



Constructing and Mining Web-scale Knowledge Graphs

Antoine Bordes (Facebook)
abordes@fb.com

Evgeniy Gabrilovich (Google)
gabr@google.com

The opinions expressed herein are the sole responsibility of the tutorial instructors and do not necessarily reflect the opinion of Facebook Inc. or Google Inc.

Technologies described might or might not be in actual use.

Acknowledgements

Thanks to Luna Dong, Matthew Gardner, Ni Lao, Kevin Murphy, Nicolas Usunier, and Jason Weston for fruitful discussions that helped improve this tutorial.

Special thanks to Philip Bohannon and Rahul Gupta for letting us use their slides on entity deduplication and relation extraction.

Outline of the tutorial

PART 1: Knowledge graphs

1. Applications of knowledge graphs
2. Freebase as an example of a large scale knowledge repository
3. Research challenges
4. Knowledge acquisition from text

PART 2: Methods and techniques

1. Relation extraction
2. Entity resolution
3. Link prediction

PART 1: KNOWLEDGE GRAPHS

The role of knowledge

- “Knowledge is Power” Hypothesis (the Knowledge Principle): “If a program is to perform a complex task well, **it must know a great deal about the world** in which it operates.”
- The Breadth Hypothesis: “To behave intelligently in unexpected situations, an agent must be capable of falling back on **increasingly general knowledge**.”



*Lenat & Feigenbaum
Artificial Intelligence 47 (1991)
“On the Threshold of Knowledge”*



Why (knowledge) graphs?

- We're surrounded by **entities**, which are connected by **relations**
- We need to store them somehow, e.g., using a **DB** or a **graph**
- **Graphs** can be processed **efficiently** and offer a convenient **abstraction**

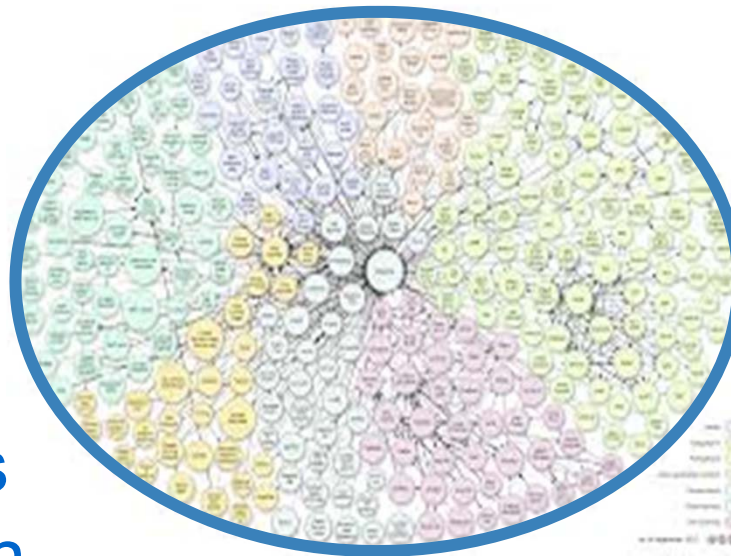
Knowledge graphs

 Freebase

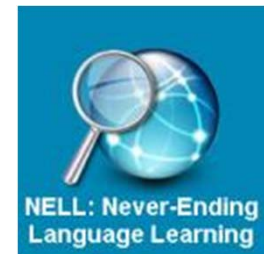

yago
select knowledge


DBpedia

Facebook's
Entity Graph



Microsoft's
Satori



*OpenIE
(Reverb, OLLIE)*

Google's
Knowledge Graph

A sampler of research problems

- **Growth:** knowledge graphs are incomplete!
 - *Link prediction:* add relations
 - *Ontology matching:* connect graphs
 - *Knowledge extraction:* extract new entities and relations from web/text
- **Validation:** knowledge graphs are not always correct!
 - *Entity resolution:* merge duplicate entities, split wrongly merged ones
 - *Error detection:* remove false assertions
- **Interface:** how to make it easier to access knowledge?
 - *Semantic parsing:* interpret the meaning of queries
 - *Question answering:* compute answers using the knowledge graph
- **Intelligence:** can AI emerge from knowledge graphs?
 - *Automatic reasoning* and planning
 - Generalization and abstraction

A sampler of research problems

- **Growth:** knowledge graphs are incomplete!
 - *Link prediction:* add relations
 - *Ontology matching:* connect graphs
 - *Knowledge extraction:* extract new entities and relations from web/text
- **Validation:** knowledge graphs are not always correct!
 - *Entity resolution:* merge duplicate entities, split wrongly merged ones
 - *Error detection:* remove false assertions
- **Interface:** how to make it easier to access knowledge?
 - *Semantic parsing:* interpret the meaning of queries
 - *Question answering:* compute answers using the knowledge graph
- **Intelligence:** can AI emerge from knowledge graphs?
 - *Automatic reasoning* and planning
 - Generalization and abstraction

Connections to related fields

- Information retrieval
- Natural language processing
- Databases
- Machine learning
- Artificial intelligence


A SAMPLER OF APPLICATIONS OF KNOWLEDGE GRAPHS

Surfacing structured results in web search











Augmenting the presentation
with relevant facts

The screenshot shows a Google search for "New York". The search bar at the top contains "New York" and the Google logo. Below the search bar, there are tabs for "Web", "Images", "News", "Maps", "Videos", and "More". The search results are displayed on the left side of the page, and a structured knowledge panel is on the right.

Search Results:

- New York - Wikipedia, the free encyclopedia**
en.wikipedia.org/wiki/New_York
New York is a state in the Northeastern and Mid-Atlantic regions of the United States. New York is the 27th-most extensive, the third-most populous, and the ...
New York City - Albany - List of cities in New York - New York metropolitan area
- New York City - Wikipedia, the free encyclopedia**
en.wikipedia.org/wiki/New_York_City
For other uses, see NYC (disambiguation) and New York, New ...
Neighborhoods - History of New York City - Nicknames - Borough
- The Official New York City Guide to NYC Attractions, Dining ...**
www.nycgo.com/
Visit NYCgo for official NYC information on travel, hotels, deals and offers like Restaurant Week, and the best restaurants, shops, clubs and cultural events.
Must-See NYC - Broadway Shows & Tickets - Events - Tours and Attractions
- The New York Times - Breaking News, World News ...**
www.nytimes.com/
There were no reports of survivors on a Malaysia Airlines flight that crashed on Thursday in eastern Ukraine near the Russian border, the scene of fighting ...
Natalie Glance and one other person +1'd this
- New York Magazine -- NYC Guide to Restaurants, Fashion ...**
nymag.com/
Daily coverage of New York's restaurants, nightlife, shopping, fashion, politics, and culture. NYMag.com is the online counterpart to New York Magazine.
- News for new york**

New York, Responding to Surge of Child Migrants, Forms ...
New York Times - by Kirk Semple - 1 hour ago
Opposition to sheltering a wave of young migrants has mounted in many communities across the country, but in New York City, the reaction has ...
- More news for new york**
- NewYork.com - Your Official Site for Travelling To and Living ...**
www.newyork.com/

Structured Knowledge Panel (Right Side):

- New York**
US State
- New York is a state in the Northeastern and Mid-Atlantic regions of the United States. New York is the 27th-most extensive, the third-most populous, and the seventh-most densely populated of the 50 United States. [Wikipedia](#)
- Capital:** Albany
- Secretary of State:** Cesar A. Perales
- Minimum wage:** 8.00 USD per hour (December 31, 2013)
- Governor:** Andrew Cuomo
- Colleges and Universities:** Cornell University, [More](#)
- Destinations**
View 45+ more

New York City

Buffalo

Long Island

Albany

Finger Lakes
- Points of interest**
View 40+ more

Statue of Liberty

Niagara Falls

Adirondack Mountains

Empire State Building

Metropolitan Museum of Art

Feedback

Surfacing facts proactively

The image shows two overlapping Google search result pages. The top page is for the search query "san francisco population". A red circle highlights the search bar containing "san francisco population". A yellow arrow points from this circle to the "Population, San Francisco, CA" result. This result includes a line graph showing population growth from 1990 to 2011, with the 2011 population of 812,826 highlighted. The source is cited as the U.S. Census Bureau. A map of San Francisco is also shown. The bottom page is for the search query "san francisco". A red circle highlights the search bar containing "san francisco". A yellow arrow points from this circle to the "San Francisco" knowledge panel on the right. This panel includes a map of San Francisco and a list of facts: Area (231.9 sq miles), Founded (June 29, 1776), Weather (59°F, Wind NE at 4 mph, 46% Humidity), Local time (Sunday 3:55 PM PT), and Population (812,826 (2011)). A red arrow points from the "Population" fact in the knowledge panel to the "Population, San Francisco, CA" result on the top page.

San Francisco Population

Population, San Francisco, CA

www.google.com/publicdata

812,826 - Jul 2011

Source: U.S. Census Bureau

San Francisco

San Francisco, officially the City and County of San Francisco, is the leading financial and cultural center of Northern California and the San Francisco Bay Area.

San Francisco

San Francisco, officially the City and County of San Francisco, is the leading financial and cultural center of Northern California and the San Francisco Bay Area.

Area: 231.9 sq miles (600.6 km²)

Founded: June 29, 1776

Weather: 59°F (15°C), Wind NE at 4 mph (6 km/h), 46% Humidity

Local time: Sunday 3:55 PM PT

Population: 812,826 (2011)

Exploratory search



The screenshot shows a Google search for "new york sightseeing". The top section features a carousel of sightseeing options near New York, NY, including Gray Line New York, City Sightseeing, Circle Line, CitySights NY, NYC & Co, The New York Pass, Rockefeller Center, The Metropolitan Museum of Art, and the Empire State Building. Below the carousel, the search results list several websites and articles, such as "Gray Line New York Sightseeing Tours, Cruises & Attractions", "New York attractions: The 50 best sights and attractions in ...", "New York City Tours and Attractions - NYC Sightseeing ...", "New York: Sightseeing in NYC - TripAdvisor", "NYC Sightseeing Tour | New York City Double Decker Tou...", "City Sightseeing New York, Hop On - Hop Off Bus Tours", and "Top 25 New York City Tours - New York Magazine". A map of New York City is also visible on the right side of the search results.

Connecting people, places and things

The image shows a Facebook page for Harvard University. At the top, there's a header with the Harvard crest and the name 'Harvard University'. Below this is a large photo of a campus scene. To the left of the photo is the Harvard crest. Below the photo, it says 'Harvard University' and '2,703,624 likes · 47,280 talking about this · 430,473 were here'. There are buttons for 'Like', 'Message', and a dropdown menu. Below the main post area, there are two sidebars. The left sidebar is titled 'People who like Harvard University' and lists several people: Paul McDonald (Engineer at Facebook), Ekaterina Skorobogatova (Works at Facebook), Gary Johnson (Corporate Development at Facebook), and Greg Marra (Product Manager at Facebook). The right sidebar is titled 'People who visited Harvard University' and lists several people: Florence Trouche (Global Client Partner at Facebook), Andrew Tulloch (Machine Learning at Facebook), Joseph Barillari (Software Engineer at Facebook), and Sheryl Sandberg (Chief Operating Officer at Facebook). Two large blue arrows point from the main page towards these two sidebars, indicating the flow of information or connections.

Harvard University

2,703,624 likes · 47,280 talking about this · 430,473 were here

Like Message

People who like Harvard University

Paul McDonald
Engineer at Facebook
Likes Harvard University, iRunFar.com and 182 others
Studied Computer Science at Harvard University '03
3 mutual friends including Clodagh Chloe Takeuchi and Serkan ...

Ekaterina Skorobogatova
Works at Facebook
Likes Harvard University, Loves Company and 3,455 others
Studied Interactive Multimedia at New York University
2 mutual friends: Amina Belghiti and Alexey Spiridonov

Gary Johnson
Corporate Development at Facebook
Likes Harvard University and 278 others
Studied at Wharton School, University of Pennsylvania '08
5 mutual friends including Jen Holmstrom and Clodagh Chloe T...

Greg Marra (马格雷)
Product Manager at Facebook
Likes Harvard University, Emmy's Spaghetti Shack and 693 others
Studied Electrical and Computer Engineering at Franklin W. Olin ...
1 mutual friend: Ledell Wu

People who visited Harvard University

Florence Trouche
Global Client Partner at Facebook
Visited Harvard University, Marché Poncelet and 317 other places
Studied at Rouen Business School '90
35 mutual friends including Michelle Gilbert and Lisa Carucci

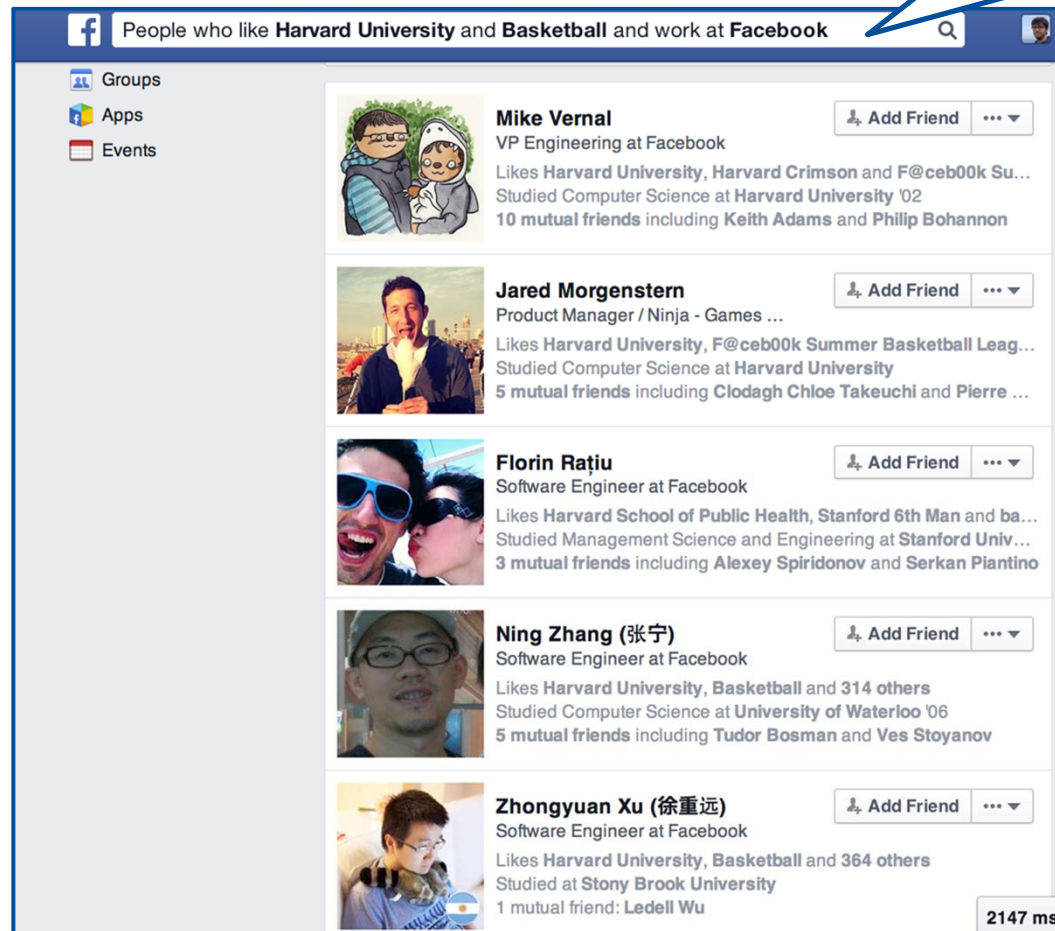
Andrew Tulloch
Machine Learning at Facebook
Visited Harvard University, City Beer Store and 90 other places
Studied Machine Learning at University of Cambridge
21 mutual friends including Jason Weston and Nicolas Vasilache

Joseph Barillari (joeb)
Software Engineer at Facebook
Visited Harvard University, Philz Coffee At Facebook and 990 others
Studied Computer Science at Harvard University '07
10 mutual friends including Tudor Bosman and Jessica Traynor

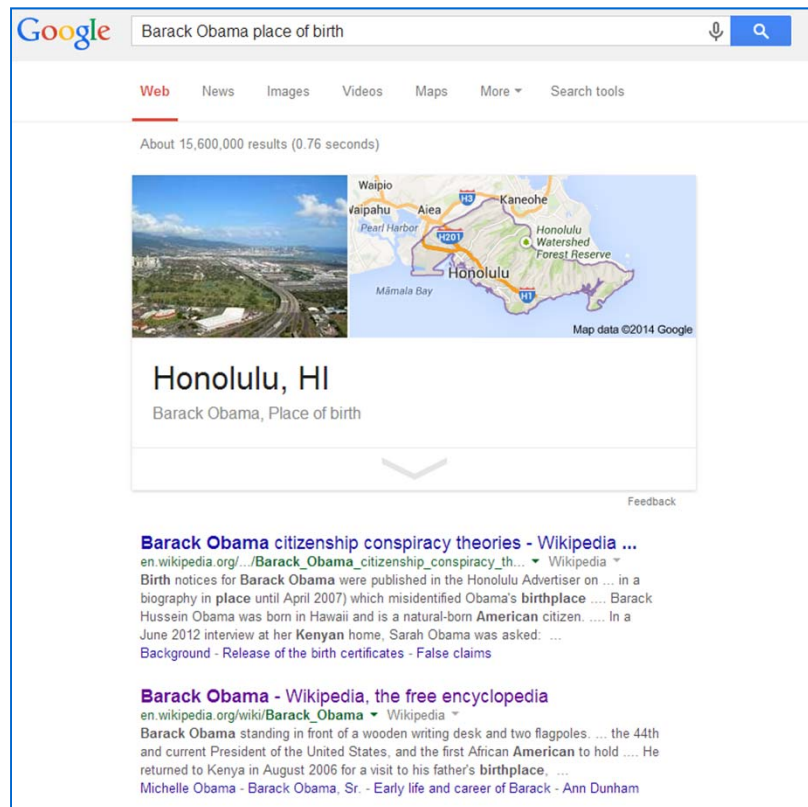
Sheryl Sandberg
Chief Operating Officer at Facebook
Visited Harvard University and 423 other places
Studied at Harvard Business School
14 mutual friends including Laurent Solly and Catalina Fries Sa...

Connecting people, places and things

Structured search within the graph



Question answering



Google

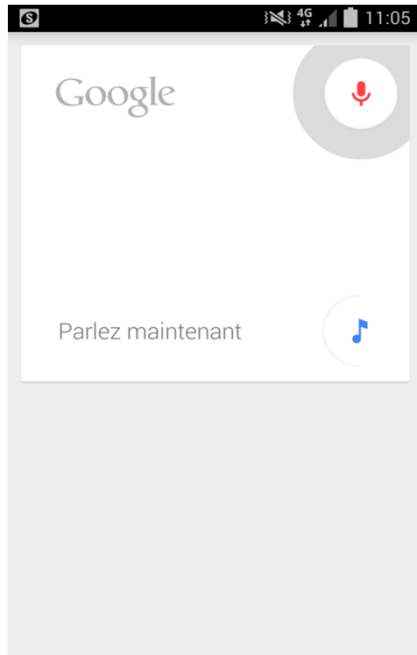


EVI
(Amazon)

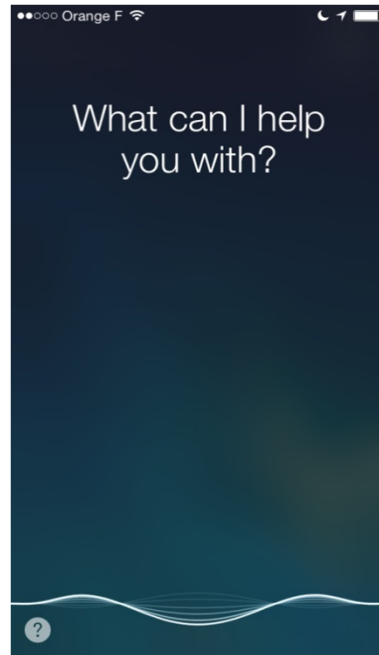


Siri
(Apple)

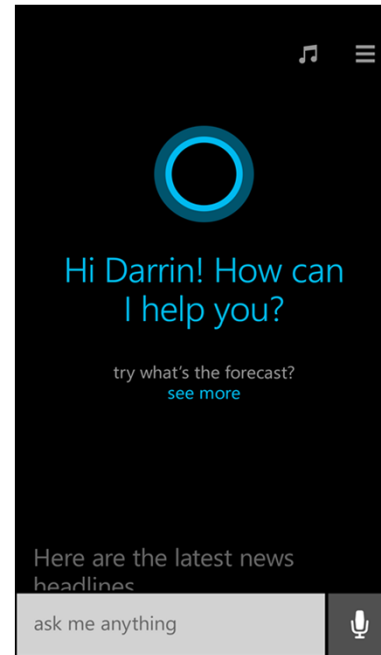
Towards a knowledge-powered digital assistant



OK Google



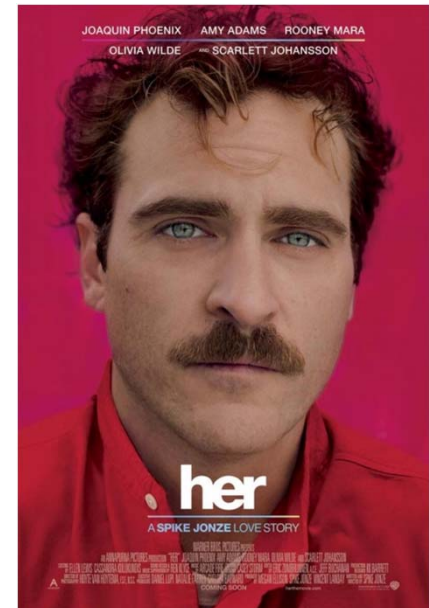
**Siri
(Apple)**



**Cortana
(Microsoft)**

- Natural way of accessing/storing knowledge
- Dialogue system
- Personalization
- Emotion

Interface revolution →



FREEBASE AS AN EXAMPLE OF A LARGE SCALE KNOWLEDGE REPOSITORY

Different approaches to knowledge representation

- Structured (e.g., Freebase or YAGO)
 - Both entities and relations come from a fixed lexicon
- Semi-structured
 - Predicates come from a fixed lexicon, but entities are strings
 - NELL used to be in this category, but is now structured (creating new entities as needed)
- Unstructured (Open IE)



- **Freebase** is an open, Creative Commons licensed repository of structured data
- **Typed entities** rather than **strings**

Person Type

Key: /people/person Includes: Topic

A person is a human being (man, woman or child) known to have actually existed. Living persons, celebrities and politicians are persons.

Table Diagram

Properties

Property	ID	Expected Type
Date of birth	/people/person/date_of_birth	/type/datetime
Place of birth	/people/person/place_of_birth	/location/location ↔
Country of nationality	/people/person/nationality	/location/country
Gender	/people/person/gender	/people/gender <i>enumerated</i>
Profession	/people/person/profession	/people/profession ↔
Religion	/people/person/religion	/religion/religion
Ethnicity	/people/person/ethnicity	/people/ethnicity ↔
Parents	/people/person/parents	/people/person ↔
Children	/people/person/children	/people/person ↔
Siblings	/people/person/sibling_s	/people/sibling_relationship
Spouse (or domestic partner)	/people/person/spouse_s	/people/marriage ↔ <i>mediated</i>
Employment history	/people/person/employment_history	/business/employment_tenu
Education	/people/person/education	/education/education ↔ <i>mediated</i>

Relations are
typed too!

The world changes, but we don't retract facts

We just add more facts!

Marriage

Mediator Type

Key: /people/marriage

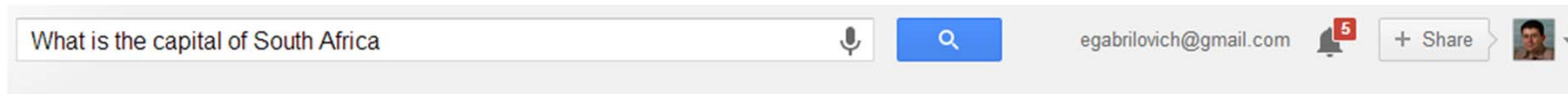
'Marriage' defines a relationship between two people. The person type uses it to store the two people in the relationship as well as a beginning and end date
[More](#)

Table	Diagram	
Properties		
Property	ID	Expected Type
Spouse	/people/marriage/spouse	/people/person ↔
From	/people/marriage/from	/type/datetime
To	/people/marriage/to	/type/datetime
Type of union	/people/marriage/type_of_union	/people/marriage_union_type ↔ <i>enumerated</i>
Location of ceremony	/people/marriage/location_of_ceremony	/location/location

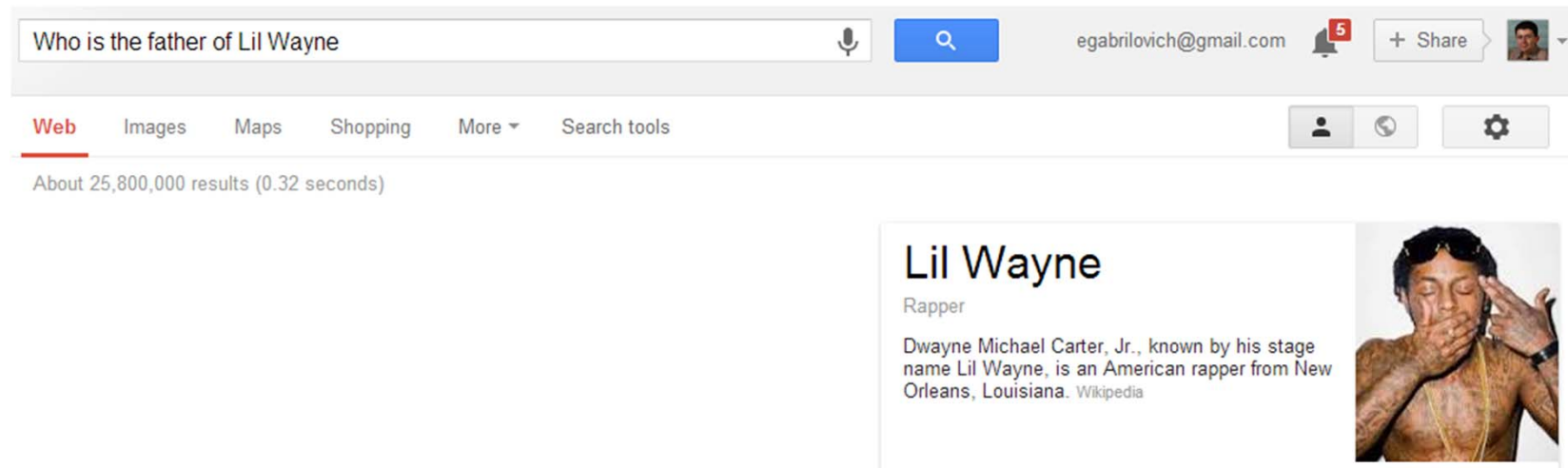
A graph of inter-related objects



Schema limitations

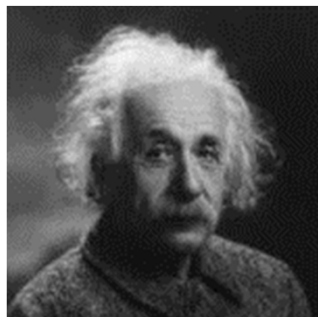


Schema limitations (cont'd)



Subject-Predicate-Object (SPO) triples

</m/0jcx, /m/04m8, /m/019xz9>



/en/albert_einstein

Albert Einstein

/en/ulm

Ulm

/people/person/place_of_birth

Place of birth

*YAGO2 uses
SPOTL tuples
(SPO + Time
and Location)*

RESEARCH CHALLENGES

Challenging research questions

- How many facts are there ? How many of them can we represent ?
- How much the boundaries of our current knowledge limit what we can learn ?
- How many facts can be potentially extracted from text ?

Limits of automatic extraction

- Freebase: **637M** (non-redundant) facts
- Knowledge Vault (automatically extracted):
302M confident facts with **Prob(true) > 0.9**
 - Of those, 223M are in Freebase (**~ 35%**)

Relations that are rarely expressed in text

Relation	% entity pairs not found	Notes
/people/person/gender	30%	Pronouns
/people/person/profession	18%	
/people/person/children and /people/person/parents	36%	
/medicine/drug_formulation/ manufactured_forms	99.9%	Sample object: "Biaxin 250 film coated tablet" (/m/0jxc5vb)
/medicine/manufactured_drug _form/available_in	99.4%	Sample subject: "Fluocinolone Acetonide 0.25 cream" (/m/0jxlbx9)
/book/author/works_written and /book/written_work/author	37%	Sample book title: "The birth day: a brief narrative of Eliza Reynolds, who died on Sunday, Oct 19, 1834" (/m/0ydpbtq)

Relations that are rarely expressed in text

Relation

/people/person/gender

/people/person/profession

/people/person/children and
/people/person/parents

/medicine/drug_formulation/
manufactured_forms

/medicine/manufactured_drug
_form/available_in

/book/author/works_written and
/book/written_work/author

Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Communications & Public Affairs

Newsroom

News Releases

Social Media Hub

Einstein in the Media

Features

Multimedia

Publications



[Home](#) > [Newsroom](#) > [News Releases](#) > [Connectomics](#)

Connectomics

Connectomics: Mapping the Neural Network Governing Male Roundworm Mating

print | subscribe

July 26, 2012 – (BRONX, NY) – In a study published today online in *Science*, researchers at [Albert Einstein College of Medicine](#) of Yeshiva University have determined the complete wiring diagram for the part of the nervous system controlling mating in the male roundworm *Caenorhabditis elegans*, an animal model intensively studied by scientists worldwide.



Scott Emmons, Ph.D.

The study represents a major contribution to the new field of connectomics – the effort to map the myriad neural connections in a brain, brain region or nervous system to find the specific nerve connections responsible for particular behaviors. A long-term goal of connectomics is to map the human “connectome” – all the nerve connections within the human brain.

Because *C. elegans* is such a tiny animal – adults are one millimeter long and consist of just 959 cells – its simple nervous system totaling 302 neurons make it one of the best animal models for understanding the millions-of-times-more-complex human brain.

The Einstein scientists solved the structure of the male worm’s neural mating circuits by developing software that they used to analyze serial electron micrographs that other scientists had taken of the region. They found that male mating requires 144 neurons – nearly half the worm’s total number – and their paper describes the connections between those 144 neurons and 64 muscles involving some 8,000 synapses. A synapse is the junction at which one neuron (nerve cell) passes an electrical or chemical signal to another neuron.

