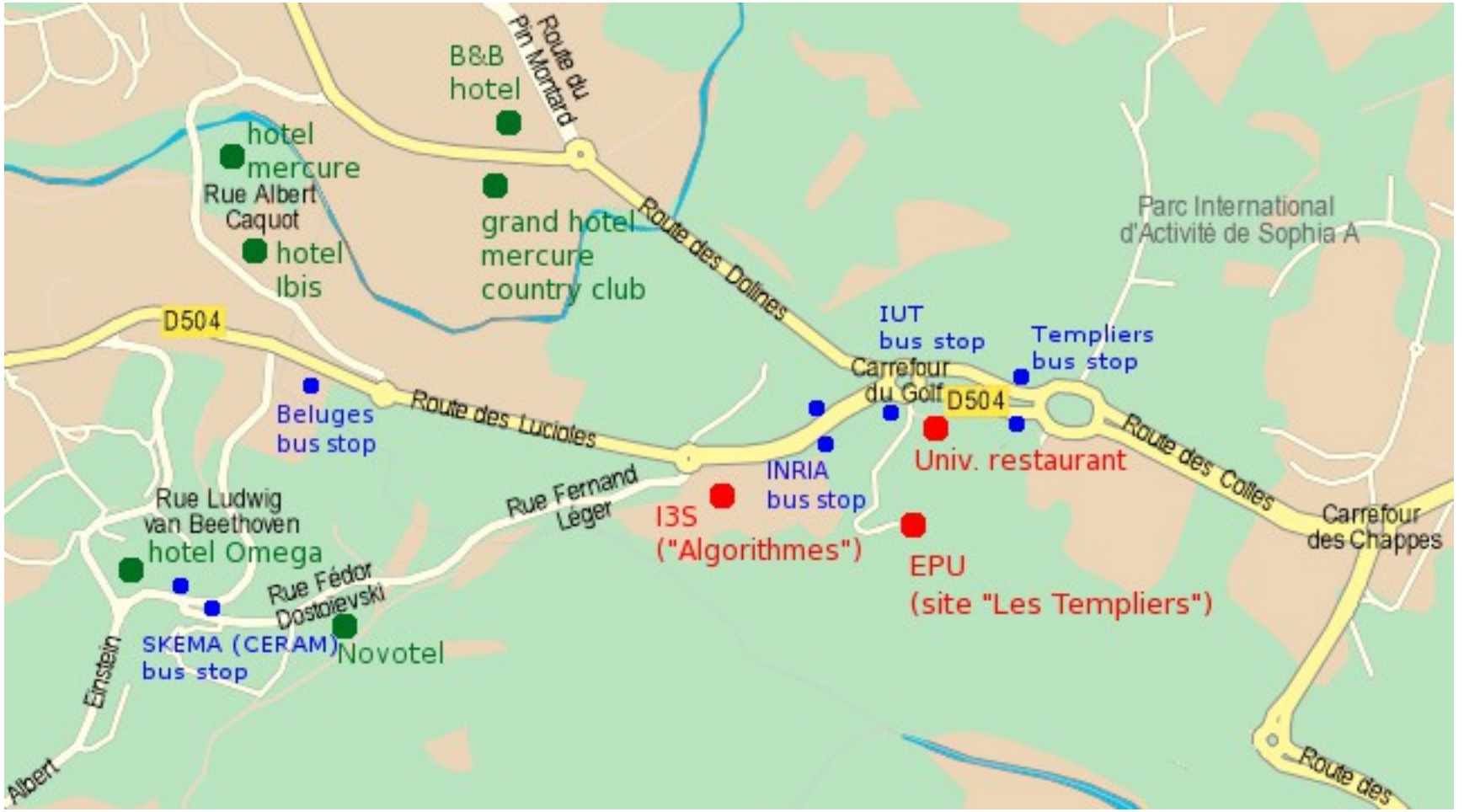
CrEDIBLE

fédération de données et de ConnaissancEs Distribuées en Imagerie BiomédicaLE

Distributed Biomedical Imaging Data and Knowledge federation

Johan Montagnat

Venue information



- All lunches served on site
- Diner at Omega hotel today, 19:30
- Taxis will be organized for returning to the airport

WiFi

- Eduroam network for those who are registered
- Unice-Hotspot with individual accounts for the others

What is CrEDIBLE?

- National project funded by CNRS (French National Research Center)
  - Part of CNRS inter-disciplinary mission
  - “MASTODONS” call (Challenge in Big Scientific Data)
  - Five partners involved in France (CNRS, INSERM, INRIA, U. Picardie, U. Lyon)
- Started in 2012, extended on a yearly basis
- Targets biomedical image resources federation
  - Data fusion, mediation, semantic alignment, querying, link with processing

What is CrEDIBLE?

- Scientific networking initiative
  - Budget for networking and dissemination
  - Opening new contact opportunities
- Follow-up on various prior initiatives from project partners
  - Medical data sharing
  - Distributed computing
  - Large-scale data analysis

