

From Big Data to Big Knowledge

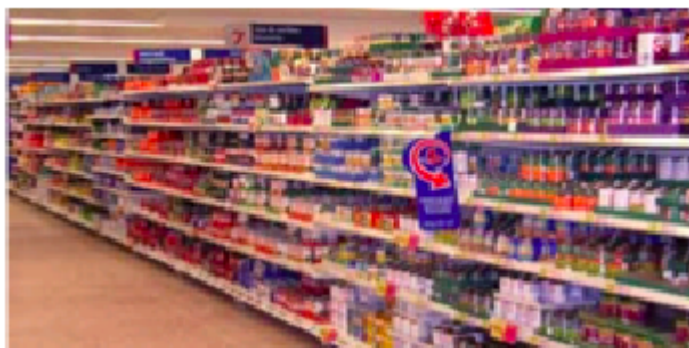
Kevin Murphy
Google Research
kpmurphy@google.com

*Joint work with Luna Dong, Evgeniy Gabrilovich,
Jeremy Heitz, Wilko Horn, Panos Ipeirotis, Ni Lao, Wei-
Lwun Lu, Thomas Strohmman, Shaohua Sun, Chun How
Tan, Robert West, Wei Zhang, and others*

Big Data is everywhere



dream



```

gacca-----
gaccacacga-----
aattcaggaccacacgacggaagagaca-----
-attcaggacaacacgaaggaagacaagttcatgtacttt
----caggaccacacgacggaagagacaagttcatgtacttt
-----accacacgacggaagagacaagttcatgtacttt
-----accacacgacggaagagacaagttcatgtacttt
-----gacgggagacaagttcatgtacttt
-----atgtacttt
    
```

From Big Data to Big Knowledge

We are drowning in information and starving for knowledge.

--- John Naisbitt.

- What does all this data “mean”?
- Words are ambiguous.
- e.g., “Taj Mahal”



- We need to move from “strings” to “things”.

Google's Knowledge Graph



- 500M nodes (entities)
- 3.5B edges (facts)
- 1500 node types
- 35k edge types
- **Extension of Freebase.com**

Knowledge Panels

The image shows a Google search interface for 'marie curie'. On the right side, a knowledge panel is displayed, providing a structured overview of her life and work. The panel includes a portrait of Marie Curie, a brief biographical text, and key facts such as her birth and death dates, her spouse, children, and education. Below the main text, there are sections for 'People also search for' and a 'Report a problem' link.

Marie Curie

Marie Skłodowska-Curie was a French-Polish physicist and chemist famous for her pioneering research on radioactivity. She was the first person honored with two Nobel Prizes—in physics and chemistry. [Wikipedia](#)

Born: November 7, 1867, [Warsaw](#)

Died: July 4, 1934, [Sancellemoz](#)

Spouse: [Pierre Curie](#) (m. 1895–1906)

Children: [Irène Joliot-Curie](#), [Ève Curie](#)

Discovered: [Radium](#), [Polonium](#)

Education: [École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris](#), [University of Paris](#)

People also search for

[Albert Einstein](#) [Pierre Curie](#) [Ernest Rutherford](#) [Louis Pasteur](#) [John Dalton](#)

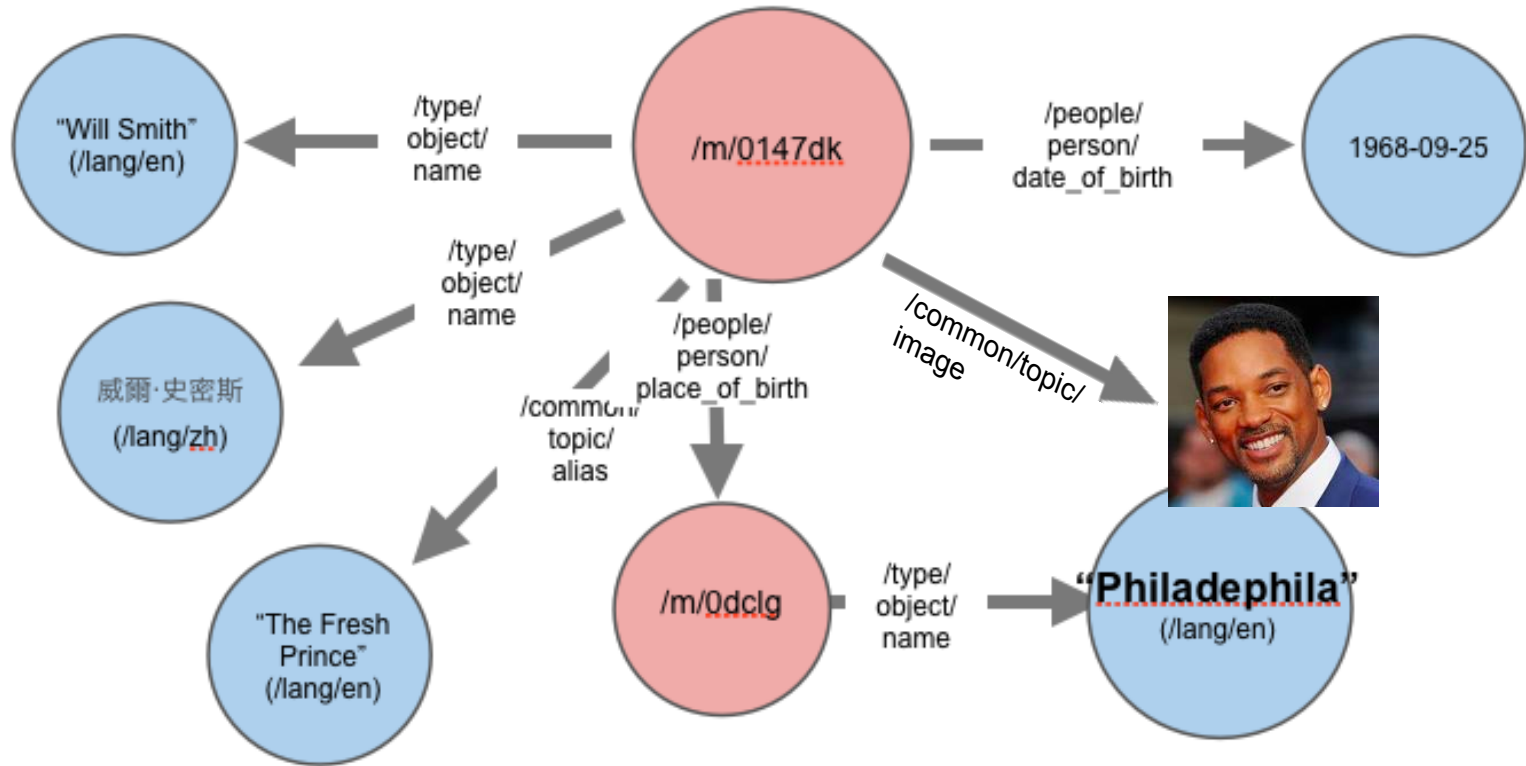
[Report a problem](#)

Freebase is created by merging many data sources



Massive entity linkage problem!

