Centre for Data Analytics

Insight

Schema-Agnostic Queries for Large-Schema Databases:

A Distributional Semantics Approach

André Freitas

PhD Viva

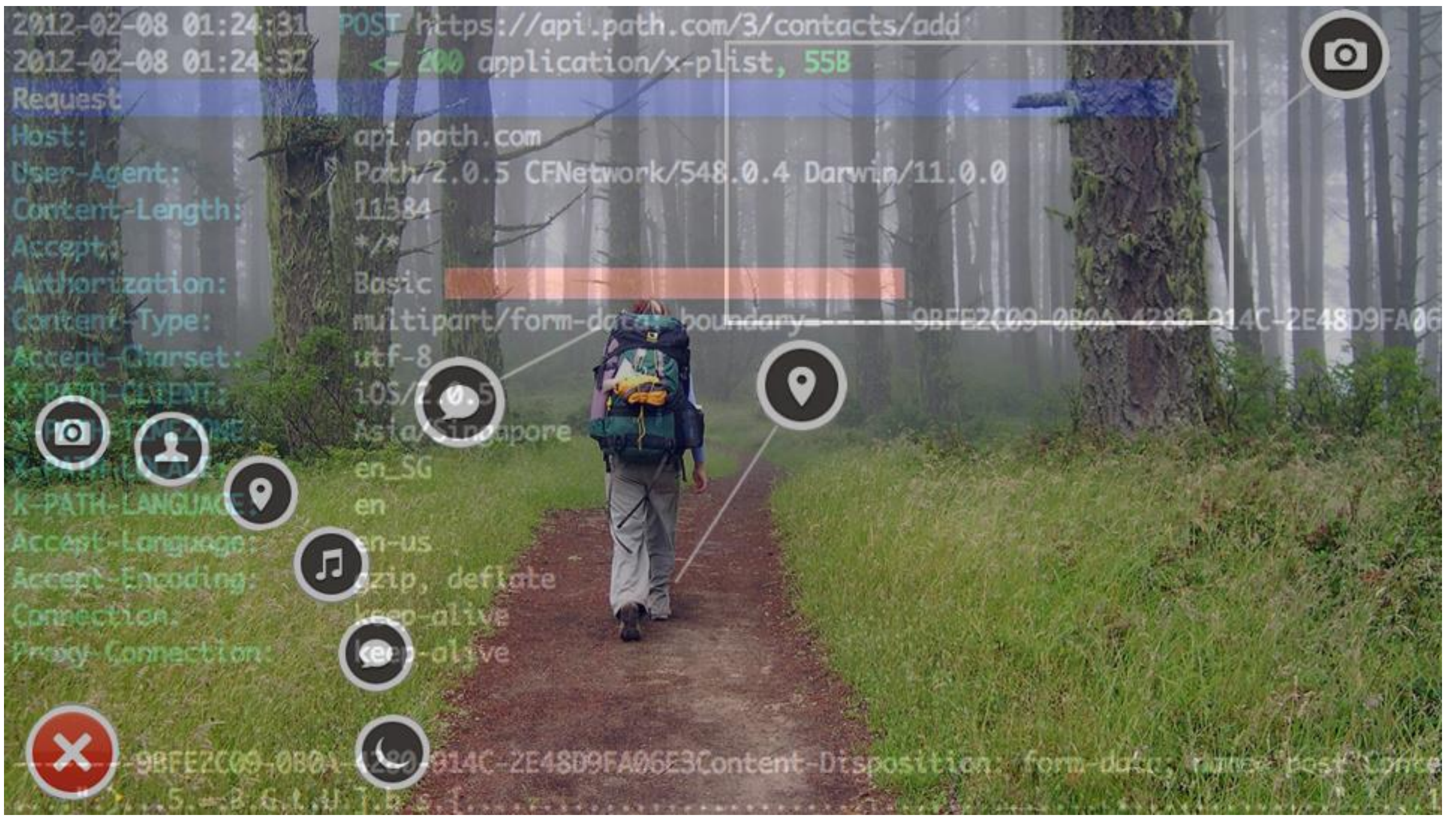
Galway - March 2nd, 2015

DERI DCU NUI Galway OÉ Gaillimh UCC UCD sfi

Motivation

Big Data

- **Vision: More complete *data-based* picture of the world for systems and users.**



Shift in the Database Landscape

- **Very-large and dynamic “schemas”.**

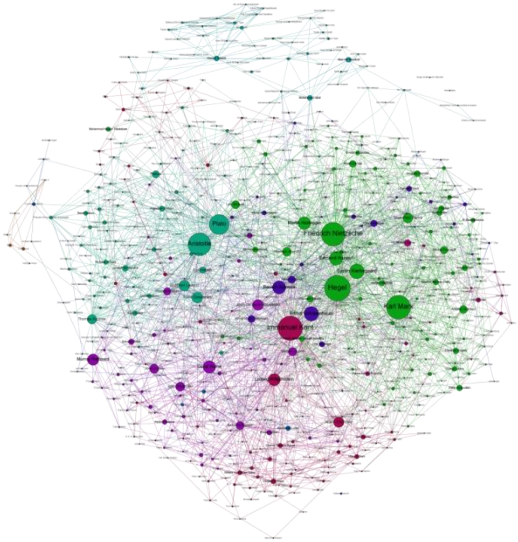
before 2000

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, Chi
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

circa 2015

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

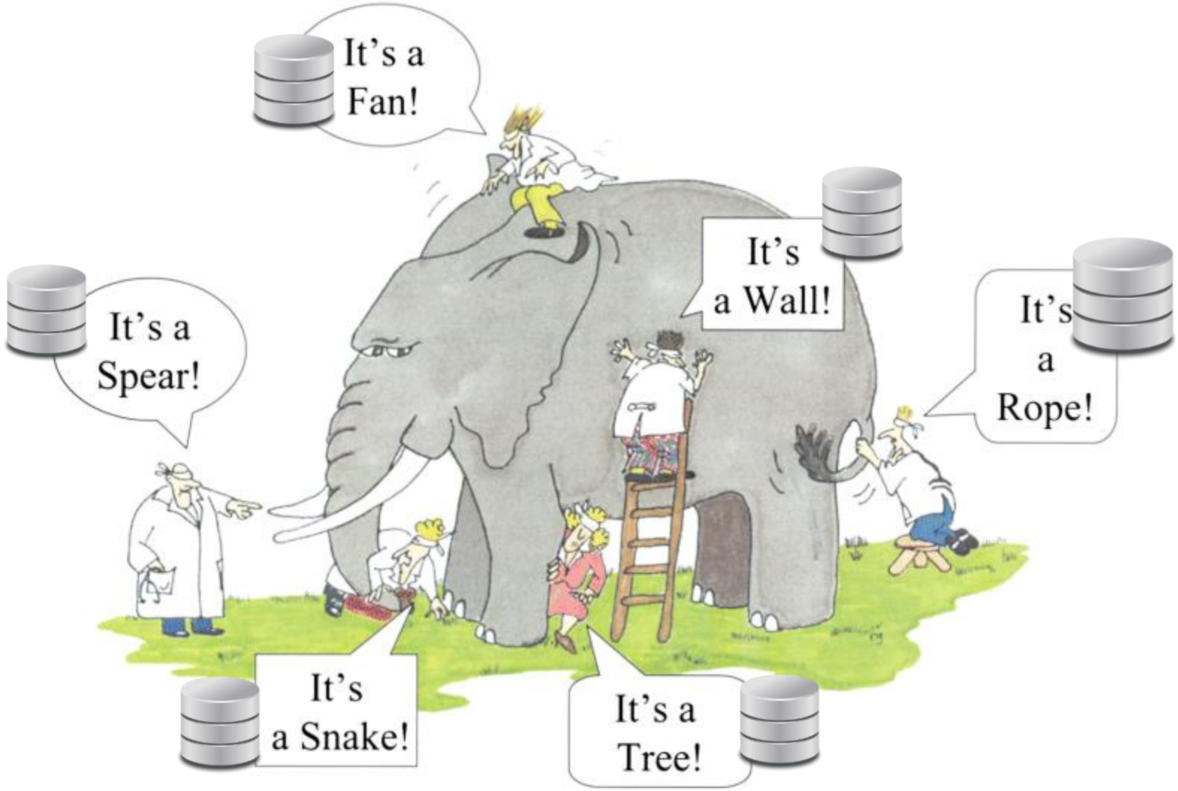


Brodie & Liu, 2010

Semantic Heterogeneity

- **Decentralized content generation.**
- **Multiple perspectives (conceptualizations) of the reality.**
- **Ambiguity, vagueness, inconsistency.**

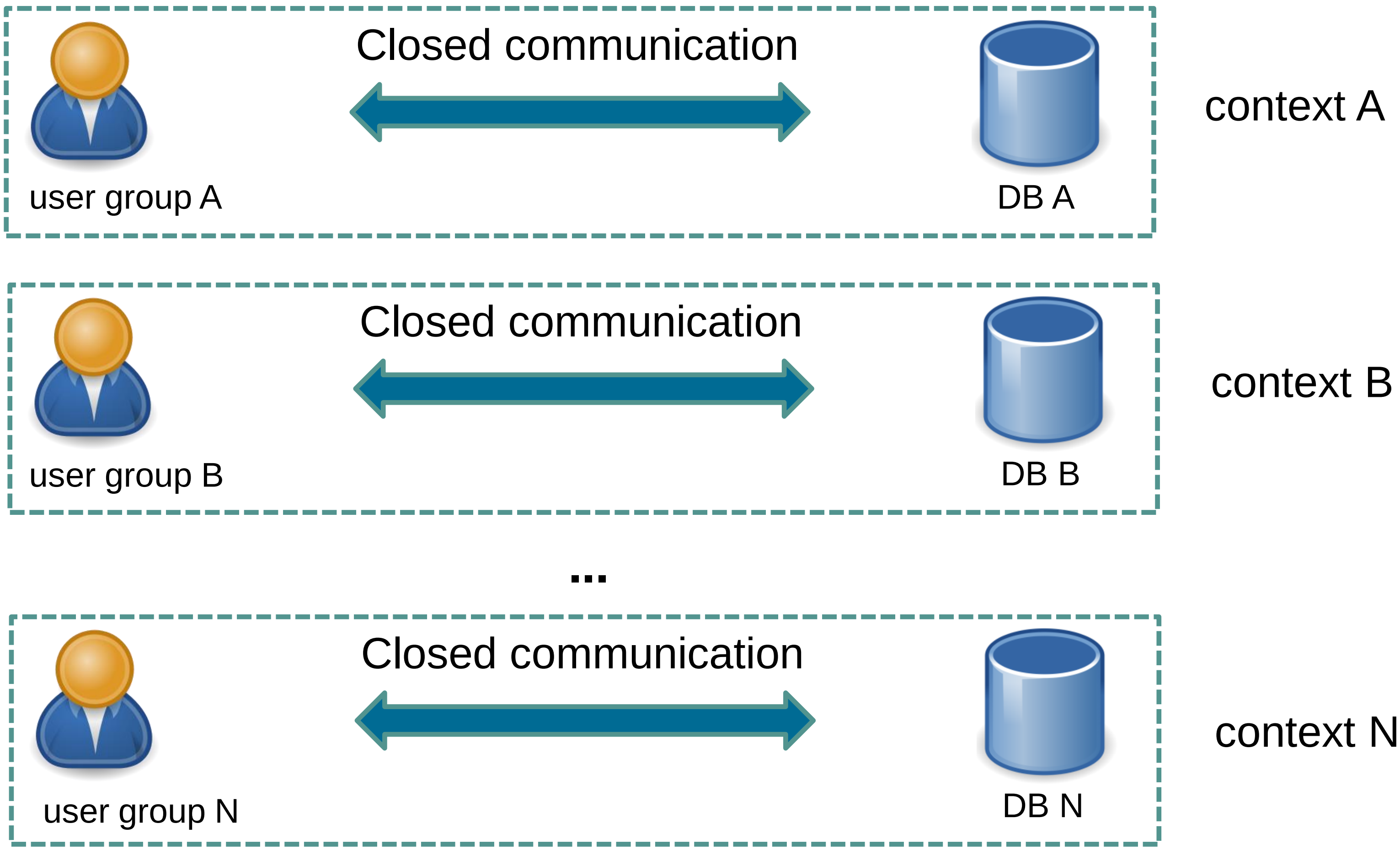
Size, Complexity, Dynamicity and Decentralisation (SCoDD)



