



max planck institut
informatik

Harvesting, Searching, and Ranking Knowledge from the Web

Gerhard Weikum

weikum@mpi-inf.mpg.de

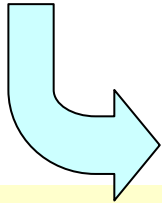
<http://www.mpi-inf.mpg.de/~weikum/>

joint work with Shady Elbassuoni, Georgiana Ifrim, Gjergji Kasneci,
Thomas Neumann, Maya Ramanath, Mauro Sozio, Fabian Suchanek

My Vision

Opportunity:

Turn the Web (and Web 2.0 and Web 3.0 ...) into the world's most comprehensive **knowledge base**



Approach:

- 1) **harvest** and combine
 - a) **hand-crafted** knowledge sources
(Semantic Web, ontologies)
 - b) **automatic** knowledge extraction
(Statistical Web, text mining)
 - c) **social** communities and **human** computing
(Social Web, Web 2.0)
- 2) express **knowledge queries**, search, and rank
- 3) everything **efficient** and **scalable**

Why Google and Wikipedia Are Not Enough

Answer „knowledge queries“ (by scientists, journalists, analysts, etc.)
such as:

- drugs or enzymes that inhibit proteases (HIV)
- connections between Thomas Mann and Goethe
- German Nobel prize winner who survived both world wars and outlived all of his four children
- how are Max Planck, Angela Merkel, Jim Gray, and the Dalai Lama related
- politicians who are also scientists

Why Google and Wikipedia Are Not Enough

Answer „knowledge queries“ (by scientists, journalists, analysts, etc.)
such as:

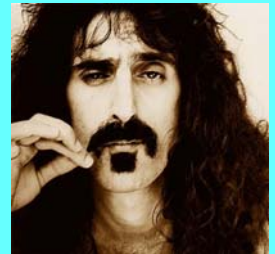
- drugs or enzymes that inhibit pro
- connections between Thomas M
- German Nobel prize winner who
and outlived all of his four children
- how are Max Planck, Angela Me
and the Dalai Lama related
- politicians who are also scientists

What is lacking?

*Information is not Knowledge.
Knowledge is not Wisdom.*

*Wisdom is not Truth
Truth is not Beauty.*

*Beauty is not Music.
Music is the best.*

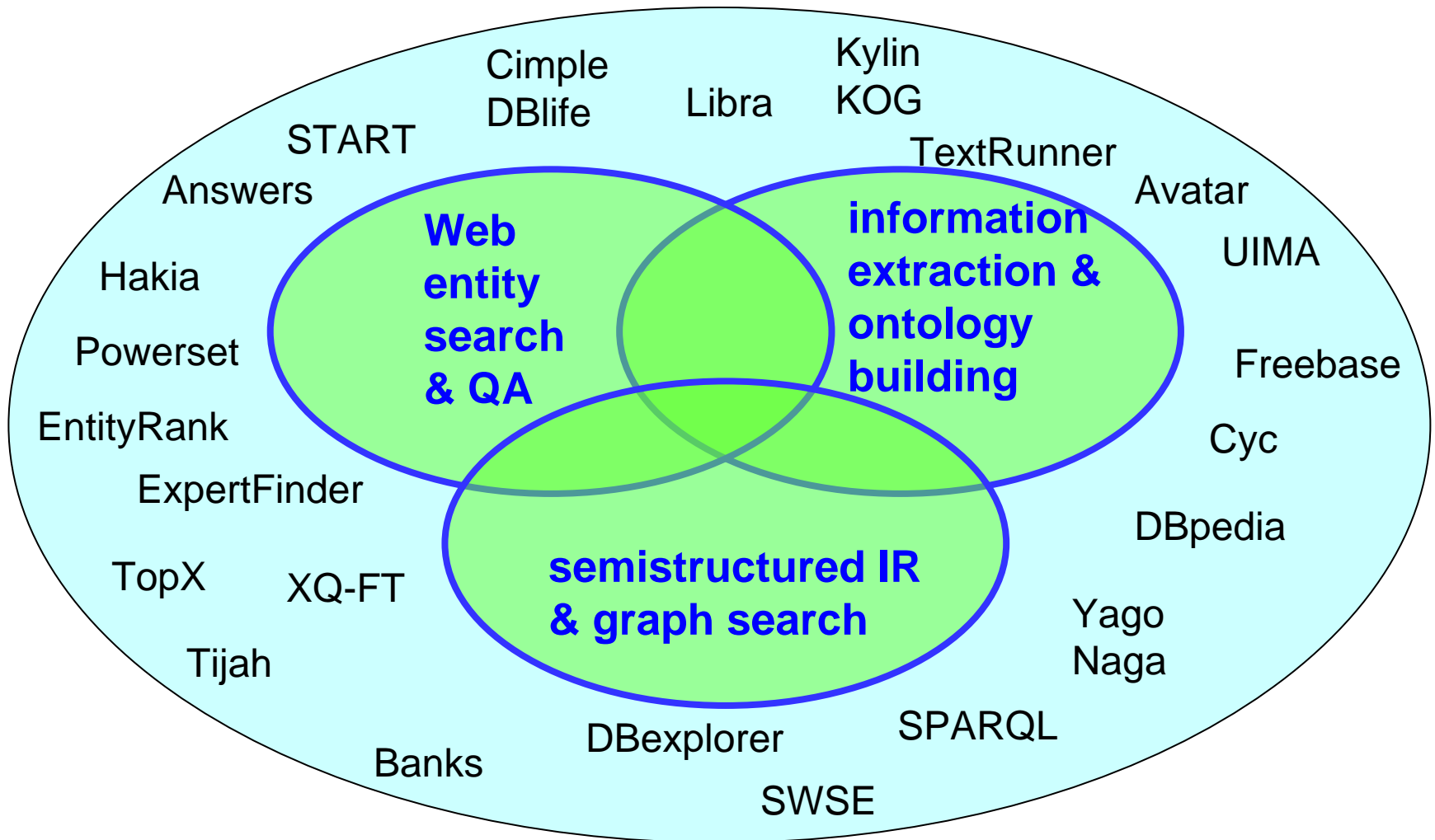


*(Frank Zappa
1940 – 1993)*

→ extract **facts** from Web pages

→ capture user intention by
concepts, entities, relations

Related Work



Relevant Projects

KnowItAll / TextRunner (UW Seattle)
IntelligenceInWikipedia (UW Seattle)
DBpedia (U Leipzig & FU Berlin)
SeerSuite (PennState)
Cimple / DBLife (U Wisconsin & Yahoo)
Avatar / System T (IBM Almaden)
Libra (MS Research Beijing)
SQoUT (Columbia U)
Wikipedia Entities (Yahoo Barcelona)
Expert Finding (U Amsterdam)
Expertise Finding (U Twente)
... **and more**
and G, Y, MS for products, locations, ...

Selected overviews in: ACM SIGMOD Record 37(4), Dec 2008

Outline

✓ Motivation

- **Information Extraction & Knowledge Harvesting (YAGO)**
- **Consistent Growth of Knowledge (SOFIE)**
- **Ranking for Search over Entity-Relation Graphs (NAGA)**
- **Efficient Query Processing (RDF-3X)**
- **Conclusion**

Information Extraction (IE): Text to Records

Max Planck

Max Karl Ernst Ludwig Planck (April 23, 1858 – October 4, 1947) was a German physicist who is considered to be the inventor of quantum theory.