Relations that are rarely expressed in text

Relation	% entity pairs not found	Notes
/people/person/gender	30%	Pronouns
/people/person/profession	18%	
/people/person/children and /people/person/parents	36%	
/medicine/drug_formulation/ manufactured_forms	99.9%	Sample object: "Biaxin 250 film coated tablet" (/m/0jxc5vb)
/medicine/manufactured_drug _form/available_in	99.4%	Sample subject: "Fluocinolone Acetonide 0.25 cream" (/m/0jxlbx9)
/book/author/works_written and /book/written_work/author	37%	Sample book title: "The birth day: a brief narrative of Eliza Reynolds, who died on Sunday, Oct 19, 1834" (/m/0ydpbtq)

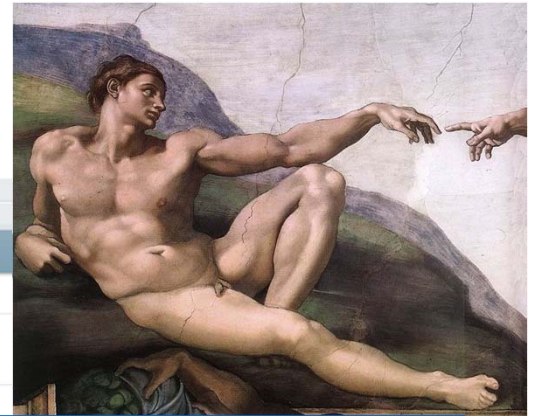
Implicitly stated information



People	/people
Person	/people/person
Date of birth	/people/person/date_of_birth
4004 BCE	
Place of birth	/people/person/place_of_birth
Garden of Eden	
Country of nationality	/people/person/nationality
This property has been flagged as having values, but those values are unknown. Remove this flag to add specific values.	
Gender	/people/person/gender
Male	
Profession	/people/person/profession
-	
Religion	/people/person/religion
-	
Ethnicity	/people/person/ethnicity
-	
Parents	/people/person/parents
This property has been flagged as having no values. Remove this flag to add new values.	
Children	/people/person/children
Children	
Azura	
Seth	
Cain	
Abel	
Awan	
Siblings	/people/person/sibling_s
Sibling	
This property has been flagged as having no values. Remove this flag to add new values.	
Spouse (or domestic partner)	/people/person/spouse
Spouse	From
Eve	
Employment history	/people/person/employment_history
Employer	

²⁰ And Adam called his wife's name Eve; because she was the mother of all living. (*Genesis 3:20*)

Implicitly stated information



People	/people
Person	/people/person
Date of birth	/people/person/date_of_birth
4004 BCE	
Place of birth	/people/person/place_of_birth
Garden of Eden	
Country of nationality	/people/person/nationality
	This property is not explicitly stated in the text.
Gender	/people/person/gender
Male	
Profession	/people/person/profession
Religion	/people/person/religion
Ethnicity	/people/person/ethnicity
Parents	/people/person/parents
Children	/people/person/children
Children	
Azura	
Seth	
Cain	
Abel	
Awan	
Siblings	/people/person/siblings
Sibling	
Spouse (or domestic partner)	/people/person/spouse
Spouse	From
Eve	
Employment history	/people/person/employment_history
Employer	

(Genesis 1)

¹ In the beginning God created the heaven and the earth.

² And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters.

³ And God said, Let there be light: and there was light.

...

(Genesis 2)

⁷ And the LORD God **formed man** of the dust of the ground, and breathed into his nostrils the breath of life; and man became a living soul.

⁸ And the LORD God **planted a garden eastward in Eden; and there he put the man whom he had formed.**

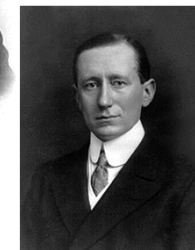
...

¹⁹ And out of the ground the LORD God formed every beast of the field, and every fowl of the air; and brought them unto **Adam** to see what he would call them: and whatsoever **Adam** called every living creature, that was the name thereof.

²⁰ And **Adam** mother of

Knowledge discovery: the long tail of challenges

- Errors in extraction (e.g., parsing errors, overly general patterns)
- Noisy / unreliable / conflicting information
- Disparity of opinion (*Who invented the radio ?*)
- Quantifying completeness of coverage



A screenshot of the Wikipedia article titled "Invention of radio". The page shows the standard Wikipedia layout with a sidebar on the left containing navigation links like "Main page", "Contents", and "Interaction". The main content area has a heading "Invention of radio" and a sub-heading "From Wikipedia, the free encyclopedia". Below this, there is a paragraph of text and a "Contents" table of contents. The table of contents lists sections such as "1 Wireless signalling methods" and "2 History of the invention of radio", with further sub-sections like "2.1 Theory of electromagnetism" and "2.1.1 Maxwell and the theoretical prediction of electromagnetic waves".

Knowledge discovery: the long tail of challenges

- Errors in extraction (e.g., parsing errors, overly general patterns)
- Noisy / unreliable / conflicting information
- Disparity of opinion (*Who invented the radio ?*)
- Quantifying completeness of coverage

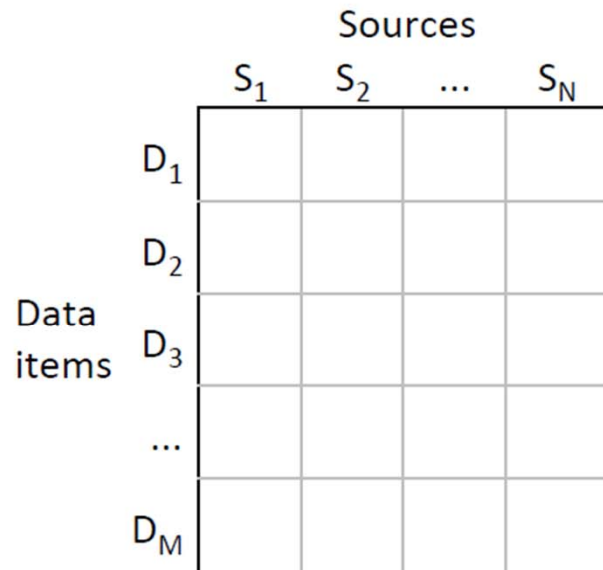


- Fictional contexts
 - `</en/abraham_lincoln,`
`/people/person/profession,`
`/en/vampire_hunter> ?`

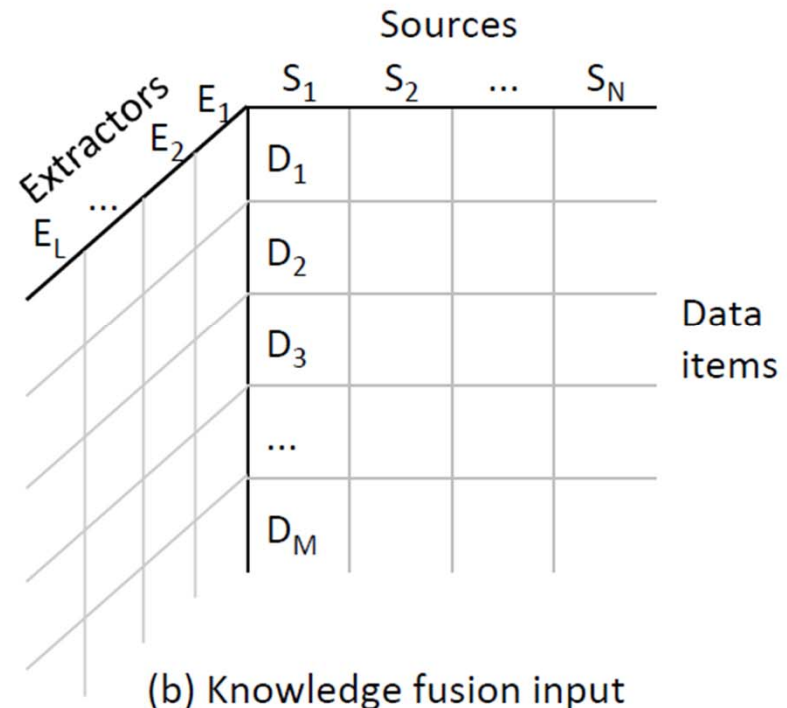


- Outright spam

Data fusion vs. knowledge fusion



(a) Data fusion input



(b) Knowledge fusion input

[Dong et al., VLDB '14]

Should we trust all sources equally ?

WIKIPEDIA The Free Encyclopedia

Article Talk Read View source View history Search

Barack Obama

From Wikipedia, the free encyclopedia

"Obama" redirects here. For other uses, see *Obama* (disambiguation).
This article is about the 44th president of the United States. For his father, see *Barack Obama, Sr.*

Barack Hussein Obama II (/ˈbəˈrɑːkˈhuːsɛnˈoʊbɑːmə/; born August 4, 1961) is the 44th and current President of the United States, the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was president of the *Harvard Law Review*. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004. He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.

In 2004, Obama received national attention during his campaign to represent Illinois in the United States Senate with his victory in the March Democratic Party primary, his keynote address at the Democratic National Convention in July, and his election to the Senate in November. He began his presidential campaign in 2007, and in 2008, after a close primary campaign against Hillary Rodham Clinton, he won sufficient delegates in the Democratic Party primaries to receive the presidential nomination. He then defeated Republican nominee John McCain in the general election, and was inaugurated as president on January 20, 2009. Nine months after his election, Obama was named the 2009 Nobel Peace Prize laureate.

During his first two years in office, Obama signed into law economic stimulus legislation in response to the Great Recession in the form of the American Recovery and Reinvestment Act of 2009 and the Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010. Other major domestic initiatives in his first term include the Patient Protection and Affordable Care Act, often referred to as "Obamacare", the Dodd–Frank Wall Street Reform and Consumer Protection Act; and the Don't Ask, Don't Tell Repeal Act of 2010. In foreign policy, Obama ended U.S. military involvement in the Iraq War, increased U.S. troop levels in Afghanistan, signed the New START arms control treaty with Russia, ordered U.S. military involvement in Libya, and ordered the military operation that resulted in the death of Osama bin Laden. He later became the first sitting U.S. president to publicly support same-sex marriage. In November 2010, the Republicans regained control of the House of

Barack Obama

44th President of the United States
Incumbent

Assumed office
January 20, 2009

Vice President Joe Biden

Preceded by George W. Bush

United States Senator from Illinois

In office
January 3, 2005 – November 16, 2008

Preceded by Peter Fitzgerald

Succeeded by Roland Burris

Member of the Illinois Senate from the 13th District

In office
January 8, 1997 – November 4, 2004

Preceded by Alice Palmer

Succeeded by Kwame Raoul

Personal details

Born Barack Hussein Obama II
August 4, 1961 (age 52)
Honolulu, Hawaii, U.S.

Political party Democratic

The Western Center For Journalism
Informing And Empowering Americans Who Love Freedom

Home Categories Blogging Tools About Polls and Petitions Contact Us Write

You are here: Home / Featured Stories / Proof Obama Born in Kenya? Obama Literary Agent Says Yes

Proof Obama Born in Kenya? Obama Literary Agent Says Yes

MAY 17, 2012 BY FLOYD BROWN 100 COMMENTS

Bretbart.com has introduced some explosive evidence showing that Obama claimed he was born in Kenya years before he became a presidential candidate. Interestingly, the editors of Bretbart still think that now Obama is telling the truth.



ALEX JONES' INFOWARS.COM BECAUSE THERE IS A WAR ON

Home Alex Jones Radio Show News Multimedia Forum News Contact Top Stories Books

Evidence Obama Born In Kenya Goes Beyond 1991 Brochure

Establishment media pulls stunt in effort to diffuse 'birther' controversy

Paul Joseph Watson

InfoWars.com

Friday, May 18, 2012

The establishment media hastily seized on yesterday's explosive story about a literary publication listing Barack Obama's birthplace as Kenya in an effort to claim that the 1991 brochure was the "origin" of the entire 'birther' issue. In reality, evidence that Obama was born in the African country is abundant.

A literary agent's promotional text for a 1991 brochure released yesterday

THE BLAZE

STORIES THEBLAZE TV RADIO MAGAZINE BLOG

HOT TOPICS: Obamacare | Ted Cruz | NSA | Education | TheBlaze TV | #2A

YAHOO! NEWS SAYS OBAMA WAS BORN IN...KENYA!

Jun. 22, 2013 12:34pm | Madeleine Morgenstern

Related: Barack Obama, Birthers, Obama Birth Certificate

Yahoo! News had to issue a correction Friday after publishing an article about President Barack Obama that called Kenya "the country of his birth."

The article, about Obama's upcoming trip to Africa, stated:

President Barack Obama makes the first extended trip to Africa of his presidency next week — but he won't be stopping at the country of his birth.

White House doesn't have 'figure on costs' of Africa trip

By Rachel Rose Hartman, Yahoo! News | The Ticket — 1 hr 40 mins ago

Email Share 40 Tweet 20 LinkedIn Share 0 Print

President Barack Obama makes the first extended trip to Africa of his presidency next week—but he won't be stopping in the country of his birth.

derich hastily claimed listing error."
al and it goes significantly
was a mistake, the listing still use a U.S. Senator. "Goderich's sixteen years, through at least four different versions of
about Obama being born in
Obama had become a Senator,
bama, "was born in Kenya."

Challenge: negative examples

- We already know a lot ... but those are only **positive** examples!
- Many ways to get negative examples ... none of them perfect ☹
 - Deleted assertions in Freebase
 - *Was the deletion justified ?* **Released ! See goo.gl/MJb3A**
 - Inconsistencies identified with manually specified rules
 - *Poor coverage*
 - Examples judged by humans **Released ! See goo.gl/MJb3A**
 - *Optimized for accuracy on the positive class*
 - Automatically create negative examples using the closed world assumption
 - *Noisy, unless applied to functional relations*
 - Feedback from Web users **Crowdsourcing**
 - *Difficult to judge automatically*

Negative examples (cont'd): feedback from Web users

Google Leonard Cohen

SIGN IN

[Leonard Cohen Home | The Official Leonard Cohen Site](#)
[www.leonardcohen.com/](#)
Official Leonard Cohen website featuring Leonard Cohen news, music, videos, album info, tour dates, and more.
[Tour](#) - [Albums](#) - [Songs From The Road \(EPK\)](#) - [News](#)

[Leonard Cohen - Wikipedia, the free encyclopedia](#)
[en.wikipedia.org/wiki/Leonard_Cohen](#)
Leonard Norman Cohen, CC GOQ (born 21 September 1934) is a Canadian Juno Award-winning singer-songwriter, musician, poet, and novelist. His work often ...
[Discography](#) - [Songs of Leonard Cohen](#) - [Hallelujah](#) - [Songs of Love and Hate](#)

[Leonard Cohen - YouTube](#)
[www.youtube.com/artist/leonard-cohen](#)
One of the most fascinating and enigmatic -- if not the most successful -- singer/songwriters of the late '60s, **Leonard Cohen** has retained an audience across...

[Leonard Cohen - Hallelujah - YouTube](#)
[www.youtube.com/watch?v=YrLk4vdY28Q](#)
Oct 3, 2009 - Uploaded by LeonardCohenVEVO
Music video by **Leonard Cohen** performing Hallelujah. (C) 2009 Sony Music Entertainment.
1,627 people +1'd this

[Leonard Cohen - Free listening, concerts, stats, & pictures at Last.fm](#)
[www.last.fm/music/Leonard+Cohen](#)
Watch videos & listen free to **Leonard Cohen**: Suzanne, So Long, Marianne & more, plus 155 pictures. **Leonard Cohen** (b. 21st September 1934 in Montréal, ...
▶ 0:30 Suzanne Songs of Leonard Cohen
▶ 0:30 Famous Blue Raincoat Songs of Love and Hate
▶ 0:30 So Long, Marianne Songs of Leonard Cohen
▶ 0:30 Hallelujah The Essential Bob Dylan

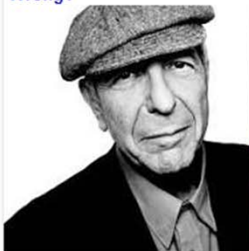
[Leonard Cohen | Music Biography, Credits and Discography | AllMusic](#)
[www.allmusic.com/artist/leonard-cohen-mn0000071209](#)
Find **Leonard Cohen** bio, songs, credits, awards, similar artists and video information on AllMusic - Cerebral yet sensual Canadian poet, novelist, and ...

[My Night With Leonard Cohen - NYTimes.com](#)
[www.nytimes.com/2013/07/18/.../my-night-with-leonard-cohen.html](#)
Jul 18, 2013 - An adventure starts with a concert and leaves a feminist wowed.


[Leonard Cohen : NPR](#)
[www.npr.org/artists/15392685/leonard-cohen](#)

Click any fact to locate it on the web. Click **Wrong?** to report a problem. You can also provide general feedback.[Cancel](#)


Wrong?




Wrong?




Wrong?




Wrong?



Wrong?



Wrong?



Wrong?

Leonard Cohen

Singer-songwriter

Leonard Norman Cohen, CC GOQ is a Canadian Juno Award-winning singer-songwriter, musician, poet, and novelist. His work often explores religion, isolation, sexuality, and personal relationships. [Wikipedia](#)

Wrong?

Wrong? Born: September 21, 1934 (age 79), Westmount, Canada





Thanks!

What's wrong with this? (optional)

Provide a URL reference with supporting evidence. (optional)

Cancel Submit

port



More Images

ino Award-winning work often explores hips. [Wikipedia](#)

Canada

ada

KDD 2014 Tutorial on Constructing and Mining Web-scale Knowledge Graphs, New York, August 24, 2014

41



[Ipeirotis & Gabrilovich, WWW 2014]

Correct Answers: 33/67 Correct (%): 49%

What is a symptom of Morgellons

Red eye

Choreoathetosis

Skin lesion

Insomnia

I don't know

Question 1 out of 10

How do you translate Dance in Russian?

Your answer:

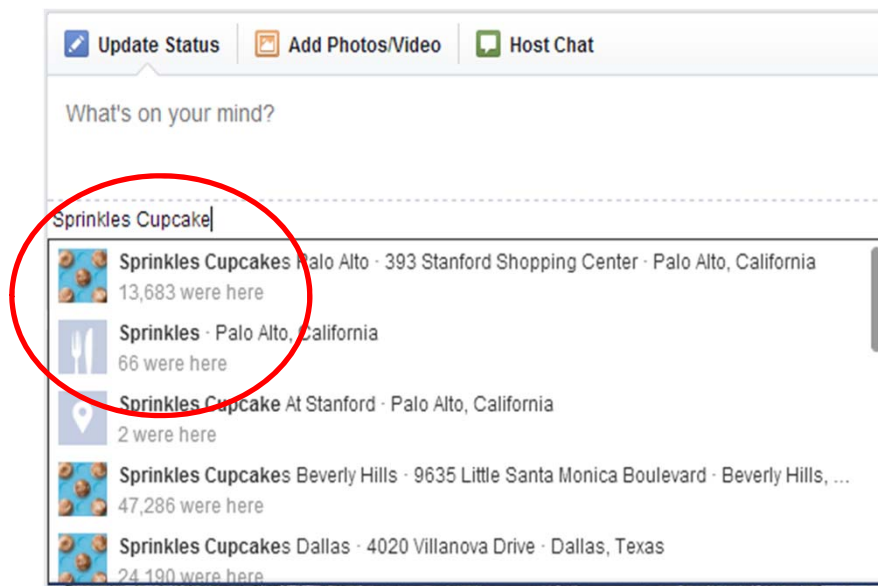
Send

I don't know

Question 1 out of 10

Entity resolution / deduplication

- Multiple mentions of the same entity is wrong and confusing.



Entity resolution / deduplication

- Multiple mentions of the same entity is wrong and confusing.

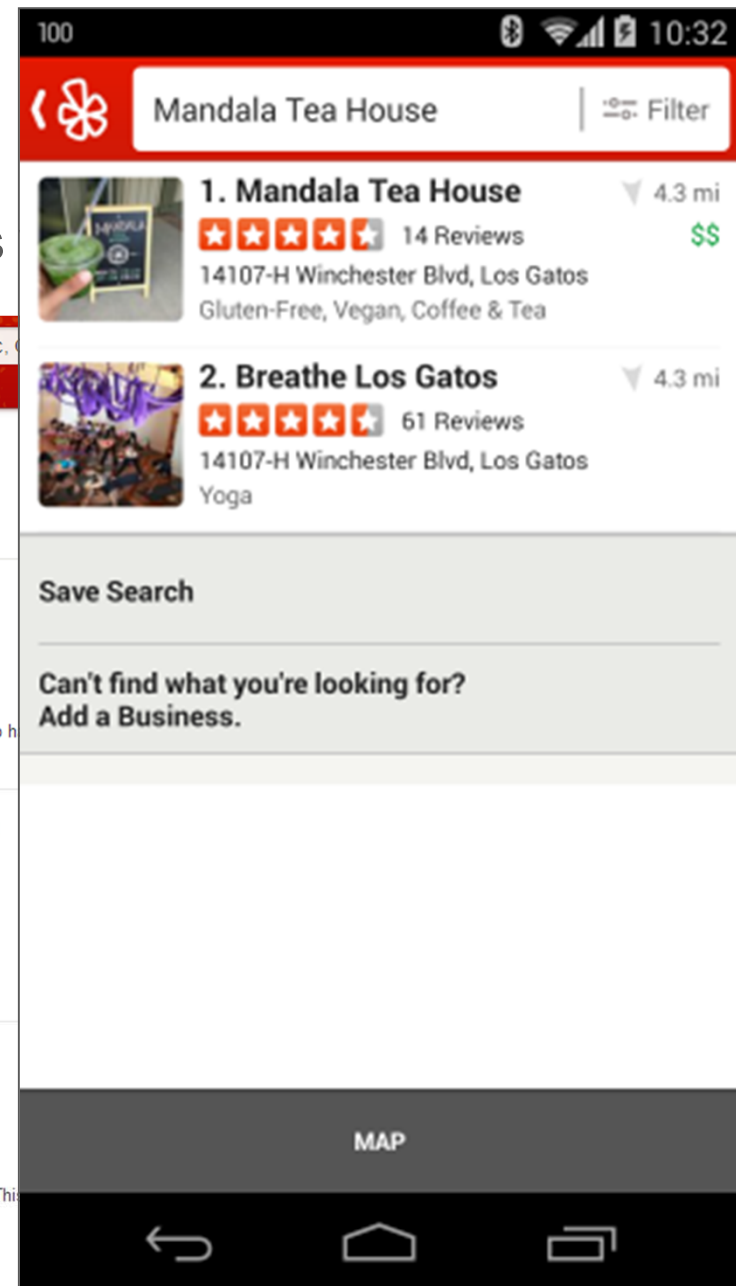
The screenshot shows a Yelp search for 'vego' in Montreal, QC, Canada. The results list three restaurants, with the first and third results circled in red to illustrate entity resolution.

Rank	Restaurant Name	Address	Phone	Reviews	Cuisine
1.	Restaurant Végo	1720 Saint-Denis Rue Montreal, QC H2X 3K6 Canada	(514) 845-2627	16 reviews	Vegetarian
2.	Resto Végo McGill	1204 Avenue McGill College Montréal, QC H3B 4J8 Canada	(514) 871-1480	5 reviews	Buffets, Vegan, Vegetarian
3.	Resto Végo St-Denis	1720 Rue Saint-denis Montreal, QC H2X 3K6 Canada	(514) 845-2627	3 reviews	Vegetarian

The map on the right shows the locations of these restaurants in Montreal, with red pins indicating their positions. The first and third results are circled in red, highlighting the issue of multiple mentions of the same entity.

Entity resolution / deduplication

- Multiple mentions of the same entity is



Commonsense knowledge

“Bananas are yellow.”



“Jasmine flowers smell good.”

“Balls bounce.”

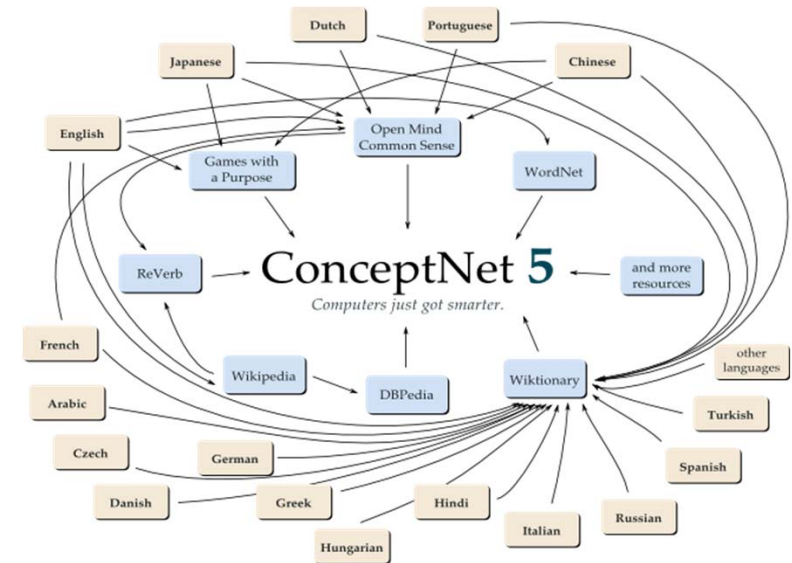


- Commonsense information is hard to collect (*too obvious*)
- Yet commonsense reasoning is often crucial

Commonsense knowledge

ConceptNet

- Nodes represent concepts (words or short NL phrases)
- Labeled relationships connecting them
saxophone → UsedFor → jazz
learn → MotivatedByGoal → knowledge



ConceptNet 5 About Wiki Downloads Search for a concept...

banana [Get /c/en/banana in JSON format](#)

banana — *IsA* → fruit
A banana is a fruit.

banana — *HasProperty* → yellow
banana is yellow.

ConceptNet 5 About Wiki Downloads Search for a concept... English Search

ball [Get /c/en/ball in JSON format](#)

ball — *CapableOf* → bounce
An activity a ball can do is bounce

ball — *CapableOf* → roll down hill
A ball can roll down hill

ball — *HasProperty* → round
A ball is round

ball — *IsA* → toy
a ball is a toy

ConceptNet (cont'd)

- ConceptNet is a **(hyper)graph**
 - Edges about the edges
- Each statement has **justifications**
 - Provenance + reliability assessment
- The graph is **ID-less**
 - Every node has all the information necessary to identify it
 - Multiple branches can be developed in parallel and later merged
 - Take the union of the nodes and edges
 - No reconciliation

[Havasi et al., RANLP '07; Speer and Havasi, LREC '12]

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

Commonsense knowledge in YAGO

- WebChild [Tandon et al., WSDM '14]
`(strawberry, hasTaste, sweet), (apple, hasColor, green)`
- Acquired from the web using semi-supervised learning
- Uses WordNet senses and web statistics to construct seeds
- Acquiring comparative commonsense knowledge from the web [Tandon et al., AAAI '14]
`(car, faster, bike), (lemon, more-sour, apple)`
 - Uses Open IE
- Earlier work: [Tandon et al., AAAI '11]
`CapableOf(dog, bark), PartOf(roof, house)`
 - Uses web n-gram data with seeds from ConceptNet

CYC

[Guha et al., CACM '90] + <http://www.cyc.com/publications>

- OpenCYC
 - 239K terms, 2M triples
- ResearchCYC
 - 500K concepts, 5M assertions, 26K relations

Multiple modalities

It should not be thought from above that Soviet
party line is necessarily disingenuous and insincere
on It should not be thought from above that Soviet
are party line is necessarily disingenuous and insincere
des on It should not be thought from above that Soviet
no are party line is necessarily disingenuous and insincere
it des on It should not be thought from above that Soviet
has are party line is necessarily disingenuous and insincere
gre it des on part of all those who put it forward many of them
ind ha no are too ignorant of outside world and mentally too
origv it dependent to question (*) self-hypnotism, and who have
thi in ha no difficulty making themselves believe what they find
pod cor gre it comforting and convenient to believe. Finally we
ind th in have the unenviable mystery as to who, if anyone, in this
tre po cor gre it actually receives accurate and unbiased
lea in th information about outside world. In atmosphere of
fu re po oriental secretiveness and conspiracy which pervades
is le in this government, possibilities for distorting or
fu re poisoning sources and currents of information are
is le infinite. The very disrespect of Russians for objective
fu re truth--indeed, their disbelief in its existence--
le in leads them to view all stated facts as instruments for
furth furtherance of one ulterior purpose or another. There
is good reason to suspect that this government is
any mil field of international law. This publication has been written with the expectation that the
the any military stories making use of it will be provided with a basic understanding of the
the any legal system governing the international community. International law is an area of
the any jurisprudence which challenges. It quite often fails to provide concise "textbook
the any answers" to problems which reach a degree of complexity far greater than that found in
evid any other legal system. Entrusted with the task of regulating the conduct of interna-
tional sovereign entities, it is a legal framework which develops on a daily basis. Its suc-
cesses go largely unrecorded, while its failures gain adverse international notoriety and
affo loss condemnation. It is a jurisdictional system particularly unsuited for complacent per-
sonalities and regressed minds. Hopefully, military stories will not view the often
only evident imposition of international law as a final weakness but as an opportunity
them afforded its practitioners to develop an efficient and viable legal system. Constructive
criticism and the ability to apply concepts and rules to practical international legal prob-
lems must be based on a working knowledge of the subject matter. The achievement of
this end underlies the purpose of this publication.

Text



Video

How to jointly acquire
knowledge from all
these sources?



Images



Speech/sounds



Artificial worlds?

Natural interfaces to knowledge

“Where is New York City?”

The image displays three different interfaces for answering the question "Where is New York City?".

Left Interface (Google Search): A screenshot of a Google search results page for the query "where is new york city". It shows a map of New York City, including Manhattan, the Bronx, and the surrounding areas. Below the map, there are links to "Images correspondant à where is new york city" and "Plus d'images pour where is new york city".

Middle Interface (Mobile App): A screenshot of a mobile application interface. At the top, it says "Where is New York City" with a "tap to edit" option. Below this, it provides a text answer: "New York City is in New York state, 135 miles (217km) south of Albany, 8.8 miles (14km) east of Newark, New Jersey." A map of the Northeast United States is shown below the text, with a blue pin indicating New York City. At the bottom, there are buttons for "Good answer" and "Bad answer", and a microphone icon for voice search.

Right Interface (Mobile App): A screenshot of another mobile application interface. It shows the same question "Where is New York City" with a "tap to edit" option. Below the question, there is a large, dark, blurred area, possibly representing a video or a large image. At the bottom, there is a microphone icon for voice search.

Natural interfaces to knowledge

“Where did Kobe Bryant play in high school?”

Google search results for "where did kobe bryant play in high school". The search bar shows the query and a microphone icon. Below the search bar are tabs for Web, News, Shopping, Videos, Images, More, and Search tools. The results show "About 1,750,000 results (0.93 seconds)".

Born on August 23, **1978 in Philadelphia, Pennsylvania**, Kobe Bryant played for the Charlotte Hornets right out of high school. He was soon traded to the **L.A. Lakers**, where he went on to win five championships and become one of the leading scorers of the NBA.

Kobe Bryant - Biography - Basketball Player - Biography.com
www.biography.com/people/kobe-bryant-10683945 Fyi

Kobe Bryant - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Kobe_Bryant Wikipedia

Kobe Bryant smiling on the bench USA vs GBR 2012.jpg ... He entered the NBA directly from **high school**, and has played for the Lakers his entire career, The teams agreed to the trade the day before the draft and the Lakers **did** not tell the ...
Shaq-Kobe feud - Sexual assault case - Joe Bryant - Idan Ravin

Kobe Bryant Stats, Bio, Career | Lakers Nation
www.lakersnation.com/kobe-bryant/ Lakers Nation

Kobe entered the NBA directly from **high school**, where he was selected 13th
Bryant did achieve an individual milestone by becoming the youngest **player** in ...

