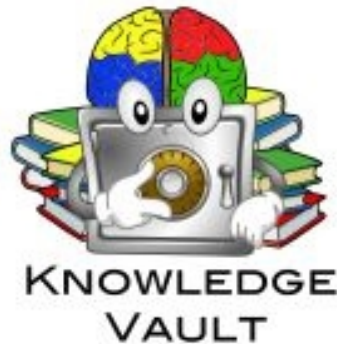


From Data Fusion to Knowledge Fusion

.....

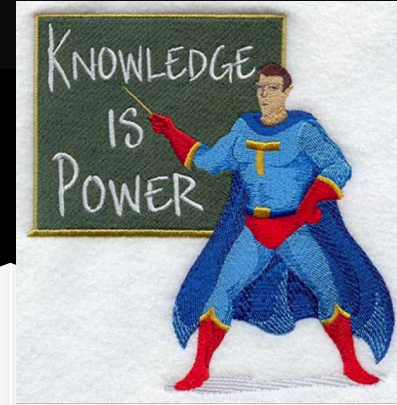
*Xin Luna Dong, Google Inc.
9/6/2014 @ APWeb '14*



SONYA: A Big Project, A Fancy Machine, And A Cute Little Girl



Knowledge Is Power



- Many Knowledge Bases (KB)



NELL: Never-Ending Language Learning



Google Knowledge Graph

.....



The most important Google story this year was the launch of the **Knowledge Graph**. This marked the shift from a first-generation Google that merely indexed the words and metadata of the Web to a next-generation Google that recognizes discrete things and the relationships between them.

- ReadWrite 12/27/2012

Using KG in Search



岳麓



Web

News

Maps

Shopping

Images

More ▾

Search tools

About 2,960,000 results (0.27 seconds)

岳麓书院: 首页

yjsy.hnu.cn/ ▾ 轉為繁體網頁 Yuelu Academy ▾

在古木参天、浓荫蔽日、山光水色的湘江西岸，有一片典雅、庄重的古建筑群，这就是为世人所瞩目的“四大书院”之一的岳麓书院。

4.5 ★★★★★ 22 Google reviews · [Write a review](#)

China, Hunan, Changsha, Yuelu, 麓山路
+86 731 8882 3764

岳麓书院- 维基百科，自由的百科全书

zh.wikipedia.org/zh-hant/岳麓书院 ▾ 轉為繁體網頁 Chinese Wikipedia ▾

岳麓书院位于中国湖南省长沙市岳麓山东麓，是中国古代四大书院之一，始建于北宋开宝九年（976年），历经宋、元、明、清各个朝代，迨及晚清（1903年）改为湖南高等...
[历史沿革](#) - [大事年表](#) - [书院学规](#) - [历代山长](#)

News for 岳麓

“小社区大民生”领创和谐岳麓

光明网 - 5 hours ago

在坐拥15个街道2个镇、81个社区的岳麓区，其中有1个街道、2个社区荣获全国和谐社区建设示范单位，有3个街道、24个社区分别荣获省、市精品、...

做客乘坐缆车高空扔垃圾长沙岳麓山很“受伤”

红网 - 21 hours ago

More news for 岳麓

岳麓书院_百度百科

baike.baidu.com/view/7288.htm ▾ 轉為繁體網頁 Baidu Baike ▾

北宋开宝九年（976），潭州太守朱洞在僧人办学的基础上，正式创立岳麓书院。嗣后，历经宋、元、明、清各代，至清末光绪二十九年（1903）改为湖南高等学堂，尔后相继...



See photos

岳麓書院

[Directions](#)

[Write a review](#)

Address: China, Hunan, Changsha, Yuelu, 麓山路

Phone: +86 731 8882 3764

Reviews

4.5 ★★★★★ 22 Google reviews

Are you the business owner?

[Feedback](#)

See results about

Yuelu District

Yuelu District is one of six urban districts of Changsha, the capital of Hunan province, China. It ...

Expanding KG— Extracting Knowledge from the Web

“In the near future, the web is going to be the master copy of human knowledge. We need to figure out ways to use that knowledge.”

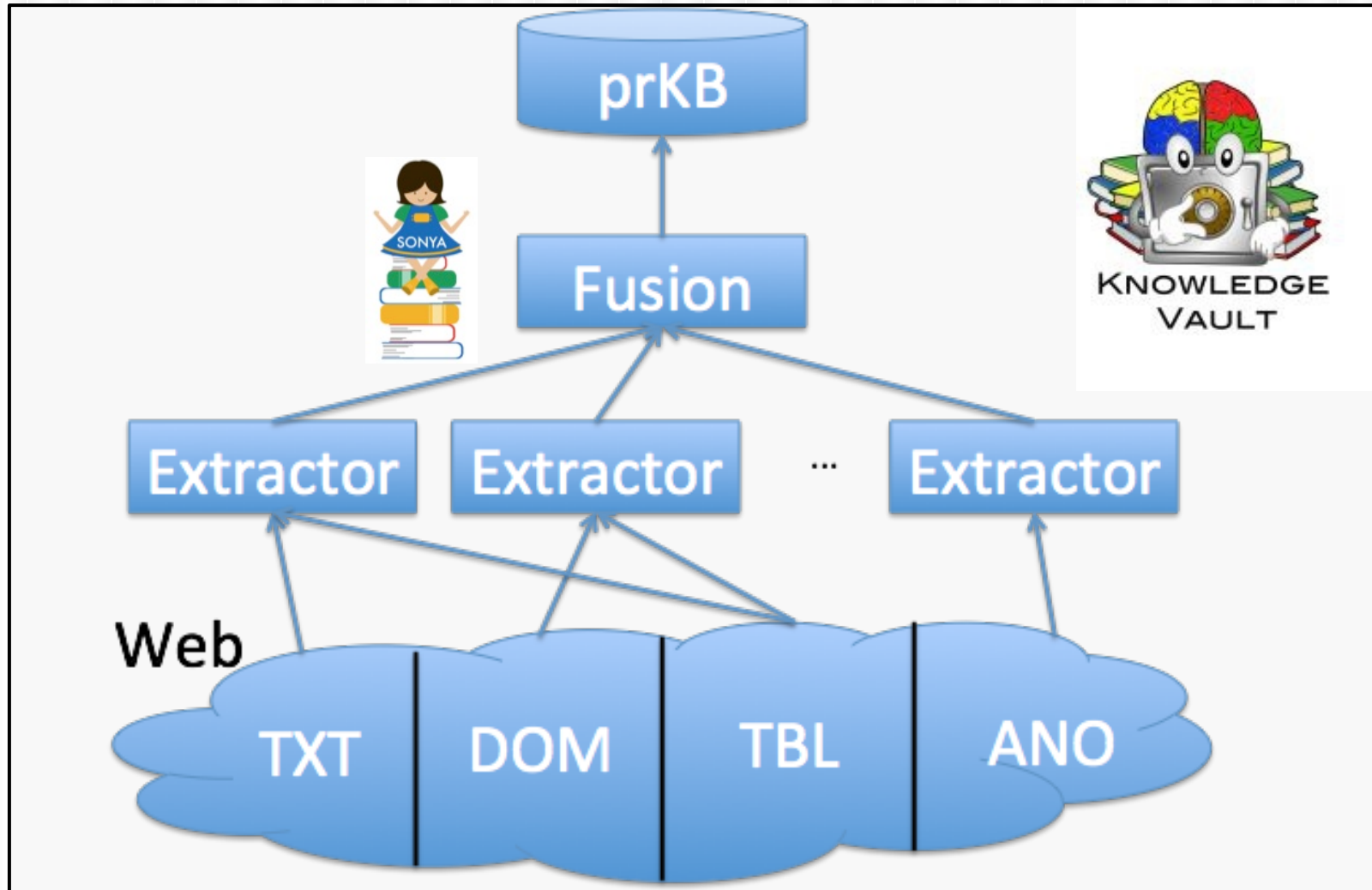
—Håkon Wium Lie

But—

- KG requires 99% accuracy for knowledge
- Web data is noisy and extraction is hard
- How to balance *coverage* and *accuracy*?

Knowledge Vault– Building a Probabilistic KB

[VLDB'2014, Sigmod'2014, KDD'2014]



News

Google's Knowledge Vault already contains 1.6 billion facts

FELICITY NELSON

SATURDAY, 23 AUGUST 2014



275



64



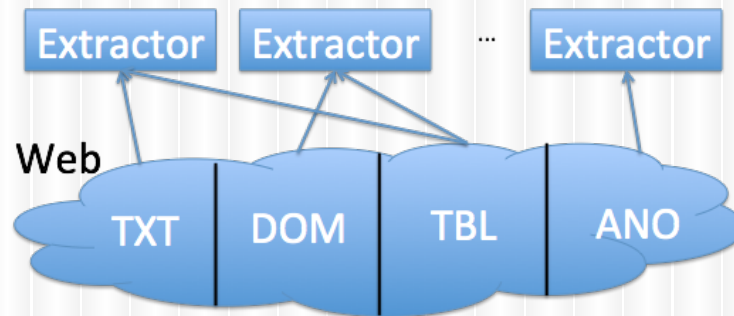
The automated, fact-harvesting bot will build up a collection of all human knowledge.

Outline

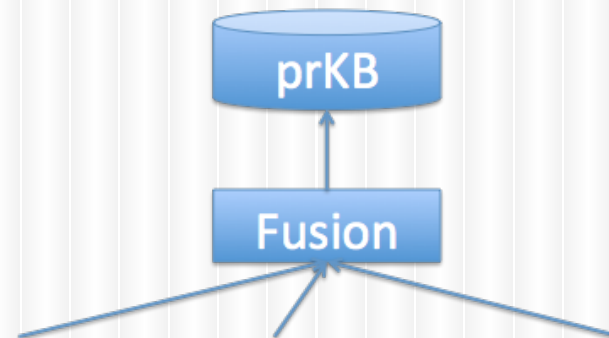
.....



Knowledge extraction



Knowledge fusion



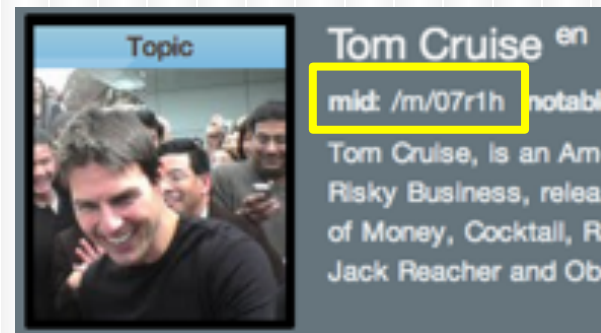
Interesting applications



Future directions

Knowledge Extraction I—Knowledge

- Triple: (subject, predicate, object)
e.g., (Tom Cruise, date_of_birth, 7/3/1962)
 - Subject—a Freebase mid
e.g., /m/07r1h
 - Predicate—predefined in Freebase; e.g., people/person/date_of_birth
 - Object—a Freebase mid, a string, a number, or a date.



Statistics for Extracted Triples

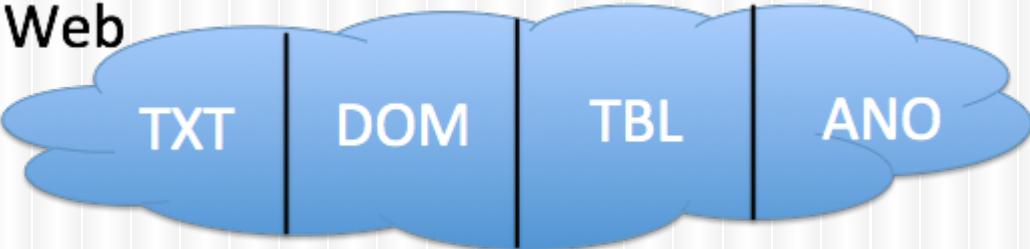
.....

- A large knowledge base

As of 11/2013

#Triples	1.6B (now 2.8B)
#Subjects (Entities)	43M
#Types	1.1K
#Predicates	4.5K
#Objects	102M

- Highly skewed data—fat heads, long tail
 - #Triples/type: 1–14M
(*location, organization, business*)
 - #Triples/entity: 1–2M (*USA, UK, CA, NYC, TX*)



Free texts

Synopsis

 [Print](#) [Cite This](#)

Born on April 15, 1452, in Vinci, Italy, Leonardo da Vinci was concerned with the laws of science and nature, which greatly informed his work as a painter, sculptor and architect. His ideas and body of work -- which included *The Last Supper*, *Leda and the Swan* and the Vitruvian Man -- influenced countless artists and made da Vinci a central figure of the Italian Renaissance.



