

# Knowledge Fusion and Knowledge-Based Trust

*Xin Luna Dong*  
*Google Inc.*

Collaboration with: *Van Dang, Evgeniy Gabrilovich, Wilko Horn,  
Camillo Lugaresi, Kevin Murphy, Shaohua Sun, Wei Zhang*

# The Many Examples I've Used About Web Data Quality

The screenshot shows the nj.com website interface. At the top, there's a navigation bar with the nj.com logo and a search bar. Below the navigation bar, there's a section for 'Find Flights from Newark to Lyon' with a link to 'Click Here for Low Fares!'. The main content area is titled 'Sunglasses in Edison, NJ' and includes a map of Edison, NJ, with a red pin marking a location. To the left of the map, there's a search box for 'Find businesses & local listings' with fields for 'What' (e.g., doctor, plumber, pizza...) and 'Where' (City OR zip code). Below the search box, there's a 'Search' button and a link to 'Is your business listed here?'. To the right of the map, there's an advertisement for 'Computer Training A+, MCSE, Networking' by PC AGE, featuring a laptop and a graduation cap. The website is displayed in a browser window with the address bar showing 'Internet' and the status bar showing '100%'.

**nj.com**

**ORBITZ** Find Flights from Newark to Lyon  
Click Here for Low Fares!  
[www.orbitz.com](http://www.orbitz.com) - [click here](#)

Site Search Search Local Business Listings  
Town, Keyword, Local Businesses, Web ID **Go**

Home News Weather Sports Entertainment Living Interact Jobs Autos Real Estate Classifieds Place an ad

Edison, NJ > Retail Shopping > Apparel > Apparel Accessories > Sunglasses > Vision 27 Inc

**Vision 27 Inc (732) 248-8889**  
1945 State Route 27, Edison, NJ 08817

Bookmark To:

**Sunglasses in Edison, NJ**  
Other cities on NJ.com

Find businesses & local listings  
**What** e.g. doctor, plumber, pizza...  
Business name or category  
**Where** City OR zip code  
Boonton, NJ  
**Search**

Is your business listed here?  
• Claim your listing for free  
• Manage your listing

Computer Training  
A+, MCSE, Networking  
3 New Jersey Location!  
**PC AGE**  
www.PCAGE.com  
Learn More  
Ads by Google

Internet 100%

# The Many Examples I've Used About Web Data Quality

The screenshot displays the Google Maps interface with a search for 'mamlouk, NYC'. The search results on the left list two entries for 'Mamlouk Restaurant' at 265 E 10th St and 211 E 4th St, both with 4.1 reviews. The right panel provides a detailed view of the restaurant at 211 E 4th St, including its address, phone number, and a list of reviews. The reviews are highly positive, praising the food, atmosphere, and service. The map shows the restaurant's location in Manhattan, near the East River and the FDR Expressway.

Google maps mamlouk, NYC Search Maps Show search options

Find businesses, addresses and places of interest. [Learn more](#)

Get Directions My Maps

Narrow by: User Rating ▼

**mamlouk, near New York, NY**

Categories: [Restaurants](#)

**Mamlouk Restaurant** ☆ - [more info](#)  
265 E 10th St, New York, NY - (212) 529-3477  
[1 review](#) - [Write a review](#)  
"The Moroccan atmosphere at this Middle Eastern restaurant located in the east ..."

**Mamlouk Restaurant** ☆ - [more info](#)  
211 E 4th St, New York, NY - (212) 529-3477  
★★★★☆ 4.1 reviews - [Write a review](#)  
"Stepping into this middle eastern restaurant is like stepping into a different ..."

Sponsored Links

[Mamlouk at Amazon](#)  
Low Prices on **Mamlouk**  
Free Shipping Available. Buy Today!  
[www.Amazon.com/Books](#)

[See all 60 results for mamlouk.](#)

Can't find what you're looking for?  
• [Add a place to the map](#)

**Mamlouk Restaurant** Print Send Link

★★★★☆ 4.1 reviews - [Write a review](#)  
211 E 4th St  
New York, NY 10009-7213  
(212) 529-3477

Get directions: [To here](#) - [From here](#)  
[Edit](#)

Overview Details (8) Reviews (41) Photos & Videos (2) User Content (16) Web Pages (36)

[Write a review](#)

★★★★★ Beautiful and Cozy - Mary L. - Jan 8, 2009  
Stepping into this middle eastern restaurant is like stepping into a different country. The food is great and the atmosphere is better. The place is filled with beautiful colors an pillows adorn every seating arrangement. ...  
Was this review helpful? [Yes](#) - [No](#)  
[More from insiderpages.com](#)

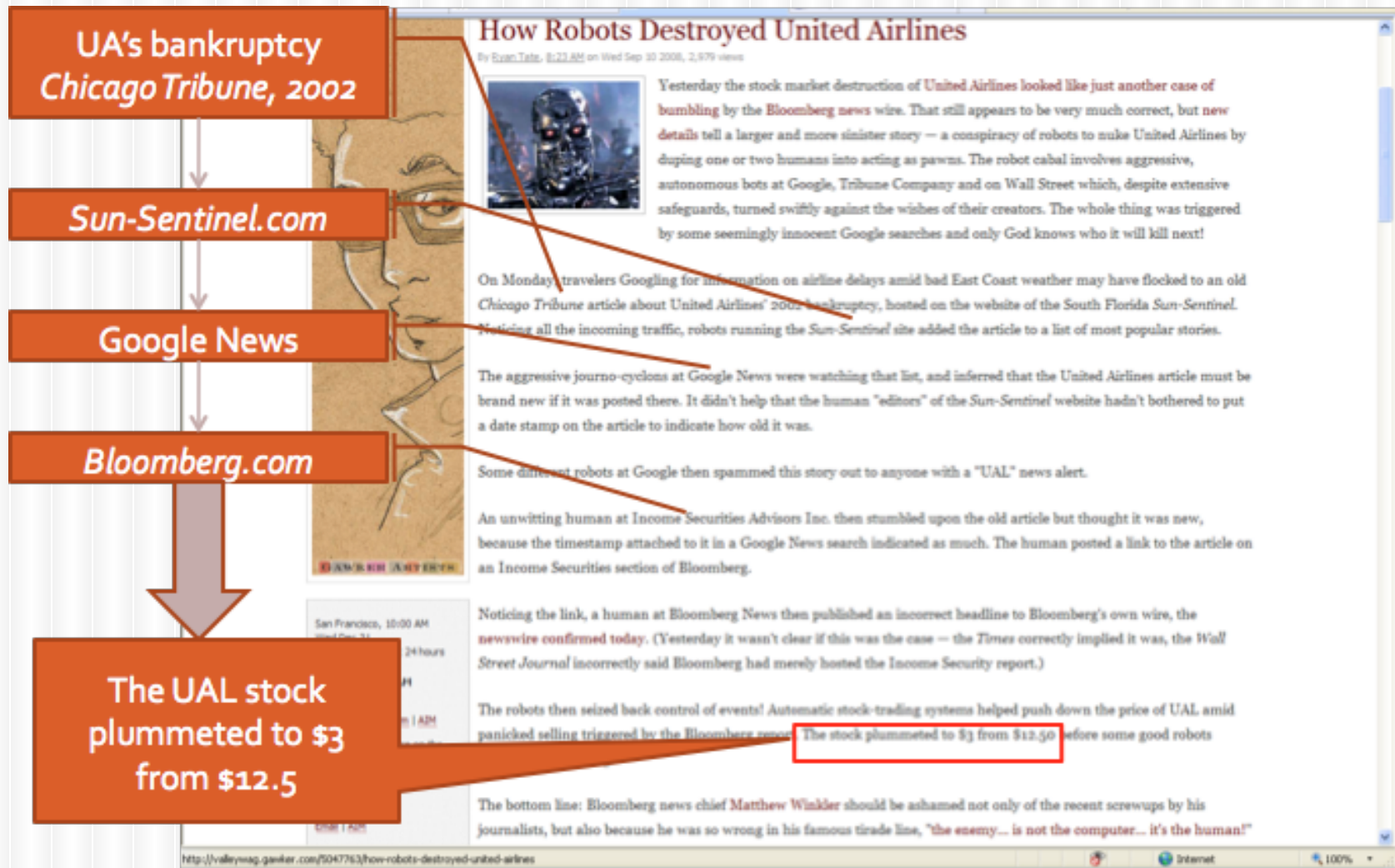
Excellent - ambiance, flavors, service - Mar 11, 2009  
I attended a large event at Mamlouk last month. The food was chosen by the hosts, so I never even saw the menu. Everything that came to the table was wonderful. I am vegetarian, so I ate none of the meat dishes, but I was delighted ...  
Was this review helpful? [Yes](#) - [No](#)  
[More from menutopia.com](#)

★★★★★ Mamlouk - ogiovetti - Nov 5, 2008  
Good Middle Eastern is a rare commodity in Manhattan. Yes, I love the falafel and gyro trucks as much as the next guy, and don't get in between me and a mouth-watering doner kebab. But here,

mamas mu... | madalein... | lonesome... | lomito, ...

Internet 100%

# The Many Examples I've Used About Web Data Quality



# The Many Examples I've Used About Web Data Quality

Telegraph.co.uk

Home News Sport Finance Comment Travel Lifestyle Culture Fas

UK World Politics Celebrities Obituaries Weird Earth Science Health News Education

HOME > NEWS > NEWS TOPICS > HOW ABOUT THAT?

## Steve Jobs obituary published by Bloomberg

An obituary of very-much-alive Apple founder Steve Jobs has been accidentally published by the respected Bloomberg business news wire.

By Matthew Moore

Last Updated: 7:05PM BST 28 Aug 2008



Steve Jobs was described as the man who 'refashioned the mobile phone' in the erroneous obituary. Photo: REUTERS

The story, marked "Hold for release – Do not use", was sent in error to the news service's thousands of corporate clients.

T Text Size

Email this article

Print this article

Share this article

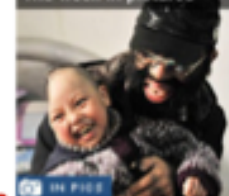
91 diggs digg it

How about that?

USA

News

The week in pictures



Pictures of the day

# The Many Examples I've Used About Web Data Quality

.....



**Maurice Jarre (1924-2009)** French Conductor and Composer

"One could say my life itself has been one long soundtrack. Music was my life, music brought me to life, and music is how I will be remembered long after I leave this life. When I die there will be a final waltz playing in my head and that only I can hear."

2:29, 30 March 2009

The screenshot shows the Wikipedia page for Maurice Jarre. At the top left is the Wikipedia logo. Below it, the text "WIKIPEDIA The Free Encyclopedia" is visible. The main heading is "Maurice Jarre". Below the heading, it says "From Wikipedia, the free encyclopedia". A red warning box states: "This is an old revision of this page, as edited by 86.42.227.123 (talk) at 02:29, 30 March 2009. It may differ significantly from the current revision." Below the warning box, there are links: "(diff) — Previous revision | Current revision (diff) | Newer revision — (diff)". A section titled "Quotes" contains the quote: "Nowadays, if a studio assumes that his film is bad, there is always an executive that gets more nervous than usual and thinks that if they change the music, the film will masterpiece. One could say my life itself has been one long soundtrack. Music was my life, music brought me to life, and music is how I will be remembered long after I leave this life. will be a final waltz playing in my head and that only I can hear."

