



max planck institut
informatik

Knowledge Harvesting from Web Sources

Part 1: Knowledge Bases and their Automatic Construction

Gerhard Weikum

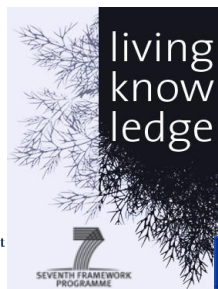
Max Planck Institute for Informatics

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

Acknowledgements



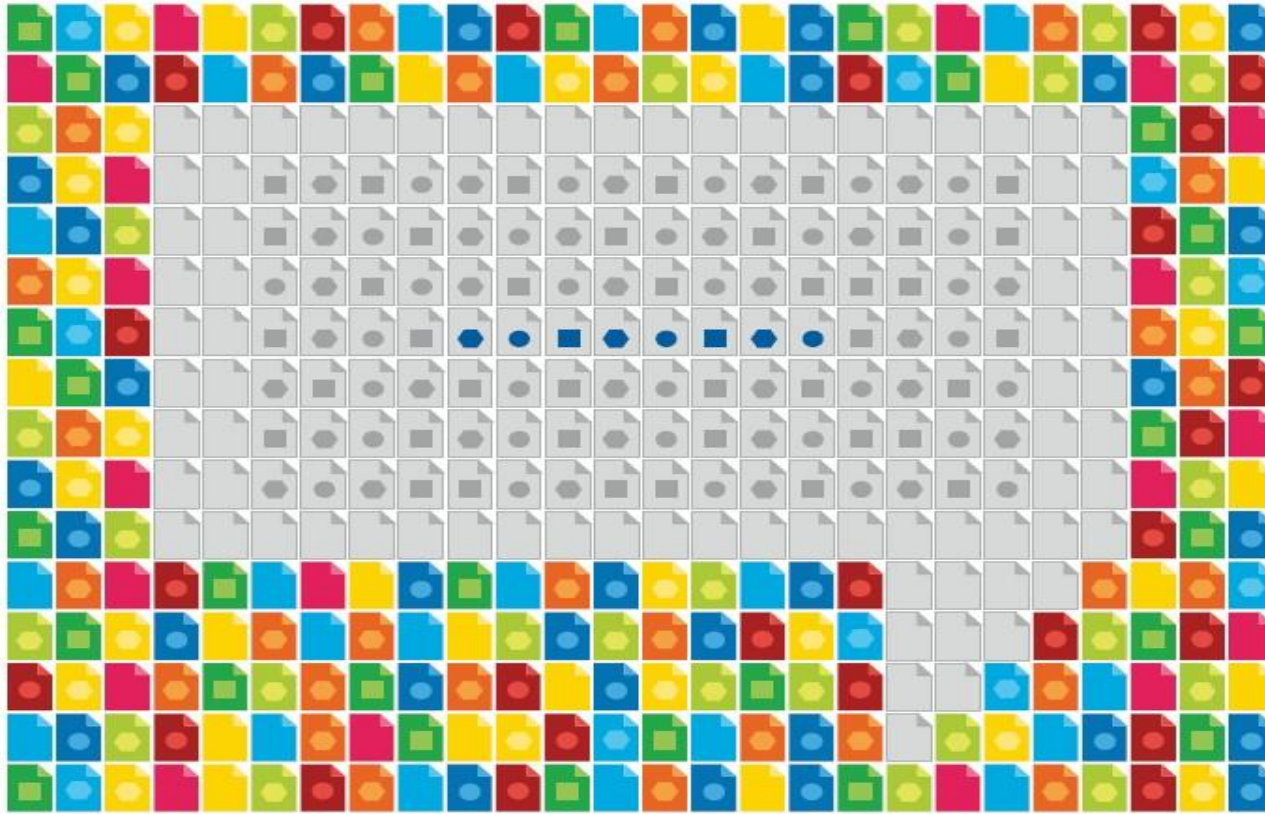
mpi max planck institut
informatik



DFG Deutsche
Forschungsgemeinschaft

Google

Goal: Turn Web into Knowledge Base



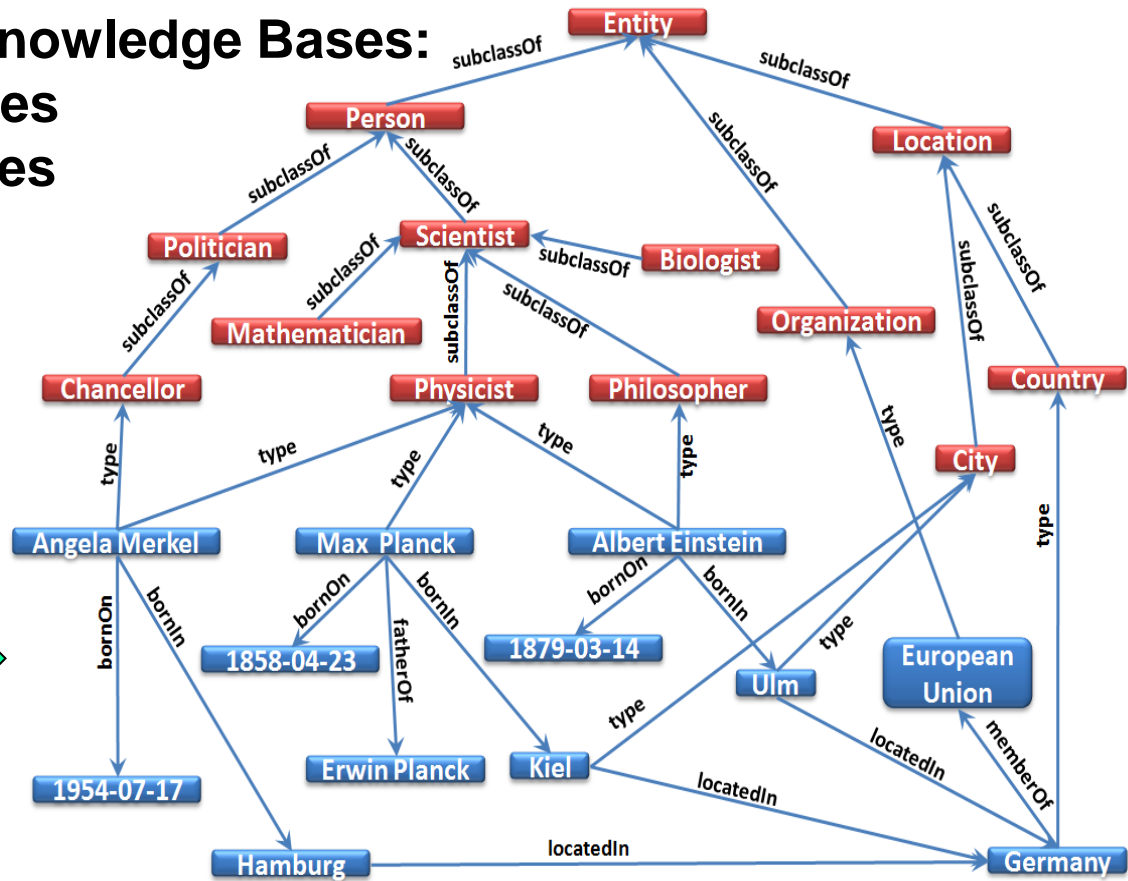
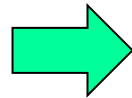
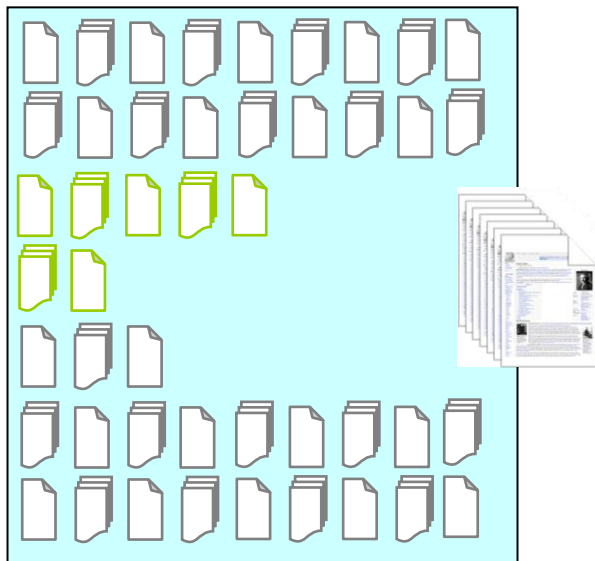
Source:
DB & IR methods for
knowledge discovery.
Communications of
the ACM 52(4), 2009

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

Approach: Harvesting Facts from Web

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



SUMO



DBLife

TextRunner



YAGO-NAGA



umbel



DBpedia



SIG.MA
SEMANTIC INFORMATION
MASHUP



IWP

True Knowledge[®]
The Internet Answer Engine[™] BETA

Carnegie Mellon

ReadTheWeb



max planck institut
informatik

freebase[™]

WolframAlpha[™] computational-
knowledge engine

Knowledge for Intelligence

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

- ★ 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?
- ★ Enzymes that inhibit HIV?
Influenza drugs for teens with high blood pressure?
...

Application 1: Semantic Queries on Web



ruddian composers

Square it

[canadian actors](#)

[8000 meter peaks](#)

[european union countries](#)

[antibiotics](#)

[maine lighthouses](#)

[hotels in chicago](#)

[canadian prime ministers](#)

[blueberry varieties](#)

[dog breeds](#)

[tennis champions](#)

[bars in pacific heights](#)

[italian provinces](#)

[Start with an empty Square](#)

[Google Home](#)

© 2011 - [Privacy](#)

Application 1: Semantic Queries on Web










Google squared labs

russian composers

Square it Add to this Square

Unsaved Share Export Save

17 items


Item Name	Image	Description	Date Of Death	Date Of Birth	Place Of Birth	Place Of Death	Add columns	Add
<input checked="" type="checkbox"/> Modest Mussorgsky		Modest Petrovich Mussorgsky (Russian: Модѣст Петровичъ Мусоргскій; 21 March [O.S. 9 March] 1839, Karevo – 28 March [O.S. 16 March] 1881, ... For many years	Mar. 28, 1881	March 21, 1839	Pskov, Russia	St. Petersburg, Russia		
<input checked="" type="checkbox"/> Igor Stravinsky		Igor Fyodorovich Stravinsky 17 June [O.S. 5 June] 1882 – 6 April 1971) was a Russian-born, naturalized French, later naturalized American composer, pianist, and	1971-04-06	Jun				
<input checked="" type="checkbox"/> Alfred Schnittke		Alfred Schnittke (Russian: Альфред Гари́евич Шни́тке (Al'fred Gariyevich Šnitke); November 24, 1934 – August 3, 1998) was a Russian and Soviet composer. Schnittke's early	1998-08-03	193				
<input checked="" type="checkbox"/> Sergei Prokofiev		In 1902, Prokofiev's mother met Sergei Taneyev, director of the Moscow Conservatory, who initially suggested that Prokofiev should start lessons in piano and composition with	1953-03-05	Apr				
<input checked="" type="checkbox"/> Nikolai Rimsky-Korsakov		Nikolai Andreyevich Rimsky-Korsakov was a Russian composer, and a member of the group of composers known as The Five. He was a master of orchestration. His best -known	June 21, 1908	Mar				
<input checked="" type="checkbox"/> Dmitri Shostakovich		Dmitri Dmitriyevich Shostakovich was a Soviet Russian composer and one of the most celebrated composers of the 20th century. Shostakovich achieved fame in the Soviet	1975-08-09	September 25, 1906	Saint Petersburg, Russia	Moscow, USSR		
<input checked="" type="checkbox"/> Sergei Rachmaninoff		Sergei Vasilievich Rachmaninoff (Russian: Серге́й Васи́льевич Рахма́нинов) (Russian pronunciation: [sʲɪrˈɡʲej rɐxˈmɐnʲɪnɐf]; 1 April 1873 – 28 March 1943) was a Russian	1943-03-28	April 1, 1873	Novgorod, Russia	Beverly Hills, California, United States		
<input checked="" type="checkbox"/> Pyotr Ilyich Tchaikovsky		Pyotr Ilyich Tchaikovsky /tʃɑːˈkɒfski/ (Russian: Пётр Ильич Чайковский) (May 7, 1840 – November 6, 1893) was a Russian composer of the Romantic era. His wide-	November 6, 1893	May 7, 1840	Votkinsk	St. Petersburg, Russia		
<input checked="" type="checkbox"/> Mily Balakirev		Mily Alexeyevich Balakirev 2 January 1837 [O.S. 21 December 1836] – 29 May [O.S. 16 May] 1910), was a Russian pianist, conductor and composer known today primarily for his	1910-05-29	1837-01-02	Nizhny Novgorod	Saint Petersburg		

St. Petersburg, Russia
Location of death for Modest Mussorgsky
[www.nndb.com](#) - [all 2 sources »](#)

Other possible values







- St** Low confidence
Date: March 16, 1881. Place of Birth: Pskov, Russia. **Place of Death: St.** Petersburg, ...
[www.bookrags.com](#) - [all 2 sources »](#)
- Worldwide** Low confidence
Location for Modest Mussorgsky
[www.youtube.com](#)
- Saint Petersburg** Low confidence
Place of death for Modest Mussorgsky
[www.freebase.com](#)
[Search for more values »](#)

