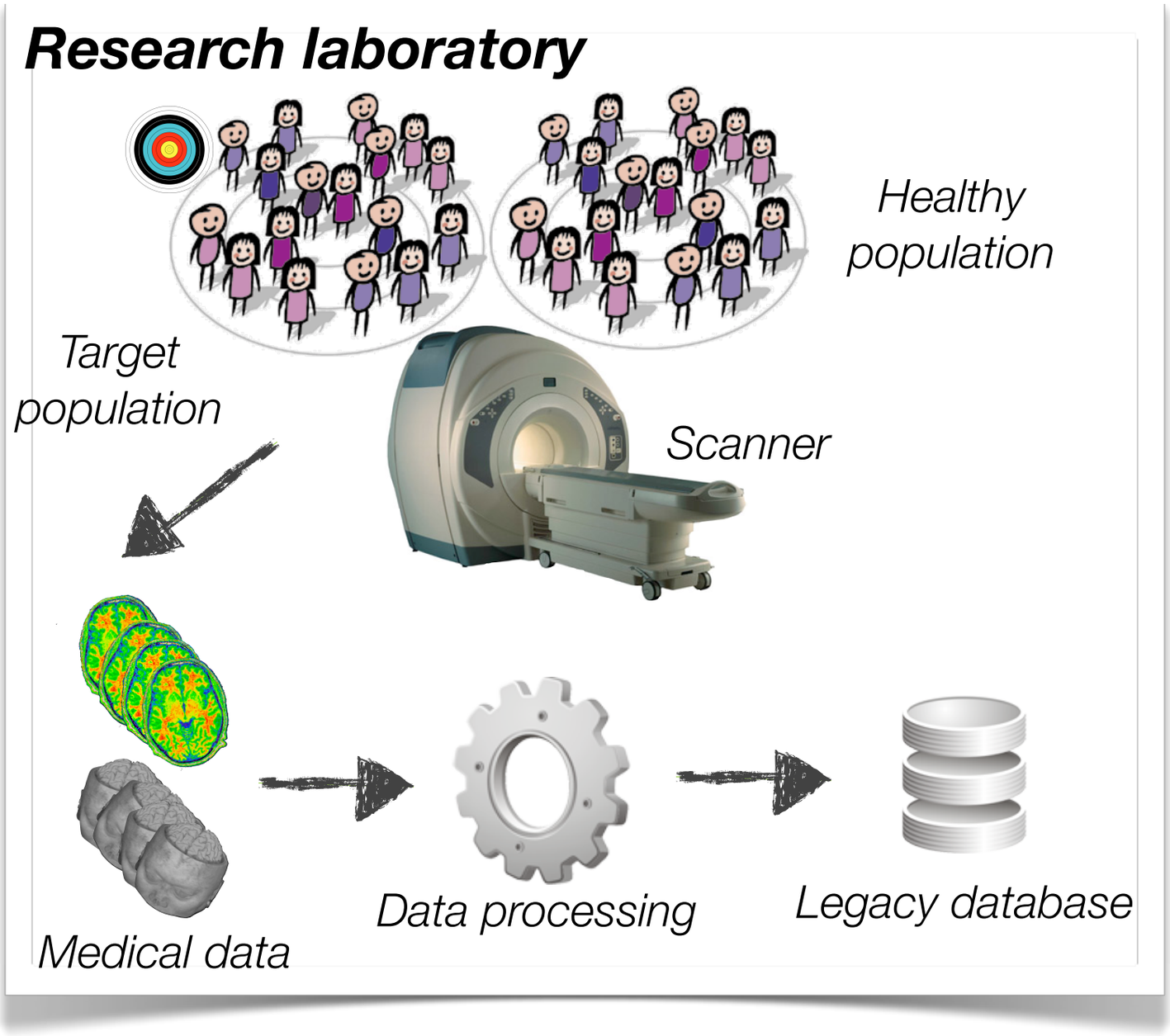
Distributed knowledge sharing and production through collaborative e-Science platforms

PhD Defense - Alban Gaignard

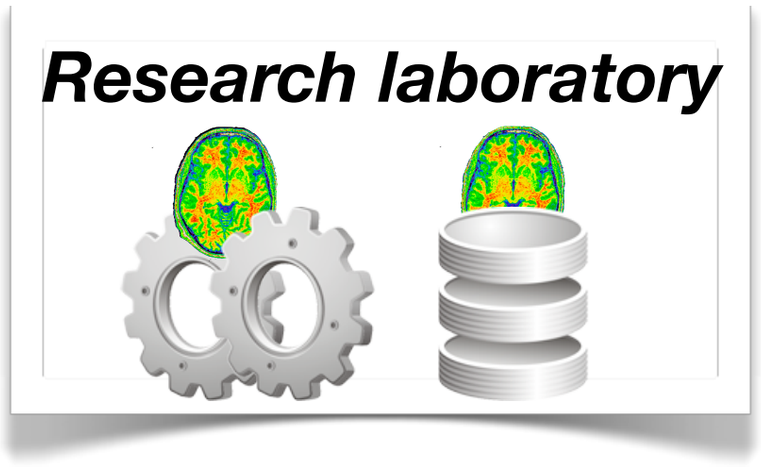
Advisor: Johan Montagnat

CNRS, University of Nice Sophia Antipolis,
I3S Laboratory, MODALIS research group

