DeFacto

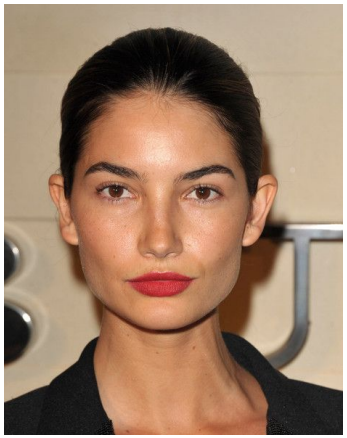
Temporal and Multilingual Deep Fact Validation

Daniel Gerber, Diego Esteves, Jens Lehmann, Lorenz Bühmann,
Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, René Speck

Institute for Applied Informatics
(InfAI)

October 19, 2016

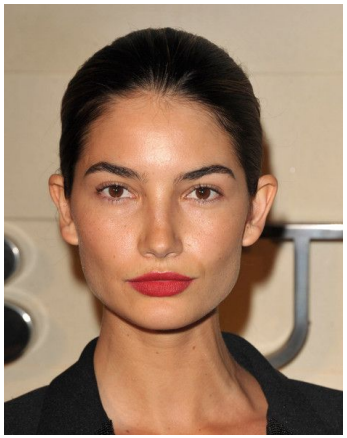
Introduction



Lily Aldridge

is married to

Introduction

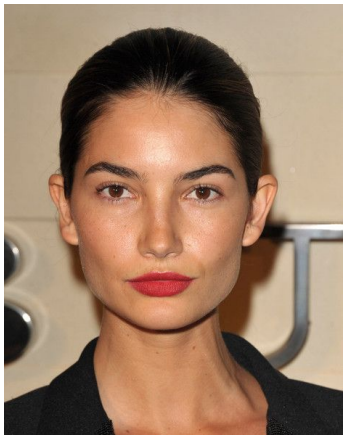


Lily Aldridge

is married to

Kings of Leon

Introduction



Lily Aldridge

is married to

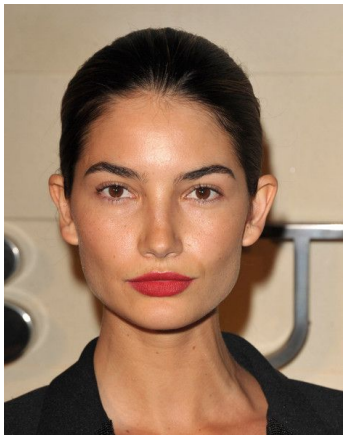
Kings of Leon

Arthur Shrewsbury

's death place is

the English Cricket Team

Introduction



Lily Aldridge

is married to

Kings of Leon

Arthur Shrewsbury

's death place is

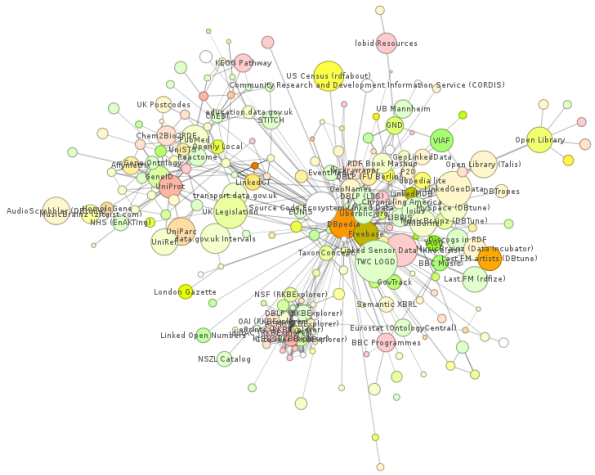
the English Cricket Team

- Michelle Obama is married to the presidency of Barack Obama
- People can be born in Lacrosse, the sport ...
- ...

Introduction

Problem

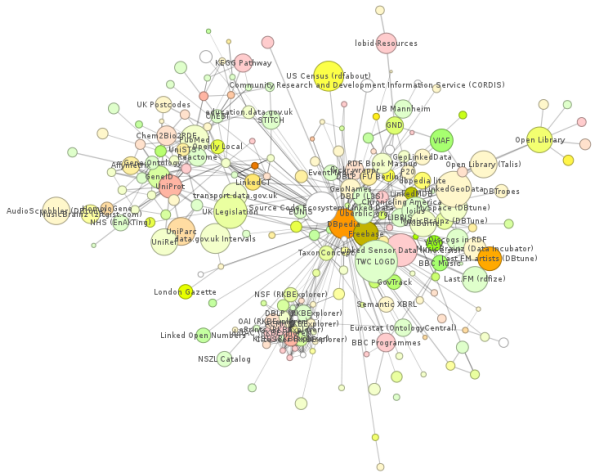
- 130+ billion facts
- Automatically generated facts
- Manual checking too tedious



Introduction

Problem

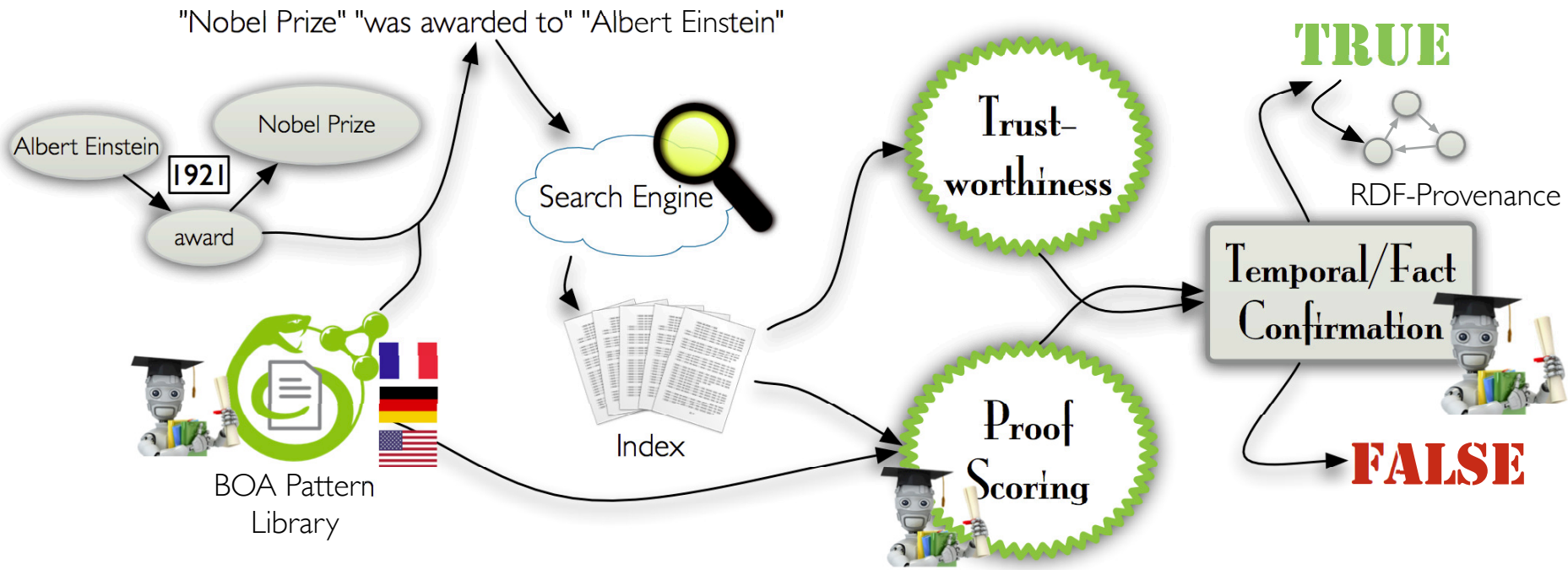
- 130+ billion facts
- Automatically generated facts
- Manual checking too tedious



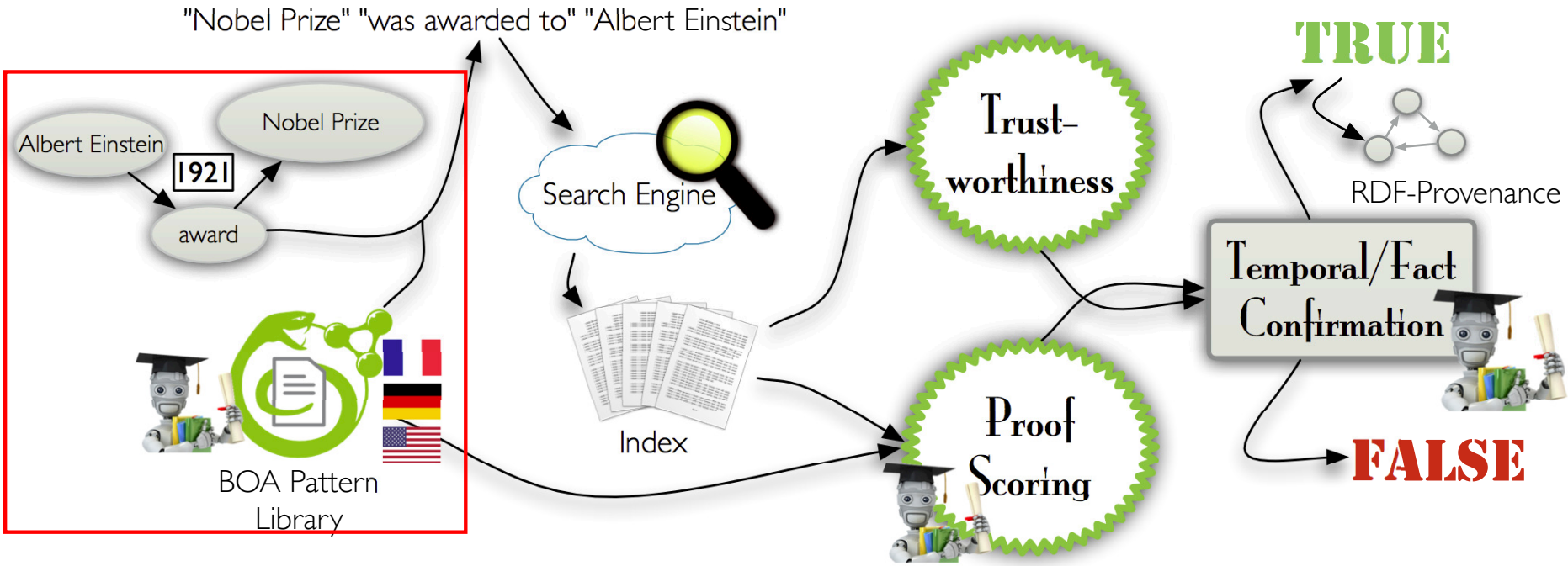
Solution

- Deep Fact Validation
- Use verbalization and ML to find evidence for facts being true
- Support temporal scoping

