

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 Strohmann, Shaohua Sun, Chun How
Tan, Robert West, Wei Zhang, and others*

Big Data is everywhere



gacc**ca**-----
gacc**cacacga**-----
aattcaggacc**cacacgacggg**aagacaa-----
-attcaggaca**aacacga**aggaaagacaagttcatgtacttt
----caggacc**cacacgacggg**tagacaagttcatgtacttt
-----acc**cacacgacggg**tagacaagttcatgtacttt
-----acc**cacacgacggg**tagacaagttcatgtacttt
-----gacgggaagacaagttcatgtacttt
-----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

Google search results for "marie curie":

- Search**: 8 personal results, 17,300,000 other results (0.32 seconds)
- Everything**: Marie Curie - Wikipedia, the free encyclopedia (en.wikipedia.org/wiki/Marie_Curie)
- Images**: Marie Skłodowska-Curie (7 November 1867 – 4 July 1934) was a French-Polish physicist and chemist famous for her pioneering research on radioactivity.
- Maps**: [Marie Curie Cancer Care](#) - Pierre Curie - Irène Joliot-Curie - Radioactive decay
- Videos**: [Marie Curie - Biography](#) - Nobelprize.org (www.nobelprize.org/nobel_prizes/physics/.../marie-curie-bio.html)
- News**: Short profile from the foundation that awards the Nobel Prize.
- Shopping**: [Marie Curie - Biography](#) - Nobelprize.org (www.nobelprize.org/nobel_prizes/chemistry/.../marie-curie-bio.html)
- Books**: Marie Curie, née Maria Skłodowska, was born in Warsaw on November 7, 1867, the daughter of a secondary-school teacher. She received a general education ...
- More**: [Images for marie curie](#) - Report images (
- Mountain View, CA**: Change location
- Any time**: Past hour, Past 24 hours, Past 2 days, Past week, Past month, Past year, Custom range...

Knowledge Panel for Marie Curie:

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

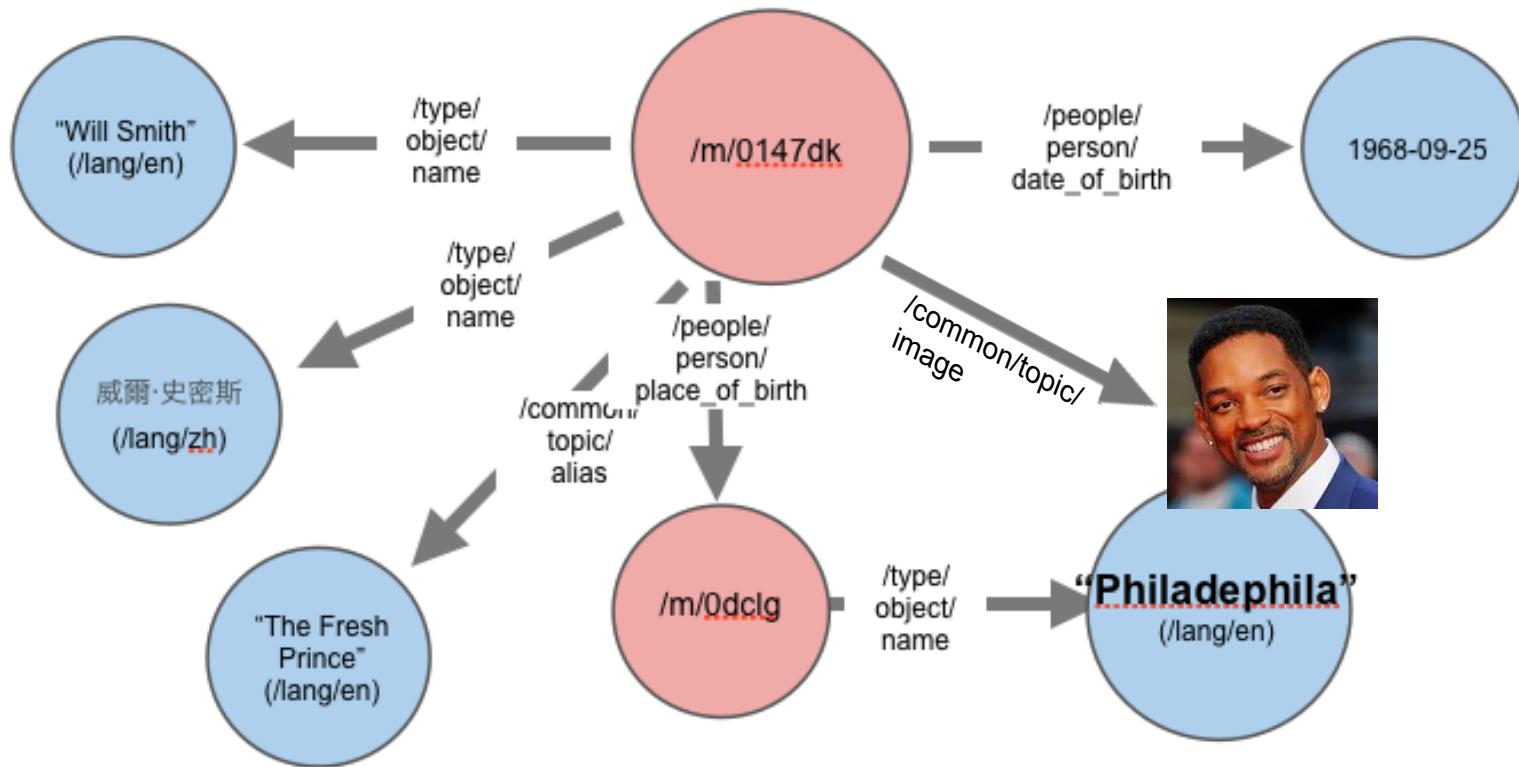
Freebase is created by merging many data sources



Massive entity linkage problem!