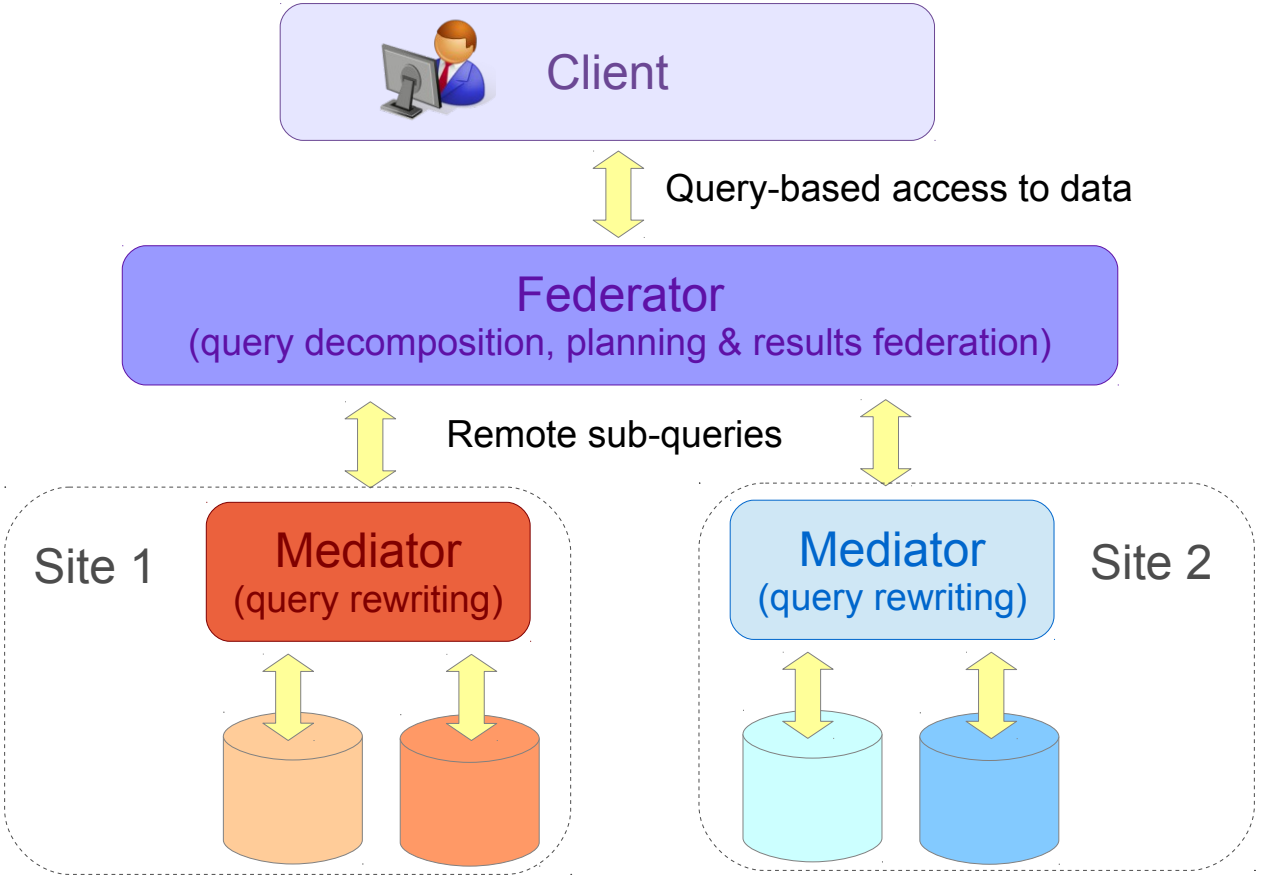
What is this workshop?

(https://credible.i3s.unice.fr/doku.php?id=2013_workshop)

- A multi-communities “think tank” opportunity
  - Intent to raise community discussions
    - 30 min. allocated per talk + panel for each session
  - Multi-disciplinary
  - Place for discussions
- How was it organized?
  - By invitation
  - Considering scientific themes of interest
  - Considering scientific challenges identified
- This year is the second edition
  - 2012 report: https://credible.i3s.unice.fr/doku.php?id=rapports

Biomedical data mediation & federation

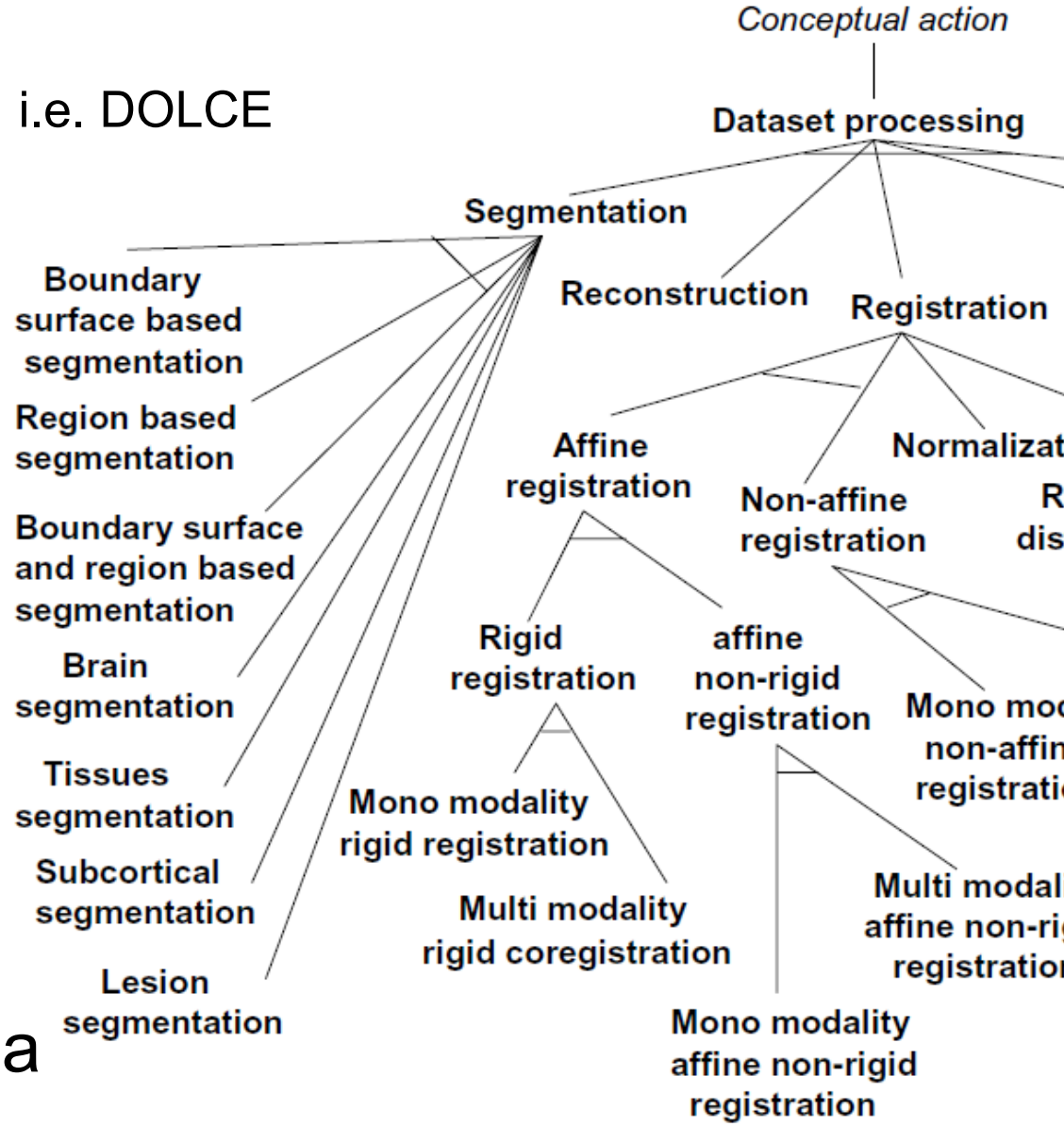
- Heterogeneous databases schema mediation
- Data federation through distributed querying and query rewriting



- Scientific experiment support platform:
  - raw data + models + processing results + models + provenance...

