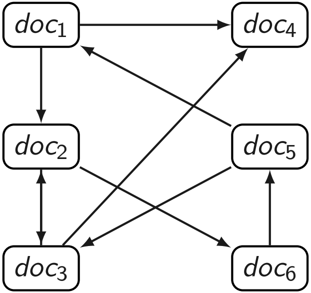
From the Web of Documents to the Web of Data

Marcelo Arenas

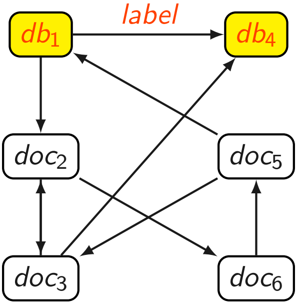
Pontificia Universidad Católica de Chile

ICWBD 2014, Goa, India

The Web of documents



But things have changed . . .



But things have changed . . .

KEGG
Kyoto Encyclopedia of Genes and Genomes
DBpedia
Linked GeoData
PubMed
United States™ Census 2010
OPEN LIBRARY
CiteSeer.IST
Scientific Literature Digital Library

A new opportunity: more structured queries

A new opportunity: more structured queries

Who is the most cited researcher in area X in country Y?

A new opportunity: more structured queries

Who is the most cited researcher in area X in country Y?

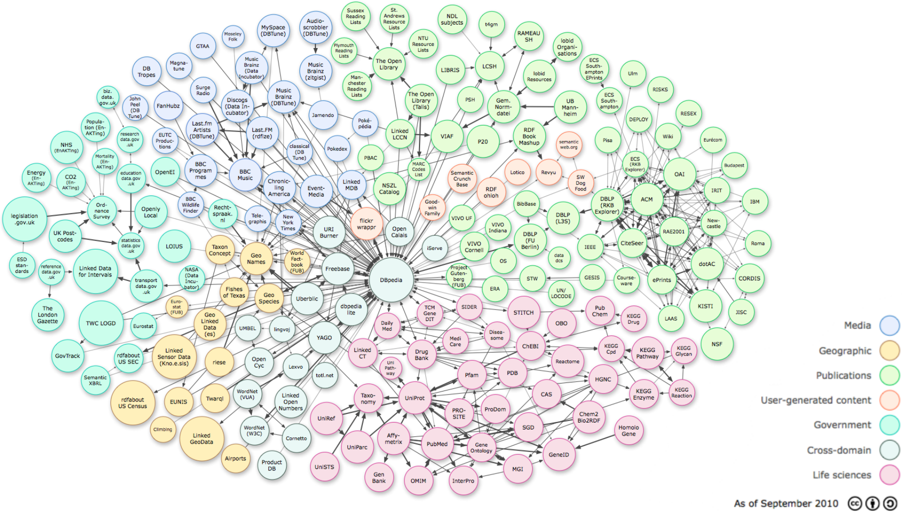
The information is on the Web, the process can be automatized:

- *Semantics:* Interpret terms "most cited", "area X", ...
- *Distribution:* Gather the needed pieces of information
- *Heterogeneity:* Integrate heterogeneous pieces of information

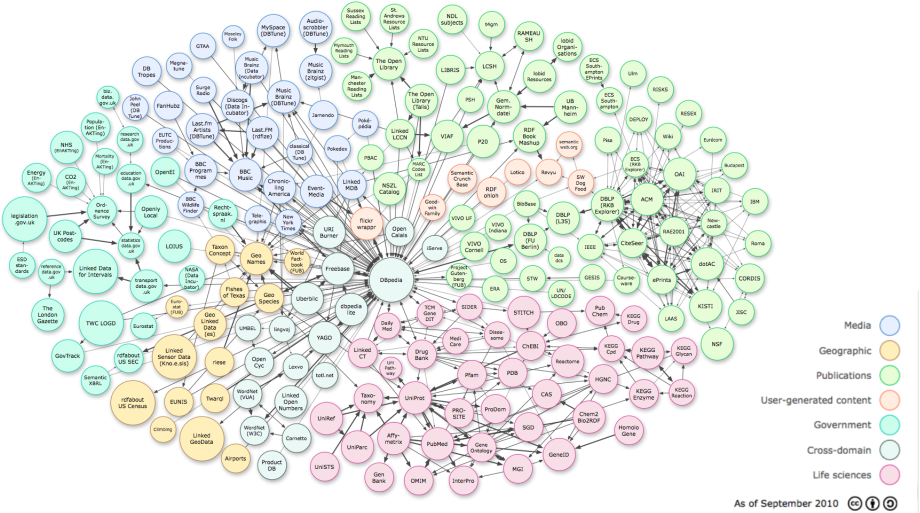
We encounter similar challenges all around the Web

KEGG
Kyoto Encyclopedia of Genes and Genomes
DBpedia
Linked GeoData
PubMed
United States™ Census 2010
OPEN LIBRARY
CiteSeer.IST
Scientific Literature Digital Library

We encounter similar challenges all around the Web



We encounter similar challenges all around the Web



How to query distributed and heterogeneous semantic data?

Data sources keep getting bigger and bigger

Some of the known techniques are falling short.

We need to develop foundations and algorithms
to take full advantage of the semantics of data at Web scale.

The Semantic Web

Semantic Web

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

[Tim Berners-Lee et al. 2001.]