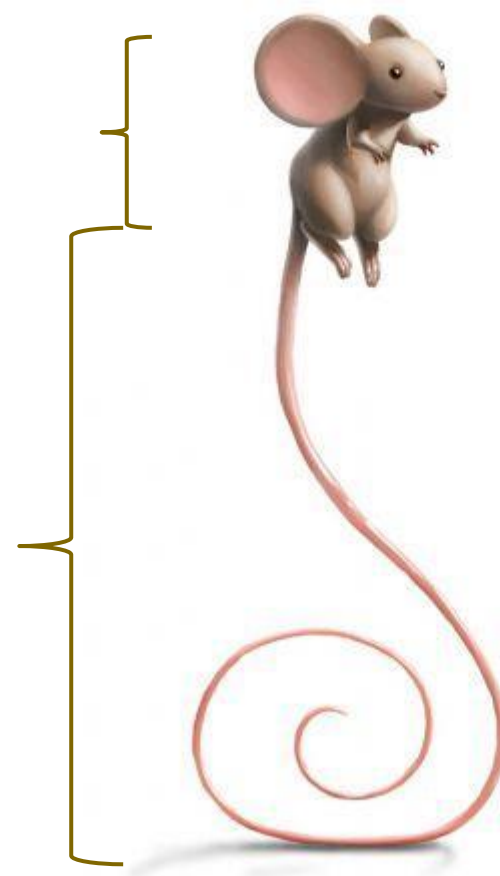
A fragment of Freebase (in RDF format)



- Freebase is large, but still very incomplete:

| Relation | % unknown in Freebase |
|----------------|--------------------------|
| Profession | 68% |
| Place of birth | 71% |
| Nationality | 75% |
| Education | 91% |
| Spouse | 92% |
| Parents | 94% |

- We need automatic knowledge base construction methods
 - cf AKBC workshop at CIKM.

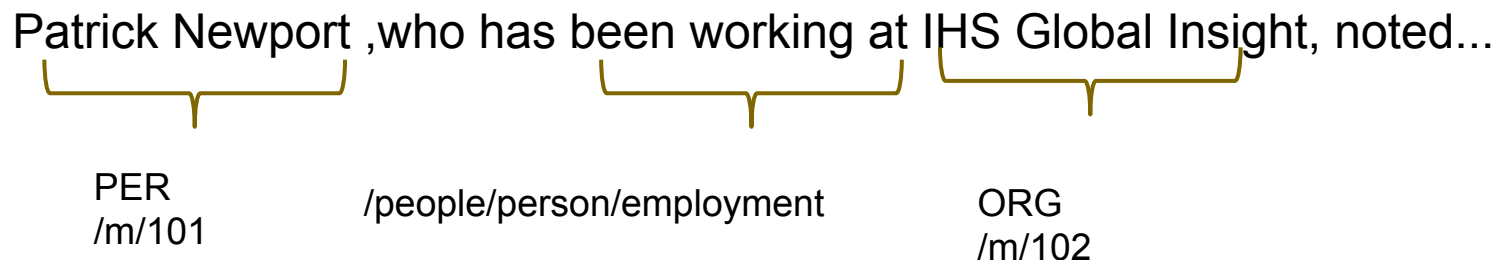


<http://www.flickr.com/photos/sandreli/4691045841/>

- From strings to things
- • Reading the web
- Asking the web
- Asking people
- Open issues

- There are many academic groups (e.g., CMU, UW, MPI) that have developed methods to extract facts from large text corpora.
- At Google, we have developed a similar system, except it is 10x bigger.
- In addition, we use “prior knowledge” to help reduce the error rate.

- Template matching methods



- Machine learning (binary classifiers trained on text / parse tree features)

$$\Pr(\text{predicate } p | \text{features } \mathbf{x}) = \frac{1}{1 + e^{-\theta^T f(\mathbf{x}, p)}}$$

Wrapper induction

① User labels data records in example pages

2. **Beginning XML, Second Edition**
by David Hunter, et al (Paperback - March 2003)
Avg. Customer Rating: ★★★★★
(Rate this item)

Usually ships in 24 hours
List Price: ~~\$30.00~~ Used & new from \$11.08
Buy new: \$27.19

② A set of extration rules are induced from the examples

1. **Head First Java, 2nd Edition**
by Kathy Sierra, Bert Bates (Paperback - February 9, 2005)
Avg. Customer Rating: ★★★★★
(Rate this item)

Usually ships in 24 hours
List Price: ~~\$44.00~~ Used & new from \$17.87
Buy new: \$29.87 In-stock Purchase: \$44.95

2. **Effective Java Programming Language Guide**
by Joshua Bloch (Paperback - June 5, 2001)
Avg. Customer Rating: ★★★★★
(Rate this item)

Usually ships in 24 hours
List Price: ~~\$44.00~~ Used & new from \$24.24
Buy new: \$29.99

WRAPPER

③ The induced wrapper can be used to extract data from pages following the same template

Title: Head First Java, 2nd Edition
Author: Kathy Sierra, Bert Bates
Format: Paperback
Date: February 9, 2005
Price: \$29.87

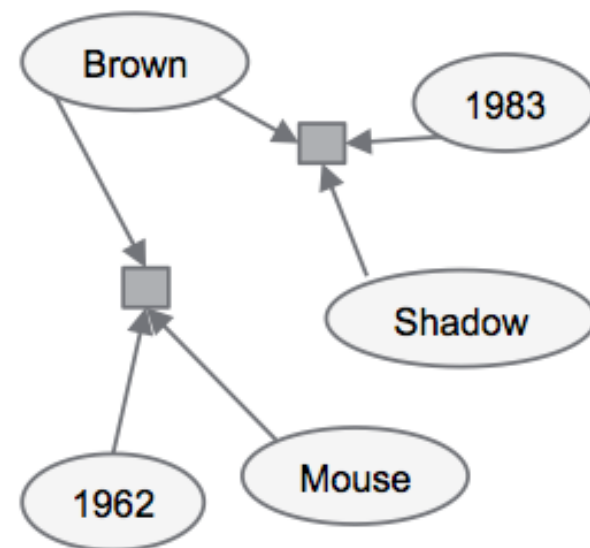
Title: Effective Java Programming Language Guide
Author: Joshua Bloch
Format: Paperback
Date: June 5, 2001
Price: \$29.69

Fact extraction from tables

Caldecott Medal

From Wikipedia, the free encyclopedia

| ◆ | Illustrator | ◆ | Title |
|------|---------------------|---|--|
| 1999 | Mary Azarian | | <i>Snowflake Bentley</i> |
| 1954 | Ludwig Bemelmans | | <i>Madeline's Rescue</i> |
| 1983 | Marcia Brown | | <i>Shadow</i> |
| 1962 | Marcia Brown | | <i>Once a Mouse</i> |
| 1955 | Marcia Brown | | <i>Cinderella, or the Little Glass Slipper</i> |
| 1943 | Virginia Lee Burton | | <i>The Little House</i> |
| 1980 | Barbara Cooney | | <i>Ox-Cart Man</i> |
| 1959 | Barbara Cooney | | <i>Chanticleer and the Fox</i> |



Need to create hidden column containing CVT or blank node, to represent the 3-tuple

Webmaster annotation



```
<section> Hello, my name is John Doe, I am a graduate research assistant at  
the University of Dreams. My friends call me Johnny.  
You can visit my homepage at <a href="http://www.JohnnyD.com">www.JohnnyD.com</a>.  
I live at 1234 Peach Drive Warner Robins, Georgia.</section>
```