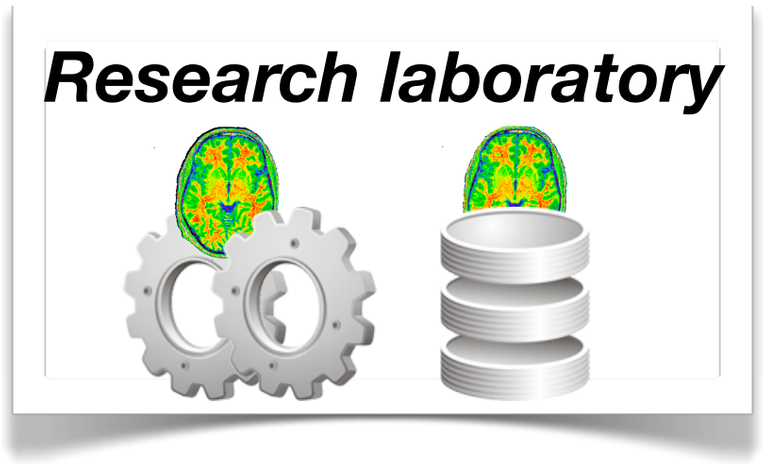
Translational research & e-Science

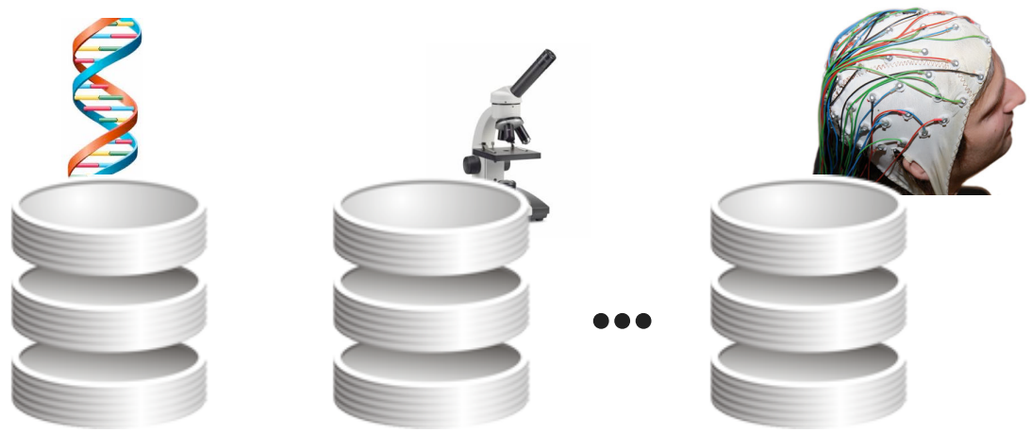


Translational research & e-Science

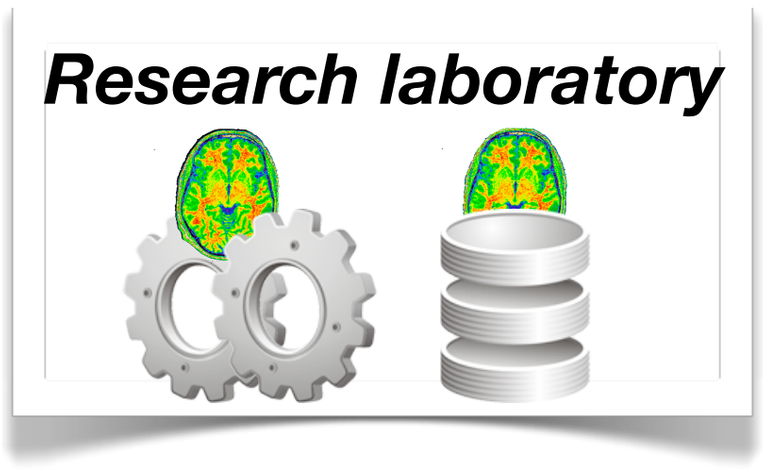


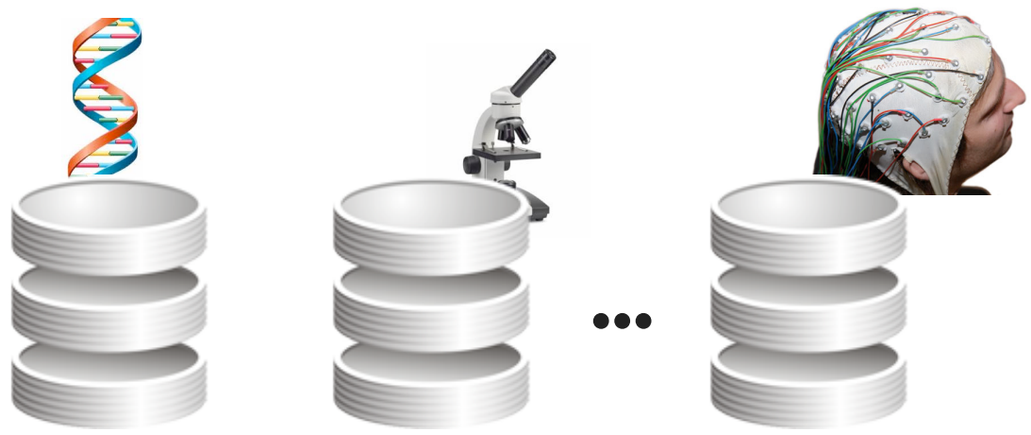
Translational research & e-Science





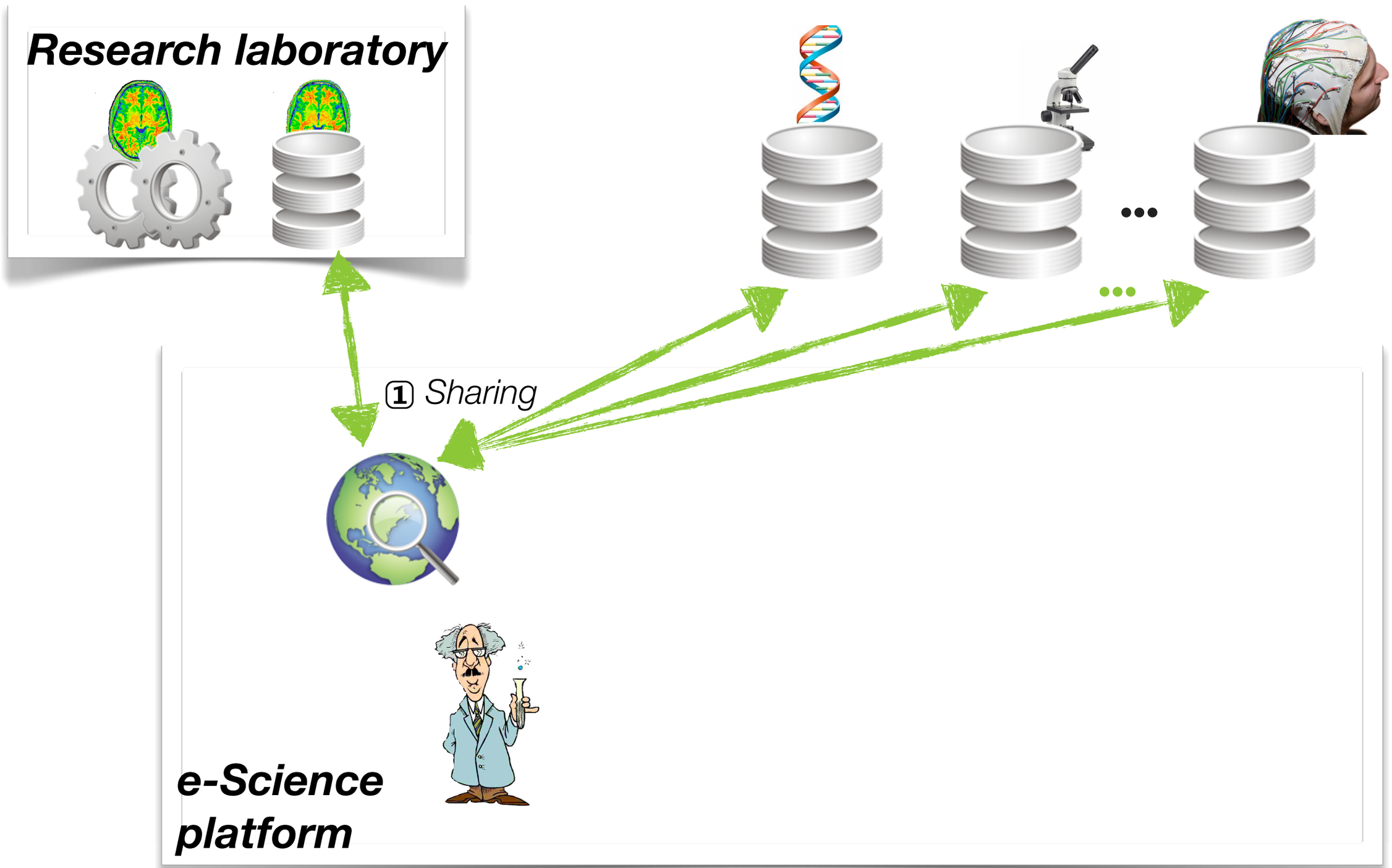
Translational research & e-Science



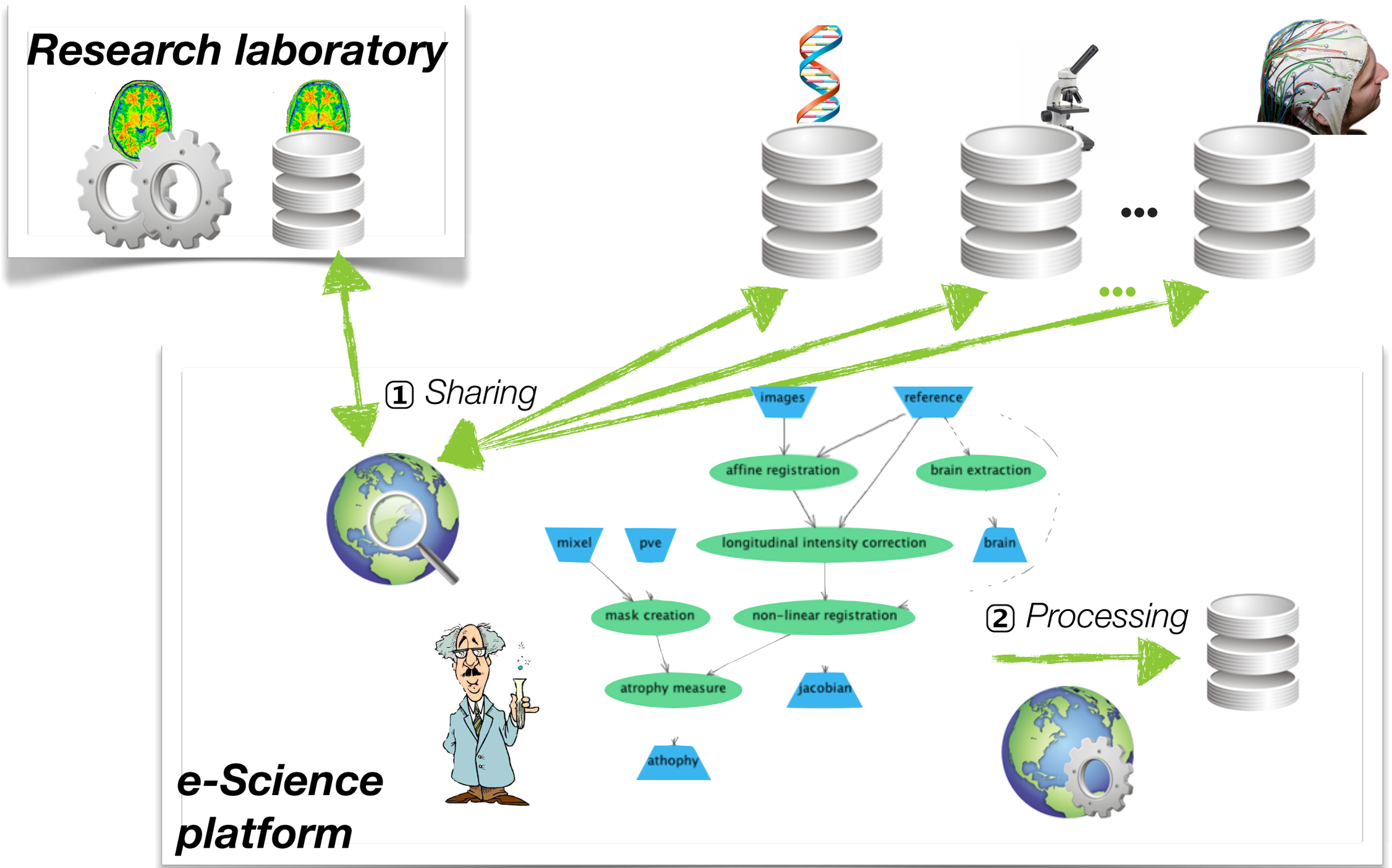


e-Science platform

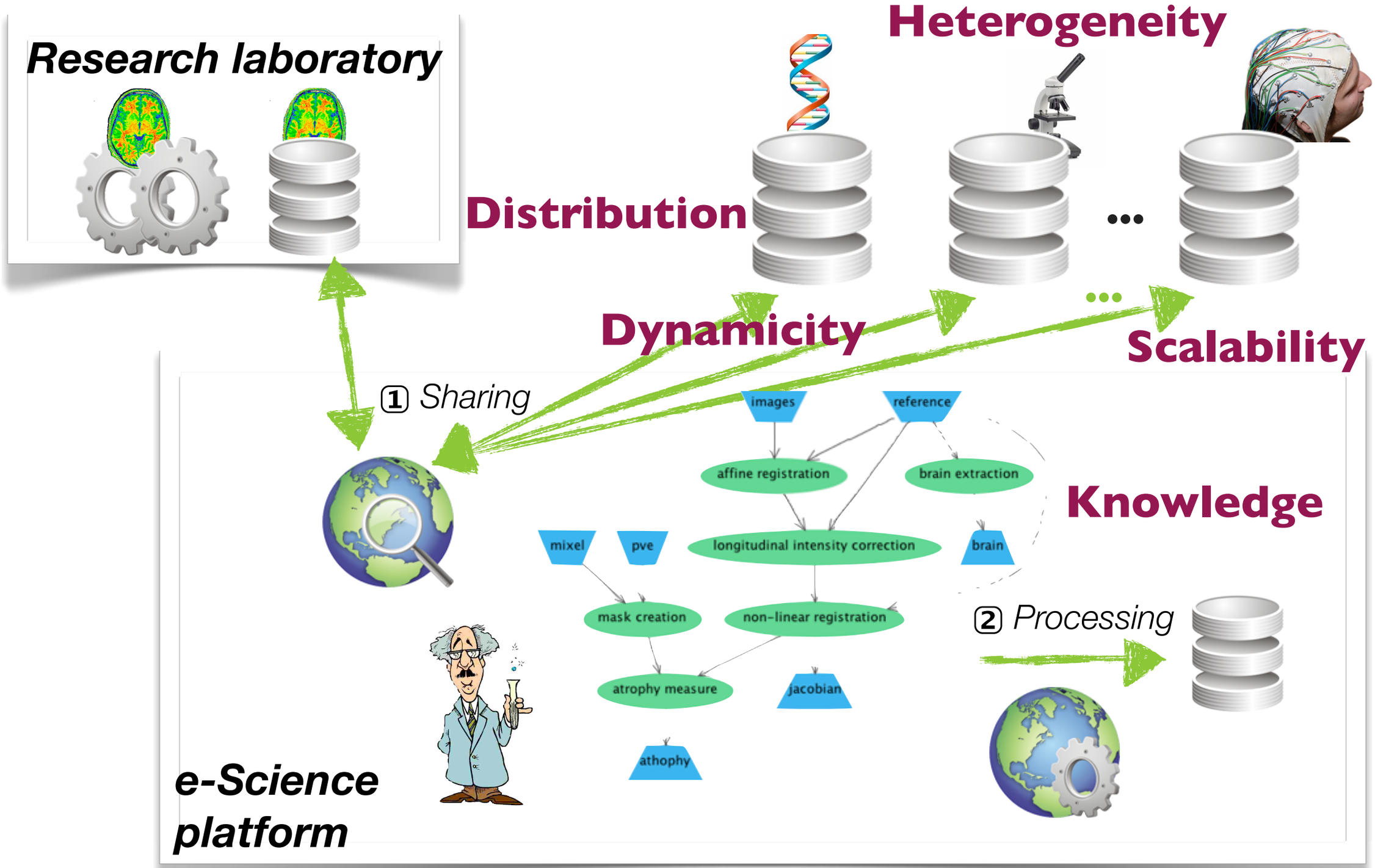
Translational research & e-Science



Translational research & e-Science



Translational research & e-Science



Challenges & Hypothesis

- **Questions:**
  - **Scalability/Distribution:** how to efficiently search over large distributed data sources ?
  - **Dynamicity/Heterogeneity:** how to cope with legacy/non-relocatable data ? how to dynamically combine several independent data sources ?
  - **Knowledge:** how to share/search for data and processing tools with high expressivity ? better results interpretation ?

Challenges & Hypothesis

- **Questions:**
  - **Scalability/Distribution:** how to efficiently search over large distributed data sources ?
  - **Dynamicity/Heterogeneity:** how to cope with legacy/non-relocatable data ? how to dynamically combine several independent data sources ?
  - **Knowledge:** how to share/search for data and processing tools with high expressivity ? better results interpretation ?

- **Hypothesis:**
  - ***H1***: *Domain ontologies*
  - ***H2***: *Data sources are distributed and autonomous*
  - ***H3***: *e-Science platforms allow to share & produce scientific resources*

Challenges & Hypothesis

- **Questions:**
  - **Scalability/Distribution:** how to efficiently search over large distributed data sources ?
  - **Dynamicity/Heterogeneity:** how to cope with legacy/non-relocatable data ? how to dynamically combine several independent data sources ?
  - **Knowledge:** how to share/search for data and processing tools with high expressivity ? better results interpretation ?

- **Hypothesis:**
  - ***H1****: Domain ontologies*
  - ***H2****: Data sources are distributed and autonomous*
  - ***H3****: e-Science platforms allow to share & produce scientific resources*

- **Scientific areas**
  - **Knowledge engineering:** reasoning on semantic description of data & processing tools
  - **e-Science:** computing infra. to process/ share/re-purpose scientific resources

Challenges & Hypothesis

- **Questions:**
  - **Scalability/Distribution:** how to efficiently search over large distributed data sources ?
  - **Dynamicity/Heterogeneity:** how to cope with legacy/non-relocatable data ? how to dynamically combine several independent data sources ?
  - **Knowledge:** how to share/search for data and processing tools with high expressivity ? better results interpretation ?

- **Hypothesis:**
  - ***H1***: *Domain ontologies*
  - ***H2***: *Data sources are distributed and autonomous*
  - ***H3***: *e-Science platforms allow to share & produce scientific resources*

- **Scientific areas**
  - **Knowledge engineering:** reasoning on semantic description of data & processing tools
  - **e-Science:** computing infra. to process/ share/re-purpose scientific resources

semantic e-Science → reducing "time-to-discovery"

Thesis Objectives

Thesis Objectives

Coherent **sharing** and **production** of **distributed knowledge** in Life-Science:

- Knowledge sharing: coping with semantic data **volume**, **distribution**, **heterogeneity**

- Knowledge production: **extracting meaningful & long-term data** from large & technical datasets

Main contributions

1. Knowledge base federation

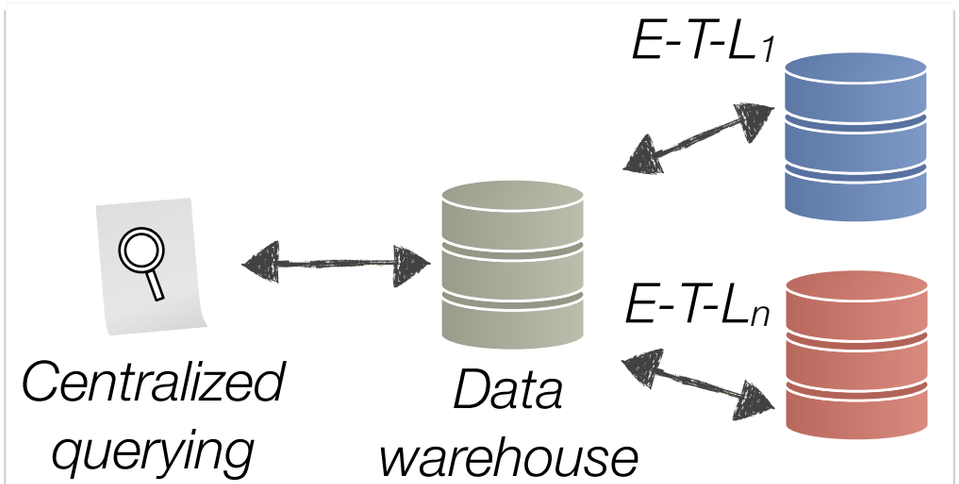
- Transparent, efficient, and expressive semantic federated querying
- Abstract Knowledge Graphs
- [Web Intelligence'12] [IC'12 workshop] [MICCAI'12 workshop]

2. Semantic Workflows

- Characterization of semantically annotated services (Nature and Role)
- Semantic experiment summaries
- [KEOD'11] [IC'10 workshop] [TMI'13] [CBMS'11]

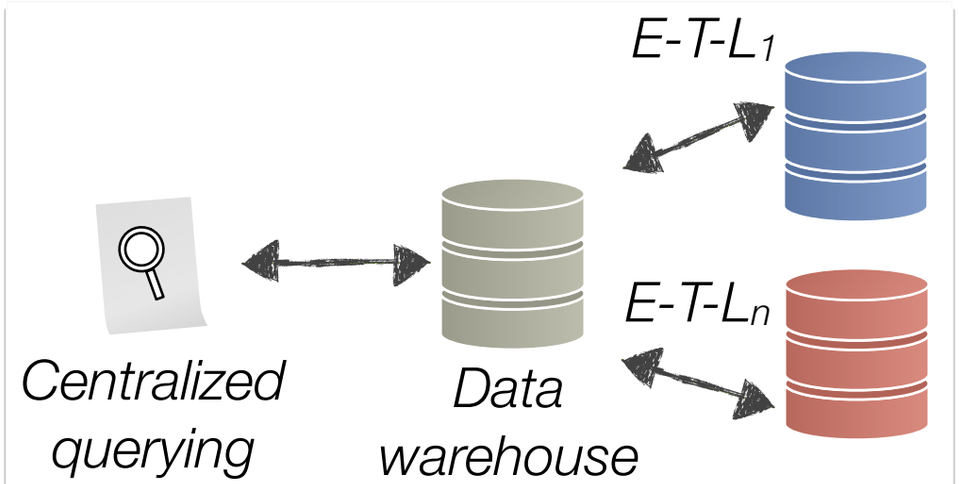
E-Science ① : data integration

E-Science ① : data integration

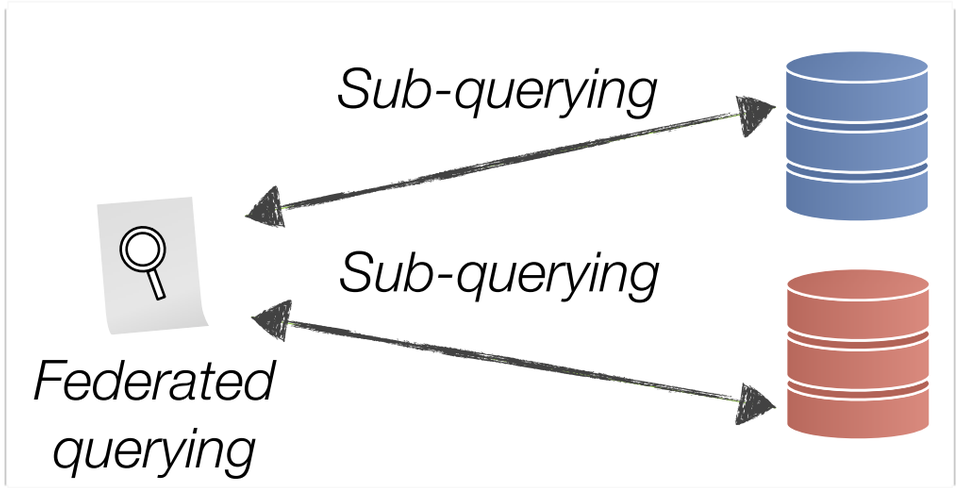


- **Materialized** Data Integration
  - **E**xtract - **T**ransform - **L**oad
    - ⊕ Efficiency
    - ⊖ Scalability
    - ⊖ Dynamicity
    - ⊖ Hardly relocatable data ?

E-Science ① : data integration



- **Materialized** Data Integration
  - **E**xtract - **T**ransform - **L**oad
    - ⊕ Efficiency
    - ⊖ Scalability
    - ⊖ Dynamicity
    - ⊖ Hardly relocatable data ?



- **Virtualized** Data Integration
  - **D**istributed **Q**uery **P**rocessing
    - ⊖ Efficiency
    - ⊕ Scalability (Load/Volume)
    - ⊕ Dynamicity
    - ⊕ Data kept at source

E-Science ① : distributed semantic querying

	DARQ	Splendid	SemWiq	Sparql-DQP	FedX	KGRAM
Distribution	+	+	+	+	+	±
Performance	-	-	?	?	++	?
Heterogeneity	-	-	++	+	-	+
Dynamicity	-	-	-	-	++	+
Expressivity	-	-	-	-	+	++

E-Science ① : distributed semantic querying

	DARQ	Splendid	SemWiq	Sparql-DQP	FedX	KGRAM
Distribution	+	+	+	+	+	-
Performance	-	-	?	?	+ +	?
Heterogeneity	-	-	+ +	+	-	+
Dynamicity	-	-	-	-	+ +	+
Expressivity	-	-	-	-	+	+ +

- **Missing** expressivity (subset of SPARQL)
  - Only SELECT queries on Basic Graph Patterns, no PATH expressions, no bound subjects for SemWiq, etc.

E-Science ① : distributed semantic querying

	DARQ	Splendid	SemWiq	Sparql-DQP	FedX	KGRAM	KGRAM-DQP
Distribution	+	+	+	+	+	-	+
Performance	-	-	?	?	+ +	?	+
Heterogeneity	-	-	+ +	+	-	+	+
Dynamicity	-	-	-	-	+ +	+	+
Expressivity	-	-	-	-	+	+ +	+ +

- **Missing** expressivity (subset of SPARQL)
  - Only SELECT queries on Basic Graph Patterns, no PATH expressions, no bound subjects for SemWiq, etc.

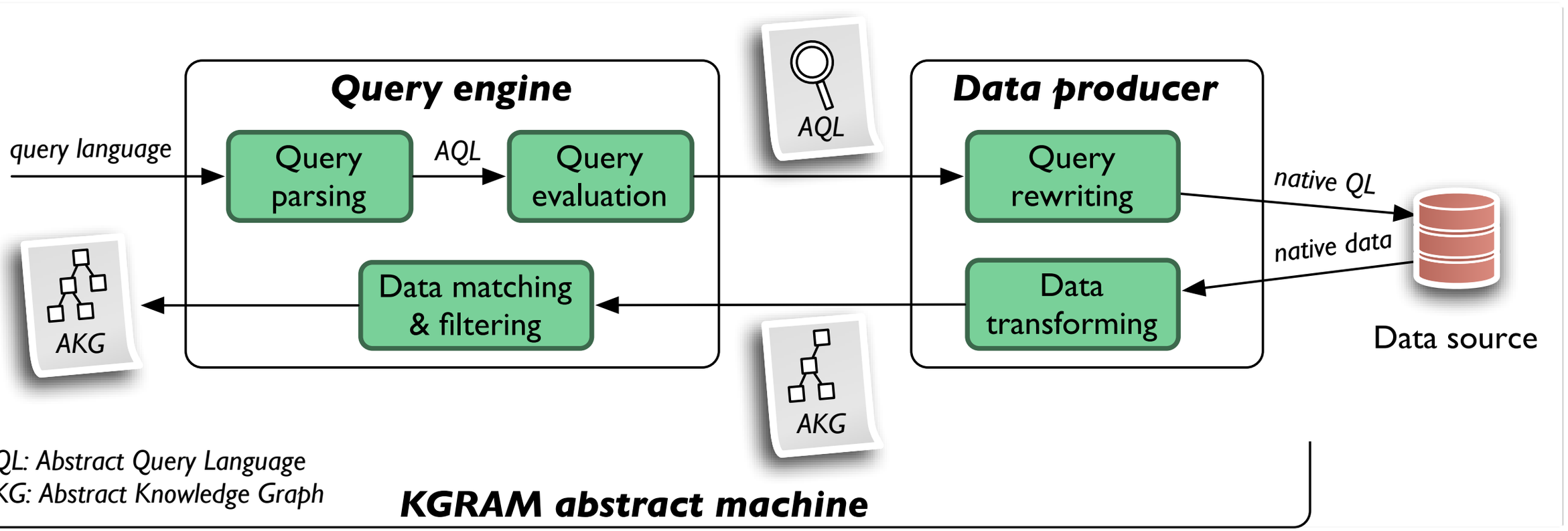
Balancing Expressivity & Performance

E-Science ① : semantic data handling with KGRAM

- Representing, querying and reasoning on Knowledge Graphs
- **Generic** engine
  - **Expressivity:** SPARQL 1.1 compliant
  - **Versatility:** several data models (RDF, XML, SQL)
  - **Reasoning:** RDFS entailments + Inference rules

