CrEDIBLE

fédération de données et de ConnaissancEs Distribuées en Imagerie BiomédicaLE

Data fusion, semantic alignment, distributed queries

Johan Montagnat
CNRS, I3S lab, Modalis team

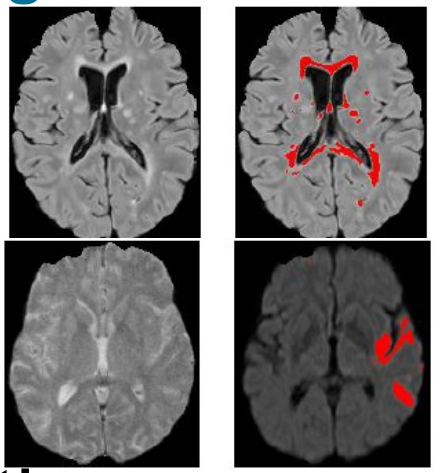
on behalf of the CrEDIBLE consortium

CNRS/UNS, laboratoire I3S (UMR7271), équipe MODALIS
INRIA/CNRS/UNS, laboratoire I3S (UMR 7271), équipe Wimmics
INSERM U1099, laboratoire LTSI, équipe MediCIS
U. Picardie, laboratoire MIS, équipe Connaissances

Motivations

- Biomedical data
  - High heterogeneity: images, clinical data, biomarkers, biology....
  - Increasing amount / number of (open) sources – **Big Data**
    - Large-scale medical studies

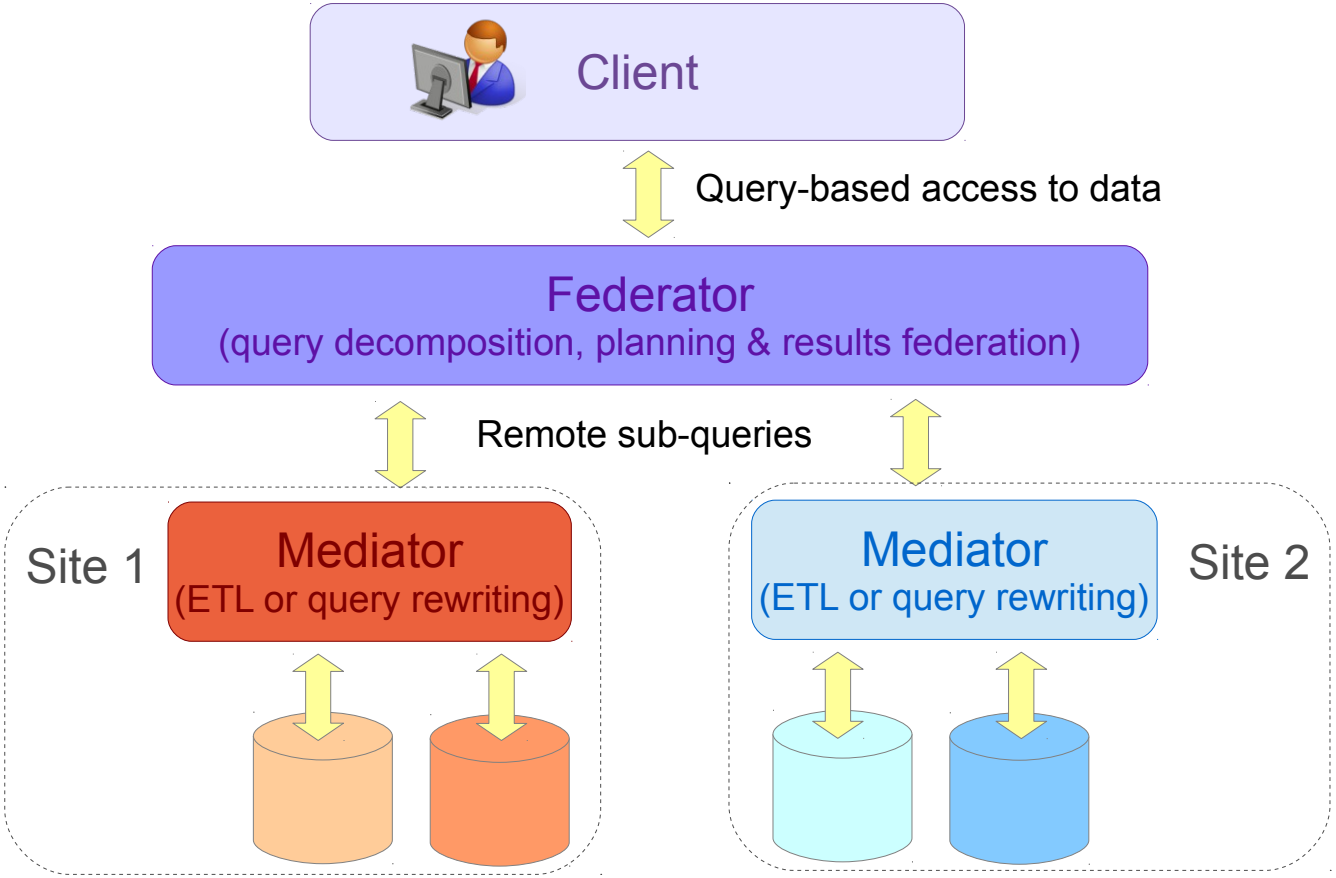
      (statistical medical studies, epidemiology...)
  - Need for cross-factors analysis – **Linked Data**
    - Data (re)analysis opportunities
    - Multicentric studies, Translational research



- Centralized approaches encounter limitations
  - Multiple data source kinds
  - Large data volumes to transfer / archive / search
  - Sensitive patient data / complex access control policies
  - Need to adopt uniform data model & format
- Data is *de facto* distributed over acquisition centers

Biomedical data mediation & federation

- Data federation through distributed querying and query rewriting



- Heterogeneous databases schema mediation
- Medical data & metadata:
  - raw data + models + processing results + models + provenance...

Domain ontology-based federation

Medical domain ontology
(reference model)

