UnSAID: Uncertainty and Structure in the Access to Intensional Data

Pierre Senellart

2 April 2014, *IC2 Lunch Seminar*

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

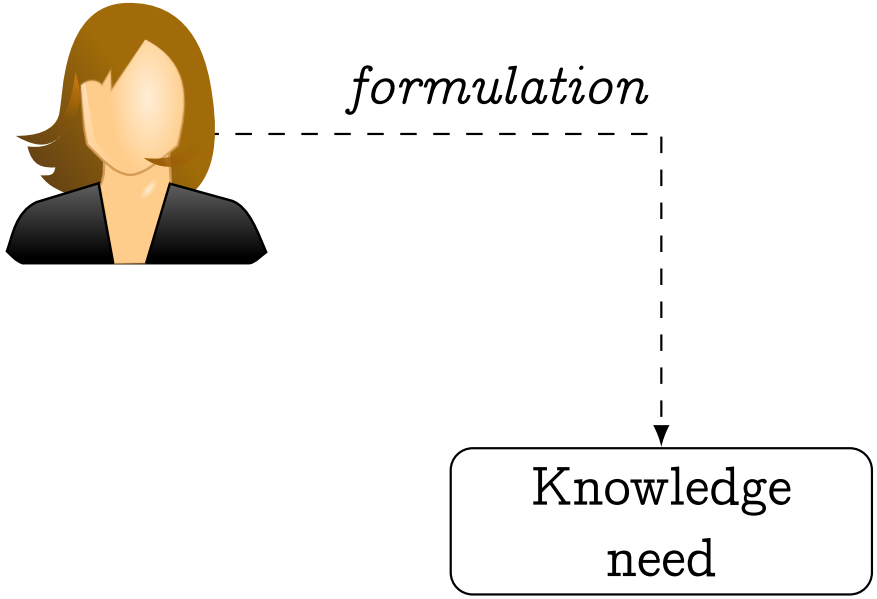
Intensional data is everywhere

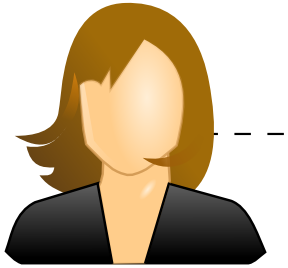
Lots of data sources can be seen as intensional: accessing all the data in the source (in extension) is impossible or very costly, but it is possible to access the data through views, with some access constraints, associated with some access cost.

- 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

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 next best access to do 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

