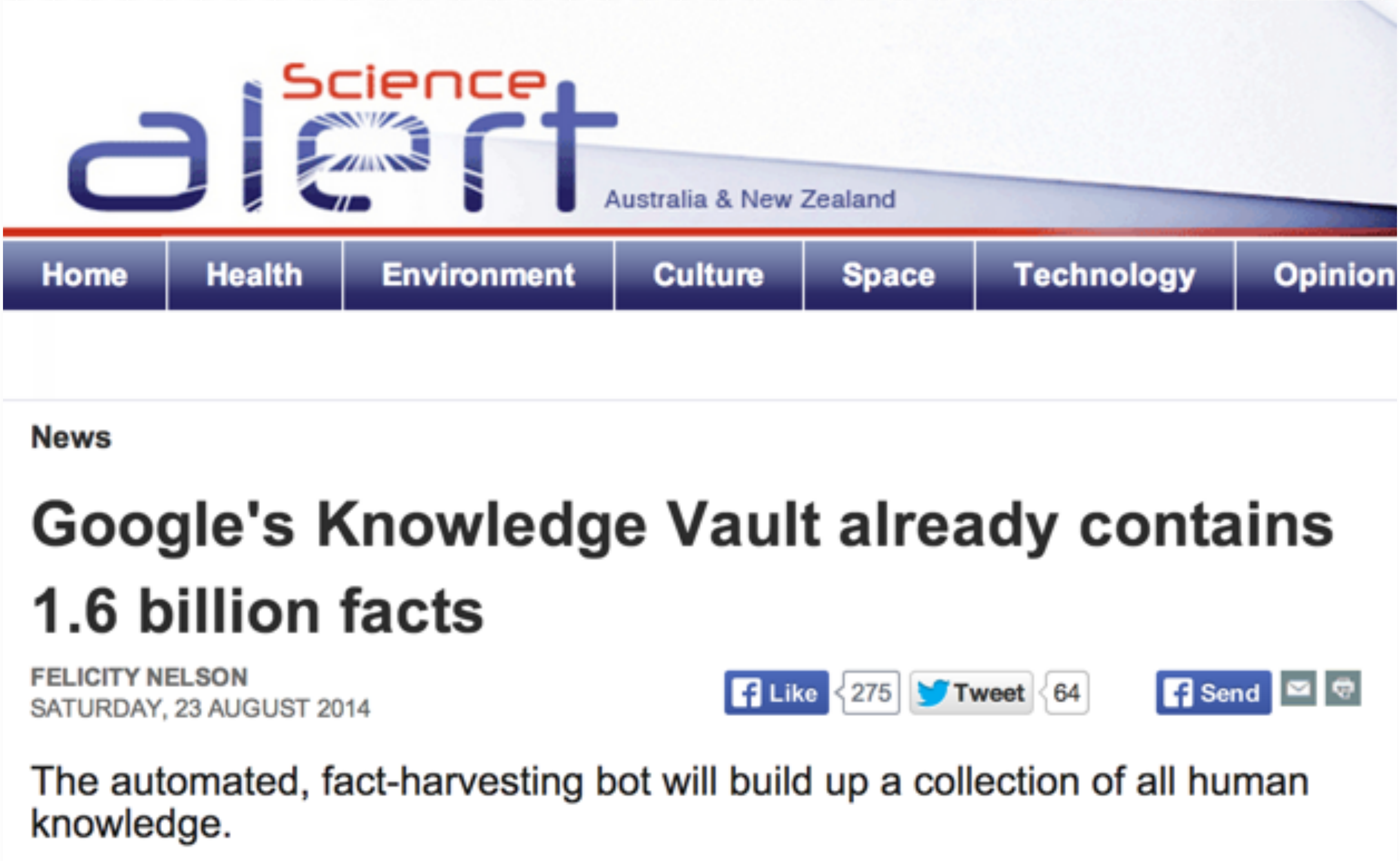
From Data Fusion to Knowledge Fusion

Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn,
Kevin Murphy, Shaohua Sun, Wei Zhang

Google Inc.