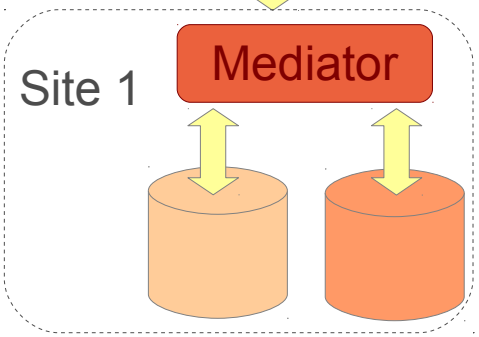
Data querying

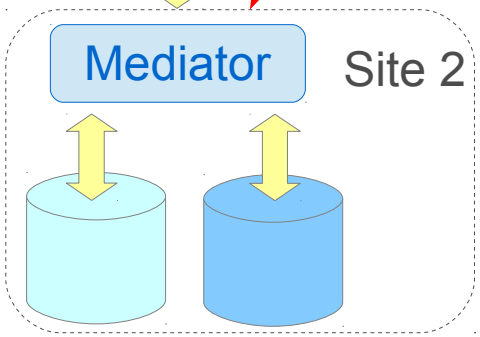
Data alignment

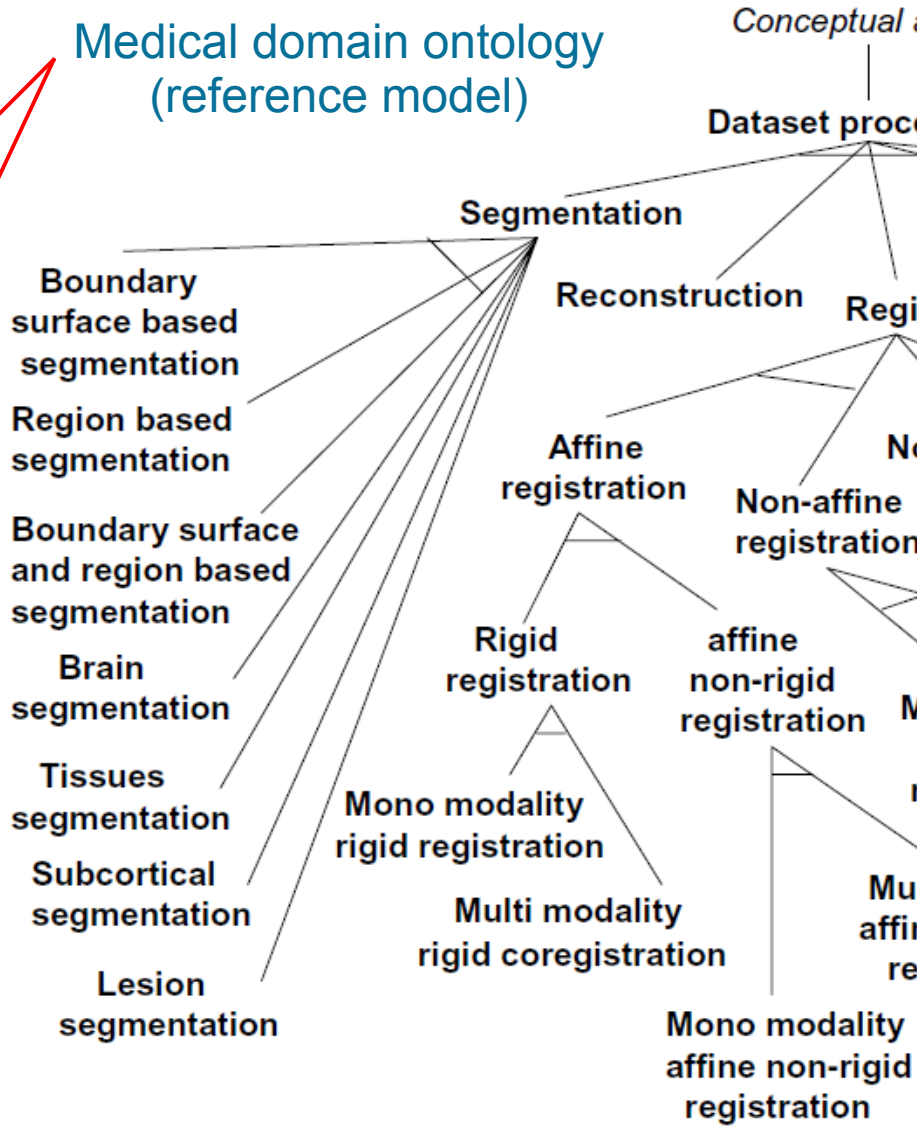
Query-based access to data

Federator

Remote sub-queries

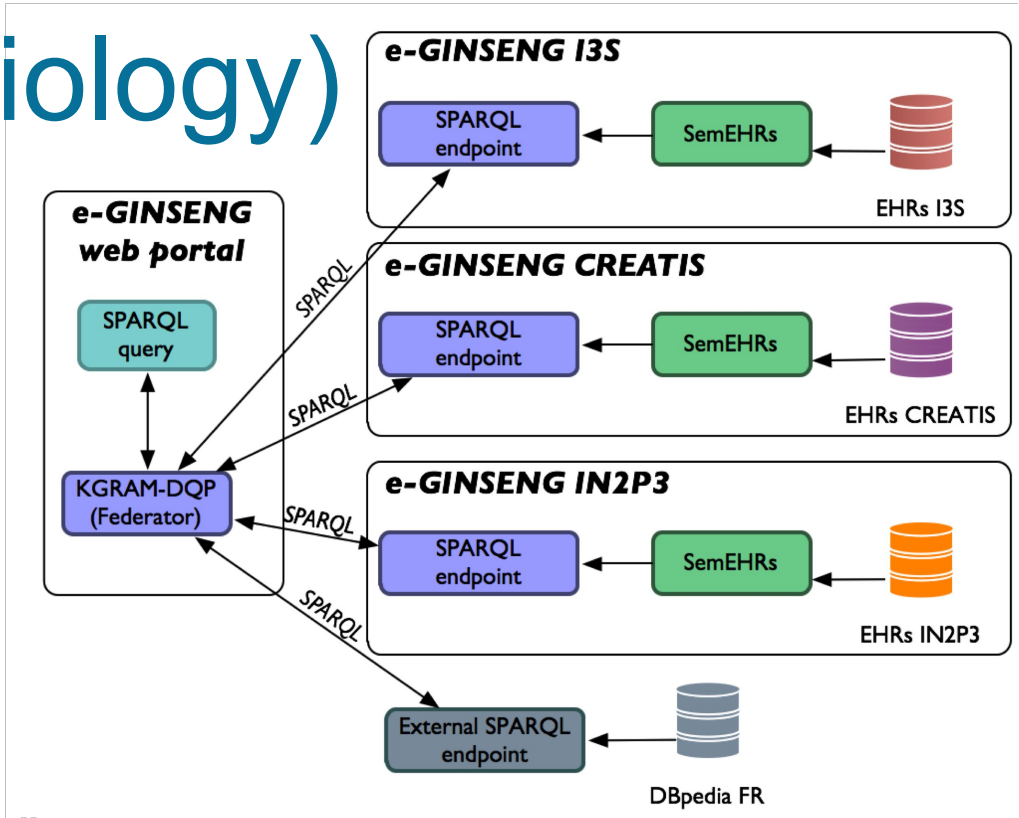


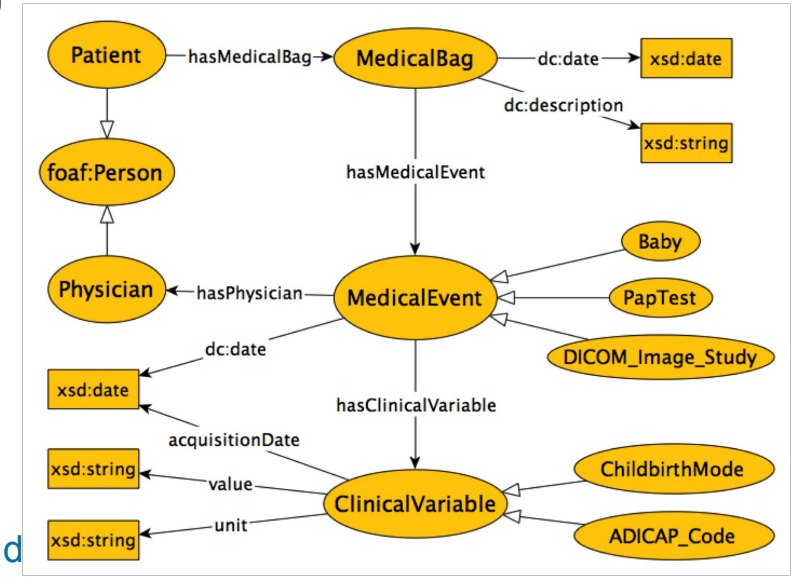




Exploitation example: ANR GINSENG (epidemiology)

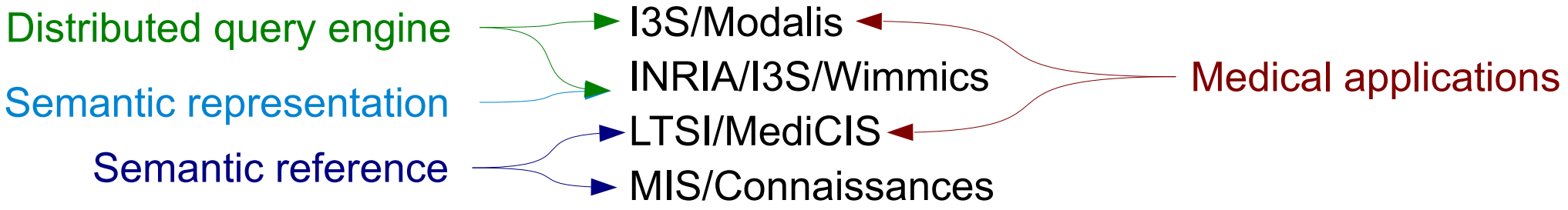
- Partnership with GINSENG consortium and Mnemotix SME
- Federation of heterogeneous epidemiology repositories
  - Multiple epidemiology data acquisition networks
  - Cross-correlation with “external” data (e.g. demographic: IGN)
- Mediation of the EHR (Electronic Health Record) data schema





Challenges and expertise

- Challenges
  - Representation of data semantics for heterogeneous data sources → biomedical ontology building
  - Data federation → distributed query engine
  - Data mediation → RDB2RDF, ontology alignment
- Partnership

