent des interactions nouvelles avec les acteurs humains, il faut les identifier et les prendre en compte dans le cadre de l'enrichissement des représentations des savoirs.

En dehors de la Data Science, la mise en évidence de la dimension sociale de l'information n'est pas nouvelle (Couzinet et al., 2001; Gardiès & Fabre, 2015) : « les premiers chercheurs qui se sont employés à développer une approche scientifique de l'information, comme Robert Escarpit, Jean Meyriat ou Robert Estivals, ont insisté sur la dialectique complexe entre la dimension intellectuelle de l'information comme relation et sa dimension matérielle en tant qu'inscription ». Or, la matérialité de l'inscription, hétérogène et bousculée par le numérique, est elle-même issue de l'action d'inscription, réalisée dans un cadre contextuel marqué par des représentations sociales préalables et dotée d'une finalité. Dans ce cadre, l'information s'oppose à la donnée « brute », en tant que trace d'une réalité externe au sujet, un contenu objectif à s'approprier dans un cadre social défini, et fait référence à la donnée « construite ».

L'état de l'art sur la qualité des données montre bien cette distinction : d'une part, la matière première utilisée est constituée de données préexistantes, internes ou externes, utilisées ou non dans le cadre de l'activité passée des experts métier, et d'autre part la chaîne de transformation des données génère en soi un ensemble d'informations nouvelles, qu'elles soient intermédiaires (métadonnées, données enrichies, structures de données, paramètres des algorithmes…) ou finales (résultats des algorithmes, indicateurs d'évaluation des résultats, évaluation du potentiel et valorisation des usages des résultats…). La vision issue du mythe Big Data, simplifiant la notion de donnée comme input initial à son essence objective, « brute », ne permet pas de prendre en considération les mécanismes de choix et de construction des informations intermédiaires utiles à la génération des connaissances.

Ces contextualisations intermédiaires constituent *a priori* une activité clé dans les projets data, et les placent dans un processus inédit de sémantisation des données, c'est-à-dire le processus de création de sens dans le cadre d'un paradigme métier existant. Il semble difficile d'isoler ces problématiques des pratiques, cadres théoriques et représentations des savoirs dans lesquels l'usage de ces connaissances aura lieu, et indispensable de comprendre la dynamique du projet data en termes de mobilisation de nouveaux rapports aux savoirs, et de modalités d'interaction entre ces savoirs avec les savoirs existants. La médiation Homme-Données doit alors jouer non seulement un rôle de facilitateur, fluidifiant la conception des usages, mais aussi celui de mise à disposition du savoir aux preneurs de décision finaux, afin qu'ils puissent comprendre les raisons de leur prise de décision. Elle doit ainsi comprendre la documentation de la dynamique des interactions qui ont conduit à l'élaboration de ce savoir. Un établissement des vecteurs de médiation, précisant leur nature (rôles des acteurs humains, place des algorithmes et des interfaces, capital de connaissances et représentations sociales associées…) et leur place dans le processus de projet data constitue alors l'un des objectifs clés de ces travaux de recherche.

***

A ce stade, l'état de l'art pointe ainsi un processus de référence des projets data, CRISP_DM, doté d'un ensemble de limites traduites par 3 dimensions clés : les indicateurs de valeur, la qualité des données et la Médiation Homme-Données. Ces dimensions multidisciplinaires semblent intimement liées au potentiel de génération de valeur des projets data, et nécessitent une confrontation au terrain. La suite de ces travaux de recherche présente les terrains et les méthodes qui permettent cette confrontation (Deuxième Partie) à travers une recherche-action débouchant sur une étude de cas multiples. Cette méthode constitue la base de la construction

des résultats (Troisième Partie), où chaque étude de cas correspond à un projet data réel et permet non seulement d'établir une critique du modèle processus de référence par le terrain, mais aussi de proposer un nouveau modèle de dispositif projet data. Ce nouveau modèle s'inscrit alors dans une visée opérationnelle tout en étant théoriquement robuste dans la mesure où il articule et précise ces 3 dimensions clés identifiées dans l'état de l'art comme insuffisantes dans le modèle de référence CRISP_DM.

« Comme jouer du violon ou du piano, penser exige une pratique quotidienne. »

***Charlie Chaplin**, Acteur, Artiste, Cinéaste, Scénariste*

Deuxième partie :
Terrains et Méthodes

Une fois posée la synthèse introductive, mettant en perspective le phénomène Big Data et les concepts clés autour des questions de cette thèse, la problématique appelle des réponses concrètes. Or, il n'est pas question ici de rentrer dans la technicité mathématique des modèles algorithmiques : il s'agit de comprendre comment se déroule, du point de vue anthropologique, et prend sens et racine un projet data en entreprise.

Pour ce faire, le travail de recherche a été réalisé en deux temps. Une première phase de pré-expérimentation, pour découvrir sous forme d'entretiens trois cas de projets, a permis de confronter les premiers éléments terrain au cadre théorique, d'approfondir la problématisation et le cadre conceptuel du sujet, de choisir les modèles de processus projet de référence, et de dégager un protocole de recherche. Mais entrouvrir la « boite noire » du projet data n'était pas suffisant : il fallait plonger au cœur de ces projets en intégrant les équipes de Quinten, société française experte en Data Science, et voir que la conception d'algorithmes était en effet issue d'un travail de co-construction, laborieux, parfois imprévisible, et marqué par une certaine spécificité d'interactions.

Ces projets data, en tant que terrain de recherche traité en première partie de ce chapitre, sont apparus en effet comme des dispositifs intéressants pour analyser l'articulation entre le mondes de l'entreprise et une société de l'écosystème, ainsi que l'impact de cette interaction sur l'entreprise, bien que le suivi de cet impact reste limité par l'absence de visibilité sur l'exploitation des usages *a posteriori*. Ils sont bornés dans le temps et directement observables, contrairement à l'étude des discours, des résultats financiers, sondages, artefacts et autres traces. Contrairement aux projets plus technologiques, comme le déploiement de Data Lakes visant essentiellement des réductions de coûts IT, ces projets orientés sur la Data Science sont marqués par une visée de création de connaissances métier et de leviers opérationnels grâce aux algorithmes, et permettent de mettre le doigt de façon plus pertinente au cœur du débat sur la construction de sens. Le choix de la recherche action, et plus globalement l'approche méthodologique décrite en seconde partie de ce chapitre, permettent de tester immédiatement les propositions d'optimisation qui émergent au cours de l'expérience. Basée sur la conduite d'étude de cas multiples (7 cas terrain dans des entreprises d'assurance et de parfum), la démarche aboutit à une proposition de modèle : non pas un modèle algorithmique, mais bien un modèle de dispositif projet data, incluant l'optimisation du processus de référence et les dimensions clés permettant de répondre aux questions principales sur la génération de valeur, la qualité des données et la Médiation Homme-Data.

1 Choix du terrain

Le choix du terrain pour ce travail de recherche est issu d'une contrainte double : l'accès limité aux projets Data Science concrets en France, et l'exigence d'un niveau d'approfondissement suffisant permis sur le terrain.

En effet, l'accès aux projets réalisés directement au sein des entreprises est avant tout limité par la volonté de préservation de la confidentialité. Elle vise à protéger l'avantage concurrentiel créé grâce à la stratégie de sélection des projets data et aux résultats des projets data réalisés, et à la capitalisation de connaissances à l'issue de leur mise en œuvre. Par ailleurs, le nombre d'échecs de ces projets est un facteur limitant le désir de communication des entreprises : selon Gartner[28] à fin 2011, plus de 85% des 500 plus grandes entreprises ne réussiront pas à mener avec succès l'exploitation des données avant 2015. Enfin, la récence du phénomène Big Data dans les entreprises en France, associée à l'opacité de ces projets, ne permet pas une identification simple dans les organisations en place des interlocuteurs directement concernés par ces projets. Ces particularités ont limité certains modes de recueil de données classiques en sciences sociales (De Bruyne et al., 1974) comme l'enquête auprès des acteurs concernés ou l'analyse documentaire, basée sur les informations réglementaires (rapports annuels, biannuels, trimestriels, organisation et objectifs du board…) et les communiqués de presse issus des sites institutionnels. La recherche de cas terrain a alors lieu à travers les réseaux professionnels auprès des entreprises, comme Inter Mutuelles Assistances qui s'est d'ores et déjà engagé en 2014 dans des projets data autour de la télématique.

Un autre choix du terrain possible était offert par les prestataires multisectoriels en Data Science, en lien direct avec les entreprises. Or, les cabinets de conseil généralistes ou spécialisés présents en France sont peu expérimentés dans ce domaine en 2013, malgré une communication institutionnelle riche sur le sujet. Seuls quelques acteurs ont réalisé des projets concrets. Parmi les grandes entreprises de conseil, BearingPoint a acquis en 2012 l'algorithme d'intelligence artificielle HyperCube, et les équipes de Data Scientists maîtrisant l'algorithme. Les cabinets

[28] "Gartner Reveals Top Predictions for IT Organizations and Users for 2012 and Beyond", 01/12/2011, http://www.gartner.com/newsroom/id/1862714 , repris sur le site de BusinessWire https://www.businesswire.com/news/home/20111201005541/en/Gartner-Reveals-Top-Predictions-Organizations-Users-2012

technologiques, comme OCTO, commençaient à développer une activité Data Science. Des cabinets de plus petite taille, comme Ekimetrics (2006), Quinten (2008), BrainCube (2008), ou encore Quantmetry (2011), étaient discrets, mais de plus en plus actifs sur le marché français.

Les premières enquêtes de 2014, sous forme d'interview des Data Scientists de Quinten et d'acteurs métier chez Inter Mutuelles Assistances, ont rapidement permis de faire émerger les premières hypothèses de travail (Chartron & Broudoux, 2015) et les concepts clés comme la médiation, mais ont présenté aussi des limites : le manque de visibilité sur le processus Data Science dans son ensemble, et en particulier sur les interactions entre les acteurs, ne permettait pas de répondre aux objectifs des travaux. Le choix de mode de recueil s'est alors dirigé vers l'observation participante, ce qui a déclenché le changement de poste professionnel en mars 2015, suite à l'opportunité d'intégrer les équipes de Quinten.

Quinten est une société spécialisée dans la valorisation des données pour de grandes entreprises. Créée en 2008, elle fait partie des plus anciens cabinets en France spécialisés en Data Science, qui constitue son cœur de métier : cette société a pu capitaliser sur plus d'une centaine de projets en 2015 (plus de 300 en 2017) pour mettre en œuvre une méthodologie Data Science éprouvée. Les clients de l'entreprise sont des Directions générales de grandes entreprises, des Directions de production, des Directions Marketing, des Responsables Innovation ou autres acteurs souhaitant définir et mettre en œuvre des améliorations sur des sujets à fort potentiel de création de valeur. Quinten emploie des Data Scientists, ayant des savoir-faire scientifiques et techniques pointus (mathématiques, statistiques et informatiques) et une expertise métier avec des spécialisations sectorielles. L'entreprise mobilise au service des projets data sa propre plateforme comprenant des technologies Small Data et Big Data, et a la capacité de concevoir des solutions algorithmiques articulant des modèles Data Science disponibles sur le marché, ou bien des algorithmes uniques et propriétaires développés en interne. Historiquement, Quinten intervient dans le secteur de la santé avec une offre techno-centrée (proposition de création de valeur grâce à sa technologie propriétaire), et compte parmi ses clients la majorité des groupes pharmaceutiques français et internationaux. En 2013, l'entreprise décide de se diversifier sur d'autres secteurs et d'enrichir son offre pour mieux répondre aux besoins de création de valeur propres à chaque entreprise. Son offre est alors large et orientée sur les capacités analytiques offertes par la Data Science au service des questionnements métier et de la prise de décision, proche du champ des possibles présenté par Gartner (voir Figure 19 et Figure 13).

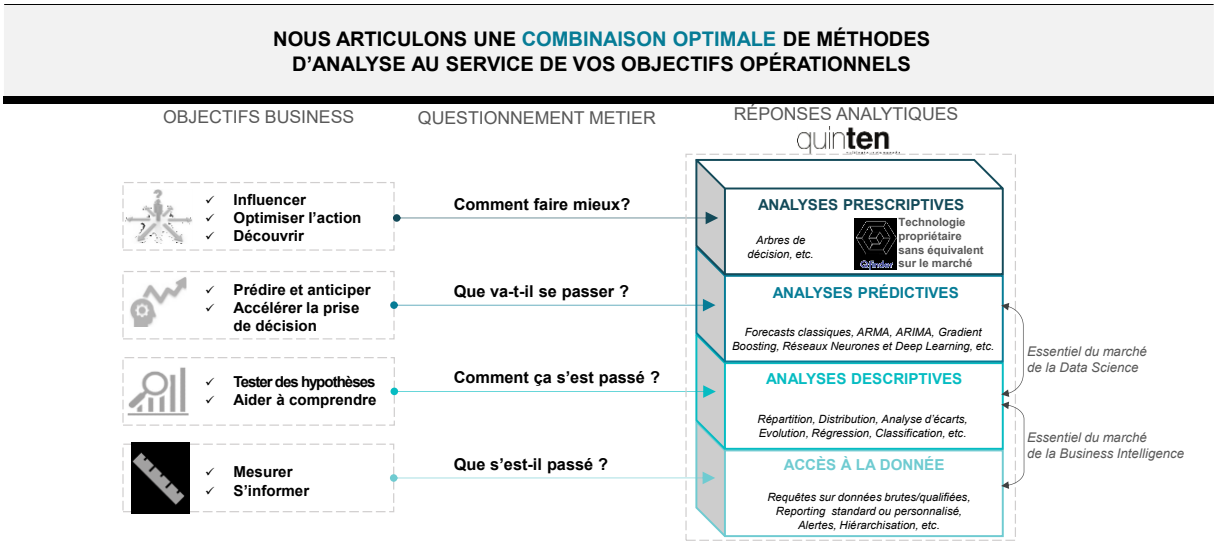


Figure 19 – Extrait de l'offre Quinten en 2015 en termes de capacités d'analyse

Les travaux se sont déroulés au sein même de l'entreprise, sous statut salarié, avec un objectif professionnel triple : apprendre et mettre en œuvre les connaissances en Data Science dans le cadre de la réalisation de projets concrets, contribuer à optimiser la création de valeur pour livrer de meilleurs résultats pour les clients, en s'appuyant notamment sur les acquis professionnels en conseil et en finance, et participer à la croissance de Quinten, en particulier en transposant l'offre technologique et méthodologique développée pour l'univers médical en offre dédiée au secteur de l'assurance. Le choix de ce terrain, délibéré et opportuniste à la fois, a été déterminant pour l'établissement de la stratégie d'observation et de la méthodologie de ces travaux de recherche, réalisés sous une double casquette « Chercheuse » et « Data Scientist ».

Ce choix est délibéré dans la mesure où il permet d'approcher concrètement l'univers d'analyse constitué des objets auxquels ces travaux de recherche se réfèrent : il s'agit des projets data réalisés dans le cadre de la recherche de la performance par les organisations dont l'activité peut être exercée indépendamment du phénomène Big Data et de la Data Science. Cet univers est difficilement dénombrable, car soumis à des problématiques de confidentialité, mais aussi de définition, comme vu dans la mise en perspective du Big Data. Un projet data est alors une situation de production de résultat ayant une finalité de création de valeur au sein d'une organisation, impliquant des concepts issus du phénomène étudié, c'est-à-dire les méthodes analytiques et les compétences en Data Science, les technologies Big Data, ou la mythologie associée au phénomène Big Data. Cet univers d'analyse d'intérêt est délibérément large dans la mesure où la définition de ce phénomène récent reste à préciser, ce qui n'est pas visé dans le cadre de ces travaux. Il ne constitue pas non plus un domaine de généralisation des résultats

comme étape préalable. En effet, cet univers d'analyse fait l'objet dans cette thèse d'une phase de pré-expérimentation, détaillée plus loin, avant d'être restreint à une population d'objets plus accessible : il s'agit de projets data réalisés par les entreprises avec intervention de Data Scientists internes ou externes à l'entreprise. Quinten, en tant que l'un des acteurs du marché français intervenant sur ces projets, donne accès à une sous-population dénombrable et précise, ce qui en fait un terrain de recherche riche en projets passés et en cours : ces derniers présentent un intérêt particulier pour cette thèse pour deux raisons. D'une part, ils constituent le reflet d'un apprentissage dans le temps, et permettent d'observer l'état de l'art des pratiques en Data Science en France. D'autre part, ils donnent la possibilité d'observer le processus étudié en situation, et ce notamment lorsque Quinten ouvre la porte à l'action directe : cette opportunité donne lieu à l'adoption des méthodologies de recherche-action, et conditionne du point de vue matériel et logistique l'échantillonnage et le traitement des observations.

2 Approche méthodologique

« À la racine du mot opportunisme, se trouve le mot portus*, le port.*

Ce mot désigne donc une manière d'arriver au port, pas toujours par le chemin que l'on prévoyait de suivre, pas toujours dans le temps prévu, et même, quelquefois, pas dans le port où l'on pensait se rendre. »

***Jacques Girin**, chercheur en gestion*

2.1 Recherche-action

Ces travaux constituent une recherche-action, au sens anglo-saxon du terme (« action research »), c'est-à-dire un processus réfléchi de résolution progressive de problèmes mené par une communauté de pratique (Wenger, 1998) composée de Data Scientists. Ils s'appuient sur une méthode dans laquelle il y a « une action délibérée de transformation de la réalité ; recherche[s] ayant un double objectif : transformer la réalité et produire des connaissances concernant ces transformations » (Hugon & Seibel, 1988). Cette recherche-action, contextuelle et par nature subjective, est basée sur des études qualitatives de cas, soit des données primaires de nature exploratoire.

Elle remplit plusieurs fonctions (Michelle et al., 2014) : d'une part, elle permet de rendre compte d'une investigation terrain à travers la description et l'explication des études de cas, d'autre part elle propose une critique des référentiels scientifiques existants sur les processus de Data Science, et enfin elle offre la possibilité d'établir une jonction entre la théorie et la pratique à travers une proposition de modélisation du processus Data Science en entreprise, cette dernière finalité ayant une visée à la fois opérationnelle et pédagogique. En effet, l'objectif de ce travail de recherche consiste à construire une représentation instrumentale du processus étudié. Les connaissances concernant la transformation de la réalité constituent des lignes directrices et les bonnes pratiques (Denscombe, 2014) pour la réalisation des projets Data Science. Ce résultat se traduit à l'issue de la thèse sous forme de modélisation du processus de projet Data Science enrichi sur 3 dimensions : la médiation entre les acteurs impliqués dans un projet Data Science, l'impact sur la qualité des données, et enfin l'impact sur les indicateurs de valeur. Ces dimensions font alors l'objet d'une confrontation permanente entre les canons

académiques et les résultats pratiques de leur application afin de dégager une zone de convergence à la fois opérationnelle et scientifiquement légitime (Cappelletti et al., 2018). Cette représentation se veut utile pour l'action, dans le sens où elle est convenable pour appréhender les interactions entre une entreprise et le phénomène Big Data dans le cadre de projets data visant la création de valeur, et de l'évolution des connaissances des acteurs engagés dans ces projets. Cette modélisation est le fruit d'un travail d'abstraction et de conceptualisation réalisées à l'issue d'une confrontation entre les modèles Data Science existants et la réalité du terrain observé, représenté par un échantillon d'études de cas décrivant les projets réalisés avec les équipes Quinten. Ils permettent de comprendre la relation complexe entre les facteurs tels qu'ils opèrent dans le contexte social particulier de Data Scientists confrontés aux problématiques des entreprises.

Cette dernière finalité inscrit ces travaux de recherche dans un paradigme constructiviste (Bertacchini, 2009; Moigne, 2012), à partir du moment où la « réalité » étudiée « est une construction intellectuelle qui dépend des prérequis conceptuels et théoriques pris comme référentiels » (Mucchielli, 2009). En effet, cette recherche vise une construction de connaissance plausible et inachevée, une proposition de consensus plus sophistiqué et plus informé que la construction de prédécesseurs, cette proposition étant créée et expérimentée au cours de la recherche grâce à l'interaction entre le chercheur et l'objet de recherche. La connaissance visée, représentée sous forme de modèle de processus Data Science, est un résultat indissociable du processus même de création de cette connaissance. Enfin, le caractère téléologique de ces travaux s'inscrit dans la recherche de pragmatisme lié à l'optimisation de la création de valeur visée par les projets Data Science, c'est-à-dire l'intention de guider les acteurs dans les pratiques conduisant à l'amélioration des résultats des projets, notamment en termes d'adéquation entre ces derniers avec les besoins liés au pilotage et à la stratégie d'entreprise. S'inscrivant dans un mouvement historique de fond qui oriente les sciences humaines et sociales vers les logiques d'action, le paradigme constructiviste appuie cette recherche-action qualitative et permet de mobiliser les critères de validité (Girod-Seville & Perret, 2002) comme l'adéquation, c'est-à-dire une forte orientation pragmatique, et l'« enseignabilité » du résultat, c'est-à-dire une transmissibilité de connaissance construite de manière projective (Charreire & Huault, 2001).

2.2 Posture du chercheur

Face à la diversité des définitions de la recherche-action (Michelle et al., 2014), il semble important de préciser le rôle du chercheur dans le cadre de cette thèse. Selon les typologies établies par Henri Desroche (Desroche, 1981), cette participation peut être définie comme intégrale, dans la mesure où elle comprend la recherche d'explication (recherche sur l'action), la recherche d'application (recherche pour l'action) et la recherche d'implication (recherche par l'action, plus précisément par l'action du chercheur dans l'action des acteurs). Cependant, ces typologies de participation n'interviennent pas au même niveau dans le temps, et ne produisent pas les mêmes résultats.

Les premiers mois de mon intervention sont marqués par la découverte de l'environnement de travail d'un Data Scientist : il s'agit d'une phase de formation aux méthodes et aux outils des acteurs impliqués, ainsi qu'aux interactions entre les catégories d'acteurs. L'impact des projets Data Science sur les entreprises clientes (création de valeur, impact sur les indicateurs, création de connaissances...) est alors observé à travers l'action des autres acteurs Data Scientists. Cependant, dès cette phase, j'ai été impliquée directement sur les projets en cours en tant qu'élément productif (structuration et documentation des données, évaluation des résultats algorithmiques, réalisation des supports de restitution des résultats, accompagnement en gestion de projet…). En effet, la stratégie analytique à mettre en œuvre était établie et pilotée par des Data Scientists expérimentés, et j'étais sous leur supervision. Par ailleurs, j'ai été force de proposition sur les méthodes et outils issus du conseil stratégique et opérationnel, sur l'expertise métier dans le secteur assurance et la fonction finance, sur les outils de documentation de processus et de gestion de projet, et autres méthodes issues de la recherche ou du conseil. La décision d'appliquer ou non mes propositions était prise par les responsables de projet. Cette phase était ainsi marquée par une recherche sur l'action et pour l'action, et par l'action mais sans pouvoir de décision.

Ensuite, la participation est progressivement devenue intégrale, avec la montée en compétences et la prise de responsabilités au sein de la société. Concrètement, cette prise de responsabilités professionnelles a consisté à évoluer d'un périmètre opérationnel de Data Scientist / Consultant à un rôle de directeur de projet Data Science. Ce rôle comportait la responsabilité du projet au sens large et l'engagement face au client, y compris du point de vue commercial. La génération de valeur par le projet ne pouvait donc plus être théorique, et la recherche de facteurs de succès au cas par cas s'avérait indispensable et légitime, ce qui a permis de tester rapidement des

hypothèses. Au-delà de la capacité de décider, ce rôle comprenait des actions productives, en complément de celles réalisées au cours de la première phase de découverte. Il s'agissait par exemple de la conception de la solution Data Science amont (identification du besoin et établissement de la stratégie d'analyse Data Science, en mobilisant un Lead Data Scientist dans le dispositif projet en cas de complexité analytique particulière), de la réalisation d'une partie du traitement des données et des analyses Data Science, ou encore de la conception des interfaces de restitution dynamiques (sauf codage en soi des applicatifs web, réalisés par les Web Développeurs de Quinten) sur les projets concernés par ce livrable. Les actions de production variaient beaucoup entre les projets, qui mobilisaient entre 1 et 8 personnes (dont moi-même). Ainsi, les études de cas présentées dans le cadre de cette thèse sont marquées par une participation variable du chercheur, précisée dans chaque situation, mais une recherche définitivement intégrale.

Plus généralement, l'un des aspects clés du rôle de Directeur de Projet sur les projets réalisés a consisté à être présent à chaque étape lors des interactions avec les interlocuteurs en entreprise, y compris en avant-vente. Cette présence a autorisé trois éléments structurants pour la thèse :

- l'identification des modalités d'interaction entre les différents intervenants du projet, ce qui a alimenté l'analyse des médiations

- l'établissement des zones d'incompréhension nécessitant un appui pédagogique

- l'évaluation de la pertinence des propositions réalisées au fur et à mesure, que ce soit en termes d'outils, de méthodes, ou d'éléments pédagogiques

Ce « feedback » récurrent permettait d'évaluer la satisfaction des demandeurs, sous forme de validations de propositions commerciales et de récolte d'évaluations qualitatives pendant et à l'issue des projets. Le dernier aspect du rôle a été l'implication dans le développement interne de la société, et notamment sur les méthodes. Or, l'intégration d'une nouvelle méthode de travail en interne correspondait à une confirmation de sa validité par la communauté de Data Scientists. Ces différents retours, de la part des Data Scientists et des acteurs en entreprise, ont guidé la construction des propositions de cette thèse.

2.3 Etude de cas multiples

Longtemps considérée comme scientifiquement contestable par les défenseurs du paradigme quantitatif, positiviste, la méthode de recherche par étude de cas s'inscrit dans le développement des approches interprétatives, dites réalistes. Bien que peu rependue en Sciences de l'Informations et de la Communication, mis à part dans quelques travaux de thèse canadiens, elle est de plus en plus utilisée en sciences sociales et administratives ainsi que dans l'éducation (Merriam, 1998) pour leur portée pédagogique, « reposant sur la conviction théorique que notre connaissance de la réalité est imparfaite et que nous ne pouvons connaitre la réalité que de notre point de vue » (Cohen & Crabtree, 2016). Elle ne permet pas de dégager des lois universelles, mais des spécificités de phénomènes étudiés. En effet, une étude de cas est définie comme « une enquête empirique qui étudie un phénomène contemporain dans son contexte de vie réelle, surtout quand les frontières entre le phénomène et le contexte ne sont pas nettement évidentes » (Yin, 1981). Cette définition pointe la pertinence d'une approche par étude de cas pour un phénomène naissant, mais aussi les difficultés à dessiner les contours d'une étude de cas. En effet, il serait utopique de croire qu'un phénomène puisse être décrit entièrement, ou qu'il soit limité à une unité isolée aux frontières claires (Dumez, 2013). Cette nécessité de discuter la frontière du cas est mise de côté dans une autre définition issue des sciences de l'éducation : « les cas sont des histoires avec un message » (Herreid, 1997). La portée narrative et didactique d'une étude de cas est alors mise en valeur. Les deux approches sont assez complémentaires à partir du moment où le cas est étudié non seulement comme une réalité complexe, autonome et dotée de propriétés, mais aussi comme une histoire, un dialogue perpétuel d'action et de contraintes entre le cas et son environnement (Abbott, 1992). Cet aspect d'intrigue, ou « plot », met en contraste l'approche par étude de cas, dotée d'une continuité causale, avec les méthodes d'analyse de populations, basée sur des mesures d'évènements discontinus, et pose insidieusement la question du véritable début de l'histoire.

A ces questionnements en termes de frontière du phénomène et de périmètre temporel s'ajoute la difficulté des chercheurs à juxtaposer des cas ou à distinguer des sous-phénomènes, ce qui met en évidence un paradoxe (Dumez, 2013) : contrairement au terme employé, une étude de cas n'est pas une unité, mais un processus de comparaison systématique. Dumez met alors en perspective trois aspects de l'étude de cas. Tout d'abord, un cas est une instanciation empirique d'une classe de phénomènes, mais aussi une démarche de construction d'une unité théorique caractérisée, qui constitue en soi un résultat et non pas un prérequis théorique. Dans ce contexte,

il faut éviter les deux extrêmes de l'opportunisme méthodique (Girin, 1989), soit en orientant rapidement du point de vue théorique une étude de cas qui s'est présentée au hasard, soit en évitant de structurer trop vite une étude de cas issue d'un choix théorique afin d'éviter le risque de circularité. Ensuite, dans la mesure où une description complète est impossible, un cas prend son sens dans l'identification des incidents, décisions, pratiques routinières, et autres unités de sens qui constituent une multiplicité de cas dans un seul cas. Cela conduit à comparer ces unités de sens entre différents cas, empiriques ou théoriques, ou de façon dynamique. Enfin, il est nécessaire de distinguer (Lijphart, 1971) parmi les cas produits, les cas « a-théoriques » (purement descriptifs), les cas interprétatifs uniques (mis en perspective avec des lois générales), les cas permettant d'infirmer (ou conforter) une théorie, les cas déviants (non expliqués par des théories existantes) et les cas permettant de générer des hypothèses. Dans les deux derniers cas, la méthode abductive est la plus appropriée pour faire émerger des pistes d'enrichissement des théories de référence, voire de nouvelles théories, à partir des observations surprenantes sur les cas. Or, les résultats issus de démarches abductives n'ont pas de validité scientifique tant qu'elles n'ont pas été mises à l'épreuve par des méthodes quantitatives, notamment grâce à la triangulation (Denzin, 2012; Jick, 1979).

Ainsi, une étude de cas permet de clarifier le contexte complexe d'application d'un concept, et non pas d'établir une théorie et des lois universelles. Il ne s'agit pas non plus d'un simple exemple. Un cas est un « système intégré [d'un] intérêt secondaire ; il joue un rôle de support, facilitant notre compréhension de quelque chose d'autre » (Denzin & Lincoln, 1994; Stake, 1994), ce qui inscrit cette méthode parmi les techniques d'analyse situationnelles (Mucchielli, 1991). Cela donne la possibilité de mobiliser ainsi « un site d'observation permettant de découvrir et de suivre à la trace des processus particuliers » (Collerette, 1997), le cas étant « lui-même accessoire », mais doté de ses propres dynamiques qui nécessitent d'être explicitées à travers une recherche méthodique. Dans le cadre d'une étude d'un cas seul, ce cas permet de pointer des comparaisons entre les éléments réels dans le temps et entre ces éléments réels et des références théoriques. Dans le cadre d'une étude de cas multiples, les cas font l'objet, en complément des comparaisons dynamiques et théoriques, de comparaison entre eux afin de faire émerger des récurrences, des constantes, des différences. En effet, cette approche est plus appropriée lorsqu'un phénomène peut se produire dans une variété de situations. Or, si l'étude de cas unique s'inscrit clairement dans les méthodes qualitatives, l'étude de cas multiples est plus ambiguë.

En effet, la multiplicité des cas peut représenter un échantillon, ce qui rapproche l'étude des exigences d'études quantitatives. Il s'agit alors d'un moyen de cueillette d'informations à analyser à travers l'agrégation. Or, l'objectif des études de cas est d'aboutir à des propositions d'enrichissements théoriques généralisables, confirmées par leur reproduction dans des contextes similaires, et non pas à une généralisation statistique à une population, basée sur une énumération de fréquences (Yin, 1984). Il ne s'agit donc pas, selon Yin d'un « échantillon » proprement dit, d'autant plus que le paradigme de recherche qualitative reste plus approprié pour valoriser les études de cas approfondies. D'autres courants situent différemment les études de cas multiples dans les paradigmes de recherche qualitative. Par exemple, la théorie ancrée (Glaser & Strauss, 1967) est une méthode de recherche inductive qui s'appuie sur un cumul de cas, à l'instar d'une collecte de données, pour faire émerger sans hypothèse préalable des concepts similaires, puis des catégories de concepts qui serviront de base pour construire des théories. Cette approche a été mobilisée par un grand nombre de travaux de recherche en Sciences de Gestion, notamment sur les problématiques organisationnelles (Eisenhardt, 1989), et a donné lieu à des méthodes plus détaillées en termes de protocole. Notamment, le nombre de cas à mobiliser successivement est lié à la saturation théorique, c'est-à-dire qu'un cas complémentaire n'apporte plus d'éléments déterminants pour la généralisation de la théorie (Glaser & Strauss, 1967; Pires, 1997). Il s'élève, d'après Eisenhardt, à 4 à 10 cas, et constitue un échantillon. La théorie produite à partir de cette approche est propice à la découverte, testable et empiriquement valide, mais reste considérée comme de moyenne portée.

Ainsi, il est possible d'aborder l'étude de cas multiples comme une méthode purement qualitative abductive qui nécessite une validation quantitative postérieure, notamment par triangulation, ou alors comme une méthode inductive. La triangulation est dans ce cadre effectuée non pas à travers le test de la théorie sur d'autres cas et les méthodes quantitatives, mais grâce à la multiplicité des données, y compris quantitatives, collectées sur chaque cas : entretiens, observations, archives... Pour Yin, cette triangulation des données analysées est valable aussi. La cueillette d'informations requises pour préparer une étude de cas provient habituellement de six sources : des documents, des archives, des entrevues, l'observation directe, l'observation participante et des objets physiques. « Le chercheur doit s'obliger à recourir à plusieurs sources d'information pour s'assurer d'avoir couvert l'objet d'analyse sous divers angles » (Collerette, 1997). Les études de cas multiples semblent alors s'inscrire difficilement dans une conception dichotomique des méthodes scientifiques, quantitatives et qualitatives. Par ailleurs, ce dualisme a d'ores et déjà été dénoncé (Bourdieu et al., 1968;

Brannen, 2005; Hammersley, 1992) dans les sciences humaines et sociales, qui se démarqueraient des sciences naturelles par le principe du « refus de propositions transhistoriques dans les théories ou les hypothèses des sciences sociales ce qui implique, pour celles-ci, une science de la détermination « contextuelle » des actions sociales » (Groulx, 1997). Ce point de vue est contesté, et la résolution de ce débat dépasse le cadre de ces travaux, qui assument une approche qualitative des études de cas.

Ce choix de l'approche qualitative est guidé par le besoin de prendre en considération la complexité des processus analysés, marqués par un caractère évolutif dans le temps étant donné la récence du phénomène étudié. En effet, ces travaux visent à mettre en évidence les particularités des projets data dans plusieurs de leurs composantes et dans la dynamique qui lie ces composantes entre elles, sans se restreindre *a priori* en termes de composantes à analyser. Chaque étude de cas constitue ainsi une situation dans laquelle cette complexité peut être observée sous forme de dynamique des facteurs, de comparaison entre ses facteurs, ainsi que d'écarts par rapport aux modèles de référence. L'étude de cas multiples est aussi envisagée dans ces travaux selon une approche qualitative approfondie, accompagnée d'une collecte de données active *in situ*. Le recueil des données a été réalisé selon une démarche compréhensive sans établissement de cadre trop restreint *a priori,* et a mobilisé un raisonnement idiographique (Charmillot & Dayer, 2007; Groulx, 1997), marqué par la subjectivité du chercheur ainsi que des acteurs intervenus sur chaque projet data étudié. « Les règles méthodologiques recommandées qui caractérisent la compétence du chercheur sont les suivantes : la comparaison des données ; la saturation ; l'utilisation de cas négatifs ; la variation et la comparaison des sources ; la durée prolongée sur le terrain ; l'imprégnation distancée. ». La collecte, ainsi que le traitement des observations, c'est-à-dire l'analyse des composantes et des dynamiques d'intérêt, est le fruit d'une stratégie d'observation itérative, tour à tour inductive, abductive et déductive (Johansson, 2003), caractérisée par un ensemble d'éléments justifiant la « plausibilité » et la « crédibilité » des données et la validité des résultats, suivant un protocole de recherche établi *ad hoc*.

2.4 Stratégie d'observation

Le protocole de recherche de ces travaux est issu d'une construction progressive au grès des opportunités saisies sur le terrain (passage en recherche-action après une première phase de découverte), de l'évolution du contexte marché du phénomène étudié (multiplication et

évolution progressive des projets data au sein des entreprises), et du va-et-vient permanent entre les concepts théoriques et les pratiques sur le terrain minutieusement documentées.

2.4.1 Un protocole construit et sous contraintes

Deux phases majeures sont à distinguer dans le déroulé de ces travaux de recherche : la pré-expérimentation et la conduite d'études proprement dite.

La première phase de pré-expérimentation s'est basée essentiellement sur un processus de recherche inductif, réalisée sous forme d'entretiens, de collecte d'observations sur le terrain et d'analyse de documents sur 3 cas de projets data. Elle a permis une description intensive et une analyse contextualisée de facteurs intervenant dans le processus étudié et non nécessairement couvert par des théories ou modèles existants : en effet, aucun cadre théorique n'a été choisi en amont de ces entretiens. L'échantillonnage de ces 3 premiers cas a eu lieu selon deux critères majeurs : les acteurs intervenus dans le dispositif et interviewés considéraient qu'il s'agissait de projets ou de résultats de projets fortement liés au phénomène Big Data, le projet comportait des éléments techniques et analytiques nouveaux pour les experts métier, et il était possible d'approcher la réalité de ces cas sans limites particulières en termes de confidentialité, notamment pour mener librement les entretiens, ouverts. Dit autrement : il s'agissait des rares cas accessibles issus du phénomène Big Data (aspects technologiques, analytiques et mythologiques), mis en perspective dans ses dimensions historiques.

Cette pré-expérimentation a donné lieu à une précision de la problématique et à une première mise en perspective théorique de l'objet de recherche, sous la forme d'un rapport présenté en conférence (voir Annexe 6 - Présentation du rapport préliminaire) et intégré dans un ouvrage collectif (Chartron & Broudoux, 2015). Elle a notamment fait apparaitre le processus global, guidé le choix du modèle de projet data de référence par établissement de similitudes, et la précision du cadre conceptuel. Plus particulièrement, elle a permis son enrichissement théorique en mettant en évidence un processus de médiation, d'objectivation de sens, qui nécessitait un changement de protocole pour être observé. Ces nouveaux éléments ont guidé le changement de posture du chercheur et le basculement vers la recherche-action, restructuré le protocole de réalisation de l'étude de cas multiples et la construction simultanée des résultats qui en découle. En synthèse, les résultats de la première phase de pré-expérimentation sont les suivants :

- Rapport préliminaire

- 1 Compte rendu par cas (3 études de cas)
- Choix du modèle de référence
- Précision et émergence des hypothèses et précision des concepts théoriques relatifs
- Conception de la démarche de recherche pour la phase suivante

La deuxième phase du déroulé de ces travaux de recherche est la conduite de l'étude de cas complète sous la forme de recherche-action au sein du cabinet Quinten. Ce processus de recherche, essentiellement abductif, a permis l'identification de faits imprévus par le modèle de référence et la reconnaissance de similarités entre les différents cas. Cette émergence de faits a conduit à l'établissement de nouvelles hypothèses permettant de mieux cibler et restituer les éléments clés du phénomène complexe étudié. Ce processus a, à son tour, ouvert la voie à l'expérimentation déductive : le test de ces hypothèses s'est déroulé sur le terrain directement, avec confrontation de propositions d'actions aux acteurs. L'évaluation qualitative de ces hypothèses a alors eu lieu grâce au recueil d'appréciations de la part des acteurs, qu'il s'agisse de des équipes de Data Scientists de Quinten ou des experts métier côté client. La généralisation des observations, sous forme de proposition d'ajustement du modèle de référence et son enrichissement, est ainsi issue d'une combinaison des trois processus de recherche (inductif, abductif, puis déductif), avec une précision du domaine d'application des propositions formulées en fonction des spécificités des cas observés.

Avant de détailler la constitution de l'échantillon et le recueil d'observations et méthode de construction des résultats, il est important de noter que la nature économique de ces projets et l'intervention active du chercheur en tant que Data Scientist pose un ensemble de contraintes dans le traitement des données de recherche. D'une part, la relation client-fournisseur ne permet pas de s'adresser ouvertement aux clients sous la casquette de chercheur, ce qui restreint la collecte des retours des experts métier aux éléments habituels échangés dans le cadre du projet (supports de présentation, mails, échanges téléphoniques et discours en réunion, décisions budgétaires, retours d'expérience...). Cette contrainte ne semble pas inadéquate avec l'objectif de recherche consistant à capter les éléments représentant l'interaction habituelle entre les acteurs concernés par ces projets. D'autre part, la réalisation des projets au rythme des opportunités commerciales, de la disponibilité des intervenants, et des contraintes économiques diverses conditionnent la temporalité des projets et impliquent que

les études de cas se superposent entre elles dans le temps et ne peuvent être vues comme un ensemble de situations chronologiquement indépendantes ou linéaires. Cet élément est pris en compte à travers la restitution chronologique de chaque cas, ce qui permet de pointer les évolutions transversales. Enfin, ces contraintes économiques et logistiques restreignent la possibilité de choisir les projets d'intervention, ce qui limite l'échantillonnage et la possibilité d'expérimenter librement au cours de chaque projet dans le cadre des travaux de recherche-action. Elles sont toutefois bien prises en compte et détaillées pour chaque étude de cas, et font partie des résultats en tant que composantes directes d'un projet Data Science visant l'optimisation de la performance d'une organisation à ressources limitées.

La modélisation progressive des résultats, soumise aux contraintes du terrain évoquées précédemment, a impacté le choix de l'échantillon final, construit de façon cumulative jusqu'à saturation, c'est-à-dire jusqu'à ce que l'addition de cas complémentaires n'ajoutât pas d'éléments significatifs nouveaux aux résultats. Ces cas complémentaires ont tout de même été réalisés sur le terrain et décrits, ayant ainsi donné l'opportunité de tester la saturation (1 cas saturant) et, plus ponctuellement, les facteurs et propositions du modèle proposé (2 cas négatifs). Le modèle plus global peut faire l'objet d'une poursuite des travaux de recherche, selon les approches quantitatives, en s'appuyant sur une grille d'analyse construite selon le modèle proposé, et donc standardisée, et pré-testée sur l'un des cas (voir Annexe 7 – Grille d'analyse des études de cas selon une approche quantitative). Cette grille permet de tester les propositions d'ajustement du modèle de référence en termes de processus de projet Data Science, ainsi que ses 3 dimensions : la médiation entre acteurs impliqués dans un projet data, la prise en compte de la qualité des données, et la valeur.

Les résultats, présentés de façon enseignable, sont discutés et contextualisés afin d'être opérationnels et perfectibles dans les périmètres et champs d'action appropriés. Ils sont constitués des éléments suivants :

- 1 Compte rendu détaillé par cas avant saturation (4 études de cas détaillés)
- 1 Compte rendu par cas après saturation (3 études de cas, dont 1 saturant et 2 négatifs)
- Synthèse comparative des cas
- Mise en évidence des similitudes et des divergences entre les cas et le modèle de référence

- Proposition de modèle ajusté et enrichi sur les 3 dimensions clés (Brizo_DS)
- Databook – Structure du Prototype
- Grille d'analyse, outil standardisé de recueil d'observations primaires, testée sur 1 cas, et destiné à une application future à des cas en dehors de Quinten pour une approche quantitative

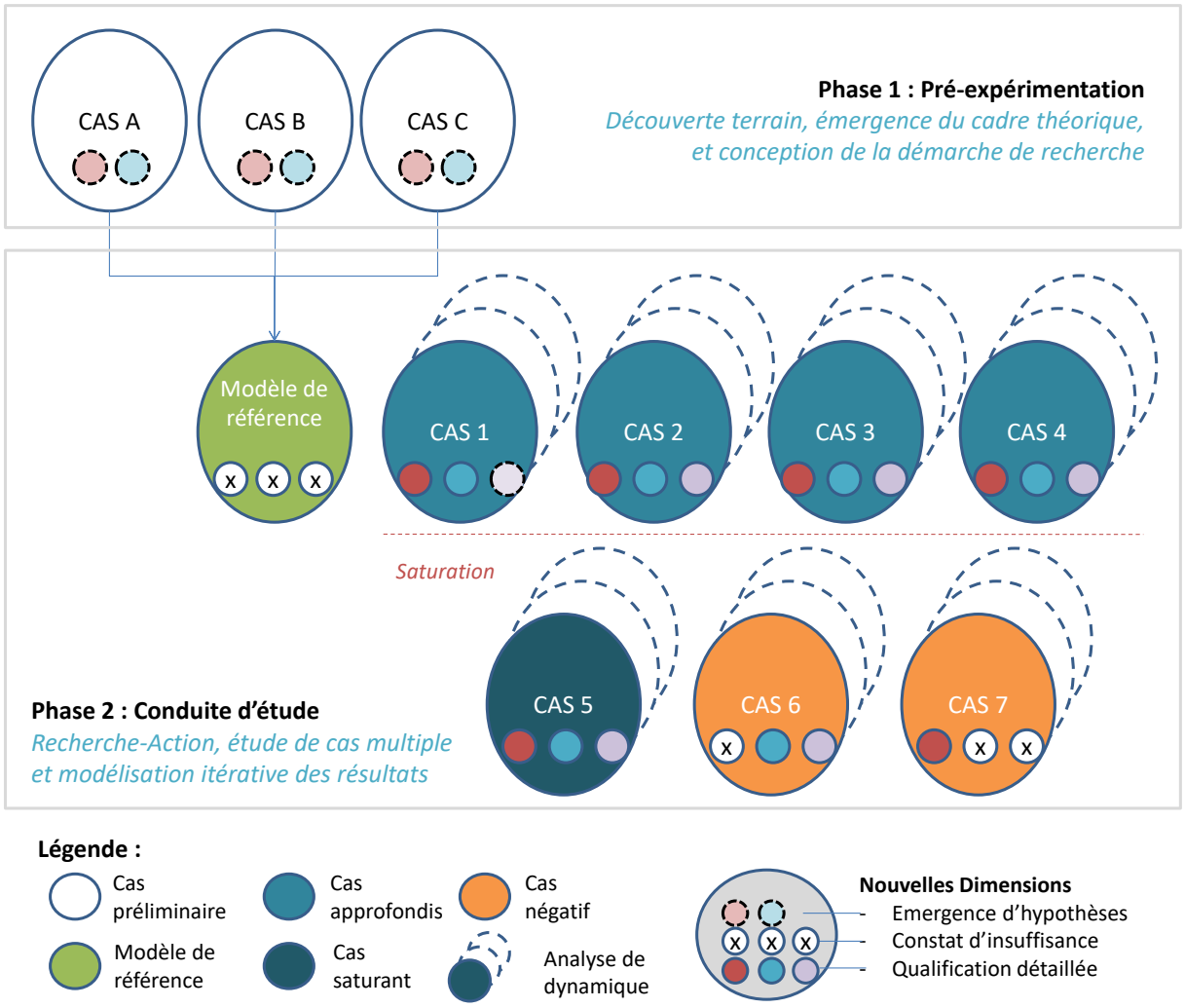


Figure 20 – Protocole de l'étude de cas multiples – *Représentation inspirée de Dumez, 2013*

Ce protocole (voir Figure 20) permet de générer des résultats plausibles et crédibles, notamment grâce à une description minutieuse du contexte, des résultats de chaque projet, et, pour les cas détaillés au cours de la recherche-action, des actions réalisées au cours des projets. Malgré un contexte d'intervention long et profondément imprégné, engagé auprès des clients et de la société Quinten, une distanciation a pu être assurée tout le long des travaux (Dudezert, 2018)

grâce à un travail itératif alternant la mise en perspective théorique et la pratique, et à la production de résultats scientifiques intermédiaires confrontés aux communautés de chercheurs (rapport intermédiaire suite à la pré-expérimentation, conférences, séminaires, deux présentations de résultats partiels, articles et interventions pédagogiques sur le sujet, échanges réguliers avec les directeurs de thèse). Enfin, la proposition de modèle résultant de ces travaux est issue d'un travail de comparaison permanent (dynamique temporelle des cas, similitudes et divergences entre les cas, similitudes et divergences par rapport au modèle de référence) jusqu'à atteindre une saturation du modèle, y compris en ajoutant des cas négatifs. Ces critères de validité du travail qualitatif ouvrent tout de même la voie à un test du modèle, qui reste provisoire jusqu'à la stabilisation du phénomène Big Data, selon des approches quantitatives, par exemple grâce à la grille d'analyse de cas d'ores et déjà testée sur un cas et à appliquer dans des contextes hors Quinten.

2.4.2 Conception de l'échantillon d'études de cas

La phase de pré-expérimentation s'appuie sur un échantillon initial de 3 cas pour préparer le cadre de recherche et le choix des cas qui constituent le cœur de ces travaux. Le premier cas est le dispositif télématique « Urgence » sur les plateaux d'assistance Inter Mutuelles Assistance[29], et illustre l'utilisation de nouvelles technologies dans le cadre de l'internet des objets, ou web sémantique, pour contribuer à la prise de décision en temps réel dans une situation d'accident routier marquée par des difficultés de recherche de causalité. Les deux derniers cas sont des exemples d'identification de tendances grâce à l'algorithme Q-Finder et autres dispositifs d'analyse de nouvelle génération et de data visualisation conçus par Quinten, appliqués à la prédiction de l'efficacité de traitement des patientes atteintes de cancer du sein triple négatif par le Centre Jean Perrin (Nabholtz et al., 2012) et à l'analyse de campagnes publicitaires par la régie M6 publicité[30,31]. L'approche terrain est basée sur des entretiens semi-structurés auprès de professionnels impliqués dans les projets (6 entretiens d'une durée d'environ 1h30), l'analyse de corpus de documents publics (rapports annuels, sites institutionnels, communication externe, études sectorielles…), le recueil et l'analyse d'informations internes

[29] http://www.ima.tm.fr/fr/notre_offre/assistance_routiere.php
[30] http://www.journaldunet.com/solutions/saas-logiciel/m6-publicite-et-datamining-avec-Quinten.shtml?utm_source=greenarrow&utm_medium=mail&utm_campaign=ml49_marchedesnaviga
[31]La correspondance de la publicité. 7 février 2014. n°15570, p. 20.

et une bibliographie interdisciplinaire. Cette pré-expérimentation donne lieu à une intervention en conférence internationale et à une publication académique (Chartron & Broudoux, 2015).

La réalisation de ces 3 cas met en lumière un ensemble d'éléments qui ont été clés pour l'établissement de la méthodologie de cette thèse. D'une part, la richesse interne de chaque cas est d'emblée identifiée comme un facteur de complexité fort. Cette complexité s'applique au processus chronologique des projets data marqué par des phases d'incertitude et d'exploration, aux compétences nouvelles et imprécises et à l'interaction entre les acteurs impliqués, à la diversité des méthodes analytiques et outils mobilisés, ainsi qu'à la variété des finalités de ces projets, mobilisant des données différentes pour des résultats propres à chaque objectif opérationnel et stratégique. Ces éléments sont perçus comme interagissant entre eux, et impactant fortement la création de sens et de valeur, ainsi que les indicateurs associés. D'autre part, l'absence de la vision sur le processus dans sa durée à travers la découverte du résultat *a posteriori* ne donne pas la possibilité d'appréhender l'alignement progressif entre les acteurs sur l'objectivation des connaissances générées et la création de leviers de valeur. Le processus d'alignement dans le cadre de projets data s'inscrit alors comme objet de recherche de référence, et appelle à un éclaircissement à travers l'approfondissement d'études de cas dans le temps, ce qui a pu être réalisé à travers le choix du terrain, mu par l'intention de répondre aux enjeux de complexité de l'objet de recherche et de compréhension approfondie du processus étudié.

La pertinence du choix de l'échantillon final est alors justifiée de façon suivante :

- Tous les projets s'inscrivent dans le phénomène Big Data tel que défini au sens large (Boyd & Crawford, 2012), c'est-à-dire mobilisant les méthodes analytiques, les technologies, et la mythologie associés

- Tous les projets étudiés impliquent une manipulation de données selon des méthodes qui n'étaient pas employées auparavant par l'entreprise, et sont réalisés par des équipes d'acteurs comprenant au moins un Data Scientist expérimenté, c'est-à-dire un « acteur social compétent » ayant pu acquérir de l'expérience intime et technique de projets data passés ayant porté leurs fruits. Ce paramètre garantit l'ancrage du cas dans la Data Science et donne la possibilité de décrire le(s) rôle(s) de Data Scientist(s) sur le projet.

- Chaque projet est reconnu comme utile, pertinent, et potentiellement créateur de valeur pour l'entreprise dans la mesure elle lui attribue un budget dans le but d'améliorer sa performance. Ce potentiel de création de valeur constitue une finalité hypothétique des projets data, qui justifie l'interdisciplinarité de cette thèse et sa posture téléologique qui consiste à croire que le phénomène impacte bien les indicateurs de valeur en entreprise. La reconnaissance de l'intérêt des résultats, c'est-à-dire la confirmation de la création de valeur, est captée à travers un retour positif de la part des entreprises, considéré comme un marqueur positif de bonnes pratiques : ce bouclage qualitatif est pris en compte dans les observations et dans les résultats.

- Etant donné le point précédent, le démarrage du projet, et donc la frontière chronologique du cas, est par convention identifié lors de la réunion de lancement du projet, soumise à la validation de l'allocation de ressources. Cette exclusion de la phase d'avant-vente, de la phase de confrontation plus large avec le phénomène Big Data, ou de l'évolution passée de l'usage visé par le projet fait alors l'objet d'une mise en contexte au cas par cas et à l'instant t. Cette frontière initiale des cas sera discutée dans les résultats.

- Tous les projets contribuent à optimiser le métier historique et le business modèle de l'entreprise. Ce critère exclut de la population d'analyse les entreprises dont l'interaction avec le phénomène constitue la création de sous-systèmes productifs aux business modèles dédiés, c'est-à-dire une recherche de bénéfice opportuniste. Il s'agit, par exemple, de la vente de ses données brutes ou bien de la prestation intellectuelle ou technique data. Ce critère de sélection justifie la notion de « métier », c'est-à-dire l'activité historique de l'entreprise ou d'une fonction de l'entreprise.

- L'implication du chercheur dans chaque projet est directe (contribution active à l'effort de production collectif) afin de justifier de sa compréhension de la complexité de chaque situation. Ce critère implique un biais de subjectivité lié d'une part à l'appartenance à l'entreprise Quinten et à ses pratiques, et d'autre part aux compétences du chercheur.

L'échantillon ainsi constitué est alors composé de l'ensemble des projets réalisés par Quinten avec la participation du chercheur sur une durée significative (entre mars 2015 et mars 2017). Il illustre certains projets, et ne prétend pas à une représentativité statistique ou exhaustivité quelconques. Il s'agit de 7 projets sur des problématiques métier variées, marqués par des spécificités fortes, et réalisés de bout en bout par les équipes Quinten avec des entreprises issues des secteurs de l'assurance et du parfum. Sur ces 7 projets, 4 font l'objet d'une analyse

approfondie et d'un compte rendu détaillé. En effet, l'ajout de cas complémentaires n'apporte pas d'éléments nouveaux pour la construction des résultats, selon le principe de saturation, ni d'élément discriminant complémentaire dans le cadre de l'analyse permettant d'apporter de nouveaux contrastes. Parmi ces 3 cas complémentaires, le premier sert à constater la saturation, et les deux suivants à tester les propositions du modèle : il s'agit de cas négatifs, dans le sens où les 3 dimensions d'enrichissement du modèle sont mises à l'épreuve, les facteurs de succès étant absents ou incomplets. En effet, le premier cas ne débouche ni sur une manipulation de données ni sur un usage direct (*a priori*, il ne génère donc pas de valeur), et le second fait l'objet d'une médiation pauvre et d'une documentation de la qualité des données insuffisante. L'ensemble des cas (voir Figure 21) est comparé à la lumière des concepts émergés, ce qui permet de montrer les critères d'applicabilité des résultats.

#	Nom du cas	Secteur
Pré-expérimentation		
Cas A	Dispositif télématique "Urgence"	Assistance
Cas B	Cancer du sein triple négatif	Santé
Cas C	Placement publicitaire	Medias
Cas réalisés et détaillés		
Cas 1	Attrition assurance santé	Assurance
Cas 2	Prévision d'activité	Parfum
Cas 3	Prévention santé et prévoyance	Assurance
Cas 4	Contrôle de non-conformité	Assurance
Cas réalisés non détaillés		
Cas 5	Sinistres lourds en dommage aux biens	Assurance
Cas 6	Prédiction de prix des agrumes	Parfum
Cas 7	Multi-équipement	Assurance

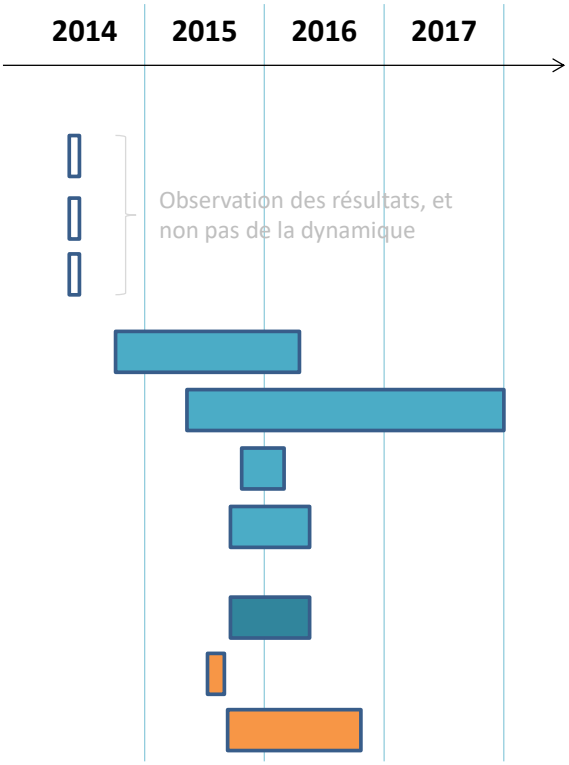


Figure 21 – Liste et enchainement des études de cas composant l'échantillon d'analyse

2.4.3 Recueil d'observations et modélisation itérative de résultats

Le recueil des observations sur chaque cas étudié est réalisé selon 3 niveaux de lecture.

Premièrement, chacun des 7 cas est mis en perspective dans son contexte propre afin de permettre au lecteur d'appréhender la situation sous l'angle de la combinaison de finalités spécifiques et des particularités des résultats attendus. Cette mise en perspective du contexte est

réalisée sous forme de description de la situation initiale et des objectifs du projet ainsi que de la situation finale à l'issue du projet : il s'agit d'une synthèse de cas. La description du contexte est complétée par une représentation systémique de la fonction impliquée dans le projet data, sous la forme d'un modèle Input-Process-Output inspiré des recherches industrielles sur la fonction R&D au sein d'entreprises (Samsonowa et al., 2009). Cette représentation met en lumière les impacts du projet sur le système à travers la modification des paramètres du processus métier, correspondant aux ressources, activités et résultats du processus métier modifiés par l'usage mise en œuvre (voir Annexe 8 - Modèles Input-Process-Output détaillés). Cette double représentation du contexte permet une compréhension approfondie et spécifique de chaque situation, et facilite la comparaison des impacts de projets data.

Deuxièmement, chacun des 4 cas approfondis est détaillé sous l'angle de l'analyse de processus, au cœur de ces travaux de recherche. La restitution des observations est alors effectuée sous forme de compte rendu, structuré de façon chronologique afin de mettre en évidence les actions et les décisions successives qui ont fait évoluer le dispositif projet. Chaque compte rendu est illustré par un ensemble d'éléments issus de la pratique des acteurs du dispositif projet, tels que des extraits de rapports intermédiaires et finaux, des extraits de supports de présentations orales et des points d'avancement, des maquettes d'interfaces homme-machine, des représentations graphiques des résultats analytiques et de leur évaluation, des représentations d'outils de travail sur les données, des mails... Ces extraits de documents sont décrits en détail afin d'établir un lien indiscutable entre la mémoire des projets partagée par les acteurs impliqués, et la mémoire des travaux de recherche construite par le chercheur sous la forme de compte rendu détaillé.

Troisièmement, l'ensemble des 10 cas (3 cas de pré-expérimentation, 4 cas détaillés, et 3 cas non détaillés) est restitué de façon synthétisée, agrégée, afin de mettre en perspective les similitudes et les divergences. La comparaison s'effectue en termes de contextes initiaux, d'application des processus identifiés comme clés, et d'atteinte des résultats. Cette comparaison a pour objectif d'alimenter les propositions de cette thèse grâce à leur mise en perspective avec l'atteinte des finalités du projet, mais aussi avec l'identification d'éléments de création de valeur inattendus, c'est-à-dire non prévus initialement lors de la décision de lancement de projet. La réintégration des cas complémentaires dans cette synthèse est un facteur de diversification en complément du cas saturant et des cas négatifs. Cette synthèse des observations constitue la maille de lecture la plus simple pour comprendre les résultats.

Enfin, pour l'un des cas, la vision qualitative et complétée par une grille d'analyse conçue *ad hoc* et revue au fur et à mesure de l'avancement des travaux (voir Annexe 7 – Grille d'analyse des études de cas selon une approche quantitative). Cette grille est une matrice qui structure de façon standardisée la collecte des observations à travers la description, minutieusement détaillé, des actions réalisées à chaque étape du projet. Cette matrice est à double entrée pour une lecture croisée des processus Data Science et de la création de sens sous l'angle de 3 dimensions : la médiation, la qualité des données, et les indicateurs (voir Figure 22). Chaque ligne de la matrice représente alors une unité d'observation ordonnée dans le temps, soit une activité réalisée au cours du projet. Une activité est considérée comme une action ou un ensemble d'actions homogènes (tâches) réalisées par les mêmes individus et donnant lieu à un résultat intermédiaire tangible, précisé dans la matrice. La structure verticale de la matrice est chronologique, ce qui permet de rendre compte des enchainements, des itérations des superpositions d'activités et des instances de prise de décision. Chaque tâche est décrite et catégorisée selon le modèle de processus Data Science ajusté. Le rôle de chaque acteur réalisant cette tâche est précisé. Chaque tâche présentant un intérêt particulier sur un axe clé (qualité des données, médiation, et impact sur les indicateurs) est commentée sous cet angle. Cette matrice d'observation permet non seulement de justifier les propositions d'ajustement du modèle de référence, mais aussi de le compléter.

Cette grille constitue un facteur de validité interne dans la mesure où sa restitution associée au contexte et à la synthèse du cas permet au lecteur d'avoir une visibilité claire sur la situation telle que vue par le chercheur. Bien que sa formalisation détaillée n'ait eu lieu que pour un seul cas, dont l'observation a commencé en dernier (c'est-à-dire presque à l'issue du travail de modélisation des résultats qui a permis de standardiser les axes de la matrice), tous des éléments qui la composent sont décrits dans les comptes rendus de l'ensemble des études de cas, c'est-à-dire que tous les axes de cette grille ont fait l'objet de l'analyse. En plus de cette vision dynamique illustrée des comptes rendus, les trois dimensions principales sont, pour chaque projet, résumées dans un chapitre dédié aux observations clés.

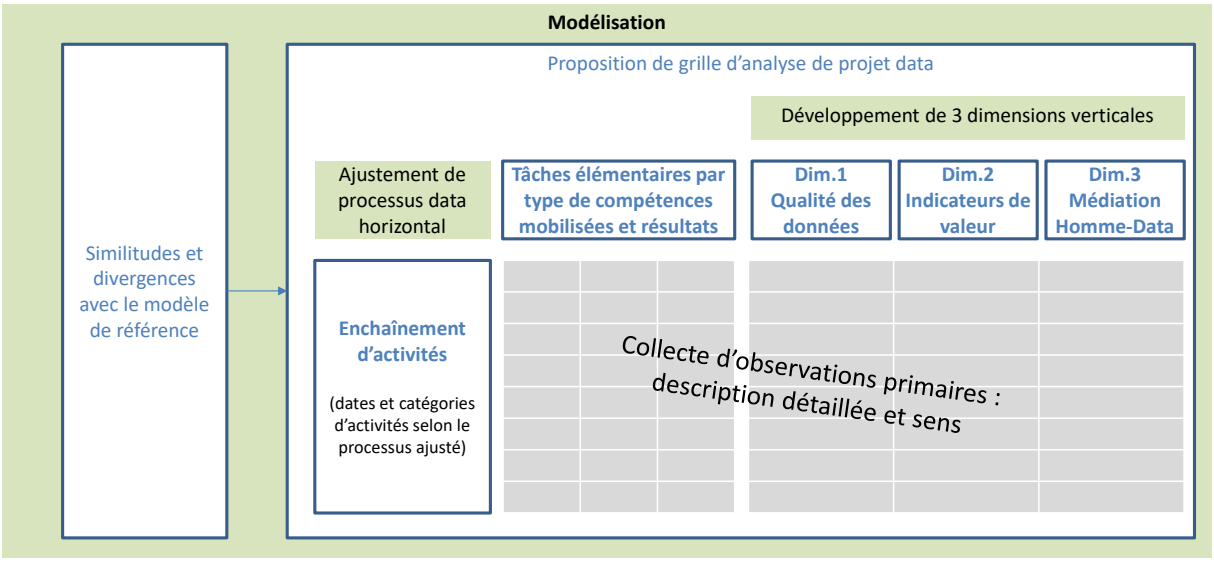


Figure 22 – Description du modèle de la grille d'analyse

L'usage de cette grille rapproche la méthode de recherche des approches quantitatives, possibles à travers l'étude de cas multiples (Yin, 1984), étant donné qu'il permet une comparaison entre les cas multiples, c'est-à-dire l'identification des constantes et des différences. Cependant, cette thèse n'a pas la prétention d'avoir recueilli un échantillon représentatif afin d'appuyer les connaissances produites selon une méthode quantitative positiviste. Par ailleurs, cette grille n'a pas constitué un filtre qui aurait empêché l'attention donné aux spécificités de chaque cas, étant donné que sa construction a été réalisée au fur et à mesure de l'avancement des travaux de recherche. Elle n'en reste pas moins une proposition de méthode d'analyse quantitative, résultant des travaux de recherche-action, applicable *a posteriori* en dehors d'une démarche de recherche-action et en dehors de Quinten, et constitue un préalable à la diversification des méthodes de validation des connaissances produites. En effet, « dans le domaine des sciences humaines et sociales, les approches quantitatives et qualitatives peuvent sans doute cohabiter et se compléter dans un même programme de recherche, que ce soit à différents stades d'une recherche ou pour examiner un objet sous des angles variés » (Collerette, 1997). Ainsi, seuls les critères de validité issus de l'approche qualitative sont pris en compte dans cette thèse.

Les résultats de cette thèse ont été issus d'une modélisation itérative, ce qui rend floue la frontière entre la restitution des observations et les résultats de l'analyse de ces observations. Ils sont constitués de 2 familles d'éléments : la description des cas et la proposition de modèle de dispositif « projet data », ajusté en termes de processus et enrichi. Si la description détaillée

des cas a une vocation narrative, dans un but de partage de connaissances entre la recherche et le monde professionnel, le modèle de dispositif « projet data » a une visée essentiellement opérationnelle dans la mesure où il contient des recommandations de pratiques utiles à la création de valeur à travers les projets data pour les organisations. Il constitue par ailleurs un ajustement de modèle de référence, développé pour le Data Mining et adapté à la réalité des projets Data Science actuels, et se veut enseignable. Ces résultats s'inscrivent pleinement dans la recherche-action, et ouvrent des voies inédites pour la généralisation grâce à des approches complémentaires, positivistes. Ils constituent d'ores et déjà une première brique issue d'un processus de généralisation analytique (Yin, 1984) constituant un corpus d'actions logiquement articulées et prenant un sens selon les axes d'enrichissement.

***

Les méthodes présentées dans cette partie, qualitatives et marquées par une posture constructiviste, permettent ainsi de bâtir les résultats suivants : les études de cas, détaillés, synthétisés et comparés sur les 3 dimensions clés identifiées dans l'état de l'art comme manquantes, un modèle de processus projet Data Science, basé sur une critique du modèle de processus de référence et une nouvelle proposition de modèle de dispositif de projet data, articulant ces 3 dimensions, et enfin une discussion des limites de ces travaux de recherche avant de conclure et de présenter les perspectives de recherche.

Troisième partie :

Résultats

1 Exposé des études de cas

Le premier chapitre de cette partie est une description de la matière première de ces travaux de recherche, c'est-à-dire un exposé de l'ensemble des études de cas. Le chapitre commence par une synthèse (voir 1.1) qui met en perspective l'ensemble des cas, leurs contextes, résultats analytiques et spécificités. Cette synthèse est complétée par des clés de lecture de la suite du chapitre, c'est-à-dire les méthodes de collecte et de restitution des observations pour chaque cas, qui correspond à un projet data complet et indépendant. Les 10 projets qui constituent cette étude de cas multiples sont alors regroupés en trois lots dont la méthode de construction est décrite en synthèse : la pré-expérimentation (voir 1.2), les cas réalisés et détaillés (voir 1.3), et les cas réalisés non détaillés (voir 1.4). La lecture de ces trois parties, détaillant chaque projet de façon individuelle, n'est pas indispensable pour passer à l'état des lieux des observations clés (voir 1.5) et au deuxième chapitre de la partie résultat présentant le modèle que ces cas permettent de bâtir.

1.1 Synthèse des études de cas

Les études de cas présentent une grande diversité en termes d'usages dans les secteurs aussi variés que l'Assurance, l'industrie du Parfum, les Médias, la Santé, et l'Assistance (voir Figure 23). Le secteur de l'assurance est le plus représenté avec cinq cas sur dix, voire six car l'Assistance peut être considérée comme l'une de activités du secteur. Cette prépondérance est expliquée par mon rôle dans la société qui comprend le développement de l'activité du secteur Assurance, mais aussi par le fait que ce secteur est très appétant à se lancer dans ce type de projets pendant la période étudiée, c'est-à-dire entre 2015 et 2017. Les enjeux sont ainsi très différents d'un cas à l'autre, le cas santé étant le moins marqué par des enjeux économiques. Les trois premiers cas (A, B et C) constituent le lot de pré-expérimentation où les observations ont été recueillies sur le terrain sous forme d'interviews, après la fin du projet. Les sept cas suivants ont été réalisés en tant que Data Scientist salarié de Quinten et acteur dans chaque projet. Seuls les quatre premiers cas sont complétés par un compte rendu détaillé, car l'apport des trois cas suivants n'a pas permis d'obtenir de nouveaux éléments dans ces travaux de recherche : il s'agit de cas de saturation, dont deux cas négatifs (6 et 7).

#	Nom du cas	Secteur	Nature du résultat analytique du projet	Caractère discriminant
Pré-expérimentation				
Cas A	Dispositif télématique "Urgence"	Assistance	Identification d'alertes automatiques sur les véhicules accidentés	Projet non réalisé par Quinten
Cas B	Cancer du sein triple négatif	Santé	Identification de critères miximisant l'efficacité du traitement médical	Finalité économique secondaire
Cas C	Placement publicitaire	Medias	Recommendation de critères pour le placement de publicités	Projet réalisé sur le long terme (poursuite des itérations)
Cas réalisés et détaillés				
Cas 1	Attrition assurance santé	Assurance	Profils de clients à risque d'attrition et score de risque individuel	Niveau d'alignement initial faible
Cas 2	Prévision d'activité	Parfum	Chiffre d'affaires et volume mensuels prédit sur 12 mois	Restitution de résultat sous forme d'applicatif métier
Cas 3	Prévention santé et prévoyance	Assurance	Typologies de comportement de consommation santé individuels par secteur d'activité	Mobilisation de la plateforme Big Data sur place
Cas 4	Contrôle de non-conformité	Assurance	Profils de contrats à risque de non-conformité et scoring associé à chaque contrat	Mesure de la performance facilitée par l'usage
Cas réalisés non détaillés				
Cas 5	Sinistres lourds en dommage aux biens	Assurance	Profils de contrats à risque de non-conformité et scoring associé à chaque contrat	RAS
Cas 6	Prédiction de prix des agrumes	Parfum	Prix d'achat d'essence d'agrumes à horizon d'un an	Projet arrêté avant la livraison les résultats analytiques
Cas 7	Multi-équipement	Assurance	Profils de client succeptibles de souscrire à une 3ème famille de contrats d'assurance	Compétences, interactions et documentation insuffisantes

Figure 23 – Synthèse des études de cas

Tous les projets, sauf le premier, ont été réalisés par les équipes de Data Scientists de Quinten, société qui possède plus de 7 années d'expérience sur ce type de projets et des compétences assez variées (métier, statistiques et informatique) au sein des profils communs de « Data Scientists », et mobilise un éventail non limité d'algorithmes et d'outils technologiques : en effet, chaque projet conduit à choisir, en interne ou dans la palette des possibilités Open Source, les algorithmes à paramétrer et les outils appropriés pour réaliser les traitements nécessaires.

L'activité de la société a évolué au cours de cette étude : l'activité historique s'appuyait essentiellement sur la mobilisation d'un algorithme propriétaire, le Q-Finder. Cet algorithme permet d'extraire d'un jeu de données des sous-populations qui maximisent un phénomène d'intérêt, c'est-à-dire identifier des contextes explicites qui concentrent un risque (décès, attrition, sinistre, non-conformité…) ou une performance (vente de produit, efficacité publicitaire, efficacité d'un traitement médical…). L'algorithme, de nature prescriptive, a été utilisé pour les deux projets observés au cours de la pré-expérimentation de façon isolée, ainsi que sur quatre projets réalisés avec ma participation. Cette récurrence n'est pas sans impact sur la culture de la société : en effet, l'algorithme n'est pas un traitement de type « boite noire » et fournit des résultats intelligibles et facilement vérifiables par les acteurs métier. Cela contribue à chercher une culture de qualité des échanges entre acteurs impliqués et de lisibilité des

résultats, ce qui pousse à aborder les autres algorithmes avec une approche similaire. Parmi les six projets utilisant une approche prescriptive, trois projets ont fait l'objet d'une approche mixte : en effet, d'autres algorithmes ont été utilisés, comme des arbres de décision ou du scoring prédictif. Les modèles prédictifs ont été utilisés en tout dans cinq projets sur les sept observés. Dans la plupart des cas, de nombreux modèles prédictifs ont été utilisés pour un seul et même projet. Enfin, tous les projets ont fait l'objet d'une application de modèles descriptifs divers et variés, allant de simples analyses mono-variées à des algorithmes à la pointe de la recherche, en complément des modèles prédictifs et prescriptifs. Ainsi, le spectre des modèles algorithmiques mobilisés est très large, et leur usage articulé est riche et différencié, ce qui permet de prendre en compte les pratiques en Data Science plus globales et non pas le simple paramétrage d'un algorithme donné. Pour rappel, l'ensemble des familles d'algorithmes utilisés en Data Science a été synthétisé par Gartner (voir Figure 13), présenté dans l'activité de Quinten sous forme de palette d'approches algorithmiques mobilisables par la société (voir Figure 19, Deuxième partie, Chapitre 1 « Choix du terrain », page 128), et décrit en Annexe 4 - Data Science et algorithmes.

Les différences analytiques ne semblent pas avoir des impacts particuliers sur les processus et les modes de fonctionnement des projets. En revanche, la nature du résultat est assez discriminante : en effet, le cas 2 (Prévision d'activité) a donné lieu au développement d'une interface métier qui a dû être déployée en Saas (Software As A Service). Cette application a été utilisée tout le long du projet, et a fait l'objet d'évolutions. Cette différence explique la durée relativement longue de ce projet. Il s'agit du seul projet observé sur le terrain ayant donné lieu à une restitution de résultat sous forme d'applicatif métier. Cependant, d'autres applicatifs ont été approchés au cours de la pré-expérimentation, à travers deux applicatifs utilisés en télématique et un applicatif de placement publicitaire. Ces trois projets sont tous marqués par leur durée, et le placement publicitaire plus particulièrement par la récurrence de l'intervention des Data Scientists (évolution de l'applicatif, mais aussi réapprentissages manuels, contrairement au projet de prévision d'activité qui a donné lieu au développement d'un algorithme auto-apprenant). Par ailleurs, trois des autres cas ont nécessité le développement d'une interface homme-données conçue sur mesure, dans un format Excel. Cette interface se distingue des applicatifs métier par sa finalité : elle n'est pas conçue pour être déployée, mais pour accompagner la prise en main des résultats par les métiers, le temps du projet. Ainsi, trois cas ont pour résultats des solutions applicatives, et trois cas ont pour résultat des recommandations d'hypothèses de travail, issues des analyses et mises à disposition dans des

interfaces lisibles. Les autres projets ont aussi généré des recommandations d'hypothèses de travail, mais présentées seulement sous forme de rapports plus classiques.

Les volumétries en jeu ont été assez variées entre ces projets, qui ont tous été réalisés sur des ordinateurs personnels des salariés de Quinten, des serveurs classiques ou des clusters de type Spark déployés en interne chez Quinten (pour rappel, ces dispositifs sont décrits en Annexe 2 - Data Lakes et Informatique Décisionnelle), sauf un projet : il s'agit du projet prévention santé prévoyance, dont l'un des objectifs était justement de tester une nouvelle plateforme Big Data qui venait d'être mise à disposition en interne chez le client. Par ailleurs, l'un des projets (attrition en assurance santé) devait être réalisé à partir des données déversées dans le Data Lake du client afin de le valoriser. Cet objectif n'a pas été atteint car les données disponibles se sont avérées inexploitables : il a fallu donc remonter jusqu'aux données sources. Ces points techniques ne présentaient finalement que des différences assez mineures par rapport aux autres projets, liés au rodage technique de la nouvelle plateforme ou à une durée plus longue de la phase de compréhension data.

En revanche, la maturité des dispositifs projet a été assez discriminante pour leur déroulement. Tout d'abord, l'absence d'alignement initial entre les équipes data et métier, mais aussi en interne entre les métiers (au cours du projet attrition en assurance santé) a généré des échanges laborieux avant la mise en place d'un système dédié d'instances d'arbitrage. Inversement la maturité relative du dispositif (projet contrôles de non-conformité), avec un expert métier possédant une expérience solide en statistiques, a permis d'avoir un usage non seulement plus mesurable que les autres, mais aussi une mise en œuvre plus rapide, avec une observation de bénéfices mesurables et plus élevés que le potentiel. Cette performance n'est cependant pas liée uniquement au niveau de maturité du dispositif, mais aussi à la nature de l'usage, au contexte de son application, ainsi qu'à la poursuite des travaux en interne en intégrant les compétences Data Science.

Plus largement, chaque cas fait l'objet d'une représentation détaillée des impacts du projet sur le fonctionnement de l'entité cliente selon le modèle IPO (Input-Process-Output) proposé dans des travaux de recherche ultérieurs (Samsonowa et al., 2009) : ces représentations illustrent la création de nouveaux usages, détaillent la nature de l'impact, et permettent de faire le lien entre ces nouveaux usages et la création de valeur. Les impacts sont détaillés en annexe pour chaque cas (voir Annexe 8 - Modèles Input-Process-Output détaillés) et sont synthétises dans le corps du modèle proposé. Les comptes rendus détaillés des cas permettent de voir les contextes et la

dynamique de génération de ces usages, tout le long du projet et de façon plus chronologique, afin de mieux percevoir les contextes d'émergence des connaissances sur lesquelles s'appuient ces usages. Cette mise en perspective permet d'orienter le processus de production vers des usages plus larges que celui qui est initialement visé par la production analytique, c'est-à-dire des usages intermédiaires et indirects. Le cas des sinistres lourds en dommage aux biens constitue dans ce sens une confirmation des propositions issues des quatre cas précédents, notamment en démontrant que le projet a permis de découvrir des pistes de mise en qualité des données avant même de livrer les résultats prévus.

Enfin, deux cas sont ici des cas négatifs. Le premier est destiné à tester la génération de valeur en absence de travail analytique, lorsque l'objectif métier est incompatible avec une approche Data Science. Et le second permet de tester la génération de connaissances et la capacité à innover en cas de Médiation Homme-Données très faible. Ainsi, les processus identifiés comme clés, c'est-à-dire la production analytique et la Médiation Homme-Données, sont mis à l'épreuve. Cet ajout de cas permet de contraster la finalité des deux processus, et de proposer l'intégration de la Médiation Homme-Données au cœur des résultats de ces travaux. Ses facteurs de succès sont issus de l'ensemble des instances de médiation vécues sur les autres projets, sous la forme de comités de projet regroupant les interlocuteurs clés et permettant un alignement progressif sur les résultats attendus et sur la capitalisation de connaissances. Ces facteurs clés sont les outils de capitalisation de connaissances soutenant ces comités, sous la forme de présentations support et de Databook, la clarification du fonctionnement des algorithmes avec partage des méthodes de construction pour garantir leur transparence, l'utilisation des interfaces homme-données et des prototypes de solutions applicatives pour s'assurer d'une représentation optimale des concepts clés, ou encore la gestion de projet, armée d'outils classiques (plannings, analyse de risques…) adaptés au contexte d'incertitude des projets data. Ces facteurs ont notamment permis des cas de création de valeur inattendus, plus particulièrement à travers leur impact sur la capacité d'innovation.

Aide à la lecture des cas :

Les dix études de cas présentées dans ce chapitre sont traitées de façon comparable. Tous les cas présentent un chapitre « Contexte et Enjeux » et un chapitre « Synthèse des résultats » afin de situer les projets dans leur contexte de lancement et de comprendre rapidement leur aboutissement. Les quatre cas réalisés et détaillés contiennent, en complément, un chapitre « Compte Rendu du projet » : ce chapitre est une description détaillée du dispositif mis en place,

du déroulé du projet dans le temps et des incertitudes rencontrées et traitées. Contrairement à la pré-expérimentation, le compte rendu est rendu possible par l'implication directe dans les projets en tant que membre de l'équipe de Data Scientists (les rôles seront précisés projet par projet). Chaque compte rendu est illustré par des éléments conçus par les Data Scientists de l'équipe projet : il ne s'agit pas de figures réalisées dans le cadre de ce travail de recherche, mais d'éléments inédits extraits du travail terrain (illustrations anonymisées tirées des dossiers, des présentations, des fichiers de travail, extractions des outils d'exploration des données, des mails...). Ces extraits sont numérotés dans l'ordre de conception. Pour faciliter la lecture des résultats, seuls deux comptes rendus sont maintenus dans le corps du chapitre (l'un correspond au projet attrition en assurance santé, qui aboutit à une génération de connaissance, l'autre au déploiement d'un applicatif métier pour la prévision d'activité), les deux autres figurant en annexe.

Si le compte rendu détaillé est réalisé de façon linéaire, chaque étude de cas possède par ailleurs un récapitulatif des observations clés dans un chapitre « Observations clés ». Ce chapitre permet de faire le lien entre le projet et les trois dimensions du modèle final : la qualité des données, les indicateurs de valeur et la Médiation Homme-Données. La qualité des données et la liste plus précise d'acteurs et de données exploitées ne sont pas abordées dans les trois cas de la pré-expérimentation, car l'observation *a posteriori* des résultats du projet ne permettait pas de mettre en évidence ces éléments. La réalisation de cas a permis de faire émerger progressivement ces dimensions, et à travers elles une grille d'analyse quantitative qui a fait l'objet d'une formalisation pour le dernier cas détaillé. Pour rappel, cette grille est décrite en annexe (voir Annexe 7 – Grille d'analyse des études de cas selon une approche quantitative) et peut être utilisée pour d'autres recueils d'observations terrain dans un second temps. Elle comprend ces trois dimensions, déclinées à travers l'enchainement chronologique des activités réalisées tout le long du projet sous la forme de tâches effectuées par les différents acteurs mobilisés.

Des compléments sont présents en annexe, notamment les illustrations des cas de pré-expérimentation tirées de la présentation des résultats intermédiaires (voir Annexe 6 - Présentation du rapport préliminaire) et les modèles IPO formalisés pour chaque cas (voir Annexe 8 - Modèles Input-Process-Output détaillés), ainsi que des illustrations tirées des Databooks des projets (voir Annexe 11 – Databook : genèse et prototypage).

1.2 Pré-expérimentation

Dans ces trois premières études de cas, l'accent est mis sur la génération de valeur, car, à ce stade des travaux de recherche, la dynamique interne du projet n'est pas observable (le modèle de référence n'est pas encore confronté à la réalité du terrain) et les premières dimensions qui enrichiront le modèle de référence ne font qu'émerger. En effet, les dimensions de valeur et de médiation sont examinées, mais les projets observés *a posteriori* restent opaques sur les traitements des données, et donc sur les enjeux liés à leur qualité. La pré-expérimentation fait l'objet d'une première présentation en colloque international à Rabat en 2015 (voir support de présentation en Annexe 6 - Présentation du rapport préliminaire) et à une publication en 2016 sous forme de chapitre dans l'ouvrage collectif qui couvre ce colloque (Chartron & Broudoux, 2015).

1.2.1 Cas A : Dispositif télématique « urgence »

1.2.1.1 Contexte et enjeux

Deuxième entreprise d'assistance en France, Inter Mutuelles Assistance est un groupement d'intérêt économique qui assiste près d'un français sur deux. Le groupe IMA distribue des contrats d'assistance auprès de professionnels (sociétés d'assurance, constructeurs automobiles, mutuelles santé, banques, groupe de presse, sociétés multinationales, grande distribution…). Les prestations d'assistance réalisées par le groupe IMA sont de nature technique et médicale 24h/24, 7j/7, à l'international. Pour réaliser ces prestations, IMA s'appuie sur 11 plateformes d'assistance et des infrastructures téléphoniques et informatiques permettant de répondre aux volumes d'activité.

Depuis plus de 15 ans, IMA a construit des partenariats avec des constructeurs comme Peugeot et Citroën[32], pour développer progressivement un dispositif qui permet d'automatiser la détection d'accidents au niveau du véhicule et de le relier directement aux plateaux d'assistance.

1.2.1.2 Synthèse des résultats

Ce projet long a débouché sur le déploiement d'un outil d'assistance et d'« Urgence » connectées. Les services « Urgence » sont en 2015 traités par une équipe dédiée de chargés d'assistance, spécialement formés au sein de leur service. Grâce aux progrès en termes

[32] Depuis janvier 2014 Renault a également choisi de confier son assistance en France à IMA

d'interopérabilité, ce dispositif relie en temps réel le véhicule équipé d'un boitier télématique autonome (BTA[33]), le constructeur et ses partenaires dont IMA, lui-même connecté aux PC autoroutiers et à terme aux secours publics[34]. Le dispositif permet de suivre trois types de déclenchement de prise de contact : Emergency Call (eCall) automatique suite à la détection de choc par les capteurs du véhicule, eCall manuel avec un bouton SOS, et Breakdown Call manuel (bCall) avec un bouton au logo constructeur en cas de panne. Ces déclenchements donnent lieu à des flux informationnels différents, les paramètres ayant été convenus lors de la conception des outils d'interface. L'intégration des données comme la géolocalisation, les identifiants du véhicule, le détail des chocs et le type de déclenchement permettent d'accélérer et de fiabiliser le recueil de données, la qualification de la situation et le choix de transfert ou de filtrage par le gestionnaire d'assistance (pour plus de détails illustratifs, et notamment les interfaces homme-machine mobilisées dans le dispositif, voir Annexe 6 - Présentation du rapport préliminaire).

1.2.1.3 Observations clés

Indicateurs de valeur :

Le paramètre de temps de traitement est optimisé par le dispositif, impactant le coût interne du dossier, un des principaux indicateurs clés de l'entreprise. Par ailleurs, la capacité à réagir en temps réel, mesurée par le nouveau paramètre de délai d'appel des secours, normalisé au niveau européen[35] avec une cible de 90 secondes, contribue à affirmer un avantage concurrentiel, et ainsi à renforcer la présence d'IMA sur ce marché. L'eCall devrait réduire le temps d'intervention des secours de 40 à 50% « et sauver 2500 vies par an » selon la Commission Européenne, qui a initié la réglementation visant à rendre l'eCall obligatoire sur les nouveaux types de véhicules légers à partir du 1er avril 2018.

Par ailleurs, le déploiement volontaire du eCall par les constructeurs a conduit à une translation vers des applications supplémentaires comme la mise sous surveillance, le tracking d'un véhicule volé, et autres en cours de développement. Ainsi, les indicateurs de valeur à l'issue du

[33] Combinaison des technologies de géolocalisation et de téléphonie mobile en lien avec le réseau multiplexe du véhicule

[34] Une interface automatique est en cours de développement : l'outil de transmission est testé début 2015 avec les SDIS (Service départemental d'incendie et de secours) et les CODIS (Centre opérationnel départemental d'incendie et de secours)

[35] Norme NF EN 16102 avril 2012

projet restent incomplets car de nombreux usages indirects n'ont pas fait l'objet d'une estimation particulière.

Médiation Homme-Données :

Ce dispositif a modifié le périmètre de l'activité d'assistance : le nombre d'appels comprend non seulement les appels décidés par le passager, mais aussi des appels déclenchés sans action de sa part, ce qui augmente les probabilités d'alerte pour les accidents avérés, mais inclut aussi des cas d'accident bénins. Chaque alerte est traitée comme une hypothèse d'accident grave. Un processus de traitement, structuré, permet de résoudre partiellement ce biais par le rejet d'hypothèse (le questionnement lors de la conversation avec l'habitacle permet de filtrer entre 85 et 90% de fausses alertes : absence de situation d'urgence ou erreurs de manipulation) et de qualifier chaque cas. Dans ce cadre, l'expertise des chargés d'assistance est fortement sollicitée avec les nouveaux paramètres : leurs habitudes, au sens humien, s'enrichissent régulièrement par des stimulations visuelles (priorisation selon l'interface « notificateur », cartes géographiques à interpréter comme indiquant une zone à risque…) et auditives (environnement du véhicule, rires d'enfant, bruit de circulation rapide, gémissements…). Ces perceptions permettent de construire un raisonnement abductif (émergence de nouvelles hypothèses d'accidents ou de dysfonctionnements), puis inductif, parfois à travers des échanges informels.

Par exemple, un chargé d'assistance a partagé l'expérience de capitalisation de connaissances suivante. Un véhicule déclenche une alarme perçue comme grave au cours de l'appel (indicateurs mécaniques liés à une casse violente affichés sur l'interface, bruits forts dans le véhicule et pas de réponse des passagers au cours de la procédure de prise d'appel...). Or, le chargé d'assistance a un doute : la géolocalisation de l'accident n'indique pas une route perçue comme particulièrement dangereuse, associée à ce type d'accidents violents. Une fois la procédure d'assistance terminée par un appel des pompiers, le chargé d'assistance échange avec ses collaborateurs et comprend que ce cas n'est pas isolé. L'analyse par l'équipe de la situation, basée sur les perceptions et la localisation des accidents, a généré une hypothèse, confirmée quelques heures plus tard : il s'agissait d'un centre d'essais de véhicule qui omettait d'éteindre le boitier télématique. Des situations similaires avaient émergé auparavant pour les centres VHU (véhicules partis à la casse sans éteindre le boitier) et autres : ces hypothèses métier ont pu partiellement être intégrés au cours des différentes évolutions du dispositif technique, de l'interface, ou de la procédure de prise d'appel.

A l'heure de l'observation, la validation formelle de l'hypothèse d'accident n'était pas indispensable, le transfert aux secours étant systématiquement appliqué en cas de doute, comme lors d'appels sans réponse (« silent call »). Cette prise de décision est conditionnée par le temps de traitement et encadrée par un ensemble de procédures pour limiter notamment la responsabilité des chargés d'assistance. Leurs perceptions complètent alors les données fournies par le dispositif et sont transmises aux secours pour rendre leur intervention plus efficiente.

1.2.2 Cas B : Cancer du sein triple négatif

1.2.2.1 Contexte et enjeux

Le centre Jean Perrin est un établissement de soins privé d'intérêt collectif financé par l'Etat et les dons pour contribuer à la lutte contre le cancer. L'activité du centre comprend des études cliniques financées par des entreprises privées (industries pharmaceutiques, biotechnologies…) sur leurs traitements, par exemple contre le cancer du sein triple négatif. Représentant 15% des tumeurs mammaires, ce cancer est particulièrement dangereux étant donné sa virulence et sa rapidité de propagation dans d'autres parties du corps[36]. Son nom provient de l'absence de trois types de récepteurs : il ne réagit ni à l'hormonothérapie ciblant les œstrogènes et la progestérone, ni aux traitements spécifiques ciblant les facteurs de croissance épidermiques humains HER2. Le centre Jean Perrin a réalisé une étude sur 47 patientes (données cliniques) pour identifier des marqueurs pour la prédiction de réponse à un traitement utilisé à ce jour (Panitumumab plus Fec 100, suivi de Docétaxel).

1.2.2.2 Synthèse des résultats

La performance, ou réponse au traitement, est traduite par sa classification de 1 à 4 sur l'échelle de Chevalier : un traitement « efficace » est alors défini comme correspondant aux indices 1 et 2, soit une absence de métastases, et concerne 46,8% de la population. Chaque profil correspond à une déclinaison de 250 variables, soit les informations sur les patientes, sur la tumeur, les modalités d'admission du traitement, les marqueurs biologiques et génétiques et autres (pour plus de détails illustratifs, voir Annexe 6 - Présentation du rapport préliminaire). Les premières analyses statistiques basées sur une hiérarchisation des variables permettent d'identifier le

[36] http://www.cbcf.org/fr-fr/central/YourDollarAtWork/ResearchSavesLives/Pages/RSL-Triple-negativeBreastCancer.aspx

niveau d'expression du marqueur de lymphocytes CD8+ comme prédictif d'un traitement efficace, avec 84% de guérison pour les 19 patientes correspondant au profil déterminé. Les cliniciens font appel à Quinten pour confirmer cette analyse et identifier des combinaisons de variables supplémentaires en abandonnant la hiérarchisation *a priori*. Le risque de fausse découverte est alors limité par la restriction de l'analyse à une complexité de 2 variables, ainsi que par l'évaluation de la robustesse des résultats (p-value hypergéométrique et intervalle de confiance).

La méthode mise en œuvre permet de faire 5 observations, ou « règles », présentées sous forme d'arbres de probabilités avec des candidats de marqueurs biologiques discriminants. Deux règles sont simples : elles permettent de confirmer la pertinence du marqueur CD8+ et identifier un autre marqueur correspondant à une sous-catégorie de CD8+. Les trois autres règles sont des corrélations de deux variables avec les seuils d'expression correspondants : CK8/18 et EGFR[37], CK8/18 et un autre marqueur corrélé à l'EGFR, ainsi que CK8/18 et Ki67. La première attire particulièrement l'attention des cliniciens avec un taux de guérison de 86% pour 14 patientes. En effet, le marqueur des cellules lumino-basales CK8/18 n'a pas été identifié comme premier nœud d'analyse, et le traitement serait anti-EGFR. Cette découverte statistique a été ainsi effectuée sans expertise médicale particulière, mais si la corrélation semble fournir une « règle » prédictive robuste, aucun lien de causalité n'est mis en évidence. Une cohorte de confirmation de ces résultats statistiques a été lancée, avec une recherche de financement à l'heure de l'observation, et les cliniciens ont émis l'hypothèse que cette corrélation est causalement liée à la nature du traitement.

