PÔLE **GLC** : **G**ÉNIE DU **L**OGICIEL ET DE LA **C**ONNAISSANCE

Research group on

Software and Knowledge engineering

Scientific leader

Mme Blay-Fornarino

http://glc.i3s.unice.fr/

Staff

~30 permanent researchers

~50 non permanent staff

(PHD students, postdocts, engineers...)

Software & Knowledge Engineering ?

First, why would you “engineer” software?

Second, what is it like to “engineer” knowledge?

...and anyway, how is this all related?

Roughly, it boils down to...

...moving from here to there

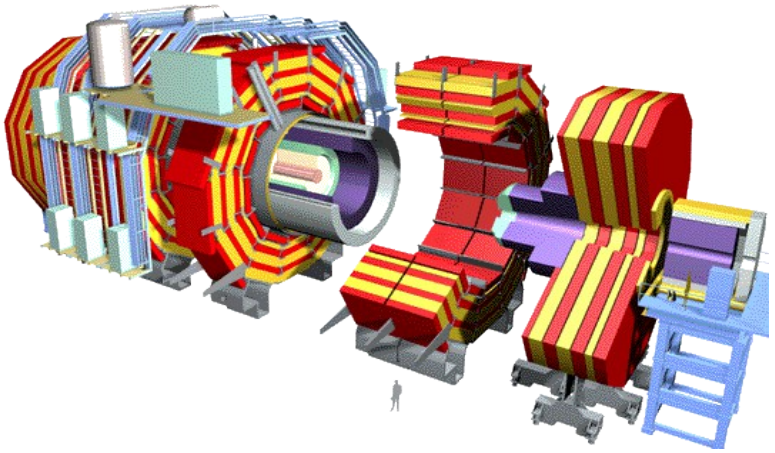
```
# Generic relations were moved in Django revision 5172
try:
    from django.contrib.contenttypes import generic
except ImportError:
    import django.db.models as generic

class Tag(models.Model):
    """
    A basic tag.
    """
    name = models.CharField(maxlength=50, unique=True,
            db_index=True, validator_list=[isTag])

    objects = TagManager()

    class Meta:
        db_table = 'tag'
        verbose_name = 'Tag'
        verbose_name_plural = 'Tags'
        ordering = ('name',)
```

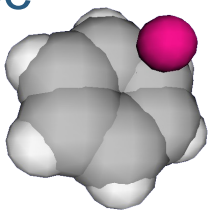
1000's lines of code
Centralized system
Client-server connections

Up to millions of computing units
Worldwide distribution
Heterogeneous components

Need for storage capacity

- High Energy Physics
  - TBs of data generated per second as the output of the Large Hadron Collider
- Life Sciences
  - Large-scale data distribution (radiology centers)
  - Human genome sequencing: from 3 B$ to 1 k$ in 12 years
- Astronomy, Astrophysics and Earth Sciences
  - Astronomical data, multi-spectral images
  - Gaia satellite to produce 10s of PBs of data
- Social networks
  - 100 Billions mails + 100 Millions tweets daily
  - In the next 2 hours, 80 years worth of video will be uploaded in youtube
- Environmental Sciences
  - All biosphere data
- Industry
  - PBs of data acquired by jet sensors per hour of flight
- ...





Need for computing power

- Data analysis needs proportional to the amount of data collected
  - Commonly, the volume of processed data is many times the volume of raw input data
- Increasing use of computer simulation
  - In silico experiments: faster, cheaper “pre-experimentation”
  - Prototypes design and testing
  - Modeling and assistance to understanding of complex proce
  - ...
- New approaches to data analysis
  - Blind search and brute force computing
  - Mining correlations out of wealths of data
  - ...

Why so much software complexity?

- Amount of data acquired is growing exponentially with
  - Doubles every 18 months
  - EBs of data generated yearly (1EBs), ZBs total (1ZBs)
  - More data stored in the last 10 years than in the rest of histo
- Storage capacity is growing exponentially with time
  - Doubles every 2 years
- Number of transistors on chip is growing exponentially with time
  - Doubles every 18 months (Moore's law), hold true for 40 yea
- Network bandwidth is growing exponentially with time
  - Internet backbone bandwidth doubles every 6 months

All exponential, yet not equal

So what is wrong?

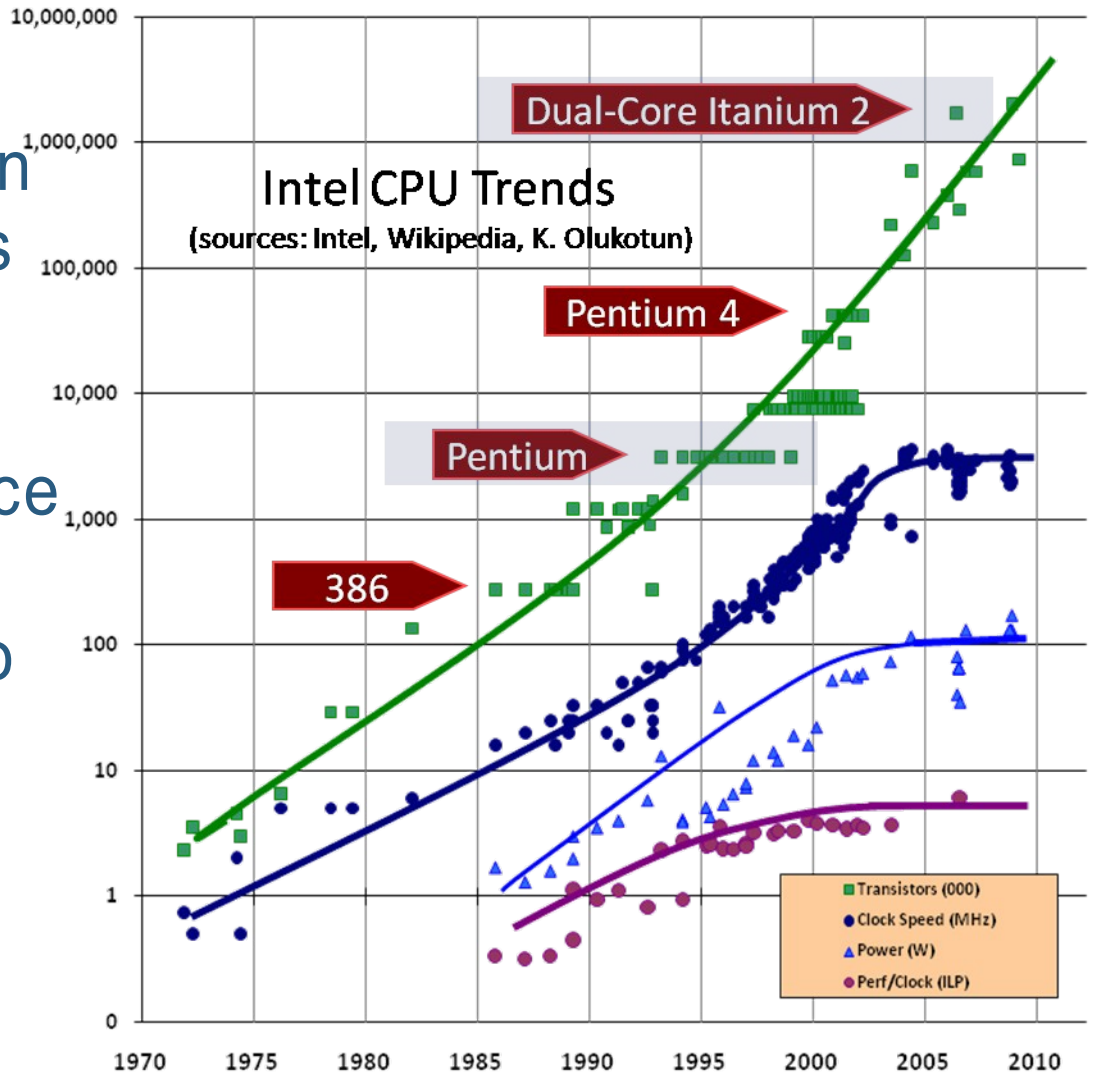
Not to confuse with processing power

Not to confuse with access rate

What about software development?

