Reasoning on Data: Challenges and Applications

Marie-Christine Rousset

Laboratoire d'Informatique de Grenoble

Université Grenoble Alpes & Institut Universitaire de France

Data are everywhere …

multi-form, multi-source, multi-scale

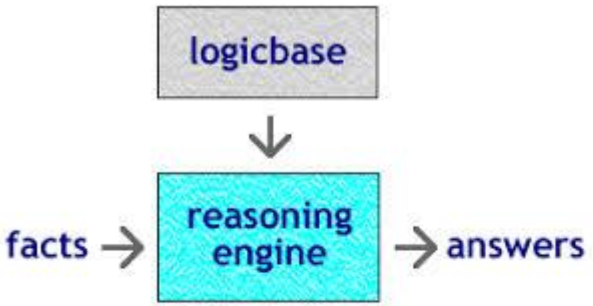
their use raises practical, theoretical and societal challenges for helping humans …

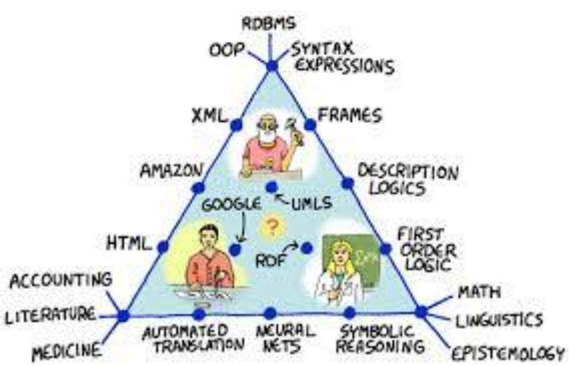
…. to :

- take decisions
- make a diagnosis
- plan actions
- do prediction
- etc …

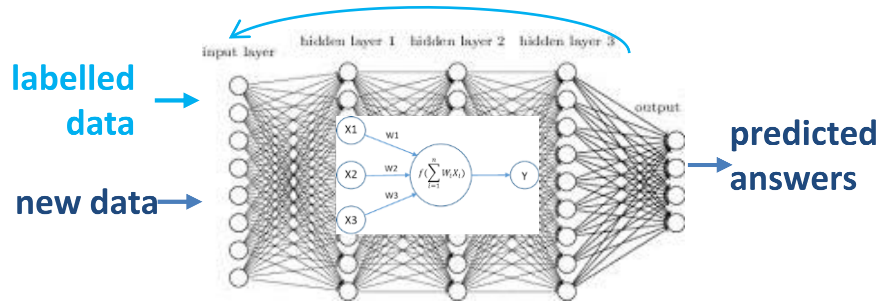
Two branches of Artificial Intelligence

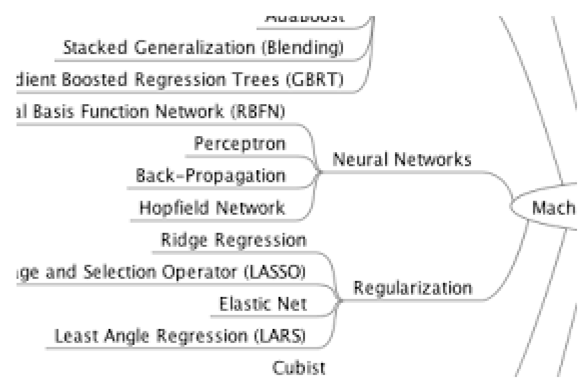
Symbolic knowledge-driven approaches





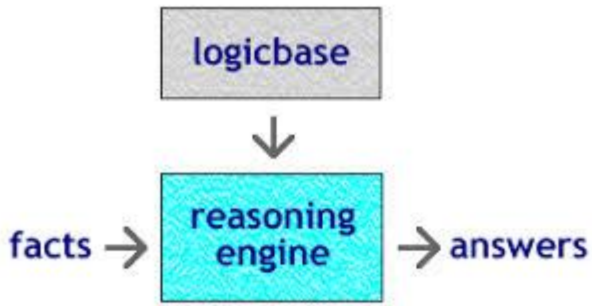
Numerical data-driven approaches



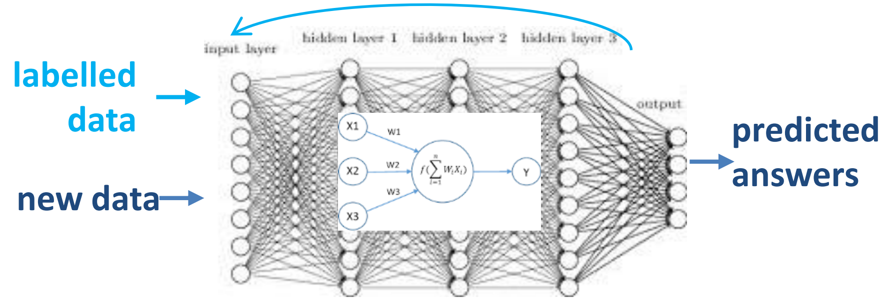


Two branches of Artificial Intelligence

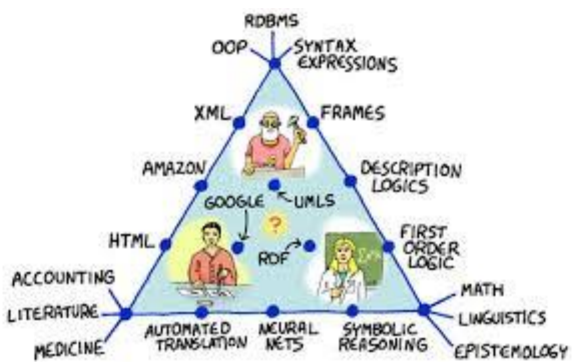
Symbolic knowledge-driven approaches



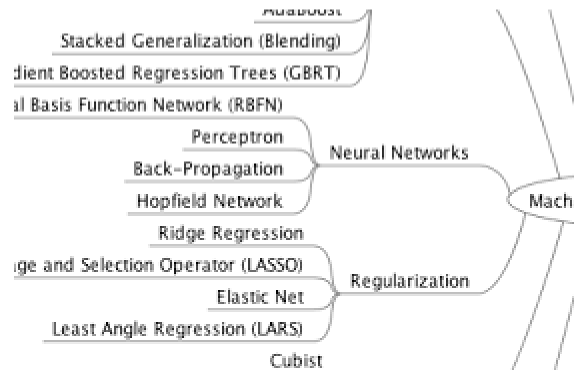
Numerical data-driven approaches



a.k.a Good-Old-Fashioned AI (GOFAI)



a.k.a Modern AI



Respective advantages and disadvantages

Explicability and transparency:

all reasoning steps to reach a conclusion are based on symbolic human readable representations

