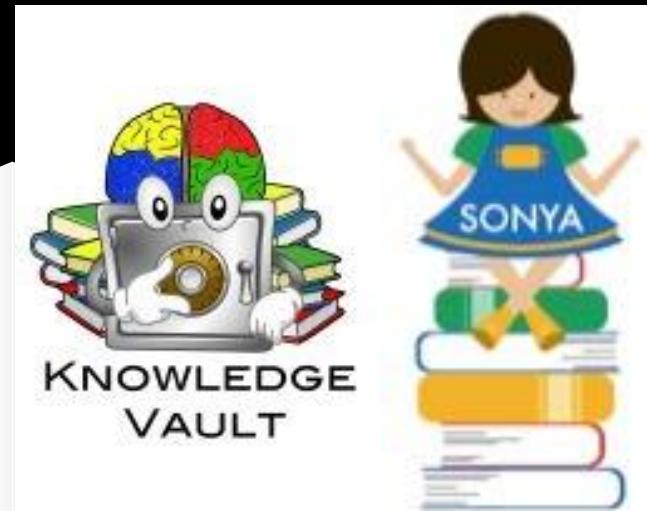


From Data Fusion to Knowledge Fusion

.....

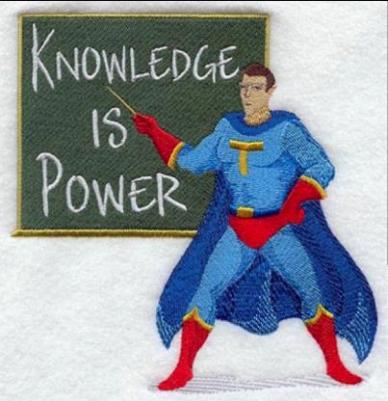
Xin Luna Dong, Google Inc.
9/13/2014 @ WISA '14



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



Knowledge Is Power



- Many Knowledge Bases (KB)



NELL: Never-Ending Language Learning

ProBase



facebook

Walmart

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

Google 搜索结果：南开

Web News Maps Shopping Images More Search tools

About 13,900,000 results (0.26 seconds)

南开大学
www.nankai.edu.cn/ ▾ Translate this page Nankai University ▾
当地时间7月12日，在拉脱维亚里加举办的第八届世界合唱比赛第一阶段比赛中，南开大学学生合唱团获得青年混声组、有伴奏宗教组、有表演民谣合唱组3项金奖。南.
4.3 ★★★★★ 8 Google reviews · Write a review

📍 94 Weijin Rd, Nankai, Tianjin, China
+86 22 2350 8219
专业学院 - 学生 - 南开大学图书馆 - 职能部门

南开大学- 维基百科，自由的百科全书
zh.wikipedia.org/zh/南开大学 ▾ Translate this page Chinese Wikipedia ▾
南开大学（简称：南开、南大、NNU），原称私立南开大学，主校区坐落于天津市南开区八里台。南开大学是1919年由近代著名爱国教育家严修、张伯苓创立的私立大学，...
南开大学附属中学 - 南开大学校钟 - 南开大学体育中心 - 南开大学主楼

南开大学首页_中国高校信息查询系统_腾讯高考频道
data.edu.qq.com/college_info/2/ ▾ Translate this page
南开大学创建于1919年，创办人是近代著名爱国教育家张伯苓和严修。抗日战争时期，南开大学与北京大学、清华大学在昆明组成举世闻名的西南联合大学，被誉为...

南开大学,分数线,专业设置_新浪院校库_新浪教育_新浪网
kaoshi.edu.sina.com.cn ▾ 所有院校 ▾ Translate this page Sina Corp ▾
南开大学是国家教育部直属重点综合性大学，是敬爱的周恩来总理的母校。南开大学创建于1919年，创办人是近代著名爱国教育家严修和张伯苓。抗日战争时期，南开...

南开大学_百度百科
baike.baidu.com/view/2967.htm ▾ Translate this page Baidu Baike ▾
南开大学被誉为“学府北辰”，是中国最著名的高等院校之一。其诞生于1919年，旧称“私立南开学校”，周恩来总理曾在此短暂就读；抗日战争时期，南开大学与北京大学、...

南开大学吧_百度贴吧
tieba.baidu.com/f?kw=南开大学 ▾ Translate this page Baidu ▾

Nankai University Directions

University in Tianjin, China

Nankai University, often known as Nankai, is a public research university located in Tianjin mainland China. Founded in 1919 by prominent educators Zhang Boling and Yan Fansun, Nankai is one of the most prestigious top universities in China. [Wikipedia](#)

Address: 94 Weijin Rd, Nankai, Tianjin, China
Phone: +86 22 2350 8219
Enrollment: 21,905 (2010)
Founder: Zhang Boling
Founded: 1919
Colors: White, Violet

See results about

Nan Kai University of Technology (University in Taiwan)
Founded: 1971 

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.”

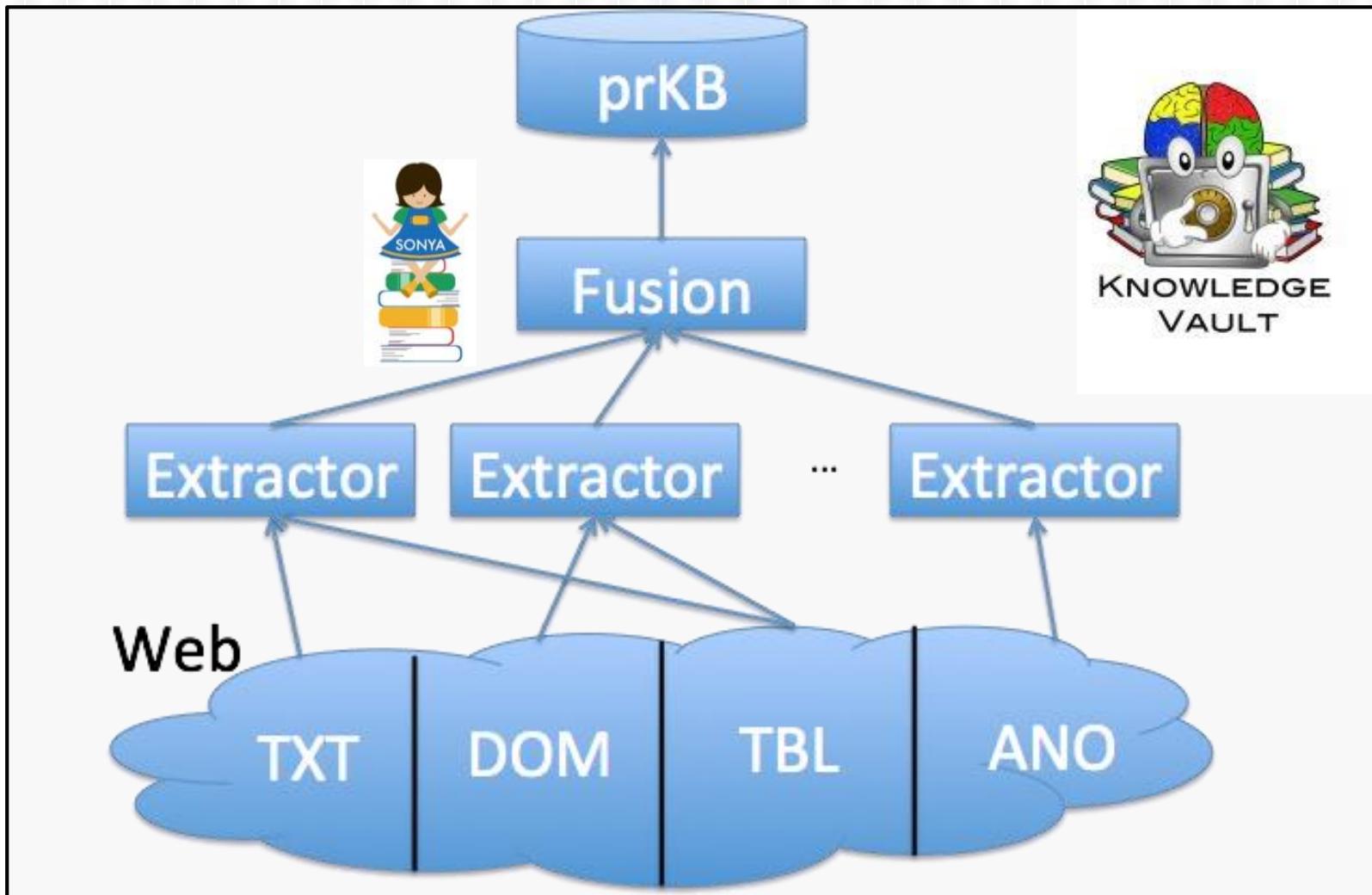
—Håkon Wium Lie

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]





Australia & New Zealand

Home

Health

Environment

Culture

Space

Technology

Opinion

News

Google's Knowledge Vault already contains 1.6 billion facts

FELICITY NELSON

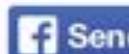
SATURDAY, 23 AUGUST 2014



275



64

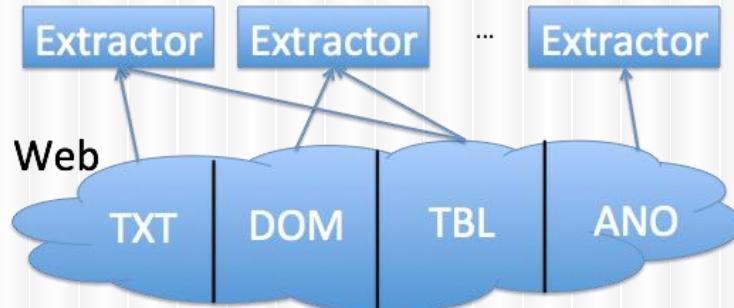


The automated, fact-harvesting bot will build up a collection of all human knowledge.

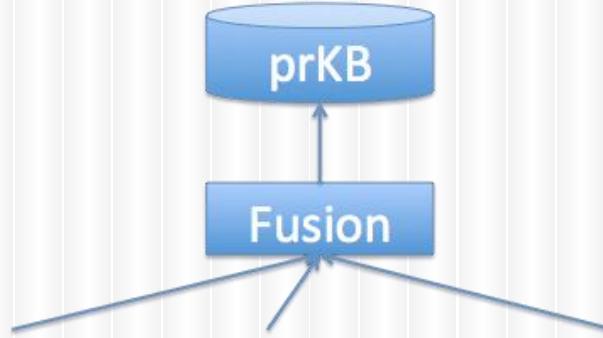
Outline



Knowledge extraction



Knowledge fusion



Interesting applications

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)*

Knowledge Extraction II–Sources

Web



Free texts

Synopsis

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, engineer, and scientist. His ideas and body of work -- which include the *Last Supper*, *Leda and the Swan* and influenced countless artists and made da Vinci a central figure of the Italian Renaissance.

Print Cite This



DOM Tr

Welcome About Me Write a Review Find Friends

Shana Thai Restaurant

143 reviews 100% Rating Details

Category: Thai (264)

311 Moffett Blvd

Ste A

Mountain View, CA 94031

(844) 940-9999

<http://www.shanathai.com>

Explore the menu:

Hours:

Mon-Sun: 11 am - 9 pm

Mon-Sun: 8 am - 10 pm

Kids Eat Free: Yes

Accepts Credit Cards: Yes

Parking: Private Lot

Aircon: Casual

Good For Groups: Yes

Price Range: \$

Takes Reservations: Yes

Deli/Salad: No

Takeout: Yes

Washer Service: Yes

Outdoor Seating: Yes

Hi-Fi: No

Good For Dinner: Yes

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 (D-R)	1801–1809	Va.	4/13/1743	7/4/1826
4.	Madison (D-R)	1809–1817			

