Semantic Knowledge Bases from Web Sources

IJCAI 2011 Tutorial

Hady W. Lauw[1], Ralf Schenkel[2], Fabian Suchanek[3], Martin Theobald[4], and Gerhard Weikum[4]

[1]Institute for Infocomm Research, Singapore
[2]Saarland University, Saarbruecken
[3]INRIA Saclay, Paris
[4]Max Planck Institute for Informatics, Saarbruecken

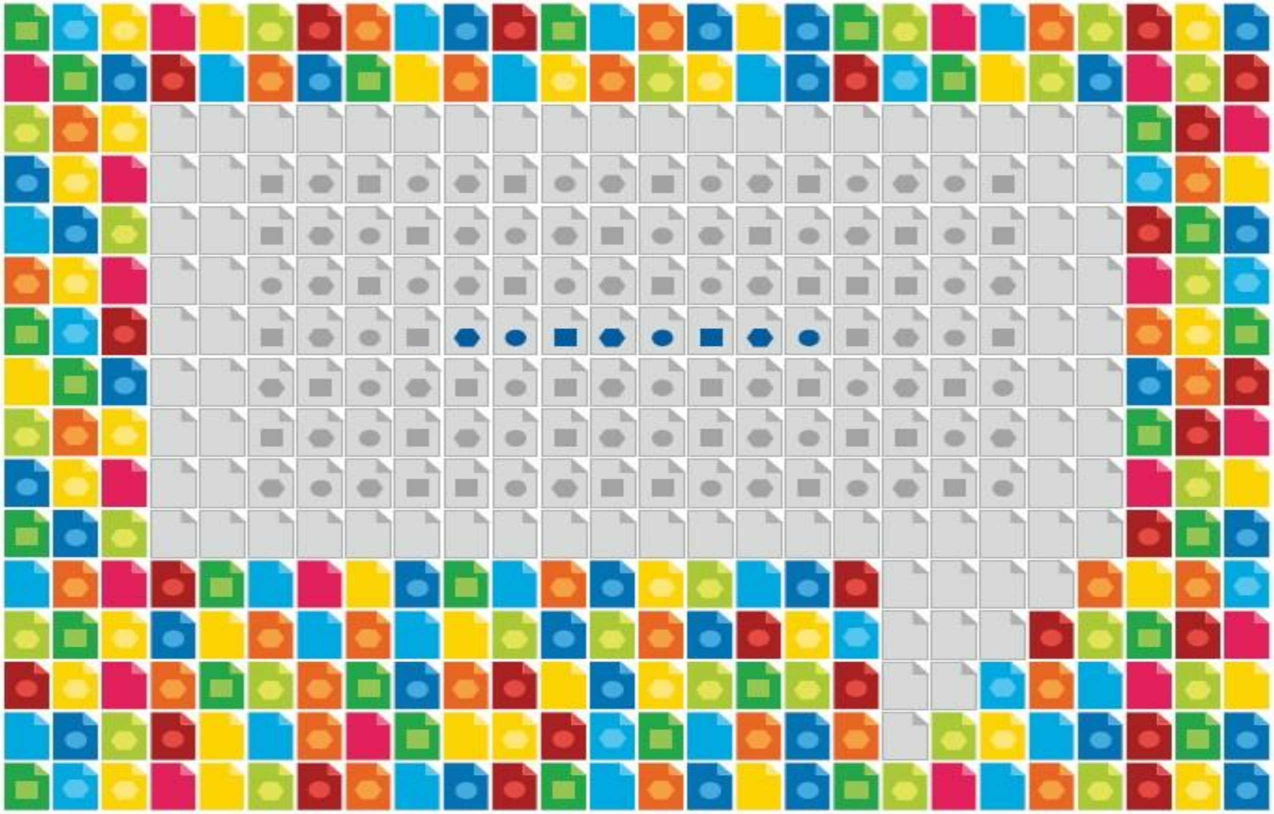
All slides for download…

http://www.mpi-inf.mpg.de/yago-naga/IJCAI11-tutorial/

Outline

- **Part I**
  - **Machine Knowledge & Intelligent Applications**
- **Part II**
  - **Knowledge Representation & Public Knowledge Bases**
- **Part III**
  - **Extracting Knowledge**
- **Part IV**
  - **Ranking and Searching**
- **Part V**
  - **Linked Data**
- **Part VI**
  - **Conclusion and Outlook**

Goal: Turn Web into Knowledge Base



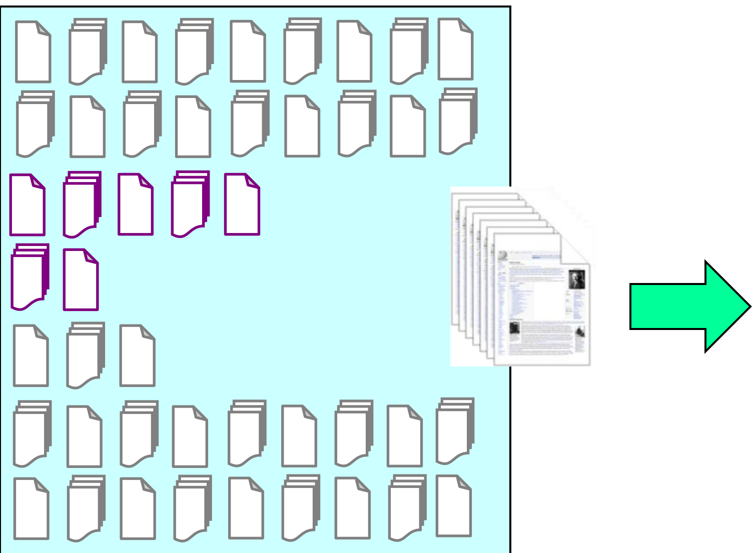
Source:
G. Weikum, G., Kasneci, M. Ramanath, F. Suchanek: DB & IR methods for knowledge discovery. Communications of the ACM 52(4), 2009

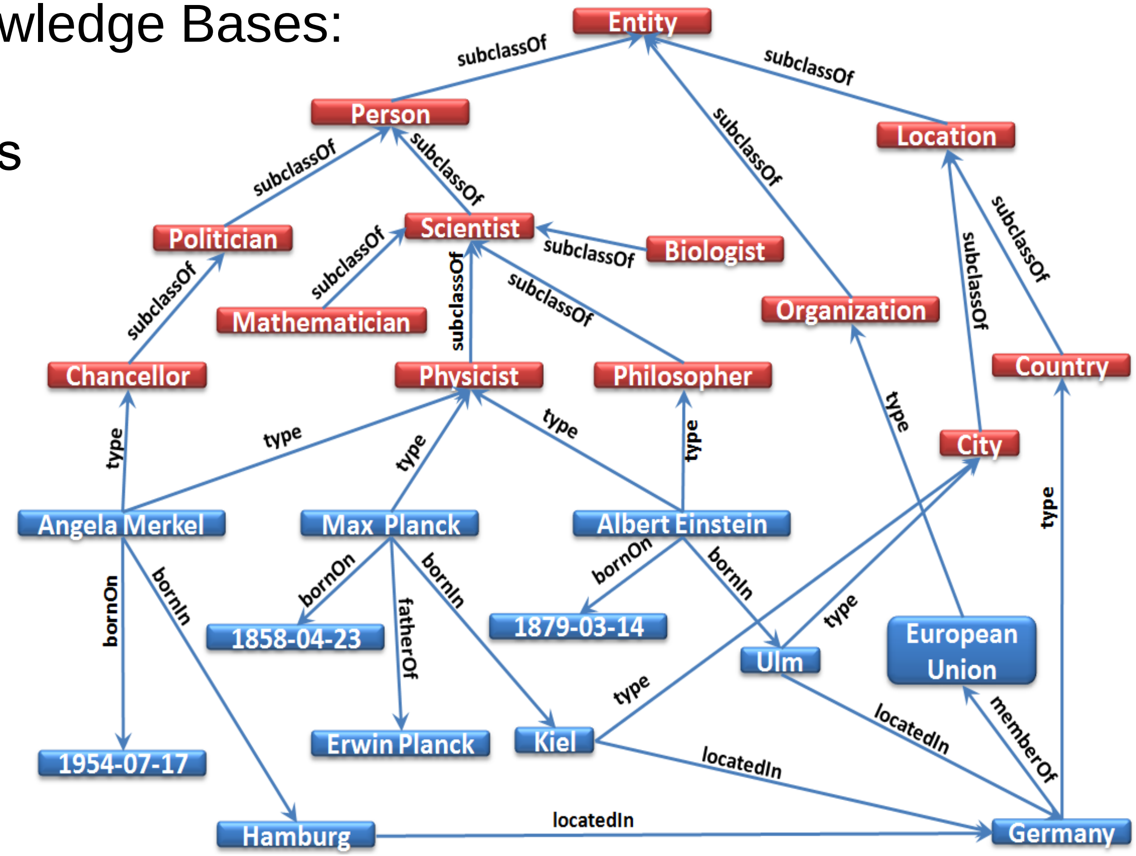
comprehensive DB of human knowledge
- **everything that Wikipedia knows**
- **everything machine-readable**
- **capturing entities, classes, relationships**

Approach: Knowledge Harvesting

Automatically Constructed Knowledge Bases:

- Mio's of individual entities
- 100 000's of classes/types
- 100 Mio's of facts
- 100's of relation types





Cyc SUMO TextRunner YAGO-NAGA DBpedia SIG.MA Semantic Information Mashup WikiNet IWP DBLife umbel WikiTaxonomy True Knowledge The Internet Answer Engine BETA Carnegie Mellon ReadTheWeb freebase WolframAlpha computational knowledge engine ProBase

Knowledge for Intelligence

- **entity recognition & disambiguation**
- **understanding natural language & speech**
- **knowledge services & reasoning for semantic apps**

  (e.g. deep question answering)
- **semantic search: precise answers to advanced queries**

  (by scientists, students, journalists, analysts, etc.)

★ **Swedish king's wife when Greta Garbo died?**

★ **FIFA 2010 finalists who played in a Champions League final?**

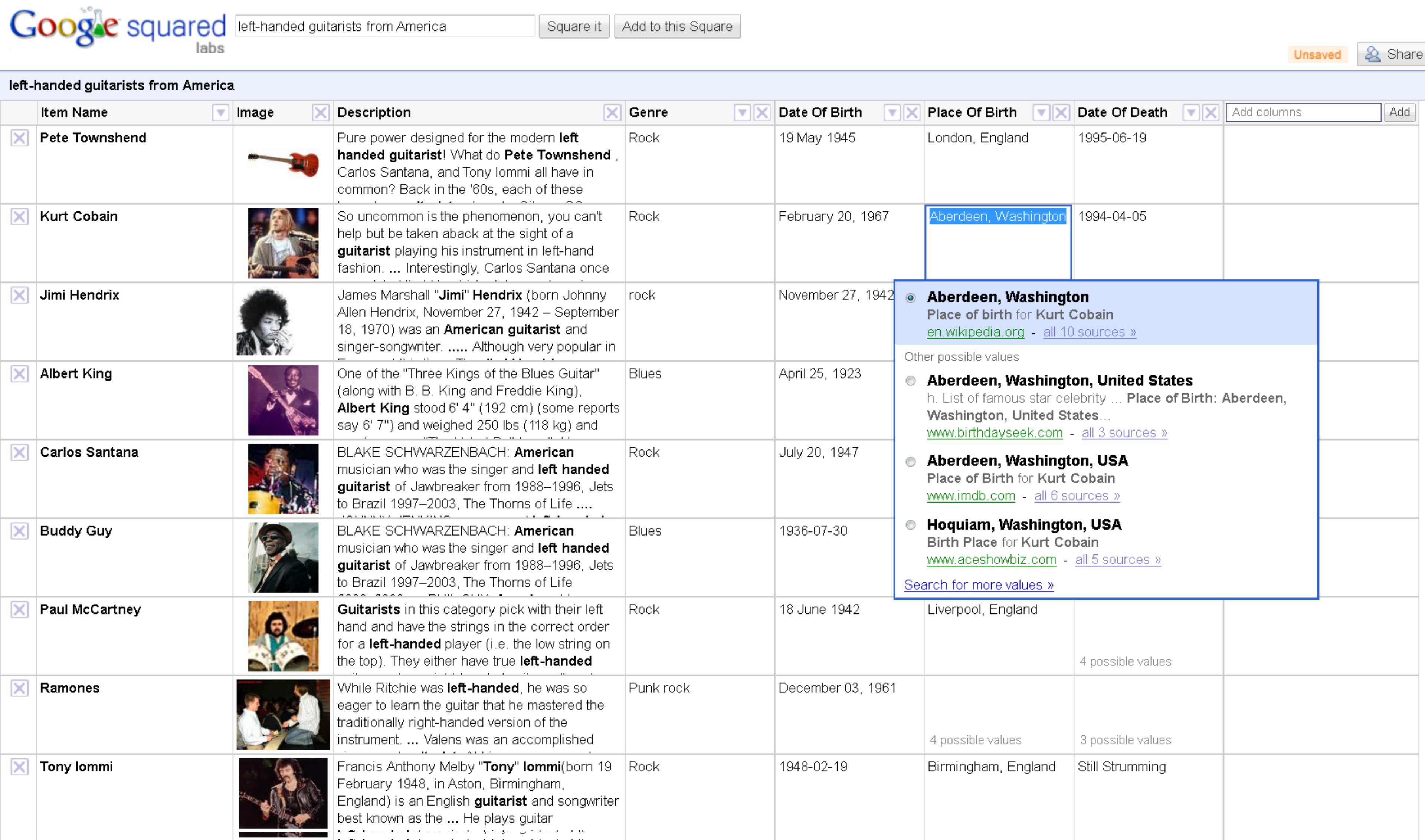
★ **Politicians who are also scientists?**

★ **Relationships between**
Max Planck, Angela Merkel, Jim Gray, and the Dalai Lama?

★ **Drugs for treating Alzheimer?**
Influenza vaccines for teens with high blood pressure?

...

Application 1: Semantic Queries on Web



Google squared labs

left-handed guitarists from America | Square it | Add to this Square

Unsaved | Share

left-handed guitarists from America

Item Name	Image	Description	Genre	Date Of Birth	Place Of Birth	Date Of Death
Pete Townshend		Pure power designed for the modern **left handed guitarist**! What do **Pete Townshend** , Carlos Santana, and Tony Iommi all have in common? Back in the '60s, each of these	Rock	19 May 1945	London, England	1995-06-19
Kurt Cobain		So uncommon is the phenomenon, you can't help but be taken aback at the sight of a **guitarist** playing his instrument in left-hand fashion. ... Interestingly, Carlos Santana once	Rock	February 20, 1967	Aberdeen, Washington	1994-04-05
Jimi Hendrix		James Marshall "**Jimi**" **Hendrix** (born Johnny Allen Hendrix, November 27, 1942 – September 18, 1970) was an **American guitarist** and singer-songwriter. ..... Although very popular in	rock	November 27, 1942		
Albert King		One of the "Three Kings of the Blues Guitar" (along with B. B. King and Freddie King), **Albert King** stood 6' 4" (192 cm) (some reports say 6' 7") and weighed 250 lbs (118 kg) and	Blues	April 25, 1923		
Carlos Santana		BLAKE SCHWARZENBACH: **American** musician who was the singer and **left handed guitarist** of Jawbreaker from 1988–1996, Jets to Brazil 1997–2003, The Thorns of Life	Rock	July 20, 1947		
Buddy Guy		BLAKE SCHWARZENBACH: **American** musician who was the singer and **left handed guitarist** of Jawbreaker from 1988–1996, Jets to Brazil 1997–2003, The Thorns of Life	Blues	1936-07-30		
Paul McCartney		**Guitarists** in this category pick with their left hand and have the strings in the correct order for a **left-handed** player (i.e. the low string on the top). They either have true **left-handed**	Rock	18 June 1942	Liverpool, England	4 possible values
Ramones		While Ritchie was **left-handed**, he was so eager to learn the guitar that he mastered the traditionally right-handed version of the instrument. ... Valens was an accomplished	Punk rock	December 03, 1961	4 possible values	3 possible values
Tony Iommi		Francis Anthony Melby "**Tony**" **Iommi**(born 19 February 1948, in Aston, Birmingham, England) is an English **guitarist** and songwriter best known as the ... He plays guitar	Rock	1948-02-19	Birmingham, England	Still Strumming

Add columns | Add

Aberdeen, Washington
Place of birth for Kurt Cobain
en.wikipedia.org - all 10 sources »

Other possible values

Aberdeen, Washington, United States
h. List of famous star celebrity ... Place of Birth: Aberdeen, Washington, United States...
www.birthdayseek.com - all 3 sources »

Aberdeen, Washington, USA
Place of Birth for Kurt Cobain
www.imdb.com - all 6 sources »

Hoquiam, Washington, USA
Birth Place for Kurt Cobain
www.aceshowbiz.com - all 5 sources »

Search for more values »

www.google.com/squared/

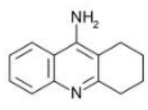
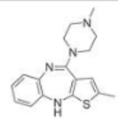
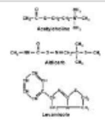
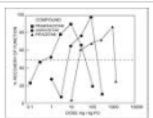
Application 1: Semantic Queries on Web

Google squared labs

drugs for treating Alzheimer | Square it | Add to this Square

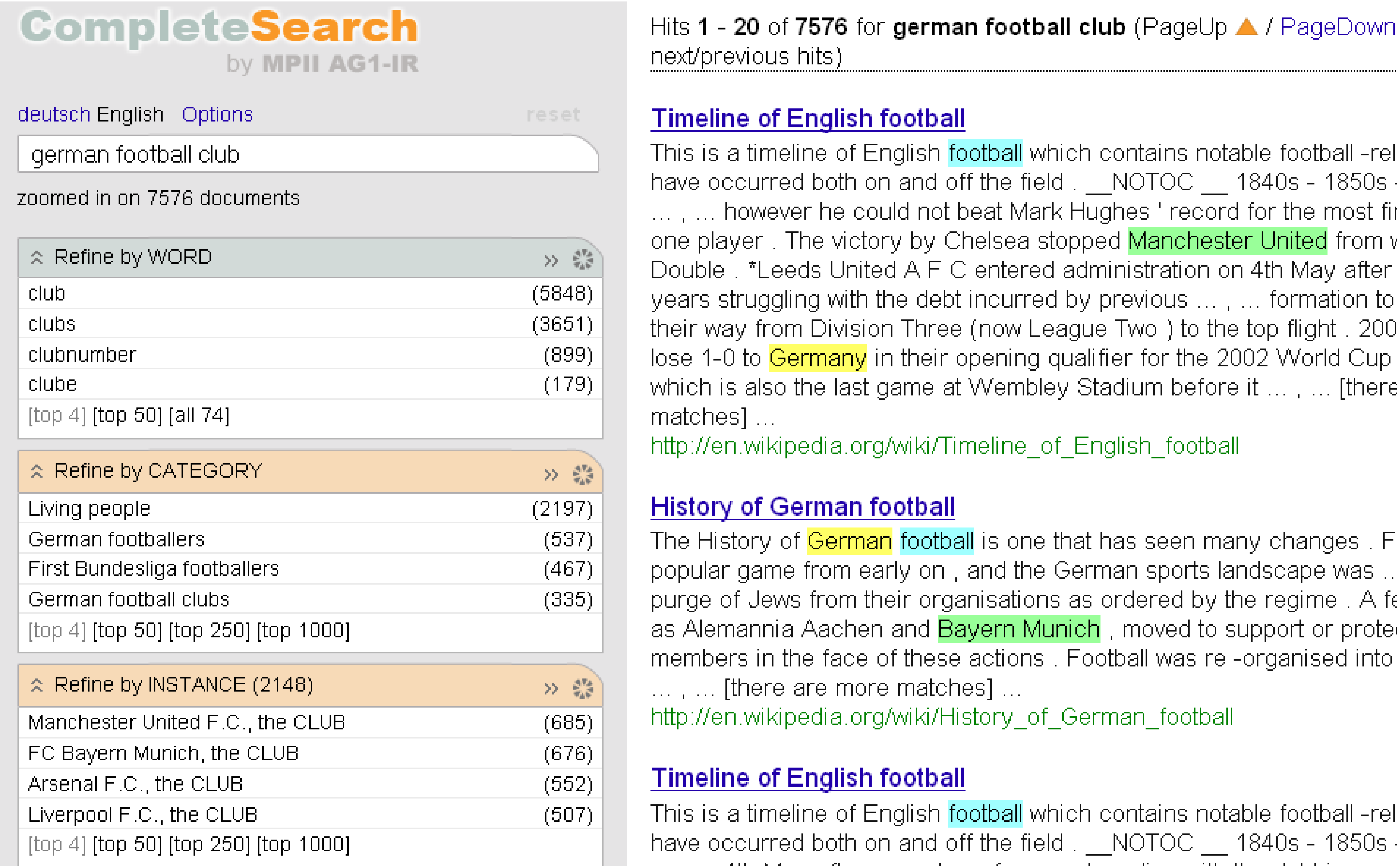
Did you mean: [drugs for treating ***Alzheimer's***]

drugs for treating Alzheimer

	Item Name	Image	Description	Cas Number	Formula	Half Life	Pubchem
	Tacrine		**Tacrine** is the first FDA approved **drug** for the treatment of **Alzheimer's** disease as safe and effective. The fear of toxicity has been exaggerated. Liver function testing could be	321-64-2	C13H14N2	2–4 hours	CID 1935
	Olanzapine		The Food and **Drug** Administration (FDA) has declied to approve Memantine (namenda) to treat mild **Alzheimer's**. **Olanzapine** (Zyprexa) **Olanzapine** (Zyprexa). Atypical Antipsychotic	132539-06-1	C17H20N4S	21–54 hours	CID 4585
	Acetylcholine		**Treating** the symptoms of **Alzheimer's** can provide patients with comfort, dignity, and independence for a longer period of time and can encourage **... Medications** called	51-84-3	C7H16NO2	approximately 2 minutes	187
	Phosphatidylserine		**Phosphatidylserine** might increase a chemical in the body called acetylcholine. **Medications** for **Alzheimer's** disease called acetylcholinesterase inhibitors also increase	8002-43-5	C13H24NO10P		445141
	Reminyl		What should I avoid while taking **Reminyl** (galantamine)? Galantamine can cause side effects that may impair your thinking or reactions. **... Reminyl** (galantamine) side	357-70-0	C17H21NO3	7 hours	1 possible value
	Vitamin E		While current **medications** cannot stop the damage **Alzheimer's** causes to brain cells, they may help lessen or stabilize symptoms for a limited time by affecting certain chemicals	59-02-9	C 29 H 50 O 2		
	Exelon		**Exelon** will not be available in generic form until Novartis' patent expires in 2014. Sources: About **Exelon** for mild to moderate **Alzheimer's** dementia. Novartis Pharmaceuticals. 2008.	123441-03-2	3 possible values	1.5 hours	77991
	Aricept		It's important to remember that while **ARICEPT** treats the symptoms of Alzheimer's disease, it is not a cure. ... Before starting on **ARICEPT** 23 mg/day, patients should be on **ARICEPT** 10	120011-70-3	2 possible values	70 hours	

[www.google.com/squared/]

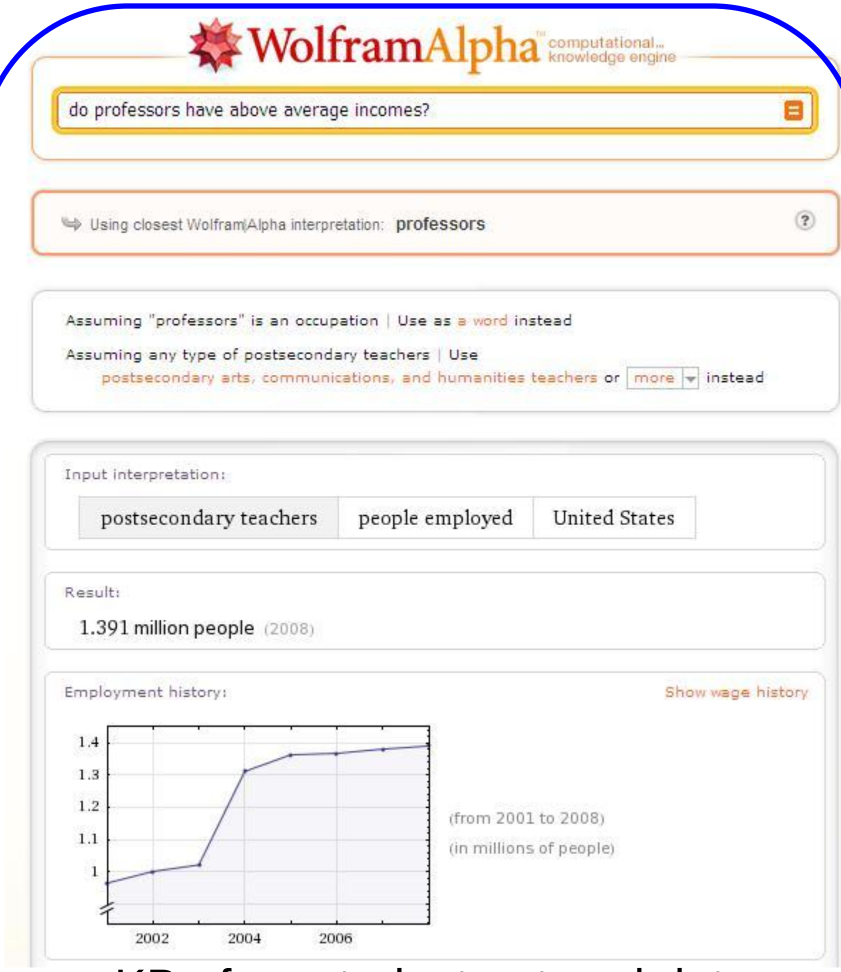
Application 1: Faceted Search



H. Bast et al.: ESTER: Efficient Search on Text, Entities, and relations, SIGIR'07

Application 2: Question Answering (QA)

- KB from Wikipedia and user edits
- translation of natural-language questions into KB queries



- KB of curated, structured data
- not just facts, but also algorithms & models

Application 2: Deep QA in NL

William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia" inspired this author's most famous novel

This town is known as "Sin City" & its downtown is "Glitter Gulch"

As of 2010, this is the only former Yugoslav republic in the EU

99 cents got me a 4-pack of Ytterlig coasters from this Swedish chain

question classification & decomposition → **knowledge back-ends**

WIKIPEDIA
The Free Encyclopedia

DBpedia

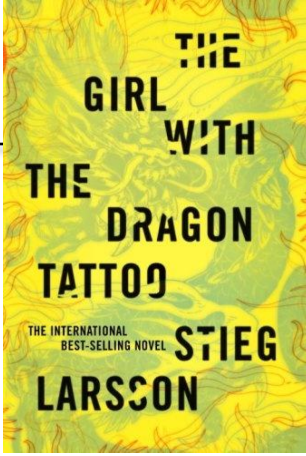
freebase

YAGO

D. Ferrucci et al.: Building Watson: An Overview of the DeepQA Project. AI Magazine, Fall 2010.

www.ibm.com/innovation/us/watson/index.htm

Application 3: Machine Reading



It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

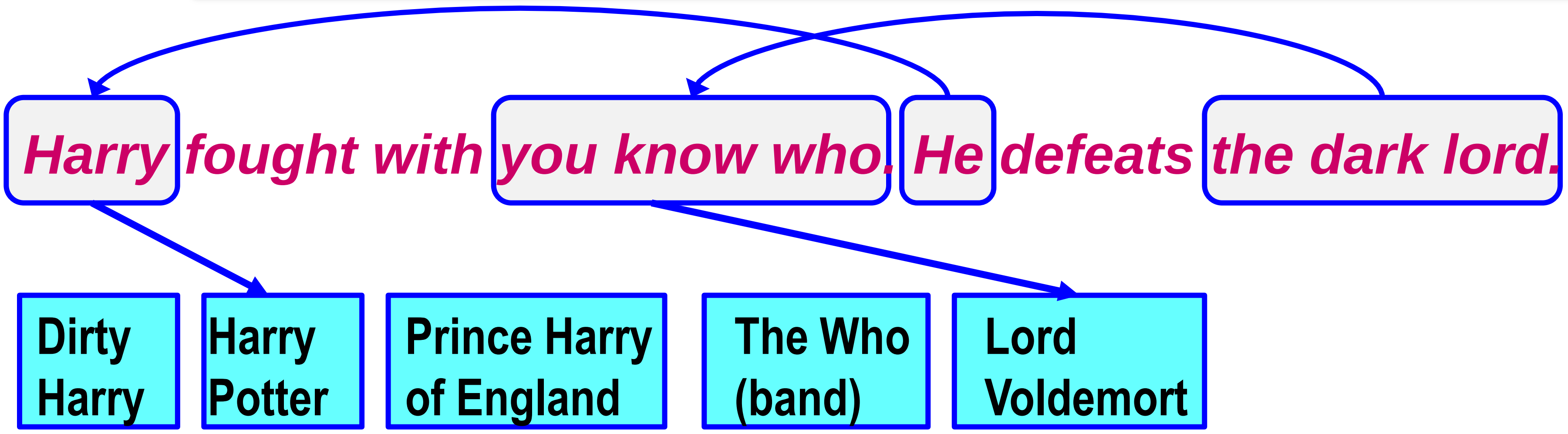
Blomkvist visits Henrik Vanger at same te on th same and of Hedeby. The old man dr... Blomkvist in by promising solid evidence against Wennerström. Blomkvist ag same spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V owns niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco uncleOf inted with the men hires the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of H...

Af same overing that Salander has hacked into his co affairWith persuade same assist him with research. They event affairWith e lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer. enemyOf

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou headOf gations for Dragan Armansky, who, in tu same ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

O. Etzioni, M. Banko, M.J. Cafarella: Machine Reading, AAAI'06
T. Mitchell et al.: Populating the Semantic Web by Macro-Reading Internet Text, ISWC'09

Named-Entity Disambiguation



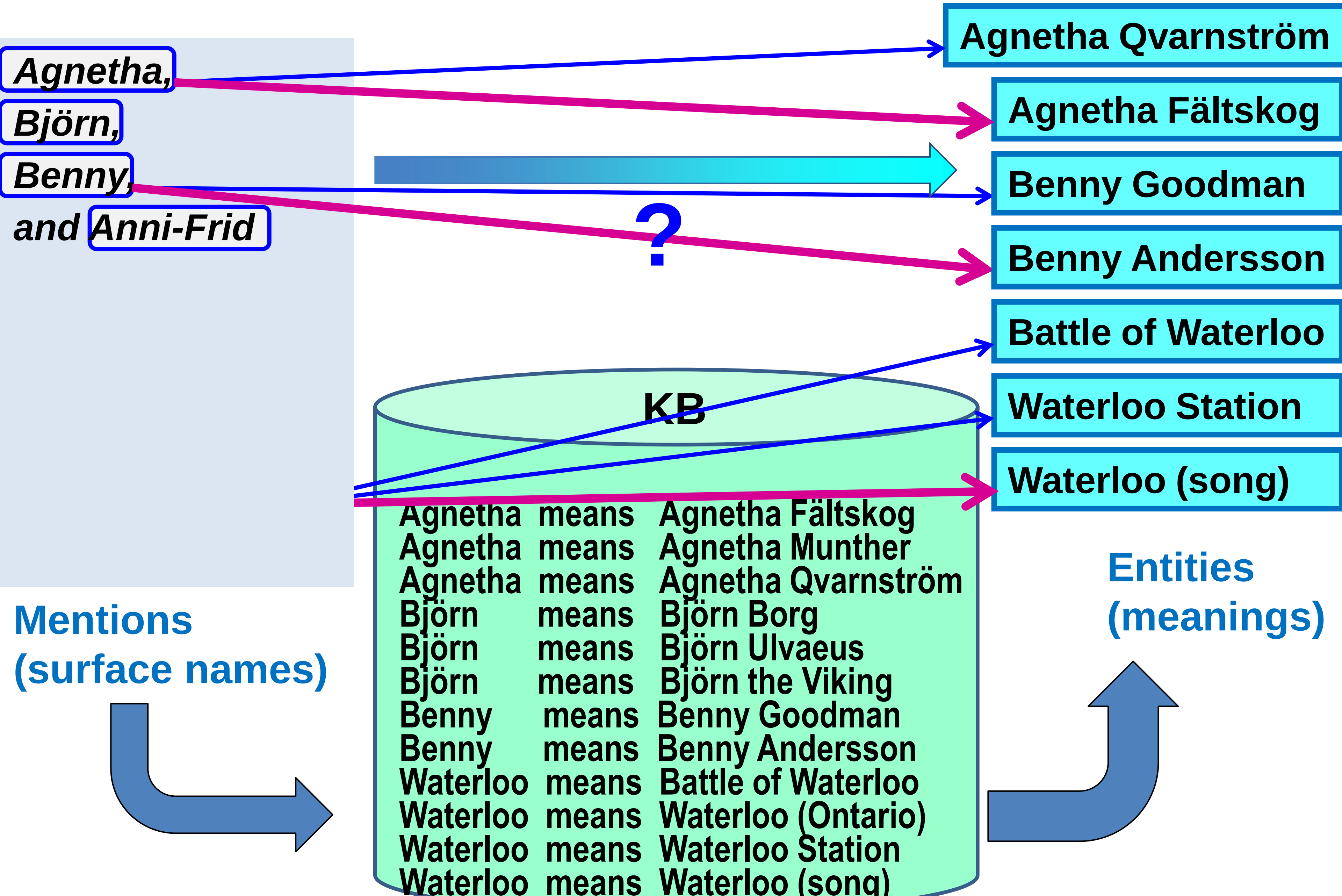
Three NLP tasks:

1) named-entity detection: segment & label by HMM or CRF (e.g. Stanford NER tagger)

2) co-reference resolution: link to preceding NP (trained classifier over linguistic features)

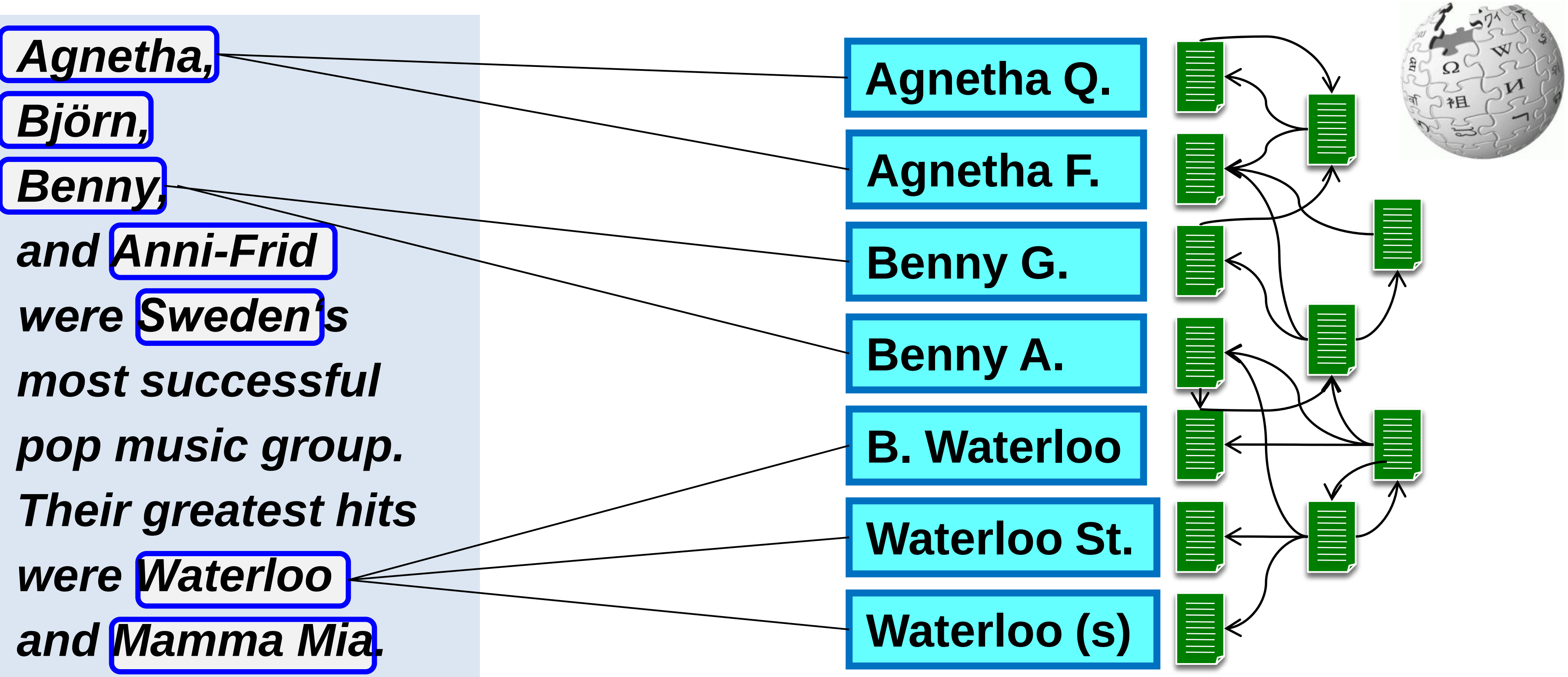
3) named-entity disambiguation: map each mention (name) to canonical entity (entry in KB)

Mentions, Meanings, Mappings



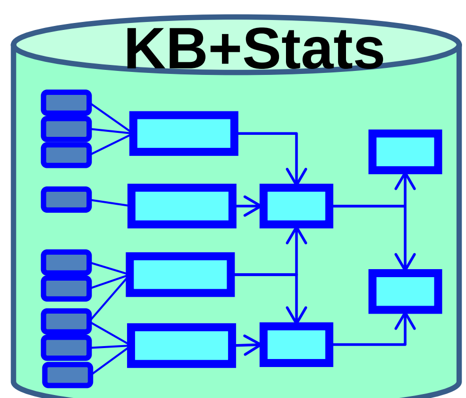
Mention-Entity Graph

weighted undirected graph with two types of nodes



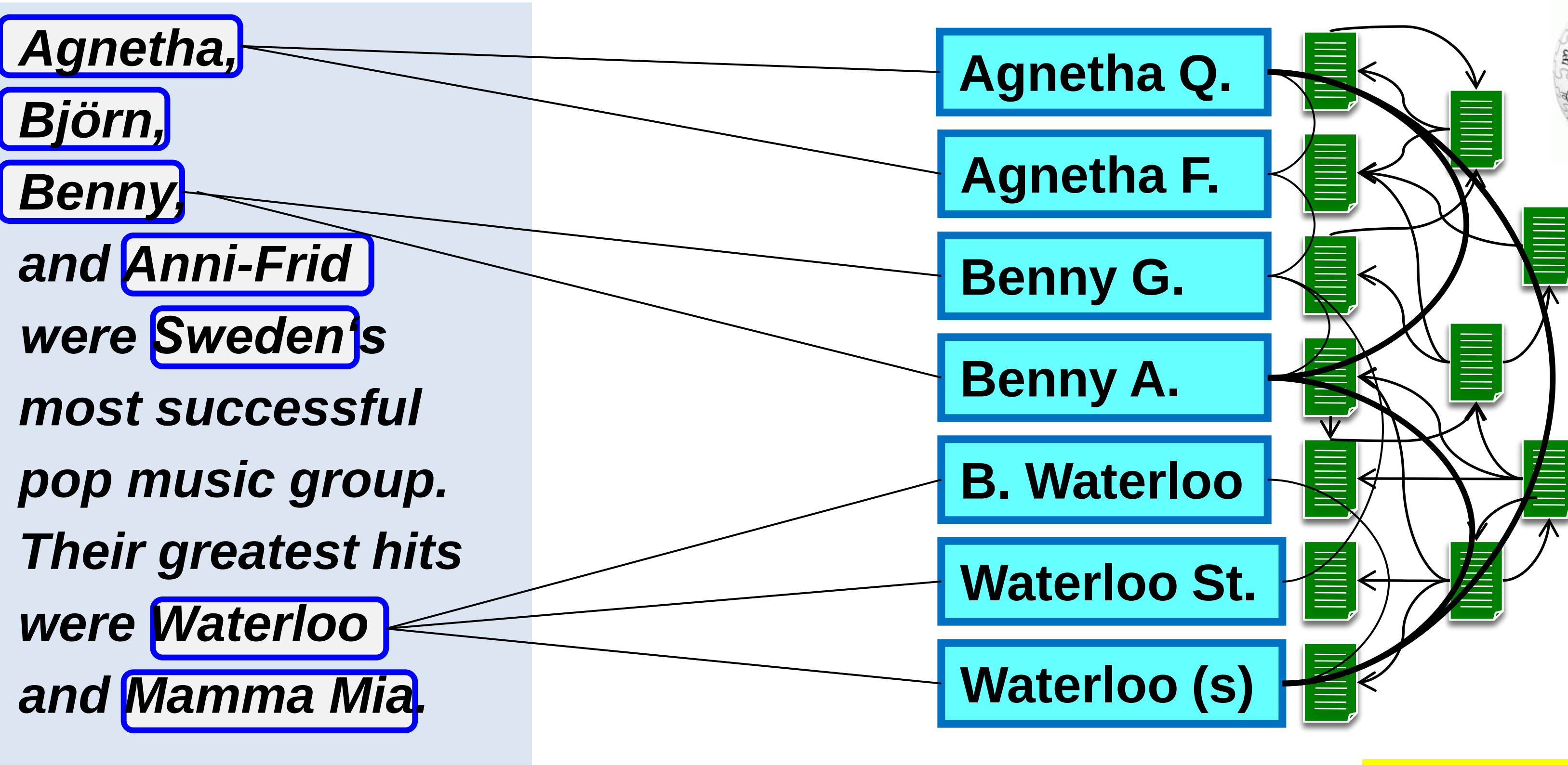
Popularity (m,e):
- freq(m,e|m)
- length(e)
- #links(e)

Similarity (m,e):
- cos/Dice/KL (context(m), context(e))



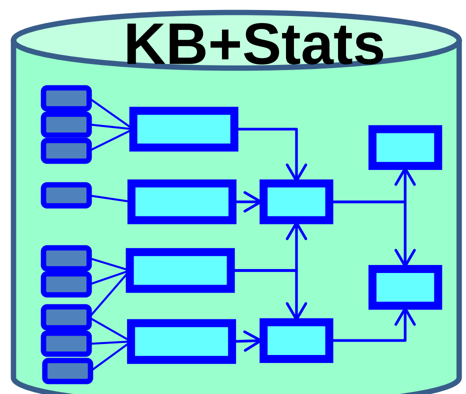
Mention-Entity Graph

weighted undirected graph with two types of nodes



Popularity (m,e):
- freq(m,e|m)
- length(e)
- #links(e)

Similarity (m,e):
- cos/Dice/KL (context(m), context(e))



Coherence (e,e‘):
- dist(types)
- overlap(links)
- overlap (anchor words)

Mention-Entity Graph

weighted undirected graph with two types of nodes

Agnetha, Björn, Benny, and Anni-Frid were Sweden's most successful pop music group. Their greatest hits were Waterloo and Mamma Mia.

Agnetha Q.

Agnetha F.

Benny G.

Benny A.

B. Waterloo

Waterloo St.

Waterloo (s)

Swedish female singers
people from Jönköping
singers
musicians

Swedish songwriters
people from Stockholm
composers
musicians

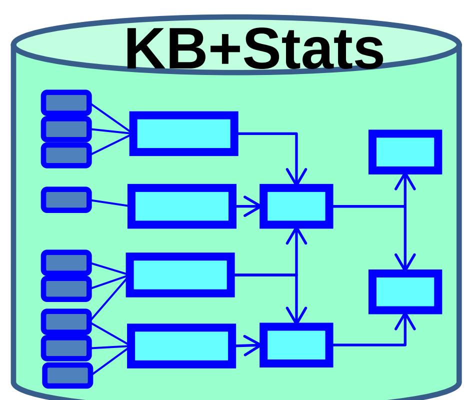
ABBA songs
#1 chart singles
songs
artifacts

Popularity (m,e):

- freq(m,e|m)
- length(e)
- #links(e)

Similarity (m,e):

- cos/Dice/KL (context(m), context(e))



Coherence (e,e'):

- dist(types)
- overlap(links)
- overlap (anchor words)

Mention-Entity Graph

weighted undirected graph with two types of nodes

Agnetha, Björn, Benny, and Anni-Frid were Sweden's most successful pop music group. Their greatest hits were Waterloo and Mamma Mia.

- Agnetha Q.
- Agnetha F.
- Benny G.
- Benny A.
- B. Waterloo
- Waterloo St.
- Waterloo (s)

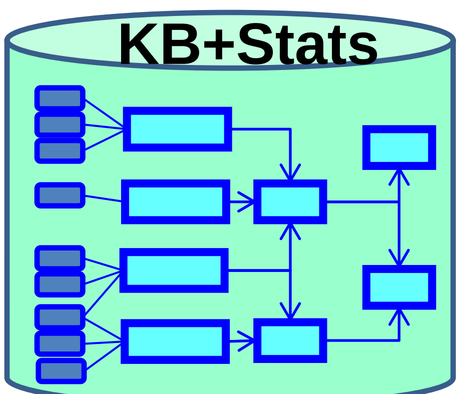
http://.../wiki/ABBA
http://.../wiki/Anni-Frid_Lyngstad
http://.../wiki/Jönköping
http://.../wiki/Eurovision_Song_Cor

http://.../wiki/ABBA
http://.../wiki/Anni-Frid_Lyngstad
http://.../wiki/Mamma_Mia!
http://.../wiki/Agnetha_Fältskog

http://.../wiki/ABBA
http://.../wiki/Eurovision_Song_Cor
http://.../wiki/Mamma_Mia!

Popularity (m,e):
- freq(m,e|m)
- length(e)
- #links(e)

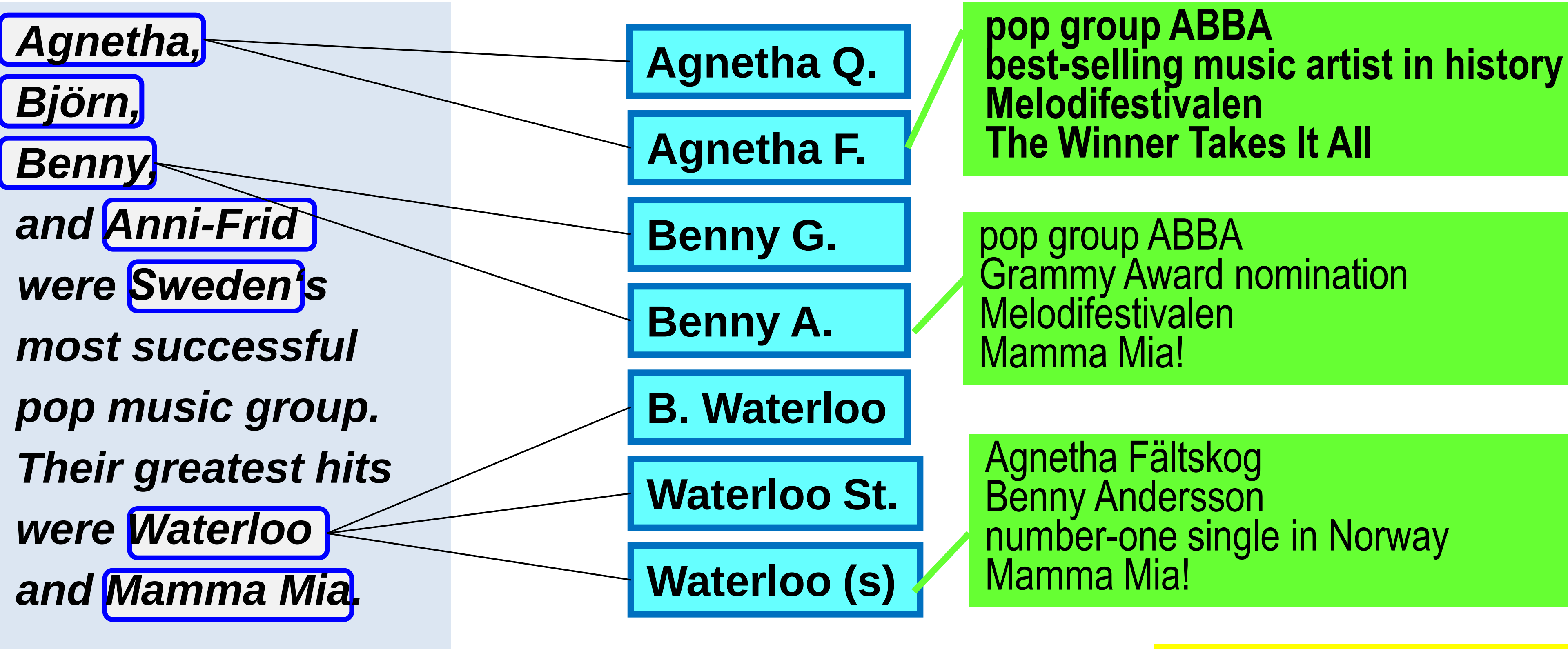
Similarity (m,e):
- cos/Dice/KL (context(m), context(e))



Coherence (e,e'):
- dist(types)
- overlap(links)
- overlap (anchor words)

Mention-Entity Graph

weighted undirected graph with two types of nodes

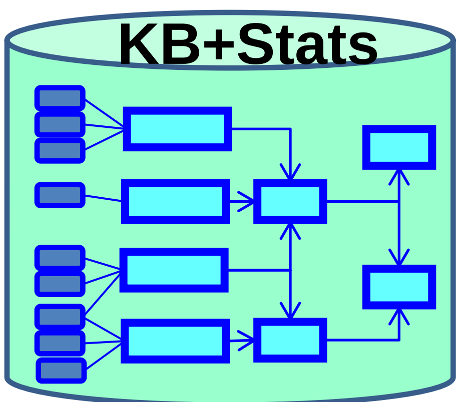


Popularity (m,e):

- freq(m,e|m)
- length(e)
- #links(e)

Similarity (m,e):

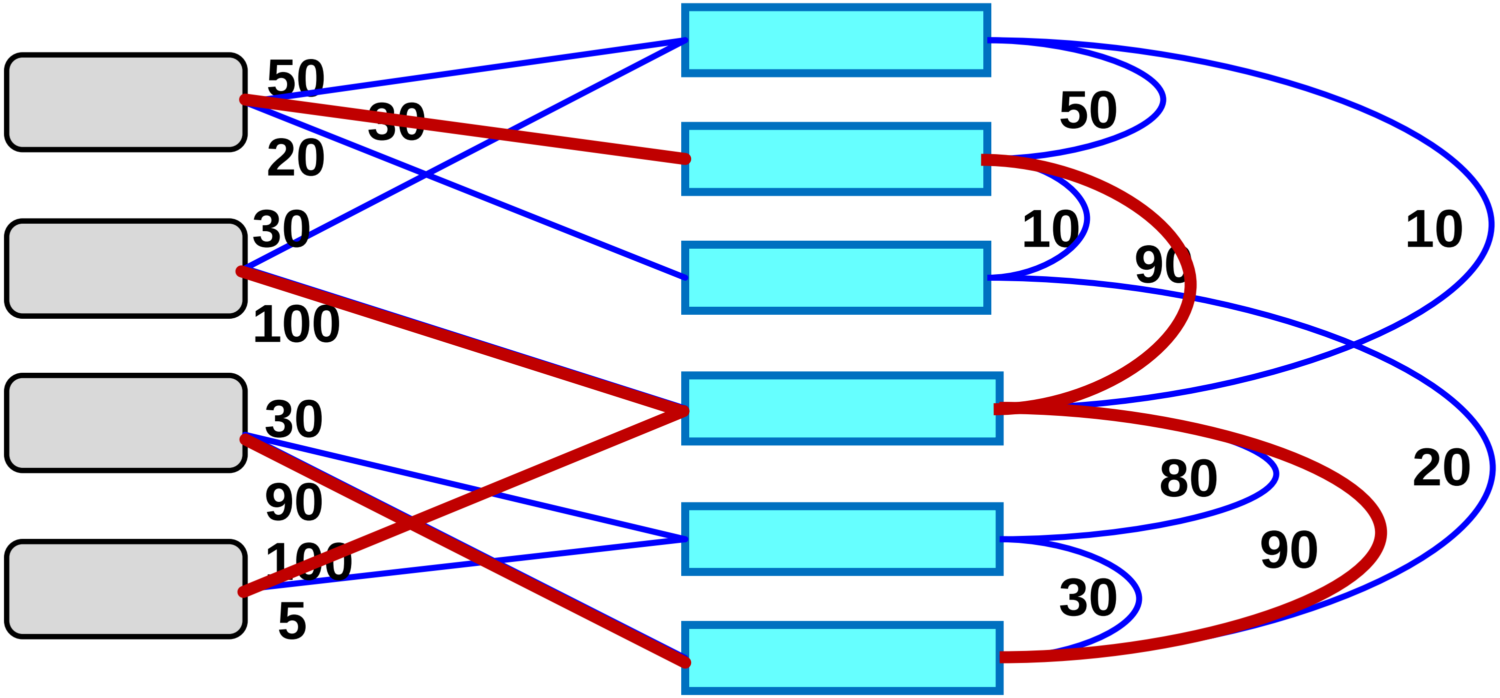
- cos/Dice/KL (context(m), context(e))



Coherence (e,e'):

- dist(types)
- overlap(links)
- overlap (anchor words)

Joint Mapping

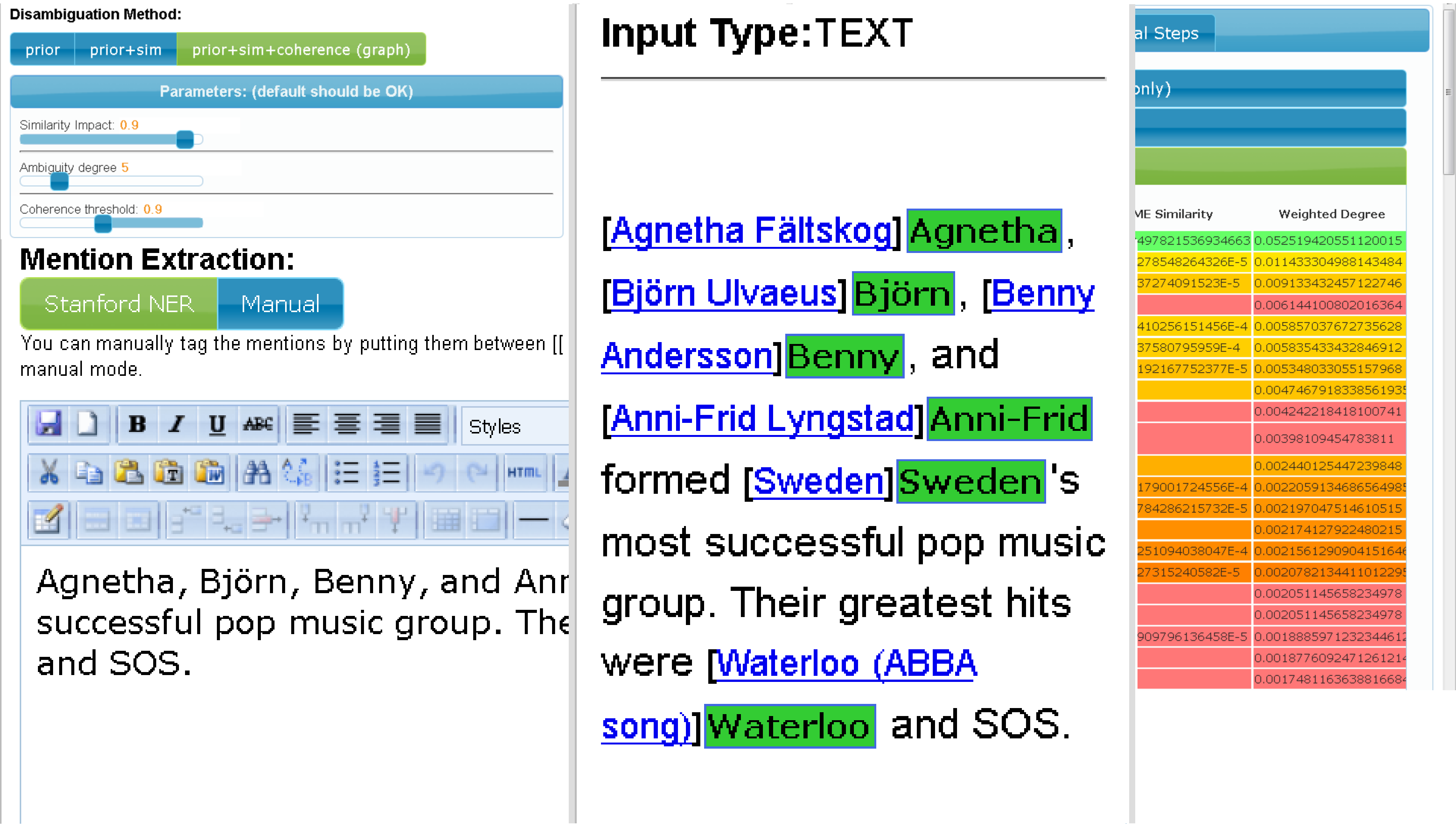


- Build **mention-entity graph** or **joint-inference factor graph** from knowledge and statistics in KB
- Compute **high-likelihood mapping** (ML or MAP) or **dense subgraph** such that: each m is **connected to exactly one** e (or **at most one** e)

K. Kulkarni et al.: Collective Annotation of Wikipedia Entities in Web Text, KDD'09
J. Hoffart et al.: Robust Disambiguation of Named Entities in Text, EMNLP'11

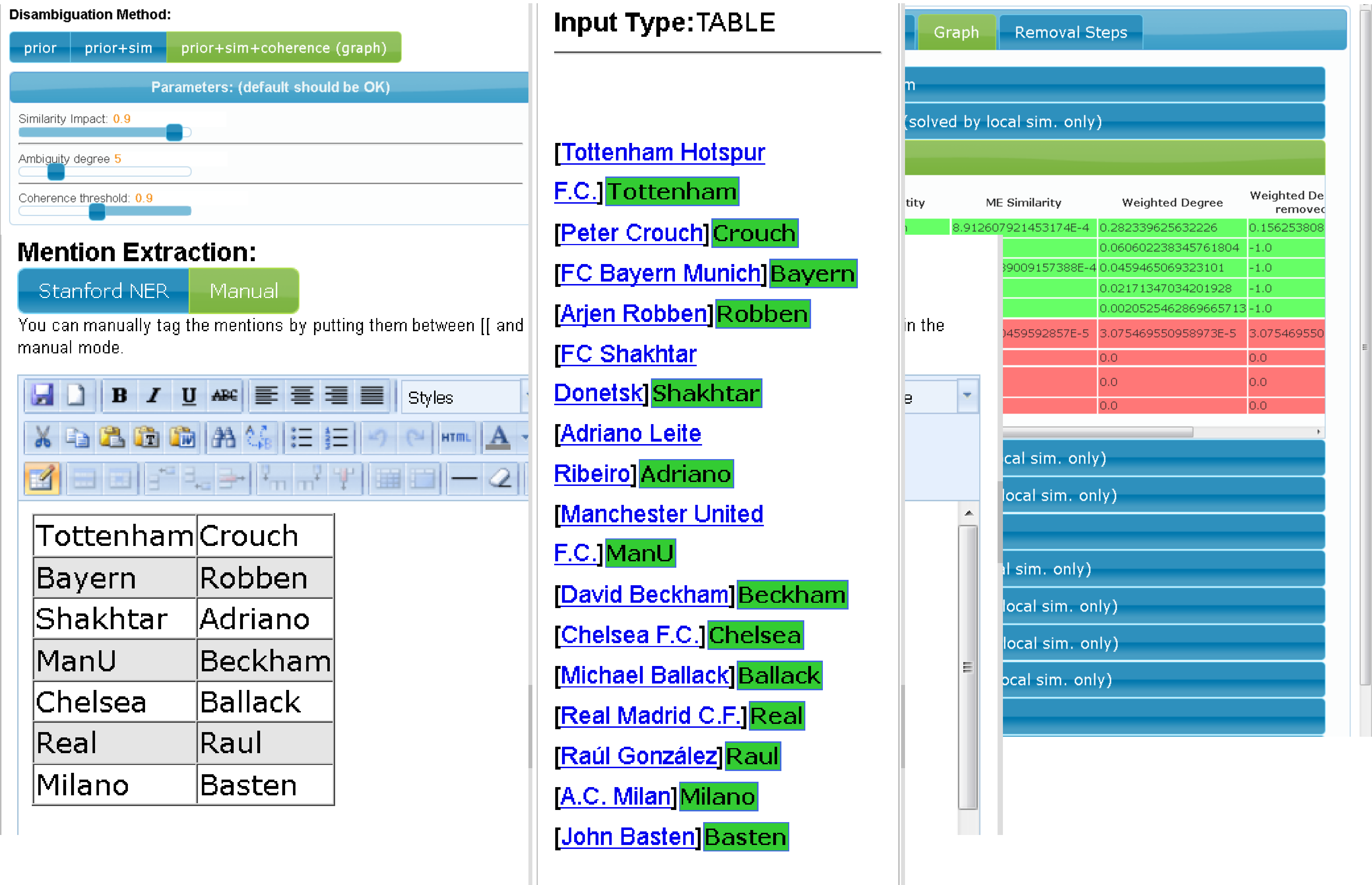
AIDA Accurate Online Disambiguation

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



AIDA Accurate Online Disambiguation

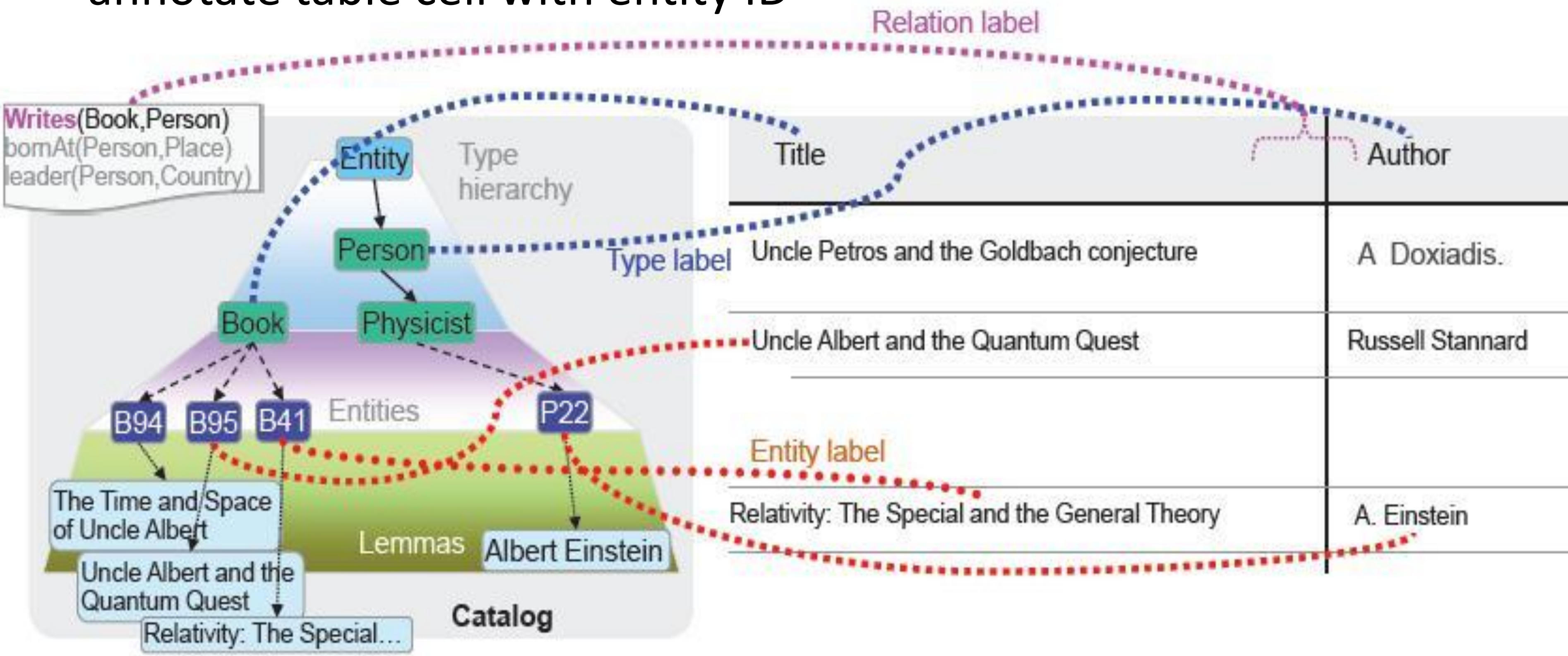
http://www.mpi-inf.mpg.de/yago-naga/aida/



Application 4: Annotation of Web Data

Given a Web table (in HTML, XML, …)

- annotate column with entity type
- annotate pair of columns with relationship type
- annotate table cell with entity ID



G. Limaye, S. Sarawagi, S. Chakrabarti: Annotating and Searching Web Tables Using Entities, Types and Relationships, PVLDB 2010

Application 4: Map Annotation

- Determine geo entities (landmarks) in vicinity, via GPS
- Show information about these entities, obtained from KB
- Smartphone and Augmented-Reality applications

C. Becker , C. Bizer: Exploring the Geospatial Semantic Web with DBpedia Mobile, J. Web Sem. 2009

Spectrum of Machine Knowledge (1)

factual:

bornIn (GretaGarbo, Stockholm), hasWon (GretaGarbo, AcademyAward), playedRole (GretaGarbo, MataHari), livedIn (GretaGarbo, Klosters)

taxonomic (ontology):

instanceOf (GretaGarbo, actress), subclassOf (actress, artist)

lexical (terminology):

means (“Big Apple“, NewYorkCity), means (“Apple“, AppleComputerCorp)
means (“MS“, Microsoft) , means (“MS“, MultipleSclerosis)

multi-lingual:

meansInChinese („乔戈里峰“, K2), meansInUrdu („ “, K2)
meansInFrench („école“, school (institution)),
meansInFrench („banc“, school (of fish))