Application 1: Semantic Queries on Web



Unsaved

russian composers from moscow 6 items

Item Name	Image	Description	Date Of Birth	
<input checked="" type="checkbox"/> Valery Gergiev		The Moscow Easter Festival was inaugurated in 2002 as a joint effort between Maestro Valery Gergiev of St. Petersburg's Mariinsky Theatre and Yuri Luzhkov, ... The Moscow Easter	1953-05-02	<div>Type your own... Language Publisher Date Of Death Year Of Death Orchestra Tchaikovsky See Died Date Hotels</div>
<input checked="" type="checkbox"/> Oxford		Here are listed some of the operas written and premiered in Russia: Vincenzo Manfredini (1737–1799) spent 12 years in Russia and died in Saint Petersburg. The son and pupil of	2 possible values	
<input checked="" type="checkbox"/> Anton Rubinstein		Anton Grigorevich Rubinstein (Russian: Анто́н Григо́рьевич Руби́нштейн, tr. Anton Grigor'evič Rubínštejn) (November 28 [O.S. November 16] 1829 – November 20 [O.S.	1829-11-28	
<input checked="" type="checkbox"/> Gennady Rozhdestvensky		Gennady Nikolayevich Rozhdestvensky (Генна́дий Никола́евич Рожде́ственский) (born May 4, 1931) is a Russian conductor. ... Rozhdestvensky was born in Moscow. His	1931-05-04	
<input checked="" type="checkbox"/> Vodka.		Vodka (Polish: wódka, Russian: водка, Ukrainian: ро́півка, Slovak: vodka , Czech: vodka , Lithuanian: degtinė, Estonian: viin, German: Wodka) is a distilled beverage. It is	July 12, 1977	
<input checked="" type="checkbox"/> Yuri Temirkanov		Yuri Khatuevich Temirkanov (Russian: Ю́рий Хату́евич Теми́рканов) (born 10 December 1938) is a Russian conductor of Circassian (Kabardian) origin. Yuri Temirkanov has been	1 possible value	
<input type="button" value="Add items"/>	<input type="button" value="Add"/>			

Not finding the right items? [Start with an empty Square.](#)

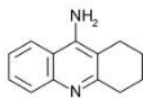
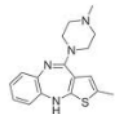
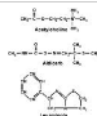
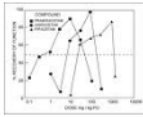




max planck institut
informatikwww.google.com/squared/

Application 1: Semantic Queries on Web



Did you mean: [drugs for treating Alzheimer's](#)

drugs for treating Alzheimer

Item Name	Image	Description	Cas Number	Formula	Half Life	Pubchem
<input checked="" type="checkbox"/> 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
<input checked="" type="checkbox"/> Olanzapine		The Food and Drug Administration (FDA) has declined to approve Memantine (namenda) to treat mild Alzheimer's . Olanzapine (Zyprexa) Olanzapine (Zyprexa). Atypical Antipsychotic	132539-06-1	C17H20N4S	21-54 hours	CID 4585
<input checked="" type="checkbox"/> 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
<input checked="" type="checkbox"/> 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
<input checked="" type="checkbox"/> 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
<input checked="" type="checkbox"/> 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		
<input checked="" type="checkbox"/> 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
<input checked="" type="checkbox"/> 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	70 hours		

www.google.com/squared/



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



freebase™

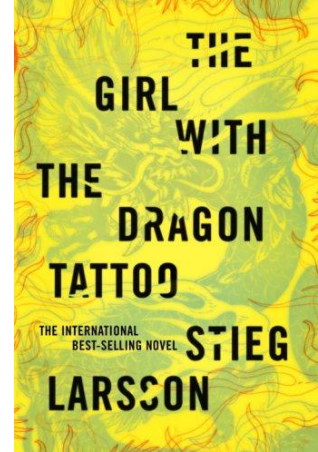


YAGO

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 the same time on the same day and of Hedeby. The old man deceives Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the men who own the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Vanger. After discovering that Salander has hacked into his cell, Blomkvist persuades Salander to assist him with research. They even have an affair, 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. A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, hires that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

Outline

✓ **Motivation**

★ **Machine Knowledge**

★ **Knowledge Harvesting**

- **Entities and Classes**
- **Relational Facts**

★ **Research Challenges**

- **Open-Domain Extraction**
- **Temporal Knowledge**

★ **Wrap-up**

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 („کے ٹو“, 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):

$\forall x: \text{human}(x) \Rightarrow \text{male}(x) \vee \text{female}(x)$

$\forall x: (\text{male}(x) \Rightarrow \neg \text{female}(x)) \wedge (\text{female}(x) \Rightarrow \neg \text{male}(x))$

$\forall x: \text{animal}(x) \Rightarrow (\text{hasLegs}(x) \Rightarrow \text{isEven}(\text{numberOfLegs}(x)))$

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))

Knowledge Representation

- **RDF** (Resource Description Framework, W3C):
subject-property-object (SPO) **triples**, binary relations structure, but no (prescriptive) schema
- **Relations**, frames
- Description logics: **OWL**, **DL-lite**
- Higher-order logics, epistemic logics

facts (RDF triples):

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

facts about facts:

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

temporal & provenance annotations

can refer to reified facts via fact identifiers

(approx. equiv. to RDF quadruples: “Color” × Sub × Prop × Obj)

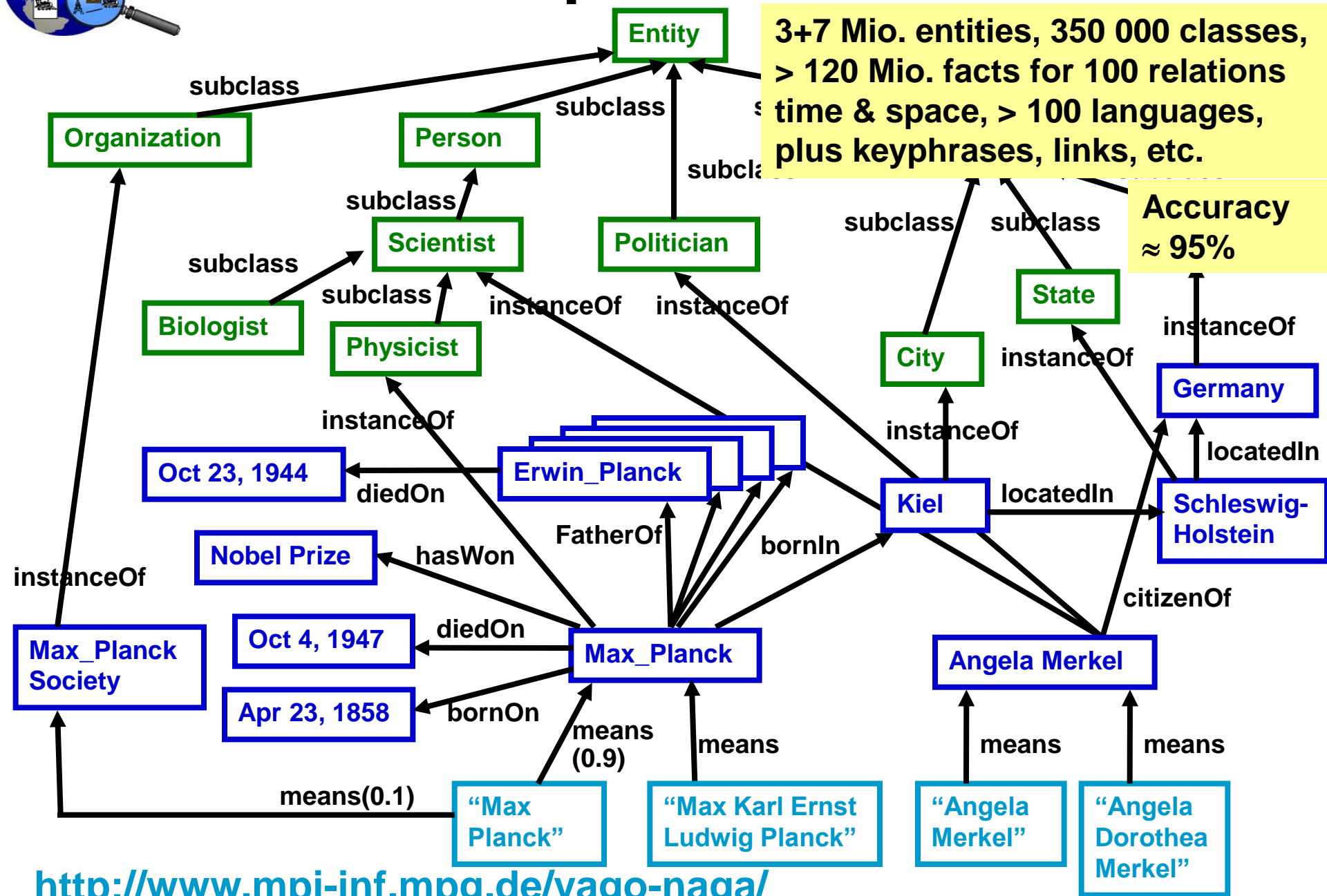


KB's: Example YAGO

(Suchanek et al.: WWW'07,
Hoffart et al.: WWW'11)

3+7 Mio. entities, 350 000 classes,
> 120 Mio. facts for 100 relations
time & space, > 100 languages,
plus keyphrases, links, etc.

Accuracy
≈ 95%





YAGO2 Knowledge Base (Nov 2010)

integrates knowledge from Wikipedia, WordNet, Geonames:
10 M entities, 350 K classes, 120+300 M facts, 95% accuracy

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

Browse YAGO2

Entity: ☒ case insensitive

Igor_Stravinsky

☒ Show transitive facts

<ul style="list-style-type: none">← Ballets by Igor Stravinsky← Igor' Fëdorovič Stravinskij← Igor Fyodorovich Stravinskij← Igor Fyodorovich Stravinsky<ul style="list-style-type: none">← Igor Stravinski← Igor Stravinskij← Igor stravinsky← Igor Stravinsky← Igor Strawinsky← Katerina Nossenko<ul style="list-style-type: none">← Stravinsky← Stravinski← Stravinskij← Stravinsky← Stravinsky Igor← Stravinsky, Igor← Stravinsky, Igor Fedorovich<ul style="list-style-type: none">← Strawinsky← Игорь Фёдорович Стравинский	means
← Igor Stravinsky	hasPreferredMeaning

hasWonPrize	Grammy Award → Grammy Lifetime Achievement Award →
hasPreferredName	Igor Stravinsky →
hasWikipediaCategory	1882 births → 1971 deaths → 20th-century classical composers → Ballet composers → Ballets Russes composers → Burials at Isola di San Michele → Grammy Award winners → Grammy Lifetime Achievement Award winners → Harvard University people → Honorary Members of the Royal Philharmonic Society → Modernist composers → National Museum of Dance Hall of Fame inductees → Naturalized citizens of the United States → Neoclassical composers → Opera composers → People from Lomonosov → People from Saint Petersburg → Ragtime composers → Royal Philharmonic Society Gold Medallists → Russian ballet → ...

Agon (ballet) →
Apollo (ballet) →
Le baiser de la fée →
Les noces →
Petrushka (ballet) →
Pulcinella (ballet) →
Scènes de ballet (Ashton) →



YAGO2 Knowledge Base (Nov 2010)

integrates knowledge from Wikipedia, WordNet, Geonames:
10 M entities, 350 K classes, 120+300 M facts, 95% accuracy

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

true&l=true&entity=Igor_Stravinsky

type	<p>20th-century classical composers → composer → musician → artist → creator → person → Ballet composers → composer → musician → artist → creator → person → Ballets Russes composers → composer → musician → artist → creator → person → Harvard University people → person → Modernist composers → composer → musician → artist → creator → person → Naturalized citizens of the United States → citizen → national → person → Neoclassical composers → composer → musician → artist → creator → person → Opera composers → composer → musician → artist → creator → person → People from Lomonosov → person → People from Saint Petersburg → person → Ragtime composers → composer → musician → artist → creator → person → Royal Philharmonic Society Gold Medallists → medalist → winner → contestant → person → Russian composers → composer → musician → artist → creator → person → Soviet immigrants to the United States → immigrant → migrant → traveler → person → artist → creator → person → causal agent → physical entity → entity → citizen → national → person → composer → musician → artist → creator → person → contestant → person → creator → person → ...</p>
influences	<p>Carla Lucero → Jono El Grande → Nicholas Lens →</p>
hasGivenName	<p>Igor →</p>
	<p>1913 premiere of → Aaron Copland → Aeolian Company → Agn...</p>

Id	Subject	Property	Object	Time	Location	Keywords
1	#627524445	Teimuraz II	type	1680-01-01	1762-01-08	Erkeli Lof

Firefox

Yago 2 spotk: A Core of Semantic Know... +

https://d5gate.ag5.mpi-sb.mpg.de/web/yagospotk/WebInterface?passedQuery=I%3A2%09S%3A\u003fp%09P%3AisLocatedIn%09O%3ARussia%3BP%3A1%09S%3A\u003f%09O%3A\u00

Google

Query

Id	Subject	Property	Object	Time	Location	Keywords
?id0:	<input type="text" value="?x"/>	<input type="text" value="isA"/>	<input type="text" value="writer"/>	<input type="text"/>	<input type="text"/>	<input type="text" value="romantic poem"/>
?id1:	<input type="text" value="?x"/>	<input type="text"/>	<input type="text" value="?p"/>	<input type="text"/>	<input type="text" value="nearby"/>	<input type="text" value="Saint Petersburg"/>
?id2:	<input type="text" value="?p"/>	<input type="text" value="isLocatedIn"/>	<input type="text" value="Russia"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
?id3:	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
?id4:	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Results

	Id	Subject	Property	Object	Time	Location	Keywords
1	#545740785	Alexander Pushkin	type	writer	1799-06-06 ↓↑, 1837-02-10 ↓↑	-	2208 Pushkin Pushkin ...
	#505944477	Alexander Pushkin	diedIn	Saint Petersburg	1837-02-10 ↓↑, 1837-02-10 ↓↑	-	2208 Pushkin Pushkin ...
	#507130749	Saint Petersburg	isLocatedIn	Russia	-	-	Zelenogorsk Trams in ...
	#8987372	Russia	means	Russia	0862-01-01 ↓↑, -	Russia	Turkey Sukhoi Superjet ...
	#971322	writer	means	writer	-	-	-



KB's: Example DBpedia

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

About: Igor Stravinsky

An Entity of Type : [Neoclassical composers](#), from Named Graph : <http://dbpedia.org>, within Data Space : dbpedia.org



Igor Fyodorovich Stravinsky (17 June 1882 – 6 April 1971) was a Russian composer, one of the most important and influential composers of 20th century music. He was a quintessential one of the 100 most influential people of the century. He became a naturalized US citizen.