Search for (keyword)

[Welcome](#) [About Me](#) [Write a Review](#) [Find](#)

Shana Thai Restaurant

 740 reviews [Rating Details](#)

Web tables & Lists

	Name and (party) ¹	Term	State of birth	Born	Died
1.	Washington (F) ³	1789–1797	Va.	2/22/1732	12/14/1799
2.	J. Adams (F)	1797–1801	Mass.	10/30/1735	7/4/1826
3.	Jefferson (DR)	1801–1809	Va.	4/13/1743	7/4/1826
4.	Madison (DR)	1809–1817	Va.	3/16/1751	6/27/1836

Annotations

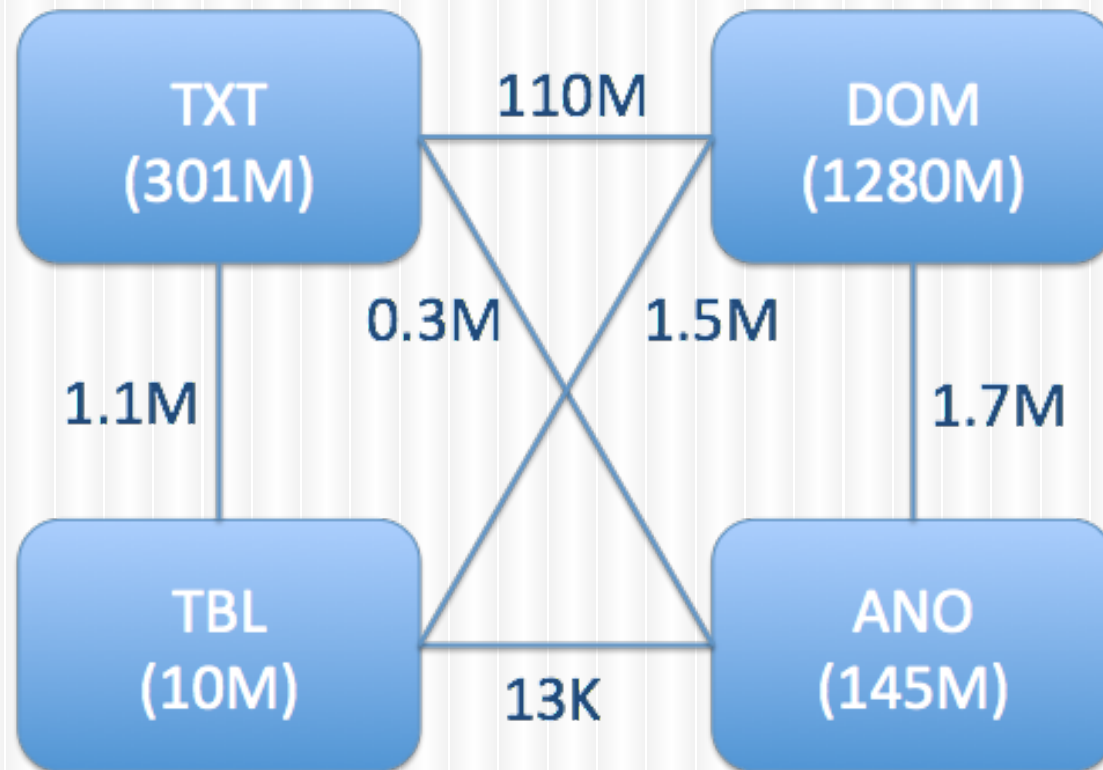
```
<h1 itemprop="name">
Tom Cruise </h1>
<span itemprop="birthDate">
7/3/1962 </span>
<span itemprop="gender">
Male </span>
```

schema.org

schema.org

Statistics for Web Sources

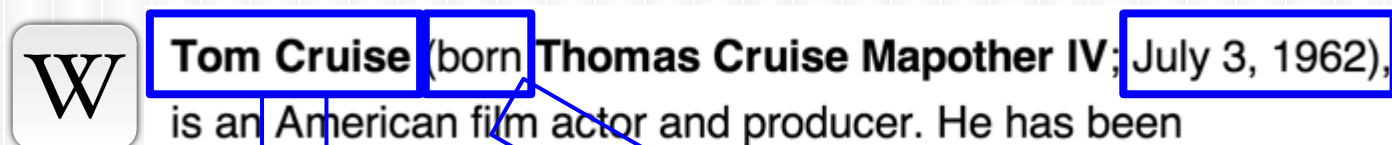
- 1B+ Webpages over the Web
- Contribution is skewed: 1- 50K



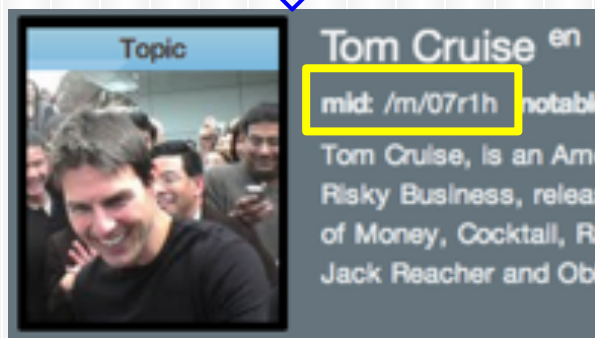
Knowledge Extraction II–Extractors

- Three tasks (any order, maybe combined)

I. Triple identification



II. Entity linkage





/people/person/date_
of_birth


III. Predicate linkage

Knowledge Extraction II–Extractors

- Texts/DOM: distant supervision

 **Tom Cruise** (born **Thomas Cruise Mapother IV**; **July 3, 1962**), is an American film actor and producer. He has been



Topic
 Tom Cruise en mid: /m/07r1h notable Tom Cruise, is an Ame Risky Business, relea of Money, Cocktail, Ra Jack Reacher and Oth
Date of birth /people/person/date_of_birth 7/3/1962

Pattern 1: X “born” Y
→ (X, /people/person
/date_of_birth, Y)

- Web tables/lists: schema mapping
- Annotations: semi-automatic mapping

Statistics for Extractors

- 12 extractors; high variety

	#Triples	#Webpages	#Patterns	Accu	Accu (conf $\geq .7$)
TXT1	274M	202M	4.8M	0.36	0.52
TXT2	31M	46M	3.7M	0.18	0.80
TXT3	8.8M	16M	1.5M	0.25	0.81
TXT4	2.9M	1.2M	0.1M	0.78	0.91
DOM1	804M	344M	25.7M	0.43	0.63
DOM2	431M	925M	No pat.	0.09	0.62
DOM3	45M	N/A	No pat.	0.58	0.93
DOM4	52M	7.8M	No pat.	0.26	0.34
DOM5	0.7M	0.5M	No pat.	0.13	No conf.
TBL1	3.1M	0.4M	No pat.	0.24	0.24
TBL2	7.4M	0.1M	No pat.	0.69	No conf.
ANO	145M	53M	No pat.	0.28	0.30

Errors Can Creep in at Every Stage

.....

Extraction error: (Obama, nationality, Chicago)



Errors Can Creep in at Every Stage

.....

Reconciliation error:
(Obama, nationality, North America)

American
President
Barack Obama



Errors Can Creep in at Every Stage

Source data error: (Obama, nationality, Kenya)

Obama born
in Kenya



Knowledge Extraction IV–Quality

- Gold standard: Freebase
- LCWA (Local Closed-World Assumption)
 - If (s,p,o) exists in FB: true
 - Otherwise,
 - If (s,p) exists in FB: false (Freebase knowledge is locally complete)
 - Otherwise: UNKNOWN
- The gold standard contains about 40% of the triples

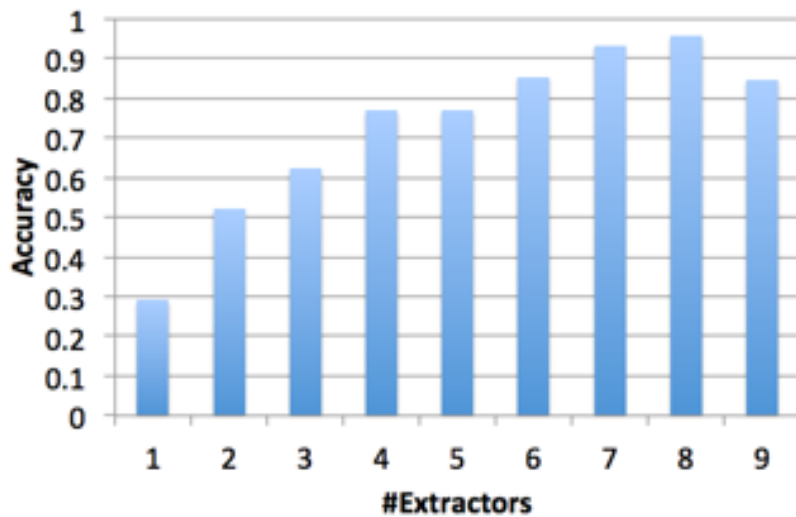
Statistics for Triple Correctness

- Overall accuracy: 30%
- Random sample on 25 false triples
 - Triple-identification errors: 11 (44%)
 - Entity-linkage errors: 11 (44%)
 - Predicate-linkage errors: 5 (20%)
 - Source-data errors: 1 (4%)

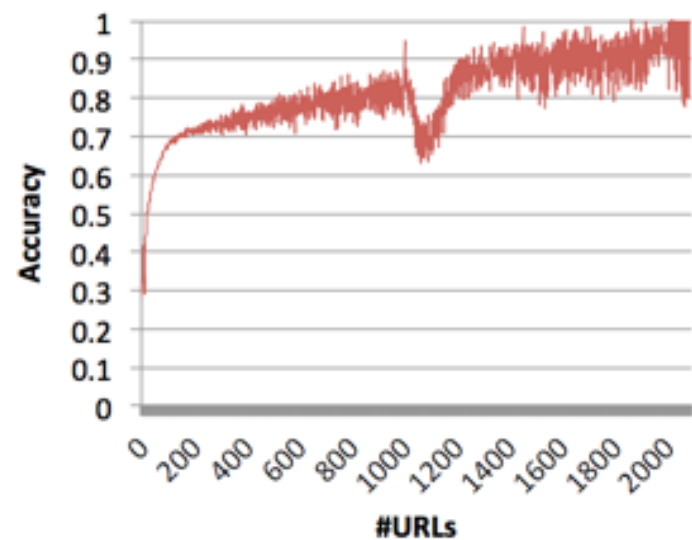
Statistics for Triple Correctness

.....

Triple Accuracy by #Extractors



Triple Accuracy by #URLs



Statistics for Extractors

- 12 extractors; high variety

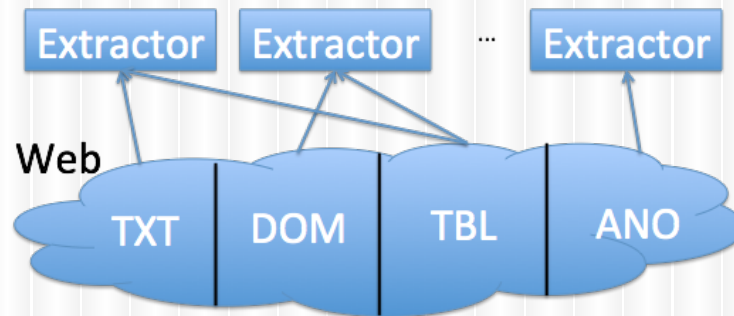
	#Triples	#Webpages	#Patterns	Accu	Accu (conf $\geq .7$)
TXT1	274M	202M	4.8M	0.36	0.52
TXT2	31M	46M	3.7M	0.18	0.80
TXT3	8.8M	16M	1.5M	0.25	0.81
TXT4	2.9M	1.2M	0.1M	0.78	0.91
DOM1	804M	344M	25.7M	0.43	0.63
DOM2	431M	925M	No pat.	0.09	0.62
DOM3	45M	N/A	No pat.	0.58	0.93
DOM4	52M	7.8M	No pat.	0.26	0.34
DOM5	0.7M	0.5M	No pat.	0.13	No conf.
TBL1	3.1M	0.4M	No pat.	0.24	0.24
TBL2	7.4M	0.1M	No pat.	0.69	No conf.
ANO	145M	53M	No pat.	0.28	0.30

Outline

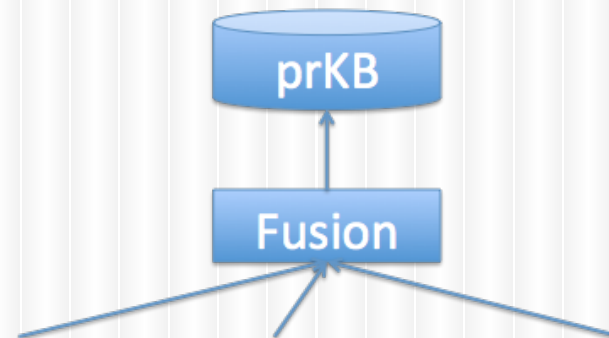
.....