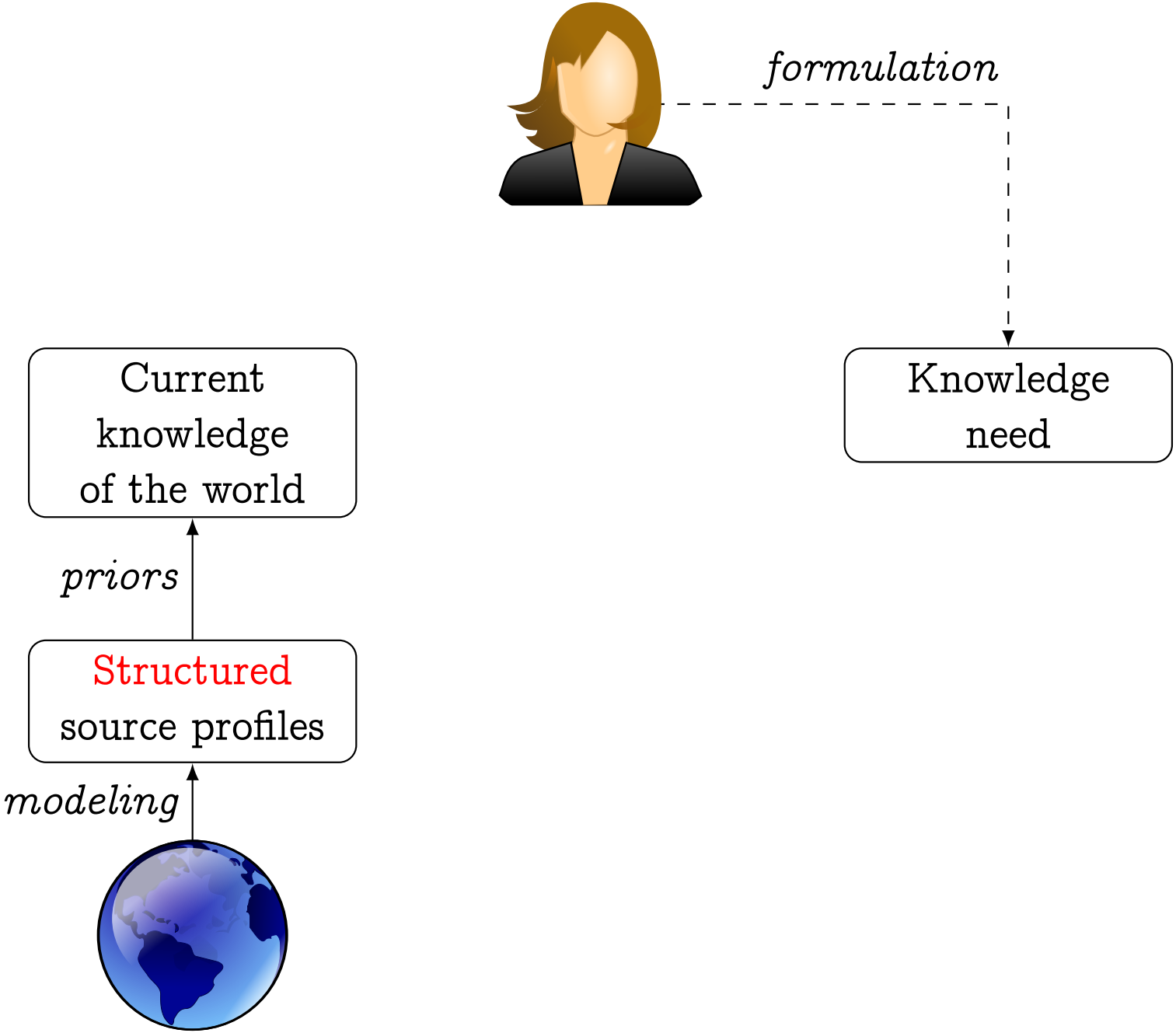


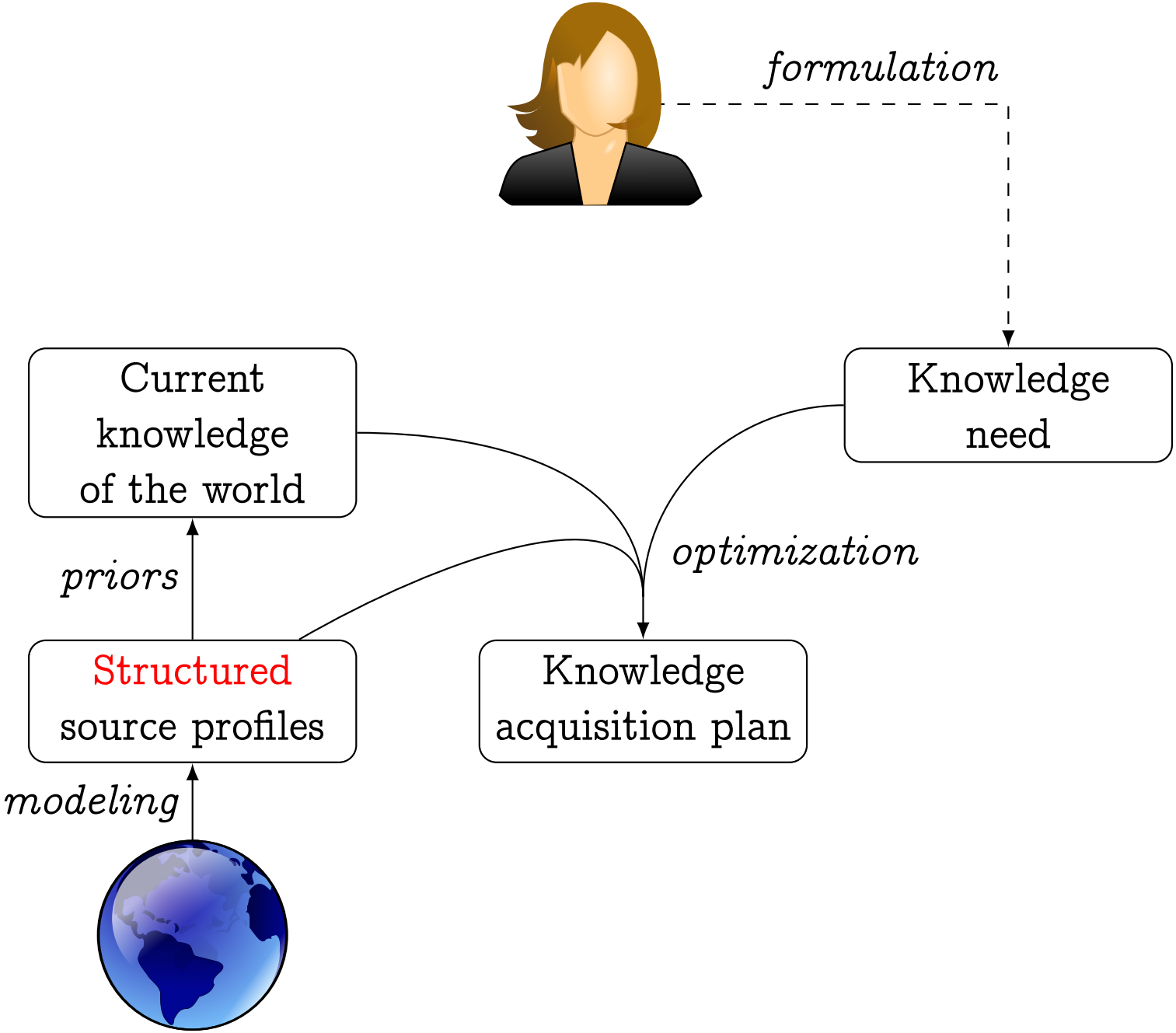
formulation

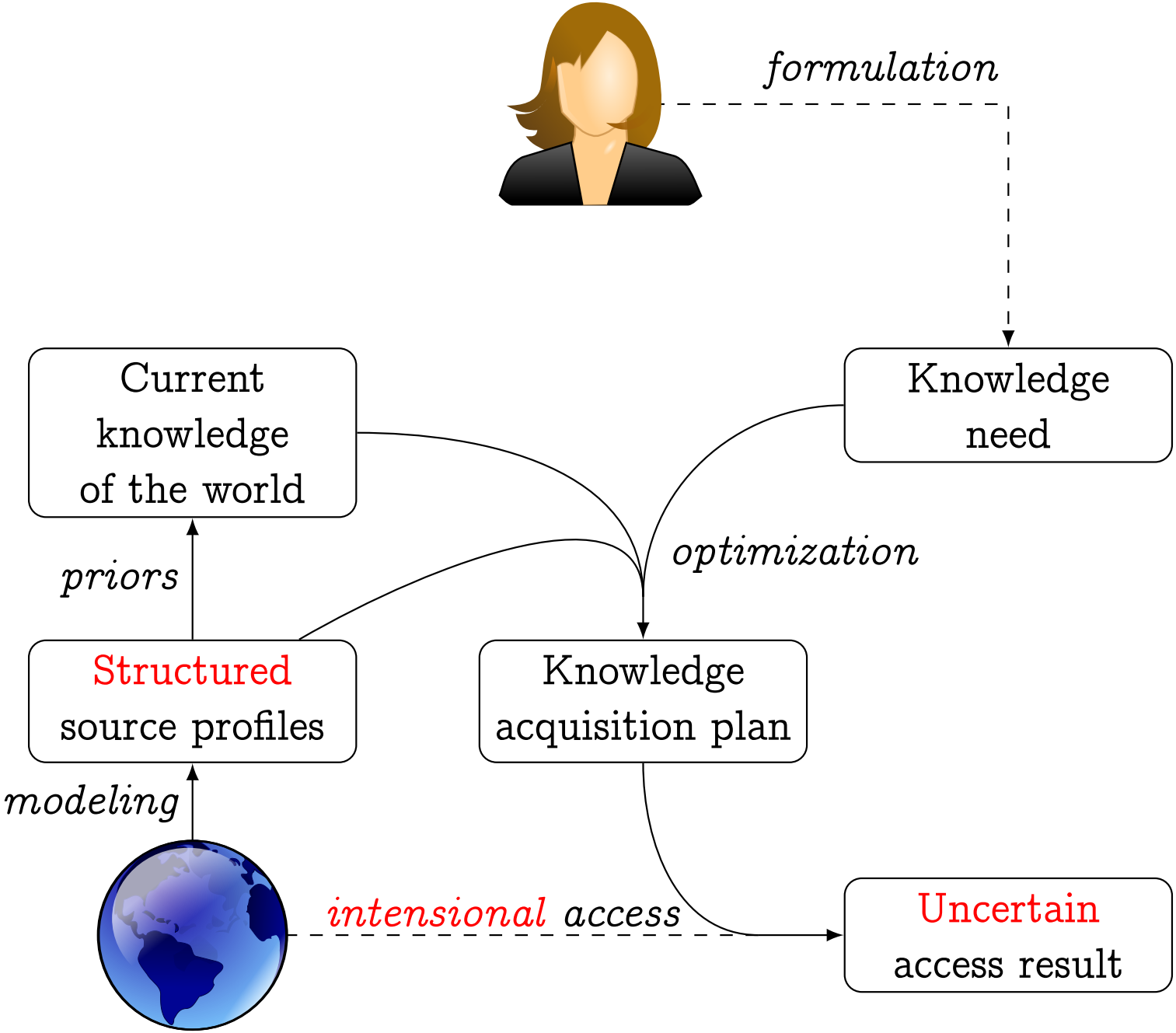
Knowledge need

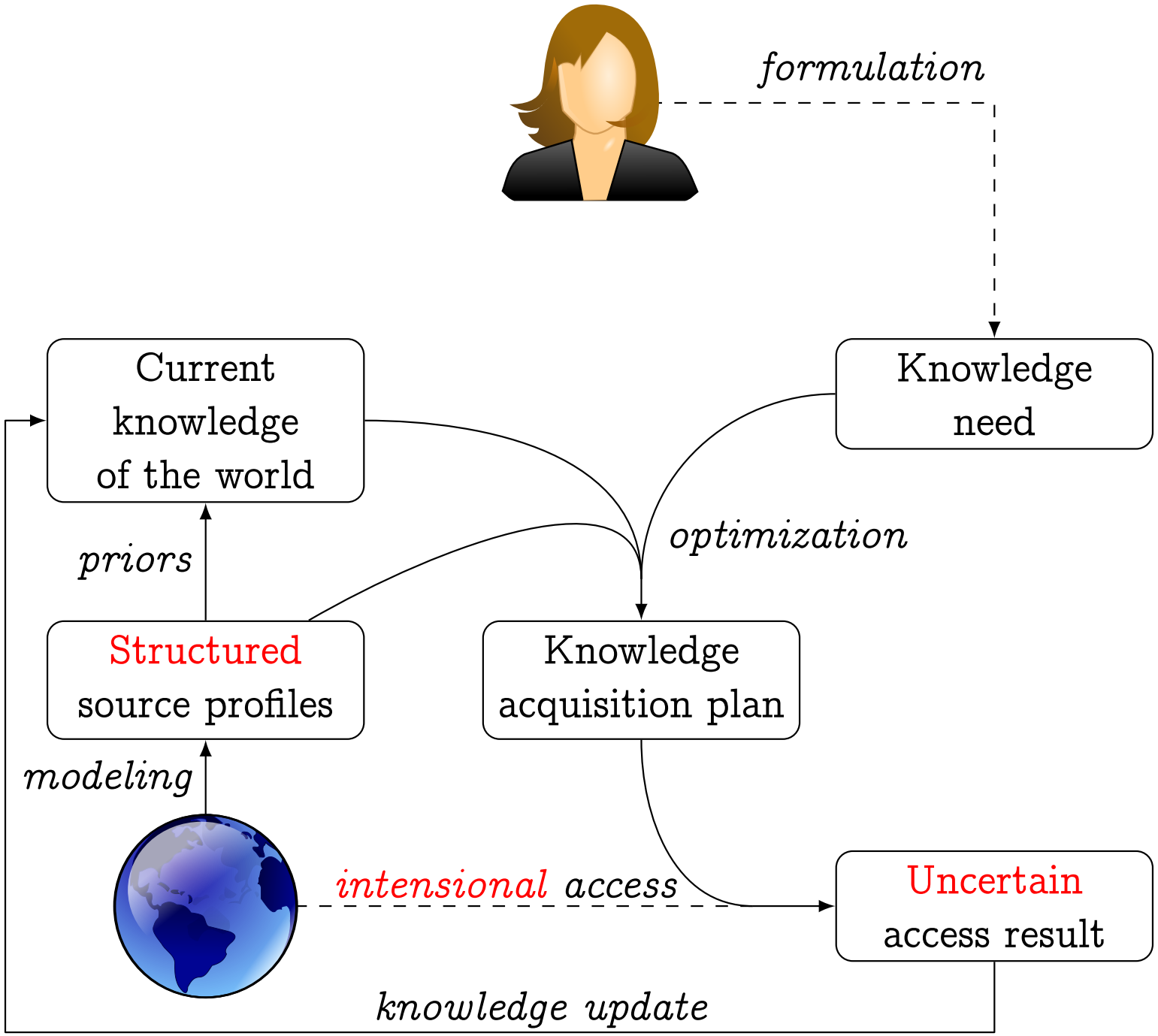
Structured source profiles

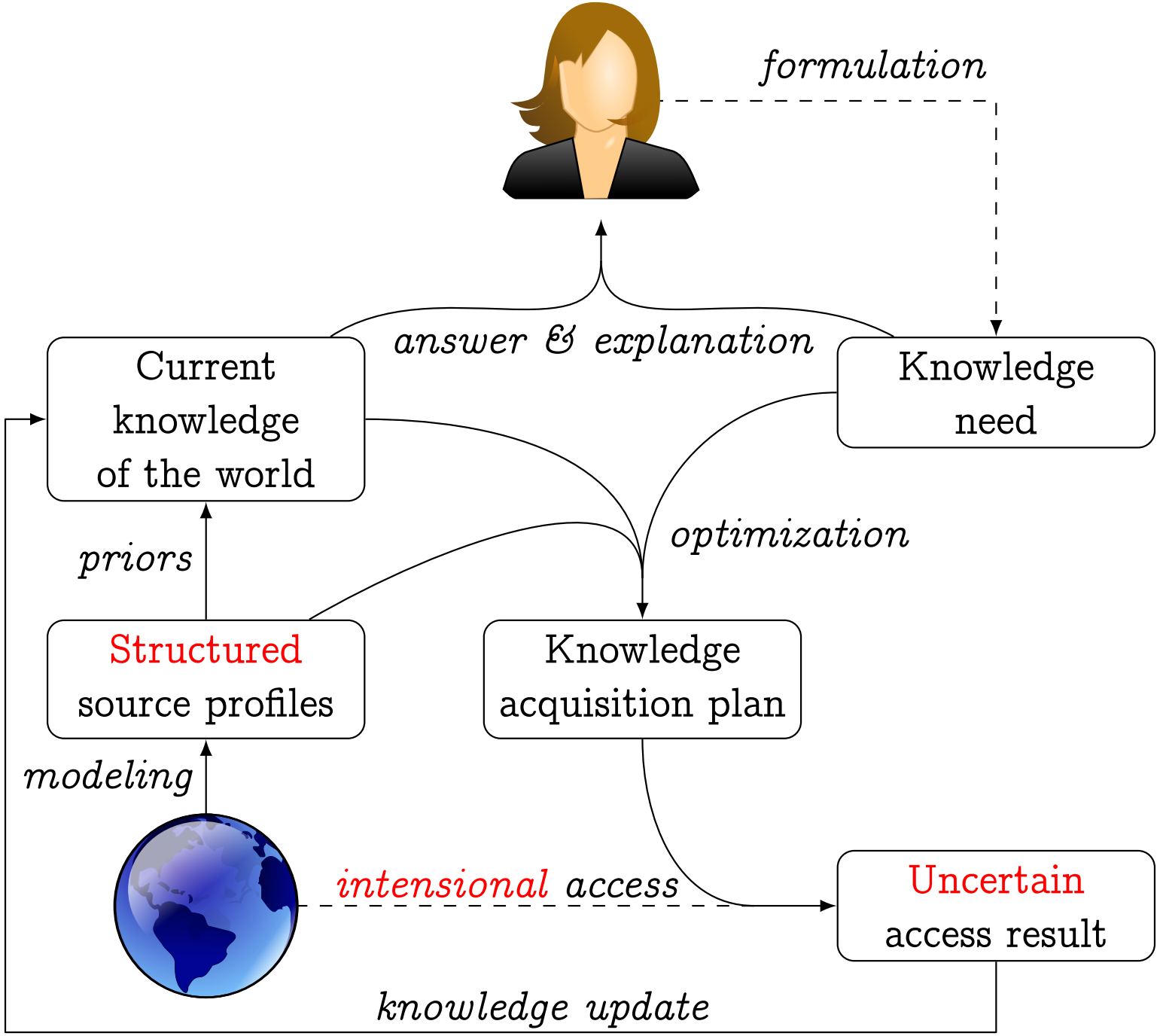
modeling


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


