Complex Data Managament

Pierre Senellart

18 June 2018, *AI Singapore – France*

DI ENS

- DI ENS: Computer science department of École normale supérieure, Paris
- Research lab joint between ENS (PSL Univ.), CNRS, Inria
- ENS: Elite school for students interested basic scientific research and academic careers
- Members:
  - $\approx$ 30 faculty
  - $\approx$ 70 undergrad/MSc students (4 years of training, 15–20 per year), highly competitive selection after a 2-year preparation
  - $\approx$ 70 PhD students, 20 post-docs
- Research areas:
  - Security and cryptography
  - Programming languages, software verification
  - Theoretical computer science, mathematical models
  - AI and data science

AI at DI ENS

- 4 research teams specializing in Artificial Intelligence and Data Science:

  Data Mathematical models of high-dimensional complex data got classification (Stéphane Mallat)

  Sierra: Optmization and machine learning (Francis Bach, Alexandre d'Aspremont, Pierre Gaillard)

  Willow: 3D computer vision (Ivan Laptev, Jean Ponce, Josef Sivic)

  Valda: Complex data management (Pierre Senellart, Serge Abiteboul, Camille Bourgaux, Olivier Cappé, Luc Segoufin, Michaël Thomazo)

AI at DI ENS

- 4 research teams specializing in Artificial Intelligence and Data Science:

  Data: Mathematical models of high-dimensional complex data got classification (Stéphane Mallat)

  Sierra: Optmization and machine learning (Francis Bach, Alexandre d'Aspremont, Pierre Gaillard)

  Willow: 3D computer vision (Ivan Laptev, Jean Ponce, Josef Sivic)

  Valda: Complex data management (Pierre Senellart, Serge Abiteboul, Camille Bourgaux, Olivier Cappé, Luc Segoufin, Michaël Thomazo)

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

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

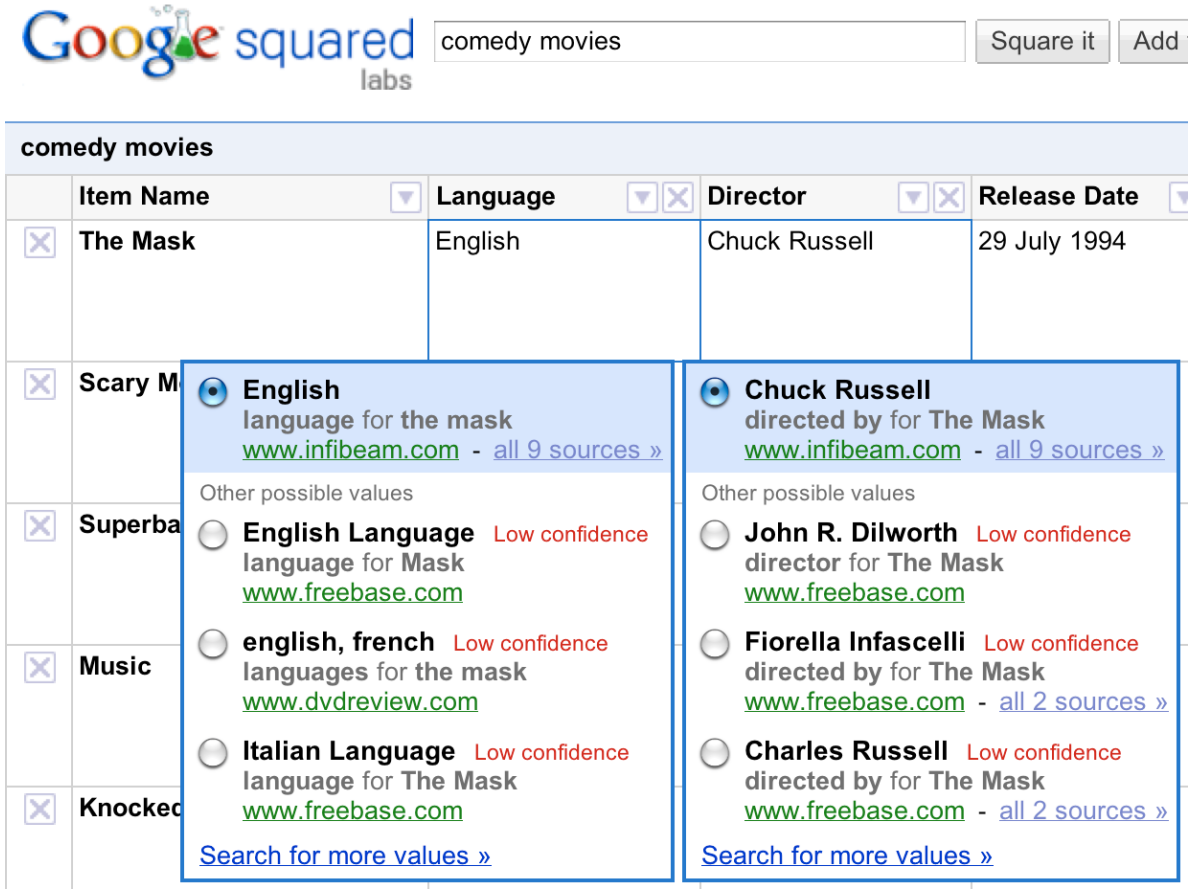
Uncertainty in Web information extraction

