Research within DBWeb Focus on Uncertainty, Structure, Intensionality

Pierre Senellart

21 January 2015, *U.S. Army Research Laboratory Visit at Télécom ParisTech*

The DBWeb team

- Research team within the INFRES (Computer Science and Networking) department of Télécom ParisTech
- The team:
  - 5 faculty members
  - 1 visiting researcher
  - 2 post-doctoral researchers
  - 8 PhD candidates
  - 1 engineer
- At the confluence of three areas of research:
  - Data management (systems and theory)
  - Web mining and social network analysis
  - Cognitive approach to artificial intelligence
- Web site: `http://dbweb.enst.fr/`

Senior Researchers (1/2)

Talel Abdessalem, Professor

Database systems, XML query processing, trust management, social networks, data evolution, recommender systems...

Jean-Louis Dessalles, Associate Professor

Artificial intelligence, cognitive modeling, language evolution, emergent phenomena, simplicity theory, temporal aspects of language...

Pierre Senellart, Professor

Database systems and theory, probabilistic databases, deep Web querying, graph mining, Web crawling, crowd data sourcing, dynamic data management...

Senior Researchers (2/2)

Mauro Sozio, Associate Professor
Distributed algorithms, social networks, peer-to-peer systems, approximation algorithms, MapReduce...

Fabian Suchanek, Associate Professor
Web information extraction, Semantic Web, ontologies, rule mining, culturonomics...

DBWeb and the Five Vs of Big Data

DBWeb and the Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

DBWeb and the Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

DBWeb and the Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

Velocity: Data produced or changing at high speed (LHC: 100 millions of collision per second), sometimes more than one is able to store

DBWeb and the Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

Velocity: Data produced or changing at high speed (LHC: 100 millions of collision per second), sometimes more than one is able to store

Veracity: Data quality very diverse; imprecise, imperfect, untrustworthy information

DBWeb and the Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

Velocity: Data produced or changing at high speed (LHC: 100 millions of collision per second), sometimes more than one is able to store

Veracity: Data quality very diverse; imprecise, imperfect, untrustworthy information

Value: Making sense of potentially very valuable data, but with a value not immediately apparent

DBWeb and the Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

Velocity: Data produced or changing at high speed (LHC: 100 millions of collision per second), sometimes more than one is able to store

Veracity: Data quality very diverse; imprecise, imperfect, untrustworthy information

Value: Making sense of potentially very valuable data, but with a value not immediately apparent

Special focus within DBWeb: Web Data (Web pages, social networks, e-commerce data, Semantic Web, Linked Open Data, etc.)

DBWeb and Volume, 2013–2015

- How to mine subgraphs of large graphs? [WSDM'2015]
- How to match subgraphs of large graphs? [VLDB'2013]

DBWeb and Variety, 2013–2015

- How to merge multilingual Wikipedias into a single semantic knowledge base? [CIDR'2015]
- How to generically model aspect in natural language? [CogSci'2014]
- How to filter irrelevant tags in tag recommender systems? [RecSys'2014]?
- How to extract information from Web data to better orchestrate Web services? [ICDE'2013]
- How to efficiently crawl redundant Web sites with varying templates [ICWE'2013]?
- How to discover semantic relations from text? [SIGMOD Record'2013]

DBWeb and Velocity, 2013–2015

- How to efficiently crawl social networks under severe rate policy limitations? [HT'2014 Best Paper Award]

DBWeb and Veracity, 2013–2015

- How to generate probabilistic models of document corpora? [ToCS'2014]
- How to manage uncertainty in version control systems [DocEng'2013]
- How to optimize queries over uncertain data? [ICDE'2013]
- How to capture uncertainty in crowdsourced data? [SIGMOD'2013]

DBWeb and Value, 2013–2015

- How to efficiently mine patterns in big knowledge bases? [WWW'2013 Best Student Paper Award; WWW'2015]
- How to canonicize existing knowledge bases? [CIKM'2014]
- How to price the value of incomplete data? [DEXA'2014]
- How to explain human investment in social networks? [Evolution'2014]
- How to explain unexpectedness in fiction? [Lit.&Ling. Comp.'2014]
- How to mine cultural information from the Web? [VLDB'2014]
- How to harvest common sense knowledge from the Web? [WSDM'2014]

Big Data Teaching in DBWeb

- Data Science track for the M. Eng. at Télécom ParisTech (first year of graduate studies)
- Data & Knowledge track of the Computer Science M. Sc. at Université Paris–Saclay (second year of graduate studies, from September 2015 on)
- Big Data Advanced Master (post-graduate diploma for young professionals)
- Data Scientist Certificate of Advanced Studies (continuing education program)

