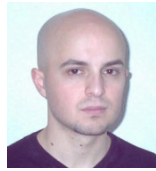
max planck institut informatik

Knowledge on the Web:

Towards Robust and Scalable Harvesting of Entity-Relationship Facts

Gerhard Weikum
Max Planck Institute for Informatics
http://www.mpi-inf.mpg.de/~weikum/

Acknowledgements



max planck institut informatik

living knowledge

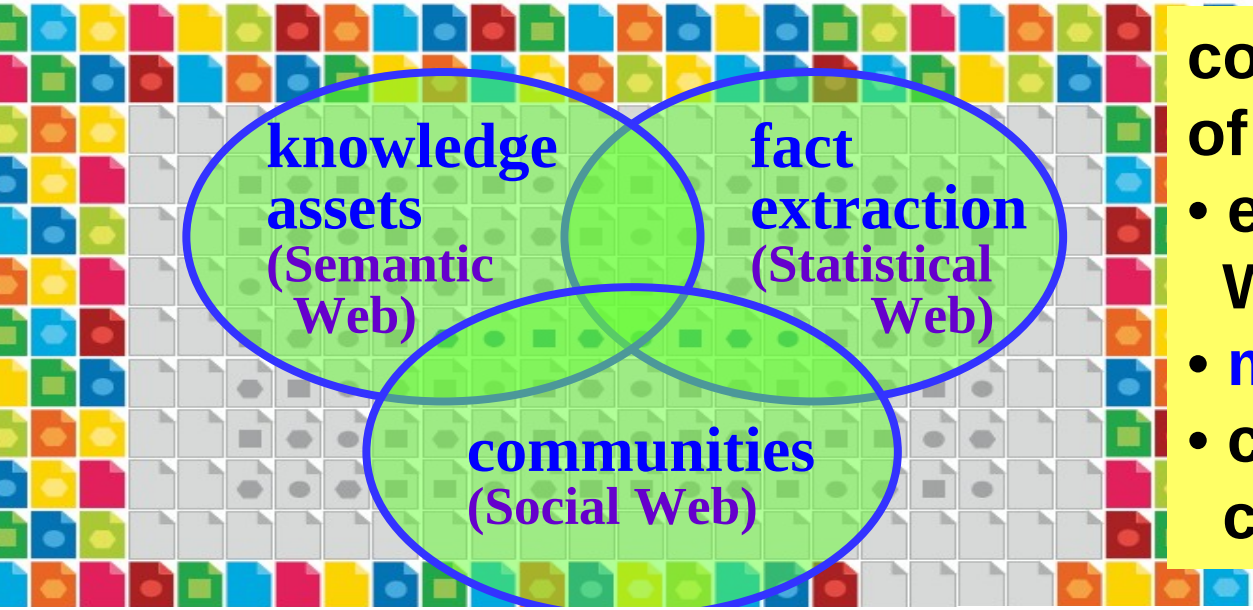
SEVENTH FRAMEWORK PROGRAMME

M2CI CLUSTER OF EXCELLENCE

DFG Deutsche Forschungsgemeinschaft

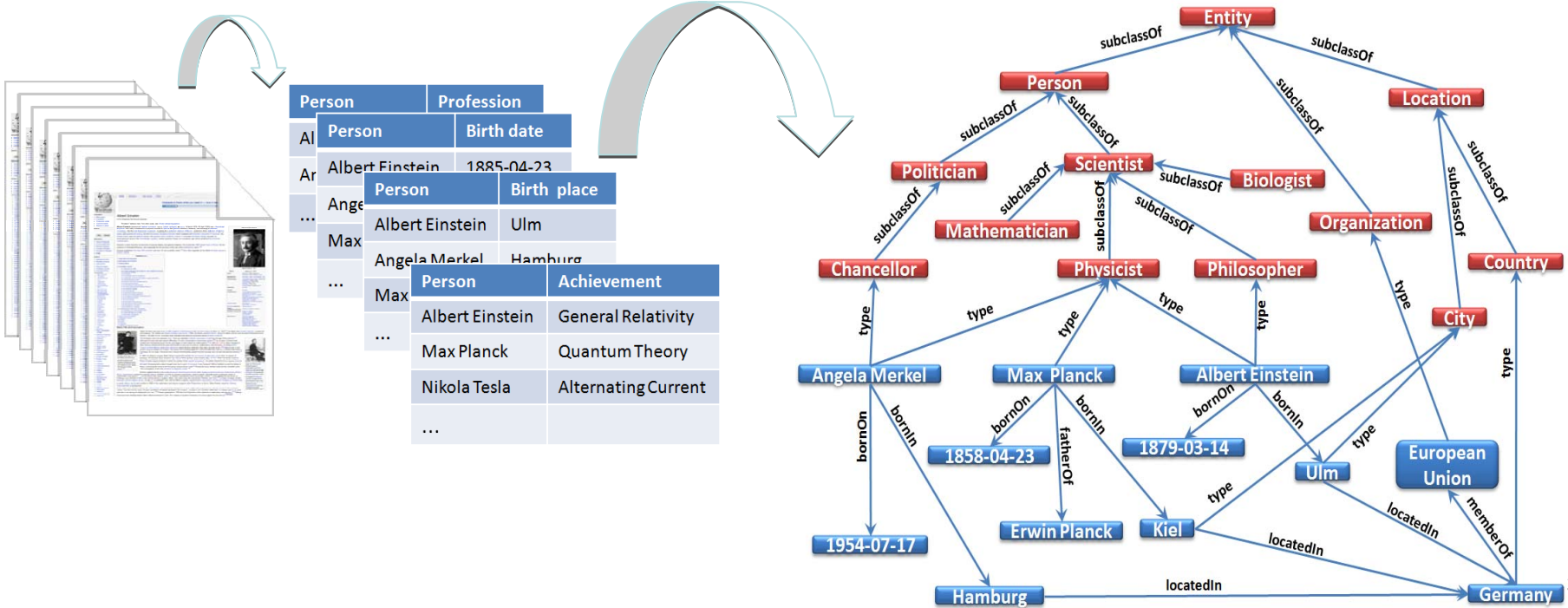
Google

Vision: Turn Web into Knowledge Base



comprehensive DB of human knowledge

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



Knowledge as Enabling Technology

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

★ German chancellor when Angela Merkel was born?

★ Japanese computer science institutes?

★ Politicians who are also scientists?

★ Enzymes that inhibit HIV?
Influenza drugs for pregnant women?

...

Knowledge Search on the Web (1)

Google squared labs

sushi ingredients | Square it | Add to this Square | Unsaved | Share

sushi ingredients

	Item Name	Image	Description	Cholesterol	Protein
☒	Nori seaweed		**Nori** (海苔) is the Japanese name for various edible **seaweed** species of the red alga Porphyra including most notably P. yezoensis and P. tenera, ...	0mg	1 possible value
☒	Ginger		With your Sushi order, you will be served some pickled **ginger**, ... Eat a slice of pickled **ginger** after each variety of sushi to cleanse your palate. ...	0 mg	1g
☒	Tuna		½ lb of fresh **tuna** fillet. 2 cups of sushi rice. ... Using your hands, take the rice and form a rectangular shape to match the **tuna** slices. ...	0mg	0g
☒	Sashimi		Sushi plate (盛り合わせ) with **sashimi** to the left. There are many sushi and **sashimi** ingredients, some traditional and some contemporary. ...	0mg	5.3g
☒	Sushi Rice Recipe		Long grain rice isn't proper for sushi because it is drier and doesn't stick together. Photo Tutorial: How to Cook **Sushi Rice** · Sushi **Recipes** ...	0mg	4 possible values
☒	Sushi-rolling Mat		SushiNow.com specializes in teaching **Sushi Rolling** as well as supplying ... As well as finding great dry goods for your Sushi (rice, rolling **mat**, nori, etc. ...		
☒	Albacore		**Albacore** Tuna Sushi **Albacore** Tuna: High-fat and rich in omega-3 fatty acids, the **albacore** ... **Albacore** tuna are commonly 80 pounds and can reach 200 pounds. ...	4 possible values	13g
☒	Unagi		**Unagi** (うなぎ) is the Japanese word for freshwater eels, especially the Japanese eel, Anguilla japonica. Saltwater eels are known as anago in Japanese. ...		