Born in Kiel, Planck started his physics studies at Munich University in 1874, graduating in 1879 in Berlin. He returned to München in 1880 to teach at the university, and moved to Kiel in 1885. There he married Marie Merck in 1886. In 1889, he moved to Berlin, where from 1892 on he held the chair of theoretical physics.

In 1899, he discovered a new fundamental constant, which is named Planck's constant, and is, for example, used to calculate the energy of a photon. Also that year, he developed his own set of units of measurement based on fundamental physical constants. One year later, he discovered the law of heat radiation, which is named Planck's Radiation Law. This law became the basis of quantum theory, which emerged later in cooperation with Albert Einstein and Niels Bohr.

Person	BirthDate	BirthPlace	...
Max Planck	4/23, 1858	Kiel	
Albert Einstein	3/14, 1879	Ulm	
Mahatma Gandhi	10/2, 1869	Porbandar	



Person	ScientificResult
Max Planck	Quantum Theory

Constant

Planck's constant



extracted facts often have confidence < 1 (DB with uncertainty)

sometimes:

confidence << 1

high computational costs

Person	Organization
Max Planck	KWG / MPG

Person	Organization
Max Planck	KWG / MPG

combine NLP, pattern matching, lexicons, statistical learning

High-Quality Knowledge Sources

General-purpose ontologies and thesauri: **WordNet** family

**200 000 concepts and relations;
can be cast into**

- **description logics or**
- **graph, with weights for relation strengths
(derived from co-occurrence statistics)**

scientist, man of science -- (a person with advanced knowledge of
=> cosmographer, cosmographer -- (a scientist knowledgeable
=> bibliotist -- (someone who engages in bibliotics)
=> biologist, life scientist -- ((biology) a scientist who studies life
=> chemist -- (a scientist who specializes in chemistry)
=> cognitive scientist -- (a scientist who studies cognitive processes)
=> computer scientist -- (a scientist who specializes in the theory and
=> geologist -- (a specialist in geology)
=> linguist, linguistic scientist -- (a specialist in linguistics)
=> mathematician -- (a person skilled in mathematics)
=> medical scientist -- (a scientist who studies disease processes)
=> microscopist -- (a scientist who specializes in research with
=> mineralogist -- (a scientist trained in mineralogy)
=> oceanographer -- (a scientist who studies physical and biological
=> paleontologist, palaeontologist, fossilist -- (a specialist in paleontology)
=> physicist -- (a scientist trained in physics)
=> principal investigator, PI -- (the scientist in charge of an experiment)
=> psychologist -- (a scientist trained in psychology)
=> radiologic technologist -- (a scientist trained in radiological technology)
=> research worker, researcher, investigator -- (a scientist who
=> social scientist -- (someone expert in the study of human social behavior)
HAS INSTANCE=> Bacon, Roger Bacon -- (English scientist and philosopher who
combustion and first used lenses to correct vision (1214-1292))
HAS INSTANCE=> Franklin, Benjamin Franklin -- (printer who invented the
the Constitution, he played a major role in the American Revolution
his research in electricity (1706-1790))
HAS INSTANCE=> Galton, Francis Galton, Sir Francis Galton -- (English
psychology, anthropology, founder of eugenics and first to use the term
HAS INSTANCE=> Harvey, William Harvey -- (English physician who discovered
ovum produced by the female of the species (1578-1634))

scientist, man of science

(a person with advanced knowledge)

=> cosmographer, cosmographer

=> biologist, life scientist

=> chemist

=> cognitive scientist

=> computer scientist

...

=> principal investigator, PI

...

HAS INSTANCE => Bacon, Roger Bacon

...

Exploit Hand-Crafted Knowledge

Wikipedia and other lexical sources

Max Karl Ernst Ludwig Planck (April 23, 1858 – October 4, 1947 in Göttingen, Germany) was a German physicist. He is considered to be the founder of quantum theory, and therefore one of the most important physicists of the twentieth century.

Contents [hide]

- 1 Life and work
 - 1.1 Early Childhood
 - 1.2 Education
 - 1.3 Academic career
 - 1.4 Family
 - 1.5 Professor at Berlin University
 - 1.6 Black-body radiation
 - 1.7 Einstein and the Theory of Relativity
 - 1.8 World War and Weimar Republic
 - 1.9 Quantum mechanics
 - 1.10 Nazi dictatorship and Second World War
- 2 Honours and medals
- 3 See also
- 4 Publications
- 5 Bibliography
- 6 External links
 - 6.1 Biographies
 - 6.2 Articles
- 7 Notes

Life and work [edit]

Early Childhood [edit]

Planck came from a traditional, intellectual family. His paternal great-grandfather and grandfather were both **theology** professors in **Göttingen**, his father was a law professor in **Kiel** and **Munich**, and his paternal uncle was a judge.

Planck was born in **Kiel** to Johann Julius Wilhelm Planck and his second wife, Emma Patzig. He was the sixth child in the family, though two of his siblings were from his father's first marriage. Among his earliest memories was the marching of Prussian and

Max Planck



Max Karl Ernst Ludwig Planck

Born	April 23, 1858 Kiel, Germany
Died	October 4, 1947 Göttingen, Germany
Residence	 Germany
Nationality	 German
Field	Physicist
Institutions	University of Kiel Humboldt-Universität zu Berlin Georg-August-Universität Göttingen
Alma mater	Ludwig-Maximilians-Universität München
Academic advisor	Philipp von Jolly

Exploit Hand-Crafted Knowledge

Wikipedia and other lexical sources



```
{{Infobox_Scientist
| name = Max Planck
| birth_date = [[April 23]], [[1858]]
| birth_place = [[Kiel]], [[Germany]]
| death_date = [[October 4]], [[1947]]
| death_place = [[Göttingen]], [[Germany]]
| residence = [[Germany]]
| nationality = [[Germany|German]]
| field = [[Physicist]]
| work_institution = [[University of Kiel]]</br>
  [[Humboldt-Universität zu Berlin]]</br>
  [[Georg-August-Universität Göttingen]]
| alma_mater = [[Ludwig-Maximilians-Universität München]]
| doctoral_advisor = [[Philipp von Jolly]]
| doctoral_students =
[[Gustav Ludwig Hertz]]</br>
...
| known_for = [[Planck's constant]],
  [[Quantum mechanics|quantum theory]]
| prizes = [[Nobel Prize in Physics]] (1918)
...
}}
```

Max Planck



Max Karl Ernst Ludwig Planck

Born	April 23, 1858 Kiel, Germany
Died	October 4, 1947 Göttingen, Germany
Residence	 Germany
Nationality	 German
Field	Physicist
Institutions	University of Kiel Humboldt-Universität zu Berlin Georg-August-Universität Göttingen
Alma mater	Ludwig-Maximilians-Universität München
Academic advisor	Philipp von Jolly

Exploit Hand-Crafted Knowledge

[Wikipedia](#), [WordNet](#), and other lexical sources

Nobel Prize in Physics: Laureates (1901-1925)