Spectrum of Machine Knowledge (2)

ephemeral (dynamic services):

wsdl:getSongs (musician ?x, song ?y), wsdl:getWeather (city?x, temp ?y)

common-sense (properties):

hasAbility (Fish, swim), hasAbility (Human, write),

hasShape (Apple, round), hasProperty (Apple, juicy),

hasMaxHeight (Human, 2.5 m)

common-sense (rules):

$\forall$ **x: human(x)** $\Rightarrow$ **male(x)** $\vee$ **female(x)**

$\forall$ **x: (male(x)** $\Rightarrow \neg$ **female(x))** $\wedge$ **(female(x))** $\Rightarrow \neg$ **male(x))**

$\forall$ **x: animal(x)** $\Rightarrow$ **(hasLegs(x)** $\Rightarrow$ **isEven(numberOfLegs(x))**

temporal (fluents):

hasWon (GretaGarbo, AcademyAward)@1955

marriedTo (AlbertEinstein, MilevaMaric)@[6-Jan-1903, 14-Feb-1919]

Spectrum of Machine Knowledge (3)

free-form (open IE):

hasWon (NataliePortman, AcademyAward)

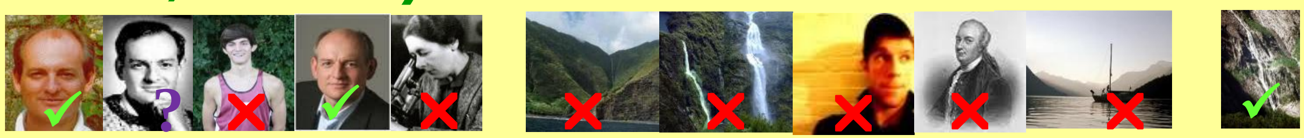
occurs („Natalie Portman“, „celebrated for“, „Oscar Award“)

occurs („Jeff Bridges“, „nominated for“, „Oscar“)

multimodal (photos, videos):

StuartRussell

JamesBruceFalls



social (opinions):

admires (maleTeen, LadyGaga), supports (AngelaMerkel, HelpForGreece)

epistemic ((un-)trusted beliefs):

believe(Ptolemy,hasCenter(world,earth)), believe(Copernicus,hasCenter(world,sun))

believe (peopleFromTexas, bornIn(BarackObama,Kenya))

This Tutorial

In this tutorial, we will explain:

- **how a knowledge base is organized**
- **which knowledge bases are publicly available**
- **how we can automatically construct knowledge bases**
- **how we can query a knowledge base and rank the results**
- **how we can deal with inter-linked knowledge bases**

We discuss:

- **fundamental models & methods**
- **state-of-the-art techniques**
- **open problems & research challenges**

Readings for Part I

- D.B. Lenat: CYC: A Large-Scale Investment in Knowledge Infrastructure. Commun. ACM 38(11): 32-38, 1995
- C. Fellbaum, G. Miller (Eds.): WordNet: An Electronic Lexical Database, MIT Press, 1998
- O. Etzioni, M. Banko, S. Soderland, D.S. Weld: Open information extraction from the web. Commun. ACM 51(12): 68-74, 2008
- G. Weikum, G. Kasneci, M. Ramanath, F.M. Suchanek: Database and information-retrieval methods for knowledge discovery. Commun. ACM 52(4): 56-64, 2009
- A. Doan, L. Gravano, R. Ramakrishnan, S. Vaithyanathan (Eds.): Special Issue on Managing Information Extraction, SIGMOD Record 37(4), 2008
- G. Weikum, M. Theobald: From information to knowledge: harvesting entities and relationships from web sources. PODS 2010
- First Int. Workshop on Automated Knowledge Base Construction (AKBC), Grenoble, 2010, http://akbc.xrce.xerox.com/
- D.A. Ferrucci, Building Watson: An Overview of the DeepQA Project. AI Magazine 31(3): 59-79, 2010
- T.M. Mitchell, J.Betteridge, A. Carlson, E.R. Hruschka Jr., R.C. Wang: Populating the Semantic Web by Macro-Reading Internet Text. ISWC 2009
- Roberto Navigli: Word sense disambiguation: A survey. ACM Comput. Surv. 41(2), 2009
- Stefano Ceri, Marco Brambilla: Search Computing: Challenges and Directions, Springer, 2010

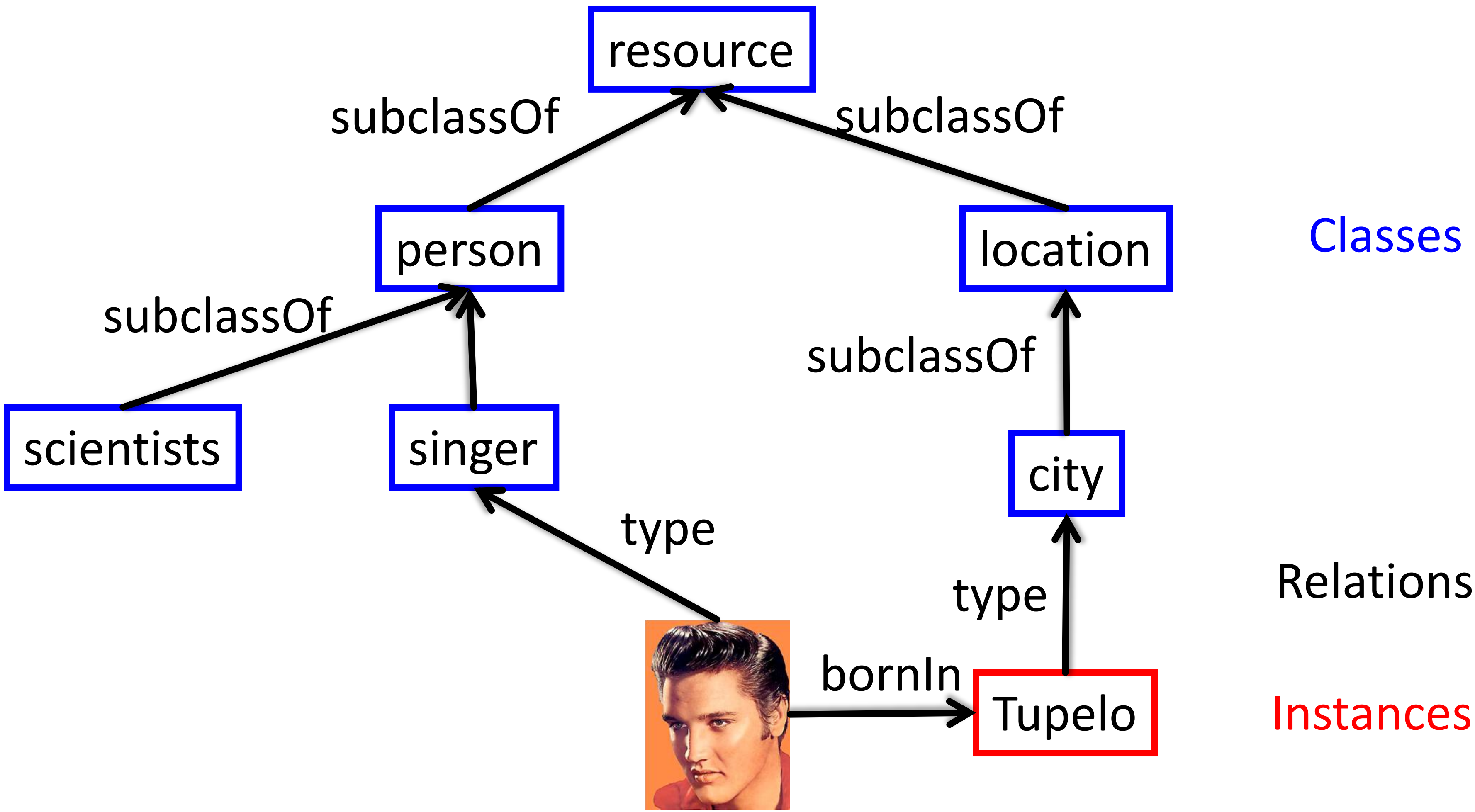
Outline

- Part I ✔
  - Machine Knowledge & Intelligent Applications
- **Part II**
  - Knowledge Representation & Public Knowledge Bases
- Part III
  - Extracting Knowledge
- Part IV
  - Ranking and Searching
- Part V
  - Linked Data
- Part VI
  - Conclusion and Outlook

Outline for Part II

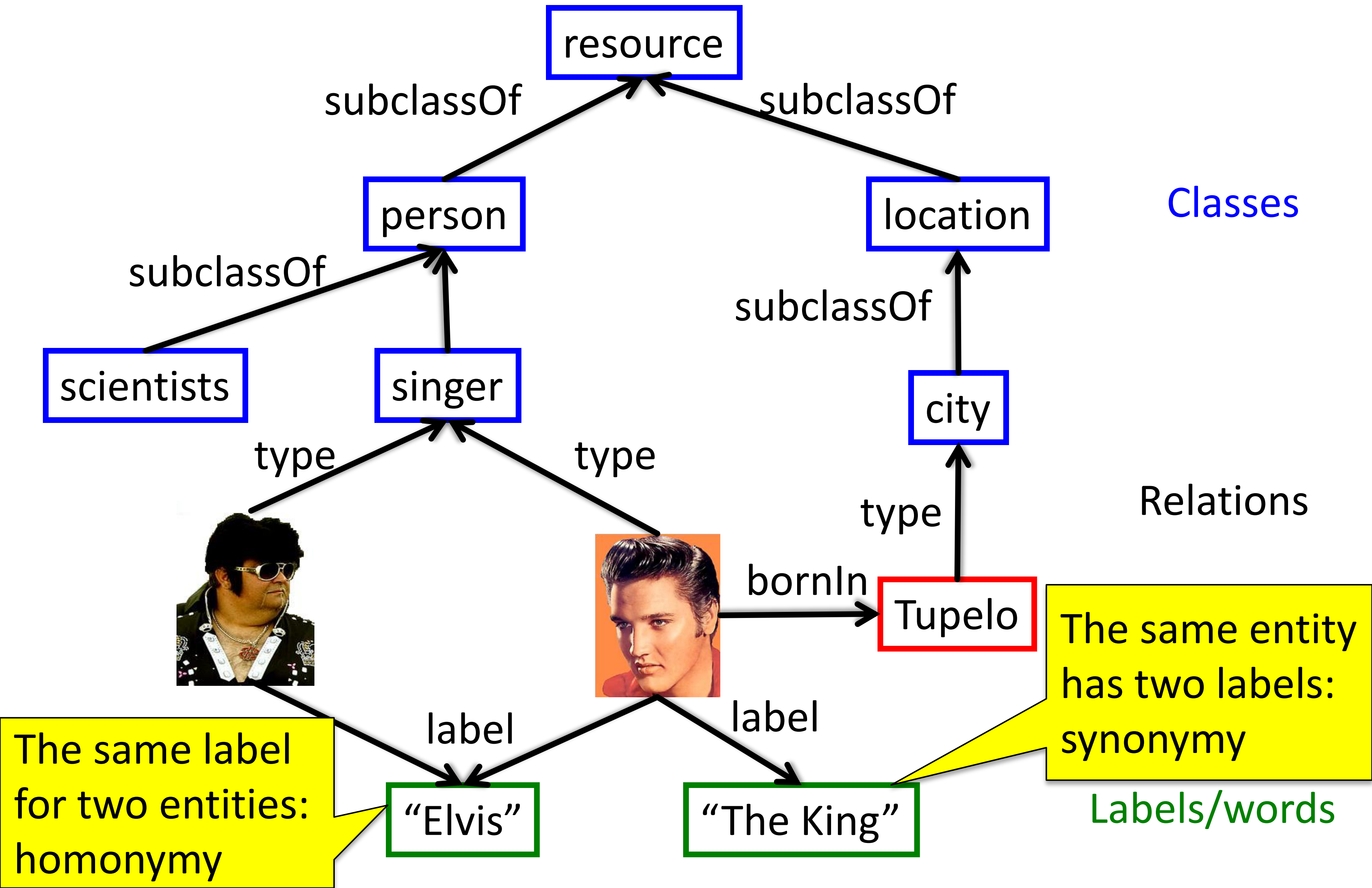
- **Knowledge Representation**
- **Public Knowledge Bases:**
  - **Manually constructed knowledge bases**
  - **Knowledge bases from Wikipedia**
  - **Knowledge bases beyond Wikipedia**

RDFS-Ontologies



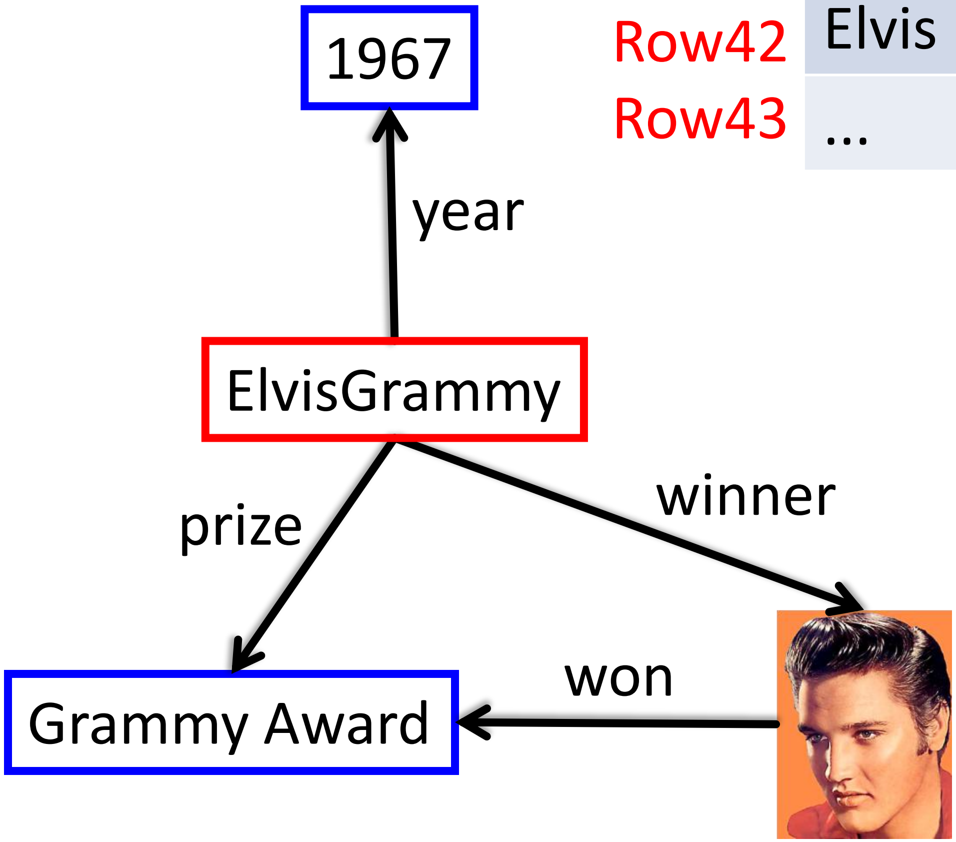
An **RDFS-ontology** can be seen as a directed labeled multi-graph, where the nodes are entities and the edges relations.

Labels


resource
subclassOf
subclassOf
person
location
Classes
subclassOf
subclassOf
scientists
singer
city
type
type
type
Relations
bornIn
Tupelo
The same entity has two labels: synonymy
label
label
The same label for two entities: homonymy
“Elvis”
“The King”
Labels/words

Event Entities

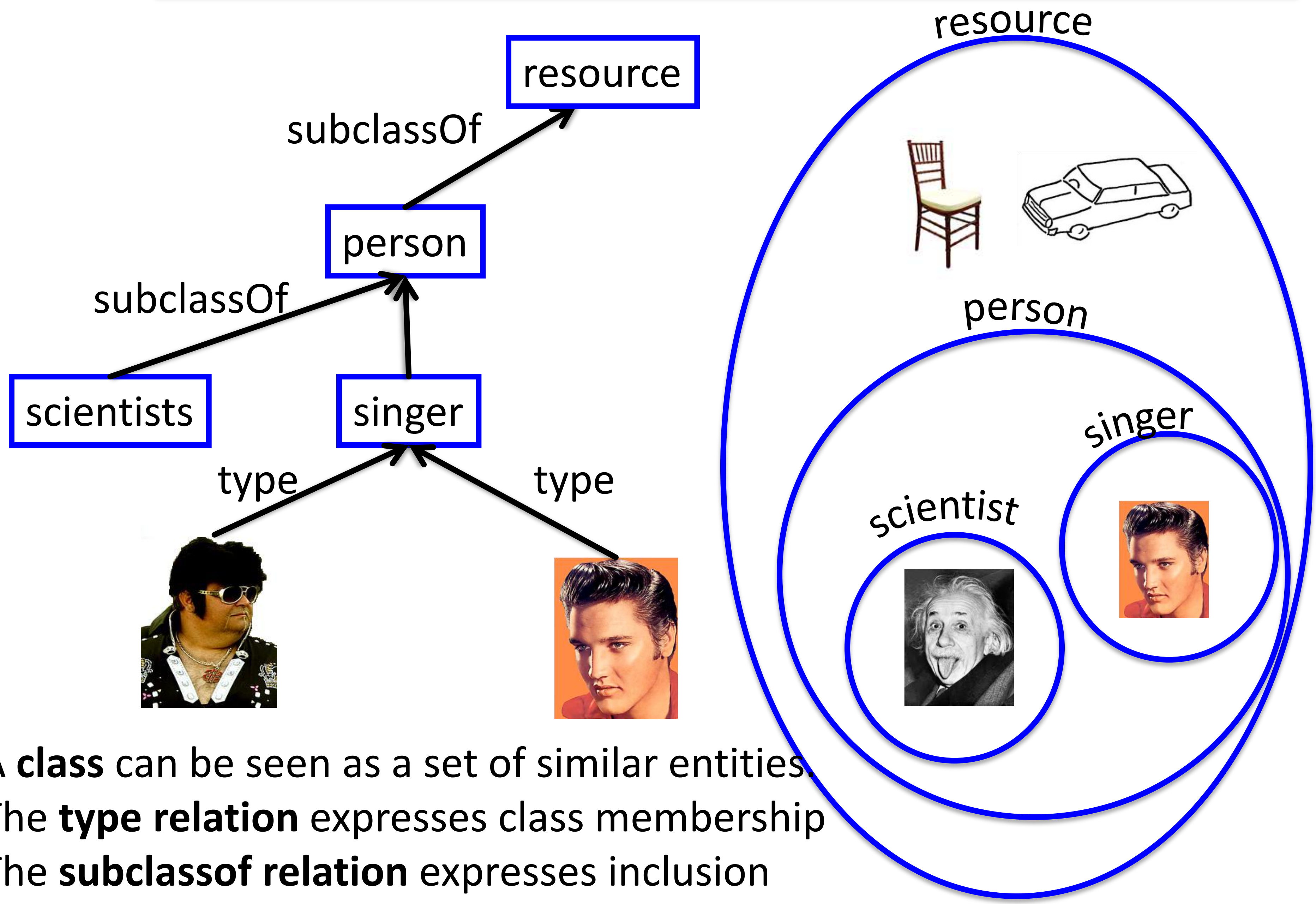
	Winner	Prize	Year
Row42	Elvis Presley	Grammy Award	1967
Row43	...	...	...



An **event entity** is an artificial entity introduced to represent an n-ary relationship

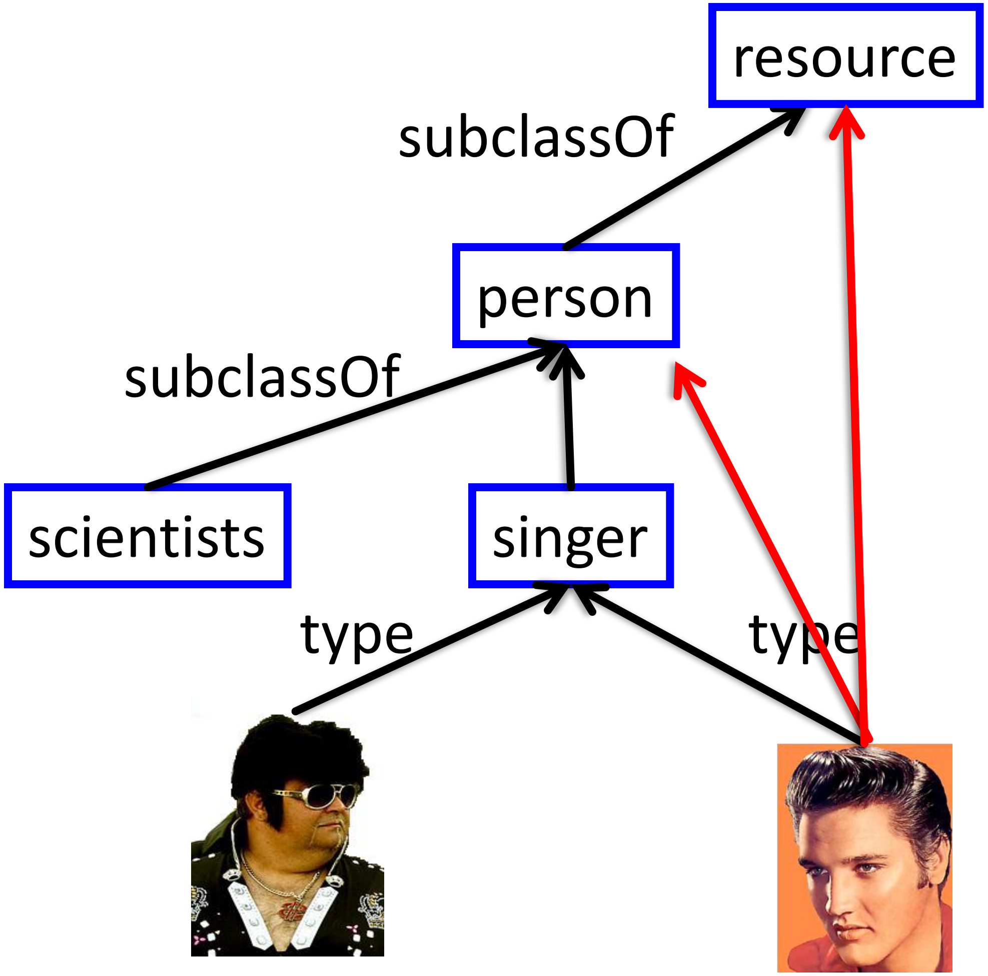
Event entities allow representing arbitrary relational data as binary graphs

Classes



A **class** can be seen as a set of similar entities.
The **type relation** expresses class membership
The **subclassof relation** expresses inclusion

Entailment



RDFS specifies **entailment rules** of the form

If the KB contains triples of this form

then add this triple

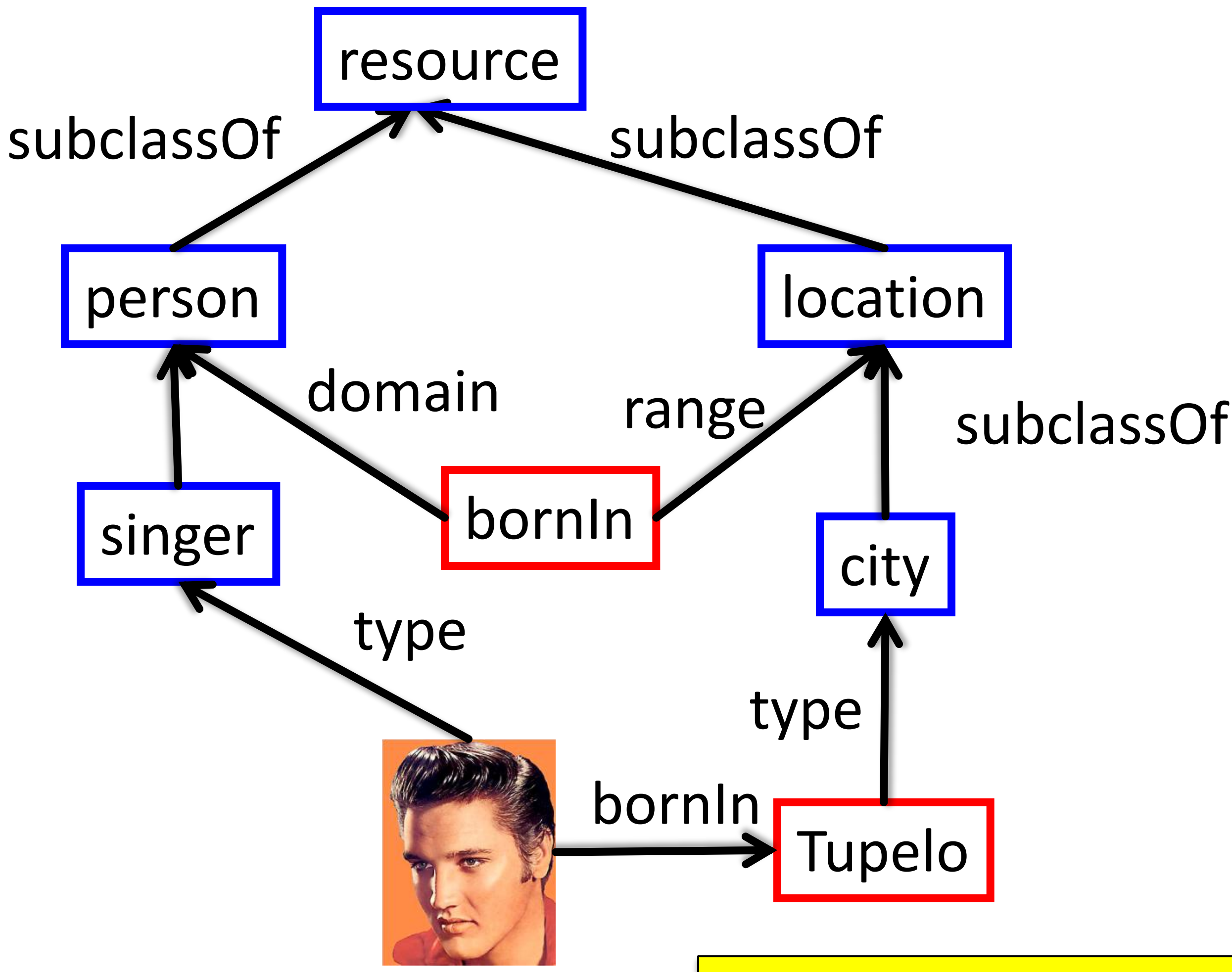
Example:

<X, type, C>
<C, subclassOf,D>

<X,type,D>

This computation terminates in polynomial time (if no blank nodes are present).

Relations

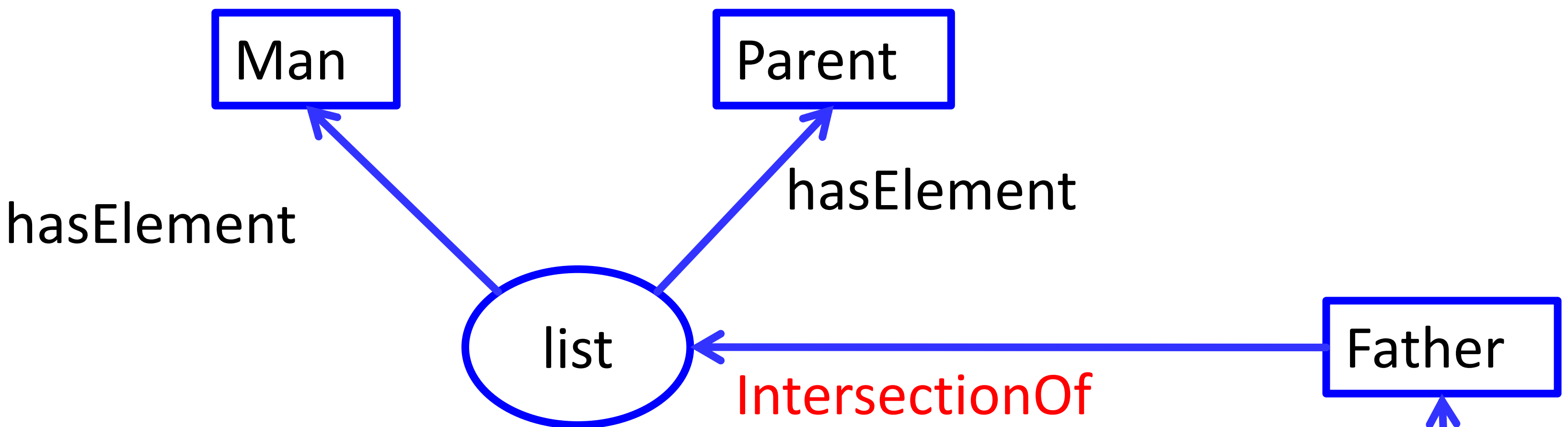


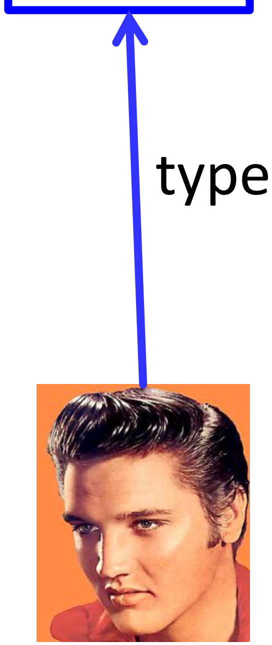
Relations are by themselves entities.
Facts can talk about relations.

Looks like higher order, but is not. Everything is just triples.

OWL

The **Web Ontology Language (OWL)** is a set of predicates with special additional semantic rules.





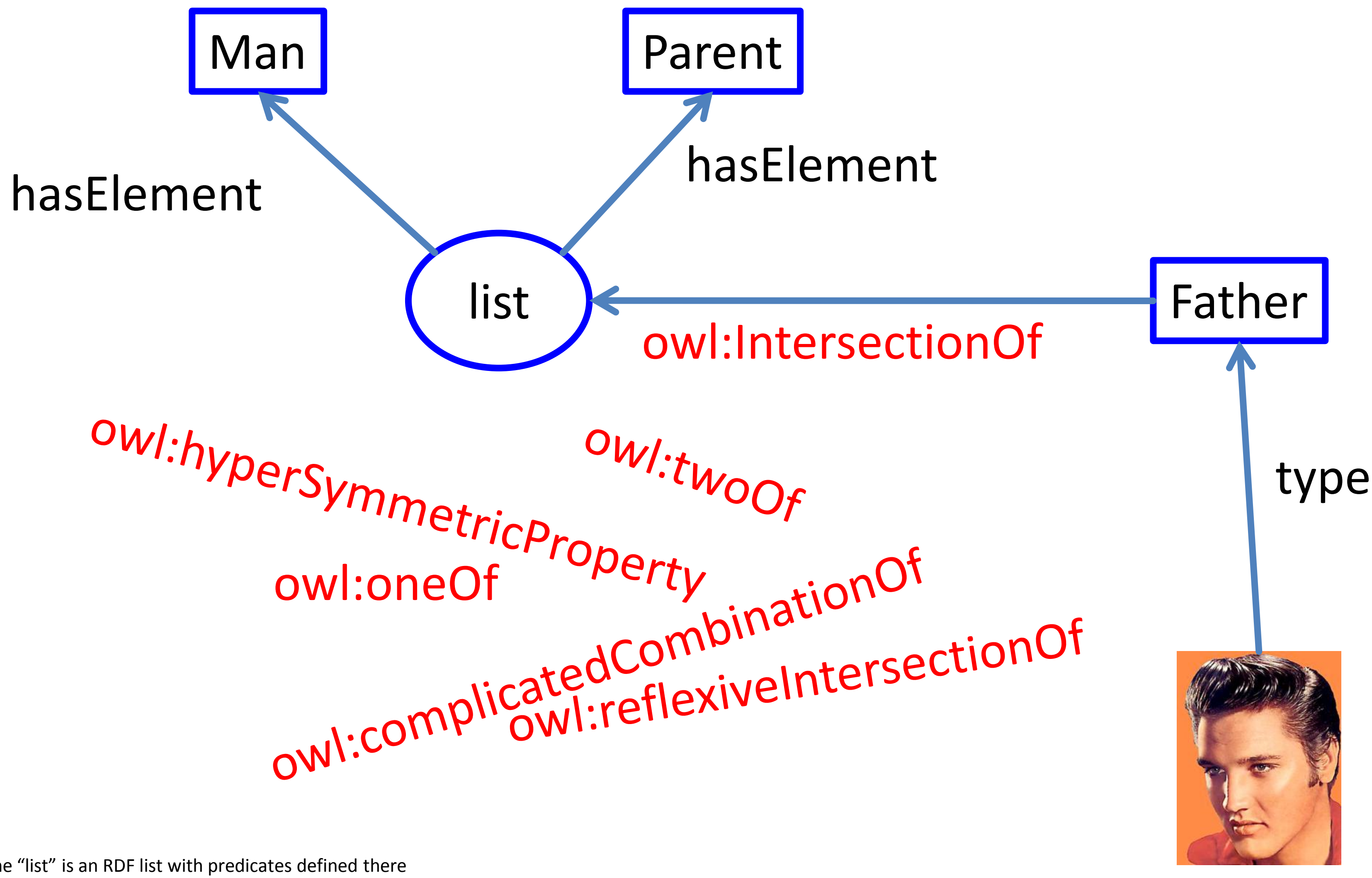
X type C
C intersectionOf LIST
LIST hasElement Z

X type Z

The “list” is an RDF list with predicates defined there

OWL Undecidability

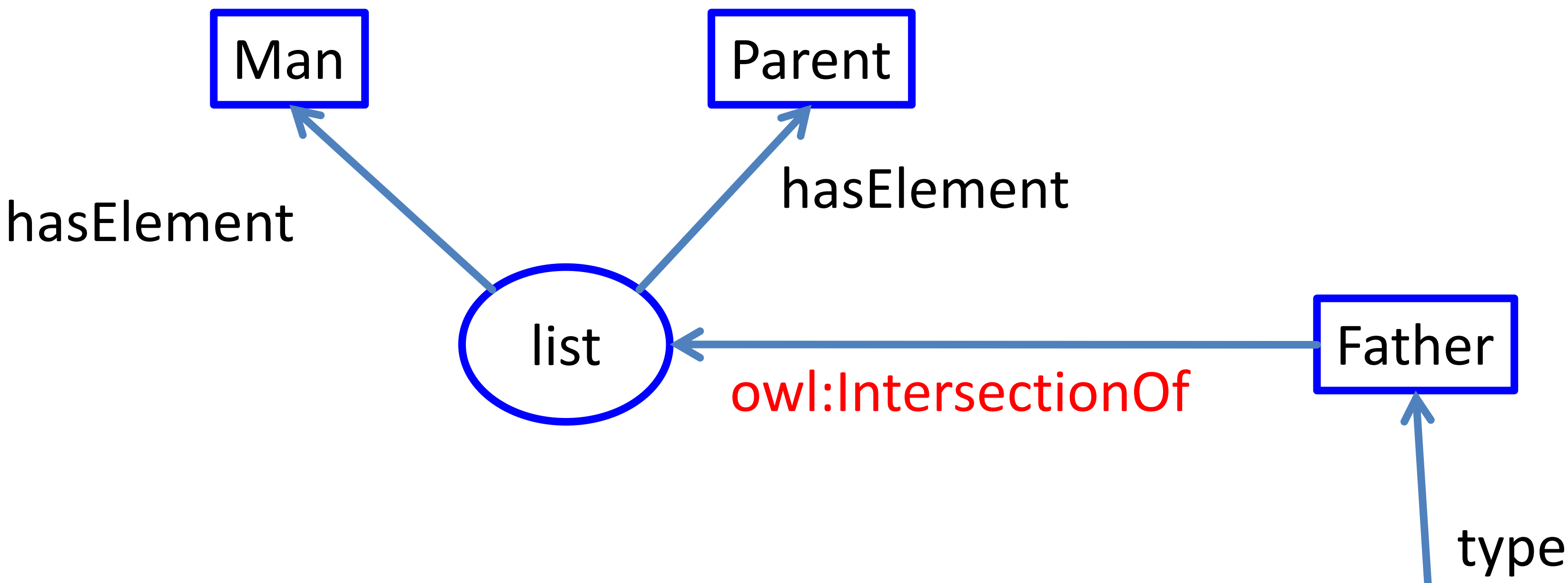
OWL defines so powerful predicates that it is **undecidable**.



The “list” is an RDF list with predicates defined there

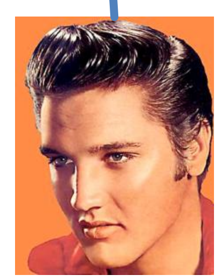
OWL-DL

There are several decideable fragments of OWL, e.g., **OWL-DL**.



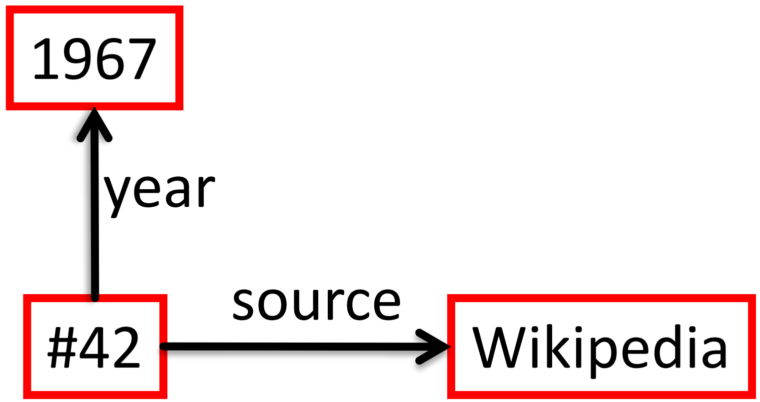
OWL-DL mirrors the Description Logic $\mathcal{SHOIN}^{(D)}$.

$$\text{father} = \text{parent} \sqcap \text{man}$$



Reification

Reification is the method of creating an entity that represents a fact.



RDFS: Summary

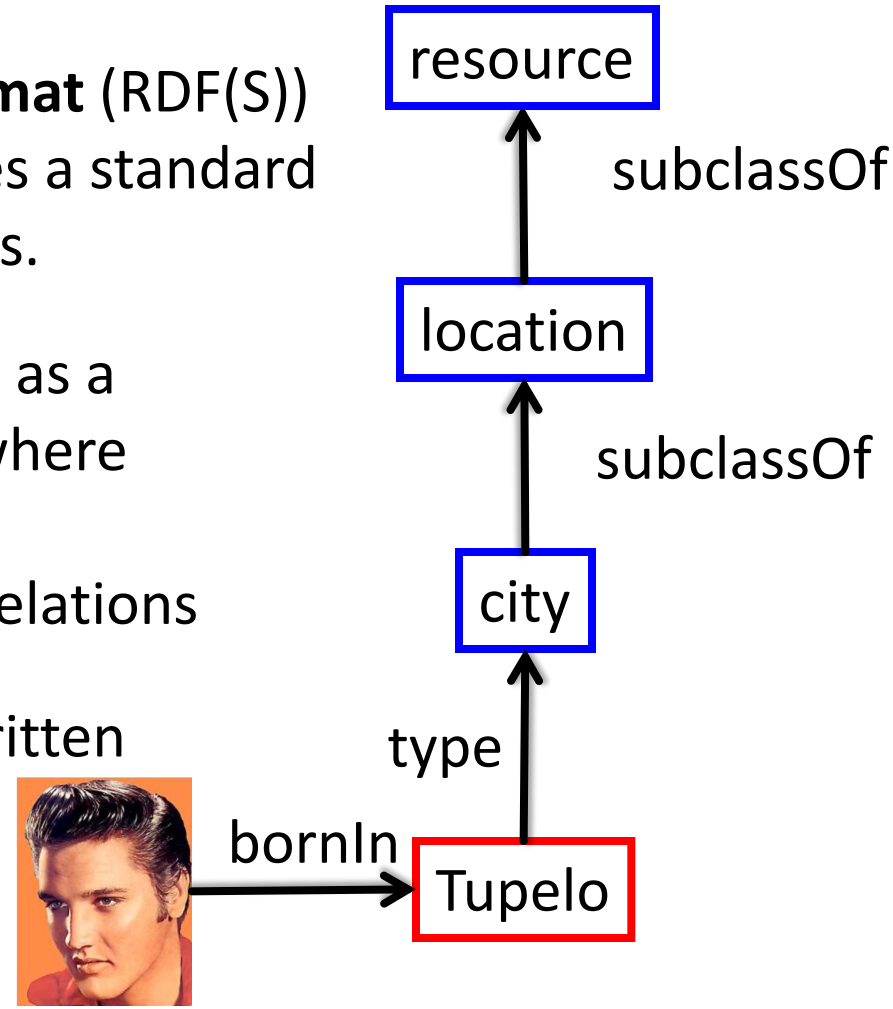
The **Resource Description Format** (RDF(S)) is a W3C standard that provides a standard vocabulary to model ontologies.

An **RDFS ontology** can be seen as a directed labeled multi-graph where

- the nodes are entities
- the edges are labeled with relations

Edges (**facts**) are commonly written

- as triples
  <Elvis, bornIn, Tupelo>
- as literals
  bornIn(Elvis, Tupelo)



[W3C RDF 2004]

Outline for Part II

- **Knowledge Representation ✔**
- **Public Knowledge Bases:**
  - **Manually constructed knowledge bases**
  - **Knowledge bases from Wikipedia**
  - **Knowledge bases beyond Wikipedia**

Cyc

What if we could make all common sense knowledge computer-processable?

Cyc project


Douglas Lenat

Cyc is a knowledge base
about common sense knowledge

- started in 1984
- driven by cycorp
- staff of 20
- goal: formalize knowledge *manually*

[Lenat, Comm. ACM 1995]

Cyc: Language

CycL is the formal language that Cyc uses to represent knowledge.
(Semantics based on First Order Logic, syntax based on LISP)

```
(#$forall ?A
   (#$implies
      (#$isa ?A #$Animal)
      (#$thereExists ?M
         (#$mother ?A ?M))))
```

Cyc project

```
(#$arity #$GovernmentFn 1)
(#$arg1Isa #$GovernmentFn #$GeopoliticalEntity)
(#$resultIsa #$GovernmentFn #$RegionalGovernment)

(#$governs (#$GovernmentFn #$Canada)  #$Canada)
```

http://cyc.com/cycdoc/ref/cycl-syntax.html

Cyc: Example of Content

#$Love

Strong affection for another agent arising out of kinship or personal ties.

guid: bd589433-9c29-11b1-9dad-c379636f7270

direct instance of: #$FeelingType

direct specialization of: #$Affection

direct generalization of: #$Love-Romantic

Cyc project

http://cyc.com/cycdoc/vocab/emotion-vocab.html#Love

Facts and axioms about: Transportation, Ecology, everyday life, chemistry, healthcare, animals, law, computer science...

If a computer network implements IEEE 802.11 Wireless LAN Protocol and some computer is a node in that computer network, then that computer is vulnerable to decryption. http://cyc.com/cyc/technology/whatiscyc_dir/maptest

Cyc: Summary

	Cyc	SUMO
Content	Common sense knowledge, axioms	Common sense knowledge, axioms
Main strength	Huge ontology, with tools	Free research project
Technique	Manual	Manual
License	proprietary, OpenCyc is Apache License V2.0	GNU GPL
Entities	500k	20k
Assertions	5m	70k
Relations	15k	
Tools	Reasoner, NL tool	Reasoner
URL	http://cyc.com	http://ontologyportal.org
References	[Lenat, Comm. ACM 1995]	[Niles, FOIS 2001]

SUMO is a research project in a similar spirit, driven by Adam Pease.

http://cyc.com/cyc/technology/whatiscyc_dir/whatsincyc in 2010 http://ontologyportal.org

WordNet

What if we could make the English language computer-processable?

George Miller,
Christiane Fellbaum

WordNet is a lexicon of the English language

- started in 1985
- driven by the Cognitive Science Laboratory, Princeton University
- written by lexicographers
- goal: formalize the English language

[Miller, CACM 1995]

WordNet: Content

Concept hierarchy

{equipment}

Concept ("synset")

Synonymous words for the concept

{ camera,
photographic camera}

{ camera,
television camera}

Gloss describing the concept

"television equipment consisting of a lens system that focuses an image on a photosensitive mosaic"

WordNet: Semantic Relations

Relation	Meaning	Examples
Synonymy (N, V, Adj, Adv)	Same sense	(camera, photographic camera) (mountain climbing, mountaineering) (fast, speedy)
Antonymy (Adj, Adv)	Opposite	(fast, slow) (buy, sell)
Hypernymy (N)	subclassOf	(camera, photographic equipment) (mountain climbing, climb)
Meronymy (N)	Part	(camera, optical lens) (camera, view finder)
Troponymy (V)	Manner	(buy, subscribe) (sell, retail)
Entailment (V)	X must mean doing Y	(buy, pay) (sell, give)

WordNet: Summary

WordNet: A lexicon of the English language.

Content	Adjectives, verbs, nouns and adverbs of the English language
Format	Visualization tool data downloadable in Prolog-like format
Main strength	High quality lexicon for English
Technique	Manual
Size	Words: 155k
	Senses: 117k
	Word-sense pairs: 207k
License	Proprietary, free use
Reference	[Miller, Comm ACM 1995]
URL	http://wordnet.princeton.edu

http://wordnet.princeton.edu/wordnet/man2.1/wnstats.7WN.html

Wikipedia

If a small number of people can create a knowledge base, how about a LARGE number of people?

Jimmy Wales

Wikipedia is a free online encyclopedia

- started in 2001
- driven by Wikimedia Foundation, and a large number of volunteers
- goal: build world's largest encyclopedia

Wikipedia: Articles and Attributes

Elvis Presley

1 Article == 1 Page == 1 Entity

Elvis Aaron Presley (January 8, 1935 – August 16, 1977) was one of the most popular American singers of the 20th century.

Full text information

Infobox:
Tabular information in the form
Attribute: Value

A page is in one or multiple categories.
Categories form a hierarchy

Categories: American Rock singers

Elvis Presley

Publicity photo for *Jailhouse Rock* (1957)

Background information	
Birth name	Elvis Aaron Presley
Born	January 8, 1935 Tupelo, Mississippi, United States
Died	August 16, 1977 (aged 42) Memphis, Tennessee, United States
Genres	Rock and roll, pop, rockabilly, country, blues, gospel, R&B
Occupations	Musician, actor
Instruments	Vocals, guitar, piano
Years active	1954–77

Wikipedia: Summary

Wikipedia: A free online encyclopedia.

Content	Entities of public interest (people, geography, music…)
Format	Full text, downloadable in XML
Main strength	Good quality, large coverage, free
Technique	Manual creation by the community
Size	Articles: 18m (3.6m in English)
	Languages: 281
License	Creative Commons Attribution-ShareAlike (CC-BY-SA)
URL	http://download.wikimedia.org/

Outline for Part II

- **Knowledge Representation ✔**
- **Public Knowledge Bases:**
  - **Manually constructed knowledge bases ✔**
  - **Knowledge bases from Wikipedia**
  - **Knowledge bases beyond Wikipedia**

Knowledge Bases from Wikipedia

Can we construct the knowledge bases automatically from Wikipedia?

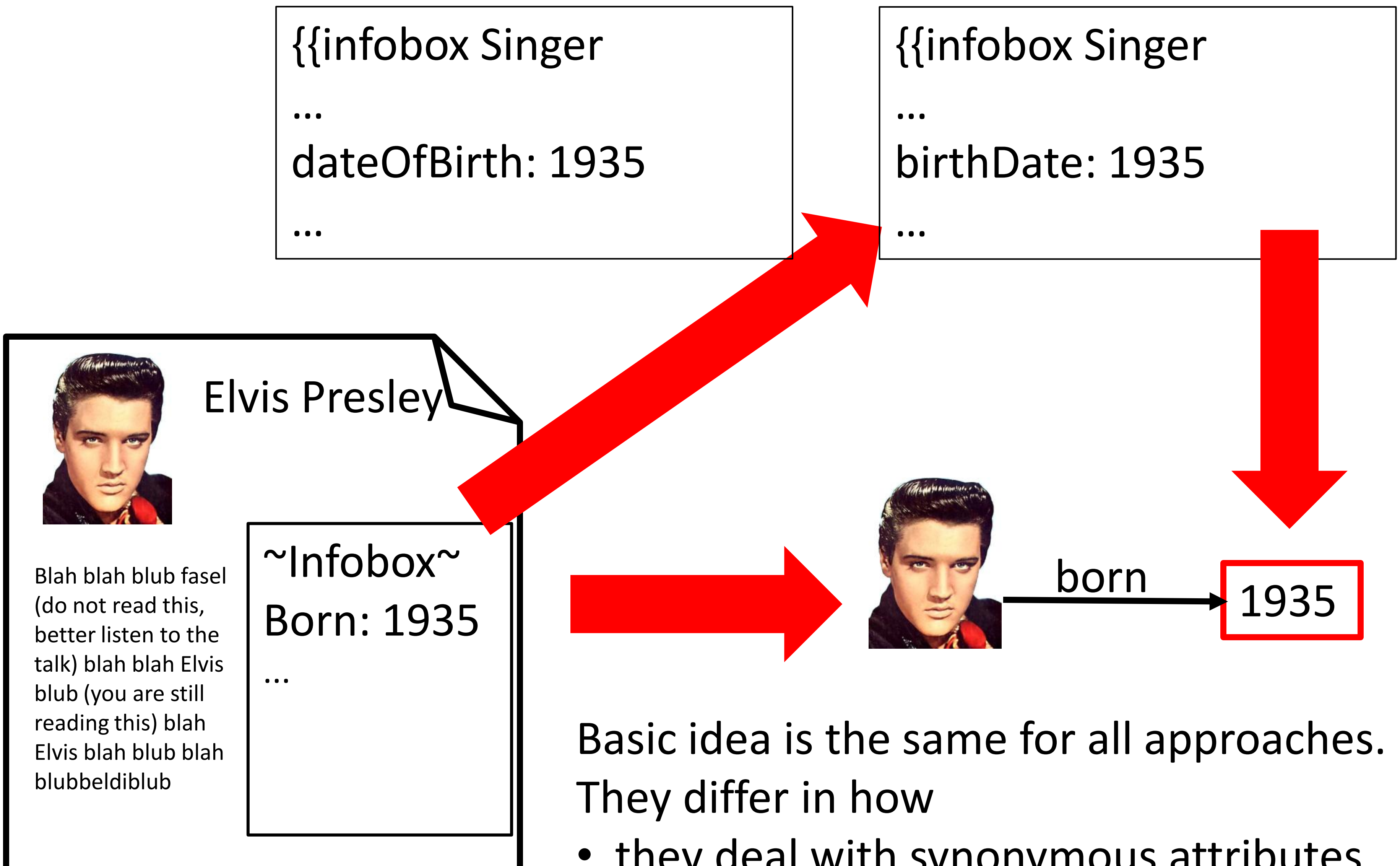
YAGO

DBpedia

Heidelberg Institute for Theoretical Studies | HITS

Freebase

Basic idea



Basic idea is the same for all approaches.
They differ in how

- they deal with synonymous attributes
- they construct a taxonomy

WikiNet

{{infobox Singer
...
dateOfBirth: 1935
...

{{infobox Singer
...
birthDate: 1935

birthDate

Category: 1935 births

birthDate

1935

WikiNet started in 2010 and is driven by the Heidelberg Institute for Theoretical studies (HITS)

Main idea: Categories and infobox attributes cross-fertilize

WikiNet: Summary

Content	Entities of public interest
Format	inverted index as plain text
Sources	Wikipedia
Main strength	Focus on multilinguality
Technique	Extraction from Wikipedia, propagation of category and infobox attributes
Size	Entities: 3 m Facts: 50m Relations: 500
License	Creative Commons BY-SA
URL	http://www.h-its.org/english/research/nlp/download/wikinet.php
References	[Nastase, LREC 2010]

DBpedia

Started in 2007, driven by Free U. Berlin, U. Leipzig, OpenLink

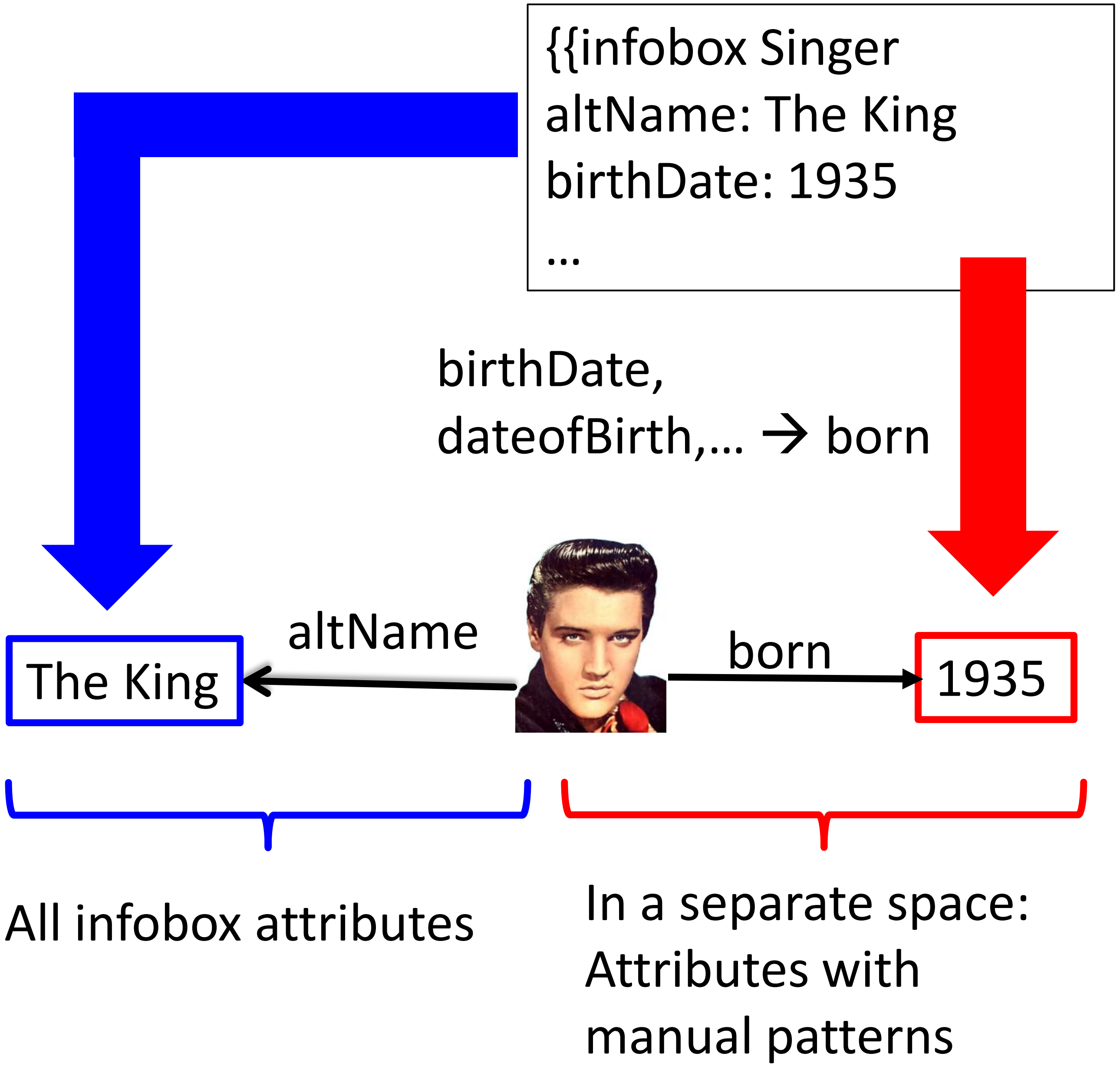
Main idea: Build a community of people who can define and curate the extraction patterns.

{{infobox Singer
...
birthDate: 1935
...

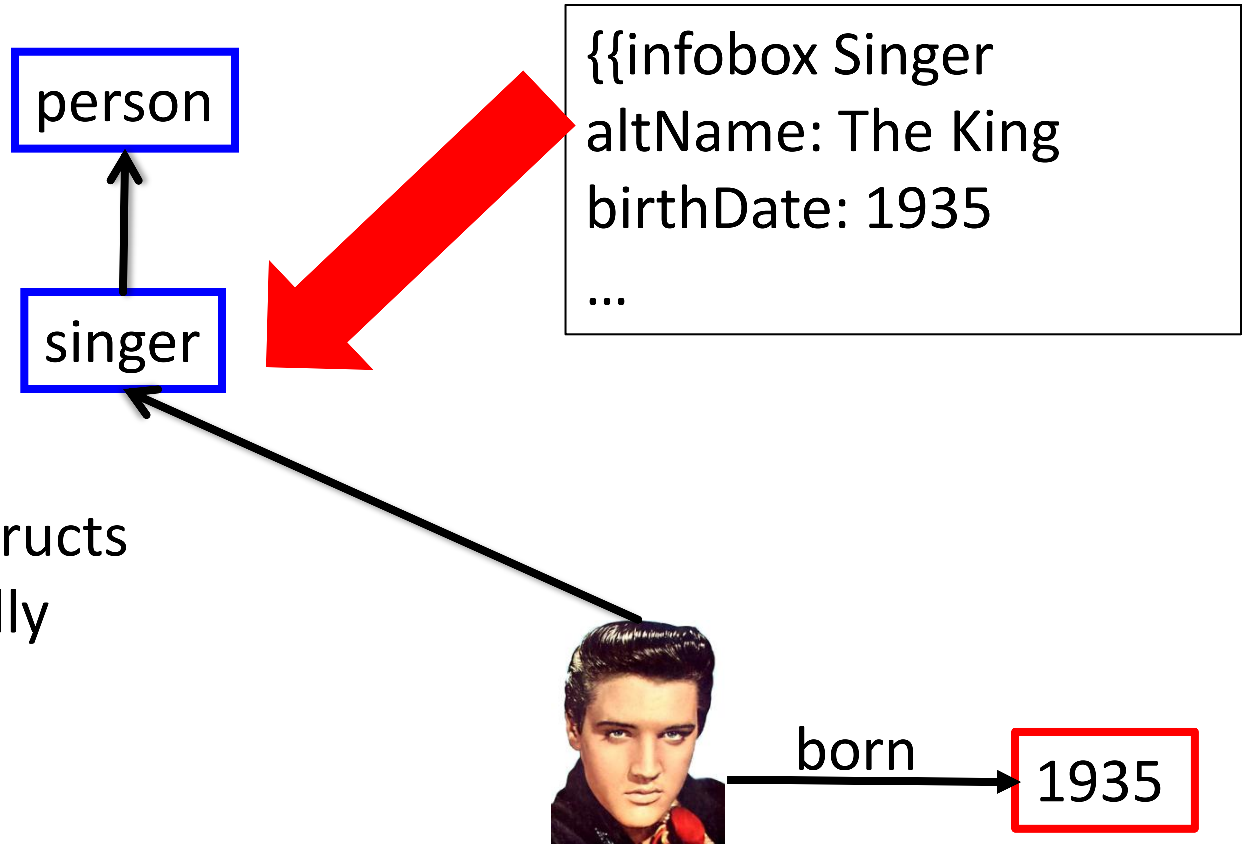
“birthDate” → born

born → 1935

DBpedia


{{infobox Singer
altName: The King
birthDate: 1935
...
birthDate,
dateofBirth,... → born
altName
born
The King
1935
All infobox attributes
In a separate space:
Attributes with
manual patterns

DBpedia

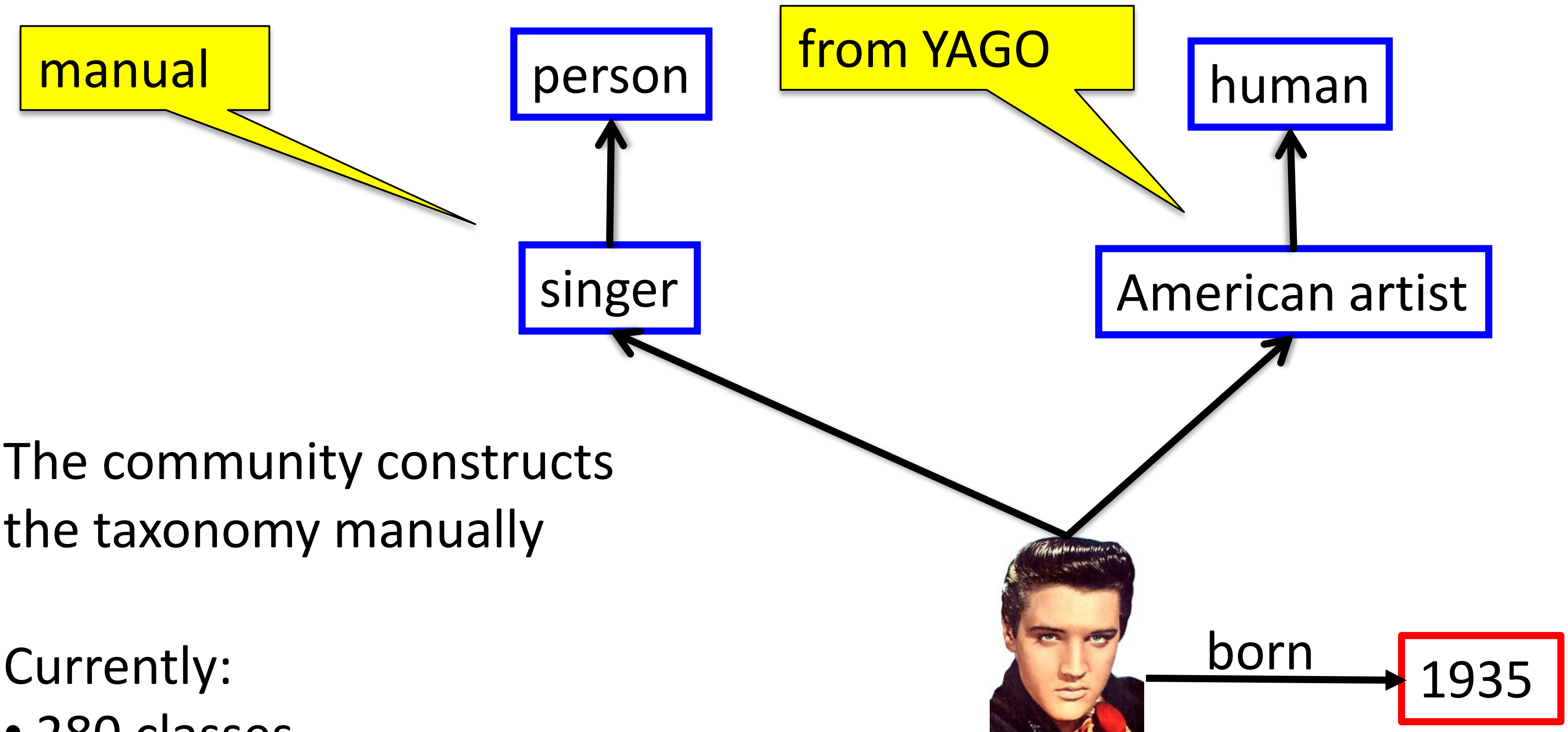


The community constructs the taxonomy manually

Currently:

- 280 classes
- covers 50% of all entities

DBpedia



The community constructs the taxonomy manually

Currently:

- 280 classes
- covers 50% of all entities

Complemented by the YAGO taxonomy

DBpedia

Content	Entities of public interest
Format	RDF, API, SPARQL
Sources	Wikipedia, YAGO/WordNet
Main strengths	Focus on coverage, interlinking with other data sets
Technique	Extraction from Wikipedia + manual supervision by the community
Size	Entities: 3.5m (in manual taxonomy: 1.7m) Facts: 670m Attributes: 9k (manually defined: 1k) Manual Classes: 280
License	CC-BY-SA & GNU FDL
URL	http://dbpedia.org
Reference	[Auer, ISWC 2007], [Bizer09, JWS 2009]

YAGO

YAGO

{{infobox Singer
...
birthDate: 1935
...

manually defined patterns

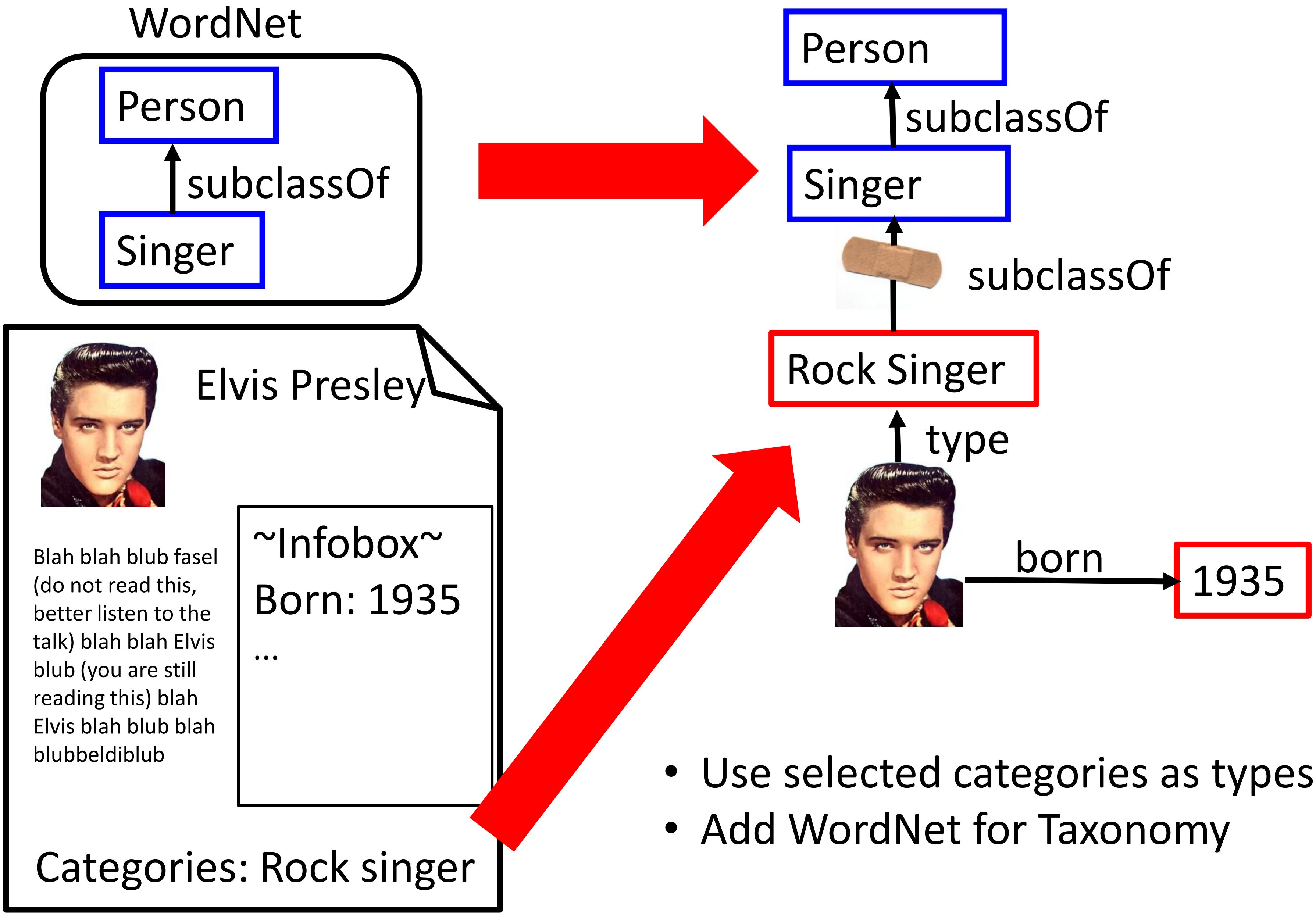
born

1935

YAGO (Yet Another Great Ontology) started as PhD thesis in 2007, now major project at the Max Planck Institute for Informatics in Germany

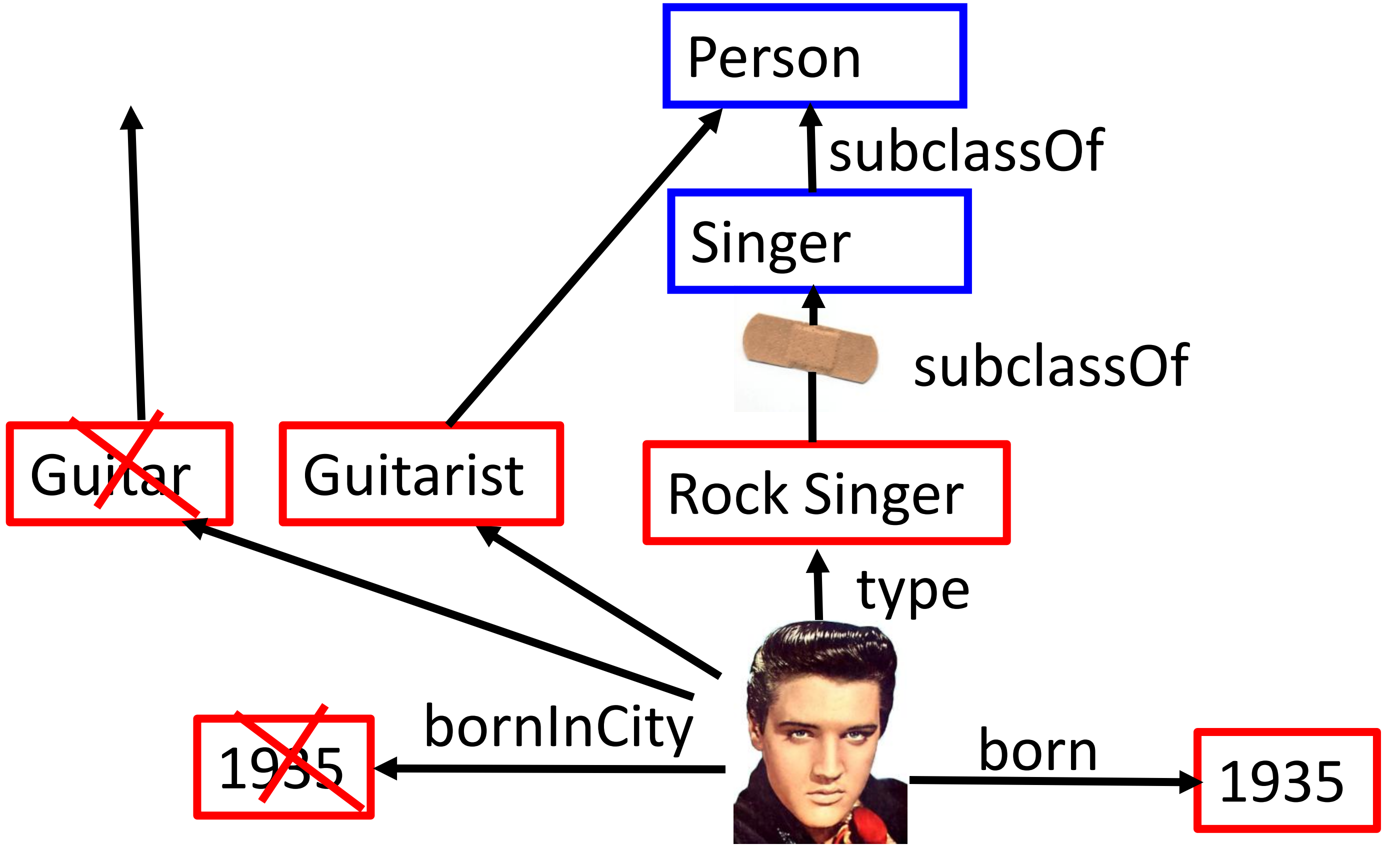
Main idea: Let the ontology check itself for precision.