Knowledge extraction



Knowledge fusion



Interesting applications

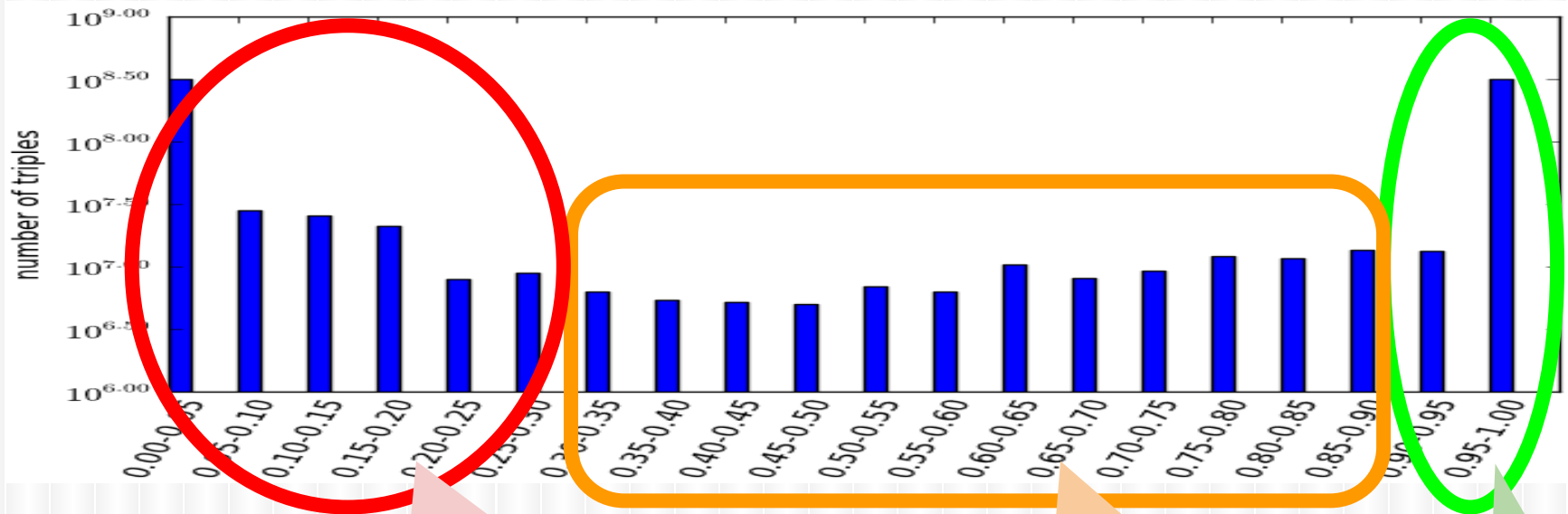


Future directions

Goal: Judge Triple Correctness

- Input: Knowledge triples and their provenances (i.e., which extractor extracts from which source)
- Output: a probability in $[0,1]$ for each triple
 - Probabilistic decisions
vs. deterministic decisions

Usage of Probabilistic Knowledge



Negative training examples, and
MANY EXCITING APPLICATIONS!!

Active learning,
probabilistic inference, etc.

Upload
to KG

Data Fusion–Definition

.....

Input

		Sources			
		S_1	S_2	...	S_N
Data items	D_1				
	D_2				
	D_3				
	...				
	D_M				

Output

		Truths
Data items	D_1	
	D_2	
	D_3	
	...	
	D_M	

Data Fusion–Intuition

.....

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

Data Fusion–Intuition

.....

	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.

Data Fusion–Intuition

.....



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

Data Fusion–Intuition

.....



	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.

Data Fusion—A Bayesian Model

.....[Dong et al., VLDB'09]

Q1. How to compute source accuracy?

- Source Accuracy: $A(S)$

$$A(S) = \text{Avg}_{v \in \bar{V}(S)} P(v)$$

- $\bar{V}(S)$ - values provided by S
- $P(v)$ - pr of value v being true

Data Fusion—A Bayesian Model

[Dong et al., VLDB'09]

Q2. How to leverage accuracy in voting?

Input:

- Data item D
- $\text{Dom}(D)=\{v_0, v_1, \dots, v_n\}$
- Observation Φ on D

Output:

$\Pr(v_i \text{ true} | \Phi)$ for each
 $i=0, \dots, n$ (sum up to 1)

According to the Bayes Rule,
we need to know $\Pr(\Phi | v_i \text{ true})$

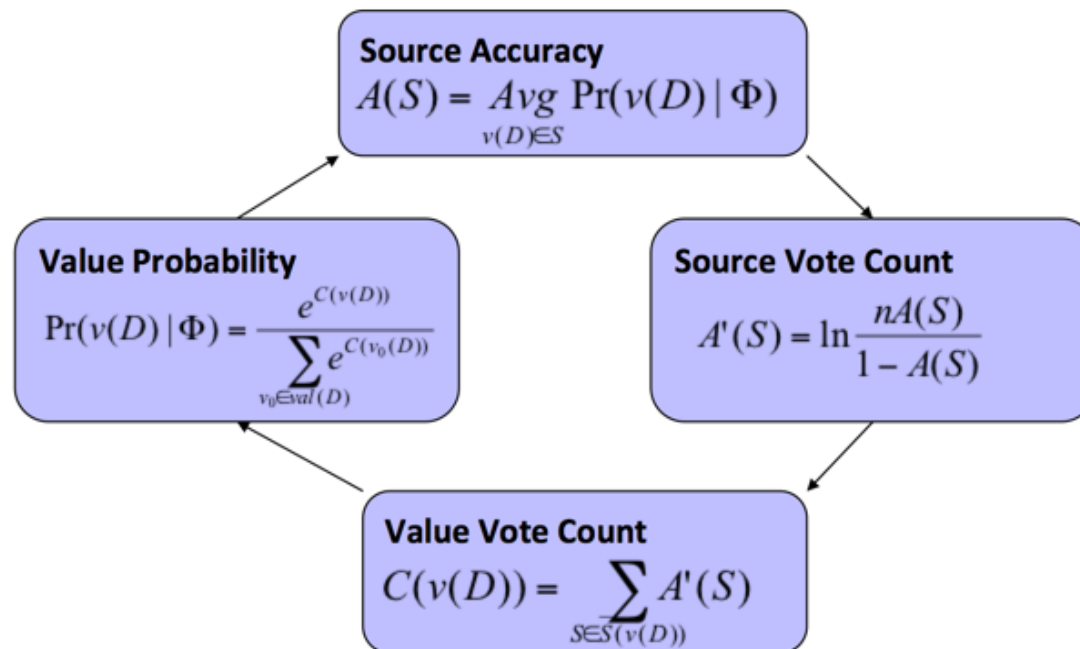
- Assuming independence of sources, we need to know $\Pr(\Phi(S) | v_i \text{ true})$
- If S provides v_i :
 $\Pr(\Phi(S) | v_i \text{ true}) = A(S)$
- If S does not provide v_i :
 $\Pr(\Phi(S) | v_i \text{ true}) = (1 - A(S))/n$

Data Fusion—A Bayesian Model

[Dong et al., VLDB'09]

Q3. How to handle interdependence between source accuracy and value pr?

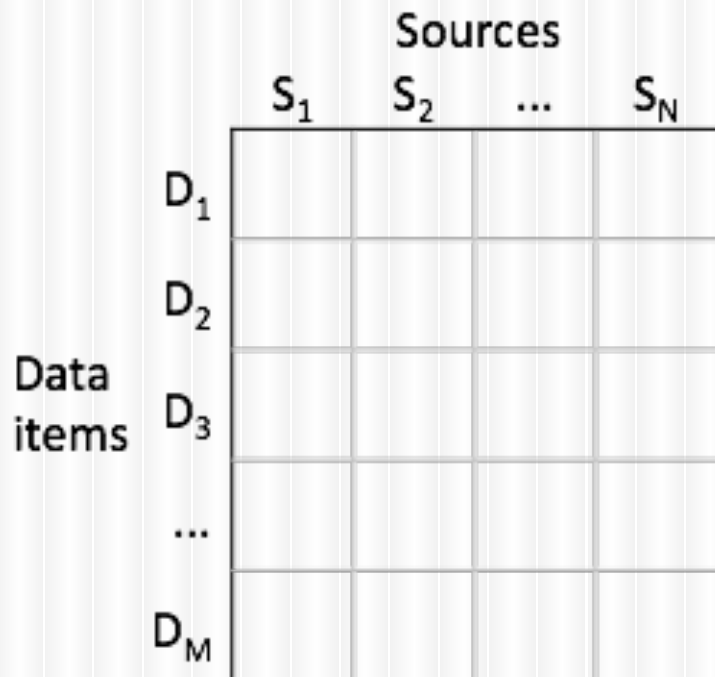
- ◆ Continue until source accuracy converges



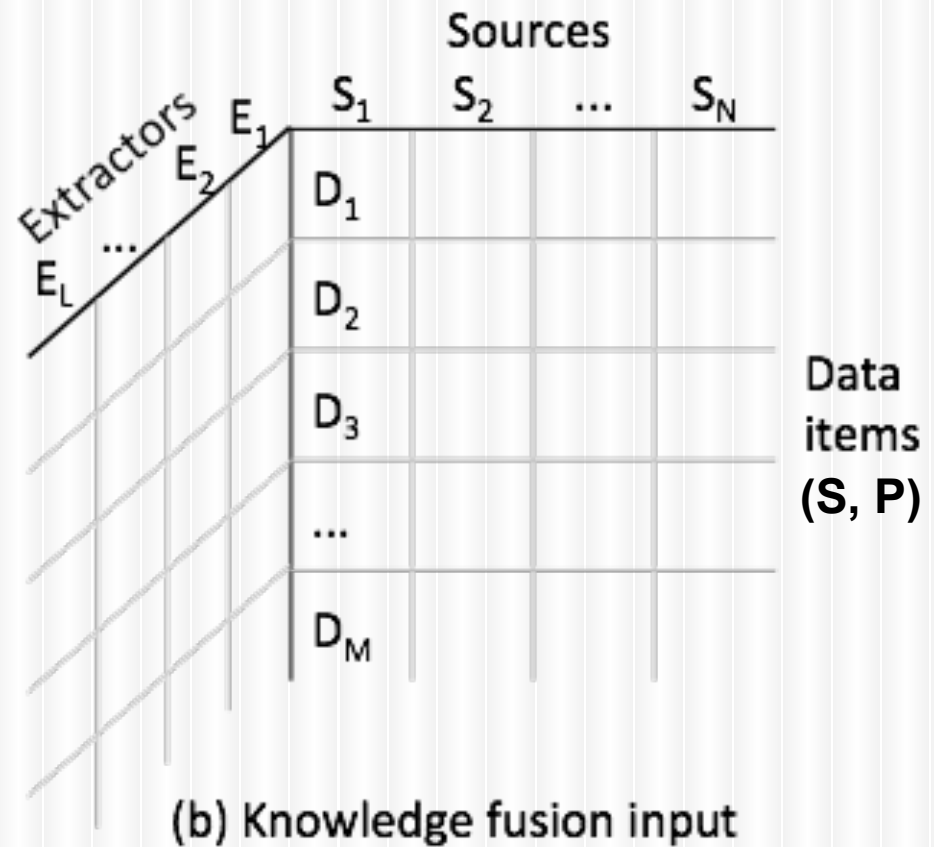
Knowledge Fusion Challenges

.....

I. Input is *three-dimensional*



(a) Data fusion input

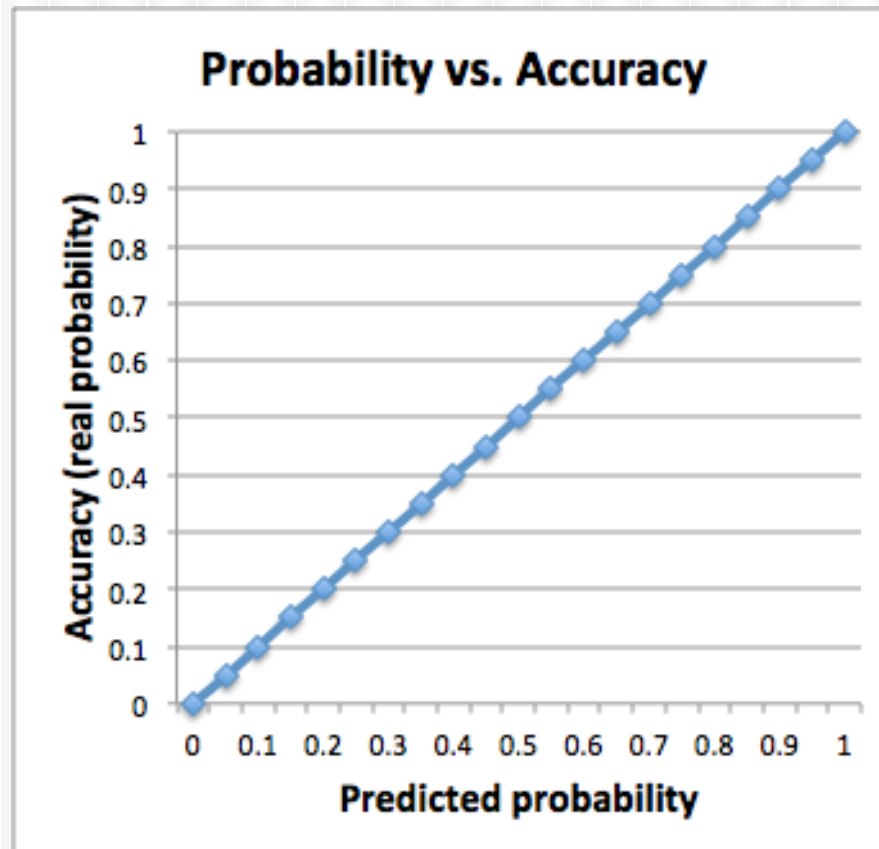


(b) Knowledge fusion input

Knowledge Fusion Challenges

.....

