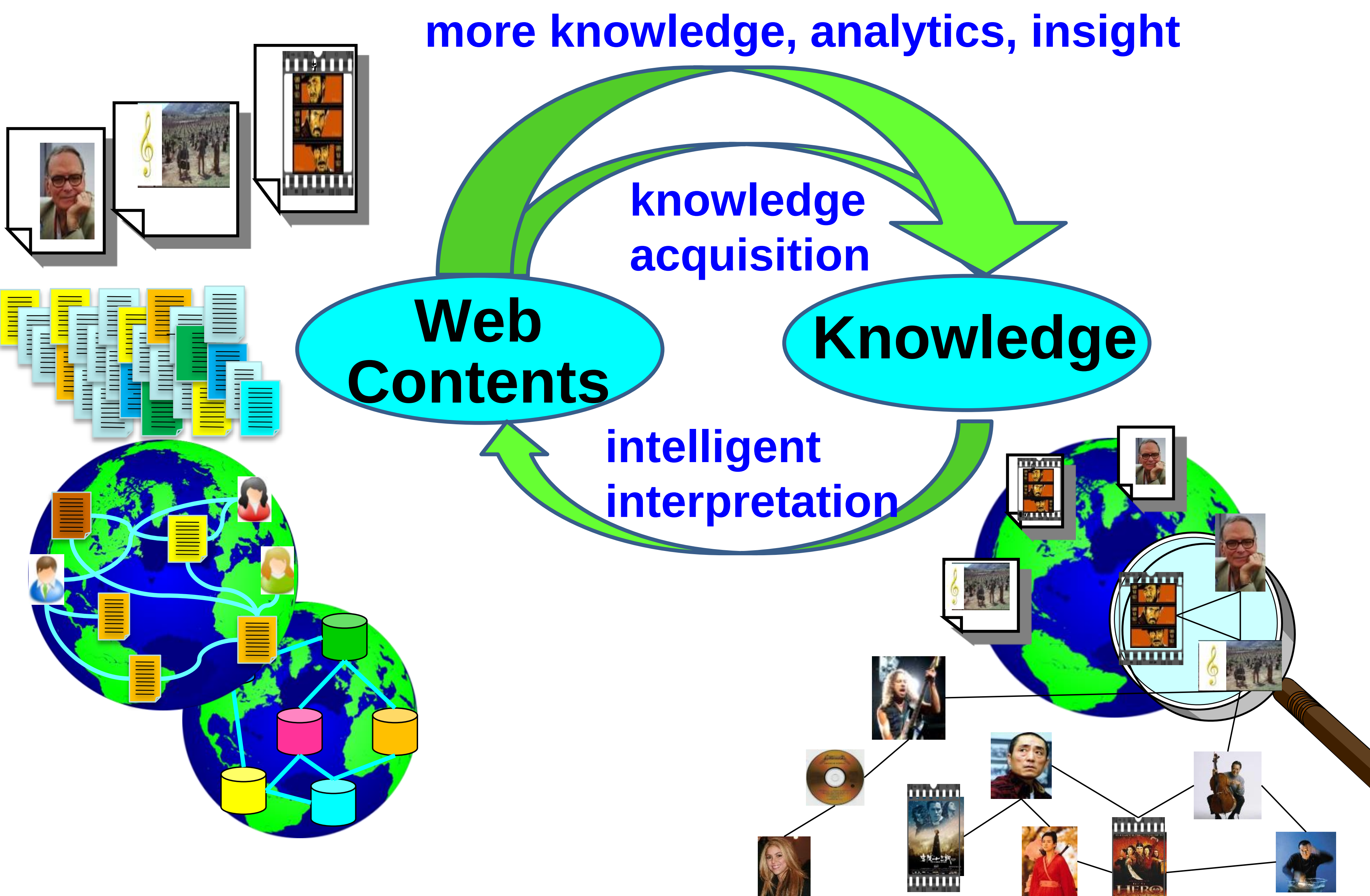
Knowledge Graphs:
from a Fistful of Triples to Deep Data and Deep Text

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

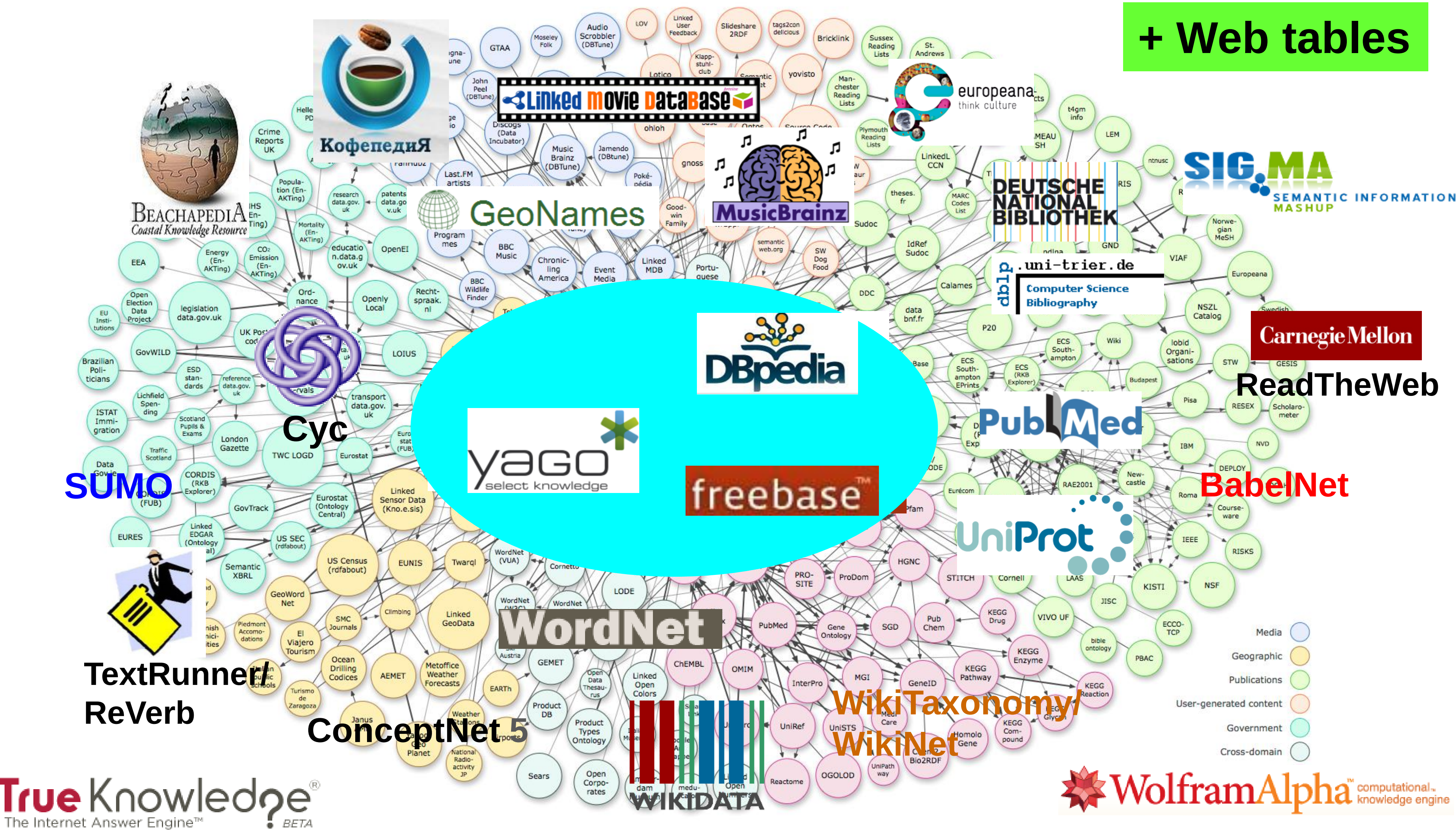
Turn Web into Knowledge Base



Web of Data & Knowledge

> 50 Bio. subject-predicate-object triples from > 1000 sources

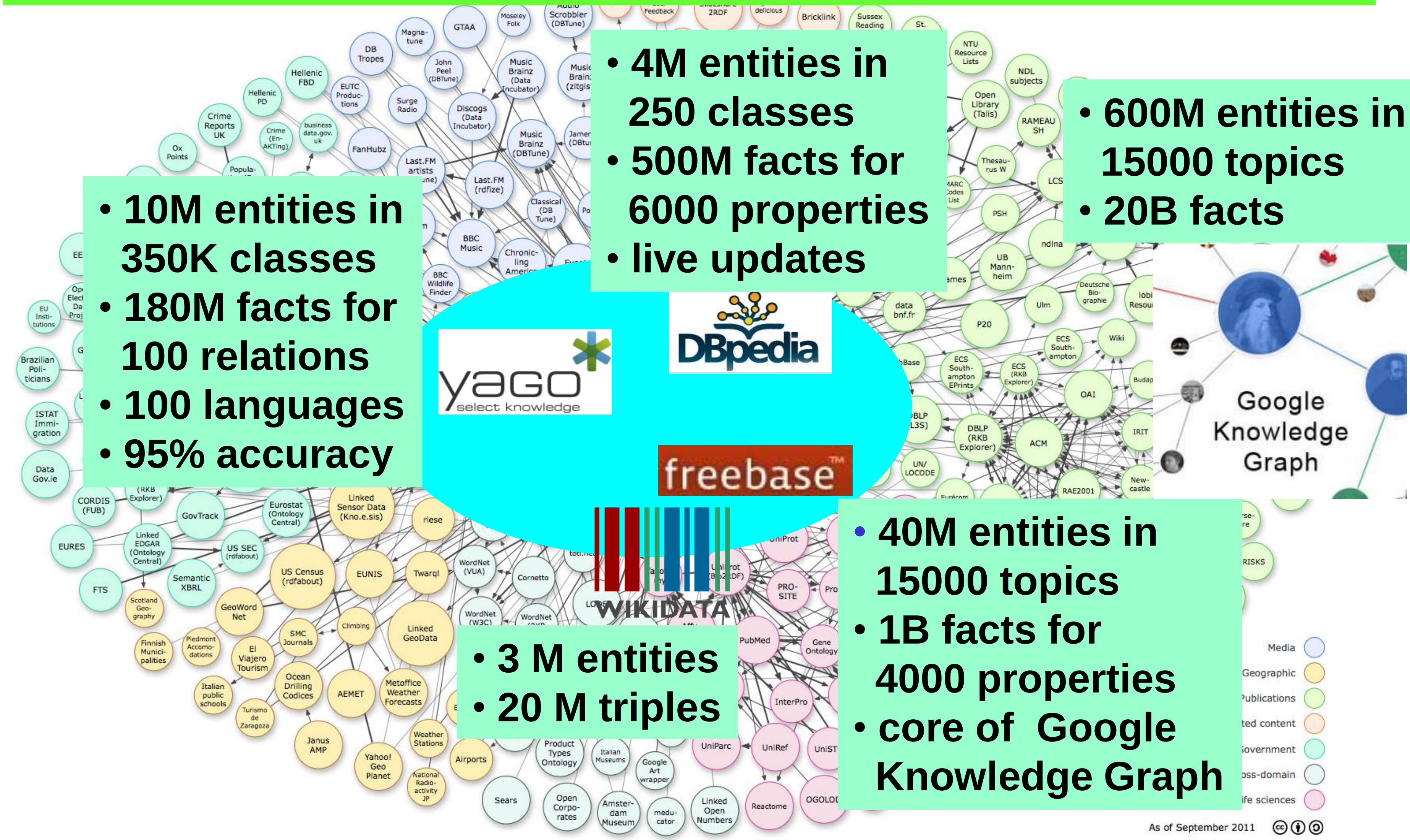
+ Web tables



http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.png

Web of Data & Knowledge

> 50 Bio. subject-predicate-object triples from > 1000 sources



- 10M entities in 350K classes
- 180M facts for 100 relations
- 100 languages
- 95% accuracy

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

- 600M entities in 15000 topics
- 20B facts

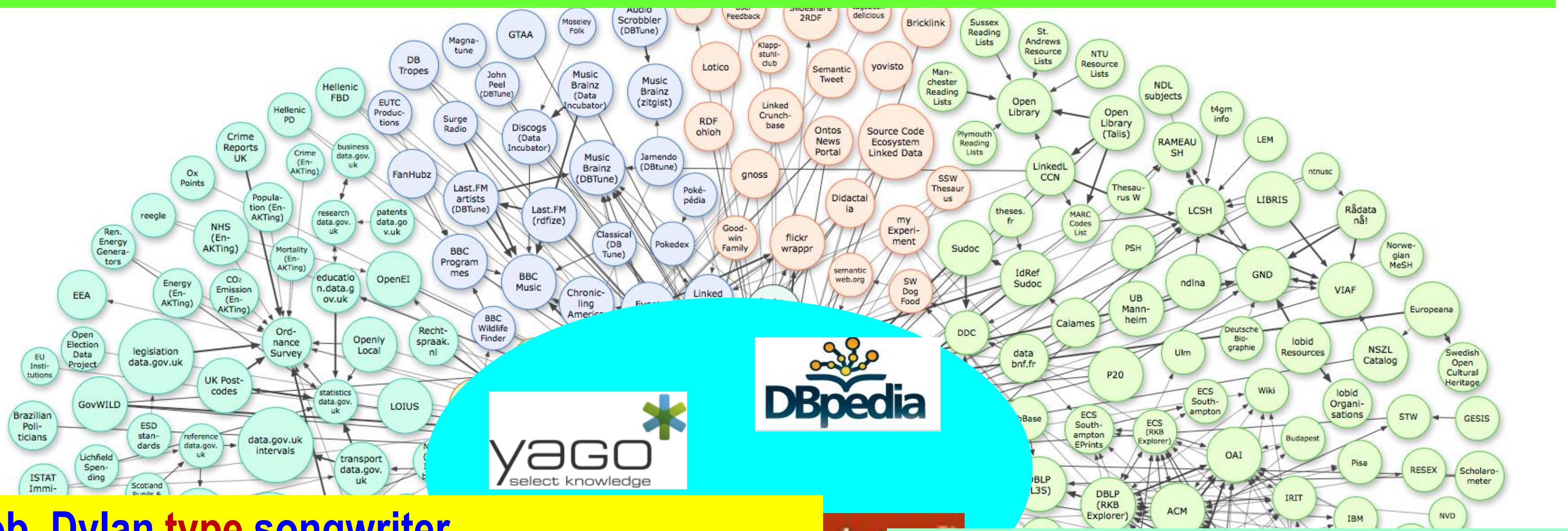
- 40M entities in 15000 topics
- 1B facts for 4000 properties
- core of Google Knowledge Graph

- 3 M entities
- 20 M triples

http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.png

Web of Data & Knowledge

> 50 Bio. subject-predicate-object triples from > 1000 sources



Bob_Dylan type songwriter
Bob_Dylan type civil_rights_activist
songwriter subclassOf artist
Bob_Dylan composed Hurricane
Hurricane isAbout Rubin_Carter
Bob_Dylan marriedTo Sara_Lownds
validDuring [Sep-1965, June-1977]
Bob_Dylan knownAs „voice of a generation“
Steve_Jobs „was big fan of“ Bob_Dylan
Bob_Dylan „briefly dated“ Joan_Baez

taxonomic knowledge

factual knowledge

temporal knowledge

terminological knowledge

evidence & belief knowledge

Knowledge Bases: Pragmatic Definition aka. Knowledge Graphs

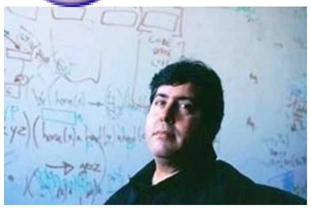
Comprehensive and semantically organized

machine-readable collection of
universally relevant or domain-specific
entities, classes, and
SPO facts (attributes, relations)

plus spatial and temporal dimensions
plus commonsense properties and rules
plus contexts of entities and facts
(textual & visual witnesses, descriptors, statistics)
plus …..

History of Digital Knowledge Bases

Cyc



WordNet

guitarist ⊂ {player,musician}
⊂ artist

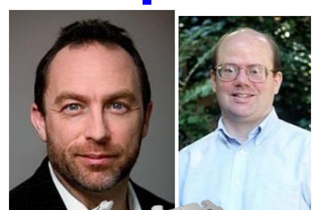
algebraist
⊂ mathematician
⊂ scientist

$\forall$ x: human(x) $\Rightarrow$
 ($\exists$ y: mother(x,y) $\wedge$
 $\exists$ z: father(x,z))

$\forall$ x,u,w: (mother(x,u) $\wedge$
 mother(x,w)
 $\Rightarrow$ u=w)

from humans for humans

Wikipedia



4.5 Mio. English articles
20 Mio. contributors

from algorithms for machines

WolframAlpha

DBpedia

yago
select knowledge

freebase™

1985 1990 2000 2005 2010

Some Publicly Available Knowledge Bases

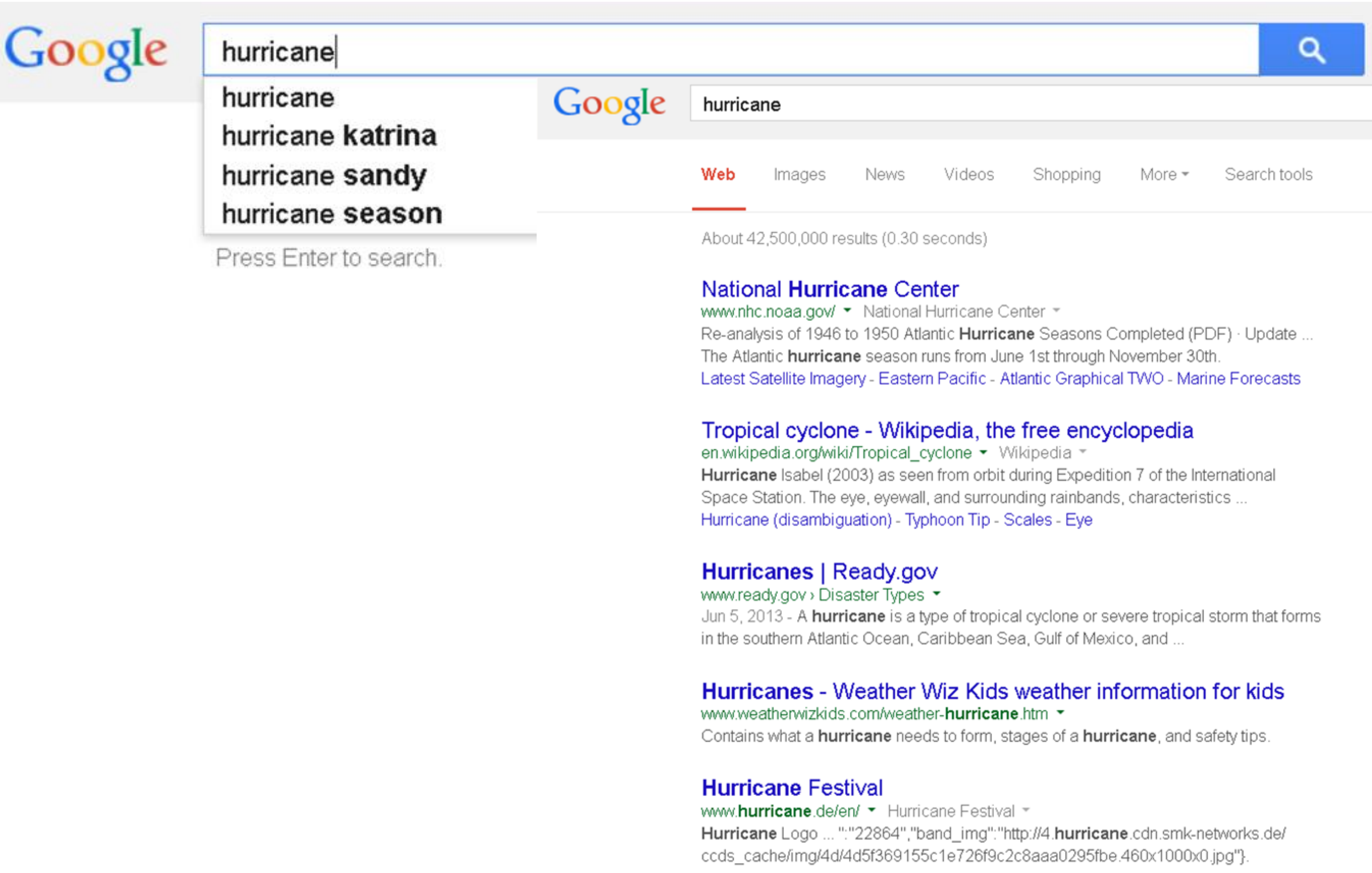
YAGO:	yago-knowledge.org
Dbpedia:	dbpedia.org
Freebase:	freebase.com
Wikidata:	www.wikidata.org
Entitycube:	entitycube.research.microsoft.com
	renlifang.msra.cn
NELL:	rtw.ml.cmu.edu
DeepDive:	deepdive.stanford.edu
Probase:	research.microsoft.com/en-us/projects/probase/
KnowItAll / ReVerb:	openie.cs.washington.edu
	reverb.cs.washington.edu
BabelNet:	babelnet.org
WikiNet:	www.h-its.org/english/research/nlp/download/
ConceptNet:	conceptnet5.media.mit.edu
WordNet:	wordnet.princeton.edu
Linked Open Data:	linkeddata.org

Knowledge for Intelligent Applications

Enabling technology for:

- **disambiguation**
  in written & spoken natural language
- **deep reasoning**
  (e.g. QA to win quiz game)
- **machine reading**
  (e.g. to summarize book or corpus)
- **semantic search**
  in terms of entities&relations (not keywords&pages)
- **entity-level linkage**
  for Big Data & Big Text analytics

Use-Case: Internet Search



Google Knowledge Graph

(Google Blog: „Things, not Strings“, 16 May 2012)

hurricane singer

weikum

Web Images Videos News Shopping More Search tools

About 7,650,000 results (0.33 seconds)

Cookies help us deliver our services. By using our services, you agree to our use of cookies.
Learn more Got it

Bob Dylan

Hurricane, Artist

Feedback

Hurricane (band) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Hurricane**_(band)
Hurricane is a 1980s heavy metal band originally featuring current Foreigner lead **vocalist** Kelly Hansen (vocals/rhythm guitar), Robert Sarzo (guitar), Tony ...
History - Current members - Past members - Discography

Kelly Hansen - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Kelly_Hansen
Kelly Hansen (born April 18, 1961) is an American **singer**, best known as the ... of Quiet Riot fame), with whom he formed the hard-rock band **Hurricane** in 1984.

Bob Dylan

Musician

Bob Dylan is an American musician, singer-songwriter, artist, and writer. He has been an influential figure in popular music and culture for more than five decades. Wikipedia

Spouse: Carolyn Dennis (m. 1986–1992), Sara Dylan (m. 1965–1977)

Children: Jakob Dylan, Desiree Gabrielle Dennis-Dylan, Anna Dylan, Jesse Dylan, Maria Dylan, Sam Dylan

Movies: Pat Garrett and Billy the Kid, Masked and Anonymous, more

Songs

Knockin' on Heaven's Door	1973	Pat Garrett & Billy the Kid
Farewell		
Forever Young	1974	Planet Waves
Make You Feel My Love	1997	Time Out of Mind
Hurricane	1976	Desire

Albums

Google Knowledge Graph: Limitations

(Google Blog: „Things, not Strings“, 16 May 2012)

Google

bob dylan cover songs

Web Videos News Shopping Images More Search tools

About 1,090,000 results (0.39 seconds)

List of artists who have covered Bob Dylan songs - Wikipedia
en.wikipedia.org/.../List_of_artists_who_have_**covered**_**Bob**_... Wikipedia
Many major recording artists have covered **Dylan's** material, some even increasing its popularity as is the case with The Byrds' **cover** version of "Mr. Tambourine ...

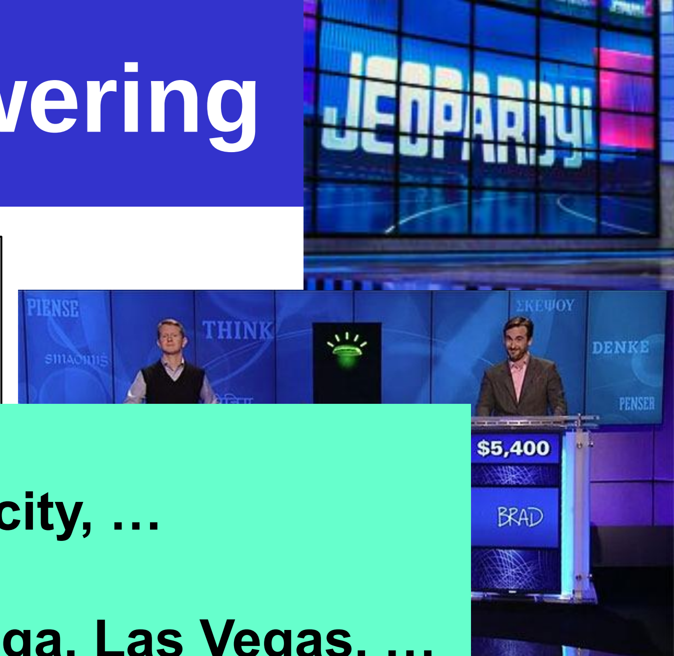
Bob Dylan's 20 Greatest Cover Versions - Mojo
www.mojo4music.com/19239/**bob-dylans**-20-greatest-**covers**/ Mojo
Mar 5, 2015 - Listen to **Bob Dylan's** 20 best **cover versions**, as selected by MOJO's experts.

Bob Dylan: His New Cover Songs Explored | MOJO
www.mojo4music.com/.../**bob-dylan**-new-**cover-songs**-explored/ Mojo
Dec 17, 2014 - Mojo explores the sources of **Bob Dylan's** new **covers** album Shadows In The Night.

50 Best Bob Dylan Covers of All Time :: Music :: Lists :: Paste
www.pastemagazine.com/.../50-best-**bob-dylan-covers**-of-all-time.html
Apr 28, 2009 - As we began to compile this list of the 50 Best **Bob Dylan Covers** of All ... There are so many transcendent moments in these 50 **songs**. Antony's ...

Bob Dylan - Second Hand Songs
secondhand**songs**.com/artist/158
Bob Dylan originally did Forever Young, All Along the Watchtower, Knockin' on ... Released on Blood on the **Tracks** (1975) ... Popular **Covers** by **Bob Dylan** ...

Use Case: Question Answering



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

Q: Sin City ?
→ movie, graphical novel, nickname for city, …
A: Vegas ? Strip ?
→ Vega (star), Suzanne Vega, Vincent Vega, Las Vegas, …
→ comic strip, striptease, Las Vegas Strip, …

This American city has two airports
named after a war hero and a WW II battle

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

WIKIPEDIA The Free Encyclopedia
DBpedia
freebase
yago select knowledge

D. Ferrucci et al.: Building Watson. AI Magazine, Fall 2010.
IBM Journal of R&D 56(3/4), 2012: This is Watson.

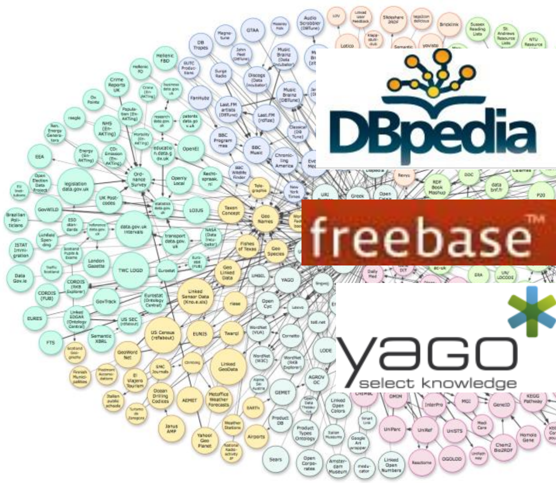
Use Case: Question Answering

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

Q: Sin City ?
→ **movie, graphical novel, nickname for city, …**
A: Vegas ? Strip ?
→ **Vega (star), Suzanne Vega, Vincent Vega, Las Vegas, …**
→ **comic strip, striptease, Las Vegas Strip, …**

question → **structured query**

Select ?t Where {
?t type location .
?t hasLabel “Sin City“ .
?t hasPart ?d .
?d hasLabel “Glitter Gulch“ . }



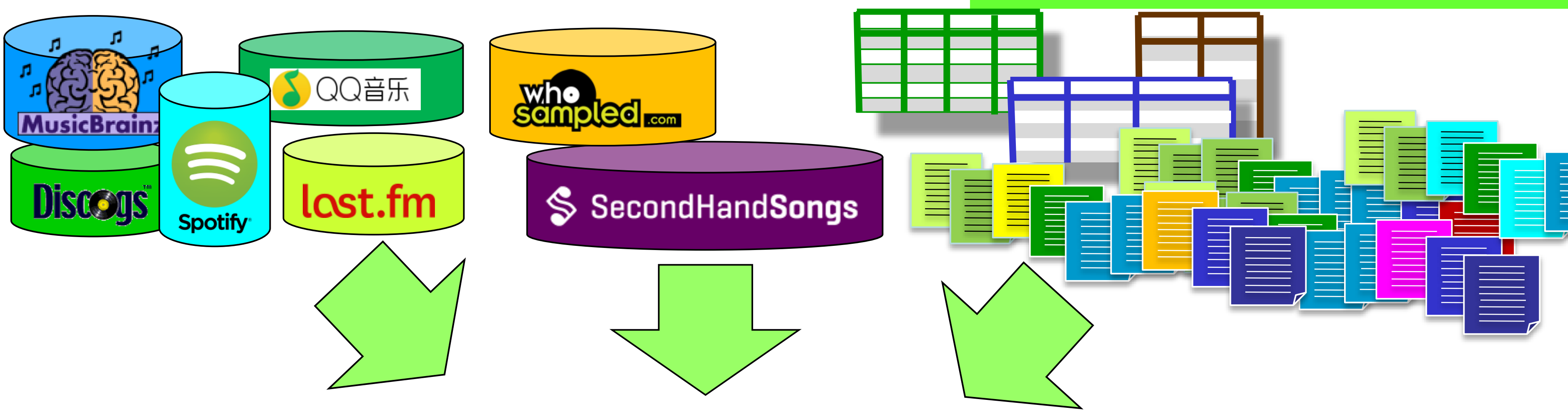
Linked Data
Big Data
Web tables

Use Case: Deep Data & Text Analytics

Who Covered Whom?

in different language, country, key, …
with more sales, awards, media buzz, …

1000's of Databases
100 Mio's of Web Tables
100 Bio's of Web & Social Media Pages



Musician	Original	Title
Elvis Presley	Frank Sinatra	My Way
Robbie Williams	Frank Sinatra	My Way
Sex Pistols	Frank Sinatra	My Way
Frank Sinatra	Claude Francois	Comme d'Habitude
Claudia Leitte	Bruno Mars	Famo$a (Billionaire)
Only Won	Bruno Mars	I wanna be an engineer
…..	…..	…..

Use Case: Deep Data & Text Analytics

Who Covered Whom?

in different language, country, key, …
with more sales, awards, media buzz, …

1000's of Databases
100 Mio's of Web Tables
100 Bio's of Web & Social Media Pages

⋈

Musician	PerformedTitle
Sex Pistols	My Way
Frank Sinatra	My Way
Claudia Leitte	Famo$a
Petula Clark	Boy from Ipanema

Musician	CreatedTitle
Francis Sinatra	My Way
Paul Anka	My Way
Bruno Mars	Billionaire
Astrud Gilberto	Garota de Ipanema

Name	Show
Petula C.	Muppets
Claudia L.	FIFA 2014

Name	Group
Sid Vicious	Sex Pistols
Bono	U2

Use Case: Deep Data & Text Analytics

Who Covered Whom?

