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

more knowledge, analytics, insight

Web Contents

Knowledge

knowledge acquisition

intelligent interpretation

Turn Web into Knowledge Base

more knowledge, analytics, insight
Web of Data & Knowledge

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

+ Web tables

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

- 3M entities
- 20M triples

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

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
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

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

∀ x,u,w: (mother(x,u) ∧
mother(x,w)
⇒ u=w)

from humans for humans

from algorithms for machines

Wikipedia

4.5 Mio. English articles
20 Mio. contributors

<table>
<thead>
<tr>
<th>Knowledge Base</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAGO</td>
<td>yago-knowledge.org</td>
</tr>
<tr>
<td>Dbpedia</td>
<td>dbpedia.org</td>
</tr>
<tr>
<td>Freebase</td>
<td>freebase.com</td>
</tr>
<tr>
<td>Wikidata</td>
<td><a href="http://www.wikidata.org">www.wikidata.org</a></td>
</tr>
<tr>
<td>Entitycube</td>
<td>entitycube.research.microsoft.com</td>
</tr>
<tr>
<td></td>
<td>renlifang.msra.cn</td>
</tr>
<tr>
<td>NELL</td>
<td>rtw.ml.cmu.edu</td>
</tr>
<tr>
<td>DeepDive</td>
<td>deepdive.stanford.edu</td>
</tr>
<tr>
<td>Probase</td>
<td>research.microsoft.com/en-us/projects/probase/</td>
</tr>
<tr>
<td>KnowItAll / ReVerb</td>
<td>openie.cs.washington.edu</td>
</tr>
<tr>
<td></td>
<td>reverb.cs.washington.edu</td>
</tr>
<tr>
<td>BabelNet</td>
<td>babelnet.org</td>
</tr>
<tr>
<td>WikiNet</td>
<td><a href="http://www.h-its.org/english/research/nlp/download/">www.h-its.org/english/research/nlp/download/</a></td>
</tr>
<tr>
<td>ConceptNet</td>
<td>conceptnet5.media.mit.edu</td>
</tr>
<tr>
<td>WordNet</td>
<td>wordnet.princeton.edu</td>
</tr>
<tr>
<td>Linked Open Data</td>
<td>linkeddata.org</td>
</tr>
</tbody>
</table>
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

hurricane
hurricane katrina
hurricane sandy
hurricane season

Press Enter to search.

About 42,500,000 results (0.30 seconds)

National Hurricane Center
www.nhc.noaa.gov - National Hurricane Center
Re-analysis of 1946 to 1950 Atlantic Hurricane Seasons Completed (PDF) - Update ...
The Atlantic hurricane season runs from June 1st through November 30th.
Latest Satellite Imagery - Eastern Pacific - Atlantic Graphical TWO - Marine Forecasts

Tropical cyclone - Wikipedia, the free encyclopedia
Hurricane Isabel (2003) as seen from orbit during Expedition 7 of the International Space Station. The eye, eyewall, and surrounding rainbands, characteristics ...
Hurricane (disambiguation) - Typhoon Tip - Scales - Eye

Hurricanes | Ready.gov
www.ready.gov - Disaster Types
Jun 5, 2013 - A hurricane is a type of tropical cyclone or severe tropical storm that forms in the southern Atlantic Ocean, Caribbean Sea, Gulf of Mexico, and ...

Hurricanes - Weather Wiz Kids weather information for kids
www.weatherwizkids.com/weather-hurricane.htm - Weather Wiz Kids
Contains what a hurricane needs to form, stages of a hurricane, and safety tips.

Hurricane Festival
www.hurricane.de/en/ - Hurricane Festival
Hurricane Logo "22864", "band_img": "http://4.hurricane.cdn.smik-networks.de/ccds_cache/img/4d/4d5f369155c1e726f9c2c8aa0295fb.e.460x1000x0.jpg";
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

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.
Google Knowledge Graph: Limitations

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

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 mojo4 music 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 mojo4 music com/.../bob-dylan-new-cover-songs-explored/ ~ Mojo

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
secondhandsongs 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

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, …

Select ?t Where {
   ?t type location .
   ?t hasLabel “Sin City“ .
   ?t hasPart ?d .
   ?d hasLabel “Glitter Gulch“ .
}
Use Case: Deep Data & Text Analytics

Who Covered Whom?

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

<table>
<thead>
<tr>
<th>Musician</th>
<th>Original</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elvis Presley</td>
<td>Frank Sinatra</td>
<td>My Way</td>
</tr>
<tr>
<td>Robbie Williams</td>
<td>Frank Sinatra</td>
<td>My Way</td>
</tr>
<tr>
<td>Sex Pistols</td>
<td>Frank Sinatra</td>
<td>My Way</td>
</tr>
<tr>
<td>Frank Sinatra</td>
<td>Claude Francois</td>
<td>Comme d‘Habitude</td>
</tr>
<tr>
<td>Claudia Leitte</td>
<td>Bruno Mars</td>
<td>Famo$a (Billionaire)</td>
</tr>
<tr>
<td>Only Won</td>
<td>Bruno Mars</td>
<td>I wanna be an engineer</td>
</tr>
<tr>
<td>. . . . . .</td>
<td>. . . . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . . . .</td>
</tr>
</tbody>
</table>
### Use Case: Deep Data & Text Analytics