Specific goals:

- Build a description language with standard semantics
  - Make semantics machine-processable and understandable
- Incorporate logical infrastructure to reason about resources
- W3C proposals: Resource Description Framework (RDF) and SPARQL

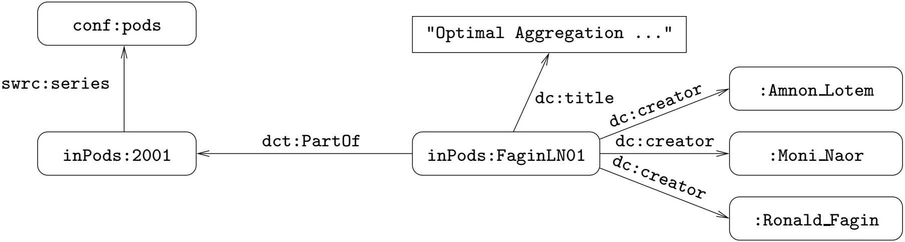
RDF in a nutshell

RDF is the framework proposed by the W3C to represent information in the Web:

- URI vocabulary
  - A URI is an atomic piece of data, and it identifies an abstract resource
- Syntax based on directed labeled graphs
  - URIs are used as node labels and edge labels
- Schema definition language (RDFS): Define new vocabulary
  - Typing, inheritance of classes and properties, . . .
- Formal semantics

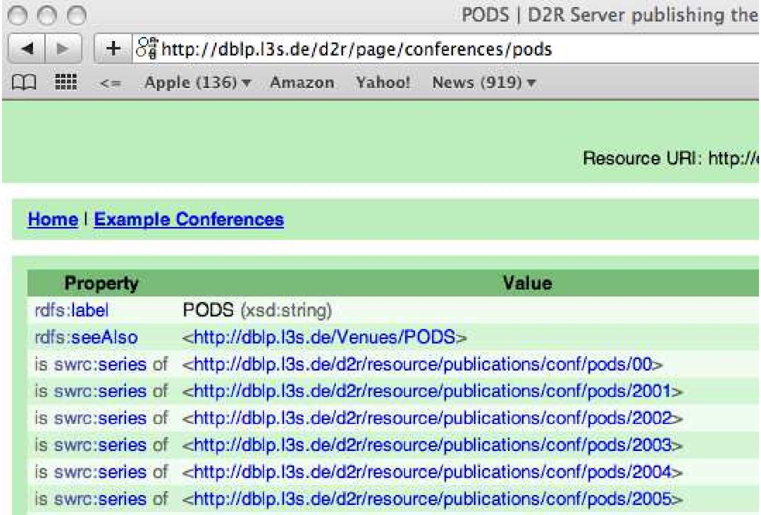
An example of an RDF graph: DBLP

```
      : <http://dblp.l3s.de/d2r/resource/authors/>
  conf: <http://dblp.l3s.de/d2r/resource/conferences/>
inPods: <http://dblp.l3s.de/d2r/resource/publications/conf/pods/>
  swrc: <http://swrc.ontoware.org/ontology#>
    dc: <http://purl.org/dc/elements/1.1/>
   dct: <http://purl.org/dc/terms/>
```



An example of a URI

`http://dblp.l3s.de/d2r/resource/conferences/pods`



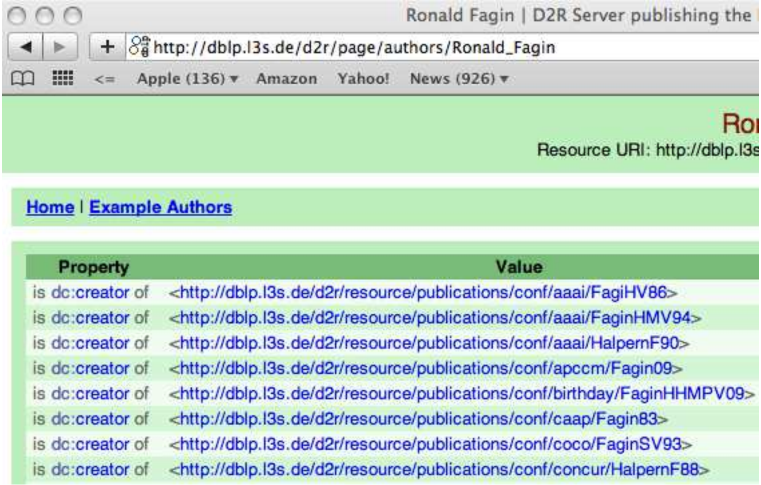
PODS | D2R Server publishing the
http://dblp.l3s.de/d2r/page/conferences/pods
<= Apple (136) ▾ Amazon Yahoo! News (919) ▾
Home | Example Conferences

Resource URI: http://

Property	Value
rdfs:label	PODS (xsd:string)
rdfs:seeAlso	<http://dblp.l3s.de/Venues/PODS>
is swrc:series of	<http://dblp.l3s.de/d2r/resource/publications/conf/pods/00>
is swrc:series of	<http://dblp.l3s.de/d2r/resource/publications/conf/pods/2001>
is swrc:series of	<http://dblp.l3s.de/d2r/resource/publications/conf/pods/2002>
is swrc:series of	<http://dblp.l3s.de/d2r/resource/publications/conf/pods/2003>
is swrc:series of	<http://dblp.l3s.de/d2r/resource/publications/conf/pods/2004>
is swrc:series of	<http://dblp.l3s.de/d2r/resource/publications/conf/pods/2005>

URI can be used for any abstract resource

`http://dblp.l3s.de/d2r/page/authors/Ronald_Fagin`



Ronald Fagin | D2R Server publishing the

http://dblp.l3s.de/d2r/page/authors/Ronald_Fagin

Ro

Resource URI: http://dblp.l3s

Home | Example Authors

Property	Value
is dc:creator of	<http://dblp.l3s.de/d2r/resource/publications/conf/aaai/FagiHV86>
is dc:creator of	<http://dblp.l3s.de/d2r/resource/publications/conf/aaai/FaginHMV94>
is dc:creator of	<http://dblp.l3s.de/d2r/resource/publications/conf/aaai/HalpernF90>
is dc:creator of	<http://dblp.l3s.de/d2r/resource/publications/conf/apccm/Fagin09>
is dc:creator of	<http://dblp.l3s.de/d2r/resource/publications/conf/birthday/FaginHHMPV09>
is dc:creator of	<http://dblp.l3s.de/d2r/resource/publications/conf/caap/Fagin83>
is dc:creator of	<http://dblp.l3s.de/d2r/resource/publications/conf/coco/FaginSV93>
is dc:creator of	<http://dblp.l3s.de/d2r/resource/publications/conf/concur/HalpernF88>

Querying RDF

Why is this an interesting problem? Why is it challenging?