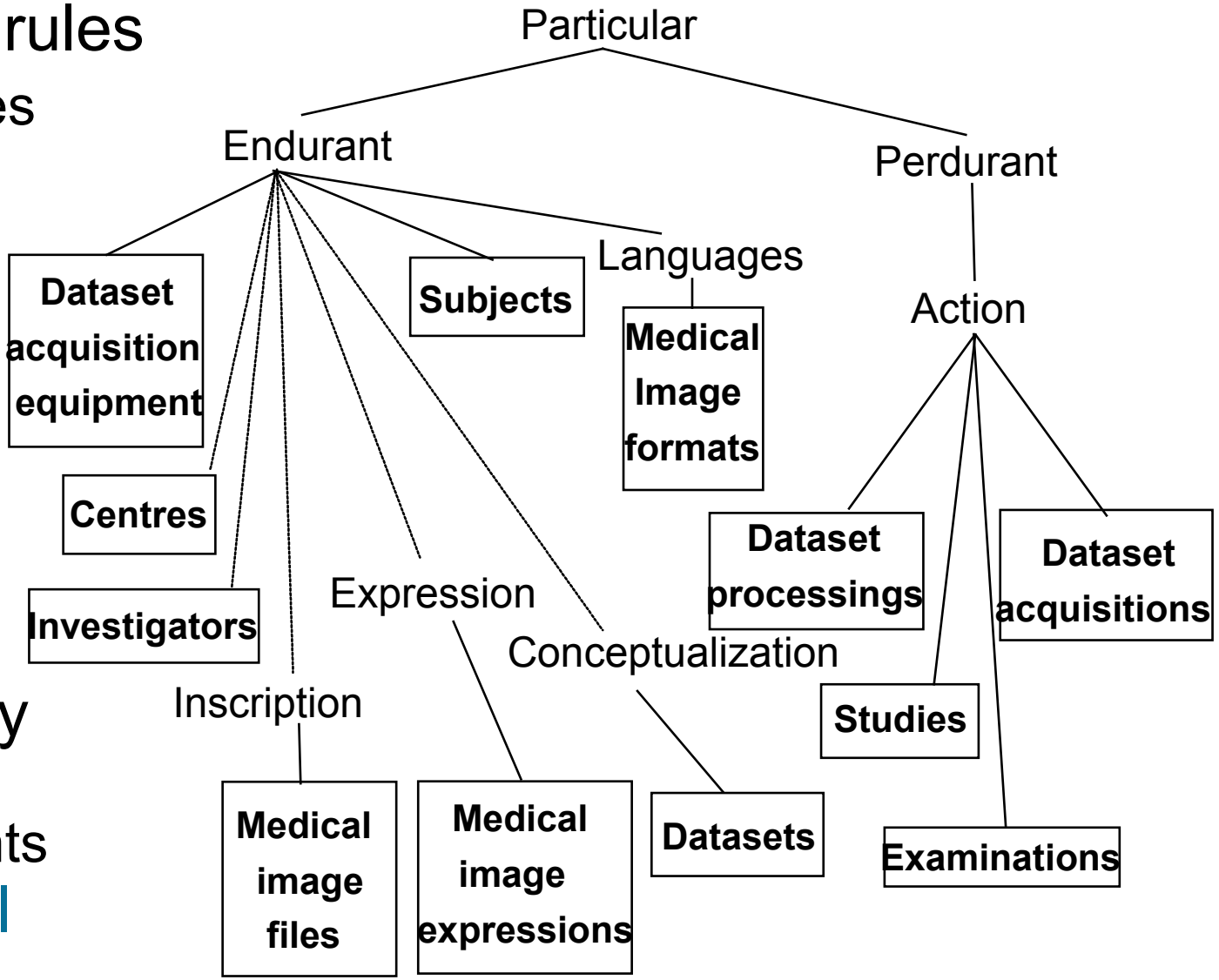


Scientific networking – annual workshop

- Objectives
  - Multidisciplinary workshop gathering international experts in biomedical data representation, semantic, distribution, federation, integration...
- October 2012
  - Semantic models usage becomes mainstream
  - Medical systems are mostly centralized but strigent need to support multi-centric studies
- October 2013
  - Many distributed data federation engines
  - Work on data partition schemes and data modeling
  - Query language expressivity limitations
- October 2014
  - Towards Web-scale databases (bilion of tuples)
  - R/W access and Semantic reasoning increasingly studied

Reference ontology

- 3-levels structure: foundational (DOLCE), core, domain
- Domain-specific rules
  - Inference abilities



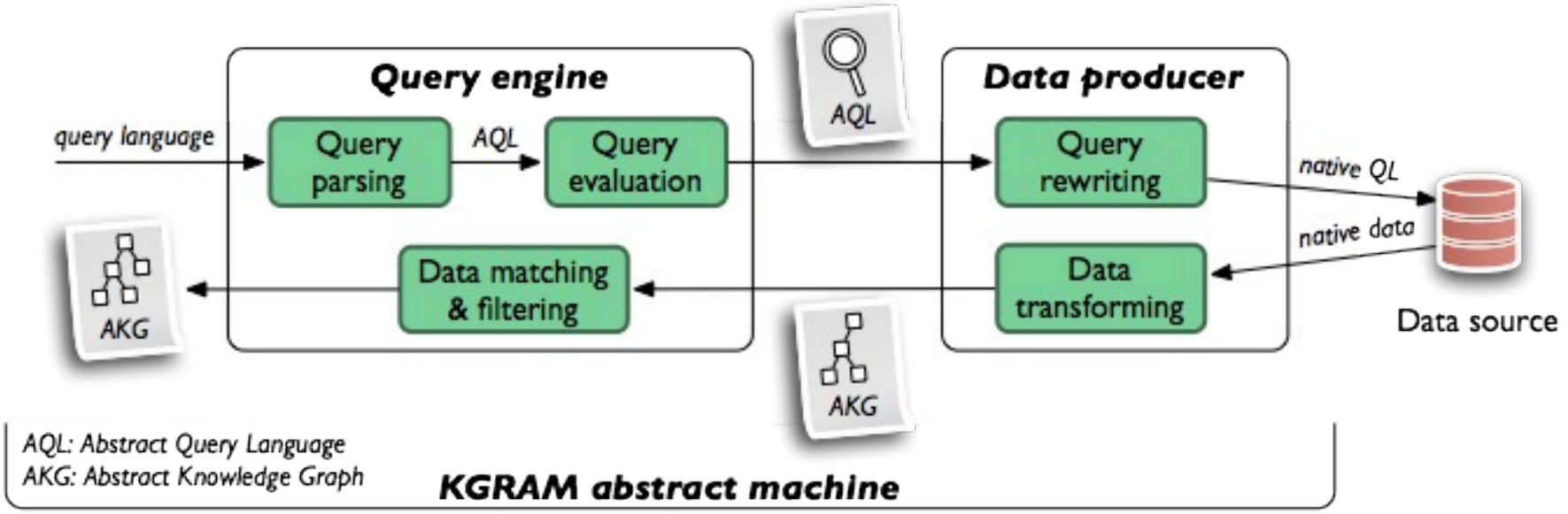
- DataTop ontology
  - Current focus on measurements
- Derived relational schema

Ontology modules

- Modularized ontology to improve reuse and lightweightness
  - ONL-MR-DA: MR Dataset Acquisition
  - ONL-DP: Data Processing
  - ONL-MSA: Mental State Assessment
  - OntoVIP: Medical Image Simulation
- Wide diffusion
  - http://bioportal.bioontology.org/ontologies

Data query and federation engine

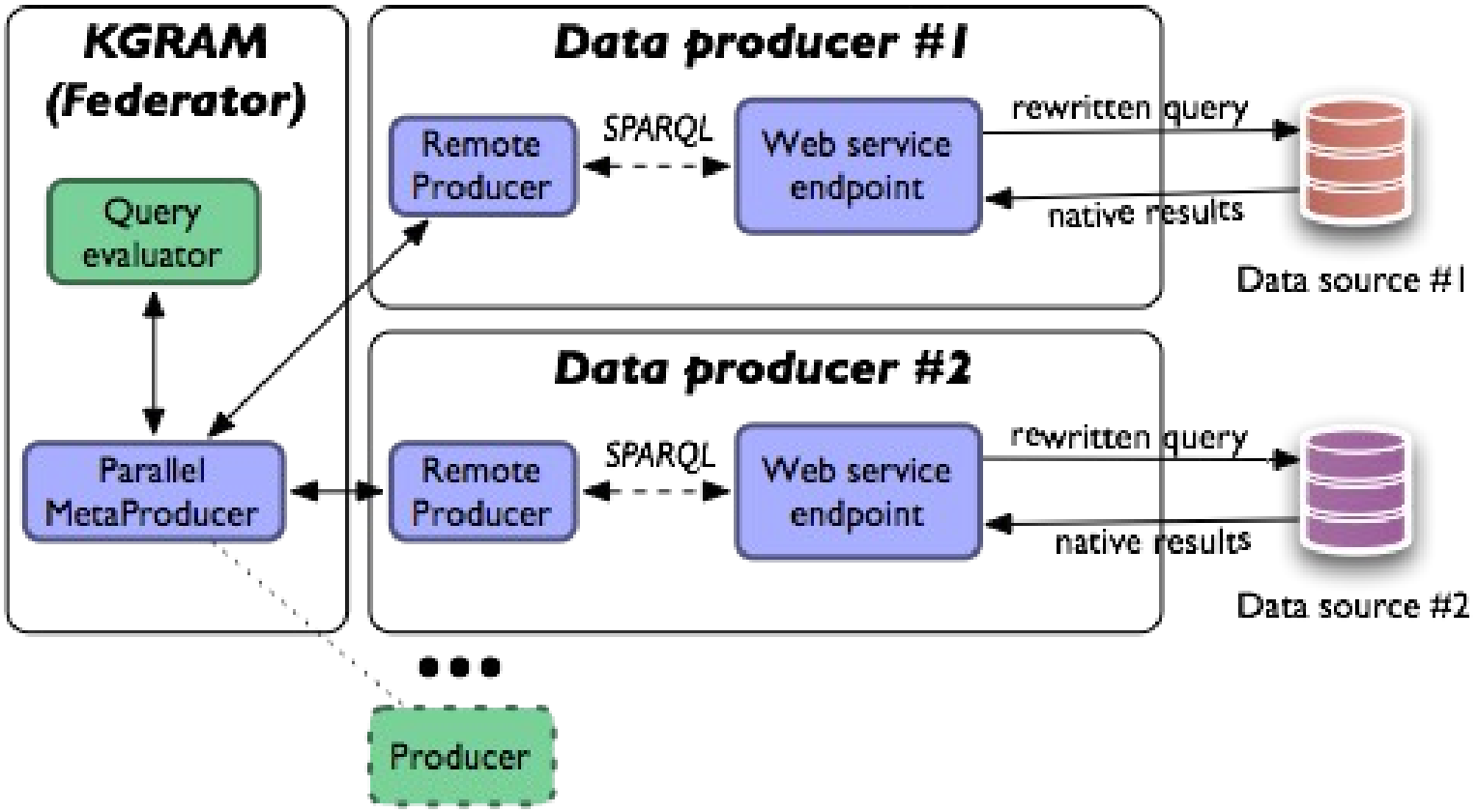
- KGRAM (Knowledge Graph Abstract Machine) Semantic query engine:
  - Full support of SPARQL1.1
  - Generic interface for heterogeneous backends
  - Flexible architecture facilitating different deployment scenarios