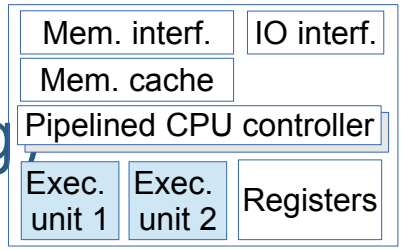
The end of sequential computing

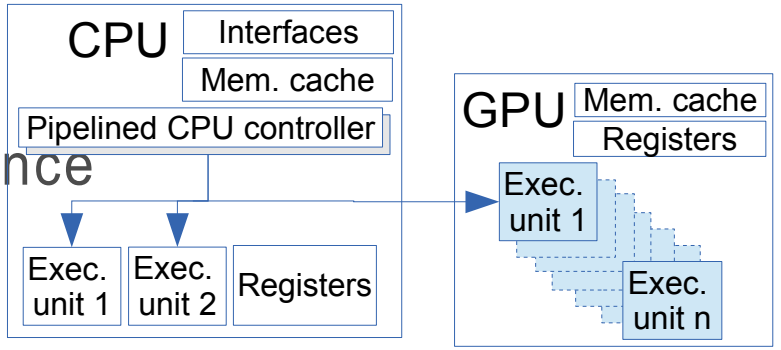
- CPU performance increase slows down
  - Moore's Law still applies to transistors
  - CPUs power consumption reached acceptable limits around 2002
  - Frequency suddenly capped as a consequence
  - ILP (Instruction Level Parallelism = capability to process multiple instructions per clock bit) faces chip complexity limitations



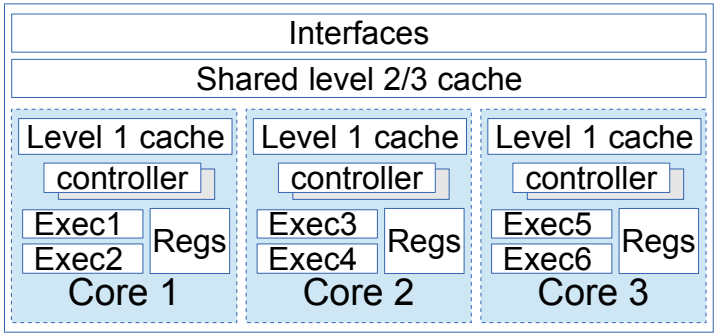
The end of sequential computing

- Multiply execution units
  - Doubling CPU execution units (hyperthreading)
    - 15-30% performance increase
  - GPU execution units (GPGPU)
    - 100s computing units
    - Very application-dependent performance





- Multiple cores on chip
  - One CPU, several cores
  - Some sharing
    - Memory caches (with coherency)
    - Shared-memory (through cache)
    - Inter-cores message passing



The end of humanly manageable data sets

- Large and distributed data sets
  - 1000s to millions of files, distributed over hundreds of places
  - Cannot be handled manually
  - Cannot be visualized
- The need to link data
  - Every data is digital
  - Data becomes openly accessible over the Web
  - Data entries may be correlated to any other, worldwide
- The need to reuse data
  - Secondary use in contexts for which data was not acquired becomes common
  - Data preservation needs commonly reach decades
  - Detailed data description is vital for secondary use

Heterogeneity and mobility

- Heterogeneity in computing resources
  - Mainframes, workstations, laptops, smartphones, specialized sensors...
- Heterogeneity in data sources
  - Data acquired in many different contexts, using different encoding, file formats, data models...
- Smart devices widely available
  - Increasing number of smart / interconnected devices: Smartphones, high-tech devices (e.g. camera), everyday obje (e.g. fridge)Internet of Things
- Ever higher mobility
  - Ubiquitous access to network and high-level functionality

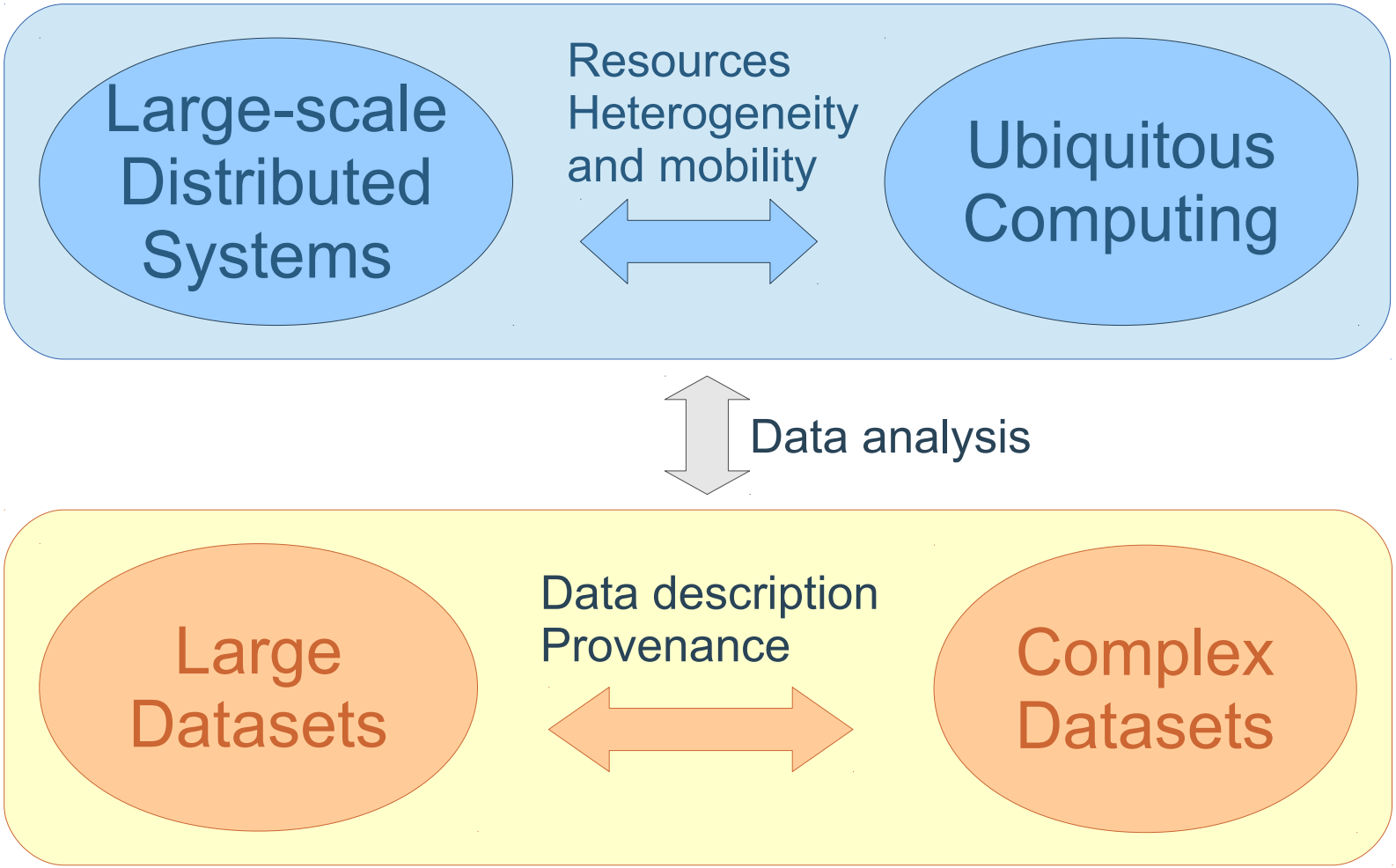
Software Engineering ?