**Who Covered Whom?**

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

### Musician vs. PerformedTitle

<table>
<thead>
<tr>
<th>Musician</th>
<th>PerformedTitle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex Pistols</td>
<td>My Way</td>
</tr>
<tr>
<td>Frank Sinatra</td>
<td>My Way</td>
</tr>
<tr>
<td>Claudia Leitte</td>
<td>Famo$a</td>
</tr>
<tr>
<td>Petula Clark</td>
<td>Boy from Ipanema</td>
</tr>
</tbody>
</table>

### Musician vs. CreatedTitle

<table>
<thead>
<tr>
<th>Musician</th>
<th>CreatedTitle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Francis Sinatra</td>
<td>My Way</td>
</tr>
<tr>
<td>Paul Anka</td>
<td>My Way</td>
</tr>
<tr>
<td>Bruno Mars</td>
<td>Billionaire</td>
</tr>
<tr>
<td>Astrud Gilberto</td>
<td>Garota de Ipanema</td>
</tr>
</tbody>
</table>

### Name vs. Show

<table>
<thead>
<tr>
<th>Name</th>
<th>Show</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petula C.</td>
<td>Muppets</td>
</tr>
<tr>
<td>Claudia L.</td>
<td>FIFA 2014</td>
</tr>
</tbody>
</table>

### Name vs. Group

<table>
<thead>
<tr>
<th>Name</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sid Vicious</td>
<td>Sex Pistols</td>
</tr>
<tr>
<td>Bono</td>
<td>U2</td>
</tr>
</tbody>
</table>
Use Case: Deep Data & Text Analytics

Who Covered Whom?

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

Big Data & Deep Text

Volume

Velocity

Variety

Veracity

Musician
Sex Pistols
Frank Sinatra
Claudia Leitte
Petula Clark

Artist
Boy from Ipanema

Artist
Astrud Gilberto

Created
My Way
My Way
Billionaire

Title
Garota de Ipanema

Volume

Velocity

Variety

Veracity

Jennifer Lopez Pitbull & Claudia Perform at FIFA Opening

Claudia Leitte feat. Travie McCoy e Bruno Mars - Famosa (Billionaire)

Billionaire space pioneer announces plan to colonise Mars with 80,000 people in two decades

India’s mission to Mars cost less than the movie Gravity
### 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**
✓ 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 travel for a fee)
  - 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 financial institutions)
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

Steve Jobs

From Wikipedia, the free encyclopedia

For the biography, see Steve Jobs (biography).

Steven Paul Jobs (/ˈdʒɒbs/; February 24, 1955 – October 5, 2011) was an American businessman and inventor widely recognized as a charismatic pioneer of the personal computer revolution. 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. 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. In 1986, he acquired the computer graphics division of Lucasfilm Ltd, which was spun off as Pixar Animation Studios. 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, making Jobs Disney’s largest individual shareholder at seven percent and a member of Disney’s Board of Directors.

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. 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 was 56 years old.
Knowledge Communities & New Opportunities

Steve Jobs

From Wikipedia, the free encyclopedia

For the biography, see Steve Jobs (biography).

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

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

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
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

Data Mining

Knowledge Harvesting

Barbara Liskov

Born: 1939 (age 71)
Nationality: American
Fields: Computer Science
Institutions: Massachusetts Institute of Technology
Alma mater: University of California, Berkeley
Doctoral advisor: John McCarthy
Notable awards: IEEE John von Neumann Medal, A.M. Turing Award

Serge Abiteboul

Fields: Computer Science
Institutions: University of California, Berkeley
Alma mater: University of Wisconsin-Madison
Doctoral advisor: Joseph Naughton, Michael Stonebraker

Jeffrey Ullman

Born: November 22, 1942 (age 77)
Citizenship: American
Nationality: American
Institutions: Columbia University, Princeton University
Known for: Abiteboul-Vianu Theorem
Fields: Computer Science
Alma mater: University of California, Berkeley
Doctoral advisor: Joseph Naughton

Joseph M. Hellerstein

Fields: Computer Science
Institutions: University of Wisconsin-Madison
Alma mater: University of Wisconsin-Madison
Doctoral advisor: Jeffrey Naughton, Michael Stonebraker

Notable awards:
- ACM SIGMOD Edgar F. Codd Innovations Award (1998)
- ACM SIGMOD Test of Time Award 2004
- Prix d'Informatique de l'Academie des Sciences (Prix EAOS) 2007
- ACM PODS Alberto O. Mendelzon Test-of-Time Award 2008

Known for: Abiteboul-Vianu Theorem

Acknowledgments:

- Alexander Birman
- Sunit Choudhuri, Evan Cohn, Alan Demers, Marcia Derr, Nabil El Djeh, Anmol Fong
- Lee Goel, Deepak Goyal, Ashish Gupta, Himanshu Gupta, Udalakshmi Gupta, Venkatesh Harinarayanan, Tamer Haset, Matthew Hochst, Daniel Hochberg, Peter Hochschid, Peter Honeyman, Edward Horvath, Gregory Hunter, Nam (Pierre) Hoyt, Hakim Jakobsen, John Kao, Marc Kaplan, Anna Karlin, Kenji Karp, Henry Kathleen Riker, Chen Li, Leonard Liu, George Leek, David Maier, Harry Maier, Alberto O. Mendelzon, Katherine Morris, Ivarul Munic, Jeffrey F. Naughton, Svetlana Nesterov, Geoffrey Phipps, Thane Plummer, Arun Rajaraman, Kenneth Ross, Eric Scheid, Yehudah Sage, Yatin Kingsay, Dilip Samet, Edrite Sciere, Reza Sethi, Alan Siegel, Howard Siegel, Albert Tamos, Howard Tidrick, Alan Van Gelder, Vladimir Vassilev, Chang (Chen) Yang, Mihae Vukanovic,
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