1.2.2.3 Observations clés

Indicateurs de valeur :

La valorisation de la génération de ces nouvelles hypothèses médicales est difficile à réaliser en absence de cadre économique classique. Cependant, les indicateurs statistiques permettent d'avoir les premiers ordres de grandeur de l'efficacité du traitement, et des coûts éventuels d'analyses à réaliser pour les patients pour confirmer la pertinence du traitement.

[37] Protéine EGFR : Epithelial Growth Factor Receptor

Médiation Homme-Données :

Ce cas montre d'une part la nécessité d'une prise en compte de connaissances métier dès le démarrage du projet (par exemple, pour l'établissement de l'indicateur de l'efficacité du traitement avec l'échelle de Chevalier) ainsi qu'au cours de l'interprétation des résultats (traduction des variables comme EGFR), et d'autre part le potentiel de génération de nouvelles connaissances métier à travers ce type de projets. Il démontre par ailleurs que les connaissances générées sont insuffisantes en soi, et nécessitent une phase d'appropriation et de validation, en termes statistiques et de sens (confirmation de causalité nécessaire à travers la recherche fondamentale).

1.2.3 Cas C : Placement Publicitaire

1.2.3.1 Contexte et enjeux

Le secteur Antenne M6 du groupe audiovisuel et multimédia français repose sur un modèle économique financé entièrement par la publicité[38]. La régie M6 Publicité[39] gère la commercialisation des espaces publicitaires des chaînes du groupe auprès des agences media : son enjeu consiste à augmenter la part de M6 dans la répartition du budget des campagnes publicitaires des annonceurs. Le pôle études de la régie effectue annuellement des analyses de tendances basées sur des données internes M6, comme les détails qualitatifs et quantitatifs des campagnes publicitaires diffusées, et des données acquises auprès d'acteurs externes, comme Kantar qui effectue des mesures de pression publicitaire agrégées par campagne, basées sur des panels représentatifs de foyers.

1.2.3.2 Synthèse des résultats

La performance d'une campagne publicitaire est caractérisée par le volume de ventes, le recrutement de clients et la fidélisation, dont les indicateurs, seuils et degré de risque sont challengés et convenus en amont entre les acteurs métier et les Data Scientists. L'analyse de l'efficacité publicitaire consiste à mettre en perspective ces performances avec les combinaisons de variables à partir de l'ensemble des données disponibles, soit plus d'une cinquantaine de variables hétérogènes sur un historique de trois ans pour un millier de campagnes publicitaires de produits de grande consommation (pour plus de détails illustratifs,

[38] Document de Référence 2013 incluant le rapport annuel du groupe M6 (http://www.groupem6.fr)
[39] http://m6pub.fr/qui-sommes-nous/la-regie/

et notamment l'interface homme-machine mobilisée dans le dispositif, voir Annexe 6 - Présentation du rapport préliminaire). Il s'agit de déclinaisons de l'indice de pression GRP[40], de données de media planning (répartition du budget publicitaire, couverture, nombre de spots publicitaires par campagne, position du spot dans l'enchainement, part du jour, durée de l'écran…), de catégories de produits, de données qualitatives du discours publicitaire, mais aussi de météo et autres. Face à la quantité de variables, l'analyse s'appuie sur le Data Mining et l'algorithme Q-Finder pour l'analyse combinatoire. La première année du projet est ainsi essentiellement marquée par le cadrage du besoin à travers la formalisation des connaissances métier, la conception des bases de données (structuration, retraitement, calcul de variables agrégées et dérivées, ajout de variables nouvelles…). Ce cadrage permet de faire les premières observations, ou « règles » corrélatives, parmi lesquelles les experts métier sélectionnent des leviers, dont des leviers opérationnels. La suite du projet est une alternance entre optimisation de la visualisation et analyse des résultats de croisement de variables avec les seuils de valeurs maximisant les critères d'efficacité des campagnes publicitaires.

Cette construction s'accompagne par une capitalisation de connaissances et une appropriation progressive des paramètres statistiques par les métiers utilisateurs de l'interface. En 2015, 70 contextes sont regroupés en 9 types, et recoupées pour proposer des hypothèses opérationnellement intéressantes et pour répondre aux spécificités d'une campagne selon 5 indicateurs détaillés. La mise en place de l'outil de visualisation « M6 Advisor » permet aux pôles études et marketing de la régie de consulter ces résultats lors de la négociation avec une agence pour argumenter une optimisation de la performance d'une campagne publicitaire, se positionnant ainsi comme conseiller de l'agence media. Cela constitue un avantage concurrentiel pour gagner des parts de marché publicitaire, indicateur clé stratégique du groupe M6. L'outil est par ailleurs adapté pour vérifier statistiquement les hypothèses des experts métier, et sa construction a donné lieu à des découvertes de corrélations parfois contre-intuitives, appelées « pépites ». La recherche de causalité et la validation des hypothèses par l'expérience ne sont pas considérées comme indispensables pour la prise de décision. Après trois ans de déploiement, M6 Publicité fait face au besoin de former l'ensemble des utilisateurs de l'interface pour maintenir les connaissances en termes de modalités de construction de résultats pour la prise de décision.

[40] Gross rating point, point de couverture brute représentant un indice de pression publicitaire sur une cible définie, soit le taux de couverture fois le taux de répétition moyen

1.2.3.3 Observations clés

Indicateurs de valeur :

Les indicateurs de performance objectivés par l'étude sont plutôt cohérents avec les standards du marché (GRP, volumes de vente, recrutement, fidélisation…), mais l'impact de l'utilisation de contextes de surperformance permet bien à M6 de proposer à ces clients un service de ciblage complémentaire qui contribue à l'avantage compétitif. Le nouveau service proposé semble d'ailleurs non seulement optimiser l'usage historique, mais d'en créer un nouveau : en effet, la régie et le marketing basculent dans un business modèle de conseil pour les agences media. Le bénéfice final généré, sous forme de parts de marché, n'est pas disponible à l'heure de l'observation.

Médiation Homme-Données :

Une fois de plus, le cas confirme l'importance de la prise en compte des connaissances métier en amont du projet, qui s'avère ici assez longue, et la nécessité de s'approprier de nouvelles hypothèses de travail sous forme de leviers. Par ailleurs, l'importance de l'interface et de la compréhension des modalités de construction des résultats, c'est-à-dire le traitement algorithmique, est mise en évidence, bien que cette prise de conscience n'impacte le projet que plusieurs années après le démarrage. Malgré cette volonté de capitalisation de connaissances, la prise de décision prime à travers l'exploitation des résultats.

1.3 Cas réalisés et détaillés

L'ensemble des cas suivants ont été réalisés en tant que l'un des « Data Scientists » de l'équipe projet. En effet, tous les collaborateurs de la société Quinten intervenant sur les projets portaient entre 2015 et 2017 ce titre, ce qui globalement séparait l'équipe projet complète entre les « Data Scientists » et les « Experts métier » issus de l'entreprise cliente. Or, les compétences réellement mobilisées, dans les deux cas, ont présenté une variabilité forte non seulement entre les projets, mais aussi au cours de chaque projet : une grande porosité des compétences a été globalement observée. Ainsi, chaque étude de cas présente plus en détails le dispositif humain et décrit les tâches réalisées par chacun, et une discussion sur les compétences et sur les titres plus génériques n'est soulevée que dans le chapitre final de ces travaux de recherche afin de rendre cette mobilisation de compétences plus transposable.

Tous les cas sont formalisés selon une structure et rédigés de façon narrative et illustrée afin de privilégier leur compréhension et la mise en perspective du processus projet observé.

1.3.1 Cas 1 : Attrition en assurance santé

1.3.1.1 Contexte et enjeux

Le marché des contrats d'assurance en santé individuelle en France croissait de 1 point tous les deux ans, et l'attrition, c'est-à-dire la part des contrats résiliés pour un assureur, se situait en 2013 entre 18 et 20%. Au sein du groupe mutualiste AAA, la branche Santé réalisait 1 milliard d'euros de chiffre d'affaires par an avec 3,5 millions de contrats individuels en santé pour un total de 11 millions de clients, appelés sociétaires, et n'échappait pas à cette évolution avec un taux d'attrition de 18%. Le groupe souhaitait mettre en place, pour l'une des enseignes, des actions de rétention ciblées et performantes afin de réduire le taux d'attrition en santé individuelle grâce à une activation pertinente des sociétaires à forte probabilité de résiliation de leurs affaires santé.

Le secteur de l'assurance constituait à ce moment l'un des terrains de jeu les plus matures en termes de connaissance des approches analytiques, grâce à son ancrage dans l'expertise en statistiques, détenue historiquement par les actuaires au sein des directions techniques et, plus récemment, par les directions marketing. Tout comme l'analyse de survenance de sinistres, la fraude ou l'appétence à acheter un produit, l'attrition était habituellement abordée grâce aux méthodes de scoring, une méthode d'analyse prédictive résultant dans l'attribution à chaque individu d'un score de risque de résiliation, souvent représenté par une probabilité de résiliation par contrat détenu par un client. Cette méthode s'appuyait sur un grand nombre de techniques prédictives variées qui pouvaient être conventionnelles, comme les modèles de régression (GLM, logistique...), ou plus élaborées, comme l'analyse de survie ou les arbres de décision. A la date de l'intervention, début 2015, le sujet d'attrition avait déjà fait l'objet chez AAA de calculs de scores prédictifs dans les outils CRM habituels, sous la forme d'un score de fragilité d'une affaire, cependant le résultat n'était pas à la hauteur des attentes. En effet, le score d'attrition était calculé par les équipes Data Mining du Marketing et restitué aux agents sur le terrain sous forme d'alertes binaires pour chaque client (pastilles de couleur rouge pour « affaire fragile » et verte pour « affaire non fragile »). Les agents n'ont pas su interpréter et utiliser cet indicateur qui ne leur donnait aucun levier explicite. Il s'agissait de l'effet « boite noire » d'un score dont la formule de calcul était opaque pour l'utilisateur et qui ne permettait pas de pointer

individuellement le contexte de risque de résiliation, c'est-à-dire les facteurs d'attrition propres à chaque affaire. La démarche analytique proposée pour ce projet par Quinten permettait alors de compenser cette difficulté d'appropriation opérationnelle grâce à une approche double, articulant le scoring habituel à une identification des facteurs ou des combinaisons de facteurs annonciateurs d'une résiliation d'affaire pour qualifier, de façon interprétable pour l'utilisateur, des profils de sociétaires à risque.

L'intervention démarrait dans un contexte où AAA avait d'ores et déjà investi dans le déploiement d'un Data Lake, regroupant l'ensemble des données sur les clients, dans un seul et même outil. AAA considérait alors ces données comme rapidement exploitables. Elles contenaient non seulement les caractéristiques du sociétaire, de ses contacts et de l'utilisation de son contrat santé, mais aussi la détention d'autres contrats (auto, habitation...). Cette détention transversale n'avait jamais été intégrée dans un score propre à une affaire santé. Le Data Lake semblait ainsi ouvrir une opportunité d'exploitation de données riches, rare en 2015 dans les institutions d'assurance organisées en silos par type de produit.

Le projet était porté en interne par la Direction Santé avec la participation de contributeurs issus de la Relation Client, du Marketing et de la Direction des Systèmes d'Information Opérationnels : chaque contributeur possédait une expertise élevée dans son périmètre d'activité et aucun n'avait été confronté à un projet data (de type Data Science) antérieur. Ce projet, mené avec les équipes de Data Scientists de Quinten, était alors vu comme un moyen de découvrir non seulement des éléments opérationnels pour l'attrition, mais aussi les méthodes de travail sur les données.

1.3.1.2 Synthèse des résultats

Suite à la collecte, à l'audit et à la structuration de plus de 500 variables, plus riches que des variables utilisées pour le scoring initial qui ne prenait en compte aucune donnée de sinistre en santé, et à l'application d'un ensemble de méthodes analytiques prescriptives et prédictives, le projet a permis d'établir 6 familles de profils de clients à risque d'attrition interprétables, et de nouveaux scores de risque de résiliation de chaque contrat santé. Les connaissances générées sur les profils ont constitué des hypothèses de priorisation de travail, et ont été utilisées par le Marketing, la Relation Client et la Direction Santé pour concevoir des leviers de réduction du taux de résiliation de contrats santé. Cela a guidé la construction de la feuille de route stratégique de la Relation Client de l'enseigne, comprenant 14 propositions d'actions

directement issues des 18 recommandations basées sur les résultats du projet (ces propositions d'action et recommandations restent confidentielles, et comprennent des campagnes de rétention personnalisées et des services nouveaux destinés aux clients à risque). Les résultats ont confirmé qu'il était possible de fournir aux agents locaux des facteurs explicites d'attrition pour chaque contrat santé, au-delà d'une pastille verte ou rouge selon le score de risque d'attrition, et les plans d'actions correspondants à chaque profil de risque. Par ailleurs, la découverte de certains profils a permis de choisir des leviers plus pertinents que l'usage terrain initialement imaginé, notamment des usages plus amont par la Direction commerciale, avec une élaboration de processus de prise en charge systématique de certains clients spécifiques, comme les plus récents. Ces leviers visaient alors une personnalisation de la relation avec les clients, et à terme la baisse de l'attrition et une économie de moyens grâce à un choix plus pertinents de leviers face à chaque facteur de risque.

1.3.1.3 Observations clés

Indicateurs de valeur :

Bien que la mise en œuvre des leviers et la mesure de leur performance n'aient pas été suivies à l'issue du projet, un ensemble d'éléments de mesure ont été prédéfinis à l'issue de l'intervention. Tout d'abord, la génération d'hypothèses de travail pour la Direction de la Relation Client, le Marketing et la Direction Santé devaient constituer un gain de temps en termes d'investigation, et diminuer le risque d'investigation infructueuse grâce à l'emploi de méthodes analytiques qui garantissent un balayage exhaustif des contextes de risques d'attrition observables avec les données existantes. La richesse de ces hypothèses évitait ainsi un travail d'investigation futur, et la documentation des données allégeait la poursuite du test d'hypothèses métier. Ensuite, la qualité des leviers proposés était jugée supérieure, que ce soit en termes de sens interprétable (facteurs explicites au lieu d'un score « boite noire ») ou de pertinence (les profils identifiés étaient à risque d'attrition beaucoup plus élevé que la moyenne). Enfin, la revue de l'indicateur même d'attrition, précisé comme un composé d'attrition volontaire (comme le départ à la concurrence, actionnable) et involontaire (comme le décès du client, n'ayant rien d'opérationnel), et applicable à des segments de clients précis et non pas à la population globale, a enrichi le pilotage de l'efficacité des leviers opérationnels. Pour finir, la capitalisation, par l'équipe projet et plus largement par AAA, a dégagé un potentiel de gains de performance par l'innovation liée aux futurs projets data.

Un retour d'expérience, réalisé par l'ensemble des contributeurs projet auprès d'un cercle large d'acteurs de AAA, a consolidé les apports principaux du projet de la façon suivante :

- Les connaissances produites étaient bien innovantes et stratégiques pour l'entreprise, et reproductibles sur de nouveaux périmètres (contrats auto, habitation...), même si leur impact financier direct restait difficilement mesurable à l'issue du projet.

- L'étape suivante souhaitée sur les données était l'automatisation de la structuration d'une base de données interne, selon la documentation produite au cours du projet, que ce soit pour la mise en œuvre des résultats directs ou pour la poursuite d'analyses sur d'autres problématiques. L'un des usages indirects du projet, issu de la capitalisation, a été la mise en place d'un référentiel de données harmonisé permettant de piloter l'activité non pas à la maille d'une affaire, mais d'un client, et du foyer du client. Ces éléments ont contribué à poser le socle de la Vision Client 360°.

- Le besoin d'une gouvernance transversale était mis en évidence, avec une implication amont indispensable de l'ensemble des contributeurs afin d'assurer l'adhésion et d'anticiper l'appropriation des résultats.

- La transmission de connaissances (concepts métier et concepts data, ainsi que la méthodologie du projet), la mise en place d'une gestion de projet dynamique avec traçabilité des échanges et des décisions (suivi hebdomadaire), et la méthode de capitalisation et de documentation des données sous la forme d'un Databook étaient vues comme des facteurs d'accélération et de qualité significatifs pour ce type de projets, reproductibles à l'échelle de l'entreprise, pour d'autres projets data.

Qualité des données :

Liste des données explorées : caractéristiques souscripteur et bénéficiaires, caractéristiques contrat santé, sinistres et prestations santé, contacts marketing, contacts commerciaux, dates et motifs de résiliation, détention d'autres contrats que la santé…

Le projet a donné lieu à une documentation précise des données mobilisées et générées au cours du travail analytique, sous forme d'un Databook, et a permis de mieux partager le sens de certaines données qui n'étaient pas comprises de la même façon par les différents contributeurs en amont du projet. L'absence imprévue de cet alignement initial sur le sens des données,

notamment contenues dans un nouveau Data Lake, a en effet provoqué un dépassement considérable des ressources du projet. Cette capitalisation a alimenté une feuille de route plus globale de la gestion de la qualité des données.

Médiation Homme-Données :

Liste des acteurs du projet : Directeur Santé, Représentant Marketing, Représentant Produits Santé, Représentant DSI (compétence de Data Stewart et Ingénierie Data), 6 Data Scientists de Quinten (dont 2 dédiés à la structuration des données (Ingénierie data), 2 spécialisés en Machine Learning, 1 profil généraliste et 1 Manager de projet ayant des connaissances métier).

Le projet a généré une montée en maturité significative des contributeurs sur les méthodes projet data, ce qui promettait une diminution des incertitudes pour les projets futurs, jugés utiles à mettre en œuvre sur des sujets en dehors de l'attrition en santé par des acteurs à qui les résultats ont été présentés au cours d'un retour d'expérience. Cette montée en maturité a été favorisé par des instances d'échange réguliers entre les différentes parties prenantes, et la documentation de l'ensemble des arbitrages réalisés au cours de ces instances. En effet, ce projet démontre l'utilité de la mise en place (tardive) de ces instantes.

Par ailleurs, les interfaces permettant l'appropriation des résultats, sous la forme de maquettes Excel complétées de présentations PowerPoint, se sont avérées très utiles, mais tout à fait perfectibles : en effet, les ateliers d'appropriations ont été laborieux et peu intuitifs pour les experts métier en absence de possibilité d'interagir plus directement avec les résultats pour sélectionner et préciser les leviers opérationnels, ce qui a présenté une charge complémentaire par rapport au budget initialement prévu.

1.3.1.4 Compte rendu du projet

Un lancement de projet marqué par des incertitudes mal identifiées.

Le projet a démarré dans un contexte d'incertitudes jugées faibles. Les contributeurs « métier », représentant le Marketing, la Direction Santé et la Relation client, affirmaient que les objectifs stratégiques du projet étaient cohérents avec la volonté du groupe à réduire l'attrition, et que les incertitudes opérationnelles étaient réduites : avant d'envisager les leviers, il s'agissait essentiellement de générer des connaissances sur les facteurs de résiliation des sociétaires ayant

un contrat santé à horizon de 3 et de 6 mois. Un contributeur expert en données internes de la DSI de AAA était intégré au projet, et estimait l'incertitude quant à l'exploitabilité des données faible suite à la mise en place du Data Lake. En effet, le Data Lake en place avait été alimenté avec les principales données de la société, et devait justement servir pour leur analyse. Enfin, les Data Scientists jugeaient la méthode d'analyse des données fiable : les scores précédents étaient d'ores et déjà statistiquement évalués comme suffisants, ce qui prouvait la faisabilité, et la méthode d'analyse de profils de clients à risque avait fait ses preuves sur des cas *a priori* transposables, comme l'analyse de risque de décès dans le cadre d'études cliniques. L'équipe projet a alors entamé directement la phase d'identification et de collecte des bases de données appropriées pour les analyses, en commençant par l'extraction d'échantillons des bases de données souscripteur et sinistres santé. L'exploration des sources existantes a fait alors émerger un ensemble de contraintes qui ont nécessité la remise en cause de la démarche et la recherche de solutions inédites.

La première contrainte majeure tenait au fait que les bases du Data Lake n'étaient pas conçues initialement pour une exploitation algorithmique sous l'angle de la prévision de l'attrition. Tout d'abord, les données sur les sociétaires, récoltées habituellement par AAA, ne convenaient pas à l'analyse temporelle, car il s'agissait de caractéristiques des sociétaires (ancienneté, détention de contrats, ville, nombre de bénéficiaires dans le foyer...) saisies à une date donnée pour remplacer les caractéristiques précédentes, ce qui ne permettaient pas de suivre l'évolution de ces caractéristiques dans le temps. Or, l'évolution de certaines caractéristiques était pressentie par les experts métier comme étant potentiellement des facteurs clés du phénomène d'attrition. Par exemple, le nombre d'enfants était pressenti comme un facteur moins important qu'une nouvelle naissance. Ainsi, la base existante ne permettait pas de détecter des évènements marquants chez le souscripteur, mais seulement ses caractéristiques actuelles ou au moment de la résiliation : toute notion de parcours de vie du souscripteur devait être éliminée du périmètre d'analyse de facteurs, au détriment de la qualité des résultats.

Deuxièmement, les données de sinistres, c'est-à-dire les consommations de prestations en santé par le souscripteur, n'ont jamais été récoltées dans le cadre de projets marketing : l'exploration des données a rapidement montré qu'il était impossible d'attribuer une prestation santé à un individu bénéficiaire en absence d'une clé directement exploitable entre les caractéristiques du souscripteur et ses consommations santé. En effet, une prestation santé était attachée à un contrat et non pas à un bénéficiaire : elle pouvait alors être consommée par le souscripteur lui-

même, par son conjoint ou par son enfant. Là encore, une détérioration potentielle de la qualité des résultats était mise en évidence.

La troisième contrainte majeure consistait dans le fait que le concept même de « résiliation » n'était pas précisément défini dans les données exploitables, et qu'il n'existait pas de consensus sur la qualification de l'attrition au sein de l'entreprise. En effet, l'attrition était vue selon l'interlocuteur comme une absence prolongée de consommation, comme une réception de lettre de demande de résiliation, ou comme une résiliation juridique du contrat (seul élément présent dans les données sous forme de date de fin de couverture du contrat). Cela posait un problème d'interprétation conséquent, et empêchait la définition de l'indicateur clé à analyser.

L'identification de ces trois contraintes a eu lieu selon un processus d'extraction des données dans le Data Lake par l'expert data, d'anonymisation (censure de l'ensemble des éléments relatifs aux individus comme les noms et les adresses, ainsi que recodification de l'ensemble des numéros de contrats et des numéros de sinistres), de transfert par portail sécurisé aux équipes de Data Scientists avec les explications générales sur le contenu des fichiers transmis, et d'exploration par les Data Scientists aidée par des échanges mail avec les interlocuteurs métier. A ce stade, les données du Data Lake, éparses et volumineuses, ont suscité un temps d'extraction et de transfert plus long que prévu auprès de l'expert data de AAA, et il semblait de plus en plus pertinent d'extraire les données directement à partir des outils source et non pas depuis le Data Lake. Mais surtout, l'échange lié à l'exploration s'enlisait : les questions ont été soumises sous forme de questionnaire aux différents experts métier, ce qui a donné lieu à des retours contradictoires entre interlocuteurs métier impliqués. Le format d'échanges n'était pas non plus efficace pour faire converger les experts métier avec l'expert data sur ces définitions, notamment la date de résiliation. Les retours contradictoires sur les questionnaires ont par ailleurs fait émerger des questionnements complémentaires sur de nouveaux sujets, et le questionnaire s'allongeait rapidement. Suite à trois itérations par mail sur le questionnaire, plusieurs extractions chronophages de données inexploitables, et face à la multiplication des incohérences entre les définitions fournies et les données envoyées ainsi qu'à la difficulté du management à valider les concepts étudiés, l'équipe de Data Scientists ont tiré une alerte d'enlisement du projet, et les itérations sous cette forme ont été arrêtées. La structure de données existantes n'étant plus considérée comme appropriée aux analyses prévues à l'origine du projet, il a été indispensable de revenir à la phase de cadrage métier et à la redéfinition de la stratégie d'analyse.

La revue du dispositif projet et l'intégration de la médiation.

Cette revue du cadre du projet a permis d'ajuster rapidement les ressources dédiées au projet : en effet, si les données étaient considérées comme directement exploitables par AAA, une phase de structuration *ad hoc* pour le projet est apparue comme nécessaire en amont des analyses prévues, et l'animation d'une instance de médiation plus pertinente qu'un échange de questionnaires par mail est apparue nécessaire afin de mettre en œuvre une redéfinition progressive du cadre, au fur et à mesure de la découverte des nouvelles contraintes. Cette médiation, sous la forme d'un dispositif de comités de projet hebdomadaires, visait alors à établir une compréhension claire entre les données explorées et le sens métier, et à piloter les ressources du projet grâce à une meilleure anticipation des incertitudes. Enfin l'équipe de Data Scientists, composée de 3 personnes sans différenciation de rôles, a été revue avec une distinction entre un ingénieur data spécialisé dans les langages permettant de traiter des données volumineuses (notamment le langage Python sur Spark), deux Machine Learners (le premier dédié aux analyses de facteurs de risque d'attrition, spécialiste de l'algorithme Q-Finder, et le second, junior, dédié au scoring), et un médiateur capable d'établir la convergence entre l'ensemble des membres de l'équipe (moi-même).

La mise en œuvre de ce nouveau dispositif a permis un processus de convergence progressive et itérative. Les sujets ont été regroupés en concepts, correspondant à une réalité métier transformable en données. Parmi les concepts étudiés se trouvaient par exemple l'objet de résiliation (client, foyer, affaire santé, ensemble des contrats d'assurance du client, population concernée par l'analyse...), le périmètre historique de résiliation (historique à prendre en compte, période de référence, date de résiliation…), les caractéristiques des contrats santé (garanties, familles de garanties, modalités contractuelles, sens des risques couverts...), les contacts (campagnes marketing, appels, courriers, mails…) ou bien le motif de résiliation (suite à un contentieux avec l'assureur, suite à un décès, attrition volontaire, motif inconnu…). Chaque concept a fait alors l'objet d'une étude de cohérence séparée donnant lieu à une définition précise et partagée par l'ensemble des interlocuteurs. L'attrition, par exemple, a été limitée au fur et à mesure aux seules résiliations volontaires et opérationnellement actionnables (élimination des décès, des contrats contentieux, des résiliations à l'initiative de la mutuelle...), soit une attrition opérationnelle de 13.8% au lieu de 18% d'attrition totale.

Chaque étude de concept était composée de 3 étapes, ponctuées de 3 ateliers de travail tripartites entre les experts métier (Marketing, Directeur Santé, utilisateurs des données, management,

acteur terrain...), les experts data (gestionnaire de référentiel, expert IT et autres acteurs, mobilisés selon la base de données analysée) et les Data Scientists :

- **Atelier de priorisation** : identification des concepts à traiter à court terme, priorisation selon l'impact espéré et les risques à mitiger, puis validation en atelier des priorités, des délais de préparation, et implication des contributeurs clés.

- **Préparation des sujets prioritaires** : Analyse des données fournies, préparation des questions pour les acteurs métier, mise en évidence des biais (par exemple, démonstration des limites d'une saisie déclarative des motifs de résiliation, avec un référentiel de motifs redondant et mal renseigné) mais aussi proposition de solutions et des arguments pour la prise de décision, puis arbitrages en atelier et rédaction de compte rendu, contenant la liste des actions décidées et écartées, et les délais de mise en œuvre. En cas d'incapacité à arbitrer, une nouvelle priorisation était réalisée pour la préparation de l'arbitrage suivant, avec renvoi des données en cas de besoin.

- **Consolidation et Capitalisation** : Pilotage de la mise en œuvre des actions décidées, suivi des délais de réalisation, et validation en atelier de la conformité des réalisations par rapport aux décisions prises et aux attentes, ainsi que capitalisation des connaissances générées.

Ces études, superposées dans le temps (c'est-à-dire « tuilées » en fonction des ressources nécessaires pour le traitement des concepts), ont donné lieu à une convergence de l'interprétation par l'ensemble des acteurs réunis, à une fixation des concepts jugés comme utiles et opérationnels pour l'analyse, à une mise à l'écart des concepts inutiles ou non opérationnels, et à l'établissement d'une liste de points d'optimisation non prioritaires. Par ailleurs, certains usages ont été identifiés comme d'ores et déjà opérationnels, sans attendre la restitution des résultats : par exemple, la revue du processus de saisie des motifs de résiliation a donné lieu à la conception d'un référentiel harmonisé, fermé et obligatoire, facile à déployer dans les systèmes existants. Cet usage, identifié comme générateur de valeur, a été mis en œuvre en parallèle de l'avancement du projet.

Ce processus de convergence a abouti à des définitions de concepts suffisantes, bien que non définitives et perfectibles : un certain degré d'incertitude a été accepté par tous les contributeurs lorsque son traitement n'était pas indispensable. Par exemple, la date de demande de résiliation

(envoi du courrier de résiliation par le souscripteur) aurait été préférable à la date d'effet de la résiliation, car un levier de rétention n'est plus activable une fois que la résiliation est demandée par le souscripteur. Or, la date de demande de résiliation n'était pas exploitable dans les données existantes : sa mise en qualité a été jugée trop couteuse à court terme par rapport à l'apport opérationnel espéré. Malgré l'incomplétude de ces définitions, elles étaient parfaitement partagées par l'ensemble de l'équipe projet, les biais mis en évidence étant contournés ou acceptés.

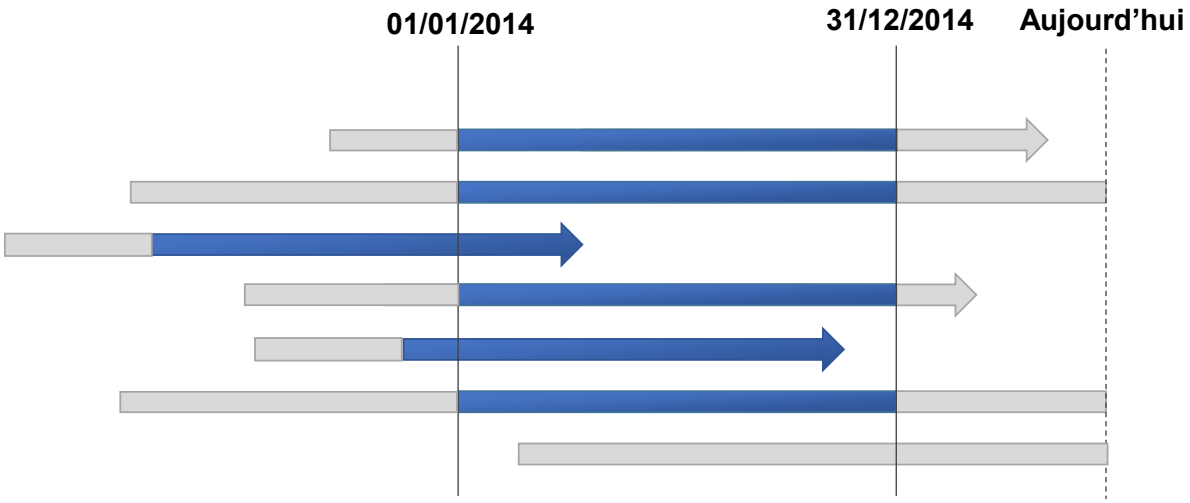
La particularité de ce processus tripartite a consisté à impliquer fortement les interlocuteurs métier dans la phase amont du projet, tout en avançant sur les chantiers d'analyse. Elle s'est avérée plus fructueuse par rapport à la démarche initiale de compréhension des données par les Data Scientists sans implication amont des métiers, consultés seulement sous forme de questionnaires. L'avantage de ce processus était de concentrer l'attention des métiers en amont des résultats au lieu de transférer cette charge en aval lors de leur interprétation, s'assurant ainsi de la justesse opérationnelle des résultats et de l'adhésion des experts dès l'origine.

Par exemple, la contrainte concernant la prise en compte de la temporalité de résiliation a nécessité une recherche approfondie de modèle approprié à cause des particularités suivantes :

- L'attrition est un phénomène saisonnier (un tiers des résiliations a lieu en janvier)
- Les leviers de prévention de l'attrition doivent prendre en compte non pas la date de résiliation, mais la date de demande de résiliation, absente dans les bases de données
- La présence irrégulière et limitée dans le temps des campagnes marketing et des actions commerciales induit des biais temporels (par exemple, les campagnes d'automne sont plus nombreuses et plus proches de la vague des résiliations en janvier que les campagnes de printemps : il s'agit d'un biais temporel à prendre en compte pour ne pas déduire à tort que les caractéristiques d'une campagne d'automne sont moins efficaces, voire « provoquent une résiliation plus rapide »)

Ces particularités, qui n'ont été que partiellement mises en évidence dans des publications de recherche en attrition en assurance (Luyang Fu & Wang, 2014), ont nécessité la recherche d'un modèle qui puisse prendre en compte l'ensemble des biais temporels induits dans ce contexte précis. Ainsi, l'anticipation de la résiliation à 3 et 6 mois, prévue initialement, a dû être écartée au profit d'un risque d'attrition immédiate, à l'issue d'un choix de solution équilibrée entre sa

complexité et sa pertinence opérationnelle, en particulier étant donné l'impossibilité opérationnelle d'anticipation sans la présence de la date de demande de résiliation dans les bases analysées. Le périmètre temporel choisi était ainsi une analyse des contrats actifs ou résiliés au cours de 2014 (c'est-à-dire une saison complète, la plus récente parmi les saisons complètes à la date de la réalisation du projet) avec un historique de prestations santé et de contacts marketing agrégé sur l'année en question. Il n'agissait pas d'une simple application ou d'un paramétrage d'algorithmes, mais bien d'une contextualisation de stratégie analytique en fonction de l'usage et de l'exploitabilité des données.



Extrait 1 ***Illustration du périmètre temporel, issue d'un comité projet*** *: pour chaque contrat santé, représenté par une flèche qui correspond à sa période de couverture, seules les informations concernant les périodes représentées en bleu sont prises en compte dans l'analyse afin d'éviter au maximum les biais propres au contexte de AAA.*

L'un des sujets traités au cours de cette phase concernait la sensibilité des données traitées dans le cadre de cette analyse (consommations santé individuelles), qui a nécessité la mise en place d'un processus d'anonymisation issu d'une recherche de méthodologie appropriée pour le volume de données en question. En effet, les méthodes d'anonymisation, nativement présentes dans Oracle, outil d'extraction de données en place chez AAA et utilisées dans la phase initiale du projet, ne permettaient pas d'anonymiser de façon exploitable le volume de données requis : il s'agit de la fonction de hachage. Cette méthode, incluse dans la fonction de base ora_hash, est utilisée habituellement pour crypter les fichiers en calculant, à partir d'une donnée d'entrée comme un identifiant de contrat ou un nom de famille, une empreinte de type « ao958al245kh60rt » non interprétable. Elle résulte en un code de 16 clés, c'est-à-dire 16

lettres ou chiffres déterminés au hasard, et peut hacher efficacement un volume de 2^16 = 65 536 enregistrements. Au-delà de ce nombre d'enregistrements, la probabilité que cette fonction de chiffrage automatique attribue le même code à deux enregistrements distincts est significative : il s'agit de collisions. Cela conduit à récupérer en sortie des codes de contrats en doublon alors que les enregistrements initiaux sont uniques. Or la plus petite base exploitée dans le cadre de ce projet, la base client, contient plus de 300 000 enregistrements. Les données anonymisées fournies ont donc comporté des doublons, ou collisions, c'est-à-dire que 2 clients différents pouvaient porter le même code anonymisé. Le travail de l'ingénieur data a alors consisté à identifier ce problème (en s'apercevant de l'existence des doublons et en analysant le fonctionnement théorique de la fonction de hachage d'Oracle) et à proposer des solutions plus efficaces à AAA. Le choix s'est alors porté vers des méthodes plus appropriées, comme MD5 et SHA256, plus performants pour le hachage de données volumineuses. Cette démarche d'investigation illustre le spectre de compétences techniques d'un ingénieur data spécialiste du traitement de données volumineuses, et l'importance de ces compétences avec la protection croissante des données personnelles au-delà du rôle (inexistant au moment du projet) d'un DPO. Au-delà de cette problématique technique d'efficacité du cryptage, certains sujets ont tout simplement été écartés du projet pour cause de contraintes juridiques, comme la recherche de liens entre le bénéficiaire du foyer et l'assuré-souscripteur. Ces arbitrages ont été rendus possibles au cours des instances de médiation tripartites grâce à la capacité des interlocuteurs à expliquer de façon pédagogique les enjeux liés à leur zone d'expertise (technique, juridique, analytique, terrain, qualité des données…), à les illustrer, notamment par des éléments construits à partir des données fournies, et à permettre les arbitrages en anticipant l'impact d'un problème sur le résultat et sur les ressources nécessaires à la mise en place des solutions alternatives face à ces problèmes.

Structuration et documentation des traitements dans un Databook

La démarche de clarification progressive des concepts a été menée de front avec une structuration au fur et à mesure des données reçues, et a abouti à la conception d'une base de données inédite, comprenant une redéfinition précise et opérationnelle de l'attrition, avec l'agrégation des données anonymisées issues de sources éparpillées et jamais réunies auparavant, enrichies grâce à la dérivation des données, notamment de l'ensemble des contacts et des prestations santé pour les ramener à la maille d'une affaire, sur une maille temporelle

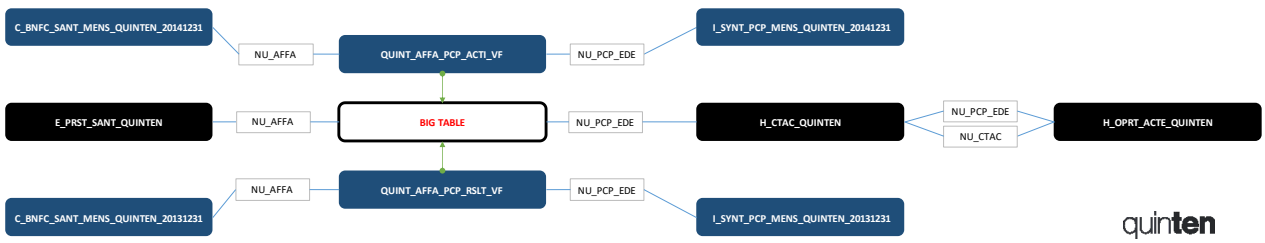
définie de façon spécifique pour le contexte. Cette base de données a été documentée dans un Databook, prototype d'outil de documentation du projet data conçu *ad hoc* au cours du projet, et comprenant les éléments suivants :

- **Description des bases de données initiales**

Les bases envoyées par le client ont été réparties en 12 lots correspondant aux envois groupés, et sont au nombre de 45, soit 1657 variables. En éliminant les tables erronées (mauvaises variables, premières versions contenant les doublons liés au hashcode…), incomplètes (oubli de variables, absence de clés…), inutiles (comme les listes de fiches client, en doublon avec les synthèses client, plus complètes), l'équipe n'a gardé que 9 tables exploitables dans le cadre du projet. Ces 9 tables correspondaient à 691 variables, dont 132 ont été exploitées dans le cadre du projet : il s'agissait de l'ensemble des variables dont l'interprétation avait donné lieu à une convergence tripartite en amont, non techniques (clés en doublon, clés inutiles), et qui ne comportaient pas de biais particuliers impossibles à redresser par une mise en qualité (par exemple, colonnes vides ou avec peu de variabilité). Cette sélection de variables constituait en soi un résultat opérationnel, car il s'agissait d'un dictionnaire de données complet, documenté, et contrôlé du point de vue de la qualité et complétude des données : l'usage de ce dictionnaire donnait dès cet instant la possibilité d'économiser du temps pour la collecte de données pour d'autres projets data.

- **Modèle conceptuel des données**

Une fois les variables sélectionnées, elles ont été agrégées au sein d'une seule et même matrice, à la granularité de l'affaire, grâce à une stratégie d'agrégation conçue *ad hoc* et décrite sous forme de modèle conceptuel simplifié.



Extrait 2 ***Illustration du modèle conceptuel des données****, documenté dans le Databook et présenté en comité projet. La matrice agrégée (BigTable) est ainsi une table, à la maille d'une*

affaire, comportant les caractéristiques du souscripteur, les éléments descriptifs des bénéficiaires, les prestations santé consommées et des contacts avec l'assureur.

- **Matrice finale**

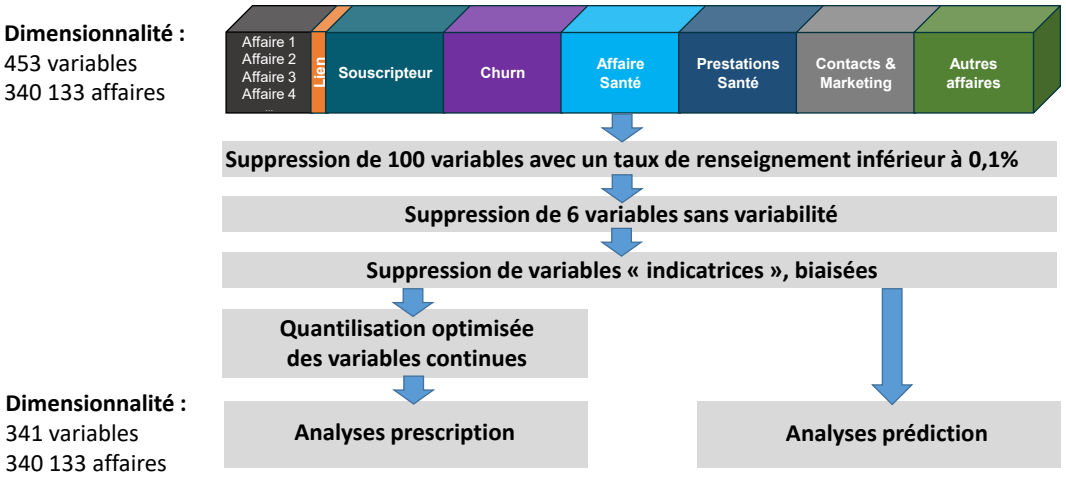
La matrice finale était constituée des 340 133 affaires santé en ligne, et de 453 variables en colonnes. Chaque variable était une dérivation d'une ou plusieurs variables d'origine, sur la maille d'une affaire. Par exemple, une date de création d'affaire permettait de calculer son ancienneté, ou alors le montant du reste à charge était calculé en fonction des montants des frais réels, des remboursements de la complémentaire et du montant des remboursements obligatoires par la sécurité sociale. Le Databook comprend la description détaillée de chaque variable, permettant de les comprendre et de les reconstituer en cas de déploiement.

Ces variables finales étaient réparties en 6 familles :

- Phénomène d'intérêt : l'attrition opérationnelle (« churn »)
- Caractéristiques du souscripteur (code sociodémographique, ancienneté du client...)
- Caractéristiques de l'affaire santé (nombre et qualité des bénéficiaires couverts, cotisations, ancienneté du contrat santé...)
- Caractéristiques des autres affaires détenues (contrats auto...)
- Caractéristiques des prestations santé (montants décomposés des prestations, nature et nombre de garanties activées, risques particuliers avérés...)
- Caractéristiques des contacts et des actions marketing passées entre le souscripteur et l'assureur (nombre des contacts de natures diverses comme les mails, courriers, sms, face à face et autres, nature des campagnes marketing reçues...)

- **Sélection des variables finales pour les analyses**

Certaines variables dérivées ont été créées, puis supprimées selon la démarche ci-dessous :



Extrait 3 ***Illustration de la sélection des variables****, documenté dans le Databook et présentée ainsi en comité projet : parmi les 453 variables de la matrice finale, 100 variables ont été supprimées pour cause de taux de renseignement trop faible ou sans variabilité (par exemple, toujours égale à « oui »). Ensuite, les variables qui constituent potentiellement les conséquences d'une attrition et non pas les causes sont supprimées : par exemple, vu que la résiliation du contrat nécessite l'envoi d'un courrier de résiliation, les contacts de type « nombre de courriers reçus par l'assureur » constituent une variable biaisée, à exclure de l'analyse des facteurs de résiliation. Enfin, cette matrice a un usage double, à la fois pour les analyses prédictives (scoring), et pour les analyses prescriptives (algorithme Q-Finder) : ces dernières nécessitent une quantillisation statistique des variables continues (remplacement des valeurs par des classes de valeur, définies de façon optimale selon leur distribution statistique et leur sens métier) afin de limiter le temps de calcul et faciliter l'interprétation des résultats.*

Il est à noter que cette dernière partie documentée dans le Databook, c'est-à-dire la sélection des variables finales, ne constitue pas une phase chronologique du projet. En effet, dès lors que la matrice finale, telle que décrite dans le Databook, a été construite, les analyses prédictives et prescriptives ont été entamées par les deux Data Scientists ayant le rôle de Machine Learners. Or, l'application de modèles basiques (premiers coups de sonde) a permis d'identifier rapidement trois éléments nouveaux : l'absence de signal dans les variables peu renseignées ou sans variabilité, un signal très fort sur certaines variables comme la réception de courrier, et les difficultés à interpréter certains résultats. Le premier constat a conduit à l'élimination de la suite des analyses, avec l'approbation des experts métier, des variables inutiles afin de simplifier les modèles. Le second a donné lieu à une investigation terrain plus poussée, notamment auprès des contributeurs du service administratif, non identifié initialement comme partie prenante du projet. En effet, les courriers de résiliation étaient compris comme exclus des données

collectées, or il s'est avéré que le processus de saisie des informations de contact ne permettait pas de les exclure correctement, car le contenu du courrier n'était pas connu au moment de la saisie de la réception, donc cette donnée contenait sans distinction des courriers de résiliation et des courriers autres (sinistres, réclamations...). La découverte à ce stade de ce manque de qualité des données a nécessité l'exclusion du périmètre de l'analyse de l'ensemble des courriers reçus, n'ayant trouvé aucun moyen de retraiter de façon efficiente tous les courriers de 2014. Enfin, les premières analyses prescriptives résultaient en un certain nombre de contextes de résiliation difficilement interprétables, comme par exemple les clients ayant eu « moins de 3 » prestations santé au cours de l'année, mais aussi « moins de 2 » et « moins de 4 ». Or la majorité de cette population avait tout simplement 0 prestations, indicateur plus facile à interpréter. Une quantillisation « mixte » entre des classes métier (aucune, peu, ou beaucoup de prestations) et les classes statistiques a alors fait le consensus, d'autant plus qu'elle accélérait le temps de calcul grâce à la simplification de la variable continue. Ainsi, la superposition de la phase d'analyse avec la phase de structuration, animée au cours des instances de médiation, a permis de rendre des modèles plus simples et plus pertinents en termes de sens, et a pointé des pistes d'optimisation de la qualité des données complémentaires.

Une fois la sélection des variables finales réalisée, après les coups de sonde, les analyses prédictives et prescriptives ont été menées de front, et ont donné lieu à des problématiques différentes d'appropriation progressive des résultats par les experts métier.

Résultat des analyses prédictives et difficultés d'interprétation d'un scoring « boite noire »

La modélisation de la probabilité de risque de résiliation pour chaque affaire santé a mobilisé un modèle prédictif de type « boite noire », connu alors pour être utilisé dans les moteurs de recherche Yahoo ou Yandex, et depuis peu pour la tarification en assurance et l'attrition des utilisateurs de cartes de crédit. Le choix de l'algorithme et la recherche des paramètres optimaux (nombre d'arbres de décision, leur profondeur, vitesse d'apprentissage...) n'ont pas fait l'objet d'une explication particulière au cours du projet, car les experts métier ne tenaient pas à internaliser la méthode, assez complexe à leur sens. Les choix de paramétrage ont alors été réalisés par le Machine Learner seul, dans le cadre de sa connaissance des librairies de modèles disponibles pour ce type de données et de sujets, et a nécessité un travail de veille et

d'expérimentation, assez court vu la présence d'une documentation scientifique riche et facilement accessible pour un Data Scientist junior.

Cet écart entre la complexité perçue par les acteurs métier et la complexité d'exécution réelle n'a pas semblé important à combler dans les circonstances du projet. La construction du modèle sur les données du projet a abouti cependant sur deux éléments partagés avec l'ensemble de l'équipe : d'une part, une matrice comprenant un score de risque de résiliation pour chaque affaire de la matrice finale, et d'autre part des indicateurs statistiques permettant d'évaluer la qualité statistique du modèle. Or, ces indicateurs statistiques étaient peu interprétables pour un non Data Scientist, ce qui a complexifié l'évaluation du modèle par les experts métier.

En effet, la modélisation s'est basée sur une démarche de découpage de la matrice complète en 2 parties :

- une base d'apprentissage, comprenant une partie des affaires santé, les variables explicatives et le phénomène d'intérêt (attrition opérationnelle),

- une base de validation, comprenant l'autre partie des affaires santé et les variables explicatives, mais sans le phénomène d'intérêt.

La base d'apprentissage servait à établir le modèle, c'est-à-dire trouver la fonction optimale qui permet de modéliser le phénomène d'intérêt affaire par affaire en fonction de ses caractéristiques, et la base de validation servait à appliquer ce modèle optimal pour générer un phénomène d'intérêt prédit pour chaque affaire. Sur cette base de validation, le Data Scientist dispose alors de l'attrition prédite pour chaque affaire (issue du modèle) et de l'attrition réelle (contenue dans la matrice complète initialement), et peut les comparer. L'écart entre le phénomène d'intérêt prédit par le modèle et le phénomène d'intérêt observé est représenté habituellement par une courbe ROC[41], présentant la sensibilité et la spécificité du modèle. La sensibilité est la capacité du modèle à bien cibler les contrats résiliés sans en oublier (maximisation de la part des positifs prédits positifs, c'est-à-dire des « vrais positifs ») : un modèle théorique qui indique que tous les contrats seront résiliés a une sensibilité de 100% car il n'en « oublie » pas. A l'opposé, la spécificité est la capacité à éviter de pointer comme à risque d'attrition des affaires non résiliées (minimisation de la part des négatifs prédits positifs, c'est-à-dire des « faux négatifs ») : un modèle théorique qui indique qu'aucun contrat ne sera

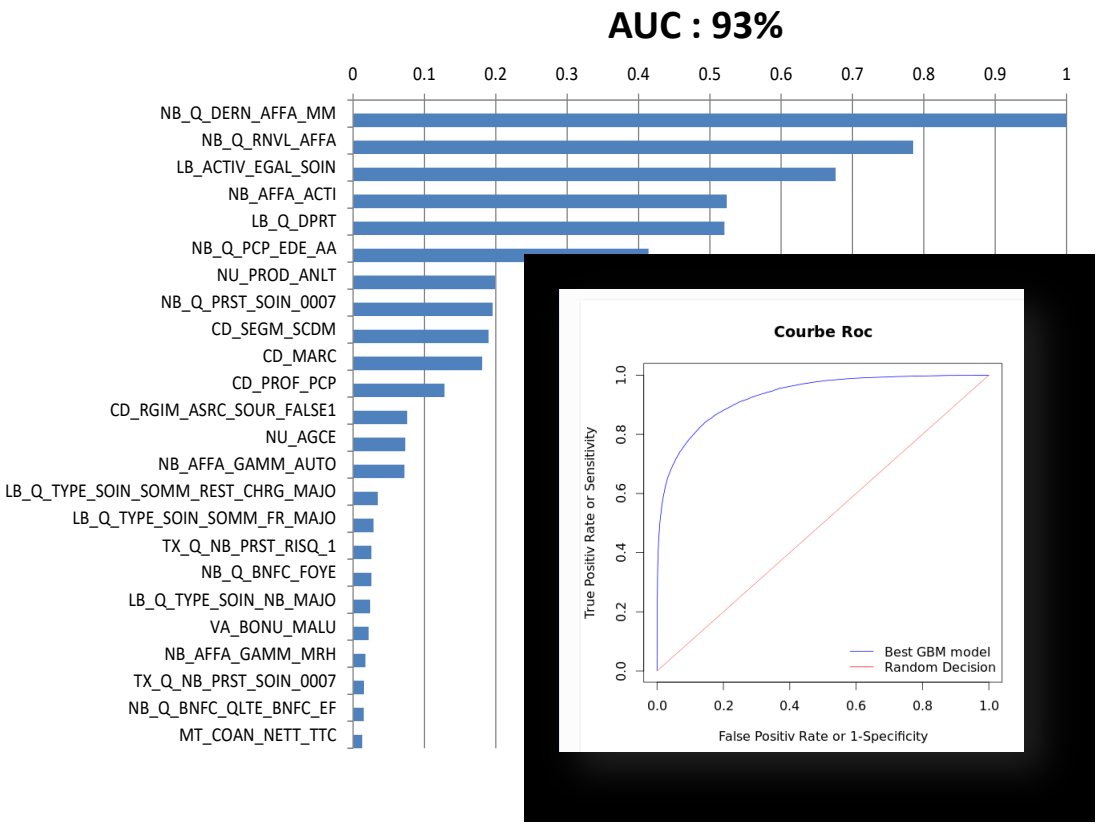
[41] Receiver Operating Characteristic

résilié a une spécificité de 100% car il ne se trompe sur aucun contrat non résilié. Dans ce cadre, la médiation entre les enjeux liés à l'attrition et les indicateurs statistiques a été indispensable, afin de reboucler avec le démarrage du projet : en effet, l'optimisation de la sensibilité du modèle est indispensable pour ne pas oublier d'agir sur ces clients fragiles et perdre le chiffre d'affaires associé, et l'optimisation de la spécificité est nécessaire pour économiser les ressources dédiées à la rétention des clients, en évitant d'agir sur des clients non fragiles, et plus globalement ne pas sur-solliciter les clients. En absence de directive particulière des experts métier, aucun des deux indicateurs n'a été privilégié, le modèle a été construit de façon équilibrée, et l'indicateur statistique final présenté est unique : il s'agit de l'AUC[42], correspondant à l'aire sous la courbe de ROC représentant la combinaison entre la sensibilité et la spécificité du modèle (voir illustration des indicateurs d'évaluation du modèle présentés dans l'Extrait 4). Dans le cadre de cet indicateur, un modèle théorique qui attribue le risque d'attrition totalement au hasard aura une valeur AUC de 50%, et un modèle théorique qui comporte un biais et donne toujours le bon résultat aura un AUC de 100%. Un AUC permet dans ce cadre de comparer juger globalement un modèle en le comparant à ces deux modèles théoriques (plus l'indicateur est proche de 100%, meilleur est le modèle), ou de comparer deux modèles entre eux. Le modèle final généré par le Data Scientist avait un AUC de 93%, soit un indicateur jugé très robuste du point de vue statistique, et suffisant pour les métiers étant donné l'évaluation de leurs modèles de scoring plus habituels réalisés par les Data Miners des équipes Marketing.

Cet indicateur, peu interprétable en termes de logique de construction, a été complété par la restitution des variables qui pèsent le plus dans le calcul du score : il s'agit d'une représentation graphique de ces pondérations ordonnées des variables (voir illustration des indicateurs d'évaluation du modèle présentés dans l'Extrait 4). Cependant, la lisibilité d'une telle représentation, parfaitement suffisante du point de vue du Machine Learner, a rapidement été remise en cause par les experts métier. En effet, si l'importance de la variable est bien lisible, la façon dont elle impactait le score ne l'est pas. Par exemple, la variable NB_Q_DERN_AFFA_MM, c'est-à-dire le nombre de mois qui s'est écoulé depuis la signature de la dernière affaire (santé ou hors santé) était bien identifiée comme importante dans le score, mais la représentation ne permettait pas de voir si le score était plus élevé si le nombre de mois était élevé, ou l'inverse. Une étude dédiée de l'impact de cette variable isolée (une seule

[42] Area Under Curve

dimension) était alors nécessaire pour confirmer que l'ancienneté de la dernière affaire jouait en faveur de la fidélité, ce qui s'éloignait du modèle prédictif car une étude monodimensionnelle ne prenait pas en compte l'interaction entre cette variable avec les autres. La médiation sur ce sujet a été rendue difficile par des cadres méthodologiques différents entre le Machine Learning (génération de modèle optimal multidimensionnel) et les analyses métier habituelles (ajout progressif de dimensions interprétables, mais isolées). Enfin, comme les scores précédemment utilisés, celui-ci ne rendait pas aisé l'interprétation du risque pour chaque affaire, et notamment l'identification de la « raison » principale pour laquelle une affaire en particulier était à risque. En effet, bien qu'on sache que globalement, certaines raisons (ancienneté du contrat santé, détention d'affaires autres que santé...) pèsent lourdement dans le scoring, le score ne pointe pas la raison affaire par affaire, c'est-à-dire ne permet pas de savoir quelle caractéristique précise influence en priorité la probabilité de résiliation. Ainsi, deux affaires ayant un risque d'attrition équivalent, par exemple de 80%, peuvent être expliquées par des variables différentes : la première par le fait que l'affaire est récente, et la seconde par le fait qu'un client ne possède pas d'autres contrats en dehors de l'affaire santé. Or, dans ce cadre, l'usage du scoring restait difficile car il ne permettait pas à un agent de construire un discours commercial adapté à chaque situation. Il s'agissait ainsi d'un score très précis, mais comportant les limites d'un scoring prédictif habituel de type boite noire, ce qui devait justement être compensé par la seconde approche analytique, prescriptive.



Extrait 4 ***Illustration des indicateurs d'évaluation du modèle prédictif****, présentée ainsi en comité de projet : la courbe ROC du modèle de scoring (en bleu) présente une optimisation harmonieuse de la sensibilité (en ordonnées) et de la spécificité (en abscisses) du modèle, et le modèle est significativement plus performant qu'un modèle théorique qui attribuerait un score de risque de résiliation au hasard (en rouge). Le score AUC du modèle s'élève en effet à 93%, et le modèle est particulièrement sensible à l'ancienneté de la dernière affaire souscrite auprès de AAA (NB_Q_DERN_AFFA), au nombre de renouvellements passés du contrat santé (NB_Q_RVNL_AFFA), à l'activation de certaines garanties comme le soin (LB_ACTIV_EGAL_SOIN), au nombre d'affaires actives (NB_AFFA_ACTI), au département de l'agence (LB_Q_DPRT)...*

Résultats des analyses prescriptives, et démarche d'émergence d'hypothèses métier

L'approche par l'analyse prescriptive a permis l'identification, la sélection et l'interprétation de 19 contextes qui maximisaient le risque de résiliation, appelés « règles » ou profils client, grâce à l'algorithme générique Q-Finder, conçu par Quinten et utilisé sur de nombreux projets de ciblage de sous-groupes. Cet algorithme permet un balayage massif de l'ensemble des variables d'une matrice d'apprentissage sous l'angle d'un phénomène d'intérêt, ici l'attrition

opérationnelle. Plus concrètement, dans un premier temps le Q-Finder est paramétré par le Machine Learner pour tester un a un tous les facteurs d'attrition possibles (toutes les variables et toutes les modalités de chaque variable) pour identifier des contextes à risque, c'est-à-dire toutes les sous-populations qui ont un taux d'attrition au moins deux fois supérieur à la moyenne. Dans un second temps, le Q-Finder est paramétré pour tester toutes les combinaisons de facteurs deux par deux, et enfin trois par trois. Ces analyses, massivement combinatoires, permettent d'explorer de façon exhaustive et sans *a priori* tous les contextes possibles d'attrition et de pointer les contextes qui s'avèrent statistiquement à risque, et ce de façon significative (rapport suffisant entre la taille de la sous-population et le risque d'attrition).

Un contexte à risque d'attrition identifié est ainsi constitué de 1, 2 ou 3 facteurs combinés, ce qui constitue une sous-population d'affaires à risque avéré. Un facteur est dans ce cadre une variable et un ensemble de modalités de cette variable : il s'agit d'un simple filtre applicable à la population globale, ce filtre étant par construction optimal. Or, une analyse exhaustive et sans *a priori* correspond à l'établissement de plusieurs centaines de contextes à risque, et ce nombre de contextes est difficile à interpréter un par un, d'autant plus que certains groupes de contextes pointaient des réalités métier similaires. Par exemple, trois contextes comme « l'ancienneté de la dernière affaire est inférieure à 12 mois », « l'ancienneté du client est inférieure à 18 mois », et « le nombre d'affaires détenues en dehors du contrat santé est égal à 0 » pointent en réalité de 3 façons des populations assez proches, et notamment tous les nouveaux clients qui viennent de signer leur premier contrat avec AAA en santé. L'interprétation d'un tel nombre d'hypothèses était impossible à soumettre aux experts métier, et une sélection plus rigoureuse a été demandée aux Data Scientists. Cette sélection s'est appuyée sur l'outil Diamond (un outil d'analyse et de visualisation de sous-populations développé par Quinten pour faciliter l'analyse des résultats du Q-Finder par les Data Scientists, et comportant un ensemble d'indicateurs et de représentations statistiques). La sélection des contextes visait à couvrir le maximum du phénomène d'intérêt (c'est-à-dire la plus grande partie des clients churners) avec le moins de contextes possibles, et en choisissant les contextes les plus diversifiées en termes de facteurs. La sélection a abouti à l'élimination d'une grande partie des contextes qui pointaient les mêmes populations, pour se concentrer sur seulement 19 contextes qui, à eux seuls, couvraient 71% de l'attrition, c'est-à-dire que 71% des affaires résiliées trouvaient des hypothèses d'explication de la cause de résiliation, ce qui était jugé parfaitement suffisant par les experts métier.

Numéro règle(ppt)	Critère(s)	Taille	Lift
Base		*340129*	*x 1.0*
Règle 1	Ancienneté de l'affaire santé < 2ans	49361	x 2.0
Règle 2	Nombre de prestation(s) = 0 Numéro de produit analytique = 451001	9294	x 2.5
Règle 3	[redacted]	18259	x 2.0
Règle 4	[redacted]	2259	x 3.4
Règle 5	Nombre d'e-mail émis en 1 an ≥ 5	11177	x 6.0
Règle 6	[redacted]	13580	x 5.3
Règle 7	[redacted]	14210	x 5.2
Règle 8	[redacted]	7409	x 3.3
Règle 9	[redacted]	7168	x 4.7
Règle 10	[redacted]	5575	x 4.3
Règle 11	[redacted]	24989	x 2.7
Règle 12	[redacted]	11596	x 4.0
Règle 13	[redacted]	29452	x 2.7
Règle 14	[redacted]	27332	x 2.3
Règle 15	[redacted]	7061	x 2.7
Règle 16	Nombre d'e-mail émis en 1 an ≥ 5 Numéro du produit source = 541	8791	x 6.3
Règle 17	[redacted]	5353	x 2.8
Règle 18	[redacted]	15617	x 3.2
Règle 19	[redacted]	67205	x 2.1
Union des règles	-	143833	x 1.8

Extrait 5 ***Illustration des 19 contextes*** *(appelés ici « règles »)* ***qui permettent d'expliquer 71% de l'attrition****, présentés ainsi en comité de projet. Chaque règle est caractérisée par un*

ensemble de filtres (critère(s) constitués de variables et de seuils, ici essentiellement confidentiels pour ne pas dévoiler la stratégie marché de l'entreprise AAA), une taille (nombre de souscripteurs concernés par ces critères) et un « lift » (coefficient multiplicateur du risque d'attrition). La base totale analysée est constituée de 340 129 affaires, et peut être filtrée de 19 façons différentes pour isoler des contextes à risque d'attrition. Par exemple (règle 1), si la base des affaires est filtrée uniquement sur les affaires santé dont l'ancienneté est inférieure à 2 ans, alors la sous-population sera constituée de 49 361 affaires, et le risque d'attrition constaté dans cette population est 2 fois plus élevé que dans la base totale, soit un lift de 2. De même (règle 2), si les affaires sont filtrées uniquement sur des cas où aucune prestation n'a été consommée, et que le produit détenu s'appelle « 451001 » (il s'agit d'un type de contrat spécifique, anonymisé), alors nous obtenons une sous-population de 9 294 affaires, présentant une attrition 2,5 fois plus élevée que dans la population totale. Ces contextes d'attrition sont plutôt explicatifs, mais peu opérationnels : si l'on cherche à mieux cibler les populations à risque, et non pas à mieux expliquer l'attrition, il faut maximiser le lift : par exemple (règle 5), si le souscripteur a reçu plus de 5 e-mails en 1 an, alors il a un risque d'attrition 6 fois plus élevé qu'en moyenne, ce risque étant encore plus élevé s'il détient le produit « 541 » (règle 16).

Une appropriation progressive des résultats des analyses prescriptives

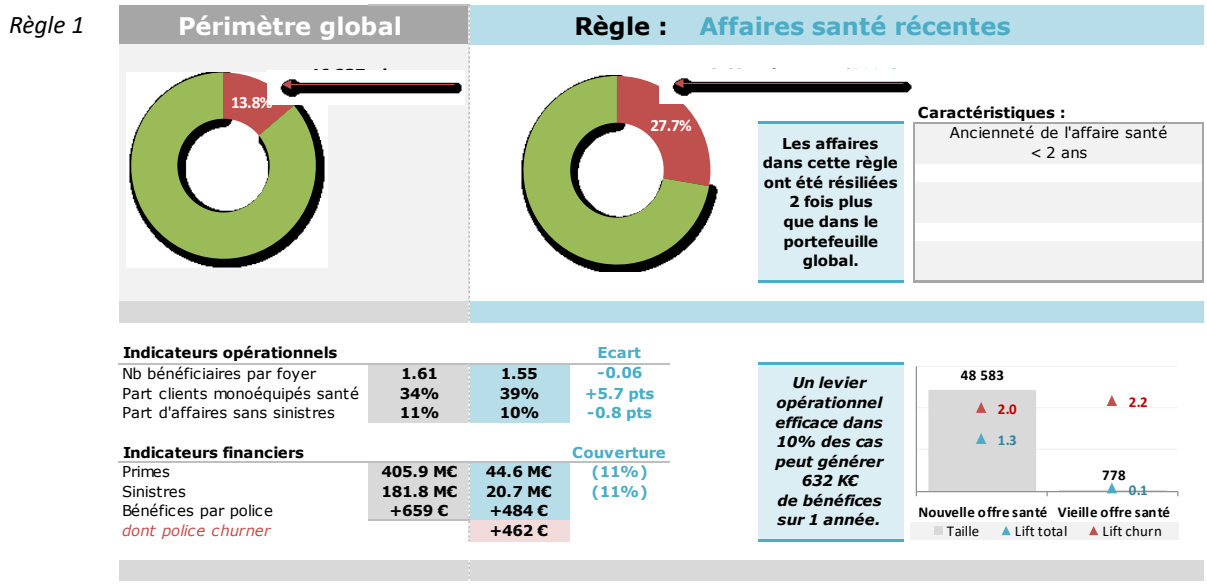
Les contextes sélectionnés par les Data Scientists ont éveillé un grand intérêt pour les experts métier, en particulier des contextes qui étaient inconnus, voire contrintuitifs. Par exemple, les experts métier ont découvert que lorsque les sociétaires étaient bien contactés par AAA, mais jamais par mail, leur risque d'attrition augmentait de 3.3 fois (règle 8). Or, le mail était perçu comme un levier plutôt moins fidélisant que le contact téléphonique ou le face à face. L'interprétation par le responsable de la relation client a permis de mettre le doigt sur la cause terrain cachée derrière ce contexte : en effet, lorsque la société connaissait peu son client, le mail du sociétaire n'était tout simplement pas renseigné, et par conséquent il ne recevait pas de mails. Ainsi, la cause de l'attrition était une mauvaise connaissance du client et non pas l'absence d'envoi de mail. Autre exemple, l'excès de sollicitations commerciales était bien pressenti comme une cause, plutôt mineure, d'attrition pour cause de ras-le-bol, cependant aucune étude existante ne pointait du doigt l'ampleur des dégâts et des seuils précis à partir desquels il est possible d'avoir un risque d'excès de sollicitation. Or, les analyses faisaient émerger des seuils précis dans pas moins de 12 contextes différents sur 19, ce qui donnait des

arguments massifs pour revoir la politique de relation client. En particulier, la sollicitation commerciale pour vendre un produit non issu de la gamme santé auprès des sociétaires fragiles qui venaient à peine d'acheter un produit santé, était identifiée comme dangereuse, alors qu'il s'agissait d'une pratique historique répandue.

Au-delà de ces découvertes, les contextes identifiés étaient en majeure partie pressentis comme des hypothèses de travail d'interprétation, et guidaient ainsi la compréhension des facteurs relatifs d'attrition. Cependant, bien que moins nombreux suite à la sélection statistique, les contextes restaient difficiles à s'approprier : à ce stade, un contexte était défini par un ensemble de facteurs intelligibles qui maximisaient le risque d'attrition, et représenté par sa taille et par son lift, soit la différence entre la concentration d'attrition dans la population totale et la concentration dans le contexte. Mais il restait peu visuel, présenté dans une liste de contextes, et ne portait pas d'autres éléments proches du cadre du travail habituel des experts métier. Cette représentation a alors été enrichie avec un ensemble d'indicateurs métier complémentaires, qui s'éloignaient des notions purement statistiques issues de l'outil Diamond des Data Scientists, comme la taille et le lift, pour intégrer des représentations plus habituelles pour les experts métier. Par exemple, l'ajout de certains indicateurs opérationnels et financiers ont largement facilité l'interprétation des contextes : il s'agissait du nombre de bénéficiaires du foyer, de la notion de mono ou multi-équipement (c'est-à-dire la détention d'un ou de plusieurs contrats d'assurance chez AAA), ou encore du montant de primes et de sinistres, ainsi que des bénéfices moyens par affaire (grossièrement estimés comme le montant des primes moins le montant de sinistres, divisés par le nombre d'affaires). Au fur et à mesure de l'avancement du projet et des échanges entre les Data Scientists et les experts métier, ces indicateurs étaient intégrés par les Data Scientists, qui commençaient à proposer eux-mêmes des indicateurs métier, comme la part d'affaires sans sinistres (contrats « dormants »). Par ailleurs, les experts métier ont voulu comprendre si chaque contexte présentait un potentiel économique s'il était activé par un levier. Or, aucun levier concret n'a été défini à ce stade face à chaque contexte. Des hypothèses de travail ont alors été fournies, basées sur l'expérience des campagnes marketing classiques : une campagne marketing est considérée comme efficace dans 10% des cas, c'est-à-dire qu'elle génère une réaction positive (non résiliation) chez un client sur 10 parmi les clients qui voulaient résilier dans la population ciblée par la campagne. Ainsi, chaque contexte a pu donner lieu à une estimation grossière de son potentiel, c'est-à-dire du gain espéré (préservation de bénéfices moyens) sur une sous-population (taille) contenant un certain nombre de churners potentiels (lift) selon l'hypothèse d'efficacité du levier. Enfin, des études marketing menées par

ailleurs montraient qu'il était important de renouveler l'offre historique, c'est-à-dire d'équiper les clients existants détenant l'ancienne offre santé avec une nouvelle offre, stratégiquement plus pertinente : cette nouvelle information a éveillé l'intérêt pour les contextes qui pouvaient toucher plus particulièrement les détenteurs de l'ancienne offre, ce qui n'était pas visible dans la liste initiale des contextes à risque.

Face au besoin de visualisation de tous ces indicateurs pour guider l'appropriation des résultats, un nouvel outil de visualisation des résultats a été conçu sous Excel. Il permettait de représenter, de façon homogène, tous les différents contextes à risque.



Extrait 6 ***Illustration de l'outil de restitution des contextes*** *(appelés ici « règles »)* ***intégrant des représentations et indicateurs utiles à l'appropriation des résultats****, présenté ainsi en comité de projet : le périmètre global (base totale analysée) est constitué de 340 129 affaires, dont 13,8% ont été résiliées, soit 46 837 « churners ». Or, si cette base est filtrée uniquement sur les affaires santé récentes (règle 1, caractérisée par « ancienneté de l'affaire santé inférieure à 2 ans), alors la sous-population sera constituée de 49 361 affaires, soit 15% de l'ensemble des affaires du portefeuille, et le risque d'attrition constaté dans cette sous-population est de 27,7%, soit 2 fois plus élevé que dans le portefeuille. Cela a représenté 13 684 churners, soit 29% de l'ensemble des affaires résiliées. Cette sous-population est similaire au portefeuille global en termes de nombre de bénéficiaires par foyer et de part d'affaires sans sinistres, cependant, il est à constater qu'elle contient plus de clients mono-équipés qu'en moyenne. Les bénéfices moyens perdus à cause des résiliations dans cette sous-population s'élèvent à 462€ par contrat. Si une campagne markéting était appliquée à cette population, et*

aurait pu probablement préserver 632 k€ de bénéfices. Cependant, cette population ne contient que très peu d'affaires correspondant à la vieille offre santé, soit 778 affaires, donc cette sous-population ne peut pas faire l'objet d'une campagne marketing de renouvellement.

L'ensemble des contextes présentés ainsi a été interprété par les experts métier, issus de la direction Relation Client, du Marketing, de la direction Santé, et du réseau des agences au cours d'ateliers d'appropriation dédiés. Ces ateliers ont permis une évaluation métier au-delà d'une évaluation statistique des résultats d'analyses prescriptives. Les contextes à risque ont été regroupés par les experts métier en 6 familles, qualifiées selon le sens métier des facteurs de résiliation qui caractérisaient les contextes composant chaque famille. Les 6 familles correspondaient alors à des concepts d'ancienneté du contrat santé, de consommation de prestations santé, d'équipement client, de récence de mouvement client, de connaissance client et enfin de pression commerciale, les deux derniers concepts étant composés de contextes contrintuitifs découverts grâce à l'analyse. Cette conceptualisation métier, plus longue que prévu en absence d'une interface plus malléable pour les experts métier, a permis de valider la pertinence stratégique des études sur ces sujets, et donnait des arguments tangibles pour la conception de leviers concrets pour chaque famille de facteurs de risque d'attrition. En effet, chaque famille a donné lieu à un établissement de leviers opérationnels afin de limiter le risque d'attrition, chaque levier portant un sens métier, une cible d'affaires, un coût, et une efficacité estimée, ce qui était nécessaire pour juger le ROI des opérations marketing (les leviers restent confidentiels dans le cadre de ces travaux). Les connaissances et les leviers opérationnels ainsi générés ont permis l'établissement d'un ensemble d'axes de réflexion stratégiques, intégrés à l'issue du projet dans le Programme Relationnel Client du groupe, sous forme rapport d'expert comprenant 18 recommandations générales ainsi que de 14 propositions d'action concrètes (résultat confidentiel). La mise en œuvre des actions et son suivi ont été réalisés par les experts métier en absence des Data Scientists à partir de 2016.

Au-delà de ces résultats portant directement sur le sujet d'attrition, le projet a permis une capitalisation très importante en termes de compétences (compréhension par les experts métier des méthodes projet data, des concepts algorithmiques, du processus d'exploration, et inversement des concepts métier, leviers opérationnels et enjeux stratégiques de l'assurance pour les Data Scientists), et de données. Plus particulièrement sur ce point, le Databook comprenait à l'issue du projet l'ensemble des pistes d'optimisation de la qualité des données identifiées au cours du projet. Par exemple, il était impossible de s'appuyer sur la date de

résiliation définie par la réception du courrier dans le cadre du projet, car le coût de la mise en qualité de cette donnée était jugé trop élevé par rapport au bénéfice à générer dans le cadre du projet. Si l'arbitrage a bien été réalisé en faveur de l'économie de ressources projet, la piste d'optimisation des modèles a aussi été identifiée, et la mise en qualité future de cette donnée a été planifiée dans une feuille de route data. La capitalisation sur la qualité des données a fourni une vingtaine de pistes de ce type, et notamment la mise en qualité des données permettant l'établissement d'une maille « foyer » au lieu d'une maille « client ». D'un autre côté, le projet a généré une documentation riche et de nouvelles données parfaitement exploitables dans d'autres situations : par exemples, le simple nombre de personnes dans le foyer du souscripteur a été identifié comme important à intégrer dans le cadre d'un autre projet data de l'assureur, la mise en place d'une Vision Client 360°. Ainsi cette capitalisation a donné des leviers en termes de gains de productivité sur les projets data de façon transversale, en dehors du sujet d'attrition.

1.3.2 Cas 2 : Prévision d'activité

1.3.2.1 Contexte et enjeux

Les prévisions de ventes permettent aux dirigeants d'entreprise d'avoir une vision future à court et long terme pour appuyer leur prise de décision et leur stratégie. Elles permettent également d'anticiper la facturation, les ressources nécessaires pour soutenir l'activité, élaborer le budget et prévoir les besoins de trésorerie. Il est donc important de réaliser des prédictions les plus réalistes possibles. Les approches utilisées dans la prédiction de ventes se résument principalement à des méthodes extrapolatives simples, réalisées dans le cadre de l'exercice budgétaire. Ainsi, dans la plupart des entreprises, les prédictions de ventes sont réalisées de façon collaborative (consolidation des prévisions intuitives remontées par les responsables de centres de profit), ou bien en calculant des moyennes mobiles, à travers un lissage exponentiel simple, ou autres méthodes (Bourbonnais, 2001) pour dégager des tendances et des saisonnalités. L'exercice est réalisé manuellement une fois par an ou par trimestre pour un exercice annuel, et plus rarement sous forme de « rolling forecast », avec une prédiction à horizon mobile (Montgomery, 2002) qui s'alimente et s'ajuste automatiquement au fur et à mesure de l'avancement de l'activité de l'entreprise.

Chez BBB, l'un des leaders mondiaux de la production des parfums et arômes, les prédictions de vente étaient effectuées par le contrôle de gestion à court terme uniquement, pour un horizon allant d'un à trois mois, en fonction du stock de commandes en cours. Un besoin d'avoir une

visibilité des ventes à long terme a été formulé par les contrôleurs de gestion, et BBB a fait appel à Quinten pour obtenir des prédictions à long terme de qualité et à différents niveaux d'observation. Le projet consistait à construire des modèles de prédiction adaptés aux données de ventes de BBB. Les modèles devaient prédire les ventes des 12 mois à venir, au niveau global, mais aussi à 76 sous-niveaux (par régions et par pays, par produit, par client) à partir des données possédées par BBB.

1.3.2.2 Synthèse des résultats

Le projet a donné lieu à la conception d'un applicatif métier contenant les données historiques de vente et les prédictions des ventes à chaque niveau de lecture (76 nœuds) sur 12 mois roulants, chaque nœud étant doté d'un niveau de confiance dans les prévisions. L'outil a été mis à disposition auprès des contrôleurs de gestion des deux Business Units (« Parfums » et « Arômes ») de l'entreprise, optimisé, et déployé dans la Business Unit « Parfums », ce qui a permis d'accélérer le processus de prévision et d'investigation, et de mieux prendre en compte la variation de l'activité pour estimer les ressources nécessaires à l'exploitation. Le résultat a doté ainsi le Contrôle de Gestion de nouveaux éléments pour communiquer avec les actionnaires, mais aussi avec les différents interlocuteurs dans la hiérarchie des forces de vente. Plus largement, le projet s'est inscrit dans une montée en maturité des équipes internes dans l'exploration de leurs données, et a alimenté les réflexions sur la capacité d'innovation dans l'entreprise grâce aux nouvelles méthodes et approches propres aux projets data.

1.3.2.3 Observations clés

Indicateurs de valeur :

L'approche innovante dans le cadre de la prédiction de chiffre d'activité était un sujet très suivi au sein de BBB. Cet exercice était effectué manuellement et était très chronophage pour des équipes qui manquaient de temps pour se concentrer sur l'analyse qualitative des chiffres. Ainsi, la valeur principale générée concernait le gain de temps des contrôleurs de gestion. Le projet n'a pas donné lieu à une estimation d'autres gains liés à la précision des prévisions, pourtant visés et atteints, comme sa contribution au développement d'une culture d'innovation.

Qualité des données :

Liste des données explorées : historique fin de facturation sur l'ensemble de la société, historique de commandes, données externes comme les devises…

Contrairement à ce qui a été pressenti au lancement du projet, les données disponibles en dehors des factures, comme les commandes ou les caractéristiques des équipes de R&D travaillant sur les nouveaux parfums pour répondre aux briefs des clients, n'ont pas pu être exploitées dans le cadre du projet, pour une question d'absence de clé pour relier les données de factures aux données d'avant-vente. Cette absence de préqualification des données a généré une perte de temps pour le projet, car les données d'avant-vente ont été auditées avant d'identifier ce problème de clé.

Par ailleurs, une recherche de corrélations a eu lieu entre les données internes (facturation) et les données externes, comme d'évolution des taux de change des différentes devises. Cette analyse a démontré que l'ajout de ces données externes coûterait plus cher (notamment en termes de mise en place et de gestion du processus de collecte) qu'il ne génèrerait de gain par l'amélioration de la précision du modèle de prévision.

Médiation Homme-Données :

Liste des acteurs du projet : Responsable Innovation, 5 Contrôleurs de Gestion et un responsable des outils décisionnels, 9 Data Scientists de Quinten (intervenus sur toute la durée du projet, dont 2 dédiés au développement de l'interface de l'applicatif).

Les experts ont trouvé l'approche innovante et ont souligné l'apport en termes de compétences statistiques et projet, avant de décider d'internaliser ces compétences. Ils ont apprécié la clarté de la stratégie d'analyse, ainsi que les résultats de prédictions. Par ailleurs, ils ont parfaitement adhéré à la démarche d'alignement mise en place au cours de la première phase du projet, et une trajectoire d'optimisation a été tracée en fonction de l'ensemble des connaissances mises en évidences et des incertitudes en suspens. Cette trajectoire d'optimisation a accentué le positionnement de l'usage opérationnel au cœur des axes prioritaires.

Plus particulièrement, l'impact de la Data Visualisation sur la qualité de l'alignement dans la phase de conception du modèle a été très appréciée dans la mesure où elle a permis de faire émerger des besoins métier, et à faire converger les équipes sur les concepts métier et statistiques difficiles à partager de façon intuitive. Au-delà des techniques de Data Visualisation, l'alignement sur des représentations sociales communes (mix entre les représentations entre le contrôle de gestion et les représentations issues de la Data Scinece) a été indispensable pour l'avancement du projet, et ce non seulement pour la construction de

l'applicatif mais aussi pour le choix des modèles. Enfin, la capitalisation de connaissances réalisée tout le long du projet a accéléré de façon significative les optimisations apportées au fur et à mesure de l'appropriation de l'applicatif.

L'une des observations clés du projet concerne par ailleurs les compétences mobilisées : en effet, l'attribution des responsabilités par type de compétences (expert métier, ingénieur data, machine learner, web développeur) ne fonctionne pas en absence d'un dosage adéquat de l'expérience sur chacune, et d'un médiateur capable de faire converger les parties prenantes. Il ne s'agit pas ici d'un problème de production, mais bien d'une capacité d'anticipation du projet, et notamment de la construction de la stratégie analytique (feuille de route de production des analyses pour le projet).

1.3.2.4 Compte rendu du projet

Un démarrage marqué par une absence d'alignement des acteurs impliqués

Lorsque le projet a été identifié comme présentant un intérêt significatif pour BBB, au cours de la phase amont (commerciale), il était porté principalement par la cellule innovation du client, qui avait pour objectif de générer des gains de temps pour les équipes de contrôle de gestion et d'infuser une culture d'innovation dans l'entreprise. La formulation de la problématique (prédire le chiffre d'affaires à 12 mois), l'explication du processus de vente, connu par les équipes de Data Scientists dans le cadre d'autres projets dans le secteur du parfum, et la quantité des données disponibles (une vingtaine de sources riches pour qualifier le processus opérationnel de vente, et notamment les caractéristiques du pipe commercial composé des contrats en cours de négociation, ainsi que l'historique de ventes de l'entreprise), ne présageait pas d'incertitudes particulièrement qui mettraient le projet en risque. Ainsi la démarche projet anticipée était constituée d'une courte phase de cadrage métier et data, puis d'une structuration des données et d'application de modèles prédictifs, et en parallèle d'un développement d'interface de visualisation des résultats. Cependant, le lancement du projet a révélé dès la première rencontre avec les experts métier (contrôleurs de gestion) et data (responsable des outils décisionnels) un élément inattendu structurant : les données opérationnelles, c'est-à-dire les caractéristiques du devis en cours de négociation, n'étaient pas rattachables aux données financières, c'est-à-dire les contrats effectivement vendus et facturés, car la jointure par le numéro de contrat face à un devis était inexistante, ce qui annulait toute possibilité d'exploiter

les données d'avant-vente pour prédire le chiffre d'affaire. Seules les données liées aux factures, portant la date, le montant, le client, le produit et le pays de vente étaient exploitables. La stratégie analytique prévue initialement était de fait obsolète, et il fallait en imaginer une nouvelle.

Le projet a ainsi démarré dans un contexte d'incertitude très élevée, en présence de 3 types d'acteurs principaux : des Experts Métier (dont les contrôleurs de gestion et moi-même), des Ingénieurs Data (responsables de la préparation des données) et un Machine Learner (responsable de la modélisation des prévisions). Les experts métier client n'ont jamais travaillé sur un projet Data Science et étaient en attente d'inspiration pour imaginer des solutions différentes de leur pratique habituelle, consistant à travailler uniquement sur les factures. Les Data Scientists n'ont jamais collaboré sur un projet de prévision d'activité financière pure, et ne comprenaient pas la façon dont le résultat allait être construit et utilisé. Aucun des acteurs n'avait la possibilité d'anticiper la stratégie analytique : en effet, l'expert métier avait besoin de comprendre si l'analyse était « faisable » pour arbitrer sur les choix métier (par exemple, le choix entre une maille de prédiction annuelle, mensuelle, ou quotidienne), les ingénieurs data avaient besoin de comprendre comment structurer les données pour qu'elles puissent être exploitables pour le Machine Learner et générer du sens pour l'expert métier, et le Machine Learner avait besoin d'avoir plus de précisions sur la question métier pour se prononcer sur la faisabilité de l'analyse et indiquer à l'ingénieur data la structure cible. Dans ce contexte, une première extraction de l'ensemble des facturations passées (une dizaine d'années d'historique, à la maille de chaque facture) a été fournie aux ingénieurs data pour qu'ils puissent établir les premiers éléments objectifs permettant de réaliser les arbitrages. Cependant, aucun ingénieur data n'avait effectué ce type d'analyse auparavant : leur métier consistait à structurer les données au service d'un modèle algorithmique, et non pas au service d'un arbitrage par les experts métier. Ils étaient ainsi incapables de prioriser l'angle sous lequel il était nécessaire de construire ces éléments, et notamment des ordres de grandeur simples. Par exemple, la multitude des axes d'analyse (par pays, par produit, par client…) était difficile à appréhender pour comprendre les axes qui comptaient le plus, ou encore l'évolution de la facturation dans le temps était difficile à représenter en absence de connaissance du pilotage financier du chiffre d'affaires (par jour, par mois, par an, en cumulé ou à l'instant t…). Le projet démarrait ainsi dans un contexte assez problématique : aucun acteur, expert sur son propre domaine, n'était expérimenté sur les métiers de l'autre, et personne ne comprenait les besoins de ses interlocuteurs. Par ailleurs, aucun acteur ne pouvait initier le mouvement : les Experts Métier,

habitués à l'investigation analytique, se heurtaient à l'impossibilité de manipuler les données, trop volumineuses pour être ouvertes dans les outils habituels de type Excel, et les Data Scientists ne comprenaient pas le sens des données fournies.

La première étape a été réorientée rapidement sur une phase de découverte du métier de contrôle de gestion et des données de facturation fournies aux Ingénieurs Data. Cette découverte a permis de mieux comprendre notamment la construction des données financières, et les besoins des contrôleurs de gestion. De nouvelles complexités se sont ajoutées à ce stade, notamment un besoin d'avoir une granularité d'analyse fine du chiffre d'affaires, sur 3 axes métier correspondant à l'organisation matricielle de l'entreprise par types de comptes client, par régions géographiques, et par types de produits. Par ailleurs, les différences de fonctionnement des entités Arôme (cycles d'activité trimestriels) et Parfum (cycles d'activité annuels avec un suivi mensuel) nécessitaient *a priori* des approches analytiques différentes en absence de granularité temporelle et d'horizon de prédiction communs. Enfin, les 3 axes d'analyse n'avaient pas la même importance sur ces deux activités de l'entreprise. Ainsi, le modèle de prévision était supposé être assez flexible pour répondre aux besoins des deux entités tout en maximisant la pertinence sur les spécificités. Les deux entités, chacune dotée de ses contrôleurs de gestion, chacune exerçant dans des pays différents, généraient un avancement à deux vitesses peu compatibles dans le cadre d'un projet commun.

En parallèle, face au besoin d'inspirer les acteurs métier avec des méthodes de prévision innovantes, le Machine Learner a proposé plusieurs types de modèles pouvant présenter un intérêt pour le projet. Les méthodes proposées englobaient notamment le Deep Learning, l'analyse par ondelettes souvent utilisée dans la compression d'images, ou encore des techniques inspirées de la modélisation de dynamiques physiques : par exemple, représentation d'une facturation nouvelle (lancement de nouveau produit sur le marché), de sa phase de croissance puis de déclin (érosion des ventes), sur le même principe qu'un modèle de mouvement de lancer de balle. Or, ces propositions étaient générées à partir de la veille externe (recherche sur les modèles prédictifs), et non pas des besoins métier qui se précisaient progressivement auprès des Ingénieurs Data. Un gouffre s'est installé progressivement entre le champ des possibles, de plus en plus ouvert du point de vue algorithmique, et les besoins métier de plus en plus précis.

La restructuration du dispositif projet et l'établissement d'une stratégie analytique

La convergence entre les 3 types d'acteurs a pu être réalisée dès lors qu'un quatrième acteur a rejoint le projet : il s'agit d'un acteur expérimenté sur les projets data et capable de faire communiquer l'Expert Métier et le Machine Learner, c'est-à-dire traduire les besoins métier en question data. L'apport clé de cette intervention a consisté à écarter définitivement la démarche par l'innovation externe (proposition de solutions en dehors des besoins métier) au profit de solutions algorithmiques plus pragmatiques, l'objectif étant de construire au plus vite un premier modèle de référence qui pourrait servir de base pour toutes les optimisations futures. Son implication a donné lieu à l'établissement d'un nouveau plan du projet, plus itératif, et à une nouvelle répartition des rôles parmi l'ensemble des acteurs impliqués. Le rôle d'ingénieur data a été préservé, en capitalisant sur la phase d'appropriation du sens de chaque base et de chaque variable sous l'angle métier. Le rôle du Machine Learner a été rendu plus opérationnel que théorique, avec une incitation plus forte à produire, puis expliquer des modèles. Et enfin, l'un des experts métier (moi-même) a dû prendre une place de médiateur, ainsi que de concepteur en avance de phase des interfaces de restitution des résultats en lien proche avec tous les autres acteurs. Un Web Développeur, secondé par un expert IT, allait par ailleurs être mobilisé pour le développement technique des interfaces une fois la conception stabilisée. Le rôle d'experts métier a été scindé en deux entre les contrôleurs de gestion côté « Aromes » et les contrôleurs de gestion côté « Parfum », tout en maintenant dans le dispositif le pôle innovation dont l'objectif était de garantir les liens synergiques entre les deux business units, ce qui a séparé le projet en deux. Ce nouveau dispositif a rendu possible les premiers arbitrages et a débloqué l'avancement.

Les contraintes émises par les experts métier ont été ajustées au fur et à mesure de l'exploration menée sur les données de ventes, de la validation des ordres de grandeur et de la visualisation des données par les experts métier. A ce stade, les contraintes ont été revues de la façon suivante :

- La prédiction portait sur le chiffre d'affaires mensuel sur 12 mois pour la business unit « Parfums », et le chiffre d'affaire trimestriel sur 4 trimestres pour la business unit « Aromes ». Ce choix était issu des impératifs économiques et opérationnels divergents entre les deux business units, et confirmé par un benchmark des modèles (détaillés plus loin) sur les deux mailles temporelles. *Dans la suite du compte rendu, seuls les travaux de recherche réalisés sur*

la business unit « Parfums » sera abordée, car la démarche a été répliqué sur la business unit « Aromes ».

- Une vente était caractérisée par un pays, par un produit et par un client, à une date donnée. Chaque dimension faisait l'objet de regroupements hiérarchiques selon trois visions (géographie, client et produit), ce qui a multiplié le nombre de modèles à construire et à optimiser (avec un modèle par nœud, soit 76 modèles)

- Il était nécessaire d'avoir des prédictions cohérentes sur les trois dimensions, c'est-à-dire que la somme des pays devait être égale à la somme des clients et des produits. Si ce choix parait évident dans une démarche déterministe qui somme les sous-parties, il ne l'est pas dans le cadre d'une modélisation probabiliste qui génère un modèle optimisé par nœud. A titre d'exemple, dans l'approche prédictive probabiliste, le Canada, les Etats Unis et l'Amérique du Nord sont traités comme 3 nœuds indépendants, donc la somme des chiffres d'affaires du Canada et des États-Unis n'est pas égale au chiffre d'affaires du continent Amérique du Nord.

- Les modèles devaient pouvoir être réajustés et réadaptés tous les mois lors de la mise à jour des données de ventes afin d'obtenir un mois de prédiction supplémentaire (alignement sur les processus de clôtures comptables). Il était donc nécessaire de construire des modèles stables dans le temps, c'est-à-dire qui auraient une bonne qualité de prédiction quelle que soit la période de prédiction et sans variabilité extrême d'un moins sur l'autre.

- Le type de données utilisables pour prédire les ventes futures était limité : la possibilité de lier une vente future à une commande existante, ou à une demande de devis (« brief » par une marque de parfum pour créer une nouvelle fragrance), était éliminée à cause de l'absence de clés dans les bases de données. Seul l'historique des ventes était ainsi exploitable jusqu'à nouvel ordre, et un chantier de revue des jointures a été identifié, et écarté à ce stade du projet.

- L'historique des ventes exploitables était constitué de 60 mois (de janvier 2010 à Juin 2015) pour prédire 12 mois, ce qui était assez limité pour une série temporelle.