- Taming software complexity
  - Distributed / parallel programming complexity
  - Heterogeneity
  - Adaptability
- Taming distributed infrastructures complexity
  - Large scale, distributed
  - Heterogeneity
- Improve performance and reliability
- Man-Machine interaction

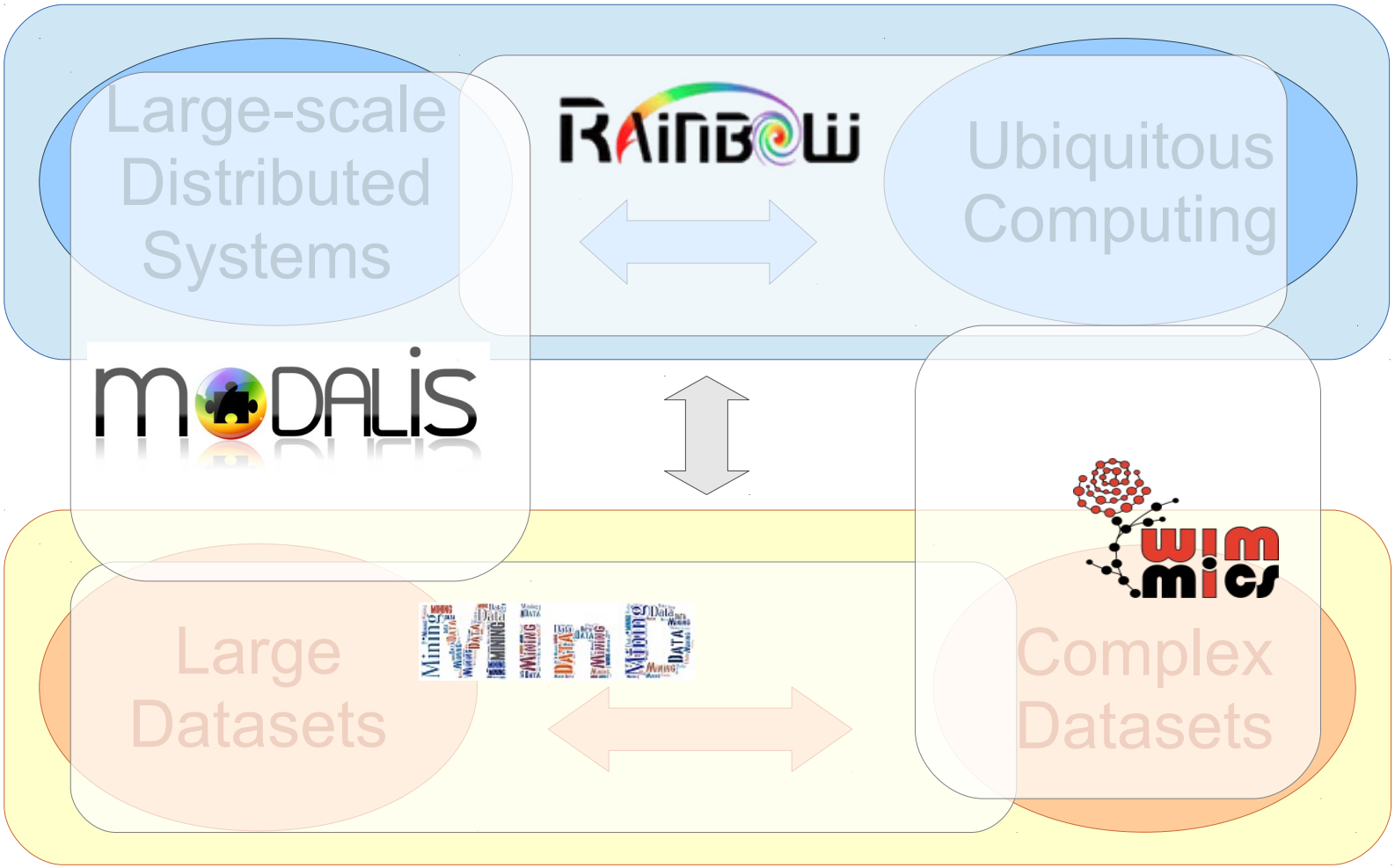
Knowledge Engineering ?

- Taming data complexity
  - Data heterogeneity
  - Data volume
- Mining data
  - Data classification & identification
  - Finding needles in haystacks
- Semantic data
  - Describing and linking data
  - Inferring new knowledge from known data

Software **AND** Knowledge Engineering ?



Software **AND** Knowledge Engineering ?



Distributed Computing

Big Data

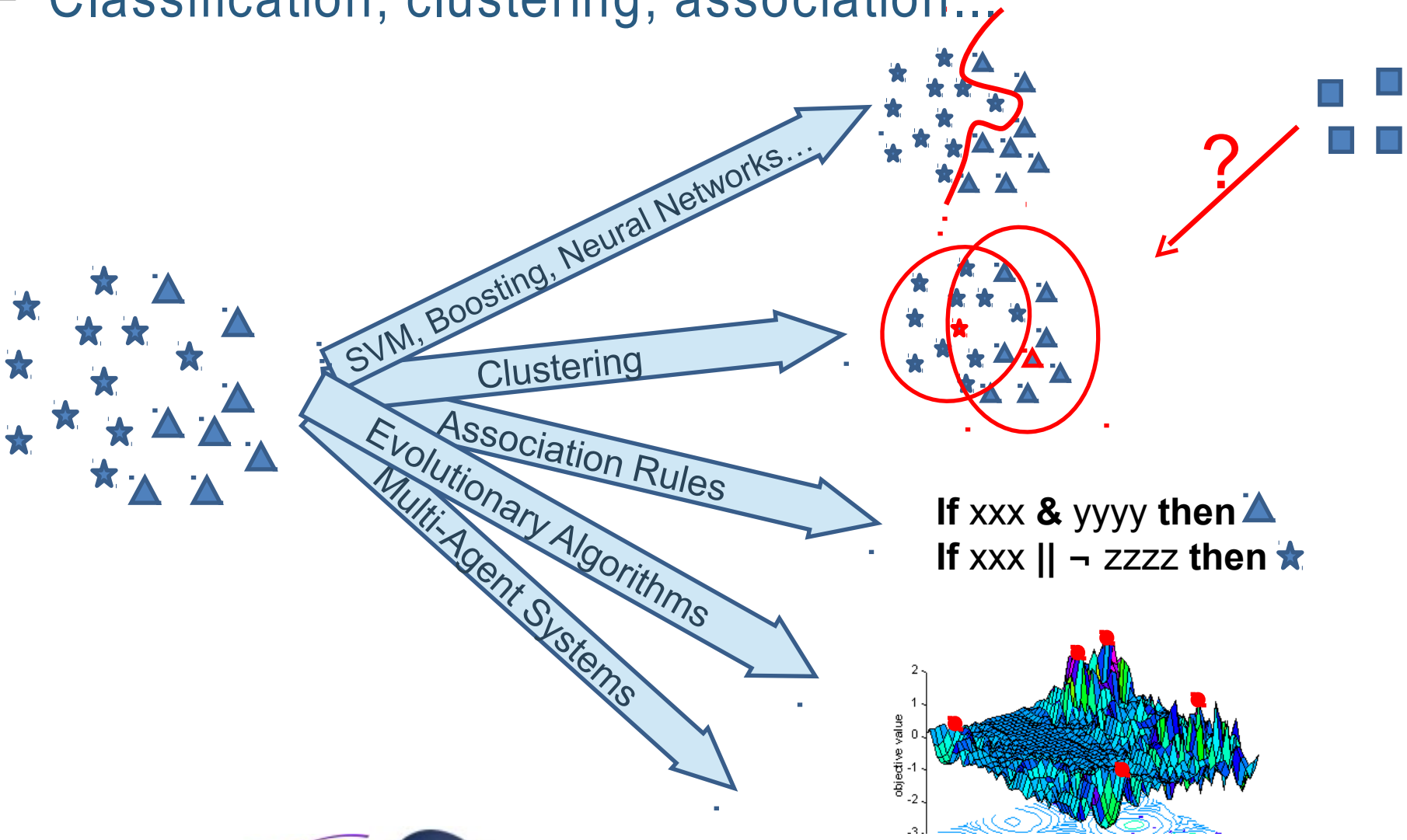
Pôle GLC = MinD + MODALIS + RAINBOW + Wimmics