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

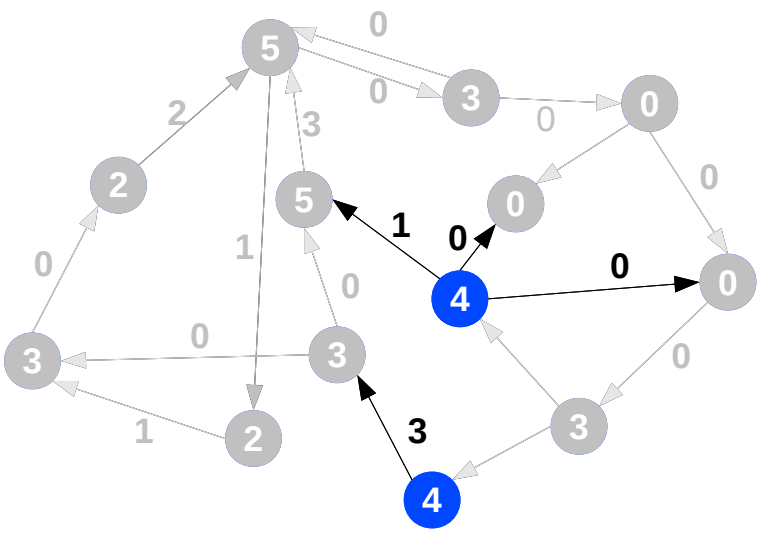

formulation
answer & explanation
Current knowledge of the world
Knowledge need
optimization
priors
Structured source profiles
Knowledge acquisition plan
modeling
intensional access
Uncertain access result