La stratégie analytique, clé d'une exploration générant la convergence et la capitalisation

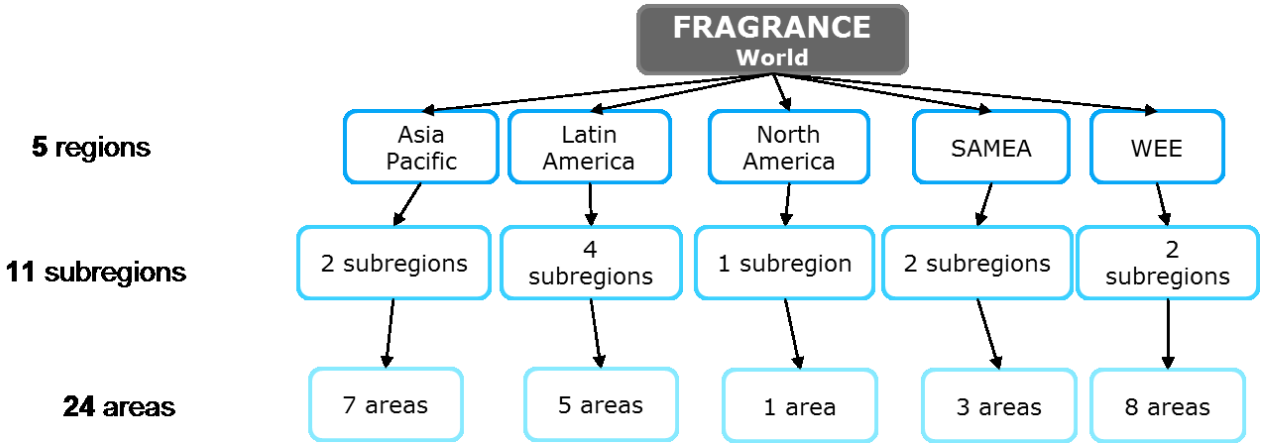
Afin de répondre à ces objectifs et contraintes, une démarche en quatre étapes a été mise en œuvre, chaque étape étant marquée par des choix de stratégies analytiques spécifiques et cohérents :

1. Trouver le format de structuration des données le plus adapté pour pouvoir extraire de l'information de qualité

A partir des données reçues, et en tenant compte de l'organisation hiérarchique des ventes, plusieurs formats de structuration ont été testés pour les modèles. Le format retenu était celui qui permettait de descendre au niveau le plus fin possible tout en permettant d'extraire du signal. Trois matrices ont été construites autour de 3 axes : l'axe géographique, l'axe client et l'axe produit. Chaque axe était composé d'un certain nombre de nœuds organisés de manière hiérarchique.

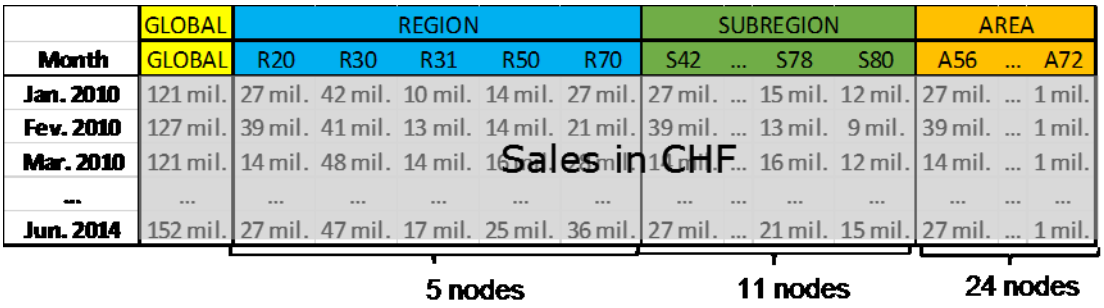
- Sur **l'axe géographique**, 41 nœuds organisés de la manière suivante :

  - Niveau 1 (niveau hiérarchique le plus haut) : 1 nœud (les ventes totales)
  - Niveau 2 : 5 nœuds (régions)
  - Niveau 3 : 11 nœuds (sous-régions)
  - Niveau 4 : 24 nœuds (zones géographiques)



Extrait 7 ***Illustration du modèle de structuration des ventes sur l'axe géographique****, issue d'un rapport intermédiaire : les ventes totales de la business unit « Parfum » (FRAGRANCE) est décomposée en 5 régions, elles-mêmes décomposées en 11 sous régions, qui à leur tour sont divisées en zones (areas). On s'affranchit ainsi de la notion de pays, qui peut constituer dans certains cas une maille trop fine, avec très peu de ventes, et donc difficilement prédictibles. La confirmation de cette structure est issue de la convergence entre les besoins métier (alignement sur l'organisation géographique des forces de vente et nécessité de descendre aux mailles fines) et la faisabilité (identification des mailles minimales qui restent statistiquement prédictibles)*

Cette structure hiérarchique a permis de construire la matrice suivante :



Extrait 8 ***Illustration de la matrice qui alimente les modèles sur l'axe géographique****, issue d'un rapport intermédiaire : les ventes sont regroupées par mois (lignes) et sont analysées sur tous les nœuds (colonnes) correspondant au modèle de structuration décrit précédemment. Chaque case correspond au montant des ventes, ici en Francs Suisses.*

- Sur **l'axe client**, 10 nœuds organisés de la manière suivante :

  - Niveau 1 : 1 nœud (ventes globales)
  - Niveau 2 : 3 nœuds (regroupement de segments de produits)
  - Niveau 3 : 6 nœuds (segments de produits)

- Sur **l'axe produit**, 27 nœuds organisés de la manière suivante :

  - Niveau 1 : 1 nœud (ventes globales)
  - Niveau 2 : 2 nœuds (type global du client)
  - Niveau 3 : 6 nœuds (taille du client)
  - Niveau 4 : 18 nœuds (client ou regroupement de clients)

2. Réaliser un benchmark de différents modèles afin de trouver le modèle de prédiction le plus adapté à chaque nœud et donnant les meilleures qualités de prédictions

Afin de trouver le modèle de prédiction le plus adapté à chacun des nœuds, plusieurs stratégies de modélisation ont été conçues et mises en œuvre, comprenant des modèles existants adaptés aux séries temporelles et d'autres modèles plus innovants d'autre part. Ces différents modèles étaient mis en compétition à travers un benchmark, et le meilleur selon le critère de sélection RMSE a été choisi (voir détails sur ce critère plus bas).

a. Découpage des matrices

Dans le cadre de ce projet, afin de contrer les biais de saisonnalités et de se mettre dans les conditions réelles de mise à jour des données, le choix s'est porté sur une méthode de découpage de matrice (matrice d'apprentissage et de validation) particulière. Ainsi, une des contraintes

était d'avoir un modèle qui puisse se réajuster et se réappliquer tous les mois lors de la mise à jour des données de ventes, pour donner un mois de prédiction supplémentaire. Il était donc important de sélectionner un modèle stable et dont la qualité de prédiction ne se dégradait pas au cours du temps, afin de garantir la confiance des utilisateurs dans la prédiction des ventes. Douze découpages par modèle et par nœud ont été réalisés, permettant de réaliser douze séries de prédiction et donnant douze erreurs de prédictions. En moyennant ces erreurs de prédictions, l'erreur globale de prédiction était obtenue pour chaque modèle à chaque nœud.

b. Types de modèles utilisés

Pour chaque nœud de chaque vision, différents types de modèles prédictifs ont été testés lors du benchmark (modèles standards ou stratégies de modélisations plus innovantes imaginées au démarrage du projet et jamais testées sur ce type de sujets dans l'état de l'art de la recherche à la date du projet). Sans rentrer dans les détails, ce projet a mobilisé un panel très riche et complexe de modèles algorithmiques, ce qui a permis de constituer en soi un socle de capitalisation conséquent en termes de Recherche et de Développement sur les séries temporelles.

c. Critère de sélection statistique

Le critère de sélection du meilleur modèle du point de vue statistique est le RMSE (Root Mean-Square Error), qui pénalise fortement les écarts importants entre valeur prédite et valeur observée. Le benchmark a permis de sélectionner la meilleure stratégie de modélisation pour chaque nœud selon la minimisation du critère RMSE. Dans une dernière étape, le modèle sélectionné était testé sur une période non utilisée dans le benchmark.

Cet indicateur statistique, classique dans le cadre de l'étude de séries temporelles, était méconnu par les experts métier et plus généralement par les utilisateurs de la prédiction des ventes, ce qui a nécessité une montée en compétences statistiques dans la mesure où il s'agissait de l'un des principaux critères d'évaluation des résultats qui devait servir de référence à l'ensemble des interlocuteurs.

3. Prédire les ventes sur la base de test : de Juillet 2014 à Juin 2015 avec le meilleur modèle sélectionné de chaque nœud grâce au Benchmark.

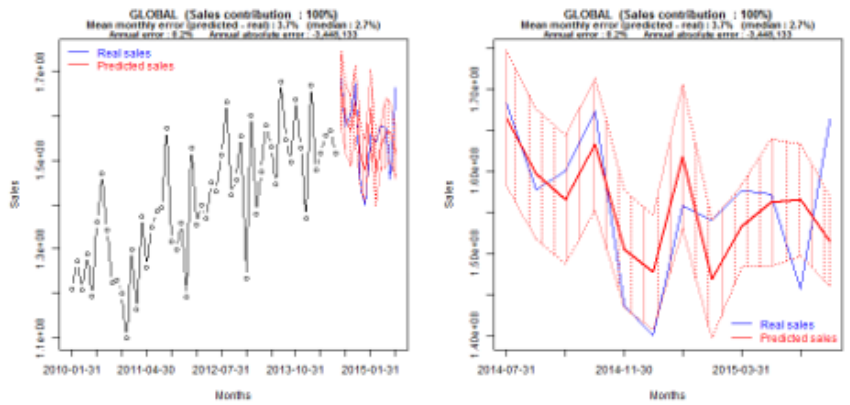
Le modèle sélectionné lors du benchmark, réalisé sur les données de Janvier 2010 à Juin 2014, allait prédire un an de ventes de Juillet 2014 à Juin 2015. Ces prédictions allaient ensuite être comparées aux ventes réellement réalisées, à travers l'erreur annuelle et l'erreur mensuelle, deux indicateurs opérationnels et dont le sens était parfaitement familier pour l'ensemble des interlocuteurs du projet, contrairement au RMSE. Cet ajout de critère métier a été clé pour l'évaluation opérationnelle du modèle et l'appropriation des résultats, et a facilité l'accord de confiance au RMSE.

4. « Reforecast »

Lorsque les prédictions étaient réalisées nœud par nœud et axe par axe, on arrivait à deux incohérences opérationnelles propres à la structure hiérarchique des séries temporelles. En effet, d'une part la somme des ventes prédites à un niveau inférieur de l'arbre (ex : les pays de l'Europe) différait de la prédiction du niveau supérieur (ex : l'Europe), et d'autre part les trois totaux de tous les nœuds, selon les visions géographique, produit et client, n'était pas identiques. Un « reforecast » en cascade a été réalisé à partir du nœud le plus agrégé (chiffre d'affaires total de la société), c'est-à-dire un ajustement de l'ensemble des modèles prédictifs des nœuds inférieurs pour que leur somme soit égale à la prédiction des nœuds supérieurs, permettant d'obtenir une cohérence des ventes prédites entre les nœuds et les niveaux. Là encore, les méthodes statistiques habituellement employées et documentées dans la recherche se sont avérées inappropriées, car aucune ne permettait de prendre en compte ces contextes métier évidents : l'inadéquation entre les approches statistiques théoriques et la réalité du besoin a une fois de plus nécessité l'élaboration d'une méthode spécifique.

Les résultats de la démarche étaient ainsi composés de l'ensemble des prédictions mensuelles sur 12 mois pour tous les nœuds et des indicateurs de confiance dans cette prédiction (extraits et fournis sous format Excel), ainsi que de l'ensemble des résultats transposés sur la base de validation et des critères statistiques et métier associés.

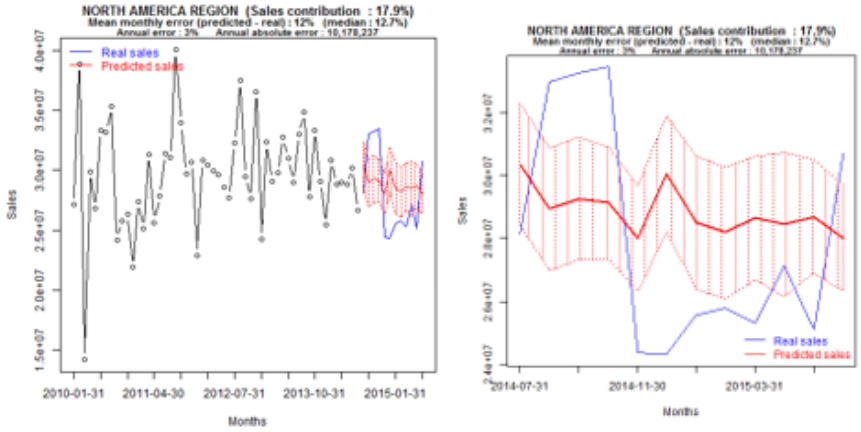
- Chiffre d'affaires global



Erreur annuelle : 0.2%

Erreur mensuelle moyenne : 3.7%

- Amérique du Nord



Erreur annuelle : 3%

Erreur mensuelle moyenne : 12%

Extrait 9 ***Illustration des critères d'évaluation des résultats****, issus d'une présentation intermédiaire des résultats : sur deux nœuds (chiffre d'affaires global et chiffre d'affaires de l'Amérique du Nord), l'historique des ventes sur la période d'observation est représenté par les courbes noires, la prédiction des ventes est représentée par les courbes rouges, accompagnées d'un intervalle de confiance, et les ventes réelles sur la période prédite sont représentées par les courbes bleues. Les représentations de droite sont des zooms sur la période prédite : on peut constater que le modèle prédit très bien les ventes globales, car les ventes réelles sont plutôt bien incluses dans la zone de confiance des prédictions, mais que le modèle prédit moins bien les ventes sur l'Amérique du Nord. Cela se voit aussi dans les critères*

d'évaluation métier : l'erreur annuelle au global est de 0,2%, alors qu'en Amérique du Nord elle est de 3%, et grimpe jusqu'à 12% si l'on s'intéresse à l'erreur mensuelle. Cependant, ces erreurs restent cohérentes : en effet, les ventes globales sont moins volatiles que les ventes en Amérique du Nord, comme on peut l'observer sur l'historique en noir.

L'importance de la Data Visualisation dans le processus de convergence

Un premier prototype d'application a été développé afin d'intégrer au fur et à mesure les résultats à une interface plus fonctionnelle pour les différents utilisateurs potentiels. Les itérations de conception que le prototype a induites dès le démarrage du projet ont été clés dans la phase de recherche du modèle approprié, et un vecteur de génération de connaissances auprès des experts métier. En effet, la Data Visualisation a permis aux experts métier de s'approprier plusieurs notions clés dans le modèle, et de formuler des besoins qu'ils n'avaient pas la capacité d'anticiper. Chaque étape de construction du prototype a été marquée par des échanges animés par le médiateur autour d'un tableau blanc (dessins réalisés en direct par les experts métier), de propositions de représentations (copies d'écran des représentations imaginées par les Data Scientists, exemples d'outils de restitution du marché) et d'échanges oraux sur les représentations sociales issues du contexte du contrôle de gestion et des statistiques.

Extrait 10 ***Illustration des propositions de restitutions des résultats*** *issues des supports d'atelier de conception de l'outil de restitution : représentations variées de la distribution des ventes par régions et par pays, des flux de facturation entre les pays, des ventes quotidiennes rapportées sur une année, et en dernier la prédiction des ventes réalisée avec un modèle de prédiction basique contenu dans l'outil de marché « Tableau » : ce dernier a constitué l'une des briques clés de la convergence vers une représentation cible des prévisions.*

Voici quelques exemples des concepts ainsi clarifiés :

- La notion de seuil de confiance statistique des ventes prédites, représenté par la zone rouge sur le graphique temporel dans le prototype : ce seuil de confiance, fixé à 90% dans les prédictions, n'a jamais été utilisé auparavant par les experts métier, et a permis de mettre en évidence des nœuds stables, pour lesquelles le modèle donnait des résultats très satisfaisants (zone de confiance étroite), et les nœuds à forte variabilité. Cet indicateur a non seulement apporté la connaissance de la variabilité des différents marchés (nœuds), mais a permis aussi de représenter cette variabilité de façon intuitive et opérationnelle, sans remplacer l'indicateur RMSE, difficile à interpréter pour un expert métier non-statisticien.

- L'évolution de la croissance annuelle prédite en fonction du mois de mise à jour des données. Ce dernier est un indicateur clé, imaginé par les experts métier pour visualiser la stabilité des prédictions dans le temps du point de vue opérationnel. Il s'agit de la croissance sur une année calendaire, liée à l'exercice budgétaire et financier, au lieu de la croissance sur l'année roulante, formulée initialement par les métiers comme prioritaire. Cet indicateur a donné lieu au développement d'un encart de visualisation spécifique dans l'outil, et à une révision de la fonction d'optimisation du modèle.

- Besoin de mettre en place un reforecast, afin de visualiser le même chiffre d'affaires sur les trois visions : en effet, ce besoin n'a pas été détecté directement, car il semblait évident pour les contrôleurs de gestion, et donc n'avait jamais fait l'objet d'une formalisation jusqu'au jour où une simple représentation graphique n'ait suscité une alerte de la part du Machine Learner.

L'applicatif a ainsi constitué un dispositif de convergence dans la démarche globale du projet grâce à sa capacité à aligner la compréhension des concepts analytiques et des concepts métier.

Extrait 11 ***Illustration de l'outil de prédiction construit pour le contrôle de gestion,*** *copie d'écran de la première version du prototype : en haut à gauche, une représentation graphique de type « sunburst » pour visualiser le chiffre d'affaires global en 2015 (en bleu eu centre) par régions (première rangée de pétales), sous-régions (deuxième rangée de pétales) et zones géographiques (troisième rangée de pétales). Cette représentation peut être basculée sur les axes par produit pet par client ou sur d'autres années grâce à la fenêtre de sélection dans la bande du haut. En cliquant sur un pétale, on obtient la représentation temporelle du chiffre d'affaires sur ce nœud, en haut à droite. Il s'agit de l'ensemble de l'historique des ventes (courbe bleue), et des ventes prédites avec leur zone de confiance (bande rouge). Ces éléments sont accompagnés d'indicateurs métier (tableau de suivi des ventes en bas de la courbe, et autres représentations opérationnelles dans les fenêtres du bas), dont l'évolution de la*

croissance annuelle selon le mois de prédiction (en bas à gauche) : cet indicateur permet de visualiser la stabilité des prédictions.

Le processus d'exploration et la capitalisation de connaissances métier

Le modèle ainsi initié a donné lieu à un ensemble de questionnements complémentaires, visant à optimiser l'usage des prédictions, et notamment réduire le niveau d'incertitude cognitive et statistique. Ces questionnements ont débouché sur une trajectoire d'optimisation en 4 étapes priorisées selon l'impact espéré de l'usage et le coût de développement.

1. Intégration des données macroéconomiques pour améliorer les résultats de prédiction

L'équipe des contrôleurs de gestion a eu l'intuition que les taux de change de certaines devises ainsi que le cours du pétrole devaient être des leviers forts de leur activité. En effet, les prix des produits de BBB étaient cotés en USD/CHF/EUR, mais les réglementations de certains pays imposaient d'effectuer les transactions en devises locales. Cela impliquait mécaniquement des fluctuations qui n'étaient pas liées à l'activité de production ou à la demande. Aussi, le caractère prédictif de variables macro-économiques exogènes aux factures, composées de 2 indices relatifs au cours pétrolier et de 18 taux de change, ont été testés. Pour cela, trois stratégies de réduction de dimension ont été mises en place, la première étant une méthode métier, et les deux suivantes statistiques :

- **Sélection métier des features** : La première méthode intuitive a été de sélectionner les « features » (devises et indices) en fonctions des nœuds sur lesquels était faite la prédiction (par exemple, sélectionner le taux de change du real brésilien pour le nœud Brésil). Cependant, en dehors de ces nœuds évidents, la sélection des features n'étaient pas intuitive (quelle devise ou combinaison de devises choisir pour le produit « Parfum de luxe » ? Pour la zone Asie Pacifique ?)

- **Réduction de dimension non supervisée (ACP)** : L'objectif de l'ACP était de réduire toute l'information contenue dans les 20 variables dans un nombre plus restreint de features afin de l'utiliser comme input du modèle.

- **Réduction de dimension supervisée (PLS)** : L'objectif était de maximiser la variance des indices et taux tout en maximisant leur corrélation avec le chiffre d'affaires

Ces différentes stratégies d'ajout de données exogènes ont été testées en input d'un modèle ARIMA, avec observation de l'impact sur le RMSE. Les résultats de prédiction ont été présentés nœud par nœud afin de générer des connaissances et d'arbitrer quant à la pertinence de l'intégration des données exogènes dans le processus de prédiction. Cette exploration a conduit à un constat qui a contredit les intuitions des contrôleurs de gestion : l'intégration des données macro-économiques n'améliorait que dans de rares cas la prévision de l'activité.

ENDUSE AXIS

GLOBAL

	ΔRMSE_bench	ΔAE _test
GLOBAL	- 626 665	2,2

UPPER SEGMENT

	ΔRMSE_bench	ΔAE _test
UNKNOWN	91 089	-34,9
FF	627 771	-0,2
CP	152 164	-0,9
OTHERS	4 168	5,6

SEGMENT

	ΔRMSE_bench	ΔAE _test
UNKOWN	97 978	-35,0
FINE FRAGRANCE	627 771	-0,2
FABRIC CARE	353 571	0,6
HOME CARE	446 983	-1,4
PERSONAL CARE	- 111 204	0,7
ORAL CARE	- 48 616	4,4
OTHERS	4 168	5,6

CUSTOMER AXIS

GLOBAL

	ΔRMSE_bench	ΔAE _test
GLOBAL	- 626 665	2,2

COMMERCIAL RESPONSIBILITY

	ΔRMSE_bench	ΔAE _test
CP	- 517 971	0,8
FF	808 505	1,3

SALES CUSTOMER TYPE

	ΔRMSE_bench	ΔAE _test
CP_Key Global Customers	282 909	1,9
CP_Key International Customers	11 993	1,1
CP_Local & Regional Customers	679 255	0,9
FF_Key Global Customers	134 635	-5,3
FF_Key International Customers	427 815	2,9
FF_Local & Regional Customers	73 920	-15,4

CUSTOMER OR GROUP OF CUSTOMERS

	ΔRMSE_bench	ΔAE _test
X275000	123 657	13,9
X723000	181 084	-1,9
X877000	416 849	1,1
small_cust_CP_k10	- 31 579	-2,2
X135000	37 567	-0,1
X751000	- 110 128	2
X800000	18 171	0,2
small_cust_CP_k20	- 222 807	-3,2
X2594	116 428	1,9
X597000	60 597	-0,1
small_cust_CP_nlr	500 968	0
X436000	68 871	20,6
small_cust_FF_k10	53 398	2,5
X100000	8 415	-0,4
X645	296 209	-0,4
X657000	48 473	16,1
small_cust_FF_k20	257 917	-0,3
small_cust_FF_nlr	73 920	-15,4

GEOGRAPHY AXIS

GLOBAL

NAME	ΔRMSE_bench	ΔAE _test
GLOBAL	- 626 518	2,2

REGION

NAME	ΔRMSE_bench	ΔAE _test
NORTH AMERICA	288 465	-5,4
WEE	930 834	-0,2
SAMEA	56 689	1,1
LATIN AMERICA	144 926	1,6
ASIA PACIFIC	- 337 786	-2,4

SUBREGION

NAME	ΔRMSE_bench	ΔAE _test
U.S.A. _ CANADA	288 465	-5,4
EUROPE	340 517	0,1
CENTRAL ASIA	92 788	-3,3
AFRICAN MIDDLE EAST	- 136 302	-0,7
SOUTH ASIA	8 329	-6,2
ANDINA	5 668	-0,5
MEXICAN	- 47 909	-1,5
BRAZILIAN	- 105 786	-4,1
ARGENTINA	161 613	-2,6
SOUTH EAST ASIA	99 217	0,1
NORTH ASIA	71 015	0,6

AREA

NAME	ΔRMSE_bench	ΔAE _test
U.S.A. _ CANADA AREA	288 465	-5,4
CENTRAL EUROPEAN AREA	102 473	-0,1
FRANCE AREA	412 356	2,8
ITALIAN AREA	107 890	3
GERMAN AREA	325 695	2,1
UK AREA	144 786	-2,6
SPANISH AREA	24 572	-4,2
TURKISH AREA	131 220	-1,5
CENTRAL ASIA AREA	92 788	-3,3
SUBSAHARAN AREA	15 763	-3,6
N. AFRICA MIDDLE EAST	56 739	0
INDIA AREA	10 651	-6,6
VENEZUELA AREA	44 167	0,5
COLOMBIAN AREA	2 249	-2,2
MEXICAN AREA	- 37 696	-2,2
BRAZILIAN AREA	- 105 786	-4,1
ARGENTINA AREA	161 613	-2,6
SINGAPORE AREA	- 9 397	1,6
THAILAND AREA	24 828	0,3
INDONESIAN AREA	203 558	8,8
OCEANIA AREA	- 11 020	-1,3
GREATER CHINA AREA	- 57 477	0,3
JAPAN AREA	- 23 390	-5,2
KOREA AREA	135	6,7

Extrait 12 *Illustration des résultats de l'exploration de la piste d'intégration de données exogènes, issue d'un rapport intermédiaire : la métrique ΔRMSE_bench représente la*

différence de RMSE entre les prédictions avec intégration des données macro-économiques et les prédictions sans intégration de ces données sur la période de benchmark, et ΔAE_test représente la même grandeur sur la période de test. Chaque indicateur est en vert lorsqu'il pointe une amélioration de la prévision par l'ajout de données exogènes, et en rouge lorsqu'il la dégrade (génération de bruit). Lorsque l'ajout de données exogènes améliore la prédiction selon les deux indicateurs à la fois, le nœud en question est représenté en vert foncé : seuls 10 nœuds ont été optimisés selon les deux métriques.

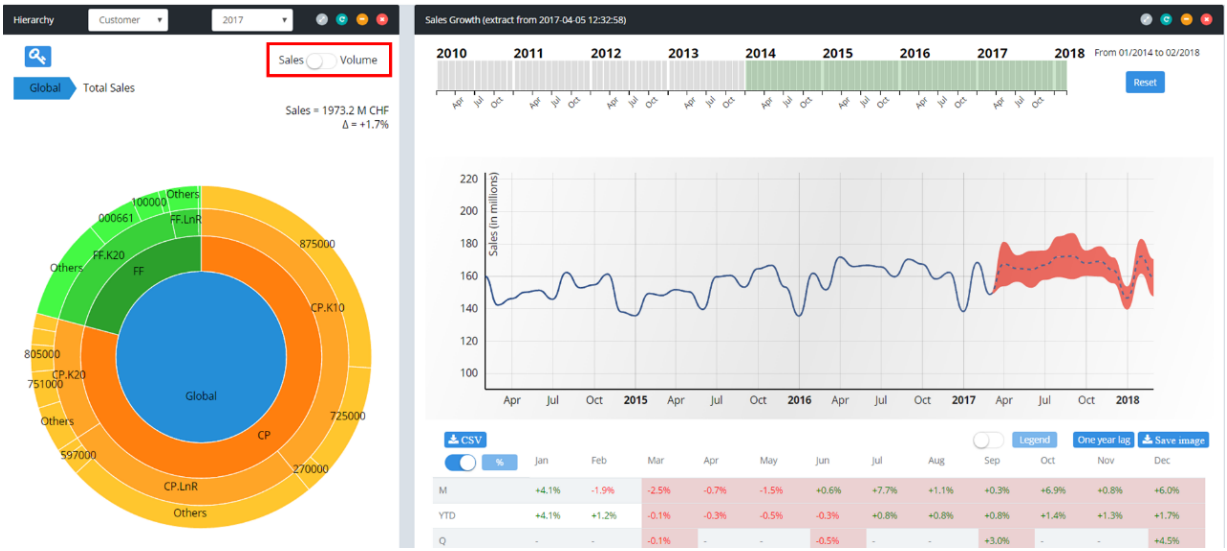
L'explication de ce manque d'optimisation tient au fait que, bien qu'explicatives des ventes, ces données macro-économiques devaient être prédites avant d'être utilisables, ce qui revenait à faire de la prédiction des ventes à partir de la prédiction des indicateurs macro-économique, cumulant ainsi des erreurs. Cette stratégie d'optimisation a ainsi été abandonnée, n'ayant été jugée ni assez robuste du point de vue statistique, ni assez opérationnelle, car elle allait impliquer un processus récurrent de collecte de données macro-économiques dès la mise en exploitation du modèle. En revanche, la génération de connaissance induite par cette phase exploratoire a été reconnue comme utile par les experts métier, que ce soit pour les 10 nœuds optimisés (sur lesquels l'intuition métier a bien été confirmée) ou pour l'absence d'optimisation sur les autres nœuds (objectivation claire de l'absence de corrélation). Ainsi, l'arbitrage en faveur de l'abandon de l'ajout de données exogènes a constitué un élément de connaissances capitalisé.

2. Prédiction des volumes

Pour la fonction contrôle de gestion, le pilotage financier (prédiction du chiffre d'affaires) s'accompagne d'un pilotage opérationnel de la production (prédiction des volumes de ventes, en kilogrammes de produits parfum vendus). L'intégralité des fonctionnalités développées pour la prédiction du chiffre d'affaires ont ainsi été transposées pour la prédiction des volumes de ventes. Les deux modèles de prédiction étant indépendants l'un de l'autre, il a fallu également faire une analyse de cohérence de prédictions entre ces deux métriques pour chaque nœud, les tendances d'évolution du chiffre d'affaires et des volumes de ventes devant être corrélées. Cette analyse s'est avérée plus complexe que prévu : en effet, elle a mis en évidence des cas où le chiffre d'affaires et les volumes n'étaient pas directement corrélés. L'analyse métier de ces cas a permis d'identifier des spécificités technologiques dans la vente des parfums (modification de la concentration de certains produits, diminuant mécaniquement le volume pour un prix équivalent) comme facteur de divergence des deux grandeurs. Cette investigation de la part du

contrôle de gestion, guidée par l'émergence d'une incohérence statistique, a permis de capitaliser une connaissance métier dans le service de contrôle de gestion, bien qu'elle existât déjà dans les entités opérationnelles directement concernées par ces produits.

S'appuyant sur la même démarche analytique que pour la prédiction du chiffre d'affaires, la qualité de la prédiction des volumes de vente est du même ordre de grandeur, avec une erreur annuelle de 0,1% sur le nœud global. Ces prédictions, de qualité jugée satisfaisante par les experts métier, ont donné lieu à une industrialisation de l'applicatif de prévision complet. Il s'agit ici d'un exemple clair d'une évolution itérative d'un applicatif, bien que sa première version était dès le départ plus évoluée et confortable qu'un MVP (Minimum Viable Product).



Extrait 13 ***Illustration de l'optimisation de la solution de prévision d'activité par l'ajout de l'indicateur de volume en kilogrammes****, copie d'écran de la deuxième version du prototype : en haut, dans l'encadré rouge, ajout d'une fonction qui permet de « switcher » d'une vision de la vente en euros (Sales) à une vision des volumes en kilogrammes (Volume).*

3. Optimisation de la prédiction sur la croissance annuelle

L'actionnariat de BBB, entreprise cotée sur le marché Suisse, suit de très près les communications financières qui ont lieu tous les trimestres et impactent le cours de l'action. Parmi elles, les communications sur les performances annuelles du groupe (publiées en février pour l'année calendaire précédente) sont les plus suivies, les plus détaillées et mises en perspective avec les prévisions et les objectifs fixés l'année antérieure. Aussi, les contrôleurs de gestion de BBB apportent une attention particulière à « l'atterrissage de fin d'année », c'est-

à-dire aux prédictions des mois restants dans l'année en cours et qui permettront de calculer la croissance de l'activité par rapport à l'année précédente. Ainsi, il fallait éviter au maximum la dégradation de la prévision de l'atterrissage au profit d'une prévision roulante, c'est-à-dire qui glisse sur 12 mois et se décale d'un mois à l'autre, prévue initialement.

Afin de prendre en compte cette contrainte, la démarche de benchmark annuel a été modifiée en appliquant une surpondération sur les coefficients RMSE des mois en fin d'année prédits, ce qui a permis de sélectionner les modèles qui minimisaient en priorité les erreurs de fin d'année. A la suite de l'optimisation sur l'atterrissage de fin d'année, les taux d'erreurs sur l'atterrissage ont été réduits de plus d'un tiers, apportant ainsi aux actionnaires une meilleure capacité d'anticipation de leurs bénéfices (les leviers permis par cette anticipation n'ont pas fait l'objet d'une étude particulière au cours du projet).

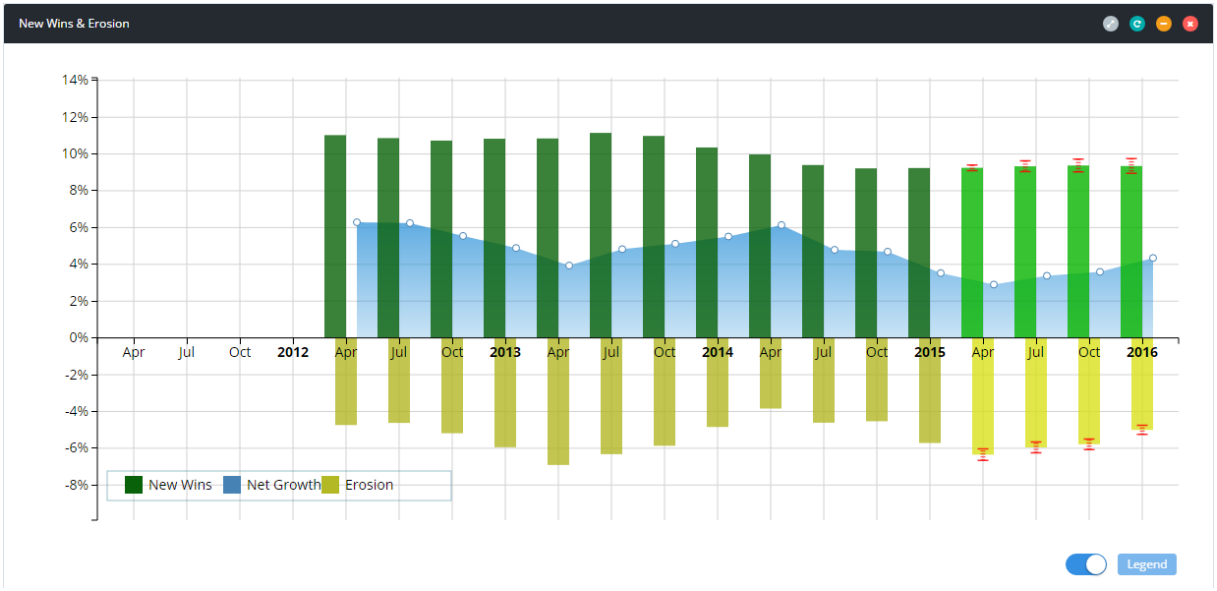
4. Prédiction des « New Wins » et de l'« Existing Business » *(pour la business unit « Arômes » uniquement)*

Comme évoqué, les besoins de pilotage et de visualisation des prédictions pour la Business Unit « Arômes » étaient spécifiques par rapport à la Business Unit « Parfums ». En effet, l'activité était pilotée selon deux notions décomposant le chiffre d'affaires :

- « New Wins » : vente de produits dont la première vente a eu lieu il y a moins d'un an et qui ont généré au moins 10kCHF (il s'agit de lancements de nouveaux produits)

- « Existing Business » : vente de tous les autres produits

Afin de tirer le maximum d'informations de ces données et des données de vente, de nouvelles variables ont été implémentées pour l'apprentissage, et quatre stratégies de modélisation possibles ont été appliquées et évaluées avec les experts métier. Sans surprise pour ces derniers, et toujours en absence d'éléments liés aux commandes en amont des ventes, les « New Wins » étaient moins prédictibles que l'« Existing Business », cependant cette optimisation a permis d'ajouter une visualisation des résultats manquante pour l'usage opérationnel. La prédiction de nouvelles grandeurs a été intégrée à l'application métier.



Extrait 14 ***Illustration de l'ajout d'une fonctionnalité spécifique de prédiction des ventes des Aromes,*** *copie d'écran de la troisième version du prototype : il s'agit de l'historique des ventes par trimestre avec l'arrivée de nouveaux produits (en vert foncé, en positif) et l'érosion des autres produits (en vert clair, en négatif). Ces entrées et sorties des ventes se compensent et expliquent la croissance nette (en bleu). Sur les 4 derniers trimestres, il s'agit d'indicateurs prédits par les modèles, avec un intervalle de confiance (moustaches rouges sur les histogrammes).*

A la suite de ces résultats, un ensemble de données supplémentaires à intégrer à terme pour mieux prédire les ventes a été listé par les experts métier, ce qui avait pour objectif d'anticiper l'amélioration de la précision de la prédiction des « New Wins ». Il s'agissait par exemple des briefs commerciaux (appels d'offres) ou des commandes (anticipation de la facturation à 3 mois). L'évaluation de l'impact de l'ajout de ces nouvelles données a été suspendue suite à la décision de ne pas industrialiser l'application pour les équipes « Arômes » à l'issue de l'année 2016, marquée par des spécificités de l'évolution du marché jamais observées auparavant, et donc imprédictibles, que ce soit par des algorithmes ou par des modèles de contrôle de gestion classiques : cette limite de l'usage a été jugée significative pour suspendre la trajectoire d'optimisation.

En conclusion, si le projet a été assez centré sur les méthodes de modélisation temporelles au cours de son démarrage, il a été **significativement recentré sur les problématiques métier** dans sa phase d'optimisation. Certaines optimisations, bien que pertinentes d'un point de vue statistique, n'étaient que peu exploitables du point de vue de l'usage. Par exemple,

l'optimisation par l'ajout de données macro-économiques sur les 10 nœuds où la corrélation a bien été prouvée avec les ventes n'a pas été jugée suffisamment intéressante pour mettre en place un processus de collecte et d'alimentation récurrente de ces données, trop coûteux par rapport au bénéfice lié au gain de précision des modèles. Autre exemple, les critères métier (suivi opérationnel sur l'atterrissage de fin d'année) ont été jugés prioritaires sur les critères statistiques (RMSE) qui ont dû être réajustés. Enfin, le développement de fonctionnalités pour faciliter l'usage, comme la visualisation du lancement de nouveaux produits « Arôme » a été priorisé au détriment de l'investissement dans l'optimisation algorithmique (meilleure prédiction des New Wins). Ce projet démontre ainsi de façon très claire que l'innovation par l'algorithmie de nouvelle génération, *a priori* innovante et inspirante, fait rapidement place à un travail algorithmique de fond, complexe et s'éloignant des modèles théoriques pour être essentiellement guidé par l'usage.

Le déploiement de l'applicatif métier a généré un ensemble de connaissances métier (variabilité des marchés, corrélation avec les données macro-économiques sur certains marchés, nécessité de maintenir un processus manuel pour des évènements imprévisibles…). Cette capitalisation a donné lieu à la volonté d'internaliser les méthodes de génération de connaissances à travers la Data Science. Cette internalisation a eu lieu d'une part sous forme de recrutement de Data Scientists par les équipes d'innovation pour une multiplication des projets data sur de nouvelles thématiques, et d'autre part sous forme de transfert de compétences en Data Science aux équipes Business Intelligence, interlocuteurs privilégiés des contrôleurs de gestion.

1.3.3 Cas 3 : Prévention santé prévoyance

1.3.3.1 Contexte et enjeux

CCC est un groupe paritaire de protection sociale français dont le chiffre d'affaires s'élève en 2014 à 3,6 milliards d'euros pour 3,9 milliards de fonds propres. Il couvre l'ensemble des besoins de protection des personnes en retraite complémentaire, santé, prévoyance et épargne. La santé collective au sein de CCC représente un portefeuille de 199 000 entreprises, dont environ 80% ont moins de 20 salariés, soit 4,7 millions d'assurés à titre collectif, et 1,8 millions d'assurés à titre individuel.

Dans le cadre d'un programme d'accélération de vente par la prévention, CCC souhaite apporter de la valeur ajoutée à ses clients (Branches professionnelles et entreprises) à travers une meilleure connaissance des risques individuels. Dans ce contexte, CCC a fait appel à l'expertise

de Quinten en termes de valorisation stratégique de données pour comprendre les comportements de consommation santé et prévention. L'intervention a eu lieu sous forme de « Preuve de concept », (Proof of Concept, ou POC) afin de permettre une validation rapide de la valeur ajoutée de l'approche sur un environnement Big Data qui venait d'être mis en place par l'assureur.

Le premier objectif de cette POC consistait à démontrer la possibilité de maîtriser les dépenses de santé en améliorant la connaissance des risques individuels pour organiser une prévention ciblée auprès des branches professionnelles et des entreprises clientes. Le second objectif était de valider la possibilité d'anticiper le risque lourd en santé et en prévoyance. Le cas d'usage proposé par Quinten pour répondre aux objectifs exprimés par les experts métier de CCC visait à générer les premières connaissances pour atteindre ces objectifs en identifiant des comportements typiques de consommation de prestation santé, et en établissant un lien de corrélation entre ces consommations typiques et le risque en prévoyance.

1.3.3.2 Synthèse des résultats

La POC a donné lieu à l'identification de profils de comportements de consommation typiques par branche d'activité, et des risques en prévoyance associés. Ces profils ont fait l'objet d'une interprétation métier (médicale, risque, relation branches) et ont été jugés intéressants dans la mesure où il s'agissait de connaissances intuitives, confirmées ou nouvelles. La POC a prouvé la richesse d'information contenue dans les données non exploitées jusque-là, et a donné lieu à un plan d'action moyen terme afin de tendre vers une application de prévention santé. Cette trajectoire d'optimisation contenait essentiellement des optimisations opérationnelles au service de la création de sens pour les métiers, et se basait sur une recherche de stratégies de modélisation complémentaires.

Cependant le projet n'a pas abouti à un usage concret, et la trajectoire d'optimisation, pourtant validée avant le changement des orientations stratégiques du groupe, n'a pas été mise en œuvre. Les compétences acquises en termes d'innovation et de gestion de projet data ont toutefois fait l'objet d'une capitalisation interne significative et de communication externe dans le milieu des mutuelles. Il s'agissait d'un projet d'apprentissage pour l'entreprise, et de recherche de preuve de concept, dont l'usage était avant tout intangible.

1.3.3.3 Observations clés

Indicateurs de valeur :

Malgré une mise en scène ROIste de ce projet sous forme de « Preuve de Concept », aucun usage direct n'a été déployé à l'issue du projet. La valeur ajoutée principale visée fut en effet la génération de connaissances (médicales), la montée en maturité sur les projets data et la prise en main d'une nouvelle plateforme technique. Or, l'évaluation de l'impact de la capitalisation de connaissances, et dans un second temps de la communication reste complexe.

Qualité des données :

Liste des données explorées : croisement de deux univers des données, la santé et la prévoyance, contenant les caractéristiques des clients, des contrats, des prestations…

Le projet a fait office d'un révélateur de l'absence de connaissance des données de l'entreprise en termes de croisement de deux univers de produits (santé et prévoyance). Ce simple croisement, sous forme de jointure, a dégagé des connaissances inédites sur le portefeuille client détenu.

Médiation Homme-Données :

Liste des acteurs du projet : Direction générale, CIL (Correspondant Informatique et Libertés), Responsable Observatoire des branches, Responsable Médecin-Conseil, Chef de projet (centre de solution décisionnel), autres contributeurs ponctuels du centre de solution décisionnel et spécialistes de certaines données, 3 Data Scientists de Quinten (dont 1 ingénieur data, 1 machine learner expert, et 1 plus généraliste et manger de projet, et deux interventions ponctuelles de médecin et d'architecte des données).

Le projet met en évidence la nécessité d'un travail de convergence entre de nombreux experts impliqués, et ce plus particulièrement au cours de la phase d'interprétation des résultats pour leur transformation en leviers opérationnels. Cette convergence a été accélérée par une représentation adaptée des résultats analytiques et par une communication pédagogique sur les résultats et les méthodes de génération de ces résultats (modèle algorithmique) : ces derniers apparaissent alors comme indissociables.

1.3.3.4 Compte rendu du projet

Voir Annexe 12 – Compte rendu cas 3 : Prévention santé prévoyance

1.3.4 Cas 4 : Contrôles de non-conformité

1.3.4.1 Contexte et enjeux

Au sein de l'Unité Technique & Produits du groupe DDD, la Direction des Contrôles a pour mission de contribuer à la maîtrise de la rentabilité et des risques techniques en contrôlant le respect des règles et des procédures édictées par les Directions techniques. Pour cela, la Direction des Contrôles a deux principaux objectifs :

- Emettre des recommandations et suivre la mise en œuvre des actions correctives

- Initier et/ou participer à l'amélioration des moyens et outils pour renforcer et sécuriser la qualité de souscription

Sur le terrain, la direction réalise environ 25 000 contrôles par an auprès de ses agents sur 4 univers : Auto, Multi Risque Habitation, Professionnels et Moyennes Entreprises. Le premier mobilise le plus de ressources en termes d'équipes de contrôle qui visitent les agences sur le terrain. Ces visites de contrôle permettent de détecter 20 à 25% de contrats non-conformes (dits NC), pour un manque à gagner en moyenne de 7% du montant souscrit, c'est-à-dire qu'un contrôle peut permettre d'identifier des contrats sur lesquels la prime est trop basse pour couvrir le risque, ce que nécessite de l'augmenter. Pour le portefeuille Auto, cela représente environ 7 millions d'euros.

Dans ce contexte, le contrôle de conformité a fait appel à Quinten pour un projet visant deux enjeux majeurs :

- Identifier les leviers de prévention en comprenant mieux les contextes dans lesquels les non-conformités sont les plus fréquentes, de manière à concevoir et déployer des plans de prévention de ces NC.

- Générer un ROI mesurable à court terme par le simple biais de contrôles plus ciblés, notamment en mobilisant moins de ressources de contrôle.

L'objectif du projet consistait à mettre en place une stratégie de ciblage des contrôles afin de libérer une partie des ressources de contrôle mobilisées sur l'univers Auto pour les affecter aux

autres univers, de rattraper les manques à gagner (bénéfices à court terme), et de définir des stratégies de prévention opérationnelles, forts d'une meilleure connaissance des contextes de non-conformité, pour rattraper les manques à gagner sur l'ensemble du portefeuille (bénéfices à moyen terme). Plus précisément, le travail à réaliser est défini, dès le démarrage du projet, par les objectifs analytiques suivants :

- Découvrir des contextes de visites de contrôle ayant donné lieu aux non conformités les plus conséquentes à travers un croisement des différentes bases existantes : ceci permettra d'analyser les résultats de ces contrôles et de définir les actions correctrices

- Prédire le risque de non-conformité de chaque visite de contrôle et ainsi allouer les ressources de manière optimale, l'objectif étant de mener des actions ciblées de diverses natures visant à renforcer et sécuriser la qualité de la souscription

1.3.4.2 Synthèse des résultats

Le projet a généré deux résultats analytiques clés : l'ensemble des contextes à risque de non-conformité et les caractéristiques explicites de ces contextes, ainsi que des scores de risque de non-conformité pour chaque contrat non contrôlé. L'articulation de ces deux résultats permettait alors d'établir des thématiques de contrôle (une thématique par contexte, défini par un ou plusieurs critères à appliquer sur l'ensemble des contrats), et de prioriser les contrats précis à contrôler dans le cadre de ces thématiques (au sein des sous-populations de contrat d'une thématique, les contrats sont classés en fonction du score de risque de non-conformité, et seuls les plus à risque sont contrôlés). Ces résultats ont permis d'identifier un bénéfice de 30% en termes de ressources dédiées aux contrôles, et un impact financier estimé à 15%. La mise en œuvre de l'usage a été immédiate à l'issue du projet, cependant cet usage direct a été complété par un second usage, indirect : la monté en maturité sur ce type de démarches analytiques a suscité un grand intérêt, et une internalisation des compétences a démarré rapidement, notamment avec le recrutement de deux Data Scientist qui ont rejoint l'équipe de contrôle. Cette internalisation de compétences a permis d'élargir la démarche à d'autres univers que l'Auto, de mettre en œuvre des optimisations complémentaires, ou encore de tester d'autres solutions data. Après 1 année d'exploitation des usages directs et indirects sur le terrain, le bénéfice est réellement mesuré s'est élevé à 50% d'impact financier au lieu des 15% estimés à l'issue du projet. La valeur de l'usage indirect a ainsi pu être estimée par différence, ce qui est plutôt rare.

Enfin, la réduction générale à terme de la non-conformité des contrats, grâce à une meilleure prévention suite à l'identification des facteurs de risque, devait générer une baisse globale des primes auto, créant ainsi un avantage compétitif pour acquérir plus de contrats. Cet impact n'a pas donné lieu à des mesures particulières, mais a permis de communiquer sur l'importance de ce type de démarches par la Direction des Contrôles.

1.3.4.3 Observations clés

Indicateurs de valeur :

Ce projet a été guidé par un indicateur de valeur clair dès le démarrage du projet (réduction des ressources de contrôle), et enrichi par un second (impact financier). Ce second indicateur a pu être non seulement atteint, mais dépassé grâce à l'internalisation de l'expertise, ce qui prouve que les usages indirects, comme la capitalisation de connaissances, peuvent dans certains cas être mesurés *a posteriori*.

Qualité des données :

Liste des données explorées : caractéristiques client et contrats auto, historique des contrôles de conformité, caractéristiques des quartiers, caractéristiques des agences et des intéressements…

L'éparpillement des données en entreprise présente l'inconvénient d'être générateur de coût pour les projets data. Inversement, les projets data permettent d'améliorer la qualité des données, par exemple en générant des données plus pertinentes et documentées ou de nouveaux modèles conceptuels, dont peuvent bénéficier les porteurs des projets.

Médiation Homme-Données :

Liste des acteurs du projet : Directrice des Risques et Responsable des conformités, 2 Data Scientists de Quinten (dont 1 plus orienté sur le Machine Learning).

Si le besoin de médiation humaine dans ce projet a été moindre, dans la mesure où certaines compétences statistiques étaient d'ores et déjà détenues par les experts métier, celle d'une médiation à travers des interfaces de manipulation des données volumineuses et complexes a clairement été identifiée comme un facteur clé de succès manquant. Cette absence a généré une charge lourde dans le travail d'appropriation des résultats par les experts métier.

1.3.4.4 Compte rendu du projet

Voir Annexe 13 – Compte rendu cas 4 : Contrôles de non-conformité

1.4 Cas réalisés non détaillés

1.4.1 Cas 5 : Sinistres lourds en dommage aux biens

1.4.1.1 Contexte et enjeux

EEE est l'un des premiers assureurs européens, et ses activités comprennent l'assurance de personnes, l'assurance de biens et responsabilité, l'assurance-crédit, l'assistance, la gestion d'actifs et la banque. Le groupe est actif en Allemagne, en France et en Italie.

Les orientations stratégiques du groupe, et notamment les initiatives MidCap, Satisfaction Client et Big Data, poussent EEE à prendre une longueur d'avance en termes de prévention dans le domaine de l'assurance Entreprise, et le groupe organise dans ce cadre un innovathon sur la prévention MidCap en juillet 2014. Dans le cadre de cette compétition interne, la proposition de création de valeur de l'équipe partenaire de Quinten retient l'attention du jury pour le projet d'utilisation du Big Data et de l'intelligence artificielle pour baisser l'occurrence et/ou le coût du sinistre.

Le projet s'inscrit dans un contexte métier suivant :

- Existence de bases de données riches mais non reliées : base contrats, sinistres, business box, rapports de visite, de sinistres…

- Prévention concentrée sur 10% des sites, soit 90 % à traiter

- Prévention majoritairement concentrée chez les ingénieurs de prévention

L'enjeu du projet consiste alors à bénéficier de la totalité des informations disponibles pour impliquer les réseaux d'EEE dans la baisse de la sinistralité à travers la prévention

L'objectif du projet consiste à améliorer le ciblage et la prédiction des risques pour systématiser une prévention personnalisée afin de baisser la fréquence des sinistres de la totalité des sites, y compris les 90% des sites non visités. Cette démarche vise à fournir à l'intermédiaire un nouveau diagnostic d'exposition aux risques pour chacun de ses clients et à animer une nouvelle prévention par tous les réseaux d'EEE.

1.4.1.2 Synthèse des résultats

Le projet a abouti à la découverte de nouveaux contextes de risques à travers un croisement des différentes bases existantes (identification de combinaisons de facteurs/contextes de risque) et à la génération d'un score de prédiction de la survenance des sinistres pour chaque contrat.

1.4.1.3 Observations clés

Indicateurs de valeur :

Le projet n'a pas donné lieu à une transformation des résultats directs en usages opérationnels, mais a eu un apport clé sur la perception de la qualité des données internes de l'entreprise.

Qualité des données :

Liste des données explorées : caractéristiques client MidCap, sinistres passés, dossiers de prévention, caractéristiques des zones géographiques…

L'apport principal a été de mettre en évidence l'absence de gestion de la qualité des données internes, et en particulier des dossiers de prévention : malgré l'existence d'un modèle standard de rapport de visite, parfaitement opérationnel pour une qualification individuelle des risques, les données ne sont pas exploitables en état pour effectuer des analyses comparatives. Ces rapports contiennent notamment des données descriptives textuelles non homogènes. Leur mise en qualité a été identifiée comme prioritaire à la suite du projet par rapport à la mise en œuvre des usages opérationnels.

Médiation Homme-Données :

Liste des acteurs du projet : Responsable MidCap, Référent data, Chef de projet, quelques contributeurs ponctuels, 2 Data Scientists de Quinten (dont 1 plus orienté sur le Machine Learning).

Ce projet a confirmé la nécessité d'une mise en place de médiation (instances d'arbitrage, représentations sociales et interfaces et capitalisation de connaissances) : leur application a conduit à un respect des charges anticipées du projet. Cependant, la communication autour du projet a fait l'effet d'un buzz non maîtrisé par l'équipe projet quant à l'usage de l'intelligence artificielle dans le traitement des données textuelles : l'exploitation de ces données, peu nombreuses et de qualité insuffisante, n'a pas donné d'éléments suffisants pour éviter la

déception. Ainsi, la médiation devrait être élargie au-delà du dispositif projet, car cette absence peut présenter des externalités négatives en cas de mauvaise prise en compte de la maturité de l'entreprise qui héberge le projet et de ses motivations.

1.4.2 Cas 6 : Prédiction des prix des agrumes

1.4.2.1 Contexte et enjeux

Le marché des parfums et arômes est soumis à une variation forte et peu prévisible du coût des matières premières, et notamment des différentes familles d'agrumes dont l'essence constitue la principale matière première du producteur. Les agrumes (citrons, oranges, et autres) présentent une grande diversité de qualité, et sont échangés sur un marché qui mêle des pratiques de négociation estivale de prix et quantités pour l'année à venir avec des achats continus au prix du marché. La Direction Achat du producteur souhaite, dans ce contexte, se doter d'une solution analytique permettant de prédire à horizon de 1 an l'évolution des prix des agrumes afin de pouvoir prendre des décisions d'achat et de stockage, et de challenger les décisions des négociateurs sur le terrain. Cette demande a paru difficilement réalisable du point de vue analytique : en effet, une prédiction d'un tel phénomène sur une période aussi longue ne présageait pas un résultat analytique satisfaisant. Cependant, la Direction Achat a tout de même souhaité faire le nécessaire pour comprendre le niveau d'incertitude et décider de réinvestir ou non dans le développement de la solution. Le projet a eu lieu sous forme d'entretiens avec les différentes parties prenantes et de qualification des données perçues comme utiles à travers un plan de collecte pour partager un diagnostic de la situation.

1.4.2.2 Synthèse des résultats

Le projet n'a pas donné lieu à la phase de structuration et de modélisation des données dans la mesure où il n'y a pas eu de convergence entre le besoin métier et les possibilités offertes par la Data Science. Cette absence de résultat cache une satisfaction client significative en termes de démystification du phénomène et de création de compétences projet data grâce au partage des critères d'exploitabilité des données perçues comme utiles. Or, la qualification des sources de données perçues comme utile a fait remonter une connaissance métier du terrain. En effet, la mobilisation d'un bénéficiaire direct du résultat attendu (négociateur des agrumes) a permis de mettre en évidence les véritables drivers du prix des agrumes sur le marché, totalement inexploitables en termes d'accès aux données et de sens : il s'agit du l'activité commerciale de l'industrie du pneu (usage d'essence d'agrumes dans la fabrication) et l'avènement de marées

noires (usage d'essence d'agrumes pour la dilution naturelle de la couche de pétrole). La collecte et le traitement récurrent des données nécessaires pour rendre ces réalités « observables » représentaient un investissement trop élevé par rapport au bénéfice attendu, et l'ampleur des incertitudes liées à ces drivers ne permettait pas de réduire le risque d'erreur de prédiction.

1.4.2.3 Observations clés

Indicateurs de valeur :

Ce cas illustre de façon radicale que le travail de production analytique n'est pas le seul à générer de la valeur au cours d'un projet data. Par ailleurs, il inscrit la capacité d'innovation comme un bénéfice potentiel direct du projet, et ce sans aucun usage déployé mais seulement à travers la création de compétences data, y compris minimales.

Qualité des données :

Liste des données explorées : évolution des prix des agrumes sur le marché, prix interne

Le travail de qualification des données, notamment externes, est un facteur fort de limitation d'investissements inutiles, mais il ne peut être réalisé de façon autonome, sans mobilisation d'hypothèses métier préalables : la valeur des données est confirmée comme dépendante de leur usage.

Médiation Homme-Données :

Liste des acteurs du projet : Responsable Achat, Acheteur, Data Analyst, 2 Data Scientists de Quinten.

La Médiation Homme-Données s'appuie sur la capitalisation de connaissances qui, à elle seule, peut présenter une valeur ajoutée clé, au-delà de tout travail analytique.

1.4.3 Cas 7 : Multi-équipement

1.4.3.1 Contexte et enjeux

Le marché de l'assurance construction est à un tournant de son histoire : le recul des cotisations, la hausse de la sinistralité, l'impact inflationniste des nouvelles réglementations, la baisse des taux d'intérêt et l'entrée en vigueur de Solvabilité II sont quelques exemples des défis à relever.

Le groupe GGG, spécialiste de l'assurance construction, souhaite se diversifier en trouvant de nouveaux relais de croissance sur les marchés hors-BTP tout en poursuivant l'enracinement sur son marché historique. Dans ce contexte, GGG fait appel à l'expertise de Quinten afin d'optimiser sa performance commerciale :

- Améliorer l'efficacité commerciale, sachant qu'un commercial coute entre 350K€ et 400K€ par an et génère en moyenne 1,2 millions d'affaires nouvelles sur 3 ans.

- Générer un chiffre d'affaires additionnel en augmentant le volume d'affaires par client

- Pérenniser les résultats sur le marché de l'assurance construction pour pouvoir se diversifier efficacement sur les marchés hors BTP

L'objectif global est alors d'augmenter le taux de saturation (multi équipement), initialement estimé à 2,1 produits par client.

Ce projet est réalisé avec un dispositif très serré : Quinten ne doit intervenir que sur l'analyse des données, et GGG garde entièrement la main sur la préparation des données et la structuration de la matrice d'apprentissage. Or, l'assureur n'a pas de compétences en Data Science, mais seulement en Business Intelligence. Par ailleurs, ce même interlocuteur, qui a l'habitude de travailler sur des sujets métier, représente aussi l'intérêt des acteurs métier.

1.4.3.2 Synthèse des résultats

Les analyses ont permis d'identifier 12 règles simples qui, à elles seules, permettent d'expliquer 65% des cas de multi-équipement. Il s'agit d'hypothèses opérationnelles pouvant être transposées à des clients non multi-équipés possédant les mêmes caractéristiques, mais aussi de jouer sur des leviers de communication pour maximiser les chances de souscription des clients non multi-équipés.

1.4.3.3 Observations clés

Indicateurs de valeur :

La mise en œuvre de nouveaux usages générés à partir des résultats du projet n'a pas été observée, bien que leur potentiel ait été partagé, sous la forme de contextes inédits favorisant le multi-équipement. Par ailleurs, aucun signe n'indique dans ce projet une valeur ajoutée quelconque liée aux usages indirects : en effet, la médiation et le travail sur la qualité des

données ont été très largement absents, ce qui n'a pas contribué à générer une capitalisation particulière.

Qualité des données :

Liste des données explorées : caractéristiques des clients et des contrats, campagnes commerciales, devis.

La documentation de la transformation des données a été lacunaire sur ce projet : seul un plan de collecte initial a été établi, et un cahier des charges pour l'extraction. Aucune sémantisation n'a eu lieu, ni une documentation des traitements réalisés. La découverte des erreurs était ainsi effectuée sur les matrices d'apprentissage directement, ainsi qu'au cours de la phase d'évaluation des résultats, et non pas en amont des analyses. Cette découverte d'erreurs a été rendue possible par la compensation de l'expertise métier par l'expérience du secteur de la part de l'équipe Quinten, mais n'a pas été suffisante pour éviter un dépassement significatif des ressources projet liées à la correction itérative des erreurs de qualité des données dans la matrice d'apprentissage fournie.

Médiation Homme-Données :

Liste des acteurs du projet : Représentant de l'équipe Business Intelligence, Directeur Commercial, 3 Data Scientists de Quinten.

La particularité de ce projet est un niveau d'interactions extrêmement faible, une quasi-absence de l'implication d'experts métier dans le projet (deux échanges de 2 heures au démarrage, puis 1 heure de restitution des résultats en fin de projet) ainsi que l'absence de coordination entre le responsable de structuration des données, pour son premier projet Data Science, et le Machine Learner. Les échanges oraux (2 ateliers de travail entre moi-même et l'interlocuteur client) et quelques échanges par mail n'ont pas compensé cette absence de communication directe.

Ce manque critique d'instances de médiation ainsi que leur inadéquation, tout comme celle des compétences pour réaliser les activités du projet, a retardé d'environ 6 mois le projet, a doublé le temps nécessaire à sa réalisation dans la mesure où la matrice d'apprentissage a fait l'objet d'un nombre important d'erreurs de construction, que ce soit en termes de données, de structure de données, ou de sens de données vis-à-vis de l'objectif de la modélisation. Enfin, l'absence

de l'interlocuteur métier n'a pas permis de corriger toutes les erreurs de sens métier, ni de confirmer la pertinence des résultats pour une transformation future en leviers opérationnels.

1.5 Etat des lieux des observations clés

A ce stade, il est nécessaire de récapituler les observations clés sur les trois dimensions principales et d'identifier les pistes d'amélioration du modèle de référence qui en découlent.

Indicateurs de valeur :

- Tous les cas ont généré des résultats analytiques aboutis et inédits, sauf le cas 6. Pourtant, l'inaboutissement du travail analytique n'enlève pas l'intérêt de ce projet. Inversement, un travail analytique abouti n'est pas nécessairement perçu comme intéressant (cas 7), ce qui pointe un éventuel sous-investissement dans la production non analytique au cours de ce projet. Il est donc nécessaire de **prévoir d'autres sources de valeur ajoutée que la production analytique**, décrite essentiellement par le processus de référence.

- Ces autres sources de valeur semblent liées à des mécanismes de capitalisation de connaissances, mises en exergue à travers la génération de connaissances métier (cas A, B, C, 1, 2, 3, 4, 5, 6, 7), un transfert de compétences en data science (cas 1, 2, 4, 6), ou encore la mise en évidence de lacunes dans les données et processus existants (cas 1, 5, 6). Ces éléments sont en effet perçus comme générateurs de valeur futurs, parfois activés et ayant des effets mesurables (cas 4) ou transformés en nouvelles activités (cas A, B, C). Ces **mécanismes de capitalisation** méritent ainsi d'être soulignés.

- Il faut distinguer les cas qui ont véritablement généré une valeur tangible (cas 4, et à moindre échelle le cas 2) de ceux où seul un potentiel de valeur a été mesuré (cas 1, 3, 5, 7), ou bien ceux où la valeur générée est difficilement mesurable (cas A, B, C, 6) car soumise à d'autres processus d'innovation auxquels le projet contribue. Globalement, il semble nécessaire d'avoir **un cadre d'évaluation plus approprié** qu'une approche par le ROI (retour sur investissement).

- Enfin, la plupart des projets n'ont pas donné lieu à une exploitation d'usage immédiatement à l'issue du projet, en présence d'incertitudes complémentaires à lever (cas B, 3, 5), et les usages ont pu évoluer dans le temps (cas A, 2, 4), au fur et à mesure de l'avancement de la production analytique : les projets data ont donc pu permettre de **lever une part d'incertitudes** et de générer des pistes d'**optimisation des usages**, arbitrés au cours ou après les projets.

Qualité des données :

- La **documentation des traitements des données** au cours des projets data est perçue comme un facteur de qualité externe (celle du résultat), interne (gain de temps), et extra-projet (capitalisation des pistes de mise en qualité des données). Elle contribue fortement à la convergence entre acteurs (cas 1, 4, 5, 6), et son absence met en péril les ressources du projet (cas 1, 2, 4, 7).

- La documentation des traitements est dynamique, marquant les **étapes intermédiaires** du travail analytique, et peut s'appuyer sur une structure générique comme un **Databook** (cas 1, 2, 4, 5, 6). Ce dernier retrace notamment des **métadonnées** clés (dont la sémantisation des données) et des **métriques propres à la construction algorithmique** (maille, phénomène d'intérêt, périmètre, drivers, et critères d'évaluation).

- L'**activité de mise en qualité** des données au cours du projet est bénéfique non seulement pour le résultat, mais aussi en dehors du projet (cas 1, 3, 4, 6). Elle est partielle, et peut ainsi révéler des lacunes en termes de qualité des données, voire permettre d'**arbitrer sur les priorités de mise en qualité** (cas 1, 2, 5, 6).

Médiation Homme-Données :

- Les compétences en Data Science mobilisées sur ces projets recouvrent une **diversité élevée de compétences** (statistique, technique, informatique, métier, ingénierie data, architecture, analyse de données, conseil, management…) et de **rôles** (Ingénieurs Data, Machine Learning, Experts Métier, Consultants, Managers de projet…). La complexité est accentuée par la fluctuation de la mobilisation de compétences et de rôles dans le temps, par les différences d'expérience et de séniorité, et par l'absence de séparation de ces compétences par individu (cas de projets à taille réduite). Une mise à plat semble indispensable pour gagner en efficience.

- La **nécessité d'interactions** est largement confirmée par tous les cas, **son absence étant génératrice de risques** sur le projet (délai et coût) et sur la qualité du résultat analytique (cas 1, 6, 7). En effet, ces instances permettent d'aligner l'ensemble des acteurs sur la même cible, et de réaliser des arbitrages sur les usages visés. Cela est particulièrement indispensable quand les usages ne sont pas clairement prédéfinis (cas 1, 3, 4, 7) ou quand les données ne sont pas maîtrisées par les acteurs demandeurs (cas 1, 2, 3, 5, 6).

- La construction des algorithmes est marquée par un ensemble de **jalons intermédiaires** qu'il est nécessaire de partager avec les parties prenantes. Ces jalons sont des instances de partage de connaissances et de documentation, de garantie d'adhésion à la cible du travail analytique, et peuvent en soi générer des connaissances utiles (cas 3, génération de connaissances stratégiques pour l'entreprise suite à l'harmonisation des tables santé et prévoyance).

- Au-delà de ces jalons sur le chemin critique du travail analytique, les interactions sont nécessaires entre tous les acteurs du dispositif, et ce à tous les stades du projet. Bien que les phases amont (ateliers de cadrage métier et data) et aval (ateliers d'interprétation, appropriation des résultats, transformation en leviers) soient plus marquées par ces échanges d'information, la régularité même des interactions, par exemple au cours des points d'avancement hebdomadaires (cas 1, 2, 3, 4, 5) est bénéfique. Cela confirme l'intérêt à **compléter le travail de production analytique avec des instances d'échange dédiées**. Il est possible d'envisager ces points comme une ouverture du projet à l'ensemble de l'entreprise, afin d'intégrer des enjeux de gouvernance (proposition cas 1) et de communication externe au dispositif (risque avéré cas 5).

- Plus l'hétérogénéité des compétences est forte, plus les instances d'interaction sont indispensables (intérêt moindre pour les instances pour le cas 4 en équipe réduite et partageant un socle de connaissances commun en statistiques et en assurance, moyen pour le cas 2 avec des contrôleurs de gestion aux profils analytiques, et très fort pour les cas 1 et 3). La compétence analytique semble discriminante, et ces **transferts de compétences** sont bénéfiques (cas 1, 2, 3, 5, 6) et très appréciés. Inversement, les projets sont générateurs d'**apprentissage métier** pour les acteurs data (cas 1, 2, 3, et cas 7 négatif), ce qui contribue à la pertinence des résultats analytiques.

- Les interactions orales ne suffisent pas : il est indispensable de documenter les échanges en prenant soin d'aligner tous les acteurs sur des **représentations sociales communes** (indicateurs, graphiques, vocabulaire…). Ce support d'échange concerne aussi les **interfaces homme-machine** dans le cas d'interactions entre l'homme et la donnée, en particulier pour l'appropriation des résultats algorithmiques (cas 1, 2, 3, 4, 5, 6). Plus particulièrement, ce point émerge du cas 4, où l'interface de manipulation des résultats algorithmiques est le principal axe d'amélioration identifié par les experts métier.

Pour approfondir la compréhension des cas, il est rappelé qu'il est possible de consulter deux comptes rendus complémentaires en Annexe 12 – Compte rendu cas 3 : Prévention santé prévoyance, et en Annexe 13 – Compte rendu cas 4 : Contrôles de non-conformité.

***

Il semble important de souligner à nouveau la grande diversité des cas usages, issus de secteurs aux enjeux distincts, mobilisant des algorithmes variés et aboutissant à des résultats de nature différente (génération de connaissances ou applicatifs). Si le processus de construction n'a pas pu être observé sur les trois cas de pré-expérimentation, il l'a été pleinement pour l'ensemble des sept cas suivants réalisés au sein des équipes de Data Scientists. Cette observation, sur la durée complète de chaque projet, fait émerger des points saillants sur les dimensions clés analysées, mis en relief grâce aux cas négatifs.

Conservant la **valeur**, le travail analytique n'est pas le seul à en générer. En effet, les projets génèrent aussi de la capitalisation de connaissances, un transfert de compétences, et mettent en évidence des lacunes dans les données et dans les processus existants. Ces éléments peuvent être transformés en usages avant l'aboutissement du projet, ce qui rend les approches par le ROI insuffisantes, d'autant plus que les projets peuvent laisser des incertitudes résiduelles nécessitant un réinvestissement. Par ailleurs, au sujet de la **qualité des données**, la documentation des traitements est perçue comme facteur de qualité externe, interne, mais aussi extra-projet. La dynamique de documentation est compatible avec une structure ad hoc du Databook. Enfin, le « Data Scientist » concentre une diversité forte de compétences et de rôles, complexifiés par une dynamique nécessitant des interactions au cours d'instances, orales et documentées, en particulier en cas de fortes incertitudes métier et data ou de fortes hétérogénéités de compétences. Ces instances servent de jalons pour la co-construction des algorithmes et pour l'émergence d'usages indirects. Cette **Médiation Homme-Données** est facilitée par des outils de capitalisation de connaissances, des représentations sociales partagées, des interfaces homme-données et une gestion de projet adaptée.

A la lumière de ces observations, le chapitre suivant présente les ajustements du modèle de référence et son inscription dans un cadre plus large donnant une place de choix à ces dimensions clés.

« Mon projet préféré ? C'est le prochain. »

***Frank Lloyd Wright**, Architecte, Artiste*

2 Modèle de dispositif projet Data Science et ses dimensions dégagées

La proposition de modèle décrit la dynamique interne transversale aux projets data et orientée sur la génération de la valeur par les usages. L'usage est défini ici délibérément comme synonyme de pratique métier, de l'action de production contextualisée des acteurs métier dans le système de l'entreprise, ces acteurs étant dotés d'un capital de connaissances et de représentations sociales. Ce choix de vocabulaire est induit par le fait que la proposition du modèle s'adresse à l'ensemble des parties prenantes d'un projet data en entreprise, et non pas seulement aux professionnels de l'information (absents des projets observés) qui s'intéressent aux traces des usages et aux détournements possibles. En effet, le projet en soi est une tentative d'actualisation des usages en tant que modèles logiques de production (Curry, Flett, et Hollingsworth, 2006), une transformation délibérée de la pratique métier à travers la conception de nouveaux outils dont les praticiens deviendront utilisateurs dans un second temps. Cette dynamique de transformation est au cœur de ces projets qui impactent les usages, et doit être mise en perspective selon les dimensions qu'elle porte, c'est-à-dire les indicateurs de valeur (des usages), la qualité des données, la Médiation Homme-Données. Ces dimensions s'inscrivent alors plutôt dans la pratique observée sur le terrain dans les projets data, en tant qu'ajustements des pratiques de la Data Science face aux modèles de référence.

La dynamique interne fait l'objet d'une mise en perspective préalable des limites du modèle de référence, c'est-à-dire des éléments non couverts par ce modèle CRISP_DM (Shearer, 2000; Wirth & Hipp, 2000). Ces divergences concernent les phases, les tâches et les résultats du modèle de référence, ainsi que les dépendances entre ces phases et la mise en lumière de la dynamique non seulement itérative, mais aussi simultanée. Face à ces limites du modèle de référence, un nouveau modèle est proposé, cohérent par rapport au premier, mais plus flexible et plus robuste afin d'intégrer les réalités incompatibles avec celui-ci. Ce modèle est profondément orienté sur l'usage pour favoriser la génération de valeur à travers le projet : l'outil (algorithme, applicatif…) n'en est pas la finalité, mais un levier de transformation. Cette génération de valeur est alors intégrée dans le processus en distinguant des activités d'arbitrage des activités de production. Cette distinction permet de tenir compte des conventions et des

négociations (Desrosières & Kott, 2005) qui jalonnent le dispositif, marqué par un niveau d'incertitudes élevé, et de rendre opérationnelle l'orientation sur l'usage à travers la prise en compte des indicateurs de bénéfices, de ressources, et de ce niveau d'incertitudes.

A l'issue d'un travail de confrontation entre la pratique et la théorie, le modèle proposé a été complétée avec des outils et représentations enseignables et à visée opérationnelle. L'articulation du processus avec les indicateurs de valeur des usages donne des clés de lecture pour soutenir les arbitrages au cours des projets. Le dispositif projet est alors indissociable du modèle proposé. La dimension de qualité des données a donné lieu à une proposition de structure de prototype de Databook, documentation de référence de la transformation des données qui sert la qualification des données et la capitalisation de connaissances. Enfin, la Médiation Homme-Données dresse les modes de médiation en fonction de la maturité du dispositif et les facteurs de succès du projet : cette structure peut servir de base pour une meilleure anticipation des ressources pour les projets futurs, de façon intelligible et contextuelle.

2.1 Modèle CRISP_DM et études de cas : analyse comparative

Deux éléments clés sont comparés dans ces travaux : le détail des phases du modèle de processus de référence ainsi que sa dynamique, et notamment les liens de dépendance. Si les phases, assez complètes, ne sont critiquées qu'à la marge sur trois inadéquations, la dynamique de référence présente non seulement une inadaptation au terrain, mais aussi des manques qui limitent le pilotage de l'ensemble du processus.

2.1.1 Critique des phases, des tâches et des résultats

Le modèle CRISP_DM s'est avéré parfaitement adapté à la description de la réalité des projets data observés sur le terrain en termes de phases : en effet, toutes les activités telles que décrites dans le modèle sont bien identifiables dans les études de cas observées (voir Figure 24). Seul le cas 6 n'a pas permis d'observer l'ensemble des phases, car il fait l'objet d'un arrêt anticipé du projet, suite à la preuve de l'inadéquation entre les besoins métier et les données disponibles.

#	Nom du cas	Respect des phases CRISP-DM					
		Compréhension métier	Compréhension des données	Préparation des données	Modélisation	Evaluation	Déploiement
Pré-expérimentation							
Cas A	Dispositif télématique "Urgence"	Non observé	Non observé	Non observé	Non observé	Non observé	Oui
Cas B	Cancer du sein triple négatif	Oui	Oui	Oui	Oui	Oui	Oui
Cas C	Placement publicitaire	Oui	Oui	Oui	Oui	Oui	Oui
Cas réalisés et détaillés							
Cas 1	Attrition assurance santé	Oui	Oui	Oui	Oui	Oui	Oui
Cas 2	Prévision d'activité	Oui	Oui	Oui	Oui	Oui	Oui
Cas 3	Prévention santé et prévoyance	Oui	Oui	Oui	Oui	Oui	Oui
Cas 4	Contrôle de non-conformité	Oui	Oui	Oui	Oui	Oui	Oui
Cas réalisés non détaillés							
Cas 5	Sinistres lourds en dommage aux biens	Oui	Oui	Oui	Oui	Oui	Oui
Cas 6	Prédiction de prix des agrumes	Oui	Oui	Non	Non	Non	Non
Cas 7	Multi-équipement	Oui	Oui	Oui	Oui	Oui	Oui

Figure 24 – Identification des phases CRISP_DM dans les études de cas

Le modèle de référence présente un degré de liberté adéquat dans le cadre de chaque phase, étant donné que les tâches préconisées ainsi que les résultats intermédiaires associés à ces tâches ne se veulent en soi ni exhaustives, ni obligatoires. Cependant, des inadéquations significatives sont constatées entre la définition des tâches du modèle CRISP_DM et la réalité observée.

2.1.1.1 Prise en compte tardive des usages

La première inadéquation concerne l'absence dans les activités du modèle de certaines tâches qui visent à inscrire l'usage métier en amont dans le processus : si le modèle comprend bien une tâche de priorisation des objectifs (formalisation de l'objectif prioritaire et des questions métier connexes relatives à cet objectif), il ne traite pas la priorisation des usages (formalisation de l'objectif opérationnel des réponses qui pourraient être apportées à l'issue du projet). Les auteurs du modèle (Shearer, 2000) citent l'exemple de l'attrition en banque pour décrire l'objectif prioritaire et les questions métier connexes : il est nécessaire de les compléter avec un objectif opérationnel (en développant ici son illustration sur le sujet de l'attrition) :

- **Objectif prioritaire** : « Retenir les clients actuels en prédisant quand ils sont enclins à se déplacer vers un concurrent »

- **Objectif opérationnel (usage prioritaire)** : Démontrer l'intérêt d'investir dans un projet de revue des coûts des guichets automatiques (le projet vise donc une réduction d'incertitudes quant à la pertinence de cet investissement parmi les leviers de rétention possibles)

- **Questions métier connexes** : « L'abaissement des coûts des guichets automatiques peut-il réduire significativement le nombre de clients à haute valeur qui partent ? »

Le modèle CRISP_DM redescend la première définition de l'usage des résultats dans la phase d'évaluation des résultats par le chef de projet. Or, cette étape impacte le projet en amont en termes de mobilisation d'interlocuteurs, de choix des outils, de sélection du périmètre et des variables identifiées comme « opérationnelles », de choix d'algorithmes, de priorisation de critères d'évaluation métier (performance opérationnelle) et de format de restitution des résultats pour l'évaluation métier. Par exemple, prendre en compte l'avis de l'utilisateur dès l'amont du projet peut éviter l'investissement dans des usages irréalistes, comme dans le cas du prix des agrumes. Elle impacte aussi le projet en aval, en accélérant la phase de déploiement, notamment en complétant la planification du pilotage avec la formalisation des indicateurs de mesure de performance opérationnelle comme dans le cas du déploiement des leviers de rétention des clients en assurance santé. Bien que l'exhaustivité des usages ne peut potentiellement émerger qu'à l'issue de la modélisation, ce qui peut conduire à un nouveau cycle d'analyse, la prise en compte tardive de l'usage dans le modèle CRISP_DM est identifié comme un risque de non opérationnalité des résultats directs, et n'est pas appliquée sur le terrain. Le fait d'écarter l'usage des phases initiales du modèle diffère par ailleurs la prise en compte, potentiellement native, de certaines notions juridiques comme l'anonymisation des données. Par exemple, bien qu'un projet data puisse être mené sur des données anonymisées, si l'usage d'un résultat anonyme ne présente pas d'intérêt opérationnel il risque de faire l'objet d'une dévalidation dans la phase de déploiement. La prise en compte de l'usage doit alors être effectuée dès la compréhension métier (comment voulez-vous utiliser les résultats ?), impacte la précision des questions métier, et constitue sur le terrain une tâche intermédiaire entre la formulation des objectifs et de la question connexe métier.

2.1.1.2 Facilitation insuffisante de l'interprétation des résultats

La seconde inadéquation est observée dans l'absence de proposition de format de restitution des résultats dans la phase de compréhension métier. Or, cette tâche est bien observée sur le terrain : son utilité est d'alimenter les échanges entre les acteurs afin d'être alignés sur les critères d'évaluation métier, et d'anticiper la production de support qui permet de faire cette évaluation. Le modèle CRISP_DM précise bien une tâche de synthèse des résultats sous l'angle des atteintes des objectifs métier : cette synthèse est réalisée par le data analyste. La réalité est bien différente : l'équipe produit un format de restitution des résultats, par exemple une

représentation de la zone de confiance sur une série temporelle de prévision d'activité, et un format de synthèse des résultats (indicateurs d'évaluation de l'atteinte des objectifs). Ces formats garantissent l'intelligibilité des résultats et accélèrent la phase d'évaluation, réalisée conjointement avec les experts métier. Le développement du format de restitution des résultats pour leur évaluation constitue une tâche en soi, dépendante des types de résultats (prédictions multidimensionnelles, les profils riches...), des critères d'évaluation métier, et de la nature du support optimisant l'activité d'évaluation métier (rapport, maquette Excel, Data Visualisation ergonomique permettant l'interaction avec les résultats, une application métier testée sur le terrain...). Ce changement méthodologique génère une nouvelle tâche, absente du modèle CRISP_DM : il s'agit de la sélection des résultats opérationnels au cours de la phase d'évaluation, fortement liée à l'usage, qu'il soit anticipé en amont, ou bien émergeant à l'issue de la modélisation.

2.1.1.3 Insuffisance de la tâche de sélection des données

La troisième inadéquation concerne la tâche de sélection des données, appartenant dans le modèle à l'activité de préparation des données. Selon le modèle CRISP_DM, il s'agit de « décider des données qui seront utilisées pour l'analyse selon plusieurs critères, y compris leur pertinence pour les objectifs de l'exploration de données, mais aussi les contraintes techniques, le volume de données ou les types de données. Par exemple, alors que l'adresse d'un individu peut être utilisée pour déterminer de quelle région l'individu est originaire, les données de numéro de rue peuvent être éliminées pour réduire la quantité de données qui doivent être évaluées. Une partie du processus de sélection des données devrait impliquer d'expliquer pourquoi certaines données ont été incluses ou exclues. C'est aussi une bonne idée de décider si un ou plusieurs sont plus importants que d'autres ».

L'observation des cas terrain contredit non pas la description, mais le positionnement de cette tâche dans le processus global : en effet, la définition de la méthode de rationalisation des critères d'inclusion et d'exclusion des variables a lieu dès la phase de compréhension métier pour guider le plan de collecte, en particulier lors de l'établissement de la cible des analyses, dans la mesure où le cadrage du champs des données constitue en soi une qualification des données, jugées « utiles » à l'analyse. Par ailleurs, les critères d'inclusion et d'exclusion sont revus lors de la vérification de la qualité des données. Au cours de cette phase, certains critères sont directement utilisés : les critères d'exclusion statistique permettent de réduire le champs des données exploitables afin d'accélérer le processus de compréhension métier des données,

les critères d'inclusion métier structurent la revue du plan de collecte et l'anticipation de la mise en qualité, et enfin les critères techniques (clés) permettent de valider la possibilité de construire la matrice d'apprentissage dans le cadre de l'usage de données issues de plusieurs sources de données, et impactent ainsi également le plan de collecte. Ensuite, les critères d'inclusion et d'exclusion sont ajustés selon la stratégie de modélisation et appliqués en amont de la construction de la matrice d'apprentissage, comme évoqué dans le modèle CRISP_DM. Puis, la matrice d'apprentissage fait elle-même l'objet d'un contrôle complet selon les critères d'inclusion et exclusion (en effet, la dérivation de certaines variables conduit à créer de nouvelles variables qui doivent respecter les critères de sélection initiaux), avant d'évoluer itérativement tout le long du processus de modélisation (découverte d'incohérences de sens ou statistiques au cours de la phase d'évaluation, tests de sensibilité statistiques...) et d'évaluation (prise en compte de contraintes opérationnelles et tests de sensibilité métier). Ainsi, la version initiale de l'évaluation des critères d'inclusion et d'exclusion est établie dès la première phase du projet (critères d'utilité, d'accessibilité des données, de perception de leur qualité avant l'audit…), et la version définitive à l'issue de la phase d'évaluation (voir exemples en annexe dans la description du Databook).

L'impact des critères d'inclusion et d'exclusion, complété par des éléments qualitatifs sur les données, permet de constituer une base de métadonnées complète dont d'utilité consiste à documenter le projet, de prioriser des pistes d'optimisation des résultats, de guider la maintenance et de servir de support de capitalisation pour les usages non couverts par le projet data. Cette base de métadonnées de référence comprend dans les cas observés la description des traitements effectués sur les données, c'est-à-dire la description de chaque variable présente dans la matrice d'apprentissage non censurée suite à l'application de critères d'inclusion liés aux besoins du modèle et à l'usage des résultats. Cette documentation permet de contribuer à la traçabilité de la prise de la décision au cours de l'usage déployé, sur la partie des données qui est utile à l'exploitation de l'usage, mais aussi de réutiliser la matrice d'apprentissage pour d'autres modèles et d'autres usages potentiels, ce qui permet de dégager des synergies futures.

Ainsi, **trois catégories d'inadéquations** sont détectées grâce à la comparaison des tâches du modèle CRISP_DM et des tâches observées sur les projets terrain (voir Figure 25). Ces inadéquations pointent les limites de fond du modèle, c'est-à-dire le **manque de prise en compte de l'usage**, **la négligence du format de restitution des résultats pour leur appropriation**, et **l'incomplétude de la capitalisation liée à la qualité des données**.

Modèle CRISP DM et études de cas : analyse comparative							
Phases CRISP DM	**Tâches**	**Résultats**	**Attrition assurance santé**	**Prévision d'activité**	**Prévention santé et prévoyance**	**Contrôle de non-conformité**	**Catégorie de l'inadéquation**
Compréhension métier	Déterminer les objectifs métier	Contexte	Réalisé	Réalisé	Réalisé	Réalisé	
		Priorisation des usages	Nouveau	Nouveau	Nouveau	Nouveau	Usage des résultats
		Objectifs métier	Réalisé	Réalisé	Réalisé	Réalisé	
		Critères de succès métier	Réalisé	Réalisé	Réalisé	Réalisé	
	Evaluer la situation	Inventaire des ressources	Réalisé	Réalisé	Réalisé	Réalisé	
		Exigences, hypothèses et contraintes	Réalisé	Réalisé	Réalisé	Réalisé	
		Risques et contingences	Réalisé	Réalisé	Réalisé	Réalisé	
		Terminologie	Réalisé	Réalisé	Réalisé	Réalisé	
		Coûts et bénéfices	Réalisé	Réalisé	Réalisé	Réalisé	
	Déterminer les cibles d'analyse	Cible d'analyse Data Mining	Réalisé	Réalisé	Réalisé	Réalisé	
		Critères de succès Data Mining	Réalisé	Réalisé	Réalisé	Réalisé	
	Sélectionner les données	Méthode de rationalisation de l'inclusion / excusion des données	Modifié	Modifié	Modifié	Modifié	Qualité des données
	Définir le format de restitution des résultats	Maquette initiale de datavisualisation	Nouveau	Nouveau	Nouveau	Nouveau	Format de restitution des résultats
	Produire le plan du projet	Plan projet	Réalisé	Réalisé	Réalisé	Réalisé	
		Evaluation iniitale des outils et techniques	Réalisé	Réalisé	Réalisé	Réalisé	
Compréhension des données	Collecter les données initiales	Rapport de collecte des données initiales	Réalisé	Réalisé	Réalisé	Réalisé	
	Décrire les données	Dictionnaire des données	Réalisé	Réalisé	Réalisé	Réalisé	
	Explorer les données	Rapport d'exploration	Réalisé	Réalisé	Réalisé	Réalisé	
	Contrôler la qualité des données	Rapport de qualité des données	Réalisé	Réalisé	Réalisé	Réalisé	
Préparation des données	Collecter l'ensemble des données	Description des données collectées	Réalisé	Réalisé	Réalisé	Réalisé	
	Nettoyer les données	Rapport de nettoyage des données	Réalisé	Réalisé	Réalisé	Réalisé	
	Construire la matrice	Attributs dérivés	Réalisé	Réalisé	Réalisé	Réalisé	
		Enregistements générés	Réalisé	Réalisé	Réalisé	Réalisé	
	Intégrer les données	Matrice d'analyse agrégée	Réalisé	Réalisé	Réalisé	Réalisé	
	Formater les données	Matrice d'analyse reformatée pour les analyses (caractères spéciaux, vides,...)	Réalisé	Réalisé	Réalisé	Réalisé	
	Documenter la préparation des données	Description des traitements de données	Nouveau	Nouveau	Nouveau	Nouveau	Qualité des données
Modélisation	Sélectionner les techniques de modélisation	Techniques de modélisation	Réalisé	Réalisé	Réalisé	Réalisé	
		Critères d'évaluation de la modélisation	Réalisé	Réalisé	Réalisé	Réalisé	
	Générer le concept de test	Test Design (processus d'apprentissage et de test)	Réalisé	Réalisé	Réalisé	Réalisé	
	Construire le modèle	Parametrages	Réalisé	Réalisé	Réalisé	Réalisé	
		Modèle(s)	Réalisé	Réalisé	Réalisé	Réalisé	
		Description du (des) modèle(s)	Réalisé	Réalisé	Réalisé	Réalisé	
	Evaluer le modèle (du point de vue statistique)	Evaluation du modèle	Réalisé	Réalisé	Réalisé	Réalisé	
		Paramétrages révisés	Réalisé	Réalisé	Réalisé	Réalisé	
Evaluation	Développer le format de restitution des résultats	Support de présentation des résultats	Nouveau	Nouveau	Nouveau	Nouveau	Format de restitution des résultats
	Evaluer les résultats du modèle	Evaluation des résultats de Data Mining par rapport aux critères de succès métier	Réalisé	Réalisé	Réalisé	Réalisé	
	Sélection de résultats opérationnels	Résultat opérationnel (ajusté)	Nouveau	Nouveau	Nouveau	Nouveau	Usage des résultats
		Modèle(s) approuvés	Réalisé	Réalisé	Réalisé	Réalisé	
	Revoir le process	Revue critique du process	Réalisé	Réalisé	Réalisé	Réalisé	
	Déterminer les étapes suivantes	Liste des actions possibles	Réalisé	Réalisé	Réalisé	Réalisé	
		Décision	Réalisé	Réalisé	Réalisé	Réalisé	
Déploiement	Planifier le déploiement	Plan de déploiement	Réalisé	Réalisé	Réalisé	Réalisé	
	Planifier le pilotage et la maintenance	Plan de pilotage et de maintenance	Réalisé	Réalisé	Réalisé	Réalisé	
	Produire le rapport final	Rapport présentation finale	Réalisé	Réalisé	Réalisé	Réalisé	
	Revue du projet	Retour d'expérience et documentation	Réalisé	Réalisé	Réalisé	Réalisé	

Figure 25 – Synthèse comparative entre le modèle CRISP_DM et la réalité terrain sous l'angle de l'observation des tâches et des résultats des tâches.

2.1.2 Critique des dépendances et de la cyclicité

Au-delà des phases, des tâches et des résultats des tâches, le modèle CRISP_DM inclut un ensemble de dépendances entre les phases. Il s'agit des « dépendances les plus fréquentes et les plus importantes » (Shearer, 2000), pouvant être à sens unique (enchainement séquentiel de deux phases) ou bien à double sens (aller-retour possible entre deux phases, constituant des blocs itératifs). Le modèle capte par ailleurs la nature cyclique du projet data, liée à l'apprentissage au cours du processus et à partir de la solution déployée qui permettent de reformuler des questions métier de façon nouvelle ou plus précise (voir Figure 26). La suggestion de dépendance et de cyclicité a l'avantage d'être facilement communicable, et s'adresse aussi bien aux interlocuteurs novices qu'aux experts pour expliquer le cheminement global des phases du projet et la nécessité d'établir des liens entre les différentes tâches. La conscience de ces itérations génère une mobilisation flexible de ressources nécessaires, et notamment des compétences variées.

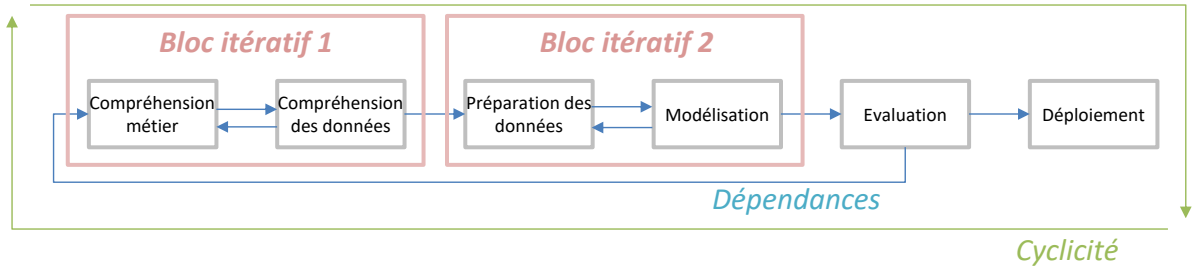


Figure 26 – Itérations et cyclicité au sein du processus CRISP_DM

Cependant, malgré sa clarté apparente et sa vocation à servir à la conduite d'un projet data, la représentation de dépendance et de cyclicité du modèle CRISP_DM présente l'inconvénient majeur d'absence d'ancrage temporel délimité (jalon de début, durée de réalisation, jalon de fin) des différentes phases, et du projet au global, ce qui la rend difficilement exploitable dans le cadre de la gestion de projet classique. Au-delà de leur absence, le modèle CRISP_DM n'est pas « confortable » pour l'établissement des jalons dans la mesure où le séquencement ne correspond pas à la réalité observée.

Tout d'abord, les projets sur le terrain pointent un ensemble de superpositions entre les phases dans le temps. Par exemple, l'aller-retour est difficilement observable entre les phases de compréhension métier et compréhension des données (premier bloc itératif) : les tâches de ce

bloc itératif sont en effet réalisées simultanément à partir du moment où les premières données sont accédées. De même, les phases de préparation des données et la modélisation (second bloc itératif) sont superposées dès que la première version de la matrice d'analyse est établie. Ces deux superpositions peuvent être admises comme captées dans le modèle CRISP_DM sous forme de dépendances itératives : il s'agirait alors d'un simple ajustement sémantique entre les notions « itération » et « simultanéité ». Cependant il ne s'agit pas de la seule superposition observée.