II. Output prs should be *well-calibrated*



Knowledge Fusion Challenges

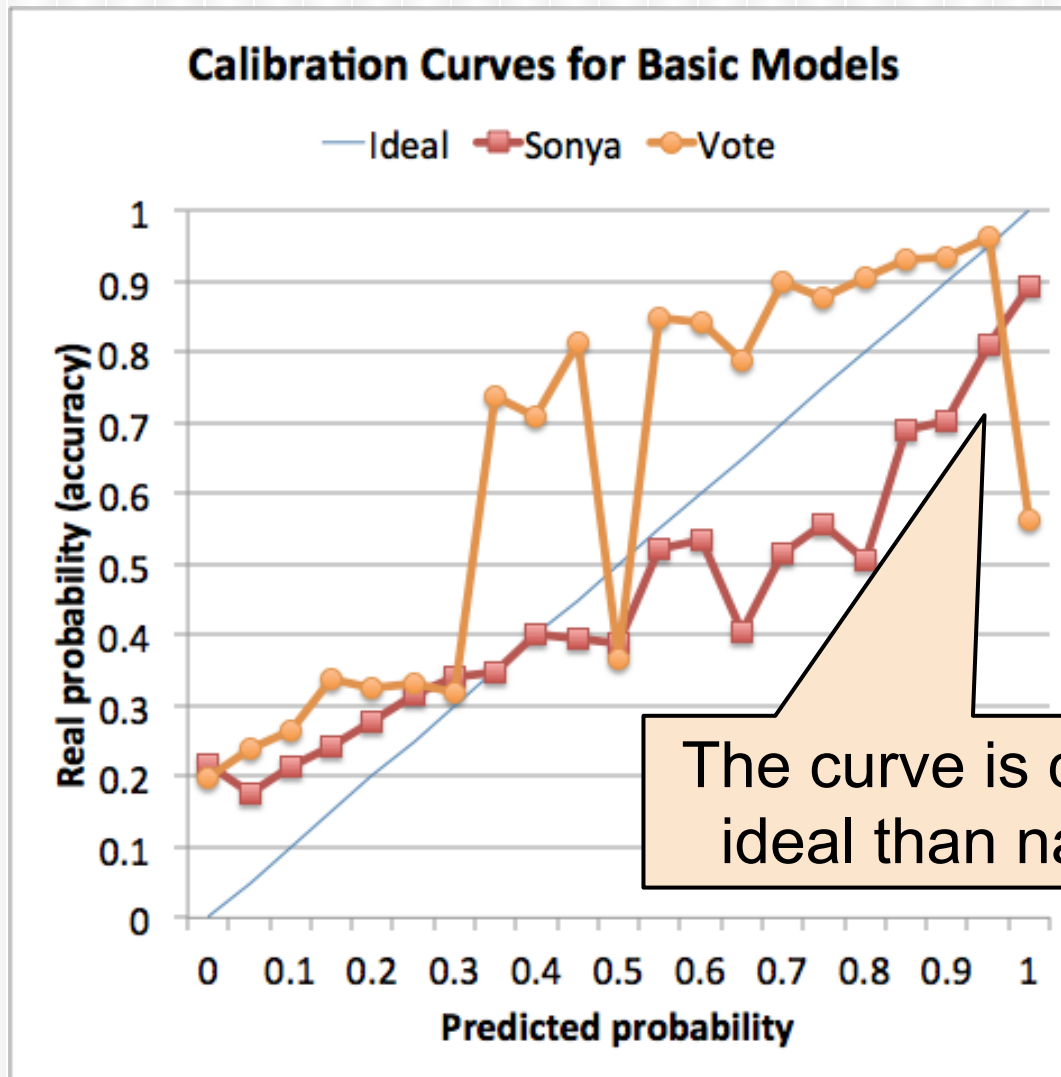
III. Data are of *Web-scale*

- Three orders of magnitude larger than currently published data-fusion applications
 - Size: 1.1TB
 - Sources: 170K→ 1B+
 - Data items: 400K→375M
 - Values: 18M→6.4B (1.6B unique)
- Data are highly skewed
 - #Triples/Data-item: 1 - 2.7M
 - #Triples/Source: 1 - 50K

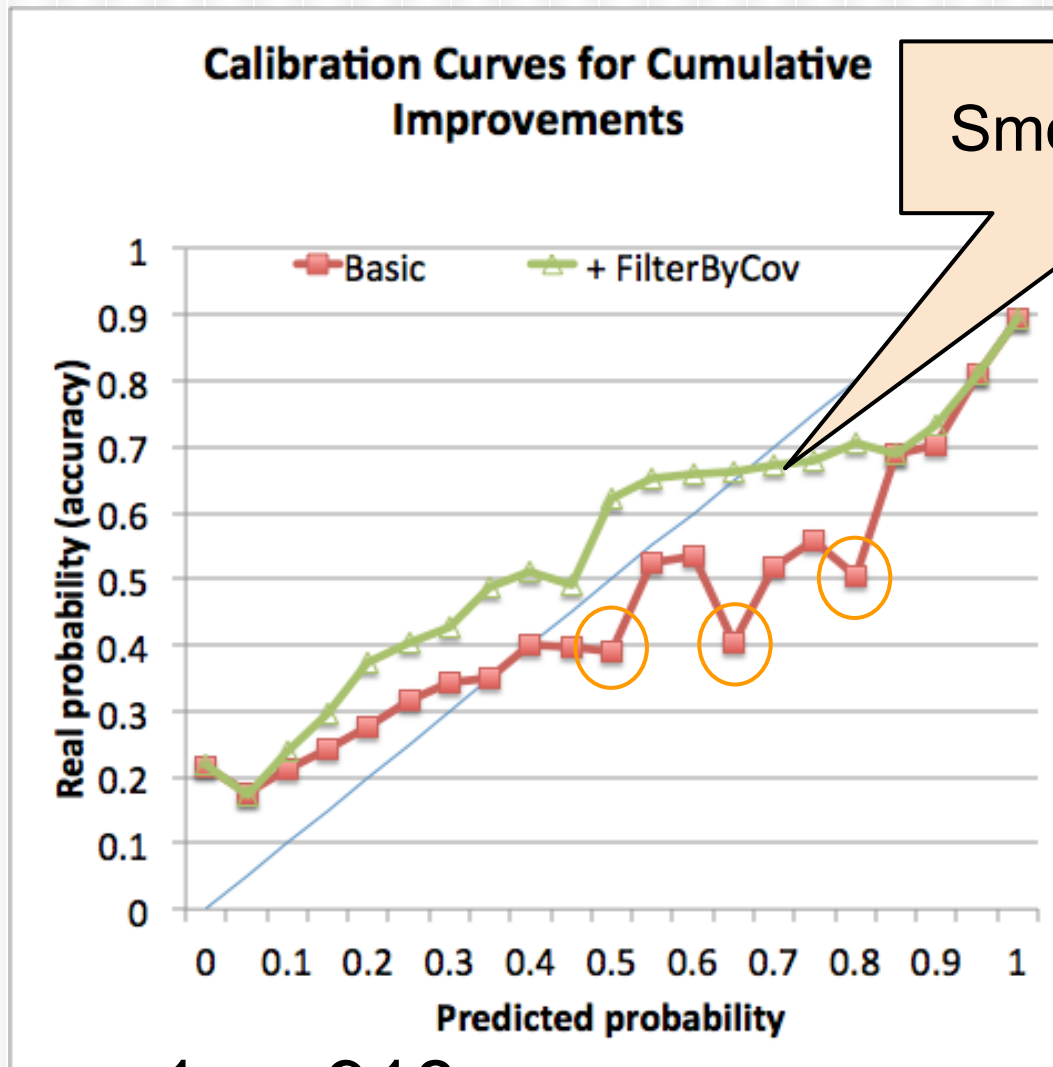
Knowledge Fusion Solutions

- Treat each (URL, Extractor) as a whole (*provenance*) for accuracy evaluation
- A series of refinements to improve probability calibration
- MapReduce Based Framework
 - Terminate in 5 rounds
 - Sample for *too big* data items or provenances

Basic Sonya Solution vs. Voting

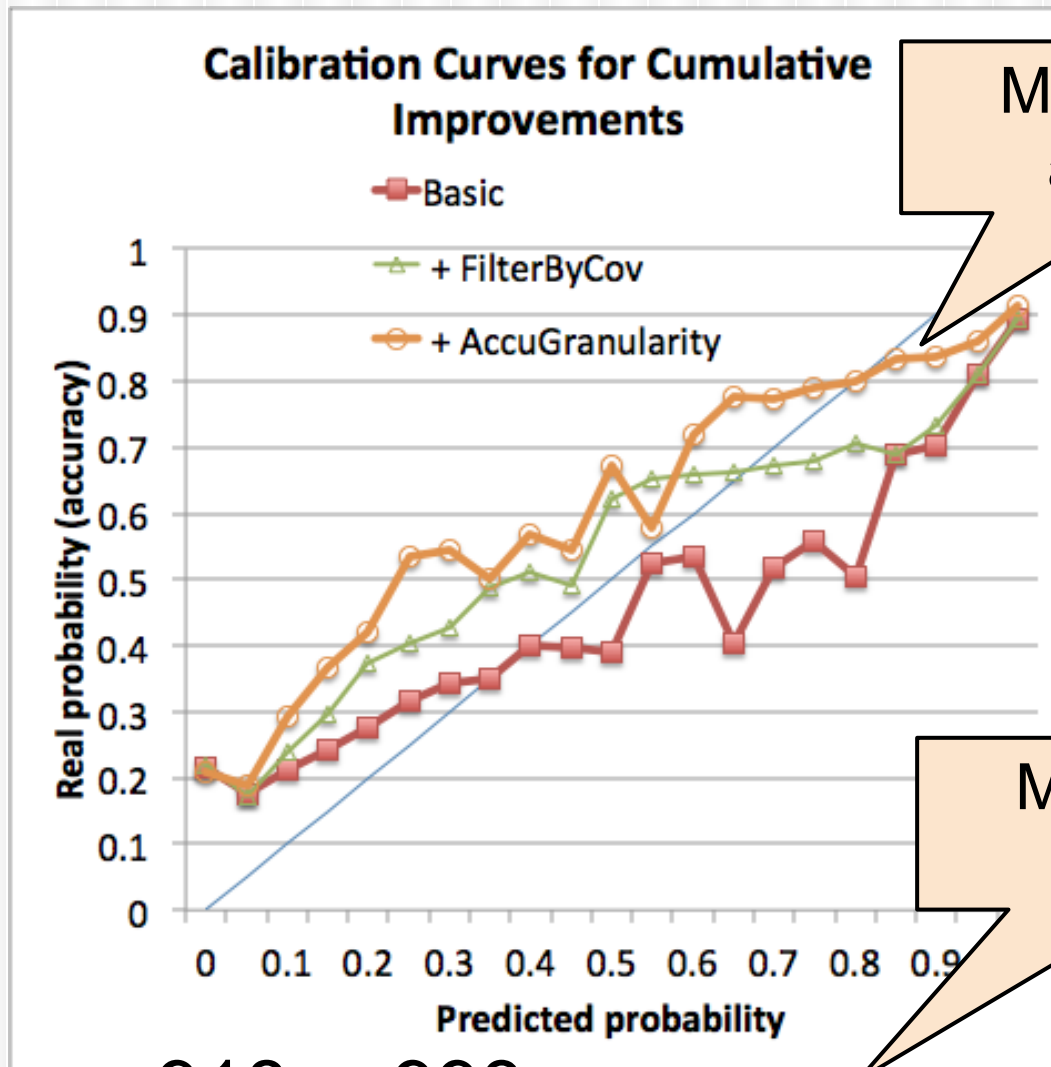


Refinement I. Ignore Low-Coverage Provenances



Coverage: $1 \rightarrow .918$

Refinement II. Granularity (URL->Site, Extractor->Pattern, Predicate)