1901: Röntgen 1902: Lorentz, Zeeman 1903: Becquerel, P. Curie, M. Curie 1904: Rayleigh
1905: Lenard 1906: Thomson 1907: Michelson 1908: Lippmann 1909: Marconi, Braun
1910: van der Waals 1911: Wien 1912: Dalén 1913: Kamerlingh Onnes 1914: von Laue
1915: W. L. Bragg, W. H. Bragg 1917: Barkla 1918: Planck 1919: Stark 1920: Guillaume
1921: Einstein 1922: N. Bohr 1923: Millikan 1924: Siegbahn 1925: Franck, Hertz


[Complete List](#) | [Laureates \(1926-1950\)](#) | [Laureates \(1951-1975\)](#) | [Laureates \(1976-2000\)](#) |
[Laureates \(2001- \)](#)

Categories: [1858 births](#) | [1947 deaths](#) | [Cornell University faculty](#) | [German Nobel laureates](#) | [German physicists](#) | [Members of the Pontifical Academy of Sciences](#) | [Nobel laureates in Physics](#) | [Particle physics](#) | [People from Kiel](#) | [Quantum theory physicists](#) | [Thermodynamicists](#) | [Humboldt University of Berlin alumni](#) | [University of Munich alumni](#)

YAGO: Yet Another Great Ontology

[F. Suchanek et al.: WWW'07]

- Turn Wikipedia into formal **knowledge base** (semantic DB); keep source pages as **witnesses**
- Exploit hand-crafted **categories** and **infoboxes**
- Represent facts as **knowledge triples**:

relation (entity1, entity2) 

(in FOL, compatible with **RDF**, OWL-lite, XML, etc.)

- Map relations into **WordNet** concept DAG

Examples:

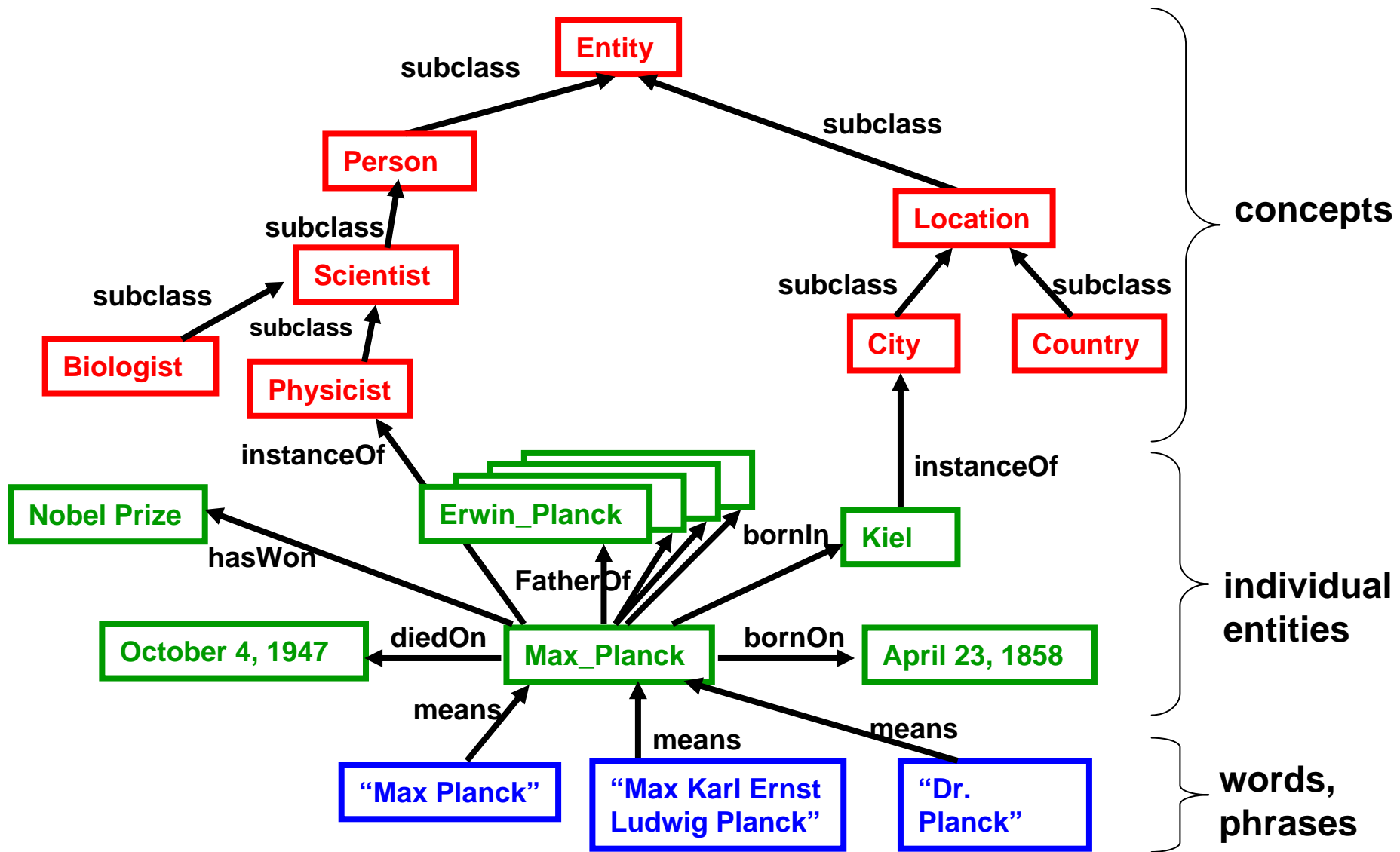


Difficulties in Wikipedia Harvesting

- **instanceOf relation**: misleading and difficult **category names**
„disputed articles“, „particle physics“, „American Music of the 20th Century“, „naturalized citizens of the United States“, ...
- **subclass relation**: mapping categories onto **WordNet classes**:
„Nobel laureates in physics“ \subset Nobel_laureates, „people from Kiel“ \subset person
- **entity name** synonyms & ambiguities:
„St. Petersburg“, „Saint Petersburg“, „M31“, „NGC224“ \rightarrow means ...
- **type (consistency) checking** for rejecting false candidates:
AlmaMater (Max Planck, Kiel) \notin Person \times University



YAGO Knowledge Base [F. Suchanek et al.: WWW 2007]

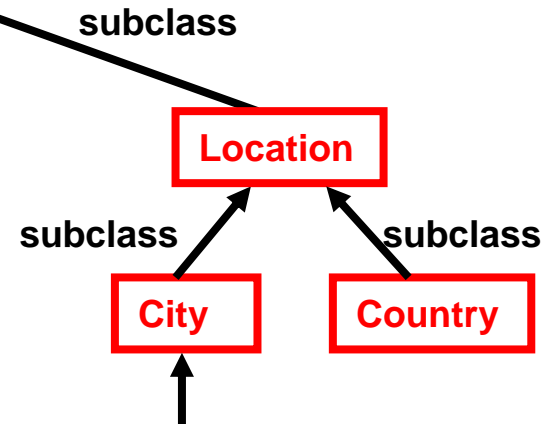


Online access and download at <http://www.mpi-inf.mpg.de/yago/>

YAGO Knowledge Base [F. Suchanek et al.: WWW 2007]

	Entities	Facts
KnowItAll	30 000	
SUMO	20 000	60 000
WordNet	120 000	80 000
Cyc	300 000	5 Mio.
TextRunner	n/a	8 Mio.
YAGO	1.9 Mio.	19 Mio.
DBpedia	1.9 Mio.	103 Mio.
Freebase	???	156 Mio.