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

Gong Li

{{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

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

(?i)IBL\|BEG\s*birth[_ ]\?place\s*="\s*(\.*?)IBL\|END" => "$0 wasBornIn @WikiLink($1)
<table>
<thead>
<tr>
<th>Relational Facts from Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>composed (&lt;musician&gt;, &lt;song&gt;)</td>
</tr>
<tr>
<td><em>Bob Dylan</em> wrote the song <em>Knockin’ on Heaven’s Door</em></td>
</tr>
<tr>
<td><em>Lisa Gerrard</em> wrote many haunting pieces, including <em>Now You Are Free</em></td>
</tr>
<tr>
<td><em>Morricone’s masterpieces</em> include the <em>Ecstasy of Gold</em></td>
</tr>
<tr>
<td><em>Dylan’s song</em> <em>Hurricane</em> was covered by <em>Ani DiFranco</em></td>
</tr>
<tr>
<td><em>Strauss’s famous work</em> was used in 2001, titled <em>Also sprach Zarathustra</em></td>
</tr>
<tr>
<td><em>Frank Zappa</em> performed a jazz version of <em>Rota’s Godfather Waltz</em></td>
</tr>
<tr>
<td><em>Hallelujah</em>, originally by <em>Cohen</em>, was covered in many movies, including <em>Shrek</em></td>
</tr>
</tbody>
</table>

Pattern-based Gathering (statistical evidence) + Constraint-aware Reasoning (logical consistency)
Pattern-based Harvesting: Fact-Pattern Duality

Task populate relation **composed** starting with **seed facts**

**Facts & Fact Candidates**

- (Dylan, Knockin)
- (Gerrard, Now)
- (Dylan, Hurricane)
- (Morricone, Ecstacy)
- (Zappa, Godfather)
- (Mann, Buddenbrooks)
- (Gabriel, Biko)
- (Puebla, Che Guevara)
- (Mezrich, Zuckerberg)
- (Jobs, Apple)
- (Newton, Gravity)

**Patterns**

- X wrote the song Y
- X wrote … including Y
- X covered the story of Y
- X has favorite movie Y
- X is famous for Y

- 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 confusuable 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:**

  Melbourne lies on the banks of the **Yarra**

  People in **Cairo** like wine from the **Yarra** valley
### 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:

\[
\forall x, y: \text{composed }(x,y) \Rightarrow \text{type}(x) = \text{musician} \\
\forall x, y: \text{composed }(x,y) \Rightarrow \text{type}(y) = \text{song} \\
\forall x, y, z: \text{composed }(x,y) \land \text{appearedIn}(y,z) \Rightarrow \text{wroteSoundtrackFor }(x,z) \\
\forall x, y, t, b, e: \text{composed }(x,y) \land \text{composedInYear }(y, t) \land \text{bornInYear }(x, b) \land \text{diedInYear }(x, e) \Rightarrow b < t \leq e \\
\forall x, y, w: \text{composed }(x,y) \land \text{composed}(w,y) \Rightarrow x = w \\
\forall x, y: \text{sings}(x,y) \land \text{type}(x, \text{singer-songwriter}) \Rightarrow \text{composed}(x,y)
\]

**consistent subset(s) of hypotheses** ("possible world(s)“, “truth“)

→ **Weighted MaxSat** solver for set of logical clauses  
→ **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)

\[
\text{spouse}(x,y) \land \text{diff}(y,z) \Rightarrow \neg \text{spouse}(x,z)
\]
\[
\text{spouse}(x,y) \land \text{diff}(w,y) \Rightarrow \neg \text{spouse}(w,y)
\]
\[
\text{spouse}(x,y) \Rightarrow \text{female}(x)
\]
\[
\text{spouse}(x,y) \Rightarrow \text{male}(y)
\]

RVs coupled by MRF edge if they appear in same clause

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

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

Joint distribution has product form over all cliques
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 \(<\text{entity1}, \text{relation}, \text{entity2}>\)

\(<\text{CarlaBruni marriedTo NicolasSarkozy}> \in \text{Person} \times \text{R} \times \text{Person}\)
\(<\text{NataliePortman wonAward AcademyAward} > \in \text{Person} \times \text{R} \times \text{Prize}\)

Open and Dynamic Knowledge Harvesting:
would like to discover new entities and new relation types
\(<\text{name1, phrase, name2}>\)

* 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 ...


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

333 answers from 977 sentences (cached)

- person (101)
- composer (91)
- music contributor (84)
- group member (79)
- award nominees (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

Open Information Extraction

?x „an affair with“ ?y

307 answers from 1015 sentences (cached)

- Whitney Houston, Jermaine Jackson (7)
- John McCain, a lobbyist (5)
- Bill Clinton, Monica Lewinsky (5)
- Jesus, Mary Magdalene (5)
- Suzanne Coleman, Bill Clinton (3)
- her mother, Tiger Woods (3)
- the medias, Barack Obama (3)
- Newt Gingrich, House (3)
- Thomas Jefferson, Sally Hemings (3)
- Saddam Hussein, Samira Shahbandar (3)
- Suzanne Coleman Reportedly, Bill Clinton (3)
- his wife, George Foreman (2)
- Clementine Churchill, Baroness Spencer-Churchill, Terence Phillip (2)
- the extraterrestrial, Hillary Rodham Clinton (2)
- an unnamed intern, John F. Kennedy (2)

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>*</td><td>°C</td></tr>;  <h4>capitals <ul><li>Y: X </li></ul> ...

