Intensional Data Management

Pierre Senellart

20 April 2018, *BioQOP meeting*

Uncertain data is everywhere

Numerous sources of uncertain data:

- Measurement errors
- Data integration from contradicting sources
- Imprecise mappings between heterogeneous schemas
- Imprecise automatic processes (information extraction, natural language processing, etc.)
- Imperfect human judgment
- Lies, opinions, rumors

Structured data is everywhere

Data is structured, not flat:

- Variety of representation formats of data in the wild:
  - relational tables
  - trees, semi-structured documents
  - graphs, e.g., social networks or semantic graphs
  - data streams
  - complex views aggregating individual information
- Heterogeneous schemas
- Additional structural constraints: keys, inclusion dependencies

Intensional data is everywhere

Lots of data sources can be seen as intensional: accessing all the data in the source (in extension) is impossible or very costly, but it is possible to access the data through views, with some access constraints, associated with some access cost.

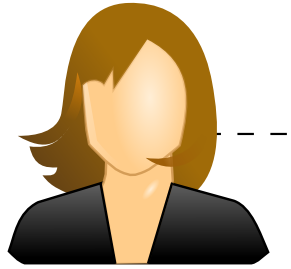
- Indexes over regular data sources
- Deep Web sources: Web forms, Web services
- The Web or social networks as partial graphs that can be expanded by crawling
- Outcome of complex automated processes: information extraction, natural language analysis, machine learning, ontology matching
- Crowd data: (very) partial views of the world
- Logical consequences of facts, costly to compute

Interactions between uncertainty, structure, intensionality

- If the data has complex structure, uncertain models should represent possible worlds over these structures (e.g., probability distributions over graph completions of a known subgraph in Web crawling).
- If the data is intensional, we can use uncertainty to represent prior distributions about what may happen if we access the data. Sometimes good enough to reach a decision without having to make the access!
- If the data is an RDF graph accessed by semantic Web services, each intensional data access will not give a single data point, but a complex subgraph.

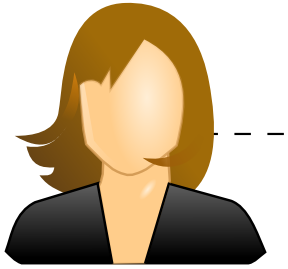
Intensional Data Management

- Jointly deal with Uncertainty, Structure, and the fact that access to data is limited and has a cost (computational, network, monetary, privacy...), to solve a user's knowledge need
- Lazy evaluation whenever possible
- Evolving probabilistic, structured view of the current knowledge of the world
- Solve at each step the problem: What is the best access to do next given my current knowledge of the world and the knowledge need
- Knowledge acquisition plan (recursive, dynamic, adaptive) that minimizes access cost, and provides probabilistic guarantees