RDF triples (entity1-relation-entity2,
subject-predicate-object)

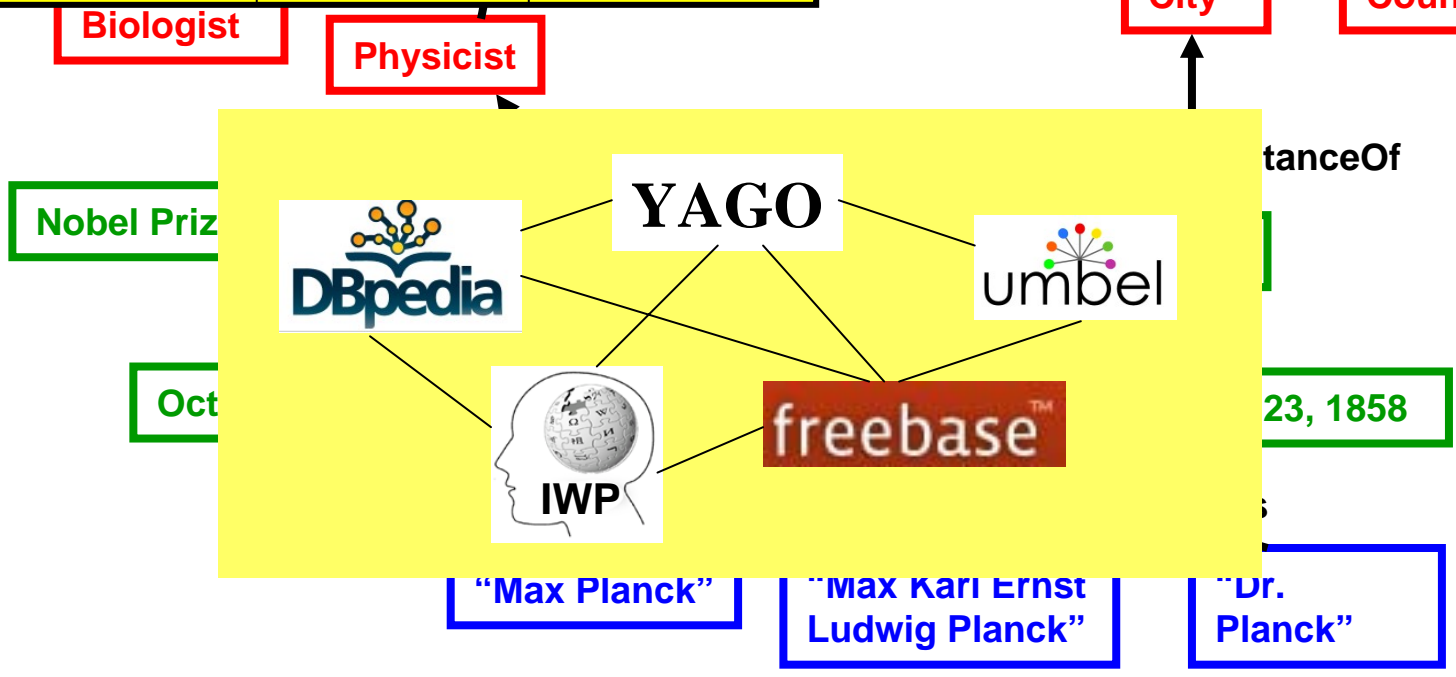


concepts

Accuracy
≈ 95%

individual
entities

words,
phrases



Online access and download at <http://www.mpi-inf.mpg.de/yago/>

Long Tail of Wikipedia

(Intelligence-in-Wikipedia Project) [Wu / Weld: WWW 2008]

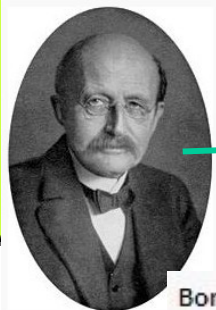
YAGO & DBpedia mappings of entities onto classes are valuable assets

Ricardo Baeza-Yates

From Wikipedia, the free encyclopedia

Ricardo Baeza-Yates (born March 21, 1961) is a Chilean computer scientist and director of the Yahoo! Research Labs at Barcelona, Spain and Santiago, Chile. His Ph.D. from the University of Waterloo was entitled *Efficient Text Searching*, supervised by Gaston Gonnet and granted in 1989.

Learning infobox attributes
→ sparse & noisy training data



Physicist

Computer Scientist

Musician

University

Scientist

Artist

Organization

Frank Zappa

Zappa's interest in composing and arranging proliferated in his last high-school years. By his final year, he was writing, arranging and conducting avant-garde performance pieces for the school orchestra.^[21] He graduated from Antelope Valley High School in 1958, and later acknowledged two of his music teachers on the sleeve of the 1966 album *Freak Out!*^[22] Due to his family's frequent moves, Zappa

Born	Apr	Died	lost at sea January 20, 2007
	Kiel,	Nationality	American
Died	Oct	Fields	Computer Science
	Gött	Institutions	IBM Tandem Computers DEC Microsoft
Nationality	Ger	Alma mater	University of California, Berkeley
Fields	Phy	Notable awards	Turing Award
Institutions	Univ		
	Univ		
	Univ		
Alma mater	Ludwig-Maximilians-Universität München		