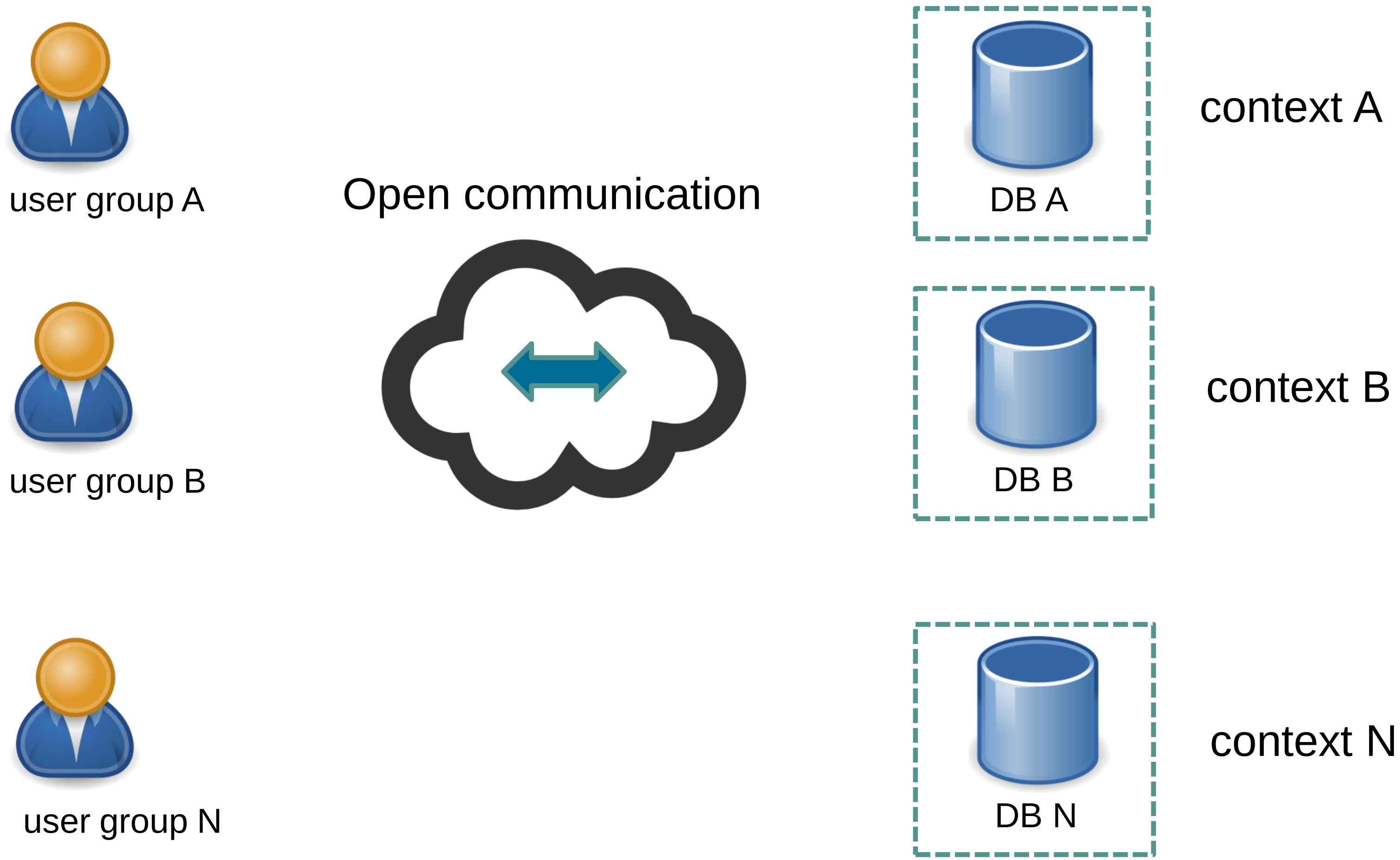
Brodie & Liu, 2010

Helland, 2012

From Closed to Open Communication

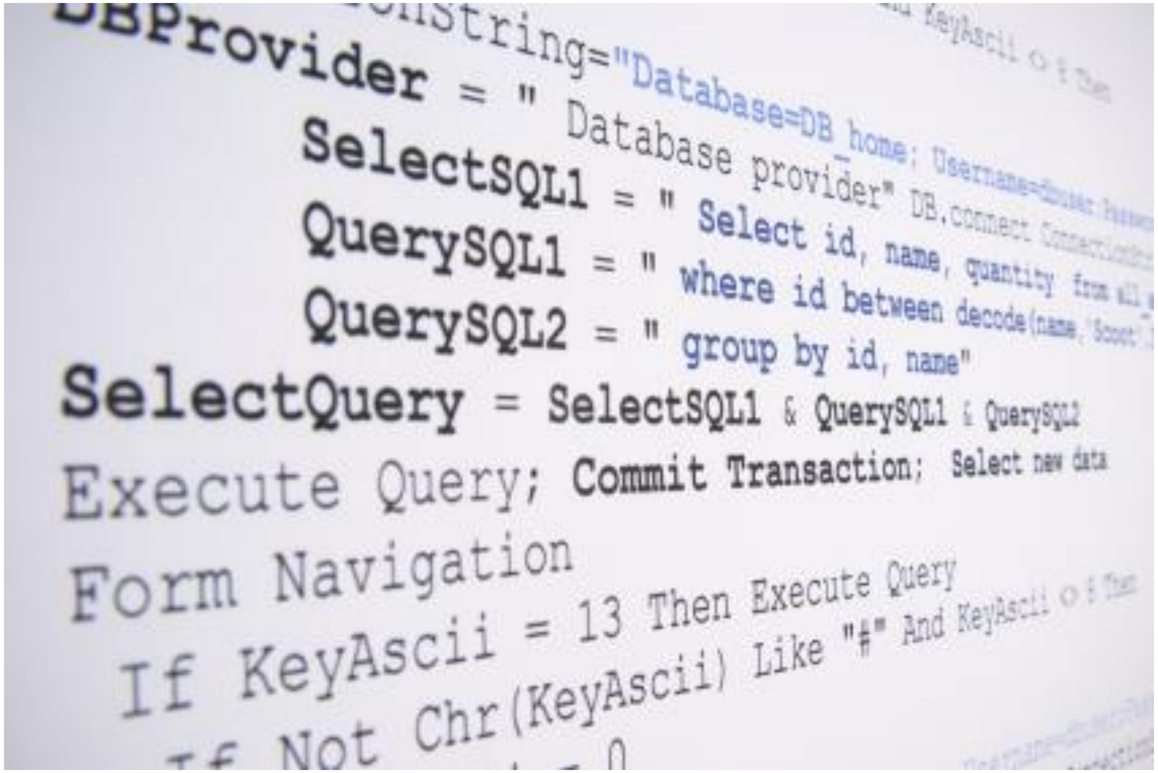


From Closed to Open Communication



Databases for a Complex World

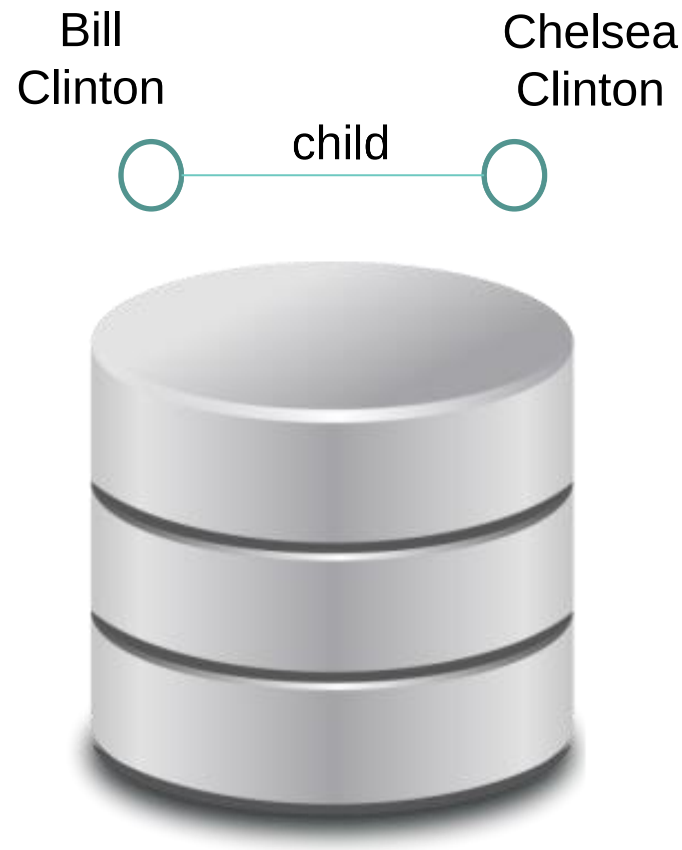
How do you **query** data on this scenario?



Schema-agnosticism

Who is the daughter of Bill Clinton?

Abstraction Layer



Schema-agnostic queries

Query approaches over structured databases which allow users satisfying complex information needs without the understanding of the representation (schema) of the database.

First-level independency (Relational Model)

"… it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and representation and organization of data on the other"

Codd, 1970

Second-level independency (Schema-agnosticism)

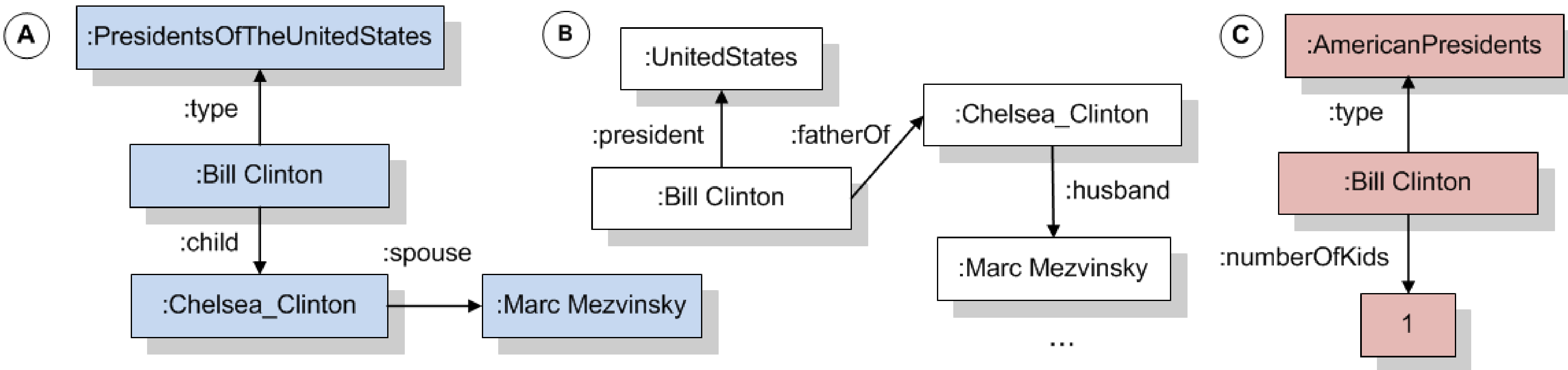
Vocabulary Problem for Databases

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

Semantic Gap

Schema-agnostic query mechanisms

Possible representations



- Abstraction level differences
- Lexical variation

- Structural (compositional) differences

Proposed Approach

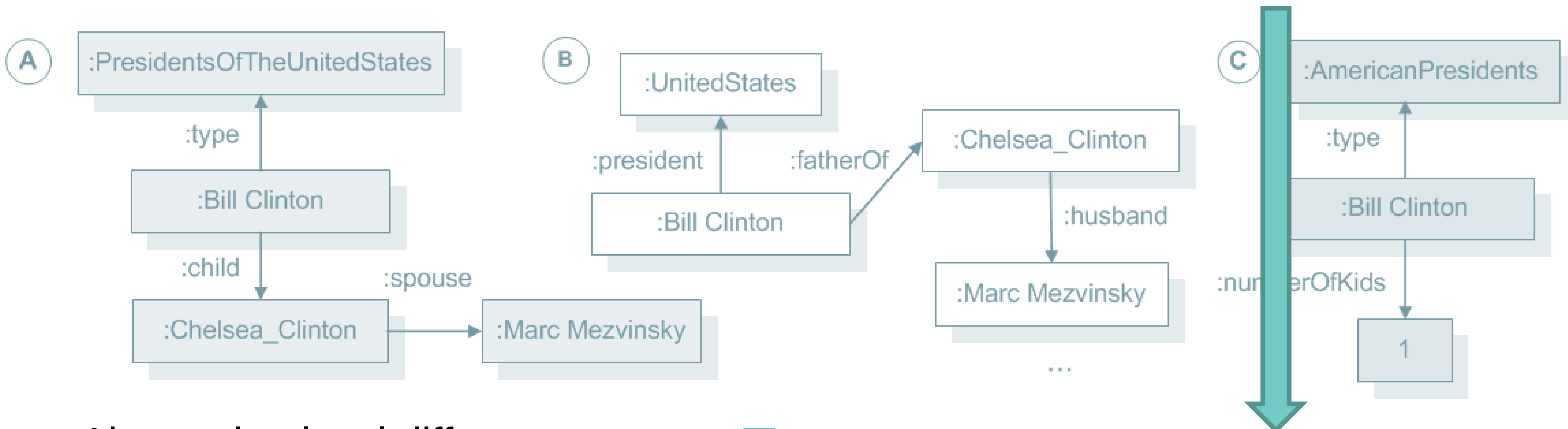
Query: Who is the daughter of Bill Clinton married to?

Semantic approximation

$\propto$

Commonsense Knowledge

Possible representations



- Abstraction level differences
- Lexical variation

Distributional Model

- Structural (compositional) differences

Compositional Model

Semantic Best-Effort

- "Too much, too fast-you need to approximate", **Helland (2011)**.
- Database results as ranked results (information retrieval perspective).

High-level Research Hypotheses

- **Hypothesis I: Distributional semantics** provides an accurate, **comprehensive** and **low maintainability** approach to cope with the abstraction level and conceptual-level dimensions of semantic heterogeneity in schema-agnostic queries over large- schema open domain datasets.

- **Hypothesis II:** The **compositional semantic model** defined by the query planning mechanism supports **expressive schema-agnostic queries** over large-schema open domain datasets.

- **Hypothesis III:** The proposed **distributional-relational structured vector space model (τ−Space)** supports the development of a **schema-agnostic query mechanism** with **interactive query execution time**, **low index construction time** and **size** and it is **scalable to large-schema open domain datasets**.

Schema-agnostic queries: General Requirements

R1. High usability & Low query construction time:

Supporting natural language queries.

R2. High query expressivity:

Path, conjunctions, disjunctions, aggregations, conditions.

R3. Accurate & comprehensive semantic matching:

High precision and recall.

R4. Low setup & maitainability effort:

Easily transportable across datasets from different domains (minimum adaptation effort/low adaptation time).

R5. Interactive & Low query execution time:

Suitable for interactive querying.

R6. High scalability:

Scalable to large datasets / a large number of datasets.

Stuckenschmidt, 2003 Wang et al., 2009 Lopez et al., 2010

Research Methodology

- Evolution of databases: demand for schema-agnosticism (Chapter I).
- Literature survey of the state-of-the-art in the problem space (Chapter III).
- Analysis and formalization of the semantic phenomena involved in the process of semantically mapping schema-agnostic queries (Chapters II, V).
- Proposal of a semantic model to support a schema-agnostic query mechanism (Chapters IV, VI, VII, VIII).
- Formulation of a schema-agnostic query mechanism (Chapter VIII).
- Evaluation of the approach and its test collection (Chapter VI).
- Analysis of the consequences of schema-agnosticism for logic programming and reasoning over incomplete knowledge bases (Chapter X).

Outline

Query-DB
Semantic Gap

Semantic Model for Schema-agnostic Databases

Schema-agnostic
Query Approach

Evaluation

Query-DB
Semantic Gap

From Semantic Tractability to Semantic Resolvability

- Semantic Tractability (Popescu et al., 2004)

- Focuses on **soundness** and **completeness** conditions for mapping natural language queries to databases.

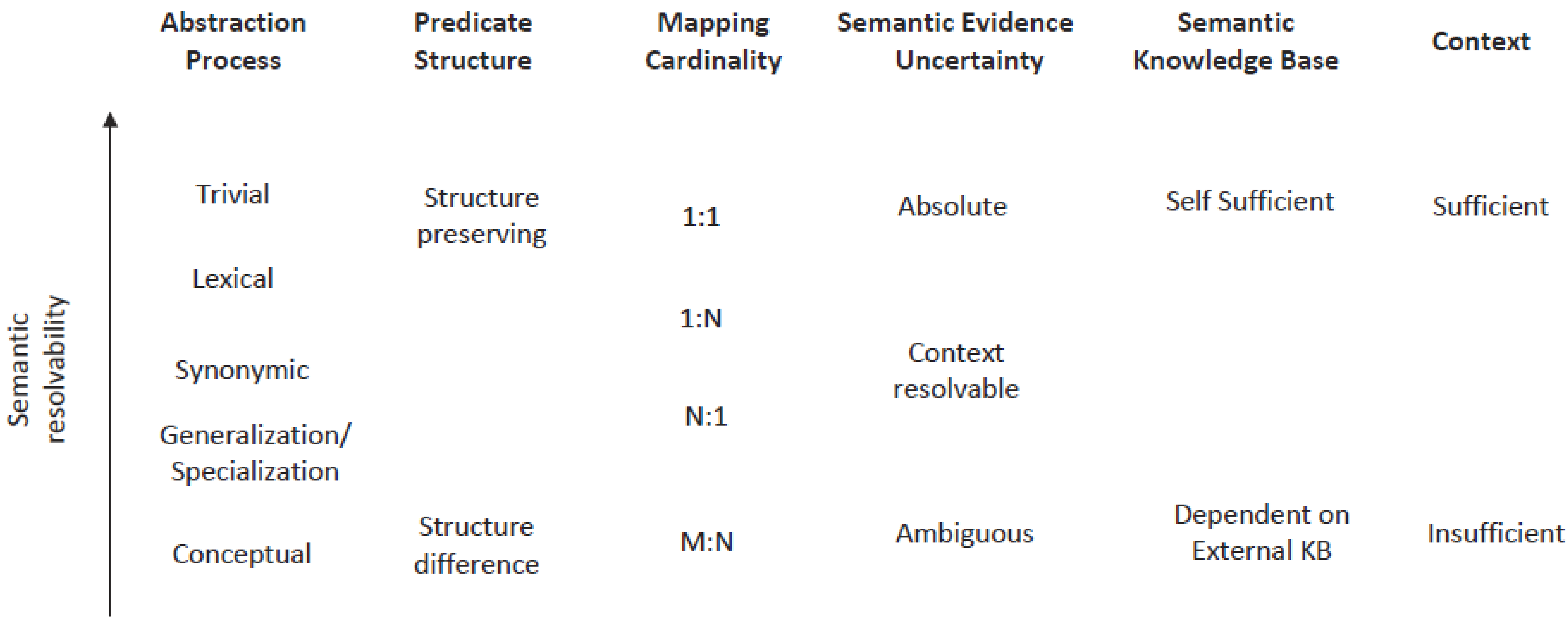
- Focuses on a restricted class of semantic mappings.

- Semantic Resolvability

- Provides a formal model for classifying query-dataset mappings for schema-agnostic queries.

George (2005) and Sheth & Kashyap (1990)

Semantic Resolvability

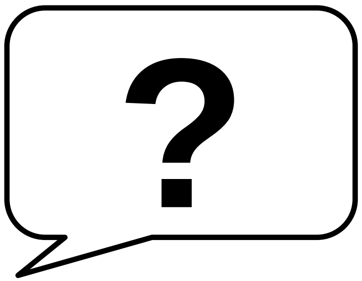


Towards an Information-Theoretical Model for Schema-agnostic Semantic Matching

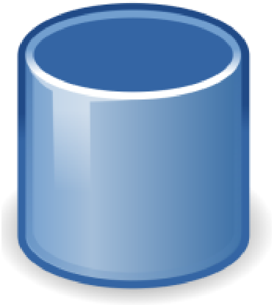
Semantic Complexity & Entropy: *Configuration space* of semantic matchings.

- Query-DB semantic gap.
- Ambiguity, synonymy, indeterminacy, vagueness.

Semantic Entropy



$H_{matching}$



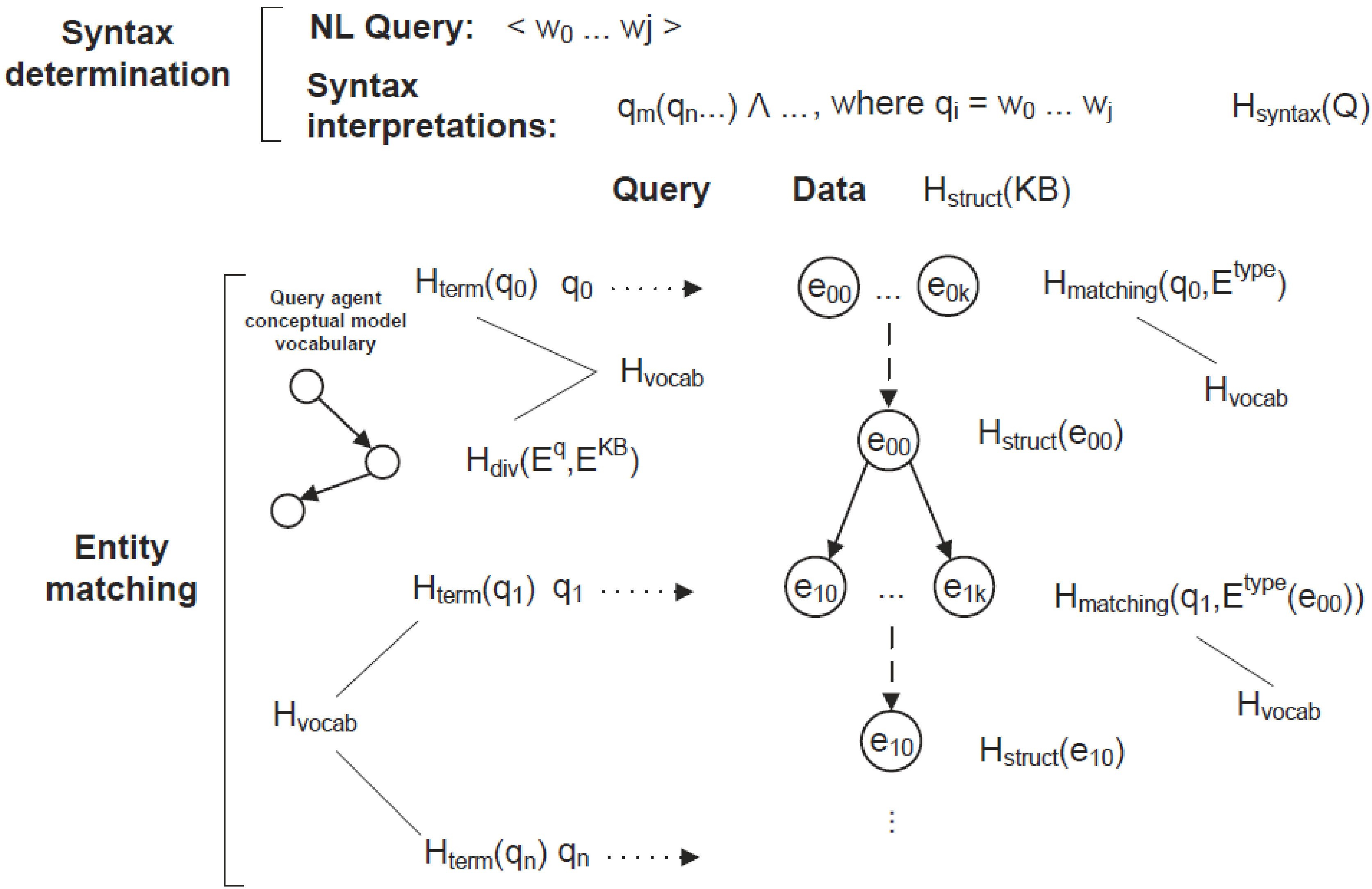
H_{syntax}

H_{struct}

H_{term}

H_{term}

Semantic Complexity & Entropy



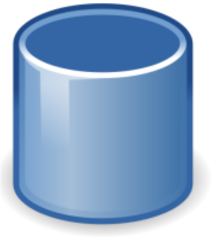
Minimizing the Semantic Entropy for the Semantic Matching

Definition of a **semantic pivot**: first query term to be resolved in the database.