- RDF graphs can be interconnected
  - URIs should be dereferenceable
- Semantics of RDF is open world
  - RDF graphs are inherently incomplete
  - The possibility of adding optional information if present is an important feature
- Vocabulary with predefined semantics
- . . .

Querying RDF: SPARQL

- SPARQL is the W3C recommendation query language for RDF (January 2008).
  - SPARQL is a recursive acronym that stands for *SPARQL Protocol and RDF Query Language*
- SPARQL is a graph-matching query language.
- A SPARQL query consists of three parts:
  - Pattern matching: optional, union, filtering, . . .
  - Solution modifiers: projection, distinct, order, limit, offset, . . .
  - Output part: construction of new triples, ...

SPARQL in a nutshell

SPARQL in a nutshell

```
SELECT ?Author
```

SPARQL in a nutshell

```
SELECT ?Author
WHERE
{



}
```

SPARQL in a nutshell

```
SELECT ?Author
WHERE
{
  ?Paper      dc:creator     ?Author .


}
```

SPARQL in a nutshell

```
SELECT ?Author
WHERE
{
  ?Paper      dc:creator     ?Author .
  ?Paper      dct:PartOf     ?Conf .

}
```

SPARQL in a nutshell

```
SELECT ?Author
WHERE
{
  ?Paper       dc:creator     ?Author .
  ?Paper       dct:PartOf     ?Conf .
  ?Conf        swrc:series    conf:pods .
}
```

SPARQL in a nutshell

```
SELECT ?Author
WHERE
{
  ?Paper      dc:creator     ?Author .
  ?Paper      dct:PartOf     ?Conf .
  ?Conf       swrc:series    conf:pods .
}
```

A SPARQL query consists of a:

SPARQL in a nutshell

```
SELECT ?Author
WHERE
{
  ?Paper      dc:creator     ?Author .
  ?Paper      dct:PartOf     ?Conf .
  ?Conf       swrc:series    conf:pods .
}
```

A SPARQL query consists of a:

Body: Pattern matching expression

SPARQL in a nutshell

```
SELECT ?Author
WHERE
{
  ?Paper      dc:creator     ?Author .
  ?Paper      dct:PartOf     ?Conf .
  ?Conf       swrc:series    conf:pods .
}
```

A SPARQL query consists of a:

- Body: Pattern matching expression
- Head: Processing of the variables

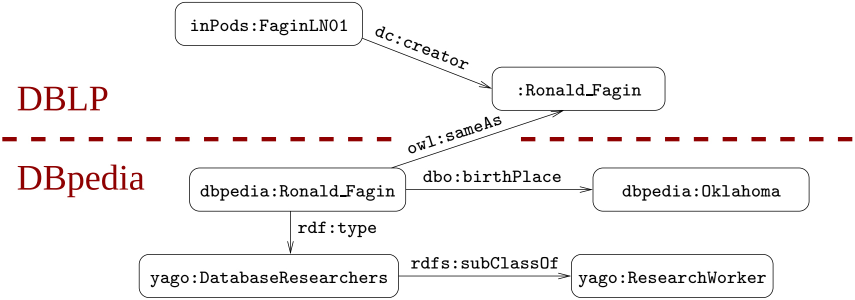
What are the challenges in implementing SPARQL?

SPARQL has to take into account the distinctive features of RDF:

- Should be able to extract information from interconnected RDF graphs
- Should be consistent with the open-world semantics of RDF
  - Should offer the possibility of adding optional information if present
- Should be able to properly interpret RDF graphs with a vocabulary with predefined semantics

Extracting information from interconnected RDF graphs

```
       : <http://dblp.l3s.de/d2r/resource/authors/>
dbpedia: <http://dbpedia.org/resource/>
    rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
   rdfs: <http://www.w3.org/2000/01/rdf-schema#>
    owl: <http://www.w3.org/2002/07/owl#>
   yago: <http://dbpedia.org/class/yago>
    dbo: <http://dbpedia.org/ontology/>
```



DBLP

DBpedia

inPods:FaginLN01

dc:creator

:Ronald_Fagin

owl:sameAs

dbpedia:Ronald_Fagin

dbo:birthPlace

dbpedia:Oklahoma

rdf:type

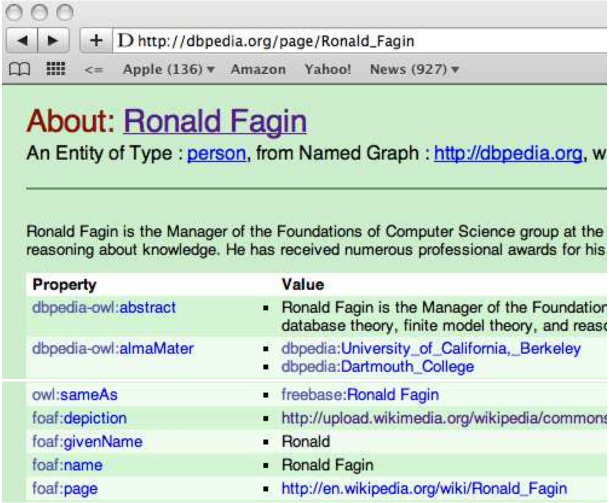
yago:DatabaseResearchers

rdfs:subClassOf

yago:ResearchWorker

Dereferenceable URIs are the glue

```
http://dbpedia.org/resource/Ronald_Fagin
```



About: Ronald Fagin

An Entity of Type : person, from Named Graph : http://dbpedia.org, w

Ronald Fagin is the Manager of the Foundations of Computer Science group at the
reasoning about knowledge. He has received numerous professional awards for his