Kobe Bryant - Biography - Basketball Player - Biography.com
www.biography.com/people/kobe-bryant-10683945 Fyi

Born on August 23, 1978 in Philadelphia, Pennsylvania, **Kobe Bryant played** for the Charlotte Hornets right out of **high school**. He was soon traded to the L.A. ...

Siri interface showing a conversation. The user asks "Where did Kobe Bryant play in high school". Siri responds with "Maybe you want something about Kobe Bryant? Try this web page: 'Kobe Bryant - Wikipedia, the free encyclopedia'." and provides a link. The user asks "Would you like to see some more results?". Siri responds with "Going away? Try asking me for the weather in Rome".

iPhone screen showing the Siri interface. The user asks "Where did Kobe Bryant play in high school". Siri responds with "I don't know where you are. But you can show me. Go to Settings, tap Privacy, tap Location Services and turn it on. Then scroll to Siri and turn that on, too." The screen then shows the "Location Services Settings" page.

Natural interfaces to knowledge

“Where did Kobe Bryant play in high school?”

The image is a composite illustrating a search process. At the top, a Google search bar contains the query "where did kobe bryant play in high school". Below the search bar, the Google results page shows "About 1,750,000 results (0.93 seconds)". The main content area displays the Wikipedia article for "Kobe Bryant". The article text is as follows:

Kobe Bean Bryant (born August 23, 1978) is an American professional [basketball](#) player for the [Los Angeles Lakers](#) of the [National Basketball Association](#) (NBA). He entered the NBA directly from high school, and has played for the Lakers his entire career, winning five [NBA championships](#). Bryant is a 16-time [All-Star](#), 15-time member of the [All-NBA Team](#), and 12-time member of the [All-Defensive team](#). As of March 2013, he ranks third and fourth^[3] on the league's [all-time postseason scoring](#) and [all-time regular season scoring lists](#), respectively.

Bryant enjoyed a successful high school basketball career at [Lower Merion High School](#) in [Pennsylvania](#), where he was recognized as the top high school basketball player in the country. He declared his eligibility for the [NBA Draft](#) upon graduation, and was selected with the 13th overall pick in the [1996 NBA Draft](#) by the [Charlotte Hornets](#), then traded to the Los Angeles Lakers. As a rookie, Bryant earned himself a reputation as a high-flyer and a fan favorite by winning the [1997 Slam Dunk Contest](#).

On the right side of the Wikipedia article, there is a photo of Kobe Bryant with the caption "Bryant at the 2012 Summer Olympics in London" and "No. 24 – Los Angeles Lakers".

Overlaid on the right side of the image is a dark blue box containing the text "Where did Kobe Bryant play in high school" and "tap to edit".

KNOWLEDGE ACQUISITION FROM TEXT

External sources of knowledge

- Text
 - Unstructured (NL text) or semi-structured (tables or pages with regular structure)
 - Relevant tasks: **entity linking, relation extraction**
- Structured knowledge bases (e.g., IMDB)
 - Relevant task: **entity resolution**

Possible approaches to knowledge acquisition from the Web

- **Unfocused**

- Start from a collection of Web pages

➔ Non-targeted (blanket) extraction

- **Focused**

- Formulate specific questions or queries, looking for missing data

- Identify (a small set of) relevant Web pages

➔ Targeted extraction

Open IE – extracting **unstructured** facts from **unstructured** sources (text)

- **TextRunner** [Banko et al., IJCAI '07], **WOE** [Wu & Weld, ACL '10]
- Limitations
 1. Incoherent extractions – the system makes independent decisions whether to include each word in the relation phrase, possibly gluing together unrelated words
 2. Uninformative extractions – those omitting critical information (e.g., “has” instead of “has a population of” or “has a Ph.D. in”)
- **ReVerb** [Fader et al., EMNLP '11] solves these problems by adding syntactic constraints
 - Every multi-word relation phrase must begin with a verb, end with a preposition and be a contiguous sequence of words)
 - Relation phrases should not omit nouns
 - Minimal number of distinct argument pairs in a large corpus

OLLIE: Open Language Learning for Information Extraction

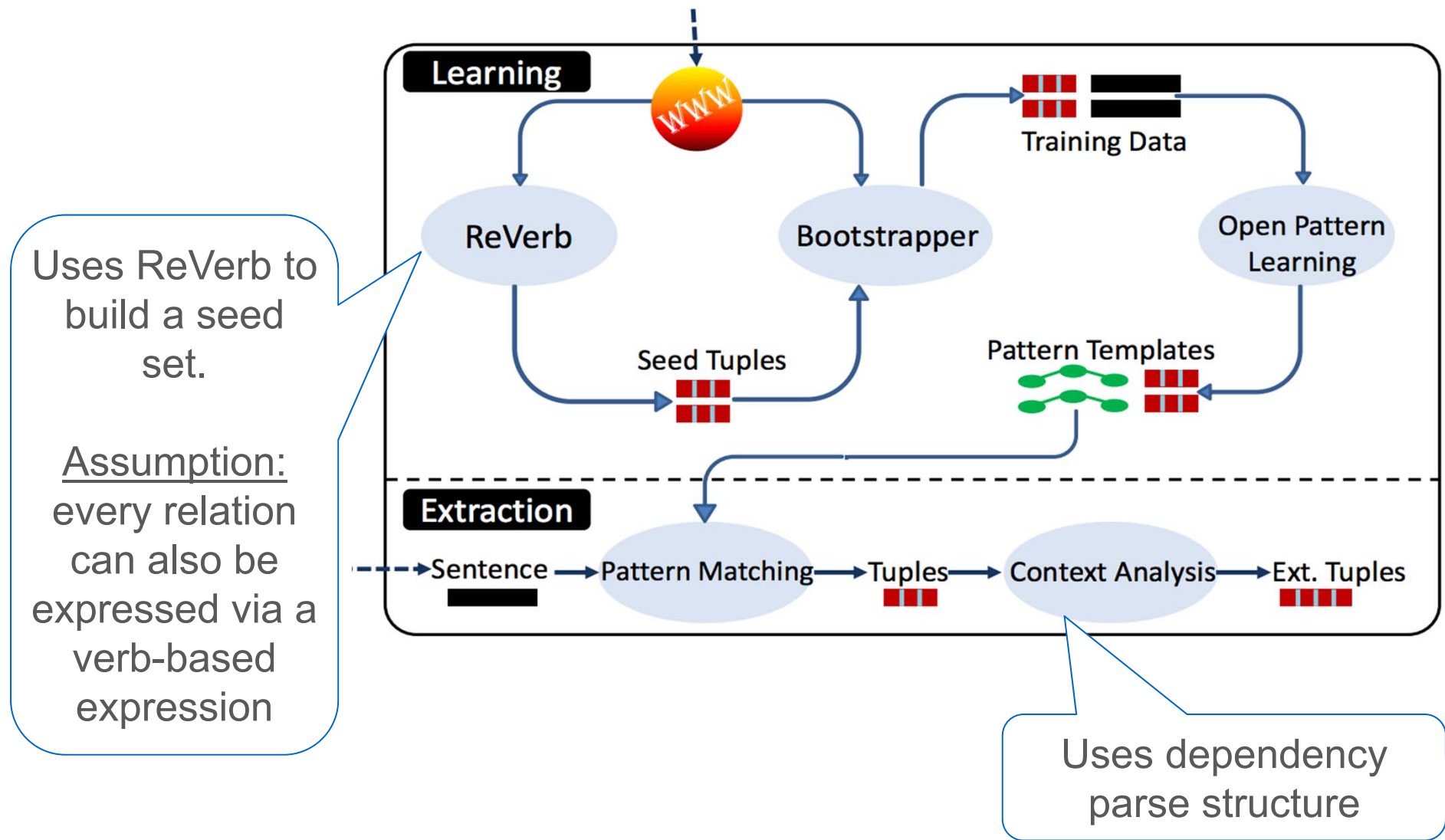
[Mausam et al., EMNLP '12]

Limitations of ReVerb

- Only extracts relations mediated by verbs
- Ignores context, potentially extracting facts that are not asserted

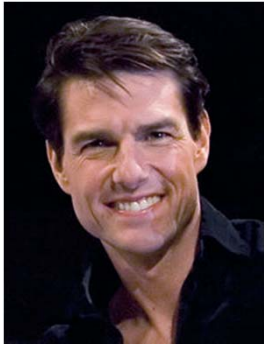
- | |
|---|
| 1. "After winning the Superbowl, the Saints are now the top dogs of the NFL."
O: (the Saints; win; the Superbowl) |
| 2. "There are plenty of taxis available at Bali airport."
O: (taxis; be available at; Bali airport) |
| 3. "Microsoft co-founder Bill Gates spoke at ..."
O: (Bill Gates; be co-founder of; Microsoft) |
| 4. "Early astronomers believed that the earth is the center of the universe."
R: (the earth; be the center of; the universe)
W: (the earth; be; the center of the universe)
O: ((the earth; be the center of; the universe)
<i>AttributedTo</i> believe; Early astronomers) |
| 5. "If he wins five key states, Romney will be elected President."
R,W: (Romney; will be elected; President)
O: ((Romney; will be elected; President)
<i>ClausalModifier</i> if; he wins five key states) |

OLLIE (cont'd)



Extracting **structured** facts from **unstructured** sources (text)

/en/tom_cruise



Thomas Cruise Mapother IV (born July 3, 1962),

/people/person/
date_of_birth

/common/topic/
alias

widely known as Tom Cruise,

/people/person/
nationality

is an American



/en/united_states

/people/person/
profession

film actor and producer.

/en/actor

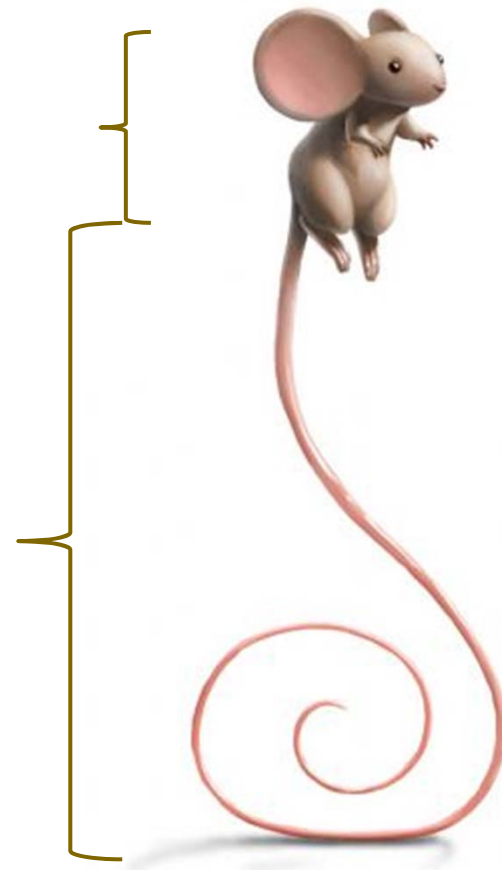
/en/film_producer

Knowledge discovery

- Relying on humans
 - Volunteer contributions at Freebase.com
 - Import of large datasets (e.g., IMDB)
 - **Head + torso**
- Automatic extraction
 - Extraction from web pages
 - **The long tail**
 - Learning patterns using known facts

“... jumped from X into Y ...”

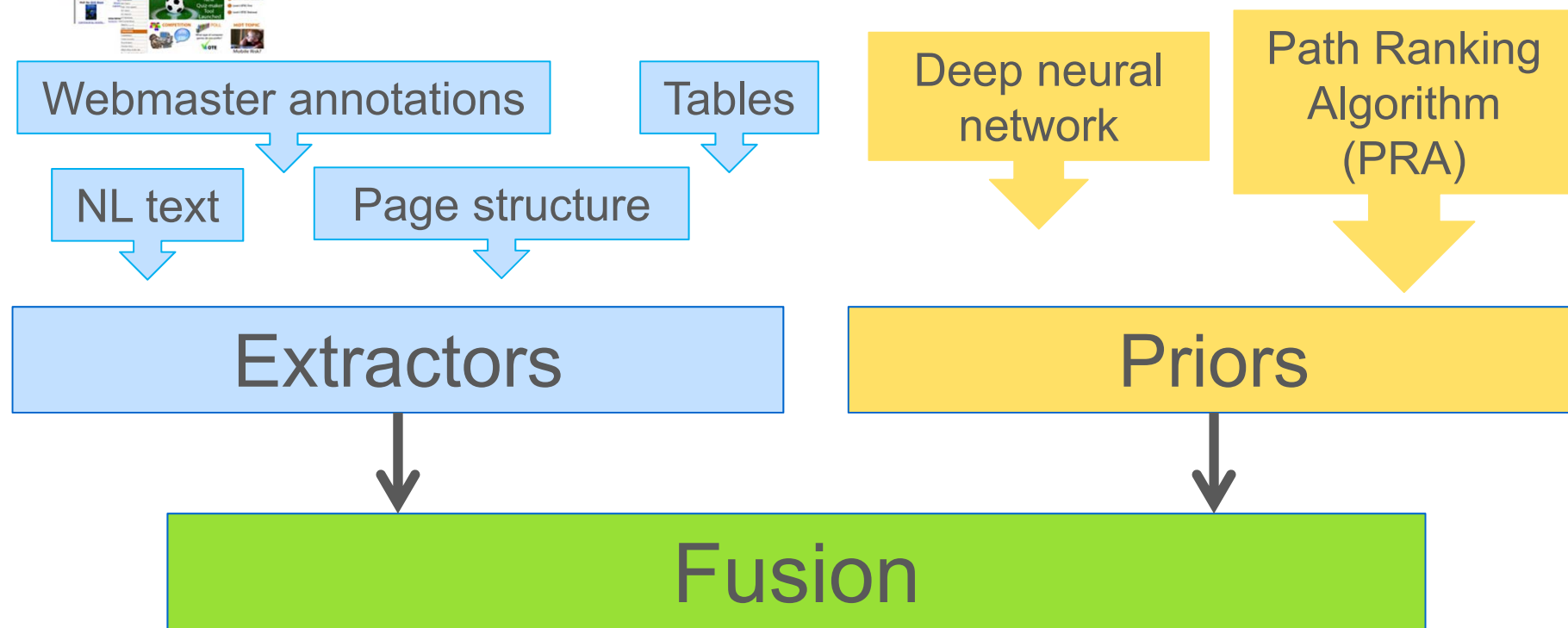
```
</en/tower_bridge,  
/transportation/bridge/body_of_water_spanned,  
/en/river_thames>
```



<http://www.flickr.com/photos/sandreli/4691045841/>



Knowledge Fusion



[Dong et al., KDD 2014]

Research 3: Statistical techniques for big data

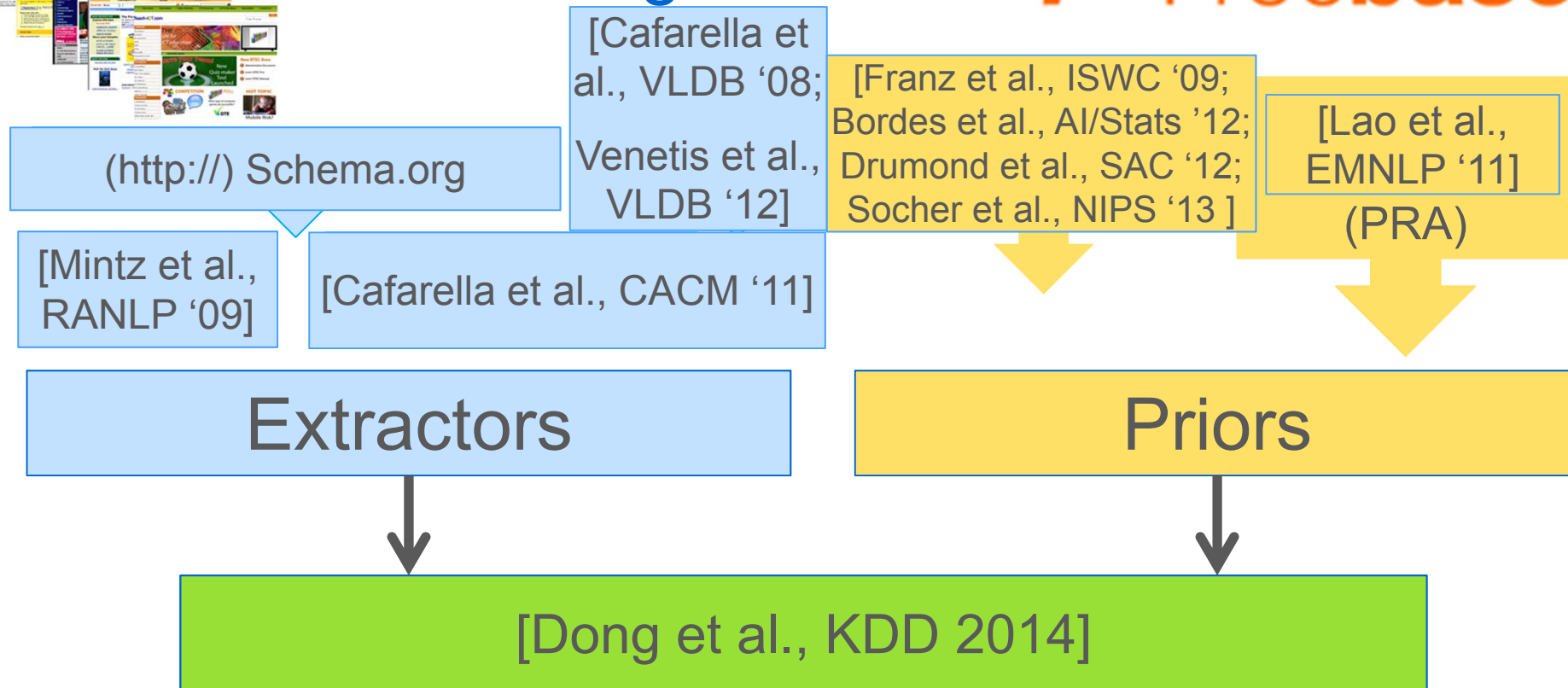
Mon, 10:30-12, Empire West



**KNOWLEDGE
VAULT**



Knowledge Fusion



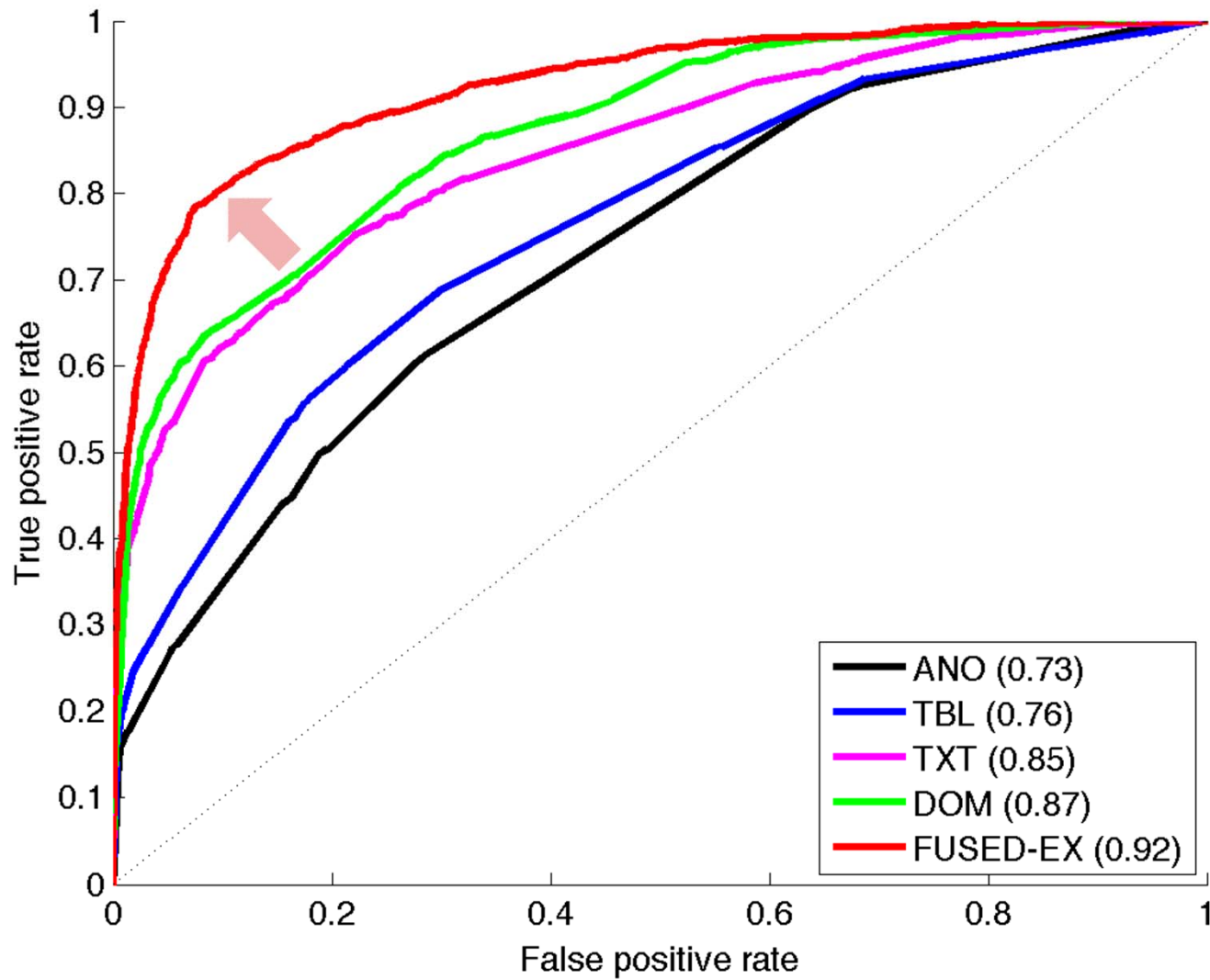
Research 3: Statistical techniques for big data

Mon, 10:30-12, Empire West

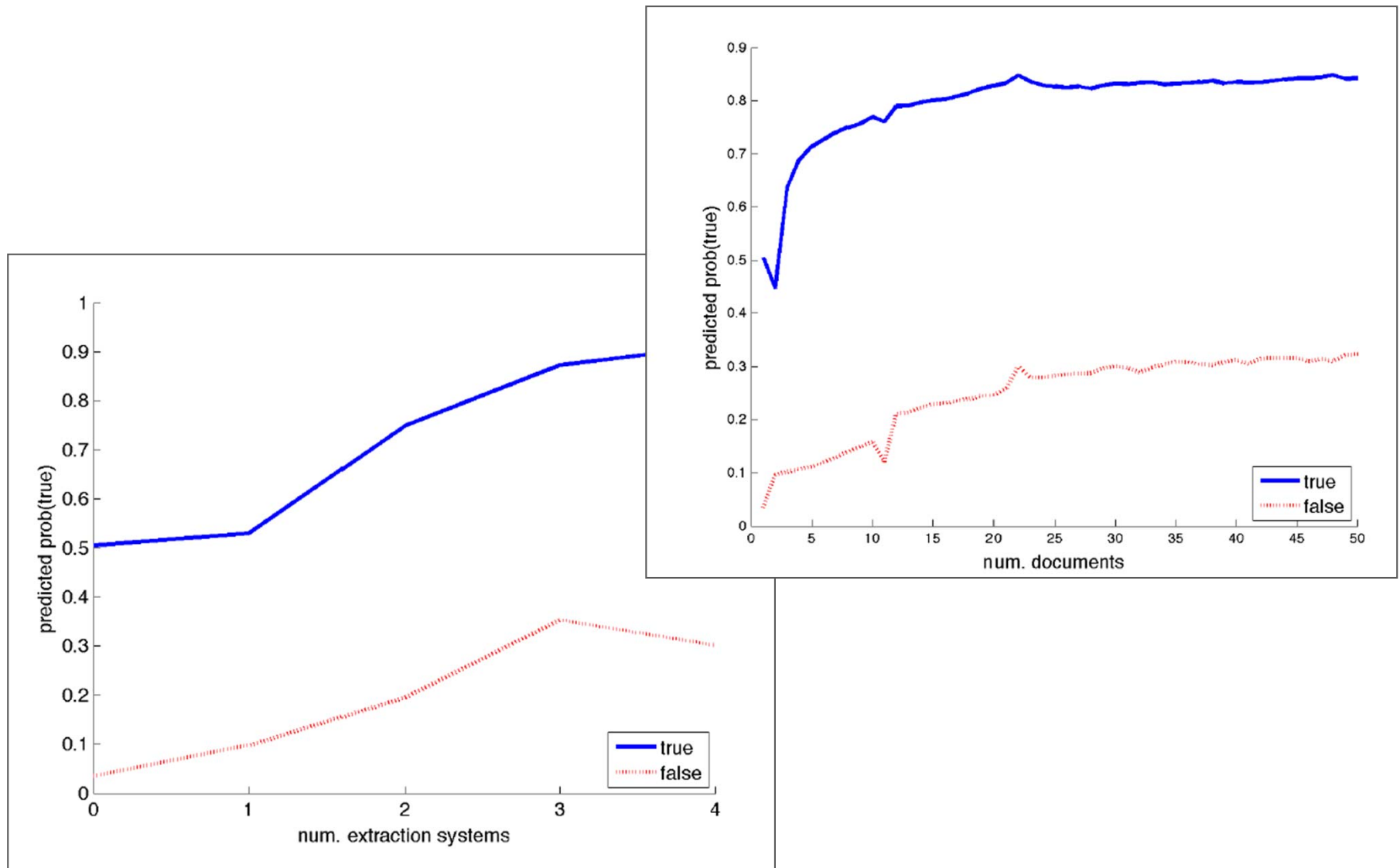


**KNOWLEDGE
VAULT**

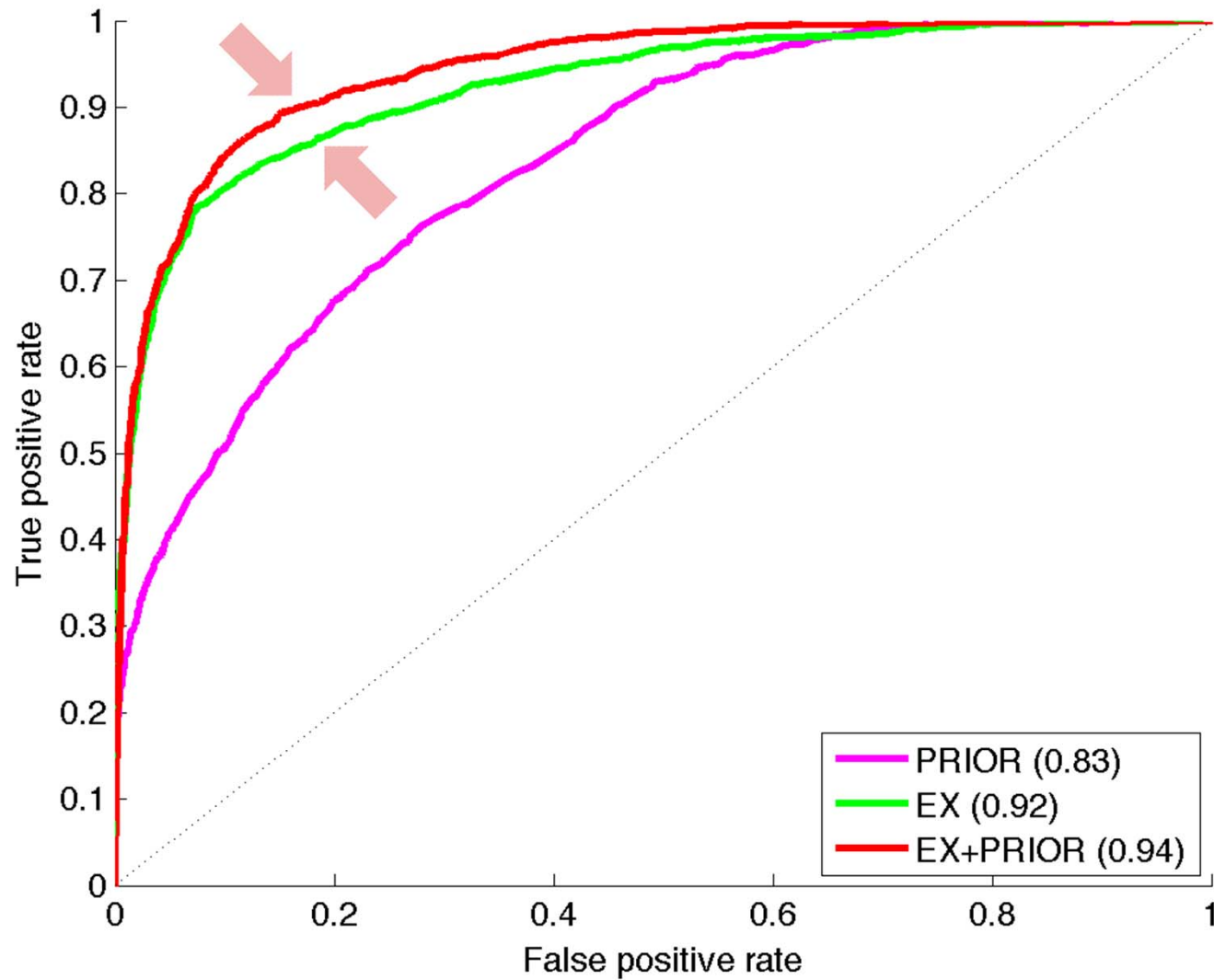
Fusing multiple extractors

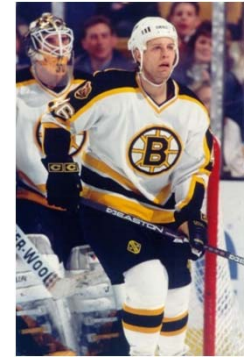


The importance of adding more evidence



Fusing extractors with priors





Example: (Barry Richter, studiedAt, UW-Madison)

“In the fall of 1989, Richter accepted a scholarship to the University of Wisconsin, where he played for four years and earned numerous individual accolades ...”

“The Polar Caps' cause has been helped by the impact of knowledgeable coaches such as Andringa, Byce and former UW teammates Chris Tancill and Barry Richter.”