- Maximizes the reduction of the semantic configuration space.

Semantic Pivots

Who is the daughter of Bill Clinton married to?



> 4,580,000

dbpedia:spouse	**dbpedia:children**	**:Bill_Clinton**
100,184	62,781	437

Minimizing the Semantic Entropy for the Semantic Matching

Definition of a **semantic pivot**: first query term to be resolved in the database.

- Maximizes the reduction of the semantic configuration space.
- Less prone to more complex synonymic expressions and abstraction-level differences.

Semantic Pivots

Who is the daughter of Bill Clinton married to?

Bill Clinton	Thomas Edward Lawrence	Paris
William Jefferson Clinton	T. E. Lawrence	City of light
William J. Clinton	Lawrence of Arabia	French capital
		Capital of France

Proper nouns tends to have high percentage of string overlap for synonymic expressions.

Minimizing the Semantic Entropy for the Semantic Matching

Definition of a **semantic pivot**: first query term to be resolved in the database.

- Maximizes the reduction of the semantic configuration space.
- Less prone to more complex synonymic expressions and abstraction-level differences.
- Semantic pivot serves as interpretation context for the remaining alignments.
- proper nouns >> nouns >> complex nominals >> adjectives , verbs.

Towards a New Semantic Model for Schema-agnostic Databases

Towards a New Semantic Model for Schema-agnostic databases

- Strategies:

- Efficient and robust semantic model **for semantic matching**.

- Semantic pivoting.

- Semantic best-effort.

Robust Semantic Model

- **Semantic approximation** (matching) is highly dependent on knowledge scale (commonsense, semantic)

Semantics

=

Formal meaning representation model
(lots of data)

+

inference model

Robust Semantic Model

- Not scalable!

1st Hard problem: Acquisition

Semantics

=

Formal meaning representation model
(lots of data)

+

inference model

Robust Semantic Model

- Not scalable!

2nd Hard problem: Consistency

Semantics

=

Formal meaning representation model

(lots of data)

+

inference model

Semantics for a Complex World

- “Most semantic models have dealt with particular types of constructions, and have been carried out under very simplifying assumptions, in true lab conditions.”
- “If these idealizations are removed it is not clear at all that modern semantics can give a full account of all but the simplest models/statements.”

Formal World

Real World

Baroni et al. 2013

Distributional Semantic Models

- Semantic Model with **low acquisition effort** (**automatically** built from text)

Simplification of the representation

- Enables the construction of comprehensive commonsense/semantic KBs

- What is the cost?

Some level of noise
(semantic best-effort)

Limited semantic model

Distributional Hypothesis

"Words occurring in similar (linguistic) contexts tend to be semantically similar"

- "He filled the *wampimuk* with the substance, passed it around and we all drunk some"

Harris, 1954 Baroni & Boleda, 2010 McDonald & Ramscar, 2001

Distributional Semantic Models (DSMs)

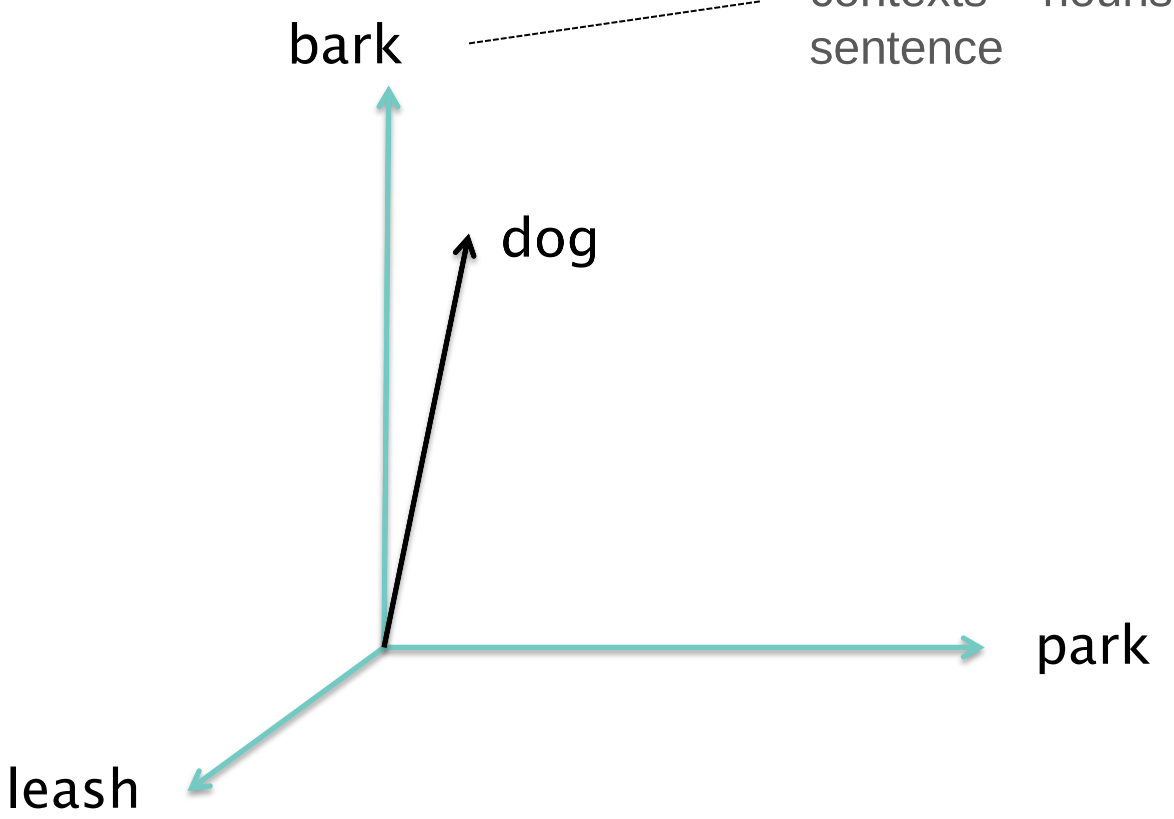
“The **dog** barked in the park. The owner of the **dog** put him on the leash since he barked.”

contexts = nouns and verbs in the same sentence

Distributional Semantic Models (DSMs)

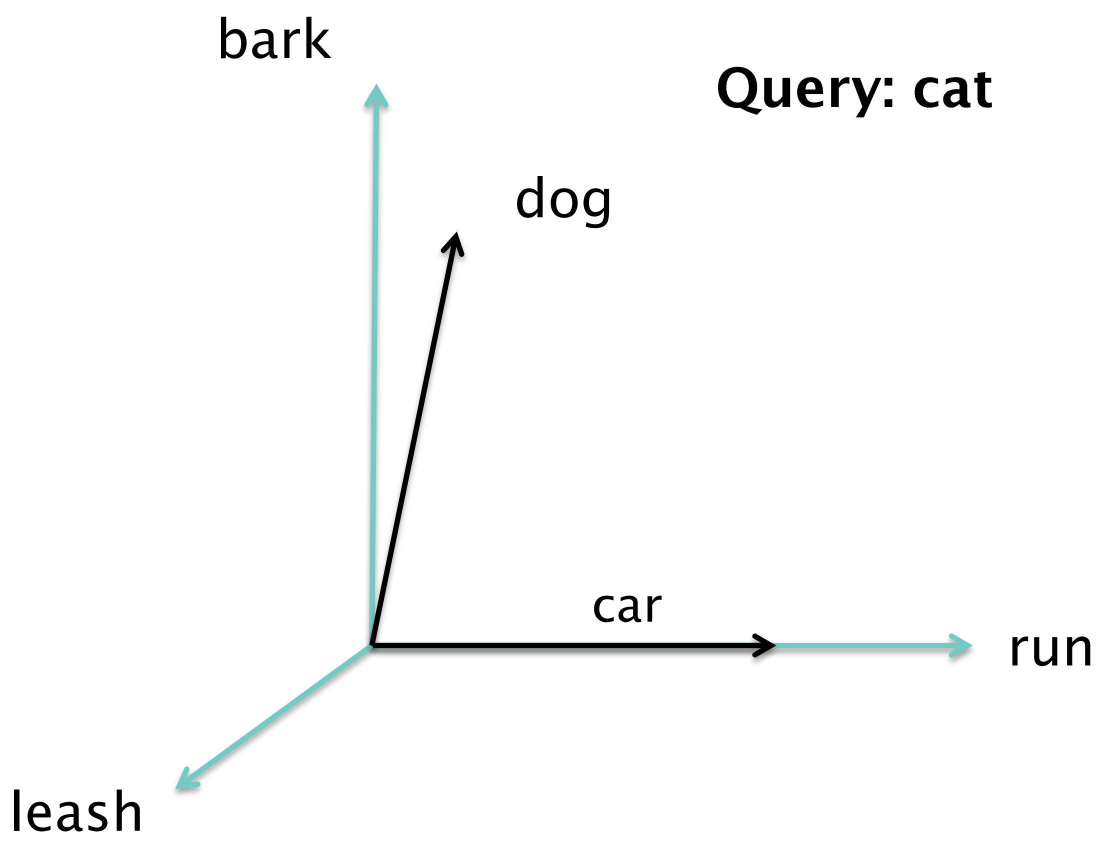
"The **dog** barked in the park. The owner of the **dog** put him on the leash since he barked."

contexts = nouns and verbs in the same sentence

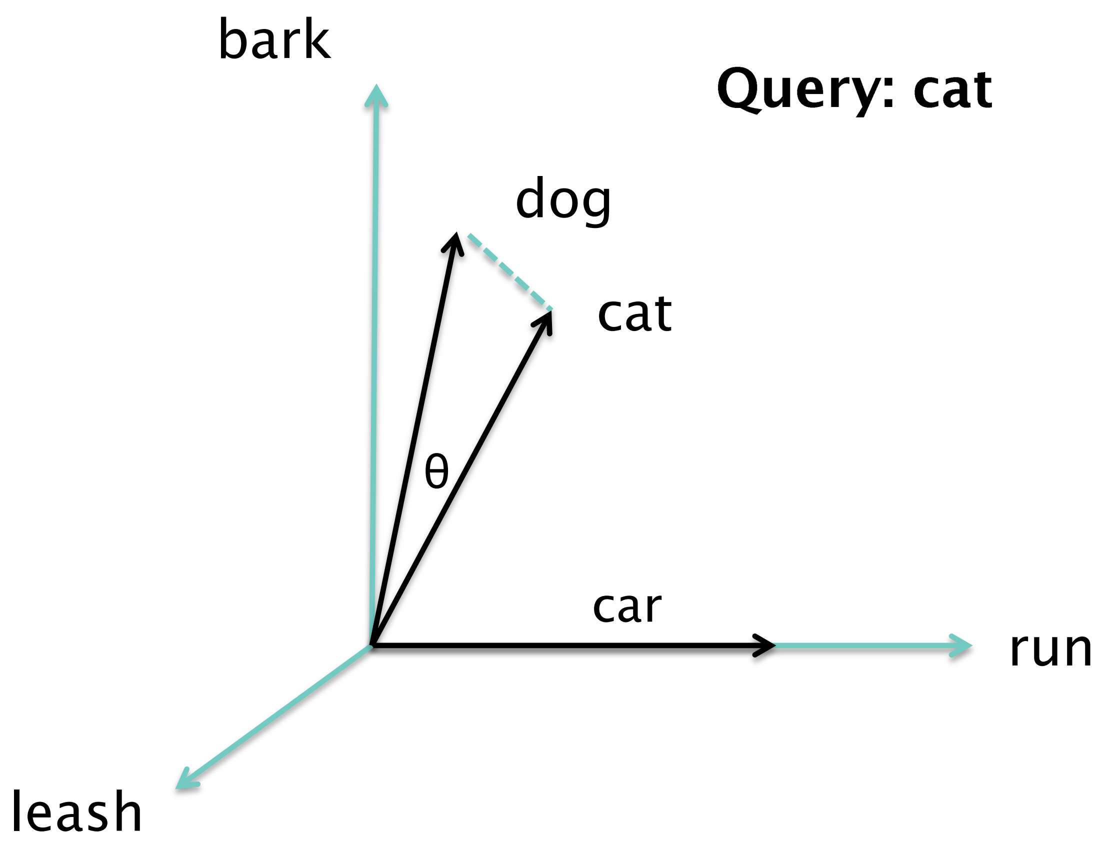


bark : 2
park : 1
leash : 1
owner : 1

Semantic Similarity & Relatedness



Semantic Relatedness



Definition of DSMs

DSMs are tuples < *T, C, R, W, M, d, S* >

T target elements, words for which the DSM provides a contextual representation.

C contexts, with which *T* co-occur.

R relation, between *T* and the contexts *C.*

W context weighting scheme.

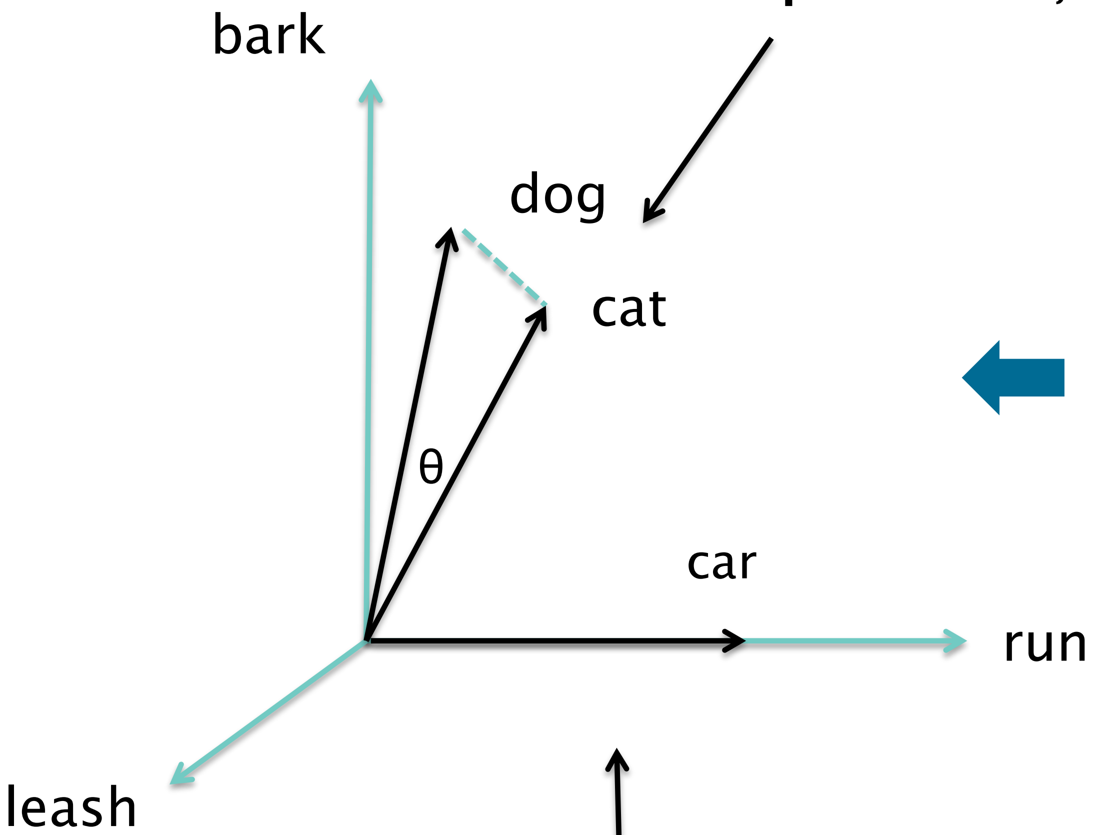
M distributional matrix, *T x C.*

d dimensionality reduction function, d *M -> M'.*

S *distance measure*, between the vectors in *M'.*

DSMs as Commonsense Reasoning

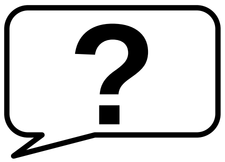
Dataset predicates , query terms



Commonsense is here

Semantic Approximation is here

Distributional Semantic Relatedness



Who is the child of Bill Clinton?