Add items | Add | or Add next 10 items

Add

Type your own...
- Color
- Calories
- Awards
- Language
- Author
- Featuring
- Nutrition
- Nutrition Facts
- Nutritional Information
- Nutritional Value

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

Knowledge Search on the Web (1)

Query: Japanese computers ? institutes ?

Google squared labs

japanese computers | Square it | Add to this Square

japanese computers

	Item Name	Image	Description	Number Of
☒	Hiragana		Typing Japanese Words in **Hiragana**, Katakana and Kanji ... If you want katakana for the word, type the word in **hiragana** and press "F7" key. ...	72
☒	Tomoyuki Nishita		**Tomoyuki Nishita** is a professor in the Department of Complexity Science and Engineering (also in the Department of Information Science) at University of ...	1 possible value
☒	Katakana		Typing Japanese Words in Hiragana, **Katakana** and Kanji ... If you want **katakana** for the word, type the word in hiragana and press "F7" key. ...	104
	Add items Add	or Add next 10 items		

Add

Type your own...

- Isbn
- Subjects
- Sister Systems
- Unicode Range
- Time Period
- Spoken Languages
- Type
- Symbol
- Tattoo

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

Knowledge Search on the Web (2)

Ranked results

Score: 0.42442708378025645

"politician" —means—> Massachusetts politicians
"scientist" —means—> American scientists
Benjamin Franklin —type—> American scientists
Benjamin Franklin —type—> Massachusetts politicians

- $x = Benjamin Franklin

Score: 0.42442708378025645

"politician" —means—> Foreign-born American politicians
"scientist" —means—> American political scientists
Zbigniew Brzezinski —type—> American political scientists
Zbigniew Brzezinski —type—> Foreign-born American politic

- $x = Zbigniew Brzezinski

Score: 0.42442708378025645

"politician" —means—> Jewish American politicians
"scientist" —means—> Jewish American scientists
Alan Greenspan —type—> Jewish American politicians
Alan Greenspan —type—> Jewish American scientists

- $x = Alan Greenspan

Query:
politicians who are also scientists ?

?x isa politician .
?x isa scientist

Results:
Benjamin Franklin
Zbigniew Brzezinski
Alan Greenspan
Angela Merkel
…

mpii

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

Knowledge Search on the Web (2)

Ranked results

Score: 3.918978454378219E-21

"politician" —means—> Overseas Chinese politicians
"scientist" —means—> Canadian political scientists
Adrienne Clarkson —isMarriedTo—> Stephen Clarks
Adrienne Clarkson —type—> Overseas Chinese poli
Stephen Clarkson —type—> Canadian political scien

- $@politician = Overseas Chinese politicians
- $@scientist = Canadian political scientists
- $X = Adrienne Clarkson
- $Y = Stephen Clarkson

Score: 1.1779824132840814E-24

"politician" —means—> Cuban politicians
"scientist" —means—> Cuban scientists
Ra&#65533;l Castro —isMarriedTo—> Vilma Esp&#65533;n
Ra&#65533;l Castro —type—> Cuban politicians
Vilma Esp&#65533;n —type—> Cuban scientists

- $@politician = Cuban politicians
- $@scientist = Cuban scientists
- $X = Ra&#65533;l Castro
- $Y = Vilma Esp&#65533;n

Query:
politicians who are married to scientists ?

?x isa politician .
?x isMarriedTo ?y .
?y isa scientist

Results (3):
[Adrienne Clarkson, Stephen Clarkson],
[Raúl Castro, Vilma Espín],
[Jeannemarie Devolites Davis, Thomas M. Davis]

mpi

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

Knowledge Search on the Web (3)

MEDIE — See *what causes cancer?* http://www-tsujii.is.s.u-tokyo.ac.jp/medie/

MEDIE is a demo system presented by Tsujii Laboratory

Semantic Search | **Keyword Search** | **GCL Search**

subject	verb	object (disease)			
	inhibit	hiv	Search!	Clear	Help

Results **1-50** for **inhibit hiv** »Show next »Show query 191.41 seconds (32.38% finished)

1. PMID: 16941340 »XML

   Gonococcal lipooligosaccharide suppresses HIV infection in human primary macrophages through induction of innate immunity .

2. PMID: 16815309 »XML

   AMD3100 has been shown to be a CXCR4 receptor antagonist , and to block HIV infection of T-tropic , X4-using , virus in vitro and in vivo .

3. PMID: 16574767 »XML

   Naltrexone inhibits alcohol-mediated enhancement of HIV infection of T lymphocytes .

4. PMID: 16574767 »XML

   We determined whether naltrexone , an opioid receptor antagonist widely used in the treatment of alcoholism , inhibits alcohol-mediated enhancement of HIV infection of T cells .

5. PMID: 16648264 »XML

   This finding suggests rhTRIM5alpha restricts HIV infection in two distinct phases : (i) altering the normal passage of the reverse-transcribing viral genome to the nucleus and (ii) targeting the reverse transcription complex to be disrupted by the proteasome .

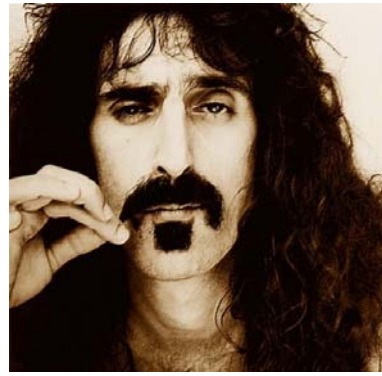
Take-Home Message

If music was invented
20 years ago
[when the Web was created],
we'd all be playing
one-string instruments.

(Udi Manber
VP Engineering
Google)

