YAGO:
Yet Another Great Ontology

Fabian M. Suchanek

(joint work with Gjergji Kasneci, Mauro Sozio and Gerhard Weikum)

(Max-Planck-Institute for Informatics, Saarbrücken/Germany)

Overview

- Motivation: Why would anybody need Ontologies?
- Building a Core Ontology: YAGO
- Extending the Core Ontology: SOFIE

The Search for Excellent Scientists

Max-Planck Institute

DFKI

The Search for Excellent Scientists

Google

scientist musician

Invisible Gorilla steals the Nobel **Prize**

...The gorilla, plus dropped food and country **music**, were honored...

new**scientist**.org/article/invisibleGorilla.htm Cached Similar pages

The Search for Excellent Scientists

Google

scientist who are musicians and won a prize

Invisible Gorilla steals the Nobel **Prize**

...The gorilla, plus dropped food and country **music**, were honored...

new**scientist**.org/article/invisibleGorilla.htm Cached Similar pages

The Search for Excellent Scientists

Google

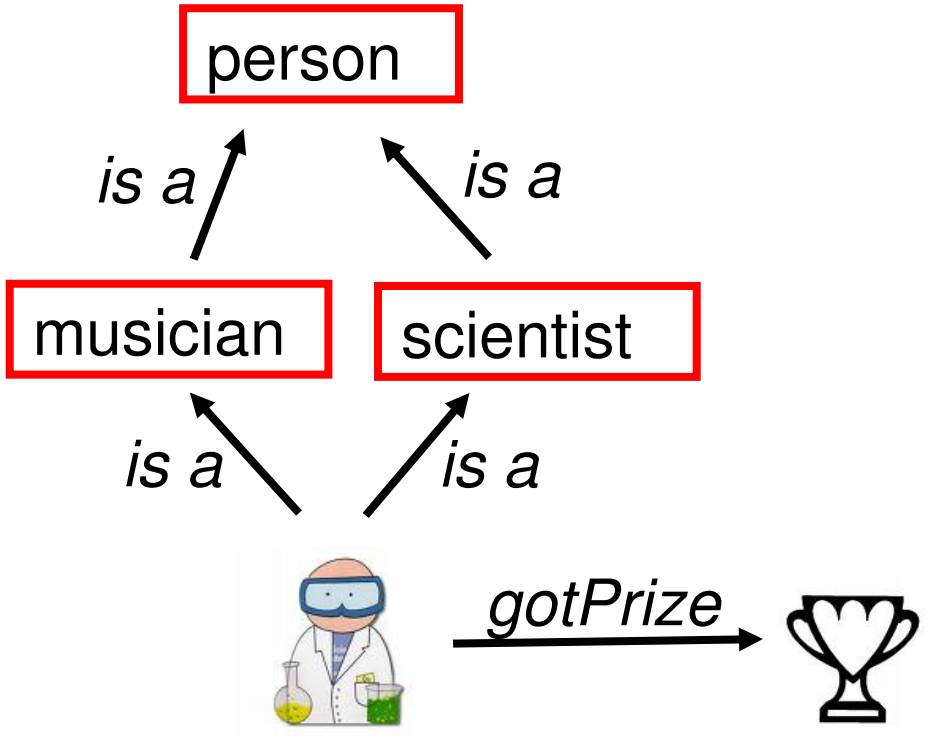
Please give me IMMEDIATELY the scientists who are...

Invisible Gorilla steals the Nobel **Prize**

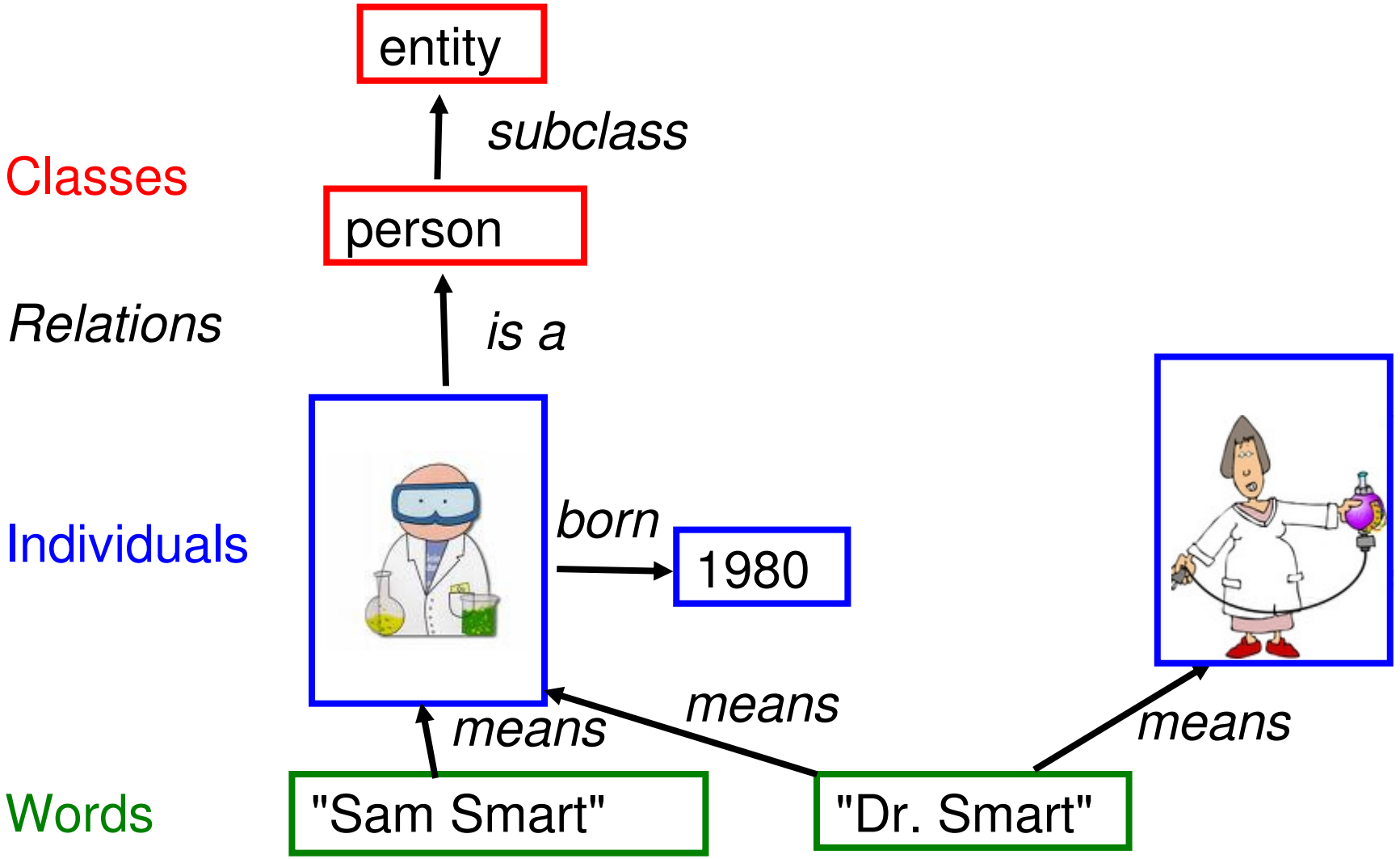
...The gorilla, plus dropped food and country **music**, were honored...