knowledge update

What this talk is about

- General overview of my current (and recent) research, through one-slide presentation of individual works
- Hopefully, emerging consistent themes
- Connections with the UnSAID problem

Plan

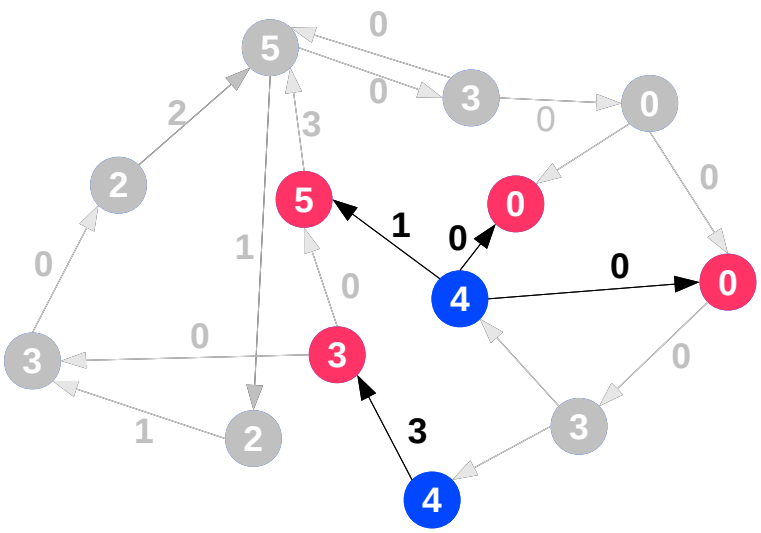
Adaptive focused crawling
