Semantische Suche

Outline

- **Knowledge-based Search – What and why**
- Existing Knowledge Bases
- Building and Maintaining Knowledge Bases
- Knowledge-Based Search

Why textual information is not enough

- **Meaning** (**semantics**) of single words and word sequences **often unclear** (unless context is known)
  Examples: **bank**, **java**, **jaguar**, **mustang**, **president**, …
- Language may have different interpretations
  - **It is very hot today in Dagstuhl.**
  - **„That‘s great weather today!“**
  - **Yesterday the head of state met her American counterpart in the capital of Germany.**

When? **Who?** **Where?**

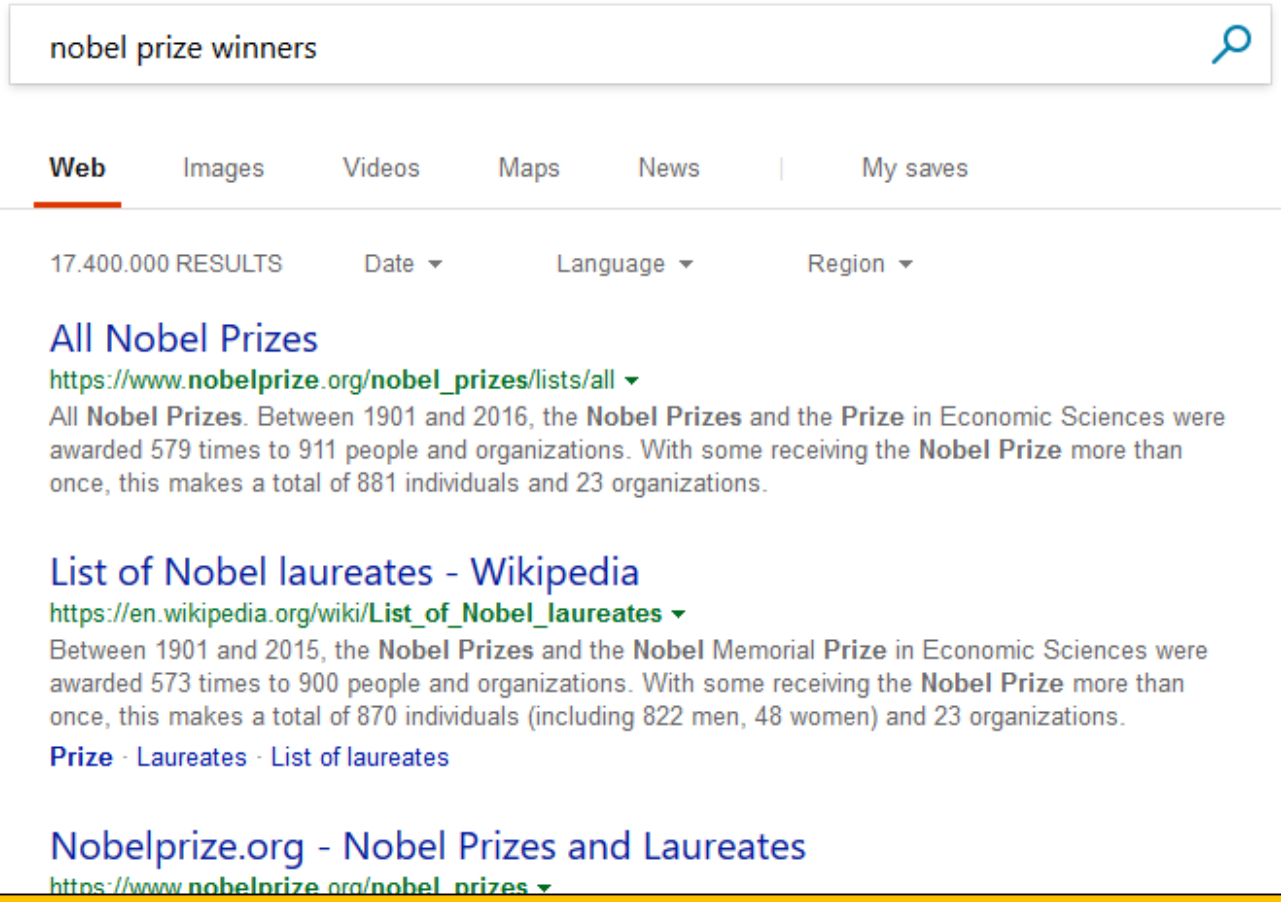
Need for factual and contextual background knowledge

- Search should not retrieve document, but **answer to question**
  Who was president of the USA when Donald Trump was born?
- Often no single document contains required information

Donald Trump was born in 1946 in New York.

Harry S. Truman was president of the United States of America from 1945 to 1953.

Why Web Search is not enough



Results are pages that satisfy the information need, but a list of persons would be much better

Why Web Search is not enough

nobel prize winners

Web Images Videos Maps News My

17.400.000 RESULTS Date Language Region

All Nobel Prizes
https://www.nobelprize.org/nobel_prizes/lists/all
All **Nobel Prizes**. Between 1901 and 2016, the **Nobel Prizes** and the **Priz**
awarded 579 times to 911 people and organizations. With some receiving th
once, this makes a total of 881 individuals and 23 organizations.

List of Nobel laureates - Wikipedia
https://en.wikipedia.org/wiki/List_of_Nobel_laureates
Between 1901 and 2015, the **Nobel Prizes** and the **Nobel** Memorial **Prize**

Additional structured query results from semantic knowledge base, but no real understanding of query

List of Nobel Peace Prize laureates - Wikipedia
https://en.wikipedia.org/wiki/List_of_Nobel_Peace_Prize_laureates
List of **Nobel** Peace **Prize** laureates ... Sixteen women have won the **Nobe**
any other **Nobel Prize**. ... "**Nobel Prize winners** by ...

Nobelpreis

Der Nobelpreis ist eine seit 1901 jährlich vergebene Auszeichnung, die der schwedische Erfinder und Industrielle Alfred Nobel gestiftet hat. In seinem Testament legte er fest, dass mit seinem Vermögen eine Stiftung gegründet werden sollte, deren Zin... +

Wikipedia Official website

Categories: Friedensnobelpreis · Nobelpreis für Physik · Nobelpreis für Literatur · Nobelpreis für Physiologie oder Medizin +

Related people

See all (50+)


Alfred Nobel
Named after


Malala Yousafzai
Award winner


Kailash Satyarthi
Award winner


Marie Curie
Award winner


Albert Einstein
Award winner

People also search for

See all (10+)


Friedensnobelpreis


Nobelpreis für Literatur


Nobelpreis für Physik


Nobelpreis für Physiologi...


Nobelpreis für Chemie

Why Web Search is not enough

Web Images Videos Maps News My saves

63.400.000 RESULTS Date Language Region

Nobel Laureates and Country of Birth - Nobel Prize
https://www.**nobelprize**.org/**nobel_prizes**/lists/countries.html
Nobel Laureates and Country of Birth. Alfred **Nobel**, the man behind the **Nobel Prize**, wrote in his will 1895: "It is my express wish that in awarding the ...

EUROPA - European Union receives Nobel Peace Prize 2012
https://**europa**.eu/.../about-eu/history/2010-today/2012/eu-**nobel**_en
The **prize** was conferred in Oslo on 10 December. Like all **winners**, the EU received the **prize** from the chairman of the Norwegian **Nobel** Committee. The EU was represented by the President of the **European** Council, the President of the **European** Commission and the President of the **European** ...

Nobel Peace prize winners - the full list | World news ...
www.theguardian.com › World news › **Nobel** peace **prize**
The **European** Union has been awarded the **Nobel** peace **prize**. This morning the **European** Union was announced as the 2012 **Nobel** peace **prize** winner. The **winners** list for 2011 was split between three ...

All Nobel Prizes
https://www.**nobelprize**.org/**nobel_prizes**/lists/all
All **Nobel Prizes**. Between 1901 and ... The **Nobel** Peace **Prize** 2012 **European** Union (EU) "for over six decades contributed to the advancement of peace and ...

Nobel Prize Winners By Country - WorldAtlas.com
www.worldatlas.com/.../top-30-countries-with-**nobel-prize-winners**.html
Nobel Prize Winners By Country. United States has the most noble **prize winners** with 353 followed by the UK with 125 and Germany with 105. The **Nobel Prize** is the most distinguished and coveted ...

Related searches

peace nobel prize winners
muslim nobel prize winners
ig nobel prize
nobel prize **org**
nobel **peace** prize
nobel prize **literature list**
nobel prize winners **list**
nobel prize winners **by country**

Very limited understanding of the query, no semantic data to show here

Question Answering, not Search

nobel prize winners

All Images News Books Videos More Settings Tools

Nobel Prize › Winners (2016)

Juan Manuel Santos Peace Prize 	David J. Thouless Physics 	Yoshinori Ohsumi Physiology or Medicine 
Bob Dylan Literature 	Fraser Stoddart Chemistry 	Bengt Holmström Economic Sciences 
Duncan Haldane Physics 	Jean-Pierre Sauvage Chemistry 	Oliver Hart Economic Sciences 
J. Michael Kosterlitz Physics 	Ben Feringa Chemistry 	

All Nobel Prizes
https://www.nobelprize.org/nobel_prizes/lists/all/
David J. Thouless, F. Duncan M. Haldane and **J. Michael Kosterlitz**. "for theoretical discoveries of topological phase transitions and topological phases of matter" ...

Nobelprize.org - Nobel Prizes and Laureates
https://www.nobelprize.org/nobel_prizes/
Most Popular Laureates. Martin Luther King Jr. Peace Prize 1964. **Albert Einstein**. Physics Prize 1921. **Niels Bohr**. Physics Prize 1922. **Marie Curie**. Physics Prize 1903. Rabindranath Tagore.

Nobel Prize
Award

The Nobel Prize is a set of annual international awards bestowed in

Anwers come directly from semantic knowledge base (for Google: *Knowledge Graph*)

Question Answering, not Search

nobel prize winners from germany

nobel prize winners from 2015

nobel prize winners from 2012

All Images News

About 5.700.000 results

Choose a year:

- The Nobel Prize in
- The Nobel Prize in
- The Nobel Prize in
- The Nobel Prize in
- The Nobel Peace
- The Sveriges Riks 2012.

The Nobel Prize in
https://www.nobelprize.o

nobel prize winners from europe

All Imag

About 14.000

German N

- Stefan W
- Thomas
- Herta M
- Harald z
- Peter Gr
- Gerhard
- Theodor
- Wolfgan

More iten

German N
https://www.research-in-germany.org/en/research-landscape/nobel-laureates.html

nobel prize winners in computer science

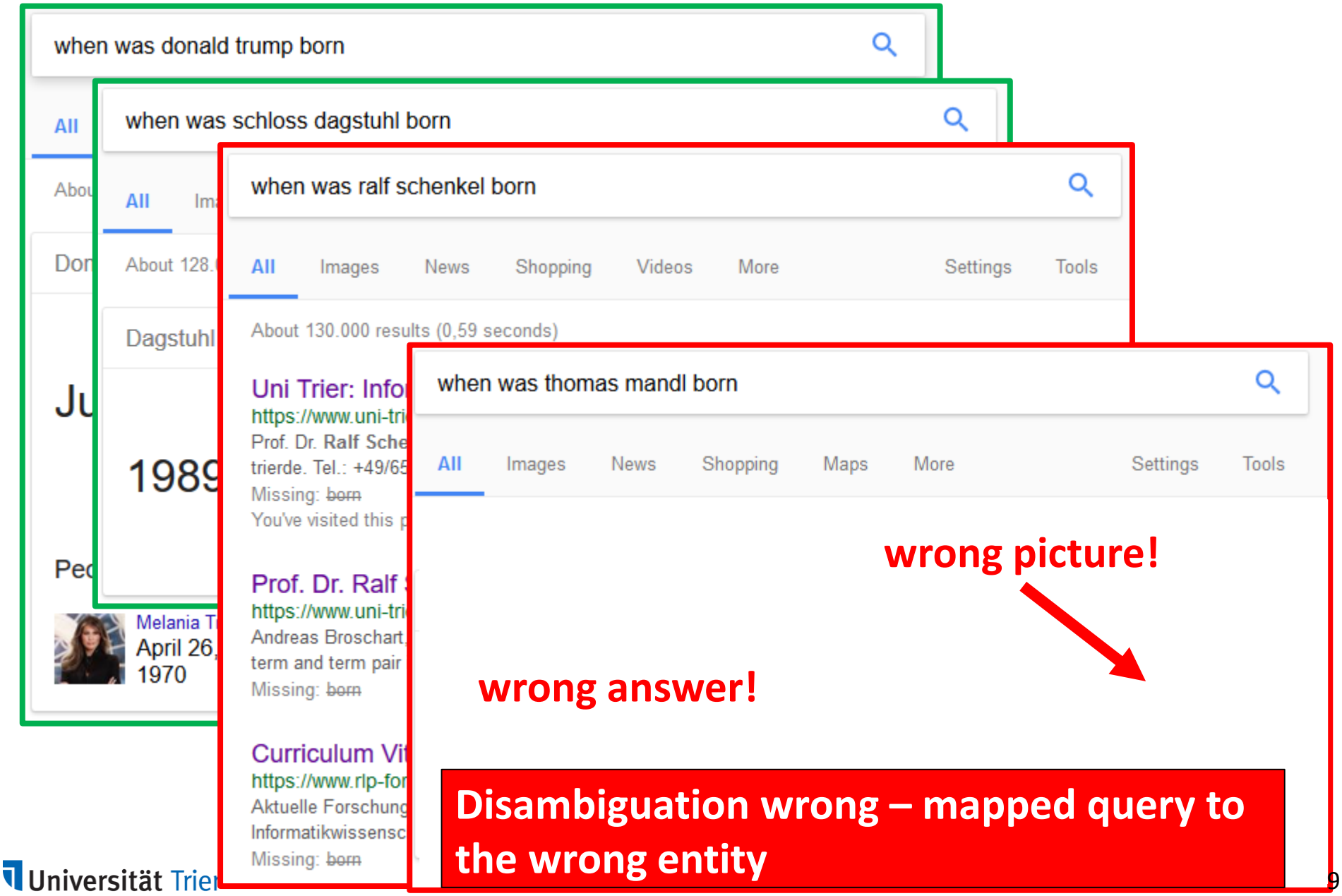
All Images News Videos Maps More Settings Tools

About 806.000 results (0,67 seconds)

The **Turing Award** is generally recognized as the highest distinction in computer science and the "Nobel Prize of computing". The award is named after Alan Turing, a British mathematician and reader in mathematics at the University of Manchester.

Turing Award - Wikipedia
https://en.wikipedia.org/wiki/Turing_Award

Question Answering on the Web



Question Answering on the Web

~ 2010

Very limited knowledge, very limited understanding of question

More complex examples

who was president of the us when the atomic bomb was dropped

All Image

About 8.980.0

Harry

Harry S. Tr
atomic bom

Who was t
https://www.q

who was president of the us when the war began

All News Images Videos Shopping More Settings Tools

Abou

Wh
US

#
32
33
34
35

46 more rows

Results extracted fro
http://retiary.org/mi

Search for: Who were the

who invented google

All Images N

Google › Founde

Larry Page

who invented the wheel

All Images Videos Shopping News More Settings Tools

About 23.400.000 results (0,43 seconds)

No one knows for sure how the **wheel** and axle was **invented**, but we do know that it was **invented** during a time when humans had already began constructing sailboats and casting metal alloys like bronze. Mar 27, 2017

Who Invented the Wheel and Axle? | Wonderopolis
https://wonderopolis.org/wonder/who-invented-the-wheel-and-axle

About this result Feedback

Knowledge Search – Wolfram Alpha

Knowledge Search – Wolfram Alpha

Who was mayor of Wadern when Barack Obama was born?

not enough facts known by the system!

http://www.wolframalpha.com

Early Research System – Google Squared

Formula one winners?

formula one winners | Square it

entourage actors	science fiction tv series	noble gases	8000 meter peaks
beatles songs	italian provinces	mountains in peru	fighter jets
dog breeds	kings of england	israeli prime ministers	small hypoallergenic dogs

Start with an empty Square

Early Research System – Google Squared

Formula one winners?

Google squared labs

formula one winners | Square it | Add to thi

Google Squared

Unsaved

formula one winners

Item Name	Image	Description	Date Of Birth	Team
Denny Hulme		Denis Clive "**Denny**" **Hulme** OBE (18 June 1936–4 October 1992) was a New Zealand car racer, the 1967 Formula One World Champion for the Brabham team. ...	1936-06-18	brabham,
Fernando Alonso		**Fernando Alonso** Díaz (born 29 July 1981) is a Spanish Formula One racing driver and a two -time World Champion, who is currently racing for Ferrari alongside ...	1981-07-29	Ferrari
Lewis Hamilton		**Lewis Hamilton** was in the wrong two weeks ago, but he was in the right under .... Who Was To Blame For The **Lewis Hamilton**-Mark Webber Shunt In Singapore? ...	1985-01-07	McLaren-
Niki Lauda		Andreas Nikolaus "**Niki**" **Lauda** (born February 22, 1949 in Vienna) is an Austrian former Formula One (F1) racing driver and three-time F1 World Champion. ...	February 22, 1949	March, BR Brabham,
Keke Rosberg		The **Keke Rosberg** Fansite. You are viewing the text version of this site. To view the full version please install the Adobe Flash Player and ensure your web	6 December, 1948	theodore, fittipaldi, w mclaren

http://www.google.com/squared/

Early Research System – Google Squared

Google squared labs

formula one winners from europe | Square it | Add to this Square

Unsa

formula one winners from europe

Item Name	Image	Description	Date Of Birth	Firs
Ayrton Senna		**Ayrton Senna** da Silva was a Brazilian racing driver and three-time Formula One world champion. He died in a crash while leading the 1994 San Marino Grand ...	21 March, 1960	198 Prix
Michael Schumacher		**Michael Schumacher** is a Formula One racing driver currently driving for Mercedes GP. Most famous for his legendary ten year-spell with Ferrari, Schumacher ...	3 January 1969	199 Prix
Fernando Alonso		**Fernando Alonso** Díaz (born 29 July 1981) is a Spanish Formula One racing driver and a two -time World Champion, who is currently racing for Ferrari alongside ...	1981-07-29	200 Prix
Juan Manuel Fangio		**Juan Manuel Fangio** (June 24, 1911 – July 17, 1995), nicknamed El Chueco ("knock-kneed") or El Maestro ("The Master"), was a racing car driver from Argentina ...	June 24, 1911	195
Niki Lauda		Andreas Nikolaus "**Niki**" **Lauda** (born February 22, 1949 in Vienna) is an Austrian former Formula One (F1) racing driver and three-time F1 World Champion. ...	February 22, 1949	197 Prix

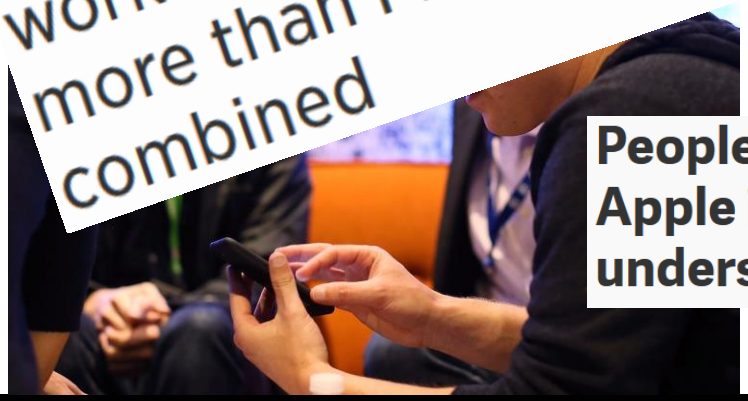
Knowledge for Intelligence

- entity recognition & disambiguation
- understanding natural language & speech
- knowledge services & reasoning for semantic apps
  (e.g. deep QA)
- semantic search: precise answers to advanced queries
  (by scientists, students, journalists, analysts, etc.)

- ★ German president when Barack Obama was born? ✓
- ★ Female soccer players from Europe?
- ★ Politicians who are also scientists?
- ★ Relationships between Carlo Ancelotti, Jürgen Klinsmann, Felix Magath, Jupp Heynckes, Giovanni Trapattoni?
- ★ Enzymes that inhibit HIV?
  Influenza drugs for teens with high blood pressure?
  ...