- Mediation interface to access relational data
  - Federated relational schema derived from the ontology

Distributed Query Processing

- Query federator decoupled from data sources
- Asynchronous querying of multiple data sources
- Query planning and parallel querying



Distributed Query Processing

- KGRAM query processing

Q

```
SELECT ?name ?date
WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
        FILTER (CONTAINS (?name, 'Bob')) }
```

Distributed Query Processing

- KGRAM query processing

Q

```
SELECT ?name ?date
WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
        FILTER (CONTAINS (?name, 'Bob')) }
```

Q1

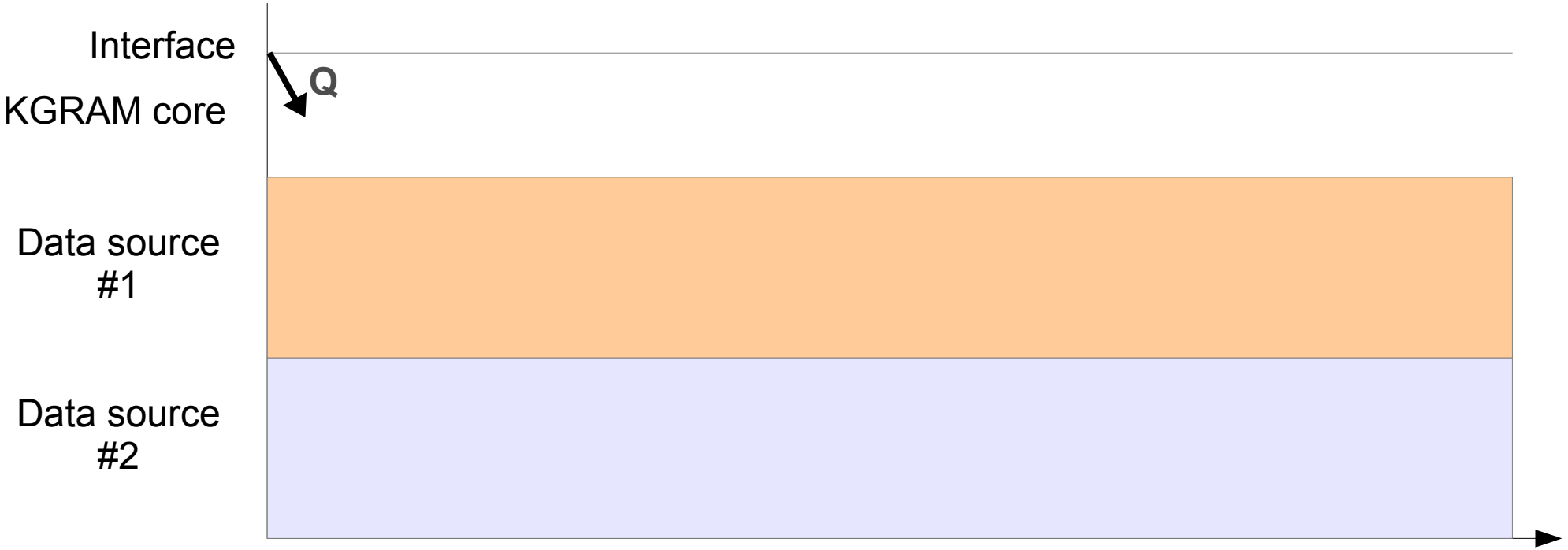
Q2

Distributed Query Processing

- KGRAM query processing

```
Q   SELECT ?name ?date
    WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
      Q1    FILTER (CONTAINS (?name, 'Bob')) }        Q2
```

- Asynchronous execution

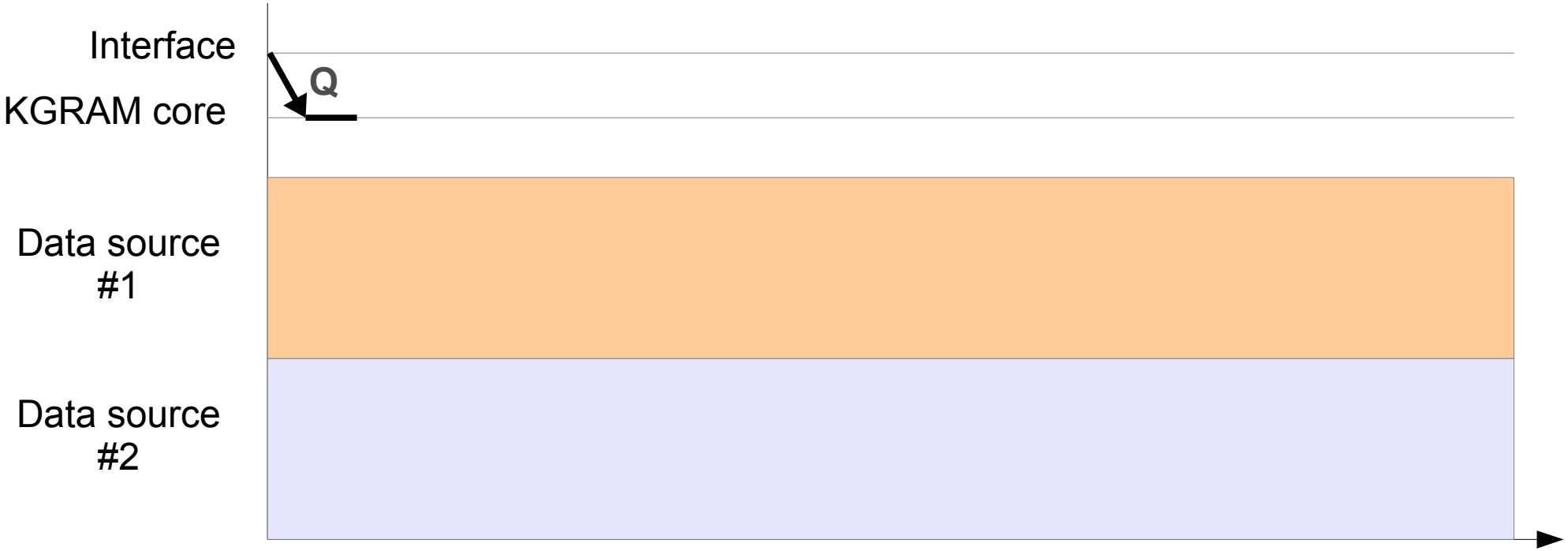


Distributed Query Processing

- KGRAM query processing

```
Q   SELECT ?name ?date
    WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
      Q1    FILTER (CONTAINS (?name, 'Bob')) }        Q2
```

- Asynchronous execution

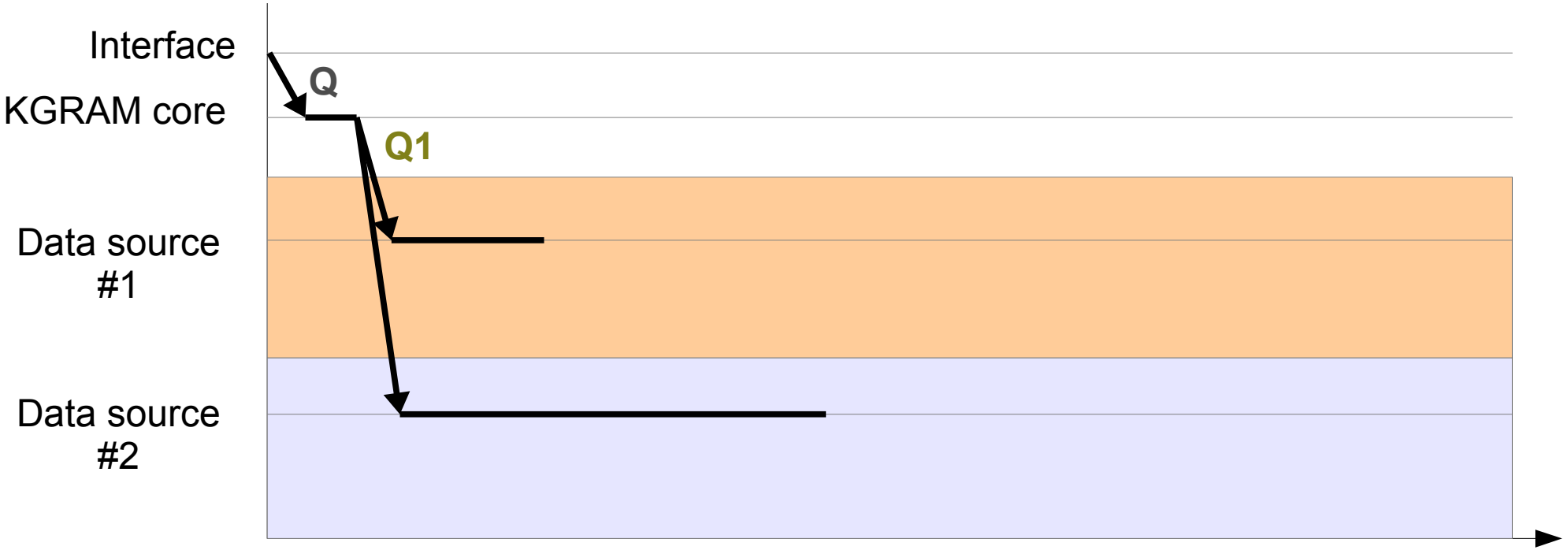


Distributed Query Processing

- KGRAM query processing

```
Q   SELECT ?name ?date
    WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
      Q1    FILTER (CONTAINS (?name, 'Bob')) }        Q2
```