Ubiquitous computing & mobility

- Context-aware applications
- High adaptability, hot reconfigurability
- Software-hardware components assembly

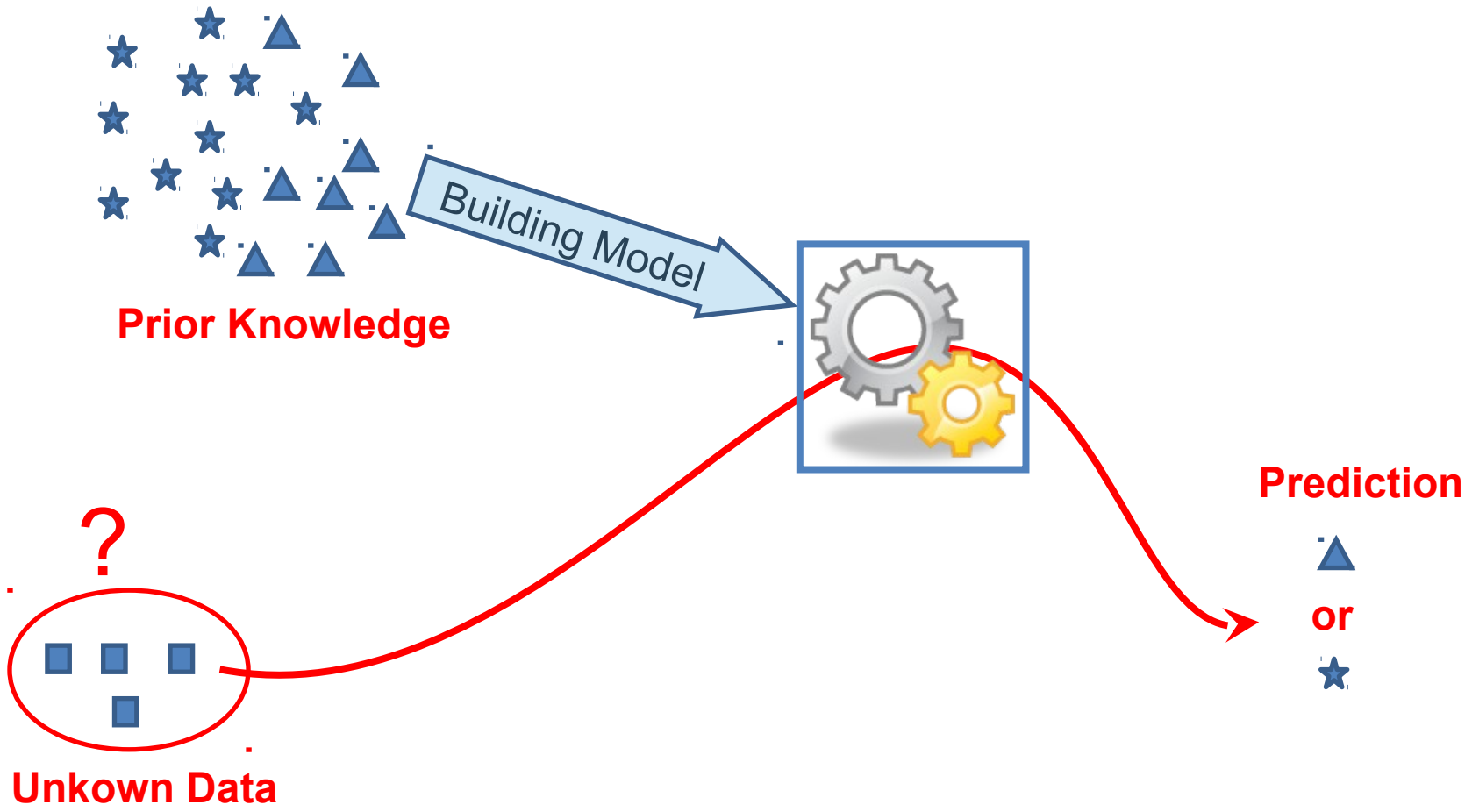
Data Mining

- Skills
  - Classification, clustering, association...