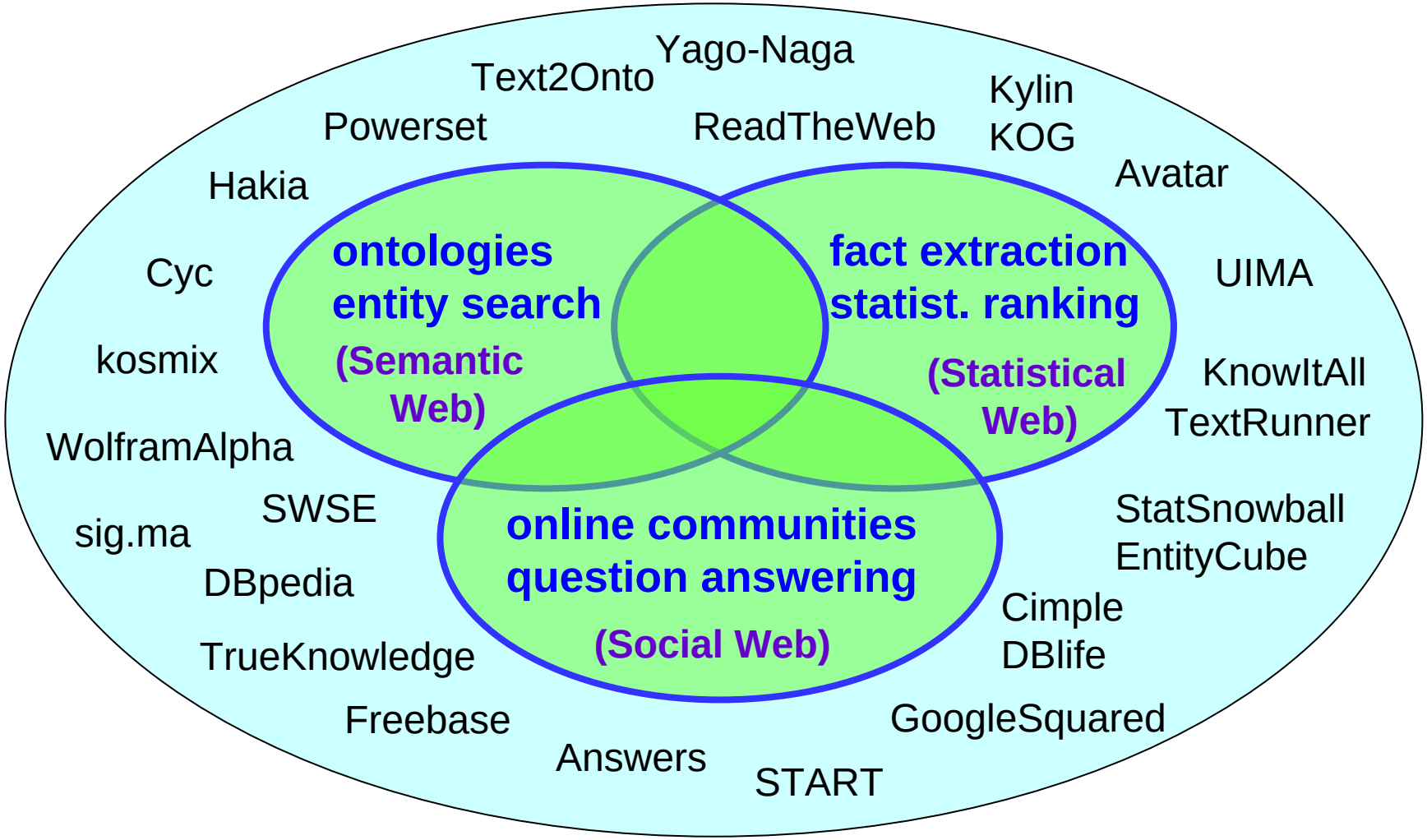
Information is not Knowledge.
Knowledge is not Wisdom.
Wisdom is not Truth
Truth is not Beauty.
Beauty is not Music.
Music is the best.

(Frank Zappa
jazz&rock musician
1940 – 1993)



→ **extract facts from Web sources**
→ **organize them in an automatically built knowledge base**
→ **answer questions in terms of entities and relations**

Related Work



Outline

- ✓ What and Why
- ★ Building a Large Knowledge Base
- ★ Consistent Growth of the Knowledge Base
- ★ Adding Multimodal Knowledge
- ★ Challenges: Scope, Scale, Robustness

Information Extraction (IE): Text to Relations

Max Planck
The Nobel Prize in Physics 1918

Max Karl Ernst Ludwig Planck was born in Kiel, Germany, on April 23, 1858, the son of Julius Wilhelm and Emma (*née* Patzig) Planck.

Planck studied at the Universities of Munich and E where his teachers included Kirchhoff and Helmho and received his doctorate of philosophy at Munic He was Privatdozent in Munich from 1880 to 1885 Associate Professor of Theoretical Physics at Kiel until 1889, in which year he succeeded Kirchhoff as Professor at Berlin University, where he remained until his ret Afterwards he became President of the Kaiser W for the Promotion of Science, a post he held unti

He was also a gifted pianist and is sa considered music as a career.
Planck was twice married. Upon his a to Associate Professor in his native to he married a friend of his childhood, I in 1909. He remarried her cousin Marga von Hös Three of his children died young, leaving him wit

bornOn (Max Planck, 23 April 1858) [0.99]
bornIn (Max Planck, Kiel) [0.9]
type (Max Planck, physicist) [0.9]
advisor (Max Planck, Kirchhoff) [0.6]
advisor (Max Planck, Helmholtz) [0.6]
AlmaMater (Max Planck, TU Munich) [0.5]
plays (Max Planck, piano) [0.7]
spouse (Max Planck, Marie Merck) [0.9]
spouse (Max Planck, Marga Hösslin) [0.8]

Person	BirthDate	BirthPlace	...
Max Planck	4/23, 1858	Kiel	

IE builds data space (with uncertain data)
- **confidence < 1 (sometimes << 1)**
- **knowledge base from many sources**
- **high computational cost**

Marie Curie	Nobel Prize in Physics
Marie Curie	Nobel Prize in Chemistry

IE: combine NLP, pattern matching, statistical learning

IE for Knowledge Harvesting

- **YAGO knowledge base from Wikipedia infoboxes & categories and integration with WordNet taxonomy**
- **NAGA search on RDF graph with entity-relationship LM for ranking**

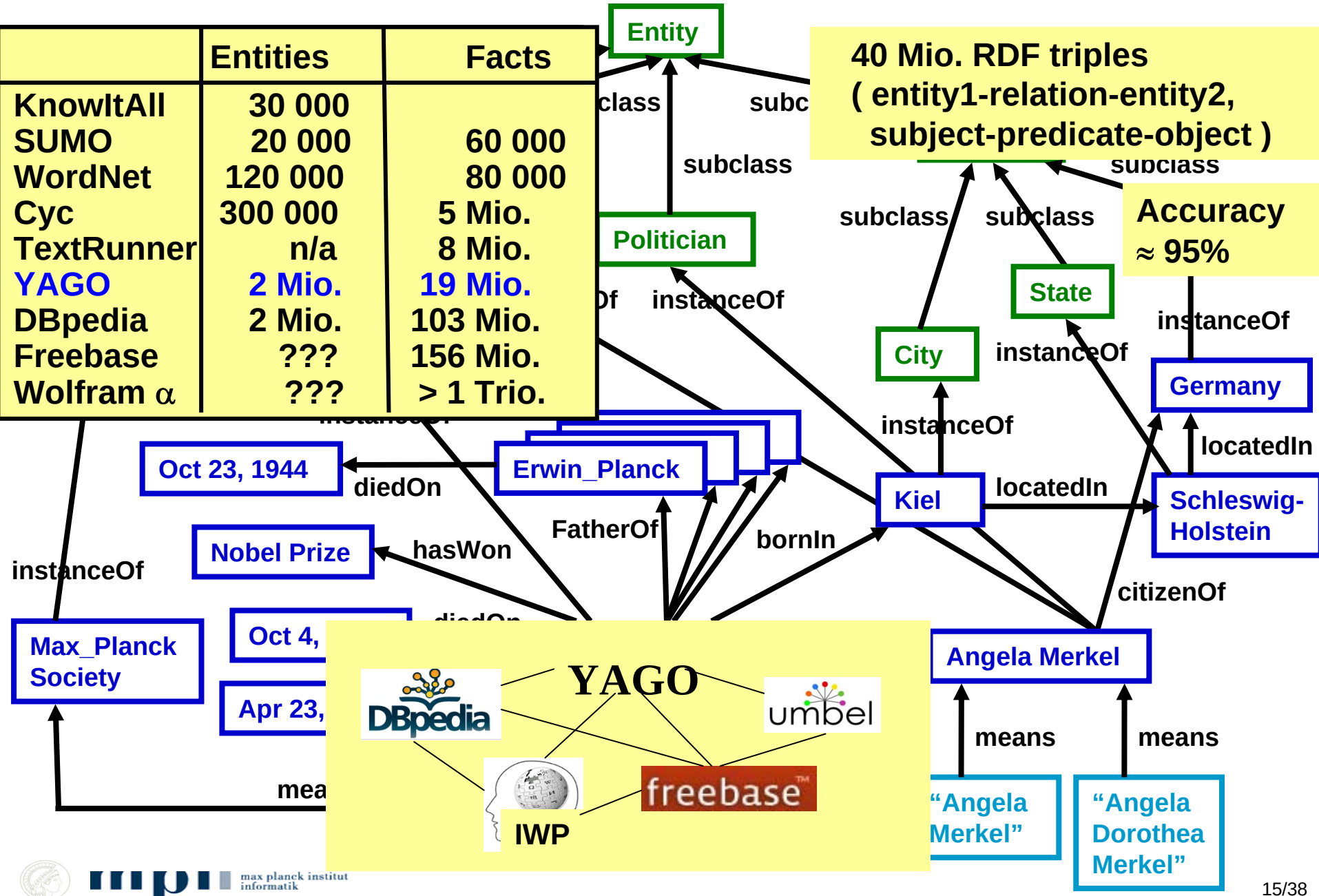
```
{{Infobox scientist
| name            = Max Planck
| birth_date      = {{birth date|1858|4|23|mf=y}}
| birth_place     = [[Kiel]], [[Holstein]]
| death_date      = {{death date and age|mf=yes|1947|10|4}}
| death_place     = [[Göttingen]], [[West Germany]]
| nationality     = [[Germany|German]]
| field           = [[Physics]]
| alma_mater      = [[Ludwig-Maximilians-Universität München]]
| work_institutions = [[University of Kiel]]<br />
        [[Humboldt University of Berlin|University of Berlin]]<br />
        [[University of Göttingen]]<br />
```