# The Many Examples I've Used About Web Data Quality

## Numerous rumors after the Japan earthquake and tsunami

*"[Please spread the word] From my friend living in Chiba Prefecture. The weather forecast says it will rain from Monday. People living around Chiba, please be careful. The explosion at the Cosmo oil refinery will cause harmful substance to rise to clouds and become toxic rain. So when you go out, take your umbrella"*

*"The creator of Pokemon died today in the #tsunami, #Japan. RIP: Satoshi Tajiri. #prayforjapan." By xCyrusAndLovato*

*"The Japa*

*Relief aid from individuals*

*In order to avoid confusion, we ask that you please refrain*

*[from Chain letters with specific bank account information for donations are getting sent around*

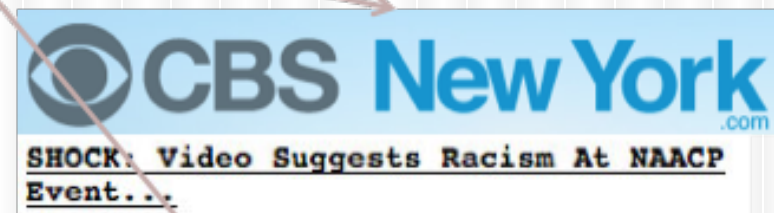
*Please Help Japan! Earthquake Weapons caused Tsunami*

# The Many Examples I've Used About Web Data Quality

.....



Posted by Andrew Breitbart  
In his blog



# Finally, My Own Example About Web Data Quality

arXiv.org > cs > arXiv:1502.03519

Computer Science > Databases

## Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, Wei Zhang

(Submitted on 12 Feb 2015)



**NewScientist**

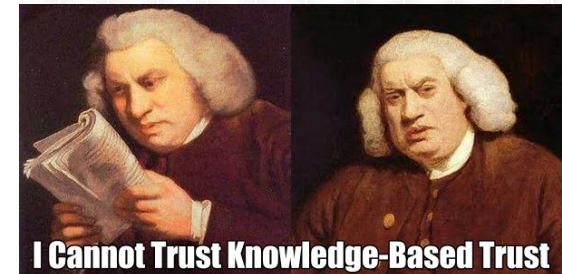
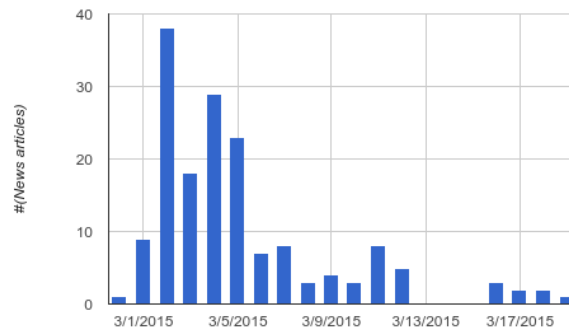
### Google wants to rank websites based on facts not links

28 February 2015 by [Hal Hodson](#)  
Magazine issue 3010. [Subscribe and save](#)

*The trustworthiness of a web page might help it rise up Google's rankings if the search giant starts to measure quality by facts, not just links*



Press Coverage



# Motivation for Knowledge-Based Trust (KBT)

---

- Providing a new perspective to evaluate Web source quality
- What we have now--*Exogenous signals*
  - Link-based
  - Search log and click-through rate
  - Web spam
- Key idea: Evaluate trustworthiness of sources by *the correctness of its factual information*--*Endogenous signals*

# Correctness of Factual Information



WIKIPEDIA  
The Free Encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikimedia Shop

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
Print/export

Languages

العربية  
Aragonés

Create account Log in

Article Talk Read View source Search

## United States

From Wikipedia, the free encyclopedia  
(Redirected from [USA](#))

*For other uses, see [US \(disambiguation\)](#), [USA \(disambiguation\)](#), and [United States \(disambiguation\)](#).*

The **United States of America** (**USA**), commonly referred to as the **United States** (**US**), **America** or simply **the States**, is a [federal republic](#)<sup>[10][11]</sup> consisting of 50 [states](#) and a [federal district](#). The 48 [contiguous states](#) and the federal district of [Washington, D.C.](#), are in central [North America](#) between [Canada](#) and [Mexico](#). The state of [Alaska](#) is the northwestern part of North America and the state of [Hawaii](#) is an [archipelago](#) in the mid-[Pacific](#). The country also has five populated and nine unpopulated [territories](#) in the Pacific and the [Caribbean](#). At 3.79 million square miles (9.83 million km<sup>2</sup>) in total and with around 317

### United States of America



Flag

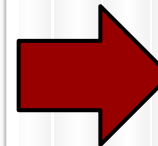


Great Seal

**Motto:**  
"In God we trust" (official)<sup>[1][2][3]</sup>  
"E pluribus unum" (Latin) (traditional)  
"Out of many, one"

**Anthem:** ["The Star-Spangled Banner"](#)





Fact 1	✓
Fact 2	✓
Fact 3	✗
Fact 4	✓
Fact 5	✗
Fact 6	✓
Fact 7	✓
Fact 8	✓
Fact 9	✓
Fact 10	✗
...	...
Accu	0.7

.....  
Why Should We Care?

# I. Tail Sources Can Be Useful

who play the stumble

Web Videos News Images Maps More Search tools


About 23,500,000 results (0.49 seconds)

Duke Robillard shows how to play the double stops in **Freddie King's** "The Stumble", that many guitar players (including myself) seem to "stumble" over. Duke plays 'm with a flatpick and uses one finger to pick the high E-string, but **Freddie** used finger picks, so he could pluck the two strings easier. Jul 17, 2010

[How to play the double stops in The Stumble - YouTube](#)  
[www.youtube.com/watch?v=POd9wJnixsA](http://www.youtube.com/watch?v=POd9wJnixsA)

Feedback

[How to play the double stops in The Stumble - YouTube](#)  
[www.youtube.com/watch?v=POd9wJnixsA](http://www.youtube.com/watch?v=POd9wJnixsA) ▾  
Jul 17, 2010 - Uploaded by Gibbo  
Duke Robillard shows how to play the double stops in **Freddie King's** "The Stumble", that many guitar ...




Good answer for an award-winning song


# I. Tail Sources Can Be Useful

who play guitar going home

Web Videos Images News Shopping More Search tools

About 24,200,000 results (0.46 seconds)

5:25 Play next Play now Going Home - YouTube  
 [www.youtube.com/watch?v=UIZtbwLoKWE](http://www.youtube.com/watch?v=UIZtbwLoKWE)  
Sep 30, 2008 - Uploaded by prann2003  
Playing Mark Knopfler's "Going Home" from OST "Local Hero"  
Stereo Audio Combine & Simplify Guitar ...

How to Play "Hold On, We're Going Home" by Drake, Majid ...  
 [www.youtube.com/watch?v=UxByes5JXUA](http://www.youtube.com/watch?v=UxByes5JXUA)  
Feb 28, 2014 - Uploaded by martyzsongs  
How to Play "Hold On, We're Going Home" by Drake, Majid  
Jordan - Acoustic Songs on Guitar .... Your a guitar goddess! ...  
New guitar Marty?

Alvin Lee Is Going Home: 'Ten Years After' Guitarist Dies ...  
[www.npr.org/.../alvin-lee-is-going-home-ten-years-after-guitarist-di...](http://www.npr.org/.../alvin-lee-is-going-home-ten-years-after-guitarist-di...) NPR  
Mar 6, 2013 - But for those of us of a certain age who wished they could play a guitar well, it's Lee's furious fretting on "I'm Going Home" — famously ...

Missing answer for a not-so-popular song

# I. Tail Sources Can Be Useful

The screenshot shows the homepage of Backingtrackguitar.com. The header features the site name and navigation links: Backing Tracks, Guestbook, and Terms Of Use. Below the header is a search bar with social media icons and a search button. A red box highlights a search result for 'going home backing track' by Dire Straits. The result includes a play button, a download link, and a table with details about the track.

going home backing track	
Download	
Information about backing track	
Artist	Dire Straits
Rating	★★★★★☆☆☆☆
Filesize	5.29 Mb
Length	00:03:51

Very precise info on guitar players but low Page Rank

# II. Popular Websites May Not Be Trustworthy

## Gossip Websites

<http://www.ebizmba.com/articles/gossip-websites>

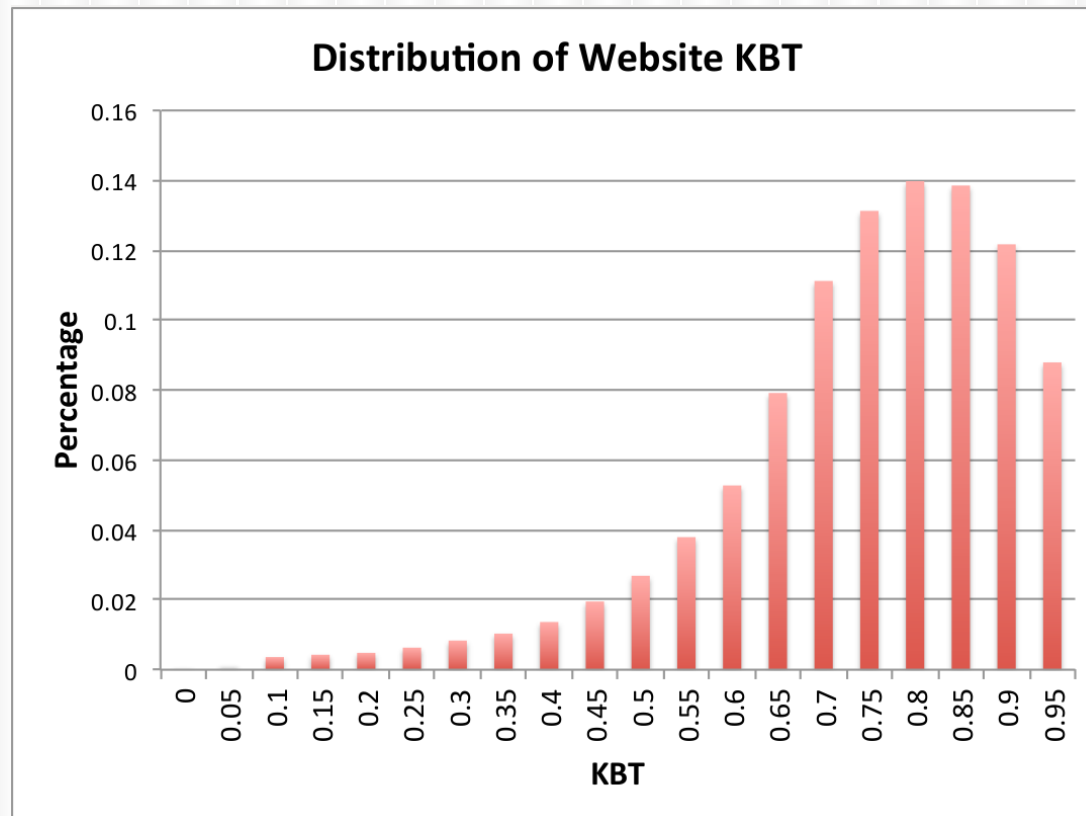
Domain
www.eonline.com
perezhilton.com
radaronline.com
www.zimbio.com
mediatakeout.com
gawker.com
www.popsugar.com
www.people.com
www.tmz.com
www.fishwrapper.com
celebrity.yahoo.com
wonderwall.msn.com
hollywoodlife.com
www.wetpaint.com