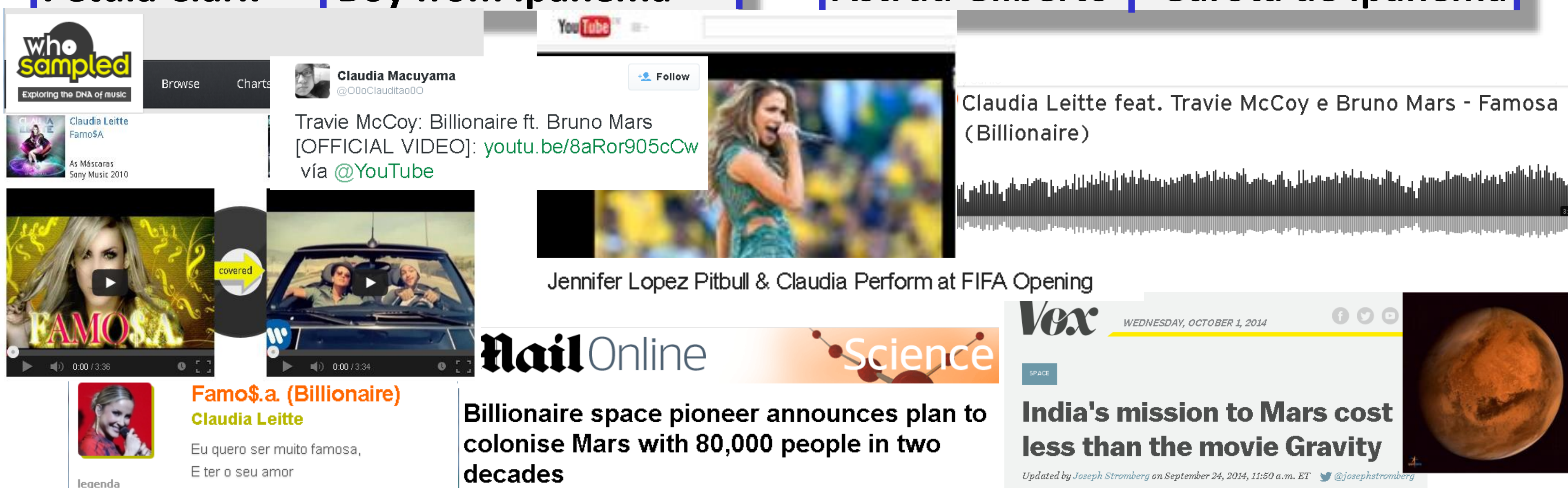
in different language, country, key, …
with more sales, awards, media buzz, …

1000's of Databases
100 Mio's of Web Tables
100 Bio's of Web &
Social Media Pages

Big Data & Deep Text

Volume
Velocity
Variety
Veracity

Musician			CreatedTitle
Sex Pistols		a	My Way
Frank Sinatr			My Way
Claudia Leit			Billionaire
Petula Clark	Boy from Ipanema	Astrud Gilberto	Garota de Ipanema



Deep Data & Text Analytics

Entertainment: Who covered which other singer?
Who influenced which other musicians?

Health: Drugs (combinations) and their side effects

Politics: Politicians‘ positions on controversial topics

Finance: Risk assessment based on data, reports, news

Business: Customer opinions on products in social media

Culturomics: Trends in society, cultural factors, etc.

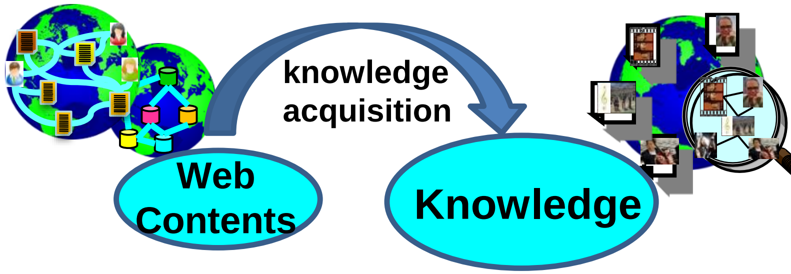
General Design Pattern:

- Identify relevant contents sources
- Identify entities of interest & their relationships
- Position in time & space
- Group and aggregate
- Find insightful patterns & predict trends

Outline

✓ Introduction

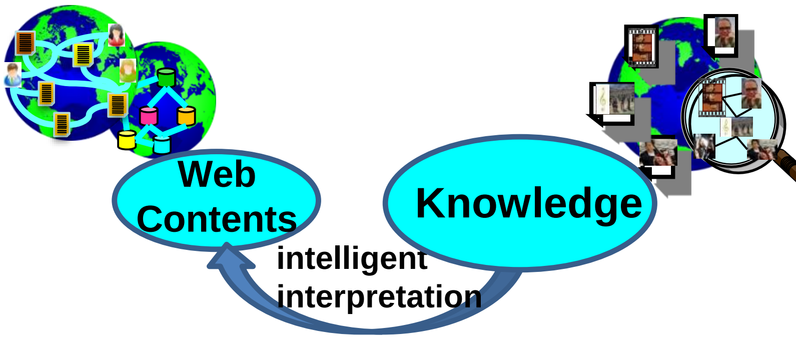
★ KG Construction



★ Refined Knowledge

★ Knowledge for Language

★ Deep Text Analytics



★ Search for Knowledge

★ Conclusion

Goal: KG of 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 (Jim Gray computer scientists),
instanceOf (BarbaraLiskov, computer scientists),
instanceOf (Barbara Liskov, female humans), …

Modern Knowledge Resources: WordNet

WordNet project (1985-now)

George Miller

Christiane Fellbaum

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

WordNet Search - 3.1
- WordNet home page - Gloss

Noun

- S: (n) **enterprise**, endeavor, endeavour (a purposeful or industrious undertaking (especially one that requires effort or boldness)) *"he had doubts about the whole enterprise"*
- S: (n) **enterprise** (an organization created for business ventures) *"a growing enterprise must have a bold leader"*
- S: (n) **enterprise**, enterprisingness, initiative, go-ahead (readiness to embark on bold new ventures)

Modern Knowledge Resources: WordNet

- direct hyponym / full hyponym
  - S: (n) giant (an unusually large enterprise) *"Walton built a retail giant"*
  - S: (n) collective (members of a cooperative enterprise)
  - S: (n) business, concern, business concern, business organization, business organisation (a commercial or industrial enterprise and the people who constitute it) *"he bought his brother's business"; "a small mom-and-pop business"; "a racially integrated business concern"*
    - direct hyponym / **full hyponym**
      - S: (n) agency (a business or organization that provides a particular service, especially the mediation of transactions between two parties)
        - S: (n) advertising agency, ad agency (an agency that designs advertisement to call public attention to its clients)
        - S: (n) credit bureau (a private firm that maintains consumer credit data files and provides credit information to authorized users for a fee)
        - S: (n) detective agency (an agency that makes inquiries for its clients)
        - S: (n) employment agency, employment office (an agency that finds people to fill particular jobs or finds jobs for unemployed people)
        - S: (n) mercantile agency, commercial agency (an organization that provides businesses with credit ratings of other firms) *"Dun & Bradstreet is the largest mercantile agency in the United States"*
        - S: (n) news agency, press agency, wire service, press association, news organization, news organisation (an agency to collects news reports for newspapers and distributes it electronically)
          - S: (n) syndicate (a news agency that sells features or articles or photographs etc. to newspapers for simultaneous publication)
        - S: (n) service agency, service bureau, service firm (a business that makes its facilities available to others for a fee; achieves economy of scale)
        - S: (n) travel agency (an agency that arranges

- S: (n) firm, house, business firm (the members of a business organization that owns or operates one or more establishments) *"he worked for a brokerage house"*
  - S: (n) corporation, corp (a business firm whose articles of incorporation have been approved in some state)
    - S: (n) conglomerate, empire (a group of diverse companies under common ownership and run as a single organization)
      - S: (n) publishing conglomerate, publishing empire (a conglomerate of publishing companies)
    - S: (n) large cap (a corporation with a large capitalization) *"he works for a large cap"*
    - S: (n) small cap (a corporation with a small capitalization) *"this annual conference is a showcase for ambitious small caps"*
    - S: (n) closed corporation, close corporation, private corporation, privately held corporation (a corporation owned by a few people; shares have no public market)
      - S: (n) family business (a corporation that is entirely owned by the members of a single family)
    - S: (n) closely held corporation (stock is publicly traded but most is held by a few shareholders who have no plans to sell)
    - S: (n) shell corporation, shell entity (a company that is incorporated but has no assets or operations)
    - S: (n) Federal Deposit Insurance Corporation, FDIC (a federally sponsored corporation that insures accounts in national banks and other

Modern Knowledge Resources: WordNet

S: (n) **enterprise** (an organization created for business ventures) *"a growing enterprise must have a bold leader"*

- *direct hyponym / full hyponym*
  - S: (n) giant (an unusually large enterprise) *"Walton built a retail giant"*
  - S: (n) collective (members of a cooperative enterprise)
  - S: (n) business, concern, business concern, business organization,

- S: (n) **entrepreneur**, enterpriser (someone who organizes a business venture and assumes the risk for it)
  - ***has instance***
    - S: (n) Gates, Bill Gates, William Henry Gates (United States computer entrepreneur whose software company made him the youngest multi-billionaire in the history of the United States (born in 1955))
    - S: (n) Sinclair, Clive Sinclair, Sir Clive Marles Sinclair (English electrical engineer who founded a company that introduced many innovative products (born in 1940))

+ focus on classes and taxonomic structure

– few or no instances (entities) of classes

- S: (n) capitalist (a person who invests capital in a business (especially a large business))
  - S: (n) person, individual, someone, somebody, mortal, soul (a human being) *"there was too much for one person to do"*
    - S: (n) organism, being (a living thing that has (or can

Knowledge Communities & New Opportunities

Jimmy Wales

Larry Sanger

Steve Jobs

From Wikipedia, the free encyclopedia

For the biography, see Steve Jobs (biography).

Steven Paul Jobs (/ˈdʒɒbz/; February 24, 1955 – October 5, 2011)[4][5] was an American businessman and inventor widely recognized as a charismatic pioneer of the personal computer revolution.[6][7] He was co-founder, chairman, and chief executive officer of Apple Inc. Jobs also co-founded and served as chief executive of Pixar Animation Studios; he became a member of the board of directors of The Walt Disney Company in 2006, following the acquisition of Pixar by Disney.

In the late 1970s, Apple co-founder Steve Wozniak engineered one of the first commercially successful lines of personal computers, the Apple II series. Jobs directed its aesthetic design and marketing along with A.C. "Mike" Markkula, Jr. and others. In the early 1980s, Jobs was among the first to see the commercial potential of Xerox PARC's mouse-driven graphical user interface, which led to the creation of the Apple Lisa (engineered by Ken Rothmuller and John Couch) and, one year later, creation of Apple employee Jef Raskin's Macintosh.

After losing a power struggle with the board of directors in 1985, Jobs left Apple and founded NeXT, a computer platform development company specializing in the higher-education and business markets. NeXT was eventually acquired by Apple in 1996, which brought Jobs back to the company he co-founded, and provided Apple with the NeXTSTEP codebase, from which the Mac OS X was developed."[8] Jobs was named Apple advisor in 1996, interim CEO in 1997, and CEO from 2000 until his resignation. He oversaw the development of the iMac, iTunes, iPod, iPhone, and iPad and the company's Apple Retail Stores.[9] In 1986, he acquired the computer graphics division of Lucasfilm Ltd, which was spun off as Pixar Animation Studios.[10] He was credited in *Toy Story* (1995) as an executive producer. He remained CEO and majority shareholder at 50.1 percent until its acquisition by The Walt Disney Company in 2006,[11] making Jobs Disney's largest individual shareholder at seven percent and a member of Disney's Board of Directors.[12][13]

In 2003, Jobs was diagnosed with a pancreas neuroendocrine tumor. Though it was initially treated, he reported a hormone imbalance, underwent a liver transplant in 2009, and appeared progressively thinner as his health declined.[14] On medical leave for most of 2011, Jobs resigned as Apple CEO in August that year and was elected Chairman of the Board. On October 5, 2011, Jobs died of respiratory arrest related to his metastatic tumor. He

Steve Jobs

Jobs holding a white iPhone 4 at Worldwide Developers Conference 2010

Born	Steven Paul Jobs February 24, 1955[1][2] San Francisco, California, U.S.[1][2]
Died	October 5, 2011 (aged 56)[2] Palo Alto, California, U.S.
Nationality	American
Alma mater	Reed College (dropped out)

Knowledge Communities & New Opportunities


Jimmy Wales


Larry Sanger

Steve Jobs

From Wikipedia, the free encyclopedia

For the biography, see Steve Jobs (biography).

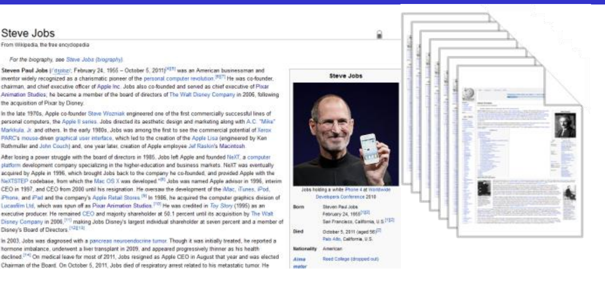
Steven Paul Jobs (/ˈdʒɒbz/; February 24, 1955 – October 5, 2011)[4][5] was an American busine
inventor widely recognized as a charismatic pioneer of the personal computer revolution.[6][7] He w
chairman, and chief executive officer of Apple Inc. Jobs also co-founded and served as chief exec
Animation Studios; he became a member of the board of directors of The Walt Disney Company i
the acquisition of Pixar by Disney.

In the late 1970s, Apple co-founder Steve Wozniak engineered one of the first commercially succ
personal computers, the Apple II series. Jobs directed its aesthetic design and marketing along w
Markkula, Jr. and others. In the early 1980s, Jobs was among the first to see the commercial pote
PARC's mouse-driven graphical user interface, which led to the creation of the Apple Lisa (engine

Born	Steven Paul Jobs February 24, 1955[1][2] San Francisco, California, U.S.[1][2]
Died	October 5, 2011 (aged 56)[2] Palo Alto, California, U.S.
Nationality	American

Categories: Steve Jobs | 1955 births | 2011 deaths | American adoptees | American billionaires | American chief executives | American computer businesspeople | American industrial designers | American inventors | American people of German descent | American people of Swiss descent | American people of Syrian descent | American technology company founders | American Zen Buddhists | Apple Inc. | Apple Inc. employees | Businesspeople from California | Businesspeople in software | Cancer deaths in California | Computer designers | Computer pioneers | Deaths from pancreatic cancer | Disney people | Internet pioneers | National Medal of Technology recipients | NeXT | Organ transplant recipients | People from the San Francisco Bay Area | Pescetarians | Reed College alumni

Automatic Knowledge Base Construction



map 300K Wikipedia categories onto 100K WordNet classes

American billionaires → **tycoon, magnate**

Technology company founders → **entrepreneur**

Apple Inc. →

Deaths from cancer → **?**

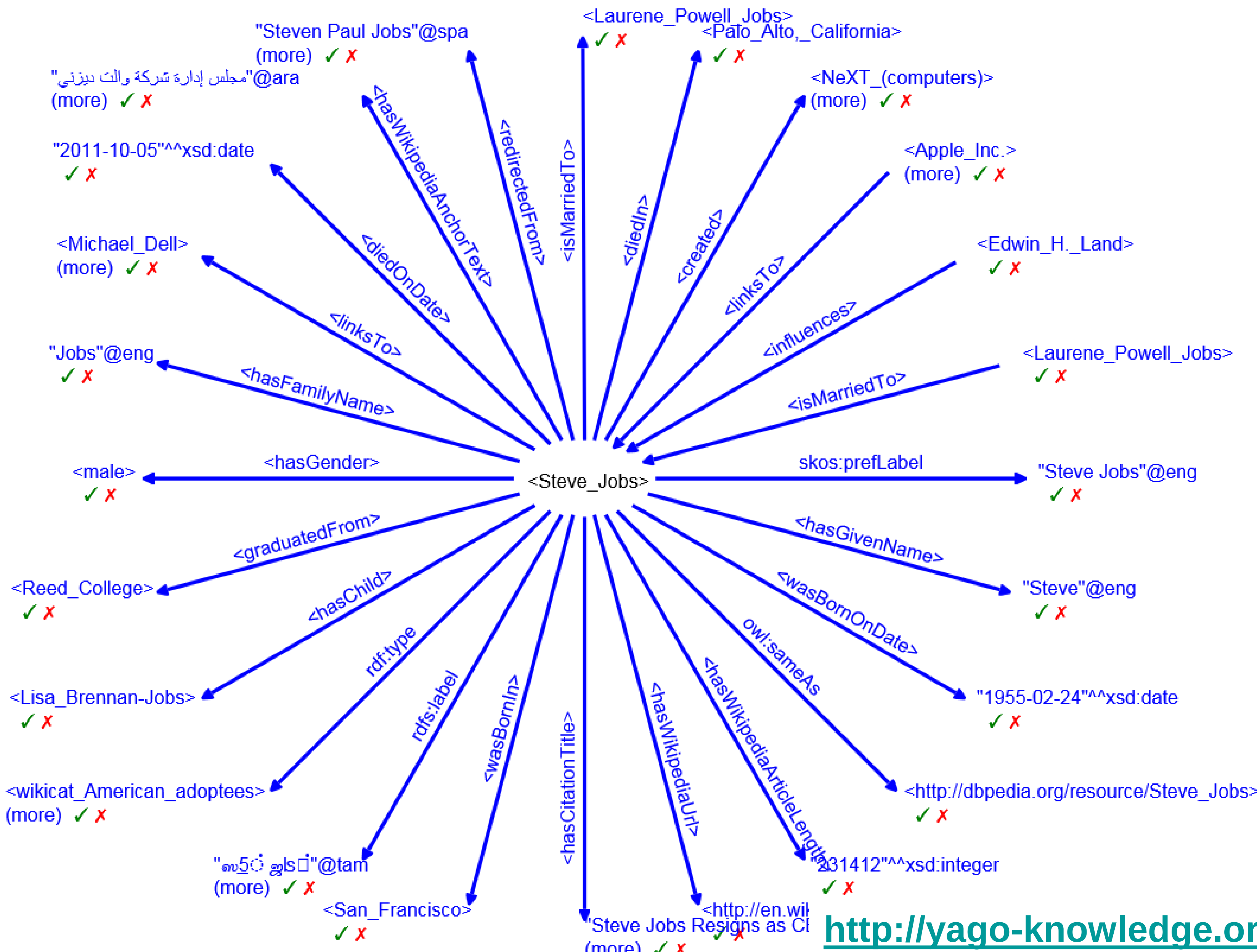
Internet pioneers → **pioneer, innovator** / **pioneer, colonist** **?**

- **Noun group parsing identifies head word**
- **Plural heuristics eliminates non-classes**
- **MostFrequentSense heuristics maps head word to WordNet**

Integrating entities from Wikipedia with classes in WordNet
→ YAGO: 10M entities, 350K classes

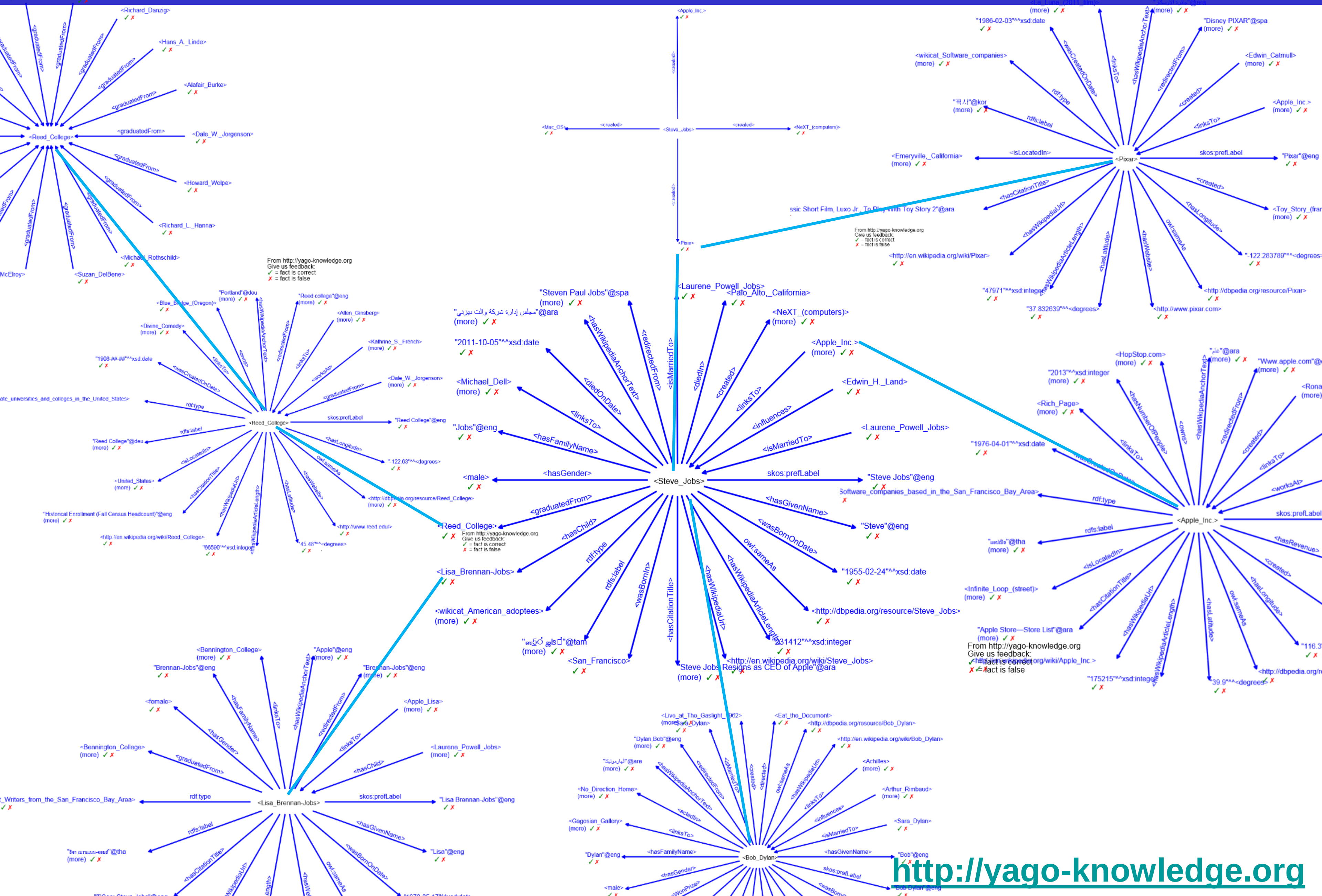
Parallel work → WikiTaxonomy (HITS / U Heidelberg)

Automatic Knowledge Graph Construction



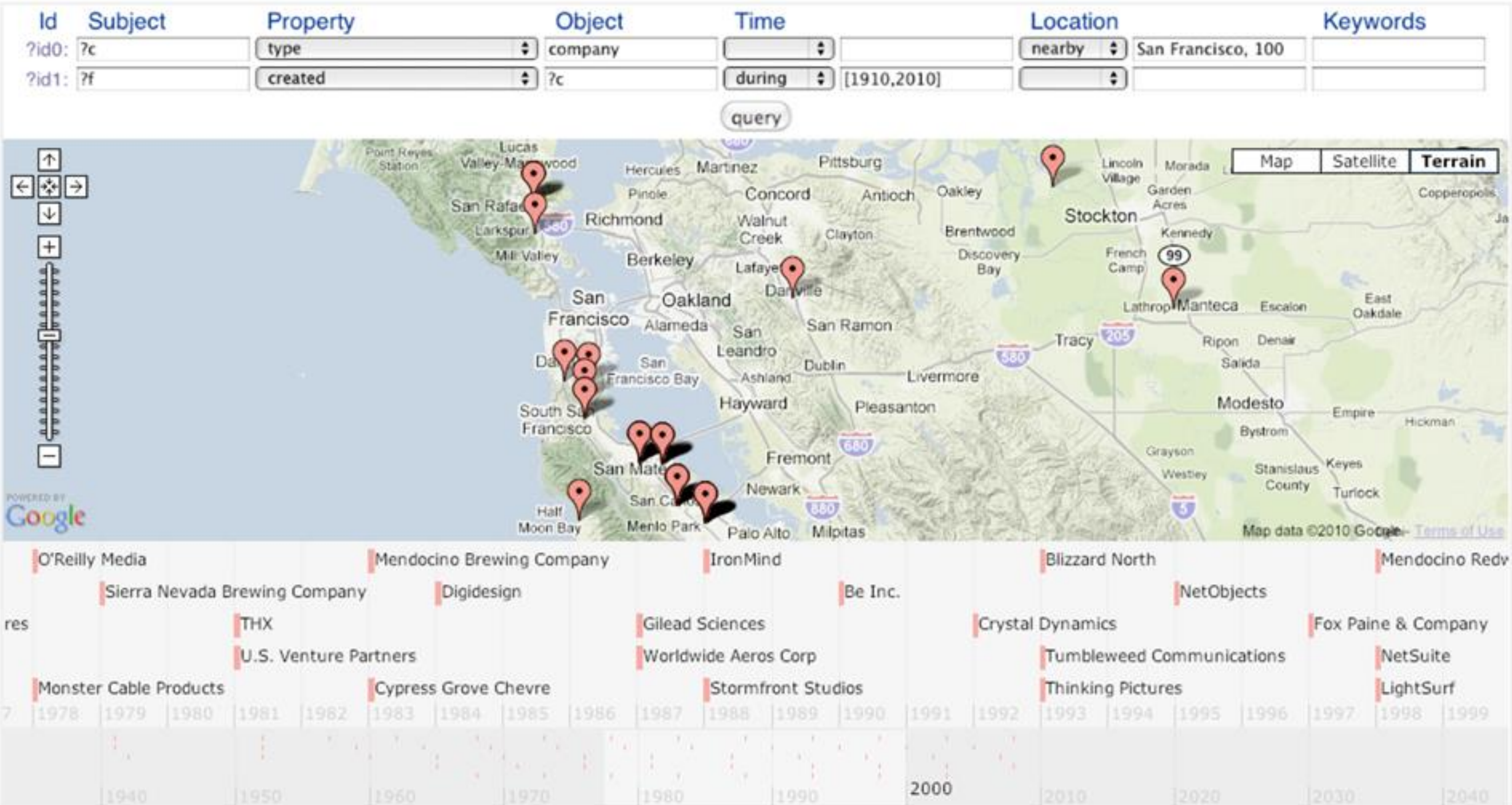
http://yago-knowledge.org

Automatic Knowledge Graph Construction



http://yago-knowledge.org

Automatic Knowledge Graph Construction



Knowledge by space and time

http://yago-knowledge.org

Classes and Entities in the Long Tail

[M. Hearst, P. Pantel, S. Ponzetto / M. Strube, D. Weld, M. Pasca, H. Wang, …]

Extract more classes and instances from

- **Wikipedia categories, infoboxes, lists, headings, edit history, etc.**

  Ex.: knownFor: database research, CEO: Jack Ma

  → **machine learning**

- **Web page contents**

  → **linguistic (Hearst) patterns & taxonomy induction**

  Ex.: database researchers such as Halevy

- **Query and click logs**

  Ex.: Jordan machine learning, Jordan computer scientist

  → **search engine statistics**

May use existing KB for distant supervision

Goal: Facts about Relationships

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), …

Knowledge Harvesting: Low-Hanging Fruit

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

	cience
Alma mater	University of Southern California
Doctoral advisor	Seymour Ginsburg
Known for	Abiteboul-Vianu Theorem
Notable awards	ACM SIGMOD Edgar F. Codd Innovations Award (1998) ACM SIGMOD Test of Time Award 2004 Prix d'informatique de l'Académie des Sciences (Prix EADS) 2007 ACM PODS Alberto O. Mendelzon Test-of-Time Award 2008 ISI highly cited researcher

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
	American
	Columbia University, Princeton University
	Arthur Bernstein, Archie McKellar
	Alexander Birman, Surajit Chaudhuri, Evan Cohn, Alan Demers, Marcia Derr, Nahed El Djabri, Amelia Fong Lochovsky, Deepak Goyal, Ashish Gupta, Himanshu Gupta, Udaiprakash Gupta, Venkatesh Harinarayan, Taher Haveliwala, Matthew Hecht, Daniel Hirschberg, Peter Hochschild, Peter Honeyman, Edward Horvath, Gregory Hunter, Nam (Pierre) Huyn, Hakan Jakobsson, John Kam, Marc Kaplan, Anna Karlin, Kevin Karplus, Henry Korth, Gabriel Kuper, Chen Li, Leonard Liu, George Lueker, David Maier, Harry Mairson, Alberto O. Mendelzon, Katherine Morris, Inderpal Mumick, Jeffrey F. Naughton, Svetlozar Nestorov, Geoffrey Phipps, Thane Plambeck, Anand Rajaraman, Kenneth Ross, Fereidoon Sadri, Yehoshua Sagiv, Yatin Saraiya, Dilip Sarwate, Edward Sciore, Ravi Sethi, Alan Siegel, Howard Siegel, Alberto Torres, Howard Trickey, Allen Van Gelder, Vasilios Vassalos, Cheng (Calvin) Yang, Mihalis Yannakakis

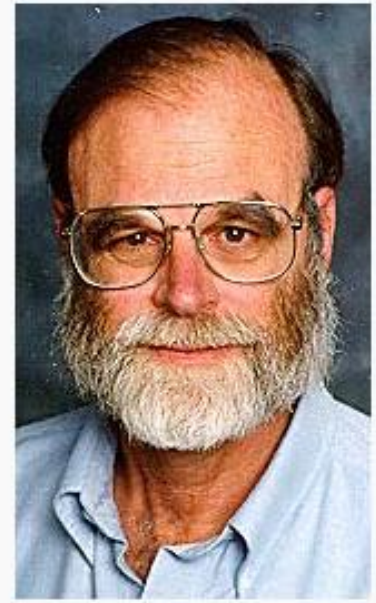
Harvesting Wikipedia Infoboxes

Jim Gray (computer scientist)

```
{{Infobox scientist
| name                  = James Nicholas "Jim" Gray
| birth_date            = {{birth date|1944|1|12}}
| birth_place           = [[San Francisco, California]]
| death_date = ('''lost at sea''')
          {{death date and age|2007|1|28|1944|1|12}}
| death_place           =
| residence             =
| citizenship           =
| nationality           = American
| ethnicity             =
| field                 = [[Computer Science]]
| alma_mater     = [[University of California, Berkeley]]
| doctoral_advisor = Michael Harrison
| known_for          = Work on [[database system|database]]
                       and [[transaction processing]] systems
| prizes                = [[Turing Award]]
| religion              =
}}
```

harvest by extraction rules:
- **regex matching**
- **type checking**

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

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

Harvesting Wikipedia Infoboxes

Gong Li

From Wikipedia, the free encyclopedia

{{Infobox Chinese-language singer and actor
| name= Gong Li
| image=Gong Li Cannes 2011.jpg
| caption= Gong Li at the [[2011 Cannes Film Festival]]
| tradchinesename = {{linktext|鞏|俐}}
| simpchinesename = {{linktext|巩|俐}}
| pinyinchinesename = Gǒng Lì
| birth_date = {{Birth date and age|1965|12|31|df=y}}
| birth_place = [[Shenyang]], China
| nationality = [[Singapore]]an
| spouse = Ooi Hoe Soeng (1996–2010)

Gong Li

Wrapper Induction:
- **generate regex rules from markup examples (with visual tools)**
- **alt. crowdsourcing**

Works with many structured web sources

harvest by extraction rules:
- **regex matching**
- **type checking**

Chinese name	鞏俐 (Traditional)
Chinese name	巩俐 (Simplified)
Pinyin	Gǒng Lì (Mandarin)
Born	31 December 1965 (age 46) Shenyang, China
Spouse(s)	Ooi Hoe Soeng (1996–2010)

```
(?i)IBL\|BEG\s*birth[_ ]?place\s*=\s*(.*?)IBL\|END“
 => "$0 wasBornIn @WikiLink($1)
```

Relational Facts from Text

composed (<musician>, <song>) **appearedIn (<song>, <film>)**

Bob Dylan wrote the song Knockin' on Heaven's Door
Lisa Gerrard wrote many haunting pieces, including Now You Are Free
Morricone's masterpieces include the Ecstasy of Gold
Dylan's song Hurricane was covered by Ani DiFranco
Strauss's famous work was used in 2001, titled Also sprach Zarathustra
Frank Zappa performed a jazz version of Rota's Godfather Waltz
Hallelujah, originally by Cohen, was covered in many movies, including Shrek

composed (Bob Dylan, Knockin' on Heaven's Door)
composed (Lisa Gerrard, Now You Are Free)
...
appearedIn (Knockin' on Heaven's Door, Billy the Kid)
appearedIn (Now You Are Free, Gladiator)
...