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-Maximilians-Universität München
Doctoral advisor	Alexander von Brill
Doctoral students	Gustav Ludwig Hertz Erich Kretschmann Walther Meißner

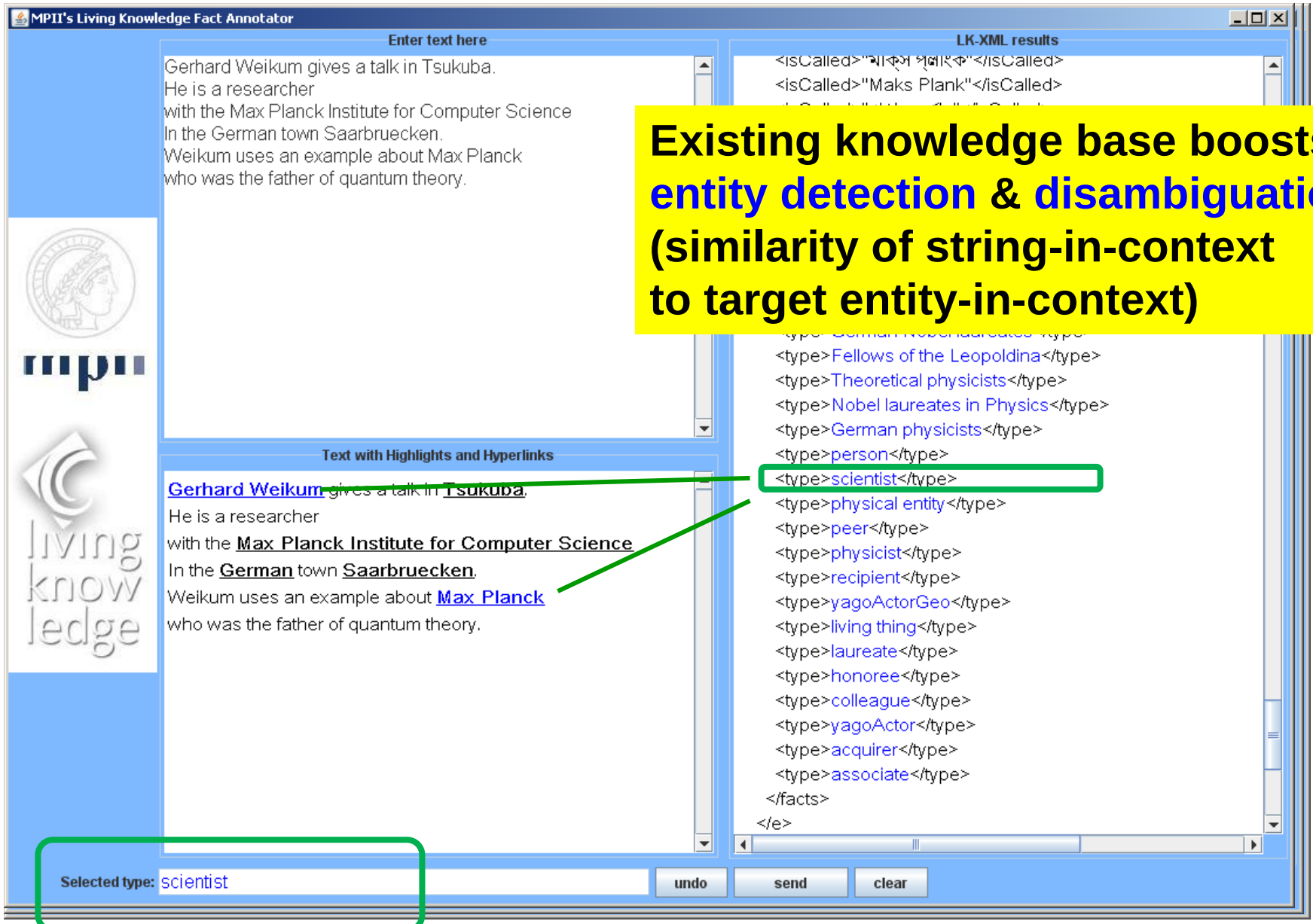
Categories: 1858 births | 1947 deaths | German Nobel laureates | German physicists | Members of the Pontifical Academy of Sciences | Members of the Prussian Academy of Sciences | Nobel laureates in Physics | Recipients of the Copley Medal | People from Kiel | People from the Province of Schleswig-Holstein | 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 | Columbia University faculty | German Christians | Religion and science | Fellows of the Leopoldina

YAGO Knowledge Base (F. Suchanek et al.: WWW'07)

	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	2 Mio.	19 Mio.
DBpedia	2 Mio.	103 Mio.
Freebase	???	156 Mio.
Wolfram α	???	> 1 Trio.