E-Science ① : semantic data handling with KGRAM

- Representing, querying and reasoning on Knowledge Graphs
- **Generic** engine
  - **Expressivity:** SPARQL 1.1 compliant
  - **Versatility:** several data models (RDF, XML, SQL)
  - **Reasoning:** RDFS entailments + Inference rules



E-Science ② : scientific workflows

- **Semantic workflow (WF) environments**
  - METEOR-S ; Taverna/FETA ; BioCatalogue ; BioMOBY
  - target WF **design/sharing**

- **WF results interpretation** through **Provenance** standards (Provenir, OPM → PROV-*)

	e-BioInfra	NeuGrid	RDFProv	ProvBase	Linked Provenance Data	Wings/ Pegasus	PaCE	Taverna/ Janus
Standards	+	-	-	-	+	-	+	+
Scalability	+	+	+	+ +	-	+	-	-
Linked Data approach	-	-	+	+	+	+/-	+	+
Domain knowledge	-	-	-	-	+	+	+	+

PROV-O published as a W3C Candidate Recommendation (11 December 2012)

E-Science ② : scientific workflows

- **Semantic workflow (WF) environments**
  - METEOR-S ; Taverna/FETA ; BioCatalogue ; BioMOBY
  - target WF **design/sharing**

- **WF results interpretation** through **Provenance** standards (Provenir, OPM → PROV-*)

	e-BioInfra	NeuGrid	RDFProv	ProvBase	Linked Provenance Data	Wings/ Pegasus	PaCE	Taverna/ Janus	NeuSem Store
Standards	+	-	-	-	+	-	+	+	+
Scalability	+	+	+	+ +	-	+	-	-	+
Linked Data approach	-	-	+	+	+	+/-	+	+	+
Domain knowledge	-	-	-	-	+	+	+	+	+

PROV-O published as a W3C Candidate Recommendation (11 December 2012)

Contribution 1

Knowledge **Sharing** (for e-Science platforms)

Efficient & expressive **sharing** of **knowledge graphs**

Efficient & expressive **sharing** of **knowledge graphs**

- **Objectives**
  - **Transparent federated** semantic engine
    - **Heterogeneity + Dynamicity**
  - Balancing **expressivity** and **performance**
    - **Distribution + Scalability + Knowledge**

Efficient & expressive **sharing** of **knowledge graphs**

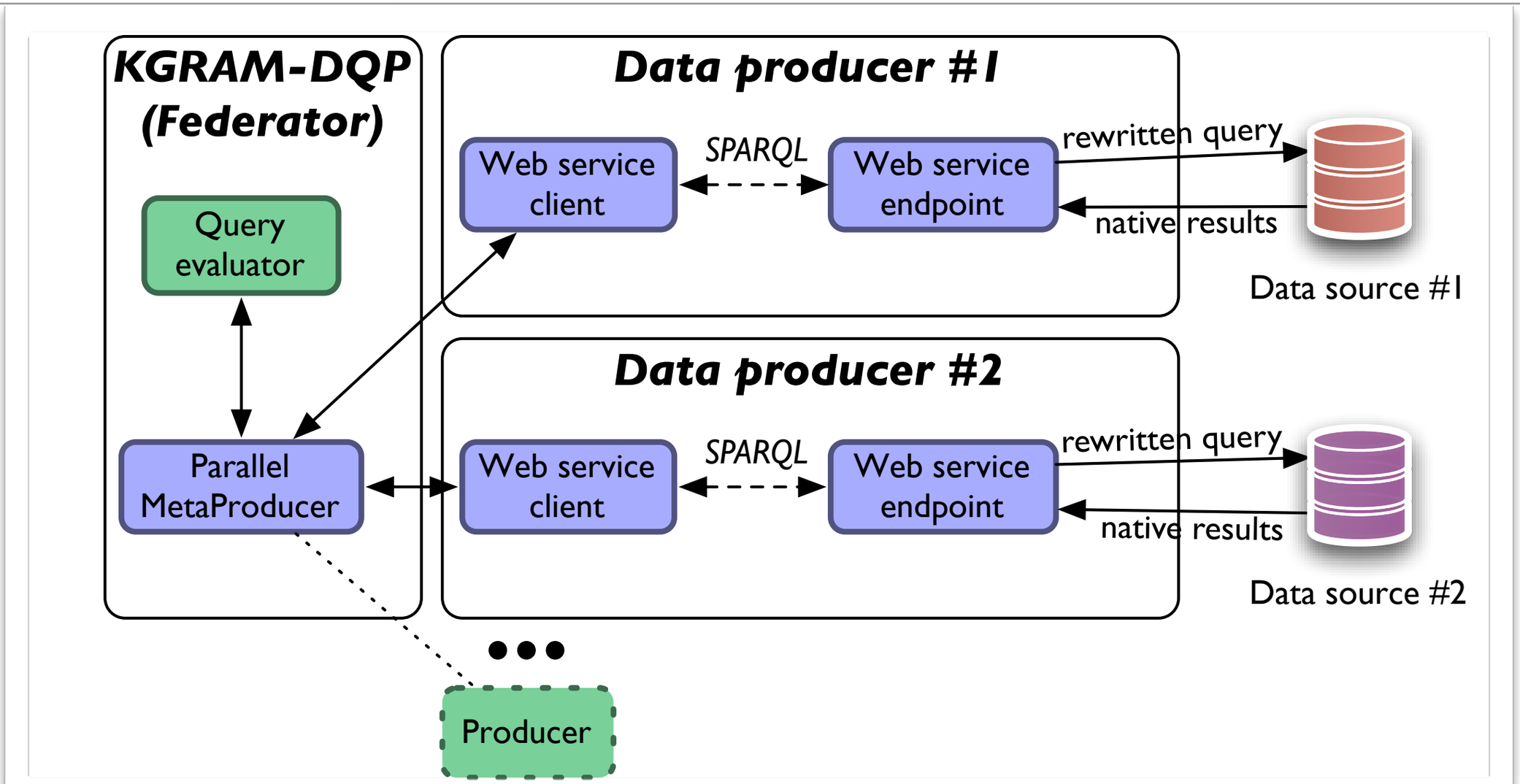
- **Objectives**
  - **Transparent federated** semantic engine
    - **Heterogeneity + Dynamicity**
  - Balancing **expressivity** and **performance**
    - **Distribution + Scalability + Knowledge**

Efficient & expressive **sharing** of **knowledge graphs**

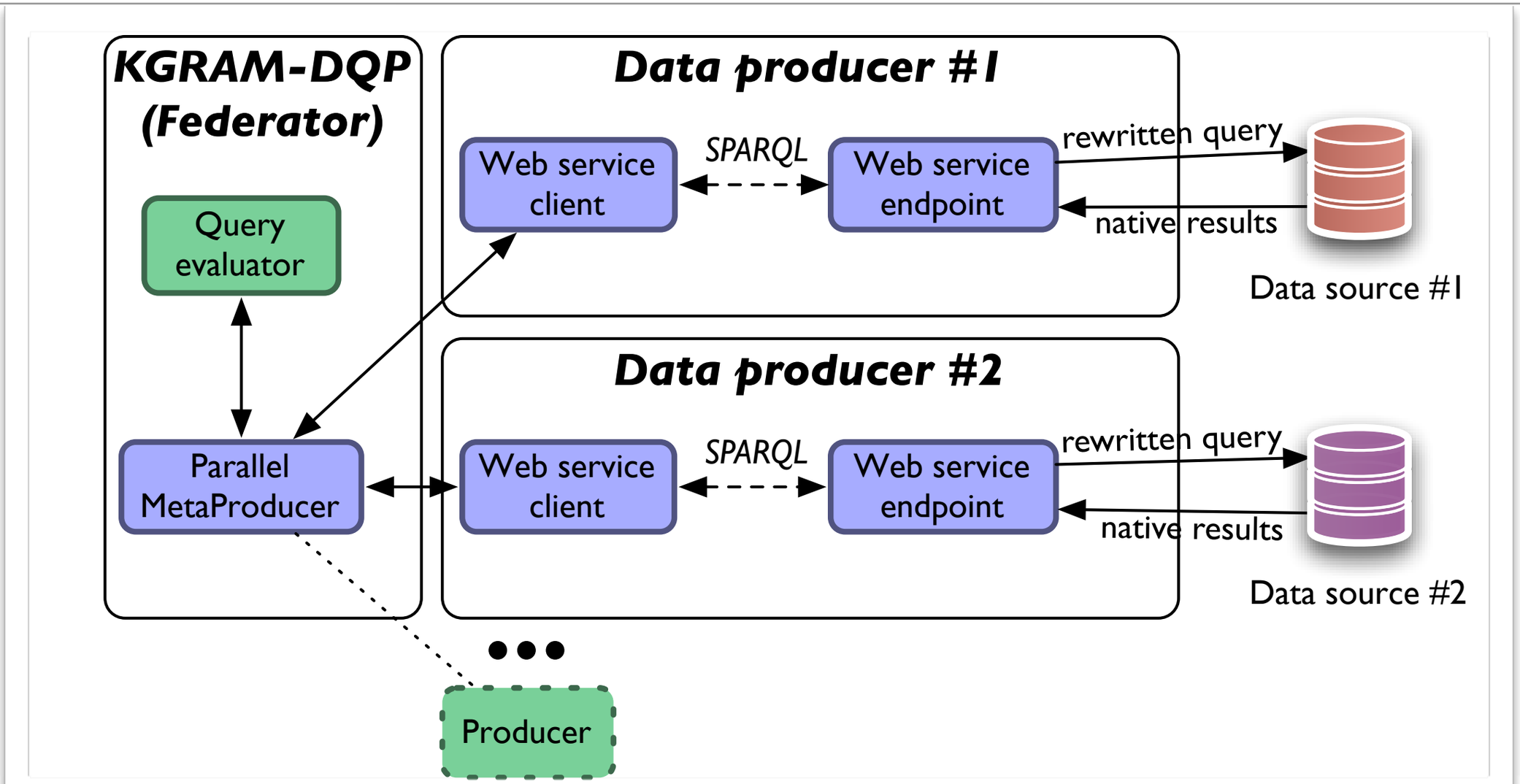
- **Objectives**
  - **Transparent federated** semantic engine
    - **Heterogeneity + Dynamicity**
  - Balancing **expressivity** and **performance**
    - **Distribution + Scalability + Knowledge**

- **Methods**
  - Abstract Knowledge Graphs
  - Distributed Query Processing techniques
  - Static and dynamic optimization

KGRAM-DQP: **d**istributed **q**uery **p**rocessing



KGRAM-DQP: **d**istributed **q**uery **p**rocessing



⊖ **Cost** of network communication

- Distributed query processing → **performance**
  - Service **parallelism** / **optimizations**

KGRAM-DQP: parallel evaluation

Data: $Producers$ the set of SPARQL endpoints,
$EdgeReq$ the set of edge requests forming the SPARQL query,
$scheduler$ a thread pool allowing parallel execution.
Result: $Results$ the set of SPARQL results.

```
1 foreach (e ∈ EdgeReq) do
2     foreach (p ∈ Producers) do in parallel
3         scheduler.submit(p.getEdges(e)) ;
4     wait for scheduler ;
5     foreach (task ∈ scheduler.getFinished()) do
6         Results ← task.getResults() ;
```

(a) Synch. barrier
(b) Pipelining

Static optimization: pushing applicable FILTERs

- **Filtering** irrelevant results the **sooner** (lighter network communications)
  - ➡ add FILTER to each single triple pattern (if applicable)

Input SPARQL query

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?x ?name ?date WHERE {
        ?x foaf:name ?name .
        ?x dbpedia:birthDate ?date .
        FILTER (CONTAINS (?name, 'Bobby␣A'))
}
```

Static optimization: pushing applicable FILTERs

- **Filtering** irrelevant results the **sooner** (lighter network communications)
  - ➡ add FILTER to each single triple pattern (if applicable)

Input SPARQL query

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?x ?name ?date WHERE {
        ?x foaf:name ?name .
        ?x dbpedia:birthDate ?date .
        FILTER (CONTAINS (?name, 'Bobby␣A'))
}
```

Rewritten sub-query

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
CONSTRUCT {?x foaf:name ?name} WHERE {
        ?x foaf:name ?name.
}
```

Static optimization: pushing applicable FILTERs

- **Filtering** irrelevant results the **sooner** (lighter network communications)
  - ➡ add FILTER to each single triple pattern (if applicable)

Input SPARQL query

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?x ?name ?date WHERE {
        ?x foaf:name ?name .
        ?x dbpedia:birthDate ?date .
        FILTER (CONTAINS (?name, 'Bobby␣A'))
}
```

Rewritten sub-query

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
CONSTRUCT {?x foaf:name ?name} WHERE {
        ?x foaf:name ?name.
}
```

Optimized sub-query

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
CONSTRUCT {?x foaf:name ?name} WHERE {
        ?x foaf:name ?name.
        FILTER (CONTAINS (?name, 'Bobby␣A'))
}
```

Dynamic optimization: pushing values

- **Avoid re-evaluation** by exploiting intermediate results (communication of already known values saved) *[Bind joins]*

  ➡ Replacing variables by their known values for each single triple pattern.

Input SPARQL query

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?x ?name ?date WHERE {
        ?x foaf:name ?name .
        ?x dbpedia:birthDate ?date .
        FILTER (CONTAINS (?name, 'Bobby␣A'))
}
```

Intermediate result

?x = http://dbpedia.org/resource/Bobby_Abel

Dynamic optimization: pushing values

- **Avoid re-evaluation** by exploiting intermediate results (communication of already known values saved) *[Bind joins]*

  ➡ Replacing variables by their known values for each single triple pattern.

Input SPARQL query

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?x ?name ?date WHERE {
        ?x foaf:name ?name .
        ?x dbpedia:birthDate ?date .
        FILTER (CONTAINS (?name, "Bobby A"))
}
```

Rewritten sub-query

```
PREFIX dbpedia: <http://dbpedia.org/ontology/>
CONSTRUCT {
        ?x dbpedia:birthDate ?date
} WHERE {
        ?x dbpedia:birthDate ?date
}
```

Intermediate result

?x = http://dbpedia.org/resource/Bobby_Abel

Dynamic optimization: pushing values

- **Avoid re-evaluation** by exploiting intermediate results (communication of already known values saved) *[Bind joins]*
  - ➡ Replacing variables by their known values for each single triple pattern.

Input SPARQL query

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?x ?name ?date WHERE {
        ?x foaf:name ?name .
        ?x dbpedia:birthDate ?date .
        FILTER (CONTAINS (?name, 'Bobby_A'))
}
```

Rewritten sub-query

```
PREFIX dbpedia: <http://dbpedia.org/ontology/>
CONSTRUCT {
        ?x dbpedia:birthDate ?date
} WHERE {
        ?x dbpedia:birthDate ?date
}
```

Intermediate result

?x = http://dbpedia.org/resource/Bobby_Abel

Optimized sub-query