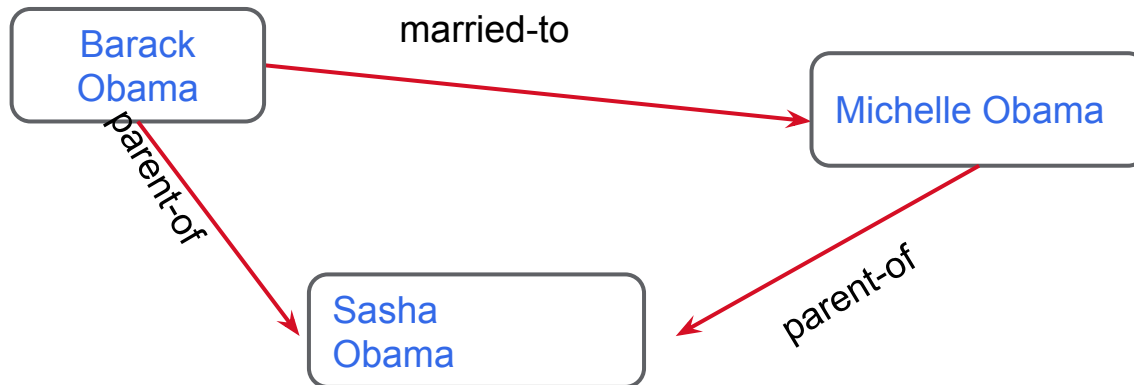
```
<section itemscope itemtype="http://data-vocabulary.org/Person">  
  Hello, my name is  
  <span itemprop="name">John Doe</span>,  
  I am a  
  <span itemprop="title">graduate research assistant</span>  
  at the  
  <span itemprop="affiliation">University of Dreams</span>.  
  My friends call me  
  <span itemprop="nickname">Johnny</span>.  
  You can visit my homepage at  
  <a href="http://www.JohnnyD.com" itemprop="url">www.JohnnyD.com</a>.  
  <section itemprop="address"  
    itemscope itemtype="http://data-vocabulary.org/Address">  
    I live at  
    <span itemprop="street-address">1234 Peach Drive</span>  
    <span itemprop="locality">Warner Robins</span>  
    ,  
    <span itemprop="region">Georgia</span>.  
  </section>  
</section>
```

Example taken from [http://en.wikipedia.org/wiki/Microdata_\(HTML\)](http://en.wikipedia.org/wiki/Microdata_(HTML))

Predicting facts given prior knowledge

- Perform association rule mining* on Freebase graph, to find noisy rules (features passed to a learned classifier).

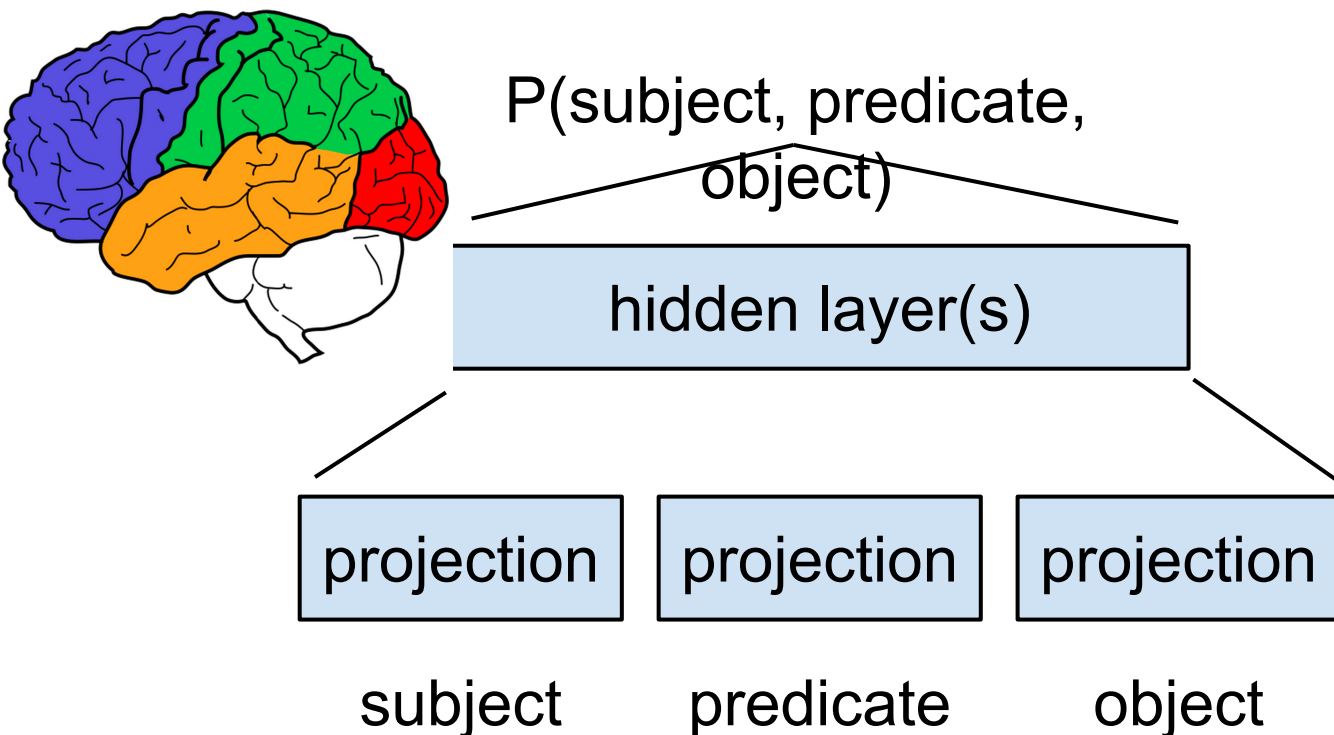
$$\forall x, y. \exists z. \text{parent-of}(x, z) \wedge \text{parent-of}(y, z) \Rightarrow \text{married}(x, y)$$



* “Random Walk Inference and Learning in A Large Scale Knowledge Base”, Ni Lao et al, 2011

A “neural” prior model

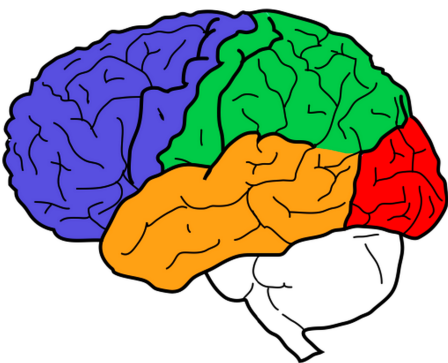
- Train a deep neural network* to predict the probability of arbitrary facts, cf. tensor factorization.



* Similar to “Learning Structured Embeddings of Knowledge Bases”, Bordes et al, 2011

A “neural” prior model - Halloween version

- Train a deep neural network* to predict the probability of arbitrary facts of tensor factorization.



proj
sub

tion
ct

Knowledge Vault* fuses all these signals together

- Data from web
 - Unstructured text
 - Semi-structured DOM trees
 - Structured WebTables
- “Prior” data from FB

