Question Answering over Linked Data

Challenges, Approaches & Trends

André Freitas, Christina Unger

Tutorial @ ESWC 2014

Goal of this Talk

- Understand the changes on the database landscape in the direction of more heterogeneous data scenarios.
- Understand how Question Answering (QA) fits into this new scenario.
- Give the fundamental pointers to develop your own QA system from the state-of-the-art.
- Coverage over depth.

Outline

- Motivation & Context
- Big Data, Linked Data & QA
- The Anatomy of a QA System
- QA over Linked Data
- Treo: Detailed Case Study
- Pythia: Detailed Case Study
- Evaluation of QA over Linked Data
- Do-it-yourself (DIY): Core Resources
- Trends
- Take-away Message

Motivation & Context

Why Question Answering?

- Humans are built-in with natural language communication capabilities.
- Very natural way for humans to communicate information needs.

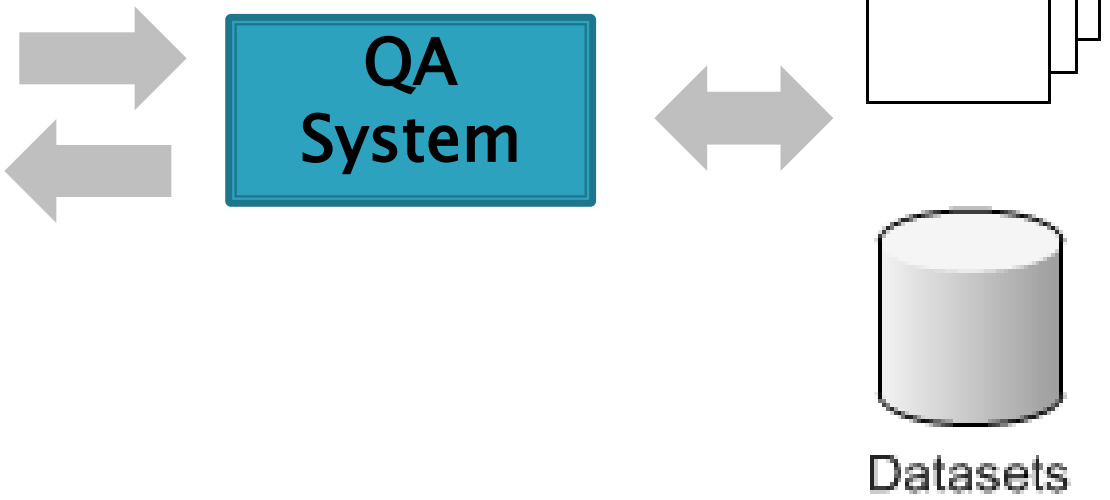
What is Question Answering?

- A research field on its own.
- Empirical bias: Focus on the development of automatic systems to answer questions.
- Multidisciplinary:
  - Natural Language Processing
  - Information Retrieval
  - Knowledge Representation
  - Databases
  - Linguistics
  - Artificial Intelligence
  - Software Engineering
  - ...

What is Question Answering?

Question: **Who is the daughter of Bill Clinton married to?**

Answer: **Marc Mezvinsky**



QA vs IR

- Keyword Search:
  - User still carries the major efforts in interpreting the data.
  - Satisfying information needs may depend on multiple search operations.
  - Answer-driven information access.
  - Input: Keyword search
    - Typically specification of simpler information needs.
  - Output: documents, data.
- QA:
  - Delegates more 'interpretation effort' to the machines.
  - Query-driven information access.
  - Input: natural language query
    - Specification of complex information needs.
  - Output: direct answer.

QA vs Databases

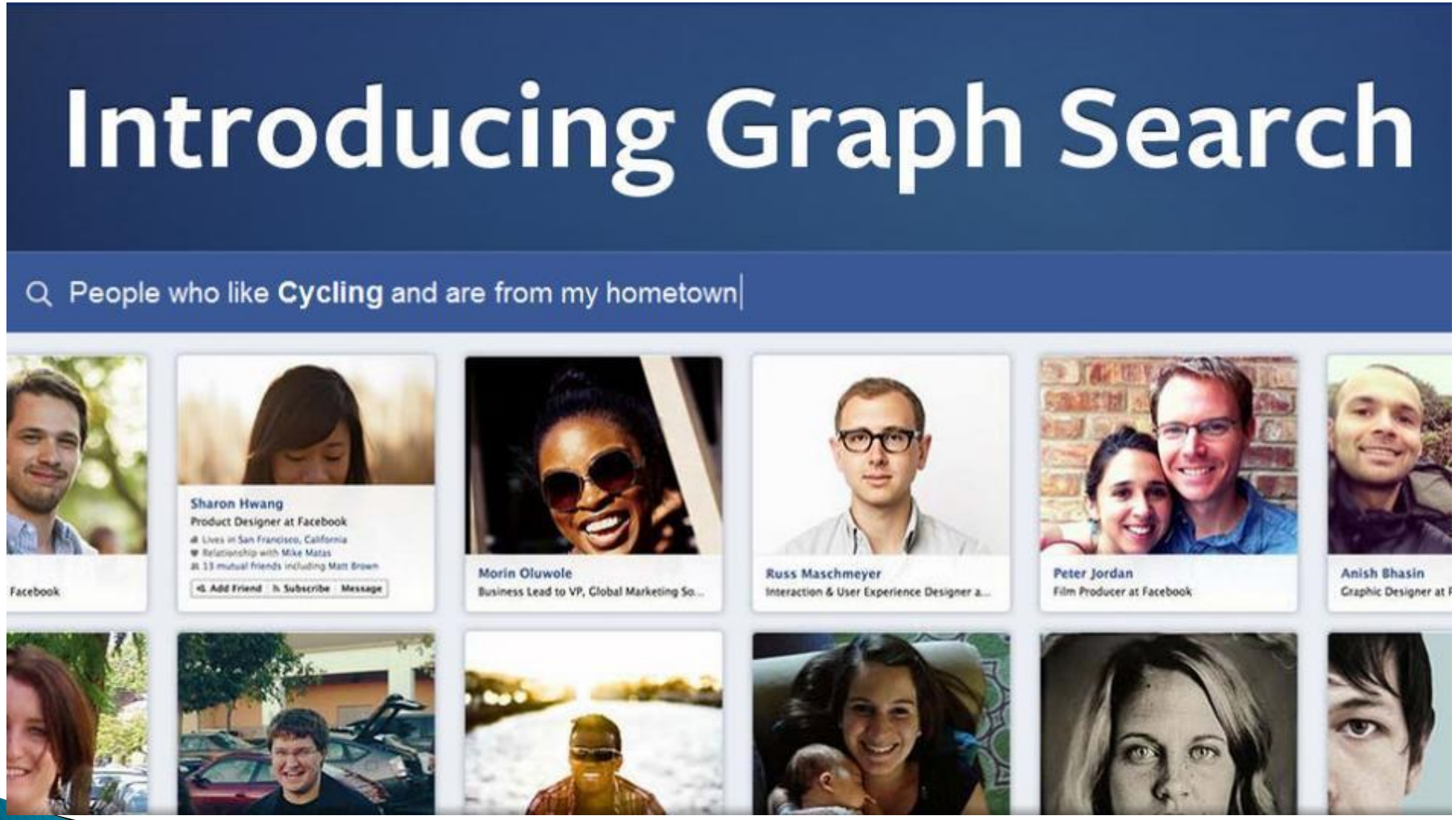
- Structured Queries:
  - A priori user effort in understanding the schemas behind databases.
  - Effort in mastering the syntax of a query language.
  - Satisfying information needs may depend on multiple querying operations.
  - Input: Structured query
  - Output: data records, aggregations, etc
- QA:
  - Delegates more ‘interpretation effort’ to the machines.
  - Input: natural language query
  - Output: direct answer

When to use?

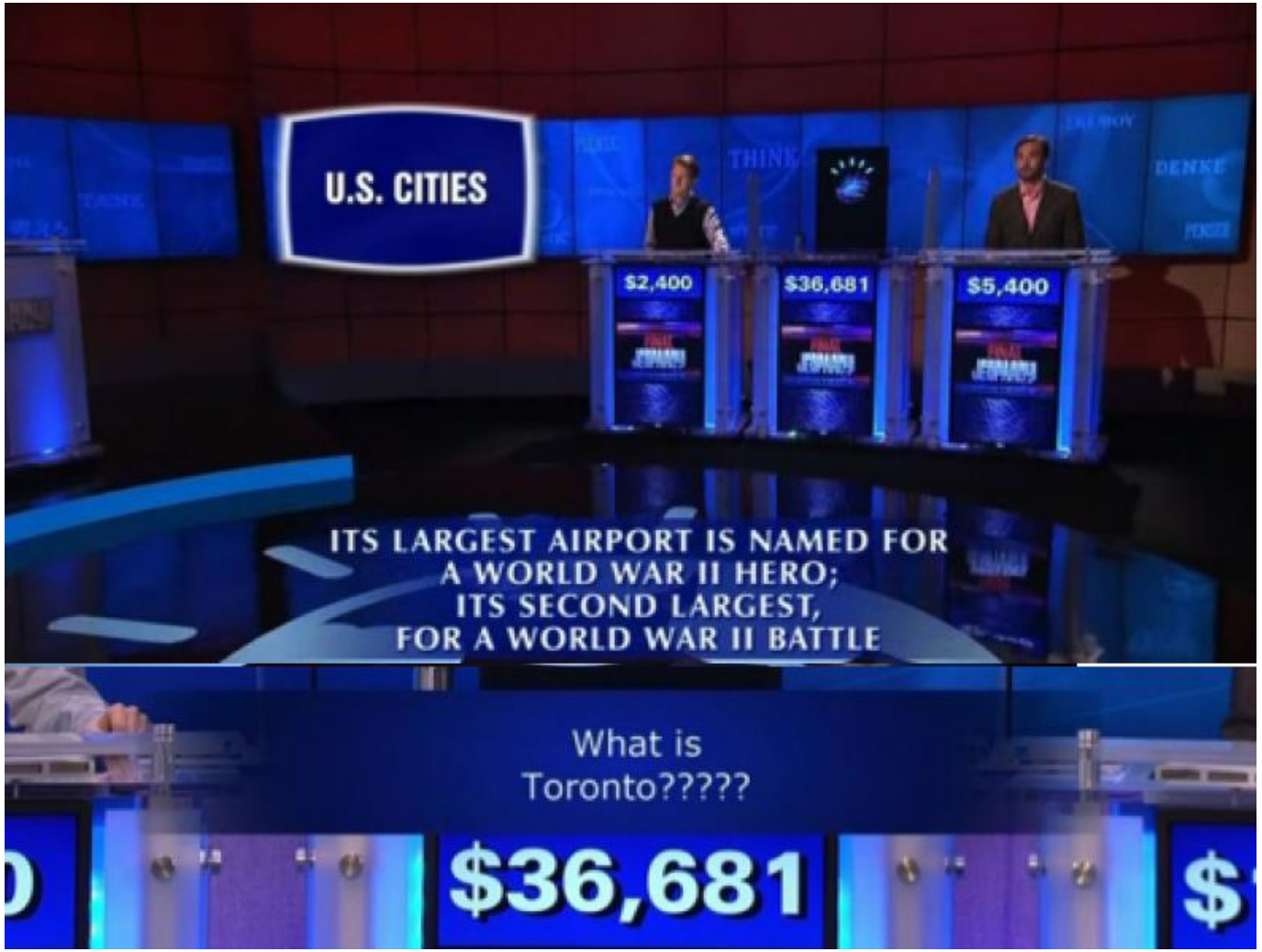
- Keyword search:
  - Simple information needs.
  - Predictable search behavior.
  - Vocabulary redundancy (large document collections, Web)
- Structured queries:
  - Precision/recall guarantees.
  - Small & centralized schemas.
  - More data volume/less semantic heterogeneity.
- QA:
  - Specification of complex information needs.
  - More automated semantic interpretation.

QA: Vision

QA: Reality (FB Graph Search)



QA: Reality (Watson)



QA: Reality (Watson)

QA: Reality (Siri)

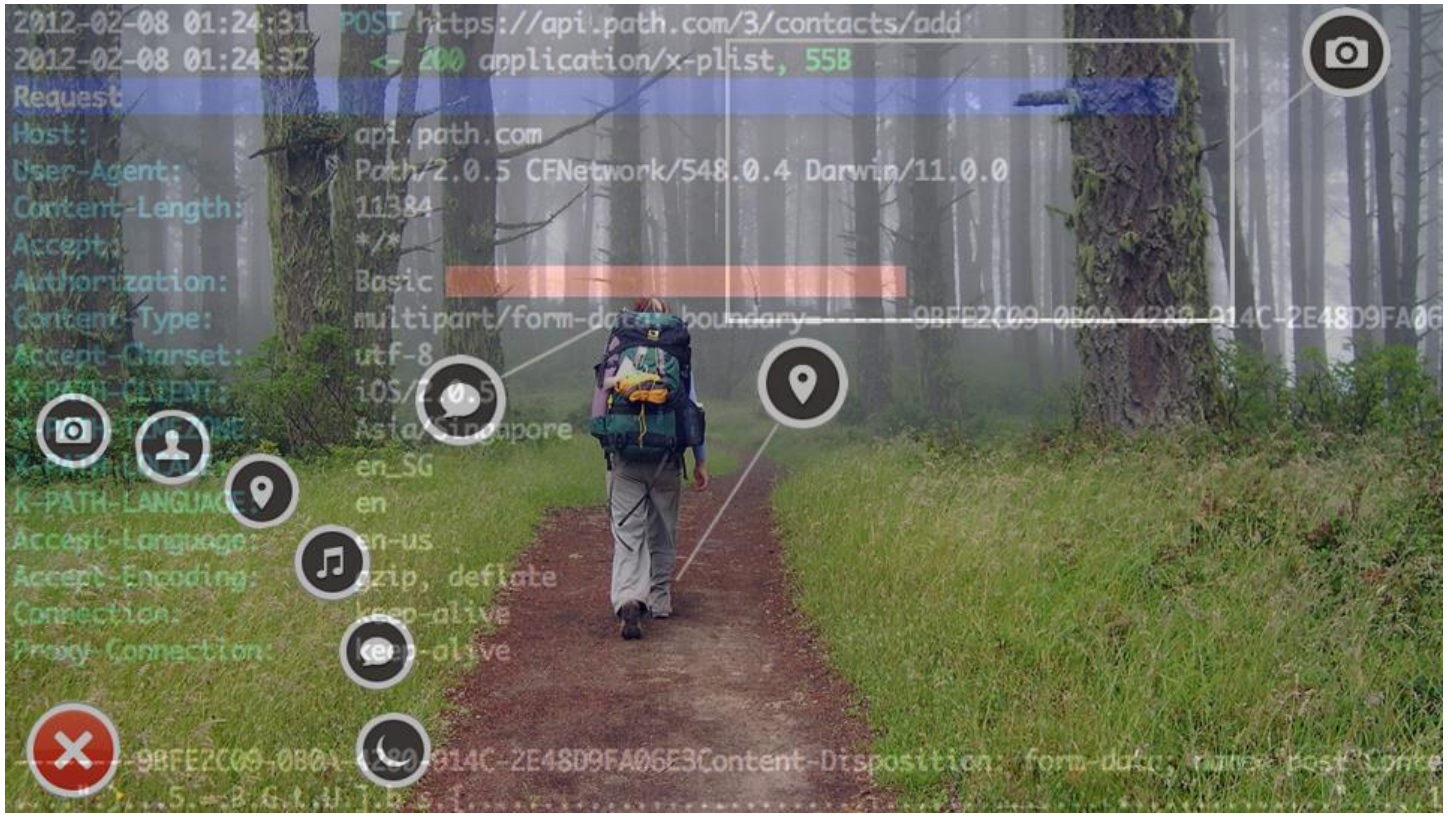
To Summarize

- QA is usually associated with the delegation of more of the 'interpretation effort' to the machines.
- QA supports the specification of more complex information needs.
- QA, Information Retrieval and Databases are complementary.

Big Data, Linked Data & QA

Big Data

- Big Data: More complete *data-based* picture of the world.



From Rigid Schema to Schemaless

- Heterogeneous, complex and large-scale databases.
- Very-large and dynamic “schemas”.

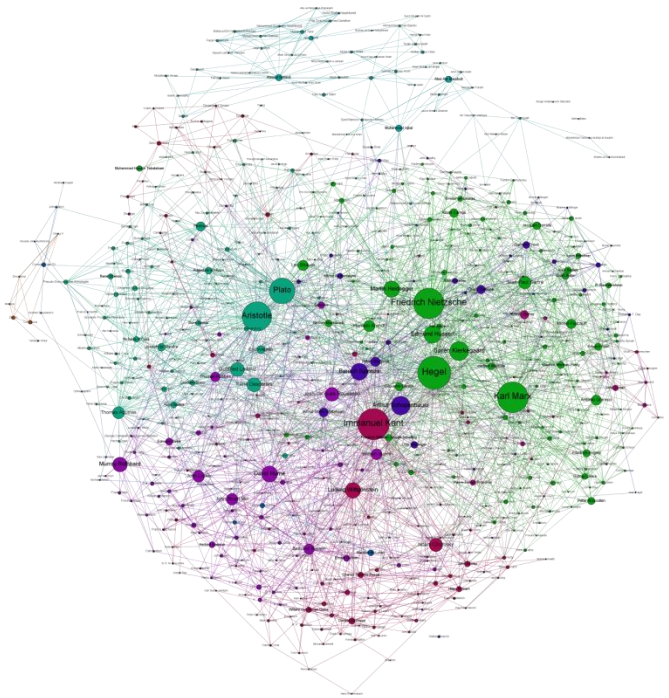
circa 2013

1,000s-1,000,000s attributes

circa 2000

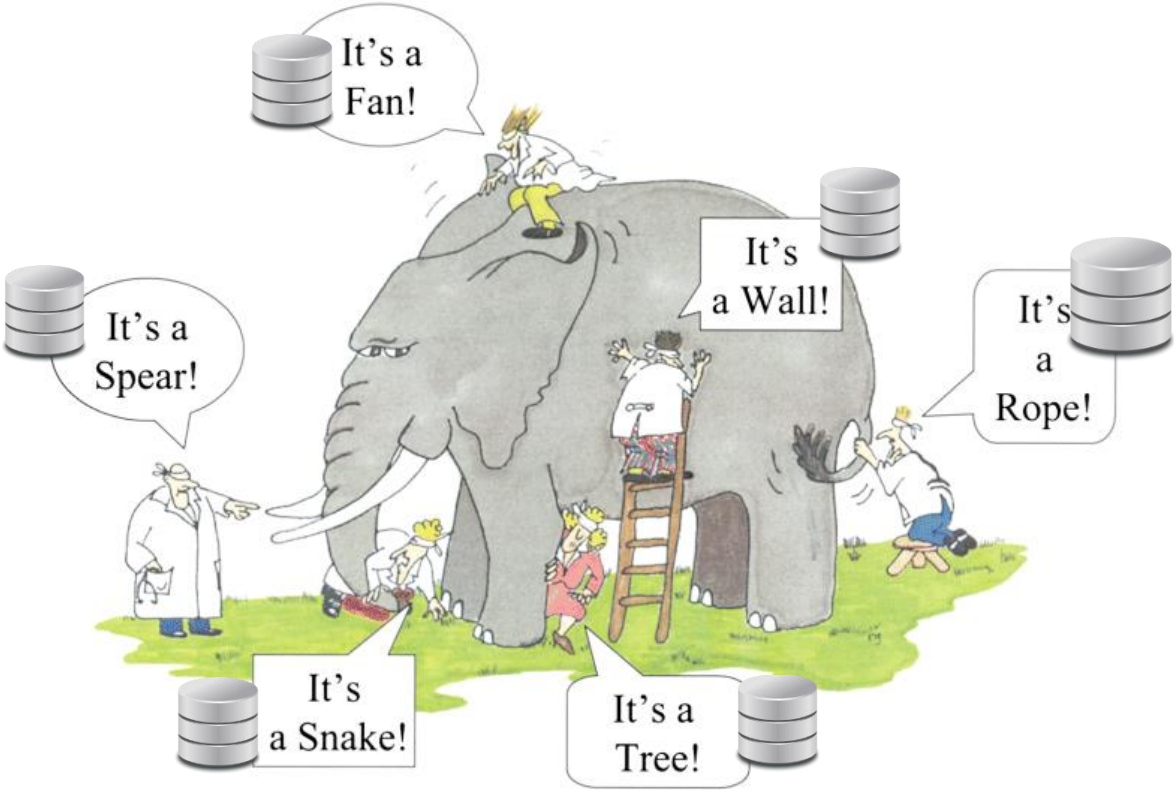
10s-100s attributes

EMP_NO	FIRST_NAME	LAST_NAME	PHONE_EXT	HIRE_DATE	DEPT...	JOB_C...	JOB_GR...	JOB_COUNT...	SALARY	FULL_NAME
2	Robert	Nelson	250	12.28.1988 12:00 am	600	VP	2	USA	105.900,00	Nelson, Robert
4	Bruce	Young	233	12.28.1988 12:00 am	621	Eng	2	USA	97.500,00	Young, Bruce
5	Kim	Lambert	22	02.06.1989 12:00 am	130	Eng	2	USA	102.750,00	Lambert, Kim
8	Leslie	Johnson	410	04.05.1989 12:00 am	180	Mktg	3	USA	64.635,00	Johnson, Leslie
9	Phil	Forest	229	04.17.1989 12:00 am	622	Mngr	3	USA	75.060,00	Forest, Phil
11	K. J.	Weston	34	01.17.1990 12:00 am	130	SRep	4	USA	86.292,94	Weston, K. J.
12	Terri	Lee	256	05.01.1990 12:00 am	000	Admin	4	USA	53.793,00	Lee, Terri
14	Stewart	Hall	227	06.04.1990 12:00 am	900	Finan	3	USA	69.482,63	Hall, Stewart
15	Katherine	Young	231	06.14.1990 12:00 am	623	Mngr	3	USA	67.241,25	Young, Katherine
20	Chris	Papadopoulos	887	01.01.1990 12:00 am	671	Mngr	3	USA	89.655,00	Papadopoulos, Chr
24	Pete	Fisher	888	09.12.1990 12:00 am	671	Eng	3	USA	81.810,19	Fisher, Pete
28	Ann	Bennet	5	02.01.1991 12:00 am	120	Admin	5	England	22.935,00	Bennet, Ann
29	Roger	De Souza	288	02.18.1991 12:00 am	623	Eng	3	USA	69.482,63	De Souza, Roger
34	Janet	Baldwin	2	03.21.1991 12:00 am	110	Sales	3	USA	61.637,81	Baldwin, Janet
36	Roger	Reeves	6	04.25.1991 12:00 am	120	Sales	3	England	33.620,63	Reeves, Roger
37	Willie	Stansbury	7	04.25.1991 12:00 am	120	Eng	4	England	39.224,06	Stansbury, Willie
44	Leslie	Phong	216	06.03.1991 12:00 am	623	Eng	4	USA	56.034,38	Phong, Leslie
45	Ashok	Ramanathan	209	08.01.1991 12:00 am	621	Eng	3	USA	80.689,50	Ramanathan, Ashc
46	Walter	Steadman	210	08.09.1991 12:00 am	900	CFO	1	USA	116.100,00	Steadman, Walter
52	Carol	Nordstrom	420	10.02.1991 12:00 am	180	PRel	4	USA	42.742,50	Nordstrom, Carol
61	Luke	Leung	3	02.18.1992 12:00 am	110	SRep	4	USA	68.805,00	Leung, Luke
65	Sue Anne	O'Brien	877	03.23.1992 12:00 am	670	Admin	5	USA	31.275,00	O'Brien, Sue Anne



Multiple Interpretations

- Multiple perspectives (conceptualizations) of the reality.
- Ambiguity, vagueness, inconsistency.