Property	Value
dbpedia-owl:abstract	▪ Ronald Fagin is the Manager of the Foundatior database theory, finite model theory, and reasc
dbpedia-owl:almaMater	▪ dbpedia:University_of_California,_Berkeley ▪ dbpedia:Dartmouth_College
owl:sameAs	▪ freebase:Ronald Fagin
foaf:depiction	▪ http://upload.wikimedia.org/wikipedia/commons
foaf:givenName	▪ Ronald
foaf:name	▪ Ronald Fagin
foaf:page	▪ http://en.wikipedia.org/wiki/Ronald_Fagin

Querying interconnected RDF graphs

Retrieve the authors that have published in PODS and were born in Oklahoma:

```
SELECT ?Author
WHERE
{
  ?Paper      dc:creator     ?Author .
  ?Paper      dct:PartOf     ?Conf .
  ?Conf       swrc:series    conf:pods .
  ?Person     owl:sameAs     ?Author .
  ?Person     dbo:birthPlace dbpedia:Oklahoma .
}
```

Retrieving optional information

Retrieve the authors that have published in PODS, and their Web pages if this information is available:

```
SELECT ?Author ?WebPage
WHERE
{
  ?Paper      dc:creator     ?Author .
  ?Paper      dct:PartOf     ?Conf .
  ?Conf       swrc:series    conf:pods .
  OPTIONAL { ?Author  foaf:homePage  ?WebPage . }
}
```

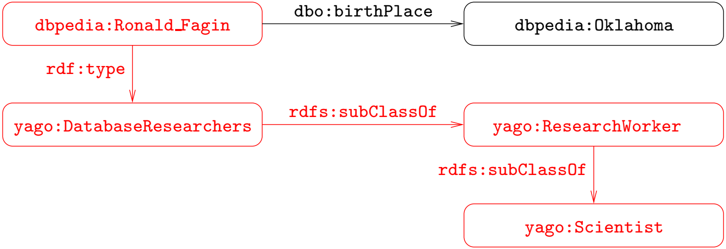
Taking into account vocabularies with predefined semantics

Retrieve the scientists that were born in Oklahoma and that have published in PODS:

```
SELECT ?Author
WHERE
{
  ?Author     rdf:type       yago:Scientist .
  ?Author     dbo:birthPlace dbpedia:Oklahoma .
  ?Paper      dc:creator     ?Author .
  ?Paper      dct:PartOf     ?Conf .
  ?Conf       swrc:series    conf:pods .
}
```

Taking into account vocabularies with predefined semantics

Retrieve the scientists that were born in Oklahoma and that have published in PODS:



The Center for Semantic Web Research

The Center for Semantic Web Research

(funded by the Millennium Scientific Initiative)

Researchers

Director

Marcelo Arenas (PUC)	*semantic Web, database theory*

Deputy director

Pablo Barcelo (UChile)	*graph databases, database theory*

Researchers

Director

Marcelo Arenas (PUC)	*semantic Web, database theory*

Deputy director

Pablo Barcelo (UChile)	*graph databases, database theory*

Associate researchers

Jorge Perez (UChile)	*semantic Web, interoperability*
Juan Reutter (PUC)	*graph databases, interoperability*
Claudio Gutierrez (UChile)	*semantic Web, graph databases*

Critical mass of young researchers

Young researchers

Renzo Angles (UTalca)	*Semantic Web*
Carlos Buil-Aranda (PUC)	*Semantic Web*
Aidan Hogan (UChile)	*Linked data*
Barbara Poblete (UChile) & Yahoo!	*Social networks*
Cristián Riveros (PUC)	*Interoperability, automata*

Graduate students

6 PhD & 3 postdocs

Strong international connections

IBM Almaden & Watson

U. of Oxford

U. of Texas at Austin

Rice U.

Microsoft Research

U. of Edinburgh

Polytechnic U. of Madrid

TU Vienna

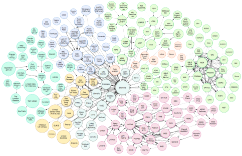
U. of Bolzano

Digital Research Enterprise Institute (DERI)

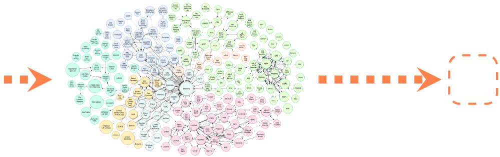
Yahoo! Research

Our Proposal

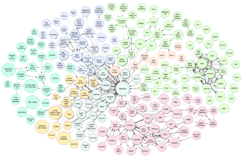
Who is the most cited researcher in area X in country Y?



Who is the most cited researcher in area X in country Y?



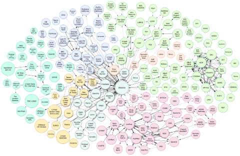
Who is the most cited researcher in area X in country Y?



SELECT researcher
FROM DataWeb ...

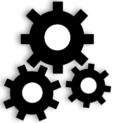
$$\forall x.\exists y.S(x) \rightarrow D(y)$$

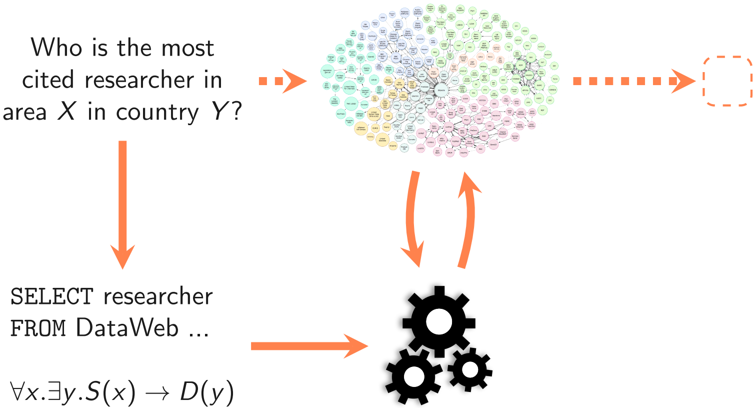
Who is the most cited researcher in area X in country Y?