- Asynchronous execution

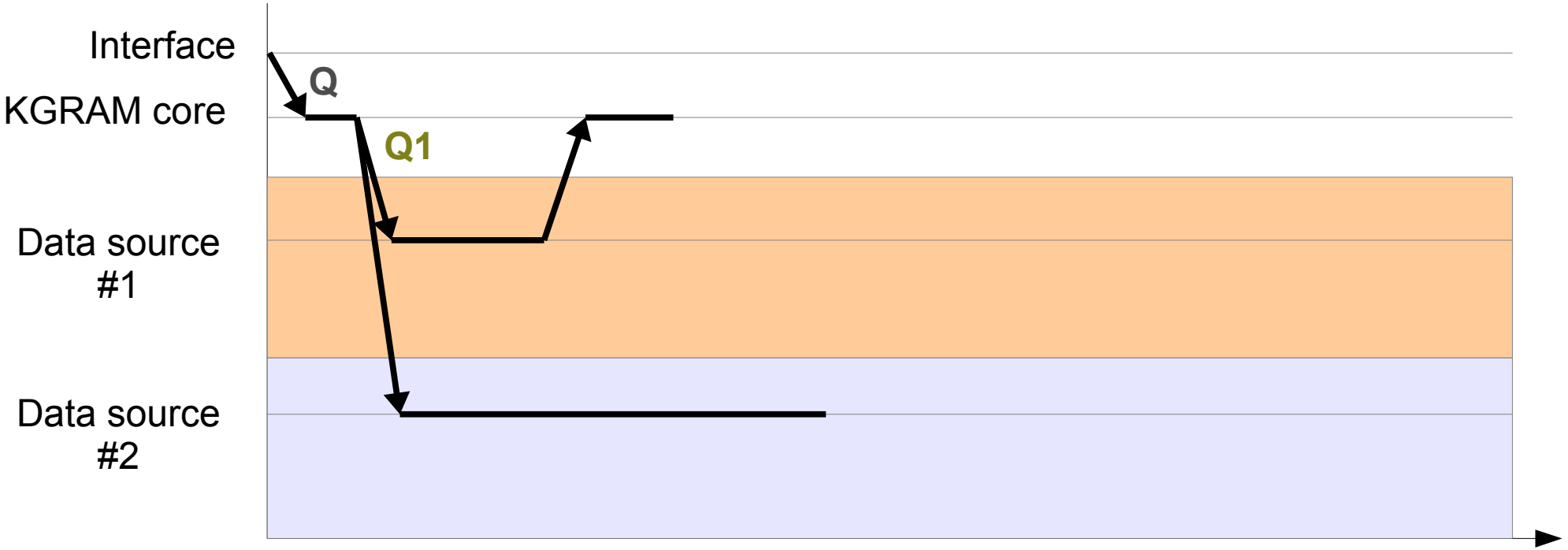


Distributed Query Processing

- KGRAM query processing

```
Q   SELECT ?name ?date
    WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
      Q1    FILTER (CONTAINS (?name, 'Bob')) }        Q2
```

- Asynchronous execution

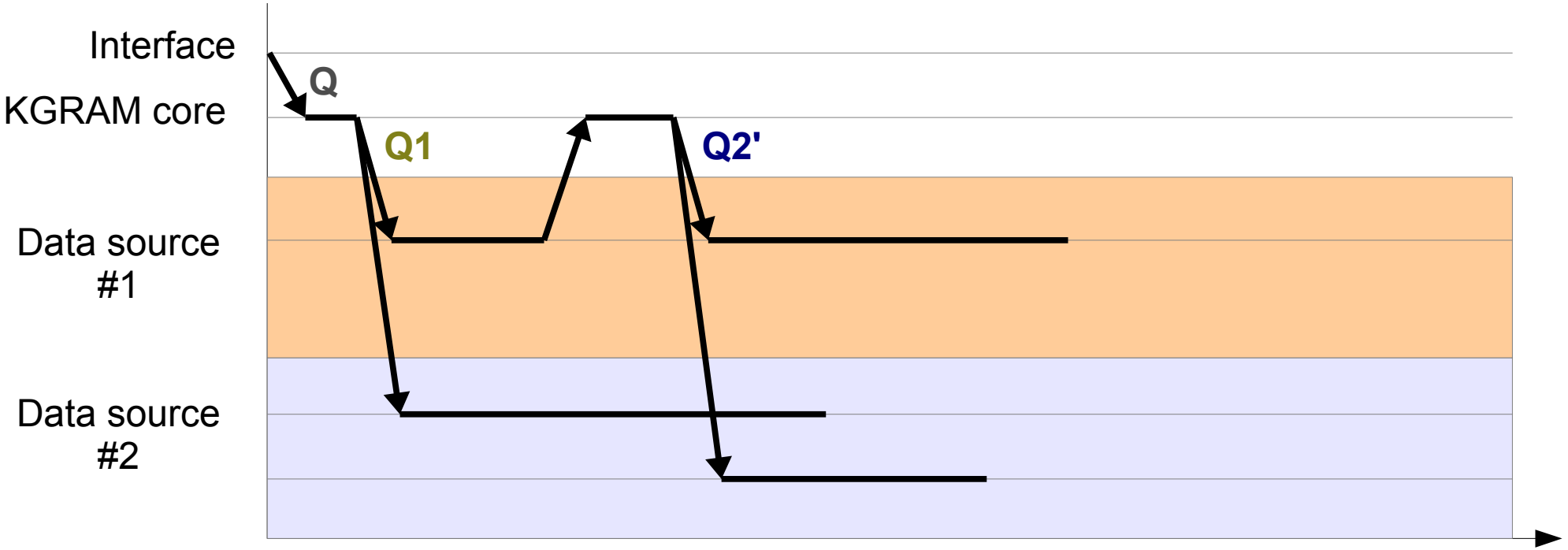


Distributed Query Processing

- KGRAM query processing

```
Q   SELECT ?name ?date
    WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
      Q1    FILTER (CONTAINS (?name, 'Bob')) }        Q2
```

- Asynchronous execution

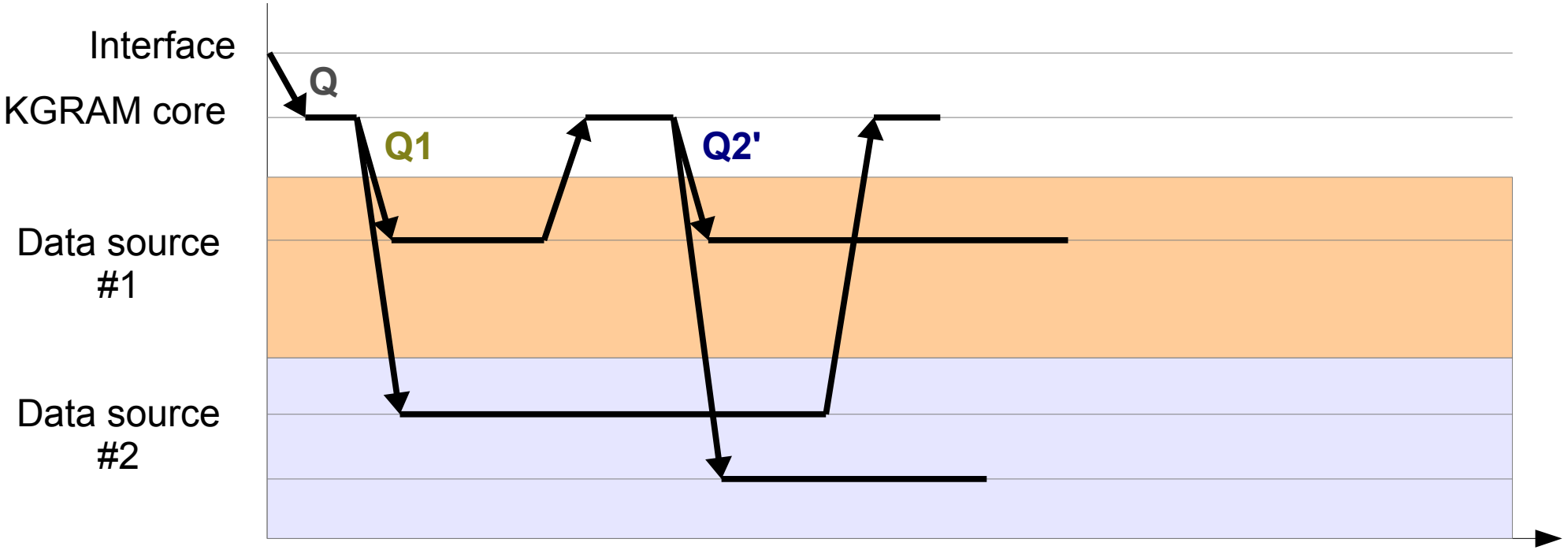


Distributed Query Processing

- KGRAM query processing

```
Q   SELECT ?name ?date
    WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
        Q1  FILTER (CONTAINS (?name, 'Bob')) }        Q2
```