Overview

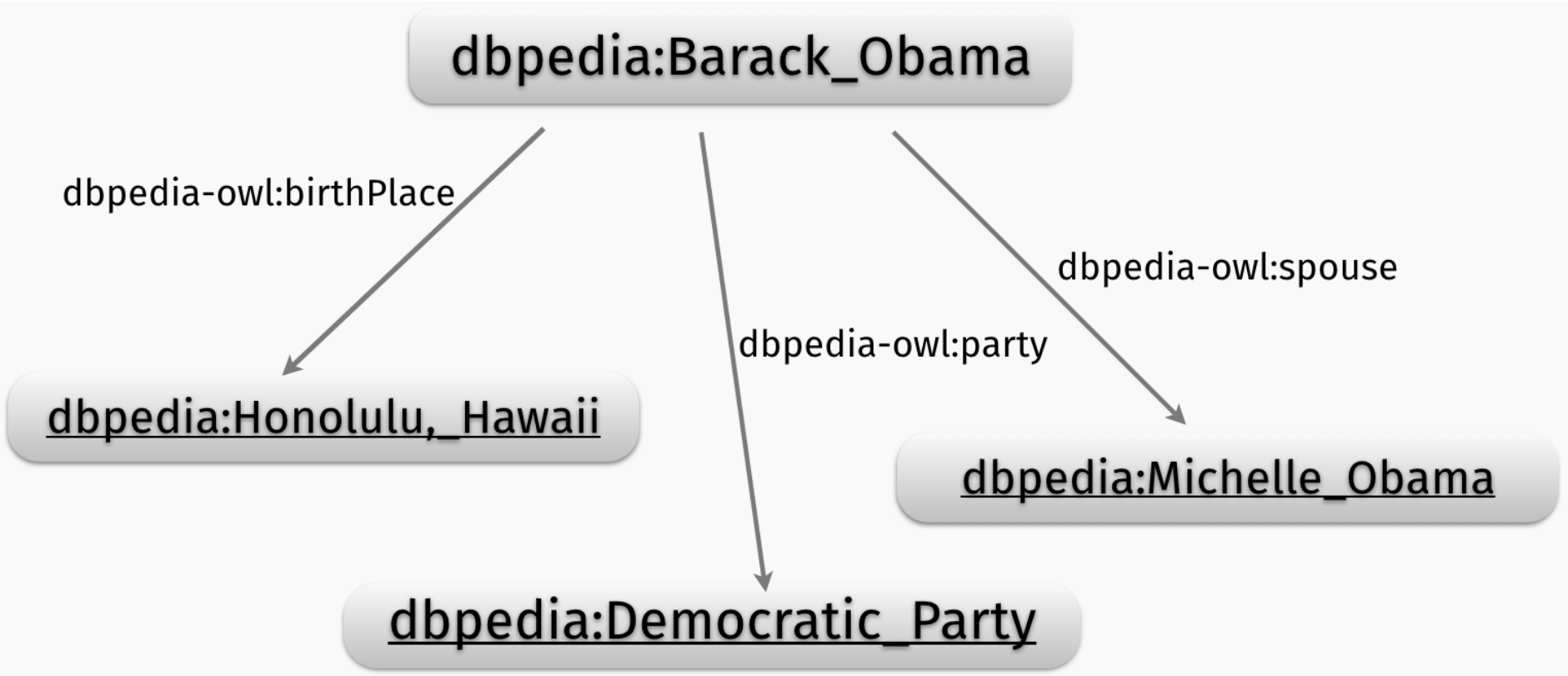

"Nobel Prize" "was awarded to" "Albert Einstein"
Albert Einstein
1921
award
Nobel Prize
Search Engine
Index
BOA Pattern Library
Trust-worthiness
Proof Scoring
Temporal/Fact Confirmation
TRUE
RDF-Provenance
FALSE

Overview


"Nobel Prize" "was awarded to" "Albert Einstein"
Albert Einstein
Nobel Prize
1921
award
BOA Pattern Library
Search Engine
Index
Trust-worthiness
Proof Scoring
Temporal/Fact Confirmation
TRUE
RDF-Provenance
FALSE

Idea

- Billions of triples available
- Apply distance learning to extract RDF

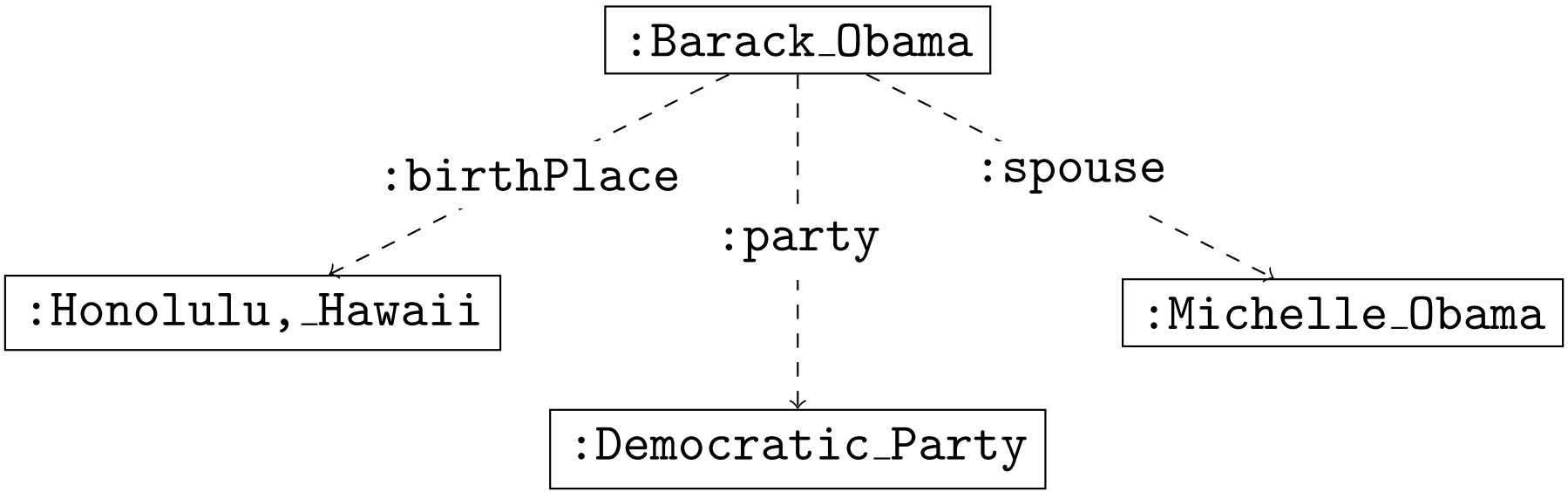


Idea

- Billions of triples available
- Apply distance learning to extract RDF

Idea

- Billions of triples available
- Apply distance learning to extract RDF



Barack Obama was born in Honolulu, Hawaii.

Barack Hussein Obama is a politician of the Democratic Party.

Obama married Michelle Robinson in 1992.

Barack Obama was born in Honolulu, Hawaii.

Barack Hussein Obama is a politician of the Democratic Party.

Obama married Michelle Robinson in 1992.

Barack Obama was born in Honolulu, Hawaii.

Barack Hussein Obama is a politician of the Democratic Party.

Obama married Michelle Robinson in 1992.

was born in

Dietrich's only child, Maria Elisabeth Sieber, *was born in* Berlin on 13 December 1924.

is a politician of the

Joseph Martin "Joschka" Fischer (born 1948-04-12) *is a politician of the* German Green Party.

married

Jackie Bouvier Kennedy Onassis who *married* John F. Kennedy was tied to the Auchinclosses via her sister's marriage into the Auchincloss family.