(Gouriten, Maniu, and Senellart 2014)



- 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
- Methodology: Use various predictors of node scores, and adaptively select the best one so far with multi-armed bandits

with 

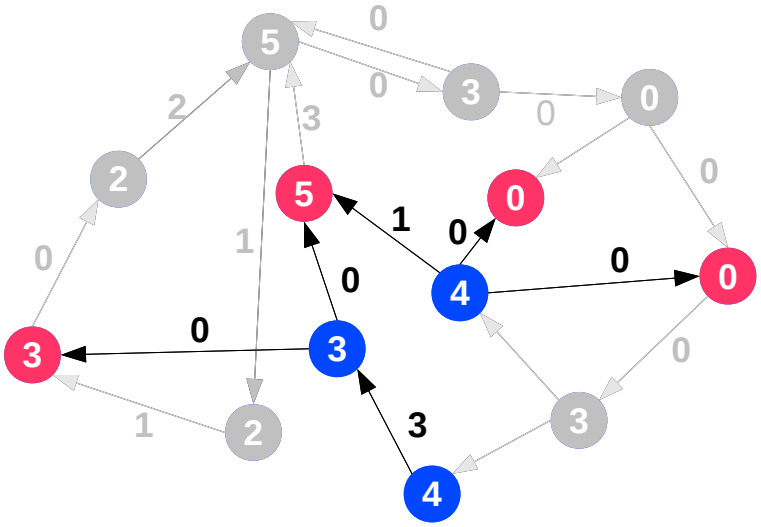
Adaptive focused crawling
(Gouriten, Maniu, and Senellart 2014)