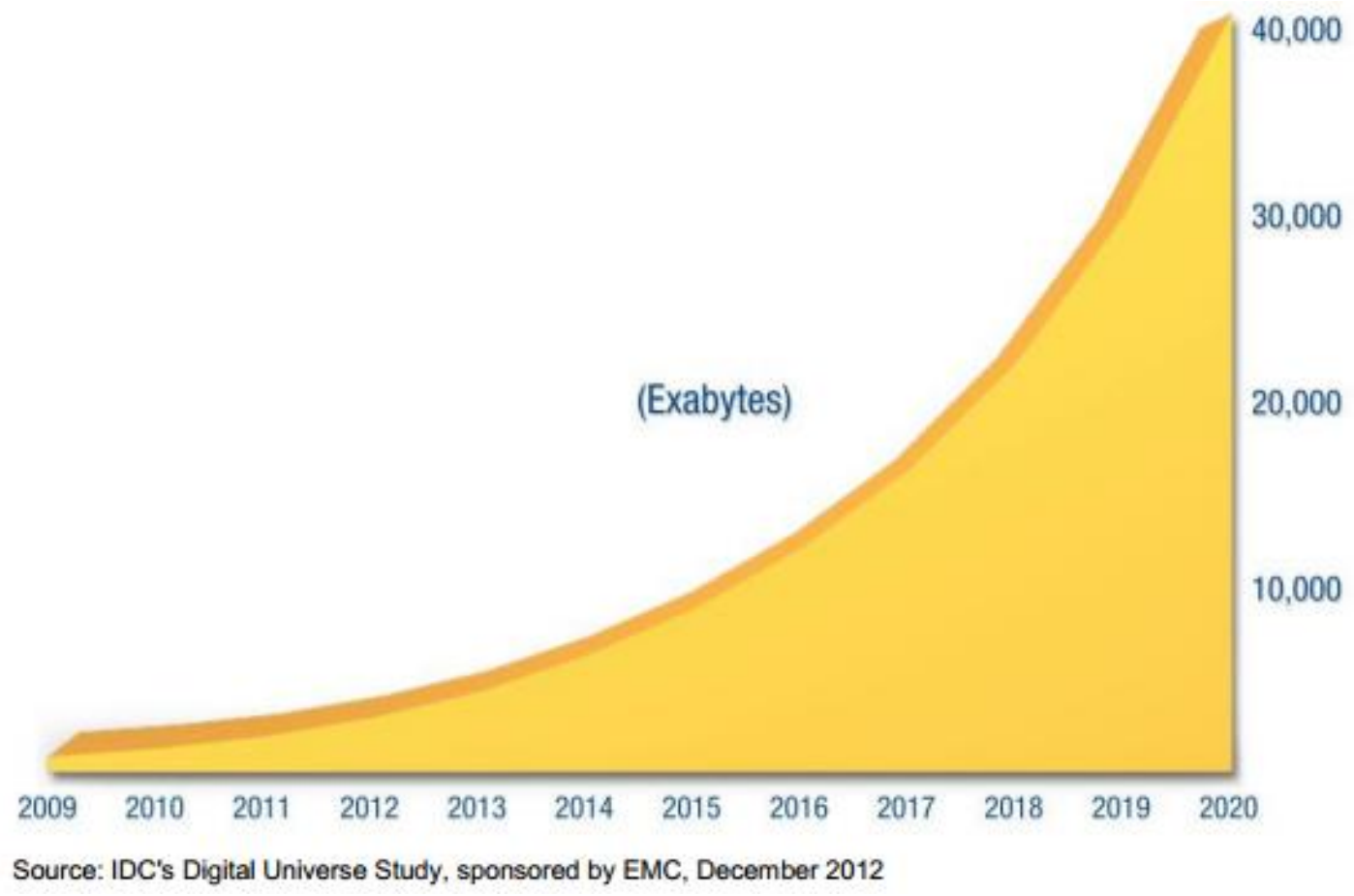


Big Data & Dataspaces

- Franklin et al. (2005): *From Databases to Dataspaces.*
- Helland (2011): *If You Have Too Much Data, then "Good Enough" Is Good Enough*.

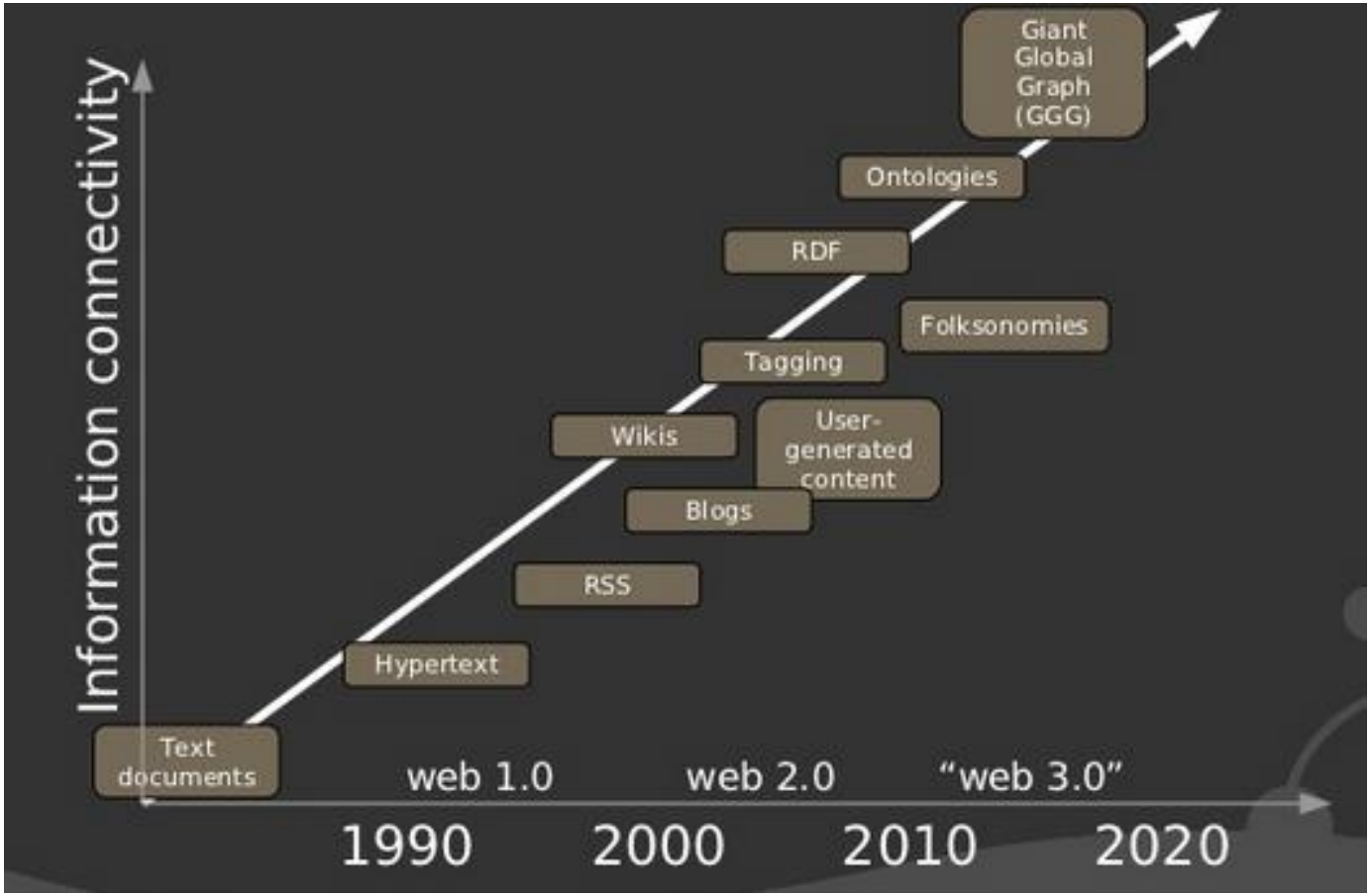
- Fundamental trends:
  - Co-existence of heterogeneous data.
  - Semantic best-effort queries.
  - Pay-as-you go data integration.
  - Co-existent query/search services.

Trend 1: Data size



Gantz & Reinsel, The Digital Universe in 2020 (2012).

Trend 2: Connectedness



Eifrem, A NOSQL Overview And The Benefits Of Graph Database (2009).

Trends 3 & 4: Semi-structured data

- Individualization of content.
- Decentralization of content generation.

Eifrem, A NOSQL Overview And The Benefits Of Graph Database (2009)

Big Data

- Volume
- Velocity
- **Variety**

– The most interesting but usually neglected dimension.

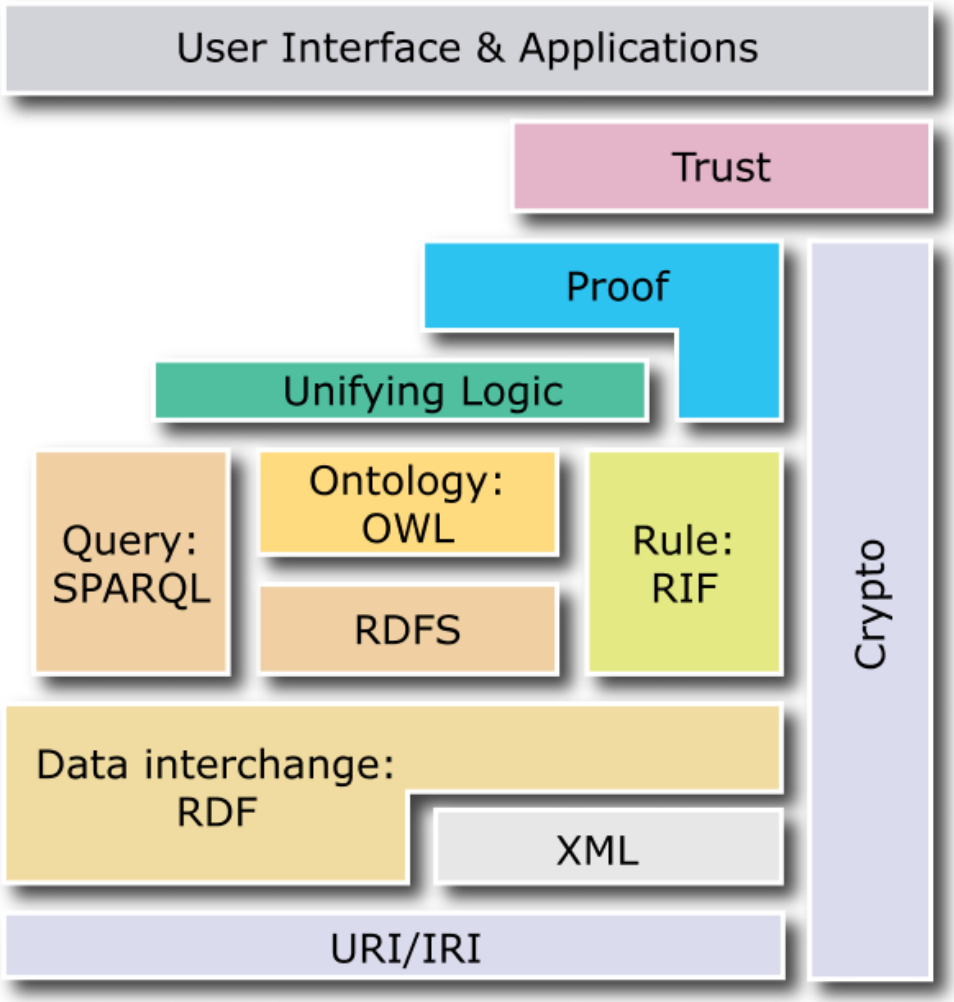
The Semantic Web Vision

2001:

- Software which is able to understand meaning (intelligent, flexible)
- Leveraging the Web for information scale

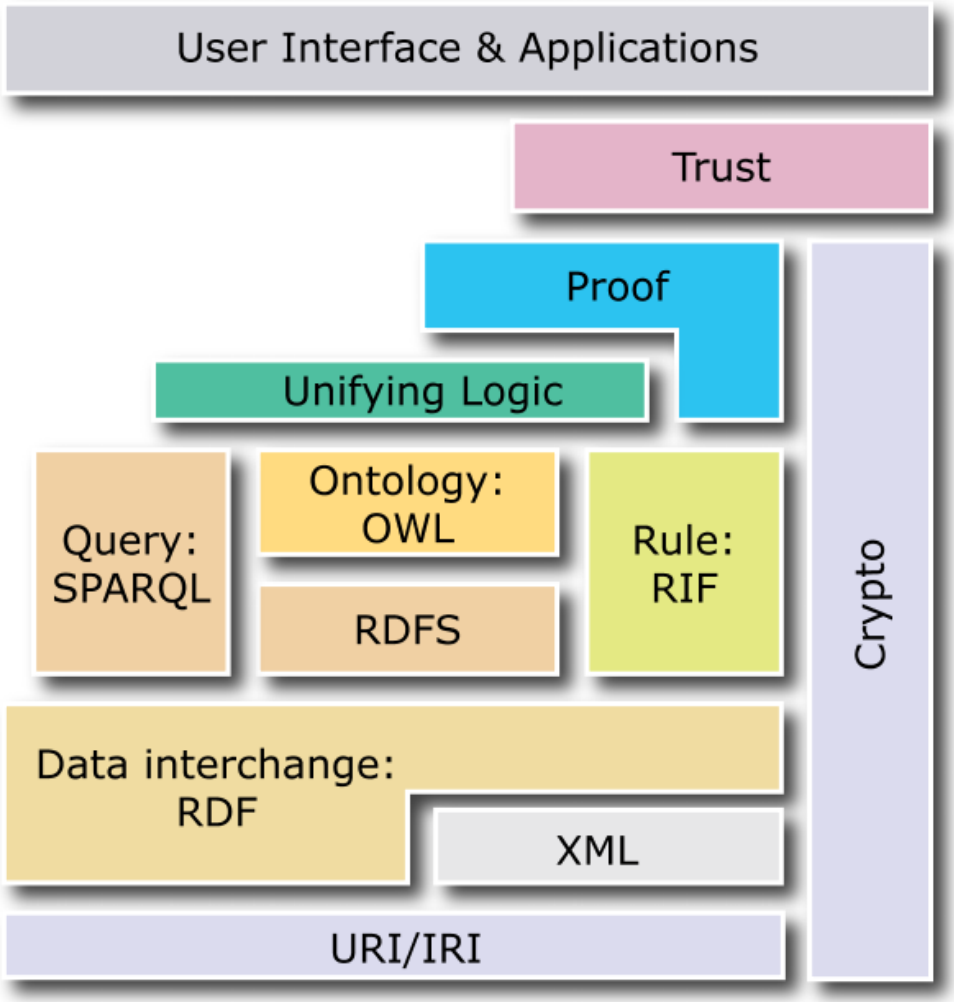
The Semantic Web Vision

- What was the plan to achieve it?
- Build a Semantic Web Stack
- Which covers both representation and reasoning



Reality Check

- Adoption:
  - No significant data growth
- Ontologies are not straightforward to build:
  - People are not familiriazed with the tools and principles
  - Difficult to keep consistency at Web scale
- Scalability



Reasoning

- Problems:
  - Consistecy
  - Scalability

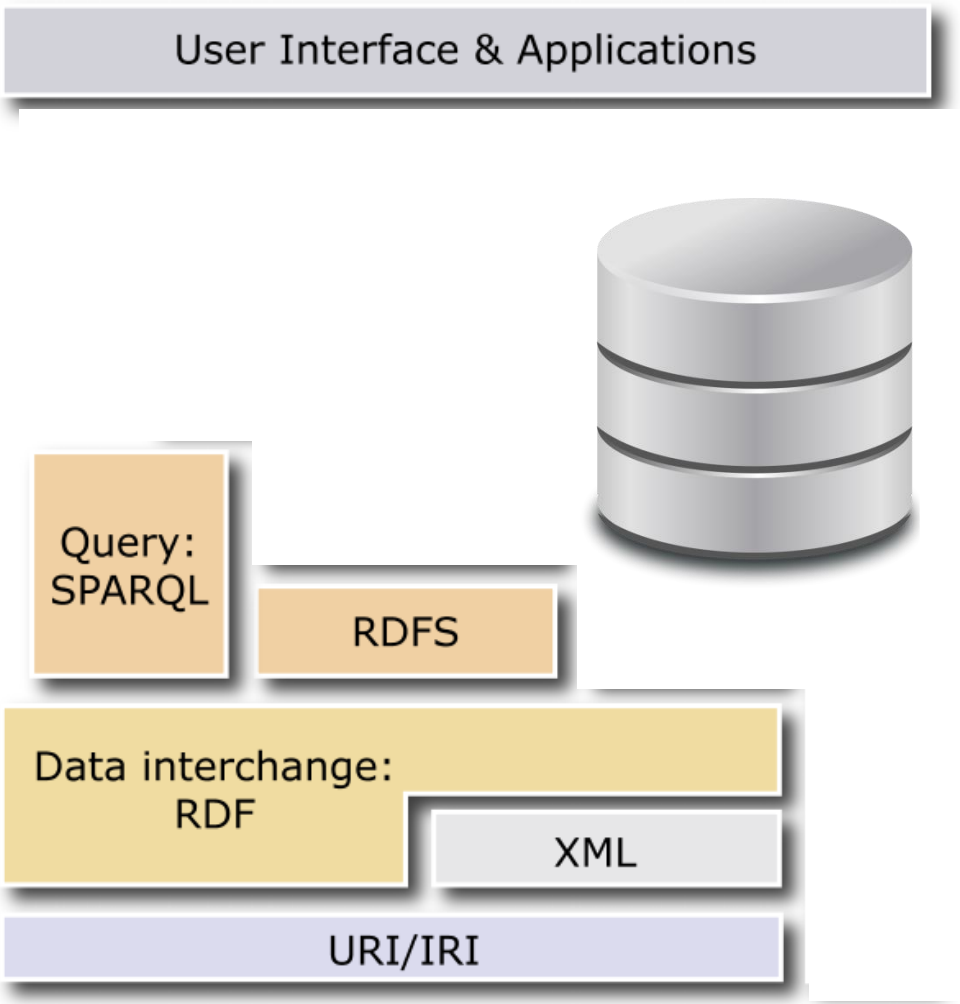
Logic World

Web World

Linked Data

2006:

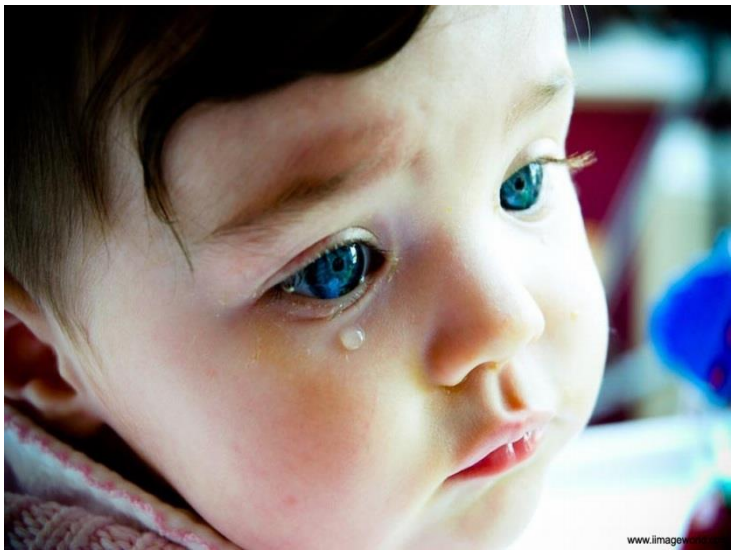
- The Web as a Huge Database
- Fundamental step for data creation



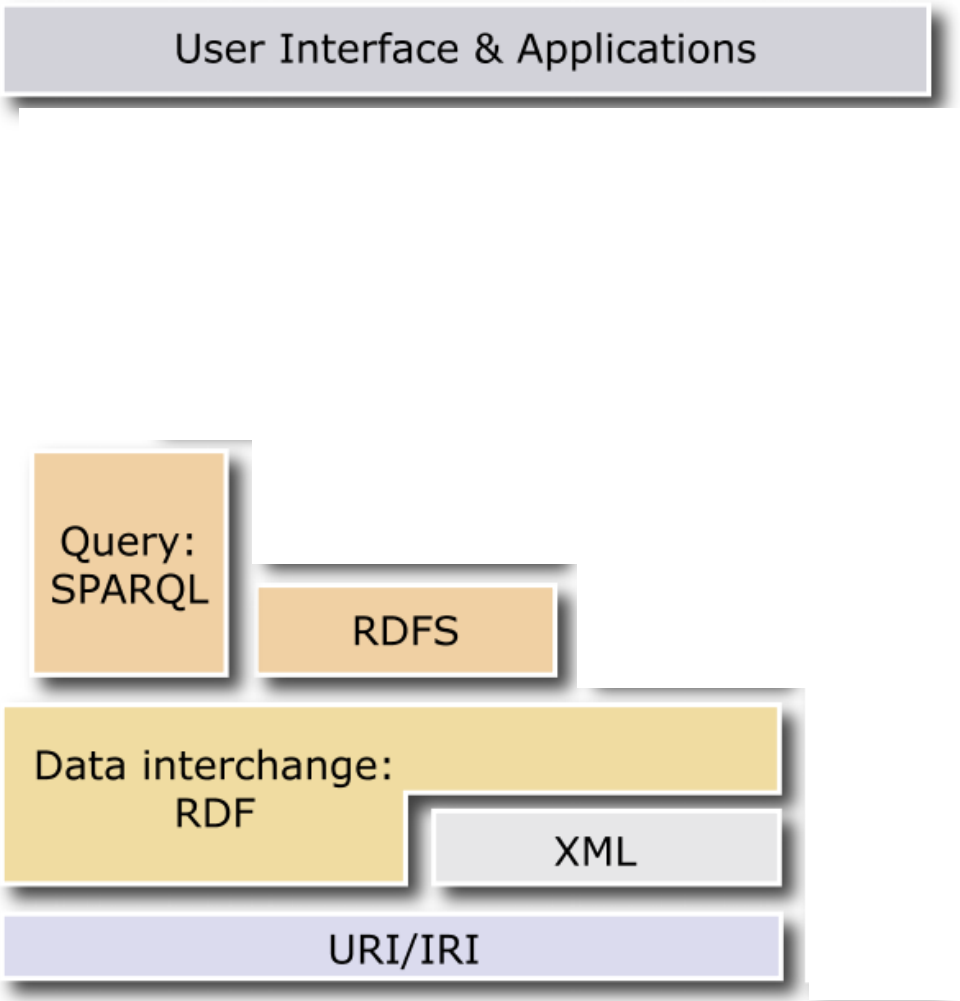
Linked Data

- Positives:
  - Solid adoption in the Open Data context (eGovernment, eScience, etc,...)
  - Existing data is relevant (you can build real applications)
- Negatives:
  - Data consumption is a problem
  - Data generation beyond databases mapping/triplification is also a problem
  - Still far from the Semantic Web vision

No Reasoning, no Fun?



- Where are the intelligence and flexibility?
- QA as a way to restore the initial vision!



Consuming/Using Linked Data

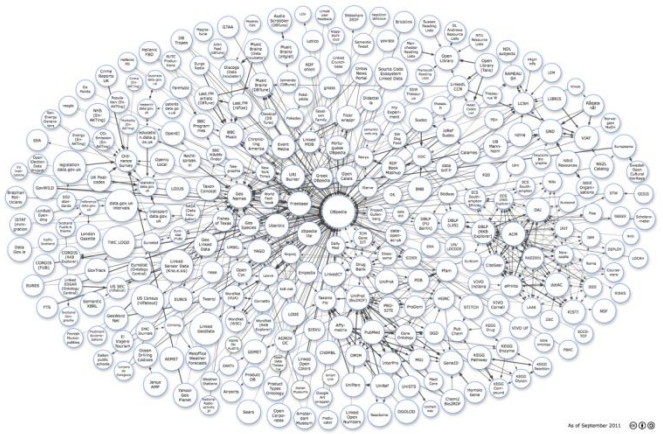
- With Linked Data we are still in the DB world

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbonto: <http://dbpedia.org/ontology/>
SELECT ?university
WHERE {
    dbpedia:Barack_Obama dbonto:spouse ?spouse .
    ?spouse dbonto:almaMater ?university .
}
```

Consuming/Using Linked Data

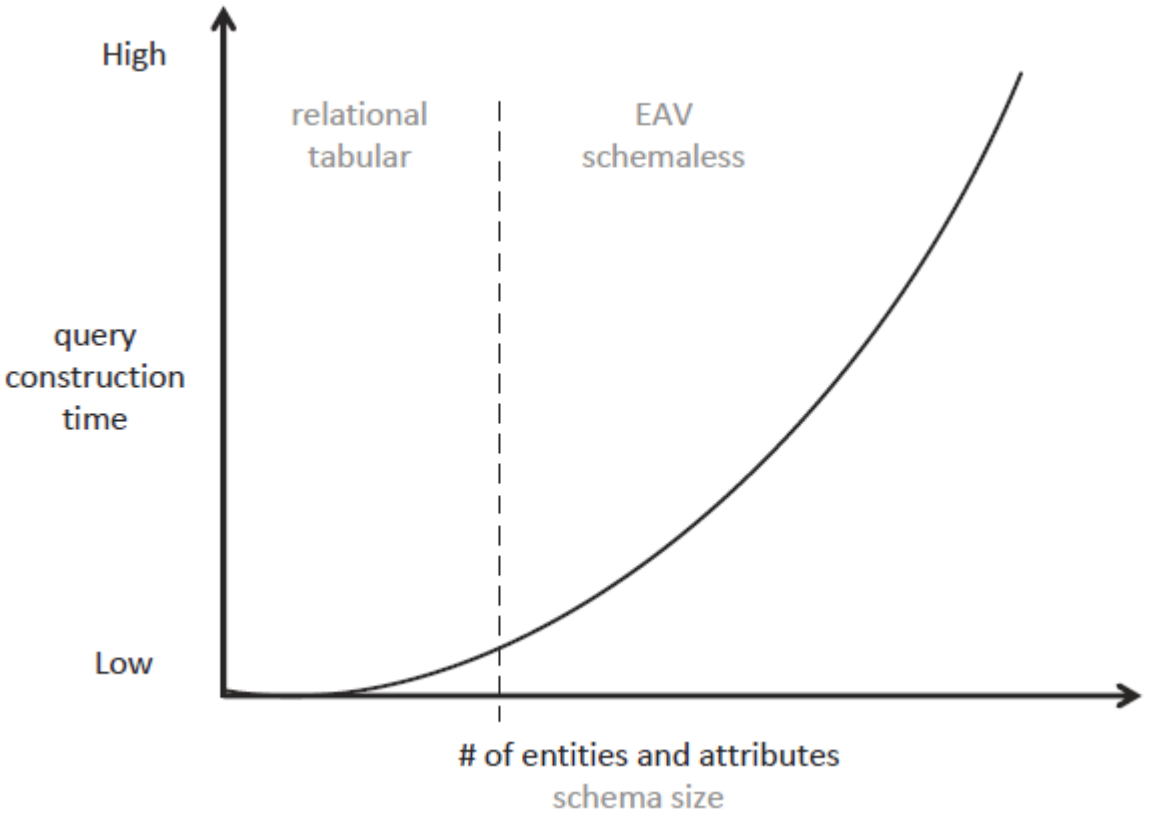
- With Linked Data we are still in the DB world
- (but slightly worse)

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbonto: <http://dbpedia.org/ontology/>
SELECT ?university
WHERE {
    dbpedia:Barack_Obama dbonto:spouse ?spouse .
    ?spouse dbonto:almaMater ?university .
}
```