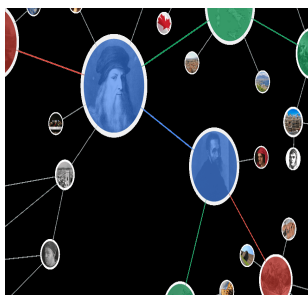
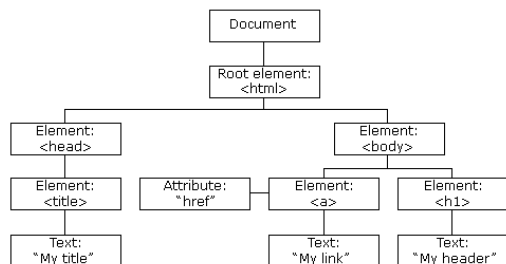
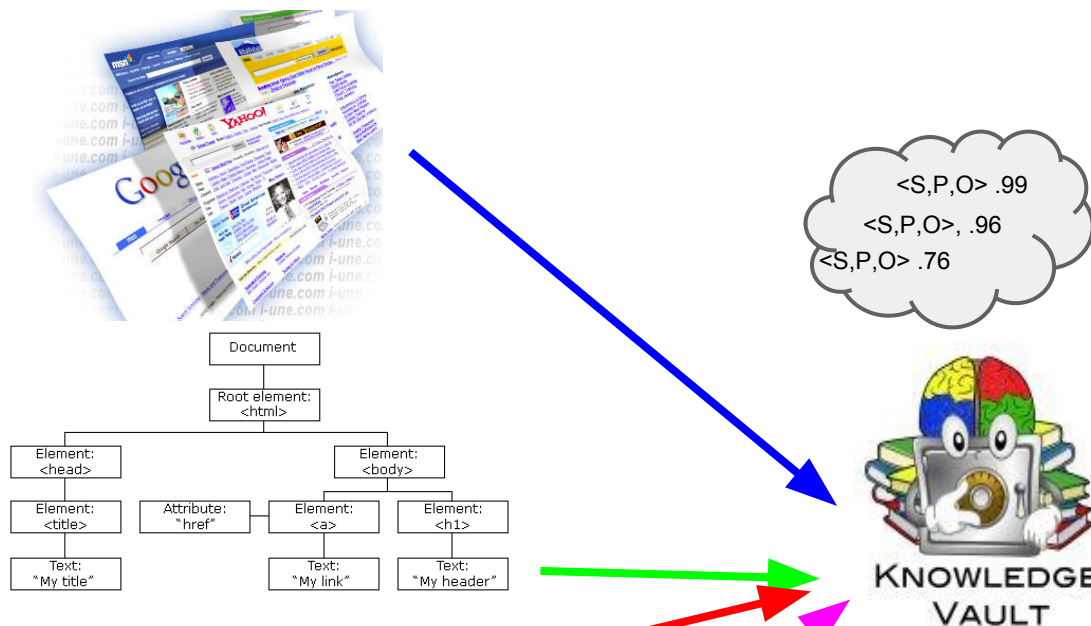
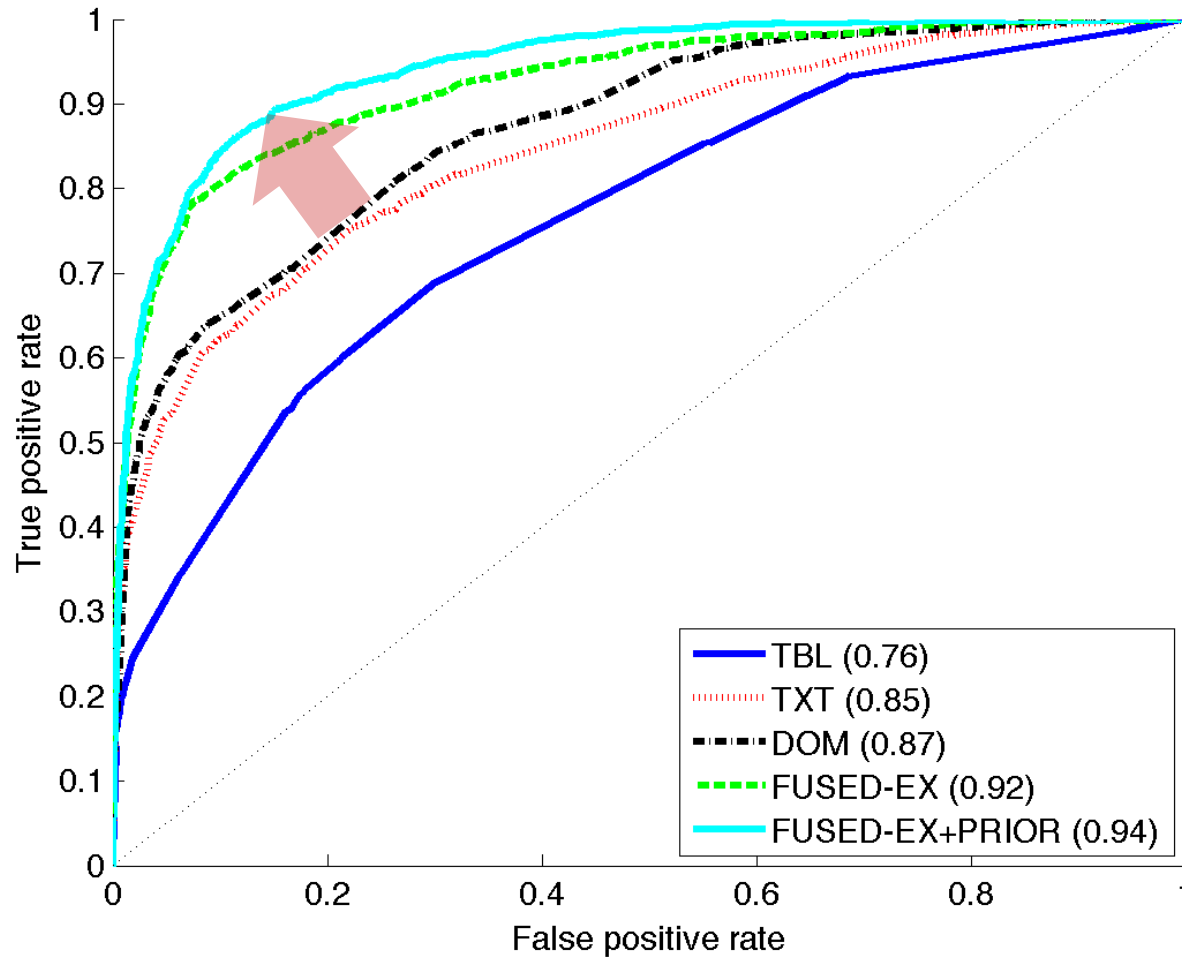


TABLE 1 MEDICAID-RELATED (MA) FUNDING-MINUS GENERAL PURPOSE REVENUE (GPR), 2009-2009 (IN MILLIONS)