Source: John Giannandrea, CIKM 2011 industry talk

A fragment of Freebase (in RDF format)



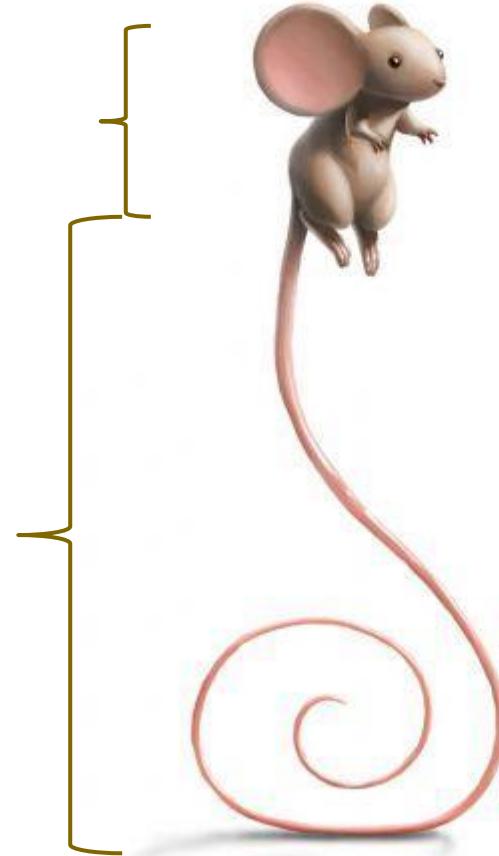
The long tail of knowledge



- 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.

Fact extraction from text



- Template matching methods

Patrick Newport ,who has been working at IHS Global Insight, noted...

PER
/m/101

/people/person/employment

ORG
/m/102

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

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

Wrapper induction

Google

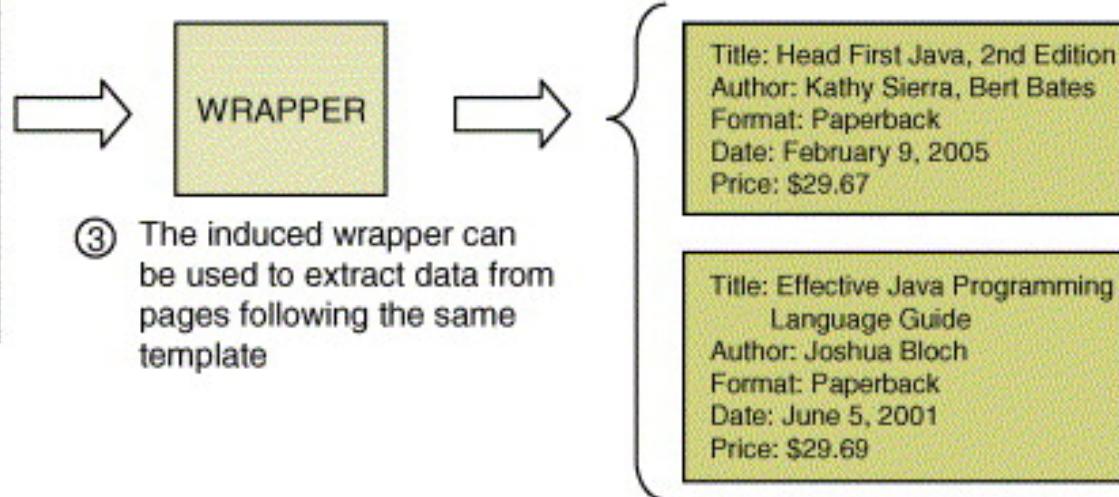
① User labels data records in example pages



② A set of extraction rules are induced from the examples

This block contains two screenshots of product listing pages. The first screenshot shows a book cover thumbnail, the title 'Head First Java, 2nd Edition', author 'Kathy Sierra, Bert Bates', format 'Paperback', date 'February 9, 2005', average rating '★★★★★', shipping info 'Usually ships in 24 hours', list price '\$30.00', buy price '\$29.67', and a 'Used & new from \$11.08' link. The second screenshot shows a book cover thumbnail, the title 'Effective Java Programming Language Guide', author 'Joshua Bloch', format 'Paperback', date 'June 5, 2001', average rating '★★★★★', shipping info 'Usually ships in 24 hours', list price '\$44.99', buy price '\$29.69', and a 'Used & new from \$24.24' link.