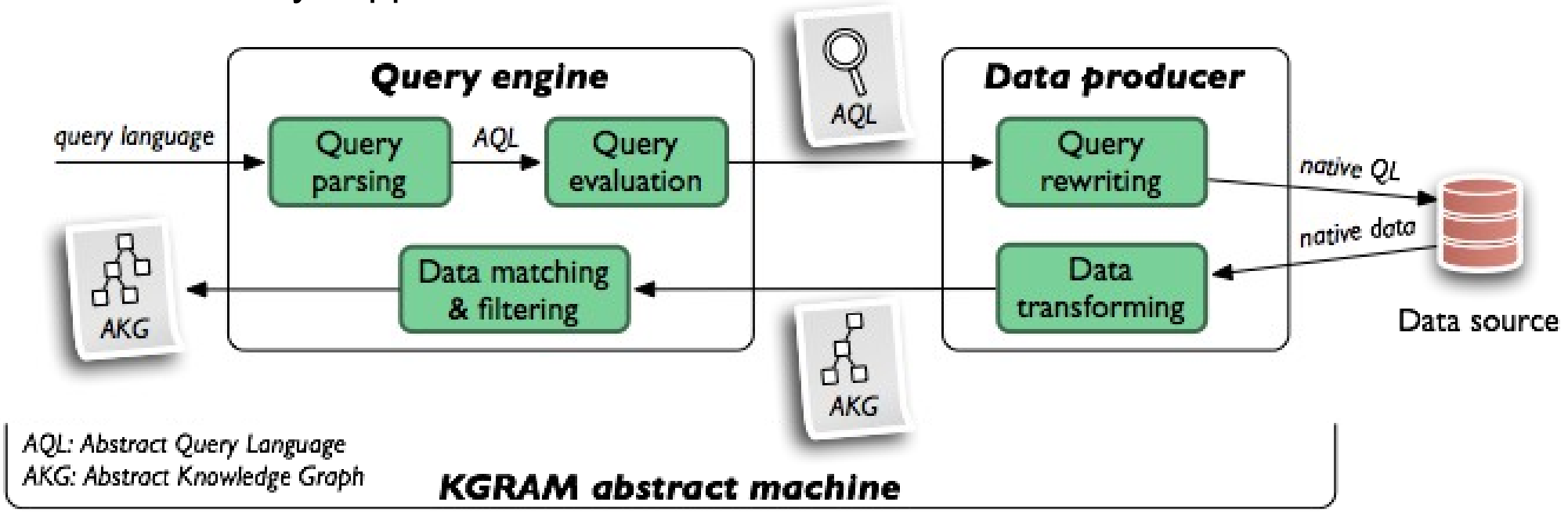
Semantic reference

- Application ontology
  - 3-levels structure
    - one Foundational ontology: i.e. DOLCE
    - Several Core ontologies
    - Several Domain ontologies
- Covering
  - Data sets
  - Data processing tools
  - Scientific measure
  - Medical context
  - Data provenance
  - ...
- Domain-specific rules
  - Inference abilities
- Derived relational schema



From relational to semantic medical stores

- Databases federation
  - Requires relational model mediation
  - Based on a semantic reference to derive the federated schema
- Based on KGRAM (Knowledge Graph Abstract Machine)
  - Semantic query engine enabling:
    - Heterogeneous data models
    - Fully supports SPARQL v1.1