D *with his wife* **R**

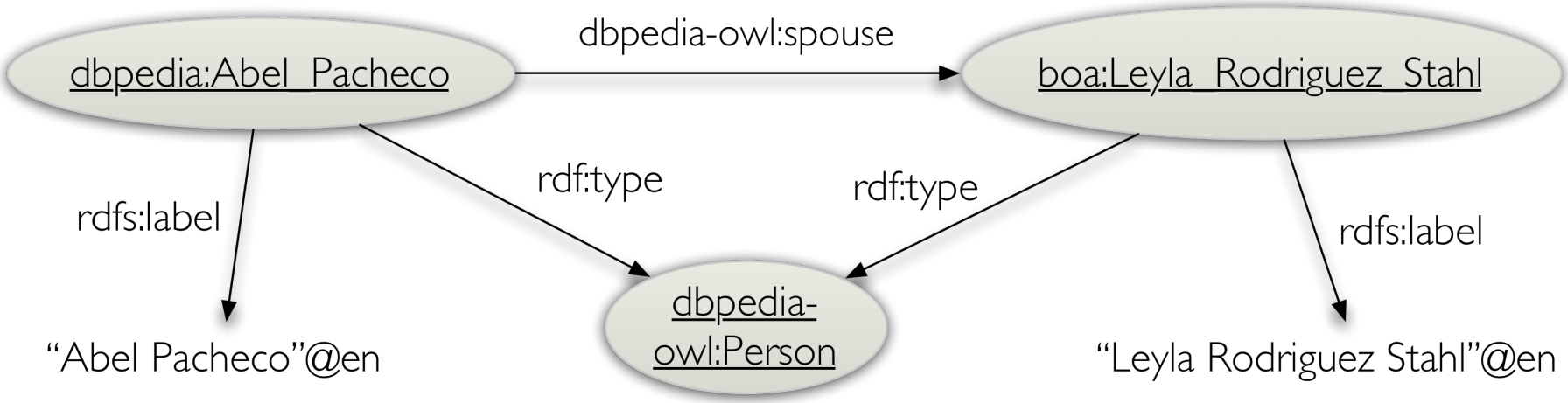
↓

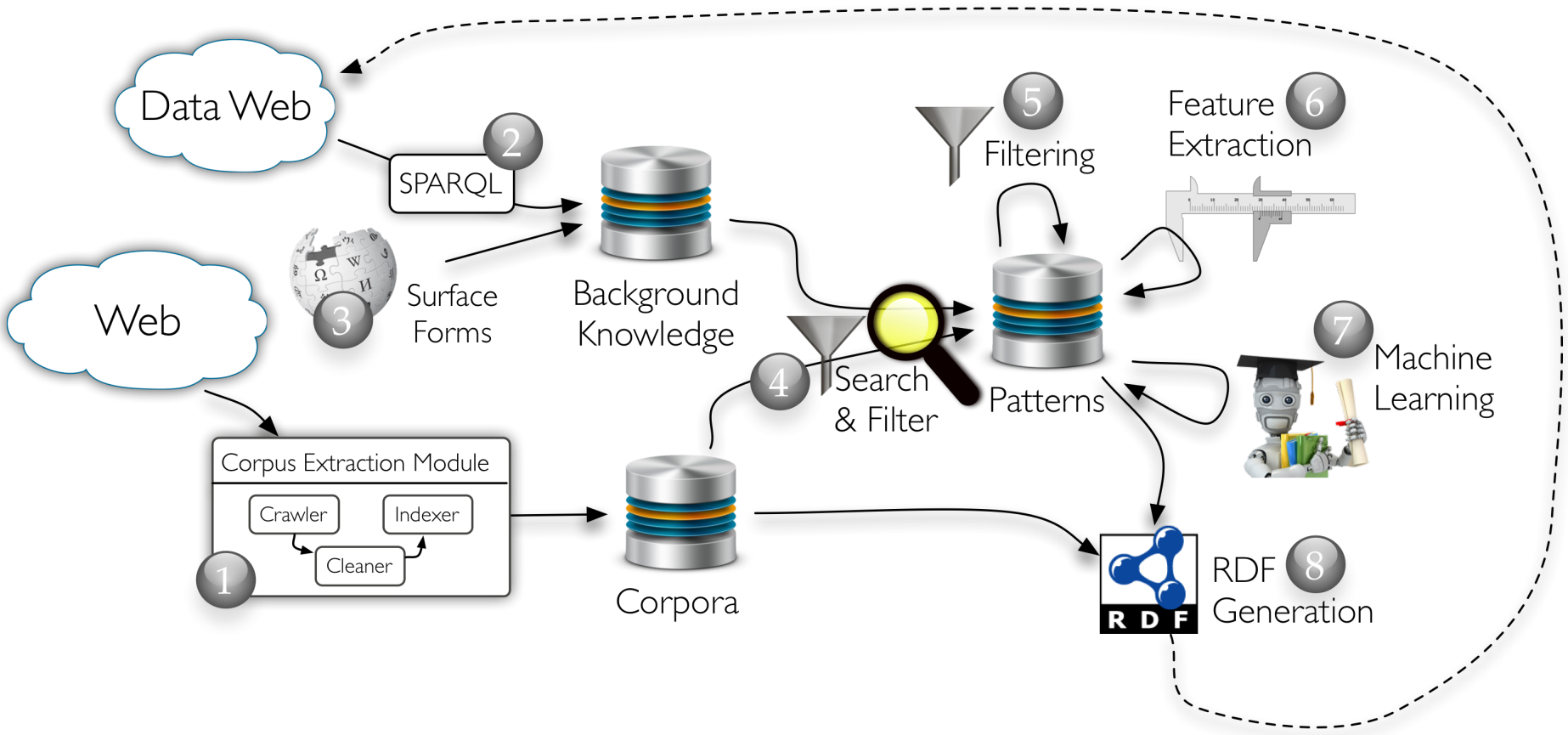
Pacheco arrived *with his wife* Leyla Rodriguez Stahl and several...

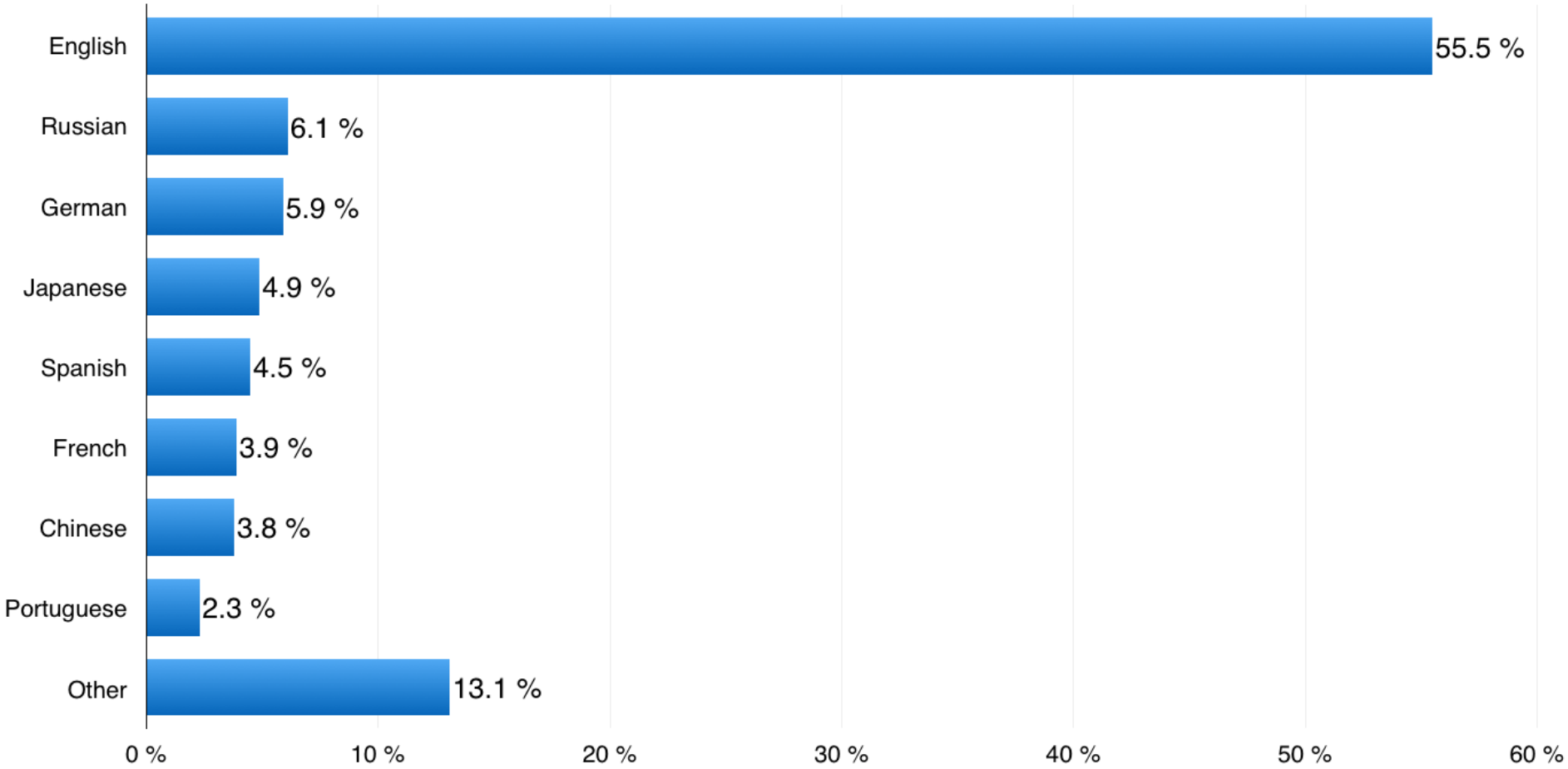
↓

Pacheco_PER arrived_O with_O his_O wife_O **Leyla_PER**
Rodriguez_PER Stahl_PER and_O several...