Pattern-based Gathering (statistical evidence) **+** **Constraint-aware Reasoning (logical consistency)**

Pattern-based Harvesting: Fact-Pattern Duality

Task populate relation composed starting with seed facts

[Brin 1998, Etzioni 2004, Agichtein/Gravano 2000]

Facts & *Fact Candidates*	**Patterns**
(Dylan, Knockin)	X wrote the song Y
(Gerrard, Now)	X wrote … including Y
(Dylan, Hurricane)	X covered the story of Y
(Morricone, Ecstasy)	X has favorite movie Y
(Zappa, Godfather)	X is famous for Y
(Mann, Buddenbrooks)	…
(Gabriel, Biko)	
(Puebla, Che Guevara)	
(Mezrich, Zuckerberg)	
(Jobs, Apple)	
(Newton, Gravity)	

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

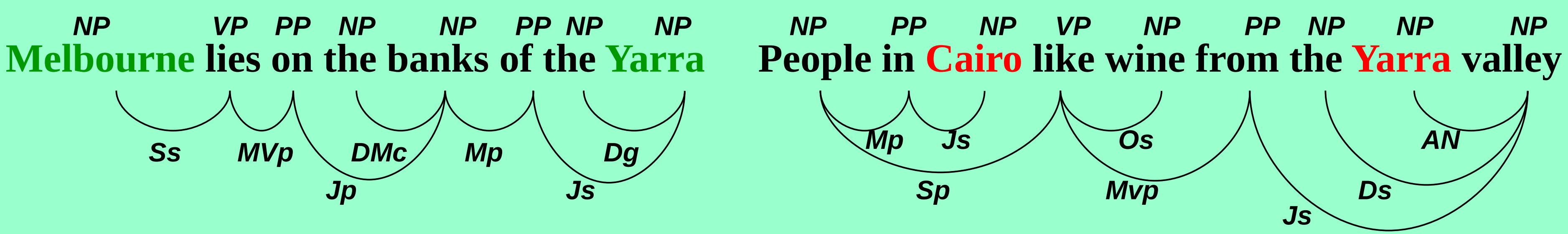
Improving Pattern Precision or Recall

- **Statistics for confidence:**
  occurrence frequency with seed pairs
  distinct number of pairs seen

- **Negative seeds for confusable relations:**
  capitalOf(city,country) → X is the largest city of Y
  pos. seeds: (Paris, France), (Rome, Italy), (New Delhi, India), ...
  neg. seeds: (Sydney, Australia), (Istanbul, Turkey), ...

- **Generalized patterns with wildcards and POS tags:**
  hasAdvisor(student,prof) → X met his celebrated advisor Y
  → X * PRP ADJ advisor Y

- **Dependency parsing for complex sentences:**



Constrained Reasoning for Logical Consistency

Use knowledge (consistency constraints) for joint reasoning on hypotheses and pruning of false candidates

Hypotheses:

composed (Dylan, Hurricane)
composed (Morricone, Ecstasy)
~~composed (Zappa, Godfather)~~
composed (Rota, Godfather)
composed (Gabriel, Biko)
~~composed (Mann, Buddenbrooks)~~
~~composed (Jobs, Apple)~~
~~composed (Newton, Gravity)~~

Constraints:

$\forall$ x, y: composed (x,y) $\Rightarrow$ type(x)=musician
$\forall$ x, y: composed (x,y) $\Rightarrow$ type(y)=song
$\forall$ x, y, z: composed (x,y) $\wedge$ appearedIn(y,z) $\Rightarrow$ wroteSoundtrackFor (x,z)
$\forall$ x,y,t,b,e: composed (x,y) $\wedge$ composedInYear (y, t) $\wedge$ bornInYear (x, b) $\wedge$ diedInYear (x,e) $\Rightarrow$ $b < t \leq e$
$\forall$ x, y, w: composed (x,y) $\wedge$ composed(w,y) $\Rightarrow$ x = w
$\forall$ x, y: sings(x,y) $\wedge$ type(x,singer-songwriter) $\Rightarrow$ composed(x,y)

consistent subset(s) of hypotheses ("possible world(s)", "truth")
$\rightarrow$ Weighted MaxSat solver for set of logical clauses
$\rightarrow$ max a posteriori (MAP) for probabilistic factor graph

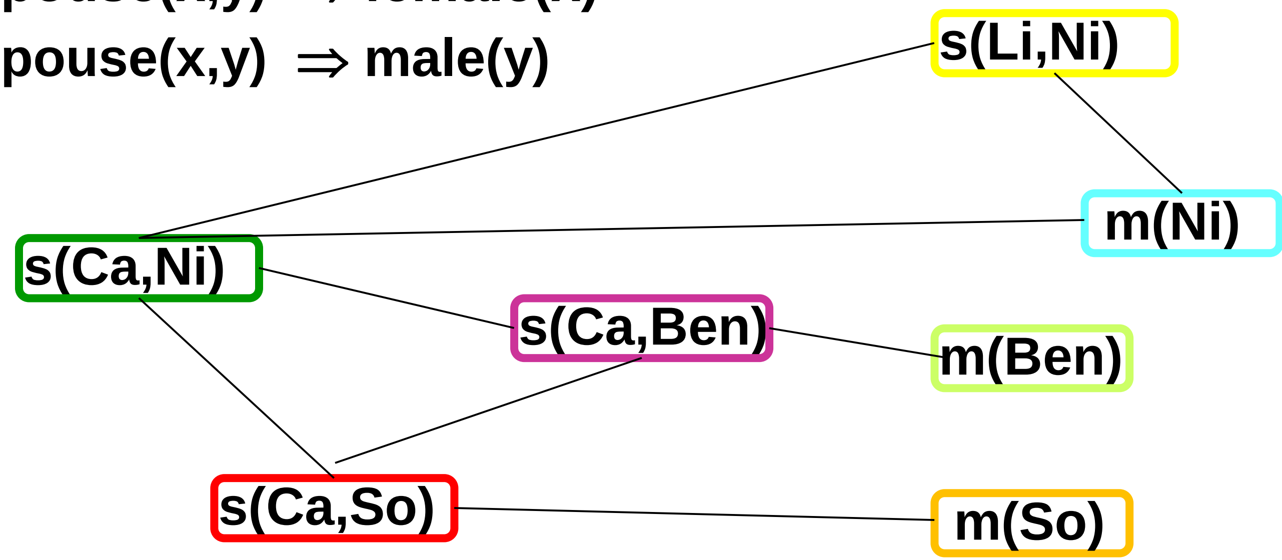
Markov Logic Networks (MLN's)

(M. Richardson / P. Domingos 2006)

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

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

s(Carla,Nick)
s(Lisa,Nick)
s(Carla,Ben)
s(Carla,Sofie)
…



RVs coupled by MRF edge if they appear in same clause

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

joint distribution has product form over all cliques

Variety of algorithms for joint inference:
Gibbs sampling, other MCMC, belief propagation, …
MAP inference equivalent to Weighted MaxSat

Related Alternative Probabilistic Models

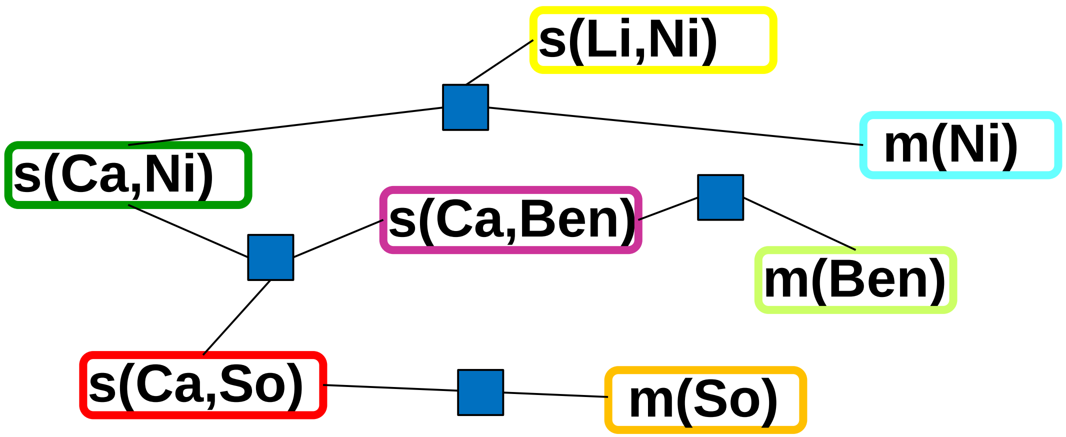
Constrained Conditional Models [Roth et al.]

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

Factor Graphs with Imperative Variable Coordination

[A. McCallum et al.]

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



Probabilistic Soft Logic (PSL) [L. Getoor et al.]

gains MAP efficiency by continuous RV's (degree of truth)

Goal: Discovering "Unknown" Knowledge

so far KB has explicit model:

- **canonicalized entities**
- **relations with type signatures <entity1, relation, entity2>**

< CarlaBruni marriedTo NicolasSarkozy> $\in$ Person $\times$ R $\times$ Person

< NataliePortman wonAward AcademyAward > $\in$ Person $\times$ R $\times$ Prize

Open and Dynamic Knowledge Harvesting:

would like to discover new entities and new relation types

<name1, phrase, name2>

Madame Bruni in her happy marriage with the French president …

The first lady had a passionate affair with Stones singer Mick …

Natalie was honored by the Oscar …

Bonham Carter was disappointed that her nomination for the Oscar …

Open IE with ReVerb [A. Fader et al. 2011, T. Lin 2012, Mausam 2012]

Consider all verbal phrases as potential relations and all noun phrases as arguments

Problem 1: incoherent extractions

"New York City has a population of 8 Mio" → <New York City, has, 8 Mio>

"Hero is a movie by Zhang Yimou" → <Hero, is, Zhang Yimou>

Problem 2: uninformative extractions

"Gold has an atomic weight of 196" → <Gold, has, atomic weight>

"Faust made a deal with the devil" → <Faust, made, a deal>

Problem 3: over-specific extractions

"Hero is the most colorful movie by Zhang Yimou"

→ <..., is the most colorful movie by, …>

Solution:

- **regular expressions over POS tags:**
  VB DET N PREP; VB (N | ADJ | ADV | PRN | DET)* PREP; etc.
- **relation phrase must have # distinct arg pairs > threshold**

Open IE with ReVerb

Open Information Extraction

?x „composed by“ ?y

Hosted by AI2 · Created at

Argument 1: ______ **Relation:** composed by **Argument 2:** ______ All ▾ Search

333 answers from **977 sentences** (cached)

all | person (101) | composer (91) | music contributor (84) | group member (79) | award nominee (44) | misc. | more types

- The Whole , several parts (23)
- **Apartment** , a living room (22)
- **Psalms** , **David** (21)
- The music , **Pritam** (18)
- Music , **Devi Sri Prasad** (16)
- **Apartment** , **Bedroom** (13)
- music , **Chakri Dynasty** (12)
- music , **Mani Sharma** (12)
- Rockstar music , India (9)
- The music , **A. R. Rahman** (9)
- your It , 03cheapairmax 11.21.2011 (9)
- **Music** , one (8)
- your It , 03cheapairmax 11.20.2011 (8)
- The music , Calixa Lavalle (8)
- The music , Sajid-Wajid (7)
- the music , **Alan Menken** (7)
- **Song** , the band (7)

http://openie.cs.washington.edu
http://openie.allenai.org

Open IE with ReVerb

http://openie.cs.washington.edu

http://openie.allenai.org

NELL: Never-Ending Language Learning

[Carlson et al. 2010, Mitchell et al. 2015]

- **Philosophy: learn entire KB „ab initio“ and continue learning**
- **Start with manually specified classes, typed relations, and seed instances, plus constraints for „coupling“**

Coupled Pattern Learner:
iteratively learns patterns for classes and relations
downtown Y; X mayor of Y // with functional mutex

Coupled SEAL (wrapper induction & set expansion):
queries Web and learns extraction patterns from lists & tables
<tr><td>X</td><td>∗ °C</td></tr>; <h4>capitals Y: X …

Coupled classifiers per class: accept/reject noun phrases based on linguistic features and context // with mutex

Rule learner: Horn clauses on classes & relations
leaderOf(X,Y) $\wedge$ city(Y) $\Rightarrow$ mayorOf(X,Y); mayorOf(X,Y) $\Rightarrow$ livesIn(X,Y)

http://rtw.ml.cmu.edu

NELL: Never-Ending Language Learning

50 Mio. SPO assertions, 2.5 Mio high confidence

Recently-Learned Facts

instance	iteration	date learned	confidence
bresaola is a visualizable thing	922	05-may-2015	96.4
francis_derwent_wood is a visual artist	922	05-may-2015	99.9
frank_g is an Australian person	922	05-may-2015	92.2
g_protein_coupled_receptor_124 is a protein	922	05-may-2015	100.0
n_butyl_benzyl_phthalate is a chemical	922	05-may-2015	100.0
chicken001 eat potatoes	926	20-may-2015	100.0
bioinformatics is an academic program at the university college	922	05-may-2015	93.8
samuel_j__palmisano is the CEO of ibm	926	20-may-2015	100.0
national001 is a company that has an office in the country czech_republic	922	05-may-2015	99.2
the companies dc and fox_news_channel compete with eachother	922	05-may-2015	98.4

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

NELL: Never-Ending Language Learning

50 Mio. SPO assertions, 2.5 Mio high confidence

NELL Knowledge Base Browser

CMU Read the Web Project

Search

log in | preferences | help/instructions | feedback

- sportsleague
- tradeunion
- nonprofitorganization
- person
  - monarch
  - astronaut
  - personbylocation
    - personnorthamerica
      - personcanada
      - personus
        - politicianus
      - personmexico
    - personeurope
    - personaustralia
    - personafrica
    - personsouthamerica
    - personasia
    - personantarctica
  - visualartist
  - model
  - scientist
  - journalist
  - female
  - actor
  - professor
  - director
  - architect
  - politician
    - politicianus
  - athlete
  - musician
  - chef
  - male
  - writer
  - ceo
  - judge
  - mlauthor
  - coach
  - celebrity

nick_cave (musician)

literal strings: NICK CAVE, nick cave, Nick cave, Nick Cave

Help NELL Learn!

NELL wants to know if this belief is correct.
If it is or ever was, click thumbs-up. Otherwise, click thumbs-down.

- nick_cave is a musician

---

categories

- musician(98.7%)
  - MBL @865 (96.9%) on 25-aug-2014 [Promotion of celebrity:nick_cave musicianinmusicartist musicartist:bad_seeds]
  - SEAL @623 (57.5%) on 10-aug-2012 [1] using nick_cave

NELL has only weak evidence for items listed in grey

- visualartist
  - SEAL @221 (50.0%) on 18-mar-2011 [1] using nick_cave
- personaustralia
  - SEAL @628 (65.7%) on 26-aug-2012 [1] using nick_cave
- celebrity
  - SEAL @347 (75.0%) on 13-jul-2011 [1 2] using nick_cave

relations

NELL has only weak evidence for items listed in grey

- agentcollaborateswitha
  - john

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

NELL: Never-Ending Language Learning

50 Mio. SPO assertions, 2.5 Mio high confidence

NELL Knowledge Base Browser
CMU Read the Web Project

Search
log in | preferences | help/instructions | feedback

- touristattractionsuchastouristattraction
- celebritysuchascelebrity
- archaeasuchasarchaea
- agriculturalproductincludingagricultura
- plantincludeplant
- actorsuchasactor
- arachnidssuchasarachnids
- mammalsuchasmammal
- architectssuchasarchitects
- companyeconomicsector
- trophywonbycoaches
- agentinvolvedwithitem
  - wineryproduceswine
  - agentworkedondrug
  - producesproducttype
  - producesproduct
    - automakerproducesmodel
  - mlauthorofsoftware
  - musicianplaysinstrument
- animaldevelopdisease
- countrylanguage
- issueofpoliticsbill
- universityoperatesinlanguage
- iteminvolvedwithagent
  - drugworkedonbyagent
  - wineproducedbywinery
  - mlsoftwareauthor
  - producedby
    - automodelproducedbymaker
  - typeproducedby
  - instrumentplayedbymusician
- dateof
  - dateatwhichexistsitem
  - dateevent
    - dateofsportsgame
    - dateofmeetingeventtitle

musicianplaysinstrument

(relation: domain musician, range musicinstrument)

Specifies that a musical instrument is played by a paritcular musician

See metadata for musicianplaysinstrument
712 instances, 1 page

instance	iteration	date learned	confidence
adam, drums	799	27-dec-2013	100.0
adam, guitar	799	27-dec-2013	100.0
bach, piano	551	19-apr-2012	100.0
bach, violin	551	19-apr-2012	100.0
barber, violin	598	21-jun-2012	100.0
bb_king, guitar	680	09-jan-2013	100.0
beethoven, piano	853	11-jul-2014	100.0
beethoven, violin	853	11-jul-2014	100.0
ben_harper, guitar	820	08-mar-2014	100.0
billie_joe_armstrong, guitar	818	03-mar-2014	100.0
brahms, piano	592	13-jun-2012	100.0
brahms, violin	503	06-feb-2012	100.0
buddy_guy, guitar	664	01-dec-2012	100.0
b_b__king, guitar	406	08-sep-2011	(Seed) 100.0
charlie, guitar	724	12-apr-2013	100.0
chopin, piano	683	15-jan-2013	100.0
copland, piano	665	05-dec-2012	100.0
david, bass	904	20-feb-2015	100.0
david, drums	904	20-feb-2015	100.0
david, guitar	904	20-feb-2015	100.0
david, keyboards	904	20-feb-2015	100.0
earl_scruggs, banjo			
eddie, guitar			

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

Paraphrases of Relations

composed (<musician>, <song>) **covered (<musician>, <song>)**

Dylan wrote his song Knockin' on Heaven's Door, a cover song by the Dead
Morricone 's masterpiece is the Ecstasy of Gold, covered by Yo-Yo Ma
Amy's souly interpretation of Cupid, a classic piece of Sam Cooke
Nina Simone's singing of Don't Explain revived Holiday's old song
Cat Power's voice is sad in her version of Don't Explain
Cale performed Hallelujah written by L. Cohen

covered by: (Amy,Cupid), (Ma, Ecstasy), (Nina, Don't), (Cat, Don't), (Cale, Hallelujah)

voice in version of: (Amy,Cupid), (Sam, Cupid), (Cat, Don't), (Cale, Hallelujа

performed: (Amy,Cupid), (Amy, Black), (Nina, Don't), (Co n, Hallelujah), (Dylan, Knockin), …

Statistical analysis yields (near-)equivalence classes of paraphrases

covered (<musician>, <song>):
cover song, interpretation of, singing of, voice in … version , …

composed (<musician>, <song>):
wrote song, classic piece of, 's old song, written by, composition of, …

PATTY: Pattern Taxonomy for Relations

[Nakashole et al.: EMNLP-CoNLL'12, VLDB'12,
Moro et al.: CIKM'12, Grycner et al.: COLING'14]

WordNet-style dictionary/taxonomy for relational phrases based on SOL patterns (syntactic-lexical-ontological)

Relational phrases are typed

<*person*> graduated from <*university*>
<*singer*> covered <*song*> <*book*> covered <*event*>

Relational phrases can be synonymous

"graduated from" ⇔ "obtained degree in * from"
"and PRP ADJ advisor" ⇔ "under the supervision of"

One relational phrase can subsume another

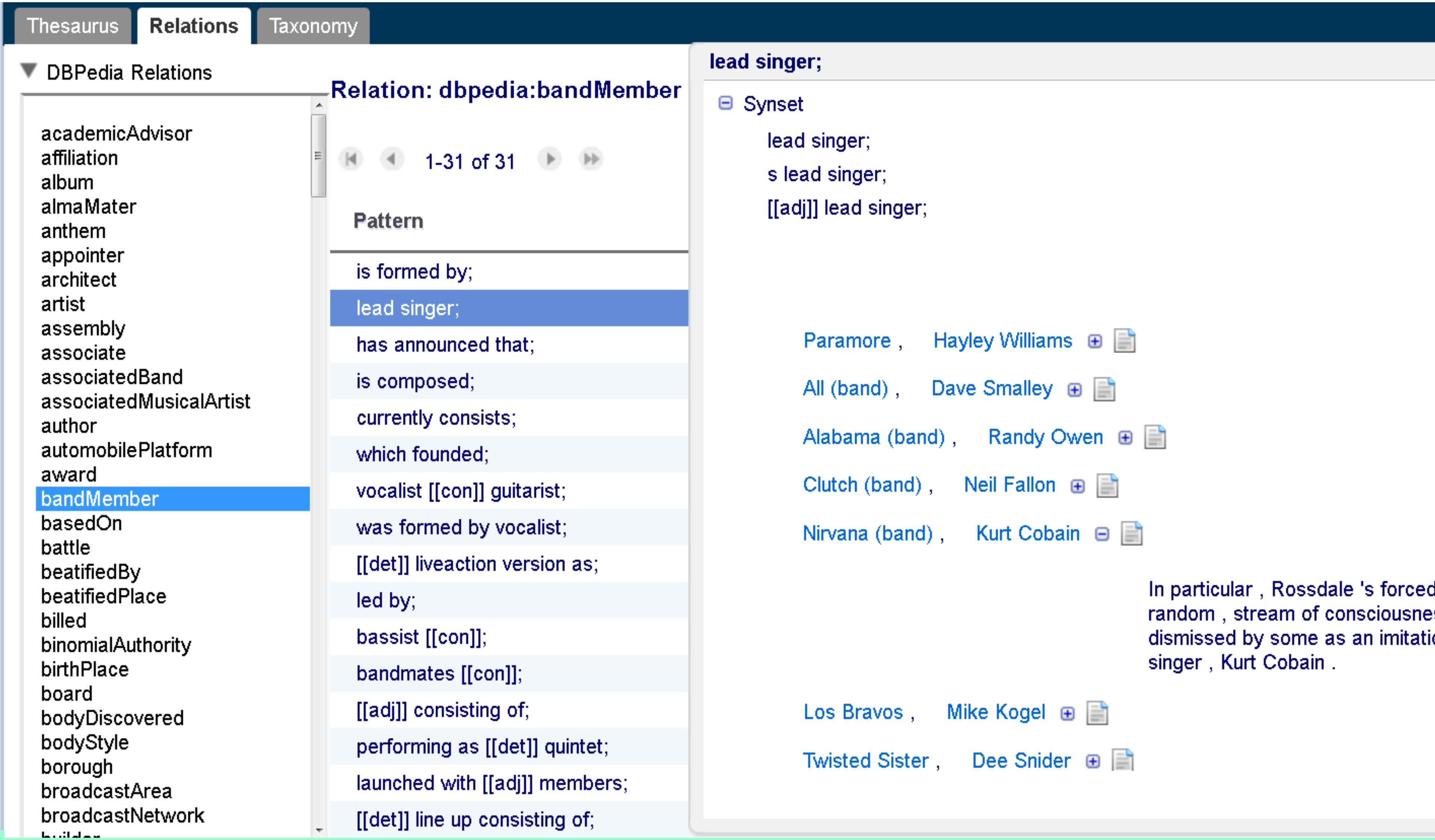
"wife of" ⇒ " spouse of"

350 000 SOL patterns from Wikipedia, NYT archive, ClueWeb

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

PATTY: Pattern Taxonomy for Relations

[N. Nakashole et al.: EMNLP 2012, VLDB 2012]



350 000 SOL patterns with 4 Mio. instances
accessible at: www.mpi-inf.mpg.de/yago-naga/patty

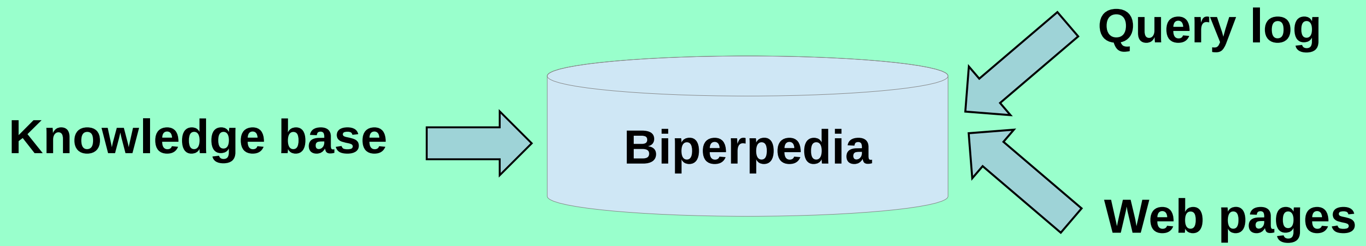
Paraphrases of Attributes: Biperpedia

[M. Gupta et al.: VLDB'14]

Motivation: understand and rewrite/expand web queries

Goal: Collect attributes (birth place, spouse, population, height, etc.)
Determine domain, range, sub-attributes, synonyms, misspellings

Ex.: capital → domain = countries, range = cities,
synonyms = capital city,
misspellings = capitol, ...,
sub-attributes = former capital, fashion capital, ...



- **Candidates from noun phrases**
  („CEO of Google", „population of Melbourne")
- **Discover sub-attributes (by textual refinement, Hearst patterns, …)**
- **Attach attributes to classes in KB: many instances in common**
- **Label attributes as numeric/text/set**
  (verbs as cues: „increasing" → numeric)

Knowledge Graphs:
Large Size, Good Coverage, High Quality, Open-World

Are We Done?

Wikipedia Test

Which salient facts about an entity are captured in infobox?

Johnny Cash

Cash in 1955

Born	J. R. Cash February 26, 1932 Kingsland, Arkansas, U.S.
Died	September 12, 2003 (aged 71) Nashville, Tennessee, U.S.
Cause of death	Diabetes mellitus
Occupation	Singer-songwriter, actor
Years active	1954–2003
Spouse(s)	Vivian Liberto (m. 1954; div. 1966) June Carter (m. 1968–2003; her death)
Children	Rosanne (1955–) Carlene (1955–) Kathy (1956–) Rosie (1958–2003) Cindy (1959–) Tara (1961–) John (1970–)

Military career

Allegiance	United States of America
Service/branch	United States Air Force
Years of service	1950–1954
Rank	Staff sergeant

Musical career

Genres	Country, rockabilly,[1] rock and roll, gospel
Instruments	Vocals, guitar
Labels	Sun, Columbia, Mercury, American, House of Cash, Legacy Recordings
Associated acts	The Tennessee Three, The Highwaymen, June Carter Cash, The Statler Brothers, The Carter Family , Waylon Jennings
Website	johnnycash.com

Notable instruments

Martin Acoustic Guitars[2]

JohnnyCash

- **longRomanticRelationshipWith**
  - **JuneCarter [1961-2003]**
- **featuredInMovie**
  - **WalkTheLine**
- **playedConcertIn**
  - **SanQuentinStatePrison**
- **adored**
  - **JackCash (elder brother)**
- **covered**
  - **One (by U2)**
  - **MercySeat (by Nick Cave)**

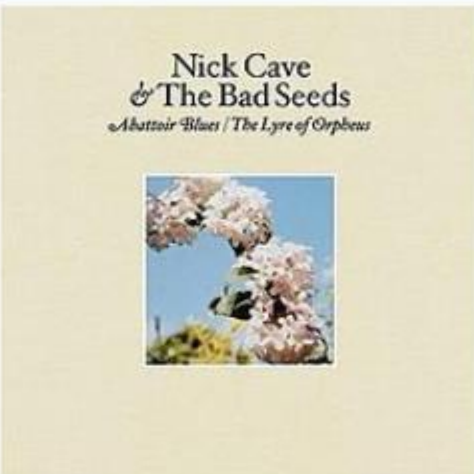
.....

not in any KG !

Wikipedia Test

Which salient facts about an entity are captured in infobox?

Abattoir Blues / The Lyre of Orpheus



Studio album by Nick Cave and the Bad Seeds

Released	20 September 2004
Recorded	March–April 2004 at Studio Ferber in Paris, France
Genre	Alternative rock
Length	82:30
Label	Mute
Producer	Nick Launay

Nick Cave and the Bad Seeds chronology

Nocturama (2003)	***Abattoir Blues / The Lyre of Orpheus*** (2004)	*B-Sides & Rarities* (2005)

Singles from *Abattoir Blues / The Lyre of Orpheus*

1. "Nature Boy"
   Released: 6 September 2004
2. "Breathless / There She Goes, My Beautiful World"
   Released: 15 November 2004
3. "Get Ready for Love"
   Released: 14 March 2005

Let the Bells Ring
- **isOnAlbum**
  - **AbbatoirBlues**
- **lyricsAbout**
  - **JohnnyCash**

O'Children
- **isOnAlbum**
  - **AbbatoirBlues**
- **featuredInMovie**
  - **HarryPotter & Deathly Hallows 1**

WarrenEllis
- **performsOnAlbum**
  - **AbbatoirBlues**

.....

not in any KG !

Watson Test

How many Jeopardy questions could be answered having solely Yago+Dbpedia+Freebase?

24-Dec-2014: http://www.j-archive.com/showgame.php?game_id=4761
Categories: Alexander the Great, Santa‘s Reindeer Party, Making Some Coin, TV Roommates, The „NFL“

- **Alexander the Great was born in 356 B.C. to King Philip II & Queen Olympias of this kingdom**
  (Macedonia)
- **Against an Indian army in 326 B.C., Alexander faced these beasts, including the one ridden by King Porus**
  (elephants)
- **In 2000 this Shoshone woman first graced our golden dollar coin**
  (Sacagawea)
- **When her retirement home burned down in this series, Sophia moved in with her daughter Dorothy and Rose & Blanche**
  (The Golden Girls)
- **Double-winged "mythical" insect**
  (dragonfly)

Lessons Learned

Size & Coverage:
pick low-hanging fruit first, then tackle difficult terrain

Scale:
pattern-centric methods scale out well
reasoning and advanced learning form bottleneck

Quality:
consistency reasoning is key asset
1 hour of domain modeling can beat 1 year of ML

Open-World:
Need to discover emerging entities, new relations, …
Quality gap between model-driven harvesting and Open IE

What’s Next

Further improve algorithms
for consistency reasoning (KG quality)
Combine with human curation (active learning)

Construct more background resources:
relational paraphrases, statistics on user queries, …

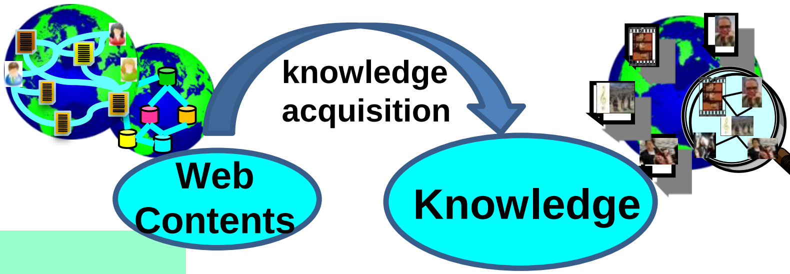
Reconcile and integrate
model-driven harvesting and Open IE

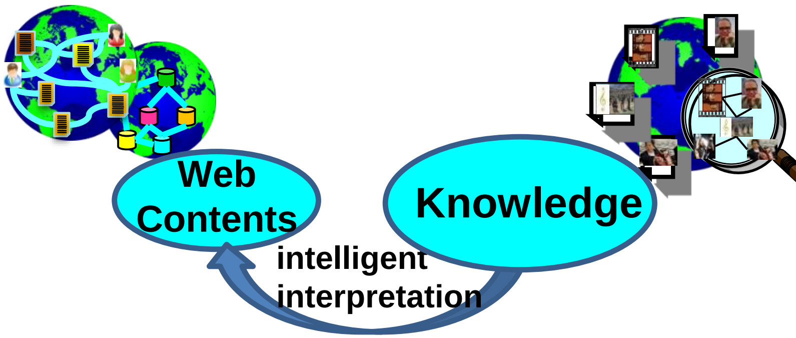
Detect and acquire truly informative facts

Domain-specific and on-the-fly KGs
(music, health, politics, …
for journalists, analystes, …)

Outline

- ✓ Introduction
- ✓ KG Construction
- ★ Refined Knowledge
- ★ Knowledge for Language
- ★ Deep Text Analytics
- ★ Search for Knowledge
- ★ Conclusion





Goal: Temporal Knowledge

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

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

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

- US president's wife when Steve Jobs died?
- students of Hector Garcia-Molina while he was at Princeton?

Temporal Knowledge is Challenging

for all people in Wikipedia (> 500 000) 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**

consistency constraints are potentially helpful:

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

Dating Considered Harmful

explicit dates vs. implicit dates

Nicolas Sarkozy

From Wikipedia, the free encyclopedia

Nicolas Sarkozy (pronounced [ni.kɔ.la saʁ.kɔ.zi] (listen), born **Nicolas Paul Stéphane Sarközy de Nagy-Bocsa**; 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 the 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.