Bill Clinton father of Chelsea Clinton

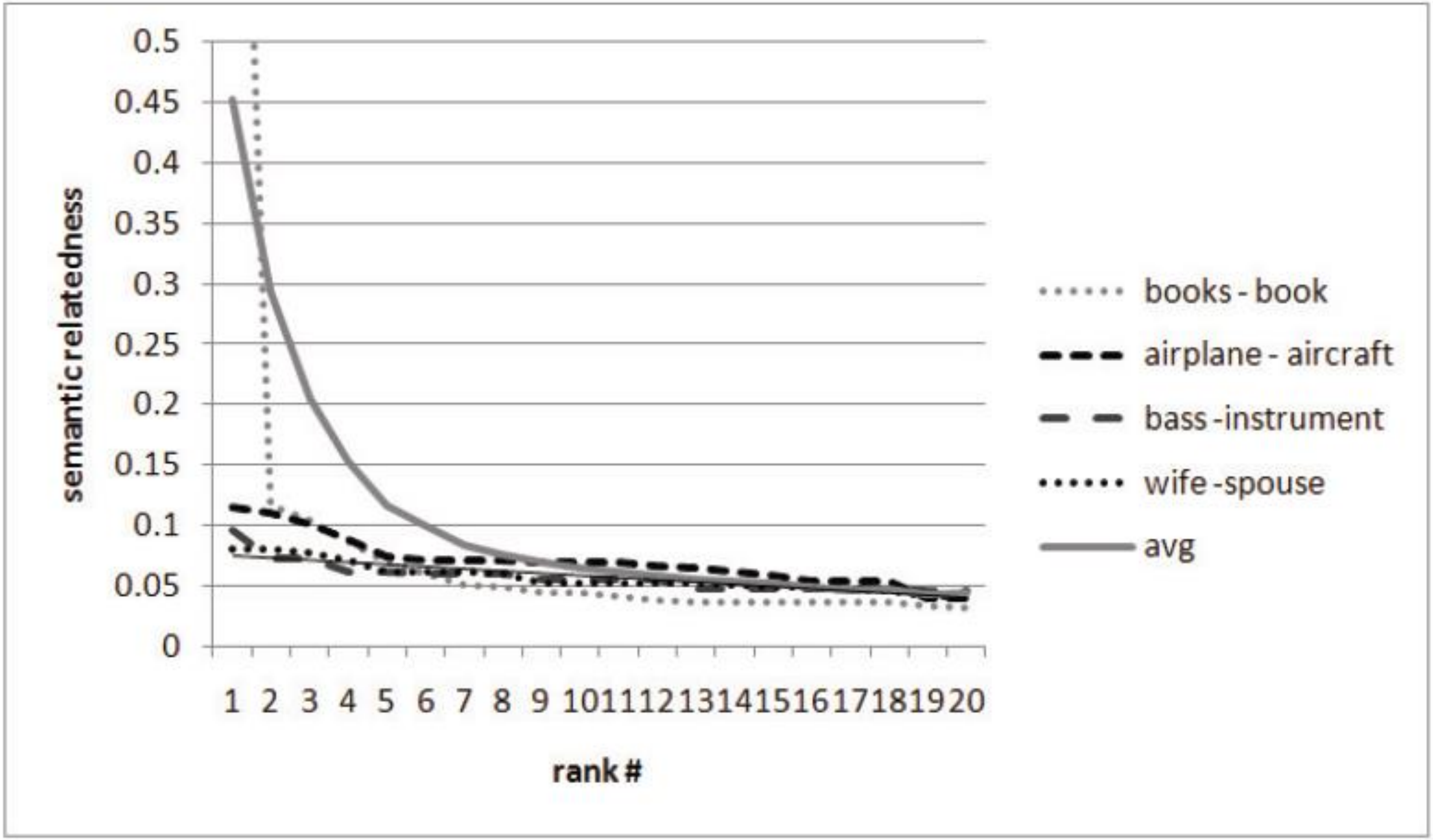
Distributional Commonsense KB (Terminology-level)

s_{rel}(childOf, fatherOf) = "0.03259"
s_{rel}(childOf, sonOf) = "0.01091"
s_{rel}(childOf, kidOf) = "0.01046"
s_{rel}(childOf, daughterOf) = "0.01059"

threshold

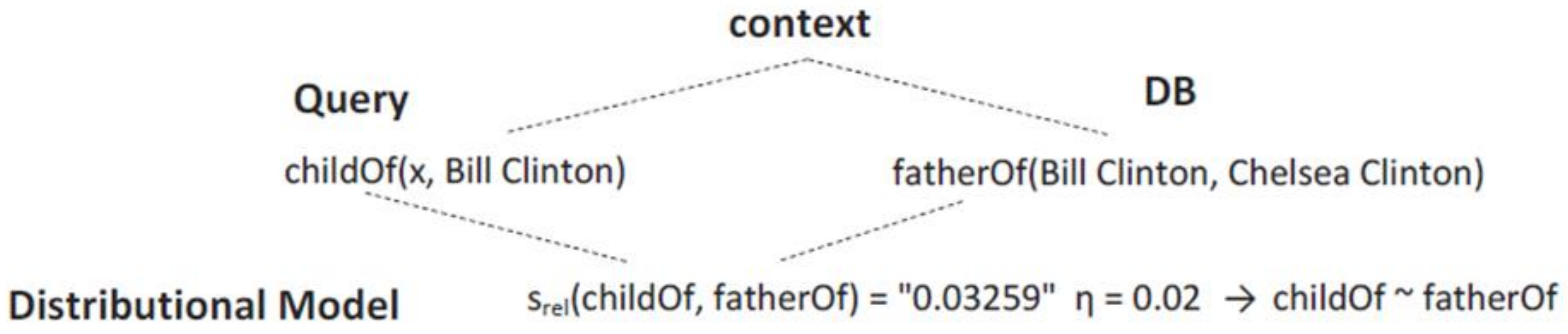
...
s_{rel}(childOf,occupation) = "0.00356"
s_{rel}(childOf, religion) = "0.00120"
s_{rel}(childOf, almaMater) = "0.0"
...

Semantic Relatedness Measure as a Ranking Function

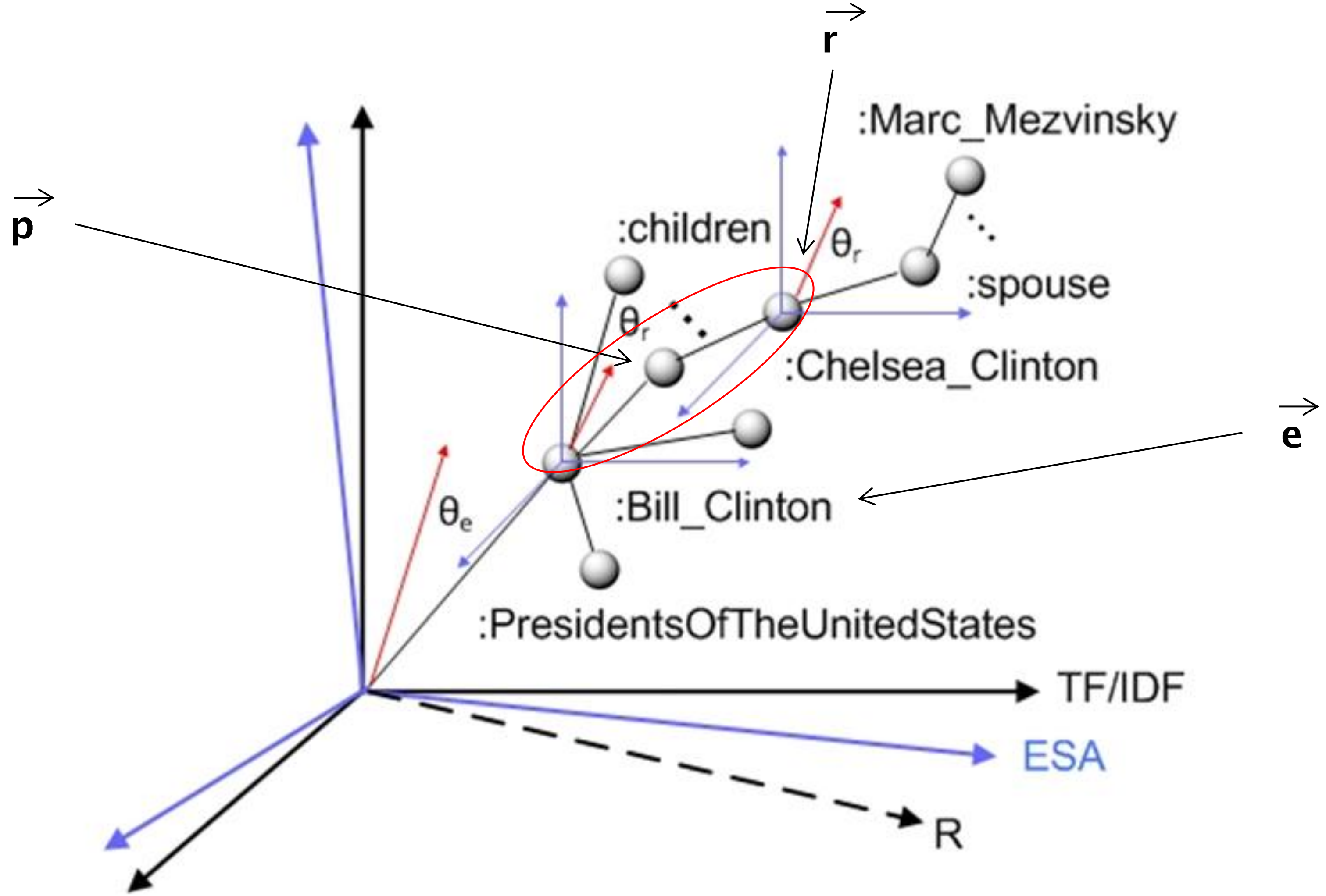


Semantic Pivoting + Distributional Semantics

- Contextual mechanism for the distributional semantic approximation.



Ƭ-Space: Hybrid Distributional-Relational Semantic Model



T-Space

Dimensional reduction mechanism!

The vector space is segmented by the semantic pivots

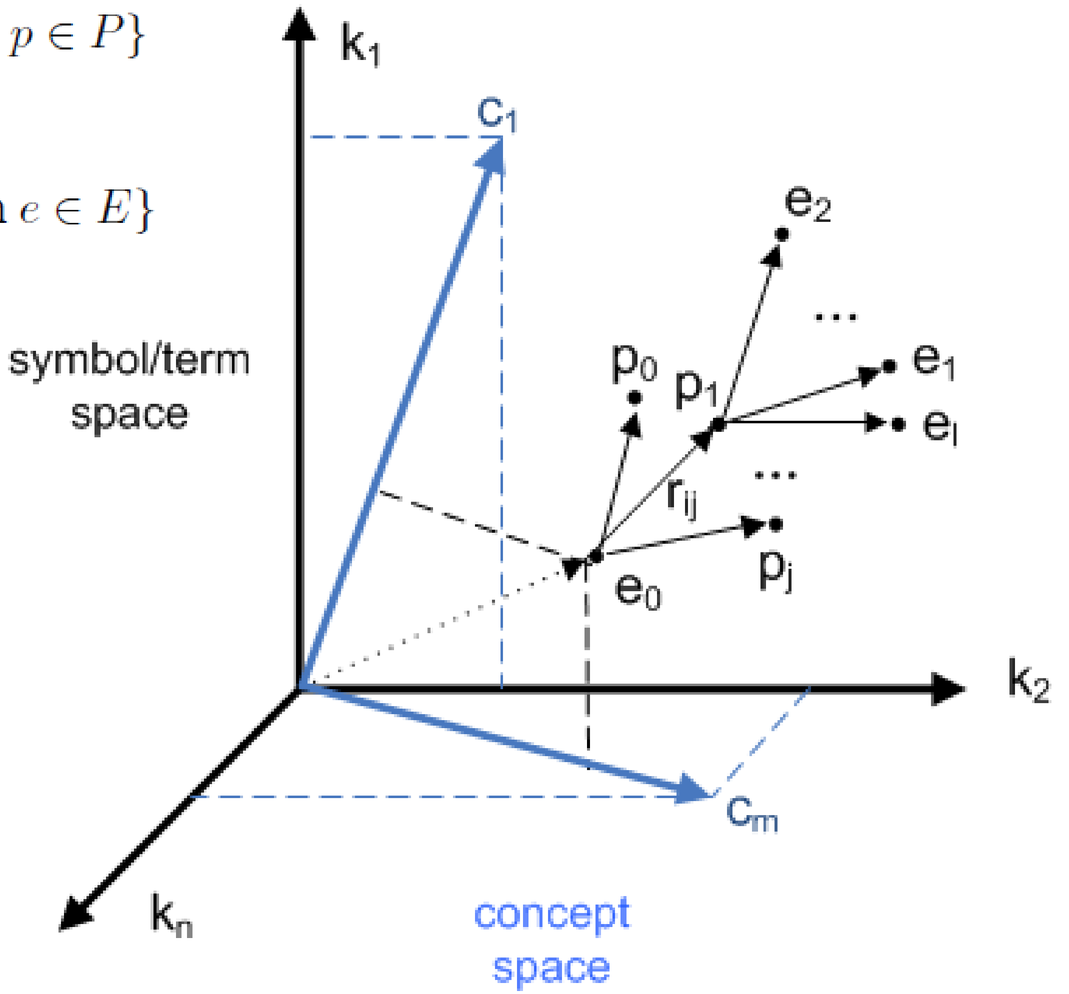
$$\vec{\mathbf{P}}_{VS^{dist}} = \{\vec{\mathbf{p}} : \vec{\mathbf{p}} = \sum_{i=1}^{t} v_i^p \vec{\mathbf{c}}_i, \text{ for each } p \in P\}$$

$$\vec{\mathbf{E}}_{VS^{dist}} = \{\vec{\mathbf{e}} : \vec{\mathbf{e}} = \sum_{i=1}^{t} v_i^e \vec{\mathbf{c}}_i, \text{ for each } e \in E\}$$

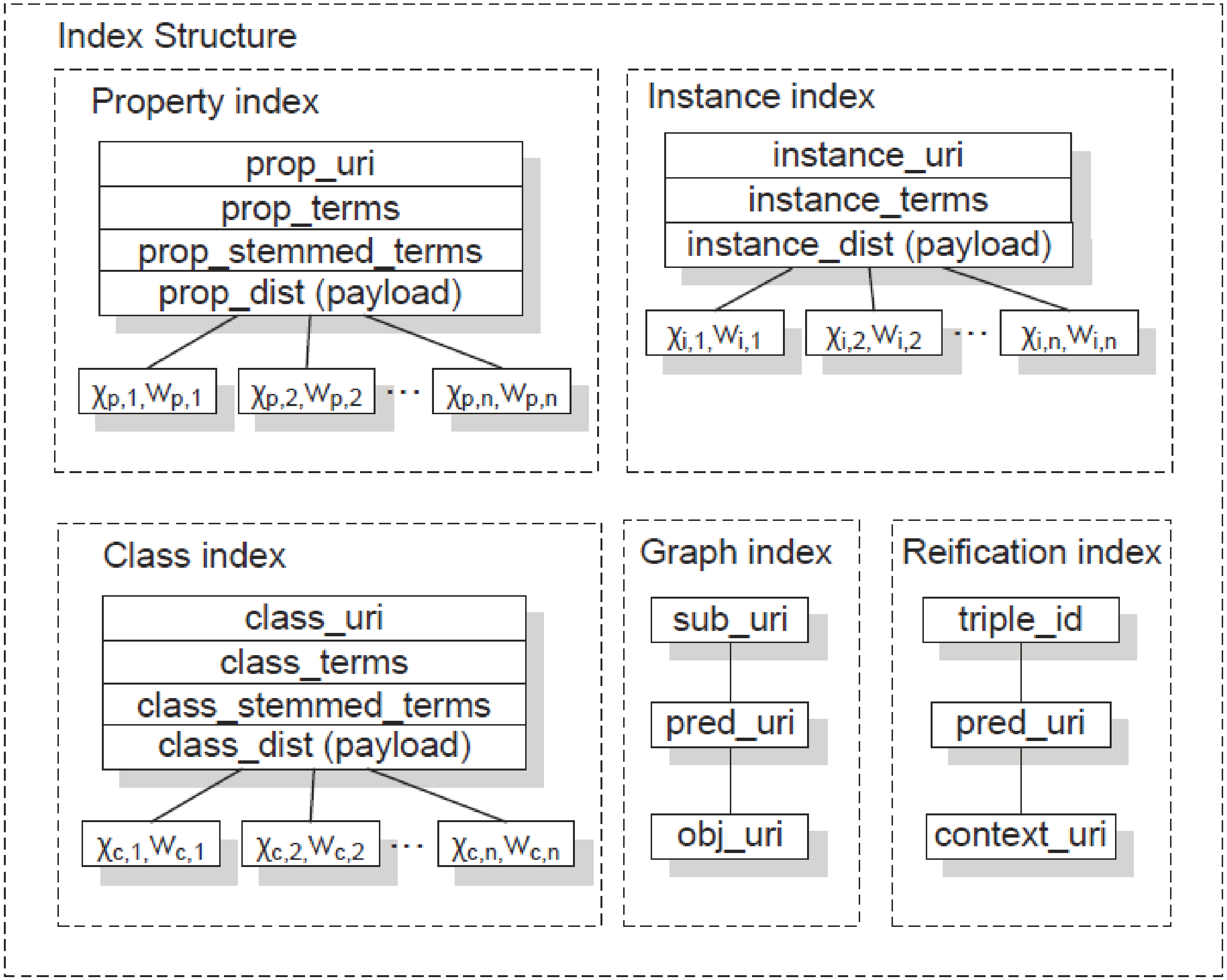
atom vector representation $\vec{\mathbf{r}}$

$p(e_1)$ $(\vec{\mathbf{p}} - \vec{\mathbf{e_1}})$

$p(e_1, e_2)$ $(\vec{\mathbf{p}} - \vec{\mathbf{e_1}}, \vec{\mathbf{e_2}} - \vec{\mathbf{p}})$