new**scientist**.org/article/invisibleGorilla.htm Cached Similar pages

Solution: An Ontology



Solution: An Ontology



Where do we get the ontology from?

recoverWithout(most_people, medication)

areUnder(0%, the_age_of_18)

support(these_findings, the_notion)

Previous Approaches:

- Assemble the ontology manua
  (WordNet, SUMO, Cyc, GeneOntology
  Problem: Usually low coverage (MPI of these)
- Use community work (Semantic W , Freebase)
  Problem: We don't know yet wh takes off
- Extract the ontology from cor (e.g. the Web)
  (Text2Onto, KnowItAll, Esp so, Snowball, LEILA, TextRunner)
  Problems:
  1. Usually low accuracy (50%-92%)
  2. Non-canonicity

Overview

- Motivation: Why would anybody need Ontologies?
- Building a Core Ontology: YAGO
- Extending the Core Ontology: SOFIE

YAGO Construction: Infoboxes

bornIn → Berlin

Exploit infoboxes

YAGO Construction: Categories

Smart, S

blah blah blub Elvis (don't read this! Better listen to the talk!) laber fasel suelz. Insbesondere, blub, texte zu, und so weiter blah blah blub Elvis laber fasel suelz. Blub, aber blah! Insbesondere, blub, texte zu, und so weiter blah blah blub Elvis laber fasel suelz. Insbesondere, blub, texte zu, und so weiter

Categories:

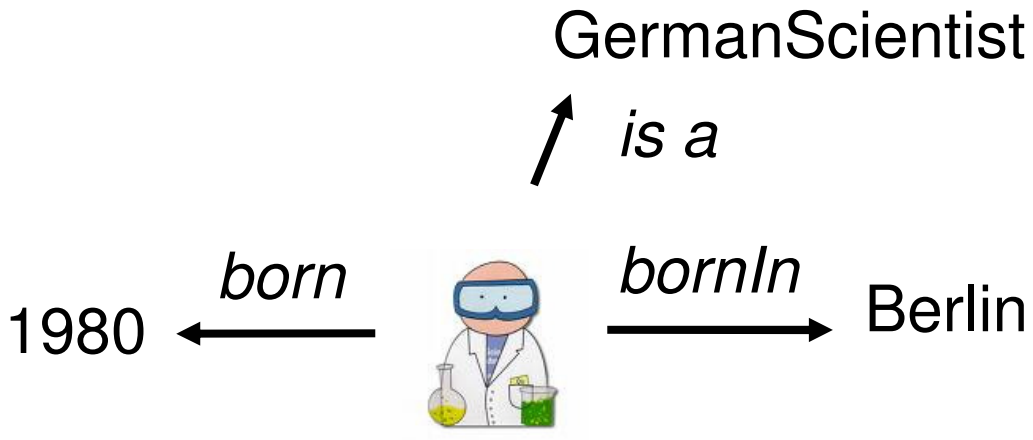
1980_births

1980 ← *born* —  — *bornIn* → Berlin

Exploit infoboxes

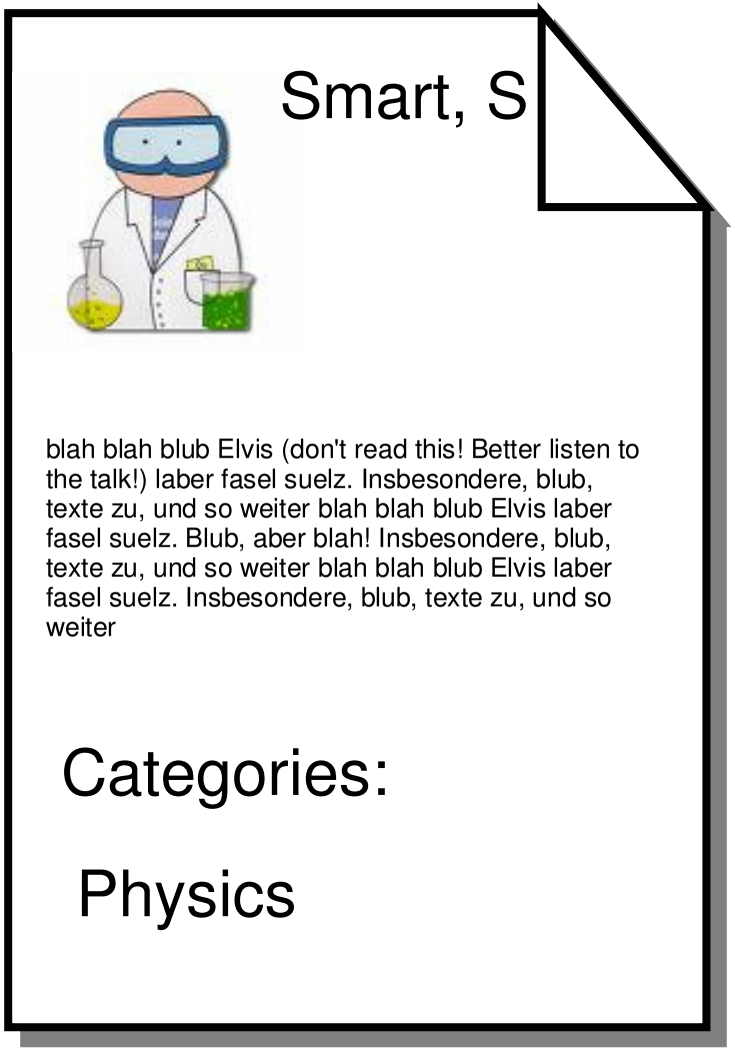
Exploit relational categories

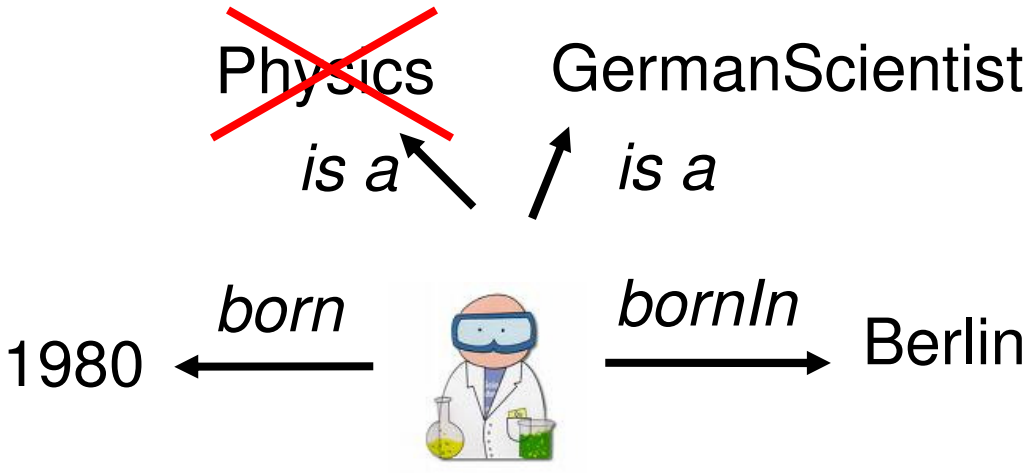
YAGO Construction: Categories



Exploit infoboxes
Exploit relational categories
Exploit conceptual categories

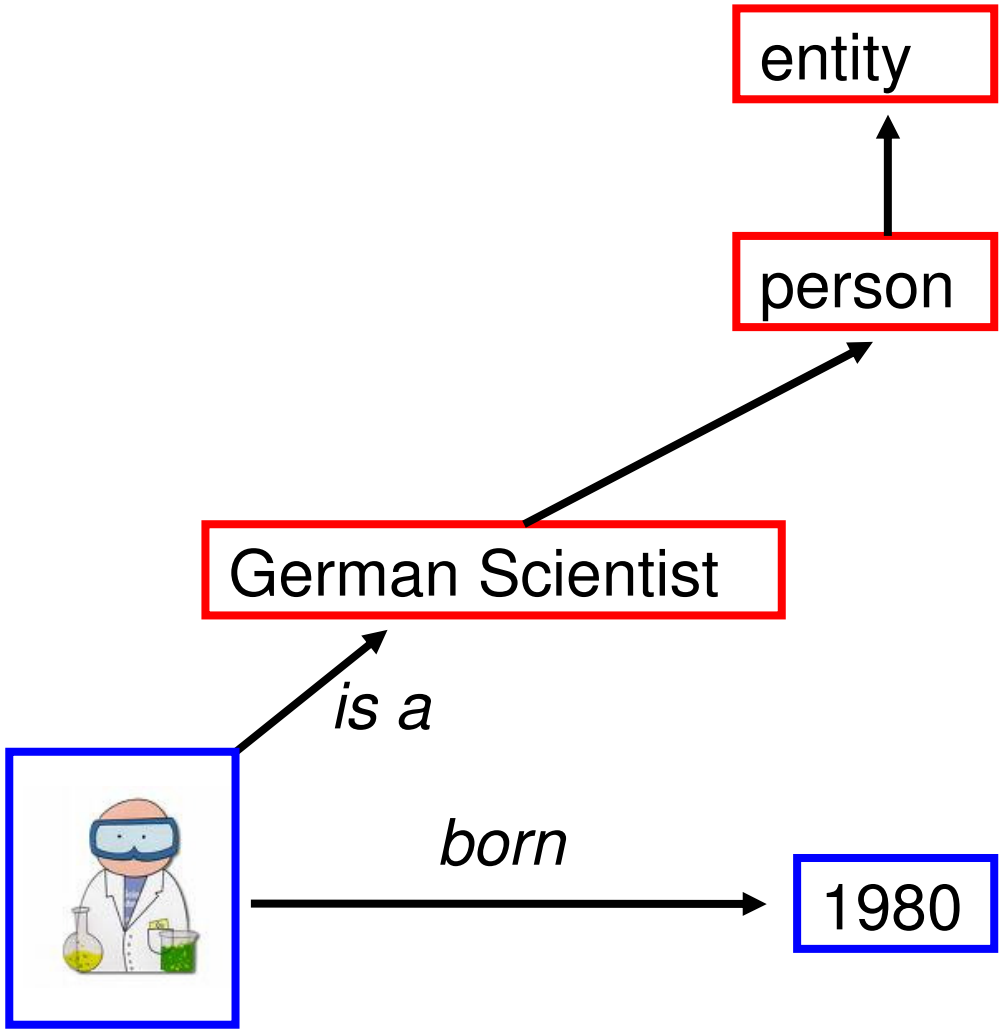
YAGO Construction: Categories





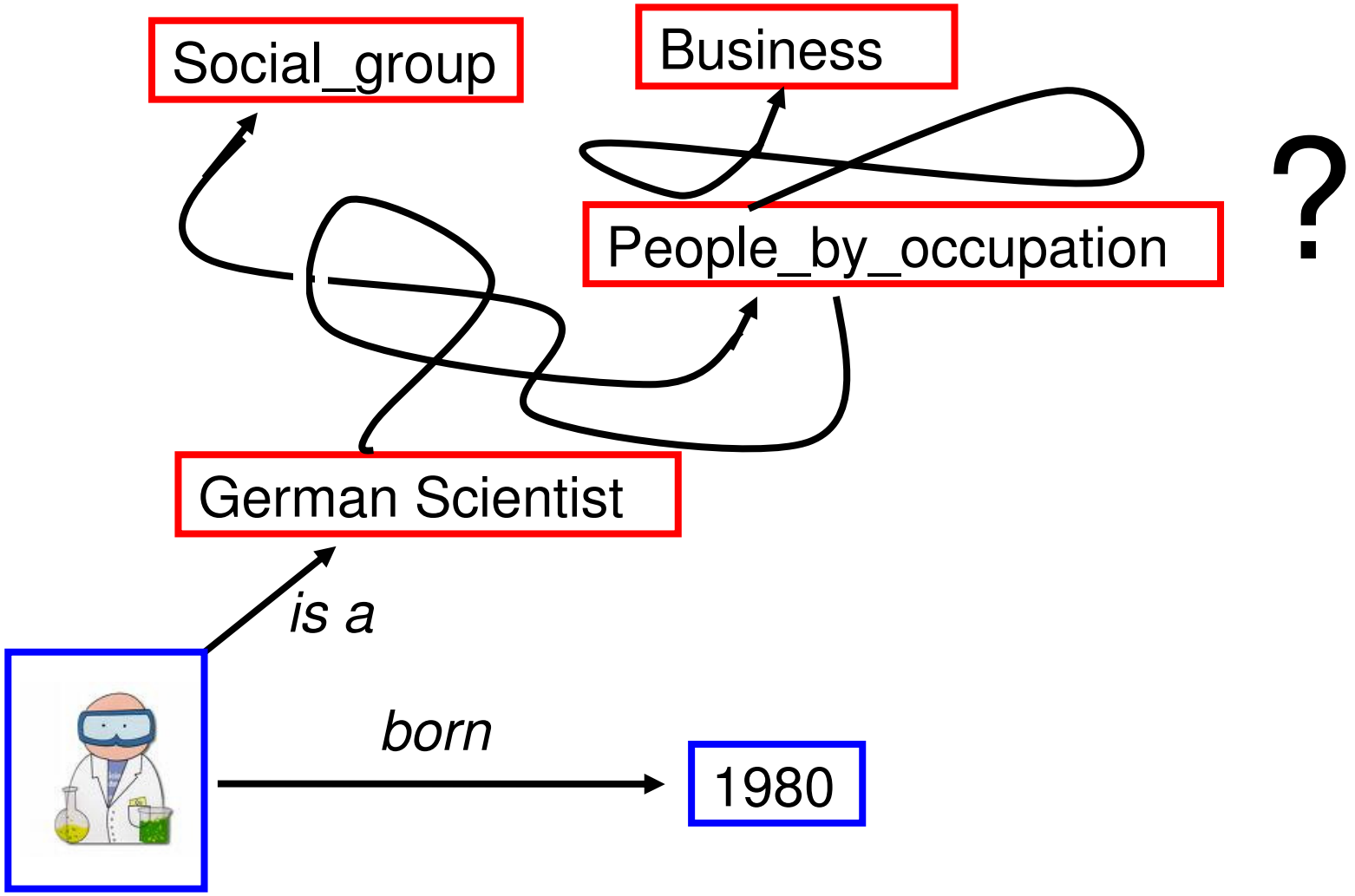
Exploit infoboxes
Exploit relational categories
Exploit conceptual categories
Avoid thematic categories