Leveraging YAGO for Entity Extraction

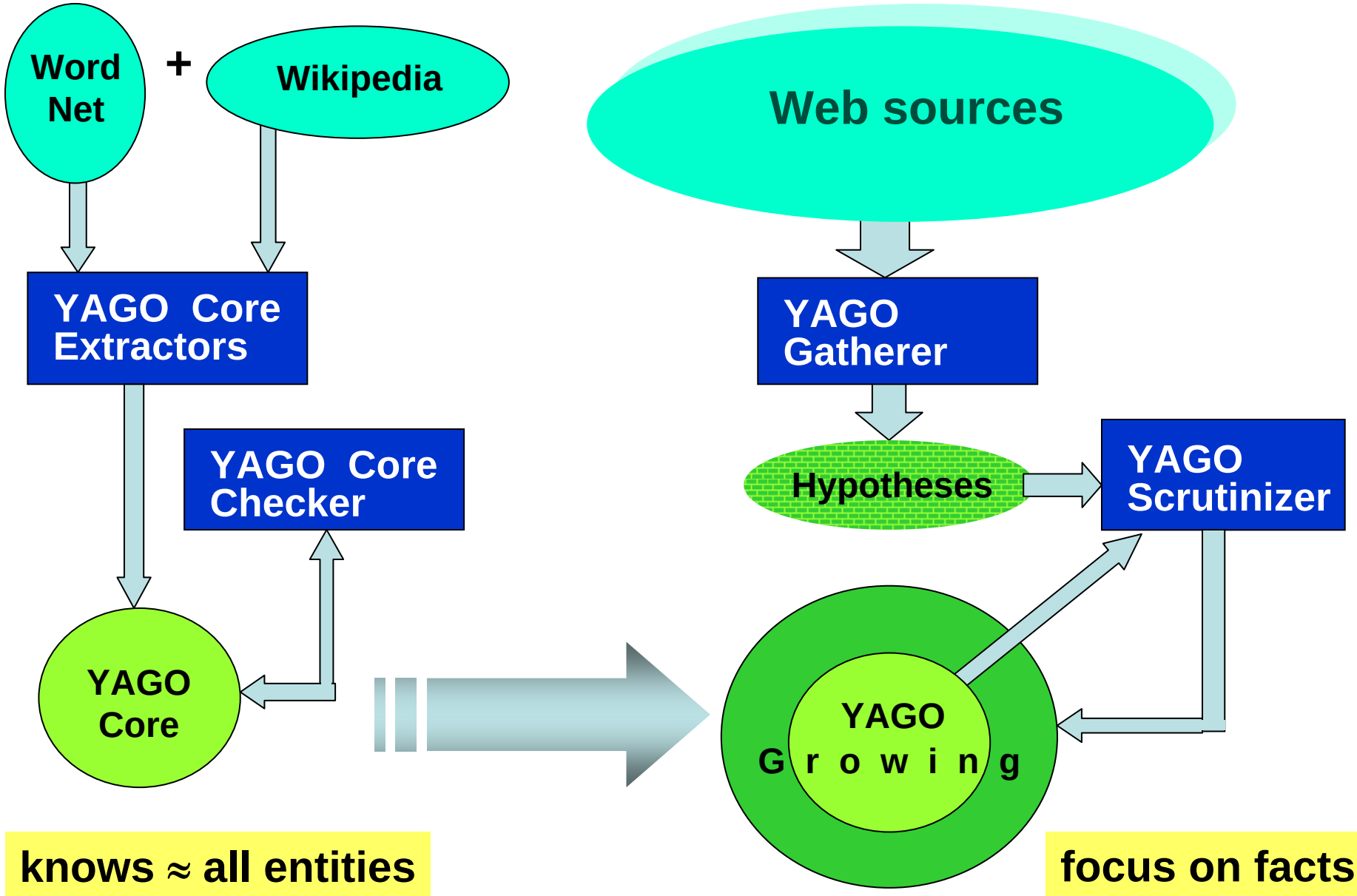


Existing knowledge base boosts entity detection & disambiguation (similarity of string-in-context to target entity-in-context)

Outline

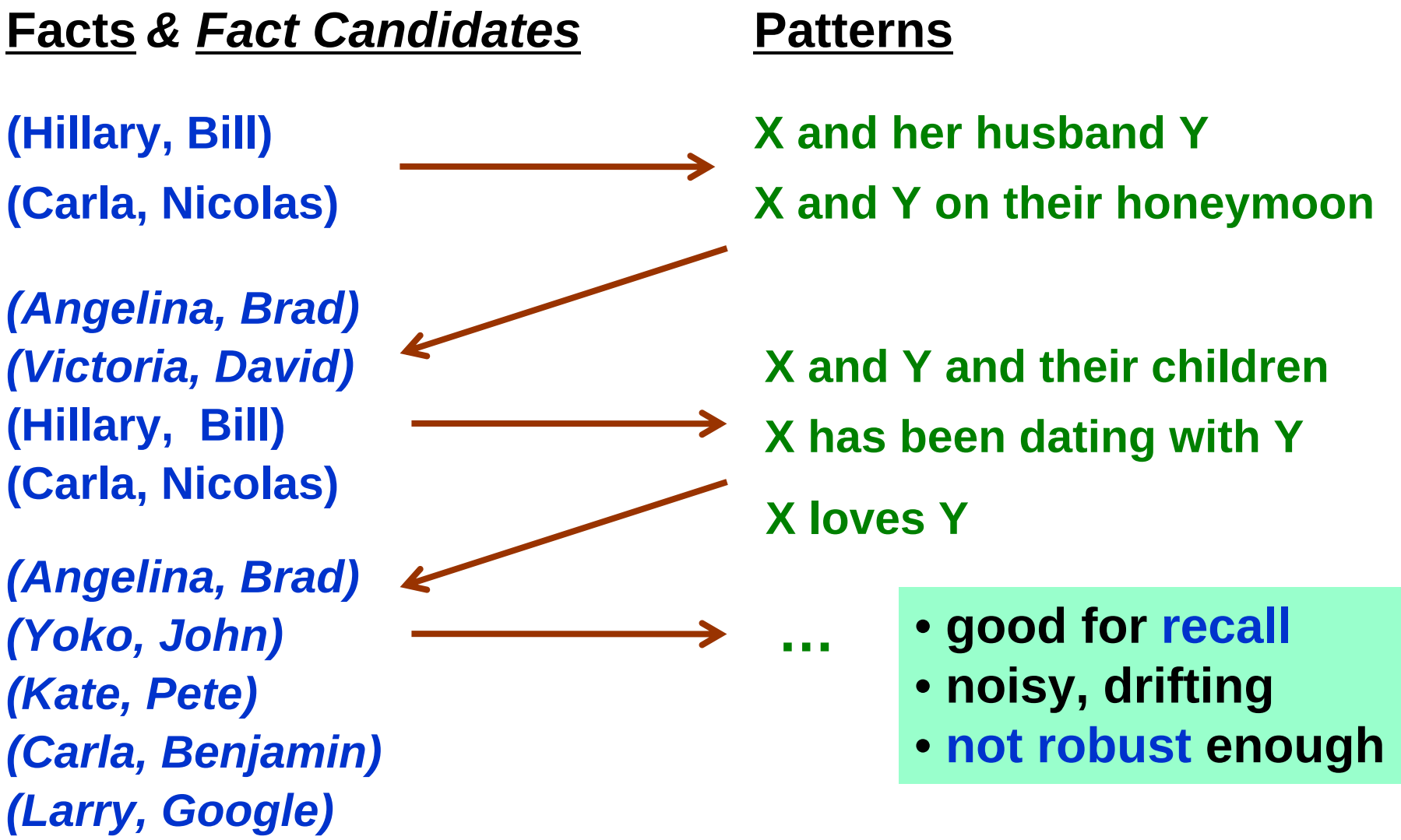
- ✓ What and Why
- ✓ Building a Large Knowledge Base
- ★ Consistent Growth of the Knowledge Base
- ★ Adding Multimodal Knowledge
- ★ Challenges: Scope, Scale, Robustness

Growing the Knowledge Base



Pattern-Based Harvesting

(Dipre, Snowball, Text2Onto, Leila, StatSnowball, etc.)



SOFIE: Self-Organizing Framework for IE

(F. Suchanek et al.: WWW'09)

Integrate methods:

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

  seeds → patterns → facts → patterns → ...