- **3.5 Mio. entities,**
- **700 Mio. facts (RDF triples)**
- **1.5 Mio. entities mapped to hand-crafted taxonomy of 259 classes with 1200 properties**
- **interlinked with Freebase, Yago, ...**

Property	Value
dbpedia-owl:birthDate	<ul style="list-style-type: none">▪ 1882-06-17 (xsd:date)
dbpedia-owl:birthPlace	<ul style="list-style-type: none">▪ dbpedia:Lomonosov_Russia▪ dbpedia:Russia
dbpedia-owl:deathDate	<ul style="list-style-type: none">▪ 1971-04-06 (xsd:date)
dbpedia-owl:deathPlace	<ul style="list-style-type: none">▪ dbpedia:New_York_City▪ dbpedia:New_York▪ dbpedia:United_States
dbpedia-owl:thumbnail	<ul style="list-style-type: none">▪ http://upload.wikimedia.org/wikipedia/en/thumb/3/33/Igor_Stravinsky_LOC_32392u.jpg/200px-Igor_Stravinsky_
dbpprop:alternativeNames	<ul style="list-style-type: none">▪ Stravinskij, Igor Fëdorovič
dbpprop:dateOfBirth	<ul style="list-style-type: none">▪ --06-17
dbpprop:dateOfDeath	<ul style="list-style-type: none">▪ 1971-04-06 (xsd:date)
dbpprop:filename	<ul style="list-style-type: none">▪ Igor Stravinsky - 3 Pieces for Clarinet Alone.ogg
dbpprop:format	<ul style="list-style-type: none">▪ dbpedia:Ogg
dbpprop:hasPhotoCollection	<ul style="list-style-type: none">▪ http://www4.wiwiwiss.fu-berlin.de/flickrwrappr/photos/Igor_Stravinsky
dbpprop:name	<ul style="list-style-type: none">▪ Stavinsky, Igor Fyodorovich
dbpprop:placeOfBirth	<ul style="list-style-type: none">▪ Lomonosov, Russia, Russia
dbpprop:placeOfDeath	<ul style="list-style-type: none">▪ New York City, New York, United States
dbpprop:shortDescription	<ul style="list-style-type: none">▪ Russian composer

KB's: Example DBpedia

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

About: Igor Stravinsky



An Entity of Type : [Neoclassical composers](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)

dcterms:subject

- [category:Honorary_Members_of_the_Royal_Philharmonic_Society](#)
- [category:People_from_Saint_Petersburg](#)
- [category:Russian_Orthodox_Christians](#)
- [category:Grammy_Award_winners](#)
- [category:Royal_Philharmonic_Society_Gold_Medallists](#)
- [category:People_from_Lomonosov](#)
- [category:Opera_composers](#)
- [category:Burials_at_Isola_di_San_Michele](#)
- [category:Russian_composers](#)
- [category:Ragtime_composers](#)
- [category:National_Museum_of_Dance_Hall_of_Fame_inductees](#)
- [category:Naturalized_citizens_of_the_United_States](#)
- [category:Soviet_immigrants_to_the_United_States](#)
- [category:20th-century_classical_composers](#)
- [category:Neoclassical_composers](#)
- [category:Russian_ballet](#)
- [category:Ballets_Russes_composers](#)
- [category:1882_births](#)
- [category:1971_deaths](#)
- [category:Harvard_University_people](#)
- [category:Grammy_Lifetime_Achievement_Award_winners](#)
- [category:Modernist_composers](#)
- [category:Ballet_composers](#)

rdf:type

- [foaf:Person](#)
- [yago:ModernistComposers](#)
- [yago:RussianComposers](#)
- [yago:BalletComposers](#)
- [yago:BalletsRussesComposers](#)
- [yago:Person100007846](#)
- [yago:OperaComposers](#)
- [yago:RagtimeComposers](#)
- [yago:HarvardUniversityPeople](#)
- [yago:NaturalizedCitizensOfTheUnitedStates](#)
- [yago:PeopleFromSaintPetersburg](#)
- [yago:RoyalPhilharmonicSocietyGoldMedallists](#)

KB's: Example DBpedia

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



About: Igor Stravinsky

An Entity of Type : [Neoclassical composers](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)

rdfs:label	<ul style="list-style-type: none">Igor StravinskyÍgor StravinskiIgor StravinskyIgor StravinskiIgor' Fëdorovič Stravinskijイーゴリ・ストラヴィンスキーIgor StravinskiIgor StravinskijÍgor StravinskiСтравинский, Игорь ФёдоровичIgor Stravinskij伊戈尔·费奥多罗维奇·斯特拉文斯基
owl:sameAs	<ul style="list-style-type: none">freebase:Igor Stravinskyhttp://zitgist.com/music/artist/c278de2c-9696-4fdf-a919-0781cd945e2chttp://data.nytimes.com/N1686270504070431933
foaf:depiction	<ul style="list-style-type: none">http://upload.wikimedia.org/wikipedia/en/3/33/Igor_Stravinsky_LOC_32392u.jpg
foaf:givenName	<ul style="list-style-type: none">Igor Fyodorovich
foaf:name	<ul style="list-style-type: none">Igor Fyodorovich Stavinsky
foaf:page	<ul style="list-style-type: none">http://en.wikipedia.org/wiki/Igor_Stravinsky
foaf:surname	<ul style="list-style-type: none">Stavinsky
is dbpedia-owl:influencedBy of	<ul style="list-style-type: none">dbpedia:Nicholas_Lensdbpedia:Carla_Lucero
is dbpedia-owl:musicComposer of	<ul style="list-style-type: none">dbpedia:La_Belle_Noiseusedbpedia:Melinda_and_Melindadbpedia:Histoire(s)_du_cinéma
is dbpedia-owl:wikiPageDisambiguates of	<ul style="list-style-type: none">dbpedia:Igordbpedia:Stravinsky_(disambiguation)
is dbpedia-owl:wikiPageRedirects of	<ul style="list-style-type: none">dbpedia:Stravinskydbpedia:Igor_Stravinskijdbpedia:Igor_Strawinskydbpedia:Igor_Fyodorovitch_Stravinskydbpedia:Stravinsky_Igor_Fedorovich

KB's: Example NELL (Carlson, Mitchell, et al.: WSDM'10, AAAI'10)

NELL Knowledge Base Browser

CMU Read the Web Project

log in | preferences | help/instructions | feedback

Search

- person
 - actor
 - athlete
 - celebrity
 - ceo
 - chef
 - coach
 - female
 - journalist
 - male
 - mlauthor
 - musician
 - politician
 - politicianus
 - scientist
 - visualartist
 - writer
 - astronaut
 - architect
 - comedian
 - model
 - judge
 - criminal
 - monarch
 - personbylocation
 - personafrika
 - personantarctica
 - personasia
 - personaustralia
 - personeurope
 - personnorthamerica
 - personus
 - politicianus
 - personmexico
 - personcanada
 - personsouthamerica
 - director
 - professor
 - location
 - website
 - blog (+)
 - building
 - hotel
 - monument
 - museum
 - restaurant
 - stadiumoreventvenue
 - airport
 - bridge

writer

(category)

See [metadata](#) for writer
9,479 instances, 2 pages: 1 [2](#) [next](#)

instance

nora_roberts	
jane_yolen	
jorge_luis_borges	
adrienne_rich	
aeschylus	
agatha_christie	
alan_moore	
albert_camus	
aldous_huxley	
alexander_pushkin	
alexandre_dumas	Alexander Pushkin
alice_hoffman	
alice_sebold	
alice_walker	
allen_ginsberg	
amy_tan	
anne_bishop	
anne_rice	
anthony_burgess	
anton_chekhov	
aristophanes	
aristotle	
arthur_agatston	
arthur_conan_doyle	
arthur_c_clarke	
arthur_c_clarke	
arthur_miller	
augustine	
austen	
author	
author_charles_dickens	
author_michael_crichton	
ayn_rand	

A writer is a person who produces written work. This category includes book authors, poets, and playwrights.

- 800 000 assertions (on entity names & relations)
- 800 classes & relations
- extracted from Web pages
- continuously growing

245	28-apr-2011	100.0
0	28-oct-2010	(Seed) 100.0
242	24-apr-2011	100.0
216	05-mar-2011	100.0
222	21-mar-2011	100.0
216	05-mar-2011	100.0
216	05-mar-2011	100.0
216	05-mar-2011	100.0
227	03-apr-2011	100.0
1	12-jan-2010	100.0
224	26-mar-2011	100.0
233	13-apr-2011	100.0
216	05-mar-2011	100.0
244	27-apr-2011	100.0
0	28-oct-2010	(Seed) 100.0
216	05-mar-2011	100.0
216	05-mar-2011	100.0
216	05-mar-2011	100.0
216	05-mar-2011	100.0
242	24-apr-2011	100.0
216	05-mar-2011	100.0
216	05-mar-2011	100.0
30	23-jan-2010	100.0
7	12-jan-2010	100.0

<http://rtw.ml.cmu.edu/rtw/kbbrowser/>

KB's: Example NELL (Carlson, Mitchell, et al.: WSDM'10, AAAI'10)

NELL Knowledge Base Browser

CMU Read the Web Project

 Search





- person
 - actor
 - athlete
 - celebrity
 - ceo
 - chef
 - coach
 - female
 - journalist
 - male
 - mlaauthor
 - musician
 - politician
 - politicianus
 - scientist
 - visualartist
 - writer
 - astronaut
 - architect
 - comedian
 - model
 - judge
 - criminal
 - monarch
 - personbylocation
 - personafrica
 - personantarctica
 - personasia
 - personastralia
 - personeurope
 - personnorthamerica
 - personus
 - politicianus
 - personmexico
 - personcanada
 - personsouthamerica
 - director
 - professor
 - location
 - website
 - blog (+)
 - building
 - hotel
 - monument
 - museum
 - restaurant
 - stadiumoreventvenue
 - airport
 - bridge

alexander_pushkin (writer)

literal strings: [Alexander Pushkin](#), [alexander pushkin](#)

Help NELL Learn!