14 out of 15 have a PageRank among top 15% of the websites

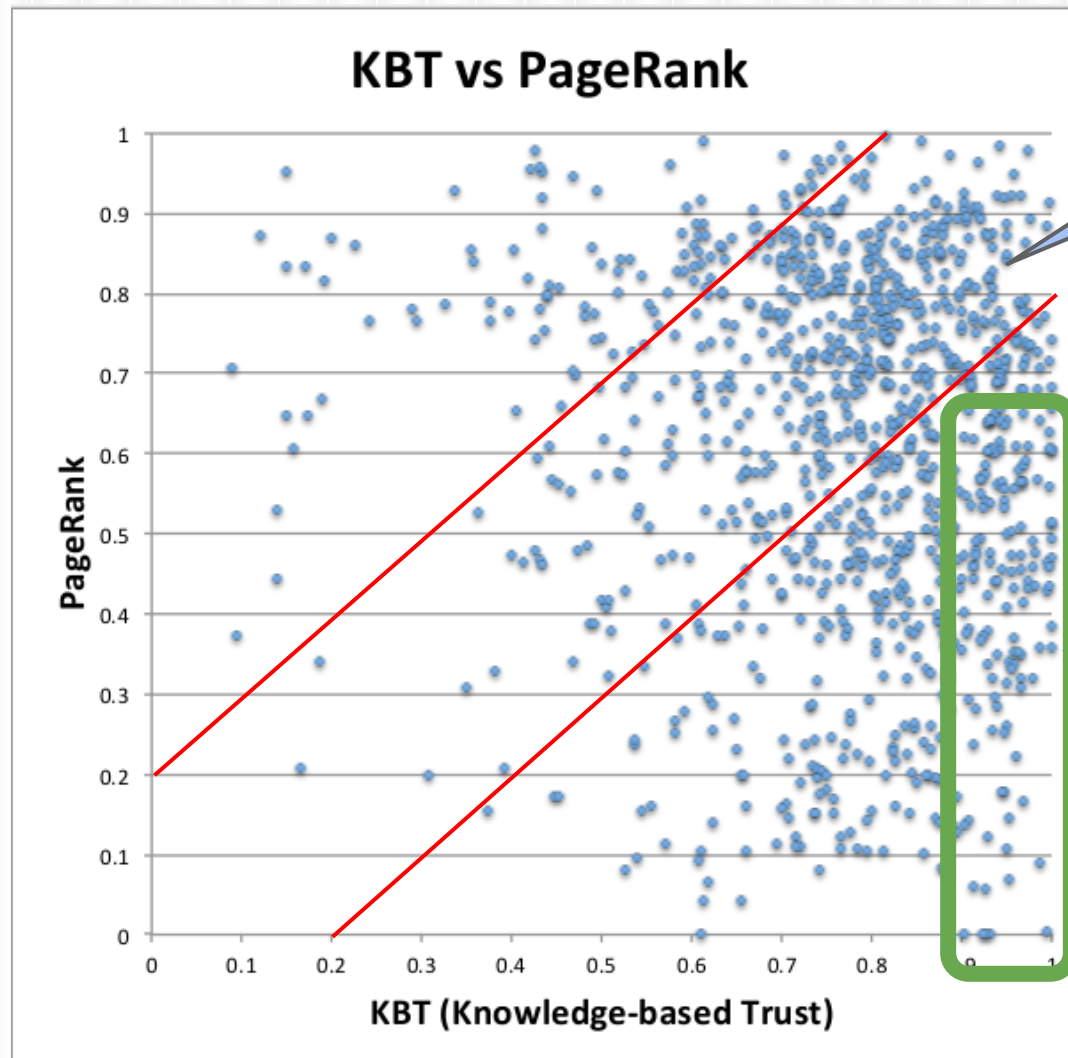
# How Can Trustworthiness Help?

# Knowledge-Based Trust (KBT)

Trustworthiness in  $[0,1]$  for 5.6M websites and 119M webpages



# Knowledge-Based Trust vs. PageRank



Correlated scores

Often tail sources w. high trustworthiness

# I. Tale Sources w. Low PageRank May Provide Valuable Info

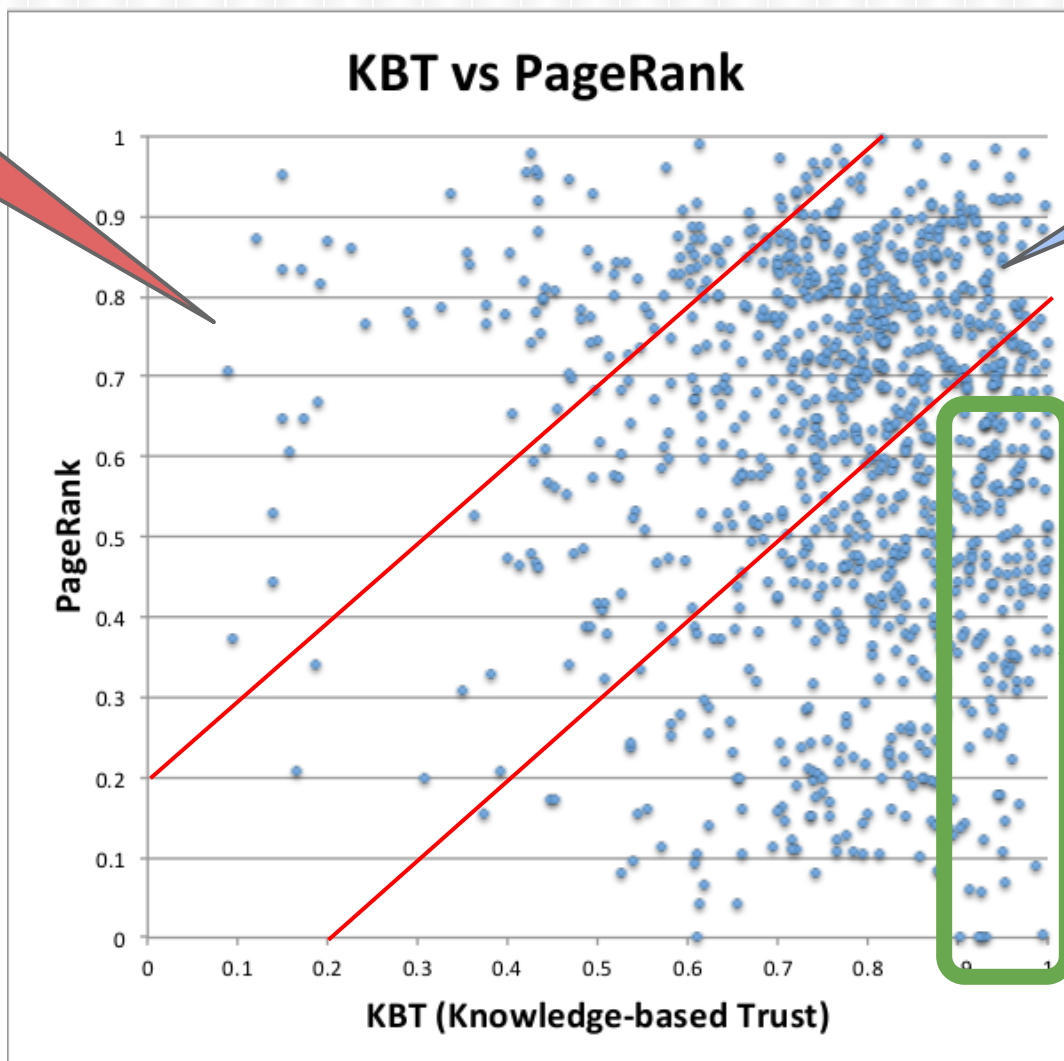
The collage displays five distinct websites, illustrating a variety of content and user interfaces:

- salary.com**: A career and salary website with a green header, navigation tabs (Salary, Job Search, Education, Career Development, Work & Life, Features, Business Products), and a central section for job seekers and employers. It includes a survey about workplace bullying and a 'Save 44% on 2 Job Postings' offer.
- WOYLAA**: A website for Korean drama episodes, featuring a blue header, a search bar, and a grid of episode listings with synopsis and subtitle information.
- Backingtrackguitar.com**: A website for guitar backing tracks, with a green header, navigation tabs (Backing Tracks, Guestbook, Terms Of Use), and a search bar. It includes a 'The Platinum Card' advertisement.
- amazonmom**: An Amazon advertisement for '20% OFF DIAPERS' featuring a baby image and a 'Learn more' button.
- UK Map Site**: A website featuring a map of the United Kingdom with various regions labeled (e.g., Scotland, Northern Ireland, Wales, England). It includes a sidebar with 'Branches' (Agricultural, Dairies, Farming, Fish, Livestock, Mixed Crops, Services, Tree & Forestry) and 'Countries' (England, Northern Ireland, Scotland, Wales).

Among 100 sampled websites, 85 are indeed trustworthy.

# Knowledge-Based Trust vs. PageRank

Often sources  
w. low accuracy



Correlated  
scores

Often tail  
sources w. high  
trustworthiness

# II. Popular Websites May Not Be Trustworthy

## Gossip Websites

<http://www.ebizmba.com/articles/gossip-websites>

Domain
www.eonline.com
perezhilton.com
radaronline.com
www.zimbio.com
mediatakeout.com
gawker.com
www.popsugar.com
www.people.com
www.tmz.com
www.fishwrapper.com
celebrity.yahoo.com
wonderwall.msn.com
hollywoodlife.com
www.wetpaint.com

14 out of 15 have a PageRank among top 15% of the websites

All have knowledge-based trust in bottom 50%

# II. Popular Websites May Not Be Trustworthy

YAHOO!  
ANSWERS

Entertainment & Music > Celebrities



## Why are British women so unattractive?

Seriously, what's with that? There are very few English chicks I would say are attractive yet so many countries who are drop-dead gorgeous.

**Update:** Catherine Zeta-Jones is from NEW ZEALAND!!!!!! Dummy!

**Update 2:** The SPICE GIRLS! Surely, you jest. Put them all together, all their best points, and then

**Update 3:** crazy\_lad wins the "moron" award for this question. He says all Americans are fat and being narrowminded! LMFAO at that IDIOT! LOL

☆ Follow 37 answers

[Are you getting your biggest tax refund?](#)

Get your taxes done right, and your biggest refund, guaranteed. Start for free today!