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) ∧ city(Y) ⇒ mayorOf(X,Y);  mayorOf(X,Y) ⇒ livesIn(X,Y)

http://rtw.ml.cmu.edu
### Recently-Learned Facts

<table>
<thead>
<tr>
<th>instance</th>
<th>iteration</th>
<th>date learned</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>bresaola is a <strong>visualizable thing</strong></td>
<td>922</td>
<td>05-may-2015</td>
<td>96.4</td>
</tr>
<tr>
<td>francis_denwent_wood is a <strong>visual artist</strong></td>
<td>922</td>
<td>05-may-2015</td>
<td>99.9</td>
</tr>
<tr>
<td>frank_g is an <strong>Australian person</strong></td>
<td>922</td>
<td>05-may-2015</td>
<td>92.2</td>
</tr>
<tr>
<td>g_protein_coupled_receptor_124 is a <strong>protein</strong></td>
<td>922</td>
<td>05-may-2015</td>
<td>100.0</td>
</tr>
<tr>
<td>n_butyl_benzyl_phthalate is a <strong>chemical</strong></td>
<td>922</td>
<td>05-may-2015</td>
<td>100.0</td>
</tr>
<tr>
<td>chicken001 eat potatoes</td>
<td>926</td>
<td>20-may-2015</td>
<td>100.0</td>
</tr>
<tr>
<td>bioinformatics is an academic program at the university college</td>
<td>922</td>
<td>05-may-2015</td>
<td>93.8</td>
</tr>
<tr>
<td>samuel_j_palmer is the CEO of ibm</td>
<td>926</td>
<td>20-may-2015</td>
<td>100.0</td>
</tr>
<tr>
<td>national001 is a company that has an office in the country czech republic</td>
<td>922</td>
<td>05-may-2015</td>
<td>99.2</td>
</tr>
<tr>
<td>the companies dc and fox_news_channel compete with eachother</td>
<td>922</td>
<td>05-may-2015</td>
<td>98.4</td>
</tr>
</tbody>
</table>

[http://rtw.ml.cmu.edu/rtw/kbbrowser/](http://rtw.ml.cmu.edu/rtw/kbbrowser/)
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 (98.9%) 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

- agentcollaborateswith
  - http://rtw.ml.cmu.edu/rtw/kbbrowser/
### musicianplaysinstrument

Specifies that a musical instrument is played by a particular musician

See [metadata](http://rtw.ml.cmu.edu/rtw/kbbrowser/) for musicianplaysinstrument

<table>
<thead>
<tr>
<th>instance</th>
<th>iteration</th>
<th>date learned</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>adam, drums</td>
<td>799</td>
<td>27-Dec-2013</td>
<td>100.0</td>
</tr>
<tr>
<td>adam, guitar</td>
<td>799</td>
<td>27-Dec-2013</td>
<td>100.0</td>
</tr>
<tr>
<td>bach, piano</td>
<td>551</td>
<td>19-Apr-2012</td>
<td>100.0</td>
</tr>
<tr>
<td>bach, violin</td>
<td>551</td>
<td>19-Apr-2012</td>
<td>100.0</td>
</tr>
<tr>
<td>barbar, violin</td>
<td>598</td>
<td>21-Jun-2012</td>
<td>100.0</td>
</tr>
<tr>
<td>bb_king, guitar</td>
<td>680</td>
<td>09-Jan-2013</td>
<td>100.0</td>
</tr>
<tr>
<td>beethoven, piano</td>
<td>853</td>
<td>11-Jul-2014</td>
<td>100.0</td>
</tr>
<tr>
<td>beethoven, violin</td>
<td>853</td>
<td>11-Jul-2014</td>
<td>100.0</td>
</tr>
<tr>
<td>ben_harper, guitar</td>
<td>820</td>
<td>08-Mar-2014</td>
<td>100.0</td>
</tr>
<tr>
<td>billie_joe_armstrong, guitar</td>
<td>818</td>
<td>03-Mar-2014</td>
<td>100.0</td>
</tr>
<tr>
<td>brahms, piano</td>
<td>592</td>
<td>13-Jun-2012</td>
<td>100.0</td>
</tr>
<tr>
<td>brahms, violin</td>
<td>503</td>
<td>06-Feb-2012</td>
<td>100.0</td>
</tr>
<tr>
<td>buddy_guy, guitar</td>
<td>684</td>
<td>01-Dec-2012</td>
<td>100.0</td>
</tr>
<tr>
<td>b_b_king, guitar</td>
<td>406</td>
<td>08-Sep-2011</td>
<td>(Seed) 100.0</td>
</tr>
<tr>
<td>charlie, guitar</td>
<td>724</td>
<td>12-Apr-2013</td>
<td>100.0</td>
</tr>
<tr>
<td>chopin, piano</td>
<td>683</td>
<td>15-Jan-2013</td>
<td>100.0</td>
</tr>
<tr>
<td>copland, piano</td>
<td>665</td>
<td>05-Dec-2012</td>
<td>100.0</td>
</tr>
<tr>
<td>david, bass</td>
<td>904</td>
<td>20-Feb-2015</td>
<td>100.0</td>
</tr>
<tr>
<td>david, drums</td>
<td>904</td>
<td>20-Feb-2015</td>
<td>100.0</td>
</tr>
<tr>
<td>david, guitar</td>
<td>904</td>
<td>20-Feb-2015</td>
<td>100.0</td>
</tr>
<tr>
<td>david, keyboards</td>
<td>904</td>
<td>20-Feb-2015</td>
<td>100.0</td>
</tr>
<tr>
<td>earl_scruggs, banjo</td>
<td>904</td>
<td>20-Feb-2015</td>
<td>100.0</td>
</tr>
<tr>
<td>eddie, guitar</td>
<td>904</td>
<td>20-Feb-2015</td>
<td>100.0</td>
</tr>
<tr>
<td>fox, drums</td>
<td>904</td>
<td>20-Feb-2015</td>
<td>100.0</td>
</tr>
<tr>
<td>fox, piano</td>
<td>904</td>
<td>20-Feb-2015</td>
<td>100.0</td>
</tr>
</tbody>
</table>