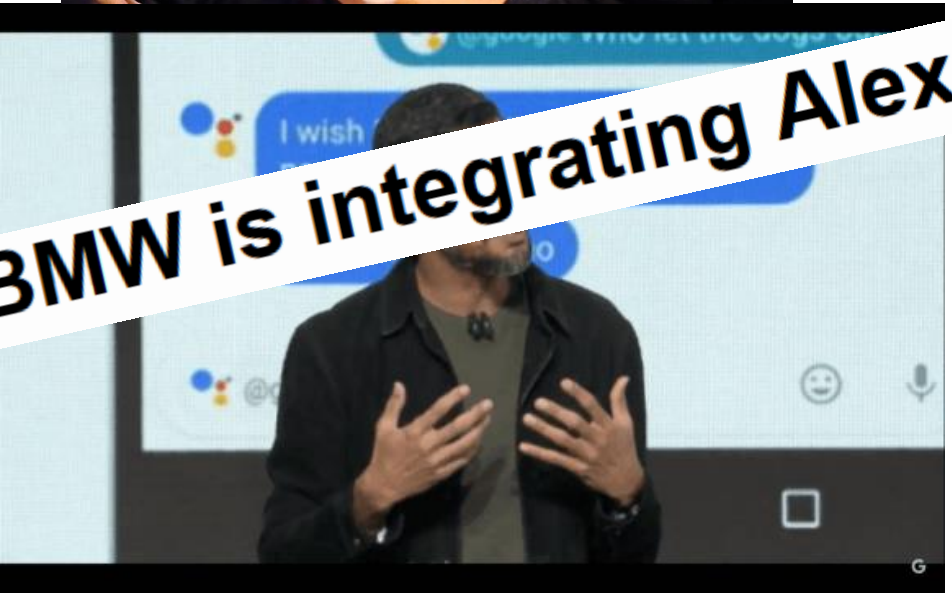
Amazon has 5,000 people working on Echo and Alexa -- more than Fitbit and GoPro combined

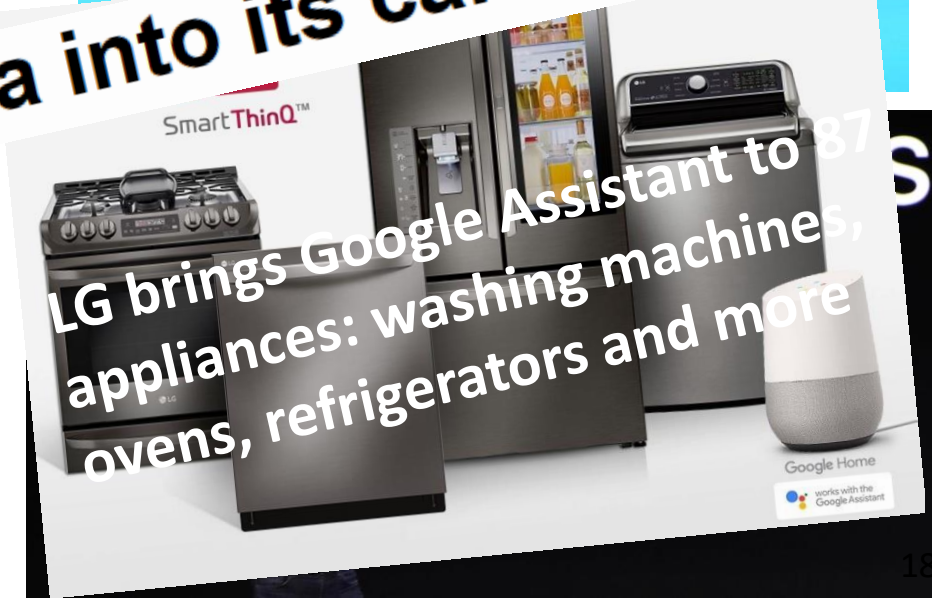
British navy warships 'to use Siri' as technology transforms warfare



People are using Siri as a therapist, so Apple is seeking engineers who understand psychology

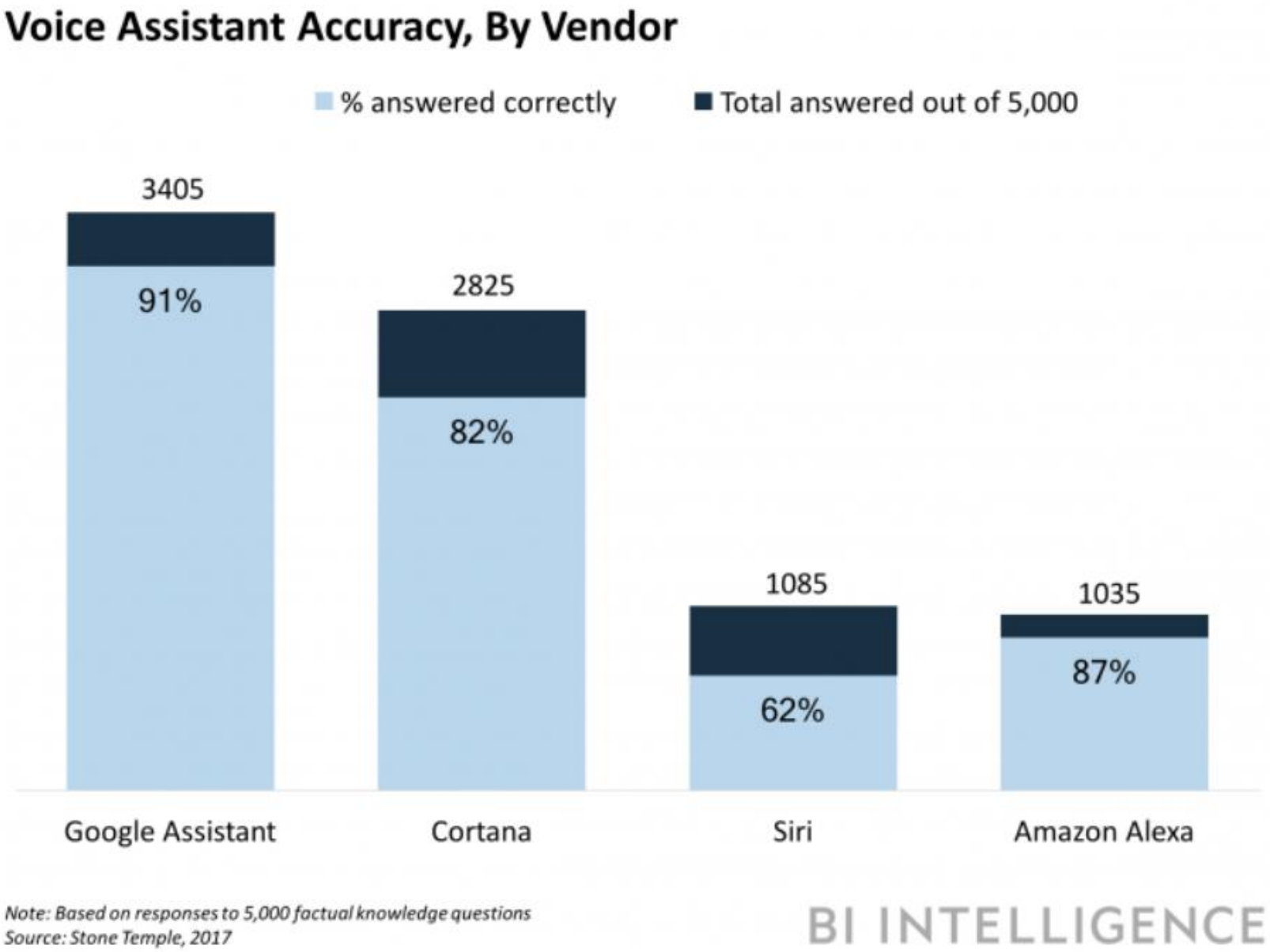
BMW is integrating Alexa into its cars in 2018





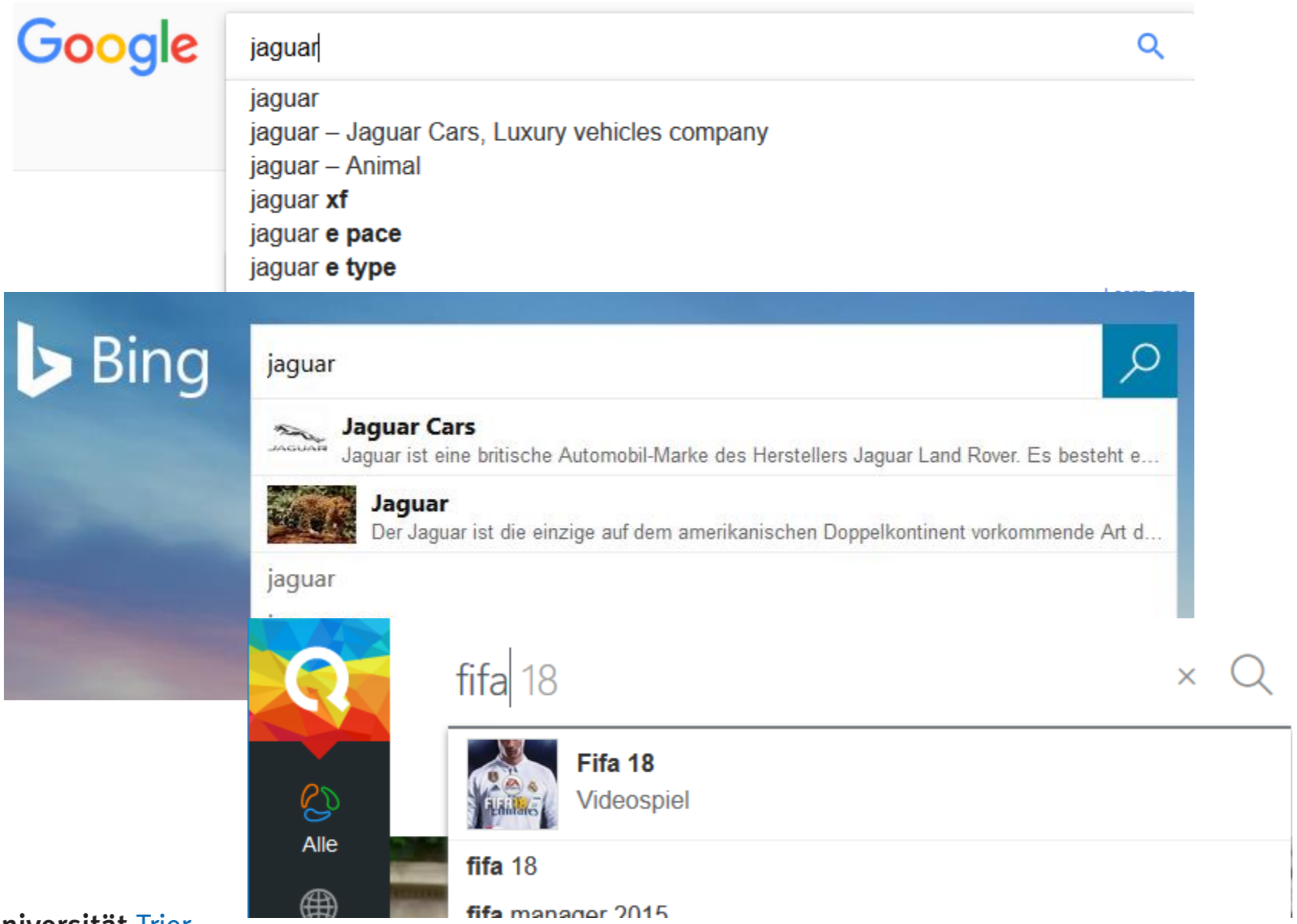
LG brings Google Assistant to 87 appliances: washing machines, ovens, refrigerators and more

Today



https://www.onmsft.com/news/cortana-trails-google-assistant-but-is-far-more-accurate-that-siri-or-alexa

Application: Query Disambiguation



Application: Infobox Construction

WIKIPEDIA
Die freie Enzyklopädie

Artikel | Diskussion | Lesen | Bearbeiten | Versionsgeschichte | Suche

Thomas Mandl

Thomas Mandl (* 7. Februar 1979 in Eisenstadt) ist ein österreichischer Fußballspieler. Aktuell spielt Mandl beim LASK Linz in der Bundesliga.

Karriere [Bearbeiten]

Der Torhüter begann seine Karriere in der burgenländischen Landesliga bei ASKÖ Hirm, wo er bereits mit 15 Jahren erster Torwart war. Nach guten Leistungen klopfte der SV Mattersburg an, wo er ab 1994 spielte. 1997 wechselte er dann zu Austria Wien. Bei der Wienern war er anfangs nur Ersatztorhüter für Franz Wohlfahrt. Nachdem Wohlfahrts Karriere dem Ende zuging, wurde er sein Na
Leistungen wurde er sogar österreichischer Nationaltorhüter

2003 kam dann der Rückschlag. Die Austria verpflichtete Joey
Australier noch durchsetzen, aber nach groben Schnitzern wurde
degradiert. Im Jänner 2004 nahm er dann das Angebot von Sturm
bekommen. Mandl wurde von den Grazern für ein halbes Jahr aus

automatically generate
by knowledge harvesting

Thomas Mandl	
Spielerinformationen	
Geburtstag	7. Februar 1979
Geburtsort	Eisenstadt, Österreich
Größe	185 cm
Position	Tor
Vereine in der Jugend	
1986–1992	ASK Hirm
1992–1997	SV Mattersburg
Vereine als Aktiver	
Jahre	**Verein**
1997–2004	Austria Wien

Outline

- Knowledge-based Search – What and Why
- **Existing Knowledge Bases**
  - **Principles**
  - **WordNet**
  - **YAGO and DBPedia**
  - **Linked Open Data**
- Building and Maintaining Knowledge Bases
- Knowledge-Based Search

What is a Knowledge Base?

- A structured representation of entities and their relationships
- In machine-readable form
- With querying capabilities
- With deductive reasoning capabilities to derive implicit facts

Entities & Classes

Which entity types (classes, unary predicates) are there?

scientists, doctoral students, computer scientists, …
female humans, male humans, married humans, …

Which subsumptions should hold

(subclass/superclass, hyponym/hypernym, inclusion dependencies)?

subclassOf (computer scientists, scientists),
subclassOf (scientists, humans), …

Which individual entities belong to which classes?

instanceOf (Surajit Chaudhuri, computer scientists),
instanceOf (BarbaraLiskov, computer scientists),
instanceOf (Barbara Liskov, female humans), …

Which names denote which entities?

means ("Lady Di", Diana Spencer),
means ("Diana Frances Mountbatten-Windsor", Diana Spencer), …
means ("Madonna", Madonna Louise Ciccone),
means ("Madonna", Madonna(painting by Edward Munch)), …

Taxonomy

Instances

Text

Relationships

Which binary relations exist between entities of which classes (type signatures)?

- hasAdvisor (Person, Person)
- graduatedAt (Person, Location)
- hasWonPrize (Person, Prize)
- bornOn (Person, Date)
- marriedTo (Person, Person)

Which instances (pairs of individual entities) are there for given binary relations with specific type signatures?

- hasAdvisor (JimGray, MikeHarrison)
- hasAdvisor (HectorGarcia-Molina, Gio Wiederhold)
- hasAdvisor (Susan Davidson, Hector Garcia-Molina)
- graduatedAt (JimGray, Berkeley)
- graduatedAt (HectorGarcia-Molina, Stanford)
- hasWonPrize (JimGray, TuringAward)
- bornOn (JohnLennon, 9Oct1940)
- diedOn (JohnLennon, 8Dec1980)
- marriedTo (JohnLennon, YokoOno)

Taxonomy

Instances

Text?

Representing Knowledge

RDF (Resource Description Framework, W3C):
subject-property-object (SPO) triples, binary relations
structure, but no (prescriptive) schema

facts (RDF triples):

1: (JimGray, hasAdvisor, MikeHarrison)
2: (SurajitChaudhuri, hasAdvisor, JeffUllman)
3: (Madonna, marriedTo, GuyRitchie)
4: (NicolasSarkozy, marriedTo, CarlaBruni)

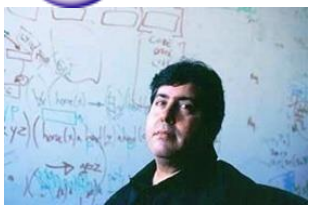
facts about facts:

5: (1, inYear, 1968)
6: (2, inYear, 2006)
7: (3, validFrom, 22-Dec-2000)
8: (3, validUntil, Nov-2008)
9: (4, validFrom, 2-Feb-2008)
10: (2, source, SigmodRecord)

temporal & provenance annotations
can refer to reified facts via fact identifiers

History of Digital Knowledge Bases

Cyc



WordNet

from humans
for humans

guitarist ⊂ {player,musician}
⊂ artist

Wikipedia

∀ x: human(x) ⇒
(∃ y: mother(x,y) ∧
∃ z: father(x,z))

4.5 Mio. English articles
20 Mio. contributors

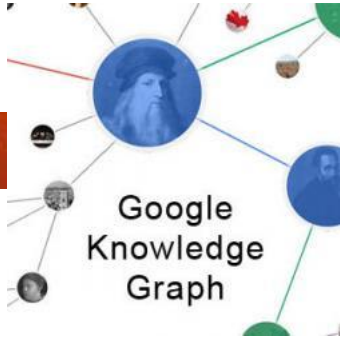
WolframAlpha

from algorithms
for machines

DBpedia

yago
select knowledge

freebase



1985 1990 2000 2005 2010

WordNet Thesaurus

WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: partner Search WordNet

change) Change

) relations, "W:" = Show Word (lexical) relations

3 concepts / classes & their synonyms (synset's)

Noun

- S: (n) spouse, **partner**, married person, mate, better half (a person's partner in marriage)
- S: (n) collaborator, cooperator, **partner**, pardner (an associate in an activity or endeavor or sphere of common interest) *"the musician and the librettist were collaborators"; "sexual partners"*
- S: (n) **partner** (a person who is a member of a partnership)

Verb

- S: (v) **partner** (provide with a partner)
- S: (v) **partner** (act as a partner) *"Astaire partnered Rogers"*

[Miller/Fellbaum 1998] http://wordnet.princeton.edu/

WordNet Thesaurus

Noun

- S: (n) spouse, **partner**, married person, mate, better half (a person's partner in marriage)
  - *direct hyponym* / *full hyponym*
    - S: (n) bigamist (someone who marries one person while already legally married to another)
    - S: (n) consort (the husband or wife of a reigning monarch)
    - S: (n) helpmate, helpmeet (a helpful partner)
    - S: (n) husband, hubby, married man (a married man; a woman's partner in marriage)
    - S: (n) monogamist, monogynist (someone who practices monogamy (one spouse at a time))
    - S: (n) newlywed, honeymooner (someone recently married)
    - S: (n) polygamist (someone who is married to two or more people at the same time)
    - S: (n) wife, married woman (a married woman; a man's partner in marriage)
  - *member holonym*
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) relative, relation (a person related by blood or marriage) *"police are searching for relatives of the deceased"; "he has distant relations back in New Jersey"*
    - S: (n) domestic partner, significant other, spousal equivalent, spouse equivalent (a person (not necessarily a spouse) with whom you cohabit and share a long-term sexual relationship)
  - *derivationally related form*
- S: (n) collaborator, cooperator, **partner**, pardner (an associate in an activity or endeavor or sphere of common interest) *"the musician and the librettist were collaborators"; "sexual partners"*
- S: (n) **partner** (a person who is a member of a partnership)

subclasses (hyponyms)