YAGO: Classes



- Use selected categories as types
- Add WordNet for Taxonomy

YAGO: Consistency Checks



Check uniqueness of functional arguments

Check domains and ranges of relations

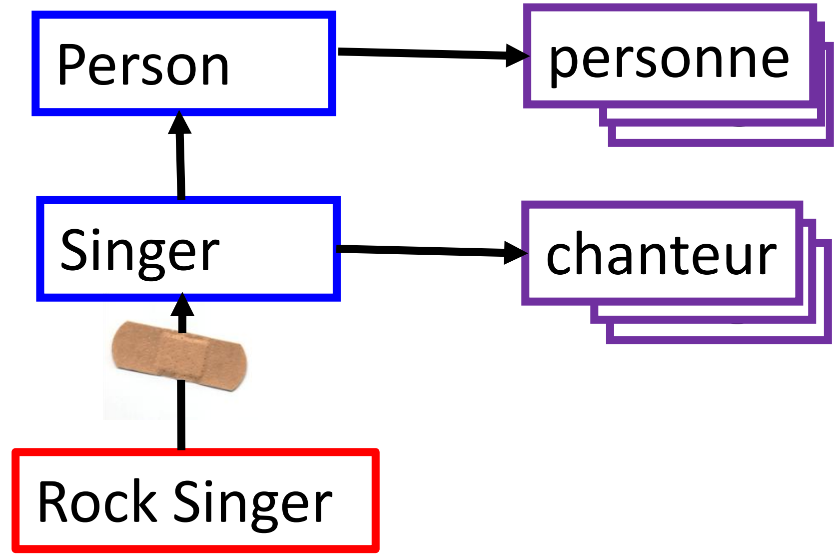
Check type coherence

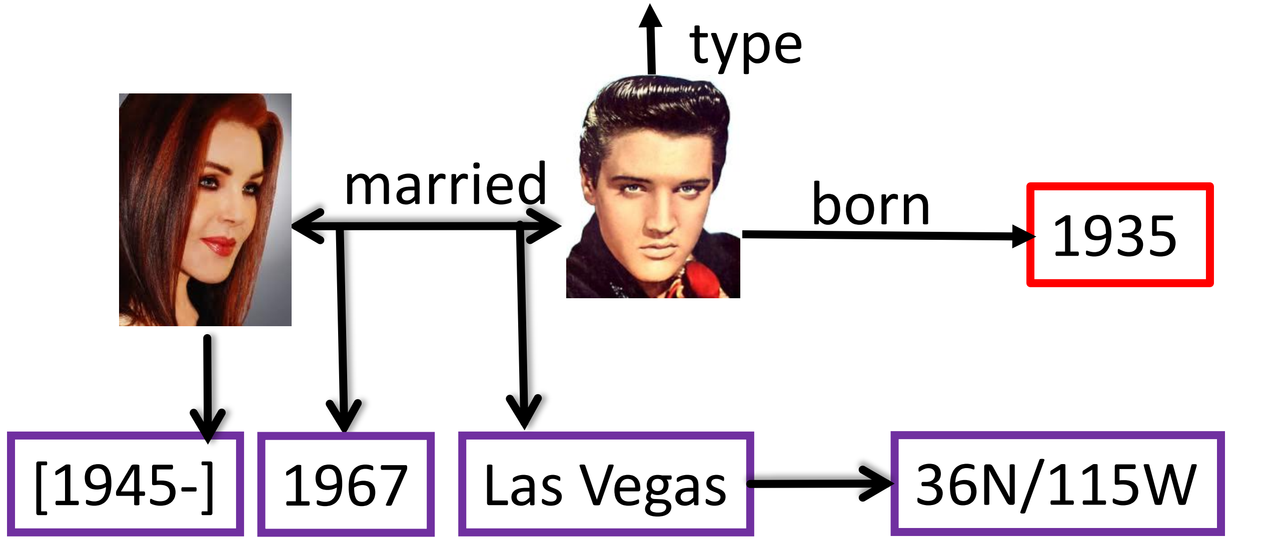
YAGO: Annotations

Adding in GeoNames

and some rule-based fact deduction

Adding in the Universal WordNet





≈ 1/3 of entities and facts have a geospatial and/or temporal annotation

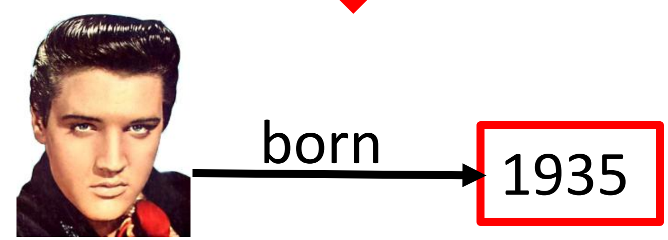
YAGO: Summary

Content	Entities of public interest
Format	TSV, RDF, XML, N3, Web Interface
Sources	Wikipedia, WordNet, Geonames
Main strength	Focus on precision, geotemporal annotations, multilingual
Precision	95%
Technique	Extraction from Wikipedia + matching with WordNet & Geonames + consistency checks
Size	Entities: 3 m (+ geonames -> 10m) Facts: 120m (+geonames -> 460m) Relations: 100, Classes: 200k, Languages: 200
License	Creative Commons BY-SA
URL	http://mpii.de/yago
References	[Suchanek, WWW 2007] [Hoffart, WWW 2011] [deMelo, CIKM 2010]

Freebase

Freebase started in 2000, driven by Metaweb, part of Google since 2010

Imports data from Wikipedia and other sources (e.g., ChefMoz, NNDB, and MusicBrainz).



Main idea:

In Wikipedia, people edit articles.
In Freebase, people edit facts.

Freebase


People
Person /people/person
edit
Date of birth:
Jan 8, 1935
edit
Place of birth:
location
containe
Tupelo
Lee Cou
Mississip
United S
edit
Country of nationality:
country
Jnite States of America
Select an item from the list:
United States of America
Country
Episcopal Church in the United States of Am
Religion
view more
Your item not in the list?
Create new Country
(Shift+Enter)
edit
Gender:
edit
Profession:
edit
Religion:
edit
Ethnicity:
edit
Parents:
edit
Children:
person
Lisa Marie Presley

Freebase: User Contribution

Edit Entities

- create new entities
- assign a new class to an entity
- add/change attributes
- connect to other entities
- upload/edit images

Edit Schema

- define new class
- specify attributes of the class
- only by creator/admin
- class is peer-reviewed & promoted by staff/admin

Review

- flag vandalism
- flag entities to be deleted
- vote on flagged content (3 unanimous vote, or expert as tie-breaker)

Data Game

- find aliases in Wikipedia
- extract dates of events from Wikipedia articles
- use Yahoo image search API

Freebase: Community

Experts

- act as tie breakers
- split entities
- “rewind” changes

Inducted by current experts.

Admins

- create new classes and attributes
- respond to community suggestions

Promoted by staff or other admins.

Members

- edit
- review,
- vote

Anyone can be a member.

Freebase: Summary

Freebase is a large collaborative knowledge base owned by Google.

Content	Entities with public information
Format	API, RDF
Construction	by the community data import from public sources
Sources	Wikipedia, Libraries, WordNet, MusicBrainz…
Main strength	free and large
Size	Facts: several millions
	Entities: 20 m
License	Creative Commons Attribution (CC-BY)
URL	http://download.freebase.com

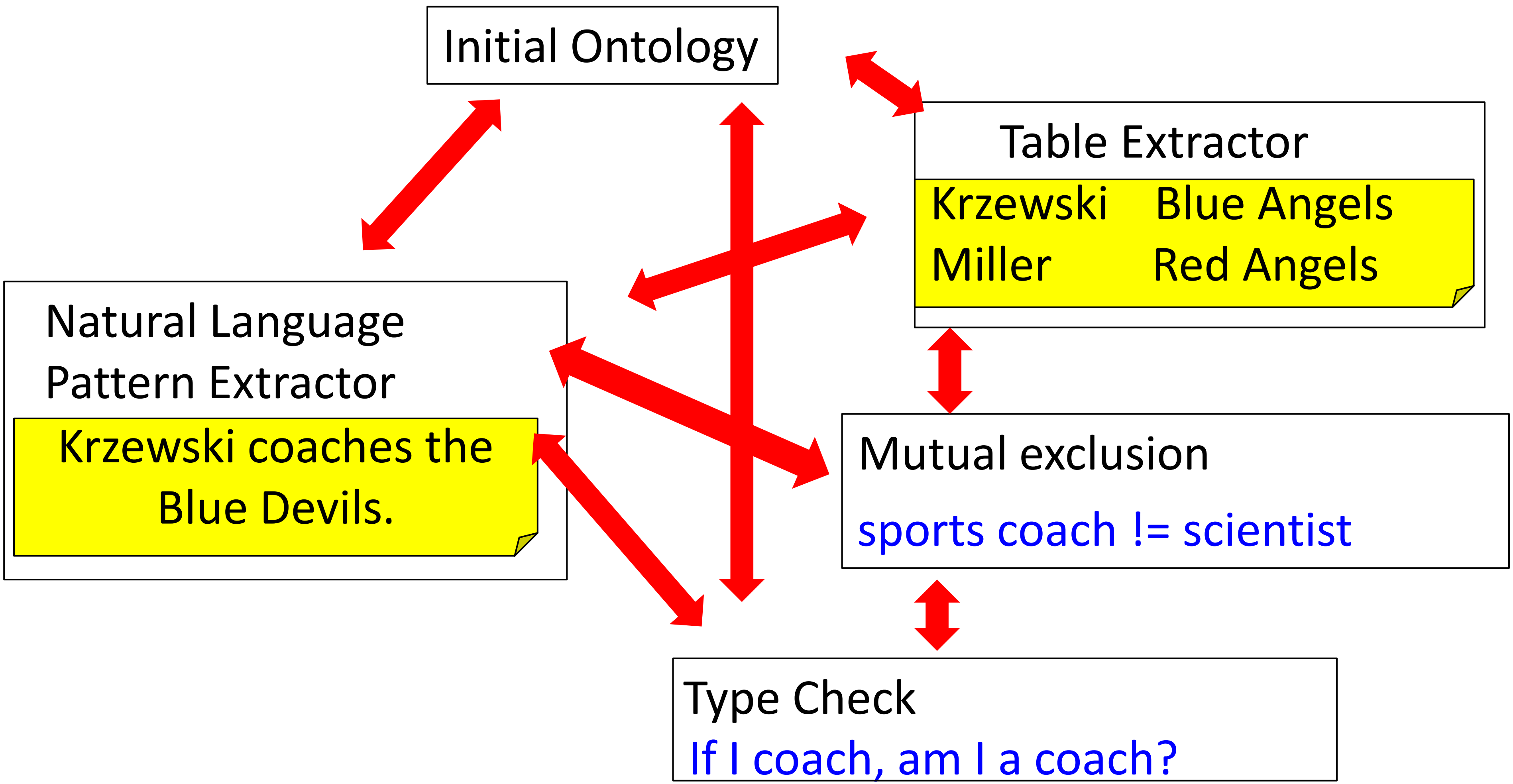
http://wiki.freebase.com/wiki/FAQ#How_big_is_Freebase.3F

Outline for Part II

– **Knowledge Representation** ✔
– **Public Knowledge Bases:**
  - **Manually constructed knowledge bases** ✔
  - **Knowledge bases from Wikipedia** ✔
  - **Knowledge bases beyond Wikipedia**

Read the Web/NELL

"Read the Web/NELL" is a project at the Carnegie Mellon University in Pittsburgh, PA, since 2009.



+ regular human feedback

Read the Web/NELL

NELL Know
CMU Read the Wel

- v
    - fung
    - plar
    - arch
    - bact
  - politica
  - color
  - language
  - programminglanguage
  - dateliteral
  - gamescore
  - nonneginteger
  - politicsissue
  - llcoordinate
- agent
  - animal
    - invertebrate
      - arthropod
        - arachnid
        - insect
        - crustacean
      - mollusk
    - vertebrate
      - amphibian
      - bird
      - fish

- arthropod (100.0%)
  - Seed
  - CPL @156 (100.0%) on 30-sep-2010 ["hind wings of _" "invertebrates , such as _" "_ swarm from" "other insects , including _" "_ marching home" "honeydew produce like _" "other insects , such as _" "_ do not eat wood" "many legs as _" "_ produce si have complete metamorphosis" "I do n't see anymore _" "ants , so _" "insecticide fo "such insects as _" "_ are the only insects" "red imported _" "insects like _" "social ir , such as _" "arthropods include _" "insect pests including _" "meaty foods like _" "_ pests , such as _" "other insects such as _" "insects , in particular _" "_ release a ph like _" "many insects , including _" "_ are social insects" "insect pests such as _" "_ a pests , including _" "arthropods , including _" "_ are beneficial insects" "_ are comm "arthropods , such as _"]
  - SEAL @151 (50.0%) on 26-sep-2010 [1]

kateretes (Seed)
mosquito (Seed)
peppered moth (Seed)
sap beetle (Seed)
tettigoniidae (Seed)
triatoma protracta (Seed)
honeylocust spider mite
grape flea beetle
blueberry leaf beetle
sugarcane moth borer
psychoda moth flies
bagworm moth
carpenterworm moths
leafcurl plum aphid
merchant grain beetle

http://rtw.ml.cmu.edu/rtw/

Read the Web/NELL

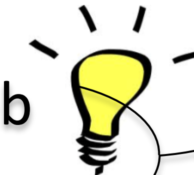
NELL is an information extraction system that runs continuously.

Content	Entities mentioned on Web pages
Format	TSV
Construction	by a perpetual extractor
Sources	The Web
Main strength	Not limited to a specific source
Size	Facts: 800k
	Categories & relations: 633
Reference	[Carlson, AAAI 2010]
URL	http://rtw.ml.cmu.edu/

http://rtw.ml.cmu.edu/rtw/overview

Wolfram Alpha

Can we compute answers to questions instead of showing Web pages?



WolframAlpha™ computational knowledge engine

Stephen Wolfram

Wolfram Alpha is a question answering system
- started in 2009
- driven by Wolfram Research
- goal: provide answers instead of Web pages

Wolfram Alpha: Content

Do professors have above average income?

- computes answers from an internal knowledge base of curated, structured data.
- stores not just facts, but also algorithms and models

True Knowledge

True Knowledge is a project similar in spirit, driven by William Tunstall-Pedoe's company.

True Knowledge
The Internet Answer Engine™ BETA

Wolfram Alpha & TrueKnowledge

	Wolfram Alpha	TrueKnowledge
Content	Facts, Algorithms, Models, Data	Entities of public interest
Sources	Public data	Wikipedia
Main strength	Computational NL queries on public data	Natural Language Query answering on public data
Technique	built-in data and algorithms, curated by experts	Extraction from Wikipedia + user feedback + consistency checks
Size	Facts: 10 trillion Algorithms: 50k	Entities: 25m Facts: 600m
License	Proprietary, access by Web form	Proprietary, access by API
URL	http://wolframalpha.com	http://trueknowledge.com

http://www.wolframalpha.com/about.html

http://trueknowledge.com

Outline for Part II

– **Knowledge Representation ✔**
– **Public Knowledge Bases:**
  - **Manually constructed knowledge bases ✔**
  - **Knowledge bases from Wikipedia ✔**
  - **Knowledge bases beyond Wikipedia ✔**

References for Part II

- Soeren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak and Zachary G. Ives: "DBpedia: A Nucleus for a Web of Open Data", ISWC 2007
- Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soeren Auer, Christian Becker, Richard Cyganiak and Sebastian Hellmann: "DBpedia - A Crystallization Point for the Web of Data", Journal of Web Semantics, 2009
- Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr. and Tom M. Mitchell: "Toward an Architecture for Never-Ending Language Learning}}, AAAI, 2010
- Gerard de Melo and Gerhard Weikum: "Towards a Universal Wordnet by Learning from Combined Evidence", CIKM, 2009
- Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum: "YAGO2: Exploring and Querying World Knowledge in Space, Context, and Many Languages", demo at WWW 2011
- D. Lenat: "CYC: A large-scale investment in knowledge infrastructure", Communications of the ACM, 1995
- G. A. Miller: "WordNet: A Lexical Database for English", Communications of the ACM Vol. 38, No. 11: 39-41, 1995
- Vivi Nastase, Michael Strube, Benjamin Boerschinger, Caecilia Zirn, and Anas Elghafari: "WikiNet: A Very Large Scale Multi-Lingual Concept Network", LREC, 2010
- I. Niles, and A. Pease: "Towards a Standard Upper Ontology", FOIS, 2001.
- F. M. Suchanek, G. Kasneci and G. Weikum: "Yago - A Core of Semantic Knowledge", WWW, 2007.
- World Wide Web Consortium: "RDF Primer. W3C Recommendation", 2004. http://www.w3.org/TR/rdf-primer/

Outline

- **Part I ✔**
  - **Machine Knowledge & Intelligent Applications**
- **Part II ✔**
  - **Knowledge Representation & Public Knowledge Bases**
- **Part III**
  - **Extracting Knowledge**
- **Part IV**
  - **Ranking and Searching**
- **Part V**
  - **Linked Data**
- **Part VI**
  - **Conclusion and Outlook**

Outline for Part III

- Domain-oriented IE vs. Open-domain IE
  - What to extract: entities, classes, binary & higher-arity relations
- Entities, Classes & Subsumptions
  - WordNet concepts, Wikipedia categories, entity disambiguation
- Pattern-based Knowledge Harvesting
  - Wrapper induction, WebTables, statistical pattern mining
- Probabilistic Extraction Models
  - HMMs, MEMMs, CRFs
- Constraints & Reasoning
  - MLNs, CCMs, FactorIE, SOFIE/PROSPERA
- Open-domain IE
  - ReadTheWeb, TextRunner, Omnivore, REVERB
- Advanced reasoning
  - Temporal/spatial annotations of facts

Two Paradigms in Information Extraction (IE)

Surajit
obtained his
PhD in CS from
Stanford University
under the supervision
of Prof. Jeff.
He later joined HP and
worked closely with
Umesh …

one or few sources

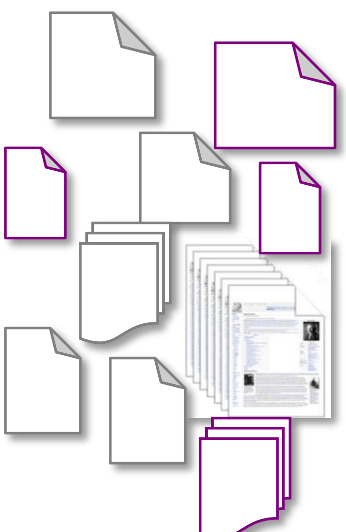
source-centric IE

1) recall !
2) precision

<Surajit> „obtained his“ <PhD>
<Surajit> „PhD in“ <CS>
<Surajit> „under supervision“ <Jeff>
<Surajit> „PhD from“ <Stanford>
<Surajit> „joined“ <HP>
<Surajit> „works with“ <Umesh>

Open-domain IE

near-human quality!



many sources

yield-centric harvesting

1) precision !
2) recall

hasAdvisor

Student	Advisor
Surajit	Jeff
Alon	Jeff
Jim	Mike
…	

almaMater

Student	University
Surajit	Stanford U
Alon	Stanford U
Jim	UC Berkeley
…	

Domain-oriented IE

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)), …

…

Binary Relations

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, 9-Oct-1940)
diedOn (JohnLennon, 8-Dec-1980)
marriedTo (JohnLennon, YokoOno)

Which additional & interesting relation types are there between given classes of entities?

competedWith(x,y), nominatedForPrize(x,y), …
divorcedFrom(x,y), affairWith(x,y), …
assassinated(x,y), rescued(x,y), admired(x,y), …

Higher-arity Relations & Reasoning

- **Time**, **location** & **provenance** annotations
- **Knowledge representation** – how do we model & store these?
- **Consistency reasoning** – how do we filter out inconsistent facts that the extractor produced? how do we quantify & manage uncertainty?

Facts (RDF triples):

1: (JimGray, hasAdvisor, MikeHarrison)

2: (SurajitSurajit, hasAdvisor, JeffJeff)

3: (Madonna, marriedTo, GuyRitchie)

4: (NicolasSarkozy, marriedTo, CarlaBruni)

5: (ManchesterU, wonCup, ChampionsLeague)

Reification:
"Facts about Facts":

6: (1, inYear, 1968)

7: (2, inYear, 2006)

8: (3, validFrom, 22-Dec-2000)

9: (3, validUntil, Nov-2008)

10: (4, validFrom, 2-Feb-2008)

11: (2, source, SigmodRecord)

12: (5, inYear, 1999)

13: (5, location, CampNou)

14: (5, source, Wikipedia)

Outline for Part III

- Domain-oriented IE vs. Open-domain IE ✔
  - What to extract: entities, classes, binary & higher-arity relations
- Entities, Classes & Subsumptions
  - WordNet concepts, Wikipedia categories, entity disambiguation
- Pattern-based Knowledge Harvesting
  - Wrapper induction, WebTables, statistical pattern mining
- Probabilistic Extraction Models
  - HMMs, MEMMs, CRFs
- Constraints & Reasoning
  - MLNs, CCMs, FactorIE, SOFIE/PROSPERA
- Open-domain IE
  - ReadTheWeb, TextRunner, Omnivore, REVERB
- Advanced reasoning
  - Temporal/spatial annotations of facts

Unary Relations – Classes, Instances, Subsumptions

- **Taxonomy construction**
  - Mapping Wikipedia categories onto Wordnet
  - Subsumption & consistency checks
  - Long tail of entities and classes
- **Entity disambiguation**
  - Individual vs. joint disambiguation

WordNet Thesaurus [Miller/Fellbaum 1998]

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

Word to search for: partner   Search WordNet

...change)   Change

3 concepts / classes & their synonyms (synsets)

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

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"*

http://wordnet.princeton.edu/

WordNet Thesaurus [Miller/Fellbaum 1998]

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 m
    - 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 th
    - S: (n) wife, married woman (a married woman; a man's partner in marr
  - *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)

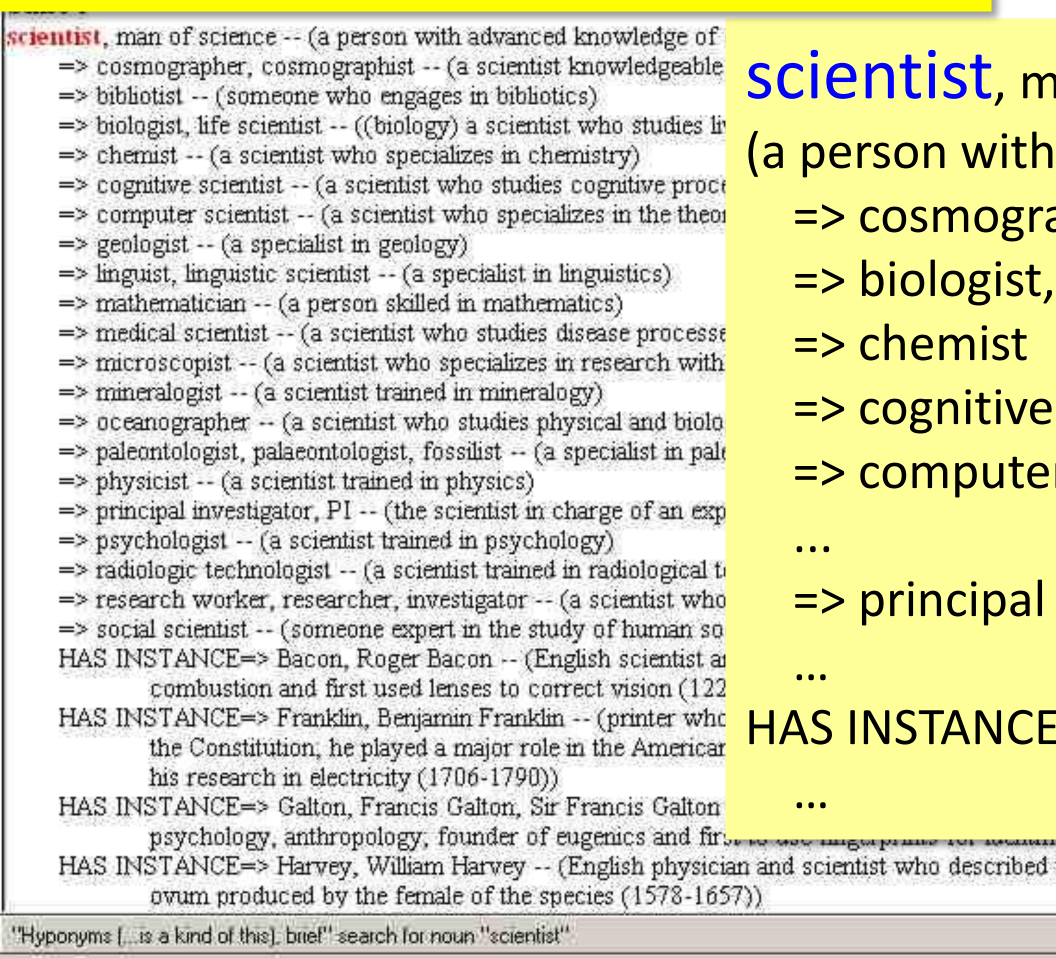
superclasses (hypernyms)

http://wordnet.princeton.edu/

WordNet Thesaurus [Miller & Fellbaum 1998]

> **100,000 classes** and lexical relations;
can be cast into
- **description logics** or
- **graph**, **with weights** for relation strengths (derived from co-occurrence statistics)

but:
only few individual entities (instances of classes)


scientist, man of science -- (a person with advanced knowledge of
=> cosmographer, cosmographist -- (a scientist knowledgeable
=> bibliotist -- (someone who engages in bibliotics)
=> biologist, life scientist -- ((biology) a scientist who studies li
=> chemist -- (a scientist who specializes in chemistry)
=> cognitive scientist -- (a scientist who studies cognitive proc
=> computer scientist -- (a scientist who specializes in the theo
=> geologist -- (a specialist in geology)
=> linguist, linguistic scientist -- (a specialist in linguistics)
=> mathematician -- (a person skilled in mathematics)
=> medical scientist -- (a scientist who studies disease processe
=> microscopist -- (a scientist who specializes in research with
=> mineralogist -- (a scientist trained in mineralogy)
=> oceanographer -- (a scientist who studies physical and biolo
=> paleontologist, palaeontologist, fossilist -- (a specialist in pal
=> physicist -- (a scientist trained in physics)
=> principal investigator, PI -- (the scientist in charge of an exp
=> psychologist -- (a scientist trained in psychology)
=> radiologic technologist -- (a scientist trained in radiological t
=> research worker, researcher, investigator -- (a scientist who
=> social scientist -- (someone expert in the study of human so
HAS INSTANCE=> Bacon, Roger Bacon -- (English scientist a
combustion and first used lenses to correct vision (122
HAS INSTANCE=> Franklin, Benjamin Franklin -- (printer wh
the Constitution; he played a major role in the America
his research in electricity (1706-1790))
HAS INSTANCE=> Galton, Francis Galton, Sir Francis Galton
psychology, anthropology; founder of eugenics and fir
HAS INSTANCE=> Harvey, William Harvey -- (English physician and scientist who described
ovum produced by the female of the species (1578-1657))
"Hyponyms (...is a kind of this), brief" search for noun "scientist"

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/

Tapping on Wikipedia Categories

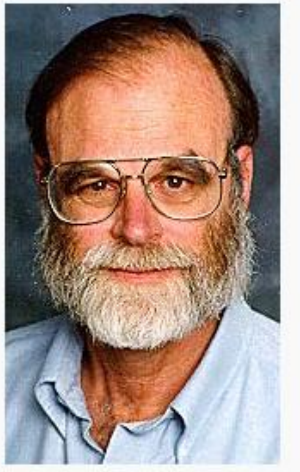
Jim Gray (computer scientist)

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

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

Tapping on Wikipedia Categories

Madonna (entertainer)

From Wikipedia, the free encyclopedia

Madonna (born **Madonna Louise Ciccone**; August 16, 1958) is an American recording artist, actress and entrepreneur. Born in Bay City, Michigan, and raised in Rochester Hills, Michigan, she moved to New York City in 1977, for a career in modern dance. After performing as a member of the pop groups Breakfast Club and Emmy, she released her debut album, *Madonna*, in 1983 on Sire Records.

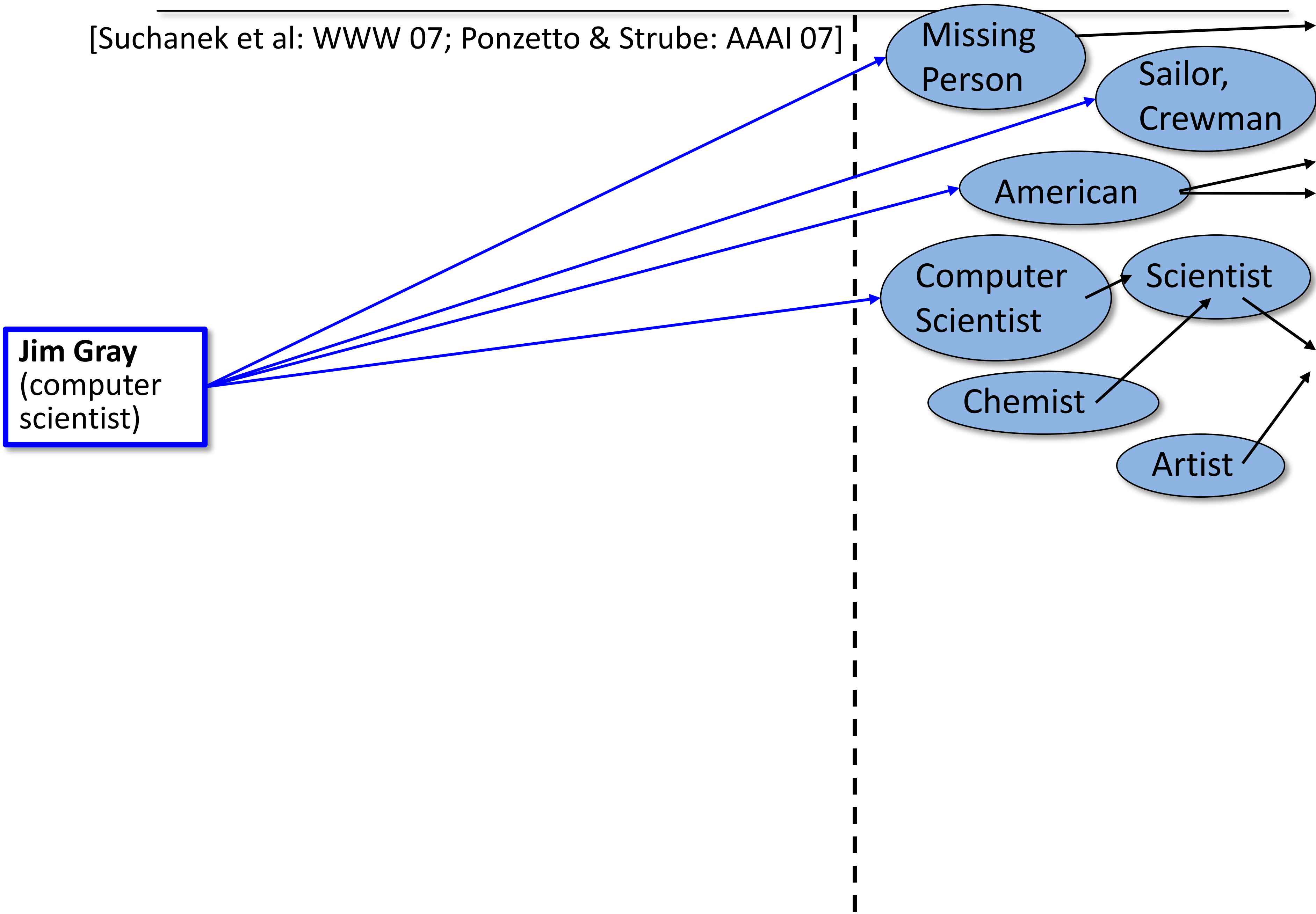
Madonna

Background information	
Birth name	Madonna Louise Ciccone
Also known as	Madonna Ciccone, Madonna Louise Veronica Ciccone
Born	August 16, 1958 (age 51) Bay City, Michigan, United States
Origin	New York, New York
Genres	Pop, dance
Occupations	Singer, songwriter, record producer, dancer, actress,

Categories: Madonna (entertainer) | 1958 births | 1980s singers | 1990s singers | 2000s singers | 2010s singers | Actors from Michigan | American businesspeople | American dance musicians | American dancers | American expatriates in the United Kingdom | American female singers | American film actors | American film producers | American musicians of Italian descent | American people of French-Canadian descent | American people of Italian descent | American philanthropists | American pop singers | American record producers | Converts to Judaism | American singer-songwriters | American writers | Best Musical or Comedy Actress Golden Globe (film) winners | BRIT Award winners | Electronica musicians | English-language singers | Female rock singers | Feminist artists | Grammy Award winners | Ivor Novello Award winners | Juno Award winners | Living people | MTV Europe Music Awards winners | MTV Video Music Awards winners | MTV Video Vanguard Award winners | Musicians from Michigan | People from Bay City, Michigan | People from Corona, Queens | People from Queens | People from Staten Island | Rock and Roll Hall of Fame inductees | University of Michigan alumni | Warner Bros. Records artists | World Music Awards winners | World record holders | Worst Actress Golden Raspberry Award winners | Worst Supporting Actress Golden Raspberry Award winners | Worst Screen Couple Golden Raspberry Award winners

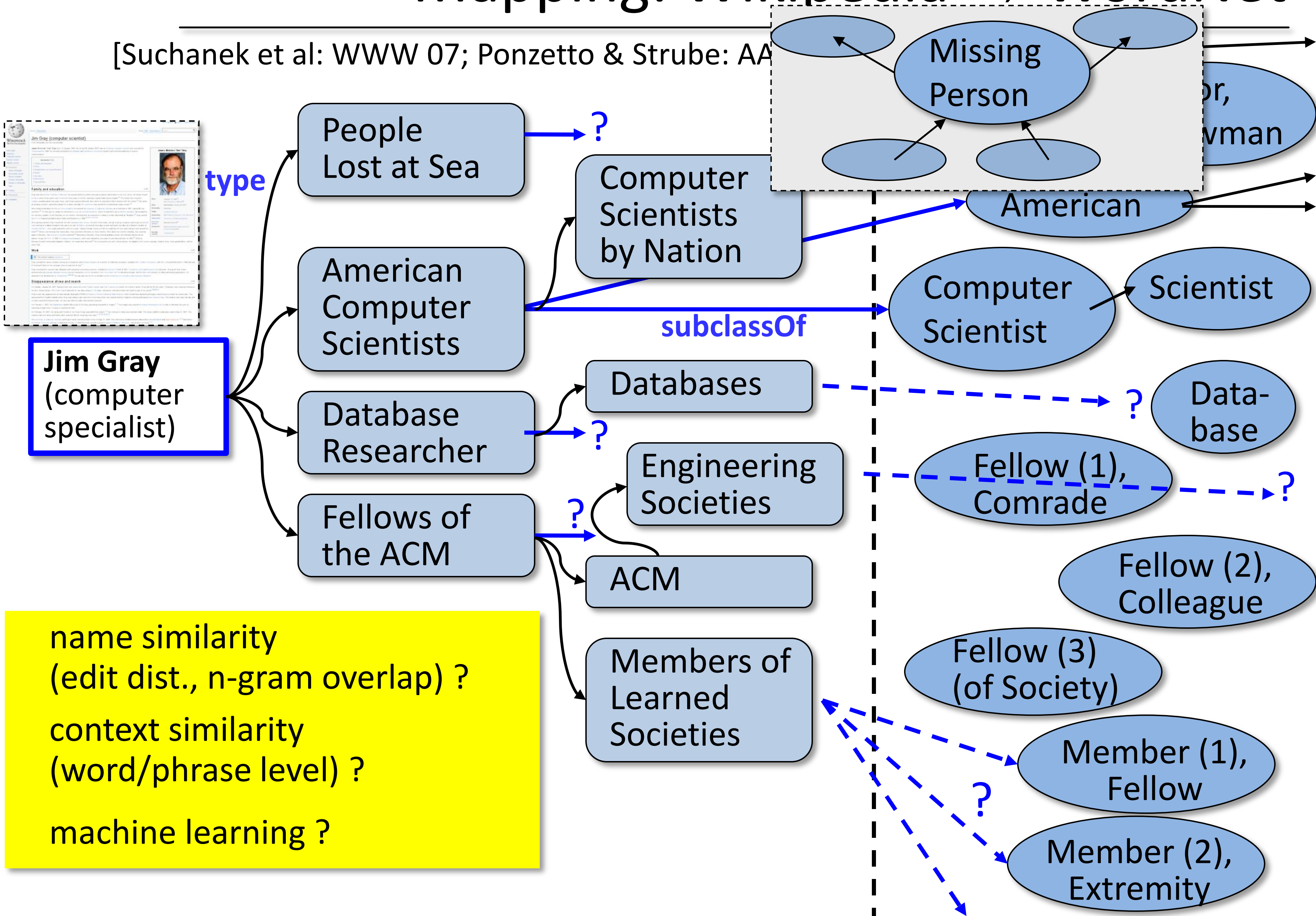
Mapping: Wikipedia → WordNet

[Suchanek et al: WWW 07; Ponzetto & Strube: AAAI 07]



Mapping: Wikipedia → WordNet

[Suchanek et al: WWW 07; Ponzetto & Strube: AA



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

context similarity
(word/phrase level) ?

machine learning ?

Mapping: Wikipedia → WordNet

[Suchanek et al: WWW 07; Ponzetto & Strube: AAAI 07]

<u>Given:</u> entity e in Wikipedia categories $c_1, \ldots, c_k$

<u>Wanted:</u> type(e,c) and subclassOf(c_i,c) for WordNet class c

<u>Problem:</u> vagueness & ambiguity of names $c_1, \ldots, c_k$

Analyzing category names → noun group parser:

$\underbrace{\text{American}}_{\text{pre-modifier}}\ \underbrace{\text{Musicians}}_{\text{head}}\ \underbrace{\text{of Italian Descent}}_{\text{post-modifier}}$

$\underbrace{\text{American Folk}}_{\text{pre-modifier}}\ \underbrace{\text{Music}}_{\text{head}}\ \underbrace{\text{of the 20th Century}}_{\text{post-modifier}}$

$\underbrace{\text{American Indy 500}}_{\text{pre-modifier}}\ \underbrace{\text{Drivers}}_{\text{head}}\ \underbrace{\text{on Pole Positions}}_{\text{post-modifier}}$

Head word is key, should be in plural for instanceOf

Mapping: Wikipedia → WordNet

[Suchanek et al: WWW 07; Ponzetto & Strube: AAAI 07]

<u>Given:</u> entity e in Wikipedia categories $c_1, \ldots, c_k$
<u>Wanted:</u> type(e,c) and subclassOf(c_i,c) for WordNet class c
<u>Problem:</u> vagueness & ambiguity of names $c_1, \ldots, c_k$

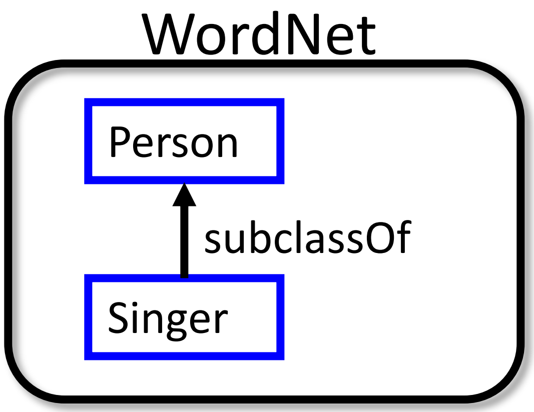
<u>Heuristic Method:</u>

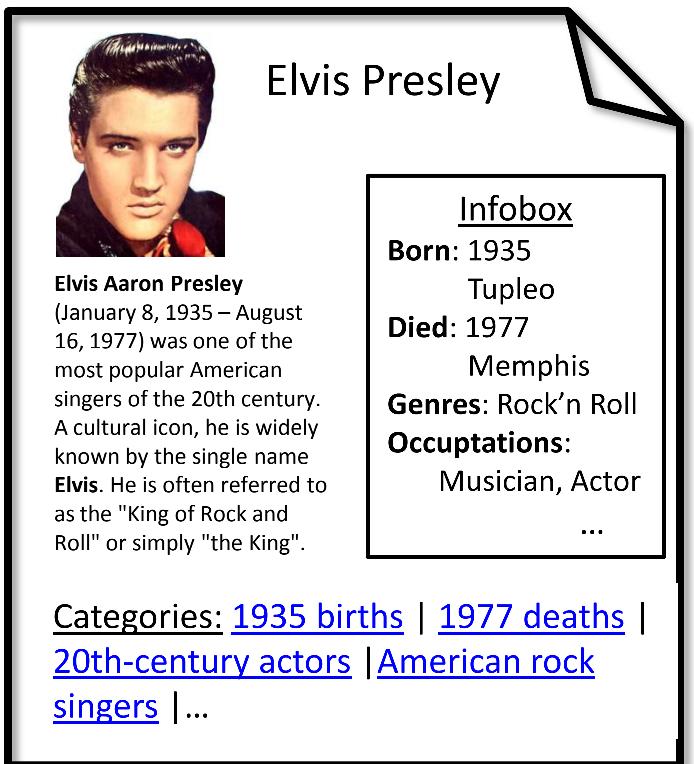
for each c_i do
 if head word w of category name c_i is plural
  1) find WordNet classes c, c', c'', … with
   synsets that contain a match of w
  2) choose best class c (from polysemous c, c', c'', …)
   and set $e \in c$
  3) expand w by pre-modifier from name c_i, returning w^+,
   and set $c_i \subseteq w^+ \subseteq c$

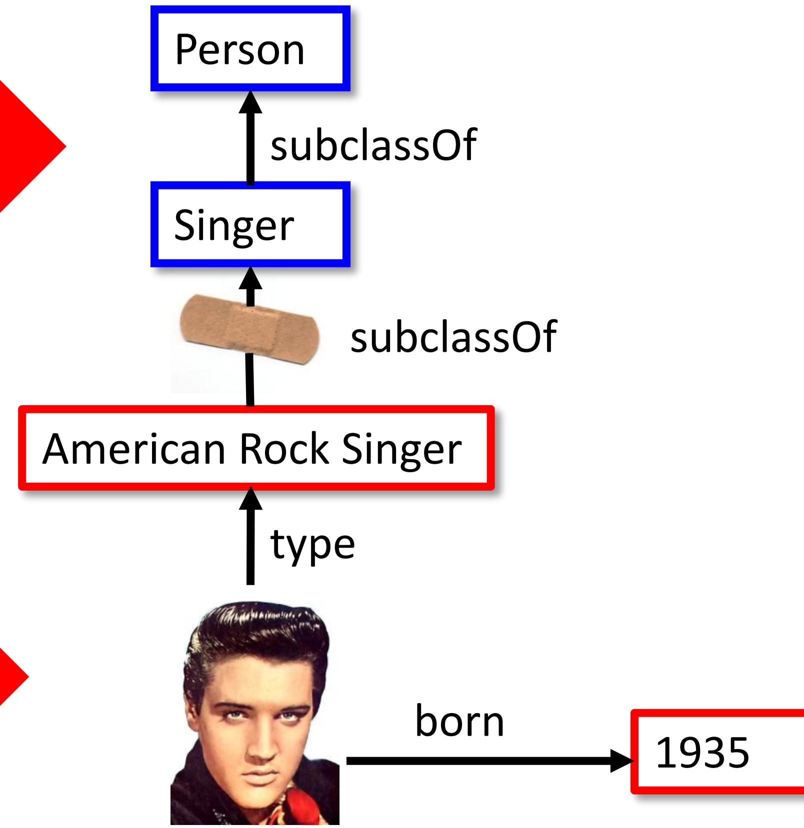
- can also derive features this way
- feed into supervised classifier

YAGO Concept Mappings

[Suchanek,Kasneci,Weikum: WWW 07, SIGMOD Rec. 08]







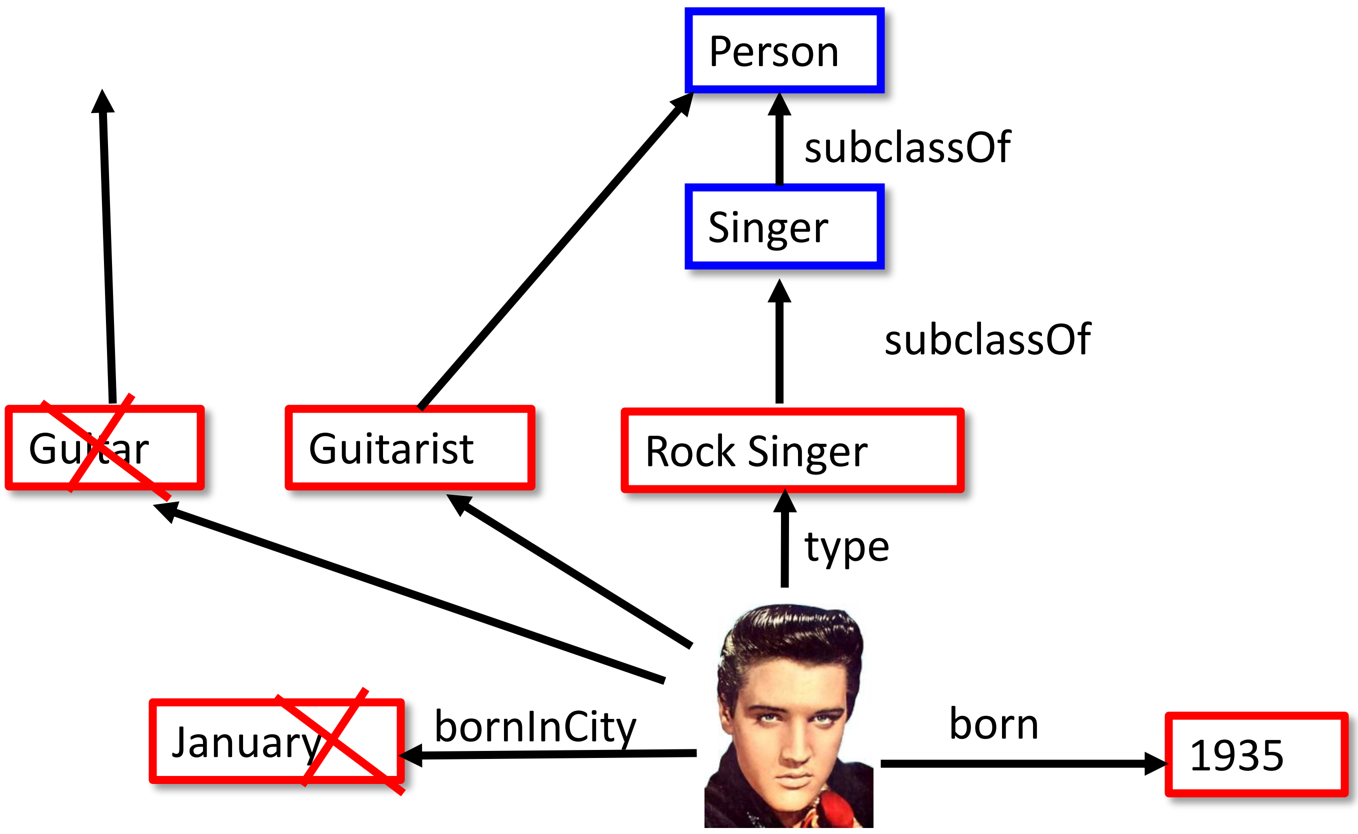
Exploit Infoboxes

Exploit conceptual categories

Add WordNet

YAGO Consistency Checks

[Suchanek,Kasneci,Weikum: WWW 07, SIGMOD Rec. 08]



Check uniqueness of entities and functional arguments

Check domains and ranges of relations

Check type coherence

Learning More Mappings

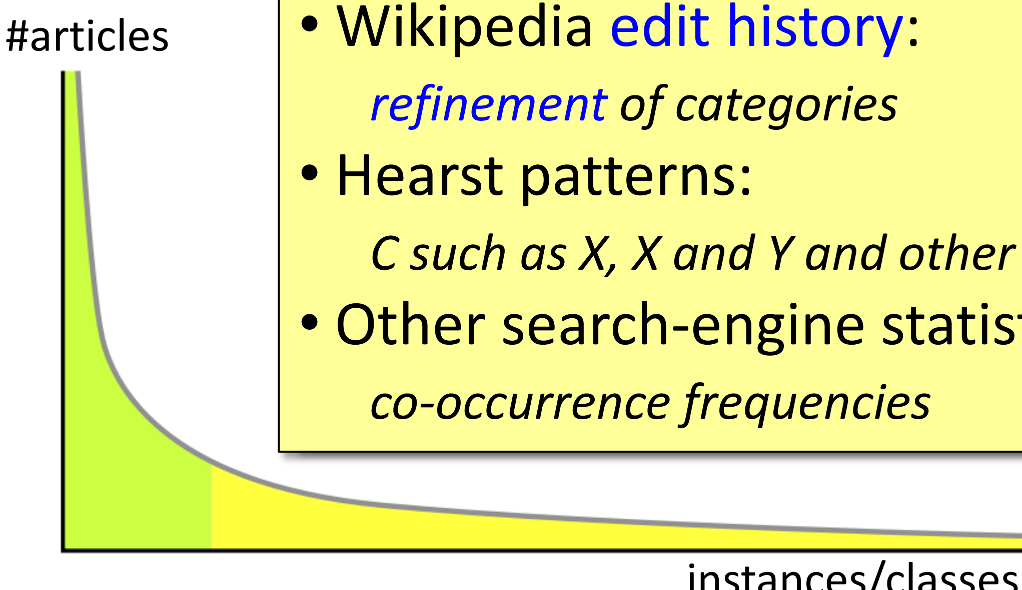
[Wu & Weld: CIKM 07, WWW 08]

Kylin Ontology Generator (KOG):

learn classifier for *subclassOf* across **Wikipedia** & **WordNet** using

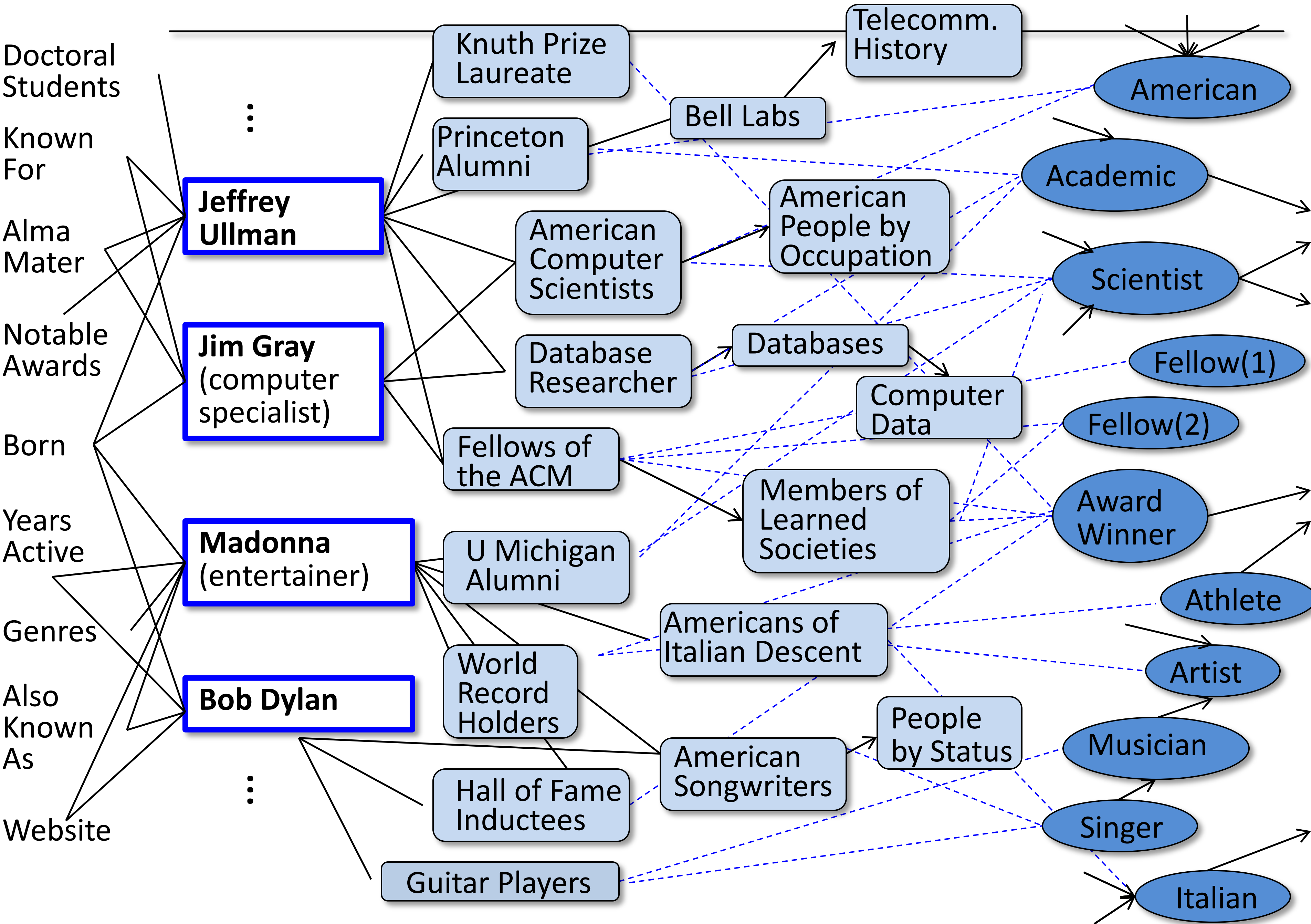
- YAGO as training data
- advanced ML methods (MLN's, SVM's)
- rich features from various sources

- Category/class name similarity measures
- Category instances and their infobox templates:
  template names, attribute names (e.g. knownFor)
- Wikipedia edit history:
  refinement of categories
- Hearst patterns:
  C such as X, X and Y and other C's, …
- Other search-engine statistics:
  co-occurrence frequencies

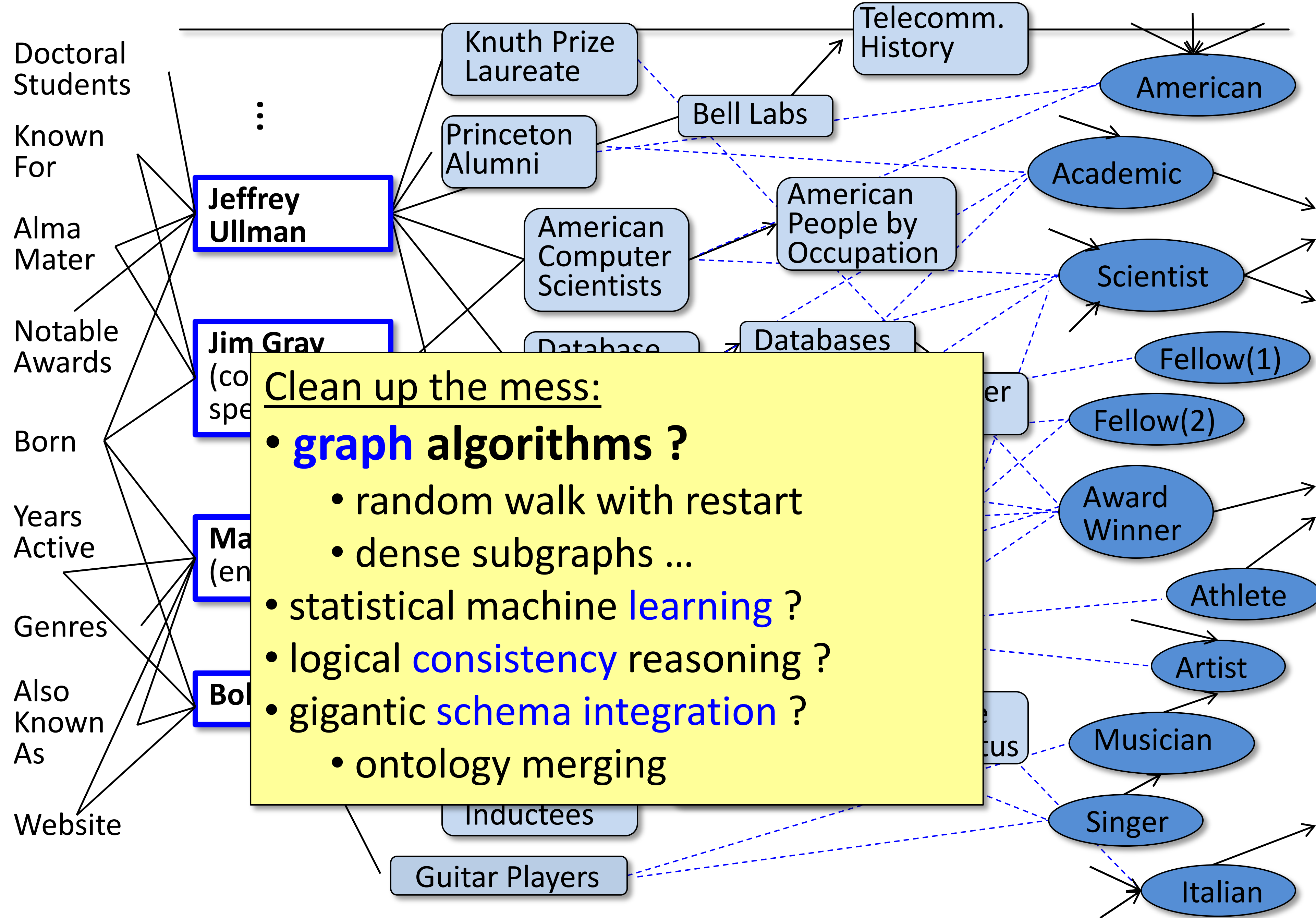


> 3 Mio. entities
> 1 Mio. w/ infoboxes
> 500 000 categories

Goal: Comprehensive & Consistent !



Goal: Comprehensive & Consistent !