```
PREFIX dbpedia: <http://dbpedia.org/ontology/>
CONSTRUCT {
        <http://dbpedia/resource/Bobby_Abel> dbpedia:birthDate ?date
} WHERE {
        <http://dbpedia/resource/Bobby_Abel> dbpedia:birthDate ?date
}
```

Experiment: large-scale **benchmarking** (1/2)

- **Objective:** performance assessment
- **Material and Methods**
  - FedBench from the FedX team (50M triples ; 7 life-science SPARQL queries)
  - Grid'5000 Computing Infrastructure
    - FedX + Fuseki endpoints
    - KGRAM-DQP

FedBench Life-Science datasets

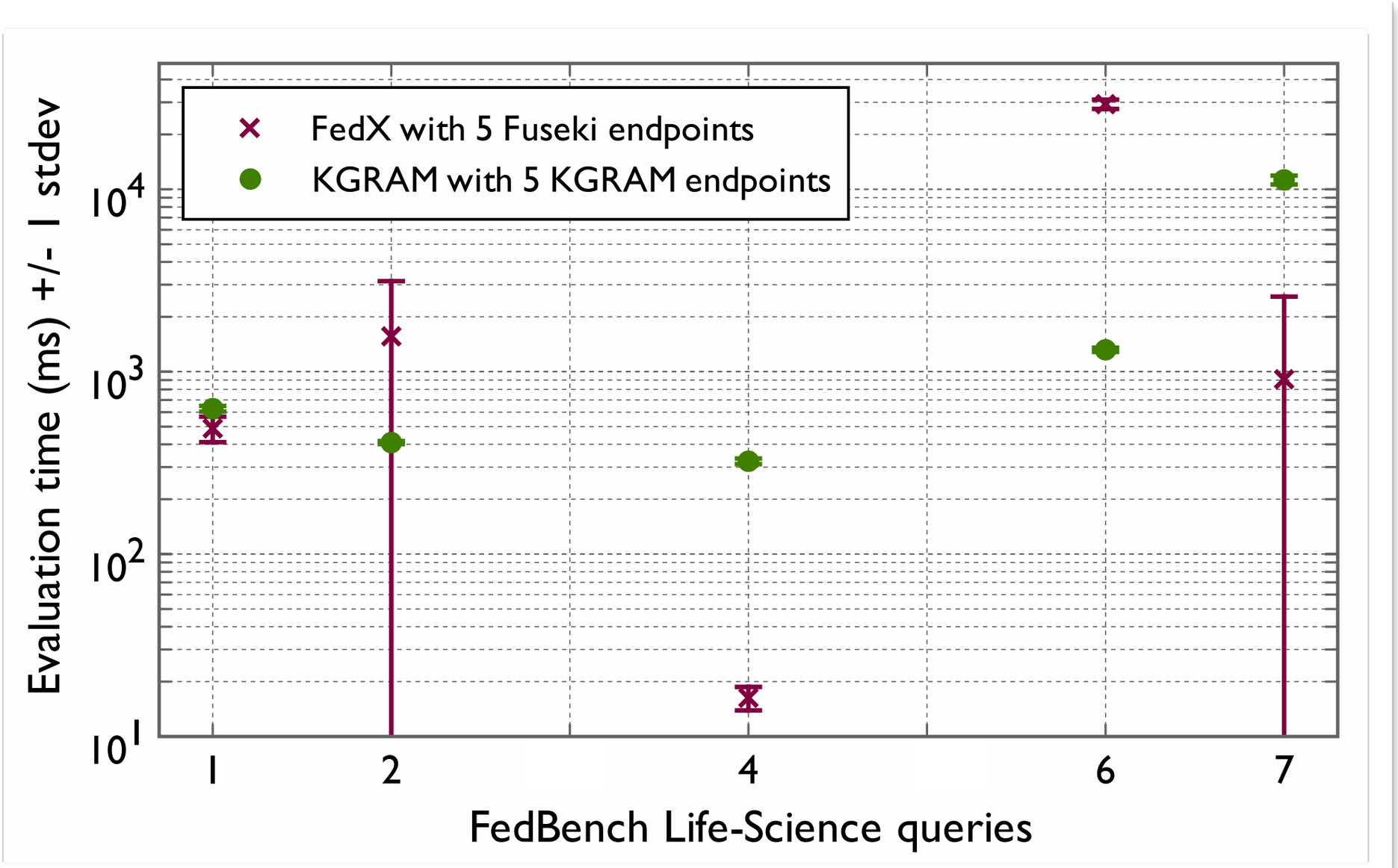
Data source	*Linked Data collection*	*Size (triples)*
#1	ChEBI	7.3M
#2	DBpedia sub-set #1	25.3M
#3	DBpedia sub-set #2	18.3M
#4	DrugBank	0.7M
#5	KEGG Drug	1M

FedBench Life-Science query #7