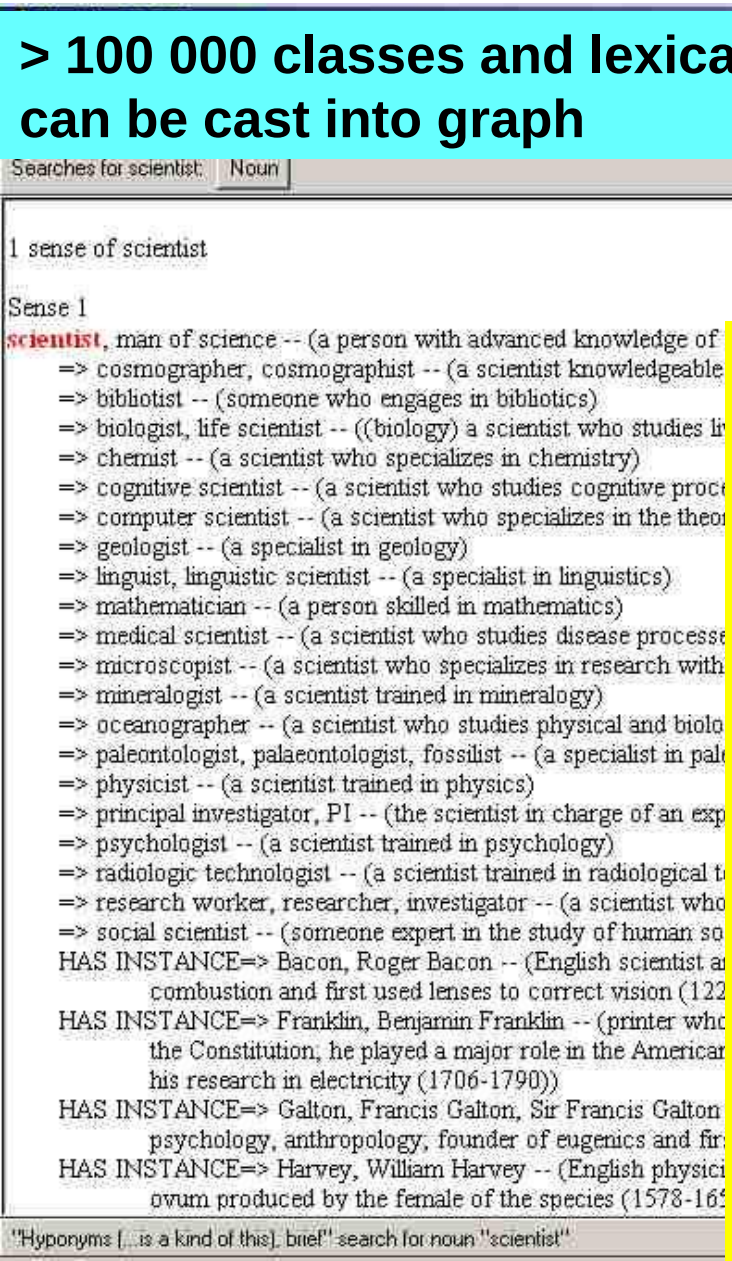
superclasses (hypernyms)

[Miller/Fellbaum 1998] http://wordnet.princeton.edu/

WordNet Thesaurus

> 100 000 classes and lexical relations; can be cast into graph

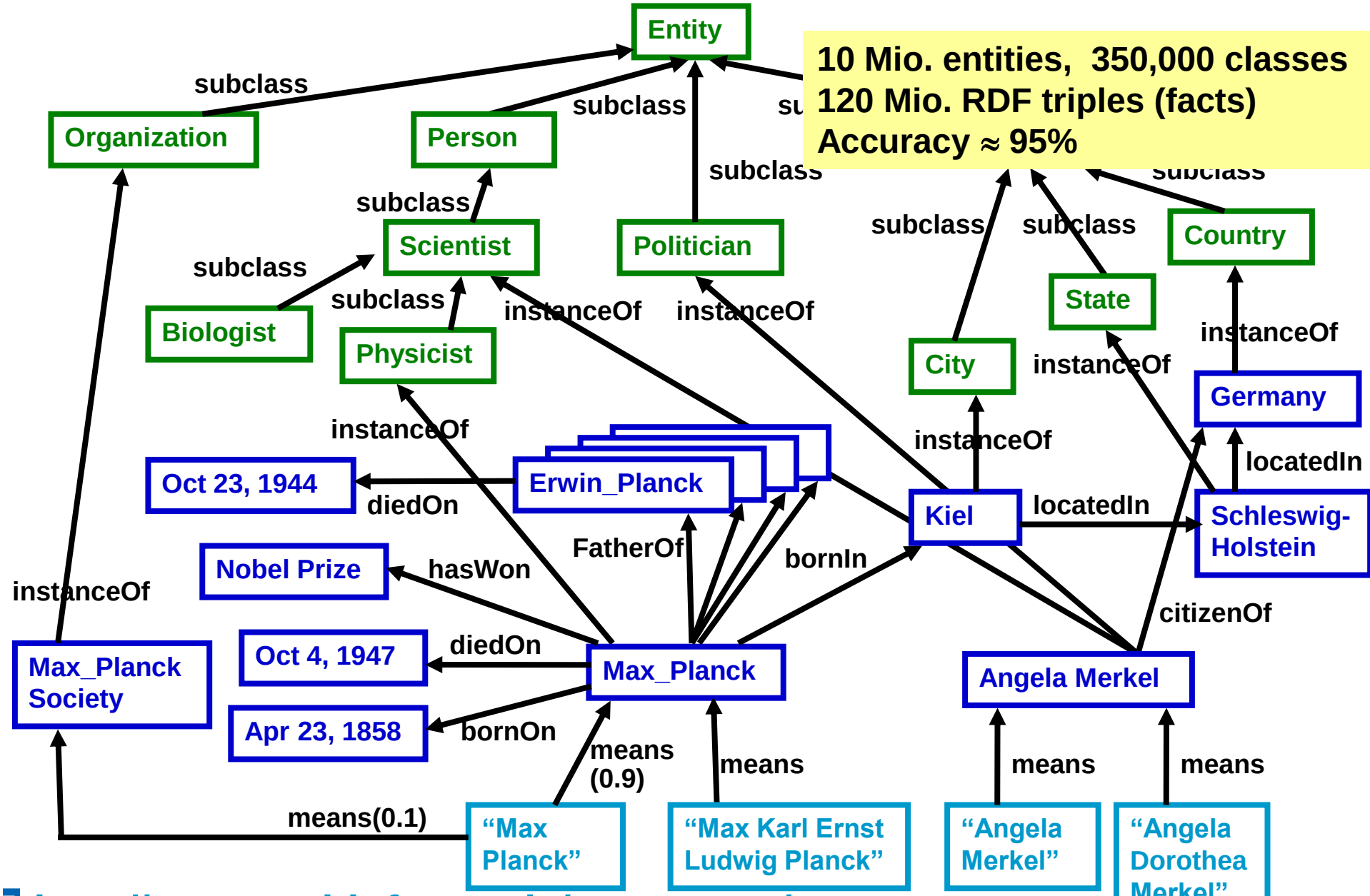
but:
only few individual entities (instances of classes)



scientist, man of science
(a person with advanced knowledge)
=> cosmographer, cosmographist
=> biologist, life scientist
=> chemist
=> cognitive scientist
=> computer scientist
…
=> principal investigator, PI
…
HAS INSTANCE => Bacon, Roger Bacon
…

http://wordnet.princeton.edu/

YAGO (Yet Another Great Ontology, Suchanek et al.: WWW'07)



10 Mio. entities, 350,000 classes
120 Mio. RDF triples (facts)
Accuracy ≈ 95%

http://www.mpi-inf.mpg.de/yago-naga/

DBpedia (Auer, Bizer, et al.: ISWC'07)

- 4.58 Mio. entities,
- 3 Bio. facts (RDF triples)
- 4.22 Mio. entities mapped to hand-crafted taxonomy of 685 classes with 2,795 properties

DBpedia Browse using Formats

About: Carlo Ancelotti

An Entity of Type : Sportmanager, from Named Graph : http://dbpedia.org, within Data Space : dbpedi

Carlo Ancelotti (* 10. Juni 1959 in Reggiolo) ist ein ehemaliger italienischer Fußballspieler und heutiger Fußballtrainer. Sowohl als Spieler als auch als Trainer gewann er zahlreiche nationale und internationale Titel, so u. a. zweimal den Europapokal der Landesmeister als Spieler und dreimal die Champions League als Trainer. Seit der Saison 2016/17 trainiert er den FC Bayern München.

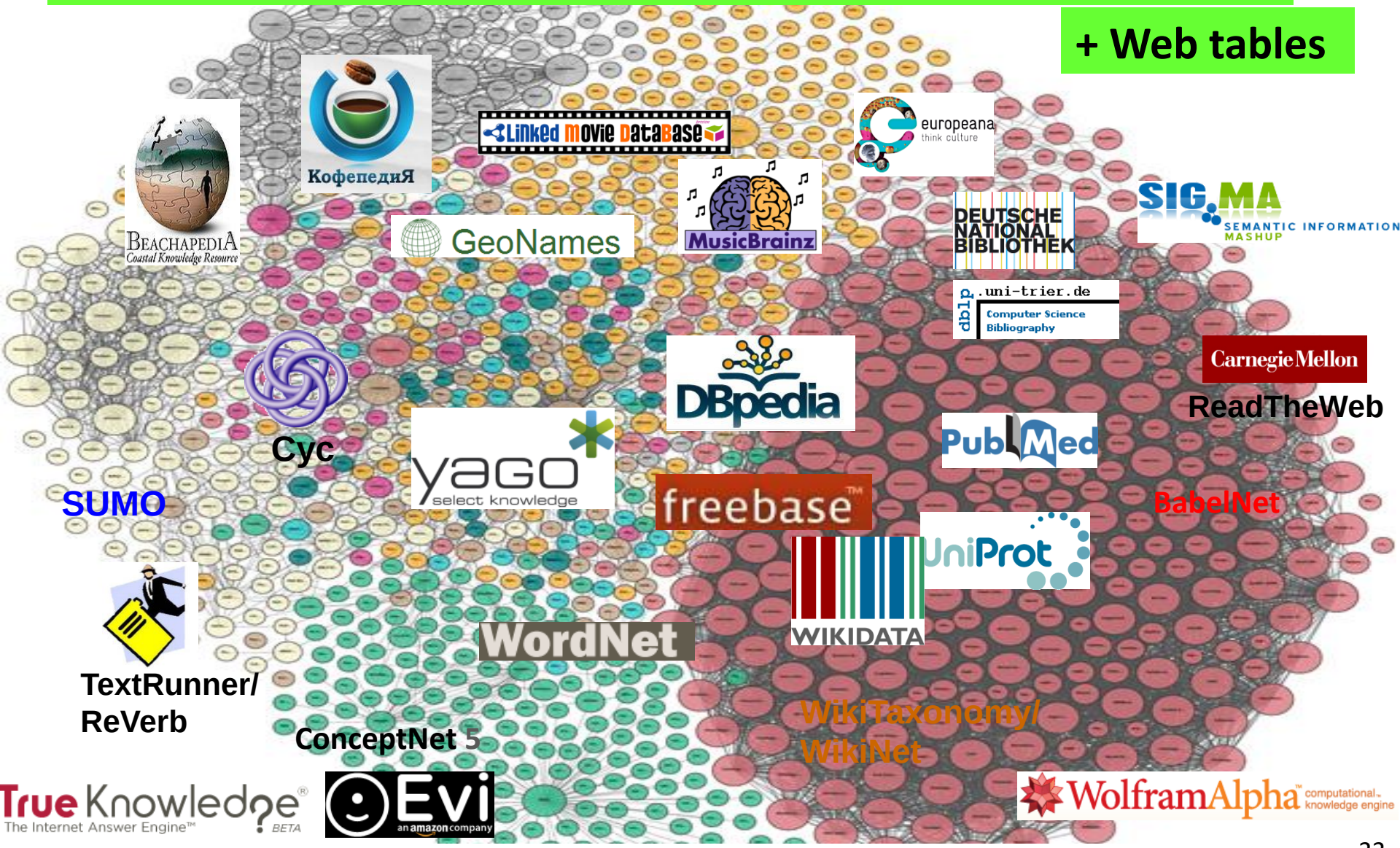
Property	Value
foaf:givenName	■ Carlo (en)
foaf:isPrimaryTopicOf	■ wikipedia-en:Carlo_Ancelotti
foaf:name	■ Carlo Ancelotti (en)
foaf:surname	■ Ancelotti (en)
is dbo:manager of	■ dbr:2009–10_Chelsea_F.C._season ■ dbr:1997–98_Parma_A.C._season ■ dbr:1996–97_Parma_A.C._season ■ dbr:1998–99_Juventus_F.C._season ■ dbr:1999–2000_Juventus_F.C._season ■ dbr:2002–03_A.C._Milan_season ■ dbr:2003–04_A.C._Milan_season ■ dbr:2005–06_A.C._Milan_season ■ dbr:2008–09_A.C._Milan_season ■ dbr:2006–07_A.C._Milan_season

http://www.dbpedia.org

Web of Data & Knowledge

> 150 Bio. subject-predicate-object triples from > 1000 sources

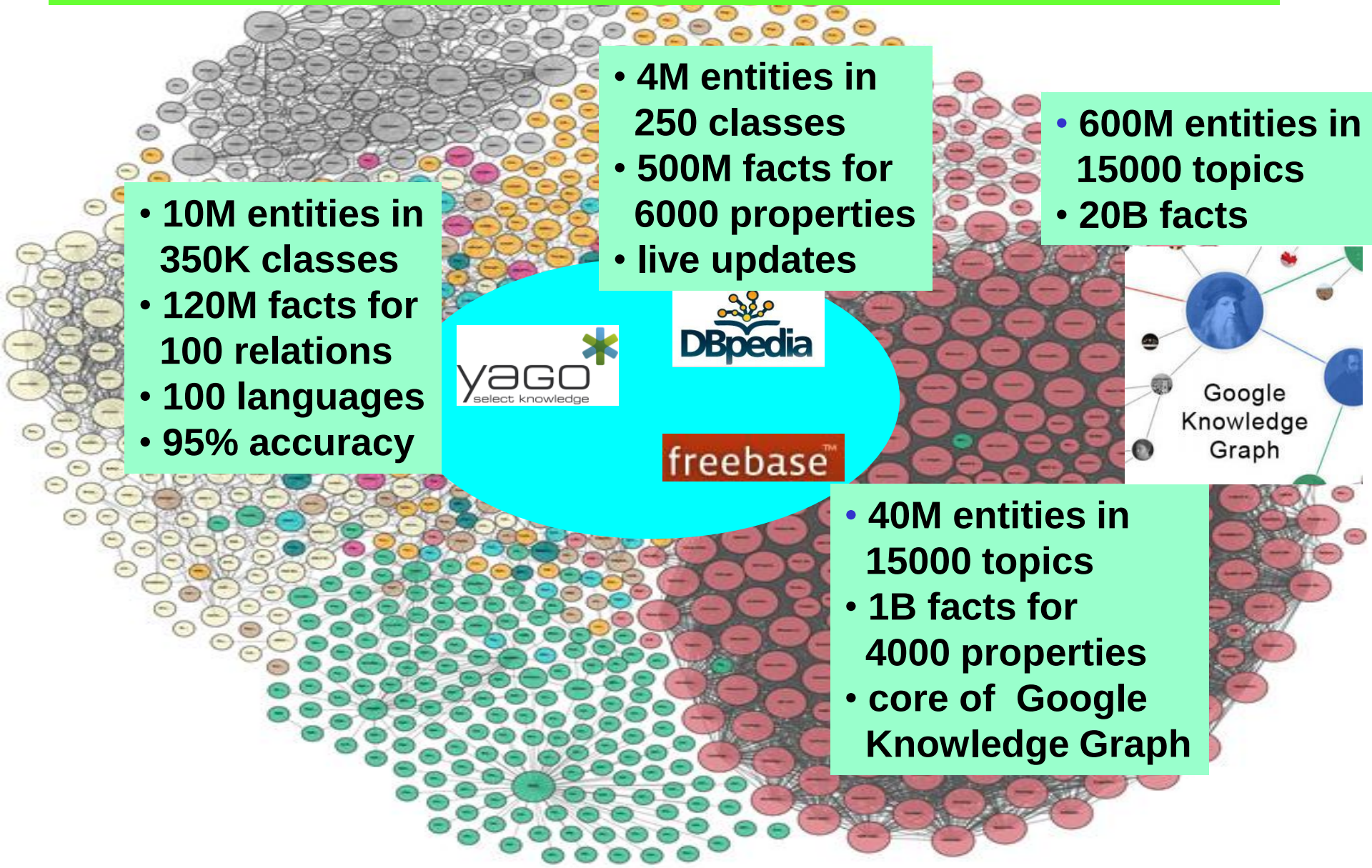
+ Web tables



Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

Web of Data & Knowledge

> 150 Bio. subject-predicate-object triples from > 1000 sources



- 10M entities in 350K classes
- 120M facts for 100 relations
- 100 languages
- 95% accuracy