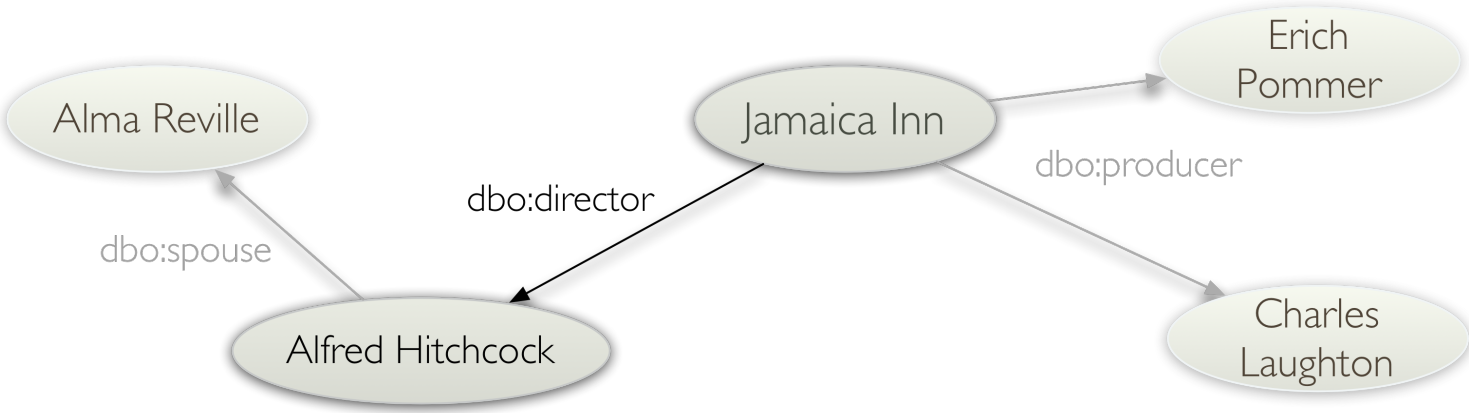
↓




Data Web
2
SPARQL
Web
3
Surface Forms
Background Knowledge
5
Filtering
Feature Extraction
6
4
Search & Filter
Patterns
7
Machine Learning
Corpus Extraction Module
Crawler
Indexer
Cleaner
1
Corpora
RDF
8
Generation

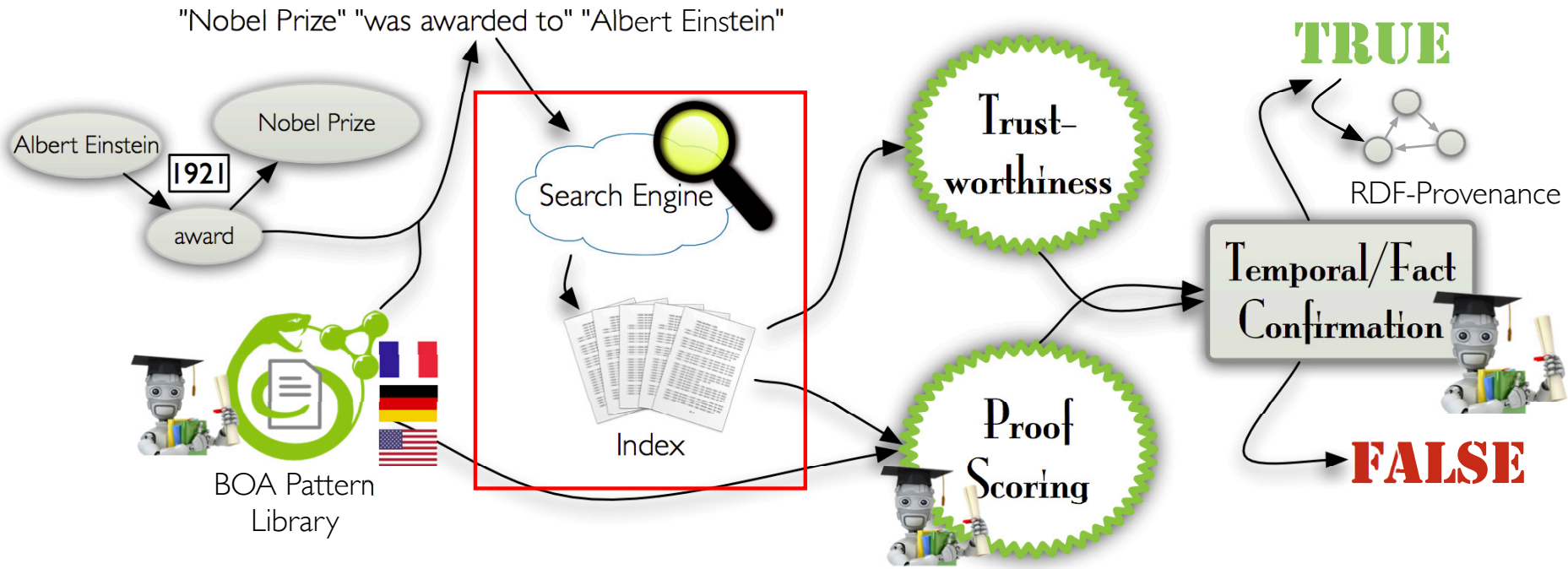

English
55.5 %
Russian
6.1 %
German
5.9 %
Japanese
4.9 %
Spanish
4.5 %
French
3.9 %
Chinese
3.8 %
Portuguese
2.3 %
Other
13.1 %
0 %
10 %
20 %
30 %
40 %
50 %
60 %

English	German	French
	publication	
R 's novel " **D**	**R** in seinem Roman " **D**	**D** est un roman **R**
R 's book " **D**	**R** in seinem Buch " **D**	**R** dans son roman **D**
R , author of " **D**	**R** in seinem Werk " **D**	**R** intitulé **D**
	marriage	
R married **D**	**D** seiner Frau **R**	**R** épouse **D**
R , his wife **D**	**D** seiner Ehefrau **R**	**R** , veuve **D**
D 's marriage to **R**	**R** und seiner Gattin **D**	**D**, la femme de **R**



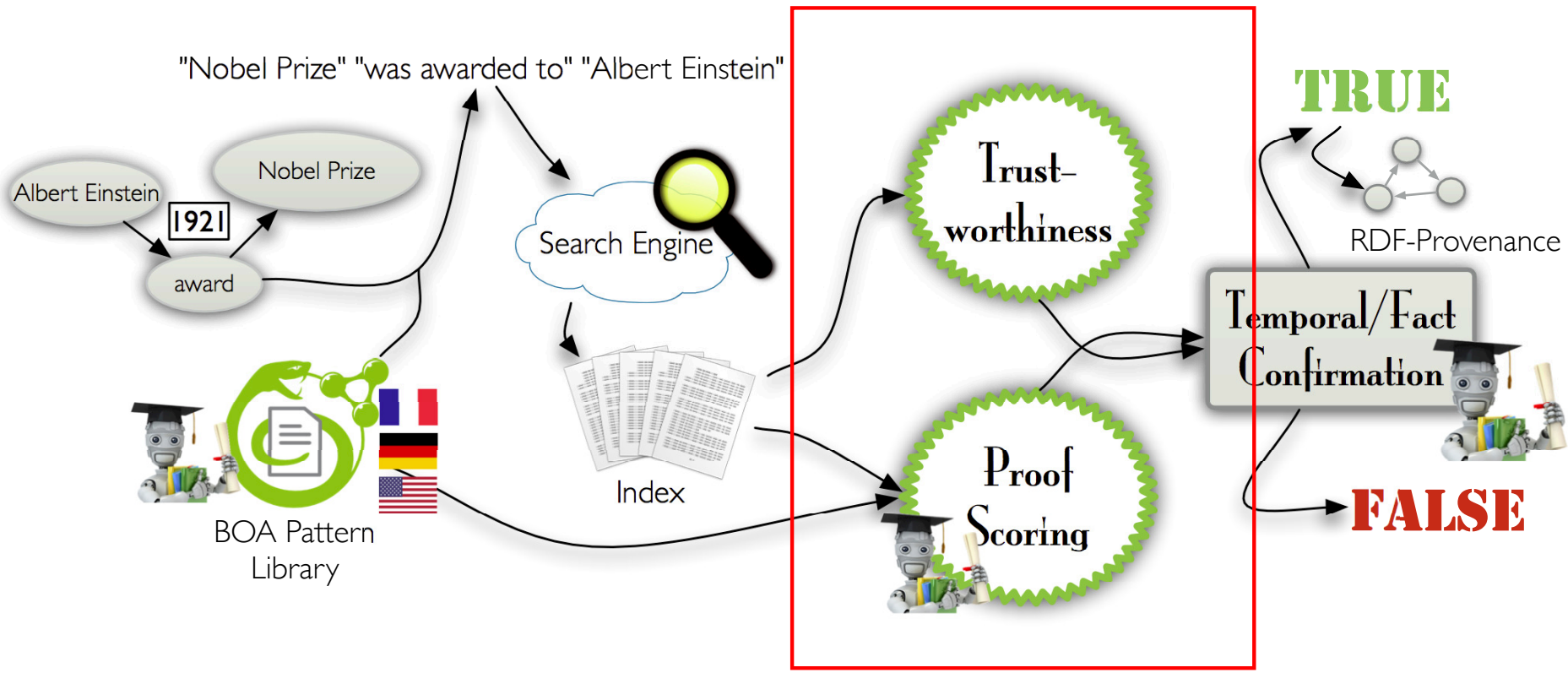
O'Hara's first major film was Alfred Hitchcock directed "Jamaica Inn" which was released in 1939, she had previously . . .