Plan

DBWeb Research

Focus: Uncertainty, Structure, Intensionality

Instances of UnSAID

Conclusion

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

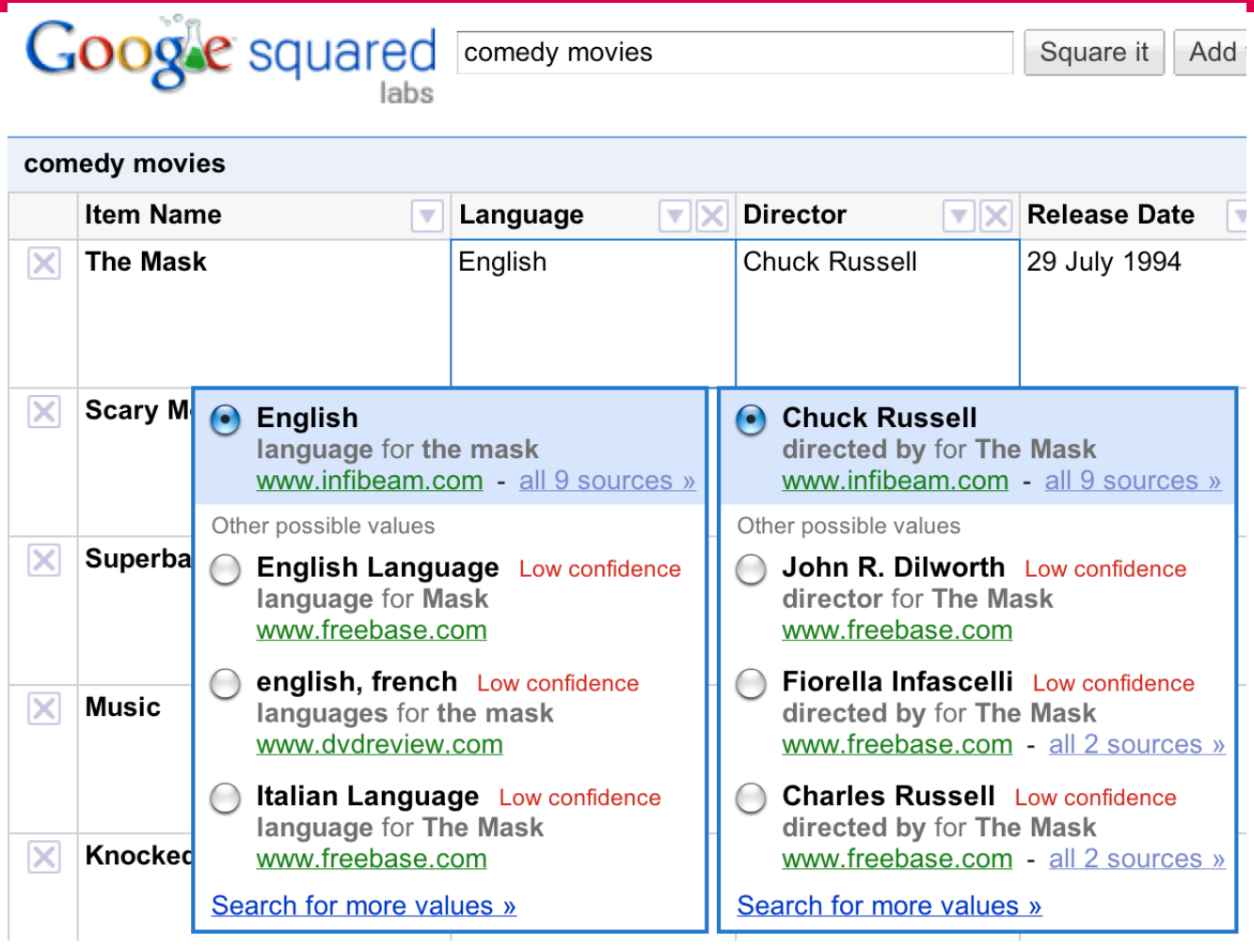
Uncertainty in Web information extraction

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, Hogue, Olteanu, and Rath 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, Kasneci, and Weikum 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 a RDF graph accessed by semantic Web services, each intensional data access will not give a single data point, but a complex subgraph.

State of the art and opportunities

Probabilistic databases cover limited structure variations, do not consider intensionality

Active and reinforcement learning deals with uncertainty and intensionality, but assumes trivial structures and simple goals

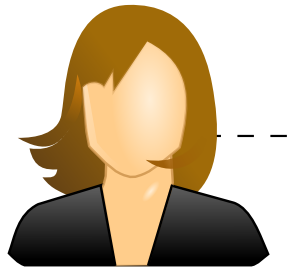
Crowdsourcing, focused crawling, deep Web crawling focus on specific applications of the uncertainty/structure/intensionality problem

Answering queries using views assumes simplistic cost models

Opportunities for Web data management systems that take all dimensions into account

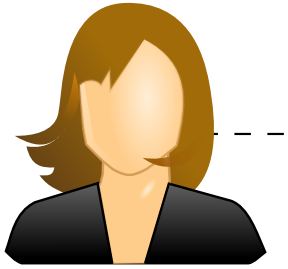
Introducing UnSAID

- Uncertainty and Structure in the Access to Intensional Data
- 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