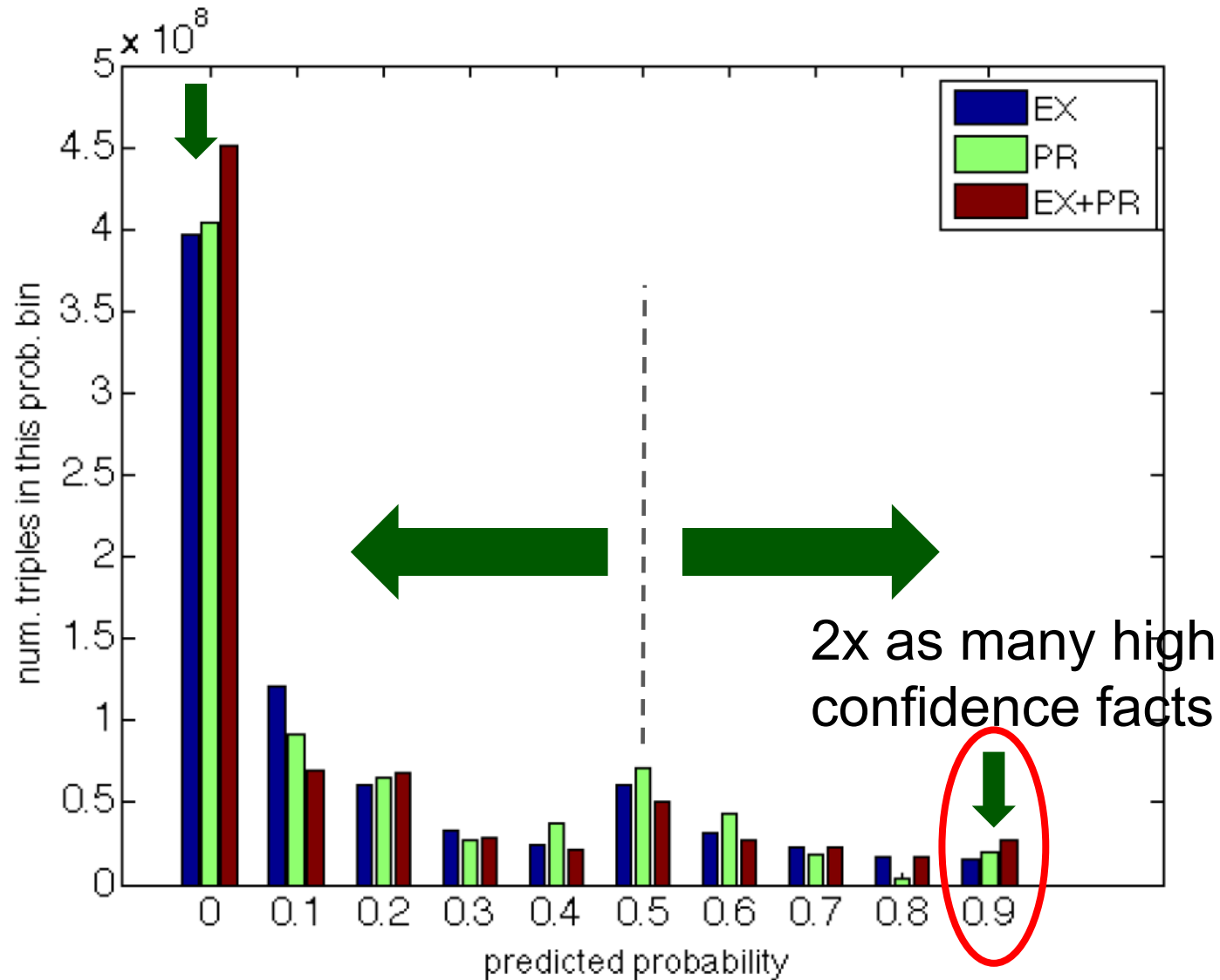
| Year | Change | MA | Change |
|--------|------------|-------|-----------|
| | \$10,948.0 | N/A | \$2,872.3 |
| FY01 | \$10,063.4 | -8.1% | \$3,649.4 |
| FY02 | \$10,020.2 | -0.4% | \$3,447.4 |
| FY03 | \$10,199.7 | 1.8% | \$3,852.7 |
| FY04 | \$10,739.3 | 5.3% | \$4,169.2 |
| FY05 | \$11,396.7 | 6.1% | \$4,270.9 |
| FY06 | \$12,030.1 | 5.6% | \$4,194.6 |
| FY07 | \$12,613.7 | 4.9% | \$4,716.8 |
| FY08 | \$13,101.1 | 3.9% | \$4,819.9 |
| FY09 | \$13,627.2 | 4.0% | \$5,114.2 |
| Total: | 24.5% | | 78.1% |

* Details in a paper submitted to WWW'14 (Dong et al)

Benefits of information fusion



Benefits of prior knowledge



Example: <Barry Richter, studied at, 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**

Prior knowledge:

<Barry Richter, born in, Madison>

<Barry Richter, lived in, Madison>

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

- From strings to things
- Reading the web
- • Asking the web
- Asking people
- Open issues

Knowledge based completion using Question Answering *

- Even after large-scale machine reading of the web, many facts are still unknown.
- We can use web-based question-answering to perform targeted completion of missing attributes (pull vs push model).
- Main issue: what questions should we ask?

* Details in a paper submitted to WWW'14 (West et al)

The importance of asking the right question



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 ...

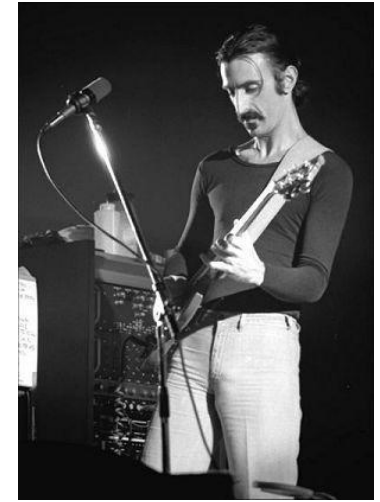
[History](#) - [Personnel](#) - [Discography](#) - [References](#)

[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](#)



The importance of asking the right question



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 ...

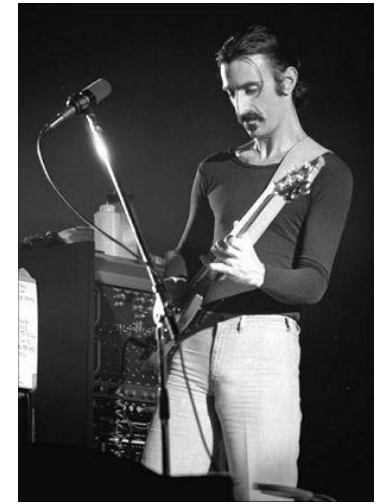
[History](#) - [Personnel](#) - [Discography](#) - [References](#)