Annotations

<h1 itemprop="name">
Tom Cruise </h1>

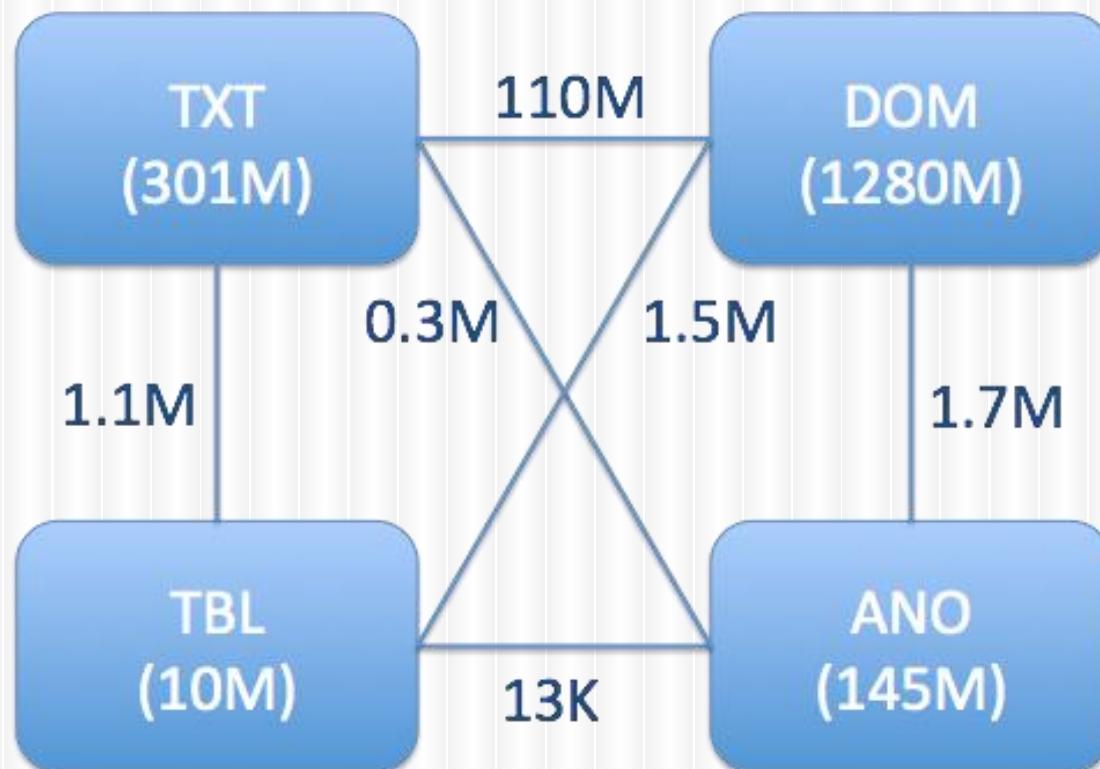
7/3/1962

Male

schema.org

Statistics for Web Sources

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



As of 11/2013

Knowledge Extraction III— Extractors

- Three tasks (any order, maybe combined)
 - I. Triple identification
 - II. Entity linkage

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

III. Predicate linkage

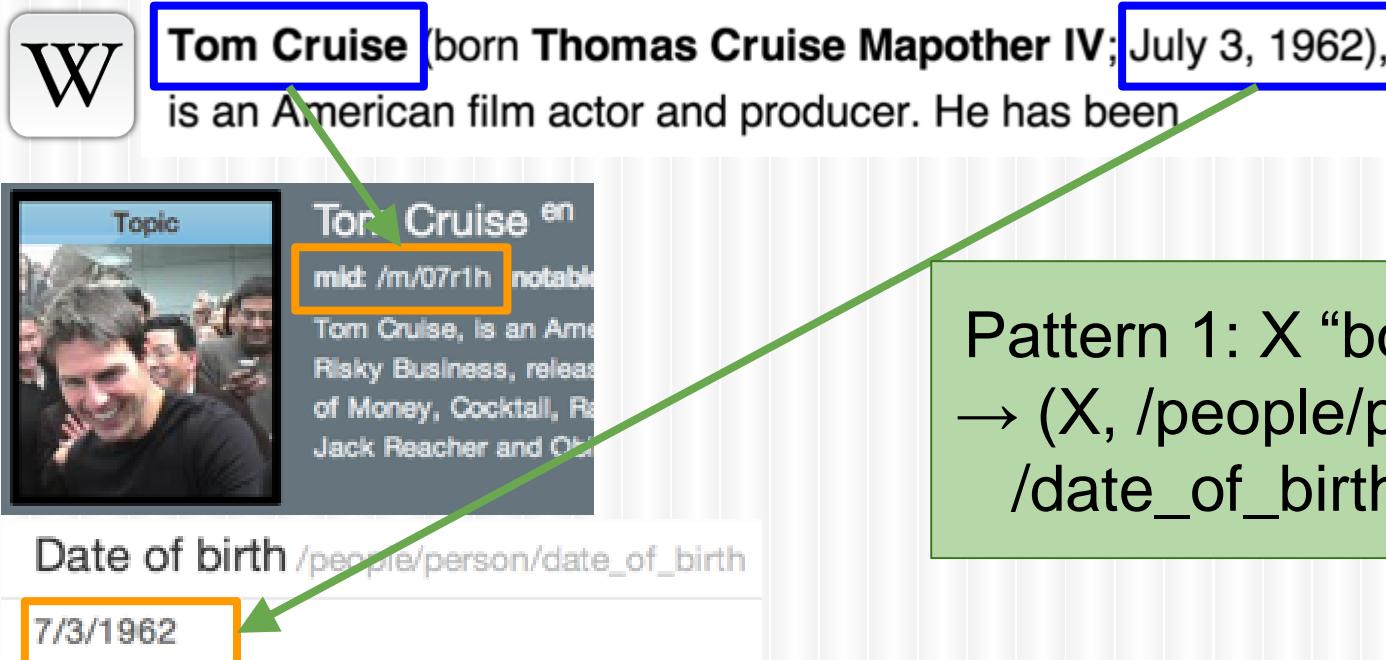


Tom Cruise
mid: /m/07r1h notable
Tom Cruise, is an Ameri
Risky Business, release
of Money, Cocktail, Ris
Jack Reacher and Obl

/people/person/date_of_birth

Knowledge Extraction III– Extractors

- Texts/DOM: distant supervision



- 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)

The screenshot shows a web browser window with the following details:

- Title Bar:** VOA Obama to Campaign For C x
- Address Bar:** www.voanews.com/content/a-13-2009-10-01-voa56/4143... ★ PDF ☰
- Page Headers:** VOA Sites by Language ▾ TOP STORIES: OBAMA, PUTIN FAIL TO RESOLVE DIFFERENCES ON SYRIA | ABE TO XI: LET'S...
- Logo:** VOA Voice of America
- Navigation Bar:** HOME USA AFRICA ASIA MIDEAST EUROPE SCIENCE & TECH HEALTH ENTE...
- Section:** News
- Article Title:** Obama to Campaign For Chicago's 2016 Olympic Bid
- Article Summary:** (The title is highlighted with a yellow box.)
- Interaction Buttons:** Print Comment Share:

Errors Can Creep in at Every Stage

Reconciliation error:
(Obama, nationality, North America)

American President Barack Obama

influences from all over the world. There's a touch of this, a little smidgeon of that, a dash of something else, like when you're cooking." On the song "The Queen Is Back", Summer reveals her wry and witty self-awareness of her musical legacy and her public persona. "I'm making fun of myself," she admits. "There's irony. It's poking fun at the idea of being called a queen. That's a title that has followed me, followed me and followed me. We were sitting and writing and that title kept popping up in my mind and I'm thinking, 'Am I supposed to write this? Is this too arrogant to write?' But people call me 'the queen,' so I guess it's ok to refer to myself as what everybody else refers to me as. We started writing the song and thought it was kind of cute and funny." Summer wrote "The Queen Is Back" and "Mr. Music" with J.R. Rotem and Evan Bogart, the son of Casablanca Records founder Neil Bogart.

On December 11, 2009, Summer performed at the Nobel Peace Prize Concert in Oslo, Norway in honor of American President Barack Obama. She was backed by the Norwegian Radio Orchestra.

2010–12: Final recordings [edit source | edit beta]

On July 29, 2010, Summer gave an interview with Allvoices.com wherein she was asked if she would consider doing an album of standards. She said, *I actually am, probably in September. I will begin work on a standards album. I will probably do an all-out dance album and a standards*

Errors Can Creep in at Every Stage

Source data error: (Obama, nationality, Kenya)

The screenshot shows a web browser window with the title bar "7 Obama Born In Kenya? His...". The URL in the address bar is "www.israelnationalnews.com/blogs/message.aspx/30...". The page content includes a banner for "Guard Your Tools for Br FREE & A", a sidebar with "Main | News | Radio | More | Op-Eds | Judaism | Video | News Briefs", and a main article titled "Obama Born In Kenya? His Grandmother Says Yes." by Tamar Yonah. A large orange callout box with the text "Obama born in Kenya" points to the article title.

Obama born in Kenya

7 Obama Born In Kenya? His...

www.israelnationalnews.com/blogs/message.aspx/30...

Main | News | Radio | More | **Op-Eds** | Judaism | Video | News Briefs

Tishrei 13, 5769, 10/12/2008

Obama Born In Kenya? His Grandmother Says Yes.

by Tamar Yonah

Someone is lying. According to Obama's Kenyan (paternal) grandmother, as well as his half-brother and half-sister, Barack Hussein Obama was born in Kenya, not in Hawaii as the Democratic candidate for president claims. His grandmother boasted that her grandson is

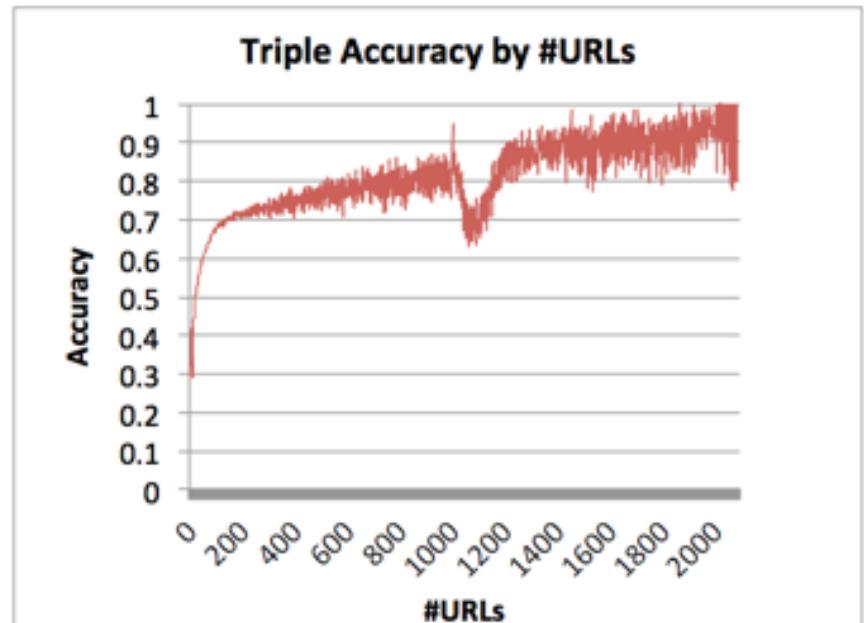
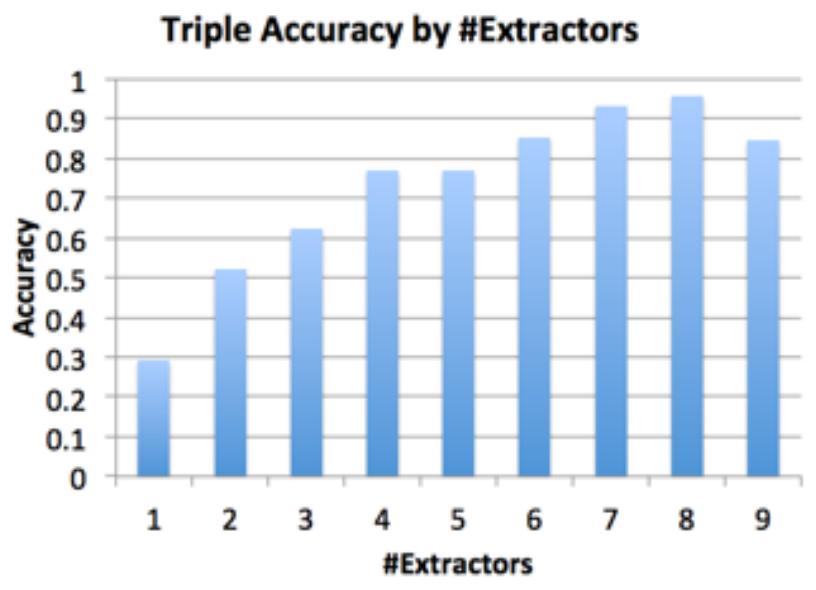
Knowledge Extraction IV–Quality