Machine-Reading Biographies

Early life

vague dates
relative dates

During Sarkozy's childhood, his father allegedly refused to give his wife
help, even though he had founded his own advertising agency and had become wealthy. The family lived in a mansion owned by Sarkozy's grandfather, Benedict Mallah, in the 17th Arrondissement of Paris. The family later moved to Neuilly-sur-Seine, one of the wealthiest

Education

narrative text
relative order

Sarkozy was enrolled in the *Lycée Chaptal*, a well regarded public midd
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,[9] but where he nonetheless obtained his *baccalauréat* in 1973. He enrolled at the *Université Paris X Nanterre*, where he graduated with an MA 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.[10] After graduating, he entered the *Institut d'Études Politiques de Paris*, better known as Sciences Po, (1979–1981) but failed to graduate[11] due to an insufficient

Methods and Tools for Explicit Dates

[M. Verhagen et al., J. Stroetgen & M. Gertz]

Temporal Taggers:

capture temporal expressions and normalize them

- **TARSQI:** http://www.timeml.org/site/tarsqi/
  based on regular expressions
- **HeidelTime:** http://heideltime.ifi.uni-heidelberg.de/heideltime/
  regex, rules, …
  supports many languages,
  captures some implicit dates (e.g. Christmas holidays)

Example: TARSQI Annotations

http://www.timeml.org/site/tarsqi/ (M. Verhagen et al.: ACL'05)

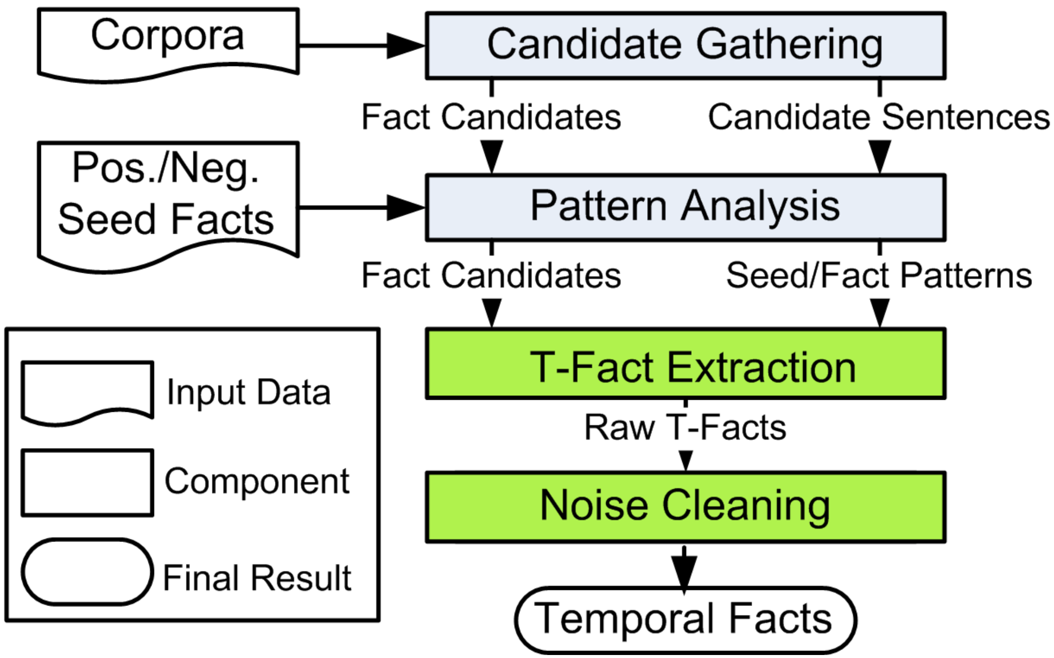
Hong Kong is poised to hold the first election in more than half <TIMEX3 tid="t3" TYPE="DURATION" VAL="P100Y">a century</TIMEX3> that includes a democracy advocate seeking high office in territory controlled by the Chinese government in Beijing. A pro-democracy politician, Alan Leong, announced <TIMEX3 tid="t4" TYPE="DATE" VAL="20070131">Wednesday</TIMEX3> that he had obtained enough nominations to appear on the ballot to become the territory's next chief executive. But he acknowledged that he had no chance of beating the Beijing-backed incumbent, Donald Tsang, who is seeking re-election. Under electoral rules imposed by Chinese officials, only 796 people on the election 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.

Temporal Facts from Text

[Y. Wang et al. 2011]

Temporal Fact = Basic Fact + Temporal Scope

1) **Candidate gathering:**
   extract pattern & entities of basic facts and time expression
2) **Pattern analysis:**
   use seeds to quantify strength of candidates
3) **Label propagation:**
   construct weighted graph of hypotheses and minimize loss function
4) **Constraint reasoning:**
   use ILP for temporal consistency



Reasoning on T-Fact Hypotheses

[Y. Wang et al. 2012, P. Talukdar et al. 2012]

Temporal-fact hypotheses:

s(Ca,Ni)@[2008,2012]{0.7}, s(Ca,Ben)@[2010]{0.8}, s(Ca,Al)@[2007,2008]{0.2}, s(Li,Ni)@[1996,2004]{0.9}, s(Li,Joe)@[2006,2008]{0.8}, …

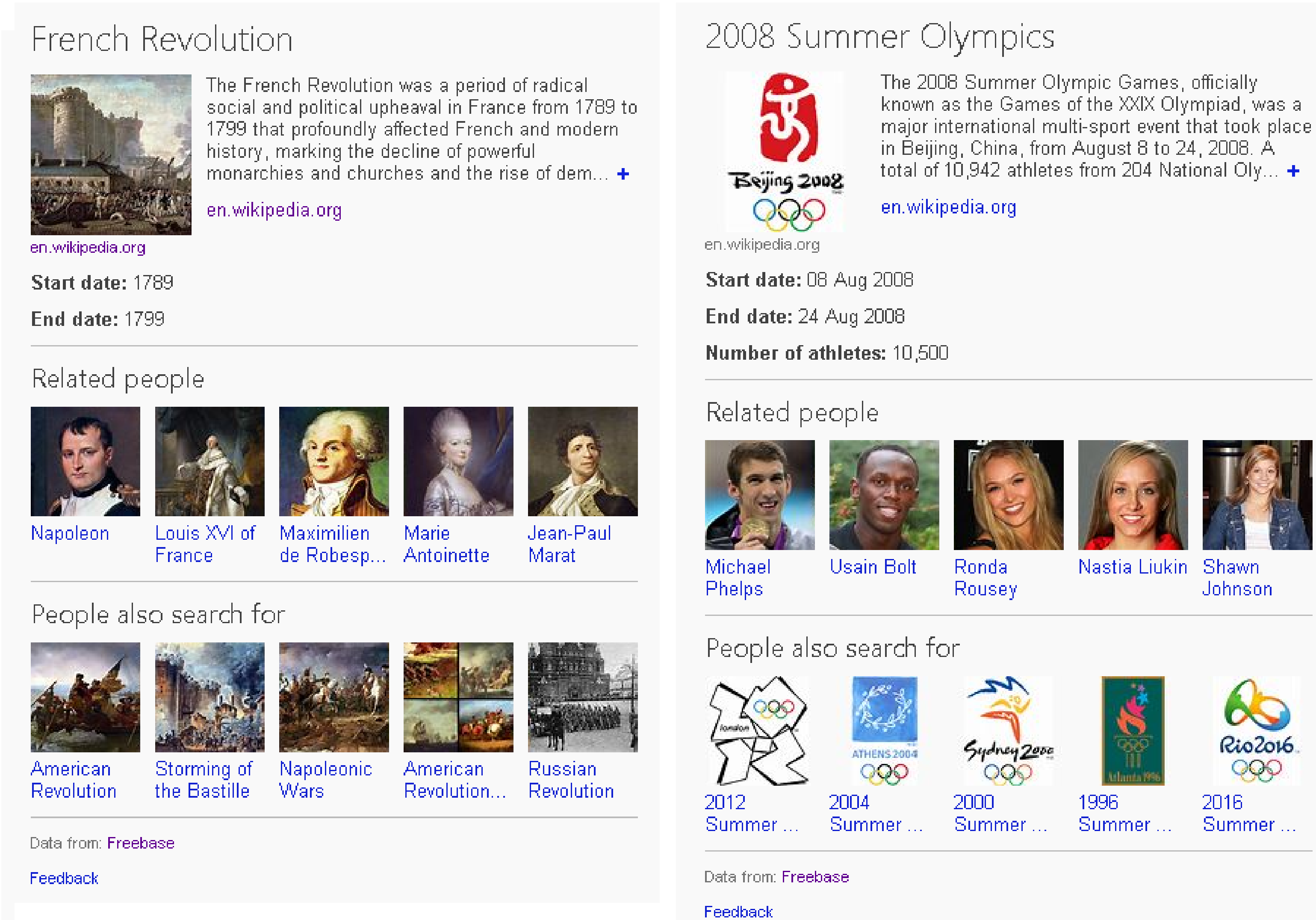
Cast into evidence-weighted logic program or integer linear program with 0-1 variables:

for temporal-fact hypotheses X_i
and pair-wise ordering hypotheses P_{ij}
maximize $\Sigma\, w_i X_i$ with constraints

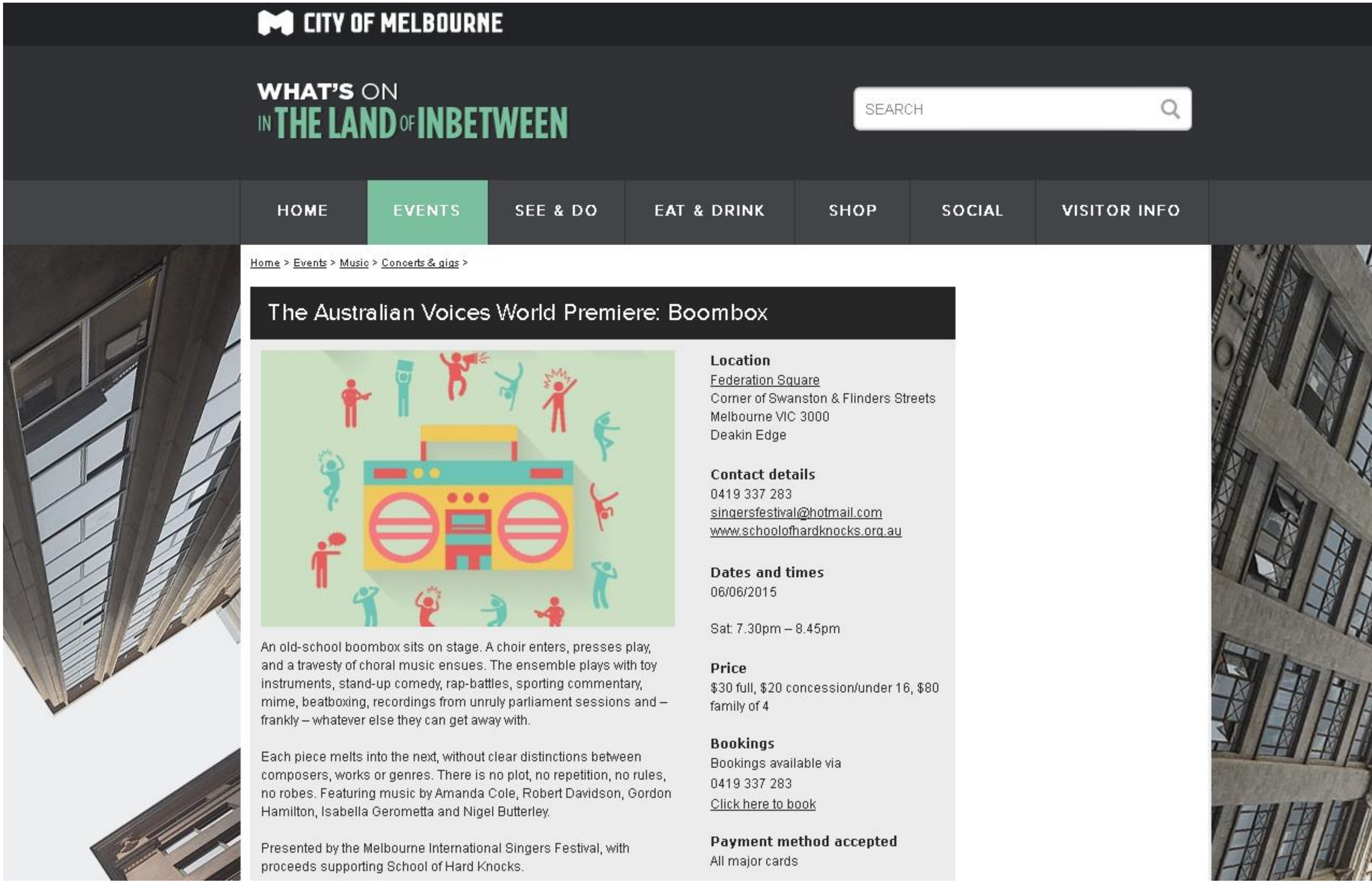
- $X_i + X_j \le 1$
  if X_i, X_j overlap in time & conflict
- $P_{ij} + P_{ji} \le 1$
- $(1 - P_{ij}) + (1 - P_{jk}) \ge (1 - P_{ik})$
  if X_i, X_j, X_k must be totally ordered
- $(1 - X_i) + (1 - X_j) + 1 \ge (1 - P_{ij}) + (1 - P_{ji})$
  if X_i, X_j must be totally ordered

Efficient ILP solvers:
www.gurobi.com
IBM Cplex
…

Events in the Knowledge Base



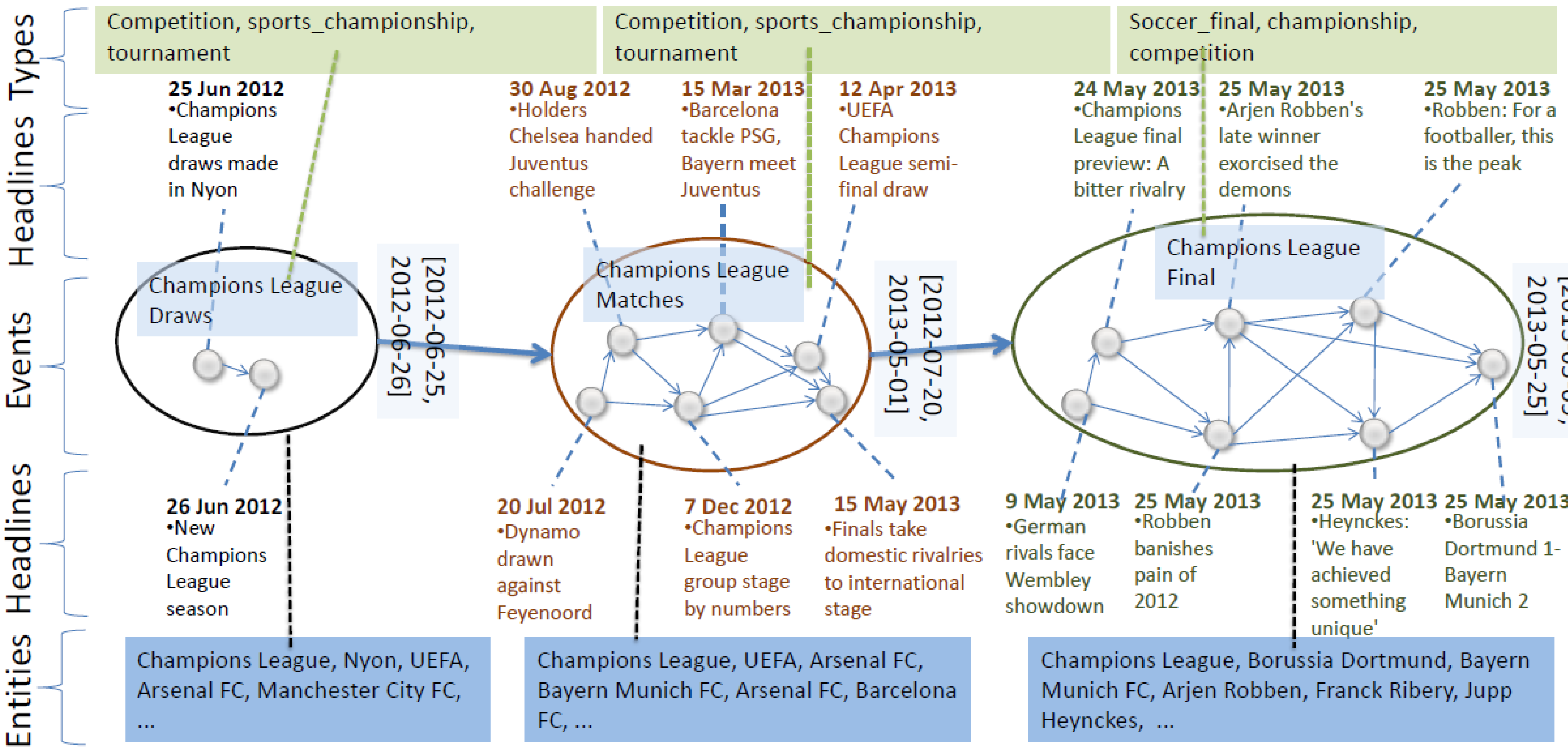
Capturing Emerging & Long-Tail Events



EVIN: Populating KB with Emerging Events

For multi-view attributed graph G compute coarsened graph G* s.t. G = G* + Δ(G,G*) with MDL

E. Kuzey et al. [WWW'14, CIKM'14]

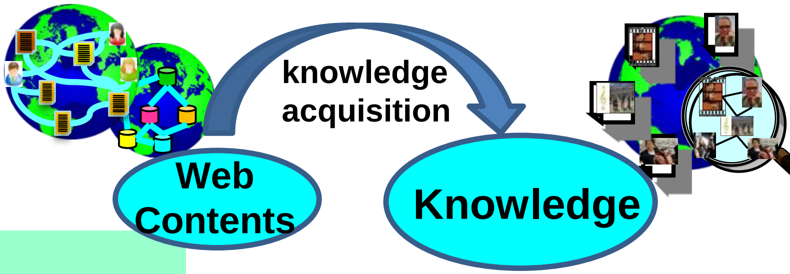


24,000 high-quality events from 300,000 news articles

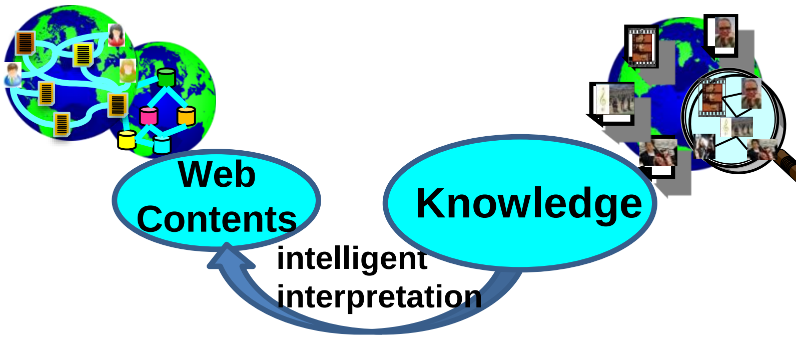
http://www.mpi-inf.mpg.de/yago-naga/evin/

Outline

- ✓ Introduction
- ✓ KG Construction
- ★ Refined Knowledge



- ★ Knowledge for Language
- ★ Deep Text Analytics
- ★ Search for Knowledge



- ★ Conclusion

Goal: Commonsense Knowledge

Every child knows that

apples are green, red, round, juicy, …
but not fast, funny, verbose, …

pots and pans are in the kitchen or cupboard, on the stove, …
but not in in the bedroom, in your pocket, in the sky, …

children usually live with their parents

But: commonsense is rarely stated explicitly
Plus: web and social media have reporting bias

rich family: 27.8 Mio on Bing
poor family: 3.5 Mio on Bing

singers: 22.8 Mio on Bing
workers: 14.5 Mio on Bing

Acquiring Commonsense Knowledge

Approach 1: Crowdsourcing
→ **ConceptNet (Speer/Havasi)**

Problem: coverage and scale

Approach 2: Pattern-based harvesting
→ **WebChild (Tandon et al.)**

Problem: noise and robustness

Crowdsourcing for Commonsense Knowledge

[Speer & Havasi 2012]

many inputs incl. WordNet, Verbosity game, etc.

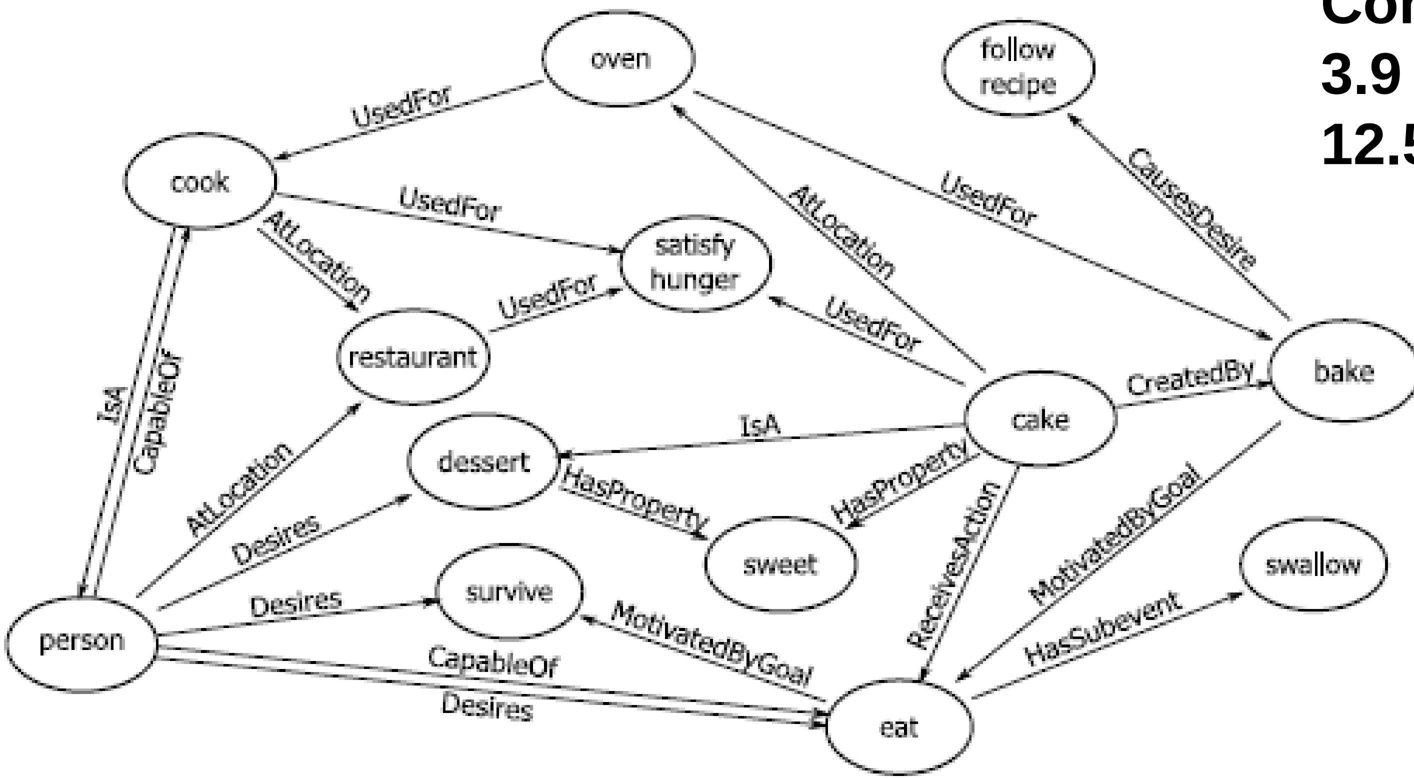
http://www.gwap.com/gwap/

Crowdsourcing for Commonsense Knowledge

[Speer & Havasi 2012]

many inputs incl. WordNet, Verbosity game, etc.

ConceptNet 5:
3.9 Mio concepts
12.5 Mio. edges



http://conceptnet5.media.mit.edu/

Pattern-Based Harvesting of Commonsense Properties

(N. Tandon et al.: AAAI 2011)

Approach 2: Start with seed facts for

apple hasProperty round
dog hasAbility bark
plate hasLocation table

Find patterns that express these relations, such as

X is very Y, X can Y, X put in/on Y, …

Apply these patterns to find more facts.

Problem: noise and sparseness of data
Solution: harness Web-scale n-gram corpora
$\rightarrow$ 5-grams + frequencies

Confidence score: PMI (X,Y), PMI (p,(XY)), support(X,Y), …
are features for regression model

WebChild: Commonsense Properties

[N. Tandon et al.: WSDM'14, AAAI'14]

Who looks hot ?

What tastes hot ?

What is hot ?

What feels hot ?

→ 4 Mio **sense-disambiguated SPO triples** for predicates:

hasProperty, hasColor, hasShape, hasTaste, hasAppearance, isPartOf, hasAbility, hasEmotion, …

- **pattern learning with seeds: high recall**
- **semisupervised label propagation: good precision**
- **Integer linear program: sense disambiguation, high precision**

https://gate.d5.mpi-inf.mpg.de/webchild/

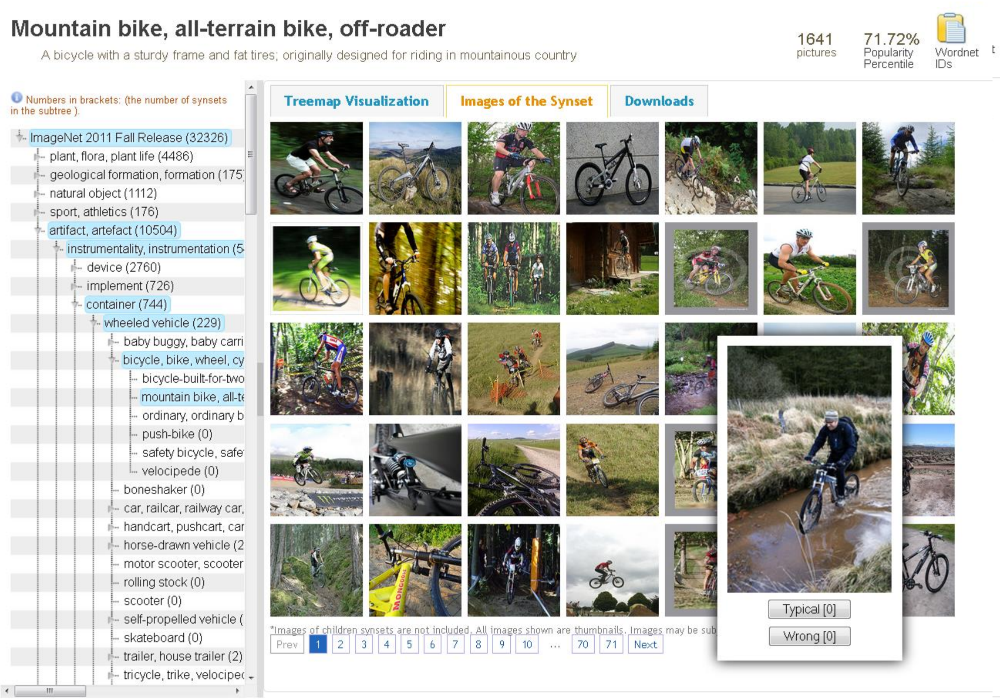
Visual Commonsense

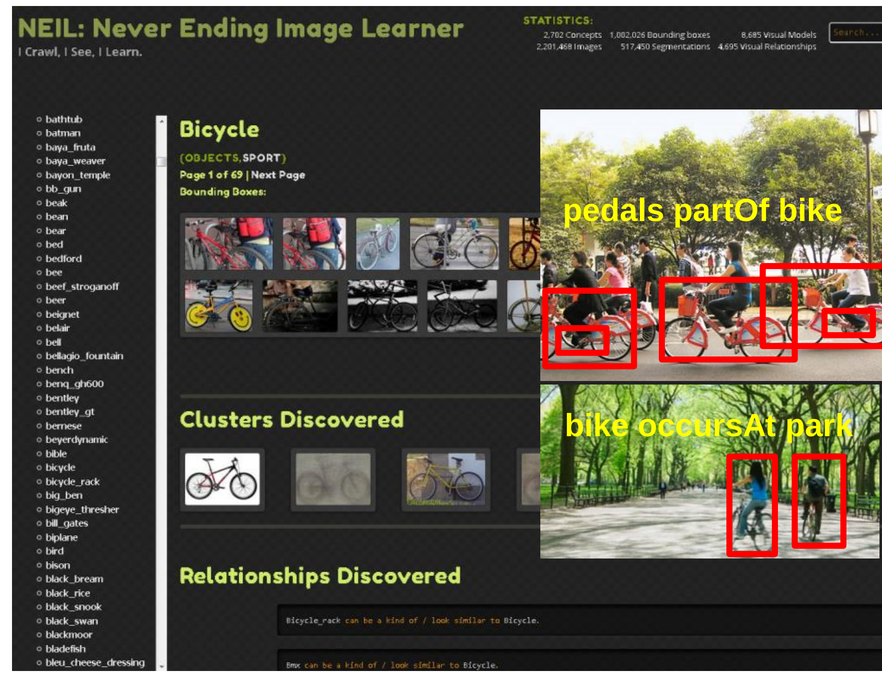
ImageNet: populate WordNet classes with many photos
[J. Deng et al.: CVPR'09]

http://www.image-net.org

NEIL: infer instances of partOf occursAt, inScene relations
[X. Chen et al.: ICCV'13]

http://www.neil-kb.com/





How:
crowdsourcing for seeds, distantly supervised classifiers, object recognition (bounding boxes) in computer vision

Commonsense for Visual Scenes

Knowlywood: **[N. Tandon et al.: WWW'15]**

Activity knowledge from movie&TV scripts, aligned with visual scenes

→ 0.5 Mio activity types with attributes: location, time, participants, prev/next

Refined part-whole relations from web&books text and image tags

→ 6.7 Mio sense-disambiguated triples for physicalPartOf, visualPartOf, hasCardinality, memberOf, substanceOf

Challenge: Commonsense Rules

Horn clauses:
can be learned by Inductive Logic Programming

$\forall$ **x,m,c: type(x,child) $\wedge$ mother(x,m) $\wedge$ livesIn(m,t)) $\Rightarrow$ livesIn(x,t)**
$\forall$ **x,m,f: type(x,child) $\wedge$ mother(x,m) $\wedge$ spouse(m,f) $\Rightarrow$ father(x,f)**

Advance rules beyond Horn clauses:
specified by human experts

$\forall$ **x: type(x,spider) $\Rightarrow$ numLegs(x)=8**
$\forall$ **x: type(x,animal) $\wedge$ hasLegs(x) $\Rightarrow$ even(numLegs(x))**
$\forall$ **x: human(x) $\Rightarrow$ ($\exists$ y: mother(x,y) $\wedge$ $\exists$ z: father(x,z))**
$\forall$ **x: human(x) $\Rightarrow$ (male(x) $\vee$ female(x))**

Commonsense: What Is It Good For?

- **How-to queries:**

  repair a bike tire, pitch a tent, cross a river, …

- **Scene search (over videos, books, diaries):**

  romantic dinner, dramatic climb, …

- **Question disambiguation:**

  hottest battles with JZ ?

  hottest place on earth ?

- **Sentiment analysis:**

  warm beer in cool bar – got the flu

  the bar was cool and the beer, too

  the bar was warm but the beer was cool

  the hot springs are very cool

Lessons Learned

Temporal knowledge:
factual knowledge must and can be positioned in time (and geo-space)

Extracting t-facts seems harder than harvesting basic facts (little low-hanging fruit here?)