instance	iteration	date learned	confidence
arabic, egypt	406	08-sep-2011	(Seed) 100.0
chinese, republic_of_china	439	24-oct-2011	100.0
chinese, singapore	421	21-sep-2011	(Seed) 100.0
english, britain	439	24-oct-2011	100.0
english, canada	439	24-oct-2011	(Seed) 100.0
english, england001	439	24-oct-2011	100.0
arabic, morocco	422	23-sep-2011	100.0
cantonese, hong_kong	406	08-sep-2011	100.0
english, uk	436	19-oct-2011	100.0
english, south_vietnam	427	27-sep-2011	99.9
french, morocco	422	23-sep-2011	99.9
greek, turkey	430	07-oct-2011	99.9

Never-ending Language Learning (NELL, CMU),

`http://rtw.ml.cmu.edu/rtw/kbbrowser/`

Uncertainty in Web information extraction



Google Squared (terminated),
screenshot from [Fink et al., 2011]

Uncertainty in Web information extraction

Subject	Predicate	Object	Confidence
Elvis Presley	diedOnDate	1977-08-16	97.91%
Elvis Presley	isMarriedTo	Priscilla Presley	97.29%
Elvis Presley	influences	Carlo Wolff	96.25%

YAGO, `http://www.mpi-inf.mpg.de/yago-naga/yago`

[Suchanek et al., 2007]

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.

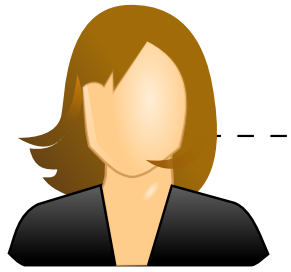
- Indexes over regular data sources
- Deep Web sources: Web forms, Web services
- The Web or social networks as partial graphs that can be expanded by crawling
- Outcome of complex automated processes: information extraction, natural language analysis, machine learning, ontology matching
- Crowd data: (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.

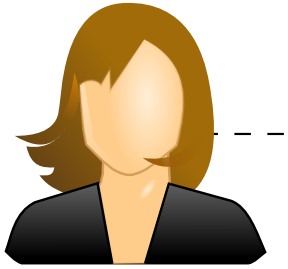
Complex Data Management

- Jointly deal with Uncertainty, Structure, and the fact that access to data is limited and has a cost, 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