→ Fused extraction confidence: **0.14**

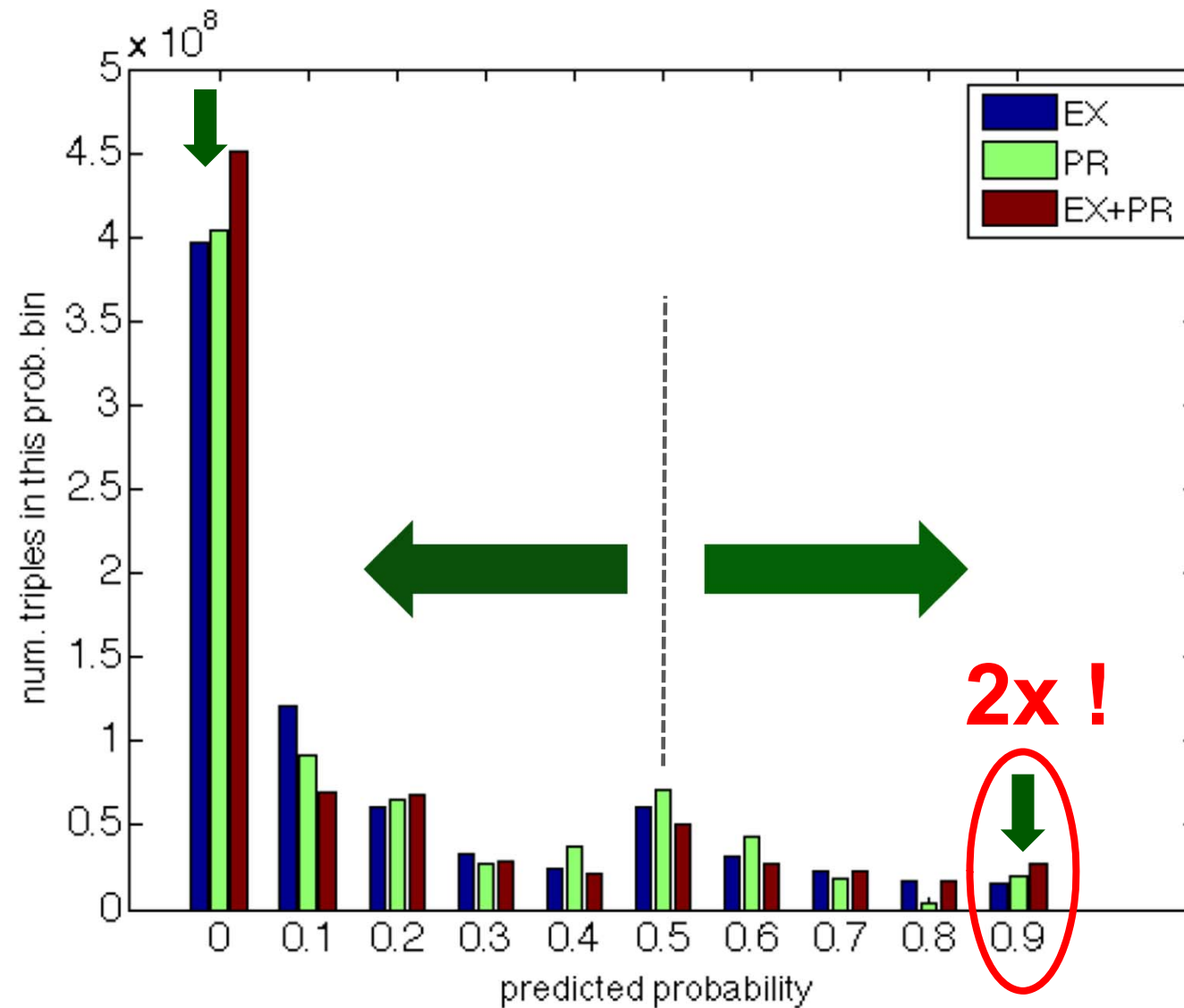


<Barry Richter, born in, Madison>

<Barry Richter, lived in, Madison>

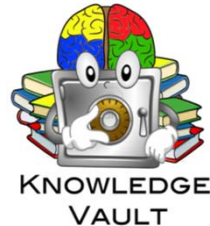
→ Final belief (fused with prior): **0.61**

The importance of prior modeling



Comparison of knowledge repositories

Total # facts in



> 2.5B

Name	# Entity types	# Entity instances	# Relation types	# Confident facts (relation instances)
<i>Knowledge Vault (KV)</i>	1100	45M	4469	302M
DeepDive [32]	4	2.7M	34	7M ^a
NELL [8]	271	5.19M	306	0.435M ^b
PROSPERA [30]	11	N/A	14	0.1M
YAGO2 [19]	350,000	9.8M	100	4M ^c
Freebase [4]	1,500	40M	35,000	637M ^d
Knowledge Graph (KG)	1,500	570M	35,000	18,000M ^e

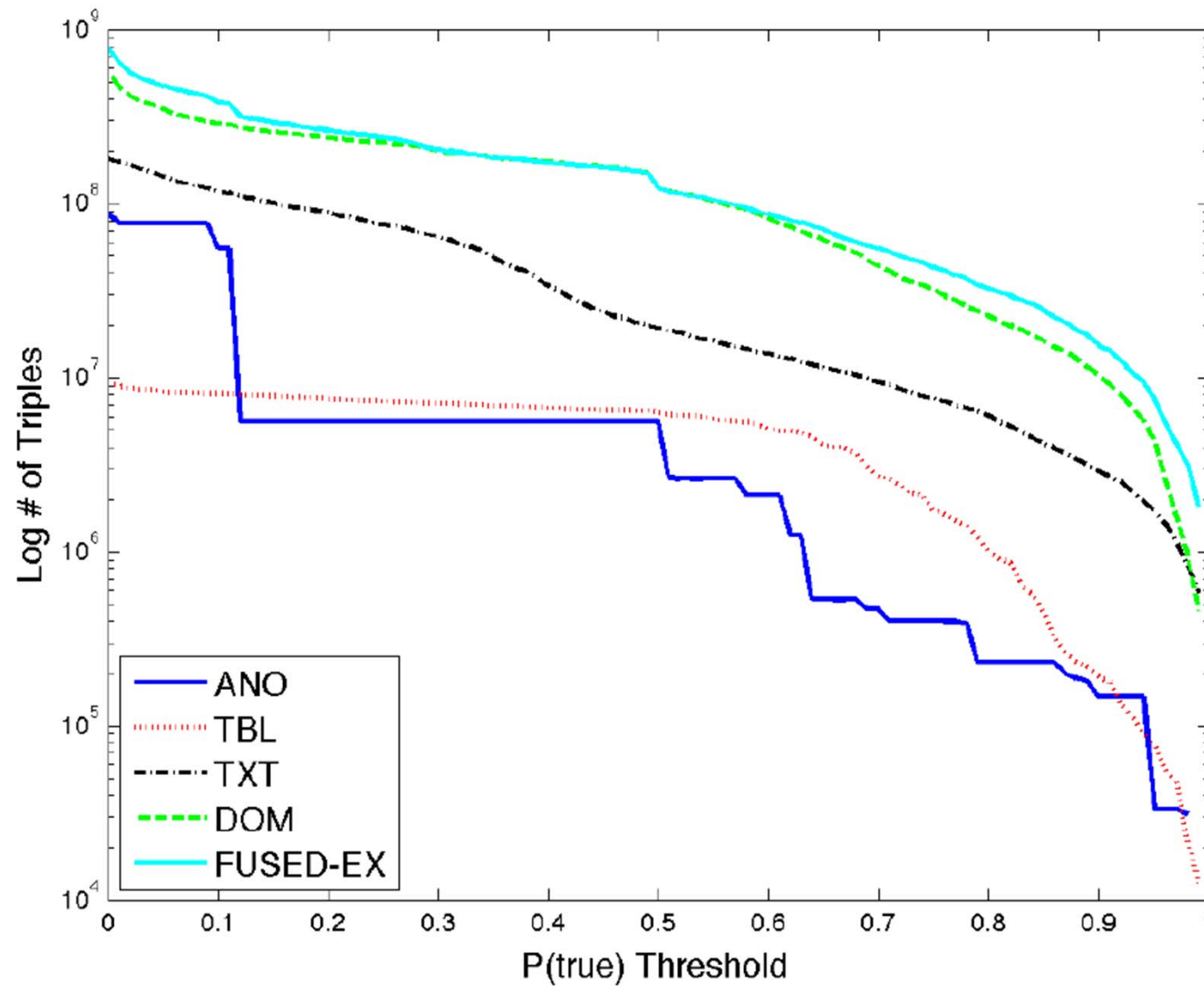
Open IE (e.g., Mausam et al., 2012)

5B assertions (Mausam, Michael Schmitz,
personal communication, October 2013)

302M with Prob > 0.9

381M with Prob > 0.7

The yield from different extraction systems



Should we trust all sources equally ?

WIKIPEDIA The Free Encyclopedia

Article Talk Read View source View history Search

Barack Obama

From Wikipedia, the free encyclopedia

"Obama" redirects here. For other uses, see *Obama* (disambiguation).
This article is about the 44th president of the United States. For his father, see *Barack Obama, Sr.*

Barack Hussein Obama II (/ˈbəˈrɑːkˈhuːsɛnˈoʊbɑːmə/; born August 4, 1961) is the 44th and current President of the United States, the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was president of the *Harvard Law Review*. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004. He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.

In 2004, Obama received national attention during his campaign to represent Illinois in the United States Senate with his victory in the March Democratic Party primary, his keynote address at the Democratic National Convention in July, and his election to the Senate in November. He began his presidential campaign in 2007, and in 2008, after a close primary campaign against Hillary Rodham Clinton, he won sufficient delegates in the Democratic Party primaries to receive the presidential nomination. He then defeated Republican nominee John McCain in the general election, and was inaugurated as president on January 20, 2009. Nine months after his election, Obama was named the 2009 Nobel Peace Prize laureate.

During his first two years in office, Obama signed into law economic stimulus legislation in response to the Great Recession in the form of the American Recovery and Reinvestment Act of 2009 and the Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010. Other major domestic initiatives in his first term include the Patient Protection and Affordable Care Act, often referred to as "Obamacare", the Dodd–Frank Wall Street Reform and Consumer Protection Act; and the Don't Ask, Don't Tell Repeal Act of 2010. In foreign policy, Obama ended U.S. military involvement in the Iraq War, increased U.S. troop levels in Afghanistan, signed the New START arms control treaty with Russia, ordered U.S. military involvement in Libya, and ordered the military operation that resulted in the death of Osama bin Laden. He later became the first sitting U.S. president to publicly support same-sex marriage. In November 2010, the Republicans regained control of the House of

Barack Obama

44th President of the United States
Incumbent

Assumed office
January 20, 2009

Vice President Joe Biden

Preceded by George W. Bush

United States Senator from Illinois

In office
January 3, 2005 – November 16, 2008

Preceded by Peter Fitzgerald

Succeeded by Roland Burris

Member of the Illinois Senate from the 13th District

In office
January 8, 1997 – November 4, 2004

Preceded by Alice Palmer

Succeeded by Kwame Raoul

Personal details

Born Barack Hussein Obama II
August 4, 1961 (age 52)
Honolulu, Hawaii, U.S.

Political party Democratic

The Western Center For Journalism
Informing And Empowering Americans Who Love Freedom

Home Categories Blogging Tools About Polls and Petitions Contact Us Write

You are here: Home / Featured Stories / Proof Obama Born in Kenya? Obama Literary Agent Says Yes

Proof Obama Born in Kenya? Obama Literary Agent Says Yes

MAY 17, 2012 BY FLOYD BROWN 100 COMMENTS

Bretbart.com has introduced some explosive evidence showing that Obama claimed he was born in Kenya years before he became a presidential candidate. Interestingly, the editors of Bretbart still think that now Obama is telling the truth.



ALEX JONES' INFOWARS.COM BECAUSE THERE IS A WAR ON

Home Alex Jones Radio Show News Multimedia Forum News Contact Top Stories Products

Evidence Obama Born In Kenya Goes Beyond 1991 Brochure

Establishment media pulls stunt in effort to diffuse 'birther' controversy

Paul Joseph Watson

InfoWars.com

Friday, May 18, 2012

The establishment media hastily seized on yesterday's explosive story about a literary publication listing Barack Obama's birthplace as Kenya in an effort to claim that the 1991 brochure was the "origin" of the entire 'birther' issue. In reality, evidence that Obama was born in the African country is abundant.

A literary agent's promotional text for a 1991 brochure released yesterday

THE BLAZE

STORIES THEBLAZE TV RADIO MAGAZINE BLOG

HOT TOPICS: Obamacare | Ted Cruz | NSA | Education | TheBlaze TV | #2A

YAHOO! NEWS SAYS OBAMA WAS BORN IN...KENYA!

Jun. 22, 2013 12:34pm | Madeleine Morgenstern

Related: Barack Obama, Birthers, Obama Birth Certificate

Yahoo! News had to issue a correction Friday after publishing an article about President Barack Obama that called Kenya "the country of his birth."

The article, about Obama's upcoming trip to Africa, stated:

President Barack Obama makes the first extended trip to Africa of his presidency next week — but he won't be stopping at the country of his birth.

White House doesn't have 'figure on costs' of Africa trip

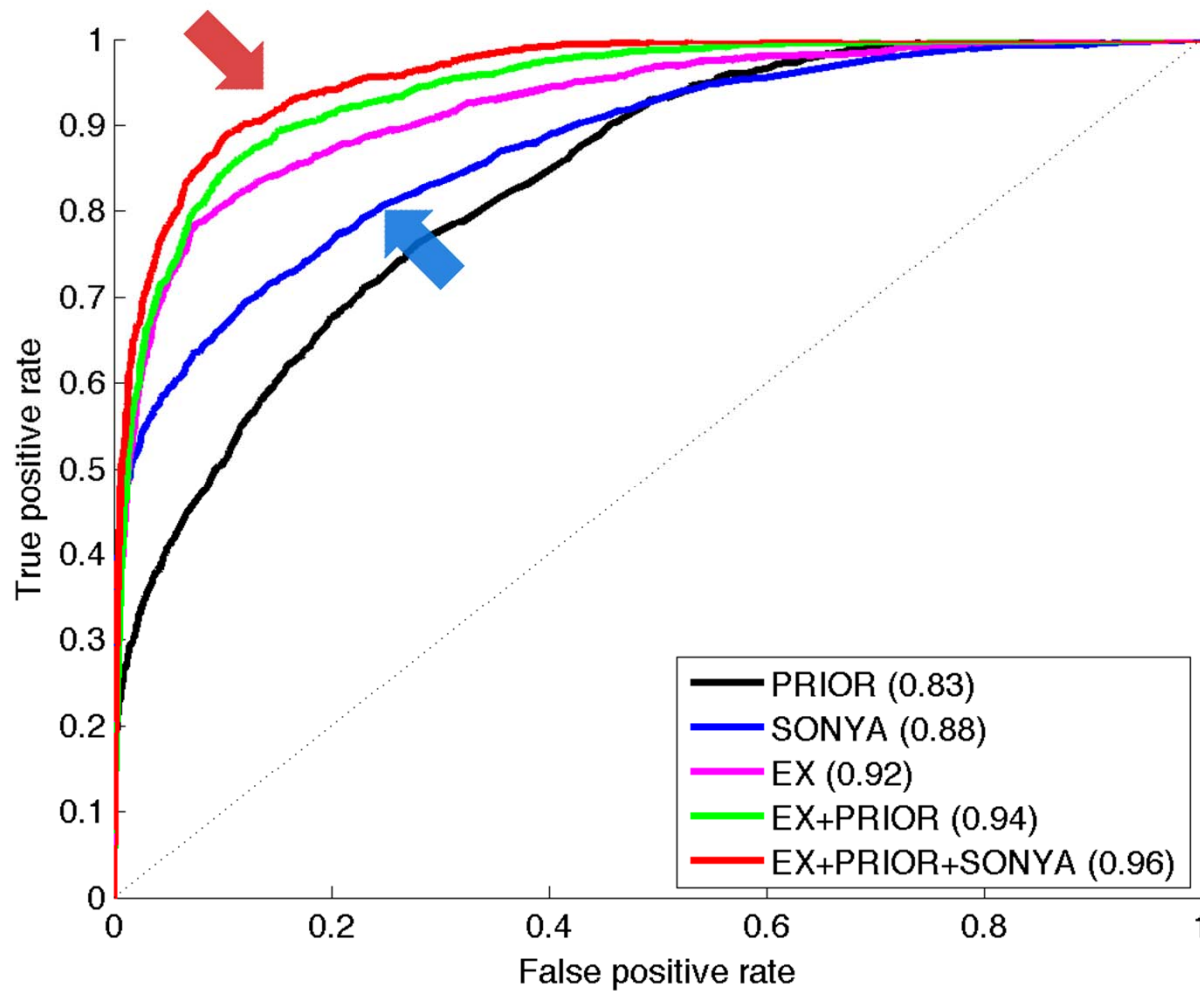
By Rachel Rose Hartman, Yahoo! News | The Ticket — 1 hr 40 mins ago

President Barack Obama makes the first extended trip to Africa of his presidency next week—but he won't be stopping in the country of his birth.

derich hastily claimed listing error."
al and it goes significantly
was a mistake, the listing still use a U.S. Senator. "Goderich's sixteen years, through at least four different versions of
about Obama being born in
Obama had become a Senator,
bama, "was born in Kenya."

Joint modeling of source and fact accuracy

[Dong et al., VLDB '09]



**Estimate
source accuracy**

**Estimate
fact accuracy**

Automatic knowledge base completion (focused extraction)

Relation % unknown in Freebase

Profession	68%
Place of birth	71%
Nationality	75%
Education	91%
Spouse	92%
Parents	94%

People /people	
Person /people/person	
Date of birth /people/person/date_of_birth	4004 BCE
Place of birth /people/person/place_of_birth	
Garden of Eden	
Country of nationality /people/person/nationality	
Gender /people/person/gender	Male
Profession /people/person/profession	



(Genesis 2)

⁸ And the LORD God **planted a garden eastward in Eden; and there he put the man whom he had formed.**

¹⁵ Then the LORD God **took the man and put him in the garden of Eden to tend and keep it.**

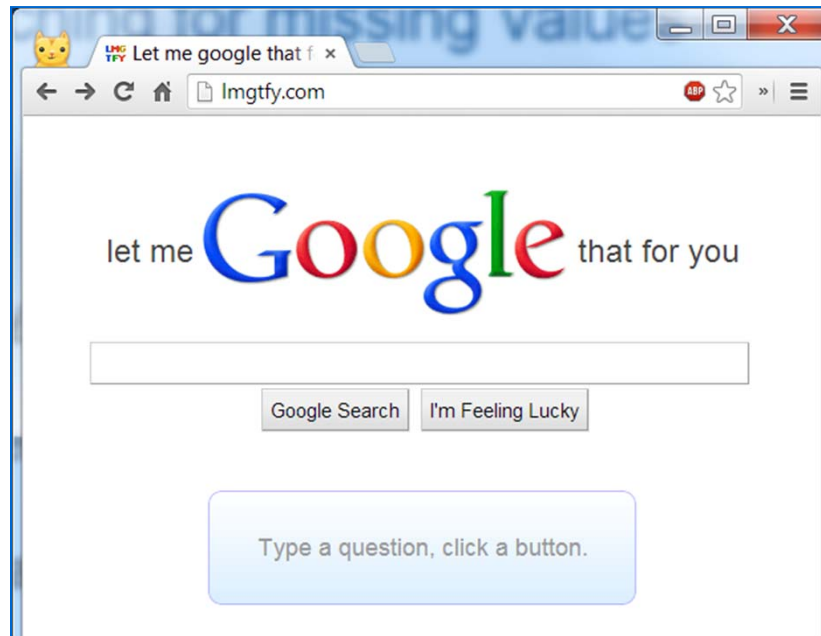
¹⁹ And out of the ground the LORD God formed every beast of the field, and every fowl of the air; and **brought them unto Adam to see what he would call them: and whatsoever Adam called every living creature, that was the name thereof.**

Employment history /people/person/employment_history

Employer

Title

Proactively searching for missing values [West et al., WWW '14]



- Mine search logs for best query templates (per relation)
- Augment queries with disambiguating information
- Thou shalt ask **in moderation**
 - Asking too much may be harmful!

The importance of query augmentation

Who is the mother of Frank Zappa



[The Mothers of Invention - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/The_Mothers_of_Invention ▼

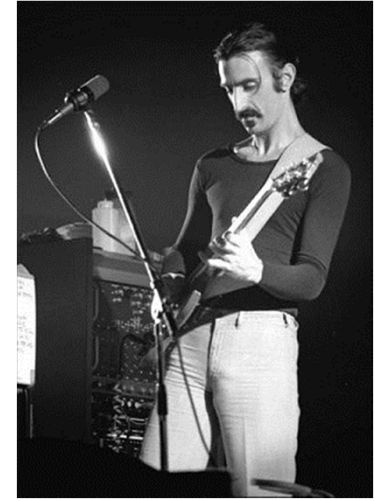
The **Mothers of Invention** were an American rock band from California that served as the backing musicians for **Frank Zappa**, a self-taught composer and ...

[History](#) - [Personnel](#) - [Discography](#) - [References](#)

[Frank Zappa - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Frank_Zappa ▼

Jump to 1970: Rebirth of The **Mothers** and filmmaking - [edit]. **Frank Zappa** in Paris, early 1970s. Later in 1970, Zappa formed a new version of The ...

[Discography](#) - [Moon Zappa](#) - [Diva Zappa](#) - [Gail Zappa](#)



Who is the mother of Frank Zappa Baltimore Maryland

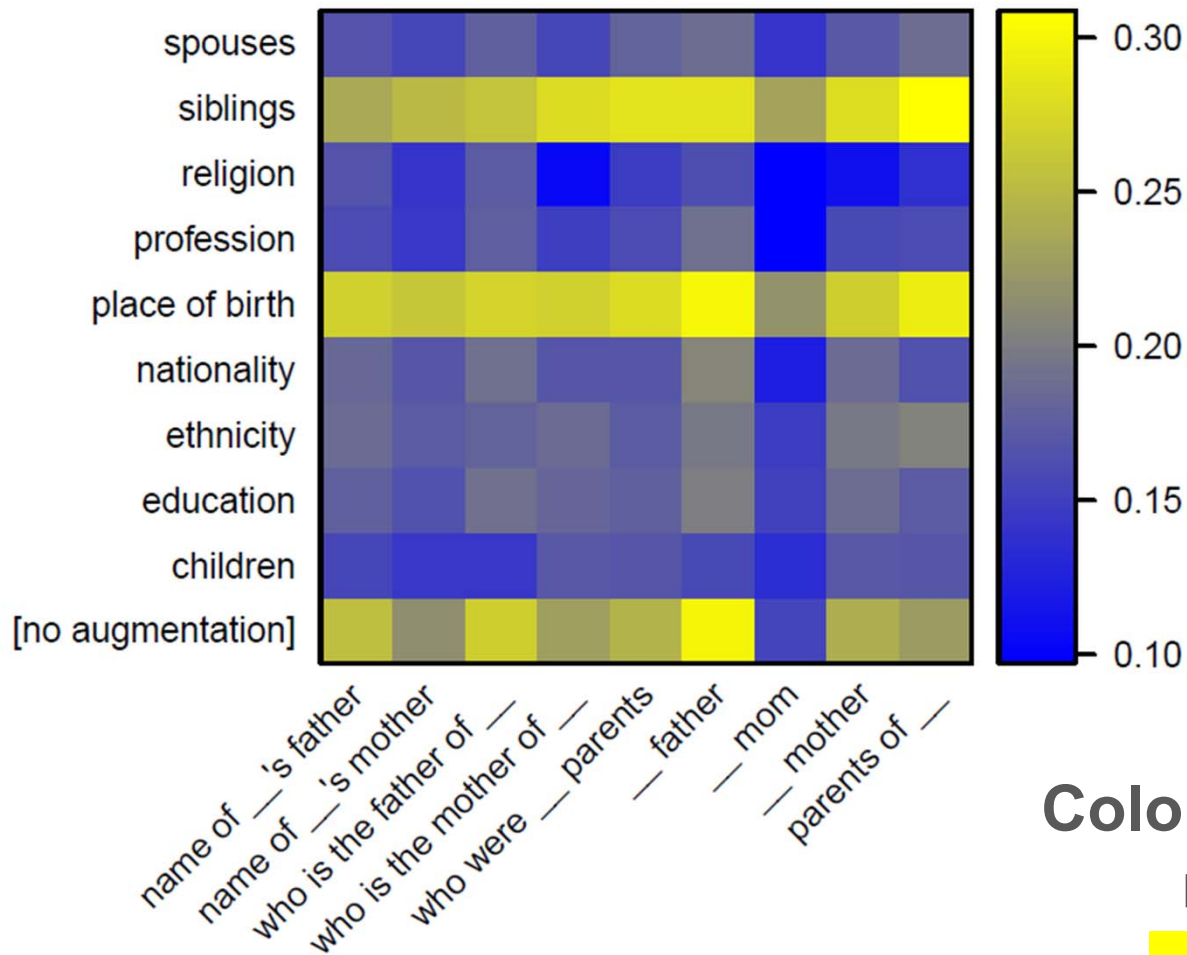


[Frank Zappa - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Frank_Zappa ▼

Frank Vincent Zappa was born in **Baltimore, Maryland**, on December 21, 1940. His **mother**, Rose Marie (née Colimore), was of Italian and French ancestry; his ...

Learning to query

/people/person/parents

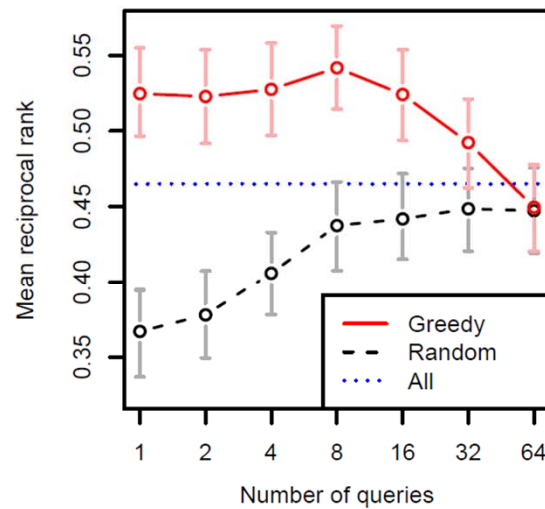


Color = mean reciprocal rank of true answer

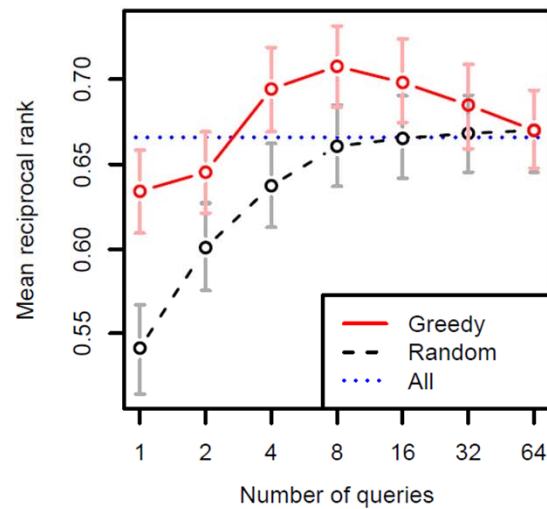
GOOD

BAD

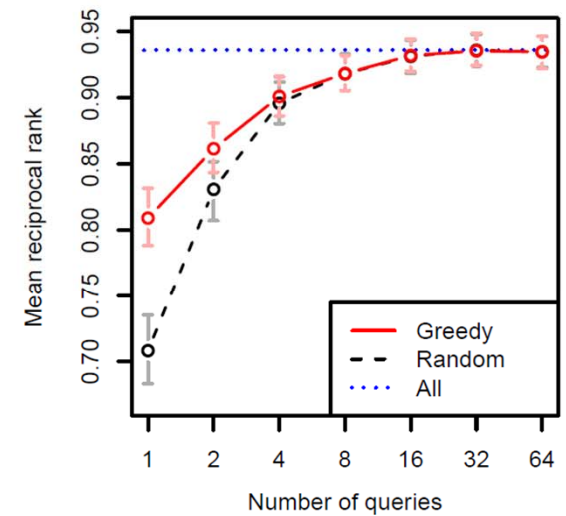
Asking the right (number of) questions



(a) SPOUSES



(b) PLACE OF BIRTH



(c) NATIONALITY

PART 2: METHODS AND TECHNIQUES

Methods and techniques

1. Relation extraction:

- Supervised models
- Semi-supervised models
- Distant supervision

2. Entity resolution

- Single entity methods
- Relational methods

3. Link prediction

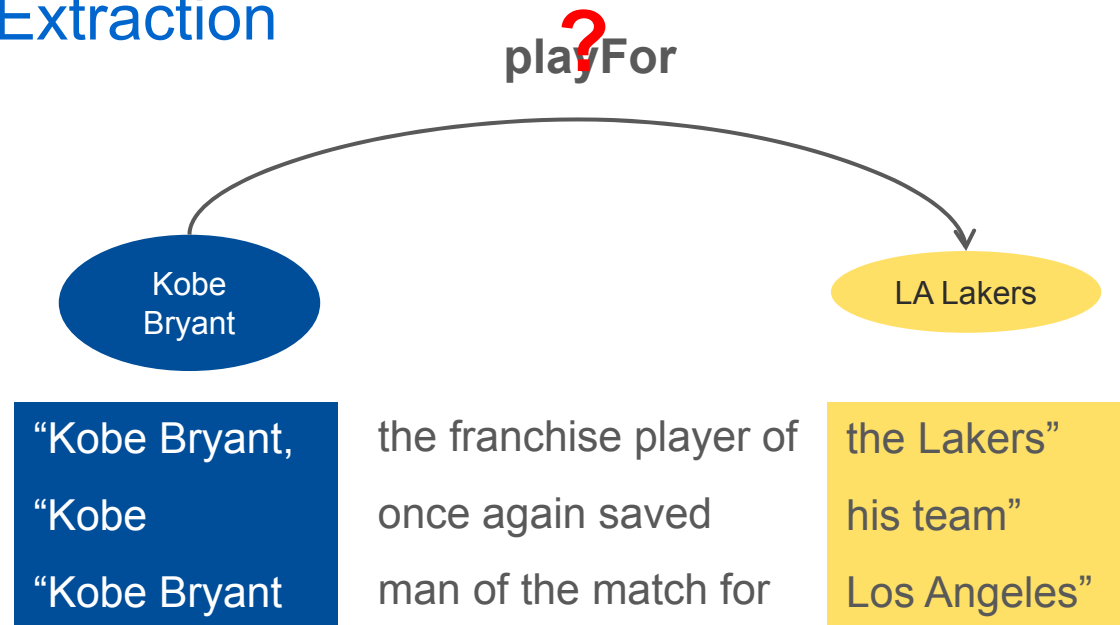
- Rule-based methods
- Probabilistic models
- Factorization methods
- Embedding models

Not in this tutorial:

- Entity classification
- Group/expert detection
- Ontology alignment
- Object ranking