<u>Clean up the mess:</u>

- **graph algorithms ?**
    - random walk with restart
    - dense subgraphs …
- statistical machine learning ?
- logical consistency reasoning ?
- gigantic schema integration ?
    - ontology merging

Long Tail of Class Instances

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

Long Tail of Class Instances

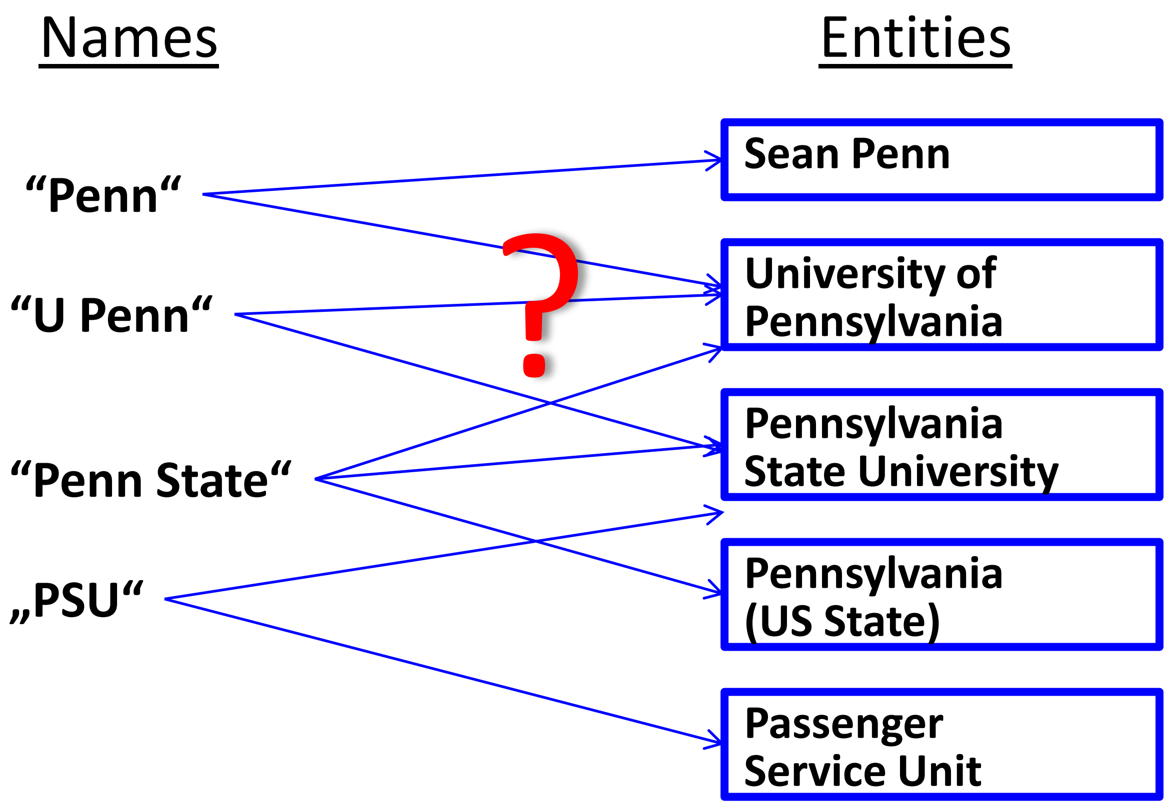
[Etzioni et al. 2004; Cohen et al. 2008; Mitchell et al. 2010]

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

- Start with seeds: a few class instances
- Find lists, tables, text snippets ("for example: …"), … 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

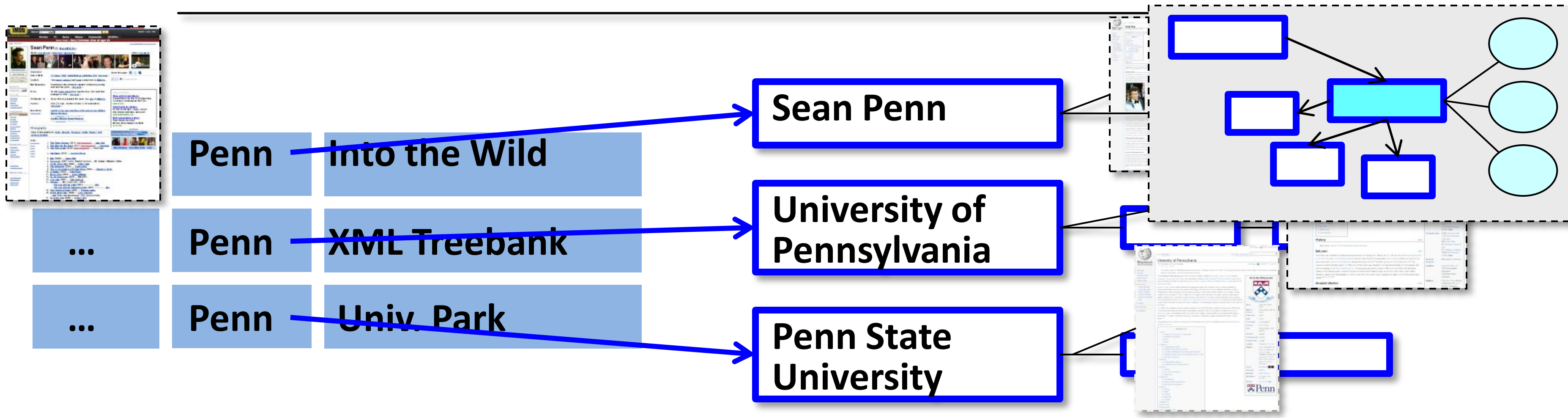
<u>But:</u>
Precision drops for classes with sparse statistics (DB profs, …)
Harvested items are names, not entities
Canonicalization (de-duplication) unsolved

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

Individual Entity Disambiguation



Typical approaches:

name similarity:
edit distances, n-gram overlap, …

context similarity: record level

context similarity: words/phrases level

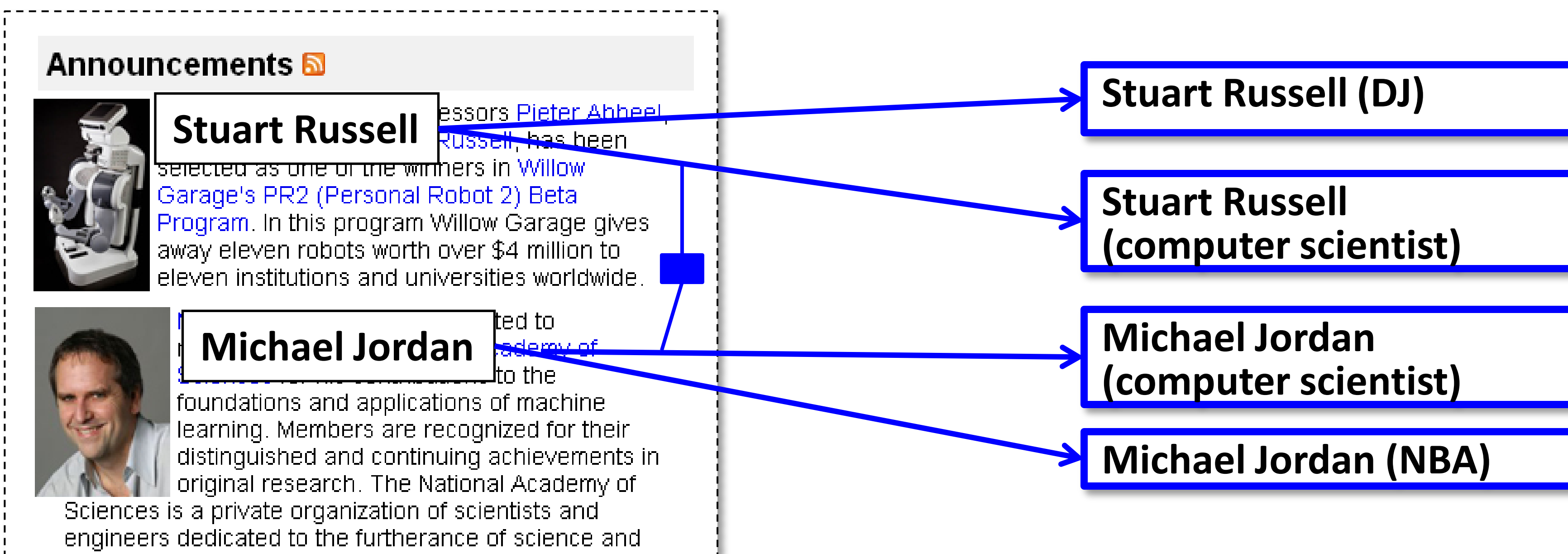
context similarity:
text around names, classes & facts around entities

Challenge: efficiency & scalability

Joint Disambiguation

[Doan et al: AAAI 05; Singla,Domingos: ICDM 07; Chakrabarti et al: KDD 09, …]

- 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 $\mu(n_1)=x_1 \in E_1$, $\mu(n_2)=x_2 \in E_2$, …
- Solve optimization problem



AIDA – Disambiguating Names in YAGO2

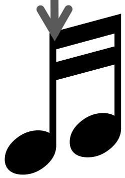
[Hoffart,Yosef,Weikum et al.: VLDB 11, EMNLP 11]

Mississippi, one of Bob's later songs, was first recorded by Sheryl on her album.

Features for Disambiguation

- Bob Hope
- Hurricane Bob
- Bob Quick
- …

Mississippi, one of Bob’s later songs, was first recorded by Sheryl on her album.



- Bob Dylan songs
- Sheryl Crow songs
- 1997 songs
- …

Prior	Similarity	Coherence
86%	0.9	Hurricane Bob
0.1%	3.1	Bob Dylan

How often did “Mississippi” link to this entity in Wikipedia?

entities related?

Objective Function

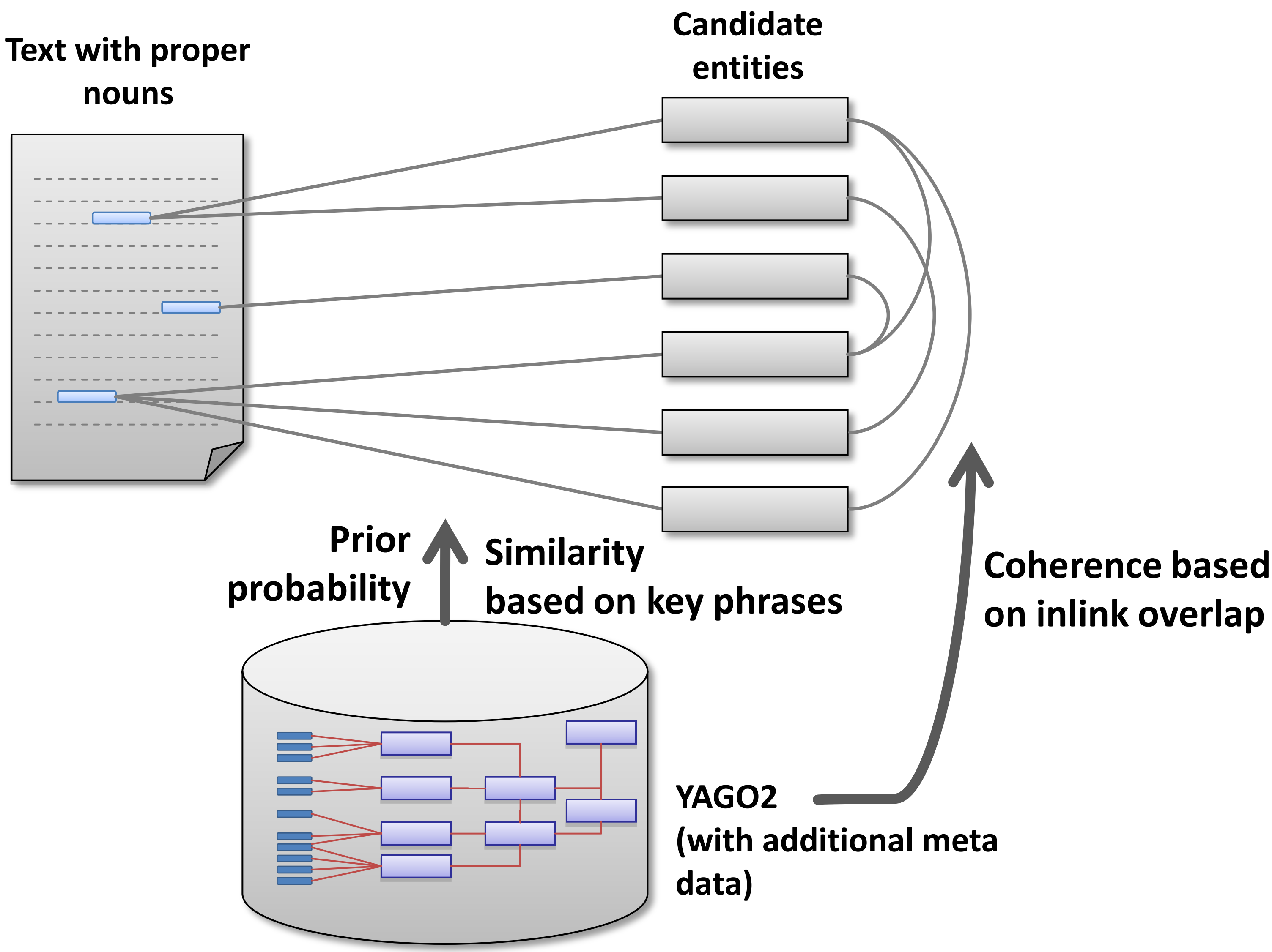
- Input
  - Mentions
    - context of mention ***ctx(m)***
    - entity candidates ***e + ctx(e)***
  - Features

Prior	*prior(m,e)*
Similarity	*sim(cxt(m),cxt(e))*
Coherence	*coh(e1,e2)*

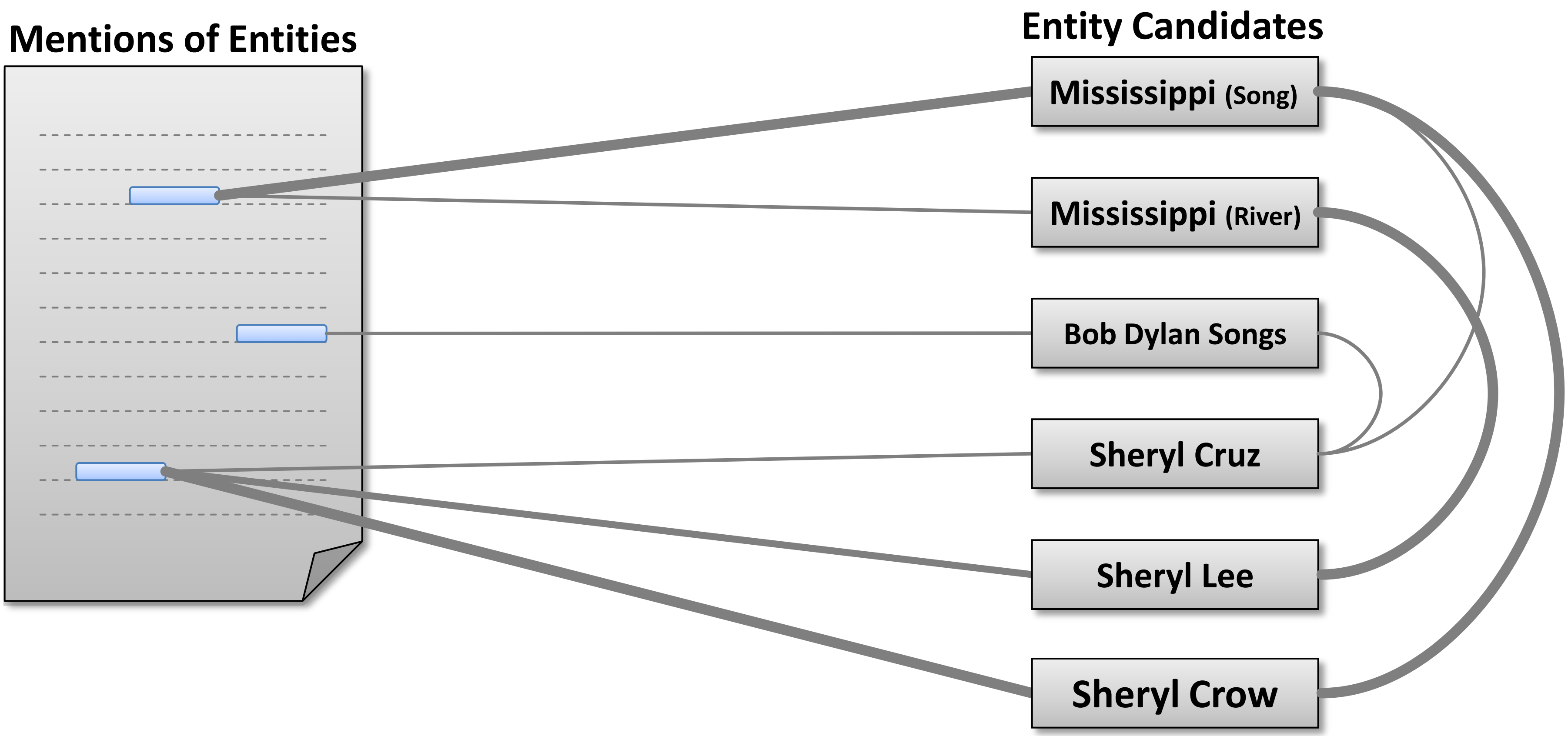
- Goal

$$\alpha \cdot \sum_{i=1}^{k} \mathrm{prior}(m_i, e_{j_i}) + \beta \cdot \sum_{i=1}^{k} \mathrm{sim}(\mathrm{cxt}(m_i), \mathrm{cxt}(e_{j_i})) + \gamma \cdot \mathrm{coh}(e_{j_1}, e_{j_2}, \ldots e_{j_k}) = \max!$$

Joint Disambiguation as Graph Problem



Graph Algorithm



Objective: Maximize the minimum weighted degree

Constraint: Keep at least one entity per mention

Outline for Part III

- Domain-oriented IE vs. Open-domain IE ✔
  - What to extract: entities, classes, binary & higher-arity relations
- Entities, Classes & Subsumptions ✔
  - WordNet concepts, Wikipedia categories, entity disambiguation
- Pattern-based Knowledge Harvesting
  - Wrapper induction, WebTables, statistical pattern mining
- Probabilistic Extraction Models
  - HMMs, MEMMs, CRFs
- Constraints & Reasoning
  - MLNs, CCMs, FactorIE, SOFIE/PROSPERA
- Open-domain IE
  - ReadTheWeb, TextRunner, Omnivore, REVERB
- Advanced reasoning
  - Temporal/spatial annotations of facts

Binary Relations – Which Sources to Pick?

- **Semi-structured data**

  The "Low-Hanging Fruit"
  - Wikipedia infoboxes & categories
  - HMTL lists & tables, etc.

- **Free text**
  - Hearst-patterns, clustering by verbal phrases
  - Natural-language processing
  - Advanced patterns & iterative bootstrapping ("Dual Iterative Pattern Relation Extraction")

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
Doctoral advisor	John McCarthy [1]
Notable awards	IEEE John von Neumann Medal, A. M. Turing Award

Serge Abiteboul

Citizenship	French
Nationality	French
Fields	Computer Science
Institutions	INRIA
Alma mater	University of Southern California
Doctoral	

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

Deterministic Pattern Matching

[Kushmerick 97; Califf & Mooney 99; Gottlob 01, …]

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)

Spouse(s)	Nicolas Sarkozy
Children	Aurélien Enthoven (with Raphaël Enthoven)

Spouse	Charles, Prince of Wales (29 July 1981 – 28 August 1996)[1]

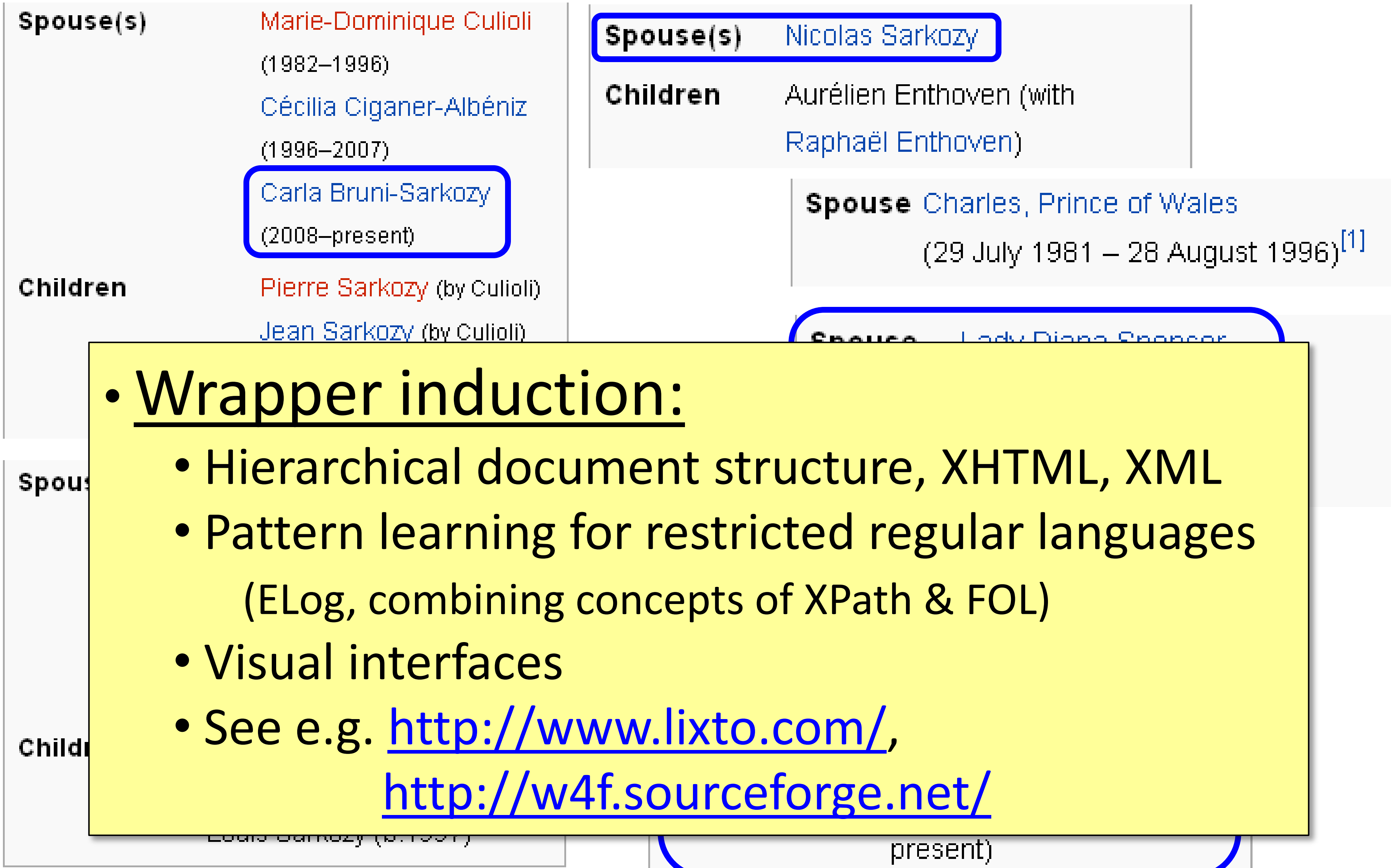
Spouse	Lady Diana Spencer 1981-1996 Camilla Parker Bowles m. 2005

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)

Spouse(s)	Lori Anne Allison (1983–1986)
Domestic partner(s)	Sherilyn Fenn (1985–1988) Winona Ryder (1989–1993) Kate Moss (1994–1998) Vanessa Paradis (1998–present)

Wrapper Induction

[Gottlob et al: VLDB 01, PODS 04,…]



- <u>Wrapper induction:</u>
  - Hierarchical document structure, XHTML, XML
  - Pattern learning for restricted regular languages (ELog, combining concepts of XPath & FOL)
  - Visual interfaces
  - See e.g. http://www.lixto.com/, http://w4f.sourceforge.net/

Tapping on Web Tables

[Cafarella et al: PVLDB 08; Sarawagi et al: PVLDB 09]

Academy Awards

(Reference:[1])

Year	Nominated work	Category	Result
1978	*The Deer Hunter*	Best Supporting Actress	Nominated
1979	*Kramer vs. Kramer*	Best Supporting Actress	Won
1981	*The*		
1982			

Academy Awards

Year	Category	Film	Result
	Academy Award for Best Actor	*Sweeney Todd: The Demon Barber of Fleet Street*	Nominated
	Academy Award for Best Actor	*Finding Neverland*	Nominated
	Academy Award for Best Actor	*Pirates of the Caribbean: The Curse of the Black Pearl*	Nominated

Academy Awards

Winner

- Best Art Direction
- Best Cinematography
- Best Makeup

Nominated

- Best Original Score
- Best Original Screenplay
- Best Foreign Language Film

Year	Winner Composer	Nominees
2000	***Crouching Tiger, Hidden Dragon*** – Tan Dun	• *Chocolat* – Rachel Portman • *Gladiator* – Hans Zimmer [3] • *Malèna* – Ennio Morricone • *The Patriot* – John Williams

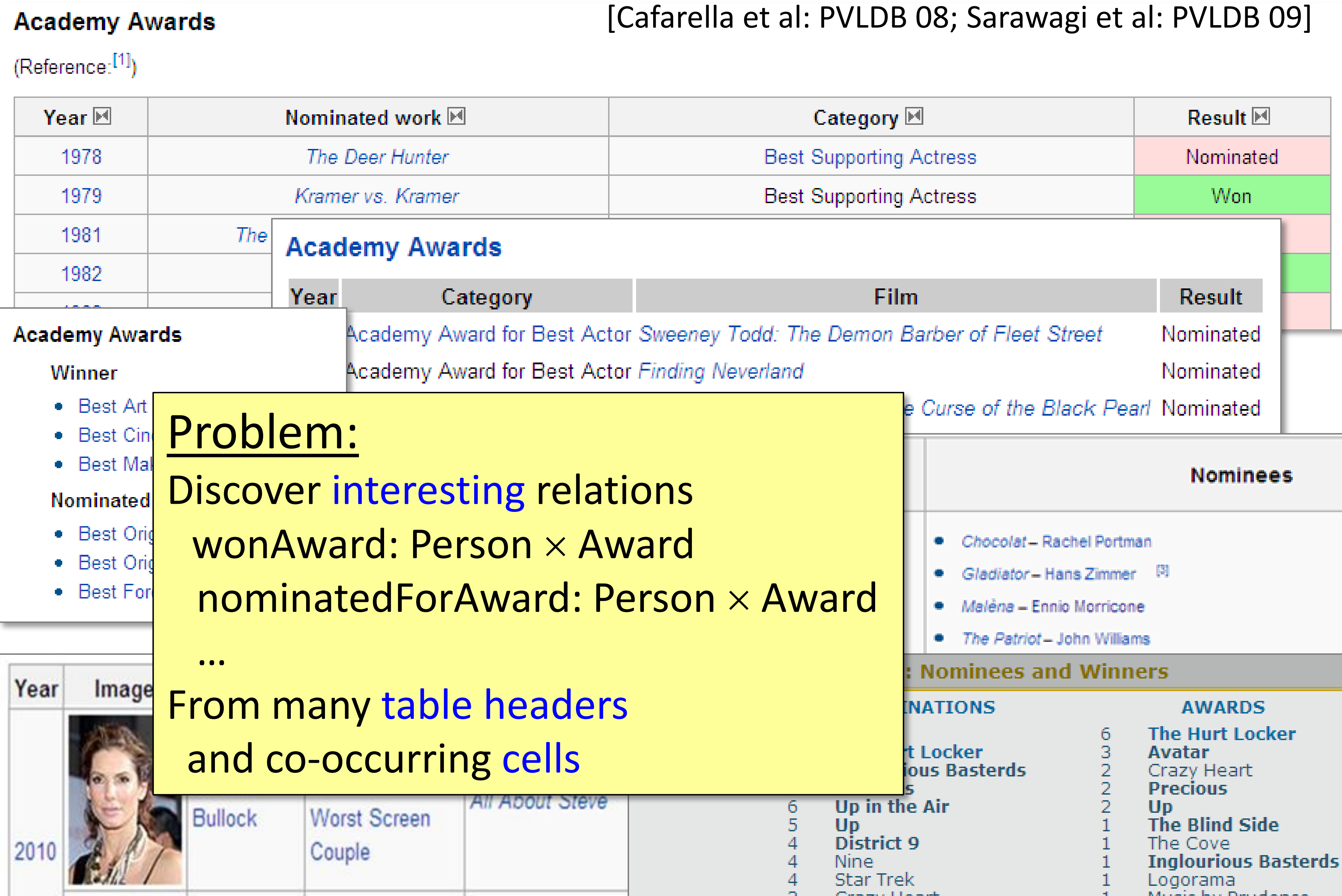
Year	Image	Recipient	Category	Film
2010		Sandra Bullock	Worst Actress Worst Screen Couple	*All About Steve*

Academy Awards (2009): Nominees and Winners

NOMINATIONS		AWARDS	
9	**Avatar**	6	**The Hurt Locker**
9	**The Hurt Locker**	3	**Avatar**
8	**Inglourious Basterds**	2	Crazy Heart
6	**Precious**	2	**Precious**
6	**Up in the Air**	2	**Up**
5	**Up**	1	**The Blind Side**
4	**District 9**	1	The Cove
4	Nine	1	**Inglourious Basterds**
4	Star Trek	1	Logorama

Tapping on Web Tables

[Cafarella et al: PVLDB 08; Sarawagi et al: PVLDB 09]



Problem:

Discover interesting relations

wonAward: Person × Award

nominatedForAward: Person × Award

…

From many table headers

and co-occurring cells

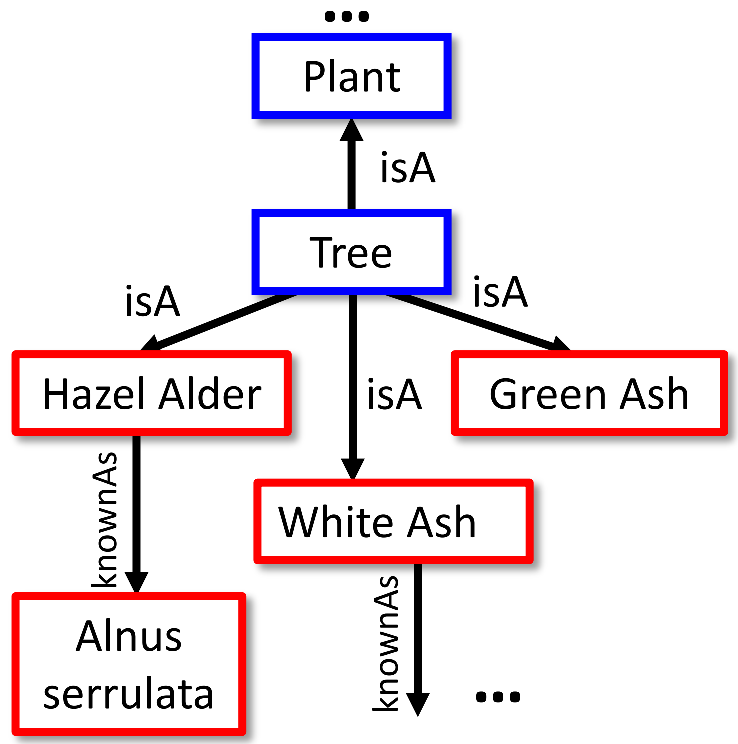
Recovering the Semantics of Web Tables

[Venetis,Halevy et al: PVLDB 11]

knownAs

Tree	Name
Hazel Alder	*Alnus serrulata*
Green Ash	*Fraxinus pennsylvanica*
White Ash	*Fraxinus americana*
Baldcypress	*Taxodium distichum*
Beech	*Fagus grandifolia*
River Birch	*Betula nigra*
Boxelder	*Acer negundo*
Red Cedar	*Juniperus virginiana*
Black Cherry	*Prunus serotina*
Sweet Cherry	*Prunus avium*
Crab Apple	*Malus coronaria*

http://www.hcforest.sailorsite.net/Elkhorn.html



Automatically enrich Web tables with **semantic annotations**

- Extract **instances of classes** using **Hearst patterns**
- Assign **most likely class labels** to columns
- Identify binary relations among pairs of columns
  → **Open IE tools** (TextRunner)

Large-scale statistics

- **100 Mio Web documents**
- **50 Mio queries** (for entity boundaries)

→ **60,000 classes** with > 10 instances

Relational Fact Extraction From Plain Text

- Hearst patterns [Hearst: COLING'92]
  - POS-enhanced regular expression matching in natural-language text

    $\langle NP_0$ {,} such as {NP_1, NP_2, … (and|or) }{,} $NP_n\rangle$

    $\langle NP_0$ {,}{NP_1, NP_2, … NP_{n-1}}{,} or other $NP_n\rangle$

    …

  "The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string."

  → isA("Bambara ndang", "bow lute")

- Noun classification from predicate-argument structures [Hindle: ACL'90]
  - Clustering of nouns by **similar verbal phrases**
  - Similarity based on **co-occurrence frequencies** (mutual information)

	beer	wine
drink	9.34	10.20
sell	4.21	3.75
have	0.84	1.38

Relational Fact Extraction From Plain Text

- Hearst patterns [Hearst: COLING'92]
  - POS-enhanced regular expression matching in natural-language text

    ⟨NP_0 {,} <u>such as</u> {NP_1, NP_2, … (<u>and</u>|<u>or</u>) }{,} NP_n⟩

    ⟨NP_0 {,}{NP_1, NP_2, … NP_{n-1}}{,} <u>or other</u> NP_n⟩

    …

  "The bow lute, <u>such as</u> the Bambara ndang, is plucked and has an individual curved neck for each string."

<u>Problem:</u>

Low recall

out of 8.6 M words only 152

occurrences of „such as" with matching noun conjugations

Difficult to extend to generic relations

(other than isA, partOf, etc.)

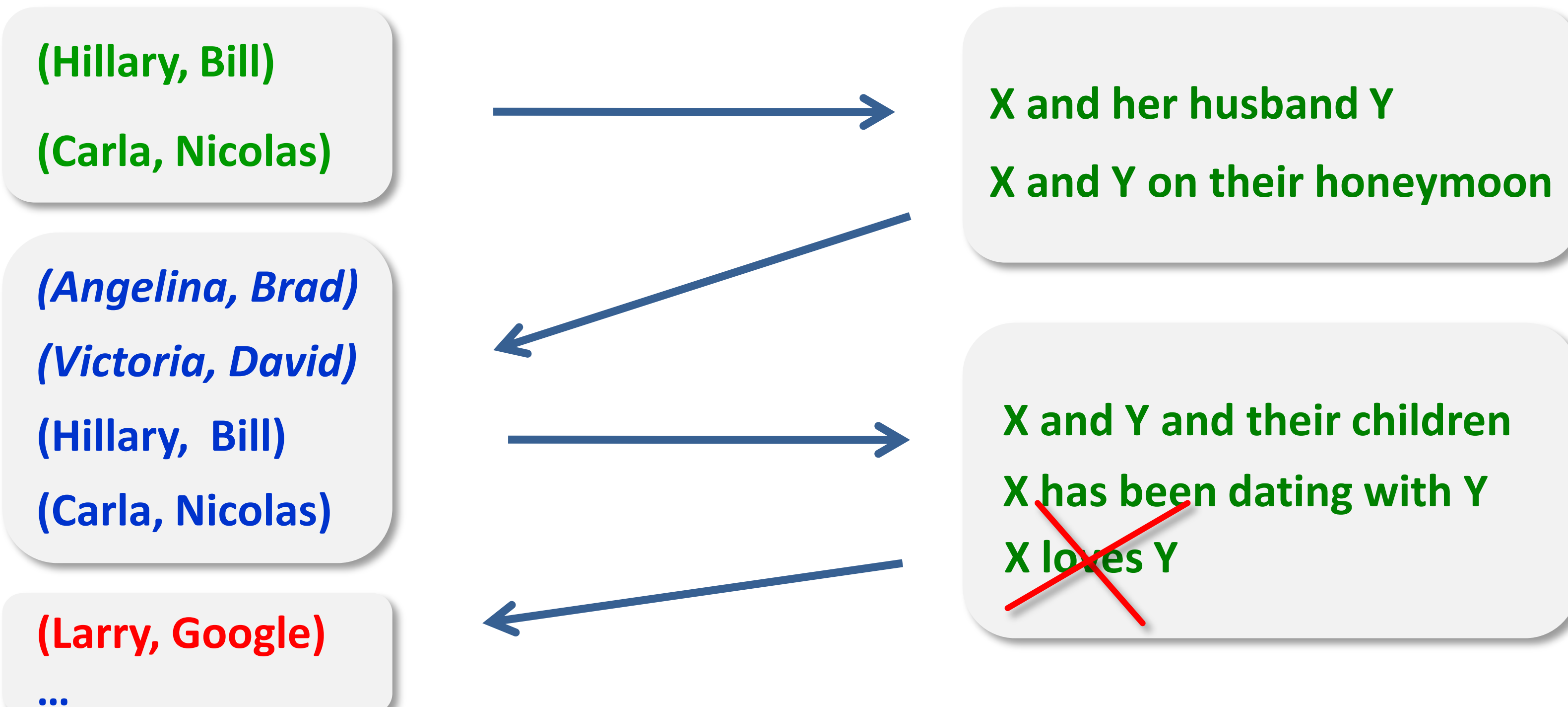
- **frequencies** (mutual information)

have	0.84	1.38

DIPRE/Snowball

[Brin: WebDB 98; Agichtein/Gravano: ACL 00, …]

- *Dual Iterative Pattern Relation Extraction* (DIPRE)
  - Semi-supervised, iterative gathering of facts and patterns
  - **Positive** & **negative** examples as **seeds** for a given **target relation**
    e.g. **+(Hillary, Bill)** **+(Carla, Nicolas)** **–(Larry, Google)**
  - Various tuning parameters for pruning low-confidence patterns and facts



DIPRE/Snowball/QXtract

[Brin: WebDB 98; Agichtein,Gravano: SIGMOD 01+03]

- *Dual Iterative Pattern Relation Extraction* (DIPRE)
  - Semi-supervised, iterative gathering of facts and patterns
  - **Positive** & **negative** examples as **seeds** for a given **target relation**

    e.g. **+(Hillary, Bill)** **+(Carla, Nicolas)** **–(Larry, Google)**
  - Various tuning parameters for pruning low-confidence patterns and facts

- Snowball/QXtract [Agichtein,Gravano: DL 00, SIGMOD 01+03]
  - Refined patterns and statistical measures
  - **>80% recall** at **>85% precision** over a large news corpus
  - Qxtract allows for **user feedback** in the iteration loop

Help from NLP: Dependency Parsing!

- Analyze **lexico-syntactic** structure of sentences
  - **Part-Of-Speech** (POS) tagging: HMMs, CRFs
  - **Dependency Parsing** (DP): probabilistic grammars
  - **Semantic Role Labeling** (SRL): map constituents onto semantic frames

  Prefer shorter dependency paths for fact candidates

Software tools:
CMU Link Parser: http://www.link.cs.cmu.edu/link/
Stanford Lex Parser: http://nlp.stanford.edu/software/lex-parser.shtml
Open NLP Tools: http://opennlp.sourceforge.net/
ANNIE Open-Source IE: http://www.aktors.org/technologies/annie/
LingPipe: http://alias-i.com/lingpipe/ (commercial license)
FrameNet: http://framenet.icsi.berkeley.edu/

Part-of-Speech Tagger Demo Results - Cognitive Computation Group - Mozilla Firefox

http://cogcomp.cs.illinois.edu/demo/pos/results.php

Part-of-Speech Tagger Demo Results

(Views)

The Part-of-Speech tagger has automatically labeled the input in the following way.

NNP/ Carla VBZ/ has VBN/ been VBN/ seen VBG/ dating IN/ with NNP/ Ben

Key:

- # - Pound sign
- $ - Dollar sign
- '' - Close double quote
- `` - Open double quote
- ' - Close single quote
- ` - Open single quote
- , - Comma
- . - Final punctuation
- : - Colon, semi-colon
- -LRB- - Left bracket
- -RRB- - Right bracket
- CC - Coordinating conjunction
- CD - Cardinal number
- DT - Determiner
- EX - Existential there
- FW - Foreign word
- IN - Preposition
- JJ - Adjective
- JJR - Comparative adjective
- JJS - Superlative adjective
- LS - List Item Marker
- MD - Modal
- NN - Singular noun
- NNS - Plural noun
- NNP - Proper singular noun
- NNPS - Proper plural noun
- PDT - Predeterminer
- POS - Possesive ending
- PRP - Personal pronoun
- PP$ - Possesive pronoun
- RB - Adverb
- RBR - Comparative adverb
- RBS - Superlative Adverb
- RP - Particle
- SYM - Symbol
- TO - to
- UH - Interjection
- VB - Verb, base form
- VBD - Verb, past tense
- VBG - Verb, gerund/present participle
- VBN - Verb, past participle
- VBP - Verb, non 3rd ps. sing. present
- VBZ - Verb, 3rd ps. sing. present
- WDT - wh-determiner
- WP - wh-pronoun
- WP$ - Possesive wh-pronoun
- WRB - wh-adverb

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

http://cogcomp.cs.illinois.edu/

Stanford Parser - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Stanford Parser

http://nlp.stanford.edu:8080/parser/index.jsp

Your query

Carla has been seen dating with Ben.

Tagging

```
Carla/NNP  has/VBZ  been/VBN  seen/VBN  dating/VBG  with/IN  Ben/NNP  ./.
```

Parse

```
(ROOT
  (S
    (NP (NNP Carla))
    (VP (VBZ has)
      (VP (VBN been)
        (VP (VBN seen)
          (S
            (VP (VBG dating)
              (PP (IN with)
                (NP (NNP Ben))))))))
    (. .)))
```

Typed dependencies

```
nsubjpass(seen-4, Carla-1)
aux(seen-4, has-2)
auxpass(seen-4, been-3)
xcomp(seen-4, dating-5)
prep(dating-5, with-6)
pobj(with-6, Ben-7)
```

http://nlp.stanford.edu:8080/parser/

Open-Domain Gathering of Facts

[Etzioni,Cafarella et al: WWW 04, IJCAI 07; Weld,Hoffman,Wu: SIGMOD-Rec 08]

Analyze **verbal phrases** between entities for **new relation types**

- unsupervised bootstrapping with short dependency paths

  Carla has been seen dating with Ben.

  Rumors about Carla indicate there is something between her and Ben.

- self-supervised **classifier** for (noun, verb-phrase, noun) triples

  … seen dating with … *(Carla, Ben), (Carla, Sofie), …*

  … partying with … *(Carla, Ben), (Paris, Heidi), …*

- build **statistics** & **prune** sparse candidates

- **group**/**cluster** candidat

  {datesWith, partiesWit

…

But: Result often is noisy
Clusters are not canonicalized relations
High recall at (relatively) low precision

Declarative Extraction Frameworks

- IBM's SystemT [Krishnamurthy et al: SIGMOD Rec. 08, ICDE 08]
  - Declarative IE in a database engine
  - SQL-style operators, cost models, full optimizer support
- DBLife/Cimple [DeRose, Doan et al: CIDR 07, VLDB 07]
  - Online community portal centered around the DB domain (regular crawls of DBLP, conferences, homepages, etc.)

DBLife: Jennifer Widom - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://dblife.cs.wisc.edu/person/Jennifer_Widom&query=Jennifer+Widom

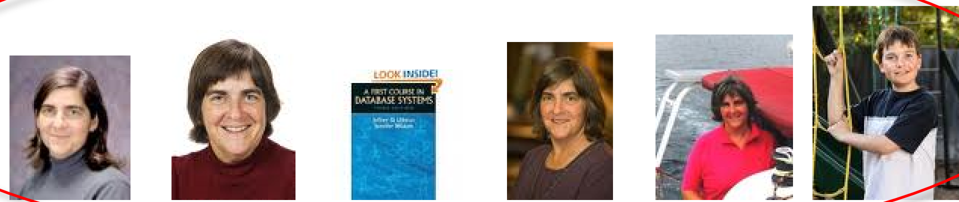
DBLife: Jennifer Widom

DBLife Jennifer Widom [Search]

Help | The Cimple Project

Jennifer Widom

Bing Citeseer DBLP Google Google Scholar Kosmix Wikipedia Yahoo!



from Google Images

more

Google Images

Recent News

- Active Database Systems cited 1 time - details
- ULDBs: Databases with Uncertainty and Lineage cited 20 times - details
- Change Detection in Hierarchically Structured Information cited 1 time - details
- Lore: A Database Management System for Semistructured Data cited 10 times - details
- The TSIMMIS Approach to Mediation: Data Models and Languages cited 8 times - details
- Memory-Limited Execution of Windowed Stream Joins cited 5 times - details
- A First Course in Database Systems cited 5 times - details
- Research Problems in Data Warehousing cited 7 times - details
- Continuous Queries over Data Streams cited 1 time - details

News Archive

DBWorld/DBLP /Google Scholar

DBLP

Sorted by Year/Conf, Year/Citation, Citation Community Statistics

	2010
161	Synthesizing view definitions from data. Anish Das Sarma, Aditya G. Parameswaran, Hector Garcia-Molina, Jennifer Widom. ICDT 2010, 89-103. Web Search BibTeX Download
160	Foundations of Uncertain-Data Integration. Parag Agrawal, Anish Das Sarma, Jeffrey D. Ullman, Jennifer Widom. PVLDB (3): 1080-1090 (2010). Web Search BibTeX Download
159	LIVE: A Lineage-Supported Versioned DBMS. Anish Das Sarma, Martin Theobald, Jennifer Widom. SSDBM 2010, 416-433. Web Search BibTeX Download
	2009
158	Schema Design for Uncertain Databases. Anish Das Sarma, Jeffrey D. Ullman, Jennifer Widom. AMW 2009. Cited by 2 Web Search BibTeX Download
157	Confidence-Aware Join Algorithms. Parag Agrawal, Jennifer Widom. ICDE 2009, 628-639. Web Search BibTeX Download
156	Representing uncertain data: models, properties, and algorithms. Anish Das Sarma, Omar Benjelloun, Alon Y. Halevy, Shubha U. Nabar, Jennifer Widom. VLDB J. (18): 989-1019 (2009). Cited by 1 Web Search BibTeX Download
155	Swoosh: a generic approach to entity resolution. Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, Jennifer Widom. VLDB J. (18): 255-276 (2009). Cited by 67 Web

Professor
http://www-db.stanford.edu/people/widom.html
Stanford University
USA
Papers cited 22,613 times
H-Index of 69

Related People
- Shivnath Babu
- Rajeev Motwani
- Utkarsh Srivastava
- Richard T. Snodgrass

more

Related Topics
- location
- data streams
- streams
- continuous queries

more

Services
- SIGMOD 2010 (SIGMOD New Initiatives Committee) [1]
- CIDR 2007 (PC) [2]
- CIDR 2007 ((Chair) [3]
- CIDR 2005 (Organization Committee) [4]

more

Related Organizations
- Stanford University
- Purdue University
- Microsoft
- INRIA

more

Talks
- Stanford University [5]
- Stony Brook University [6]
- University of Arizona [7]

Tutorials
- SIGMOD 2005 [8]

Homepages/ DBLP/ DBWorld/ Google Scholar

Pattern-Based Harvesting Summary

Facts & Fact Candidates

(Hillary, Bill)
(Carla, Nicolas)

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

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

Patterns

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,** but often noisy/drifting
- **not robust** enough for high precision

Outline for Part III

- Domain-oriented IE vs. Open-domain IE ✔
  - What to extract: entities, classes, binary & higher-arity relations
- Entities, Classes & Subsumptions ✔
  - WordNet concepts, Wikipedia categories, entity disambiguation
- Pattern-based Knowledge Harvesting ✔
  - Wrapper induction, WebTables, statistical pattern mining
- Probabilistic Extraction Models
  - HMMs, MEMMs, CRFs
- Constraints & Reasoning:
  - MLNs, CCMs, FactorIE, SOFIE/PROSPERA
- Open-domain IE
  - ReadTheWeb, TextRunner, Omnivore, REVERB
- Advanced reasoning
  - Temporal/spatial annotations of facts

Applications for Sequence Labeling

- Part-of-Speech (**POS**) tagging
- Named Entity Recognition (**NER**)
- Various specialized labeling tasks
  - e.g. **Blogs**, **emails**, **tweets**
  - lists and fields with regular structures:

    news articles, **citations**, **HTML tables**, etc.

Probabilistic Extraction Models

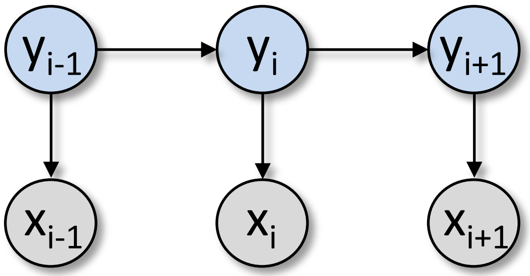
- Hidden Markov Models (HMMs)
  [Rabiner: IEEE 89; Sutton,McCallum: MIT Press 06]
  - Markov Chain (directed)
  - Generatively trained based on **P(X,Y)**
  - Maximum (Log-)likelihood principle for training

- Maximum Entropy Markov Models (MEMMs)
  [McCallum,Freitag,Pereira: ICML 00]
  - Markov Chain (directed)
  - Discriminatively trained using **P(Y|X)**
  - Maximum Entropy principle for training

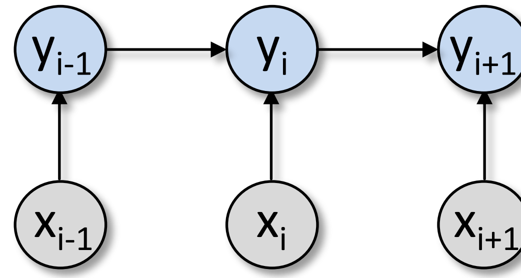
- Conditional Random Fields (CRFs)
  [Lafferty,McCallum,Pereira: ML 01; Sarawagi,Cohen: NIPS 04]
  - Markov Random Field (undirected)
  - Discriminatively trained using **P(Y|X)**
  - Maximum (Log-)likelihood principle for training

X: Observations (tokens)
Y: Labels (POS, NE, etc.)

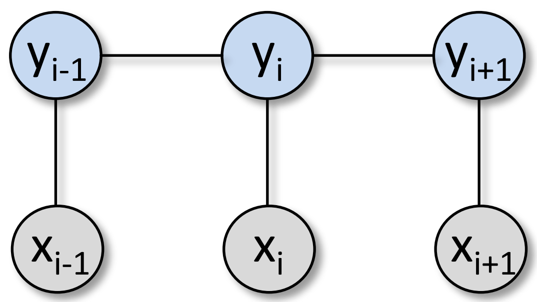
Probabilistic Models for Sequence Labeling



Hidden Markov Model (HMM)



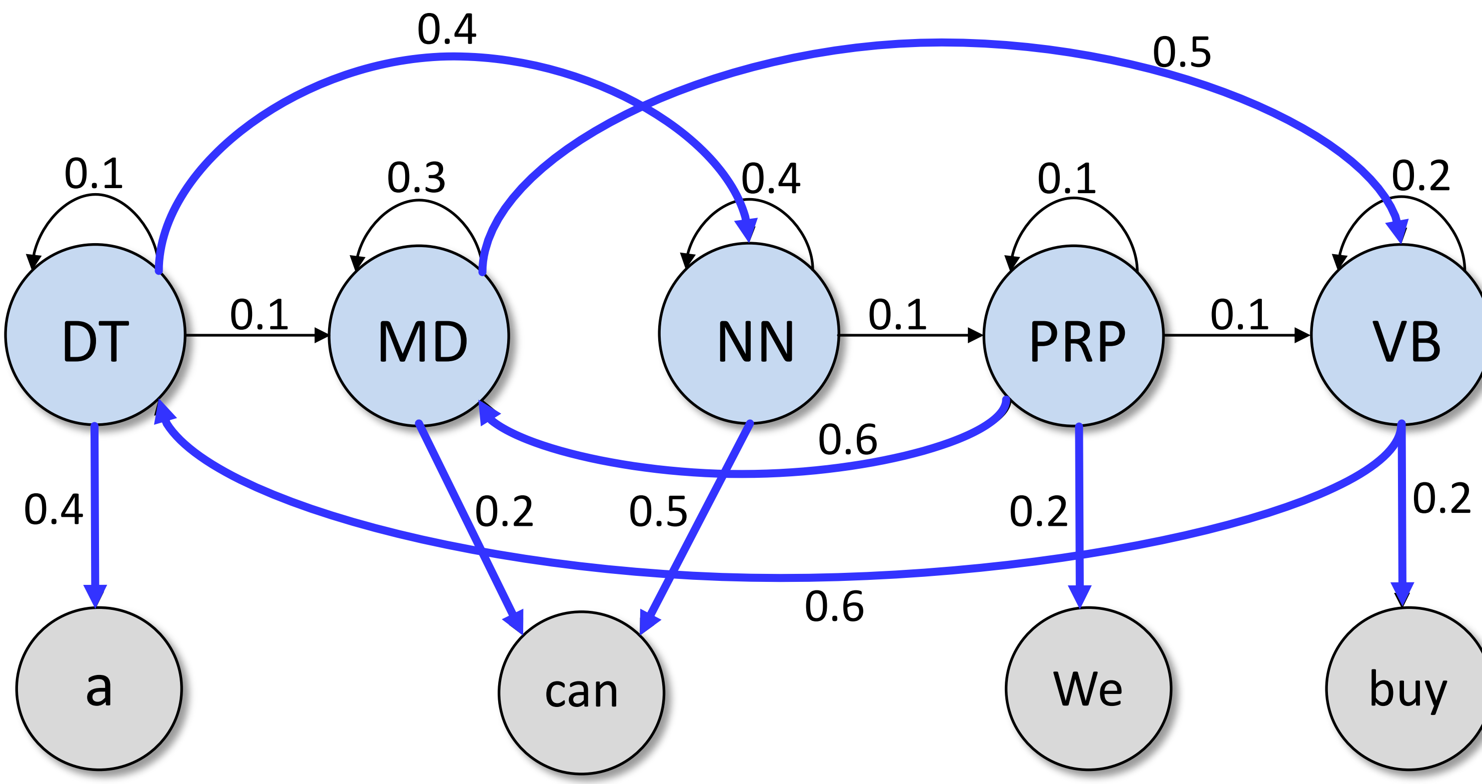
Maximum Entropy Markov Model (MEMM)



Conditional Random Field (CRF)

X: Observations (tokens)
Y: Labels (POS, NE, etc.)

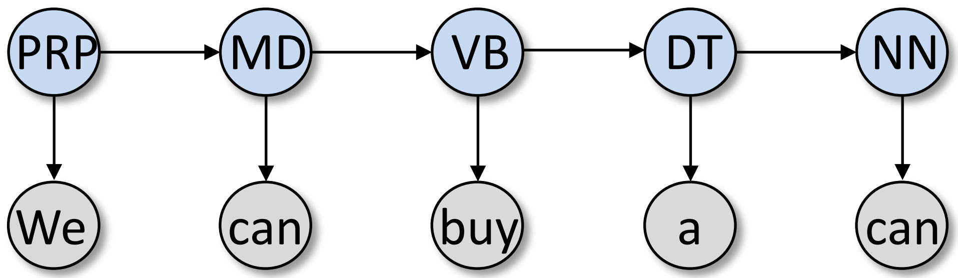
Hidden Markov Models – HMMs



- Part-Of-Speech tagging example with an HMM
  - How to find the best sequence of POS tags for "**We can buy a can**" *efficiently*?

HMMs: Inference & Learning

[Lawrence L. Rabiner, Proc. IEEE 88]



Given: observations X, labels Y, transition probabilities $P(y_i | y_{i-1})$ and $P(x_i | y_i)$

Compute inductively:

$$\alpha_1(i) = \pi_i b_i(x_1)\,, \quad 1 \le i \le N$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(x_{t+1})\,, \quad \begin{array}{l} 1 \le t \le T-1, \\ 1 \le j \le N \end{array}$$

$$P(X|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

(**initial state probabilities π_i , transition probabilities a_{ij}, observation probabilities $b_i(x_t)$** usually estimated from a large annotated training corpus)

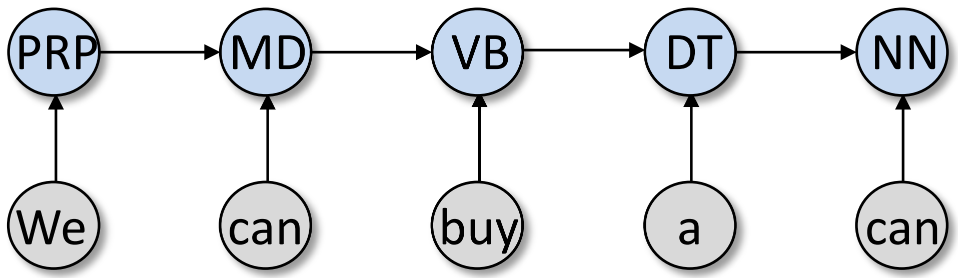
Probability of an observation:
Forward/backward algorithm

Most likely sequence:
Trellis diagram/Viterbi

Weight learning:
Baum-Welch/
Expectation Maximization (EM)

Maximum Entropy Markov Models – MEMMs

[McCallum,Freitag,Pereira: ICML 00]



Given: Observations X, labels Y, transition probabilities $P_{y'}(y|x)$

$$P_{y'}(y|x) = \frac{1}{\boxed{Z(x,y')}} \exp\left(\sum_i \boxed{\lambda_i} \boxed{f_i(x,y)}\right)$$

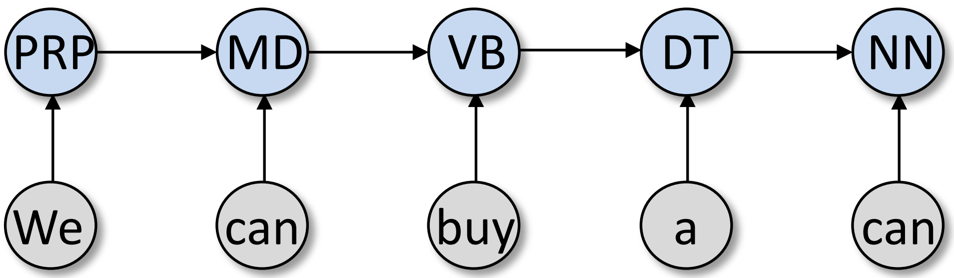
feature functions

parameters

normalizing constant

Maximum Entropy Markov Models – MEMMs

[McCallum,Freitag,Pereira: ICML 00]



Given: Observations X, labels Y, transition probabilities $P_{y'}(y|x)$

$$P_{y'}(y|x) = \frac{1}{Z(x, y')} \exp\left(\sum_i \lambda_i f_i(x, y)\right)$$

Feature functions f_i

$$f_1(x, y) = \begin{cases} 1 \text{ if } \mathit{suffix}(x) = \text{"ing" and } y = \text{VB} \\ 0 \text{ otherwise} \end{cases}$$

$$f_2(x, y) = \begin{cases} 1 \text{ if } \mathit{UpperCase}(x) \text{ and } y = \text{NN} \\ 0 \text{ otherwise} \end{cases}$$

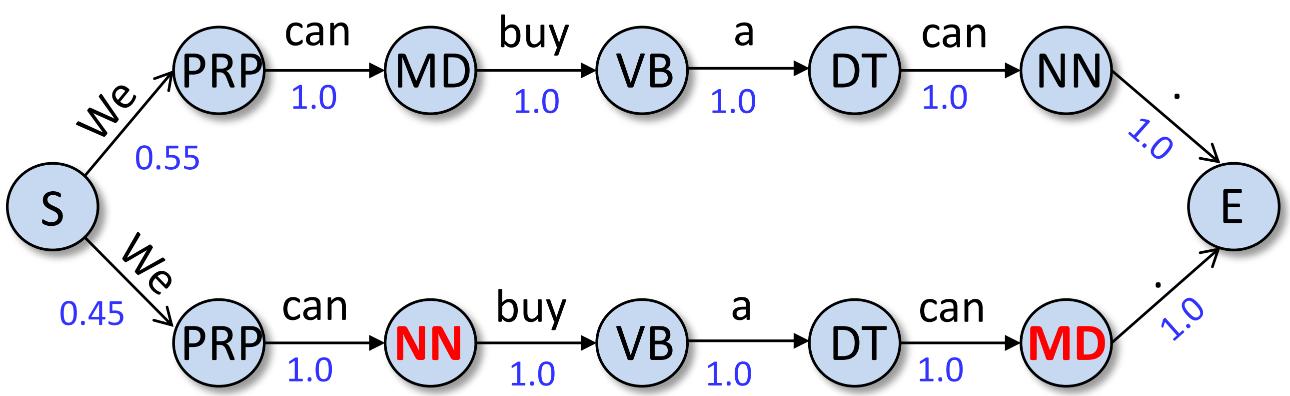
...

Training and **inference** similar to regular HMMs

begins-with-number	contains-question-mark
begins-with-ordinal	contains-question-word
begins-with-punctuation	ends-with-question-mark
begins-with-question-word	first-alpha-is-capitalized
begins-with-subject	indented
blank	indented-1-to-4
contains-alphanum	indented-5-to-10
contains-bracketed-number	more-than-one-third-space
contains-http	only-punctuation
contains-non-space	prev-is-blank
contains-number	prev-begins-with-ordinal
contains-pipe	shorter-than-30

Features used for Blog post segmentation

Directed Models and Label Bias

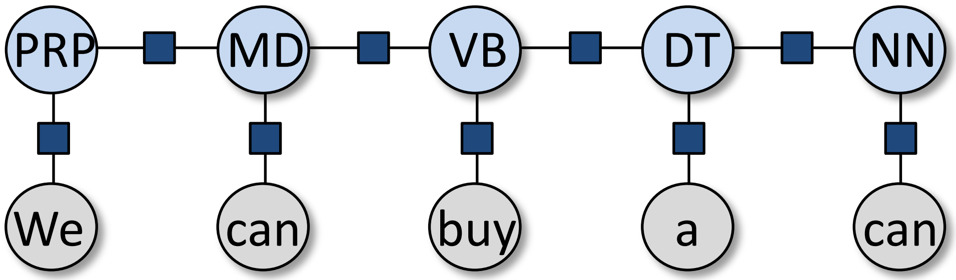


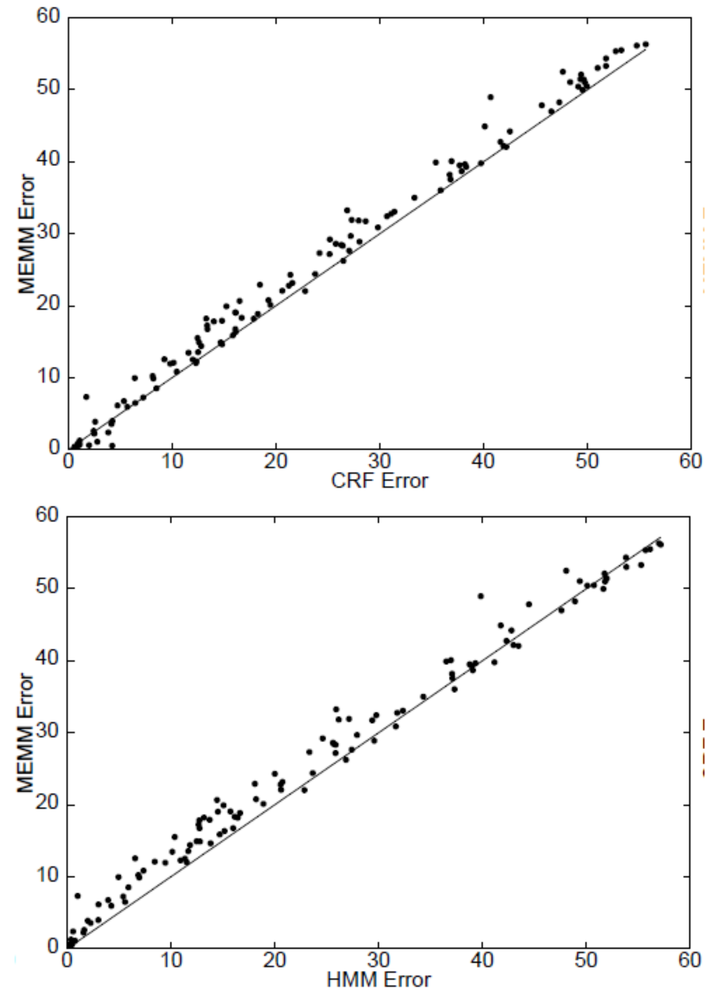
- Top path and bottom paths are almost equally likely
- Difference only at initial transition

→ States with low-entropy transitions (in the extreme case: a single transition) effectively ignore their observations

Conditional Random Fields – CRFs

[Lafferty,McCallum,Pereira: ML 01]





Given:

Factor graph G=(V,E) with

- Random variables V = (X U Y)
- Factors E U V (potential functions)

$$P_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\exp\left(\sum_{e\in E,i}\lambda_i f_i(e,\mathbf{x},\mathbf{y}|_e) \; + \; \sum_{v\in V,i}\lambda_i g_i(v,\mathbf{x},\mathbf{y}|_v)\right)$$

- Exact inference, e.g., via **forward/backward**, **Viterbi**, or **variable elimination**
- Various EM techniques for training

- Plots of 2x2 error rates for synthetic data runs

CRF Extensions

- Semi-Markov CRFs
  [Sarawagi,Cohen: NIPS 04]
  – Identify entire **subsequences** with the same label
  – Additional cost for inference remains linear in the maximum label length

- Joint Training over "Fusion Graphs"
  [Gupta,Sarawagi: CoRR 10, WSDM 11]
  – Merge overlapping sequences from **multiple sources** into a single graph structure
  – Train individual CRFs using features from the merged graph
  – Can learn CRFs from **very few training examples** (~4)

Outline for Part III

- Domain-oriented IE vs. Open-domain IE ✔
  - What to extract: entities, classes, binary & higher-arity relations
- Entities, Classes & Subsumptions ✔
  - WordNet concepts, Wikipedia categories, entity disambiguation
- Pattern-based Knowledge Harvesting ✔
  - Wrapper induction, WebTables, statistical pattern mining
- Probabilistic Extraction Models ✔
  - HMMs, MEMMs, CRFs
- Constraints & Reasoning
  - MLNs, CCMs, FactorIE, SOFIE/PROSPERA
- Open-domain IE
  - ReadTheWeb, TextRunner, Omnivore, REVERB
- Advanced reasoning
  - Temporal/spatial annotations of facts

More Ontological Rigor!