- 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



Statistics for Extractors

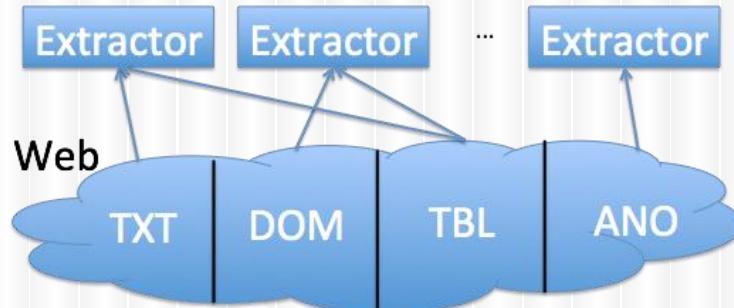
- 12 extractors; high variety

	#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

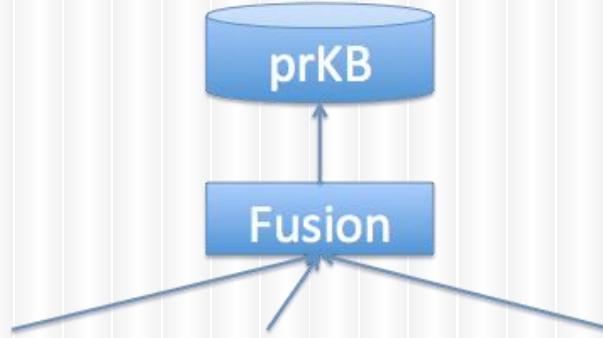
Outline



Knowledge extraction



Knowledge fusion

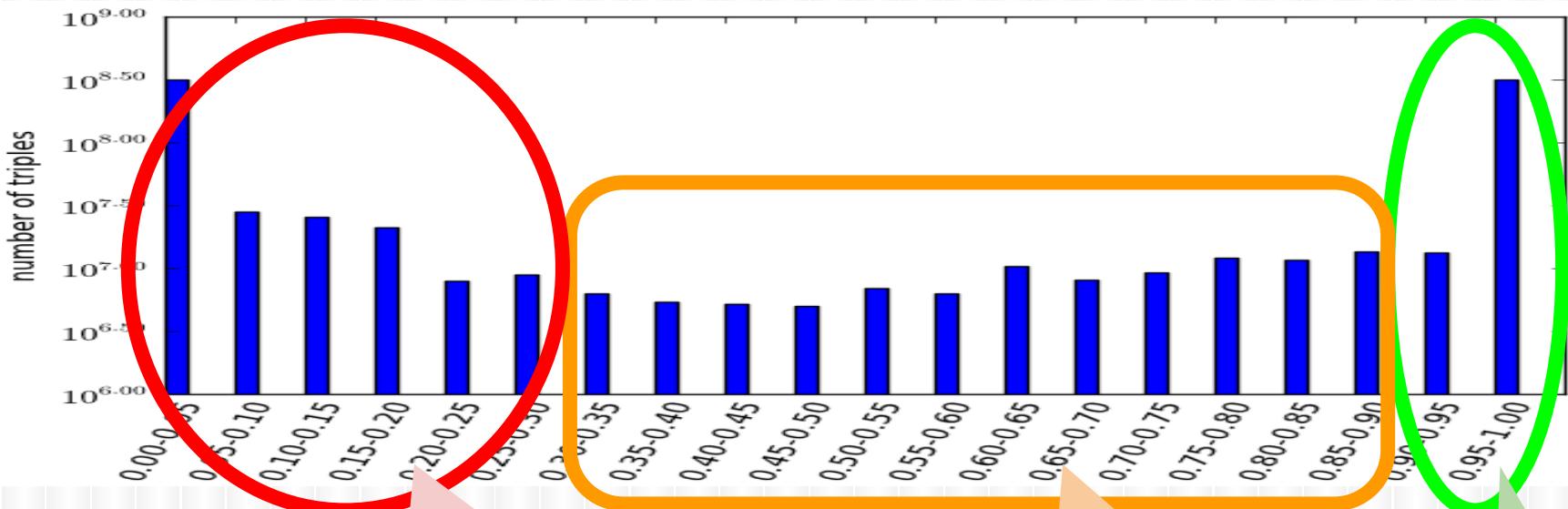


Interesting applications

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	\dots	S_N
Data items	D_1				
	D_2				
	D_3				
	\dots				
	D_M				

Output

		Truths			
		D_1	D_2	\dots	D_M
Data items	D_1				
	D_2				
	D_3				
	\dots				
	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



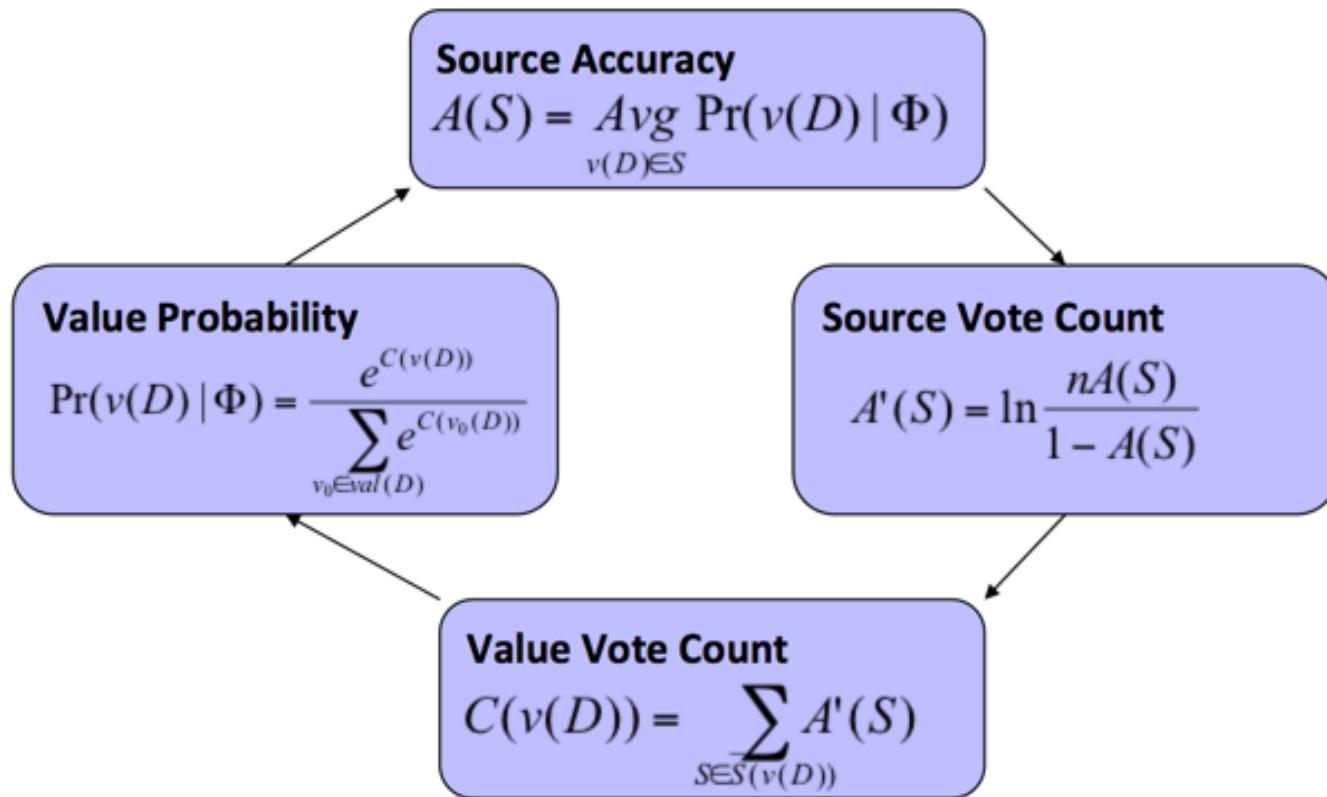
	Src1	Src2	Src3
Jagadish	UM	ATT	UM
Dewitt	MSR	MSR	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

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

Data Fusion–A Bayesian Model

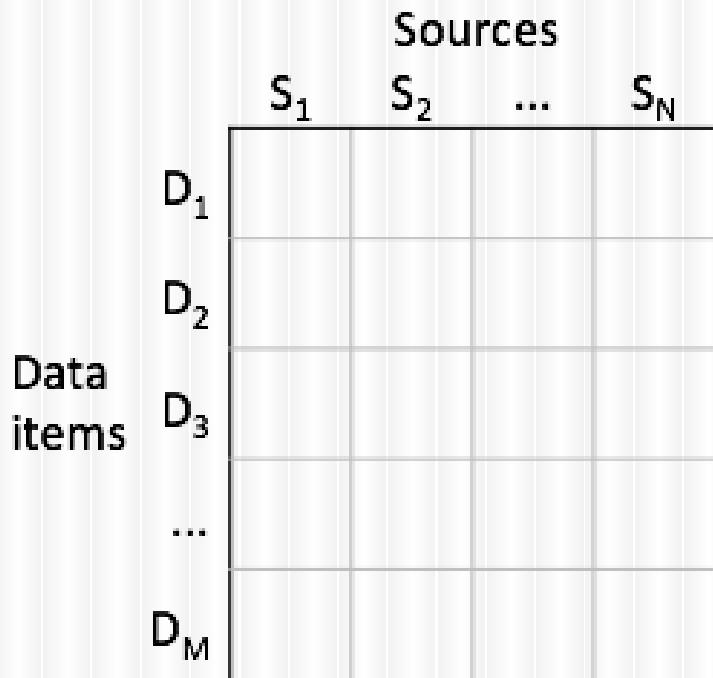
[Dong et al., VLDB'09]