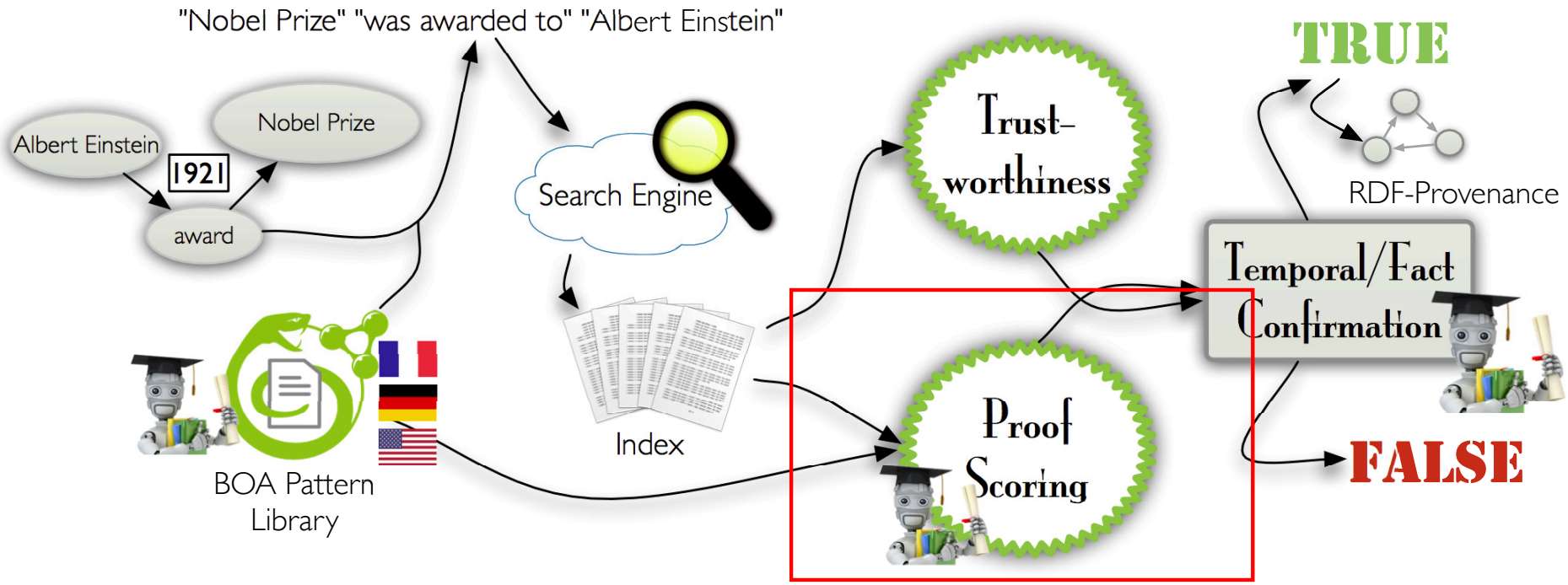
`examiner.com`


"Nobel Prize" "was awarded to" "Albert Einstein"
Albert Einstein
1921
award
Nobel Prize
BOA Pattern Library
Search Engine
Index
Trust-worthiness
Proof Scoring
Temporal/Fact Confirmation
RDF-Provenance
TRUE
FALSE

DeFacto

Trustworthiness and Proof Scoring




"Nobel Prize" "was awarded to" "Albert Einstein"
Albert Einstein
1921
award
Nobel Prize
Search Engine
Index
BOA Pattern Library
Trust-worthiness
Proof Scoring
Temporal/Fact Confirmation
TRUE
RDF-Provenance
FALSE

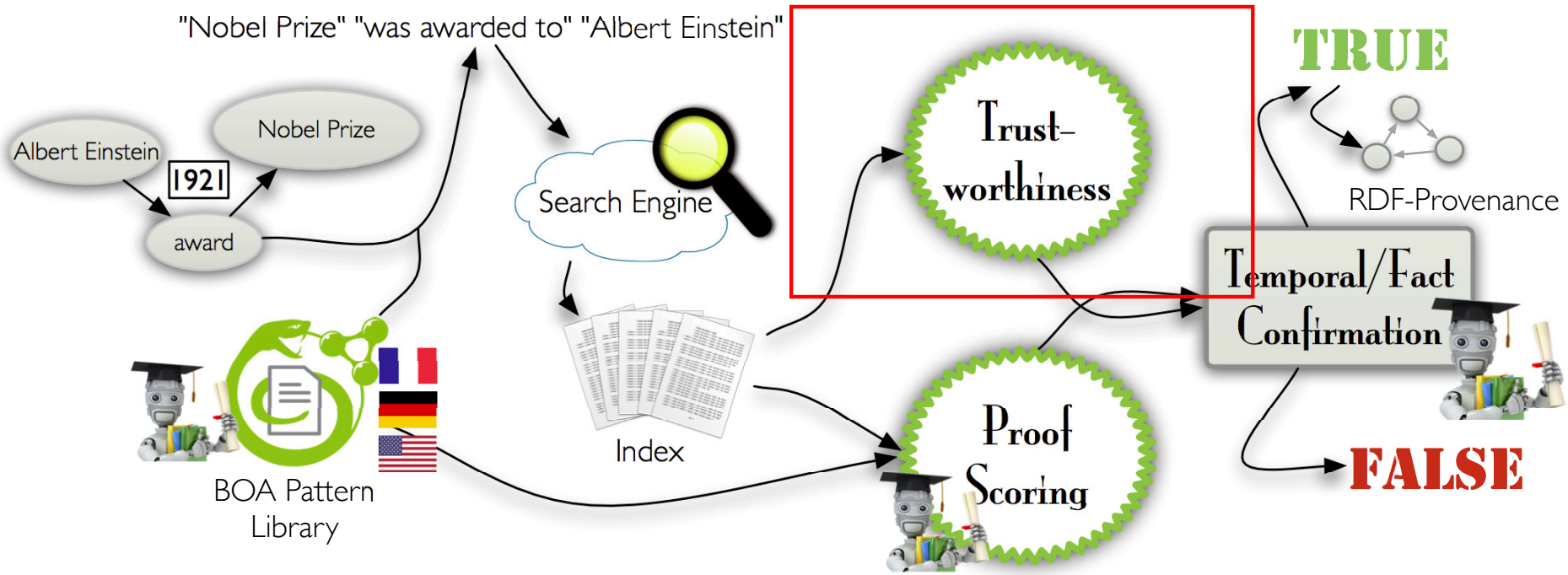
Idea

- Search for surface forms $\lambda(s)$ and $\lambda(o)$ for triple (s, p, o)
- if $dist(e_1, e_2) < \vartheta$ then extract proof features
- Example: Albert Einstein was awarded the Nobel Prize in 1921

Idea

- Search for surface forms $\lambda(s)$ and $\lambda(o)$ for triple (s, p, o)
- if $dist(e_1, e_2) < \vartheta$ then extract proof features
- Example: Albert Einstein was awarded the Nobel Prize in 1921

Feature	Value
BOA Pattern	0, 1
BOA Score	[0,1]
Entity_{dist}	$[0, \vartheta]$
Wordnet_{sim}	[0,1]
Frequency	[1,n]
Title_{sim}	[0,1]
Punctuation	0, 1
Text	Vector
Predicate	Word


"Nobel Prize" "was awarded to" "Albert Einstein"
Albert Einstein
Nobel Prize
1921
award
Search Engine
Index
BOA Pattern Library
Trust-worthiness
Proof Scoring
Temporal/Fact Confirmation
TRUE
RDF-Provenance
FALSE

Idea

- Approximate reliability of document (Nakamura et al., 2007)
- Rely on distributional features
- Example: Albert Einstein was awarded the Nobel Prize in 1921

Idea

- Approximate reliability of document (Nakamura et al., 2007)
- Rely on distributional features
- Example: Albert Einstein was awarded the Nobel Prize in 1921

Feature	Explanation
Topic majority in search results	Number of pages in search results with similar topics
Topic majority im WWW	Number of websites with similar topics
Pagerank	Google Page Rank of website

Idea

- Approximate reliability of document (Nakamura et al., 2007)
- Rely on distributional features
- Example: Albert Einstein was awarded the Nobel Prize in 1921

Feature	Explanation
Topic majority in search results	Number of pages in search results with similar topics
Topic majority im WWW	Number of websites with similar topics
Pagerank	Google Page Rank of website
$\lvert\{p \mid p \in P\}\rvert$	Number of proofs found in website
Number of search results	Total number of all queries (approx.)
rdfs:domain, *rdfs:range*	Binary validation of domain and range
Background knowledge	Co-occurrence classes and predicates

Facts are not always true ...