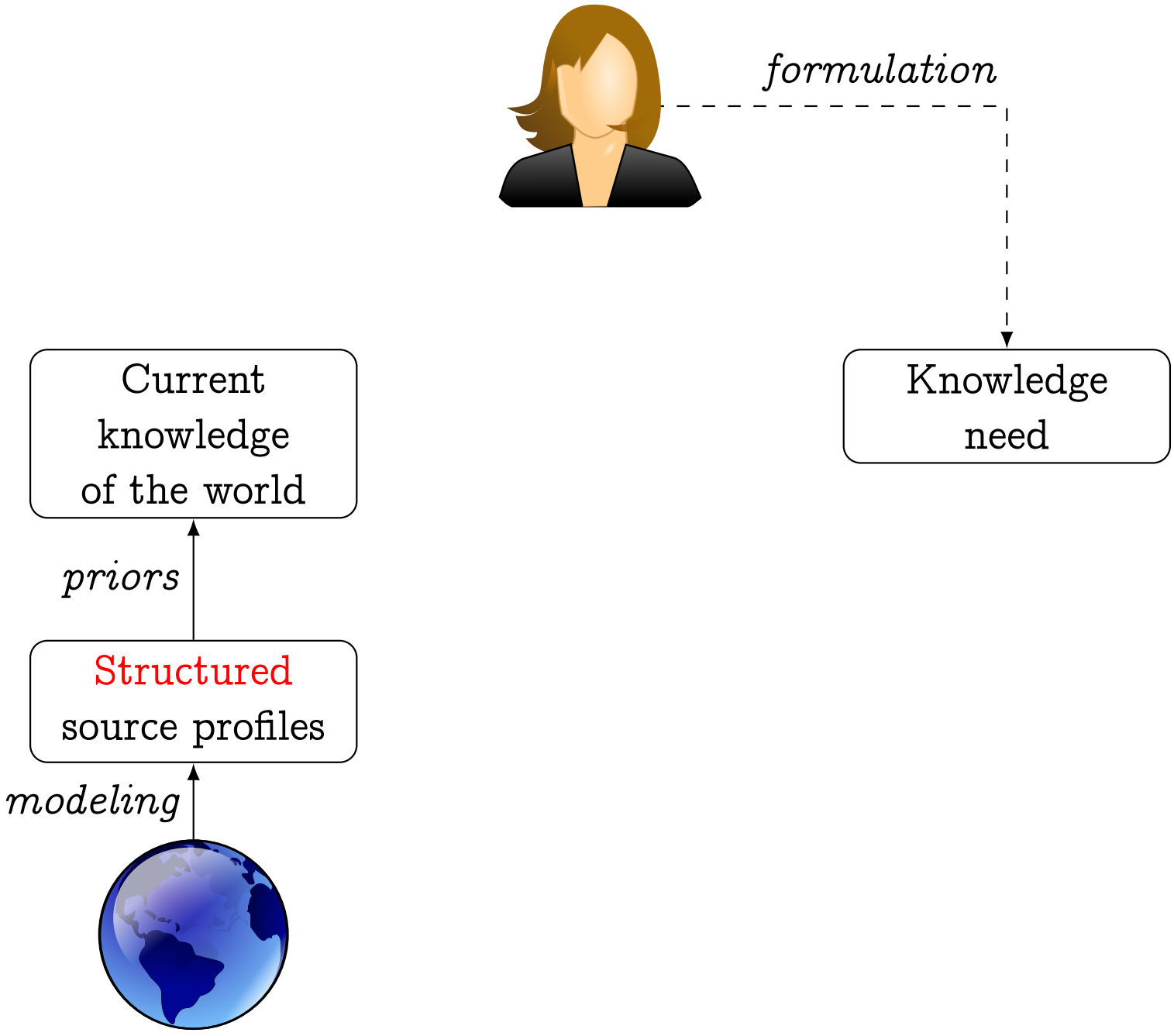


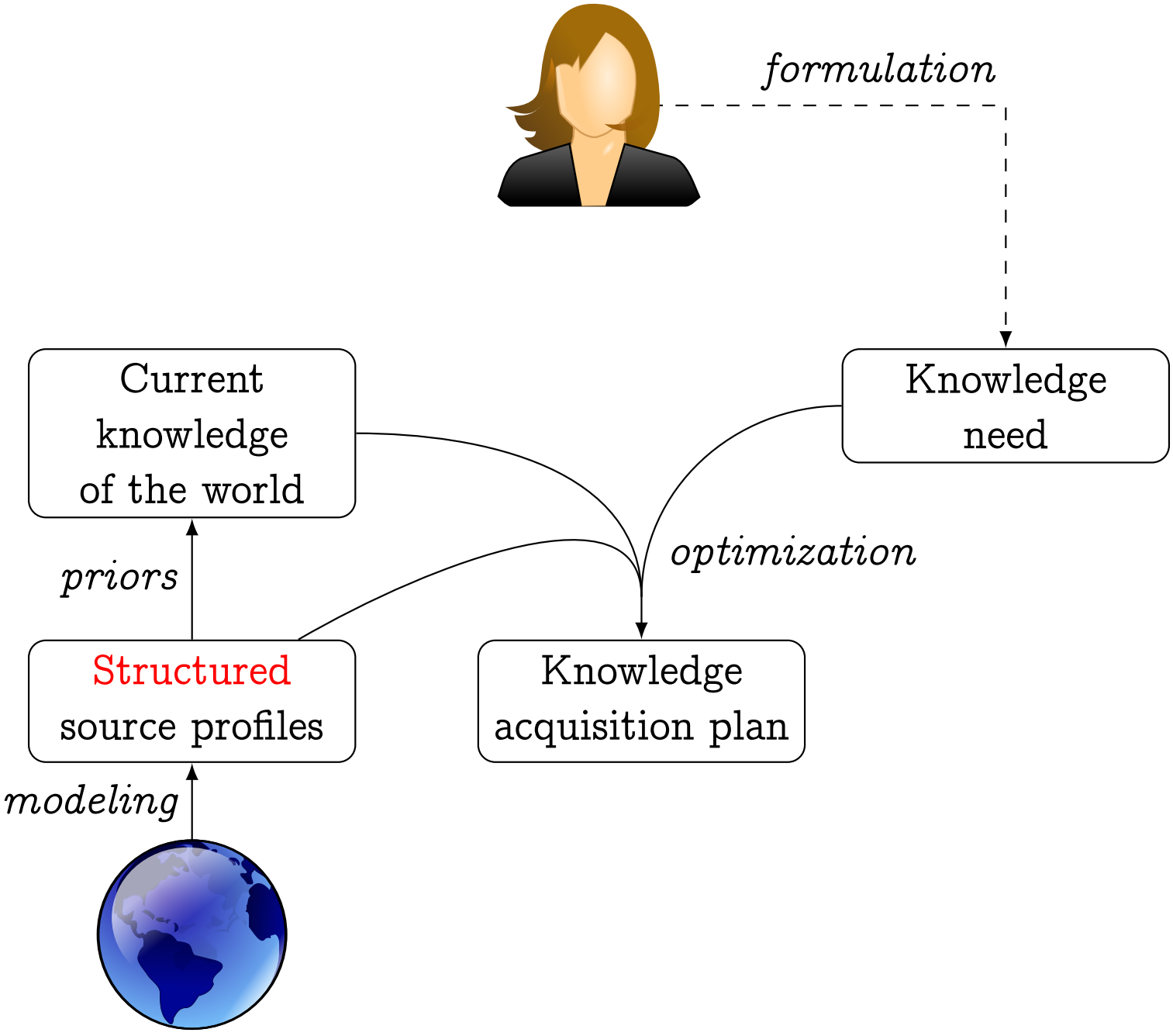
formulation

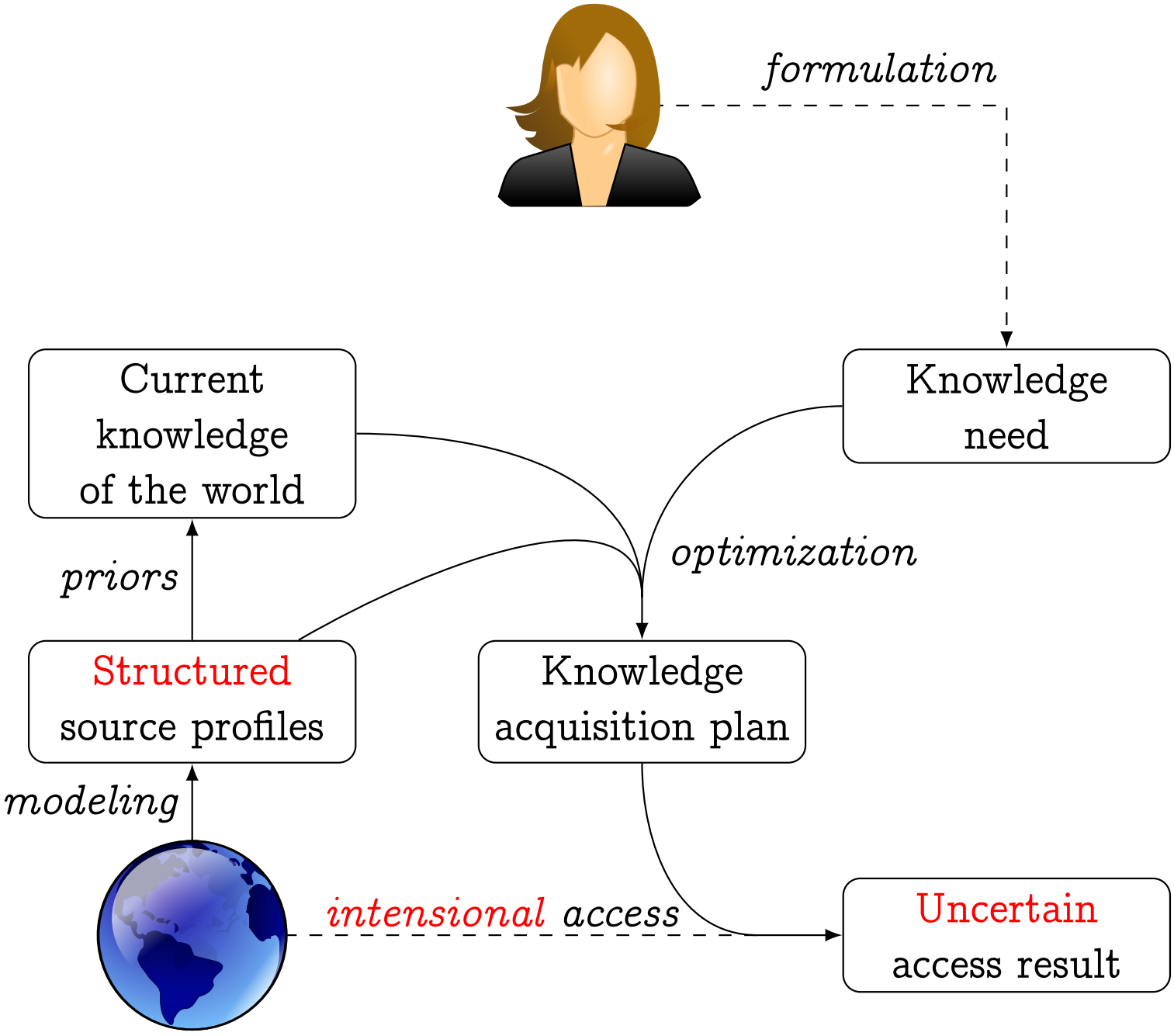
Knowledge need

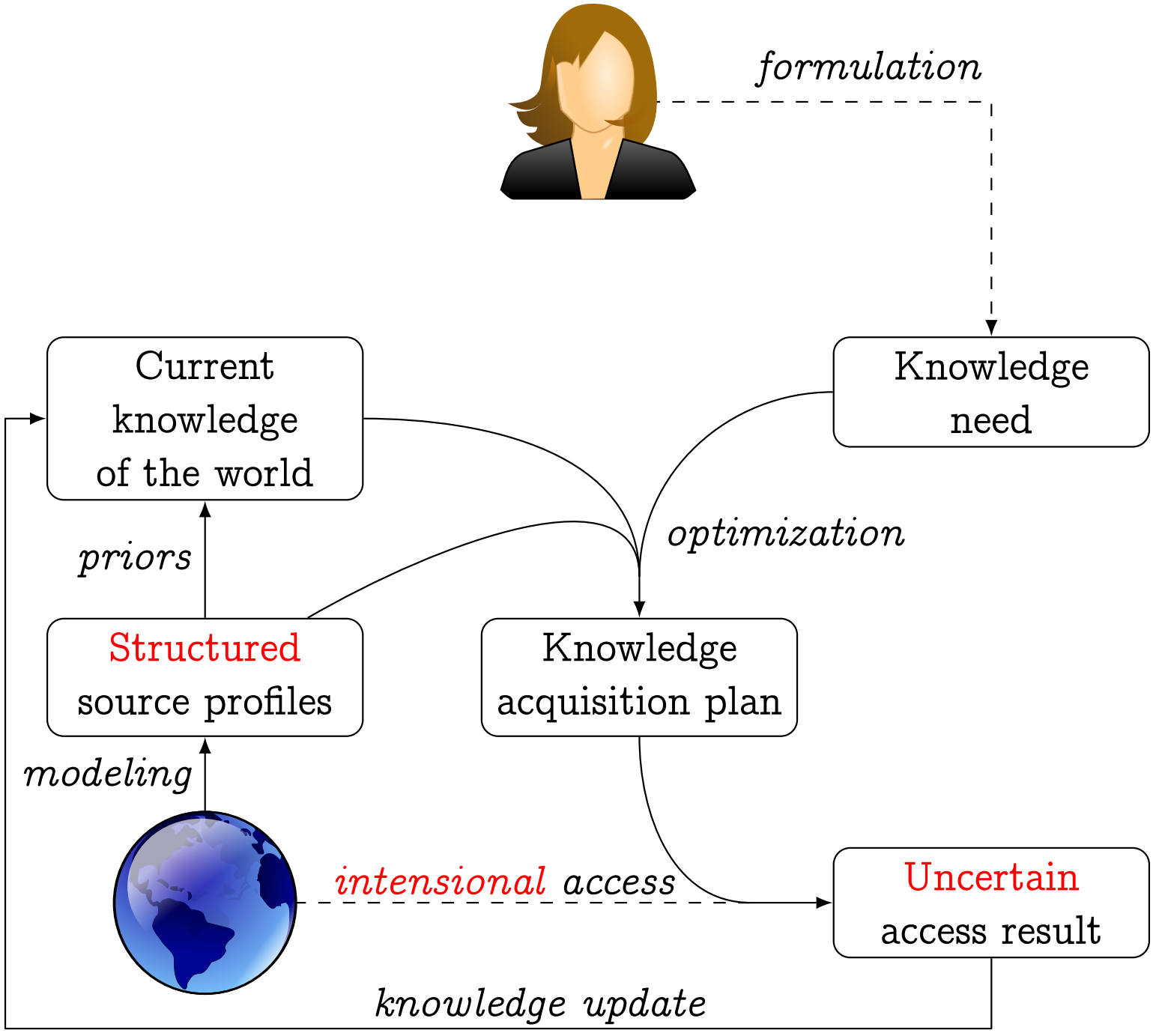
Structured source profiles

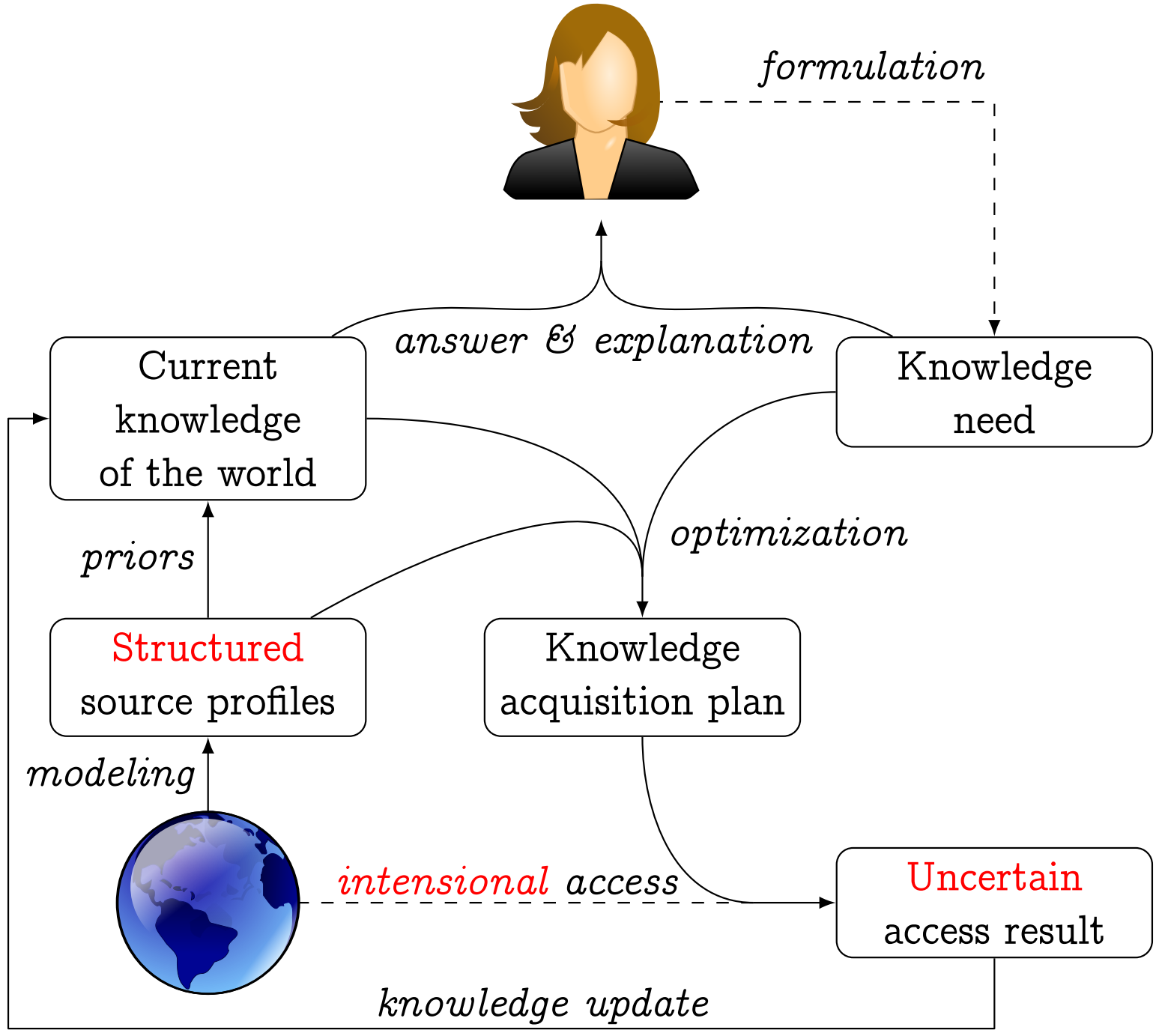
modeling


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