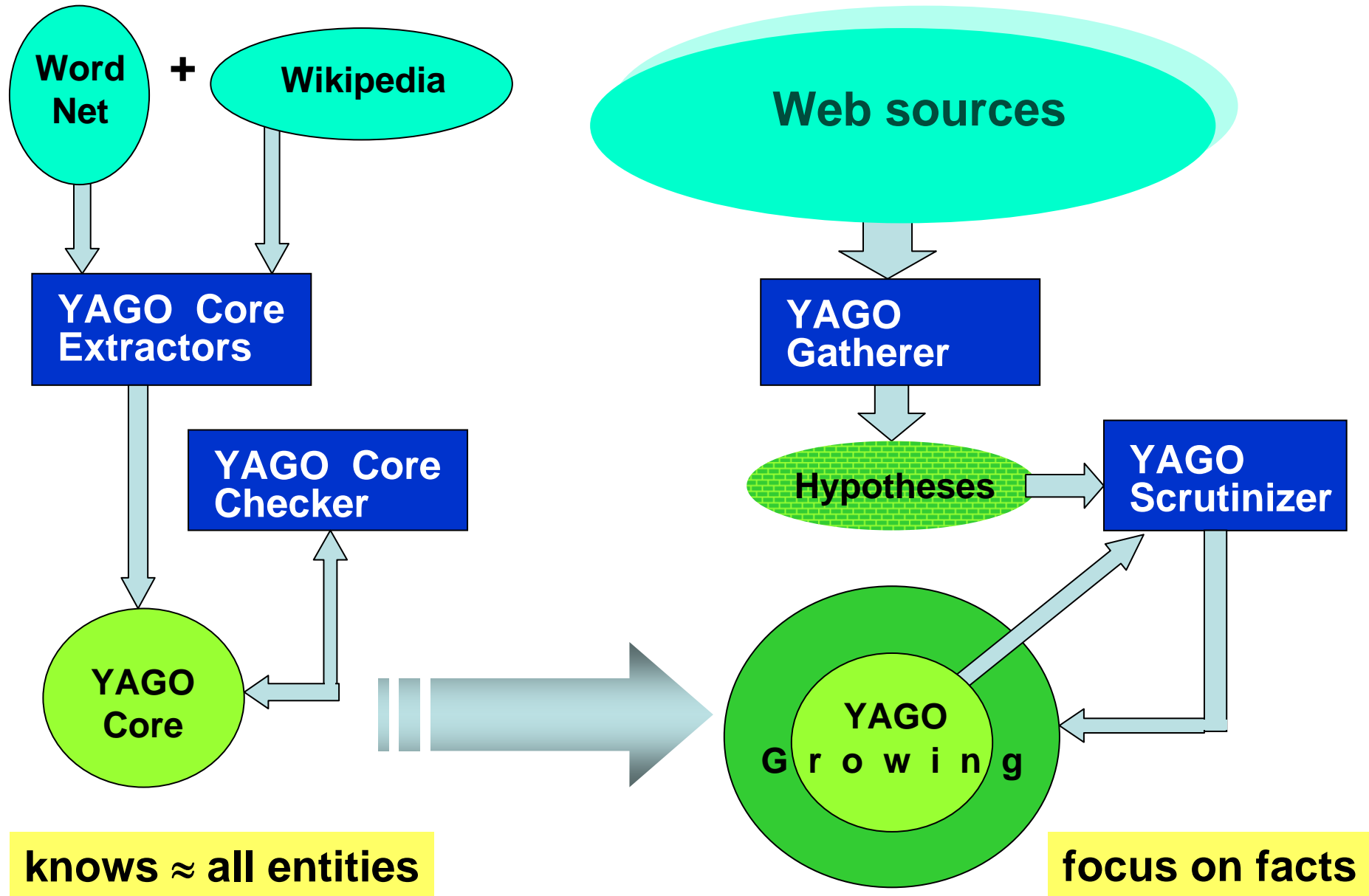
Outline

✓ Motivation

✓ Information Extraction & Knowledge Harvesting (YAGO)

- Consistent Growth of Knowledge (SOFIE)
- Ranking for Search over Entity-Relation Graphs (NAGA)
- Efficient Query Processing (RDF-3X)
- Conclusion

Maintaining and Growing YAGO



knows \approx all entities

focus on facts

SOFIE: Self-Organizing Framework for IE

[F. Suchanek et al.: WWW 2009]

Reconcile

- textual/linguistic **pattern-based IE** with **statistics**

seeds → patterns → facts → patterns → ...

- declarative **rule-based IE** with **constraints**

functional dependencies: *hasCapital* is a function

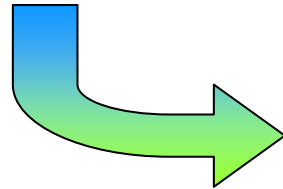
inclusion dependencies: *isCapitalOf* \subseteq *isCityOf*



From Facts to Patterns to Hypotheses

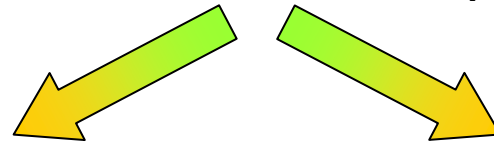
Facts

- Spouse (HillaryClinton, BillClinton)
- Spouse (MelindaGates, BillGates)
- Spouse (AngelaMerkel, JoachimSauer)



Patterns

- occurs (*X and her husband Y*, AngelaMerkel, JoachimSauer) [4]
- occurs (*X and her husband Y*, MelindaGates, BillGates) [2]
- occurs (*X and her husband Y*, CarlaBruni, NicolasSarkozy) [3]
- occurs (*X married to Y*, MelindaGates, BillGates) [2]
- occurs (*X loves Y*, LarryPage, Google) [5]



Hypotheses

- Spouse (CarlaBruni, NicolasSarkozy)
- Spouse (LarryPage, Google)
- Spouse (AngelaMerkel, UlrichMerkel)

- expresses (*and her husband*, Spouse)
- expresses (*married to*, Spouse)
- expresses (*loves*, Spouse)

Adding Consistency Constraints

Facts

Spouse (HillaryClinton, BillClinton)
 Spouse (MelindaGates, BillGates)
 Spouse (AngelaMerkel, JoachimSauer)

Patterns

occurs (X and her husband Y, AngelaMerkel, JoachimSauer) [4]
 occurs (X and her husband Y, MelindaGates, BillGates) [2]
 occurs (X and her husband Y, CarlaBruni, NicolasSarkozy) [3]
 occurs (X married to Y, MelindaGates, BillGates) [2]
 occurs (X loves Y, LarryPage, Google) [5]

Hypotheses

Spouse (CarlaBruni, NicolasSarkozy) expresses (and her husband, Spouse)
 Spouse (LarryPage, Google) expresses (married to, Spouse)
 Spouse (AngelaMerkel, UlrichMerkel) expresses (loves, Spouse)

$\text{occur}(P, X, Y) \wedge \text{expresses}(P, \text{Spouse}) \Rightarrow \text{Spouse}(X, Y)$
 $\text{occur}(P, X, Y) \wedge \text{Spouse}(X, Y) \Rightarrow \text{expresses}(P, \text{Spouse})$
 $\text{Spouse}(X, Y) \wedge Y \neq Z \Rightarrow \neg \text{Spouse}(X, Z)$
 $\text{Spouse}(X, Y) \Rightarrow \text{Type}(X, \text{Person}) \wedge \text{Type}(Y, \text{Person})$

Constraints



Representation by Clauses

Spouse (HillaryClinton, BillClinton)

occurs (*X and her husband Y,*

Angela Merkel, JoachimSauer) [4]

Clauses connect facts, patterns, hypotheses, constraints

Treat **hypotheses as variables**, facts as constants:

$(\neg 1 \vee \neg A \vee 1)$, $(\neg 1 \vee \neg A \vee B)$, $(\neg 1 \vee \neg C)$, $(\neg D \vee E)$, $(\neg D \vee F)$, ...

Clauses can be weighted by pattern statistics

Solve **weighted Max-Sat** problem:

assign truth values to variables s.t.

total weight of satisfied clauses is max!