Le premier bloc itératif peut se superposer avec le second dans le cas d'une intégration progressive de données complémentaires, liée notamment à l'incertitude quant au besoin d'en intégrer et aux délais nécessaires de collecte ou de mise en qualité : en effet, un enrichissement de la matrice d'apprentissage de base est possible jusqu'à ce que le modèle soit satisfaisant, c'est-à-dire à l'issue de la phase d'évaluation de génération de valeur. Par ailleurs, la tâche d'évaluation du modèle, incluse par nature dans la phase de modélisation, peut se superposer avec la phase d'évaluation métier. Cette superposition a lieu notamment en cas de déficience du modèle analytique, comme pour le cas d'analyse de multi-équipement : il s'agit d'un processus d'investigation qui permet d'identifier si la déficience provient du modèle (l'issue est alors la correction de la matrice d'apprentissage ou du modèle) ou bien d'un aspect de la réalité métier captée par le modèle (l'issue est alors la fin de projet ou la reformulation de la question métier). Enfin, la phase de déploiement semble être conditionnée par la validation de l'intérêt des résultats au cours de l'évaluation. Or, l'émergence de résultats, exploitables et à forte valeur, peut avoir lieu à chaque phase du projet : il s'agit d'usages indirects liés aux connaissances découvertes, ou alors d'usages directs issus des phases intermédiaires (accès à de nouvelles données, construction de nouvelles structures et référentiels de données...). Ce dernier point renvoie à l'intégration de la notion de l'usage en amont du projet, comme vu précédemment, et aux réajustements successifs du projet. La préparation du déploiement se superpose aussi avec les activités de compréhension métier et d'appropriation des résultats pour les usages.

Ces superpositions des phases, en tant qu'actions de production inscrites dans le temps, ne correspondent pas à un écart de la méthode : elles sont justifiées et créatrices de valeur, et une séparation des tâches peut conduire à des dérives. Le modèle CRISP_DM doit ainsi faire l'objet d'une déstructuration au service de la gestion de projet afin de mettre en évidence les superpositions de phases (voir Figure 27), ainsi que l'accumulation de l'impact du projet sur les usages.

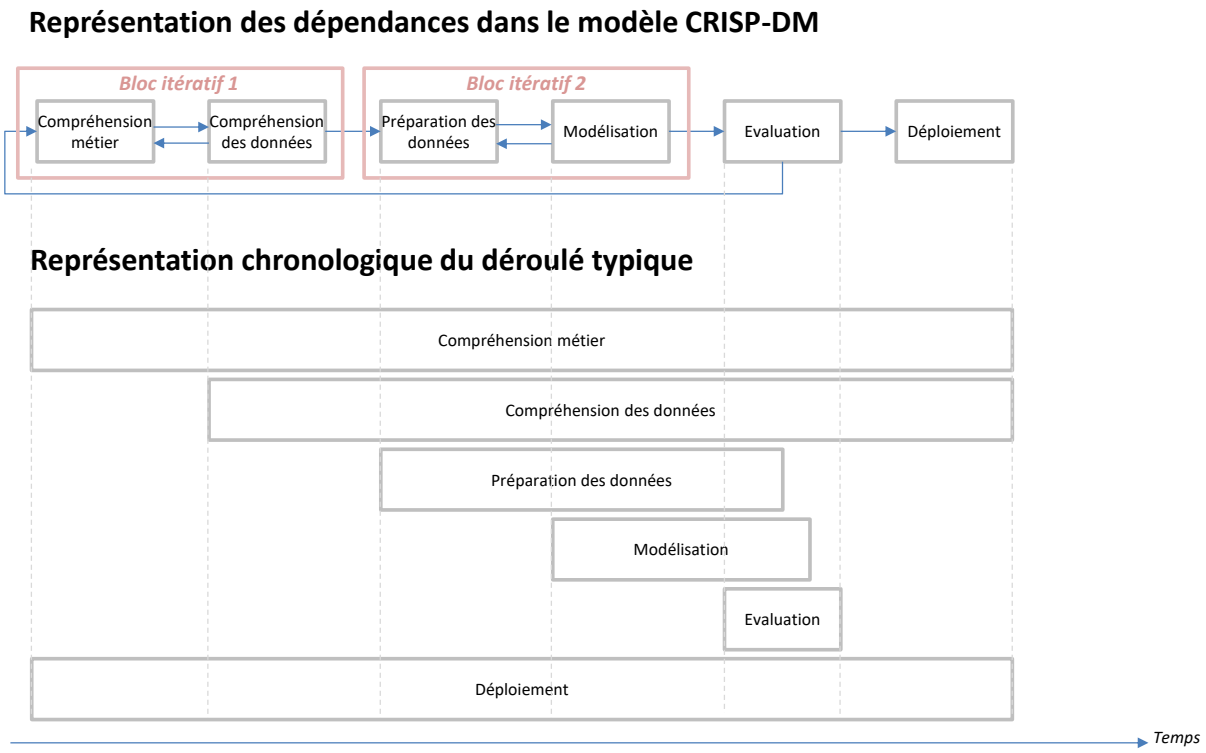


Figure 27 – Synthèse comparative entre le modèle CRISP_DM et la réalité terrain sous l'angle de l'observation des superpositions chronologiques des phases

Par ailleurs, si le modèle CRISP_DM met bien en évidence l'estimation initiale des bénéfices et des risques, il ne comporte pas de relations de dépendance impactant ces estimations au cours du projet. Il s'agit d'une réévaluation des bénéfices et des risques au fur et à mesure que les incertitudes sont levées. Seules les dépendances en termes de ressources sont suggérées, chaque tâche se référant à une action de production qui présente un certain coût. Or, seule la conjonction des 3 éléments (bénéfices, incertitudes et ressources) permet d'assurer la clôture des tâches, en évitant l'enlisement dans des tâches réitérées sans valeur ajoutée démontrée. Au-delà de l'impact négatif de cette incomplétude du modèle sur la gestion de projet par le coût, ce manque de liens de dépendance causale ne donne pas les clés d'une gestion de projet par le résultat. L'inconfort du modèle est attribuable à l'arrêt de la phase de compréhension métier et données, et à sa séparation de la notion d'usage, renvoyé en fin de cycle, ce qui empêche la prise en compte des jalons intermédiaires jusqu'au bout de la phase d'évaluation. Ce séquencement produit un ensemble de redondances itératives qui prêtent à confusion.

Enfin, en absence de jalonnement clair, la méthode projet induite par le modèle CRISP_DM ne permet pas d'avoir des éléments de mesure de la performance du projet, que ce soit pour le

contrôle de productivité interne de chaque phase, du nombre d'itérations, ou de la valeur externe des résultats. Cela nuit à l'industrialisation de la méthode, c'est-à-dire à son application à une échelle plus large qu'un seul projet data en soi et ses propres itérations. Or, ce besoin de méthode de gestion de portefeuilles de projets data est croissant sur le terrain, que ce soit dans le cadre de l'accélération de l'innovation ou bien pour la gestion des ressources techniques et humaines dédiées.

En résumé, le modèle doit être ajusté en termes de dynamique temporelle des activités, complété avec un **jalonnement d'instances de réévaluation** des bénéfices, des risques et des ressources, et rendu ainsi plus commode pour une gestion de projet (classique, concourante, agile ou autre) transposable à une gestion de portefeuille de projets.

2.2 Proposition de modèle de dispositif de projet data : Brizo_DS

Le modèle de dispositif de projet data proposé à l'issue de ces travaux de recherche est fondamentalement **orienté vers la génération de valeur par l'exploitation des usages qu'il permet de clarifier**. Le modèle va au-delà du processus de production analytique pour décrire le dispositif complet, comprenant l'ensemble des agencements et dynamiques entre acteurs humains et non humains, et inclut ainsi le processus de production pour l'ancrer dans un cadre d'évaluation et de mesure de la valeur par un jalonnement d'instances de Médiation Homme-Données qui visent la maturation des usages, qu'ils soient directs ou indirects. Le dispositif projet data apparait ainsi comme un investissement, un vecteur de réduction des incertitudes qui caractérisent la valeur générée par les usages, jusqu'à ce que leur niveau soit jugé acceptable pour la mise en exploitation. Le modèle est ainsi caractérisé par (voir Figure 28) :

- une orientation sur l'**usage** (génération de valeur par l'exploitation d'usages directs et indirects)

- une cadre d'**arbitrage** (évaluation et mesure des bénéfices, des ressources et des incertitudes liées au projet et à l'usage qu'il vise à construire, sous la forme d'un jalonnement transversal par des instances de Médiation Homme-Données)

- un **dispositif « projet data »**, doté de ressources agencées et d'une dynamique décrite par le modèle de processus proposé dans ces travaux (processus CRISP_DM ajusté)

Ce code couleur sert de repère dans l'ensemble des figures de qui illustrent les propositions de ces travaux de recherche.

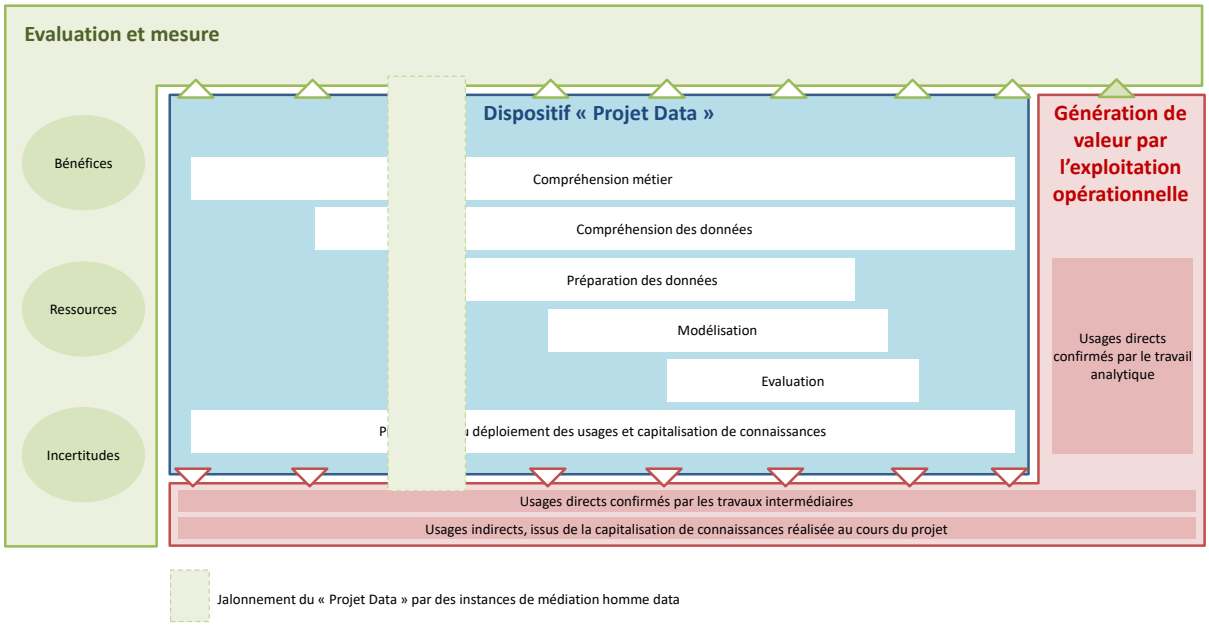


Figure 28 – Brizo_DS, modèle de dispositif projet data

Au-delà des initiales des 3 indicateurs de référence (Bénéfices, Ressources et Incertitudes), le nom du modèle est inspiré de la déesse grecque Brizo, porteuse de rêves prophétiques et protectrice des marins : l'art de prédire l'avenir par les songes est en effet un atout de taille pour un projet Data Science, par nature exploratoire dans un environnement incertain, et visant à arriver à bon port, c'est-à-dire sur un usage générateur de valeur, anticipé ou non.

2.2.1 Orientation sur usage

L'orientation du dispositif sur l'usage signifie une prise de conscience collective que toute activité du projet doit avoir une visée opérationnelle intelligible. Elle s'enracine sur la capacité à reconnaitre ce qui est utile à l'entreprise et aux utilisateurs, et ce à tous les stades du projet. Le choix du terme est délibérément réalisé au profit des Sciences de Gestion afin que le modèle reste intelligible et non ambigu pour les acteurs impliqués en entreprise : il s'agit de la même définition qu'une « pratique » pour un professionnel de l'information, c'est-à-dire un usage contextualisé doté d'une finalité d'exploitation.

2.2.1.1 Nouveauté des usages

Si la nature les usages possibles est marquée par une grande diversité, dépendante des secteurs (cœur d'activité, marché…), des entreprises (positionnement stratégique, choix tactiques…), des fonctions exécutantes (productives, support…) et des utilisateurs, ils restent parfaitement comparables grâce à la mobilisation d'une définition plus générique. Un usage est une production, à partir de ressources (Inputs), au cours d'activités déterminées et mesurables (Process), de produits ou de connaissances (Outputs) (Curry et al., 2006; Samsonowa et al., 2009). Les ressources comprennent l'ensemble des représentations sociales, d'outils ou encore de capital de connaissances préalables. Un usage peut dans ce cadre être considéré comme nouveau s'il se distingue des précédents par un ou plusieurs éléments issus de ces trois catégories.

La cartographie détaillée des impacts des projets data (voir synthèse en Figure 29 et le détail en voir Annexe 8 - Modèles Input-Process-Output détaillés) démontre bien l'existence de nouveaux usages selon cette définition, car dans tous les projets au moins l'une des trois catégories est impactée :

- **Ressources (Inputs)** : Ajout des ressources nécessaires à la réalisation des activités du système traitant sous forme de données et informations, de compétences data, parfois de nouveaux outils.

- **Activités (Process)** : Modification des activités opérées par le système traitant (changement de processus des activités d'anticipation et de prise de décision, de production ou de reporting) et parfois la création de nouvelles activités. Aucune disparition d'activités historiques n'a été constatée dans ce cadre de cette étude.

- **Résultats (Outputs)** : Génération de nouveaux produits informationnels destinés à alimenter l'activité d'autres acteurs, création d'éléments de communication et dans certains cas création de produits et services nouveaux ou de meilleure qualité.

La valeur générée par ces nouveaux usages peut alors faire l'objet d'un processus de mesure à travers une boucle de feedback qui permet d'évaluer l'impact sur l'**efficacité** du système traitant, c'est-à-dire de la communauté d'acteurs qui exploite l'usage. Un nouvel usage vise

alors l'optimisation de l'efficacité, voire la création d'un sous-système dédié[43]. L'efficacité peut être interne ou externe. L'efficacité interne est mesurée par des indicateurs de performance des activités comme les gains en productivité (gain de temps ou de réactivité, baisse du coût des ressources, l'accélération des activités d'investigation et diminution du risque d'investigations infructueuses, accélération des prises de décision grâce un accès simplifié aux informations utiles…). L'efficacité externe est mesurée par des indicateurs de performance des résultats des activités (amélioration de la qualité des produits ou des services fournis, décisions plus éclairées, diffusion de connaissances et de produits informationnels plus pertinents…). L'amélioration de l'efficacité du système traitant produit des **bénéfices** à l'échelle de l'entreprise. Ils sont mesurés par des indicateurs de performance stratégiques. Ces bénéfices (Atamer & Calori, 2003) sont de l'ordre financier (croissance, baisse des coûts, obtention de financement, économie d'investissements inutiles, baisse de risques…) ou non financier, à travers la plus-value sociale (développement de partenariats, satisfaction client, capacité à innover, voire la diminution de la mortalité à l'échelle de la société grâce à la santé personnalisée ou la télématique).

Or, ces bénéfices peuvent aussi être générés de façon indépendante de l'efficacité du système traitant (voir Figure 29). Notamment, la capacité d'innovation de l'entreprise est tout à fait identifiée à travers un impact sur les compétences data, transparent pour les indicateurs d'efficacité (Cas 6). Inversement, l'absence de nouvelles compétences data ne semble pas créer de capacité d'innovation (Cas 7). Cette situation complexifie une évaluation linéaire et causale des bénéfices, habituellement utilisée pour piloter les différentes fonctions qui composent l'entreprise et leur investissement de ressources dans les projets (Lebas, 1995). Elle mobilise plutôt un mécanisme d'investissement dans un avantage informationnel, que ce soit au sens polémique (détention d'informations non uniques) ou concurrentiel (détention d'informations uniques). Cette mise en perspective permet de mieux situer les projets data, à la fois comme générateurs direct de capacité d'innovation par l'apport de compétences data, et comme générateurs d'usages exploitables et de capital de connaissances, qui à leur tour seront transformés en usages exploitables si la capacité d'innovation est suffisante.

[43] les cas de création de nouveaux business modèles n'ont pas été abordés dans ces travaux, mais restent compatibles avec le modèle proposé

Synthèse des impacts de projets data												
		Cas A	Cas B	Cas C	Cas 1	Cas 2	Cas 3	Cas 4	Cas 5	Cas 6	Cas 7	
Element impacté	Nature de l'impact	Dispositif Télématique Urgences	Cancer du sein triple négatif	Placement publicitaire	Attrition assurance santé	Prévision d'activité	Prévention santé et prévoyance	Contrôle de non-conformité	Sinistres lourds en dommage aux biens	Prédiction de prix des agrumes	Multi-équipement	Nombre de cas consernés (sur 10)
Inputs	Données/information	oui	oui	oui	oui	oui	oui	oui	oui	non	oui	9
	Compétences (data)	non	non	oui	oui	oui	oui	oui	oui	oui	non	7
	Outils	oui	non	oui	non	oui	oui	non	non	non	non	4
Process	Nouvelles activités	oui	non	oui	non	oui	oui	non	non	non	non	4
	Modification des activités	oui	oui	oui	oui	oui	oui	oui	oui	non	oui	9
	Abandon d'activité	non	non	non	non	non	non	non	non	non	non	0
Output	Produits / Services	oui	non	oui	non	non	non	non	non	non	non	2
	Produits informationnels	oui	oui	oui	oui	oui	oui	oui	oui	non	oui	9
	Communication	oui	oui	non	oui	oui	oui	oui	non	non	non	6
Impact autres acteurs		oui	oui	oui	oui	oui	oui	oui	oui	non	oui	9
Outcomes	Croissance	oui	oui	oui	oui	oui	non	oui	non	non	oui	7
	Rentabilité	oui	non	non	oui	oui	oui	oui	oui	non	non	6
	Plus-value sociale	oui	oui	non	oui	oui	oui	oui	oui	oui	non	8
	dont capacité à innover	Non observé	Non observé	Non observé	oui	oui	oui	oui	oui	oui	non	6

Figure 29 – Synthèse comparative des impacts des projet data

Pour une aide à la lecture, se référer aux impacts détaillés des cas en Annexe 8 - Modèles Input-Process-Output détaillés. Par exemple, le cas 1 (voir Figure 62 – Modélisation IPO : Impacts du projet « Attrition en Assurance Santé ») est marqué par un impact sur les ***Inputs*** *(nouvelles compétences sur le processus Data Science et connaissances en statistiques, nouvelles données comme les profils de clients à risque d'attrition, les scores de risque, mais aussi les métadonnées du Databook…), sur le* ***Process*** *(priorisation de l'analyse des hypothèses de travail pour la conception de leviers opérationnels…) et sur les* ***Outputs*** *(rapport d'expert sur les actions stratégiques à inclure dans le Programme relation client, retour d'expérience partagé en interne sur les bonnes pratiques d'un projet data…), et vise un impact sur d'autres acteurs (comme les agents locaux) et sur les* ***Outcomes*** *en termes de réduction d'attrition (impact sur la croissance et la rentabilité) ou de satisfaction client (impact sur la plus-value sociale) et de capacité à innover.*

Les motivations du lancement d'un projet data sont déterminées par le contexte de l'entreprise et par sa capacité à innover. Elles peuvent être basées sur des besoins opérationnels plus ou moins définis, mais aussi sur une dynamique d'isomorphisme, de mimétisme des acteurs concurrents, une perception de manque à gagner, une recherche d'inspiration par la fouille des données, une opportunité engendrée par l'écosystème Big Data, une volonté de communication externe ou interne sur l'activité data, ou encore une montée en maturité interne à travers l'apprentissage. La diversité, et souvent le manque de clarté de ces motivations dans les entreprises qui se cherchent à l'heure du buzz (voir Annexe 9 - L'internalisation des usages dérivant du recours à l'Intelligence Artificielle dans les entreprises), distingue ces projets exploratoires des projets plus traditionnels initiés dans le cadre d'un besoin opérationnel défini. En effet, si les projets traditionnels visent en priorité des usages directs aux bénéfices

anticipables, les projets Data Science font face à un niveau d'incertitudes élevé sur cet usage, et peuvent privilégier le processus de découverte de nouveaux usages ou la capitalisation de connaissance en vue de déploiement d'usages indirects.

2.2.1.2 Usages directs et indirects

Les nouveaux usages peuvent être anticipés en amont du projet, ou bien être découverts au cours du projet dans le cadre de la fouille des données. Ils sont distingués entre les usages **directs** et **indirects**. Les usages directs sont issus d'un travail de préparation du déploiement au cours du projet data. Ils résultent des travaux de production analytique, livrés à l'issue du projet, ou alors des travaux de production intermédiaire, livrés tout le long du projet (documents, données, métadonnées, informations…). Ces travaux permettent de confirmer (ou d'infirmer) la pertinence d'un usage en apportant des preuves analytiques. A lumière de ces preuves et de la confiance qui leur sont accordées, l'usage est jugé comme acceptable et traduit en termes opérationnels. Contrairement aux usages directs, les usages indirects sont assimilables à des externalités positives. Ils ne font pas l'objet d'un travail de préparation du déploiement, cependant la **capitalisation de nouvelles connaissances** réalisée au cours du projet permet de percevoir la possibilité d'une transformation future de cette connaissance en usages bénéfiques. Cette conceptualisation des usages constitue un facteur de différenciation entre le processus traditionnel de Data Mining, qui ne vise que le déploiement des usages directement confirmés par les travaux analytiques, en omettant les usages indirects et les usages directs issus des travaux intermédiaires du processus (voir Figure 30).

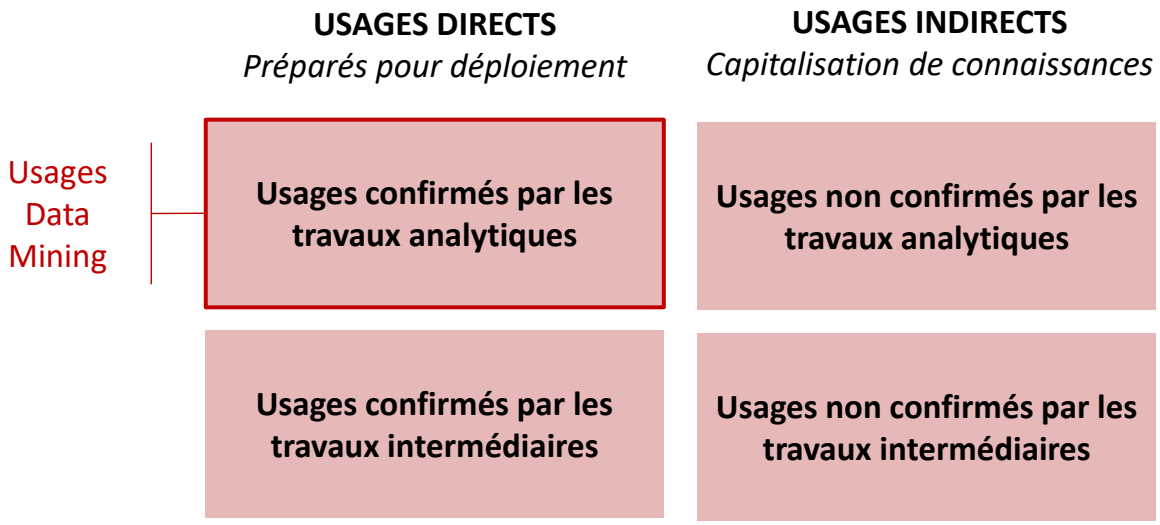


Figure 30 – Emergence et confirmation d'usages à travers un projet Data Science

Par exemple, les usages confirmés par les travaux analytiques sont les leviers de rétention des clients à risque d'attrition, de contrôle de contrats par thématiques, ou encore l'accélération de la production des prévisions d'activité ou la proposition de nouveaux services comme la prise en charge par l'assistance en cas d'urgence. Les usages confirmés par les travaux intermédiaires concernent l'exploitation de nouveaux référentiels de données ou de modèles conceptuels (par exemple, l'exploitation de la matrice d'apprentissage du projet attrition comme base client 360° servant à alimenter la connaissance client, indépendamment de l'attrition), ou encore la mise en place de feuilles de route de mise en qualité des données. Les usages non confirmés par les travaux analytiques comprennent typiquement l'élimination de la volonté de prédire le prix des agrumes à 1 an : il est alors nécessaire de capitaliser cette connaissance pour éviter un investissement futur dans cet usage. Enfin, les usages non confirmés par les travaux intermédiaires concernent la capacité à réaliser de nouveaux projets (transfert de compétences, gains de productivité sur les projets data suite à une qualification des données ou une montée en maturité…) et tous les usages éliminés au fur et à mesure de l'avancement des projets, notamment suite aux arbitrages sur les périmètres d'analyse : lorsqu'un périmètre a été identifié comme moins prioritaire, il pourra éventuellement donner lieu à un nouveau projet pour poursuivre sa confirmation, à condition que les raisons de cet arbitrage aient bien été capitalisées au cours du projet sous la forme d'un capital de savoirs.

2.2.1.3 L'interaction comme vecteur de convergence sur les usages

A la lumière de l'état de l'art, pointant la nécessité de clarifier la nature des interactions indispensables à la convergence sur un résultat utile, ainsi que de la distinction confirmée entre les usages directs et indirects, il est possible de synthétiser les flux informationnels qui marquent un projet data sous la forme d'un modèle conceptuel dédié (voir Figure 31). Il s'agit d'un schéma de convergence entre la dimension intellectuelle du projet (savoirs préexistants et représentations sociales) et sa dimension matérielle (données brutes) pour générer d'abord des informations contextualisées, puis des connaissances issues d'un travail d'interprétation. Ces connaissances ont alors une portée opérationnelle dans le cadre de la prise de décision au cours d'un usage donné. Cependant, les connaissances peuvent avoir aussi une portée générative de savoirs (il s'agit de la capitalisation de savoirs, voire de nouvelles représentations sociales, au cours d'un processus d'appropriation). Par ailleurs, l'ensemble du processus de traitement qui permet d'aboutir aux informations contextualisées génère aussi les inscriptions qui se cumulent aux données brutes préexistantes. Enfin, l'usage, en tant que seul levier de génération de valeur, permet de propager celle-ci à la fois au capital de savoirs et de données. Ce modèle conceptuel

permet de s'affranchir d'une vision trop linéaire du travail de construction cognitive d'un résultat orienté sur l'usage direct, et d'envisager une construction progressive de capital intellectuel et matériel pour des usages futurs. Elle suggère ainsi deux finalités d'interactions possibles : la convergence sur un usage direct, et l'apprentissage dans le cadre de la capitalisation de savoirs.

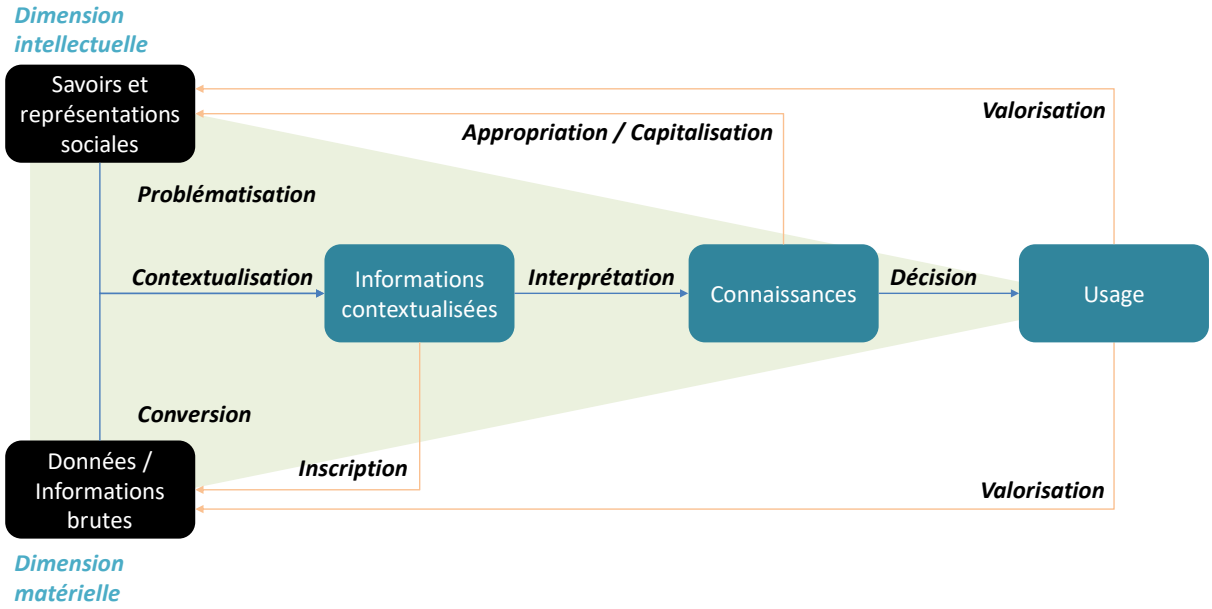


Figure 31 – Modèle conceptuel des interactions visant la convergence sur les usages directs et la génération des savoirs à partir des données.

La première finalité correspond aux interactions d'alignement sur l'usage direct : elles sont fortement dépendantes de la capacité du dispositif à converger au cours de la contextualisation du projet et de l'interprétation de ces résultats. Il s'agit par exemple des phases de définition et d'évaluation des critères de performance d'un modèle, comme, dans le cas 2, les zones de confiance dans les prédictions d'activité (convergence entre la définition statistique et la représentation graphique). La seconde vise en priorité l'enrichissement du capital social de savoirs et des données exploitables pour de nouveaux usages. Elle s'appuie sur un travail d'inscription de nouvelles informations contextualisées, et d'appropriation de l'ensemble des connaissances acquises au cours du projet (voir Figure 32). Si les interactions d'alignement garantissent la qualité externe du projet data et visent des gains de valeur par la prise de décision directe, l'apprentissage social garantit la qualité interne des projets (data ou non) suivants, dans la mesure où il améliore la capacité à percevoir des incertitudes, c'est-à-dire les évènements qui peuvent impacter négativement une activité ou un résultat d'activité.

		Finalité de l'interaction	
		Alignement sur le résultat *(usage direct)*	**Apprentissage social** *(usage indirect)*
Nature de l'interaction	**Contextualisation**	Capacité à poser un problème métier pertinent et à convertir les données brutes en informations liées à la problématique	Capacité à transcrire cette nouvelle contextualisation des informations sous forme d'inscriptions réutilisables pour d'autres usages
	Interprétation	Capacité à traduire les informations contextualisées en connaissances utiles à une prise de décision dans le cadre de l'usage visé	Capacité à traduire les informations contextualisées en connaissances capitalisables pour générer de nouveaux usages

Figure 32 – Nature et finalité des flux informationnels principaux

La génération de connaissances, en tant que résultat qualitatif du projet data et facteur de progrès au-delà de la recherche de la performance, apparait sur le terrain dès l'étude de cas préliminaires comme une résultante de l'amplification du capital de connaissances grâce à la multiplication et l'aspect inédit des observations possibles. Les natures de ces observations sont diverses : il s'agit de nouvelles représentations des données stimulant la perception des acteurs métiers, de résultats sous forme de nouvelles variables ou de leurs corrélations. Les observations en question peuvent être ignorées, constituer une confirmation ou rejet d'une prédiction issue d'une hypothèse proposée par les métiers au préalable, ou bien donner lieu soit à la transformation par les métiers en nouvelle hypothèse, soit à la formulation d'une hypothèse qui pointerait la cause de cette observation[44]. Cette diversité de mécanismes montre une association hybride sur le terrain entre des raisonnements déductif et inductif, mais aussi abductif, le choix de méthode étant effectué par les acteurs métier selon la situation et leur besoin de progression. Celui-ci subit cependant une restriction en termes de ressources : le projet doit être réalisé dans le cadre d'un budget ou de délais limités, et la création de connaissances est alors soumise à la pression des délais de convergence vers un résultat acceptable, suffisant pour la prise de décision. L'environnement du système étudié contraint et influe ainsi sur la capitalisation de connaissances : en fonction de la pression de ces contraintes, la progression s'arrêtera sur un résultat suffisamment opérationnel, ou bien sera poursuivie par l'étude des hypothèses complémentaires générées à partir des observations écartées pour permettre la progression du projet.

[44] Nesvijevskaia (2015), La controverse épistémologique Big Data face à la réalité de l'appropriation de nouveaux paramètres par les acteurs métier en entreprise. In Chartron, Broudoux, *Big data - Open data, Quelles valeurs ? Quels enjeux ?* (p. 137‑149). De Boeck Supérieur.

L'appropriation de ces observations a lieu au cours de deux grandes phases marquées par des interactions fortes entre les acteurs du projet. La première correspond au début du projet : ces interactions entre acteurs métier, les Data Scientists et les données permettent la convention (Desrosières, 2008) et la sélection des paramètres et des modèles de traitement de l'information. Elles sont partiellement captées dans les modèles existants à travers l'identification d'itérations entre les activités de compréhension business et compréhension data, ainsi que d'établissement des objectifs et des hypothèses. Elle cible en priorité la compréhension de la pratique métier, et a une fonction d'anticipation des travaux à réaliser. Dans ce cadre, le Data Scientist (au sens large d'acteur clé du projet) joue un rôle de médiateur pour animer cette interaction. La seconde correspond à la restitution des résultats : des médiations entre humains et objets « non humains » (résultats d'algorithmes, interfaces…), au sens de l'absorption du lien social dans la machine (Latour, 1994), ont lieu au cours du partage des résultats pour leur transformation connaissances utiles à l'aide à la décision, automatisée au non. Ces deux situations sont créatrices de savoirs et traduisent un raisonnement de l'acteur métier qui a choisi d'affronter un défi, un mouvement de réconciliation dialectique hégélien, impliquant les connaissances métier, statistiques et techniques.

Ces cycles peuvent être répétés : l'itération globale est perçue sur le terrain comme un progrès si elle conduit à une convergence vers de meilleurs usages. Il ne s'agit pas d'itérer sur l'optimisation d'un modèle algorithmique pour résoudre la même question, malgré ce que peut suggérer la cyclicité globale du modèle CRISP_DM, mais bien de redéfinir l'usage cible grâce à de nouvelles connaissances, comme le montre l'évolution de l'applicatif de prévision d'activité (cas 2). Or, cette itération positive est conditionnée à la capitalisation des savoirs : en absence de celle-ci, il existe un risque de déperdition de connaissances au profit de la prise de décision dans le cadre de l'usage direct, anticipé initialement. Il est donc indispensable de mettre en place des instances d'interaction pour favoriser les arbitrages entre cette capitalisation des savoirs et la poursuite de l'objectif initial, et par conséquent la convergence vers un usage optimal.

2.2.2 Indicateurs clés : bénéfices, ressources et incertitudes

L'état de l'art démontre que la valeur d'usage est à globalement considérée comme la plus appropriée pour juger l'impact des projets data sur les entreprises, et que ces usages restent marqués par un certain niveau d'incertitudes. Par ailleurs, un projet étant un dispositif pilotable, il est utile d'identifier les indicateurs clés qui permettent de réaliser les arbitrages qui marquent

l'avancement du projet. Ainsi, le lancement et l'évolution du dispositif « projet data » sont soumis à l'acceptation d'une cohérence entre les bénéfices potentiels des usages futurs et de ressources à mobiliser pour leur conception, en prenant en compte de façon adaptée les spécificités liées aux incertitudes du projet. L'arbitrage s'articule alors autour de 3 indicateurs de valeur (voir Figure 33) : les bénéfices (potentiellement générés par l'exploitation des différents usages), les ressources (investissement dans le dispositif « projet data ») et les incertitudes qui indiquent un niveau de confiance dans l'estimation des deux premiers indicateurs.

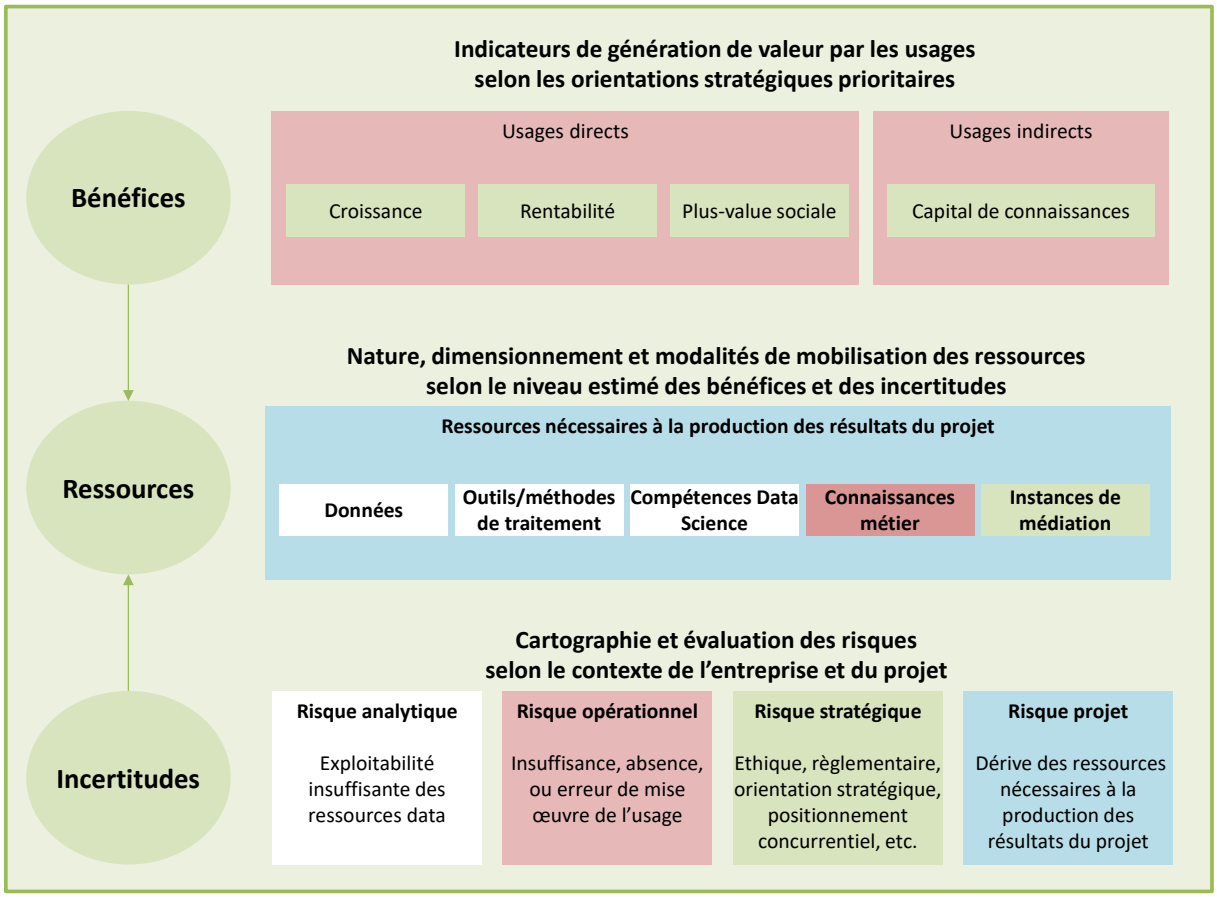


Figure 33 – Evaluation et mesure de la valeur : bénéfices, ressources et incertitudes

2.2.2.1 Bénéfices

Les bénéfices correspondent aux gains générés par l'exécution des usages issus du projet data. Ces gains peuvent être de natures différentes et complémentaires, et leur valorisation dépend des orientations stratégiques prioritaires de l'entreprise (Atamer & Calori, 2003). Les orientations peuvent prioriser la **croissance de l'activité** (création de produits pour gagner de nouveaux de marchés, croissance du chiffre d'affaire, financements...), la **rentabilité**

(identification des sources de baisse de coûts possibles, investissements plus rentables, efficacité de ciblage pour la mise en œuvre de leviers opérationnels, amélioration de la productivité ou accélération de tâches métier et support, y compris les tâches de prise de décision et de gestion des connaissances...) ou la **plus-value sociale** (confort des parties prenantes comme les salariés ou les actionnaires, communication interne et externe, produits et services contribuant à l'amélioration de la satisfaction client, capacité d'innovation...). Il s'agit des bénéfices directs. Ils ne sont générés que par l'exploitation des usages, dépendent de la nature de la valeur qu'ils produisent et des ressources nécessaires à son exploitation, et impliquent ainsi de prévoir des instruments de mesure, qualitatifs ou quantitatifs, appropriés. Les ressources nécessaires à l'exploitation des usages directs comprennent généralement des informations utiles à la décision issues des travaux du projet, mais aussi des applicatifs et outils métier liées à l'usage et des compétences métier et support, articulées au sein d'un processus d'exploitation approprié. Certains usages peuvent entrer en exploitation au cours du projet, ce qui permet de mesurer les bénéfices réalisés avant la fin du projet.

A ces bénéfices directs s'ajoute un potentiel de **bénéfices indirects lié à la capitalisation de connaissances** (Manyika et al., 2011; Mayer-Schoenberger & Cukier, 2014), c'est-à-dire l'ensemble des bénéfices qui seront générés par des usages futurs permis par le savoir construit au cours du dispositif. Cette estimation de bénéfice est beaucoup plus difficile à évaluer : elle se fait habituellement de façon relative au marché, à l'expérience passée, ou de façon purement qualitative. La somme des bénéfices liés aux usages directs et indirects constitue le bénéfice potentiel total.

2.2.2.2 Ressources

Face au bénéfice potentiel des usages directs et indirects, la mobilisation d'un dispositif « projet data » est un investissement initial dans d'un ensemble des ressources. Ces ressources sont évaluées et confrontées avec les bénéfices potentiels afin de converger vers un dimensionnement optimal du dispositif. Les projets mobilisent des ressources pour plusieurs objectifs :

- Réaliser les travaux de **production analytique**, c'est-à-dire la mise en œuvre d'une stratégie de résolution d'un problème par l'analyse des données pour générer une connaissance utile. Cette activité mobilise les données qui font l'objet d'une transformation ainsi que des outils, des méthodes analytiques et les compétences en Data Science.

- **Anticiper les usages**, c'est-à-dire prévoir les modalités de leur exploitation en précisant le mode opératoire de génération de gain et les ressources associées. L'anticipation des usages mobilise lourdement les connaissances métier (et support au métier si l'exploitation de l'usage le nécessite) et impacte la stratégie de la production analytique

- Animer les **instances de médiation** entre hommes et données qui jalonnent le processus. Les ressources spécifiques de ces instances seront précisées plus loin.

Ces activités ont pour objectif de produire des résultats exploitables optimaux dans le cadre de nouveaux usages encore incertains au démarrage du projet. C'est donc un processus actif de réduction d'incertitudes.

2.2.2.3 Incertitudes

Le concept d'incertitude est choisi ici pour plusieurs raisons. Tout d'abord, il fait référence à l'ensemble des risques que peut comporter un projet : à l'heure où les projets data n'ont pas fait l'objet de standardisation de pratiques sur le marché, et que leurs risques spécifiques n'ont pas été cartographiés, il est préférable de parler d'incertitudes (risques encore méconnus en termes de probabilité de survenance et d'impact). Ensuite, le concept peut avoir une définition plus radicale, knightienne : l'incertitude est alors liée à un nouvel usage où non seulement l'avenir n'est pas connu, mais il ne peut l'être au démarrage du projet. Par exemple, il est impossible de savoir si un algorithme détectera un contexte d'efficacité d'un traitement médical nouveau, ni si le contexte détecté sera activable pour l'administration du traitement. Il s'agit d'une spécificité des projets data qui visent la découverte de connaissances : il est, en théorie, impossible de prévoir l'usage tant que cette découverte n'a pas abouti à l'issue du projet. En pratique, les deux types d'incertitude sont interdépendants, ce qui pousse à privilégier l'utilisation du terme « incertitude » tout en employant le terme de risque lorsque celui-ci a été d'ores et déjà été observé ou qu'il peut être utilisé pour une cartographie. L'objectif ici est de proposer des pistes minimales de mitigation des risques sans toutefois restreindre l'incertitude qui caractérise le projet aux risques gérables.

En complément des incertitudes classiques d'un projet en entreprise, le dispositif « projet data » est en effet caractérisé par, *a minima*, quatre facteurs de risque spécifiques. Tout d'abord, la mythologie du phénomène Big Data et la pression du mimétisme induisent des motivations imprécises de la part des commanditaires des projets : elles sont souvent erronées, trop ou pas assez ambitieuses. C'est une difficulté à définir l'utilité de l'usage lui-même pour l'entreprise.

Cela génère des incompréhensions, voire des conflits d'intérêt entre les parties prenantes. Ensuite, le caractère exploratoire des usages (volonté de découverte d'usages, d'inspiration) présente des difficultés que ne portent par les projets qui répondent à un besoin opérationnel formulé en amont. De plus, le caractère exploratoire des données ajoute une complexité à l'analyse de la pertinence et de faisabilité, c'est-à-dire de l'exploitabilité de la ressource « donnée » pour la transformation en information utile. Enfin, un dispositif « projet data » est polarisé entre ses composantes métier (enjeux stratégiques et opérationnels, usages, connaissances métier…) et data (données, algorithmes, compétences Data Science, outils de traitement des données…), et ces pôles ne sont pas toujours convergents dès l'amont du projet, ce qui nécessite la mise en place d'une médiation appropriée.

Chacun de ces facteurs spécifiques peut ainsi engendrer des risques, dont une partie a été observée (voir Annexe 10 – Risques observés sur les projets data) et a fait l'objet d'une cartographie des risques préliminaire. Elle se veut illustrative, et non pas exhaustive, et ce plus particulièrement face à la récence du phénomène. Il est ainsi préférable d'envisager ces facteurs non pas comme des générateurs de risque, mais comme des générateurs d'incertitudes, où l'on ignore encore la probabilité de survenance d'un risque. Les risques d'ores et déjà observés ont été regroupés au sein de 4 familles d'incertitudes principales (voir Figure 34) :

- **Les risques analytiques** : Incertitude à produire une information utile à la prise de décision dans le cadre de l'usage

- **Les risques opérationnels** : Incertitudes à produire des usages exploitables correctement.

- **Les risques stratégiques** : Incertitudes à produire des usages utiles ou bénéfiques pour l'entreprise. Ce risque dépasse le cadre du système traitant l'usage, et englobe des enjeux liés à la valorisation même des bénéfices, mais aussi à la stratégie de l'entreprise, à son l'image et à sa conformité dans l'environnement externe, que ce soit d'un point de vue métier (usage non conforme aux règles et à l'éthique des pratiques) ou data (traitement des données non conforme aux règlementations en vigueur et aux pratiques acceptées par le marché). L'occurrence d'un évènement de cet ordre impacte non seulement le système traitant qui exploite l'usage, mais l'entreprise plus globalement. Il modifie l'assiette de bénéfices, voire génère une perte.

- **Les risques projet** : Incertitude à livrer des résultats de qualité suffisante dans les coûts et les délais impartis. Elle est liée à l'ampleur probable des réajustements liés aux trois incertitudes précédentes.

Facteurs de risques spécifiques

Nature de risque	Imprécision des motivations	Caractère exploratoire des usages	Caractère exploratoire des données	Manque de médiation homme / donnée
Risque analytique Incertitudes liées à l'utilité des informations à produire	Difficultés d'accès aux données	Difficultés à définir le champs d'investigation et les critères de jugement des résultats	Qualité ou quantité insuffisante des données pour générer le résultat visé (absence de signal)	Difficultés à donner du sens aux données, et inversement de trouver les données qui représentent un concept métier
Risque opérationnel Incertitudes liées à l'exploitabilité des usages	Difficulté à mobiliser les connaissances métier pour juger la pertinence de l'exploitation des usages	Difficultés à anticiper les bénéfices des usages et les ressources nécessaires à leur exploitation	Données non exploitables dans le cadre du processus de de l'usage (modalité d'accès, de traitement, de visualisation,,..)	Sous-exploitation ou erreurs d'utilisation liées à une mauvaise compréhension des l'information produites
Risque stratégique Incertitudes liées à l'utilité et au bien fondé des usages à produire	Complexité de la priorisation des usages et déconnexion de ce choix de la finalité stratégique de l'entreprise	Difficultés à décliner les objectifs stratégiques sous la forme d'usages métier	Données non exploitables dans le cadre de l'activité de l'entreprise	Nature des résultats déconnectée ou incompatible avec les enjeux stratégiques
Risque projet Incertitudes liées à la dérive des ressources dédiées au projet	Dispositif inadapté en absence d'arbitrage entre les usages directs et indirects	Changement de cible au cours du projet, (gaspillage) et difficultés à faire atterrir un projet (enlisement).	Données non exploitables dans le cadre du projet, ou dérives liées à la collecte et à la mise en qualité des données	Erreurs d'arbitrage impactant les ressources du projet

Figure 34 – Cartographie des risques spécifiques au dispositif de projet data

2.2.2.4 Cadre d'évaluation

Ainsi, les arbitrages se basent sur 3 types d'indicateurs : les bénéfices des usages, les ressources du projet, et les incertitudes liées à l'évaluation de ces premiers. Un cadre d'articulation entre ces indicateurs permet de mieux comprendre les enjeux (et les difficultés) des arbitrages réalisés au cours des projets data.

En effet, l'émergence initiale des attentes conduit à admettre la possibilité d'un bénéfice (**B**) qui sera généré par un usage permis par le résultat d'un projet data mobilisant des ressources (**R**). Il s'agit essentiellement des gains espérés des usages directs, comme le gain de temps en nombre de salariés pour les contrôleurs dans le cas 4 du contrôle conformité en cas, et du budget pour le projet, c'est-à-dire le coût de la prestation et le temps passé sur le projet pour les acteurs internes. Cependant, l'usage n'est pas encore certain au démarrage du projet : les bénéfices B

sont dotés d'un certain niveau d'incertitude face aux risques stratégiques (**I_{RS}**, ou erreur de valorisation du bénéfice) et opérationnels (**I_{RO}**, méconnaissance de l'exploitabilité de l'usage, par défaut équivalente au bénéfice total au cas où l'usage s'avère non opérationnel). Dans l'exemple du cas 4, le potentiel de gain est en effet inconnu, et la capacité à modifier la procédure de contrôle non plus, car ces pistes opérationnelles dépendent des résultats de l'analyse.

Les ressources du dispositif « projet data » (R) sont dotées elles aussi d'un niveau d'incertitude liée aux risques projet (**I_{RP}**, erreur de valorisation des ressources projet). Toujours dans le même exemple, le prestataire évalue un budget, mais n'est pas certain que le temps pour la réalisation soit correct : il réinvestira si nécessaire pour couvrir ce risque. Les ressources projet R contiennent des ressources dédiées aux travaux de production analytique (**R_{PA}**), nécessaires et spécifiques à ces projets data. L'objectif du travail de production analytique consiste à réduire les incertitudes liées au risque analytique (**I_{RA}**, méconnaissance de l'utilité de l'information à produire pour l'usage). Cela veut dire que la réduction de cette incertitude impacte l'incertitude opérationnelle, grâce à la confirmation analytique de la pertinence à mettre en œuvre un usage. Or, elle ne l'annule pas : en effet, un usage parfaitement confirmé du point de vue analytique, mais non opérationnel pour d'autres raisons, génèrera toujours un bénéfice nul. Par exemple, dans le cas 3, l'algorithme identifie bien un signal significatif sur le fait que les femmes qui tombent enceintes ont plus de risques d'être invalides, et donc de générer de l'absentéisme, or il est tout à fait inapproprié de mettre en œuvre un quelconque levier de prévention en entreprise qui pourrait mener vers une discrimination à l'embauche ou une dévalorisation de la maternité. Par ailleurs, si le projet est inscrit dans des incertitudes stratégiques trop importantes (pouvant aller jusqu'à générer des pertes, par exemple si l'application d'un usage atteint l'image de l'entreprise, qui perd par conséquent des parts de marché), la mise en œuvre de l'usage confirmé du point de vue analytique sera décevante (voire dangereuse).

Il est donc nécessaire de compléter le dispositif avec des ressources destinées à articuler la réduction des incertitudes stratégique, opérationnelle et projet avec la production analytique : il s'agit du reste des ressources du dispositif du projet (**R_{Δ}**), consommé au cours des instances de médiation. Ainsi, un cadre d'évaluation global (**E_{v0}**) (voir Figure 35) peut être utilisé tout le long du projet pour l'articulation des indicateurs clés.

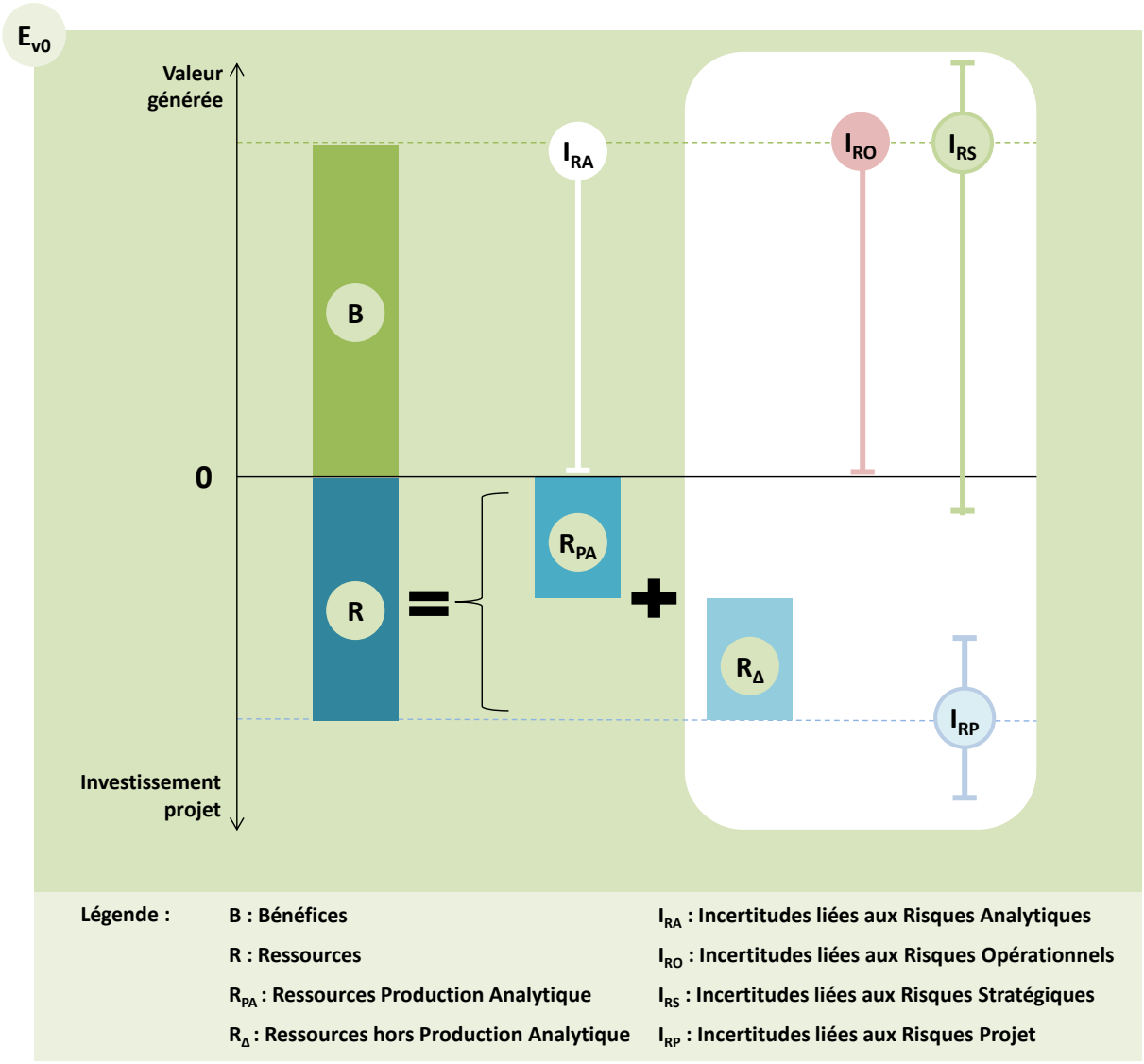


Figure 35 – Cadre d'évaluation du dispositif « projet data »

Ce cadre d'évaluation est destiné à être utilisé à chaque instance d'arbitrage (**E_{v1}**, **E_{v2}...**, **E_{vf}**), l'évaluation finale étant équivalente au cadre de mesure de la valeur (**M_{v0}**) à générer par l'exploitation de l'usage (direct ou indirect) confirmé par le projet. Il s'agit de la **maturité de l'usage**. Une fois ce cadre posé, le dispositif projet data apparait doté d'une finalité plus subtile qu'une génération de valeur mesurée par un simple retour sur investissement : il s'agit de lever un maximum d'incertitudes, dont l'incertitude analytique spécifique à ce dispositif, qui empêchent l'exploitation des usages optimaux, ceux-ci restant éventuellement marqués par des incertitudes résiduelles.

2.2.3 Processus de réduction d'incertitudes

L'allocation des ressources dans un projet data est ainsi étroitement liée à la nature des incertitudes à lever avant d'exploiter les usages optimaux. Le projet apparait comme un travail de production consommateur de ces ressources, à la fois informatique et cognitif (Mayère, 1990; McAfee & Brynjolfsson, 2012; Provost & Fawcett, 2013b), visant de façon performative la convergence vers une valeur optimale des usages directs et indirects. Les observation clés sur le terrain appellent à distinguer dans le processus global (Chapman, 1999; Shearer, 2000; Wirth & Hipp, 2000) les ressources dédiées à la production de résultats analytiques (levée des incertitudes analytiques tout le long de la chaine de valeur du traitement de la donnée visant l'usage direct), les ressources liées à la réduction des incertitudes métier (stratégiques et opérationnelles) et enfin les ressources nécessaires au pilotage du projet, déterminant les tactiques de réallocations des ressources dans ces projets mouvants. Enfin, dès lors que les projets data sont marqués par une dynamique de capitalisation sur les usages que ces incertitudes caractérisent, il est nécessaire de mettre en perspective les externalités positives de cette capitalisation, notamment en termes de gestion de portefeuille de projets data.

2.2.3.1 Réduction d'incertitudes analytiques

Le travail de production analytique proprement dit, spécifique aux projets data, est la **mise en œuvre d'une stratégie analytique anticipée** pour lever une incertitude analytique identifiée et générer une connaissance utile. Le travail de production analytique consomme les ressources de production analytique (données, compétences Data Science, outils et méthodes analytiques) et suit un chemin critique marqué par un ensemble de dépendances en termes **traitements de la ressource « données »**.

2.2.3.1.1 Livrables intermédiaires de la production analytique

Chaque étape de ce chemin critique est conditionnée à l'aboutissement du traitement de la « donnée » au cours du travail de production de l'étape précédente, ce qui permet d'établir un ensemble de livrables intermédiaires qui jalonnent la production analytique.

- Périmètre d'investigation

Les concepts métiers sont cartographiés et structurés pour l'exploration sur un périmètre restreint par l'usage direct anticipé. L'utilité des concepts métier est traduite en utilité analytique (objet d'analyse, phénomène d'intérêt, attributs…). Chaque élément du périmètre

est alors qualifié par l'attribution d'un ensemble de critères d'inclusion et d'exclusion métier, ce qui permet de réduire le champ des données possibles, théoriquement infini, uniquement aux concepts potentiellement utiles : c'est un cadrage métier du champ d'investigation. Le cadrage métier permet ainsi de réduire une partie des incertitudes liées au risque analytique : en effet, les hypothèses métier guident la recherche de d'informations utiles dans les données pour éviter la collecte et le traitement de données connues d'avance comme inutiles.

- Qualification des données source

Les données explorées dans le cadre d'un projet data sont collectées (requête, achat, génération de données brutes…) ou construites à partir de données source (enrichissement, agrégation…). Les données source sont jugés comme à collecter ou non dans le cadre du projet, selon leur valeur contributive perçue pour l'usage. Les concepts métier et data font l'objet d'un travail de traduction afin d'identifier les données les plus pertinentes à l'observation d'un concept métier jugé utile. Cette traduction inclut un travail de qualification, de validation d'intégrité et de cohérence, et peut provoquer l'évolution des concepts métier et la mise en évidence des solutions d'optimisation de la qualité (nettoyage, restructuration…). La collecte, l'exploration des données et leur qualification permettent ainsi d'éliminer les données non exploitables dans le cadre de la recherche du signal, c'est-à-dire de poursuivre l'alimentation des critères d'inclusion et d'exclusion sous l'angle de la **faisabilité**. Ce travail conduit à établir alors une structure cible pour la préparation des données, et contribue à réduire le risque analytique grâce au cadrage data du champ d'investigation.

- Matrice d'apprentissage

Il s'agit du résultat du travail de préparation des données collectées, et se présente sous forme de matrice remplie par les données transformées selon le modèle structuration. La matrice d'apprentissage est issue des tâches de collecte de l'ensemble des données, de sélection, de nettoyage, de construction de la matrice, d'intégration des données dans la matrice, et de reformatage de cette dernière. Elle lève des incertitudes analytiques complémentaires quant à l'exploitabilité des données.

- Résultat analytique

Le résultat analytique comprend l'ensemble des nouvelles données générées par les apprentissages réalisés au cours de la modélisation, c'est-à-dire les données, modèles et

métadonnées générées par les algorithmes, y compris les paramètres des algorithmes et les critères d'évaluation des résultats techniques. Il est issu des tâches de construction et d'évaluation statistique du modèle, et permet de mesurer le niveau de signal dans la matrice d'apprentissage. Ce travail lève les incertitudes analytiques liées à l'exploitation algorithmique des données, et peut conduire à revoir la matrice d'apprentissage pour optimiser le signal.

- Résultat utile

Il s'agit de la traduction de l'ensemble des éléments du résultat analytique en termes métier, et comprend ainsi un enrichissement métier final ainsi qu'une représentation des données résultat du travail analytique sous forme d'indicateurs et de visualisation interprétables pour la validation selon les critères d'évaluation anticipés. En cas de découverte de signal imprévu, les critères d'évaluation peuvent être enrichis, avec une revue potentielle des concepts métier mobilisés. A l'issue de la validation, le signal sera considéré comme information utile, et la réduction de l'incertitude analytique suffisante pour que l'information puisse être envisagée comme une ressource potentielle pour un usage.

- Résultats opérationnels

Une information utile fait l'objet d'une capitalisation de connaissances, et, si l'usage qui la mobilise est confirmé, d'une préparation pour le déploiement de l'usage. Les résultats opérationnels comprennent ainsi l'ensemble des retraitements de données, modèles et métadonnées nécessaires pour garantir leur intégration optimale, parfois automatisée, dans un usage jugé exploitable, en dehors des retraitements réalisés dans les livraisons des étapes intermédiaires. Les données sont complétées avec le processus de leur exploitation future (dont la distinction entre usage direct et indirect), ainsi que par les indicateurs de mesure de l'incertitude analytique pour le pilotage au cours de l'exploitation de l'usage direct.

Ainsi, chaque livrable intermédiaire du travail de production analytique consomme des ressources et constitue un vecteur de réduction de l'incertitude analytique. L'objectif final de ce travail est de juger la pertinence d'utiliser une connaissance construite à partir d'un ensemble de données brutes dans le cadre d'un usage envisagé. Il vise la confirmation de la pertinence de l'usage du point de vue analytique : dans ce cas, l'incertitude analytique restante est considérée comme acceptable pour l'exploitation de l'usage (voir Figure 36).

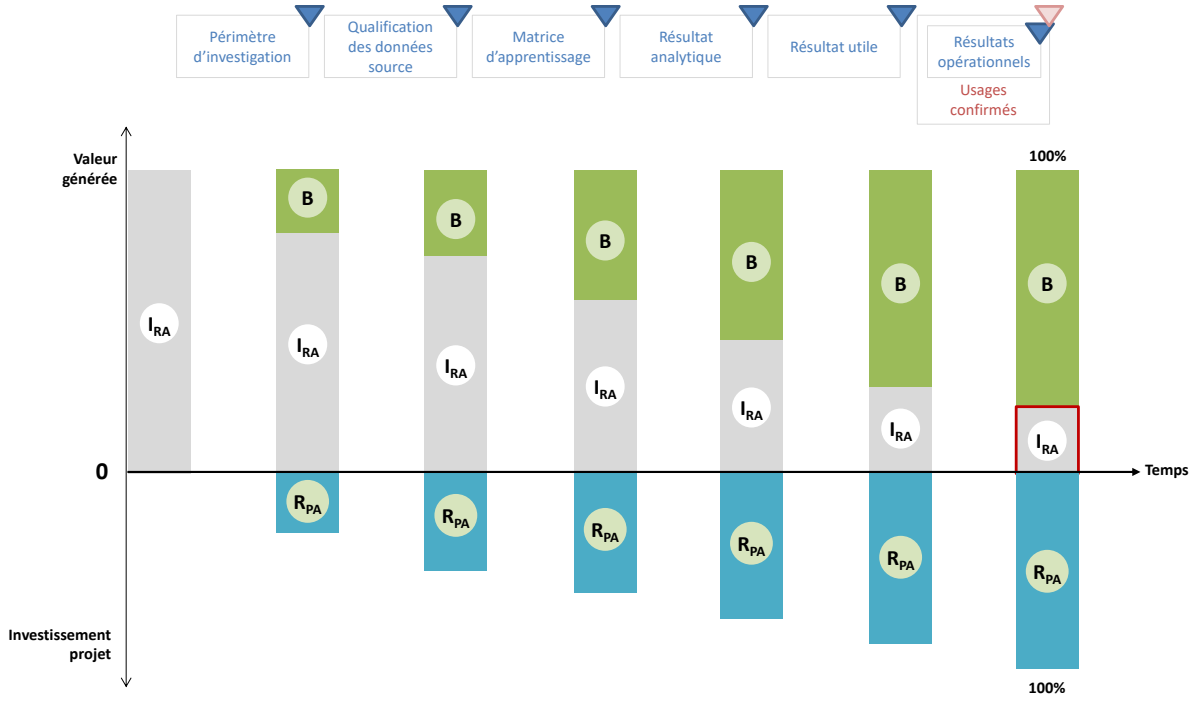


Figure 36 – Jalonnement de la réduction de l'incertitude liée au risque analytique par le travail de production analytique

2.2.3.1.2 Chemin de traitement des données et gestion des versions

La production des résultats intermédiaires peut être séquentielle (chemin critique simple) ou superposée, lorsque des données intermédiaires sont produites de façon progressive et donnent lieu à des **versions de livrables intermédiaires**. Le processus de production analytique distingue en effet les livrables des étapes intermédiaires, et les livrables intermédiaires au sein de chaque étape. Chaque livraison nouvelle fait l'objet d'une revue par l'instance de médiation pour juger du caractère définitif du livrable, et présente une opportunité de découvrir des éléments susceptibles d'alimenter des usages imprévus.

De plus, l'étape d'évaluation constitue un passage clé dans la mesure où elle permet de transformer un résultat analytique en résultat utile, ce qui conditionne la capitalisation de connaissances et la préparation du déploiement de l'usage. Cela conduit à distinguer un chemin critique non abouti (absence de validation de l'utilité du résultat) d'un chemin critique complet, abouti. Dans ce cadre, le chemin critique minimal est réalisé dans le cas où chaque livrable des étapes intermédiaires est jugé comme un livrable définitif, et où l'évaluation donne lieu à la validation de l'utilité du résultat : il s'agit alors d'une production séquentielle, proche des processus de projets plus classiques, moins marqués par les incertitudes (voir Figure 37). Il est

à noter que si le chemin critique minimal peut naturellement être perçu comme optimal du point de vue des ressources, ce n'est pas nécessairement le cas dans les projets data exploratoires, à condition d'assurer la transformation maximale des livraisons intermédiaires en usages découverts.

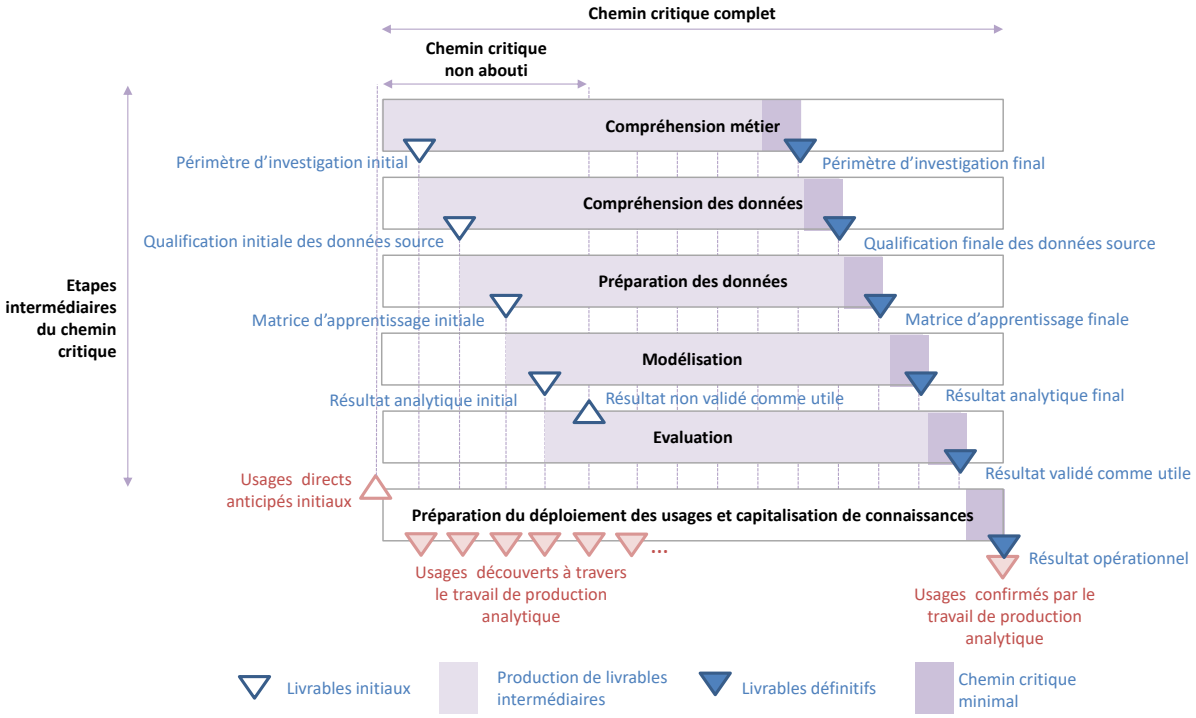


Figure 37 – Chemin de traitement des données au cours de la production analytique

Contrairement à la présentation cyclique du modèle de processus CRISP_DM, cette vision du chemin de traitement des données rend la production analytique pilotable dans le temps et compatible non seulement avec une gestion projet séquentielle, mais aussi avec les méthodes concourantes (pour les projets complexes en termes d'hétérogénéité de compétences mobilisées) ou agiles (pour les équipes plus réduites), à condition de suivre le versionning des livrables intermédiaires et leur impact sur les livrables dépendants.

2.2.3.2 Réduction d'incertitudes métier

En parallèle du travail de production analytique, le dispositif met en œuvre un travail de production qui vise la réduction des incertitudes stratégiques et opérationnelles. Cette réduction d'incertitudes nécessite la mobilisation de méthodes classiques, issues du champ de l'audit des risques, du conseil stratégique et opérationnel, de l'analyse des processus, de l'organisation, de

développement d'applicatifs, de gestion de connaissances... Elle est mise en œuvre au cours des phases de compréhension métier (stratégique) et de la préparation des usages (opérationnelle).

- La réduction des incertitudes stratégiques est réalisée au cours de la phase de **compréhension métier**. Elle produit un cadre de valorisation des bénéfices, selon des critères conformes aux choix stratégiques de l'entreprise dans son environnement organisationnel, économique, règlementaire, éthique... Elle permet de juger la pertinence des problématiques analytiques et des usages découverts au cours du projet grâce à la priorisation des enjeux métier, ainsi que de poser un ensemble de contraintes au dispositif pour éviter les évènements susceptibles d'affecter négativement l'entreprise. Elle garantit ainsi la pertinence du **cadrage stratégique** de la stratégie analytique et assure ainsi la bonne **problématisation** métier dans le processus de contextualisation du dispositif.

- La réduction des incertitudes opérationnelles est quant à elle réalisée au cours de la phase de **préparation du déploiement des usages et de capitalisation de connaissances**. Elle produit un cadre d'évaluation des bénéfices, et définit les ressources nécessaires à l'exploitation des usages et les gains potentiels générés. Elle garantit la pertinence du **cadrage opérationnel** de la stratégie analytique par la définition des contraintes liés à l'exploitation des usages, qu'il s'agisse de contraintes informatives (sens des résultats, pertinence et fraîcheur de l'information, capacité à capitaliser les connaissances…), techniques (compatibilité de l'algorithme cible avec les outils qui serviront pour l'usage…), humaines (adoption de l'usage, capacité à prendre une décision en fonction du résultat…), ou économiques (pilotage de valeur d'usage, coûts d'exploitation…). En effet, elle répond aux besoins d'adhérence de l'usage dans les pratiques métier, et fournit donc les éléments nécessaires à la prise en compte native de cette adhérence dans les choix de stratégie analytique.

Les cadrages stratégiques et opérationnels constituent le cadrage métier, c'est-à-dire l'établissement d'un cadre qui contraint l'anticipation et l'ajustement les ressources du dispositif et contribue ainsi à la construction de la stratégie analytique qui guide le travail de production analytique. Cette construction suit une logique d'anticipation verticale descendante (« top down »), c'est-à-dire que les besoins métier stratégiques permettent la priorisation des usages, de ces usages découlent les critères d'évaluation des résultats métier, qui sont eux même traduits en critères statistiques et guident le choix des modèles algorithmiques candidats. Cette anticipation des modèles guide à son tour la structure cible des données, et donc la préparation des données, ce qui conditionne les axes selon lesquels les données devront être comprises.

Ainsi, le cadrage métier guide l'ensemble des livrables des phases intermédiaires du projet et débouche sur un plan de collecte cible qui alimentera la compréhension des données.

Or, ce cadrage métier n'est pas suffisant au dispositif puisqu'il ne permet pas d'estimer les ressources nécessaires à la mise en œuvre de la stratégie analytique pour chaque phase : ces ressources sont en effet contraintes par l'exploitabilité des données, ou la notion de faisabilité, et la capacité à prévoir les scenarii de production analytique afin la guider. L'analyse de faisabilité est réalisée au cours de la phase de **compréhension des données**, qui consiste à établir les risques analytiques et à réduire les incertitudes liées à ces risques au cours des travaux de production analytique. Elle permet de poser un ensemble de contraintes au dispositif pour éviter les évènements susceptibles d'affecter négativement la pertinence des résultats analytiques, et garantit ainsi le **cadrage data** de la stratégie analytique. Ce cadrage data comprend l'anticipation des ressources nécessaires à chaque phase pour la réalisation du chemin critique, et suit une logique d'anticipation des ressources verticale montante (« bottom up ») : à partir de la compréhension des données sont déduits le temps de retraitement (mise en qualité, structuration...), les compétences nécessaires, et les outils. Les caractéristiques de la matrice d'apprentissage cible guideront les choix de méthode d'analyse algorithmique, et notamment la stratégie d'itération entre la préparation des données et la modélisation. Les critères d'évaluation statistiques induits par les modèles nécessiteront un travail de traduction en indicateurs métier pour permettre l'évaluation. Enfin, ces résultats imposeront des contraintes sur l'usage, et devront être traduits en indicateurs de valeur pour l'entreprise.

Ainsi, l'établissement de la stratégie analytique suit une **logique anticipatoire contextualisée à double dynamique**, marqué par un mouvement de convergence polarisé entre la compréhension métier (« top down ») et la compréhension des données (« bottom up »), et animé par les instances de médiation (voir Figure 38). Ces instances de médiation, verticales, permettent de juger et de faire évoluer le dispositif projet, et sont alors à distinguer de la phase d'évaluation, qui constitue une rotule entre les critères de jugement métier et data du résultat analytique, et s'inscrit dans la logique temporelle de la production analytique. Les instances de médiation partagent cependant avec la phase d'évaluation les critères d'évaluation des résultats métier et data : en effet, l'avancement du projet est marqué par une amélioration des critères, c'est-à-dire par la réduction de l'incertitude analytique, ce qui constitue un indicateur d'avancement de la production du résultat informationnel transversal à l'ensemble du dispositif, dans le temps et à travers les phases. Il est ainsi indispensable d'établir les critères de référence

(par exemple, issus des usages existants dans l'entreprise, des références externes, ou une estimation initiale). Enfin, les instances de médiation partagent avec la phase d'évaluation les méthodes de représentation des résultats, qu'il s'agisse d'indicateurs, de présentations graphiques, de cartographies, voire d'interfaces dynamiques de restitution des résultats développés *ad hoc*.

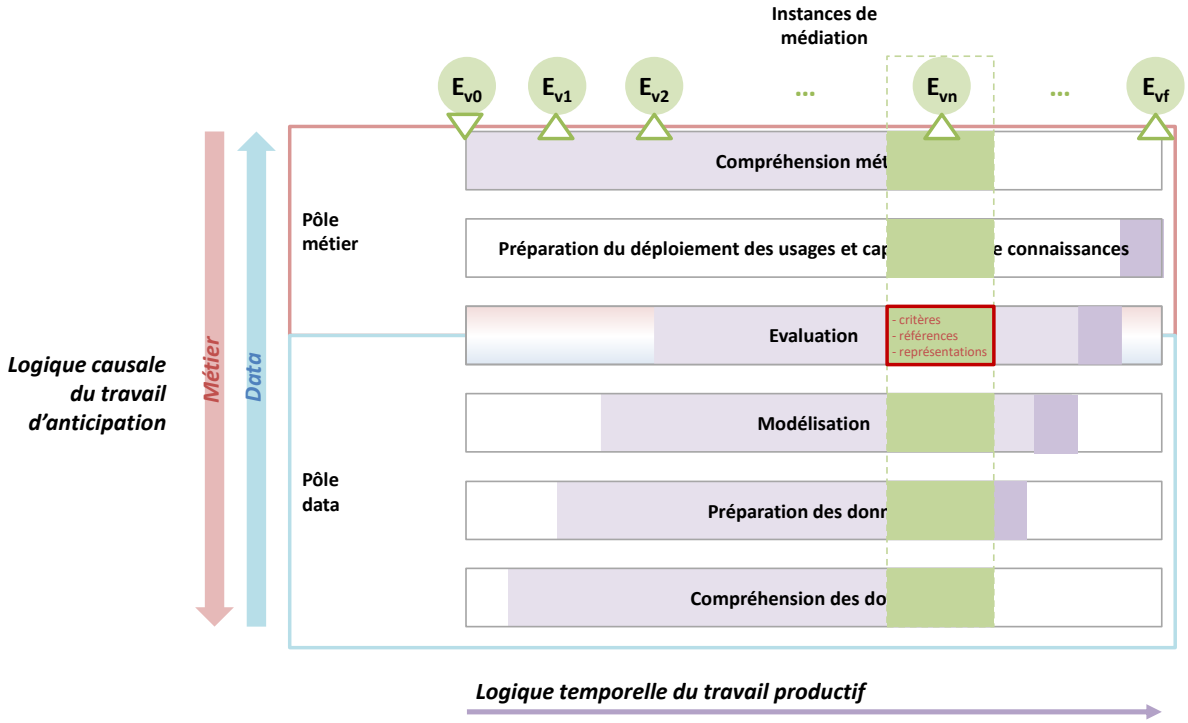


Figure 38 – Logiques anticipatoire et temporelle du dispositif data polarisé

La mise à plat de ces travaux de production et des logiques anticipatoires et temporelles ne s'affranchit pas du besoin de mettre en évidence la dynamique d'arbitrage qui doit permettre au dispositif projet de s'ajuster, en réallouant les ressources face aux perturbations propres au travail dans un univers incertain.

2.2.3.3 Dynamique de réévaluation des incertitudes projet

La première évaluation du dispositif projet ($\mathbf{E_{v1}}$) représente un état de référence de la **maturité du dispositif projet data**. Elle est dépendante de son capital de savoirs initial qui permet la perception de l'incertitude projet et détermine la capacité à réaliser l'évaluation, et donc la qualité des hypothèses de la nature et du niveau d'incertitudes à lever. Le capital de savoirs initial a été généré précédemment par un ensemble d'interactions d'apprentissage et représente l'expérience passée du dispositif. Le capital de connaissances s'accumulera au fur et à mesure

de la montée en maturité du dispositif au cours du projet. Un dispositif sera considéré (*a posteriori*) comme parfaitement mature lorsque le capital de connaissance aura été suffisant pour que la perception de l'incertitude projet et l'arbitrage qui en découle précèdent systématiquement tout risque avéré. Par exemple, une équipe de Data Scientists et de médecins qui a déjà traité plusieurs études cliniques se trompera moins en prévoyant le budget pour une nouvelle étude.

Mais la perception de l'incertitude projet n'est pas suffisante : il faut pouvoir agir en ajustant le dispositif pour faire face aux risques potentiels. En effet, l'évaluation est effectuée au cours d'une instance de médiation qui génère une convergence vers un dimensionnement du dispositif. Cette convergence s'appuie sur la quantification, c'est-à-dire la convention et la mesure (Desrosières & Kott, 2005), des risques liés à ces incertitudes, ainsi que sur la capacité à prendre les décisions qui pourront résoudre une partie ou la totalité de l'incertitude projet. Cette phase de quantification, qui peut être assimilée à une gestion de risques, nécessite ainsi la mobilisation des deux pôles (métier et data) du dispositif, et une génération de flux d'informations circulantes pour le partage du sens utile à la décision. Il s'agit de la mise en œuvre des interactions d'alignement entre les acteurs du dispositif (voir

Figure 39), visant à trouver un équilibre convenable entre la qualité externe du résultat final visé (Bénéfices B, dotés d'un niveau d'incertitudes acceptables I_{RS}, I_{RO} et I_{RA}) et la qualité interne du dispositif (Ressources R, dotées d'un niveau d'incertitudes acceptables I_{RP}).

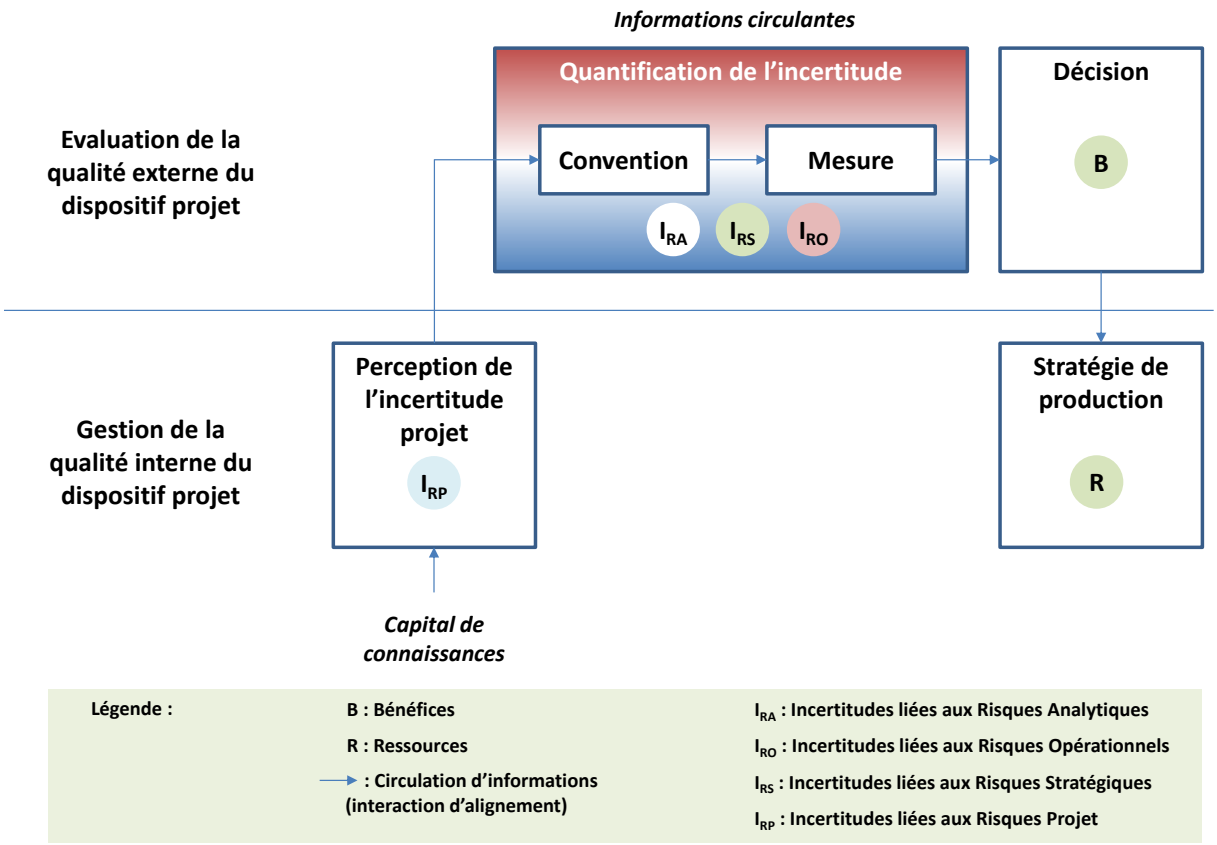


Figure 39 – Schéma de levée d'incertitudes projet au cours d'une instance de médiation

La qualité interne du dispositif apparait alors comme une conséquence de la capacité à percevoir les incertitudes projet, à susciter un arbitrage selon la qualité externe, puis à en déduire la mobilisation des ressources de production. Par exemple, lorsque l'équipe du cas 2 a découvert l'impossibilité d'exploiter les données des commandes pour anticiper les ventes (en effet, l'audit de ces données durait plus longtemps que prévu sans aboutir à une solution), une alerte a été remontée en point d'avancement pour convenir d'un changement de stratégie analytique, d'en évaluer l'impact sur la pertinence de l'usage visé (réduit), et de confirmer en arbitrage que le bénéfice espéré restait intéressant face au changement d'allocations de ressources du projet (changement de stratégie analytique).

2.2.3.4 Tactiques d'allocation de ressources

Lorsque les incertitudes sont jugées faibles, la logique d'anticipation métier « top down », réalisée au cours de la phase de compréhension métier, est suivie d'une étude de faisabilité « bottom up » au cours de la phase de compréhension des données. Cette convergence initiale aboutit sur le plan de collecte définitif et sur une stratégie d'analyse figée, ainsi que sur un lancement de la préparation des usages dotés d'une valorisation qui n'évaluera que peu au cours

du projet. Les instances de médiation sont alors réduites aux évaluations réalisées sur les livrables des phases intermédiaires, supposés définitifs dès la version initiale. Il s'agit d'une méthode de gestion de projet séquentiel, suivant le chemin critique minimal, selon un mode taylorien de gestion des ressources. L'échange informationnel au cours de ces projets est alors réduit à une mobilisation initiale du capital de connaissances sur la phase de définition des besoins et de la faisabilité, puis un transfert d'informations amont (anticipation) et aval (livrables) à chaque phase, ainsi que de la documentation de l'ensemble des travaux du projet.

Lorsque les incertitudes sont significatives, ce mode de fonctionnement n'est pas approprié, car des évènements imprévus impactent un dispositif projet peu flexible au cours des phases de réalisation, provoquant la dérive des ressources (surconsommation et délais) ou la dévalidation du résultat au cours de la phase d'évaluation. Par ailleurs, l'évolution des usages, notamment découverts, ne permet pas la modification de l'objectif et le redimensionnement du dispositif. Il est alors nécessaire de mettre en place une gestion de projet plus flexible (de type ingénierie concurrente ou agile) qui parallélise les phases de production, et fluidifie les échanges informationnels, qu'il s'agisse des informations amont (ajustement de la stratégie analytique) ou aval (revue de versions intermédiaires des livrables). Ce mode de fonctionnement priorise des livraisons à forte valeur d'usage et rapides à produire pour la livraison de versions intermédiaires, afin de permettre le démarrage et l'ajustement des phases dépendantes sur le chemin critique, et accélère ainsi le processus tout en laissant la possibilité à l'usage d'évoluer au cours du projet. Par ailleurs, ce mode de fonctionnement favorise les expertises data avec une **logique causale ascendante proactive** : en effet, celle-ci épouse le chemin critique, et produit des propositions d'usages au lieu de questionner les besoins métier. En outre, cela répond à la demande des experts métier qui souhaitent privilégier la découverte d'informations et d'usages, et permet d'explorer simultanément les possibilités analytiques (« techniques ») et les usages (valeur métier). Enfin, ce fonctionnement augmente significativement la part des ressources dédiées aux instances de médiation ($\mathbf{R_A}$) par rapport aux ressources de production analytique ($\mathbf{R_{PA}}$), et nécessite une expertise poussée pour anticiper les incertitudes. En cas d'incertitudes extrêmes et d'absence de capital de connaissances suffisant pour l'anticipation, c'est-à-dire d'une immaturité forte du dispositif, les méthodes agiles en petite équipe sont les plus appropriées, à condition de s'appuyer sur des expertises fortes de gestion de projet sur ce mode qui pousse l'innovation et de privilégier la montée en maturité à la production de résultat direct.

Ainsi, deux extrémités de choix s'offrent au dispositif de projet data : une tactique d'allocation de ressources séquentielle, ou une allocation de ressources simultanée. La première est essentiellement adoptée en cas de **recherche de garantie de résultat**, sans bénéficier de l'opportunité de découverte d'usages indirects. A l'inverse, la seconde tactique est celle de **« payer pour voir »**, c'est-à-dire limiter les pertes en envisageant d'emblée un projet data comme un investissement à perte, en misant seulement sur la génération des usages indirects, c'est-à-dire un capital de connaissances qui permettra, de façon itérative, d'identifier des usages stratégiquement et opérationnellement acceptables. Les alternatives à ces deux choix extrêmes sont nombreuses : ne rien faire, innover sans passer par un projet data, ou encore bénéficier de la seconde tactique à cout marginal par rapport à la première. Ces choix de posture déterminent les modes de pilotage et de partage de risque, tout en s'inscrivant dans un cadre d'évaluation commun qui permet de mettre en cohérence les indicateurs pour réaliser des arbitrages, y compris pour changer de posture.

Au-delà de ces choix de mode de gestion de projet, se pose la question de réaffectation de ressources à compétences équivalentes. En effet, la différence fondamentale entre en modèle de référence et le modèle dispositif Brizo_DS est cette séparation entre la production analytique et la production non analytique qui comprend l'ensemble de cette logique anticipatoire. Or, le cadrage métier, le cadrage data ou l'évaluation des modèles peuvent contenir un travail sur les données, mobilisant des ressources similaires à celles qui sont nécessaires au travail de production analytique au cœur du projet, voire les mêmes données, les mêmes outils et les mêmes individus. Par exemple, pour déterminer si le périmètre de l'analyse algorithmique d'attrition doit porter sur le marché des particuliers ou des professionnels, il faut soit le demander directement à un sachant métier porteur des choix stratégiques, soit établir les ordres de grandeur de ces marchés, dans le but d'alimenter les arbitrages que fera ce sachant métier et qui impacteront la stratégie analytique. Pour établir ces ordres de grandeur, une requête est demandée à un Data Scientist. Or il s'agit non pas d'une compétence de Machine Learner, mais de celle d'un Business Analyst. Cette analyse de données se distingue foncièrement du travail de production analytique par sa finalité : en effet, il ne s'agit pas de réduire des incertitudes analytiques liées à l'information finale exploitée par l'usage, mais les incertitudes liées à l'anticipation de la production analytique par une précision des hypothèses métier. De même, construire une interface de restitution des analyses mobilise des compétences de développement d'applicatif (UX/UI, MOA, Web Développement, BI…).

En cela, la réduction des incertitudes stratégique et opérationnelle, comprenant potentiellement des analyses de données *ad hoc* ou des compétences pointues dans des domaines liée directement à chaque usage, est perçue comme externe à la production analytique. Ce travail sur les données ne s'inscrit donc pas dans la temporalité du chemin critique de la production analytique, mais bien dans les instances de médiation, en tant que moyen d'appréhender les incertitudes métier. Le modèle Brizo_DS permet ainsi de s'affranchir (conceptuellement, si ce n'est en pratique) de cette pression sur les Data Scientists dits « moutons à cinq pattes » en séparant les compétences mobilisables entre la production analytique, la capacité à anticiper la stratégie analytique (mobilisation d'expérience dans le cadre des instances de médiation), et la contribution à la levée des autres incertitudes (stratégique, opérationnelle et projet). Il donne ainsi des leviers de réallocation de ressources humaines en fonction des compétences et selon le type d'incertitude nécessaire à lever.

La multitude des incertitudes possibles et le besoin de mobiliser les ressources data sur l'ensemble des tâches, que ce soit dans le cadre de la production analytique ou des instances de médiation, tendent à privilégier une gestion du projet data parallélisée et rompt la bijection entre les individus et les rôles qu'ils tiennent dans les projets. Cette méthode s'articule aussi bien avec le développement agile des applications métier qu'avec les méthodes classiques du conseil et de l'audit, à condition de ne pas les restreindre au pôle métier, mais bien aux instances de médiation tout le long du projet. Enfin, cette méthode est propice à l'émergence d'usages à travers ces projets exploratoires.

2.2.3.5 Suite et élargissement : vers une gestion de portefeuille de projets data

Le dispositif « projet data » est clos à partir du moment où l'ensemble des usages directs validés par le travail analytique est prêt pour la mise en exploitation, et où toutes les connaissances nécessaires à la génération d'usages indirects sont capitalisées, ce qui inclut les tâches de documentation des méthodes et des livrables de chaque phase et la synthèse de l'ensemble des décisions prises au cours des instances de médiation, restituées au cours du retour d'expérience.

Ce dernier point est clé dans le modèle. En effet, au-delà des usages directs exploitables, le dispositif génère un ensemble d'éléments valorisables de façon indirecte. Il s'agit d'usages « dépriorisés », nécessitant des analyses métier ou data complémentaires avant la mise en exploitation, des informations jugées utiles mais non transformées en usages opérationnels, des données, métadonnées et modèles intermédiaires réutilisables, et plus globalement du cumul de

connaissances (savoirs, expertise issue de l'apprentissage) contribuant à la montée en maturité du dispositif. La génération de valeur indirecte est soumise l'existence de moyens de capitalisation et de traçabilité des arbitrages entre une exploitation directe et future, y compris en cas d'abandon de pistes d'investigations.