`SELECT` researcher
`FROM` DataWeb ...

$\forall x.\exists y.S(x) \rightarrow D(y)$

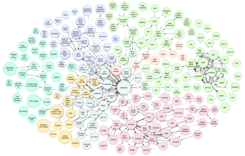




Who is the most cited researcher in area X in country Y?

SELECT researcher
FROM DataWeb ...

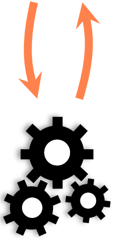
$\forall x.\exists y.S(x) \rightarrow D(y)$



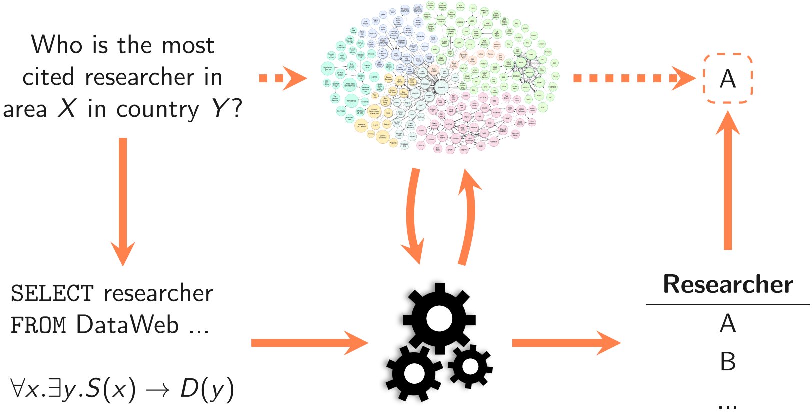
Who is the most cited researcher in area X in country Y?

SELECT researcher
FROM DataWeb ...

$$\forall x.\exists y.S(x) \rightarrow D(y)$$



Researcher
A
B
...



Who is the most cited researcher in area X in country Y?

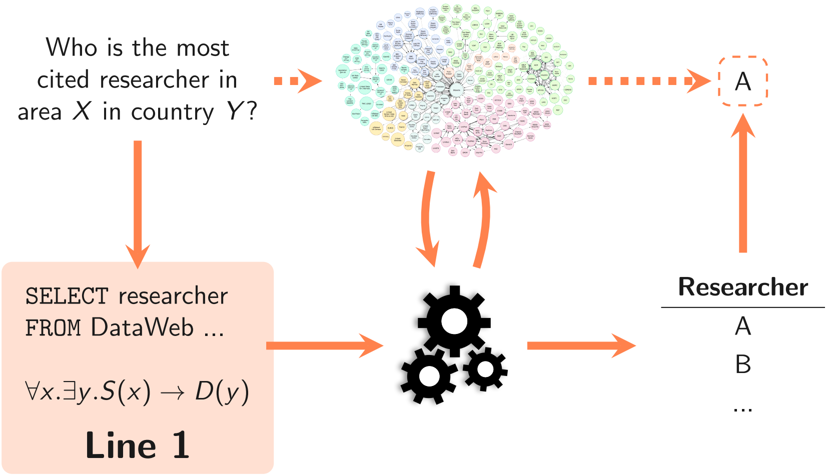
A

`SELECT` researcher
`FROM` DataWeb ...

$\forall x.\exists y.S(x) \rightarrow D(y)$

Researcher
A
B
...

Identifying the right language for querying semantic data at Web scale



Who is the most cited researcher in area X in country Y?

A

`SELECT` researcher
`FROM` DataWeb ...

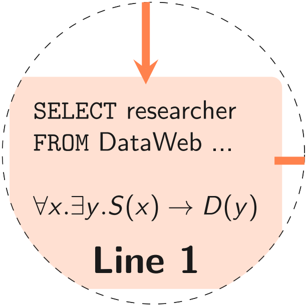
$\forall x.\exists y.S(x) \rightarrow D(y)$

Line 1

Researcher
A
B
...

Identifying the right language for querying semantic data at Web scale

- logic - finite model theory
- automata theory
- computational complexity

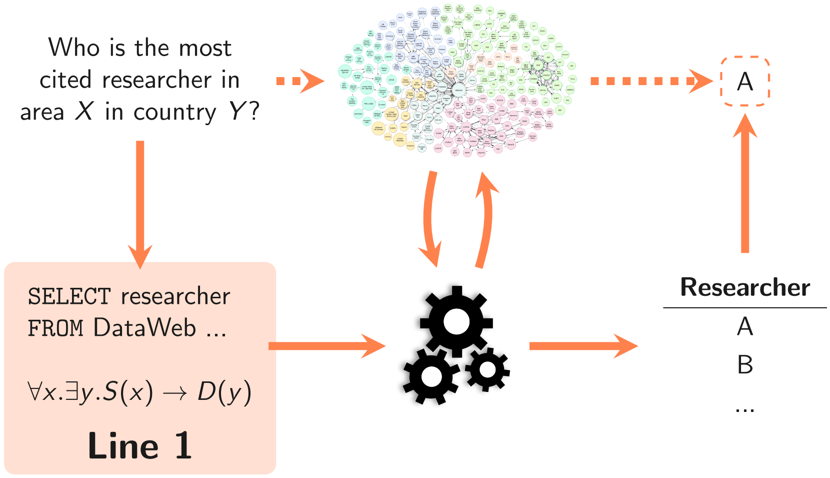


SELECT researcher
FROM DataWeb ...

$\forall x . \exists y . S(x) \rightarrow D(y)$

Line 1

Identifying the right language for querying semantic data at Web scale



Who is the most cited researcher in area X in country Y?