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



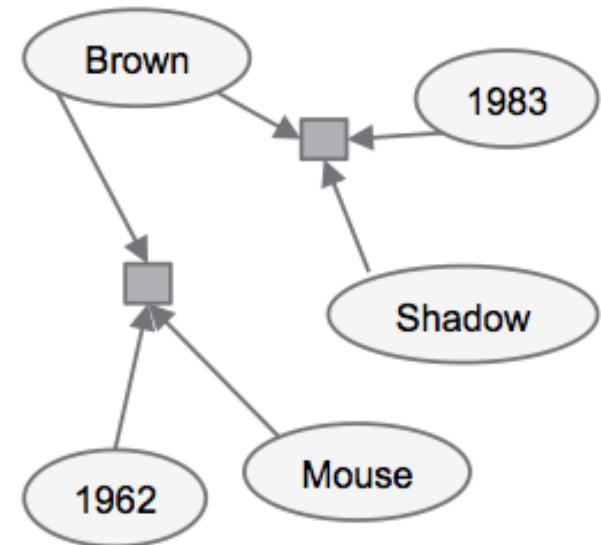
Fact extraction from tables

Google

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>
```

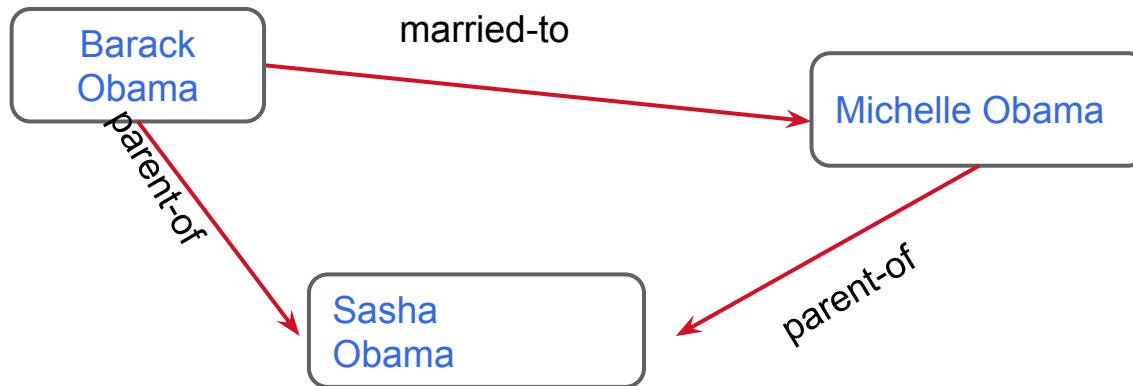
Example taken from [http://en.wikipedia.org/wiki/Microdata_\(HTML\)](http://en.wikipedia.org/wiki/Microdata_(HTML))

Predicting facts given prior knowledge



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

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

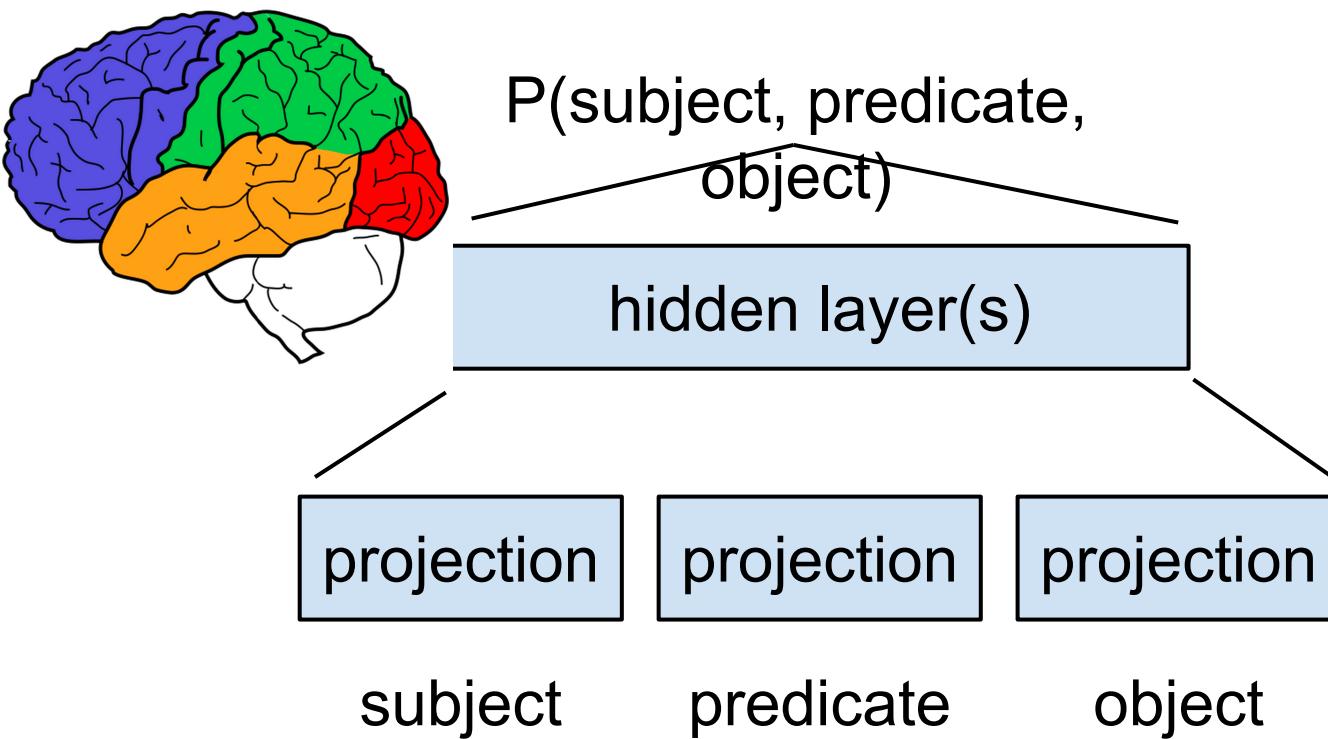


* “Random Walk Inference and Learning in A Large Scale Knowledge Base”, Ni Lao et al, 2011

A “neural” prior model



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

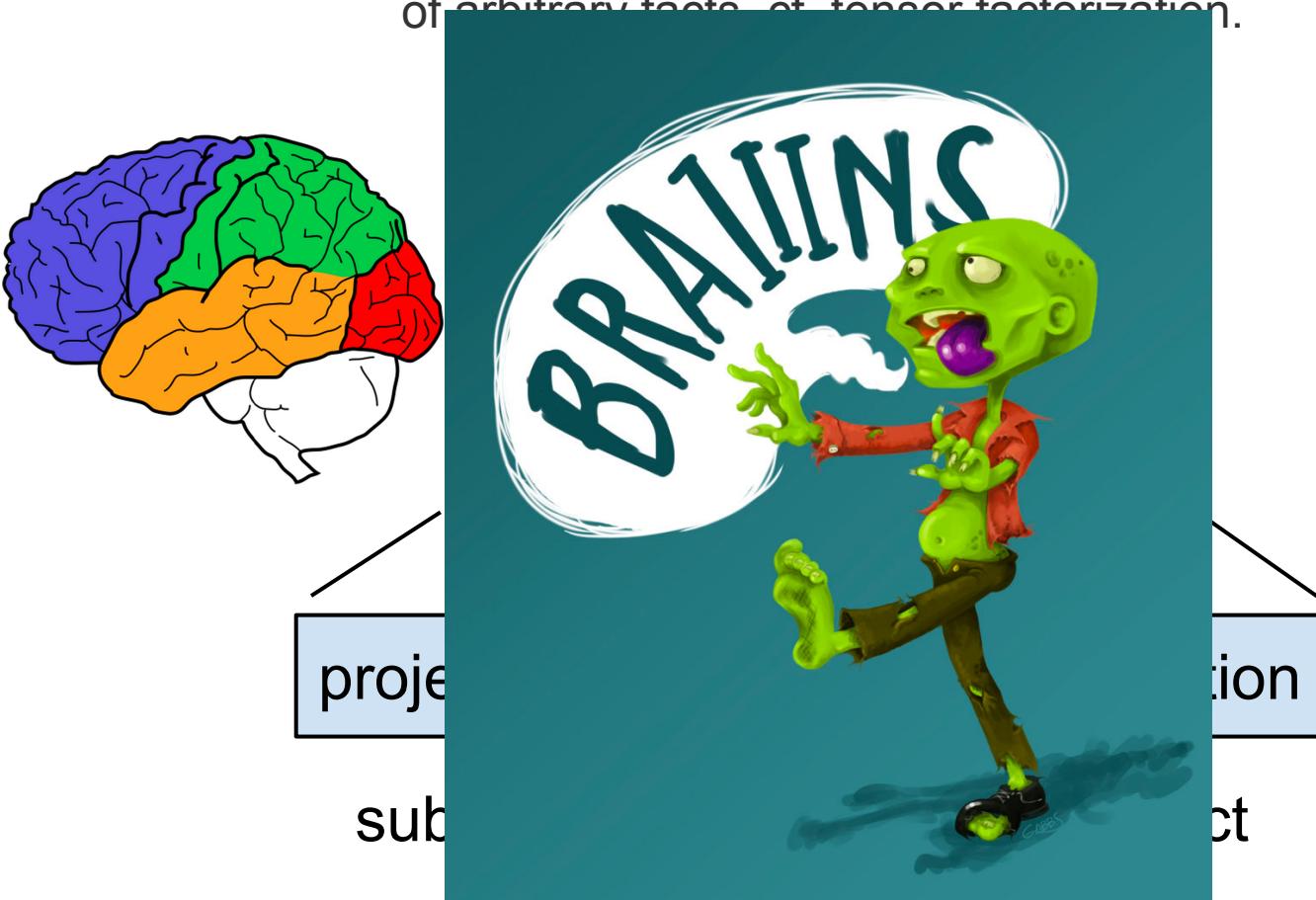


* Similar to “Learning Structured Embeddings of Knowledge Bases”, Bordes et al, 2011

A “neural” prior model - Halloween version

Google

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



Knowledge Vault* fuses all these signals together

- Data from web

- Unstructured text
- Semi-structured
DOM trees
- Structured
WebTables

- “Prior” data from FB

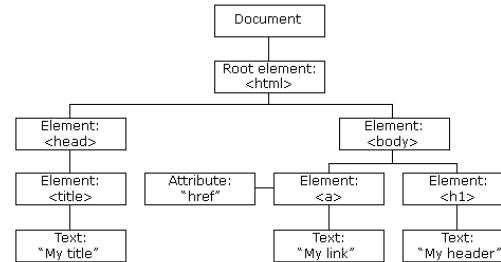