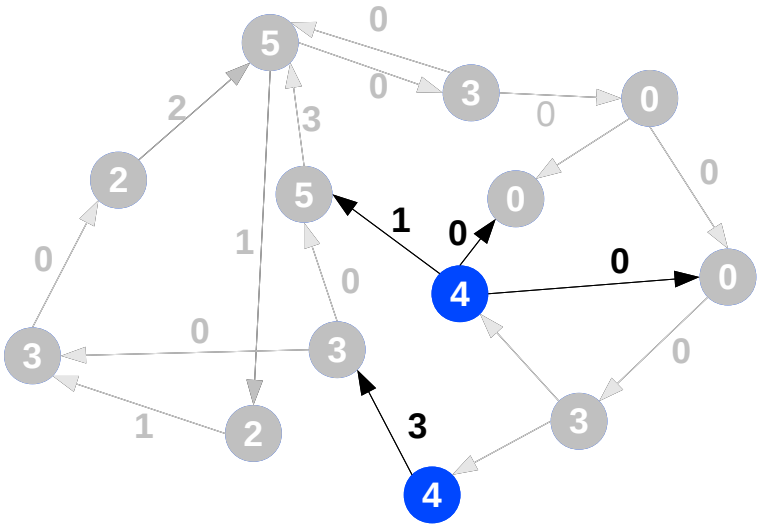

formulation
Current knowledge of the world
Knowledge need
optimization
priors
Structured source profiles
Knowledge acquisition plan
modeling
intensional access
Uncertain access result


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
optimization
priors