Robustness and scalability:

- the rules and knowledge have to be hand coded … but more and more work on learning rules from data
- the generic reasoning algorithms may have a high computational complexity (atleast in the worst-case)

Automated Reasoning

- Problem studied in Mathematics, Logic and Informatics
  - Many decidability and complexity results coming from decades of research in the KR&R community
  - Several inference algorithms and implemented reasoners

- The key point
  - first-order-logic is appropriate for knowledge representation
  - but **full first-order-logic is not decidable**

$\Rightarrow$ the game is to find restrictions to design:

- decidable fragments of first-order-logic
- expressive enough for modeling useful knowledge or constraints

Key logic-based knowledge representation formalisms

- **Rules:** logical foundation of **expert systems**
  - the first successful and commercial AI systems (in the 1970s)
    - human expertise in a specific domain is captured as **a set of if-then rules**
    - given **a set of input facts, the inference engine** triggers **relevant rules** to build a chain of reasoning arriving to a particular conclusion
  - extended to fuzzy rules to deal with uncertain reasoning
- **Conceptual graphs**: a **graphical representation of logic**
  - logical formalism focused on representing individuals by their classes and relations (> mid-eighties)
    - originated from semantic networks (introduced to represent meaning of sentences in natural language)
  - reasoning algorithms based on **graph operations**
    - directly applicable to Linked Data for querying RDF knowledge bases (RDF graphs constrained by RDFS statements)
- **Description logics**: logical foundation of **ontologies** and **the Semantic Web**

  (started in the early 1990s)

Ontologies

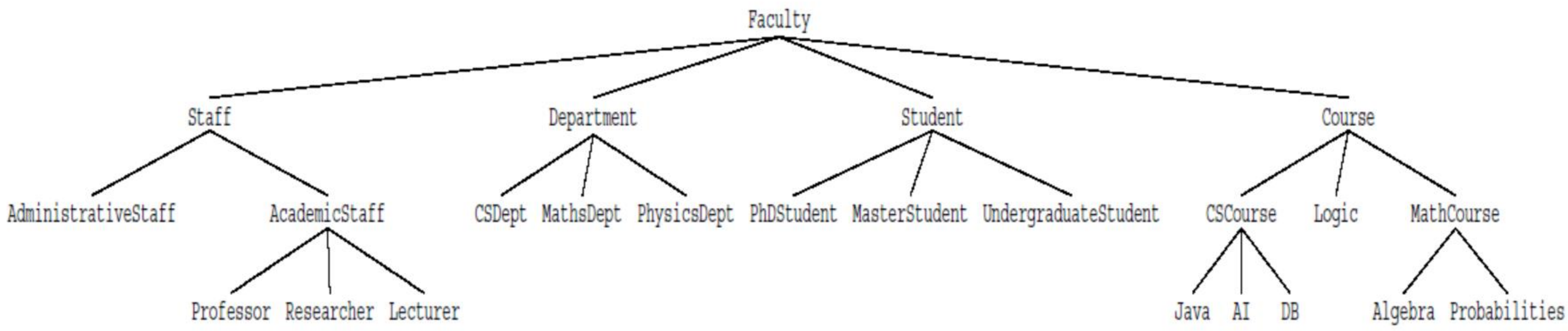
- A formal specification of a domain of interest
  - a vocabulary (classes and properties)
  - enriched with statements that constrain the meaning of the terms used in the vocabulary
    - *java* can be a *dance*, an *island*, a *programming language* or a *course*
    - the statement *java* <u>is a subclass of</u> *CS Courses* makes clear the corresponding meaning for java: it is a course
- With a logical semantics
  - Ontological statements are axioms in logic

⇒ a conceptual yet computational model of a particular domain of interest.

- computer systems can then base decisions on reasoning about domain knowledge.
- humans can express their data analysis needs using terms of a shared vocabulary in their domain of interest or of expertise

Example

A taxonomy (graphical representation of subclass constraints)



\+ set of properties with constraints on their domain and range

TeachesIn (Academic Staff, Courses)

TeachesTo (Academic Staff, Students)

Manager (Staff , Departments)

\+ additional constraints (not expressible in RDFS but in OWL)

Student disjoint from *Staff*

Only *Professors* or *Lecturers* may teach to *Undergraduate Students*

Every *Department* must have a unique *Manager* who must be a *Professor*

Query answering over data through ontologies

- A **reasoning** problem
  - Ontological statements can be used to infer new facts and deduce answers that could not be obtained otherwise
  - Subtlety: some inferred facts can be partially known

From the constraint “a professor teaches at least one master course”

∀x (Professor(x) => ∃ y Teaches(x,y), MasterCourse(y))

and the fact:

Professor(dupond) (RDF syntax: **<dupond, type, Professor>**)

it can be inferred the two following incomplete “facts” :

Teaches(dupond, v) , MasterCourse(v)

i.e, in RDF notation, two RDF triples with blank nodes:

<dupond, Teaches, _v> , <_v, type, MasterCourse>

Reasoning: a tool for checking data inconsistency

- Some ontological statements can be used as **integrity** constraints

  "a professor cannot be a lecturer" ; "a course must have a responsible"

  $\forall$x (Professor(x) => ¬ Lecturer(x))

  $\forall$x (Course(x) => $\exists$ y ResponsibleFor(y,x))

  "a master course is taught by a single teacher"

  "only professors can be responsible of courses that they have to teach"

  $\forall$x $\forall$y (Course(x), ResponsibleFor(y,x) => Professor(y), Teaches(y,x))

- Subtlety: showing data inconsistency may require **intricate reasoning** on different rules, constraints and facts

  The facts: **Lecturer (jim), Teaches(jim, c431) , MasterCourse(c431)**

  + the above integrity constraints

  + the rule $\forall$**x (MasterCourse(x) => Course(x))** leads to an inconsistency

Description Logics

- A family of class-based logical languages for which reasoning is decidable
  - Provides algorithms for reasoning on (possibly complex) logical constraints over unary and binary predicates
- This is exactly what is needed for handling ontologies
  - in fact, the OWL constructs come from Description Logics
- A fine-grained analysis of computational complexity with surprising complexity results
  - $\mathcal{ALC}$ is EXPTIME–complete