Spouse (LarryPage, Google)

expresses (*married to, Spouse*)

Spouse (AngelaMerkel, UlrichMerkel)

expresses (*loves, Spouse*)

occur (and her husband, AngelaMerkel, JoachimSauer)
 \wedge expresses (and her husband, Spouse) \Rightarrow Spouse (AngelaMerkel, JoachimSauer)

occur (and her husband, CarlaBruni, NicolasSarkozy)
 \wedge expresses (and her husband, Spouse) \Rightarrow Spouse (CarlaBruni, NicolasSarkozy)

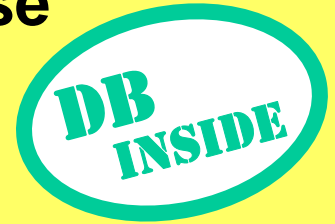
Spouse (AngelaMerkel, JoachimSauer) \Rightarrow \neg Spouse (AngelaMerkel, UlrichMerkel)

Spouse (LarryPage, Google) \Rightarrow Type (LarryPage, Person) \wedge Type (Google, Person)

SOFIE: Consistent Growth of YAGO

[F. Suchanek et al.: WWW 2009]

- self-organizing framework for **scrutinizing hypotheses** about new facts, enabling **automated growth** of the knowledge base
- unifies **pattern-based IE**, **consistency checking** and **entity disambiguation**



Experimental evidence:

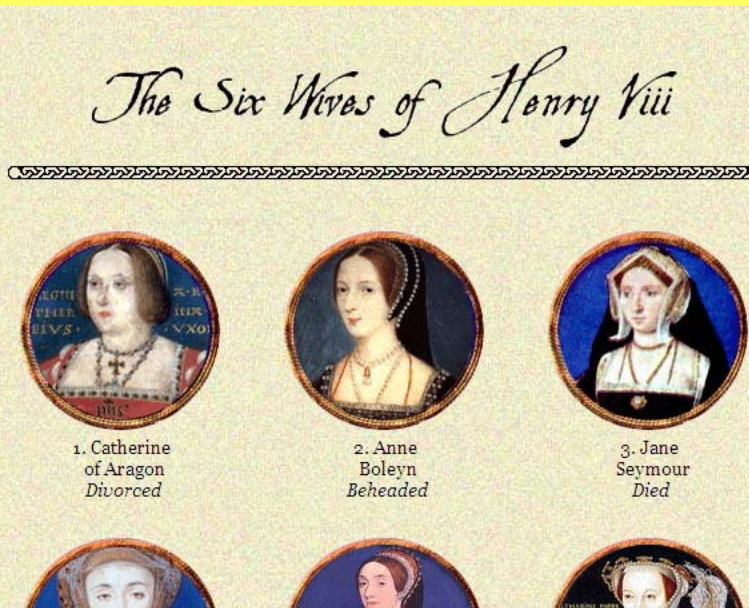
- **input:** biographies of 400 US senators, 3500 HTML files
- **output:** birth/death date&place, politicianOf (state)
- **run-time:** 7 h parsing, 6 h hypotheses, 2 h weighted Max-Sat
- **precision:** 90-95 %, except for death place
- **discovered patterns:**
 - politicianOf: X was a * of Y, X represented Y, ...
 - deathDate: X died on Y, X was assassinated on Y, ...
 - deathPlace: X was born in Y

Open Issues

- **Temporal Knowledge:**
temporal validity of all facts (spouses, CEO's, etc.)
- **Total Knowledge:**
all possible relations („Open IE“), but in canonical form
worksFor, employedAt, isEmployeeOf, ... → affiliation
- **Multimodal Knowledge:**
photos, videos, sound, sheetmusic of
entities (*people, landmarks, etc.*) and
facts (*marriages, soccer matches, etc.*)
- **Scalable Knowledge Gathering:**
high-quality IE at the rate at which
news, blogs, Wikipedia updates are produced !

Scalability: Benchmark Proposal

for **all people** in Wikipedia (100,000's) gather **all spouses**, incl. divorced & widowed, and corresponding **time periods!**
>95% accuracy, >95% coverage, in one night



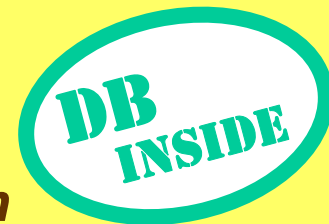
Nicolas Sarkozy

Born	28 January 1955 (age 53) Paris, France
Birth name	Nicolas Paul Stéphane Sarközy
Political party	RR (?–2002) UMP (2002–)
Spouse	Marie-Dominique Culioli (div.) Cécilia Ciganer-Albéniz (div.) Carla Bruni
Children	Pierre (by Culioli) Jean (by Culioli) LOUIS (by Ciganer-Albéniz)
Residence	Élysée Palace

redundancy of sources helps, stresses **scalability** even more

consistency constraints are potentially helpful:

- functional dependencies: $\{husband, time\} \rightarrow wife$
- inclusion dependencies: $marriedPerson \subseteq adultPerson$
- age/time/gender restrictions: $birthdate + \Delta < marriage < divorce$



Outline

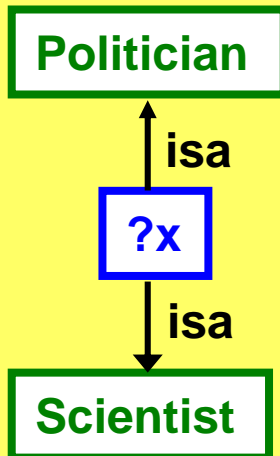
- ✓ **Motivation**
- ✓ **Information Extraction & Knowledge Harvesting (YAGO)**
- ✓ **Consistent Growth of Knowledge (SOFIE)**
- **Ranking for Search over Entity-Relation Graphs (NAGA)**
- **Efficient Query Processing (RDF-3X)**
- **Conclusion**

NAGA: Graph Search with Ranking

[G. Kasneci et al.: ICDE 2008, ICDE 2009]

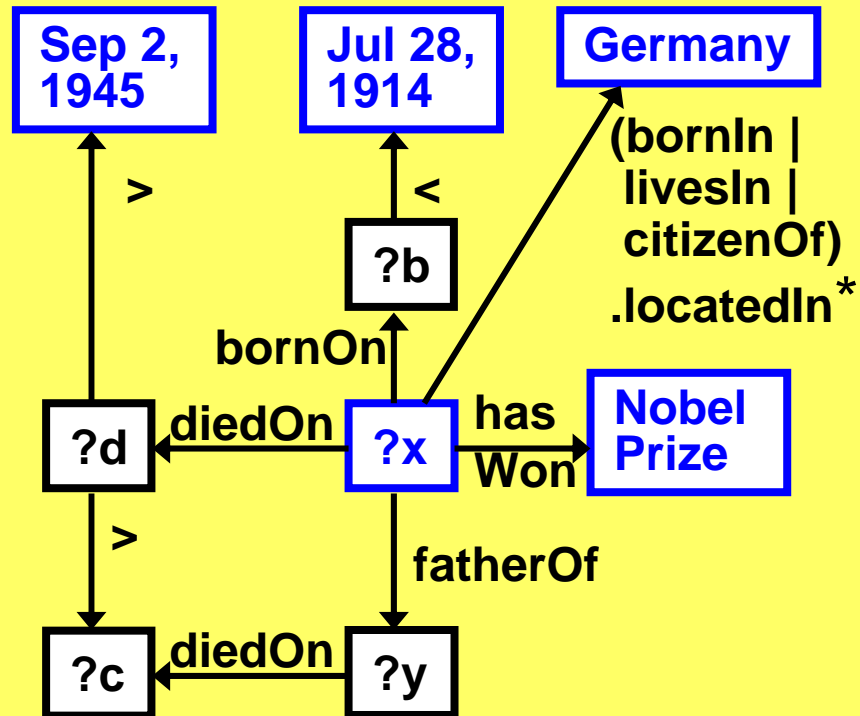
Graph-based search on knowledge bases with built-in **ranking** based on **confidence** and **informativeness**

Simple query



*?x isa Politician .
?x isa Scientist*

Complex query (with regular expr.)