A

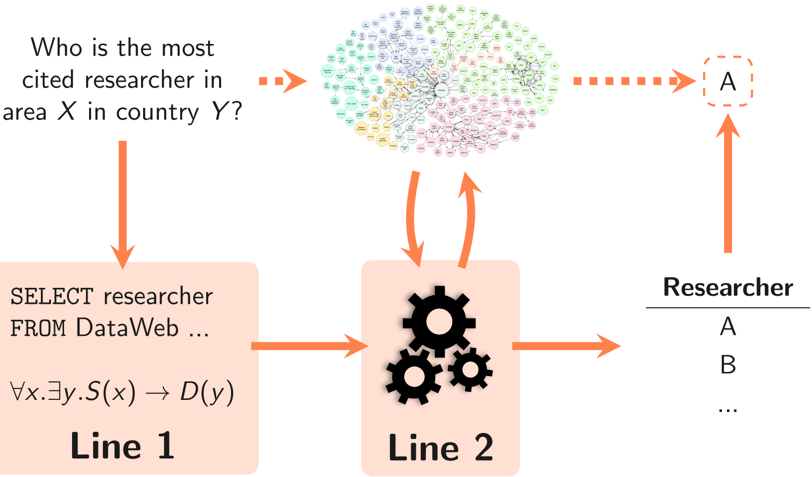
SELECT researcher
FROM DataWeb ...

$\forall x.\exists y.S(x) \rightarrow D(y)$

Line 1

Researcher
A
B
...

Obtaining relevant information, efficiently



Who is the most cited researcher in area X in country Y?

A

SELECT researcher FROM DataWeb ...

$\forall x.\exists y.S(x) \rightarrow D(y)$

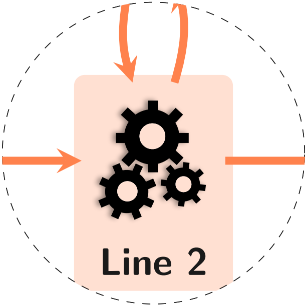
Line 1

Line 2

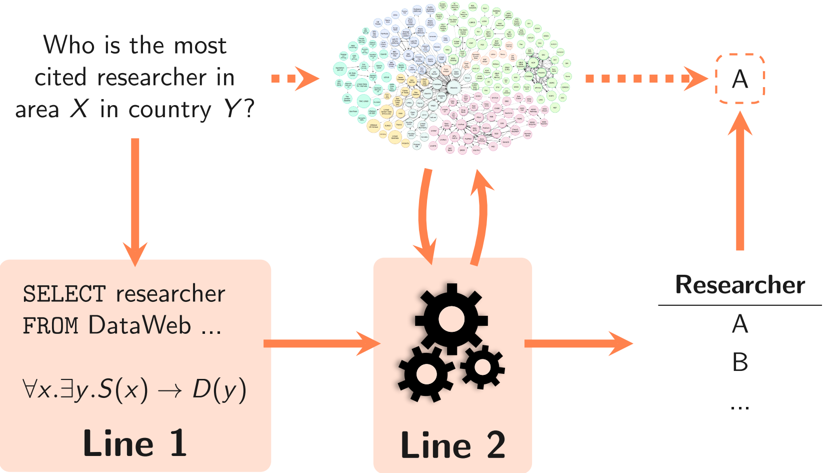
Researcher
A
B
...

Obtaining relevant information, efficiently

- data structures, indexing
- query optimization
- (hyper)tree decomposition
- computational complexity



Obtaining relevant information, efficiently



Who is the most cited researcher in area X in country Y?

A

SELECT researcher
FROM DataWeb ...

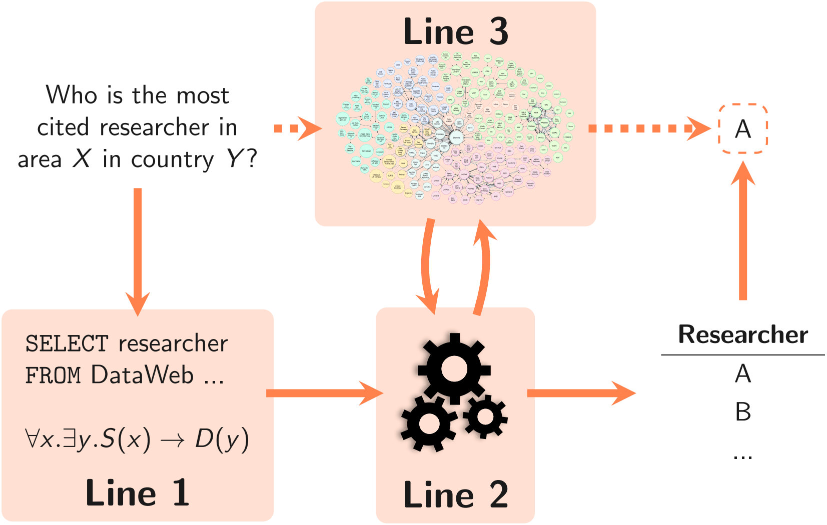
$\forall x . \exists y . S(x) \rightarrow D(y)$

Line 1

Line 2

Researcher
A
B
...

Taking advantage of the structure of the data



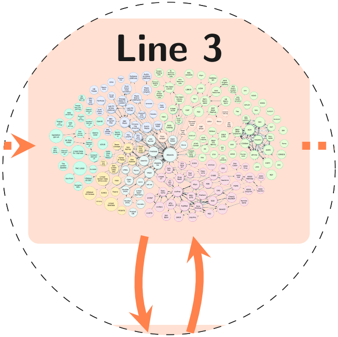
Line 3
Who is the most cited researcher in area X in country Y?
A
Line 1
Line 2

SELECT researcher
FROM DataWeb ...