- Reasoning for **pattern**/**fact consistency** using first-order logical constraints
  - Markov Logic Networks
  - Constrained Conditional Models
  - FactorIE
  - SOFIE/PROSPERA

→ **Canonical** entities & **typed** relations

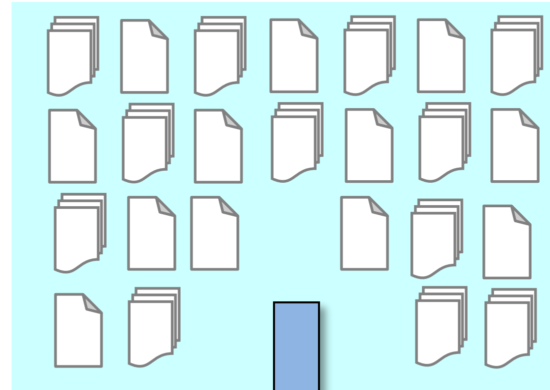
French Marriage Problem

marriedTo: person × person

$\forall x,y,z$: marriedTo(x,y) $\wedge$ marriedTo(x,z) $\Rightarrow$ y=z

~~**marriedTo_French: person × person**~~

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 constraints/reasoning

Reasoning about Fact Candidates

Use **consistency constraints** to prune false candidates

First-order-logic rules (restricted):

spouse(x,y) ∧ diff(y,z) ⇒ ¬spouse(x,z)

spouse(x,y) ∧ diff(w,y) ⇒ ¬spouse(w,y)

spouse(x,y) ⇒ f(x)

spouse(x,y) ⇒ m(y)

Grounded 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, or by learning)
→ uncertain / probabilistic data
→ compute marginal probabilities of grounded atoms being “true“

Markov Logic Networks

[Richardson, Domingos: ML 2006]

Map logical constraints & fact candidates
into **probabilistic graphical model**: Markov Random Field (MRF)

FOL rules w/ weights:

$s(x,y) \wedge diff(y,z) \Rightarrow \neg s(x,z)_{5.9}$ $\quad s(x,y) \Rightarrow f(x)_{100}$ $\quad f(x) \Rightarrow \neg m(x)_{100}$

$s(x,y) \wedge diff(w,y) \Rightarrow \neg s(w,y)_{7.5}$ $\quad s(x,y) \Rightarrow m(y)_{100}$ $\quad m(x) \Rightarrow \neg f(x)_{100}$

Facts/entities:

s(Carla,Nicolas)
s(Cecilia,Nicolas)
s(Carla,Ben)
s(Carla,Sofie)
...

Grounding: Literal → Boolean Var
Reasoning: Literal → Binary RV

Grounding:

$\neg$s(Ca,Nic) $\vee$ $\neg$s(Ce,Nic)

$\neg$s(Ca,Nic) $\vee$ $\neg$s(Ca,Ben)

$\neg$s(Ca,Nic) $\vee$ $\neg$s(Ca,So)

$\neg$s(Ca,Ben) $\vee$ $\neg$s(Ca,So)

$\neg$s(Ca,Ben) $\vee$ $\neg$s(Ca,So)

$\neg$s(Ca,Nic) $\vee$ m(Nic)

$\neg$s(Ce,Nic) $\vee$ m(Nic)

$\neg$s(Ca,Ben) $\vee$ m(Ben)

$\neg$s(Ca,So) $\vee$ m(So)

Markov Logic Networks

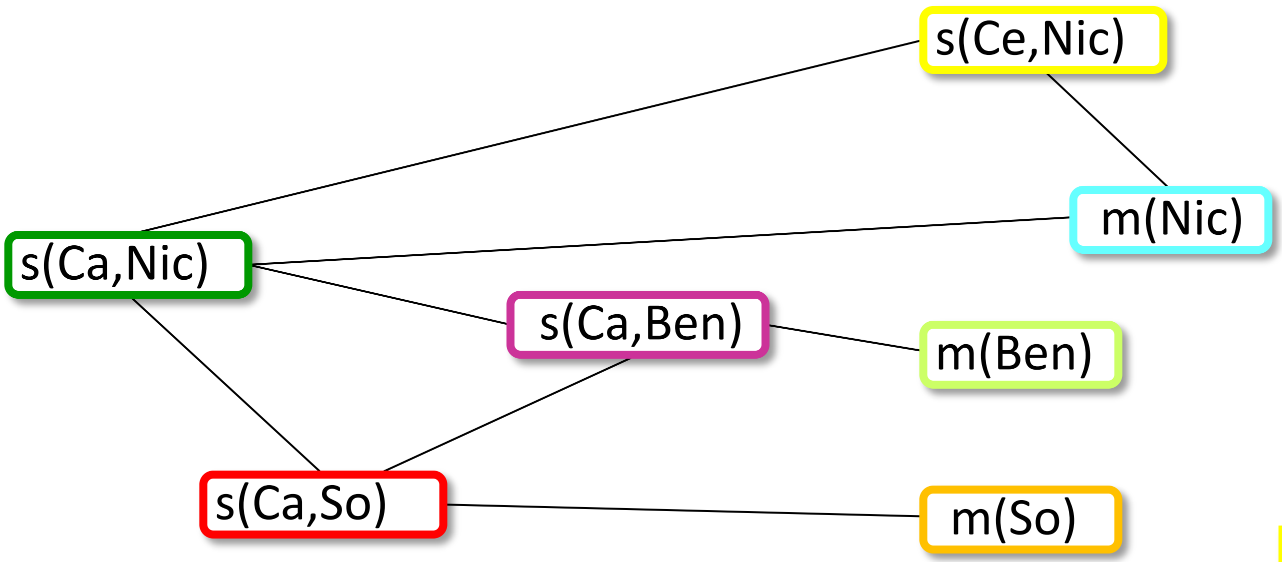
[Richardson, Domingos: ML 2006]

Map logical constraints & fact candidates
into **probabilistic graphical model**: Markov Random Field (MRF)

$s(x,y) \wedge diff(y,z) \Rightarrow \neg s(x,z)_{5.9}$ $\quad s(x,y) \Rightarrow f(x)_{100}$ $\quad f(x) \Rightarrow \neg m(x)_{100}$

$s(x,y) \wedge diff(w,y) \Rightarrow \neg s(w,y)_{7.5}$ $\quad s(x,y) \Rightarrow m(y)_{100}$ $\quad m(x) \Rightarrow \neg f(x)_{100}$

s(Carla,Nicolas)
s(Cecilia,Nicolas)
s(Carla,Ben)
s(Carla,Sofie)
…



RVs coupled by MRF edge if they appear in same clause

MRF assumption:
$P[X_i | X_1..X_n] = P[X_i | MB(X_i)]$

joint distribution
has product form
over all cliques

Markov Logic Networks

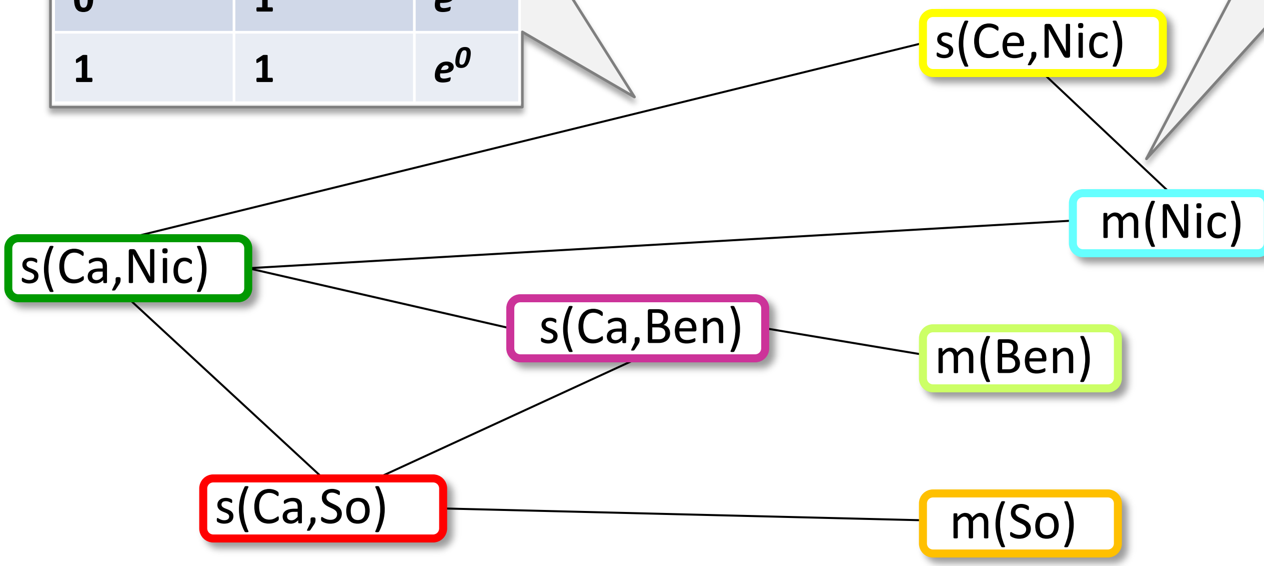
[Richardson, Domingos: ML 2006]

$\neg$ s(Ca,Nic) $\vee$ $\neg$s(Ce,Nic)

s(Ca,Nic)	s(Ce,Nic)	ϕ_i
0	0	$e^{5.9}$
1	0	$e^{5.9}$
0	1	$e^{5.9}$
1	1	e^{0}

$\neg$ s(Ce,Nic) $\vee$ m(Nic)

s(Ce,Nic)	m(Nic)	ϕ_i
0	0	e^{100}
1	0	e^{0}
0	1	e^{100}
1	1	e^{100}



RVs coupled by MRF edge if they appear in same clause

MRF assumption:
$P[X_i | X_1..X_n]=P[X_i | MB(X_i)]$
joint distribution has product form over all cliques

$$P(X=x) = \frac{1}{Z}\exp\left(\sum_i w_i n_i(x)\right) = \frac{1}{Z}\prod_i \phi_i(x_{\{i\}})^{n_i(x)}$$

Markov Logic Networks

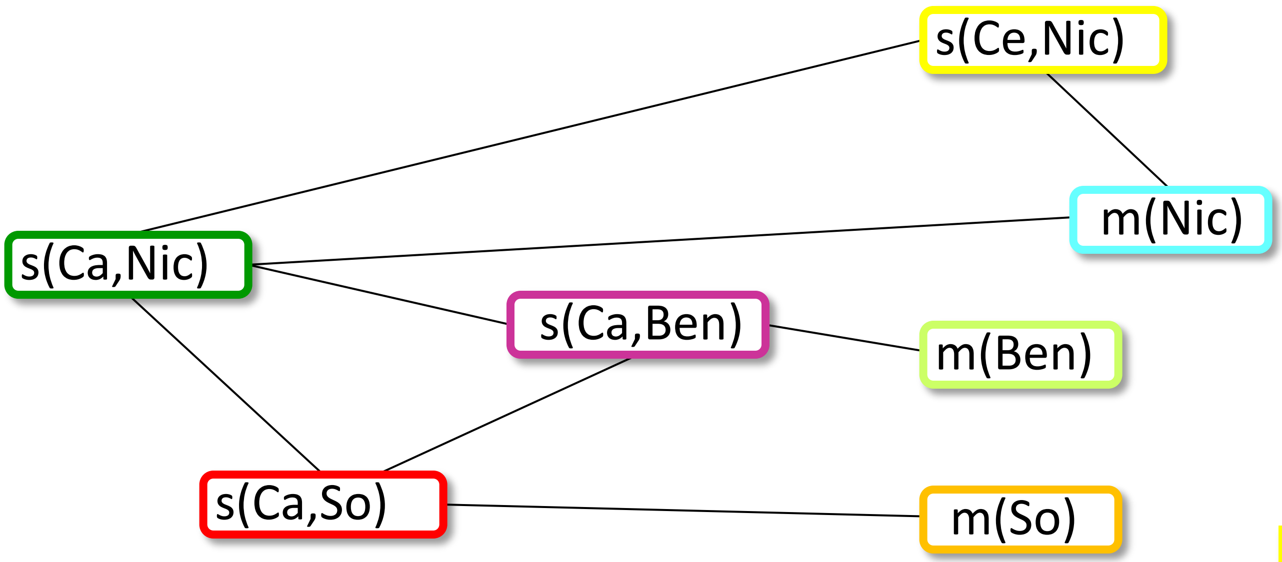
[Richardson, Domingos: ML 2006]

Map logical constraints & fact candidates
into **probabilistic graphical model**: Markov Random Field (MRF)

$\mathbf{s(x,y) \wedge diff(y,z) \Rightarrow \neg s(x,z)}_{5.9}$ $\mathbf{s(x,y) \Rightarrow f(x)}_{100}$ $\mathbf{f(x) \Rightarrow \neg m(x)}_{100}$

$\mathbf{s(x,y) \wedge diff(w,y) \Rightarrow \neg s(w,y)}_{7.5}$ $\mathbf{s(x,y) \Rightarrow m(y)}_{100}$ $\mathbf{m(x) \Rightarrow \neg f(x)}_{100}$

s(Carla,Nicolas)
s(Cecilia,Nicolas)
s(Carla,Ben)
s(Carla,Sofie)
...



RVs coupled by MRF edge if they appear in same clause

MRF assumption:
$P[X_i | X_1..X_n] = P[X_i | MB(X_i)]$

joint distribution has product form over all cliques

Variety of algorithms for joint inference:
MCMC (Gibbs sampling, MC-SAT), belief propagation, stochastic MaxSat, …

Markov Logic Networks

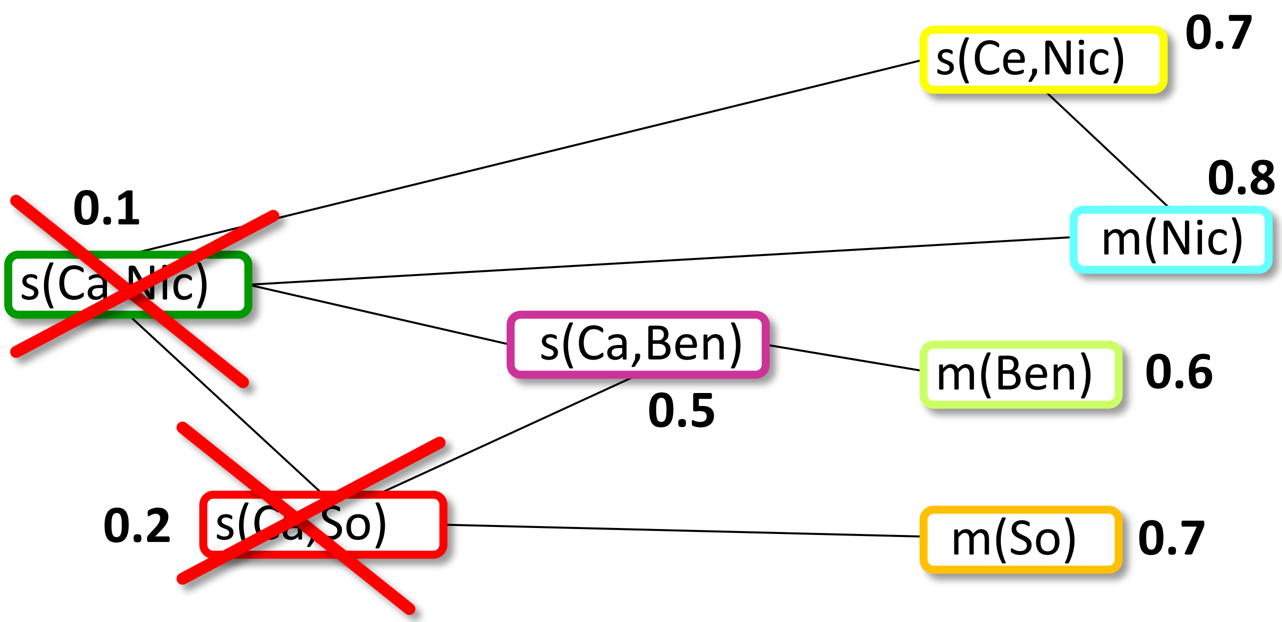
[Richardson, Domingos: ML 2006]

Map logical constraints & fact candidates
into **probabilistic graphical model**: Markov Random Field (MRF)

$\mathbf{s(x,y) \wedge diff(y,z) \Rightarrow \neg s(x,z)_{5.9}}$ $\mathbf{s(x,y) \Rightarrow f(x)_{100}}$ $\mathbf{f(x) \Rightarrow \neg m(x)_{100}}$

$\mathbf{s(x,y) \wedge diff(w,y) \Rightarrow \neg s(w,y)_{7.5}}$ $\mathbf{s(x,y) \Rightarrow m(y)_{100}}$ $\mathbf{m(x) \Rightarrow \neg f(x)_{100}}$

s(Carla,Nicolas)
s(Cecilia,Nicolas)
s(Carla,Ben)
s(Carla,Sofie)
…



Consistency reasoning: prune low-confidence facts
StatSnowball [Zhu et al: WWW'09], BioSnowball [Liu et al: KDD'10]
EntityCube, MSR Asia: http://entitycube.research.microsoft.com/

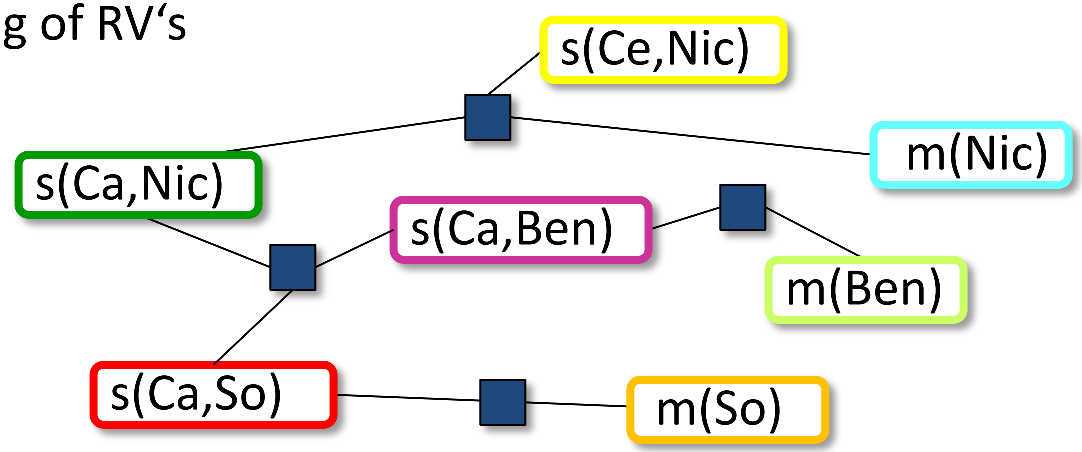
Related Alternative Probabilistic Models

Constrained Conditional Models [Roth et al. 2007]

log-linear classifiers with constraint-violation penalty
mapped into <u>Integer Linear Programs</u>

Factor Graphs with Imperative Variable Coordination [McCallum et al. 2008]

RV's share "factors" (joint feature functions)
generalizes MRF, BN, CRF, ...
inference via advanced MCMC
flexible coupling & constraining of RV's



Software tools:
alchemy.cs.washington.edu
code.google.com/p/factorie/
research.microsoft.com/en-us/um/cambridge/projects/infernet/

FactorIE

[McCallum et al.: NIPS 08, ECML/PKDD 09, VLDB 10]

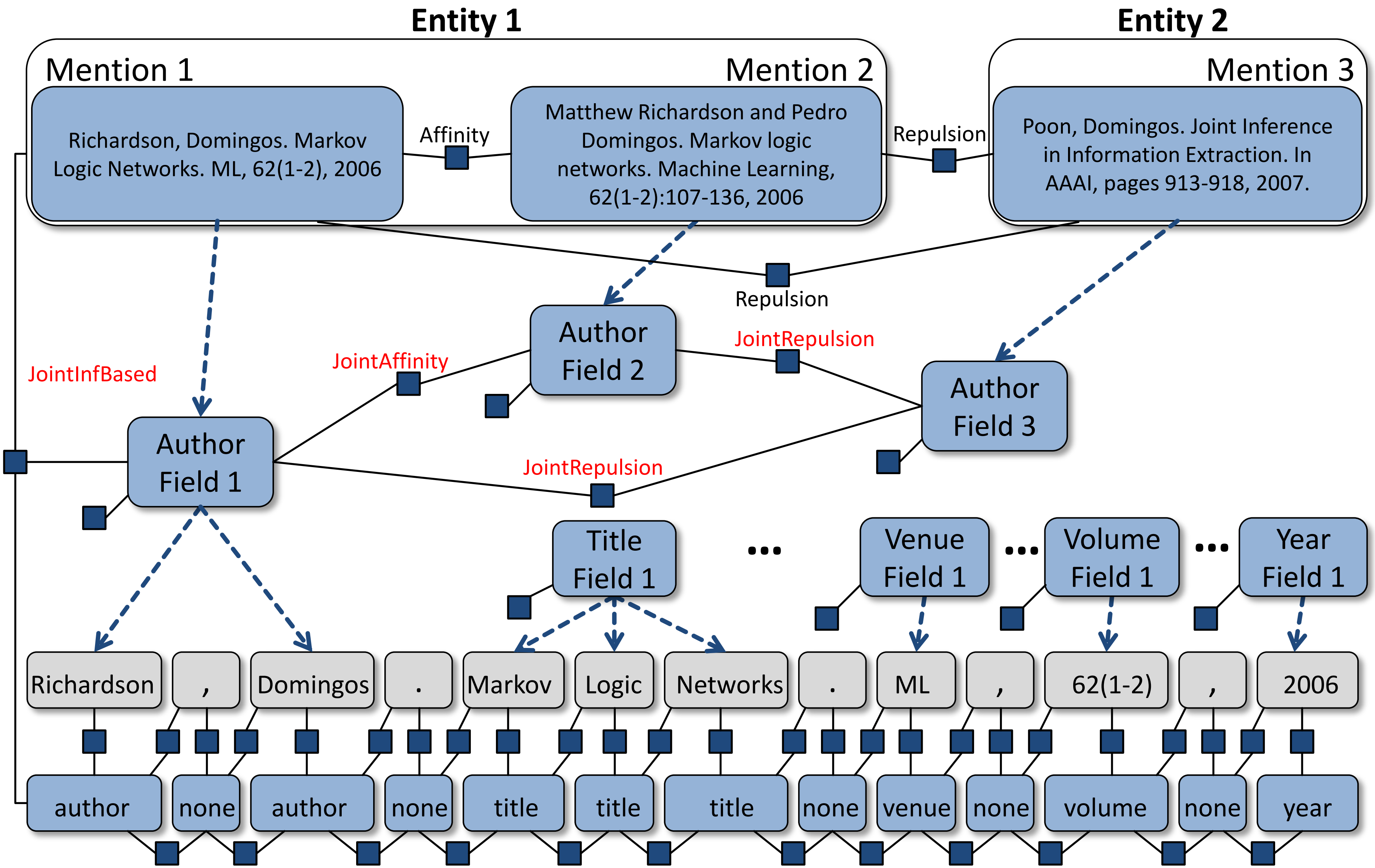
- **Imperatively Defined Factor graphs (IDF)**
  - Object-oriented , imperative programming language (Scala)
- **Open-source toolsuite for *deployable* probabilistic modeling**
  - Markov Logic, CRFs, MCMC, weight learning, etc.
  - Scalable DB backend

http://code.google.com/p/factorie/

```
object LogicDemo1 {
def main(args:Array[String]) : Unit = {

// Define entity, attribute and relation types
class Person (val name:String)
 extends ItemizedObservation[Person] with Entity[Person] {
   object smokes extends BooleanVariable with Attribute
   object cancer extends BooleanVariable with Attribute
 ...}
object Friends extends Relation[Person,Person];

// Define the model
val model = new Model (
Forany[Person] { p => p.cancer } * 0.1,
Forany[Person] { p => p.smokes ==> p.cancer } * 2.0
Forany[Person] { p => p.friends.smokes <==> p->Smokes } * 1.5  )

// Create the data
val amy = new Person("Amy"); amy.smokes := true
val bob = new Person("Bob");
Friends(amy,bob); Friends(bob,amy)

// Do 2000 iterations of sampling, gathering sample counts every 20 iterations
val inferencer = new VariableSamplingInferencer(
   new VariableSettingsSampler[BooleanVariable](model))
inferencer.burnIn = 100; inferencer.iterations = 2000; inferencer.thinning = 20
val marginals = inferencer.infer(List(bob.cancer, bob.smokes)) }}
```

Bidirectional Joint Segmentation & Disambiguation

[Singh,Schultz,McCallum: ECML 09; Poon,Domingos: AAAI 07]



SOFIE: Reasoning for KB Growth

[Suchanek,Sozio,Weikum: WWW 09]

YAGO Facts

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

+

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:

- KB facts are true; fact candidates & patterns $\rightarrow$ hypotheses
- known entities and typed relations
- grounded constraints $\rightarrow$ clauses with hypotheses as vars
- cast into Weighted Max-Sat with weights from pattern stats
- customized approximation algorithm
- unifies: fact/candidate consistency, pattern goodness, entity disambiguation

www.mpi-inf.mpg.de/yago-naga/sofie/

SOFIE: Facts & Patterns Consistency

[Suchanek,Sozio,Weikum: WWW'09]

Constraints to connect **facts**, **fact candidates** & **patterns**

pattern-fact duality:

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

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

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

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

functional dependencies:

$\text{spouse}(x,y)$: $x \rightarrow y$, $y \rightarrow x$

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:

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

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

www.mpi-inf.mpg.de/yago-naga/sofie/

SOFIE: Facts & Patterns Consistency

[Suchanek,Sozio,Weikum: WWW 09]

Constraints to connect facts, fact candidates & patterns

pattern-fact duality:

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

$\text{occurs}(p,x,y) \wedge R(x,y) \Rightarrow$ expresse

- **Grounded** into large propositional Boolean formula in CNF
- **Max-Sat solver** for joint inference (complete truth assignment to all candidate patterns & facts)

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

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

functional dependencies:

$\text{spouse}(x,y)$: $x \rightarrow y$, $y \rightarrow x$

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:

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

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

www.mpi-inf.mpg.de/yago-naga/sofie/

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) [-1]
spouse (Rebecca, David) [-1]
spouse (Victoria, Tom) [-1]

expresses (*X and her husband Y*, Spouse) [-1]
expresses (*X Y and their children*, Spouse) [-1]
expresses (*X dating with Y*, Spouse) [-1]

Clauses

$\forall x,y,z,w$: R(x,y) $\wedge$ R(x,z) $\Rightarrow$ y=z
spouse (Victoria, David) $\wedge$ spouse (Rebecca, David)

$\forall x,y,z,w$: R(x,y) $\wedge$ R(w,y) $\Rightarrow$ x=w
spouse (Victoria, David) $\wedge$ spouse (Victoria, Tom)

...

$\forall x,y$: R(x,y) $\Rightarrow$ R(y,x)
occurs (husband, Victoria, David) $\wedge$ expresses (husband, Spouse) [60]
... $\Rightarrow$ Spouse (Victoria, David)

$\forall p,x,y$: occurs (p, x, y) $\wedge$ expresses (p, R) $\Rightarrow$ R (x, y)
occurs (dating, Rebecca, David) $\wedge$ expresses (dating, Spouse) [20]
$\Rightarrow$ Spouse (Rebecca, David)

...

$\forall p,x,y$: occurs (p, x, y) $\wedge$ R(x,y) $\Rightarrow$ expresses (p, R)
occurs (husband, Victoria, David) $\wedge$ Spouse (Victoria, David) [60]
$\Rightarrow$ expresses (husband, Spouse)

...

Soft Rules vs. Hard Constraints

Enforce FD's (mutual exclusion) as hard constraints:

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

Combine with weighted constraints
No longer regular MaxSat
Constrained & weighted MaxSat

Generalize to other forms of constraints:

Hard constraint

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

Soft constraint

firstPaper(x,p) $\wedge$ firstPaper(y,q) $\wedge$
author(p,x) $\wedge$ author(p,y)) $\wedge$
inYear(p) > inYear(q) + 5years
$\Rightarrow$ hasAdvisor(x,y) [0.6]

Open issue for arbitrary constraints
(e.g., Datalog-style deductive grounding
vs. „open-world" Markov Logic)
$\rightarrow$ **Rethink reasoning!**

Pattern Harvesting, Revisited

[Suchanek et al: KDD 06; Nakashole et al: WebDB 10, WSDM 11]

using narrow & dropping nasty patterns loses recall !

narrow / nasty / noisy patterns:

X and his famous advisor Y

X jointly developed the method with Y

X carried out his doctoral research in math under the supervision of Y

using noisy patterns loses precision & slows down reasoner !

POS-lifted n-gram itemsets as patterns:

X { PRP ADJ advisor} Y

X { ADJ developed method} Y

X { carried out PRP doctoral research [IN NP] [DET] supervision [IN] } Y

confidence & support weights, using seeds and counter-seeds:

seeds: (MosheVardi, CatrielBeeri), (JimGray, MikeHarrison)

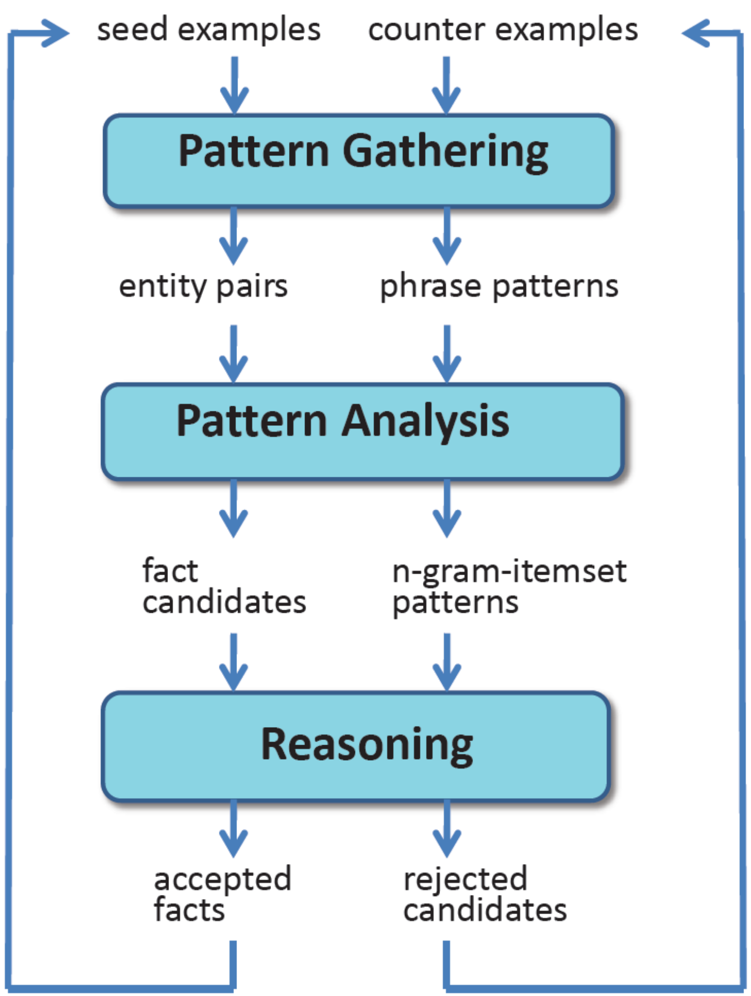
counter-seeds: (MosheVardi, RonFagin), (AlonHalevy, LarryPage)

→ confidence of pattern p ~ (#p with seeds – #p with counter-seeds)

→ support of pattern p ~ frequency of p

PROSPERA Architecture

[Nakashole,Theobald,Weikum: WebDB 10; WSDM 11]



- **Gathering: Enhanced Hearst patterns**
  - POS-enriched n-grams
  - Pattern-fact duality
  - Disambiguation of entities based on “means” and “type” in YAGO
- **Analysis: Refined pattern weights**
  - Carefully chosen **seeds** and **counter seeds** (closed set of target relations)
  - Thresholds for **confidence** & **support**
- **Reasoning: Scalable (distributed) extraction & consistency reasoning**
  - **MapReduce** functions for extraction & gathering of statistics
  - SOFIE-based, **distributed MaxSat solver** + graph partitioning
  - Experiments on large Web corpus w/**500 Mio** documents

Trivially Parallel: Pattern Mining

FUNCTION map(i, P_i)
    List $N \leftarrow$ generateNgrams(P_i)
    FOR $n_i \in N$ **DO**
        emit(n_i,1)

FUNCTION reduce(n_i, $[v1, v2, v3, ...]$)
    $support \leftarrow 0$
    FOR $v_i \in [v1, v2, v3, ...]$ **DO**
        $support \leftarrow support + v_i$
    IF $support \geq MINSUPPORT$
        emit(n_i, $support$)

- Frequent n-gram patterns

FUNCTION map(i, $[e_1, p, e_2]$)
    IF isSeedPattern(p)
        FOR $r \in R$ **DO**
            SeedOccurrence $O \leftarrow [r, e_1, e_2]$
            emit($p.id$, O)

FUNCTION reduce($p.id$, $[O1, O2, O3, ...]$)
    List $L \leftarrow \{\ \}$
    FOR $O \in [O1, O2, O3, ...]$ **DO**
        L.append(O)
    emit(p.id, L)

- Seed pattern occurrences & confidences

Harder to Parallelize: Consistency Reasoning

- Distributed reasoning is non-trivial!
  - Constraints impose dependencies
  - Facts and pattern candidates are **vertices**
  - Literals in a grounded constraint form **cliques**
- Min-cut two-phase algorithm
  - Randomized approximation

    [Karypis et al. 98, Karger 96]

    1: **coarsen graph**

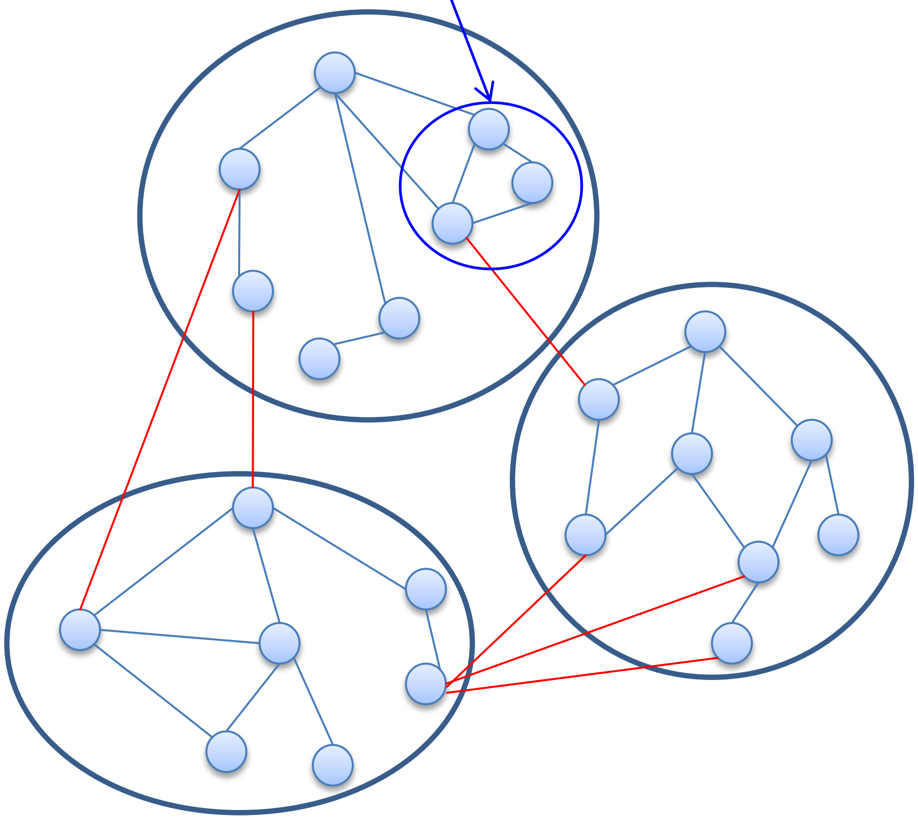
    2: **partition the coarser graph**

    → minimize the weight of the cut edges

    → keep partitions balanced

occurs(“husband”, Victoria, David) ∧ Spouse(Victoria, David)

⇒ expresses(“husband”, Spouse)

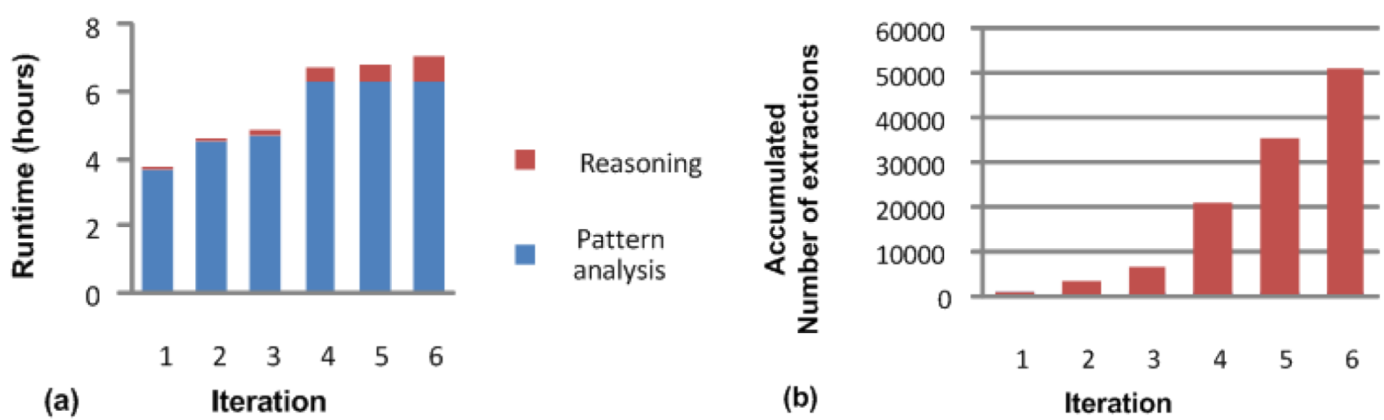


PROSPERA Results

[Nakashole et al : WSDM 11]

Relation	# Extractions			Precision			Precision@1000
	PROSPERA-6	NELL-6	NELL-66	PROSPERA-6	NELL-6	NELL-66	PROSPERA-6
AthletePlaysForTeam	14,685	29	456	82%	100%	100%	100%
CoachCoachesTeam	1,013	57	329	88%	100%	100%	n/a
TeamPlaysAgainstTeam	15,170	83	1,068	89%	96%	99%	100%
TeamWonTrophy	98	29	397	94%	88%	68%	n/a
AthletePlaysInLeague	3,920	2	641	94%	n/a	n/a	n/a
TeamPlaysInLeague	1,920	62	288	89%	n/a	n/a	n/a
AthleteWonTrophy	10	n/a	n/a	90%	n/a	n/a	n/a
CoachCoachesInLeague	676	n/a	n/a	99%	n/a	n/a	n/a
TeamMate	19,666	n/a	n/a	86%	n/a	n/a	100%

Table 1: Performance comparison between PROSPERA and NELL on sports relations



ClueWeb-2009 corpus: ~500 Mio English Web documents

Outline for Part III

- Domain-oriented IE vs. Open-domain IE ✔
  - What to extract: entities, classes, binary & higher-arity relations
- Entities, Classes & Subsumptions ✔
  - WordNet concepts, Wikipedia categories, entity disambiguation
- Pattern-based Knowledge Harvesting ✔
  - Wrapper induction, WebTables, statistical pattern mining
- Probabilistic Extraction Models ✔
  - HMMs, MEMMs, CRFs
- Constraints & Reasoning ✔
  - MLNs, CCMs, FactorIE, SOFIE/PROSPERA
- Open-domain IE
  - ReadTheWeb, TextRunner, Omnivore, REVERB
- Advanced reasoning
  - Temporal/spatial annotations of facts

Open-Domain IE, History

- KnowItAll (Web), Kylin/Kog (Wikipedia-centric)

  [Etzioni,Cafarella et al: WWW 04; Wu,Weld et al: CIKM 07, SIGMOD-Rec. 08, WWW 2008]

- TextRunner, Omnivore, REVERB

  [Cafarella,Banko,Etzioni,Soderland et al: AAAI 06, NAACL-HLT 07, SIGMOD Rec. 07, IJCAI 07, CIDR 09, NAACL 10, EMNLP 11]

- ReadTheWeb, NELL

  [Carlson,Cohen,Mitchell et al: NAACL-HLT Ws. 09, AAAI 10, WSDM 10]

Open-domain IE, Methodology

- Information extraction from free text with **limited or no** assumptions about **domain knowledge**
- **NLP** techniques: POS, DP, SRL
- Unsupervised **clustering** or semi-supervised **classifiers**
- **Bootstrapping** loops: pattern/fact duality
- **No** or **limited** (periodic) **human supervision**
- Extract a large number of "**beliefs**" or "**assertions**"

ReadTheWeb

[Carlson,Mitchell et al: NAACL-HLT Ws. 09, WSDM 10, AAAI 10]

- Coupled Semi-Supervised Learning for Information Extraction

- Ontological backbone
  - Closed set of categories & typed relations
  - Seeds/counter seeds
  - Open set of predicate arguments
- Coupled learners
  - Coupled pattern/relation extractor
  - Coupled SEAL
- NELL
  - Constantly running over a large Web corpus since January 2010 (200 Mio pages Web crawl)
  - Periodic human supervision

athletePlaysForTeam (Athlete, SportsTeam)

athletePlaysForTeam (Alex Rodriguez, Yankees)

athletePlaysForTeam (Alexander_Ovechkin, Penguins)

ReadTheWeb

[Carlson,Mitchell et al: NAACL-HLT Ws. 09, WSDM 10, AAAI 10]

- Coupled Semi-Supervised Learning for Information Extraction

- **Coupled output constraints**
  - For $f_1(x_1) \rightarrow y_1$ and $f_2(x_1) \rightarrow y_2$
  - Restrict output y_1 and y_2 (e.g. $f_1(x) \rightarrow f_2(x)$ for functional dependencies, mut.-ex.)
- **Compositional constraints**
  - For $f_1(x_1) \rightarrow y_1$ and $f_2(x_1, x_2) \rightarrow y_2$
  - Restrict y_1, y_2 to valid pairs (special case: type checking)
- **Multi-view agreement**
  - Co-training classifiers $f_1(x_1) \rightarrow y$ and $f_2(x_2) \rightarrow y$
- **Constraints employed for experiments**
  - Mutual-exclusiveness predicates
  - Type checking
  - Label-agreement

Coupled Pattern Learner (CPL)

For i = 1,..,∞ **do**

- **For each** predicate *p* **do**
  - **Extract** new candidates instances/contextual patterns of *p* using recently promoted instances
  - **Filter** candidates that violate constraints
  - **Rank** candidate instances/patterns
  - **Promote** top candidates for next round

Meta-Boostrap Learner (MBL)

For i = 1,..,∞ **do**

- **For each** predicate *p* **do**
  - **For each** extractor *e* **do**

    Extract new candidates for *p* using *e* with recently promoted instances
  - **Filter** candidates that violate mutual-exclusion or type constraints
  - **Promote** candidates that were extracted by all extractors

ReadTheWeb

[Carlson,Mitchell et al: WSDM 10]

	Precision (%)					Promoted Instances (#)				
Predicate	CPL	UPL	CSEAL	SEAL	MBL	CPL	UPL	CSEAL	SEAL	MBL
CompanyAcquiredCompany	97	77	-	-	-	93	230	0	0	0
AthletePlaysForTeam	100	93	100	76	100	9	269	4	17	96
AthletePlaysInLeague	-	78	100	57	-	0	18	14	82	0
AthletePlaysSport	100	47	100	100	100	83	258	1	1	109
CEOOfCompany	100	100	-	100	100	18	18	0	1	1
CityLocatedInCountry	93	57	100	100	100	185	787	9	577	136
CityLocatedInState	100	70	100	93	100	76	194	34	537	54
CoachCoachesInLeague	-	-	0	-	-	0	0	1	0	0
CoachCoachesTeam	100	100	-	-	100	324	668	0	0	6
CompanyIsInEconomicSector	93	97	-	-	-	583	889	0	0	0
CompanyCompetesWithCompany	100	67	-	-	-	28	123	0	0	0
CompanyHasOfficeInCity	-	63	-	100	-	0	526	0	4	0
CompanyHasOfficeInCountry	-	90	-	-	-	0	195	0	0	0
CompanyHeadquarteredInCity	50	53	100	100	-	2	532	1	2	0
LeaguePlaysGamesInStadium	-	-	-	100	-	0	0	0	177	0
CompanyProducesProduct	97	93	-	-	100	54	215	0	0	8
ProductInstanceOfProductType	73	67	-	-	-	153	484	0	0	0
SportUsesSportsEquipment	33	3	100	87	33	15	1330	5	15	6
StadiumLocatedInCity	100	20	77	70	90	7	600	200	554	56
StateHasCapitalCity	60	70	-	73	-	266	188	0	495	0
StateLocatedInCountry	97	40	100	97	100	194	1299	46	653	61
TeamHasHomeStadium	100	87	100	100	100	97	208	179	106	92
TeamPlaysAgainstTeam	100	80	-	-	-	238	2088	0	0	0
TeamHasHomeCity	-	57	-	93	100	0	680	0	29	11
TeamPlaysInLeague	100	67	100	100	100	7	255	104	749	23
TeamPlaysSport	-	70	100	100	100	0	177	30	30	37
TeamWonAwardTrophyTournament	90	70	-	-	-	128	262	0	0	0
Average	89	69	91	91	95	95	463	23	149	26
Weighted Average	91	61	92	90	99					

Beliefs learned over a 200 million pages Web corpus after 10 iterations

NELL: Never-Ending Language Learning

[Carlson,Mitchell et al: AAAI 10]

- Constantly online since January 2010
  - Many hundreds of iterations

- More **Coupled Learners**
  - Coupled Pattern Learner
    e.g., ***mayor of X***, ***X plays for Y***
  - Coupled SEAL
    Set expansion & wrapper induction algorithm
  - Coupled Morphological Classifier
    Regression model for morphological features of noun phrases
  - First-order Rule Learner (based on FOIL)
    e.g., ***athleteInLeague(X, NBA) ⇒ athletePlaysSport(X, basketbal)***

- More **mutual-exclusion** constraints using seeds/counter seeds and “mutex-relations”

NELL: Never-Ending Language Learning

[Carlson,Mitchell et al: AAAI 10]

Predicate	Web URL	Extraction Template
academicField	http://scholendow.ais.msu.edu/student/SholSearch.Asp	`  [X] -`
athlete	http://www.quotes-search.com/d_occupation.aspx?o=+athlete	`<a href='d_author.aspx?a=[X]'>-`
bird	http://www.michaelforsberg.com/stock.html	`<option>[X]</option>`
bookAuthor	http://lifebehindthecurve.com/	`</li> <li>[X] by [Y] –`

SEAL wrappers

Probability	Consequent	Antecedents
0.95	athletePlaysSport(X, basketball)	$\Leftarrow$ athleteInLeague(X, NBA)
0.91	teamPlaysInLeague(X, NHL)	$\Leftarrow$ teamWonTrophy(X, Stanley Cup)
0.90	athleteInLeague(X, Y)	$\Leftarrow$ athletePlaysForTeam(X, Z), teamPlaysInLeague(Z, Y)
0.88	cityInState(X, Y)	$\Leftarrow$ cityCapitalOfState(X, Y), cityInCountry(X, USA)
† 0.62	newspaperInCity(X, New York)	$\Leftarrow$ companyEconomicSector(X, media), generalizations(X, blog)

Deduction rules

Predicate	Feature	Weight
mountain	LAST=peak	1.791
mountain	LAST=mountain	1.093
mountain	FIRST=mountain	-0.875
musicArtist	LAST=band	1.853
musicArtist	POS=DT_NNS	1.412
musicArtist	POS=DT_JJ_NN	-0.807
newspaper	LAST=sun	1.330
newspaper	LAST=press	1.276
newspaper	LAST=university	-0.318
university	LAST=college	2.076
university	PREFIX=uc	1.999
university	LAST=university	1.745
university	FIRST=college	-1.381
visualArtMovement	SUFFIX=ism	1.282
visualArtMovement	PREFIX=journ	-0.234
visualArtMovement	PREFIX=budd	-0.253

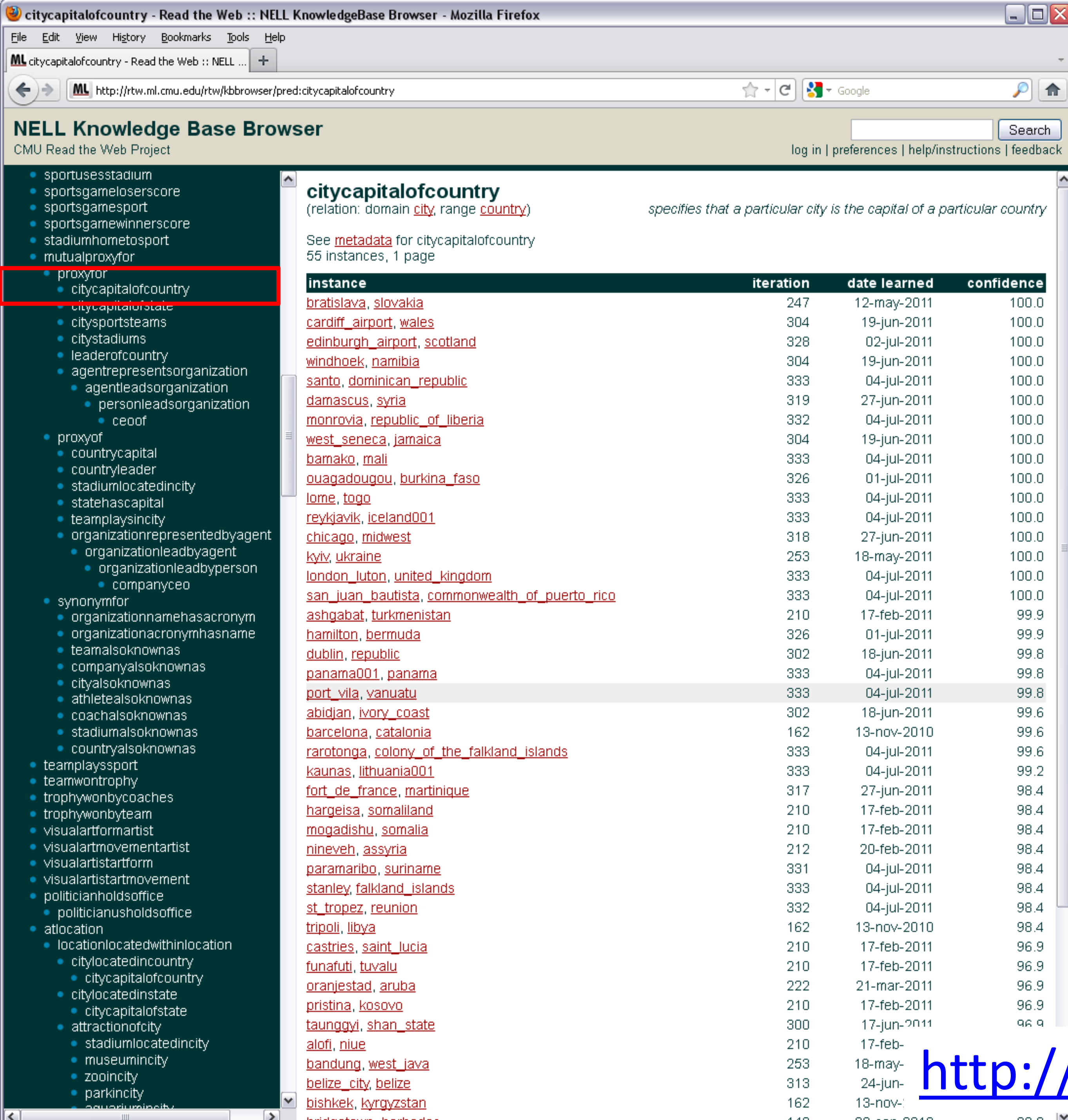
Morphological features & weights

Predicate	Pattern
emotion	hearts full of X
beverage	cup of aromatic X
newspaper	op-ed page of X
teamPlaysInLeague	X ranks second in Y
bookAuthor	Y classic X

Extraction patterns

NELL

Example Output



citycapitalofcountry - Read the Web :: NELL KnowledgeBase Browser - Mozilla Firefox

http://rtw.ml.cmu.edu/rtw/kbbrowser/pred:citycapitalofcountry

NELL Knowledge Base Browser
CMU Read the Web Project

log in | preferences | help/instructions | feedback

- sportsusesstadium
- sportsgameloserscore
- sportsgamesport
- sportsgamewinnerscore
- stadiumhometosport
- mutualproxyfor
  - proxyfor
    - citycapitalofcountry
    - citycapitalofstate
    - citysportsteams
    - citystadiums
    - leaderofcountry
    - agentrepresentsorganization
      - agentleadsorganization
        - personleadsorganization
          - ceoof
  - proxyof
    - countrycapital
    - countryleader
    - stadiumlocatedincity
    - statehascapital
    - teamplaysincity
    - organizationrepresentedbyagent
      - organizationleadbyagent
        - organizationleadbyperson
          - companyceo
  - synonymfor
    - organizationnamehasacronym
    - organizationacronymhasname
    - teamalsoknownas
    - companyalsoknownas
    - cityalsoknownas
    - athletealsoknownas
    - coachalsoknownas
    - stadiumalsoknownas
    - countryalsoknownas
- teamplayssport
- teamwontrophy
- trophywonbycoaches
- trophywonbyteam
- visualartformartist
- visualartmovementartist
- visualartistartform
- visualartistartmovement
- politicianholdsoffice
  - politicianusholdsoffice
- atlocation
  - locationlocatedwithinlocation
    - citylocatedincountry
      - citycapitalofcountry
    - citylocatedinstate
      - citycapitalofstate
    - attractionofcity
      - stadiumlocatedincity
      - museumincity
      - zooincity
      - parkincity

citycapitalofcountry
(relation: domain city, range country) *specifies that a particular city is the capital of a particular country*

See metadata for citycapitalofcountry
55 instances, 1 page

instance	iteration	date learned	confidence
bratislava, slovakia	247	12-may-2011	100.0
cardiff_airport, wales	304	19-jun-2011	100.0
edinburgh_airport, scotland	328	02-jul-2011	100.0
windhoek, namibia	304	19-jun-2011	100.0
santo, dominican_republic	333	04-jul-2011	100.0
damascus, syria	319	27-jun-2011	100.0
monrovia, republic_of_liberia	332	04-jul-2011	100.0
west_seneca, jamaica	304	19-jun-2011	100.0
bamako, mali	333	04-jul-2011	100.0
ouagadougou, burkina_faso	326	01-jul-2011	100.0
lome, togo	333	04-jul-2011	100.0
reykjavik, iceland001	333	04-jul-2011	100.0
chicago, midwest	318	27-jun-2011	100.0
kyiv, ukraine	253	18-may-2011	100.0
london_luton, united_kingdom	333	04-jul-2011	100.0
san_juan_bautista, commonwealth_of_puerto_rico	333	04-jul-2011	100.0
ashgabat, turkmenistan	210	17-feb-2011	99.9
hamilton, bermuda	326	01-jul-2011	99.9
dublin, republic	302	18-jun-2011	99.8
panama001, panama	333	04-jul-2011	99.8
port_vila, vanuatu	333	04-jul-2011	99.8
abidjan, ivory_coast	302	18-jun-2011	99.6
barcelona, catalonia	162	13-nov-2010	99.6
rarotonga, colony_of_the_falkland_islands	333	04-jul-2011	99.6
kaunas, lithuania001	333	04-jul-2011	99.2
fort_de_france, martinique	317	27-jun-2011	98.4
hargeisa, somaliland	210	17-feb-2011	98.4
mogadishu, somalia	210	17-feb-2011	98.4
nineveh, assyria	212	20-feb-2011	98.4
paramaribo, suriname	331	04-jul-2011	98.4
stanley, falkland_islands	333	04-jul-2011	98.4
st_tropez, reunion	332	04-jul-2011	98.4
tripoli, libya	162	13-nov-2010	98.4
castries, saint_lucia	210	17-feb-2011	96.9
funafuti, tuvalu	210	17-feb-2011	96.9
oranjestad, aruba	222	21-mar-2011	96.9
pristina, kosovo	210	17-feb-2011	96.9
taunggyi, shan_state	300	17-jun-2011	96.9
alofi, niue	210	17-feb-	
bandung, west_java	253	18-may-	
belize_city, belize	313	24-jun-	
bishkek, kyrgyzstan	162	13-nov-	
bridgetown, barbados	142	03-sep-2010	93.8

Browse by **concepts** & **relations**

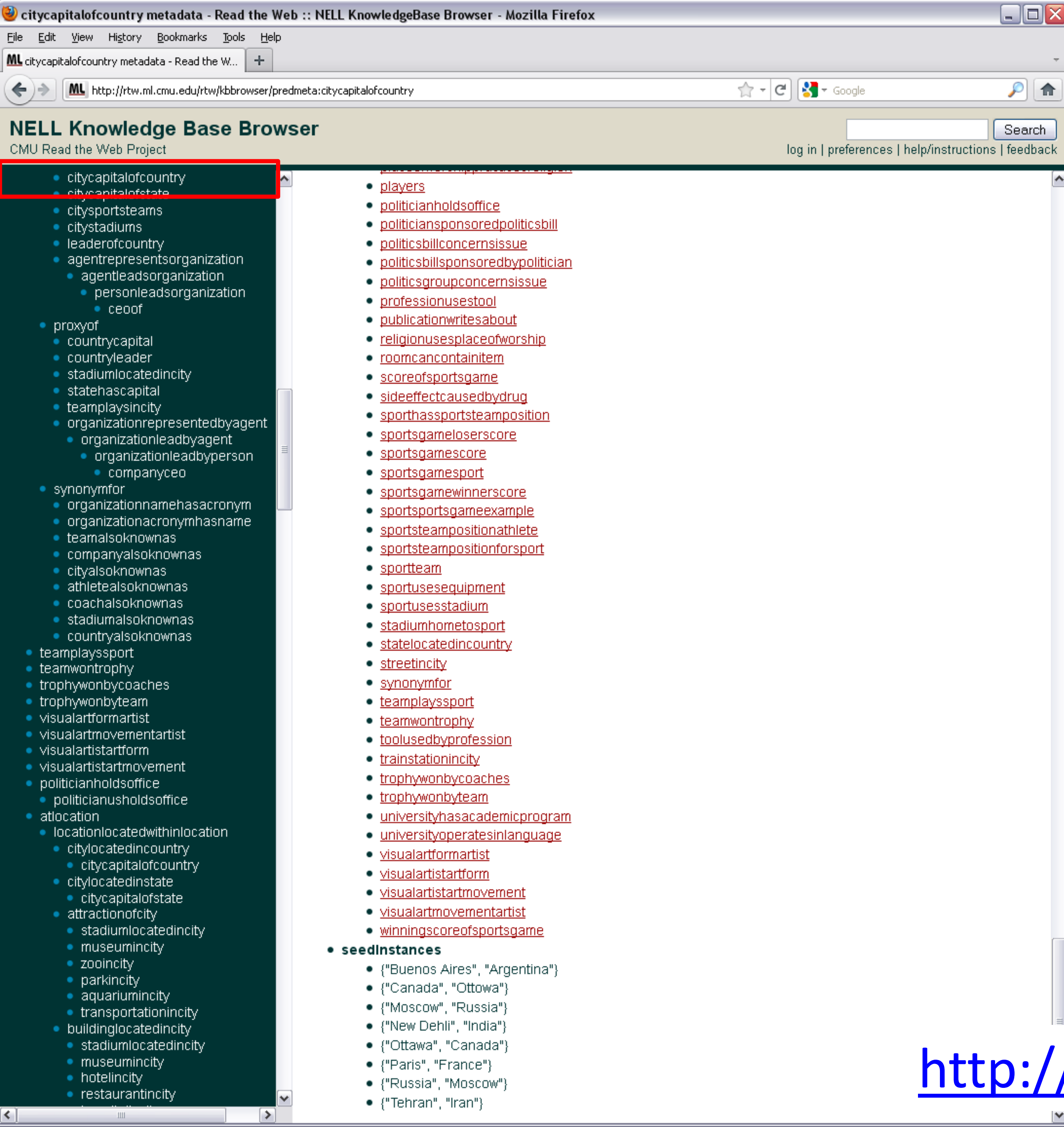
CityCapitalOfCountry

- 55 high-confidence instances

http://rtw.ml.cmu.edu/

NELL

Example Output



Metadata

- Domain/range
- Extraction patterns
- Freebase Ids
- Negative examples
- Seed examples
- Mutex-relations

…

http://rtw.ml.cmu.edu/

TextRunner

[Cafarella,Banko,Etzioni,Soderland et al: AAAI 06, NAACL-HLT 07, SIGMOD Rec. 07, IJCAI 07]

Machine Reading

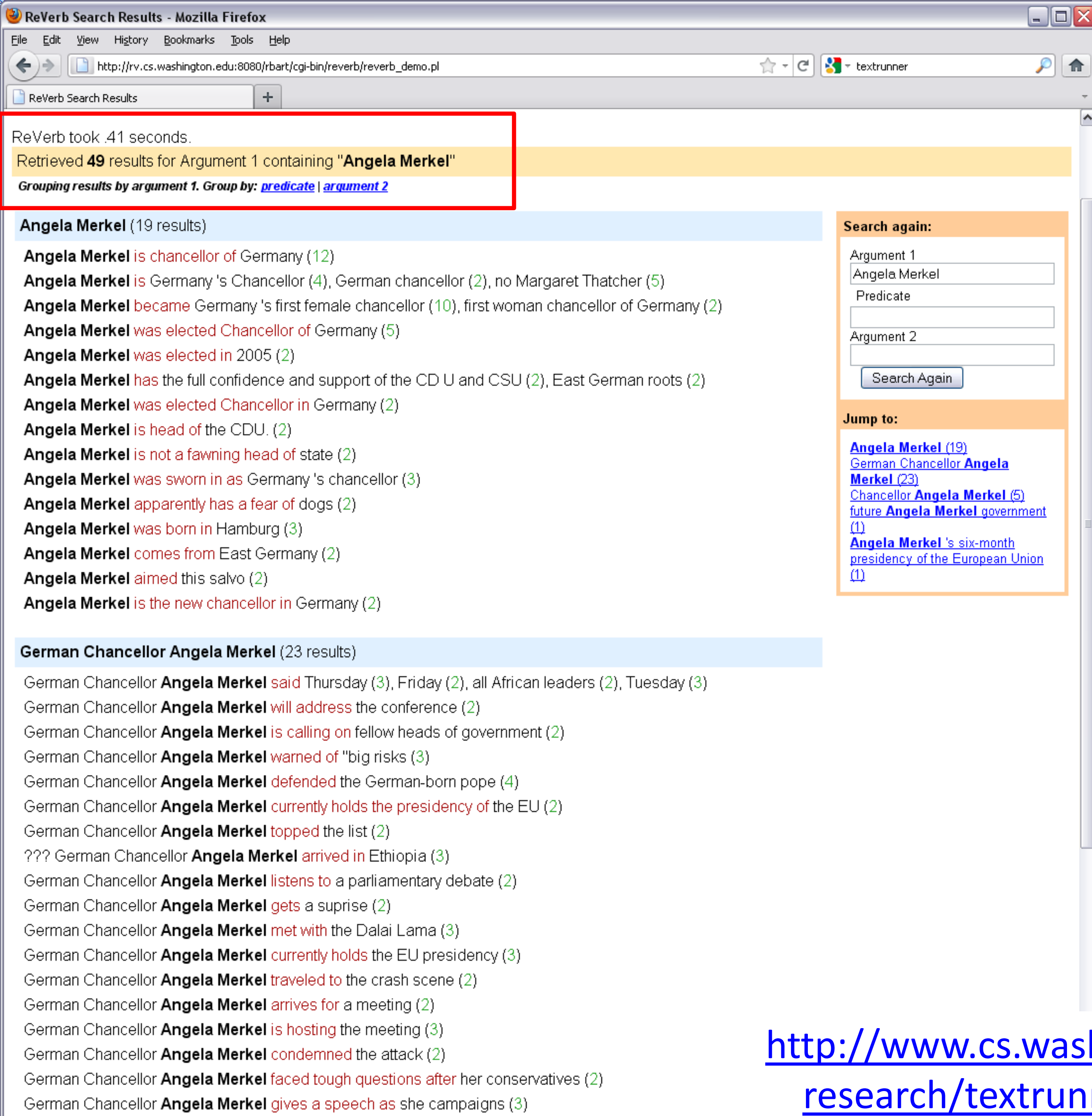
- Automatic, unsupervised understanding of text
- Open set of entities and relations (→ assertions)

- **Single-pass extractor**
  - POS tagging, extract triplets of NP x VP x NP
- **Self-supervised classifier**
  - Trained on POS sequences from trustworthy sentences
- **Synonym resolution**
  - Unsupervised clustering of nouns and verbal phrases
- **Query interface**
  - Issue structured keyword over triplet patterns

- REVERB
  - Syntactic & lexical constraints on verbs and nouns; open-source release [Fader,Soderland,Eztioni: EMNLP 11]
  - Deep NLP: semantic role labeling for Open-IE [Christensen,Mausam,Soderland,Eztioni: NAACL 10]

TextRunner

Example Ouput



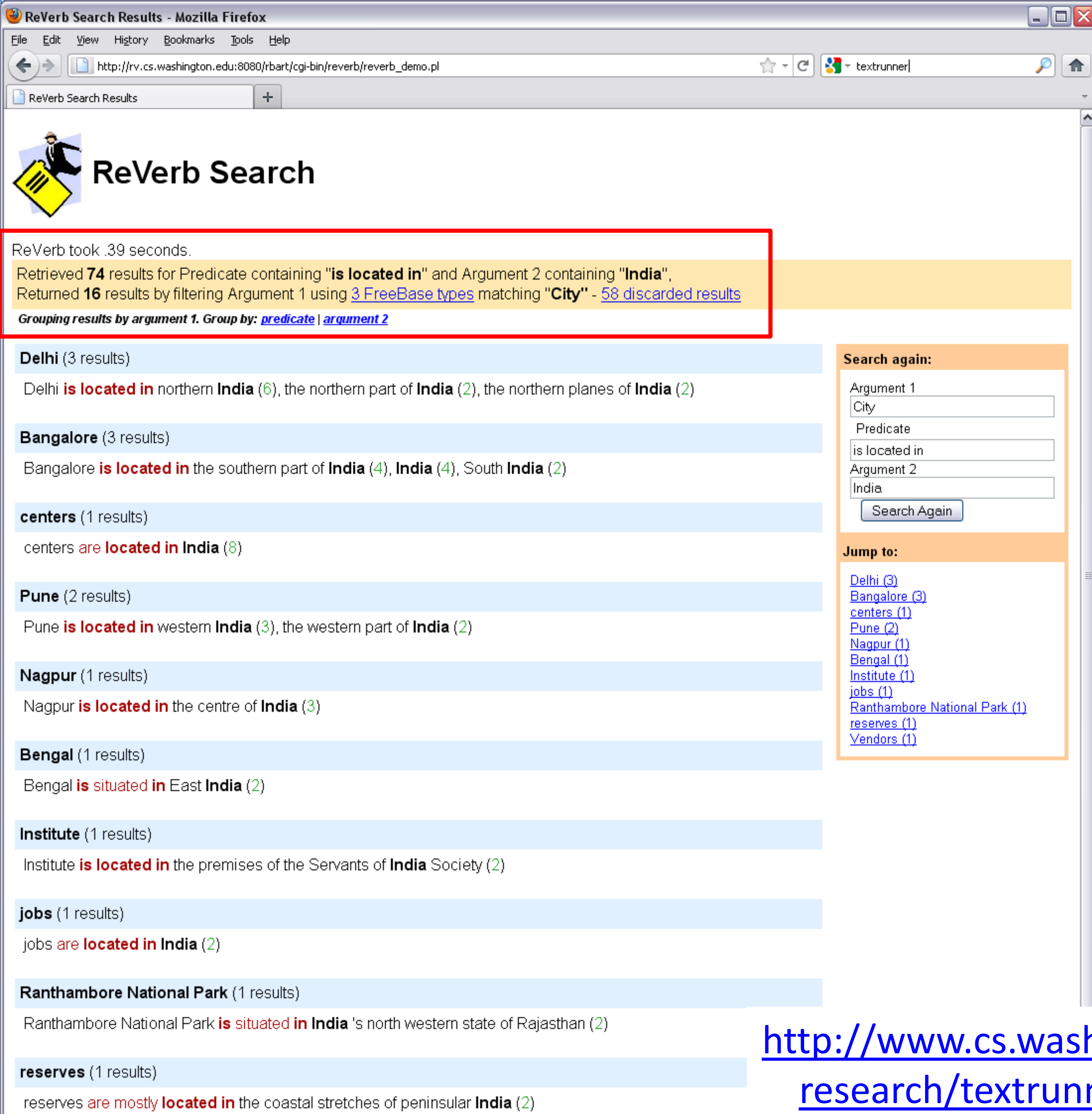
Triplets grouped by **Arg1** X **Pred** X **Arg2**

- Entity names & predicates matched but not canonicalized

http://www.cs.washington.edu/research/textrunner/reverbdemo.html

TextRunner

Example Ouput



Can filter arguments

by FreeBase concepts

http://www.cs.washington.edu/research/textrunner/reverbdemo.html

Omnivore

[Cafarella: CIDR 07]

Extract and query a comprehensive Web database

- Combines extractors from
  - KnowItAll
  - TextRunner
  - WebTables
  - Weak associations (unknown relations)

    e.g., Mike_Cafarella <?> Alon_Halevy

- SQL query interface
  - Individual sources unified into a relational schema for on-the-fly querying

Outline for Part III

- Domain-oriented IE vs. Open-domain IE ✔
  - What to extract: entities, classes, binary & higher-arity relations
- Entities, Classes & Subsumptions ✔
  - WordNet concepts, Wikipedia categories, entity disambiguation
- Pattern-based Knowledge Harvesting ✔
  - Wrapper induction, WebTables, statistical pattern mining
- Probabilistic Extraction Models ✔
  - HMMs, MEMMs, CRFs
- Constraints & Reasoning ✔
  - MLNs, CCMs, FactorIE, SOFIE/PROSPERA
- Open-domain IE ✔
  - ReadTheWeb, TextRunner, Omnivore, REVERB
- Advanced reasoning
  - Temporal/spatial annotations of facts

Higher-arity Relations – Space & Time

- YAGO2 Numbers

	Just Wikipedia	Incl. Gazetteer Data
#Relations	104	114
#Classes	364,740	364,740
#Entities	2,641,040	9,804,102
#Facts	**601,984,236**	**2,231,699,989**
- base relations	120,639,022	461,893,127
- types & classes	8,649,652	15,716,697
- space, time & proven.	**472,695,562**	**1,754,090,165**
Size (CSV format)	**23.4 GB**	**121 GB**

estimated **precision > 95%**
(for basic relations excl. space, time & provenance)

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

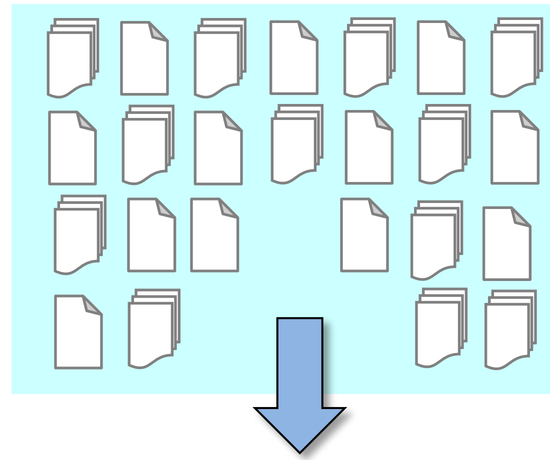
French Marriage Problem (Revisited)

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)

Challenge: Temporal Knowledge Harvesting

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

Consistency constraints are potentially helpful:

- functional dependencies: *{husband, time}* $\rightarrow$ *{wife, time}*
- inclusion dependencies: *marriedPerson* $\subseteq$ *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 Raphael 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

ame, see *Sárközi (surname)*.