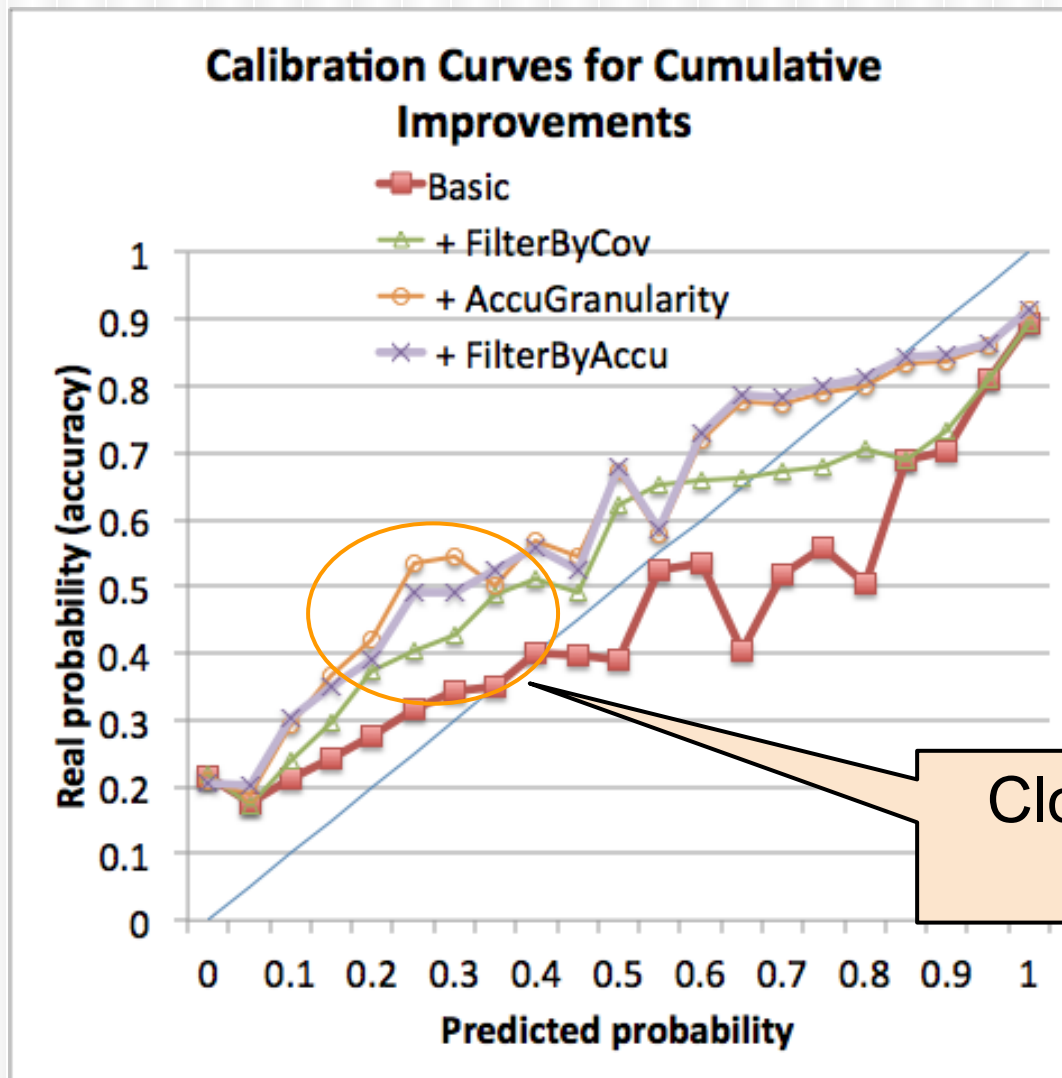


Much higher accuracy

Much higher coverage

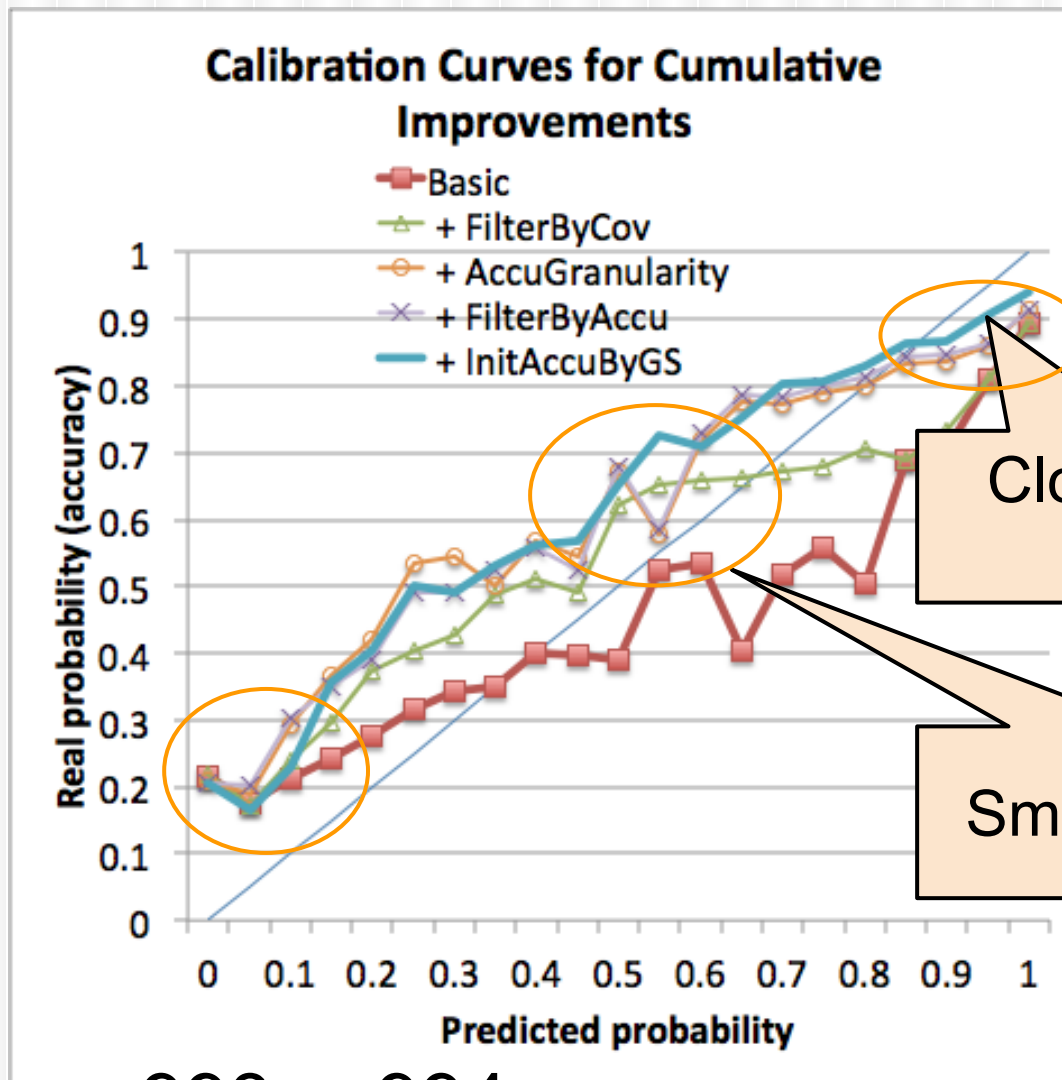
Coverage: .918 → .993

Refinement II. Ignore Low-Accuracy Provenances



Closer to ideal curve

Refinement IV. Initiate Provenance Accuracy by FB

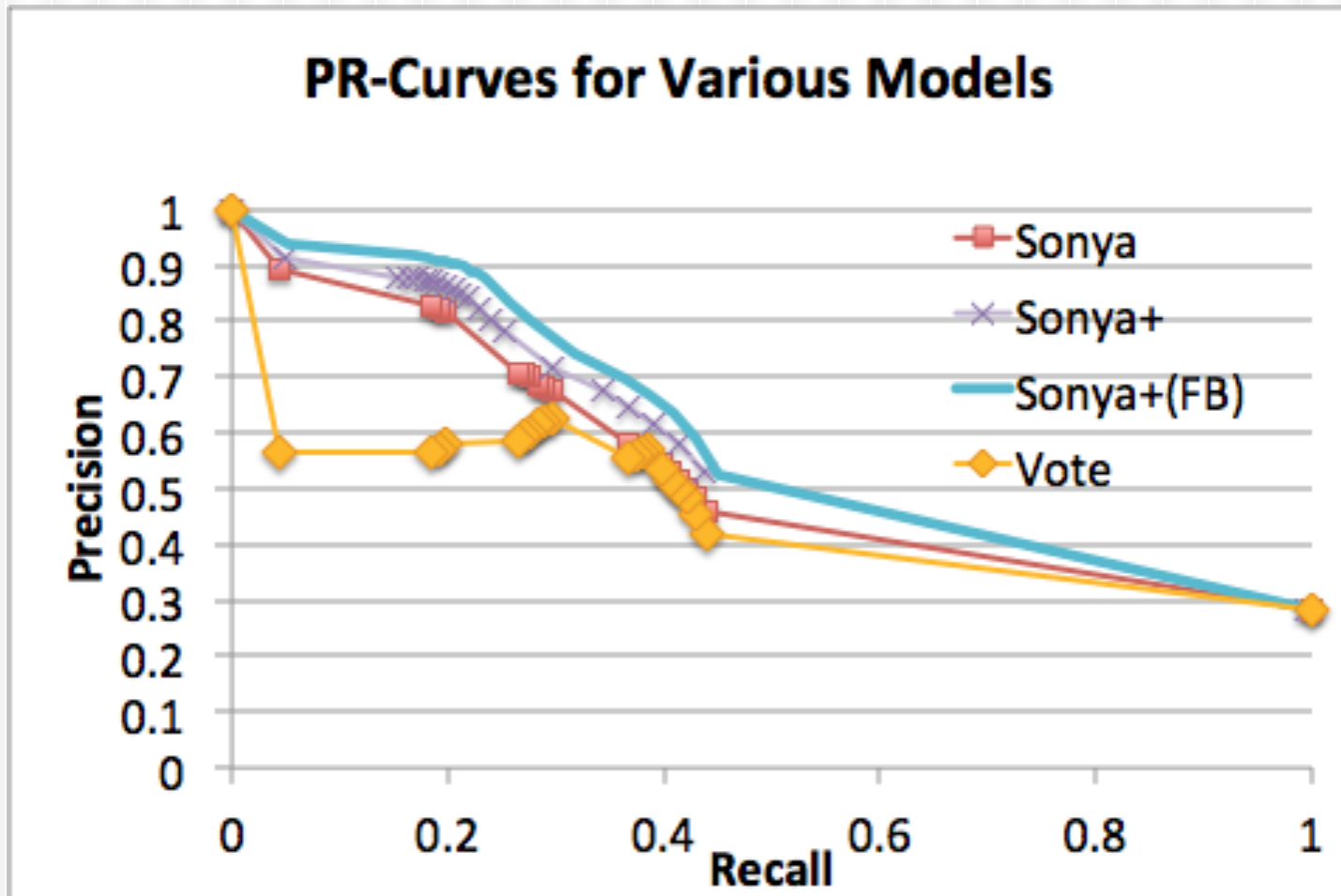


Closer to ideal curve

Smoother curve

Coverage: .993 → .994

Precision-Recall Curve



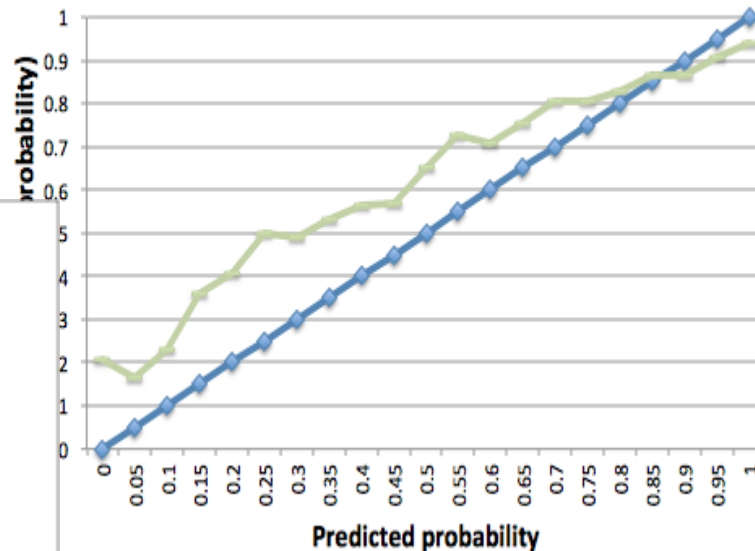
Analysis of Errors

False Negatives



- Multiple truths (13)
- Specific/general value (7)

Calibration Curves



False Postives

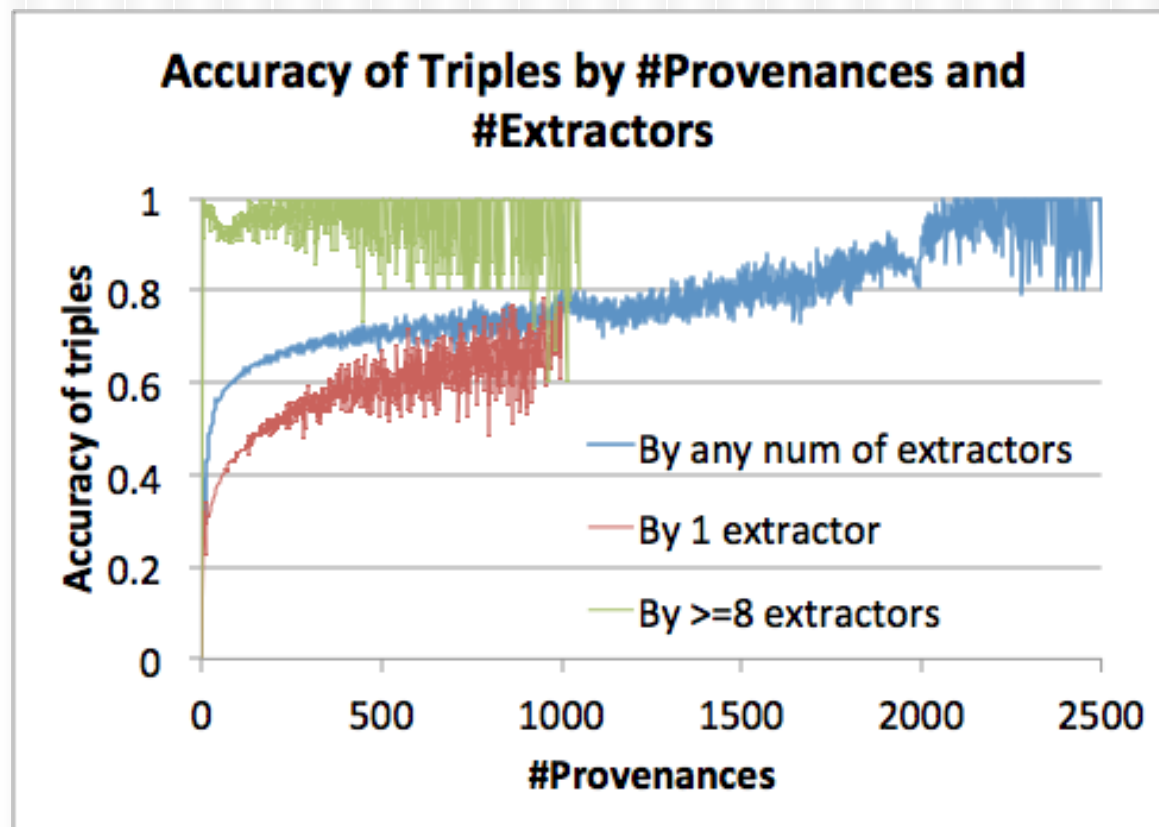


- Common extraction error (8)
- Closed-world assumption (10)
- Wrong value in Freebase (1)
- Hard to judge (1)

Future Directions!!!