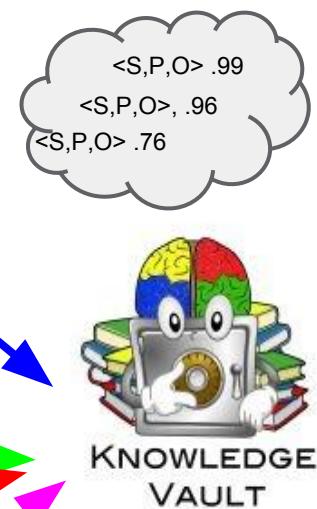


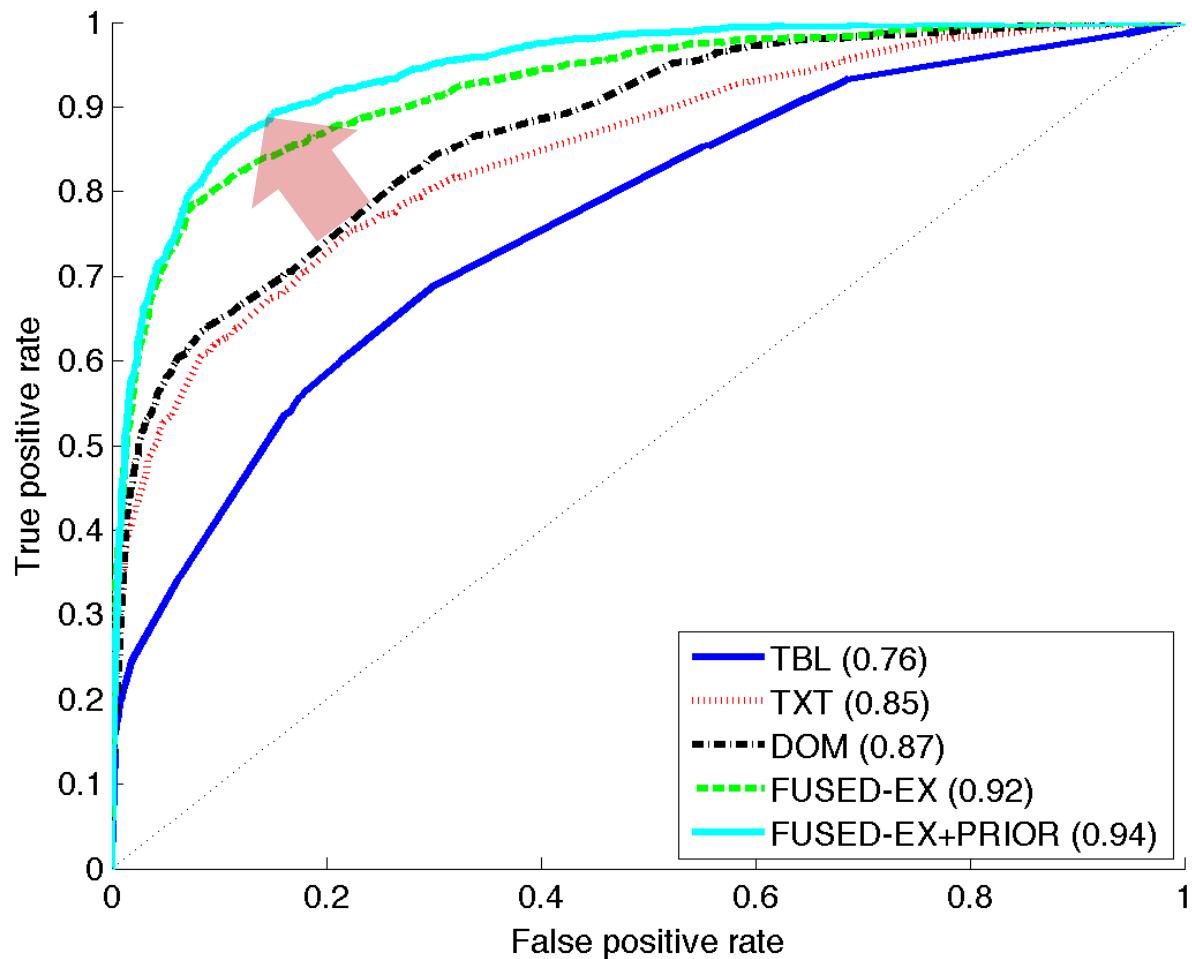
TABLE 1 MEDICAID-RELATED (MA) FUNDING—NET OF GENERAL PURPOSE REVENUE (GPR), 2000–2010 (IN MILLIONS)				
Year	Change	MA	Change	
\$10,946.0	N/A	\$2,872.3	N/A	
FY01	\$10,063.4	-8.1%	\$3,649.4	27.1%
FY02	\$10,020.2	-0.4%	\$3,447.4	-5.5%
FY03	\$10,199.7	1.6%	\$3,852.7	11.8%
FY04	\$10,739.3	5.3%	\$4,169.2	8.2%
FY05	\$11,396.7	6.1%	\$4,270.9	2.4%
FY06	\$12,030.1	5.6%	\$4,194.6	-1.8%
FY07	\$12,613.7	4.9%	\$4,716.8	12.4%
FY08	\$13,101.1	3.9%	\$4,819.9	2.2%
FY09	\$13,627.2	4.0%	\$5,114.2	6.1%
Total:	24.5%	78.1%		

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



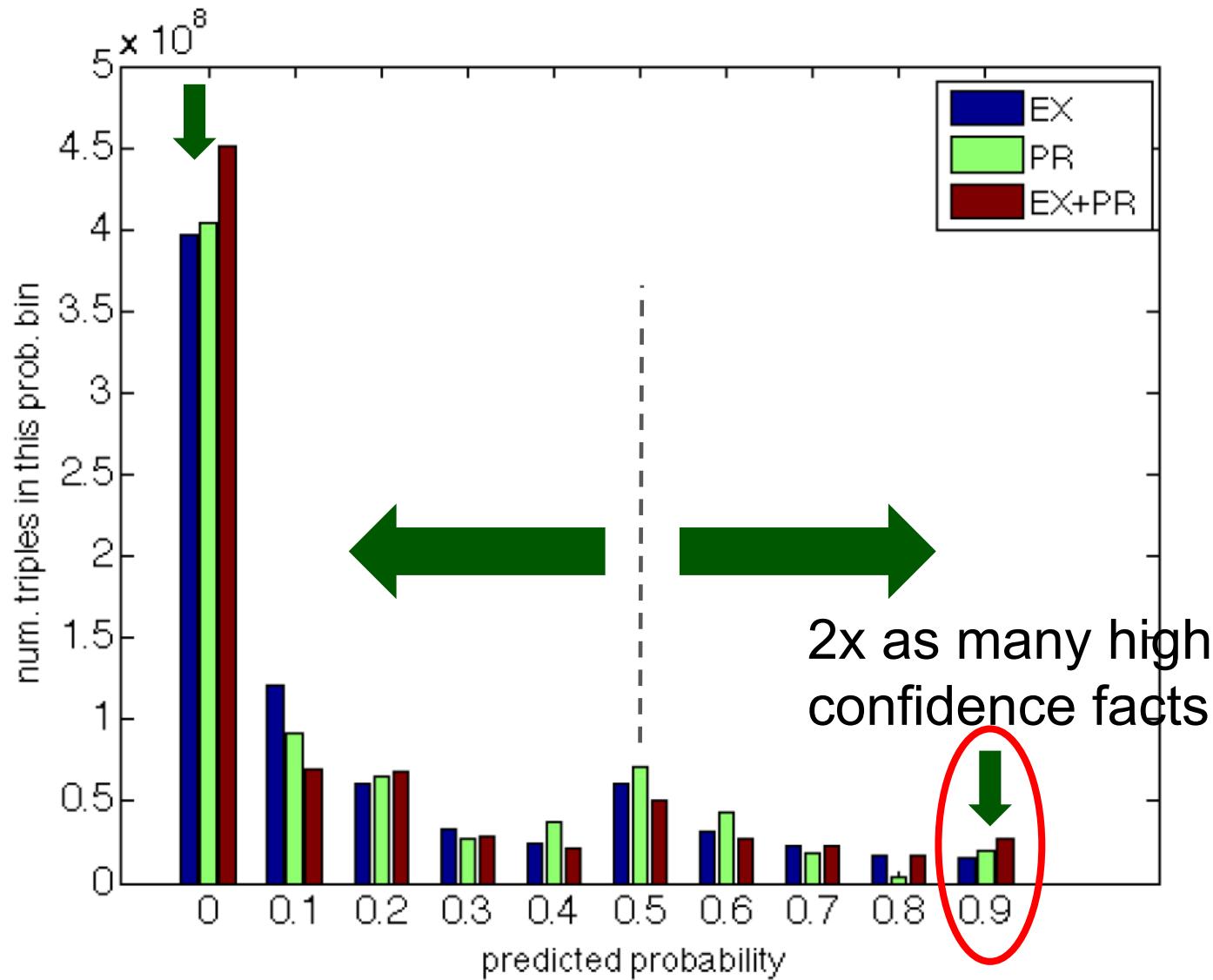
Benefits of information fusion

Google



Benefits of prior knowledge

Google



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

Google

[The Mothers of Invention - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wik/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 ...

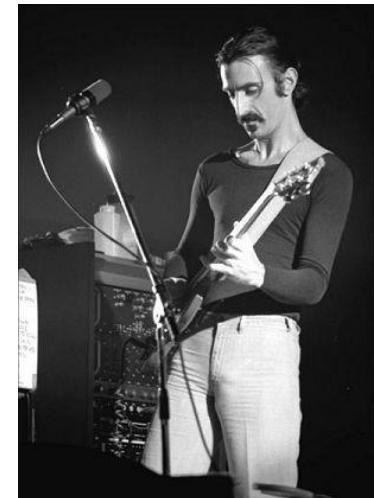
[History](#) [Personnel](#) [Discography](#) [References](#)

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

en.wikipedia.org/wik/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