YAGO Construction: Upper Model

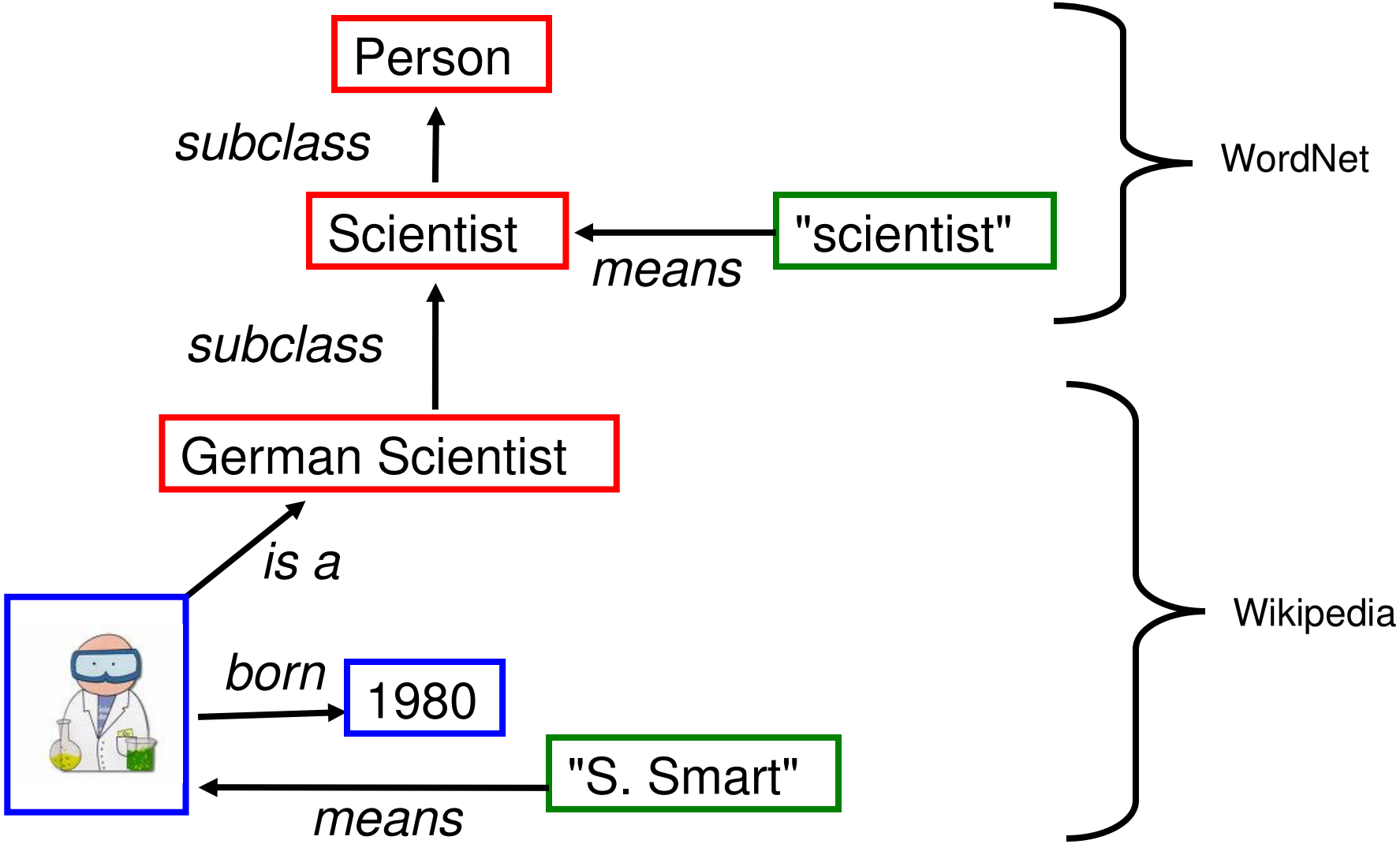


?

YAGO Construction: Upper Model



YAGO Construction: Upper Model



YAGO: Relations

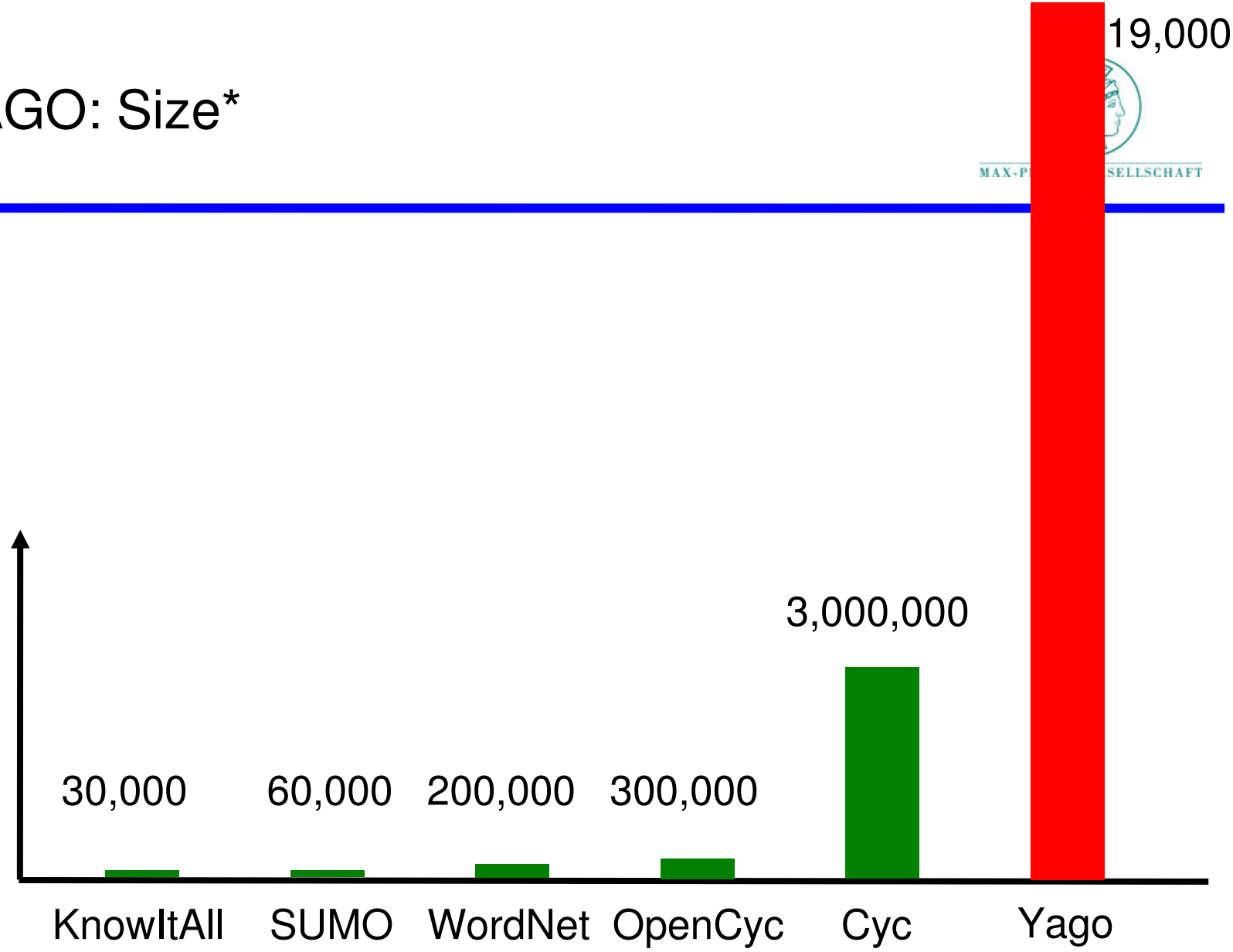
is a
familyNam
givenNam
bornOnDate
diedOnDate
bornIn
diedIn
locatedIn

Manual evaluation:
95% correct

establishedOnDate
isMarriedTo
hasPopulation
asHeight
asWeight
hasInflation
actedIn
...