[http://rtw.ml.cmu.edu/rtw/kbbrowser/](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, Hallelujah)

performed: (Amy,Cupid), (Amy, Black), (Nina, Don‘t),
(Cohen, Hallelujah), (Dylan, Knockin), ...

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, ...
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

lead singer;

Paramore, Hayley Williams
All (band), Dave Smalley
Alabama (band), Randy Owen
Clutch (band), Neil Fallon
Nirvana (band), Kurt Cobain

In particular, Rossdale’s forced stream of consciousness dismissed by some as an imitation singer, Kurt Cobain.

Los Bravos, Mike Kogel
Twisted Sister, Dee Snider
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 $\rightarrow$ 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“ $\rightarrow$ numeric)
Knowledge Graphs: Large Size, Good Coverage, High Quality, Open-World

Are We Done?
Which salient facts about an entity are captured in infobox?

<table>
<thead>
<tr>
<th>Johnny Cash</th>
</tr>
</thead>
</table>

**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–)  
Barbara (1959–)  
John (1970–)

**Military career**

<table>
<thead>
<tr>
<th>Allegiance</th>
<th>United States of America</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service/branch</td>
<td>United States Air Force</td>
</tr>
<tr>
<td>Years of service</td>
<td>1950–1954</td>
</tr>
<tr>
<td>Rank</td>
<td>Staff sergeant</td>
</tr>
</tbody>
</table>

**Musical career**

<table>
<thead>
<tr>
<th>Genres</th>
<th>Country, rockabilly,[1] rock and roll, gospel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instruments</td>
<td>Vocals, guitar</td>
</tr>
<tr>
<td>Labels</td>
<td>Sun, Columbia, Mercury, American, House of Cash, Legacy Recordings</td>
</tr>
</tbody>
</table>

**Website**  
johnnycash.com

**Notable instruments**  
Martin Acoustic Guitars[2]
Which salient facts about an entity are captured in infobox?

Let the Bells Ring
isOnAlbum
AbbatoirBlues

lyricsAbout
JohnnyCash

O‘Children
isOnAlbum
AbbatoirBlues

featuredInMovie
HarryPotter &
Deathly Hallows 1

WarrenEllis
performsOnAlbum
AbbatoirBlues

…..

not in any KG!
How many Jeopardy questions could be answered having solely Yago+Dbpedia+Freebase?


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, analysts, ... )
✓ 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) [Dec2000, 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 → wife
- inclusion dependencies: marriedPerson ⊆ adultPerson
- age/time/gender restrictions: birthdate + Δ < marriage < divorce

<table>
<thead>
<tr>
<th>Political party</th>
<th>Spouse</th>
<th>Children</th>
<th>Residence</th>
<th>Alma mater</th>
<th>Occupation</th>
<th>Religion</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMP (2002–2012)</td>
<td>Cécilia Ciganer-Albéniz (div.)</td>
<td>Jean (by Culioli), Louis (by Ciganer-Albéniz)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Carla Bruni</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dating Considered Harmful

Nicolas Sarkozy
From Wikipedia, the free encyclopedia

Nicolas Sarkozy (pronounced [ni.kɔ.la saʁ.kɔ.zi] (listen), born Nicolas Paul Stéphane Sarkozy de Nagy-Bocsá, 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

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

Sarkozy was enrolled in the Lycée Chaptal, a well-regarded public middle school in Paris's 8th arrondissement, where he failed his sixième. His family then sent him to the Cours Saint-Louis de Monceau, a private Catholic school in the 17th arrondissement, where he was reportedly a mediocre student, 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. After graduating, he entered the Institut d'Études Politiques de Paris, better known as Sciences Po, (1979–1981) but failed to graduate due to an insufficient
Temporal Taggers:

capture temporal expressions and normalize them

- TARSQI: [http://www.timeml.org/site/tarsqi/](http://www.timeml.org/site/tarsqi/)
  based on regular expressions
- HeidelTime: [http://heideltime.ifi.uni-heidelberg.de/heideltime/](http://heideltime.ifi.uni-heidelberg.de/heideltime/)
  regex, rules, ...
supports many languages,
captures some implicit dates (e.g. Christmas holidays)
Hong Kong is poised to hold the first election in more than half a century that includes a democracy advocate seeking high office in territory controlled by the Chinese government in Beijing. A pro-democracy politician, Alan Leong, announced 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 election. It will be the first contested election for chief executive since Britain returned Hong Kong to China in 1997. Mr. Tsang, an able administrator who took office during the early stages of a sharp economic upturn in 2005, 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 another five years, said Mr. Leong, the former chairman of the Hong Kong Bar Association.
Temporal Facts from Text