[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

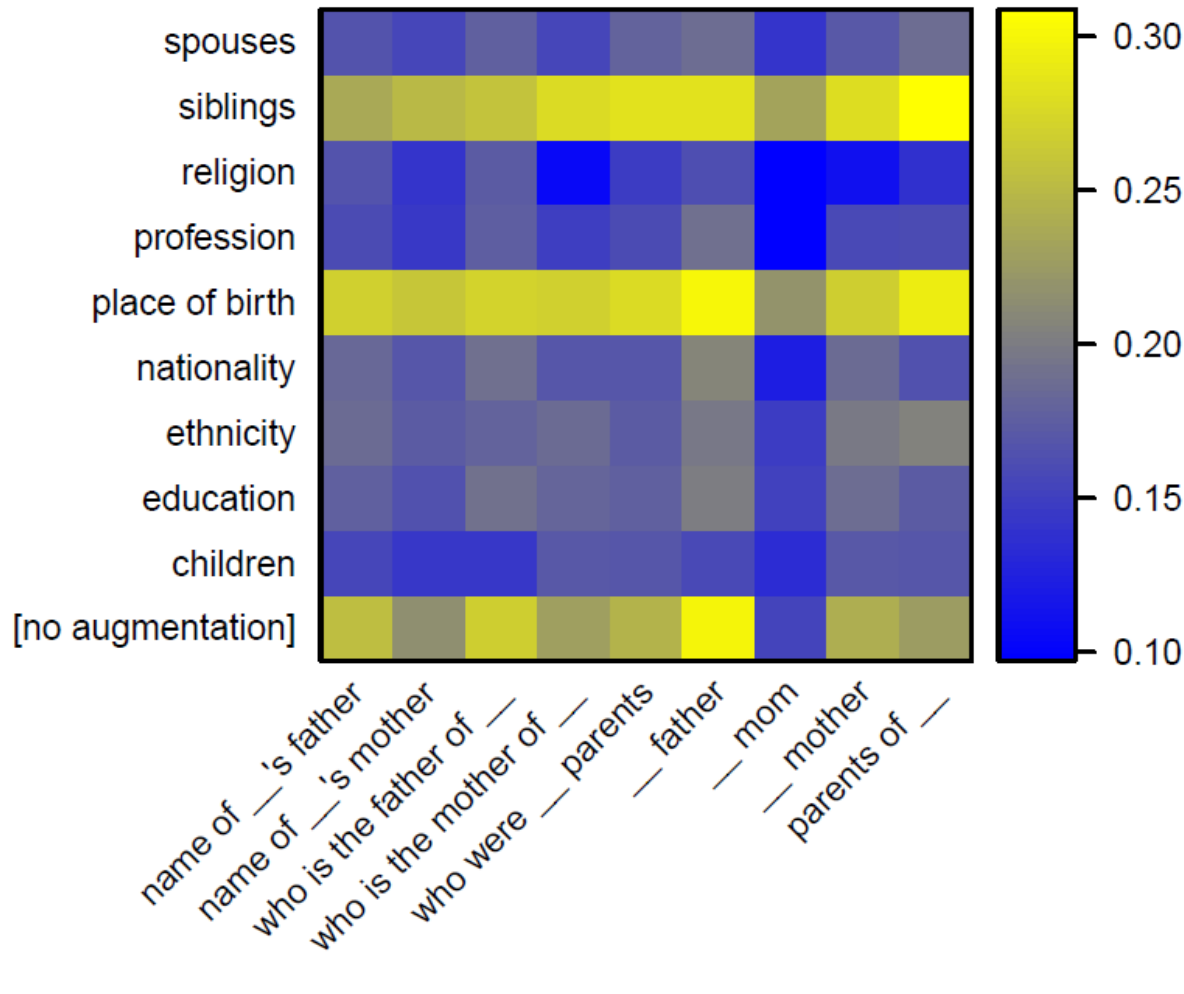


[Frank Zappa - Wikipedia, the free encyclopedia](#)

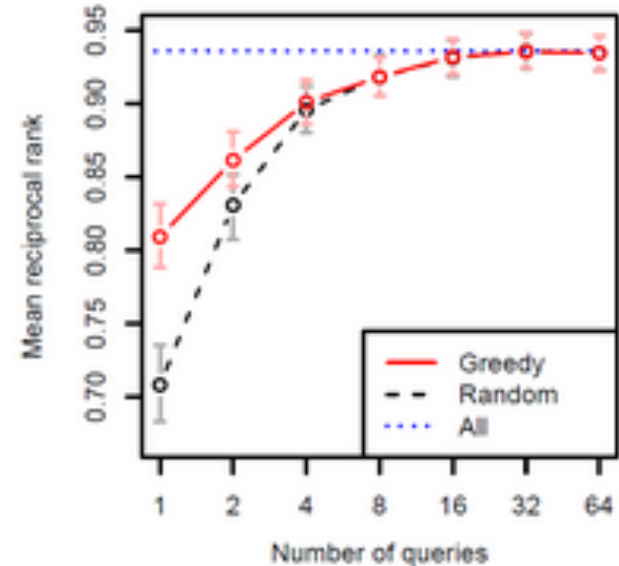
en.wikipedia.org/wiki/Frank_Zappa ▾

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 which questions to ask



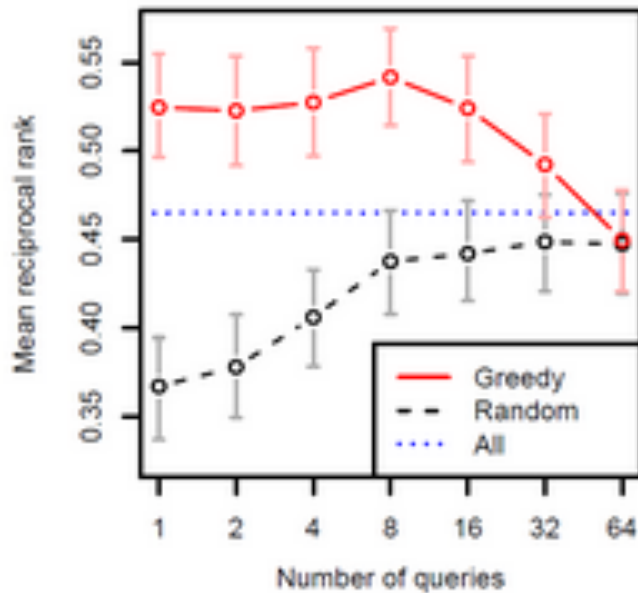
How many questions should we ask?



(c) NATIONALITY

Performance increases,
then plateaus

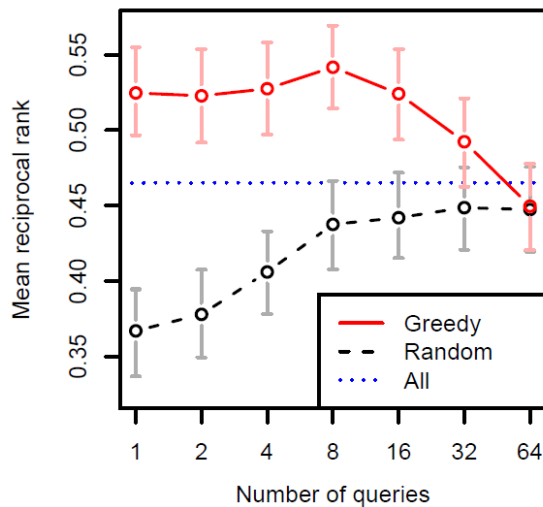
Asking too many questions can hurt performance



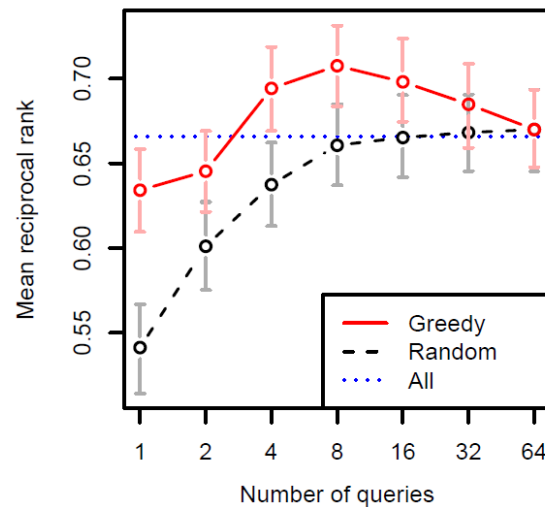
(a) SPOUSES

Performance gets worse

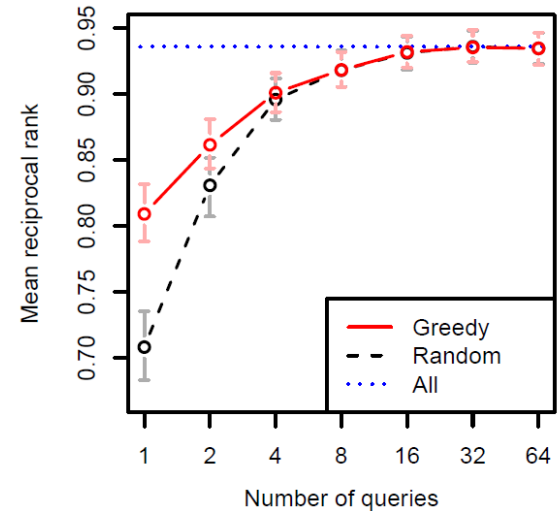
Why does performance differ?