```
SELECT $drug $transform $mass WHERE {
        { $drug <http://www4.wiwiss.fu−berlin.de/drugbank/resource/drugbank/affectedOrganism>
                                                        'Humans_and_other_mammals'.
          $drug <http://www4.wiwiss.fu−berlin.de/drugbank/resource/drugbank/casRegistryNumber> $cas .
          $keggDrug <http://bio2rdf.org/ns/bio2rdf#xRef> $cas .
          $keggDrug <http://bio2rdf.org/ns/bio2rdf#mass> $mass
              FILTER ( $mass > '5' )
        }
        OPTIONAL {$drug <http://www4.wiwiss.fu−berlin.de/drugbank/resource/drugbank/biotransformation>
            $transform .}
}
```

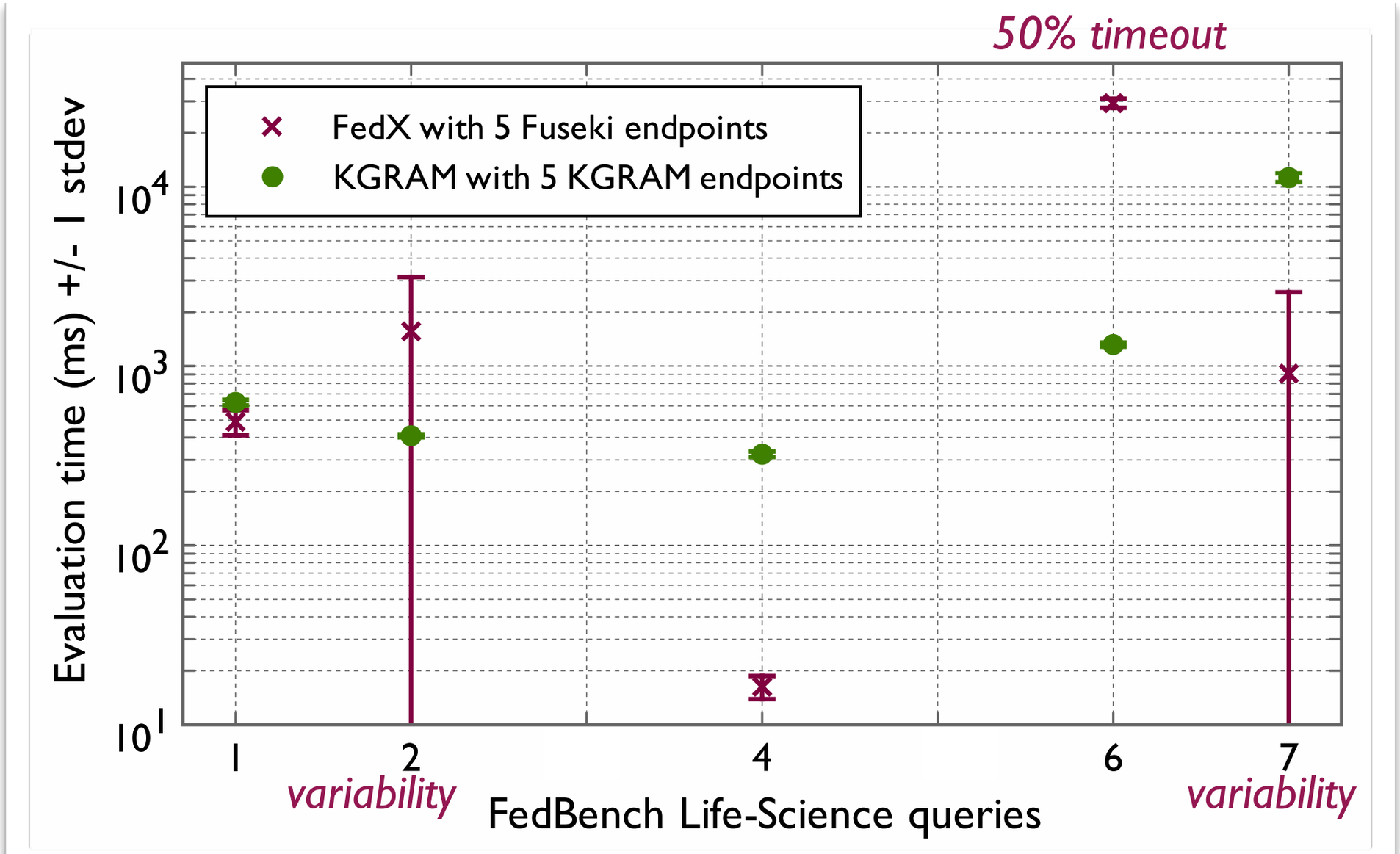
Experiment: large-scale **benchmarking** (2/2)

- Real distributed computing infrastructure
- Mean evaluation time over 10 runs



Experiment: large-scale **benchmarking** (2/2)

- Real distributed computing infrastructure
- Mean evaluation time over 10 runs



Highlights & short-term perspectives

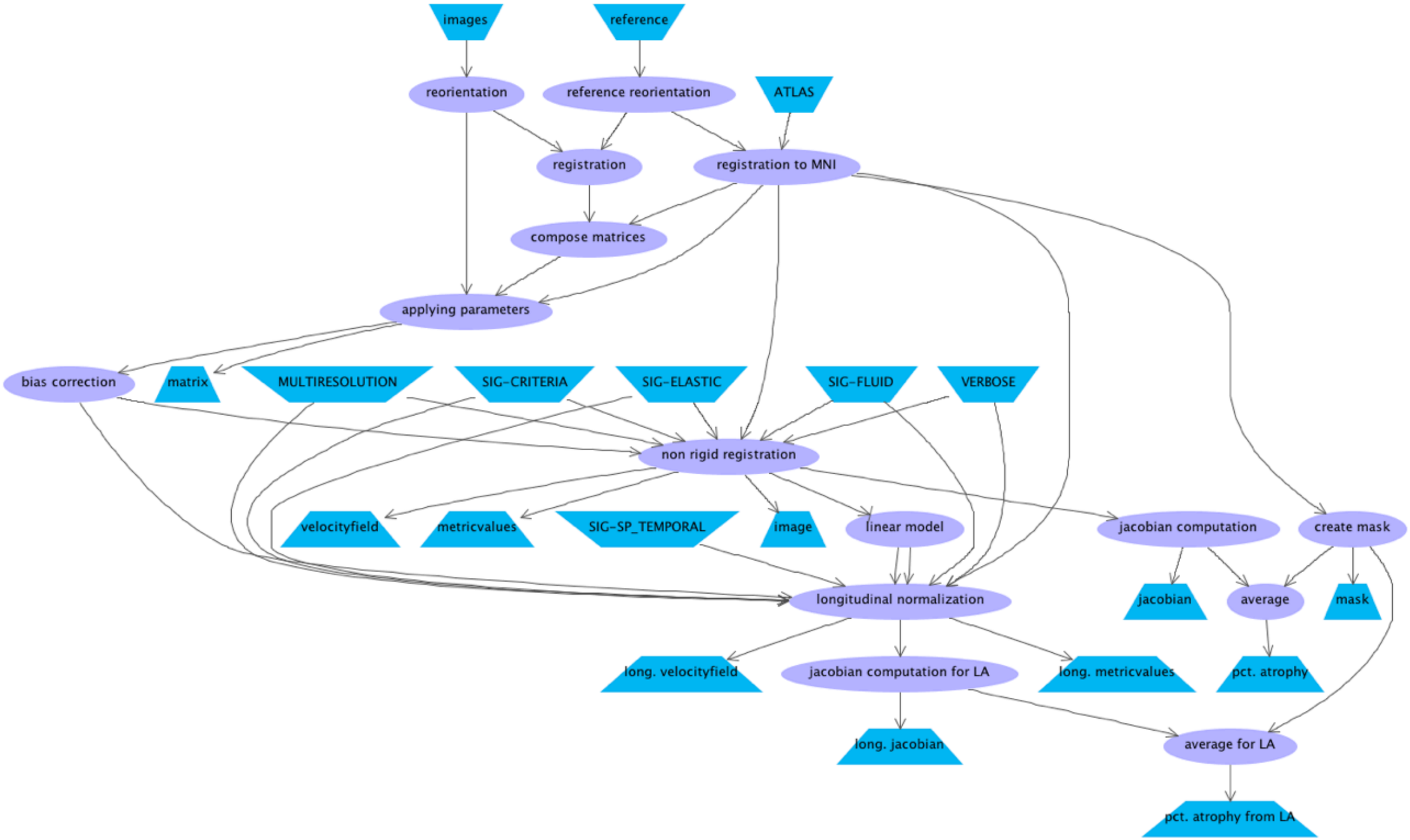
- **Highlights**
  - Transparent federated semantic querying **[Distribution /Dynamicity]**
    - No prior knowledge on data source content
  - Performances between DARQ / Splendid and FedX **[Scalability]**
  - Expressive approach: SPARQL 1.1 support (Optional, Negation, Property path, aggregates) **[Knowledge]**

- **Short-term perspectives**
  - Coarse-grain DQP (dynamic triple pattern grouping in SERVICE clauses)
    - Prototype algorithm, but possibly ineffective (query planing) **[Scalability]**
  - Relational database mediation
    - Prototype SQL data producer in KGRAM-DQP **[Heterogeneity]**

Contribution 2

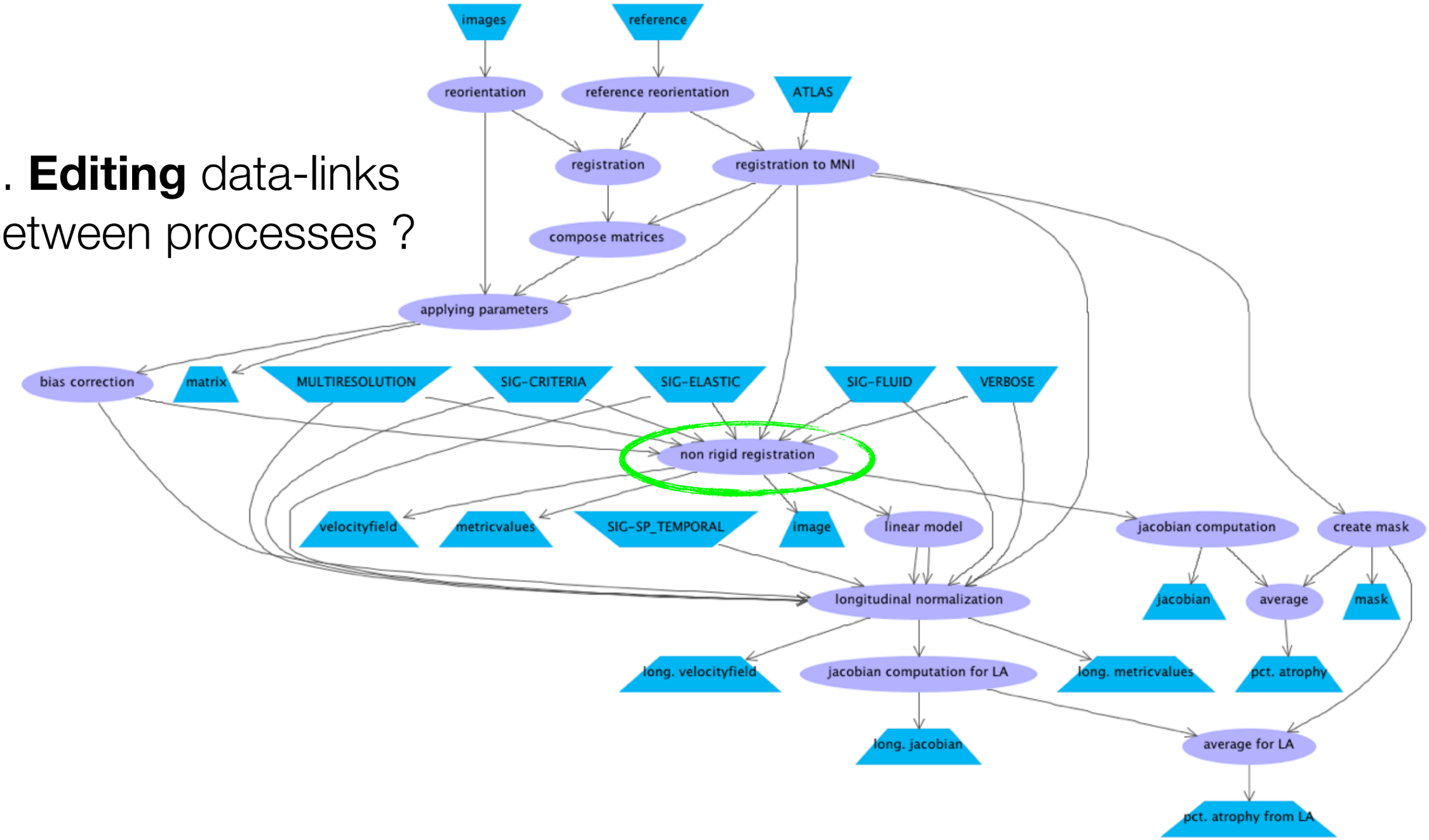
Knowledge **Production** (for e-Science platforms)

Scientific workflow issues



Scientific workflow issues

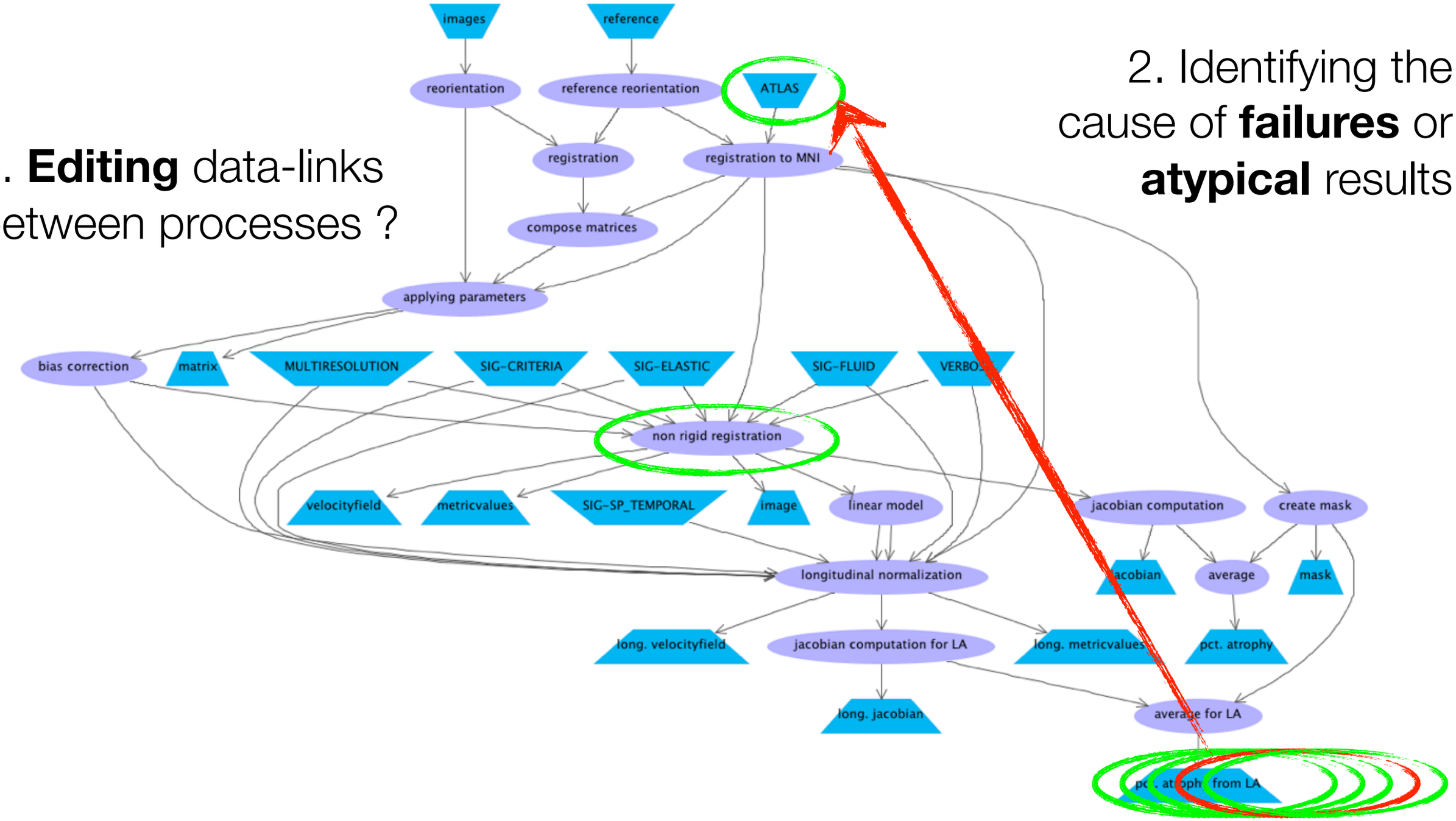
1. **Editing** data-links between processes ?



Scientific workflow issues

1. **Editing** data-links between processes ?

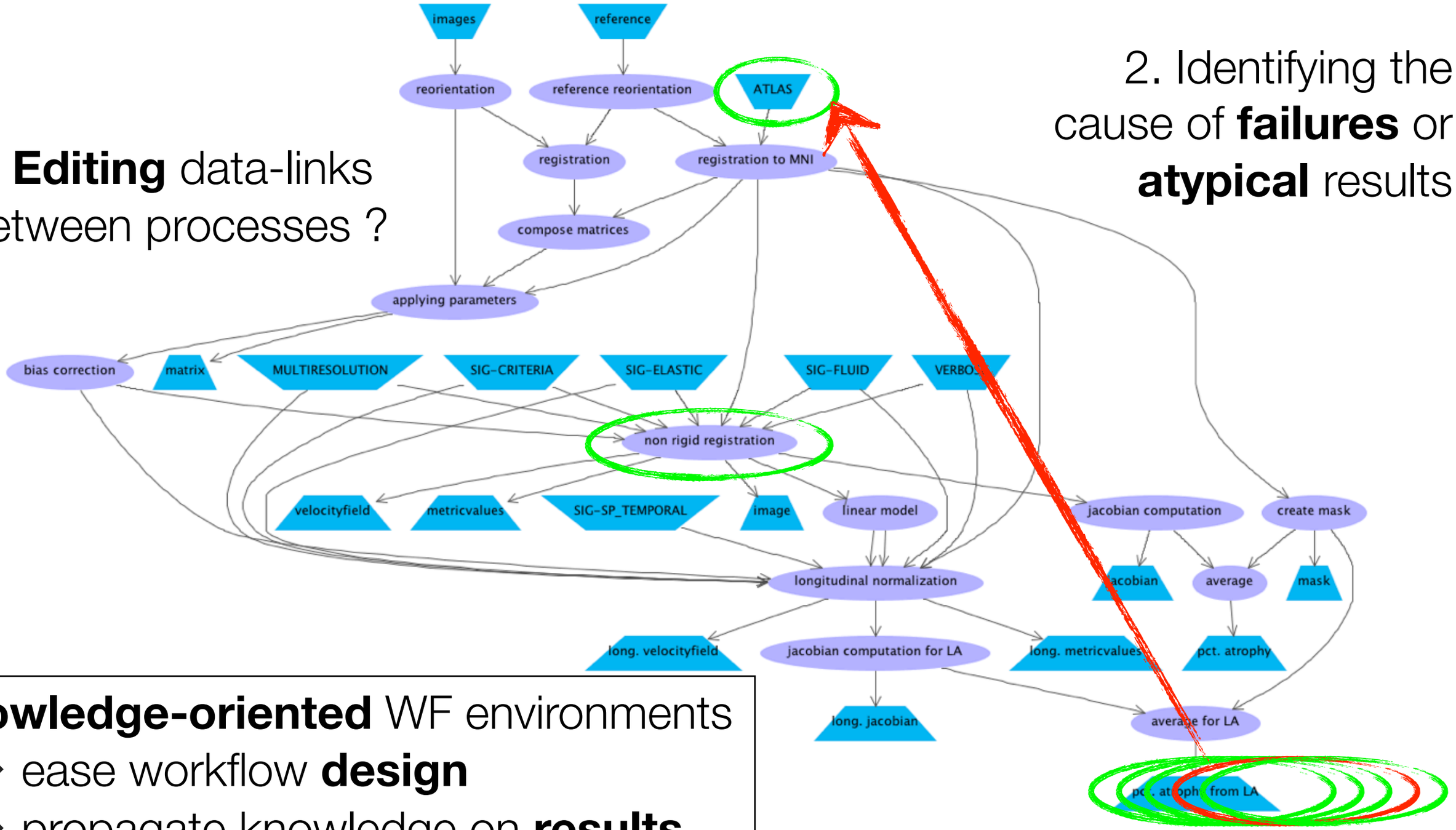
2. Identifying the cause of **failures** or **atypical** results



Scientific workflow issues

1. **Editing** data-links between processes ?

2. Identifying the cause of **failures** or **atypical** results



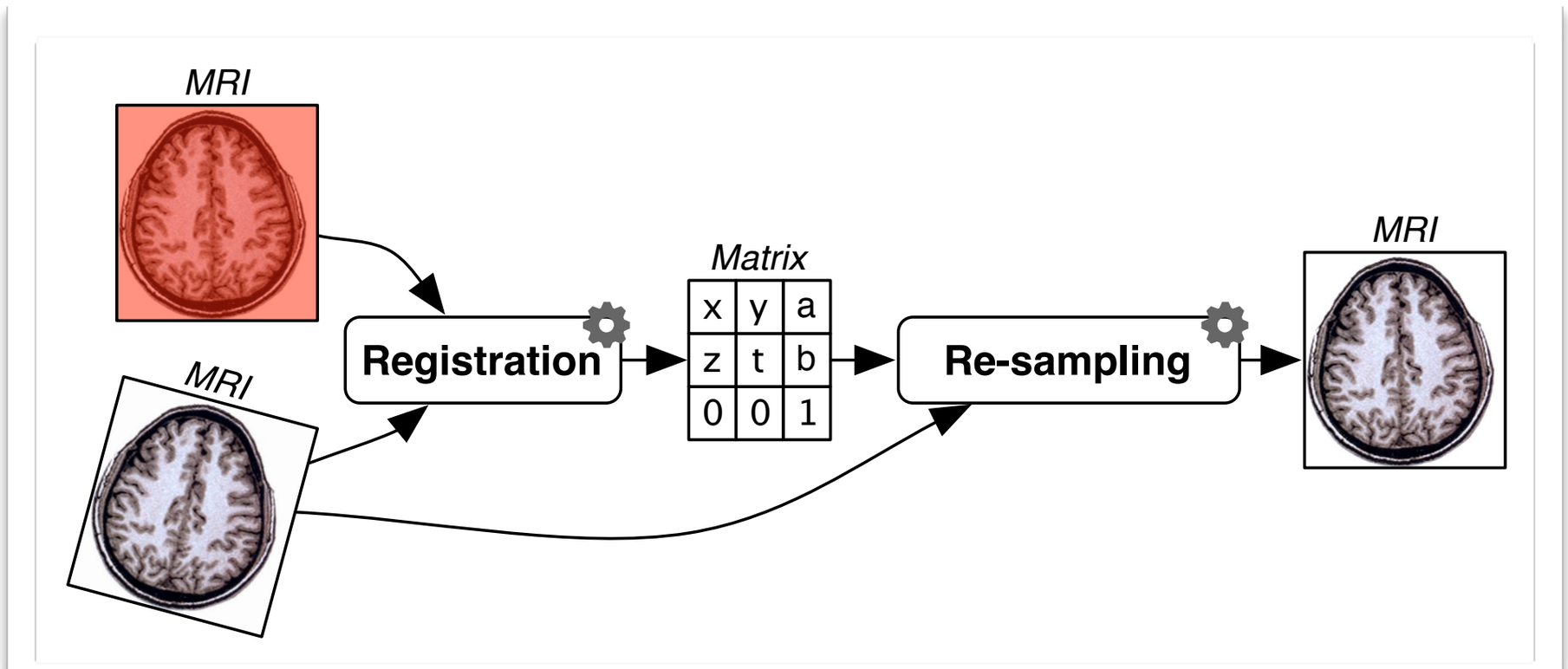
Knowledge-oriented WF environments

→ ease workflow **design**