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

[Y. Wang et al. 2011]
Reasoning on T-Fact Hypotheses

Temporal-fact hypotheses:
\[ s(Ca,Ni)@[2008,2012]{0.7}, \ s(Ca,Ben)@[2010]{0.8}, \ s(Ca,Al)@[2007,2008]{0.2}, \ s(Li,Ni)@[1996,2004]{0.9}, \ s(Li,Joe)@[2006,2008]{0.8}, \ldots \]

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 \( \sum w_i X_i \) with constraints

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

Efficient ILP solvers:
www.gurobi.com
IBM Cplex
…
Events in the Knowledge Base

French Revolution

The French Revolution was a period of radical social and political upheaval in France from 1789 to 1799 that profoundly affected French and modern history, marking the decline of powerful monarchies and churches and the rise of democracy.

Start date: 1789
End date: 1799

Related people:
- Napoleon
- Louis XVI of France
- Maximilien de Robespierre
- Marie Antoinette
- Jean-Paul Marat

People also search for:
- American Revolution
- Storming of the Bastille
- Napoleonic Wars
- American Revolution
- Russian Revolution

Data from: Freebase
Feedback

2008 Summer Olympics

The 2008 Summer Olympic Games, officially known as the Games of the XXIX Olympiad, was a major international multi-sport event that took place in Beijing, China, from August 8 to 24, 2008. A total of 10,942 athletes from 204 National Olympic Committees participated.

Start date: 08 Aug 2008
End date: 24 Aug 2008
Number of athletes: 10,500

Related people:
- Michael Phelps
- Usain Bolt
- Ronda Rousey
- Nastia Liukin
- Shawn Johnson

People also search for:
- 2012 Summer Olympics
- 2004 Summer Olympics
- 2000 Summer Olympics
- 1996 Summer Olympics
- 2016 Summer Olympics

Data from: Freebase
Feedback
The Australian Voices World Premiere: Boombox

Location
Federation Square
Corner of Swanston & Flinders Streets
Melbourne VIC 3000
Deakin Edge

Contact details
0418 037 283
singersfestival@hotmail.com
www.schoollforhardknocks.org.au

Dates and times
06/06/2016
Sat 7.30pm – 8.45pm

Price
$30 full, $20 concession
Under 18, 800 family of 4

Bookings
Bookings available via
0418 037 283
Click here to book

Payment method accepted
All major cards

An old-school boombox sits on stage. A choir enters, presses play, and a medley of choral music ensues. The ensemble plays with toy instruments, store-up comedy, rap battles, sporting commentary, mime, beatboxing, recordings from unruly parliament sessions and—frankly—whatever else they can get away with.

Each piece melds into the next, without clear distinctions between composers, works or genres. There is no plot, no repetition, no rules, no robes. Featuring music by Amanda Cole, Robert Davidson, Gordon Hamilton, Isabella Geromutto and Nigel Butterley.

Presented by the Melbourne International Singers Festival, with proceeds supporting School of Hard Knocks.
EVIN: Populating KB with Emerging Events

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

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

E. Kuzey et al. [WWW’14, CIKM’14]

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
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
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/
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
→ 5-grams + frequencies

Confidence score: PMI (X,Y), PMI (p,(XY)), support(X,Y), ... are features for regression model
WebChild: Commonsense Properties

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/

Mountain bike, all-terrain bike, off-roader
A bicycle with a sturdy frame and fat tires, originally designed for riding in mountainous country

NEIL: Never Ending Image Learner
Crawl, See, I Learn.

Bicycle

<table>
<thead>
<tr>
<th>Clusters Discovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bike</td>
</tr>
<tr>
<td>Bicycle, bike</td>
</tr>
<tr>
<td>Pedals</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relationships Discovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedals partOf bike</td>
</tr>
<tr>
<td>Bike occursAt park</td>
</tr>
</tbody>
</table>

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

∀ x,m,c: type(x,child) ∧ mother(x,m) ∧ livesIn(m,t) ⇒ livesIn(x,t)
∀ x,m,f: type(x,child) ∧ mother(x,m) ∧ spouse(m,f) ⇒ father(x,f)

Advance rules beyond Horn clauses:
specified by human experts

∀ x: type(x,spider) ⇒ numLegs(x)=8
∀ x: type(x,animal) ∧ hasLegs(x) ⇒ even(numLegs(x))
∀ x: human(x) ⇒ (∃ y: mother(x,y) ∧ ∃ z: father(x,z))
∀ x: human(x) ⇒ (male(x) ∨ 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

Hurricane about Carter, is on Bob's Desire.