Commonsense:
acquiring what every child knows is amazingly hard for machines

What's Next

Advanced algorithms for t-fact extraction:
web scale, relative t-expressions, narrative texts

Dynamically acquire knowledge of
emerging events in news and social media

Understand, support and automate
the long-term maintenance of the KG:
knowledge life-cycle

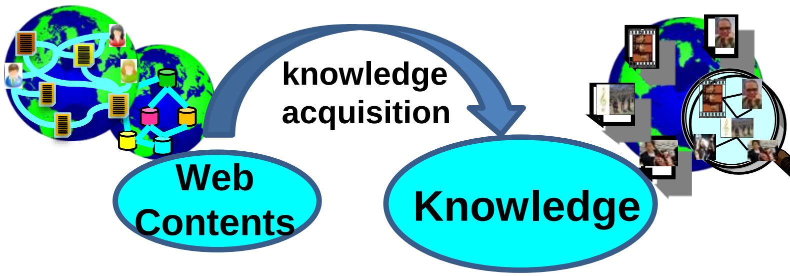
Acquire more and cleaner commonsense:
sophisticated properties, advanced rules, visual scenes

Use cases for temporal and commonsense knowledge:
time-aware info needs, how-to queries, …

Outline

✓ Introduction

✓ KG Construction

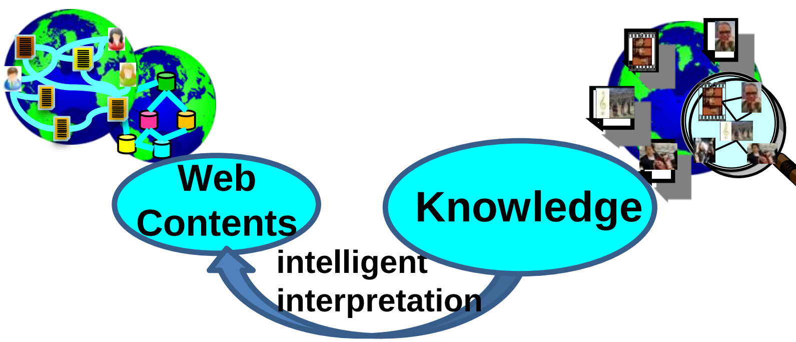


✓ Refined Knowledge

★ **Knowledge for Language**

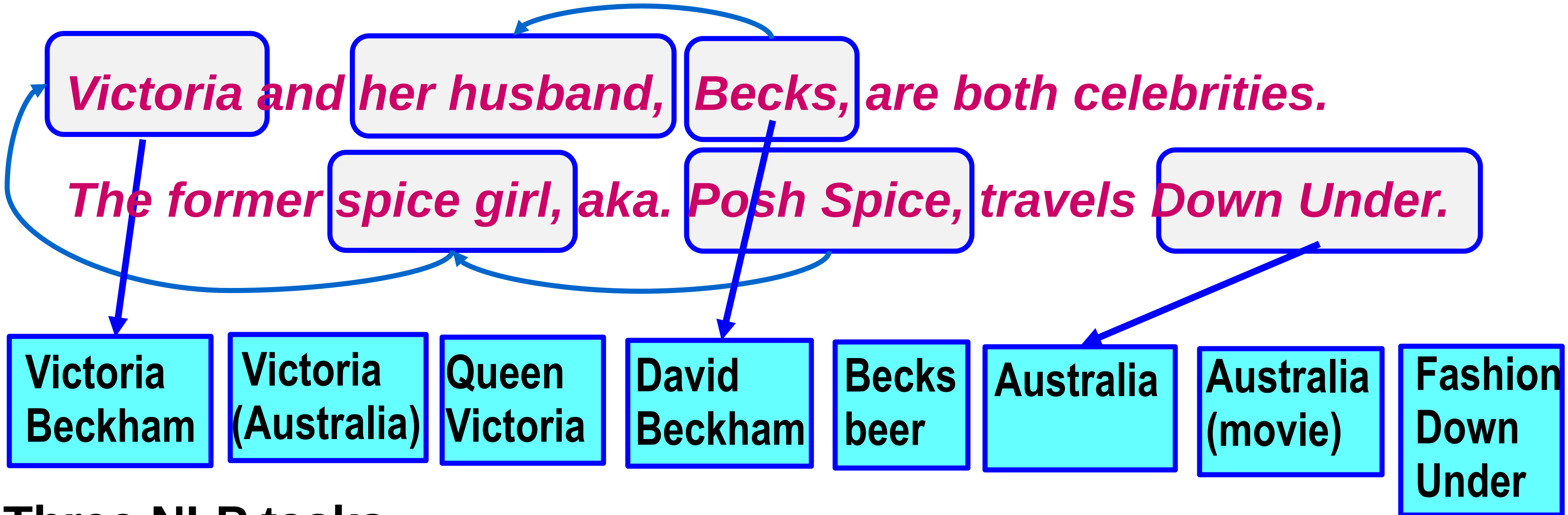
★ Deep Text Analytics

★ Search for Knowledge



★ Conclusion

Goal: Entities, not Names



Three NLP tasks:

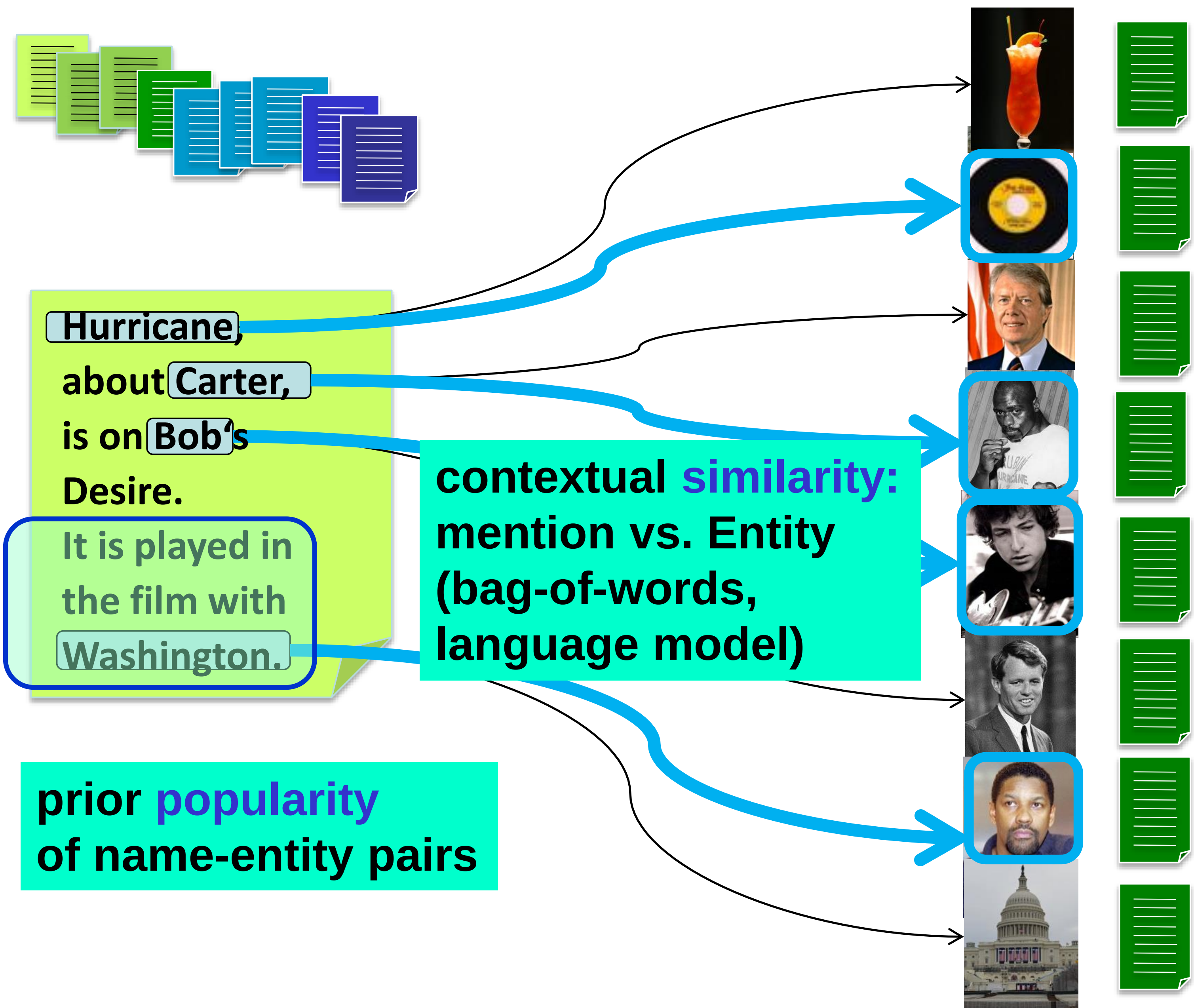
1) named-entity detection: segment & label by HMM or CRF (e.g. Stanford NER tagger)

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)

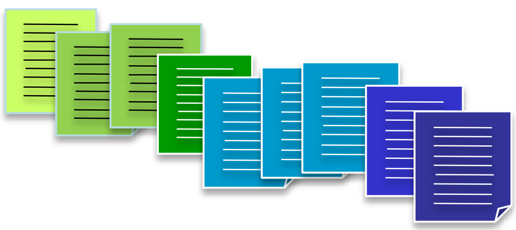
tasks 1 and 3 together: NERD

Named Entity Recognition & Disambiguation

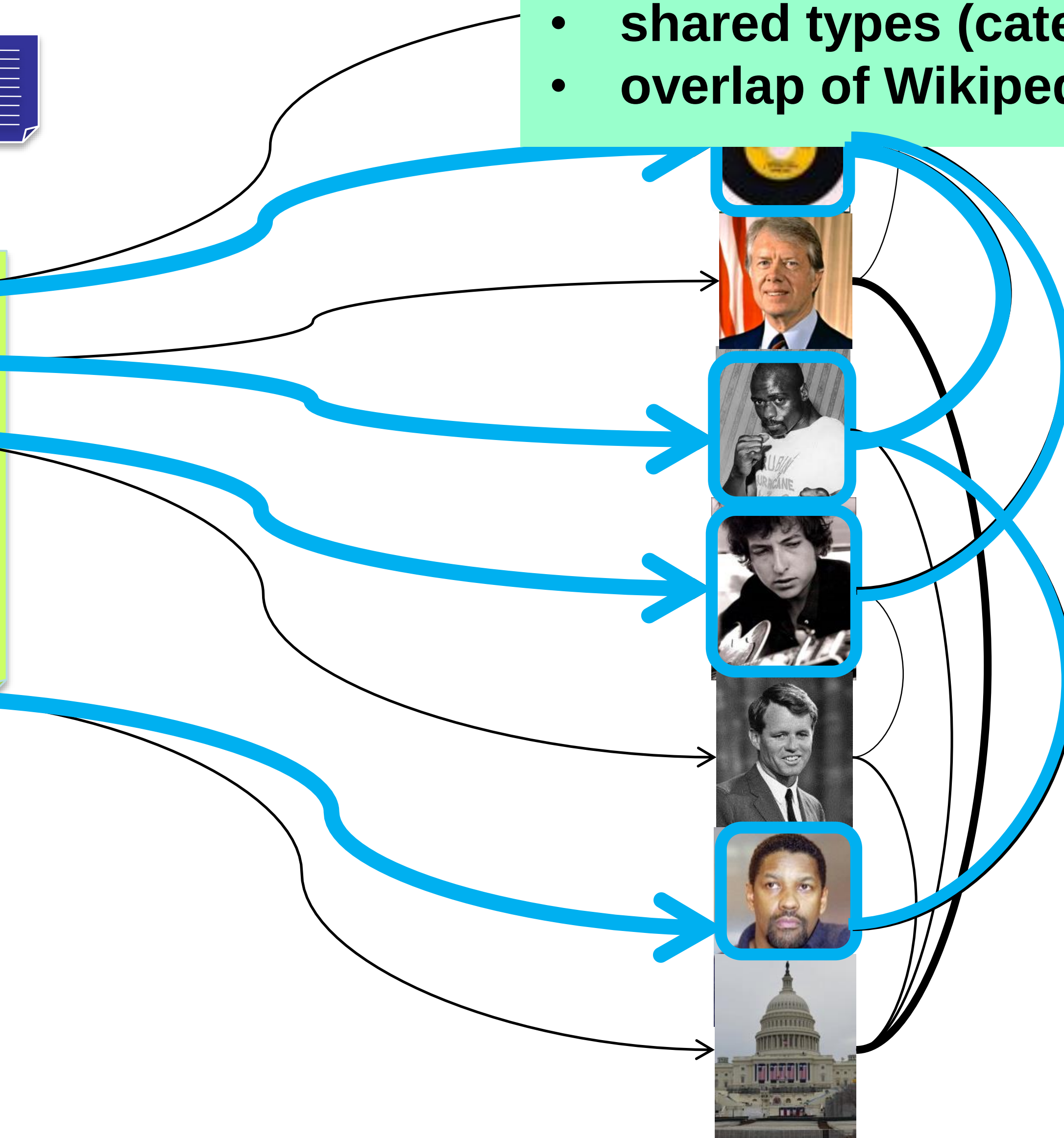


Named Entity Recognition & Disambiguation

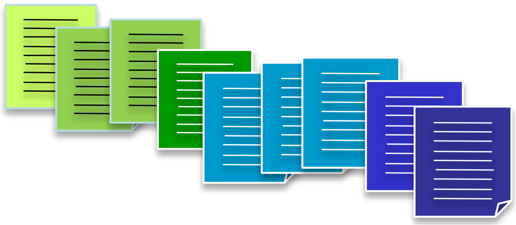
Coherence of entity pairs:
- semantic relationships
- shared types (categories)
- overlap of Wikipedia links



Hurricane,
about Carter,
is on Bob's
Desire.
It is played in
the film with
Washington.



Named Entity Recognition & Disambiguation



Coherence: (partial) overlap of (statistically weighted) entity-specific keyphrases

Hurricane, about Carter, is on Bob's Desire. It is played in the film with Washington.

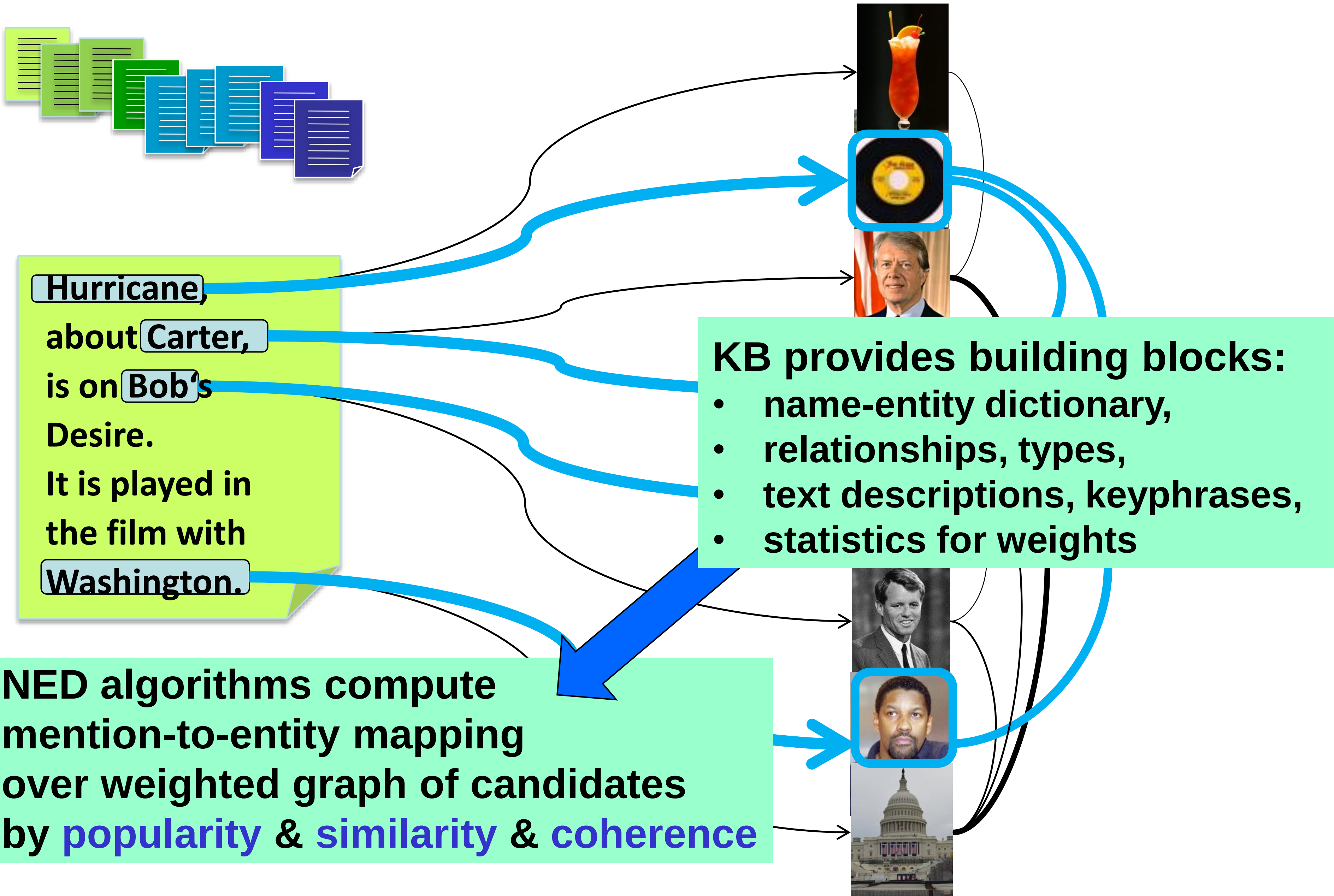
racism protest song
boxing champion
wrong conviction

racism victim
middleweight boxing
nickname Hurricane
falsely convicted

Grammy Award winner
protest song writer
film music composer
civil rights advocate

Academy Award winner
African-American actor
Cry for Freedom film
Hurricane film

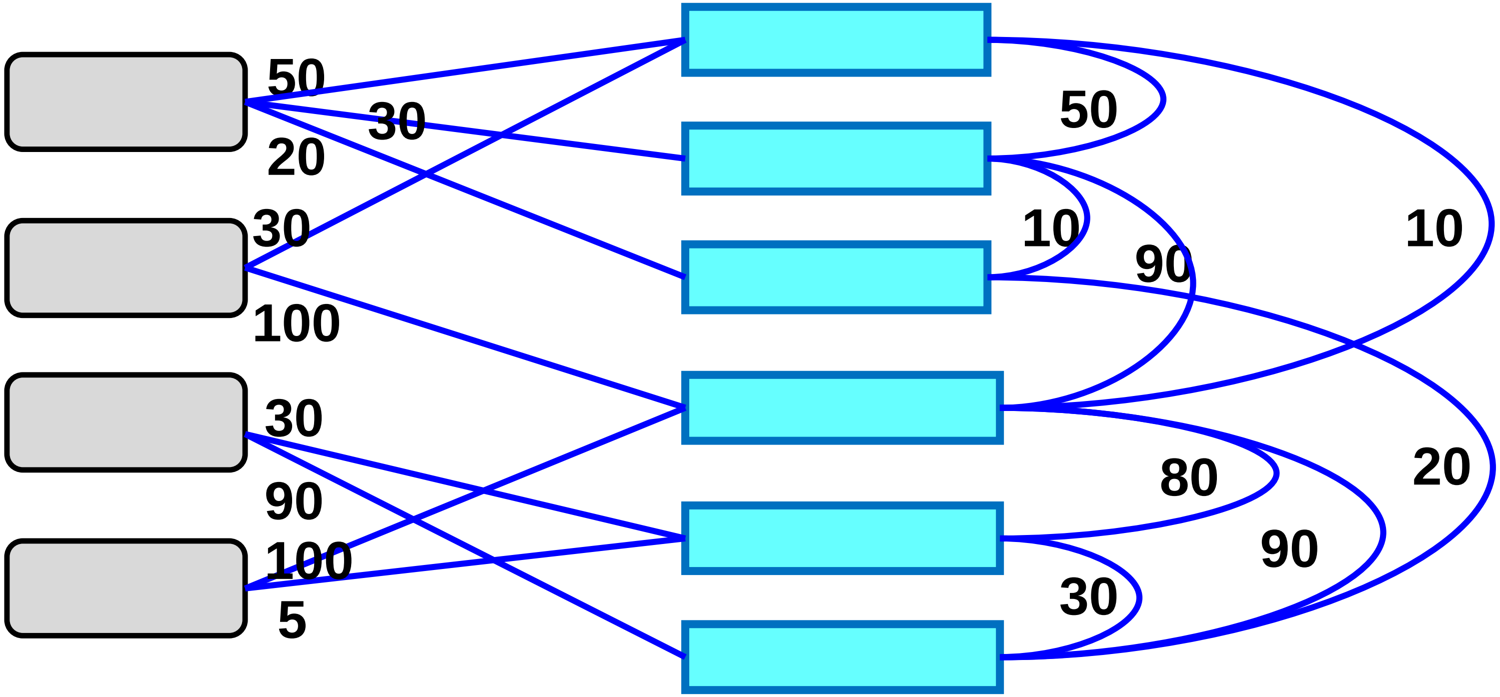
Named Entity Recognition & Disambiguation



KB provides building blocks:
- name-entity dictionary,
- relationships, types,
- text descriptions, keyphrases,
- statistics for weights

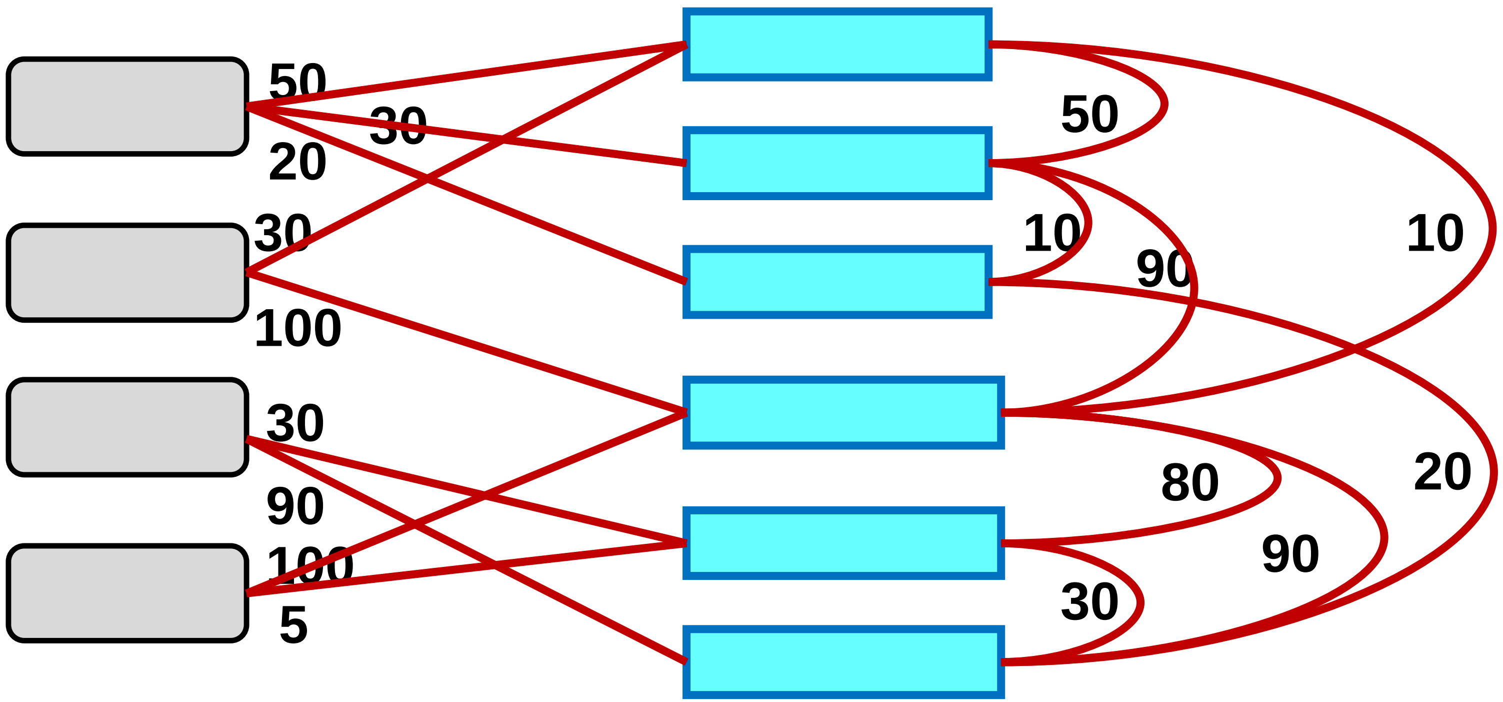
NED algorithms compute mention-to-entity mapping over weighted graph of candidates by popularity & similarity & coherence

Joint Mapping of Mentions to Entities



- **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)**

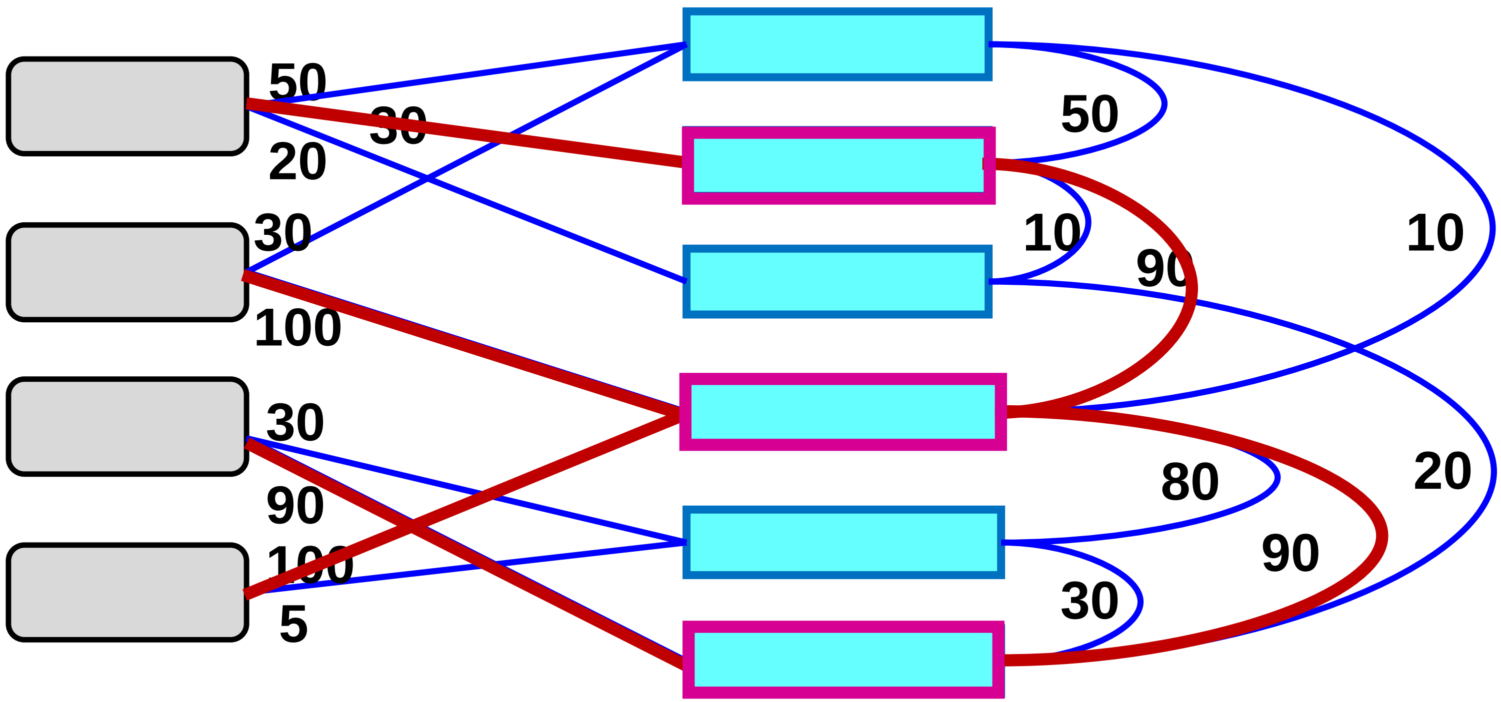
Joint Mapping: Prob. Factor Graph



Collective Learning with Probabilistic Factor Graphs
[Chakrabarti et al.: KDD'09]:

- **model P[m|e] by similarity and P[e1|e2] by coherence**
- **consider likelihood of P[m1 … mk | e1 … ek]**
- **factorize by all m-e pairs and e1-e2 pairs**
- **use MCMC, hill-climbing, LP etc. for solution**

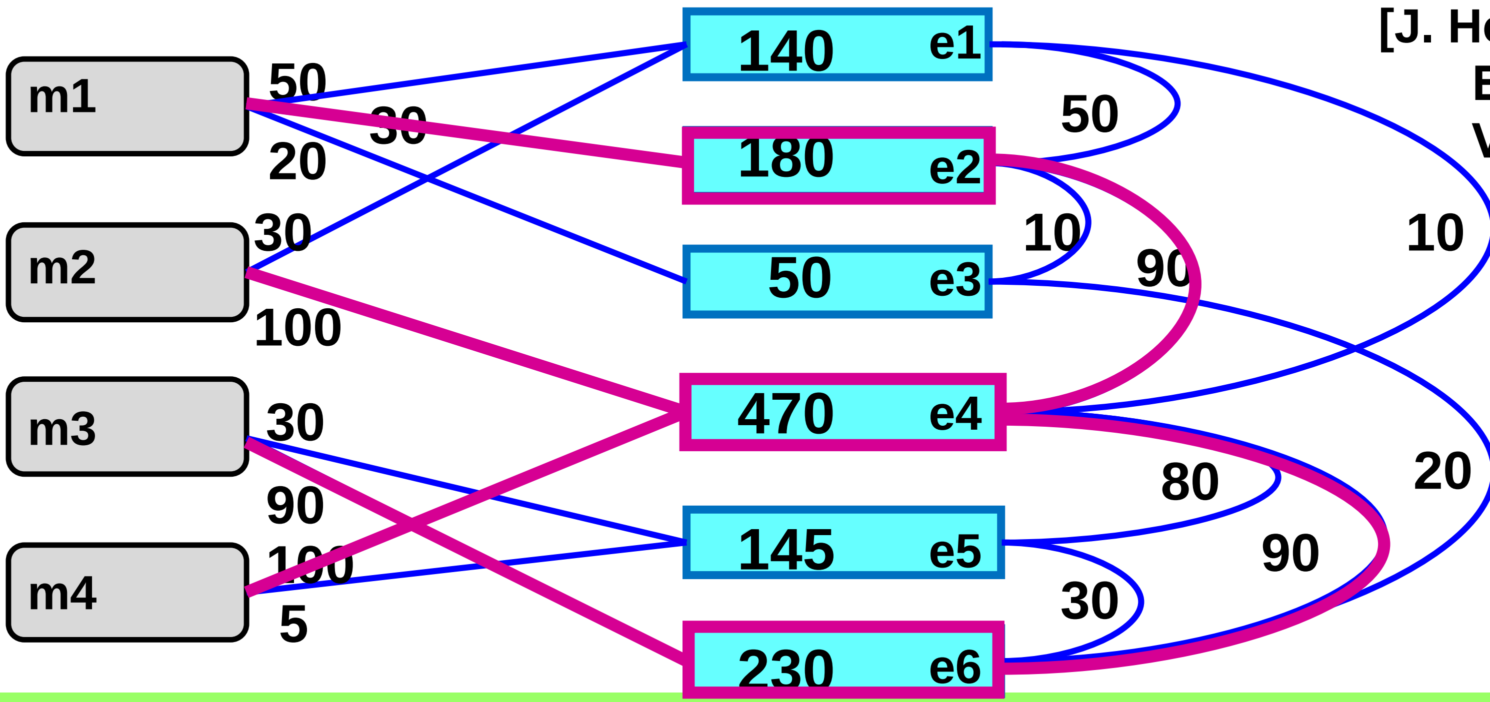
Joint Mapping: Dense Subgraph



- Compute **dense subgraph** such that:
  each m is **connected to exactly one** e (or **at most one** e)
- NP-hard → approximation algorithms
- Alt.: feature engineering for similarity-only method

  [Bunescu/Pasca 2006, Cucerzan 2007,
  Milne/Witten 2008, Ferragina et al. 2010 …]

Coherence Graph Algorithm



[J. Hoffart et al. EMNLP'11, VLDB'12]

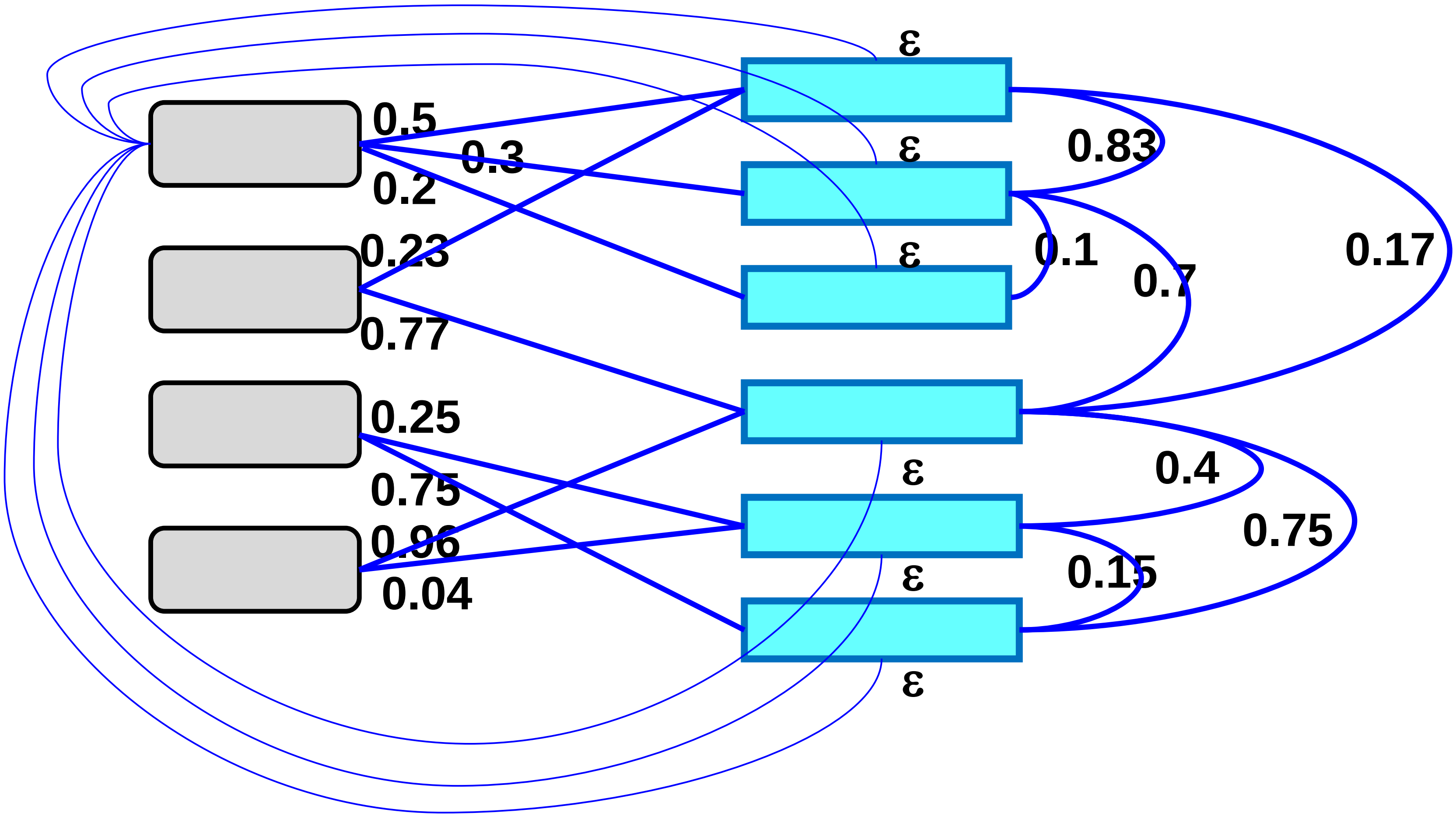
- Compute dense subgraph to
  maximize min weighted degree among entity nodes

  such that:

  each m is connected to exactly one e (or at most one e)
- Approx. algorithms (greedy, randomized, …), hash sketches, …
- 82% precision on CoNLL'03 benchmark
- Open-source software & online service AIDA

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

Random Walks Algorithm



- for each mention run random walks with restart (like Personalized PageRank with jumps to start mention(s))
- rank candidate entities by stationary visiting probability
- very efficient, decent accuracy

Long-Tail and Emerging Entities

Cave composed haunting songs like Hallelujah, O Children, and the Weeping Song.