RELATION EXTRACTION

Relation Extraction

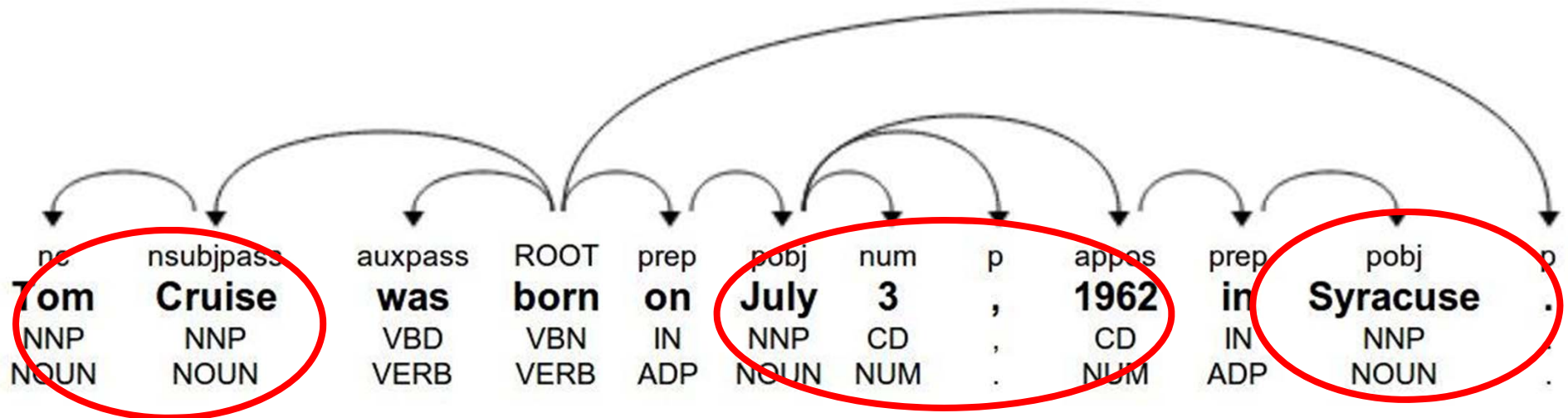


- Extracting semantic relations between sets of [grounded] entities
- Numerous variants:
 - Undefined vs **pre-determined set of relations**
 - **Binary** vs n-ary relations, facet discovery
 - Extracting temporal information
 - Supervision: {**fully**, un, **semi**, **distant**}-supervision
 - Cues used: only lexical vs **full linguistic features**

Supervised relation extraction

- Sentence-level labels of relation mentions
 - "**Apple** CEO **Steve Jobs** said.." => (SteveJobs, CEO, Apple)
 - "**Steve Jobs** said that **Apple** will.." => NIL
- Traditional relation extraction datasets
 - ACE 2004
 - MUC-7
 - Biomedical datasets (e.g BioNLP challenges)
- Learn classifiers from +/- examples
- Typical features: context words + POS, dependency path between entities, named entity tags, token/parse-path/entity distance

Examples of features



X was born on DDDD in Y

- **DFP**: X <nsubjpass / born prep> on pobj> DATE prep> in pobj> Y
- **NER**: X = PER, Y = LOC
- **POS**: X = NOUN, NNP; Y = NOUN, NNP
- **Context**: born, on, in , "born on"

Supervised relation extraction

- Used to be the “traditional” setting [Riloff et al., 06; Soderland et al., 99]
- **Pros**
 - High quality supervision
 - Explicit negative examples
- **Cons**
 - **Very expensive to generate supervision**
 - Not easy to add more relations
 - Cannot generalize to text from different domains

Semi-supervised relation extraction

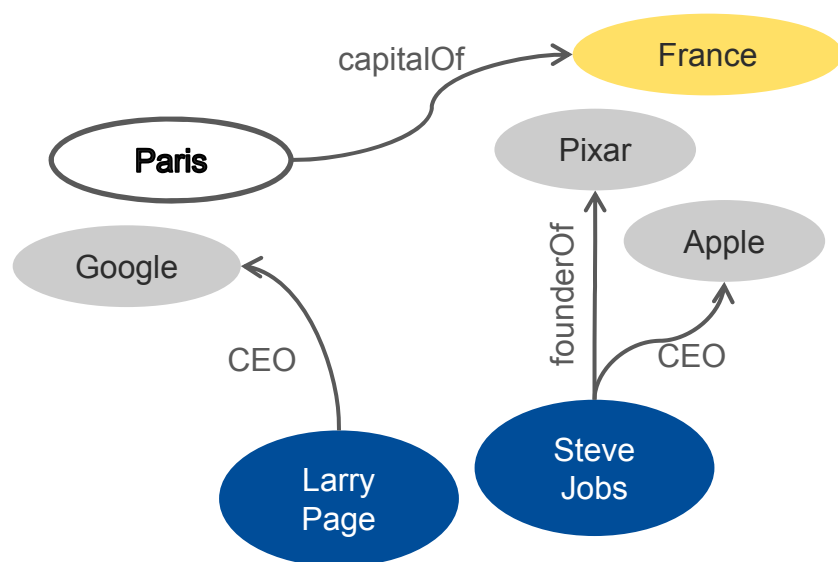
- **Generic algorithm**
 1. Start with **seed triples / golden seed patterns**
 2. **Extract patterns that match** seed triples/patterns
 3. Take the **top-k** extracted patterns/triples
 4. **Add to seed** patterns/triples
 5. Go to 2
- Many published approaches in this category:
 - Dual Iterative Pattern Relation Extractor [Brin, 98]
 - Snowball [Agichtein & Gravano, 00]
 - **TextRunner [Banko et al., 07] – almost unsupervised**
- Differ in pattern definition and selection

TextRunner [Banko et al., 07]

- Almost unsupervised
 - **Relations not fixed:** does not follow Knowledge Graph schema (**growing**)
 - No labeled data
 - Mostly unlabeled text
 - Uses heuristics to **self-label** a starting corpora (using a parser), such as
 - Path length $< k$
 - Path does not cross sentence-like boundaries like relative clauses
 - Neither entity is a pronoun
- Self-training
 - Generate +/- examples \rightarrow learn classifier
 - **Extract new relation mentions** using this classifier
 - **Generate triples from aggregated mentions**, assign probabilistic score using [Downey et. al., 2005]
- Later improved in Reverb [Fader et al., 11]

Distantly-supervised relation extraction

- Existing knowledge base + unlabeled text → generate examples
 - Locate pairs of related entities in text
 - Hypothesizes that the relation is expressed



Google CEO Larry Page announced that...

Steve Jobs has been Apple for a while...

Pixar lost its co-founder Steve Jobs...

I went to Paris, France for the summer...

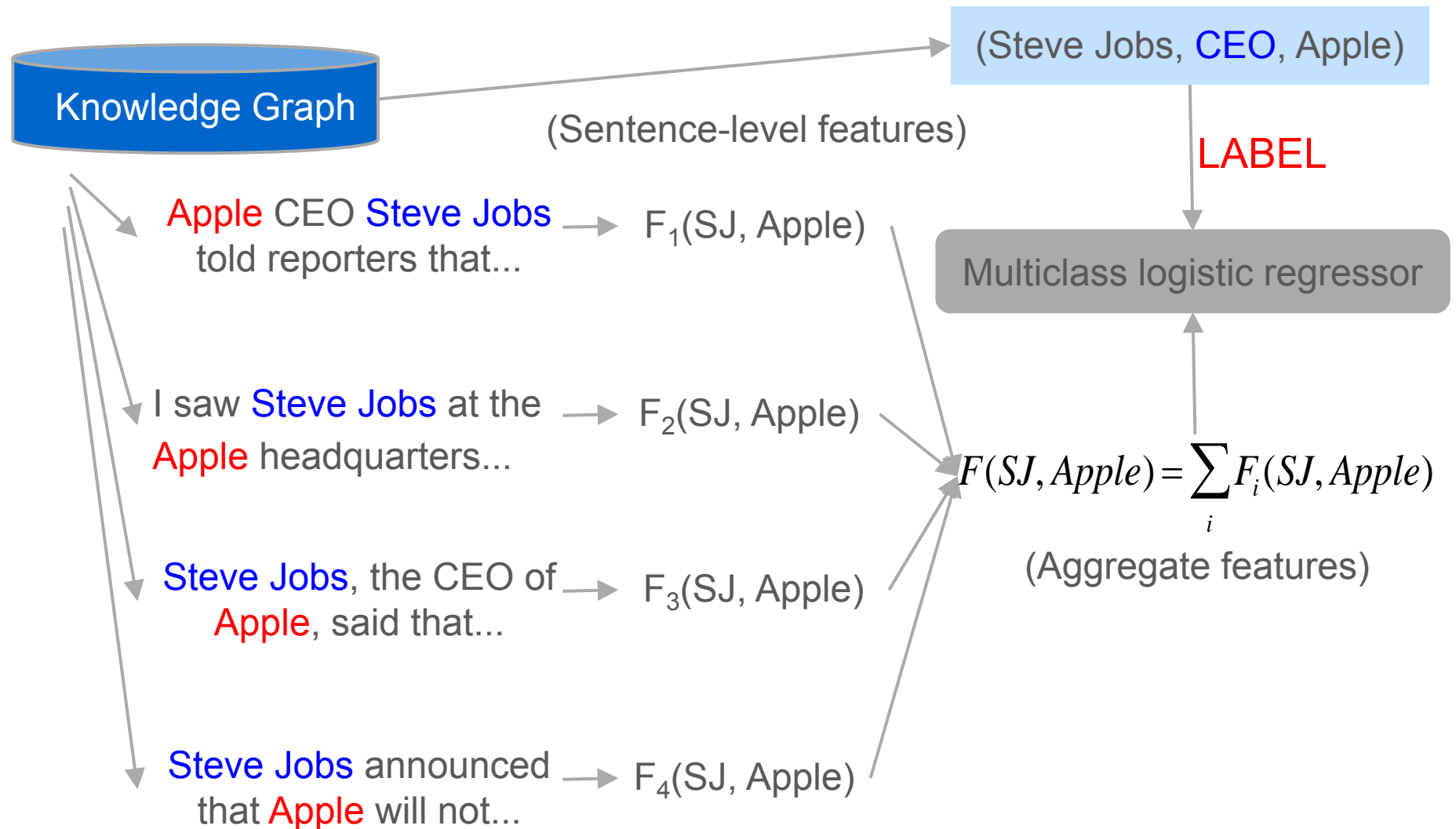
Distant supervision: modeling hypotheses

Typical architecture:

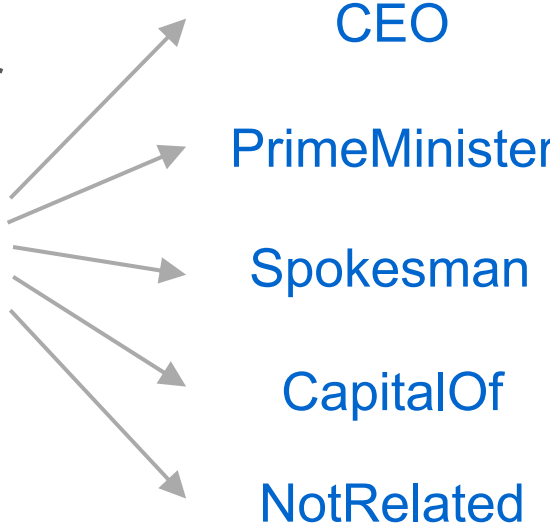
1. Collect many pairs of entities co-occurring in sentences from text corpus
2. If 2 entities participate in a relation, several hypotheses:
 1. **All** sentences mentioning them express it [Mintz et al., 09]

“**Barack Obama** is the 44th and current President of **the US**.” → (BO, employedBy, USA)

[Mintz et al., 09]



[Mintz et al., 09]

- Classifier: multiclass logistic regressor
(Steve Jobs, Apple, *AggFeatures*)

```
graph LR; Input["(Steve Jobs, Apple, AggFeatures)"] --> CEO; Input --> PM["PrimeMinister"]; Input --> Spokesman; Input --> CapitalOf; Input --> NotRelated;
```

 - Negative examples
 - Randomly sample **unrelated entity pairs occurring in the same sentence**
 - > 98% such pairs actually unrelated

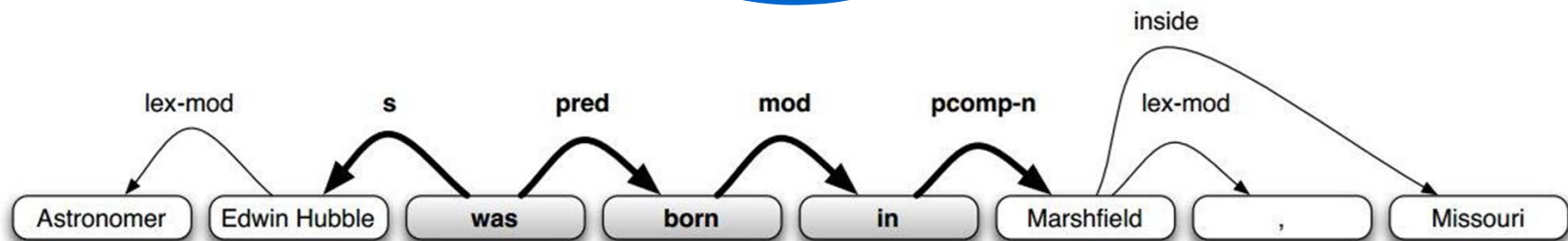
Sentence-level features

- **Lexical:** words in between and around mentions and their parts-of-speech tags (conjunctive form)
- **Syntactic:** dependency parse path between mentions along with side nodes
- **Named Entity Tags:** for the mentions
- **Conjunctions** of the above features
 - Distant supervision is used on to lots of data → sparsity of conjunctive forms not an issue

Sentence-level features

Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[.]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]
Syntactic	[]	PER	[\uparrow_s was \downarrow_{pred} born \downarrow_{mod} in $\downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Edwin Hubble $\downarrow_{lex-mod}$]	PER	[\uparrow_s was \downarrow_{pred} born \downarrow_{mod} in $\downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Astronomer $\downarrow_{lex-mod}$]	PER	[\uparrow_s was \downarrow_{pred} born \downarrow_{mod} in $\downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[]	PER	[\uparrow_s was \downarrow_{pred} born \downarrow_{mod} in $\downarrow_{pcomp-n}$]	LOC	[$\downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\downarrow_{lex-mod}$]	PER	[\uparrow_s was \downarrow_{pred} born \downarrow_{mod} in $\downarrow_{pcomp-n}$]	LOC	[$\downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\downarrow_{lex-mod}$]	PER	[\uparrow_s was \downarrow_{pred} born \downarrow_{mod} in $\downarrow_{pcomp-n}$]	LOC	[$\downarrow_{lex-mod}$,]
Syntactic	[]	PER	[\uparrow_s was \downarrow_{pred} born \downarrow_{mod} in $\downarrow_{pcomp-n}$]	LOC	[\downarrow_{inside} Missouri]
Syntactic	[Edwin Hubble $\downarrow_{lex-mod}$]	PER	[\uparrow_s was \downarrow_{pred} born \downarrow_{mod} in $\downarrow_{pcomp-n}$]	LOC	[\downarrow_{inside} Missouri]
Syntactic	[Astronomer $\downarrow_{lex-mod}$]	PER	[\uparrow_s was \downarrow_{pred} born \downarrow_{mod} in $\downarrow_{pcomp-n}$]	LOC	[\downarrow_{inside} Missouri]

Table 3: Features for 'Astronomer Edwin Hubble was born in Marshfield Missouri'.



Examples of top features

Relation	Feature type	Left window	NE1	Middle	NE2	Right window
/architecture/structure/architect	LEX \hookleftarrow		ORG	, the designer of the	PER	
	SYN	designed \uparrow_s	ORG	\uparrow_s designed $\downarrow_{bu-subj}$ by \downarrow_{pcn}	PER	\uparrow_s designed
/book/author/works_written	LEX		PER	s novel	ORG	
	SYN		PER	\uparrow_{pcn} by \uparrow_{mod} story \uparrow_{pred} is \downarrow_s	ORG	
/book/book_edition/author_editor	LEX \hookleftarrow		ORG	s novel	PER	
	SYN		PER	\uparrow_{nn} series \downarrow_{gen}	PER	
/business/company/founders	LEX		ORG	co - founder	PER	
	SYN		ORG	\uparrow_{nn} owner \downarrow_{person}	PER	
/business/company/place_founded	LEX \hookleftarrow		ORG	- based	LOC	
	SYN		ORG	\uparrow_s founded \downarrow_{mod} in \downarrow_{pcn}	LOC	
/film/film/country	LEX		PER	, released in	LOC	
	SYN	opened \uparrow_s	ORG	\uparrow_s opened \downarrow_{mod} in \downarrow_{pcn}	LOC	\uparrow_s opened
/geography/river/mouth	LEX		LOC	, which flows into the	LOC	
	SYN	the \downarrow_{det}	LOC	\uparrow_s is \downarrow_{pred} tributary \downarrow_{mod} of \downarrow_{pcn}	LOC	\downarrow_{det} the

Distant supervision: modeling hypotheses

Typical architecture:

1. Collect many pairs of entities co-occurring in sentences from text corpus
2. If 2 entities participate in a relation, several hypotheses:
 1. **All** sentences mentioning them express it [Mintz et al., 09]
 2. **At least one** sentence mentioning them express it [Riedel et al., 10]

“**Barack Obama** is the 44th and current President of **the US**.” → (BO, employedBy, USA)

“**Obama** flew back to **the US** on Wednesday.” → (BO, employedBy, USA)

[Riedel et al., 10]

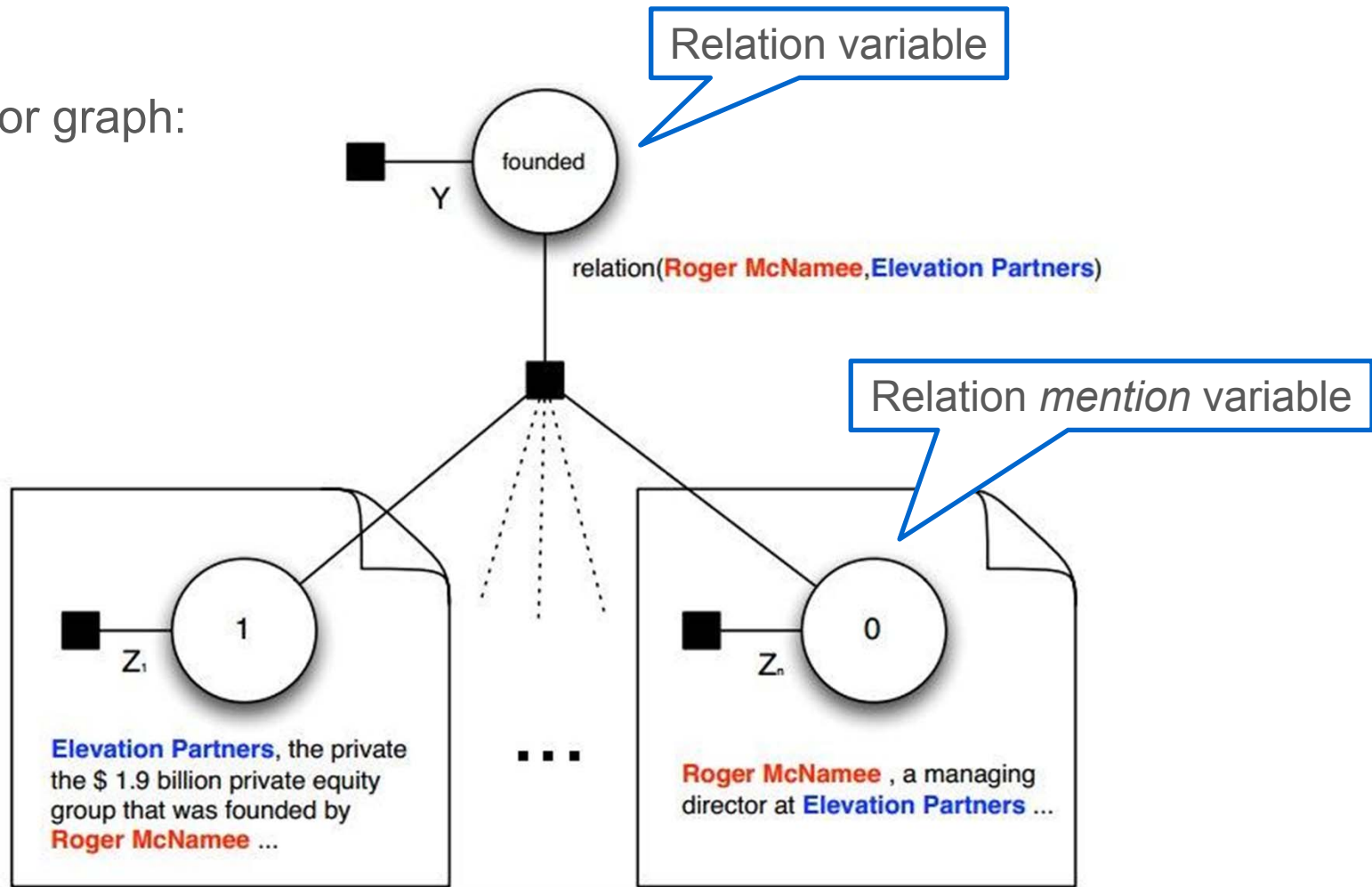
- Every mention of an entity-pair does not express a relation

Relation Type	New York Times	Wikipedia
nationality	38%	20%
place_of_birth	35%	20%
contains	20%	10%

- Violations more in news than encyclopediac articles
- Assert triple from only a few mentions, not all

[Riedel et al., 10]

- Factor graph:



- Multiple-instance setting

Distant supervision: modeling hypotheses

Typical architecture:

1. Collect many pairs of entities co-occurring in sentences from text corpus
2. If 2 entities participate in a relation, several hypotheses:
 1. **All** sentences mentioning them express it [Mintz et al., 09]
 2. **At least one** sentence mentioning them express it [Riedel et al., 10]
 3. **At least one** sentence mentioning them express it and 2 entities can express **multiple relations** [Hoffmann et al., 11] [Surdeanu et al., 12]

“**Barack Obama** is the 44th and current President of **the US**.” → (BO, employedBy, USA)

“**Obama** flew back to **the US** just Wednesday’s” → (BO, ✗ (BO, bornIn, USA))



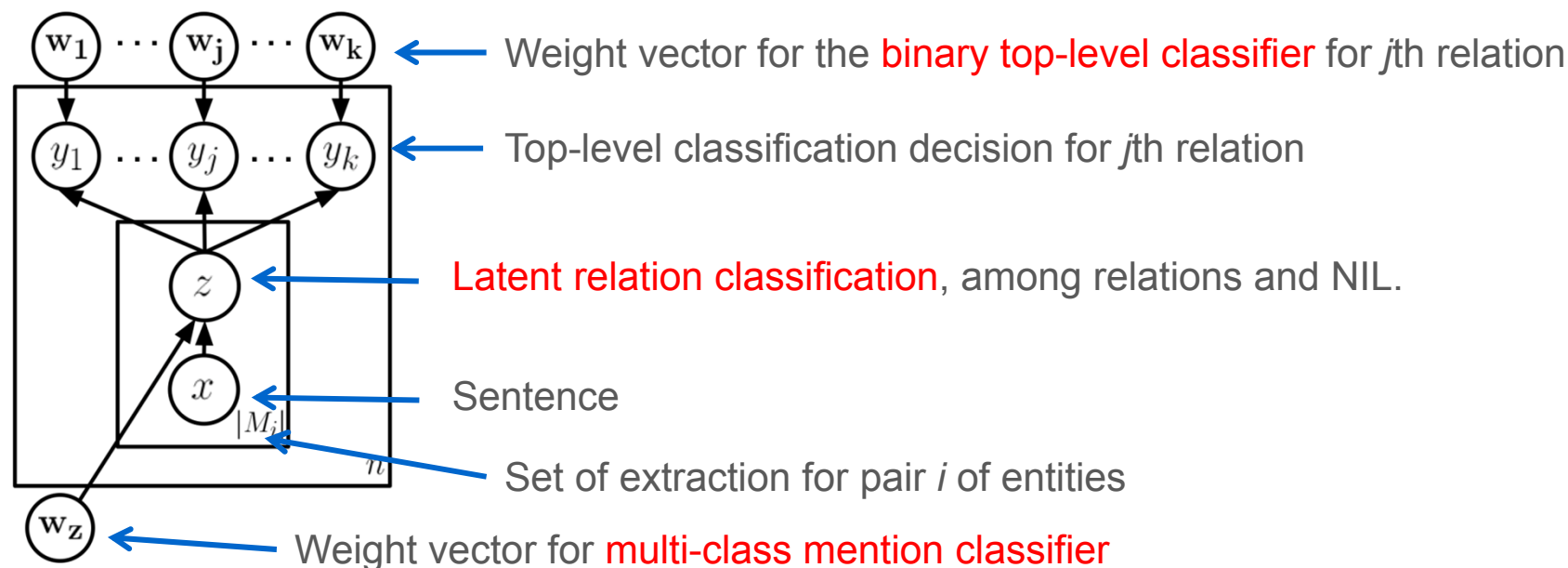
[Surdeanu et al., 12]

- Relation extraction is a multi-instance multi-label problem.

“**Barack Obama** is the 44th and current President of **the US**.” \rightarrow (BO, *employedBy*, USA)

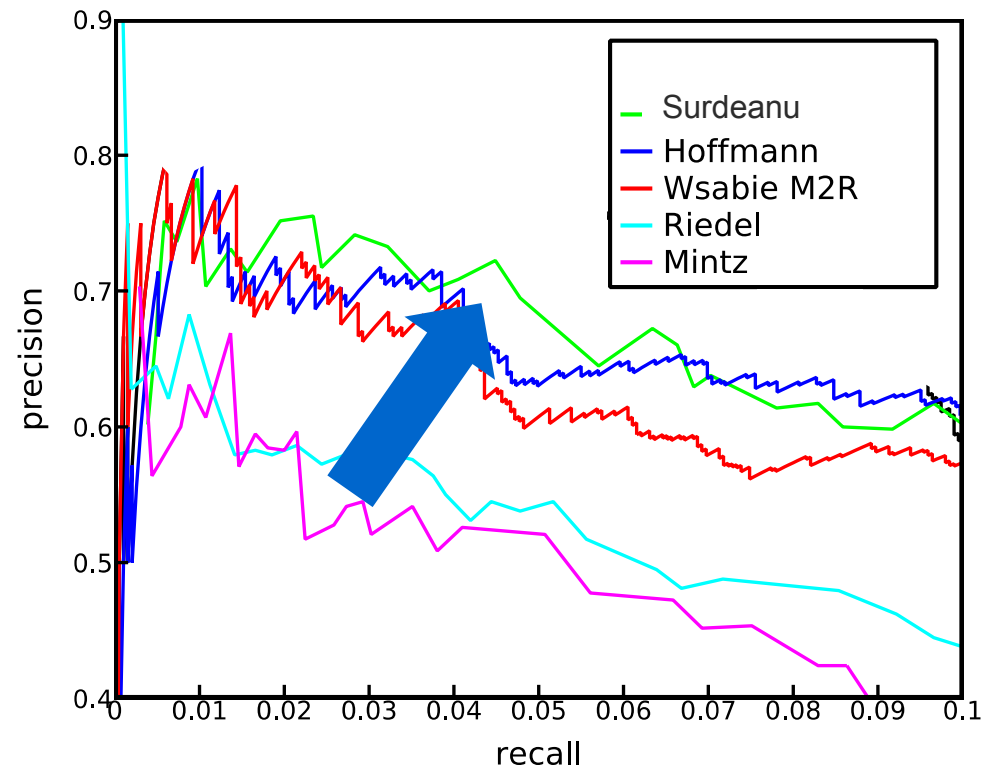
“**Obama** was born in **the US** just as he always said.” \rightarrow (BO, *bornIn*, USA)

“**Obama** flew back to **the US** on Wednesday.” \rightarrow *NIL*



- Training via EM with initialization with [Mintz et al., 09]

Relaxing hypotheses improves precision



Precision-recall curves on extracting from New York Times articles to Freebase [Weston et al., 13]

Distant supervision

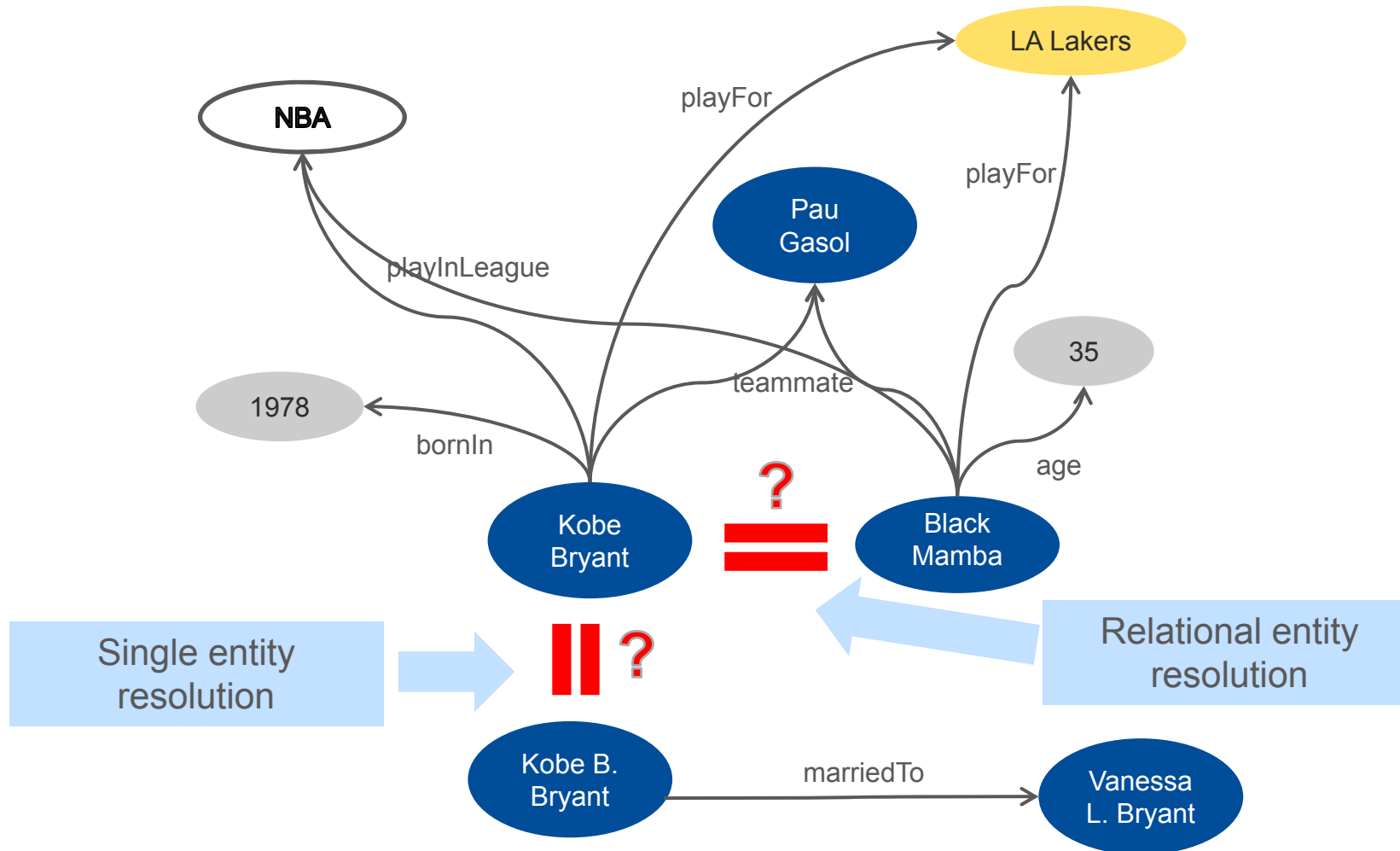
- **Pros**
 - Can scale to the web, as no supervision required
 - Generalizes to text from different domains
 - Generates a lot more supervision in one iteration
- **Cons**
 - Needs high quality entity-matching
 - Relation-expression hypothesis can be wrong
 - Can be compensated by the extraction model, redundancy, language model
 - Does not generate negative examples
 - Partially tackled by matching unrelated entities

Plenty of extensions

- Using **language models** [Downey et al., 07]
 - Do two entities seem to express a given relation, given the context?
- Joint **relation extraction + other NLP tasks**
 - Named Entity tagging [Yao et al., 10]
 - Possibly with entity resolution and/or coreference
- Jointly + **repeatedly training** multiple extractors [Carlson et. al., 10]
- **Unsupervised** extraction [Poon & Domingos, 10]
- **Jointly perform relation extraction and link prediction** [Bordes et al., 12; Weston et al., 13; Riedel et al., 13]

ENTITY RESOLUTION

Entity resolution

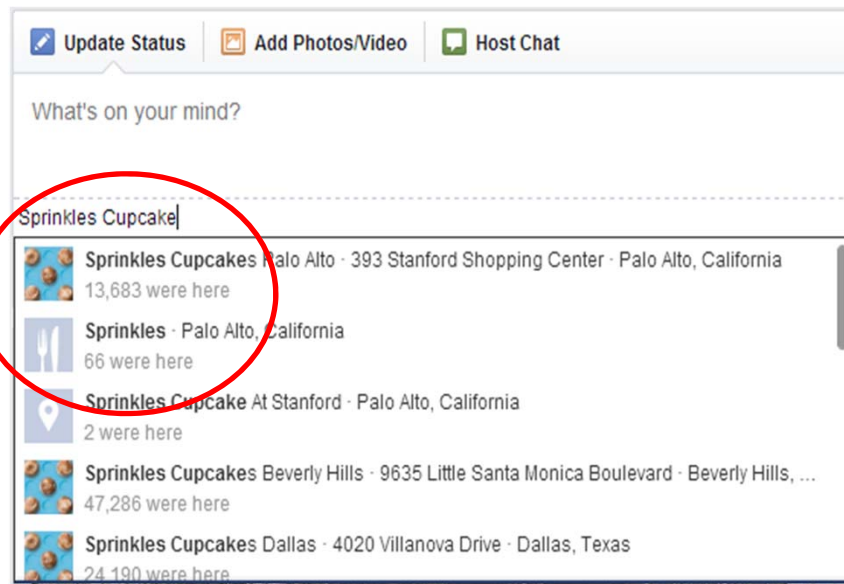


Single-entity entity resolution

- Entity resolution **without using the relational context** of entities
- Many **distances/similarities** for single-entity entity resolution:
 - Edit distance (Levenshtein, etc.)
 - Set similarity (TF-IDF, etc.)
 - Alignment-based
 - Numeric distance between values
 - Phonetic Similarity
 - Equality on a boolean predicate
 - Translation-based
 - Domain-specific

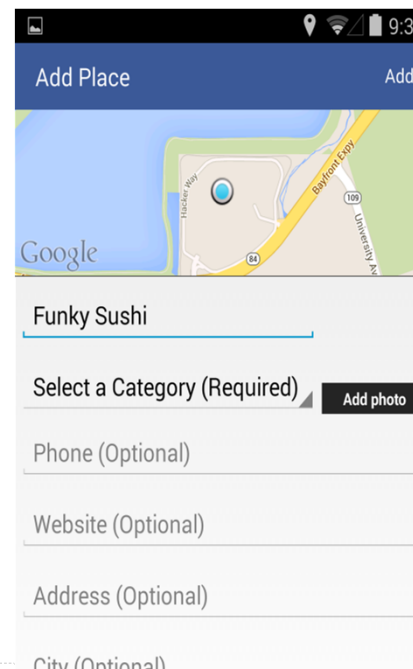
Case study: deduplicating places [Dalvi et al., 14]

- Multiple mentions of the same place is wrong and confusing.