- Asynchronous execution

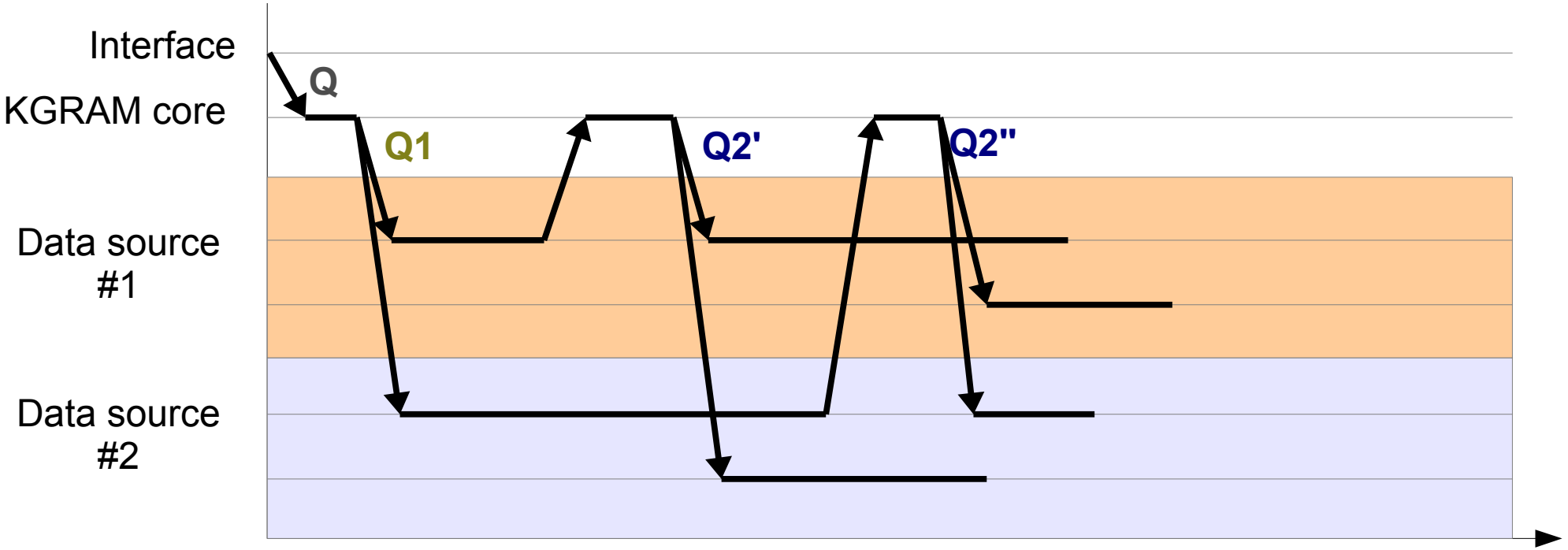


Distributed Query Processing

- KGRAM query processing

```
Q   SELECT ?name ?date
    WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
      Q1    FILTER (CONTAINS (?name, 'Bob')) }        Q2
```

- Asynchronous execution

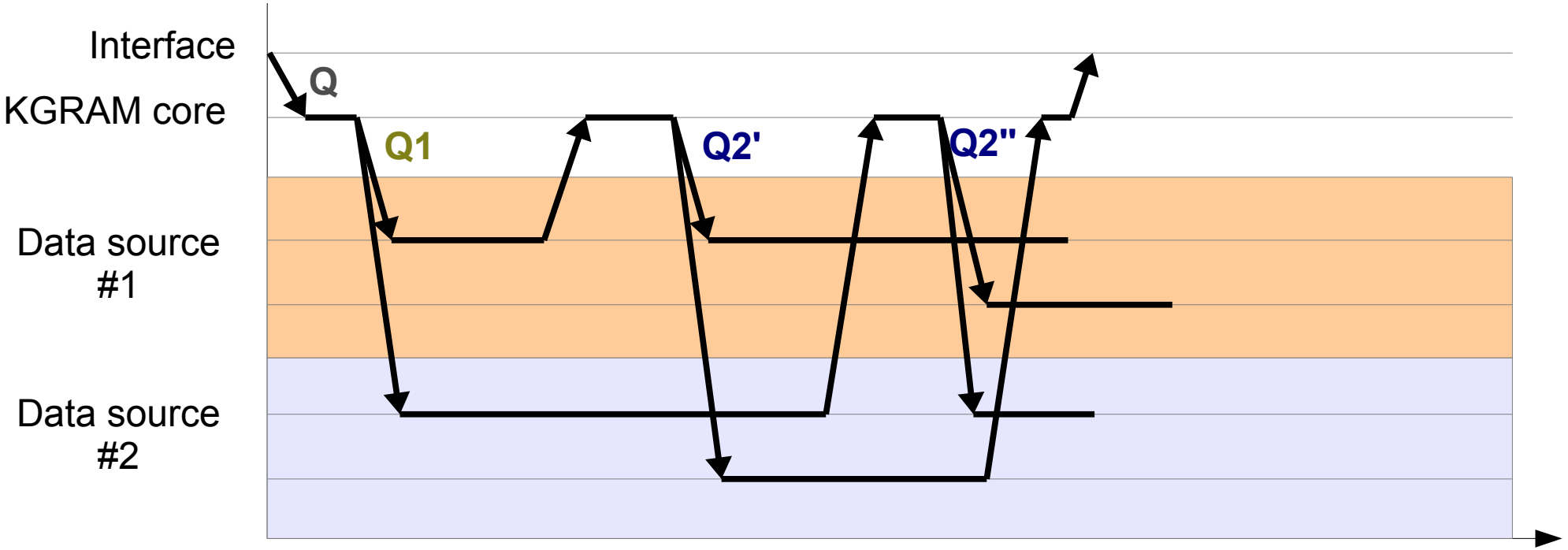


Distributed Query Processing

- KGRAM query processing

```
Q   SELECT ?name ?date
    WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
      Q1    FILTER (CONTAINS (?name, 'Bob')) }        Q2
```

- Asynchronous execution

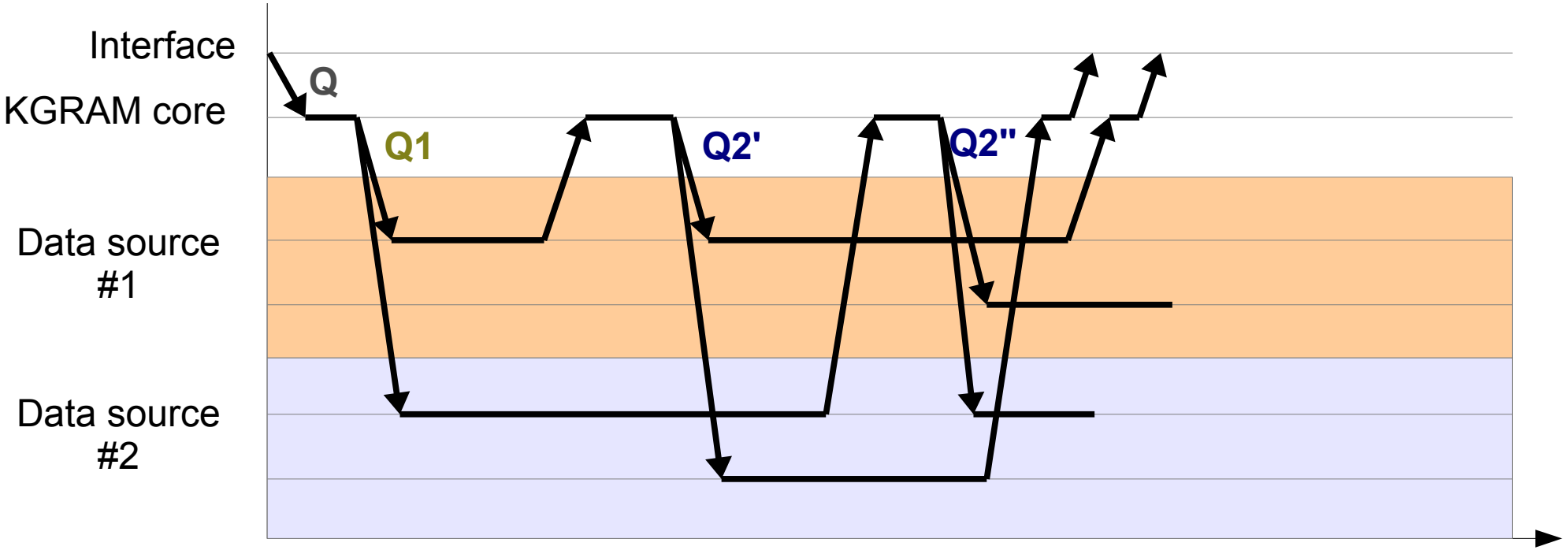


Distributed Query Processing

- KGRAM query processing

```
Q   SELECT ?name ?date
    WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
      Q1    FILTER (CONTAINS (?name, 'Bob')) }        Q2
```