Big Data: Structured queries

Query construction size x schema size



10-100
attributes

10^3-10^6s
attributes

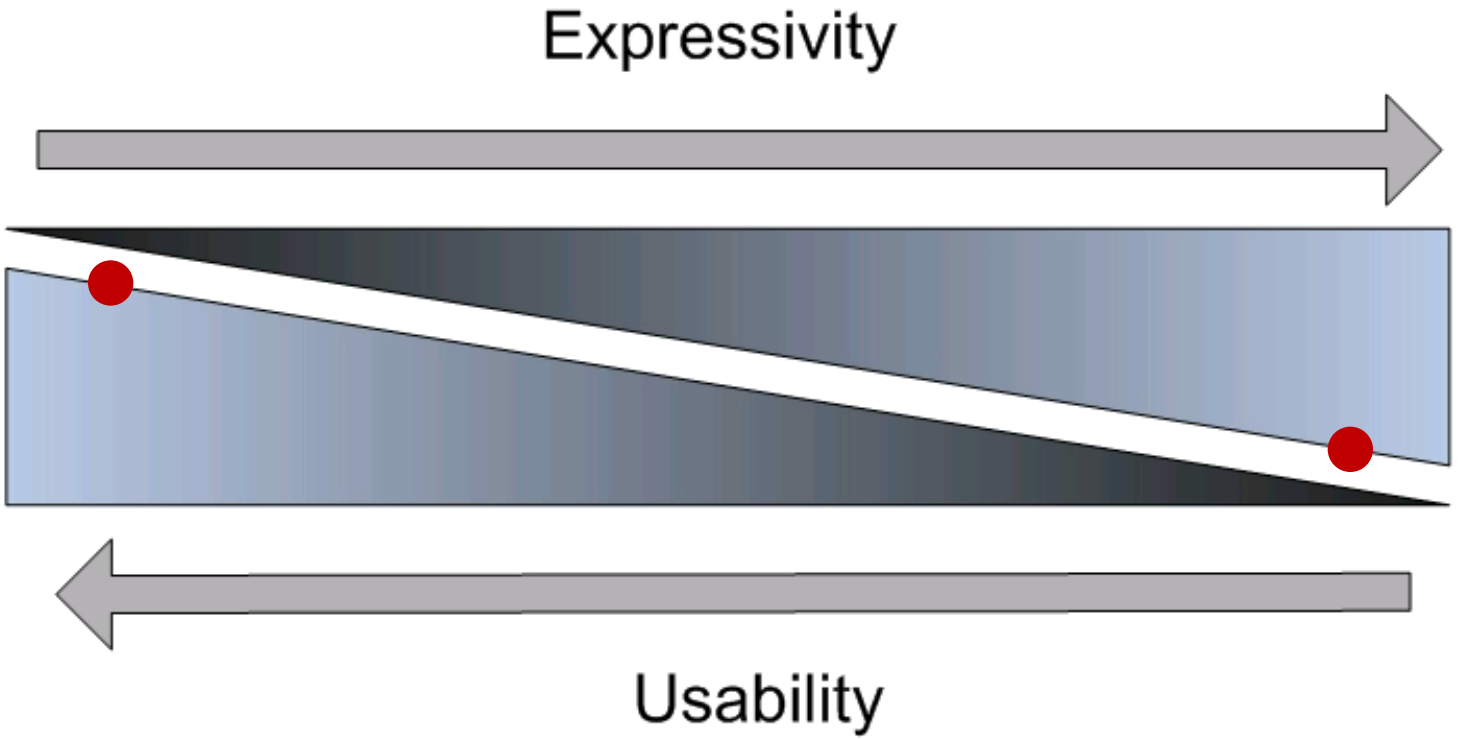
Linked Data

- Data Model Features:
  - Graph-based data model
  - Extensible schema
  - Entity-centric data integration
- Specific Features:
  - Designed over open Web standards
  - Based on the Web infrastructure (HTTP, URIs)

QA for Linked Data

- Addresses practical problems of data accessibility in a data heterogeneity scenario.
- A fundamental part of the original Semantic Web vision.

Query/Search Spectrum



Keyword search | Entiy-centric search | … | Visual query interfaces | Structured queries

Adapted from Kauffman et al (2009)

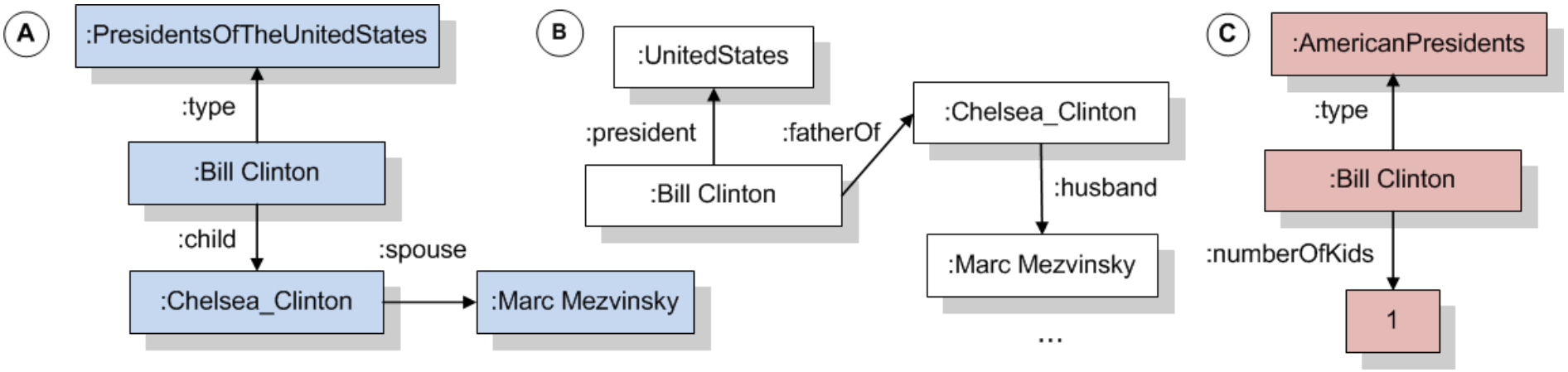
Vocabulary Problem for Databases

Who is the daughter of Bill Clinton married to?

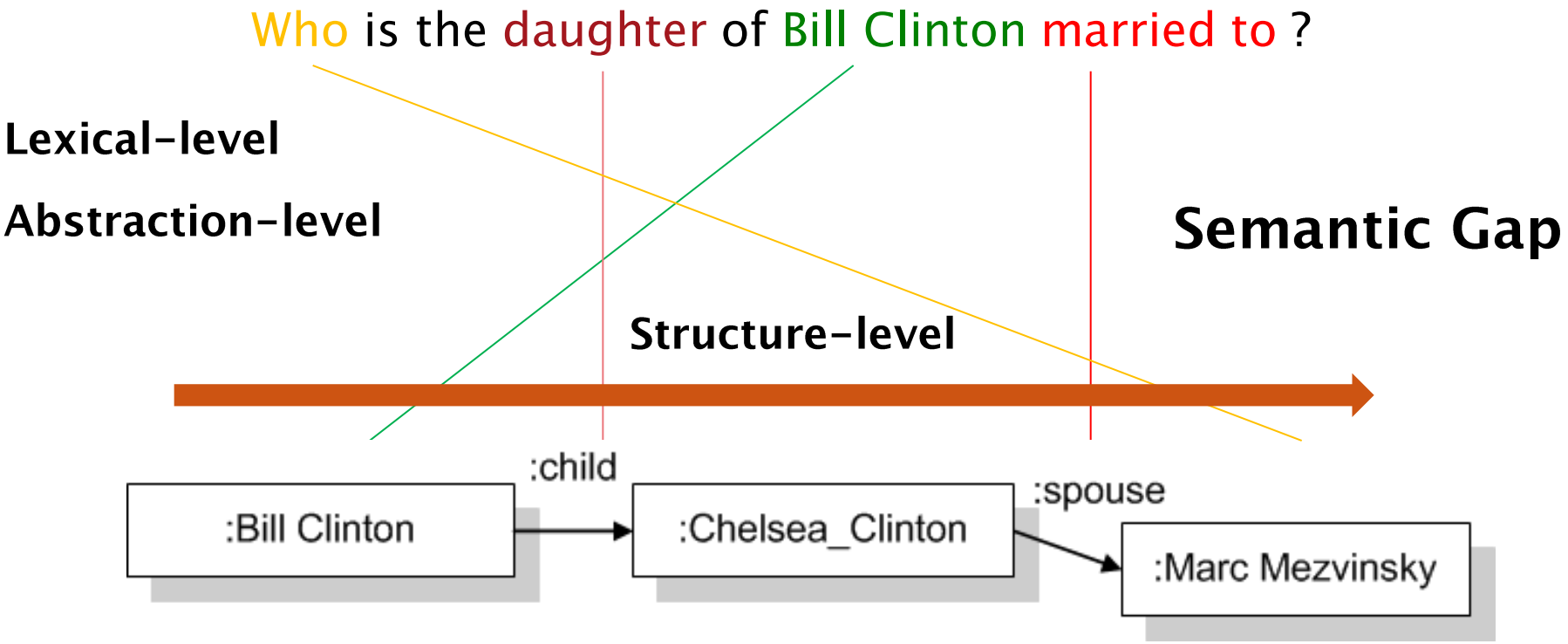
Semantic Gap

Schema-agnostic queries

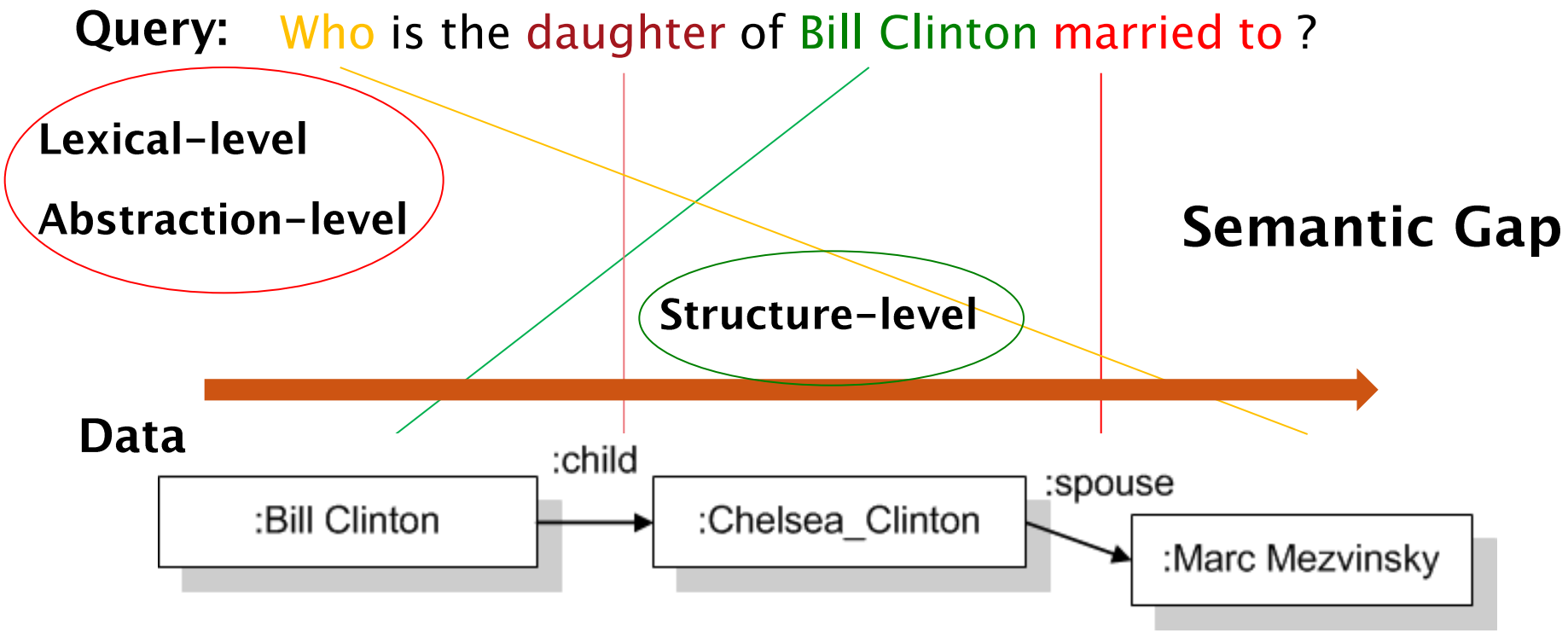
Possible representations



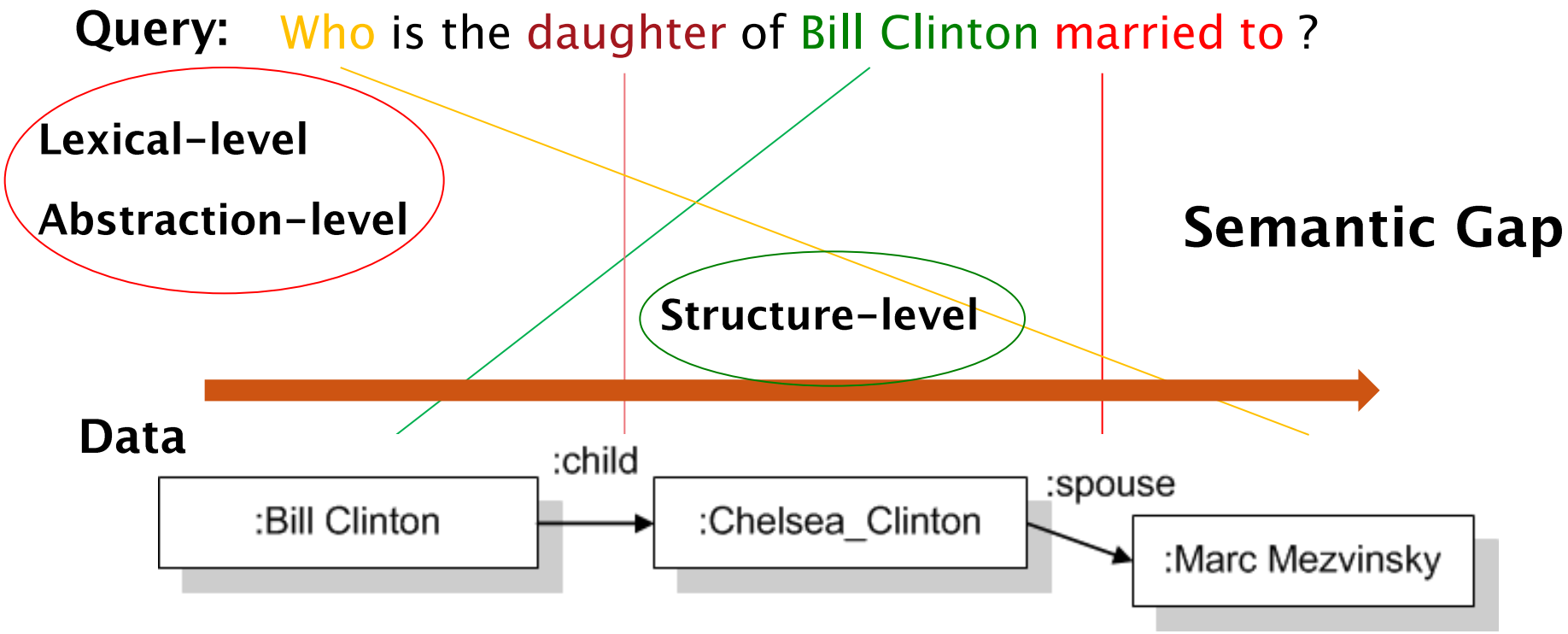
Vocabulary Problem for Databases



Vocabulary Problem for Databases



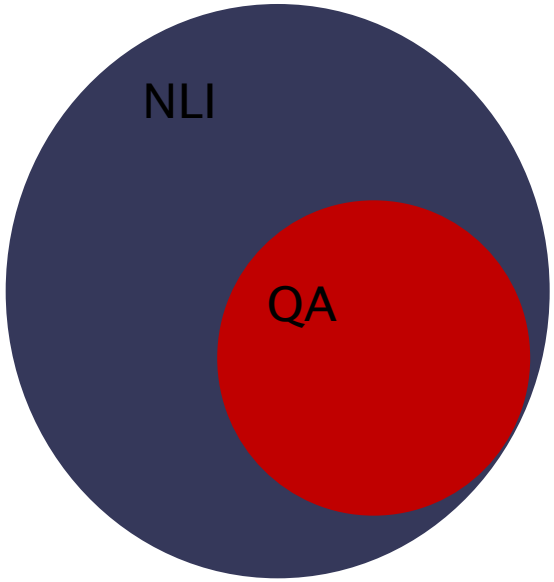
Vocabulary Problem for Databases



Popescu et al. (2003): Semantic tractability

NLI & QAs

- Natural Language Interfaces (NLI)
  - Input: Natural language queries
  - Output: Either processed or unprocessed queries
    - Processed: Direct answers.
    - Unprocessed: Database records, text snippets, documents.



To Summarize

- Schema size and heterogeneity represent a fundamental shift for databases.
- Addressing the associated data management challenges (specially querying) depends on the development of principled semantic models for databases.
- QA/Natural Language Interfaces (NLIs) as schema-agnostic query mechanisms.

The Anatomy of a QA System

QA4LD Requirements

- High usability:
  - Supporting natural language queries.
- High expressivity:
  - Path, conjunctions, disjunctions, aggregations, conditions.
- Accurate & comprehensive semantic matching:
  - High precision and recall.
- Low maintainability:
  - Easily transportable across datasets from different domains (minimum adaptation effort/low adaptation time).
- Low query execution time:
  - Suitable for interactive querying.
- High scalability:
  - Scalable to a large number of datasets (Organization-scale, Web-scale).

Basic Concepts & Taxonomy

- Categorization of questions and answers.
- Important for:
  - Understanding the challenges before attacking the problem.
  - Scoping the system.

- Based on:
  - Chin-Yew Lin: Question Answering.
  - Farah Benamara: Question Answering Systems: State of the Art and Future Directions.

Terminology: Question Phrase

- The part of the question that says what is being asked:
  - Wh-words:
    - who, what, which, when, where, why, and how
  - Wh-words + nouns, adjectives or adverbs:
    - •"which party …", "which actress …", "how long …", "how tall …".

Terminology: Question Type

- Useful for distinguishing different processing strategies
  - ***FACTOID***: *"Who is the wife of Barack Obama?"*
  - ***LIST***: *"Give me all cities in the US with less than 10000 inhabitants."*
  - ***DEFINITION***: *"Who was Tom Jobim?"*
  - ***RELATIONSHIP***: *"What is the connection between Barack Obama and Indonesia?"*
  - ***SUPERLATIVE***: *"What is the highest mountain?"*
  - ***YES-NO***: *"Was Margaret Thatcher a chemist?"*
  - ***OPINION***: *"What do most Americans think of gun control?"*
  - ***CAUSE & EFFECT***: *"Why did the revenue of IBM drop?"*

Terminology: Answer Type

- The class of object sought by the question
  - **Person** (from “Who …”)
  - **Place** (from “Where …”)
  - **Process & Method** (from “How …”)
  - **Date** (from “When …”)
  - **Number** (from “How many …”)
  - **Explanation & Justification** (from “Why …”)

Terminology: Question Focus & Topic

- Question focus is the **property** or **entity** that is being sought by the question
  - "In which **city** Barack Obama was born?"
  - "*What is the **population** of Galway?*"
- Question topic: What the question is generally about :
  - "*What is the height of Mount Everest?*" (geography, mountains)
  - "Which organ is affected by the Meniere's disease?" (medicine)

Terminology: Data Source Type

- Structure level:
  - Structured data (databases)
  - Semi-structured data (e.g. comment field in databases, XML)
  - Free text
- Data source:
  - Single (Centralized)
  - Multiple
  - Web-scale

Terminology: Domain Type

- Domain Scope:
  - Open Domain
  - Domain specific
- Data Type:
  - Text
  - Image
  - Sound
  - Video
- Multi-modal QA

Terminology: Answer Format

- Long answers
  - Definition/justification based.
- Short answers
  - Phrases.
- Exact answers
  - Named entities, numbers, aggregate, yes/no

Answer Quality Criteria

- **Relevance**: The level in which the answer addresses users information needs.
- **Correctness**: The level in which the answer is factually correct.
- **Conciseness**: The answer should not contain irrelevant information.
- **Completeness**: The answer should be complete.
- **Simplicity**: The answer should be simple to be interpreted by the data consumer.
- **Justification**: Sufficient context should be provided to support the data consumer in the determination of the query correctness.

Answer Assessment

- **Right**: The answer is correct and complete.
- **Inexact**: The answer is incomplete or incorrect.
- **Unsupported**: the answers does not have an appropriate justification.
- **Wrong**: The answer is not appropriate for the question.

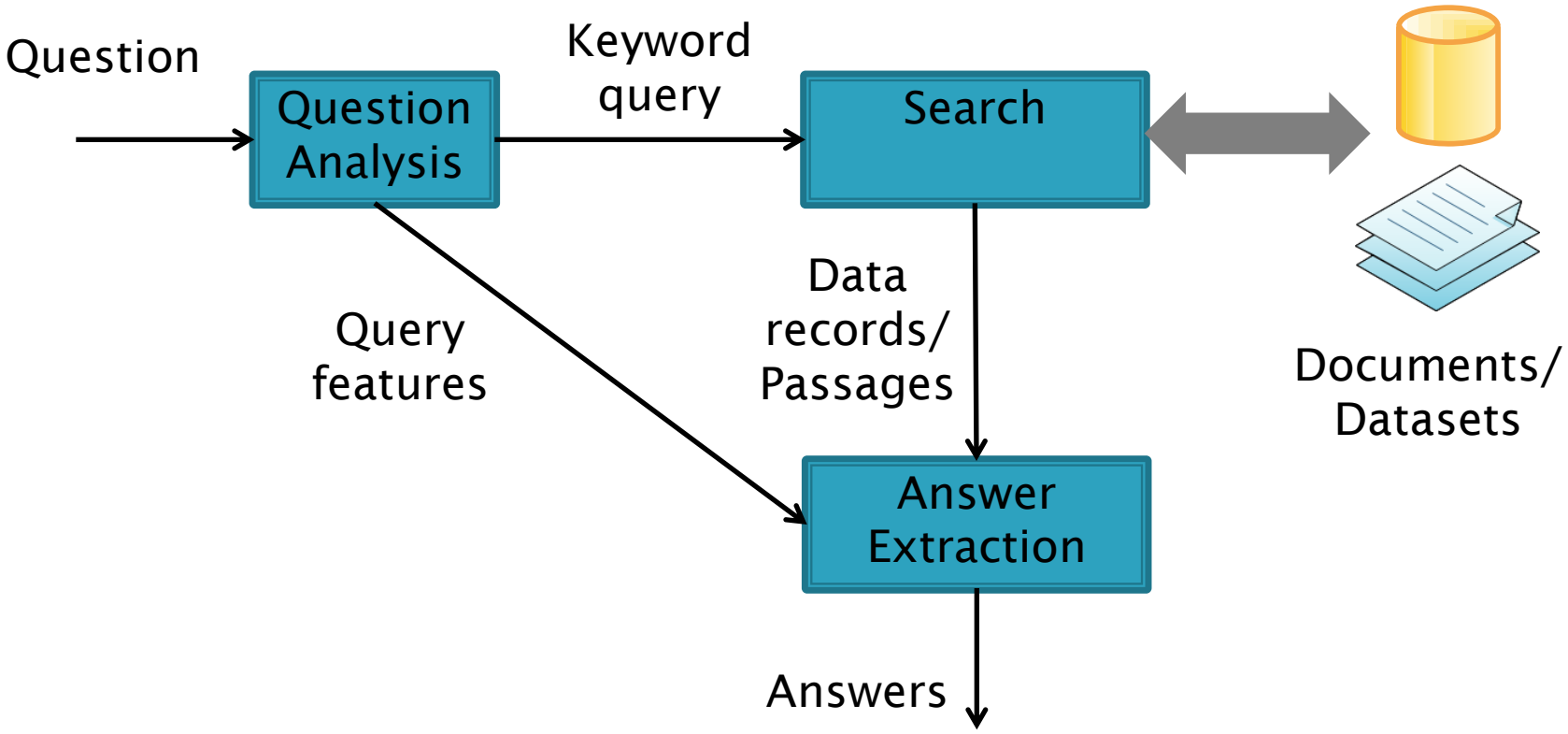
Answer Processing

- **Simple Extraction**: cut and paste of snippets from the original document(s) / records from the data.
- **Combination**: Combines excerpts from multiple sentences, documents / multiple data records, databases.
- **Summarization**: Synthesis from large texts / data collections.
- **Operational/functional**: Depends on the application of functional operators.
- **Reasoning**: Depends on the an inference process.