Origin of duplicates

- Duplicates are often created during check-in:
 - Different spellings
 - GPS Errors
 - Wrong checkins
- Frequently, these duplicates have:
 - few attribute values
 - names were typed hurriedly



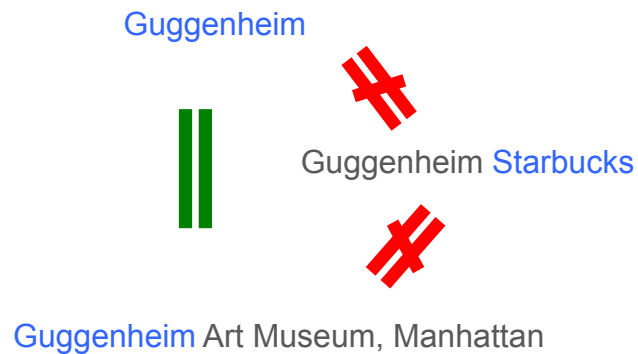
Effectively matching place names is hard

Good Matches (Help Recall)		Bad Matches (Hurt Precision)	
Guggenheim Art Museum Manhattan	Guggenheim	Guggenheim	Guggenheim Starbucks
DishDash	Dish Dash Restaurant	Central Park Café (NYC)	Central Park Restaurant (NYC)
Ippudo New York	Ipudo	Glen Park	Glen Canyon Park
Central Park Café (Sunnyvale)	Central Park Restaurant (Sunnyvale)		

- Easy to find cases where the “bad match” pair is more similar than the “good match” pair
- Existing similarity metrics (TF-IDF, Levenshtein, Learned-weight edit distance, etc.) generally fail to handle this level of variability

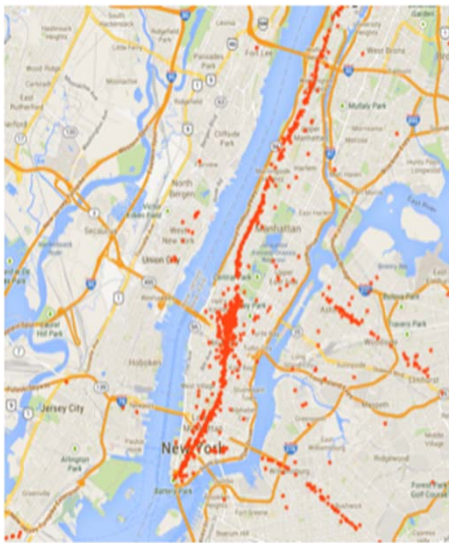
Idea 1: core words

- A core word = a word a human would use to refer to the place, if only a single word were allowed
- **Goal:** try to identify the core word, use it for comparisons



Idea 2: spatial context model

- Tokens **vary in importance** based on geographic context
 - Central Park is common/meaningless in NYC
- **Goal:** filter out context-specific tokens when matching names



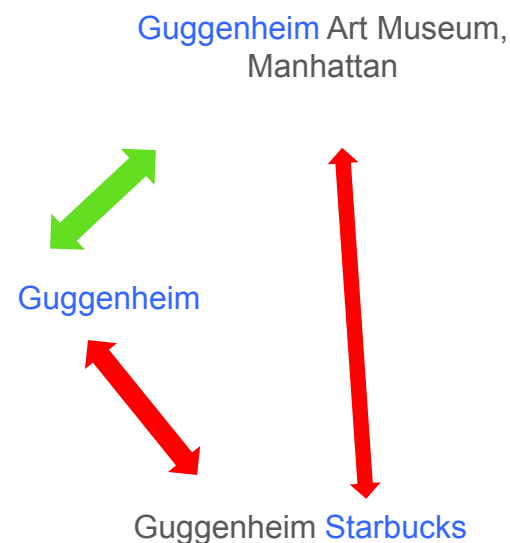
Broadway



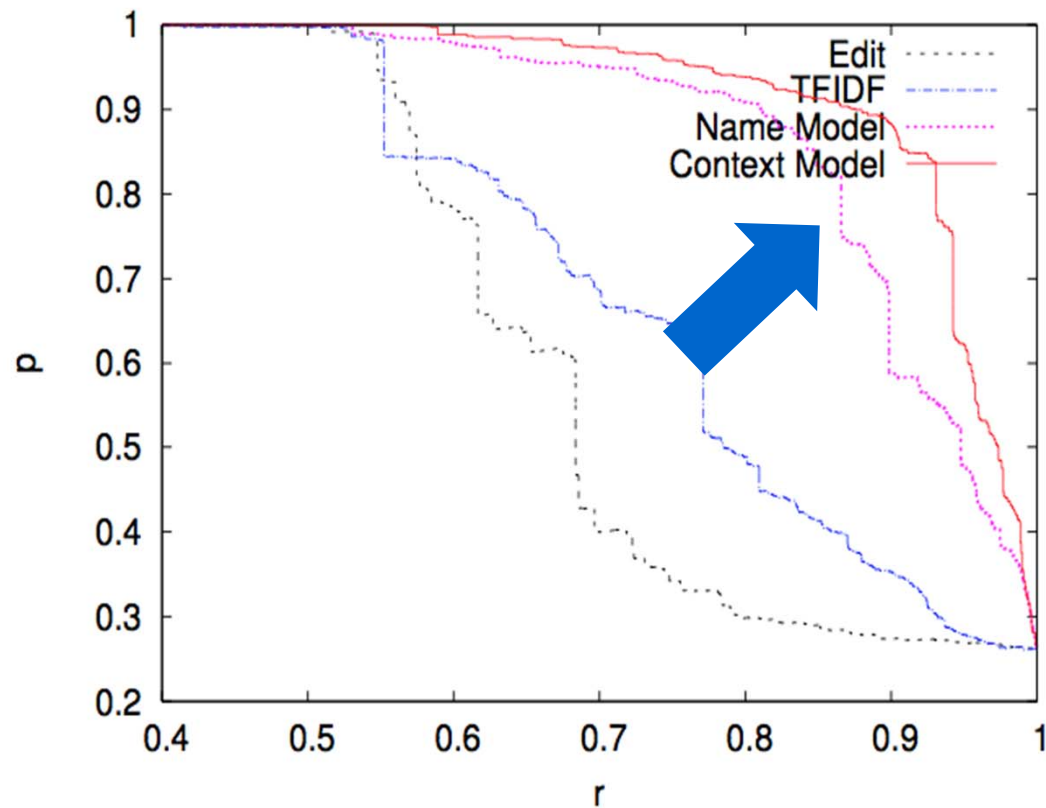
Times Square

Convert into an edit distance

- We match N_1 with N_2 given:
 - Core words model
 - Spatial contextual model
- Treat N_1 , N_2 as bag of words, and require:
 - Core words match
 - Any words that match are either core or background in both N_1 and N_2
- Extend this to Levenshtein-like edit distance

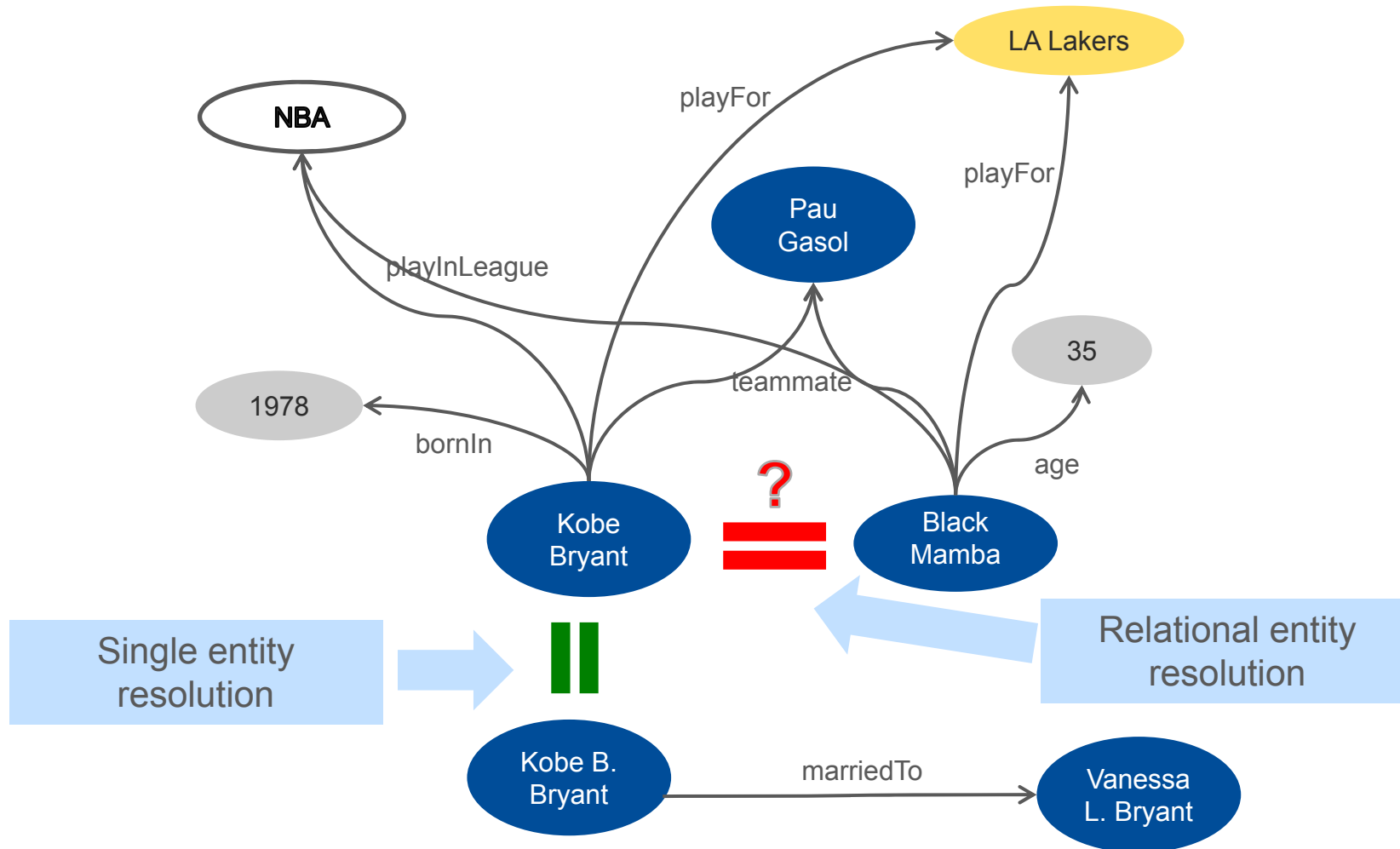


Deduplication results [Dalvi et al., 14]



- Edit: Levenshtein distance between place names
- TF-IDF: cosine similarity of TF-IDF weighted vector of overlapping names

Entity resolution

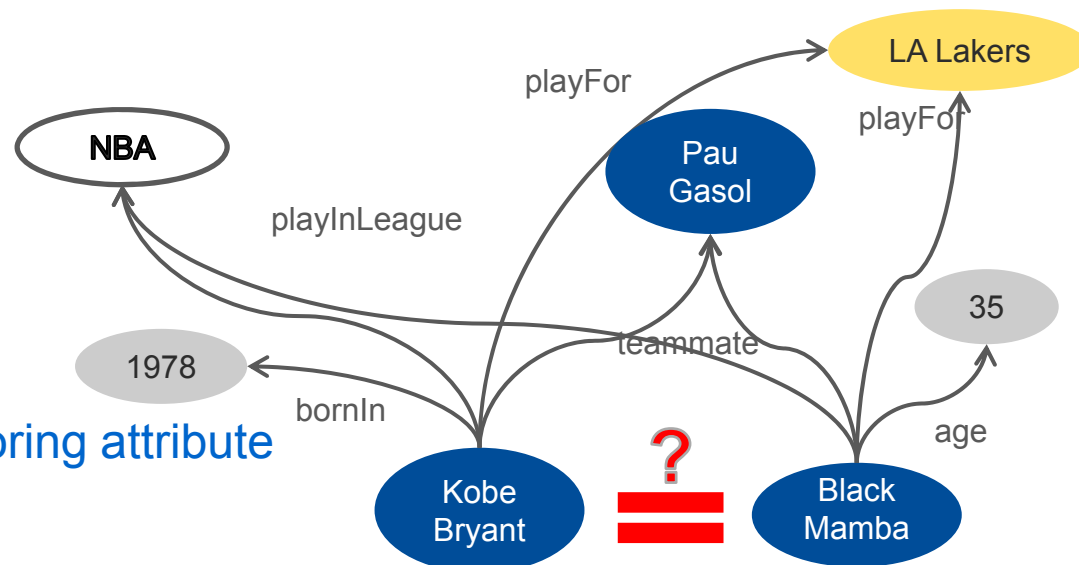


Relational entity resolution – Simple strategies

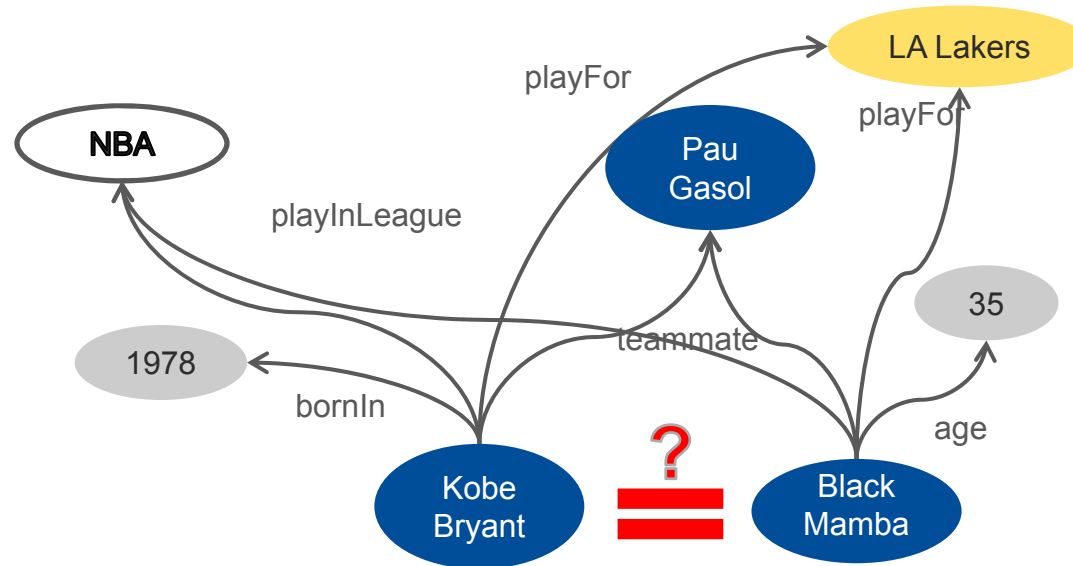
- Enrich model with **relational features** → **richer context** for matching

- **Relational features:**

- Value of **edge or neighboring attribute**
- Set **similarity measures**
 - *Overlap/Jaccard*
 - *Average similarity* between set members
 - *Adamic/Adar*: two entities are more similar if they share more items that are overall less frequent
 - *SimRank*: two entities are similar if they are related to similar objects
 - *Katz score*: two entities are similar if they are connected by shorter paths



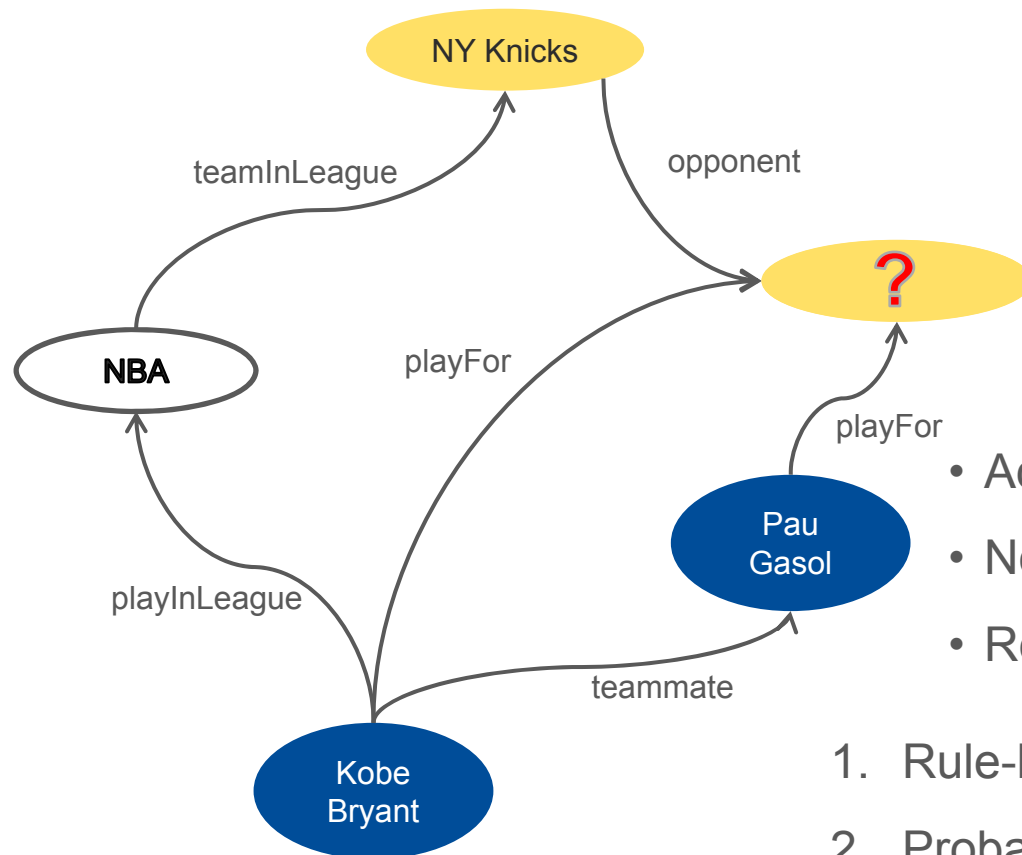
Relational entity resolution – Advanced strategies



- Dependency graph approaches [Dong et al., 05]
- Relational clustering [Bhattacharya & Getoor, 07]
- **Probabilistic Relational Models** [Pasula et al., 03]
- **Markov Logic Networks** [Singla & Domingos, 06]
- **Probabilistic Soft Logic** [Broecheler & Getoor, 10]

LINK PREDICTION

Link prediction

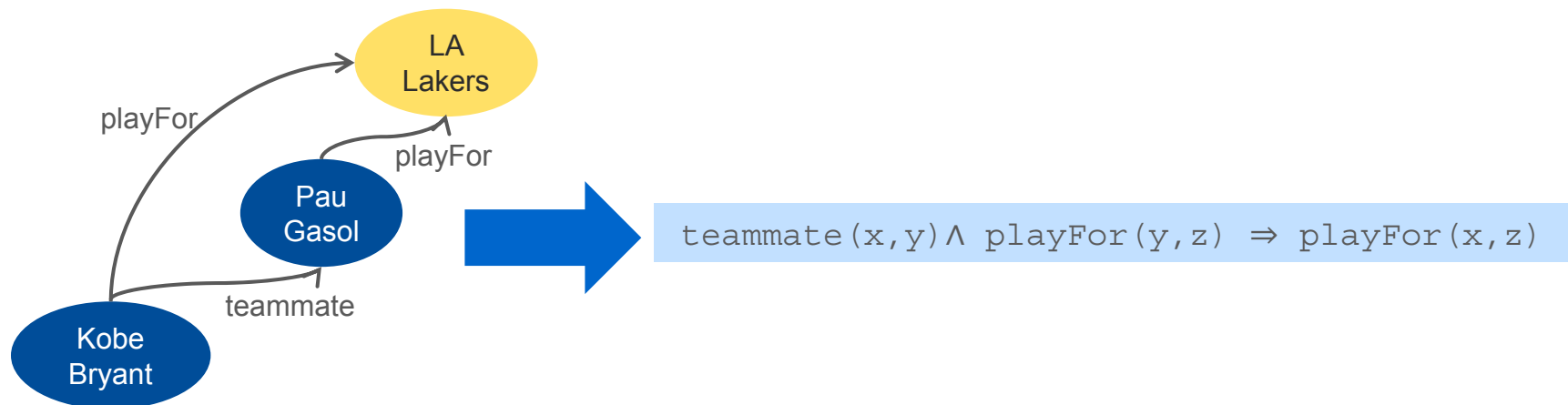


- Add knowledge from existing graph
- No external source
- Reasoning within the graph

1. Rule-based methods
2. Probabilistic models
3. Factorization models
4. Embedding models

First Order Inductive Learner

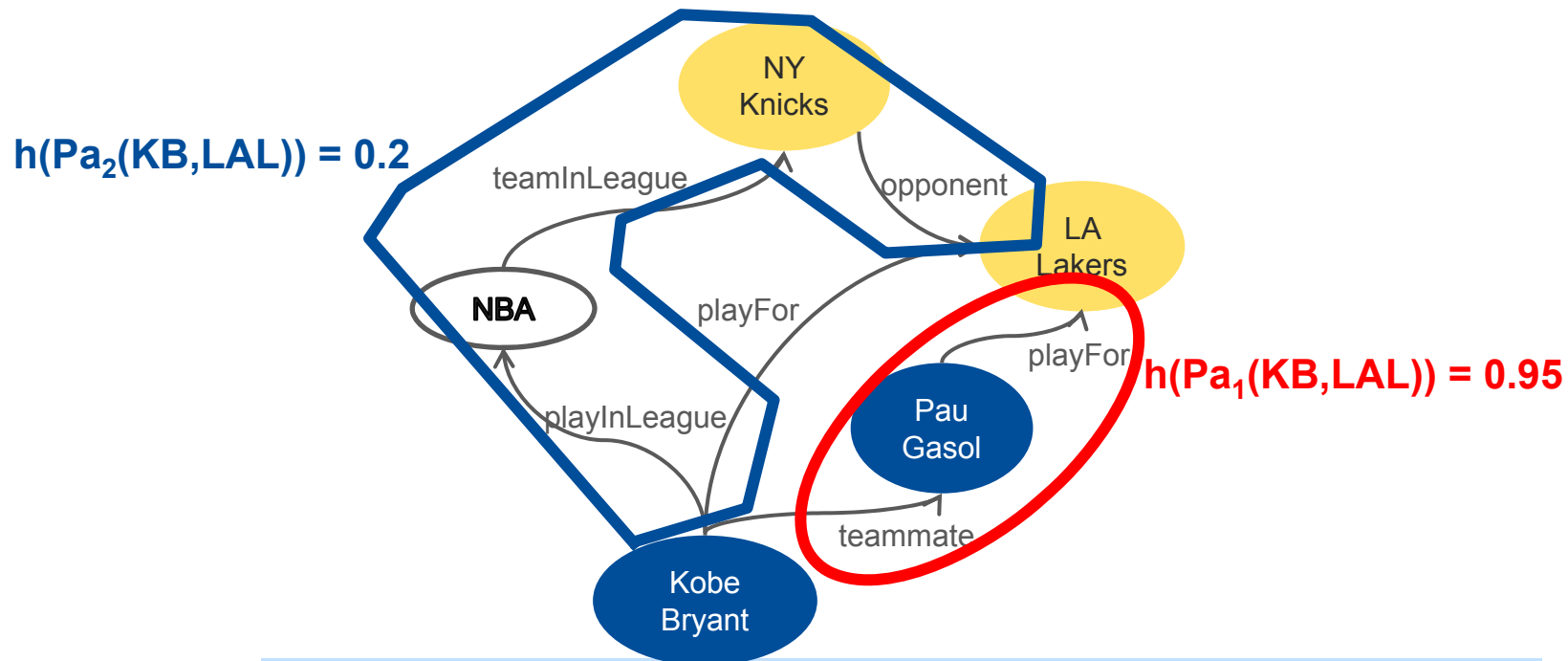
- FOIL learns function-free Horn clauses:
 - given positive negative examples of a concept
 - a set of background-knowledge predicates
 - FOIL inductively generates a logical rule for the concept that cover all + and no -



- Computationally expensive:** huge search space large, costly Horn clauses
- Must **add constraints** \rightarrow high precision but low recall
- Inductive Logic Programming: **deterministic and potentially problematic**

Path Ranking Algorithm [Lao et al., 11]

- Random walks on the graph are used to **sample paths**
- Paths are weighted with **probability of reaching target from source**
- Paths are used as ranking experts in a scoring function



$$S(KB, playFor, LAL) = \sum_{i \in paths} \theta_{playFor}^i h(pa_i(KB, LAL))$$

Link prediction with scoring functions

- A scoring function alone does not grant a decision
- **Thresholding:** determine a threshold θ

$(KB, \text{playFor}, LAL)$ is *True* iff $S(KB, \text{playFor}, LAL) > \theta$

- **Ranking:**
 - The most likely relation between **Kobe Bryant** and **LA Lakers** is:

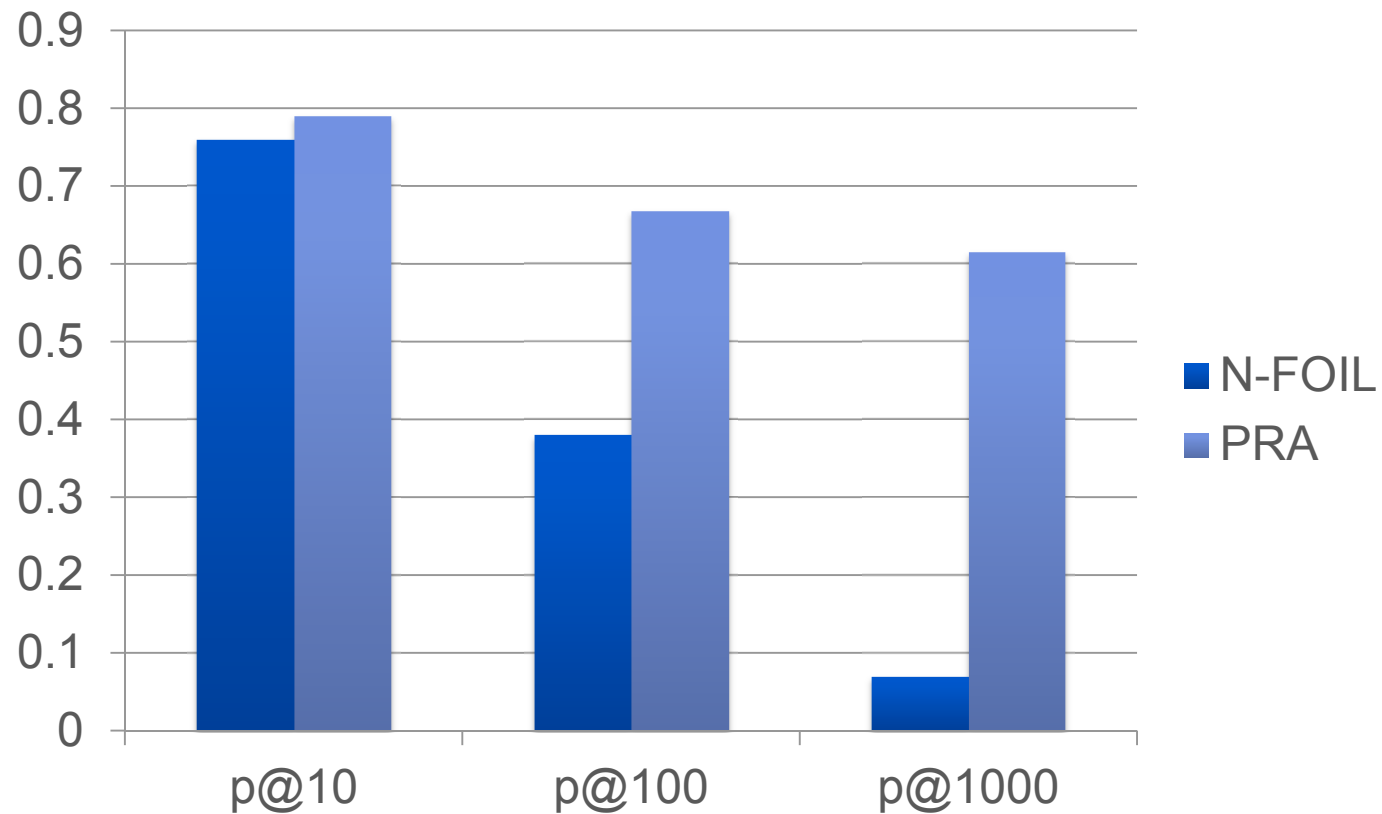
$$rel = \arg \max_{r' \in rels} S(KB, r', LAL)$$