- 4M entities in 250 classes
- 500M facts for 6000 properties
- live updates

yago
select knowledge

DBpedia

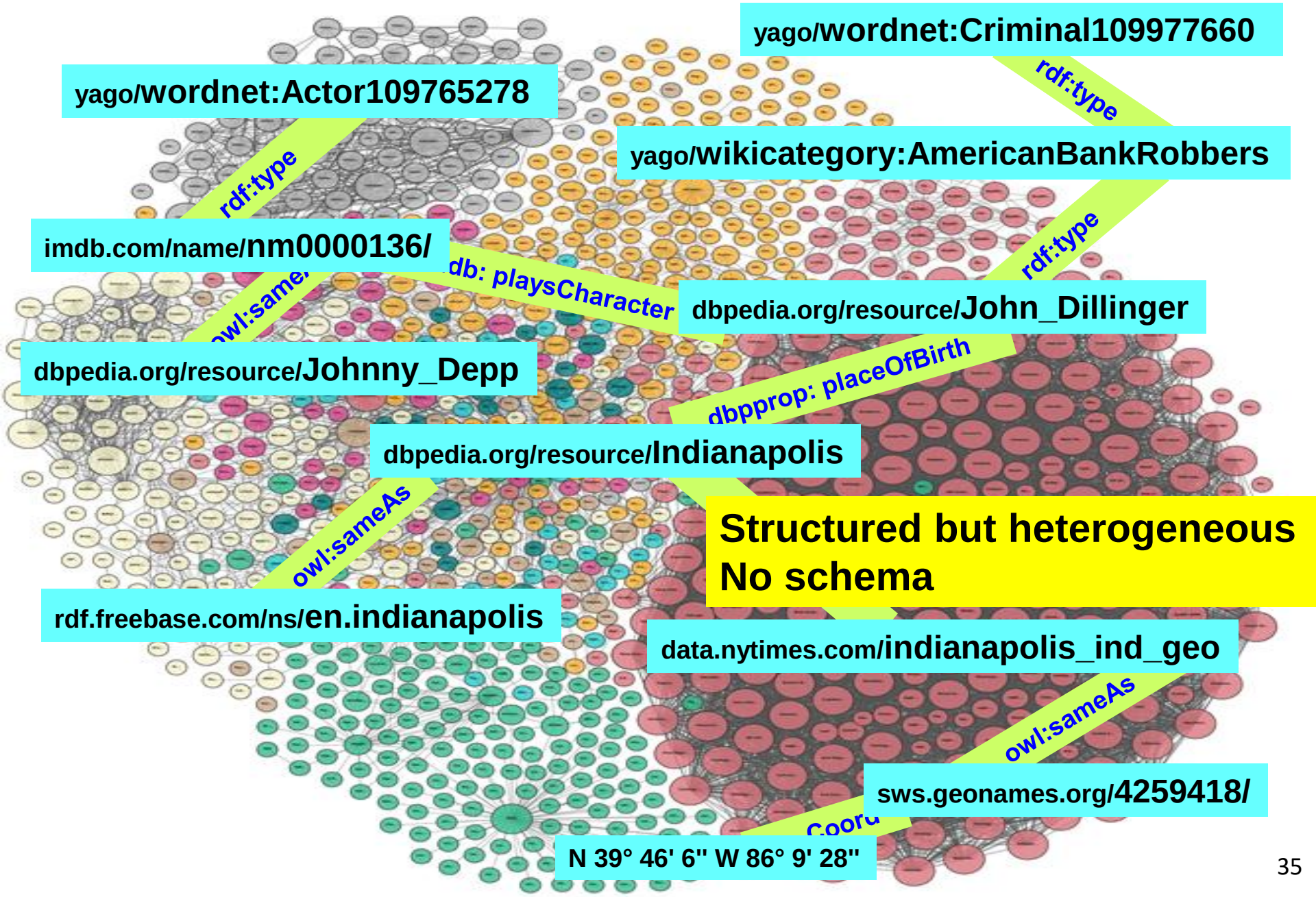
freebase™

- 600M entities in 15000 topics
- 20B facts

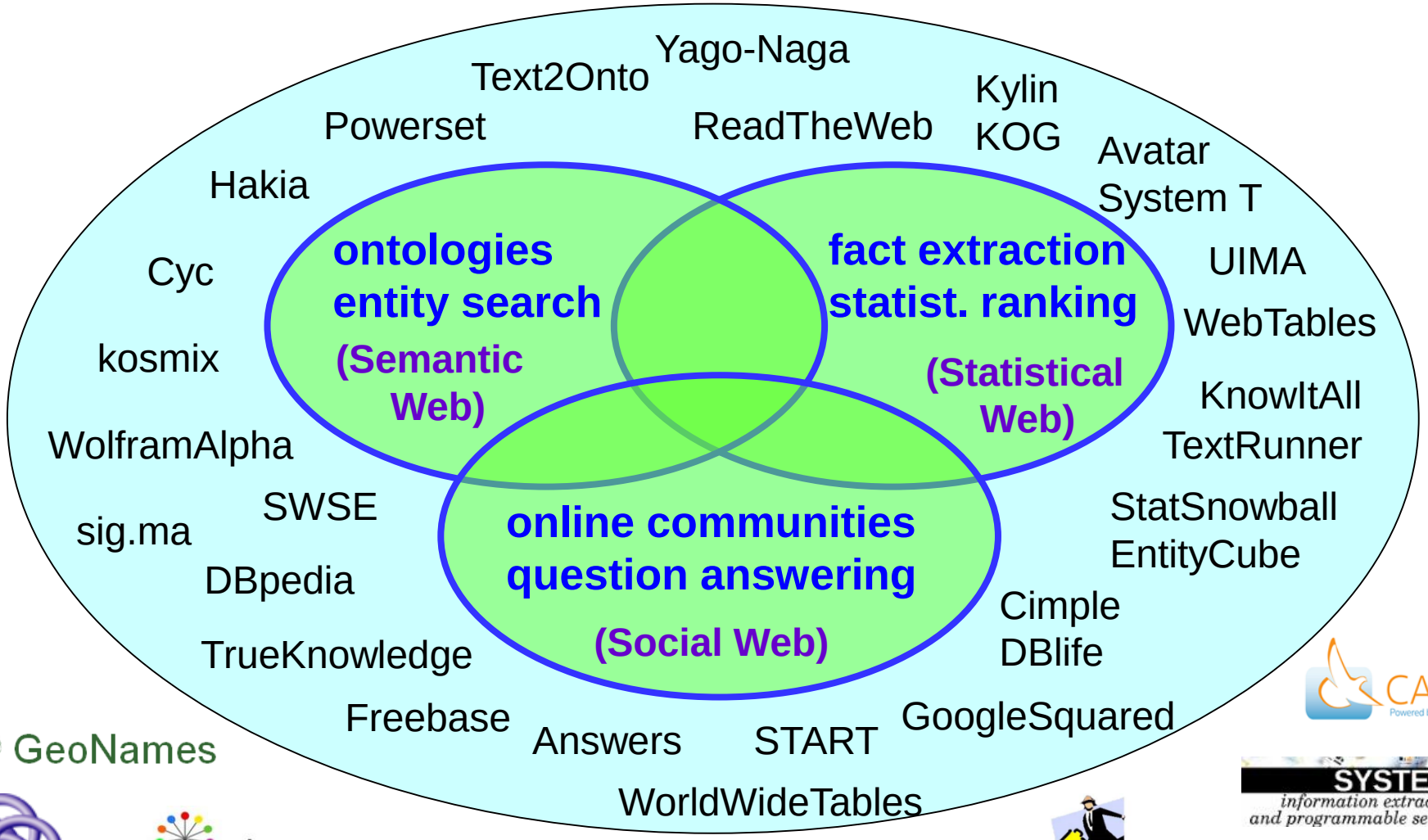
Google Knowledge Graph

- 40M entities in 15000 topics
- 1B facts for 4000 properties
- core of Google Knowledge Graph

Distributed Structure: Linking Open Data



Huge number of knowledge-enabled systems



GeoNames

CALAIS Powered by Thomson Reuters

Cyc

umbel

DBLife

DBpedia

YAGO-NAGA

IWP

TextRunner

SYSTEM information extraction and programmable search

SIG.MA SEMANTIC INFORMATION MASHUP

Carnegie Mellon

ReadTheWeb

WolframAlpha computational knowledge engine

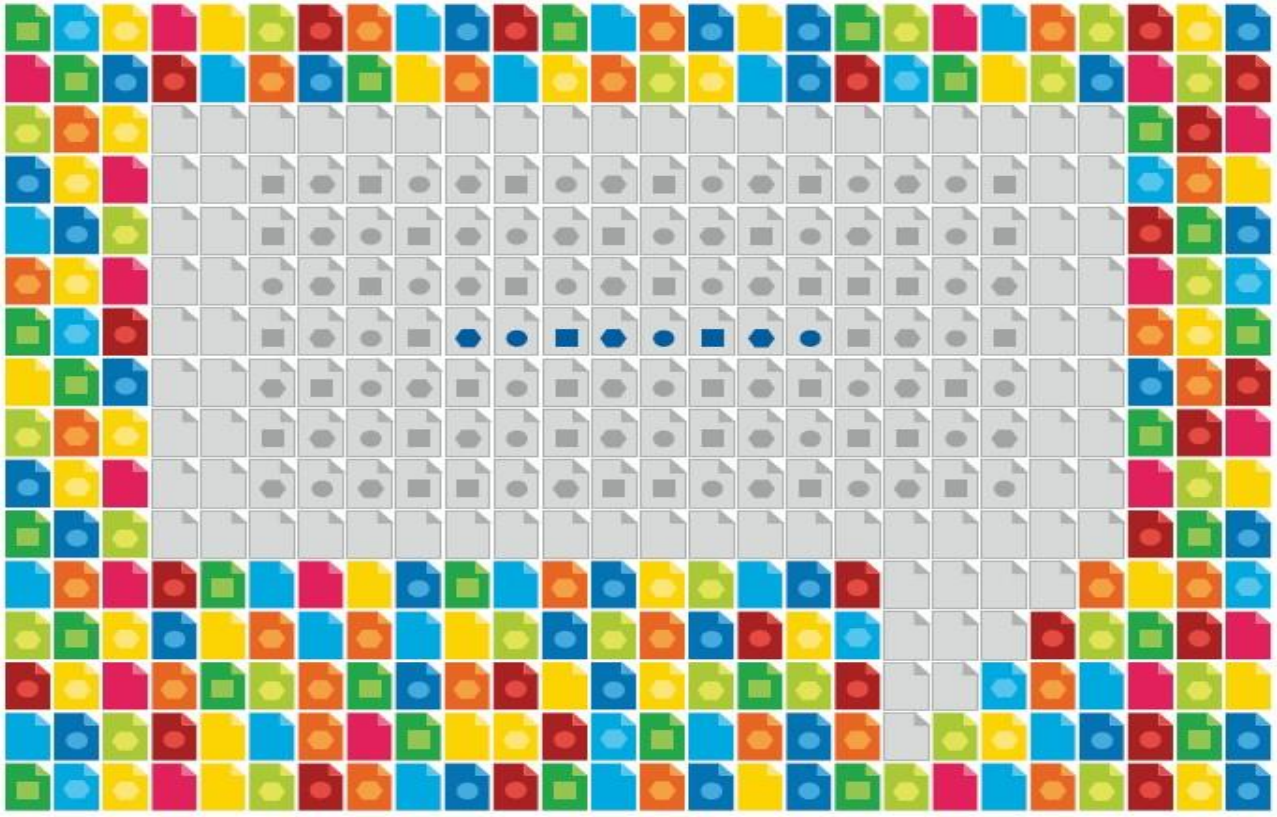
freebase

True Knowledge The Internet Answer Engine BETA

Outline

- Knowledge-based Search – What and Why
- Existing Knowledge Bases
- **Building and Maintaining Knowledge Bases**
  - **Entities & Classes**
  - Relationships
- Knowledge-Based Search

Goal: Turn Web into Knowledge Base

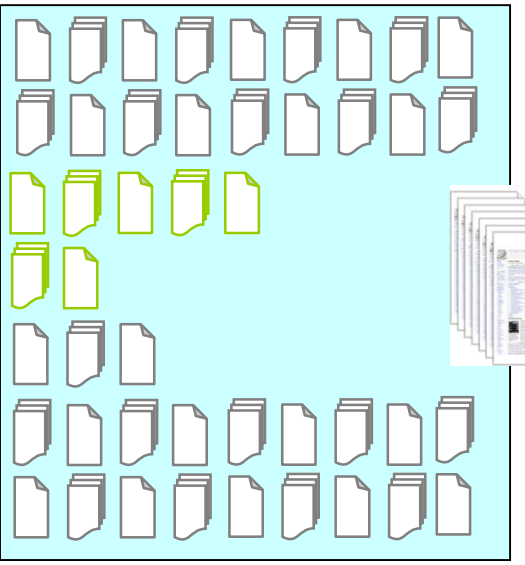


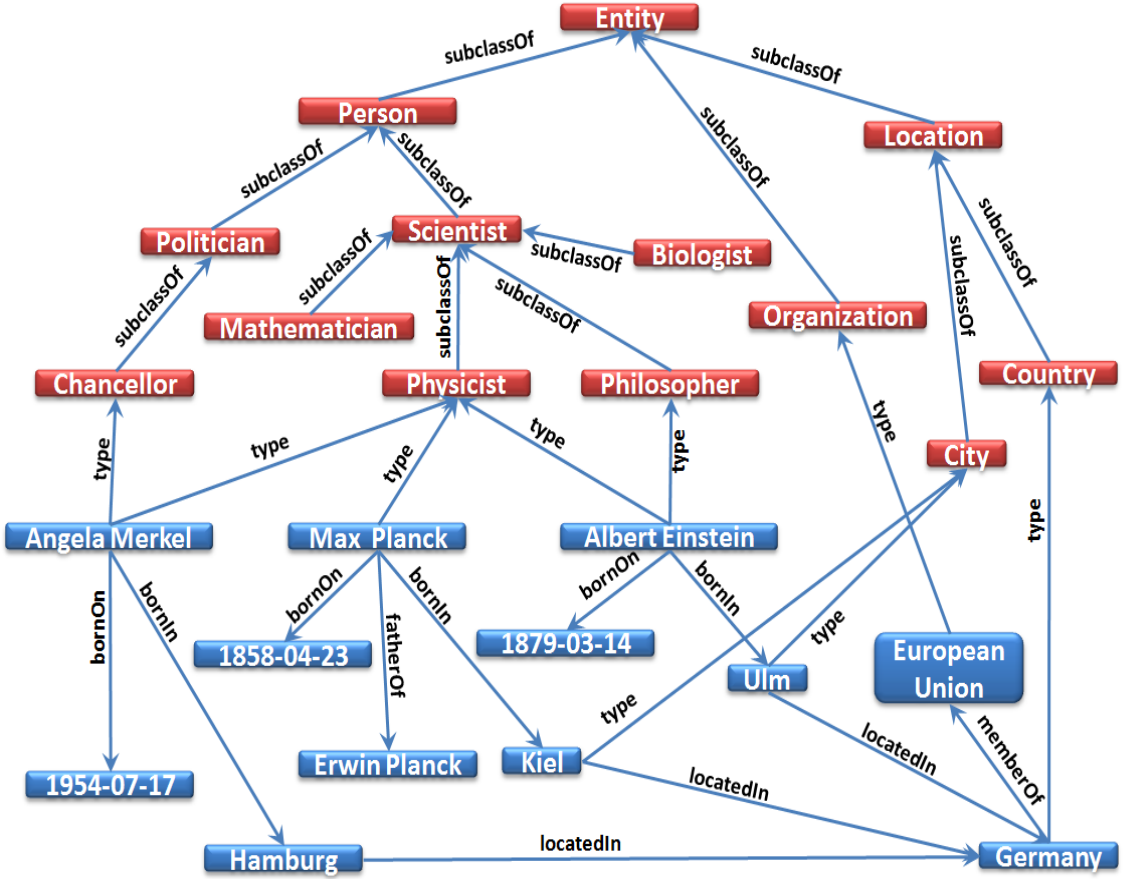
Source:
DB & IR methods for knowledge discovery. Communications of the ACM 52(4), 2009

comprehensive DB of human knowledge

- **everything that Wikipedia knows**
- **everything machine-readable**
- **capturing entities, classes, relationships**

Approach: Harvesting Facts from Web





Cyc

TextRunner

YAGO-NAGA

DBpedia

SIGMA SEMANTIC INFORMATION MASHUP

DBLife

IWP

umbel

True Knowledge The Internet Answer Engine BETA

Carnegie Mellon ReadTheWeb

Universität Trier

freebase

WolframAlpha computational knowledge engine

Tapping on Wikipedia Categories

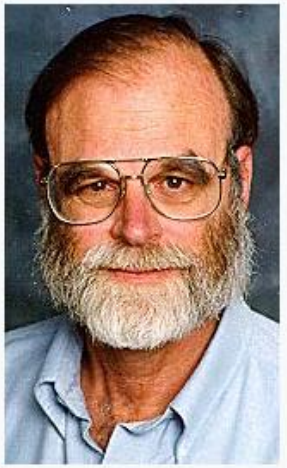
Jim Gray (computer scientist)

From Wikipedia, the free encyclopedia

James Nicholas "Jim" Gray (born 12 January 1944, lost at sea 28 January 2007) was an American computer scientist who received the Turing Award in 1998 "for seminal contributions to database and transaction processing research and technical leadership in system implementation."

Contents [hide]

James Nicholas "Jim" Gray



Born	January 12, 1944[1] San Francisco, California[2]
Died	**(lost at sea)** January 28, 2007
Nationality	American
Fields	Computer Science
Institutions	IBM, Tandem Computers, DEC, Microsoft
Alma mater	University of California, Berkeley
Doctoral advisor	Michael Harrison[2]
Known for	Work on database and transaction processing systems
Notable awards	Turing Award

Categories: Members of the National Academy of Sciences | American computer scientists | Fellows of the Association for Computing Machinery | Microsoft employees | DEC people | Database researchers | SIGMOD Edgar F. Codd Innovations Award winners | Turing Award laureates | 1944 births | 2007 deaths | People lost at sea | University of California, Berkeley alumni

Tapping on Wikipedia Categories

Max Planck

From Wikipedia, the free encyclopedia

"Planck" redirects here. For other uses, see Planck (disambiguation).

Max Planck (April 23, 1858 – October 4, 1947) was a German physicist. He is considered to be the founder of the quantum theory, and thus one of the most important physicists of the twentieth century. Planck was awarded the Nobel Prize in Physics in 1918.

Contents [hide]

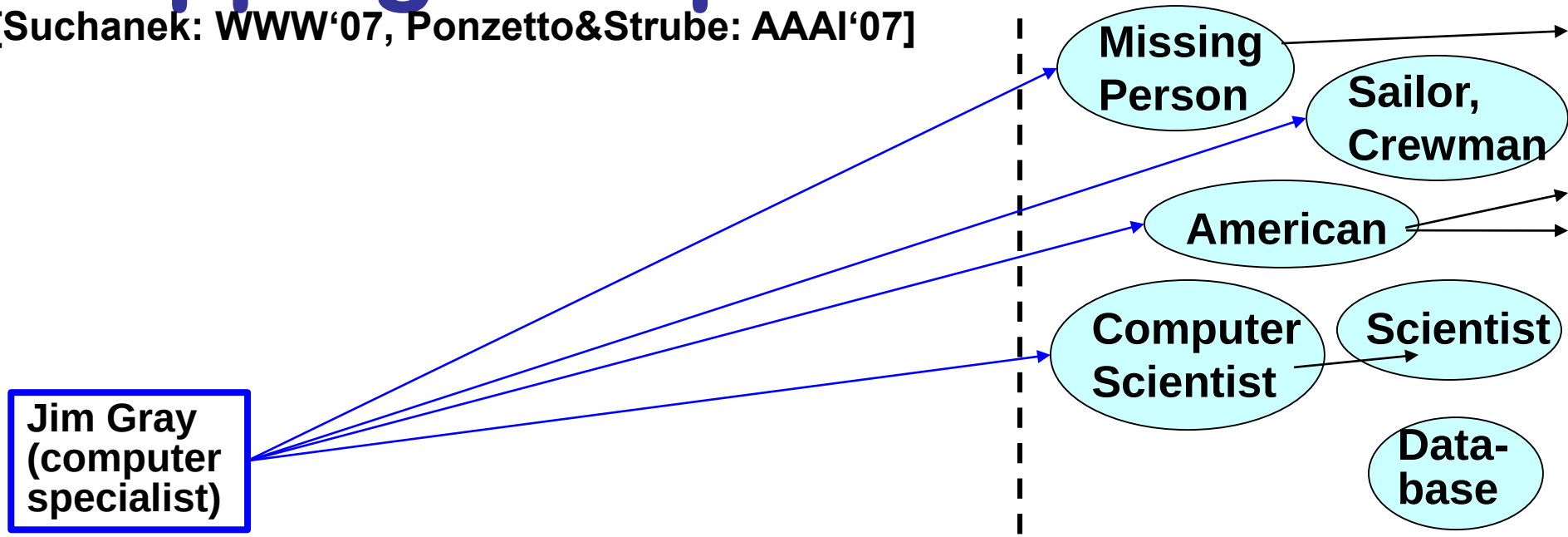
Max Planck

Born	April 23, 1858 Kiel, Holstein
Died	October 4, 1947 (aged 89) Göttingen, West Germany
Nationality	German
Fields	Physics
Institutions	University of Kiel University of Berlin University of Göttingen Kaiser-Wilhelm-Gesellschaft
Alma mater	Ludwig Maximilian University of

Categories: German Nobel laureates | German physicists | Members of the Pontifical Academy of Sciences | Members of the Prussian Academy of Sciences | Nobel laureates in Physics | Recipients of the Copley Medal | People from Kiel | People from the Province of Schleswig-Holstein | Quantum physicists | Recipients of the Pour le Mérite (civil class) | Theoretical physicists | Thermodynamicists | University of Munich alumni | University of Munich faculty | Humboldt University of Berlin alumni | Humboldt University of Berlin faculty | University of Kiel faculty | German Christians | Religion and science | Fellows of the Leopoldina | 1858 births | 1947 deaths

Mapping: Wikipedia → WordNet

[Suchanek: WWW'07, Ponzetto&Strube: AAAI'07]



Mapping: Wikipedia → WordNet

[Suchanek 2007, Ponzetto 2007]

Jim Gray (computer specialist)

instanceOf

People Lost at Sea

American Computer Scientists

Database Researcher

Fellows of the ACM

Computer Scientists by Nation

Databases

Engineering Societies

ACM

Members of Learned Societies

?

?

?

?

Missing Person

American

Computer Scientist

subclassOf

Scientist

?

Data-base

Fellow (1), Comrade

?

Fellow (2), Colleague

Fellow (3) (of Society)

?

Member (1), Fellow

Member (2), Extremity

name similarity
(edit dist., n-gram overlap) ?

context similarity
(word/phrase level) ?

machine learning ?

Mapping: Wikipedia → WordNet

[Suchanek 2007, Ponzetto 2007]

Given: entity **e** in Wikipedia categories $c_1, \ldots, c_k$

Wanted: **instanceOf(e,c)** and **subclassOf(c_i,c)** for WN class **c**

Problem: vagueness & ambiguity of **names** $c_1, \ldots, c_k$

Analyzing category names → noun group parser:

American Musicians of Italian Descent

pre-modifier head post-modifier

American Folk Music of the 20th Century

pre-modifier head post-modifier

American Indy 500 Drivers on Pole Positions

pre-modifier head post-modifier

Head word is key, should be in plural for instanceOf

Mapping Wikipedia Entities to WordNet Classes

[Suchanek 2007, Ponzetto 2007]

Given: entity e in Wikipedia categories $c_1, \ldots, c_k$

Wanted: instanceOf(e,c) and subclassOf(c_i,c) for WN class c

Problem: vagueness & ambiguity of names $c_1, \ldots, c_k$

Heuristic Method:

for each c_i do

if head word w of category name c_i is plural

{

1) match w against synsets of WordNet classes

2) choose best fitting class c and set $e \in c$

3) expand w by pre-modifier and set $c_i \subseteq w^+ \subseteq c$

}

tuned conservatively: high precision, reduced recall

Long Tail of Class Instances

Automatically create sets of items from a few examples.

Enter a few items from a set of things. (example)
Next, press *Large Set* or *Small Set* and we'll try to predict other items in the set.