- Asynchronous execution

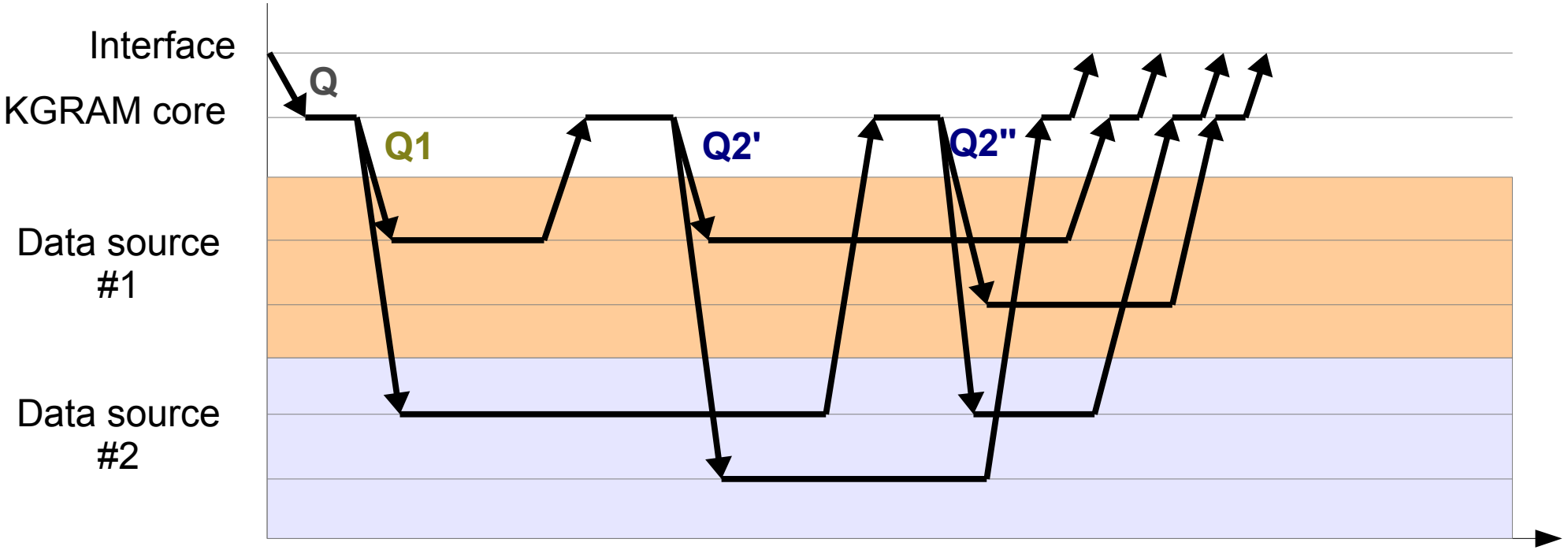


Distributed Query Processing

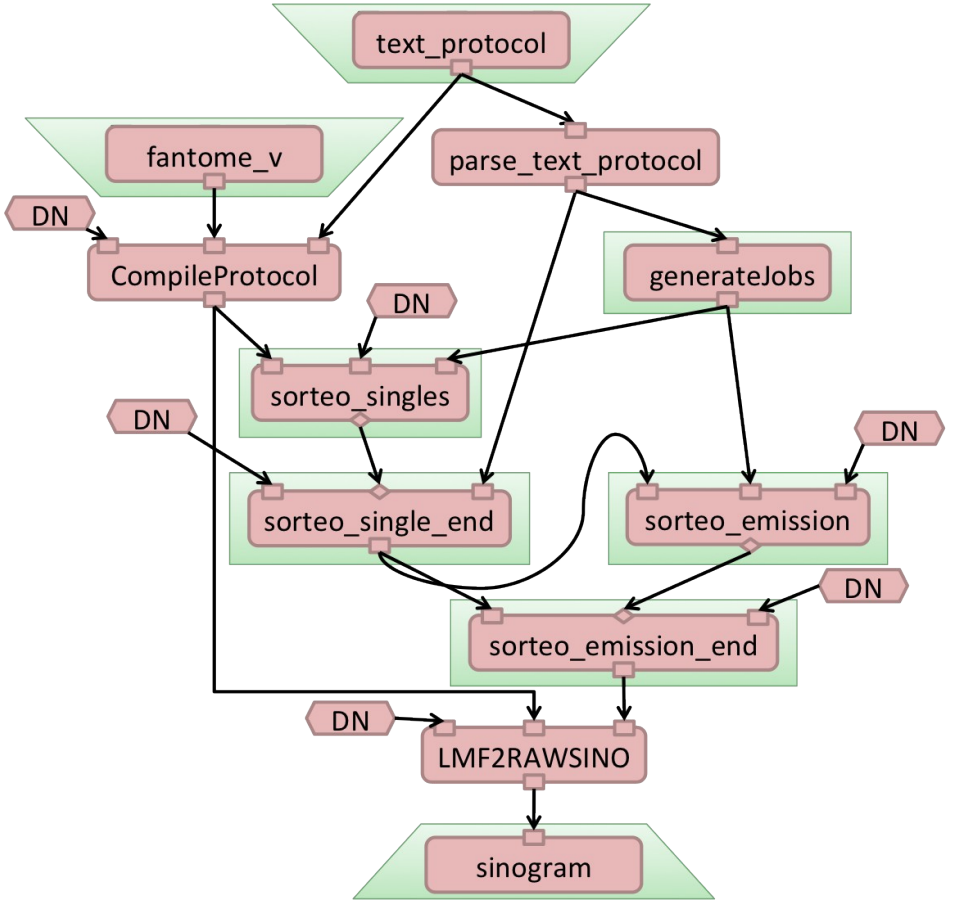
- KGRAM query processing

```
Q   SELECT ?name ?date
    WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
      Q1    FILTER (CONTAINS (?name, 'Bob')) }        Q2
```