Complexity of the QA Task

- **Semantic Tractability** (Popescu et al., 2003): Vocabulary distance between the query and the answer.
- **Answer Locality** (Webber et al., 2003): Whether answer fragments are distributed across different document fragments/documents or datasets/dataset records.
- **Derivability** (Webber et al, 2003): Dependent if the answer is explicit or implicit. Level of reasoning dependency.
- **Semantic Complexity**: Level of ambiguity and discourse/data heterogeneity.

Main Components



Main Components

- **Question Analysis**: Includes question parsing, extraction of core features (NER, answer type, etc).
- **Search** (i.e. passage retrieval): Pre-selection of fragments of text (sentences, paragraphs, documents) and data (records, datasets) which may contain the answer.
- **Answer Extraction**: Processing of the answer based on the passages.

QA over Linked Data

QA Systems (Case Studies)

- Aqualog & PowerAqua (Lopez et al. 2006)
- ORAKEL (Cimiano et al, 2007)
- QuestIO & Freya (Damljanovic et al. 2010)
- Pythia (Unger & Cimiano, 2011)
- Treo (Freitas et al. 2011, 2014)

PowerAqua (Lopez et al. 2006)

- Key contribution: semantic similarity mapping.
- Terminological Matching:
  - WordNet-based
  - Ontology Based
  - String similarity
  - Sense-based similarity matcher
- Evaluation: QALD (2011).
- Extends the AquaLog system.

WordNet

Noun

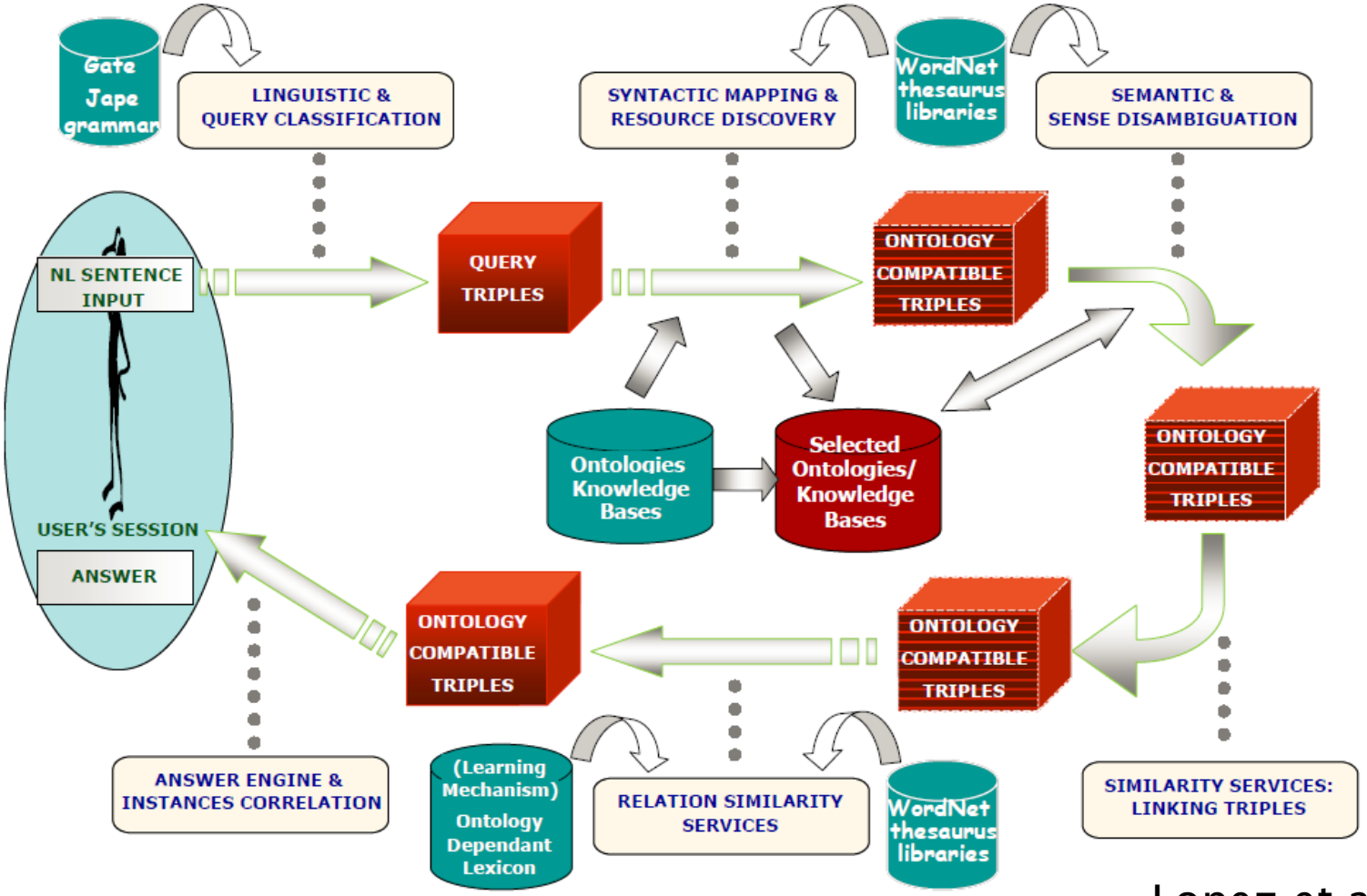
- S: (n) **power**, powerfulness (possession of controlling influence) *"the deterrent power of nuclear weapons"; "the power of his love saved her"; "his powerfulness was concealed by a gentle facade"*
  - ***direct hyponym*** / *full hyponym*
  - *attribute*
    - S: (adj) powerful (having great power or force or potency or effect) *"the most powerful government in western Europe"; "his powerful arms"; "a powerful bomb"; "the horse's powerful kick"; "powerful drugs"; "a powerful argument"*
    - S: (adj) powerless (lacking power)
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) quality (an essential and distinguishing attribute of something or someone) *"the quality of mercy is not strained"--Shakespeare*
  - *antonym*
    - W: (n) powerlessness [Opposed to: power, powerfulness] (the quality of lacking strength or power; being weak and feeble)
  - *derivationally related form*

Sense-based similarity matcher

- Two words are strongly similar if any of the following holds:
  - 1. They have a synset in common (e.g. “human” and “person”)
  - 2. A word is a hypernym/hyponym in the taxonomy of the other word.
  - 3. If there exists an allowable “is-a” path connecting a synset associated with each word.
  - 4. If any of the previous cases is true and the definition (gloss) of one of the synsets of the word (or its direct hypernyms/hyponyms) includes the other word as one of its synonyms, we said that they are highly similar.

Lopez et al. 2006

PowerAqua (Lopez et al. 2006)



Lopez et al. 2006

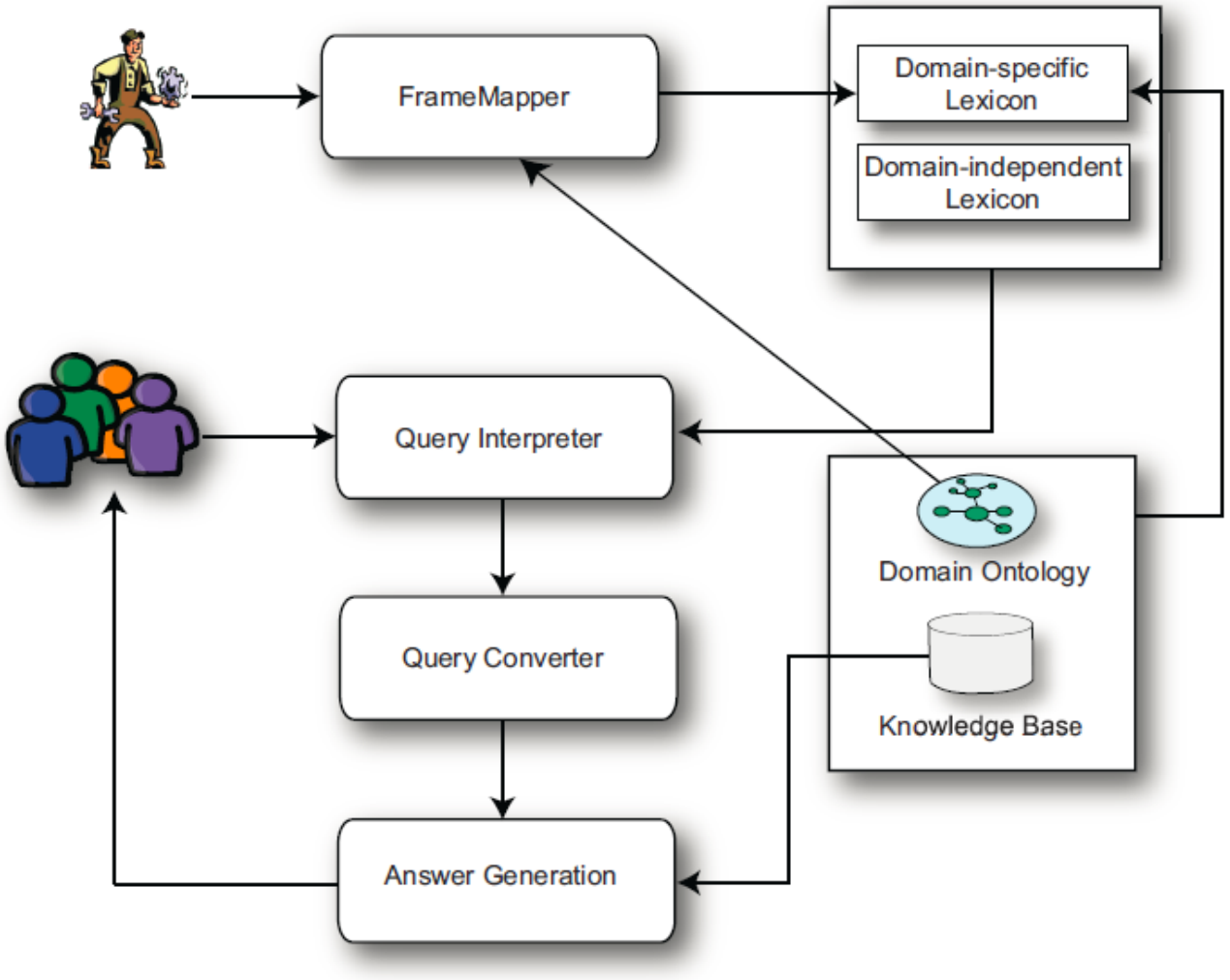
ORAKEL (Cimiano et al. 2007)

- Key contribution:
  - FrameMapper lexicon engineering approach.
  - Parsing strategy: Logical Description Grammars (LDGs).
    - LDG is inspired by Lexicalized Tree Adjoining Grammars (LTAGs).
    - An important characteristic of these trees is that they encapsulate all syntactic/semantic arguments of a word.
- Terminological Matching:
  - FrameMapper
- Evaluation:
  - Dataset: domain-specific
  - Dimensions: Relevance, Performance, lexicon construction

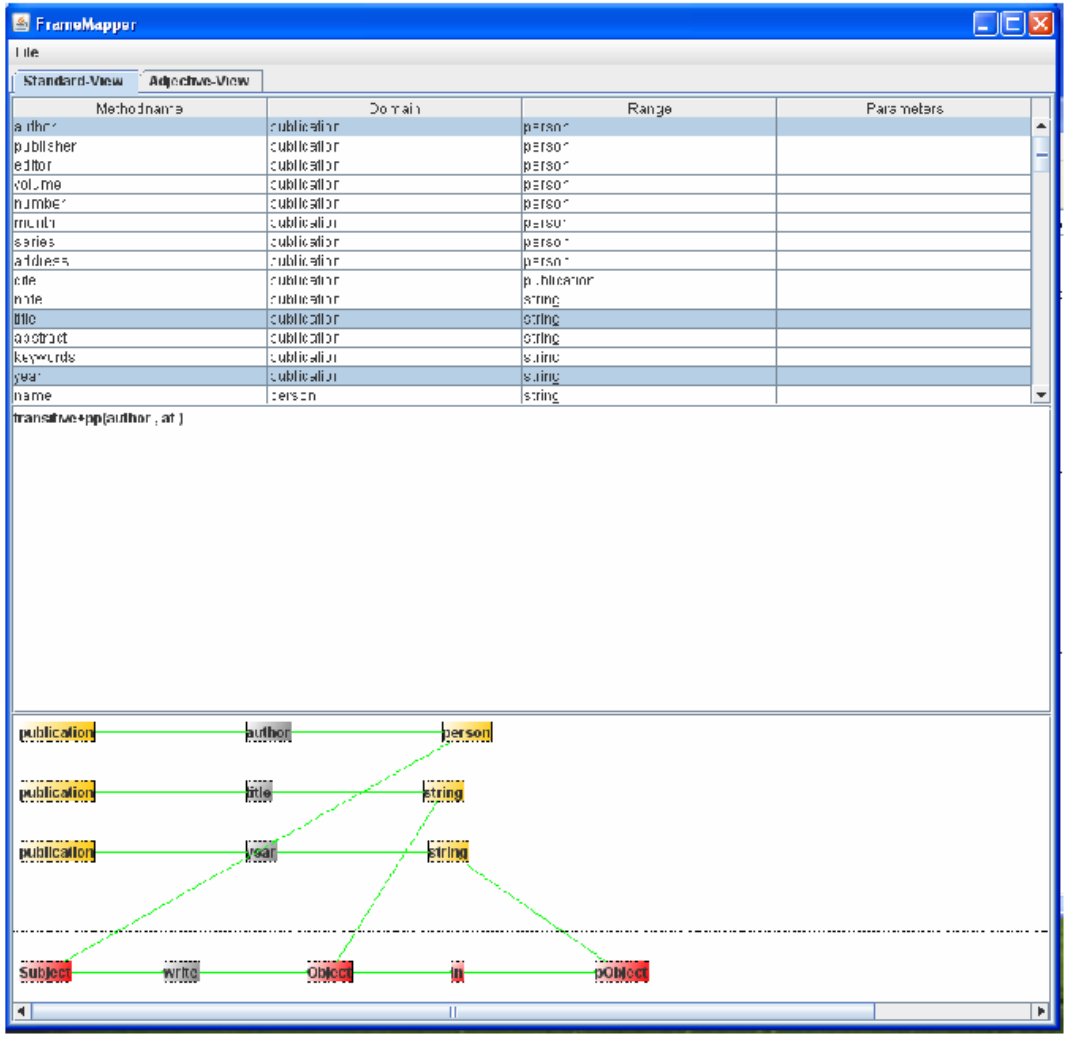
ORAKEL (Cimiano et al. 2007)

- More sophisticated parsing strategy: Logical Description Grammars (LDGs).
- LDG is inspired by Lexicalized Tree Adjoining Grammars (LTAGs).
- An important characteristic of these trees is that they encapsulate all syntactic/semantic arguments of a word.

ORAKEL (Cimiano et al. 2007)



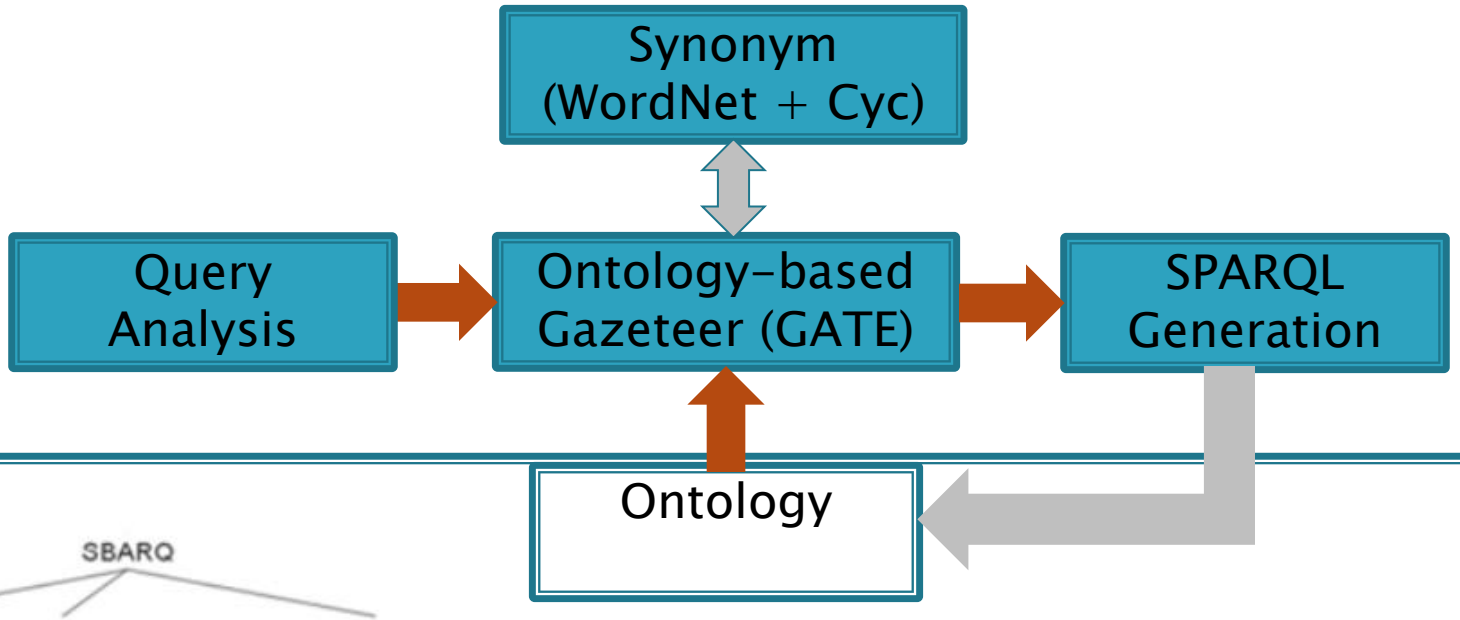
ORAKEL (Cimiano et al. 2007)

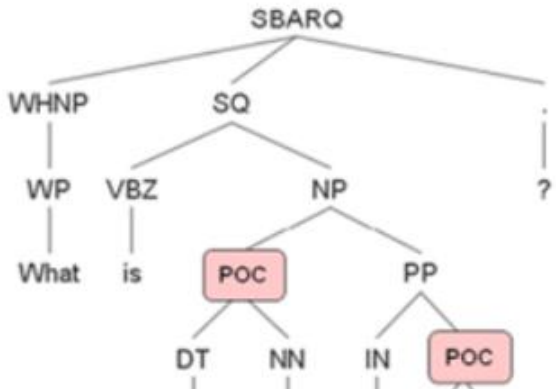


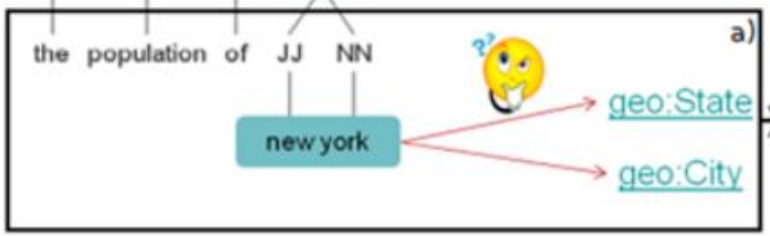
Freya (Damljanovic et al. 2010)

- Key contribution: user interaction (dialogs) for terminological matching.
- Terminological Matching:
  - WordNet & Cyc (synonyms)
  - String similarity (Monge Elkan + Sundex)
  - User feedback
- Extends the QuestIO system.
- Evaluation: Mooney (2010), QALD (2011)

Freya (Damljanovic et al. 2010)

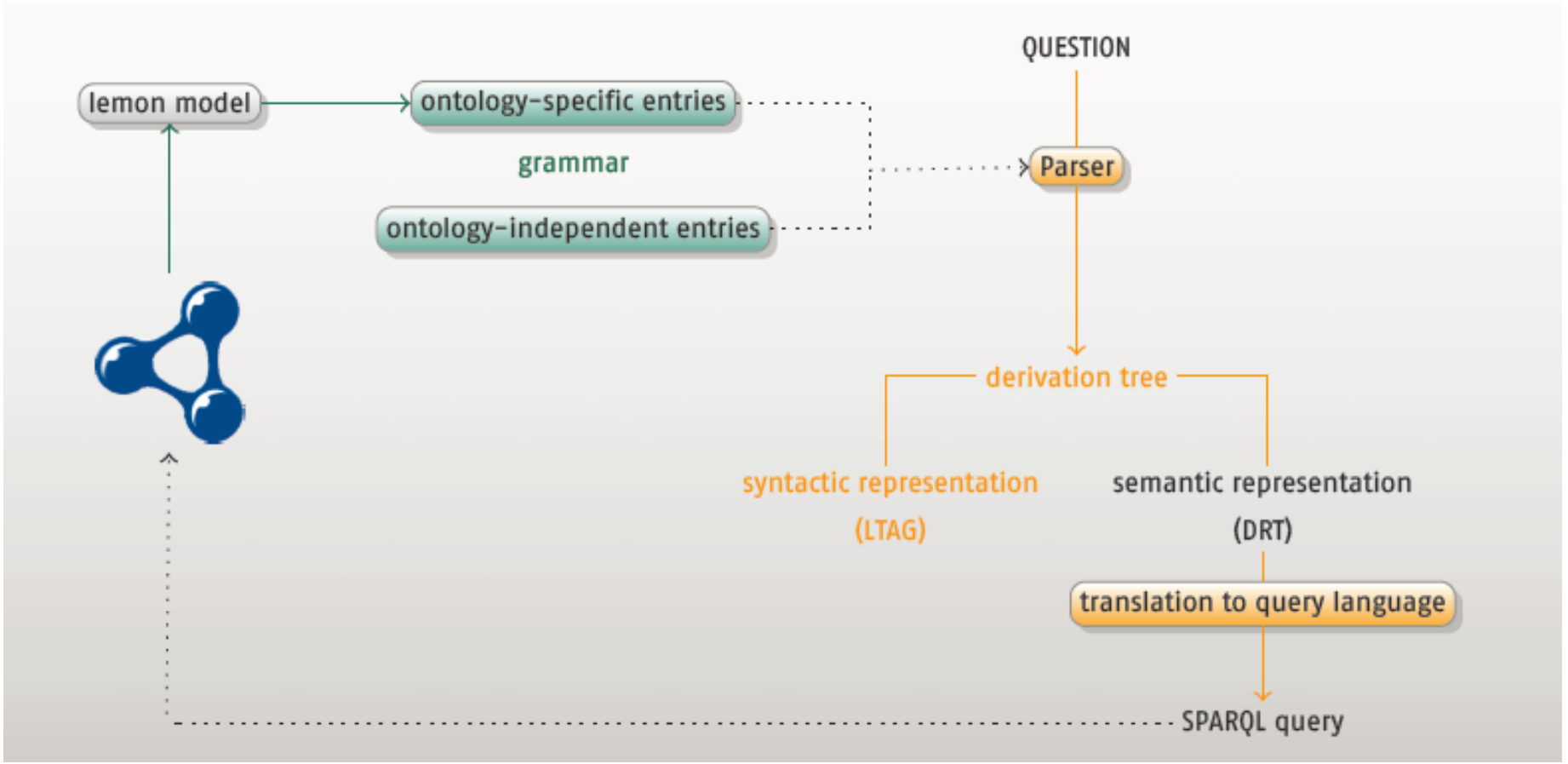






Pythia: Detailed Case Study

Architecture



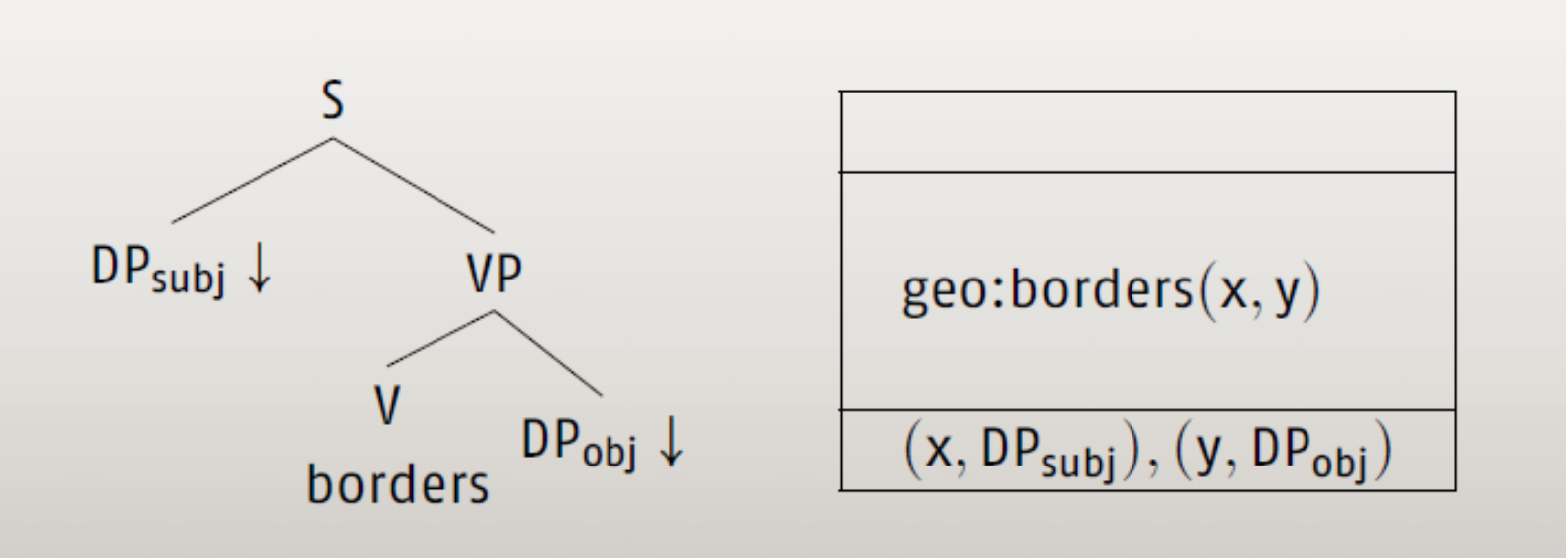
Grammar

The grammar consists of two parts:

- Ontology-specific entries
  - content words and phrases corresponding to concepts and properties in the ontology
    - automatically generated from lemon model
- Ontology-independent entries
  - function words
    - quantifiers (some, every, two)
    - wh-words (who, when, where, which, how many)
    - negation (not)
  - manually specified and re-usable for all domains

Grammar entries

- Grammar entries are pairs of
  - a syntactic representation (LTAG tree)
  - a semantic representation (DUDE)

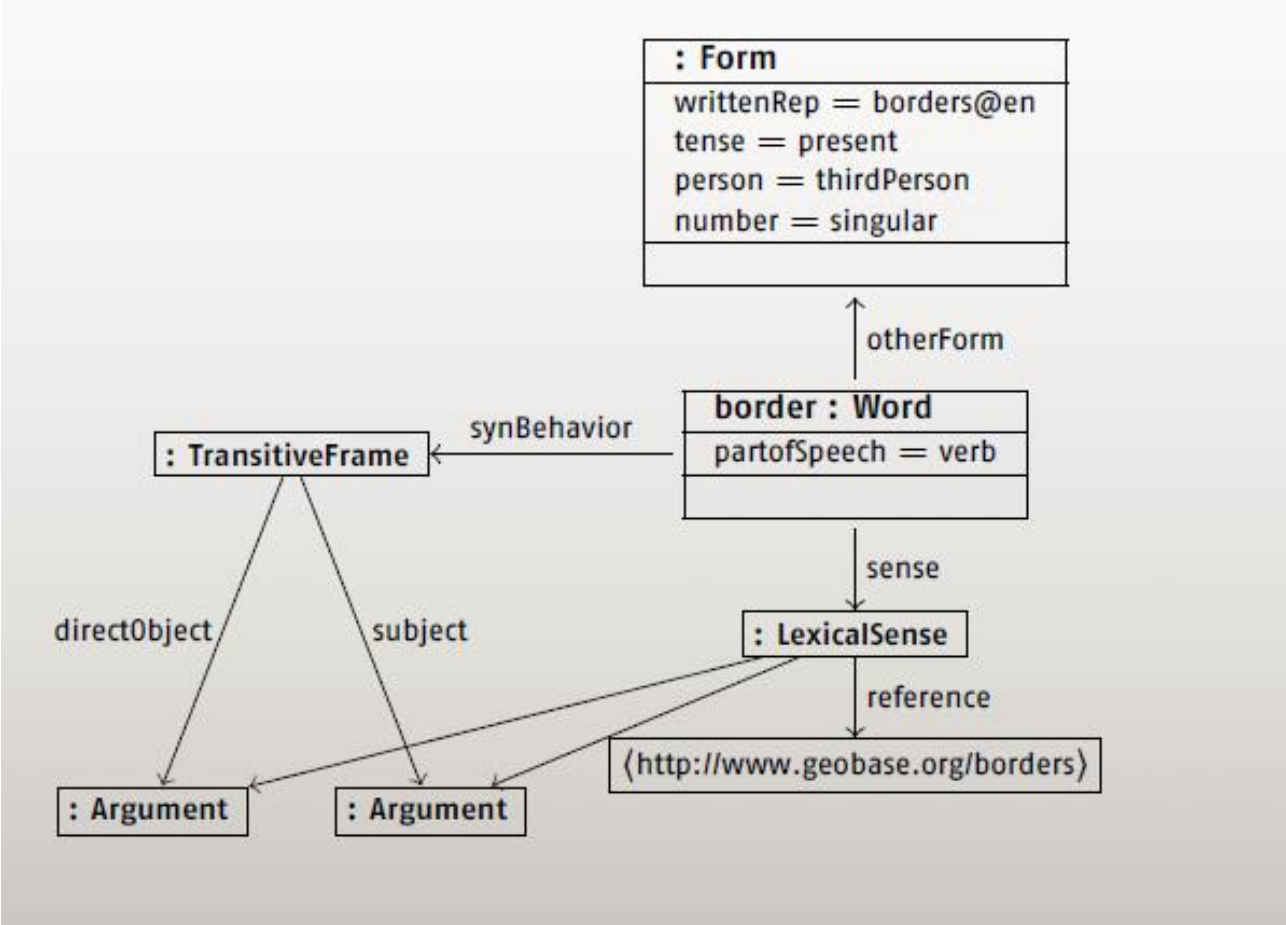


Grammar generation

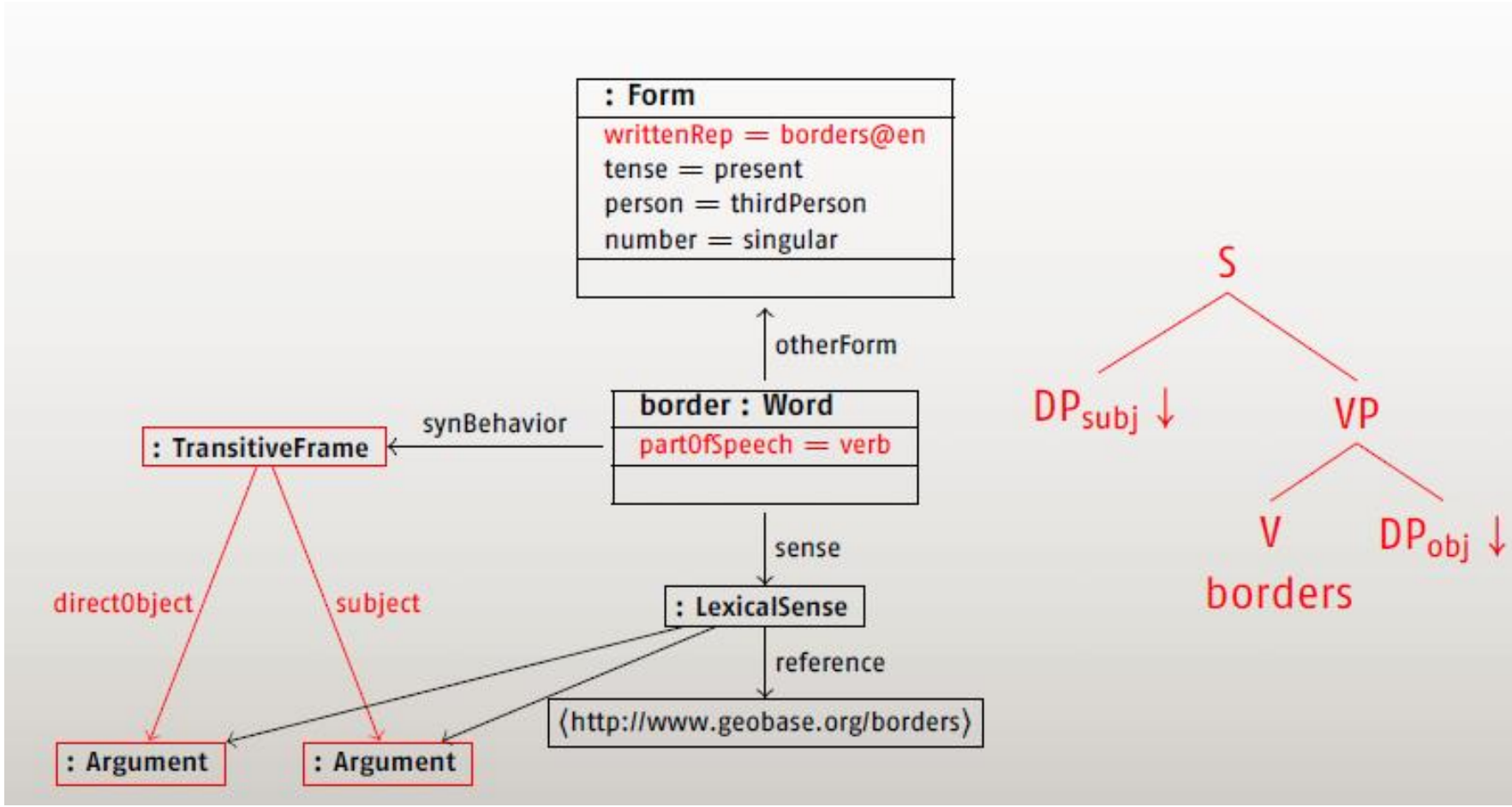
- Step 1: Manual construction of a lemon model for the ontology
  - i.e. specifying how concepts and relations are verbalized
  - (syntactic category, morphological features, required arguments,semantic properties)

- Step 2: Automatic grammar generation
  - by means of general methods that create grammar entries from lemon entries

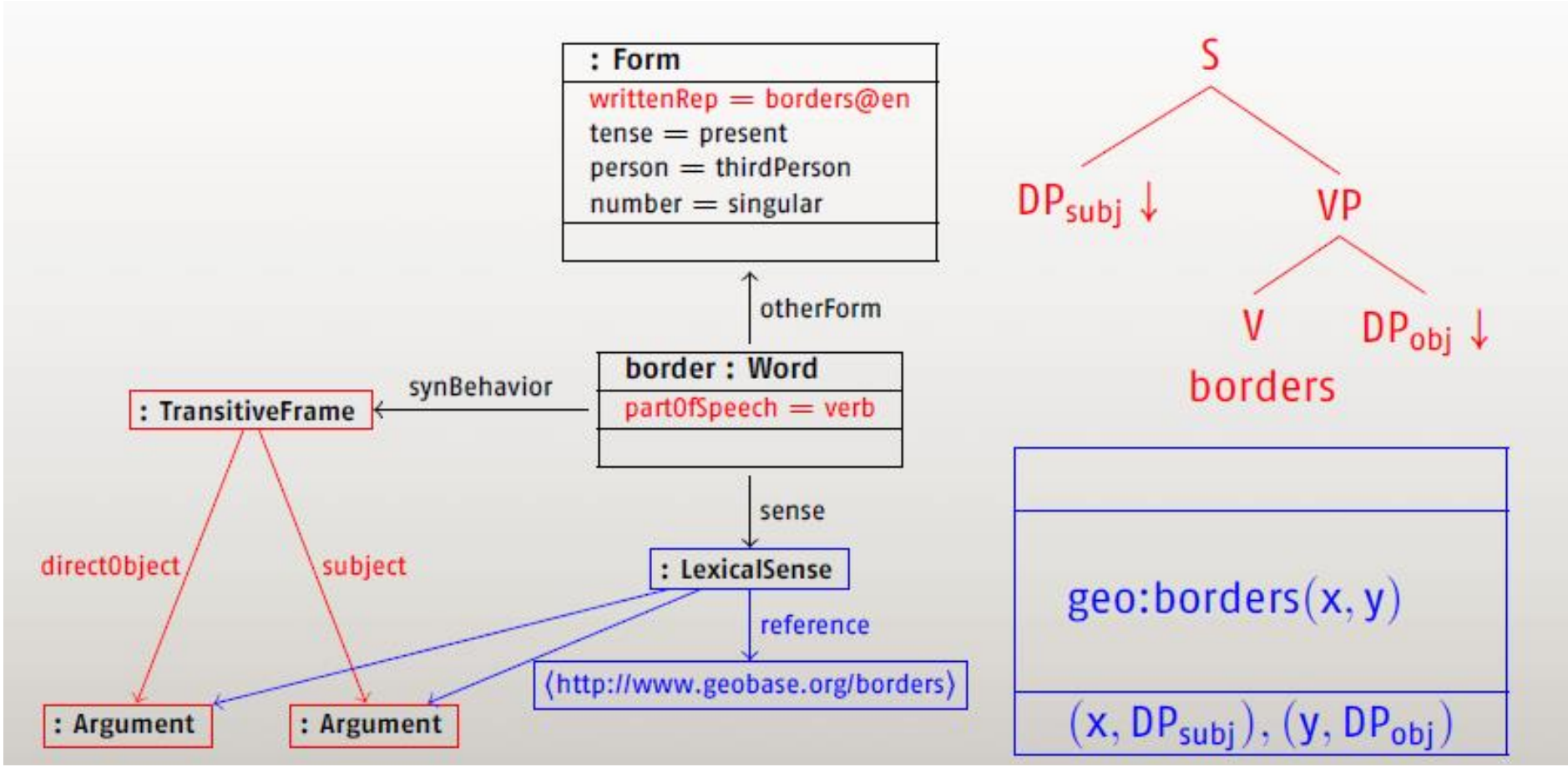
Example: ‘borders’



Example: ‘borders’

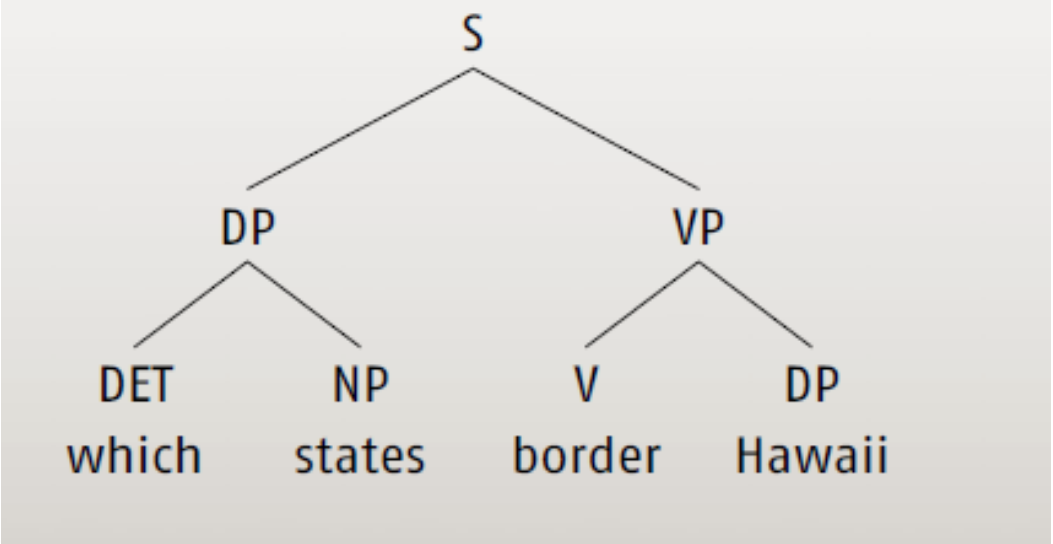


Example: ‘borders’



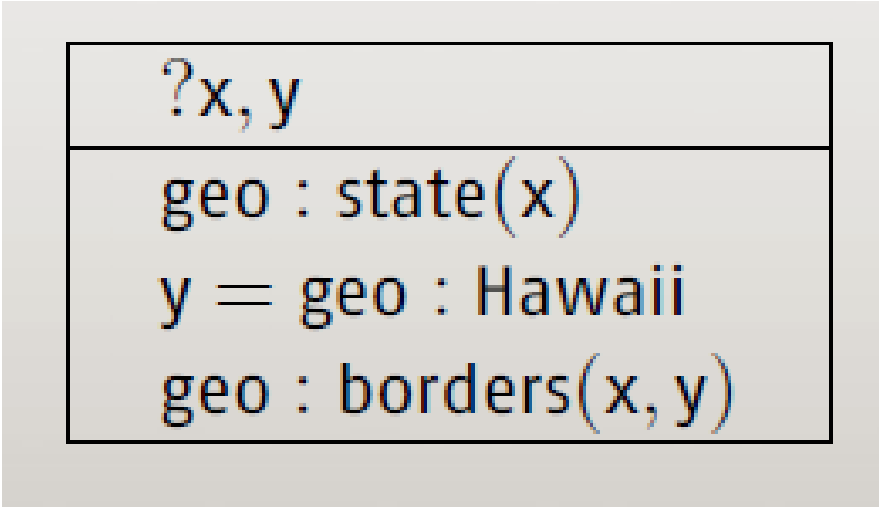
Parsing & interpretation

- Parsing along the lines of the Earley-type parser devised by Schabes & Joshi (1988) – the result is an LTAG derivation tree.
- Next, syntactic and semantic composition rules apply in tandem in order to construct an LTAG derived tree…



Parsing & interpretation

- Parsing along the lines of the Earley-type parser devised by Schabes & Joshi (1988) – the result is an LTAG derivation tree.
- Next, syntactic and semantic composition rules apply in tandem in order to construct an LTAG derived tree…
- …together with an according Discourse Representation Structure



Parsing & interpretation

- Parsing along the lines of the Earley-type parser devised by Schabes & Joshi (1988) – the result is an LTAG derivation tree.
- Next, syntactic and semantic composition rules apply in tandem in order to construct an LTAG derived tree…
- …together with an according Discourse Representation Structure
- …which is then translated into a SPARQL query

```
PREFIX geo:  <http://www.geobase.org/>
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?x
WHERE  {   ?x rdf:type geo:state .
           ?x geo:borders geo:hawaii . }
```

Treo: Detailed Case Study

Tréo

Treo (Irish): Direction, path

Treo (Freitas et al. 2011, 2012)

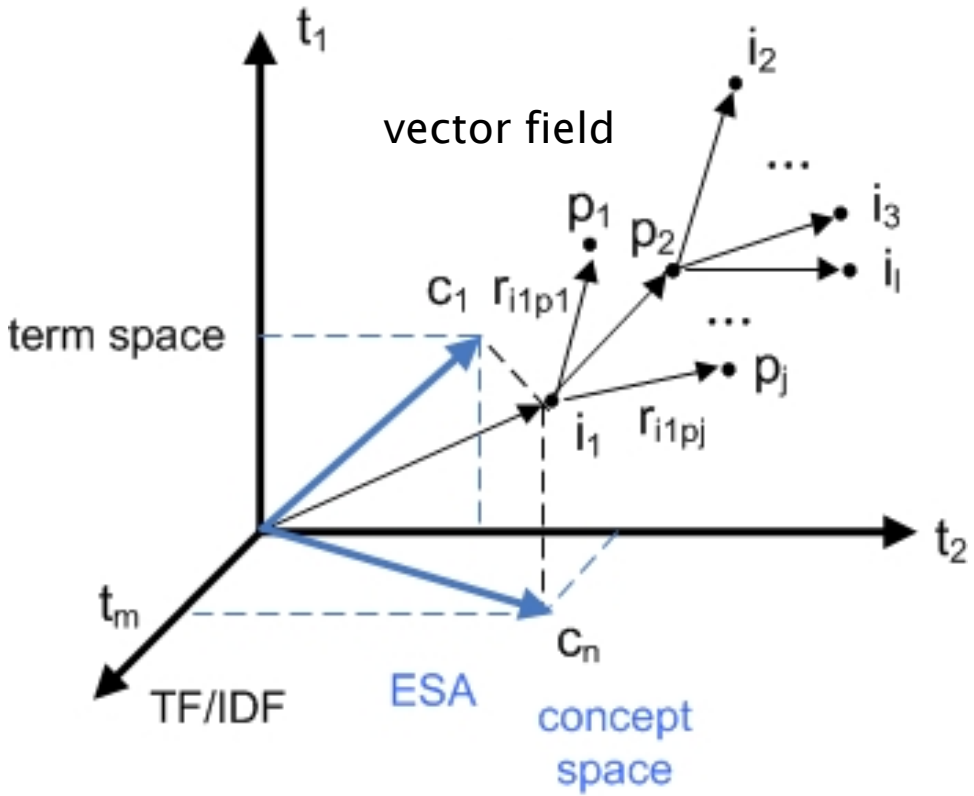
- Key contribution:
  - Distributional semantic relatedness matching model
  - Distributional relational space.
- Terminological Matching:
  - Explicit Semantic Analysis (ESA)
  - WordNet
  - String similarity + node cardinality
- Evaluation: QALD (2011)

Core Elements of Treo

- Hybrid model: database/IR/QA.
- Ranked query results.
- A distributional VSM for representing and semantically processing relational data was formulated: T-Space.
  - Similar in motivation to Cohen's predication space.
- Distributional semantic relatedness as a primitive operation.

T-Space

ESA


t1
i2
vector field
...
i3
p1
p2
il
c1
ri1p1
term space
...
pj
i1
ri1pj
t2
tm
cn
TF/IDF
ESA
concept
space

NL Query: Who is the daughter of Bill Clinton married to ?

Results

Chelsea Clinton's spouse is Marc Mezvinsky

Bill Clinton's daughter is Chelsea Clinton

Query Pre-Processing (Question Analysis)

- Transform natural language queries into triple patterns

"Who is the daughter of Bill Clinton married to?"

Query Pre-Processing (Question Analysis)

- Step 1: POS Tagging
  - Who/WP
  - is/VBZ
  - the/DT
  - daughter/NN
  - of/IN
  - Bill/NNP
  - Clinton/NNP
  - married/VBN
  - to/TO
  - ?/.

Query Pre-Processing (Question Analysis)

- Step 2: Core Entity Recognition
  - Rules-based: POS Tag + TF/IDF

Who is the daughter of Bill Clinton married to?

(PROBABLY AN INSTANCE)

Query Pre-Processing (Question Analysis)

- Step 3: Determine answer type
  - Rules-based.

Who is the daughter of Bill Clinton married to?

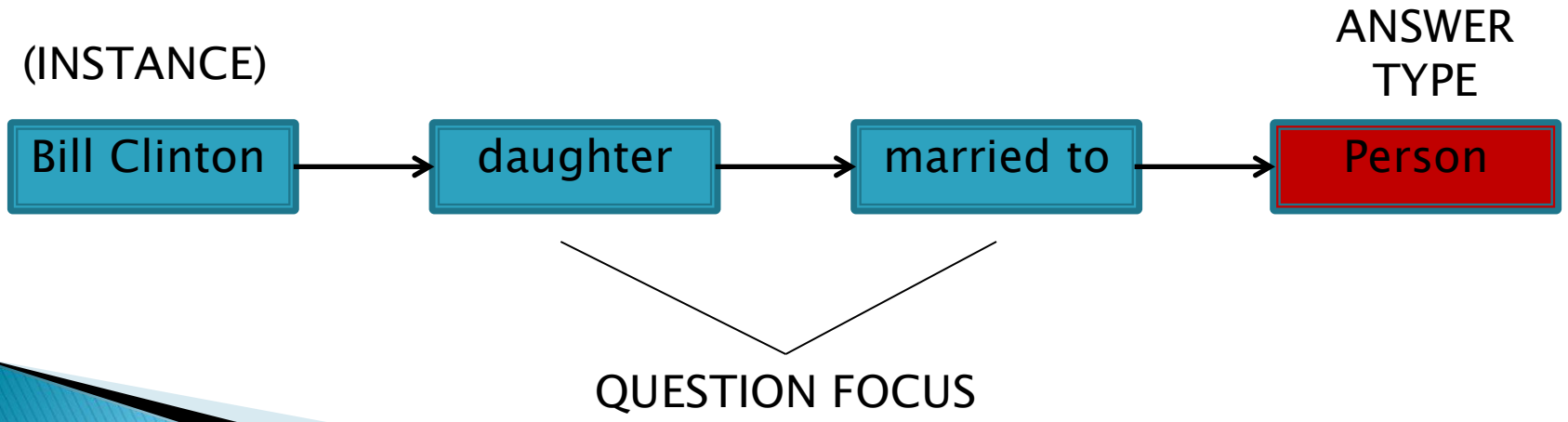
(PERSON)

Query Pre-Processing (Question Analysis)

- Step 4: Dependency parsing
  - dep(married-8, Who-1)
  - auxpass(married-8, is-2)
  - det(daughter-4, the-3)
  - nsubjpass(married-8, daughter-4)
  - prep(daughter-4, of-5)
  - nn(Clinton-7, Bill-6)
  - pobj(of-5, Clinton-7)
  - root(ROOT-0, married-8)
  - xcomp(married-8, to-9)

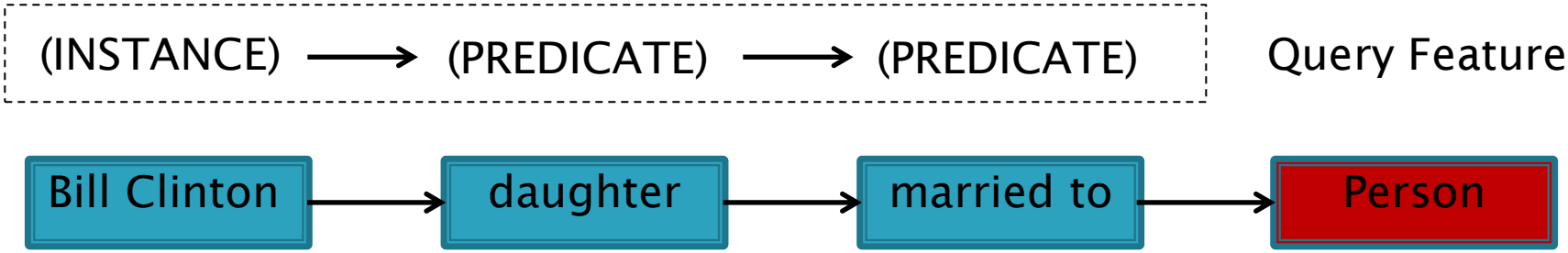
Query Pre-Processing (Question Analysis)

- Step 5: Determine Partial Ordered Dependency Structure (PODS)
  - Rules based.
    - Remove stop words.
    - Merge words into entities.
    - Reorder structure from core entity position.