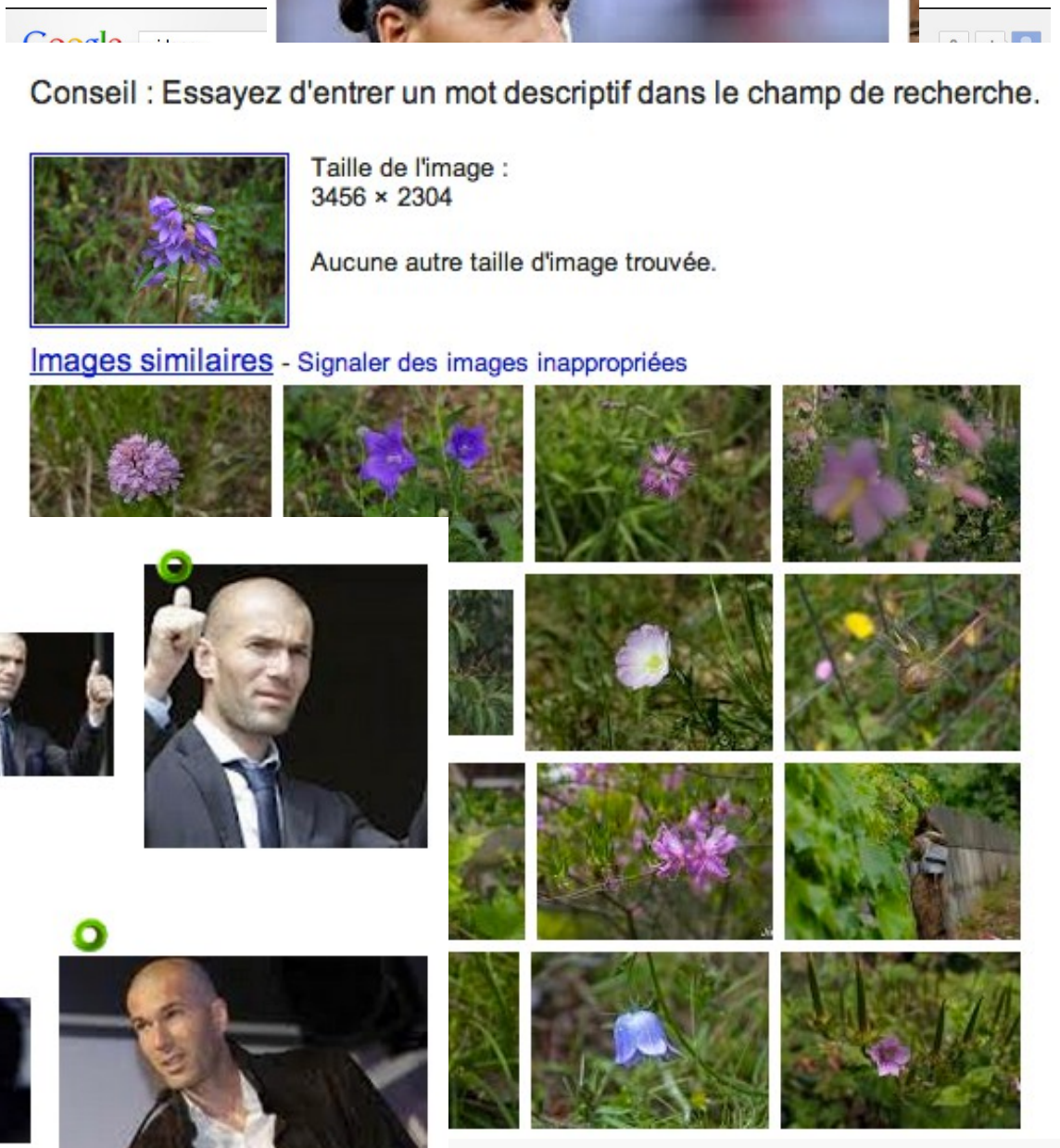
Decision making

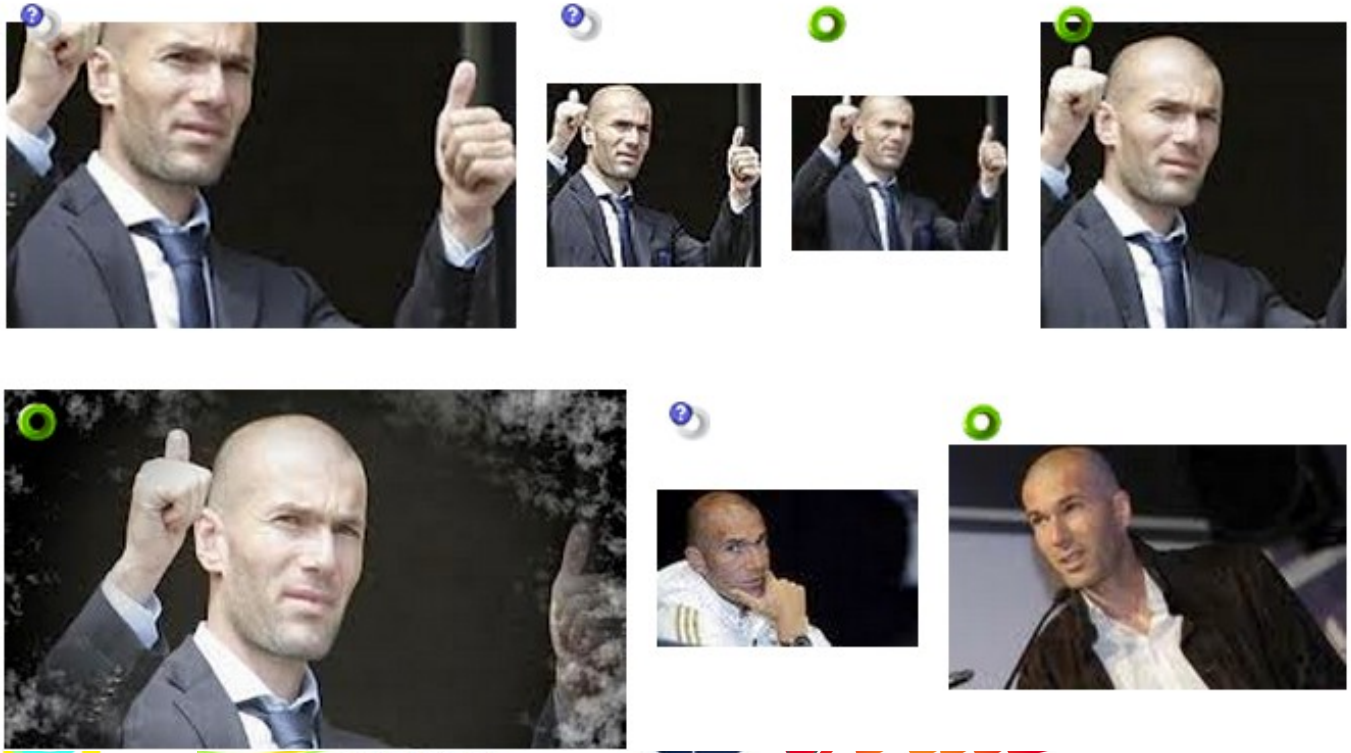
- Learning, Prediction



Application ex

- Image search engin



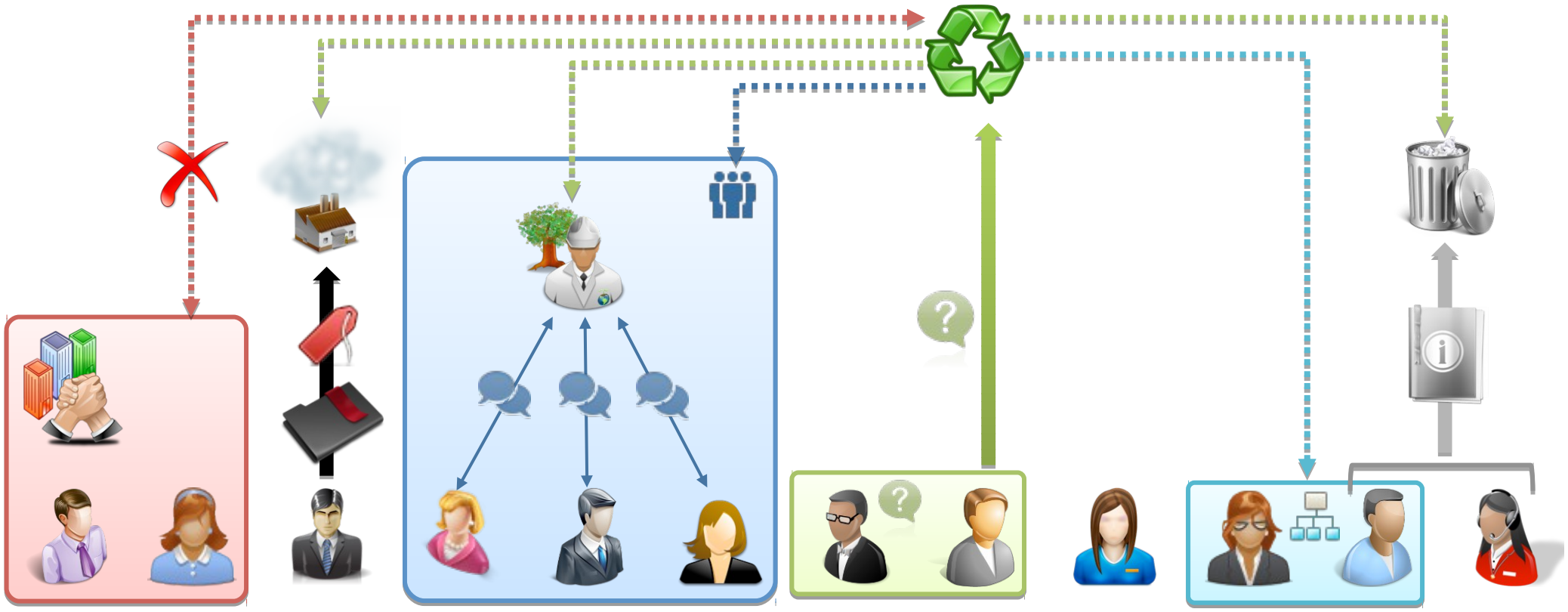


Methods and domains

- Methods
  - Evolutionary Algorithms
  - Decision Trees and Random Forests
  - Support Vector Machines
  - Multi-Agent Systems
  - Boosting
  - Neural networks
  - Galois Lattice
  - Naïve Bayes
  - …
- Domains
  - Text mining
  - Vision/Image/Robotic
  - Health
  - Hydrology
  - Biology
  - Transportation
  - Sensor Networks
  - Chemistry (Perfumes)
  - Cognition
  - …