- Lorsque l'exploitation future de l'usage est dépendante uniquement d'une exploration métier préalable, l'usage sort du dispositif polarisé métier-data et constitue une **découverte d'hypothèse métier**. Celle-ci pourra faire l'objet d'une exploration plus classique au sein d'un dispositif projet métier dédié, et le bénéfice généré à terme sera mis en perspective avec le dispositif initial qui a permis l'élaboration de l'hypothèse.

- En cas de décision d'abandon de l'exploration, le bénéfice étant inférieur aux ressources nécessaires pour la poursuite du projet, l'usage indirect est valorisé à l'économie des investissements futurs et génère des économies de ressources et de temps dans la gestion de de portefeuille de projets data, et contribue à la montée en maturité du dispositif.

- Lorsque les incertitudes ne sont pas suffisamment réduites dans l'immédiat par le dispositif projet pour la mise en exploitation de l'usage, mais le bénéfice attendu reste supérieur aux ressources nécessaires, l'usage donnera lieu à la mise en place d'un nouveau dispositif projet data. La notion d'usage indirect rend alors possible l'évolution du dispositif « projet data » en dispositif de génération d'usages récurrent et évolutif, doté de ressources ajustables selon les potentiels comparés des usages capitalisés, ce qui permet de mettre en œuvre une **gestion de portefeuille de projets data** (voir Figure 40). Tout élément de capitalisation de connaissances impacte directement l'ensemble des projets data du dispositif « Portefeuille de projets data »

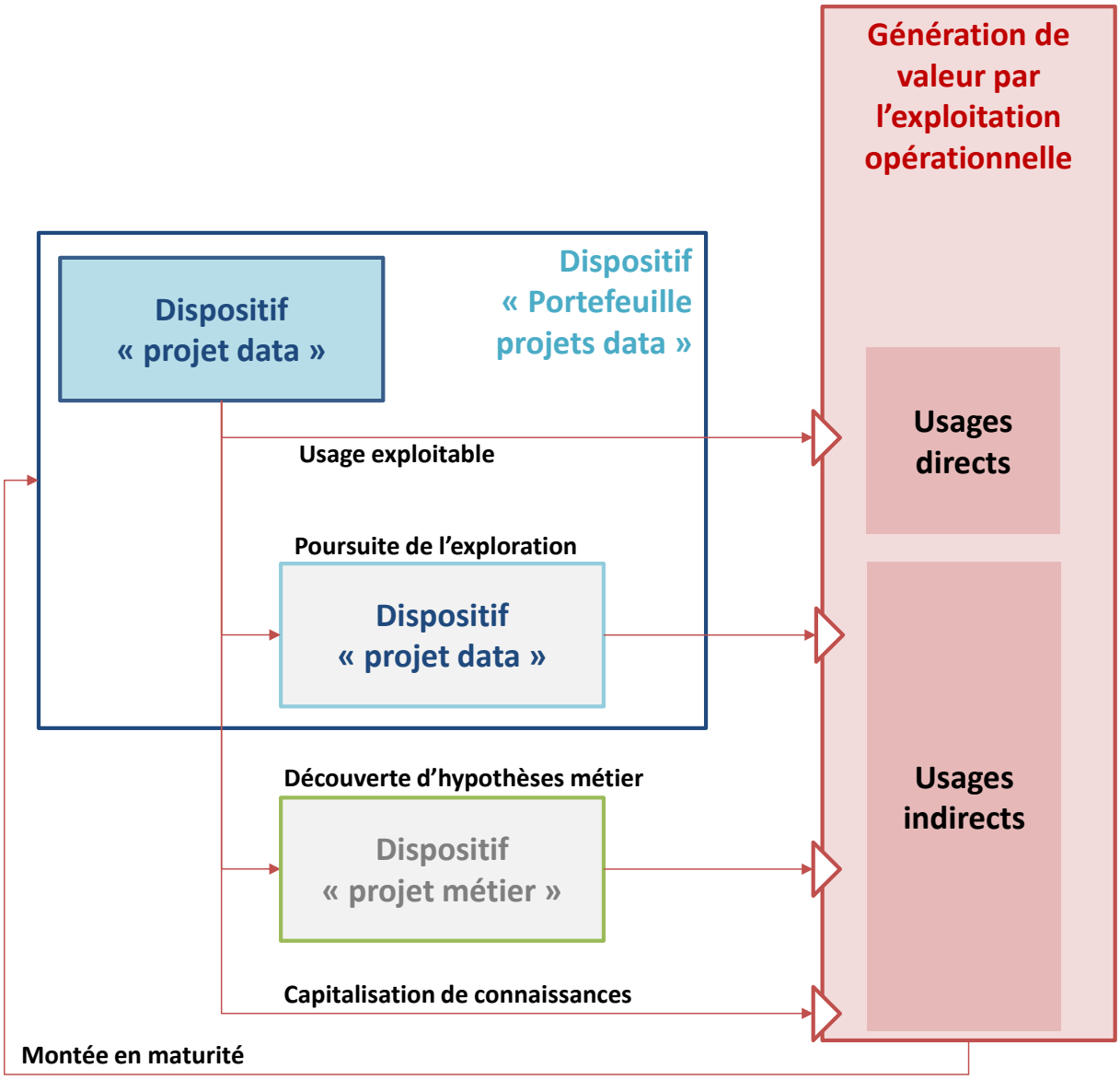


Figure 40 – Génération de valeur directe et indirecte par un projet data au sein d'un dispositif « portefeuille de projets data »

Le concept de dispositif « Portefeuille de projets data » résout la difficulté de la prise en compte de la cyclicité du processus global et présente l'avantage de ne pas omettre l'apparition de nouveaux usages tout le long du processus d'exploration, ainsi que l'exploration des usages incertains en dehors d'un dispositif « projet data ». Il assure une continuité entre les différents projets, que ce soit pour l'évolution d'un usage donné ou la génération d'autres usages, et des synergies de ressources qui évitent un gaspillage résultant d'une vision centrée sur un projet isolé.

A la lumière de cette proposition de modèle, un dispositif « projet data », voire « portefeuille projets data », peut être jugé en fonction de son efficacité, de sa qualité interne (productivité) et externe (utilité des informations produites pour les usages directs et indirects). Or, l'identification des premiers facteurs de risque de ces projets pointe déjà les causes des incertitudes principales qu'il est nécessaire d'éviter pour gagner en efficacité. Les facteurs liés à l'imprécision des motivations des demandeurs semblent intimement liés à la récence du phénomène Big data et à une **acculturation insuffisante** des acteurs métier. Le caractère exploratoire des usages en découle, de façon plus opérationnelle, et nécessite une mise en perspective des **usages possibles par type de secteur ou de fonction** (attrition en assurance, études cliniques en santé, prévision d'activité en contrôle de gestion…) et une compréhension des **usages transversaux** (accélération et pertinence de la prise de décision, capacité à innover, génération de connaissances, qualification des données…). Le caractère exploratoire des données et le manque de Médiation Homme-Données manquent de cadre théorique décliné en fonction des spécificités liées à ces projets. Ces deux facteurs constituent les dimensions approfondies dans la suite de ces travaux (voir Figure 41, dont les illustrations plus détaillées ont été fournies en Figure 34 dans le chapitre 2.2.2.3 de la Troisième partie page 254), qui proposent des cadres et des solutions concrètes à mettre en œuvre pour réduire les incertitudes des projets data : le Databook et le cadre de Médiation Homme-Données.

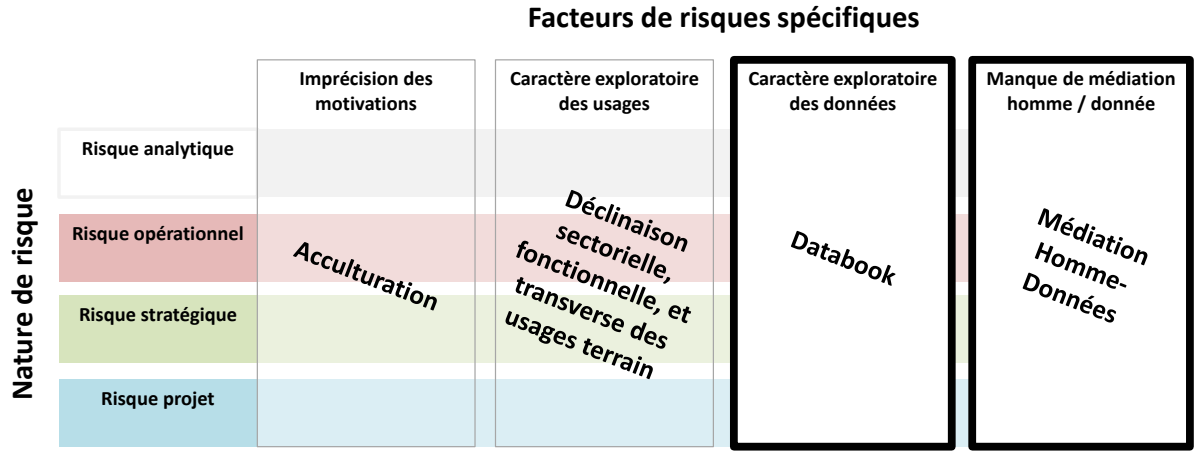


Figure 41 – Le Databook et la Médiation Homme-Données : proposition de leviers au service de la réduction des incertitudes

2.3 Qualité des données

La génération d'informations utiles est issue du processus de contextualisation, qui constitue une convergence entre les savoirs des acteurs métier (dimension intellectuelle des connaissances capitalisées qui permet la problématisation de l'information) et la conversion des données et informations (Harrathi & Calabretto, 2006). Le dispositif « projet data », jalonné des arbitrages interactifs, se situe au cœur de ce processus, que ce soit à travers la production analytique ou les instances de médiation. Il s'impose à la fois comme :

- Un processus consommateur de données comme ressource première, dont l'efficience interne dépendant de leur qualité initiale (par exemple, lorsque le projet prévision d'activité, cas 2, s'est heurté au fait que les données de commandes étaient inexploitables en absence de clé pour les relier aux factures)

- Un processus de mise en qualité des données source, notamment au cours de la phase de préparation des données, et de qualification, mobilisant un ensemble hétérogène de contributeurs à l'évaluation de la qualité des données dans le cadre des usages (par exemple, lors de la sémantisation complète des données sous la forme de dictionnaires pour le cas 1 en attrition assurance santé)

- Un générateur de données, de modèles et d'algorithmes au cours de la production analytique, qualifiées par des métadonnées issues d'une grille d'évaluation propre au projet, et un générateur de documents non analytiques propres au projet (voir illustrations de modèles de données documentées dans le Databook en Annexe 11 – Databook : genèse et prototypage)

- Un moyen d'objectiver la gestion de la qualité des données grâce à l'identification des usages indirects, dont ceux qui nécessitent la poursuite de la mise en qualité des données (par exemple au cours du cas 5 lorsque le projet a permis de mettre en évidence le besoin de mettre en qualité les rapports de visite pour mieux analyser les sinistres en dommage aux biens)

Les études de cas permettent de mettre en évidence de façon exploratoire l'impact de ces projets sur les données, les modèles de données, et les métadonnées, tout le long du processus d'exécution du projet, afin d'établir un lien entre les données et l'usage, c'est-à-dire la valorisation des données. Le jalonnement des activités de production analytique est le squelette de l'articulation entre les données et les usages, et permet de cartographier les impacts du projet sur la qualité des données. La qualification progressive des données au rythme des versions des

livrables intermédiaires représente alors un état d'avancement de chaque activité, ce qui constitue une articulation tangible entre les instances de médiation et le travail de production analytique. Si la construction des livrables analytiques est propre à chaque projet, tout comme les usages, les jalons peuvent en revanche être documentés selon un cadre commun, présentant un ensemble d'éléments rigides liés à la finalité des activités au sein du processus, et un degré de flexibilité qui correspond au cadre d'évaluation propre à chaque dispositif.

2.3.1 Databook : documentation dynamique de la qualification des données

Le Databook représente ce cadre de documentation et, à la manière d'un livre, raconte l'histoire de la transformation des données en information au cours du projet. Il s'agit d'un dispositif de capitalisation de connaissances, d'un objet matériel et humain qui sert à tracer la mémoire du projet data, et par conséquent le modèle algorithmique qui en résulte. Il s'apparente aux outils de documentation de données et de connaissances plus habituels, comme les modules de gestion de métadonnées intégrés par les éditeurs ou les bases de connaissances théoriques de type IKR (Hofmann & Tierney, 2009) ou de librairies de connaissances réutilisables entre différents projets d'entreprise (Bernard & Tichkiewitch, 2008). En absence de dispositifs de documentation dédiés aux projets data, le Databook est en effet inspiré d'ingénierie des supports numériques de connaissances chers au Knowledge Management et aux Sciences de l'Information et de la Communication, cherchant à établir des attributs standards d'éléments de connaissance (Zacklad et al., 2007) et à suivre les processus de capitalisation, de partage, de création de connaissances, ainsi que l'apprentissage et la sélection et d'évaluation d'informations utiles (Ermine, 2003). Ainsi, le Databook utilisé et présenté ici est un outil nouveau, improvisé sur le terrain face à un besoin de documentation et en absence de dispositif adéquat, qui va plus loin qu'une documentation figée des données en épousant le processus complet d'un projet data et ses spécificités, dont une dynamique complexe de construction d'algorithmes. Il a une utilité non seulement à l'issue du projet, comme documentation de référence indispensable à l'usage (Pavel & Serris, 2016), mais aussi pendant le projet, comme support aux arbitrages à chaque jalon du projet.

Prototype issu de la confrontation d'un cadre théorique (mise en évidence du travail de construction cognitif, importance des métadonnées et nécessité d'inclure l'algorithme comme nouvel objet dans les processus de qualité des données) et d'un réel besoin terrain qui a émergé dès le début de ces travaux de recherche, le Databook a fait l'objet d'un test à la fois par moi-même sur les études de cas présentés ici, et par des Data Scientists sur d'autres projets (voir

Annexe 11 – Databook : genèse et prototypage). Il s'agit d'un recueil structuré d'informations descriptives sur la construction du livrable final. Ces informations sont métier (sens, utilité perçue…), statistiques et techniques, et comprennent l'évaluation de chaque donnée mobilisée selon un ensemble des critères de qualité construit selon le contexte. Les critères de qualité portent sur les données source et livrables analytiques intermédiaires, ainsi que leur état d'avancement face aux incertitudes identifiées. Il est ainsi à la fois utile à la gestion du dispositif, grâce à la traçabilité de l'avancement du livrable analytique final par la formalisation du flux informationnel convergent, et à la traçabilité des transformations intermédiaires nécessaires à la construction du livrable définitif. C'est, par conséquent, la documentation de référence pendant et après la mobilisation du dispositif, destinée à l'ensemble des contributeurs de l'équipe projet et aux utilisateurs des résultats, en particulier en cas de déploiement des résultats (le Databook sert alors de cahier des charges pour la conversion des données brutes en informations utiles de façon industrielle). Ce cadre de documentation de la qualité des données constitue ainsi une structure utile à la valorisation des données, par attribution rétroactive de la valeur liée à l'usage à chaque élément data mobilisé dans la construction amont de cet usage.

Cette structure est articulée autour d'une suite de modules organisés en 2 parties (voir Figure 42) : 4 modules introductifs qui permettent d'ancrer le Databook dans le contexte du projet (ces onglets servent et documentent les instances de médiation) et 7 modules qui documentent les livrables analytiques intermédiaires tout le long du projet. Ces modules permettent de suivre l'avancement de la résolution des incertitudes sur le chemin critique au cours des instances de médiation, ce qui permet de réajuster les ressources nécessaires à la production. Plus spécifiquement, chaque module représente les rotules, l'interfaçage entre les phases, ce qui constitue une vision assez inédite par rapport à des propositions de documentation à plat de chaque étape. Parmi ces 7 modules, 2 sont dédiés à la documentation de la valeur potentielle de l'usage, pour chaque usage direct et pour l'ensemble des usages indirects. Ces deux modules donnent un cadre à la valorisation des données traitées au cours de la production analytique et au suivi de l'avancement de la maturité des usages.

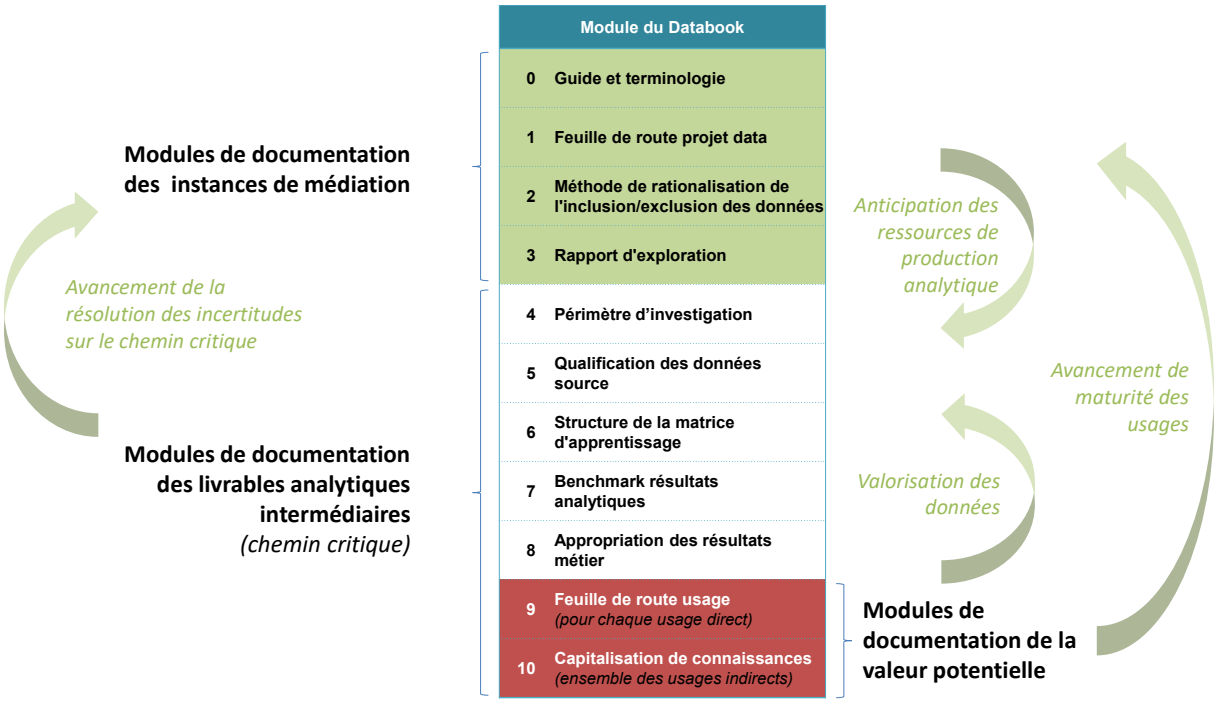


Figure 42 – Structure du prototype de Databook utilisé dans les études de cas

Les modules de documentation des livrables analytiques intermédiaires permettent de dresser une liste structurée d'objets, un ensemble de critères de qualification liés aux types d'incertitudes réduites par la qualification, et un statut d'avancement de la qualification pour chaque élément. Les statuts d'avancements alimentent les modules de documentation des instances de médiation, qui fournissent elles-mêmes les critères et les méthodes de qualification. La structure est compatible avec des modes de gestion de projet différentes (séquentielle, concourante, agile…), ainsi qu'avec le modèle CRISP_DM ajusté dans ces travaux de recherche, grâce à l'établissement de lien clair entre les livrables et les modules (voir Figure 43), permettant ainsi de distinguer les livrables nécessaires à la gestion de projet data au cours des instances de médiation, et les livrables liés à la production analytique.

Module du Databook		Livrables associés au module, selon le modèle CRISP_DM ajusté
0	Guide et terminologie	Terminologie
1	Feuille de route projet data	Contexte, priorisation des usages, objectifs métier, critères de succès métier, inventaire des ressources, exigences, hypothèses et contraintes, risques et contingences, coûts et bénéfices, maquette initiale de datavisualisation, plan projet,évaluation initiale des outils et techniques
2	Méthode de rationalisation de l'inclusion/exclusion des données	Méthode de rationalisation de l'inclusion / excusion des données
3	Rapport d'exploration	Rapport d'exploration
4	Périmètre d'investigation	Cible d'analyse Data Mining, Critères de succès Data Mining
5	Qualification des données source	Rapport de collecte des données et description des données collectées, dictionnaire des données, rapport de qualité des données
6	Structure de la matrice d'apprentissage	Rapport de nettoyage des données, attributs dérivés, enregistrements générés, matrice d'analyse agrégée, matrice d'analyse refrormatée pour les analyses, description des traitements des données
7	Benchmark résultats analytiques	Techniques de modélisation, critères d'évaluation de la modélisation, Test Design, paramétrages, modèle(s), description du (des) modèle(s), évaluation du modèle, paramétrages révisés
8	Appropriation des résultats métier	Support de présentation des résultats, évaluation des résultats par rapport aux critères de succès métier, résultat opérationnel, revue critique du process, liste des actions possibles, décision
9	Feuille de route usage (pour chaque usage direct)	Plan de déploiement, plan de pilotage et de maintenance
10	Capitalisation de connaissances (ensemble des usages indirects)	Retour d'expérience et documentation

Figure 43 – Modules du Databook et livrables associés

Enfin, le Databook établit un **lien tangible** entre la réduction des incertitudes et les phases du projet grâce à la traçabilité des critères de qualification externes (incertitudes analytique, opérationnelle, et stratégique) et internes (incertitude projet, à logique anticipatoire).

Ce lien est décrit en détail sous la forme d'une grille de critères de qualification de chaque concept à qualifier (Figure 44).

Module du Databook	Phase pilotée grâce au module	Objet qualifié	Critères de qualification externe des données			Phase déclenchée par le module	Critères de qualification interne des données
			Analytiques	Opérationnels	Stratégiques		Projet
0 **Guide et terminologie**	Instances de médiation					Compréhension métier	
1 **Feuille de route projet data**		**Critères de succès métier**					**Priorités du projet et critères d'arbitrage partagés**
2 **Méthode de rationalisation de l'inclusion/exclusion des données**		**Critères d'inclusion et d'exclusion**	**Qualité réelle et souhaitée pour la production analytique**	**Qualité souhaitée pour l'exploitation de l'usage**	**Qualité théorique et perçue**		**Qualité souhaitée pour pour le projet,** sous la contrainte des ressources
3 **Rapport d'exploration**		**Critères d'évaluation**	**Niveau d'incertitude en termes de risques analytiques**	**Niveau d'incertitude en termes de risques opérationnels**	**Niveau d'incertitude en termes de risques stratégiques**		**Ressources nécessaires,** comprenant le cout de l'acquisition et du traitement des données, y compris le temps, les outils, les compétences, etc.)
4 **Périmètre d'investigation**	**Compréhension métier**	**Concepts métier perçus comme utiles**	Utilité dans la construction analytique (objet d'analyse, phénomène d'intérêt, périmètre temporel, attributs bruts et dérivés, filtres, clés,…)	Utilité dans l'exploitation opérationnelle (contribution à la décision, natures et délais d'actions possibles, modalités d'historisation,…)	Pertinence des concepts selon les critères de succès métier (ordres de grandeur des bénéfices, règles de calcul,…)	**Compréhension des données**	Ressources pour la collecte et l'exploration des données
5 **Qualification des données source**	**Compréhension des données**	**Données perçues comme exploitables**	Exploitabilité techique, accessibilité, intégrité, cohérence, complétude, actualité, lisibilité, précision, biais et qualité des données relative à la finalité de l'analyse	Exploitabilité technique des données brutes (volumétrie, formats,…) et ressources nécessaires pour leur usage	Sens métier et lisibilité des données brutes (dictionnaire), censure stratégique (exclusions, anonymisations,…)	**Préparation des données**	Ressources nécessaires à la préparation des données, dont le travail de mise en qualité et la structuration, et à la modélisation
6 **Structure de la matrice d'apprentissage**	**Préparation des données**	**Données qualifiées comme exploitables** (sélectionnées et construites)	Nature de l'utilisation pour la modélisation (découpage de base d'apprentissage, validation et test, transformations statistiques, usages pour différents modèles…) et utilité pour le résultat analytique (poids, sensibilités, correlations,...)	Exploitabilité technique des données construites et ressources nécessaires à la structuration pour l'usage	Sens métier et lisibilité des données construites (dictionnaire)	**Modélisation**	Ressources nécessaires à la modélisation
7 **Benchmark résultats analytiques**	**Modélisation**	**Modèles de données explorés**	Performance des modèles, niveau de confiance statistique dans les résultats, valeur des critères d'évaluation statistiques	Exploitabilité opérationnelle des algotithmes (complexité, fréquence ou seuils de réapprentissage, maintenance des modèles…)	Sens métier et lisibilité des modèles construits (dictionnaire, processus de généraition des modèles, processus de génération des résultats,...)	**Evaluation**	Ressources nécessaires à l'évaluation des résultats, dont leur restitution
8 **Appropriation des résultats métier**	**Evaluation**	**Résultats évalués**	Retraitements réalisés pour permettre l'évaluation métier (ajout d'indicateurs de lecture, changements de format de présentation des résultats…)	Performance visée des usages et niveau de confiance dans l'application opérationnelle	Sens métier des critères statistiques, lisibilité, évaluation des critères de succès métier, confiance dans le niveau des bénéfices	**Préparation du déploiement des usages et capitalisation de connaissances**	Ressources nécessaires à la préparation des usages et à la capitalisation
9 **Feuille de route usage** *(pour chaque usage direct)*	**Préparation du déploiement des usages et capitalisation de connaissances**	**Résultats opérationnels**	Retraitements techniques nécessaires à l'exploitation (intégration dans un applicatif, automatisations, contrôles, sécurisation des données,...) et identification du niveau de contribution à la génération de valeur	Exploitation opérationnelle des résultats (gouvernance, habilitations, organisation, outils mobilisés, processus d'exploitation, processus de maintenance,...) et identification du coût de la donnée pour l'usage	Indicateurs de mesure de génération de bénéfices par les usages directs	**Exploitation des usages directs**	Bénéfices potentiels des usages directs : impact du dispositif sur la valeur des données
10 **Capitalisation de connaissances** *(ensemble des usages indirects)*		**Ensemble des résultats du projet** (données, métadonnées, modèles de données, livrables non analytiques)	Retraitements techniques nécessaires à la capitalisation de connaissances contenues dans les livrables analytiques (intégration dans un applicatif, sécurisation des données,...) et identification du niveau de leur contribution à la génération de valeur	Processus de mise à disposition des connaissances contenues dans ces résultats, et identification du coût de la l'exploitation de ces connaissances	Indicateurs de mesure de génération de bénéfices par les usages indirects	**Exploitation des usages indirects**	Bénéfices potentiels des usages indirects : impact du dispositif sur la valeur des données, et objectivation de la mise en qualité future

Figure 44 – La qualification des données au service de la réduction des incertitudes

Cette grille se lit de la façon suivante :

1. ***Chaque module permet de piloter une phase du projet**. Les 4 premiers guident les instances de médiation, le périmètre d'investigation guide la compréhension métier, la qualification de données source guide la compréhension des données, et ainsi de suite.*

2. ***Chaque module sert à qualifier un concept.***

    a. *Les premiers modules permettent de définir et de qualifier les critères de succès métier, les critères d'inclusion et d'exclusion des données, et les critères d'évaluation des résultats analytiques.*

    b. *Les modules suivants servent à qualifier tous concepts utiles dans la chaîne de transformation de la donnée : d'abord un concept métier perçu comme utile, puis la donnée qui le représente et qui parait exploitable, puis la donnée confirmée comme exploitable...*

3. ***La qualification interne des concepts se fait en fonction de sa capacité à réduire les incertitudes stratégiques, opérationnelles et analytiques.** Par exemple, un modèle peut être jugé en fonction du niveau de confiance statistique dans les résultats (incertitude analytique), ou en fonction de la facilité à les maintenir (incertitude opérationnelle).*

4. ***Une fois qu'un concept est qualifié, il est utilisable pour le module suivant.** Ainsi, le module « structure de la matrice d'apprentissage », qui guide la phase de préparation des données, déclenche la phase de modélisation. Dans ce cadre, la qualification des données exploitables détermine les ressources nécessaires à la phase suivante de modélisation. Elle réduit donc les incertitudes projet en permettant de mieux anticiper les phases suivantes.*

Cette grille décrit ainsi le contenu et la méthode de gestion de l'incertitude par la qualité des données. Elle est conceptuelle, et sert de référence théorique complétée à la lumière du modèle de dispositif data proposé dans ces travaux de recherche. Physiquement, le Databook a été partiellement développé pour être testé sur le terrain au cours des études de cas, ainsi que par d'autres Data Scientists, et a fait l'objet d'un retour d'expérience documenté, montrant son utilité et ses limites. La spécificité de chaque projet nécessite en effet un choix dédié des critères de qualification, ce qui impose un développement d'outil à la fois très flexible, tout en

préservant les éléments reconnus comme à forte valeur ajoutée. Le résultat de cette double approche, à la fois théorique et pratique, est une structure de prototype de Databook, utilisée et adaptée activement par les équipes Quinten.

Le Databook n'a pas fait l'objet d'études comparatives avec des outils sur le marché, cependant la connaissance de ces outils par les Data Scientists au moment des projets n'a pas permis trouver des équivalents pour combler les besoins de traçabilité des flux informationnels, en particulier en cas de gestion de projet concourante. Au cours du projet, il permet à l'ensemble des contributeurs d'avoir un accès aux flux informationnels cumulés au moment de la consultation, comme l'interprétation des données et des regroupements de variables, les sources et les interlocuteurs clés, le statut de chaque indicateur et de chaque variable dans le processus, les ordres de grandeur de référence, l'historique des arbitrages... Sa finalité première dans le cadre des projets reste la **facilitation des instances de médiation**, et non pas la capitalisation sur l'ensemble des travaux de qualification des données. Or, le Databook constitue un outil à finalités plus larges (voir Figure 45), sous-exploitées dans le cadre des études de cas. Il est ainsi nécessaire de poursuivre le développement des fonctionnalités de ce prototype afin de permettre son usage en dehors des projets.

Finalités		Bénéficiaires	Fonctionnalités offertes par le Databook
Efficience interne	Projet data	*Equipe projet data, sponsors du projet*	- Facilitation des instances de médiation
	Portefeuille de projets data	*CDO, Management d'équipes data,…*	- Capitalisation de connaissances nécessaires à la maturité des dispositifs, dont la connaissances des données et des algorithmes - Anticipation des usages indirects nécessitant la poursuite du travail analytique
Efficience externe	Usages directs	*Utilisateurs métier, fonctions support,…*	- Traçabilité de la construction de l'usage (traitement des données et arbitrages réalisés au cours de la conception de l'usage)
	Usages indirects	*Knowledge Manager, R&D, Innovation…*	- Mise à disposition d'une base de connaissances utile à la génération de nouveaux usages, y compris en dehors des projets data
Valorisation du patrimoine de données		*Direction générale, direction financière*	- Valorisation combinatoire des données, des modèles, et des métadonnées par les bénéfices liés à leur mobilisation dans les différents usages
Gouvernance de la qualité des données		*DPO, Data Manager, CDO,…*	- Génération et évaluation des critères de qualité des données (métadonnées), et suivi du niveau d'incertitudes - Identification des propriétaires (data owner, data steward,…) - Identification des données caractère personnel, et autres restrictions règlementaires - Documentation des processus de génération des données, y compris issues des algorithmes - Construction des référentiels et des dictionnaires des données
Coût de la mise à disposition des données		*DSI*	- Dimensionnement des outils nécessaires à l'usage - Dimensionnement des outils nécessaires à l'exploration - Anticipation de l'évolution du parc applicatif lié aux nouveaux usages

Figure 45 – Finalités et bénéficiaires des fonctionnalités d'un Databook

A travers le Databook, le dispositif projet data s'inscrit au cœur de la gestion de la qualité des données de l'entreprise. Il pose notamment un cadre nouveau pour la qualification de l'**utilité des informations** sous le prisme du projet data, en distinguant notamment des utilités des informations intermédiaires et des informations finales utilisées dans le cadre de la prise de décision. Il peut alors alimenter des problématiques plus larges, par exemple au cours de déploiement de Data Lakes.

En effet, deux « écoles » semblent s'affronter sur le terrain. La première école, portée notamment par les éditeurs, consiste à prioriser la mise en place technique d'un Data Lake en entreprise, puis de le mettre à disposition des équipes data et métier afin de valoriser les données qui l'alimentent à travers la génération d'usages. Cette école privilégie la sécurisation de l'accessibilité des données et leur historisation, au détriment de l'utilité des données contenues, ce qui complexifie le travail exploratoire et amplifie la perception du manque de qualité des données. La deuxième école, plus portée par les Data Scientists, positionne les dispositifs projet data en amont de la mise en place des Data Lakes. Cette école privilégie l'usage et l'objectivation rapide de la qualité des données, ce qui s'explique notamment par le processus du projet : si de nombreuses données sont explorées, seules les données les plus contributives sont identifiées comme utiles à l'usage, ce qui rend dans certains cas le Data Lake inutile à l'exploitation des résultats. La séparation entre un environnement technique dédié à l'exploration, et un autre dédié au déploiement des usages peut alors permettre de déconnecter les investissements techniques afin de dimensionner au mieux les ressources IT en fonction de la qualification des données au cours du projet. Les entreprises qui ont choisi cette démarche semblent déprioriser les investissements dans les technologies Data Lake au profit de la mise en qualité des données internes, cependant ce constat reste restreint à un nombre trop faible de cas.

Un autre exemple tient à la mise en évidence de l'importance des données internes de l'entreprise. En effet, l'exploration des données internes conduit à mener un travail de qualification approfondi, et une génération de données inédites. Face à la richesse des résultats construits sur les données internes seules, enrichies et qualifiées, les entreprises considèrent en effet comme moins prioritaire l'exploration des données externes, plus coûteuses à obtenir et à mettre en qualité que les données internes. Cette logique s'appuie sur la capacité à remonter d'un indicateur nouveau (utile) aux données sources, pour les qualifier d'utiles à leur tour. Or, cette utilité justifie un investissement dans une mise en qualité plus approfondie. Dans les

projets étudiés, seules les données externes mobilisées par l'entreprise pour des usages existants ont fait l'objet d'une intégration comme ressource, aucune donnée nouvelle externe n'ayant été acquise. En revanche, l'un des projets a nécessité la structuration sous forme de bases de données d'un ensemble de données internes issues de rapports Word : ce travail manuel de mise en qualité des données, pour les passer de « non exploitables » à « exploitables », a été identifié comme axe d'optimisation plus prioritaire que le déploiement des résultats du projet en soi. Cela confirme le manque de capitalisation de connaissances et la possibilité de s'appuyer sur le dispositif pour réaliser des travaux de mise en qualité.

2.3.2 Gouvernance des données et métriques propres aux algorithmes

La gouvernance de la qualité des données, identifié comme clé dans l'état de l'art (Loshin, 2010; Mariko, 2016; Nesme & Cottin, 2017a) apparait comme un élément qui a manqué aux dispositifs étudiés, marqués par un engagement insuffisant des ressources dédiées à ces activités : le manque de vision transversale et long terme conduit alors à une sous-capitalisation des connaissances, bien que celle-ci soit plus théorique que perceptible sur le terrain. Très concrètement, cela veut dire que les experts métier utilisent volontiers les outils de qualification et de documentation des données pour suivre l'avancement du projet, mais ne savent pas comment les partager et les utiliser en dehors. Cette sous-capitalisation est expliquée par la diversité des expertises mobilisées (difficultés à converger sur des critères de qualité partagés), l'absence de contributeurs transversaux pour garantir les méthodes de capitalisation des connaissances (knowledge manager…), l'absence de l'implication des contributeurs intéressés par la valorisation du patrimoine des données, comme des directions financières, et enfin un cadre d'intervention en tant que prestataire au forfait (projets limités dans le temps), ce qui ne génère pas le besoin de capitaliser des connaissances au profit d'une gestion de portefeuille de projets data partageant la même matière première.

Ainsi, en absence d'ancrage fort du dispositif dans la gouvernance des données de l'entreprise, le potentiel de connaissances des données, pourtant jugé important comme le confirment les projets d'attrition santé, d'analyse des sinistres ou encore la prévention santé-prévoyance, est réduit à l'efficience interne du dispositif, c'est-à-dire une ressource utile aux instances de médiation. Le capital de connaissances est alors moins prioritaire que la construction du résultat analytique au service de la prise de décision dans le cadre de l'usage direct, plus « mesurable ». Mais au-delà de la capitalisation de connaissances, ce manque d'ancrage dans la gouvernance des données de l'entreprise laisse entrevoir des risques liés au déploiement des algorithmes

conçus au cours du projet. Dans ce sens, le Databook intègre les premières métriques dédiées à l'optimisation de ce capital de connaissances et au cahier des charges du déploiement des algorithmes, mais reste à mettre en lien avec la gouvernance des données au sens plus large dans l'entreprise. Il s'agit ici de faire converger, sur un terrain encore vierge d'algorithmes comme modèles de données spécifiques à qualifier (Fox et al., 1994), trois visions sur les métadonnées : une vision performative, liée aux usages auxquels répondent les algorithmes, une vision informatique (Berti-Equille, 2012) visant à établir des attributs de qualité communs, et enfin une vision cognitive, dont l'objectif est de donner du sens et de « redocumenter » le contexte, le processus et les arbitrages successifs qui ont mené à la construction d'un algorithme (Broudoux & Scopsi, 2011).

Ces premières métriques proposées ici permettent de circonscrire la question métier à laquelle devra répondre un algorithme et la traduire en concepts data à chaque étape du traitement. Elles sont spécifiques aux projets data dans le sens où elles décrivent la composition d'un algorithme, et constituent ses « métadonnées » principales pour rendre intelligibles et plus transparent son fonctionnement. Les métriques en question présentent un intérêt particulier pour la gouvernance des données, car elles pointent l'ensemble des transformations des données en résultat analytique, ce qui permet de documenter les algorithmes déployés. Cette documentation est orientée sur 6 fonctions principales que peut jouer, de façon cumulative, chaque donnée dans la construction d'un algorithme (voir Figure 46).

Fonction de la donnée dans la construction de l'algorithme	Définition	Exemple pour l'algorithme de score de risque d'attrition
Objet de l'analyse	Il s'agit de la maille de la matrice d'apprentissage et de la restitution des résultats	Contrat d'assurance (numéro de l'affaire)
Périmètre de l'objet d'analyse	Sous-population d'individus analysés	Clients possédant des contrats de santé individuels ayant été actifs au cours de l'année 2014
Attributs de l'objet d'analyse	Caractéristiques (dits également drivers, patterns…) de l'objet analysé	Caractéristiques des clients et de leurs contrats, de leur consommation en santé et des contacts avec la société
Phénomène d'intérêt	Pour le algorithmes supervisés, attribut clé (performance ou risque) qu'il faut prédire ou expliquer	Attrition opérationnelle (résiliation de contrat santé subi et potentiellement activable), c'est à dire le churn
Critères d'évaluation statistique	Indicateur statistique permettant de comparer les modèles pour choisir le meilleur	Sensibilité (capacité à détecter les contrats churners) et spécificité (capacité à éviter de mal prédire les churners)
Critères d'évaluation métier	Indicateurs complémentaires non spécifiques à l'algorithme mais facilitant son appropriation	Calcul du montant global de primes ou de nombre de bénéficiaires par sous-population à risque

Figure 46 – Proposition de métriques clés pour documenter le traitement algorithmique

Dans le cadre de déploiement à plus grande échelle d'algorithmes au sein d'une organisation, notamment si le nombre d'algorithmes est élevé et si les données sources sont croisées, il sera indispensable de tracer le traitement des données dans la construction des algorithmes : l'utilisation de ces métadonnées permet d'indiquer la fonction principale des données pour chaque algorithme concerné. Ainsi, une modification du processus de traitement d'une donnée en particulier pourra donner lieu à une anticipation de son impact sur tous les algorithmes déployés. Ce sujet parait indispensable à intégrer dans la gouvernance globale des données en entreprise, et ce plus particulièrement à l'heure d'une régulation croissante de l'usage des données personnelles.

2.4 Dispositif de Médiation Homme-Données

Le dispositif de Médiation Homme-Données correspond, au sein d'un projet ou portefeuille de projets data, au sous-ensemble regroupant l'ensemble des acteurs humains et non humains (Latour, 1994) impliqués dans les interactions d'alignement permettant la génération de connaissances utiles aux instances de médiation. Il pourvoit la dynamique de convergence du processus de conception (Olivesi, 2014), polarisé entre la donnée explorée et l'homme qui définit le cadre de son exploitation future. La notion de « médiation » résulte de la méconnaissance *a priori* des informations constructibles à partir des données, et de la méconnaissance des possibilités d'actions déclenchées par les décisions basées sur ces informations. Ces méconnaissances sont directement liées au niveau d'incertitudes métier (stratégiques et opérationnelles) et data (analytiques).

Les incertitudes métier sont considérées comme faibles lorsque le praticien est capable de formuler et de justifier l'action envisagée (usage) et sa valeur générée, le cadre de la prise de décision (processus métier, outils, délais de décision...), et la nature de l'information nécessaire à la prise de décision, y compris ses critères de jugement et de confiance. Elles sont fortes lorsque ces questions sont en suspens : le praticien est alors en attente de recommandation ou d'inspiration, que ce soit à travers les usages de ses concurrents, les possibilités offertes par les solutions techniques, ou bien l'aide à l'interprétation des données. L'incertitude métier implique le risque de mauvaise formulation du besoin.

Quant à l'incertitude data, elle est faible lorsque les données sont qualifiées (sens, caractéristiques, processus de construction...), contextualisées, intelligibles pour la prise de décision, valorisables par l'usage, et traitables sur toute la chaîne de valeur (Bertino et al., 2011; Delecroix, 2005; H. G. Miller & Mork, 2013) avec des technologies maîtrisées. L'incertitude data forte est marquée par la difficulté à juger l'exploitabilité des données pour un usage, à commencer par leur accessibilité, leur sens, et la méthode de transformation en information utile à la prise de décision. Elle implique le risque de ne pas générer d'information utile.

Lorsque les incertitudes métier et data sont faibles, la médiation consiste à établir la transmission de l'information vers son utilisateur. Cette transmission se rapproche d'une interaction simple entre objets déterminés, comme lors d'une lecture. En cas d'incertitude data plus forte, la donnée nécessite d'être convertie pour répondre au besoin de l'usage, comme lors de l'établissement d'une requête sur demande formulée par praticien, ou bien au cours d'un

projet de Data Mining doté dès le départ d'un objectif métier, adressable ou non par l'exploration de la donnée. Une incertitude métier forte nécessite, quant à elle, une problématisation, c'est-à-dire une précision du besoin et une validation de l'utilité des informations mises à disposition. Enfin lorsque les incertitudes métier et data sont toutes les deux élevées, l'information utile est issue d'un processus de détermination mutuelle du besoin et de la donnée : il s'agit du mode de médiation le plus complexe, nécessitant à la fois un travail de contextualisation et d'appropriation des informations. Ces 4 modes de médiations privilégiés se transposent facilement à l'idée d'une polarisation du dispositif, et peuvent être représentées sous forme de matrice de maturité croisée (voir Figure 47), où la maturité, en tant qu'accumulation d'expérience et capacité à anticiper les risques, est l'inverse de l'incertitude.

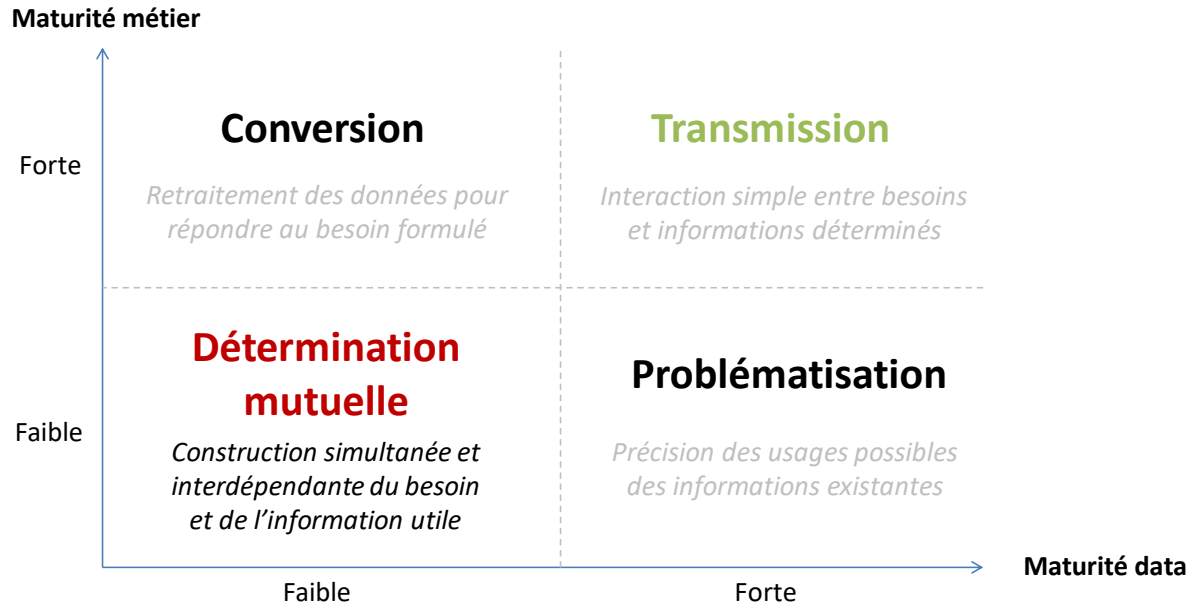


Figure 47 – Modes de Médiation Homme-Données selon la maturité du dispositif

Ce prisme de représentation des incertitudes s'appuie sur l'évaluation initiale de la maturité du dispositif polarisé, et simplifie le choix de la nature du résultat attendu : en effet, le projet peut prioriser la réduction des incertitudes métier, ou bien les incertitudes data, ce qui détermine les ressources dédiées aux instances de médiation. Face au cas le plus complexe, représentatif de l'essentiel des cas d'usages de ces travaux de recherche, le dispositif de Médiation Homme-Données doit générer une dynamique de convergence, sous forme d'arbitrages qui constituent des prises de décision portant sur le projet, en s'appuyant sur une articulation de 4 éléments (voir Figure 48): le capital de connaissances, les algorithmes, les représentations sociales et la gestion de projet.

- **Le capital de connaissances** est l'ensemble des savoirs métier et data disponibles pour un arbitrage, ainsi que l'historique des arbitrages passés. Il représente la maturité du dispositif projet, c'est-à-dire sa capacité à s'autoévaluer, à anticiper la nature et le niveau des incertitudes qui le déterminent, et à choisir les voies optimales de résolution des incertitudes. Il contient l'ensemble des compétences nécessaires à l'anticipation de la production réalisée au cours du projet. Le capital de connaissances peut être porté par des individus, par des documents et par des outils de gestion de connaissances (dont le Databook), à condition que ceux-ci soient intellectuellement accessibles pour les décideurs au sein du projet. Le capital de connaissances est mobilisé à chaque instance de médiation, et s'enrichit au fur et à mesure de la convergence, contribuant à la transition culturelle de l'ensemble des acteurs impliqués.

- **Les algorithmes** constituent la chaîne de transformation de la donnée en information, sous forme d'une suite finie et non ambiguë d'instructions permettant d'aboutir à un résultat à partir de données fournies en entrée. Issus d'un paramétrage humain ou machine, les algorithmes visent soit le résultat final du projet, soit une aide à la décision au cours d'une instance de médiation. Dans les deux cas, ce sont des objets techniques alimentés, modélisés et évalués au cours d'un processus de construction jalonnée par un ensemble d'arbitrages humains marqués par des versions intermédiaires des éléments qui les composent. Ces jalons sont indispensables au co-design et constituent des « prises » pour l'établissement de la relation avec l'ensemble des acteurs impliqués, garantissant ainsi la transparence de l'algorithme.

- **Les représentations sociales** sont des conventions permettant les interactions entre les acteurs humains et les données grâce à leur mise à disposition de façon compréhensible et utile à la prise de décision. Elles sont nécessaires à la définition du besoin et à la contextualisation des données, mais aussi à l'appropriation des résultats intermédiaires ou finaux. Ces représentations sociales comprennent le vocabulaire, les symboles, les indicateurs, les dessins et supports de présentations, mais encore les Data Visualisations et les interfaces homme-machine. La convention des représentations sociales est un préalable à une instance de médiation, garantissant son efficacité. 3 niveaux de Data Visualisation sont constatés dans la pratique : la visualisation des données et des résultats du projet au service des interactions d'alignement entre les acteurs humains et la donnée (création de sens), l'interface de traitement de données et des algorithmes destinée aux Data Scientists (outil de productivité), et l'interface métier destinée à appuyer la prise de décision et embarquant les algorithmes et résultats du traitement des données (aide à la décision). Si la nature de l'activité des Data Scientists nécessite la

deuxième, qui fait par construction partie de l'essentiel des technologies récentes, les deux autres ne semblent pas systématiquement abordées. Pourtant, les enjeux liées à ces éléments semblent clés, notamment en ce qui concerne leur conception : en effet, les choix de types de Data Visualisation (listes, tableaux, arbres, diagrammes, réseaux cartes…), du niveau de détail (agrégats statistiques ou métier, périmètre de restitution…), la sélection des indicateurs et des mises en perspectives multidimensionnelles sont autant de retranscriptions subjectives, de convertissements des données en informations contextualisées et interprétables, mais aussi de dispositifs d'appropriation de connaissances, et donc de représentations sociales de savoirs au service de la génération de valeur.

- **La gestion de projet** est l'ensemble des moyens mis en œuvre pour organiser la mobilisation des ressources du dispositif projet data. Elle comprend notamment les choix de gouvernance, des méthodes de management, des processus, des outils d'évaluation et de pilotage des coûts. Elle permet de cadrer l'inventaire des risques et la priorisation des solutions de résolution en vue d'attendre la finalité du projet. Enfin, elle garantit la pertinence logistique, temporelle et spatiale de la communication entre acteurs humains et non humains, permettant aux instances de médiation d'avoir lieu. La gestion de projet ne semble pas présenter de spécificités particulières qui la distingueraient des pratiques dans d'autres secteurs (projets de conseil et d'audit, gestion concourante dans la conception automobile, conception d'applications web en mode agile…), mais doit être adaptée au niveau des incertitudes. En effet, lorsque le niveau d'incertitudes est bas, la gestion de projet tend vers des processus séquentiels en cascade ; lorsqu'il est élevé, l'incapacité à anticiper fait privilégier les méthodes agiles et la construction de MVP ; enfin, une capitalisation de connaissances suffisante appliquée en contexte de mobilisation de compétences mixtes favorise l'adoption de méthodes concourantes.

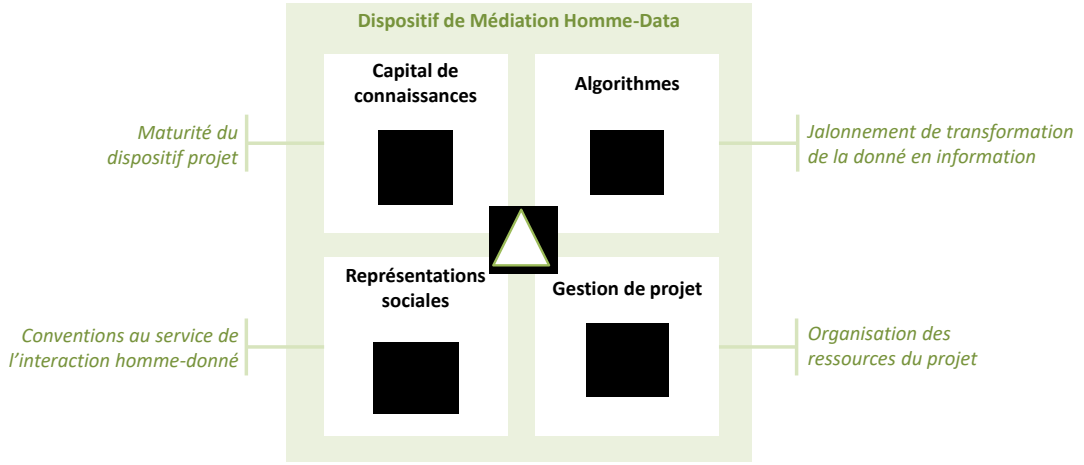
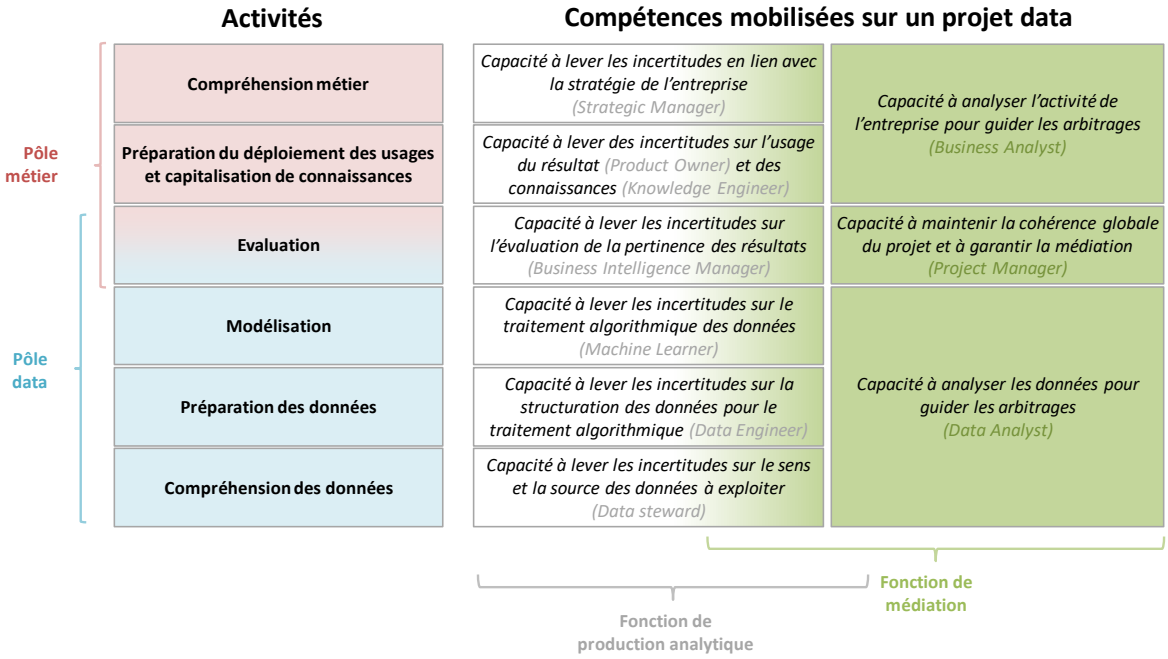


Figure 48 – Les 4 éléments principaux du dispositif de Médiation Homme-Données

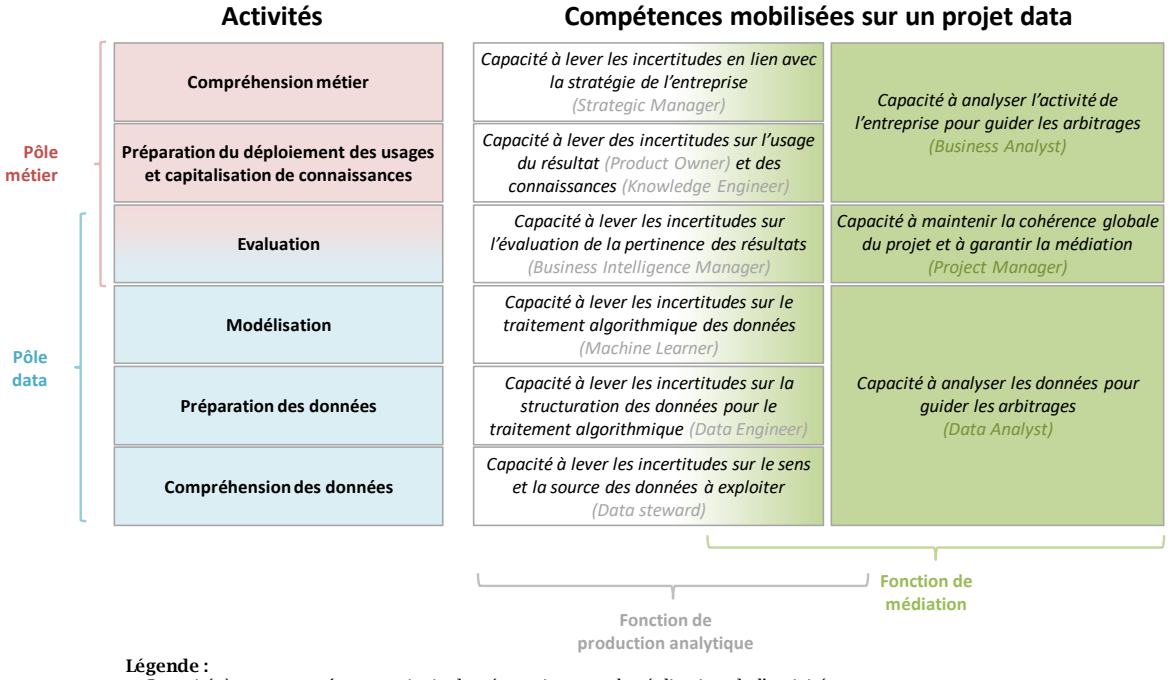
Chaque élément du dispositif est porteur d'un univers qui lui est propre, et peut se doter de ressources dédiées qui viennent alors compléter les ressources nécessaires au travail de production analytique (Hofmann & Tierney, 2009). La spécificité des multiples activités de production analytique nécessite en effet des compétences propres, ainsi que la capacité à anticiper leur exécution en coopération avec l'ensemble des acteurs du dispositif de médiation. Cela soulève les problématiques de montée en maturité du dispositif, des compétences et des niveaux d'expertise, de co-apprentissage et de conscientisation collective, et enfin des responsabilités des acteurs humains, dont les Data Scientists et les praticiens métier. En effet, les instances de médiation s'appuient sur un travail de production (hors chemin critique de la production analytique) réalisé par des analystes métier et data, mais surtout sur l'expertise de l'ensemble des acteurs qui portent la responsabilité de l'exécution de chaque activité du projet. La convergence est alors dépendante de l'expérience de chaque responsable d'activité pour la perception des incertitudes propres à son activité, et placée sous la responsabilité d'une

direction de projet capable d'organiser le dialogue entre l'ensemble des parties prenantes (voir



Légende :
« *Capacité à…* » : **compétence** principale nécessaire pour la réalisation de l'activité
« *(Machine Learner)* » : titre typique du **rôle** sur le projet data, indépendamment des individus mobilisés

Figure 49).



Légende :
« *Capacité à…* » : **compétence** principale nécessaire pour la réalisation de l'activité
« *(Machine Learner)* » : titre typique du **rôle** sur le projet data, indépendamment des individus mobilisés

Figure 49 – Cartographie des compétences mobilisées au cours d'un projet data typique

Cette cartographie met en évidence un besoin de panacher les compétences à mobiliser au cours d'un projet data science, et reste totalement indépendante de la taille du projet, du jalonnement temporel, des choix d'organisation, y compris hiérarchique, et du niveau d'incertitudes qui marque le dispositif. La spécificité des compétences est liée uniquement au chemin critique de la production analytique, et le panachage dépend essentiellement des modes de médiation nécessaires. Dans le cas particulier de besoin de détermination mutuelle (incertitudes fortes en termes d'usages et de données), le rôle du Project Manager, agent médiateur pur et transversal aux pôles data et métier, est clé pour garantir l'alignement de l'ensemble des acteurs : il porte la responsabilité de la médiation, et peut s'appuyer sur des compétences de Data Analyst et des Business Analyst pour produire des informations nécessaires aux instances d'arbitrage. Dans tous les cas, toutes les autres compétences nécessaires à la production analytique (au sens de la réalisation des tâches de production « tangible » du résultat du projet) doivent contribuer aussi à la médiation grâce à leur capacité (plus « intangible », bien que mobilisant des méthodes et supports matériels) à anticiper le résultat de leur propre activité et à le communiquer aux autres parties prenantes pour permettre leur expression d'expertise.

Dans les cas extrêmes, un projet peut mobiliser un seul Data Scientist qui porte l'ensemble des compétences (il alterne alors les compétences sur le chemin critique de la production analytique et réalise seul les arbitrages), et un projet complexe peut réunir des acteurs aux expertises isolées et pointues, voire des équipes dédiées dotées de leurs propres spécificités en termes de compétences. C'est ainsi qu'un projet data visant le déploiement d'une solution applicative peut mobiliser des équipes de développeurs sous le Product Owner, qu'un Project Manager peut s'appuyer sur des compétences plus spécifiques à la gestion de projet, qu'un Data Stewart peut représenter une fonction support data plus large (organisation complexe de gouvernance de données, comprenant notamment le Data Protection Officer en cas de traitement de données personnelles), ou qu'une équipe de Machine Learners peut s'organiser sous la responsabilité d'un Lead Machine Learning expérimenté ou spécialiste d'un type de modèles algorithmiques. Enfin, il semble commun pour un Data Scientist prestataire ou appartenant à un Datalab interne d'assumer tous les rôles sauf celui du Strategic Manager, porté par un demandeur métier, à condition que celui-ci soit capable d'exercer sa fonction de médiation minimale.

La cartographie reste ainsi parfaitement malléable et compatible avec des pratiques plus spécifiques, et le niveau d'incertitudes du projet. La première évaluation des incertitudes comprend l'estimation du niveau de maturité du dispositif, c'est-à-dire le capital de

connaissances initial et la capacité des acteurs impliqués à le mobiliser. Cette évaluation guide l'équilibrage et l'organisation du dispositif, peut s'appuyer sur des outils de diagnostic[45] mis en place par le responsable de la médiation (porteur du rôle de Project Manager) et aboutit sur l'identification des pistes d'apprentissage nécessaires, assurées au cours du projet au sein des instances de médiation sous la forme du développement du capital de connaissances.

La réévaluation des incertitudes réalisée par le dispositif de Médiation Homme-Données grâce à cette mobilisation transversale impacte les usages directs et indirects tout le long du projet data (voir Figure 50). En effet, il révèle à chaque instance de médiation des informations nouvelles pour les acteurs impliqués. Ces informations peuvent être complémentaires, ou conflictuelles avec le capital de connaissances précédent, et donc avec la définition précédente de l'usage cible. L'arbitrage réalisé au cours de chaque instance de médiation consiste alors à éliminer cette nouvelle information (maintien du dispositif projet par absence de prise en compte de la nouvelle information), établir un compromis (adaptation du dispositif avec prise en compte de l'information), accepter l'information (restructuration du dispositif, voire abandon), ou alors résoudre la confusion (poursuite des recherches d'informations complémentaires pour arbitrer à nouveau, dans le cadre du dispositif de médiation ou hors projet). De nouveaux usages peuvent ainsi être générés, que ce soit des usages directs (par compromis ou acceptation) ou indirects (en cas de confusion, nécessitant des investigations complémentaires en dehors du dispositif projet). Par ailleurs des usages peuvent être infirmés, si un nouvel usage est accepté comme meilleur : ce procédé doit être capitalisé afin d'éviter la reproduction des travaux inutiles. Seule l'élimination de l'information nouvelle, issue d'une imperfection du dispositif de médiation, est alors dangereuse pour la qualité de l'usage visé. Ces procédés face à la découverte d'informations expliquent le besoin de souplesse du dispositif et accentuent la nécessité de documenter les arbitrages, c'est-à-dire des choix de procédés face aux flux informationnels internes au projet.

[45] Voir Annexe 9 - L'internalisation des usages dérivant du recours à l'Intelligence Artificielle dans les entreprises, ou l'exemple d'outil d'évaluation initiale des compétences développé pour Quinten : http://blog.quinten-france.com/questionnaire-diagnostic-big-data-quelques-minutes/

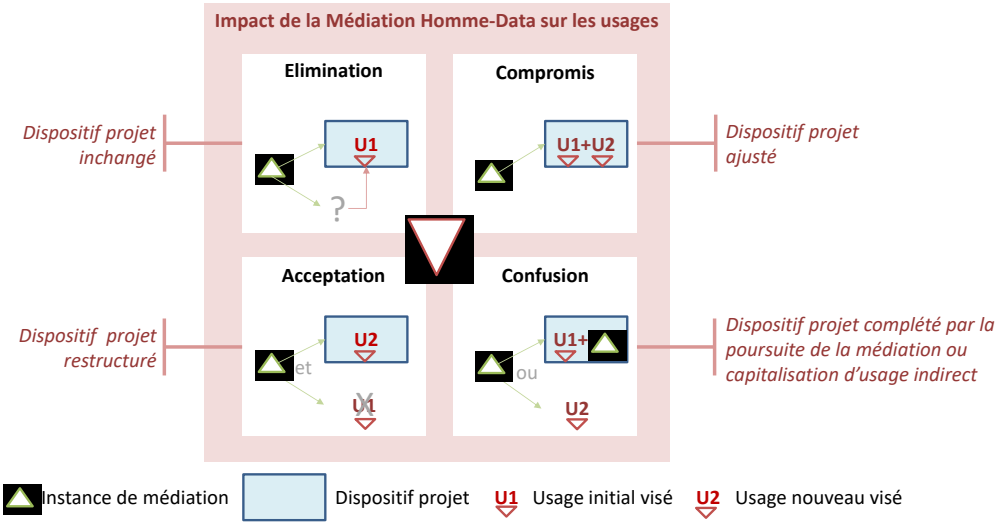


Figure 50 – Impact de la Médiation Homme-Données sur les usages visés par le dispositif projet data

Ainsi, le dispositif de Médiation Homme-Données contribue à la génération de la valeur et à la maîtrise des ressources du dispositif projet data. La maturité du dispositif projet data, et plus particulièrement celle de la Médiation Homme-Données, constitue le principal vecteur de réduction des incertitudes projet, aux côtés de la production analytique. En effet, si l'interaction d'apprentissage constitue un levier potentiel sur le coût d'un projet data (coût de l'expertise, productivité, nombre d'acteurs mobilisés pour un « bundle » de compétences suffisant...), l'interaction d'alignement présente un intérêt particulier car elle s'inscrit au cœur de la création de valeur pour l'organisation par sa capacité à impacter les usages. Cette distinction porte aussi bien sur la nature des savoirs capitalisés à l'issue de ces interactions : les premiers constituent des ressources complémentaires et des gains de productivité pour la réalisation d'autres projets data, alors que les seconds irriguent plus largement la culture et l'activité de l'entreprise. Le lien entre les interactions d'apprentissage et d'alignement n'est établi que lorsqu'il s'agit de mobiliser des expertises, c'est-à-dire l'expérience passée sur chaque activité, au service de la médiation.

La montée en maturité du dispositif est directement liée à la cartographie des compétences, et renvoie au paradoxe actuel du rôle du Data Scientist : en effet, ce métier est à ce jour mal défini et englobe plus ou moins l'ensemble des compétences cartographiées. Un Data Scientist est alors un individu qui détient un « bundle » de compétences, dont l'ensemble des expertises nécessaires à la médiation. Sa montée en maturité provoque un double mouvement d'évolution

des compétences : une spécialisation sur chaque activité par l'apprentissage (par exemple, une montée en expertise sur certains algorithmes de Machine Learning), et une accumulation de compétences généralistes à travers la contribution aux instances de médiation. Avec la complexité croissante des projets data et le gain d'expérience, les individus ont le choix de s'orienter vers des rôles d'expert contributeur (pour chaque activité du chemin critique de la production analytique) ou de Project Manager, médiateur généraliste capable d'animer la convergence entre experts. Cette évolution assure la montée du niveau d'exigences et de complexité des projets data.

La professionnalisation du rôle de médiateur humain est ainsi conditionnée à la capitalisation de connaissances issues des activités de contribution aux instances de médiation. Sa responsabilité est avant tout de garantir l'efficience des instances de médiation au cours du projet data (verticales dans le modèle Brizo_DS), et donc la cohérence du dispositif dans un contexte marqué par les incertitudes : cette cohérence tient à la capacité de l'ensemble des contributeurs experts à arbitrer progressivement sur leur zone de responsabilité spécifique (horizontales dans le modèle Brizo_DS). Les incertitudes métier et data sont en effet placées sous la responsabilité des experts contributeurs : ainsi, lorsque le dispositif projet data prend fin, l'exploitation de l'usage n'est plus conditionnée que par les incertitudes restantes, ce qui place les contributeurs experts en première ligne en termes de responsabilités, en tant que décideurs passés au cours des instances de médiation. Ce partage de responsabilité entre les contributeurs experts et le médiateur doit alors être traçable et documenté, en particulier dans le cadre de l'assurance de la transparence des algorithmes embarqués dans les usages et de la mesure de la valeur générée par ces usages.

Ainsi la Médiation Homme-Données s'inscrit dans des enjeux plus larges que le simple support à la convergence au sein d'un projet ou d'un portefeuille de projets data. Elle constitue un vecteur de responsabilisation, de transparence, et d'évaluation du potentiel de valeur par les nouveaux usages.

***

A la lumière des limites théoriques du modèle de référence CRISP_DM et des observations sur les dix cas d'usage terrain, ce travail de recherche aboutit ainsi à une proposition de nouveau modèle de dispositif de projet, BRIZO_DS, qui encapsule un modèle de processus CRISP_DM ajusté (revue de la dynamique, livrables intermédiaires complétés et catégorisés) pour l'ancrer

dans la génération de valeur par les usages et l'enrichir avec la qualité des données et avec 4 facilitateurs de la Médiation Homme-Données (capitalisation de connaissances, livrables intermédiaires et versionning documentés dans un Databook, représentations sociales comprenant la sémantisation, les symboles, ou encore les interfaces homme-donnée, et enfin la gestion de projet selon le niveau d'incertitudes et jalonnée par des instances de médiation). Ce modèle est enseignable, opérationnel, assez flexible pour être conciliable avec les différents courants de gestion de projet, et compatible avec une gestion de portefeuille de projets data, générant des usages directs et indirects.

3 Discussion des limites de ces travaux de recherche

Ces travaux de recherche sont marqués par la diversité des cas terrain en termes de secteurs (santé, assurance, industrie du parfum…), de fonctions des entreprises (finance, achats, production, R&D, conformités, marketing…) de sujets et de contextes d'intervention. Les projets menés ont mobilisé une vingtaine de profils de « Data Scientists » aux compétences différentes et aux niveaux d'expérience variés : les Data Scientists les plus expérimentés ont plus de 10 années de pratique dans ce domaine d'activité récent, qui ne se développe de façon massive que depuis seulement 2011 en France. Les méthodes de travail choisies et les approches algorithmiques ont permis d'envisager un éventail à spectre large des projets data, sans se restreindre à un modèle algorithmique particulier. Par ailleurs, les technologies mobilisées, et notamment les solutions choisies au cours de la phase d'exploration des données volumineuses (Spark…), semblent représentatives du travail réalisé au sein d'autres entreprises françaises par les équipes de Data Scientists sur les phases de conception de solutions data. Cette richesse du travail qualitatif approfondi sur le terrain justifie l'ajustement du modèle de référence, et donne la possibilité d'envisager l'élargissement à un nouveau modèle, généralisable et ancré dans la pratique. Cependant, le terrain reste porteur de certaines spécificités (culture managériale et évolution de l'offre chez Quinten), la recherche action ne s'affranchit pas de la subjectivité des propositions, et le marché n'a pas fini de se stabiliser.

3.1 Spécificités du terrain chez Quinten

Cette richesse du terrain semble limitée par la spécificité de l'environnement de travail de l'entreprise Quinten. En effet, la société est marquée par une culture propre, basée notamment sur un style managérial spécifique aux fondateurs, et plus particulièrement à la Direction des Opérations qui centralise l'ensemble des projets. Il s'agit historiquement d'un management persuasif et paternaliste, avec une implication forte du management sur les arbitrages réalisés au cours des projets et reposant sur des collaborateurs de confiance qui bénéficient d'une autonomie suffisante pour ne pas subir des procédures strictes. Ce style managérial privilégie la co-construction et la mixité des profils dans une équipe projet, et reste assez imperméable à l'évolution des méthodes en dehors de la société, et notamment des méthodes « à la mode ». Ces travaux restent ainsi limités en termes d'observation de méthodes dites agiles, bien que le modèle proposé soit théoriquement compatible avec celles-ci. La société commence en 2017 à

appliquer ces méthodes dites agiles au sein des équipes de développement d'applicatifs, ce qui correspond à une nouvelle activité sur laquelle Quinten ne possède pas un recul aussi significatif que sur les projets Data Science. En attendant, le besoin d'appliquer les méthodes agiles en Data Science n'est pas ressenti dès lors que l'équipe comprend des ressources expérimentées : les méthodes concourantes ou en cascade restent privilégiées. L'application des méthodes agiles par des équipes Data Science est même perçue comme plus symptomatique d'une immaturité du dispositif projet (incapacité à anticiper par manque d'expérience et volonté d'apprendre en faisant) que comme un véritable choix de méthode pour ses avantages, telle qu'elle est appliquée dans les milieux expérimentés de développement d'applicatifs pour favoriser la création et établir un lien plus direct entre les pratiques des utilisateurs et les développeurs pour la conception d'une interface. Ce point reste en suspens en absence d'observations approfondies.

Par ailleurs, la société est historiquement marquée par une approche inverse des sujets data : ayant développé un algorithme supervisé non paramétrique, le Q-Finder, Quinten a été marqué par une approche commerciale « en push », c'est-à-dire à une proposition de génération de valeur par l'algorithme, guidant le choix de l'usage, et non pas par l'usage qui guide le choix algorithmique. Cette approche commerciale est alors compensée par une phase de cadrage de l'usage dès le démarrage du projet, avec une implication forte des experts métier, et a été progressivement diminuée au profit d'une approche conseil, y compris dans les phases d'avant-vente. Cependant, cette spécificité ne semble pas limitante dans la mesure où le processus projet, développé autour de l'algorithme propriétaire, n'a pas été bousculé par la montée en puissance de l'utilisation d'autres algorithmes du marché, observée dans les projets. Bien au contraire, il semble que l'utilisation de l'algorithme propriétaire, marqué par un fonctionnement privilégiant la production de résultat explicites sous forme de règles métier interprétables, ait guidé la construction d'une méthode de projet data qui répond de façon plus pertinente à la montée du besoin de la transparence des algorithmes.

Enfin, Quinten prend, en 2015, le tournant du marché vers les produits, incluant les interfaces de restitution des résultats, et ce pour deux raisons. Tout d'abord, le marché est alors en attente de solutions sous forme d'applicatifs métier, embarquant les algorithmes dans un usage opérationnel. Cette attente est perçue comme une opportunité commerciale qui peut permettre de transformer des projets d'analyse ponctuels, proches d'une activité de bureau d'études, à une offre produit récurrente, proche d'une activité d'éditeur de solutions sur mesure. Ces solutions

font alors l'objet de la mise en place d'une équipe de développeurs. La seconde raison est la prise de conscience de l'importance de l'interface de restitution des résultats algorithmiques. En effet, les projets Data Science sont parfois complexes, et l'appropriation des résultats est difficile en s'appuyant seulement sur des supports papier, pdf, ou csv. Les interlocuteurs métier ont besoin de manier, d'interagir avec les données et les résultats pour s'assurer de leur sens. Plus particulièrement, les résultats de l'algorithme propriétaire peuvent être difficiles à interpréter en cas d'identification d'un nombre important de règles métier : leur interprétation est alors largement facilitée grâce à la mise à disposition d'une nouvelle interface de Data Visualisation, plus ergonomique et intuitive que celle qui était destinée auparavant aux seuls Data Scientists. Que ce soit dans le cadre du développement de solutions métier sur mesure ou de l'aide à l'interprétation des résultats, la Data Visualisation apparait comme un vecteur de convergence fondamental entre les parties prenantes du projet, et contribuent à la co-construction des usages.

Au cours des années observées, Quinten ne faisait pas partie des acteurs qui intervenaient sur des sujets de changement de Business Modèle des entreprises. Ces projets ont commencé à émerger dans l'éventail de l'offre seulement à partir de 2017. A priori, le modèle n'est pas inadéquat pour couvrit ce type d'usages, cependant d'autres mécanismes devraient être ajoutées dans le dispositif, comme la valorisation marché de l'information, la prise en compte de la concurrence, ou encore l'intégration de partenaires et des investisseurs. Ainsi, la culture managériale, l'algorithme historique, la spécificité de l'emploi d'outils de Data Visualisation et l'offre de Quinten ont influencé le modèle proposé dans ces travaux de recherche, et peuvent constituer des biais de sélection significatifs : il serait approprié de confronter le modèle proposé à des environnements différents, comme des Datalabs internes aux entreprises ou alors des cabinets concurrents, en France ou à l'étranger, pour confirmer sa robustesse.

3.2 Limites de la recherche action

La recherche action a, elle aussi, grandement impacté le résultat. En effet, ma connaissance limitée sur la conception d'algorithmes a nécessité une compensation par d'autres compétences développées par le passé, afin d'être productive sur les projets en tant que salariée de la société. Or, ce passé est fortement marqué par une approche métier, notamment les métiers financiers, ainsi que par les méthodes issues du conseil. Ce passé pousse à considérer les projets data comme un moyen de construire un avantage compétitif pour une entreprise, que ce soit sous forme de génération de connaissance ou de leviers opérationnels. L'usage est ainsi placé en

avant, tout comme dans des projets de transformation habituels, qu'ils touchent aux organisations, aux processus, aux outils IT ou autres. Ainsi, la complémentarité entre l'approche conseil et la Data Science a largement contribué aux interactions observées sur les projets. Cette posture, à l'interface entre la production Data Science et le besoin métier émergeant, a permis par ailleurs la conception d'outils qui fluidifient la communication, comme le Databook ou les supports de restitution des résultats embarquant l'adaptation des représentations sociales métier, y compris les éléments de traduction de résultats statistiques en valeur pour l'entreprise. L'approche conseil et métier, orientée sur la génération de valeur, est complétée par la pratique de la conduite de changement sur les projets passés, avec une prise en compte systématique des utilisateurs, et la mise en place de dispositifs de communication et de transfert de connaissances.

De plus, la pratique de différents modes projets m'a permis de mettre rapidement cette compétence au service des projets data. Cela s'est traduit par la mise en place d'instances projet, habituelles en conseil et inexistantes dans la pratique Quinten, comme les Project Management Office, les outils de pilotage projet, le jalonnement des offres commerciales (Go/No Go), et l'animation des instances d'arbitrage récurrentes sur les projets complexes. L'efficacité de ces méthodes a été prouvée par des gains de productivité internes, et la satisfaction des acteurs métier face à l'apparition des instances d'arbitrage, plus ou moins fréquentes selon le contexte.

Ainsi, si la recherche action a été un choix pertinent pour l'observation du terrain et pour la possibilité de valider des hypothèses au fur et à mesure de leur émergence, le risque de biais liés à la subjectivité du modèle, liée à mon expérience en conseil transposé à la Data Science, peut constituer une limite dans son application, et notamment dans l'exercice d'une Médiation Homme-Données en absence d'une expérience en conseil, métier, valeur et projet. Cette limite ne semble pas constituer une barrière particulière à l'adoption des pratiques au sein de Quinten, puisque les méthodes propres au conseil sont en cours de transposition aux pratiques de l'entreprise sur les autres projets en Data Science. L'absorption de ces apports, subjectifs, constitue en soi un axe de validation de la pertinence des hypothèses du modèle, en tant que reconnaissance de l'intérêt par l'équipe des Data Scientists. Par ailleurs, l'acceptation de nouveaux projets par les clients et leurs feedbacks sont autant de marqueurs d'intérêt pour la génération de valeur produite par cette approche. Cependant, il reste difficile de distinguer l'apport individuel (en tant que cadre dirigeant de la société) de l'apport du modèle, ce qui nécessite une application du modèle par d'autres praticiens. Si la reproductibilité des résultats semble difficile, étant donné la posture en recherche action, aucune raison n'est identifiée à ce

stade pour empêcher la validation du modèle proposé par une application sur d'autres projets en Data Science.

3.3 Un marché non stabilisé

Enfin, ces travaux de recherche s'inscrivent dans une temporalité sur le moyen terme (3 ans de pratique terrain) qui ne peut être représentative de l'évolution du marché. Les entreprises clientes sont immatures sur ces sujets, et se transforment : cela explique le besoin de faire évoluer le modèle de référence, mais ne décharge du devoir de confirmer la pertinence du modèle dans les années qui viennent. Les technologies évoluent, elles aussi, et peuvent conduire à dépasser certaines pistes de réflexion, grâce notamment à la facilitation de la médiation par de nouvelles solutions de traçabilité ou d'appropriation des algorithmes. Enfin, le manque de ressources expérimentées disponibles sur le marché peut remettre en cause le jugement sur la diversité des compétences mobilisées sur ces projets, et la pertinence des rôles établis sur les projets observés. Cependant, le modèle proposé, et notamment les quatre dimensions de la Médiation Homme-Données, semble approprié pour guider dès aujourd'hui la montée en maturité des entreprises, le développement des technologies, et la structuration des programmes pédagogiques pour la formation des ressources. Sa flexibilité, liée à l'absence de distinction entre les acteurs humains et non humains, peut s'adapter à l'évolution des technologies et des compétences. Son orientation sur la valeur devrait être compatible avec les besoins des entreprises, que ce soit à travers la mise en place des usages opérationnels ou la capitalisation de connaissances. Enfin, sa structure rapproche le dispositif data de dispositifs projet plus classiques grâce à l'ajout d'éléments issus de la gestion de projet, de la conduite de changement, et l'orientation sur la génération de valeur. Cela simplifie son application et ouvre la voie non seulement à la gestion de portefeuille de projets data, mais aussi au rattachement de ce portefeuille de projets data aux portefeuilles de projets classiques. Ce rapprochement semble inexorable avec la démystification progressive du phénomène Big Data et la montée en maturité en France sur ce sujet.

Conclusions et perspectives de recherche

Que constate-t-on en descendant sur le terrain, en France, au milieu de la première décennie du buzz Big Data ? Une grande diversité des maturités et des attentes face au phénomène, et des projets très prometteurs, mais pas dans le sens espéré.

En effet, les usages en entreprise sont bel et bien impactés : au passage d'un projet data, les praticiens métier se dotent de nouveaux outils, de nouvelles compétences, d'informations utiles à l'activité. Ils accélèrent ou perfectionnent leurs activités historiques et leurs produits, et plus particulièrement les produits informationnels, et capitalisent des connaissances, parfois même inédites. Ces usages ne sont pas révolutionnaires pour leur activité, et les projets qui vont jusqu'au bout de la mesure de la valeur générée se comptent sur le bout des doigts. Dans ces cas-là, marqués par une maturité suffisante des experts métier et une expérience analytique, la performance des usages peut même dépasser les attentes, ce qui en effet ouvre des perspectives prometteuses. En revanche, ces projets contribuent à démystifier peu à peu les opportunités offertes par le Big Data et la Data Science, à renforcer la capacité d'innovation et la capitalisation de connaissances. Dans ce contexte, parler d'une génération de valeur indéniable est difficile, mais pourtant approprié. Il est nécessaire d'assumer cette génération de valeur peu quantifiable, et de la favoriser au maximum pour monter en maturité. Or, les dispositifs de projet data, basés sur des modèles de processus assez anciens, comme CRISP_DM (Chapman, 1999; Shearer, 2000) et en pleine évolution sur le terrain, ne sont pas encore optimaux, ni pour produire de façon efficiente des résultat analytiques utiles, ni pour contribuer à la capitalisation de connaissances. Décollant difficilement de leurs fondements techniques et analytiques et des processus de production analytique, robustes mais trop autocentrés, les projets data ont encore du mal à s'ancrer dans les pratiques métier.

Cette convergence est pourtant perfectible, et l'observation de 3 années de projets data permet de dégager des outils théoriques et opérationnels pour y parvenir.

1 Un nouveau modèle de dispositif « projet data » : Brizo_DS

A la lumière de 4 années d'observation terrain, la réponse à la question initiale est évidente : les processus théoriques existants, propres aux projets data, sont bien appliqués sur le terrain, mais sont contournés, et donc peuvent être rendus plus efficaces. Le modèle global de dispositif de projet data, Brizo_DS, proposé à l'issue de ces travaux de recherche action, est fondamentalement orienté vers la génération de valeur par l'exploitation des usages qu'il permet de concevoir. Le modèle s'appuie sur un processus CRISP_DM ajusté, orienté sur l'usage et jalonné par des instances de Médiation Homme-Données, et va au-delà du processus pour décrire le dispositif dans son ensemble, en tant qu'agencement d'acteurs humains et non humains.

Les ajustements du processus CRISP_DM sont issus d'un processus de maturation des équipes mobilisées sur les projets data depuis 2008, que ce soit à travers l'appropriation des nouvelles technologies ou à la prise en compte de l'évolution des besoins des entreprises. En effet, les besoins des métiers ont évolué suite à la mise en œuvre des premiers projets data : les travaux réalisés par les équipes data dédiées, parfois dans la foulée d'une nomination de Chief Data Officer, ont souvent porté des fruits du point de vue statistique, mais n'ont pas donné lieu à la mise en œuvre opérationnelle des résultats. Les équipes expriment alors le besoin d'ajuster la méthode projet pour inscrire les résultats de façon plus pertinente et durable dans les pratiques opérationnelles. L'évolution des technologies donne, quant à elle, la possibilité d'exploiter directement l'ensemble des données disponibles sans passer par une phase d'exploration sur un échantillon réduit : cela implique le transfert de la phase de sélection des données en amont du projet afin de réduire le risque de submersion au cours de la compréhension du sens des données initiales. Elle implique aussi le besoin des entreprises, et en particulier des Directions Informatiques, de mieux comprendre les enjeux de valorisation des données liées au déploiement de nouveaux moyens de stockage de l'information comme les Data Lakes, et donc de mettre en place de nouvelles méthodes de gestion de la qualité des données, notamment lorsqu'il s'agit de données « désilotées » impliquant des responsables de qualité issus de métiers différents. Enfin, la profusion des outils de Data Visualisation personnalisables à coût réduit et

utilisables avec des volumes de données significatifs s'inscrivent dans les opportunités nouvelles à prendre en compte pour le partage des résultats. Or, il s'agit d'un levier d'appropriation des résultats par leurs utilisateurs, ce qui répond à la nécessité initiale d'enraciner les projets data dans le métier.

Dans ce contexte, la prise en compte des représentations sociales, à la frontière entre les possibilités offertes par les outils décisionnels de dernière génération et l'intelligibilité de l'information au service de la prise de décision, est indispensable. A la lumière de ces résultats, l'ajustement du modèle CRISP_DM apparait comme indispensable pour donner des prises sur le processus aux métiers jusqu'à présent désintermédiés, comme l'informatique et le décisionnel, et inclure la notion de création de valeur métier et les technologies Big Data au cœur d'un processus. Cette évolution fait alors basculer le processus qui décrit le Data Mining traditionnel vers la « Data Science ».

Au-delà de l'amélioration de la pertinence du modèle de référence dans le contexte actuel, le modèle global Brizo_DS permet de situer le dispositif projet data dans une temporalité plus large en entreprise, en l'ancrant en tant que vecteur de réduction des incertitudes dans la génération de nouveaux usages. Cet ancrage répond à un besoin de démystification du processus exploratoire auprès de l'ensemble des contributeurs, fluidifie le pilotage du dispositif et son intégration dans le portefeuille des projets de l'entreprise au même titre que des projets plus classiques, et justifie son alignement sur les méthodes de génération de valeur, ce qui favorise la capacité du dispositif à mobiliser les ressources qui lui sont nécessaires. En effet, le modèle proposé rompt le caractère autocentré du processus de référence pour le rendre plus adéquat avec les pratiques de l'entreprise bénéficiaire des résultats.

La conception du modèle a été réalisée en observant 3 dimensions clés, insuffisantes dans les modèles théoriques de référence : la prise en compte d'indicateurs de valeur, de la qualité des données et de la médiation entre les hommes et les données. La confirmation de la pertinence des choix de ces dimensions peut aujourd'hui être réalisée dans d'autres contextes d'exécution de projets data. Cette confirmation, notamment du point de vue quantitatif, peut être réalisée grâce à la mise en place d'un dispositif d'observation de projets data, en dehors de Quinten, selon une méthode d'observation d'ores et déjà définie et testée sur l'un des cas observés. Simple à utiliser pour un participant ou un observateur de projet data (voir Annexe 7 – Grille d'analyse des études de cas selon une approche quantitative), elle peut permettre de confirmer, infirmer ou faire évoluer le modèle proposé et ses dimensions. En outre, elle permet d'établir

des indicateurs quantitatifs pour mettre en évidence le poids et la nature de l'impact des facteurs de succès afin de prioriser au mieux les pistes de recherche et les investissements dans les projets.

Encore faut-il être conscient de la valeur qu'il peut véritablement générer pour investir.

2 La valeur des projets data

L'observation de cas terrain confirme indéniablement la transformation des processus de production en entreprise à travers les projets data qui impactent les ressources, les activités et les résultats de ces activités à l'échelle des fonctions bénéficiaires. Cette approche factuelle rend les projets data assimilables à tout projet classique, avec les mêmes difficultés à mesurer la valeur des connaissances générées. Elle n'est pas suffisante pour appréhender les spécificités de la valeur générée : en effet, ces projets constituent de formidables leviers de réduction d'incertitudes sur les usages métier, ainsi que des vecteurs de valorisation des données. Cette valeur est capitalisable, à condition de documenter le processus de construction cognitive.

2.1 La valeur de la réduction d'incertitudes

Trois apports majeurs des projets data sont identifiables dans les projets étudiés : la conception d'une solution algorithmique pour accélérer la prise de décision (l'usage est d'ores et déjà identifié, mais une incertitude existe quant à sa faisabilité analytique), la génération de connaissances (l'imagination de nouveaux usages est alors conditionnée à la découverte de nouvelles informations utiles, là où les méthodes et techniques de découverte habituelles ont été épuisées), et la montée en maturité plus radicale de l'entreprise qui souhaite améliorer sa capacité à innover (dans ce cas-là, elle ne sait pas définir l'usage, et ne sait pas si un projet data serait la meilleure façon d'y arriver). Les projets sont alors *a priori* assimilables à de la R&D ou à du conseil stratégique et opérationnel, par leur capacité à lever des doutes dans un contexte incertain.

Mais lorsque le mythe sur l'immense valeur dormante des données se déverse sur un terrain d'ores et déjà sous pression financière, la volonté de mesure quantitative est exacerbée, tout comme la croyance dans la capacité d'une machine à tout quantifier, y compris le capital de connaissances et la plus-value sociale. Or, si un algorithme peut accélérer un apprentissage en l'automatisant à l'ère du « Machine Learning », il ne peut aller jusqu'à la qualification des connaissances générées en termes d'utilité sans une intervention humaine. Cette quantification se heurte à la capacité de l'homme à identifier et juger ce qui est utile.

Deux inconnues sont clés. L'usage, envisagé lors du lancement du projet ou imaginé comme résultant d'une découverte par la fouille des données, est-il utile pour l'entreprise ? L'information contenue dans les données serait-elle utile à la prise de décision ? Si le travail de production analytique peut parfaitement répondre à la seconde question, il ne peut aborder la première. Or, la première justifie habituellement le lancement d'un projet. A la lumière de ce constat, un projet data ne peut pas être considéré comme un projet générateur de bénéfices classique, pilotable à travers un ROI ou un autre indicateur similaire, mais comme un vecteur de réduction d'incertitudes. Cette réduction doit être faite à la fois côté métier (sur l'usage) et côté data (sur l'information), et ce de façon convergente. Cette réduction de l'incertitude possède alors un coût divisible en deux parties : les coûts liés à la réduction d'incertitudes analytiques et la médiation qui permet de faire converger les résultats sur un usage exploitable et bénéfique. Le niveau de maturité du dispositif constitue un point de départ pour décider les proportions de l'investissement dans ces deux activités. Jusqu'à présent, l'estimation *a priori* du coût de ces activités reste difficile, en particulier en absence de références externes. La justesse de l'estimation du coût du projet, à mettre en perspective avec les bénéfices, est donc directement liée à l'expérience.

Les cartographies des risques spécifiques aux projets data n'ayant pas été établies, et peu d'entreprises ayant expérimenté assez de projets pour parler véritablement d'une connaissance des risques, il s'agit plutôt d'incertitudes, plus ou moins perceptibles selon le niveau d'expérience des acteurs concernés par les usages et par le travail analytique. Si la notion d'incertitude est en soi une brique clé du modèle proposé, elle est à ce stade synthétique, c'est-à-dire catégorisée en 4 éléments (stratégique, opérationnel, analytique et projet). Or, la complexité croissante des projets, leur transversalité, et la multiplication des retours d'expérience devraient conduire à une appréhension plus détaillée de ces incertitudes, et bien évidemment à l'identification des compétences capables de les percevoir et de les résoudre. Cette décomposition des incertitudes s'inscrira alors dans un cadre d'analyse de risques plus commun. Elle s'adresse aujourd'hui à la fois aux chefs de projet data (pour identifier les contributeurs) et aux membres contributeurs (pour percevoir la part de l'incertitude, appréhender les facteurs de risque, et anticiper les impacts et la gravité afin de guider les arbitrages projet selon leur champ d'expertise). Il s'agit en particulier des risques règlementaires ou juridiques, clés dans un contexte d'évolution des règlementations propres à chaque secteur d'activité ou liées à l'usage des données en entreprises, et notamment des données personnelles. Certains acteurs en entreprise, comme les DPO, sont d'ores et déjà

identifiables comme contributeurs dans ces projets. Plus largement, l'éthique de l'utilisation des algorithmes est un message qui reste à diffuser auprès de l'ensemble des contributeurs afin qu'ils puissent détecter rapidement des dérives potentielles à chaque étape du projet data, en amont de la mise en exploitation des usages.