=>any sound and complete inference algorithm for reasoning on most of the subsets of constraints expressible in OWL may take an exponential time (in the worst-case)

"only professors or lecturers may teach to undergraduate students"

$\forall x \forall y$ (TeachesTo(x,y), UndergraduateStudent(y) => Professor(x) $\vee$ Lecturer(x))

$$\exists TeachesTo.UndergraduateStudent \sqsubseteq Professor \sqcup Lecturer$$

The same game again…

- Find restrictions on the logical constructs and/or the allowed axioms in order to:
  - design sublanguages for which reasoning is in P

    EL, **DL-Lite**
  - expressive enough for modeling useful constraints over data
- **DL-Lite: a good trade-off**
  - captures the main constraints used in databases and in software engineering
  - extends **RDFS** (the formal basis of OWL2 QL profile)
  - specially designed for answering queries over ontologies to be **reducible to answering queries over RDBMS with same data complexity** (atleast for the fragment of union of conjunctive queries)

Reducibility to query reformulation

Query answering and data consistency checking can be performed in two separate steps:

- a **query reformulation step**
  - **reasoning** on the ontology (and the queries)
  - **independent of the data**

$\Rightarrow$ a set a queries: the reformulations of the input query

- an **evaluation step**
  - of the (SPARQL) query reformulations on the (RDF) data
  - independent of the ontology

$\Rightarrow$ Main advantage

- makes possible to use an SQL or SPARQL engine
- thus taking advantage of well-established query optimization strategies supported by standard relational DBMS

Focus of the remaining of my talk

Focus 1

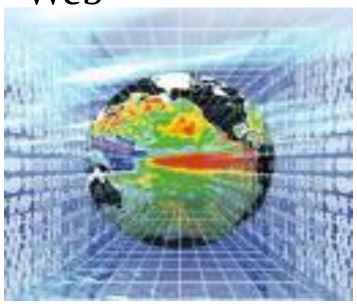
Ontology-based reasoning for data integration

Focus 2

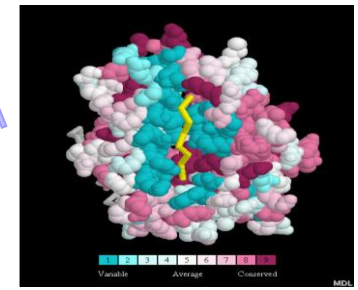
Rule-based reasoning for data linkage

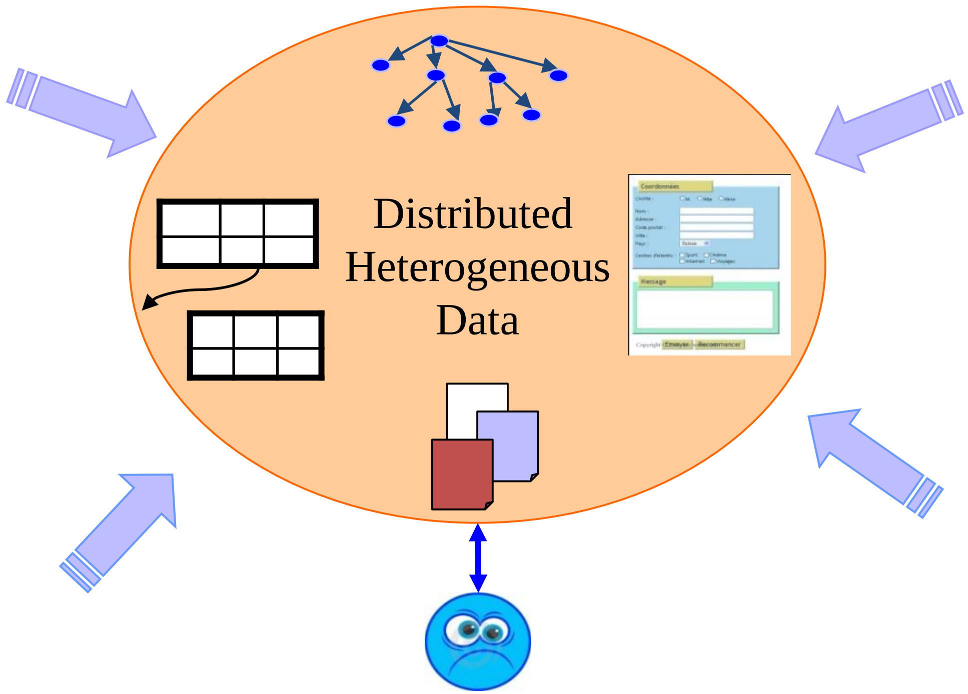
Data Integration

Web

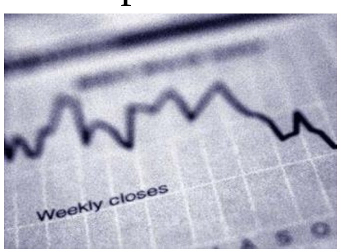


Sciences





Enterprise

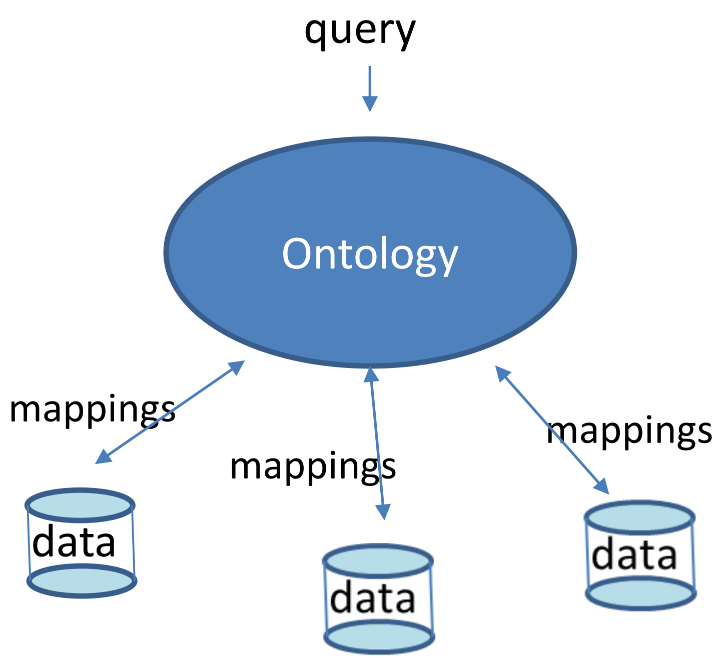


Administration

a difficult challenge !