NELL wants to know if
If they are or ever were, click thum

- [alexander_pushkin](#) is a [writer](#)  
- [alexander_pushkin](#) is a [European person](#)  
- the [latitude and longitude](#) of [alexander_pushkin](#) is [38.899560000](#)

categories

- [writer](#)(100.0%)
 - SEAL @190 (75.0%) on 16-jan-2011 [[1](#) [2](#)] using alexande
 - CMC @308 (99.9%) on 20-jun-2011 [PREFIX=ale FULL_V
 - CPL @215 (100.0%) on 26-feb-2011 ["novel by _" "novella _" _'s Eugene Onegin" "authors such as _" _'s novella" "short story by _" "same name by _" "favorite writers are _"
- [person](#)(100.0%)
 - SEAL @227 (75.0%) on 01-apr-2011 [[1](#) [2](#)] using alexande
- [personeurope](#)(99.9%)
 - SEAL @180 (93.8%) on 16-dec-2010 [[1](#) [2](#) [3](#) [4](#)] using alexa
 - CPL @215 (93.8%) on 26-feb-2011 ["novella by _" "narrati alexander_pushkin
 - CMC @182 (77.7%) on 04-jan-2011 [WORDSHAPE=aaaa: alexander_pushkin

NELL has only weak ev

- [scientist](#)
 - CPL @196 (87.5%) on 02-feb-2011 ["famous pupil was _"

relations

- [latitudeandlongitude](#)
 - [38.8995600000000-77.0448888888889](#) (100.0%)
 - LatLong @170 (1 statue@38.900,-7

search results for: 'Pushkin'

- [pushkin](#) (museum)
- [pushkin](#) (writer)
- [a_s_pushkin](#) (scientist)
- [a_s_pushkin](#) (writer)
- [a_pushkin](#) (writer)
- [alexander_pushkin](#) (monarch)
- [alexander_pushkin](#) (musician)
- [alexander_pushkin](#) (writer)
- [pushkin_museum](#) (museum)
- [aleksandr_pushkin](#) (writer)
- [alexander_s_pushkin](#) (writer)
- [aleksandr_pushkin](#) (personeurope)
- [aleksandr sergeyevich_pushkin](#) (scientist)
- [poet_alexander_pushkin](#) (writer)
- [kelly_lynn_pushkin](#) (celebrity)
- [aleksei_musin_pushkin](#) (monarch)
- [alexander_pushkin_statue](#) (monument)
- [pushkin_fine_arts_museum](#) (museum)
- [the_pushkin_museum_of_fine_arts](#) (museum)
- [pushkin_museum_of_fine_arts](#) (museum)
- [russian_poet_alexander_pushkin](#) (personeurope)
- [each_year_mary_and_alexander_pushkin](#) (monarch)
- [a_s_pushkin_the_complete_collection_of_works](#) (visual
- [alexander_pushkin_museum_of_fine_arts](#) (museum)
- [pushkin_state_museum_of_fine_arts](#) (museum)
- [pushkinsky_bridge](#) (bridge)
- [katya_pushkina](#) (model)
- [pushkinsky_pedestrian_bridge](#) (bridge)

<http://rtw.ml.cmu.edu/rtw/kbbrowser/>

Outline

✓ **Motivation**

✓ **Machine Knowledge**

★ **Knowledge Harvesting**

- **Entities and Classes**
- **Relational Facts**

★ **Research Challenges**

- **Open-Domain Extraction**
- **Temporal Knowledge**

★ **Wrap-up**

WordNet Thesaurus [Miller/Fellbaum 1998]

WordNet Search - 3.0 - [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

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

) 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*"

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

subclasses
(hyponyms)

superclasses
(hypernyms)

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)

Sense 1

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 who

the Constitution; he played a major role in the American

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 physi

ovum produced by the female of the species (1578-165

"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 [\[hide\]](#)

- [1 Family and education](#)
- [2 Work](#)
- [3 Disappearance at sea and search](#)
- [4 Books](#)
- [5 See also](#)
- [6 References](#)
- [7 External links](#)

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 [\[hide\]](#)

- 1 Life and career
 - 1.1 Academic career
 - 1.2 Family
 - 1.3 Professor at Berlin University
 - 1.4 Black-body radiation
 - 1.5 Einstein and the theory of relativity
 - 1.6 World War and Weimar Republic
 - 1.7 Quantum mechanics
 - 1.8 Nazi dictatorship and The Second World War
- 2 Religious view
- 3 Honors and awards

Max Planck

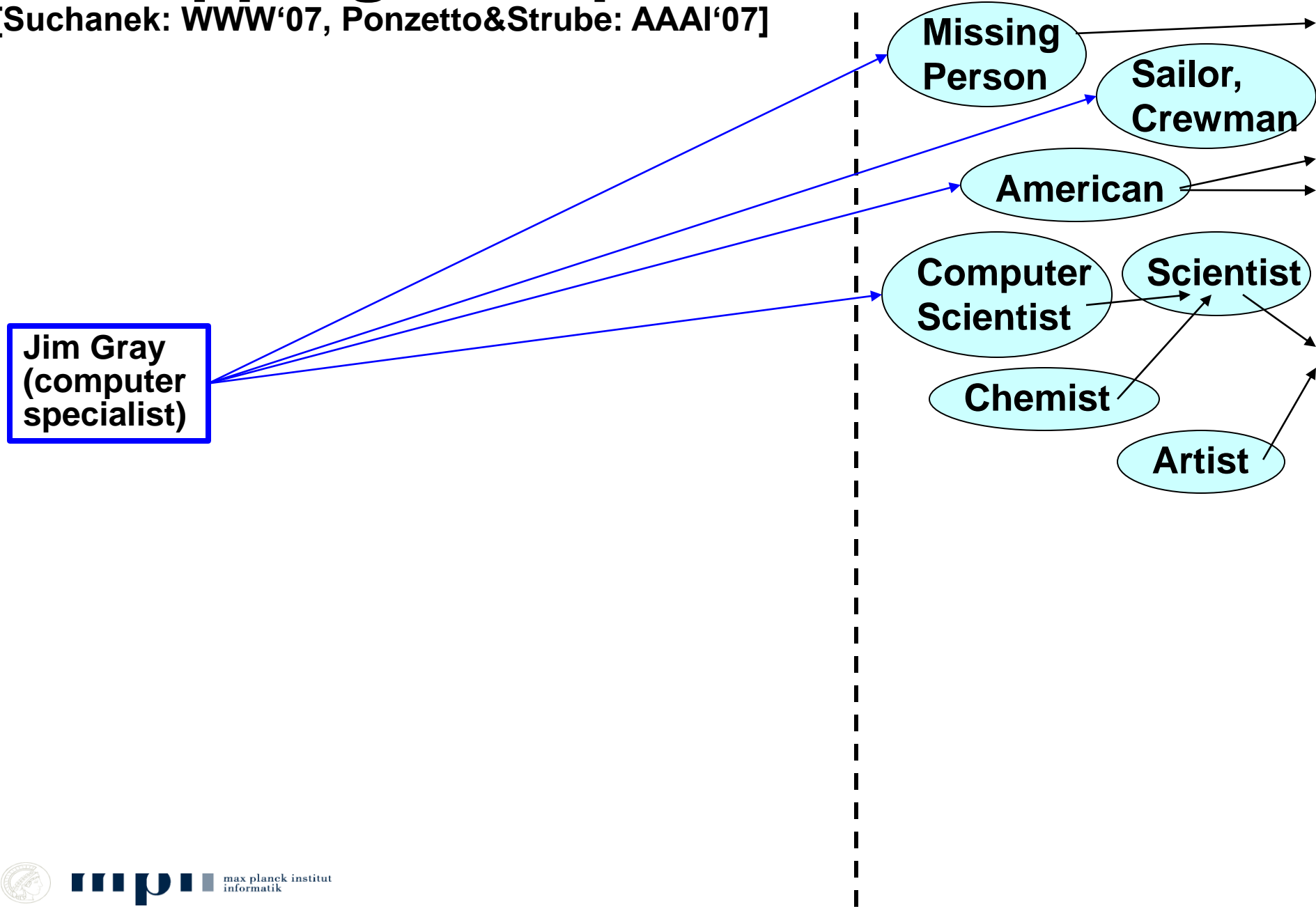


Born	April 23, 1858 Kiel, Holstein
Died	October 4, 1947 (aged 89) Göttingen, West Germany
Nationality	German
Fields	Physics
Institutions	University of Kiel University of Berlin University of Göttingen Kaiser-Wilhelm-Gesellschaft
Alma mater	Ludwig Maximilian University of

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

Mapping: Wikipedia → WordNet

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



Mapping: Wikipedia → WordNet

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



Jim Gray
(computer specialist)

instanceOf

People
Lost at Sea

Computer
Scientists
by Nation

American
Computer
Scientists

Database
Researcher

Fellows of
the ACM

Databases

Engineering
Societies

ACM

Members
of Learned
Societies

Missing
Person

American

Computer
Scientist

Scientist

Database

Fellow (1),
Comrade

Fellow (2),
Colleague

Fellow (3)
(of Society)

Member (1),
Fellow

Member (2),
Extremity

name similarity
(edit dist., n-gram overlap) ?
context similarity
(word/phrase level) ?
machine learning ?

Mapping: Wikipedia → WordNet

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

Given: entity **e** in Wikipedia categories **c_1, \dots, c_k**
Wanted: **instanceOf(e,c)** and **subclassOf(c_i, c)** for WN class **c**
Problem: vagueness & ambiguity of names **c_1, \dots, c_k**

Analyzing category names → **noun group parser**:

American Musicians of Italian Descent
pre-modifier head post-modifier

American Folk Music of the 20th Century
pre-modifier head post-modifier

American Indy 500 Drivers on Pole Positions
pre-modifier head post-modifier

Head word is key, should be in **plural** for instanceOf

Mapping Wikipedia Entities to WordNet Classes

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

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

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

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

Heuristic Method:

for each c_i do

if **head word w** of category name c_i is plural

{

1) **match w against synsets** of WordNet classes

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

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

}

tuned conservatively: high precision, reduced recall

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

Learning More Mappings [Wu & Weld: 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

Long Tail of Class Instances

[Discuss](#) [Terms of Use](#)



Automatically create sets of items from a few examples.

Enter a few items from a set of things. ([example](#))

Next, press *Large Set* or *Small Set* and we'll try to predict other items in the set.

-
-
-
-
-

([clear all](#))

Examples:

[green, purple, red](#) [chicken dance, macarena, ymca](#) [alexander, gladiator, troy](#) [hilary duff, kelly clarkson](#) [more...](#)

labs.google.com - [All About Google](#)

©2007 Google

Long Tail of Class Instances

Predicted Items
tolstoy
pushkin
leo tolstoy
anna karenina
gogol
drama
danielle steel
dostoevsky
maxim gorky
russia
fyodor dostoevsky
anton chekhov
ivan turgenev
paulo coelho
dan brown
ernest hemingway
dostojevski
alexander pushkin

john steinbeck
russian literature
lermontov
stephen king
cs lewis
madame bovary
bible
the idiot
mark twain
mikhail bulgakov
fyodor dostoyevsky
nikolai gogol
susanna tamaro
edward said
dirty dancing
albert camus
shakespeare
romance novel
jack london
george orwell
fiction
authors

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**

But:

Precision drops for classes with **sparse statistics** (IR profs, ...)

Harvested items are **names**, **not entities**

Canonicalization (de-duplication) unsolved

Outline

✓ **Motivation**

✓ **Machine Knowledge**

★ **Knowledge Harvesting**

- **Entities and Classes**
- **Relational Facts**

★ **Research Challenges**

- **Open-Domain Extraction**
- **Temporal Knowledge**

★ **Wrap-up**

Tapping on Wikipedia Infoboxes

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 [\[hide\]](#)

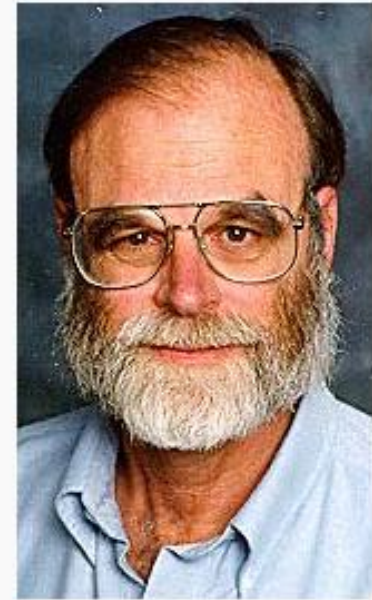
- [1 Family and education](#)
- [2 Work](#)
- [3 Disappearance at sea and search](#)
- [4 Books](#)
- [5 See also](#)
- [6 References](#)
- [7 External links](#)

harvest by
extraction rules:

- regex matching
- type checking

`(?i)IBL\|BEG\s*awards\s*=\s*(.*?)IBL\|END"`
`=> "$0 hasWonPrize @WikiLink($1)"`

James Nicholas "Jim" Gray



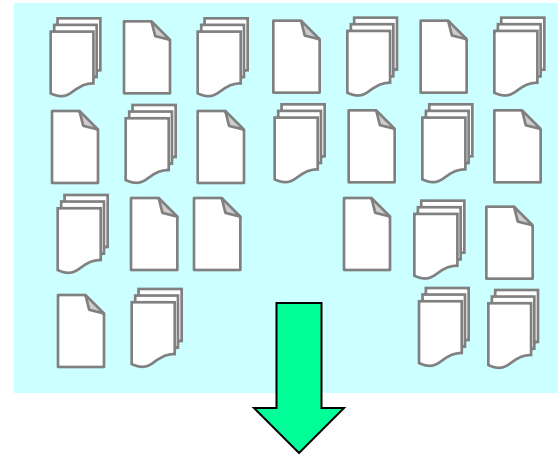
Born	January 12, 1944 ^[1] San Francisco, California ^[2]
Died	(lost at sea) January 28, 2007
Nationality	American
Fields	Computer Science
Institutions	IBM, Tandem Computers, DEC, Microsoft
Alma mater	University of California, Berkeley
Doctoral advisor	Michael Harrison ^[2]
Known for	Work on database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

French Marriage Problem



facts in KB:

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



new facts or fact candidates:

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

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

Pattern-Based Harvesting

(Hearst 92, Brin 98, Agichtein 00, Etzioni 04, ...)

Facts & Fact Candidates

Patterns

(Hillary, Bill)

(Carla, Nicolas)

(Angelina, Brad)

(Victoria, David)

(Hillary, Bill)

(Carla, Nicolas)

(Yoko, John)

(Kate, Pete)

(Carla, Benjamin)

(Larry, Google)

(Angelina, Brad)

(Victoria, David)

X and her husband Y

X and Y on their honeymoon

X and Y and their children

X has been dating with Y

X loves Y

...