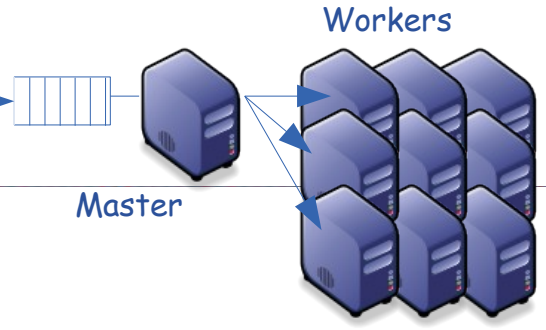
Socio-semantic networks

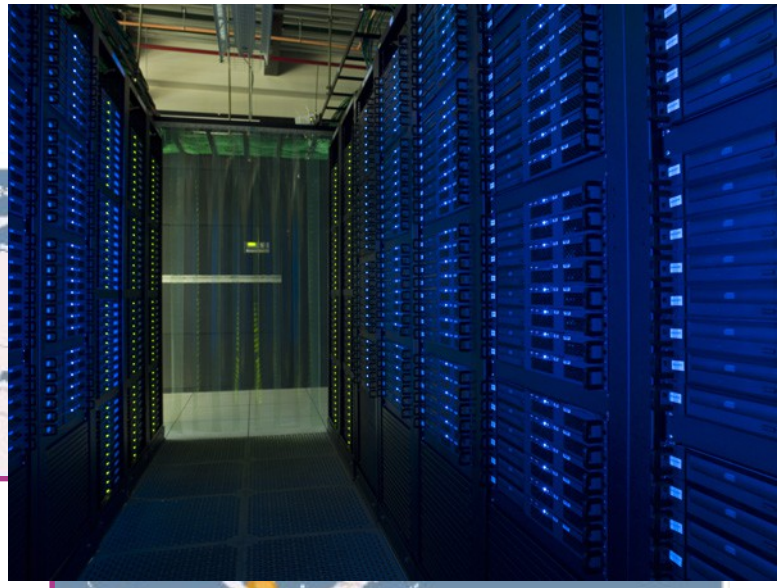
- Combining formal semantics and social semantics on t Web



Large-Scale Distributed Systems

- Distributed systems
  - Cluster Computing
  - Grid Computing
  - Cloud Computing





Size of the infrastructure today:

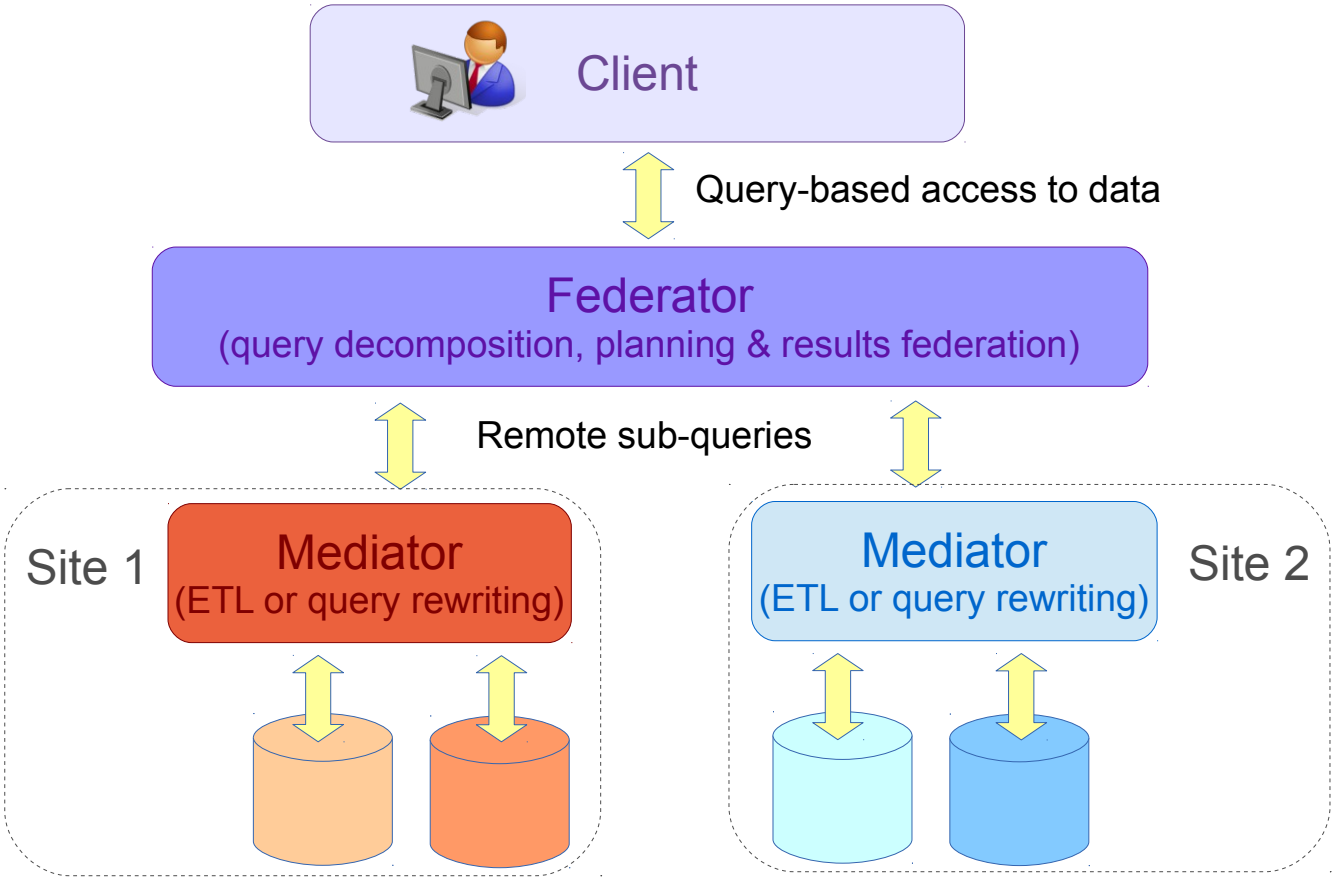
- **> 22 000 registered users**
- **> 300 sites**
- **> 330 000 CPU cores**
- **> 230 PB disk + 170 PB tape**
- **> 40 M jobs jobs/month**

Research themes

- Software architectures
  - Composition and evolution
  - Variability: software product lines
- Large-scale distributed computing
  - Performance optimization
  - Reliability
- Data-driven applications
  - Scientific workflows
  - Semantic data representation