Domain ontology + mappings:

the semantic glue between heterogeneous data sources



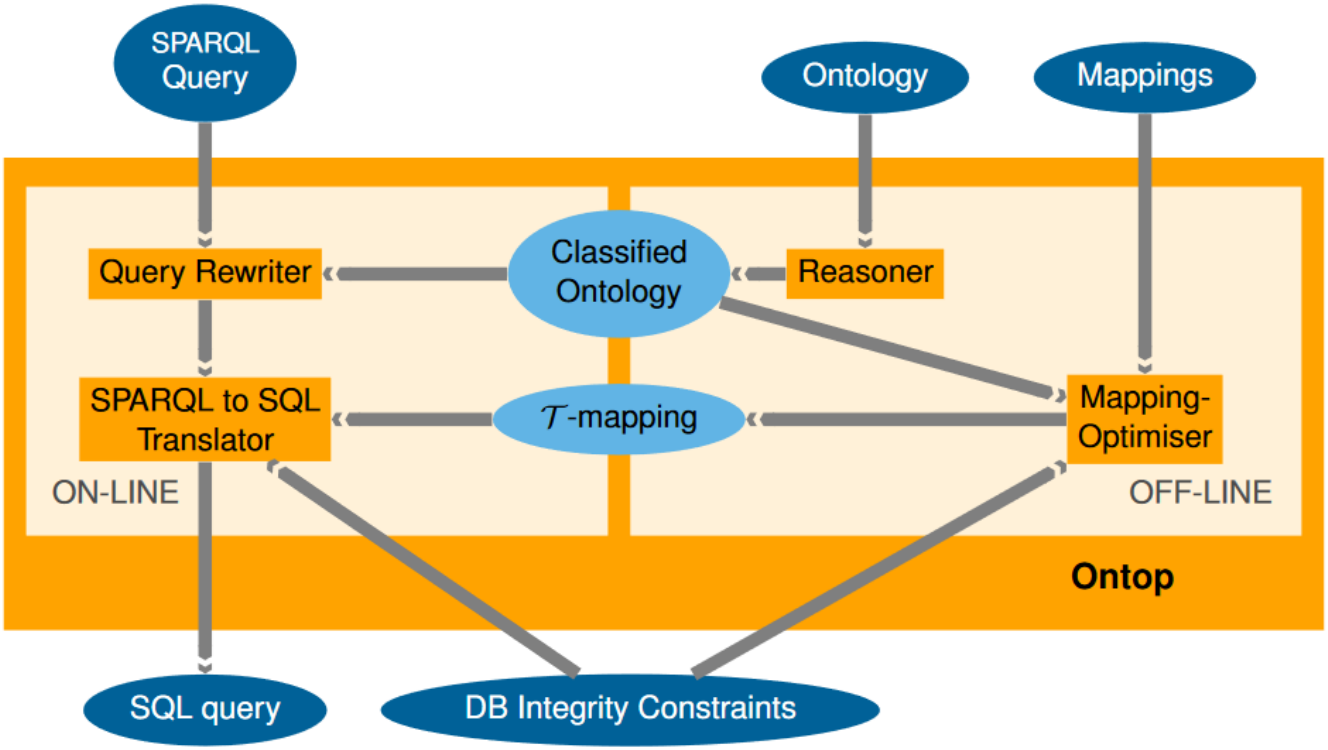
Two main algorithmic approaches

1. **Answering queries by query rewriting :**
   - query reformulation using ontologies (backward reasoning)
   - query translation using mappings
2. **Answering queries by data materialization:**
   - Data extraction and transformation using mappings (e.g., from relational to RDF)
   - Data saturation (forward reasoning on data and ontological statements)

The complexity and feasability in practice depend on the languages used for expressing the queries, the mappings and the ontology

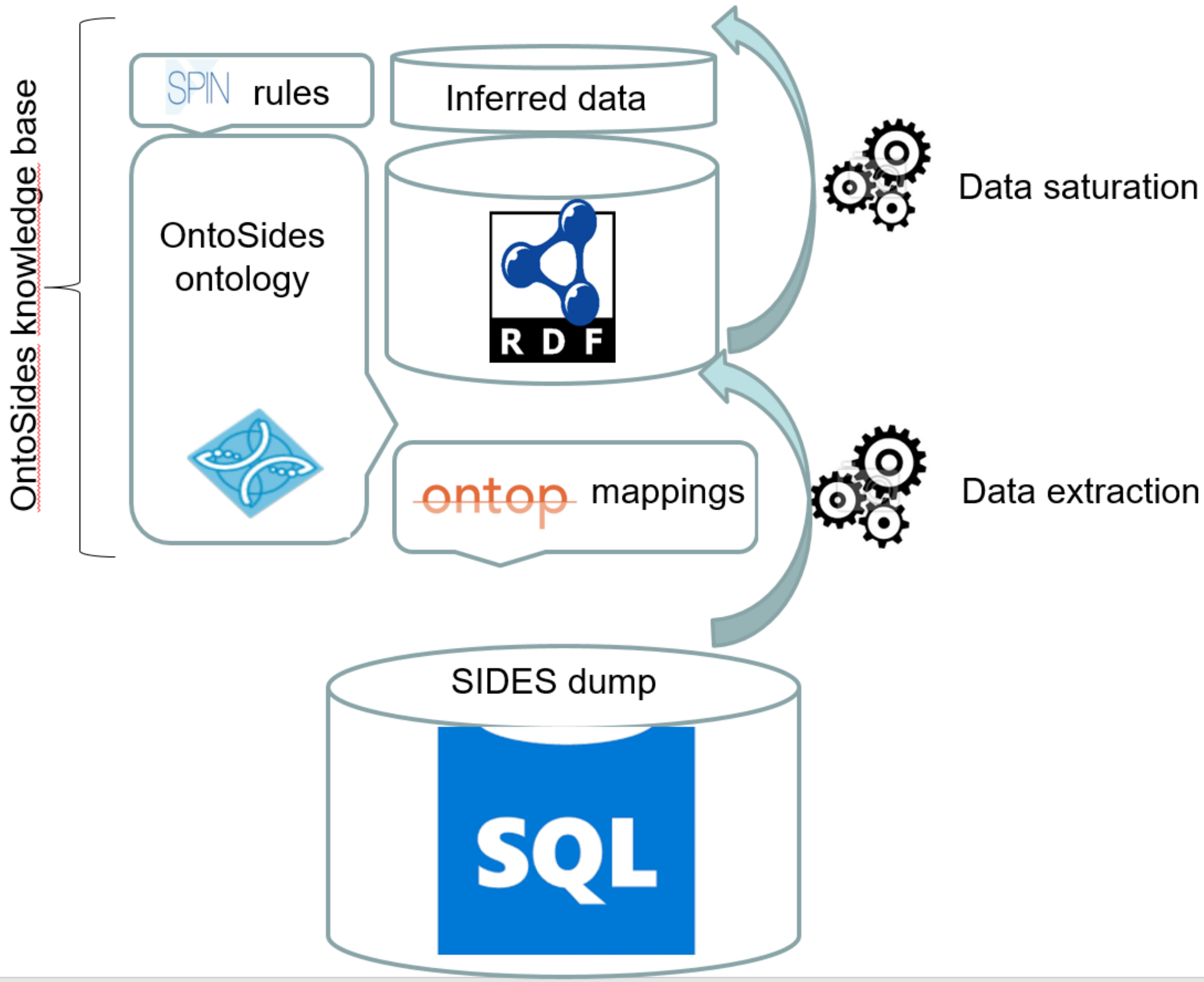
Ontop: a framework for a virtual approach of OBDQ