T-Space Index Structure



Schema-agnostic Query Approach

Approach Overview

Schema-agnostic Query

Query Analysis

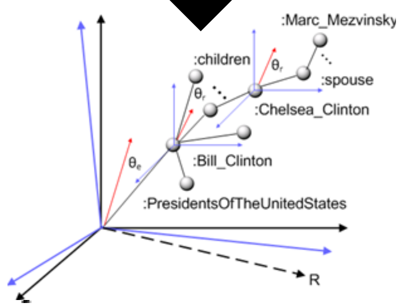
Query Features

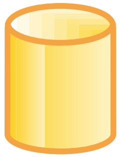
Query Planner

Query Plan

Core semantic approximation & composition operations

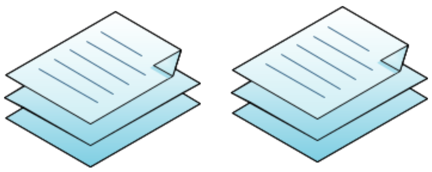
T-Space





Database

Distributional semantics



Commonsense knowledge

Large-scale unstructured data

Approach Overview

Schema-agnostic Query

Query Analysis

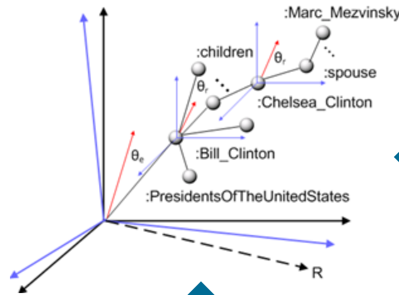
Query Features

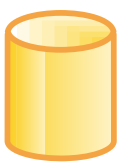
Query Planner

Query Plan

Core semantic approximation & composition operations

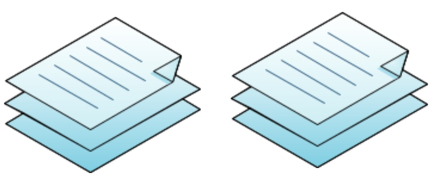
T-Space





RDF Data

Explicit Semantic Analysis (ESA)



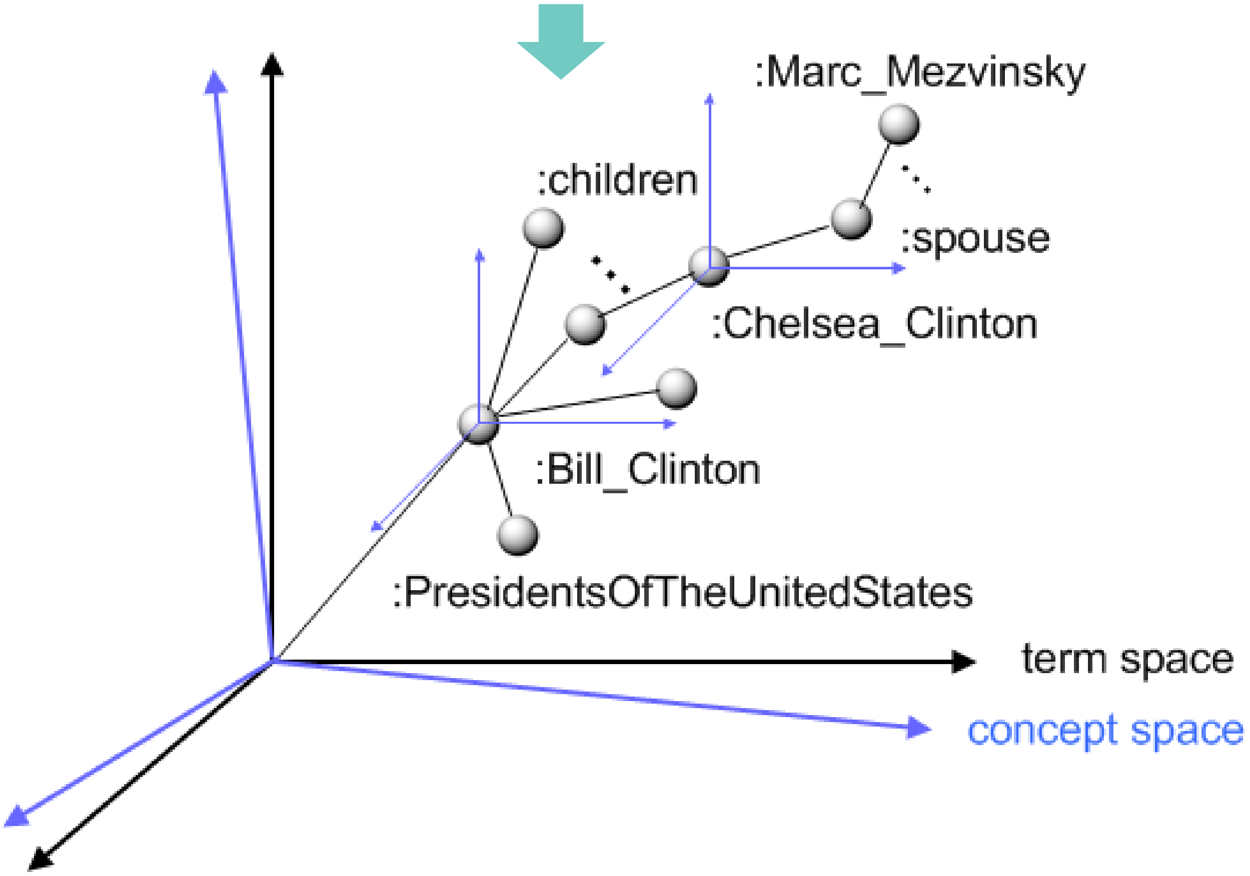
Wikipedia

Commonsense knowledge

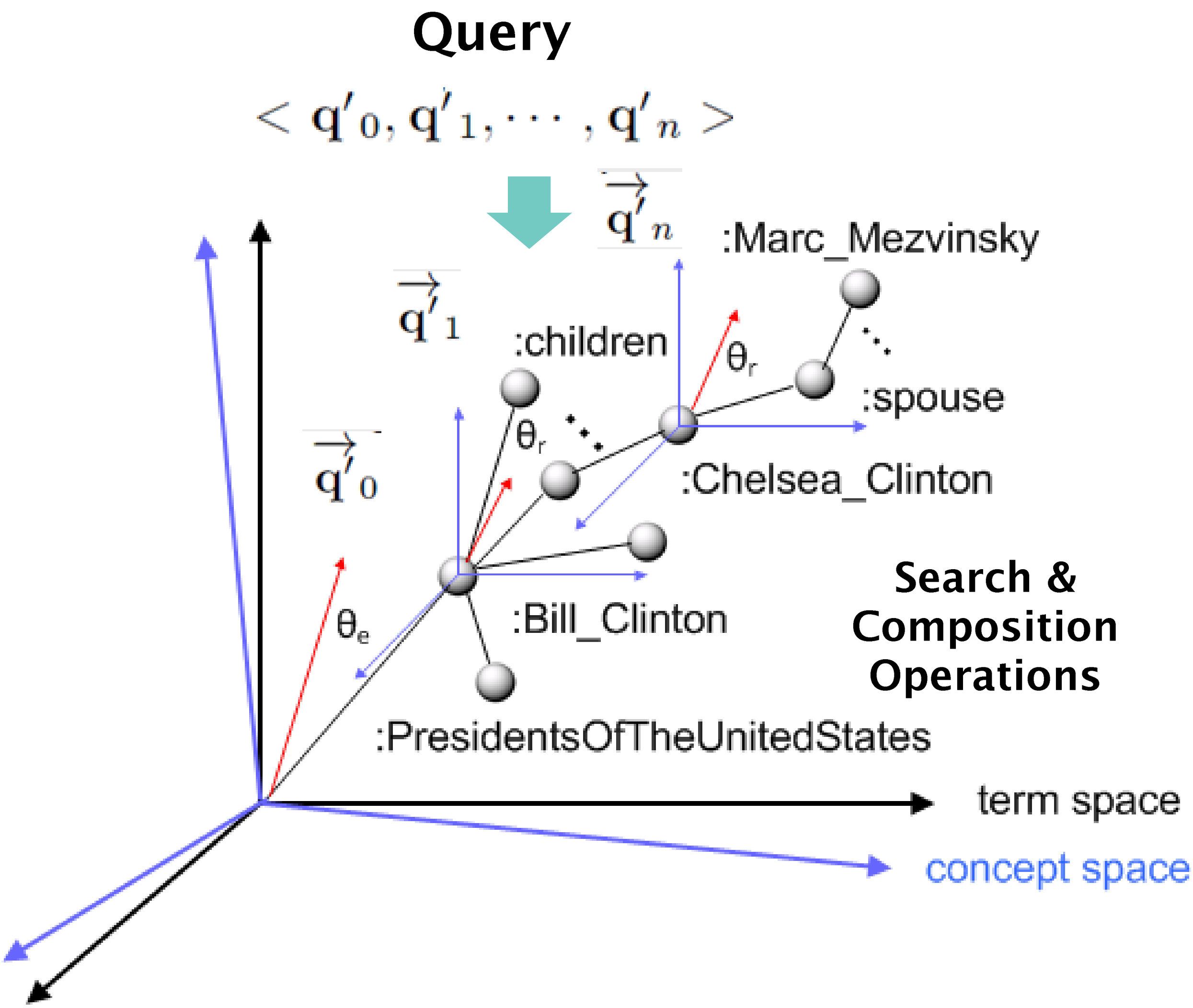
Core Operations

Query

$$< \mathbf{q'}_0, \mathbf{q'}_1, \cdots, \mathbf{q'}_n >$$



Core Operations



Search and Composition Operations

- **Instance search**
  - Proper nouns
  - String similarity + node cardinality
- **Class (unary predicate) search**
  - Nouns, adjectives and adverbs
  - String similarity + Distributional semantic relatedness
- **Property (binary predicate) search**
  - Nouns, adjectives, verbs and adverbs
  - Distributional semantic relatedness

$$sr(\overrightarrow{\mathbf{q}'}_1, \overrightarrow{\mathbf{p}}_0) \geq \eta$$

- **Navigation**

$$< (\overrightarrow{\mathbf{q}'}_1 - \overrightarrow{\mathbf{p}}_1), (\overrightarrow{\mathbf{q}'}_2 - \overrightarrow{\mathbf{p}}_2), \cdots, (\overrightarrow{\mathbf{q}'}_n - \overrightarrow{\mathbf{p}}_n) >$$

- **Extensional expansion**
  - Expands the instances associated with a class.
- **Operator application**
  - Aggregations, conditionals, ordering, position
- **Disjunction & Conjunction**
- **Disambiguation dialog (instance, predicate)**

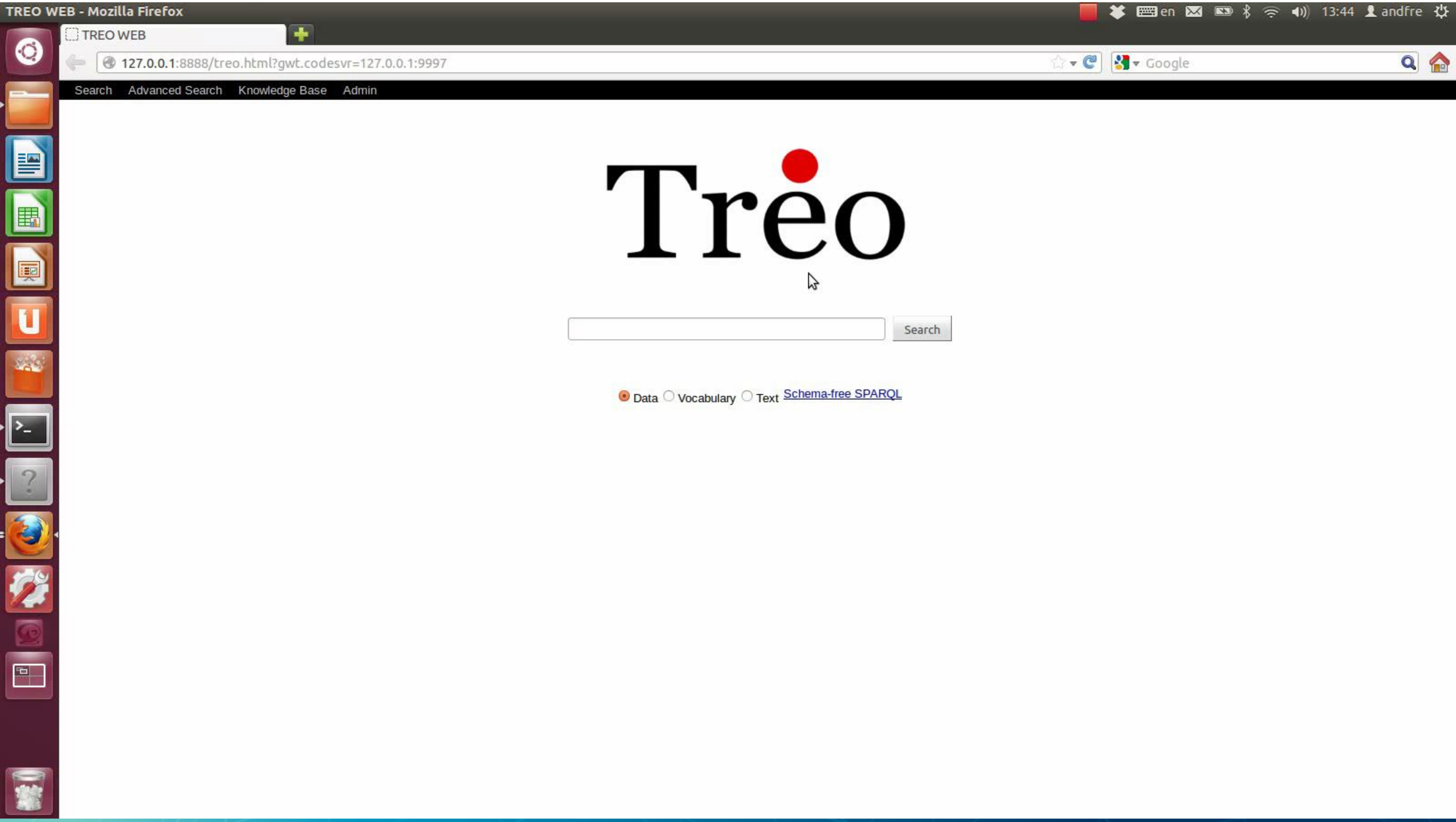
Does it work?

Addressing the Vocabulary Problem for Databases (with Distributional Semantics)

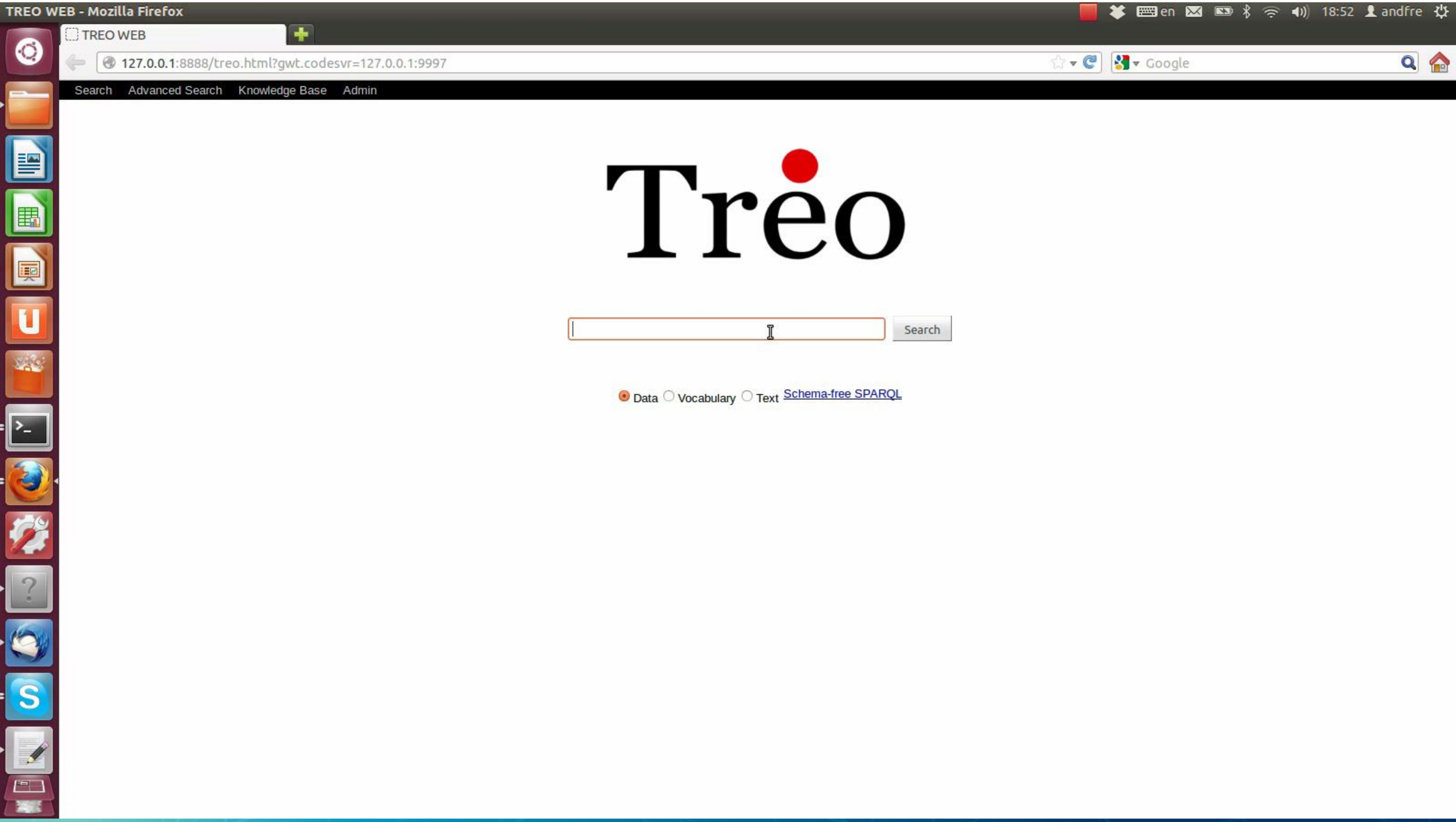
Treo

Gaelic: direction

Simple Queries (Video)



More Complex Queries (Video)



Terminology-level Search (Video)

Query Pre-Processing (Query Analysis)

- Transform natural language queries into triple patterns.