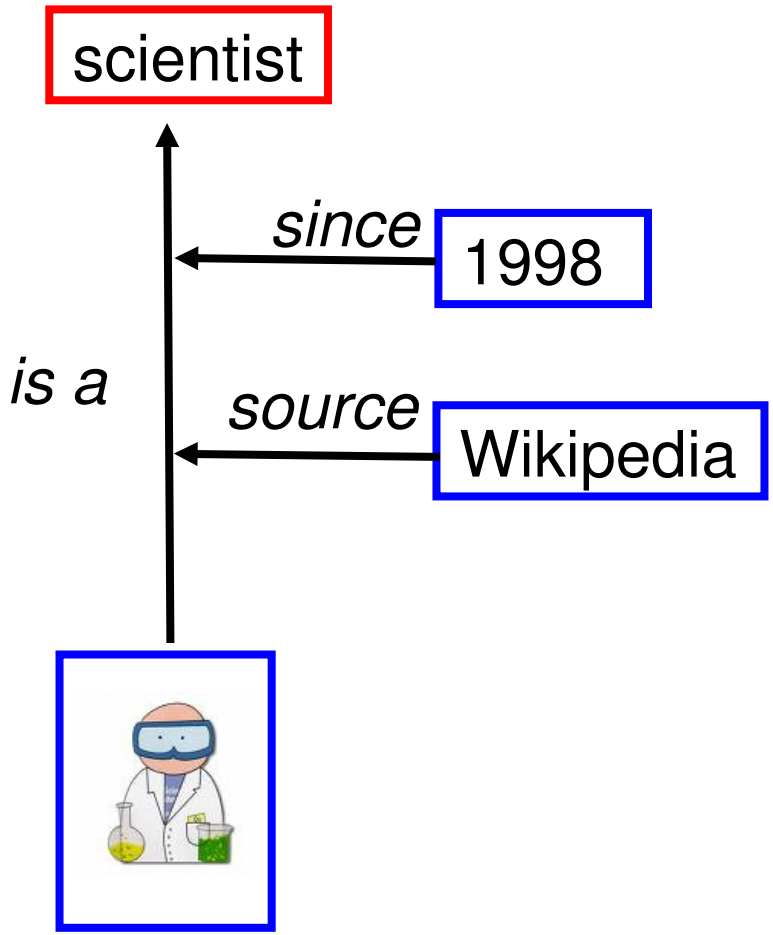
90 relations

YAGO: Size*



* Publicly available ontologies with a quality guarantee. Size is not correlated with usefulness.

YAGO Model: Why binary is not enough




```
#1 (Sam, is_a, scientist)
#2 (#1, since, 1998)
#3 (#1, source, Wikipedia)
```

YAGO Model: Formal view

A YAGO ontology over

- a set of relations $\mathtt{R}$
- a set of common entities $\mathtt{C}$
- a set of fact identifiers $\mathtt{I}$

is a function

$$\mathtt{I} \rightarrow (\mathtt{R} \cup \mathtt{C} \cup \mathtt{I}) \times \mathtt{R} \times (\mathtt{R} \cup \mathtt{I} \cup \mathtt{C})$$

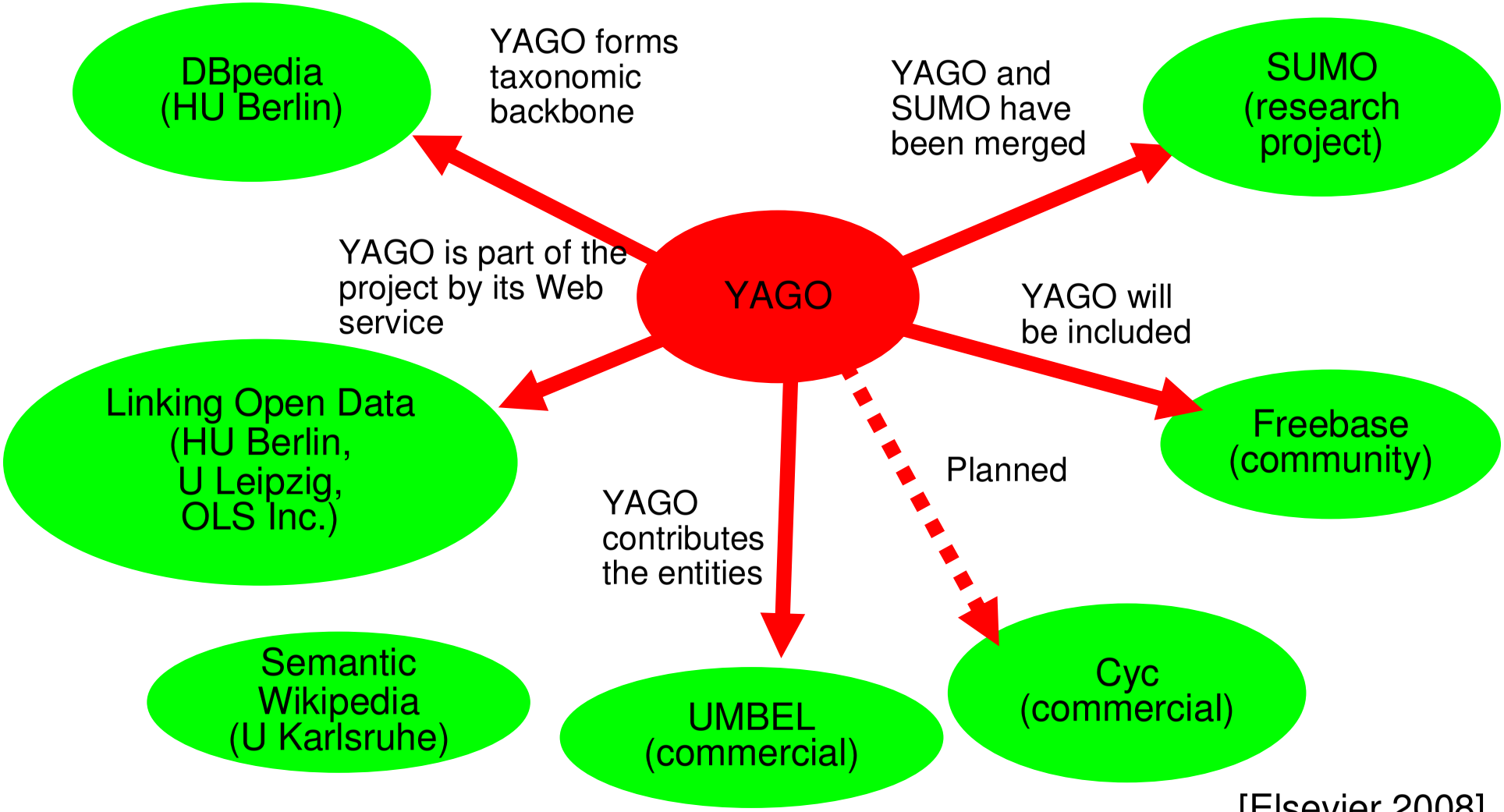
```
#1 (Sam, is_a, scientist)
#2 (#1, since, 1998)
#3 (#1, source, Wikipedia)
```