- An open source system for querying relational data sources through an ontology using SPARQL
  - support SPARQL 1.0 (BGP queries, i.e., conjunctive queries)

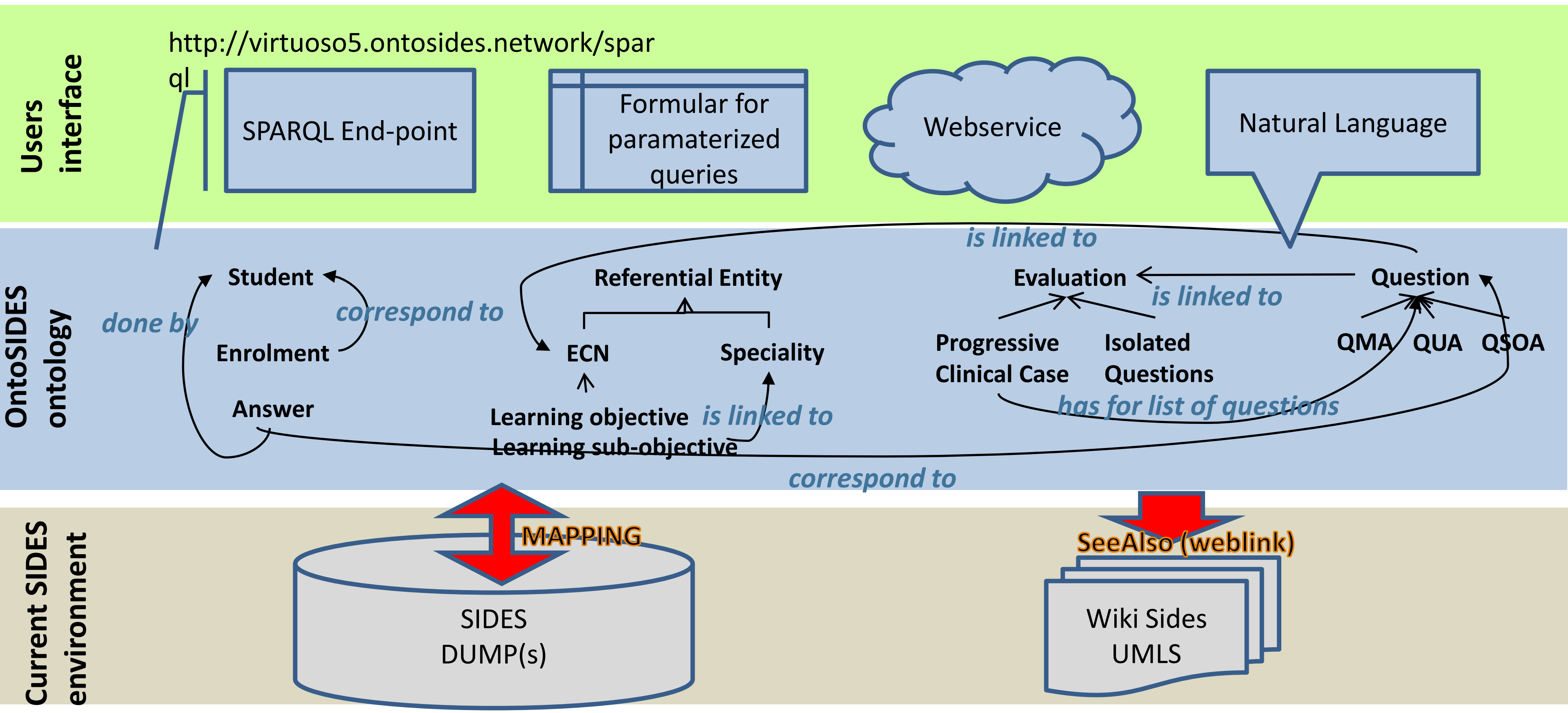


D. Calvanese, B. Cogrel, E. G. Kalayci, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, and G. Xiao. ***OBDA with the ontop framework.*** *In 23rd Italian Symposium on Advanced Database Systems, SEBD 2015, Gaeta, Italy, June 14-17, 2015., pages 296–303, 2015*

An architecture for a materialized approach of OBDQ



SIDES 3.0: AI-driven Education in Medicine



ANR-16-DUNE-0002

OntoSIDES knowledge graph

- The OBDA layer of SIDES 3.0
  - describes **training** and **assessments activities** performed by more than **145,000 students** in Medicine **over almost 6 years**
    - exams and training tests are made of **multiple choices questions**
    - students **answers** are described at the granularity of **time-stamped clicks of answers** done by students for choosing among the proposals of answers (correct or distractors) associated to questions

$\Rightarrow$6,5 billions triples with almost 400 millions clicks coming from the answers of students to almost 1,4 million questions.