“Who is the daughter of Bill Clinton married to?”

Query Pre-Processing (Query 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 (Query Analysis)

- **Step 2: Semantic Pivot Recognition**
  - Rules-based: POS Tags + 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**

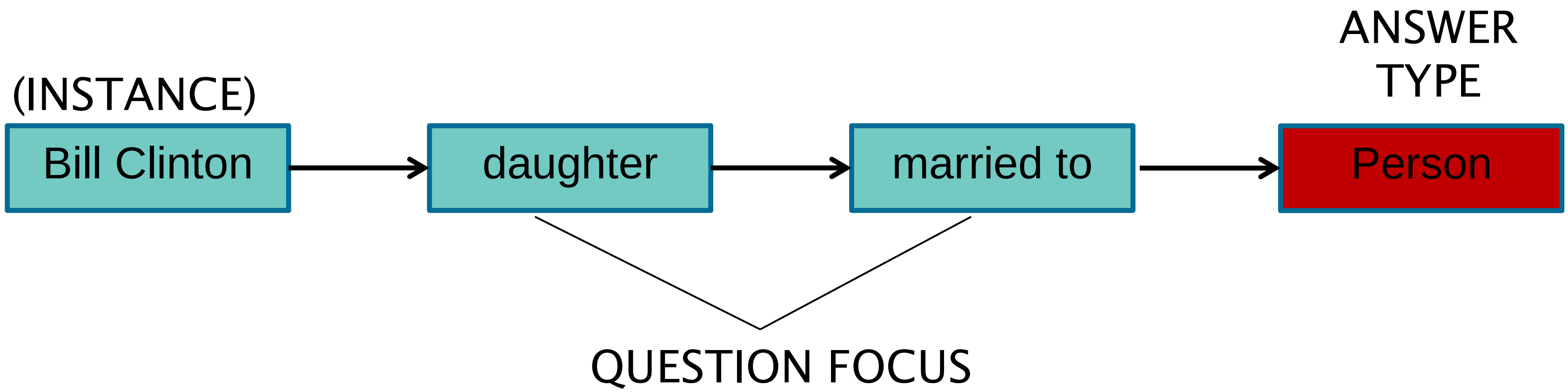
- dep(married-8, Who-1)
- auxpass(married-8, is-2)
- det(daughter-4, the-3)
- nsubjpass(married-8, daughter-4)
- prep(daughter-4, of-5)
- nn(Clinton-7, Bill-6)
- pobj(of-5, Clinton-7)
- root(ROOT-0, married-8)
- xcomp(married-8, to-9)

Query Pre-Processing (Question Analysis)

- **Step 5: Determine Partial Ordered Dependency Structure (PODS)**

  - Rules based.

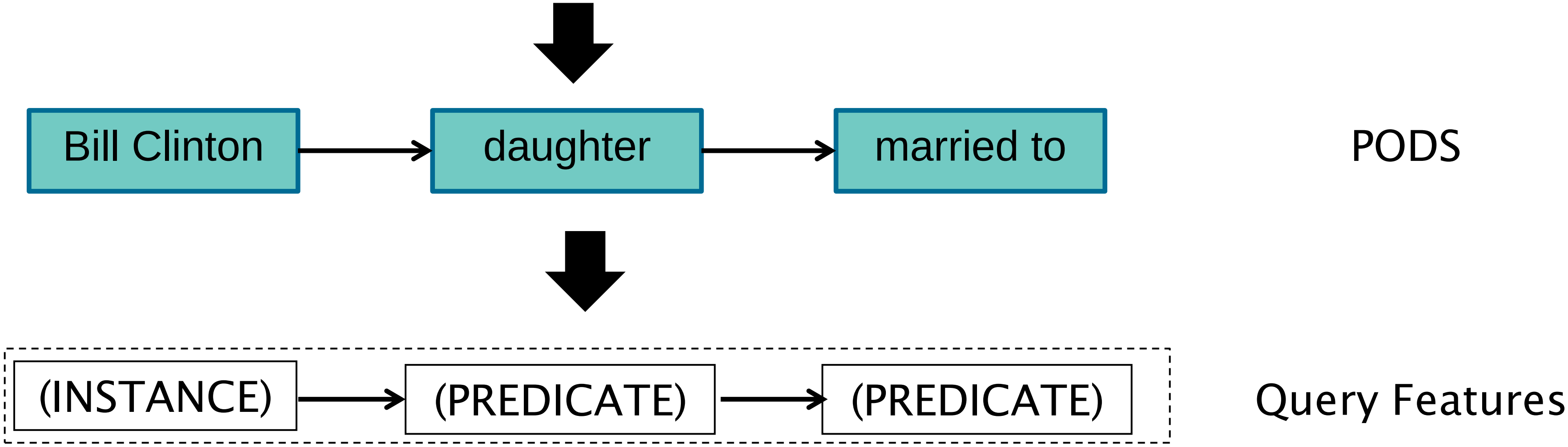
- Remove stop words.
- Merge words into entities.
- Reorder structure from core entity position.



Question Analysis

Transform natural language queries into triple patterns

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



Query Plan

Map *query features* into a *query plan*.

A *query plan* contains a sequence of *core operations*.

(INSTANCE) → (PREDICATE) → (PREDICATE)

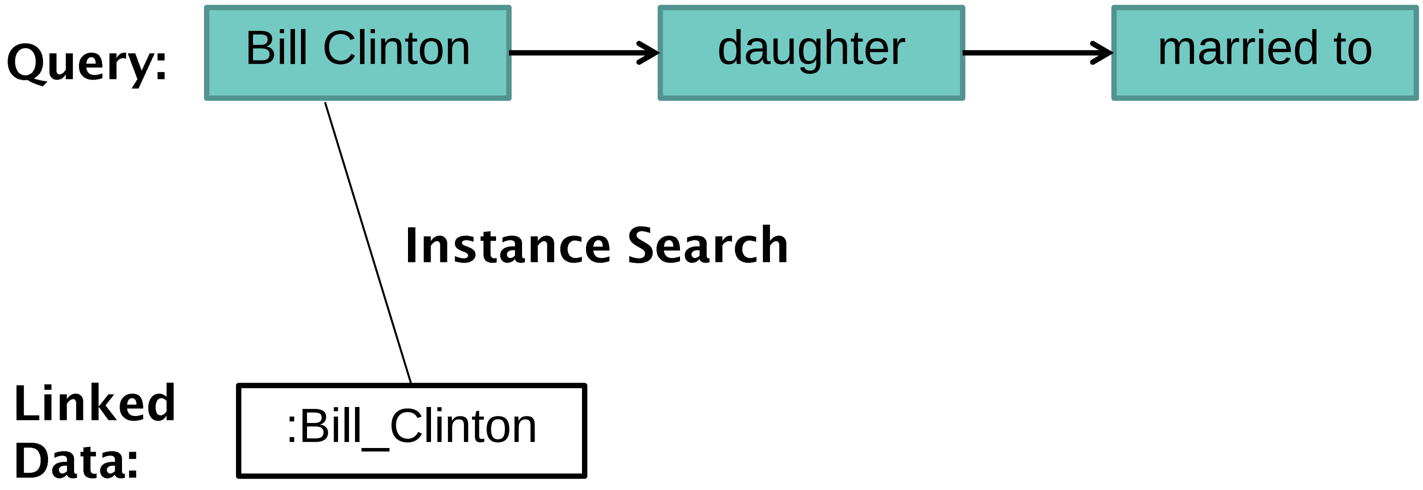
Query Features

- (1) INSTANCE SEARCH (Bill Clinton)
- (2) p_1 <- SEARCH PREDICATE (Bill Clintion, daughter)
- (3) e_1 <- NAVIGATE (Bill Clintion, p_1)
- (4) p_2 <- SEARCH PREDICATE (e_1, married to)
- (5) e_2 <- NAVIGATE (e_1, p_2)