Distributed Query Processing

- KGRAM query processing

Q

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

- 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
```

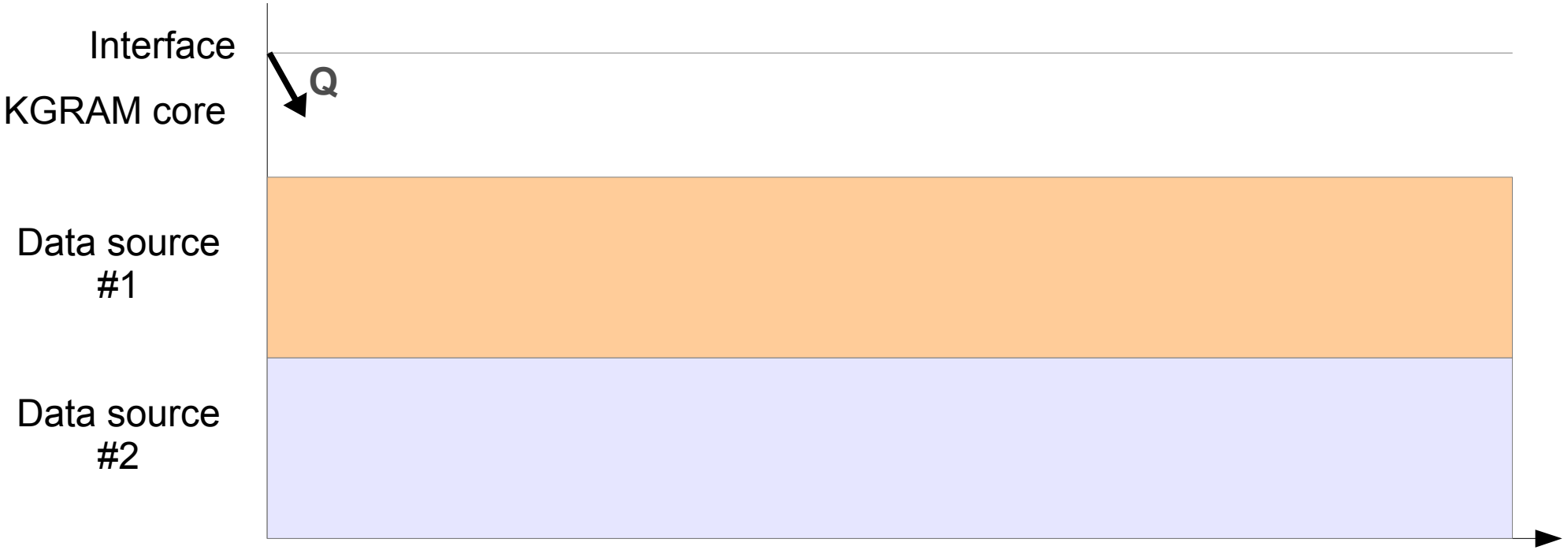
- Asynchronous execution

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

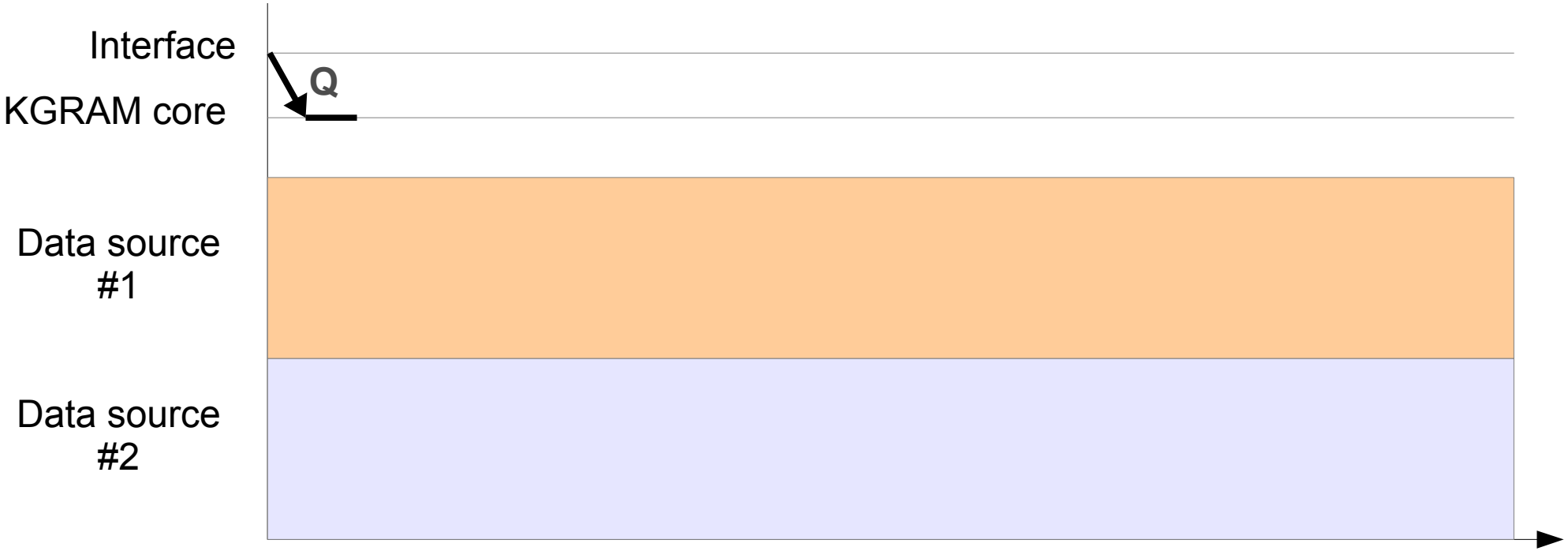


Distributed Query Processing

- KGRAM query processing

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

- Asynchronous execution

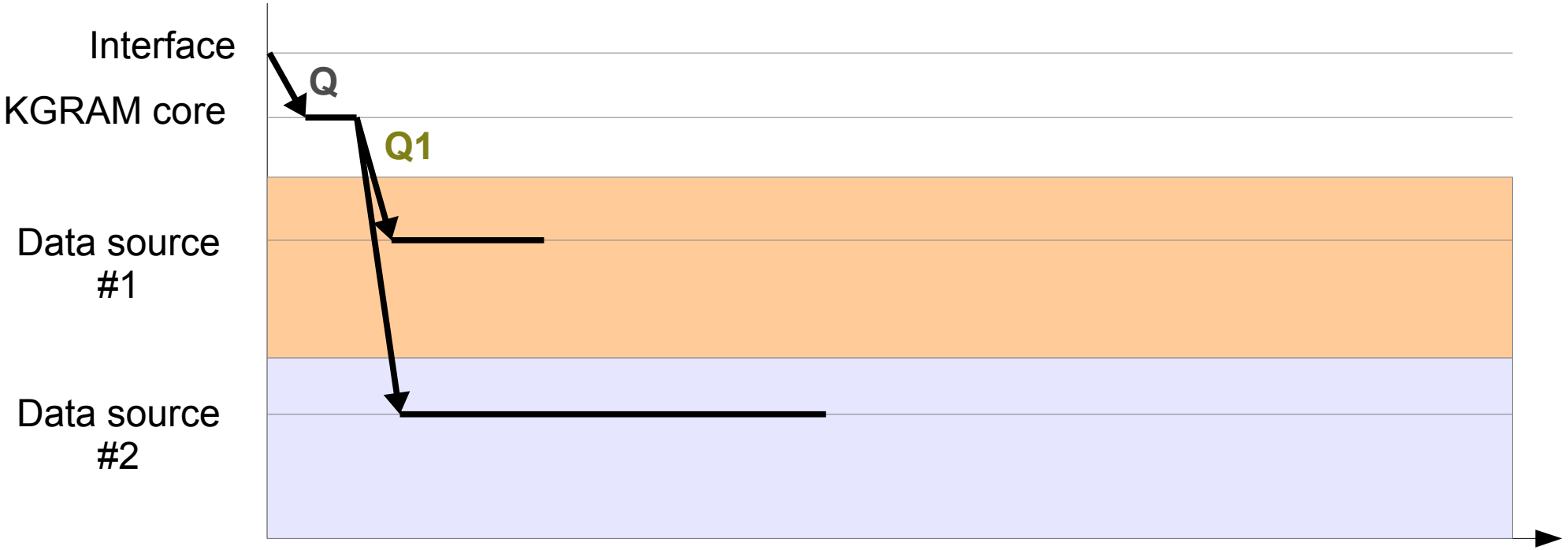


Distributed Query Processing

- KGRAM query processing

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

- Asynchronous execution

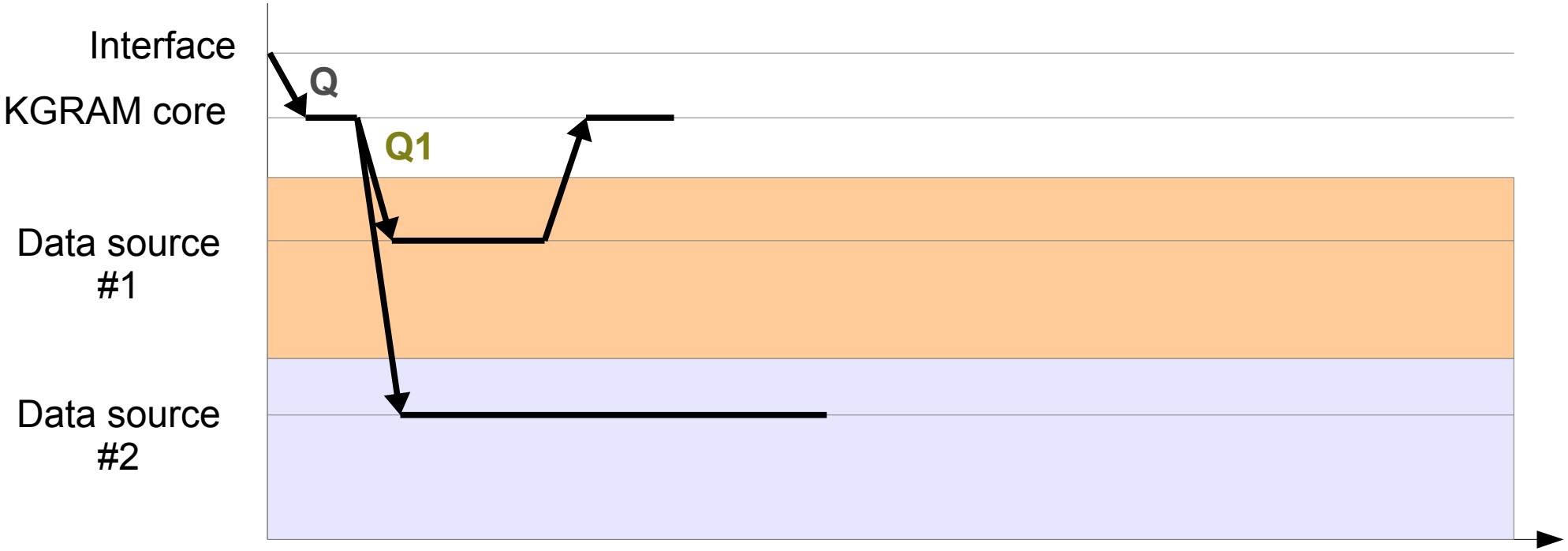


Distributed Query Processing

- KGRAM query processing

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

- Asynchronous execution

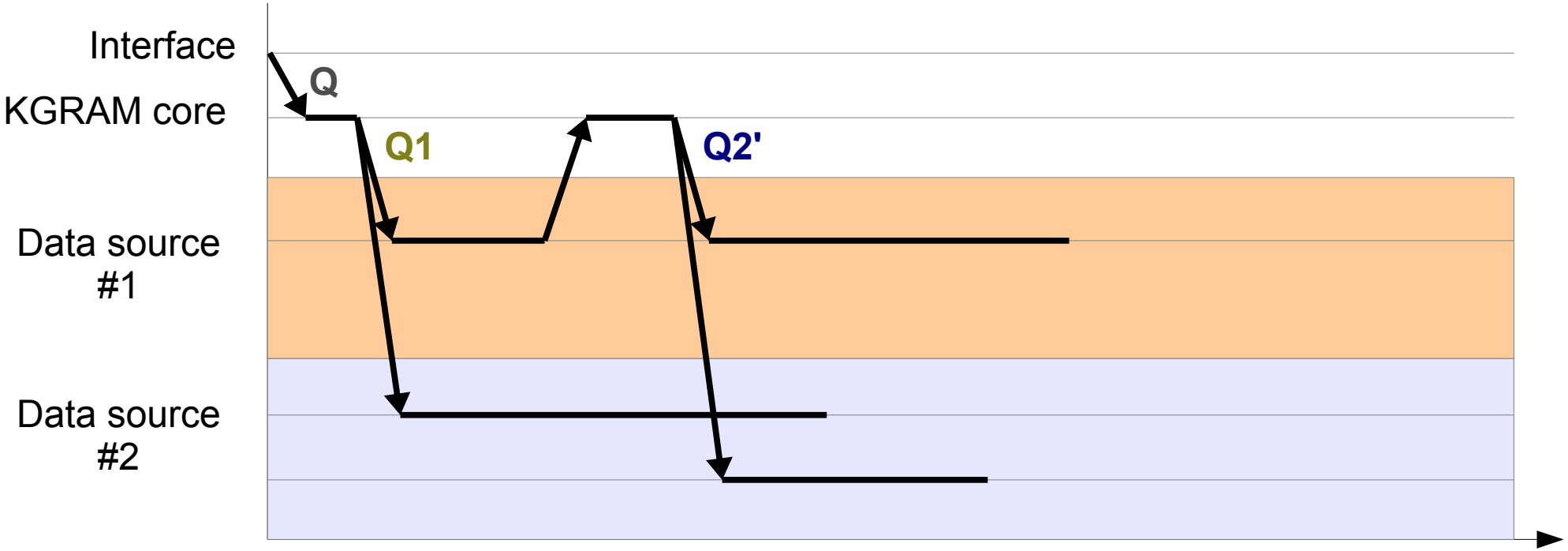


Distributed Query Processing

- KGRAM query processing

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

- Asynchronous execution