- 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
- Methodology: Use various predictors of node scores, and adaptively select the best one so far with multi-armed bandits

with 

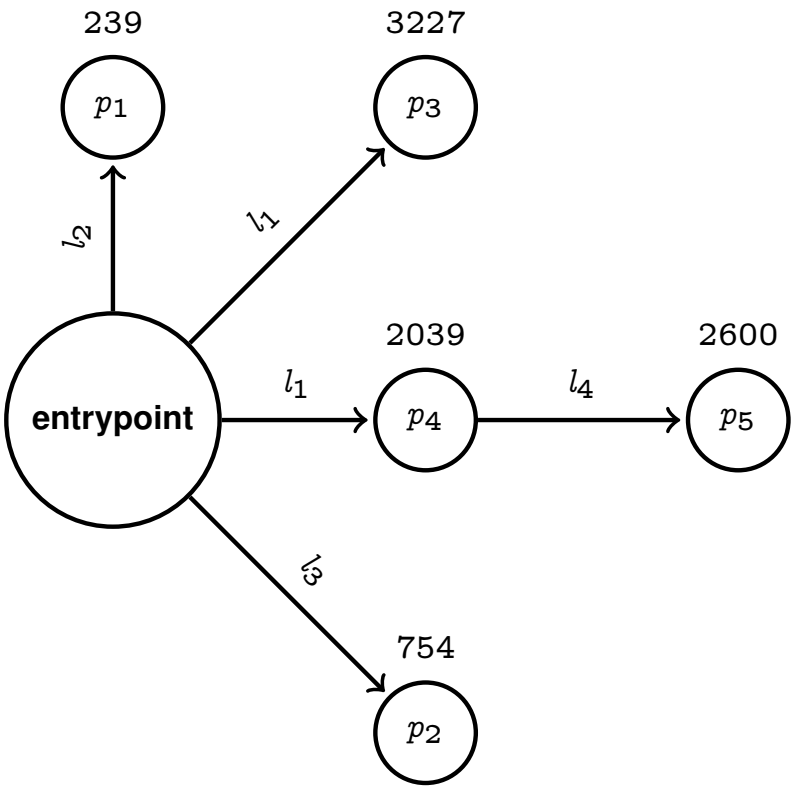
Adaptive focused crawling
(Gouriten, Maniu, and Senellart 2014)



- 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
- Methodology: Use various predictors of node scores, and adaptively select the best one so far with multi-armed bandits