- The most likely **team** for **Kobe Bryant** is:

$$obj = \arg \max_{e' \in ents} S(KB, \text{playFor}, e')$$

- **As prior** for extraction models (cf. Knowledge Vault)
- **No calibration of scores** like probabilities

Random walks boost recall



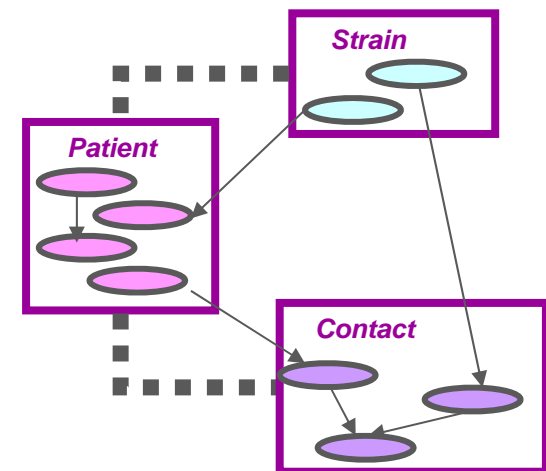
Precision of generalized facts for three levels of recall (Lao et al. 11)

Probabilistic Relational Models [Friedman et al., 99]

- **Probabilistic Relational Models** are directed graphical models that can handle link and feature uncertainty
- Probabilistic inference to predict links but also **duplicates, classes, clusters, etc.** based on **conditional probability distributions**

- **Limitations:**

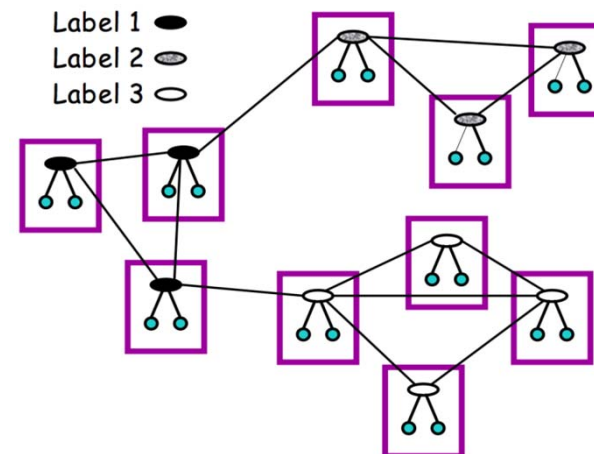
- **Careful construction:** must avoid cycles
- Generative process that **models both observations and unknowns**
- **Tractability** issues



Relational Markov Networks [Taskar et al., 02]

- Discriminative model: performs inference over the unknowns only

- Discriminant function: $P(X = x) = \frac{1}{Z} \exp(\sum_i w_i f_i(x))$



- Drawbacks:**
 - 1 feature for each state of each clique (**large**)
 - MAP estimation with belief propagation (**slow**)

Markov Logic Networks [Richardson & Domingos, 06]

- Knowledge graph = set of **hard constraints** on the set of possible worlds
 - Markov logic make them **soft constraints**
 - When a world violates a formula, it becomes **less probable but not impossible**

• Formulas

- **Constants**: KB, LAL, NBA
- **Variables**: x, y ranging over their domains (person, team, etc.).
- **Predicates**: `teammate(x, y)`
- **Atom**: `teammate(KB, x)`
- **Ground atom**: `teammate(KB, PG)`

Number of true groundings of formula i

Weight of formula i

- A **Markov Logic Network** (w, F) is a set of weighted first-order formulas
 - Probability of a grounding x :
 - **Higher weight** \square **stronger constraint**

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_{i \in F} w_i n_i(x)\right)$$

Probabilistic Soft Logic [Bach et al., 13]

- Framework where rules have continuous truth values
- Atoms like `teammate(KB, x)` are continuous random variables
- Each predicate has a weight like in MLNs
- Probability of a grounding:

The diagram illustrates the formula for the probability density over an interpretation I . The formula is $f(I) = \frac{1}{Z} \exp[-\sum_{r \in R} \lambda_r (d_r(I))^p]$. Callouts explain the components: 'Rule's weight' points to λ_r ; 'Rule's distance to satisfaction' points to $d_r(I) = \max(0, I_{r, \text{body}} - I_{r, \text{head}})$; 'Probability density over interpretation I' points to $f(I)$; 'Normalization constant' points to Z ; 'Set of ground rules' points to R ; and 'Distance exponent in {1, 2}' points to p .

$$f(I) = \frac{1}{Z} \exp[-\sum_{r \in R} \lambda_r (d_r(I))^p]$$

Rule's weight

Rule's distance to satisfaction
 $d_r(I) = \max(0, I_{r, \text{body}} - I_{r, \text{head}})$

Probability density over interpretation I

Distance exponent in {1, 2}

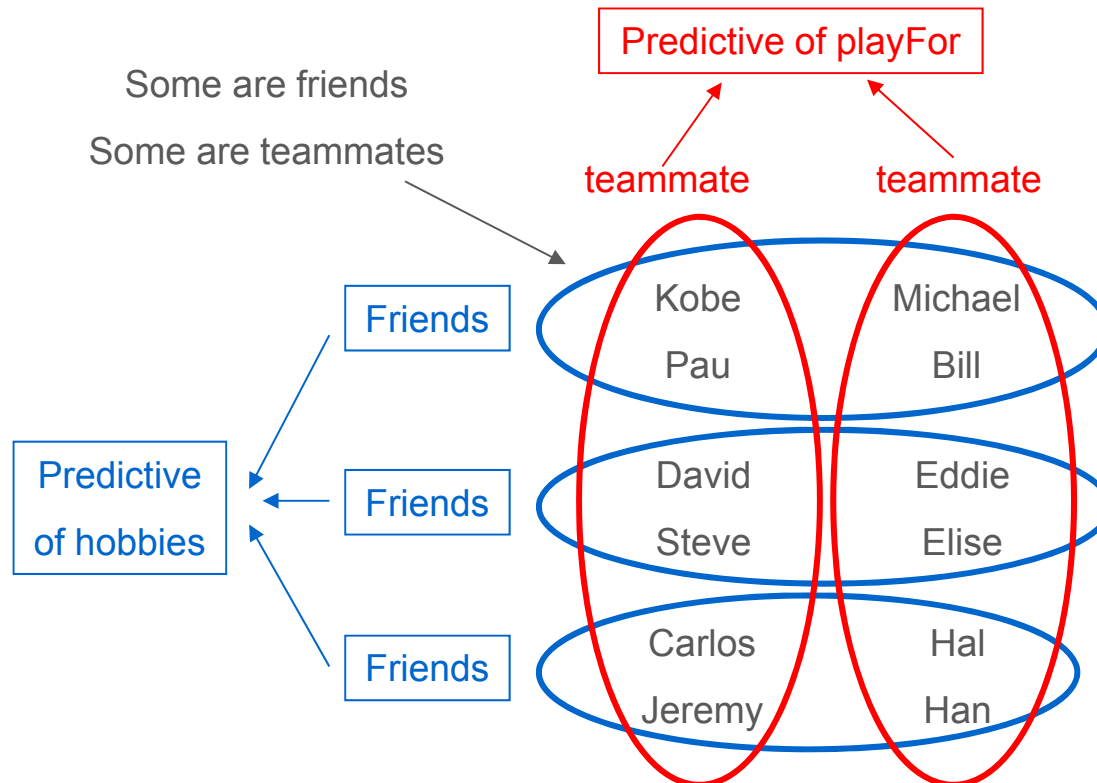
Normalization constant

Set of ground rules

- Inference is very tractable: convex optimization problem.

Multiple Relational Clustering [Kok & Domingos, 07]

- Hypothesis:** **multiple clusterings** are necessary to fully capture the interactions between entities



Multiple Relational Clustering [Kok & Domingos, 07]

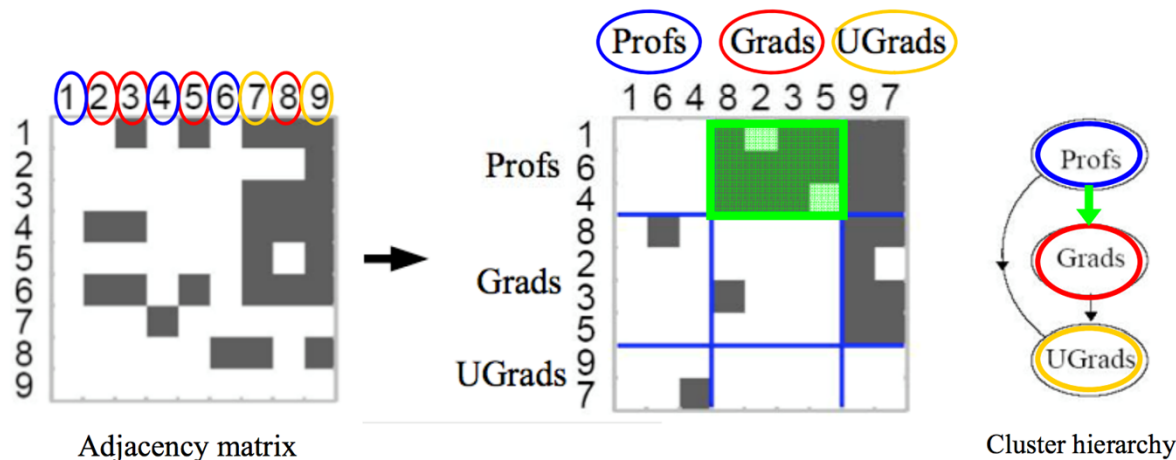
- **Markov Logic framework:**
 - Create an unary predicate for each cluster e.g. `cluster22(x)`
 - Multiple partitions are learnt together
 - Use connections:
 - Cluster relations by entities they connect and vice versa
 - Use types:
 - Cluster objects of same type
 - Cluster relations with same arity and argument types
- Learning by greedy search and multiple restarts maximizing posterior
- Link prediction is determined by evaluating truth value of grounded atoms such as `playFor(KB, LAL)`

Stochastic Blockmodels [Wang & Wong, 87]

- **Blockmodels:** learn partitions of entities and of predicates
 - **Partition entities/relations** into subgroups based on equivalence measure.
 - For each pair of positions **presence or absence of relation**.
 - **Structural equivalence:** entities are structurally equivalent if they have identical relations to and from all the entities of the graph
- **Stochastic blockmodels:**
 - Underlying probabilistic model
 - **Stochastic equivalence:** two entities or predicates are stochastically equivalent if they are “**exchangeable**” w.r.t. the **probability distribution**

Infinite Relational Models [Kemp et al., 05]

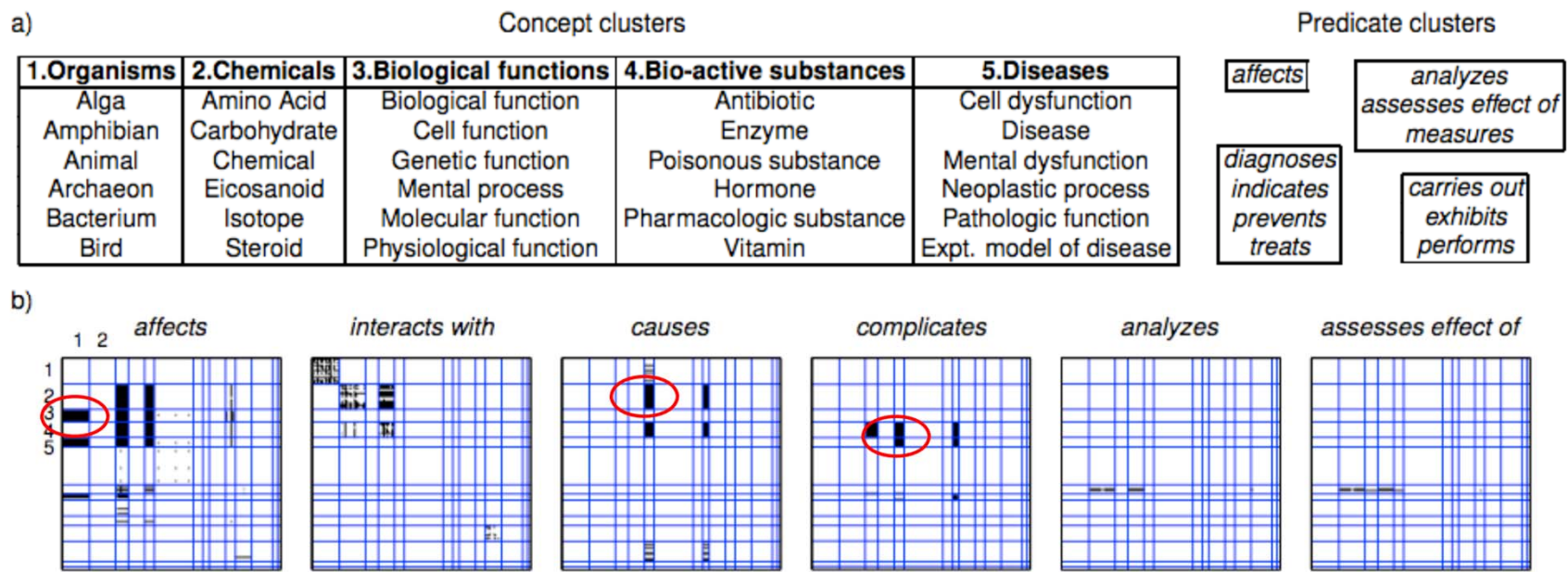
- **Infinite:** number of clusters increases as we observe more data
- **Relational:** it applies to relational data



- Prior assigns a probability to all possible partitions of the entities
- Allow number of clusters to adjust as we observe more data
- **Chinese Restaurant Process:** each new object is assigned to an existing cluster with probability proportional to the cluster size.

Example

- Semantic network with 135 concepts and 49 binary predicates.
- Finds 14 entities clusters and 21 predicate clusters



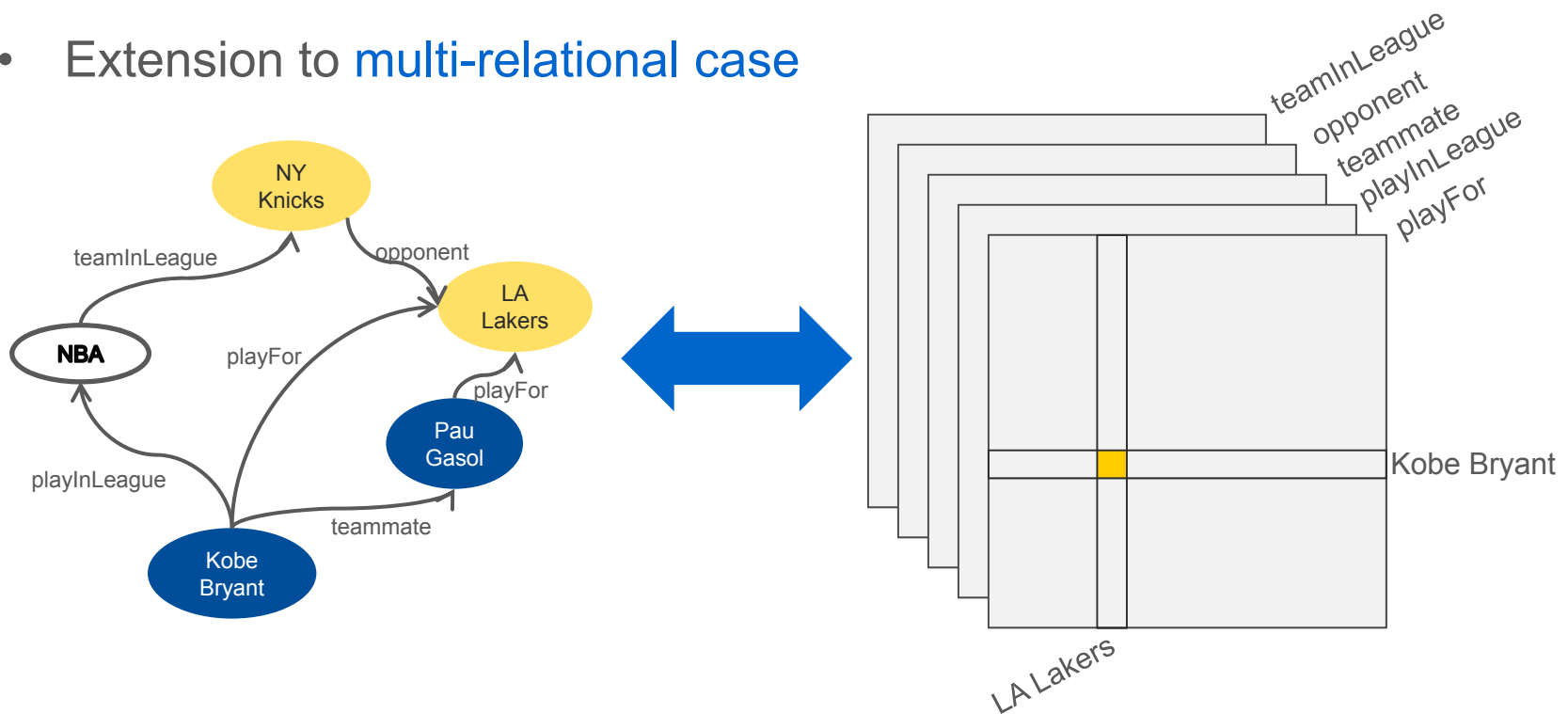
- Scalability issues with very large knowledge graphs

Variants of SBMs

- Mixed membership stochastic block models [Airoldi et al., 08]
- Nonparametric latent feature relational model [Miller et al., 09]
- Hybrid with **tensor factorization** [Sutskever et al., 09]

Factorization methods

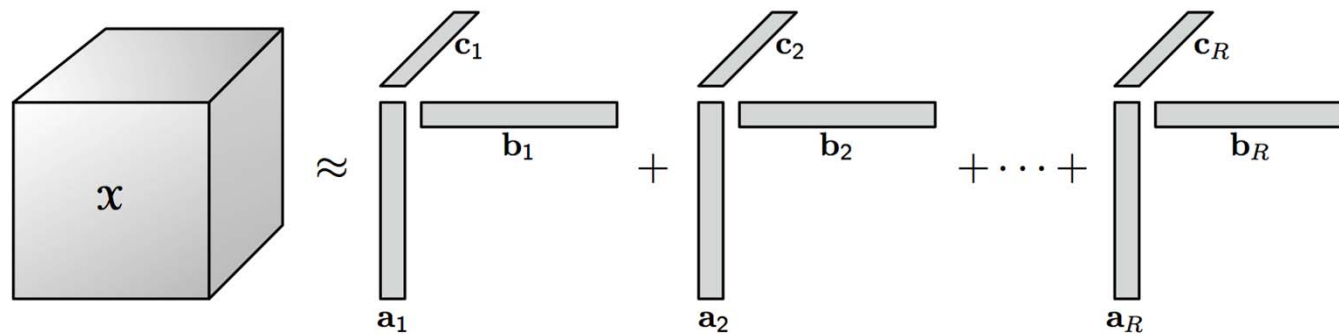
- **Matrix factorization** is successful: collaborative filtering, recommendation, etc.
- Extension to **multi-relational case**



- **Collective matrix factorization** or tensor factorization

Tensor factorization

- Many methods available: PARAFAC, Tucker, DEDICOM, etc.
- Example of **PARAFAC** [Harschman, 70]



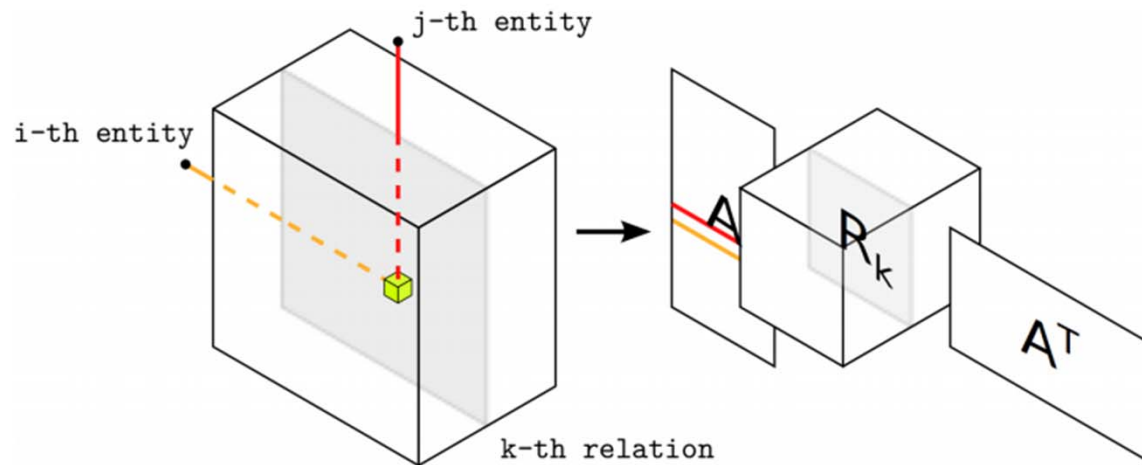
- Decomposition as a **sum of rank-one tensors**

$$S(KB, playFor, LAL) = \sum_{i=1}^R a_{KB}^i b_{LAL}^i c_{playFor}^i$$

- A , B and C are learned by alternating least squares
- **Does not take advantage of the symmetry of the tensor**

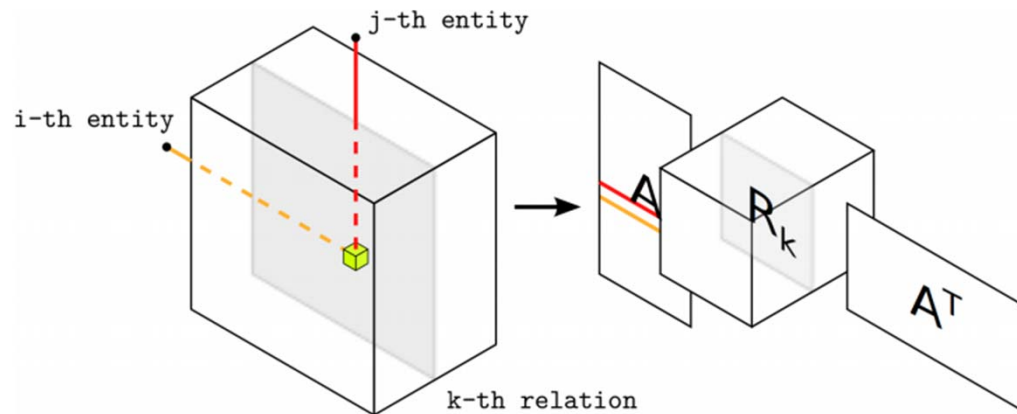
RESCAL [Nickel et al., 11]

- Collective matrix factorization inspired by DEDICOM



- A single matrix **A** stores latent representations of entities (vectors)
- Matrices **R_k** store latent representations of relations
- Score: $S(KB, playFor, LAL) = a_{KB} R_{playFor} a_{LAL}^T$

RESCAL [Nickel et al., 11]

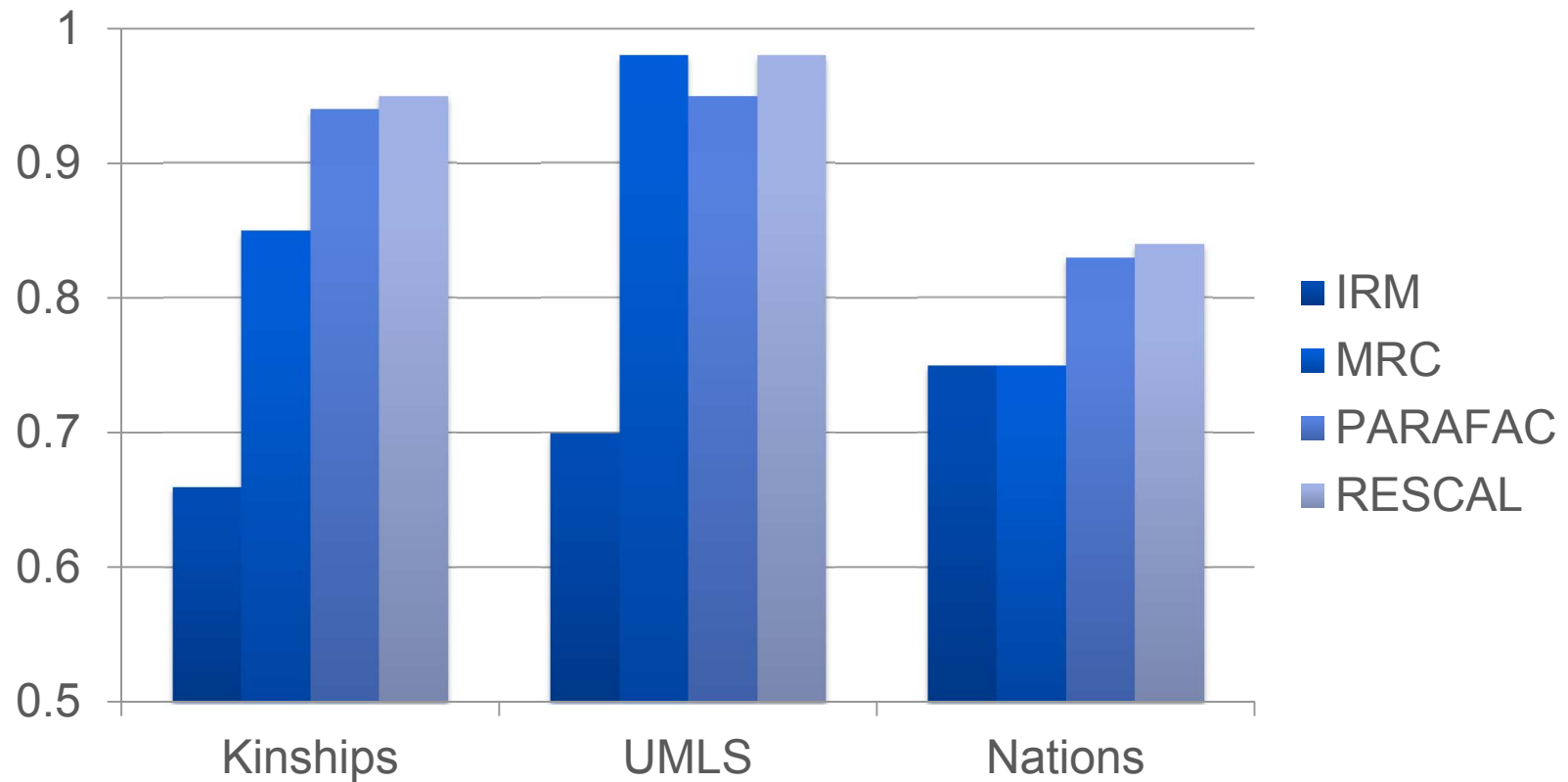


- Training with **reconstruction objective**:

$$\min_{A,R} \frac{1}{2} \left(\sum_k \|X_k - AR_k A^T\|_F^2 \right) + \lambda_A \|A\|_F^2 + \lambda_R \sum_k \|R_k\|_F^2$$

- Optimization with **alternating least squares** on A and R_k
- Faster than PARAFAC.