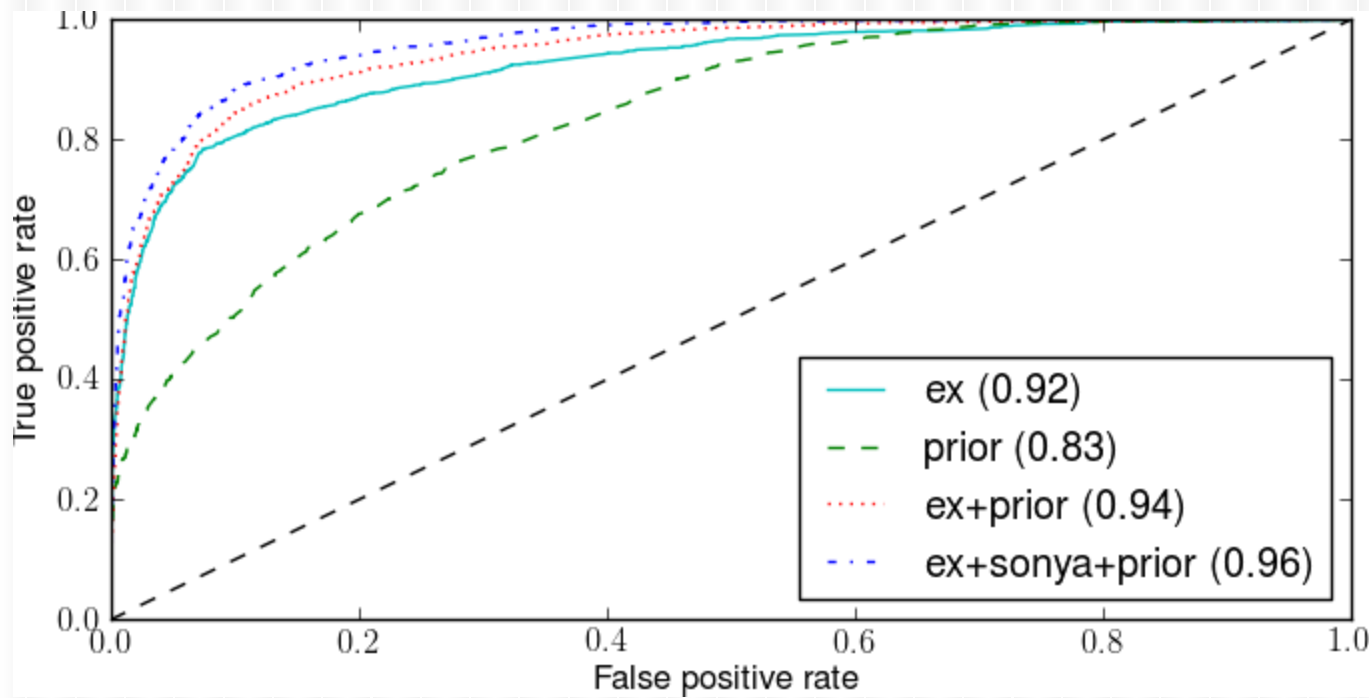
One Inherent Limitation

Cannot distinguish errors from extractors and from sources



Other Fusion Techniques

- Ex: Adaboost learning from extractions
- Prior: (A, parent_of, C), (B, parent_of, C)
→ (A, spouse_of, B)

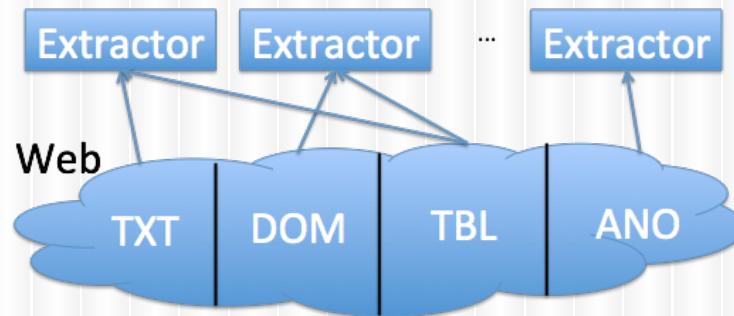


Outline

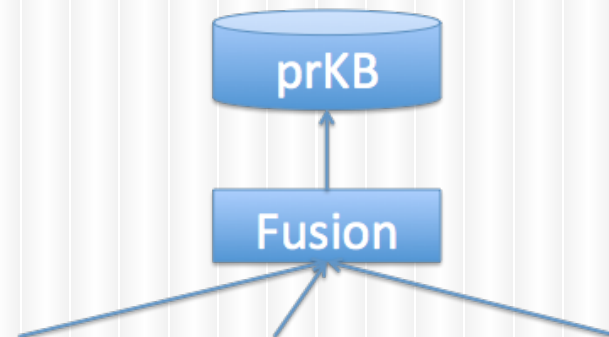
.....



Knowledge extraction



Knowledge fusion

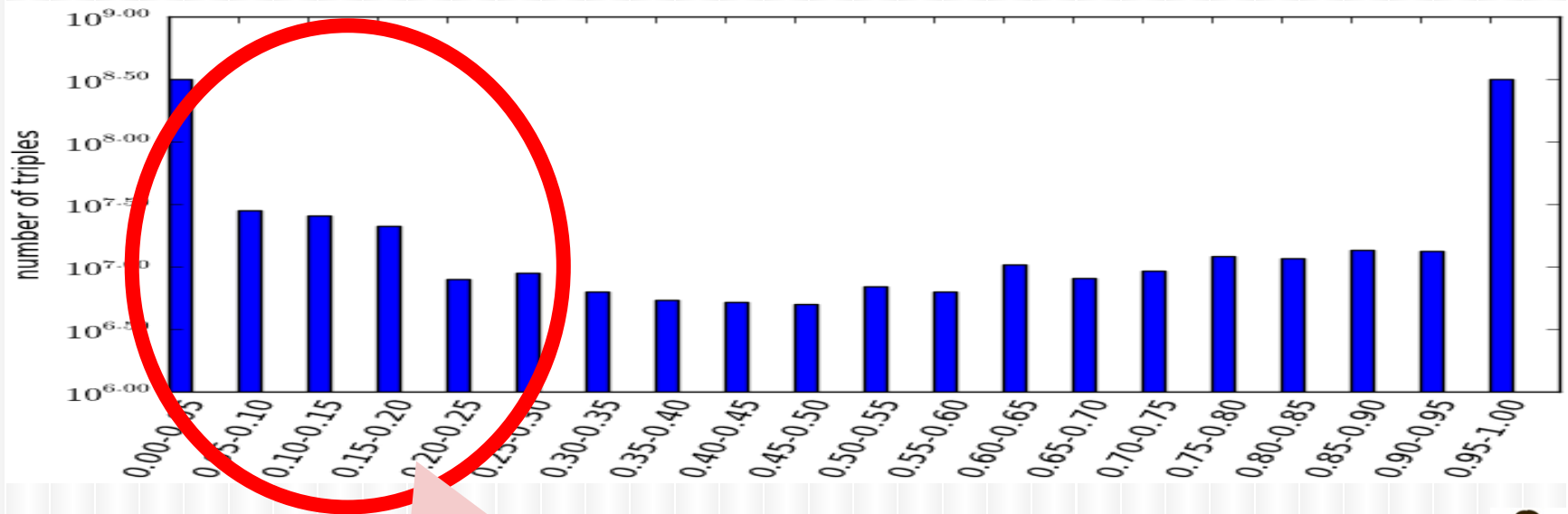


Interesting applications



Future directions

Usage of Probabilistic Knowledge



Negative training examples, and
MANY EXCITING APPLICATIONS!!

- Source errors:
trustworthiness evaluation
- Extraction errors:
data abnormality diagnosis



Application I. A New Angle to Evaluate Web Source Quality

- What we have now
 - Page Rank: links between Websites/Webpages
 - Log based: search log and click-through rate
 - Web spam
 - etc.

Popular Sources w. High Page Rank May Spread Gossip

14 out of 15 Gossip Websites have high page rank

<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

Tale Sources w. Low Page Rank May Provide Valuable Info

salary.com Search for Salaries Jobs Enter a job title Enter a city or postal code

Follow Us @Salary RSS Feed Podcast Facebook LinkedIn YouTube

Salary Job Search Education Career Development Work & Life Features Business Products

Help Me Negotiate I am an Employee or Individual I am an Employer or Business Job or Employee Salary Reports

Contemplating A Move Job Search Salary Information Job-Specific Competencies Compensation Subscriptions Save 44% on 2 Job Postings Save now monster

Over the past 12 months, how frequently have you been bullied at work? Being bullied includes things like being threatened, having rumors spread about you, being attacked verbally or physically, and being excluded from a group on purpose.

☐ Never ☐ Once or twice a month ☐ Once or twice over the past 12 months ☐ Once or twice a week ☐ Once or twice every few months ☐ Almost every day at work

Submit

Kraft Singles KRAFT SINGLES now have NO ARTIFICIAL PRESERVATIVES

Home News Features Donation Add Company Premium Account Contact

Branches

- Agricultural
- Dairies
- Farming
- Fish
- Livestock
- Mixed Crops
- Services
- Tree & Forestry

Countries

- England
- Northern Ireland
- Scotland
- Wales

Regions

- East Midlands
- East of England
- Greater London
- Merseyside
- North East England
- North West England
- North Wales
- Northern Ireland
- Scotland
- Scotland Central
- Scotland North
- Scotland South
- South East

Swim With Dolphins
At Beautiful Blue Lagoon Island. 10% Off Your Reservation Today!

Kate Middleton Photos
Senior Executive Jobs
Document Management

Welcome to the information portal.
Although this information portal is still in development and programming,

amazonmom 20% OFF DIAPERS Learn more

Home About Us Contact Us Privacy Policy Disclaimer Sitemap Drama List Search here...

WOYLAA
All Korean drama episodes english subtitle and RAW

BIG MAN GLORIOUS DAY HOTEL KING ANGEL EYES WONDERFUL DAYS EMPRESS KI

BIG MAN EPISODE 2 ENG SUB
Woylaa.Com - Big Man Episode 2
Synopsis : a drama that tells the story of a man named Kim J...

BIG MAN EPISODE 1 ENG SUB
Woylaa.Com - Big Man Episode 1
Synopsis : a drama that tells the story of a man named Kim J...

HOTEL KING EPISODE 6 ENG SUB
Woylaa.Com - Hotel King Episode 6
Synopsis : A drama tells about love between a man named ...

GLORIOUS DAY EPISODE 2 ENG SUB
Woylaa.Com - Glorious Day Episode 2
Synopsis : A drama tells a story of a woman named Han So...

EMPRESS KI EPISODE 51 ENG SUB
Woylaa.Com - Empress Ki Episode 51
Synopsis : a drama which has a theme of the loves and b...

EMPRESS KI EPISODE 50 ENG SUB
Woylaa.Com - Empress Ki Episode 50
Synopsis : a drama which has a theme of the loves and b...

WONDERFUL DAYS EPISODE 20 ENG SUB
Woylaa.Com - Wonderful Days Episode 20
Synopsis : A drama that tells the story of a young ma...

ANGEL EYES EPISODE 6 ENG SUB
Woylaa.Com - Angel Eyes Episode 6
Synopsis : A drama tells about love story between Park Don...

RECENT DRAMAS

- BIG MAN EPISODE 2 ENG SUB
- EMPRESS KI EPISODE 51 ENG SUB
- BIG MAN EPISODE 1 ENG SUB
- EMPRESS KI EPISODE 50 ENG SUB
- HOTEL KING EPISODE 6 ENG SUB
- WONDERFUL DAYS EPISODE 20 ENG SUB
- GLORIOUS DAY EPISODE 2 ENG SUB
- ANGEL EYES EPISODE 6 ENG SUB
- GLORIOUS DAY EPISODE 1 ENG SUB
- ANGEL EYES EPISODE 5 ENG SUB

POPULAR DRAMAS

- EMPRESS KI EPISODE 50 ENG SUB
- EMPRESS KI EPISODE 49 ENG SUB
- CUNNING SINGLE LADY EPISODE 15 ENG SUB
- CUNNING SINGLE LADY EPISODE 16 ENG SUB

Backingtrackguitar.com

Backing Tracks Guestbook Terms Of Use

AdChoices Guitar Pro Guitar Tabs MIDI Guitar Guitar Jam

66 Search

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

bestbackingtracks...

Looking for Guitar Backing Tracks? Buy our Guitar Backing Track Album

A backing track is an audio or MIDI recording that musicians play or sing along to in order to add parts to their music which would be impractical to perform live.

We have collected over 2000 backing tracks for you. You can listen them online and download any backing track for free.

No registration required.

Made by fans for fans.