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

Reasoning about Fact Candidates

Use **consistency constraints** to prune false candidates

FOL rules (restricted):

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

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

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

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

ground atoms:

$\text{spouse}(\text{Hillary}, \text{Bill})$
 $\text{spouse}(\text{Carla}, \text{Nicolas})$
 $\text{spouse}(\text{Cecilia}, \text{Nicolas})$
 $\text{spouse}(\text{Carla}, \text{Ben})$
 $\text{spouse}(\text{Carla}, \text{Mick})$
 $\text{spouse}(\text{Carla}, \text{Sofie})$

$f(\text{Hillary})$	$m(\text{Bill})$
$f(\text{Carla})$	$m(\text{Nicolas})$
$f(\text{Cecilia})$	$m(\text{Ben})$
$f(\text{Sofie})$	$m(\text{Mick})$

Rules reveal inconsistencies

Find **consistent subset(s)** of atoms
("possible world(s)", "the truth")

Rules can be **weighted**

(e.g. by fraction of ground atoms that satisfy a rule)

→ **uncertain / probabilistic data**

→ compute prob. distr. of subset of atoms being the truth

Markov Logic Networks (MLN's)

(M. Richardson / P. Domingos 2006)

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

$s(x,y) \wedge \text{diff}(y,z) \Rightarrow \neg s(x,z)$

$s(x,y) \wedge \text{diff}(w,y) \Rightarrow \neg s(w,y)$

$s(x,y) \Rightarrow f(x)$

$s(x,y) \Rightarrow m(y)$

$f(x) \Rightarrow \neg m(x)$

$m(x) \Rightarrow \neg f(x)$

$s(\text{Carla}, \text{Nicolas})$

$s(\text{Cecilia}, \text{Nicola})$

$s(\text{Carla}, \text{Ben})$

$s(\text{Carla}, \text{Sofie})$

...

Grounding:

$\neg s(\text{Ca}, \text{Nic}) \vee \neg s(\text{Ce}, \text{Nic})$

$\neg s(\text{Ca}, \text{Nic}) \vee \neg s(\text{Ca}, \text{Ben})$

$\neg s(\text{Ca}, \text{Nic}) \vee \neg s(\text{Ca}, \text{So})$

$\neg s(\text{Ca}, \text{Ben}) \vee \neg s(\text{Ca}, \text{So})$

$\neg s(\text{Ca}, \text{Ben}) \vee \neg s(\text{Ca}, \text{So})$

Literal \rightarrow Boolean Var
Literal \rightarrow binary RV

$\neg s(\text{Ca}, \text{Nic}) \vee m(\text{Nic})$

$\neg s(\text{Ce}, \text{Nic}) \vee m(\text{Nic})$

$\neg s(\text{Ca}, \text{Ben}) \vee m(\text{Ben})$

$\neg s(\text{Ca}, \text{So}) \vee m(\text{So})$

Markov Logic Networks (MLN's)

(M. Richardson / P. Domingos 2006)

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

$s(x,y) \wedge \text{diff}(y,z) \Rightarrow \neg s(x,z)$

$s(x,y) \Rightarrow f(x)$

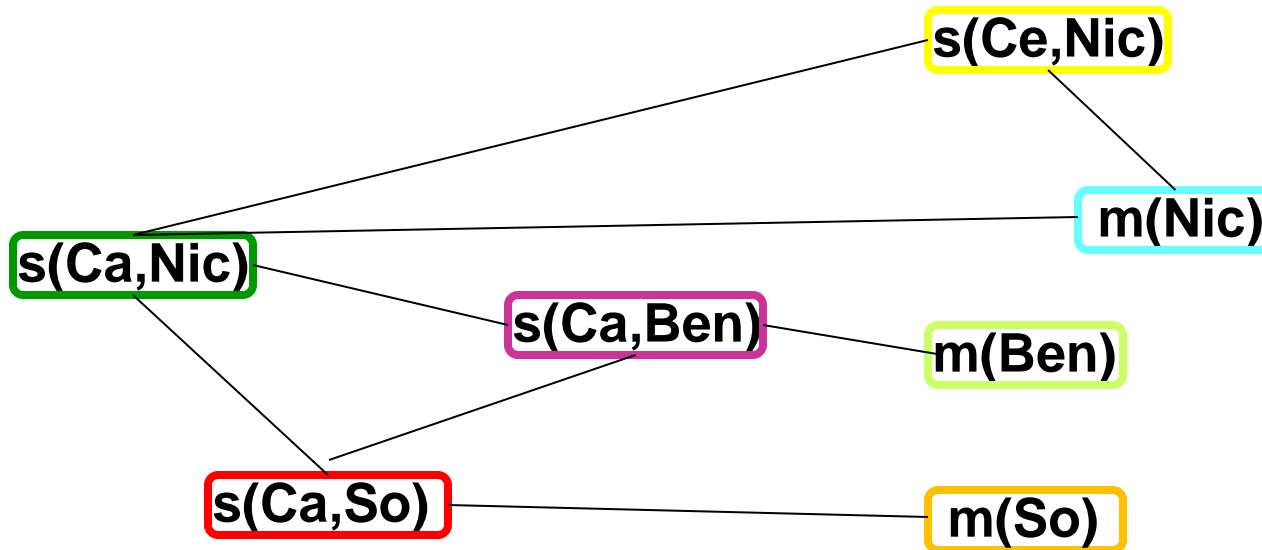
$f(x) \Rightarrow \neg m(x)$

$s(x,y) \wedge \text{diff}(w,y) \Rightarrow \neg s(w,y)$

$s(x,y) \Rightarrow m(y)$

$m(x) \Rightarrow \neg f(x)$

$s(\text{Carla}, \text{Nicolas})$
 $s(\text{Cecilia}, \text{Nicola})$
 $s(\text{Carla}, \text{Ben})$
 $s(\text{Carla}, \text{Sofie})$
...



RVs coupled
by MRF edge
if they appear
in same clause

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

Variety of algorithms for joint inference:
Gibbs sampling, other MCMC, belief propagation,
randomized MaxSat, ...

Related Alternative Probabilistic Models

Constrained Conditional Models [D. Roth et al. 2007]

log-linear classifiers with constraint-violation penalty
mapped into Integer Linear Programs

Factor Graphs with Imperative Variable Coordination

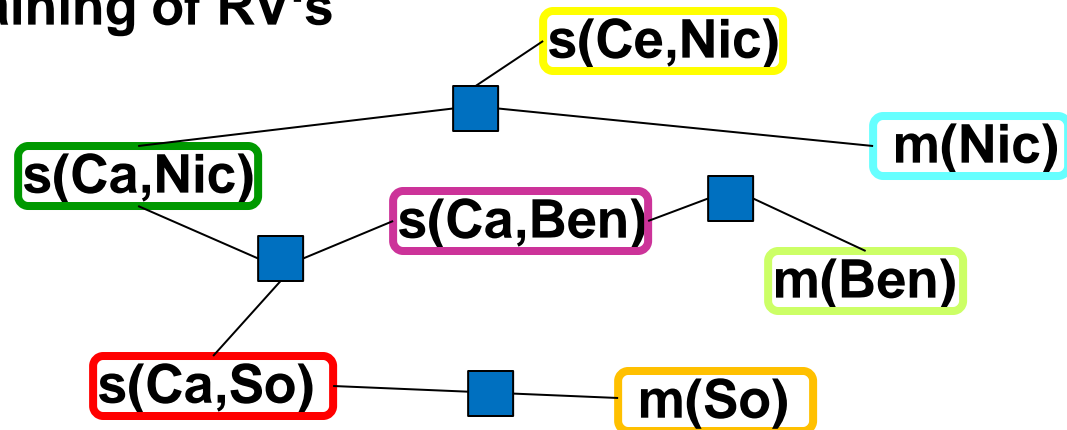
[A. McCallum et al. 2008]

RV's share "factors" (joint feature functions)

generalizes MRF, BN, CRF, ...

inference via advanced MCMC

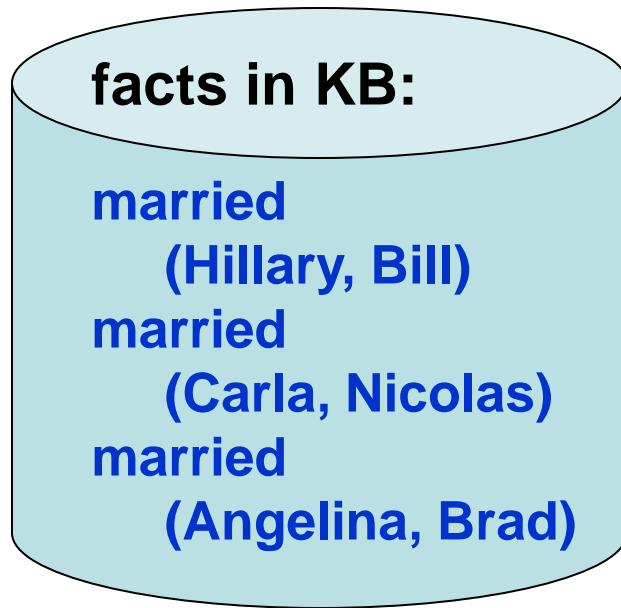
flexible coupling & constraining of RV's



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

Reasoning for KB Growth: Direct Route

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



+

new fact candidates:

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

?

patterns:

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

Direct approach:

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

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

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

Facts & Patterns Consistency with SOFIE

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

constraints to connect facts, fact candidates, patterns

pattern-fact duality:

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

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

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

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:

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

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

Pattern Harvesting Revisited

(N. Nakashole et al.: WebDB'10,)

using narrow &
dropping nasty
loses **recall** !

narrow / nasty / noisy patterns:

X and his famous advisor Y

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

X jointly developed the method with Y

using noisy
loses **precision** &
slows down MaxSat

POS-lifted n-gram itemsets as patterns:

X { his doctoral research, under the supervision of } Y

X { PRP ADJ advisor } Y

X { PRP doctoral research, IN DET supervision of } Y

confidence weights, using seeds and counter-seeds:

seeds: (ThomasHofmann, JoachimBuhmann), (JimGray, MikeHarrison)

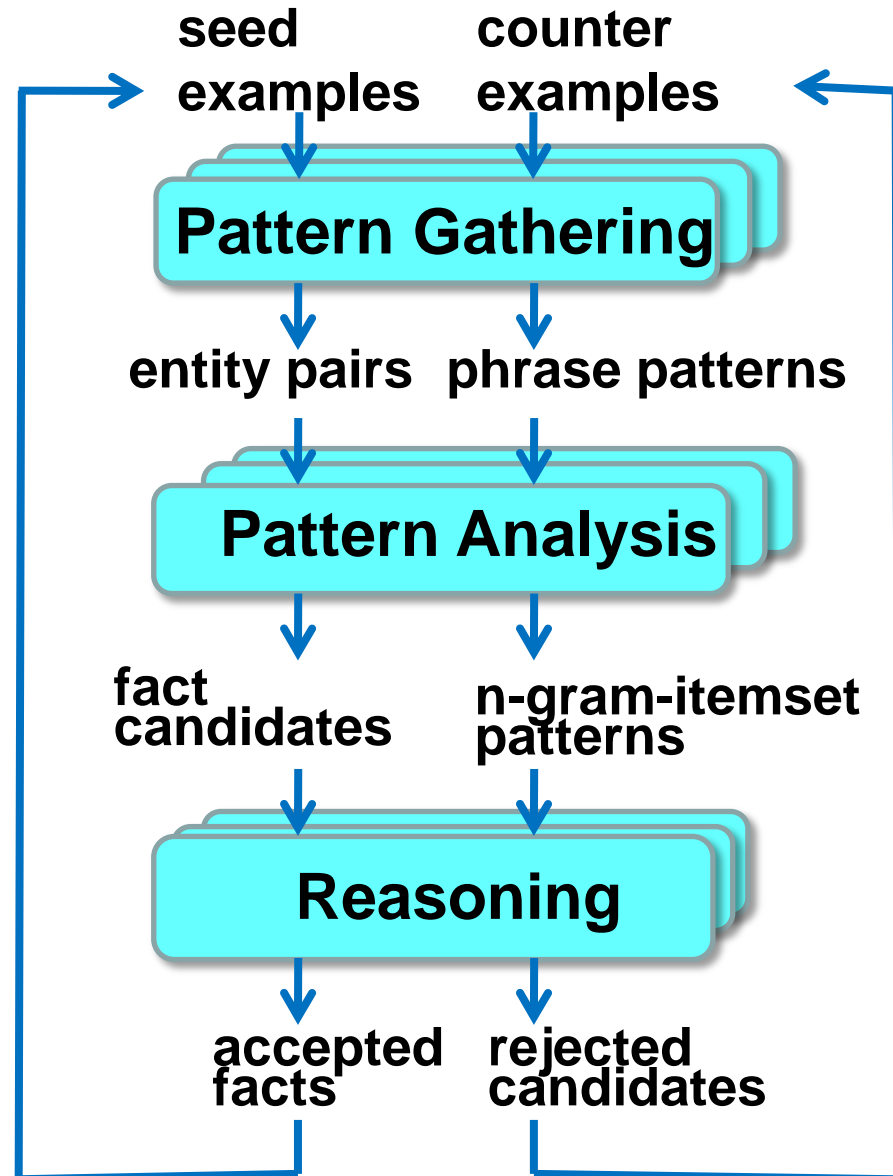
counter-seeds: (BernhardSchölkopf, AlexSmola), (AlonHalevy, LarryPage)

→ confidence of pattern $p \sim \#p \text{ with seeds} - \#p \text{ with counter-seeds}$