Structured source profiles
Knowledge acquisition plan
modeling
intensional access
Uncertain access result
knowledge update

Plan

Adaptive focused crawling
(Gouriten, Maniu, and Senellart 2014)

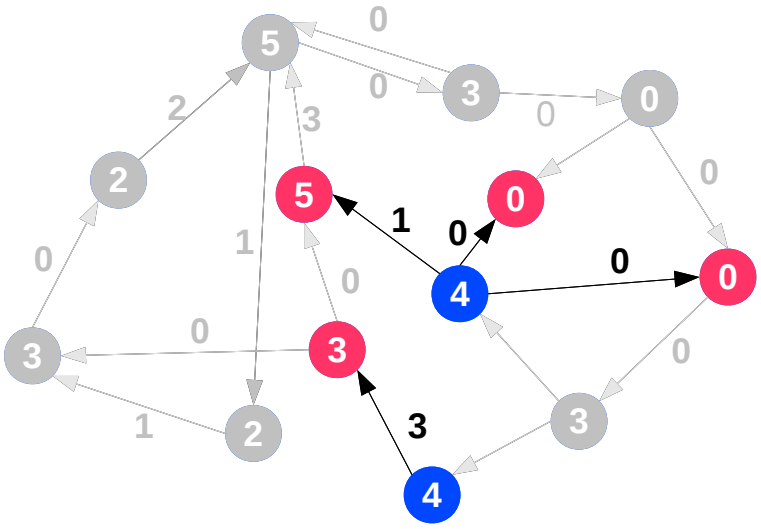


- Problem: Efficiently crawl nodes in a graph such that total score is high
- Challenge: The score of a node is unknown till it is crawled; heavy rate limitations on the number of possible requests
- Methodology: Use various predictors of node scores, and adaptively select the best one so far with multi-armed bandits

with

TELECOM ParisTech

Adaptive focused crawling
(Gouriten, Maniu, and Senellart 2014)