The Platinum Card® from American Express with up to \$200 in annual Fee Credits annually

PLUS, EARN 40,000 POINTS*

AMEX CARD 3759 81234567890101

Tale Sources w. Low Page Rank May Provide Valuable Info

A screenshot of a Google search interface. The search bar at the top contains the text "who play Boulevard of Broken Dreams". Below the search bar, there are tabs for "Web", "Videos", "Images", "Shopping", "News", "More", and "Search tools". The "Web" tab is selected. Below the tabs, it says "About 17,600,000 results (0.35 seconds)". The first search result is a Wikipedia entry titled "Boulevard of Broken Dreams (Green Day song) - Wikipedia ...". The snippet of the Wikipedia entry reads: "Boulevard of Broken Dreams" is a song by American punk rock band **Green Day**. It was released as the second single from their seventh album, American Idiot. The song was written by **Green Day**, with lyrics by lead singer **Billie Joe Armstrong**. Below the snippet is a "Feedback" link. The second search result is a YouTube video titled "How to Play Boulevard of Broken Dreams by Green Day On ...". The video thumbnail shows a man playing an acoustic guitar. The video details include the URL "www.youtube.com/watch?v...", the upload date "Jun 8, 2010", the uploader "mahalodotcom", and a description "Check out Bas Rutten's Liver Shot on MMA Surge: http://bit.ly/MMASurgeEp1 ...".

who play Boulevard of Broken Dreams

Web Videos Images Shopping News More Search tools


About 17,600,000 results (0.35 seconds)

"Boulevard of Broken Dreams" is a song by American punk rock band **Green Day**. It was released as the second single from their seventh album, American Idiot. The song was written by **Green Day**, with lyrics by lead singer **Billie Joe Armstrong**.

[Boulevard of Broken Dreams \(Green Day song\) - Wikipedia ...](#)
[en.wikipedia.org/.../Boulevard_of_Broken_Dreams_\(Green_D...](http://en.wikipedia.org/.../Boulevard_of_Broken_Dreams_(Green_D...) Wikipedia

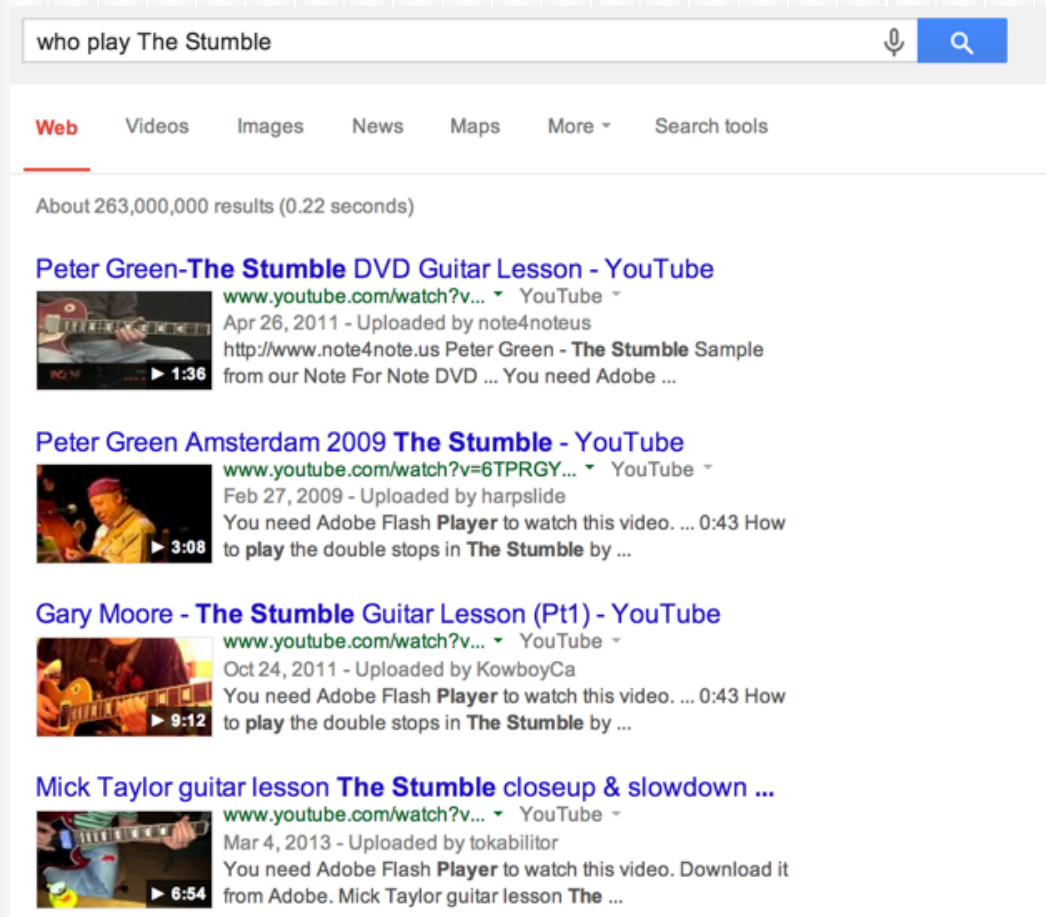
Feedback

[How to Play Boulevard of Broken Dreams by Green Day On ...](#)

 www.youtube.com/watch?v... YouTube
Jun 8, 2010 - Uploaded by mahalodotcom
Check out Bas Rutten's Liver Shot on MMA Surge:
<http://bit.ly/MMASurgeEp1> ...

Good WebAnswer for an award-winning song

Tale Sources w. Low Page Rank May Provide Valuable Info



Missing WebAnswer for a not-so-popular song

Tale Sources w. Low Page Rank May Provide Valuable Info

Backingtrackguitar.com

Backing Tracks Guestbook Terms Of Use

UP TO **70% OFF**
home décor

AdChoices ▶ [Guitar For](#) ▶ [Drum Tracks](#) ▶ [Old Guitar](#) ▶ [Free Guitar](#)

f t Q+ B p d i

Search

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

the stumble backing track

Download

Information about backing track

Artist	King Freddie
Rating	★★★★★☆☆☆
Filesize	3.66 Mb
Length	00:04:00

GET INTO IT!
Experience the Bahamas' #1 Attraction

Very precise info on guitar players but low Page Rank

Application I. A New Angle to Evaluate Web Source Quality



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
ᠠᠷᠠᠩᠭᠡᠨᠢᠰ

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](#)^{[1][11]} 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 km²) in total and with around 317

United States of America



Flag

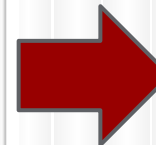


Great Seal

Motto:

"In God we trust" (official)^{[1][2][3]}
"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

Application I. A New Angle to Evaluate Web Source Quality



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
Açèh
Адыгэбзэ
Afrikaans
Alemannisch
አማርኛ
Ænglisc
Англис
العربية
Aragonés
ᠠᠷᠠᠨᠤᠯᠤᠰ

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](#)^{[10][11]} 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 km²) in total and with around 317

United States of America



Flag

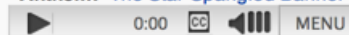


Great Seal

Motto:

"In God we trust" (official)^{[1][2][3]}
"E pluribus unum" (Latin) (traditional)
"Out of many, one"

Anthem: "The Star-Spangled Banner"



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

Triple 1	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
...	...
Accu	0.7

Application I. A New Angle to Evaluate Web Source Quality



The screenshot shows the Wikipedia article for the United States. The sidebar on the left includes links for Main page, Contents, Featured content, Current events, Random article, Donate to Wikipedia, and Wikimedia Shop. The main content area starts with the title "United States" and a redirect notice from "USA". It then provides a brief overview of the country, mentioning its status as a federal republic, its 50 states and federal district, and its geographical location. A red arrow points from the article to the table on the right.

	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
...
Accu	0.73	

Extraction and Triple Correctness

Example. (Obama, nationality, ?)

(Obama, nationality, Bolivarianism) (many many such objects)

- 3 extractions ($Pr_{ext}Corr=0.01$)

<http://mathaba.net/news/?x=631316>

<http://www.laht.com/article.asp?ArticleId=329187&CategoryId=10717>

<http://www.iamericas.org/en/about-ioa/presidents-corner>

Chávez Calls on Obama to Join Him in the Socialist Revolution

"Come on, Obama, align yourself with us on the way to socialism!" said the Venezuelan leader who this week also expropriated plants and lands of Venezuela's Polar, US firm Cargill, and Irish firm Smurfit. "Come on, it's the only way!"

By Jeremy
Morgan
Latin American
Herald Tribune
staff

CARACAS —
President Hugo
Chávez invited
President Barack



- $Pr_{triple}Corr=0$

Extraction and Triple Correctness

Example. (Obama, nationality, ?)

(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/-/ksxaqx/-/index.html>

US will respect winner of
Kenya election, Obama says

SHARE BOOKMARK PRINT RATING☆☆☆☆

- $Pr_tripleCorr=0$ (not enough support)

Extraction and Triple Correctness

Example. (Obama, nationality, ?)

(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 and you can follow him on Twitter @timothy_stanley.

 Follow 15.6K followers



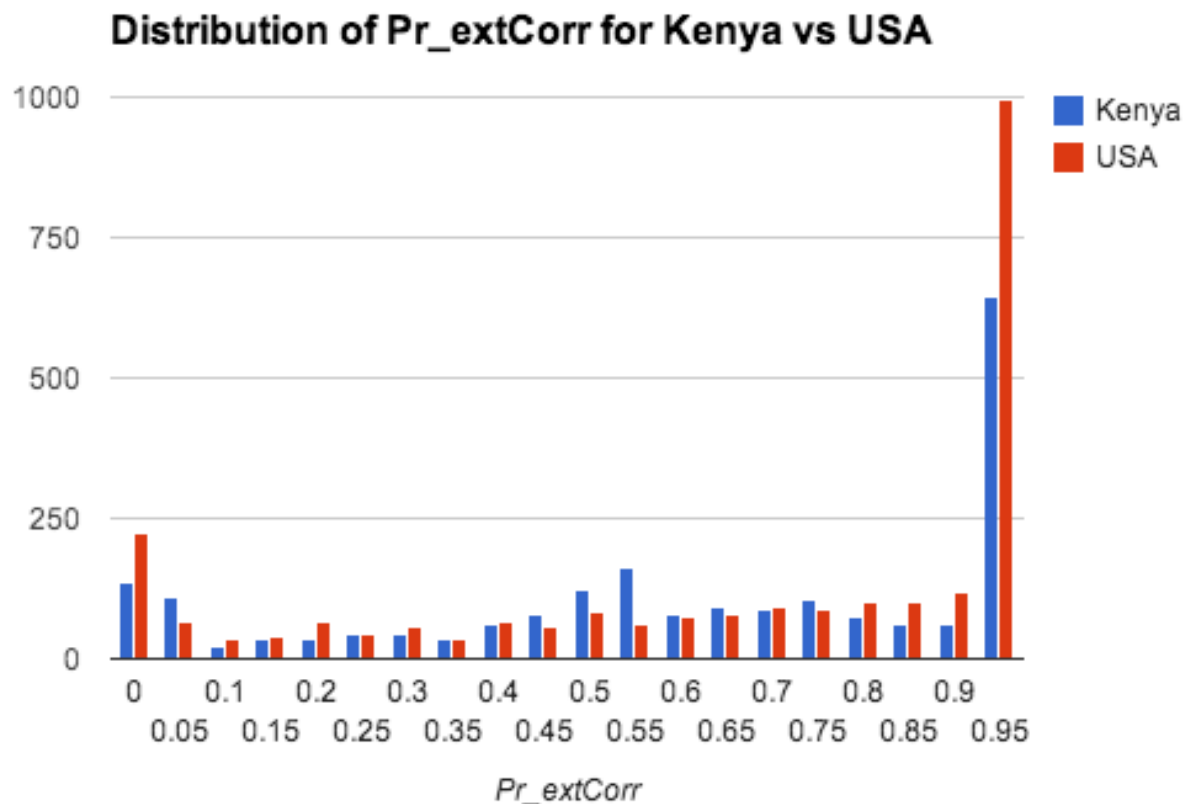
Barack Obama's life story contains 'myth, not truth', says biographer. So why did the media report it as truth?

- $Pr_tripleCorr=1$ (Higher support)

Extraction and Triple Correctness

Example. (Obama, nationality, ?)

Distribution of providers for Kenya and USA



Sonya Trustworthiness Score

Domain	#Triples	Sonya Score
www.eonline.com	12,871	0.363
perezhilton.com	46,912	0.427
radaronline.com	3,530	0.489
www.zimbio.com	2,464,452	0.530
mediatakeout.com	131	0.531
gawker.com	6,055	0.567
www.popsugar.com	1,805	0.576
www.people.com	16,886	0.585
www.tmz.com	8,149	0.621
www.fishwrapper.com	14	0.622
celebrity.yahoo.com	11,187	0.677
wonderwall.msn.com	2,524	0.684
hollywoodlife.com	4,536	0.689
www.wetpaint.com	19,284	0.730

Many gossip Web sites DO provide quite a lot of wrong factual information

Sonya Trustworthiness Score

- Example for (URL, Predicate)

URL: <https://ibirthdayworld.blogspot.com/2010/03/celebrity-birthdays-on-march-22.html>

Predicate: date_of_birth

- #Facts = 42; Trustworthiness = 0.95



Source: e.g., http://en.wikipedia.org/wiki/World_Chess_Championship_2013 or en.wikipedia.org

Pred: e.g., [/people/person/place_of_birth](#)

Max results to display:

Support threshold:

Source	Pred	Num_of_Triples	Accuracy
http://ibirthdayworld.blogspot.com/2010/03/celebrity-birthdays-on-march-22.html	/people/person/date_of_birth	42	0.953

Sonya Trustworthiness Score

.....

Celebrity Birthdays On March 22



<-Marcel Marceau

Below are 124 famous people born on March 22.

Browse [Gift Ideas](#) - Browse [Ecards](#)

The names in brackets below are duplicate entries.

Aaron North was born on March 22, 1979. American guitarist.

Amy Stud was born on March 22, 1986. English singer-songwriter and musician.