- Stanford
- Princeton
- Penn State
-
-

Classic IR problem: Entity List Completion (TREC, INEX, …)

Long Tail of Class Instances

Predicted Items
penn state
stanford
princeton
ucla
harvard
mit
usc
yale
columbia
cornell
berkeley
duke
georgetown
michigan
arizona
washington
dartmouth
oregon
nyu
california
brown
chicago
northwestern
caltech
virginia
penn

Long Tail of Class Instances

[Etzioni 2004, Wang 2008, Carlson 2010, He 2011]

State-of-the-Art Approach (e.g. SEAL):

- Start with seeds: a few class instances
- Find lists, tables, text snippets ("for example", „such as"), … that contain one or more seeds
- Extract candidates: noun phrases from vicinity
- Gather co-occurrence stats (seed&cand, cand&className pairs)
- Rank candidates
  - point-wise mutual information, …
  - random walk (PR-style) on seed-cand graph

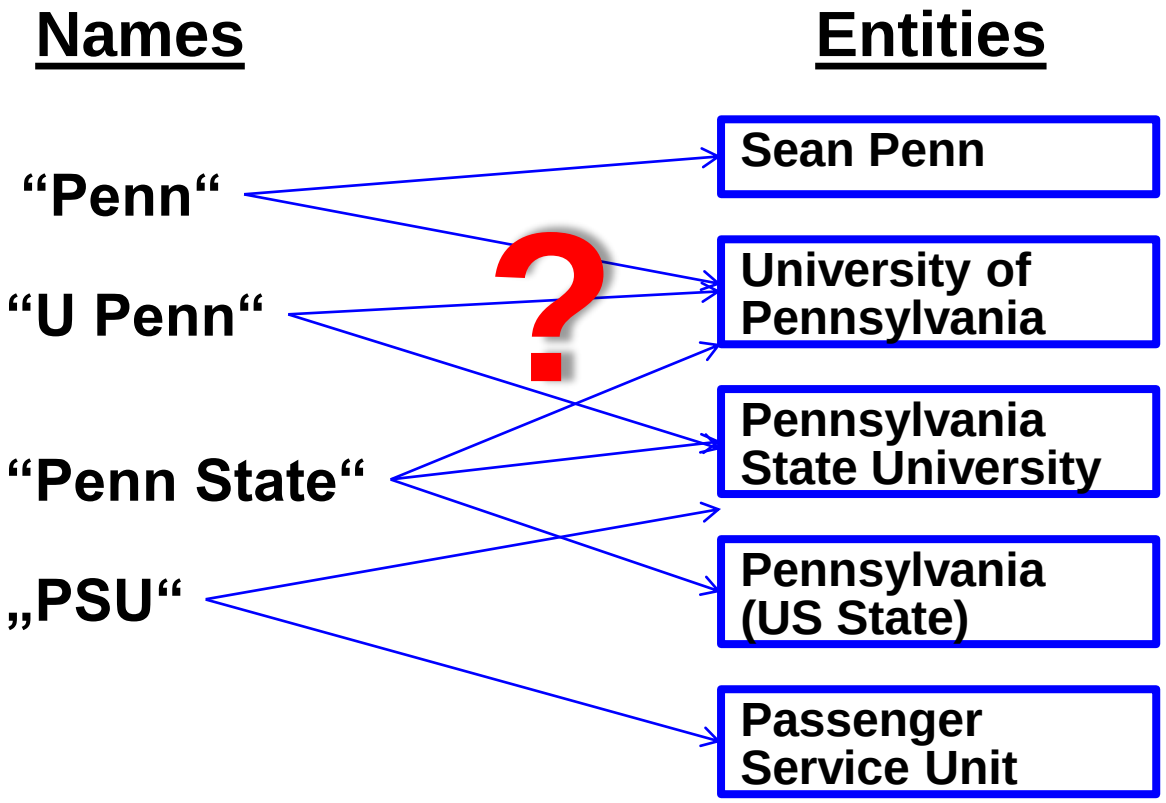
But:

Precision drops for classes with sparse statistics (DB profs, …)

Harvested items are names, not entities

Canonicalization (de-duplication) unsolved

Individual Entity Disambiguation

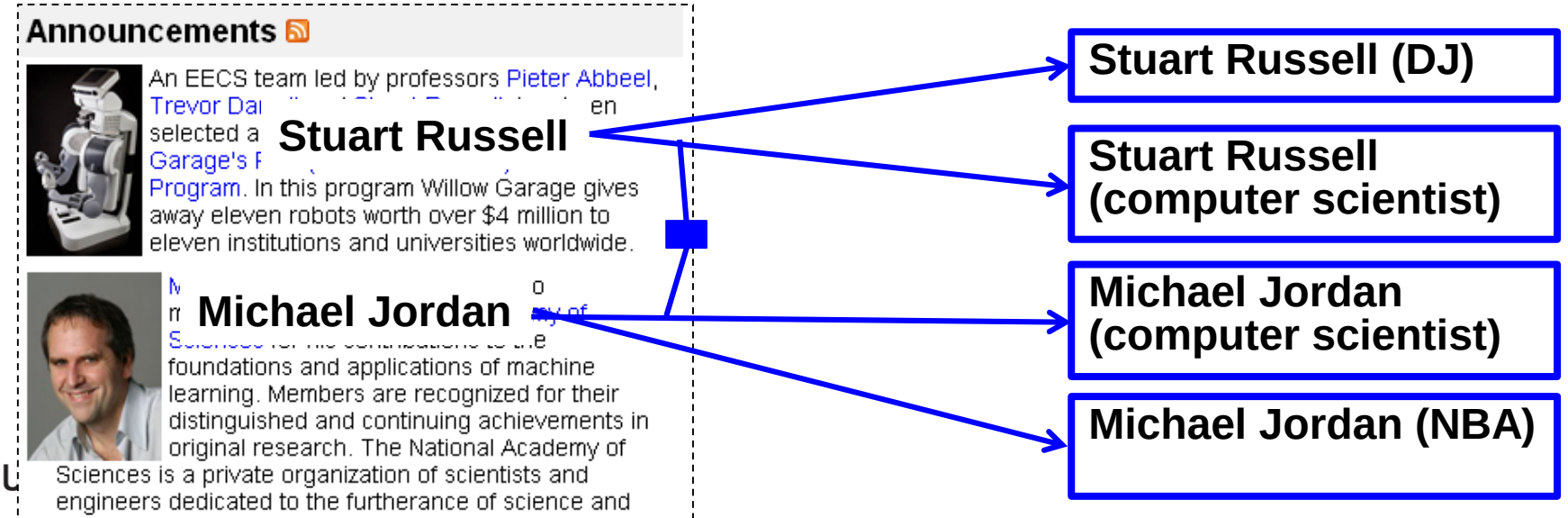


- ill-defined with zero context
- known as record linkage for names in record fields
- Wikipedia offers rich candidate mappings: disambiguation pages, re-directs, inter-wiki links, anchor texts of href links

Collective Entity Disambiguation

[McCallum 2003, Madhavan 2005, Bhattacharya 2007, Poon 2007, Kulkarni 2009, Hoffart 2011]

- Consider a set of names $\{n_1, n_2, \ldots\}$ in same context and sets of candidate entities $E1 = \{e_{11}, e_{12}, \ldots\}$, $E2 = \{e_{21}, e_{22}, \ldots\}$, …
- Define joint objective function (e.g. likelihood for prob. model) that rewards coherence of mappings $n_i \rightarrow e_{ij}$
- Solve optimization problem



Outline of the Talk

- Knowledge-based Search: What and Why
- **Building Knowledge Bases**
  - Entities & Classes
  - **Relationships**
- Searching Knowledge Bases

Picking Low-Hanging Fruit (First)

Héctor García-Molina

Born	Monterrey, Nuevo León, Mexico
Residence	United States
Nationality	Mexican
Fields	Computer Science
Institutions	Stanford University
Alma mater	ITESM
Doctoral advisor	Gio Wiederhold [1]
Doctoral students	Robert Abbott, Boris Kogan, Narayanan Shivakumar
Known for	Distributed databases
Notable awards	1999 ACM SIGMOD Edgar F. Codd Innovations Award

Barbara Liskov

Born	1939 (age 70–71)
Nationality	American
Fields	Computer Science
Institutions	Massachusetts Institute of Technology
Alma mater	University of California, Berkeley Stanford University

Kurt Mehlhorn

Born	August 27, 1949 (age 60)[1] Ingolstadt, Germany[1]
Residence	Germany
Fields	Computer Science
Institutions	Universität des Saarlandes
Alma mater	Ph.D. Cornell University, 1974[1]
Doctoral advisor	Robert Constable[2]
Known for	LEDA
Notable awards	Leibniz Prize, Konrad Zuse Medal, EATCS Award

Joseph M. Hellerstein

Fields	Computer Science
Institutions	University of California, Berkeley
Alma mater	University of Wisconsin–Madison
Doctoral advisor	Jeffrey Naughton, Michael Stonebraker

Jeffrey Ullman

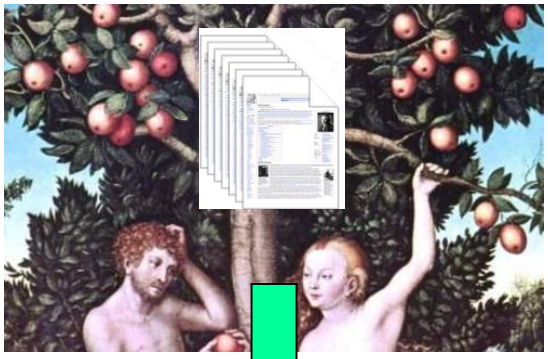
Born	November 22, 1942 (age 67)
Citizenship	American
Nationality	American
Alma mater	Columbia University, Princeton University
Doctoral advisor	Arthur Bernstein, Archie McKellar
Doctoral students	Alexander Birman, Surajit Chaudhuri, Evan Cohn, Alan Demers, Marcia Derr, Nahed El Djabri, Amelia Fong Lochovsky, Deepak Goyal, Ashish Gupta, Himanshu Gupta, Udaiprakash Gupta, Venkatesh Harinarayan, Taher Haveliwala, Matthew Hecht, Daniel Hirschberg, Peter Hochschild, Peter Honeyman, Edward Horvath, Gregory Hunter, Nam (Pierre) Huyn, Hakan Jakobsson, John Kam, Marc Kaplan, Anna Karlin, Kevin Karplus, Henry Korth,

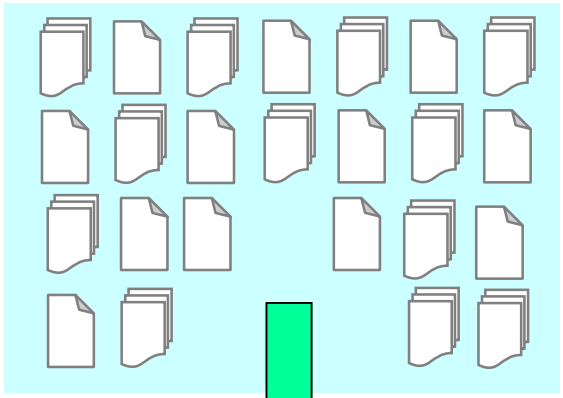
Deterministic Pattern Matching

[Kushmerick 1997, Califf 1999, Baumgartner 2001, …]

- **Regular expressions matching**
- **Wrapper induction (grammar learning for restricted regular languages)**
- **Well understood**

French Marriage Problem





facts in KB:

married
(Hillary, Bill)
married
(Carla, Nicolas)
married
(Angelina, Brad)

new facts or fact candidates:

married (Cecilia, Nicolas)
married (Carla, Benjamin)
married (Carla, Mick)
married (Michelle, Barack)
married (Yoko, John)
married (Kate, Leonardo)
married (Carla, Sofie)
married (Larry, Google)

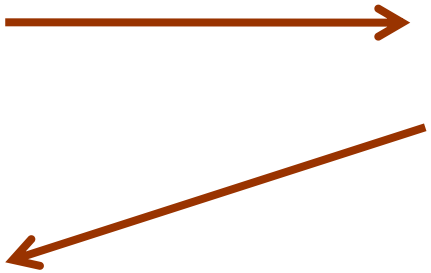
1) for recall: pattern-based harvesting
2) for precision: consistency reasoning

Pattern-Based Harvesting

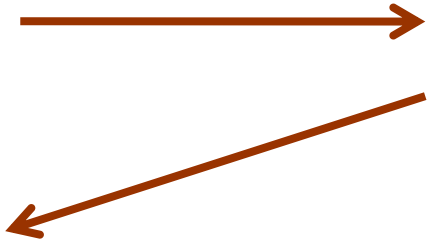
(Hearst 1992, Brin 1998, Agichtein 2000, Etzioni 2004, …)

Facts & *Fact Candidates* | **Patterns**

(Hillary, Bill)
(Carla, Nicolas)



(Angelina, Brad)
(Victoria, David)
(Hillary, Bill)
(Carla, Nicolas)



(Yoko, John)
(Kate, Pete)
(Carla, Benjamin)
(Larry, Google)
(Angelina, Brad)
(Victoria, David)

X and her husband Y
X and Y on their honeymoon
X and Y and their children
X has been dating with Y
X loves Y
…

- **good for recall**
- **noisy, drifting**
- **not robust enough for high precision**

Statistics yield pattern assessment

Support of pattern p:

$$\frac{\text{\# occurrences of p with seeds}}{\text{\# occurrences of all patterns with seeds}}$$

Confidence of pattern p:

$$\frac{\text{\# occurrences of p with seeds}}{\text{\# occurrences of p}}$$

Confidence of fact candidate (e1,e2):

$$\Sigma_p \text{ freq(e1,p,e2)*conf(p)} \;/\; \Sigma_p \text{ freq(e1,p,e2)}$$

$$\text{or: PMI (e1,e2)} = \log \frac{\text{freq(e1,e2)}}{\text{freq(e1) freq(e2)}}$$

- gathering can be iterated,
- can promote best facts to additional seeds for next round

Negative Seeds increase precision

(Ravichandran 2002; Suchanek 2006; ...)

Problem: Some patterns have high support, but poor precision:

X is the largest city of Y — for isCapitalOf (X,Y)

joint work of X and Y — for hasAdvisor (X,Y)

Idea: Use positive and negative seeds:

pos. seeds: (Paris, France), (Rome, Italy), (New Delhi, India), ...

neg. seeds: (Sydney, Australia), (Istanbul, Turkey), ...

Compute the confidence of a pattern as:

$$\frac{\text{\# occurrences of p with pos. seeds}}{\text{\# occurrences of p with pos. seeds or neg. seeds}}$$

- can promote best facts to additional seeds for next round
- can promote rejected facts to additional counter-seeds
- works more robustly with few seeds & counter-seeds

Generalized patterns increase recall

(N. Nakashole 2011)

Problem: Some patterns are too narrow and thus have small recall:

X and his celebrated advisor Y
X carried out his doctoral research in math under the supervision of Y
X received his PhD degree in the CS dept at Y
X obtained his PhD degree in math at Y

Idea: generalize patterns to n-grams, allow POS tags

X { his doctoral research, under the supervision of} Y
X { PRP ADJ advisor } Y
X { PRP doctoral research, IN DET supervision of} Y

Compute n-gram-sets by frequent sequence mining

Compute match quality of pattern p with sentence q by Jaccard:

$$\frac{|\{\text{n-grams} \in p\} \cap \{\text{n-grams} \in q\}|}{|\{\text{n-grams} \in p\} \cup \{\text{n-grams} \in q\}|}$$

=> Covers more sentences, increases recall

Deep Parsing makes patterns robust

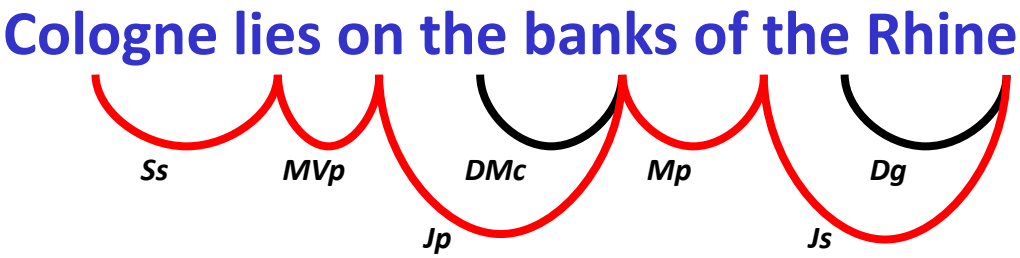
(Bunescu 2005 , Suchanek 2006, …)

Problem: Surface patterns fail if the text shows variations

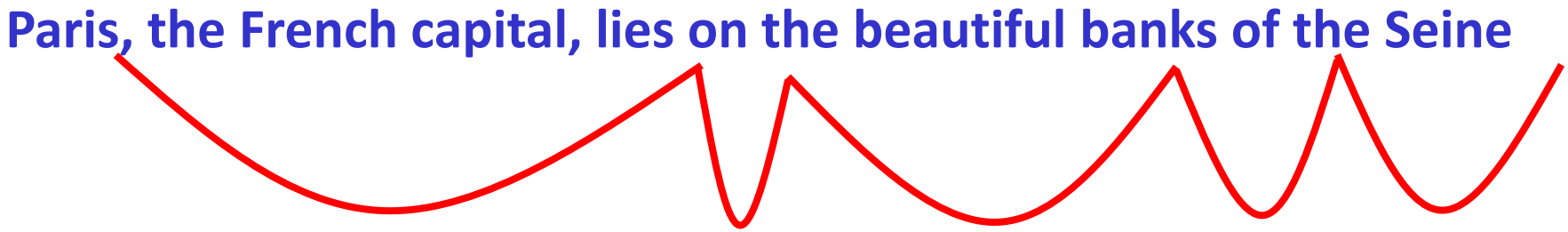
Cologne lies on the banks of the Rhine.

Paris, the French capital, lies on the beautiful banks of the Seine.

Idea: Use deep linguistic parsing to define patterns



Deep linguistic patterns work even on sentences with variations



Reasoning about Fact Candidates

Use consistency constraints to prune false candidates

FOL rules (restricted):

$spouse(x,y) \wedge diff(y,z) \Rightarrow \neg spouse(x,z)$

$spouse(x,y) \wedge diff(x,w) \Rightarrow \neg spouse(w,y)$

$spouse(x,y) \Rightarrow f(x)$ $\quad spouse(x,y) \Rightarrow m(y)$

$spouse(x,y) \Rightarrow (f(x) \wedge m(y)) \vee (m(x) \wedge f(y))$

ground atoms:

spouse(Hillary,Bill)
spouse(Carla,Nicolas)
spouse(Cecilia,Nicolas)
spouse(Carla,Ben)
spouse(Carla,Mick)
spouse(Carla, Sofie)

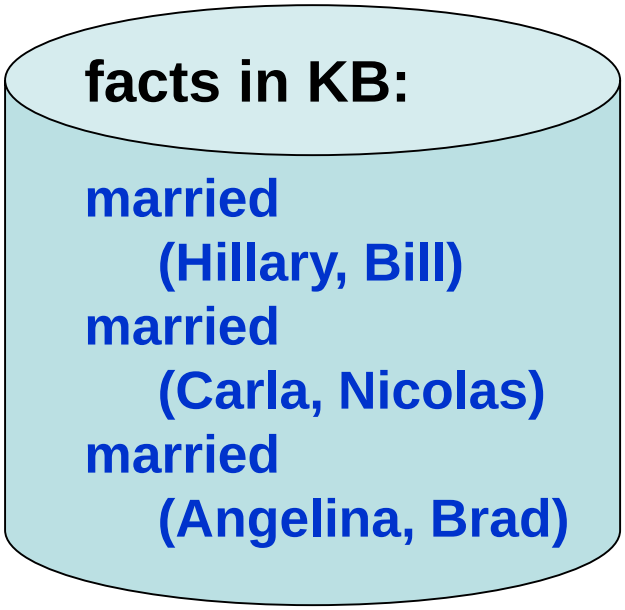
f(Hillary)	m(Bill)
f(Carla)	m(Nicolas)
f(Cecilia)	m(Ben)
f(Sofie)	m(Mick)

Rules reveal inconsistencies
Find consistent subset(s) of atoms
("possible world(s)", "the truth")

Rules can be weighted
(e.g. by fraction of ground atoms that satisfy a rule)
→ uncertain / probabilistic data
→ compute prob. distr. of subset of atoms being the truth

Reasoning for KB Growth

(Suchanek 2009)



+

new fact candidates:

- married (Cecilia, Nicolas)
- married (Carla, Benjamin)
- married (Carla, Mick)
- married (Carla, Sofie)
- married (Larry, Google)

?

patterns:

- X and her husband Y
- X and Y and their children
- X has been dating with Y
- X loves Y

Direct approach:

1. facts are true; fact candidates & patterns → hypotheses
   grounded constraints → clauses with hypotheses as vars
2. type signatures of relations greatly reduce #clauses
3. cast into Weighted Max-Sat with weights from pattern stats
   customized approximation algorithm

unifies: fact cand consistency, pattern goodness, entity disambig.