  (Hillary, Bill) → X and her husband Y → (Carla, Nicolas), (Carla, Mick) →
- declarative **rule-based IE** with **constraints**

  functional dependencies: *marriedTo is a function*

  inclusion dependencies: *presidentOf ⊆ citizenOf*

Address problems:

- pattern selection *(„and her husband", „has been dating", ...)*
- reasoning on mutual consistency of facts
- entity disambiguation *(„Merkel" → AngelaMerkel, MaxMerkel, ...;*

  „MPI" → MaxPlanckInstitute, MessagePassingInterface)

Unified solution by Weighted Max-Sat solver

(high accuracy and much faster than MCMC for prob. graphical models)

SOFIE Example

Facts

Spouse (HillaryClinton, BillClinton)

Spouse (CarlaBruni, NicolasSarkozy)

Patterns

occurs (*X and her husband Y*, Hillary, Bill) [100]

occurs (*X Y and their children*, Hillary, Bill) [40]

occurs (*X and her husband Y*, Victoria, David) [60]

occurs (*X dating with Y*, Rebecca, David) [20]

occurs (*X dating with Y*, Victoria, Tom) [10]

Hypotheses

Spouse (Victoria, David)

Spouse (Rebecca, David)

Spouse (Victoria, Tom)

expresses (*and her husband*, Spouse)

expresses (*and their children*, Spouse)

expresses (*dating with*, Spouse)

Clauses

$\forall x,y,z,w: R(x,y) \wedge R(x,z) \Rightarrow y=z$ (*at* $\neg R(x,y) \vee \neg R(x,z)$)

$\forall x,y,z,w: R(x,y) \wedge R(w,y) \Rightarrow x=w$ (*at* $\neg R(x,y) \vee \neg R(w,z)$)

...

$\forall x,y: R(x,y) \Rightarrow R(y,x)$

$\forall x,y: R(x,y) \Rightarrow R(y,x)$

$\forall x,y: R(x,y) \Rightarrow R(y,x)$

$\forall x,y: R(x,y) \Rightarrow R(y,x)$

occurs (husband, Victoria, David) $\wedge$ expresses (husband, Spouse)

... $\Rightarrow$ Spouse (Victoria, David)

occurs (dating, Rebecca, David) $\wedge$ expresses (dating, Spouse)

$\forall p,x,y: \text{occurs}(p,x,y) \wedge \text{expresses}(p,R) \Rightarrow R(x,y)$

$\forall p,x,y: \text{occurs}(p,x,y) \wedge \text{expresses}(p,R) \Rightarrow R(x,y)$

$\forall p,x,y: \text{occurs}(p,x,y) \wedge \text{expresses}(p,R) \Rightarrow R(x,y)$

$\forall p,x,y: \text{occurs}(p,x,y) \wedge \text{expresses}(p,R) \Rightarrow R(x,y)$

occurs (husband, Victoria, David) $\wedge$ expresses (Victoria, David)

$\Rightarrow$ Spouse (Rebecca, David)

...

occurs (husband, Victoria, David) $\wedge$ expresses (Victoria, David)

$\forall p,x,y: \text{occurs}(p,x,y) \wedge \text{expresses}(p,R) \Rightarrow R(x,y)$

$\forall p,x,y: \text{occurs}(p,x,y) \wedge \text{expresses}(p,R) \Rightarrow R(x,y)$

$\forall p,x,y: \text{occurs}(p,x,y) \wedge \text{expresses}(p,R) \Rightarrow R(x,y)$

$\forall p,x,y: \text{occurs}(p,x,y) \wedge \text{expresses}(p,R) \Rightarrow R(x,y)$

$\Rightarrow$ expresses (husband, Spouse)

...

Reasoning on Hypotheses by Weighted-Max-Sat Solver

- **Clauses** (propositional logic formulae consisting of conjunctions of disjunctions of positive or negative literals) **connect facts, patterns, hypotheses, constraints**
- Treat **hypotheses** (literals) **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!
  $\rightarrow$ **NP-hard, but good approximation algorithms**

SOFIE Example

Spouse (HillaryClinton, BillClinton)

Spouse (CarlaBruni, NicolasSarkozy)

occurs (*X and her husband Y*, Hillary, Bill) [100]
occurs (*X Y and their children*, Hillary, Bill) [40]
occurs (*X and her husband Y*, Victoria, David) [60]
occurs (*X dating with Y*, Rebecca, David) [20]
occurs (*X dating with Y*, Victoria, Tom) [10]

Spouse (Victoria, David) A
Spouse (Rebecca, David) B
Spouse (Victoria, Tom) C

expresses (*and her husband*, Spouse) D
expresses (*and their children*, Spouse) E
expresses (*dating with*, Spouse) F

Spouse (Victoria, David) $\Rightarrow$ $\neg$ Spouse (Rebecca, David) $\neg A \vee \neg B$
Spouse (Victoria, David) $\Rightarrow$ $\neg$ Spouse (Victoria, Tom) $\neg A \vee \neg C$
...
occurs (husband, Victoria, David) $\wedge$ expresses (husband, Spouse)
$\Rightarrow$ Spouse (Victoria, David) $\neg 1 \vee \neg D \vee A$
occurs (dating, Rebecca, David) $\wedge$ expresses (dating, Spouse)
$\Rightarrow$ Spouse (Rebecca, David) $\neg 1 \vee \neg F \vee B$
...

**Wanted: truth assignment for A, B, C, ...
with maximal total weight of satisfied clauses**

...

Consistent Growth of Knowledge

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

- highly related to methods based on Markov Logic Networks, joint learning with constraints
- but SOFIE does not compute joint probability distribution, much faster than Monte-Carlo Markov-Chain methods

Outline

- ✓ What and Why
- ✓ Building a Large Knowledge Base
- ✓ Consistent Growth of the Knowledge Base
- ★ Adding Multimodal Knowledge
- ★ Challenges: Scope, Scale, Robustness

What's Wrong With This?

navigation
- Main page
- Contents
- Featured content
- Current events
- Random article

search

Go Search

interaction
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

languages
- Deutsch
- Polski
- Русский

Max Planck Institute for Computer Science

From Wikipedia, the free encyclopedia

Coordinates: 49°15'28"N 7°2'44"E

The **Max Planck Institute for Informatics** (MPII) is devoted to cutting-edge research in computer science with a focus on algorithms and their applications in a broad sense. The research ranges from foundations (algorithms and complexity, programming logics) to a variety of application domains (computer graphics, geometric computation, constraint solving, computational biology). It is part of the Max-Planck-Gesellschaft, Germany's largest society for fundamental research. The objective of the institute is to achieve an effect equally through scientific publications, software systems and by training the next academic generation. The research institutes of the Max Planck Society have a national and international reputation as "Centres of Excellence" for pure research.

The institute consists of five departments and two research groups:

- The Algorithms and Complexity Department is headed by Prof. Dr. Kurt Mehlhorn,
- The Programming Logics Department was headed by Prof. Dr. Harald Ganzinger († June 3, 2004)
- The Department Computational Biology and Applied Algorithmics is headed by Prof. Dr. Thomas Lengauer, Ph.D.
- The Computer Graphics Department is headed by Prof. Dr. Hans-Peter Seidel
- The Databases and Information Systems Department is headed by Prof. Dr. Gerhard Weikum
- Research Group Automation of Logic is headed by Dr. Christoph Weidenbach

Among other things, numerous awards indicate the success of the institute since it was founded. Professor Kurt Mehlhorn (in 1986) and Professor Hans-Peter Seidel (in 2003) received the Gottfried Wilhelm Leibniz Prize, in 2003 Professor Thomas Lengauer received the Konrad-Zuse-Medal, and in 2004 Professor Harald Ganzinger received the Herbrand Award.