Knowledge Graphs

- Modern knowledge representation formalism based on RDF data model
  - more flexible than the relational model
    - ✓ No strict separation between schema and instances
  - adapted to data/knowledge sharing between distributed data sources over the Web
    - ✓ the basis of Linked Open Data and the Semantic Web
- a set of triples <subject, property, object/value>
  - subject, property and object are URIs (http Uniform Resource Identifiers)
  - **dereferencable URIs** (pointers to Web pages) versus **local URIs**
  - value is a literal (string, integer, date, boolean)

RDF modeling **multiple choice questions** in OntoSides

Q30986 has_for_textual_content **"Concernant la péritonite appendiculaire, donnez la ou les propositions exactes :"** ;
is_linked_to_the_medical_speciality **digestive_surgery**
has_for_proposal_of_answer **prop98552 [** has_for_textual_content **"les signes infectieux sont présents d'emblée » ;**
has_for_correction **« true »]**
prop98553 [has_for_textual_content
"il n'y a pas de défense abdominale ou de contracture" ;
has_for_correction **« false »]**
prop98604[has_for_textual_content
"elle peut se présenter comme une occlusion fébrile" ;
has_for_correction **« true»]**
prop98605[has_for_textual_content **"il n'y a pas de pneumopéritoine" ;**
has_for_correction **« true»]**
prop98606[has_for_textual_content **« le traitement est chirurgical" ;**
has_for_correction **« true»]**

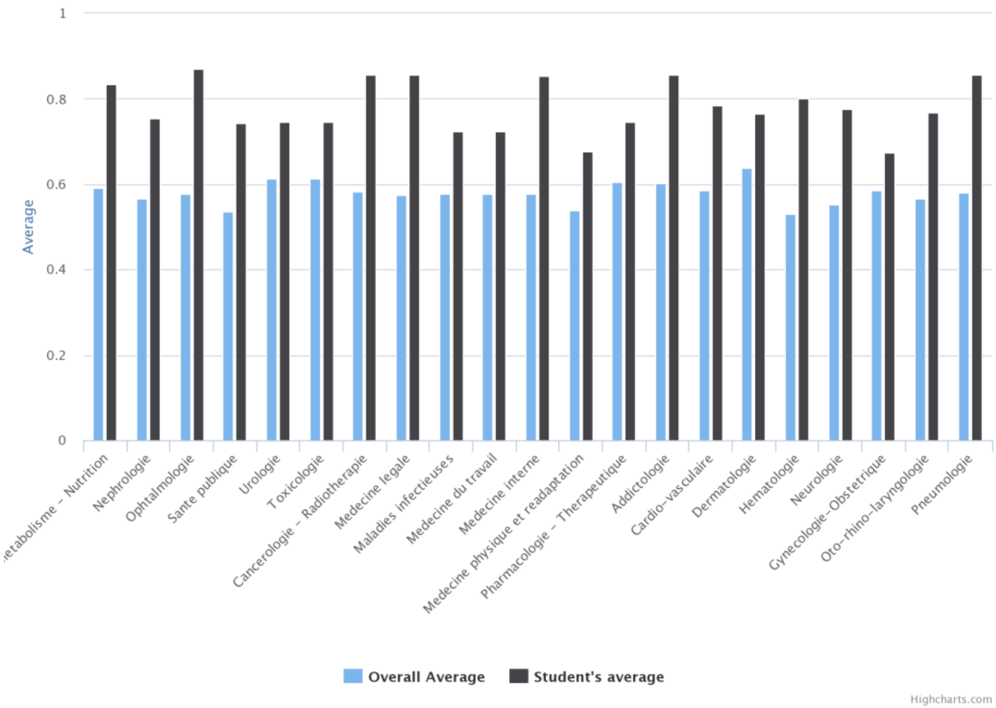
Tractable reasoning on knowledge graphs

- Simple Knowledge
  - RDFS + Datalog rules
  - OntoSides ontology:
    - 52 classes and 50 properties
    - 1400+ instances (medical specialties, official items of the ECN programme)
    - 12 rules
- Big Data:
  - associated with a powerful query language (SPARQL)
  - OntoSides KG:
    - 400 millions clicks of answer for 1,2 million multiple choice questions
    - 145 000 students

=> Explainable and Personalized Data Analytics

Illustration :

comparison of a given student's average results with average results of all students by medical specialty




```
SELECT ?specialty  ?globalAverage  ?studentAverage
WHERE  {
 {  SELECT ?specialty ( AVG(?result) AS ?globalAverage)
  WHERE { ?answer sides:has_for_result ?result .
              ?answer sides:done_by ?student .
              ?answer sides:correspond_to_a_question ?q .
              ?q sides:is_linked_to_the_medical_speciality ?specialty . }
    GROUP BY ?specialty } .
 { SELECT ?specialty (AVG(?result) AS ?studentAverage)
  WHERE { ?answer sides:has_for_result ?result .
              ?answer sides:done_by sides:etu12402 .
              ?answer sides:correspond_to_a_question ?q .
              ?q sides:is_linked_to_the_medical_speciality ?specialty .}
  GROUP BY ?specialty} .
}
```

Aggregated queries (SPARQL 1.1)

- not supported by query rewriting approaches
- **requires data completeness**

Knowledge graph completion

- A problem of increasing interest for which several **supervised** and **unsupervised** techniques have been investigated
  - can be modeled as a **classification** or a **matching** problem
    - depending on the available textual description of the target entities and the availability of training data
- Automatic inference of missing facts from existing ones
  - between **questions** and **medical specialties** or **learning objectives**
    - **13% questions** have been explicitly **linked** by their authors **to medical specialties**
    - **12% questions linked to learning objectives** (items listed in the French national medical reference program)

Experimental results for classification

Dataset	Classifier	Hits@1	Hits@2	Hits@5	Hits@10	MRR
Dataset1	Naive Bayes classifier	73.8%	83.1%	84.2%	84.3%	79.9%
	Maximum Entropy classifier	75.1%	88.9%	95.4%	96.8%	84%
	CNN classifier	76.4%	89.4%	96.3%	98.5%	85.2%
Dataset2	Naive Bayes classifier	56.4%	64.8%	67.8%	67.9%	61.5%
	Maximum Entropy classifier	68%	81.7%	90.6%	93.6%	78.2%
	CNN classifier	66.4%	78.9%	88.8%	93.4%	76%

Dataset1: 149145 questions -> 31 medical specialties
Dataset2: 144708 questions -> 362 learning objectives
Hits@k (Precision at k): average number of times a correct result appears in the top-k answers
MRR (Mean Reciprocal Rank): average of the rank inverses of the first correct answer