→ propagate knowledge on **results**

Design issues

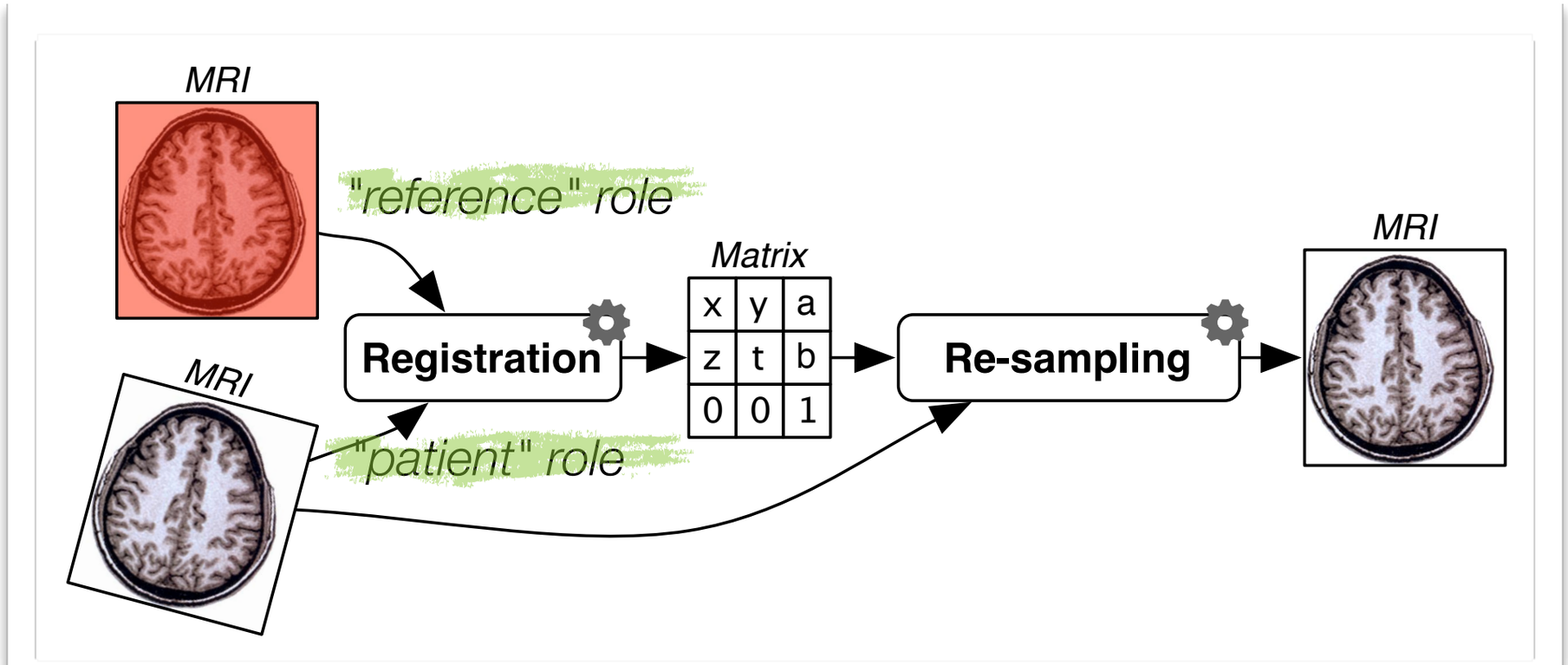
- Workflow **design issue**, close-up:



- Several **natures** of treatment or data, **not explicit** at technical level
- **Only** considering **nature**: **ambiguity**

Design issues

- Workflow **design issue**, close-up:

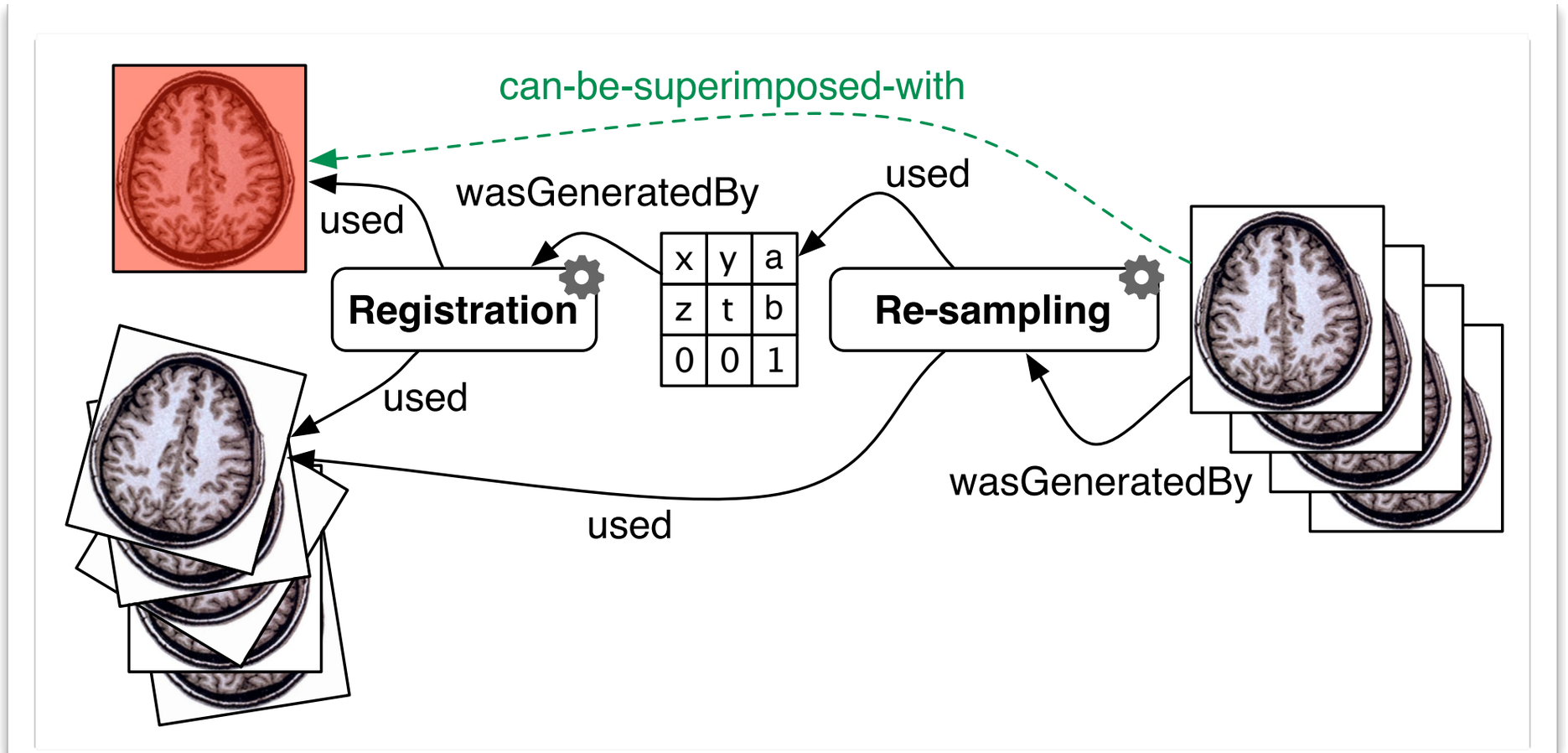


- Several **natures** of treatment or data, **not explicit** at technical level
- **Only** considering **nature**: **ambiguity**

➡ need for **Roles** to relate data to processing tools !

Runtime issues

- **Results exploitation** issue, close-up:



Need for **non-ambiguous** service annotations to **produce new domain-specific** statements

Issues & Objectives

Issues:

(i) How to **explicit** the **semantics** of data processing ?

(ii) How to **benefit** from this **knowledge** ...

- at experiment **design-time** ?
- at experiment **runtime** ?

Objectives:

(i) ↘ complexity of **designing** an e-Science experiment (workflow) ;

(ii) ↗ exploitation of **results** produced during data-intensive experiments.

Methods

- Several kinds of knowledge:
    - **Technical** knowledge (OWL-S, OPM) ;
    - **Domain** knowledge:
        1. **Nature** of data and services ;
        2. **Role** of data from the service point of view.

- Our contribution:
    1. Domain-specific **Role** Taxonomy: clarifying bindings between technical **service descriptions** and **domain concepts** ;
    2. Produce new **valuable knowledge** through **inferences** along platform exploitation.

- Supported by the **OntoNeuroLOG domain ontology** and the **OPM** provenance ontology.

Methods

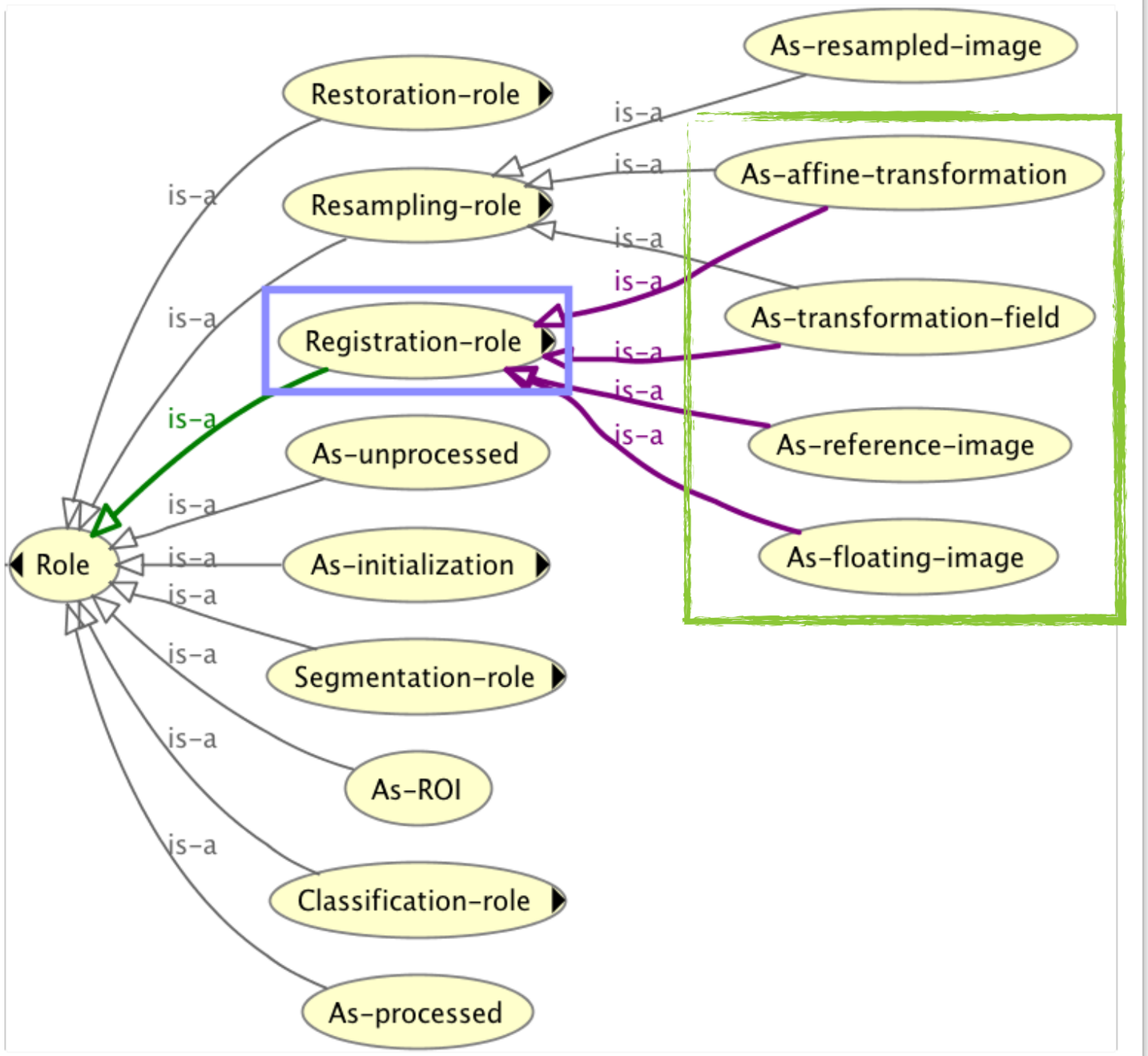
- Several kinds of knowledge:
    - **Technical** knowledge (OWL-S, OPM) ;
    - **Domain** knowledge:
        1. **Nature** of data and services ;
        2. **Role** of data from the service point of view.

- Our contribution:
    1. Domain-specific **Role** Taxonomy: clarifying bindings between technical **service descriptions** and **domain concepts** ;
    2. Produce new **valuable knowledge** through **inferences** along platform exploitation.

- Supported by the **OntoNeuroLOG domain ontology** and the **OPM** provenance ontology.

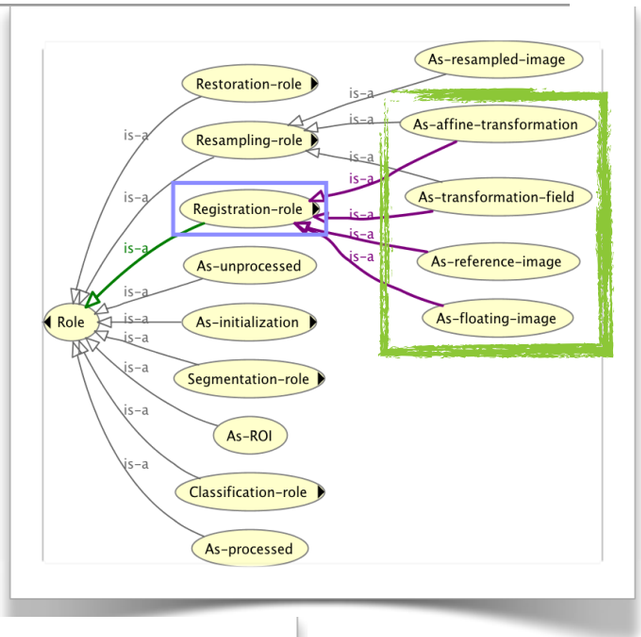
Neuroimaging **Role** taxonomy