In the context of cutting edge research in Germany, the institute joins the Max Planck Institute for Software Systems (MPI-SWS), the German Research Centre for Artificial Intelligence (DFKI) and the entire Computer Science department at the Saarland University in being involved in the Internationales Begegnungs- und Forschungszentrum für Informatik. This also offers many possibilities for obtaining further scientific qualifications.

Research School

The International Max Planck Research School for Computer Science (IMPRS-CS) is the graduate school of the MPII and the MPI-SWS. It was founded in 2000 and offers a fully funded PhD-Program (in cooperation with Saarland University). Dean is Prof. Dr. Gerhard Weikum.

External links

- Max Planck Institute for Informatics Homepage
- International Meeting Place and Research Centre for Computer Science in Dagstuhl Castle
- International Max Planck Research School for Computer Science

Multimodal Knowledge

type **(MPI, ScientificOrganization)**
fullName **(MPI, Max Planck Institute for Informatics)**
inField **(MPI, Computer Science)**
partOf **(MPI, Max Planck Society)**
foundingDirector **(MPI, Kurt Mehlhorn)**

or

K2 (Knowledge Kaleidoscope): Photos of Named Entities

Challenges:

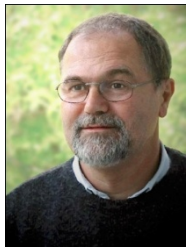
- **Long Tail:**
  non-famous but notable entities

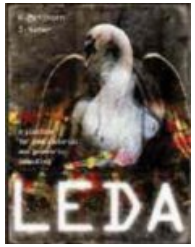
- **Diversity:**
  variety of different views, different ages, etc.

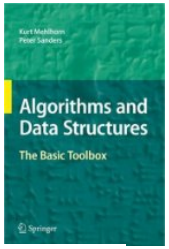
- **Scale:**
  all entities with Wikipedia article (known to YAGO)
  all entities mentioned in Wikipedia articles

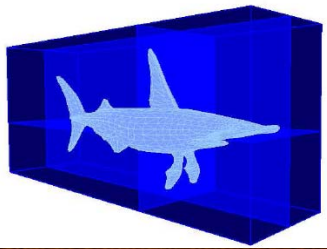
Gathering & Ranking Photos by Image Search Engines

q: Kurt Mehlhorn

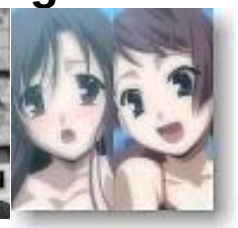


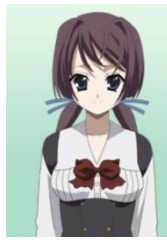


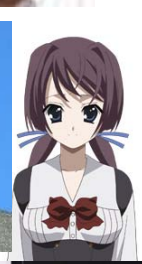




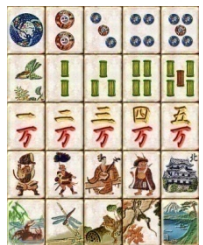
q: Kitsuregawa San



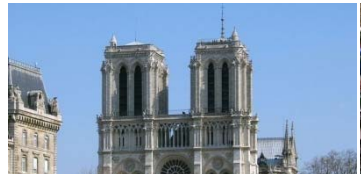




q: Fujiyama



q: Notre Dame des Cyclistes



Knowledge-based Photo Harvesting

(Bilyana Taneva et al.: WSDM 2010)

- **generate expanded queries q_i for entity e**
  using affiliation, knownFor, wonAward, etc.;
  e.g.: Kitsuregawa University Tokyo, Kitsuregawa Hash Join, Kitsuregawa Sigmod Award, etc.

- **run queries and retrieve photos p from top-k results (k=100)**

- **combine results by rank-based weighted voting (learn weights w_i from training entities)**

$$score(p,e) = \sum_{i=1}^{m} w_i \left(k + 1 - rank(p, q_i(e))\right) / k$$

- **consider visual similarities (using SIFT)**

$$score(p,e) = \sum_{i=1}^{m} w_i \sum_{x \in top\,k(q_i(e))} sim(p,x) \frac{k + 1 - rank(x, q_i(e))}{k}$$

- **rank results, cluster by similarity**

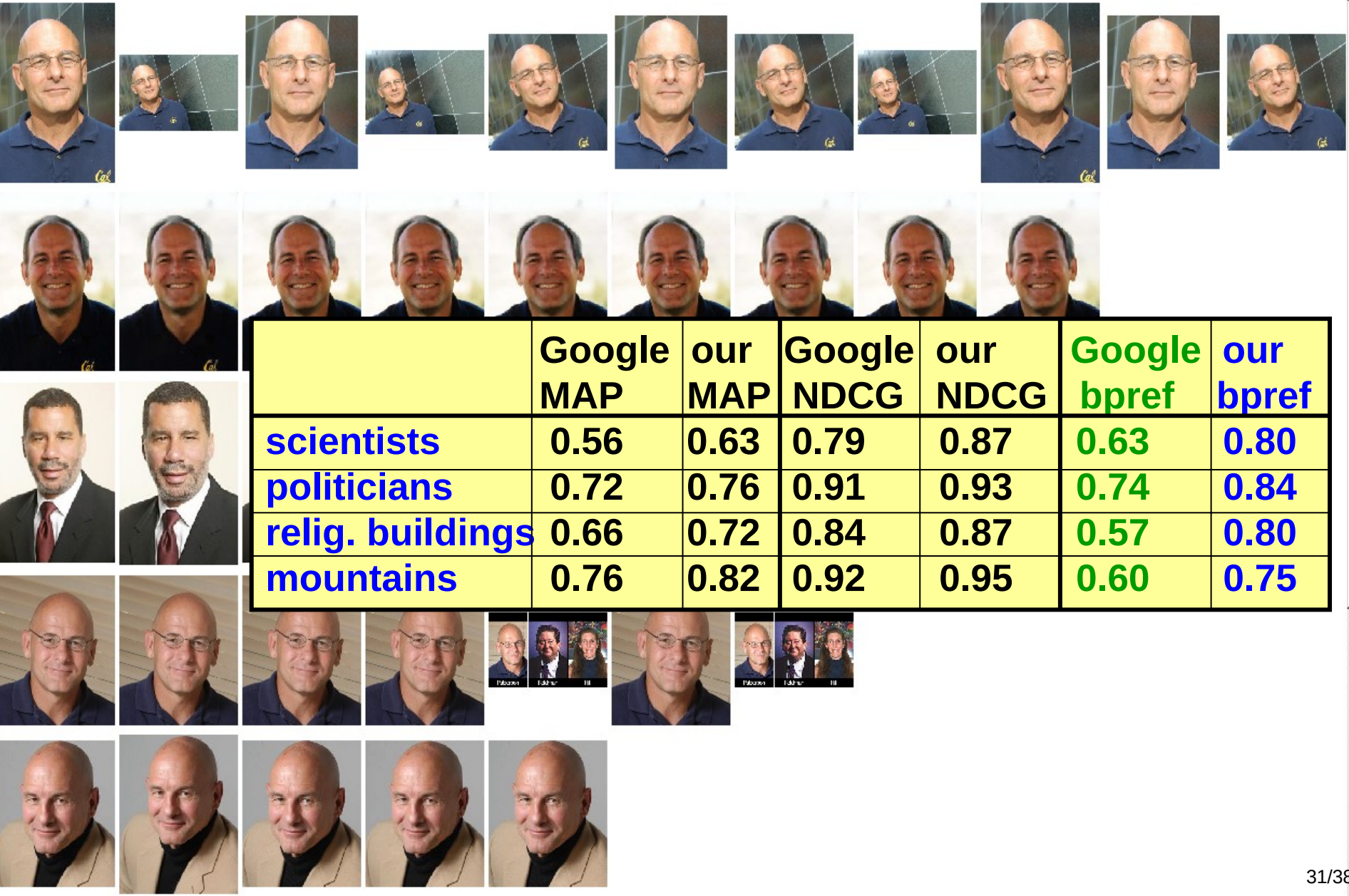
David Patterson

David Patterson Berkeley

David Patterson RISC

David Patterson ACM

our method



	Google MAP	our MAP	Google NDCG	our NDCG	Google bpref	our bpref
scientists	0.56	0.63	0.79	0.87	0.63	0.80
politicians	0.72	0.76	0.91	0.93	0.74	0.84
relig. buildings	0.66	0.72	0.84	0.87	0.57	0.80
mountains	0.76	0.82	0.92	0.95	0.60	0.75