PROSPERA: Prospering Knowledge with Scalability, Precision, Recall

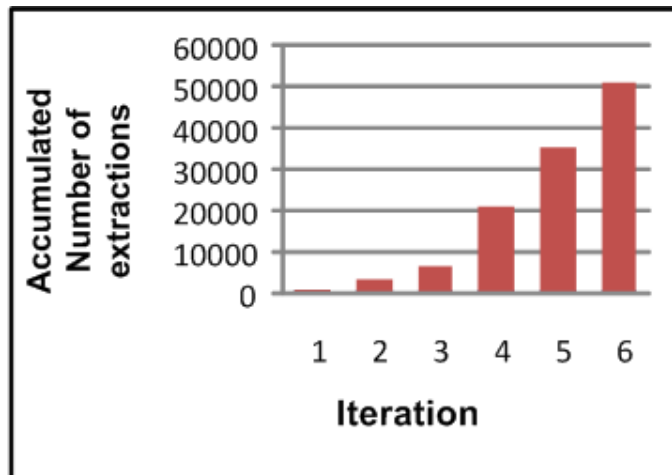
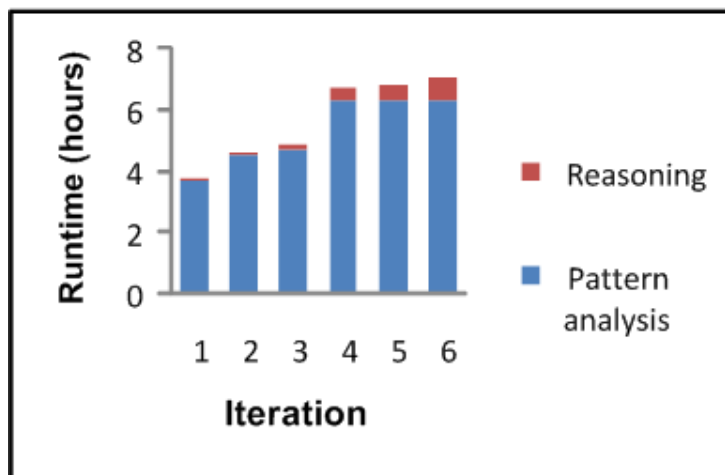
(N. Nakashole et al.: WSDM'11)

- feedback loop for higher recall
- all stages parallelizable on MapReduce platform



Web-Scale Experiments [N. Nakashole et al.: WSDM'11]

- on ClueWeb'09 corpus (500 Mio. English Web pages)
- with Hadoop cluster of 10x16 cores and 10x48 GB memory
- 10 seed examples, 5 counter examples for each relation



Relation	PROSPERA			ReadTheWeb [CMU]	
	#Facts	Precision	Prec@1000	#Facts	Precision
AthletePlaysForTeam	14685	82%	100%	456	100%
TeamPlaysAgainstTeam	15170	89%	100%	1068	99%
TeamMate	19666	86%	100%	---	---
FacultyAt	4394	96%	100%	---	---

Outline

✓ **Motivation**

✓ **Machine Knowledge**

✓ **Knowledge Harvesting**

- **Entities and Classes**
- **Relational Facts**

★ **Research Challenges**

- **Open-Domain Extraction**
- **Temporal Knowledge**

★ **Wrap-up**

Discovering New Relation Types

Targeted (Domain-Oriented) Gathering of Facts: Entity × Relation × Entity

< Carla_Bruni marriedTo Nicolas_Sarkozy >,
< Natalie_Portman wonAward Academy_Award >,
...

Explorative (Open-Domain) Gathering of „Assertions“: Name × Pattern × Name

< „Carla Bruni“ „had affair with“ „Mick Jagger“ >,
< „First Lady Carla“ „had affair with“ „Stones singer Mick“ >,
< „Madame Bruni“ „happy marriage with“ „President Sarkozy“ >,
< „Jeff Bridges“ „expected to win“ „Oscar“ >,
< „Coen Brothers“ „celebrated for“ „Oscar Award“ >,
...

Open-Domain Gathering of Assertions

[O. Etzioni et al. 2007, F. Wu et al. 2010]

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 (CRF)** 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** candidates for **new relation types** and their **facts**
{*datesWith*, *partiesWith*}, {*affairWith*, *flirtsWith*}, {*romanticRelation*}, ...

But: result is **noisy**
clusters are **not canonicalized** relations
far from near-human-quality

Open IE Example: TextRunner / ReVerb



ReVerb Search

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

ReVerb took 13.32 seconds.

Retrieved **242** results for Predicate containing "**has won**" and Argument 2 containing "**prize**"

Grouping results by predicate. Group by: [argument 2](#) | [argument 1](#)

has won (157 results)

book (28), film (25), no one (25), **50 more... has won prizes**
Al Gore (41), IPCC (6), Carter (6), **9 more... has won the Nobel Peace Prize**
Al Gore (15), highly successful novelist (6), Chinese writer (5), **11 more... has won the Nobel Prize**
book (11), author (3), scholarly work (2), **10 more... has won the Pulitzer Prize**
Coetzee (5), book (4), Peter Carey (2), **2 more... has won the Booker Prize**
Artist Mark Wallinger (3), Mark Wallinger (2), Simon Starling (2) **has won the Turner Prize**
James Ehnes (2), Work (2), school (2) **has won numerous awards and prizes**
recipient (7), Tickets (2) **has won a cash prize**
Katherine Taylor (2), His poetry (2) **has won a Pushcart Prize**
Bryan D. Dietrich (3) **has won the Paris Review Poetry Prize**
Woodward (3) **has won the Wynne Prize**
British playwright Harold Pinter (2) **has won the 2005 Nobel Literature Prize**
Al Gore (2) **has won the 2007 Nobel Peace Prize**
Argentine poet Juan Gelman (3) **has won the Cervantes prize**
Mr Bean (2) **has won a church fete raffle 's top prize**
No woman (2) **has ever won the Economics Prize**
film (2) **has already won five international prizes**
No other journalist (2) **has won the Ford Prize**
Alice Munro (3) **has won the Giller Prize**
Robert Nye (2) **has won the Hawthornden Prize**
his translator Daniel Hahn (2) **have won the Independent Foreign Fiction Prize 2007**

Search again:

Argument 1

Predicate

has won

Argument 2

prize

Jump to:

[has won \(157\)](#)
[will have a chance to win \(16\)](#)
[have a chance to win \(17\)](#)
[will have the opportunity to win \(15\)](#)
[has an equal chance of winning \(6\)](#)
[have the opportunity to win \(7\)](#)
[has won a number of \(5\)](#)
[should have won \(4\)](#)
[has become the first woman to win \(2\)](#)
[has a chance at winning \(1\)](#)
[must not have won \(2\)](#)
[has not yet won \(2\)](#)
[have gone on to win \(1\)](#)
[have the opportunity to win one of \(1\)](#)
[had a fair shot at winning \(1\)](#)

Open IE Example: TextRunner / ReVerb

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

ReVerb took 19.89 seconds.

Retrieved **296** results for Predicate containing "**wrote**"

Grouping results by predicate. Group by: [argument 2](#) | [argument 1](#)

wrote (222 results)

poet (36), Students (31), people (17), **197 more... wrote** a poem
Pablo Neruda (2), African-American poet (2), Sappho (2) **wrote** love poems
Shakespeare (11) **wrote** plays and poems
Dickinson (5), Emily Dickinson (2) **wrote** 1,775 poems
hundreds of Asian American men (4) **are writing** books and poems
students (2), Marion Brown (2), Stuart (2) **began writing** stories and poems
Keats (2) **wrote** 150 poems
Students (2) **also wrote** letters and poems
Shakespeare (2) **wrote** many poems and plays
Computer hackers (2) **had written** source code poems
Emily Dickinson (3) **wrote** 1800 poems
Students (2) **wrote** a Fall poem
students (3) **then wrote** Haiku poems
Scottish poet (2) **wrote** many many poems and songs
Ginsberg (2) **wrote** the classic long poem Kaddish
Have students (3) **wrote** poems or stories
Composers (2) **wrote** tone poems
Tennyson (2) **wrote** two versions of the poem

can write (9 results)

Anyone (7), students (5), Children (8), **6 more... can write** a poem

will write (5 results)

Students (10), someone (2) **will write** an acrostic poem
Students (5) **will write** two poems
Students (2) **will write** a haiku poem

Search again:

Argument 1

Predicate

wrote

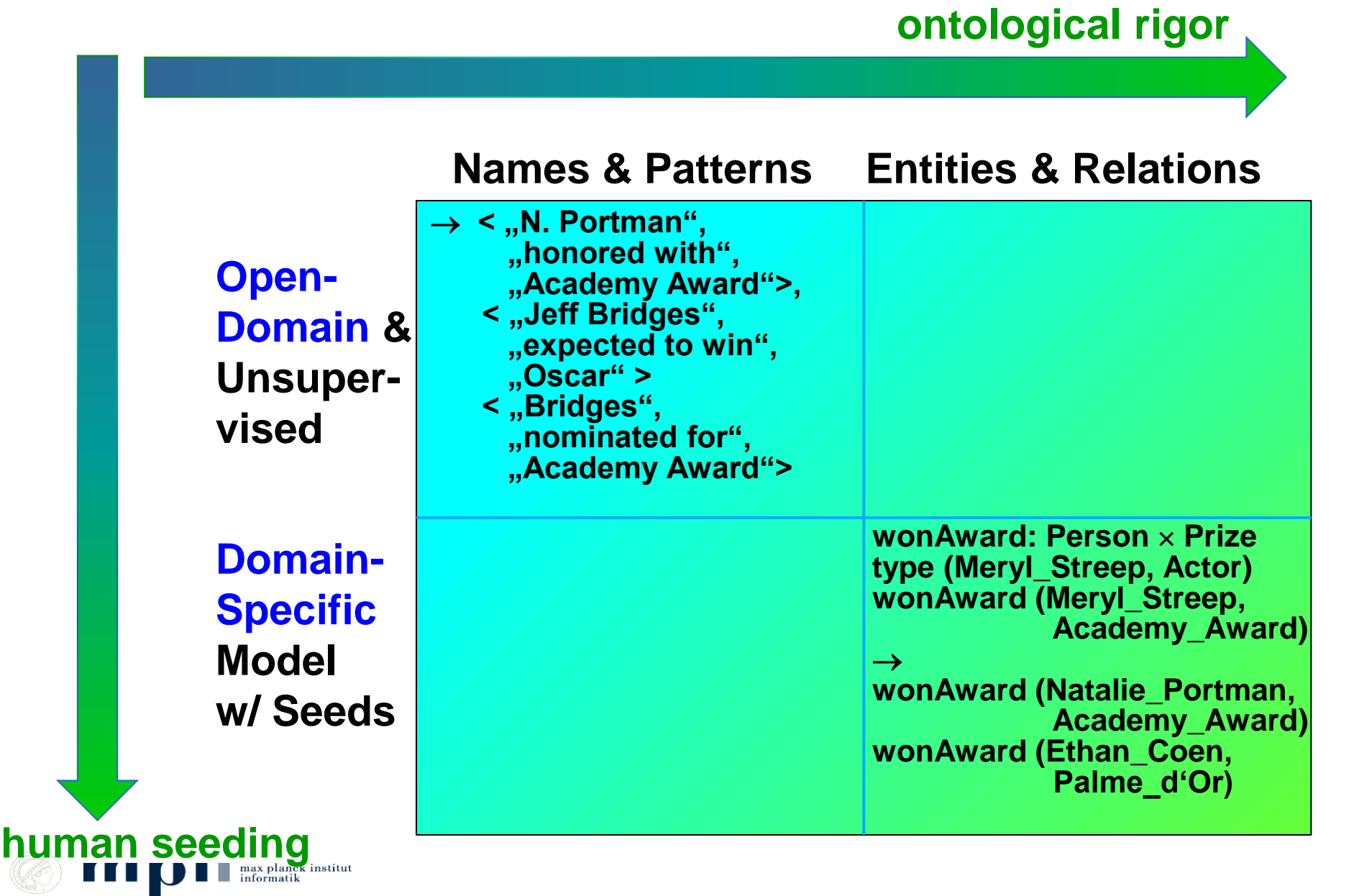
Argument 2

poem

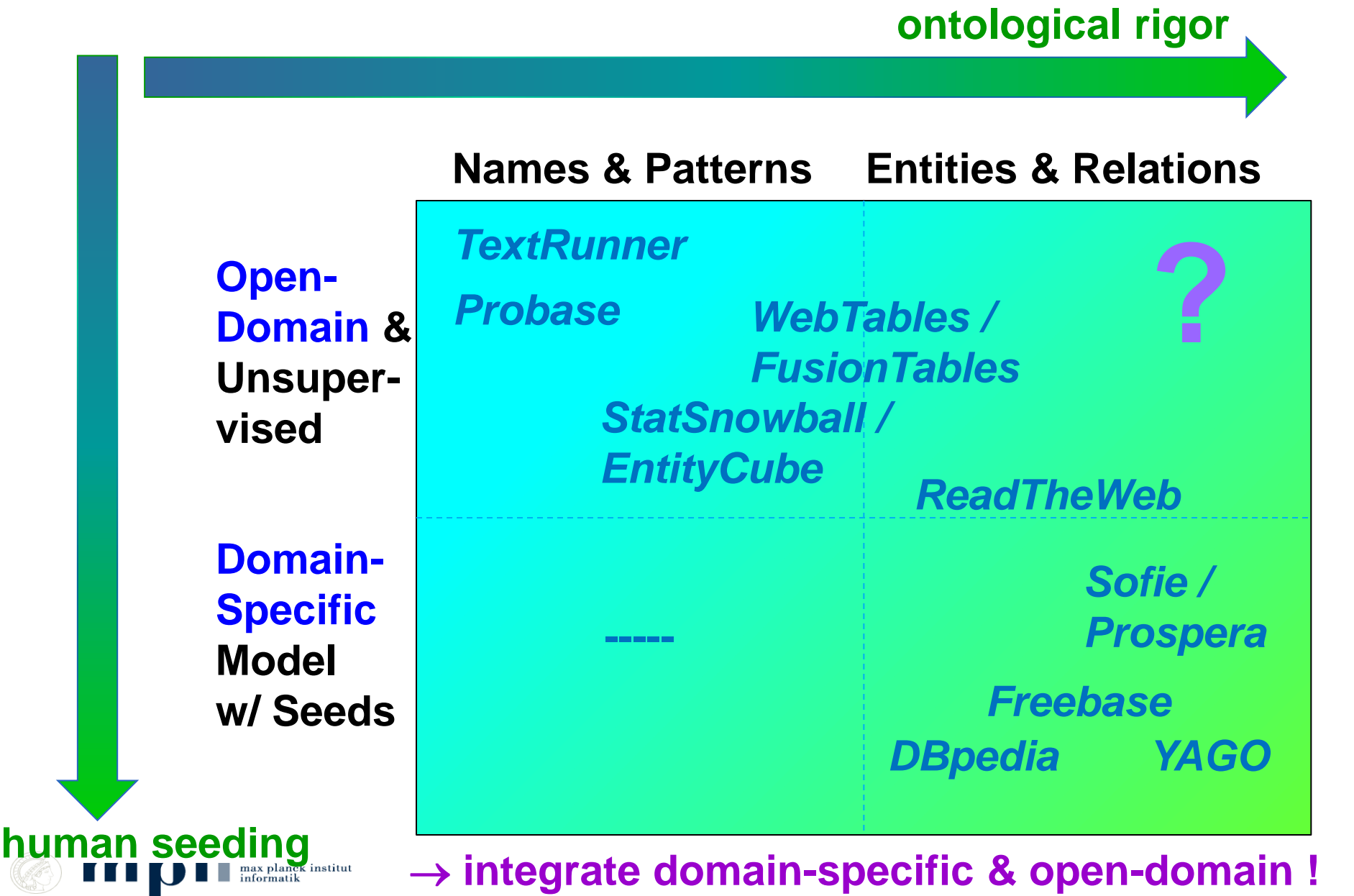
Jump to:

[wrote \(222\)](#)
[can write \(9\)](#)
[will write \(5\)](#)
[is written as \(5\)](#)
[wrote in \(9\)](#)
[does not write \(2\)](#)
[wrote some of \(4\)](#)
[has been writing \(2\)](#)
[has written thousands of \(1\)](#)
[would write \(2\)](#)
[write all kinds of \(2\)](#)
[wrote more than \(2\)](#)
[wrote over \(2\)](#)
[wrote a total of \(1\)](#)
[was written to \(1\)](#)
[can write the first line of \(1\)](#)
[always enjoyed writing \(1\)](#)
[have been writing lots of \(1\)](#)
[had begun writing \(1\)](#)
['ve never written \(2\)](#)
[started writing \(1\)](#)
[will write a variety of \(1\)](#)
[wrote hundreds of \(2\)](#)
[wrote scores of \(1\)](#)
[wrote thousands of \(1\)](#)
[wrote a variety of \(1\)](#)
[were asked to write \(1\)](#)

Challenge: Unify Targeted & Explorative Methods



Challenge: Unify Targeted & Explorative Methods



Outline

✓ **Motivation**

✓ **Machine Knowledge**

✓ **Knowledge Harvesting**

- **Entities and Classes**
- **Relational Facts**

★ **Research Challenges**

- **Open-Domain Extraction**
- **Temporal Knowledge**

★ **Wrap-up**

As Time Goes By: Temporal Knowledge

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

marriedTo (Madonna, Guy) [22Dec2000, Dec2008]

capitalOf (Berlin, Germany) [1990, now]

capitalOf (Bonn, Germany) [1949, 1989]

hasWonPrize (JimGray, TuringAward) [1998]

graduatedAt (HectorGarcia-Molina, Stanford) [1979]

graduatedAt (SusanDavidson, Princeton) [Oct 1982]

hasAdvisor (SusanDavidson, HectorGarcia-Molina) [Oct 1982, forever]

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

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

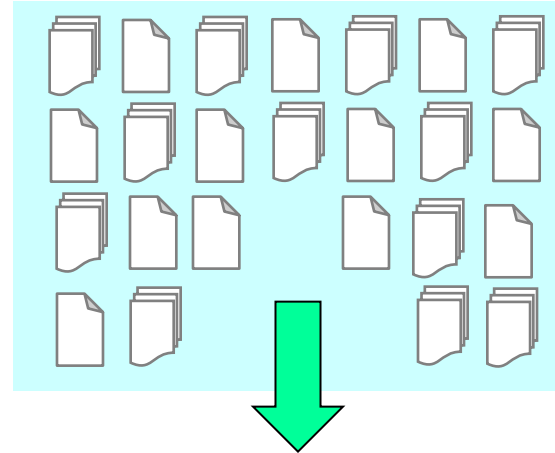
students of Hector Garcia-Molina **while** he was at Princeton?

French Marriage Problem

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

facts in KB

- 1: married
(Hillary, Bill)
 - 2: married
(Carla, Nicolas)
 - 3: married
(Angelina, Brad)
- validFrom (2, 2008)



new fact candidates:

- 4: married (Cecilia, Nicolas)
- 5: married (Carla, Benjamin)
- 6: married (Carla, Mick)
- 7: divorced (Madonna, Guy)

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

validUntil (4, 2007)

Challenge: Temporal Knowledge

for **all people** in Wikipedia (300 000) gather **all spouses**,
incl. divorced & widowed, and corresponding **time periods!**
>95% accuracy, >95% coverage, in one night

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

28 January 1955 (age 53)
Paris, France

Nicolas Paul Stéphane
Sarközy

Political party
RR (?–2002)
UMP (2002–)

Spouse
Marie-Dominique Culioli
(div.)
Cécilia Ciganer-Albéniz
(div.)
Carla Bruni

Children
Pierre (by Culioli)
Jean (by Culioli)
LOUIS (by Ciganer-Albéniz)

Residence
Élysée Palace

Alma mater
University of Paris X:
Nanterre

Occupation
Lawyer

Religion
Roman Catholic



1. Catherine
of Aragon
Divorced



2. Anne
Boleyn
Beheaded



3. Jane
Seymour
Died



consistency constraints are potentially helpful:

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

Difficult Dating

Nicolas Sarkozy



President of France

Incumbent

Assumed office

Born 28 January 1955 (age 55)
Paris, France

Political party Union for a Popular Movement (2002–present)

Other political affiliations Rally for the Republic (1976–2002)

Spouse(s) Marie-Dominique Culioli (1982–1996)
Cécilia Ciganer-Albéniz (1996–2007)
Carla Bruni-Sarkozy (2008–present)

Children Pierre Sarkozy (by Culioli)
Jean Sarkozy (by Culioli)
Louis Sarkozy (by Ciganer-Albéniz)

Residence Élysée Palace

Alma mater Paris X University Nanterre

Profession Lawyer

Religion Roman Catholicism

Cécilia Attias

First Lady of France

In office

16 May 2007 – 10 October 2007

President Nicolas Sarkozy
Preceded by Bernadette Chirac
Succeeded by Carla Bruni

Born November 12, 1957 (age 52)
Boulogne-Billancourt, France

Spouse(s) Jacques Martin (m. 1984–1989)
Nicolas Sarkozy (m. 1996–2007)
Richard Attias (m. 2008–present)

Children Judith Martin (b.1984)
Jeanne-Marie Martin (b.1987)
Louis Sarkozy (b.1997)



Wife of the President of the French Republic

Incumbent

Assumed office
2 February 2008

President Nicolas Sarkozy
Preceded by Cécilia Ciganer-Albéniz

Born 23 December 1967 (age 42)
Turin, Italy

Birth name Carla Gilberta Bruni Tedeschi

Nationality Italian, French^[1]

Spouse(s) Nicolas Sarkozy

Children Aurélien Enthoven (with Raphaël Enthoven)

Charles

Prince of Wales; Duke of Rothesay (more)



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

Issue

Prince William of Wales
Prince Harry of Wales

Full name

Charles Philip Arthur George

House Maternal: House of Windsor
Paternal: House of Schleswig-Holstein-Sonderburg-Glücksburg

Father Prince Philip, Duke of Edinburgh

Mother Elizabeth II

Born 14 November 1948 (age 61)
Buckingham Palace, London

Signature

Religion Christian (Church of England)

Diana

Princess of Wales; Duchess of Rothesay



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

Issue

Prince William of Wales
Prince Henry of Wales

Full name

Diana Frances Spencer^[N 1]

House House of Windsor

Father John Spencer, 8th Earl Spencer

Mother Frances Shand Kydd

Born 1 July 1961
Park House, Sandringham, Norfolk

Died 31 August 1997 (aged 36)
Pitié-Salpêtrière Hospital, Paris, France

Burial Althorp, Northamptonshire

Madonna



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

Background information

Birth name Madonna Louise Ciccone

Guy Ritchie



Guy Ritchie, September 2008

Born Guy Stuart Ritchie
10 September 1968 (age 41)
Hatfield, Hertfordshire, England

Occupation Filmmaker, Screenwriter

Years active 1995–present

Spouse(s) Madonna (2000–2008)
(divorced)

(Even More Difficult) Implicit Dating

explicit dates vs.
implicit dates relative to other dates

Nicolas Sarkozy

From Wikipedia, the free encyclopedia

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

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

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

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

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

(Even More Difficult) Relative Dating

vague dates
relative dates

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 6th 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]}

Framework for T-Fact Extraction

(Y. Wang et al.: EDBT'10, X. Ling et al.: AAAI'10, Y. Wang et al.: CIKM'11)

- 1) **represent temporal scopes** of facts
in the presence of incompleteness and uncertainty
- 2) **gather & filter candidates** for t-facts:
extract **base facts** $R(e1, e2)$ first; then
focus on sentences with $e1$, $e2$ and **date** or **temporal phrase**
- 3) **aggregate & reconcile** evidence from observations
- 4) **reason** on joint constraints about facts and time scopes

Joint Reasoning on Facts and T-Facts

(M. Theobald et al.: MUD'10, M. Dylla et al.: BTW'11)

Combine & reconcile t-scopes **across different facts**

constraint:

marriedTo (m) is an injective function at any given point

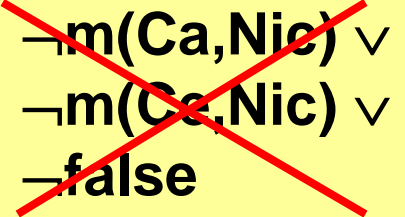
$\forall X, Y, Z, T1, T2:$

$$\begin{aligned} & m(X,Y) \wedge m(X,Z) \wedge \\ & \text{validTime}(m(X,Y),T1) \wedge \text{validTime}(m(X,Z),T2) \\ & \Rightarrow \neg \text{overlaps}(T1, T2) \end{aligned}$$

after grounding:

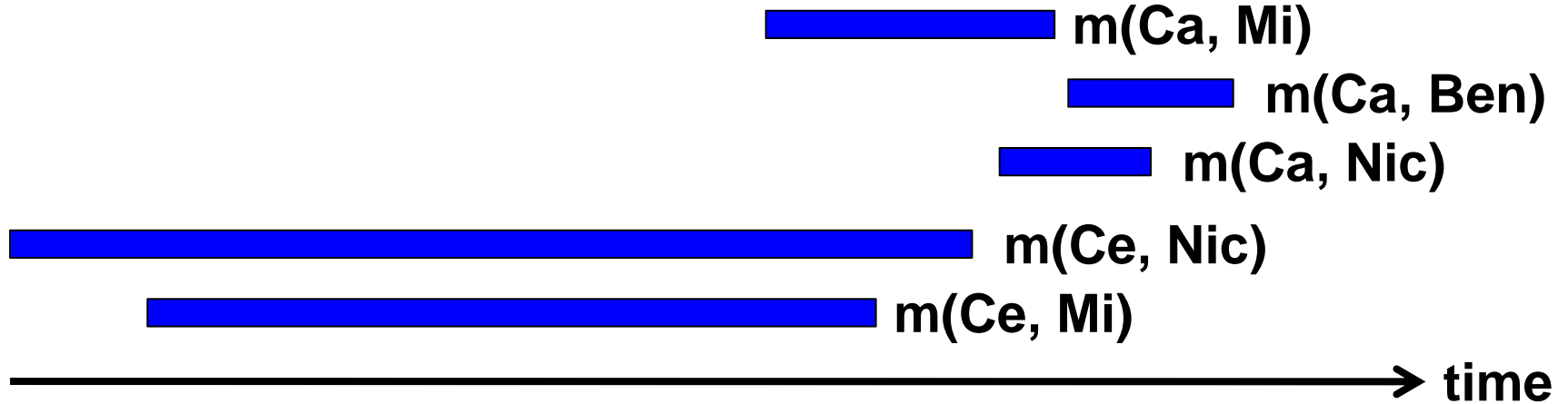
$$\begin{aligned} & m(\text{Carla}, \text{Nicolas}) \wedge m(\text{Cecilia}, \text{Nicolas}) \\ & \Rightarrow \neg \text{overlaps}([2008,2010], [1996,2007]) \end{aligned}$$

$$\begin{aligned} & m(\text{Carla}, \text{Nicolas}) \wedge m(\text{Carla}, \text{Benjamin}) \\ & \Rightarrow \neg \text{overlaps}([2008,2010], [2009,2011]) \end{aligned}$$