Cette intégration de l'éthique dans la pratique des contributeurs semble aujourd'hui freinée par la méconnaissance des algorithmes : le manque de maturité sur ce concept, sa mystification par le discours ambiant, et sa complexité empêchent les contributeurs de se positionner en tant que tireur d'alarme. Certains risques propres aux aspects purement analytiques de ces projets, comme le sur-apprentissage ou les facteurs de confusion qui compliquent des liens de causalité, ont pourtant déjà été soulevés (Provost & Fawcett, 2013a). Ces facteurs spécifiques peuvent rendre les usages inopérants, voire dangereux. A l'heure où la CNIL soulève la réflexion sur les enjeux éthiques et sociétaux liés à l'usage des algorithmes[46], tout en posant les principes de loyauté et de vigilance, il est nécessaire d'aborder ce sujet de façon pédagogique et dédiée au cours des projets. En effet, la transparence des algorithmes n'a pas à être perçue comme une contrainte imposée par le régulateur : c'est une incitation à la coopération. L'identification de ce sujet permet de consacrer les ressources nécessaires à cette question, et aborder les algorithmes comme le fruit d'une relation entre acteurs dont la forme de collaboration conditionne la valeur de leur usage (voir Annexe 5 - Transparence des algorithmes).

Ainsi, ces travaux de recherche dressent une cartographie des risques très exploratoire qui nécessite d'être confrontée à l'expérience d'autres Data Scientists et des métiers afin de dégager de façon plus éclairée les facteurs de risque, les impacts sur les projets, et les solutions possibles pour les prévenir ou les contrer. La solution n'est en effet pas perçue seulement comme analytique, car elle est aussi et surtout liée à l'usage. Cette cartographie pourrait constituer un outil de pilotage du dispositif projet data, et s'inscrirait parfaitement dans le modèle d'évaluation basé sur une anticipation itérative des incertitudes métier et des incertitudes data pour arbitrer sur les réajustements de ressources éventuelles en fonction des bénéfices attendus.

Aux deux incertitudes liées à l'utilité des usages pour l'entreprise et à l'utilité des informations, une troisième s'ajoute : quelle est l'utilité des usages basés sur IA en entreprises à l'échelle d'une société ? Cette question reste en suspens à l'issue de ces travaux, très orientés sur les

[46] https://www.cnil.fr/fr/comment-permettre-lhomme-de-garder-la-main-rapport-sur-les-enjeux-ethiques-des-algorithmes-et-de

entreprises en tant qu'organisations dans un marché concurrentiel. Pourtant, elle présente un intérêt indéniable et mérite d'être examinée sous l'angle du rôle que les secteurs d'activité jouent dans une société. Par exemple, comme le formule François Ewald au cours d'une intervention sur l'Assurance début 2018, quel rôle peut jouer l'Intelligence Artificielle pour innover l'Assurance en tant que finalité sociale ? La médecine personnalisée ? La surveillance ? Comment le concept d'incertitude se décline dans ces domaines et peut-il être réinventé à la lumière des possibilités offertes par les projets data, vecteurs de réduction d'incertitudes ? En absence de penseurs et d'experts sur ces questions dans les projets en entreprise, la finalité des projets reste très orientée sur l'optimisation des pratiques métier existantes, et non pas sur leur renouvellement.

2.2 Databook : une mémoire de la dynamique de construction des algorithmes

En attendant que les incertitudes soient levées sur ces incertitudes, et que les incertitudes deviennent des risques cartographiés pour ne laisser que celles liées à la découverte d'un signal utile dans les données, ces travaux de recherche-action incluent une proposition de cadre simplifié pour piloter les dispositifs data selon des indicateurs simples de bénéfices, de ressources et d'incertitudes. L'un des facteurs de risque identifiés concerne plus particulièrement le manque de gestion de qualité des données, et plus particulièrement des nouveaux modèles de données sous la forme d'algorithmes. Ces travaux formulent une proposition d'outil concret, le Databook, permettant la capitalisation de connaissances par la documentation de la dynamique de qualification des données tout le long du processus de leur transformation en résultat analytique utile. Il pose un cadre de gestion des nouvelles métadonnées correspondant aux métriques propres aux algorithmes, car sa structure le rend compatible avec toutes les étapes d'un projet Data et avec le modèle global proposé. Il constitue par ailleurs un outil privilégié de la Médiation Homme-Données, servant de support au cours des instances d'arbitrage au cours du projet data. Cet outil reste à ce jour à confronter avec des outils similaires sur le marché, et à optimiser en termes techniques et fonctionnels, bien qu'il soit d'ores et déjà utilisé en pratique au sein des projets data et fasse partie intégrante des formations des Data Scientists à la méthode projet.

Si les bénéficiaires actuels du Databook sont restreints aux membres de l'équipe projet, il semble parfaitement convenu pour un usage plus large en entreprise, et notamment par les Data

Stewards et autres métiers intervenant en support transversal des sujets liés à la qualité des données, à leur gouvernance, et à leur mise à disposition au profit des utilisations internes. Les Knowledge Managers peuvent y trouver une source d'informations riche et évolutive, en particulier si cet outil de qualification des données est adapté au contexte de l'entreprise et comporte l'ensemble des représentations sociales nécessaires à l'intelligibilité de cette matière première. Le principe du Databook peut être utilisé pour l'amélioration des outils informatiques dédiés aux traitements des données, par exemple dans le cadre de la qualification des données en amont ou en aval d'un Data Lake. Enfin, il est envisagé comme une base évolutive vers un processus de valorisation financière rétroactive des données en tant qu'actif de l'entreprise : il trace en effet tout le processus de transformation d'une donnée en usage générant de la valeur, donc si un usage fait l'objet d'une mesure financière, celle-ci peut être redescendue à la maille de la donnée brute.

Face à la diversité des utilisations possibles, il semble nécessaire de confronter la structure du prototype de l'outil aux besoins des parties prenantes possibles, en commençant par les responsables de la gestion de la qualité des données en entreprise, absents des projets data vécus sur le terrain. Cette confrontation pourrait donner lieu à une mise à niveau en termes d'état de l'art de la pratique avant de faire l'objet d'un échange plus élargi à d'autres acteurs, comme les Knowledge Managers, la DSI et la Finance.

En attendant cette confrontation théorique et pratique, l'usage immédiat du Databook reste la documentation des projets data, voire des portefeuilles de projets data, ainsi qu'une aide à la Médiation Homme-Données. Le Databook dans les projets data en entreprise serait alors une proposition concrète de dispositif qui continuerait à documenter l'ensemble des choix humains réalisés tout le long de la conception du l'algorithme, et contribuerait à la transparence de celui-ci dans le cadre de l'usage qu'il supporte.

3 Médiation Homme-Données

Contrairement à un certain discours mystificateur, le terrain démontre qu'il existe bel et bien une interaction forte entre les hommes, aux compétences variées, et entre l'homme en général et la donnée au cours de la construction d'un algorithme. Bien que cette implication d'acteurs variés dans les projets data ne soit pas une nouveauté en soi, les interactions ne sont traitées qu'en surface à ce jour, et les résultats de ces travaux de recherche-action ouvrent des voies plus opérationnelles à la mise en place des interactions entre les pôles métier et data. Cette interaction est présentée comme une dialectique à dynamique double : une capitalisation progressive de connaissances qui permet la montée en maturité du dispositif complet et améliore sa performance, et des interactions d'alignement au service d'une convergence vers des résultats de qualité supérieure pour alimenter des usages opérationnels. La distinction de ces interactions est clé dans la mesure où elles ne partagent pas les mêmes finalités et ne mobilisent pas les ressources et méthodes équivalentes.

Ces interactions sont généralisées sous forme de Médiation Homme-Données à **4 facilitateurs** : capital de connaissances, algorithmes, représentations sociales et gestion de projet. Cela donne un cadre à la fois opérationnel pour le choix des méthodes et des ressources, et théorique pour la poursuite de la précision des dimensions clés du dispositif de médiation. En tant que cadre opérationnel, la Médiation Homme-Données s'adresse au management des équipes data pour l'optimisation de leur dispositif portefeuille projets data à court et long terme, ainsi qu'aux consultants et chefs de projets qui garantissent la médiation entre acteurs impliqués sur ces projets. En tant que cadre théorique, la Médiation Homme-Données donne la main aux spécialistes des Sciences de l'Information et de la Communication pour la poursuite du développement des modalités de médiation dans les projets Data Science, ce qui rend l'approche interdisciplinaire sur ce dispositif hétérogène en termes de fondements historiques et de compétences.

Le modèle proposé et la Médiation Homme-Données semblent par ailleurs constituer un socle pour une meilleure adaptation des méthodes de gestion de projet à des contextes de maturité et d'incertitudes différents. En effet, le fait de situer la mesure du niveau d'incertitudes en amont du choix du dispositif, puis de les faire évoluer progressivement, permet de mettre en place des méthodes de gestion de projet les plus performantes dans le contexte évolutif du projet Data.

Ce point est particulièrement clé dans un contexte de développement croissant des méthodes agiles, qui constituent des symptômes d'un manque de maturité des dispositifs en place : bien que ces méthodes semblent parfaitement appropriées en cas d'incertitudes élevées, en particulier sur l'usage final des résultats, elles peuvent être limitantes dans les projets complexes mobilisant des acteurs expérimentés. En effet, elles priorisent la co-construction progressive à la capacité d'anticipation des risques sur un plus long terme, qui sera de plus en plus appropriée au fur et à mesure de la montée en compétences des acteurs impliqués dans ces projets. Le choix de la méthode de gestion de projet devient ainsi non pas un préalable à la montée en maturité, mais bien un outil dépendant du niveau de maturité initial, ce qui ne peut la rendre que plus efficiente. Par ailleurs, l'inscription du modèle dans la génération d'usages opérationnels dotés d'un niveau de bénéfices potentiels, la prise en compte de la capitalisation et la mise en évidence des synergies potentielles au sein d'un portefeuille de projets permettent de générer les gains de productivité et améliorer la pertinence économique du dispositif. Cette posture s'adresse principalement aux gestionnaires de projets, et implique leur propre montée en compétences sur la connaissance des avantages et inconvénients des méthodes de gestion projet afin de pouvoir guider le choix de celles-ci.

Mais Médiation Homme-Données a potentiellement une portée plus large que ses aspects fonctionnels. A l'heure où de nouveaux modèles d'entreprise voient le jour pour prendre des niches de marché de plus en plus conséquentes, et que ces acteurs possèdent des données uniques, sont « data-driven » par naissance, et détiennent des fonds de plus en plus conséquents, la compréhension de la valeur de ses propres données pour une entreprise est clé. En effet, les données, en tant qu'informations uniques procurent un avantage concurrentiel, alors que les informations non uniques ne procurent pas cet avantage, bien qu'elles permettent d'être plus réactif et performant (avantage polémique). La médiation entre les praticiens historiques et leurs propres données est alors clé pour tirer les meilleurs avantages. De plus, l'utilisation d'algorithmes pour construire des informations uniques à partir de données non uniques, mais intelligemment combinées, semble ouvrir des voies intéressantes pour les organisations. Enfin, la médiation permet de concevoir ces nouveaux acteurs de l'écosystème non pas comme des concurrents, mais comme des partenaires potentiels qui peuvent intégrer le dispositif et la dialectique qui fait naître ces avantages compétitifs. En absence de preuves sur le caractère disruptif de ces nouveaux modèles ayant des conséquences catastrophiques sur les métiers historiques, l'appropriation des usages intégrant l'Intelligence Artificielle semble constituer une opportunité non seulement défensive, mais conciliatrice par anticipation.

La Médiation Homme-Données apporte par ailleurs une visibilité nouvelle sur les **compétences** mobilisées sur ces projets, parfois partagées entre leurs activités de production analytique et la médiation. Ces travaux de recherche peuvent donc intéresser plus indirectement les ressources humaines grâce à une meilleure compréhension des compétences mobilisables dans ce type de dispositif, en particulier lors du passage à échelle pour une gestion de portefeuille de projets data. Au-delà de l'utilité d'une vision globale qui permet de constituer des ensembles de compétences complets pour un dispositif projet donnée, d'ores et déjà développé dans un certain nombre de travaux de recherche, le mécanisme de montée en maturité apporte une dimension peu couverte. Or, ce sujet semble primordial, à la lumière de la découverte du terrain, dans la mesure où il conditionne non seulement la mise en place de formations et de dynamiques de transferts de compétences, mais aussi les carrières typiques des acteurs contributeurs, et notamment des Data Scientists au sens large. La mise en parallèle des activités de production et l'existence des instances de médiation transversales pointe en effet plusieurs possibilités de développement de compétences : d'une part, il s'agit d'une voie de spécialisation sur une activité donnée (par exemple, la montée de l'expertise en ingénierie data sur la préparation des données), d'autre part d'une voie de polarisation (métier ou data), ou encore la capacité croissante à animer la médiation transversale (profil plus généraliste avec un éventail large de compétences sur toutes les activités, et des compétences plus pointues sur les méthodes de médiation, et notamment sur ses 4 facilitateurs).

A ce stade, les compétences en Data Science sur le marché ne sont pas stabilisées, cependant, il est probable que deux types de profils divergent rapidement : les experts capables d'anticiper les risques sur leur domaine de compétences et de mettre leur expérience au profit d'une meilleure planification et définition des méthodes sur les projets, et des profils plus techniques qui seraient capables de produire des résultats selon une feuille de route définie par les premiers, ainsi que de suivre de façon récurrente les solutions mises en exploitation. Au-delà de la gestion des carrières en Data Science, la montée en maturité des équipes métier sur ces sujets leur permettra de mieux comprendre les nouvelles opportunités d'usages offertes et ainsi de formuler des besoins, voire les intégrer dans leur pratique quotidienne. Rien n'indique à ce stade une transformation radicale des métiers historiques, mais il semble probable qu'un écart puisse se creuser entre les experts métier au fait des pratiques en Data Science, et les autres, ce qui peut pénaliser les derniers. Au-delà de cette pénalisation, l'implication des métiers dans la phase de conception des modèles algorithmiques garantit le maintien de la logique causale des décisions assistées par la Data Science, ainsi que celle des schémas de responsabilités. A ce

jour, certains métiers historiques de la donnée semblent particulièrement touchés par ce phénomène, comme des contrôleurs de gestion, des biostatisticiens, des consultants ou encore des actuaires, à qui la Data Science peut apporter des outils complémentaires à leurs méthodes analytiques classiques. La mise en évidence de la dynamique de montée en compétences par l'apprentissage au cours des projets semble ainsi critique pour la capitalisation des connaissances par les équipes, leur productivité, et plus globalement la valorisation du capital humain.

Ainsi, la Médiation Homme-Données constitue à l'issue de ces travaux un apport majeur, à la fois du point de vue opérationnel et managérial que théorique pour les Sciences de l'Information et de la Communication. Son insertion dans le modèle de dispositif projet data ouvre la voie à une appropriation du sujet par la discipline, encore en retrait des projets data, pour exercer au mieux son rôle de médiateur dans la capture optimale de la valeur potentielle et du sens en s'appuyant sur les 4 facilitateurs transdisciplinaires. Elle contribue enfin au renouvellement des défis des SIC sur le terrain de l'objectivation des savoirs à travers la co-construction dialectique des algorithmes et des usages qu'ils appuient.

4 Pistes de recherche

Les résultats de ces travaux peuvent intéresser plusieurs types d'acteurs en entreprise (voir Figure 51), ce qui confirme l'intérêt d'une approche interdisciplinaire du sujet. L'ajustement du modèle CRISP_DM s'adresse à tous les acteurs pouvant intervenir sur les projets data, le modèle global peut permettre de mieux comprendre le positionnement de ces projets dans les entreprises, et enfin, la richesse des pistes d'exploitation de la Médiation Homme-Données peut permettre aux Sciences de l'Information et de la Communication de s'approprier durablement le champ d'investigation ouvert, explorer et développer ces premières pistes du point de vue théorique et opérationnel. Elles semblent par ailleurs pouvoir mettre en lumière plus amplement les facteurs favorisant la compréhension mutuelle des enjeux liés au dispositif projet data par les acteurs en jeu.

		Direction	Experts métier	Utilisateurs	Knowledge Management	Data Gouv.	Ressources humaines	Business Intelligence	Directions informatiques	Project Management	Data Scientists
Ajustement CRISP_DM	Prise en compte amont de l'usage des résultats, de la qualification des données et du format de restitution des résultats	+	++	++	+	++		++	++	++	++
Modèle global proposé	Orientation du dispositif sur l'usage par la réduction d'incertitudes	++	++	++	+	+	+	+	+	++	++
	Jalonnement par des réévaluations au cours des instances de médiation	++	++	++	+	+	+	+	+	++	++
	Documentation et de valorisation de la qualité des données (Databook)	+	++		++	++		++	++	++	++
Médiation Homme Données	Capitalisation de connaissances	+	++	+	++	++	++	+	+	++	++
	Algorithmes		++	++	++	++		+	+	++	++
	Gestion de projet	+	+	+	+	+	+	+	+	++	++
	Représentations sociales	+	++	++	++	+		++		++	++

++ Apport perçu comme significatif

+ Apport perçu restant à confirmer

Figure 51 – Perception des opportunités d'application des résultats par les acteurs en entreprise

Si l'état de l'art dresse un parallèle significatif entre les projets data et la gestion documentaire, plus particulièrement la pratique de l'Intelligence Economique et le Knowledge Management en entreprise, il s'agit ici d'un apport double pour ces champs. D'une part, ces travaux pointent, une fois de plus (Dudezert & al., 2015; Chastenet de Géry, 2018), la possibilité d'armer les professionnels de l'information avec de nouveaux outils ou de devenir eux-mêmes demandeurs de projets data, au même titre qu'une équipe de contrôle de gestion ou de marketing. D'autre part, la mise en évidence des dynamiques d'arbitrage humain et de construction de sens confirment les messages forts de la profession autour de l'inadéquation entre un discours porteur de mythes sur le remplacement des sachant et des décisionnaires par des machines (Moinet & Alloing, 2016). Or, ces professionnels de l'information, étonnamment absents de tous les cas traités dans ces travaux, semblent les plus conceptuellement armés pour traduire ces éléments, et notamment la Médiation Homme-Données, en tant que contributeurs clés dans les projets data en général en entreprise. Par ailleurs, si le phénomène Big Data plonge ces racines historiques dans des intérêts de contrôle des populations par les Etats pour trouver une déclinaison en entreprise sous forme de projets data, l'Intelligence Economique fait de même à travers sa proximité avec le Renseignement : l'élargissement des pistes de recherche, depuis la sphère de veille et d'étude concurrentielle en entreprise aux enjeux plus sociétaux, reste à faire à ce stade des travaux.

A l'heure de la rédaction de ces conclusions, cette perception d'utilité se base sur 4 années de pratique de projets data chez Quinten, et les retours d'expérience des professionnels et des chercheurs en dehors de la société confirment cet intérêt, d'autant plus que les directives de recherche pour l'année 2019 comprennent en priorité l'Intelligence Artificielle et les Sciences Humaines et Sociales[47]. A court terme, trois voies semblent possibles pour avancer de façon applicative à partir de ces résultats :

1. Evaluer la pertinence de l'ajustement du modèle de référence et des 3 dimensions d'enrichissement, à savoir les indicateurs de valeur, la qualité des données et la Médiation Homme-Données, notamment grâce à l'approche quantitative permise par la grille de collecte d'observations établie et annexée.

[47] *Plan d'action 2019*, ANR, 26 juillet 2018, http://www.agence-nationale-recherche.fr/fileadmin/documents/2018/Plan-d-action-ANR-2019.pdf

2. Confronter la structure du prototype de Databook aux professionnels de la gestion de la qualité des données, puis à des bénéficiaires potentiels plus larges. Cette piste peut donner lieu à un développement d'outil dédié, et à terme d'un standard de documentation pour le traitement de l'information sur les projets data. Cette documentation du travail analytique est vue comme clé à l'heure où la règlementation européenne impose des règles de transparence des algorithmes, visant l'élimination des biais, d'erreurs et de choix de construction non conformes.

3. Etablir une cartographie des risques de référence applicable à un projet Data Science, à la fois avec des acteurs métier, les Data Scientists, et potentiellement des régulateurs.

Ces trois pistes semblent s'inscrire parfaitement dans les travaux plus récents sur la transparence et le caractère « auditable » des systèmes algorithmiques, initiés par l'Etat (notamment à travers les travaux de l'INRIA et son projet TransAlgo[48]) ou encore des acteurs privés comme les cabinets d'audit et de conseil[49], et ce plus particulièrement au sujet des voitures autonomes qui soulèvent de nombreuses question éthiques et morales dans des sociétés aux valeurs hétérogènes et évolutives.

Au-delà de ces pistes immédiates, ces travaux de recherche mettent en lumière la nécessité de se pencher sur deux sujets clés (évoqués dans le chapitre 2.2.3.5 de la Troisième partie, page 271, et illustrées par la Figure 41) : l'acculturation au phénomène, permettant de préciser les motivations de ces projets en entreprise, et la déclinaison sectorielle, fonctionnelle et plus transversale des usages terrain, afin de limiter le caractère exploratoire des usages. L'Intelligence Economique et le Knowledge Management semblent constituer des disciplines appropriées pour appréhender ces sujets, que ce soit en entreprise ou plus largement en termes de recherche, afin de faire un contrepoids aux discours commerciaux de l'écosystème Big Data, toujours en pleine expansion.

En complément de l'investigation sur l'intérêt opérationnel du modèle et des outils dédiés pour les différents groupes de bénéficiaires, une autre approche intéressante pour l'enrichissement de ces travaux serait le sens épistémologique du « projet » (Bachelard, 1934b) à travers l'idée d'objectivation : « l'objectivité n'est que le produit d'une objectivation correcte ». En effet, les

[48] *TransAlgo : évaluer la responsabilité et la transparence des systèmes algorithmiques*, INRIA, 4 avril 2018, https://www.inria.fr/actualite/actualites-inria/transalgo

[49] *L'Intelligence artificielle : quelle place pour la morale ?,* PWC, consulté en avril 2019, https://www.pwc.fr/fr/decryptages/transformation/ia-quelle-place-pour-la-morale.html

itérations de construction des résultats au cours des projets data contredisent l'un des arguments du mythe Big Data, en montrant que ce ne sont pas les données qui conduisent à des connaissances objectives, mais une convention, une médiation impliquant des acteurs humains. Ces itérations ne remettent pas en cause la possibilité d'une objectivation, correcte ou non, en amont d'une « boite noire ». Cet enjeu semble majeur, puisqu'il pointe la légitimité, voire la responsabilité de l'ensemble des acteurs impliqués dans ce processus d'objectivation, source de capitalisation de connaissances et justification de la prise de décision permise à travers l'usage. Cette légitimité s'appuie alors sur les savoirs subjectifs métier, statistiques, techniques ou autres nécessaires pour le fonctionnement du dispositif. Or, la mobilisation de ressources externes au métier, et, dans certains cas, la diffusion de l'usage au-delà de ces concepteurs, comporte un risque de provoquer un déséquilibre entre les « sachants » ayant participé à la conception de l'usage et des utilisateurs finaux. Ce déséquilibre peut être caché par un l'algorithme qui semble s'insinuer dans les rapports de « Pouvoir-Savoir » (Foucault, 2014) comme un nouvel acteur non humain soi-disant neutre, et réside dans l'accentuation d'une nouvelle forme de « l'économie de la pensée » (Mach, 1921) à travers la simplification du savoir. Dans le cadre de la croissance de la quantité d'informations pour la prise de décision en entreprises, cette mécanique semble proche du développement des techniques de transmission efficiente du savoir à l'ère de l'émergence de la division du travail, éveille d'ores et déjà des craintes de la part des dirigeants et des salariés[50] et doit être accompagnée de façon appropriée dans les années à venir.

[50] IA et capital humain : quels défis pour les entreprises ?, étude du Comptoir MM de Malakoff Médéric en partenariat avec le BCG, 12 septembre 2018, http://www.lecomptoirmm.com/ia-numerique/intelligence-artificielle-capital-humain-defis-entreprises/

« On m'accuse de présenter la société ou la technique comme des machines qui fonctionnent sans l'homme, mais c'est bien cela que m'a appris l'échec de toutes les résistances, de toutes les révolutions, de toutes les volontés, de toutes les analyses, de toutes les proclamations, de tous les programmes. »

***Jacques Ellul**, historien et sociologue.*

Les résultats ce des travaux sont ainsi riches d'une mise en perspective historique et interdisciplinaire, de rapports approfondis sur les études de cas en France, et apportent des réponses inédites, et parfois inattendues, aux questions de recherche. En effet, si les projets ne sont pas aussi disruptifs pour l'entreprise que ce qu'elle en espère parfois, ils impactent bien de façon positive son fonctionnement, mais cet impact est souvent mal identifié, sous-valorisé et perfectible. Ces réponses sont porteuses de propositions opérationnelles qui vont au-delà de simples constats, et s'ouvrent sur une évolution plus large des champs disciplinaires principaux que ces travaux explorent.

Les propositions opérationnelles s'adressent aux acteurs impliqués dans les projets data en entreprise. Les études des cas mettent en lumière les limites du modèle de processus de référence, le CRISP_DM, autocentré, insuffisamment orienté sur l'usage métier et négligeant la valorisation des usages indirectement produits par le dispositif, dont la capitalisation de connaissances et la valorisation de la qualité des données. Le premier résultat de cette étude est une proposition d'ajustement du modèle de référence, opérationnel en Data Mining, pour un usage en Data Science. Le deuxième est une proposition de modèle englobant le dispositif projet data dans une perspective plus large, encadré par l'évaluation des bénéfices attendus, des incertitudes et des ressources nécessaires à la levée de ces incertitudes, et par la mise en exploitation d'usages directs et indirects. Le modèle global, polarisé entre le métier et la donnée, est jalonné par des instances de médiation qui guident la convergence vers un résultat cumulé optimal. Il comprend un outil de capitalisation et de valorisation de la qualité des données, sous forme préliminaire de Databook. Enfin, la Médiation Homme-Données est inscrite au cœur de ce modèle comme un facteur d'optimisation qualifié et opérationnel, grâce à l'identification de 4 dimensions clés : la capitalisation de connaissances, les algorithmes, les représentations sociales et la gestion de projet.

Ces résultats complètent les travaux sur les processus de projet data qu'ils rendent plus efficients, cumulables au sein d'un portefeuille optimisé de projets data, et comparables à d'autres projets plus classiques tout en préservant les spécificités analytiques liées à la découverte de connaissances et d'usages. Cette intégration de spécificités est un terrain nouveau pour les Sciences de Gestion, et peut trouver des résonnances dans les champs aussi variés que les ressources humaines, la gestion de projet, l'optimisation de la performance des entreprises, l'innovation ou encore la stratégie des entreprises face à cet écosystème naissant. Deux voies de questionnement sont à distinguer et à approfondir dans ces différents champs : la possibilité

d'utiliser l'Intelligence Artificielle en tant que dispositif asservi aux finalités de chaque champ, et plus particulièrement les usages basés sur la Data Science liés aux fonctions et aux secteurs que ces champs supportent (par exemple, les usages en RH, l'optimisation des indicateurs de performances considérés comme métrique de phénomène d'intérêt…), et la contribution de chaque champ aux expertises transversales mobilisées pour la mise en place des usages (par exemple, la mobilisation RH au service des dispositifs Homme Data en entreprise, ou l'utilisation des indicateurs de performance pour estimer les bénéfices des usages et permettre les arbitrages au cours d'un projet data).

Mais au-delà de cette portée opérationnelle, chère à la recherche action, et de son appropriation par les Sciences de Gestion, le modèle Brizo_DS, et plus particulièrement le concept de Médiation Homme-Données qu'il porte, ouvre un terrain de jonction interdisciplinaire intéressant et des pistes de recherche complémentaires pour les Sciences de l'Information et de la Communication. Il pointe en effet la nécessité de renouveler les modalités de la médiation pour favoriser au maximum la génération de sens grâce à une proximité entre le dispositif projet et les besoins du terrain et ses pratiques, à un interfaçage entre les hommes, porteurs de représentations sociales évolutives, et les données au sens large incluant les algorithmes, et au développement du rôle des agents médiateurs. Les propositions opérationnelles ouvrent le champ de la médiation documentaire à une médiation data, comportant ses spécificités de gestion des métadonnées liées à l'introduction des algorithmes, objets co-construits et porteurs de signification dont il est indispensable de garder la trace afin d'éviter des dérives d'usages, d'ores et déjà craintes. Si les réactions dans cette discipline sont souvent critiques, et ce à bon escient face à un discours mystificateur et éloignant *a priori* l'homme de la création de connaissances à l'ère du Big Data, la réalité observée tend plutôt à infirmer le mythe et à témoigner le besoin d'assurer la médiation avec les métiers et de maintenir une transversalité des compétences.

Bibliographie

Les références bibliographiques ne contiennent les liens URL que pour les ouvrages électroniques, les articles de sites web et les thèses consultables en ligne. Les références les plus centrales des travaux de cette thèse sont en gras.

ABBOTT A. (1992). What Do Cases Do? Some Notes on Activity in Sociological Analysis. In *What is a Case? Exploring the Foundations of Social Inquiry*. New York, Cambridge University Press.

ABDEL A, & EL SHEIKH R. (2011). *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications*. USA, IGI Global.

ACKOFF R L. (1967). Management Misinformation Systems. *Management Science*, Vol. 14, No. 4, p. B-147, décembre.

AÏM R. (2011). *Filippo Brunelleschi : Le dôme de Florence, paradigme du projet*. Paris, Hermann.

AKRICH M, CALLON M, & LATOUR B. (2006). *Sociologie de la traduction : textes fondateurs*. Paris, Presses des Mines.

ALNOUKARI M, ALZOABI Z, & HANNA S. (2008). Applying adaptive software development (ASD) agile modeling on predictive data mining applications: ASD-DM methodology. In *2008 International Symposium on Information Technology* (Vol. 2, p. 1-6). Kuala Lumpur, Malaysia, IEEE.

AMATO É A. (2015). Enjeux et opportunités de la datavisualisation : interagir avec les données. *I2D – Information, données & documents*, Vol. 52, No. 2, p. 34-35, juillet.

ANDERSON C. (2008, juin 23). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, juin. Consulté à l'adresse http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

ARRUABARRENA B. (2015). Datavisualisation : des données à la connaissance. *I2D – Information, données & documents*, Vol. 52, No. 2, p. 28-29, juillet.

ARRUABARRENA B, KEMBELLEC G, & CHARTRON G. (2019, mars). « Data littératie & SHS : développer des compétences pour l'analyse des données ». Présenté à CODATA - Data Value Chain, Val d'Europe.

ATAMER T, & CALORI R. (2003). *Diagnostic et décisions stratégiques*. Paris, Dunod.

AZEVEDO A, & SANTOS M F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. In *Proceedings of the IADIS European Conference on Data Mining* (p. 182-185). Amsterdam, The Netherlands.

BACHELARD G. (1934a). *La formation de l'esprit scientifique. Contribution à une psychanalyse de la connaissance objective*. Paris, Librairie philosophique J. Vrin.

BACHELARD G. (1934b). *Le nouvel esprit scientifique* (10e édition 1968). Paris, Presses Universitaires de France.

BARNES T J. (2013). Big data, little history. *Dialogues in Human Geography*, Vol. 3, No. 3, p. 297-302, novembre.

BARTRAM P. (2013, mai 19). The value of data. Consulté 8 août 2016, à l'adresse http://www.fm-magazine.com/feature/depth/value-data#

BATTISTI M. (2017). La qualité, une obligation impérieuse. *I2D – Information, données & documents*, Vol. 53, No. 4, p. 1-1, janvier.

BATTISTI M, CHAPOY É, MERRIEN D, HAETTIGER M, COTTART O J, BOURRION D, & BALIGAND M-P. (2010). Des pratiques professionnelles renouvelées. *Documentaliste-Sciences de l'Information*, Vol. 47, No. 2, p. 44-55, juin.

BÉRA M. (2011). Les nouveaux énoncés de la modélisation prédictive à très grand nombre de variables. *Risques*, No. 45, mars.

BÉRA M. (2014, mai). « Comprendre le datamining et les méthodologies scientifiques associées ». Séminaire présenté à Big Data, fouilles de données dans les domaines scientifiques, Conservatoire National des Arts et Métiers, Paris.

BÉRA M, & MÉCHOULAN E. (1999). *La machine Internet*. Odile Jacob.

BERNARD A, & TICHKIEWITCH S. (2008). *Methods and Tools for Effective Knowledge Life-Cycle-Management*. Springer Science & Business Media.

BERTACCHINI Y. (2009). *Petit Guide à l'usage de l'Apprenti-Chercheur en Sciences Humaines & Sociales ESSAI Epistémologie & Méthodologie de Recherche en Sciences de l'Information & de la Communication*. Toulon, Presses Technologiques.

BERTI L. (1999). Quality and Recommendation of Multi-Source Data for Assisting Technological Intelligence Applications. In *Database and Expert Systems Applications* (p. 282-291). Springer, Berlin, Heidelberg.

BERTI-EQUILLE L. (2012). *La qualité et la gouvernance des données : Au service de la performance des entreprises*. Paris; Cachan, Hermes Science Publications.

BERTINO E, BERNSTEIN P, AGRAWAL D, DAVIDSON S, DAYAL U, FRANKLIN M, … WIDOM J. (2011). Challenges and Opportunities with Big Data. Cyber Center Publications.

BESSET C. (2011). « L'usage des médias sociaux par les musées : potentiel et réalisations. » (Mémoire de master). HEC, Paris.

BIERNAT E, & LUTZ M. (2015). *Data Science : fondamentaux et études de cas*. Eyrolles.

BLADT J, & FILBIN B. (2013, mars 27). A Data Scientist's Real Job: Storytelling. *Harvard Business Review*, mars.

BOHLE S. (2013, juin 12). What is E-science and How Should it be Managed? *SciLogs - Scientific and Medical Libraries*, juin.

BOLLIER D. (2010). « The Promise and Peril of Big Data ». Washington, DC: The Aspen Institute.

BOMSEL O. (2007). *Gratuit ! : Du déploiement de l'économie numérique*. Paris, Folio.

BONENFANT M, MÉNARD M, MONDOUX A, & OUELLET M. (2014). Big Data and Governance. In *A Digital Janus: Looking Forward, Looking Back* (p. 185-195). Inter-Disciplinary Press.

BOUOUCHMA K. (2016, mai 2). Réussir son projet Big Data Science en 5 étapes. *CEBI – Conseil des Experts Business Intelligence*, mai.

BOURBONNAIS R. (2001). *Prévision des ventes*. Université de Paris-Dauphine.

BOURDIEU P, CHAMBOREDON J-C, & PASSERON J-C. (1968). *Le métier de sociologue*. Paris, De Gruyter Mouton.

BOURDONCLE F. (2014). Peut-on créer un écosystème français du Big Data ? *Le journal de l'école de Paris du management*, Vol. 108, No. 4, p. 8-15, juillet.

BOUSTANY J, BROUDOUX E, & CHARTRON G. (2014). *La médiation numérique : renouvellement et diversification des pratiques.* (Vol. 1). De Boeck and ADBS.

BOUTINET J-P. (2012). *Anthropologie du projet*. Paris, Presses Universitaires de France.

BOYD D, & CRAWFORD K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, Vol. 15, No. 5, p. 662-679.

BRAHA D (Éd.). (2001). *Data Mining for Design and Manufacturing: Methods and Applications*. Springer.

BRAHIM W. (2017). L'approche processus. *I2D – Information, données & documents*, Vol. 53, No. 4, p. 37-38, janvier.

BRANNEN J. (2005). Mixing Methods: The Entry of Qualitative and Quantitative Approaches into the Research Process. *International Journal of Social Research Methodology*, Vol. 8, No. 3, p. 173-184, juillet.

BROUDOUX É, & SCOPSI C. (2011). Introduction. *Études de communication*, No. 36, p. 9-22, juin.

BROWN B, CHUI M, & MANYIKA J. (2011). Are you ready for the era of 'big data'? *McKinsey Quarterly*, octobre.

BROWNLEE J. (2013, novembre 24). A Tour of Machine Learning Algorithms. Consulté à l'adresse https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

BRYNJOLFSSON E, HITT L M, & KIM H H. (2011). « Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? » (SSRN Scholarly Paper No. ID 1819486). Rochester, NY: Social Science Research Network.

BRYSON S, KENWRIGHT D, COX M, ELLSWORTH D, & HAIMES R. (1999). Visually Exploring Gigabyte Data Sets in Real Time. *Commun. ACM*, Vol. 42, No. 8, p. 82–90, août.

BULINGE F, & BOUTIN É. (2015). Le renseignement comme objet de recherche en SHS : le rôle central des SIC. *Communication Organisation*, Vol. n° 47, No. 1, p. 179-195, octobre.

BULINGE F, & MOINET N. (2013). L'intelligence économique : un concept, quatre courants. *Securite et strategie*, Vol. 12, No. 1, p. 56-64.

BULINGE F, & MOINET N. (2016). Introduction. *Hermes, La Revue*, Vol. 76, No. 3, p. 14-21, novembre.

CAIRE D, CAMICIOTTI L, HEITMANN S, LONIE S, RACCA C, RAMJI M, & XU Q. (2017). « Data Analytics and Digital Financial Services » (Paper) (p. 160). International Finance Corporation (IFC),.

CALISTE J P, & BOURRET C. (2013). Contribution à une analyse typologique des processus : De la conformité à l'agilité. In *QUALITA2013*. Compiègne, France.

CAMICIOTTI L, & RACCA C. (2015). *Creare valore con i Big Data. Gli strumenti, i processi, le applicazioni pratiche* (1 edizione). Milano, Edizioni LSWR.

CAPPELLETTI L, SAVALL H, & BUONO A F. (2018). *La Recherche-Intervention dans les Entreprises et les Organisations : de la conception à la publication*. Charlotte, NC, Information Age Publishing.

CARDON D. (2015). *A quoi rêvent les algorithmes : Nos vies à l'heure des big data*. Paris, Seuil. Consulté à l'adresse https://www.babelio.com/livres/Cardon-A-quoi-revent-les-algorithmes--Nos-vies-a-lheure/780549

CHAPMAN P. (1999). « The CRISP-DM User Guide ». Présenté à Brussels SIG Meeting, NCR Systems Engineering Copenhagen.

CHARLOT B. (2005). *Du rapport au savoir: éléments pour une théorie*. Anthropos.

CHARMILLOT M, & DAYER C. (2007). Démarche compréhensive et méthodes qualitatives: clarifications épistémologiques. *Recherches qualitatives*, Vol. 3, p. 126–139.

CHARREIRE S, & HUAULT I. (2001, septembre). Le constructivisme dans la pratique de recherche : une évaluation à partir de seize thèses de doctorat. *Finance Contrôle Stratégie*, Vol. 4, No. 3, p. 31-55, septembre.

CHARTRON G, & BROUDOUX E. (2015). *Big data - Open data, Quelles valeurs ? Quels enjeux ?* De Boeck Supérieur.

CHASTENET DE GÉRY G. (2018). *Le knowledge management : un levier de transformation à intégrer*. De Boeck Superieur.

CHAU P Y K, & TAM K Y. (1997). Factors Affecting the Adoption of Open Source Systems: An Exploratory Study. *MIS Quarterly*, Vol. 21, No. 1, p. 1-24.

COFFMAN K, & ODLYZKO A. (1998). The work of the encyclopedia in the age of electronic reproduction. *First Monday*, Vol. 3, No. 10, octobre.

COHEN D, & CRABTREE B. (2016). « Qualitative Research Guidelines Project | Critical or Subtle Realist Paradigm | Critical or Subtle Realist Paradigm » (Qualitative Research Guidelines Project). Robert Wood Johnson Foundation.

COLLERETTE P. (1997, septembre). L'étude de cas au service de la recherche. *Recherche en soins infirmiers*, No. 50, p. 81-88, septembre.

CONWAY D. (2010, septembre 30). The Data Science Venn Diagram. Consulté à l'adresse http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

COTTIN M. (2017). Prendre des décisions fondées sur des preuves. *I2D – Information, données & documents*, Vol. 53, No. 4, p. 41-42, janvier.

COTTIN M, & NESME M-F. (2017). La qualité : variations autour d'une notion essentielle, Quality: variations on an essential notion. *I2D – Information, données & documents*, Vol. 53, No. 4, p. 28-29, janvier.

COUZINET V, RÉGIMBEAU G, & COURBIÈRES C. (2001). *Sur le document : notion, travaux et propositions*. Paris: ADBS Editions.

COX M, & ELLSWORTH D. (1997a). Application-controlled Demand Paging for Out-of-core Visualization. In *Proceedings of the 8th Conference on Visualization '97* (p. 235–ff.). Phoenix, Arizona, USA, IEEE Computer Society Press.

COX M, & ELLSWORTH D. (1997b). Managing big data for scientific visualization. *ResearchGate*, janvier.

CURRY A, FLETT P, & HOLLINGSWORTH I. (2006). *Managing Information & Systems: The Business Perspective*. Routledge.

DACOS M, & MOUNIER P. (2010). IV. L'édition numérique. In *L'édition électronique* (p. 67-87). Paris, La Découverte.

DALBIN S. (2007). Thésaurus et informatique documentaires. *Documentaliste-Sciences de l'Information*, Vol. 44, No. 1, p. 42-55.

DAVENPORT T H, & PATIL D J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, Vol. 90, No. 5, p. 70–76, octobre.

DE BRUYNE P, HERMAN J, DE SCHOUTHEETE M, & LADRIÈRE J. (1974). *Dynamique de la recherche en sciences sociales: les pôles de la pratique méthodologique*. Paris, Presses Universitaires de France.

DE MARGERIE V. (2008). Organisation de la gouvernance et stratégie d'entreprise : état des lieux des 120 premières sociétés françaises cotées. *Management & Avenir*, No. 17, p. 66-82, mars.

DEAN J, & GHEMAWAT S. (2004). MapReduce: Simplified Data Processing on Large Clusters. In *Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6* (p. 10). Berkeley, CA, USA, USENIX Association.

DEAN J, & GHEMAWAT S. (2010, janvier 19). United States Patent: 7650331 - System and method for efficient large-scale data processing.

DECLERCK R P, DEBOURSE J P, & NAVARRE C. (1983). *Méthode de direction générale: le management stratégique*. Hommes et Techniques.

DELECROIX B. (2005). « La mesure de la valeur de l'information en Intelligence Economique » (Thèse). Université de Marne-La-Vallée. Consulté à l'adresse Base de Connaissance AEGE. (http://www.bdc.aege.fr)

DELEUZE G, & GUATTARI F. (1972). *L'anti-OEdipe : Capitalisme et schizophrénie* (1re éd.). Paris, Les Editions de minuit.

DENNING P J. (1990). The Science of Computing: Saving All the Bits. *American Scientist*, Vol. 78, No. 5, p. 402-405.

DENSCOMBE M. (2014). *The Good Research Guide* (5e éd.). Berkshire, England, Open University Press.

DENZIN N K. (2012). Triangulation 2.0. *Journal of Mixed Methods Research*, Vol. 6, No. 2, février. doi:https://doi.org/10.1177/1558689812437186

DENZIN N K, & LINCOLN Y S. (1994). *Handbook of qualitative research*. Thousand Oaks, Sage Publications.

DESROCHE H. (1981). La recherche coopérative comme recherche–action (p. 9–48). Présenté à Recherche–Action, Chicoutimi, UQAC.

DESROSIÈRES A. (2008). *Pour une sociologie historique de la quantification : L'Argument Statistique I*. Paris, Presses des Mines.

DESROSIÈRES A, & KOTT S. (2005). Quantifier. *Genèses*, Vol. no 58, No. 1, p. 2-3.

DIAKOPOULOS N, & KOLISKA M. (2017). Algorithmic Transparency in the News Media. *Digital Journalism*, Vol. 5, No. 7, p. 809-828, août.

DIEBOLD F X. (2012). « On the Origin(s) and Development of the Term "Big Data" » (SSRN Scholarly Paper No. ID 2152421). Rochester, NY: Social Science Research Network.

DIENES I. (1994). *National Accounting of Information*. Budapest, SNIA.

DOMINGOS P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Penguin UK.

DOUCET C. (2010). *La qualité*. Paris, Presses Universitaires de France.

DUDEZERT A., FAYARD P. & OIRY E. (2015). Astérix and the Knowledge Management 2.0: An exploration of KMS 2.0 appropriation by the myth of Gallic Village, *Systèmes d'Information et Management* Vol. 20, No. 1, p.31-59, juin.

DUDEZERT A. (2018). « Mener une recherche ancrée sur le terrain : la dynamique du collier de perles », In *Les méthodes de recherche du DBA (*pp. 415-429). Caen, EMS.

DUDEZERT A. (2018). *La transformation digitale des entreprises*. Paris, La Découverte.

DUFF A S. (2000). *Information Society Studies*. Routledge Research in Information Technology and Society.

DULL T. (2015, septembre). Data Lake vs Data Warehouse: Key Differences. Consulté à l'adresse http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html

DUMEZ H. (2013). Qu'est-ce qu'un cas, et que peut-on attendre d'une étude de cas ? *Le Libellio d'AEGIS*, Vol. 9, No. 2, p. 13-26.

EDMUNDS A, & MORRIS A. (2000). The problem of information overload in business organisations: a review of the literature. *International Journal of Information Management*, Vol. 20, No. 1, p. 17-28, février.

EDVINSSON L, & MALONE M. (1999). *Le capital immatériel de l'entreprise*. Maxima.

EISENHARDT K M. (1989). Building Theories from Case Study Research. *The Academy of Management Review*, Vol. 14, No. 4, p. 532-550, octobre.

ELLIS D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation*, Vol. 45, No. 3, p. 171-212, mars.

ERMINE J-L. (2003). *La gestion des connaissances*. Hermes Lavoisier.

ESPAIGNET S, RAMATOULAYE F, & LAURENCEAU A. (2003). « Pertinence de l'idée de désintermédiation documentaire » (Mémoire de recherche DCB). École nationale supérieure des sciences de l'information et des bibliothèques.

EZRATTY O. (2017, octobre 19). Les usages de l'intelligence artificielle. Consulté à l'adresse https://www.oezratty.net/wordpress/2017/usages-intelligence-artificielle-ebook/

FALQUE-PIERROTIN I, MAHJOUBI M, & VILLANI C. (2017). « Comment permettre à l'Homme de garder la main ? Rapport sur les enjeux éthiques des algorithmes et de l'intelligence artificielle ». CNIL. Consulté à l'adresse https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf

FANET H. (2008). « Nano-electronique. Impact sur les architectures et le logiciel ». CEA.

FAYOL H. (1916). *General principles of management*. Consulté à l'adresse http://docentiold.unimc.it/docenti/ernesto-tavoletti/2011/economia-e-gestione-delle-imprese-2011/lettura-2/at_download/Fayol.pdf

FAYYAD U, PIATETSKY-SHAPIRO G, & SMYTH P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, Vol. 39, No. 11, p. 27–34.

FERNANDEZ A. (2000). *Les nouveaux tableaux de bord des décideurs: le projet décisionnel dans sa totalité*. Editions d'Organisation.

FERNANDEZ A. (2013). *Les nouveaux tableaux de bord des managers : Le projet Business Intelligence clés en main* (6e édition). Eyrolles.

FIOL M, JORDAN H, & SULLÀ E. (2004). *Renforcer la cohérence d'une équipe: diriger et déléguer à la fois*. Paris, Dunod.

FISHER R A. (1937). *The Design of Experiments*. Edinburgh, Oliver and Boyd.

FLECK L. (1935). *Genesis and Development of a Scientific Fact*. University of Chicago Press.

FOUCAULT M. (2014). *Surveiller et punir. Naissance de la prison*. Editions Gallimard.

FOX C, LEVITIN A, & REDMAN T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*, Vol. 30, No. 1, p. 9‑19, janvier.

FRICKÉ M. (2014). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, juin. doi:10.1002/asi.23212

GALAUP X (Éd.). (2017). *Développer la médiation documentaire numérique*. Villeurbanne, Presses de l'enssib.

GALTON F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, Vol. 15, p. 246‑263.

GARDIÈS C, & FABRE I. (2015). Médiation des savoirs : de la diffusion d'informations numériques à la construction de connaissances, le cas d'une « classe inversée ». *Distances et médiations des savoirs*, Vol. 3, No. 12, décembre. doi:10.4000/dms.1240

GAREL G. (2003). Pour une histoire de la gestion de projet. *Gérer et comprendre*, Vol. 74, No. 1, p. 77–89.

GAREL G, GIARD V, & MIDLER C. (2003). Management de projet et gestion des ressources humaines. In *L'encyclopédie de la gestion des ressouces humaines* (p. 818‑843). Vuibert.

GEORGE P M. (2002). *Le Management Cockpit. Des tableaux de bord qui vont à l'essentiel.* Editions d'Organisation.

GILLIES J, & CAILLIAU R. (2000). *How the Web was Born: The Story of the World Wide Web*. Oxford, Oxford University Press.

GIRIN J. (1989). L'opportunisme méthodique dans les recherches sur la gestion des organisations. Présenté à Journée d'étude la recherche-action en action et en question, École Centrale de Paris, Collège de systémique, AFCET.

GIROD-SEVILLE M, & PERRET V. (2002). Les critères de validité en sciences des organisations : les apports du pragmatisme. In *Questions de méthodes en sciences de gestion* (EMS, p. 315‑333). N. Mourgues & alii (Dir).

GLASER B, & STRAUSS A. (1967). The discovery of grounded theory. *Weidenfield & Nicolson, London*, p. 1–19.

GOMEZ P Y. (1996). *Le gouvernement de l'entreprise* (InterEditions). Paris.

GRAHAM M. (2012, septembre 3). Big data and the end of theory? *The Guardian*.

GRANVILLE V. (2014, juillet 24). 16 analytic disciplines compared to data science. Consulté à l'adresse http://www.datasciencecentral.com/profiles/blogs/17-analytic-disciplines-compared

GRAY J. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Tony Hey, Stewart Tansley, and Kristin Tolle). Redmond, Washington, Microsoft Research. Consulté à l'adresse http://research.microsoft.com/en-us/collaboration/fourthparadigm/

GROULX L-H. (1997). Querelles autour des méthodes. *Socio-anthropologie*, No. 2, octobre. doi:10.4000/socio-anthropologie.30

GSCHWIND T, & HAUSWIRTH M. (1999). NewsCache: A High Performance Cache Implementation for Usenet News. In *Proceedings of the Annual Conference on USENIX Annual Technical Conference* (p. 16–16). Berkeley, CA, USA, USENIX Association.

GUERRA F. (2007). *Pilotage stratégique de l'entreprise. Le rôle du tableau de bord prospectif*. De Boeck.

GUO Y. (2013, mars 15). Big Data, Big Science, Big Collaboration: Delivering Connected R&D for Better Value. Consulté à l'adresse http://www.scientificcomputing.com/article/2013/03/big-data-big-science-big-collaboration-delivering-connected-rd-better-value

HAMMERSLEY M. (1992). *What's Wrong with Ethnography? : Methodological Explorations*. Psychology Press.

HARRATHI R, & CALABRETTO S. (2006). Un modèle de qualité de l'information (p. 299‑304). Présenté à EGC'2006, Lille.

HERREID C F. (1997). What is a Case? - Bringing to Science Education the Established Teaching Tool of Law and Medicine, Vol. 27, No. 2, p. 3.

HEY T, TANSLEY S, & TOLLE K (Éd.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery* (1 edition). Redmond, Washington, Microsoft Research.

HEY T, & TREFETHEN A E. (2005). Cyberinfrastructure for e-Science. *Science*, Vol. 308, No. 5723, p. 817‑821, mai.

HILBERT M. (2012). Info Capacity| Introduction—How to Measure "How Much Information"? *International Journal of Communication*, Vol. 6, No. 0, p. 14, avril.

HOFMANN M, & TIERNEY B. (2003). The Involvement of Human Resources in Large Scale Data Mining Projects. In *Proceedings of the 1st International Symposium on Information and Communication Technologies* (p. 103–109). Dublin, Ireland, Trinity College Dublin.

HOFMANN M, & TIERNEY B. (2009). An enhanced data mining life cycle. In *2009 IEEE Symposium on Computational Intelligence and Data Mining* (p. 109‑117).

HUGON M-A, & SEIBEL C. (1988). *Recherches impliquées. Recherche action : le cas de l'éducation.* Bruxelles, De Boeck Wesmael.

JARKE M, JEUSFELD M A, PETERS P, & POHL K. (1997). Coordinating distributed organizational knowledge. *Data & Knowledge Engineering*, Vol. 23, No. 3, p. 247‑268, septembre.

JEANNERET Y. (2007). Usages de l'usage, figures de la médiatisation. *Communication & Langages*, Vol. 151, No. 1, p. 3‑19.

JICK T D. (1979). Mixing Qualitative and Quantitative Methods: Triangulation in Action. *Administrative Science Quarterly*, Vol. 24, No. 4, p. 602‑611.

JIHJAH M. (2015). Usages ou pratiques : une (simple) querelle de mots ? *La Revue des Médias*. Consulté à l'adresse http://larevuedesmedias.ina.fr/usages-ou-pratiques-une-simple-querelle-de-mots

JODELET D. (2003). *Les représentations sociales*. Paris, Presses Universitaires de France.

JOHANSSON R. (2003). Case study methodology. In *the International Conference on Methodologies in Housing Research*. Stockholm.

JOUËT J. (1990). L'informatique « sans le savoir ». *Culture Technique*, No. 21, p. 215‑222.

JULES A. (2017). La polyvalence du leader. *I2D – Information, données & documents*, Vol. 53, No. 4, p. 34‑35, janvier.

JULES A, & LEBIGRE L. (2013). Présentation. *Documentaliste-Sciences de l'Information*, Vol. 50, No. 1, p. 22‑22, avril.

JURAN J, & GODFREY A B. (1999). Quality handbook. *Republished McGraw-Hill*.

KAPLAN R, & NORTON D. (1996). *The Balanced Scorecard: Translating Strategy Into Action*. Harvard Business Press.

KASKINEN J. (2007). Creating a Best-in-Class KPI Program. *Strategic Finance*, Vol. 89, No. 4, p. 29, octobre.

KAST R. (1993). *La théorie de la décision*. Paris, La Découverte.

KEMBELLEC G, CHARTRON G, & SALEH I. (2014). *Recommender Systems*. John Wiley & Sons.

KENWRIGHT D. (1999). Automation or interaction: what's best for big data? In *Visualization '99. Proceedings* (p. 491‑495). San Francisco, CA, USA, IEEE.

KETTENRING J R. (2001). « Massive Data Sets… Reflections On a Workshop » (Compte Rendu) (p. 5). Telcordia Technologies, Inc.

KHALIL C. (2011, décembre 6). « Les méthodes "agiles" de management de projets informatiques : une analyse "par la pratique" » (Thèse). Télécom ParisTech, Paris. Consulté à l'adresse https://pastel.archives-ouvertes.fr/pastel-00683828/document

KISH L B. (2002). End of Moore's law: thermal (noise) death of integration in micro and nano electronics. *Physics Letters A*, Vol. 305, No. 3–4, p. 144‑149, décembre.

KITCHIN R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, Vol. 1, No. 1, janvier. doi:10.1177/2053951714528481

KNUTH D E. (1969). *The Art of Computer Programming* (Vol. 1). Addison-Wesley Publishing Company.

KUHN T S. (1996). *The Structure of Scientific Revolutions, 3rd Edition* (3rd edition). Chicago, IL, The University of Chicago Press.

KURGAN L A, & MUSILEK P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, Vol. 21, No. 01, p. 1–24, mars.

KWON O, LEE N, & SHIN B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*, Vol. 34, No. 3, p. 387‑394, juin.

LAMOUROUX M, & FERCHAUD B. (2008). Journée d'étude ADBS. L'impact du numérique sur l'évolution des modes de travail. *Documentaliste-Sciences de l'Information*, Vol. 43, No. 3, p. 242‑246, décembre.

LANEY D. (2001, février 6). 3-D Data Management: Controlling Data Volume, Velocity and Variety. *Application Delivery Strategies by META Group Inc.*, Vol. 949, février.

LATOUR B. (1994). On Technical Mediation - Philosophy, Sociology, Genealogy. *Common Knowledge*, Vol. 3, No. 2, p. 29‑64.

LAVILLE F. (2000). La cognition située. Une nouvelle approche de la rationalité limitée. *Revue économique*, Vol. 51, No. 6, p. 1301‑1331.

LEBAS M. (1995). Oui, il faut définir la performance. *Revue Française de Comptabilité*, No. 275, p. 52‑57, août.

LENSKI S V. (2002). Example of a reconstruction of evolution of the genetic code (GC). *BGRS*, p. 179‑181.

LESK M. (1997). How Much Information Is There In the World? Consulté à l'adresse http://www.lesk.com/mlesk/ksg97/ksg.html

LIJPHART A. (1971). Comparative Politics and the Comparative Method. *American Poliical Science Review*, Vol. 65, No. 3, p. 682‑693, septembre.

LIPPMAN S A, & RUMELT R P. (1982). Uncertain Imitability: An Analysis of Interfirm Differences in Efficiency under Competition. *The Bell Journal of Economics*, Vol. 13, No. 2, p. 418.

LORINO P. (1997). *Méthodes et pratiques de la performance - Le pilotage par les processus et les compétences* (Éditions d'Organisation).

LOSHIN D. (2010). *Master Data Management*. Morgan Kaufmann.

LUCE R D, & RAIFFA H. (1989). *Games and decisions*. Courier Corporation.

LUYANG FU, & WANG H. (2014). CAS: Estimating Insurance Attrition Using Survival Analysis. *Variance*, Vol. 8, No. 1, p. 55‑72.

LYON L, & TAKEDA K. (2012, octobre). « What is a Data Scientist ? (...Data Scientists in the Wild...) ». Microsoft eScience Workshop, Chicago. Consulté à l'adresse http://research-srv.microsoft.com/en-us/um/redmond/events/escience2012/Liz_Lyon.pdf

MACH E. (1921). *Die Mechanik in Ihrer Entwicklung: Historisch-kritisch Dargestellt*. Brockhaus.

MANOVICH L. (2012). Trending: The Promises and the Challenges of Big Social Data. *The University of Minnesota Press*.

MANYIKA J, CHUI M, BROWN B, BUGHIN J, DOBBS R, ROXBURGH C, & HUNG BYERS A. (2011). « Big data: The next frontier for innovation, competition, and productivity ». McKinsey Global Institute. Consulté à l'adresse https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

MARBAN O, MARISCAL G, & SEGOVIA J. (2009). A Data Mining & Knowledge Discovery Process Model. In *Data Mining and Knowledge Discovery in Real Life Applications* (p. 438). Julio Ponce et Adem Karahoca.

MARIKO D. (2016). « Le Master Data Management (MDM) et la qualité des données de l'entreprise : synergies digitales et collaboratives ». INTD-CNAM.

MARR B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. John Wiley & Sons.

MARRON B A, & DE MAINE P A D. (1967). Automatic Data Compression. *Commun. ACM*, Vol. 10, No. 11, p. 711–715, novembre.

MARTIN J. (1991). *Rapid Application Development*. New York : Toronto : New York, Macmillan USA.

MARTORY B, & PIERRAT C. (1996). *La gestion de l'immatériel*. Nathan.

MARTRE H, CLERC P, & HARBULOT C. (1994). Intelligence économique et stratégie des entreprises. *Rapport du Commissariat Général au Plan, Paris, La Documentation Française*, Vol. 17.

MAYÈRE A. (1990). *Pour une Économie de L'Information*. C.N.R.S. Editions. doi:10.3917/cnrs.mayer.1990.01

MAYER-SCHOENBERGER V, & CUKIER K. (2014). *Big Data - La révolution des données est en marche* (Robert Laffont). Consulté à l'adresse http://www.laffont.fr/site/big_data_&100&9782221144510.html

MAYER-SCHÖNBERGER V, & CUKIER K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt.

MCAFEE A, & BRYNJOLFSSON E. (2012). Big Data: The Management Revolution. *Harvard Business Review*, octobre.

MCMASTER T, & WASTELL D. (2005). The Agency of Hybrids: Overcoming the Symmetrophobic Block. *Scandinavian Journal of Information Systems*, Vol. 17, No. 1, p. 175‑182.

MEINDL J D. (2003). Beyond Moore's Law: The Interconnect Era. *Computing in Science & Engineering*, Vol. 5, No. 1, p. 20‑24, janvier.

MERRIAM S B. (1998). *Qualitative Research and Case Study Applications in Education. Revised and Expanded from « Case Study Research in Education. ».* Jossey-Bass Education Series & Jossey-Bass Higner Education Series.

MICHELLE G G L-H, GOYETTE G, & HÉBERT-LESSARD M. (2014). *La Recherche-Action: Ses Fonctions, Ses Fondements et Son Instrumentation*. PUQ.

MILLEKER A. (2014, septembre 5). Le Data Scientist. Consulté 18 mai 2014, à l'adresse http://www.bigdatamonkeys.fr/le-data-scientist/

MILLER A C, MERKHOFER M W, HOWARD R A, MATHESON J E, & RICE T R. (1976). « Development of Automated Aids for Decision Analysis ». Arlington, Virginia: The Defense Advanced Research Projects Agency.

MILLER H G, & MORK P. (2013). From Data to Decisions: A Value Chain for Big Data. *IT Professional*, Vol. 15, No. 1, p. 57‑59, janvier.

MILLER H J. (2010). The Data Avalanche Is Here. Shouldn't We Be Digging? *Journal of Regional Science*, Vol. 50, No. 1, p. 181‑201.

MOIGNE J-L L. (2012). *Les épistémologies constructivistes* (4e éd.). Paris, Presses Universitaires de France.

MOINET N, & ALLOING C. (2016). Les signaux faibles : du mythe à la mystification. *Hermes, La Revue*, Vol. 76, No. 3, p. 86‑92, novembre.

MONTEIL J-M, & TRUCHOT D. (1986). Dynamique sociale et systèmes de formation. *Revue française de pédagogie*, Vol. 77, p. 100‑103.

MONTGOMERY P. (2002, février). Effective rolling forecasts. *Strategic Finance*, p. 41‑44.

MOORE G E. (1965). Cramming more components onto integrated circuits. *Electronics*, Vol. 38, avril.

MUCCHIELLI A. (1991). *Les méthodes qualitatives*. Paris, Presses Universitaires de France.

MUCCHIELLI A. (2009). *Dictionnaire des méthodes qualitatives en sciences humaines*. Armand Colin.

NABHOLTZ J M, DAUPLAT M M, ABRIAL C, WEBER B, MOURET-REYNIER M A, GLIGOROV J, … GUIU S. (2012). Is it possible to predict the efficacy of a combination of Panitumumab plus FEC 100 followed by docetaxel (T) for patients with triple negative breast cancer (TNBC)? Final biomarker results from a phase II neoadjuvant trial. *Cancer Research*, Vol. 72, No. 24 Supplement 3.

NARAYANAN N E. (2015). *Statistics* (2e éd.). PHI Learning.

NASCIMENTO G S do, & OLIVEIRA A A de. (2012). An Agile Knowledge Discovery in Databases Software Process. In *Data and Knowledge Engineering* (p. 56‑64). Springer, Berlin, Heidelberg.

NATIONAL RESEARCH COUNCIL. (1996). *Massive Data Sets: Proceedings of a Workshop* (The National Academies Press). Washington, DC. Consulté à l'adresse http://www.nap.edu/openbook.php?record_id=5505

NAUMANN F, & ROLKER C. (2000). Assessment Methods for Information Quality Criteria. In *IQ*.

NESME M-F. (2017). L'implication des collaborateurs. *I2D – Information, données & documents*, Vol. 53, No. 4, p. 36‑37, janvier.

NESME M-F, & COTTIN M. (2017a). Les principes de la qualité appliqués aux activités info-documentaires, Quality principles applied to information management activities. *I2D – Information, données & documents*, Vol. 53, No. 4, p. 30‑31, janvier.

NESME M-F, & COTTIN M. (2017b). L'orientation client/utilisateur. *I2D – Information, données & documents*, Vol. 53, No. 4, p. 32‑33, janvier.

NOYER J-M, & CARMES M. (2014). L'irrésistible montée de l'algorithmique : méthodes et concepts en SHS. *Les Cahiers du numérique*, Vol. 10, No. 4, p. 63‑102, novembre.

ODEH S, & CHARTRON G. (2016). Acteurs et économie des métadonnées du livre en France : analyse et avenir. *Documentation et bibliothèques*, Vol. 62, No. 1, p. 21‑32.

OLIVESI S. (2014). *Sciences de l'information et de la communication: Objets, savoirs, discipline*. PUG.

O'NEIL C. (2018). *Algorithmes : la bombe à retardement*. Les arènes.

OUELLET M, MONDOUX A, MÉNARD M, BONENFANT M, & RICHERT F. (2014). « "Big Data", gouvernance et surveillance ». Cahiers du CRICIS. Consulté à l'adresse http://www.cricis.uqam.ca/wp-content/uploads/2016/05/CRICIS_CAHIERS_2014-1.pdf

PAJOT P. (2016). Les statistiques face à la malédiction des grandes dimensions. *La Recherche*, No. 513, août.

PANISSIER S. (2007, novembre 13). « Positionnement d'un centre de documentation dans le management de l'information : méthodologie et impact sur l'activité documentaire. Le cas du centre de documentation de la Direction du développement des médias des Services du Premier Ministre ». Institut national des techniques de la documentation du CNAM. Consulté à l'adresse https://memsic.ccsd.cnrs.fr/mem_00000624/document

PAVEL I, & SERRIS J. (2016). « Modalités de régulation des algorithmes de traitement des contenus » (p. 63). Conseil Général de l'Economie.

PEARSON K. (1900). Mathematical Contributions to the Theory of Evolution. VII. On the Correlation of Characters not Quantitatively Measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, Vol. 195, p. 1‑405.

PÉREZ R. (2010). *La gouvernance de l'entreprise*. Paris, La Découverte.

PIATETSKY-SHAPIRO G. (1994). An Overview of Knowledge Discovery in Databases: Recent Progress and Challenges. In *Rough Sets, Fuzzy Sets and Knowledge Discovery* (p. 1‑10). Springer, London.

PIGNI F, PICCOLI G, & WATSON R. (2016). Digital Data Streams. *California Management Review*, Vol. 58, No. 3, p. 5‑25, mai.

PIRES A. (1997). Échantillonnage et recherche qualitative: essai théorique et méthodologique. In *La recherche qualitative. Enjeux épistémologiques et méthodologiques* (p. 113–169).

POOL I D S. (1984). *Communications Flows: A Census in the United States and Japan*. North-Holland.

PORAT M U. (1977). « The Information Economy: Definition and Measurement. » Washington, DC: Office of Telecommunications.

PORTER M E. (1989). From Competitive Advantage to Corporate Strategy. In D. Asch & C. Bowman (Éd.), *Readings in Strategic Management* (p. 234‑255). Macmillan Education UK.

PORTER M E. (1998). *Competitive Advantage of Nations: Creating and Sustaining Superior Performance*. Simon and Schuster (ed. 2011).

PRESSMAN R S. (2005). *Software Engineering: A Practitioner's Approach*. Palgrave Macmillan.

PRÉVOT-HUBERT M. (2004). Les professionnels de l'information en France. *Documentaliste-Sciences de l'Information*, Vol. 41, No. 3, p. 182‑186.

PROVOST F, & FAWCETT T. (2013a). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, Vol. 1, No. 1, p. 51‑59, février.

PROVOST F, & FAWCETT T. (2013b). *Data Science for Business: What you need to know about data mining and data-analytic thinking* (1 edition). Sebastopol, Calif., O'Reilly Media.

QUETELET A. (1849). *Letters Addressed to H.R.H. the Grand Duke of Saxe Coburg and Gotha: On the Theory of Probabilities, as Applied to the Moral and Political Sciences*. C. & E. Layton.

RAIFFA H, BELL D E, & TVERSKY A. (1988). *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. New York, Cambridge University Press.

RANGER C, RAGHURAMAN R, PENMETSA A, KOZYRAKIS C, & BRADSKI G. (2007). Evaluating MapReduce for Multi-core and Multiprocessor Systems. In *2007 IEEE 13th International Symposium on High Performance Computer Architecture* (p. 13‑24). Scottsdale, AZ, USA, IEEE.

RATNER B. (2004). *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*. CRC Press.

RHEINBERGER H-J. (2014). *Introduction à la philosophie des sciences*. Paris, La Découverte.

ROBBINS S, & JUDGE T. (2011). *Comportements organisationnels* (14e éd.). Pearson.

ROHANIZADEH S S, & MOGHADAM M B. (2009). A proposed Data Mining Methodology and its application to industrial procedures, p. 37‑50.

ROSENBLUETH A, WIENER N, & BIGELOW J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, Vol. 10, p. 18‑24.

ROSNAY J de. (1991). *Le Macroscope: vers une vision globale*. Paris, Éditions du Seuil.

ROUVROY A, & BERNS T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation. *Reseaux*, Vol. n° 177, No. 1, p. 163‑196, mai.

SALAÜN J-M. (2007). La redocumentarisation, un défi pour les sciences de l'information. *Études de communication. langages, information, médiations*, No. 30, p. 13‑23, octobre.

SALAÜN J-M, MICHEL J, BATTISTI M, HORN F, BOMSEL O, & CHANTEPIE P. (2011). Économie de l'information : les fondamentaux. *Documentaliste-Sciences de l'Information*, Vol. 48, No. 3, p. 24‑35, novembre.

SAMSONOWA T, BUXMANN P, & GERTEIS W. (2009). Defining kpi sets for industrial research organizations - a performance measurement approach. *International Journal of Innovation Management*, Vol. 13, No. 02, p. 157‑176, juin.

SCHALLER R R. (1997). Moore's law: past, present and future. *IEEE Spectrum*, Vol. 34, No. 6, p. 52‑59, juin.

SEGARAN T, & HAMMERBACHER J. (2009). *Beautiful Data: The Stories Behind Elegant Data Solutions*. O'Reilly Media, Inc.

SENGE P M. (1990). *The fifth discipline: the art and practice of the learning organization*. Doubleday/Currency.

SHANNON C E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*.

SHAPIRO C, & VARIAN H R. (1998). *Information Rules. A Strategic Guide to the Network Economy*. Boston, Massachusetts, Harvard Business School Press.

SHARMA S. (2013). Big Data Landscape. *International Journal of Scientific and Research Publications*, Vol. 3, No. 6, p. 861, juin.

SHEARER C. (2000, Fall). The CRISP-DM model : the new blueprint for data mining. *Journal of Data Warehousing*, p. 13‑22.

SIMON H A. (1994). The bottleneck of attention: Connecting thought with motivation. In *Integrative views of motivation, cognition, and emotion* (Vol. 41, p. 1‑21). Lincoln, NE, US, University of Nebraska Press.

SIMON P. (2013). *Too Big to Ignore: The Business Case for Big Data*. Wiley.

SIMONNOT B. (2012). *L'accès à l'information en ligne : Moteurs, dispositifs et médiations*. Hermès Lavoisier.

SMUTS HON J. C. (1927). *Holism And Evolution*. Macmillan And Company Limited.

STAKE R E. (1994). Case studies. In *Handbook of qualitative research* (p. 236‑247). Sage Publications.

STEIN B, & MORRISON A. (2014). « The enterprise data lake: Better integration and deeper analytics » (Technology Forecast: Rethinking integration No. 1). PWC.

SUTTER E. (2005). Le management de l'information: présentation commentée du document de normalisation. *L'Essentiel sur...*, p. 50‑185.

TACHEAU O. (1998). « Bibliothèque publique et multiculturalisme aux Etats-Unis. Jalons pour repenser la situation française » (Memoire d'Etude pour le Diplôme de conservateur de bibliothèque). Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques, France.

TAMBE P. (2012). *How the IT Workforce Afects Returns to IT Innovation: Evidence from Big Data Analytics*. New York University's Stern School.

TAYLOR R W, & LICKLIDER J C R. (1968). *The Computer as a Communication Device*. Science and Technology.

TEMPL M. (2012). Correlation between indicators over time in thematic maps. *Austrian Journal of Statistics*, Vol. Volume 41, No. Number 1, p. 67‑79.

TJOMSLAND I A ; et al. (1980). *Digest of Papers: The Gap between MSS Products and User Requirements, Fourth IEEE Symposium on Mass Storage Systems, April 15-17, 1980, Regency Hotel Denver, Co. IEEE Catalog No. 80CH1581-8.* (1St Edition edition). National Center for Atmospheric Research/NSF.

TOFFLER A. (1984). *Future Shock* (Reissue). New York, Bantam.

TURCK M. (2016, février 1). Is Big Data Still a Thing? (The 2016 Big Data Landscape). Consulté 8 août 2016, à l'adresse http://mattturck.com/2016/02/01/big-data-landscape/

VAN DER AALST W M P. (2014). Data Scientist: The Engineer of the Future. In K. Mertins, F. Bénaben, R. Poler, & J.-P. Bourrières (Éd.), *Enterprise Interoperability VI* (p. 13‑26). Cham, Springer International Publishing.

VARIAN H R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, Vol. 28, No. 2, p. 3‑28, mai.

VARIAN H R, & LYMAN P. (2003). *How Much Information?* UC Berkeley.

VILMINKO-HEIKKINEN R, & PEKKOLA S. (2017). Master data management and its organizational implementation: An ethnographical study within the public sector. *Journal of Enterprise Information Management*, Vol. 30, No. 3, p. 454–475.

VON BERTALANFFY L. (2012). *Théorie générale des systèmes*. Paris, Dunod.

VON NEUMANN J, & MORGENSTERN O. (1945). *Theory of games and economic behavior*. Princeton University Press Princeton.

WANG Randolph Y., ANDERSON T E, & PATTERSON D A. (1999). Virtual Log Based File Systems for a Programmable Disk. In *Proceedings of the Third Symposium on Operating Systems Design and Implementation* (p. 29–43). Berkeley, CA, USA, USENIX Association.

WANG Richard Y. (1998). A Product Perspective on Total Data Quality Management. *Commun. ACM*, Vol. 41, No. 2, p. 58–65, février.

WEGMANN G. (2008). Comparaison Balanced Scorecard - Navigator. In *Indicateurs et tableaux de bord* (p. 1‑11). Afnor éditions.

WEISS S M, & INDURKHYA N. (1998). *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann.

WENGER E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. New York, Cambridge University Press.

WIENER N. (1948). *Cybernetics Or Control and Communication in the Animal and the Machine*. Paris, (Hermann & Cie) & Camb. Mass. (MIT Press).

WIRTH R, & HIPP J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (p. 29–39).

YIN R K. (1981). The Case Study Crisis: Some Answers. *Administrative Science Quarterly*, Vol. 26, No. 1, p. 58‑65.

YIN R K. (1984). *Case study research: design and methods* (Vol. 5). Sage Publications.

ZACKLAD M, CAHIER J-P, BÉNEL A, ZAHER L, LEJEUNE C, & ZHOU C. (2007). Hypertopic: une métasémiotique et un protocole pour le Web socio-sémantique. In *Actes des 18eme journées francophones d'ingénierie des connaissances* (p. 13). Francky Trichet.

Annexes

Annexe 1 – Courte histoire du Big Data et des algorithmes

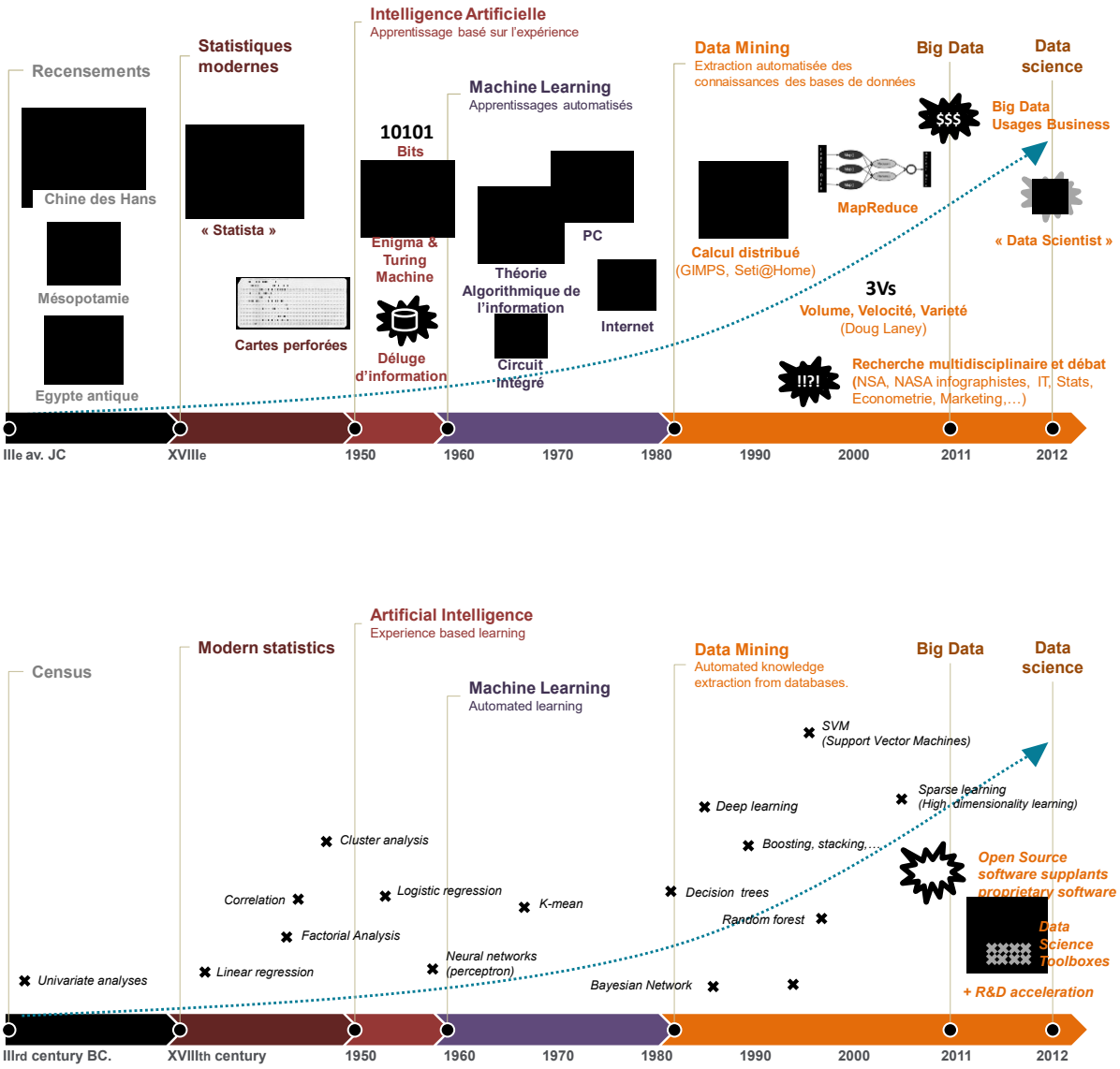


Figure 52 – Un aperçu de l'histoire du Big Data et de la Data Science

Annexe 2 - Data Lakes et Informatique Décisionnelle

L'une des technologies les plus emblématiques du phénomène Big Data est le *Data Lake*. Il s'agit d'un réservoir de stockage de données capable de collecter les données issues de sources hétérogènes, dans leur format natif et sans pré-catégorisation, et de servir de socle pour procéder à des requêtes rapides directement sur ces sources de données, structurées ou non (Not Only SQL, ou NoSQL). Les méthodes et technologies mobilisées (Hadoop, Spark...) rendent ainsi le stockage de gros volumes de données abordable et flexible et déstabilisent les principes de l'Informatique Décisionnelle.

En effet, historiquement l'interface permettant aux décideurs métier de s'approprier les informations et indicateurs de pilotage est constituée d'un ensemble de solutions informatiques. Il s'agit de l'Informatique Décisionnelle, ou Business Intelligence (BI), qui permet la production des reporting, tableaux de bord et autres interfaces de restitution d'informations utiles à la prise de décision. L'informatique décisionnelle fait son apparition dans les entreprises dans les années 70 sous forme d'infocentres, avec un envoi de requêtes sur les serveurs de production, gérant l'information opérationnelle, c'est-à-dire prenant en charge les actions métier au jour le jour, afin de restituer les informations contenues dans ces derniers. Progressivement, une isolation est établie entre les flux d'information en production et les flux d'information décisionnelle afin d'éviter des risques de sécurité et de ralentissement du dispositif de production. Le dispositif décisionnel permet alors l'analyse de données « à froid » afin d'appuyer l'apprentissage, le pilotage de l'activité et la stratégie de l'entreprise.

La production de restitutions d'informations décisionnelles s'appuie sur une architecture IT composée d'éléments propres à chaque étape de la chaîne de traitement de l'information qui s'étale de la collecte à la restitution de celle-ci. La première étape consiste à extraire les données issues des différentes sources d'information, les transformer et les charger au sein d'un entrepôt de données : il s'agit de l'ETL, ou *Extract, Transform and Load*. Un entrepôt de données (ou datawarehouse) transversal et complet doit garantir la qualité des données, c'est-à-dire leur intégrité, des règles de gestion cohérentes, et une historisation stable (principe de non-volatilité, absent des systèmes de production où une nouvelle donnée opérationnelle remplace la précédente). Les données collectées peuvent alimenter des comptoirs de données plus restreints (ou datamarts), ces derniers étant des bases de données relationnelles destinées à répondre à des besoins métiers précis, et donc limitées en termes de périmètre de données. L'historique de données y est réduit au périmètre opérationnel, et les données sont agrégées selon les modalités

définies en amont pour la prise de décision. Les données font ensuite l'objet de requêtes : il s'agit de sélectionner, trier, regrouper, répartir, ou réaliser divers calculs sur les données afin d'obtenir un résultat sous forme de liste ou d'indicateur. Enfin, la visualisation de ce résultat peut prendre la forme d'une table, d'un hypercube, d'une représentation graphique ou d'une Data Visualisation plus sophistiquée, l'objectif étant de permettre une compréhension efficiente du résultat par un expert métier.

L'Informatique Décisionnelle traditionnelle est limitée à plusieurs niveaux. D'une part, la multiplicité des sources de données n'est pas prise en compte dans la décision : seules les bases de données relationnelles sont contenues dans un entrepôt classique. Elles sont constituées de tables (lignes et colonnes) reliables entre elles selon un modèle logique de données. Or, la décision des experts s'appuie en réalité sur un ensemble plus vaste de données, qui comprend des données structurées, semi-structurées (mots clés, balises de métadonnées, logs, fichiers CSV ou XML…), non structurées (emails, documents Word ou PDF, Excel, HTML...), voire binaires (images, son ou vidéo). D'autre part, les experts métiers sont contraints par les datamarts, limités à des usages prédéfinis en termes de nature et de niveau d'agrégation des données, à se restreindre à leur silo de données, ce qui empêche d'appuyer une décision éclairée de façon transversale et ne permet pas l'émergence d'usages et de connaissances nouvelles en dehors du scope opérationnel prédéfini. Par ailleurs, les traitements techniques sont lourds et fonctionnent par nature sous forme de batch et non pas en flux. Cela n'apporte qu'une réponse limitée au besoin d'accéder à l'information la plus récente afin de supporter la prise de décision opérationnelle en temps réel, voire d'automatiser la prise de décision. Enfin, les analyses standard embarquées dans les outils de Business Intelligence sont restreintes à l'accès à l'information, sa transformation, et son analyse descriptive : elles ne couvrent pas les opportunités nouvelles offertes par le développement d'algorithmes prédictifs ou prescriptifs, particulièrement utiles dans le cadre du pilotage de la stratégie et la prise de décision en entreprise. La solution Data Lake apporte des réponses concrètes à ces problématiques et ouvre ainsi la voie à de nouveaux usages métier.

Cependant, cette technologie comporte elle aussi des limites : contrairement à un entrepôt de données structuré transversal (Datawarehouse) ou dédié à un ensemble d'utilisateurs (Datamart), un Data Lake ne garantit pas l'intégrité des données, manque de maturité en termes de cadre de sécurité et de gouvernance, et n'élimine pas le besoin de structurer les données stockées pour permettre les analyses. Par ailleurs, le succès rencontré par cette solution s'est

heurté au cours des dernières années à l'écart entre le discours commercial et la réalité terrain : si la collecte des données est en effet accélérée et à moindre coût, la phase de création de valeur reste sujette à la mobilisation d'experts métier et data pour concevoir les nouveaux usages et structurer les données permettant la réalisation des analyses et leur restitution. L'appropriation de ces technologies, notamment sous l'angle des langages de programmation, a nécessité un ensemble d'ajustements progressifs en termes de compétences ou bien de développement de facilitateurs. Par exemple, Appache Hive est développé par Facebook pour faciliter le travail de leurs salariés sur les Data Lakes en ajoutant une couche Datawarehouse permettant des requêtes proches de SQL (HiveQL), en langage logique. Ou bien Apache Pig, en langage procédural (description d'étapes de transformation des données) est développé par Yahoo pour faciliter la création et l'exécution de jobs MapReduce spécifiques, afin de s'émanciper du langage Java. Les compétences data restent à ce jour encore rares sur le marché, étant donné la nouveauté technologique.

L'implication des experts business est quant à elle conditionnée à la compréhension des nouvelles possibilités offertes par le progrès en termes de possibilités analytiques. Cette situation est similaire à la mise en place de l'Intelligence Décisionnelle traditionnelle : si ce dispositif a été annoncé à usage des experts métiers directement, seuls 20 à 25% des utilisateurs métiers l'utilisent (Dull, 2015), l'essentiel des utilisateurs restant des experts IT. La différence principale réside dans le fait que les analyses réalisées sur cette base sont plus avancées que celles permises par les outils de Business Intelligence standard : la possibilité de mixer les sources structurées et non structurées apportant des promesses d'enrichissement des matrices d'apprentissage, la profusion de méthodes d'analyses prédictives, prescriptives et autres nécessite la mobilisation de compétences statistiques plus spécifiques pour la formalisation de l'objectif de l'analyse, et donc pour la structuration des données nécessaire pour ces analyses. La mise en place de Data Lakes (voir Figure 53) implique alors la présence de Data Scientists en tant qu'interface humaine entre les données et les experts métier, et qu'acteur de production de valeur dans le cadre de conception et mise en œuvre de nouveaux cas d'usage métier.

Figure 53 – Illustration d'un Data Lake – *Source : The enterprise Data Lake: Better integration and deeper analytic, PwC Technology Forecast* (Stein & Morrison, 2014)

L'Intelligence Décisionnelle traditionnelle est alors doublement bousculée par les solutions Data Lake, couplées aux algorithmes issus de la Data Science et aux outils de restitution ergonomiques et performants : d'une part, elle est dépassée face aux opportunités de nouveaux usages qui s'affranchissent des limites de la Business Intelligence traditionnelle et court-circuitent la chaîne de traitement de la donnée, et d'autre part elle doit se restructurer en termes de compétences métier et data au service de la production de valeur dans le cadre de ces cas d'usage.

Par ailleurs, l'adoption des technologies comme les Data Lakes repose la question de la gestion de la qualité des données sous un nouvel angle. Selon une étude de Kwon, Lee et Shin sur l'adoption des technologies Big Data, les perspectives théoriques de gestion de qualité de l'information ainsi que de l'expérience de l'usage des données les entreprises devraient avoir l'intention d'adopter des technologies analytiques Big Data (Kwon et al., 2014). L'expérience de l'usage des données est alors basée sur la théorie de management par les ressources (Ressource Based View) et l'isomorphisme externe et interne. L'analyse empirique démontre

que cette intention est amplifiée par la compétence à maintenir la qualité des données internes, mais aussi par les expériences passées positives (création de bénéfice) avec les données externes. Cependant, contre toute attente, si l'entreprise a eu des expériences passées positives en utilisant seulement des données internes, son intention à investir dans ces technologies est entravée. En effet, les premières confrontations avec le terrain confirment que les entreprises privilégient la mise en qualité et la valorisation de leurs données existantes à l'investissement technologique ou à l'achat de données externes.

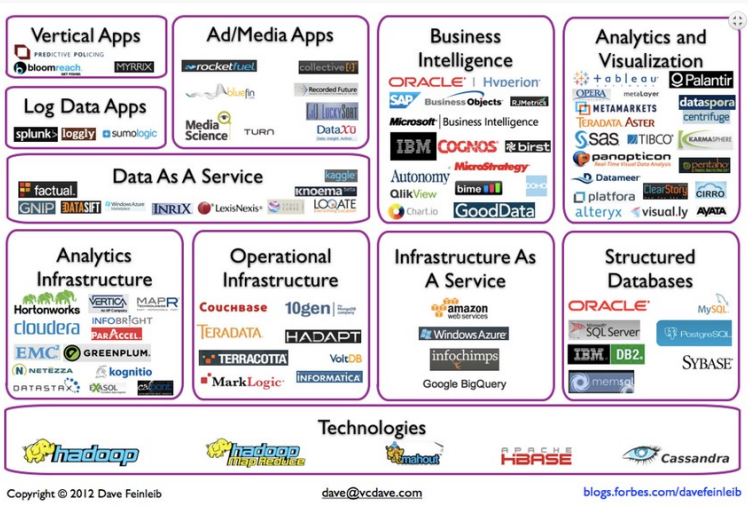
Les premières confrontations avec le terrain pointent par ailleurs un autre thème récurrent : les entreprises ayant investi dans les technologies de type Data Lake se retrouvent en difficulté pour démontrer l'intérêt de cet investissement. Cette observation, confirmée par d'autres sources (Stein & Morrison, 2014), montre progressivement l'erreur des organisation qui ont mis en place des Data Lakes complets et chargés de données : ces dernières sont inexploitables en absence de qualification (gestion des métadonnées) et de définition d'usages cibles. Au-delà des coûts de ces technologies, les échecs génèrent de la déception face aux promesses initiales. Les entreprises se tournent alors vers les Data Scientists pour mener des projets data sur ces environnements afin de démontrer leur valeur. Cette sollicitation peut avoir lieu à des stades différents d'exploitation de ces technologies : soit en amont des réflexions sur la mise en place d'un Data Lake, soit dès la création du socle technique (chargement des premières données internes), soit au cours de la décision d'ajout de données externes, parfois payantes, soit une fois les données chargées.

Annexe 3 - Ecosystème Big Data

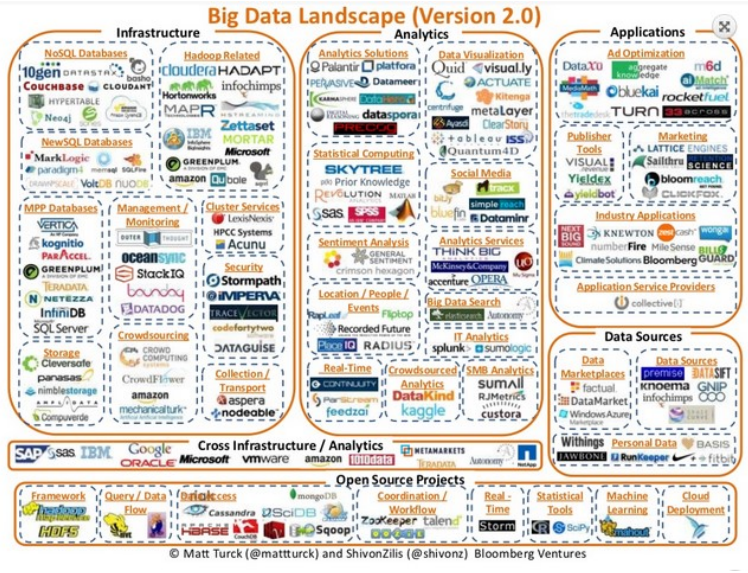
L'écosystème Big Data, en pleine croissance (Bourdoncle, 2014; Turck, 2016), est constitué d'une communauté d'organisations et de moyens technologiques (voir Figure 54). Il interagit avec les systèmes existants sous forme de flux transactionnels, matériels (produits et services) et informationnels. La survie de l'écosystème est alors dépendante de l'existence d'un marché, soit de la volonté des entreprises et des institutions en place à mettre en œuvre des projets qui dynamisent les flux d'interaction avec l'écosystème Big Data. Actuellement, l'écosystème est en plein développement, soutenu par un ensemble de mesures gouvernementales (la commission innovation 2030 en France désigne le Big Data comme l'une des sept ambitions stratégiques en 2013) sous forme de campagnes d'évangélisation, d'investissements financiers, de soutien à l'innovation, d'adaptations règlementaires et de formation académique de profils adaptés, les Data Scientists : cette aide est destinée autant aux acteurs de l'écosystème Big Data qu'à ses clients de tout secteur.

L'écosystème Big Data est composé historiquement des acteurs majeurs, au cœur de la génération massive des données et de la création de valeur à partir de celles-ci. Il s'agit des GAFA (Google, Apple, Facebook, Amazon) et autres géants du web américains (Yahoo, Twitter, LinkedIn...). Leur capacité à innover, appuyée sur des ressources et des méthodes, permet d'impulser l'éclosion de l'écosystème du point de vue des technologies et des usages. La stratégie de ces géants comprend la mise à disposition de certaines technologies en Open Source. Les compétences s'exportent, et des start-ups nativement ancrées sur ce sujet gagnent en autonomie. Les fournisseurs existants de services et de produits s'alignent alors en intégrant de nouveaux modules dans leur offre. L'écosystème initié ainsi atteint en 2016 un seuil de maturité suffisant pour permettre aux entreprises de mettre en place des programmes de transformation plus ambitieux que les preuves de concepts réalisées au cours des premières années du phénomène (Sharma, 2013). Il permet en effet de couvrir l'ensemble de *la chaîne de valeur* (Porter, 1998) du traitement de l'information, que ce soit en termes de technologies ou de prestations. La chaîne de valeur de la donnée (H. G. Miller & Mork, 2013) correspond au processus de création de valeur à partir de la donnée, et comprend les étapes de transformation des données ainsi que les activités de support transversales, c'est-à-dire l'infrastructure nécessaire (développement ou mise à disposition des technologies appropriées…), la gestion des ressources humaines (structures de formation ou de recrutement des *Data Scientists*, prestations intellectuelles…), la recherche et le développement...

July 2012



July 2013



Suite de la figure page suivante.