Query Pre-Processing (Question Analysis)

- Step 5: Determine Partial Ordered Dependency Structure (PODS)
  - Rules based.
    - Remove stop words.
    - Merge words into entities.
    - Reorder structure from core entity position.



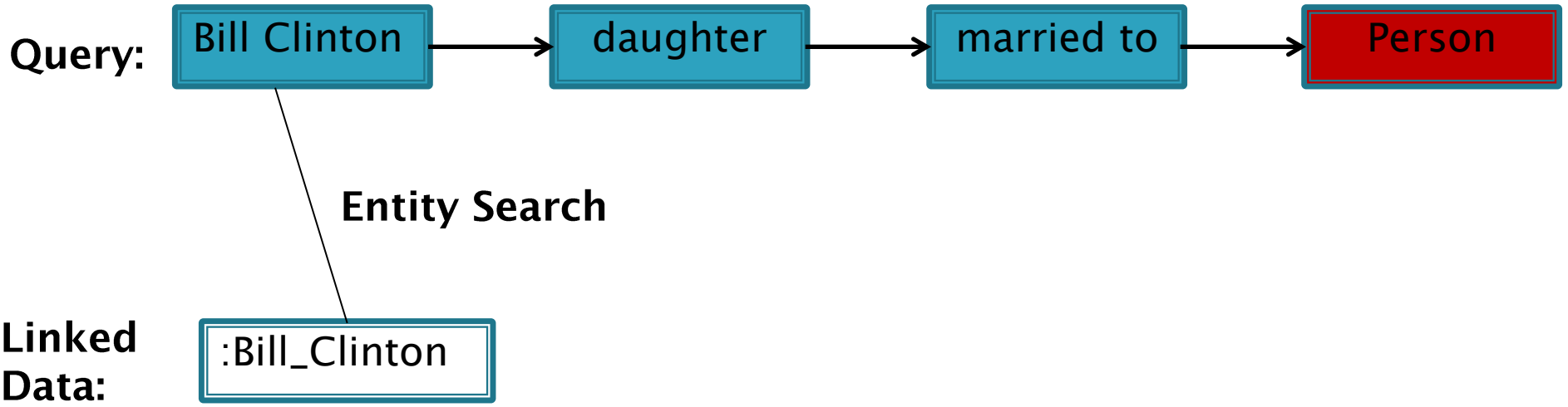
Query Planning

- Map *query features* into a *query plan*.
- A *query plan* contains a sequence of:
  - Search operations.
  - Navigation operations.

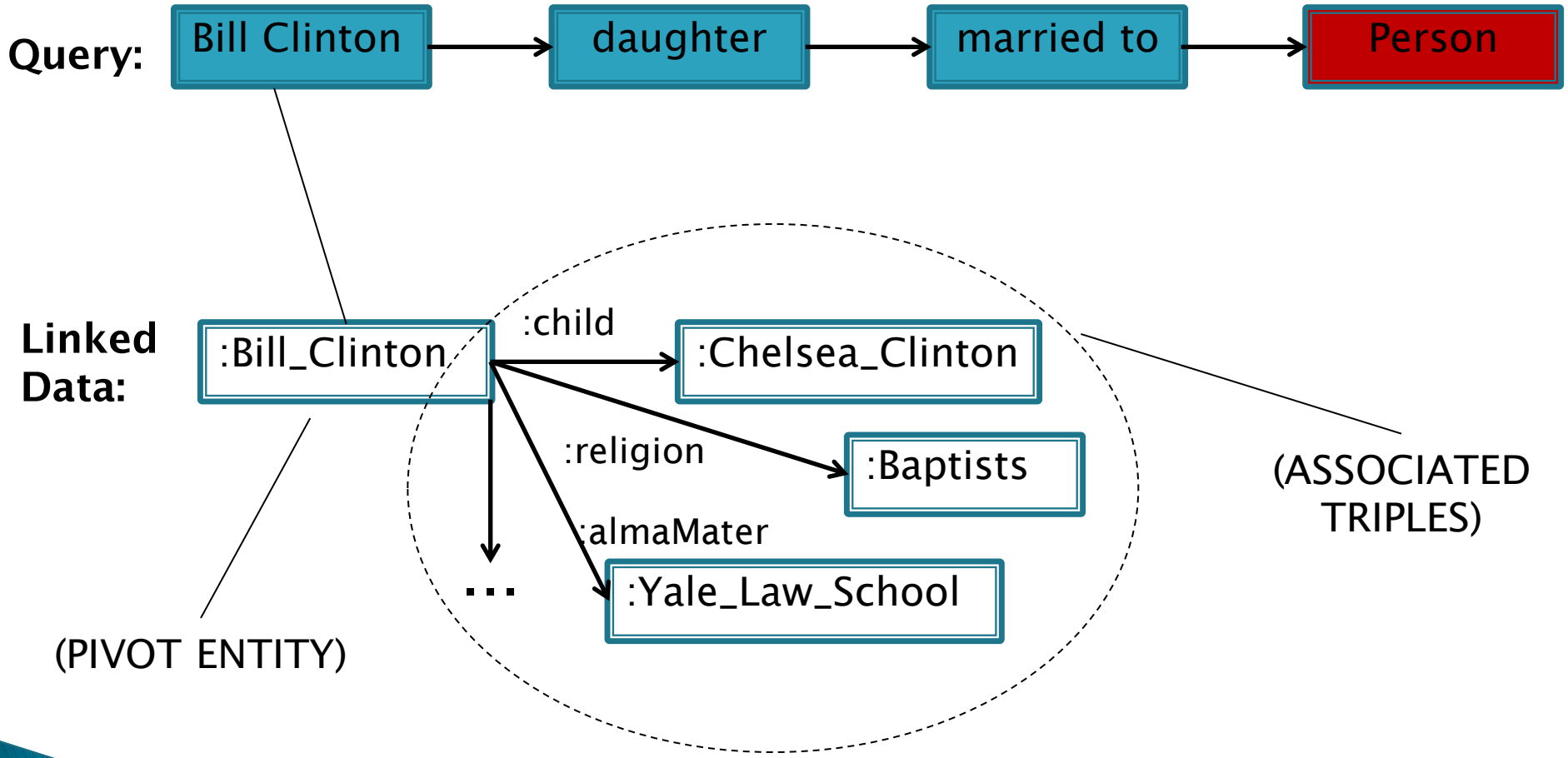
(INSTANCE) ⟶ (PREDICATE) ⟶ (PREDICATE) Query Features

- (1) INSTANCE SEARCH (Bill Clinton)
- (2) DISAMBIGUATE ENTITY TYPE
- (3) GENERATE ENTITY FACETS
- (4) p_1 <- SEARCH RELATED PREDICATE (Bill Clintion, daughter) Query Plan
- (5) e_1 <- GET ASSOCIATED ENTITIES (Bill Clintion, p_1)
- (6) p_2 <- SEARCH RELATED PREDICATE (e_1, married to)
- (7) e_2 <- GET ASSOCIATED ENTITIES (e_1, p_2)
- (8) POST PROCESS (Bill Clintion, e_1, p_1, e_2, p_2)

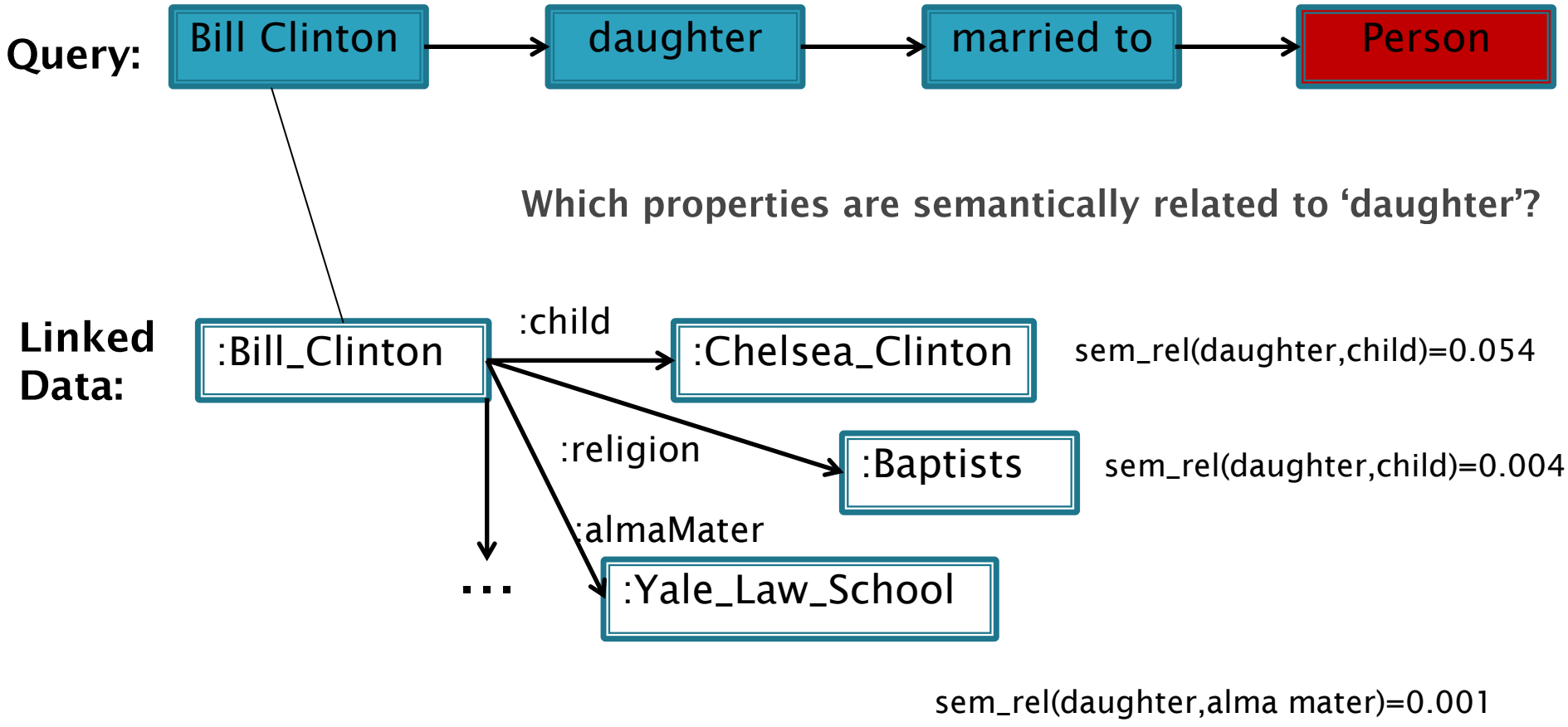
Core Entity Search



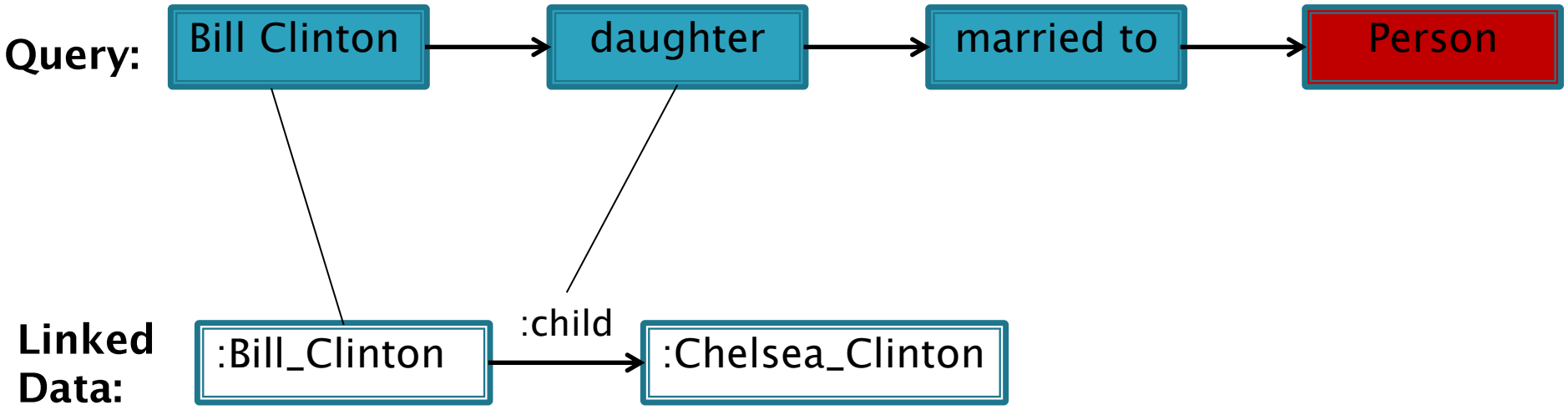
Distributional Semantic Search



Distributional Semantic Search



Distributional Semantic Search



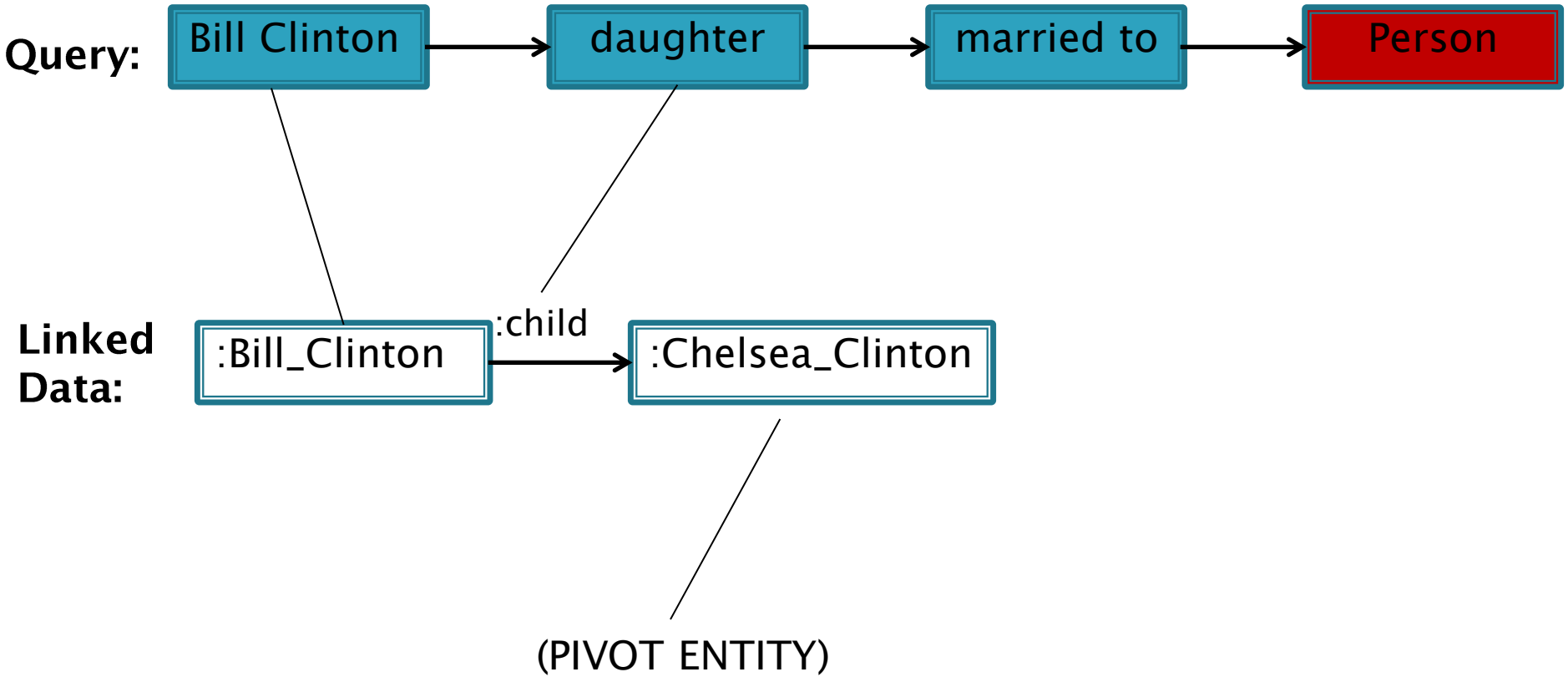
Distributional Semantic Relatedness

- Computation of a measure of "semantic proximity" between two terms.
- Allows a semantic approximate matching between *query terms* and *dataset terms.*
- It supports a commonsense reasoning-like behavior based on the knowledge embedded in the corpus.

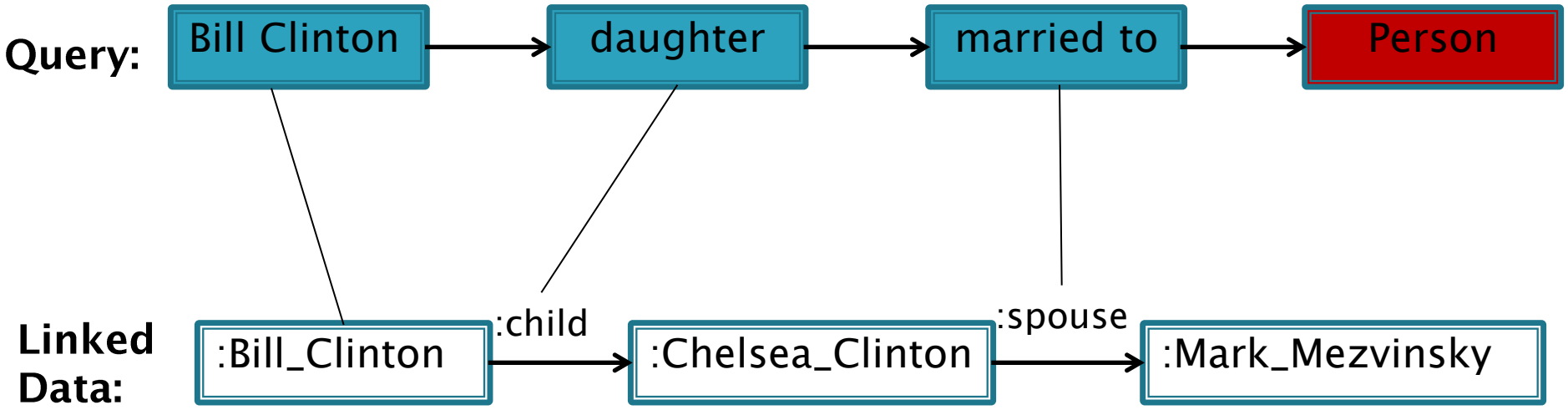
Distributional Semantic Search

- Use distributional semantics to semantically match query terms to predicates and classes.
- Distributional principle: Words that co-occur together tend to have related meaning.
  - Allows the creation of a comprehensive semantic model from unstructured text.
  - Based on statistical patterns over large amounts of text.
  - No human annotations.
- Distributional semantics can be used to compute a semantic relatedness measure between two words.

Distributional Semantic Search



Distributional Semantic Search



Semantic Relatedness

- Computation of a measure of "semantic proximity" between two terms
- Allows a semantic approximate matching between *query terms* and *dataset terms*
- It supports a reasoning-like behavior based on the knowledge embedded in the corpus

Distributional Semantic Search

- Use distributional semantics to semantically match query terms to predicates and classes
- Distributional principle: Words that co-occur together tend to have related meaning
  - Allows the creation of a comprehensive semantic model from unstructured text
  - Based on statistical patterns over large amounts of text
  - No human annotations
- Distributional semantics can be used to compute a semantic relatedness measure between two words

Results

ch Advanced Search Knowledge Base Admin

Tréo Who is the daughter of Bill Clinton married t Search

"Who is the daughter of Bill Clinton married to ?"

Answer

Chelsea Clinton spouse Marc Mezvinsky

Bill Clinton child Chelsea Clinton

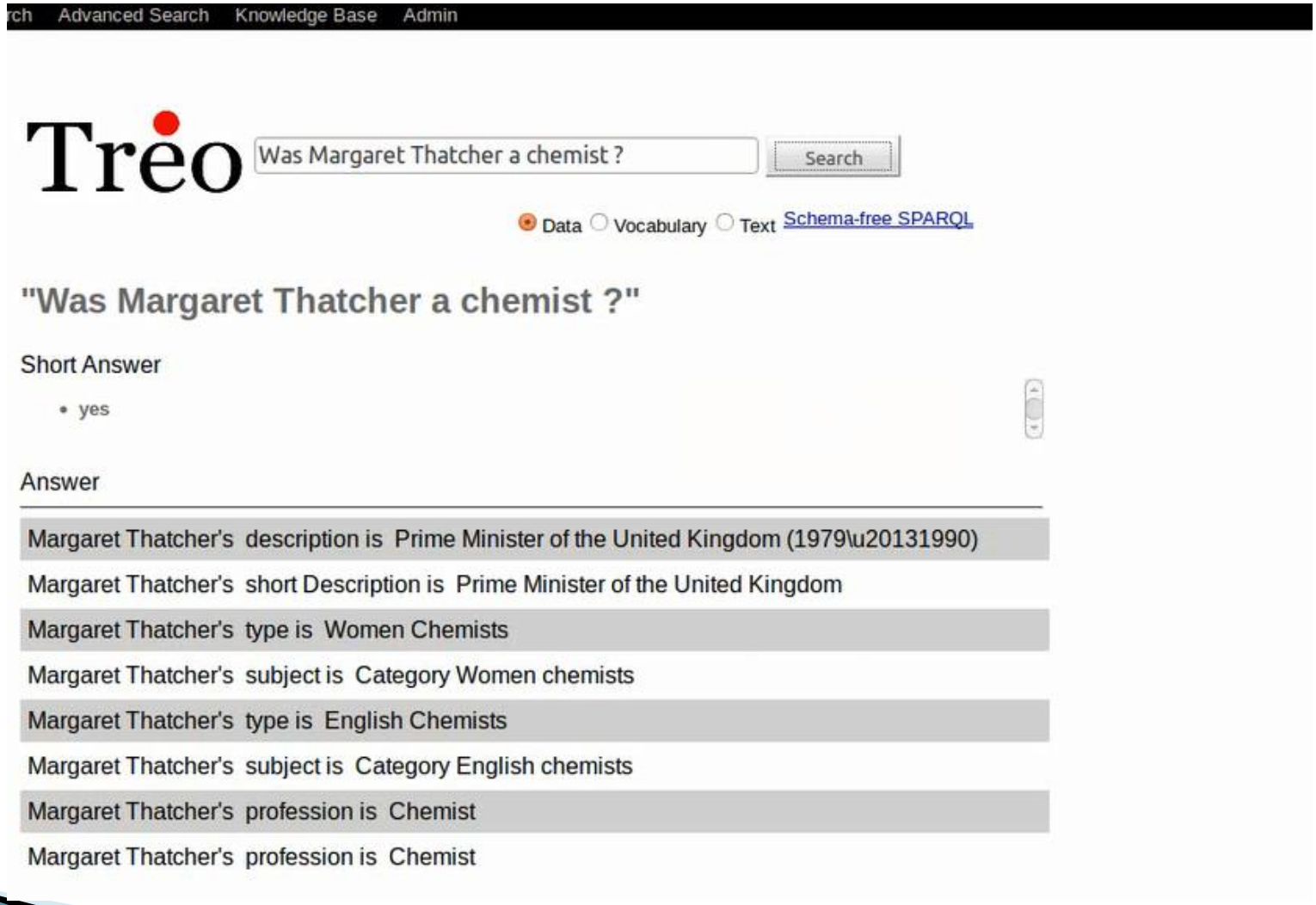
Bill Clinton children Chelsea Clinton

William Jefferson Blythe, Jr. child Bill Clinton

Virginia Clinton Kelley child Bill Clinton

Virginia Clinton Kelley children Bill Clinton

Post-Processing



Second Query Example

What is the highest mountain?