TurboTax Sponsored

[California Programs Contribute to STEM Careers](#)

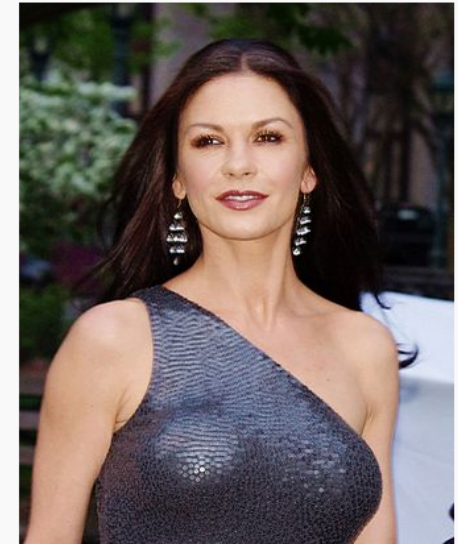
California's public school system contributes to STEM careers by offering science-centric activities

CBS Local Sponsored



Catherine Zeta-Jones

CBE




Zeta-Jones at the 2012 Tribeca Film Festival

<b>Born</b>	Catherine Zeta Jones 25 September 1969 (age 45) <a href="#">Swansea, Glamorgan, Wales</a>
<b>Nationality</b>	<a href="#">Welsh</a>
<b>Citizenship</b>	<a href="#">Britain</a>
<b>Alma mater</b>	<a href="#">Arts Educational Schools, London</a>
<b>Occupation</b>	<a href="#">Actress</a>
<b>Years active</b>	1981–present
<b>Spouse(s)</b>	<a href="#">Michael Douglas</a> (m. 2000)
<b>Children</b>	2


# III. Website Recommendation by Vertical

.....

Sula <input type="text" value="/food/cheese"/>			
Site	Accuracy	# Triples	Score 
en.wikipedia.org	0.76	1,184	5.36
www.cheesewiki.com	0.84	230	4.55
www.ranker.com	0.93	105	4.34
scratchpad.wikia.com	0.92	101	4.26
www.marcellathecheesemonger.com	0.90	109	4.25
www.cheeseplatesf.com	0.90	114	4.25
cheesetique.com	0.93	83	4.12
www.lafromagerie.co.uk	0.93	62	3.87
adrian1974fulga.wordpress.com	0.90	71	3.83
epicurefoodscorp.com	0.78	131	3.81
www.gourmetfoodstore.com	0.83	91	3.77
www.sheridanscheesemongers.com	0.85	71	3.65
www.sfgate.com	0.89	54	3.56
cheeseandchampagne.com	0.90	48	3.50
www.sciencedirect.com	0.69	161	3.50
about-france.com	0.98	34	3.48
www.cookipedia.co.uk	0.70	132	3.44
www.cheese.com	0.86	50	3.39
www.pennmac.com	0.76	87	3.38

# III. Website Recommendation by Vertical

.....

Sula	<input type="text" value="/computer"/>		
Site	Accuracy	# Triples	Score 
ru.wikipedia.org	0.85	3,597	6.98
ja.wikipedia.org	0.73	3,032	5.88
en.wikipedia.org	0.53	36,489	5.57
www.opensourcesoftwaredirectory.com	0.91	323	5.26
wpedia.goo.ne.jp	0.63	2,846	5.03
www.ranker.com	0.81	308	4.66
packages.gentoo.org	0.65	942	4.45
uk.wikipedia.org	0.73	420	4.43
freecode.com	0.57	2,085	4.36
www.file.net	0.65	756	4.34
gpo.zugaina.org	0.65	711	4.28
gentoobrowse.randomdan.homeip.net	0.80	208	4.27
whatis.techtarget.com	0.56	1,417	4.08
www.starringthecomputer.com	0.87	101	4.02
www.linuxlinks.com	0.67	411	4.02
companies.findthebest.com	0.81	135	3.95
www.fileinfo.com	0.53	1,770	3.95
file.downloadatoz.com	0.62	571	3.92
www.zwodnik.com	0.55	1,187	3.92
www.system-tray-cleaner.com	0.70	273	3.92
bitnami.com	0.70	266	3.91

.....  
Now, How to Compute KBT?

# Key Idea in KBT



Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikimedia Shop

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
Print/export

Languages  
Acèh  
Адыгэбзэ  
★ Afrikaans  
Alemannisch  
አማርኛ  
Ænglisc  
Англис  
★ العربية  
★ Aragonés  
ᠠᠷᠠᠩᠭᠡᠳᠡ

Create account Log in

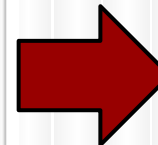
Article Talk Read View source Search

## United States

From Wikipedia, the free encyclopedia  
(Redirected from [USA](#))

*For other uses, see [US \(disambiguation\)](#), [USA \(disambiguation\)](#), and [United States \(disambiguation\)](#).*

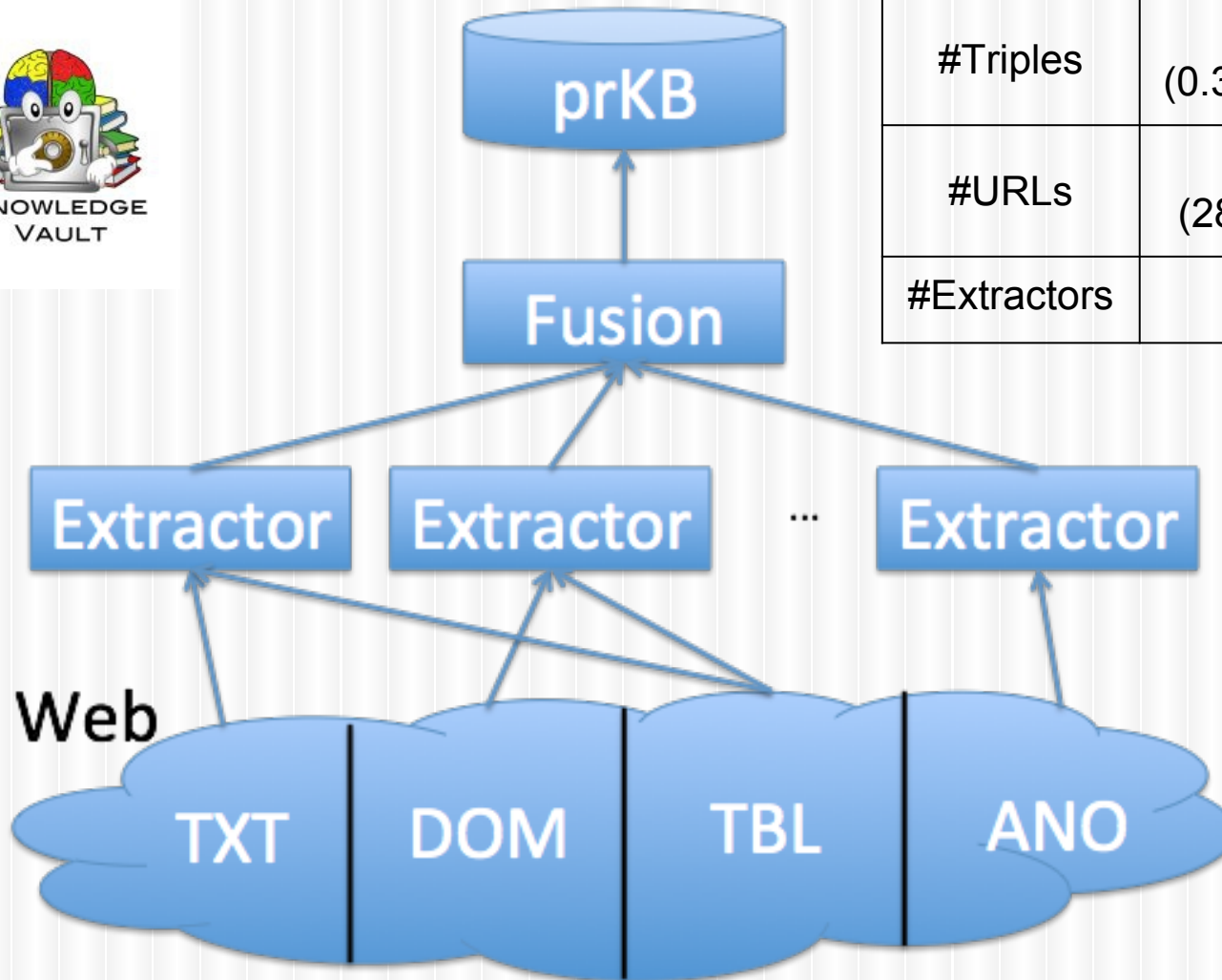
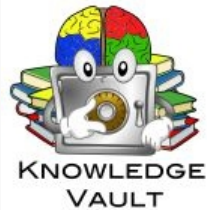
The **United States of America** (**USA**), commonly referred to as the **United States** (**US**), **America** or simply **the States**, is a [federal republic](#)<sup>[10][11]</sup> consisting of 50 [states](#) and a [federal district](#). The 48 [contiguous states](#) and the federal district of [Washington, D.C.](#), are in central [North America](#) between [Canada](#) and [Mexico](#). The state of [Alaska](#) is the northwestern part of North America and the state of [Hawaii](#) is an [archipelago](#) in the mid-[Pacific](#). The country also has five populated and nine unpopulated [territories](#) in the Pacific and the [Caribbean](#). At 3.79 million square miles (9.83 million km<sup>2</sup>) in total and with around 317



Fact 1	✓
Fact 2	✓
Fact 3	✗
Fact 4	✓
Fact 5	✗
Fact 6	✓
Fact 7	✓
Fact 8	✓
Fact 9	✓
Fact 10	✗
...	...
<b>Accu</b>	<b>0.7</b>

# Knowledge Vault– Probabilistic Knowledge Fusion.....

[SIGKDD, 2014]  
[VLDB, 2014]



#Triples	3.0B (0.3B w. $pr \geq 0.7$ )
#URLs	2.5B (28M Websites)
#Extractors	16

# KV Makes This Possible



Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikimedia Shop

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
Print/export

Languages  
Acèh  
Адыгэбзэ  
★ Afrikaans  
Alemannisch  
አማርኛ  
Ænglisc  
Англис  
★ العربية  
★ Aragonés  
ᠠᠷᠠᠩᠭᠡᠳᠡ

Create account Log in

Article Talk Read View source Search

## United States

From Wikipedia, the free encyclopedia  
(Redirected from [USA](#))

*For other uses, see [US \(disambiguation\)](#), [USA \(disambiguation\)](#), and [United States \(disambiguation\)](#).*