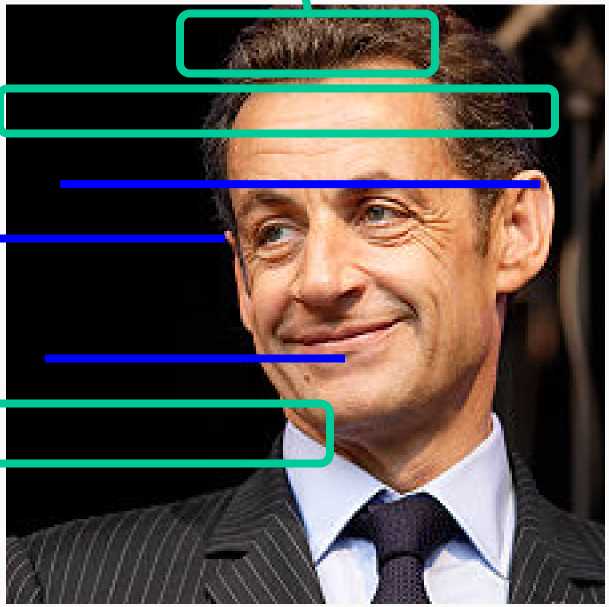
kozi] (help·info)), born **Nicolas Paul Stéphane Sarközy de Nagy-Bocsa** on
resident of the French Republic and *ex officio* Co-Prince of Andorra. He
efeating Socialist Party candidate Ségolène Royal 10 days earlier.

Union for a Popular Movement (UMP). Under Jacques Chirac's presidency he
erre Raffarin's (UMP) first two governments (from May 2002 to March 2004),
Raffarin's last government (March 2004 to May 2005) and again Minister of
rnment (2005–2007).

council of the Hauts-de-Seine department from 2004 to 2007 and mayor of
mmunes of France from 1983 to 2002. He was Minister of the Budget in the
decessor of the UMP) during François Mitterrand's last term.

he French economy.[1][2][3] He has pledged to revive the work ethic, promote
reign affairs he has promised a strengthening of the entente cordiale with the
with the United States.[5] He married Carla Bruni-Sarkozy on 2 February 2008

Nicolas Sarkozy



(Even More Difficult) Implicit Dating

vague dates
relative dates

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]

narrative text
relative order

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]

TARSQI: Extracting Time Annotations

http://www.timeml.org/site/tarsqi/

[Verhagen et al: ACL'05]

Hong Kong is poised to hold the first election in more than **half** **<TIMEX3 tid="t3" TYPE="DURATION" VAL="P100Y">a century</TIMEX3>** that includes a democracy advocate seeking high office in territory controlled by the Chinese government in Beijing. A pro-democracy politician, Alan Leong, announced **<TIMEX3 tid="t4" TYPE="DATE" VAL="20070131">Wednesday</TIMEX3>** that he had obtained eno

appear on the ballot to become the territory's next chief executive. But h

he had no chance of beating the Beijing-backed incumbent, Donald Tsang

election. Under electoral rules imposed by Chinese officials, only 796 pe

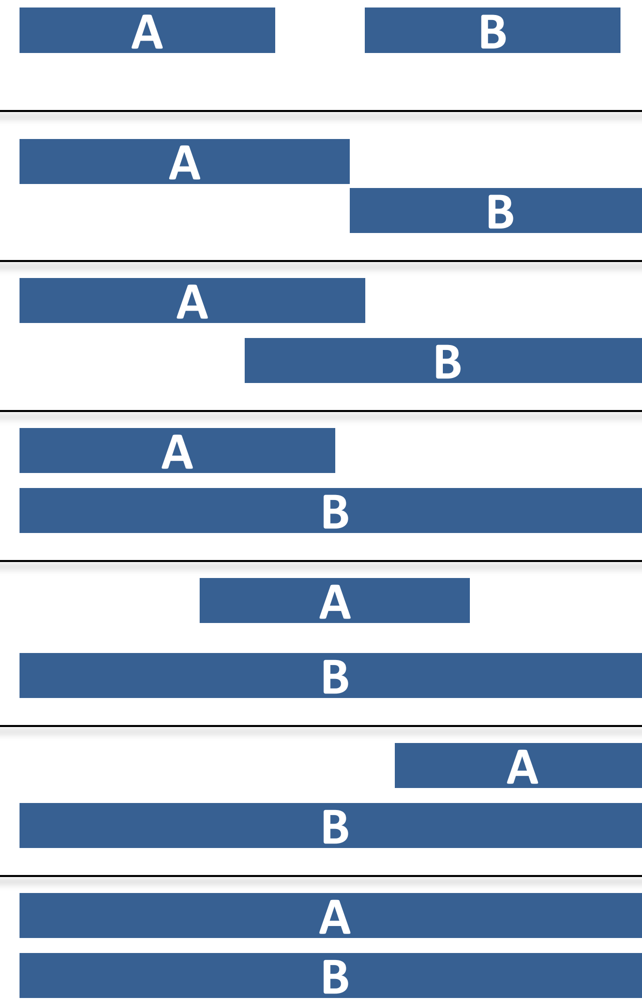
extraction errors!

committee – the bulk of them with close ties to mainland China – will be allowed to vote in the **<TIMEX3 tid="t5" TYPE="DATE" VAL="20070325">March 25</TIMEX3>** election. It will be the first contested election for chief executive since Britain returned Hong Kong to China in **<TIMEX3 tid="t6" TYPE="DATE" VAL="1997">1997</TIMEX3>**. Mr. Tsang, an able administrator who took office during the early stages of a sharp economic upturn in **<TIMEX3 tid="t7" TYPE="DATE" VAL="2005">2005</TIMEX3>**, is popular with the general public. Polls consistently indicate that three-fifths of Hong Kong's people approve of the job he has been doing. It is of course a foregone conclusion – Donald Tsang will be elected and will hold office for **<TIMEX3 tid="t9" beginPoint="t0" endPoint="t8" TYPE="DURATION" VAL="P5Y">another five years </TIMEX3>**, said Mr. Leong, the former chairman of the Hong Kong Bar Association.

13 Relations between Time Intervals

[Allen, 1984; Allen & Hayes 1989]

A Before B	B After A
A Meets B	B MetBy A
A Overlaps B	B OverlappedBy A
A Starts B	B StartedBy A
A During B	B Contains A
A Finishes B	B FinishedBy A
A Equal B	

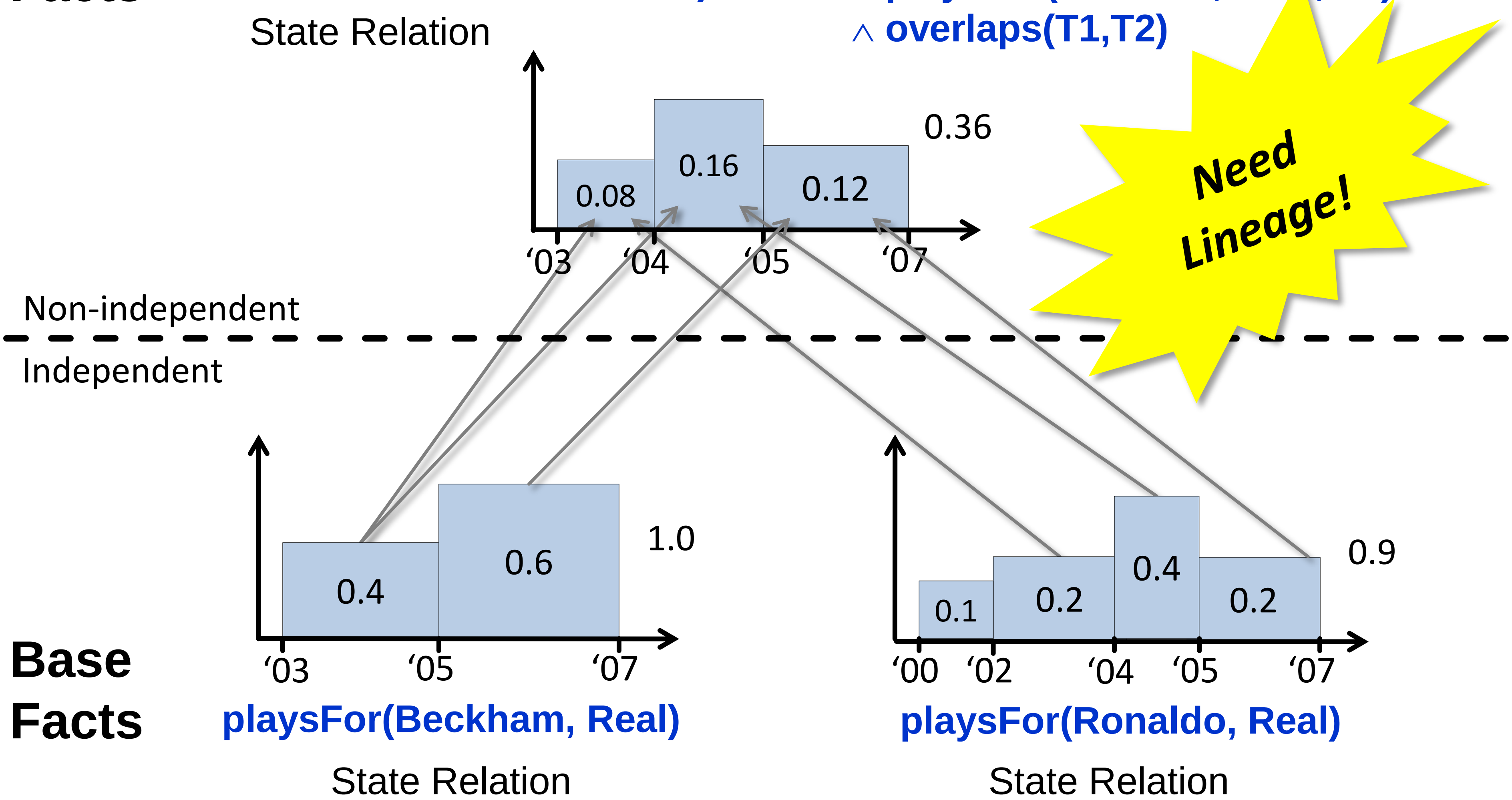


Possible Worlds in Time

[Wang,Yahya,Theobald: MUD Ws. 10]

Derived Facts

teamMates(Beckham, Ronaldo) ← **playsFor(Beckham, Real, T1) ∧ playsFor(Ronaldo, Real, T2) ∧ overlaps(T1,T2)**



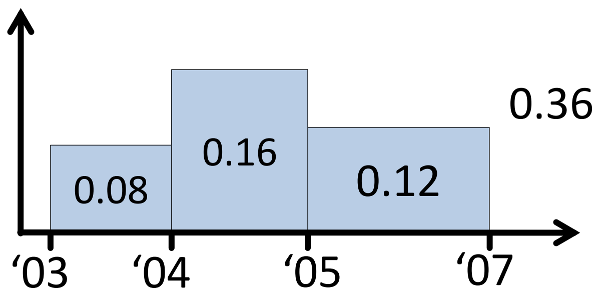
Possible Worlds in Time

[Wang,Yahya,Theobald: MUD Ws. 10]

Derived Facts

teamMates(Beckham, Ronaldo) ← **playsFor(Beckham, Real, T1)**
∧ **playsFor(Ronaldo, Real, T2)**
∧ **overlaps(T1,T2)**

State Relation



Need Lineage!

Non-independent

In

B

Fa

- **Closed** and **complete** representation model (incl. lineage)
  → Stanford Trio project [Widom: CIDR'05, Benjelloun et al: VLDB'06]
- **Interval computation** remains **linear** in the number of bins
- **Confidence computation** per bin is **#P-complete**
- In general requires possible-worlds-based sampling techniques (**Luby-Karp** for DNF**, MCMC**-style sampling, etc.)

Outline for Part III

- Domain-oriented IE vs. Open-domain IE ✔
  - What to extract: entities, classes, binary & higher-arity relations
- Entities, Classes & Subsumptions ✔
  - WordNet concepts, Wikipedia categories, entity disambiguation
- Pattern-based Knowledge Harvesting ✔
  - Wrapper induction, WebTables, statistical pattern mining
- Probabilistic Extraction Models ✔
  - HMMs, MEMMs, CRFs
- Constraints & Reasoning ✔
  - MLNs, CCMs, FactorIE, SOFIE/PROSPERA
- Open-domain IE ✔
  - ReadTheWeb, TextRunner, Omnivore, REVERB
- Advanced reasoning ✔
  - Temporal/spatial annotations of facts

Open Problems and Challenges in IE (I)

High precision & high recall at affordable cost

robust, statistical pattern mining & analysis

parallel processing and distributed reasoning, lazy / lifted inference, …

Types and constraints

soft rules & hard constraints, rich DL, beyond CWA

explore & understand different families of constraints

Declarative, self-optimizing workflows

incorporate pattern & reasoning steps into IE queries/programs

Scale, dynamics, life-cycle

grow & maintain KB with near-human-quality over long periods

Open-domain knowledge harvesting

turn names, phrases & table cells into entities & relations

Open Problems and Challenges in IE (II)

Temporal Querying (Revived)

temporal query language (T-SPARQL?)
confidence weights & ranking

Gathering Implicit and Relative Time Annotations

biographies & news, relative orderings
aggregate & reconcile observations

Incomplete and Uncertain Temporal Scopes

incorrect, incomplete, unknown begin/end dates
vague dating

Scalable Consistency Reasoning

extended MaxSat, probabilistic Datalog, graphical models,
scale-up Markov Logic, etc.
for resolving inconsistencies on uncertain facts & uncertain time

Readings for Part III

- E. Agichtein, L. Gravano, J. Pavel, V. Sokolova, A. Voskoboynik. Snowball: a prototype system for extracting relations from large text collections. SIGMOD, 2001.
- E. Alfonseca, M. Pasca, E. Robledo-Arnuncio: Acquisition of instance attributes via labeled and related instances. SIGIR 2010.
- James Allen. Towards a general theory of action and time. Artif.Intell., 23(2), 1984.
- M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni. Open information extraction from the web. IJCAI, 2007.
- R. Baumgartner, S. Flesca, G. Gottlob. Visual web information extraction with Lixto. VLDB, 2001.
- S. Brin. Extracting patterns and relations from the World Wide Web. WebDB, 1998.
- M. J. Cafarella: Extracting and Querying a Comprehensive Web Database. CIDR 2009.
- M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, Y. Zhang. WebTables: exploring the power of tables on the web. PVLDB, 1(1), 2008.
- M. E. Califf, R. J. Mooney. Relational learning of pattern-match rules for information extraction. AAAI, 1999.
- A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr., T. M. Mitchell: Toward an Architecture for Never-Ending Language Learning. AAAI 2010
- A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr., T. M. Mitchell. Coupled semi-supervised learning for information extraction. WSDM, 2010
- J.Christensen, Mausam, S. Soderland, O. Etzioni: Semantic Role Labeling for Open Information Extraction. NAACL 2010.
- P. DeRose, W. Shen, F. Chen, Y. Lee, D. Burdick, A. Doan, R. Ramakrishnan. DBLife: A community information management platform for the database research community. CIDR, 2007.
- A. Doan, L. Gravano, R. Ramakrishnan, S. Vaithyanathan. (Eds.). Special issue on information extraction. SIGMOD Record, 37(4), 2008.
- O. Etzioni, Mi.Banko, M. J. Cafarella: Machine Reading. AAAI 2006.
- O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates. Web-scale information extraction in KnowItAll. WWW, 2004.
- G. Gottlob, C. Koch, R. Baumgartner, M. Herzog, S. Flesca. The Lixto data extraction project - back and forth between theory and practice. PODS, 2004.
- R. Gupta, S. Sarawagi: Answering Table Augmentation Queries from Unstructured Lists on the Web. PVLDB, 2(1), 2009.
- R. Gupta, S. Sarawagi: Joint training for open-domain extraction on the web: exploiting overlap when supervision is limited. WSDM 2011.
- M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. COLING, 1992.
- D. Hindle. Noun classification from predicate-argument structures. ACL, 1990.

Readings for Part III

- J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, Bi. Taneva, S. Thater, G. Weikum. Robust Disambiguation of Named Entities in Text. EMNLP 2011.
- R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, S. Vaithyanathan, H. Zhu. SystemT: a system for declarative information extraction. SIGMOD Record, 37(4), 2008.
- S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti. Collective Annotation of Wikipedia Entities in Web Text. KDD, 2009.
- N. Kushmerick. Wrapper induction: efficiency and expressiveness. Artif. Intell., 118(1-2), 2000.
- J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. ML, 2001.
- X. Liu, Z. Nie, N. Yu, J.-R. Wen. BioSnowball: automated population of Wikis. KDD, 2010.
- A. McCallum, D. Freitag, F. Pereira. Maximum Entropy Markov Models for Information Extraction and Segmentation. ICML 2000.
- A. McCallum, K. Schultz, S. Singh. FACTORIE: Probabilistic Programming via Imperatively Defined Factor Graphs. NIPS, 2009.
- N. Nakashole, M. Theobald, G. Weikum. Find your Advisor: Robust Knowledge Gathering from the Web. WebDB, 2010.
- S.P. Ponzetto, M. Strube: Deriving a Large-Scale Taxonomy from Wikipedia. AAAI 2007.
- H. Poon, P. Domingos: Joint Inference in Information Extraction. AAAI 2007.
- L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, 1989.
- M. Richardson and P. Domingos. Markov Logic Networks. ML, 2006.
- D. Roth, W. Yih. Global Inference for Entity and Relation Identification via a Linear Programming Formulation. MIT Press, 2007.
- S. Sarawagi. Information extraction. Foundations and Trends in Databases, 1(3), 2008.
- S. Sarawagi, W. W. Cohen. Semi-Markov conditional random fields for information extraction. NIPS, 2004.
- W. Shen, X. Li, A. Doan. Constraint-Based Entity Matching. AAAI, 2005.
- S. Singh, K. Schultz, A. McCallum: Bi-directional Joint Inference for Entity Resolution and Segmentation Using Imperatively-Defined Factor Graphs. ECML/PKDD (2) 2009.
- P. Singla, P. Domingos. Entity resolution with Markov Logic. ICDM, 2006.
- P. P. Talukdar, F. Pereira. Experiments in Graph-based Semi-Supervised Learning Methods for Class-Instance Acquisition. ACL 2010.
- Y. Song, H. Wang, Z. Wang, H. Li, W. Chen. Short Text Conceptualization using a Probabilistic Knowledgebase. IJCAI 2011.
- F. M. Suchanek, M. Sozio, G. Weikum. SOFIE: a self-organizing framework for information extraction. WWW, 2009.
- F. M. Suchanek, G. Ifrim, G. Weikum. Combining linguistic and statistical analysis to extract relations from web documents. KDD, 2006.

Readings for Part III

- C. Sutton, A. McCallum. An Introduction to Conditional Random Fields for Relational Learning. MIT Press, 2006.
- P. Venetis, A.Y. Halevy, J. Madhavan, M. Pasca, W. Shen, F. Wu, G. Miao, C. Wu: Recovering Semantics of Tables on the Web. PVLDB 4(9), 2011.
- R. C. Wang, W. W. Cohen. Language-independent set expansion of named entities using the web. ICDM, 2007.
- R. C. Wang, Wi. W. Cohen: Iterative Set Expansion of Named Entities Using the Web. ICDM 2008.
- Y. Wang, M. Yahya, M. Theobald. Time-aware Reasoning in Uncertain Knowledge Bases. VDLB/MUD, 2010.
- D. S. Weld, R. Hoffmann, F. Wu. Using Wikipedia to bootstrap open information extraction. SIGMOD Record, 37(4), 2008.
- M. L. Wick, A. McCallum, G. Miklau: Scalable Probabilistic Databases with Factor Graphs and MCMC. PVLDB 3(1).
- F. Wu, D. S. Weld. Autonomously semantifying Wikipedia. CIKM, 2007.
- F. Wu, D. S. Weld. Automatically refining the Wikipedia infobox ontology. WWW, 2008.
- A. Yates, M. Banko, M. Broadhead, M. J. Cafarella, O. Etzioni, S. Soderland. TextRunner: Open information extraction on the web. HLT-NAACL, 2007.
- M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, G. Weikum: AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables. VLDB 2007.
- J. Zhu, Z. Nie, X. Liu, B. Zhang, J.-R. Wen. StatSnowball: a statistical approach to extracting entity relationships. WWW, 2009.

Outline for Part IV

- Querying Knowledge Bases

  A short overview of SPARQL

  Extensions to SPARQL

- Searching and Ranking Entities
- Searching and Ranking Facts
- Advanced Query Interfaces

SPARQL

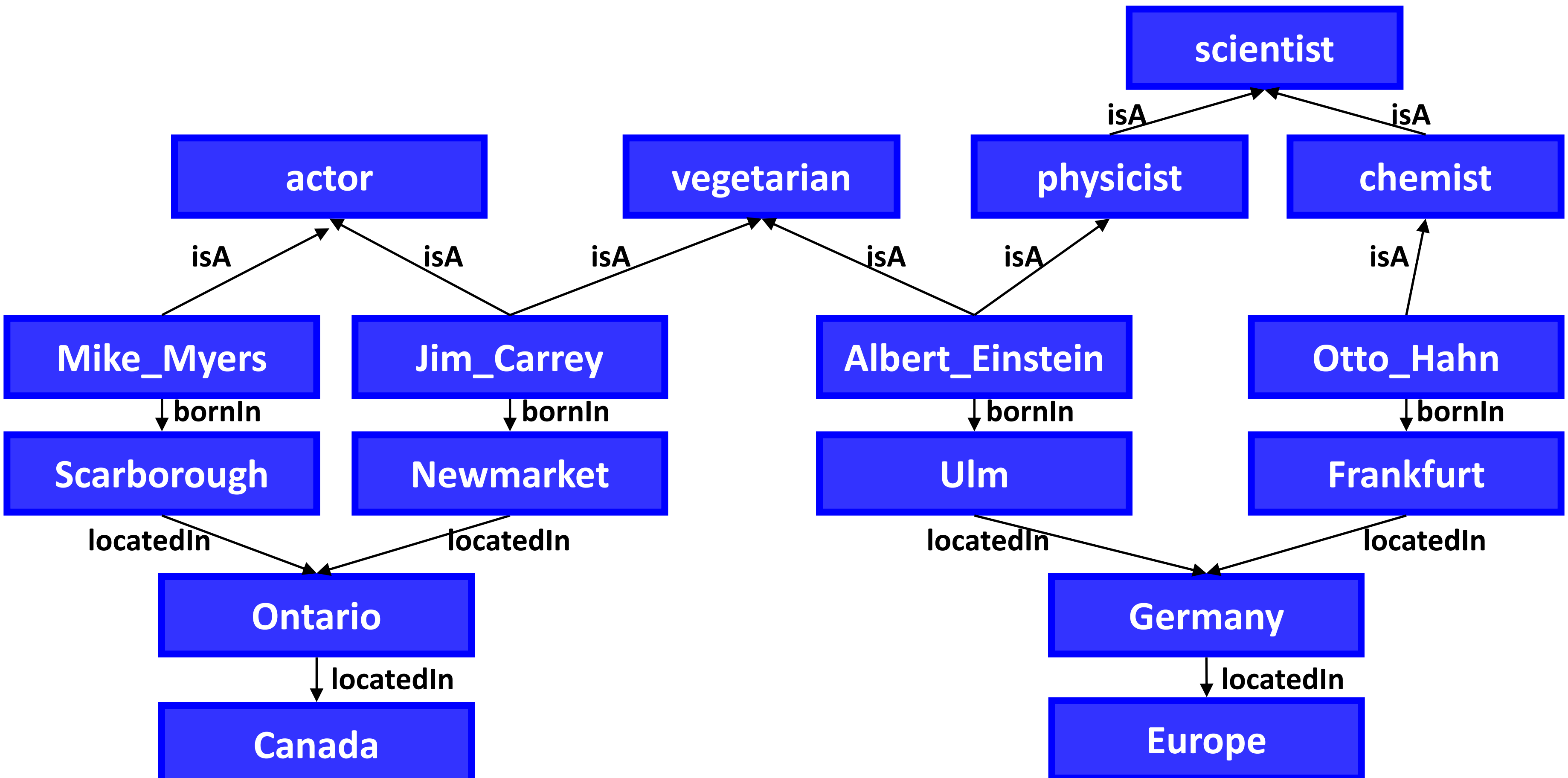
- Query language for RDF from the W3C
- Main component:
  - **select-project-join** combination of **triple patterns**

graph pattern queries on the knowledge base

SPARQL – Example

Example query:

Find all actors from Ontario (that are in the knowledge base)



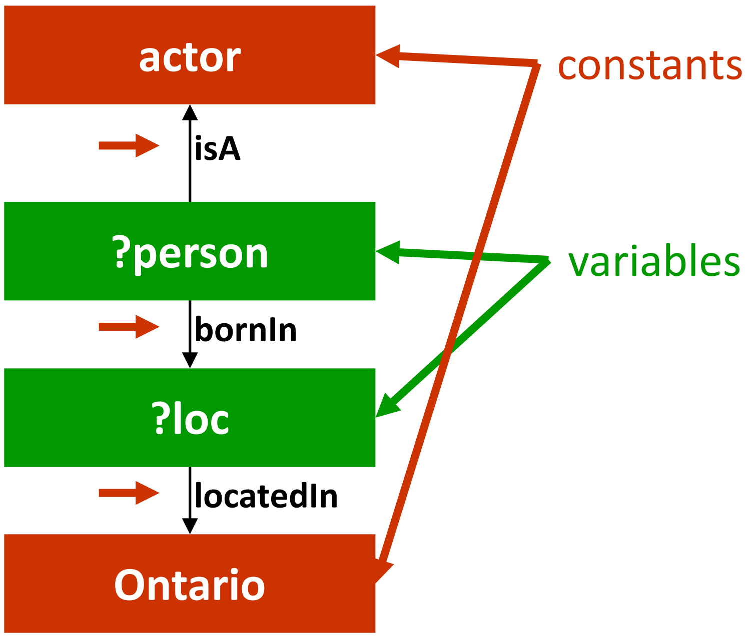
SPARQL – Example

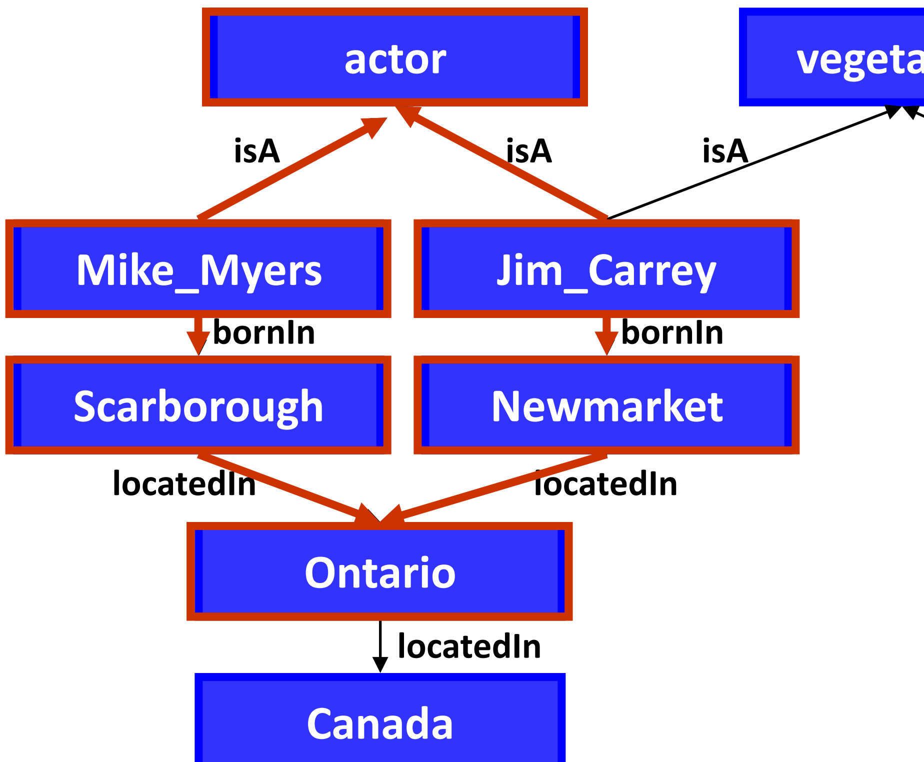
Example query:

Find all actors from Ontario (that are in the knowledge base)

```
SELECT ?person WHERE ?person isA actor. ?person bornIn ?loc.
                     ?loc locatedIn Ontario.
```

Find **subgraphs** of this form:





SPARQL – More Features

- Eliminate duplicates in results

```
SELECT DISTINCT ?c WHERE {?person isA actor. ?person bornIn ?loc.
                 ?loc locatedIn ?c}
```

- Return results in some order

```
SELECT ?person WHERE {?person isA actor. ?person bornIn ?loc.
                      ?loc locatedIn Ontario} ORDER BY DESC(?person)
```

  with optional **LIMIT n** clause

- Optional matches and filters on bounded vars

```
SELECT ?person WHERE {?person isA actor.
                      OPTIONAL{?person bornIn ?loc}.
                      FILTER (!BOUND(?loc))}
```

- More operators: **ASK**, **DESCRIBE**, **CONSTRUCT**

SPARQL: Extensions from W3C

W3C SPARQL 1.1 draft:

- Aggregations (`COUNT`, `AVG`, …) and grouping
- Subqueries
- Negation: syntactic sugar for
  ```
  OPTIONAL {?x … }
  FILTER(!BOUND(?x))
  ```
- Expressions in SELECT clause:
  ```
  SELECT (?a+?b) as ?sum
  ```
- Label constraints on paths:
  ```
  ?x foaf:knows/foaf:knows/foaf:name ?name
  ```
- More functions and operators

SPARQL: Extensions from Research (1)

More complex graph patterns:

- Transitive paths [Anyanwu et al., WWW07]

```
SELECT ?p, ?c WHERE {
?p isA scientist .
?p ??r ?c. ?c isA Country. ?c locatedIn Europe .
PathFilter(cost(??r) < 5).
PathFilter (containsAny(??r,?t )). ?t isA City. }
```

- Regular expressions [Kasneci et al., ICDE08]

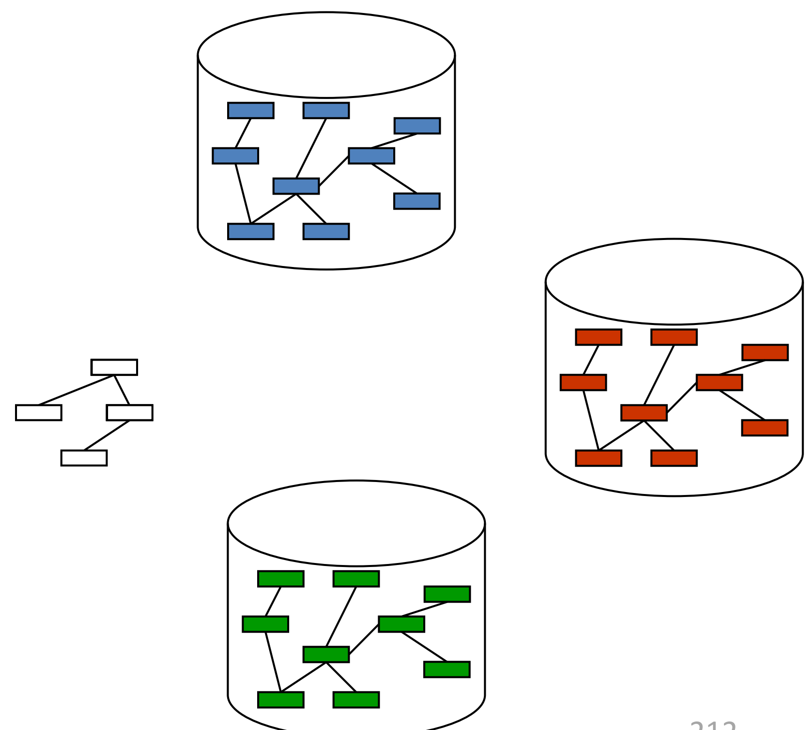
```
SELECT ?p, ?c WHERE {
?p isA ?s. ?s isA scientist.
?p (bornIn | livesIn | citizenOf) locatedIn* Europe.}
```

Now mostly covered by the SPARQL 1.1 Query proposal

SPARQL: Extensions from Research (2)

Queries over federated RDF sources:

- Determine distribution of triple patterns as part of query (for example in ARQ from Jena)
- Automatically route triple predicates to useful sources



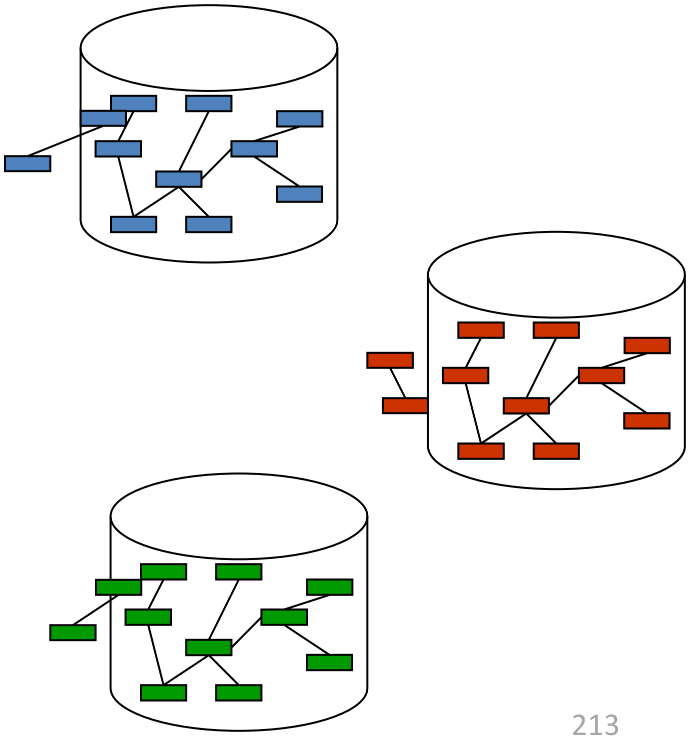
SPARQL: Extensions from Research (2)

Queries over federated RDF sources:

- Determine distribution of triple patterns as part of query (for example in ARQ from Jena)
- Automatically route triple predicates to useful sources

Potentially requires mapping of identifiers from different sources

SPARQL 1.1 will support explicit federation of sources



RDF+SPARQL: Systems

- BigOWLIM
- OpenLink Virtuoso
- Jena with different backends
- Sesame
- OntoBroker
- SW-Store, Hexastore, RDF-3X (no reasoning)

System deployments with $>10^{11}$ triples

(see **`http://esw.w3.org/LargeTripleStores`**)

More details on systems in our tutorial at the Reasoning Web Summer School: „Database foundations for scalable RDF processing“

Outline for Part IV

- Querying Knowledge Bases ✔
- Searching and Ranking Entities
  Entity Importance: Graph Analysis
  Entity Search: Language Models
- Searching and Ranking Facts
- Advanced Query Interfaces

Why ranking is essential

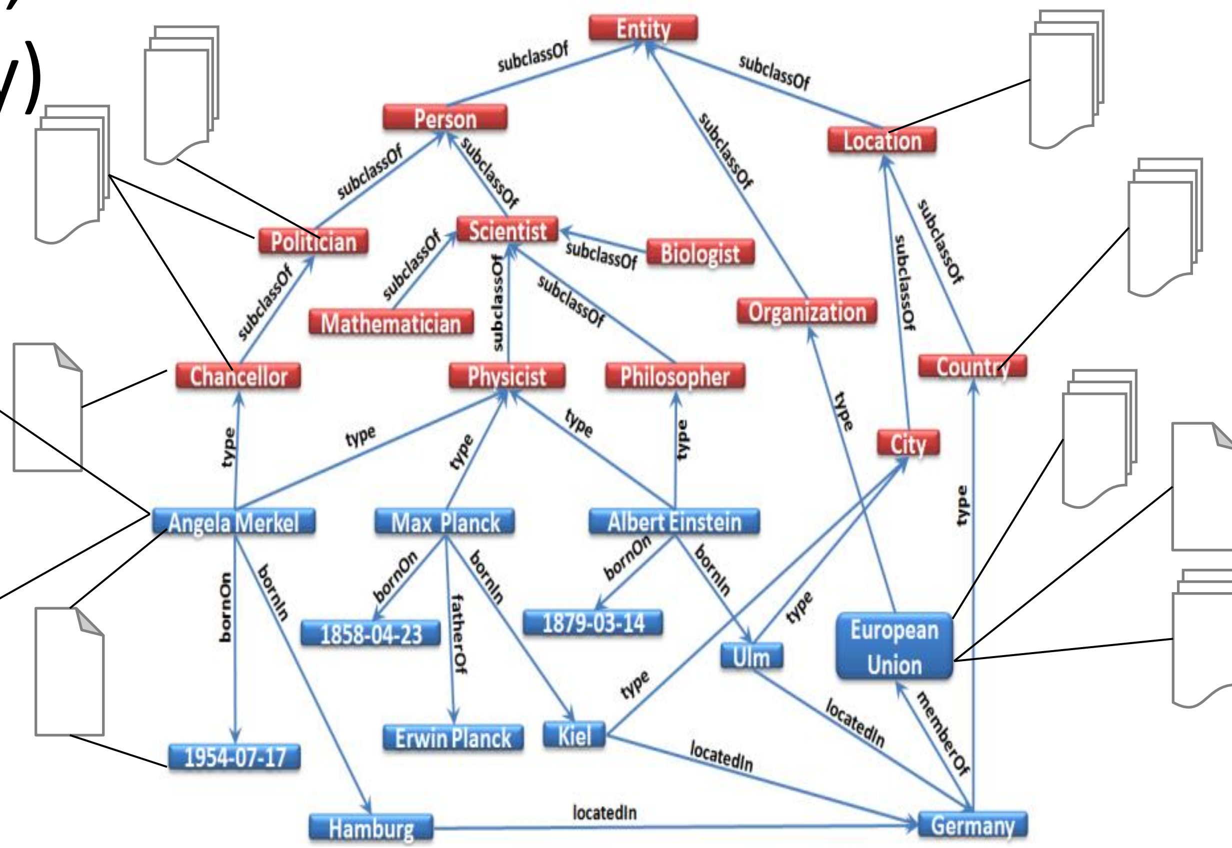
- Queries often have a huge number of results:
  - scientists from Canada
  - conferences in Toronto
  - publications in databases
  - actors from the U.S.
- Ranking as integral part of search
- Huge number of app-specific ranking methods: paper/citation count, impact, salary, …
- Need for **generic**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

Extending Entities with Keywords

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

⇒Associate entities with terms in those documents, keywords in URIs, literals, …
(**context** of entity)

chancellor Germany
scientist election
Stuttgart21 Guido
Westerwelle France
Nicolas Sarkozy



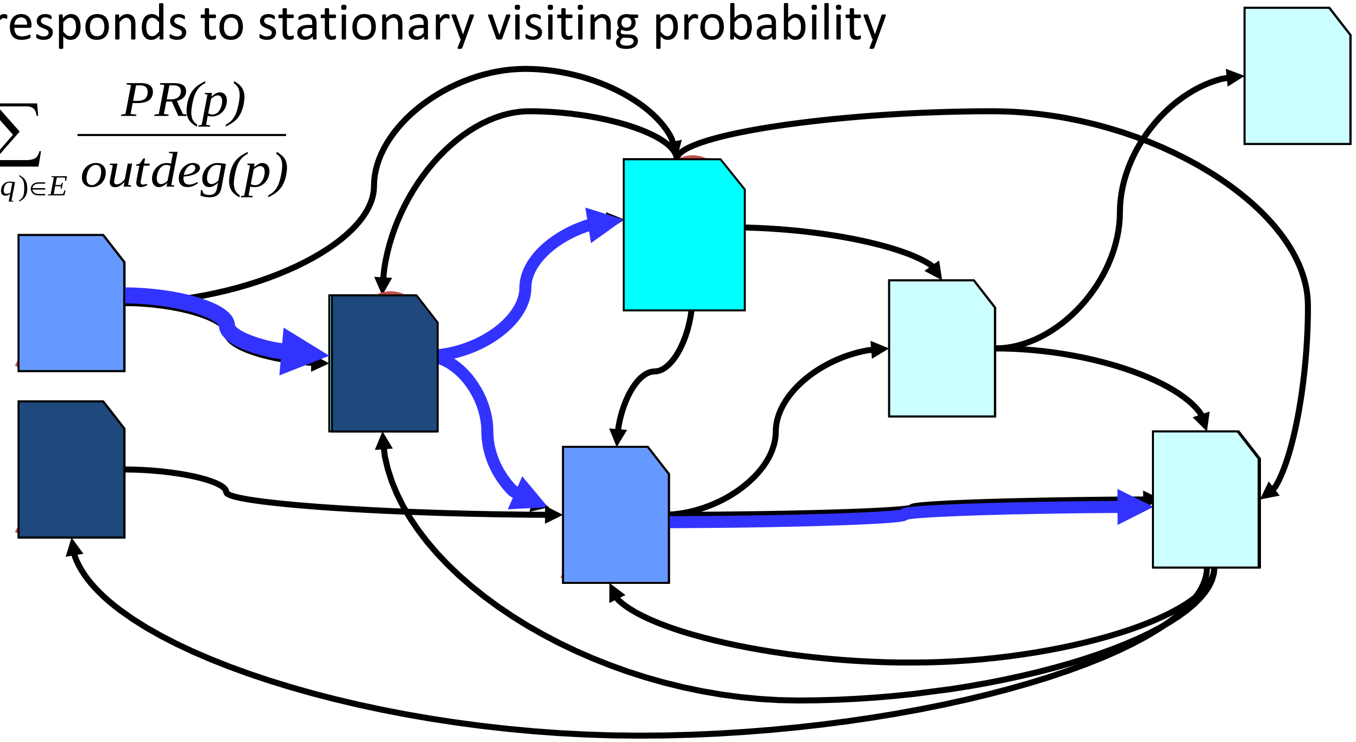
Digression 1: Graph Authority Measures

Idea: incoming links are endorsements & increase node authority, authority is higher if links come from high-authority nodes

Important instance: PageRank [Brin&Page, 1998]

- random walk of the Web graph: uniformly random choice of links, random jumps
- Authority of a page corresponds to stationary visiting probability

$$PR(q) = \frac{\varepsilon}{|V|} + (1-\varepsilon) \cdot \sum_{(p,q)\in E} \frac{PR(p)}{outdeg(p)}$$



Easy application to RDF data (with different weights for different relations):
ObjectRank (Balmin et al., 2004), EntityRank (Cheng et al., 2007), TripleRank (Franz et al., 2009)

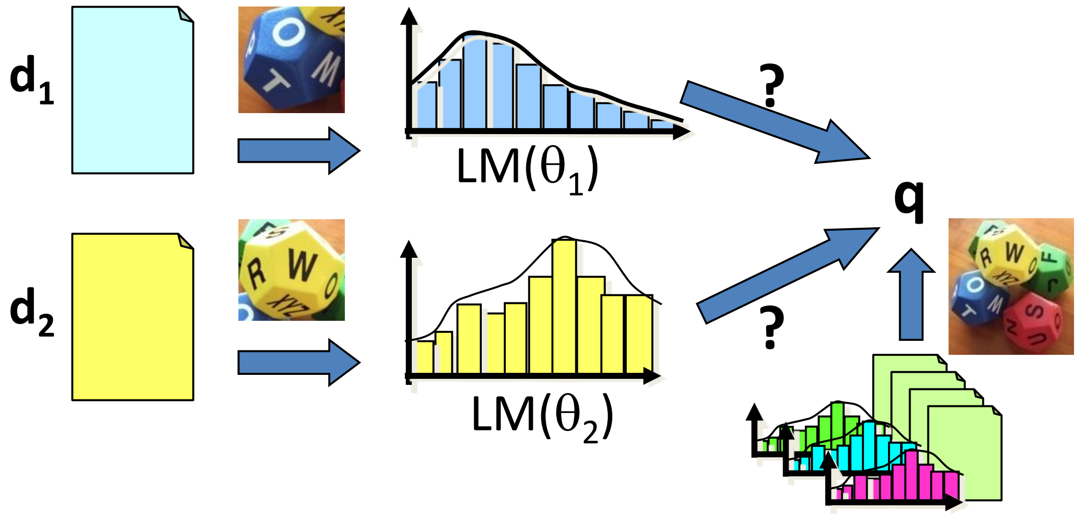
Keyword-Based Entity Search: Principles

Combine several paradigms:

- **Graph-based authority** measure to determine important entities
- **Keyword search** on associated terms to determine candidate entities
- Ranking can **combine** entity importance with keyword-based score

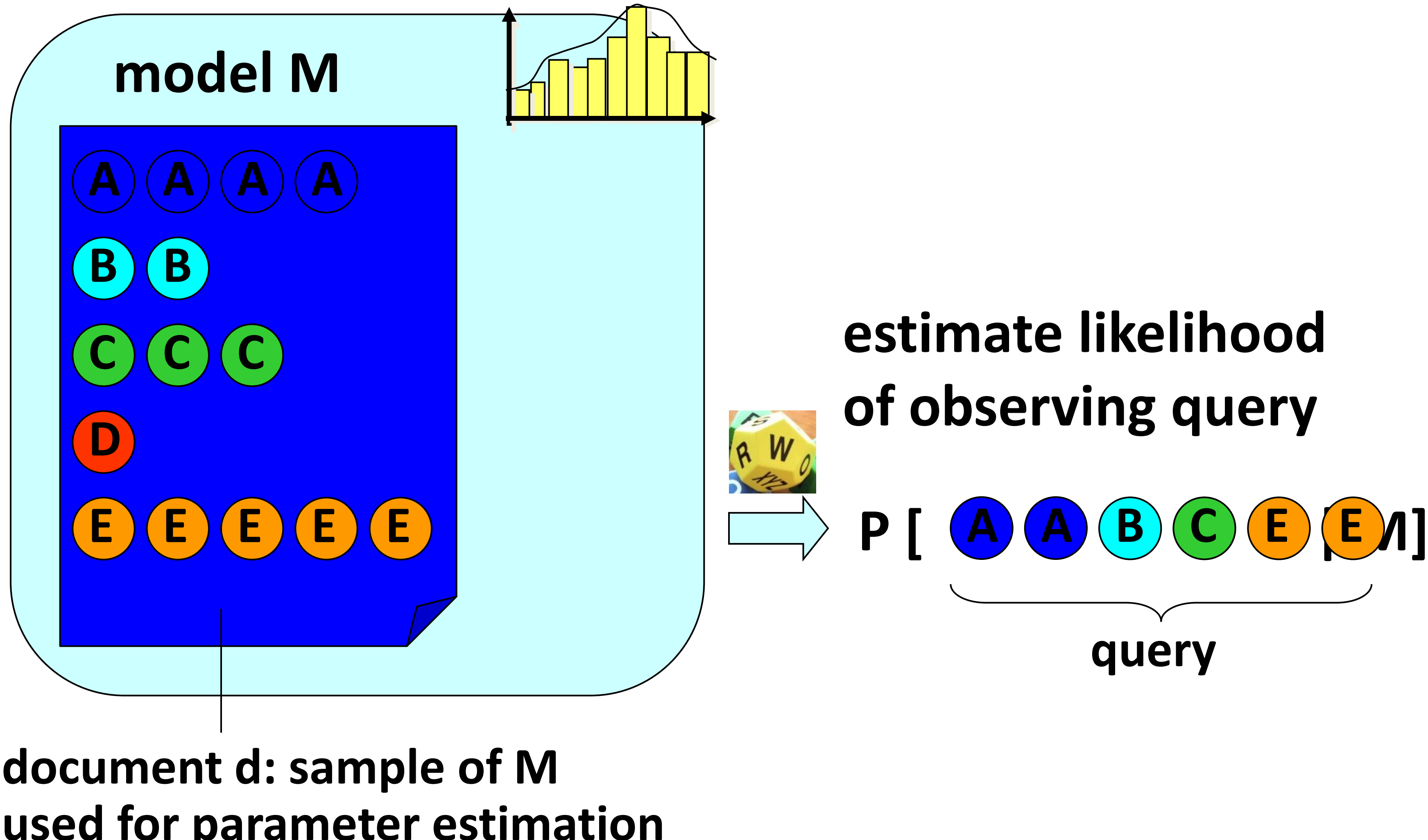
Digression 2: Language Models (LMs)

State-of-the-art model in text retrieval

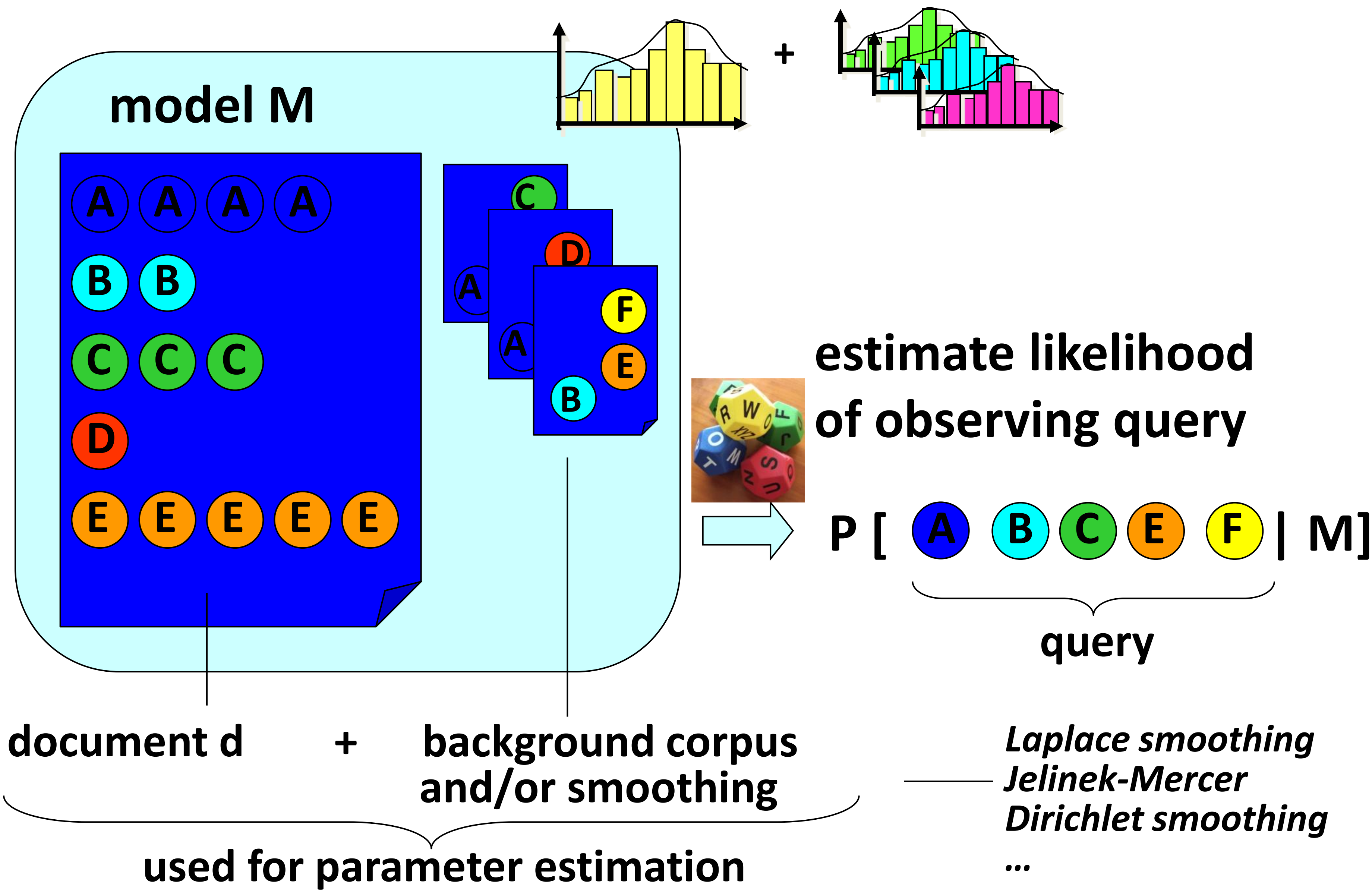


- each **document d_i** has LM: **generative probability distribution** of terms with parameter θ_i
- query q viewed as **sample** from $LM(\theta_1)$, $LM(\theta_2)$, …
- estimate **likelihood P[q | $LM(\theta_i)$]** that q is sample of LM of document d_i (q is „generated by“ d_i)
- rank by **descending likelihoods** (best „explanation“ of q)

Language Models for Text: Example



Language Models for Text: Smoothing



Entity Search with LM Ranking

query: keywords → answer: entities

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

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

query q: ***„French player who won world championship“***

candidate entities:

e1: David Beckham

e2: Ruud van Nistelroy

e3: Ronaldinho

e4: Zinedine Zidane

e5: FC Barcelona

played for ManU, Real, LA Galaxy
David Beckham champions league
England lost match against France
married to spice girl …

Zizou champions league 2002
Real Madrid won final ...
Zinedine Zidane best player
France world cup 1998 ...

weighted by conf.

[Z. Nie et al.: WWW’07]

Outline for Part IV

- Querying Knowledge Bases ✔
- Searching and Ranking Entities ✔
- Searching and Ranking Facts

  General ranking issues

  NAGA-style ranking

  Language Models for facts
- Advanced Query Interfaces

What makes a fact „good"?

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 estimation from:

- frequency in answer
- frequency on 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

How can we implement this?

Confidence:

Prefer results that are likely correct
- accuracy of info extraction
- trust in sources
  (authenticity, authority)

empirical accuracy of IE
PageRank-style estimate of trust
combine into:
max { accuracy (f,s) * trust(s) |
s ∈ witnesses(f) }

Informativeness:

Prefer results with salient facts
Statistical estimation from:
- frequency in answer
- frequency on Web
- frequency in query log

PageRank-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

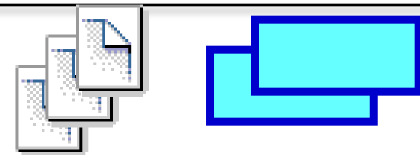
Conciseness:

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., …]

LMs: From Entities to Facts

Document / Entity LM's



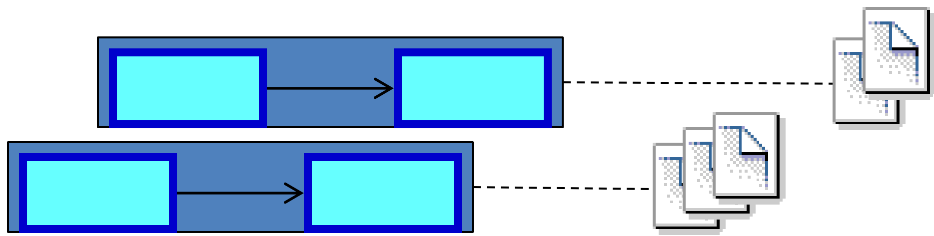
LM for doc/entity: prob. distr. of words

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 !***

LMs for Triples and Triple Patterns

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 FCBarcelona
Zidane p RealMadrid
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))

LMs for Composite Queries

q: Select ?x,?c Where {?x bornIn France . ?x playsFor ?c . ?c in UK . }

P [Henry bI F,
Henry p Arsenal,

$\sim \frac{2}{6}$

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

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

queries q with subqueries $q_1 \dots q_n$

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

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

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

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

- f21: Zidane bI F 200
- f22: Tidjani bI F 20
- f23: Henry bI F 200
- f24: Ribery bI F 200
- f25: Drogba bI F 30
- f26: Drogba bI IC 100
- F27: Zidane bI ALG 50

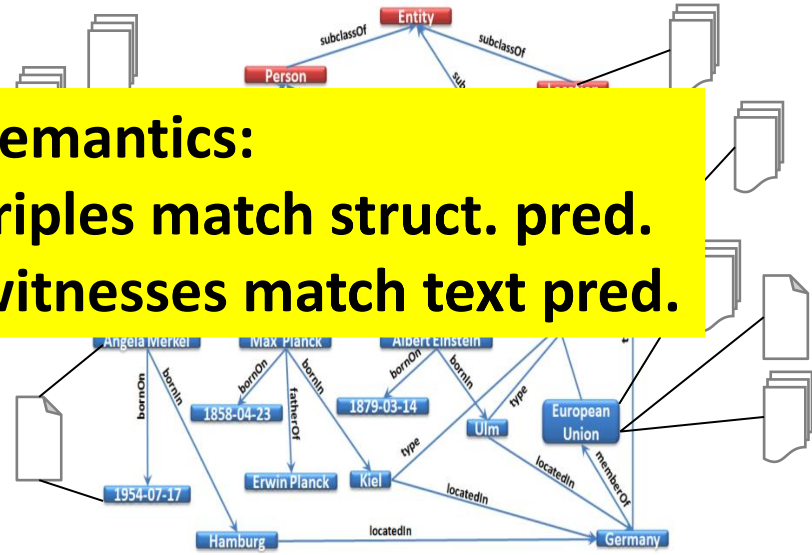
- f1: Beckham p ManU 200
- f7: Zidane p ASCannes 20
- f8: Zidane p Juventus 200
- f9: Zidane p RealMadrid 300
- f10: Tidjani p ASCannes 10
- f12: Henry p Arsenal 200
- f13: Henry p FCBarca 150
- f14: Ribery p Bayern 100
- f15: Drogba p Chelsea 150

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

Extensions: Keywords

Problem: not everything is triplified

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

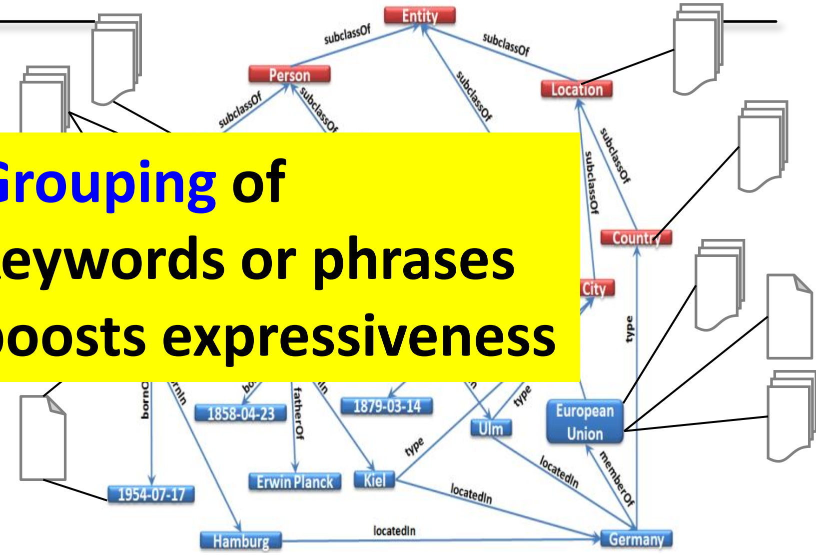


Semantics:
triples match struct. pred.
witnesses match text pred.