It is played in the film with Washington.

contextual similarity: mention vs. Entity (bag-of-words, language model)

prior popularity of name-entity pairs
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

**Hurricane**, a racism protest song, is on Bob's Desire. It is played in the film with Washington.

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

- **r**acism protest song boxing champion wrong conviction
- **r**acism 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

Coherence: (partial) overlap of (statistically weighted) entity-specific keyphrases
Named Entity Recognition & Disambiguation

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

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)
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 \ldots mk | e1 \ldots ek]
- factorize by all m-e pairs and e1-e2 pairs
- use MCMC, hill-climbing, LP etc. for solution
• 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
  
Coherence Graph Algorithm

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

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

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

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.

\[ KP(\text{new}) = KP(\text{name}) - \bigcup_e KP(\text{e}) \]

with statistical weights

[J. Hoffart et al.: WWW’14]
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

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

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
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.
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.
NERD on Tables

Disambiguation Method:
- prior
- prior+sim
- prior+sim+coherence

Parameters: (default should be OK)
- Prior-Similarity-Coherence balancing ratio: prior VS. sim. balance = 0.4
  (prior+sim.) VS. coh. balance 0.6
- Ambiguity degree 5

Mention Extraction:
- Stanford NER
- Manual

You can manually tag the mentions by putting them in the manual mode.
Entity Matching in Structured Data

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

<table>
<thead>
<tr>
<th>Musician</th>
<th>Song</th>
<th>Year</th>
<th>Listeners</th>
<th>Charts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinatra</td>
<td>My Way</td>
<td>1969</td>
<td>435 420</td>
<td></td>
</tr>
<tr>
<td>Sex Pistols</td>
<td>My Way</td>
<td>1978</td>
<td>87 729</td>
<td></td>
</tr>
<tr>
<td>Pavarotti</td>
<td>My Way</td>
<td>1993</td>
<td>4 239</td>
<td></td>
</tr>
<tr>
<td>C. Leitte</td>
<td>Famo$a</td>
<td>2011</td>
<td>272 468</td>
<td></td>
</tr>
<tr>
<td>B. Mars</td>
<td>Billionaire</td>
<td>2010</td>
<td>218 116</td>
<td></td>
</tr>
</tbody>
</table>

**Sinatra debuts cover of ‘Bang Bang’ by Cher/Nancy**

**Travie McCoy: Billionaire ft. Bruno Mars**

[OFFICIAL VIDEO]: youtu.be/8aRor905cCw via @YouTube

**Jennifer Lopez Pitbull & Claudia Perform at FIFA Opening**

**Frank Sinatra’s 1963 Playboy interview**

Playboy: Alright, let’s start with the most basic question: Are you a religious man? Do you believe in God?

Sinatra: Well, that’s for do-gooders. I think I can speak up my religious feelings in a couple of paragraphs. First, I believe in you and me. I’m like Albert Schweitzer and Bertrand Russell and Albert Einstein in that I have a respect for life — in any form. I believe in nature, in the birds, the sea, the...
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*

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
**MarketWatch**


... 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 ...
From chocolate king to president of Ukraine Outside the Box


... 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 natural gas giant, said the company has withdrawn December’s discount ...
From chocolate king to president of Ukraine. Outside the Box. AFP/Getty Images Petro Poroshenko received support in his campaign for president from boxer Vitali Klitschko. Poroshenko is leading in polls by a mile. The only question seems to be whether he will win an outright majority on May 25 or require a runoff a few weeks later. Yet he has attracted little attention around the world and remains a studiously vague figure to his own electorate. That’s partly the media’s fault. Men in ski masks waving Kalashnikovs, like the so-called separatist rebels in Eastern Ukraine, are an irresistible draw for reporters — and more so if they succeed in provoking actual bloodshed. Other stories, like who becomes president in about a dozen days, fall by the wayside. But Poroshenko see full article »
Deep Text Search & Analytics

French footballers
date: 2014-05-04
percentage: 1.218%
absolute value (not smoothed): 33

Brazilian footballers
date: 2014-05-04
percentage: 0.647%
absolute value (not smoothed): 34

German footballers
date: 2014-05-04
percentage: 0.624%
absolute value (not smoothed): 26

https://stics.mpi-inf.mpg.de
Are patients with primary hypothyroid observational study.

Mital A1, Dharmalingam M2, Revarthy A3.

Author Information
1 Department of Endocrinology, Medanta - The Medicity, Gurgaon, India. 2 Department of Endocrinology, Bangalore Endocrinology and Diabetes Centre, Bangalore, India. 3 Medical Affairs, Metabolics and Endocrinology, Medanta - The Medicity, Gurgaon, India.

Abstract
BACKGROUND: A large proportion of patients with Stimulating Hormone (TSH) values. There is a possible thyroid hormone treatment.

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, observational study in adult patients with primary hypothyroidism on treatment with levothyroxine for at least 2 m 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 150 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 ± 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 ± 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.
I was diagnosed 2 years with hypothyroidism. My TSH was 6.0. I tried Synthroid, Levothyroxine, and Armour. I seemed to have side effects to all of them, regardless of the dosing. Side effects included: nausea, palpitations, shortness of breath, weakness, insomnia, etc. I had all these and more. I have taken Synthroid, Plaquenil, Celexa, and a few other antidepressants, although what seem like withdrawals. My TSH is now 8.0, my T4 is 3.0, but I am still have the above symptoms which tend to come and go. Told to go on T3 and then go back to T4. I have no quality of life.
Deep Data & Text Analytics: Side Effects of Drug Combinations

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)

Structured Expert Data

Social Media


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**
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.