- **All the classifiers perform better on Dataset1 than on Dataset2**
  - the number of classes for Dataset2 is more than 10 times the number of classes for Dataset1 for almost the same number of items to classify
- **Naive Bayes outperformed by Maximum Entropy and CNN**
- Maximum Entropy gives slightly better results than CNN classifier on Dataset2
- **In more than 96% (93%) of the cases, the correct medical specialties (learning objectives) are returned in the top-10 answers**

Focus of the remaining of my talk

Focus 1

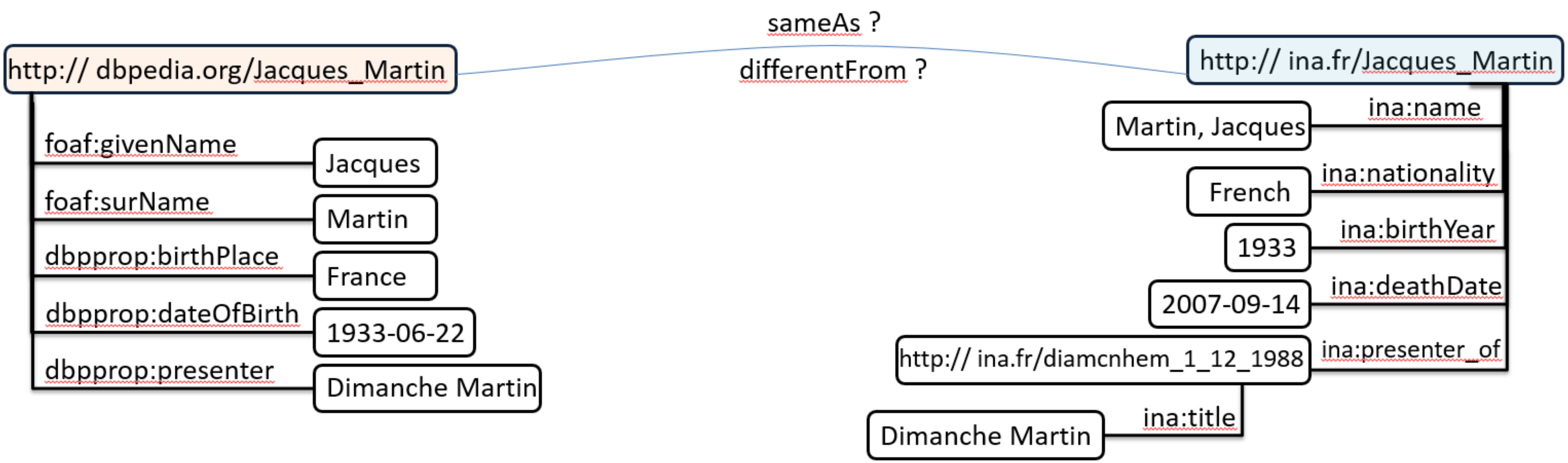
Ontology-based reasoning for data integration

Focus 2

Rule-based reasoning for data linkage

Data linkage

- Deciding whether two URIs refer to the same real-world entity across data sources



- Crucial task for data fusion and enrichment

- A hot topic in Linked Open Data

- Also related to data privacy

Existing approaches

- **Numerical methods based** on aggregating similarities between values of some relevant properties
  - Specification through linkage rules (e.g., in Silk and LIMES) of:
    1. the properties to consider within the descriptions of individuals,
    2. the similarity functions to use for comparing their respective values,
    3. the functions for aggregating these similarity values
  - Linkage rules: defined manually or learned automatically
  - **Main weakness: no formal semantics and no rule chaining**
- **Symbolic methods** based on logical rules **equipped with full reasoning**
  - Translation of schema constraints into logical rules
  - Logical inference of **sameAs facts**
  - **Main weakness: not robust to incomplete and/or noisy data**

    $\Rightarrow$100% precision but risk of low recall

Probabilistic Datalog (*) revisited to reason with uncertain data and rules

- A simple extension of Datalog in which rules and facts are associated with **symbolic probabilistic events**
- Logical inference and probability computation are separated
  - **Step 1 (ProbFR)** : computation for each inferred fact of its **provenance** (the **boolean combination** of all the events associated with the input facts and rules involved in its derivation)
    - exponential in the worst-case
    - by-passed by a practical bound on the number of conjuncts in the provenances and a priority given to the most probable rules and facts
  - **Step 2:** computation of the probabilities of the inferred facts
    - from their provenances in which each event of input facts and rules is assigned **a probabilistic weight**
    - based on independence and disjointness assumptions to make it feasible

(*) N. Fuhr, Probabilistic models in information retrieval, The Computer Journal, 1992

Illustrative Example

Rules: uncertain rules are in red, certain rules are in blue

$r_1 : (?x\ sameName\ ?y) \Rightarrow (?x\ sameAs\ ?y)$
$r_2 : (?x\ sameName\ ?y), (?x\ sameBirthDate\ ?y) \Rightarrow (?x\ sameAs\ ?y)$
$r_3 : (?x\ marriedTo\ ?z), (?y\ marriedTo\ ?z) \Rightarrow (?x\ sameAs\ ?y)$
$r_4 : (?x\ sameAs\ ?z), (?z\ sameAs\ ?y) \Rightarrow (?x\ sameAs\ ?y)$

Facts: uncertain facts are in red, certain facts are in blue

$f_1 : (i_1\ sameName\ i_2)$ $f_2 : (i_1\ sameBirthDate\ i_2)$ $f_3 : (i_2\ marriedTo\ i_3)$

$f_4 : (i_4\ marriedTo\ i_3)$ $f_5 : (i_2\ sameName\ i_4)$

Provenance of inferred facts

Inferred facts	Provenance	Uncertainty Provenance
$(i_2\ sameAs\ i_4)$	$(e(r_1) \wedge e(f_5)) \vee (e(r_3) \wedge e(f_3) \wedge e(f_4))$	$\top$
$(i_1\ sameAs\ i_2)$	$(e(r_1) \wedge e(f_1)) \vee (e(r_2) \wedge e(f_1) \wedge e(f_2))$	$e(r_2) \wedge e(f_1)$
$(i_1\ sameAs\ i_4)$	$e(r_4) \wedge Prov((i_1\ sameAs\ i_2))$ $\wedge Prov((i_2\ sameAs\ i_4))$	$e(r_2) \wedge e(f_1)$

Illustrative Example (cont.)