:Tom_Cruise :spouse :Katie_Holmes>

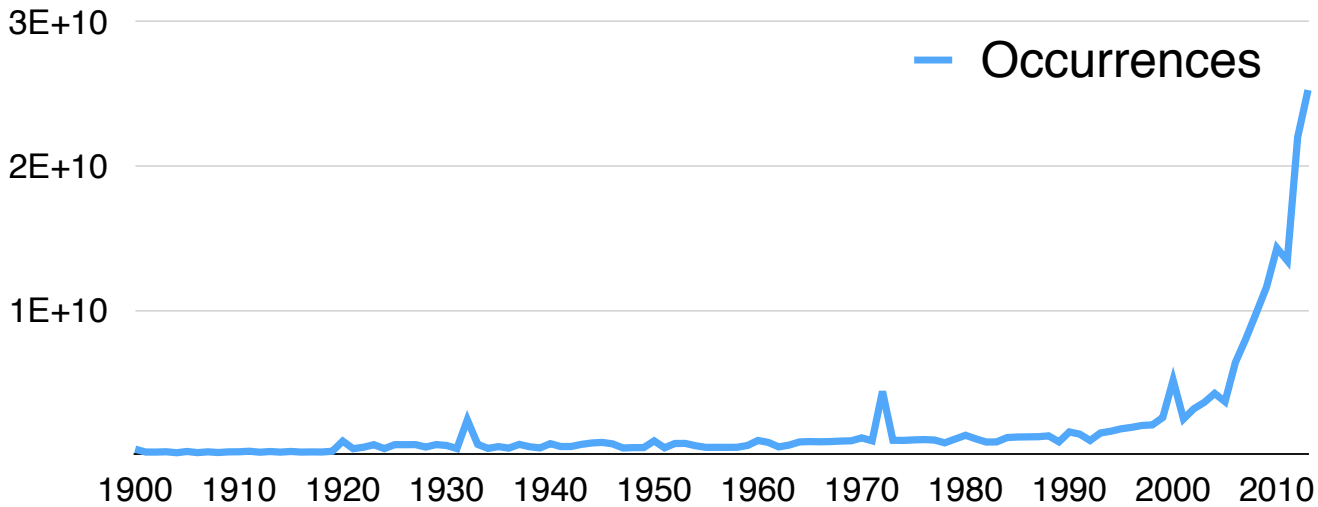
:Franz_Beckenbauer :plays_for_club :New_York_Cosmos

Facts are not always true ...

:Tom_Cruise :spouse :Katie_Holmes>
:Franz_Beckenbauer :plays_for_club :New_York_Cosmos

Extension to temporal checking

- Use distribution of years across the Web
- Extraction of years using patterns

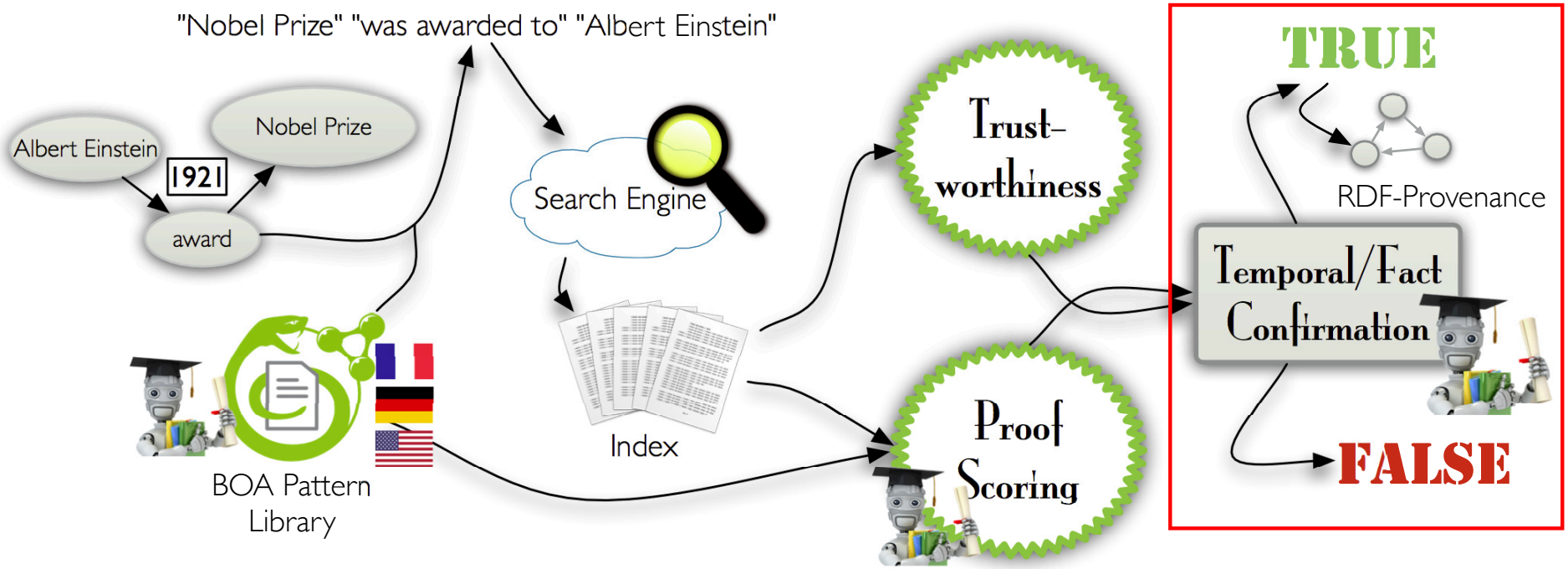


- Search year literals between within context window
- Learn patterns for valid interval
- Use year distribution to normalize score

```
[0-9]{4}\s*(/|-|--)\s*[0-9]{4}
[Ff]rom [0-9]{4} until [0-9]{4}
[bB]etween (the years) [0-9]{4} and [0-9]{4}
[0-9]{4} bis einschließlich [0-9]{4}
[zZ]wischen (den Jahren) [0-9]{4} und [0-9]{4}
[dD]urant la période [0-9]{4} - [0-9]{4}
[eE]ntre les années [0-9]{4} et [0-9]{4}
...
```

DeFacto

Fact Confirmation



- Dataset: FactBench
- Key Performace Indicators
  - Accuracy
  - Precision
  - Recall
  - F-measure
- Machine learning approaches
  - J48
  - Naïve Bayes
  - Support Vector Machine (with sequential optimization)
  - ...

Factbench

Overview

- Benchmark for fact checking
- 10 common relations
- 1,500 facts including validity period
- 750 in train and 750 in test set

Award
persons who received a nobel prize (timepoint, freebase)

Birth
birth place and date of a person (timepoint, dbpedia)

Death
death place and date of a person (timepoint, dbpedia)

Foundation Place
place and date of a company's foundation (timepoint, freebase)

Leader
presidents of countries (timespan, dbpedia)

NBA Team
team associations of NBA players (timespan, dbpedia)

Publication Date
author of a book and it's publication date (timepoint, freebase)

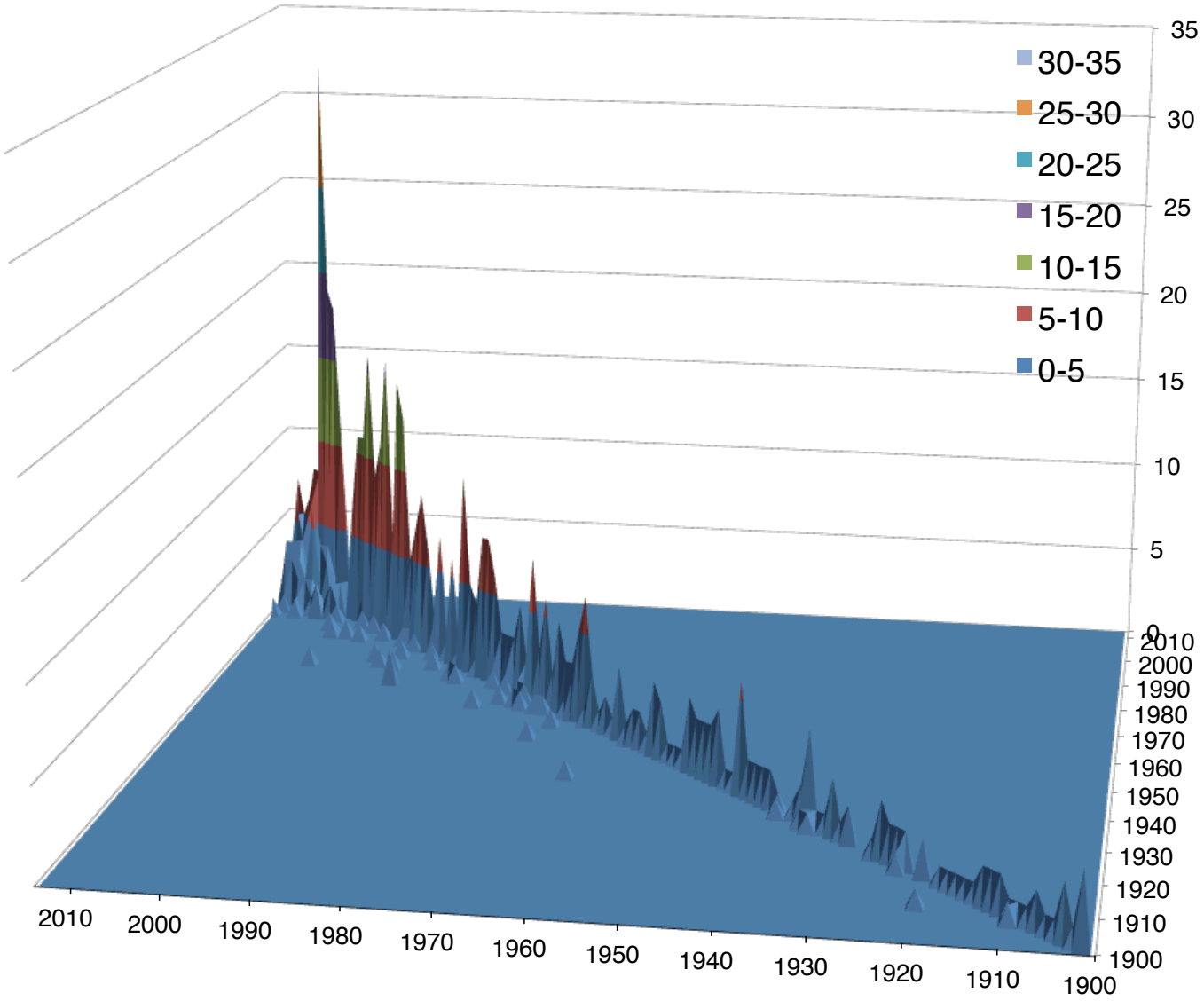
Spouse
marriage of two persons (timespan, freebase)