- wikipedia/Good_Luck_Cave
- wikipedia/Nick_Cave
- wikipedia/Hallelujah_Chorus
- wikipedia/Hallelujah_(L_Cohen)
- last.fm/Nick_Cave/Hallelujah
- wikipedia/Children_(2011 film)
- last.fm/Nick_Cave/O_Children
- wikipedia/Weeping_(song)
- last.fm/Nick_Cave/Weeping_Song

[J. Hoffart et al.: CIKM'12]

Long-Tail and Emerging Entities

Cave composed haunting songs like Hallelujah, O Children, and the Weeping Song.

- wikipedia/Good_Luck_Cave
  - Gunung Mulu National Park
  - Sarawak Chamber
  - largest underground chamber
- wikipedia/Nick_Cave
  - Bad Seeds
  - No More Shall We Part
  - Murder Songs
- wikipedia/Hallelujah_Chorus
  - Messiah oratorio
  - George Frideric Handel
- wikipedia/Hallelujah_(L_Cohen)
  - Leonard Cohen
  - Rufus Wainwright
  - Shrek and Fiona
- last.fm/Nick_Cave/Hallelujah
  - eerie violin
  - Bad Seeds
  - No More Shall We Part
- wikipedia/Children_(2011 film)
  - South Korean film
- last.fm/Nick_Cave/O_Children
  - Nick Cave & Bad Seeds
  - Harry Potter 7 movie
  - haunting choir

em

duet

KO (p,q) = $\frac{\sum_t min(weight(t\ in\ p), weight(t\ in\ q))}{\sum_t max(weight(t\ in\ p), weight(t\ in\ q))}$

KORE (e,f) ~ $\sum_{p \in e, q \in f} KO(p,q)^2 \times min(weight(p\ in\ e), weight(q\ in\ f))$

implementation uses min-hash and LSH

[J. Hoffart et al.: CIKM‘12]

Long-Tail and Emerging Entities

Cave's brand-new album contains masterpieces like Water's Edge and Mermaids.

wikipedia.org/Good_Luck_Cave

Gunung Mulu National Park
Sarawak Chamber
largest underground chamber

wikipedia.org/Nick_Cave

Bad Seeds
No More Shall We Part
Murder Songs

…/Water's Edge Restaurant

excellent seafood
clam chowder
Maine lobster

…/Water's Edge (2003 film)

Nathan Fillion
horrible acting

any OTHER „Water's Edge"

all phrases minus keyphrases of known candidate entities

…/Mermaid's Song

Pirates of the Caribbean 4
My Jolly Sailor Bold
Johnny Depp

…/The Little Mermaid

Walt Disney
Hans Chrisitan Andersen
Kiss the Girl

„Mermaids"

all phrases minus keyphrases of known candidate entities

KP(new) = KP(name) – $\cup_e$ KP(e)
with statistical weights
[J. Hoffart et al.: WWW'14]

NERD Online Tools

J. Hoffart et al.: EMNLP 2011, VLDB 2011
http://mpi-inf.mpg.de/yago-naga/aida/
P. Ferragina, U. Scaella: CIKM 2010
http://tagme.di.unipi.it/

R. Isele, C. Bizer: VLDB 2012
http://spotlight.dbpedia.org/demo/index.html

Reuters Open Calais: http://viewer.opencalais.com/

Alchemy API: http://www.alchemyapi.com/api/demo.html

S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti: KDD 2009
http://www.cse.iitb.ac.in/soumen/doc/CSAW/

D. Milne, I. Witten: CIKM 2008
http://wikipedia-miner.cms.waikato.ac.nz/demos/annotate/

L. Ratinov, D. Roth, D. Downey, M. Anderson: ACL 2011
http://cogcomp.cs.illinois.edu/page/demo_view/Wikifier
D. Ceccarelli, C. Lucchese,S. Orlando, R. Perego, S. Trani. CIKM 2013
http://dexter.isti.cnr.it/demo/
A. Moro, A. Raganato, R. Navigli. TACL 2014
http://babelfy.org

some use Stanford NER tagger for detecting mentions
http://nlp.stanford.edu/software/CRF-NER.shtml

NERD at Work

https://gate.d5.mpi-inf.mpg.de/webaida/

Hurricane, a protest song about Carter, is on Bob's Desire.
Scarlet plays the violin on this piece. In the movie, Washington plays the boxer.

Disambiguate

Input Type:TEXT **Overall runtime:**33 sec(s)

Hurricane [Hurricane (Bob Dylan song)], a protest song about **Carter** [Rubin Carter], is on **Bob** [Bob Dylan]'s **Desire** [Desire (Bob Dylan album)]. **Scarlet** [Scarlet Rivera] plays the violin on this piece. In the movie, **Washington** [Denzel Washington] plays the boxer.

select knowledge

Run Information | Graph | Removal Steps

- 0: Hurricane
- 32: Carter
- 46: Bob
- 52: Desire
- 62: Scarlet
- 116: Washington

NERD at Work

https://gate.d5.mpi-inf.mpg.de/webaida/

Disambiguation Method:

prior | prior+sim | prior+sim+coherence

Parameters

Prior-Similarity-Coherence balancing ratio:

prior VS. sim. balance = 0.13 **(prior+sim.) VS. coh.** balance 0.71

Ambiguity degree 10

Coherence robustness test threshold: 0.9

Entities Type Filters:

Mention Extraction:

Stanford NER | Manual

You can manually tag the mentions by putting them between [[and]]. HTML Tables are automatcially disambiguated in the manual mode.

Fast Mode:

Enabled

Hurricane, a protest song about Carter, is on Bob's Desire.
Scarlet plays the violin on this piece. In the movie, Washington plays the boxer.

Disambiguate

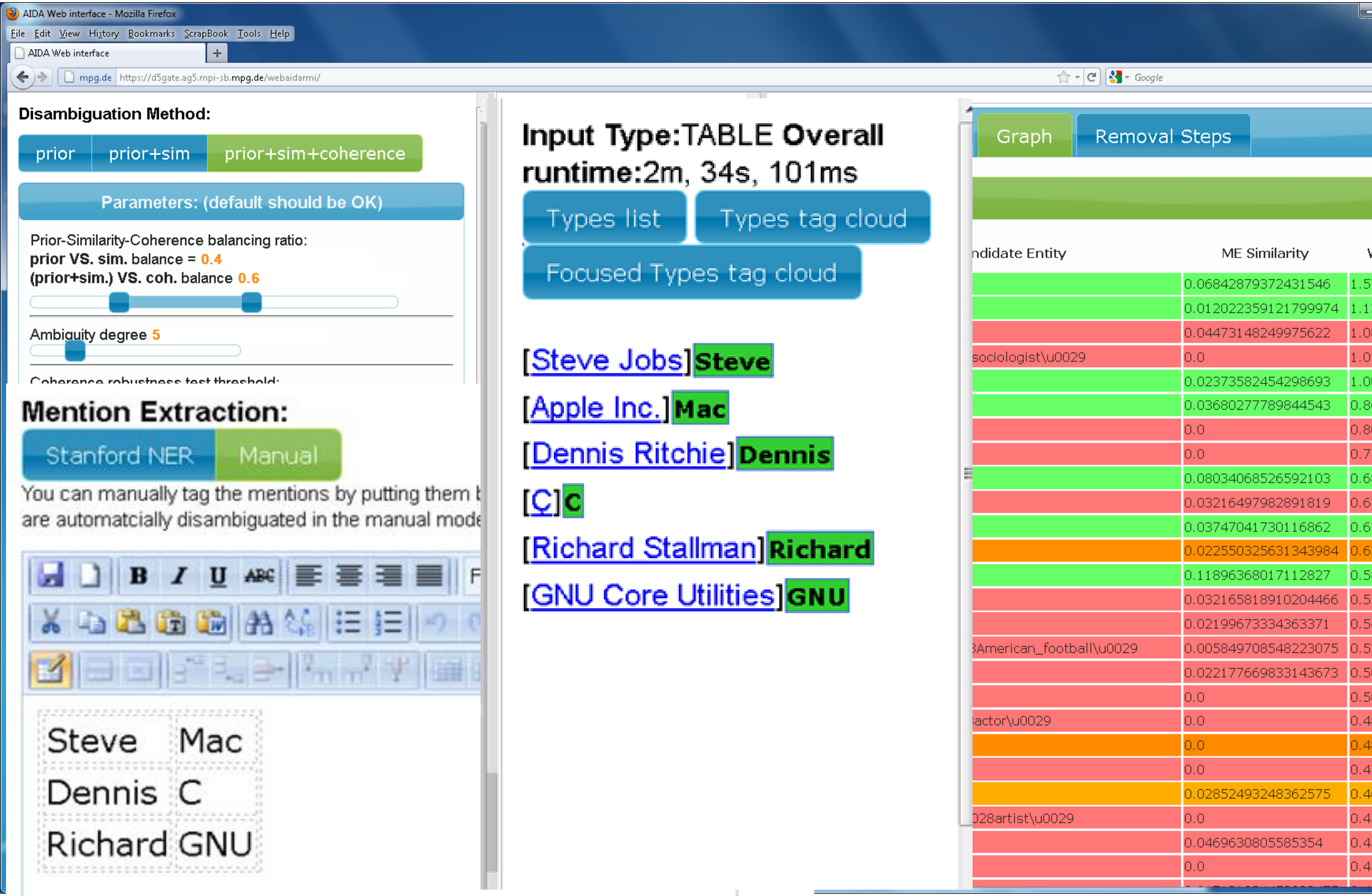
Input Type:TEXT **Overall runtime:**33 sec(s)

Hurricane [Hurricane (Bob Dylan song)], a protest song about **Carter** [Rubin Carter], is on **Bob** [Bob Dylan]'s **Desire** [Desire (Bob Dylan album)]. **Scarlet** [Scarlet Rivera] plays the violin on this piece. In the movie, **Washington** [Denzel Washington] plays the boxer.

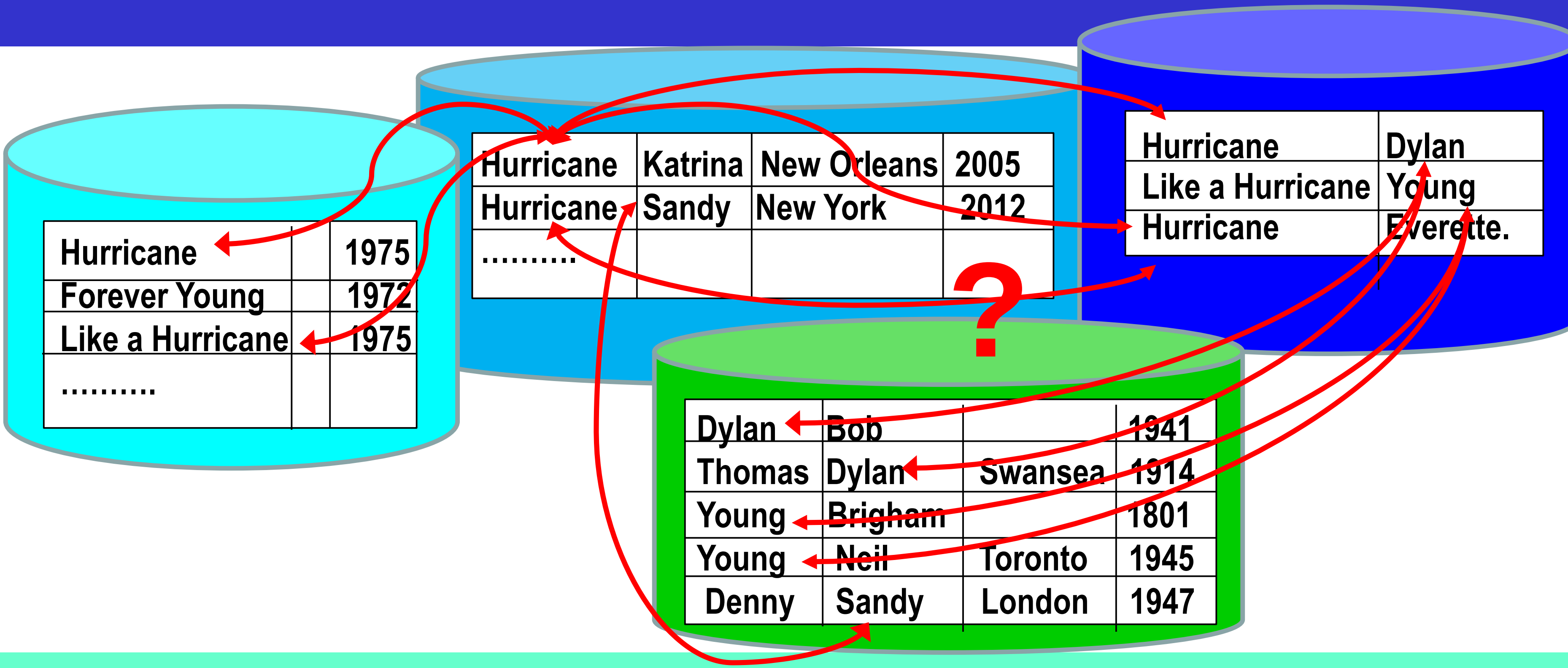
32: Carter

	Candidate Entity	ME Similarity	Weighted Degree	Weighted Degree when removed/final	Connected Entities
Info	Rubin Carter	0.007440300887298156	0.3672384453830128	0.017696436920930227	199 Show
Info	Joe Carter	0.0	0.3281050927116556	0.3281050927116556	188 Show
Info	Jimmy Carter	0.01103638778377025	0.3025790075617965	0.013351114882815256	320 Show
Info	Gary Carter	0.0021657937926300736	0.27194405292066054	0.27194405292066054	159 Show
Info	Paul Carter (baseball)	0.0	0.19680276201878621	0.19680276201878621	87 Show
Info	Vince Carter	4.1435682855787666E-4	0.1281591894396449	0.1281591894396449	88 Show
Info	Jay-Z	0.00730218654460134	0.12814442111832083	0.011735882716700024	137 Show
Info	Carter Elliott	0.0	0.1118463610679272	0.1118463610679272	47 Show
Info	Lance Carter	0.0	0.110008842052524	0.110008842052524	55 Show
Info	Steve Carter (baseball)	0.0	0.1005279520503617	0.1005279520503617	46 Show
Info	Chris Carter (right-handed hitter)	0.0	0.09913125899246221	0.09913125899246221	50 Show
Info	Arnold Carter	0.0	0.09623832488634608	0.09623832488634608	42 Show
Info	Howie Carter	0.0	0.09575478704689618	0.09575478704689618	40 Show
Info	Chris Carter (left-handed hitter)	3.774760610665208E-4	0.09537978696432067	0.09537978696432067	45 Show
Info	Nick Carter (baseball)	0.0	0.09167177180852937	0.09167177180852937	39 Show
Info	Sol Carter	0.0	0.09135182831121434	0.09135182831121434	38 Show
Info	Helena Bonham Carter	8.590379156735183E-4	0.09124507304617609	0.09124507304617609	68 Show
Info	Benny Carter	0.001310040883999477	0.09089849194529637	0.09089849194529637	67 Show
Info	Jeff Carter (pitcher)	0.0	0.09074559389855853	0.09074559389855853	40 Show
Info	Anthony Carter (American football)	4.080916063142848E-4	0.08487224122114082	0.08487224122114082	50 Show
Info	Ron Carter	0.006379385398268004	0.08444139387442567	0.010422108122627302	67 Show

NERD on Tables



Entity Matching in Structured Data



entity linkage:

- **structured counterpart to text NERD**
- **key to data integration**
- **essence of Big Data variety&veracity**
- **long-standing problem, very difficult, unsolved**

H.L. Dunn: Record Linkage. *American Journal of Public Health* 36 (12), 1946
H.B. Newcombe et al.: Automatic Linkage of Vital Records. *Science* 130 (3381), 1959

Goal: Linking Big Data & Deep Text

Musician	Song	Year	Listeners	Charts	. . .
Sinatra	My Way	1969	435 420		
Sex Pistols	My Way	1978	87 729		
Pavarotti	My Way	1993	4 239		
C. Leitte	Famo$a	2011	272 468		
B. Mars	Billionaire	2010	218 116		
. . . . .	. . . . .				

Lessons Learned

NERD lifts text to entity-centric representation: nearly as valuable and usable as a structured DB

KG is key asset for high-quality NERD:
entity descriptions, semantic relations, keyphrases, statistics

Best NERD methods can capture also long-tail and emerging entities

What's Next

High-throughput NERD: semantic indexing

Low-latency NERD: speed-reading

popular vs. long-tail entities, general vs.specific domain

Leverage deep-parsing features & semantic typing

example: ***Page played Kashmir on his Gibson***

subj obj mod

Short and difficult texts:

tweets, headlines, etc.

fictional texts: novels, song lyrics, TV sitcoms, etc.

Handle newly emerging entities in KG life-cycle

General WSD for classes, relations, general concepts

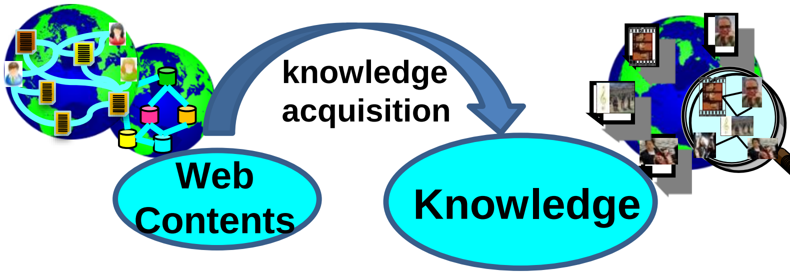
for Web tables, lists, questions, dialogs, summarization, …

Outline

✓ Introduction

✓ KG Construction

✓ Refined Knowledge

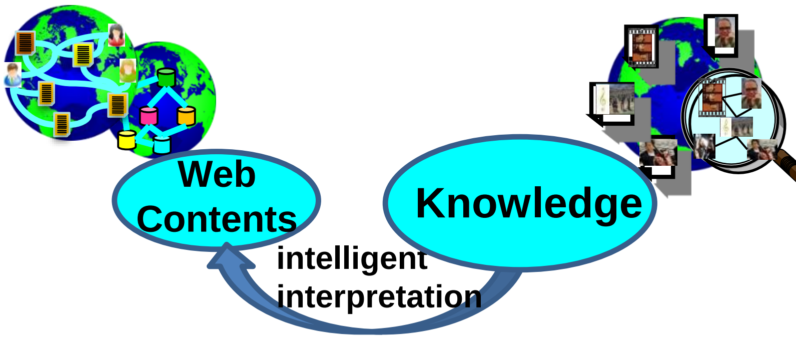


✓ Knowledge for Language

★ **Deep Text Analytics**

★ Search for Knowledge

★ Conclusion



Goal: Deep Data & Text Analytics

Entertainment: Who covered which other singer?
Who influenced which other musicians?

Health: Drugs (combinations) and their side effects

Politics: Politicians‘ positions on controversial topics

Finance: Risk assessment based on data, reports, news

Business: Customer opinions on products in social media

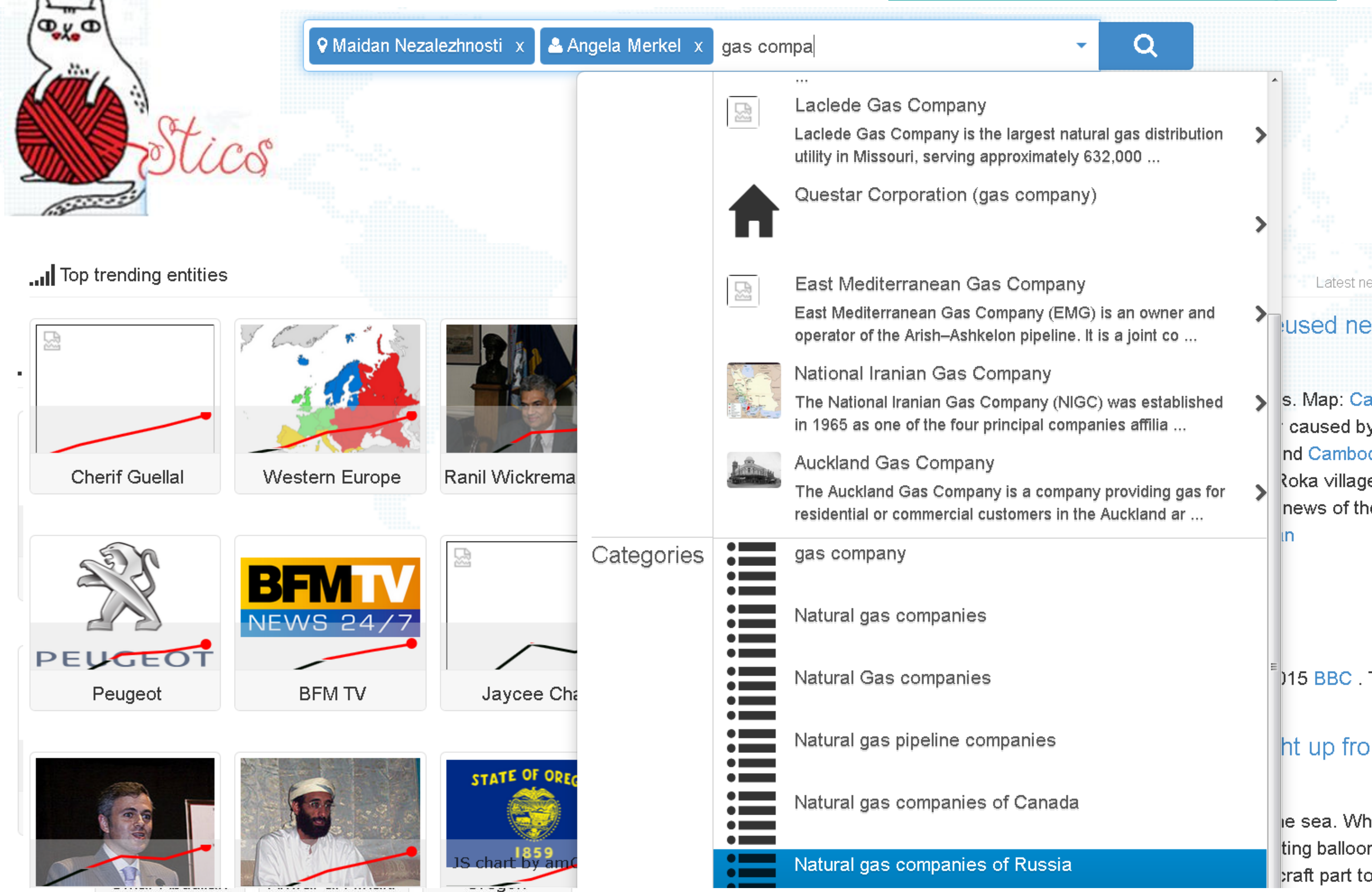
Culturomics: Trends in society, cultural factors, etc.

General Design Pattern:

- Identify relevant contents sources
- Identify entities of interest & their relationships
- Position in time & space
- Group and aggregate
- Find insightful patterns & predict trends

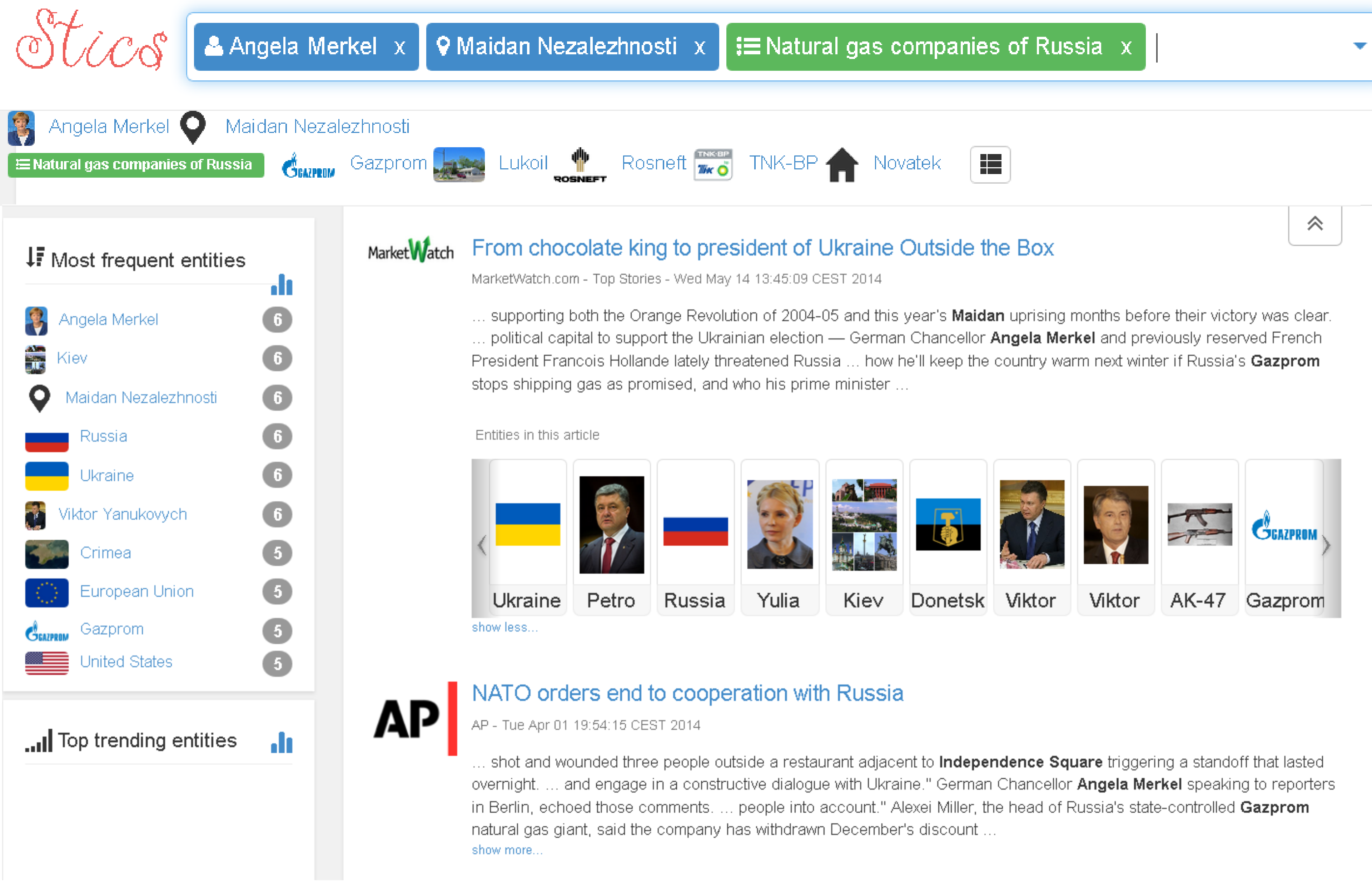
Deep Text Search & Analytics

https://stics.mpi-inf.mpg.de



Deep Text Search & Analytics

https://stics.mpi-inf.mpg.de



Deep Text Search & Analytics

https://stics.mpi-inf.mpg.de

Stics

Angela Merkel x | Maidan Nezalezhnosti x | Natural gas companies of Russia x

Angela Merkel Maidan Nezalezhnosti

Natural gas companies of Russia Gazprom Lukoil Rosneft TNK-BP Novatek

From chocolate king to president of Ukraine Outside the Box

MarketWatch.com - Top Stories - Wed May 14 13:45:09 CEST 2014

… supporting both the Orange Revolution of 2004-05 and this year's **Maidan** uprising months before their victory was clear. … political capital to support the Ukrainian election — German Chancellor **Angela Merkel** and previously reserved French President Francois Hollande lately threatened Russia … how he'll keep the country warm next winter if Russia's **Gazprom** stops shipping gas as promised, and who his prime minister …

NATO orders end to cooperation with Russia

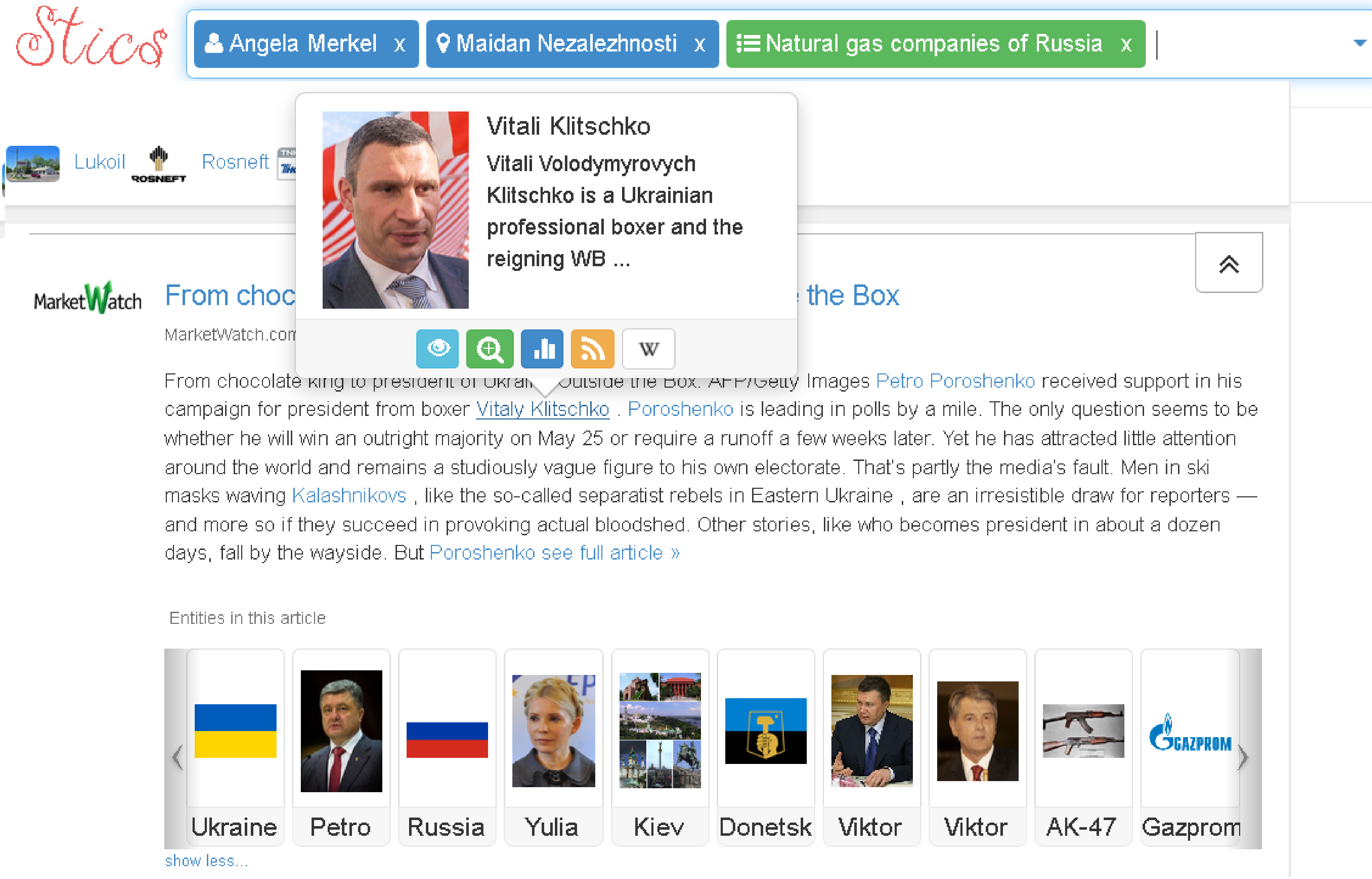
AP - Tue Apr 01 19:54:15 CEST 2014

… shot and wounded three people outside a restaurant adjacent to **Independence Square** triggering a standoff that lasted overnight. … and engage in a constructive dialogue with Ukraine." German Chancellor **Angela Merkel** speaking to reporters in Berlin, echoed those comments. … people into account." Alexei Miller, the head of Russia's state-controlled **Gazprom** natural gas giant, said the company has withdrawn December's discount …

show more…

Deep Text Search & Analytics

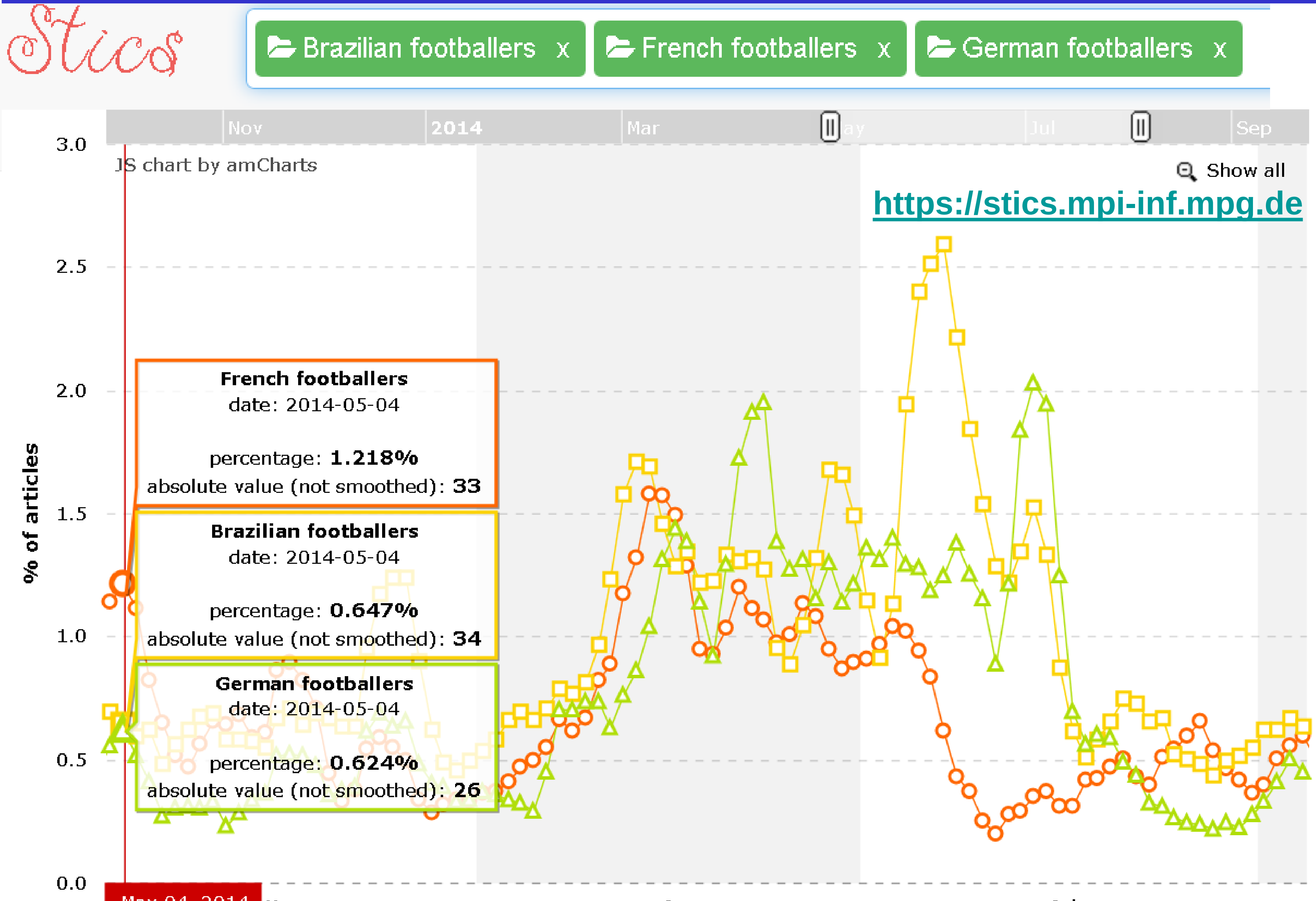
https://stics.mpi-inf.mpg.de



Deep Text Search & Analytics

https://stics.mpi-inf.mpg.de

Deep Text Search & Analytics



Machine Reading of Scholarly Papers

https://gate.d5.mpi-inf.mpg.de/knowlife/

[P. Ernst et al.: ICDE'14]

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Display Settings: Abstract

Indian J Endocrinol Metab. 2014 Jan;18(1):83-8. doi: 10.410

Are patients with primary hypothyroid observational study.