(a) SPOUSES



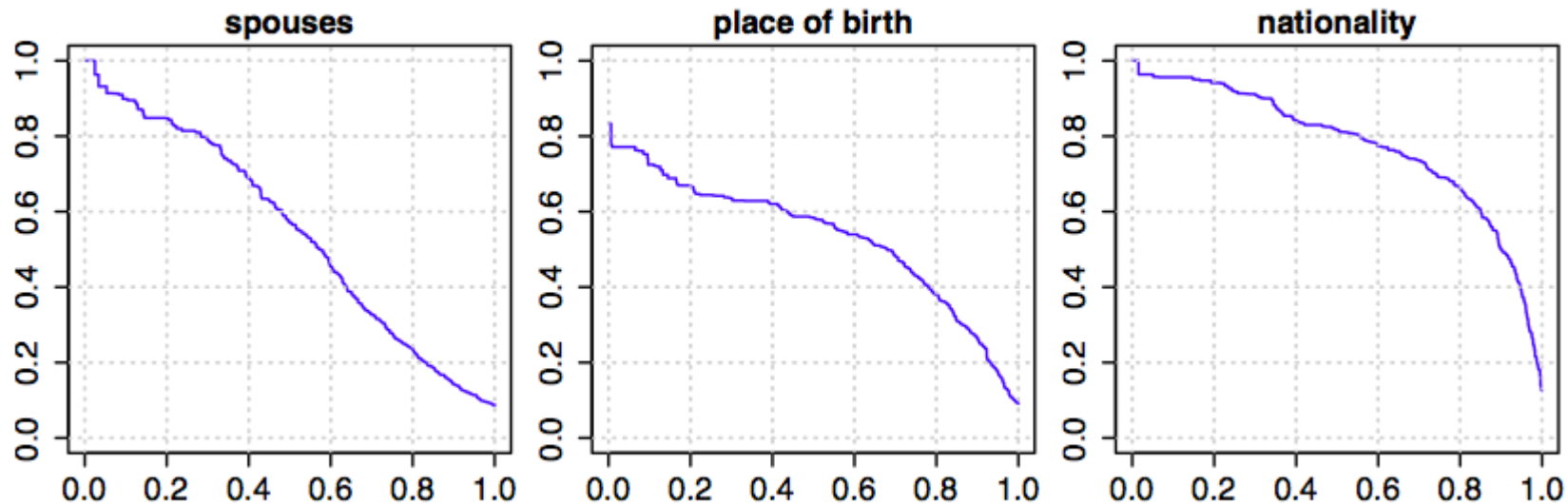
(b) PLACE OF BIRTH



(c) NATIONALITY

Open class

Closed class



- About 25% of the high confidence facts were not discovered by the “read the web” approach.
- Accuracy is higher for closed-class predicates.

- From strings to things
- Reading the web
- Asking the web
- • Asking people
- Open issues


Freebase is community generated/ edited



Topic

Bruce Lee ^{en}

mid: /m/099d4 notable type: /marial_arts/martial_artist



Bruce Lee was a Chinese American martial artist, Hong Kong Cantonese opera star Lee Hoi-Chuen. He is widely considered the greatest martial artist of all time, and a pop culture icon of the 20th century. He is often cited as the most famous Chinese American. He was born in San Francisco on 27 November 1940 to parents from Hong Kong and appeared in several films as a child actor. Lee moved to Los Angeles to pursue his interest in martial arts. His Hong Kong and Hollywood-produced films were instrumental in creating a widespread interest in Chinese martial arts in the West in the 1970s. The world. Wikipedia [-]

People /people

Person /people/person

Date of birth /people/person/date_of_birth

11/27/1940

Place of birth /people/person/place_of_birth

- Chinatown

Edit of nationality /people/person/nationality

Delete es of America

Edit localized...

Gender /people/person/gender

Male

Profession /people/person/profession

- Actor
- Screenwriter
- Film Director
- Martial Arts Instructor

Knowledge panel feedback



georgia o'keeffe artist



Georgia O'Keeffe

Artist

Georgia Totto O'Keeffe was an American artist. Born near Sun Prairie, Wisconsin, O'Keeffe first came to the attention of the New York art community in 1916. [Wikipedia](#)

Born: November 15, 1887, [Sun Prairie, WI](#)

Died: March 6, 1986, [Santa Fe, NM](#)

Nationality: American

Period: American modernism

Spouse: [Alfred Stieglitz](#) (m. 1924–1946)

Artwork: [Cow's Skull: Red, White, and Blue](#), [Sky Above Clouds IV](#), [More](#)

Knowledge panel feedback



georgia o'keeffe artist

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



Wrong?

Georgia O'Keeffe

Artist

Georgia Totto O'Keeffe was an American artist. Born near Sun Prairie, Wisconsin, O'Keeffe first came to the attention of the New York art community in 1916. [Wikipedia](#)

Wrong?

Wrong? **Born:** November 15, 1887, Sun Prairie, WI

Wrong? **Died:** March 6, 1986, Santa Fe, NM

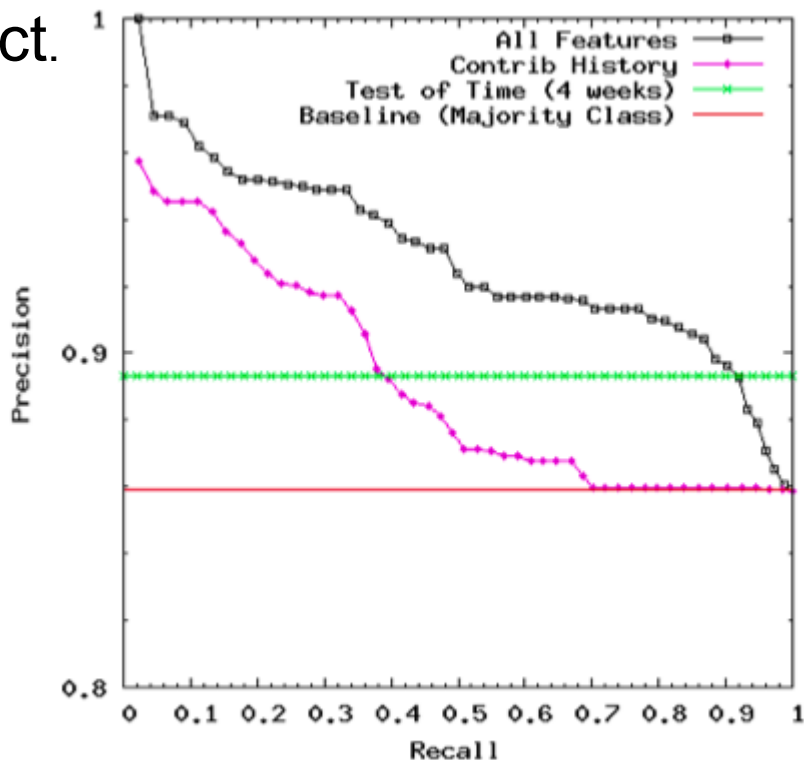
Wrong? **Nationality:** American

Wrong? **Period:** American modernism

Wrong? **Spouse:** Alfred Stieglitz (m. 1924–1946)

Wrong? **Artwork:** Cow's Skull: Red, White, and Blue, Sky Above Clouds IV, [More](#)

Use a binary classifier, trained on features derived from user contribution history, to predict the probability the contribution is correct.