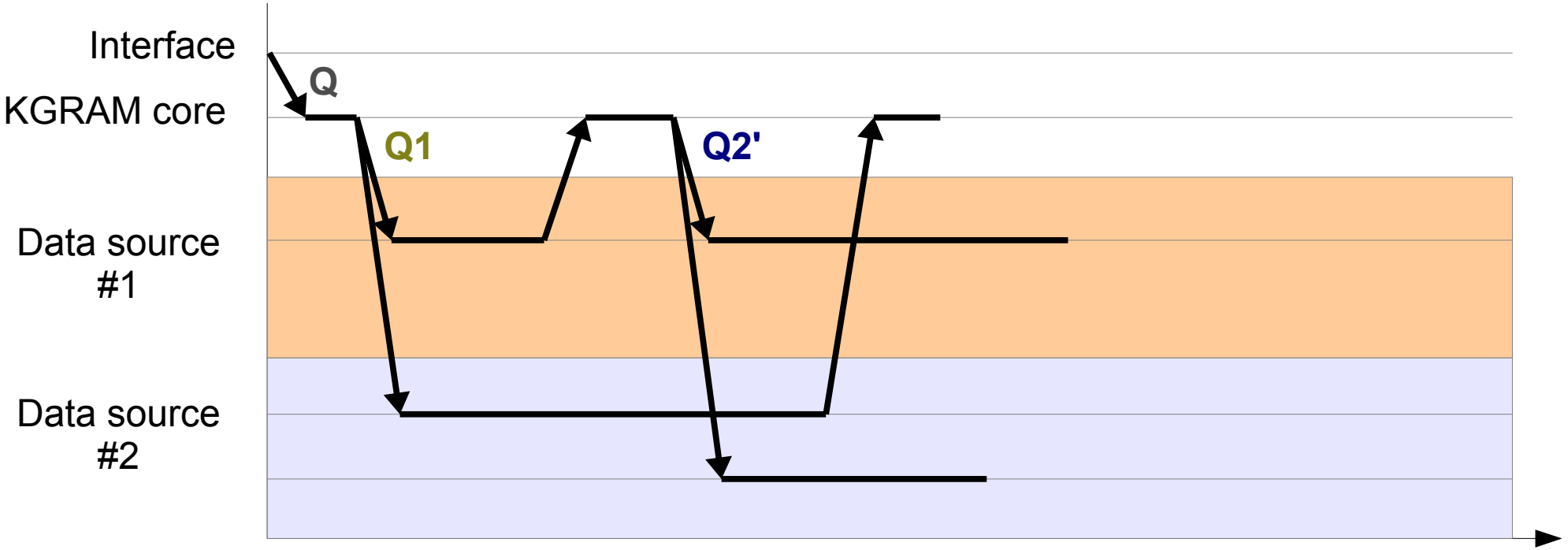


Distributed Query Processing

- KGRAM query processing

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

- Asynchronous execution

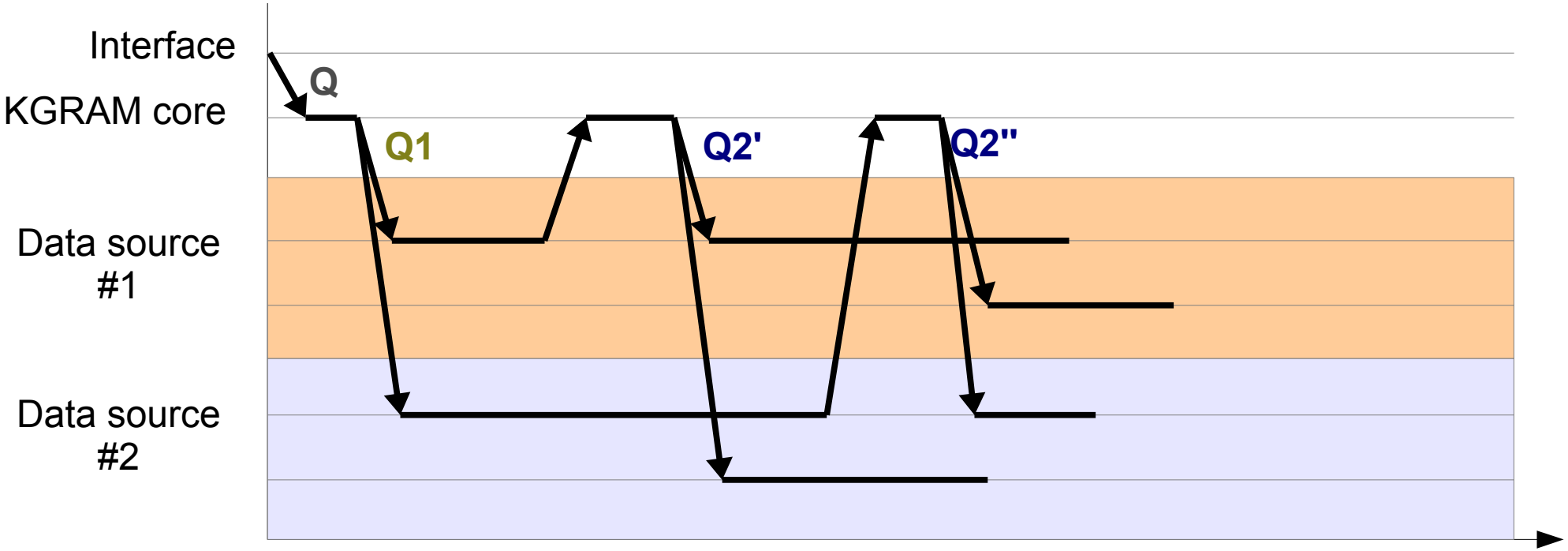


Distributed Query Processing

- KGRAM query processing

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

- Asynchronous execution

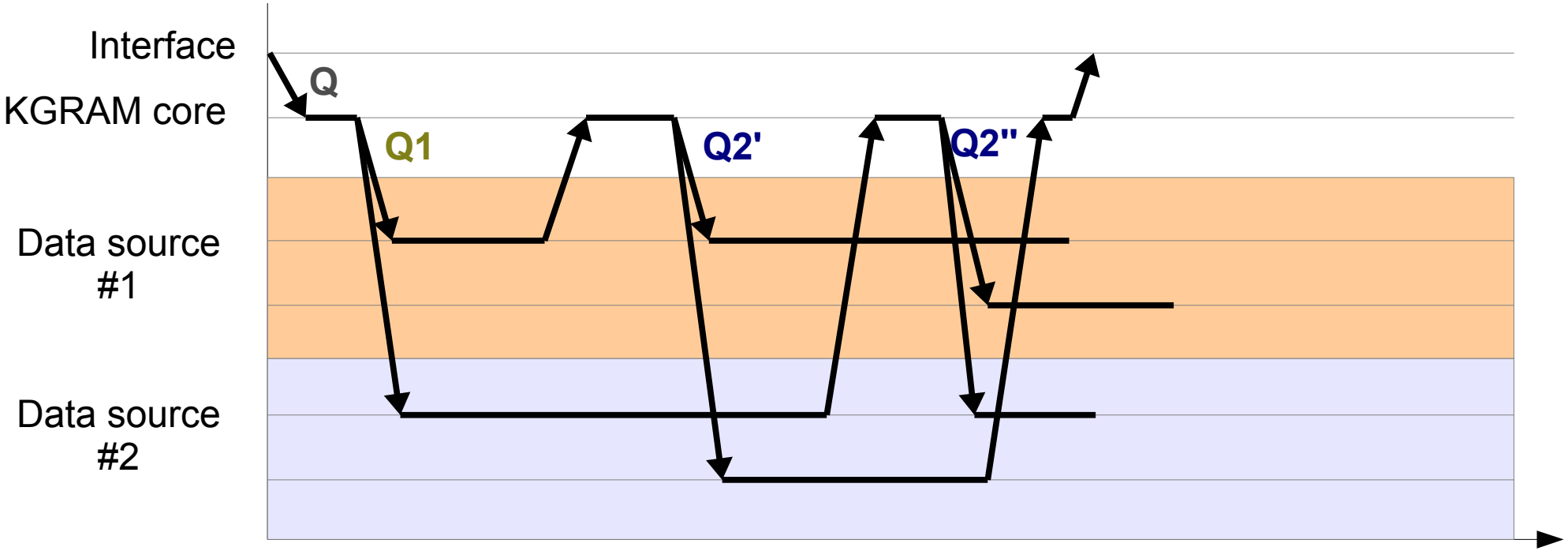


Distributed Query Processing

- KGRAM query processing

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

- Asynchronous execution

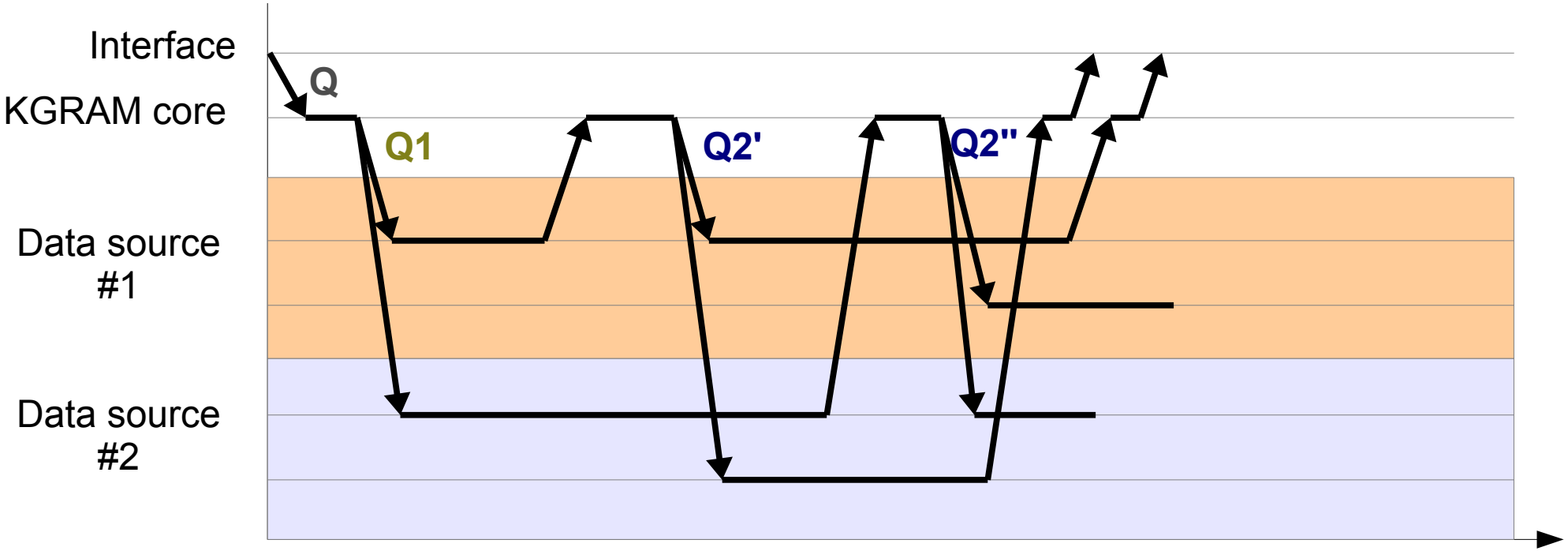


Distributed Query Processing

- KGRAM query processing

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

- Asynchronous execution

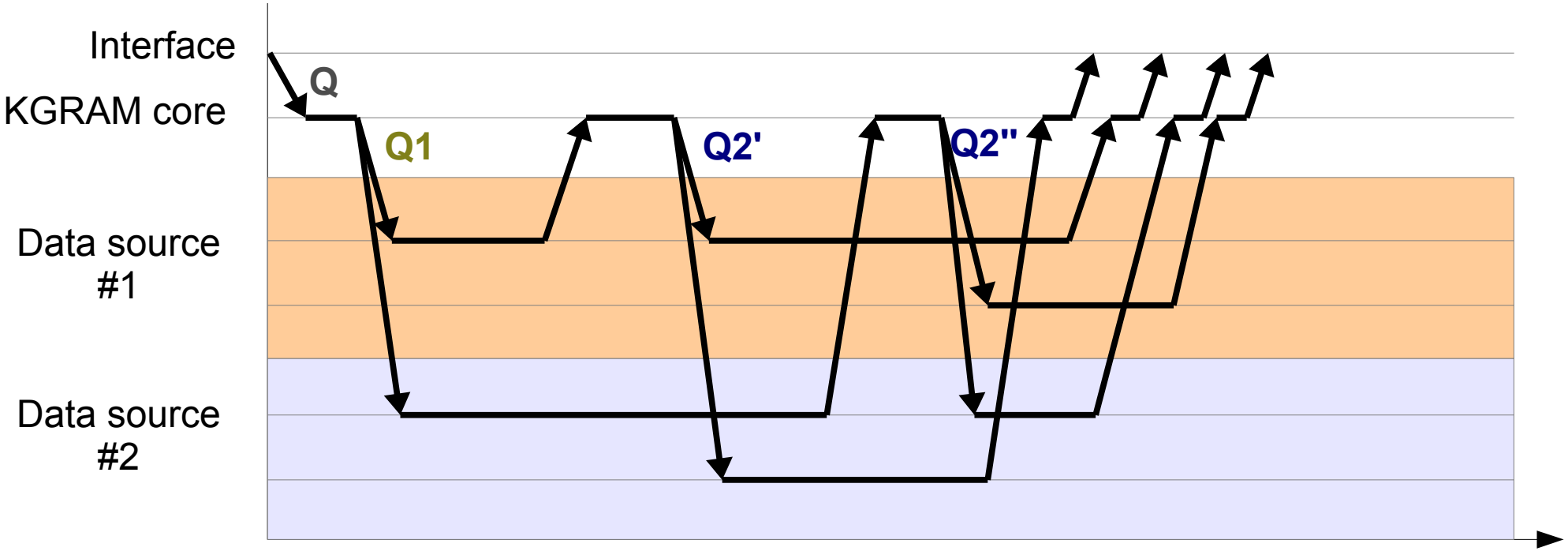


Distributed Query Processing

- KGRAM query processing

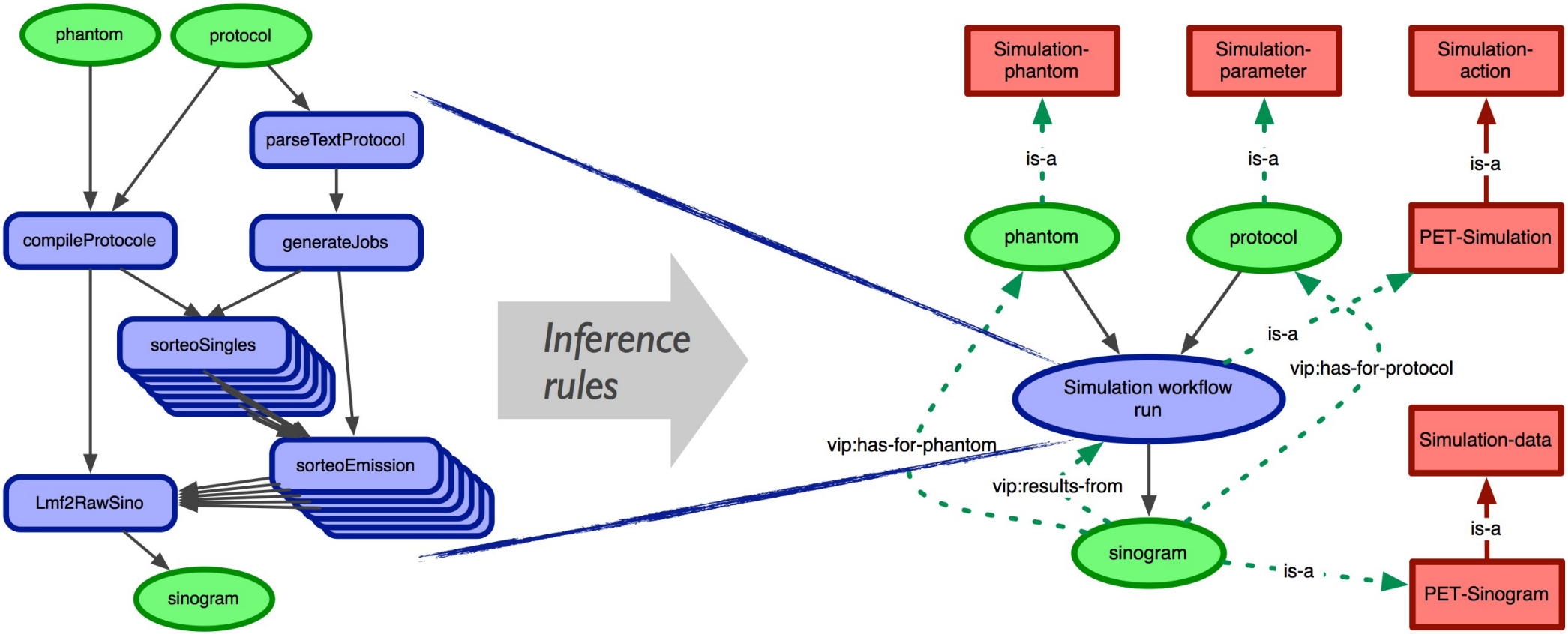
```
Q   SELECT ?name ?date
    WHERE { ?x foaf:name ?name . ?x dbpedia:birthDate ?date .
      Q1    FILTER (CONTAINS (?name, 'Bob')) }        Q2
```