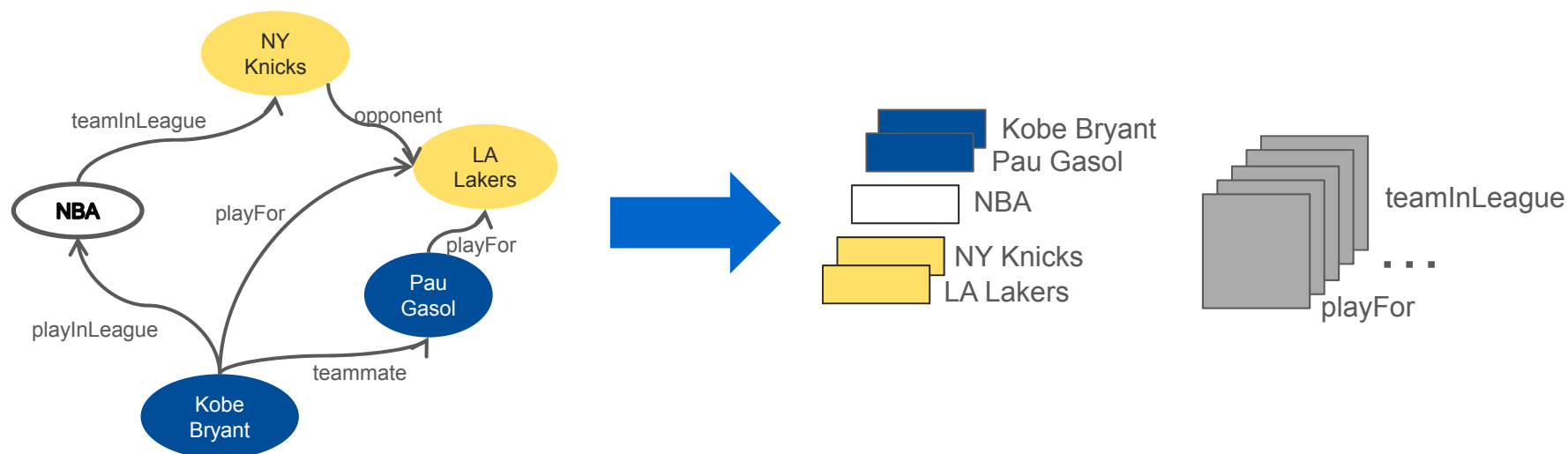
Factorization outperforms clustering



F1-score in link prediction on 3 benchmarks (Nickel et al. 11)

Embedding models

- Related to Deep Learning methods
- Entities are vectors (low-dimensional sparse)
- Relation types are operators on these vectors



- Embeddings trained to define a **similarity score** on triples such that:

$$S(KB, playFor, LAL) > S(KB, playFor, NYK)$$

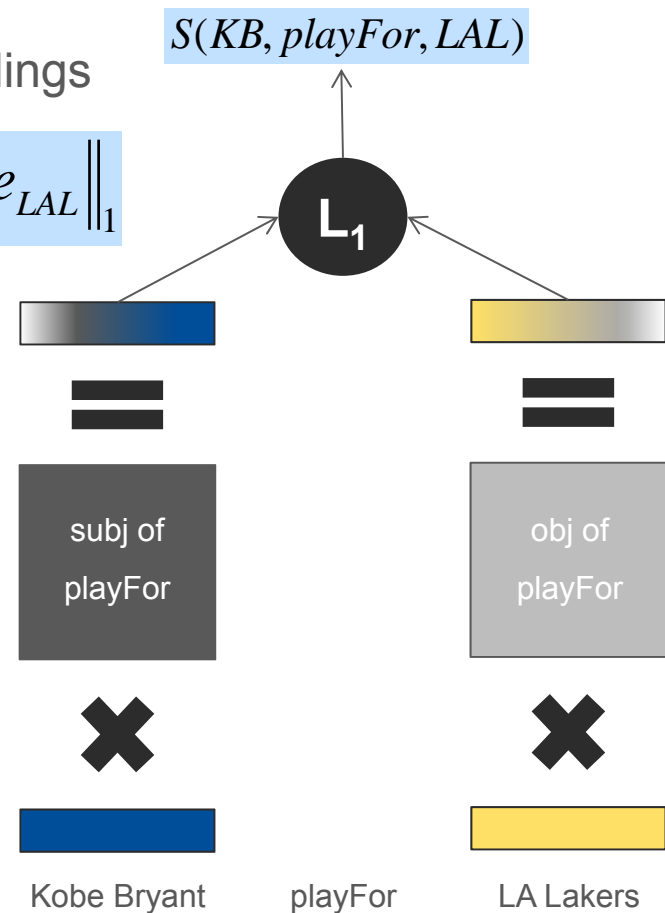
Training embedding models

- Training by ranking triples from the KG vs negative (generated)
- For each triple from the training set such as `(KB, playFor, LAL)`:
 1. Unobserved facts (false?) are sub-sampled:
 - `(Kobe Bryant, opponent, LA Lakers)`
 - `(Kobe Bryant, playFor, NY Knicks)`
 - `(NBA, teammate, LA Lakers)`
 - Etc...
 2. It is checked that the similarity score of the true triple is lower:
$$S(KB, playFor, LAL) > S(KB, playFor, NYK) + 1$$
 3. **If not**, parameters of the considered triples are updated.
- Optimization via Stochastic Gradient Descent

Structured Embeddings [Bordes et al., 11]

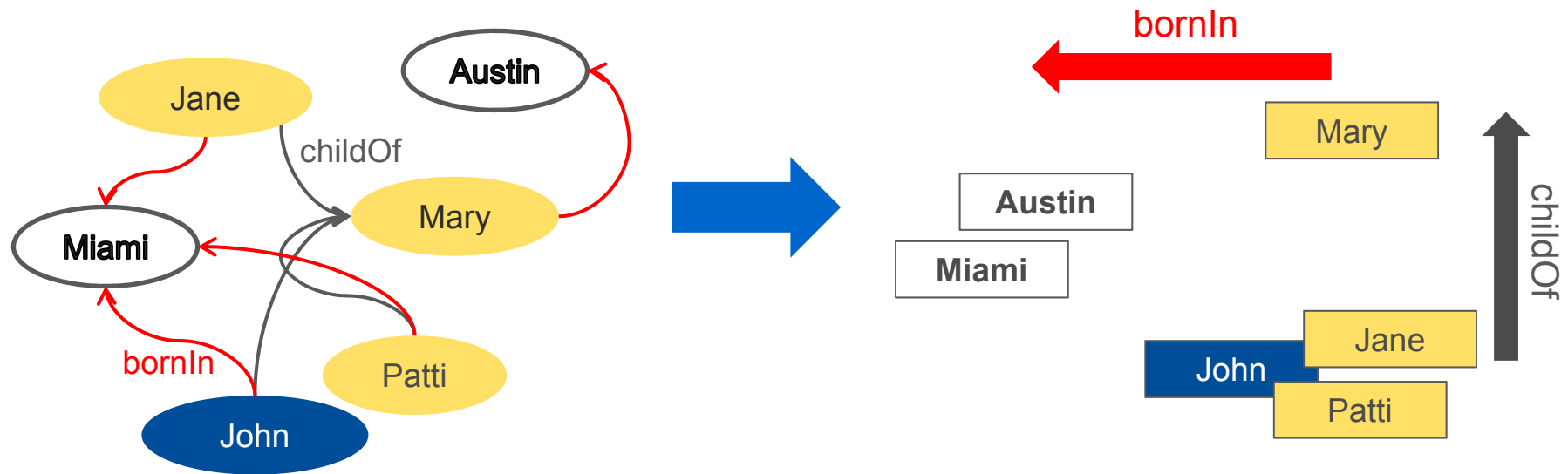
- Each entity = 1 vector
- Each relation = 2 matrices
- Score: L1 distance between projected embeddings

$$S(KB, playFor, LAL) = \left\| M_{playFor}^{sub} e_{KB} - M_{playFor}^{obj} e_{LAL} \right\|_1$$



Translating Embeddings [Bordes et al. 13]

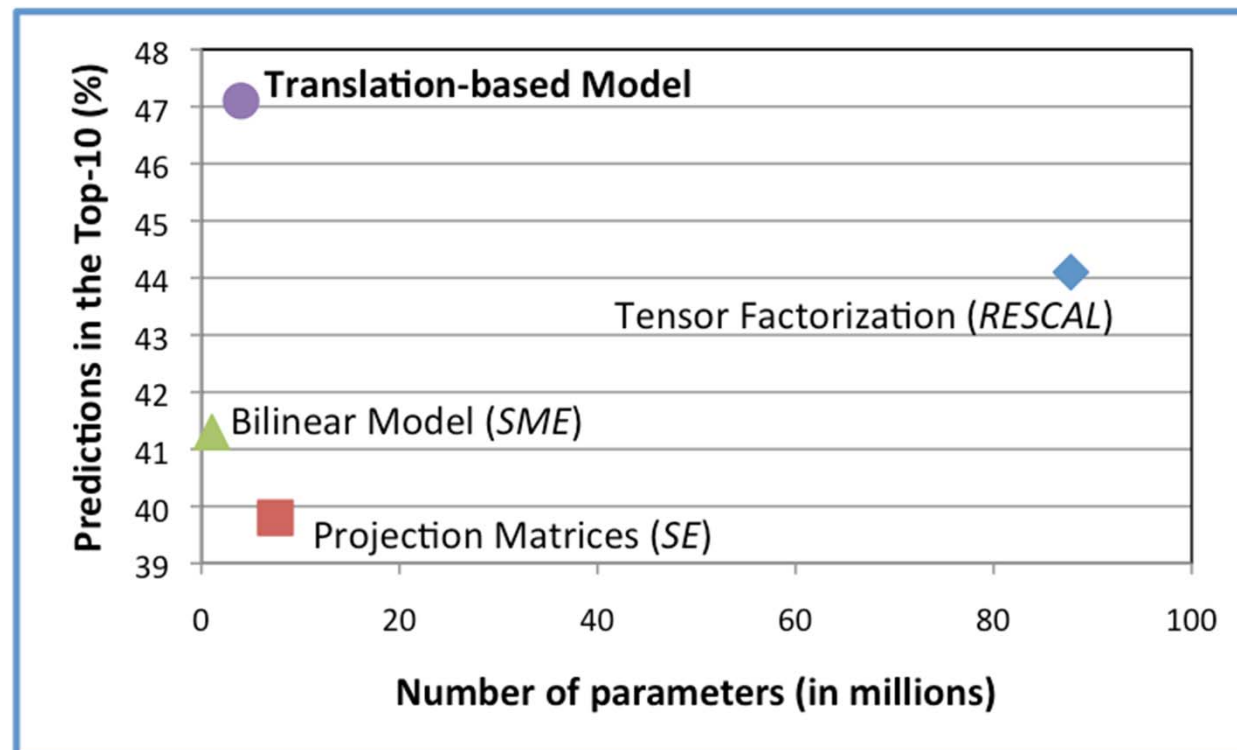
- Simpler model:* relation types are translation vectors



$$S(john, bornIn, miami) = \|e_{john} + e_{bornIn} - e_{miami}\|_2$$

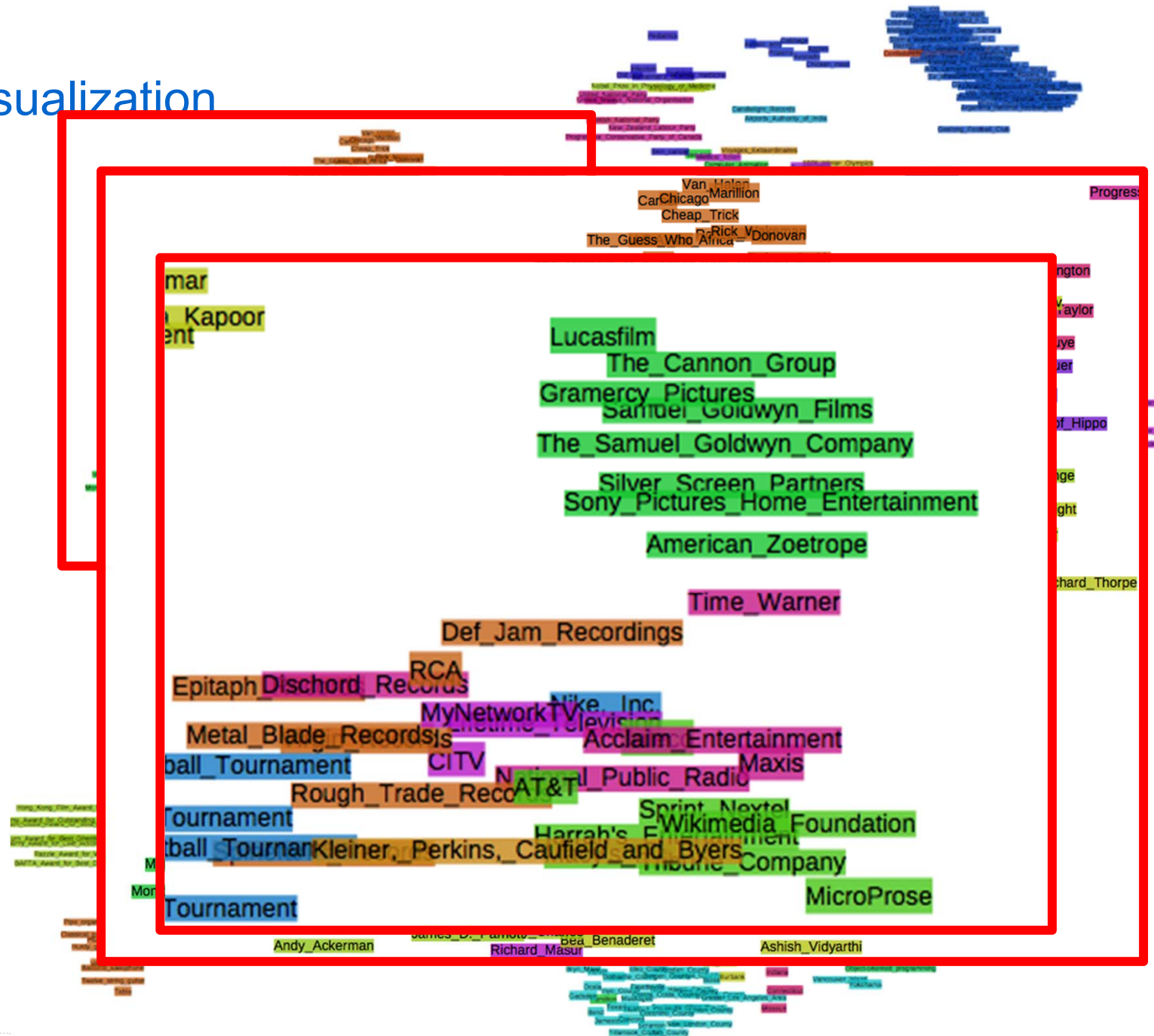
- Much fewer parameters (1 vector per relation).

The simpler, the better



Ranking object entities on a subset of Freebase [Bordes et al. 13]

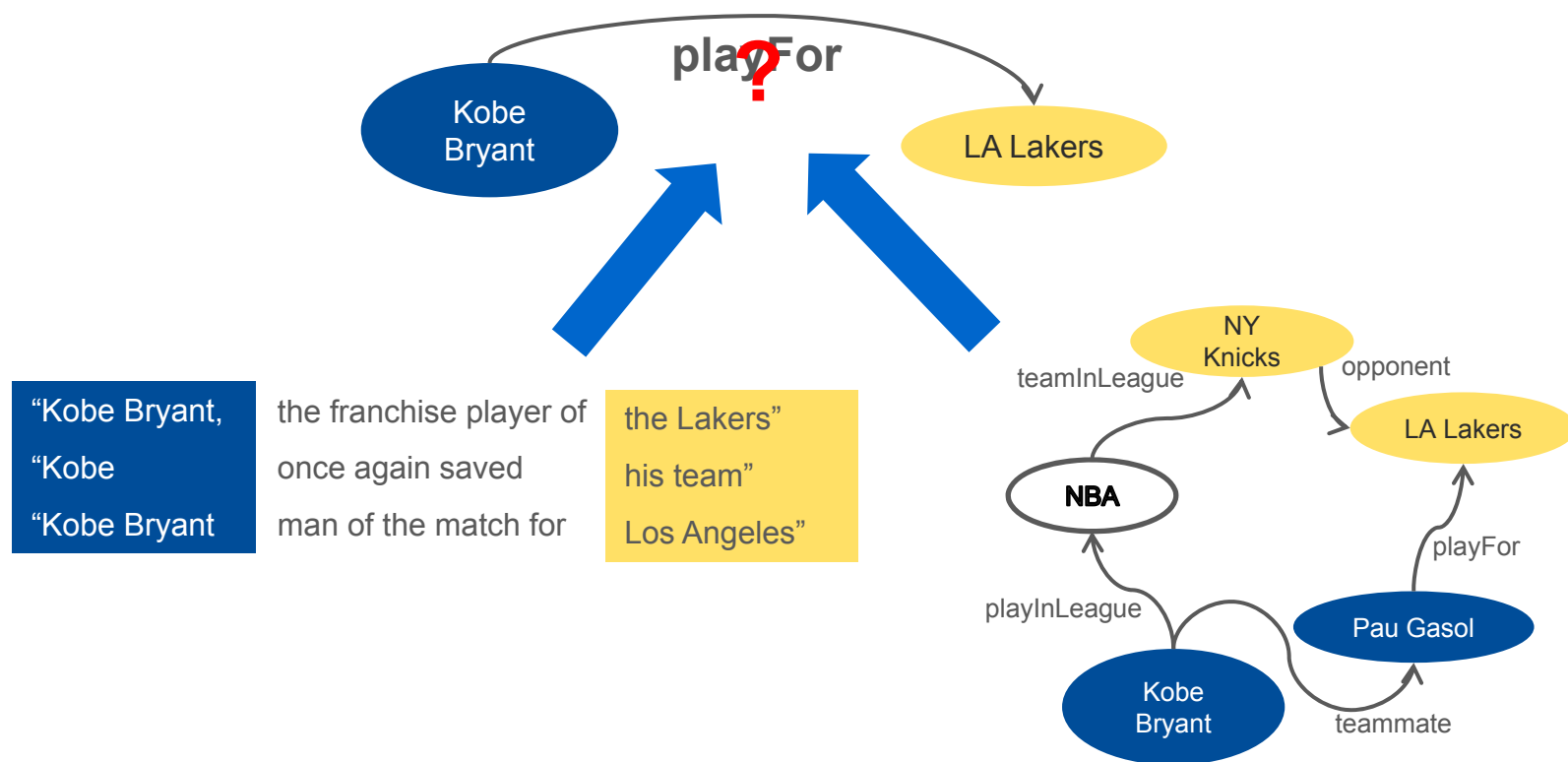
Visualization



Using knowledge graph and text together

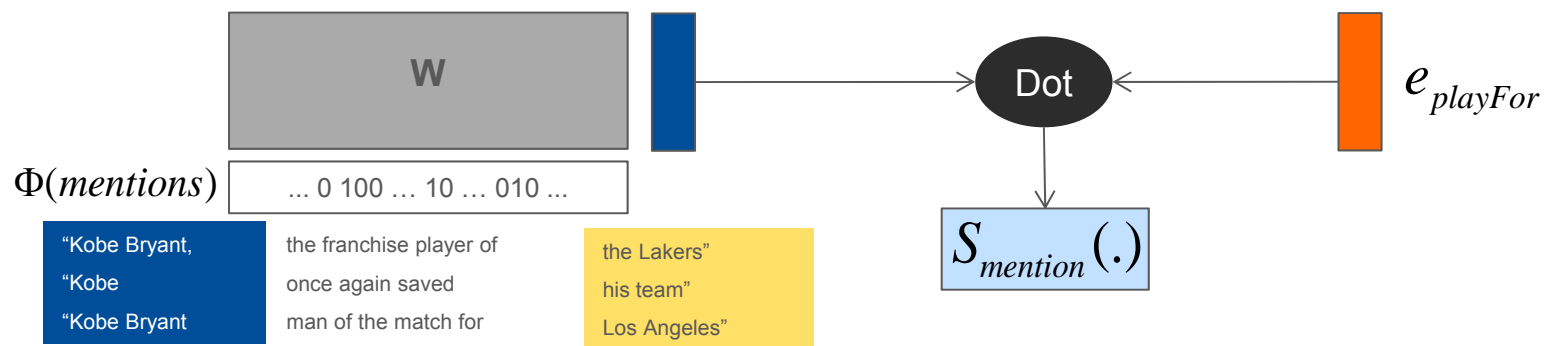
Why not merging **relation extraction** and **link prediction** in the same model?

Extracted facts should agree **both with the text and the graph!**

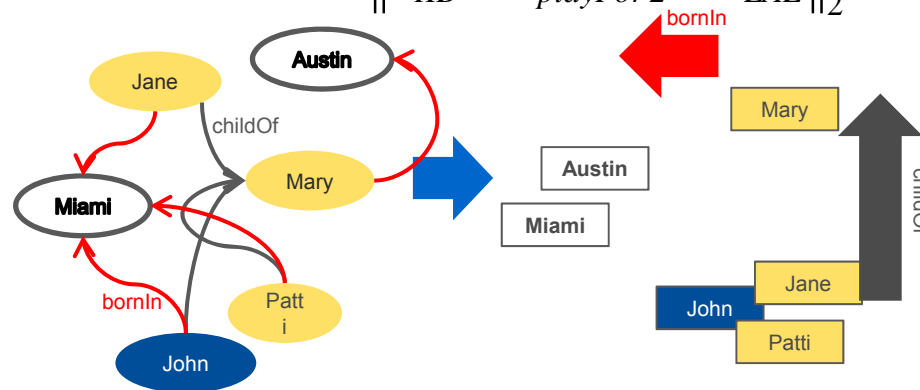


Joint embedding models [Bordes et al., 12; Weston et al., 13]

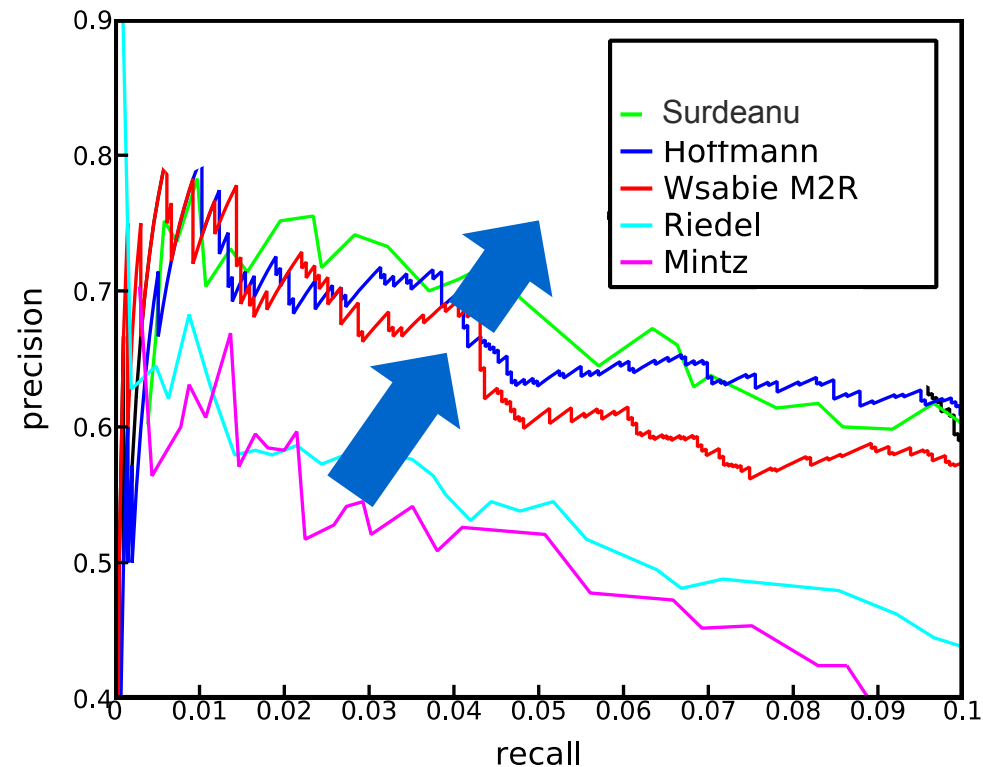
- Combination of two scores: $S(.) = S_{text}(.) + S_{FB}(.)$ (trained separately)
- $S_{text}(KB, playFor, LAL) = \langle W^T \Phi(m), e_{playFor1} \rangle$ inspired by WSABIE (Weston et al., 10)



- $S_{FB}(KB, playFor, LAL) = \|e_{KB} + e_{playFor2} - e_{LAL}\|_2$ (translating embeddings)



Using stored information improves precision even more



Precision-recall curves on extracting from New York Times articles to Freebase [Weston et al., 13]

Universal schemas [Riedel et al., 13]

- Join in a **single learning problem** link prediction and relation extraction
- The same model can score triples made of entities linked with:
 - **extracted surface forms from text**
 - **predicates from a knowledge base**

	<i>X-professor-at-Y</i>	<i>X-historian-at-Y</i>	<i>employee(X,Y)</i>	<i>member(X,Y)</i>	
Ferguson, Harvard		1	1	1	Train
Oman, Oxford	1	1			
Firth, Oxford	0.95	1	0.97	0.95	Test
Gödel, Princeton	1	0.05	0.93	0.97	
	Surface Patterns		KB Relations		

Universal schemas [Riedel et al., 13]

- Combination of three scores: $S(.) = S_{mention} (.) + S_{FB} (.) + S_{neighbors} (.)$

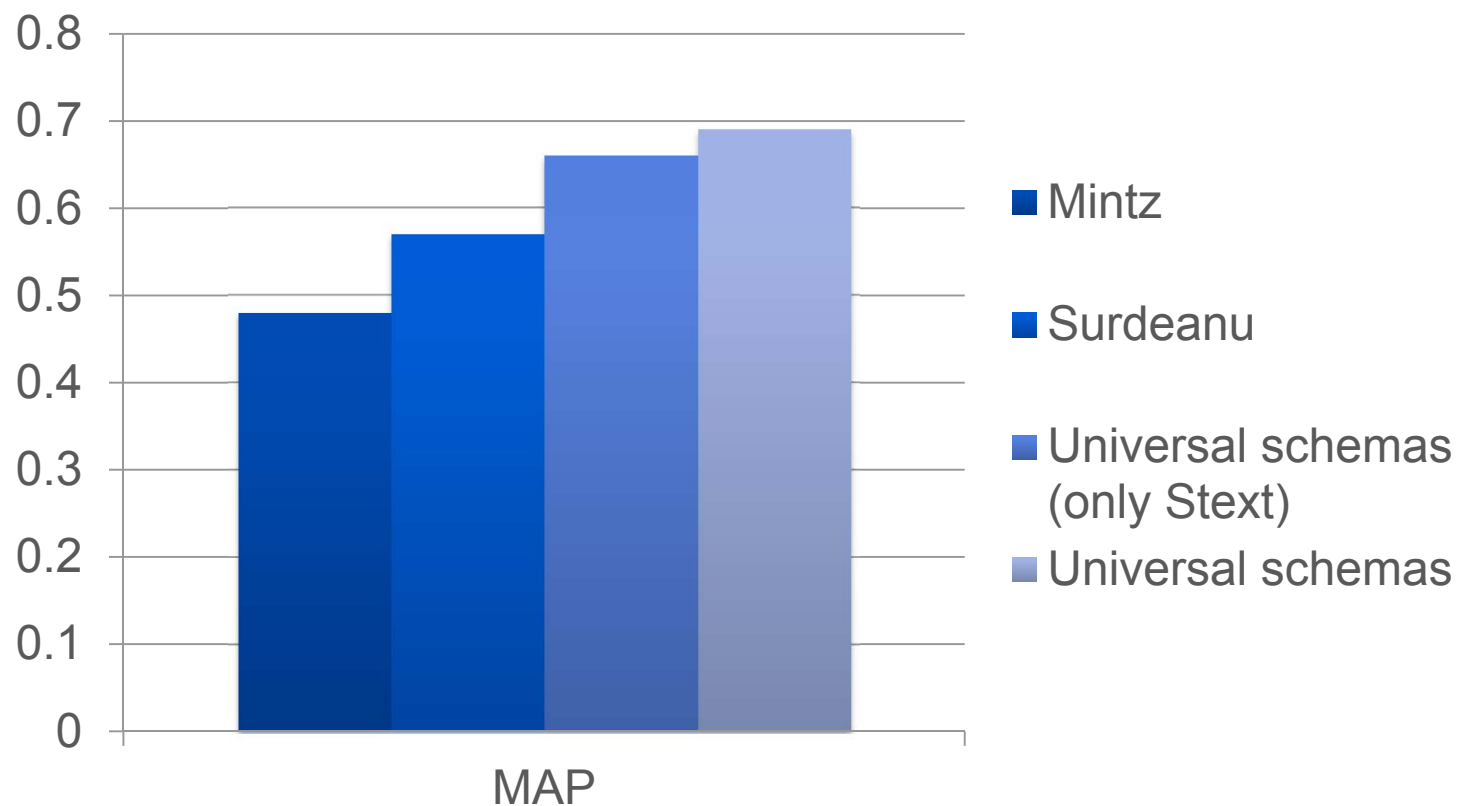
$$S_{mention}(KB, playFor, LAL) = \langle e_{mention}, e_{playFor1} \rangle$$

$$S_{FB}(KB, playFor, LAL) = \langle e_{playFor2}^{sub}, e_{KB} \rangle + \langle e_{playFor2}^{obj}, e_{LAL} \rangle$$

$$S_{neighbors}(KB, playFor, LAL) = \sum_{\substack{(KB, rel', LAL) \\ rel' \neq playFor}} w_{rel'}^{playFor}$$

- Embeddings for **entities, relations and mentions**.
- Training by **ranking observed facts versus others** and making updates using Stochastic Gradient Descent.

Using stored information (still) improves precision



Weighted Mean Averaged Precision on a subset of relations of Freebase [Riedel et al. 13]

RESOURCES

Related tutorial – here at KDD (later today) !

Bringing **Structure** to **Text**: Mining Phrases, Entity Concepts, Topics & Hierarchies

by Jiawei Han, Chi Wang and Ahmed El-Kishky

Today, 2:30pm

Relevant datasets

- Wikipedia
 - http://en.wikipedia.org/wiki/Wikipedia:Database_download
- Freebase
 - <https://developers.google.com/freebase/data>
- YAGO
 - <http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/>
- DBpedia
 - <http://wiki.dbpedia.org/Datasets>
- OpenIE/Reverb
 - <http://reverb.cs.washington.edu/>

Relevant competitions, evaluations, and workshops

- Knowledge Base Population (KBP) @ TAC
<http://www.nist.gov/tac/2014/KBP/>
- Knowledge Base Acceleration (KBA) @ TREC
<http://trec-kba.org/>
- Entity Recognition and Disambiguation (ERD) Challenge @ SIGIR 2014
<http://web-ngram.research.microsoft.com/erd2014/>
- INEX Link the Wiki track
http://link.springer.com/chapter/10.1007/978-3-642-23577-1_22
- CLEF eHealth Evaluation Lab
http://link.springer.com/chapter/10.1007/978-3-642-40802-1_24

Relevant competitions, evaluations, and workshops (cont'd)

- Named Entity Extraction & Linking (NEEL) Challenge (#Microposts2014)

<http://www.scc.lancs.ac.uk/microposts2014/challenge/>

- LD4IE 2014 Linked Data for Information Extraction

<http://trec-kba.org/>

Tutorials

- Entity linking and retrieval tutorial (Meij, Balog and Odijk)
 - <http://ejmeij.github.io/entity-linking-and-retrieval-tutorial/>
- Entity resolution tutorials (Getoor and Machanavajjhala)
 - http://www.umiacs.umd.edu/~getoor/Tutorials/ER_VLDB2012.pdf
 - <http://lings.cs.umd.edu/projects/Tutorials/ER-AAAI12/Home.html>
- Big data integration (Dong and Srivastava)
 - http://lunadong.com/talks/BDI_vldb.pptx
- Tensors and their applications in graphs (Nickel and Tresp)
 - <http://www.cip.ifi.lmu.de/~nickel/iswc2012-learning-on-linked-data/>
- Probabilistic soft logic (Bach et Getoor)
 - <http://psl.umiacs.umd.edu/>

Data releases from Google

1. Automatic annotation of ClueWeb09 and ClueWeb12 with Freebase entities (**800M documents, 11B entity mentions**)
2. Similar annotation of several TREC query sets (**40K queries**)
3. Human judgments of relations extracted from Wikipedia (**50K instances, 250K human judgments**)
4. Triples deleted from Freebase over time (**63M triples**)

Mailing list:

goo.gl/MJb3A



SUMMARY

Knowledge is crucial yet difficult to acquire

- Knowledge is **crucial** for many AI tasks
- Knowledge acquisition
 - From **experts**: slow and mostly reliable
 - From **non-experts**: faster and not always reliable
 - **Automated**: fastest and most scalable, yet noisiest
- Knowledge availability
 - **A lot** can be found online
 - **A lot** cannot be found
 - **A lot** cannot be extracted using today's methods

Where we are today

- We can **extract a lot of knowledge from text** and model its correctness
- **Enforcing structure** makes the extraction problem easier yet imposes limitations
- **Leveraging existing knowledge repositories** helps a lot

Next steps

- We need **new extraction methods**, from **new sources**
- Extracting from **modalities other than text** appears promising yet mostly unexplored

Plenty to be learned, problems are far from solved!

- Vibrant research area
- Numerous open research questions



*This is a perfect time
to work in this area!*