formulation

Knowledge need

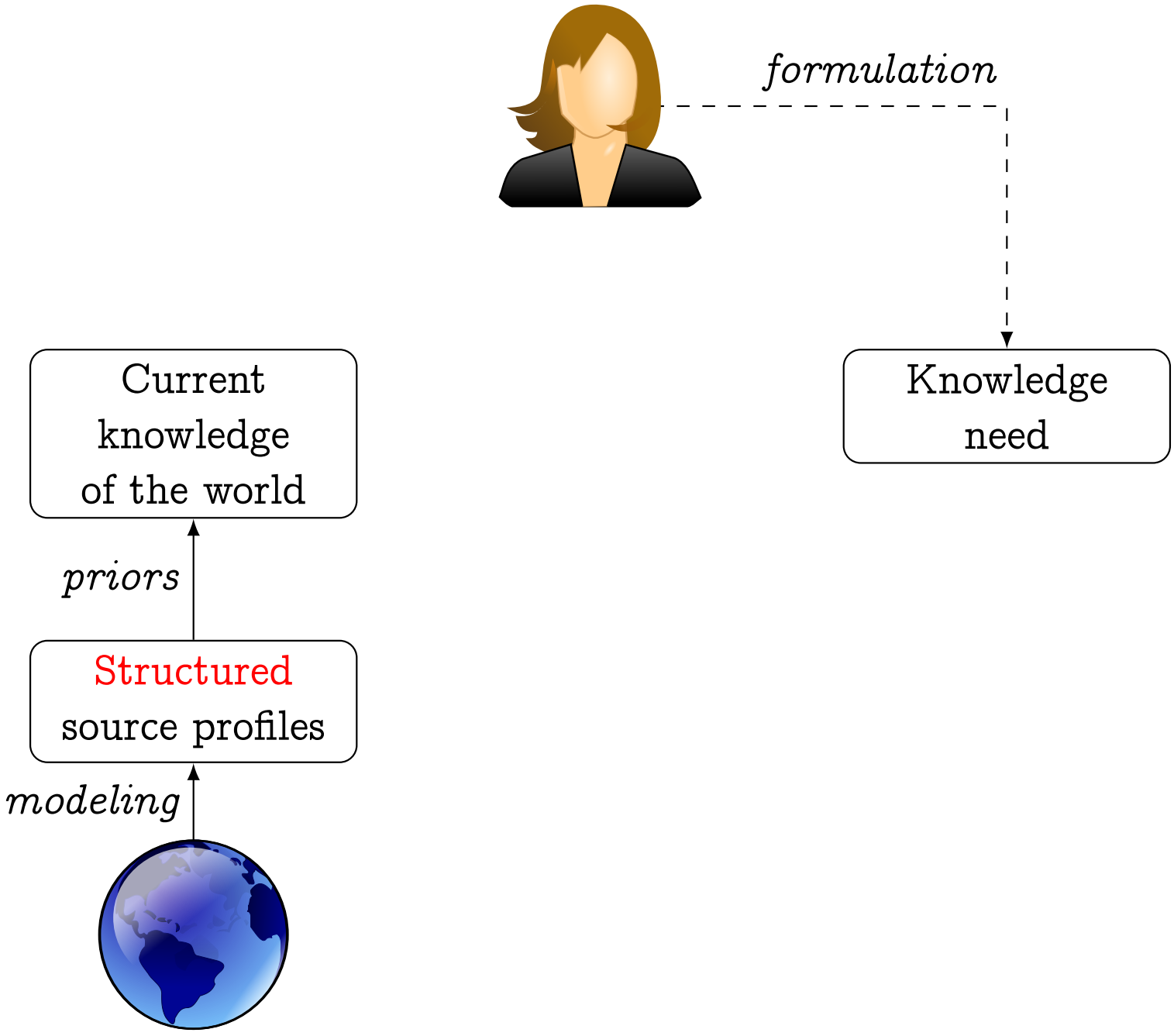


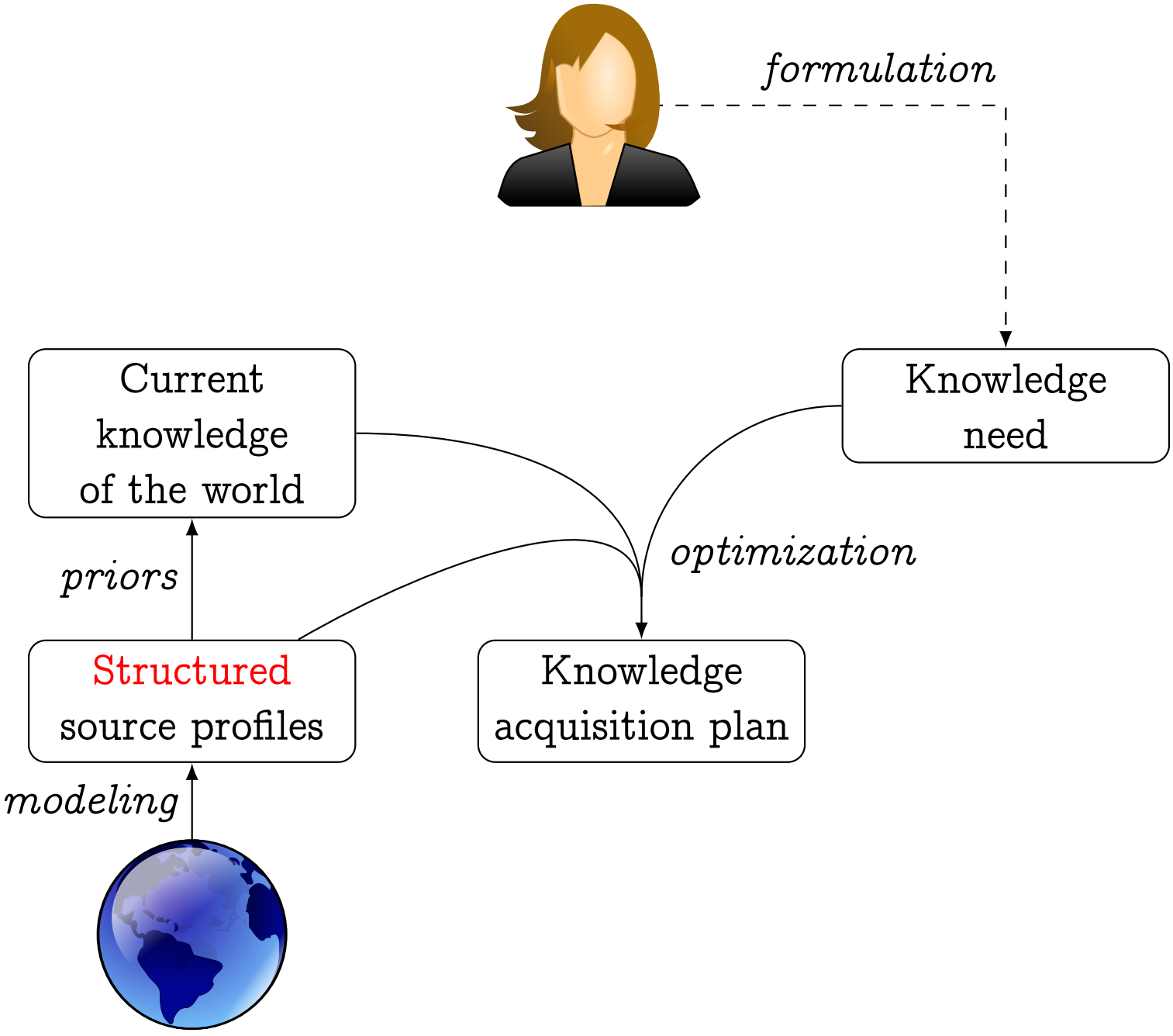
formulation

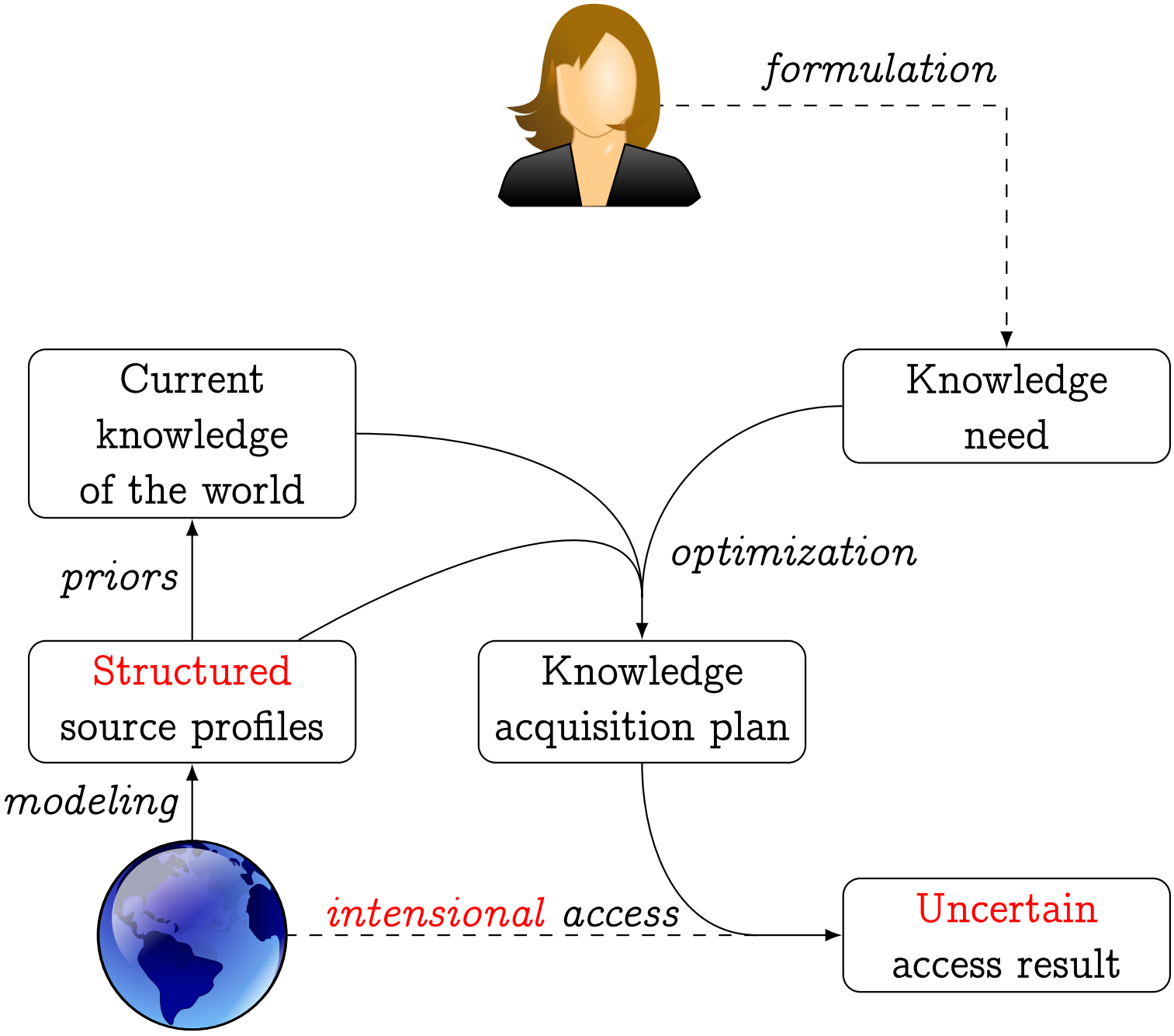
Knowledge need

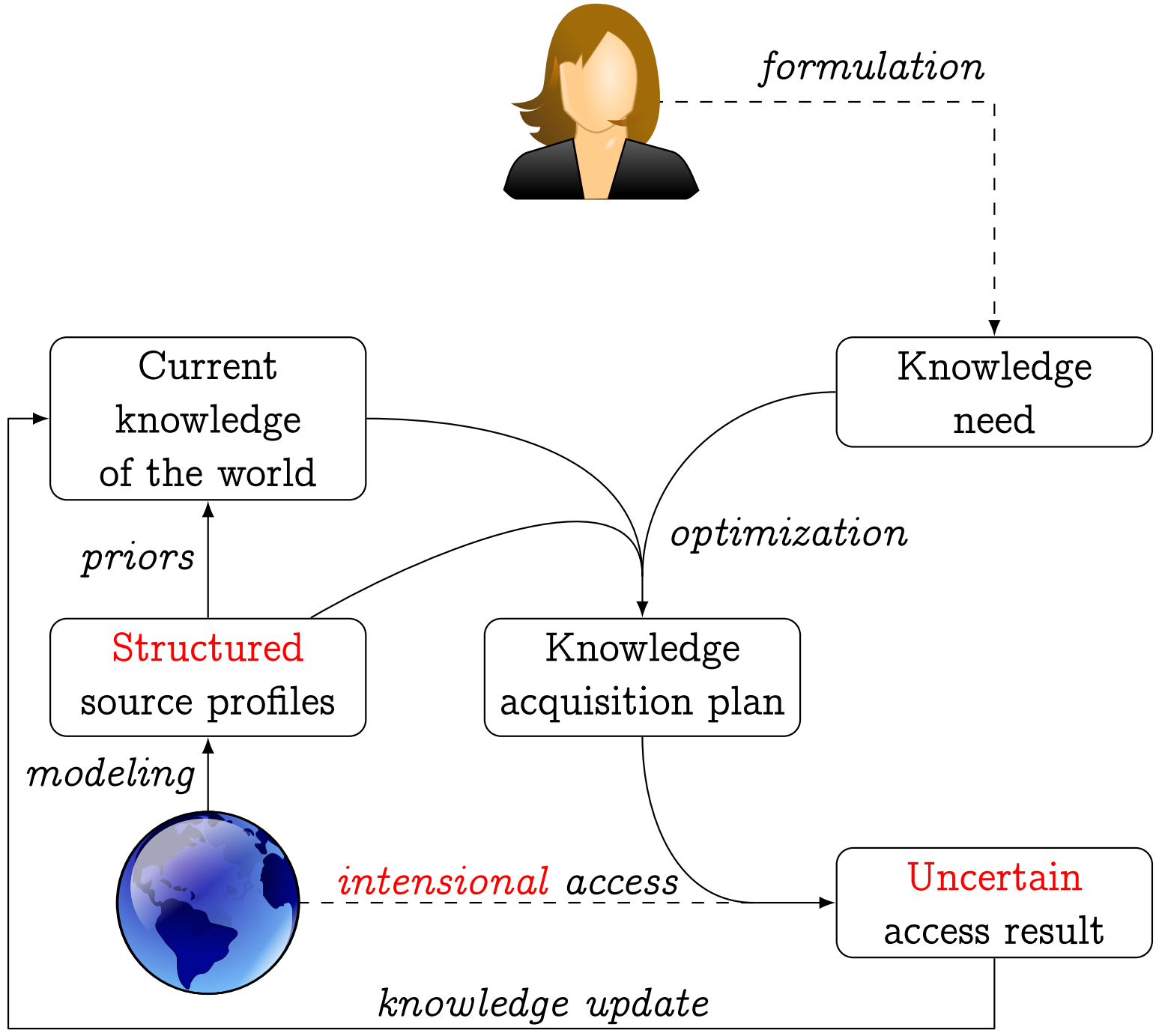
Structured source profiles

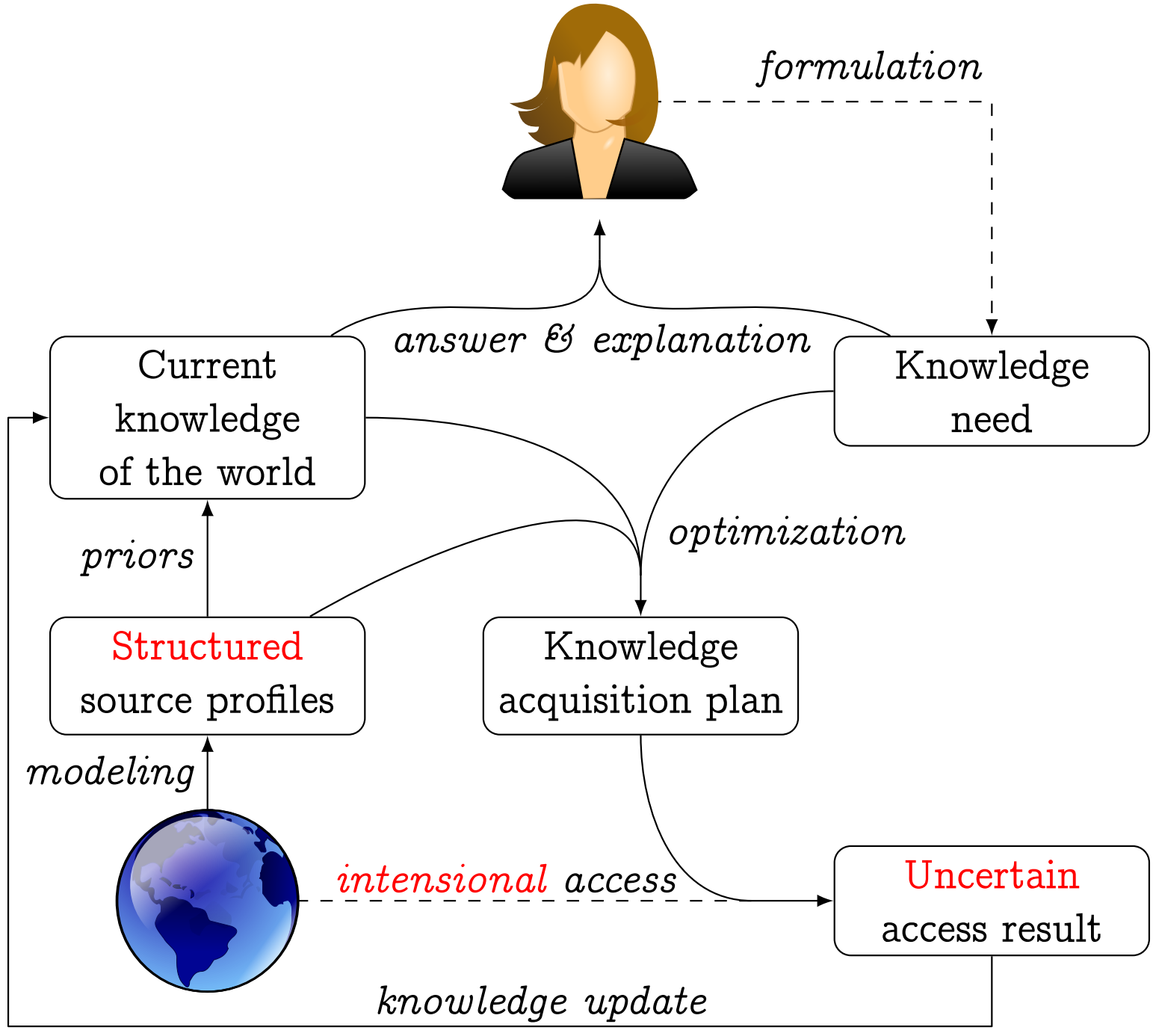
modeling


formulation
Knowledge need
Current knowledge of the world
priors
Structured source profiles
modeling


formulation
Current knowledge of the world
Knowledge need
priors
optimization
Structured source profiles
Knowledge acquisition plan
modeling


formulation
Current knowledge of the world
Knowledge need
optimization
priors
Structured source profiles
Knowledge acquisition plan
modeling
intensional access
Uncertain access result


formulation
Current knowledge of the world
Knowledge need
priors
optimization
Structured source profiles
Knowledge acquisition plan
modeling
intensional access
Uncertain access result
knowledge update


formulation
answer & explanation
Current knowledge of the world
Knowledge need
priors
optimization
Structured source profiles
Knowledge acquisition plan
modeling
intensional access
Uncertain access result
knowledge update

BioQOP: Privacy as Intensionality

- In the differential privacy framework, every query over the data has a cost: the privacy budget used by the query
- Trade-off between utility (quality, certainty of the result) and privacy cost
- Answering a complex knowledge need (query) requires combining multiple accesses, while minimizing the total privacy cost and maximizing the overall utility
- Do not necessarily neglect other forms of cost (computational cost, network cost in distributed computation)
- Do not necessarily neglect other forms of uncertainty (uncertainty in the input data, quality of classification algorithms)