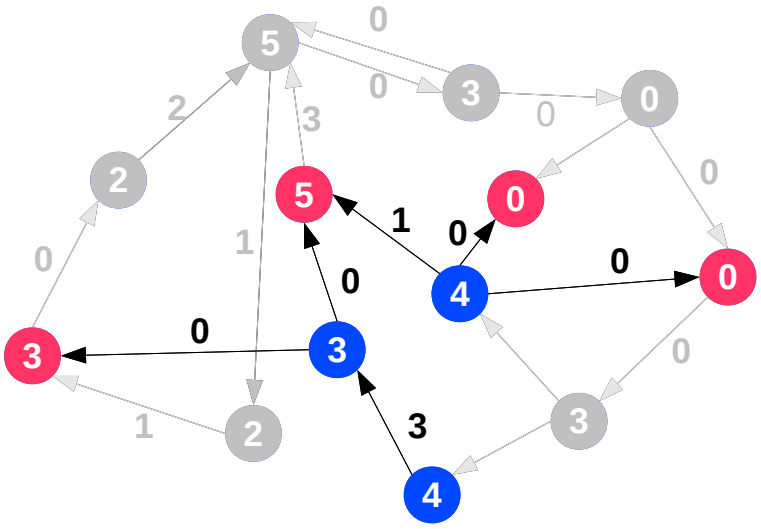


- Problem: Efficiently crawl nodes in a graph such that total score is high
- Challenge: The score of a node is unknown till it is crawled; heavy rate limitations on the number of possible requests
- Methodology: Use various predictors of node scores, and adaptively select the best one so far with multi-armed bandits

with

TELECOM ParisTech

Adaptive focused crawling
(Gouriten, Maniu, and Senellart 2014)

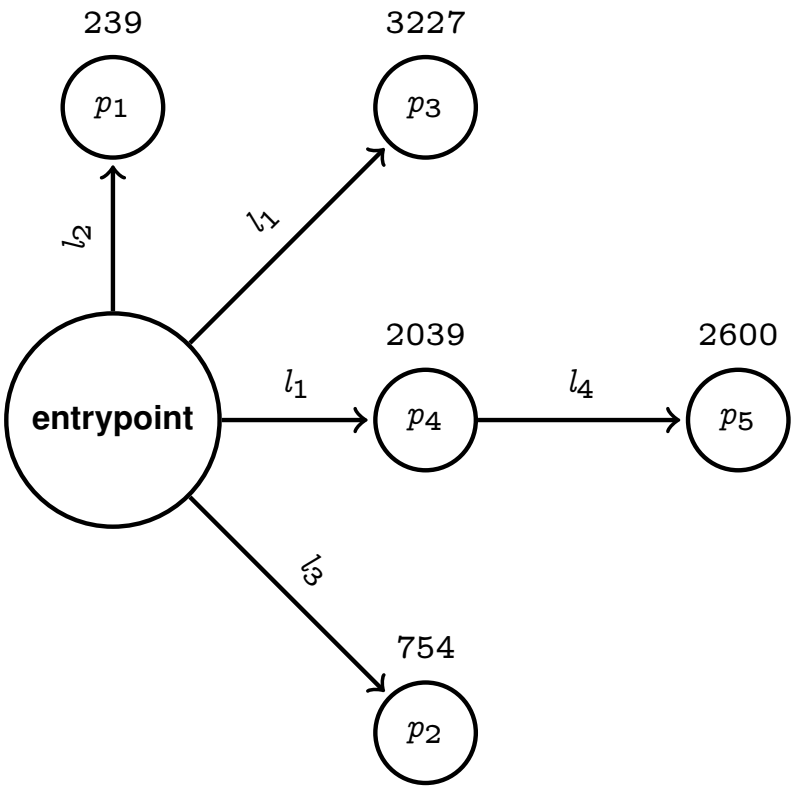


- Problem: Efficiently crawl nodes in a graph such that total score is high
- Challenge: The score of a node is unknown till it is crawled; heavy rate limitations on the number of possible requests
- Methodology: Use various predictors of node scores, and adaptively select the best one so far with multi-armed bandits

with

TELECOM ParisTech

Adaptive Web application crawling

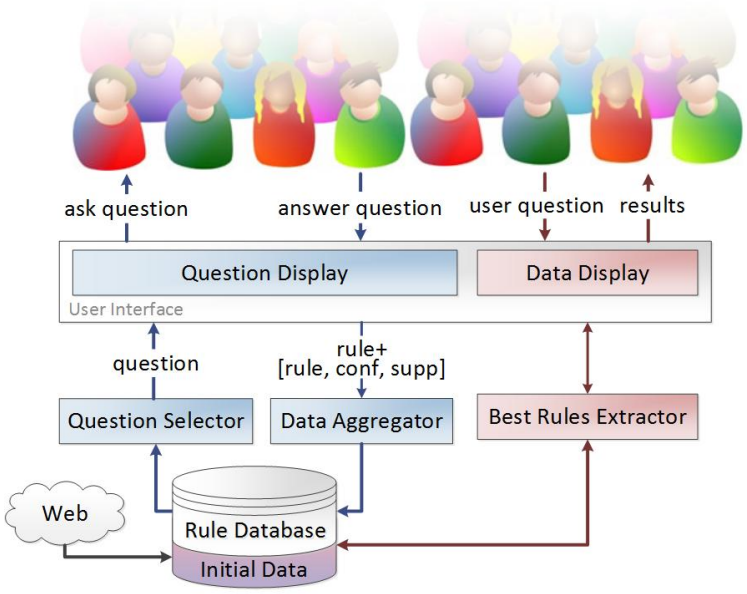


- Problem: Optimize the amount of distinct content retrieved from a Web site w.r.t. the number of HTTP requests
- Challenge: No way to know a priori where the content lies on the Web site
- Methodology: Sample a small part of the Web site and discover optimal crawling patterns from it