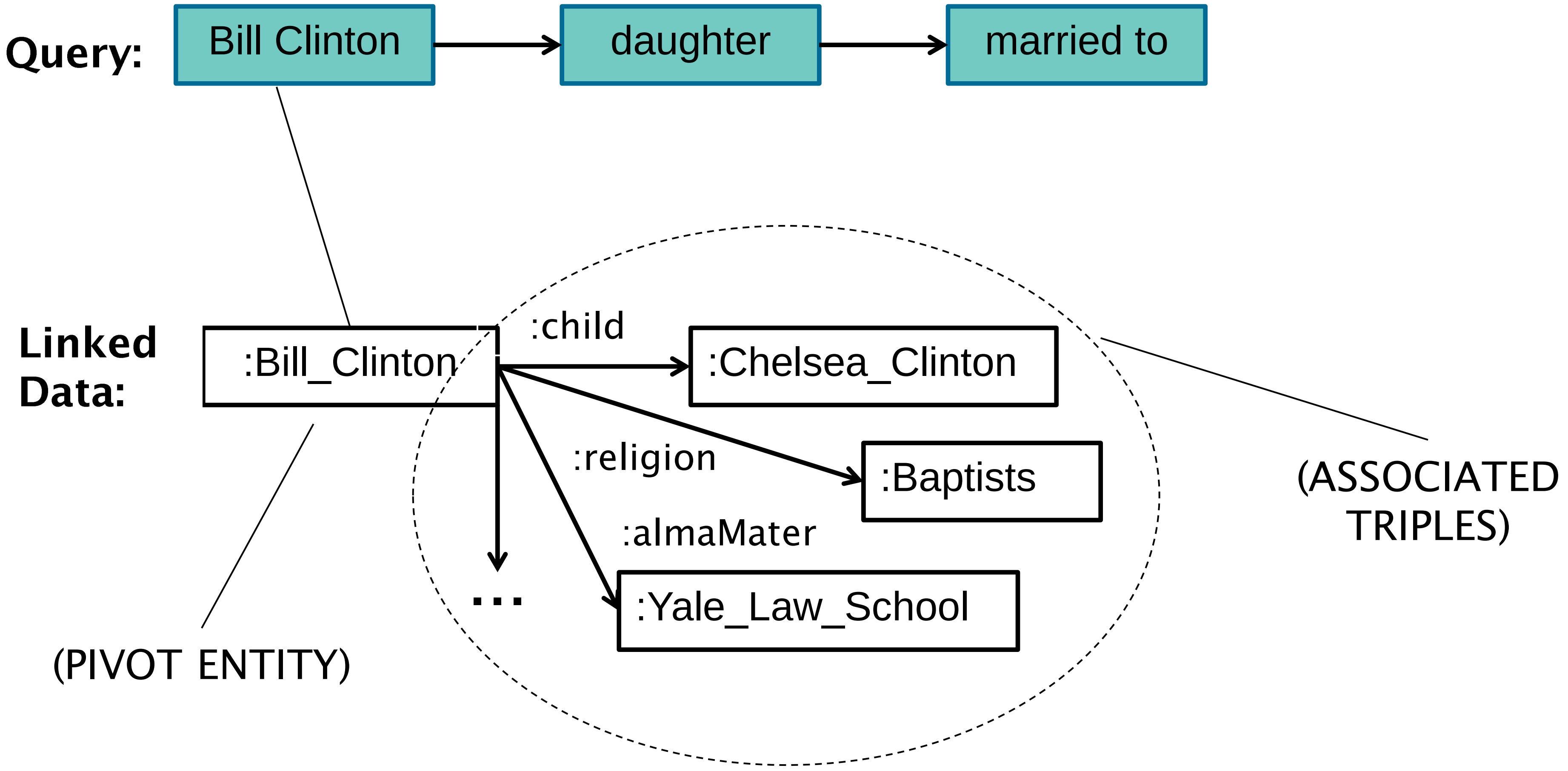
Query Plan

Query Plan Execution

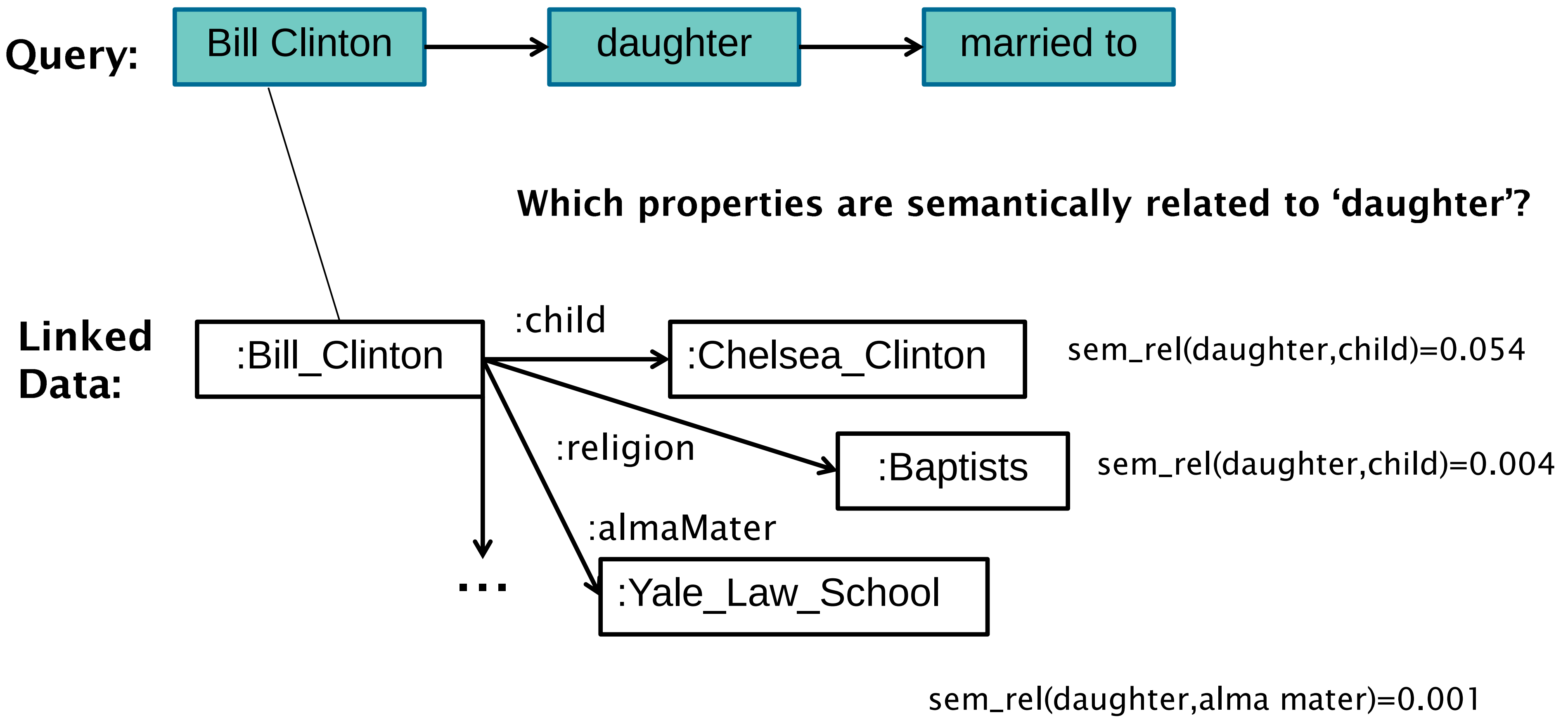
Instance Search



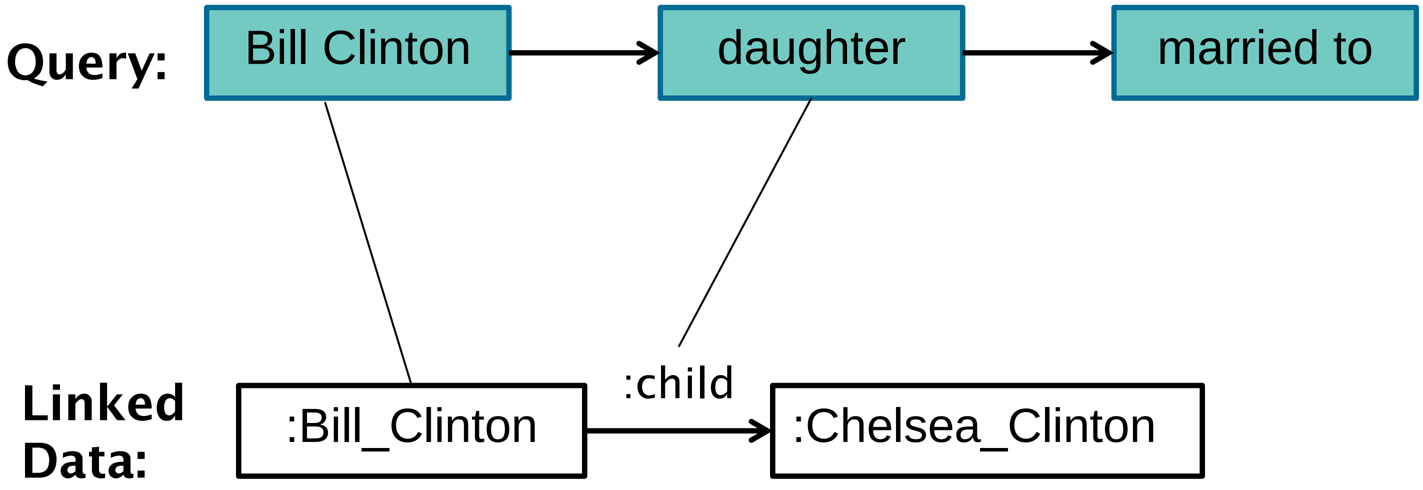
Predicate Search



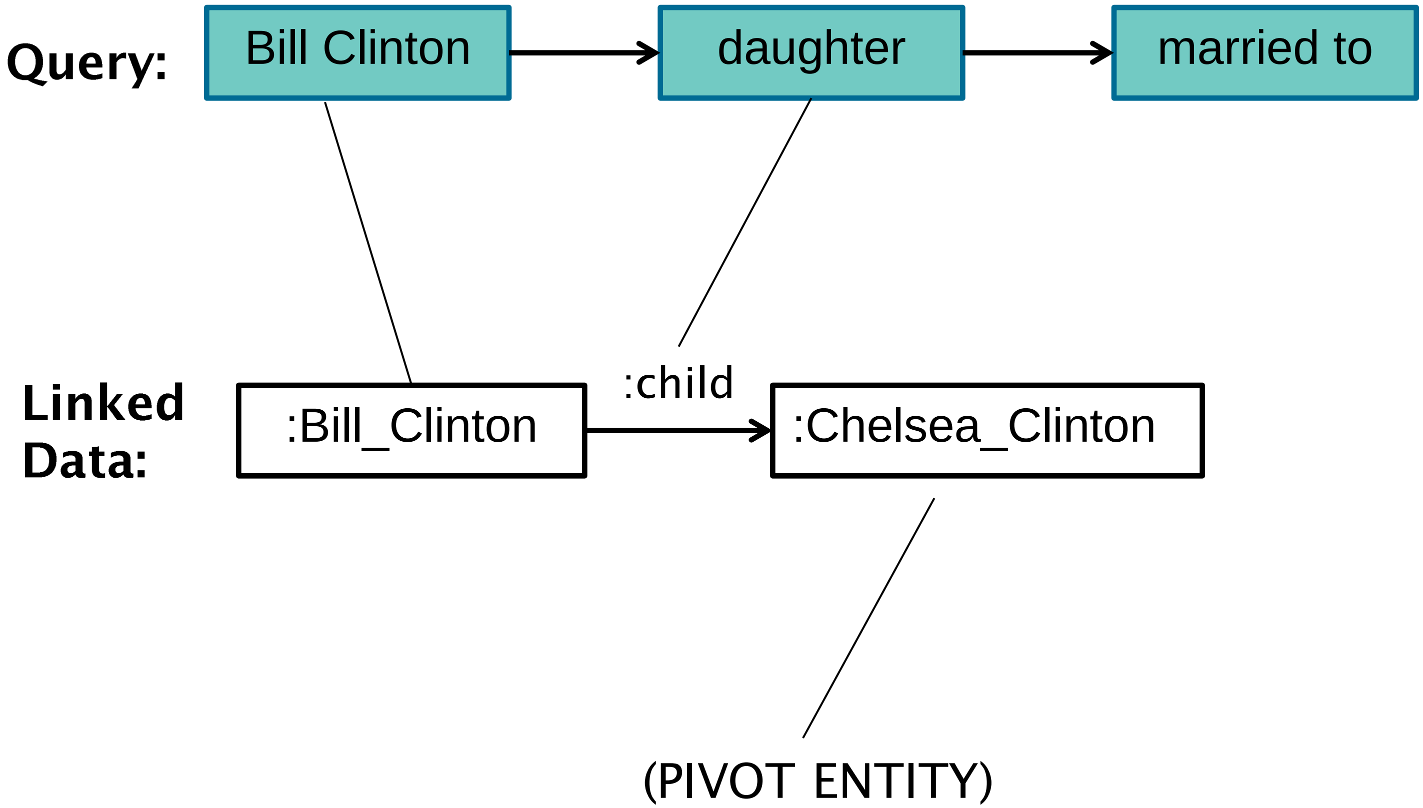
Predicate Search



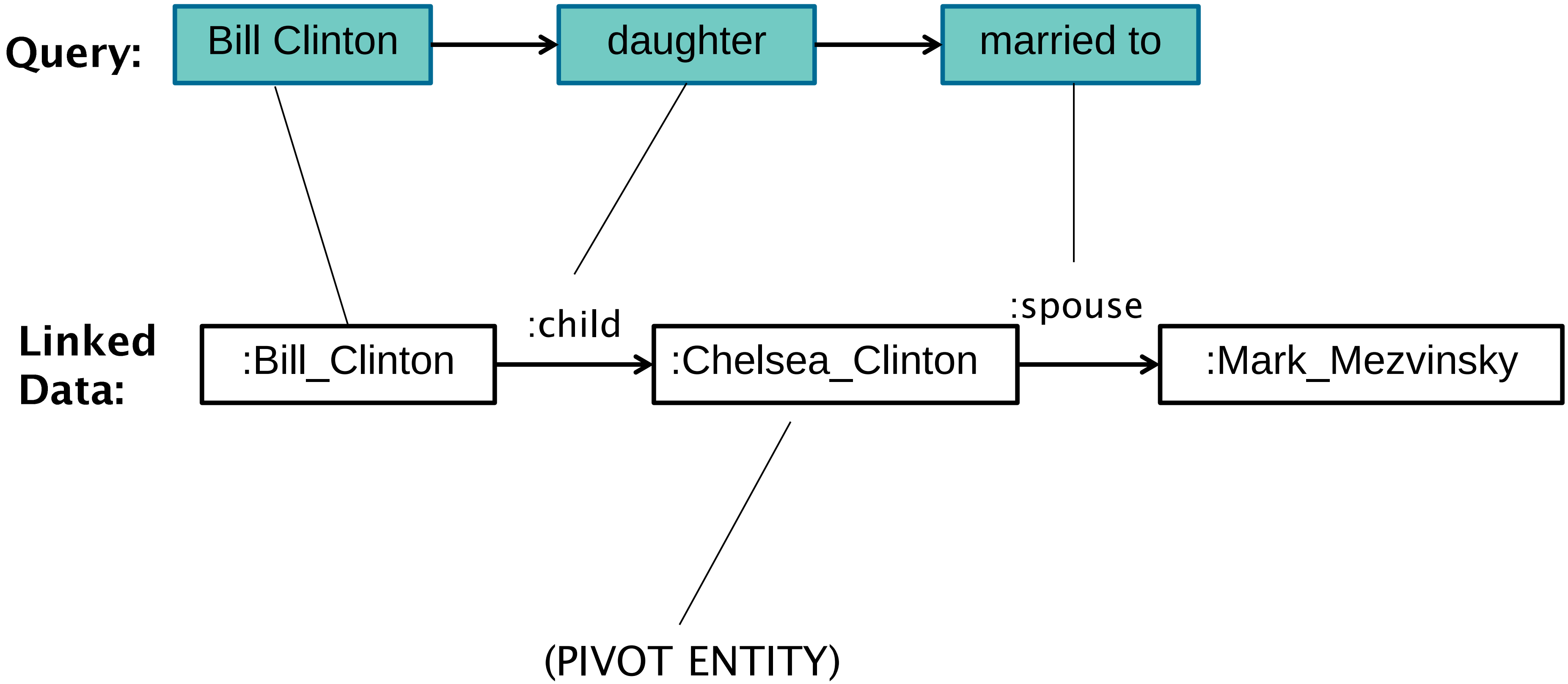
Navigate


Query:
Bill Clinton
daughter
married to
Linked Data:
:Bill_Clinton
:child
:Chelsea_Clinton

Navigate


Query:
Bill Clinton
daughter
married to
Linked Data:
:Bill_Clinton
:child
:Chelsea_Clinton
(PIVOT ENTITY)

Predicate Search



Results

ch Advanced Search Knowledge Base Admin

Treo

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

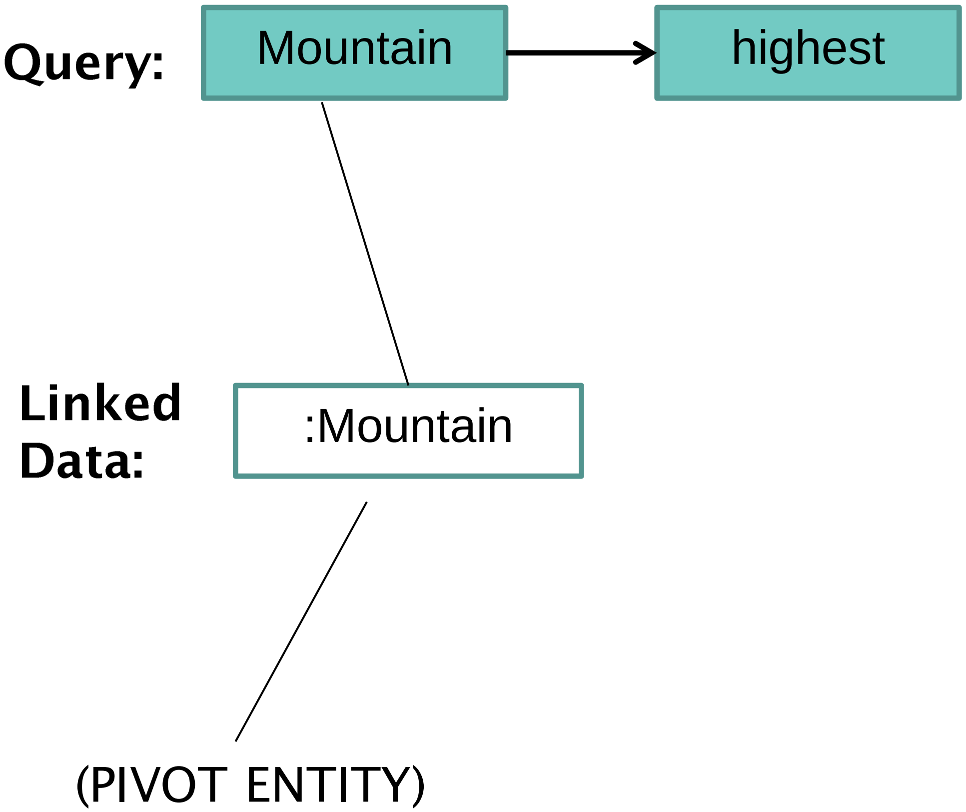
Bill Clinton children Chelsea Clinton

William Jefferson Blythe, Jr. child Bill Clinton

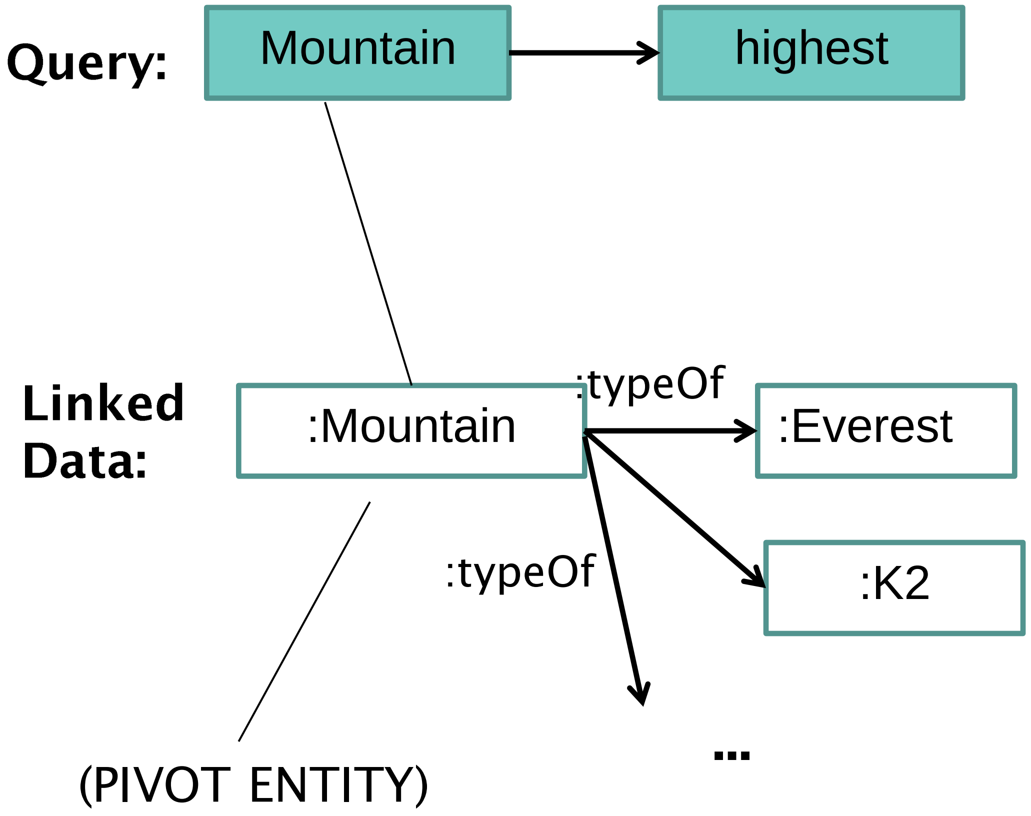
Virginia Clinton Kelley child Bill Clinton