Starring
actors who starred in films (timepoint, dbpedia)

Subsidiary
companies and their subsidiaries (timepoint, freebase)

FactBench

Time Distribution



- Manually curated positive examples
- Negative examples generated automatically using random selected Triples $t = (s, p, o)$

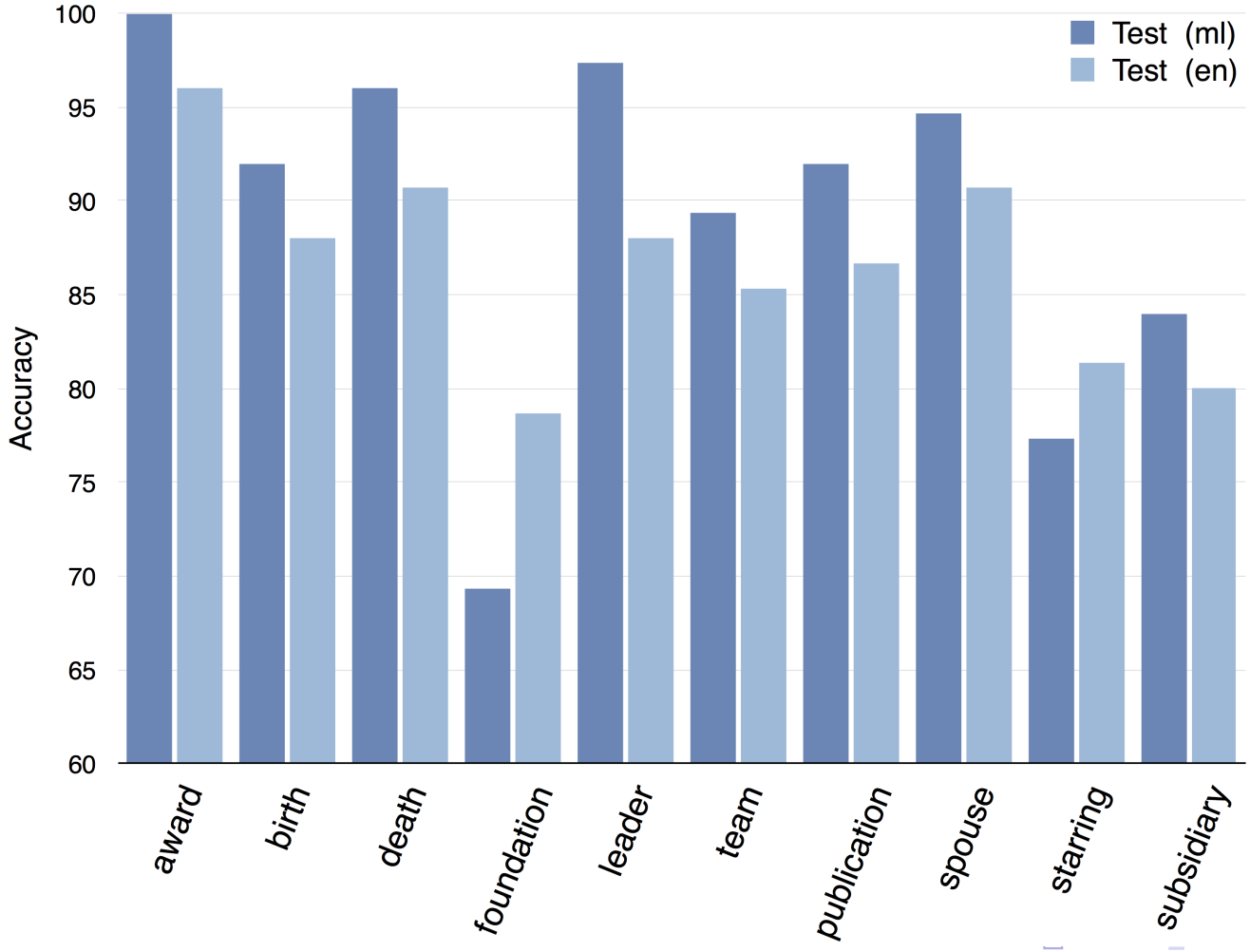
domain	Replace $s \rightarrow (s', p, o)$
range	Replace $o \rightarrow (s, p, o')$
domainrange	Replace s and $o \rightarrow (s', p, o')$
property	Replace $p \rightarrow (s, p', o)$
random	Replace s, p and $o \rightarrow (s', p', o')$
mix	Random sample of 20% of each of the above

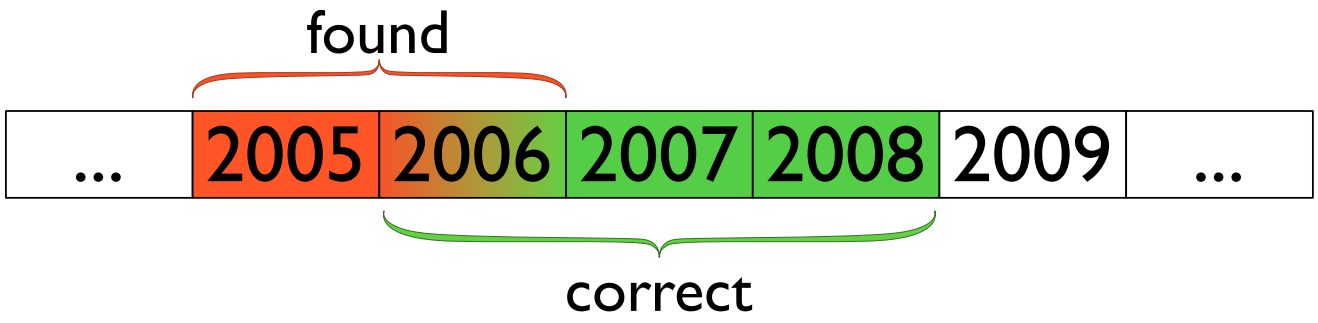
	Domain			Range		
	P	R	F_1	P	R	F_1
J48	0.898	0.897	0.897	0.909	0.909	0.909
SimpleLogistic	0.890	0.890	0.890	0.880	0.880	0.880
NaiveBayes	0.837	0.812	0.808	0.852	0.833	0.830
SMO	0.861	0.854	0.853	0.852	0.833	0.830

	DomainRange			Property		
	P	R	F_1	P	R	F_1
J48	0.910	0.910	0.910	0.786	0.708	0.687
SimpleLogistic	0.889	0.889	0.889	0.653	0.649	0.646
NaiveBayes	0.861	0.845	0.843	0.620	0.613	0.608
SMO	0.853	0.836	0.834	0.673	0.646	0.632

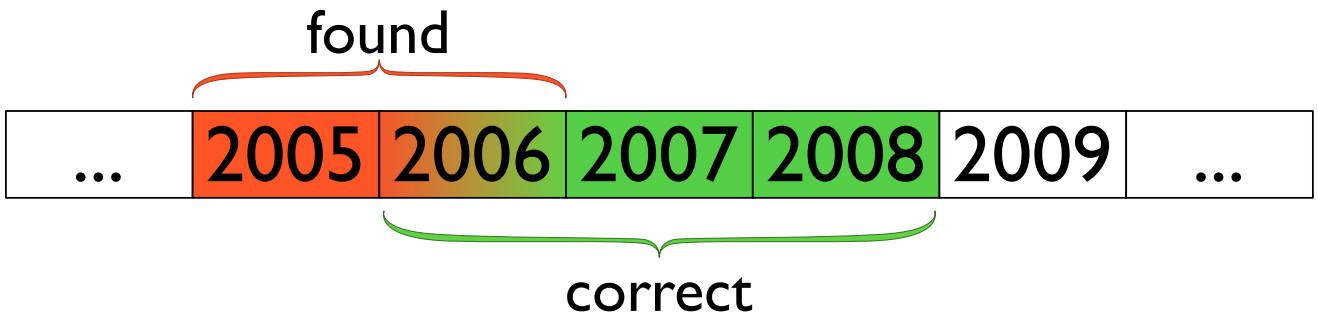
	Random			Mix		
	P	R	F_1	P	R	F_1
J48	0.910	0.909	0.909	0.850	0.849	0.849
SimpleLogistic	0.879	0.878	0.878	0.810	0.802	0.799
NaiveBayes	0.851	0.841	0.839	0.789	0.787	0.787
SMO	0.864	0.843	0.841	0.817	0.769	0.756