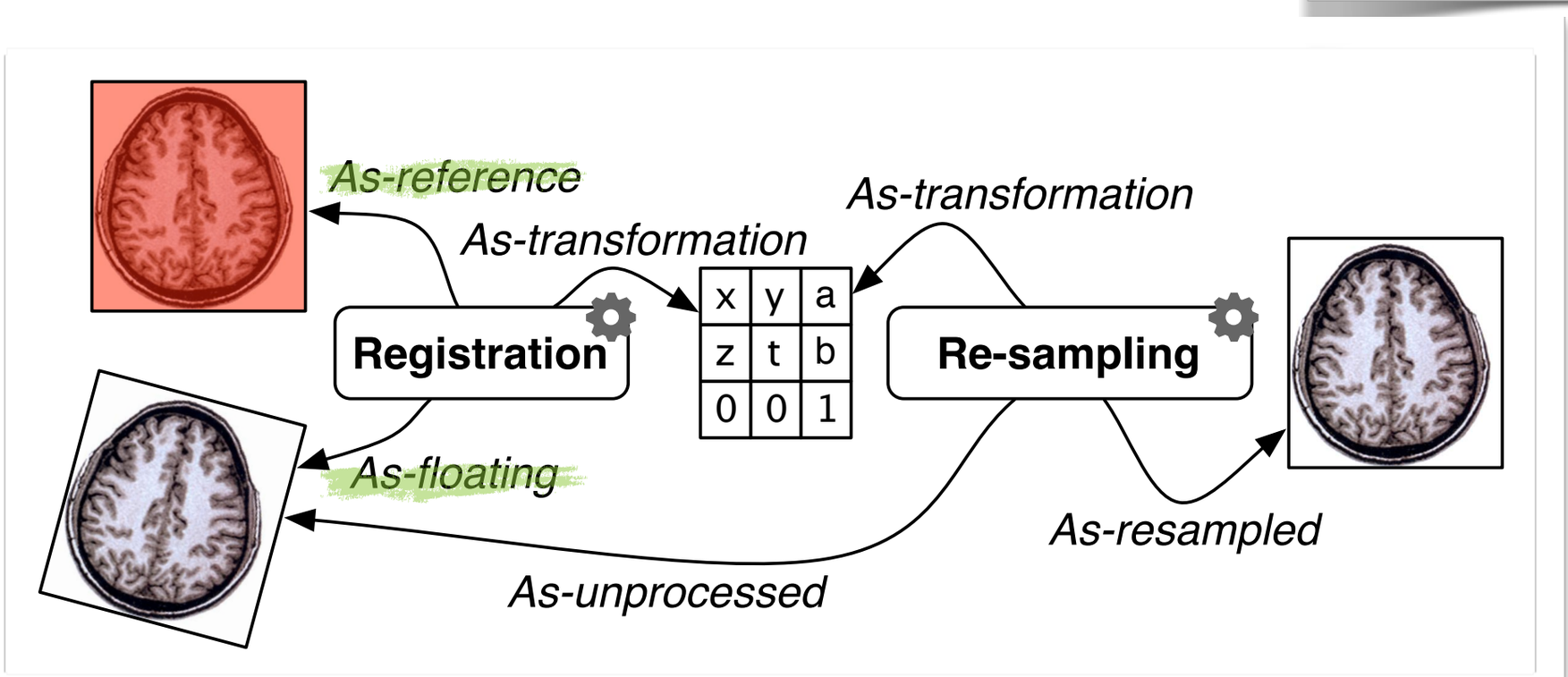
- Domain-specific extension of the **OPM Role class**
- **Roles** to **disambiguate** the annotation of **service parameters**.



Neuroimaging **Role** taxonomy

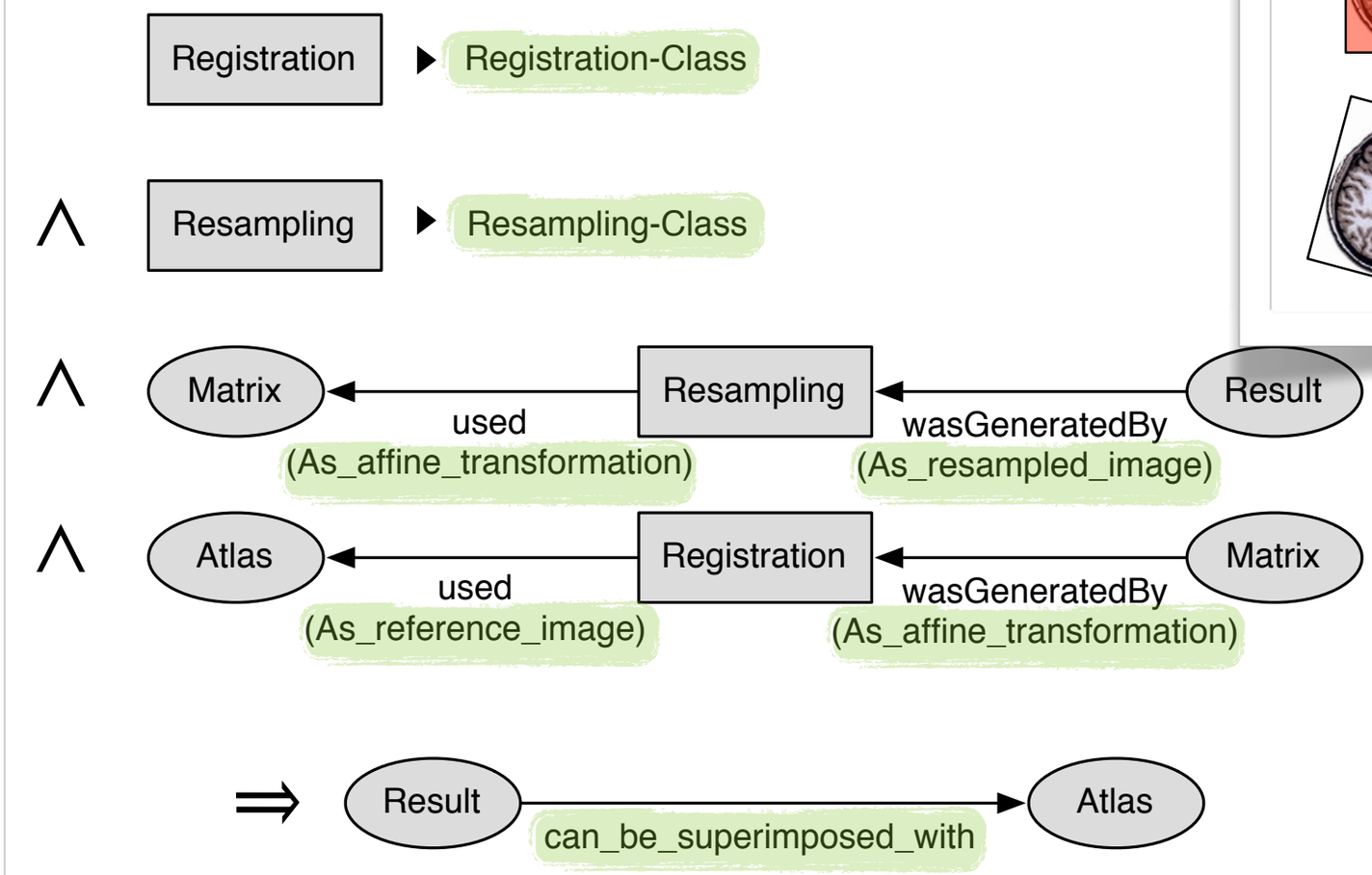
- Domain-specific extension of the **OPM Role class**
- **Roles** to **disambiguate** the annotation of **service parameters**.

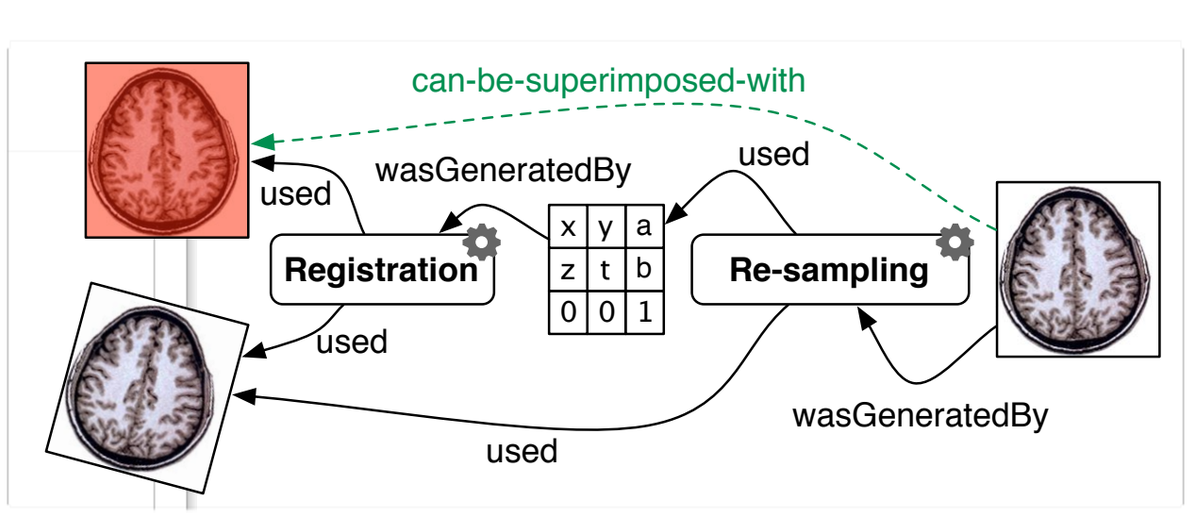




Inference rule example

- Inference rules to produce semantic annotations

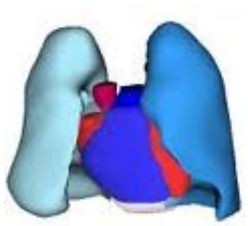


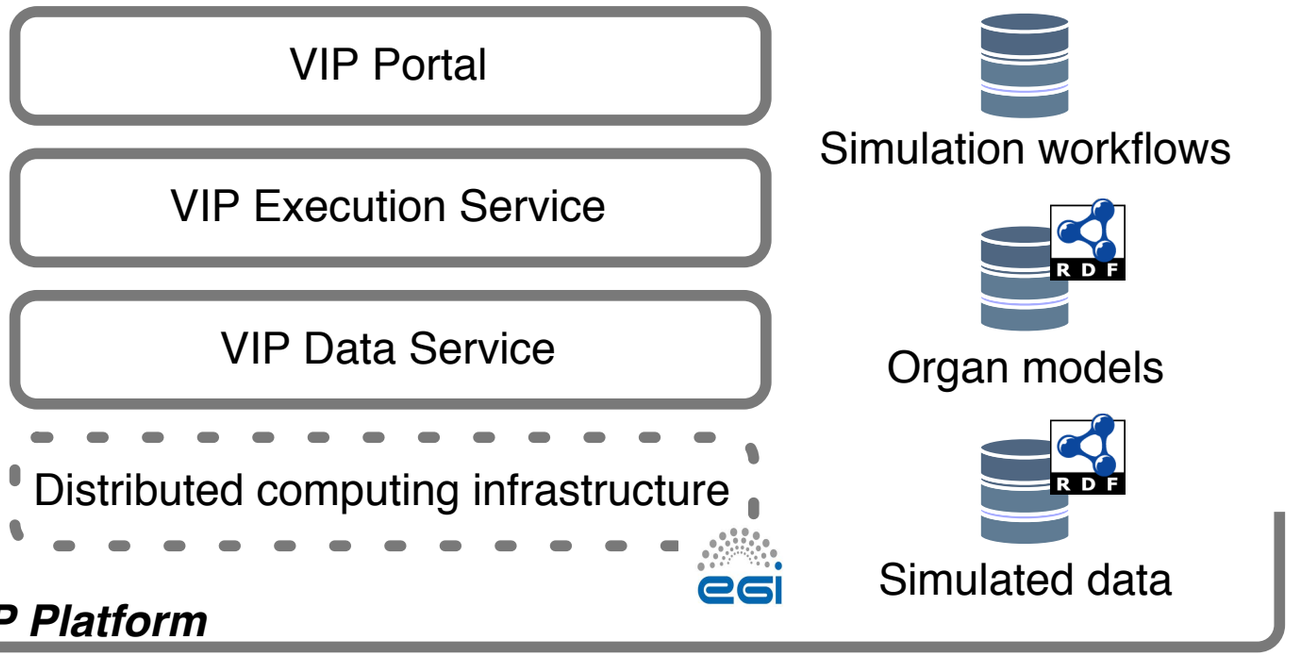


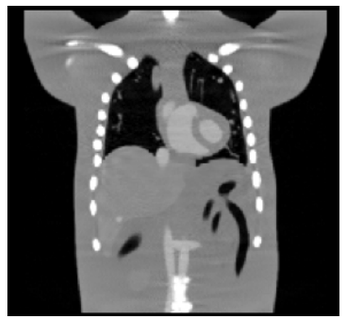
provenance-based
knowledge propagation

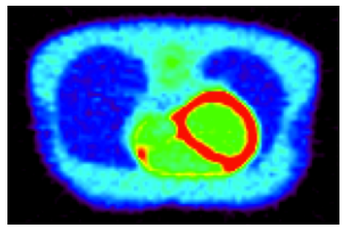
Experiment: inferring VIP experiment summaries (real-life)

- **Objectives:**
  - Inferring meaningful experiment summaries from WF runs & domain knowledge
  - Coping with provenance as distributed Linked Data
- **Material & Methods:**
  - VIP e-Science platform (Moteur WF engine ; OntoVIP ontology)
  - Service annotations (Roles), OPM provenance, Inference rules



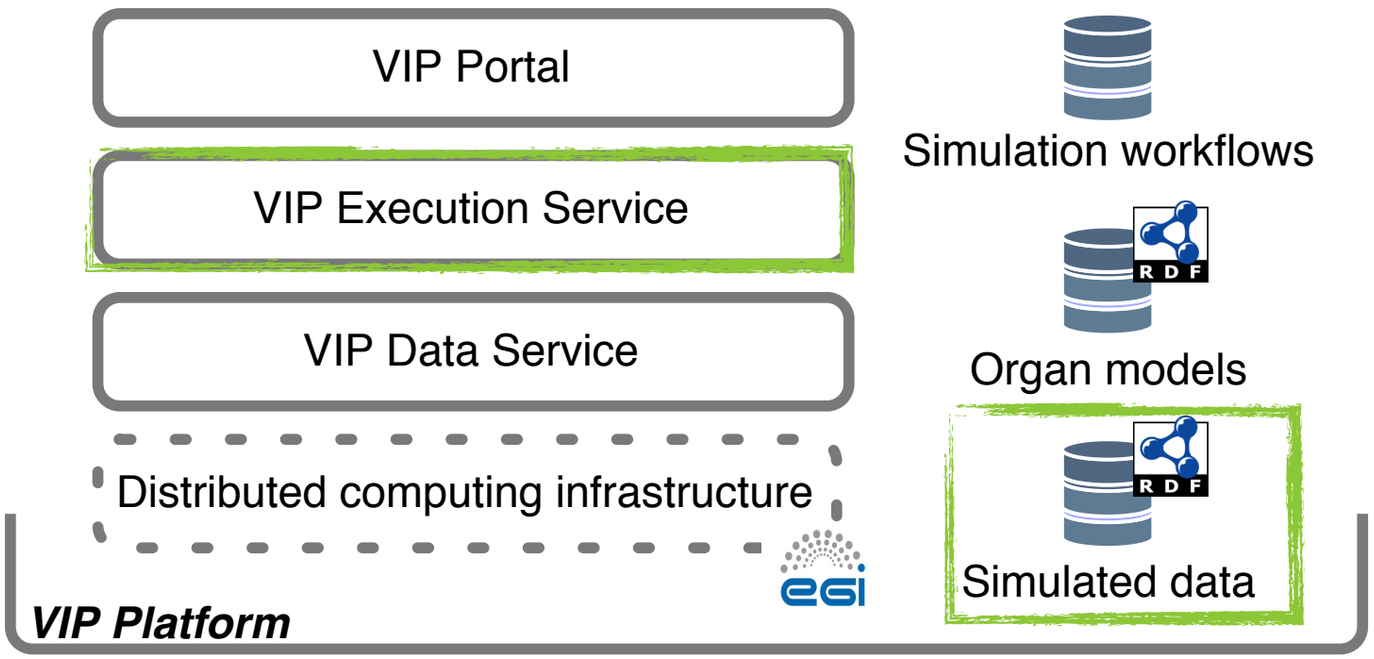


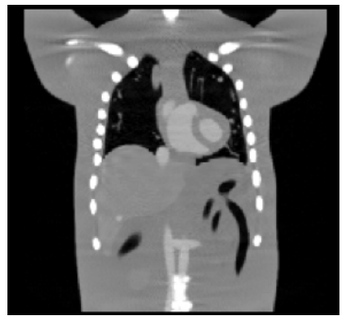


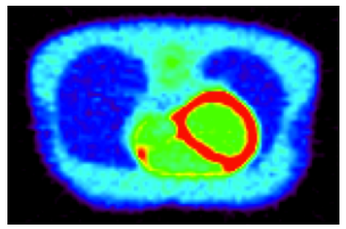


Experiment: inferring VIP experiment summaries (real-life)

- **Objectives:**
  - Inferring meaningful experiment summaries from WF runs & domain knowledge
  - Coping with provenance as distributed Linked Data
- **Material & Methods:**
  - VIP e-Science platform (Moteur WF engine ; OntoVIP ontology)
  - Service annotations (Roles), OPM provenance, Inference rules

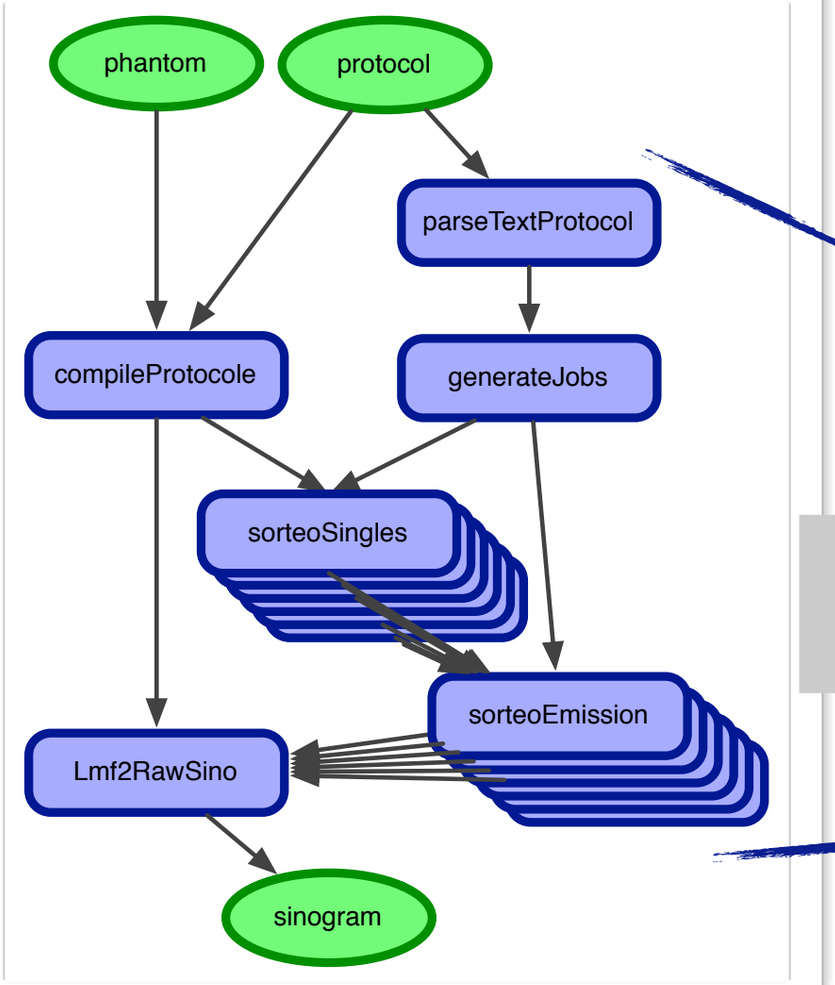






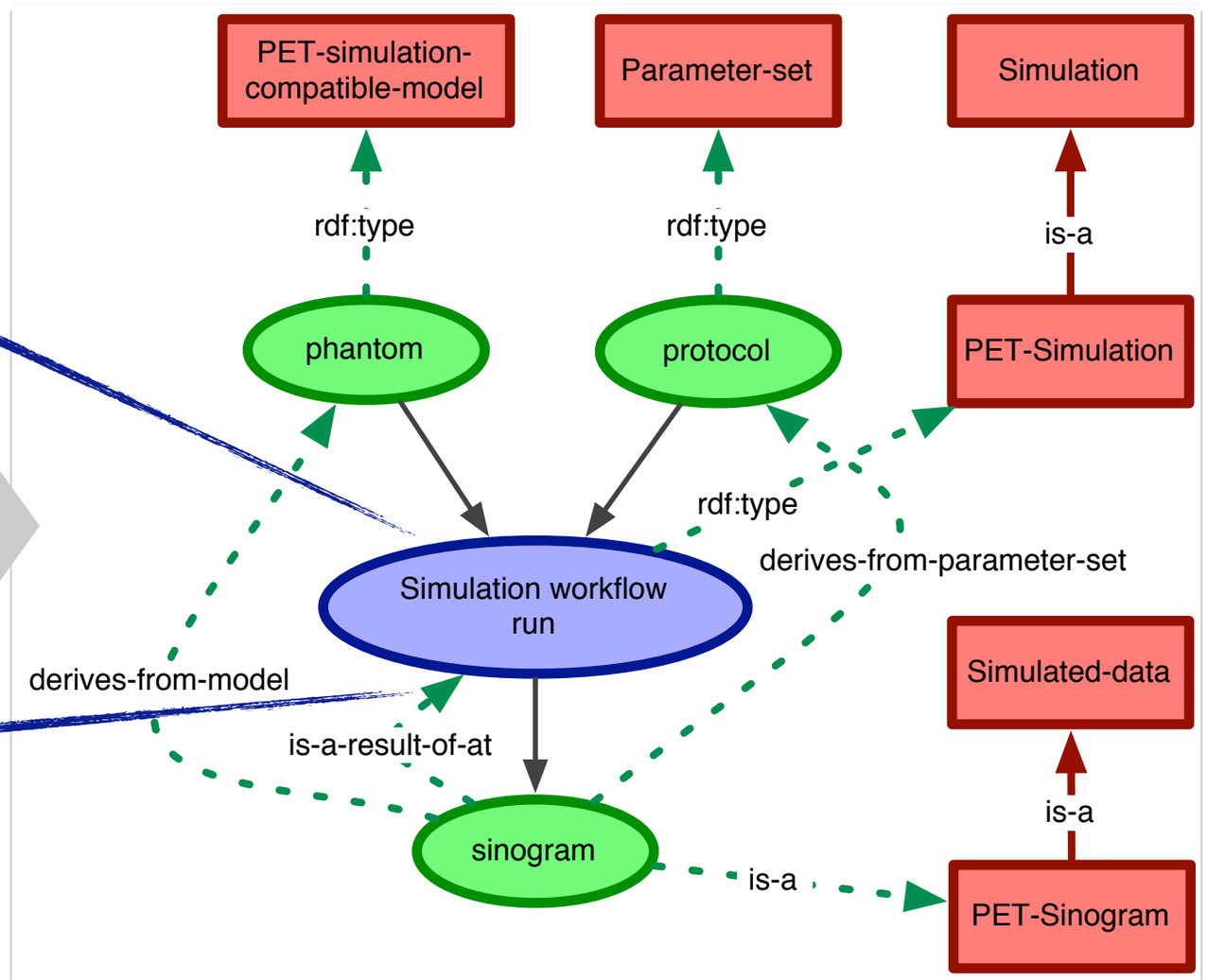
Inferring VIP experiment summaries (real-life)

Fine-grained & technical provenance



Inference rules

Coarse-grained & meaningful provenance

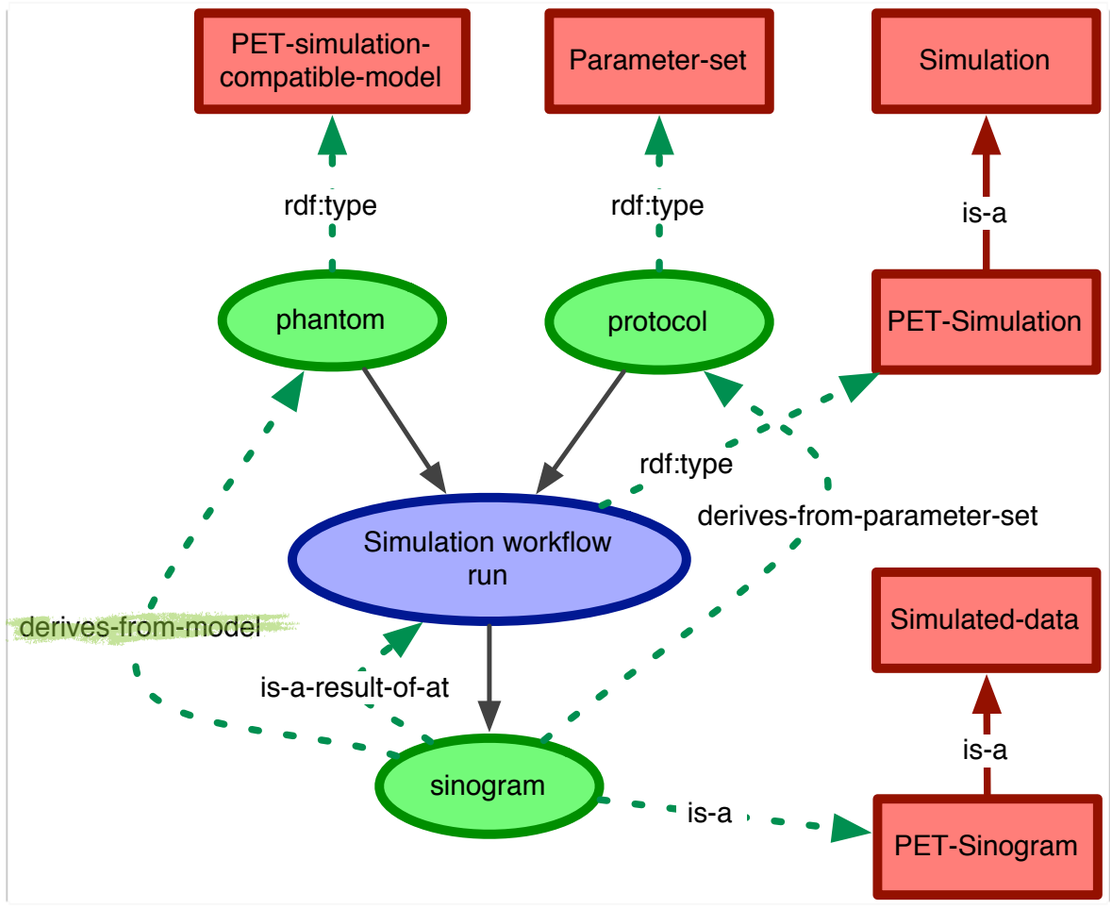


Inferring VIP experiment summaries: **material & methods**

Inference rule:

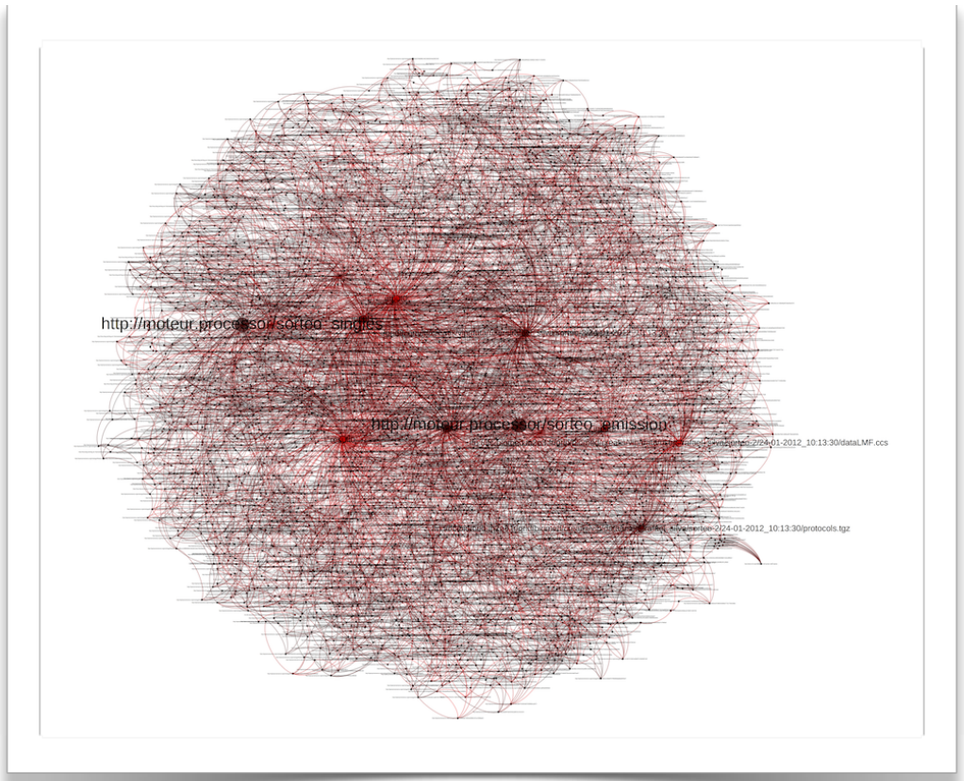
```
CONSTRUCT {
        ?out vip−model:derives−from−model ?inPhantom
        #...
} WHERE {
        ?agent (iec:refers−to/rdf:type)
                vip−simulation:image−reconstruction−simulator−component .
        ?wcb opmo:cause ?agent .
        ?wcb opmo:effect ?x .
        ?x rdf:type opmv:Process .
        ?wgb opmo:cause ?x .
        ?wgb opmo:effect ?out .