$\forall x. \exists y. S(x) \rightarrow D(y)$

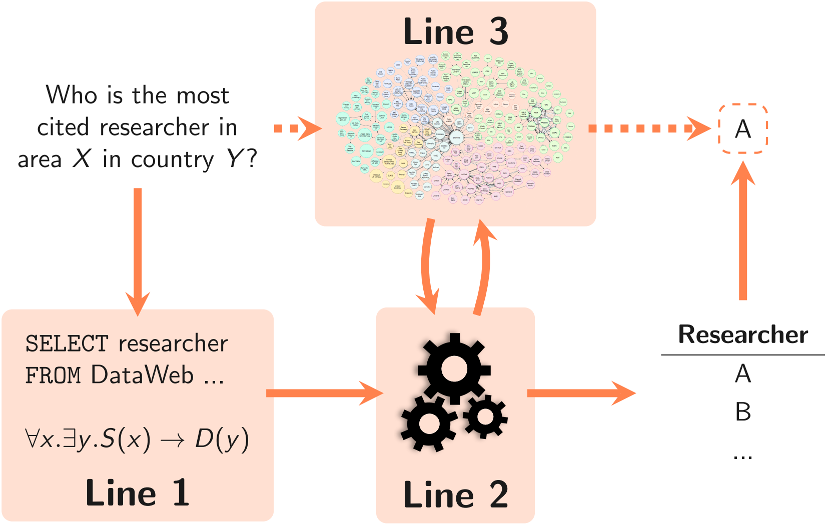
Researcher
A
B
...

Taking advantage of the structure of the data



- graph theory
- network theory
- data dependency theory

Taking advantage of the structure of the data



Who is the most cited researcher in area X in country Y?

```
SELECT researcher
FROM DataWeb ...
```

$\forall x.\exists y.S(x) \rightarrow D(y)$

Line 1

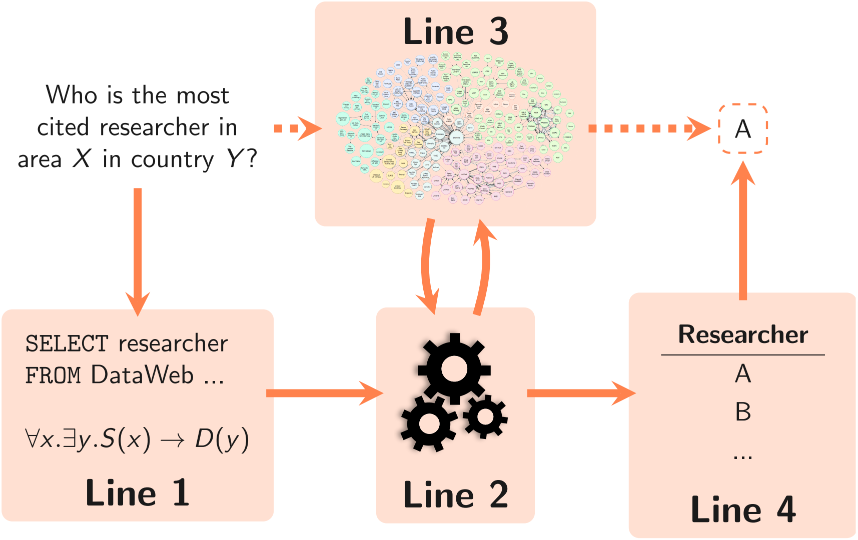
Line 2

Line 3

A

Researcher
A
B
...

Approximating answers when exact evaluation is infeasible



Who is the most cited researcher in area X in country Y?

`SELECT` researcher
`FROM` DataWeb …

$\forall x.\exists y.S(x) \rightarrow D(y)$

Line 1

Line 2

Line 3

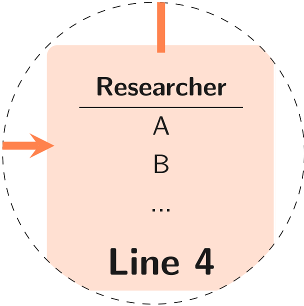
A

Researcher
A
B
…

Line 4

Approximating answers when exact evaluation is infeasible

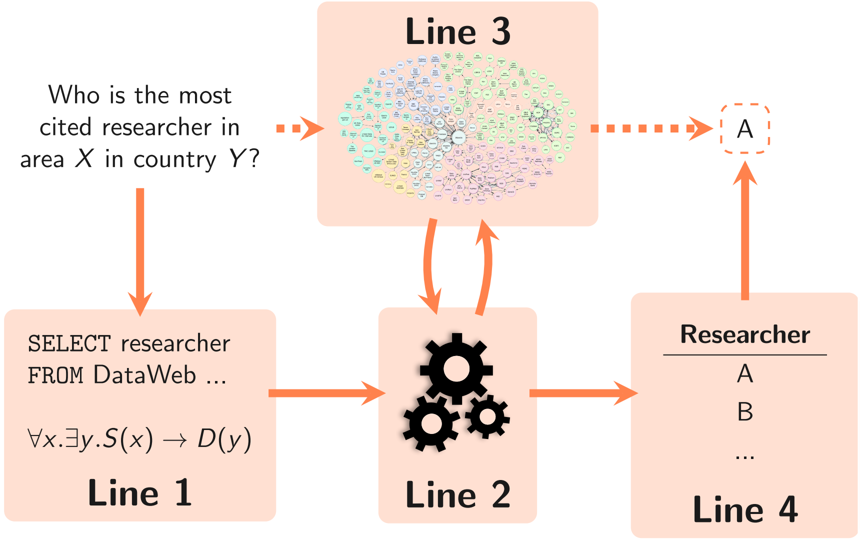
- graph theory
- approximation algorithms
- computational complexity



Researcher
A
B
...

Line 4

Approximating answers when exact evaluation is infeasible



Who is the most cited researcher in area X in country Y?

Line 3

A

```
SELECT researcher
FROM DataWeb ...
```

$\forall x.\exists y.S(x) \rightarrow D(y)$

Line 1

Line 2

Researcher
A
B
...

Line 4

(Some of) Our Projects

Publication of RDF Data

Translation of relational data into RDF

- Definition of a direct mapping, W3C standard: `http://www.w3.org/TR/rdb-direct-mapping`
- Study of fundamental notions such as information preservation, query preservation, . . . [SAM12]

Publication of RDF Data

Translation of relational data into RDF

- Definition of a direct mapping, W3C standard: `http://www.w3.org/TR/rdb-direct-mapping`
- Study of fundamental notions such as information preservation, query preservation, . . . [SAM12]

Generation of new RDF datasets from existing databases.