1. Using multiple languages better in most cases
2. Close to 100% accuracy (award, multilingual)





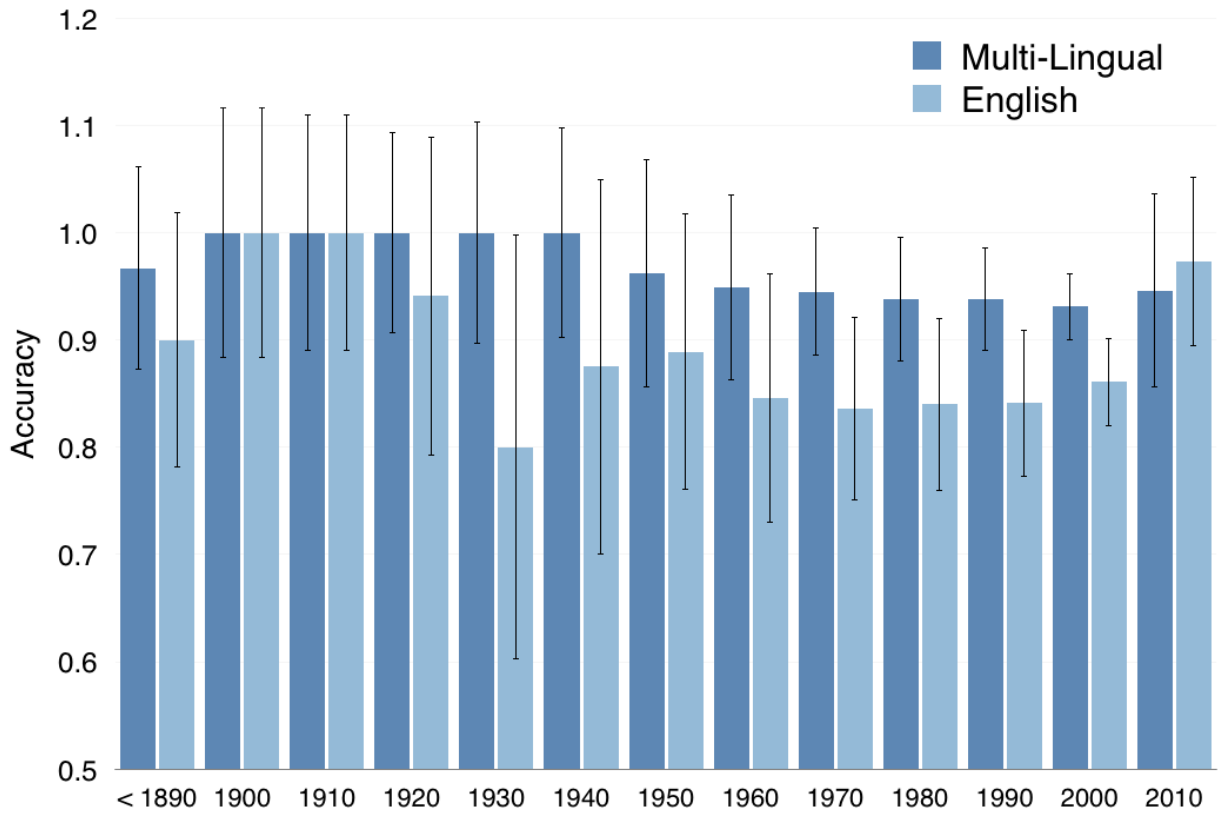
$$P = \frac{|\{correct\} \cap \{found\}|}{|\{found\}|} \quad R = \frac{|\{correct\} \cap \{found\}|}{|\{correct\}|}$$



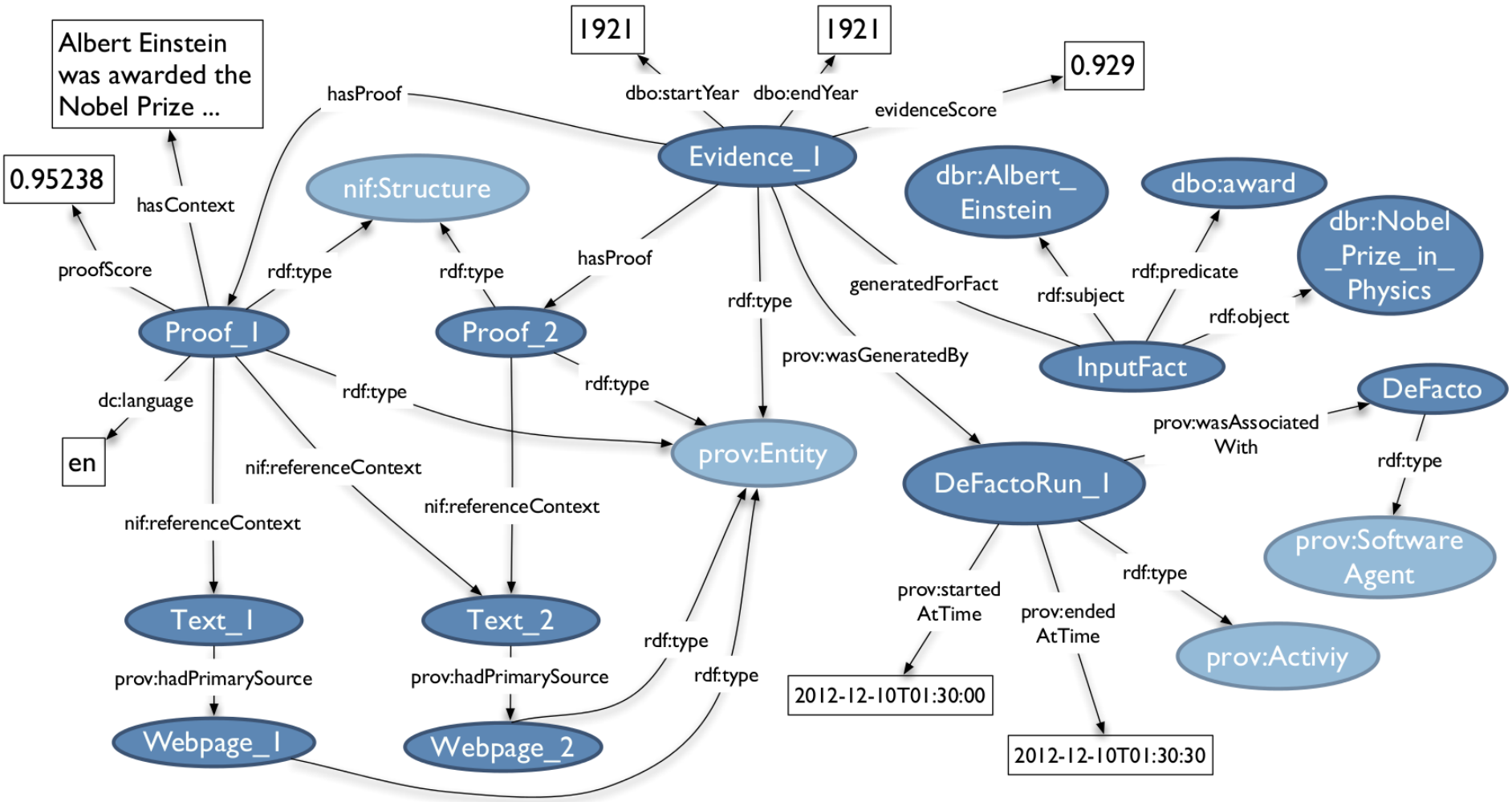
$$P = \frac{|\{correct\} \cap \{found\}|}{|\{found\}|} \quad R = \frac{|\{correct\} \cap \{found\}|}{|\{correct\}|}$$

- Precision $= 1/2$
- Recall $= 1/3$
- F-measure $= 2/5$

1. Using multiple languages significantly better
2. Increase from 86.53% to 89.2% on average
3. +6.5% for time points and +6.9% time spans



Output


Albert Einstein was awarded the Nobel Prize ...
1921
1921
0.929
hasProof
dbo:startYear
dbo:endYear
evidenceScore
0.95238
hasContext
nif:Structure
Evidence_1
dbr:Albert_Einstein
dbo:award
dbr:Nobel_Prize_in_Physics
proofScore
rdf:type
rdf:type
hasProof
rdf:type
generatedForFact
rdf:subject
rdf:predicate
rdf:object
Proof_1
Proof_2
InputFact
prov:wasGeneratedBy
DeFacto
dc:language
rdf:type
rdf:type
prov:wasAssociated With
en
prov:Entity
nif:referenceContext
DeFactoRun_1
rdf:type
nif:referenceContext
nif:referenceContext
prov:Software Agent
prov:started AtTime
prov:ended AtTime
rdf:type
Text_1
Text_2
rdf:type
rdf:type
prov:Activiy
prov:hadPrimarySource
prov:hadPrimarySource
2012-12-10T01:30:00
2012-12-10T01:30:30
Webpage_1
Webpage_2

Conclusion and Future Work

- Summary
  - Presented DeFacto, a framework for checking RDF facts
  - Performs best with J48 classifier
  - Achieves between 80% and 100% accuracy on FactBench relations

Conclusion and Future Work

- Summary
  - Presented DeFacto, a framework for checking RDF facts
  - Performs best with J48 classifier
  - Achieves between 80% and 100% accuracy on FactBench relations
- Future Work
  - Extend FactBench
  - Combination with Deep Learning
  - Integration of more languages
  - Reference corpus for faster search

That's all Folks!

thank you!

HOBBIT

Holistic Benchmarking of Big Linked Data

Axel Ngonga
AKSW Research Group
Augustusplatz 10, Room P905
04109 Leipzig, Germany
ngonga@informatik.uni-leipzig.de
@NgongaAxel