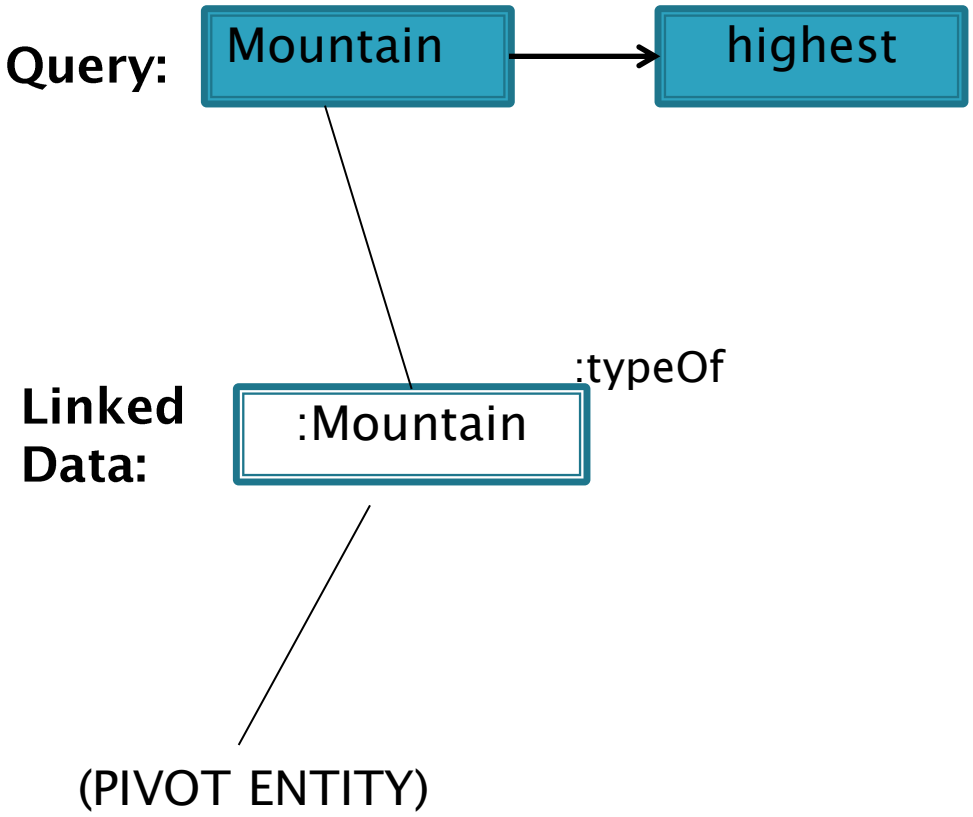
mountain – highest

(CLASS) ⟶ (OPERATOR)

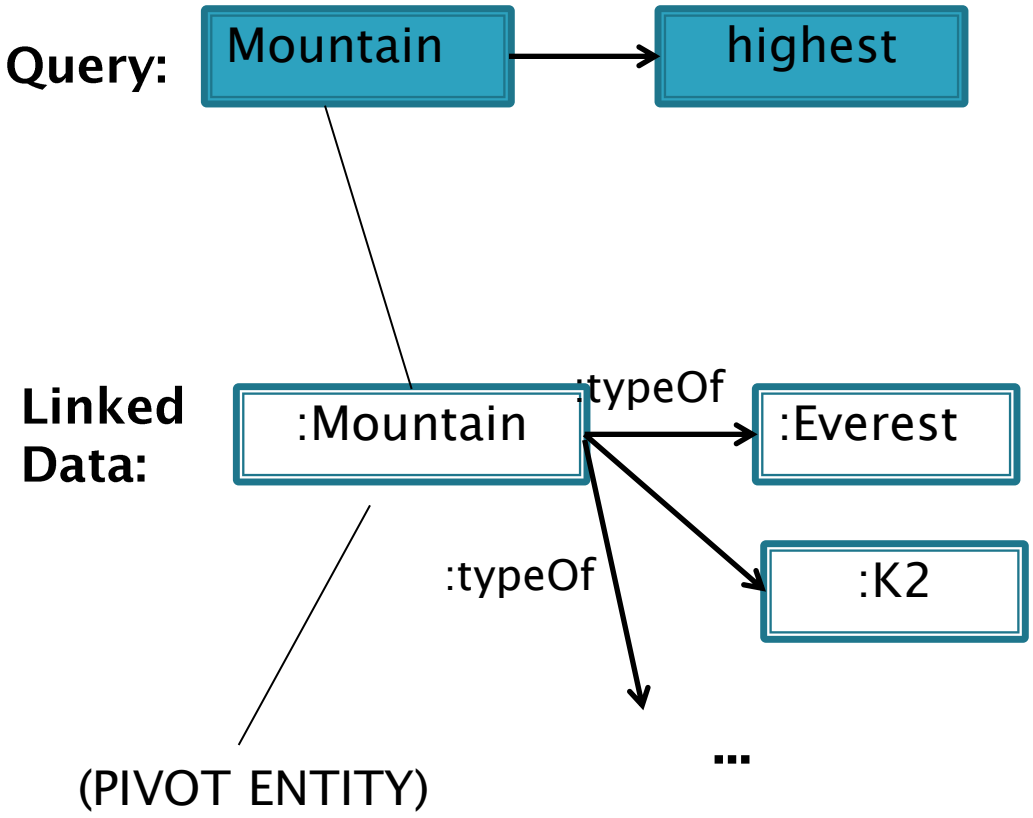
PODS

Query Features

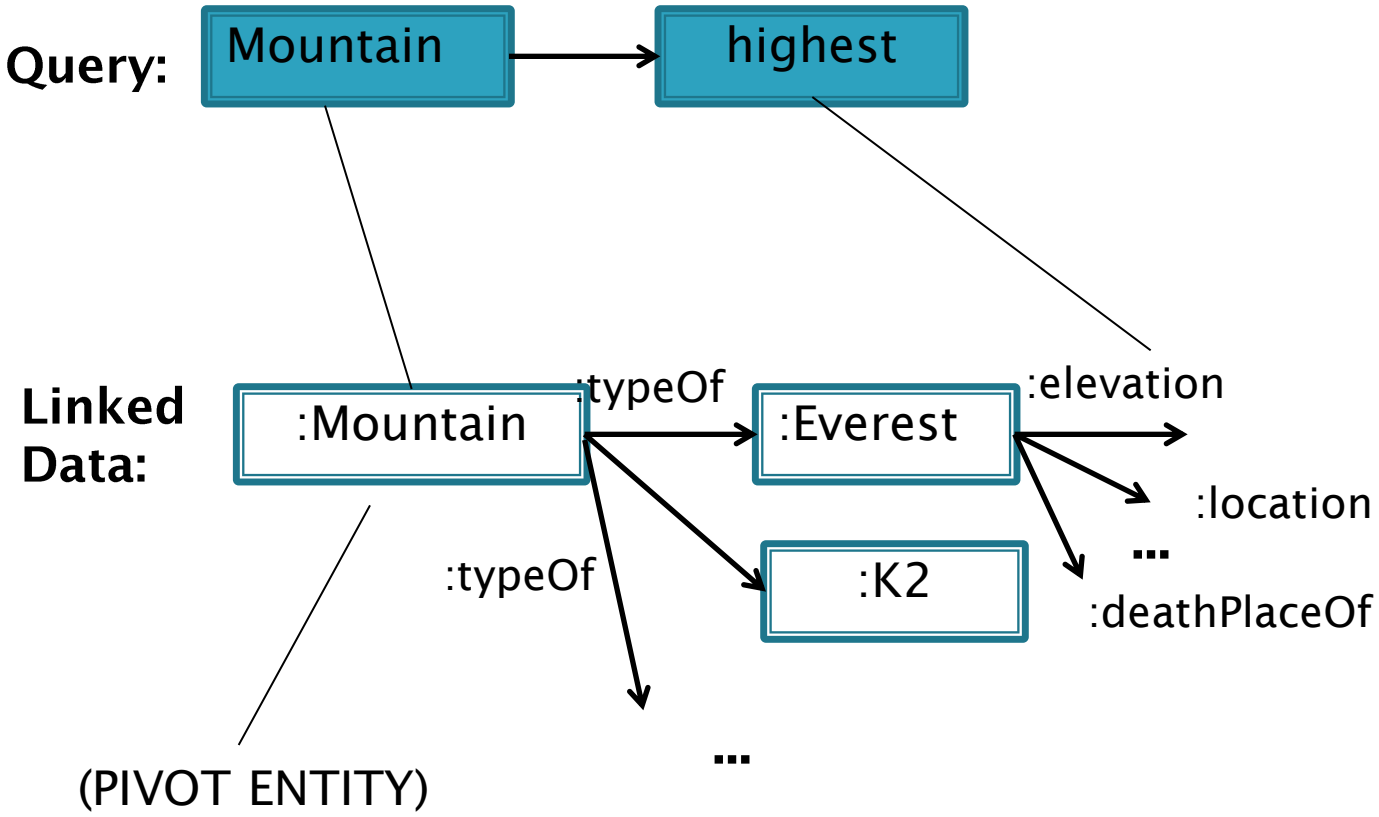
Entity Search



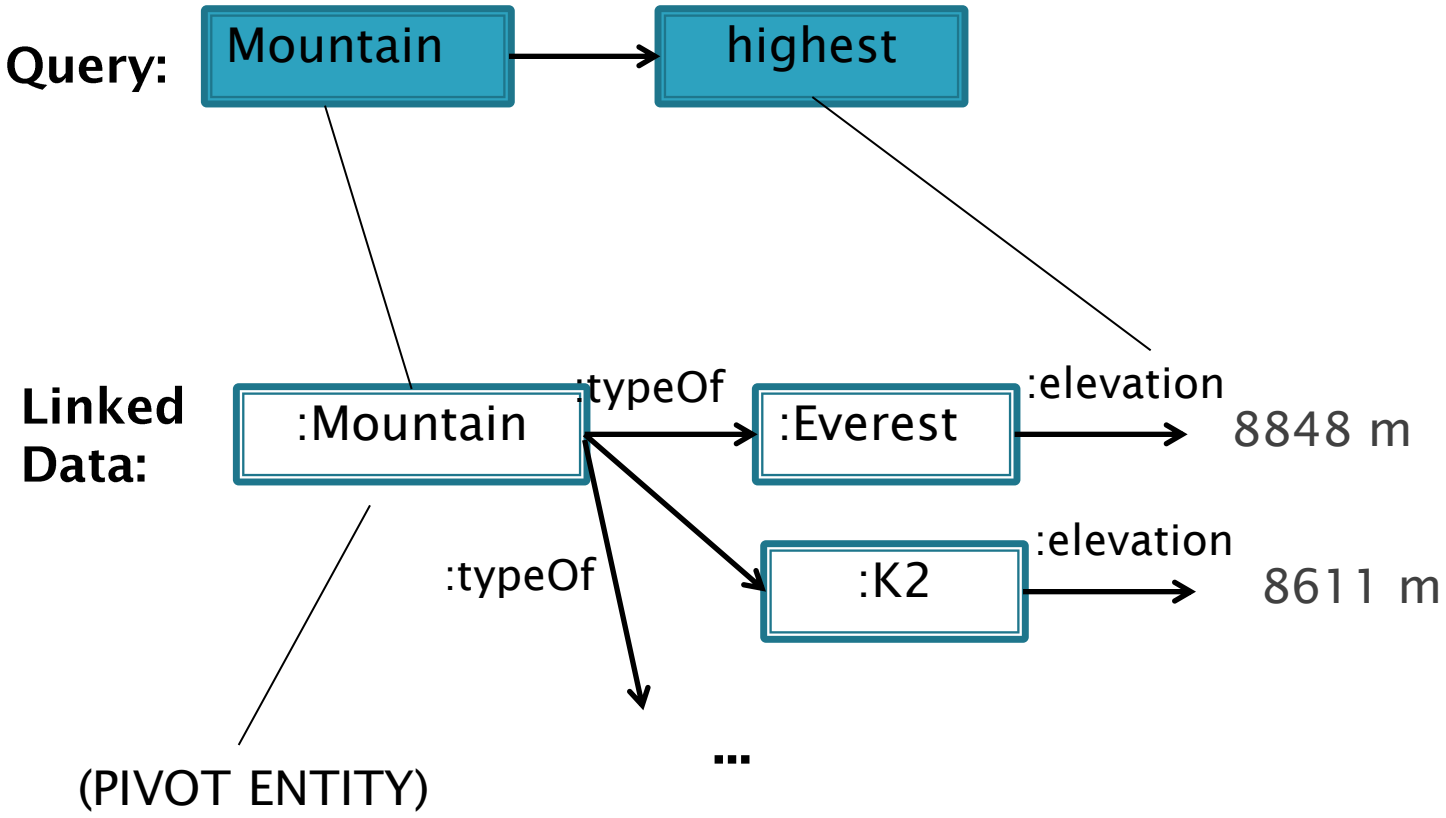
Extensional Expansion



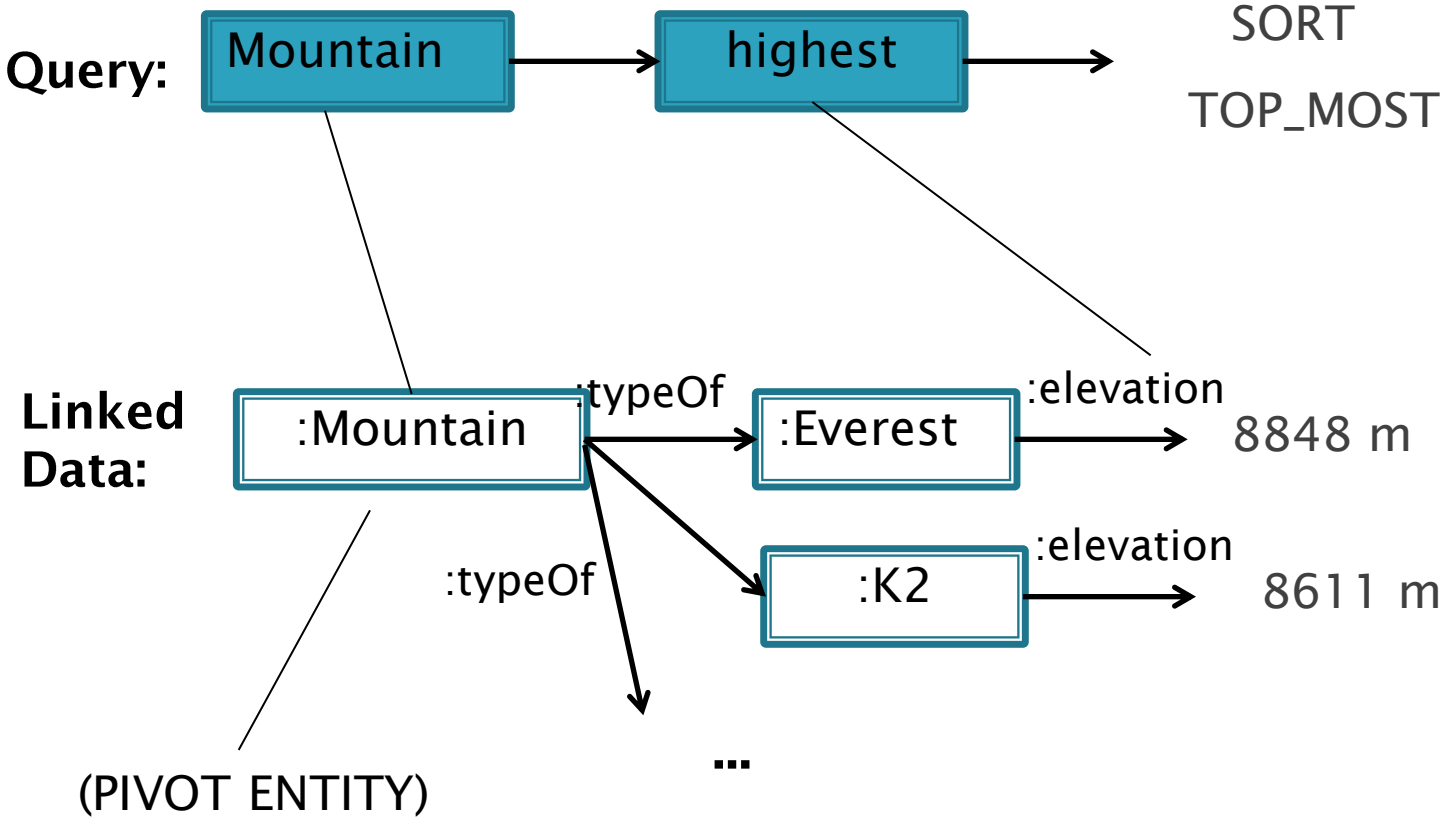
Distributional Semantic Matching



Get all numerical values



Apply operator functional definition



Results

ch Advanced Search Knowledge Base Admin

Treo What is the highest mountain ? Search

"What is the highest mountain ?"

Answer

Mount Everest elevation 8848.0

From Exact to Approximate

- Semantic approximation in databases (as in any IR system): semantic best-effort.
- Need some level of user disambiguation, refinement and feedback.
- As we move in the direction of semantic systems we should expect the need for principled dialog mechanisms (like in human communication).
- Pull the the user interaction back into the system.

From Exact to Approximate

User Feedback

Search Advanced Search Knowledge Base Admin

Treo Give me all actors starring in Batman Begins Search

Data Vocabulary Text Schema-free SPARQL

"Give me all actors starring in Batman Begins ?"

Answer

Batman Begins's starring is Michael Caine

Batman Begins's starring is Liam Neeson

Batman Begins's starring is Katie Holmes

Batman Begins's starring is Gary Oldman

Batman Begins's starring is Cillian Murphy

Batman Begins's starring is Morgan Freeman

Batman Begins's starring is Morgan Freeman

Batman Begins's starring is Cillian Murphy

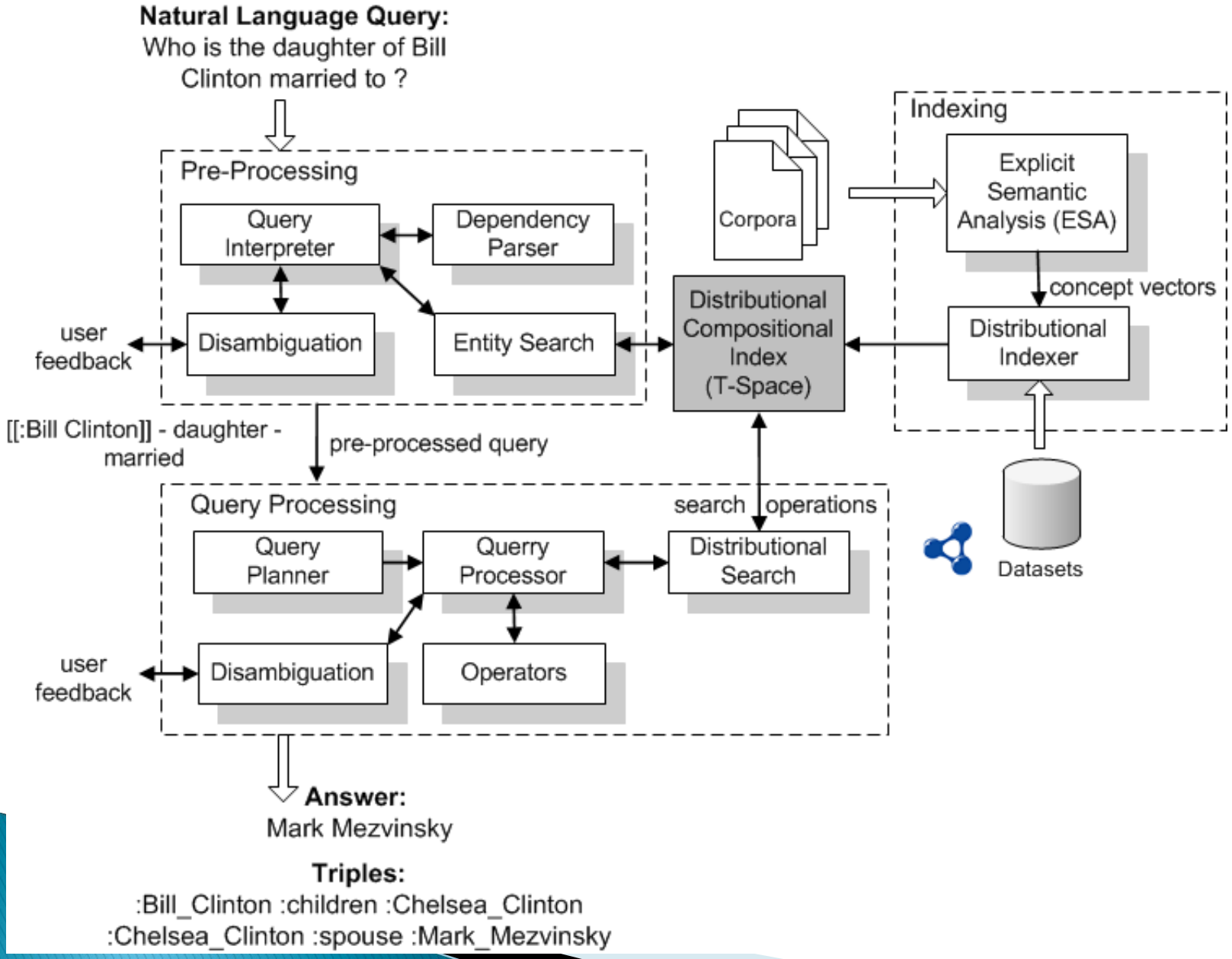
Batman Begins's starring is Gary Oldman

Batman Begins's starring is Michael Caine

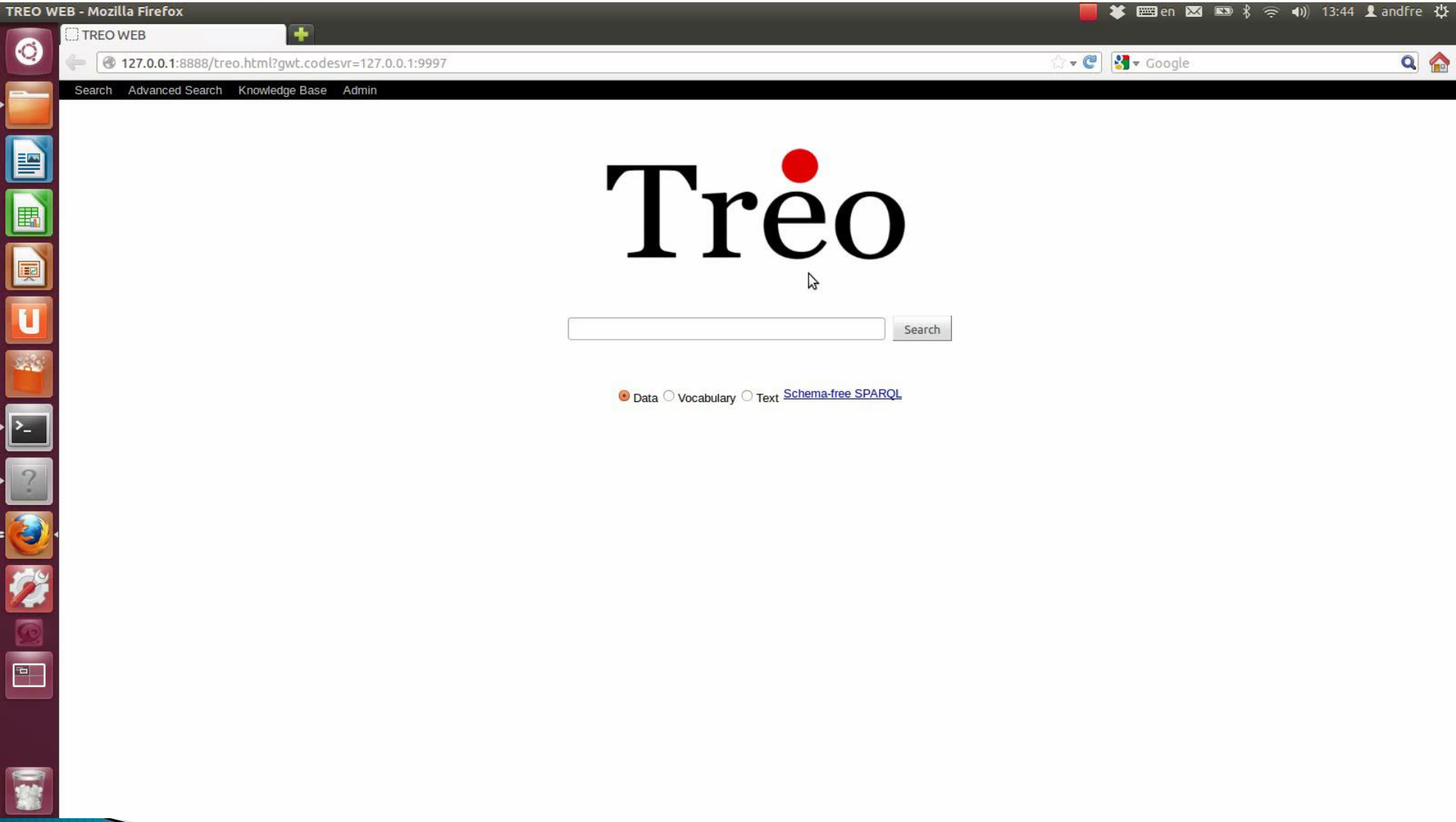
Batman Begins's starring is Christian Bale