Google

Who is the mother of Frank Zappa



[The Mothers of Invention - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wik... ▾](https://en.wikipedia.org/wik...)

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 ...

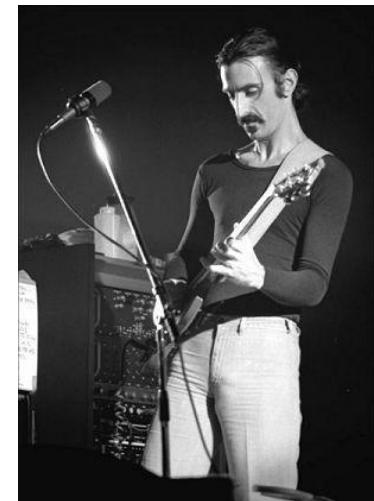
[History](#) - [Personnel](#) - [Discography](#) - [References](#)

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

[en.wikipedia.org/wik... ▾](https://en.wikipedia.org/wik...)

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



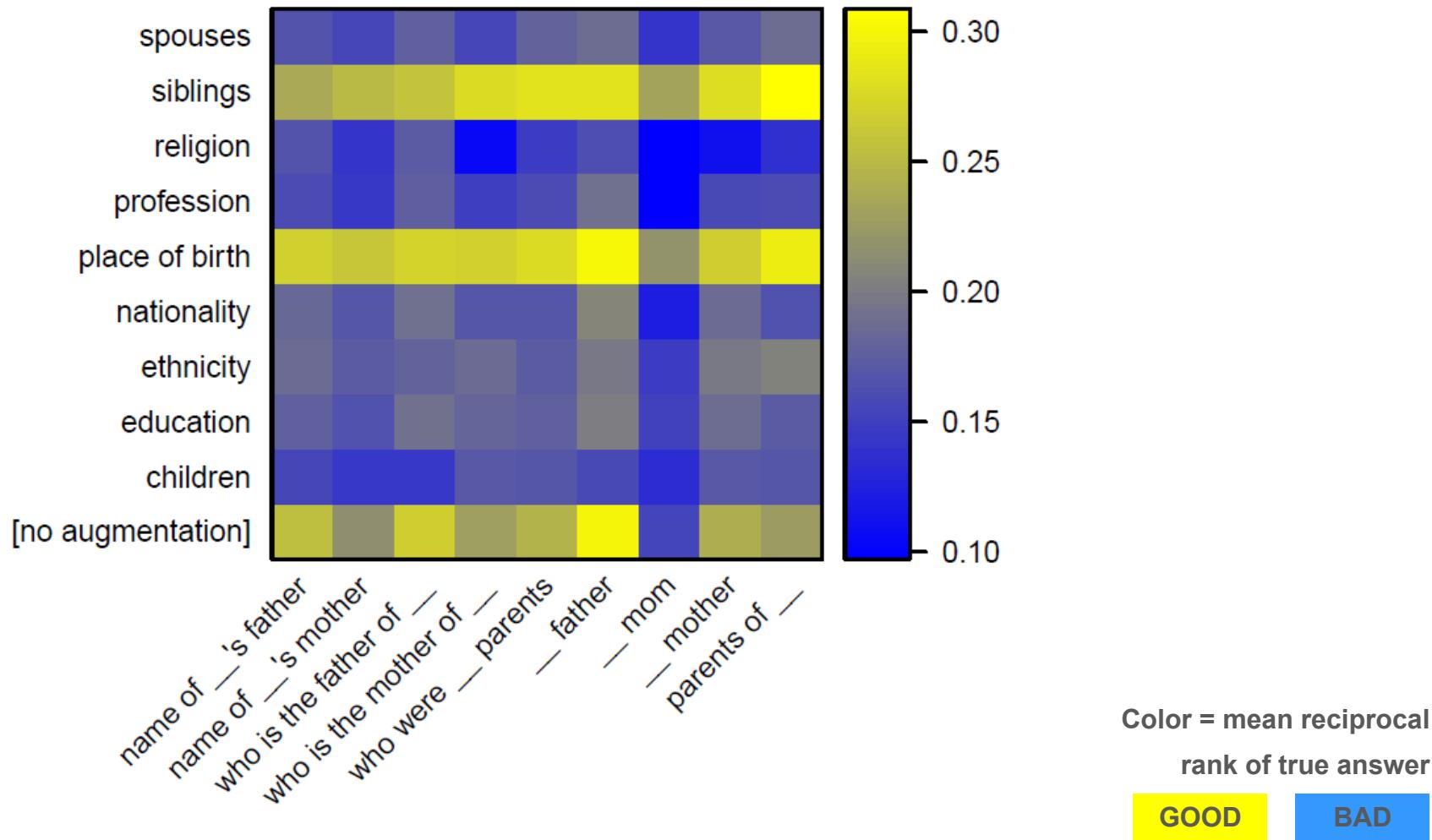
[Frank Zappa - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wik... ▾](https://en.wikipedia.org/wik...)

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 ...

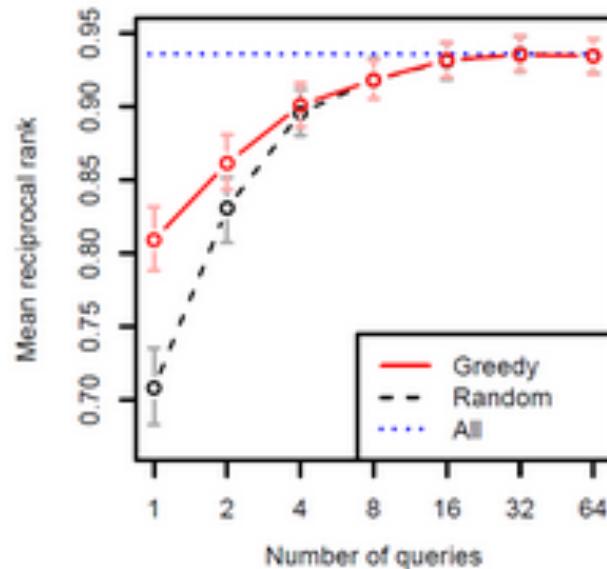
Learning which questions to ask

Google



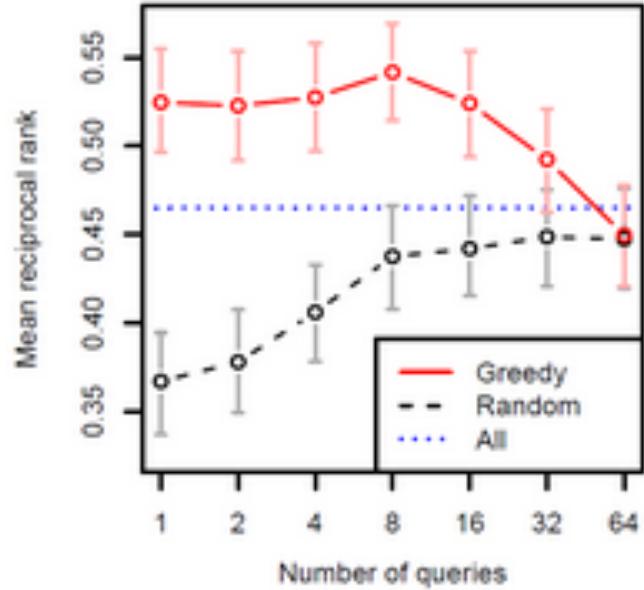
How many questions should we ask?

Google



(c) NATIONALITY

Performance increases,
then plateaus

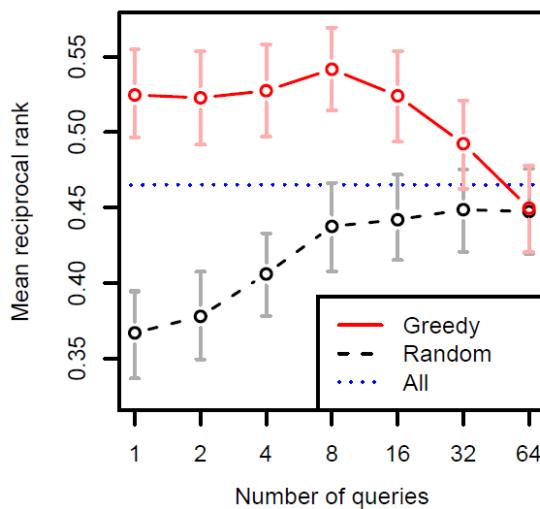


(a) SPOUSES

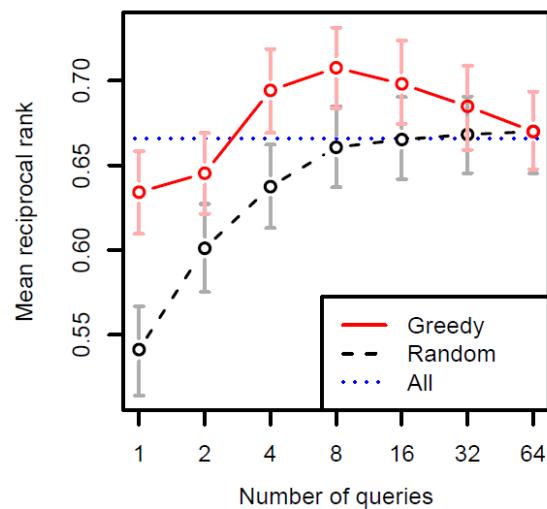
Performance gets worse

Why does performance differ?

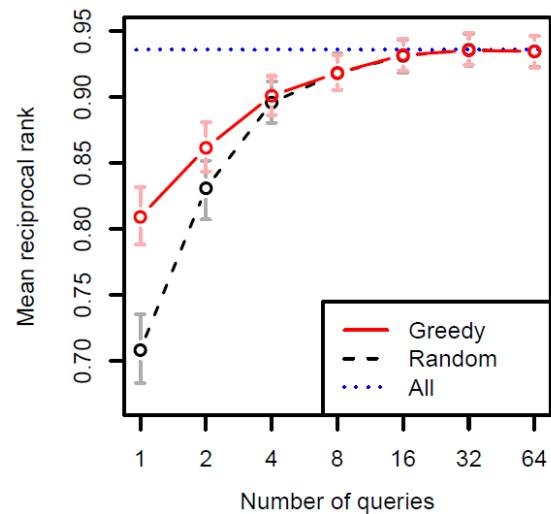