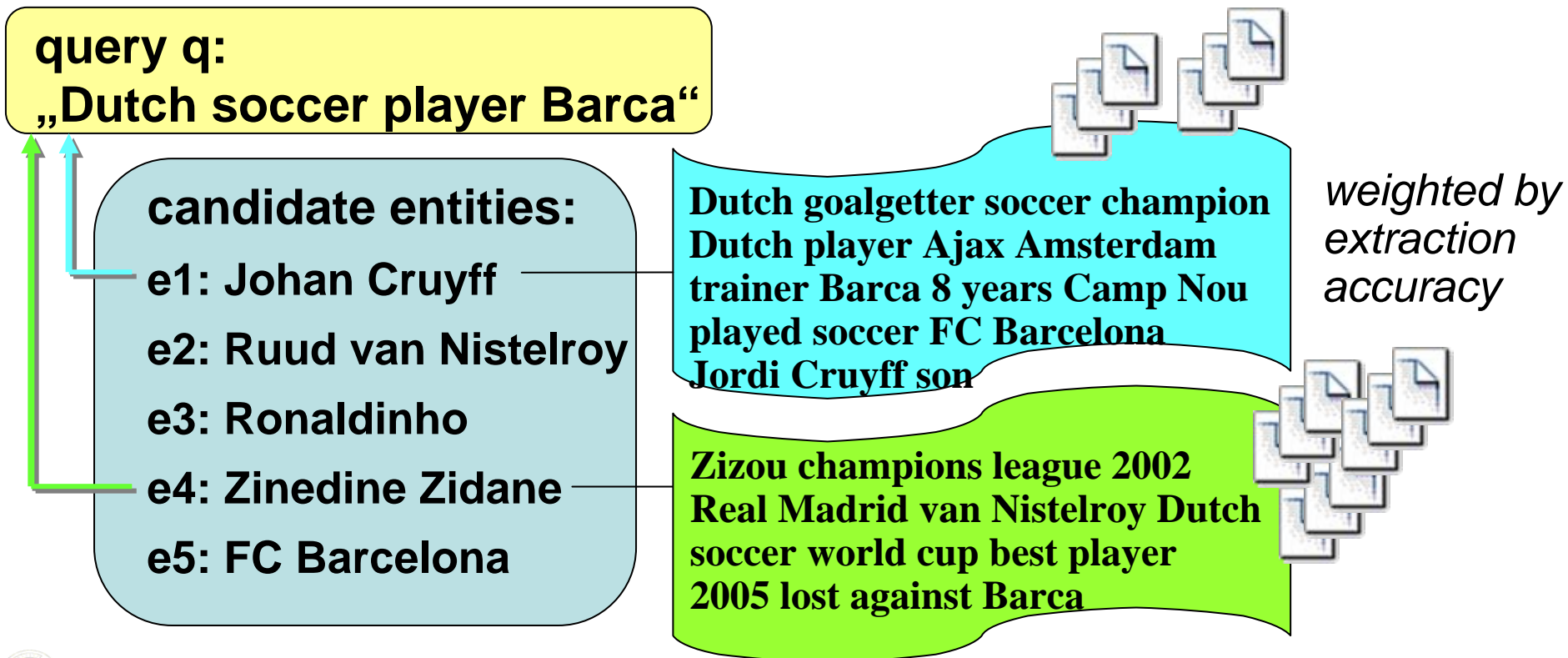
*?x hasWon NobelPrize . ?x fatherOf ?y
?x bornOn ?b . FILTER (?b < Jul-28-1914)
?x diedOn ?d*

Statistical Language Models (LM's) for Entity Ranking

[work by U Amsterdam, MSR Beijing, U Twente, Yahoo Barcelona, ...]

$$\text{score}(e, q) = \lambda P[q | e] + (1 - \lambda) P[q] \sim \prod \frac{P[q_i | e_i]}{P[q_i]} \sim \text{KL}(\text{LM}(q) | \text{LM}(e))$$

LM (entity e) = prob. distr. of words seen in context of e



LM for Fact (Entity-Relation) Ranking

$$\text{score}(f, q) = \lambda P[q | f] + (1 - \lambda) P[q] \sim \prod \frac{P[q_i | f_i]}{P[q_i]} \sim \text{KL}(\text{LM}(q) | \text{LM}(f))$$

query q

q_1 :
?x hasWon NobelPrize

q_2 :
?x bornIn Germany

fact pool for candidate answers

f1: Einstein hasWon NobelPrize	200
f2: Gruenberg hasWon NobelPrize	50
f3: Gruenberg hasWon JapanPrize	20
f4: Vickrey hasWon NobelPrize	50
f5: Cerf hasWon TuringAward	100
f6: Einstein bornIn Germany	100
f7: Gruenberg bornIn Germany	20
f8: Goethe bornIn Germany	200
f9: Schiffer bornIn Germany	150
f10: Vickrey bornIn Canada	10
f11: Cerf bornIn USA	100

instantiation
(user interests)

LM(q_1):

Einstein hasWon NP → 200/300

Gruenberg hasWon NP → 50/300

Vickrey hasWon NP → 50/300

plus smoothing

LM(q_2):

Einstein bornIn G → 100/470

Gruenberg bornIn G → 20/470

Goethe bornIn G → 200/470

Schiffer bornIn G → 150/470

witnesses
may be weighted
by confidence

NAGA Example



Score: 6.247457440558111E-7

```
"scientist" —means—> scientist_110560637
Benjamin_Franklin —type—> Massachusetts_politicians
"politician" —means—> politician_110451263
American_scientists —subClassOf—> scientist_110560637
Benjamin_Franklin —type—> American_scientists
Massachusetts_politicians —subClassOf—> politician_110451263
```

- \$@politician = politician_110451263
- \$@scientist = scientist_110560637
- \$X = Benjamin_Franklin

Score: 3.185850362140424E-7

```
"scientist" —means—> scientist_110560637
"politician" —means—> politician_110451263
Paul_Wolfowitz —type—> American_political_scientists
American_political_scientists —subClassOf—> scientist_110560637
Paul_Wolfowitz —type—> Jewish-American_politicians
Jewish-American_politicians —subClassOf—> politician_110451263
```

- \$@politician = politician_110451263
- \$@scientist = scientist_110560637
- \$X = Paul_Wolfowitz

Score: 1.121658976926192E-7

```
Angela_Merkel —type—> German_scientists
"scientist" —means—> scientist_110560637
German_Christian_Democrat_politicians —subClassOf—>
  politician_110451263
Angela_Merkel —type—> German_Christian_Democrat_politicians
"politician" —means—> politician_110451263
German_scientists —subClassOf—> scientist_110560637
```

- \$@politician = politician_110451263
- \$@scientist = scientist_110560637
- \$X = Angela_Merkel

Query:

**?x isa politician
?x isa scientist**

Results:

**Benjamin Franklin
Paul Wolfowitz
Angela Merkel**

...