~~Christian Bale's type is English Child Actors~~

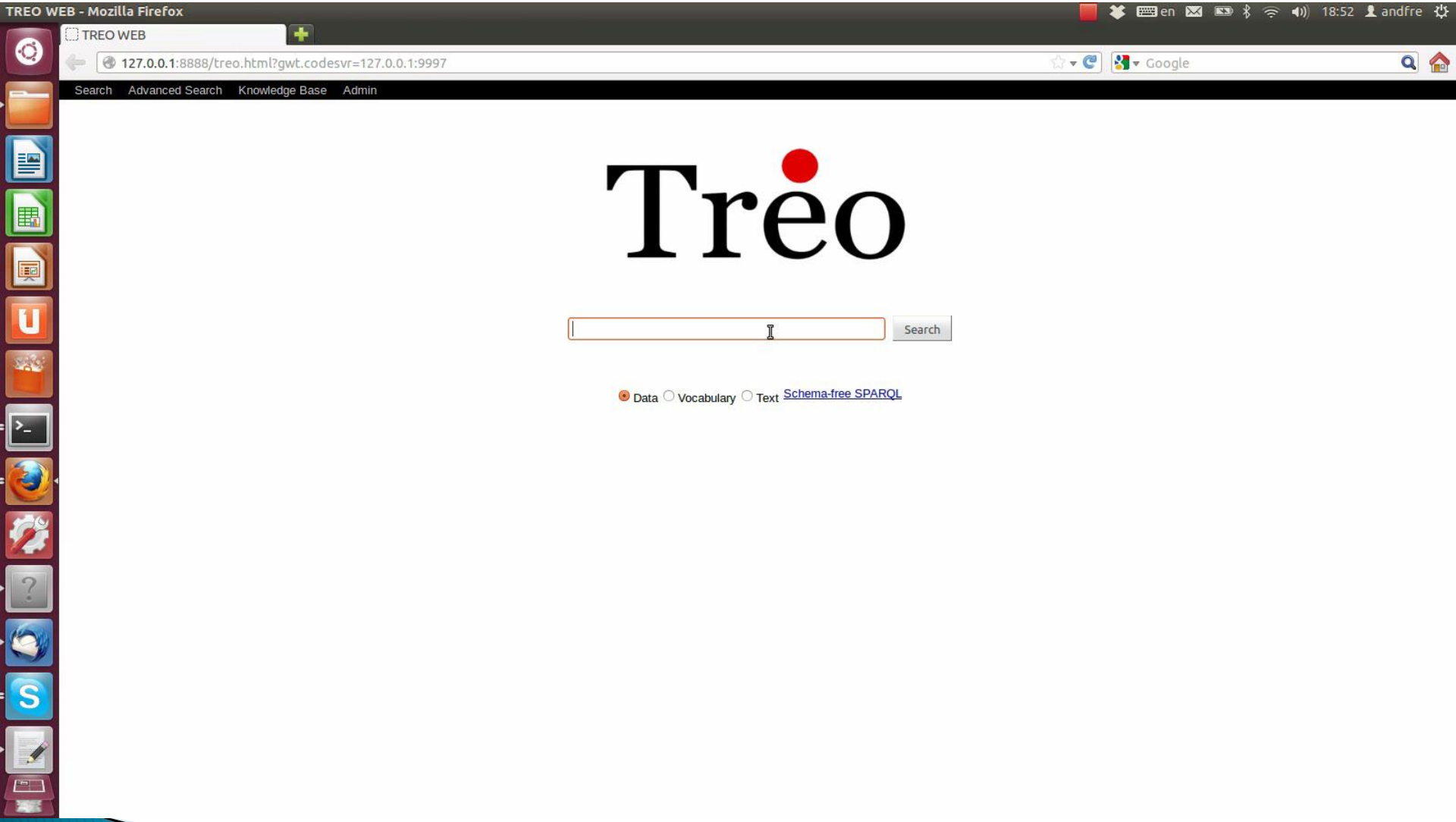
Treo Architecture



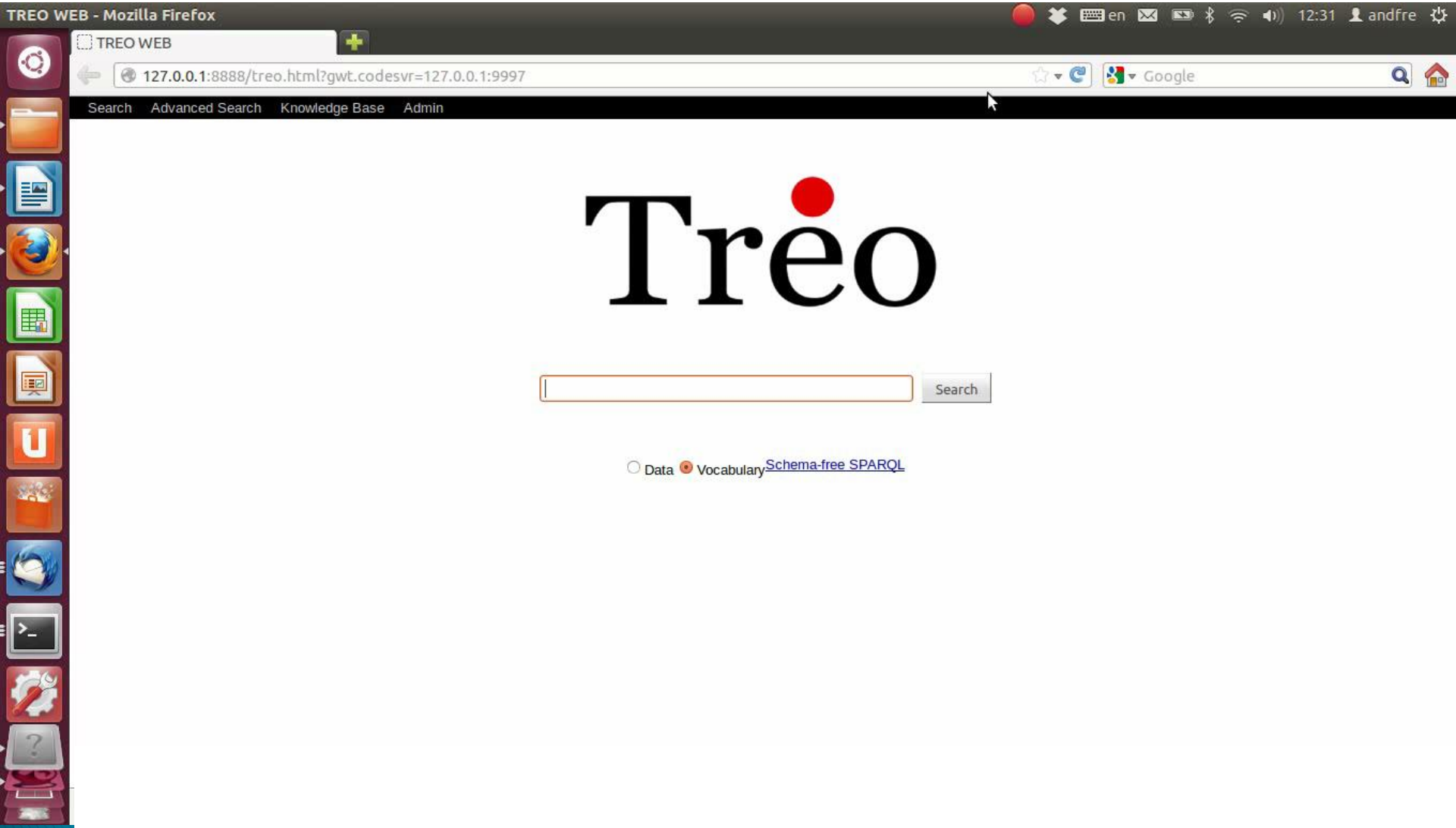
Simple Queries (Video)



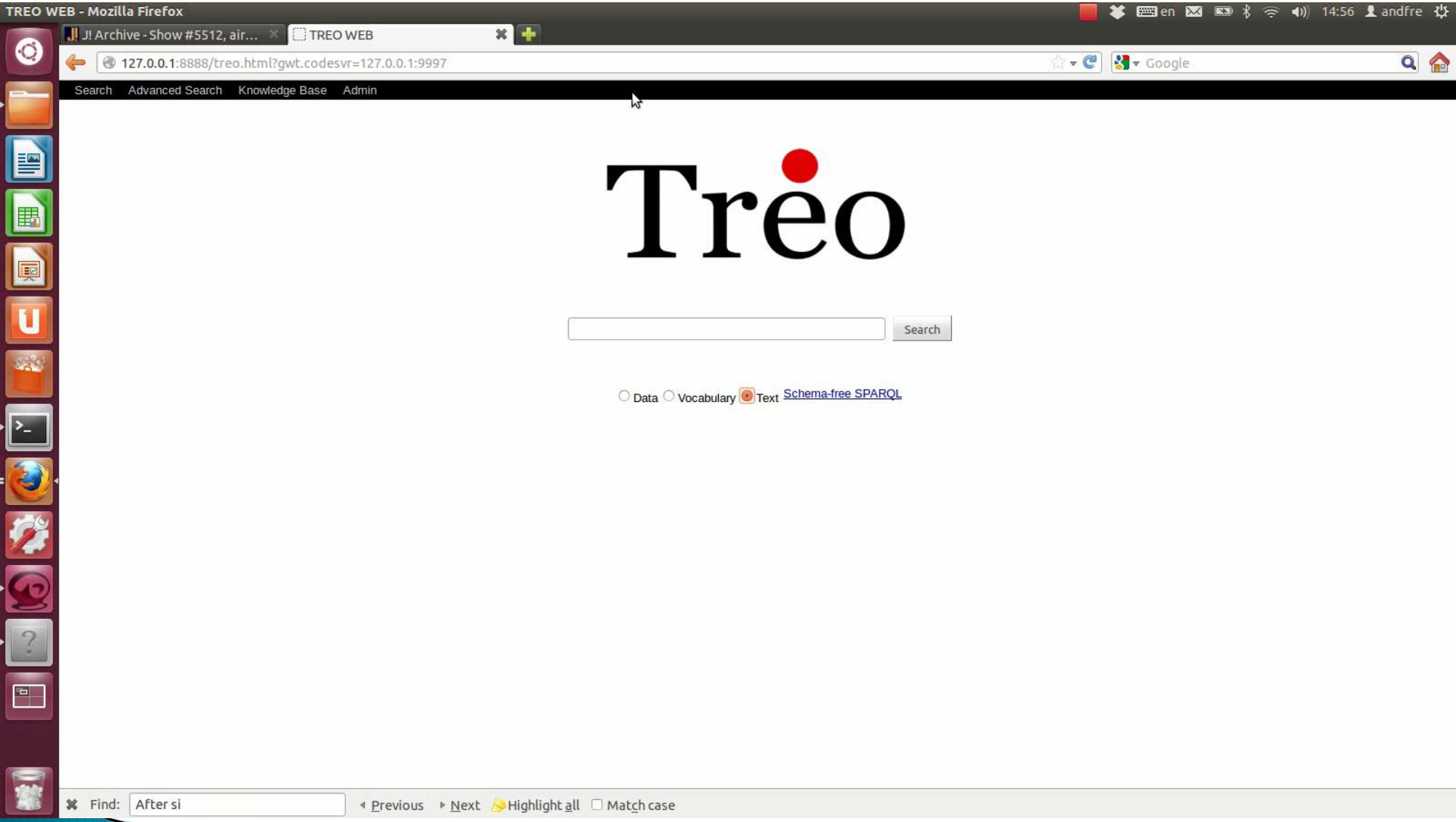
More Complex Queries (Video)



Vocabulary Search (Video)



Treo Answers Jeopardy Queries (Video)



Video Links:

- Introducing Treo: Talk to your Data
  - http://www.youtube.com/watch?v=Zor2X0uoKsM
- Treo: Do it your own (DIY) Jeopardy Question Answering Engine
  - http://www.youtube.com/watch?v=Vqh0r8GxYe8
- Treo: Semantic Search over Schema & Vocabularies
  - http://www.youtube.com/watch?v=HCBwSV1mTdY

Evaluation of QA over Linked Data

Evaluation Campaigns

- Test Collection
  - Questions
  - Datasets
  - Answers (Gold-standard)
- Evaluation Measures

Recall

- Measures how complete is the answer set.
- The fraction of relevant instances that are retrieved.

$$\text{Recall} = \frac{\text{number of correct system answers}}{\text{number of gold standard answers}}$$

- Which are the Jovian planets in the Solar System?
  - Returned Answers:
    - Mercury
    - Jupiter
    - Saturn
  - Gold-standard:
    - Jupiter
    - Saturn
    - Neptune
    - Uranus

Recall = 2/4 = 0.5

Precision

- Measures how accurate is the answer set.
- The fraction of retrieved instances that are relevant.

$$\text{Precision} = \frac{\text{number of correct system answers}}{\text{number of system answers}}$$

- Which are the Jovian planets in the Solar System?
  - Returned Answers:
    - Mercury
    - Jupiter
    - Saturn
  - Gold-standard:
    - Jupiter
    - Saturn
    - Neptune
    - Uranus

Recall = 2/3 = 0.67

Mean Reciprocal Rank (MRR)

- Measures the ranking quality.
- The Reciprocal-Rank (1/r) of a query can be defined as the rank r at which a system returns the first relevant entity.

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}$$

- Which are the Jovian planets in the Solar System?
  - Returned Answers:
    - Mercury
    - Jupiter
    - Saturn
  - Gold-standard:
    - Jupiter
    - Saturn
    - Neptune
    - Uranus

rr = 1/2 = 0.5

Mean Average Interpolated Precision (MAiP)

- Computes the Interpolated-Precision at a set of n standard recall levels (1%, 10%, 20%, etc).
- Average-Interpolated-Precision (AiP) is a single-valued measure that reflects the performance of a search engine over all the relevant results.
- Mean-Average-Interpolated-Precision (MAiP) that reflects the performance of a system over all the results.

$$Mean\text{-}Average\text{-}Interpolated\text{-}Precision(MAiP) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} iP(RL_j)$$

where $|Q|$ is the total number of topics,
m_j is the total number of relevant results for topic q_j,
RL_j is the ranked list of results returned for topic q_j.

Normalized Discounted Cumulative Gain (NDCG)

- Discounted-Cumulative-Gain (DCG) uses a graded relevance scale to measure the gain of a system based on the positions of the relevant entities in the result set.
- This measure gives a lower gain to relevant entities returned in the lower ranks to that of the higher ranks.

$$NDCG(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)}-1}{log_2(1+m)}$$

where $|Q|$ is the total number of topics,
$R(j,e)$ is the binary relevance score obtained for an individual result of topic j,
Z_{jk} is the normalization factor,
k is the rank at which NDCG is calculated.

Test Collections

- Question Answering over Linked Data (QALD-CLEF)
- INEX Linked Data Track
- BioASQ
- SemSearch

QALD

QALD-1, ESWC (2011)

- Datasets:
  - Dbpedia 3.6 (RDF)
  - MusicBrainz (RDF)
- Tasks:
  - Training questions: 50 questions for each dataset
  - Test questions: 50 questions for each dataset

http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=challenge&q=1

QALD-1, ESWC (2011)

```
<question id="123">
  <string>Which caves have more than 100 entrances?</string>
  <query>
   PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
   PREFIX onto:  <http://dbpedia.org/ontology/>
   PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
   SELECT ?uri ?string
   WHERE {
    ?uri rdf:type onto:Cave .
    ?uri onto:numberOfEntrances ?entrance .
    FILTER (?entrance > 100) .
    OPTIONAL { ?uri rdfs:label ?string .  }
    FILTER (lang(?string) = "en") }
  </query>
  <answers>
    <answer>
     <uri>http://dbpedia.org/resource/Kanheri_Caves</uri>
     <string>Kanheri Caves</string>
    </answer>
    <answer>
     <uri>http://dbpedia.org/resource/Ox_Bel_Ha_Cave_System</uri>
     <string>Ox Bel Ha Cave System</string>
   </answer>
  </answers>
</question>
```

QALD-1, ESWC (2011)

- Which presidents were born in 1945?
- Who developed the video game World of Warcraft?
- List all episodes of the first season of the HBO television series The Sopranos!
- Who produced the most films?
- Which mountains are higher than the Nanga Parbat?
- Give me all actors starring in Batman Begins.
- Which software has been developed by organizations founded in California?
- Which companies work in the aerospace industry as well as on nuclear reactor technology?
- Is Christian Bale starring in Batman Begins?
- Give me the websites of companies with more than 500000 employees.
- Which cities have more than 2 million inhabitants?

QALD-2, ESWC (2012)

- Datasets:
  - Dbpedia 3.7 (RDF)
  - MusicBrainz (RDF)
- Tasks:
  - Training questions: 100 questions for each dataset
  - Test questions: 50 questions for each dataset

http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=challenge&q=1

QALD-3, CLEF (2013)

- Datasets:
  - Dbpedia 3.7 (RDF)
  - MusicBrainz (RDF)
- Tasks:
  - Multilingual QA
    - Given a RDF dataset and a natural language question or set of keywords in one of six languages (English, Spanish, German, Italian, French, Dutch), either return the correct answers, or a SPARQL query that retrieves these answers.
  - Ontology Lexicalization

http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=task1&q=3

INEX Linked Data Track (2013)

- Focuses on the combination of textual and structured data.
- Datasets:
  - English Wikipedia (MediaWiki XML Format)
  - DBpedia 3.8 & YAGO2 (RDF)
  - Links among the Wikipedia, DBpedia 3.8, and YAGO2 URI's.
- Tasks:
  - Ad-hoc Task: return a ranked list of results in response to a search topic that is formulated as a keyword query (144 search topics).
  - Jeopardy Task: Investigate retrieval techniques over a set of natural-language Jeopardy clues (105 search topics – 74 (2012) + 31 (2013)).

https://inex.mmci.uni-saarland.de/tracks/lod/

INEX Linked Data Track (2013)

```
<topic id="2012374" category="Politics">
  <jeopardy_clue>
    Which German politician is a successor of another politician
    who stepped down before his or her actual term was over,
    and what is the name of their political ancestor?
  </jeopardy_clue>
  <keyword_title>
   German politicians successor other stepped down before
   actual term name ancestor
  </keyword_title>
  <sparql_ft>
    SELECT ?s ?s1 WHERE {
      ?s rdf:type <http://dbpedia.org/class/yago/GermanPoliticians>.
      ?s1 <http://dbpedia.org/property/successor> ?s.
      FILTER FTContains (?s, "stepped down early").
    }
  </sparql_ft>
</topic>
```

SemSearch Challenge

- Focuses on entity search over Linked Datasets.
- Datasets:
  - Sample of Linked Data crawled from publicly available sources (based on the Billion Triple Challenge 2009).
- Tasks:
  - Entity Search: Queries that refer to one particular entity. Tiny sample of Yahoo! Search Query.
  - List Search: The goal of this track is select objects that match particular criteria. These queries have been hand-written by the organizing committee.

http://semsearch.yahoo.com/datasets.php#

SemSearch Challenge

- List Search queries:
  - republics of the former Yugoslavia
  - ten ancient Greek city
  - kingdoms of Cyprus
  - the four of the companions of the prophet
  - Japanese-born players who have played in MLB where the British monarch is also head of state
  - nations where Portuguese is an official language
  - bishops who sat in the House of Lords
  - Apollo astronauts who walked on the Moon

SemSearch Challenge

- Entity Search queries:
  - 1978 cj5 jeep
  - employment agencies w. 14th street
  - nyc zip code
  - waterville Maine
  - LOS ANGELES CALIFORNIA
  - ibm
  - KARL BENZ
  - MIT

- Datasets:
  - PubMed documents

- Tasks:
  - 1a: Large-Scale Online Biomedical Semantic Indexing
    - Automatic annotation of PubMed documents.
    - Training data is provided.
  - 1b: Introductory Biomedical Semantic QA
    - 300 questions and related material (concepts, triples and golden answers).

Baseline

Balog & Neumayer, A Test Collection for Entity Search in DBpedia (2013).

Model	INEX-XER		TREC Entity		SemSearch ES		SemSearch LS		QALD-2		INEX-LD		Total	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
LM	.1672	.2618	.0970	.1294	.3139	.2508	.1788	.1907	.1067	.0507	.1057	.2360	.1750	.1816
MLM-tc	.1585	.2345$^{\triangledown}$	.0855$^{\triangledown}$	.1176	.3541$^{\blacktriangle}$	.2838$^{\blacktriangle}$	.1738	.1744	.0989$^{\triangledown}$	.0507	.1044	.2320	.1813$^{\triangle}$	.1847
MLM-all	.1589	.2273	.0641	.0882	.3010	.2454	.1514	.1581	.1204	.0593	.0857$^{\triangledown}$	.1850$^{\blacktriangledown}$	.1668	.1639$^{\blacktriangledown}$
PRMS	.1897$^{\triangle}$	.2855	.1206	.1706	.3228	.2515	.1857	.2093	.1050	.0693$^{\triangle}$	.0840$^{\blacktriangledown}$	.2030$^{\blacktriangledown}$	.1764	.1862
BM25	.1830	.2891	.0882	.1000	.3262	.2562	.1785	.2116	.1184	.0657	.1178	.2470	.1856	.1936
BM25F-tc	.1720$^{\blacktriangledown}$	.2655$^{\blacktriangledown}$	.0848	.0882	.3337$^{\triangle}$	.2631$^{\triangle}$	.1718	.2163	.1067$^{\triangledown}$	.0621	.1169	.2490	.1820$^{\triangledown}$	.1922
BM25F-all	.1810	.2836	.0824$^{\triangledown}$	.0824	.3286	.2585	.1789	.2163	.1189	.0686	.1155	.2470	.1855	.1942

Going Deeper

- Metrics, Statistics, Tests – Tetsuya Sakai (IR)
  - http://www.promise-noe.eu/documents/10156/26e7f254-1feb-4169-9204-1c53cc1fd2d7
- Building test Collections (IR Evaluation – Ian Soboroff)
  - http://www.promise-noe.eu/documents/10156/951b6dfb-a404-46ce-b3bd-4bbe6b290bfd

Do-it-yourself (DIY): Core Resources

Corpora: Wikipedia

- High domain coverage:
  - ~95% of Jeopardy! Answers.
  - ~98% of TREC answers.
- Wikipedia is entity-centric.
- Curated link structure.
- Where to use:
  - Construction of distributional semantic models.
  - As a commonsense KB
- Complementary tools:
  - Wikipedia Miner

Linked Datasets

- DBpedia: Instances and data.
- YAGO: Classes and instances.
- Freebase: Instances.
- CIA Factbook: Data.

<http://dbpedia.org/class/yago/19th-centuryPresidentsOfTheUnitedStates>
<http://dbpedia.org/class/yago/4th-centuryBCGreekPeople>
<http://dbpedia.org/class/yago/OrganizationsEstablishedIn1918>
<http://dbpedia.org/class/yago/PeopleFromToronto>
<http://dbpedia.org/class/yago/JewishAtheists>
<http://dbpedia.org/class/yago/CountriesOfTheMediterraneanSea>
<http://dbpedia.org/class/yago/TennisPlayersAtThe1996SummerOlympics>
<http://dbpedia.org/class/yago/OlympicGoldMedalistsForTheUnitedStates>
<http://dbpedia.org/class/yago/WorldNo.1TennisPlayers>
<http://dbpedia.org/class/yago/PeopleAssociatedWithTheUniversityOfZurich>
<http://dbpedia.org/class/yago/SwissImmigrantsToTheUnitedStates>
<http://dbpedia.org/class/yago/1979VideoGames>

Linked Datasets

- DBpedia: Instances and data.
- YAGO: Classes and instances.
- Freebase: Instances.
- CIA Factbook: Data.

- Where to use:
  - As a commonsense KB.

Dictionaries

- WordNet: Large Lexical Database
- Wikitionary: Dictionary

- Where to use:
  - Query expansion.
  - Semantic similarity.
  - Semantic relatedness.
  - Word sense disambiguation.

- Complementary tools:
  - WordNet::Similarity

Parsers

- Stanford Parser: POS-Tagger, Syntactic & Dependency Trees.

- Where to use:
  - Question Analysis.

Named Entity Recognition/Linking

- Stanford NER.
- DBpedia Spotlight

- Where to use:
  - Question Analysis

Search Engines

- Lucene & Solr: Advanced search engine framework.

Distributional Semantics

- EasyESA: High-performance Explicit Semantic Analysis (ESA) framework.
- Semantic Vectors.
- Where to use:
  - Semantic relatedness & similiarity.
  - Word Sense Disambiguation.

Linked Data Extraction

- Fred: Represents the extracted data using ontology patterns.
- Graphia: Extracts contextualized Linked Data graphs.

Trends

Research Topics/Opportunities

- #1: Merge Linked Data extraction into QA4LD.
- #2: Explore complex dialog/context in QA4LD tasks.
- #3: Advance the use of distributional semantic models on QA.
- #4: Open QA pipelines.
- #5: Multilingual QA.
- #6: Creation of multimodal QA approaches.
- #7:Integration of reasoning (deductive, inductive, counterfactual, abductive ...) on QA approaches and test collections.
- #8: Development of QA approaches/tasks which use both structured and unstructured data.

Take-away Message

- Big Data/complex dataspaces **demand** new principled semantic approaches to cope with the scale and heterogeneity of data.
- Information systems in the future will depend on semantic technologies.
- Part of the Semantic Web/AI vision can be addressed today with a multi-disciplinary perspective:
  - Linked Data, IR and NLP
- You can build your own IBM Watson-like application.
- Both data and tools are available and ready to use: the main barrier is the mindset.
- Still a very active research area.

References

[1] Eifrem, A NOSQL Overview And The Benefits Of Graph Database (2009)

[2] Idehen, Understanding Linked Data via EAV Model (2010).

[3] Kaufmann & Bernstein, How Useful are Natural Language Interfaces to the Semantic Web for Casual End-users? (2007)

[4] Chin-Yew Lin, Question Answering.

[5] Farah Benamara, Question Answering Systems: State of the Art and Future Directions.

[6] Freitas et al., *Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches and Trends*, 2012.

[7] Freitas et al, *A Distributional Structured Semantic Space for Querying RDF Graph Data.*, 2012.

[8] Freitas et al, *A Distributional Approach for Terminological Semantic Search on the Linked Data Web*, 2012.

[9] Freitas et al, A Semantic Best-Effort Approach for Extracting Structured Discourse Graphs from Wikipedia., 2012

[10]Freitas et al., *Answering Natural Language Queries over Linked Data Graphs: A Distributional Semantics Approach*,, 2013.

[11] Freitas et al., *Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches and Trends*, 2012.

References

[12] Cimiano et al., Towards portable natural language interfaces to knowledge bases, 2008.

[13] Lopez et al., PowerAqua: fishing the semantic web, 2006.

[14] Damljanovic et al., Natural Language Interfaces to Ontologies: Combining Syntactic Analysis and Ontology-based Lookup through the User Interaction, 2010

[16] Unger et al. Template-based Question Answering over RDF Data, 2012.

[17] Cabrio et al., QAKiS: an Open Domain QA System based on Relational Patterns, 2012.

[18] How Useful Are Natural Language Interfaces to the Semantic Web for Casual End-Users?, 2007.

[19] Popescu et al.,Towards a theory of natural language interfaces to databases., 2003