Fact & Pattern Consistency with SOFI

(Suchanek 2009, Nakashole 2011)

constraints to connect facts, fact candidates, patterns

pattern-fact duality:

$\text{occurs}(p,x,y) \wedge \text{expresses}(p,R) \wedge \text{type}(x)=\text{dom}(R) \wedge \text{type}(y)=\text{rng}(R) \Rightarrow R(x,y)$
$\text{occurs}(p,x,y) \wedge R(x,y) \wedge \text{type}(x)=\text{dom}(R) \wedge \text{type}(y)=\text{rng}(R) \Rightarrow \text{expresses}(p,R)$

name(-in-context)-to-entity mapping:

$\neg\, \text{means}(n,e1) \vee \neg\, \text{means}(n,e2) \vee \ \ldots$

functional dependencies:

spouse(X,Y): $X \rightarrow Y$, $Y \rightarrow X$

relation properties:

asymmetry, transitivity, acyclicity, …

type constraints, inclusion dependencies:

$\text{spouse} \subseteq \text{Person} \times \text{Person}$

$\text{capitalOfCountry} \subseteq \text{cityOfCountry}$

domain-specific constraints:

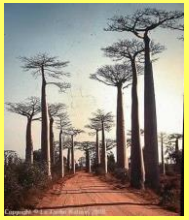
$\text{bornInYear}(x) + 10\text{years} \leq \text{graduatedInYear}(x)$

$\text{hasAdvisor}(x,y) \wedge \text{graduatedInYear}(x,t) \wedge \text{graduatedInYear}(y,s) \Rightarrow s < t$

Practicality at Web Scale ?

Need to tackle long tail of entities beyond Wikipedia

Need robust & efficient name-entity disambiguation



Pattern-based gathering:
suitable for (Map-Reduce) distribution & parallelism

Consistency reasoning:
difficult to parallelize w/o sacrificing quality

Open-domain harvesting of new relation types

Life-cycle of KB growth & maintenance

Outline

- Knowledge-based Search – What and Why
- Existing Knowledge Bases
- Building and Maintaining Knowledge Bases
- **Knowledge-Based Search**

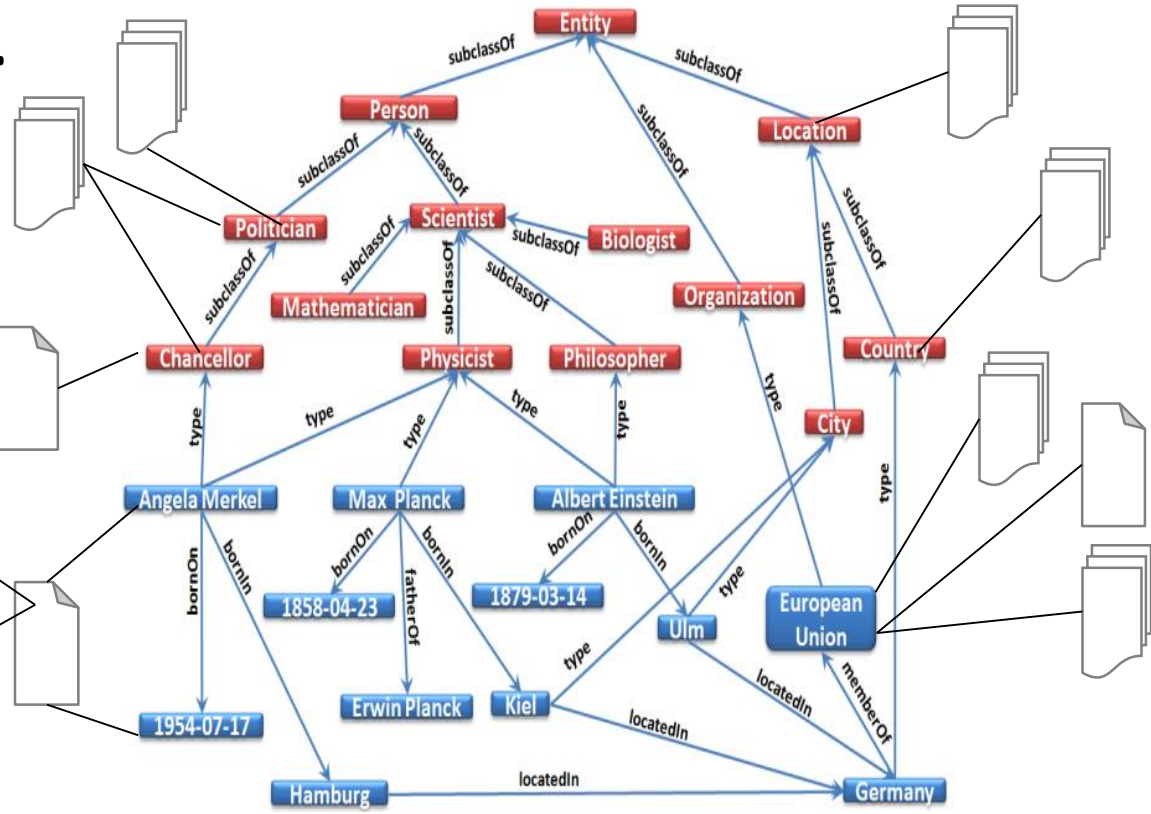
Variants of Search

	Query	Results
Document Search	Keywords	Documents
Entity Search	Keywords	Entities
Relationship Search	Set of entities	Connecting facts
Example-Based Search	Entities	Entities
Structured Search	Structured Query	Entity Sets

Extending Entities with Keywords

Remember: entities **occur in facts in documents**

⇒Associate entities with terms in those documents, keywords in facts connected to the entity, keywords in URIs, … (**context** of entity)



chancellor
Germany
scientist election
Stuttgart21

Structured Queries: SPARQL

SPJ combinations of **triple patterns**
(triples with S,P,O replaced by variable(s))

```
Select ?p, ?c Where {
?p instanceOf Composer .
?p bornIn ?t . ?t inCountry ?c . ?c locatedIn Europe .
?p hasWon ?a .?a Name AcademyAward . }
```

Semantics:
return all bindings to variables that match all triple patterns
(subgraphs in RDF graph that are isomorphic to query graph)

+ filter predicates, duplicate handling, RDFS types, etc.

```
Select Distinct ?c Where {
?p instanceOf Composer .
?p bornIn ?t . ?t inCountry ?c . ?c locatedIn Europe .
?p hasWon ?a .?a Name ?n .
?p bornOn ?b .  Filter (?b > 1945)  . Filter(regex(?n, "Academy") . }
```

Querying the Structured Web

Structure but no schema: SPARQL well suited

flexible subgraph matching

wildcards for properties (relaxed joins):

```
Select ?p, ?c Where {
?p instanceOf Composer .
?p ?r1 ?t .  ?t ?r2 ?c .  ?c isa Country . ?c locatedIn Europe .  }
```

SPARQL 1.1: regular expressions

```
Select ?p, ?c Where {
?p instanceOf Composer .
?p (bornIn | livesIn | citizenOf) / locatedIn* Europe .  }
```

also reverse edges, optional elements, …

Why ranking is essential

- Queries often have a huge number of results:
  - scientists from Germany
  - seminars in Dagstuhl
  - publications in information retrieval
  - actors from the U.S.
- Ranking as integral part of search
- Huge number of app-specific ranking methods: paper/citation count, impact, salary, …
- Need for generic ranking

Ranking Criteria

Confidence:

Prefer results that are likely correct

- accuracy of info extraction
- trust in sources (authenticity, authority)

bornIn (Jim Gray, San Francisco) from „Jim Gray was born in San Francisco“ (en.wikipedia.org)

livesIn (Michael Jackson, Tibet) from „Fans believe Jacko hides in Tibet“ (www.michaeljacksonsightings.com)

Informativeness:

Prefer results with salient facts

Statistical LM with estimations from:

- frequency in answer
- frequency in corpus (e.g. Web)
- frequency in query log

q: Einstein isa ?

Einstein isa scientist

Einstein isa vegetarian

q: ?x isa vegetarian

Einstein isa vegetarian

Whocares isa vegetarian

Diversity:

Prefer variety of facts

E won … E discovered … E played …

E won … E won … E won … E won …

Conciseness:

Prefer results that are tightly connected

- size of answer graph
- cost of Steiner tree

Einstein won NobelPrize

Bohr won NobelPrize

Einstein isa vegetarian

Cruise isa vegetarian

Cruise born 1962 Bohr died 1962

Ranking Approaches

Confidence:

Prefer results that are likely correct

- accuracy of info extraction
- trust in sources (authenticity, authority)

empirical accuracy of IE
PR/HITS-style estimate of trust
combine into:

max { accuracy (f,s) * trust(s) |
s ∈ witnesses(f) }

Informativeness:

Prefer results with salient facts

Statistical LM with estimations from:

- frequency in answer
- frequency in corpus (e.g. Web)
- frequency in query log

PR/HITS-style entity/fact ranking
[V. Hristidis et al., S.Chakrabarti, …]

or

IR models: tf*idf … [K.Chang et al., …]
Statistical Language Models

Diversity:

Prefer variety of facts

Statistical Language Models

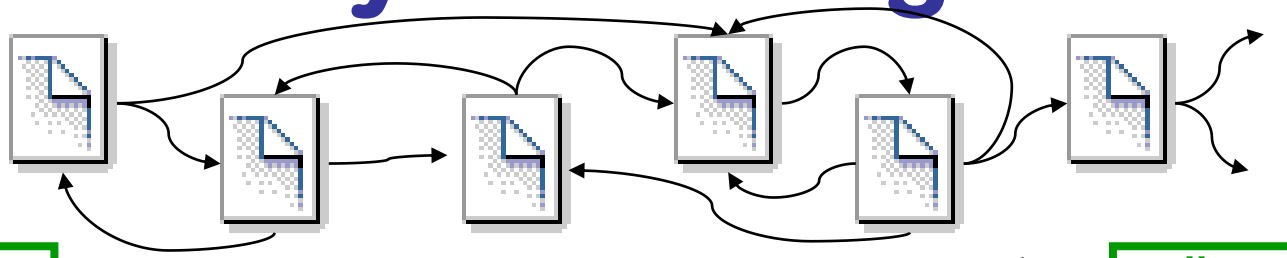
Conciseness:

Prefer results that are tightly connected

- size of answer graph
- cost of Steiner tree

graph algorithms (BANKS, STAR, …)
[J.X. Yu et al., S.Chakrabarti et al., B. Kimelfeld et al., A. Markovetz et al., B.C. Ooi et al., G.Kasneci et al., …]

PR/HITS-style Ranking of Entities



disk drives

usedIn

giant magneto-resistance

discovered

Wolf Prize

Peter Gruenberg

Nobel Prize

hasWon

Turing Award

hasWon

instanceOf

physicist

instanceOf

computer scientist

invented

2nd price auctions

usedIn

online ads

usedIn

Internet

usedIn

TCP/IP

TU Darmstadt

William Vickrey

workedAt

Princeton

Albert Einstein

degreeFrom

ETH Zurich

honDoct

UCLA

degreeFrom

Stanford

Vinton Cerf

workedAt

Google

spinoff

IT company

university

subclassOf

subclassOf

organization

...

...

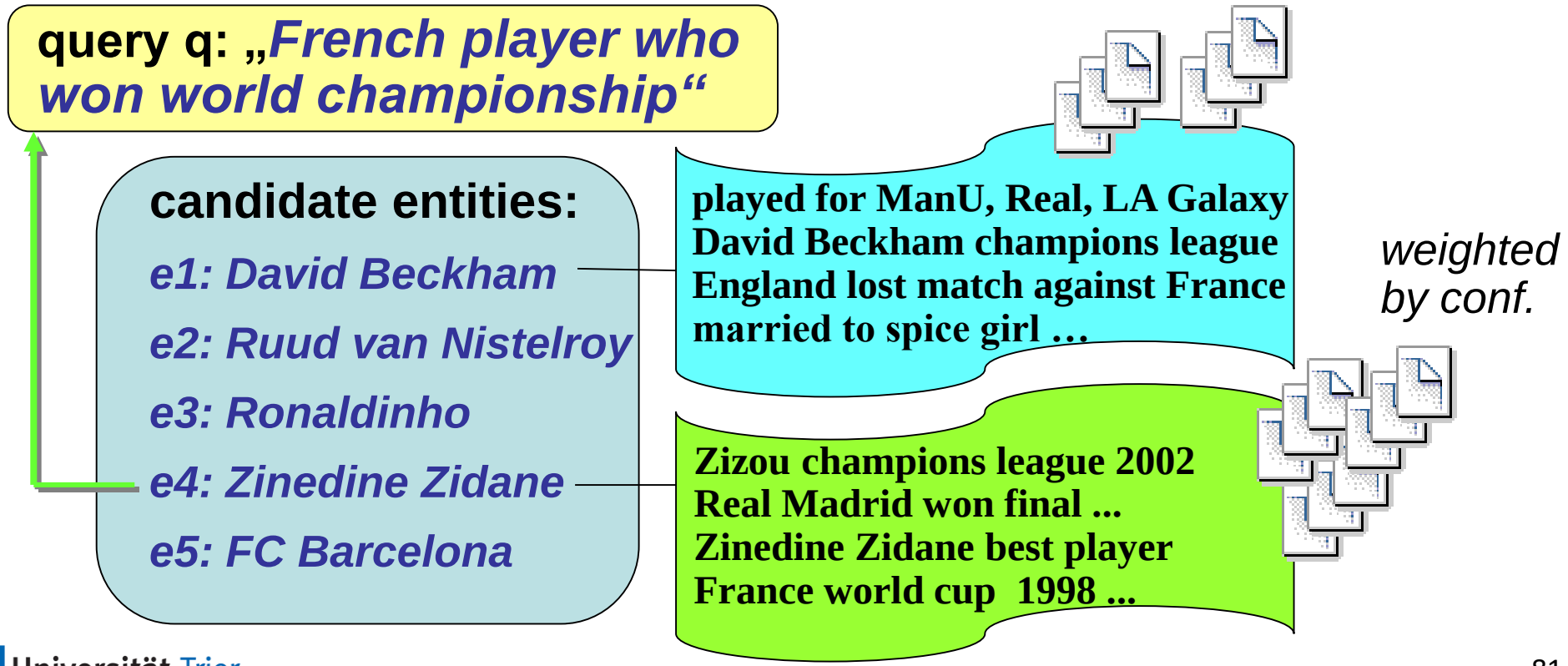
Entity Search with LM Ranking

[Nie 2007, ...]

query: keywords → answer: entities

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

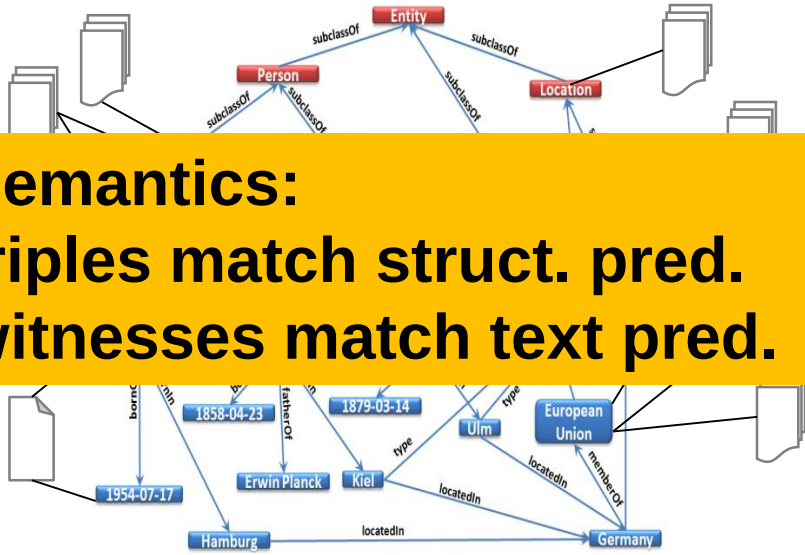
LM (entity e) = prob. distr. of words seen in context of e



Querying Facts & Text

Problem: not everything is triplified

- **Consider witnesses/sources (provenance meta-facts)**
- **Allow text predicates with each triple pattern (à la XQ-FT)**



Semantics:
triples match struct. pred.
witnesses match text pred.

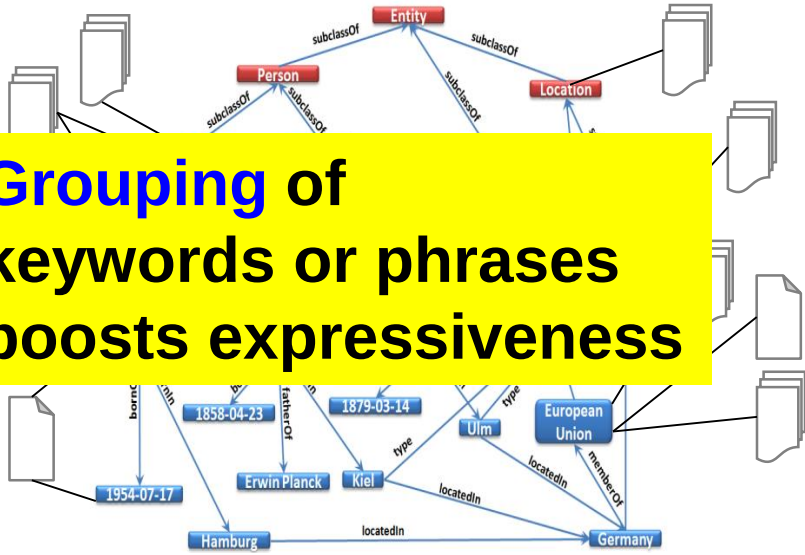
European composers who have won the Oscar, whose music appeared in dramatic western scenes, and who also wrote classical pieces ?

```
Select ?p Where {
?p instanceOf Composer .
?p bornIn ?t . ?t inCountry ?c . ?c locatedIn Europe .
?p hasWon ?a .?a Name AcademyAward .
?p contributedTo ?movie [western, gunfight, duel, sunset] .
?p composed ?music [classical, orchestra, cantata, opera] . }
```

Querying Facts & Text

Problem: not everything is triplified

- **Consider witnesses/sources (provenance meta-facts)**
- **Allow text predicates with each triple pattern (à la XQ-FT)**



Grouping of keywords or phrases boosts expressiveness

French politicians married to Italian singers?

```
Select ?p1, ?p2 Where {
?p1 instanceOf politician [France] .
?p2 instanceOf singer [Italy] .
?p1 marriedTo ?p2 . }
```

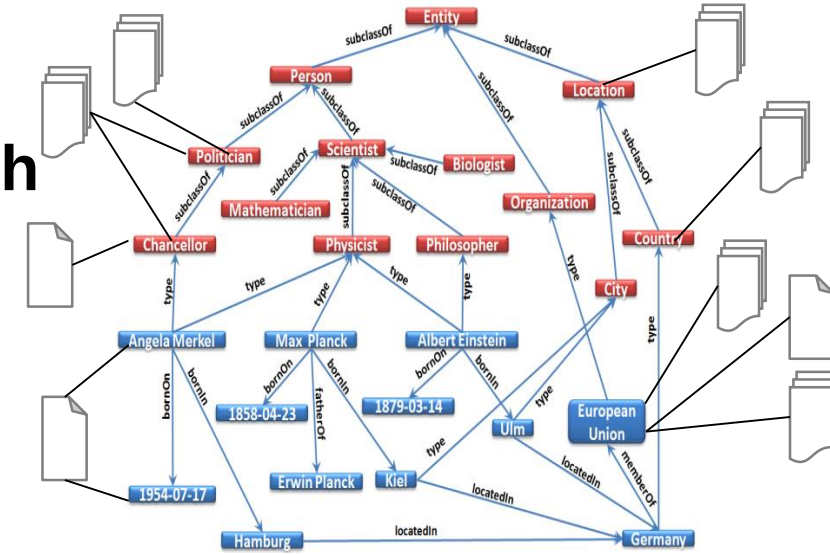
```
Select ?p1, ?p2 Where {
?p1 instanceOf ?c1 [France, politics] .
?p2 instanceOf ?c2 [Italy, singer] .
?p1 marriedTo ?p2 . }
```

CS researchers whose advisors worked on the Manhattan project?

```
Select ?r, ?a Where {
?r instanceOf researcher [“computer science“] .
?a worksOn [“Manhattan project“] .
?r hasAdvisor ?a . }
```

Relatedness Queries

Schema-agnostic keyword search (on RDF or relational DB) becomes a special case



denotes transitive connection

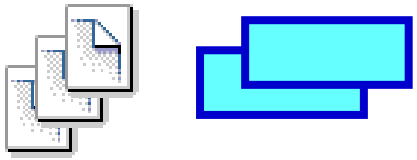
Relationship between Angela Merkel, Jim Gray, Dalai Lama?

```
Select ??p1, ??p2, ??p3 Where {
        AngelaMerkel ??p1 JimGray .
        JimGray ??p2 DalaiLama .
        DalaiLama ??p3 AngelaMerkel . }
```

```
Select ??p1, ??p2, ??p3 Where {
?e1 ?r1 ?c1 ["Angela Merkel"] .
?e2 ?r2 ?c2 ["Jim Gray"] .
?e3 ?r3 ?c3 ["Dalai Lama"] .
?e1 ??p1 ?e2 .
?e2 ??p2 ?e3 .
?e3 ??p3 ?e1 . }
```

LM‘s: from Entities to Facts

Document / Entity LM‘s



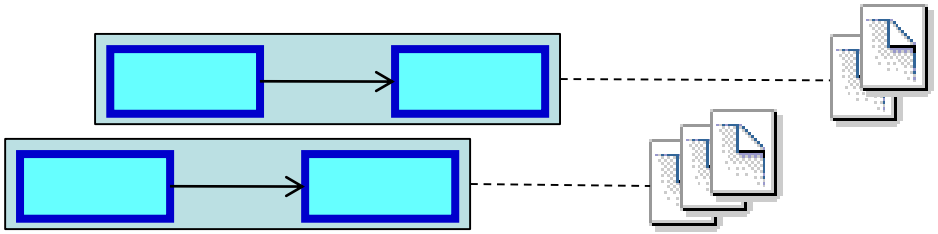
LM for doc/entity: prob. distr. of words

LM for query: (prob. distr. of) words

LM‘s: rich for docs/entities, super-sparse for queries

richer query LM with query expansion, etc.