Mithal A[1], Dharmalingam M[2], Tewari N[3].

Author information

[1]Department of Endocrinology, Medanta - The Me
[2]Department of Endocrinology, Bangalore Endocri
[3]Medical Affairs, Metabolics and Endocrinology, A

Abstract

BACKGROUND: A large proportion of patients wo Stimulating Hormone (TSH) values. There is a pau receiving levothyroxine treatment.

AIM: To assess the percentage of primary hypoth least 2 m.

MATERIALS AND METHODS: A cross-sectional, with levothyroxine for at least 2m was undertaken the subjects and their quality of life was assessed the study subjects during the same visit for asses

RESULTS: A total of 1950 subjects (mean age 41 the study. Of the 1925 subjects in whom TSH valu were over-treated (TSH < 0.4mIU/L). The mean do were compliant/moderately compliant to thyroxine on serum TSH. Subjects with abnormal TSH had to physical health (P = 0.0763).

Back

KnowLife - One-Stop Health Portal

Documents	Entities	Text Annotation	U

- Entities
- Unhighlight all
- Procedures
- Physiology
- Geographic Areas
- Genes & Molecular Sequen...
- Activities & Behaviors
- Occupations
- Living Beings
- Chemicals & Drugs
- Disorders
- Anatomy
- Concepts & Ideas
- Facts

AIM:

To assess the percentage of primary hypothyroid patients with abnormal thyroid function despite being prescribed levothyroxine for at least 2 m.

MATERIALS AND METHODS:

A cross-sectional, single visit, observational study in adult patients with primary hypothyroidism on treatment with levothyroxine for at least 2m was undertaken across 10 cities in India. Compliance to thyroxine therapy was assessed by interviewing the subjects and their quality of life was assessed by administering the SF-36 questionnaire. TSH levels were correlated with the current dose of levothyroxine. A random blood sample (5ml) was drawn from the study subjects during the same visit for assessing serum TSH levels. A total of 1950 subjects (mean age 41.4 ? 11.17 years; female 81.2 %, male 18.8 %) with primary hypothyroidism were enrolled in the study.

RESULTS:

The mean dose of thyroxine in this study was 1.23 ?g / kg/day (? 0.85). Of the 1925 subjects in whom TSH values were available, 808 (41.97 %) were under-treated (TSH > 4 mIU/L) and 243 (12.62 %) were over-treated (TSH < 0.4 mIU/L). Age and autoimmune hypothyroidism were the factors that had significant impact on serum TSH. Majority of subjects (90.79 %) were compliant/moderately compliant to thyroxine therapy.

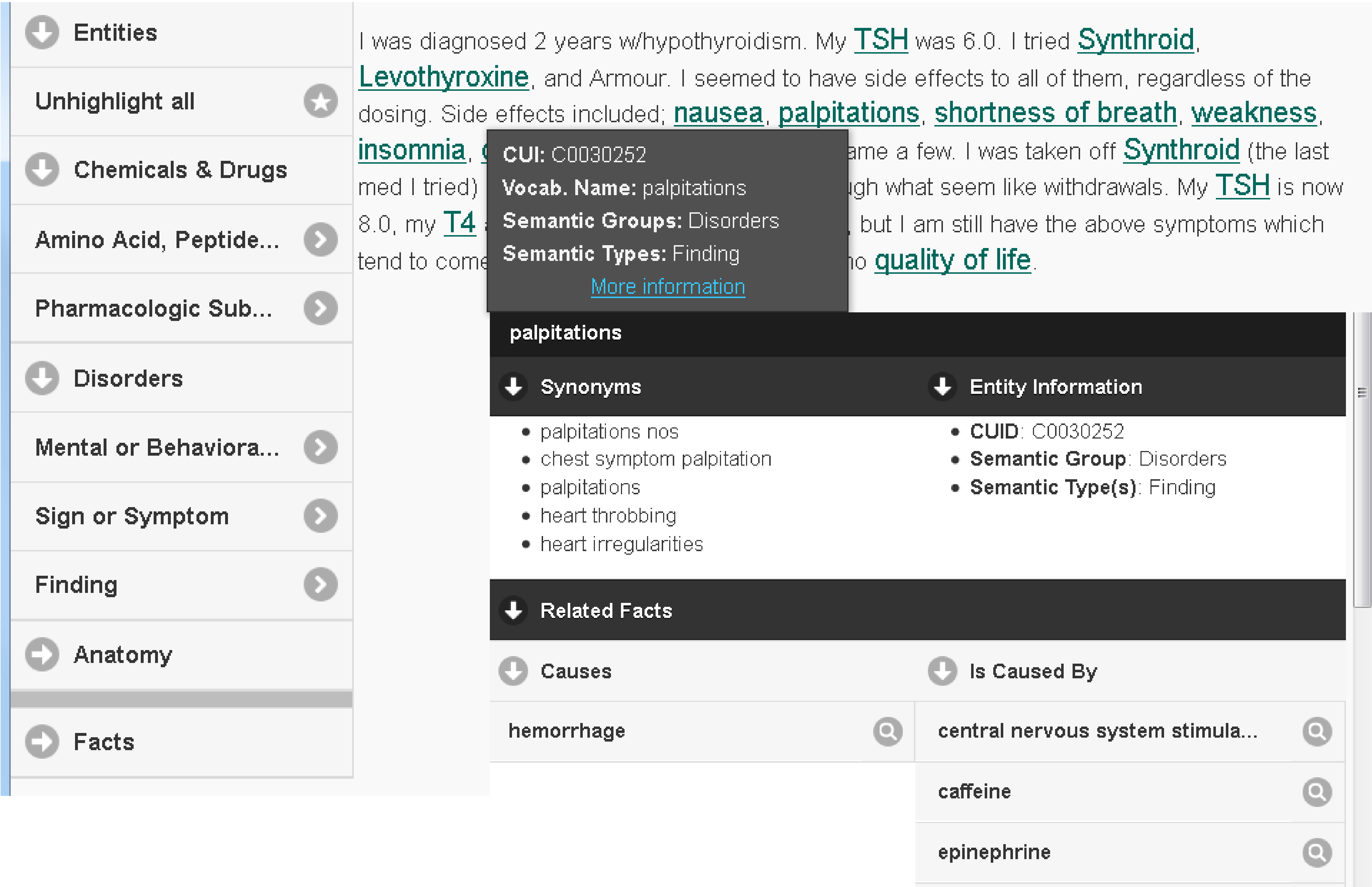
CONCLUSION:

Subjects with abnormal TSH had significantly lower scores for role limitation due to emotional problems (P = 0.0278) and due to physical health (P = 0.0763). The mean daily dose of thyroxine (*1.23 ?g / kg ? 0.85) was less than the recommended full* replacement dose. This study concluded that around half (54 %) of known hypothyroid subjects had out-of-range serum TSH despite being treated with levothyroxine for at least 2m.

Machine Reading of Health Forums

https://gate.d5.mpi-inf.mpg.de/knowlife/

[P. Ernst et al.: ICDE'14]



Deep Data & Text Analytics:
Side Effects of Drug Combinations

Structured Expert Data

Social Media

Deeper insight from both expert data & social media:

- *actual* side effects of drugs
- … and drug *combinations*
- risk factors and complications of (wide-spread) diseases
- alternative therapies
- aggregation & comparison by age, gender, life style, etc.

Showstoppers today:

- (In)Credibility of User Statements (Veracity)
- Diversity & Ambiguity of Relational Phrases (Variety)

http://dailymed.nlm.nih.gov http://www.patient.co.uk

Veracity: Where the Truth Lies

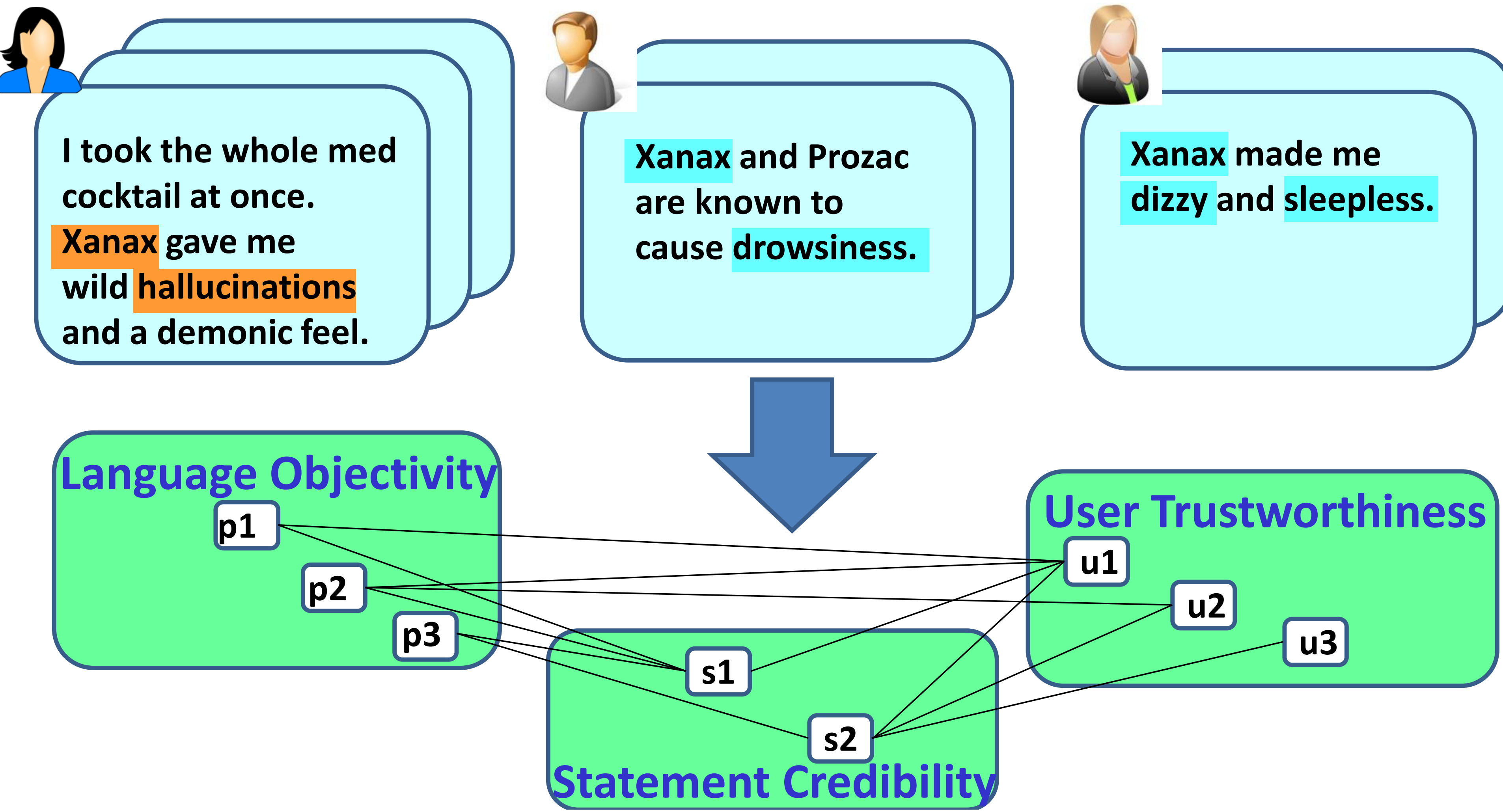
Assess credibility of statements / claims on the Internet and the trustworthiness of their sources

- **Search results:** love affairs of Hillary Clinton ?
- **Biased news:** Merkel hates Greece
- **KB contents:** Cerf & Berners-Lee invented the Internet
  Einstein invented rock'n roll
- **Social media:** Obamacare requires microchip implant
  Snowden works for Al Quaida
- **Health communities:** Xanax causes dizziness
  Xanax cures cancer

important for
KB curation, IE quality, ranking, explanation, trust in info & users

KB can help to find alternatives & to check consistency

Veracity in Health Communities



hasSideEffects (Xanax, hallucinations)

hasSideEffects (Xanax, dizziness)

hasSideEffects (Xanax, insomnia)

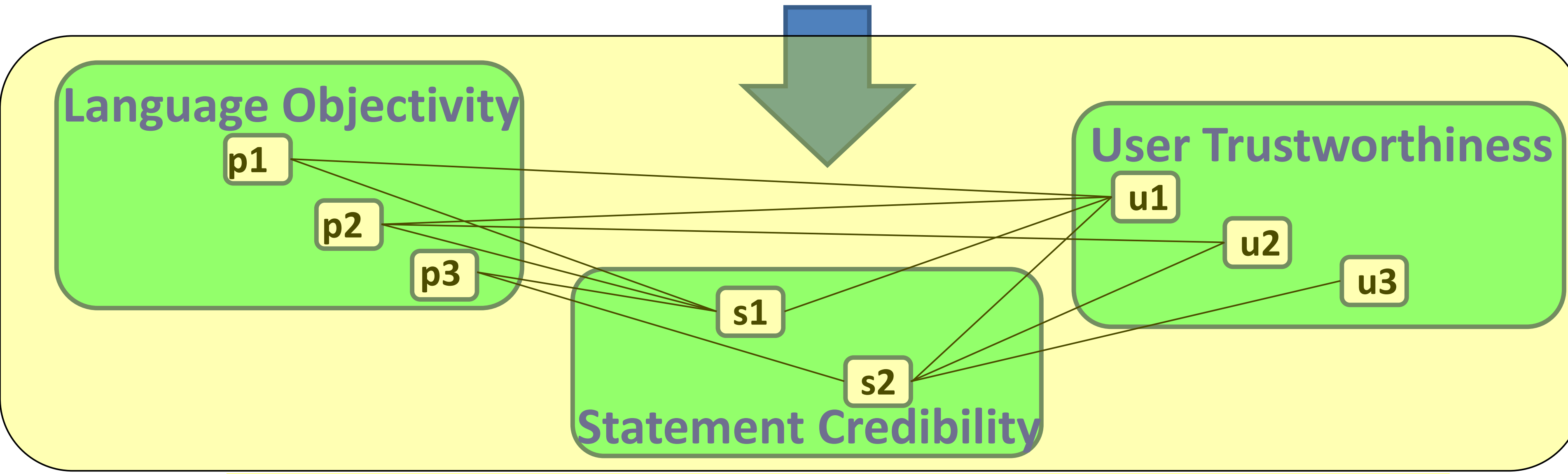
Veracity in Health Communities

[S. Mukherjee et al.: KDD'14]

I took the whole med cocktail at once. Xanax gave me wild hallucinations and a demonic feel.

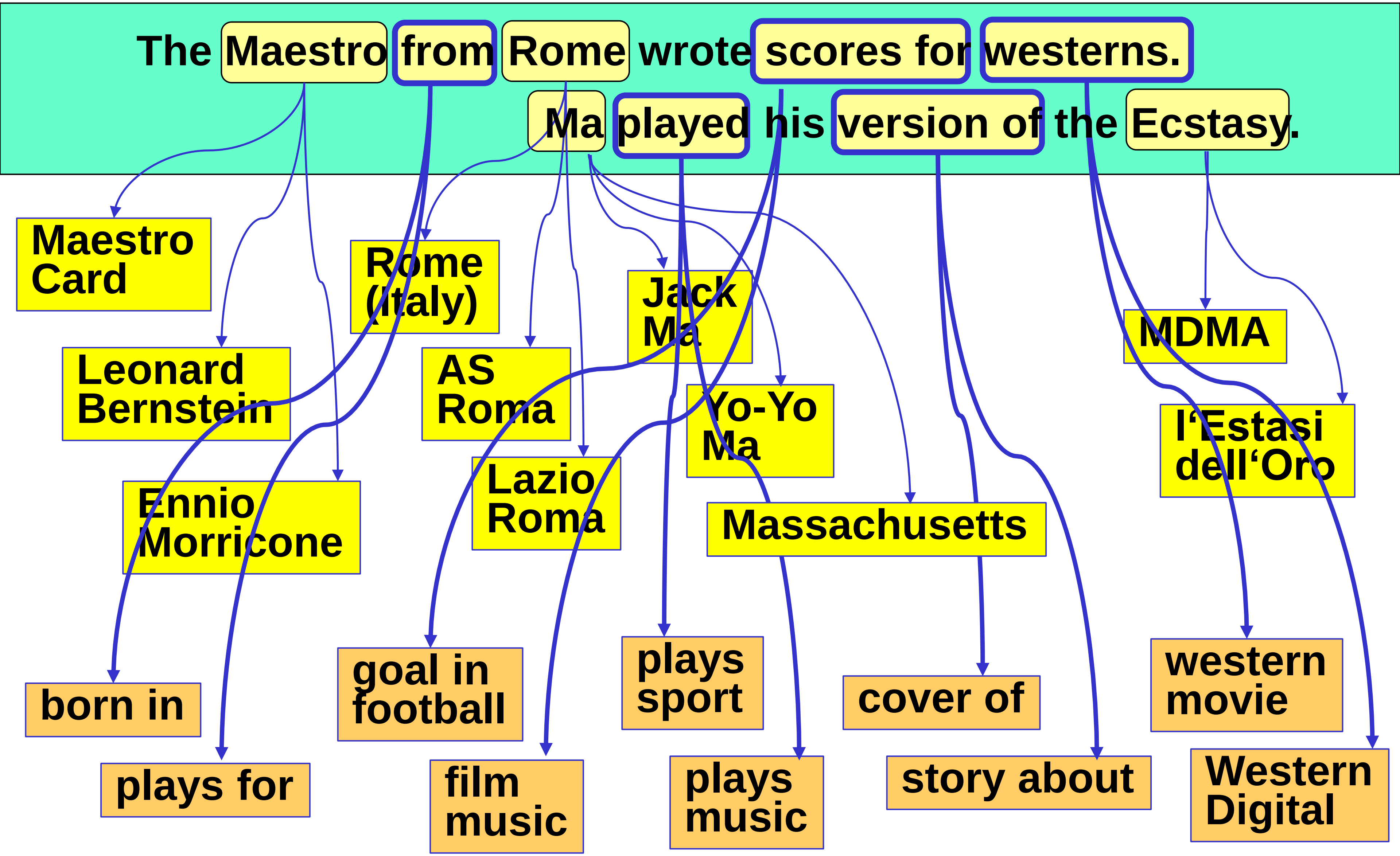
Xanax and Prozac are known to cause drowsiness.

Xanax made me dizzy and sleepless.



joint reasoning with probabilistic graphical model
(semi-supervised heterogeneous CRF with EM-style inference)

Variety: Phrases for Entities, Classes, Relations



Paraphrases of Relations

composed: musician × song **covered: musician × song**

Dylan wrote a sad song Knockin‘ on Heaven‘s Door, a cover song by the Dead
Morricone ‘s masterpiece is the Ecstasy of Gold, covered by Yo-Yo Ma
Amy‘s souly interpretation of Cupid, a classic piece of Sam Cooke
Nina Simone‘s singing of Don‘t Explain revived Holiday‘s old song
Cat Power‘s voice is haunting in her version of Don‘t Explain
Cale performed Hallelujah written by L. Cohen

SOL patterns over words, wildcards, POS tags, semantic types:

<musician> wrote ∗ ADJ piece <song>

Sequence Mining with Type Lifting
(N. Nakashole et al.: EMNLP'12, VLDB'12)

Relational phrases are typed:

<*singer*> covered <*song*>
<*book*> covered <*event*>

Relational synsets (and subsumptions):

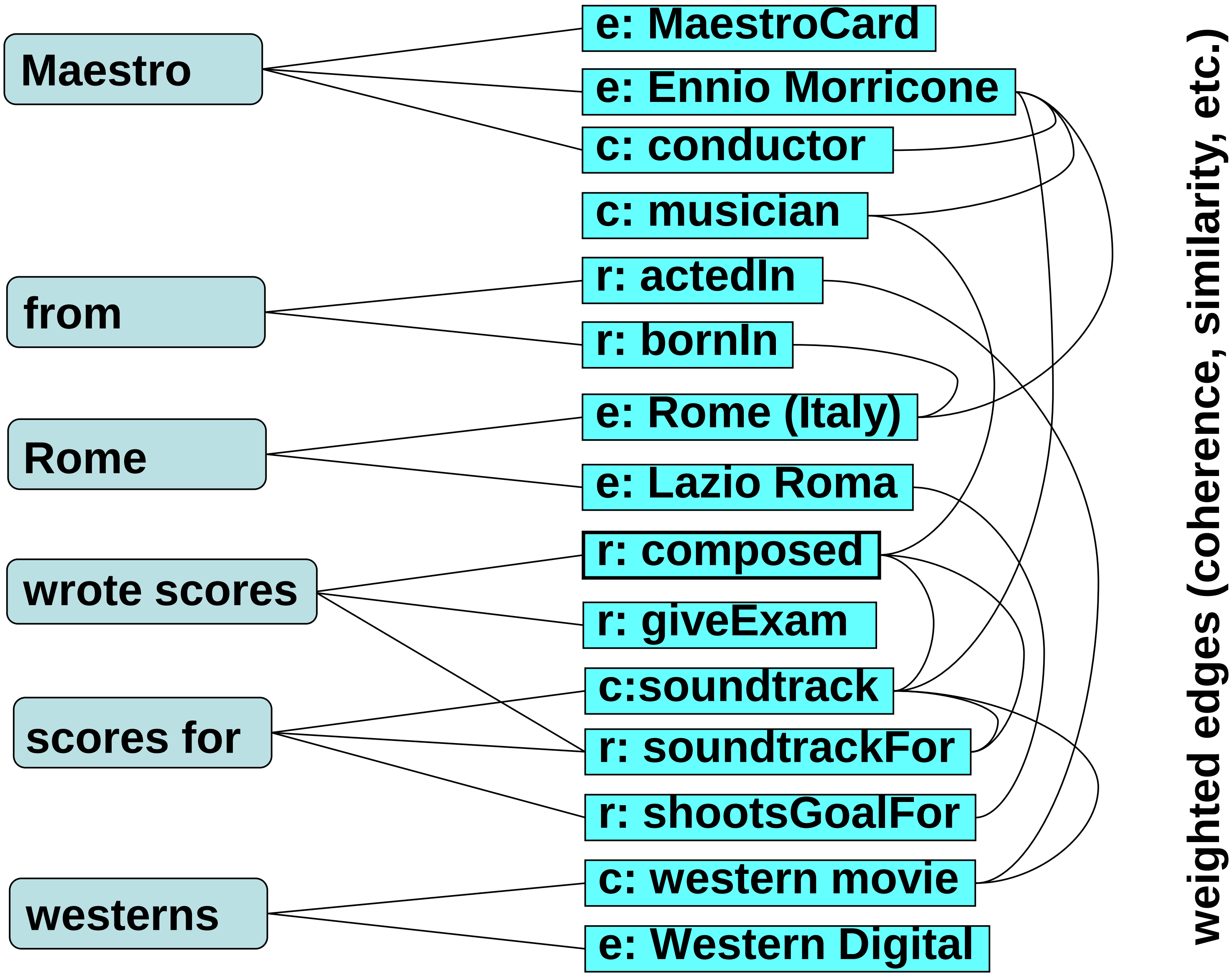
covered: cover song, interpretation of, singing of, voice in ∗ version, …

composed: wrote, classic piece of, ‘s old song, written by, composed, …

350 000 SOL patterns from Wikipedia: http://www.mpi-inf.mpg.de/yago-naga/patty/

Disambiguation for Entities, Classes & Relations

(M. Yahya et al.: EMNLP'12, CIKM'13; J. Berant et al.: EMNLP'13 …)



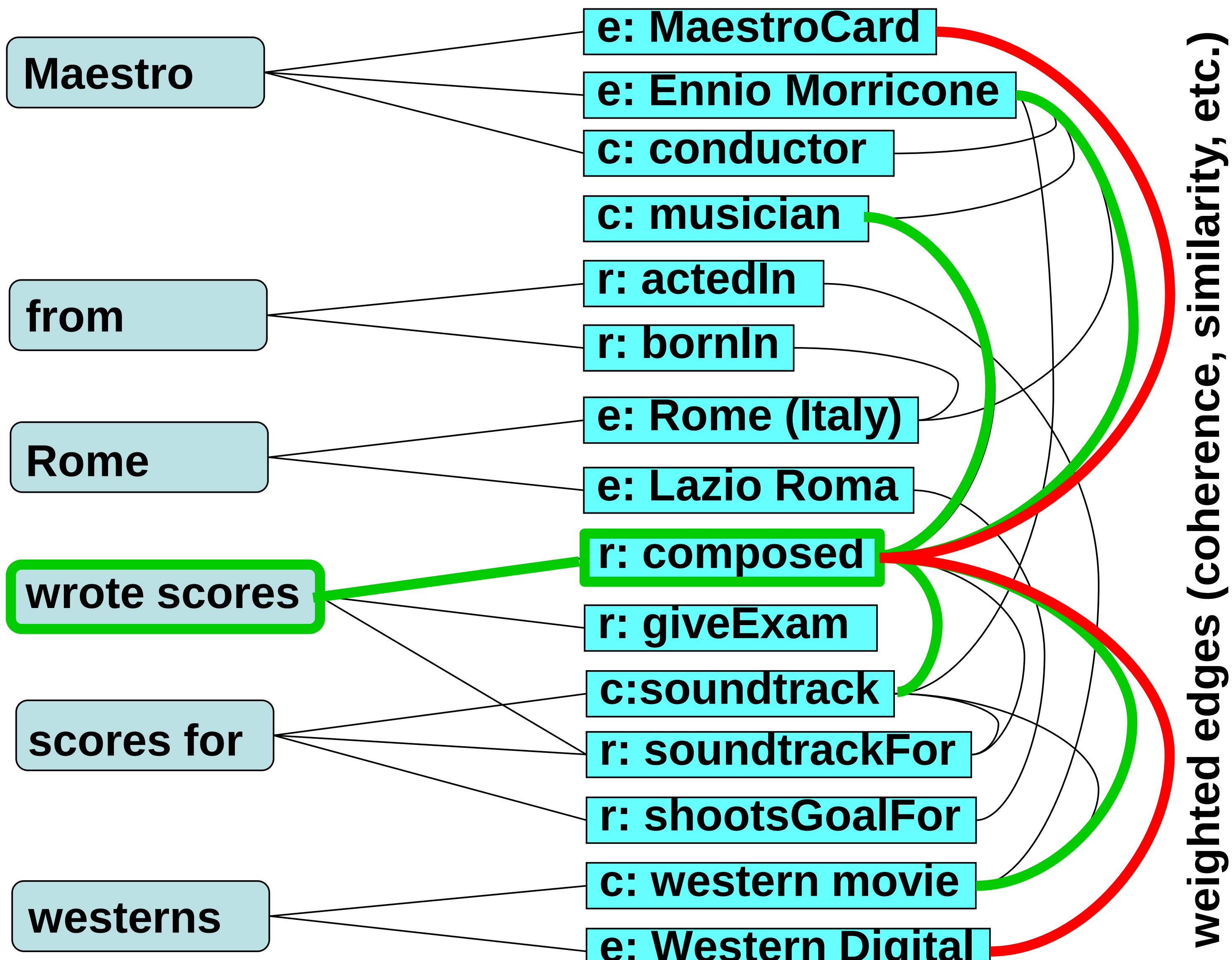
ILP optimizers like Gurobi solve this in seconds

Combinatorial Optimization by ILP (with type constraints etc.)

Disambiguation for Entities, Classes & Relations

(M. Yahya et al.: EMNLP'12, CIKM'13; J. Berant et al.: EMNLP'13 …)

ILP
optimizers
like Gurobi
solve this
in seconds



Combinatorial Optimization by ILP (with type constraints etc.)