with

TELECOM ParisTech

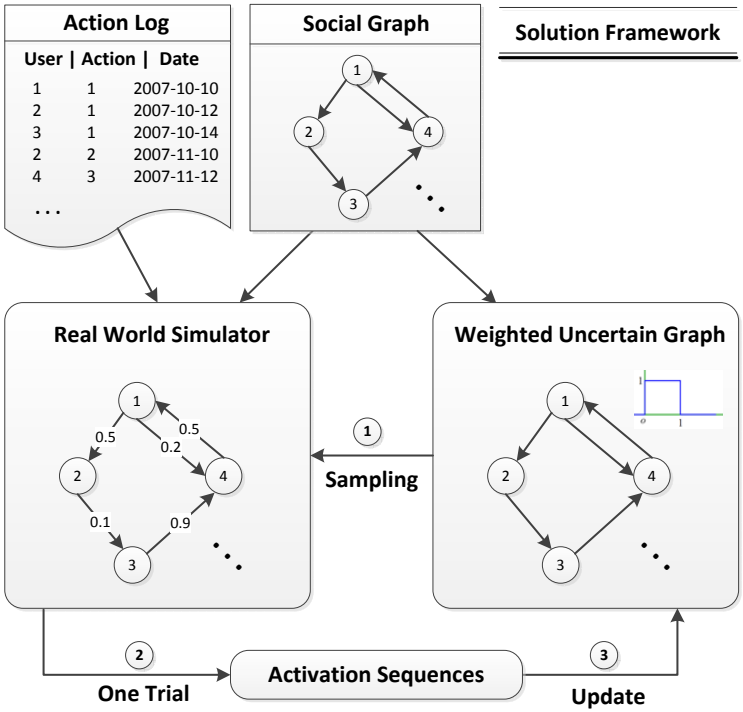
Optimizing crowd queries under order



- Problem: Given a query, what is the next best question to ask the crowd when crowd answers are constrained by a partial order
- Challenge: Order constraints make questions not independent of each other
- Methodology: Construct a polytope of admissible regions and uniformly sample from it to determine the impact of a data item

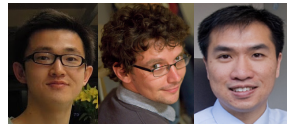
with 

Online influence maximization



- Problem: Run influence campaigns in social networks, optimizing the amount of influenced nodes
- Challenge: Influence probabilities are unknown
- Methodology: Build a model of influence probabilities and focus on influent nodes, with an exploration/exploitation trade-off

with



TELECOM ParisTech

Query answering under uncertain rules

$\text{Pope}(X) \Rightarrow \text{BuriedIn}(X, \text{Rome}) \quad (98\%)$

$\text{LocatedIn}(X, \text{Lombardy}) \Rightarrow$
$\quad \text{BelongsTo}(X, \text{AustrianEmpire}) \quad (45\%)$

$\text{Pope}(\text{PiusXI})$

$\text{BornIn}(\text{PiusXI}, \text{Desio})$

$\text{LocatedIn}(\text{Desio}, \text{Lombardy})$

$\exists X, \text{BornIn}(X, Y) \wedge$
$\quad \text{BelongsTo}(Y, \text{AustrianEmpire}) \wedge$
$\quad \text{BuriedIn}(X, \text{Rome})?$

- Problem: Determine efficiently the probability of a query being true, given some data and uncertain rules over this data
- Challenge: Produced facts may be correlated, the same facts can be generated in different ways, probability computation is hard in general...
- Methodology: Find restrictions on the rules (guarded?) and the data (bounded tree-width?) that make the problem tractable

with

TELECOM ParisTech

Plan

DBWeb Research

Focus: Uncertainty, Structure, Intensionality

Instances of UnSAID

Conclusion

What's next?

- So far, we have tackled individual aspects or specializations of the UnSAID problem
- Now we need to consider the general problem, and propose general solutions
- There is a strong potential for uncovering the unsaid information from the Web
- Strong connections with a number of research areas: active learning, reinforcement learning, adaptive query evaluation, etc. Inspiration to get from these areas.
- Everyone is welcome to join the effort!

What's next?

- So far, we have tackled individual aspects or specializations of the UnSAID problem
- Now we need to consider the general problem, and propose general solutions
- There is a strong potential for uncovering the unsaid information from the Web
- Strong connections with a number of research areas: active learning, reinforcement learning, adaptive query evaluation, etc. Inspiration to get from these areas.
- Everyone is welcome to join the effort!

Merci.

Ba, M. L., S. Montenez, T. Abdessalem, and P. Senellart (2014). *Extracting, Integrating, and Visualizing Uncertain Web Information about Moving Objects*. Preprint available at http://pierre.senellart.com/publications/ba2014extracting.pdf.

Fink, R., A. Hogue, D. Olteanu, and S. Rath (2011). "SPROUT^2: a squared query engine for uncertain web data". In: *SIGMOD*.

Gouriten, G., S. Maniu, and P. Senellart (2014). *Scalable, Generic, and Adaptive Systems for Focused Crawling*. Preprint available at http://pierre.senellart.com/publications/gouriten2014scalable.pdf.

Maniu, S., R. Cheng, and P. Senellart (2014). *ProbTree: A Query-Efficient Representation of Probabilistic Graphs*. Preprint available at http://pierre.senellart.com/publications/maniu2014probtree.pdf.

Suchanek, F. M., G. Kasneci, and G. Weikum (2007). "YAGO: A Core of Semantic Knowledge. Unifying WordNet and Wikipedia". In: *WWW*, pp. 697–706. ISBN: 978-1-59593-654-7.

Tang, R., A. Amarilli, P. Senellart, and S. Bressan (2014). *Get a Sample for a Discount: Sampling-Based XML Data Pricing*. Preprint available at http://pierre.senellart.com/publications/tang2014get.pdf.