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] . }

Extensions: Keywords



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

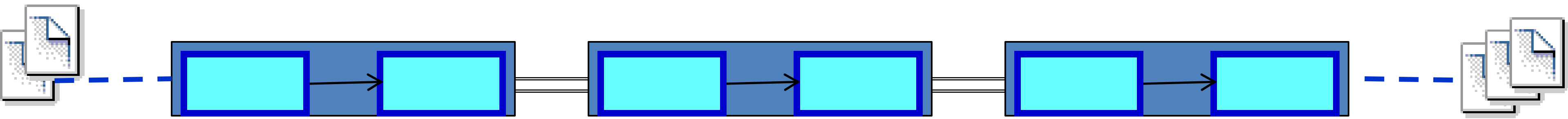
CS researchers whose advisors worked on the Manhattan project?

```
Select ?r, ?a Where {
?r instOf researcher ["computer science"] .
?a workedOn ?x ["Manhattan project"] .
?r hasAdvisor ?a . }
```

```
?r ?p1 ?o1 ["computer science"] .
?a ?p2 ?o2 ["Manhattan project"] .
?r ?p3 ?a . }
```

LMs 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 $w_1 \ldots w_m$

results are still n-tuples of triples t_i

$LM(q_i)$: $P[\text{triple } t_i \mid w_1 \ldots w_m] = \prod_k \beta P[t_i \mid w_k] + (1-\beta) P[t_i]$

LM(answer f_i) analogous

$KL(LM(q) \mid LM(\text{answer } f_i)) = \sum_i KL(LM(q_i) \mid LM(f_i))$

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

Extensions: Query Relaxation

$q^{(4)}$: … Where {?x bornIn IC . ?x p ?c . ?c in UK . }}

[Zidane bI F,
Zidane p Real,

[Drogba bI IC,
Drogba p Chelsea,

[Drogba resOf F,
Drogba p Chelsea,

[Drogba bI IC,
Drogba p Chelsea,
Chelsea in UK]

$LM(q^*) = \lambda\, LM(q) + \lambda_1\, LM(q^{(1)}) + \lambda_2\, LM(q^{(2)}) + \ldots$

replace e in q by $e^{(i)}$ in $q^{(i)}$:
precompute P:=LM (e ?p ?o)
and Q:=LM ($e^{(i)}$?p ?o)
set $\lambda_i \sim 1/2\,(KL(P|Q) + KL(Q|P))$

replace r in q by $r^{(i)}$ in $q^{(i)}$ $\rightarrow$ LM (?s $r^{(i)}$?o)
replace e in q by ?x in $q^{(i)}$ $\rightarrow$ LM (?x r ?o)
…

f21: Zidane bI F	200
f22: Tidjani bI F	20
F23: Henry bI F	200
F24: Ribery bI F	200
F26: Drogba bI IC	100
F27 Zidane bI ALG	50

f1: Beckham p ManU	200
f7: Zidane p ASCannes	20
f9: Zidane p Real	300
f10: Tidjani p ASCannes	10
f12: Henry p Arsenal	200
f15: Drogba p Chelsea	150

…	…00
f32: Arsenal in UK	160
f33: Chelsea in UK	140

LM's of e, r, …
are prob. distr.'s
of triples !

Extensions: Diversification

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 $f_1 \ldots f_k$ 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

Outline for Part IV

- Querying Knowledge Bases ✔
- Searching and Ranking Entities ✔
- Searching and Ranking Facts ✔
- Advanced Query Interfaces

  Natural Language Queries

  Incremental Query Construction

  Visual Query Interfaces

What we have seen so far

Two main paradigms for querying:

- **Keywords** for entity search:
  very easy to use, but not very powerful
- **Structured languages** (SPARQL):
  usable by experts only, but very powerful

Need for more powerful paradigms below the complexity of SPARQL

Other Query Interfaces

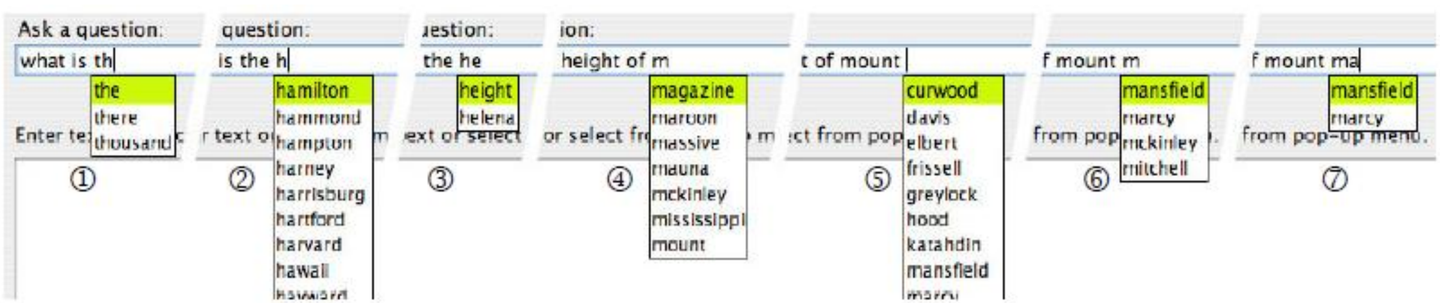
- Domain-specific form-based interfaces
- Natural language queries, QA
- Incremental Query construction
  - Faceted Browsing
  - Active Learning
- Visual SPARQL Query Builders

Natural Language Queries

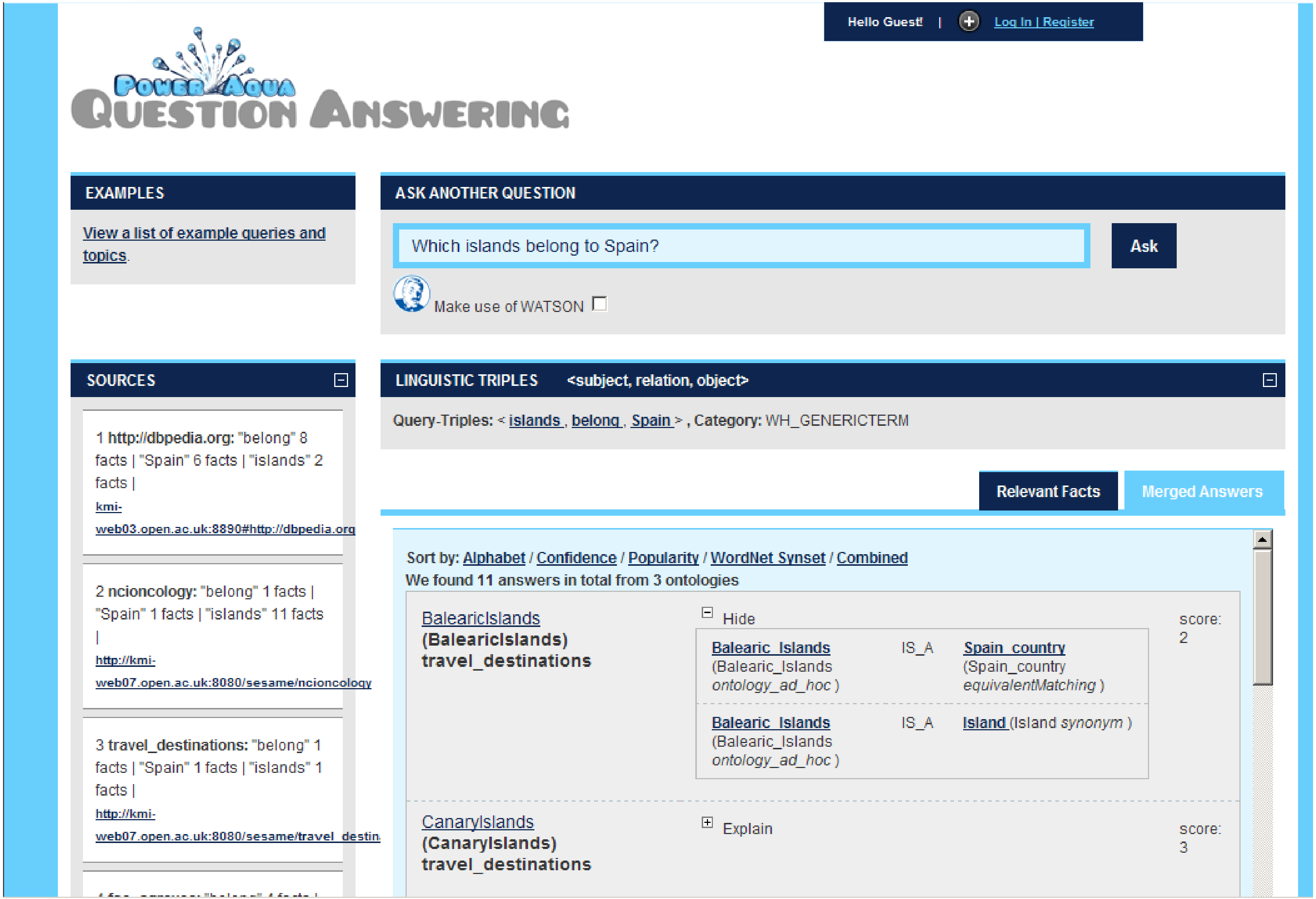
Paradigm:

Allow queries in plain English

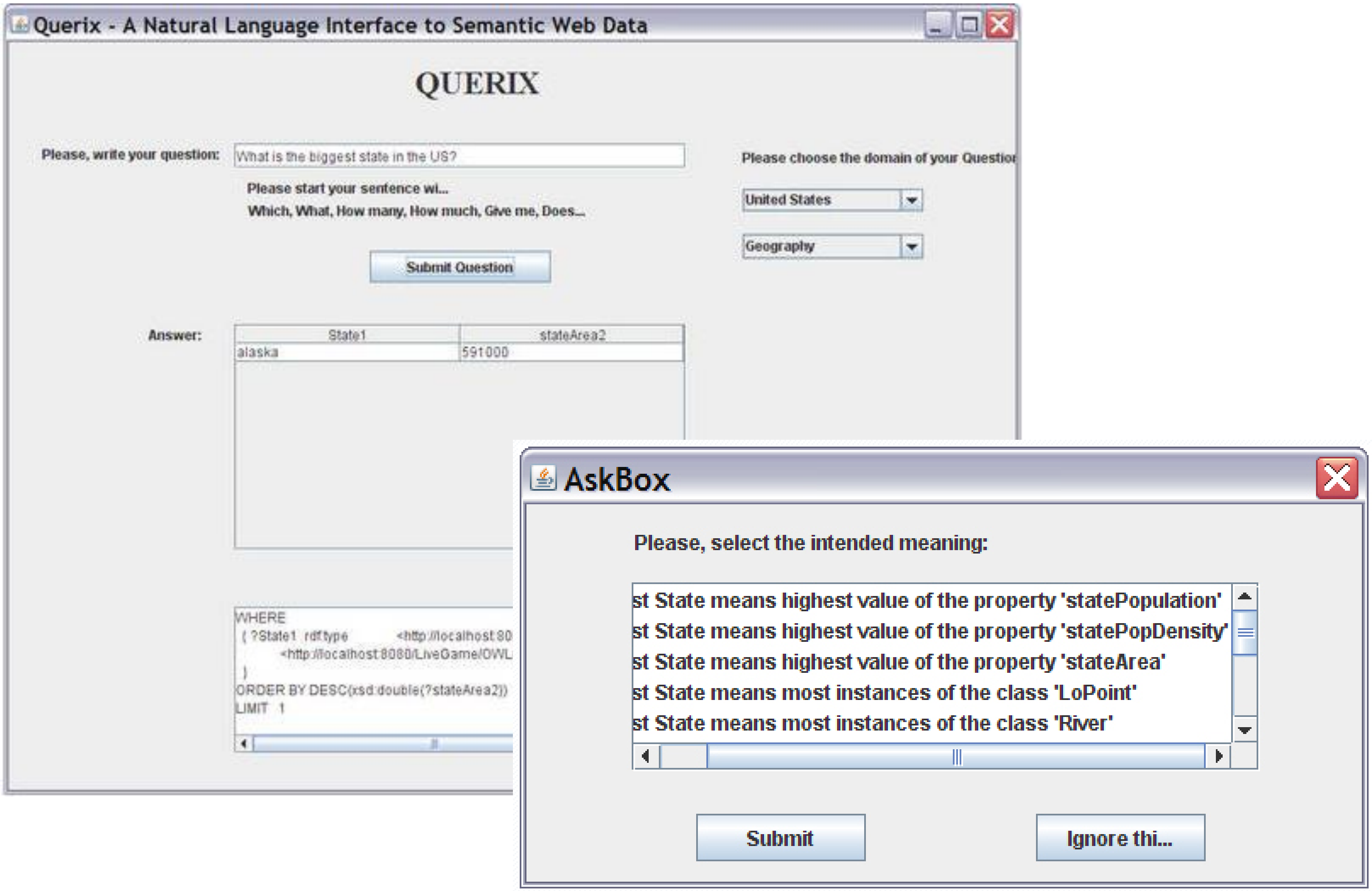
- Map (groups of) keywords to triple patterns, based on existing triples: [NLP-Reduce, PowerAqua]
  "Find a restaurant that is in Barcelona"
  ⇒ `?r isIn Barcelona. ?r isA restaurant`
- Extract query skeleton from syntax tree, heuristic match to known patterns: [Querix]
  "What is a restaurant in Barcelona"
  ⇒ 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-semantic-web/querix/

Natural Language Queries

Pros:

- Intuitive to use
- No schema knowledge necessary

Cons:

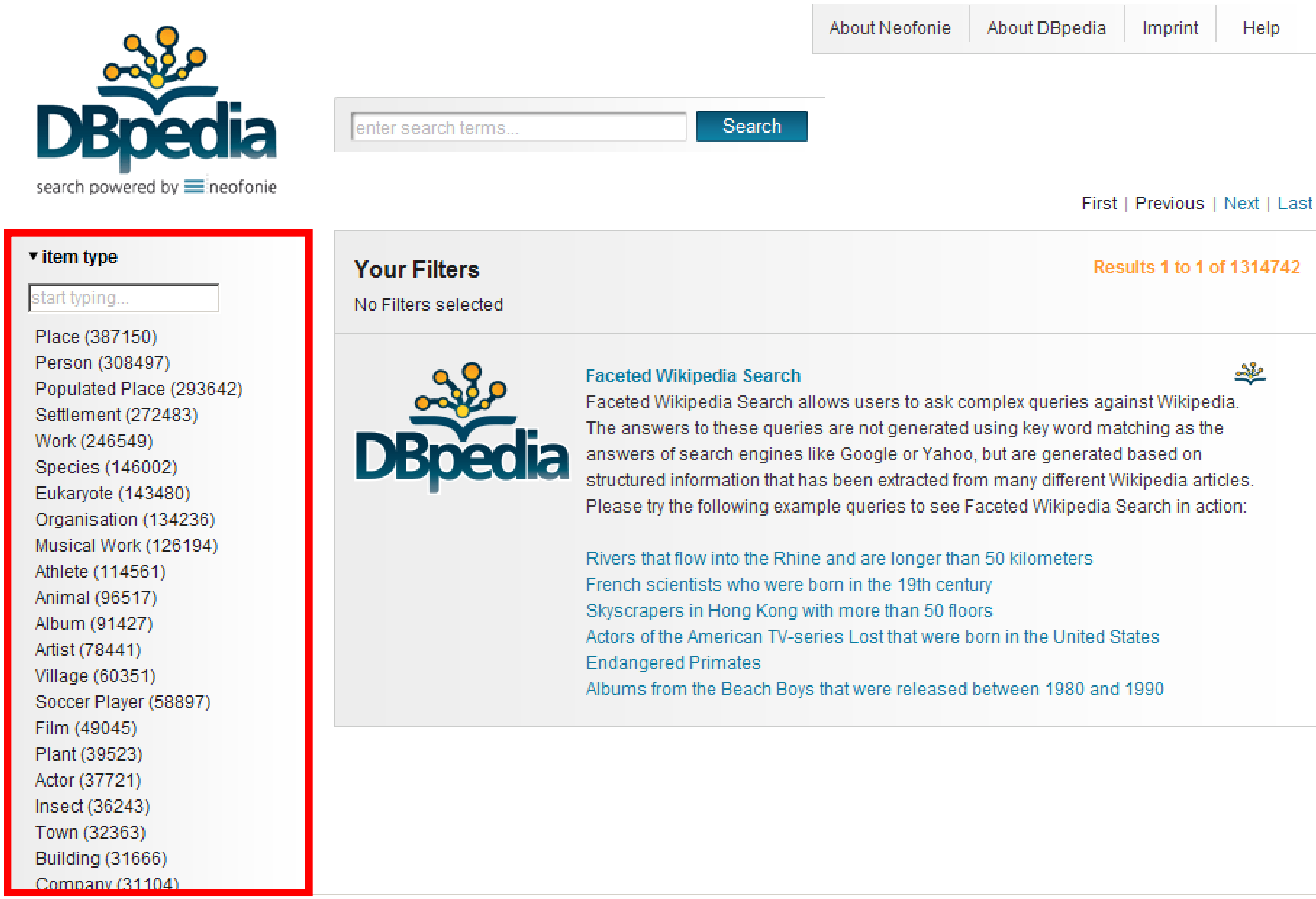
- Often domain-specific
- Finding good query formulation often hard
- Result quality often poor

Faceted Search

Paradigm:

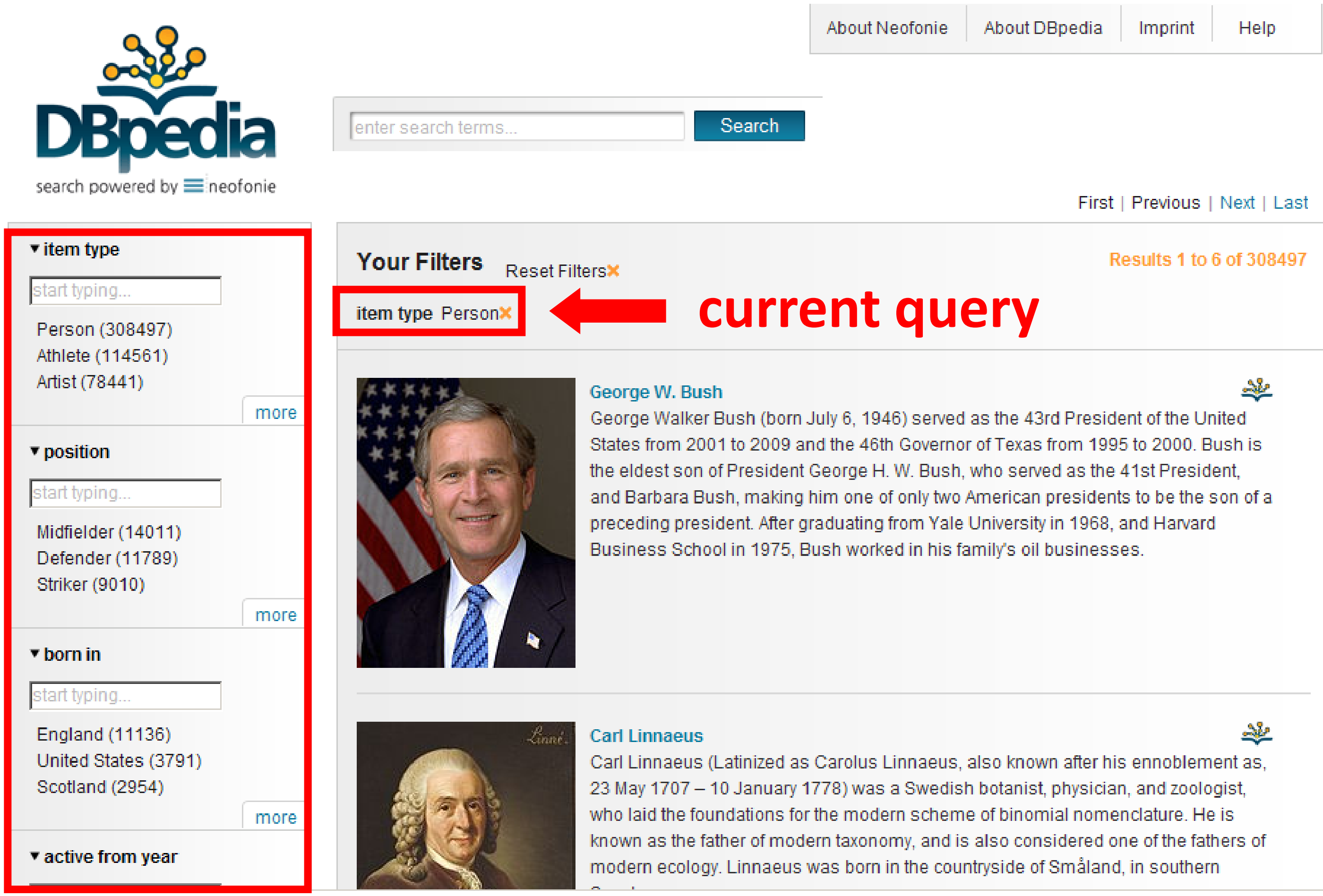
- Incremental refinement of entity-level query
- **Facets**: common properties of many results of (current) query with potential to reduce number of results

Faceted Search: http://dpbedia.neofonie.de/



Initial selection of entity type (candidates sorted by frequency)

Faceted Search: http://dpbedia.neofonie.de/



current query

Best results for current query

Possible refinements („facets“): propert-value pairs

Current SPARQL query:
SELECT ?x WHERE
{?x isA Person.}

Faceted Search: http://dpbedia.neofonie.de/

DBpedia
search powered by neofonie

About Neofonie | About DBpedia | Imprint | Help

enter search terms... Search

First | Previous | Next | Last

▾ item type

start typing...

Person (308497)
Athlete (114561)
Artist (78441)
Soccer Player (58897)
Actor (37721)
Musical Artist (25579)
Politician (19226)
Office Holder (16483)
Gridiron Football Player (15231)
Military Person (13432)
Baseball Player (11136)
American Football Player (8937)
Writer (8719)
Soccer Manager (8485)
Scientist (8465)
Ice Hockey Player (6302)
Cricketer (6082)
Cleric (5351)
British Royalty (4084)
Member Of Parliament (3939)
College Coach (3494)
Rugby Player (3022)

Your Filters Reset Filters✖

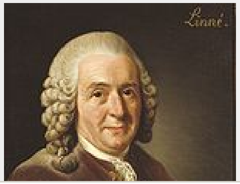
Results 1 to 6 of 308497

item type Person✖

current query

George W. Bush

George Walker Bush (born July 6, 1946) served as the 43rd President of the United States from 2001 to 2009 and the 46th Governor of Texas from 1995 to 2000. Bush is the eldest son of President George H. W. Bush, who served as the 41st President, and Barbara Bush, making him one of only two American presidents to be the son of a preceding president. After graduating from Yale University in 1968, and Harvard Business School in 1975, Bush worked in his family's oil businesses.



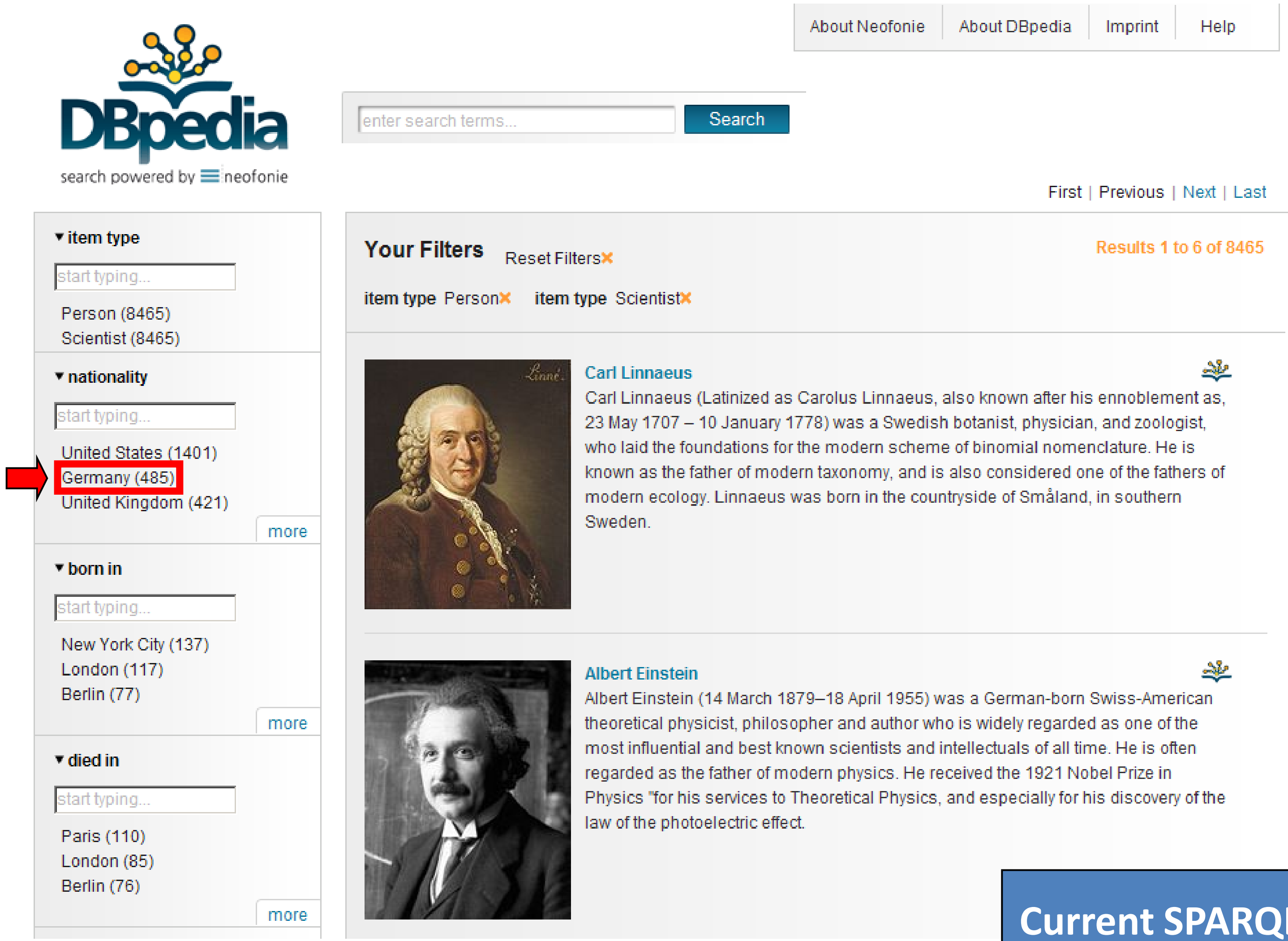
Carl Linnaeus

Carl Linnaeus (Latinized as Carolus Linnaeus, also known after his ennoblement as, 23 May 1707 – 10 January 1778) was a Swedish botanist, physician, and zoologist, who laid the foundations for the modern scheme of binomial nomenclature. He is known as the father of modern taxonomy, and is also considered one of the fathers of modern ecology. Linnaeus was born in the countryside of Småland, in southern

Best results for current query

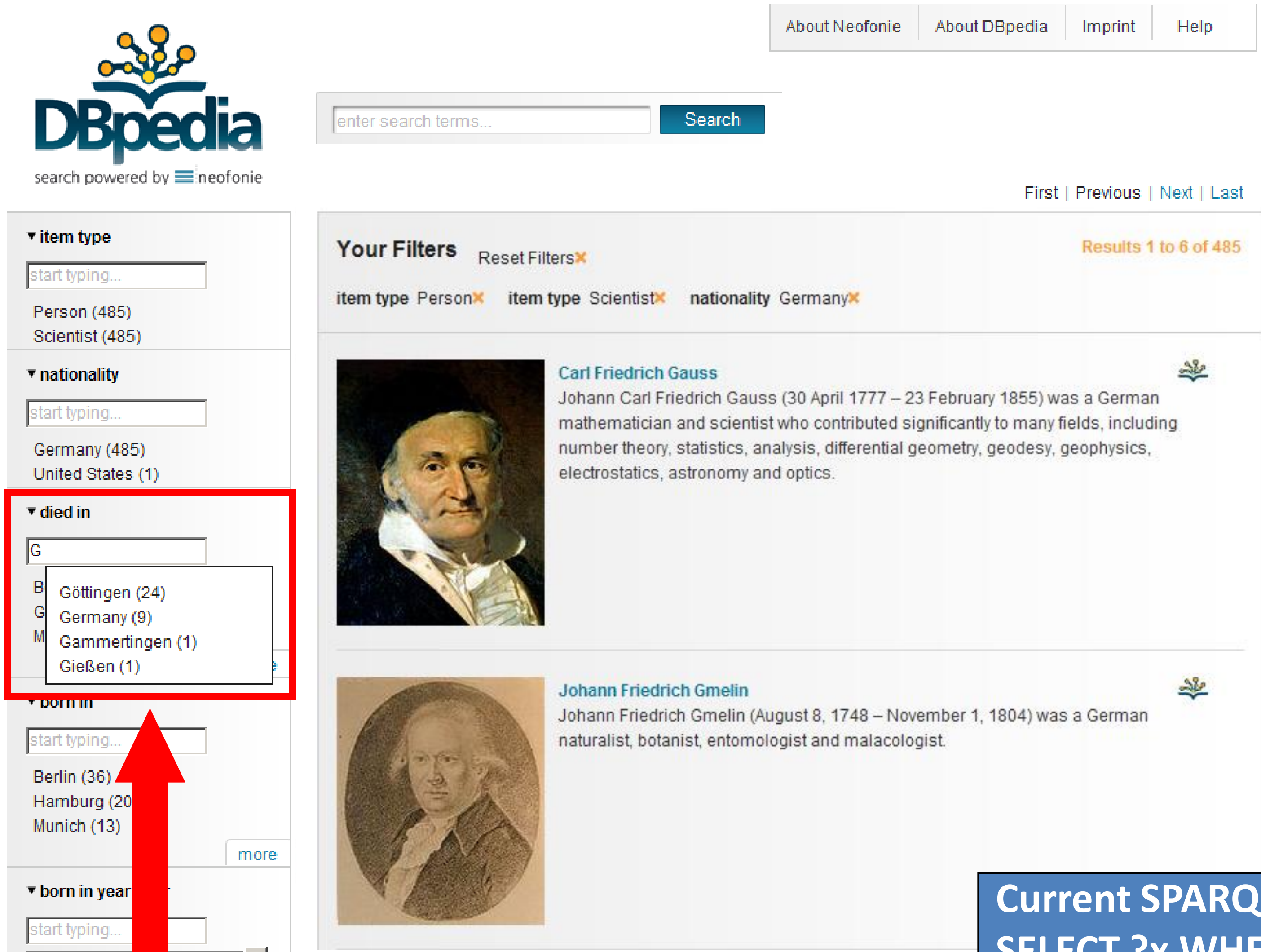
Current SPARQL query:
SELECT ?x WHERE
{?x isA Person.}

Faceted Search: http://dpbedia.neofonie.de/



Current SPARQL query:
SELECT ?x WHERE
{?x isA Person. ?x isA Scientist.}

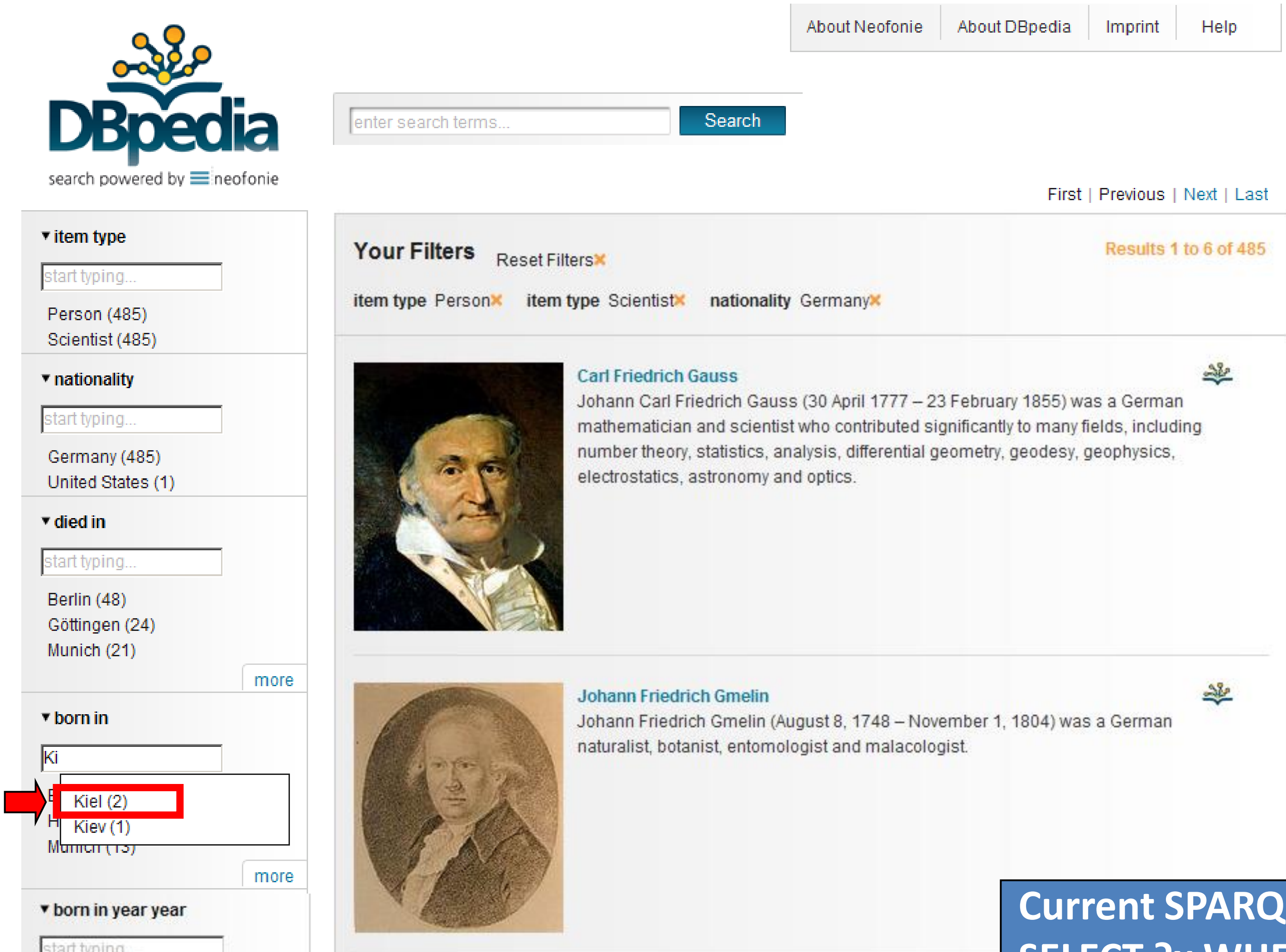
Faceted Search: http://dpbedia.neofonie.de/



Current SPARQL query:
SELECT ?x WHERE
{?x isA Person. ?x isA Scientist.
hasNationality German.}

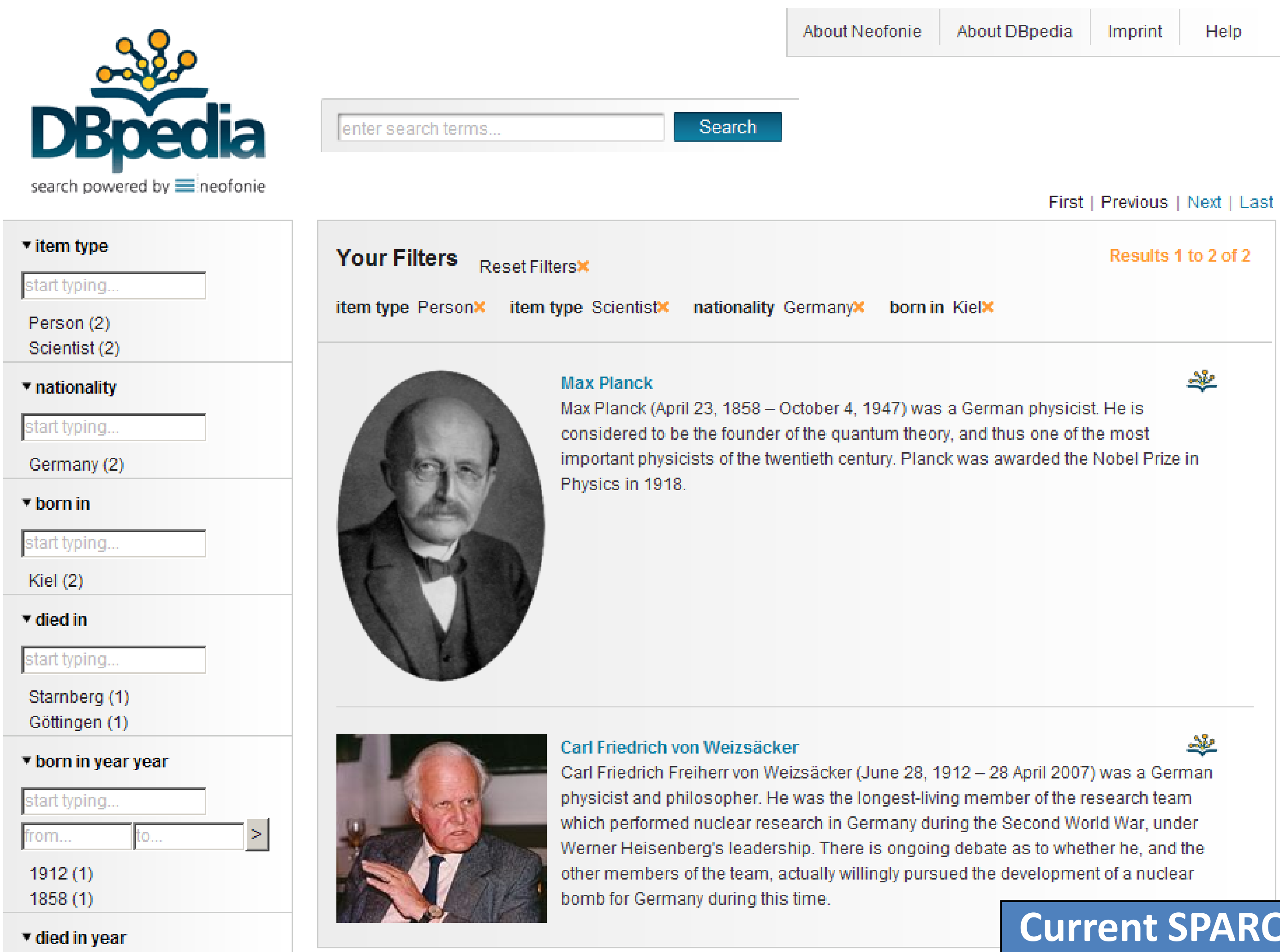
Textual facet values with completion

Faceted Search: http://dpbedia.neofonie.de/



Current SPARQL query:
SELECT ?x WHERE
{?x isA Person. ?x isA Scientist.
hasNationality German.}

Faceted Search: http://dpbedia.neofonie.de/



Current SPARQL query:
SELECT ?x WHERE
{?x isA Person. ?x isA Scientist.
hasNationality German. bornIn Kiel.}

Pros:

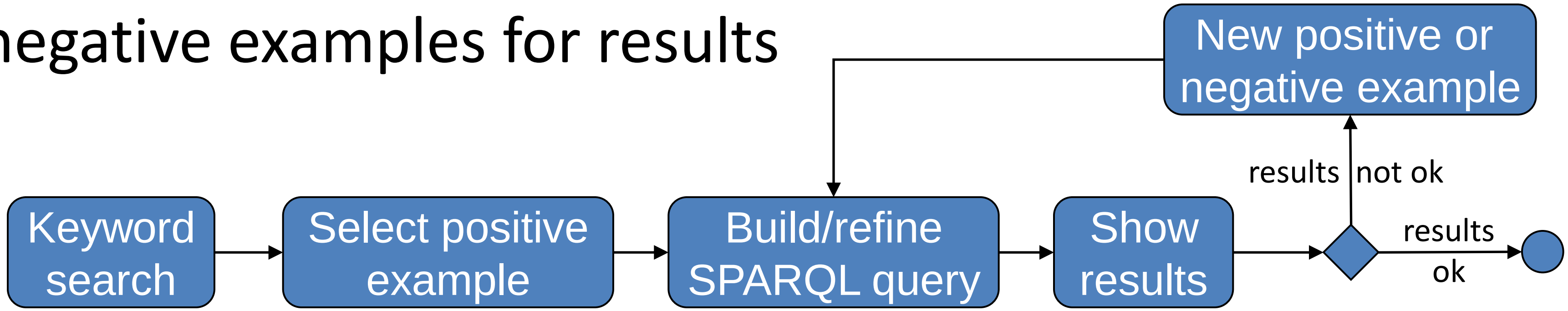
- Intuitive to use
- No schema knowledge necessary
- Quickly leads to results

Cons:

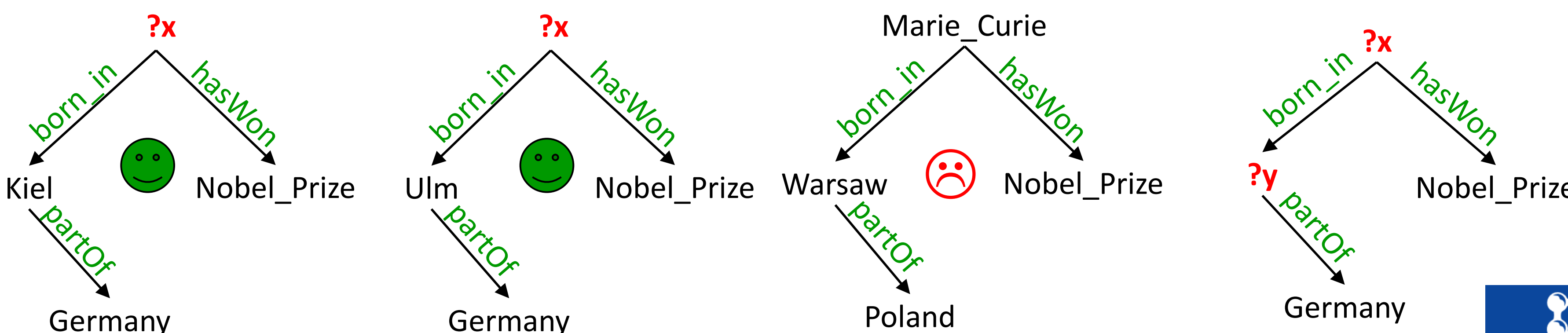
- Only few facets visible at each step
- Required facets sometimes not shown
- Limited to properties of entities, cannot create queries with more than one variable

AutoSPARQL: Learning Queries from Examples

- Goal: Generate SPARQL query from few positive and negative examples for results



- Build query tree from data graph around each example, find minimal subsuming subquery:



http://autosparql.dl-learner.org/
[Lehmann & Bühmann, ESWC2011]

AutoSPARQL Queries made Easy

Active Learning from Examples

Pros:

- Very easy to use, query refinement „on the fly“
- No schema knowledge necessary
- Quickly leads to results

Cons:

- Can require many steps until good query is found
- Limited to entity-centric queries

Visual Query Formulation

Paradigm:

Incremental construction of query by adding and refining constraints in a graphical interface

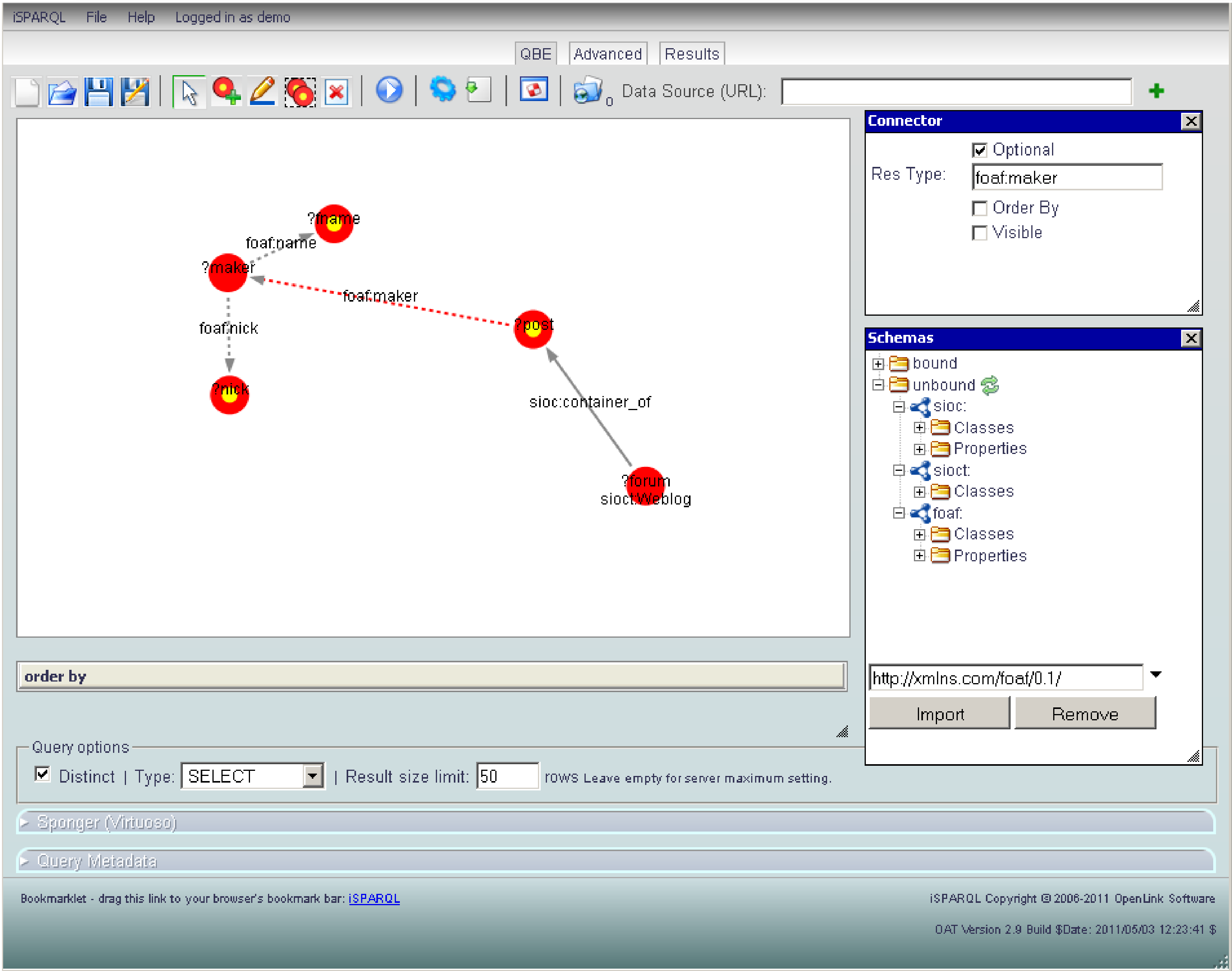
Example Systems:

- iSPARQL, **http://dbpedia.org/isparql/**
- Nitelight [Russell and Smart, ISWC 2008]
- Konduit [Möller et al., ISWC 2008]
- DSpace [Koutsomitropoulos et al., ESWC 2011]

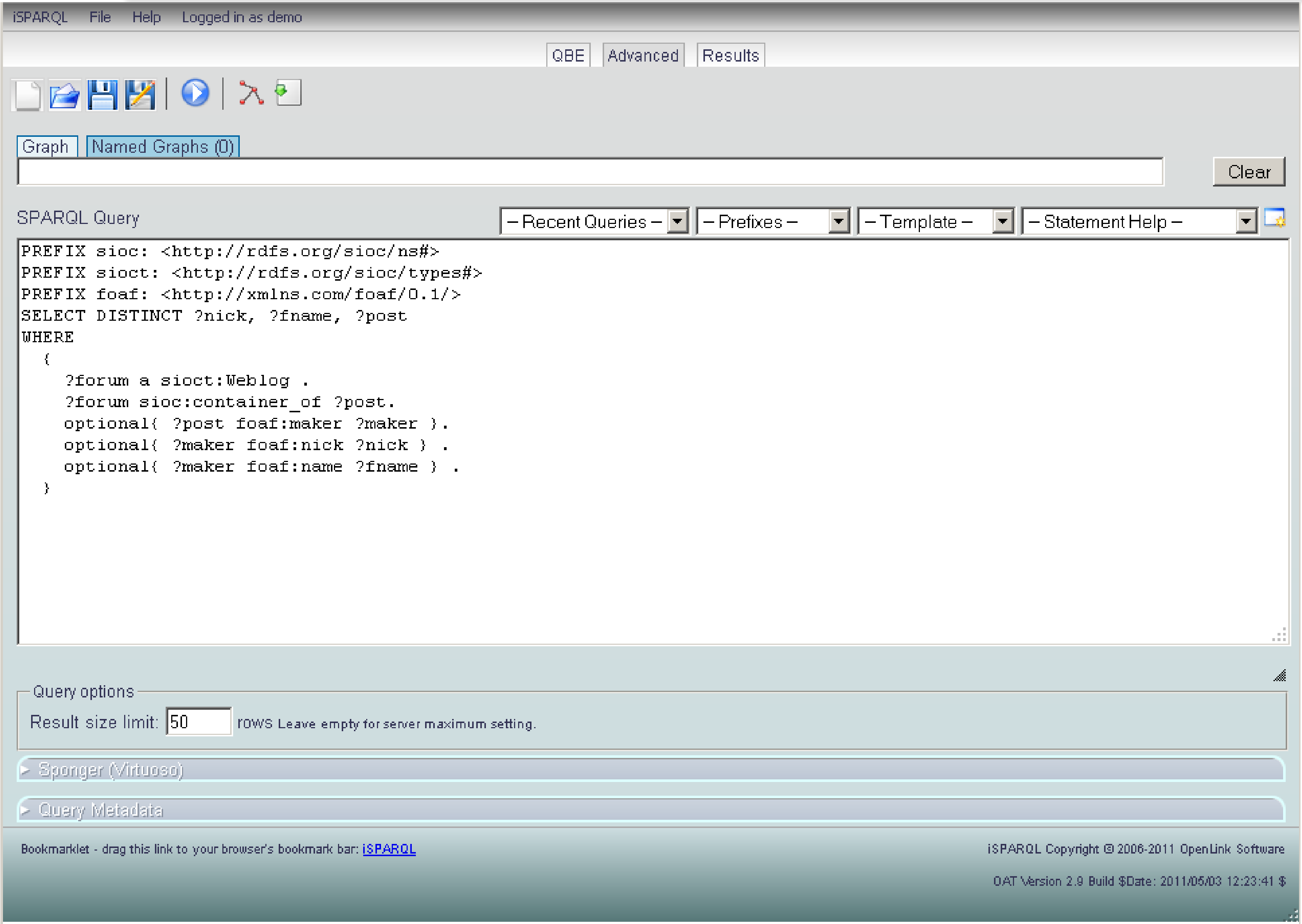
Common features:

point&click, easy access to relations and schema (lists, auto-completion)

iSPARQL, http://dbpedia.org/isparql/



iSPARQL, http://dbpedia.org/isparql/



Visual Query Formulation

Pros:

- Full expressiveness of SPARQL
- Schema knowledge provided by the system
- Leads to very precise queries

Cons:

- Not useful for non-expert users

Which interface is best (for casual users)?

Comparative study 1: [Tran et al., ESWC 2010]

- Not much difference for entity queries
- Faceted Search not very useful when searching for an attribute of an entity
- Users liked Active Learning most

Comparative study 2: [Kaufmann and Bernstein, ISWC 2007]

- Full natural language questions most popular
- Visual Query Builder: fewest steps, longest time, highest failure rate

Outline for Part IV

- Querying Knowledge Bases ✔
- Searching and Ranking Entities ✔
- Searching and Ranking Facts ✔
- Advanced Query Interfaces ✔

Open Problems and Challenges – Part IV

- **Unified ranking** for queries with keywords and structure
- **User Interfaces** for non-experts
  - Support to formulate structured queries
  - General-purpose NLP systems
  - Output of complex results beyond entities

Readings for Part IV

Part IV.1:

- **SPARQL Query Language for RDF**, W3C Recommendation, 15 January 2008, http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/
- **SPARQL New Features and Rationale**, W3C Working Draft, 2 July 2009, http://www.w3.org/TR/2009/WD-sparql-features-20090702/
- **SPARQL 1.1 Query Language**, W3C Working Draft, 12 May 2011, http://www.w3.org/TR/2011/WD-sparql11-query-20110512/
- **SPARQL 1.1 Federation Extensions**, W3C Working Draft, 1 June 2010, http://www.w3.org/TR/2010/WD-sparql11-federated-query-20100601/
- Kemafor Anyanwu, Angela Maduko, Amit P. Sheth: **SPARQ2L: towards support for subgraph extraction queries in RDF databases**. WWW Conference, 2007
- Gaurav Bhalotia, Arvind Hulgeri, Charuta Nakhe, Soumen Chakrabarti, S. Sudarshan: **Keyword Searching and Browsing in Databases using BANKS**. ICDE, 2002
- Gjergji Kasneci, Fabian M. Suchanek, Georgiana Ifrim, Maya Ramanath, Gerhard Weikum: **NAGA: Searching and Ranking Knowledge**. ICDE, 2008
- Thomas Neumann, Gerhard Weikum: **The RDF-3X engine for scalable management of RDF data**. VLDB Journal 19(1), 2010
- Nicoleta Preda, Gjergji Kasneci, Fabian M. Suchanek, Thomas Neumann, Wenjun Yuan, Gerhard Weikum: **Active knowledge: dynamically enriching RDF knowledge bases by web services**. SIGMOD Conference, 2010

Part IV.2:

- Sergey Brin, Lawrence Page: **The Anatomy of a Large-Scale Hypertextual Web Search Engine**. Computer Networks 30(1-7), 1998
- Andrey Balmin, Vagelis Hristidis, Yannis Papakonstantinou: **ObjectRank: Authority-Based Keyword Search in Databases**. VLDB, 2004
- Soumen Chakrabarti: **Dynamic personalized pagerank in entity-relation graphs**. WWW Conference, 2007
- Thomas Franz, Antje Schulz, Sergej Sizov, Steffen Staab: **TripleRank: Ranking Semantic Web Data by Tensor Decomposition**. ISWC, 2009
- Tao Cheng , Xifeng Yan , Kevin Chen-Chuan Chang: **EntityRank: searching entities directly and holistically**. VLDB, 2007
- Krisztian Balog, Edgar Meij, Maarten de Rijke: **Entity Search: Building Bridges between Two Worlds**. WWW, 2010
- Zaiqing Nie, Yunxiao Ma, Shuming Shi, Ji-Rong Wen, Wei-Ying Ma: **Web object retrieval**. WWW Conference, 2007
- Desislava Petkova, W. Bruce Croft: **Hierarchical Language Models for Expert Finding in Enterprise Corpora**. ICTAI, 2006
- Pavel Serdyukov, Djoerd Hiemstra: **Modeling Documents as Mixtures of Persons for Expert Finding**. ECIR, 2008
- ChengXiang Zhai: **Statistical Language Models for Information Retrieval**. Morgan & Claypool Publishers, 2008
- Djoerd Hiemstra: **Language Models**. Encyclopedia of Database Systems, 2009

Part IV.3:

- Shady Elbassuoni, Maya Ramanath, Ralf Schenkel, Marcin Sydow, Gerhard Weikum: **Language-model-based ranking for queries on RDF-graphs**. CIKM, 2009
- Shady Elbassuoni, Maya Ramanath, Gerhard Weikum: **Query Relaxation for Entity-Relationship Search**. ESWC, 2011
- Vagelis Hristidis, Heasoo Hwang, Yannis Papakonstantinou: **Authority-based keyword search in databases**. ACM Transactions on Database Systems 33(1), 2008
- Gjergji Kasneci, Maya Ramanath, Mauro Sozio, Fabian M. Suchanek, Gerhard Weikum: **STAR: Steiner-Tree Approximation in Relationship Graphs**. ICDE, 2009
- Alexandra Poulovassilis and Peter T. Wood: **Combining Approximation and Relaxation in Semantic Web Path Queries**. ISWC, 2010

Part IV.4:

- Esther Kaufmann, Abraham Bernstein, Lorenz Fischer: **NLP-Reduce: A "naive" but Domain-independent Natural Language Interface for Querying Ontologies.** ESWC, 2007
- Knud Möller, Oszkar Ambrus, Laura Josan, Siegfried Handschuh: **A Visual Interface for Building SPARQL Queries in Konduit**. ISWC, 2008
- Alistair Russell, Paul R. Smart: **NITELIGHT: A Graphical Editor for SPARQL Queries**. ISWC, 2008
- Jens Lehmann and Lorenz Bühmann: **AutoSPARQL: Let Users Query Your Knowledge Base**. ESWC, 2011
- Thanh Tran, Haofen Wang, Sebastian Rudolph, Philipp Cimiano: **Top-k Exploration of Query Candidates for Efficient Keyword Search on Graph-Shaped (RDF) Data**. ICDE, 2009.
- Esther Kaufmann, Abraham Bernstein: **How useful are natural language interfaces to the Semantic Web for casual end-users**? ESWC, 2007
- Thanh Tran, Tobias Mathäß, Peter Haase: **Usability of Keyword-Drive Schema-Agnostic Search**. ESWC, 2010
- Gideon Zenz et al.: **From keywords to semantic queries - Incremental query construction on the semantic web**. Journal Web Semantics 7(3),2009

Outline for Part V

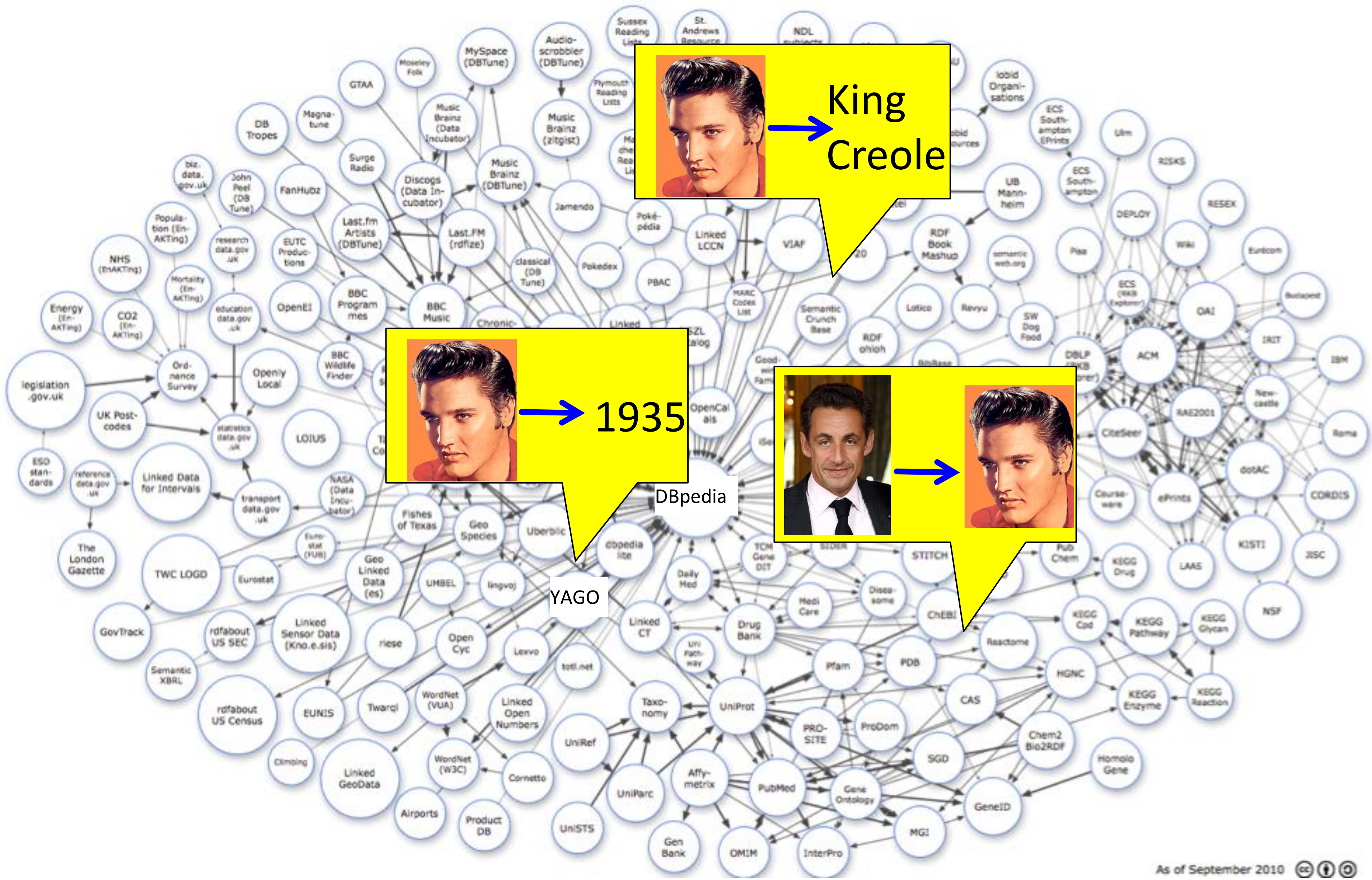
- **URIs & Dereferenceable URIs**
- **Shared Vocabularies**
- **Linked Data**
- **The Semantic Web and the Web**

There's not only DBpedia & YAGO

DBpedia

YAGO

There's a whole Web of Ontologies



URIs

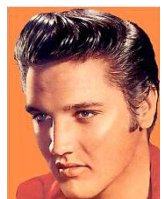
Goal: Identify entities uniquely, worldwide
The same entity can have multiple identifiers,
but the same identifier shall always mean the same entity.

[W3C, URI, 2004]

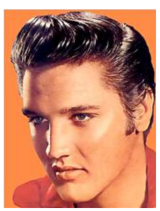
URL-like URIs

A **Uniform Resource Identifier** (URI) is a string of characters used to identify a name or a resource on the Internet

URIs can be like URLs



http://elvis.org/me



http://imitators.org/Elvis/FG17

http://onto.com/people/singers/EP

URL-like URIs

		Age
http://elvis.org/me	Identifies the person, not Internet-accessible	76
http://elvis.org/index.html	Identifies a file, Internet-accessible	5

URL-like URIs

http://imitators.org/Elvis/FG17

World-wide unique mapping to domain owner

in the responsibility of the domain owner

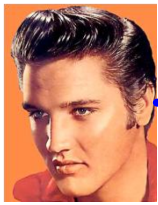
⇒ There should be no URI with two meanings

⇒ People can invent all kinds of URIs

- a company can create URIs to identify its products
- an organization can assign sub-domains and each sub-domain can define URIs
- individual people can create URIs from their homepage
- people can create URIs from any URL for which they have exclusive rights to create URIs

Triples with URIs

Every entity name and relation name is expressed by a URI:



won

Elvis Presley

Grammy Award

http://elvis.org/himself http://inria.fr/rdf/dta#won http://g-a.com/prize

=> Facts become triples of URIs

Namespace Prefixes

A **namespace prefix** is an abbreviation for the prefix of a URI.

@prefix elvis: http://elvis.org/
@prefix inria: http://inria.fr/rdf/dta#
@prefix grammy: http://g-a.com/

http://elvis.org/himself → elvis:himself

http://inria.fr/rdf/dta#won → inria:won

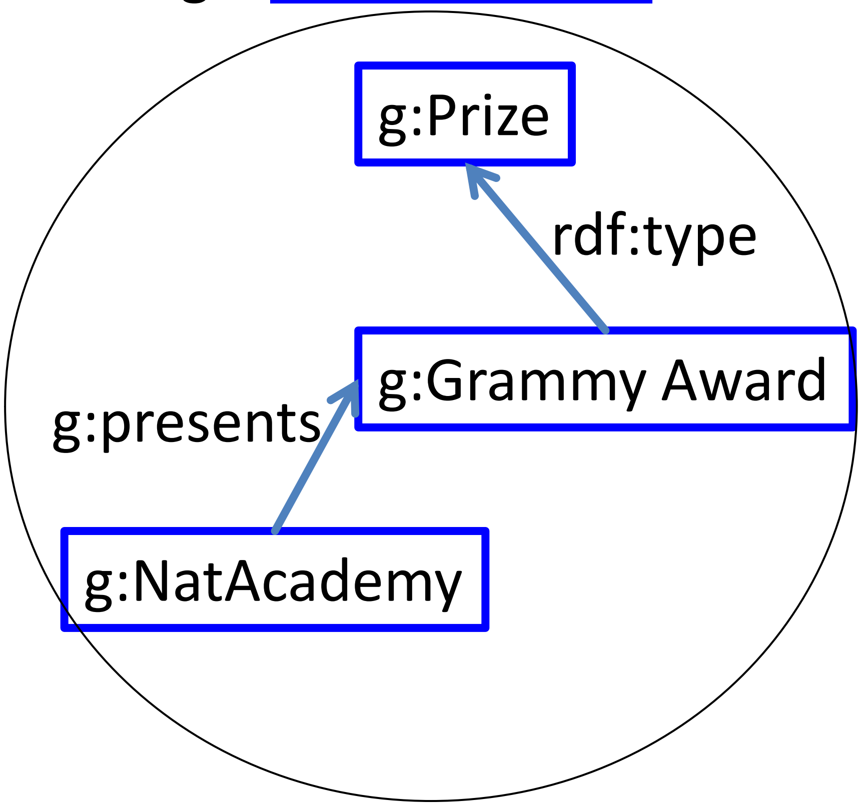
http://g-a.com/prize → grammy:prize

A URI abbreviated this way is called a **qname**.