We can talk about

- facts **(#1, source, Wikipedia)**
- additional arguments **(#1, since, 1998)**
- relations **(time, hasRange, time_interval)**

Still: Decideable Consistency

A Hitchhiker's Guide to Ontology



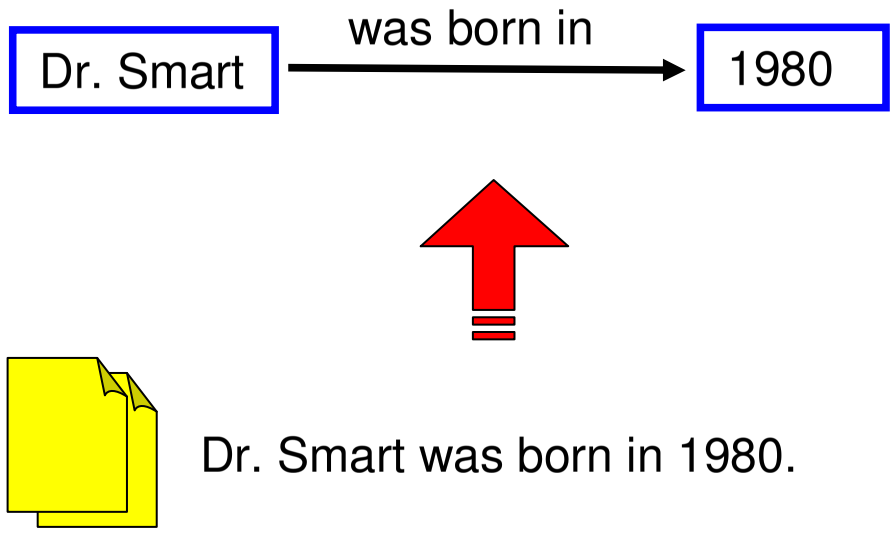
[Elsevier 2008]

Extending the Ontology

Our first approach:

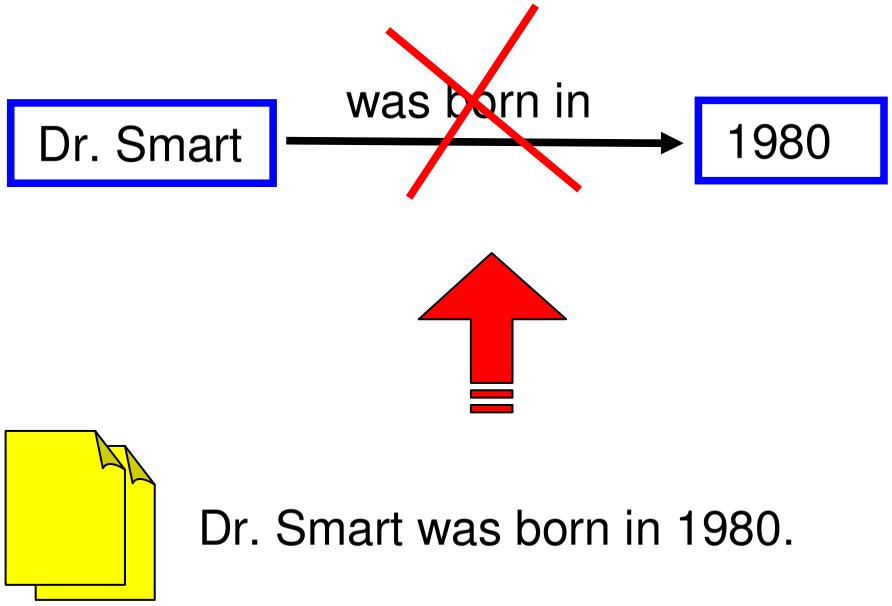
LEILA - Combining Linguistic and Statistical Analysis [SIGKDD 2006]

Worked well, but was slow.



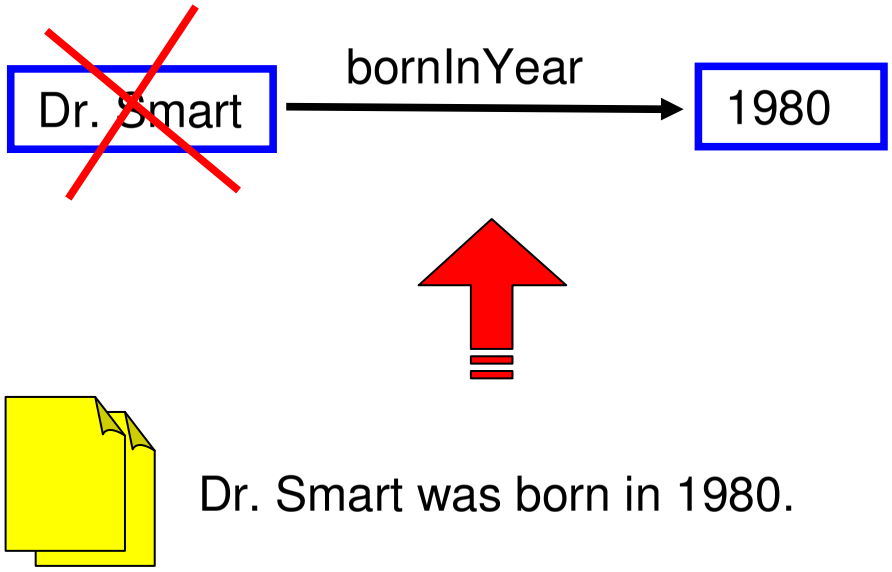
Extending the Ontology

bornInYear(Person, Year)