Google



(a) SPOUSES



(b) PLACE OF BIRTH



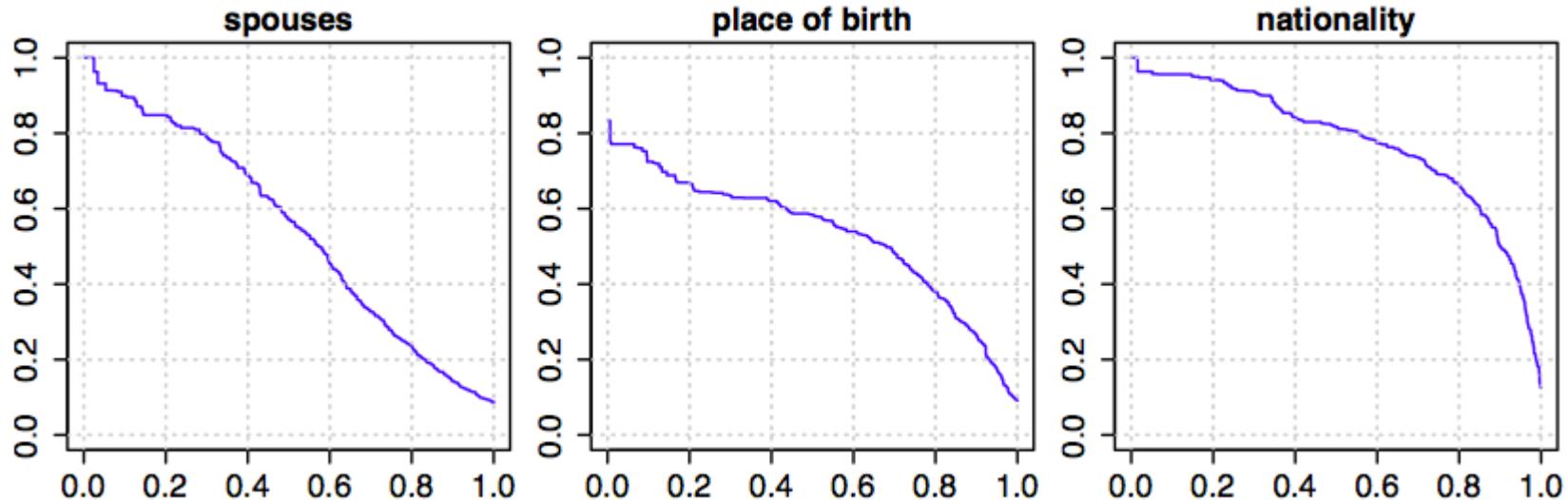
(c) NATIONALITY

Open class

Closed class

Precision-recall curves

Google



- 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: /martial_arts/martial_artist

Bruce Lee was a Chinese American martial artist, Hong Kong Cantonese opera star Lee Hoi-Chuen. He is widely considered one of the most influential figures in martial arts history. 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 Hong Kong at age 18 and became a master of Wing Chun. He taught martial arts. His Hong Kong and Hollywood-produced films helped to bring Chinese martial arts to the West and increased interest in Chinese martial arts in the 1970s. The Bruce Lee Foundation continues his legacy around 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 United States of America

Edit localized... []

Gender /people/person/gender

Male

Profession /people/person/profession

Actor

Screenwriter

Film Director

Martial Arts Instructor

Knowledge panel feedback

Google

Google

georgia o'keeffe artist



[More images](#)

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

Google