- Asynchronous execution



Knowledge exploitation for data analysis

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



Workshop schedule

- Wednesday 2nd (afternoon)
  - 14:30 Session on data repositories and use of clinical data
  - 19:30 Diner at hotel Omega
- Thursday 3rd
  - 9:00 Session on Biomedical ontologies
  - 14:00 Session on Data mediation
  - 16:15 Session on Data federation
- Friday 4th (morning)
  - 9:00 Session on Knowledge graphs and reasoning

Data repositories for secondary use of clinical and research data

- Conveners: Bernard Gibaud, Alban Gaignard
- Scientific questions:
  - Data indexing, precision of the vocabulary?
  - Data provenance, amount of data needed?
  - Data federation level required? Data sources to federate? Existing data model?
  - Access control, dealing with multiple access policies; what data to control?

Biomedical ontologies

- Conveners: Bernard Gibaud, Gilles Kassel
- Scientific questions
  - Modelling mesurable related entities and their relationships
    - Modeling of complex observation data such as images
    - Modeling time varying phenomena
  - Relation with foundational ontologies
  - Relation to existing ontologies and compatibility issues

Data mediation

- Conveners: Franck Michel, Johan Montagnat
- Scientific questions
  - Are taxonomies and ontologies the appropriate reference models?
  - Is SPARQL the appropriate query language?
  - How to mediate various data sources?
    - Periodic, statical transformation (ETL)
    - Query on-the-fly
  - How to ensure access control in an heterogeneous deployment?
    - At coarse-grain (data repository)
    - At fine-grain (entity within a data repository)

Data federation

- Conveners: Johan Montagnat, Olivier Corby, Alban Gaignard
- Scientific questions
  - Trade-off between expressiveness and performance of query language? Is SPARQL 1.1 appropriate?
  - Performance impact of distribution?
  - Scalability? Experimentations at large-scale?
  - Reliability? Impact of lower reliability? Partial results?

Graphs and reasoning

- Conveners: Catherine Faron-Zucker, Olivier Corby
- Scientific questions
  - How to process large RDF graphs?
  - Trade-off between the amounts of data to process and the reasoning capabilities of the system.
  - Other scalability opportunities when addressing data querying
  - Large graphs visualization

Let us get started

- Session 1: Data repositories for secondary use of clinical and research data (14:30 – 18:15)
  - Marc Cuggia (U. Rennes 1, France), Indexing of medical data
  - Maryann Martone (UCSD, USA), Experience with the development and operation of the Neuroscience Information Framework (NIF) portal
  - Serena Villata (INRIA, FR), Applying open data provenance and licensing to biomedical data
  - coffee break
  - Khalid Belhajjame (U. Manchester, UK), Research Objects: Preserving Scientific Workflows and Their Provenance
  - Charles Marion (Kitware), Visualization and analysis of medical data through the Internet
  - Panel discussion. Moderators: Bernard Gibaud, Alban Gaignard