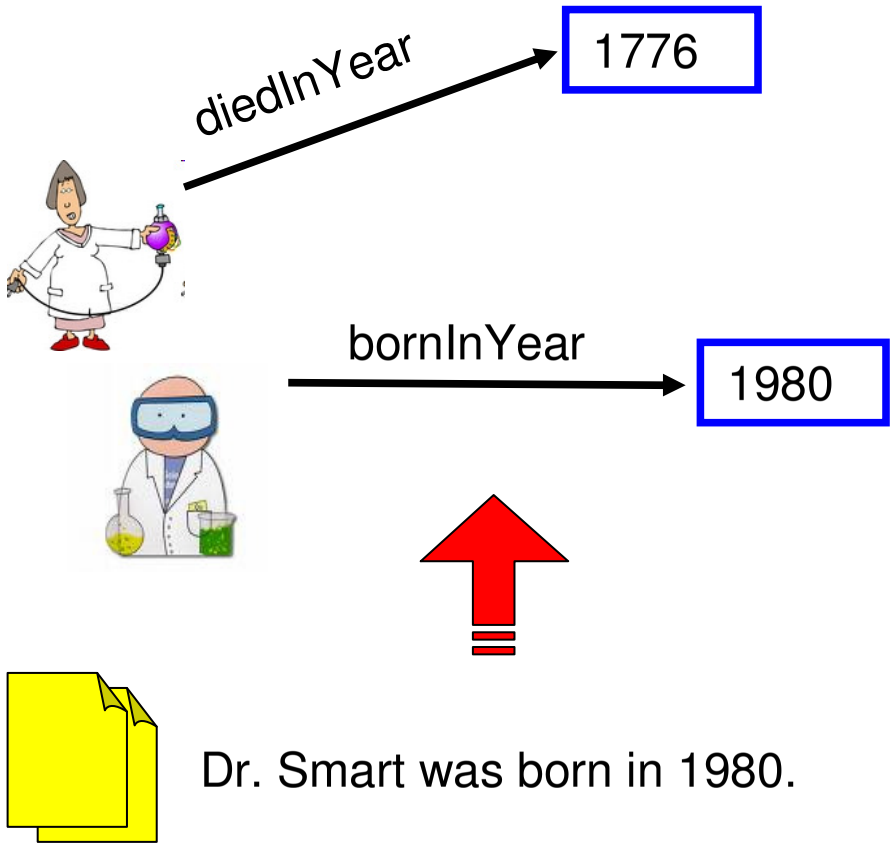
Extending the Ontology

1. Mapping patterns to relations



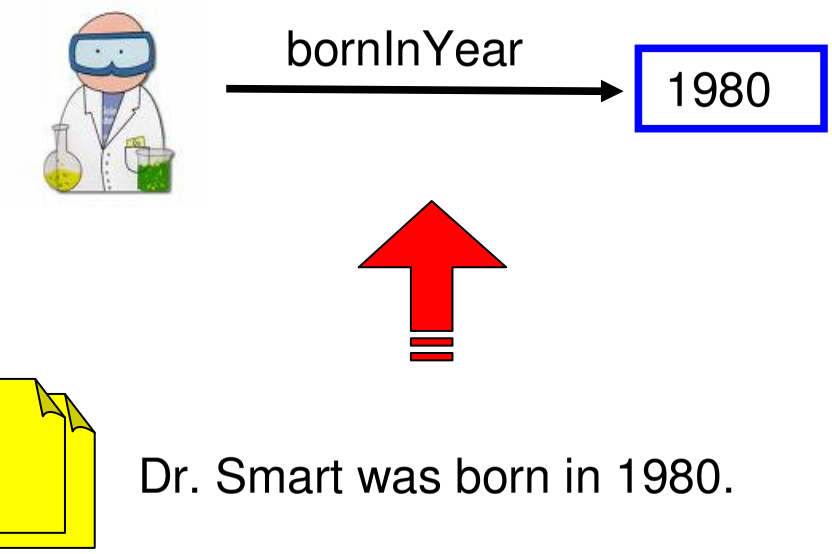
Extending the Ontology

1. Mapping patterns to relations
2. Disambiguating entity names



Extending the Ontology

1. Mapping patterns to relations
2. Disambiguating entity names
3. Performing logical reasoning



SOFIE: A Unifying Framework

New !

MAX-PLANCK-GESELLSCHAFT

1. Mapping patterns to relations
2. Disambiguating entity names
3. Performing logical reasoning

bornInYear → 1937

+

„Elvis was born in 1937.“

=

„X was born in Y“
is a good pattern for
bornInYear

SOFIE: A Unifying Framework

1. Mapping patterns to relations
2. Disambiguating entity names
3. Performing logical reasoning

„X was born in Y“
is a good pattern for
bornInYear

+

„Dr. Smart was born in 1980.“

=

bornInYear → 1980

SOFIE: A Unifying Framework

New !

MAX-PLANCK-GESELLSCHAFT

1. Mapping patterns to relations
2. Disambiguating entity names
3. Performing logical reasoning

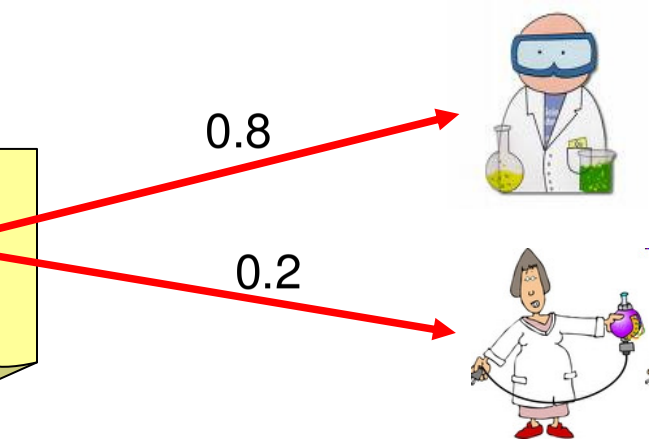
r(x,y) /\ occurs(p,x,y) => isGoodPattern(p,r)

isGoodPattern(p,r) /\ occurs(p,x',y') => r(x',y')

disambiguate(„Dr. Smart“, Sam_Smart)[0.8]

disambiguate(„Dr. Smart“, Lisa_Smart)[0.2]

„...The world as such, I would like to say – even though some will contradict – is not as it seems. As Dr. Smart pointed out in his ground-breaking paper „The world according to Smart“, the world rather seems not what it seems...“



SOFIE: A Unifying Framework

1. Mapping patterns to relations
2. Disambiguating entity names
3. Performing logical reasoning

r(x,y) /\ occurs(p,x,y) => isGoodPattern(p,r)

isGoodPattern(p,r) /\ occurs(p,x',y') => r(x',y')

disambiguate(„Dr. Smart“, Sam_Smart)[0.8]

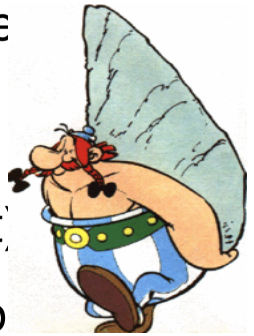
bornInYear(x,b) /\ diedInYear(x,d) => b<d

It's all just logical formulae with weights

Find truth values for the literals so that a maximal number of formulae is happy!

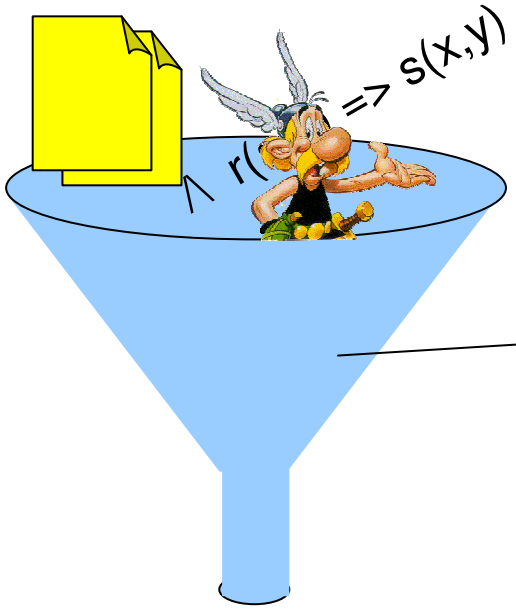
[PhD]