with 

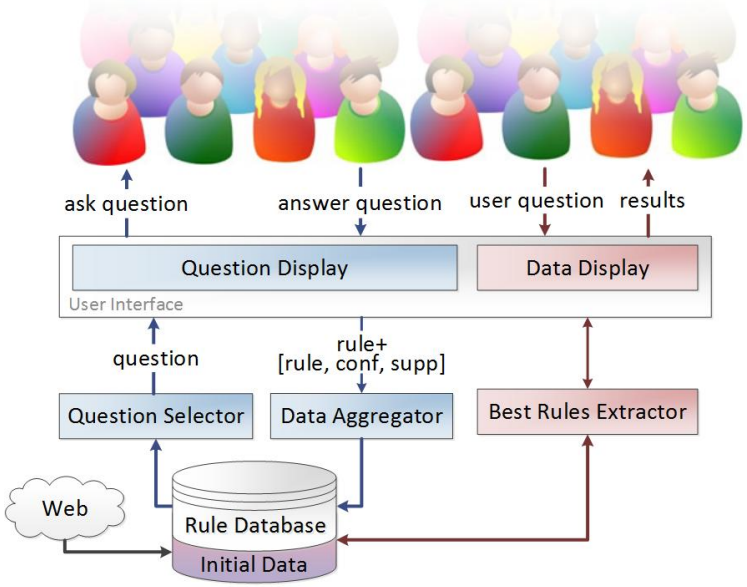
Adaptive Web application crawling



- 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 

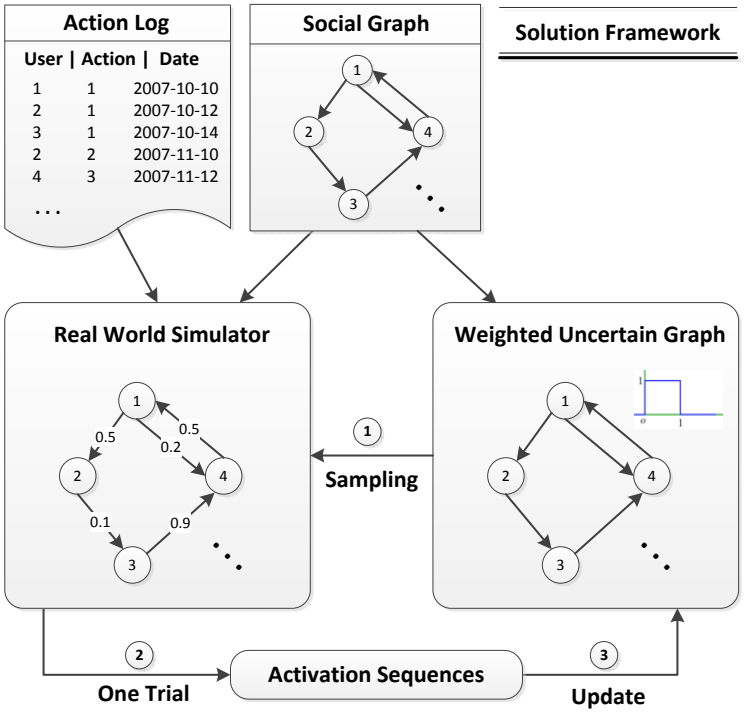
Optimizing crowd queries under order



- 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

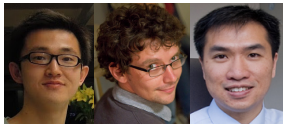
with 

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



Query answering under uncertain rules

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

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

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

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

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

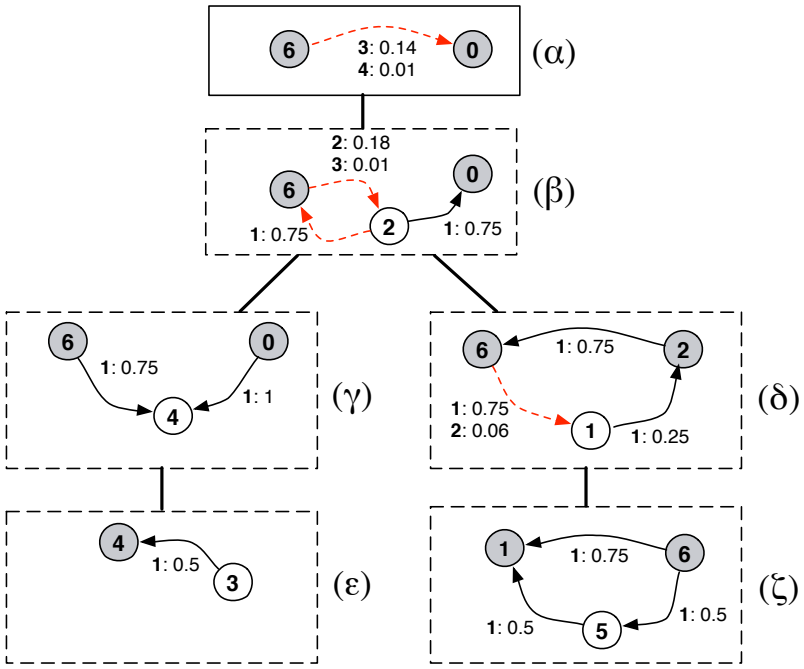
$\exists X, \text{BornIn}(X, Y) \wedge$
$\quad \text{BelongsTo}(Y, \text{AustrianEmpire}) \wedge$
$\quad \text{BuriedIn}(X, \text{Rome})?$

- 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

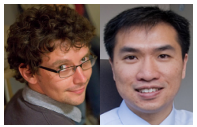
Plan

Efficient querying of uncertain graphs
(Maniu, Cheng, and Senellart 2014)



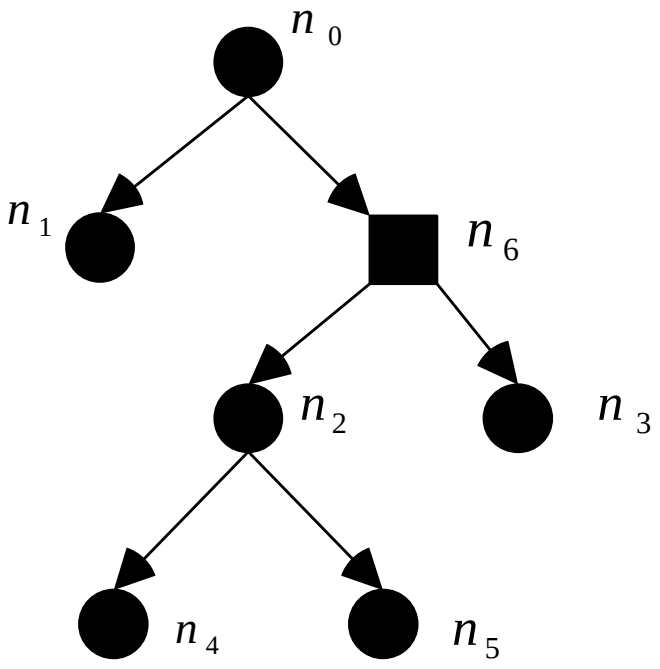
- Problem: Optimize query evaluation on probabilistic graphs
- Challenge: Probabilistic query evaluation is hard, and standard indexing techniques for large graphs do not work
- Methodology: Build a tree decomposition that preserves probabilities and run the query on this tree decomposition

with



TELECOM ParisTech

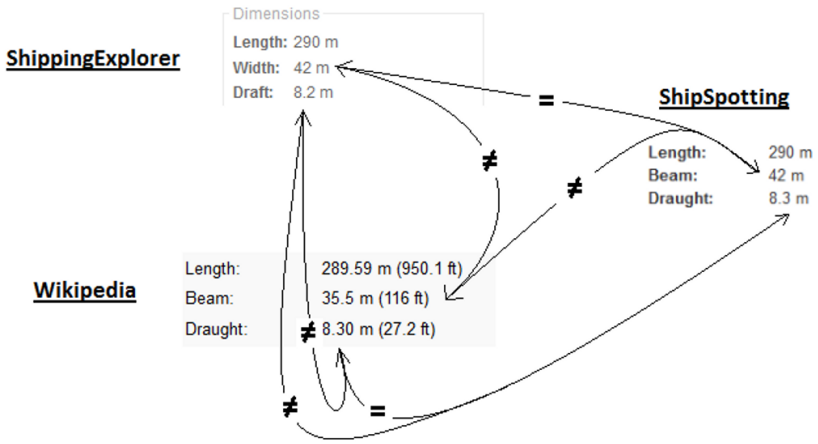
Uniform sampling of XML documents
(Tang, Amarilli, Senellart, and Bressan 2014)



- Problem: Sample a subtree of fixed size/characteristics uniformly at random from a tree, e.g., for data pricing reasons
- Challenge: Naive top-down sampling does not work, will result in biased sampling depending on the tree structure
- Methodology: Bottom-up annotation of the tree recording distribution information, followed by top-down sampling