Andreas Johnson was born on March 22, 1970. Swedish pop and rock singer-songwriter and musician.

Andrew Lloyd Webber was born on March 22, 1948. British composer of musicals.

Angelo Badalamenti was born on March 22, 1937. American composer.

Anja Kling was born on March 22, 1970. German actress.

Annabelle Apsion was born on March 22, 1963. English actress.

Anne Hyde was born on March 22, 1638. Wife of James II of England.

Anthony van Dyck was born on March 1599. Flemish Baroque artist.

Armin Hary was born on March 22, 1937. German athlete.

Avraham Fried was born on March 22, 1959. American singer-songwriter and musical entertainer.

Sonya Trustworthiness Score

.....

- Example for (URL, Predicate)

URL: <https://ibirthdayworld.blogspot.com/2010/03/celebrity-birthdays-on-march-22.html>

Predicate: date_of_birth

- #Facts = 42; Trustworthiness = 0.95

- Mistake: Anne Hyde (the URL says: 3/22/1638; Wiki/KG says: 3/12/1637)

Anne Hyde

Anne Hyde was Duchess of York and Albany as the first wife of James, Duke of York, later King James II and VII. Originally Anglican, her father was a lawyer. [Wikipedia](#)



Born: March 12, 1637, [Windsor, United Kingdom](#)

Died: March 31, 1671, [London, United Kingdom](#)


Spouse: [James II of England](#) (m. 1660–1671)

Children: [Anne, Queen of Great Britain](#), [Mary II of England](#), [More](#)

Parents: [Edward Hyde, 1st Earl of Clarendon](#), [Frances Hyde, Countess of Clarendon](#)

Siblings: [Laurence Hyde, 1st Earl of Rochester](#), [Henry Hyde, 2nd Earl of Clarendon](#)

Application II. Provide An X-Ray for Extracted Data

- Goal: Help users analyze errors, changes and abnormalities in data 
- Intuitions: cluster errors by features and return clusters with top error rates

Application II. Provide An X-Ray for Extracted Data

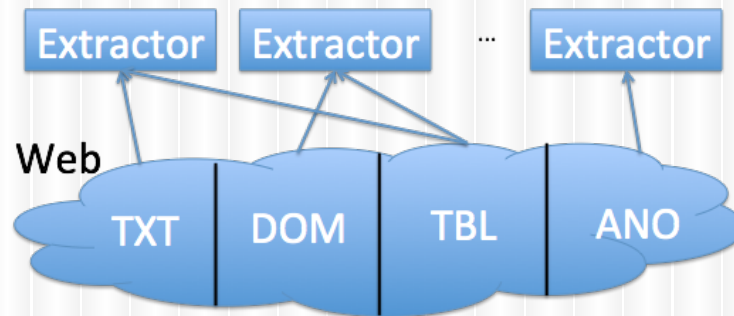
- Cluster 1.
 - Feature: (besoccor.com, date_of_birth, 1986_02_18)
 - #Triples: 630; Errs: 100%
 - Reason: default value
- Cluster 2.
 - Feature: (ExtractorX, pred: namesakes, obj:the county)
 - #Triples: 4878; Errs: 99.8%
 - E.g., [Salmon P. Chase, namesakes, The County]
 - Contexts: *The county* was named for Salmon P. Chase, former senator and govenor of Ohio
 - Reason: Unresolved coreference

Outline

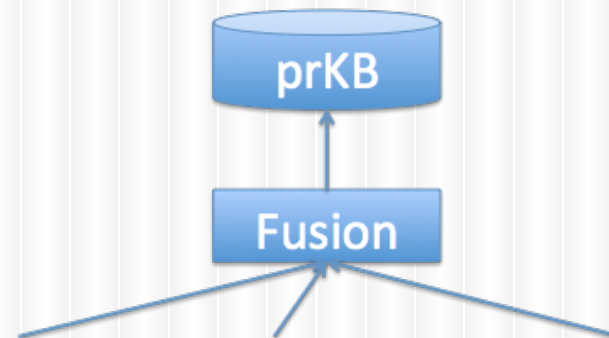
.....



Knowledge extraction



Knowledge fusion



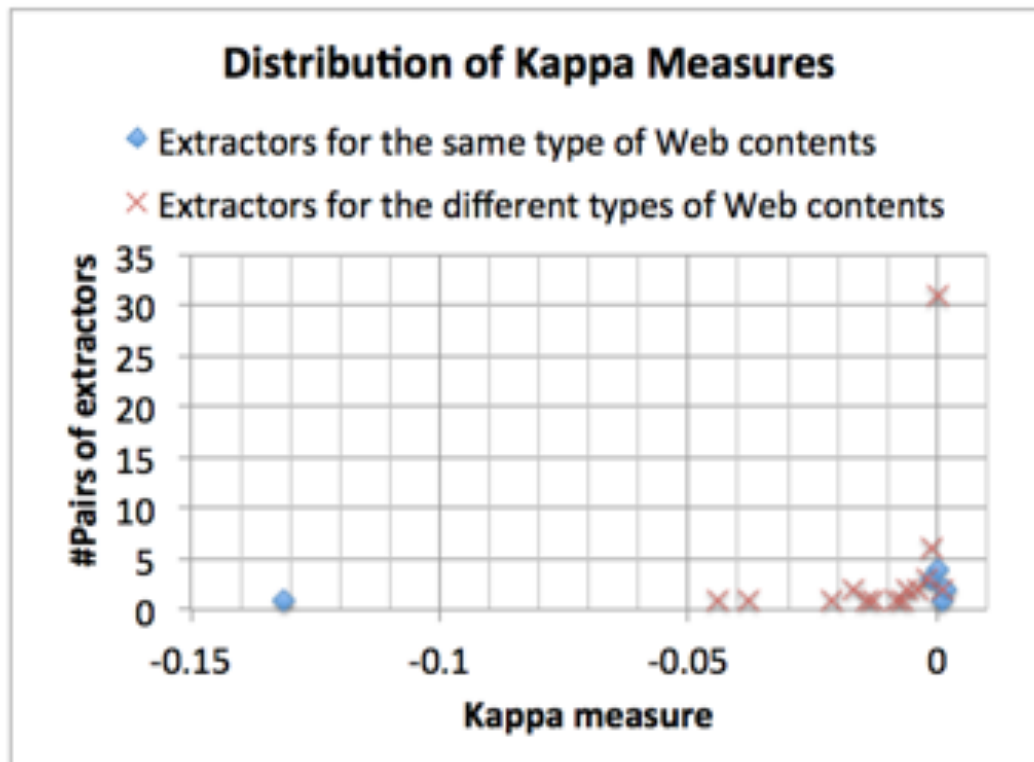
Interesting applications



Future directions

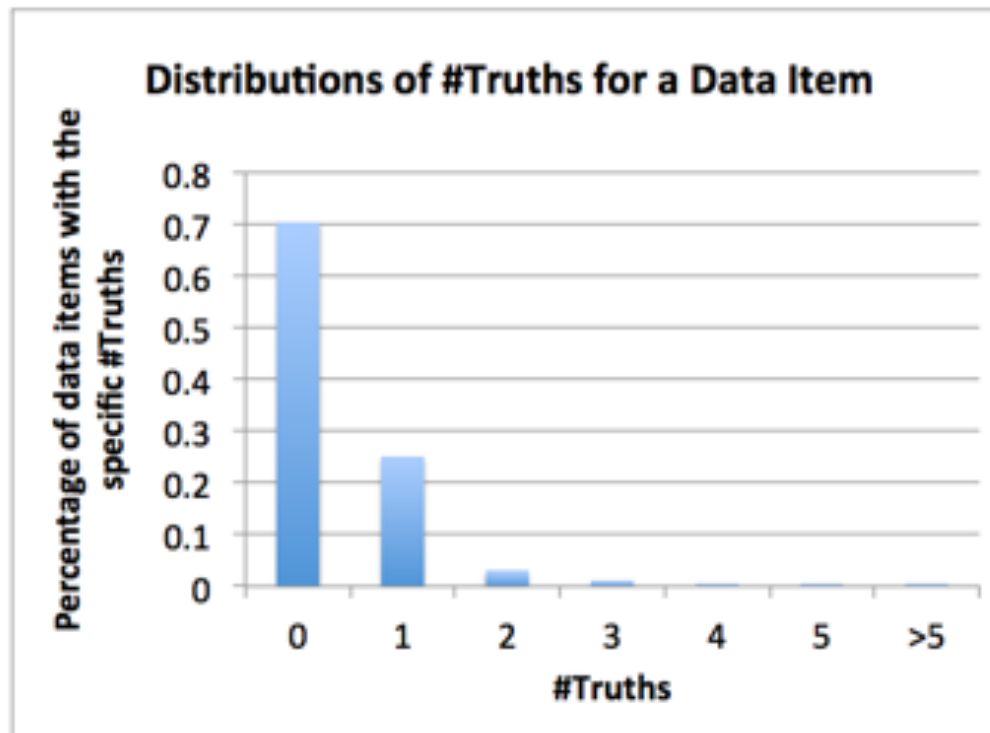
Future Directions: Remove the Assumptions One by One

Assumption I. Independence between pairs of provenances (i.e., (URL, extractor))



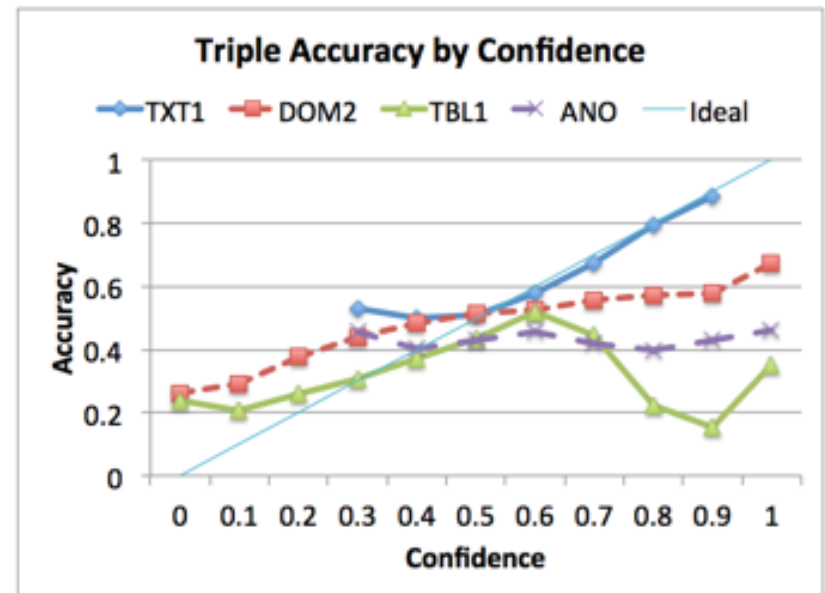
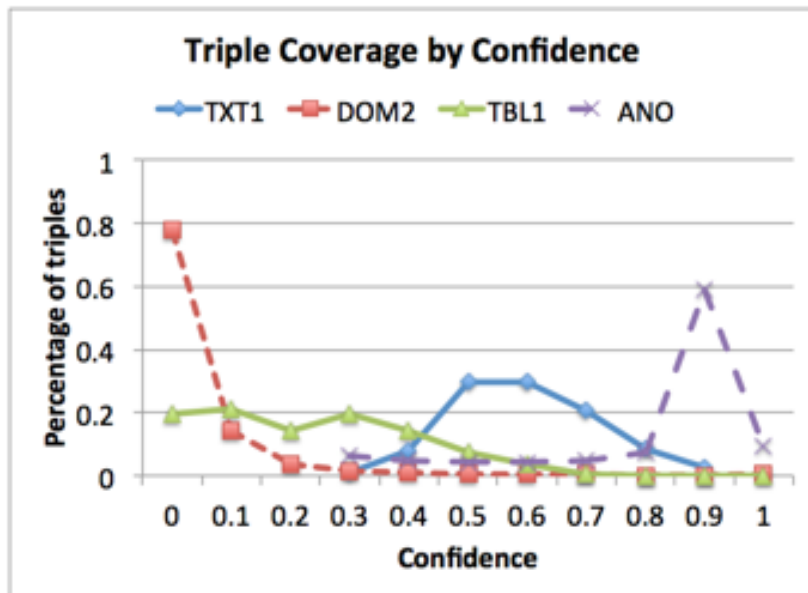
Future Directions: Remove the Assumptions One by One

Assumption II. Single true object for each (sub, pred)



Future Directions: Remove the Assumptions One by One

Assumption III. Extractions are deterministic



Future Directions: Remove the Assumptions One by One

Assumption IV. Values (objects) are categorical

Assumption V. We have enough data to judge accuracy of each source

Assumption VI. Local closed-world assumption in evaluation

Future Directions: Remove the Assumptions One by One

Assumption VII. Global closed-world assumption--consider only existing entities and predicates in FB

WE NEED SOMETHING NEW!!!



TAKE AWAYS

- A new area--Knowledge Fusion
- We can solve KF problem fairly well by adapting DF methods
- Many interesting future directions for KF!
- Many exciting applications for the prKB!!

Acknowledgement

.....



Evgeniy Gabrilovich (Manager, need to say anything?)



Jeremy Heitz (Strongest supporter)



Wilko Horn (Strictest code reviewer)



Kevin Murphy (Intelligent consultant)



Shaohua Sun (Critical representer to the outside world)



Wei Zhang (Fearless explorer of new ideas)

THANK YOU!

Questions?