SOFIE: A Unifying Framewo



SOFIE: A Unifying Framework

Weighted MAX SAT Problem



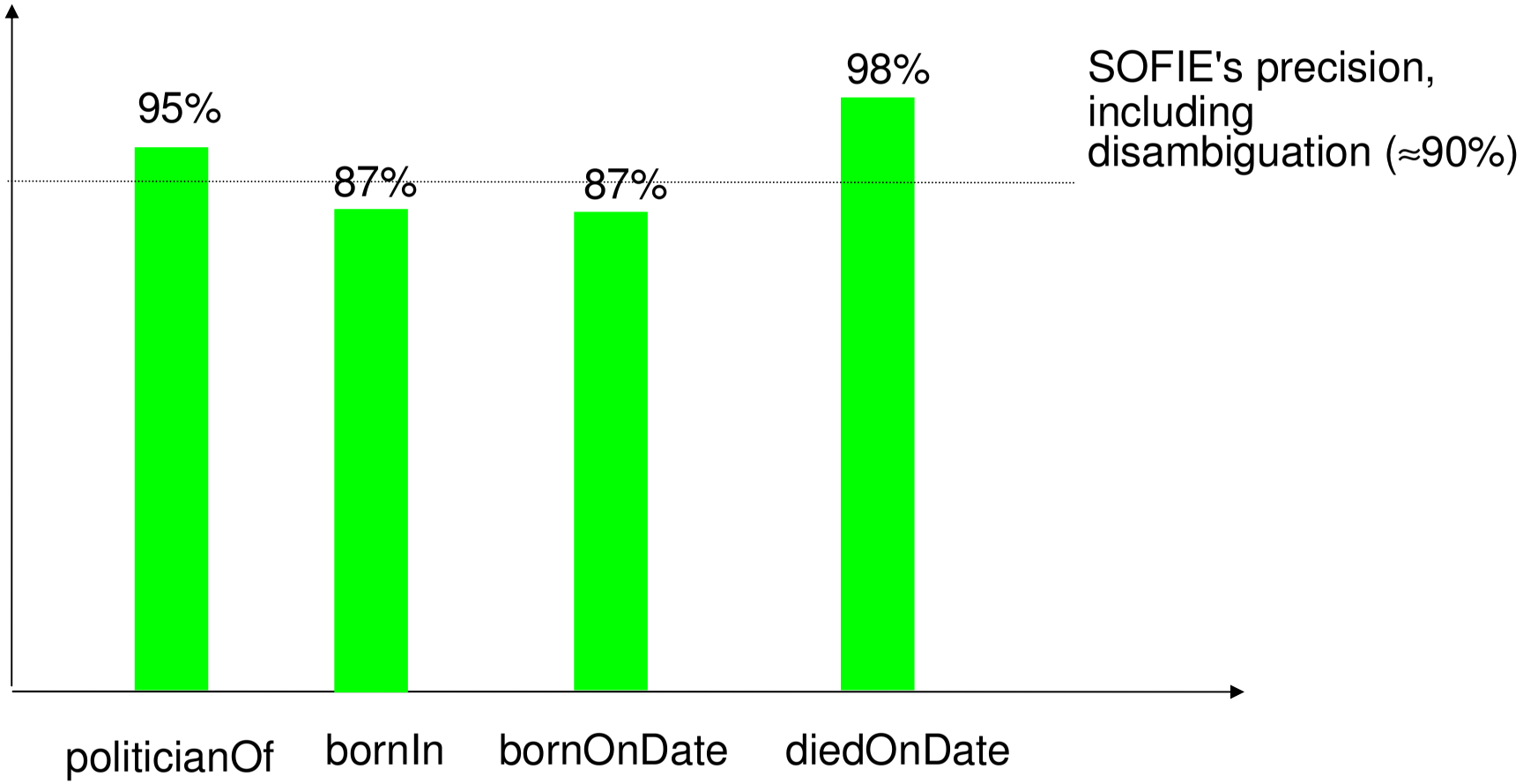

```
Algorithm
Functional MAX SAT

FOR i=1 TO 42
...
NEXT i
```

Polynomial time

Approximation Guarantee

SOFIE: A Unifying Framework

Precision values on 3700 biography documents downloaded from the Web

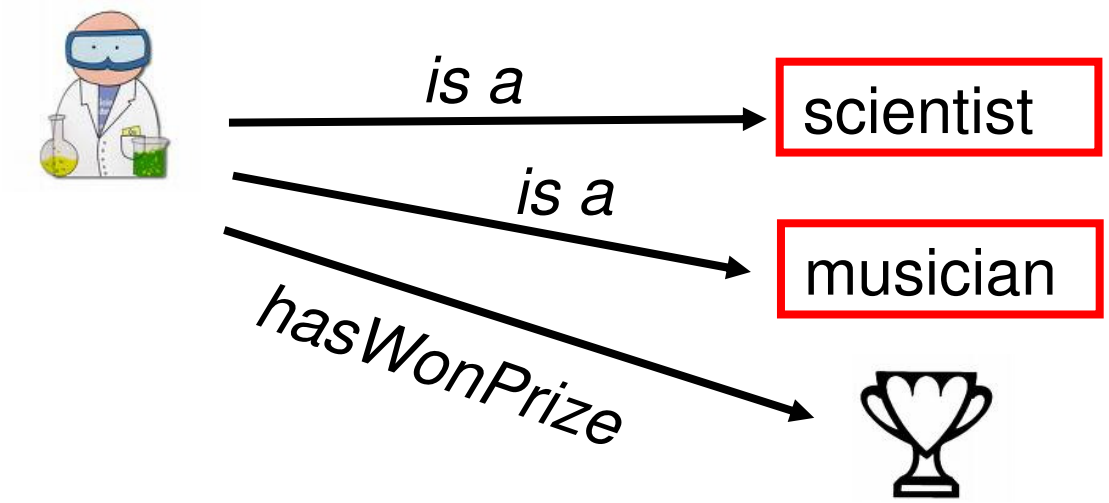


The Excellent Scientist

We're not there yet...

...but YAGO can already help us with the original question:

Which scientist was also a musician and has won a prize?



The Excellent Scientist

We're not there yet...

...but YAGO can already help us with the original question:

Which scientist was also a musician and has won a prize?

X	isa	scientist
X	isa	musician
X	hasWonPrize	Y

(DEMO)

Conclusion

- YAGO is a large ontology
- Ontological knowledge can help in many applications
- SOFIE uses logical reasoning to extend YAGO

- They do exist

http://mpii.de/yago

References

[SIGKDD 2006] Fabian M. Suchanek, Georgiana Ifrim and Gerhard Weikum
"Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents"
International Conference on Knowledge Discovery and Data Mining (SIGKDD 2006)

[WWW 2007] Fabian M. Suchanek, Gjergji Kasneci and Gerhard Weikum
"YAGO - A Core of Semantic Knowledge"
International World Wide Web conference (WWW 2007)

[Elsevier 2008] Fabian M. Suchanek, Gjergji Kasneci and Gerhard Weikum
"YAGO - A Large Ontology from Wikipedia and WordNet"
Elsevier Journal of Web Semantics 2008

[PhD] Fabian M. Suchanek
"Automated Construction and Growth of a Large Ontology"
PhD thesis, see http://mpii.de/~suchanek

The SOFIE part is also published as
Fabian M. Suchanek, Mauro Sozio, Gerhard Weikum
„SOFIE – A Self-Organizing Framework for Information Extraction“
Technical report, see http://mpii.de/~suchanek
Submitted to the International World Wide Web conference (WWW 2009)