- ◆ Continue until source accuracy converges

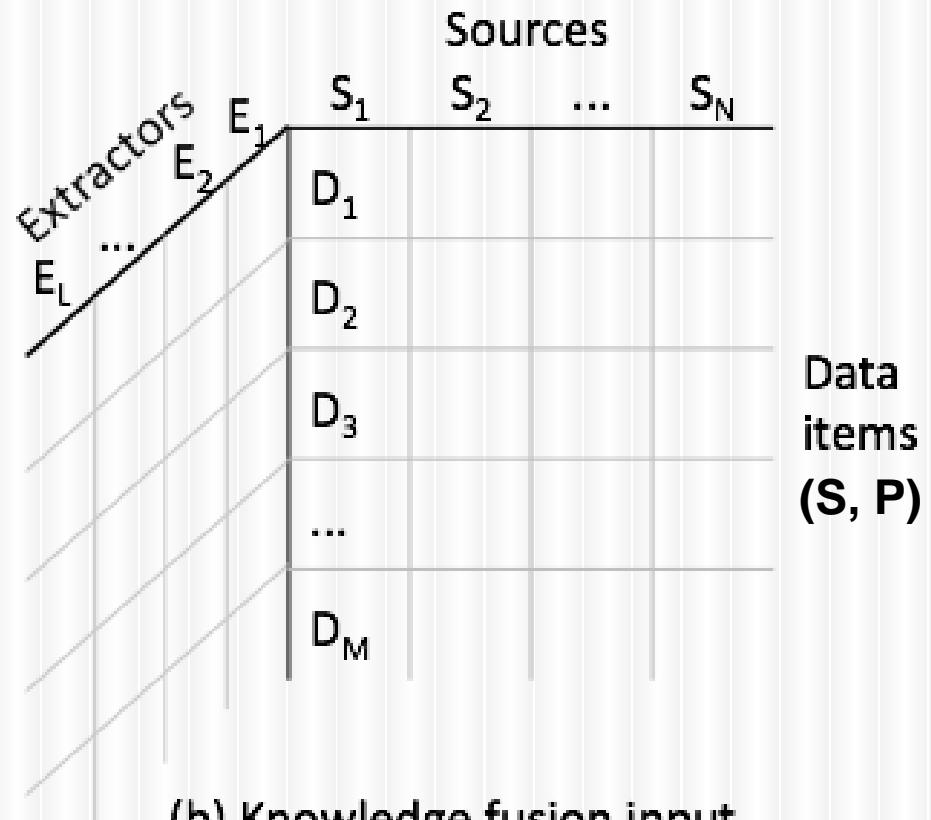


Knowledge Fusion Challenges

I. Input is *three-dimensional*



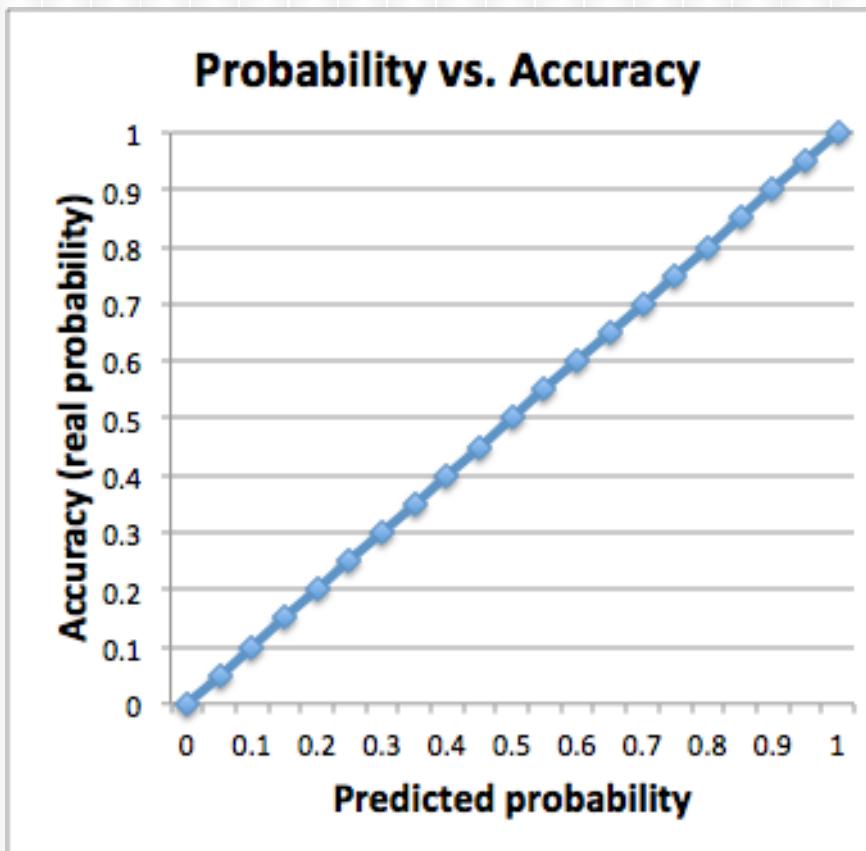
(a) Data fusion input



(b) Knowledge fusion input

Knowledge Fusion Challenges

II. Output prs should be *well-calibrated*



Knowledge Fusion Challenges

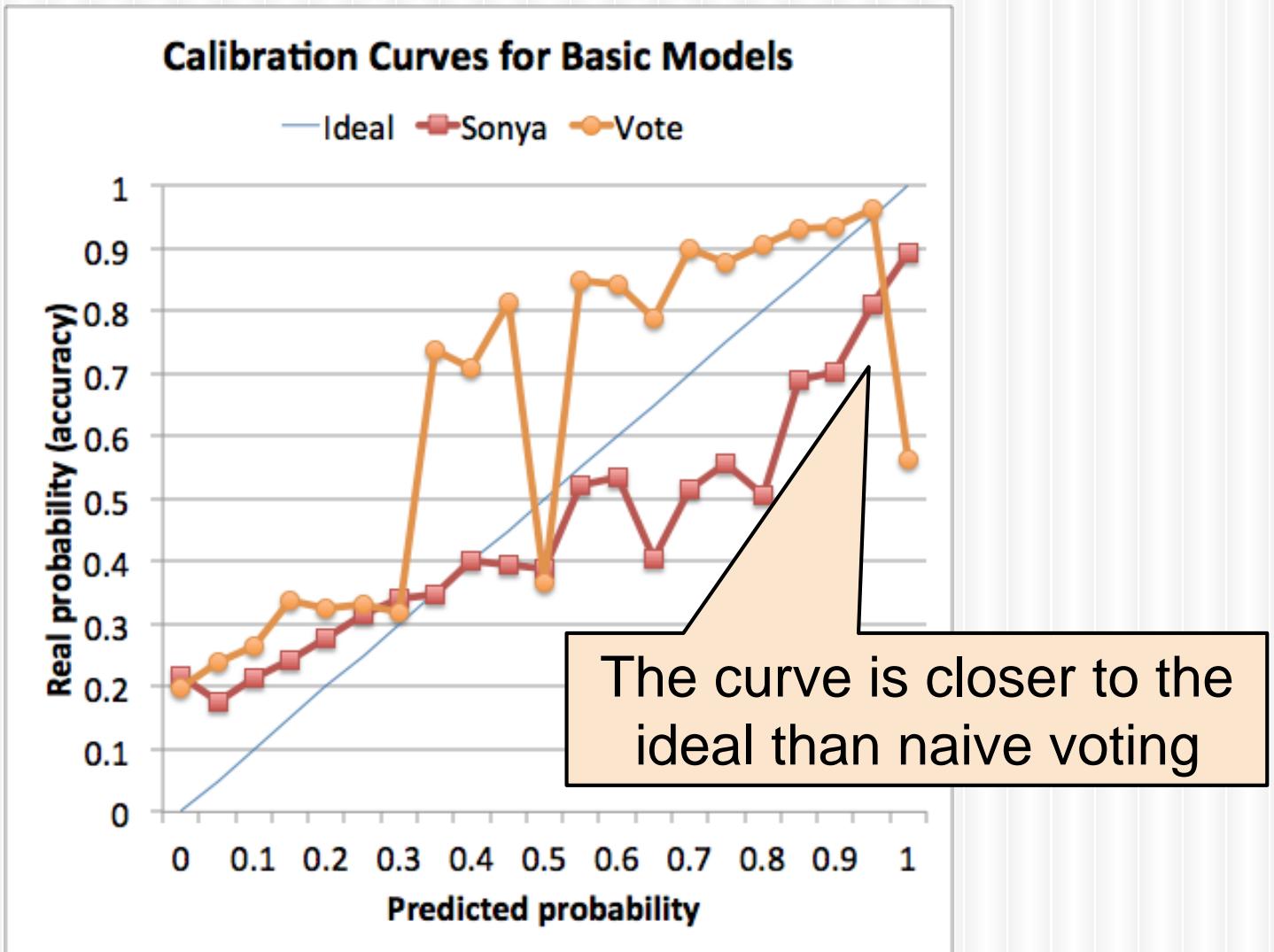
III. Data are of *Web-scale*

- 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
 - Sample for *too big* data items or provenances

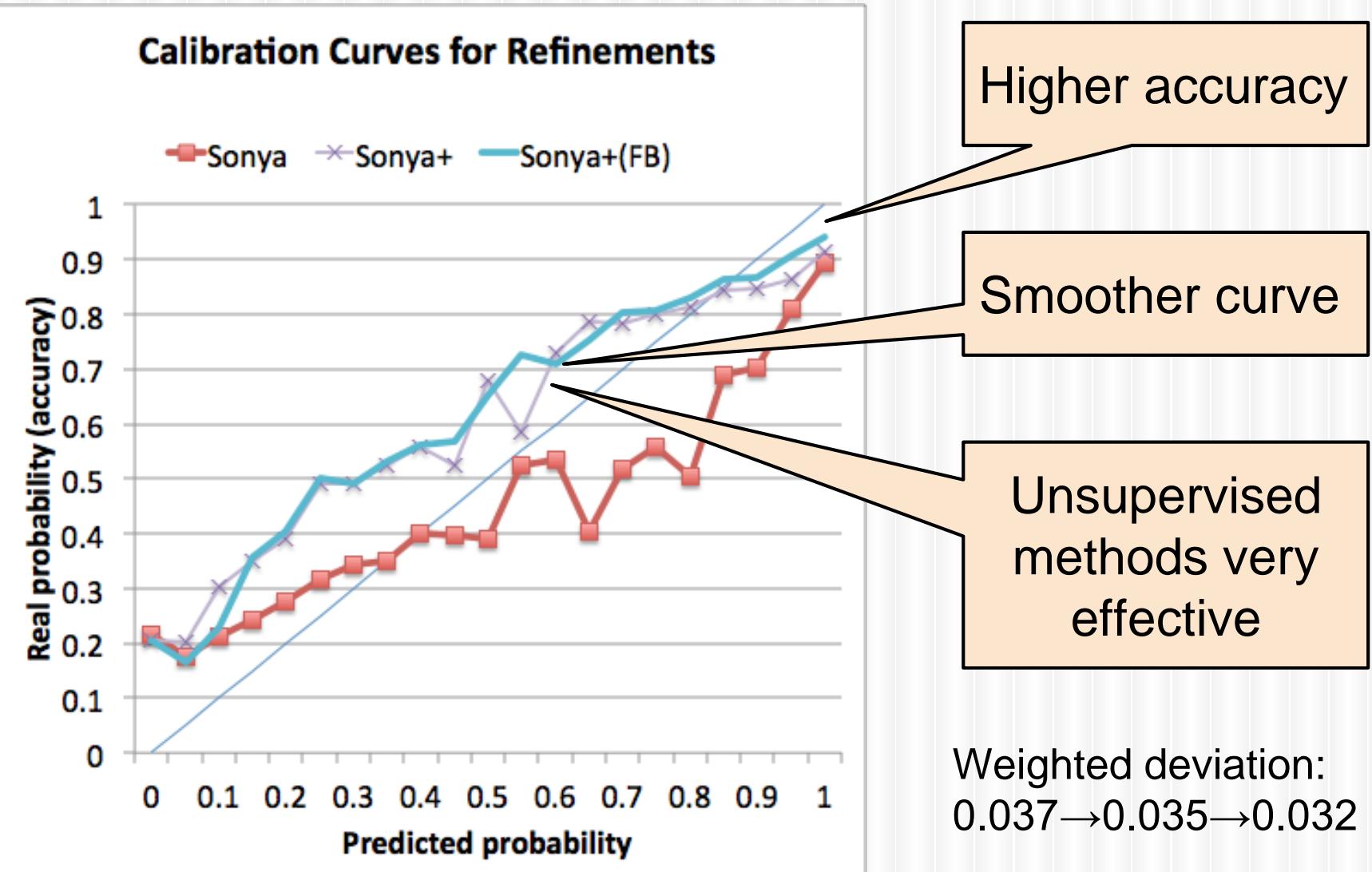
Basic Sonya Solution vs. Voting



Refinements

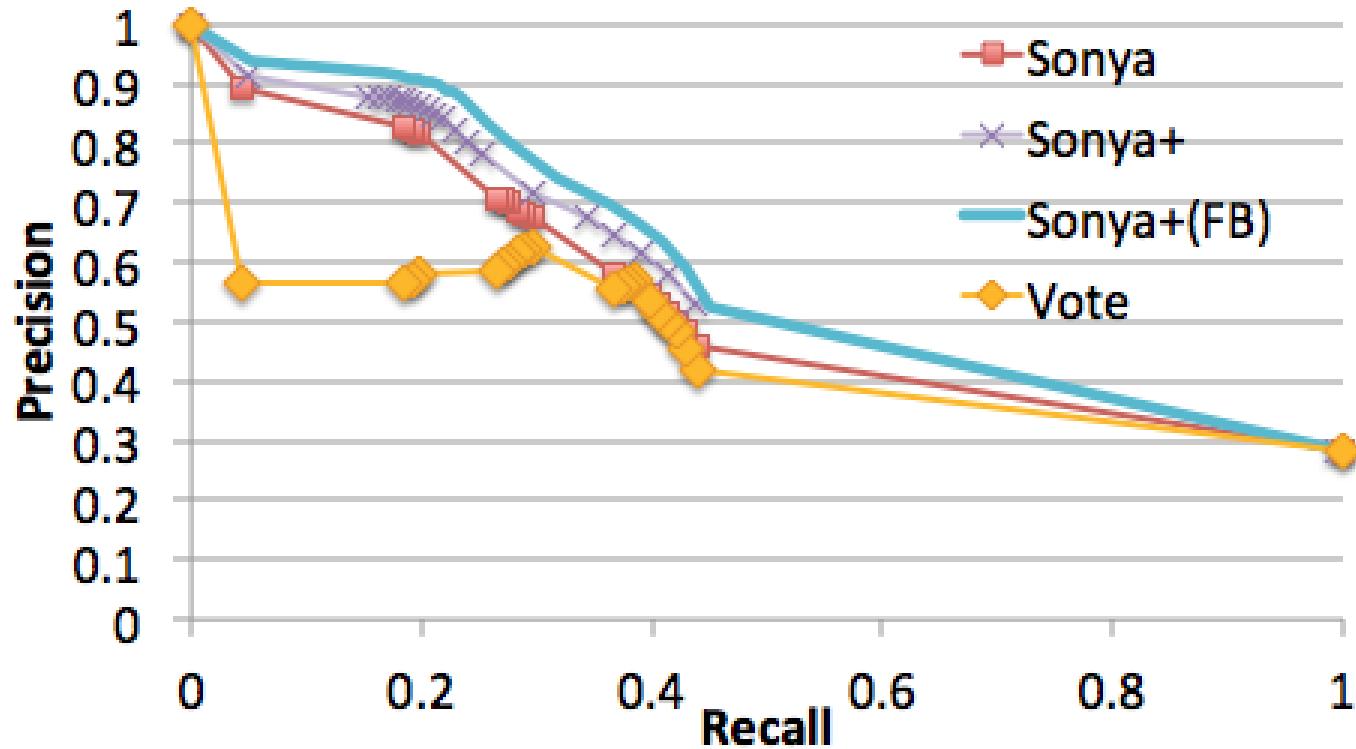
- I. Ignore low-coverage provenances
 - II. Granularity (URL->Site,
Extractor->Pattern, Predicate)
 - III. Ignore low-accuracy provenances
 - IV. Initiate provenance accuracy by FB
-
- +I, II, III. Sonya+ : unsupervised
 - +IV. Sonya+(FB) : semi-supervised