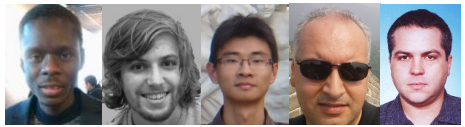
with 

Truth discovery on heterogeneous data

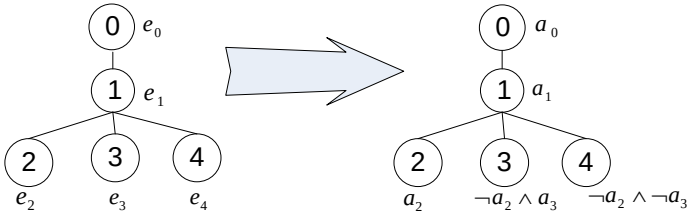


- Problem: Determine true values from integrated semi-structured Web sources
- Challenge: Semi-structured data on the Web is contradictory and copied from source to source
- Methodology: Estimate the truth and determine copy patterns between sources, not only of base facts but of subtrees of the data

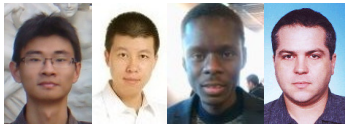
with



Probabilistic XML conditioning



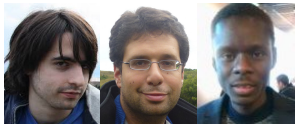
- Problem: Incorporate a logical constraint into a probabilistic XML database
- Challenge: Constraining is not a standard probabilistic database operation, NP-hard in general
- Methodology: Identify tractable subcases

with 

Provenance and Order
(Amarilli, Ba, Deutch, and Senellart 2013)

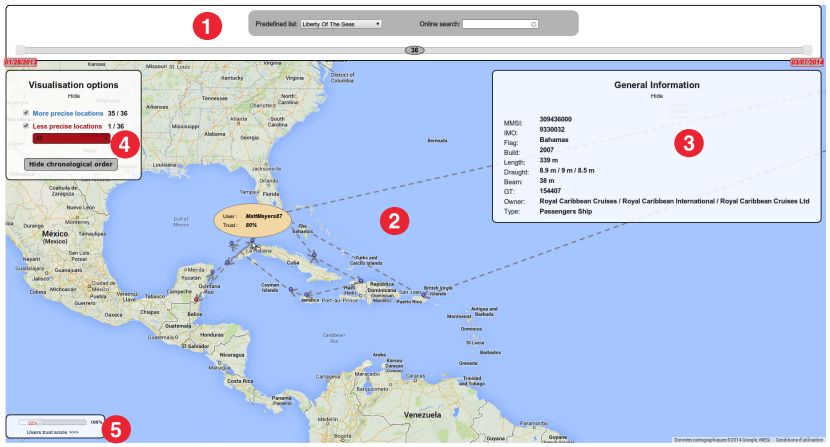
Read	Ord	Ann
Light	$(t_0 \cdot [0 < Y_0]) \otimes 1 + (t_1 \cdot [1 < Y_0]) \otimes 1$	r_0
Door	$(t_0 \cdot [0 < Y_0 + Y_1]) \otimes 1 + (t_1 \cdot [1 < Y_0 + Y_1]) \otimes 1$	r_1
Light	$(t_0 \cdot [0 < Y_0 + Y_1]) \otimes 1 + (t_1 \cdot [1 < Y_0 + Y_1]) \otimes 1 + (t_2 \cdot [1 < Y_0 + Y_1 + Y_2]) \otimes 1$	r_2
Beep	$(r_0 \cdot [Y_0 \leqslant 0]) \otimes 1 + (r_1 \cdot [Y_0 + Y_1 \leqslant 0]) \otimes 1 + (r_2 \cdot [Y_0 + Y_1 + Y_2 \leqslant 0]) \otimes 1$	t_0
Door	$t_0 \otimes 1 + (r_0 \cdot [Y_0 \leqslant 1]) \otimes 1 + (r_1 \cdot [Y_0 + Y_1 \leqslant 1]) \otimes 1 + (r_2 \cdot [Y_0 + Y_1 + Y_2 \leqslant 1]) \otimes 1$	t_1

- Problem: Clean semantics for nondeterministic order-aware queries
- Challenge: Existing data manipulation languages treat order in an ad hoc manner, with no compositional semantics
- Methodology: Order-aware type system for strict static checking, and individual data-level provenance annotations for dynamic analysis of allowed ordered operations

with 

Plan

Social Sensing of Moving Objects
(Ba, Montenez, Abdessalem, and Senellart 2014)



- Problem: Infer trajectories and meta-information of moving objects from Web and social Web data
- Challenge: Uncertainty and inconsistency in extracted information
- Methodology: Data cleaning by filtering incorrect locations, and truth discovery to identify reliable sources

with 

Smarter urban mobility

- Problem: Smart and adaptive recommendations for mobility in cities (transit, bike rental, car, etc.)
- Challenge: Should take into account personal information (calendar, etc.), past trajectories, public information about transit and traffic
- Methodology: Map GPS tracks to routes of public transport, learn route patterns, infer destination while in transit and provide push suggestions

with

Plan

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.

Amarilli, A., M. L. Ba, D. Deutch, and P. Senellart (Dec. 2013). *Provenance for Nondeterministic Order-Aware Queries*. Preprint available at http://pierre.senellart.com/publications/amarilli2014provenance.pdf.

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

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

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