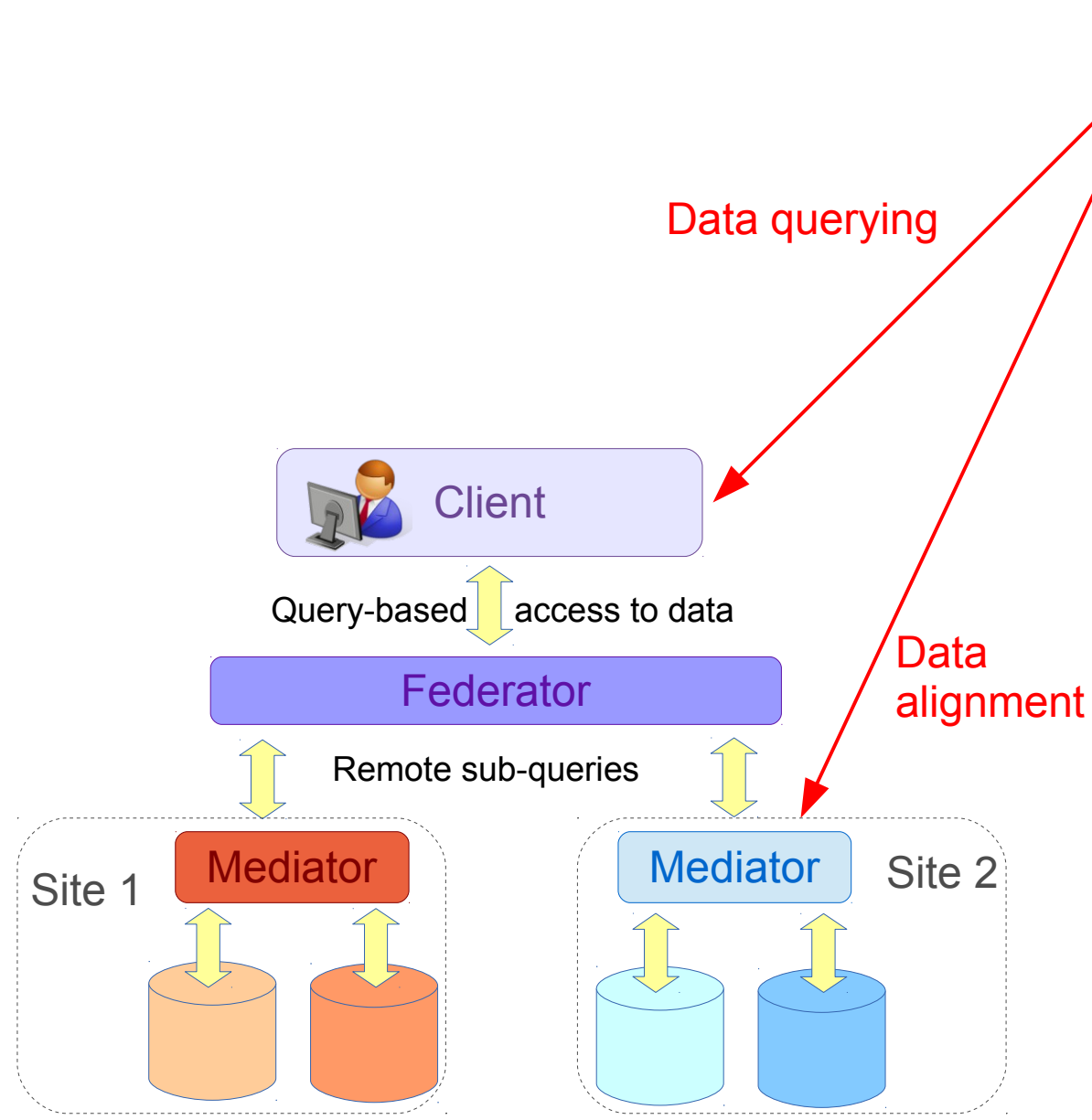
Distributed biomedical data stores

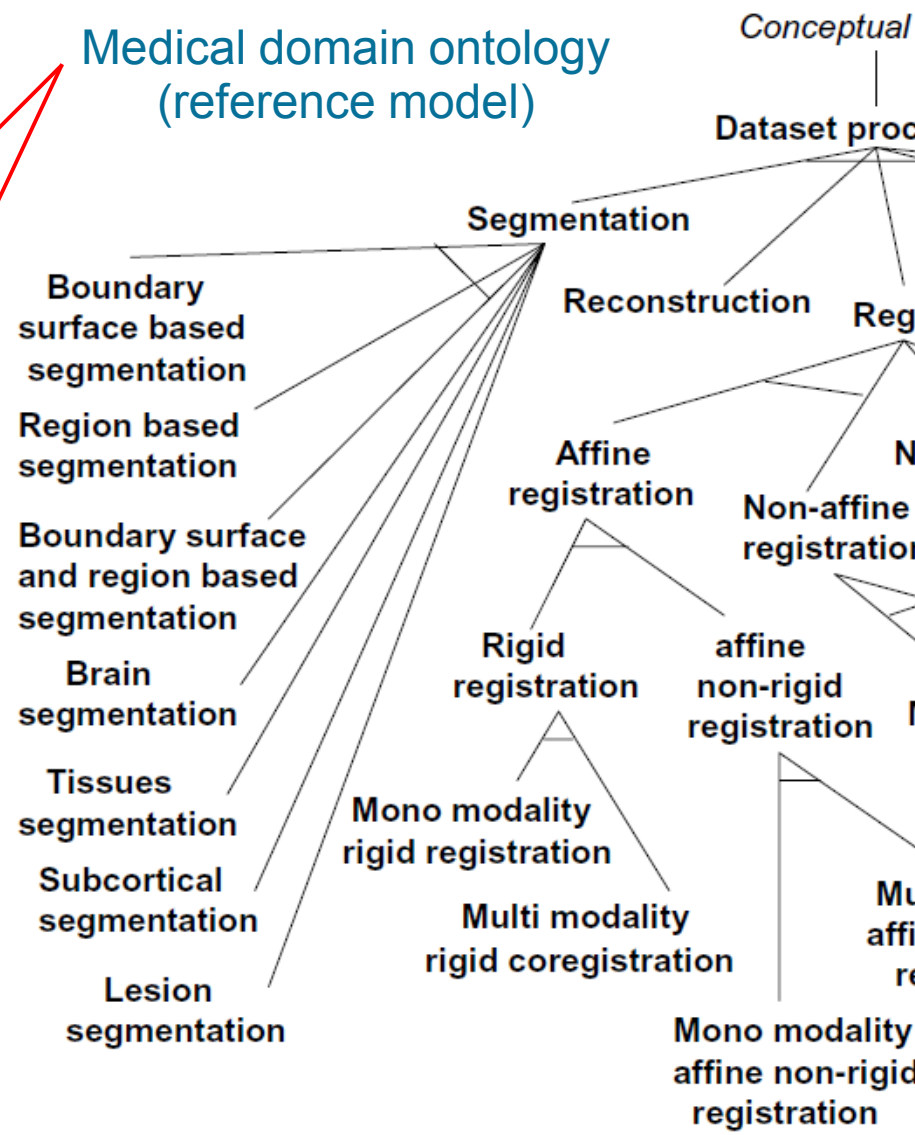
- Data federation through distributed querying