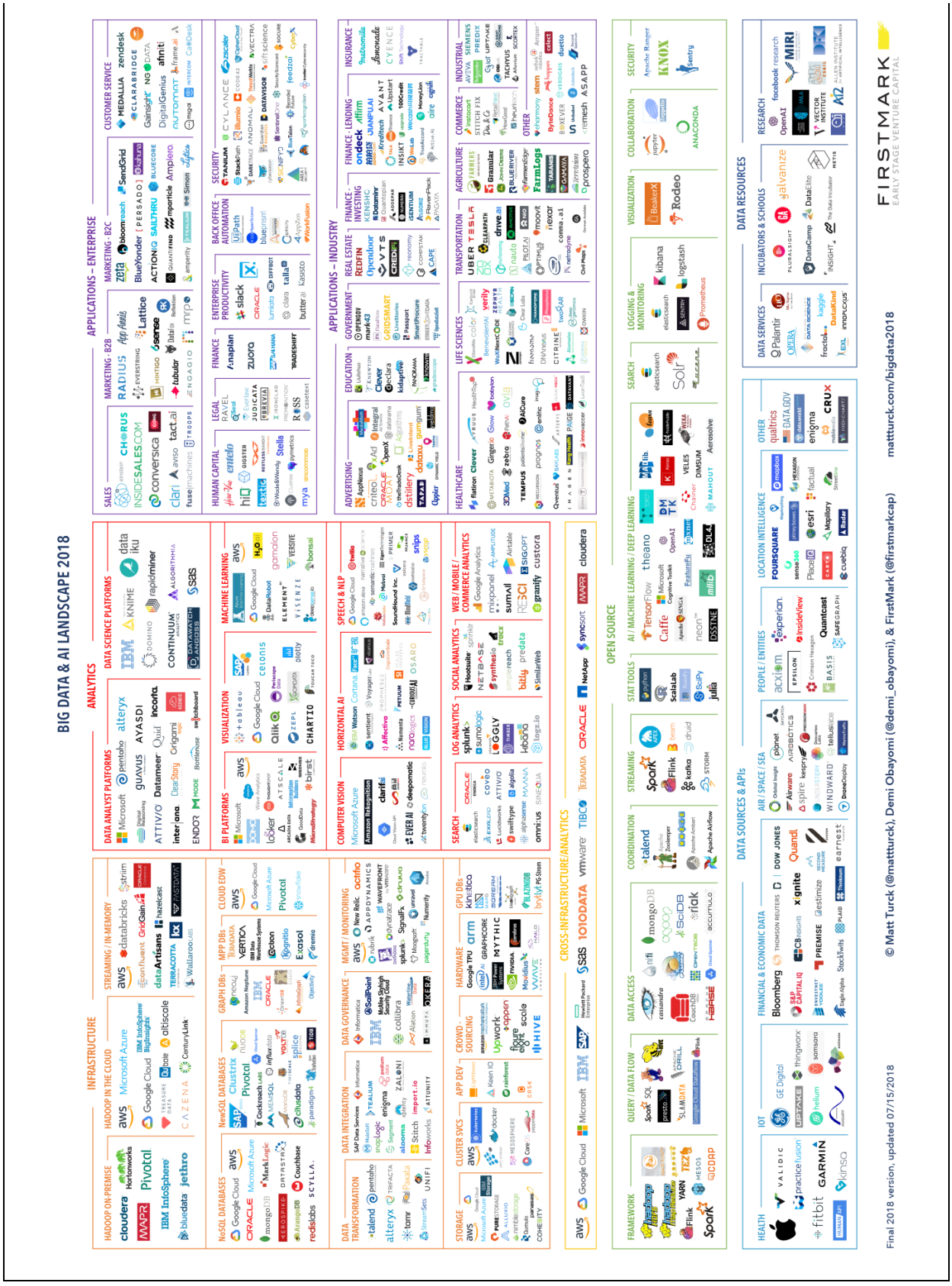


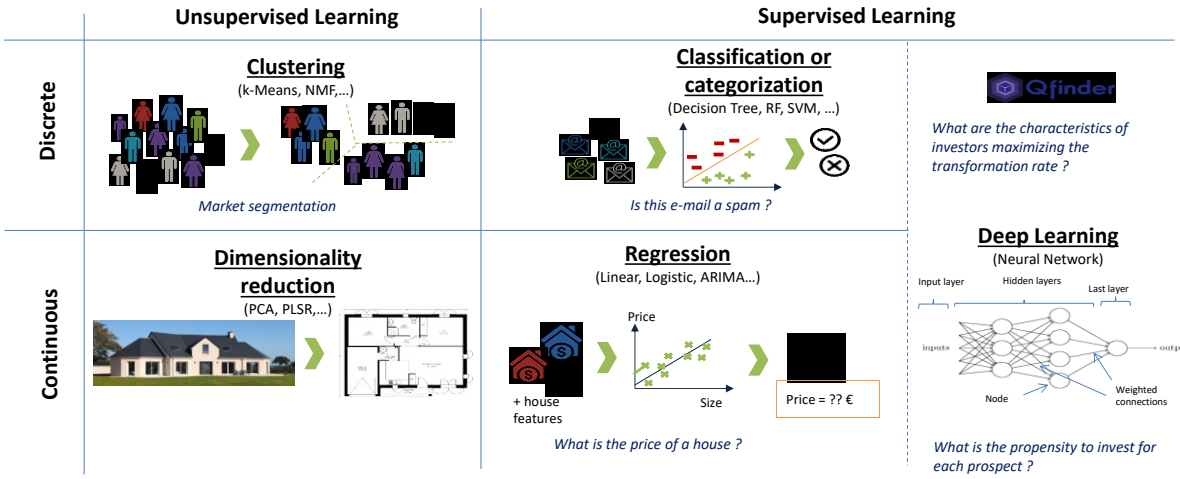
Figure 54 – Evolution de l'écosystème Big Data – *Sources : Heliocor*[51] *et Matt Turck*

[51] http://heliocor.com/solutions/compliance-big-data-tooling/an-analytic-advantage/#prettyPhoto

Annexe 4 - Data Science et algorithmes

Les Data Scientists constituent une nouvelle notion, au cœur du phénomène Big Data. Le terme désigne un nouveau métier, rare et à la mode (Davenport & Patil, 2012), comprenant 3 compétences principales (Conway, 2010; Milleker, 2014) : la connaissance mathématique et statistique, une expertise métier ou scientifique de fond qui permet d'imaginer et de mettre en œuvre une stratégie de valorisation des données et une démarche d'extraction de connaissances utiles, ainsi qu'un goût prononcé pour l'informatique (« Hacking »). Le rôle du Data Scientist, appelé « l'ingénieur du futur » (van der Aalst, 2014), consiste à utiliser les données disponibles pour construire des produits et services afin d'aider les organisations à formuler et atteindre des objectifs (Segaran & Hammerbacher, 2009). Il fait partie des métiers reconnus et clés pour le numérique, et fait référence aux compétences d'un individu maîtrisant les composantes de la *Data Science*, processus d'extraction de connaissances à partir des données. La Data Science est en soi proposée comme nouvelle discipline (van der Aalst, 2014), dérivée soit des statistiques, soit de l'informatique. Initialement appelée « datalogie », elle se situe en effet à l'intersection des deux disciplines, et s'appuie sur un ensemble d'institutions de communication (Data Science Journal, International Journal of Data Science and Analytics...) et d'éducation dédiées. Si la Data Science n'a pas attendu le phénomène Big Data pour exister, elle a connu un succès croissant avec le buzz sur le sujet, provoquant parfois des confusions, et ce notamment grâce aux progrès et l'accessibilité croissante des méthodes et technologies du Machine Learning et l'Intelligence Artificielle. La définition de la Data Science comme discipline est critiquée pour son manque de différenciation avec les statistiques, cependant elle semble comprendre un ensemble de caractéristiques propres beaucoup plus larges : au-delà de ses fondements théoriques et pratiques liés aux statistiques et à l'informatique, la Data Science inclut par exemple le Data Mining (Granville, 2014), la Data Visualisation (Lyon & Takeda, 2012; Turck, 2016), ou le Storytelling (Bladt & Filbin, 2013), soit la capacité à faire parler les données, ce qui en fait un concept pluridisciplinaire. Cette richesse renvoie à la difficulté à reboucler avec les compétences concrètes attendues de la part des Data Scientists. Se profile alors la nécessité de travailler au sein d'équipes de Data Science, mobilisant des acteurs aux rôles différents. La Data Science est appliquée en 2016 au service de secteurs économiques variés et de disciplines scientifiques diverses afin de créer des connaissances et de la valeur en mobilisant notamment des techniques de *Machine Learning* et d'Intelligence Artificielle.

Le Machine Learrning est une discipline visant l'étude et la création algorithmique, issu de la convergence des méthodes statistiques et des outils de programmation informatique, évoluant vers l'Intelligence Artificielle, c'est-à-dire la simulation de l'intelligence humaine. Il comprend un ensemble de techniques visant la détection de tendances (prédiction, sélection de variables…), et plus généralement l'apprentissage dans des ensembles de données. Les techniques communes comprennent les techniques comme la classification, les arbres de régression, ou encore le Deep Learning basé sur les progrès réalisés dans l'analyse des réseaux neuronaux. Les tâches réalisées grâce au Machine Learning peuvent être non supervisées ou supervisées. Dans le premier cas, il s'agit d'un apprentissage automatisé pour dégager une structure des données, comme dans le cas d'une segmentation classique de marché. Dans le second, il s'agit d'apprendre les liens qui existent entre un phénomène d'intérêt (donné comme output par le « superviseur »), et les données d'entrée : cet apprentissage résulte par exemple dans un score de risque. D'autres catégories d'apprentissage, intermédiaires ou complémentaires, existent, comme l'apprentissage semi-supervisé ou par renforcement. Les méthodes de Machine Learning peuvent par ailleurs être classées selon la nature du résultat. Il s'agit de classification (détection d'attributs permettant de ranger un élément dans une catégorie, comme les mails spam et non spam), de régression (le résultat est alors non pas discret, mais continu, comme pour un score), de clustering (identification de catégories, inconnues à l'avance, contrairement à la classification), ou encore la réduction de dimensions (simplification des attributs, comme pour le traitement automatique du langage naturel). Chaque catégorie comprend un ensemble de méthodes qui possèdent leurs propres caractéristiques intrinsèques, et sont choisies ou articulées par le Data Scientist pour répondre à l'objectif des analyses et aux particularités des données (volumes, qualité, nature...). Ainsi, l'une des façons de classer les algorithmes, particulièrement utilisée par les Data Scientists qui font face à un choix entre plusieurs familles algorithmiques possibles, est de les distinguer entre les algorithmes supervisés et non supervisés, ou encore entre les algorithmes qui donnent des résultats continus ou discrets, bien que ces frontières ne soient pas stables et discriminantes, comme le montre l'illustration ci-dessous (Extrait 15), issue d'un support pédagogique de Quinten à destination des Clients.



Extrait 15 ***Illustration des familles d'algorithmes*** *en fonction de leur résultat (supervisé ou non, discret ou continu) issue d'une présentation pédagogique de Quinten en 2015*

La disponibilité de cette richesse des choix algorithmiques est en soi assez récente. En effet, habituellement un algorithme est construit, par des laboratoires de recherche ou des acteurs privés, puis mis à disposition des entreprises, de façon gratuite ou payante, pour inspirer de nouveaux cas d'usages. C'est le cas notamment du Deep Learning, qui a été très à la mode au cours des dernières années, sous l'impulsion du développement des réseaux neurones artificiels multicouches, ou encore de la médiatisation des victoires, dès 2015, du programme de Google DeepMind, AlphaGo, face à Fan Hui et à Lee Sedol, champions européen et mondial du jeu de go. Aujourd'hui, la mise à disposition d'un nombre croissant d'algorithmes en Open Source, leur ajustement par les utilisateurs, et la commercialisation de plateformes regroupant ces algorithmes variés retourne le processus de choix : il ne s'agit pas de trouver un usage intéressant pour un algorithme nouveau, puis de l'optimiser progressivement dans le cadre de cet usage, mais de choisir un ou une combinaison d'algorithmes disponibles face à un usage prédéterminé. Ce déplacement des phases est loin d'être anodin dans les projets data, en particulier lorsque cela nécessite de caractériser l'utilité des algorithmes disponibles, et non pas l'éligibilité des usages.

Parmi les algorithmes disponibles, de nombreux regroupements sont aujourd'hui réalisés, l'un des plus simples étant proposé par Pedro Domingos (Domingos, 2015). Il décrit les familles d'algorithmes comme des « tribus », correspondant plus exactement aux différents courants de recherche qui travaillent sur l'intelligence artificielle. Les cinq tribus décrites sont les Symbolistes (inspirés des courants logiques et philosophiques, basés sur la déduction inverse),

les Bayésiens (inspirés des statisticiens et de l'approche par l'interférence probabiliste), les Analogiseurs (inspirés de la psychologie et des méthodes comme l'astuce du noyaux des machines de Kernel), les Evolutionnaires (inspirés de la biologie et de la programmation générique), et enfin les Connectionistes (inspirés de la neuroscience et des méthodes comme la rétropropagation du gradient). Si les trois premières tribus reproduisent des fonctionnements assez intuitifs de raisonnement humain et donnent des résultats explicites (décompositions logiques, niveau de confiance, sous-groupes ressemblants…), des deux dernières approches, très médiatisées au cours de dernières années par les acteurs comme Google, sont typiquement associés aux boites noires car le traitement des informations entre la donnée d'entrée et le résultat est particulièrement complexe.

Dans tous les cas, ces méthodes permettent de dégager une généralisation, tout en respectant la théorie d'apprentissage statistique, à partir d'une expérience représentée par l'ensemble de ses observations sous forme d'une *matrice d'apprentissage*. Une matrice d'apprentissage constitue la structure comprenant la matière première du Data Scientist : les données. Elle est généralement constituée d'un ensemble de lignes, correspondant à l'objet de l'analyse, c'est-à-dire sa granularité (ex : une ligne par client), d'un ensemble de colonnes input, correspondant aux attributs, ou caractéristiques de l'objet (ex : âge, sexe, nombre de contrats détenus, département...), et, dans le cas d'un apprentissage supervisé, d'une colonne output, c'est-à-dire le phénomène d'intérêt (ex : le client a-t-il résilié son contrat ?). La construction et l'utilisation de cette matrice d'apprentissage est déterminée par le Data Scientist en fonction de l'objectif de l'analyse, et donc des méthodes d'apprentissage mobilisées. Elle est par définition structurée, contrairement aux sources de données qui l'alimentent et qui peuvent être structurées ou non. Dans l'exemple illustré ci-dessus (voir Figure 55), il s'agit d'une structuration regroupant plusieurs sources (bases de données) différentes au sein d'une entreprise. Dans le cas de text mining, il s'agit par exemple de matrice d'occurrence dans un corpus de documents (lignes) d'un ensemble de termes (attributs en colonnes). Dans le cadre de traitement d'images, mobilisant les méthodes comme les réseaux neurones et SVM, il s'agit d'une base d'images (lignes), de leurs caractéristiques, comme la résolution, l'éclairage, la luminance, le contraste, ou les couleurs (attributs en colonnes), et de valeur de sortie, par exemple la présence d'une catégorie de véhicule militaire (colonne output).

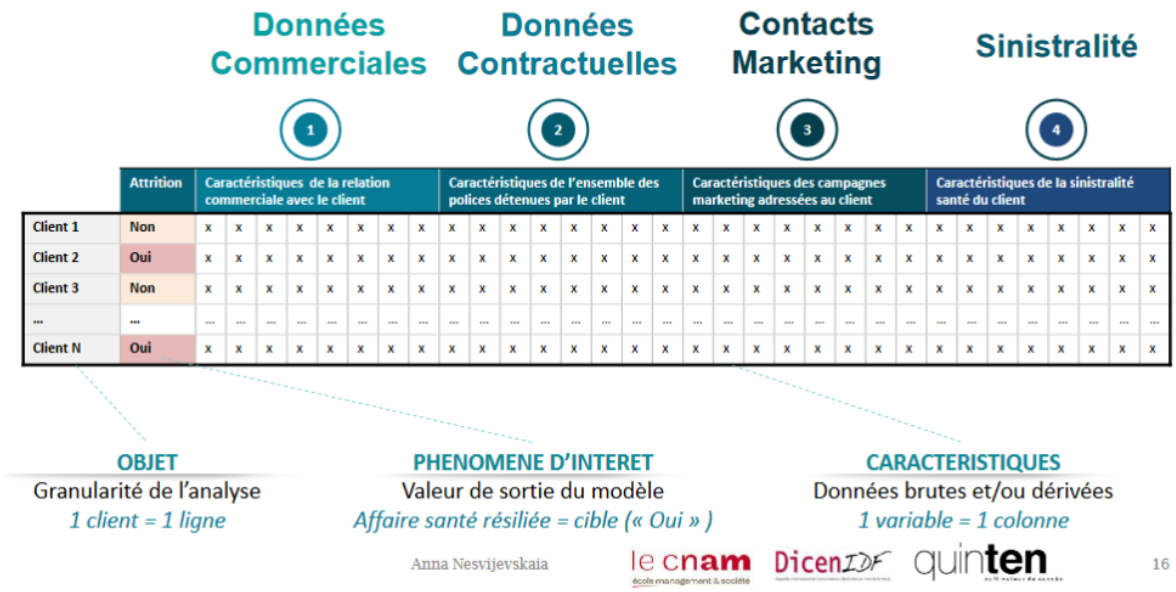


Figure 55 – Illustration d'une matrice d'apprentissage issue de la structuration des données – *Source : Dunoyer et Nesvijevskaia, Conférence Big Data Open Data, Nancy, 2016*

Le processus de génération d'un algorithme à partir de la matrice d'apprentissage diffère de la nature des méthodes d'analyse choisies. Généralement, plusieurs stratégies de modélisation sont apprises sur un premier sous-ensemble des données (il s'agit de la matrice d'apprentissage à proprement parler), confrontées sur un deuxième sous-ensemble des données pour sélectionner la meilleure stratégie de modélisation (il s'agit alors de la base de validation), puis cette meilleure stratégie de modélisation est évaluée sur un troisième sous-ensemble pour juger de sa pertinence (base de test). Le modèle qui en résulte peut alors être appliqué à une nouvelle observation, par exemple à un nouveau client qui possède ses propres caractéristiques, et donne un résultat avec un certain niveau de confiance associé (voir Figure 56).

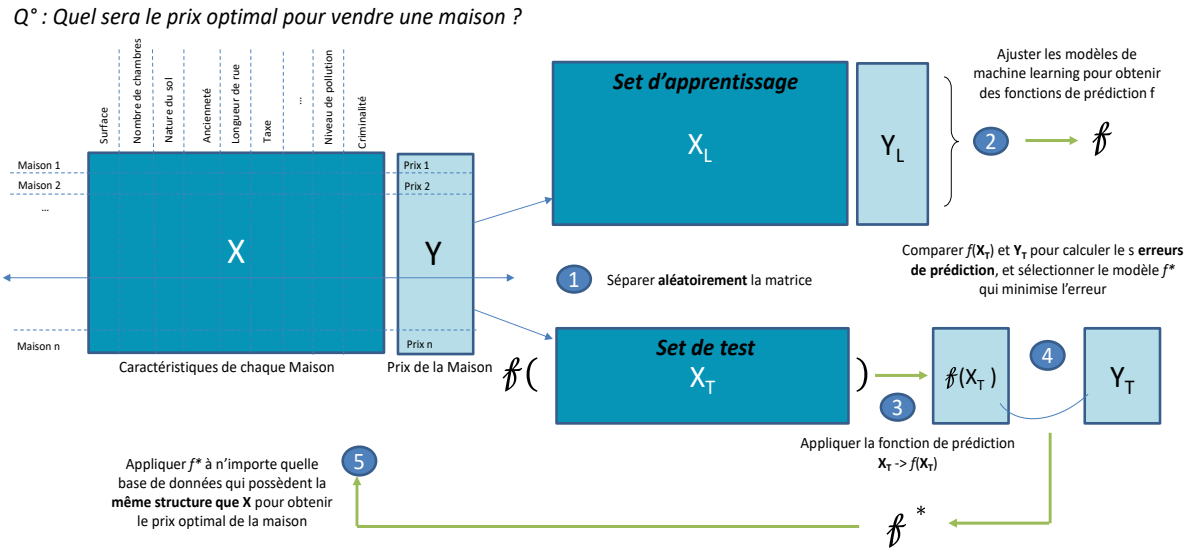


Figure 56 – Illustration du processus d'apprentissage algorithmique – *Source : Dunoyer et Nesvijevskaia, Conférence Big Data Open Data, Nancy, 2016*

Lorsque cela est nécessaire pour l'usage, la génération du meilleur modèle ou son ajustement peuvent être automatisés : il s'agit alors d'un modèle auto-apprenant. Un modèle évalué comme suffisamment pertinent pour un usage donné peut donner lieu à un déploiement, c'est-à-dire la mise à disposition récurrente de l'ensemble de la chaîne de valeur utile à la production du résultat pour l'usage en question, basée potentiellement sur la mobilisation d'acteurs de l'écosystème Big Data et de technologies spécifiques. Ce déploiement comprend alors les différents éléments qui composent le modèle généré par le Data Scientists (structures, données, générateurs de modèles, modèles, résultats des modèles…).

Annexe 5 - Transparence des algorithmes

Anna Nesvijevskaia, Chapitre issu du « Livre Blanc – Être assuré en 2030 » de l'Ecole Polytechnique d'Assurance[52]

Qu'est-ce que la transparence des algorithmes dans le monde de l'assurance ?

La transparence des algorithmes n'a pas à être perçue comme une contrainte imposée par le régulateur : c'est une incitation à la coopération. En effet, les algorithmes ne sont pas des machines, mais le fruit d'une relation entre acteurs dont la forme de collaboration conditionne la valeur de leur usage.

La pratique de l'assurance en 2030 sera résolument truffée d'algorithmes ! Les usages imaginés, et d'ores et déjà en cours de déploiement, sont innombrables : de la création de nouveaux produits basés sur l'IOT jusqu'à la détection des fraudeurs, en passant par la prospection qualifiée et optimisée tous canaux confondus ou l'accélération de la gestion des sinistres, sans oublier bien évidemment la tarification et la protection de portefeuille par l'anticipation des résiliations. L'intelligence artificielle s'empare progressivement des champs de prise de décision humaine, et ce au cœur des activités clés des institutions d'assurance. Les équipes de Data Scientists prennent de l'ampleur auprès des équipes historiques des actuaires et du marketing. Ainsi, lorsque la CNIL soulève la réflexion sur les enjeux éthiques et sociétaux liés à l'usage des algorithmes[53], tout en posant les principes de loyauté et de vigilance, le secteur de l'assurance est l'un des premiers intéressés.

Si la maturité des citoyens face au concept d'« algorithme » est à la traine, peu d'entreprises semblent préparées aussi bien que les assureurs aux questions relatives au traitement des données, qui irrigue historiquement l'analyse des risques. Pourtant, la maturité des acteurs internes parait, elle aussi, perfectible. Cela se traduit par un certain malaise, des attentes quelquefois fantasmatiques, des incompréhensions, des réticences et parfois un désengagement des praticiens métier des projets data internes, et même de mauvaises pratiques. Le résultat ne se fait pas attendre : les projets s'enlisent dans des explorations qui n'aboutissent ni à des usages

[52] « Livre Blanc : Être assuré en 2030 ! » 2018. http://www.epassurances.fr/livreblanc/livre-blanc-etre-assure-en-2030-epa

[53] https://www.cnil.fr/fr/comment-permettre-lhomme-de-garder-la-main-rapport-sur-les-enjeux-ethiques-des-algorithmes-et-de

réalistes ni à des bénéfices tangibles, voire conduisent à des dérives. Ces dernières ne tardent pas à se faire remarquer par les citoyens et les régulateurs, comme les pratiques agressives ou discriminantes de certains assureurs britanniques[54]. A ce jour, de nombreux assureurs français ont investi dans des projets data, ne serait-ce que pour s'offrir leur propre montée en maturité, et il est temps de faire le point sur le concept d'« algorithme ». Ici, il n'est pas question d'un outil incontournable et révolutionnaire dans la pratique métier, ni d'une menace américaine ou chinoise. Il est question du rôle offert aux praticiens français dans la construction de cet inexorable objet technique pour l'assurance de demain.

1. Un algorithme n'est pas un robot.

Qu'est-ce qu'un algorithme ? Il s'agit, d'après la définition simple de la CNIL, d'une « une suite finie et non ambiguë d'instructions permettant d'aboutir à un résultat à partir de données fournies en entrée ». Trois questions sont alors posées pour mieux le comprendre... Quelles sont les données source ? Quelles sont les instructions ? Quel est le résultat obtenu et à quoi sert-il ? Jusqu'ici, tout est plutôt simple. Cela semble se compliquer lorsque le phénomène Big Data impacte de front les trois réponses, mais qu'en est-il vraiment ?

La baisse des coûts de traitement, la mise en données massive et la facilitation des accès enrichissent les référentiels de données habituels. La prise en compte de sources inexploitées auparavant, internes ou externes, conduit à repenser le périmètre d'analyse d'un risque ou d'une performance. Ces sources contiennent potentiellement des données personnelles, des variables discriminantes, sans oublier des données de mauvaise qualité pouvant biaiser les résultats. Les nouvelles données nécessitent alors une qualification, qu'elle soit technique, analytique, ou simplement métier. Par ailleurs, il faut construire, nettoyer, agréger, dériver, enrichir, malaxer les données source pour en extraire la signification, à commencer par les données internes, souvent issues de plusieurs silos et fonctions de l'entreprise. Des praticiens, capables de comprendre le sens des données, sont ainsi impliqués dans la sélection des données dès le démarrage d'un projet data, sans attendre que les Data Scientists n'aient déniché un signal, statistiquement fort intéressant, mais incompréhensible ou éthiquement inexploitable. Ce travail conjoint assure la qualité de la matière première pour alimenter l'algorithme. Ainsi, la

[54] https://www.cnet.com/roadshow/news/uk-insurance-company-charges-more-based-on-email-domain-first-name/

transparence les données en entrée est, de toute évidence, à la main des concepteurs des algorithmes, c'est-à-dire les Data Scientists et praticiens métier.

La suite d'instructions de l'algorithme fait, elle aussi, l'objet d'une évolution de fond. En effet, les instructions sont traditionnellement paramétrées par des individus, sous forme de règles métier notamment. Le Machine Learning permet, quant à lui, de générer un paramétrage optimisé, selon un processus d'apprentissage automatisé, basé sur un ensemble de données dédié, distinct des données sur lequel l'algorithme résultant sera appliqué. Le rôle de la machine suscite alors de la fascination ou de la méfiance. Démystifions ce sujet : jusqu'à nouvel ordre, si la machine peut bien fournir des instructions optimisées, elle nécessite elle-même un paramétrage humain pour se lancer dans les apprentissages. Les bibliothèques disponibles, y compris en Open Source, pour la génération d'algorithmes sont aujourd'hui d'une richesse sans précédent et de plus en plus accessibles, au point où ce n'est pas l'algorithme qui prime, mais son choix parmi les algorithmes possibles. Les algorithmes de Machine Learning s'étendent des plus simples aux plus complexes, parfois à tel point que les concepteurs ne peuvent pas identifier eux-mêmes le cheminement logique des instructions générées, et encore moins les piloter dans le temps. La volonté de transparence des algorithmes peut amener à privilégier les méthodes explicites, en particulier lorsque celles-ci ne dégradent pas la précision souhaitée des résultats. Ce choix, associé à un discours pédagogique sur les méthodes de Machine Learning, garantit alors l'absence de l'effet « boite noire » auprès des utilisateurs internes des résultats, des régulateurs et des clients.

Enfin, qu'en est-il de l'usage du résultat de l'algorithme ? Le résultat doit être une information utile à une prise de décision, automatisable ou non. Il s'agit certainement de l'élément le plus important à prendre en compte dans la conception d'un algorithme, puisqu'il qu'il guide le choix des données et des méthodes de Machine Learning utilisées, et surtout la création de valeur par le nouvel usage proposée. Or, l'éventail des usages possibles ne cesse de s'agrandir, allant des pratiques les plus saines aux plus déloyales, délibérés ou inconscientes. Plus alarmant encore, lorsque les praticiens se reposent sur une fouille des données sans a priori pour découvrir de nouvelles hypothèses métier, il est tout à fait légitime de se poser la question de la maturité et de l'éthique des interprètes de ces hypothèses pour l'imagination des usages imprévus. Une double implication des acteurs métier est alors clé dans la phase de conception des algorithmes : en amont, un engagement actif dans le choix des objectifs et des usages, et, en aval, une appropriation approfondie des résultats pour décider en toute connaissance de cause

s'ils sont acceptables pour l'usage. Sans surprises, la loyauté des algorithmes est avant tout une loyauté des concepteurs des usages mobilisant ces algorithmes.

Ainsi, la conception d'un algorithme, jalonnée de décisions humaines, est très loin d'être le fruit d'un robot.

2. La valeur est dans la co-construction

La CNIL demande à douter des algorithmes. Dans ces circonstances, douter ne consiste pas à juger un résultat statistique, mais à s'impliquer dès l'amont du projet dans les choix structurants des données, des méthodes de génération d'algorithmes, et des usages. Douter implique d'exiger la documentation des technologies et des méthodes analytiques, la capitalisation et le partage des connaissances ainsi que la rigueur éthique de chaque acteur impliqué dans la conception, marquée par des étapes d'arbitrage. Or, les experts métier ne peuvent arbitrer qu'à condition d'être informés de façon intelligible par les experts des données, eux-mêmes informés des risques éthiques, généraux ou propres au secteur de l'assurance, afin de pouvoir tirer l'alerte face aux dérives possibles des usages. Les Data Scientists doivent ainsi travailler en étroite collaboration avec les experts métier, c'est-à-dire les directions générales, les actuaires, les responsables marketing, les financiers, les juristes, les Knowledge Managers, et tout autre contributeur désireux de façonner la conception de l'algorithme ou impacté par son usage.

Mais alors, comment délimiter les responsabilités des concepteurs et des utilisateurs ? L'intérêt d'un algorithme est la transformation de données brutes en informations utiles à la réduction des incertitudes face à une prise de décision. Si aujourd'hui la responsabilité est portée par un décisionnaire, demain elle sera séparée en deux : d'une part, le décisionnaire, c'est-à-dire l'utilisateur de l'information fournie par l'algorithme, et d'autre part le concepteur de l'algorithme. Prenons un exemple simplifié de la prévention de l'attrition client selon un score de churn, où un algorithme fournit une liste de clients à risque élevé de résiliation de contrat : un agent ne portera plus que la responsabilité opérationnelle de la rétention des clients ciblés selon la liste restreinte, alors que les concepteurs de l'algorithme porteront celle des critères de ciblage. Le partage et la concentration des responsabilités ne paraissent acceptables qu'à condition que l'autonomie de décision des utilisateurs soit garantie par une symétrie informationnelle avec les concepteurs. Sans cette autonomie, les utilisateurs verraient leur capital de connaissances stagner, voire fondre, leur périmètre d'action se réduire, et leur responsabilité individuelle croitre. L'application du principe de loyauté semble donc devoir

franchir d'abord le pas des utilisateurs internes, citoyens vigilants comme les autres, avant d'être pleinement opérant auprès des clients et de la communauté.

***

La transparence des algorithmes ne doit donc pas être considérée seulement comme une requête complémentaire du régulateur. Elle conditionne avant tout la création de valeur par la pertinence économique et démocratique des usages, et la capitalisation des connaissances métier internes. Elle ne peut avoir lieu qu'à travers l'engagement pédagogique des Data Scientists et la prise de conscience des concepteurs métier et des utilisateurs internes, pour accompagner pas à pas la monté en maturité du secteur de l'assurance d'ici 2030 sur le concept, pas si mystérieux, d'« algorithme ».

Anna Nesvijevskaia

Directrice Associée chez Quinten

Annexe 6 - Présentation du rapport préliminaire

Extraits de la présentation du rapport préliminaire de cette thèse, intitulé « La controverse épistémologique Big Data face à la réalité de l'appropriation de nouveaux paramètres par les acteurs métier en entreprise », présenté en colloque international *Big Data, Open Data : quelles valeurs, quels enjeux ?* à Rabat en 2015 et ayant donné lieu à la publication d'actes de colloque sous forme d'ouvrage collectif (Chartron & Broudoux, 2015) : illustrations des 3 cas préliminaires de ces travaux de thèse.

Cas A : Dispositif télématique « Urgence » :

Appréhension de nouvelles observations

- Adaptation des acteurs métier au nouveau dispositif
  - Réorganisation de l'espace de travail (bureaux dédiés disposés par ordre de prise d'appel et par langue)
  - La priorisation est facilitée par les outils, mais reste à la main des chargés d'assistance
  - Mise en place et formation aux nouveaux outils et procédures

« Notificateur » partagé

Outils individuels

Appréhension de nouvelles observations

- Développement de compétences métier nouvelles pour l'analyse d'observations nouvelles (perceptions)
  - Auditives (« silent call », environnement du véhicule,…)
  - Géo-spatiales (géolocalisation)

Ecran de travail de l'outil « Urgence » (tableau de bord)

Cas B : Cancer du sein triple négatif :

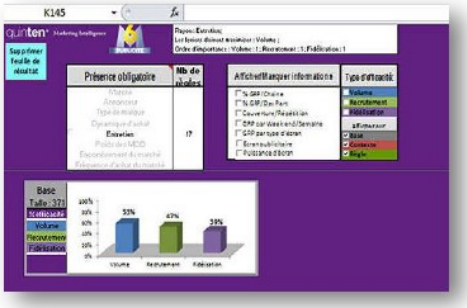
Description du dispositif étudié

Réponse pathologique	N= 47 patientes
Classification de Chevallier	
Réponse complète (classe 1 et 2)	22 (46.8%)
Réponse incomplète (classe 3 et 4)	25 (53.2%)
Variables pour chaque profil - Informations sur les patientes - Informations sur la tumeur - Modalités d'admission du traitement - Marqueurs biologiques - Marqueurs génétiques - Etc.	Env. 250

Cas C : Placement publicitaire :

Déploiement : automatisation et appropriation

- Déploiement de produit de conseil basé sur un argumentaire statistique **inductif** pour l'optimisation d'espace publicitaire (sans recherche de causalité ni nécessité d'intuition métier)



- Enrichissement progressif de contextes par de nouvelles données
- Apparition de la nécessité de formation des utilisateurs finaux

Annexe 7 - Grille d'analyse des études de cas selon une approche quantitative

La grille d'analyse a fait l'objet d'une construction itérative au fur et à mesure de l'avancement de ces travaux de recherche.

Tout d'abord, elle correspondait à un outil assez naturel de prise de notes, c'est-à-dire de collecte d'observations primaires avant leur transcription sous forme de compte rendu détaillé. Dans ce sens, chaque session de prise de notes correspondait à une activité nouvelle qu'il était nécessaire de décrire, ce qui a induit une structure chronologique et la notion d'ordre des activités. Plus particulièrement, la description des activités de chaque intervenant sur les projets, surtout pour des projets mobilisant un nombre important d'intervenants (jusqu'à 8 Data Scientists dans l'équipe projet, et jusqu'à 15 contributeurs métier) nécessitait une séparation visuelle des observations dans le cadre de l'identification des compétences nécessaires à la réalisation des projets, et des interactions existantes entre les acteurs. Une interaction est alors représentée en termes de participation simultanée de plusieurs individus à la réalisation d'une activité. Ensuite, l'analyse comparée des cas a permis de faire émerger progressivement les trois dimensions d'enrichissement du modèle : la qualité des données, les indicateurs de valeur, et la médiation. Si cette dernière était partiellement couverte par la visualisation de la richesse des interactions, les deux autres étaient jusque-là « noyées » dans la prise de notes et les comptes rendus.

La collecte d'observations primaires a été réalisée en ajoutant ces deux dimensions lors de la réalisation du dernier cas commencé (prévention santé prévoyance). Ainsi, chaque activité est ordonnée, détaillée en termes de tâches des acteurs mobilisés en interaction, donne lieu à un résultat tangible, et peut avoir des impacts sur la qualité des données ou les indicateurs de valeur (voir Figure 57 ou document complet disponible en ligne en suivant le lien suivant : https://drive.google.com/drive/folders/1T1-IIM9-decM_wpOatBTvZOswg8EkN-h?usp=sharing). La grille organisée ainsi permettait par ailleurs d'associer facilement le déroulé de l'étude de cas et les illustrations de ce déroulé.

Ordre de l'activité	Catégorie de l'activité (phase)	Tâches accomplies par chaque interlocuteur						
		Direction Générale	Responsable Observatoire des branche	Responsable Medecin-Conseil	Chef de projet - Centre de Solution Décisionnel	Expert métier assurance / strat data / PMO	Ingénieur Data	Machine learner
1	Compréhension métier	x	Expression des objectifs stratégiques dans un Appel d'Offre, réunion de présentation, et priorisation des pistes d'analyse	Consultation pour l'expression des besoins	Contribution à l'expression du besoin et coordination des acteurs impliqués	Recueil des besoins et animation des échanges pour la priorisation des pistes d'analyse	x	x
2	Compréhension métier	x	x	x	x	Elaboration d'une méthode pour réaliser une étude de cas répondant aux objectifs prioritaires : traduction des besoins aux machine learners, animation du choix entre les méthodes d'analyse proposées, puis abribrage entre les méthodes selon la cohérence avec les besoins	x	Proposition de méthodes analytiques par 4 machine learners avec des spécialisations différentes, et description des résultats possibles selon la stratégie d'analyse choisie
3	Compréhension métier	x	Ajustement et validation de la pertinence des leviers opérationnels	x	x	Imagination de leviers opérationnels possibles pour illustrer les résultats possibles de l'étude de cas	x	x

Résultats	Impact indicateurs de valeur	Impact Qualité des données
Priorisation des objectifs macro du projet	A ce stade, les experts métier et data n'ont pas de visibilité sur les indicateurs de performance des résultats, ni sur le périmètre d'analyse et les ordres de grandeur. Une estimation est émise : l'analyse portera sur 500 000 ouvrants-droit, à la croisée du portefeuille santé et du portefeuille prévoyance, et contiendra un historique de consommation santé long, lié à la réalité médicale du sujet (traitements longue durée et évolution possible des pathologies)	Les données santé et prévoyance n'ont jamais été croisées. Aucune analyse en termes de qualité de données ne permet d'affirmer qu'il est possible de le faire. En revanche, les portefeuilles sont connus comme volumineux et ne peuvent être traités avec des outils traditionnels : une plateforme "Big Data" a été mise en place pour effectuer les analyses. Le projet doit permettre de tester cette nouvelle plateforme.
Choix de la méthode d'analyse NMF (non negative matrix factorisation) pour établir des typologies de consommation santé représentatifs et explicites	La méthode choisie devra donner lieu à l'émergence de typologies de consommation : leur nombre est inconnu, leur nature aussi, et donc leur caractère opérationnel (possibilité de mettre en œuvre des leviers opérationnels) est à valider selon des critères inconnus à ce stade. Ainsi, le potentiel de création de valeur à travers le levier de la prévention est inconnu. En revanche, la méthode vise une création de connaissance exhaustive des grandes familles de consommation : le critère d'évaluation choisi est l'interprétabilité.	x
Un exemple fictif de levier opérationnel est proposé en illustration des résultats attendus	L'ajustement de l'illustration du levier opérationnel se base sur la priorisation d'un indicateur par l'expert métier : il faut viser des comportements santé couteux. L'indicateur de prix est inscrit comme axe de sélection pour la suite des analyses.	La priorisation des indicateurs opérationnels se traduit par une priorisation anticipée des données à mettre en qualité (prix et indicateurs financiers issus de la base des remboursements)

Figure 57 – Grille d'analyse détaillée : test sur le cas « Prévention santé prévoyance », extrait des 3 premières lignes

Dimensions : Ordre de l'activité, Catégorie de l'activité, Tâches accomplies par chaque interlocuteur (6 interlocuteurs client et 4 interlocuteurs prestataire), Résultat de l'activité, Impact Qualité des données, et Impact Indicateurs de valeur.

A la lumière des derniers résultats de la modélisation de processus, cette grille s'avère limitée par la notion d'ordre et une prise en compte incomplète des éléments de médiation. En effet, la notion d'ordre des activités a été choisie pour mieux percevoir les séquences et les itérations du modèle CRISP_DM. Or, l'observation du terrain montre que les activités peuvent être superposées, cette grille n'est donc pas appropriée pour rendre compte de ces superpositions. Il est donc nécessaire de l'ajuster en remplaçant la colonne de l'ordre de l'activité par deux colonnes de dates (Date de début de l'activité et Date de fin de l'activité), ce qui facilite d'autant plus la prise de notes tout en gardant la notion d'ordre selon la chronologie des dates de démarrage. En ce qui concerne la médiation, elle peut être complétée par les éléments clés identifiés au cours de ces travaux, comme le capital de connaissances (modalités d'échange de connaissances, outils de capitalisation…), les algorithmes (traitements analytiques permettant la transformation des données en information), les représentations sociales (interfaces utilisées, convention de représentation des indicateurs…) et la gestion de projet (modalités d'organisation des ressources projet). Ainsi, la grille optimale (voir Figure 58, utile pour mener une analyse comparative quantitative sur les projets data (notamment hors Quinten), se présente selon un modèle de grille d'analyse ajusté, permettant de recueillir les observations primaires sous la forme de description d'activités d'un projet data au fur et à mesure de son avancement, chaque activité étant détaillée selon les dimensions suivantes (colonnes), permettant ainsi d'établir des indicateurs quantitatifs qui pourraient confirmer ou informer le modèle proposé, ou de le faire évoluer, notamment en mettant en évidence le poids et la nature de l'impact des facteurs de succès (voir Figure 58).

<table>
<tr><th>Concept clé</th><th colspan="2">Données brutes à collecter</th><th>Indicateurs à produire</th></tr>
<tr><td rowspan="7">Processus</td><td colspan="2">Date de début de l'activité</td><td rowspan="2">Ordre des activités</td></tr>
<tr><td colspan="2">Date de fin de l'activité</td></tr>
<tr><td colspan="2">Catégorie de l'activité</td><td>Superposition, chemin critique…</td></tr>
<tr><td rowspan="4">Tâches accomplies par chaque interlocuteur (autant de dimensions que de rôles)</td><td>Rôle 1</td><td rowspan="4">Niveau de complexité des interactions (fréquence par rôle, ampleur en nombre de rôles mobilisés, rôles centraux) …</td></tr>
<tr><td>Rôle 2</td></tr>
<tr><td>Rôle 3</td></tr>
<tr><td>Rôle 4</td></tr>
<tr><td colspan="2">Résultat de l'activité</td><td>Nombre de versions, qualité…</td></tr>
<tr><td rowspan="4">Médiation Homme-Données</td><td colspan="2">Capitalisation de connaissances</td><td rowspan="6">Importance de chaque facteur à mettre en perspective avec la qualité des résultats de chaque activité</td></tr>
<tr><td colspan="2">Algorithmes</td></tr>
<tr><td colspan="2">Représentations sociales</td></tr>
<tr><td colspan="2">Gestion de projet</td></tr>
<tr><td>Qualité des données</td><td colspan="2">Impact et documentation</td></tr>
<tr><td>Indicateurs de valeur</td><td colspan="2">Impact et mode de valorisation</td></tr>
</table>

Figure 58 – Grille d'analyse détaillée pour étude quantitative de cas multiples

Annexe 8 - Modèles Input-Process-Output détaillés

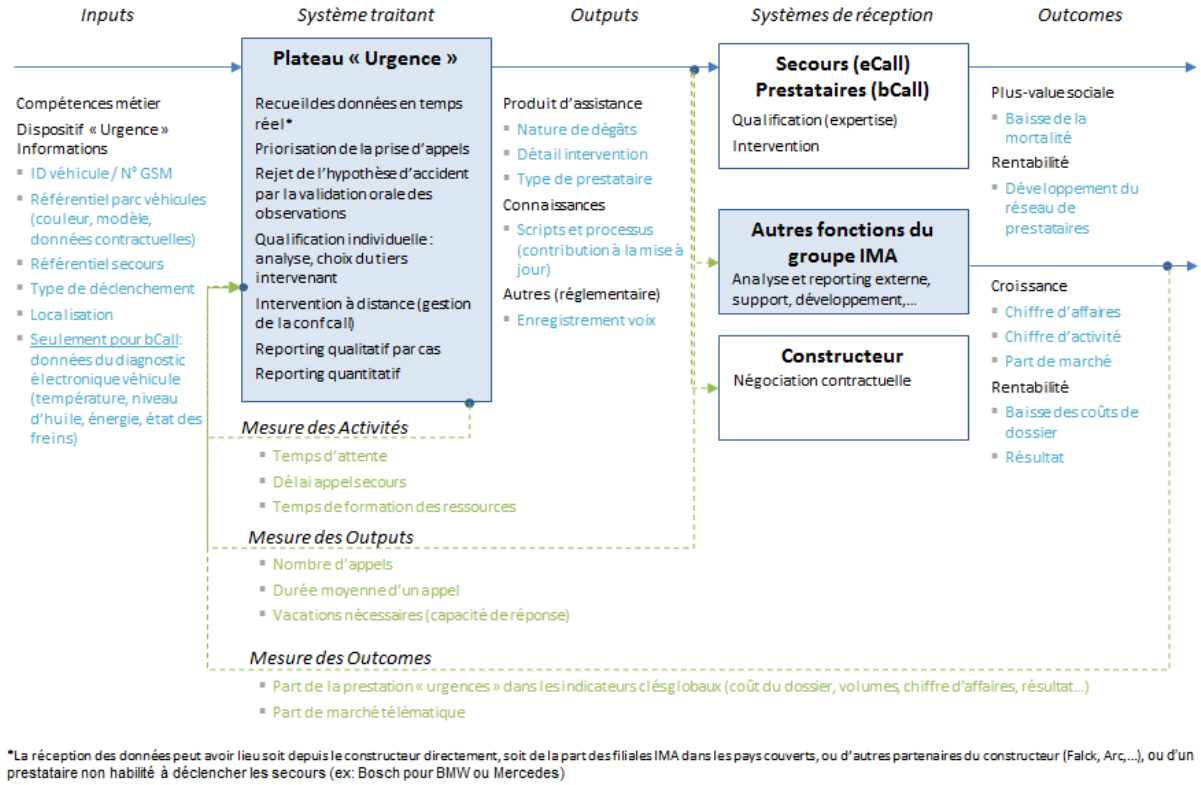


Figure 59 – Modélisation IPO : Impacts du projet « Dispositif Télématique Urgences »

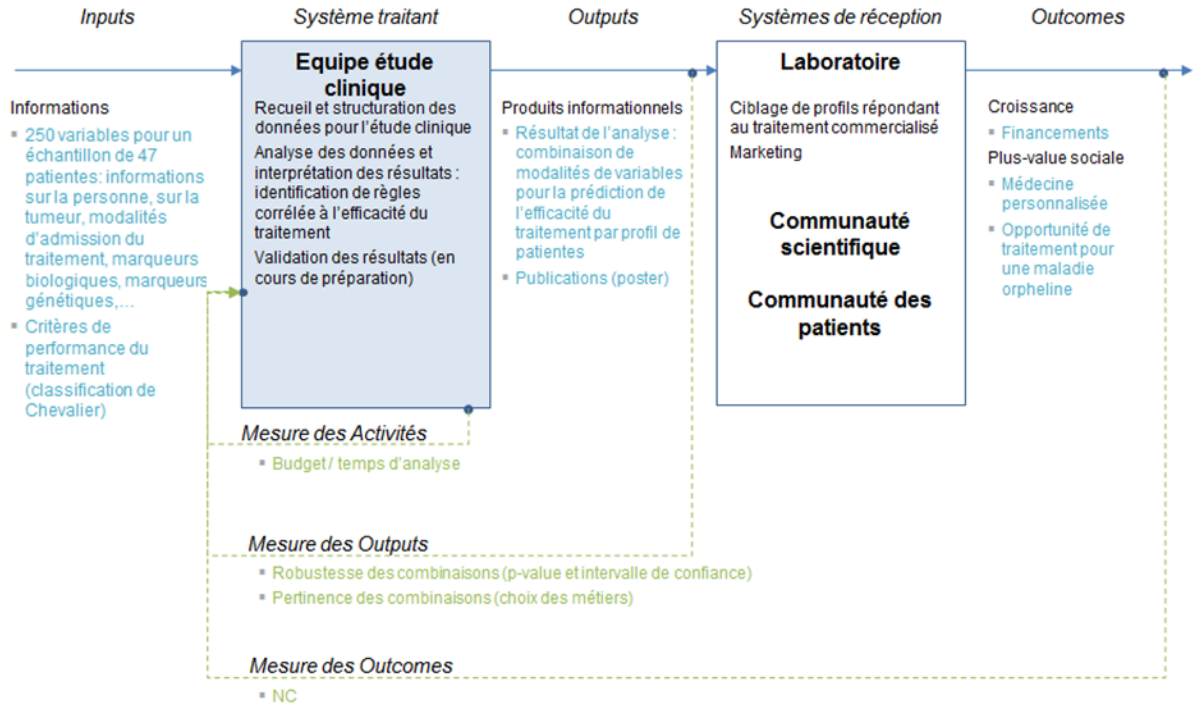


Figure 60 – Modélisation IPO : Impacts du projet « Cancer du sein triple négatif »

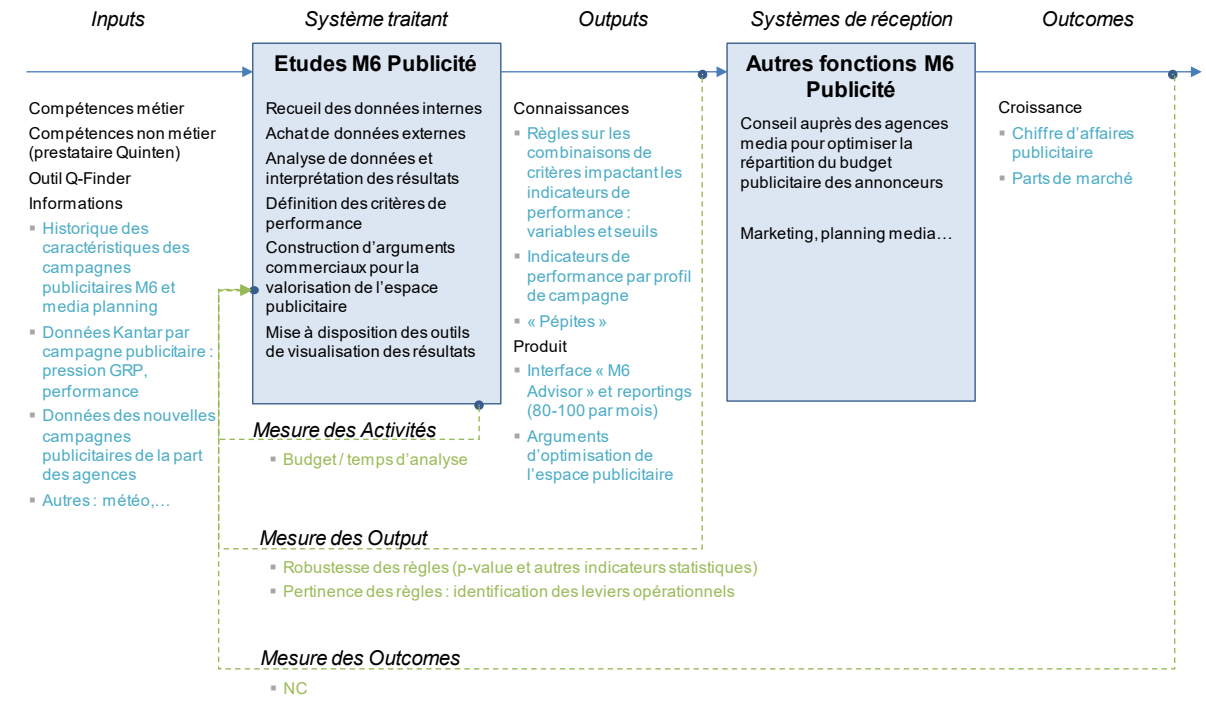


Figure 61 – Modélisation IPO : Impacts du projet « Placement publicitaire »

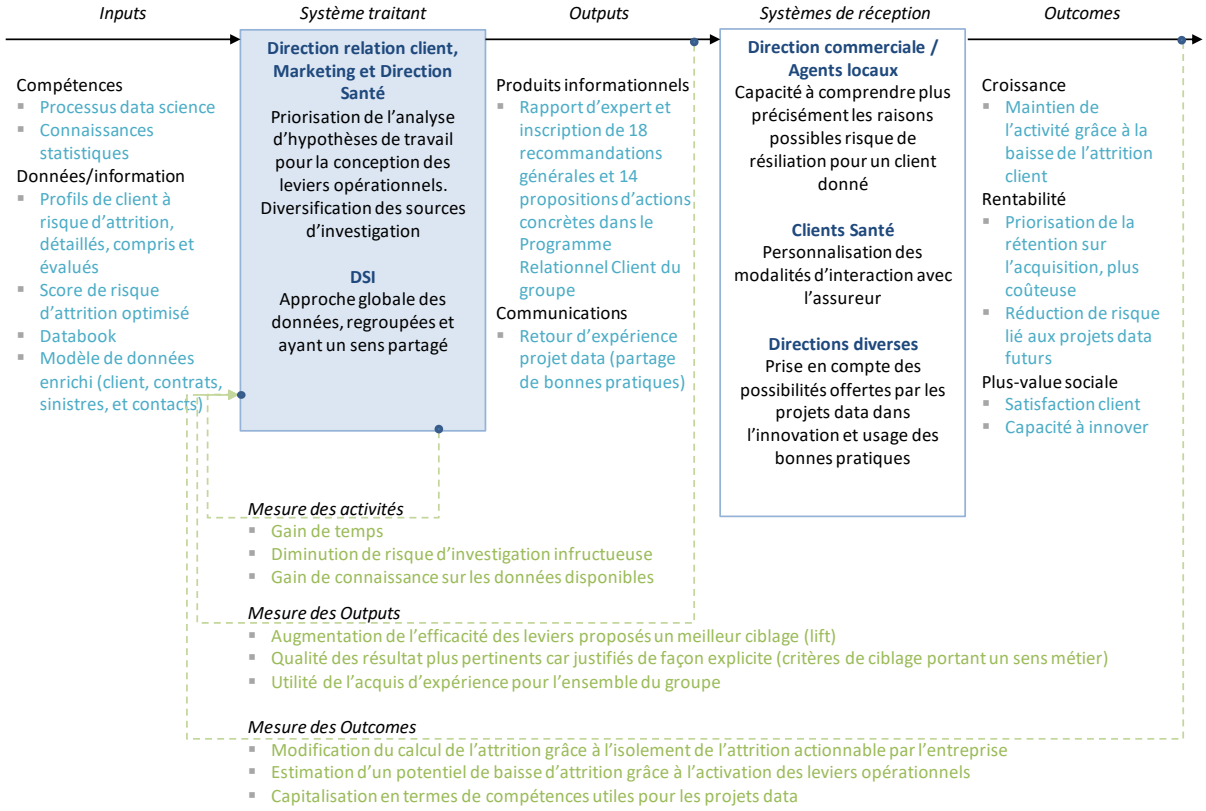


Figure 62 – Modélisation IPO : Impacts du projet « Attrition en Assurance Santé »

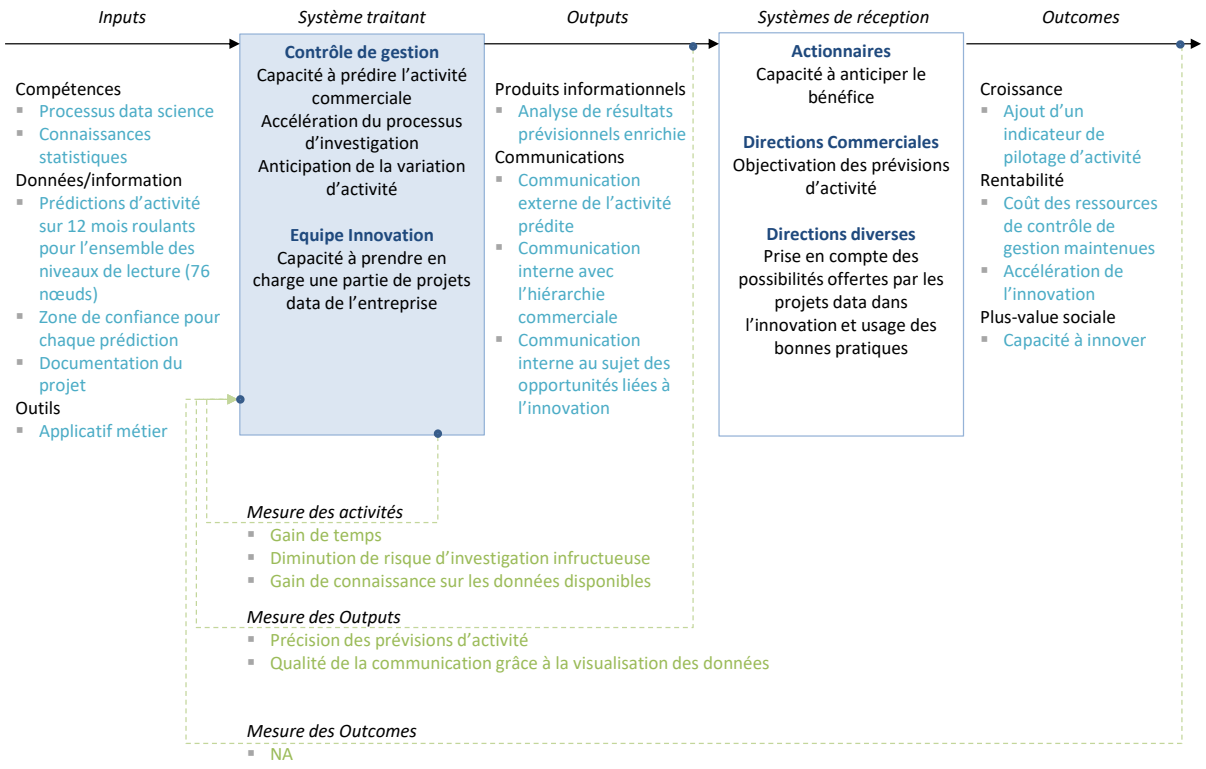


Figure 63 – Modélisation IPO : Impacts du projet « Prédiction d'activité »

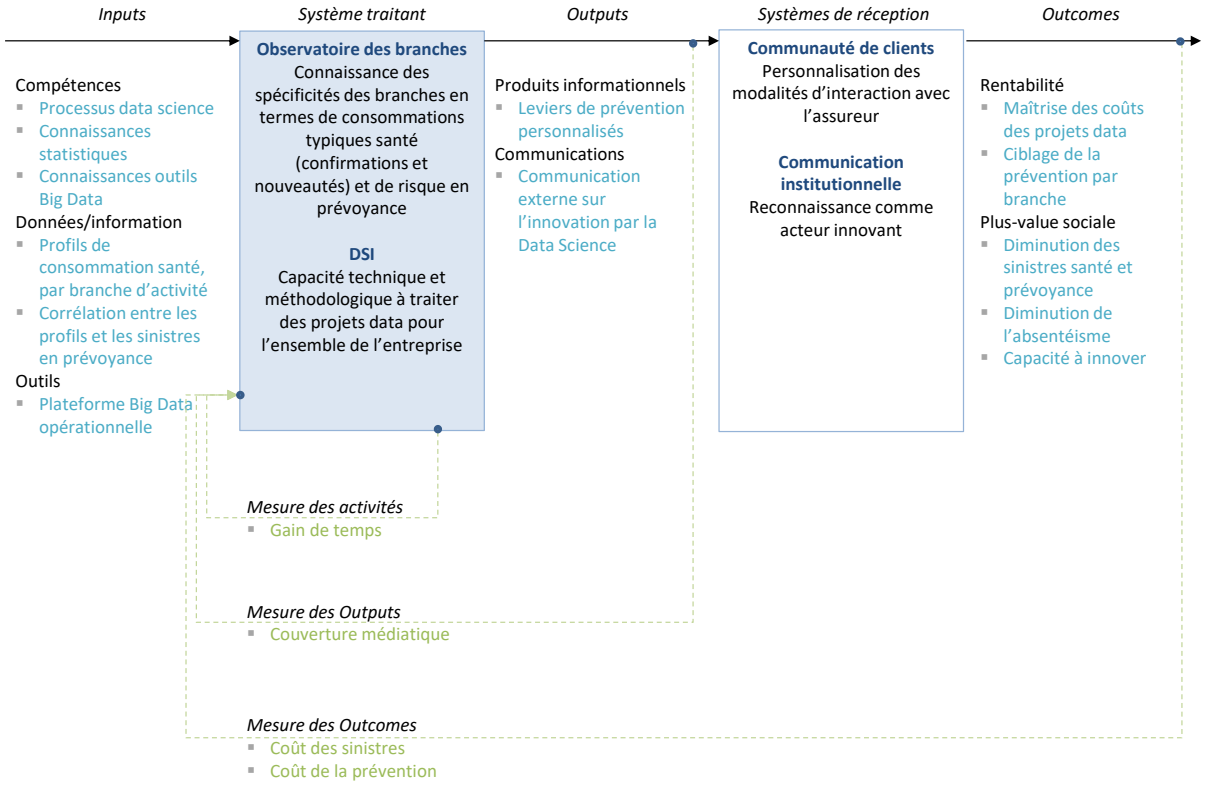


Figure 64 – Modélisation IPO : Impacts du projet « Prévention Santé Prévoyance »

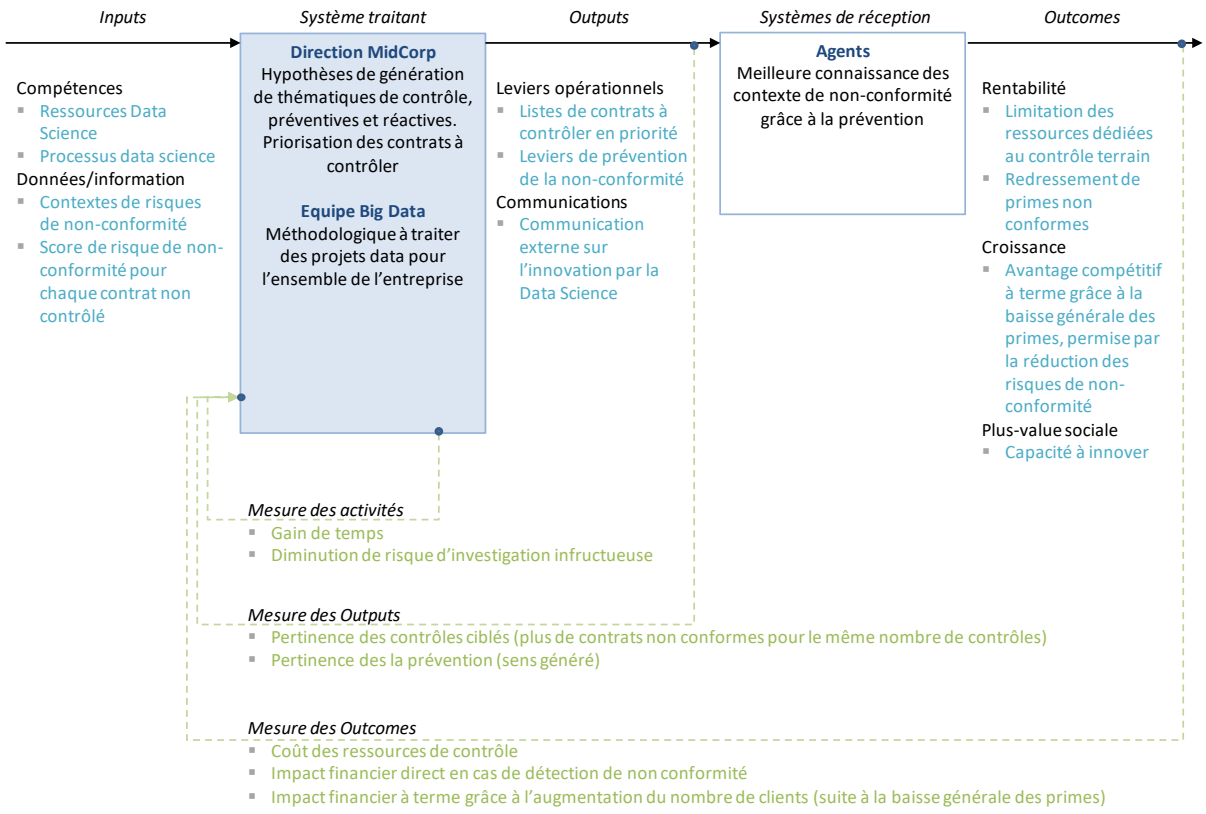


Figure 65 – Modélisation IPO : Impacts du projet « Contrôles de non-conformité »

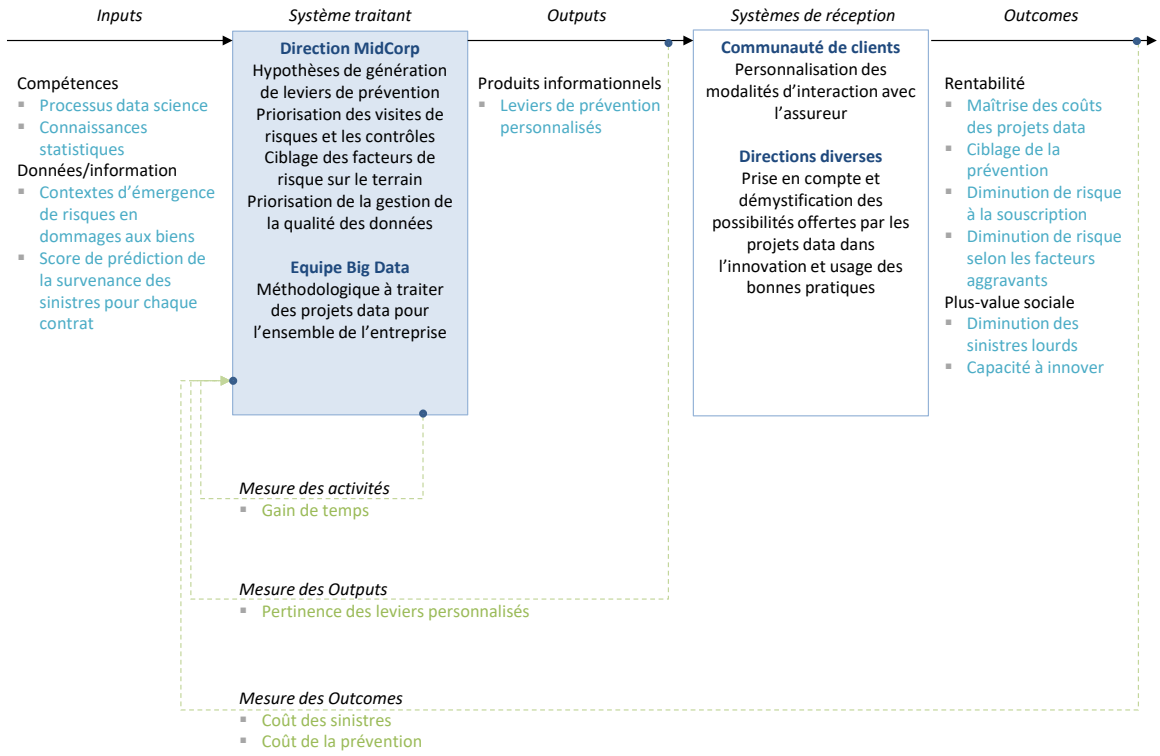


Figure 66 – Modélisation IPO : Impacts du projet « Sinistres lourds en dommage aux biens »

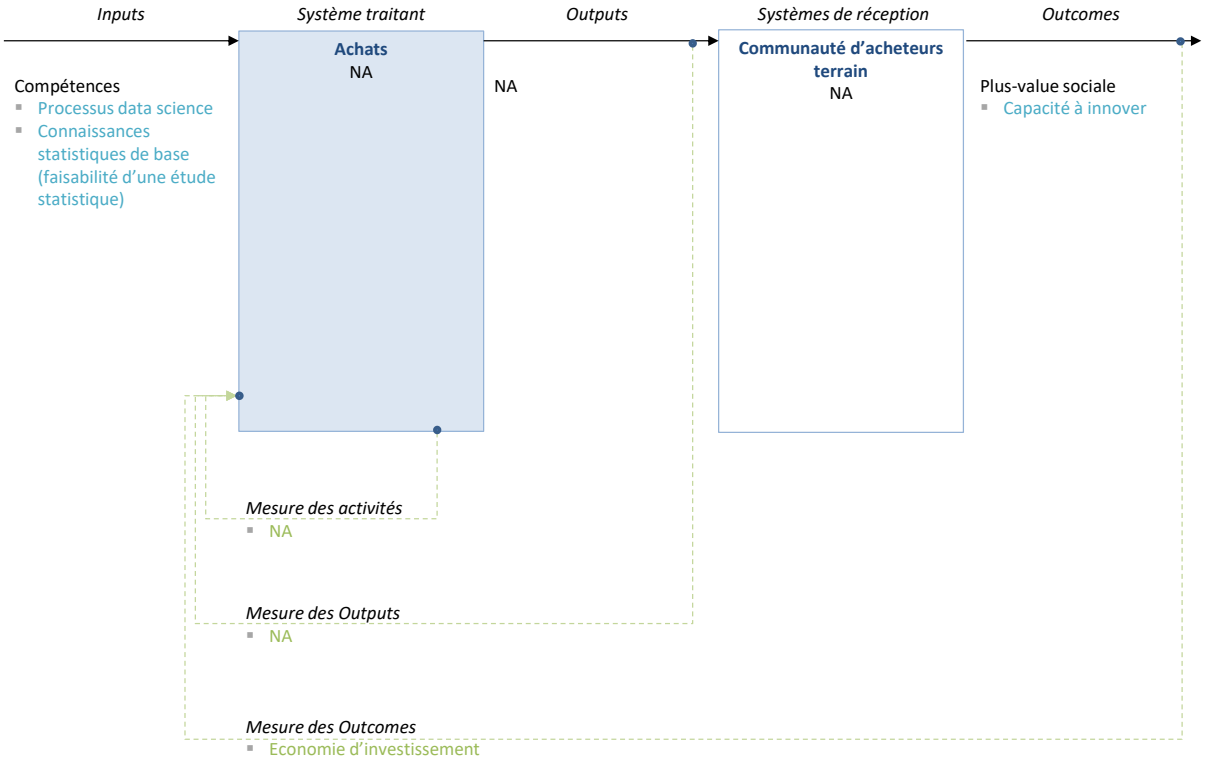


Figure 67 – Modélisation IPO : Impacts du projet « Prédiction des prix des agrumes »

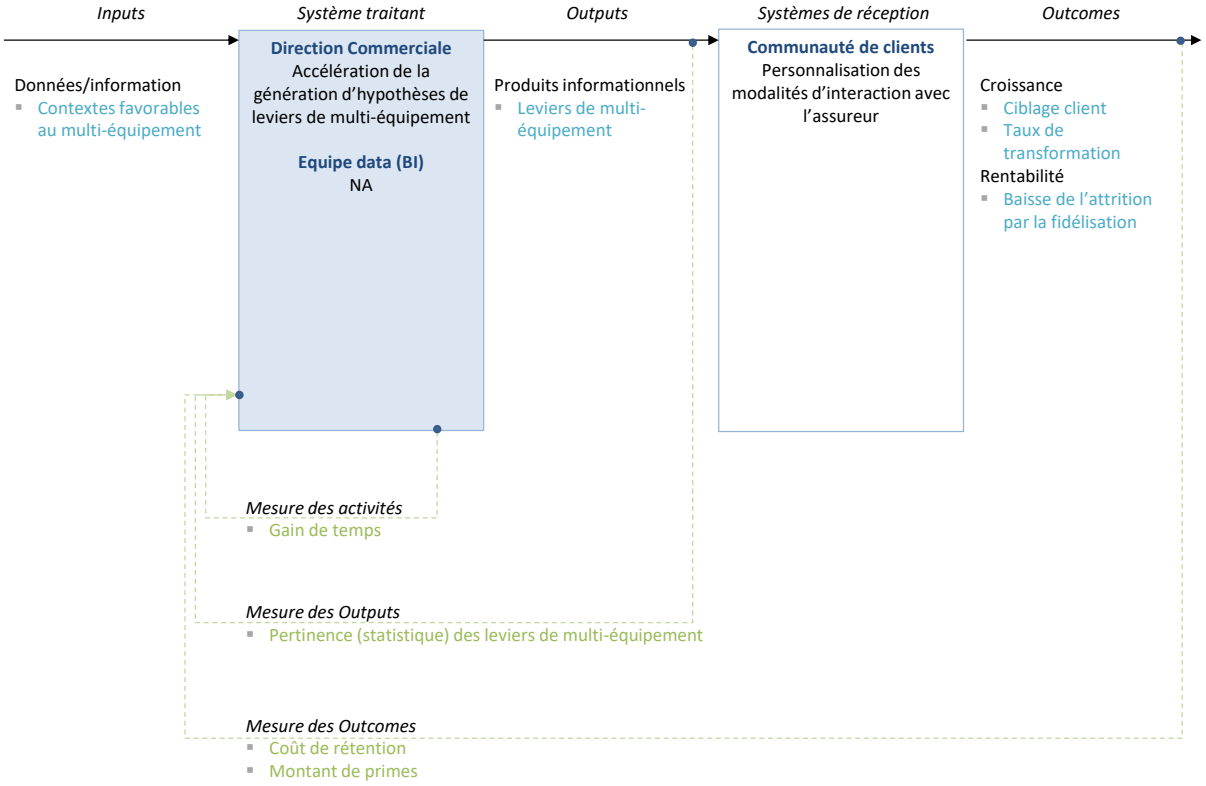


Figure 68 – Modélisation IPO : Impacts du projet « Multi-équipement »

Annexe 9 - L'internalisation des usages dérivant du recours à l'Intelligence Artificielle dans les entreprises

Extrait de l'intervention : Anna Nesvijevskaia, salon Big Data, mars 2018

Les experts métier se sont emparés du Big Data et de l'Intelligence Artificielle depuis plusieurs années. Après la vague des usages révolutionnaires, des "Preuves de Concept" et des solutions sur l'étagère, le phénomène est en train de s'ancrer dans le quotidien des entreprises. Les réticences des uns reculent tandis que les attentes se font plus réalistes avec la montée en maturité des équipes internes. A présent, les experts métier les plus mûrs veulent reprendre la main et inscrire ces innovations au cœur de l'activité de l'entreprise. Cette internalisation est marquée par des changements organisationnels structurants, par l'implication directe des décideurs métier et IT dès l'amont des projets, et par une valorisation tangible des usages déployés. Comment mieux accompagner cette dynamique ? Comment monter un véritable portefeuille de projets data, générateur d'une valeur en cohérence avec la stratégie de l'entreprise ?

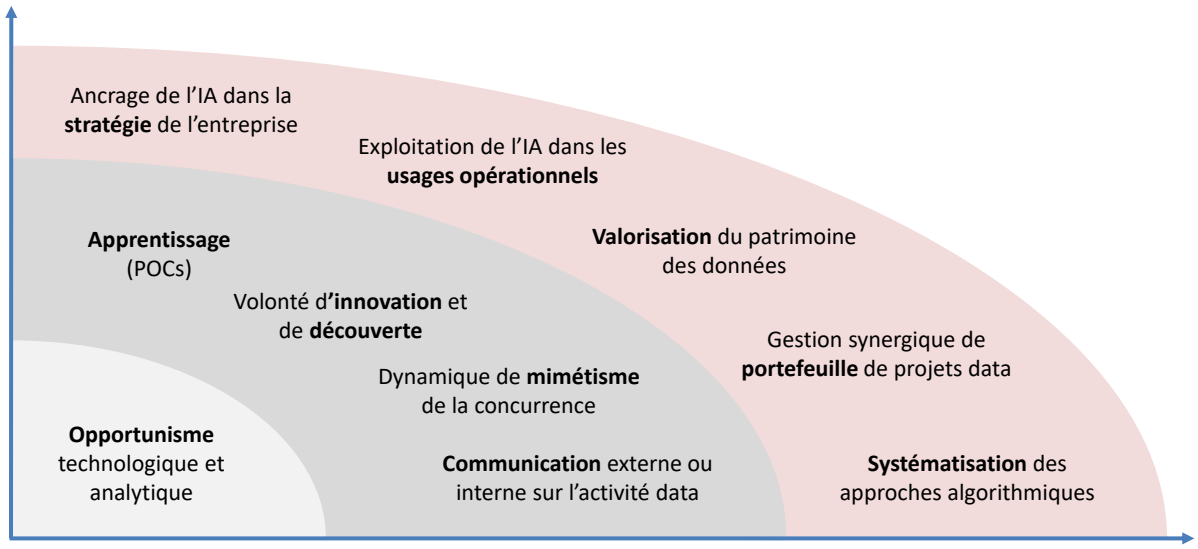


Figure 69 – Les motivations des demandeurs de projets Big Data en 2018

Les motivations des demandeurs (voir Figure 69) évoluent progressivement, au fur et à mesure que leur niveau de maturité face au phénomène Big Data monte. Les premières confrontations avec le phénomène ont souvent fait l'objet de projets opportunistes, notamment grâce à une activité commerciale forte des acteurs de l'Ecosystème, proposant notamment des interventions gratuites et des tests de leurs solutions technologies et analytiques. Puis les motivations évoluent

vers un désir de monter en maturité, de capitaliser des connaissances, de se mettre au niveau des concurrents ou de communiquer sur sa capacité à innover. Il s'agit de motivations visant des usages indirectement liés à la production de résultats analytiques. Enfin, seules quelques entreprises en France commencent à percevoir dans le sujet une véritable orientation sur la génération de valeur directe, en inscrivant les usages des solutions technologies et analytiques au cœur de leurs activités.

La montée en maturité vers des usages directement bénéfiques doit s'appuyer sur la démystification du phénomène Big Data, la clarification de l'offre de l'Ecosystème, et être effectuée en interne sur 6 axes (voir Figure 70) :

1. Une déclinaison de la stratégie de l'entreprise en usages pertinents, et conformes au niveau de maturité de l'entreprise

2. Un choix de solutions orientées sur les utilisateurs, assurant une appropriation opérationnelle optimale

3. Une garantie de l'adoption technologique des outils d'exploration de données, des outils d'exploitation des usages, et les outils de collecte et de stockage

4. Une identification et organisation pertinente des compétences, avec une prise de hauteur sur

    - le niveau d'expérience (capacité à anticiper les risques et à proposer des solutions)

    - les profils (métier, math/stat, projet, technique…)

    - le management (CDO / Directions métier / Directions IT)

    - et l'équilibre entre les ressources internes et externes

5. Une valorisation du patrimoine de données

    - Actif valorisé (Données utiles à la prise de décision et générant un avantage concurrentiel)

    - Actif non valorisé (Données disponibles, perçues comme utiles à la prise de décision mais sous-exploitées)

    - Actif potentiel (Données non disponibles, perçues comme utiles)

6. Une véritable émergence de culture de la donnée, réaliste, éthique, visant une qualité de service pour les clients et les collaborateurs, et dotée d'une gouvernance claire

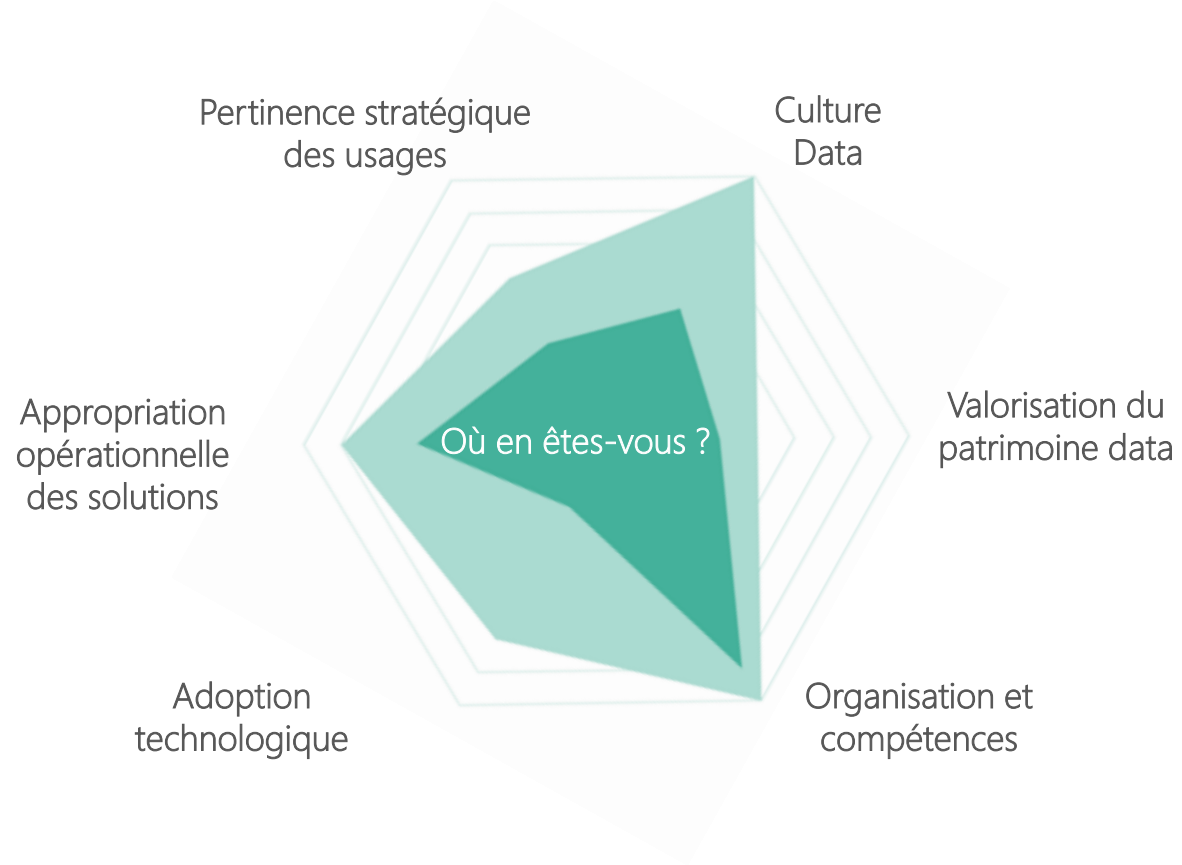
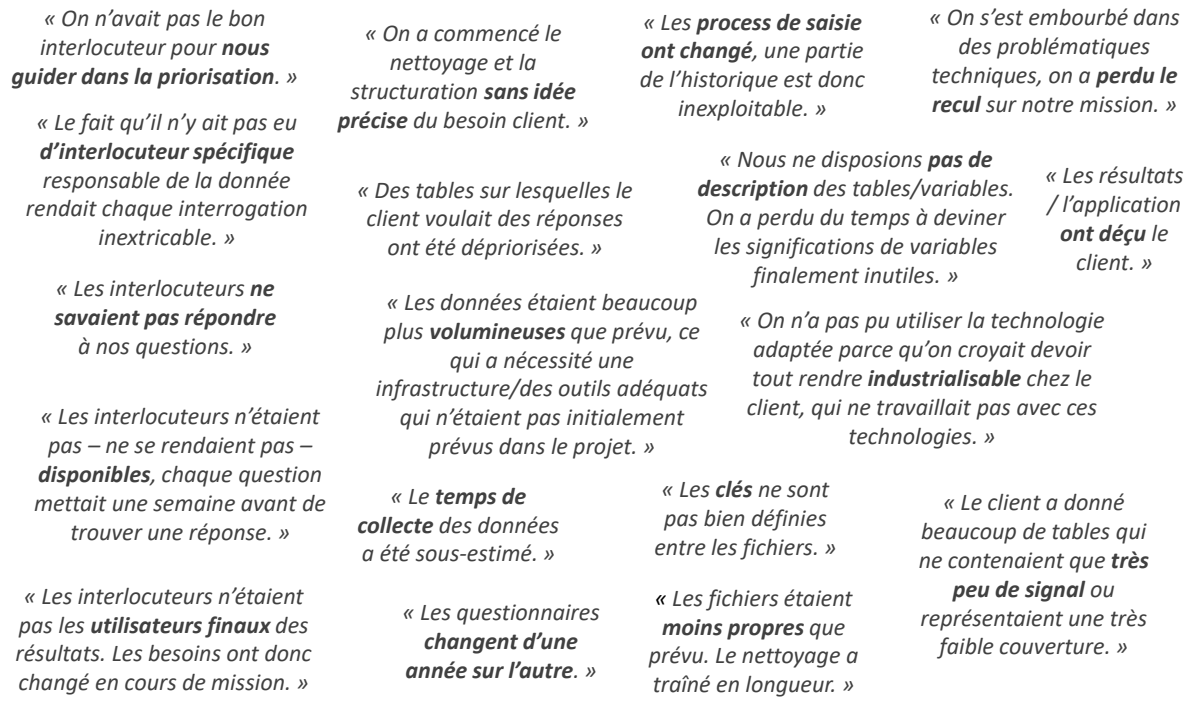


Figure 70 – Axes de montée en maturité proposées aux entreprises, sous la forme d'une matrice de maturité

Annexe 10 - Risques observés sur les projets data

Les risques présentés dans l'illustration de verbatim ci-dessous (Extrait 16) sont issus d'une enquête interne réalisée en 2018 par Orel Hacman, Data Scientist chez Quinten, auprès des équipes de Data Scientists ainsi que de Guillaume Bourdon (Fondateur de Quinten), Lucas Davy (Directeur des Opérations), Rami Fayed (Directeur de projet), Anna Nesvijevskaia (Directrice de projet), et Charles-Henri Ginter (Manager de projet).

L'enquête visait à établir les bonnes pratiques dans le cadre de l'avant-vente, visant des missions plus sereines (diminution préliminaire des incertitudes, localisation du risque restant, jalonnement de la phase de diagnostic et du cadrage des données) et des propositions commerciales plus pertinentes (affinage des objectifs et des livrables, optimisation du dimensionnement de la mission, et clarification des modalités d'intervention).



Extrait 16 ***Illustration des verbatim des Data Scientists*** *chez Quinten issus de l'enquête interne 2018 visant à identifier les risques clés sur les projets data*

Annexe 11 - Databook : genèse et prototypage

L'histoire du Databook s'est inscrite de façon inattendue dans ces travaux de recherche dès la confrontation aux premiers cas d'étude sur le terrain, le cas « Attrition Santé ». En effet, face à la complexité du sujet en termes de variété des données (nombreuses sources, nombreuses extractions, y compris erronées, sens non partagé par tous les acteurs, traitements complexes…), les arbitrages projets étaient d'une qualité insuffisante pour converger vers un résultat utile. Cette situation a rapidement fait émerger une nouvelle hypothèse de l'importance de la qualité des données et de la documentation des traitements associés pour l'efficacité interne et externe du dispositif projet, et, sur le terrain, un besoin de documenter les données traitées pour pouvoir avancer. Cette situation a donné lieu à une action double : poser un cadre théorique de la documentation d'un projet data, et réaliser un prototype de documentation pour l'utiliser et le tester sur le terrain.

1. Un cadre théorique à l'épreuve du terrain

Etant donné le choix du cadre de ces travaux de recherche et l'état de l'art, le modèle CRISP_DM a servi de référence pour dresser l'ensemble des documents qui jalonnent le projet data et permettent d'en garder la trace. Ces documents ont été classés, à la lumière du processus de production analytique, entre le résultat du travail de transformation des données comme matières brute, et le travail de documentation de cette transformation. Ces derniers ont été regroupés pour établir une base de travail théorique sur le prototype du Databook, constitué d'un ensemble de modules (correspondant à des onglets d'un fichier Excel crée pour l'occasion). La mise en pratique sur le premier cas d'usage ainsi que sur les suivants a permis de confronter le prototype théorique avec les besoins du terrain, et de se concentrer ainsi sur le développement et le test des modules les plus urgents, soit 6 modules initiaux (voir Figure 71).

Utilisation des modules du prototype Databook au cours des études de cas									
Modules		Cas 1 Attrition assurance santé	Cas 2 Prévision d'activité	Cas 3 Prévention santé et prévoyance	Cas 4 Contrôle de non-conformité	Cas 5 Sinistres lourds en dommage aux	Cas 6 Prédiction de prix des agrumes	Cas 7 Multi-équipement	Nombre de cas ayant mobilisé le module	Commentaire
Numéro	Nom									
0	Guide et terminologie	oui	oui	oui	oui	oui	non	oui	6	Séparé en 2 onglets et complété par les index de variables
1	Feuille de route projet data	oui	oui	oui	oui	oui	oui	oui	7	Simplification maximale
2	Méthode de rationalisation de l'inclusion/exclusion des données	non	non	non	non	non	non	non	0	**Partage de méthode oral**
3	Rapport d'exploration	non	non	non	non	non	non	non	0	**Rapports ppt et excel séparés**
4	Périmètre d'investigation	oui	oui	oui	oui	oui	non	non	5	Utile, mais pas pratique
5	Qualification des données source	oui	oui	oui	oui	oui	oui	non	6	Séparé en 2 onglets : soures et variables
6	Structure de la matrice d'apprentissage	oui	oui	oui	oui	oui	non	non	5	Séparé en 2 onglets : MLD et variables
7	Benchmark résultats analytiques	non	non	non	non	non	non	non	0	**Rapports ppt séparés ou outils Data Science dédiés**
8	Appropriation des résultats métier	non	non	non	non	non	non	non	0	**Rapports ppt, excel, ou solutions applicatives séparés**
9	Feuille de route usage	non	non	non	non	non	non	non	0	**Rapports ppt séparés**
10	Capitalisation de connaissances	oui	oui	non	oui	oui	non	non	4	Utile, à condition d'être communiquée plus activement

Figure 71 – Cadre théorique du Databook et sa mise en pratique dans les études de cas

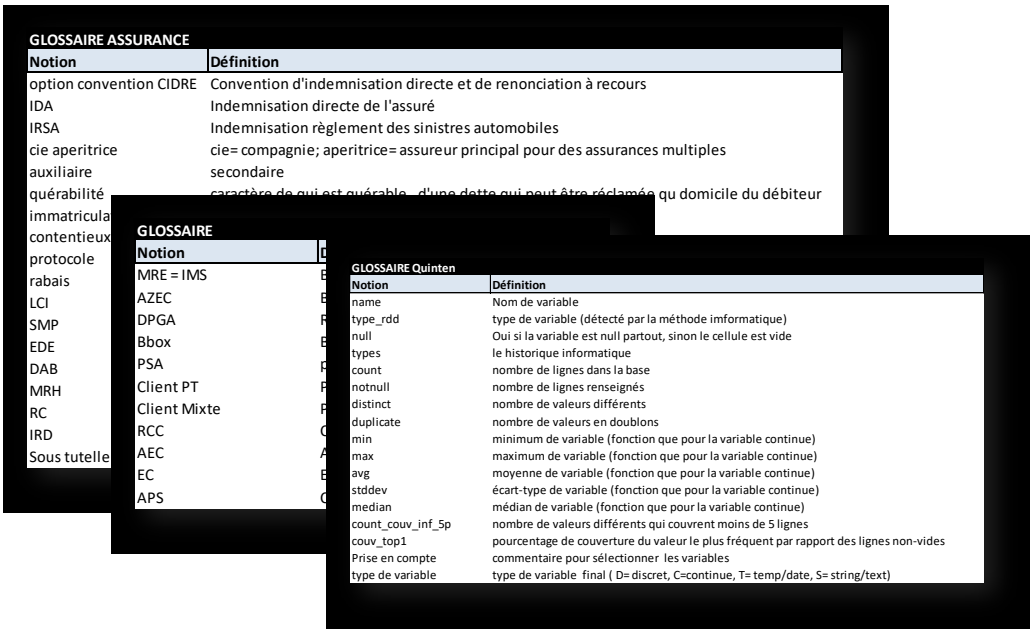
2. Détail des retours par module

Chaque module est ici décrit d'abord selon son aspect théorique, puis mis en perspective avec son illustration terrain, si testé.

0 - Le guide et la terminologie : aide à la lecture et à l'utilisation du Databook, et glossaires propres à chaque domaine d'expertise (vocabulaire métier, concepts algorithmiques, acronymes...) partagés au sein du dispositif projet data.

Il était était une fois… votre Databook Quinten	
Qu'est-ce qu'un Databook Quinten ?	
	Un Databook est un recueil d'informations structurantes, réalisé au cours de l'intervention de Quinten. Il décrit la façon dont Quinten structure les données reçues pour effectuer l'ensemble de ses analyses. Le Databook est un document de référence, destiné à tous les contributeurs du projet et aux utilisateurs des résultats de ce projet. Il se veut pédagogue, partagé et inspirant : l'histoire qu'il raconte est avant tout la vôtre.
Que comporte le Databook Quinten ?	
	Au-delà du préambule, le Databook Quinten est composé en chapitres, qui racontent l'histoire du projet co-construit pas à pas avec le client.
Project Map	Cet éditorial sert de référence au projet, de mise en perspective avec les délais, interlocuteurs, objectifs, étapes et volumétries en jeu
Périmètre	Il s'agit de l'exposition de l'histoire, de l'accroche. Qui est au cœur de l'intrigue ? Quel est notre contexte ? Que cherche-t-on à comprendre ? Ces concepts sont traduits en termes de données concrètes, et les conséquences de ces choix sont listées.
Variables recues	Chaque réception de nouvelles informations est un événement de l'histoire du projet. Quel est l'historique de réception de données, la nature des tables, la liste et la nature des données ? Quelles données présentent un intérêt pour l'analyse ?
Modèle conceptuel data	Comment l'histoire est-elle ficelée ? Quelle est l'organisation et quels sont les liens entre toutes les informations récoltées et utilisées ?
Variables finales	Chute de l'histoire, il s'agit de la description détaillée de la construction de la table qui servira à l'ensemble des analyses. Cette documentation est en soi une source d'inspiration pour de nouveaux projets.
Elements de capitalisation	La conclusion de l'histoire n'est qu'une ouverture à d'autres sujets. Tous les points d'optimisation potentiels, détectés au cours du projet, sont listés pour un usage futur.
Index construction variables	Les notes de bas de page sont regroupées au sein de l'index : il s'agit de l'ensemble des tables intermédiaires permettant de passer des données reçues aux variables finales.
Qui est l'auteur du Databook Quinten ?	
	Le Databook est une synthèse de travail réalisé par Quinten avec vos équipes projet. Il est produit par Quinten, mais peut être complété par le client par la suite (description de l'extraction des données envoyées à Quinten, sources, requêtes, etc.).