9/2/2014 @ *VLDB'14*



News

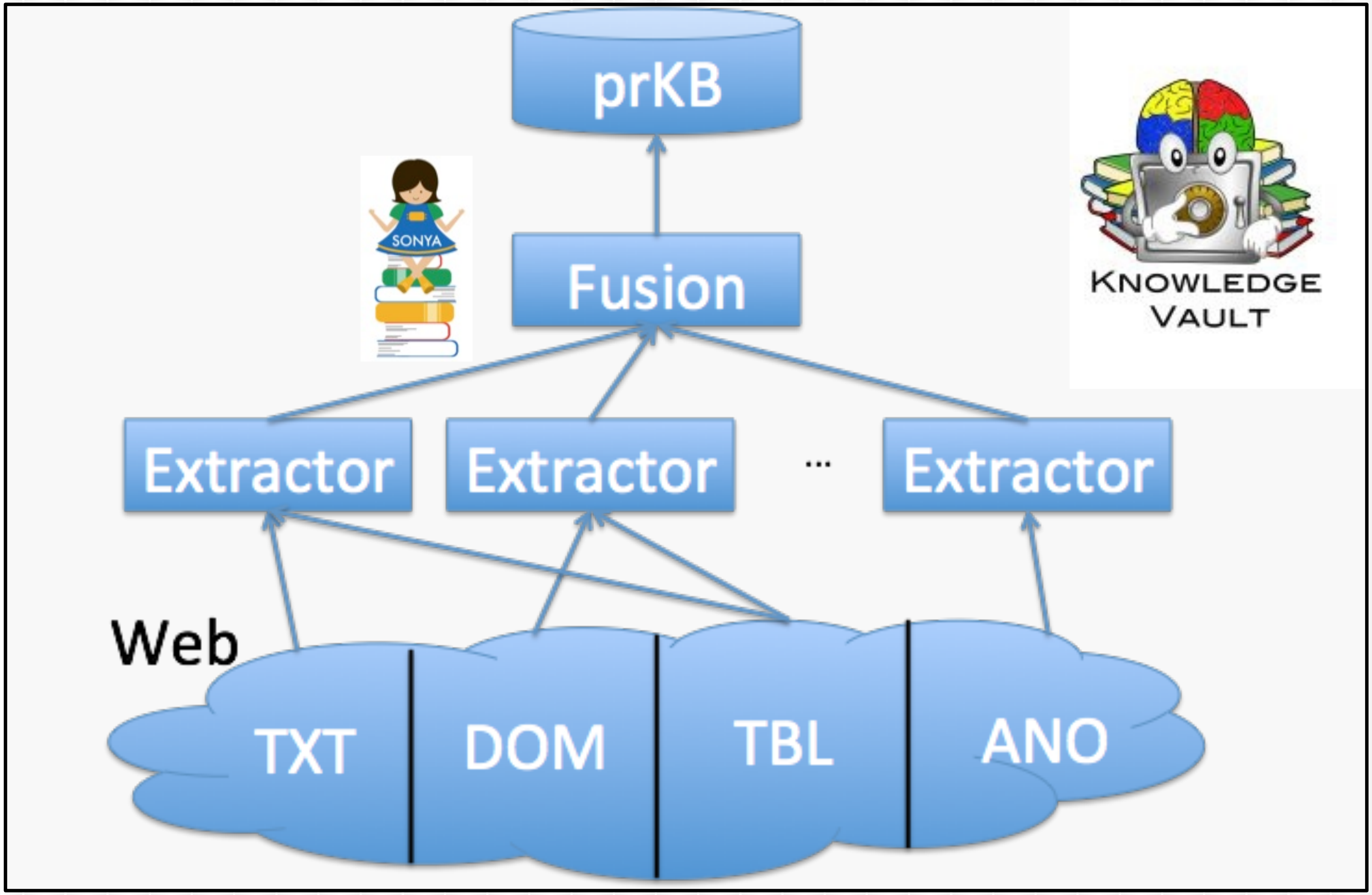
Google's Knowledge Vault already contains 1.6 billion facts

FELICITY NELSON
SATURDAY, 23 AUGUST 2014

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

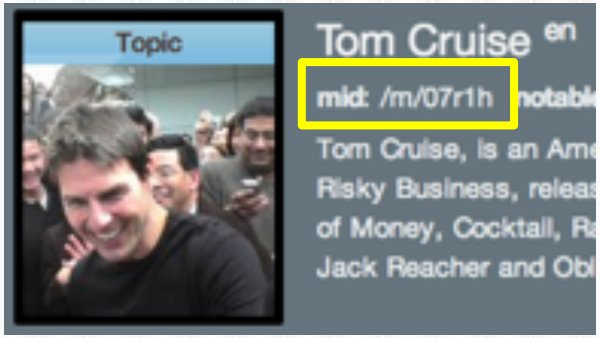
Knowledge Vault– Building a Probabilistic KB

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



Extracted 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.