Outline

- ✓ **Motivation**
- ✓ **Information Extraction & Knowledge Harvesting (YAGO)**
- ✓ **Consistent Growth of Knowledge (SOFIE)**
- ✓ **Ranking for Search over Entity-Relation Graphs (NAGA)**
- **Efficient Query Processing (RDF-3X)**
- **Conclusion**

Scalable Semantic Web: Pattern Queries on Large RDF Graphs

schema-free RDF triples: subject-property-object (SPO)

example: *Einstein hasWon NobelPrize*

SPARQL triple patterns: Select ?p,?c Where {

?p isa scientist . ?p hasWon NobelPrize .

?p bornIn ?t . ?t inCountry ?c . ?c partOf Europe}

large join queries, unpredictable workload,

difficult physical design, difficult query optimization

AllTriples

S	P	O
Einstein	hasWon	Nobel
Einstein	bornIn	Ulm
Ronaldo	hasWon	FIFA
Spain	partOf	Europe
France	partOf	Europe
...

Person

S	hasWon	bornIn	..
Einstein	Nobel	Ulm	..
Ronaldo	FIFA	Rio	..
..

hasWon

S	O
Einstein	Nobel
Ronaldo	FIFA
...	...

Country

S	partOf	capital	..
..

bornIn

S	O
...	...

Semantic-Web engines (Sesame, Jena, etc.)
did not provide scalable query performance

Scalable Semantic Web: RDF-3X Engine

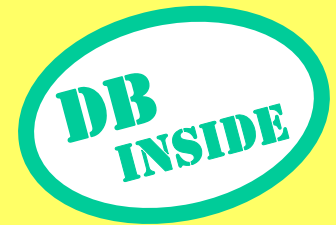
[T. Neumann et al.: VLDB'08]

- RISC-style, tuning-free system architecture
- map literals into ids (dictionary) and precompute **exhaustive indexing** for SPO triples:

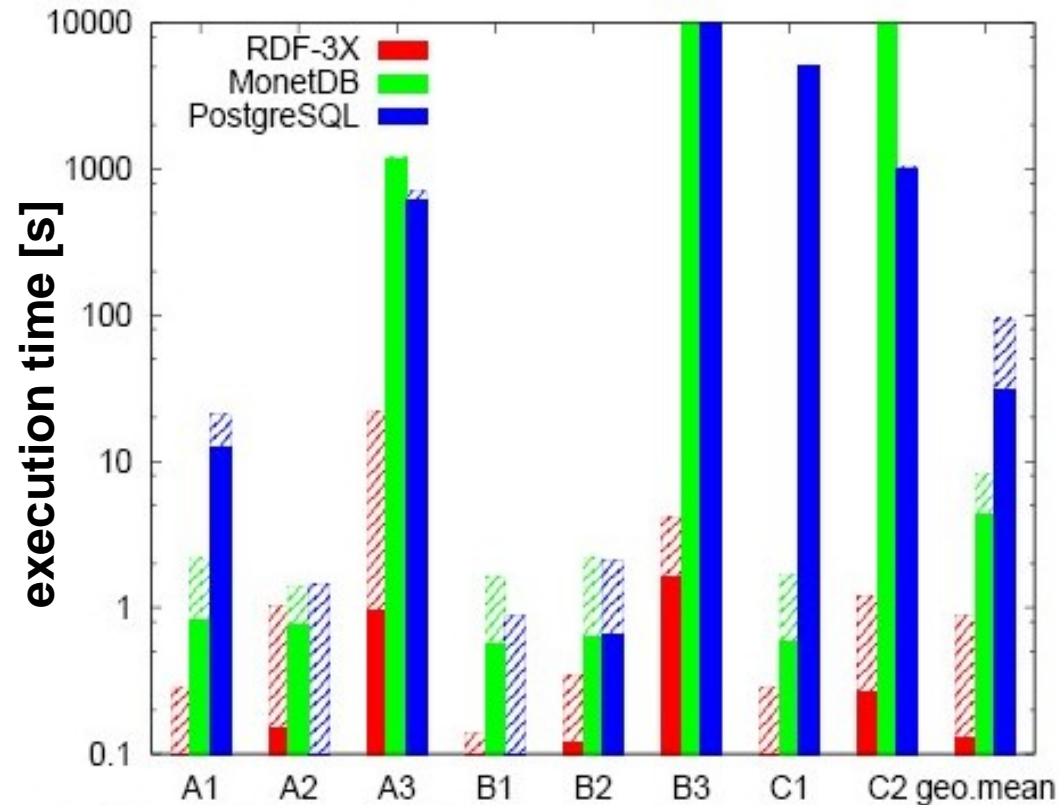
SPO, SOP, PSO, POS, OSP, OPS,
SP*, PS*, SO*, OS*, PO*, OP*, S*, P*, O*

very high compression

- efficient **merge joins** with order-preservation
- **join-order optimization**
by dynamic programming over subplan \times result-order
- **statistical synopses** for accurate result-size estimation



Performance Experiments



**Librarything
social-tagging excerpt
(36 Mio. triples)**

Benchmark queries such as:

```
Select ?t Where {  
  ?b hasTitle ?t .  
  ?u romance ?b .  
  ?u love ?b .  
  ?u mystery ?b .  
  ?u suspense ?b .  
  ?u crimeNovel ?c .  
  ?u hasFriend ?f .  
  ?f ... }
```

books tagged with
romance, love,
mystery, suspense
by users who
like crime novels
and have friends
who ...

RDF-3X on PC (2 GHz, 2 GB RAM, 30 MB/s disk) compared to:

- column-store (for property tables) using **MonetDB**
- triples store (with selected indexes) using **PostgreSQL**

similar results on **YAGO**, **Uniprot** (845 Mio. triples) and **Billion-Triples**

Outline

- ✓ **Motivation**
 - ✓ **Information Extraction & Knowledge Harvesting (YAGO)**
 - ✓ **Consistent Growth of Knowledge (SOFIE)**
 - ✓ **Ranking for Search over Entity-Relation Graphs (NAGA)**
 - ✓ **Efficient Query Processing (RDF-3X)**
- **Conclusion**

Take-Home Message

- turn Wikipedia, Web, news, literature, ... into comprehensive **knowledge base** of facts
→ YAGO core

- reconcile **rule-based** & **pattern-based** info extraction (Semantic-Web & Statistical-Web) with **consistency constraints**
→ YAGO growth with SOFIE

- enable search & ranking over **entity-relation graphs**
→ NAGA, RDF-3X

*Information is not Knowledge.
Knowledge is not Wisdom.
Wisdom is not Truth
Truth is not Beauty.
Beauty is not Music.
Music is the best.*



(Frank Zappa, 1940 – 1993)



Technical Challenges

- Handling Time

- **extracting** temporal attributes
- **reasoning** on validity times of facts
- **life-cycle** management of KB

- Scalable Performance

- high-quality **dynamic IE** at the rate of news/blogs/Wikipedia updates
- „Marital Knowledge“ **benchmark**

- Query Language and Ranking

- querying **expressive** but **simple** (Sparql-FT ?)
- **LM**-based ranking vs. **PR/HITS**-style vs. **learned scoring** from user behavior
- **efficient top-k** queries on ER graphs

... and more

Thank You !