Extrait 17 ***Illustration du guide d'utilisation du Databook*** *(Module 0)*



GLOSSAIRE ASSURANCE

Notion	Définition
option convention CIDRE	Convention d'indemnisation directe et de renonciation à recours
IDA	Indemnisation directe de l'assuré
IRSA	Indemnisation règlement des sinistres automobiles
cie aperitrice	cie= compagnie; aperitrice= assureur principal pour des assurances multiples
auxiliaire	secondaire
quérabilité	caractère de qui est quérable - d'une dette qui peut être réclamée qu domicile du débiteur
immatricula	
contentieux	
protocole	
rabais	
LCI	
SMP	
EDE	
DAB	
MRH	
RC	
IRD	
Sous tutelle	

GLOSSAIRE

Notion	D
MRE = IMS	E
AZEC	E
DPGA	R
Bbox	E
PSA	P
Client PT	P
Client Mixte	P
RCC	C
AEC	A
EC	E
APS	C

GLOSSAIRE Quinten

Notion	Définition
name	Nom de variable
type_rdd	type de variable (détecté par la méthode imformatique)
null	Oui si la variable est null partout, sinon le cellule est vide
types	le historique informatique
count	nombre de lignes dans la base
notnull	nombre de lignes renseignés
distinct	nombre de valeurs différents
duplicate	nombre de valeurs en doublons
min	minimum de variable (fonction que pour la variable continue)
max	maximum de variable (fonction que pour la variable continue)
avg	moyenne de variable (fonction que pour la variable continue)
stddev	écart-type de variable (fonction que pour la variable continue)
median	médian de variable (fonction que pour la variable continue)
count_couv_inf_5p	nombre de valeurs différents qui couvrent moins de 5 lignes
couv_top1	pourcentage de couverture du valeur le plus fréquent par rapport des lignes non-vides
Prise en compte	commentaire pour sélectionner les variables
type de variable	type de variable final (D= discret, C=continue, T= temp/date, S= string/text)

Extrait 18 ***Illustration de la terminologie*** *(Module 0)*

1 - La feuille de route projet data : numéro de version ou date de l'instance de médiation, liste organisée d'indicateurs décrivant la maturité du dispositif au moment de l'instance de médiation, comprenant la priorisation des enjeux et des objectifs opérationnels, le poids des critères de succès métier, la liste et la description des ressources (y compris les outils, les rôles et responsabilités des acteurs...), la nature et niveau des incertitudes identifiées, l'état de résolution et les solutions (méthodes, ressources, séquencement, planning...). Cette feuille de route peut être aisément associée aux outils de pilotage de projets plus classiques utilisés par le Project Management Office.

[redacted]

Projet Prévoyance [redacted]

Dates clés

Date de lancement du projet :	**10/11/2015**	Version du Databook :	**V4**
Date prévue de livraison finale :	**15/02/2016**	Date d'envoi de la version :	**24/02/2016**
Flexibilité :	**15 jours**		

Acteurs du projet

Sponsor :	[redacted]
Chef de projet client :	[redacted]
Interlocuteurs client -**Data :**	[redacted]
Interlocuteurs client -**Data métier :**	[redacted]
Chef de projet Quinten :	**Anna Nesvijevskaia**
Equipe projet Quinten :	Dian Kang
Responsable commercial Quinten :	**Anna Nesvijevskaia**

Projet Prévoyance MidCorp - L'essentiel

Contexte

[redacted]est le premier assureur [redacted] Le chiffre d'affaires du groupe s'élève à [redacted] en 2013, et ses activités comprennent l'assurance de personnes, l'assurance de biens et responsabilité, l'assurance crédit, l'assistance, la gestion d'actifs et la banque. Le groupe est actif [redacted]. Les orientations stratégiques du groupe, et notamment les [redacted] poussent [redacted] à prendre une longueur d'avance en termes de prévention dans le domaine de l'assurance Entreprise, et le groupe organise dans ce cadre [redacted] juillet 2014. Dans le cadre de cette compétition, la proposition de création de valeur de l'équipe rouge retient l'attention du jury pour le projet d'utilisation du Big Data et de l'intelligence artificielle pour baisser l'occurrence et/ou le coût du sinistre. Le projet s'inscrit dans un contexte métier suivant : Existence de 5 bases de données riches mais non reliées : base clients, base sinistres, base contrats, business box, rapports de visite Prévention concentrée sur 10% des sites, soit 90 % à traiter Prévention majoritairement concentrée chez les ingénieurs de prévention L'enjeu du projet consiste alors à bénéficier de la totalité des informations disponibles pour impliquer les réseaux d[redacted]dans la baisse de la sinistralité à travers la prévention

Enjeux

Réduire les sinistres en fréquence
Eviter la transformation d'un sinistre en grave

Enjeu reformulé lors du diagnostic : se concentrer sur les sinistres lourds (>[redacted] €)

Objectifs du projet et résultats attendus

L'objectif du projet consiste à améliorer le ciblage et la prédiction des risques pour systématiser une prévention personnalisée afin de baisser la fréquence des sinistres de la totalité des sites, y compris les 90% des sites non visités : Découverte de nouveaux contextes d'émergence de risques à travers un croisement des différentes bases existantes (identification de combinaisons de facteurs/contextes de risques) Prédiction de la survenance des sinistres Cette démarche vise à fournir à l'intermédiaire un nouveau diagnostic d'exposition aux risques pour chacun de ses clients et à Animer d'une nouvelle prévention par tous les réseaux d[redacted]

Extrait 19 ***Illustration de la feuille de route projet data*** *(Module 1)*

2 - La méthode de rationalisation de l'inclusion/exclusion des données : liste organisée des critères d'inclusion et d'exclusion des données, modalités possibles de ces critères, et méthode de sélection des données qui en résulte. Il s'agit d'une méthode combinatoire qui associe les indicateurs de qualité des données issus des champs analytique, opérationnel et stratégique, aux

indicateurs de qualité liés à l'efficience interne du projet, c'est-à-dire le coût du traitement au sein du dispositif sous contrainte de ressources.

Convenu à l'oral sur le terrain, ou présenté sous forme de slides, et ajouté directement dans les modules suivants

3 - Le rapport d'exploration : ensemble d'éléments utiles à la compréhension des phases intermédiaires de construction (référentiels, résultats d'analyses descriptives préliminaires…) réalisés dans le cadre des instances de médiation et ayant conduit à un arbitrage sur la stratégie analytique. Le rapport d'exploration constitue ainsi le cadre d'évaluation et justifie les croix des critères d'inclusion et d'exclusion ainsi que les critères de succès prioritaires.

Réalisé sous la forme de rapports PowerPoint ou dossiers Excel séparés

Ces premiers modules permettent ainsi de documenter l'ensemble des arbitrages réalisés au cours des instances de médiation. Ils sont alimentés par l'état d'avancement de l'ensemble des phases du projet et guident la compréhension métier grâce au cadre d'évaluation.

Les modules suivants servent de documentation des livrables analytiques intermédiaires, et représentent quant à eux chaque jalon du chemin critique en permettant de suivre les versions grâce à l'avancement de la qualification des données au cours de chaque phase.

4 - Le périmètre d'investigation : liste organisée des concepts métier traduits en concepts data sous forme de ressources data possibles, état de convergence de la traduction des objectifs métier en cible d'analyse selon le niveau de qualification réalisé, et contribution de chaque concept métier au résultat opérationnel.

Statut du concept	
M	Validation métier nécessaire
D	Validation data nécessaire
A	Concept acté

Périmètre d'analyse						
	Statut	Concept	Information métier	Conséquence sur le modèle	Décision détaillée et traduction data	Date de décision	Valideur
Objet de l'analyse							
		Maille d'analyse					
	M		XXXX	XXXX	XXXX	XXXX	XXXX
		Périmètre de l'analyse					
	M		XXXX	XXXX	XXXX	XXXX	XXXX
			XXXX	XXXX	XXXX	XXXX	XXXX
Phénomène d'intérêt							
		Variable de sortie :					
	M		XXXX	XXXX	XXXX	XXXX	XXXX
	M		XXXX	XXXX	XXXX	XXXX	XXXX
Caractéristiques							
		Périmètre Sites / Contrats / Clients					
	M		XXXX	XXXX	XXXX	XXXX	XXXX
	M		XXXX	XXXX	XXXX	XXXX	XXXX
	M		XXXX	XXXX	XXXX	XXXX	XXXX

Extrait 20 ***Illustration du périmètre d'investigation*** *(Module 4) – non détaillé pour des questions de complexité et de confidentialité*

5 - La qualification des données source : liste organisée des données (données, tables, sources, référentiels…) perçues comme exploitables et faisant l'objet de l'investigation, et traduites en concepts métier, état d'avancement de la collecte et de compréhension de chaque donnée selon le niveau de qualification réalisé, et niveau de contribution de chaque variable et chaque source de données au résultat opérationnel.

Plan de collecte													
Informations sur les fichiers reçus													
N° de Lot	Date de demande estimée	Date de transmission	Expéditeur	Interlocuteur(s) clé(s)	Base de données	Table / File	File format	Commentaire	Granularité (1 ligne=…)	Nombre de lignes	Volumétrie en Go (compressé)	Profondeur historique	Jointure(s)
	1	10/11/2015	16/11/2015	[redacted]	MRE (IMS)	sinmre15	sas7bdat	Sinistres [redacted] (IMS, base his	NOSIN + CDGARSII	416355	148Mo	10 ans (01/2	NOPOL
	1	10/11/2015	16/11/2015		MRE (IMS)	sinmre15	sas7bdat	Sinistres [redacted] (IMS, base his	NOSIN + CDGARSII	416355	148Mo	10 ans (01/2	NOPOL
	1	10/11/2015	16/11/2015		MRE (IMS)	ipfmre15	sas7bdat	Portefeuille [redacted] (IMS, base	NOPOL	71160	208Mo	10 ans (01/2	NOPOL
	1	10/11/2015	16/11/2015		AZEC (V5)	portef_quinten_azec	csv	Portefeuille [redacted]	POLICE + RISQUE	9983	8Mo	10 ans (01/2	POLICE
	1	10/11/2015	16/11/2015		AZEC (V5)	sinistres_quinten_azec	csv	Sinistres [redacted]	SINISTRE + GAR_SI	57237	23Mo	10 ans (01/2	POLICE
	1	10/11/2015	16/11/2015		AZEC (V5)	primes_quinten_azec	csv	Primes [redacted]	police + exercice +	317543	28Mo	10 ans (01/2	POLICE
	2	18/12/2015	03/02/2016		Base Mono et Multi sites	Base Mono Sites MRE 12 2015	.xlsx	MASSIVEMENT MULTISITE	Police	339	42 Ko	3 ans	POLICE
	3	18/12/2015	12/02/2016		multi et visites multiples	multi et visites multiples	.xlsx	multi et visites multiples	multi-sites	106	58	3 ans	POLICE
	3	18/12/2015	12/02/2016		Base des Rapports Visite	Base DPGR VB 10 02 2016	.xlsx	Base des Rapports Visite	Rapport de visite	1689	7986 Ko	3 ans	POLICE
	4	21/12/2015	05/01/2016		Business Box	Dictionnaire Données Datama	.xlsx	Dictionnaire Données Datamart				3 ans (01/2013 à 01/2015)	
	5	18/12/2015	A recevoir		Business Box	XX		[redacted] - Datamart DATA INTE	Ipolice			3 ans (01/20:	POLICE
	5	18/12/2015	A recevoir					- Datamart DATA INTE	LLIGENCE			3 ans (01/20:	SOURCE_INITIAL,
	5	18/12/2015	A recevoir					- Datamart DATA INTE	Contrat			3 ans (01/20:	SOURCE_INITIAL,
	5	18/12/2015	A recevoir					- Datamart DATA INTE	LLIGENCE			3 ans (01/20:	SOURCE_INITIAL,
	5	18/12/2015	A recevoir					- Datamart DATA INTE	Client PT			3 ans (01/20:	SOURCE,
	5	18/12/2015	A recevoir					- Datamart DATA INTE	Client MIXTE			3 ans (01/20:	SOURCE_MIXTE,
	5	18/12/2015	A recevoir					- Datamart DATA INTE	Client MIXTE			3 ans (01/20:	SOURCE,
	5	18/12/2015	A recevoir					- Datamart DATA INTE	Siret Prospectable [redacted]			3 ans (01/20:	SIRET
	5	18/12/2015	A recevoir					- Datamart DATA INTE	Contrat Actif de client Siretisé			3 ans (01/20:	SOURCE_INITIAL,
	5	18/12/2015	A recevoir					- Datamart DATA INTE	LLIGENCE			3 ans (01/20:	SIRET,
	5	18/12/2015	A recevoir					- Datamart DATA INTE	client Siretisé actif			3 ans (01/20:	SIRET
	5	18/12/2015	A recevoir					- Datamart DATA INTE	Client avec Siret qualifié/Client avec détenteur de prc			3 ans (01/20:	NU_ODS_PER_ROL

Extrait 21 ***Illustration de la qualification des sources****, sous la forme d'un* ***plan de collecte*** *(Module 5)*

suite

Data Framing - Orange Oil Price prediction							
Concepts			**Dimensions**				
Category	**Detailed Concept**	**Metrics**	**Geographical granularity**	**Product granularity**	**Historical depth**	**Timeline Granularity**	**Publication pediodicity**
Market Price	Oranges	$	Global	Orange Oil	10 years	Monthly	
	Oranges	$		Fresh Oranges	10 years	Monthly	Monthly
				Processed	10 years	Monthly	Monthly
	Citrus	$					
	Juice	$		All Juices	20 Years	Yearly	Yearly
	Futures	$		Frozen concer	N/A	Daily	Daily
Weather	Existing	Min	Brazil Florida I	N/A	10 years	Monthly	Monthly
		Max	Brazil Florida I	N/A	10 years	Monthly	Monthly
		Précipitati	Brazil Florida I	N/A	10 years	Monthly	Monthly
	Possible (commercial data)						
	Events		Brazil Florida India China		10 years	On spot	On spot

suite

Source nature						Integration Process					
Name	**Description**	**Confidence level (/10)**	**Availability**	**Table Names**	**Table Formats**	**Collection facility**	**Spécificities**	**Necessary Treatments**	**terms of impact**	**Facility of integration**	**collection for the project**
Mintec/[redacted]									High	High	High
Mintec/									(Expert imput n	to be confirmed	Low
Mintec									(Expert imput n	to be confirmed	Low
									(Expert imput n	to be confirmed	Low
Euromonitor	Client Login N	5							Low	to be confirmed	Low
Bloomberg	Only one acc	9							(Expert imput n	to be confirmed	Low
MDA Weather	Weather data	8			Excel				High	to be confirmed	Medium
MDA Weather	Weather data	8			Excel				High	to be confirmed	Medium
MDA Weather	Weather data	8			Excel				High	to be confirmed	Medium
									to be confirmef	to be confirmed	Low
Digimind	News trackin	8							High	to be confirmed	Medium

Extrait 22 ***Illustration de la qualification des sources****, sous la forme d'un extrait de* ***liste de concepts qualifiés*** *(Module 5)*

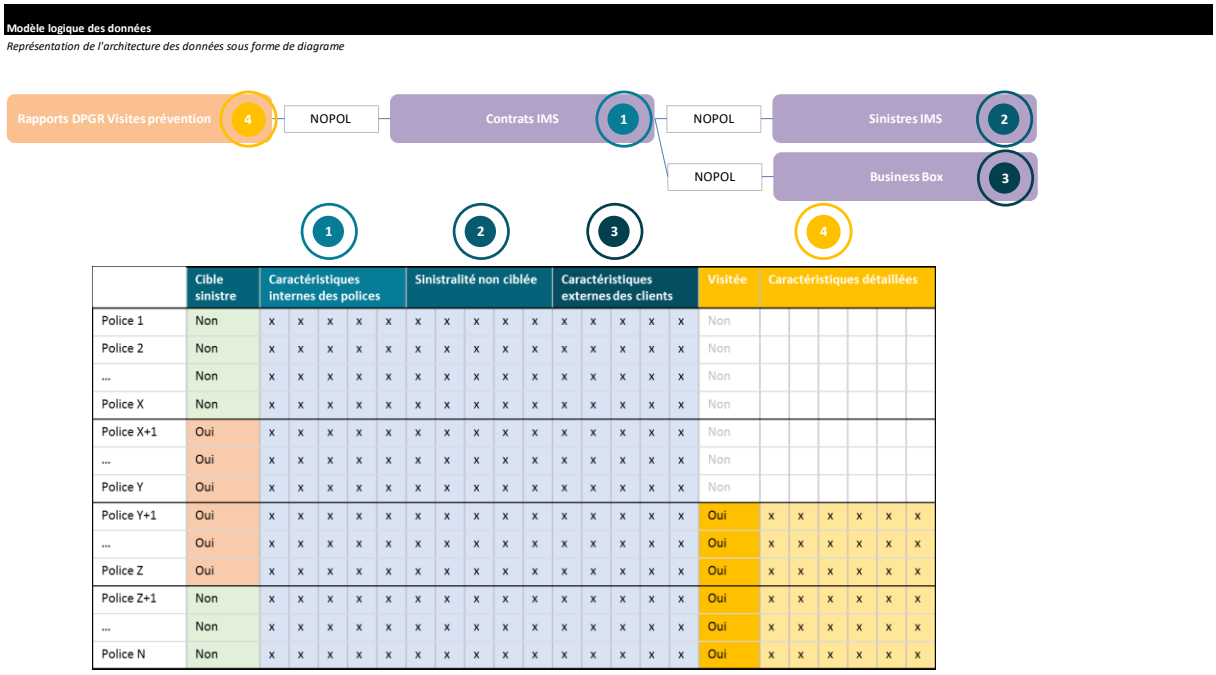
Statut des Variables	
Exclue	Variables exclues de l'analyse
Incluse	Variables incluses dans l'analyse : voire nature d'usage
Nouvel extract attendu	Variables / tables nécessitant un extract complémentaire
?	En cours d'analyse

Tableau des données reçues													
Informations sur les fichiers reçus			Informations sur les variables reçues								Nature de l'usage de la variable		
N° de Lot	Table / File	Jointure(s)	Variable	Explication	TYPE de variable (2 = Texte - 1 = Num.)	Priorité métier	DONNEE NOMINATIVE	Statut	test FORGE	Brute	Dérivée	Clé	Autre
1	sinmre15	NOPOL	TOPCRAC	?	?	non	non	Exclue	Non - vide				
1	sinmre15	NOPOL	TOPASSTR	?	?	non	non	Exclue	Non - vide				
1	sinmre15	NOPOL	chargecie	Charge	?	non	non	Exclue	OK	?			
1	sinmre15	NOPOL	regcie	Règlement	?	non	non	Exclue	OK	?			
1	sinmre15	NOPOL	reccie	Recours encaissés	?	non	non	Exclue	OK	?			
1	sinmre15	NOPOL	rapcie	Restant à payer	?	non	non	Exclue	OK	?			
1	ipfmre15	NOPOL	DTMJNE	DATE DE MISE A JOUR DU SEGMENT (AQQQ)	?	non	non	Exclue	OK				
1	ipfmre15	NOPOL	NOPOL	NUMERO DE POLICE (ANCIEN) X(14)	?		1 non	Incluse	OK	x		x	
1	ipfmre15	NOPOL	NOINT	NUMERO DE L INTERMEDIAIRE	?		1 non	Incluse	OK	x			
1	ipfmre15	NOPOL	CDPOLE	CODE POLE	?		1 non	Incluse	OK	x			
1	ipfmre15	NOPOL	CMARCH	CODE MARCHE	?		1 non	Incluse	OK	x			
1	ipfmre15	NOPOL	CDREG	CODE REGION	?	non	non	Exclue	OK				
1	ipfmre15	NOPOL	CDPROD	CODE PRODUIT	?		1 non	Incluse	OK		x		
1	ipfmre15	NOPOL	CSEGT	CODE SEGMENT	?		1 non	Exclue	Non - Modalité Fixe				
1	ipfmre15	NOPOL	CSSEGT	CODE SOUS SEGMENT	?	non	non	Exclue	OK				
1	ipfmre15	NOPOL	DTRESILP	DATE DE RESILIATION POLICE	?		1 non	Incluse	OK	x	x		
1	ipfmre15	NOPOL	DTTRAMVT	DATE TRAITEMENT DERNIER MVT DU TRAITE	?	non	non	Exclue	OK				
1	ipfmre15	NOPOL	NOAVEDER	DERNIER NUMERO D'AVENANT	?	non	non	Exclue	OK				
1	ipfmre15	NOPOL	NOPOLORI	NO DE POLICE COMPAGNIE PRECEDENTE	?	non	non	Exclue	OK				
1	ipfmre15	NOPOL	NOCIE	NUMERO DE COMPAGNIE	?	non	non	Exclue	Non - Modalité Fixe				

Extrait 23 ***Illustration de la qualification des variables*** *(Module 5) : la qualification progressive comprend le sens métier (« Explication »), l'utilité jugée a priori par les métiers,*

l'exploitabilité du point de vue règlementaire, la qualification statistique (« Test Forge »), et la façon dont la variable est utilisée par le modèle. Le « Test Forge » correspond au résultat de la qualification statistique de la variable grâce à une solution interne développée par Thomas Schott (CTO de Quinten). Il permet d'exclure notamment des variables vides, à faible variabilité, détecter des doublons, calculer les indicateurs simples comme le min, le max, la moyenne, le nombre de variables... Ainsi, le statut de la variable est issu d'une combinaison de critères d'exclusion et d'inclusion métier et statistiques, en amont du lancement des modèles algorithmiques.

6 - La structure de la matrice d'apprentissage : liste des variables brutes et construites le comprend la matrice d'apprentissage, qualification métier et technique (volumétries, formats…) de chaque variable, utilité dans a stratégie analytique (Maille, valeur de sortie si apprentissage supervisé, driver, variable intermédiaire, filtres pour la génération des matrices d'apprentissages/test/validation, variables de restitution des résultats…), état d'avancement de la construction et niveau de contribution de chaque variable au résultat opérationnel.



	Cible sinistre	Caractéristiques internes des polices					Sinistralité non ciblée					Caractéristiques externes des clients					Visitée	Caractéristiques détaillées					
Police 1	Non	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Non						
Police 2	Non	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Non						
...	Non	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Non						
Police X	Non	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Non						
Police X+1	Oui	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Non						
...	Oui	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Non						
Police Y	Oui	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Non						
Police Y+1	Oui	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Oui	x	x	x	x	x	x
...	Oui	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Oui	x	x	x	x	x	x
Police Z	Oui	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Oui	x	x	x	x	x	x
Police Z+1	Non	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Oui	x	x	x	x	x	x
...	Non	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Oui	x	x	x	x	x	x
Police N	Non	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Oui	x	x	x	x	x	x

Extrait 24 ***Illustration de la structure de la matrice d'apprentissage*** *(modèle logique et de sa représentation en instance d'arbitrage) (Module 6)*

Nomenclature des variables :
TX_Q_XXXX_XXXX : taux
DT_Q_XXXX_XXXX : dates
BC_Q_XXXX_XXXX : booléenne (0 ou 1)
CD_Q_XXXX_XXXX : code
MT_Q_XXXX_XXXX : montant
NB_Q_XXXX_XXXX : nombre
NU_Q_XXXX_XXXX : numéro
LB_Q_XXXX_XXXX : libellé
MM_Q_XXXX_XXXX : mois (de 1 à 12)
AA_Q_XXXX_XXXX : année

Nature des variables
? = Variables en attente
I ou K = Dérivée première (source : variables COVEA)
DS = Dérivée seconde (source : une variable dérivée première)
DZ = Dérivée complexe (sources multiples COVEA et dérivées Quinten)
DM = Variable à découper par modalité (_X = modalités)
Voir onglet "Index Construction Variables" pour plus de détails

Tableau des variables finales

#col	INTITULE_VARIABLE	Type de variable	Groupe	Description	N	Usage Prédiction	Usage Prescription	Spécificité des règl	Quantilisation
1	LB_Q_VS	Discret	Churn	Variable de Sortie : Churner Oui/Non (voir onglet périmètre)	DZ	Oui	Oui		
2	NU_AFFA	Discret	Affaire santé	Numéro d'affaire santé	K	Non	Non		
3	CD_TYPE_AFFA	Discret	Affaire santé	Code type affaire santé	K	Non	Non		
4	CD_ADHE	Discret	Souscripteur	code adhérent	I	Non	Non		
5	CD_CR	Discret	Affaire santé	Code Centre de Responsabilité	I	Non	Non		
6	DT_EFFE_AFFA	Date	Affaire santé	Date de début d'affaire	I	Non	Non		
7	DT_FIN_AFFA	Date	Technique	Date de fin de l'affaire (par défaut)	K	Non	Non		
8	DT_START	Date	Technique	Date de début de décompte des prestations santé	K	Non	Non		
9	DT_STOP	Date	Technique	Date de fin de décompte des prestations santé	K	Non	Non		
10	DT_SAIS_EVNM	Date	Technique	Date de churn (si non churner : "None")	K	Non	Non		
11	DT_EFFE_EVNM_SOUR	Date	Churn	Renseignée pour les churners, correspond à la date de churn	I	Non	Non		
12	CD_MOTI_RSLT_CONT	Discret	Churn	Code du motif de résiliation du contrat santé	I	Non	Non		
13	BC_ANNU	Discret	Technique	Annulation de la résiliation	K	Non	Non		
14	NU_PCP_EDE	Discret	Souscripteur	Numéro de souscripteur associé à l'affaire santé	K	Non	Non		
15	SSAA	Continue	Technique	Année de l'extraction des données	K	Non	Non		
16	NU_MOIS	Discret	Technique	Mois de l'extraction des données	K	Non	Non		
17	CD_TYPE_EVNM_EDE	Discret	Technique	Evènement S28 = résiliation sens [redacted], "None"=non résiliation au sens	K	Non	Non		
18	NB_Q_RNVL_AFFA	Continue	Affaire santé	Nombre de renouvellements, ie Ancienneté de l'affaire santé (Churner	DZ	Oui	Oui		QT_6_
19	CD_MARC	Discret	Affaire santé	Code marché du souscripteur de l'affaire	I	Oui	Oui	Oui	
20	CD_RGIM_ASRC_SOUR_01	Discret	Affaire santé	Existance d'un bénéficiaire 1 (REGIME GENERAL, VOLONTAIRE,PERS)	I	Oui	Oui		
21	CD_RGIM_ASRC_SOUR_02	Discret	Affaire santé	Existance d'un bénéficiaire 2 (EXPLOITANTS AGRICOLES (AMEXA))	I	Oui	Oui		
22	CD_RGIM_ASRC_SOUR_03	Discret	Affaire santé	Existance d'un bénéficiaire 3 (PROFESSION INDEPENDANTE (AMPI), soit TI		Oui	Oui		
23	CD_RGIM_ASRC_SOUR_60	Discret	Affaire santé	Existance d'un bénéficiaire 60 (REGIME LOCAL ALSACE-MOSELLE)	I	Oui	Oui		

Extrait 25 ***Illustration de la structure de la matrice d'apprentissage*** *(Module 6) : liste exhaustive des variables générées pour la matrice d'apprentissage, et qualification (nature et type, groupe et sens métier, prise en compte par les deux algorithmes utilisés, et spécificités de traitement. Toutes les variables générées ont du faire l'objet d'une standardisation de nomenclature réalisée spécifiquement pour le projet.*

7 - Le benchmark des résultats analytiques : liste organisée des modèles et des algorithmes mobilisés dans la phase de modélisation, description de leur nature, de la méthode et des paramétrages des analyses algorithmiques, indicateurs d'évaluation statistique et métier des résultats liés à chaque modèle, risques analytiques résiduels, état d'avancement de la modélisation et niveau de contribution de chaque modèle au résultat opérationnel.

Réalisé sous la forme de rapports PowerPoint ou dossiers Excel séparés, mais aussi dans les outils de Data Science dédiés (R, Python…)

8 - L'appropriation des résultats métier : liste organisée des résultats présentés pour évaluation métier, interprétation, méthode d'évaluation métier mise en œuvre, traduction des résultats en usages possibles, état d'avancement de l'appropriation des résultats et décision de validation, et contrition de chaque résultat évalué au résultat opérationnel.

Réalisé sous la forme de rapports PowerPoint, de dossiers Excel séparés, ou d'applicatifs métier développés sur mesure

9 - La feuille de route usage : pour chaque usage direct du projet, liste organisée des résultats jugés utiles, complétées d'indicateurs décrivant la maturité de l'usage, avec son propre contexte,

ses enjeux et ses objectifs, les bénéfices visés par l'usage, la description des ressources (rôles et responsabilités des acteurs métier et support, applicatifs métier…) et des risques analytiques, opérationnels et stratégiques résiduels (état de résolution des incertitudes), les contributions prévues le leur séquencement, planning, et indicateurs d'avancement par chantier. Cette feuille de route peut aussi être associée aux outils de pilotage de projets plus classiques utilisés par le Project Management Office, et dépend des spécificités opérationnelles de l'usage (techniques, humaines, processus d'exploitation...). Par exemple, un déploiement d'applicatif métier peut comprendre un cadre de pilotage agile par sprint, ou bien la transformation des résultats en modèles de données selon la méthode MERISE pour le développement technique.

Réalisé sous la forme de rapports PowerPoint séparés

10 - La capitalisation de connaissances : liste organisée de l'ensemble livrables intermédiaires du projet, qu'ils soient analytiques ou non, leur description et leur finalité potentielle en dehors du dispositif pour des usages indirects (dont les optimisations possibles des usages directs, comprenant des éléments de mise en qualité des données écartés du travail de production analytique au cours du projet), les ressources nécessaires à la transformation de ces livrables en usages et nature des bénéfices attendus.

Synthèse des éléments de capitalisation pour COVEA

Points soulevés	Date	Commentaires	Statut
❖Définition du churn			
•Date de churn : identifier la date de saisie de la résiliation pour ajuster le moment de la décision du churn (plus opérationnel) (dans l'idéal, il faut prenre la date de courrier de demande de résiliation, mais non trouvé dans les bases)	20/05/2015		
•Penser à ajouter la variable "ancienneté de l'affaire" en mois lors de la mise à jour des modèles afin de prendre en compte les nouveaux contrats, aujourd'hui exclus de l'analyse			
•Décès : vérifier la cohérence de la table vh.h_fich_pers_phys	27/05/2015		
•requête avec toutes les affaires qui ont soit un souscripteur qui a une date de décès renseignée dans la table vh.h_fich_pers_phys, soit un motif de résiliation avec CD_MOTI_RSLT_CONT=81			
•Code marché 99 : analyser pourquoi il existe des cas où le code marché est égal à 99 pour des affaires actives	15/07/2015		
❖Périmètre			
•Enrichir l'analyse avec des données sur le secteur d'activité / CA / nombre de salariés / autres pour les professionnels	21/05/2015		
•Elargir le périmètre temporel (à définir)	21/05/2015		
•Non unicité de la clé souscripteur : 8040 clients ont plusieurs affaires (de 2 à 7 affaires), dont seulement 2896 affaires résiliées - une analyse supplémentaire de ces doublons est nécessaire pour nettoyer la base source des affaires en doublon qui n'existent pas	27/05/2015		
A ce jour, 2 cas ont été analysée par [redacted] :			
Famille avec n enfants et n affaires au lieu d'une affaire couvrant toute le foyer			
Client avec 2 contrats dont 1 n'existe plus dans P9 ⍰ erreur dans la base de données source			
L'analyse reste à approfondir pour détecter tous les cas d'erreur correspondant au deuxième cas ou autres cas non identifiés à ce jour.			
❖Autres			
•Les sous-groupes dont le score d'attrition a tendance à augmenter seront étudiés dans le cadre d'une optimisation future du modèle (besoin d'avoir une profondeur d'historique sur le score de prédiction)	13/05/2015		
•Envisager une mise en cohérence du référentiel de motifs de résiliation dans le SI de [redacted] (2 bénéfices : data quality et opportunité de rétention)	30/04/2015		
•Le passage à la granularité souscripteur, voire foyer, est envisageable pour la suite du projet	28/07/2015		
•Si le taux de rensignement des caractéristiques client n'est pas idéal (vides), il reste tout à fait possible de créer une variable complémentaire qui correspond au niveau de connaissance client (taux de remplissage). Cette variable reste à définir à ce jour	10/11/2015		
•Sujet Multi-équipement : plusieurs règles pointent sur la possibilité de mettre en place des leviers opérationnels en termes de multi-équipement, et non pas en termes de limitation de l'attrition. Ces analyses méritent d'être approfondies avec des données complémentaires (sinistrlité sur les autres contrats, mise en évidence de l'évolution d'un client dans le temps) et les objectifs devront être reformulés. Les règles qui aujourd'hui ne sont pas directement transformables (risque maternité, bonus/malus) seront particulièrement analysées dans ce cadre.	26/02/2016		

Extrait 26 ***Illustration de la capitalisation de connaissances*** *(Module 10), montrant les pistes d'usages indirects (poursuite de projet data ou lancement de nouveau projet métier).*

3. Retours d'expérience des utilisateurs en dehors des études de projet

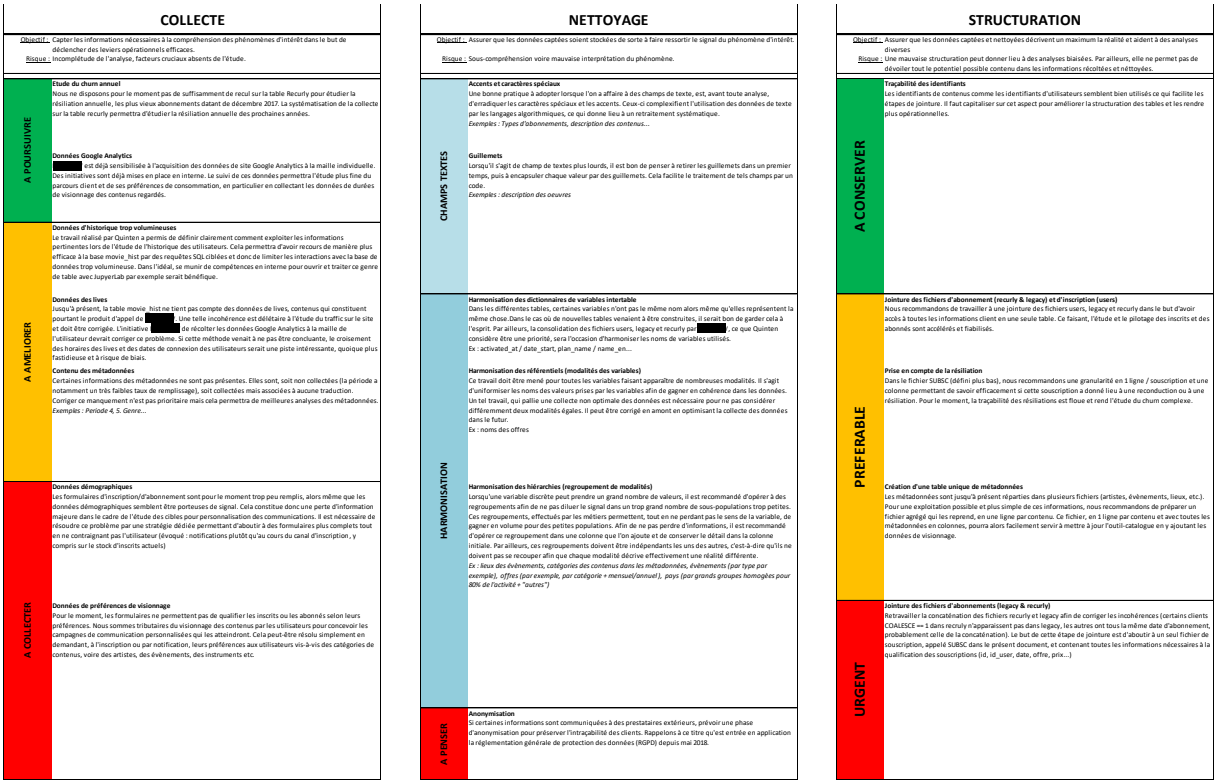
Le retour d'expérience des utilisateurs du Databook commence ici par des extraits de formation à l'utilisation du Databook, construite à partir du prototype et de son usage sur les projets Quinten entre 2015 et 2018 par les différentes équipes de Data Scientists. La formation est réalisée par deux Data Scientists (Mathilde Berthelot et Alexandra Chiorean) auprès de l'ensemble des équipes Quinten en 2018, et filmée. Cette formation avait pour objectif de décrire le Databook aux nouveaux arrivants, d'expliquer son utilité, et de partager les bonnes pratiques de son utilisation par les responsables, les contributeurs et les experts métier tout le long du projet.

Cette formation est le résultat d'une utilisation quasi-systématique du prototype de Databook au cours des projets data réalisés par Quinten depuis le démarrage de ces travaux de recherche, y compris sur des projets qui dans lesquels je n'ai joué aucun rôle. Elle traduit un retour d'expérience des utilisateurs Data Scientists, et met en évidence autant l'apport de l'outil (reconnaissance qualitative de son utilité, vérifiée par l'expérience) que ses limites. L'une des limites principales du prototype est sa complexité : en effet, le Databook adresse trois objectifs à la fois (gagner en efficience interne, assurer et piloter la qualité des données, gagner en efficience externe), alors qu'ils peuvent ne pas être prioritaires pour un projet donné. Aucun projet n'a donc mobilisé l'exhaustivité des modules de façon complète. La contrainte des ressources dédiées aux projets en est à l'origine, mais pas seulement. Si l'utilité du Databook est bien perçue sur les projets complexes en termes de variété des données, elle l'est moins dans le cadre d'analyses mobilisant une seule source de données, ou des données standardisées. Par exemple, une étude clinique mobilisant des algorithmes d'Intelligence artificielle nécessite moins la documentation du plan de collecte, vu que la source des données est unique et assez similaire d'une étude à l'autre. Par ailleurs, dans certains contextes de projets, des modules se sont avérés insuffisants :

- Les modules utilisés pour les instances de médiation constituaient des doublons avec des outils de gestion de projet, plus habituels et plus complets.

- Les modules « Feuille de route projet data », « Périmètre d'exploration » et « Capitalisation de connaissances » étaient jugés trop inconfortables pour la lecture, l'utilisation et le partage, notamment par rapport à un rapport Power Point classique. Le premier était préservé pour suivre les versions du Databook et des arbitrages, et pour garder la trace du contexte et du dispositif projet.

En revanche, la « Capitalisation de connaissances » a donné lieu une nouvelle proposition par une équipe de Data Scientists en restreignant l'usage du module plus particulièrement sur la gestion de la qualité des données. Cette proposition a fait l'objet d'une application sur un projet réalisé pour une télévision en ligne, et a permis de constituer un équivalent de cahier des charges préliminaire au déploiement des usages indirects. Ce cahier des charges a débouché sur une mise en œuvre par l'équipe data du client. Le retour du client a été très positif sur ce point, dans la mesure où la mise en qualité des données recommandée a conduit à la génération de nouvelles connaissances métier, ayant produit des usages concrets comme une optimisation de ciblage de prospects dans le cadre d'envoi de newsletters.



Extrait 27 ***Illustration de la proposition de l'évolution*** *du Module 10 (Capitalisation de connaissances) vers une « Feuille de route data », visant le déploiement des usages indirects*

Par ailleurs, la « Terminologie » (module 0) n'a pas été utilisée en dehors des cas étudiés dans le cadre de ces travaux. En effet, les spécificités de terminologie client sont captées essentiellement à l'oral, ou bien sous la forme de tables d'index annexées au Databook lorsque les termes correspondent à des jeux de données. De plus, toute variable fait l'objet d'une qualification en termes de sens, et tout terme utile une traduction en termes de données : cette activité de « traduction » est d'ores et déjà couverte par les modules dédiés, comme ceux de la

qualification des données source. En revanche, tous les termes nouveaux, convenus au cours de la production analytique sont bien présentés dans le Databook : il s'agit alors d'un contournement de l'usage du module au profit d'une documentation de termes nouveaux, et donc d'une capitalisation de connaissances. La finalité de ce contournement va plus loin : en effet, l'onglet se substitue à la « Feuille de route des usages » en documentant les développements nécessaires pour la mise en exploitation des usages directs générés à l'issue du projet data.

STATUT D'UN UTILISATEUR

Exemples donnés au 31/05

STATUT	DEFINITION	TABLES	IMPLEMENTATION
NOUVEAU	Utilisateur non-abonné inscrit au cours le mois de mai	users, subsc*	**users['date_joined']** entre 01-05 et 31-05 & **subsc['date_abonnement']** == NULL
INSCRIT	Utilisateur non-abonné inscrit avant le 1er mai ou ayant résilié son abonnement avant le 31 mai	users, subsc*	**users['date_joined']** < 01-05 & **subsc['date_abonnement']** == NULL
ABONNE	Utilisateur abonné au 31 mai (date d'abonnement renseignée antérieure au 31 mai et pas de date de résiliation renseignée)	subsc*	**subsc['date_abonnement']** < 31-05 & **subsc['date_resiliation']** == NULL

** Subsc est la table consolidée issue de recurly et legacy contenant toutes les données pertinentes (voir onglet Capitalisation Data - Section Structuration)*

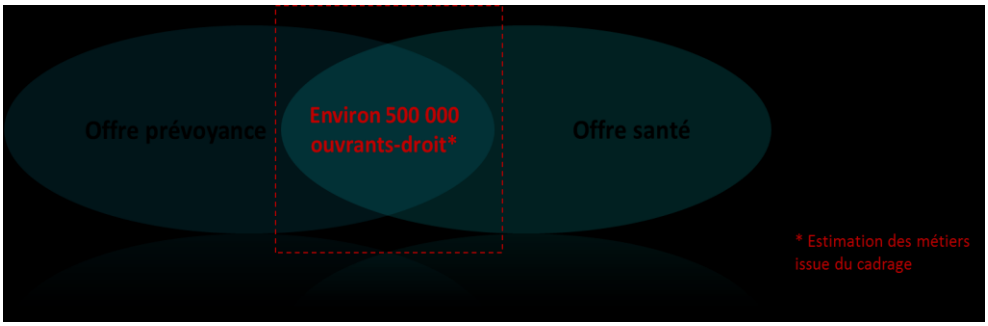
Extrait 28 ***Illustration d'un contournement du module « Terminologie »*** *en documentation de la terminologie nouvelle générée au cours du projet.*

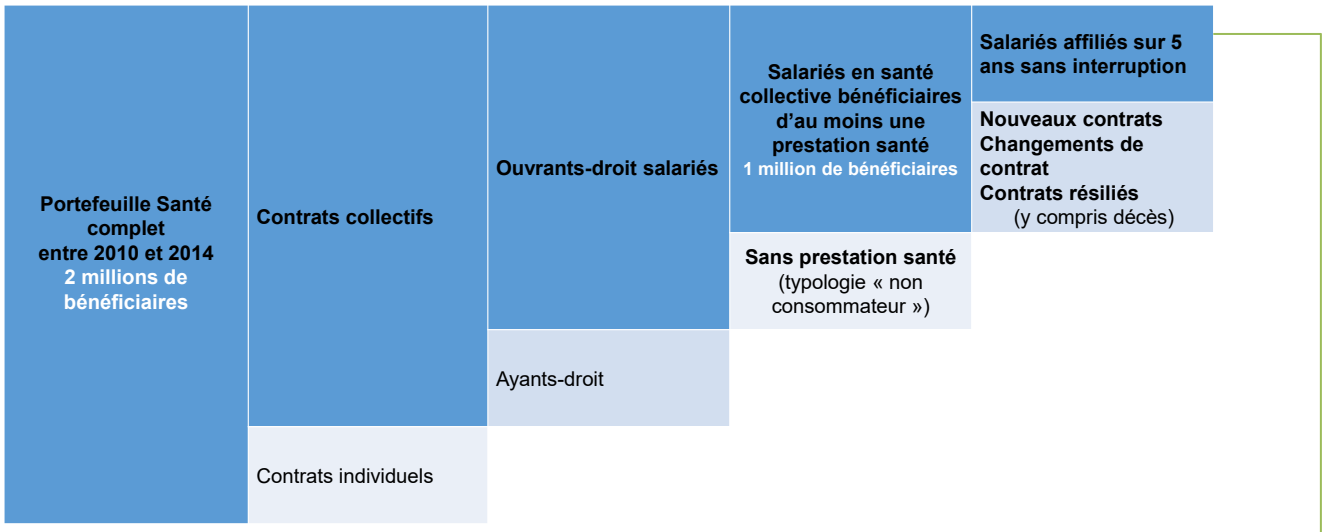
Cette confrontation du prototype du Databook au terrain, dans le cadre des études de cas ou en dehors, l'inscrit comme un outil pertinent pour la gestion de la qualité des données au cours du projet, pour son usage dans le cadre du suivi d'avancement des livrables analytiques intermédiaires au cours des instances de médiation, et pour la capitalisation de connaissances sous la forme de feuille de route data, générant des usages indirects. Son utilisation a permis d'identifier les concepts qualifiés, les critères minimaux qui permettent de les qualifier, et les raisons pour lesquels cette évaluation est clé. En effet, la qualification des données au cours de la production analytique constitue en soi un vecteur de réduction d'incertitudes.

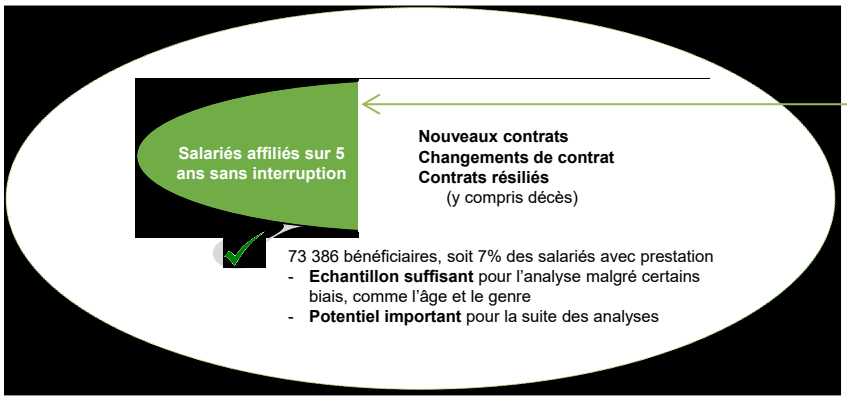
L'évolution de ce prototype reste possible, notamment à travers la modification des formats des modules jugés peu confortables et l'ajout des modules manquants, tout en préservant la spécificité de son utilisation dans le cadre des projets data et en intégrant l'état de l'art des pratiques de la gestion de la qualité des données.

Annexe 12 - Compte rendu cas 3 : Prévention santé prévoyance

La démarche adoptée pour atteindre les résultats consistait à établir un cadrage stratégique du sujet avec les experts métier, puis un cadrage data avec les experts data. La difficulté principale à ce stade consistait à réunir des bases de données qui n'avaient jamais été réunies auparavant, c'est-à-dire les données de l'offre santé et celles de l'offre prévoyance. La recherche de jointures a nécessité un temps de travail long, mobilisant des ressources techniques nouvelles comme la plateforme Big Data permettant de traiter des volumétries importantes de données, d'autant plus que les ordres de grandeur de référence étaient absentes en termes de volumétries de clients (dits « ouvrants-droits ») détenant à la fois un contrat santé et un contrat prévoyance, ou encore en termes de nombre de prestations santé ou prévoyance sur ce périmètre d'analyse, et les ordres de grandeur établis à partir des données fournies ne correspondaient pas aux estimations métier intuitives. Ainsi, cette première phase de structuration a dû être transformée en une phase d'analyse de périmètre, qui en soi a présenté un grand intérêt pour les experts data et métier.







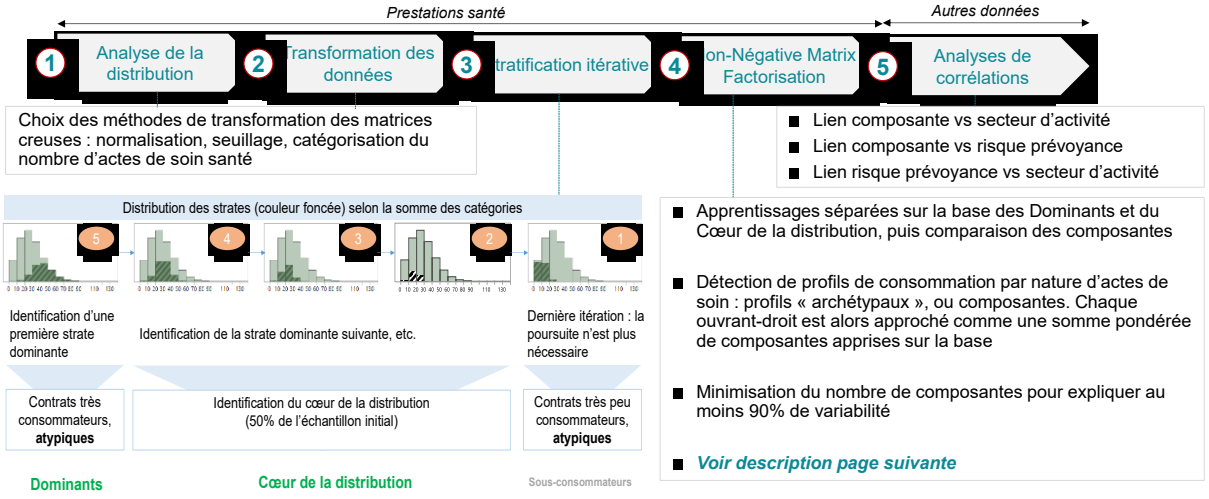
Extrait 29 ***Illustration de l'écart entre les estimations métier intuitives et l'analyse de périmètres construite à partir des données****, issue d'un rapport intermédiaire du projet : au lieu d'avoir environ 500000 ouvrants-droits détenteurs à la fois d'un contrat santé et d'un contrat de prévoyance, il s'agissait de seulement 73386 bénéficiaires directement analysables dans le cadre du projet. Ce périmètre est construit par élimination, dans l'assiette complète des bénéficiaires des contrats santé sur 4 ans, des contrats individuels (l'usage visant essentiellement des leviers opérationnels par branche et par entreprise, et non pas par individu), des ayants-droit (seul les salariés ouvrants-droit présentent un intérêt du point de*

vue de la prévoyance, notamment parce que ce contrat couvre l'absentéisme au travail), les cas où aucune consommation en santé n'est présente (il s'agit alors d'une typologie de consommation à part, facile à isoler sans aucun modèle), et enfin les cas où le contrat santé n'a pas été stable sur 5 années d'observation (en effet, la santé peut couvrir des risques longs sur la durée, toute période d'observation incomplète introduisant des biais significatifs).

Les analyses ont débuté par l'exploration des données santé ainsi délimitées, avant d'être corrélées avec les données prévoyance. La démarche d'analyse était la suivante :

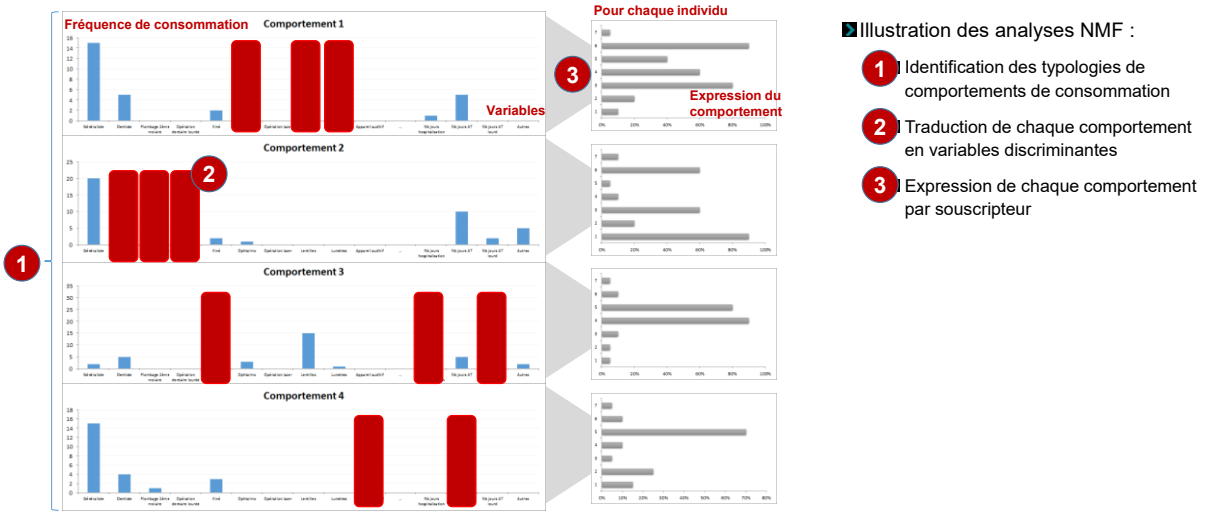
1. Décompte de toutes les prestations santé **par nature d'acte**
2. Mise en évidence des **typologies de consommations santé** (association d'actes et quantités consommées sur 5 ans)

Les résultats de l'analyse consistent à établir pour chaque individu du portefeuille l'expression des comportements de consommation santé typiques, soit une signature de comportement qui lui est propre. La méthode analytique pour générer ces résultats était particulièrement complexe et réalisée par un Machine Learner expert international sur le sujet, Paul Fogel. Elle a nécessité un travail de vulgarisation de fond dès les premières phases d'application, ce travail de médiation étant réalisé conjointement par le chef de projet (mobilisation de compétences pédagogiques, compréhension du niveau de maturité et des représentations sociales propres aux experts métier) le Machine Learner (mobilisation de son expertise sur la stratégie analytique et l'application des méthodes algorithmiques).

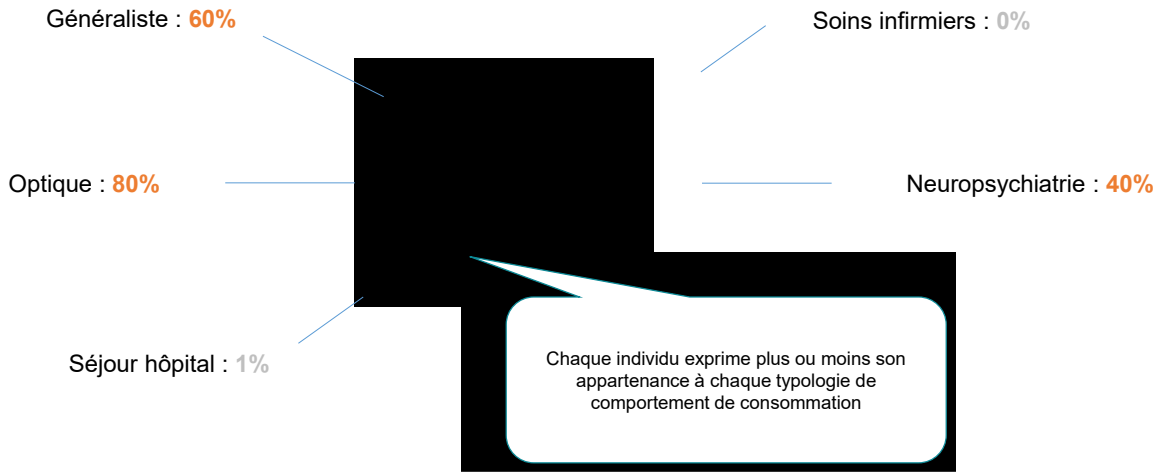


Extrait 30 ***Illustration de la méthode d'analyse de comportements santé,*** *issue d'une présentation intermédiaire : la stratégique analytique finale a été constituée de 5 étapes dont*

les objectifs et le déroulé ont été expliqués aux acteurs métier afin de favoriser l'appropriation des résultats.



Extrait 31 ***Illustration de l'étape d'analyse 4 (Non-negative Matrix Factorization)*** *issue de cette même présentation intermédiaire : l'ensemble des indicateurs et des représentations statistiques ont été confrontés aux experts métiers afin de s'assurer de leur bonne compréhension de cette méthode d'analyse, très inhabituelle pour eux.*



Extrait 32 ***Illustration des résultats attendus de l'analyse*** *issue de cette même présentation intermédiaire : il s'agit de l'expression par un individu donné de chaque type de consommation santé, ce qui permet d'éviter de segmenter les bénéficiaires par pathologie dans une situation où un même bénéficiaire peut avoir sa signature de consommation propre avec une appartenance plus ou moins grande à plusieurs typologies possibles. A ce stade, les typologies sont inventées, et non pas issues de l'analyse, afin de projeter les experts métier dans la démarche et de favoriser l'appropriation des résultats futurs.*

Le résultat livré comprend 20 typologies de comportements de consommation santé, identifiés sans *a priori* métier. L'ensemble des comportements santé, y compris les plus surprenants au premier abord, ont été interprétés par les médecins et autres experts métier, comme le montre un échange mail envoyé au cours de cette phase d'interprétation. Leur lisibilité a permis de pointer des comportements de consommation santé non identifiés à ce stade, comme les associations de consommations de soins qui correspondent aux troubles musculosquelettiques ou les burnouts (pathologie restant à confirmer à l'issue de l'analyse). L'interprétation des pathologies a nécessité de faire appel à des compétences complémentaire, et notamment faire intervenir des médecins (voir Extrait 33).

Bonjour [redacted].

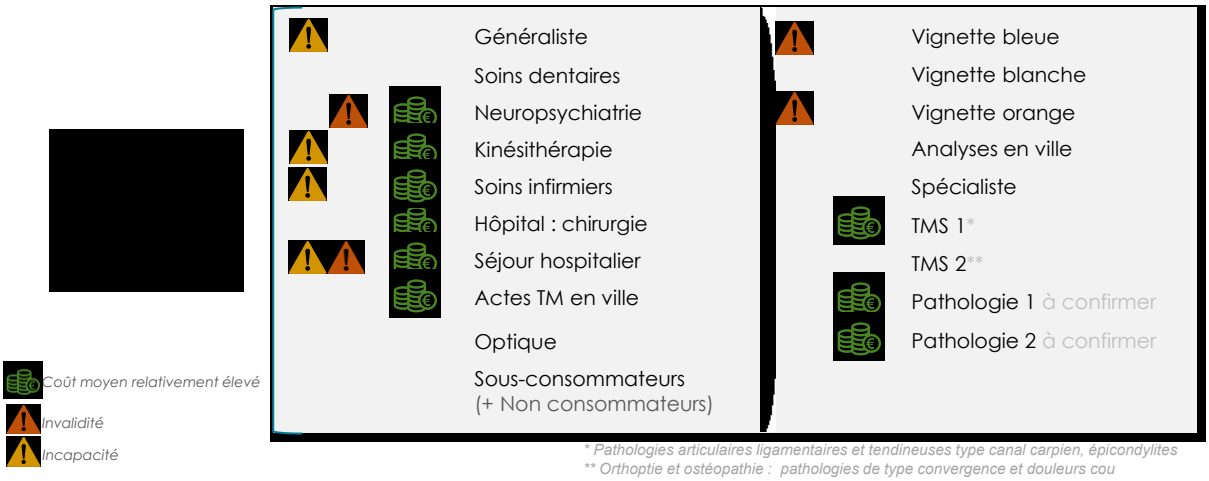
Suite à la restitution des résultats la semaine dernière, je me suis renseignée de mon côté sur le couple Ostéopathie/Orthoptie qui nous avait étonné, et discuté avec notre médecin au sein de Quinten. Il a émis l'hypothèse que cette association de comportements de consommation santé correspond non pas aux médecines alternatives, mais bien aux troubles musculo-squelettiques, en particulier ceux liés au maintien du cou pour les personnes travaillant avec les ordinateurs. Ce qui l'a mis sur cette voie est le périmètre : le fait d'avoir exclu les jeunes (et donc les enfants) avec la réduction du périmètre temporel et le filtre sur les ouvrants-droit élimine toute la cible naturelle de l'orthoptie, et ne laisse que les adultes souffrant de ce type de TMS.

J'ai trouvé aussi cet article pour illustrer ce point : http://www.osteopathe-ayena.fr/osteopathie-et-orthoptie--traitement-des-troubles-de-la-convergence_ad7.html

Extrait 33 ***Illustration du processus d'interprétation des résultats de l'analyse*** *issue des échanges de mails entre les acteurs du projet.*

3. Analyse des corrélations entre les comportements santé identifiés et les **risques lourds en prévoyance**

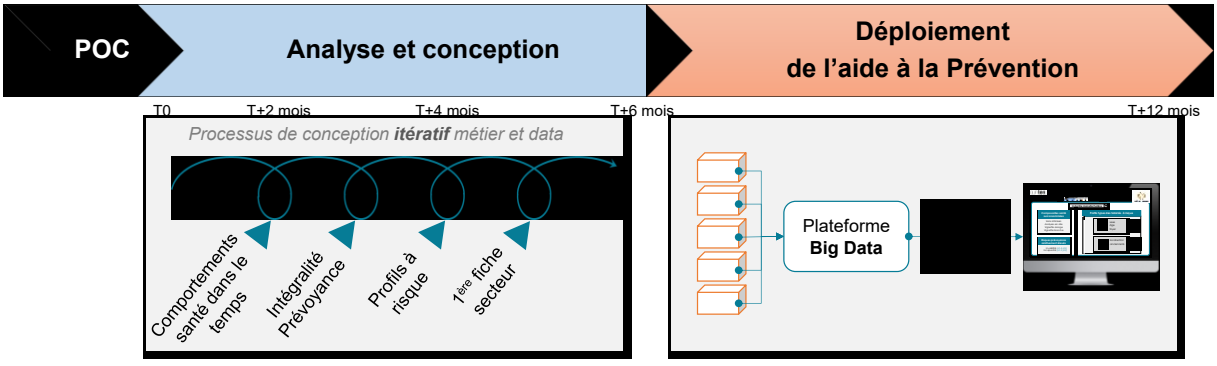
Chacun des comportements de consommation santé ainsi identifié a pu être mis en perspective avec les données prévoyance, plus précisément l'invalidité et l'incapacité. Le décès a été écarté à ce stade par construction (élimination des contrats résiliés à la date d'analyse, avec une durée inférieure à 5 ans, ce qui comprenait tous les décès survenus au cours de la période observée).



Extrait 34 ***Illustration des résultats de l'analyse*** *issue de la présentation finale : chaque comportement de consommation typique est qualifié en termes de coûts pour l'assureur, et associé à un risque d'invalidité ou à un risque d'incapacité.*

Ces comportements santé ont pu par ailleurs être mis en perspective avec les différentes branches qui constituent les canaux principaux en termes de prévention. Ainsi, chaque peut être plus ou moins touchée par les 20 comportements de consommation santé identifiés.

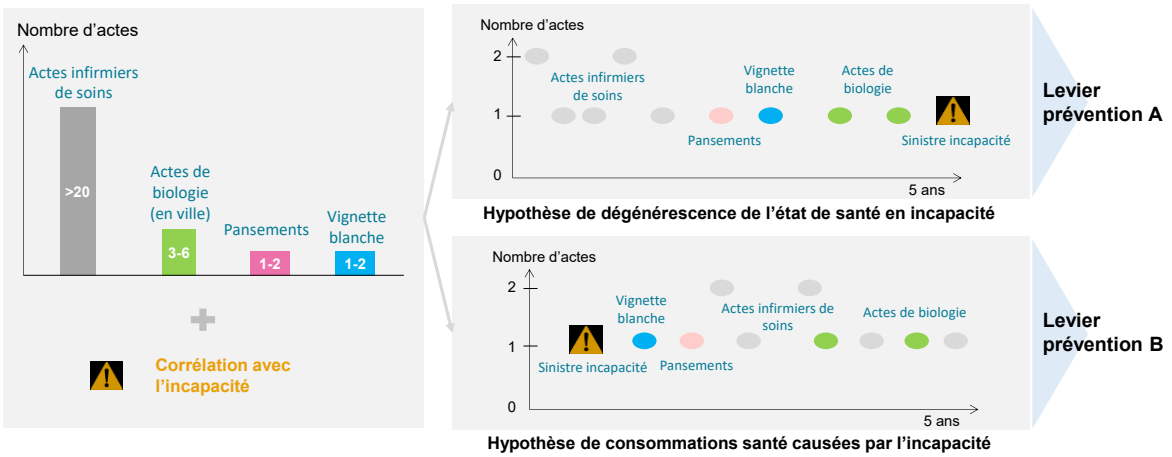
La POC a prouvé la richesse d'information contenue dans les données non exploitées jusque-là, et a donné lieu à un plan d'action moyen terme afin de tendre vers une application de prévention santé.



Extrait 35 ***Illustration de la feuille de route pour la poursuite du POC*** *issue de la présentation finale des résultats : les premiers moins devaient être consacrés à un processus de conception itératif impliquant les acteurs métier et data autour de 4 thématiques identifiées comme*

prioritaires avant de donner lieu à une phase de déploiement de solution métier utile à la prévention santé et prévoyance.

La limite principale des analyses à date consistait à ne pas pouvoir identifier l'ordre dans lequel les soins ont été consommés. Cette information nécessitait une seconde itération d'analyse sur chaque comportement jugé comme à fort enjeu par les métiers, mobilisant de nouvelles méthodes analytiques. Cette itération suivante du projet pouvait aboutir à un ajustement des leviers opérationnels.



Extrait 36 ***Illustration des limites des résultats et de la piste d'optimisation*** *issue de la présentation finale des résultats : ici il s'agit d'une typologie de consommation spécifique, les soins infirmiers (combinaison d'actes infirmiers, d'actes de biologie, d'achat de pansements et de médicaments à vignette blanche) corrélée à un risque d'incapacité (immobilisation de l'ayant-droit). Or, cette corrélation ne permet pas à ce stade de savoir si les soins infirmiers précèdent l'immobilisation, ou la suivent. L'hypothèse de travail consistait à dire que les deux cas étaient possibles, et que dans ces deux cas les leviers de prévention seraient différents. Cette hypothèse nécessitait une itération analytique complémentaire.*

Ce projet a pointé ainsi trois éléments clés complémentaires sur les projets data. D'une part, le travail de cadrage en amont s'est avéré comme capable à lui seul de générer des connaissances nouvelles sur une réalité métier et de lever des doutes quant à l'ampleur des enjeux liés à l'usage visé. Tout comme dans le cas de la redéfinition du concept d'attrition qui permettait d'isoler l'attrition activable, opérationnelle, de l'attrition structurelle liée à des facteurs non activables comme les décès des clients, il s'agissait pourtant d'une connaissance indirectement liée à l'objectif du projet. Ce constat a guidé dans d'autres contextes la mise en place d'une démarche

plus riche, en amont des phases classiques de structuration et d'analyse algorithmiques de données, sous la forme d'un diagnostic approfondi. L'objectif de cette phase est alors de mieux comprendre le contexte lorsque le niveau d'incertitudes est élevé avant de se prononcer sur la pertinence (enjeux et faisabilité) de la poursuite des analyses. Ce vecteur de réduction d'incertitudes sur le projet est alors l'objectif premier du diagnostic, contrairement aux phases suivantes qui portent plus directement sur la levée d'incertitudes sur l'usage visé. D'autre part, la complexité de certaines approches analytiques nécessitent une mobilisation importante en termes de pédagogie, tout le long du projet et non pas uniquement au moment de la restitution finale. Enfin, bien que génératrice de valeur indirecte à travers la capitalisation de connaissances utiles et inédites, la méthode exploratoire peut s'avérer itérative et induire des phases d'exploration complémentaires, plus proches d'une démarche de R&D que d'une démarche projet, ce qui est parfois difficilement réalisable dans le cadre de projets aux ressources contraintes comme des POC.

Annexe 13 - Compte rendu cas 4 : Contrôles de non-conformité

Le lancement : une phase de diagnostic consolidant la stratégie analytique

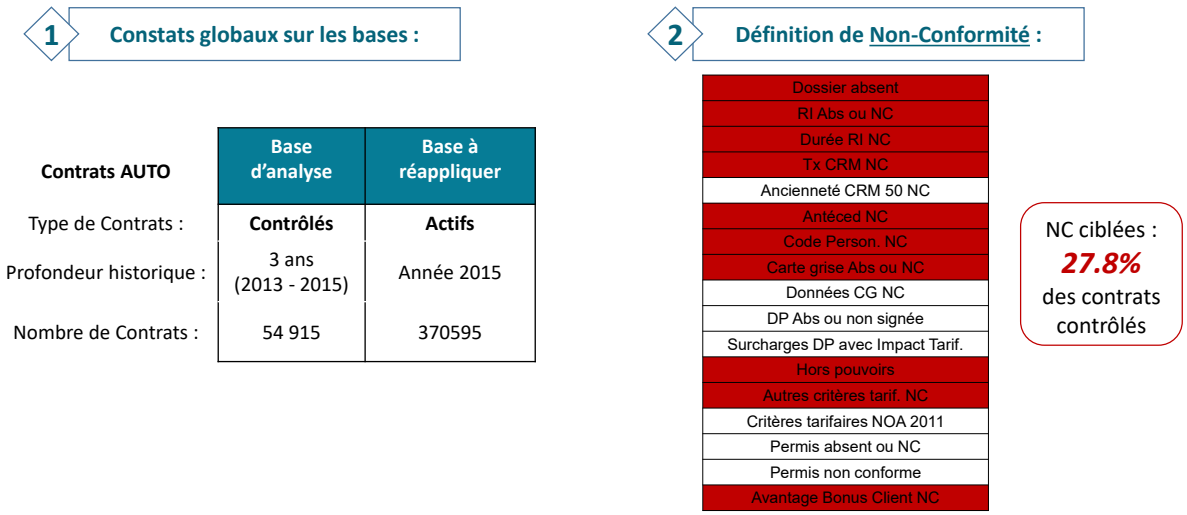
Les échanges dans la phase de diagnostic, réalisé par trois experts métier de la direction des contrôles et le chef de projet (moi-même) ont permis de préciser les concepts suivants :

- L'analyse portait sur l'ensemble des contrats Auto contrôlés au cours des 3 dernières années, soit environ 60000 contrats. Les résultats doivent être réappliqués sur l'ensemble des contrats auto actifs.

- La non-conformité (NC) était définie comme l'existence d'au moins une NC sur un contrat parmi 10 non conformités ciblées. En effet, la vérification d'un contrat donné pouvait donner lieu à 17 de types de NC (absence de dossier ou du rapport d'information, non-conformité de la carte grise...), dont seuls 10 présentaient un enjeu en termes d'impact financier ou étaient opérationnelles, c'est-à-dire possibles à corriger.

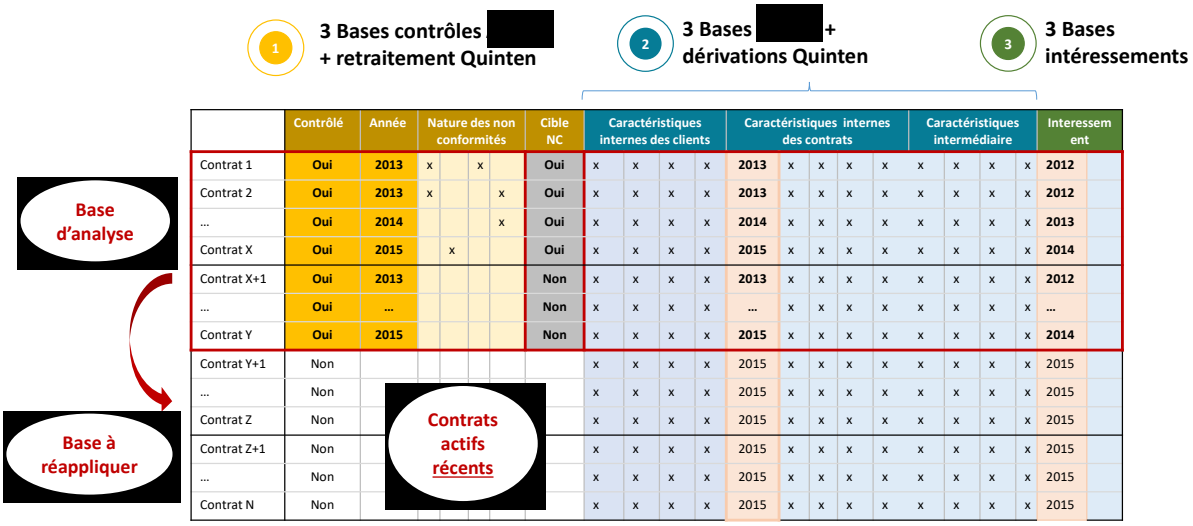
- Les corrélations entre NC (présence régulière de combinaisons de plusieurs types de NC) étaient écartées à l'analyse suite à une « coup de sonde » analytique qui a rapidement montrée que ces corrélations n'étaient pas significatives, c'est-à-dire que les NC avaient bien un caractère indépendant entre elles.

La cible est alors un phénomène qui touche environ 28% des contrats, c'est-à-dire que sur l'ensemble des contrats contrôlés sur les 3 années analysées, presqu'un tiers avaient été jugés non conformes, car au moins une NC sur les 10 était détectée.



Extrait 37 ***Illustration du périmètre d'analyse****, issue de la restitution du diagnostic : le périmètre est ici défini par la nature de l'objet d'analyse (contrats Auto), la profondeur de l'analyse (3 ans d'historique), le nombre de contrats dans la base qui servira d'analyse (contrats contrôlés) et dans la base qui servira à la restitution des résultats (Contrats actifs en 2015 et non contrôlés), et enfin par la nature du phénomène d'intérêt (10 types de NC, en rouge, parmi les 17 NC possibles), touchant à 28% des contrats.*

Au-delà de la définition précise du périmètre d'analyse, un audit complet des données a été réalisé. Plus concrètement, les premiers échanges métier ont permis d'établir un plan de collecte des données, les données définies ont été extraites et soumises au chef de projet, qui les a auditées et a déterminé quelles données étaient exploitables, quel sens avait chaque variable, et comment il fallait les mobiliser pour la poursuite des analyses. Etant donné que ce type d'analyse n'avait jamais été réalisé auparavant, un modèle conceptuel nouveau a été conçu *ad hoc* pour cette exploration. Ce modèle permettait de centraliser, année par année, l'ensemble des bases de données contenant les caractéristiques de contrôles de contrats contrôlés et leur résultat, l'ensemble des caractéristiques des contrats auto tels que saisis dans les systèmes de production, ainsi que les bases contenant les informations sur les agences qui avaient souscrit ces contrats, dont l'intéressement de tous les agents concernés. A ce stade l'intégration à l'étude de données externes, comme des données sociodémographiques, ont été écartées afin de privilégier l'exploration directe des données existantes et éviter un investissement jugé non négligeable pour construire les jointures entre ces données externes et les données internes. L'ensemble des cet audit, comprenant le périmètre d'exploration, le plan de collecte, le sens de chaque variable, et le modèle conceptuel des données a été documenté dans le Databook. Ce dernier a servi ainsi d'un cahier des charges pour la mise en œuvre de la stratégie analytique.



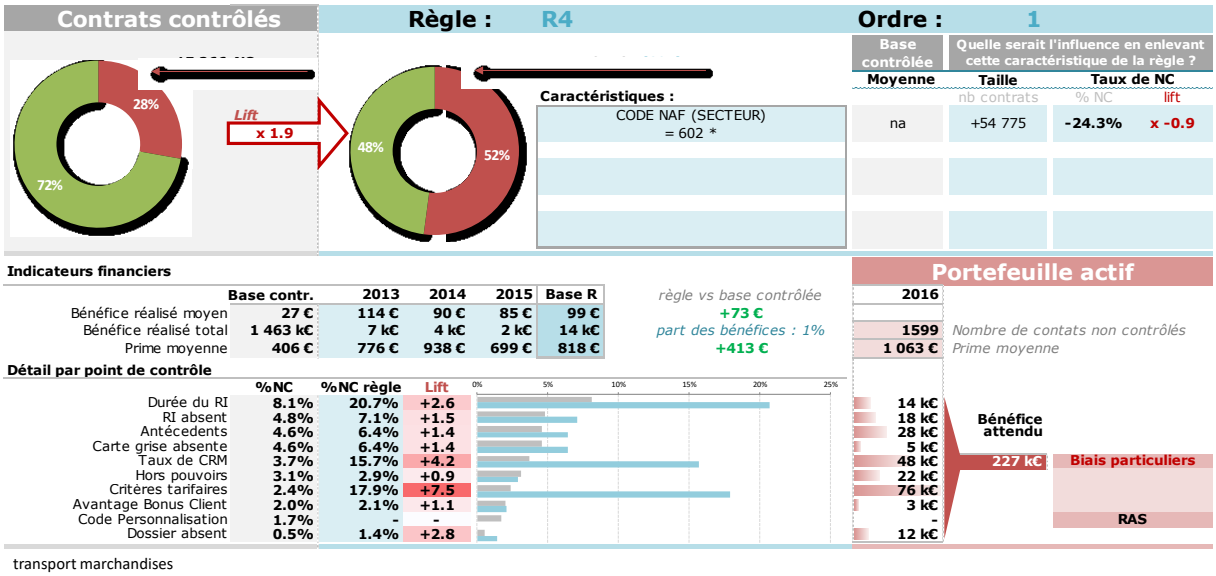
	Contrôlé	Année	Nature des non conformités				Cible NC	Caractéristiques internes des clients				Caractéristiques internes des contrats				Caractéristiques intermédiaire				Interessement
Contrat 1	Oui	2013	x		x		Oui	x	x	x	x	2013	x	x	x	x	x	x	x	2012
Contrat 2	Oui	2013	x			x	Oui	x	x	x	x	2013	x	x	x	x	x	x	x	2012
...	Oui	2014				x	Oui	x	x	x	x	2014	x	x	x	x	x	x	x	2013
Contrat X	Oui	2015		x			Oui	x	x	x	x	2015	x	x	x	x	x	x	x	2014
Contrat X+1	Oui	2013					Non	x	x	x	x	2013	x	x	x	x	x	x	x	2012
...	Oui	...					Non	x	x	x	x	...	x	x	x	x	x	x	x	...
Contrat Y	Oui	2015					Non	x	x	x	x	2015	x	x	x	x	x	x	x	2014
Contrat Y+1	Non							x	x	x	x	2015	x	x	x	x	x	x	x	2015
...	Non							x	x	x	x	2015	x	x	x	x	x	x	x	2015
Contrat Z	Non							x	x	x	x	2015	x	x	x	x	x	x	x	2015
Contrat Z+1	Non							x	x	x	x	2015	x	x	x	x	x	x	x	2015
...	Non							x	x	x	x	2015	x	x	x	x	x	x	x	2015
Contrat N	Non							x	x	x	x	2015	x	x	x	x	x	x	x	2015

Extrait 38 ***Illustration du modèle conceptuel conçu en amont de la structuration des données****, issue de la restitution du diagnostic : le modèle regroupe tous les contrats de la période d'observation et actuellement en portefeuille, les éléments des contrôlés passés (en jaune), les caractéristiques des contrats (en bleu) et les caractéristiques des agents (intéressements, en vert). La base d'analyse sert de base d'apprentissage, et les contrats actifs non contrôlés récents servent à constituer les cibles opérationnelles en réappliquant les résultats générés sur les contrats déjà contrôlés.*

Les difficultés d'appropriation des résultats par les experts métier en absence de compétences internes et d'outils de manipulation des résultats

Le Databook, accompagné d'explications et des supports de présentation aux interlocuteurs métier dans le cadre du diagnostic, ont été soumis à un Data Scientist qui a pu dans un premier temps réaliser la structuration selon le modèle conceptuel, et lancer les analyses prescriptives pour identifier sans *a priori* métier tous les contextes à risque de non-conformité grâce à l'algorithme Q-Finder (dont le fonctionnement a été décrit dans le cadre du projet attrition). Les analyses Q-Finder réalisées sur la matrice d'apprentissage (base d'analyse) ont mis en évidence 1 397 contextes (règles de complexité 1, 2 ou 3, c'est-à-dire composées de &, 2 ou 3 facteurs de risque combinés) qui pouvaient expliquer la non-conformité des contrats auto. La richesse de ces contextes pouvait être expliquée en grande partie par la diversité des natures de non conformités, soit 10 points de contrôle, correspondant à des sous-phénomènes inclus dans la notion de non-conformité.

26 règles ont été sélectionnées grâce à l'outil Diamond d'analyse de sous-groupes, et regroupées par thème pour représenter la diversité des contextes, que ce soit en termes d'indicateurs métier majeurs (montant des bénéfices potentiel) ou d'indicateurs complémentaires (lift, taille, bénéfice moyen, montant prime, point de contrôle…). Ces indicateurs métier ont été construits conjointement avec le chef de projet qui a joué le rôle de médiateur avec les experts métier. La sélection des règles d'intérêt (du point de vue opérationnel) nécessitait à ce stade une interprétation métier.



Extrait 39 ***Illustration de la représentation de l'une des règles issues des apprentissages Q-Finder****, issue d'une présentation intermédiaire : parmi les 53 915 contrats contrôlés, 28% sont non conformes. En appliquant la règle 4, qui correspond à tous les contrats signés avec un client dont l'activité est le transport de marchandises (code NAF = 602), on obtient 140 contrats, dont 73 ont été identifiés comme non conformes. Cela représente un lift de 1,9, c'est-à-dire que la non-conformité dans cette sous-population constitue presque le double de la non-conformité moyenne. Ce lift est particulièrement fort lorsqu'on s'intéresse à la non-conformité de type « Critères Tarifaires » : dans ce contexte à risque, il y a 7.5 fois plus de risques d'identifier un non-respect des critères tarifaires dans le contrat. En réappliquant cette règle sur l'ensemble des contrats récents et non contrôlés (cadre rouge en bas à droite), on obtiendrait 1599 contrats à contrôler, avec un bénéfice attendu de 227 K€, si les lifts et les primes moyennes étaient équivalentes.*

La recherche de contextes d'intérêt opérationnel a nécessité une prise en main approfondie des résultats de la part des experts métier, ce qui a démontré les limites de l'usage des outils classiques de type Excel et Access. En effet, la sélection opérationnelle des règles s'appuie sur

la possibilité de représenter les intersections et unions de règles et les indicateurs associés, les règles étant appliquées à l'ensemble des contrats à date. La représentation classique des intersections des règles sous forme de matrice de convergence n'est pas appropriée pour ce nombre de règles potentiellement opérationnelles : il aurait été nécessaire de mettre à disposition une interface dédiée pour permettre cette appropriation des résultats. Ce point constitue le principal apport de ce projet du point de vue processus de recherche. Il ne s'agit pas ici d'un besoin d'applicatif métier opérationnel, comme ce fut le cas de l'outil de prévision d'activité pour les contrôleurs de gestion, mais d'un outil d'aide à l'appropriation des résultats des algorithmes. Cet outil a depuis été développé par Quinten pour les projets similaires afin d'éviter des phases d'appropriation longues et complexes par les experts métier.

	Taille	R61	R1231	R36	R55	R302	R1210	R1109	R43	R1063	R1032	R1156	R1037	R230	R13	R72	R1017	R1160	R141	R218	R1029	R1023	R5	R114	R106	R4	R231
R61	21851		30%	25%	18%	15%	20%	15%	14%	11%	12%	10%	9%	11%	5%	6%	8%	9%	2%	3%	2%	2%	1%	1%	1%	0%	0%
R1231	9561	69%		24%	17%	16%	24%	29%	13%	23%	15%	10%	16%	10%	6%	10%	7%	6%	3%	2%	2%	4%	1%	1%	1%	0%	1%
R36	7842	70%	29%		16%	13%	57%	16%	17%	13%	16%	18%	9%	31%	6%	6%	11%	7%	3%	4%	2%	3%	1%	1%	1%	0%	1%
R55	7058	57%	23%	18%		17%	14%	15%	14%	11%	13%	9%	7%	9%	2%	7%	8%	5%	2%	3%	2%	2%	3%	1%	0%	0%	1%
R302	6647	49%	23%	16%	18%		14%	27%	12%	16%	12%	7%	17%	8%	2%	1%	6%	4%	2%	5%	4%	3%	1%	1%	1%	1%	0%
R1210	6151	72%	37%	73%	16%	15%		18%	17%	15%	16%	16%	13%	30%	8%	7%	10%	7%	3%	4%	2%	3%	1%	2%	1%	0%	0%
R1109	5956	55%	46%	21%	17%	30%	19%		17%	53%	26%	10%	27%	9%	3%	14%	9%	6%	3%	2%	2%	4%	1%	1%	1%	0%	0%
R43	5078	62%	25%	27%	20%	16%	20%	20%		17%	20%	10%	8%	11%	5%	5%	15%	6%	2%	4%	2%	2%	2%	1%	1%	1%	0%
R1063	4263	57%	51%	23%	18%	25%	22%	74%	20%		41%	10%	22%	10%	3%	12%	15%	5%	3%	2%	2%	4%	1%	1%	1%	0%	1%
R1032	4061	64%	35%	30%	23%	20%	24%	38%	25%	43%		14%	11%	21%	5%	9%	23%	7%	2%	3%	2%	3%	1%	2%	2%	0%	1%
R1156	3466	60%	28%	41%	18%	14%	28%	17%	14%	13%	16%		9%	19%	5%	6%	13%	7%	3%	4%	2%	3%	3%	2%	2%	0%	2%
R1037	3259	58%	46%	22%	15%	35%	24%	49%	13%	28%	14%	9%		8%	3%	16%	6%	6%	3%	1%	1%	3%	1%	1%	0%	0%	0%
R230	3024	77%	31%	80%	21%	17%	61%	17%	19%	14%	28%	21%	9%		5%	6%	14%	8%	2%	7%	2%	3%	1%	4%	3%	1%	2%
R13	2313	45%	23%	20%	6%	4%	20%	8%	10%	5%	10%	8%	4%	7%		5%	8%	5%	44%	1%	0%	1%	1%	0%	1%	0%	0%
R72	2200	56%	42%	21%	21%	3%	20%	37%	12%	24%	16%	9%	24%	9%	5%		8%	6%	5%	0%	0%	2%	1%	1%	1%	0%	0%
R1017	2100	78%	30%	39%	26%	18%	29%	25%	37%	30%	44%	21%	9%	20%	8%	8%		8%	2%	5%	3%	2%	3%	5%	2%	0%	3%
R1160	1950	100%	31%	28%	18%	14%	22%	17%	16%	12%	14%	12%	10%	12%	6%	6%	9%		3%	4%	2%	3%	2%	1%	0%	1%	1%
R141	1005	45%	27%	20%	14%	10%	15%	18%	10%	12%	9%	10%	10%	6%	100%	10%	5%	5%		1%	1%	3%	1%	0%	1%	0%	1%
R218	951	77%	23%	31%	20%	32%	26%	15%	21%	9%	14%	14%	5%	21%	2%	1%	11%	7%	1%		21%	9%	1%	4%	2%	5%	0%
R1029	892	46%	17%	15%	15%	29%	13%	10%	11%	7%	10%	7%	3%	8%	1%	0%	6%	4%	1%	22%		7%	1%	1%	1%	5%	0%
R1023	836	55%	44%	26%	19%	24%	23%	27%	11%	19%	14%	12%	11%	12%	4%	6%	5%	7%	4%	10%	7%		1%	1%	1%	1%	1%
R5	390	78%	26%	26%	56%	14%	23%	19%	23%	9%	13%	25%	4%	7%	6%	8%	18%	9%	2%	2%	1%	2%		0%	0%	0%	0%
R114	246	85%	32%	41%	16%	25%	38%	22%	29%	22%	39%	34%	9%	47%	2%	6%	46%	9%	1%	15%	4%	4%	0%		11%	1%	0%
R106	163	68%	28%	59%	15%	32%	38%	17%	37%	19%	40%	50%	7%	52%	9%	8%	26%	4%	4%	14%	4%	5%	0%	17%		1%	0%
R4	140	67%	25%	21%	18%	54%	19%	18%	17%	6%	11%	9%	5%	14%	0%	1%	6%	6%	0%	34%	31%	7%	1%	1%	1%		0%
R231	138	55%	36%	40%	53%	9%	12%	18%	1%	19%	28%	49%	6%	38%	7%	4%	38%	8%	5%	1%	1%	6%	0%	0%	0%	0%	

Extrait 40 ***Illustration d'une matrice de convergence classique****, issue d'une restitution intermédiaire : il s'agit d'une représentation excel des croisements entre les règles. Ainsi, la règle R4 (avant dernière ligne) est composée de 140 contrats dans la base d'apprentissage, mais ces 140 contrats sont aussi contenus à 67% dans la règle R61, ou encore à 25% dans la règle R1231. Cela veut dire qu'un seul et même contrat peut être non conforme pour plusieurs raisons, car il appartient à plusieurs contextes de risque. Cette matrice est difficile à manipuler en absence d'une représentation graphique plus dynamique qui pourrait accélérer l'investigation des experts métier.*

Par ailleurs, l'établissement des modèles prédictifs a fait l'objet d'un benchmarking de plusieurs algorithmes prédictifs, dont un modèle conçu de façon spécifique pour le projet : il s'agit d'un modèle intégrant ces mêmes règles en tant que variables. Les règles sont alors utilisées comme « drivers » de prédiction de non-conformité. Cette intégration, réalisée suite à l'établissement des règles et à l'identification des dimensions caractéristiques des règles, a conduit à établir une

projection des règles et une identification des degrés de proximité : cette phase met en évidence 10 familles de règles. Une fois les familles établies, chaque contrat fait l'objet d'un enrichissement en termes d'appartenance à une règle, et cette nouvelle variable est injectée dans le modèle prédictif, ce qui améliore sa performance. Cette modélisation a été réalisée selon une démarche de R&D par la Data Scientist accompagnée d'un Machine Learner plus expérimenté, et a permis de générer une capitalisation interne sur les modèles prédictifs. Cependant, cela n'a pas fourni d'éléments particuliers aux experts métier pour mieux comprendre le score, qui est tout de même resté une « boite noire ».

Le résultat du modèle prédictif était ainsi un score de risque de non-conformité pour chaque contrat. Il était modalisé pour être sensible aux non conformités, mais son usage était régi par le choix du seuil opérationnellement acceptable. Par exemple, en fixant un seuil de 20% aux probabilités de non-conformité (dès que la probabilité de non-conformité dépassait 20%, le contrat était prédit « non conforme »), le modèle prédisait 81% des non conformités observées, ce qui permettait de ne pas en oublier beaucoup. Cependant, cette sensibilité était au détriment de la spécificité : parmi les contrats qui étaient bien conformes, 61% étaient prédits comme non conformes. Ce choix de seuil privilégiait ainsi la détection de contrats non conformes à l'économie de contrôles. Il pouvait toutefois être ajusté selon l'effet opérationnel recherché : cette proximité entre les critères d'évaluation du modèle et la réalité opérationnelle a été comprise par les experts métier, et a constitué un argument complémentaire pour recruter rapidement un Data Scientist en interne pour travailler directement dans les équipes avec les experts métier.

Au final, environ 82% contrats prédits comme non conformes en 2016 appartenaient à un contexte (une règle explicite), et étaient ainsi « expliqués », ce qui permettait de mettre en place des thématiques de contrôle explicites. L'articulation des résultats du point de vue opérationnel a constitué le cœur du travail des experts métier : en effet, l'objectif a été d'une part d'établir des thématiques de contrôle de non conformités (selon l'appartenance à une ou plusieurs règles) et d'autre part de définir des seuils de probabilité de risque selon les ressources de contrôle disponibles sur le terrain. Ce travail aurait également nécessité des outils de manipulation de données plus appropriés de leurs outils habituels comme Excel.

Malgré cette phase d'appropriation des résultats et de transformation en leviers opérationnels assez complexe et chronophage pour les experts métier, les résultats ont permis d'établir une nouvelle stratégie de contrôle de non conformités à l'échelle de l'entreprise, basée sur les

résultats du projet. L'impact estimé de cette nouvelle stratégie de contrôle consistait à économiser 30% des ressources de contrôle des contrats auto pour les basculer sur le contrôle d'autres produits. Les contrôles auto gardés permettent par ailleurs de générer un gain de 15% en termes d'impact financier.

Capitalisation : des retombées indirectes significatives et un investissement dans les outils d'aide à l'investigation

Tout d'abord, la montée en maturité des équipes métier sur les méthodes propres aux projets data, ainsi que les résultats prometteurs, ont conduit à accélérer l'internalisation des compétences grâce à un recrutement de Data Scientists directement au sein de l'équipe de contrôle. Cette internalisation des compétences a rapidement conduit à un élargissement du périmètre d'analyse à l'ensemble du périmètre, et non pas seulement aux produits auto. Par ailleurs, de nouveaux outils du marché ont été mis en place pour accélérer le travail d'investigation, et surtout d'appropriation des résultats. Cette dynamique a permis de multiplier les bénéfices estimés à l'issue du projet : en effet, l'impact financier réel mesuré à l'issue d'une année d'exploitation de l'usage généré par le projet, s'élevait à 50% au lieu des 15% estimés initialement.

Par ailleurs, Quinten a pu confirmer grâce à ce projet l'importance de la mise en place d'outils d'aide à l'appropriation des résultats, notamment pour traiter des jeux de règles Q-Finder, leur application sur les données actualisées, ou avoir la possibilité de croiser des règles et de visualiser les indicateurs clés sur ces règles, union de règles et intersections de règles. Cette conclusion a été confirmée par des projets similaires dans le secteur de l'assurance plus largement, mais aussi des laboratoires pharmaceutiques. Elle a donné lieu au développement d'un outil de restitution dédié, appelé Cristal, et destiné notamment à des interlocuteurs métier assez matures face à la data, qui ont ainsi la possibilité de manipuler les données et les résultats pour mieux se les approprier.

Table des matières

le cnam Anna NESVIJEVSKAIA DicenIDF

Phénomène Big Data en entreprise : processus projet, génération de valeur et Médiation Homme-Données.

Résumé

Le Big Data, phénomène sociotechnique porteur de mythes, se traduit dans les entreprises par la mise en place de premiers projets, plus particulièrement des projets de Data Science. Cependant, ils ne semblent pas générer la valeur espérée. La recherche-action menée au cours de 3 ans sur le terrain, à travers une étude qualitative approfondie de cas multiples, pointe des facteurs clés qui limitent cette génération de valeur, et notamment des modèles de processus projet trop autocentrés. Le résultat est (1) un modèle ajusté de dispositif projet data (Brizo_DS), ouvert et orienté sur les usages, dont la capitalisation de connaissances, destiné à réduire les incertitudes propres à ces projets exploratoires, et transposable à l'échelle d'une gestion de portefeuille de projets data en entreprise. Il est complété par (2) un outil de documentation de la qualité des données traitées, le Databook, et par (3) un dispositif de Médiation Homme-Données, qui garantissent l'alignement des acteurs vers un résultat optimal.

Big Data, Data Science, Intelligence Artificielle, Qualité, Médiation, Stratégie, Valeur, Entreprises, Humain-Machine, Capitalisation, Projet Data, Homme-Donnée, Conduite de projet, Indicateurs de valeur

Résumé en anglais

Big Data, a sociotechnical phenomenon carrying myths, is reflected in companies by the implementation of first projects, especially Data Science projects. However, they do not seem to generate the expected value. The action-research carried out over the course of 3 years in the field, through an in-depth qualitative study of multiple cases, points to key factors that limit this generation of value, including overly self-contained project process models. The result is (1) an open data project model (Brizo_DS), orientated on the usage, including knowledge capitalization, intended to reduce the uncertainties inherent in these exploratory projects, and transferable to the scale of portfolio management of corporate data projects. It is completed with (2) a tool for documenting the quality of the processed data, the Databook, and (3) a Human-Data Mediation device, which guarantee the alignment of the actors towards an optimal result.

Big Data, Data Science, Artificial Intelligence, Quality, Mediation, Strategy, Value, Business, Human-Machine, Capitalization, Data Project, Human-Data, Project Management, Value Metrics