Virginia Clinton Kelley children Bill Clinton

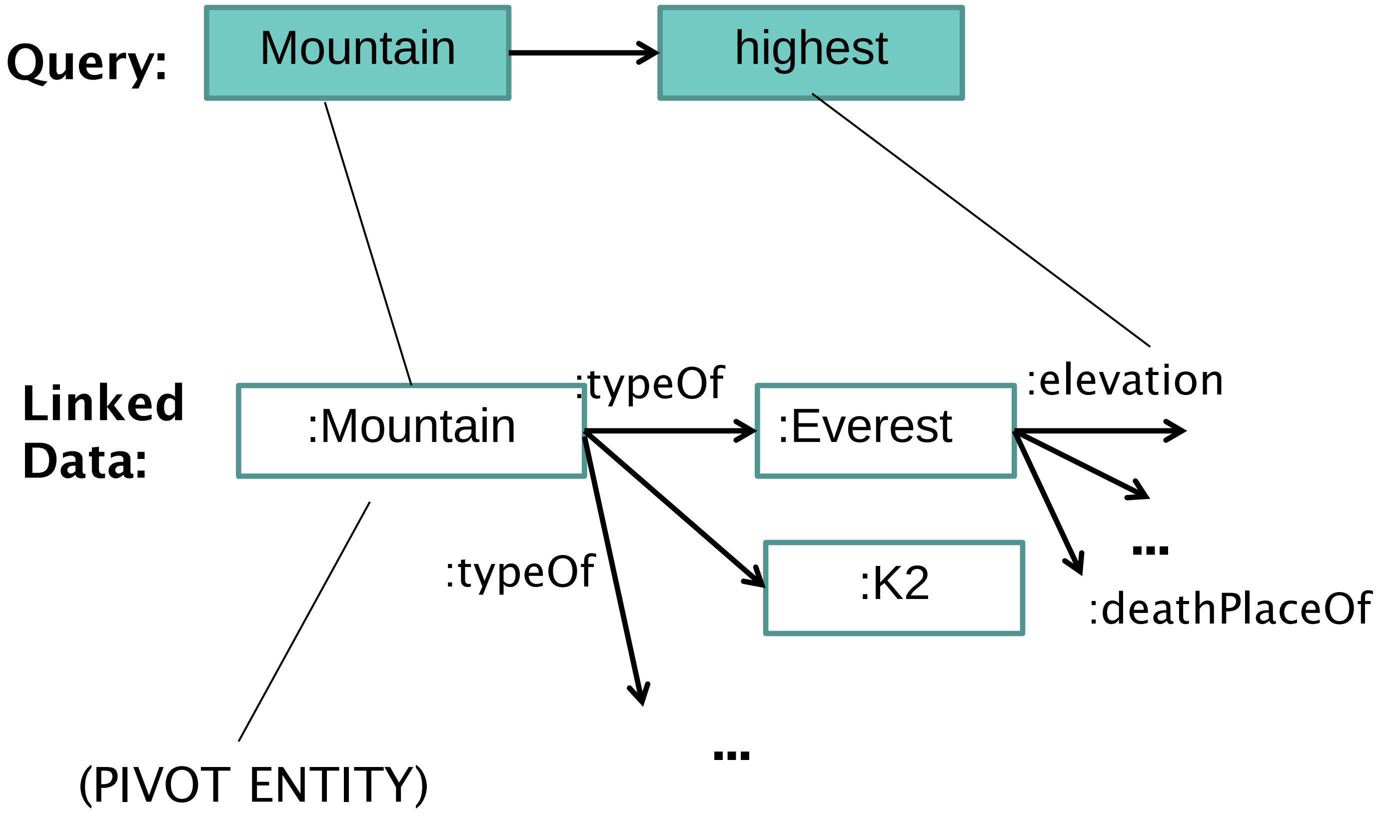
Class (Unary Predicate) Search



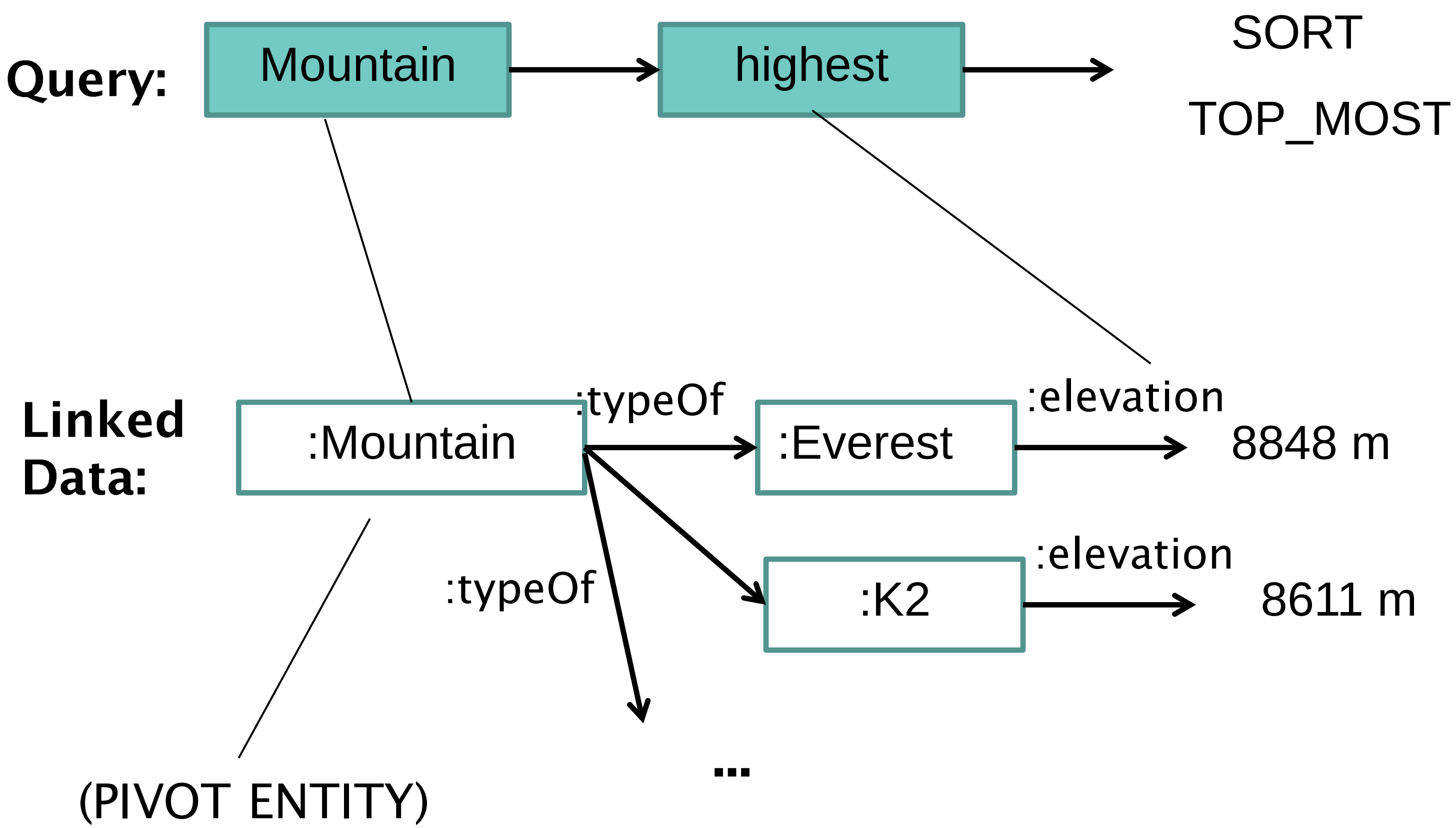
Extensional Expansion



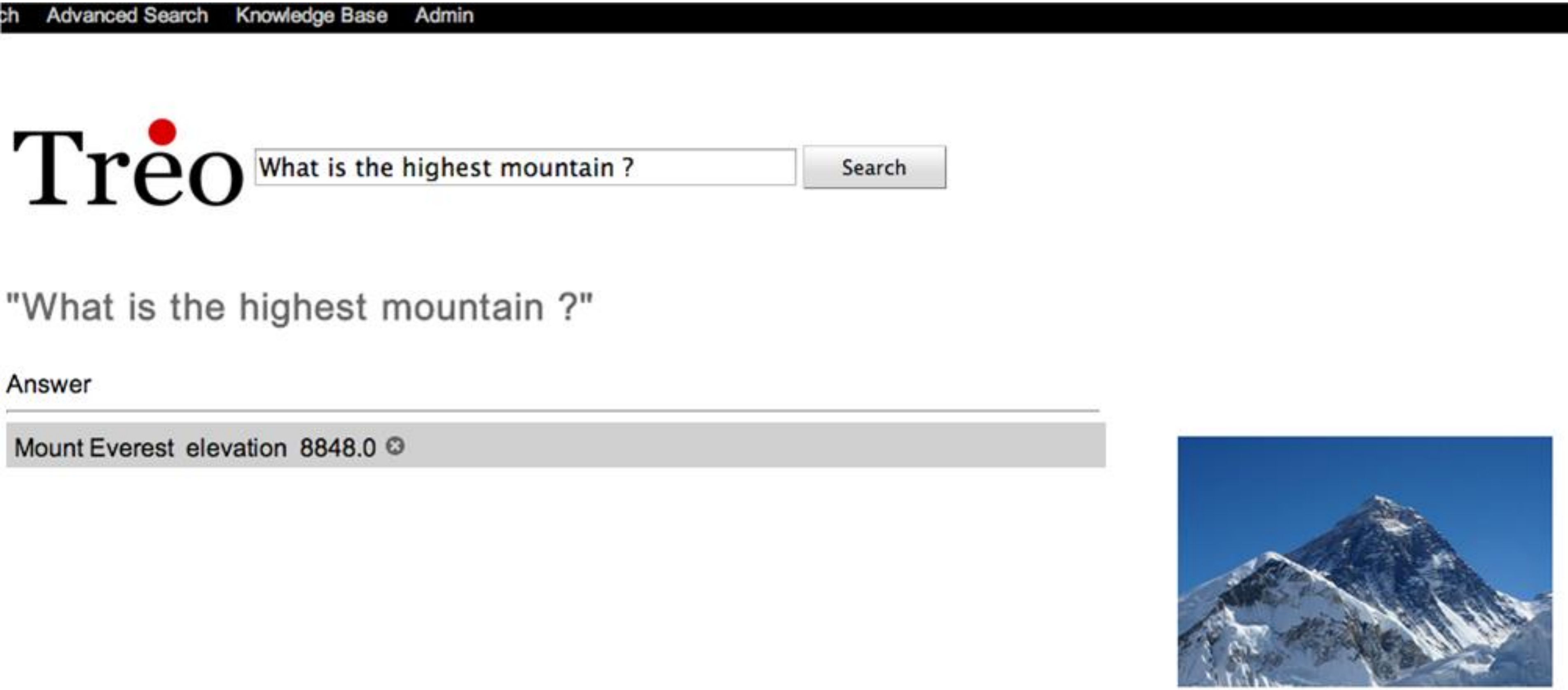
Distributional Semantic Matching



Application of the functional definition of the operator



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

Evaluation

Test Collection

- Test Collection: QALD 2011.
- DBpedia 3.6.
- Two test sets (76/50) natural language queries.

Dataset (DBpedia 3.6 + YAGO classes):
45,768 properties
288,316 classes
9,434,677 instances
128,071,259 triples

Test Collection Analysis

- QALD 2011, DBpedia 3.6

Hypotheses

Test Collection Requirements	QALD 2011 Coverage
Dataset size & semantic heterogeneity	High
Comprehensive query set	Medium-high
Query-Dataset semantic gap	High
Realistic & Representative query set collection	Medium-high

Relevance of Results

- R1. High usability
- R2. High query expressivity
- R3. Accurate & comprehensive semantic matching

Precision, recall, mean reciprocal rank, % of answered queries.

Relevance

Relevance			
Avg. Precision	Avg. Recall	MRR	% of queries answered
0.62	0.81	0.49	80%

Accurate semantic matching for a semantic best-effort scenario

Ranking in the second position in average

Medium-high query expressivity / coverage

Comparative Analysis

System	Avg. R	MAP	% answered queries
Treo	**0.79**	**0.63**	**79%**
PowerAqua	0.54	0.63	48%
FREyA	0.48	0.52	54%
Unger et al.	0.63	0.61	-

- Better recall and query coverage compared to baselines with equivalent precision.
- More comprehensive semantic matching.

Evaluating Terminology-level Semantic Matching

- R3. Accurate & comprehensive semantic matching

Quantitative evaluation: P@5, P@10, MRR, % of the queries answered, comparative evaluation using string matching and WordNet-based query expansion.

Evaluating Terminology-level Semantic Matching

Avg. Precision@5	Avg. Precision@10	MRR	% of queries answered
0.732	0.646	0.646	92.25%

Approach	% of queries answered
ESA	92.25%
String matching	45.77%
String matching + WordNet QE	52.48%

- Distributional semantics provides a more comprehensive semantic matching with medium-high precision

Performance & Adaptability

- R4. Low maintainability/adaptability effort
- R5. Low query execution time
- R6. High scalability

Avg. 1.52 s (simple queries)
Avg. 8.53 s (all queries)

Measure	value
Avg query execution time (ms)	8,530
Avg. entity search time (ms)	3,495
Avg. predicate search time (ms)	3,223
Avg. number of search operations per query	2.70
Avg. index insert time per triple (ms)	5.35
Avg. index size per triple (bytes)	250
Dataset adaptation effort (minutes)	0.00
Dataset specific semantic enrichment effort per query (secs)	0.00
Dataset specific semantic enrichment effort (minutes)	0.00

- Interactive query execution time

- Indexing size overhead (20% of the dataset size)
- Significant overhead in indexing time.

- Low adaptability effort

Requirements Coverage

Requirement	Coverage	Suitability of the Evaluation Setup
High usability & Low query construction time	High	Usability not explicitly covered in the evaluation - Intrinsic to open natural language interfaces
High query expressivity	Medium-high	Medium-High
Accurate semantic matching	Medium-high	High
Comprehensive semantic matching	High	High
Low setup & maintainability effort	High	High
Interactive search & Low query-execution time	Medium-high	High
High scalability	Medium	Medium-low

Beyond Schema-Agnosticism

- **How does schema-agnosticism affect programming?**
- *Towards An Approximative Ontology-Agnostic Approach for Logic Programs*, FOIKS 2014.
- **How the distributional-relational model can be used for reasoning over incomplete knowledge bases?**
- *A Distributional Semantics Approach for Selective Reasoning on Commonsense Graph Knowledge Bases*, NLDB 2014.

Dissemination

Dissemination Summary

- Initial approach & early evaluation (2011)
- Ƭ-Space Knowledge Representation Model (2011, 2012, 2013)
- Compositional Model (2012, 2013, 2014)
- Full evaluation (2014)
- Demonstrations (2013)
- Hybrid QA (2013)
- Extensions: Approximate Reasoning (2014)
- Formalization of Schema-agnosticism (2014,2015)
- 3 Tutorials, 2 Workshops & 1 Semantic Web Challenge (2013, 2014, 2015)

Publications

André Freitas, Edward Curry, *Natural Language Queries over Heterogeneous Linked Data Graphs: A Distributional-Compositional Semantics Approach*, In Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI), Haifa, 2014.

André Freitas, Rafael Vieira, Edward Curry, Danilo Carvalho, João Carlos Silva, *On the Semantic Representation and Extraction of Complex Category Descriptors*, In Proceedings of the 19th International Conference on Applications of Natural Language to Information Systems (NLDB), Montpellier, 2014.

André Freitas, João C. P. da Silva, Sean O'Riain, Edward Curry, Distributional Relational Networks, AAAI Fall Symposium, Arlington, 2013.

André Freitas, Edward Curry, *Do it yourself (DIY) Jeopardy QA System*, In Proceedings of the 12th International Semantic Web Conference (ISWC), Sydney, 2013.

André Freitas, João Carlos Pereira Da Silva, Edward Curry, Paul Buitelaar, A Distributional Semantics Approach for Selective Reasoning on Commonsense Graph Knowledge Bases, In Proceedings of the 19th International Conference on Applications of Natural Language to Information Systems (NLDB), Montpellier, 2014.

João C. Pereira da Silva, André Freitas, Towards An Approximative Ontology-Agnostic Approach for Logic Programs, In Proceedings of the Eighth International Symposium on Foundations of Information and Knowledge Systems (FoIKS), Bordeaux, 2014.

Publications

André Freitas, Juliano Efson Sales, Siegfried Handschuh, Edward Curry, How hard is this query? Measuring the Semantic Complexity of Schema-agnostic Queries, IWCS 2015.

André Freitas, João Carlos Pereira Da Silva, Edward Curry, On the Semantic Mapping of Schema-agnostic Queries: A Preliminary Study, Workshop of the Natural Language Interfaces for the Web of Data (NLIWoD), 13th International Semantic Web Conference (ISWC), Rival del Garda, 2014.

André Freitas, Siegfried Handschuh, Edward Curry, Distributional-Relational Models: Scalable Semantics for Databases, AAAI Spring Symposium, Knowledge Representation & Reasoning Track, Stanford, 2014.

André Freitas, João C. Pereira da Silva, Semantics at Scale: When Distributional Semantics meets Logic Programming, ALP Newsletter, 2014.

André Freitas, Edward Curry, Siegfried Handschuh, Towards a Distributional Semantic Web Stack, 10th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2014), 13th International Semantic Web Conference (ISWC), Rival del Garda, 2014.

Publications

André Freitas, Edward Curry, João Gabriel Oliveira, João C. Pereira da Silva, Sean O'Riain, *Querying the Semantic Web using Semantic Relatedness: A Vocabulary Independent Approach*. Data & Knowledge Engineering (DKE) Journal, 2013.

André Freitas, Fabricio de Faria, Sean O'Riain, Edward Curry, *Answering Natural Language Queries over Linked Data Graphs: A Distributional Semantics Approach*, In Proceedings of the 36th Annual ACM SIGIR Conference, Dublin, Ireland, 2013.

André Freitas, Sean O'Riain and Edward Curry, *A Distributional Semantic Search Infrastructure for Linked Dataspaces*, In Proceedings of the 10th Extended Semantic Web Conference (ESWC), Montpellier, France, 2013.

André Freitas, Sean O'Riain and Edward Curry, *Crossing the Vocabulary Gap for Querying Complex and Heterogeneous Databases: A Distributional-Compositional Semantics Perspective*, 3rd Workshop on Data Extraction and Object Search (DEOS), 29th British National Conference on Databases (BNCOD), Oxford, UK, 2013.

André Freitas, João C. Pereira da Silva, Danilo S. Carvalho, Sean O'Riain, Edward Curry, Representing Texts as Contextualized Entity-Centric Linked Data Graphs, 12th International Workshop on Web Semantics and Web Intelligence (WebS 2013), 24th International Conference on Database and Expert Systems Applications (DEXA), Prague, 2013.

André Freitas, Edward Curry, João Gabriel Oliveira, Sean O'Riain, *Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches and Trends*. IEEE Internet Computing, Special Issue on Internet-Scale Data, 2012.

Publications

André Freitas, Edward Curry, João Gabriel Oliveira, Sean O'Riain, *A Distributional Structured Semantic Space for Querying RDF Graph Data.* International Journal of Semantic Computing (IJSC), 2012.

André Freitas, Sean O'Riain, Edward Curry, *A Distributional Approach for Terminological Semantic Search on the Linked Data Web*. 27th ACM Applied Computing Symposium, Semantic Web and Its Applications Track, 2012.

André Freitas, João Gabriel Oliveira, Edward Curry, Sean O'Riain, *A Multidimensional Semantic Space for Data Model Independent Queries over RDF Data*. In Proceedings of the 5th International Conference on Semantic Computing (ICSC), 2011.

André Freitas, João Gabriel Oliveira, Sean O'Riain, Edward Curry, João Carlos Pereira da Silva, *Querying Linked Data using Semantic Relatedness: A Vocabulary Independent Approach.* In Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB) 2011.

André Freitas, João Gabriel Oliveira, Sean O'Riain, Edward Curry, João Carlos Pereira da Silva, *Treo: Combining Entity-Search, Spreading Activation and Semantic Relatedness for Querying Linked Data*, In 1st Workshop on Question Answering over Linked Data (QALD-1) Workshop at 8th Extended Semantic Web Conference (ESWC), 2011.

André Freitas, João Gabriel Oliveira, Sean O'Riain, Edward Curry, João Carlos Pereira da Silva*, Treo: Best-Effort Natural Language Queries over Linked Data,* In Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB), 2011.

Core Contributions

- Definition and evaluation of the schema-agnostic query approach based on distributional semantics.
- Creation of a natural language interface(NLI)/question answering(QA) system over RDF data.
- Definition of a distributional-relational semantic representation model (Ƭ-Space) and the associated index model.
- Discussion of two application scenarios for the proposed semantic approximation model on logic programming and commonsense reasoning over incomplete knowledge bases.

Limitations

- Lack of ranking and backtracking of multiple query plans.
- Lack of evaluation of the suitability of distributional semantic models for domain specific datasets.
- Lack of evaluation over multiple datasets.
- Lack of scalability evaluation.

Future Research Directions

- Investigation of uncertainty models for distributional-relational models.
- Formalization of the distributional-relational algebra & query optimization approaches.
- More comprehensive and systematic comparative study of distributional semantic models for open domain schema-agnostic queries.

Conclusions

Problem: Schema-agnosticism is fundamental for large-schema databases.

Approach:

- The compositional-distributional model supports a schema-agnostic query mechanism over a large schema (open domain) database.
- Semantic pivoting + distributional semantics + semantic-best-effort.

Evaluation:

- Semantic matching
  - Avg. recall=0.81, map=0.62, mrr=0.49
- Expressivity
  - 80% of queries answered
- Interactive query execution time
  - Avg. 1.52 s (simple queries) – 8.53 s (all queries) / query
- Better recall and query coverage compared to baselines with equivalent precision to the second best-performing approach.
- Low adaptation effort.

Any (Schema-agnostic) Queries?