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?



Wrong?



Wrong?



Wrong?



Wrong?



Wrong?



Wrong?



Wrong?



Wrong?



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

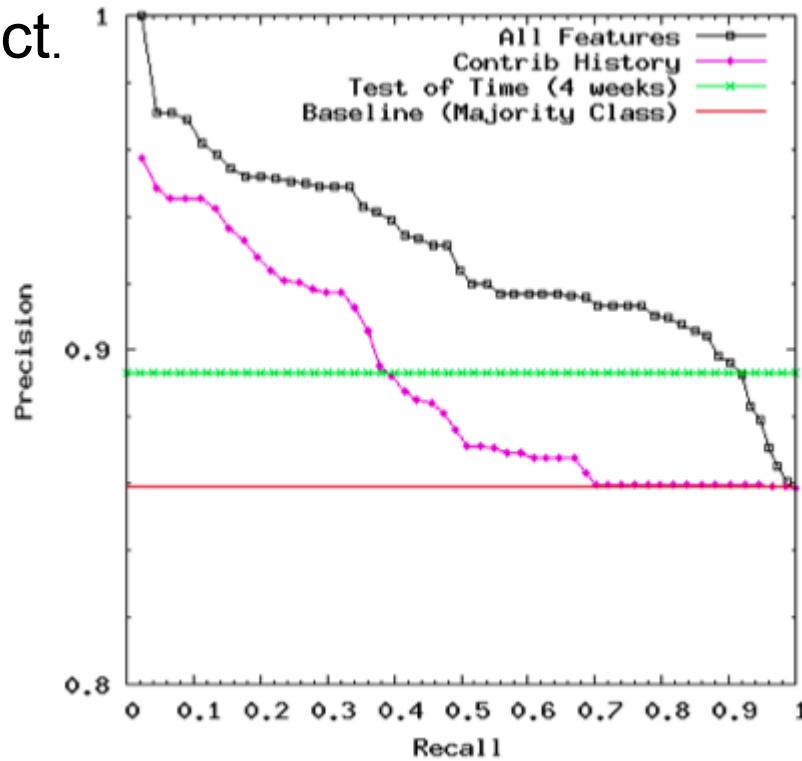
Wrong? **Spouse:** Alfred Stieglitz (m. 1924–1946)

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

Knowing who to trust*

Google

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)

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.

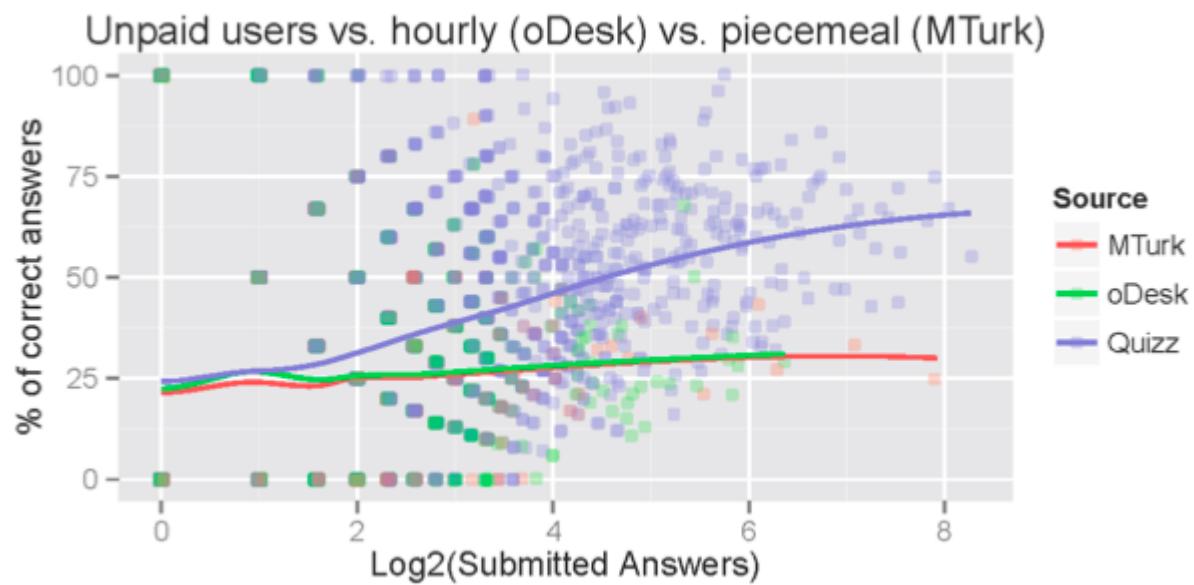
[Quiz on disease symptoms](#)
Test how well you can recognize various disease symptoms
www.quizz.us

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

What is a symptom of [Morgellons](#)

- Red eye
- Choroathetosis
- 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



/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
Ænglisc
Արցախ
العربيّة

Aragonés

অসমীয়া

Asturianu

Avareč

Aap

Aymar aru

Azerbaiyancanca

Bamanankan

বাষ্ণু

Bahasa Banjar

ବାନ୍ଦାମ୍ବୁଦ୍ଧ

Basa Banyumasan

Башҡортса

Беларуская

Беларускія (тарашкевіца)

ଓଡ଼ିଆ

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 (/*bərək huu'zɛn ou'bə meɪ/; 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

Create account Log in

Article Talk Read View source View history Search

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

MAY 17, 2012 BY FLOYD BROWN 100 COMMENTS

Facebook Share 14.3K 14.3K 34 Twitter 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.



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 Equipping Americans Who Love Freedom

Home Categories Blogging Tools About Polls and Petitions Contact Us Write

Proof Obama Born in Kenya? Obama Literary Agent Says Yes

MAY 17, 2012 BY FLOYD BROWN 100 COMMENTS

Facebook Share 14.3K 14.3K 34 Twitter 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.

Barack Obama

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

MEDIA

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

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

Related: Barack Obama, Birthers, Obama Birth Certificate

Facebook Share 802 16.8K 34 LinkedIn 6 Email 0

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 Twitter 26 Facebook Share 14.3K 14.3K 34 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> ?



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.