Outline

- ✓ What and Why
- ✓ Building a Large Knowledge Base
- ✓ Consistent Growth of the Knowledge Base
- ✓ Adding Multimodal Knowledge
- ★ Challenges: Scope, Scale, Robustness

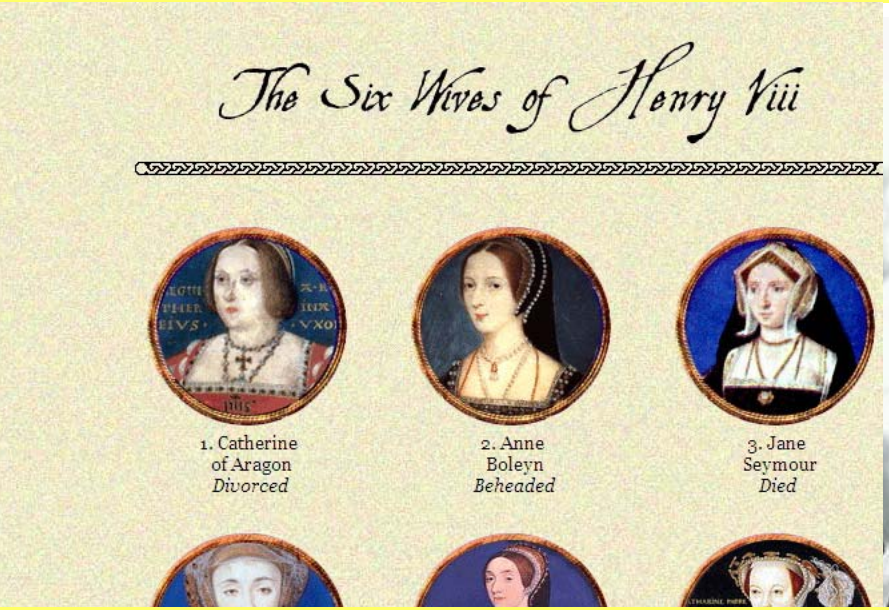
Challenges: Scope, Scale, Robustness

- **Temporal Knowledge:**
  temporal validity of all facts ***(spouses, CEO's, etc.)***
- **Multilingual Knowledge:** via cross-lingual Wikipedia links etc.
  Rome ↔ Roma ↔ Rom ↔ Řím ↔ रोम ↔ ರೋಮ್
  Moment (Stochastik) → Moment (math) → Momento estándar
- **Multimodal Knowledge:** photos & videos of
  entities ***(people, landmarks, etc.)*** and
  facts ***(weddings, award ceremonies, soccer matches, etc.)***
- **Active Knowledge:** on-demand coupling with Web Services
  for „live" facts ***(ratings, charts, sports feeds, etc.)***
- **Diverse Knowledge:** diversity of facts/facets/views of entities
- **Scalable Knowledge Gathering:**
  high-quality extraction at the rate at which
  news, publications, Wikipedia updates are produced

Scale: 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



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:** $birthdate + \Delta < marriage < divorce$

Robustness: Patterns & Reasoning

Easy to optimize either one of recall or precision alone:

- **recall → pattern-based harvesting (fast & furious IE)**
- **precision → rigorous consistency reasoning**

Challenge lies in reconciling both recall & precision

Some ideas:

- **richer patterns, richer pattern statistics**
- **negative seed facts**
- **more and richer constraints**
- **efficiency & scalability: (map-reduce) parallelism (some parts embarrasingly parallel, others very difficult)**

Scope: Temporal Knowledge

Nicolas Sarkozy

From Wikipedia, the free encyclopedia

Early life

During Sarkozy's childhood, his father refused to give his wife's family any financial help, even though he had fo
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 Arrondissemen. The family later moved o Neuilly-sur-Seine, one of t
wealthiest communes of the Île-de-France *région* immediately west of the 17th Arrondissement just outside of
According to Sarkozy, his staunchly Gaullist grandfather was more of an influence on him than his father, whor
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
years, he felt inferior in relation to his wealthier classmates.[19] "What made me who I am now is the sum of all
humiliations suffered during childhood", he said later.[19]

Education

Sarkozy was enrolled in the *Lycée*
arrondissement, where he failed hi
Catholic school in the 17th arrondi
obtained his *baccalauréat* in 1973
Private law, and later with a DEA
student movement and was still a
right-wing student organization, in
cleaner.[21] After graduating, he en
due to an insufficient command of
business and family law,[23] and w

private
etheless
Master in
May '68
oined the
Air Force
duate
izing in

Relationships

[edit]

Marie-Dominique Culioli

[edit]

Sarkozy married his first wife, Marie-Dominique Culioli, on 23 September 1982; her father was a pharmacist from Vico (a village north of Ajaccio, Corsica). They had two sons, Pierre (born in 1985), now a hip-hop producer,[27] and Jean (born in 1986) now a regional councillor in the city of Neuilly-sur-Seine, France. Sarkozy's best man was the prominent

Minister of the Budget

In office
29 March 1993 – 10 May 1995

Prime Minister Edouard Balladur

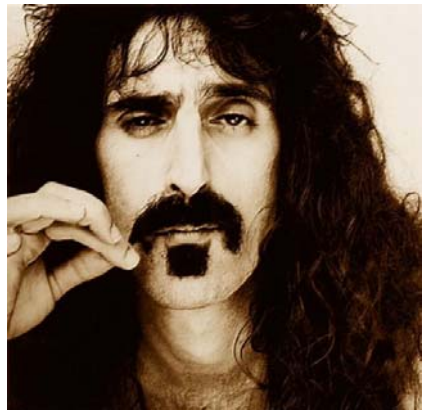
- **different resolutions**
- **missing dates**
- **relative dates**
- **adverbial phrases**
- **vague time periods**
- **temporal refinement**

extracting, aggregating, and reasoning on temporal scopes of facts from many sources is a major challenge

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)

Summary

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)

- **Distill entities & relations from Web pages to automatically build a large knowledge base**
- **knowledge (base) enables more (& better) knowledge**

Outlook: Knowledge Harvesting at Web Scale

Grand Challenge:

as literature, news & blogs are being produced,

- **„read“ everything, detect entities, extract relations,**
- **confirm old knowledge & obtain new knowledge**
  - **→ new facts**
  - **→ new relation types**
  - **→ temporal evolution of entities & facts**
  - **→ opinionated statements & diversity**
  - **→ multimodal footage**

Grand Opportunities:

machine-processable, comprehensive KB can enable or boost

- **semantic Web search: precise answers**
- **context-sensitive machine translation**
- **situation-aware human-computer dialogs**
- **machine reasoning and value-added knowledge services**

どうも有り難うございます

Domo Arigato Gozaimasu!