* Details in a paper submitted to WSDM'14 (Tan et al)

Asking the right people*



Place an ad asking users to take a quiz.
Use ad optimization system to figure out
which kinds of users to show the ad to.

[Quiz on disease symptoms](#)
Test how well you can recognize
various disease symptoms
www.quiz.us

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

What is a symptom of Morgellons

Red eye

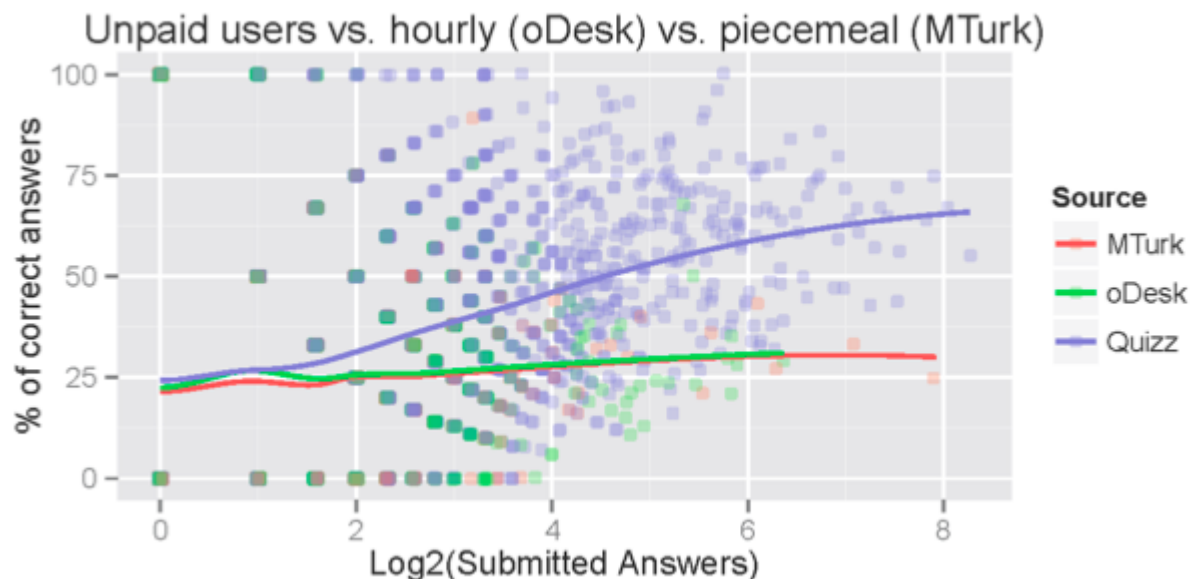
Choreoathetosis

Skin lesion

Insomnia

I don't know

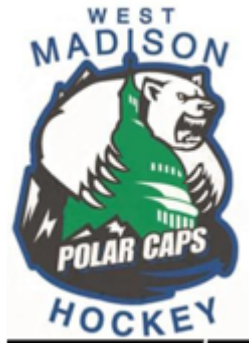
Question 1 out of 10



* Details in a paper submitted to WWW'14 (Ipeirotis and Gabrilovich)

- From strings to things
- Reading the web
- Asking the web
- Asking people
- • Open issues

New entities



/m/?

“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.”

/m/02ql38b



40M entities in Freebase, still missing many!

/people/person/education . /education/educational_institute

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 ...”

/people/person/?

35k types of relations in Freebase, still missing many!

Joanne Schieble was just twenty-three and attending graduate school in Wisconsin when she learned she was pregnant. Her father didn't approve of her relationship with a Syrian-born graduate student, and social customs in the 1950s frowned on a woman having a child outside of marriage. To avoid the glare, Schieble moved to San Francisco and was taken in by a doctor who took care of unwed mothers and helped arrange adoptions. Originally, a lawyer and his wife agreed to adopt the new baby. But when the child was born on February 24, 1955, they changed their minds. Clara and Paul Jobs, a modest San Francisco couple with some high school education, had been waiting for a baby. When the call came in the middle of the night, they jumped at the chance to adopt the newborn, and they named him Steven Paul.

<Joanne Schieble, /people/person/parents, Steve Jobs>

<Steve Jobs, /people/person/date-of-birth, 2/24/55>

Assessing trustworthiness of sources



WIKIPEDIA
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia
- Interaction
 - Help
 - About Wikipedia
 - Community portal
 - Recent changes
 - Contact page
- Toolbox
- Print/export
- Languages

- Acèh
- Afrikaans
- Alemannisch
- አማርኛ
- English
- ᲛᲗᲚᲘᲗ
- العربية
- Aragonés
- АԥсӀӀ
- Asturianu
- АваргаӀ
- Аԥар
- Aymar aru
- Azərbaycanca
- Bamanankan
- বাংলা
- Bahasa Banjar
- Беларуская
- Basa Banyumasan
- Башҡортса
- Беларуская (тарашкевіца)
- भोजपुरी

Create account Log in

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 (/ˈbɑːrək huːseɪn oʊˈbɑːmɑː/; 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 |
| <div>44th President of the United States</div> |
| <div><div><div><div></div></div></div><div>Incumbent</div></div> |
| <div><div><div><div></div></div></div><div>Assumed office</div></div> <div>January 20, 2009</div> |
| <div><div><div><div></div></div></div><div>Vice President</div></div> <div>Joe Biden</div> |
| <div><div><div><div></div></div></div><div>Preceded by</div></div> <div>George W. Bush</div> |
| <div><div><div><div></div></div></div><div>United States Senator from Illinois</div></div> |
| <div><div><div><div></div></div></div><div>In office</div></div> <div>January 3, 2005 – November 16, 2008</div> |
| <div><div><div><div></div></div></div><div>Preceded by</div></div> <div>Peter Fitzgerald</div> |
| <div><div><div><div></div></div></div><div>Succeeded by</div></div> <div>Roland Burris</div> |
| <div><div><div><div></div></div></div><div>Member of the Illinois Senate from the 13th District</div></div> |
| <div><div><div><div></div></div></div><div>In office</div></div> <div>January 8, 1997 – November 4, 2004</div> |
| <div><div><div><div></div></div></div><div>Preceded by</div></div> <div>Alice Palmer</div> |
| <div><div><div><div></div></div></div><div>Succeeded by</div></div> <div>Kwame Raoul</div> |
| <div><div><div><div></div></div></div><div>Personal details</div></div> |
| <div><div><div><div></div></div></div><div>Born</div></div> <div>Barack Hussein Obama II</div> <div>August 4, 1961 (age 52)</div> <div>Honolulu, Hawaii, U.S.</div> |
| <div><div><div><div></div></div></div><div>Political party</div></div> <div>Democratic</div> |



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

Share 14.3K | 1 | 24 | Share 14.8K | Tweet 416 | Email 0

Breitbart.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 Breitbart still think that now Obama is telling the truth.

The screenshot shows a news article on The Blaze website. The article title is "YAHOO! NEWS SAYS OBAMA WAS BORN IN...KENYA!". The author is Madeleine Morgenstern, dated Jun. 22, 2013 12:34pm. The article has 802 tweets, 16.8K shares, and 34 likes. The article text is partially visible, mentioning "Goderich hastily claimed listing error..." and "Yahoo! News had to issue a correction Friday after publishing an article about President Barack Obama that called Kenya '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 26 | Share 1 | 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**.

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



Summary



1. Knowledge Vault is the largest repository of automatically extracted structured knowledge on the planet.
2. We can extract more information by asking the right questions from the web and/or people.
3. We are only extracting a small fraction of the facts on the web.