Calibration Curve



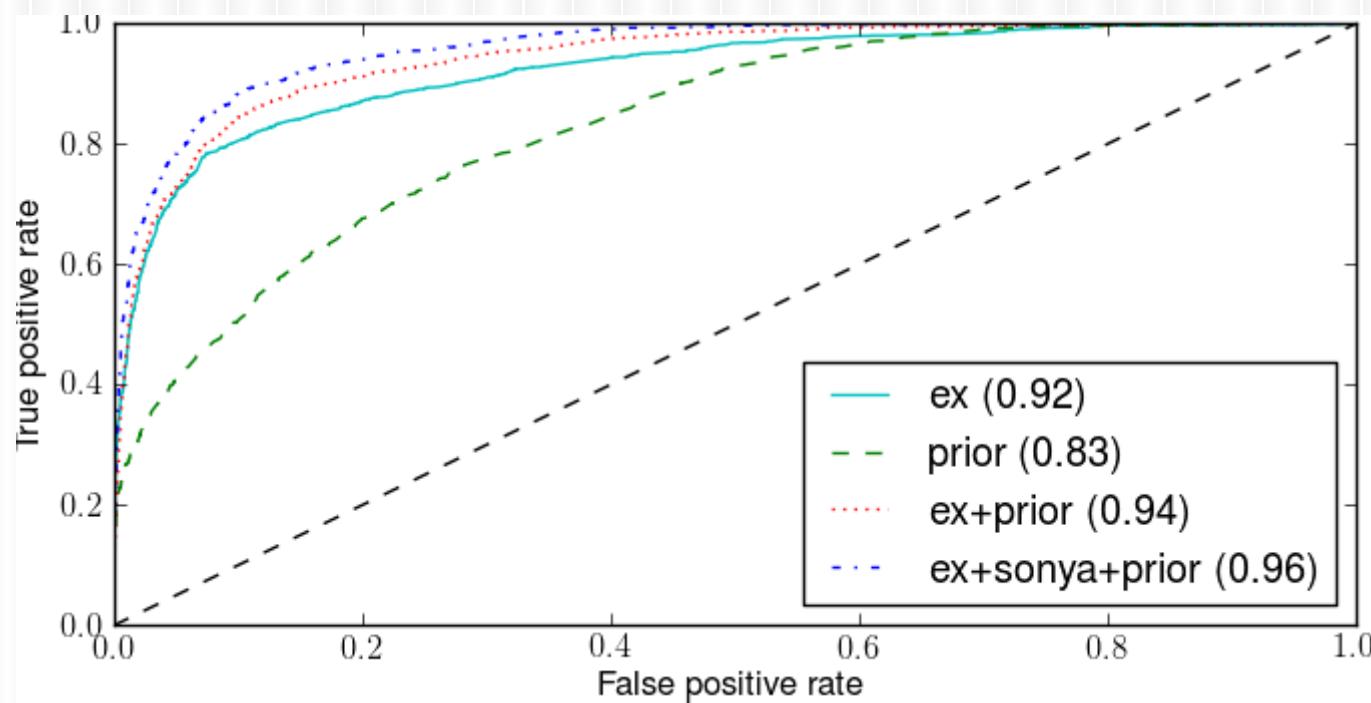
Precision-Recall Curve

PR-Curves for Various Models



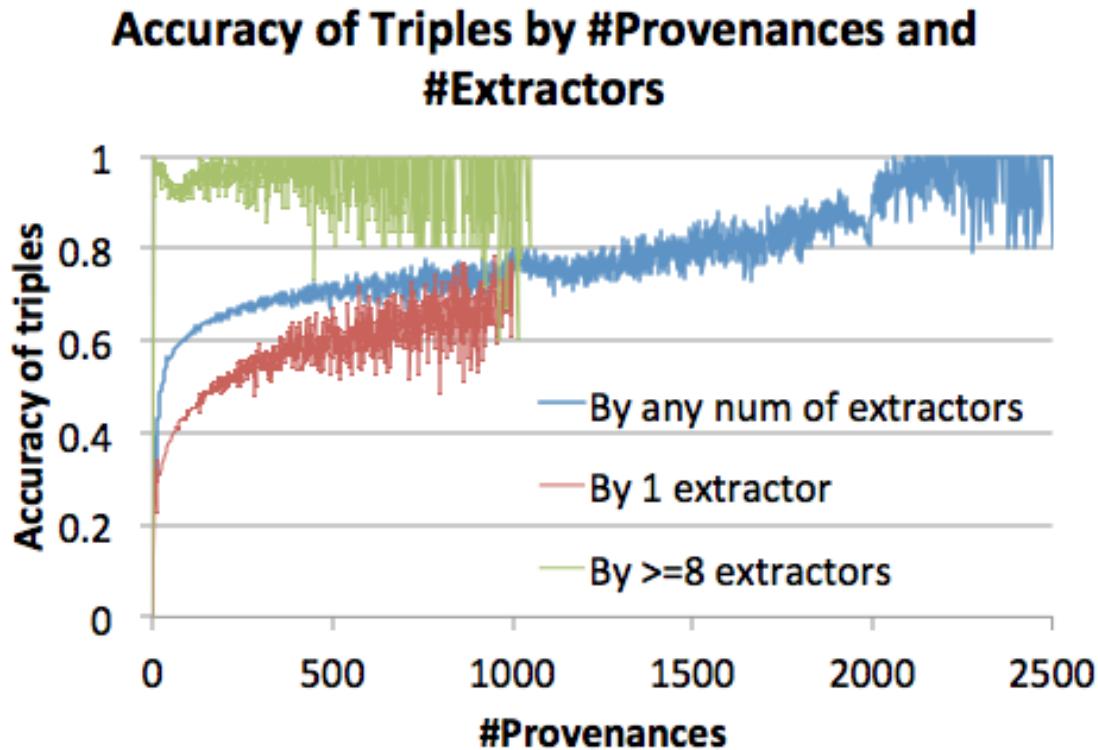
Other Fusion Techniques

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



One Inherent Limitation

Cannot distinguish errors from extractor and from sources

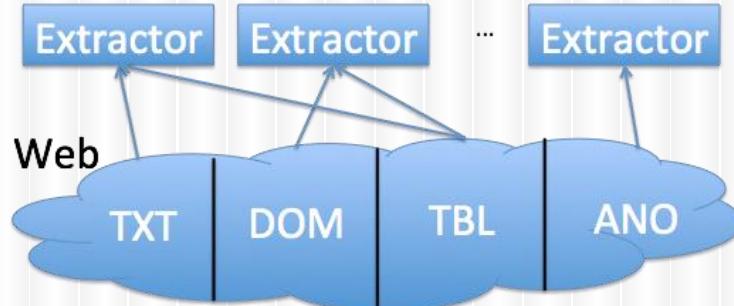


Outline



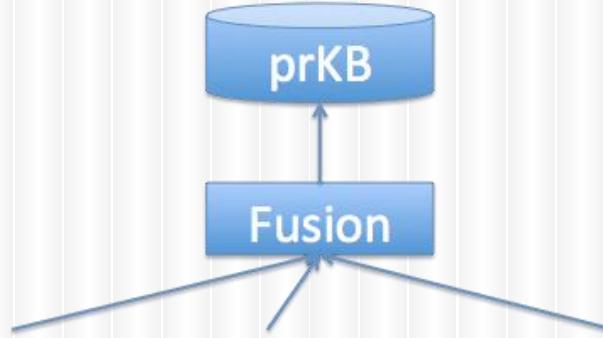
KNOWLEDGE
VAULT

Knowledge extraction



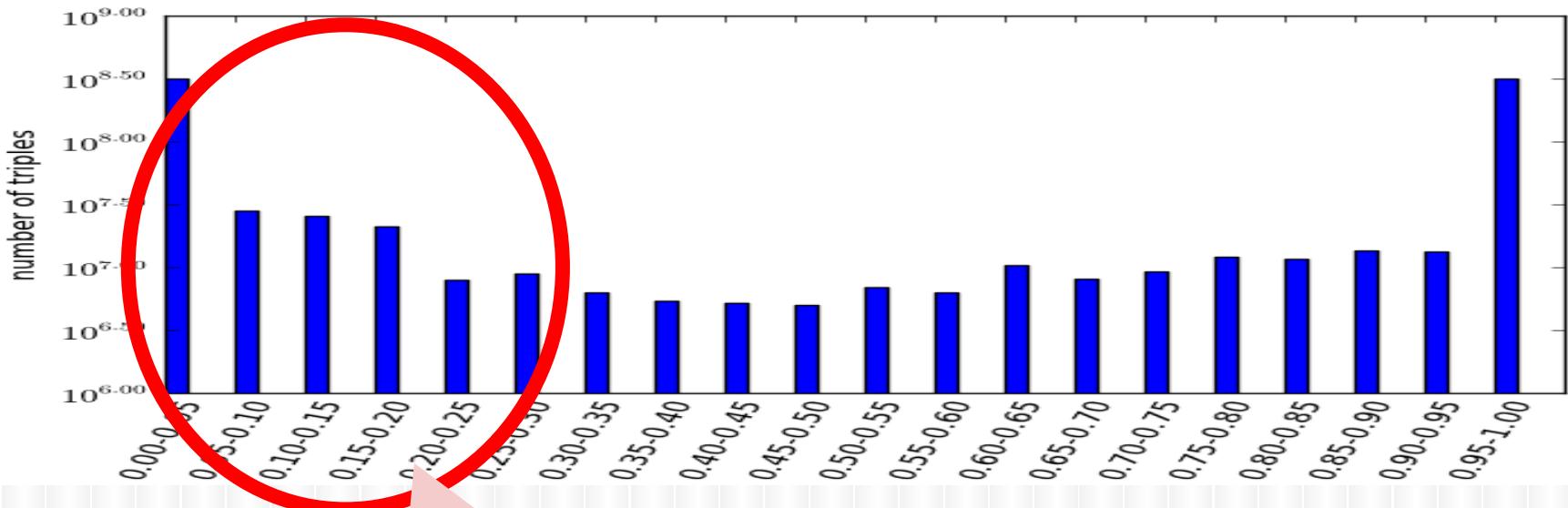
KNOWLEDGE
VAULT

Knowledge fusion



Interesting applications

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

Browse Salaries Follow Us @Salary RSS Feed Podcast Facebook LinkedIn YouTube

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

I am an Employee or Individual

I am an Employer or Business

Help Me Negotiate Contemplating A Move Job Search Salary Information Job-Specific Competencies Compensation Subscriptions

Job or Employee Salary Reports 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 over the past 12 months
 Once or twice every few months
 Once or twice a month
 Once or twice a week
 Almost every day at work

Submit

KRAFT SINGLES {now have} NO ARTIFICIAL PRESERVATIVES discover more

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
- Northern Ireland
- Scotland
- Scotland Central
- Scotland North
- Scotland South
- South East England

Swim With Dolphins dolphincounter... 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...

WOYLAAC All Korean drama episodes english subtitle and RAW

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

RECENT DRAMAS

- BIG MAN EPISODE 2 ENG SUB
- EMPERESS KI EPISODE 51 ENG SUB
- BIG MAN EPISODE 1 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

POPULAR DRAMAS

- EMPERESS KI EPISODE 50 ENG SUB
- EMPERESS 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

Guitar Backing Tracks 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 Airline Fee Credits annually • PLUS, EARN 40,000 POINTS •



Tale Sources w. Low Page Rank May Provide Valuable Info

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/w/index.php?title=Boulevard_of_Broken_Dreams_\(Green_Day_song\)&oldid=9000000](https://en.wikipedia.org/w/index.php?title=Boulevard_of_Broken_Dreams_(Green_Day_song)&oldid=9000000) ▾ 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

who play The Stumble

Web Videos Images News Maps More Search tools

About 263,000,000 results (0.22 seconds)

Peter Green-The Stumble DVD Guitar Lesson - YouTube
 www.youtube.com/watch?v=... ▾ YouTube ▾
Apr 26, 2011 - Uploaded by note4noteus
http://www.note4note.us Peter Green - The Stumble Sample from our Note For Note DVD ... You need Adobe ...