- Asynchronous execution



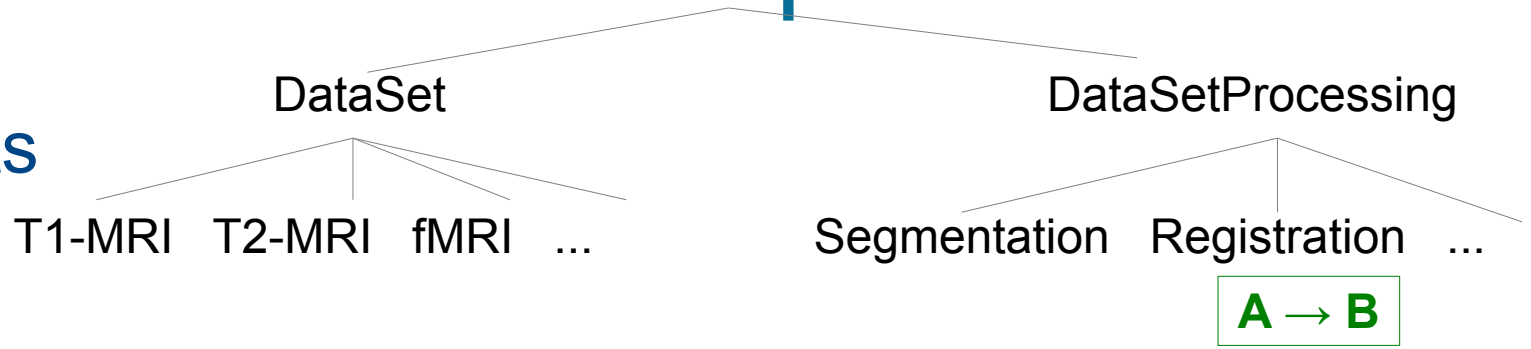
Data analysis workflows



Data-driven compute-intensive workflow engine MOTEUR2

Provenance capture

- Ontology
  - Concepts & **Rules**



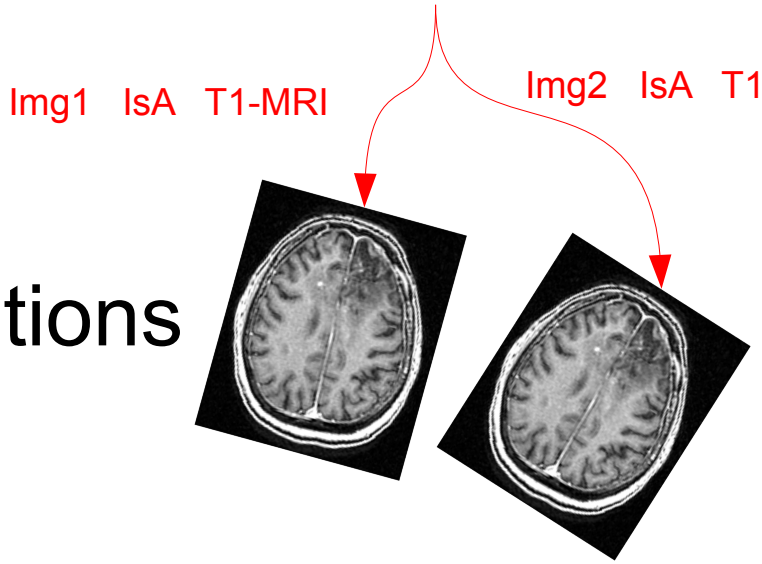
- Annotations

- Processing

Provenance capture

- Ontology
  - Concepts & **Rules**

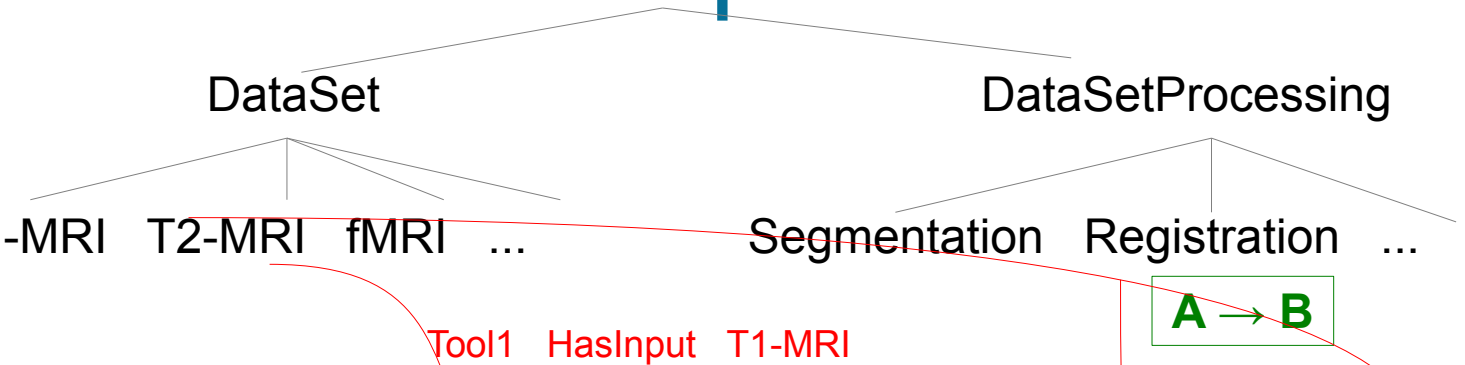
- Annotations



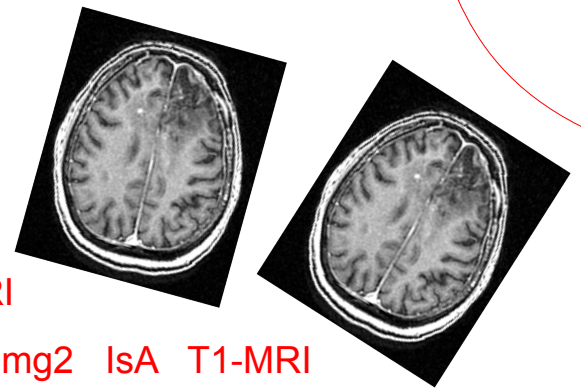
- Processing

Provenance capture

- Ontology
  - Concepts & **Rules**



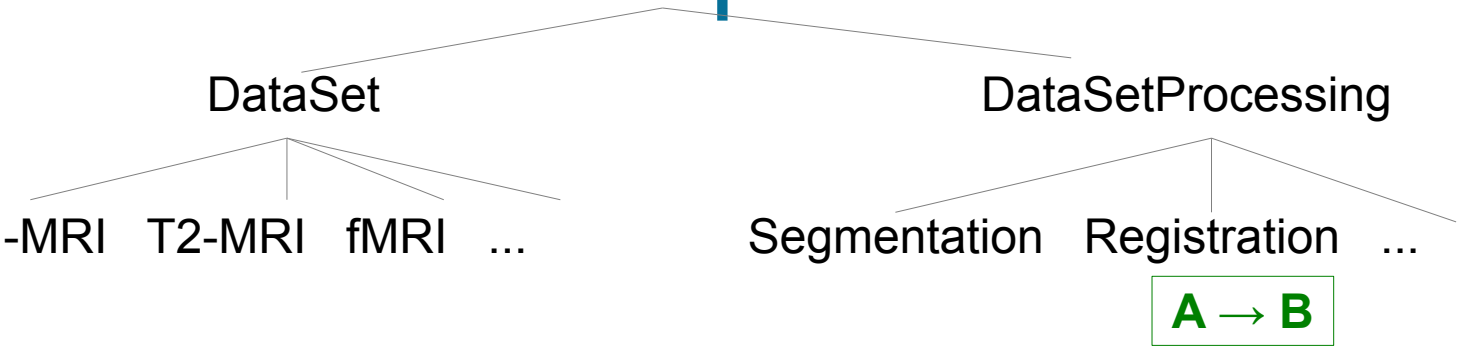
- Annotations



- Processing

Provenance capture

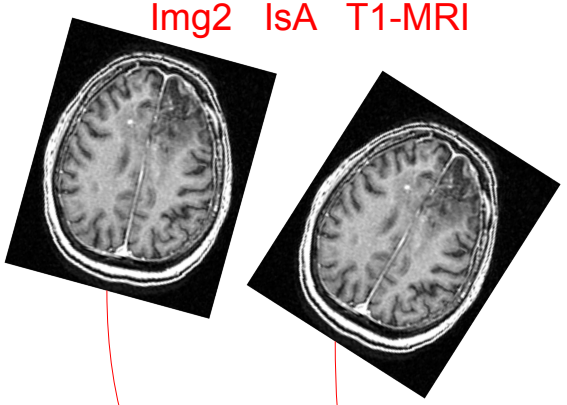
- Ontology
  - Concepts & **Rules**



- Annotations

Img1 IsA T1-MRI
Img2 IsA T1-MRI

Tool1 HasInput T1-MRI
Tool1 HasOutput Transfo
Tool1 IsA Registration



- Processing

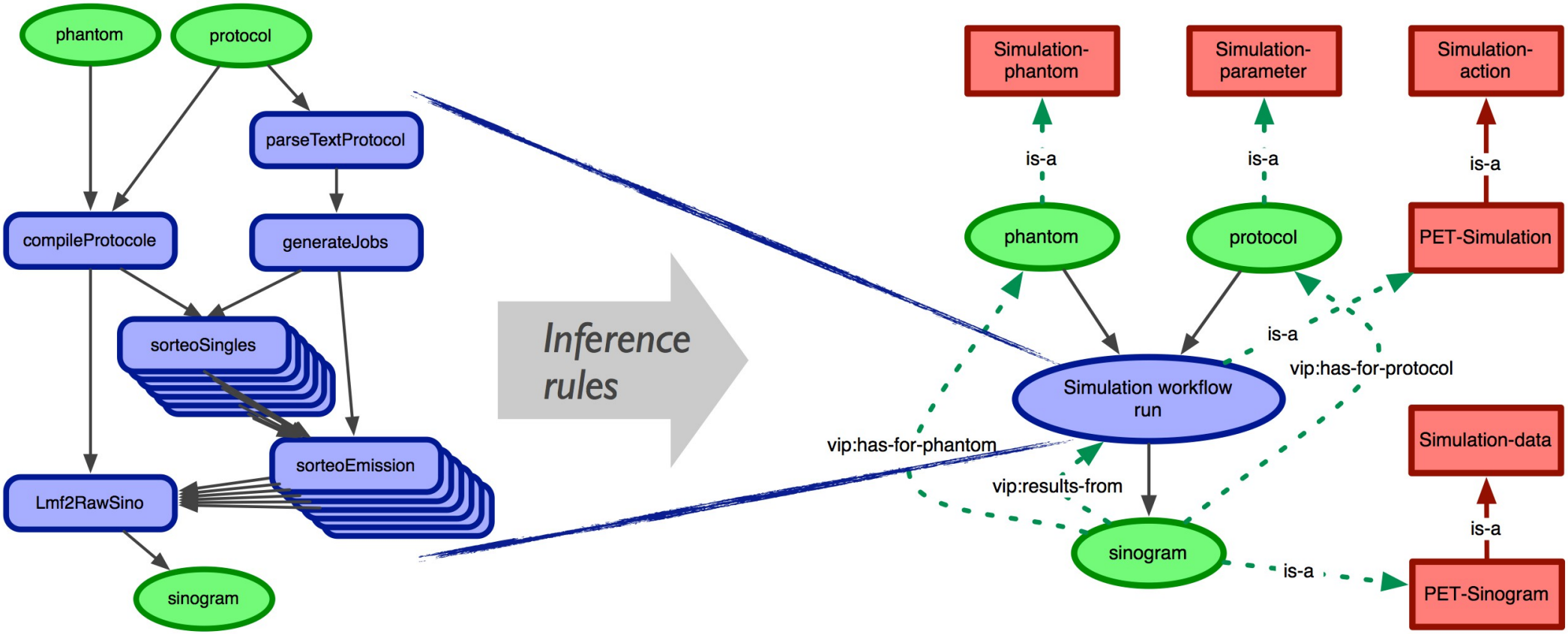
Img1 IsProcessedBy Tool1
Img2 IsProcessedBy Tool1

Tool1 Produced Transfo1

Transfo1 IsA GlobalTransfo

Provenance summarization

- Fine-grained annotation traces generated at run-time
- Summary generated by inference rules application
  - Produce relevant and human-tractable experiment summaries



Conclusions

- Query-based data federation grounded on semantic web standards (SPARQL, RDF, RDFS)
  - Emphasis on query language expressivity
  - Support for both horizontal and vertical data partitioning
  - Broad applicability (given that ontologies are available)
- Ontology-based
  - Reference model for data alignment and query terms
- Towards support of Multi-Centric studies
  - Heterogeneous databases (data models, database engines)
  - Inherently distibuted data sets
  - Cross-domains (translational research) support through Linked Data