Rules: uncertain rules are in red, certain rules are in blue

$r_1 : (?x\ sameName\ ?y) \Rightarrow (?x\ sameAs\ ?y)$

$r_2 : (?x\ sameName\ ?y), (?x\ sameBirthDate\ ?y) \Rightarrow (?x\ sameAs\ ?y)$

$r_3 : (?x\ marriedTo\ ?z), (?y\ marriedTo\ ?z) \Rightarrow (?x\ sameAs\ ?y)$

$r_4 : (?x\ sameAs\ ?z), (?z\ sameAs\ ?y) \Rightarrow (?x\ sameAs\ ?y)$

Facts: uncertain facts are in red, certain facts are in blue

$f_1 : (i_1\ sameName\ i_2)$ $f_2 : (i_1\ sameBirthDate\ i_2)$ $f_3 : (i_2\ marriedTo\ i_3)$

$f_4 : (i_4\ marriedTo\ i_3)$ $f_5 : (i_2\ sameName\ i_4)$

Computation of the inferred facts probabilities

Inferred facts	Uncertainty Provenance	Probability
$(i_2\ sameAs\ i_4)$	$\top$	1
$(i_1\ sameAs\ i_2)$	$e(r_2) \wedge e(f_1)$	$Pr(e(r_2)) \times Pr(e(f_1))$
$(i_1\ sameAs\ i_4)$	$e(r_2) \wedge e(f_1)$	$Pr(e(r_2)) \times Pr(e(f_1))$

Illustrative Example (cont.)

Rules: uncertain rules are in red, certain rules are in blue

$r_1 : (?x\ sameName\ ?y) \Rightarrow (?x\ sameAs\ ?y)$

$r_2 : (?x\ sameName\ ?y), (?x\ sameBirthDate\ ?y) \Rightarrow (?x\ sameAs\ ?y)$

$r_3 : (?x\ marriedTo\ ?z), (?y\ marriedTo\ ?z) \Rightarrow (?x\ sameAs\ ?y)$

$r_4 : (?x\ sameAs\ ?z), (?z\ sameAs\ ?y) \Rightarrow (?x\ sameAs\ ?y)$

Facts: uncertain facts are in red, certain facts are in blue

$f_1 : (i_1\ sameName\ i_2)$ $f_2 : (i_1\ sameBirthDate\ i_2)$ $f_3 : (i_2\ marriedTo\ i_3)$

$f_4 : (i_4\ marriedTo\ i_3)$ $f_5 : (i_2\ sameName\ i_4)$

Computation of the inferred facts probabilities

Inferred facts	Uncertainty Provenance	Probability
$(i_2\ sameAs\ i_4)$	$\top$	1
$(i_1\ sameAs\ i_2)$	$e(r_2) \wedge e(f_1)$	0.8×0.9
$(i_1\ sameAs\ i_4)$	$e(r_2) \wedge e(f_1)$	0.8×0.9

Experiments: interlinking DBpedia and MusicBrainz

Size and number of entities in the two datasets

Class	DBpedia	MusicBrainz
Person	1,445,773	385,662
Band	75,661	197,744
Song	52,565	448,835
Album	123,374	1,230,731
Number of RDF triples	73 millions	112 millions

86 rules from which 50 are certain and 36 are uncertain

ID	Rules
sameAsBirthDate	$(?x$:solrPSimilarName $?l)$, $(?y$ skos:myLabel $?l)$, $(?x$ dbo:birthDate $?date)$, $(?y$ mb:beginDateC $?date)$ $\Rightarrow (?x$:sameAsPerson $?y)$
sameAsMemberOfBand	$(?x$:solrPSimilarName $?l)$, $(?y$ skos:myLabel $?l)$, $(?y$ mb:member_of_band $?gr2)$, $(?gr2$ skos:myLabel $?lg)$, $(?gr1$ dbp:members $?x)$, $(?gr1$:solrGrSimilarName $?lg)$ $\Rightarrow (?x$:sameAsPerson $?y)$

Experimental results

Gain of rule chaining

43,923 links not discovered by Silk among the **144,467 sameAs links discovered by ProbFR** between DBpedia and MusicBrainz

Gain of using uncertain rules for improving recall without losing much in precision (precision and recall estimated on samples)

	DBpedia and MusicBrainz					
	Only certain rules			All rules		
	P	R	F	P	R	F
Person	1.00	**0.08**	0.15	1.00	**0.80**	0.89
Band	1.00	**0.12**	0.21	0.94	**0.84**	0.89
Song	-	-	-	0.96	0.74	0.84
Album	-	-	-	1.00	0.53	0.69

Gain of exploiting probabilities to filter out wrong sameAs links

	P	R	F
$\text{Band}_{\geqslant 0.90}$	1.00	0.80	0.89
$\text{Song}_{\geqslant 0.60}$	1.00	0.54	0.72

Lessons learnt and perspectives

Probabilistic Datalog: a good trade-off for reasoning with uncertainty in Linked Data

Some restrictions compared to general probabilistic logical frameworks (e.g., Markov Logic)

- uncertain formulas restricted to Horn rules and ground facts
- probabilities computed for inferred facts only

Better scalability and more transparency

- explanations on probabilistic inference for end-users
- useful traces for experts to set-up the rules probabilities

Future work

ANR ELKER project

- A method to set up automatically the threshold for filtering the probabilistic sameAs facts to be retained
- A backward-reasoning algorithm on probabilistic rules for importing on demand useful data from external sources

Concluding message

- Semantic Web standards, data and applications are there, due to the simplicity and flexibility of the RDF data model
- Promising applications are emerging for which reasoning on data is central
  - Fact checking
  - Interactive and personalized data exploration and analytics
- Many challenges remain
  - to handle at large scale incomplete and uncertain data

Combining numerical and symbolic AI is hard …

but worthwhile to investigate more deeply

for robustness and explainability

Joint work with many persons

Mustafa Al Bakri, Mohannad Almasri, Manuel Atencia, Shadi Baghernezhad, Jérôme David, Loic Druette (Univ. Lyon) , Fabrice Jouanot, Cyril Labbé, Steffen Lalande (INA), Behrooz Omidvar, Olivier Palombi (LADAF, LJK), Adam Sanchez, Federico Ulliana (LIRMM),…

THANKS