- Heterogeneous databases schema mediation

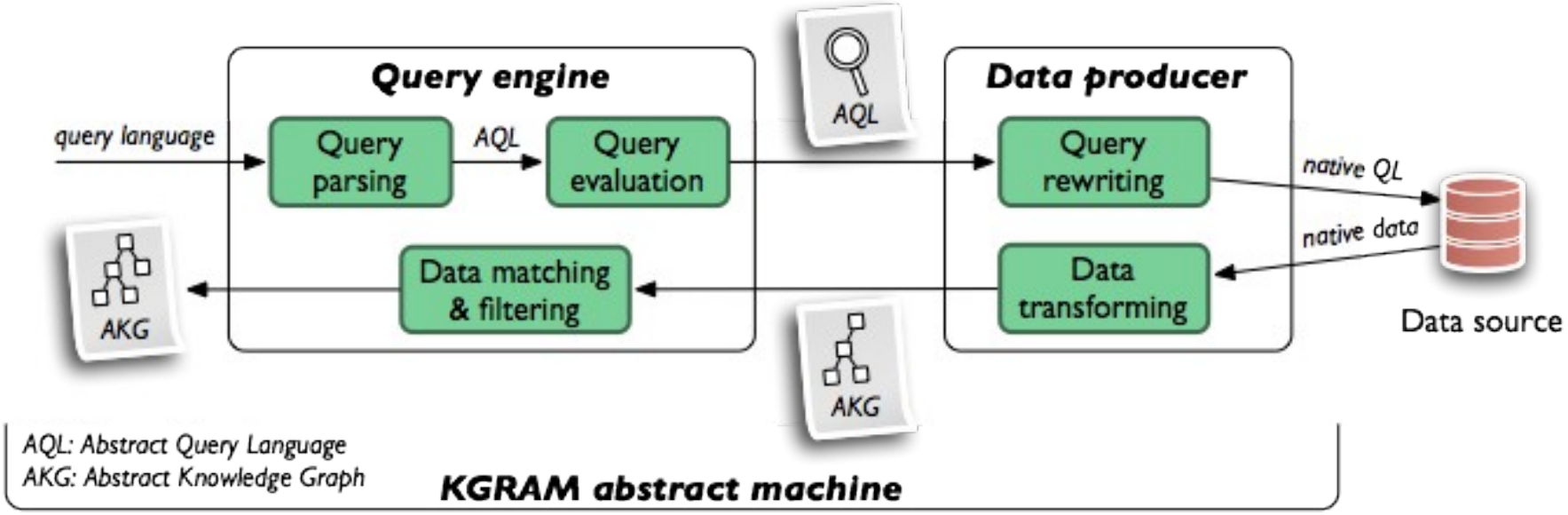
Ontology-based data model





Data query & federation engine

- KGRAM (Knowledge Graph Abstract Machine) Semant query engine



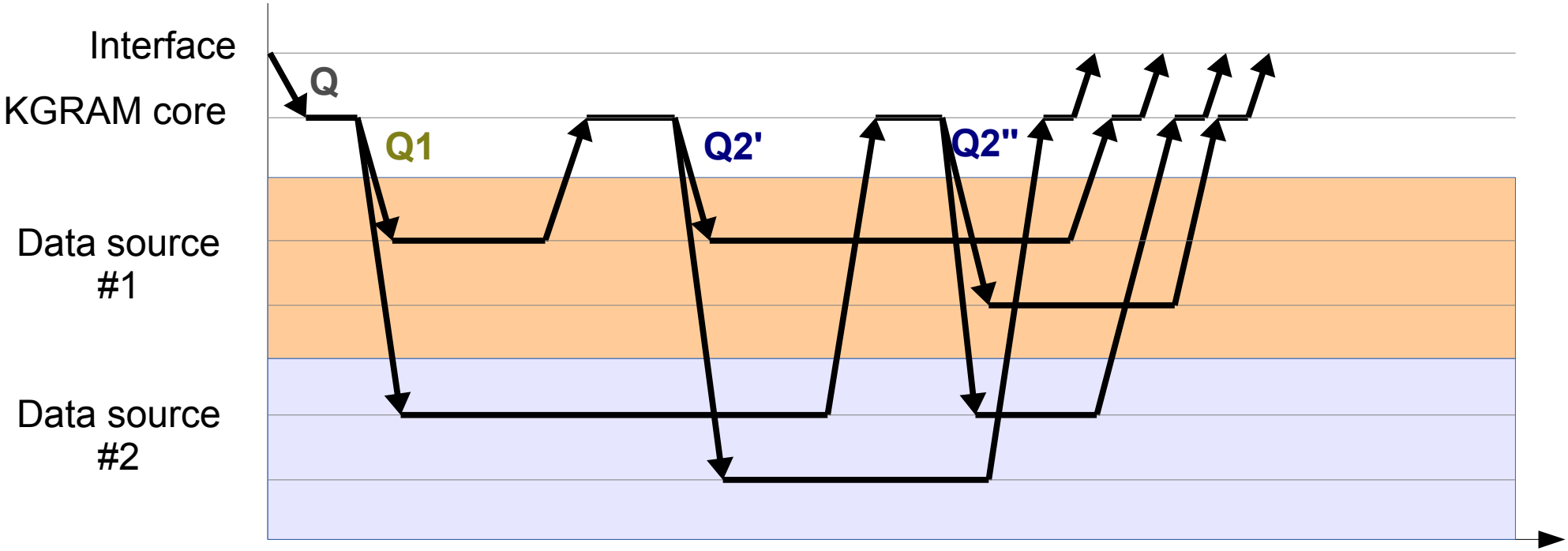
- Distributed Query Processing engine

Distributed Query Processing

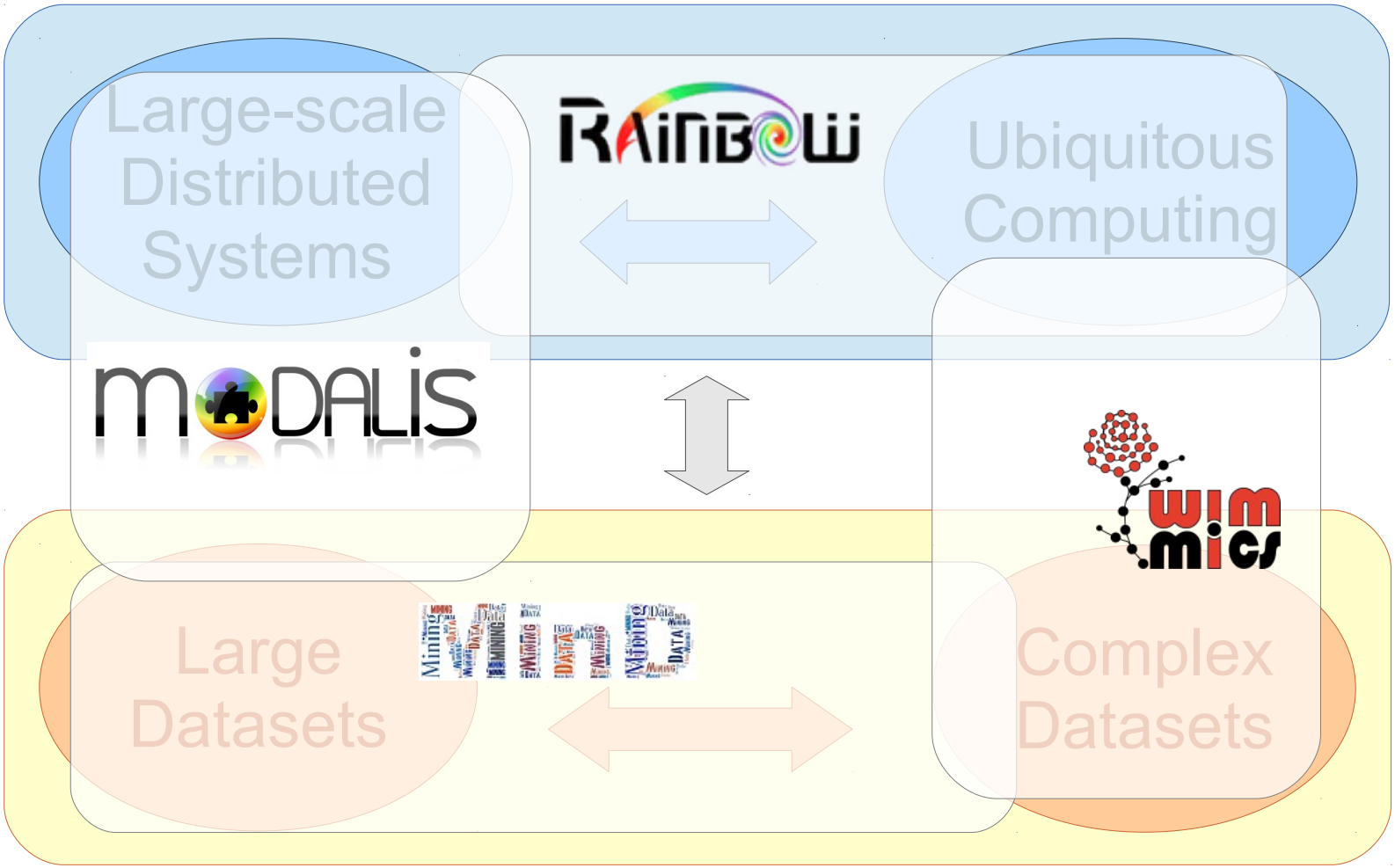
- KGRAM query processing

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

- Asynchronous execution



Software and Knowledge Engineering



Distributed Computing

Big Data

Pôle GLC = MinD + MODALIS + RAINBOW + Wimmics

GLC ecosystem

from Bachelors to PhD (College, Bachelor, Masters, Engineering school) ; Masters specializations ; ...

International scientific communications at the highest level, **standardization**, technology transfer to start-ups and industry majors

17 externally funded research projects (national research agency, government, EU and international, industry)

in-situ, in vitro and in-silico experiments, field testing ; Prototypes, Business solutions

National and international collaborations