        ?agent2 (iec:refers−to/rdf:type)
                vip−simulation:parameters−generation−simulator−component .
        ?wcb2 opmo:cause ?agent2 .
        ?wcb2 opmo:effect ?y .
        ?y rdf:type opmv:Process .

        ?used1 opmo:cause ?inPhantom .
        ?used1 opmo:effect ?y .
        ?used1 opmo:role/rdfs:label ?techRolePhantom .

        ?agent2 ws:has−input ?inPortPhantom .
        ?inPortPhantom (iec:refers−to/rdf:type)
                vip−model:geometrical−phantom−object−model .
        ?inPortPhantom rdfs:comment ?techRolePhantom .

        ?inPhantom opmo:avalue ?vInPhantom .
        ?vInPhantom opmo:content ?cInPhantom .
        #...
}
```

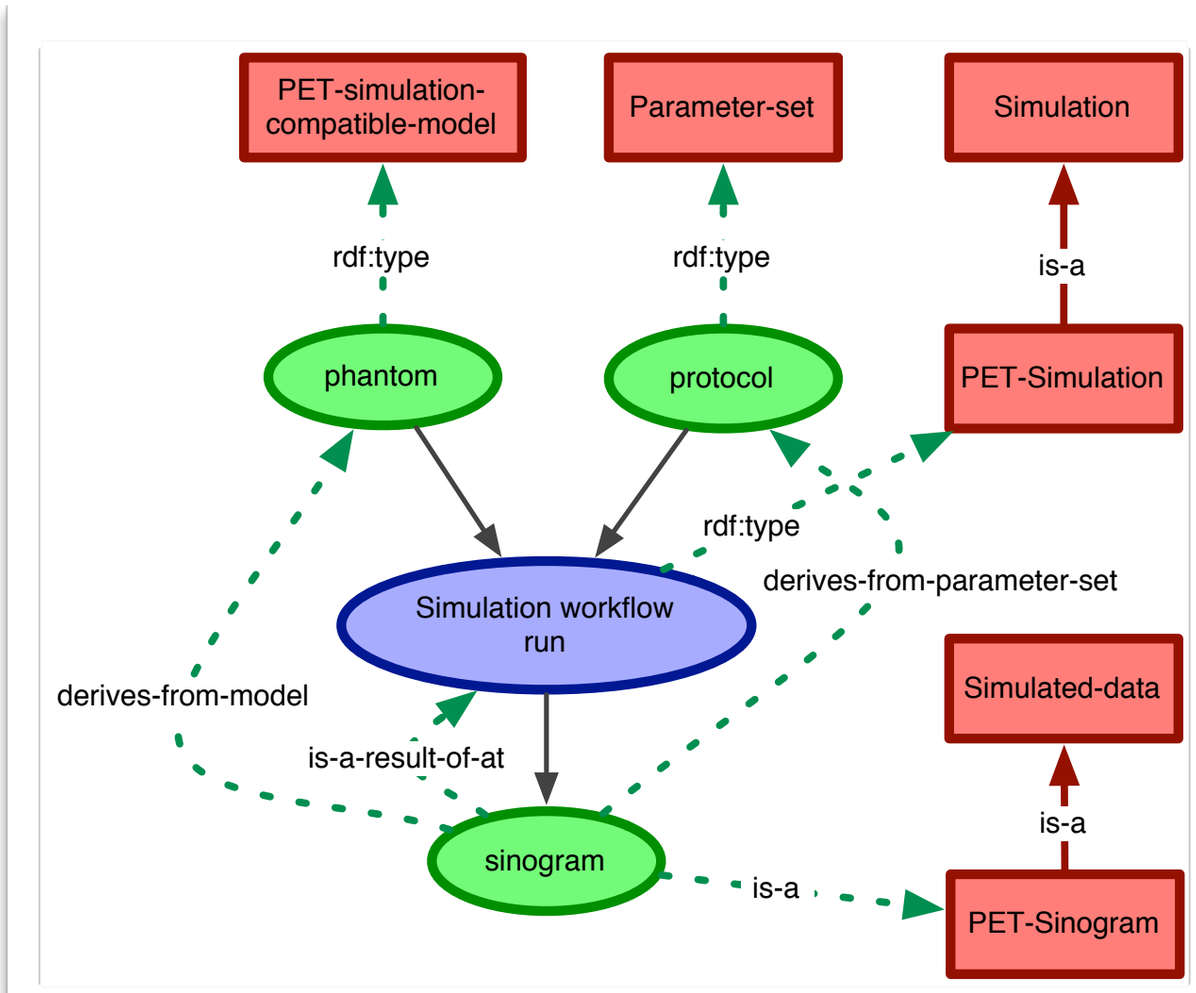
Inferred meaningful experiment summary:



Inferring VIP experiment summaries: **results**

- Semantic experiment summaries :





BIG fine-grained, ***meaningless*** *provenance*

FEW ***meaningful*** *statements*

results **Interpretation**

Inferring VIP experiment summaries: **results**

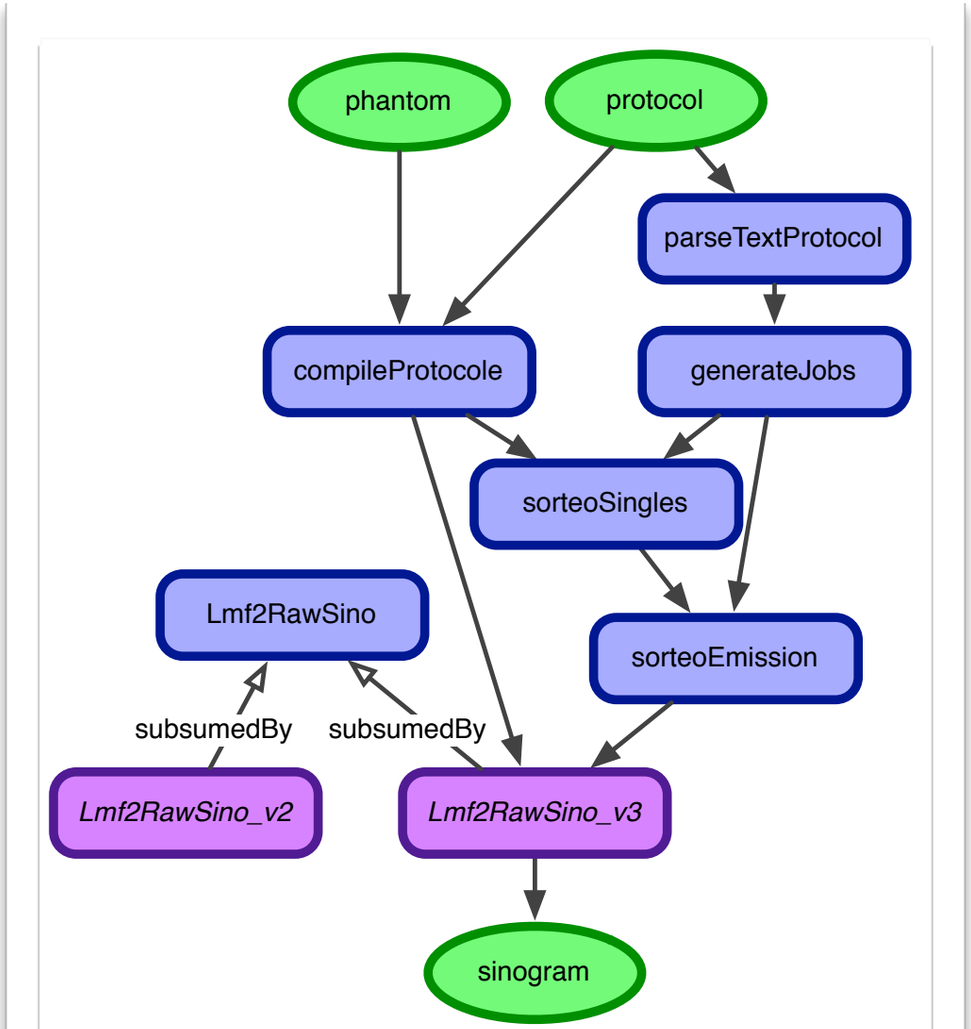
✓ Distributed **linked** provenance **data** & **inference rules**

- Grid'5000 infrastructure (3 OPM data sources) + KGRAM-DQP

✓ **Reusable** inference rules

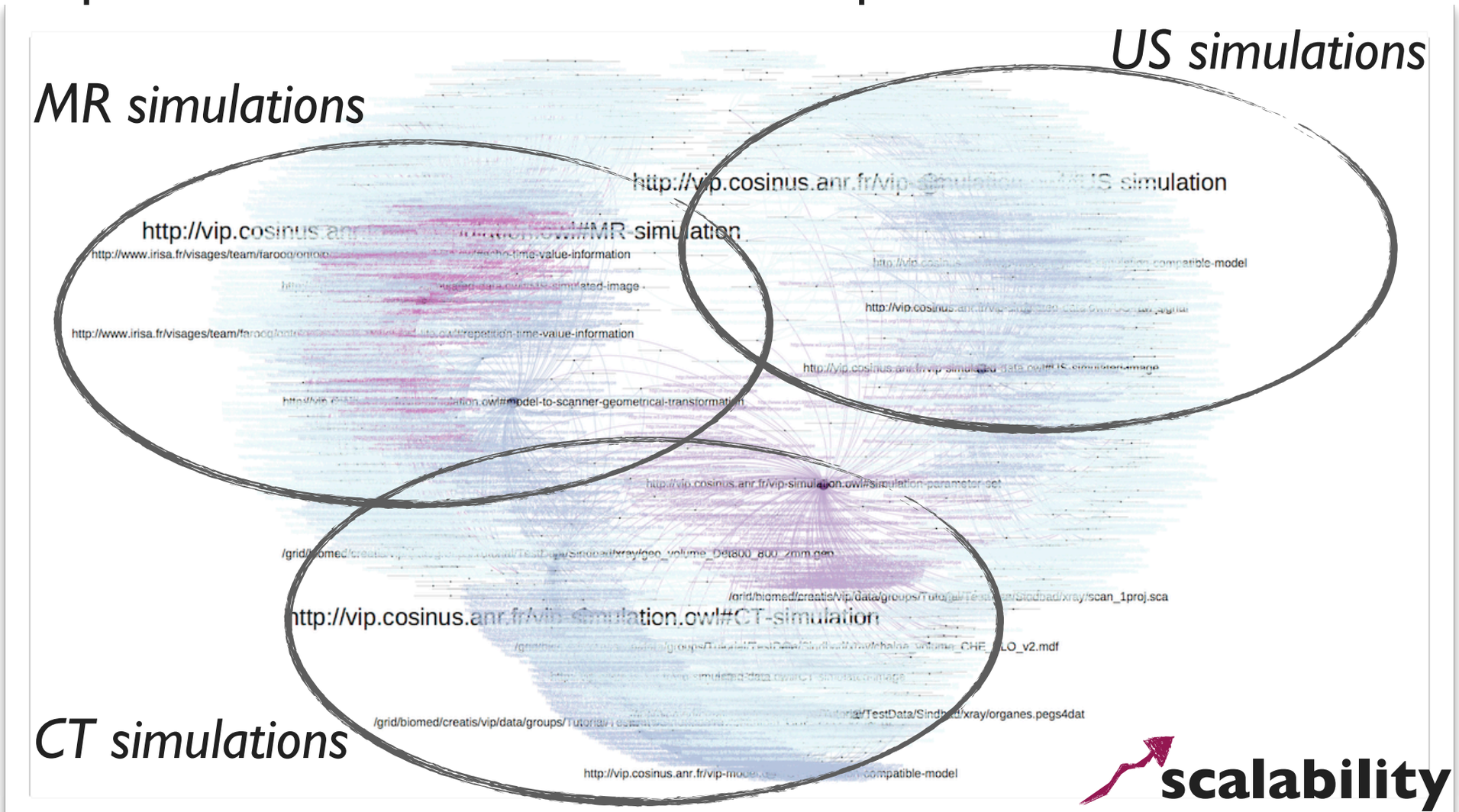
✓ adapt to simulator component evolutions

- do not adapt to workflow structure evolutions



Inferring VIP experiment summaries: **results**

- 1 week of VIP operation / 18 possible inference rules:
  - 118 Simulations (15K triples each) → 1.7 M triples
  - 118 Experiments summaries → 2656 triples



Highlights & short-term perspectives

- **Highlights**
  - Clear delineation between **Role** and **Natural concepts**
  - Domain ontology at **workflow design-time** and **run-time**
  - **Scalable annotation** of analyzed data through **semantic experiment summaries**
  - **Reusable** inference rules

- **Short-term perspectives**
  - Integration of neuro-imaging roles in a sound domain ontology
  - From OPM ontology to PROV-O
  - Publishing experiment summaries as Linked Open Data

Summary

- Enhance **e-Science platforms** with **Knowledge Engineering** (and **Semantic Web** technologies)
  - **Scalable** and **expressive Knowledge Sharing** approach through distributed query processing techniques and abstract knowledge graphs
  - **Smart Knowledge Production**: "few but meaningful data"

- **Deployment** into real-life platforms
  - **2 softwares**: NeuSemStore and KGRAM-DQP
    in production in **2 ANR projects** : NeuroLOG and VIP

Future directions

1. Towards high **performance** federated semantic querying:
   - triple pattern grouping & query planning
   - "Elastic" SPARQL endpoint for massive knowledge graphs

2. Towards highly **expressive** federated semantic querying
   - FedBench extensions with more expressive queries
   - Towards distributed reasoning (optimal plan for inferences ? materialization ?)

3. Towards **versatile** and **reliable** knowledge base federations
   - R2RML-based mediation of SQL databases
   - generalized provenance, from processed data to the originating data sources (explanation)

4. Towards reduced **information overload** in e-Science
   - Semantic experiment summaries & (goal-driven) conceptual workflows [Cerezo *et al.*, 2011]
     - Eased inference rules design by relying on WF goals
     - Annotated data to help in WF design

Merci !

alban.gaignard@cnrs.fr

- O. Corby, A. Gaignard, C. Faron Zucker, J. Montagnat. ***KGRAM versatile data graphs querying and inference engine***, **WI'12** (International Conference on Web Intelligence), Macao, 2012.
- A. Gaignard, J. Montagnat, B. Wali, B. Gibaud. ***Characterizing semantic service parameters with Role concepts to infer domain-specific knowledge at runtime***, **KEOD'11** (International Conference on Knowledge Engineering and Ontology Development), Paris, 2011.
- A. Gaignard, J. Montagnat, C. Faron Zucker, O. Corby. ***Semantic Federation of Distributed Neurodata***, **MICCAI-DCICTAI workshop** (Data- and Compute-Intensive Clinical and Translational Imaging Applications), Nice, 2012.
- A. Gaignard, J. Montagnat, C. Faron Zucker, O. Corby. ***Fédération multi-sources en neurosciences : intégration de données relationnelles et sémantiques***, **IC'12** (Ingénierie des Connaissances), **workshop** "Ingénierie des connaissances pour l'inter-opérabilité sémantique en e-Santé", Paris, 2012.
- T. Glatard, C. Lartizien, B. Gibaud, R. Ferreira da Silva, G. Forestier, F. Cervenansky, M. Alessandrini, H. Benoit-Cattin, O. Bernard, S. Camarasu-Pop, N. Cerezo, P. Clarysse, A. Gaignard, P. Hugonnard, H. Liebgott, S. Marache, A. Marion, J. Montagnat, J. Tabary and D. Friboulet. ***A Virtual Imaging Platform for multi-modality medical image simulation***, IEEE Transactions on Medical Imaging (**TMI**), 32 (1), pages 110-118, 2013.