Triple LM‘s



LM for facts: (degen. prob. distr. of) triple

LM for queries: (degen. prob. distr. of) triple pattern

LM‘s: apples and oranges

- ***expand query variables by S,P,O values from DB/KB***
- ***enhance with witness statistics***
- ***query LM then is prob. distr. of triples !***

LM's for Triples and Triple Patterns

[Kasneci 2008, Elbassuoni 2009]

triple patterns (queries q):

q: **LM(q) + smoothing**

q: Beckham p ManU	200/550
q: Beckham p Real	300/550
q: Beckham p Galaxy	20/550
q: Beckham p Milan	30/550

q: ?x p ASCannes

Zidane p ASCannes	20/30
Tidjani p ASCannes	10/30

q: ?x p ?y

Messi p FCBarcel
Zidane p RealMad
Kaka p ACMilan
...

q: Cruyff ?r FCBarcelona

Cruyff playedFor FCBarca	200/500
Cruyff playedAgainst FCBarca	50/500
Cruyff coached FCBarca	250/500

triples (facts f):

f1: Beckham p ManchesterU	200
f2: Beckham p RealMadrid	300
f3: Beckham p LAGalaxy	20
f4: Beckham p ACMilan	30
F5: Kaka p ACMilan	300
F6: Kaka p RealMadrid	150
f7: Zidane p ASCannes	20
f8: Zidane p Juventus	200
f14: Ribery p BayernMunich	100
f15: Drogba p Chelsea	150
f16: Casillas p RealMadrid	20

witness statistics Σ: 2600

LM(q): {t → P [t | t matches q] ~ #witnesses(t)}
LM(answer f): {t → P [t | t matches f] ~ 1 for f}
smooth all LM's
rank results by ascending KL(LM(q)|LM(f))

LM's for Composite Queries

q: Select ?x,?c Where { France ml ?x . ?x p ?c . ?c in UK . }

P [F ml Henry,
Henry p Arsenal,

P [F ml Drogba,
Drogba p Chelsea,
Chelsea in UK]

$$\sim \frac{30}{650} \cdot \frac{150}{2600} \cdot \frac{140}{500}$$

queries q with subqueries $q_1 \ldots q_n$

results are n-tuples of triples $t_1 \ldots t_n$

LM(q): $P[q_1 \ldots q_n] = \prod_i P[q_i]$

LM(answer): $P[t_1 \ldots t_n] = \prod_i P[t_i]$

KL(LM(q)|LM(answer)) = $\sum_i$ KL(LM(q_i)|LM(t_i))

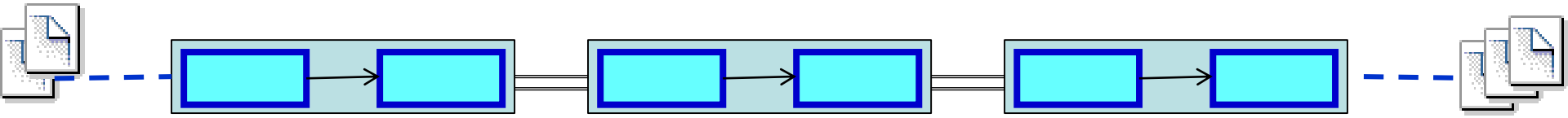
f21: F ml Zidane	200
f22: F ml Tidjani	20
f23: F ml Henry	200
f24: F ml Ribery	200
f25: F ml Drogba	30
f26: IC ml Drogba	100
f27 ALG ml Zidane	50

f1: Beckham p ManU	200
f7: Zidane p ASCannes	20
f8: Zidane p Juventus	200
f9: Zidane p RealMadrid	300
f10: Tidjani p ASCannes	10
f12: Henry p Arsenal	200
f13: Henry p FCBarca	150
f14: Ribery p Bayern	100
f15: Drogba p Chelsea	150

f31: ManU in UK	200
f32: Arsenal in UK	160
f33: Chelsea in UK	140

LM's for Keyword-Augmented Queries

q: Select ?x, ?c Where {
France ml ?x *[goalgetter, "top scorer"]* **.**
?x p ?c .
?c in UK *[champion, "cup winner", double]* **. }**



subqueries q_i with keywords w_1 … w_m

results are still n-tuples of triples t_i

LM(q_i): P[triple t_i | w_1 … w_m] = $\prod_k$ β P[t_i | w_k] + (1–β) P[t_i]

LM(answer f_i) analogous

KL(LM(q)|LM(answer f_i)) = $\sum_i$ KL (LM(q_i) | LM(f_i))

result ranking prefers (n-tuples of) triples whose witnesses score high on the subquery keywords

Result Diversification

[Carbonell 1998]

q: Select ?p, ?c Where { ?p isa SoccerPlayer . ?p playedFor ?c . }

1 Beckham, ManchesterU
2 Beckham, RealMadrid
3 Beckham, LAGalaxy
4 Beckham, ACMilan
5 Zidane, RealMadrid
6 Kaka, RealMadrid
7 Cristiano Ronaldo, RealMadrid
8 Raul, RealMadrid
9 van Nistelrooy, RealMadrid
10 Casillas, RealMadrid

1 Beckham, ManchesterU
2 Beckham, RealMadrid
3 Zidane, RealMadrid
4 Kaka, ACMilan
5 Cristiano Ronaldo, ManchesterU
6 Messi, FCBarcelona
7 Henry, Arsenal
8 Ribery, BayernMunich
9 Drogba, Chelsea
10 Luis Figo, Sporting Lissabon

rank results f_1 ... f_k by ascending

$$\delta\ KL(LM(q) \mid LM(f_i)) - (1-\delta)\ KL(LM(f_i) \mid LM(\{f_1..f_k\} \setminus \{f_i\}))$$

implemented by greedy re-ranking of f_i's in candidate pool

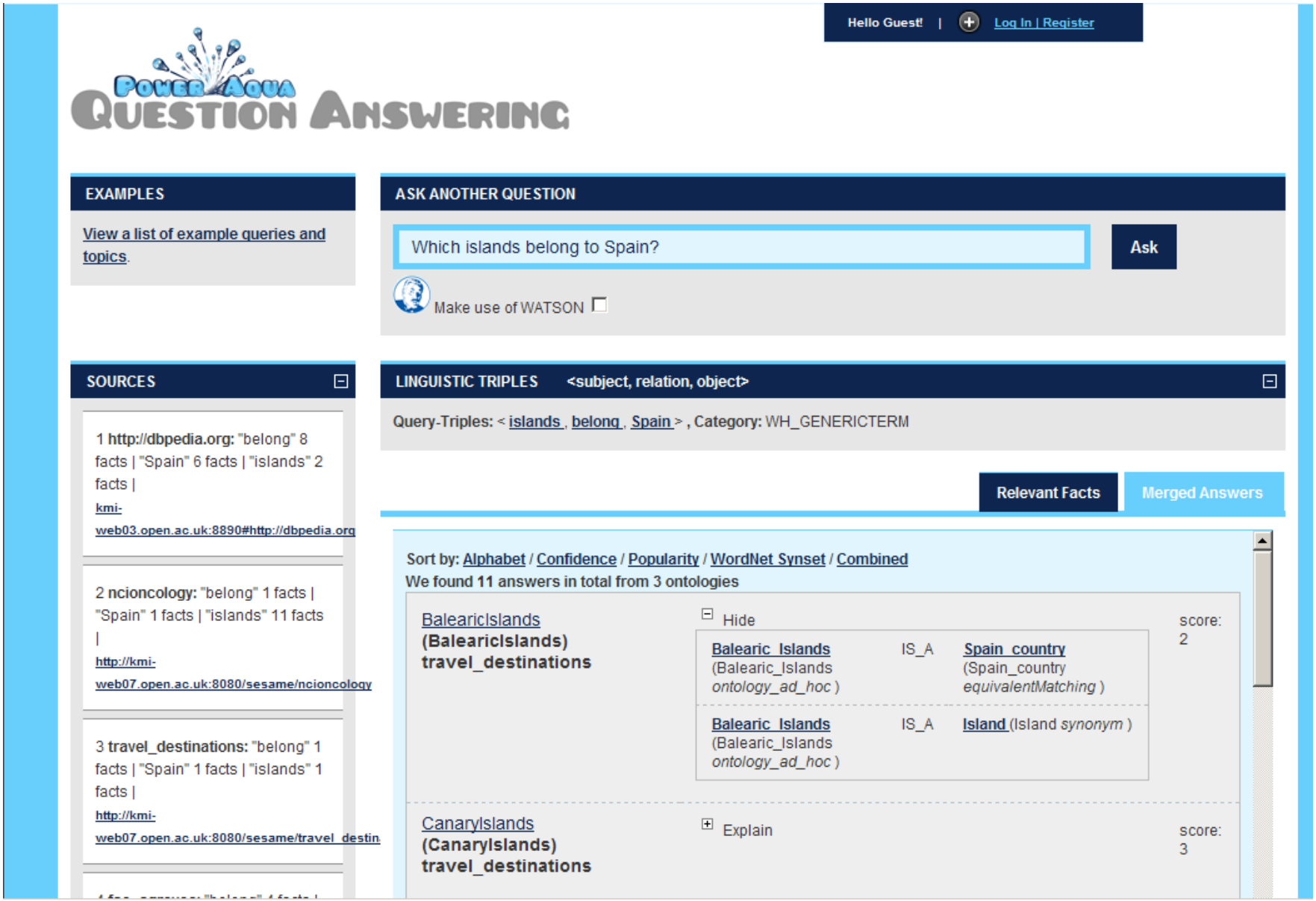
Natural Language Queries

Paradigm:

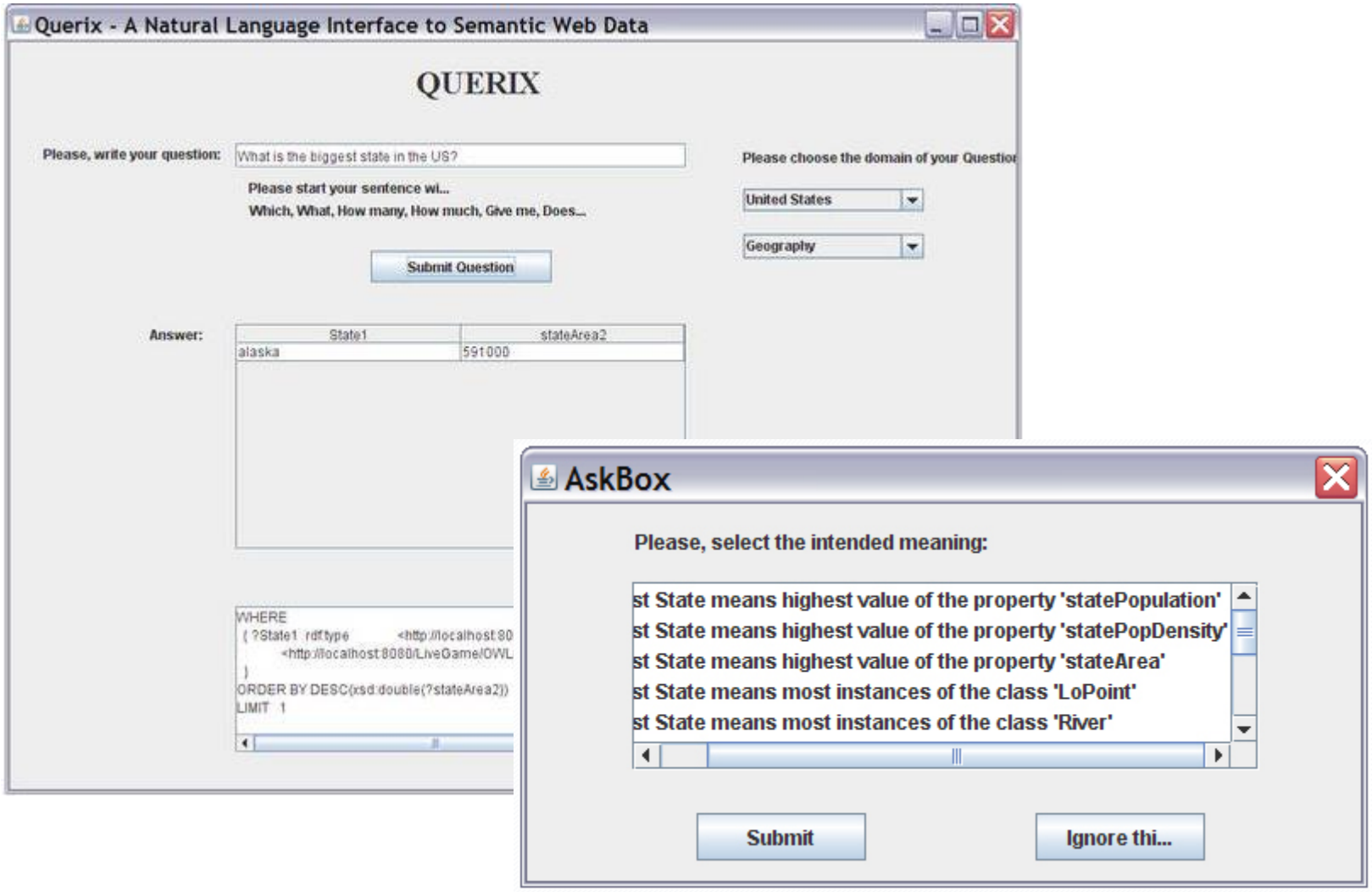
Allow queries in plain English

- Map (groups of) keywords to triple patterns, based on existing triples: [NLP-Reduce, PowerAqua]
  "Find a restaurant that is in Barcelona"
  ⇒ `?r isIn Barcelona. ?r isA restaurant`
- Extract query skeleton from syntax tree, heuristic match to known patterns: [Querix]
  "What is a restaurant in Barcelona"
  ⇒ Q-V-N-P-N
- Ambiguities resolved by user interaction or by automated methods
- Controlled language: Present possible continuation of query based on grammar [Ginseng]

Example: PowerAqua (Open University, UK)



Example: Querix (Uni Zurich)



http://www.ifi.uzh.ch/ddis/research/talking-to-the-sema

Natural Language Queries

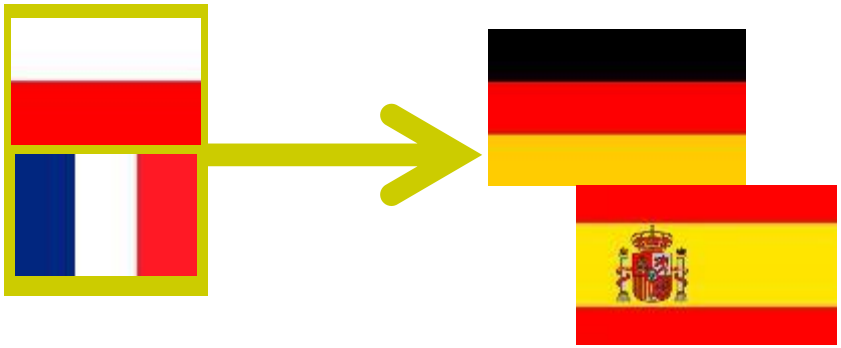
Pros:

- Intuitive to use
- No schema knowledge necessary

Cons:

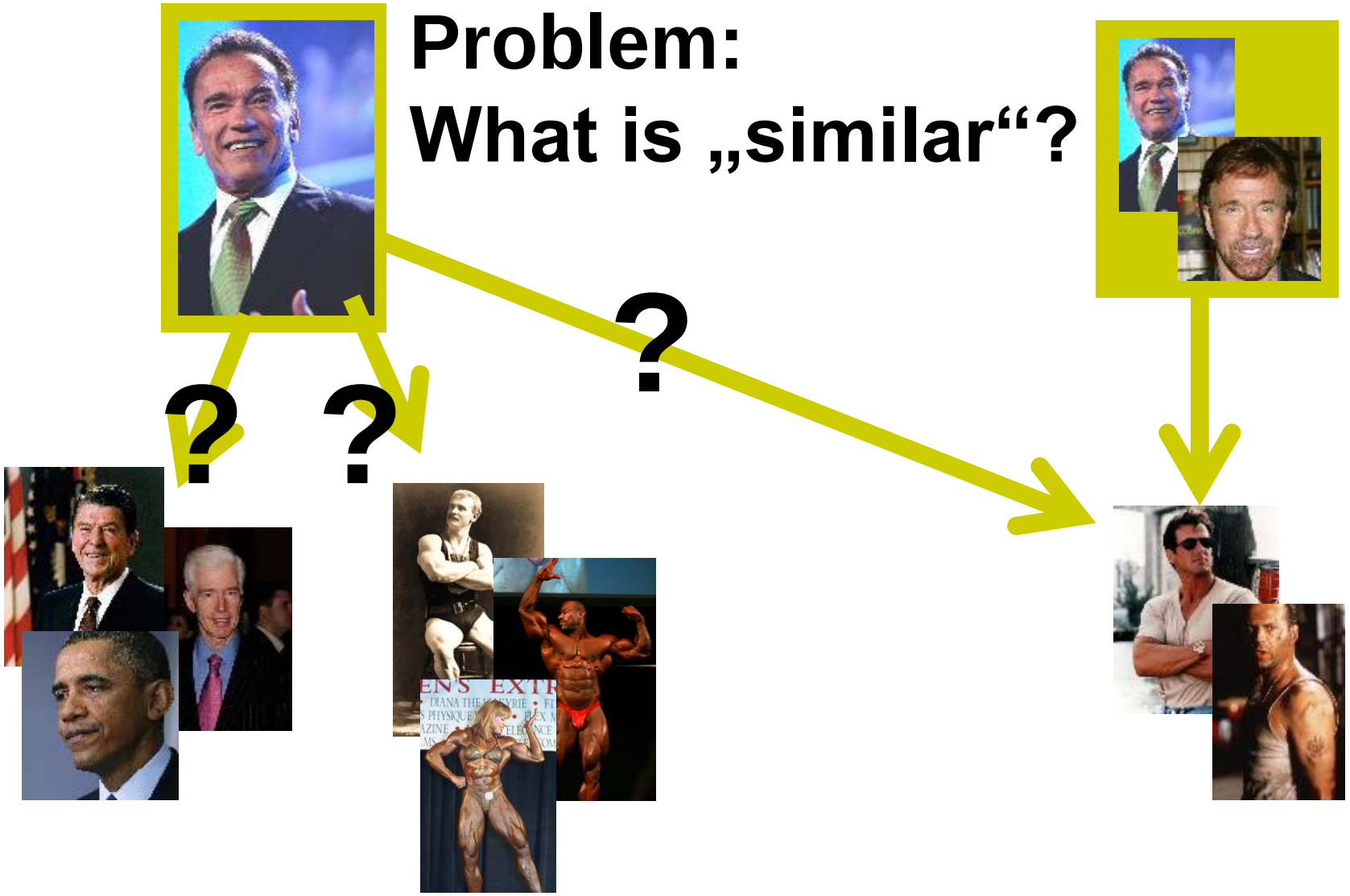
- Often domain-specific
- Finding good query formulation often hard
- Result quality often poor