The **United States of America** (**USA**), commonly referred to as the **United States** (**US**), **America** or simply **the States**, is a [federal republic](#)<sup>[10][11]</sup> consisting of 50 [states](#) and a [federal district](#). The 48 [contiguous states](#) and the federal district of [Washington, D.C.](#), are in central [North America](#) between [Canada](#) and [Mexico](#). The state of [Alaska](#) is the northwestern part of North America and the state of [Hawaii](#) is an [archipelago](#) in the mid-[Pacific](#). The country also has five populated and nine unpopulated [territories](#) in the Pacific and the [Caribbean](#). At 3.79 million square miles (9.83 million km<sup>2</sup>) in total and with around 317

United States of America


Flag


Great Seal

Motto:  

"In God we trust" (official)<sup>[1][2][3]</sup>

"E pluribus unum" (Latin) (traditional)

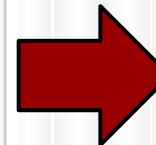
"Out of many, one"

Anthem: ["The Star-Spangled Banner"](#)

0:00 CC

MENU





Fact 1	✓
Fact 2	✓
Fact 3	✗
Fact 4	✓
Fact 5	✗
Fact 6	✓
Fact 7	✓
Fact 8	✓
Fact 9	✓
Fact 10	✗
...	...
<b>Accu</b>	<b>0.7</b>

# KV Makes This Possible



Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikimedia Shop

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
Print/export

Languages  
 Acèh  
 Адыгэбзэ  
 Afrikaans  
 Alemannisch  
 አማርኛ  
 Ænglisc  
 АӀсӀӀа  
 العربية  
 Aragonés  
 العربية

Create account Log in

Article Talk Read View source Search

## United States

From Wikipedia, the free encyclopedia  
(Redirected from [USA](#))

*For other uses, see [US \(disambiguation\)](#), [USA \(disambiguation\)](#), and [United States \(disambiguation\)](#).*

The **United States of America** (**USA**), commonly referred to as the **United States** (**US**), **America** or simply **the States**, is a federal republic<sup>[10][11]</sup> consisting of 50 [states](#) and a federal district. The 48 contiguous states and the federal district of [Washington, D.C.](#), are in central [North America](#) between [Canada](#) and [Mexico](#). The state of [Alaska](#) is the northwestern part of North America and the state of [Hawaii](#) is an [archipelago](#) in the mid-Pacific. The country also has five populated and nine unpopulated [territories](#) in the Pacific and the [Caribbean](#). At 3.79 million square miles (9.83 million km<sup>2</sup>) in total and with around 317

United States of America



Flag



Great Seal

Motto:

"In God we trust" (official)<sup>[1][2][3]</sup>

"E pluribus unum" (Latin) (traditional)

"Out of many, one"

Anthem: "The Star-Spangled Banner"

0:00

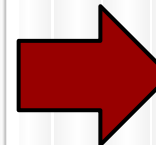
CC

⏮

⏭

MENU





Triple 1	1.0
Triple 2	0.9
Triple 3	0.3
Triple 4	0.8
Triple 5	0.4
Triple 6	0.8
Triple 7	0.9
Triple 8	1.0
Triple 9	0.7
Triple 10	0.2
...	...
<b>Accu</b>	<b>0.7</b>

# Challenges

Create account Log in

Article Talk Read View source Search

## United States

From Wikipedia, the free encyclopedia  
(Redirected from USA)

For other uses, see [US \(disambiguation\)](#), [USA \(disambiguation\)](#), and [United States \(disambiguation\)](#).

The **United States of America** (**USA**), commonly referred to as the **United States** (**US**), **America** or simply **the States**, is a federal republic<sup>[10][11]</sup> consisting of 50 states and a federal district. The 48 contiguous states and the federal district of [Washington, D.C.](#), are in central [North America](#) between [Canada](#) and [Mexico](#). The state of [Alaska](#) is the northwestern part of North America and the state of [Hawaii](#) is an [archipelago](#) in the mid-Pacific. The country also has five populated and nine unpopulated [territories](#) in the Pacific and the [Caribbean](#). At 3.79 million square miles (9.83 million km<sup>2</sup>) in total and with around 317

**United States of America**

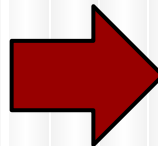
 

Flag Great Seal

**Motto:**  
"In God we trust" (official)<sup>[1][2][3]</sup>  
"E pluribus unum" (Latin) (traditional)  
"Out of many, one"

**Anthem:** "The Star-Spangled Banner"  
0:00  MENU





Triple 1	1.0
Triple 2	0.9
Triple 3	0.3
Triple 4	0
Triple 5	0.4
Triple 6	0.8
Triple 7	0.9
Triple 8	1.0
Triple 9	0.7
Triple 10	0.2
...	...
<b>Accu</b>	<b>0.7</b>

1. How to decide if a triple is indeed claimed by the source instead of an *extraction error*?

# Extractions Can Be Wrong

.....

- (Obama, nationality, Kenya)

2087 extractions:

- Example of a correct extraction

<http://beforeitsnews.com/obama-birthplace-controversy/2013/04/alabama-supreme-court-chief-justice-roy-moore-to-preside-over-obama-eligibility-case-2458624.html>

2006: Obama In Kenya: I Am So Proud To Come Back Home - [VIDEO HERE](#).

2007: Michelle Obama Declares Obama Is Kenyan And America Is Mean - [VIDEO HERE](#).

2008: Michelle Obama Declares Barack Obama's Home Country Is Kenya - [VIDEO HERE](#).

FLASHBACK: Obama Is The Original Birther! Obama In 1991 Stated In His Own Bio He Was Born In Kenya. [DETAILS HERE](#).

- Example of a wrong extraction

<http://www.monitor.co.ug/News/National/US+will+respect+winner+of+Kenya+election++Obama+says/-/688334/1685814/-/ksxagx/-/index.html>

US will respect winner of  
Kenya election, Obama says

SHARE BOOKMARK PRINT RATING ☆☆☆☆☆

# Extractions Can Be Wrong

.....

- (Obama, nationality, USA)

2481 extractions:

- Example of a correct extraction

<http://www.dogonews.com/2009/10/9/a-nobel-prize-for-our-awesome-president>

- Example of a wrong extraction

<http://blogs.telegraph.co.uk/news/timstanley/100169248/barack-obamas-life-story-contains-myth-not-truth-says-biographer-so-why-did-the-media-report-it-as-truth/>

## Tim Stanley

Dr Tim Stanley is a historian of the United States. His new book about Hollywood politics is out in May. His personal website is [www.timothystanley.co.uk](http://www.timothystanley.co.uk) and you can follow him on Twitter [@timothy\\_stanley](https://twitter.com/timothy_stanley).

 Follow 15.6K followers



Barack Obama's life story contains 'myth, not truth', says biographer. So why did the media report it as truth?

# Break-Down for Error Reasons

---

- Random sample on 25 false triples
  - Triple-identification errors: 11 (44%)  
*E.g., taking part of album name as album artist*
  - Entity-linkage errors: 11 (44%)
  - Predicate-linkage errors: 5 (20%)
  - Source-data errors: 1 (4%)
- Extraction errors dominate
- We should not penalize a source for extraction errors

# Challenges



WIKIPEDIA  
The Free Encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikimedia Shop

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
Print/export

Languages  
Acèh  
Адыгэбзэ  
Afrikaans  
Alemannisch  
አማርኛ  
Ænglisc  
Англис  
العربية  
Aragonés

Create account Log in

Article Talk Read View source Search

## United States

From Wikipedia, the free encyclopedia  
(Redirected from USA)

For other uses, see [US \(disambiguation\)](#), [USA \(disambiguation\)](#), and [United States \(disambiguation\)](#).

The **United States of America** (**USA**), commonly referred to as the **United States** (**US**), **America** or simply **the States**, is a [federal republic](#)<sup>[10][11]</sup> consisting of 50 [states](#) and a [federal district](#). The 48 [contiguous states](#) and the federal district of [Washington, D.C.](#), are in central [North America](#) between [Canada](#) and [Mexico](#). The state of [Alaska](#) is the northwestern part of North America and the state of [Hawaii](#) is an [archipelago](#) in the mid-[Pacific](#). The country also has five populated and nine unpopulated [territories](#) in the Pacific and the [Caribbean](#). At 3.79 million square miles (9.83 million km<sup>2</sup>) in total and with around 317

**United States of America**

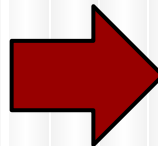
 

Flag Great Seal

**Motto:**  
"In God we trust" (official)<sup>[1][2][3]</sup>  
"E pluribus unum" (Latin) (traditional)  
"Out of many, one"

**Anthem:** "The Star-Spangled Banner"  
0:00  MENU



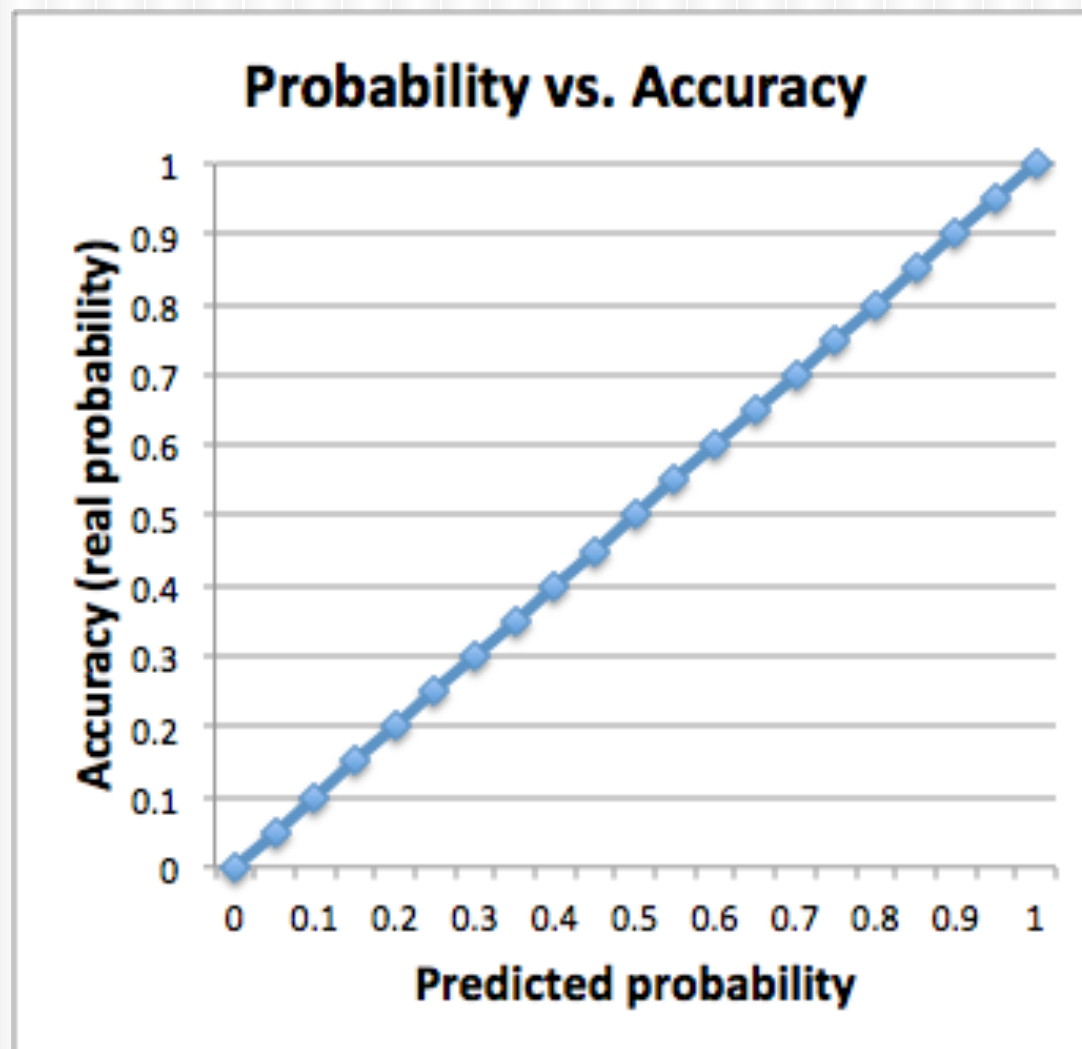


Triple 1	1.0
Triple 2	0.9
Triple 3	0.3
Triple 4	0
Triple 5	0.4
Triple 6	0.8
Triple 7	0.9
Triple 8	1.0
Triple 9	0.7
Triple 10	0.2
...	...
<b>Accu</b>	<b>0.7</b>