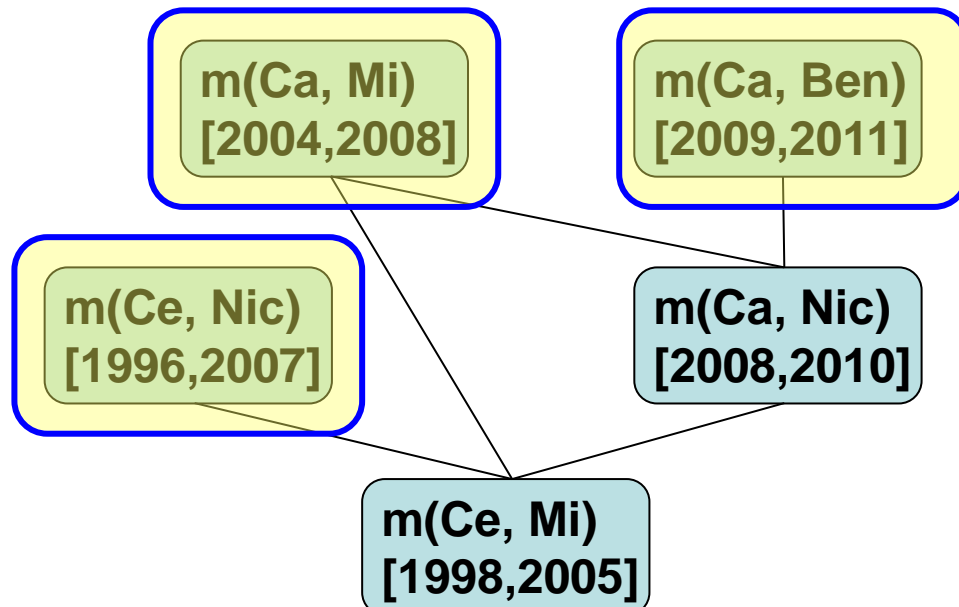

$$\begin{aligned} & \neg m(\text{Ca}, \text{Nic}) \vee \\ & \neg m(\text{Ce}, \text{Nic}) \vee \\ & \neg \text{false} \end{aligned}$$

$$\begin{aligned} & \neg m(\text{Ca}, \text{Nic}) \vee \\ & \neg m(\text{Ca}, \text{Ben}) \vee \\ & \neg \text{true} \end{aligned}$$

Joint Reasoning on Facts and T-Facts

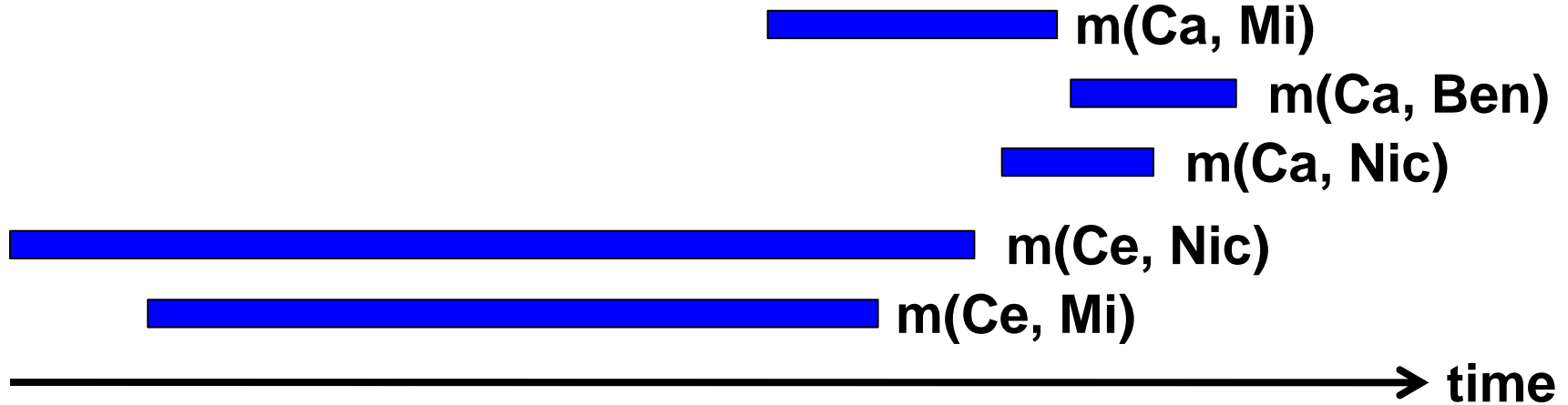


Conflict graph:

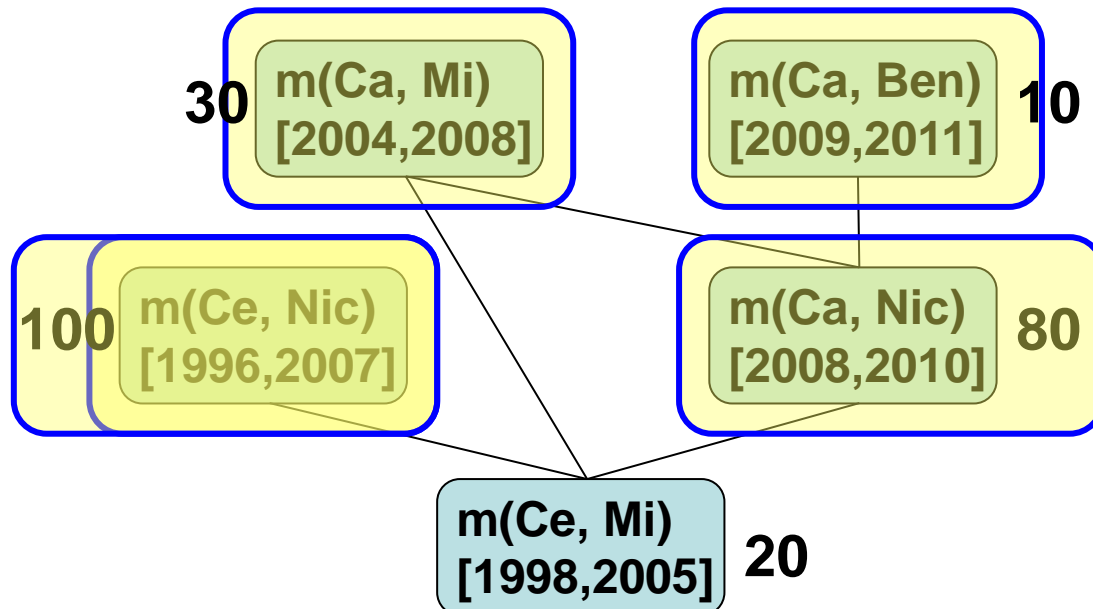


Find **maximal independent set**:
subset of nodes
w/o adjacent pairs
with (evidence-)
weighted nodes

Joint Reasoning on Facts and T-Facts



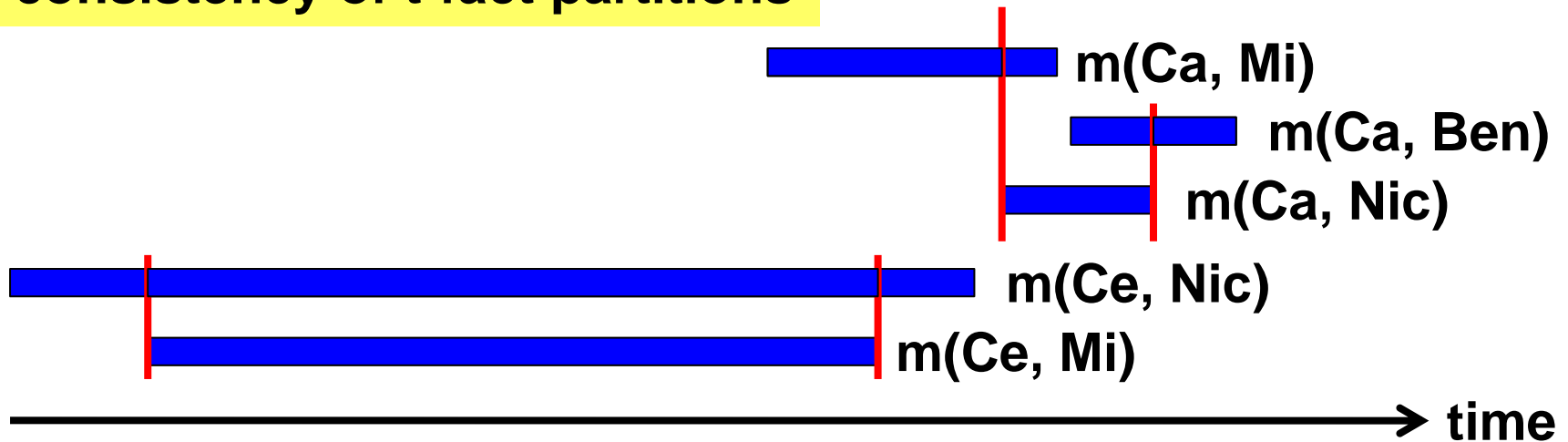
Conflict graph:



Find **maximal independent set**:
subset of nodes
w/o adjacent pairs
with (evidence-)
weighted nodes

Joint Reasoning on Facts and T-Facts

alternative approach:
split t-scopes and reason on
consistency of t-fact partitions



Outline

- ✓ **Motivation**
- ✓ **Machine Knowledge**
- ✓ **Knowledge Harvesting**
 - **Entities and Classes**
 - **Relational Facts**
- ✓ **Research Challenges**
 - **Open-Domain Extraction**
 - **Temporal Knowledge**

★ **Wrap-up**

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

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

Recommended Readings (General)

- 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

Recommended Readings (Specific)

- F.M. Suchanek, G. Kasneci, G. Weikum: Yago: a core of semantic knowledge. WWW 2007
- J. Hoffart, F.M. Suchanek, K. Berberich, et al.: YAGO2: exploring and querying world knowledge in time, space, context, and many languages. WWW 2011
- S. Auer, C. Bizer, et al.: DBpedia: A Nucleus for a Web of Open Data. ISWC 2007
- S.P. Ponzetto, M. Strube: Deriving a Large-Scale Taxonomy from Wikipedia. AAAI 2007
- F. Wu, D.S. Weld: Automatically refining the wikipedia infobox ontology. WWW 2008
- A. Carlson et al.: Toward an Architecture for Never-Ending Language Learning. AAAI 2010
- F.M. Suchanek et al.: SOFIE: a self-organizing framework for information extraction. WWW 2009
- J. Zhu et al: StatSnowball: a statistical approach to extracting entity relationships. WWW 2009
- P. Domingos, D. Lowd: Markov Logic: An Interface Layer for Artificial Intelligence. 2009
- S. Riedel, L. Yao, A. McCallum: Modeling relations and their mentions without labeled text. ECML 2010
- Y.S. Chan, D. Roth: Exploiting Background Knowledge for Relation Extraction. COLING 2010
- M. Banko, M.J. Cafarella, S. Soderland, et al.: Open Information Extraction from the Web. IJCAI 2007
- A. Fader, S. Soderland, O. Etzioni: Identifying Relations for Open Information Extraction, EMNLP 2011
- P.P. Talukdar, F. Pereira: Experiments in Graph-Based Semi-Supervised Learning Methods for Class-Instance Acquisition. ACL 2010
- R. Wang, W.W. Cohen: Language-independent set expansion of named entities using the web. ICDM 2007
- P. Venetis, A. Halevy, et al.: Recovering Semantics of Tables on the Web, VLDB 2011
- F. Niu, C. Re, A. Doan, et al.: Tuffy: Scaling up Statistical Inference in Markov Logic Networks using an RDBMS, VLDB 2011
- X. Ling, D.S. Weld: Temporal Information Extraction. AAAI 2010
- Y. Wang, M. Zhu, L. Qu, M. Spaniol, G. Weikum: Timely YAGO: harvesting, querying, and visualizing temporal knowledge from Wikipedia. EDBT 2010
- Y. Wang, L. Qu, B. Yang, M. Spaniol, G. Weikum: Harvesting Facts from Textual Web Sources by Constrained Label Propagation. CIKM 2011

Thank You!

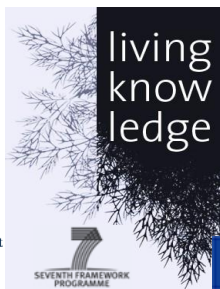


max planck institut
informatik

Thank You!



mpi max planck institut
informatik



DFG Deutsche
Forschungsgemeinschaft

Google