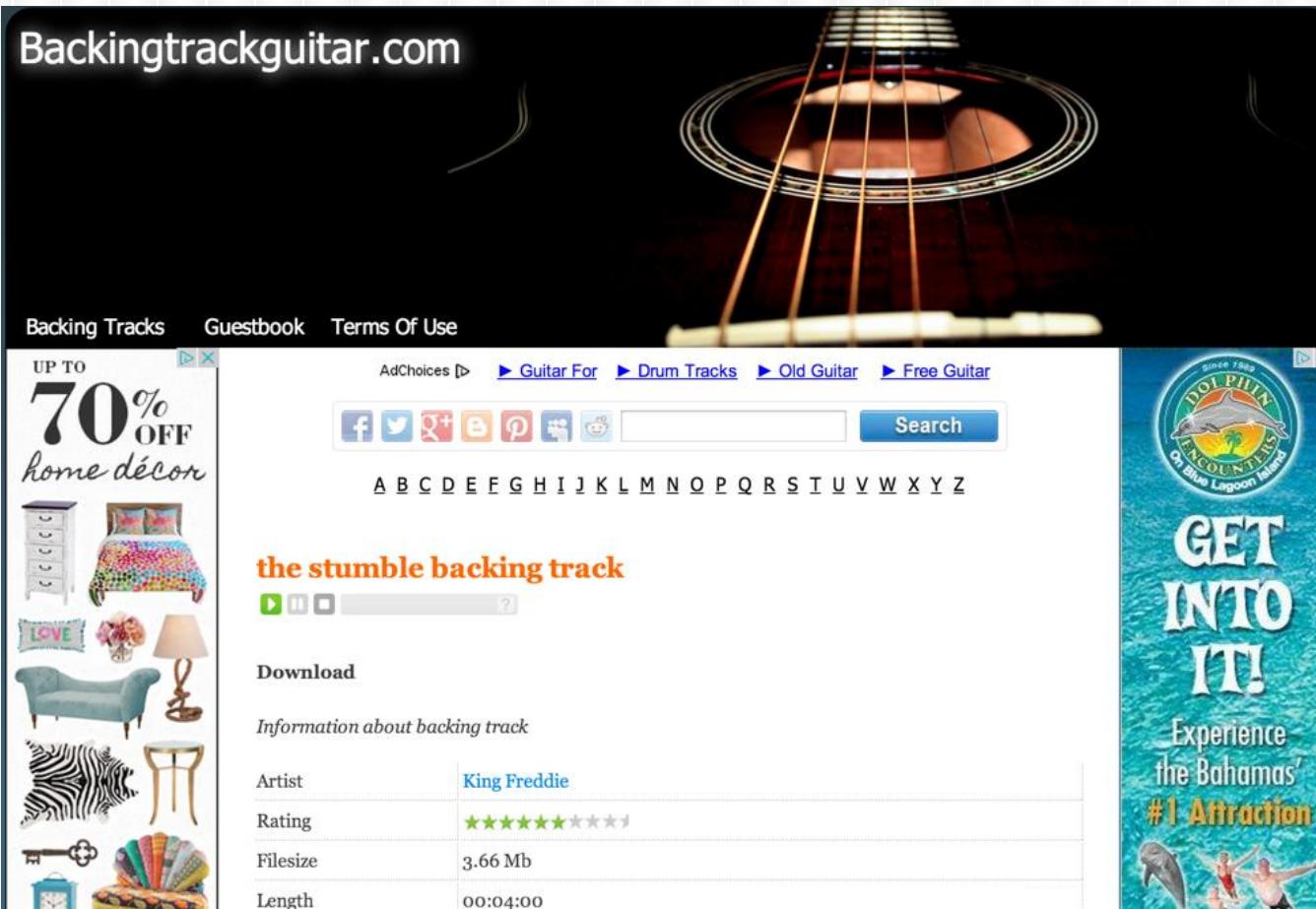
Peter Green Amsterdam 2009 The Stumble - YouTube
 www.youtube.com/watch?v=6TPRGY... ▾ YouTube ▾
Feb 27, 2009 - Uploaded by harpslide
You need Adobe Flash Player to watch this video. ... 0:43 How to play the double stops in The Stumble by ...

Gary Moore - The Stumble Guitar Lesson (Pt1) - YouTube
 www.youtube.com/watch?v=... ▾ YouTube ▾
Oct 24, 2011 - Uploaded by KowboyCa
You need Adobe Flash Player to watch this video. ... 0:43 How to play the double stops in The Stumble by ...

Mick Taylor guitar lesson The Stumble closeup & slowdown ...
 www.youtube.com/watch?v=... ▾ YouTube ▾
Mar 4, 2013 - Uploaded by tokabillitor
You need Adobe Flash Player to watch this video. Download it from Adobe. Mick Taylor guitar lesson The ...

Missing WebAnswer for a not-so-popular song

Tale Sources w. Low Page Rank May Provide Valuable Info



The screenshot shows the homepage of Backingtrackguitar.com. At the top, there's a large image of a guitar. Below it, the website's navigation menu includes "Backing Tracks", "Guestbook", and "Terms Of Use". A sidebar on the left offers a 70% discount on home decor. The main content area features a link to "the stumble backing track" and a "Download" button. To the right, there's a search bar and a sidebar for "Dolphin Encounters".

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 g+ b p 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

DOLPHIN ENCOUNTERS Since 1982 On Blue Lagoon Island

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



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

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

Red arrow pointing from the Wikipedia page to the evaluation table.

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



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

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

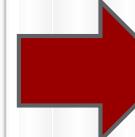
Motto:
"In God we trust" (official)^{[1][2][3]}
"E pluribus unum" (Latin) (traditional)
"Out of many, one"

Anthem: "The Star-Spangled Banner"

 0:00   MENU



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



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

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



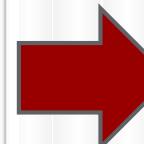
Flag Great Seal

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

Anthem: "The Star-Spangled Banner"

 0:00  





	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	

Sonya Trustworthiness Score

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

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

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: (besoccer.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

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?

Data Fusion–A Bayesian Model

[Dong et al., VLDB'09]

Q1. How to compute source accuracy?

- Source Accuracy: $A(S)$

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

- $\bar{V}(S)$ - 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 $\Pr(\Phi | v_i \text{ true})$

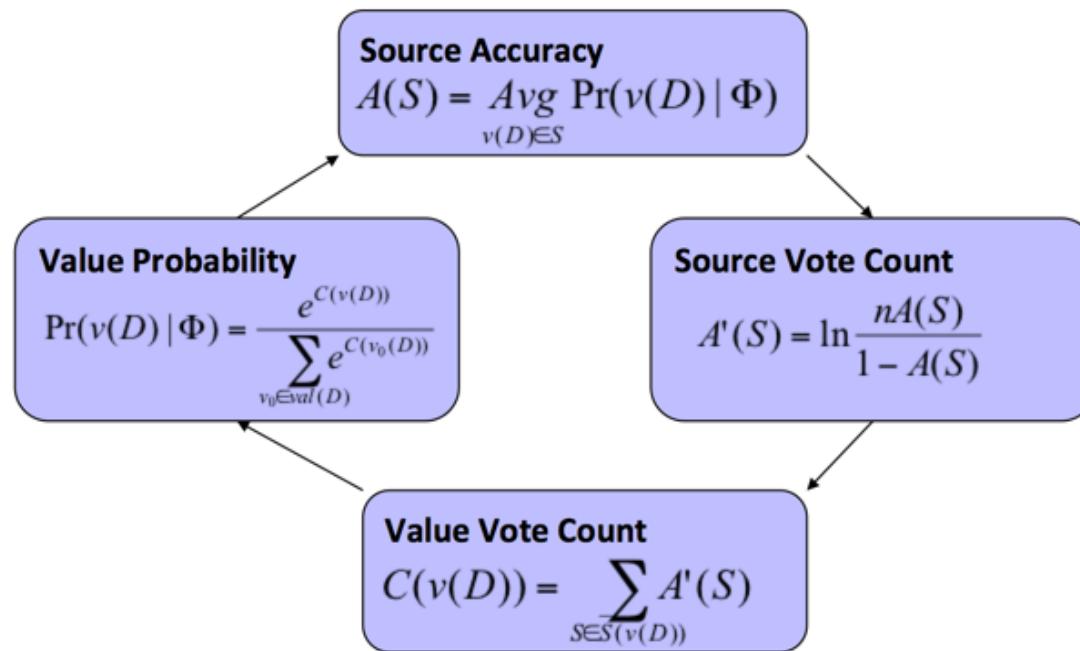
- Assuming independence of sources, we need to know $\Pr(\Phi(S) | v_i \text{ true})$
- If S provides v_i :
 $\Pr(\Phi(S) | v_i \text{ true}) = A(S)$
- If S does not provide v_i :
 $\Pr(\Phi(S) | v_i \text{ true}) = (1 - A(S)) / n$

Data Fusion–A Bayesian Model

[Dong et al., VLDB'09]

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

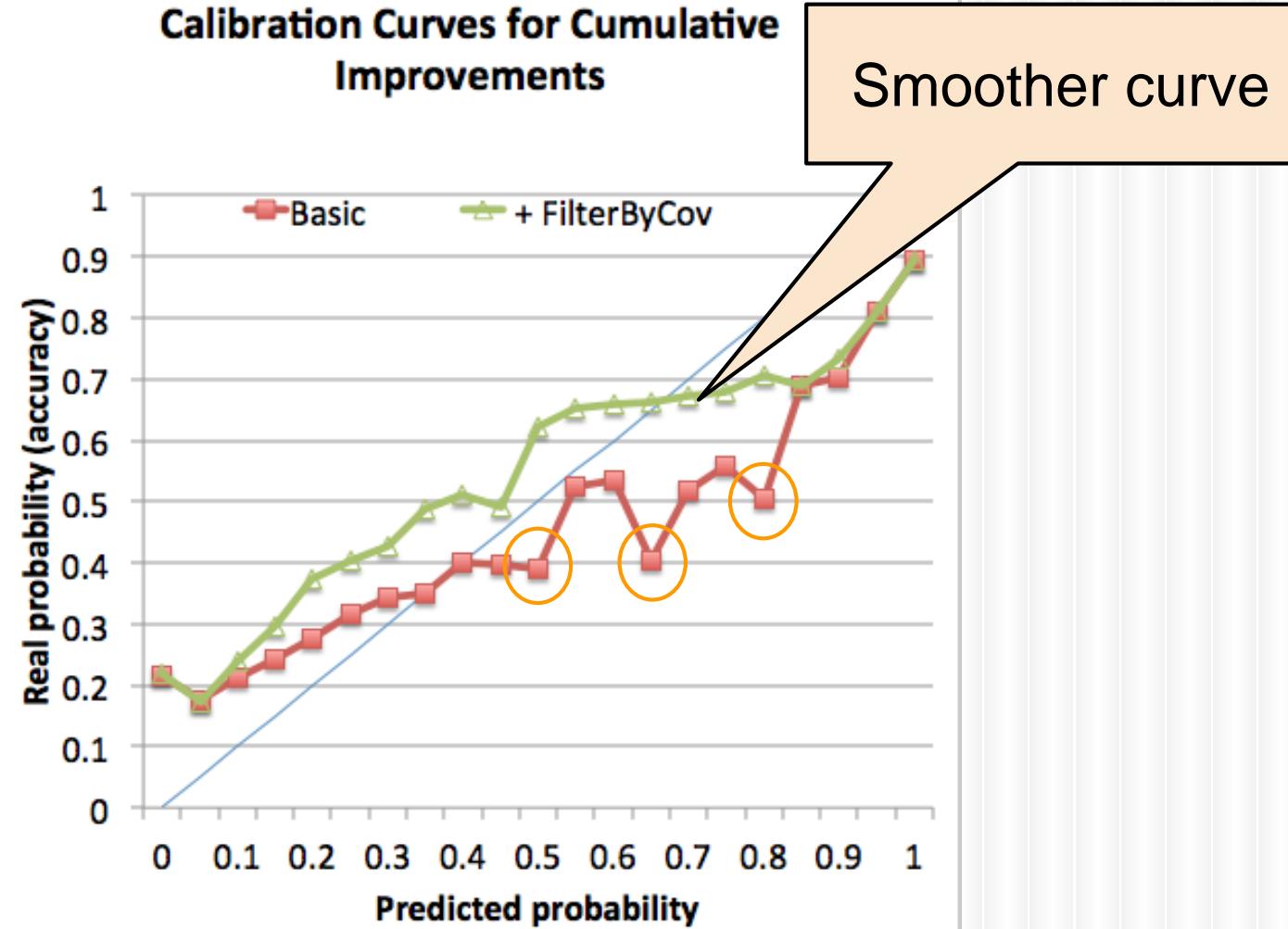
- ◆ Continue until source accuracy converges