1. How to decide if a triple is indeed claimed by the source instead of an *extraction error*?

2. How to compute well-calibrated triple probability?

# Triple Probabilities Should be Well-Calibrated



# Challenges



WIKIPEDIA  
The Free Encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikimedia Shop

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
Print/export

Languages  
Acèh  
Адыгэбзэ  
Afrikaans  
Alemannisch  
አማርኛ  
Ænglisc  
Англис  
العربية  
Aragonés

Article Talk Read View source Search

## United States

From Wikipedia, the free encyclopedia  
(Redirected from [USA](#))

For other uses, see [US \(disambiguation\)](#), [USA \(disambiguation\)](#), and [United States \(disambiguation\)](#).

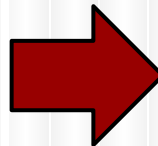
The **United States of America** (**USA**), commonly referred to as the **United States** (**US**), **America** or simply **the States**, is a federal republic<sup>[10][11]</sup> consisting of 50 states and a federal district. The 48 contiguous states and the federal district of [Washington, D.C.](#), are in central [North America](#) between [Canada](#) and [Mexico](#). The state of [Alaska](#) is the northwestern part of North America and the state of [Hawaii](#) is an archipelago in the mid-Pacific. The country also has five populated and nine unpopulated territories in the Pacific and the Caribbean. At 3.79 million square miles (9.83 million km<sup>2</sup>) in total and with around 317

United States of America

Flag Great Seal

Motto:  
"In God we trust" (official)<sup>[1][2][3]</sup>  
"E pluribus unum" (Latin) (traditional)  
"Out of many, one"

Anthem: "The Star-Spangled Banner"

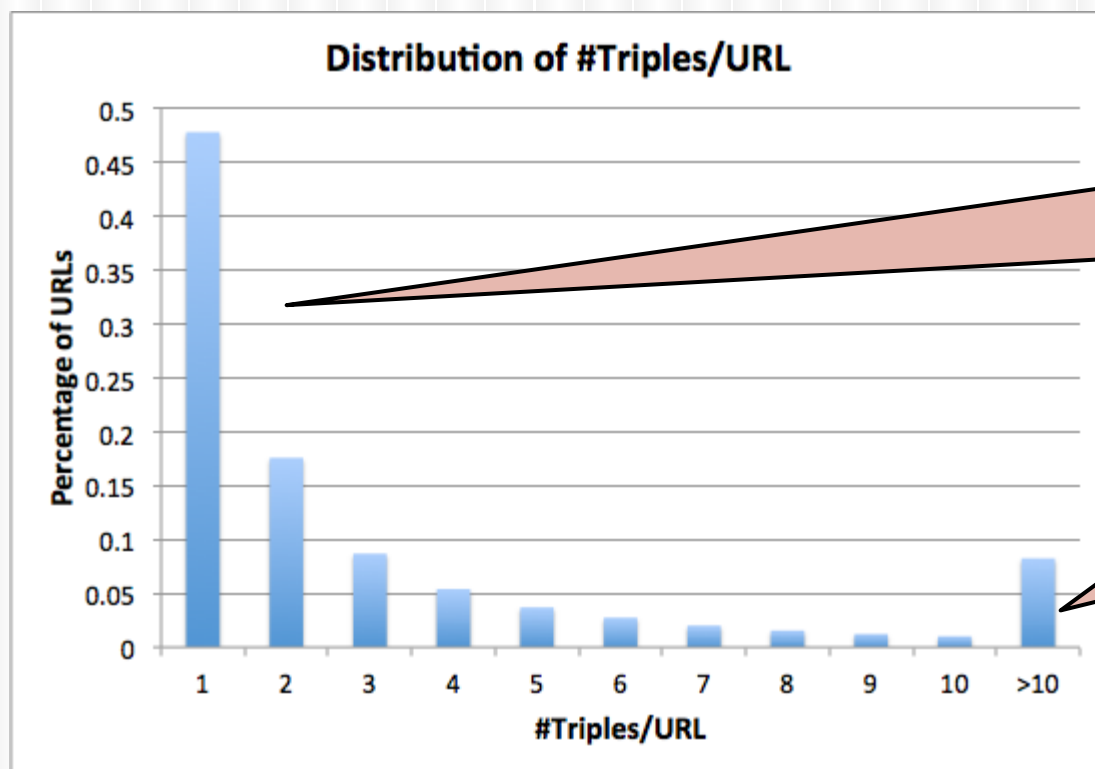
Triple 1	1.0
Triple 2	0.9
Triple 3	0.3
Triple 4	0
Triple 5	0.4
Triple 6	0.8
Triple 7	0.9
Triple 8	1.0
Triple 9	0.7
Triple 10	0.2
...	...
<b>Accu</b>	<b>0.7</b>

1. How to decide if a triple is indeed claimed by the source instead of an *extraction error*?

2. How to compute well-calibrated triple probability?

3. How to handle too large sources or too small sources?

# Heavy Head And Long Tail



74% URLs have  
at most 5  
*extracted* triples

Largest URLs  
have 50K triples

- Too large sources: Bottleneck in parallelization
- Too small sources: Not enough triples for trustworthiness evaluation

# KBT Strategies

.....

1. Graphical model--predict at the same time
  - a. extraction correctness
  - b. triple correctness
  - c. source accuracy
  - d. extractor precision/recall
2. Un(Semi-) supervised learning (Bayesian)
  - a. leverage source/extractor agreements
  - b. trust a source/extractor w. high quality
3. Source/extractor hierarchy
  - a. Break down “large” sources
  - b. Group “small” sources

# High-Level Intuition

[VLDB, 2009]

.....

## Researcher affiliation

	Src1	Src2	Src3
Jagadish	UM	ATT	UM
Dewitt	MSR	MSR	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

# High-Level Intuition

[VLDB, 2009]

## Researcher affiliation


	Src1	Src2	Src3
Jagadish	UM	ATT	UM
Dewitt	MSR	MSR	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

Voting--Trust the majority.

# High-Level Intuition

[VLDB, 2009]

## Researcher affiliation



	Src1	Src2	Src3
Jagadish	UM	ATT	UM
Dewitt	MSR	MSR	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

# High-Level Intuition

[VLDB, 2009]

.....

Researcher affiliation



	Src1	Src2	Src3
Jagadish	UM	ATT	UM
Dewitt	MSR	MSR	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

Quality-based--Give higher votes to more accurate sources.

# What About Extractions

[Sigmod, 2014]

## Extracted Harry Potter actors/actresses

Harry Potter	Ext1	Ext2	Ext3
Daniel	✓	✓	✓
Emma	✓		✓
Rupert	✓	✓	
Jonny	✓		
Eric			✓

# What About Extractions

[Sigmod, 2014]

## Extracted Harry Potter actors/actresses

Harry Potter	Ext1	Ext2	Ext3
Daniel	✓	✓	✓
Emma	✓		✓
Rupert	✓	✓	
Jonny	✓		
Eric			✓

Voting--Trust the majority.

# What About Extractions

[Sigmod, 2014]

## Extracted Harry Potter actors/actresses

Harry Potter	Ext1 (high rec)	Ext2 (high prec)	Ext3 (med prec/rec)
Daniel	✓	✓	✓
Emma	✓		✓
Rupert	✓	✓	
Jonny	✓		
Eric			✓

# What About Extractions

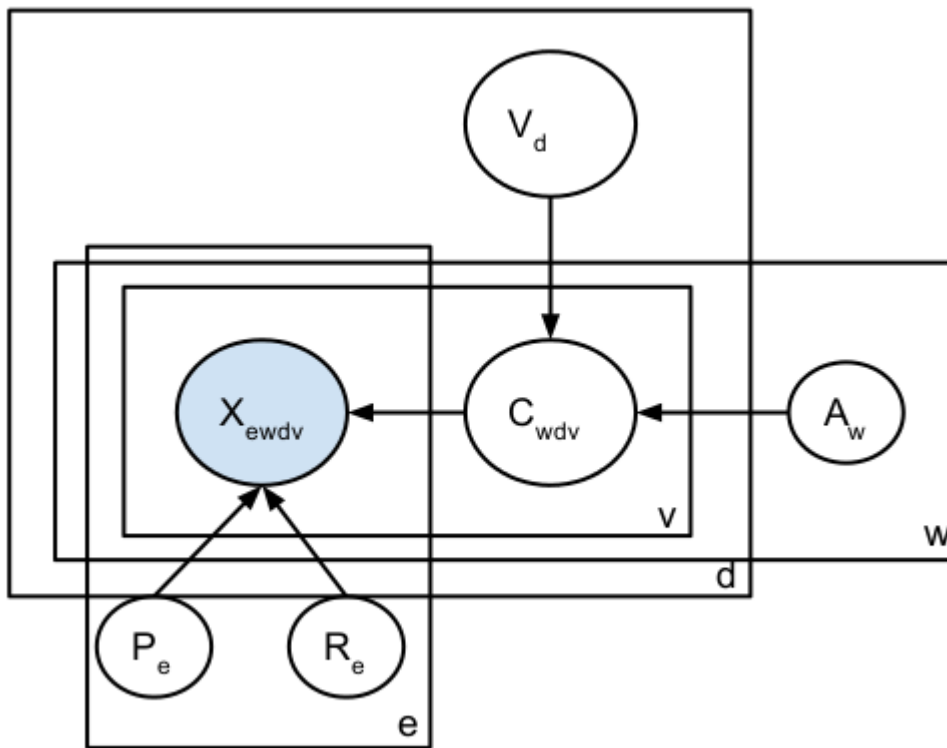
[Sigmod, 2014]