Example-Based Search



Input: Set of entities
Output: (Ranked) list
of 'similar' entities
Similarity defined by shared
semantic properties

Disambiguitiy of Query



Maximal Aspects

[Metzger 2017]

?x type Actor

?x memberOf Democrats

?x type AmericanPolitician

?x type Politician

Compound Aspect A

Basic Aspect

Entity Set E(A) of A

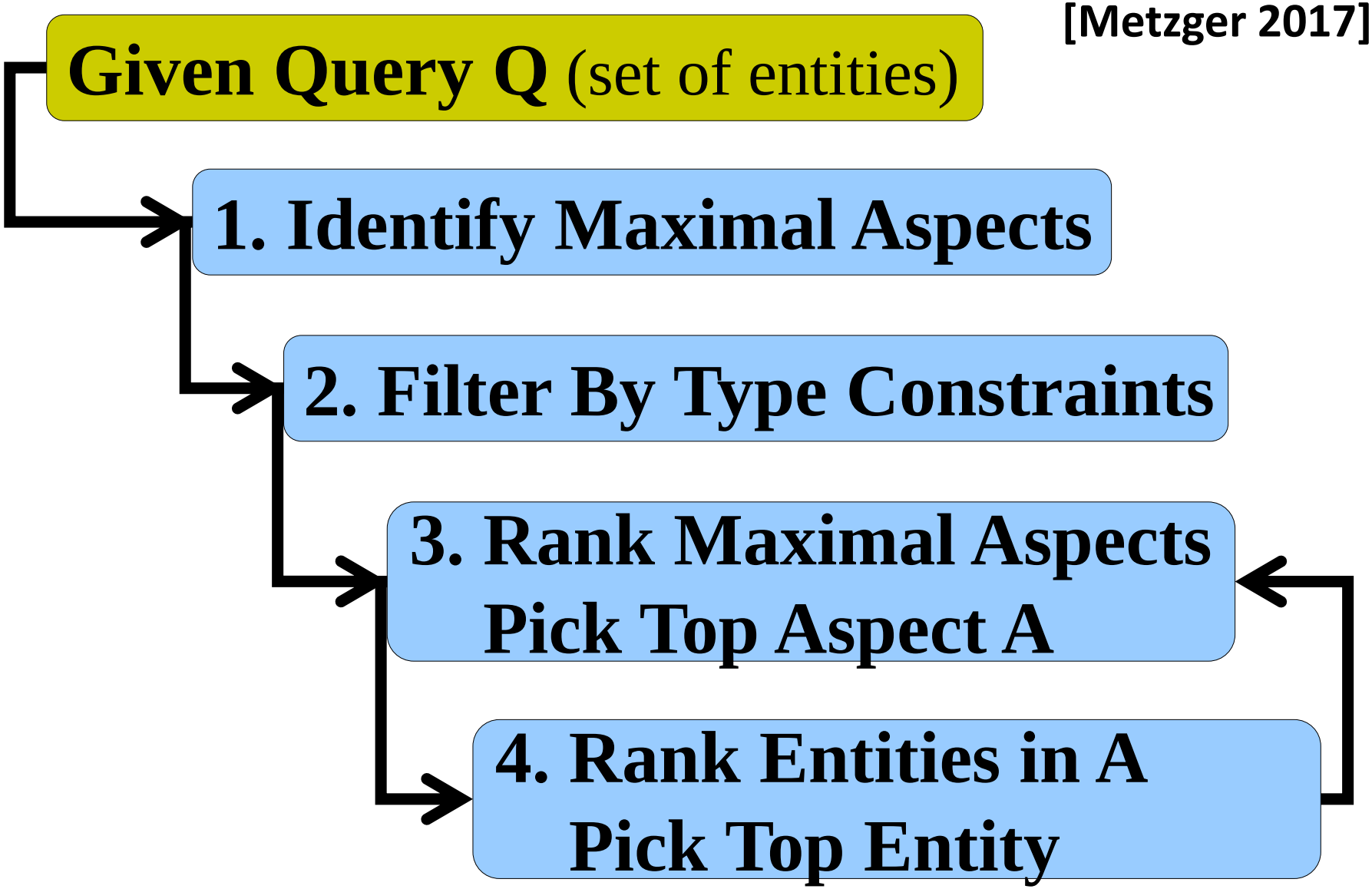
Maximal Aspect:

1) $E(A)/Q \neq \emptyset$

2) Adding any basic aspect violates 1)

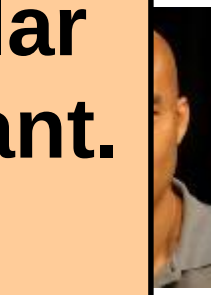
QBEES: General Approach

[Metzger 2017]



QBEES: Entity Selection

- Which one is more relevant?
- Assumption: the more popular the more likely to be relevant.
- Two variants:
  - Wikipedia clicks
  - Random walk on the graph




?x type Actor
?x actedIn The_Expendables
?x actedIn ?


Outline

- Knowledge-based Search – What and Why
- Existing Knowledge Bases
- Building and Maintaining Knowlegde Bases
- Knowledge-Based Search
- **Temporal Knowledge**

Temporal Knowledge

Which facts for given relations hold
at what time point or during which time intervals ?

marriedTo (Madonna, Guy) [22Dec2000, Dec2008]
capitalOf (Berlin, Germany) [1990, now]
capitalOf (Bonn, Germany) [1949, 1989]
hasWonPrize (JimGray, TuringAward) [1998]
graduatedAt (HectorGarcia-Molina, Stanford) [1979]
graduatedAt (SusanDavidson, Princeton) [Oct 1982]
hasAdvisor (SusanDavidson, HectorGarcia-Molina) [Oct 1982, forever]

How can we query & reason on entity-relationship facts
in a "time-travel" manner - with uncertain/incomplete KB ?

US president when Barack Obama was born?
students of Hector Garcia-Molina while he was at Princeton?

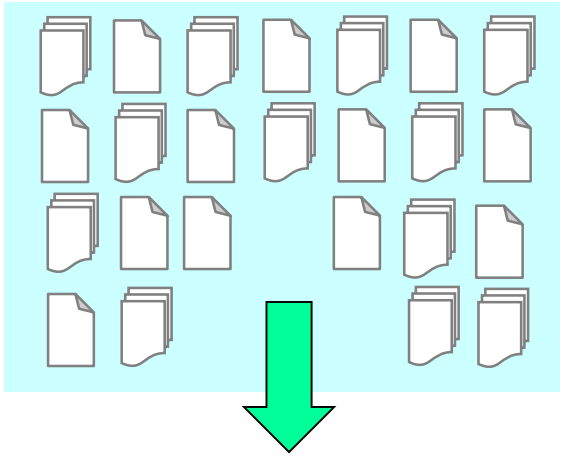
French Marriage Problem

JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

facts in KB

1: married
(Hillary, Bill)
2: married
(Carla, Nicolas)
3: married
(Angelina, Brad)

validFrom (2, 2008)



new fact candidates:

4: married (Cecilia, Nicolas)
5: married (Carla, Benjamin)
6: married (Carla, Mick)
7: divorced (Madonna, Guy)
8: domPartner (Angelina, Brad)

validFrom (4, 1996) validUntil (4, 2007)
validFrom (5, 2010)
validFrom (6, 2006)
validFrom (7, 2008)

Issues & Challenges

- Grow the KB and keep it consistent (near-human quality)?

- Combine new facts/events with old facts?
- When to overwrite/invalidate facts in the KB?
- When to create new facts/versions with temporal scope?

- How to gather temporal scopes of facts?
- Aggregate and reconcile temporal cues for a fact?
- Reason on consistency of temporal scopes across facts?

- When to actively acquire new facts for an entity or class?
- Dealing with new entities? New relation types?

Challenge: Temporal Knowledge

for all people in Wikipedia (300 000) gather all spouses, incl. divorced & widowed, and corresponding time periods!

>95% accuracy, >95% coverage, in one night

1) **recall: gather temporal scopes for base facts**
2) **precision: reason on mutual consistency**

	28 January 1955 (age 53) Paris, France
	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
Alma mater	University of Paris X: Nanterre
Occupation	Lawyer
Religion	Roman Catholic

consistency constraints are potentially helpful:

- **functional dependencies:** ***husband, time → wife***
- **inclusion dependencies:** ***marriedPerson ⊆ adultPerson***
- **age/time/gender restrictions:** ***birthdate + Δ < marriage < divorce***

Difficult Dating

Nicolas Sarkozy

President of France

Incumbent

Assumed office

Born	28 January 1955 (age 55) Paris, France
Political party	Union for a Popular Movement (2002–present)
Other political affiliations	Rally for the Republic (1976–2002)
Spouse(s)	Marie-Dominique Culioli (1982–1996) Cécilia Ciganer-Albéniz (1996–2007) Carla Bruni-Sarkozy (2008–present)
Children	Pierre Sarkozy (by Culioli) Jean Sarkozy (by Culioli) Louis Sarkozy (by Ciganer-Albéniz)
Residence	Élysée Palace
Alma mater	Paris X University Nanterre
Profession	Lawyer
Religion	Roman Catholicism

Cécilia Attias

First Lady of France

In office
16 May 2007 – 10 October 2007

President	Nicolas Sarkozy
Preceded by	Bernadette Chirac
Succeeded by	Carla Bruni
Born	November 12, 1957 (age 52) Boulogne-Billancourt, France
Spouse(s)	Jacques Martin (m. 1984–1989) Nicolas Sarkozy (m. 1996–2007) Richard Attias (m. 2008–present)
Children	Judith Martin (b.1984) Jeanne-Marie Martin (b.1987) Louis Sarkozy (b.1997)

Wife of the President of the French Republic

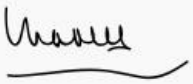
Incumbent

Assumed office
2 February 2008

President	Nicolas Sarkozy
Preceded by	Cécilia Ciganer-Albéniz
Born	23 December 1967 (age 42) Turin, Italy
Birth name	Carla Gilberta Bruni Tedeschi
Nationality	Italian, French[1]
Spouse(s)	Nicolas Sarkozy
Children	Aurélien Enthoven (with Raphaël Enthoven)

Charles

Prince of Wales; Duke of Rothesay (more)

Spouse	Lady Diana Spencer m. 1981; div. 1996 Camilla Parker Bowles m. 2005
Issue	Prince William of Wales Prince Harry of Wales
Full name	Charles Philip Arthur George
House	Maternal: House of Windsor Paternal: House of Schleswig-Holstein-Sonderburg-Glücksburg
Father	Prince Philip, Duke of Edinburgh
Mother	Elizabeth II
Born	14 November 1948 (age 61) Buckingham Palace, London
Signature	
Religion	Christian (Church of England)

Diana

Princess of Wales; Duchess of Rothesay

Spouse	Charles, Prince of Wales (29 July 1981 – 28 August 1996)
Issue	Prince William of Wales Prince Henry of Wales
Full name	Diana Frances Spencer[N 1]
House	House of Windsor
Father	John Spencer, 8th Earl Spencer
Mother	Frances Shand Kydd
Born	1 July 1961 Park House, Sandringham, Norfolk
Died	31 August 1997 (aged 36) Pitié-Salpêtrière Hospital, Paris, France
Burial	Althorp, Northamptonshire

Madonna

Madonna at the premiere of *I Am Because We Are* in 2008.

Background information

Birth name	Madonna Louise Ciccone

Guy Ritchie

Guy Ritchie, September 2008

Born	Guy Stuart Ritchie 10 September 1968 (age 41) Hatfield, Hertfordshire, England
Occupation	Filmmaker, Screenwriter
Years active	1995–present
Spouse(s)	Madonna (2000–2008) (divorced)

(Even More Difficult) Implicit Dating

explicit dates vs.
implicit dates relative to other dates

Nicolas Sarkozy

From Wikipedia, the free encyclopedia

"Sarkozy" redirects here. For the surname, see Sárközi (surname).

Nicolas Sarkozy ((pronounced [nikɔla saʁkɔzi] (help·info)), born **Nicolas Paul Stéphane Sarkozy de Nagy-Bocsa** on 28 January 1955) is the 23rd and current President of the French Republic and *ex officio* Co-Prince of Andorra. He assumed the office on 16 May 2007 after defeating Socialist Party candidate Ségolène Royal 10 days earlier.

Before his presidency, he was leader of the Union for a Popular Movement (UMP). Under Jacques Chirac's presidency he served as Minister of the Interior in Jean-Pierre Raffarin's (UMP) first two governments (from May 2002 to March 2004), then was appointed Minister of Finances in Raffarin's last government (March 2004 to May 2005) and again Minister of the Interior in Dominique de Villepin's government (2005–2007).

Sarkozy was also president of the General council of the Hauts-de-Seine department from 2004 to 2007 and mayor of Neuilly-sur-Seine, one of the wealthiest communes of France from 1983 to 2002. He was Minister of the Budget in the government of Édouard Balladur (RPR, predecessor of the UMP) during François Mitterrand's last term.

Sarkozy is known for wanting to revitalize the French economy.[1][2][3] He has pledged to revive the work ethic, promote new initiatives and fight intolerance.[1] In foreign affairs he has promised a strengthening of the entente cordiale with the United Kingdom[4] and closer cooperation with the United States.[5] He married Carla Bruni-Sarkozy on 2 February 2008 at the Élysée Palace in Paris.

Even More Difficult) Relative Dating

vague dates
relative dates

narrative text
relative order

Early life

During Sarkozy's childhood, his father refused to give his wife's family any financial help, even though he had founded his own advertising agency and had become wealthy. The family lived in a small mansion owned by Sarkozy's grandfather, Benedict Mallah, in the 17th Arrondissement. The family later moved to Neuilly-sur-Seine, one of the wealthiest communes of the Île-de-France *région* immediately west of the 17th Arrondissement just outside of Paris. According to Sarkozy, his staunchly Gaullist grandfather was more of an influence on him than his father, whom he rarely saw. Sarkozy was, accordingly, raised Catholic.[18]

Sarkozy said that being abandoned by his father shaped much of who he is today. He also has said that, in his early years, he felt inferior in relation to his wealthier classmates.[19] "What made me who I am now is the sum of all the humiliations suffered during childhood", he said later.[19]

Education

Sarkozy was enrolled in the *Lycée Chaptal* a state-funded public middle and high school in Paris's 8th arrondissement, where he failed his *sixième*. His family then sent him to the *Cours Saint-Louis de Monceau*, a private Catholic school in the 17th arrondissement, where he was reportedly a mediocre student,[20] but where he nonetheless obtained his *baccalauréat* in 1973. He enrolled at the *Université Paris X Nanterre* where he graduated with a Master in Private law, and later with a DEA degree in Business law. Paris X Nanterre had been the starting place for the May '68 student movement and was still a stronghold of leftist students. Described as a quiet student, Sarkozy soon joined the right-wing student organization, in which he was very active. He completed his military service as a part time Air Force cleaner.[21] After graduating, he entered the *Institut d'Études Politiques de Paris* (1979–1981) but failed to graduate due to an insufficient command of the English language.[22] After passing the bar, he became a lawyer specializing in business and family law,[23] and was one of Silvio Berlusconi's top French advocates.[24][25][26]

References (1)

[Rebele 2016] Thomas Rebele, Fabian M. Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, Gerhard Weikum: YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. International Semantic Web Conference (2) 2016: 177-185

[Suchanek 2007] Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum: Yago: a core of semantic knowledge. WWW 2007: 697-706

[Weikum 2009] Gerhard Weikum, Gjergji Kasneci, Maya Ramanath, Fabian M. Suchanek: Database and information-retrieval methods for knowledge discovery. Commun. ACM 52(4): 56-64 (2009)

[Fellbaum 1998] Christiane Fellbaum (ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[Hoffart 2011] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, Gerhard Weikum: Robust Disambiguation of Named Entities in Text. EMNLP 2011: 782-792

[Metzger 2017] Steffen Metzger, Ralf Schenkel, Marcin Sydow: QBEES: query-by-example entity search in semantic knowledge graphs based on maximal aspects, diversity-awareness and relaxation. Intelligent Information Systems, https://doi.org/10.1007/s10844-017-0443-x

[Lehmann 2015] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, Christian Bizer: DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6(2): 167-195 (2015)

[Auer 2007] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary G. Ives: DBpedia: A Nucleus for a Web of Open Data. ISWC/ASWC 2007: 722-735

[Elbassuoni 2011] Shady Elbassuoni, Maya Ramanath, Ralf Schenkel, Marcin Sydow, Gerhard Weikum: Language-model-based ranking for queries on RDF-graphs. CIKM 2009: 977-986

References (2)

[Yahya 2013] Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, Gerhard Weikum: Robust question answering over the web of linked data. CIKM 2013: 1107-1116

[McCallum 2003] Andrew McCallum, Khashayar Rohanimanesh, Charles Sutton: Dynamic Conditional Random Fields for Jointly Labeling Multiple Sequences. NIPS Workshop on Syntax, Semantics, and Statistics. 2003

[Madhavan 2005] Jayant Madhavan, Philip A. Bernstein, AnHai Doan, Alon Y. Halevy: Corpus-based Schema Matching. ICDE 2005: 57-68

[Bhattacharya 2007] Indrajit Bhattacharya, Lise Getoor: Collective entity resolution in relational data. TKDD 1(1): 5 (2007)

[Poon 2007] Hoifung Poon, Pedro M. Domingos: Joint Inference in Information Extraction. AAAI 2007: 913-918

[Kulkarni 2009] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, Soumen Chakrabarti: Collective annotation of Wikipedia entities in web text. KDD 2009: 457-466

[Etzioni 2004] Oren Etzioni et al.: Web-scale information extraction in knowitall: (preliminary results). WWW 2004: 100-110

[Wang 2008] Richard C. Wang, William W. Cohen: Iterative Set Expansion of Named Entities Using the Web. ICDM 2008: 1091-1096

[He 2011] Yeye He, Dong Xin: SEISA: set expansion by iterative similarity aggregation. WWW 2011: 427-436

[Carlson 2010] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., Tom M. Mitchell: Toward an Architecture for Never-Ending Language Learning. AAAI 2010

[Ponzetto 2007] Simone Paolo Ponzetto, Michael Strube: Deriving a Large-Scale Taxonomy from Wikipedia. AAAI 2007: 1440-1445

References (3)

[Kushmerick 1997] Nicholas Kushmerick, Daniel S. Weld, Robert B. Doorenbos: Wrapper Induction for Information Extraction. IJCAI (1) 1997: 729-737

[Califf 1999] Mary Elaine Califf, Raymond J. Mooney: Relational Learning of Pattern-Match Rules for Information Extraction. AAAI/IAAI 1999: 328-334

[Baumgartner 2001] Robert Baumgartner, Sergio Flesca, Georg Gottlob: Visual Web Information Extraction with Lixto. VLDB 2001: 119-128

[Hearst 1992] Marti A. Hearst: Automatic Acquisition of Hyponyms from Large Text Corpora. COLING 1992: 539-545

[Brin 1998] Sergey Brin: Extracting Patterns and Relations from the World Wide Web. WebDB 1998: 172-183

[Agichtein 2000] Eugene Agichtein, Luis Gravano: *Snowball*: extracting relations from large plain-text collections. ACM DL 2000: 85-94

[Ravichandran 2002] Deepak Ravichandran, Eduard H. Hovy: Learning surface text patterns for a Question Answering System. ACL 2002: 41-47

[Suchanek 2006] Fabian M. Suchanek, Georgiana Ifrim, Gerhard Weikum: Combining linguistic and statistical analysis to extract relations from web documents. KDD 2006: 712-717

[Nakashole 2011] Ndapandula Nakashole, Martin Theobald, Gerhard Weikum: Scalable knowledge harvesting with high precision and high recall. WSDM 2011: 227-236

[Bunescu 2005] Razvan C. Bunescu, Raymond J. Mooney: A Shortest Path Dependency Kernel for Relation Extraction. HLT/EMNLP 2005: 724-731

[Suchanek 2009] Fabian M. Suchanek, Mauro Sozio, Gerhard Weikum: SOFIE: a self-organizing framework for information extraction. WWW 2009: 631-640

[Nie 2007] Zaiqing Nie, Yunxiao Ma, Shuming Shi, Ji-Rong Wen, Wei-Ying Ma: Web object retrieval. WWW 2007: 81-90

[Carbonell 1998] Jaime G. Carbonell, Jade Goldstein: The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. SIGIR 1998: 335-336