Knowledge Fusion Solutions

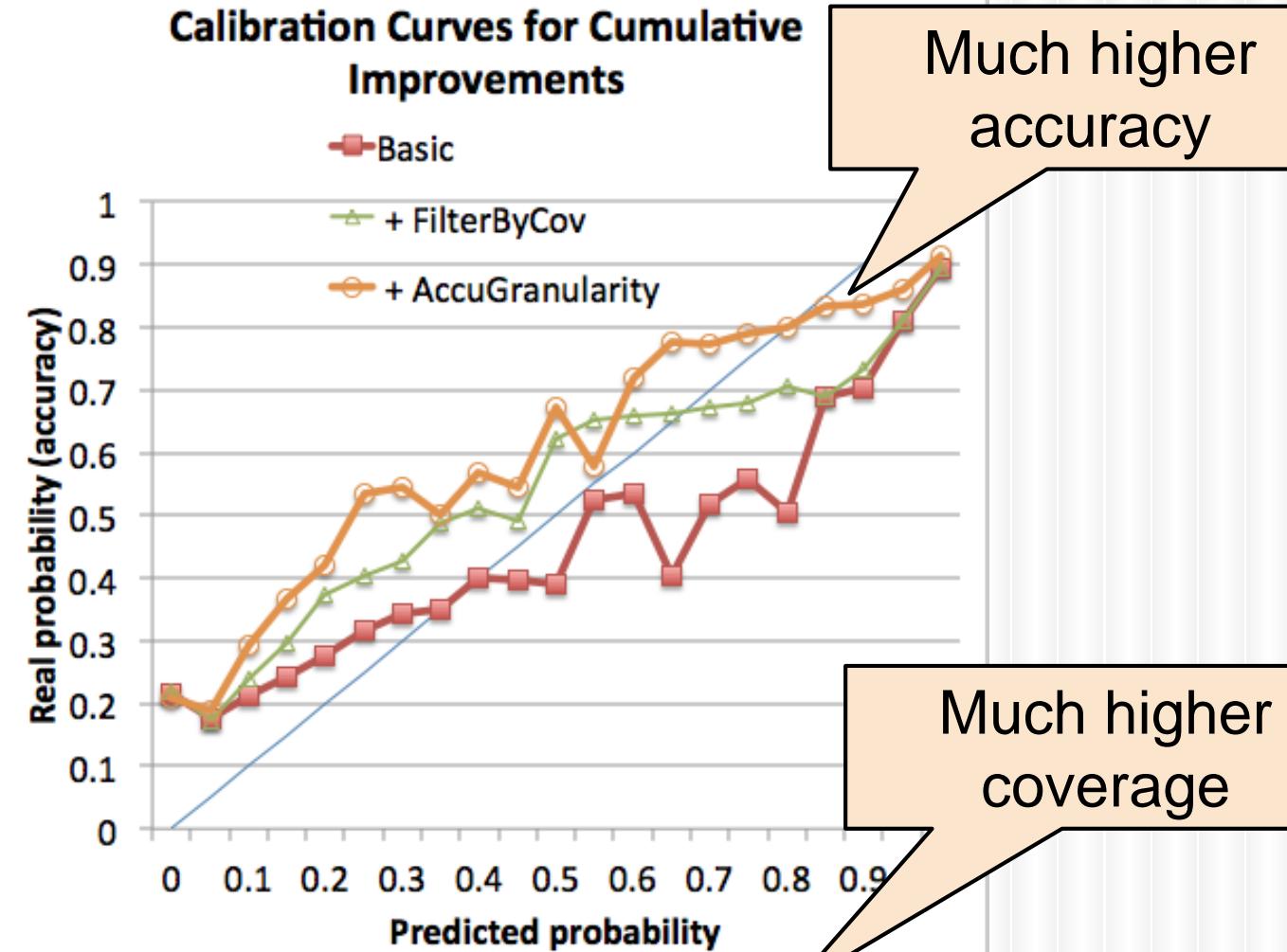
- Treat each (URL, Extractor) as a whole (*provenance*) for accuracy evaluation
- A series of refinements to improve probability calibration
- MapReduce Based Framework
 - Terminate in 5 rounds
 - Sample for *too big* data items or provenances

Refinement I. Ignore Low-Coverage Provenances

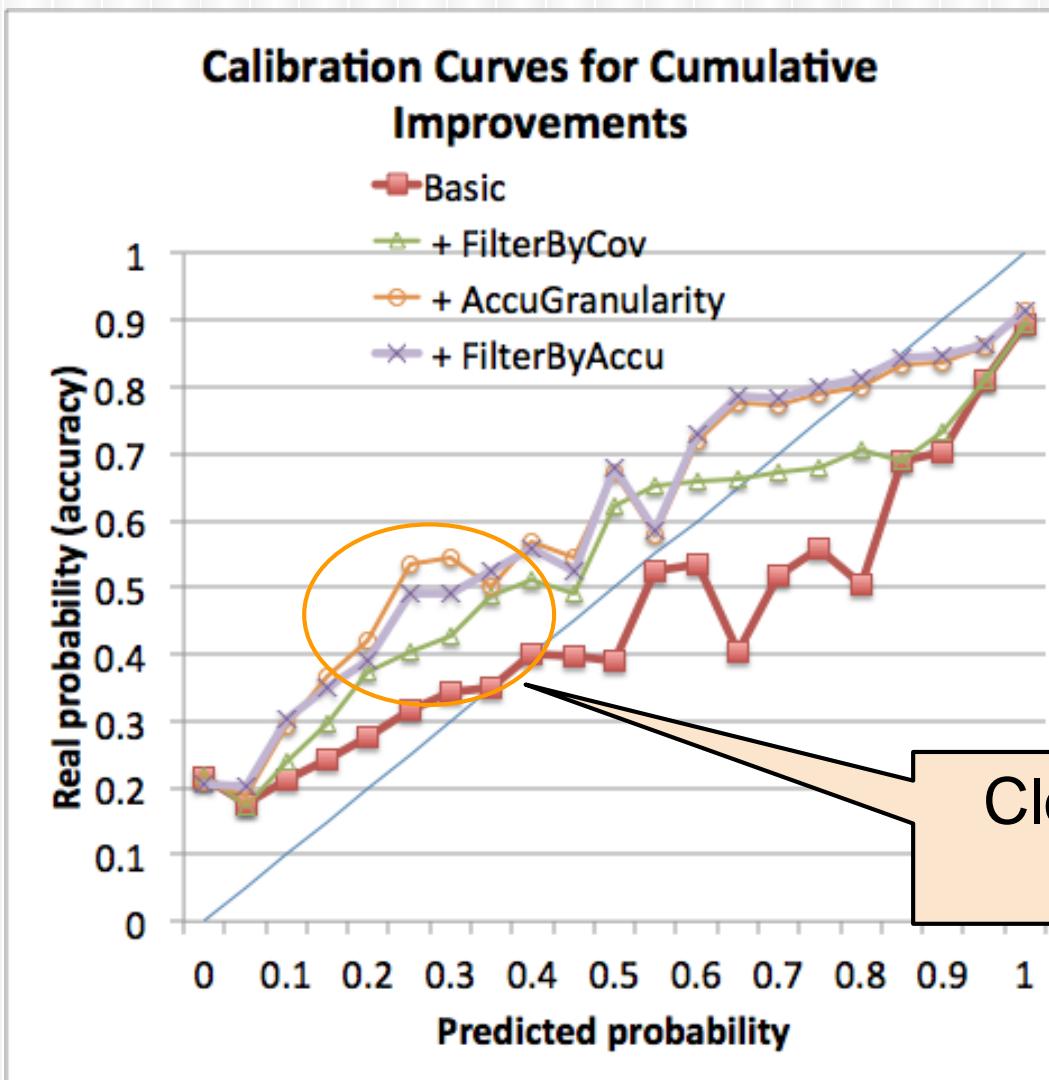


Coverage: 1 → .918

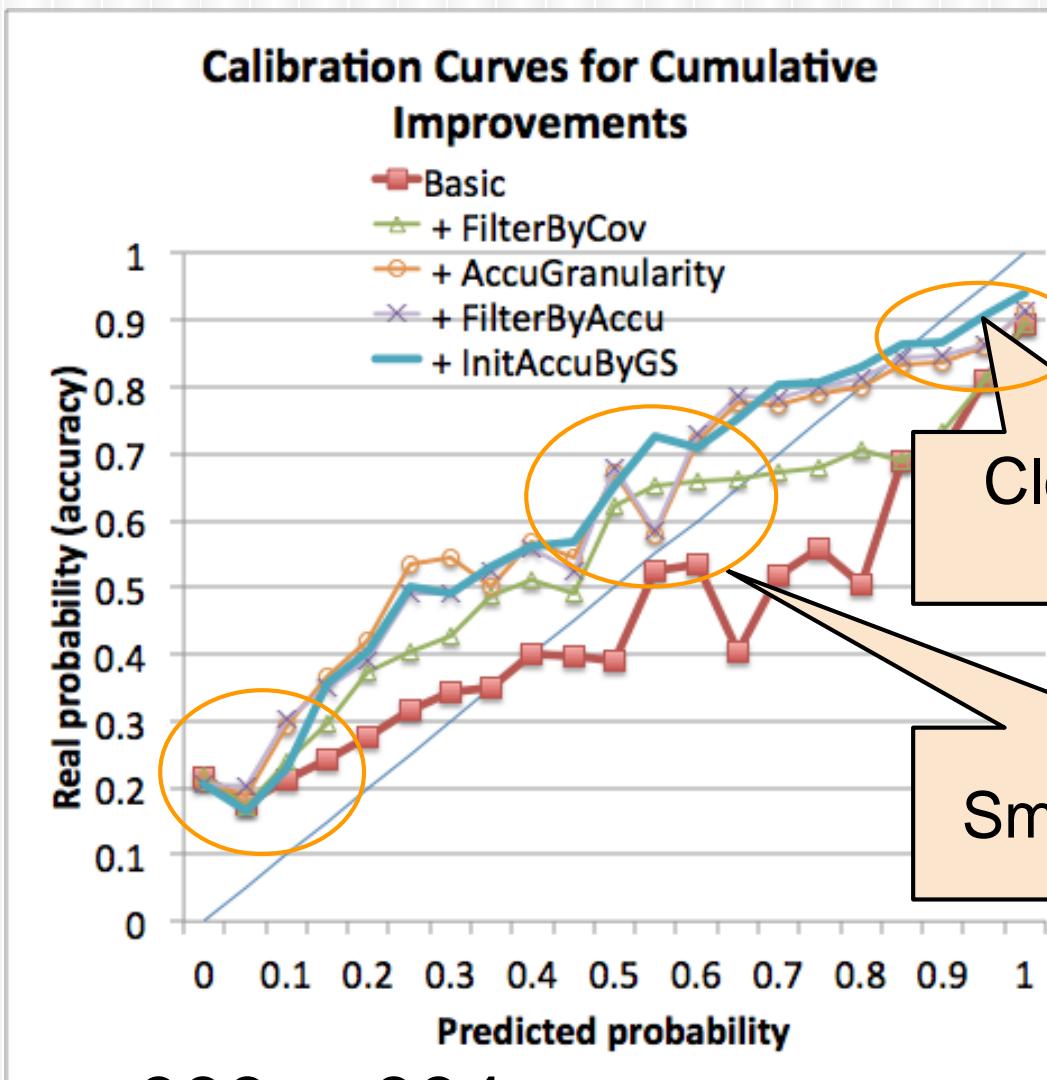
Refinement II. Granularity (URL->Site, Extractor->Pattern, Predicate)



Refinement III. Ignore Low-Accuracy Provenances

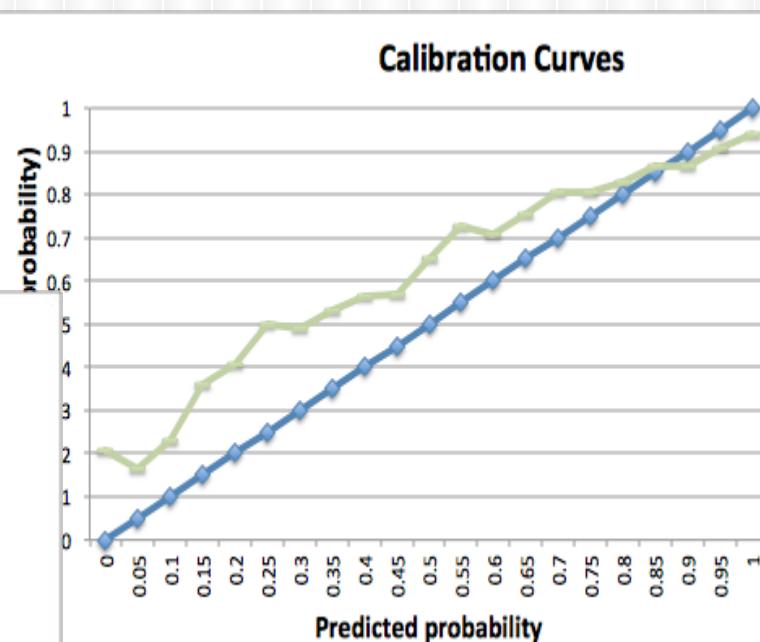


Refinement IV. Initiate Provenance Accuracy by FB



Analysis of Errors

False Negatives



False Positives



Future Directions!!!

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

Result I. Extraction and Triple Correctness

Example 1. (Obama, nationality, ?)

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

- 3 extractions
(Pr_provide=0.01)
<http://mathaba.net/news/?x=631316>
<http://www.laht.com/article.asp?ArticleId=329187&CategoryId=10717>
<http://www.iamicas.org/en/about-ioa/presidents-corner>
- Pr_true=0

Result I. Extraction and Triple Correctness

Example 1. (Obama, nationality, ?)