- Definition of a *declarative language* for HTML to RDF translation

Publication of RDF Data

Publication of public data

- Materialization of transparency law

  Design and (first) implementation of `http://www.gobiernotransparentechile.cl` and `http://datos.gob.cl`

- Scientific data from CONICYT: `http://datoscientificos.cl`

Study of the structure of RDF data

Study of the structuredness of RDF data [ADFKS14]

- Definition of a framework for specifying structuredness functions
- Study of the structure refinement problem

Study of the use of anonymous objects (blank nodes) in RDF data [HAMP]

- Reduction of the complexity of several reasoning problems

Storage of RDF data

Compression of RDF data [FMGPA13]

- HDT: defines header information, a dictionary, and the actual triples structure (`http://www.rdfhdt.org`)
- W3C submission: `http://www.w3.org/Submission/2011/03`

Study of Web query languages

Development of new benchmarks (`http://www.ldbc.eu`)

- To compare systems, and promote the development of new technologies

Study of the expressiveness of different query languages [AGP14,AP11,B13,BRV14,BLR14]

- What can and cannot be expressed in these languages
- What needs to be added to meet user requirements
- Study of new functionalities

Study of Web query languages

Development of query recommendation algorithms

- Definition of query extension and restriction
- Study of query logs (DBPedia, KEGG, ...)

Development of query evaluation algorithms

Indexing: Compression of RDF data [FMGPA13]

Incremental evaluation of SPARQL queries

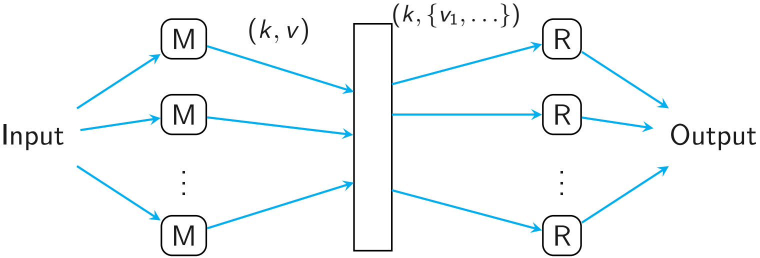
- Development of algorithms, heuristics and data structures to efficiently updating answers to queries, in highly dynamic environments

Optimization and distribution of SPARQL queries [BHUV13,BACP13]

- Use of SPARQL endpoints

Development of query evaluation algorithms: MapReduce

- MapReduce has been a popular framework for parallel programming
- Very simple and useful language for engineers/programmers
- Good for optimizing massive parallel architectures



MapReduce drawbacks

- Not all problems are parallelizable
- What are the classes of problems that are optimizable in this framework?

Development of query evaluation algorithms: MapReduce

- Understand the computational power of the MapReduce framework

- Identify features of SPARQL that can be computed efficiently in this framework

- Extend/restrict SPARQL to exploit massive parallel architectures

Development of query approximation algorithms

Development meaningful notions of approximation [BLR13]

- Yield to efficient query evaluation algorithms
- Useful in applications in which data is massive and finding interconnection patterns is important (e.g. social networks, crime-detection networks, etc)

Thank you!

Bibliography

[ADFKS14] M. Arenas, G. Diaz, A. Fokoue, A. Kementsietsidis and K. Srinivas: A Principled Approach to Bridging the Gap between Graph Data and their Schemas. To appear in 40th International Conference on Very Large Data Bases (VLDB 2014), 2014

[AGP14] M. Arenas, G. Gottlob and A. Pieris: Expressive Languages for Querying the Semantic Web. To appear in 33rd ACM Symposium on Principles of Database Systems (PODS 2014), 2014

[AP11] M. Arenas and J. Perez: Querying Semantic Web Data with SPARQL?. In 30th ACM Symposium on Principles of Database Systems (PODS 2011), pages 305–316, 2011

[B13] P. Barcelo: Querying graph databases. In 32nd ACM Symposium on Principles of Database Systems (PODS 2013), pages 175–188, 2013

Bibliography

[BRV14] P. Barcelo, M. Romero and M. Vardi: Does Query Evaluation Tractability Help Query Containment? In 33rd ACM Symposium on Principles of Database Systems (PODS 2014), 2014

[BLR14] P. Barcelo, L. Libkin and J. Reutter: Querying Regular Graph Patterns. Journal of the ACM, 61(1), 2014

BLR12] Barcelo, L. Libkin and M. Romero: Efficient approximations of conjunctive queries. In 31st ACM Symposium on Principles of Database Systems (PODS 2012), pages 249–260, 2012

[BACP13] C. Buil-Aranda, M. Arenas, O. Corcho, A. Polleres: Federating queries in SPARQL 1.1: Syntax, semantics and evaluation. Journal of Web Semantics 18(1): 1–17, 2013

Bibliography

[BHUV13] C. Buil-Aranda, A. Hogan, J. Umbrich, P.-Y. Vandenbussche: SPARQL Web-Querying Infrastructure: Ready for Action? International Semantic Web Conference (ISWC 2013), pages 277–293, 2013

[FMGPA13] J. D. Fernandez, M. A. Martinez-Prieto, C. Gutierrez, A. Polleres, M. Arias: Binary RDF Representation for Publication and Exchange (HDT). Journal of Web Semantics 19: 22–41, 2013

[HAMP] A. Hogan, M. Arenas, A. Mallea and A. Polleres: Everything You Always Wanted to Know About Blank Nodes". To appear in Journal of Web Semantics

[SAM12] J. F. Sequeda, M. Arenas and D. Miranker: On Directly Mapping Relational Databases to RDF and OWL. In 21st International Conference on World Wide Web (WWW 2012), pages 649–658, 2012