Tom Cruise en

mid: /m/07r1h notabl

Tom Cruise, is an Ame
Risky Business, releas
of Money, Cocktail, Ra
Jack Reacher and Obl

Observation I.

Yes, We CAN Build A Large KB from the Web! :-)

Statistics for Extracted Triples

- A large knowledge base

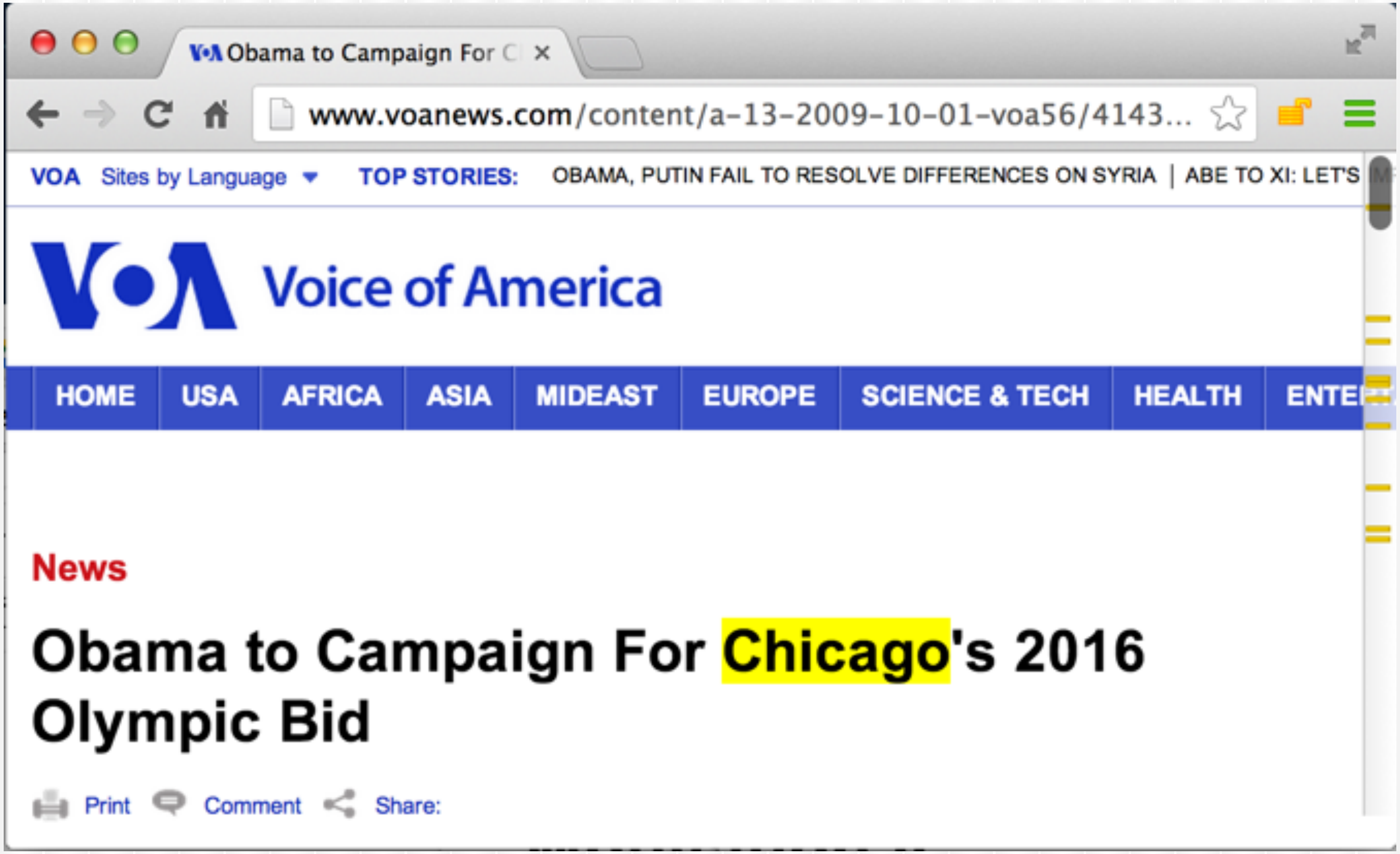
As of 11/2013

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

Observation II.
But, A Lot of Mistakes :-(

Errors Can Creep in at Every Stage

Extraction error: (Obama, nationality, Chicago)