(Obama, nationality, Kenya)

- 2087 extractions:
 - Example of a correct extraction ($\text{Pr_provide}=0.792$):
<http://beforeitsnews.com/obama-birthplace-controversy/2013/04/alabama-supreme-court-chief-justice-roy-moore-to-preside-over-obama-eligibility-case-2458624.html>
 - Example of a wrong extraction
($\text{Pr_provide}=0.130$): <http://www.monitor.co.ug/News/National/US+will+respect+winner+of+Kenya+election++Obama+says/-/688334/1685814/-/ksxagx/-/index.html>
- $\text{Pr_true}=0$ (not enough support)

Result I. Extraction and Triple Correctness

Example 1. (Obama, nationality, ?)

(Obama, nationality, USA)

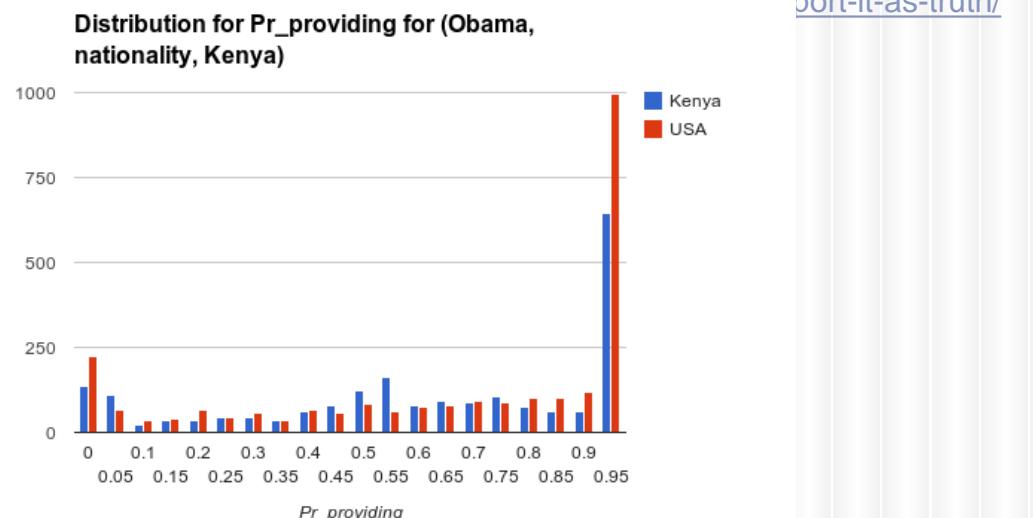
- 2481 extractions:

- Example of a correct extraction ($\text{Pr_provide}=0.999$):

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

- Example of a wrong extraction
($\text{Pr_provide}=0.261$): <http://blogs.telegraph.co.uk/news/timstanley/100169248/barack-obamas-life-story-co>

- $\text{Pr_true}=1$



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: <http://ibirthdayworld.blogspot.com/2010/03/celebrity-birthdays-on-march-22.html> e.g., http://en.wikipedia.org/wiki/World_Chess_Championship_2013 or en.wikipedia.org

Pred: /people/person/date_of_birth 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

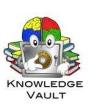
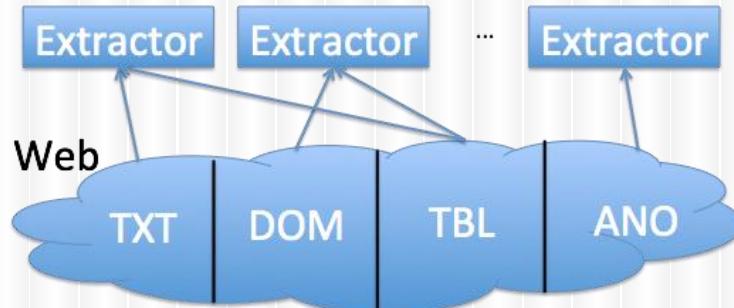
Siblings: Laurence Hyde, 1st Earl of Rochester, Henry Hyde, 2nd Earl of Clarendon



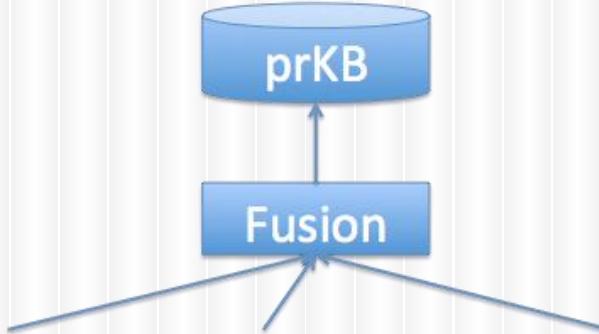
Outline



Knowledge extraction



Knowledge fusion



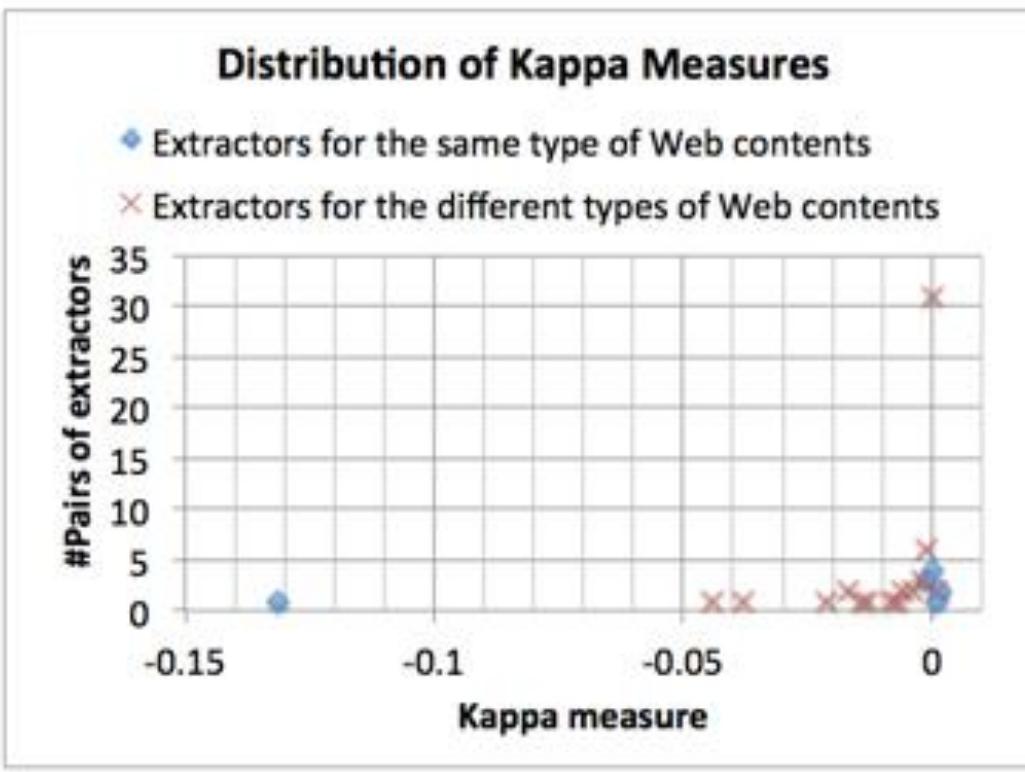
Interesting applications



[. Future directions

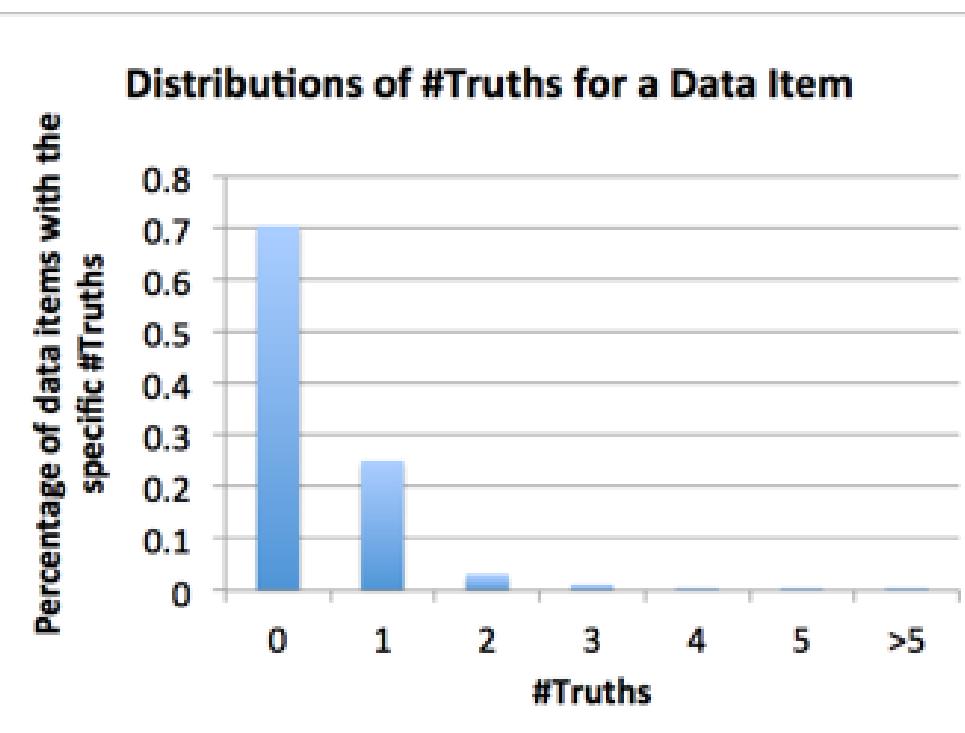
Future Directions: Remove the Assumptions One by One

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



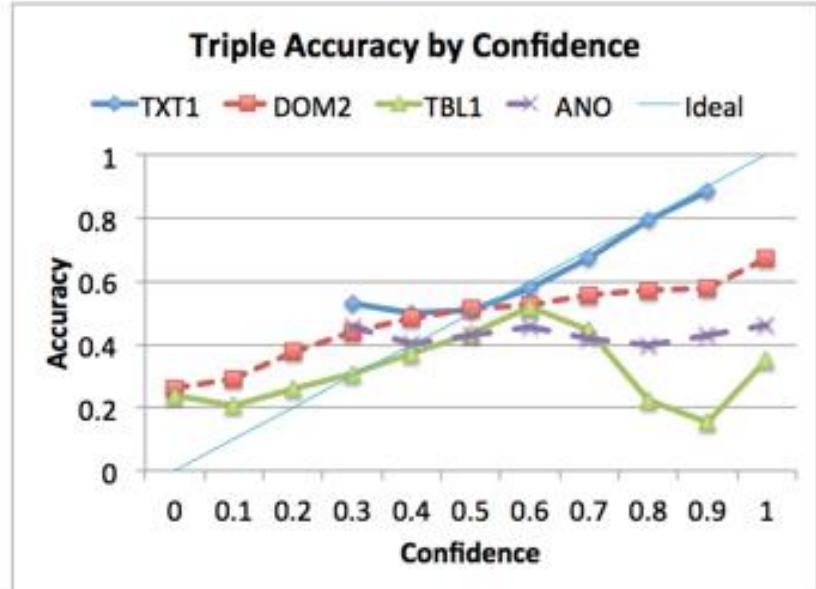
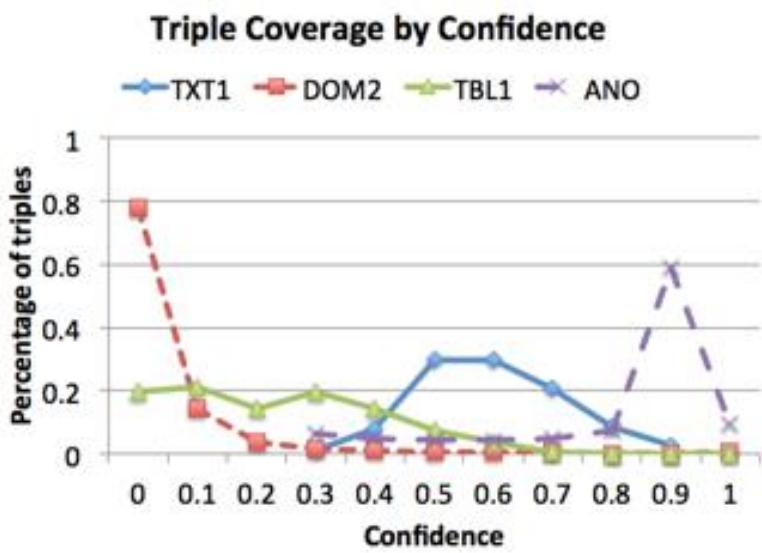
Future Directions: Remove the Assumptions One by One

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!!!