## Extracted Harry Potter actors/actresses

Harry Potter	Ext1 (high rec)	Ext2 (high prec)	Ext3 (med prec/rec)
Daniel	✓	✓	✓
Emma	✓		✓
Rupert	✓	✓	
Jonny	✓		
Eric			✓

Quality-based--More likely to be correct if extracted by high-precision sources; more likely to be wrong if not extracted by high-recall sources

# Graphical Model [VLDB, 2015]



## Observations

- $X_{ewdv}$ : whether extractor  $e$  extracts from source  $w$  the  $(d,v)$  item-value pair

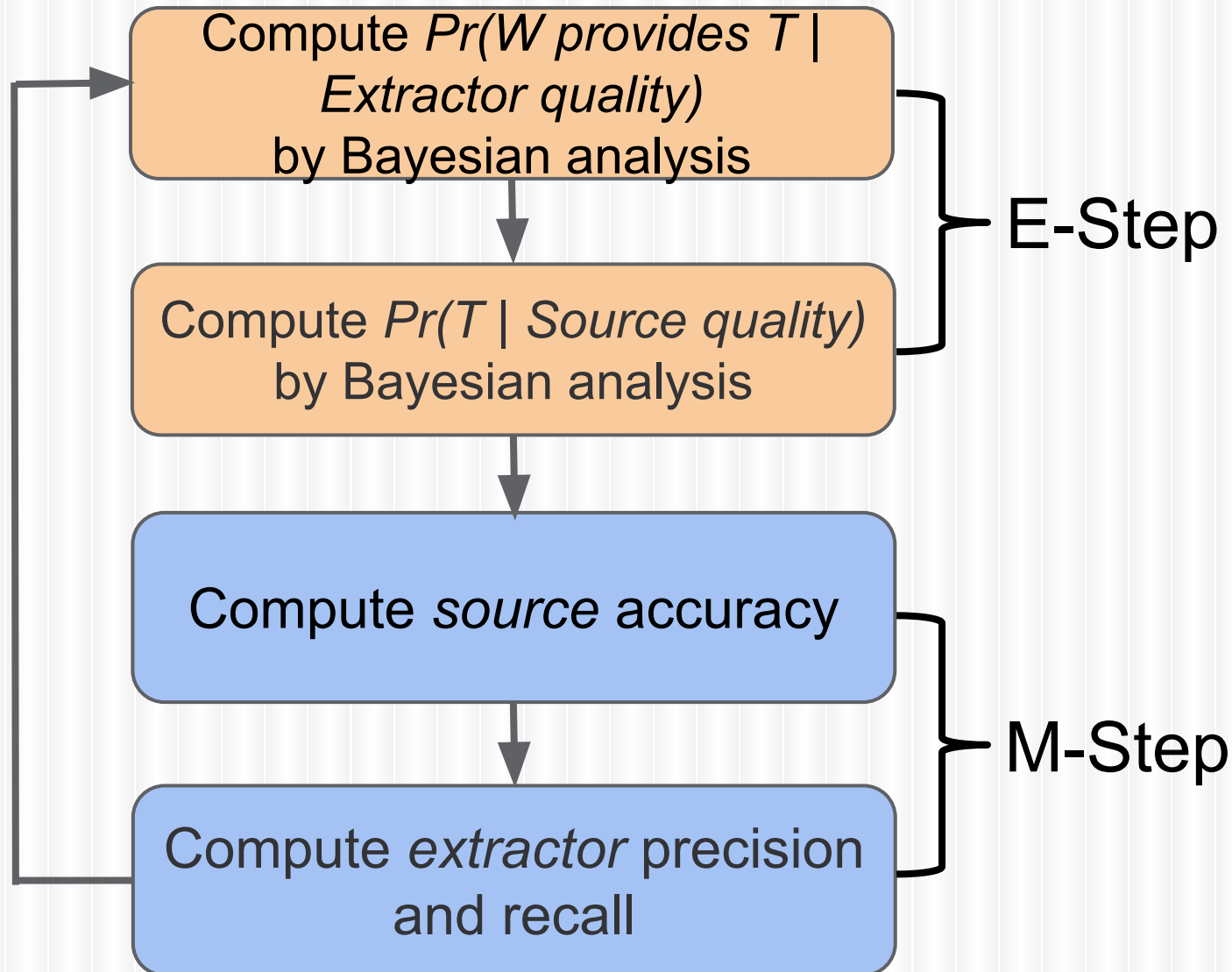
## Latent variables

- $C_{wdv}$ : whether source  $w$  indeed provides  $(d,v)$  pair
- $V_d$ : the correct value(s) for  $d$

## Parameters

- $A_w$ : **Trust** of source  $w$
- $P_e$ : Precision of extractor  $e$
- $R_e$ : Recall of extractor  $e$

# Algorithm [VLDB, 2015]



# Web Source Trustworthiness



Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikimedia Shop

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
Print/export

Languages  
Acèh  
Адыгэбзэ  
Afrikaans  
Alemannisch  
አማርኛ  
Ænglisc  
Аңгсәә  
العربية  
Aragonés  
ᠠᠷᠠᠩᠭᠠᠨ

Create account Log in

Article Talk Read View source Search

## United States

From Wikipedia, the free encyclopedia  
(Redirected from [USA](#))

For other uses, see [US \(disambiguation\)](#), [USA \(disambiguation\)](#), and [United States \(disambiguation\)](#).

The **United States of America** (**USA**), commonly referred to as the **United States** (**US**), **America** or simply **the States**, is a federal republic<sup>[10][11]</sup> consisting of 50 [states](#) and a [federal district](#). The 48 [contiguous states](#) and the federal district of [Washington, D.C.](#), are in central [North America](#) between [Canada](#) and [Mexico](#). The state of [Alaska](#) is the northwestern part of North America and the state of [Hawaii](#) is an [archipelago](#) in the mid-[Pacific](#). The country also has five populated and nine unpopulated [territories](#) in the Pacific and the [Caribbean](#). At 3.79 million square miles (9.83 million km<sup>2</sup>) in total and with around 317

### United States of America



Flag

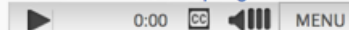


Great Seal

#### Motto:

"In God we trust" (official)<sup>[1][2][3]</sup>  
"E pluribus unum" (Latin) (traditional)  
"Out of many, one"

#### Anthem: "The Star-Spangled Banner"



	Triple Corr	Extraction Corr
Triple 1	1.0	1.0
Triple 2	0.9	1.0
Triple 3	0.3	1.0
Triple 4	0.8	1.0
Triple 5	0.4	0.9
Triple 6	0.8	0.9
Triple 7	0.9	0.8
Triple 8	1.0	0.2
Triple 9	0.7	0.1
Triple 10	0.2	0.1
...	...	...
<b>Accu</b>	<b>0.73</b>	

# Predicting Extraction and Triple Correctness

- (Obama, nationality, Kenya)

2087 extractions:

- Example of a correct extraction ( $Pr\_extCorr=0.792$ )

<http://beforeitsnews.com/obama-birthplace-controversy/2013/04/alabama-supreme-court-chief-justice-roy-moore-to-preside-over-obama-eligibility-case-2458624.html>

2006: Obama In Kenya: I Am So Proud To Come Back Home - [VIDEO HERE](#).

2007: Michelle Obama Declares Obama Is Kenyan And America Is Mean - [VIDEO HERE](#).

2008: Michelle Obama Declares Barack Obama's Home Country Is Kenya - [VIDEO HERE](#).

FLASHBACK: Obama Is The Original Birther! Obama In 1991 Stated In His Own Bio He Was Born In Kenya. [DETAILS HERE](#).

- Example of a wrong extraction ( $Pr\_extCorr=0.130$ )

<http://www.monitor.co.ug/News/National/US+will+respect+winner+of+Kenya+election++Obama+says/-/688334/1685814/-/ksxagx/-/index.html>

US will respect winner of  
Kenya election, Obama says

SHARE BOOKMARK PRINT RATING☆☆☆☆

- $Pr\_tripleCorr=0$  (not enough support)

# Predicting Extraction and Triple Correctness

- (Obama, nationality, USA)  
2481 extractions:
  - Example of a correct extraction ( $Pr\_extCorr=0.999$ )  
<http://www.dogonews.com/2009/10/9/a-nobel-prize-for-our-awesome-president>
  - Example of a wrong extraction ( $Pr\_extCorr=0.261$ )  
<http://blogs.telegraph.co.uk/news/timstanley/100169248/barack-obamas-life-story-contains-myth-not-truth-says-biographer-so-why-did-the-media-report-it-as-truth/>

## Tim Stanley

Dr Tim Stanley is a historian of the United States. His new book about Hollywood politics is out in May. His personal website is [www.timothystanley.co.uk](http://www.timothystanley.co.uk) and you can follow him on Twitter @timothy\_stanley.

 Follow 15.6K followers

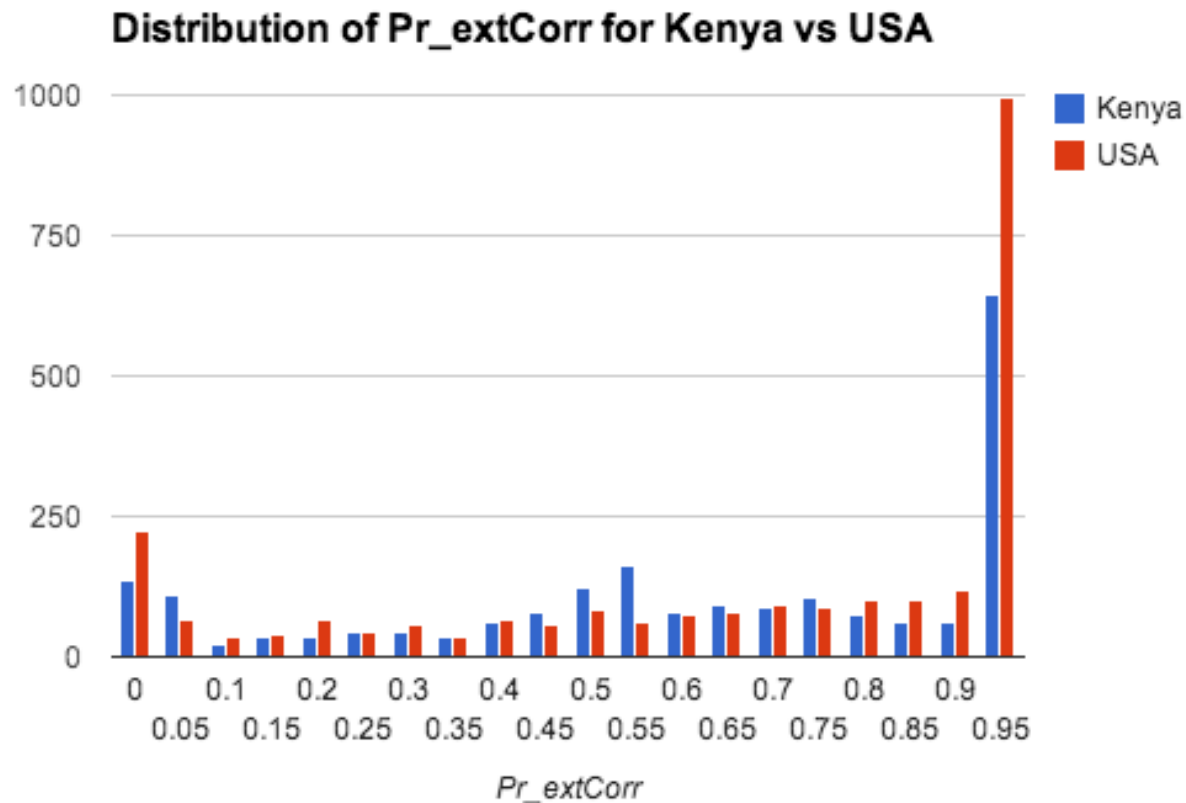


Barack Obama's life story contains 'myth, not truth', says biographer. So why did the media report it as truth?