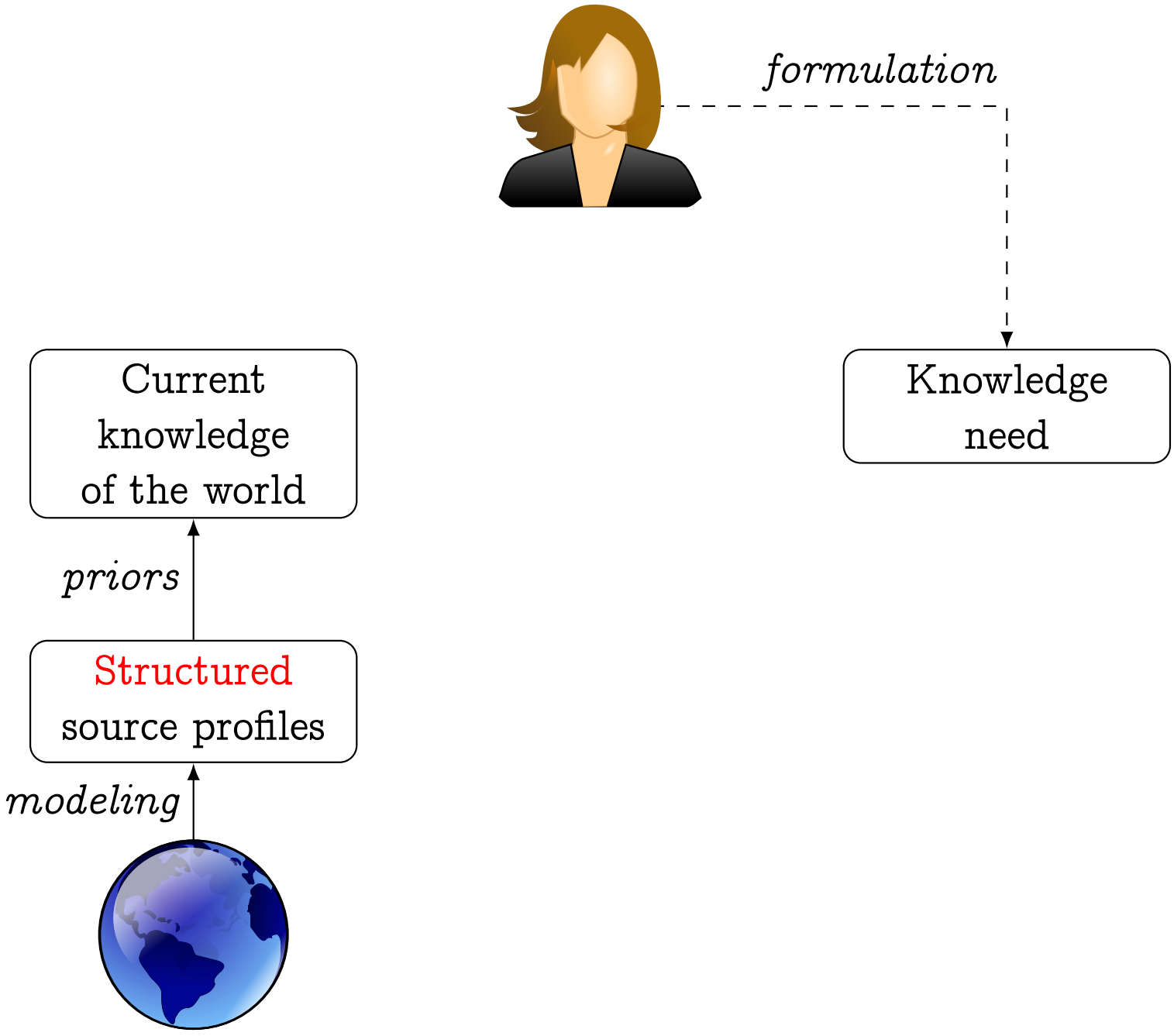


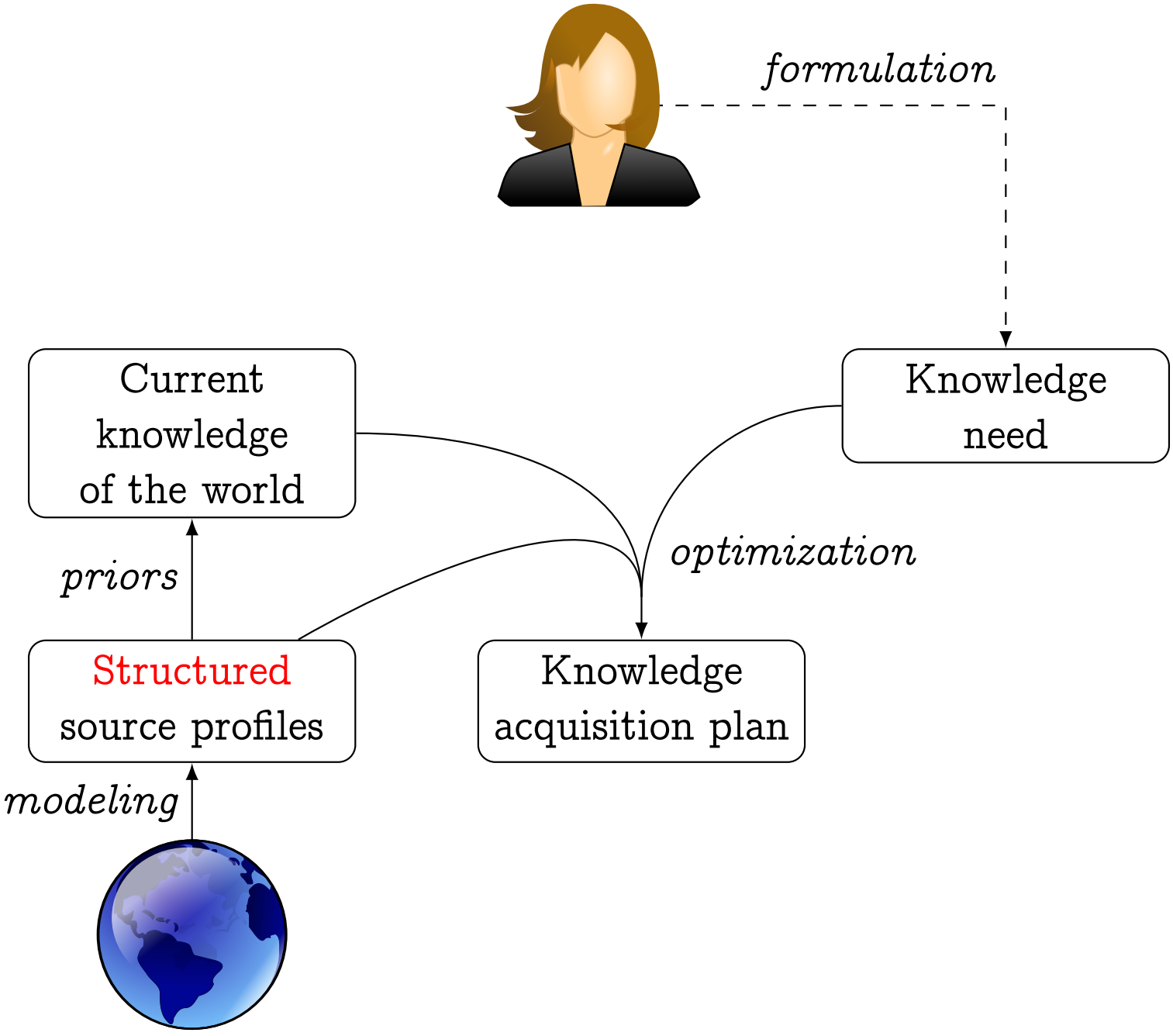
formulation

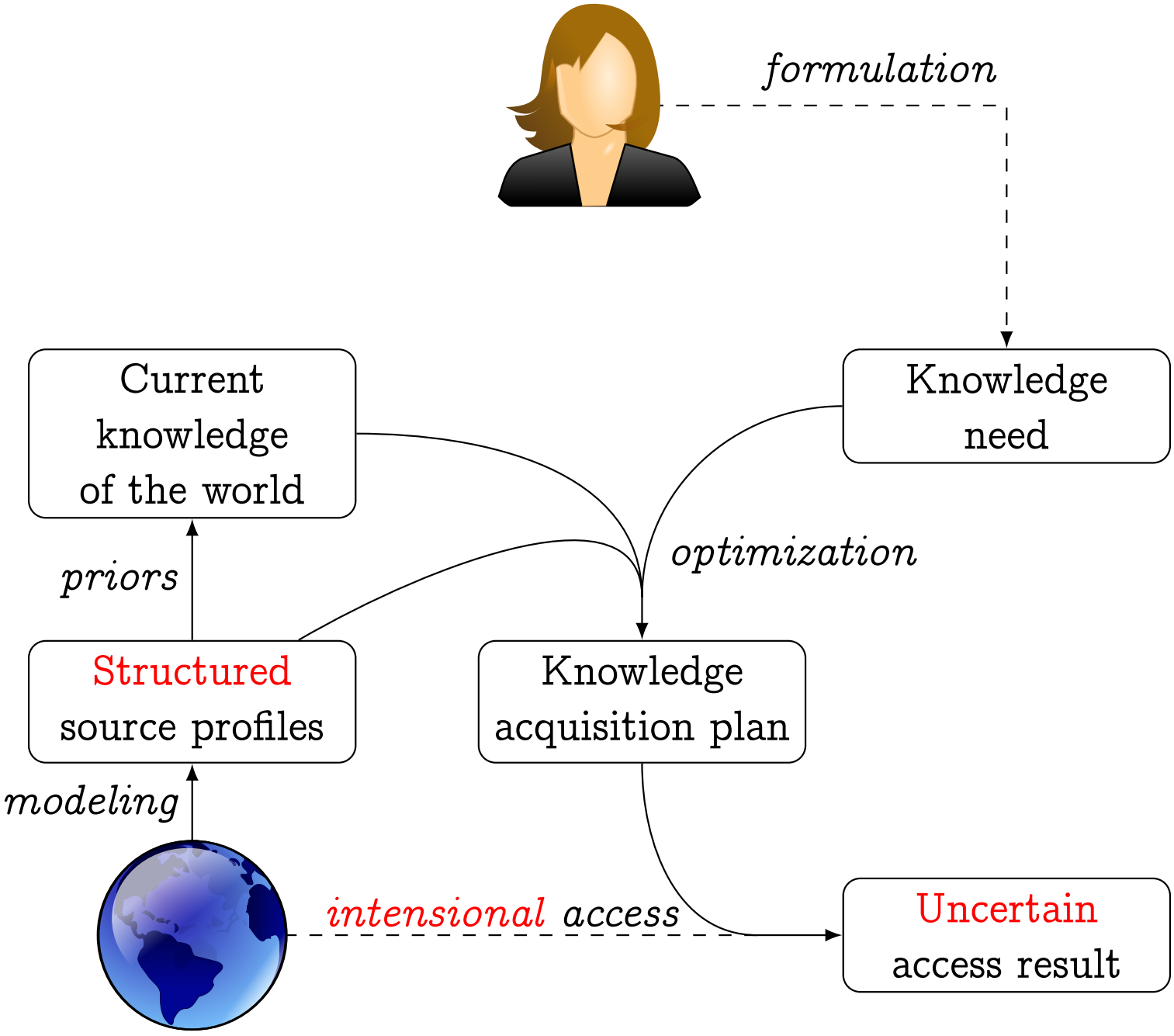
Knowledge need

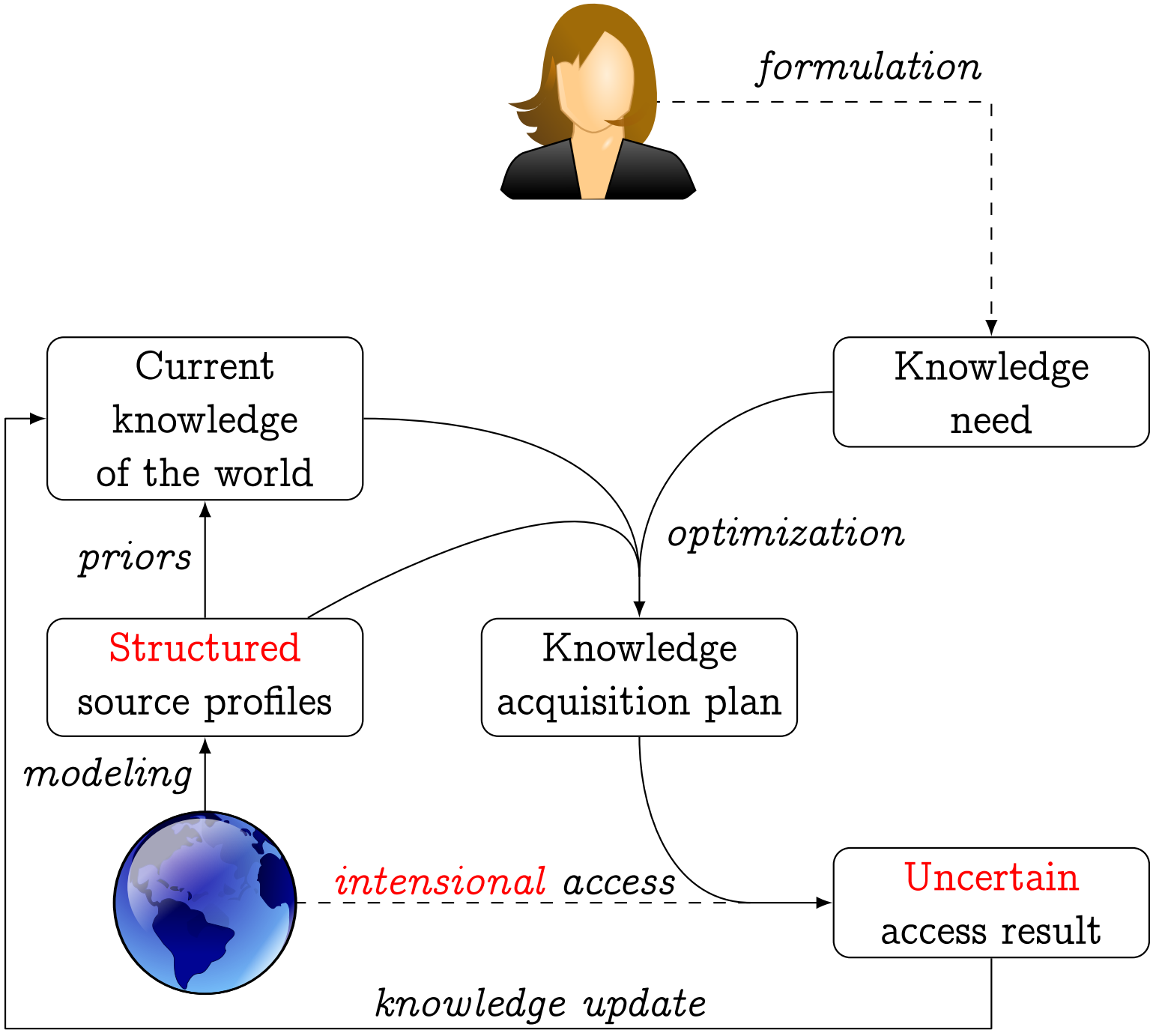
Structured source profiles

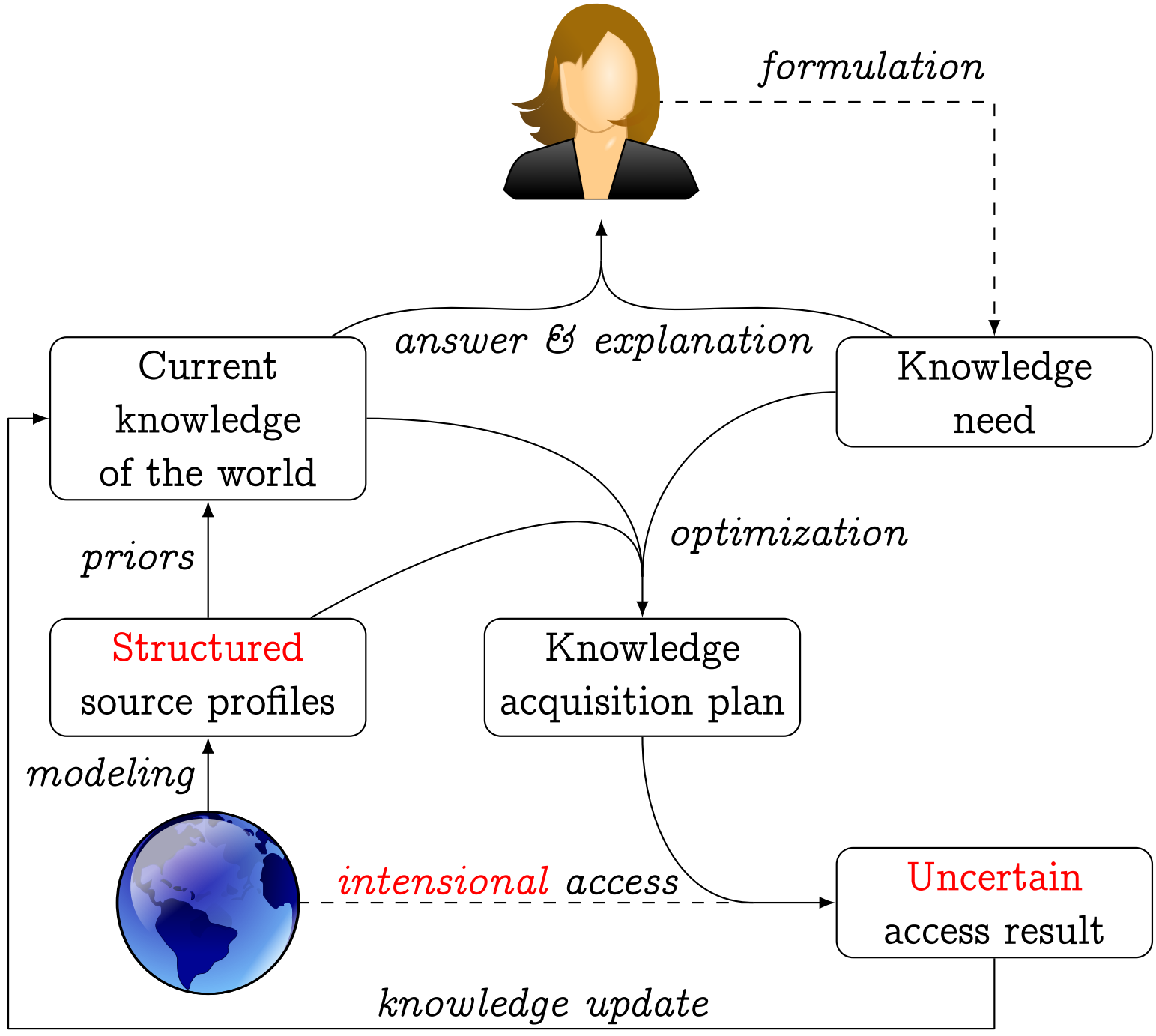
modeling


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

Valda Research on Complex Data Management

- Structured data management: extraction, integration, cleaning of real-world data: Web data (semi-structured data, text, social networks...), personal data (GPS trajectories, calendar data, emails...)
- Uncertain data management: models for incomplete and probabilistic data, algorithms and tools for tracking the provenance of data, assessing the veracity of data
- Intensional data management: reinforcement learning (MDP, bandits), crowdsourcing, data-intensive workflows
- Unique positioning of research:

  Theory: database theory, symbolic AI, knowledge representation, algorithms, bounds

  Database systems: Implementation of practical database systems (ProvSQL, Thymeflow...)

  Real-world data: Mining, analysis, learning from real-world data

Merci.

Bibliography I

Robert Fink, Andrew Hogue, Dan Olteanu, and Swaroop Rath. SPROUT^2: a squared query engine for uncertain web data. In *SIGMOD*, 2011.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A core of semantic knowledge. Unifying WordNet and Wikipedia. In *WWW*, pages 697–706, 2007. ISBN 978-1-59593-654-7.