---

**Language Objectivity**

- p1
- p2
- p3

---

**User Trustworthiness**

- u1
- u2
- u3

---

**Statement Credibility**

- s1
- s2

---

- hasSideEffects (Xanax, hallucinations)
- hasSideEffects (Xanax, dizziness)
- hasSideEffects (Xanax, insomnia)
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.**

---

**Language Objectivity**

- p1
- p2
- p3

**User Trustworthiness**

- u1
- u2
- u3

**Statement Credibility**

- s1
- s2

---

**joint reasoning with probabilistic graphical model**

(semicolon-supervised heterogeneous CRF with EM-style inference)
Maestro from Rome wrote scores for westerns. The Maestro from Rome wrote scores for westerns. Ma played his version of the Ecstasy.

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>
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/
wrote scores for westerns from Rome: Ennio Morricone composed for MaestroCard en: MaestroCard

r: actedIn
r: bornIn
r: composed
r: giveExam
r: soundtrackFor
r: shootsGoalFor
r: soundtrackFor
r: westernMovie
r: Western Digital

Combinatorial Optimization by ILP (with type constraints etc.)

ILP optimizers like Gurobi solve this in seconds

Disambiguation for Entities, Classes & Relations

(M. Yahya et al.: EMNLP’12, CIKM’13; J. Berant et al.: EMNLP’13 …)
wrote scores for westerns from Rome:

- MaestroCard: e: MaestroCard
- Ennio Morricone: e: Ennio Morricone
- conductor: c: conductor
- musician: c: musician
- actedIn: r: actedIn
- bornIn: r: bornIn
- Rome: e: Rome (Italy)
- Lazio Roma: e: Lazio Roma
- composed: r: composed
- giveExam: r: giveExam
- soundtrack: c: soundtrack
- soundtrackFor: r: soundtrackFor
- shootsGoalFor: r: shootsGoalFor
- western: c: western movie
- Western Digital: e: Western Digital

Combinatorial Optimization by ILP (with type constraints etc.)

ILP optimizers like Gurobi solve this in seconds

Disambiguation for Entities, Classes & Relations
(M. Yahya et al.: EMNLP’12, CIKM’13; J. Berant et al.: EMNLP’13 …)
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

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

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

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

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

- **Enzymes that inhibit HIV?**
  - Antidepressants that interfere with blood-pressure drugs?
  - German philosophers influenced by William of Ockham?
Who composed scores for westerns and is from Rome?

- Casual ? (form-based)
- Textual ? (keywords)
- Visual ? (point&drag)
- Sparql ? (query language)
- Natural ? (natural language)
Who composed scores for westerns and is from Rome?

S

Elisa
Aida
Vendetta
Elisa

P

created
contributedTo
type
bornIn

O

Aida
Vendetta
westernMovie
Rome
Who composed scores for westerns and is from Rome?

```
Select ?x Where {
  ?x created ?s .
  ?s contributesTo ?m .
  ?m type westernMovie .
  ?x bornIn Rome .}
```
Who composed scores for westerns and is from Rome?

q: composer Rome scores westerns

composer (creator of music)
Rome (Italy)
Media Composer video editor

Rome (NY)
Lazio Roma

film music
goal in football
western movies
western world

Western Digital
Western (airline)
Western (NY)

... born in ...
... plays for ...
... used in ...
... recorded at ...
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

scores for westerns

is from Rome
Semantic Parsing:
from Triploids to SPO Triple Patterns

Map names into entities or classes, phrases into relations

Who composed scores

scores for westerns

Who is from Rome

?x composed ?s
?x type composer
?s type music
?s contributesTo ?y
?y type westernMovie
?x bornIn Rome
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.)
Which composer wrote scores for films and was awarded the Oscar?

Structured Query

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

Try it out

YAGO 2 spot1x

Query

<table>
<thead>
<tr>
<th>Id</th>
<th>Subject</th>
<th>Property</th>
<th>Object</th>
<th>Time</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>?id1</td>
<td>?x</td>
<td>created</td>
<td>?y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>?id2</td>
<td>?y</td>
<td>type</td>
<td>wordnet_composer_109947232</td>
<td></td>
<td></td>
</tr>
<tr>
<td>?id3</td>
<td>?y</td>
<td>type</td>
<td>wordnet_movie_106613686</td>
<td></td>
<td></td>
</tr>
<tr>
<td>?id4</td>
<td>?x</td>
<td>hasWonPrize</td>
<td>Academy_Award</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://www.mpi-inf.mpg.de/yago-naga/deanna/
Who covered (songs by) European film music composers?

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

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!

Boyfriend Tracker Free
Android Aplicativos Ponto Com - July 30, 2014
Tools

Install  Add to Wishlist

⚠️ You don't have any devices

★★★★☆ (35)
Entity Linking: Privacy at Stake

Zoe
• publish & recommend

female 29y Jamame

Cry Freedom
Nive Nielsen

Social network

Internet

Search

Discussion & seek help

Synthroid tremble
Addison disorder

Levothroid shaking
Addison’s disease
Nive concert
Greenland singers
Somalia elections
Steve Biko

Female 25-30 Somalia

Online forum

Facebook

Patient.co.uk
Privacy Adversaries

Linkability Threats:

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

**Internet**

**Social network**

**Online forum**

**Search engine**
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 → 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
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
Take-Home Message

Knowledge, analytics, insight

Discover New Knowledge
Common-sense

Temporal Knowledge

Web Contents
Knowledge

Semantic Search
Deep Text Analytics
Privacy & Trust

knowledge acquisition
intelligent interpretation

more knowledge, analytics, insight