Storing data

RDF data is usually stored on a server

Namespace
g = http://g-a.com



The server at
http://g-a.com
stores:

@prefix g: http://g-a.com
@prefix rdf: http://www.w3.org/...

g:GrammyAward
rdf:type g:Award

g:NatAcademy
g:presents g:GrammyAward

Try this

Cool URIs

A URI is not necessarily **dereferenceable** (i.e., it cannot be accessed online)

http://g-a.com/GrammyAward => NOT FOUND

... but it *can be* dereferenceable. This means that if I access the URL, the server responds with an RDF snippet:

```
@prefix g:     http://g-a.com
@prefix rdf:   http://www.w3.org/1999/02/22-rdf-syntax-ns#

g:GrammyAward              rdf:type     g:Award
http://elvis.com/elvis     g:won        g:GrammyAward
```

Try this out: rdf:type = http://www.w3.org/1999/02/22-rdf-syntax-ns#type

$\Rightarrow$ URIs can be “clicked” (followed) [W3C, Cool URIs, 2008]

Cool URIs

```
@prefix e:      http://elvis.com
@prefix rdf:    http://www.w3.org/1999/02/22-rdf-syntax-ns#

e:elvis                 rdf:type      e:singer
e:elvis                 e:born        1935
```

Server at http://elvis.com

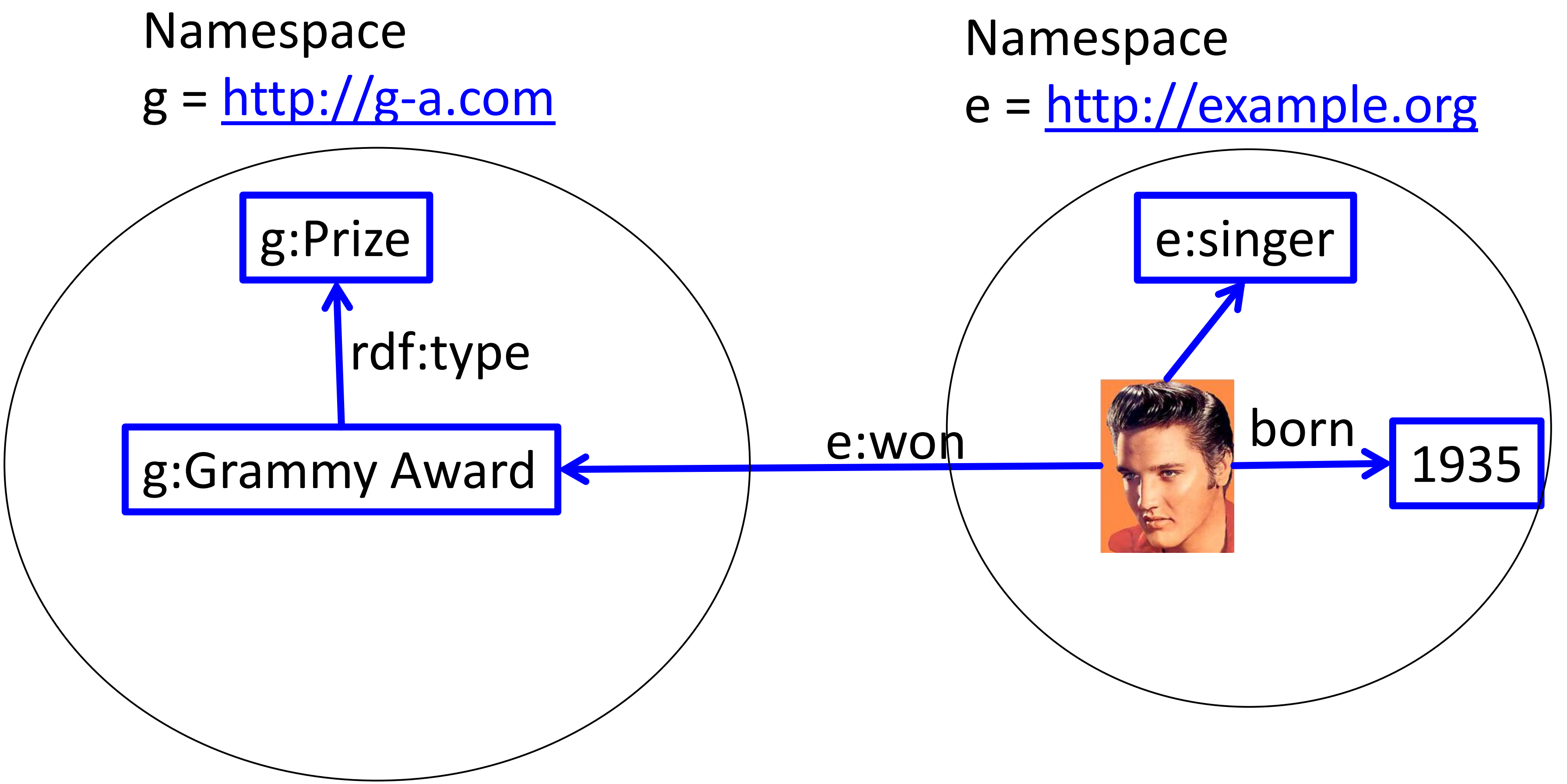
```
@prefix g:      http://g-a.com
@prefix rdf:    http://www.w3.org/1999/02/22-rdf-syntax-ns#

g:GrammyAward                rdf:type      g:Award
http://elvis.com/elvis       g:won         g:GrammyAward
```

Server at http://g-a.com

$\Rightarrow$ The RDF graph becomes traversable

We’re all one Graph

If two RDF graphs share one node, they are actually 1 graph.



A machine can follow the links and retrieve more information in the neighboring ontology.

Outline

- **URIs & Dereferenceable URIs ✔**
- **Shared Vocabularies**
- **Linked Data**
- **The Semantic Web and the Web**

Standard Vocabulary

A number of standard vocabularies have evolved

rdf: The basic RDF vocabulary
http://www.w3.org/1999/02/22-rdf-syntax-ns#

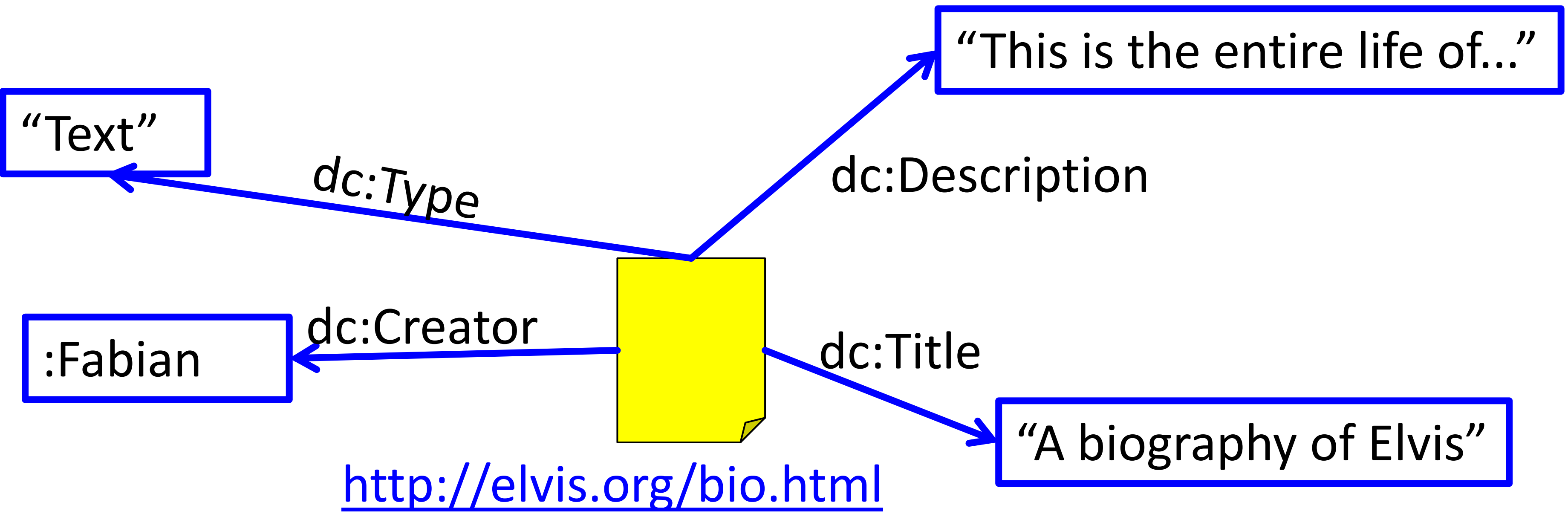
rdfs: RDF Schema vocabulary
http://www.w3.org/1999/02/22-rdf-syntax-ns#

Standard vocabulary provided by the W3C:

- type,
- subclassOf,
- Property,
- Class
- label
- ….

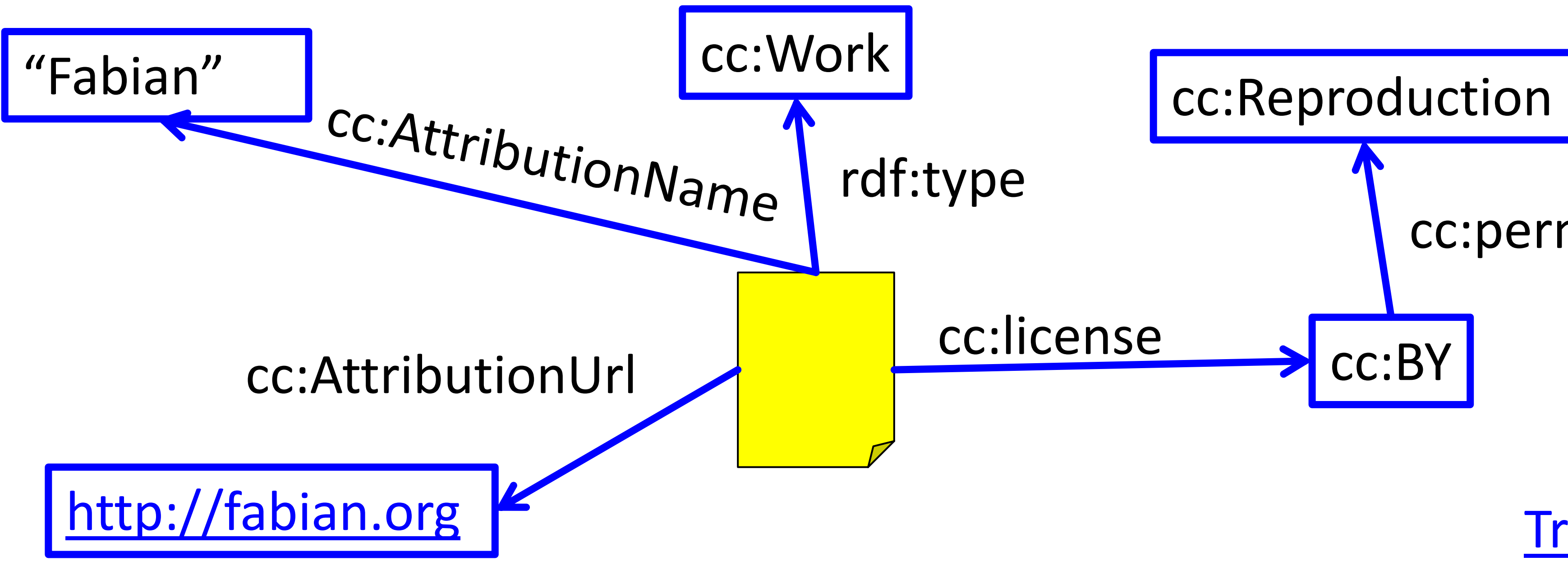
Dublin Core

dc: Dublin Core (predicates for describing documents)
http://purl.org/dc/elements/1.1/



Creative Commons

cc: Creative Commons (types of licences)
http://creativecommons.org/ns#



Try this

Creative Commons defines very popular licenses, notably

- CC-BY: Free for reuse, just give credit to the author
- CC-BY-NC: Free for reuse, give credit, non-commercial use only
- CC-BY-ND: Free for reuse, give credit, do not create derivative works

Schema.org

schema: Defined by Microsoft + Google + Yahoo
for „everything on the Web“, http://schema.org

Thing > Person

A person (alive, dead, undead, or fictional).

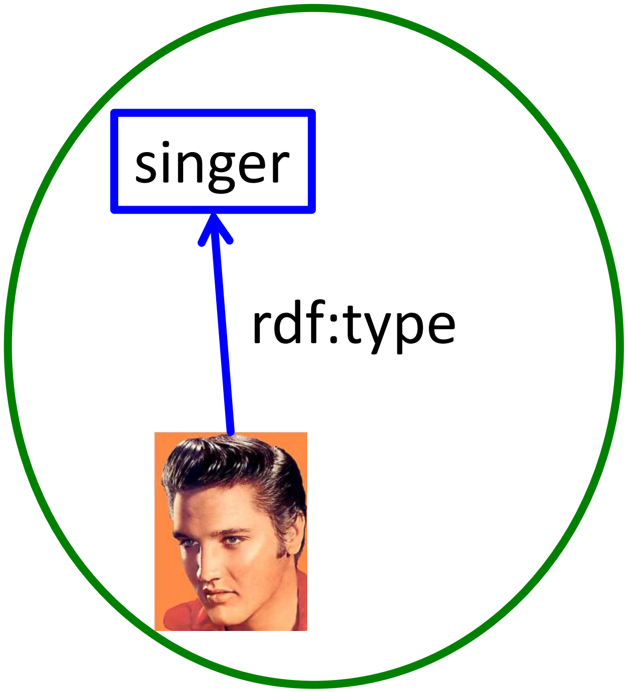
Property	Expected Type	Description
Properties from Thing		
description	Text	A short description of the item.
image	URL	URL of an image of the item.
name	Text	The name of the item.
url	URL	URL of the item.
Properties from Person		
address	PostalAddress	Physical address of the item.
affiliation	Organization	An organization that this person is affiliated with. For example,
alumniOf	EducationalOrganization	An educational organizations that the person is an alumni of.
awards	Text	Awards won by this person or for this creative work.
birthDate	Date	Date of birth.

Outline

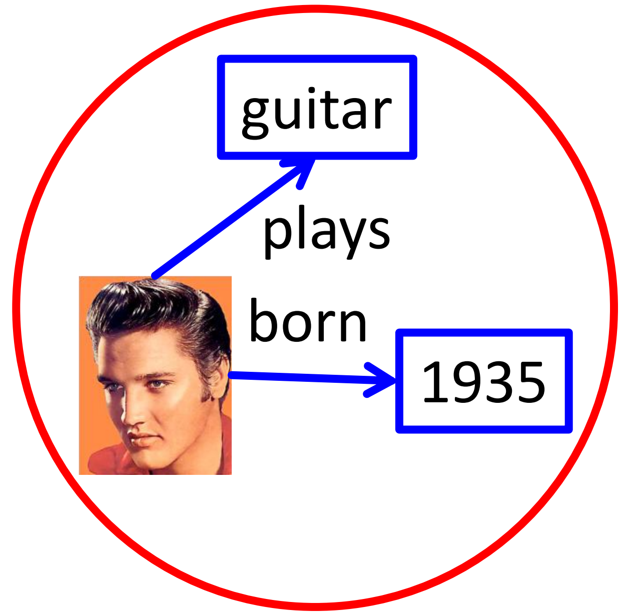
- **URIs & Dereferenceable URIs ✔**
- **Shared Vocabularies ✔**
- **Linked Data**
- **The Semantic Web and the Web**

Linked Data Problem

Many ontologies talk about the same entity with different URIs.



Elvisopedia
(http://elvisopedia.org/)

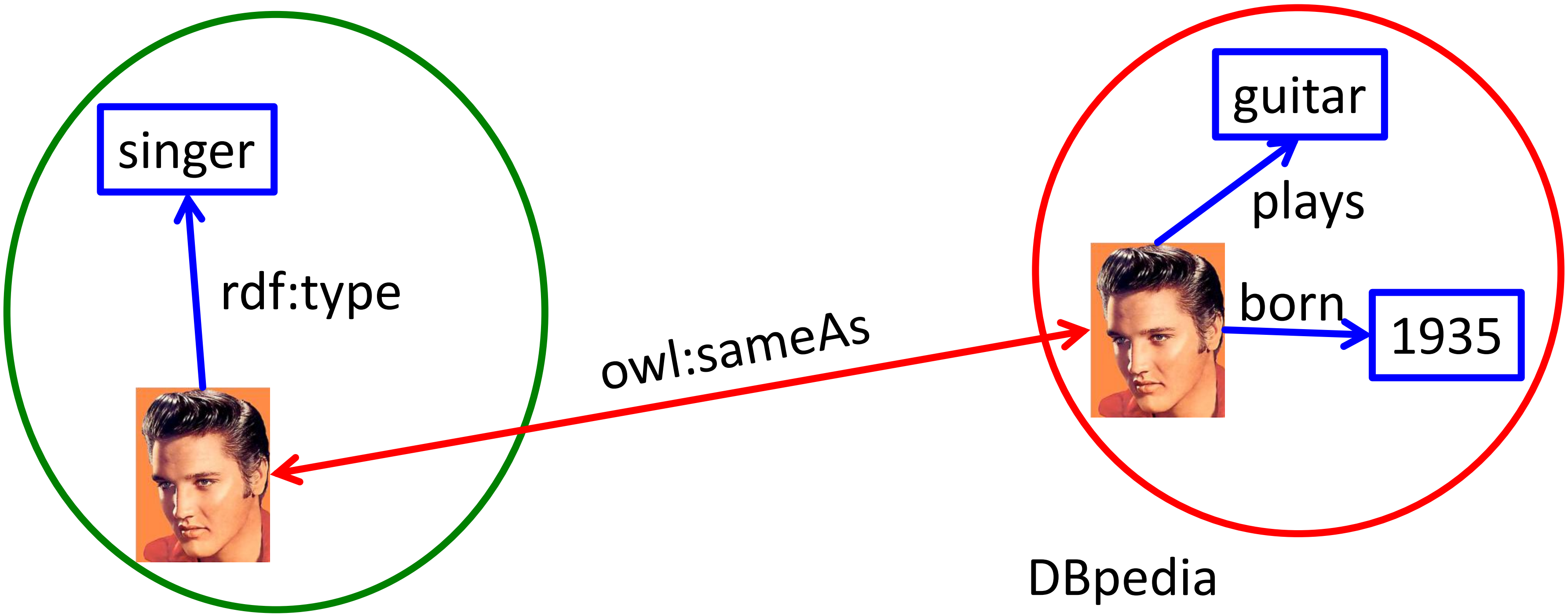


DBpedia
(http://dbpedia.org/)

This is bad, because we cannot join the information.

Linked Data Solution

OWL provides vocabulary to link equivalent entities



DBpedia
(http://dbpedia.org/)

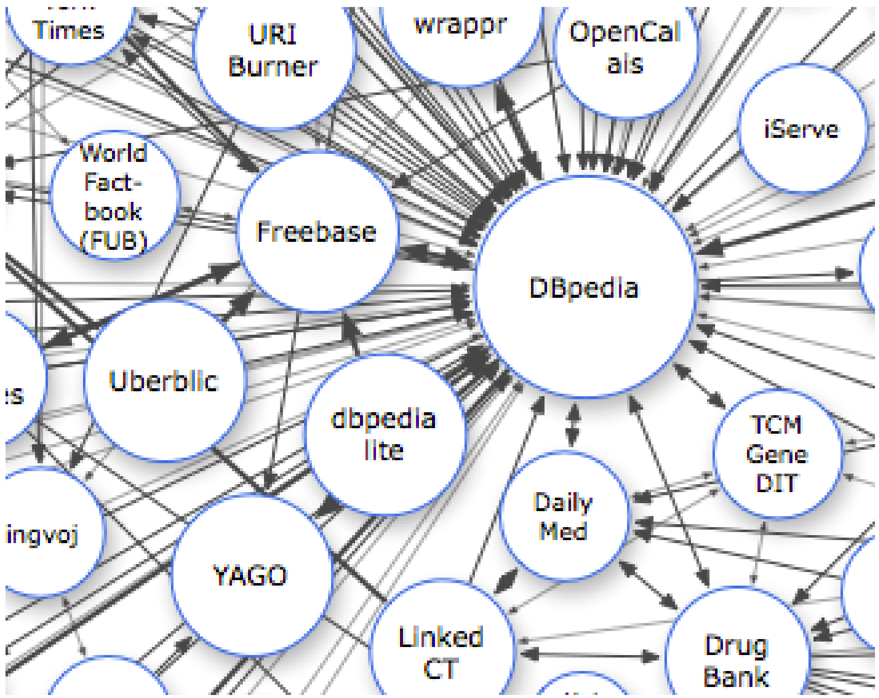
Elvisopedia
(http://elvisopedia.org/)

http://elvisopedia.org/Elvis owl:sameAs http://dpbedia.org/Elvis

The Linking Data Project

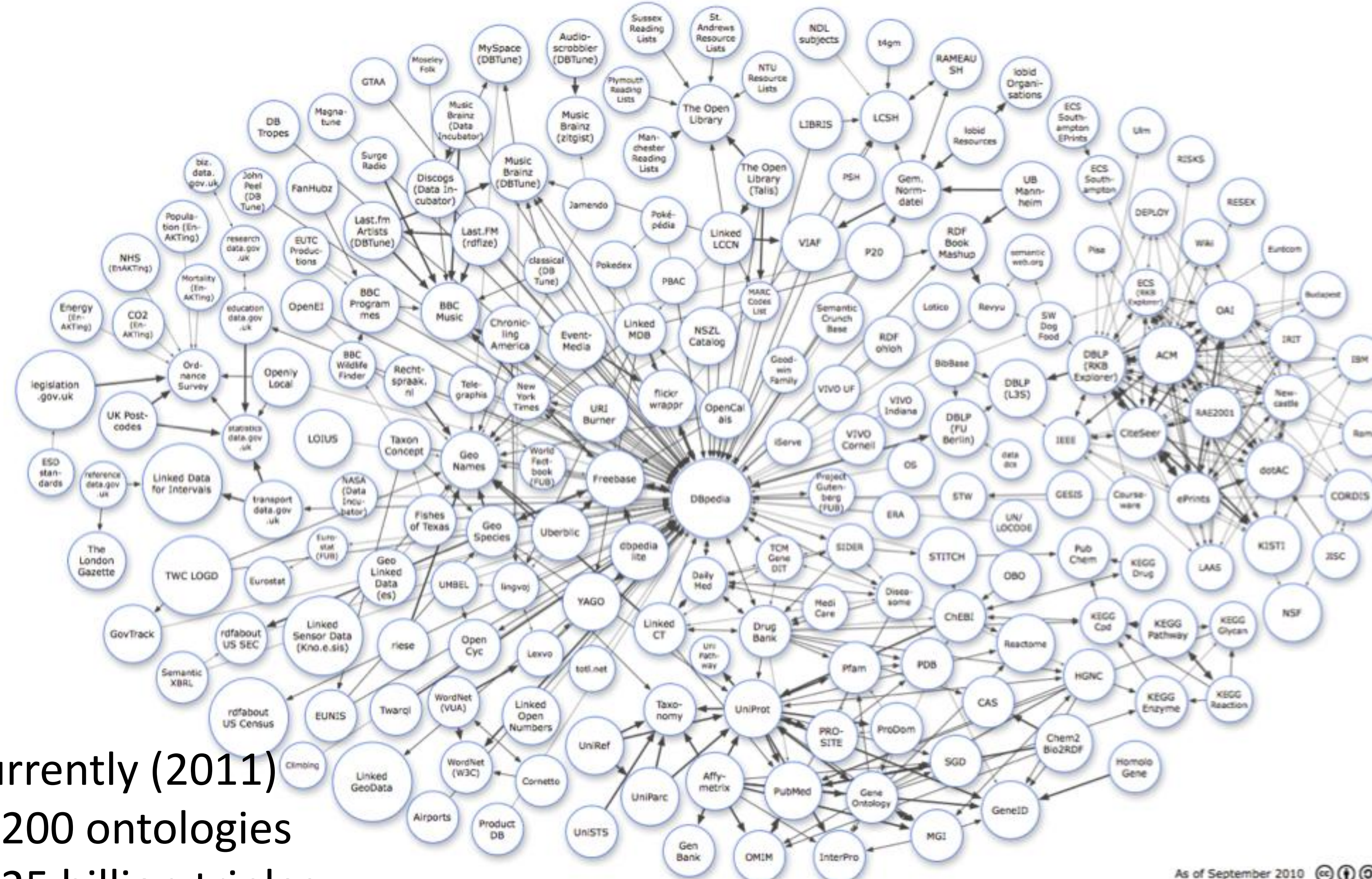
The **Linking Open Data Project** aims to interlink all open RDF data sources into one gigantic RDF graph (link).

LINKINGOPENDATA
W3C SWEO Community Project



[Bizer, JSWIS 2009]

The Linked Data Cloud



Currently (2011)

- 200 ontologies
- 25 billion triples
- 400m links

http://richard.cyganiak.de/2007/10/lod/imagemap.html

Existing Ontologies

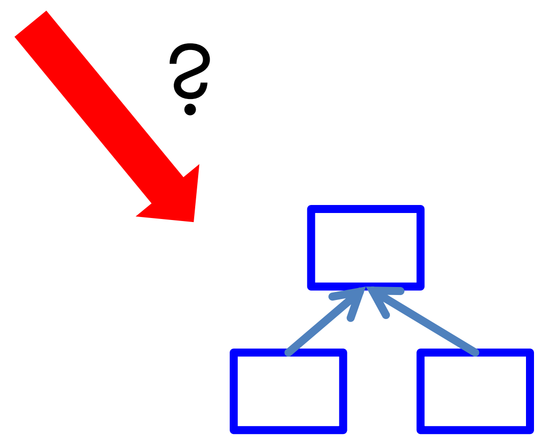
The existing ontologies in the Linked Data Cloud include (http://www4.wiwiss.fu-berlin.de/lodcloud/)

- US census data
- BBC music database
- Gene ontologies
- DBpedia general knowledge, + YAGO, + Cyc etc.
- UK government data
- geographical data in abundance
- national library catalogs (USA, Germany etc.)
- publications (DBLP)
- commercial products
- all Pokemons
- ...and many more

Outline for Part V

- **URIs & Dereferenceable URIs** ✔
- **Shared Vocabularies** ✔
- **Linked Data** ✔
- **The Semantic Web and the Web**

And the rest of the Web?



Microdata

Microdata is a W3C standard to annotate HTML 5 pages with RDF data.

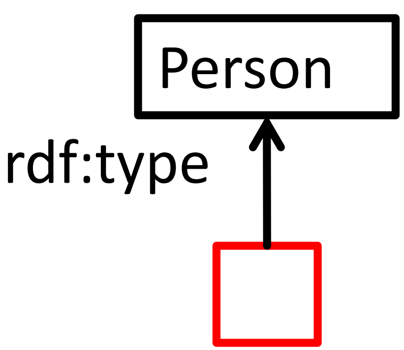
```
<div>

  Martin Thunderbird<br>
  Researcher in Rock'N'Roll Music of 1935-1977<br>
  3764 Presley Boulevard<br>
  Memphis, Tennessee
</div>
```

Creating an Entity

Makes the red box an entity

The type of this entity is “Person”

```
<div itemscope itemtype=“http://schema.org/Person”>

   Martin Thunderbird<br>
   Researcher in Rock’N’Roll Music of 1935-1977<br>
   3764 Presley Boulevard<br>
   Memphis, Tennessee
</div>
```

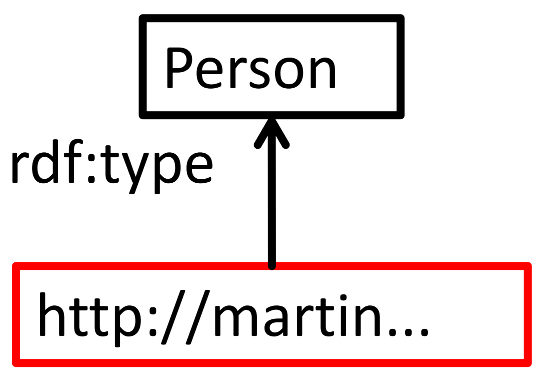


Naming an Entity

Specifies the URI of the entity.

This means we can talk about any entity in the Web of Linked Data!

```
<div itemscope itemtype=http://schema.org/Person
      itemid= http://martin.thunderbird.org/me >
  Martin Thunderbird<br>
  Researcher in Rock’N’Roll Music of 1935-1977<br>
  3764 Presley Boulevard<br>
  Memphis, Tennessee
</div>
```

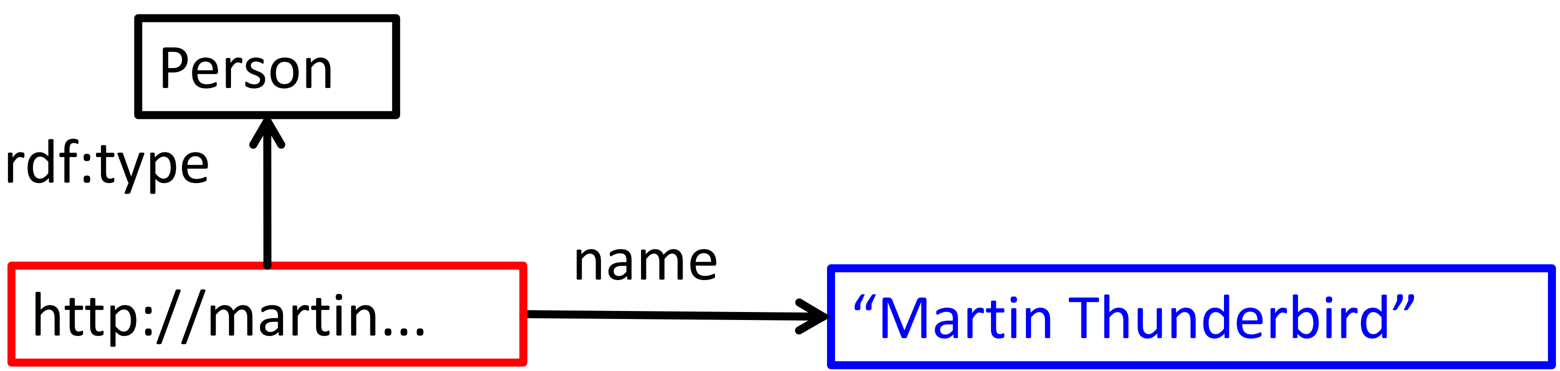


Item Properties

Statements are constructed with “itemprop=“

Text becomes a string node in RDF

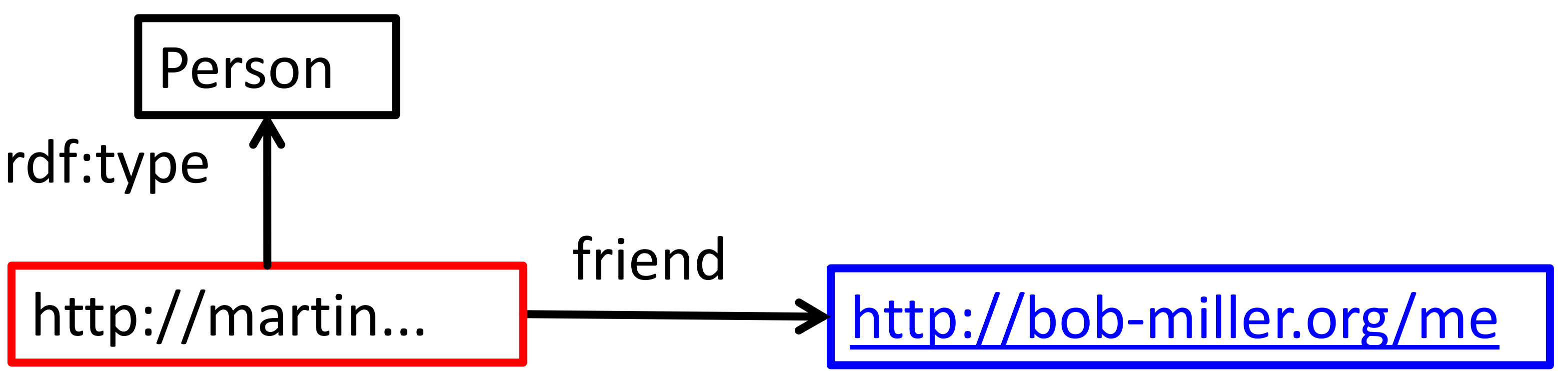
```
<div iten   ope itemtype=http://schema.o    Person
       itemid= http://martin.thunderbird.org   me >
   <span itemprop=“name”>Martin Thunderbird</span><br>
   Researcher in Rock’N’Roll Music of 1935-1977<br>
   3764 Presley Boulevard<br>
   Memphis, Tennessee
</div>
```



Item Properties with URIs

Links given by “itemprop” + “href” become a URI node in RDF

```
<div itemscope itemtype=http://schema.org/Person
      itemid= http://martin.thunderbird.org/me >

          My only friend is
          <a itemprop=“friend” href= http://bob-miller.org/me >
             Bob Miller</a>
</div>
```



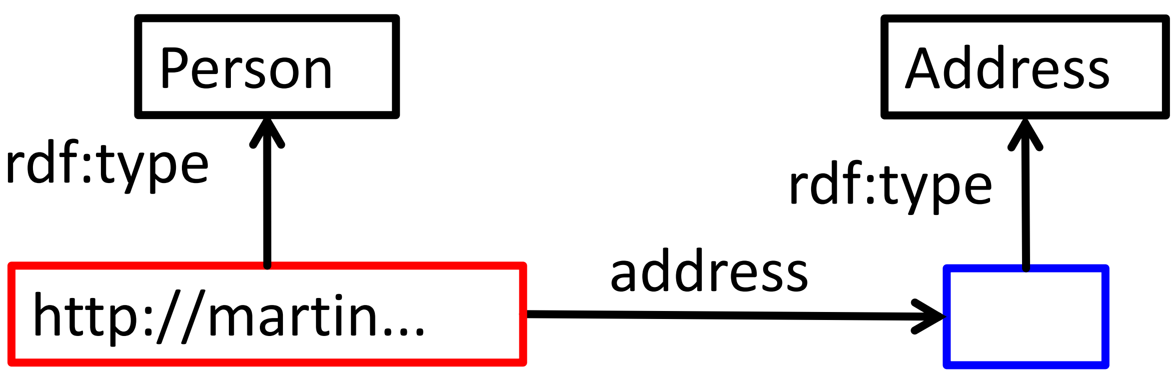
Inner Nodes

“itemprop” creates a link between the outer node and the inner node.

itemscope + itemtype creates a new node

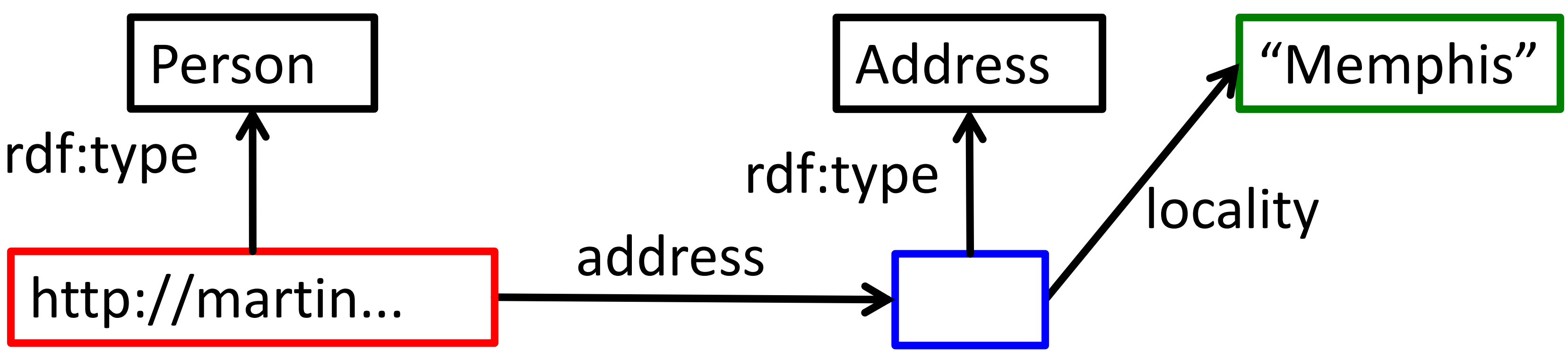
<div ite ope itemtype=http://schem org/Person
iten = http://martin.thunderbi .org/me >

</div>



Inner Nodes

```
<div itemscope itemtype=http://schema.org/Person
      itemref= http://martin.thunderbird.org/me >

 <span itemprop="address" itemscope itemtype=http://.../Address>
      <span itemprop="locality">Memphis</span>
 </span>
</div>
```

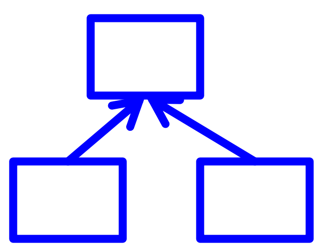


Microdata Summary

Microdata is a W3C standard to annotate HTML 5 pages with RDF data.

Advantages:

- Grass root appeal
  (everybody can start annotating pages)
- No data duplication
  (all data in one file)
- Publisher independence
  (everybody can use his own attributes)

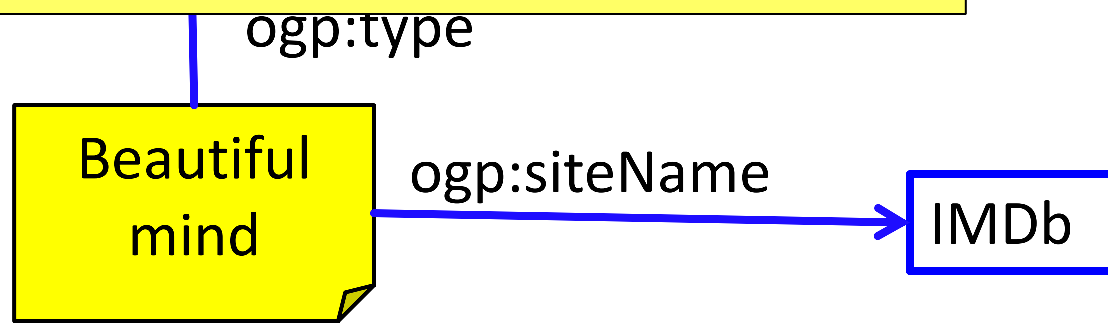


[W3C Microdata 2011]

ML

www.imdb.com/title/tt0268978/

```
<html>
   …
   <meta itemprop=‘type’ content='movie' />
   <meta itemprop ='app_id' content=‘123' />
   …
</html>
```

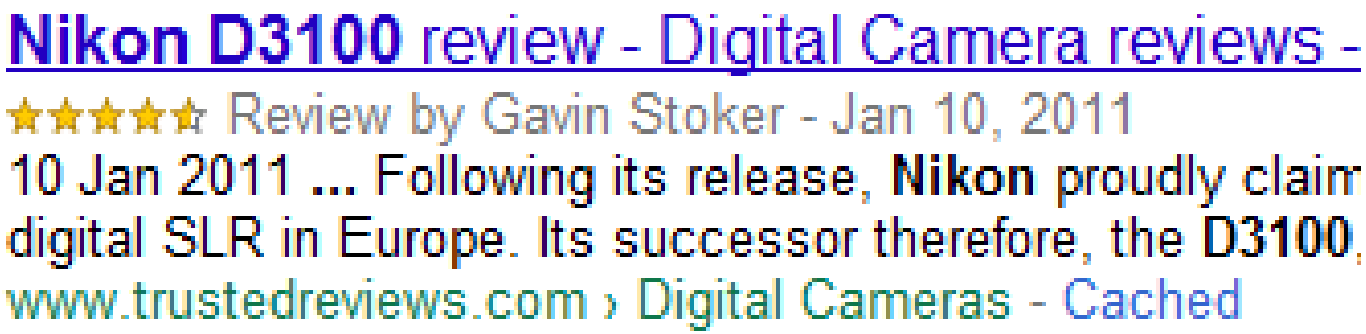


RDF data following the Open Graph Protocol is often embedded in HTML pages, thus allowing the Facebook LIKE button to work.

Search Engines & Annotated HTML

Google, Microsoft and Yahoo have agreed (!) on a common schema (http://schema.org).

It allows annotating HTML pages with meta-information that will show up in “rich snippets”.



Try it out

Schema.org is for the description of people, places, institutions, movies, documents, etc…

Outline for Part V

- **URIs & Dereferenceable URIs ✔**
- **Shared Vocabularies ✔**
- **Linked Data ✔**
- **The Semantic Web and the Web ✔**

References

References

- [W3C URI 2004] W3C: “Architecture of the World Wide Web, Volume One” Recommendation 15 December 2004, http://www.w3.org/TR/webarch/
- [W3C CoolURIs 2008] W3C: “Cool URIs for the Semantic Web” Interest Group Note 03 December 2008, http://www.w3.org/TR/cooluris/
- [W3C Microdata, 2011] W3C: “HTML Microdata”, Working Draft 25 May 2011, http://www.w3.org/TR/microdata/
- [Bizer, JSWIS 2009] C.Bizer, T.Heath, T.Berners-Lee: “Linked data – the story so far” International Journal on Semantic Web and Information Systems, 5(3):1–22, 2009.
- [Tummarello ISWC 2007] G. Tummarello, R. Delbru, E. Oren: “Sindice.com: Weaving the Open Linked Data” ISWC/ASWC 2007:

Outline

- **Part I** ✔
  - **Machine Knowledge & Intelligent Applications**
- **Part II** ✔
  - **Knowledge Representation & Public Knowledge Bases**
- **Part III** ✔
  - **Extracting Knowledge**
- **Part IV** ✔
  - **Ranking and Searching**
- **Part V** ✔
  - **Linked Data**
- **Part VI**
  - **Conclusion and Outlook**

Summary

In this tutorial, we have explained:

- **how a knowledge base is organized**
- **which knowledge bases are publicly available**
- **how we can automatically construct knowledge bases**
- **how we can query a knowledge base and rank the results**
- **how we can deal with inter-linked knowledge bases**

We discussed:

- **fundamental models & methods**
- **state-of-the-art techniques**
- **open problems & research challenges**

Spectrum of Machine Knowledge (1)

factual:

bornIn (GretaGarbo, Stockholm), hasWon (GretaGarbo, AcademyAward), playedRole (GretaGarbo, MataHari), livedIn (GretaGarbo, Klosters)

taxonomic (ontology):

instanceOf (GretaGarbo, actress), subclassOf (actress, artist)

lexical (terminology):

means (“Big Apple“, NewYorkCity), means (“Apple“, AppleComputerCorp)
means (“MS“, Microsoft) , means (“MS“, MultipleSclerosis)

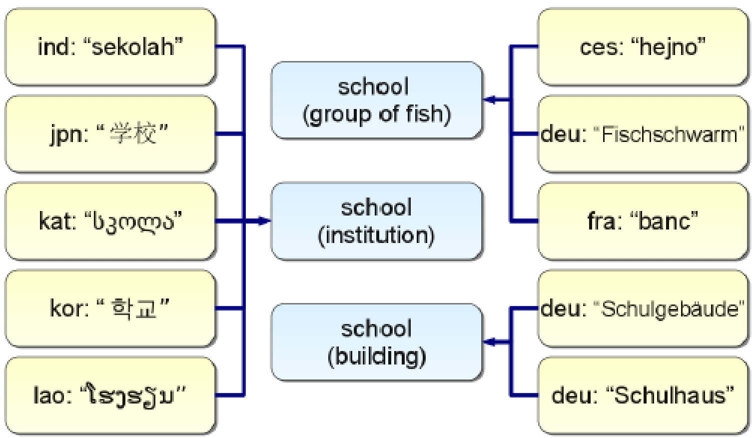
multi-lingual:

meansInChinese („乔戈里峰“, K2), meansInUrdu („ “, K2)
meansInFrench („école“, school (institution)),
meansInFrench („banc“, school (of fish))

Multilingual Lexical Knowledge

WordNet in ca. 50 languages, only English is big
several 1000 languages spoken/written in this world

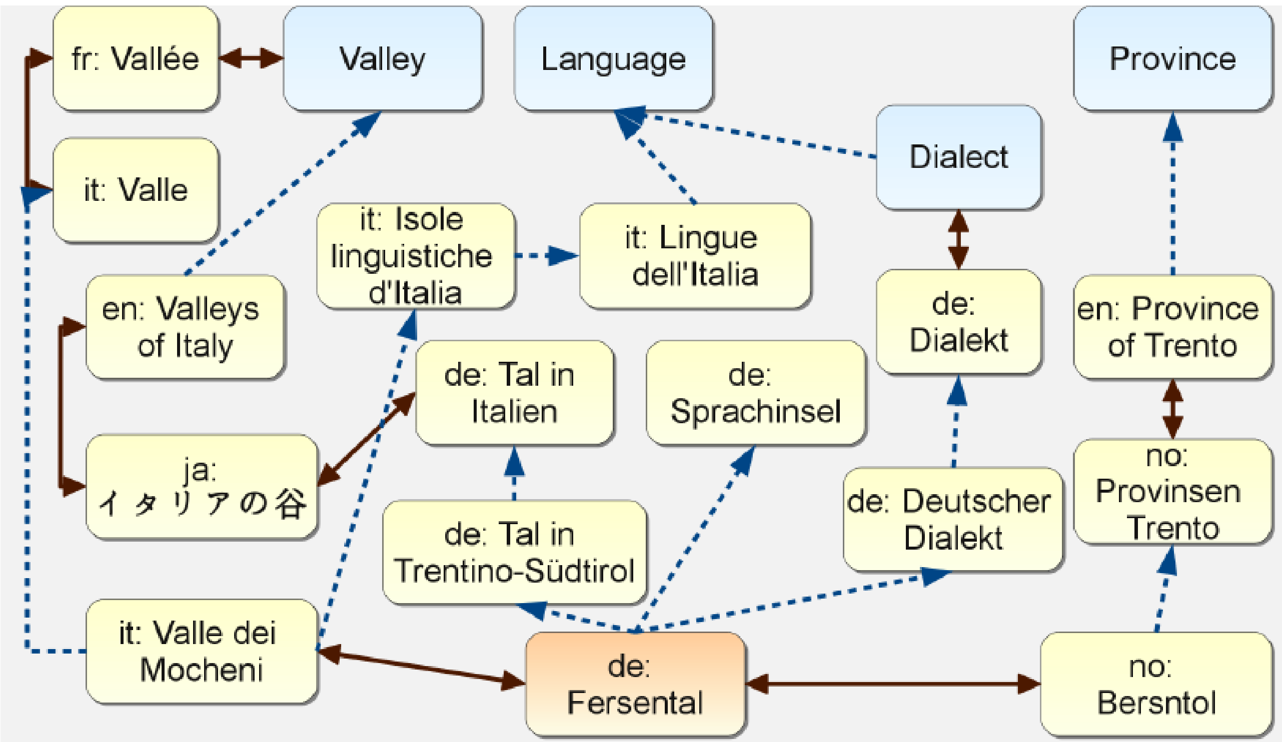
- UWN (de Melo: CIKM'09): 800 000 words, 200 languages, 120 000 senses
- PanDictionary (Mausam: AAAI'10): 10 Mio. words, 1000 languages, 80 000 senses
- WikiNet (Nastase: LREC'10): 3 Mio. concepts, 100 languages



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

Knowledge from Many Languages

- Integrate entities across Wikipedia editions
- Derive taxonomic and factual knowledge



Identify good edges:
min cost for dropping
equivalence evidence +
distinctness evidence
→ ILP, LP relexation,
random walks, etc.

once cleaned, multilingual links and categories yield additional instanceOf and subclassOf facts

(de Melo: CIKM'10)

Spectrum of Machine Knowledge (2)

ephemeral (dynamic services):

wsdl:getSongs (musician ?x, song ?y), wsdl:getWeather (city?x, temp ?y)

common-sense (properties):

hasAbility (Fish, swim), hasAbility (Human, write),

hasShape (Apple, round), hasProperty (Apple, juicy),

hasMaxHeight (Human, 2.5 m)

common-sense (rules):

$\forall$ **x: human(x)** $\Rightarrow$ **male(x)** $\vee$ **female(x)**

$\forall$ **x: (male(x)** $\Rightarrow \neg$ **female(x))** $\wedge$ **(female(x))** $\Rightarrow \neg$ **male(x))**

$\forall$ **x: animal(x)** $\Rightarrow$ **(hasLegs(x)** $\Rightarrow$ **isEven(numberOfLegs(x))**

temporal (fluents):

hasWon (GretaGarbo, AcademyAward)@1955

marriedTo (AlbertEinstein, MilevaMaric)@[6-Jan-1903, 14-Feb-1919]

Spectrum of Machine Knowledge (3)

free-form (open IE):

hasWon (NataliePortman, AcademyAward)

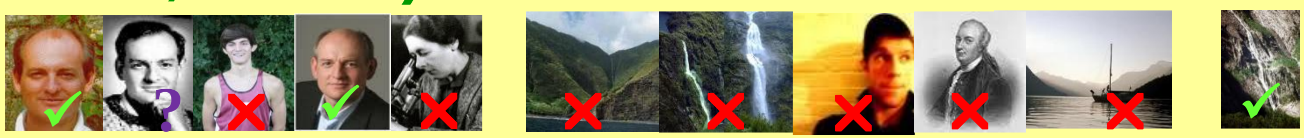
occurs („Natalie Portman“, „celebrated for“, „Oscar Award“)

occurs („Jeff Bridges“, „nominated for“, „Oscar“)

multimodal (photos, videos):

StuartRussell

JamesBruceFalls



social (opinions):

admires (maleTeen, LadyGaga), supports (AngelaMerkel, HelpForGreece)

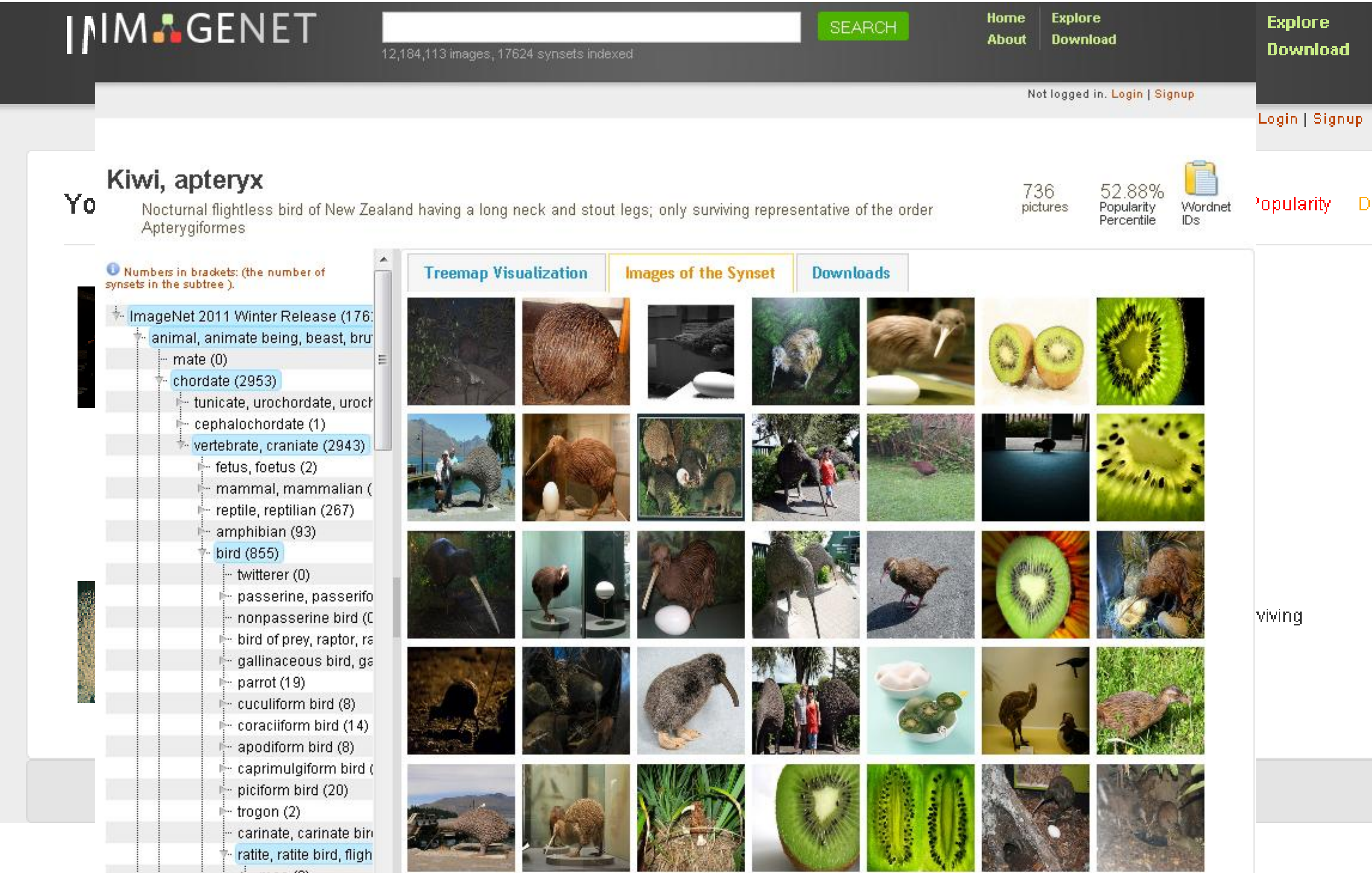
epistemic ((un-)trusted beliefs):

believe(Ptolemy,hasCenter(world,earth)), believe(Copernicus,hasCenter(world,sun))

believe (peopleFromTexas, bornIn(BarackObama,Kenya))

ImageNet: Visual WordNet

http://www.image-net.org/ [J. Deng et al.: CVPR 2009]



ImageNet: Visual WordNet

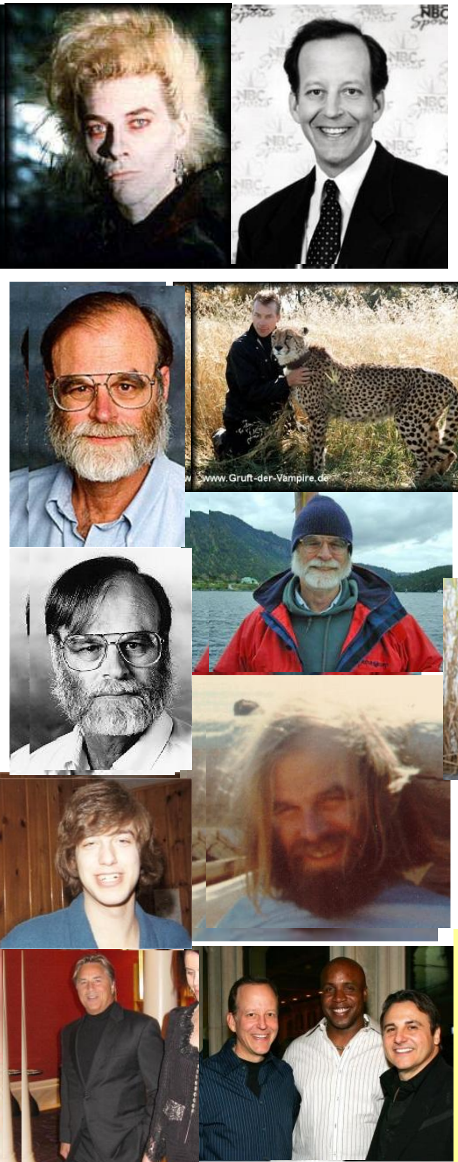
http://www.image-net.org/

[J. Deng et al.: CVPR 2009]

Photos of Entities in the Long Tail

[B. Taneva et al.: WSDM'10]

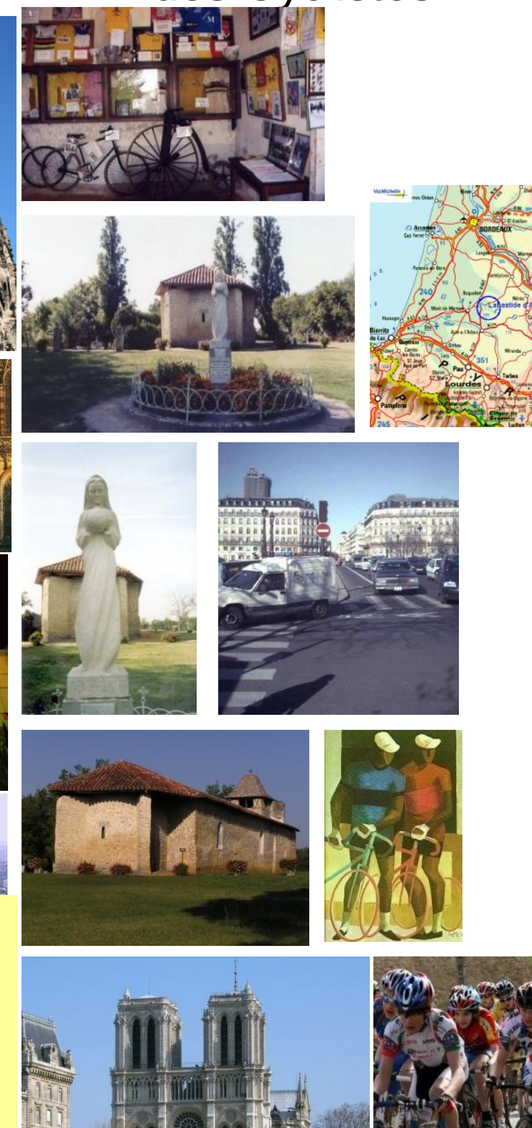
q: Jim Gray



q: Natassa Ailamaki

q: Barcelona cathedral

q: Notre Dame des Cyclistes



- tackle ambiguity & rarity
- aim for precision, recall, diversity
  → exploit knowledge on entity

David Patterson	David Patterson Berkeley	David Patterson RISC	David Patterson ACM	combined method

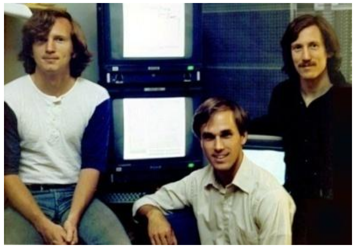


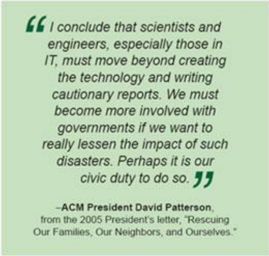


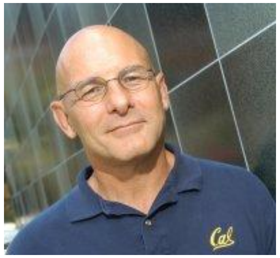


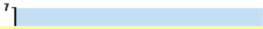
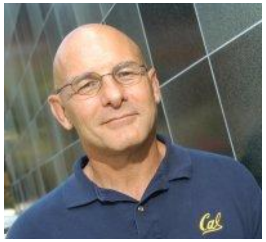










- exploit KB:
  - worksFor (DavidPatterson, Berkeley)
  - invented (DavidPatterson, RISC)
  - presidentOf (DavidPatterson, ACM)
  - wonAward (DavidPatterson, …)
- generate expanded queries
- combine results by voting

KB Building: Achievements & Challenges

Entities & Classes

strong success story, some problems left:

- large taxonomies of classes with individual entities
- long tail calls for new methods
- entity disambiguation remains grand challenge

Relationships

good progress, but many challenges left:

- recall & precision by patterns & reasoning
- efficiency & scalability
- soft rules, hard constraints, richer logics, …
- open-domain discovery of new relation types

Temporal Knowledge

widely open (fertile) research ground:

- uncertain / incomplete temporal scopes of facts
- joint reasoning on ER facts and time scopes

KB Applications: Achievements & Challenges

Search & QA

good progress on entity awareness; next challenges:

- coping with entities in the long tail
- querying relational facts for knowledge-intensive QA
- compelling UI for QA input/output (speech, visual, …)
- composable services (e.g. API for Sparql+text+time+…)

Ranking & Recommendation

progress on statistical ranking; problems remaining:

- ranking for relational queries and QA results
- consideration of diversity, trust, provenance
- aggregation of uncertain statements

Contextualization, Disambiguation & Linkage

key to all of this, remains challenging

Broad Application Areas

Web 2.0, mobile, multimodal, digital humanities, health, biology, …

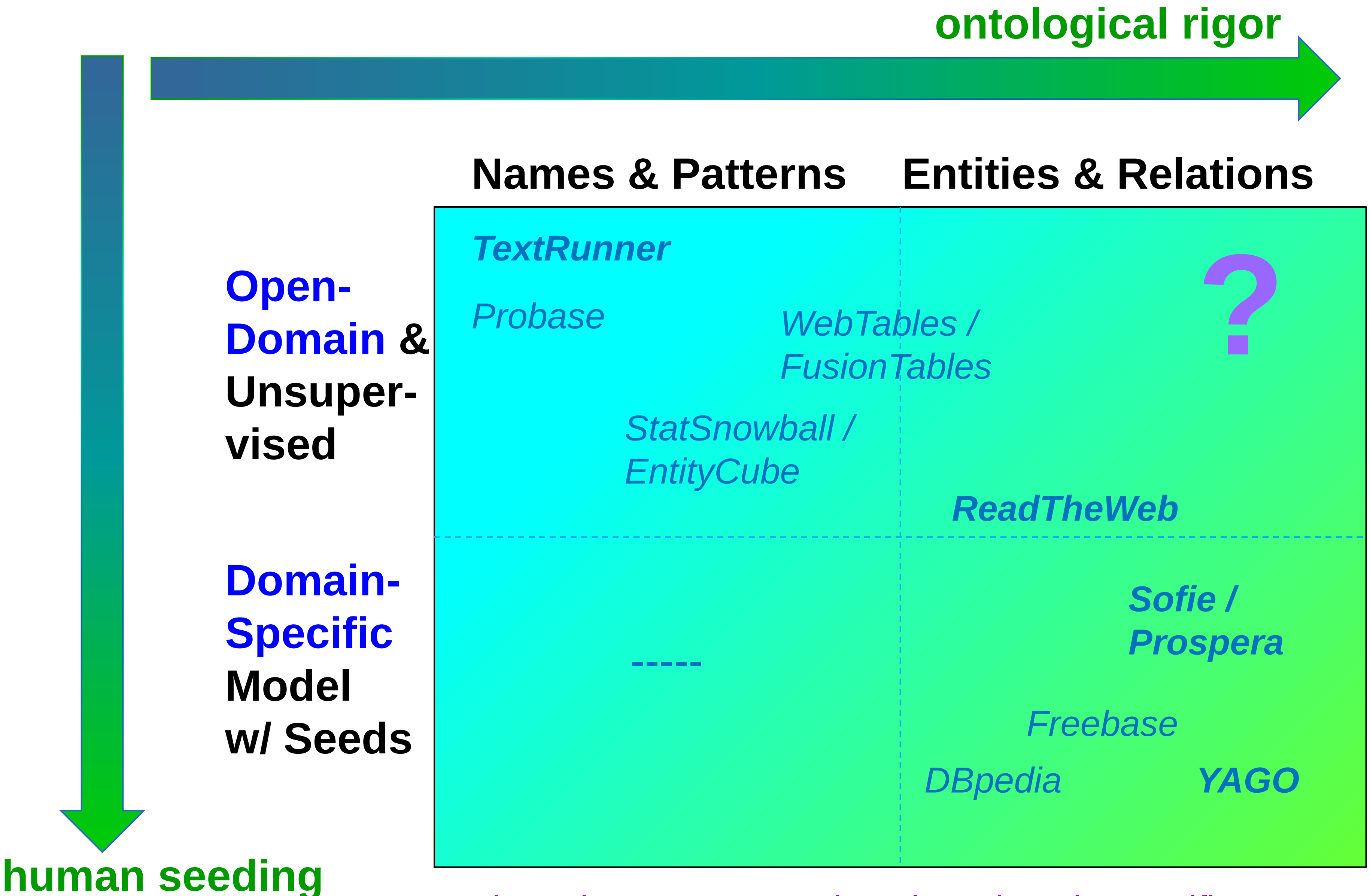
Grand Challenge: Web-Scale KB Construction

ontological rigor

human seeding

	Names & Patterns	Entities & Relations
Open-Domain & Unsupervised	→ < „N. Portman“, „honored with“, „Academy Award“>, < „Jeff Bridges“, „expected to win“, „Oscar“ > < „Bridges“, „nominated for“, „Academy Award“>	
Domain-Specific Model w/ Seeds		wonAward: Person × Prize type (Meryl_Streep, Actor) wonAward (Meryl_Streep, Academy_Award) → wonAward (Natalie_Portman, Academy_Award) wonAward (Ethan_Coen, Palme_d‘Or)

Grand Challenge: Web-Scale KB Construction



→ aim to integrate open-domain & domain-specific IE !

Overall Take-Home

Historic opportunity:
revive Cyc vision, make it real & large-scale !
challenging, but high pay-off

Explore & exploit synergies between
semantic, statistical, & social Web methods:
statistical evidence + logical consistency !

For DB / AI / IR / NLP / Web researchers:

- efficiency & scalability
- constraints & reasoning
- killer app for uncertain data management (prob. DB)
- search & ranking for RDF + text
- text (& speech) disambiguation
- knowledge-base life-cycle: growth & maintenance

Outline

- Part I ✔
  - Machine Knowledge & Intelligent Applications
- Part II ✔
  - Knowledge Representation & Public Knowledge Bases
- Part III ✔
  - Extracting Knowledge
- Part IV ✔
  - Ranking and Searching
- Part V ✔
  - Linked Data
- Part VI
  - Conclusion and Outlook ✔

The End

The slides are available at

http://www.mpi-inf.mpg.de/yago-naga/IJCAI11-tutorial/

Feel free to contact us with further questions





Hady Lauw
Institute for
Infocomm Research
Singapore
http://hadylauw.com

Fabian M. Suchanek
INRIA Saclay
Paris
http://suchanek.name

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

Ralf Schenkel
Saarland University
Saarbrücken
http://people.mmci.uni-saarland.de/~schenkel/

Martin Theobald
Max-Planck Institute
for Informatics
Saarbrücken
http://www.mpi-inf.mpg.de/~mtb/

Thanks

erc

SEVENTH FRAMEWORK PROGRAMME

living know ledge

M2CI CLUSTER OF EXCELLENCE

DFG Deutsche Forschungsgemeinschaft

Google