Lessons Learned

If data is the new oil, then text is the new chocolate:
web, news, journals, social media

Entity view of text has enormous benefit:
lifts text (almost) on par with structured data
and opens up all kinds of analytics

Entity-Relationship view is ultimately needed,
but relational paraphrases are harder to deal with

Deep text analytics enables business apps and insight:
journalists, market&media analysts, company scouts, …
[Gartner etc: from $2Bio in 2014 to $20Bio by 2020]

What's Next

Opportunity: combine text & data for deeper insight

Huge research avenue:
analytics (OLAP, KDD, ML, …) for text & data

Challenge Veracity:
cope with highly varying trust and credibility

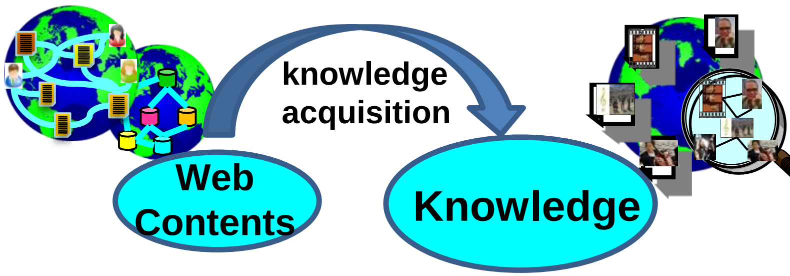
Challenge Variety:
cope with high diversity of names and phrases

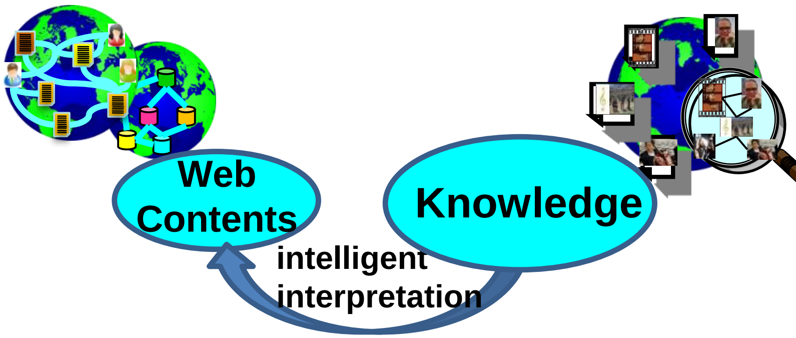
Plenty of algorithmic and scalability challenges

Outline

- ✓ Introduction
- ✓ KG Construction
- ✓ Refined Knowledge

- ✓ Knowledge for Language
- ✓ Deep Text Analytics
- ★ **Search for Knowledge**
- ★ Conclusion





Goal: Semantic Search with Entities, Classes, Relationships

properties of entity

- US president when Barack Obama was born?
  Nobel laureate who outlived two world wars and all his children?

instances of classes

- Politicians who are also scientists?
  European composers who won the Oscar?
  Chinese female astronauts?

relationships

- FIFA 2014 finalists who played in a Champions League final?
  German football clubs that won against Real Madrid?

multiple entities

- Commonalities & relationships among:
  John Lennon, Heath Ledger, King Kong?

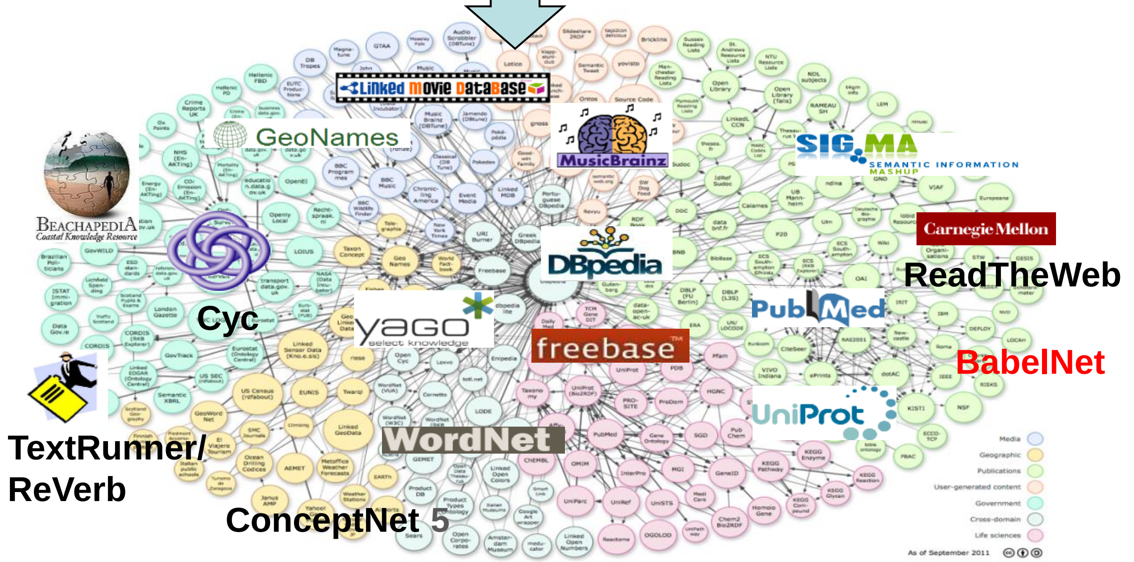
applications

- Enzymes that inhibit HIV?
  Antidepressants that interfere with blood-pressure drugs?
  German philosophers influenced by William of Ockham?

Querying the Web of Data & Knowledge

Who composed scores for westerns and is from Rome?

Casual ?	**(form-based)**
Textual ?	**(keywords)**
Visual ?	**(point&drag)**
Sparql ?	**(query language)**
Natural ?	**(natural language)**

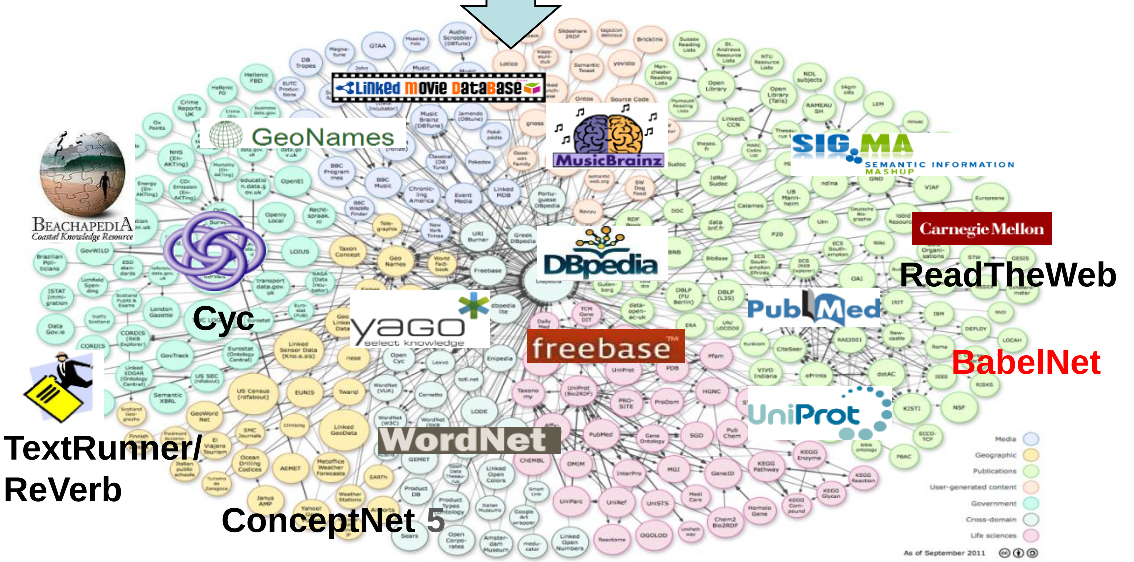


Linked Data
Big Data
Web tables

Querying the Web of Data: Casual?

Who composed scores for westerns and is from Rome?

S	P	O
Elisa	created	*Aida*
Aida	contributedTo	*Vendetta*
Vendetta	type	westernMovie
Elisa	bornIn	Rome

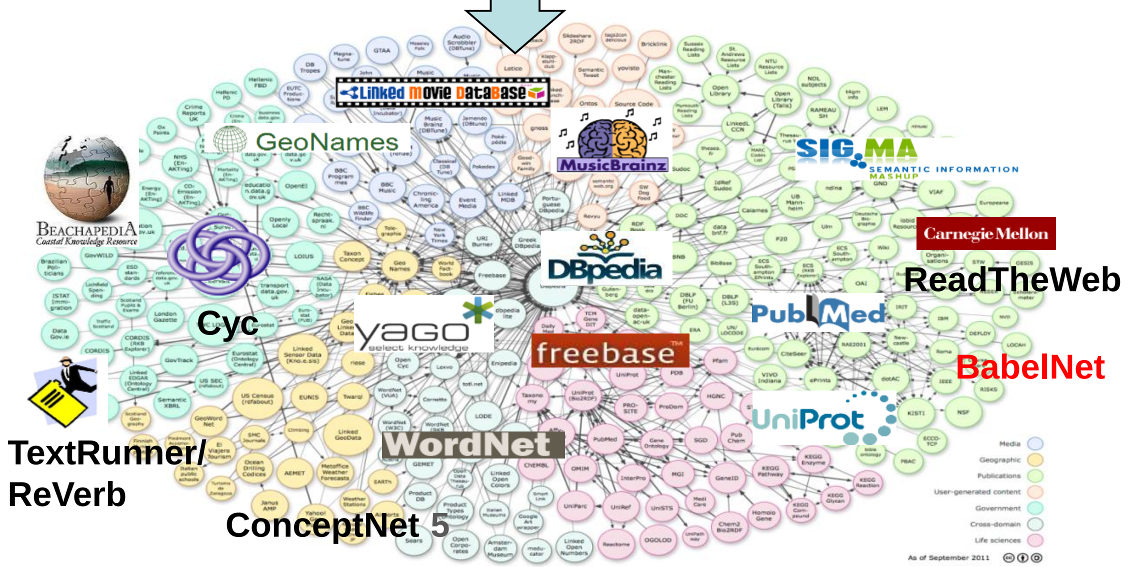


Linked Data
Big Data
Web tables

Querying the Web of Data: Sparql?

Who composed scores for westerns and is from Rome?

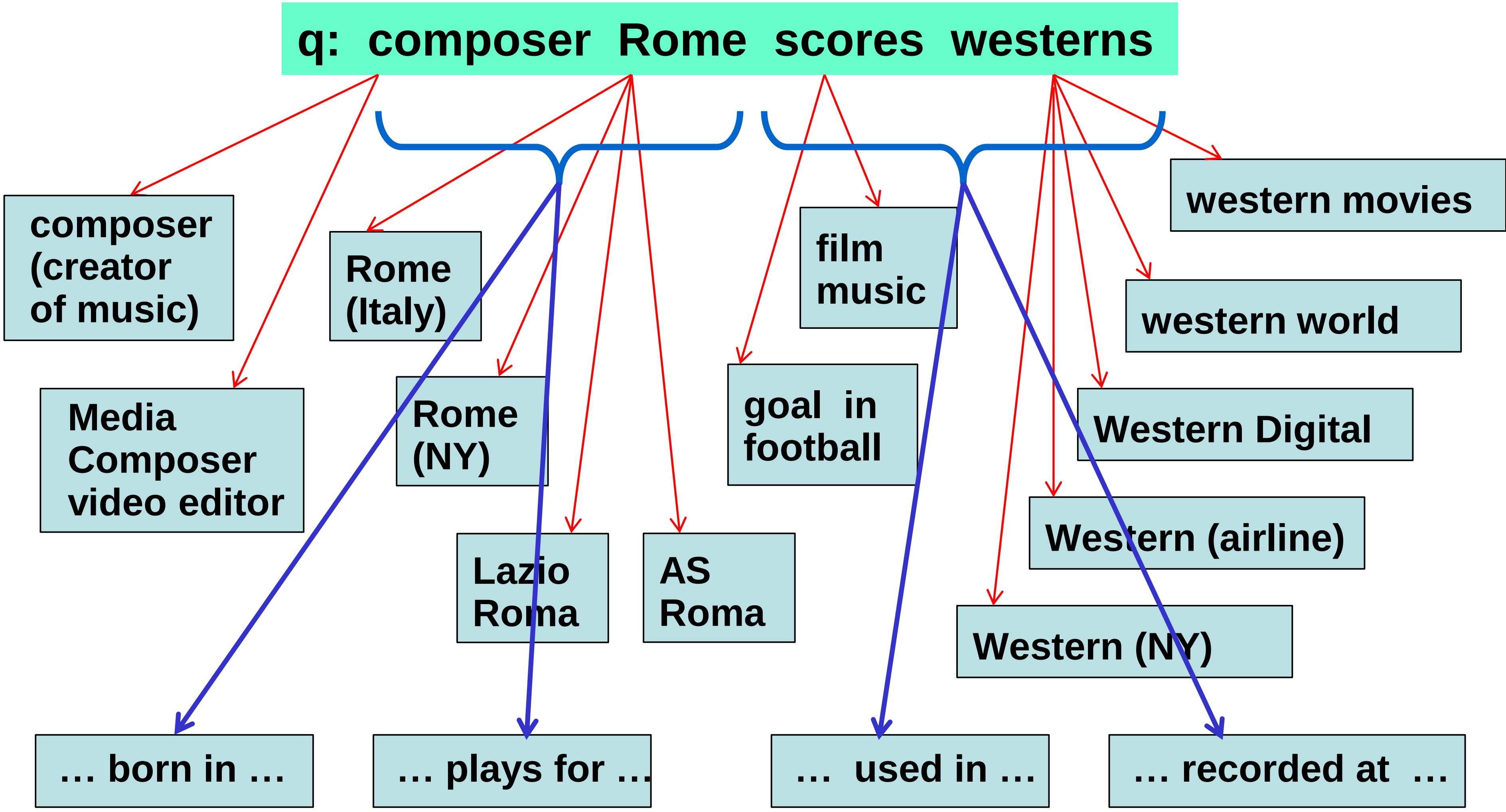
```
Select ?x Where {    ?x  created  ?s .
                     ?s  contributesTo  ?m .
                     ?m  type  westernMovie .
                     ?x  bornIn  Rome . }
```



Linked Data
Big Data
Web tables

Querying the Web of Data: Textual?

Who composed scores for westerns and is from Rome?



Querying the Web of Data: Natural! From Questions to Queries

- **dependency parsing to decompose question**
- **mapping of phrases onto entities, classes, relations**
- **generating SPO triploids (later triple patterns)**

Who composed scores for westerns and is from Rome?

Who composed scores **is from Rome**

scores for westerns

Semantic Parsing: from Triploids to SPO Triple Patterns

Map names into entities or classes, phrases into relations

Who composed scores ⇒ **?x created ?s**

?x type composer

?s type music

scores for westerns ⇒ **?s contributesTo ?y**

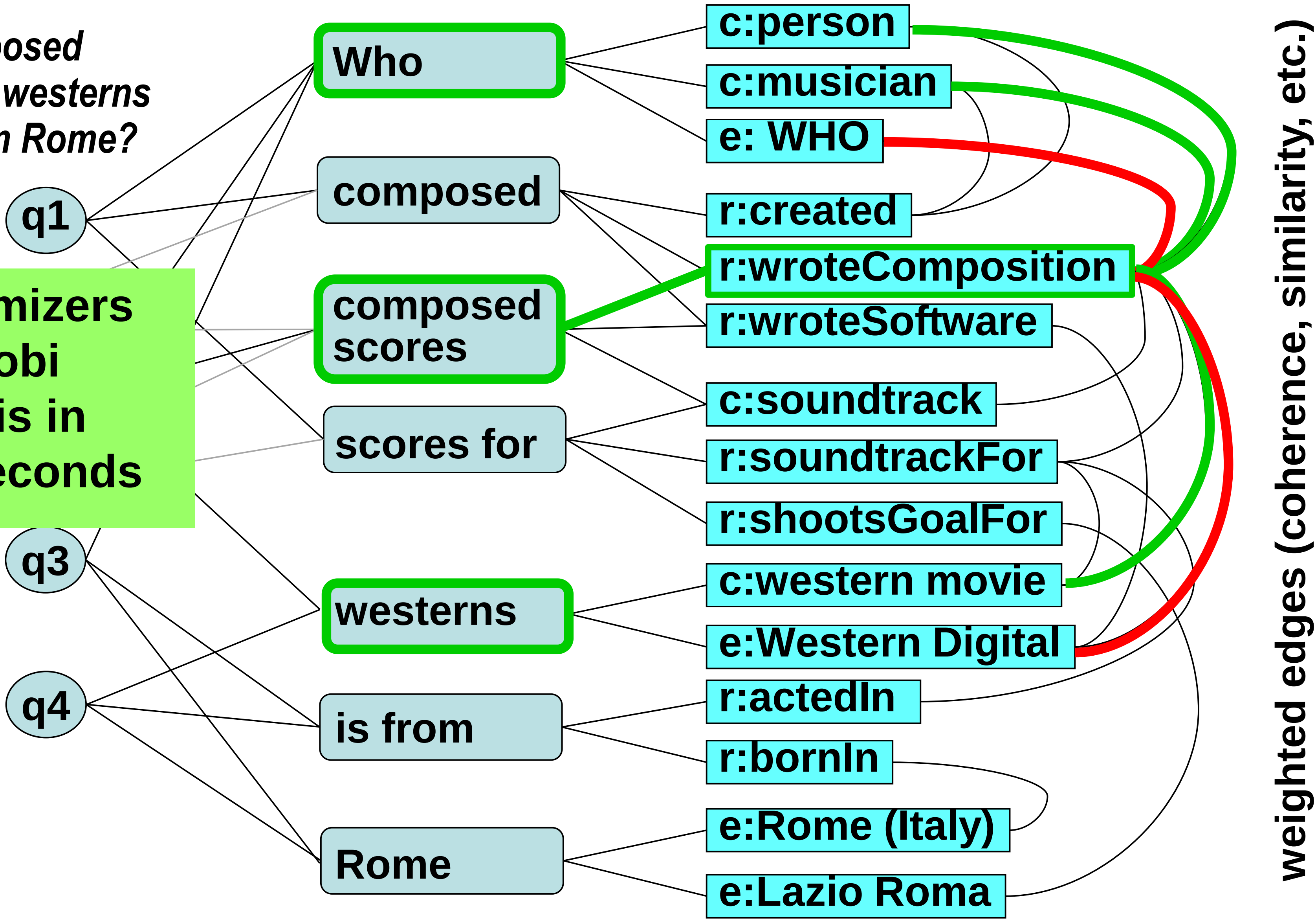
?y type westernMovie

Who is from Rome ⇒ **?x bornIn Rome**

Disambiguation Mapping

[M.Yahya et al.: EMNLP'12, CIKM'13]

Who composed scores for westerns and is from Rome?



ILP optimizers like Gurobi solve this in 1 or 2 seconds

Combinatorial Optimization by ILP (with type constraints etc.)

Prototype for Question-to-Query-based QA

DEANNA

Which composer wrote scores for films and was awarded the Oscar?

Submit

Show Sample Questions ▪ Show Advanced Options

Structured Query

```
?x created ?y .
?x type wordnet_composer_109947232 .
?y type wordnet_movie_106613686 .
?x hasWonPrize Academy_Award
```

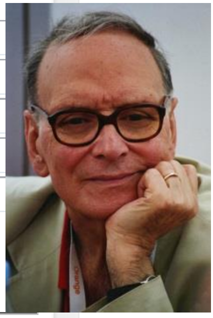
Try it out

YAGO 2 spotlx

Query

Id	Subject	Property	Object	Time		Location	
?id0:	?x	created	?y				
?id1:	?x	type	wordnet_composer_10				
?id2:	?y	type	wordnet_movie_10661:				
?id3:	?x	hasWonPrize	Academy_Award				
?id4:							

query



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

Querying the Web of Data: Sparql Again

Who covered (songs by) European film music composers?

SPOX quads

Select ?x, ?y Where
?x performed ?y .
?y type Music .
?z ?r ?y {“composed“, “created“} .
?z type Musician {“film music“, “composer“} .
?z ??q ?t {“Europe“}

Combining text & data predicates ?
Automatic query relaxation ?
Suitable result ranking ?
Interactive exploration ?
Efficient implementation - at Big-Data scale ?

Querying the Web of Data: Sparql Again

Who covered (songs by) European film music composers?

Select ?x, ?y Where
?x performed ?y .
?y type Music .
?z “composed“ ?y .
?z type “film music composer“ .
?z “birthplace“ Europe

relaxed triples ?

graph language ?

data & text cube ?

Combining text & data predicates ?
Automatic query relaxation ?
Suitable result ranking ?
Interactive exploration ?
Efficient implementation - at Big-Data scale ?

Lessons Learned

Natural Language is most natural UI for searching data & knowledge bases

Sparql-like query language appropriate as API

Translating questions into queries becomes viable and essential

Combine SPO triples with text
Automatically relax queries

What's Next

Make question-to-query translation (semantic parsing) robust and versatile

Automatically generate SPOX query from speech input in discourse

Consider multimodal input:
speech, touch, gesture, gaze, facial expression

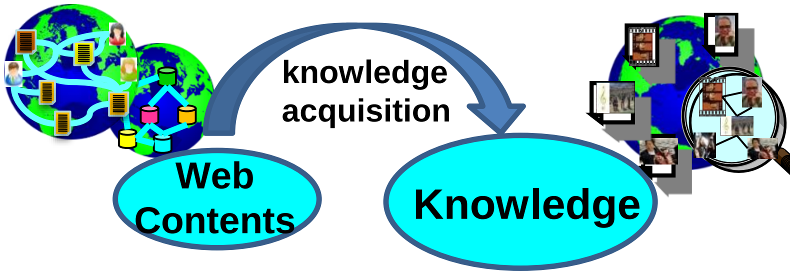
Cope with spatial / temporal / sentiment / belief modifiers

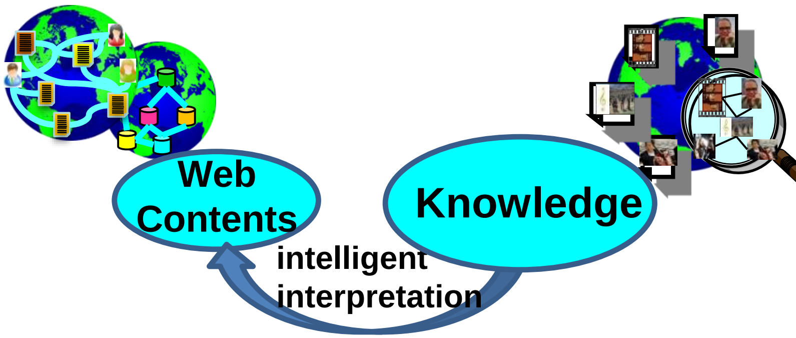
What did George Orwell write after 1984?
What did Bob Dylan write after 2010?
What did Nick Cave write after Grinderman?

Outline

- ✓ Introduction
- ✓ KG Construction
- ✓ Refined Knowledge

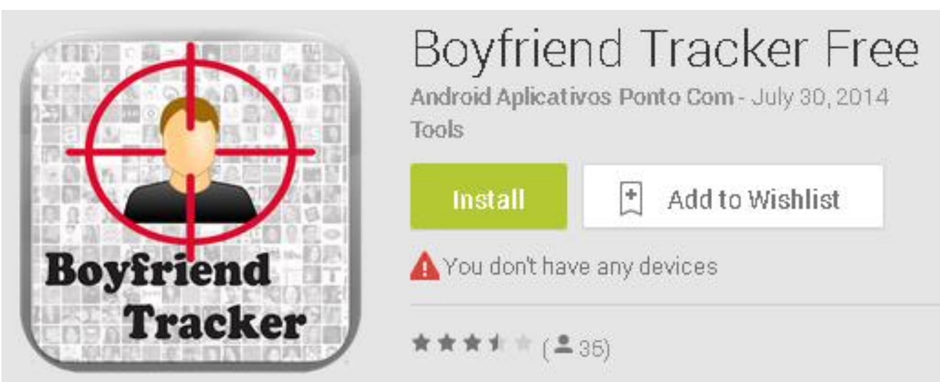
- ✓ Knowledge for Language
- ✓ Deep Text Analytics
- ✓ Search for Knowledge
- ★ **Conclusion**



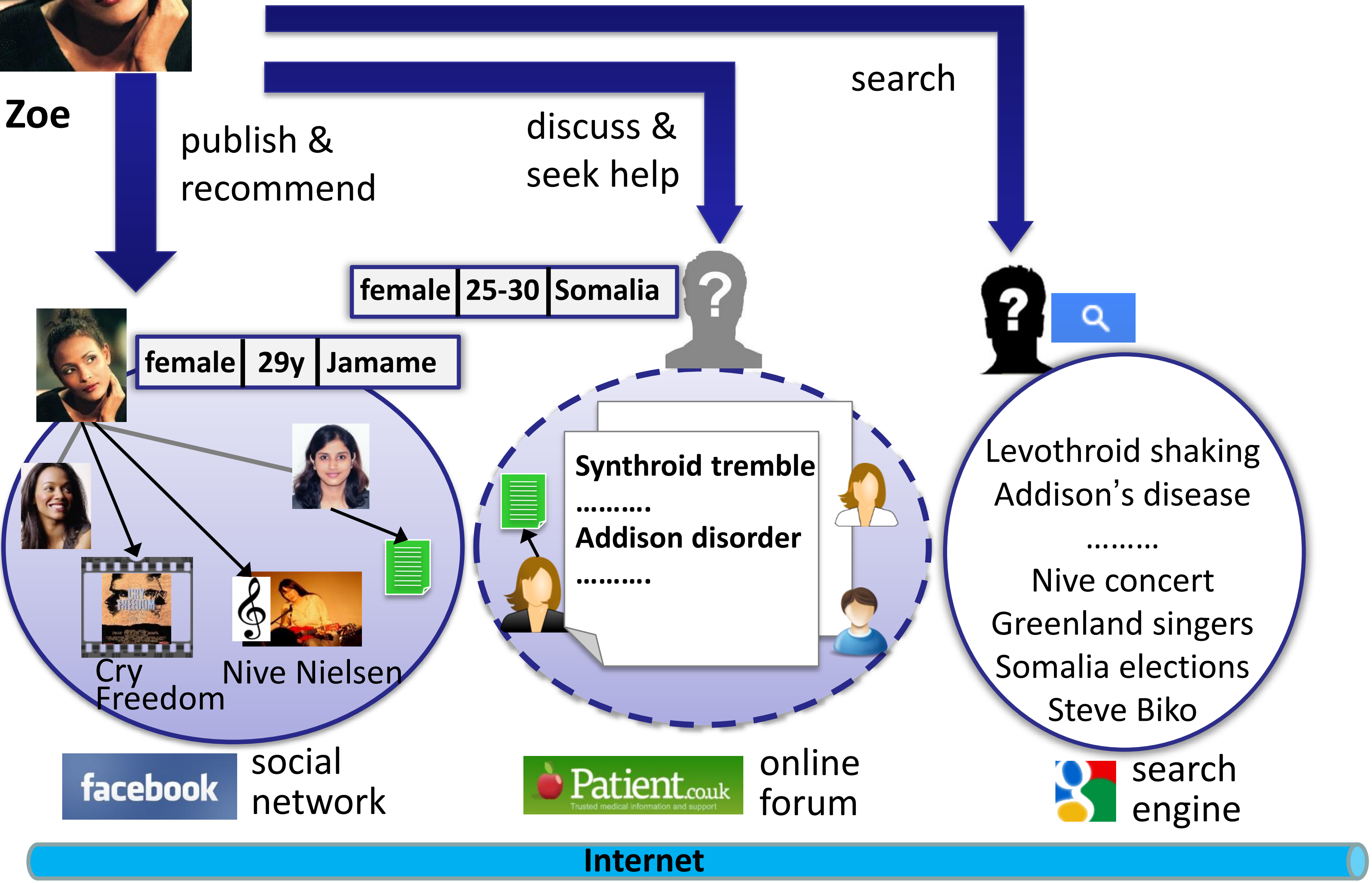


The Dark Side of Digital Knowledge

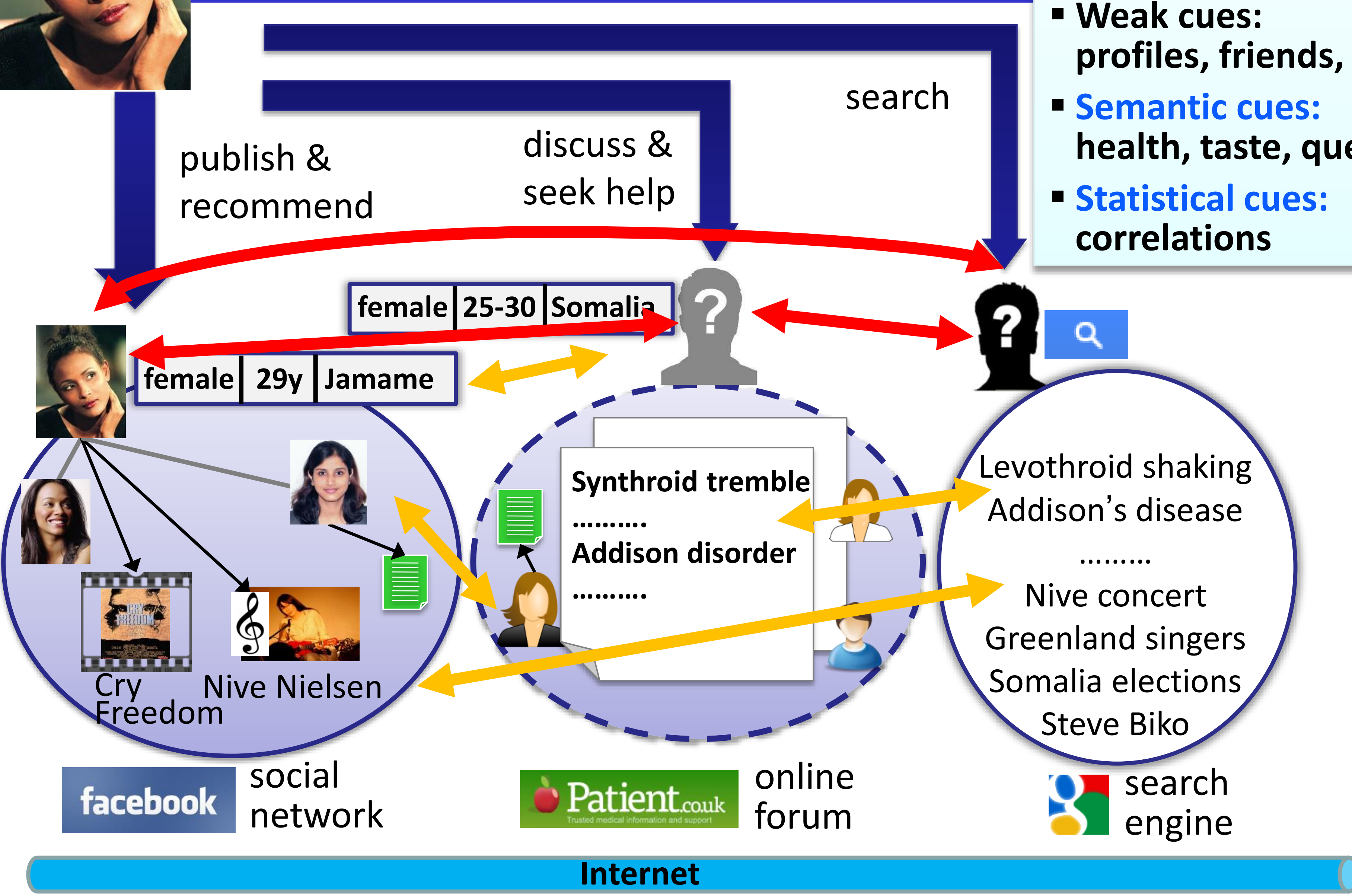
Nobody interested in your research? We read your papers!



Entity Linking: Privacy at Stake



Privacy Adversaries



Linkability Threats:

- **Weak cues:** profiles, friends, etc.
- **Semantic cues:** health, taste, queries
- **Statistical cues:** correlations

Privacy in the Age of Digital Knowledge

Established privacy models

- **Data:** single database
- **Adversary:** computationally powerful, but agnostic
- **Goal:** anonymity guarantee
- **Measures:** data coarsening, perturbation, limit queries

Today‘s user behavior & risks

- **Data & User:** wide contents, **social**, **agile**, **longitudinal**
- **Adversary:** **world knowledge** and **probabilistic inference**
- **Goal:** **alert**&**advise**, **bound risk**
- **Measures:** estimate risk, rank “target users“, switch ids $\rightarrow$ **Privacy Advisor** tool

Summary

- **Knowledge Graphs from Web are Real, Big & Useful: Key Assets for Intelligent Applications**
- **Harvesting Methods Viable at Web Scale for Entities& Classes and for Extracting Relational Facts**
- **Refined Knowledge acquired about Time & Commonsense**
- **NERD Lifts Text to Level of Structured DB and Enables Deep Search & Analytics**
- **Research Challenges & Opportunities:**
  scale & robustness; temporal, multimodal, commonsense;
  open & real-time knowledge discovery; search & analytics …
- **Models & Methods from Different Communities:**
  DB, Web, AI, IR, NLP

References

see comprehensive lists in

Fabian Suchanek and Gerhard Weikum:
Knowledge Bases in the Age of Big Data Analytics,
Tutorial at VLDB 2014

Fabian Suchanek and Gerhard Weikum:
Knowledge Harvesting in the Big-Data Era,
Tutorial at SIGMOD 2013

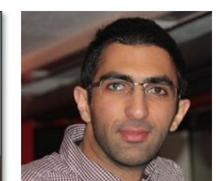
Acknowledgements

erc
European Research Council

M2CI
CLUSTER OF EXCELLENCE

DFG Deutsche Forschungsgemeinschaft

Google



mpi max planck institut informatik

Take-Home Message