- $Pr\_tripleCorr=1$  (higher support)

# Predicting Extraction and Triple Correctness

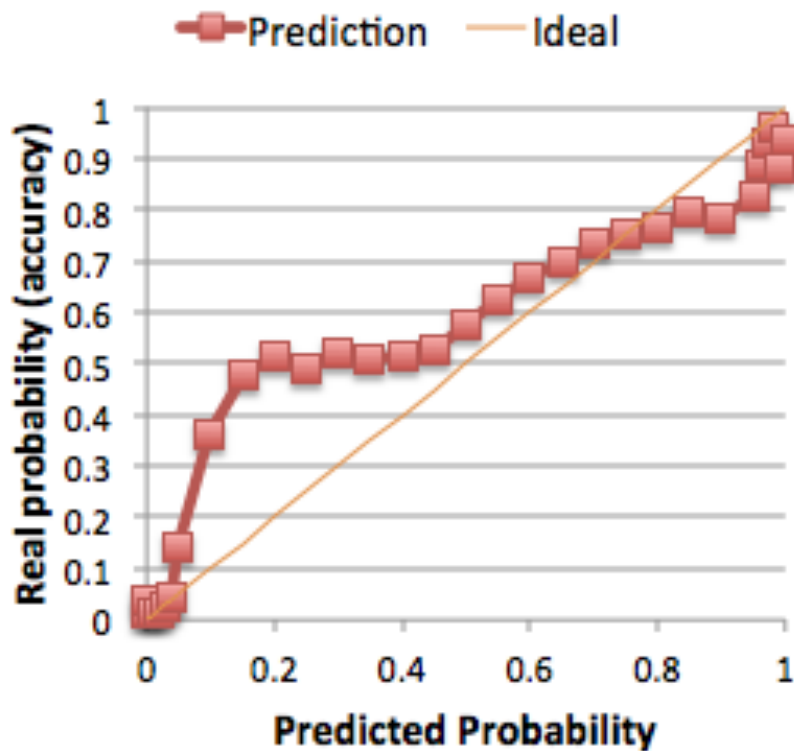
## Distribution of providers for Kenya and USA



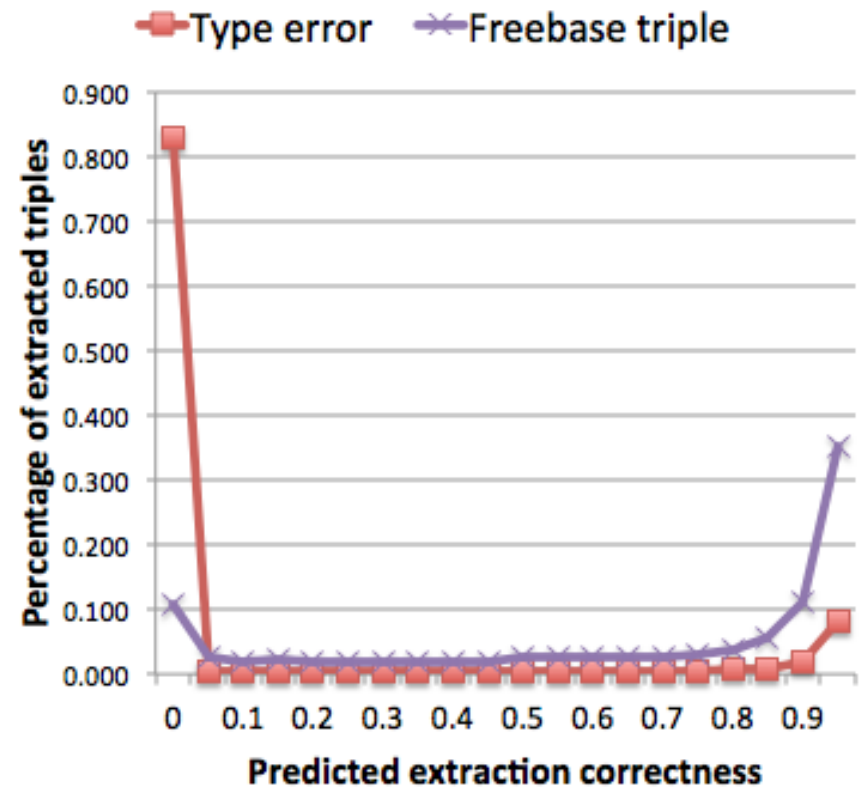
# Predicting Extraction and Triple Correctness

.....

## Triple Correctness Prediction



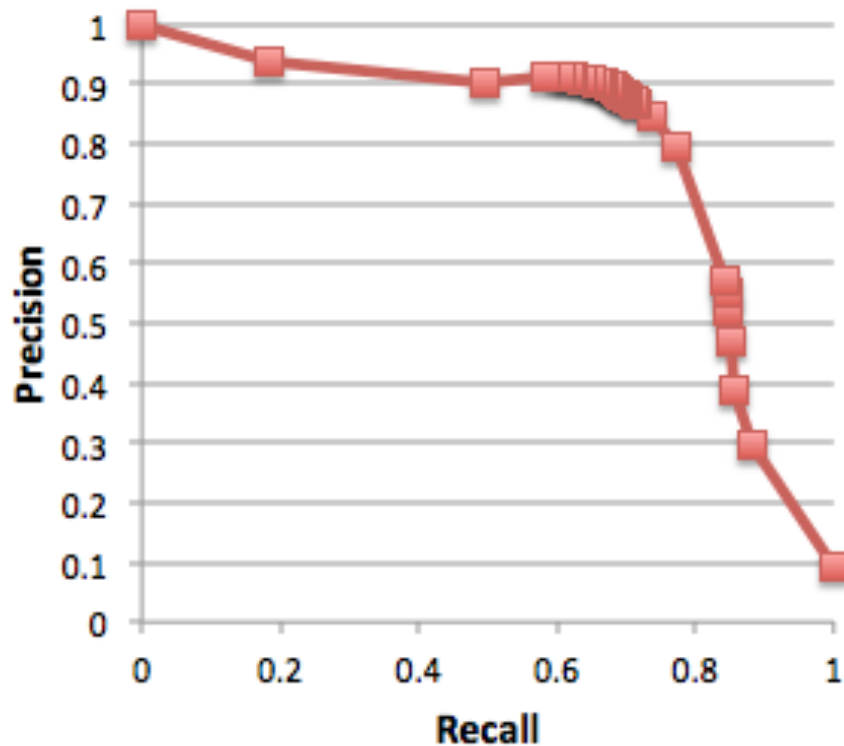
## Extraction Correctness Prediction



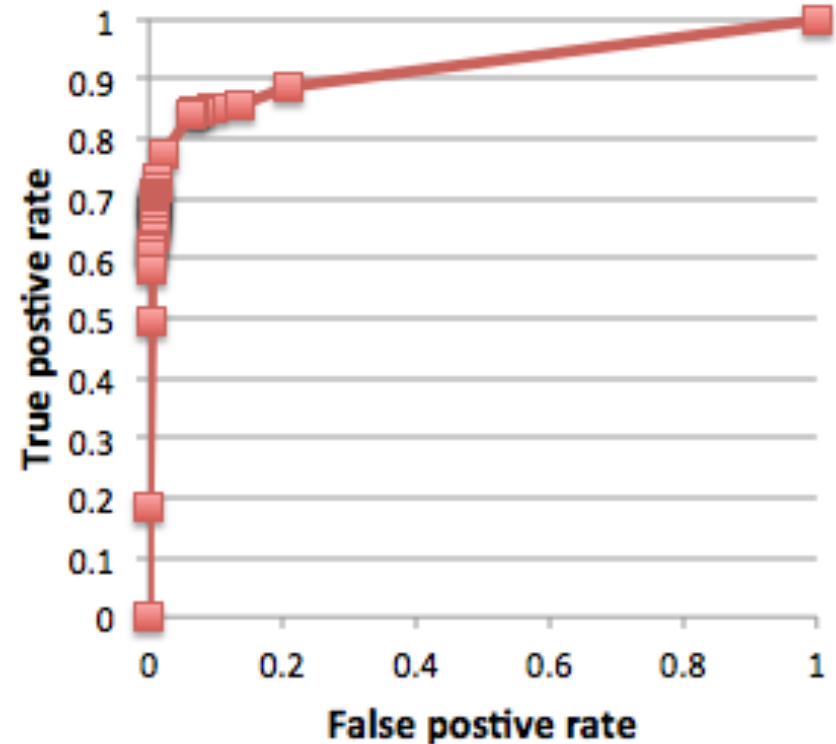
# Predicting Triple Correctness

.....

**PR-Curves**



**ROC-Curves**



But, How Ready Are We  
in Using KBT?

# Input Readiness

---

1. Extraction is still very sparse
  - a. 74% URLs each contributes fewer than 5 triples
  - b. We compute reliable KBT for <20% websites and <<5% webpages
2. Extraction is of low quality
  - a. Overall accuracy is as low as 11.5%
  - b. Low accuracy for some good sources because of undetected extraction errors
3. Inbalance between texts and semi-structured data
  - a. Among 100 sample sources w. KBT over 0.9, 95% are mainly semi-structured data

# Technique Readiness

.....

1. Single-truth assumption
  - a. Pro: filters large amount of noise
  - b. Con: often does not hold in practice
2. Value similarity and hierarchy  
e.g., San Francisco vs. Oakland
3. Copy detection: Existing techniques not applicable because of web-scale
4. Triple filtering
  - a. Triples irrelevant to source topic
  - b. “Trivial” triples, well-known facts
5. Simple vs. sophisticated facts
6. Identify different perspectives

# Customer Readiness

.....

1. Quality prediction for websites/pages  
Combination w. other signals, e.g., PageRank
2. Source recommendation
  - a. Head:  
Well-known  
authoritatives
  - b. Long tails:  
Limited support

Call to arms --  
Leave NO Valuable Data  
Behind



# THANK YOU!

**NOT** [www.knowledgebasedtrust.org](http://www.knowledgebasedtrust.org)  
[www.knowledgebasedtrust.com](http://www.knowledgebasedtrust.com)