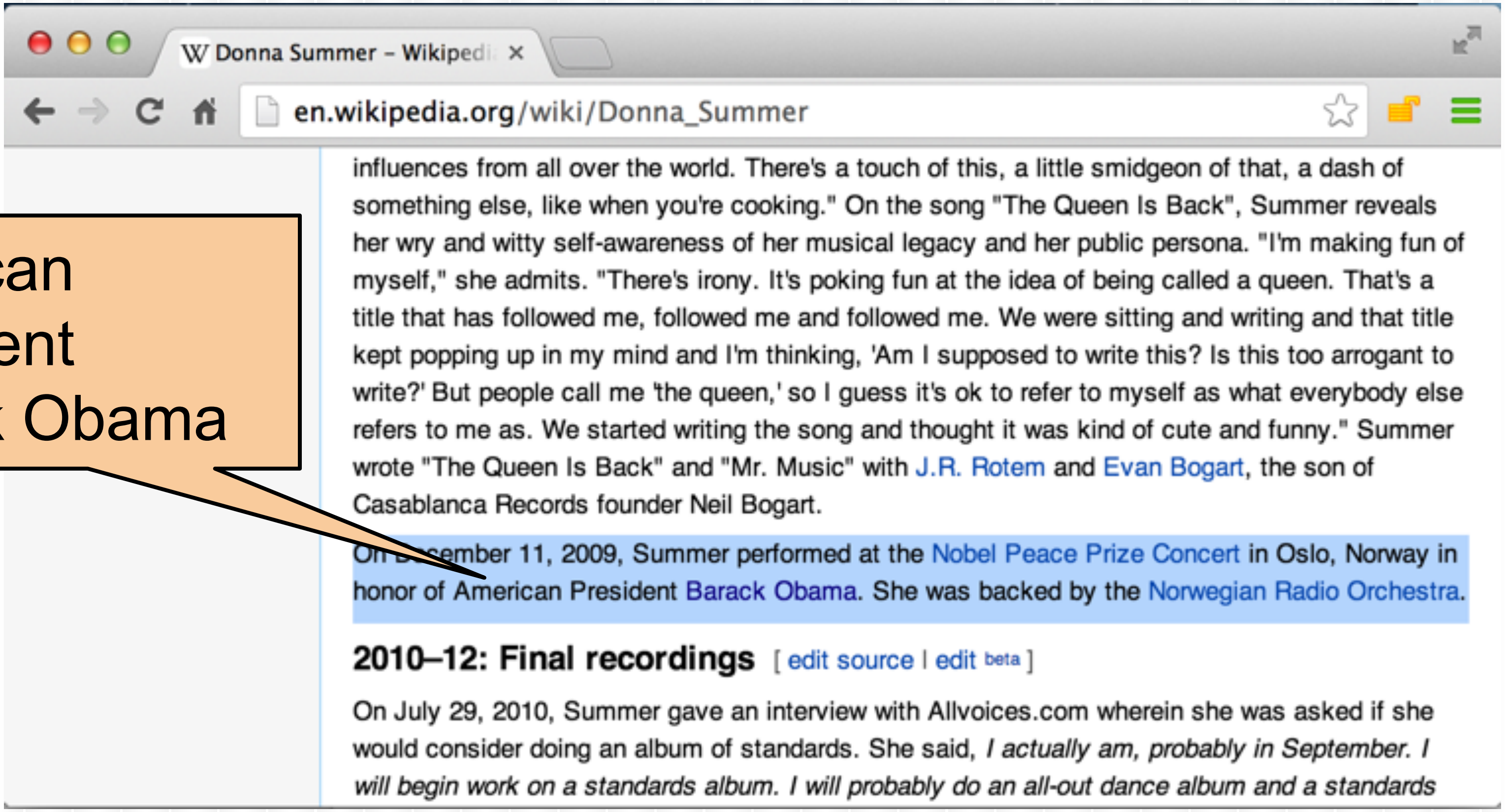
VOA Obama to Campaign For C ×
www.voanews.com/content/a-13-2009-10-01-voa56/4143...
VOA Sites by Language
TOP STORIES:
OBAMA, PUTIN FAIL TO RESOLVE DIFFERENCES ON SYRIA | ABE TO XI: LET'S
VOA Voice of America
HOME
USA
AFRICA
ASIA
MIDEAST
EUROPE
SCIENCE & TECH
HEALTH
ENTE
Print
Comment
Share:

News

Obama to Campaign For Chicago's 2016 Olympic Bid

Errors Can Creep in at Every Stage

Reconciliation error:
(Obama, nationality, North America)



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*

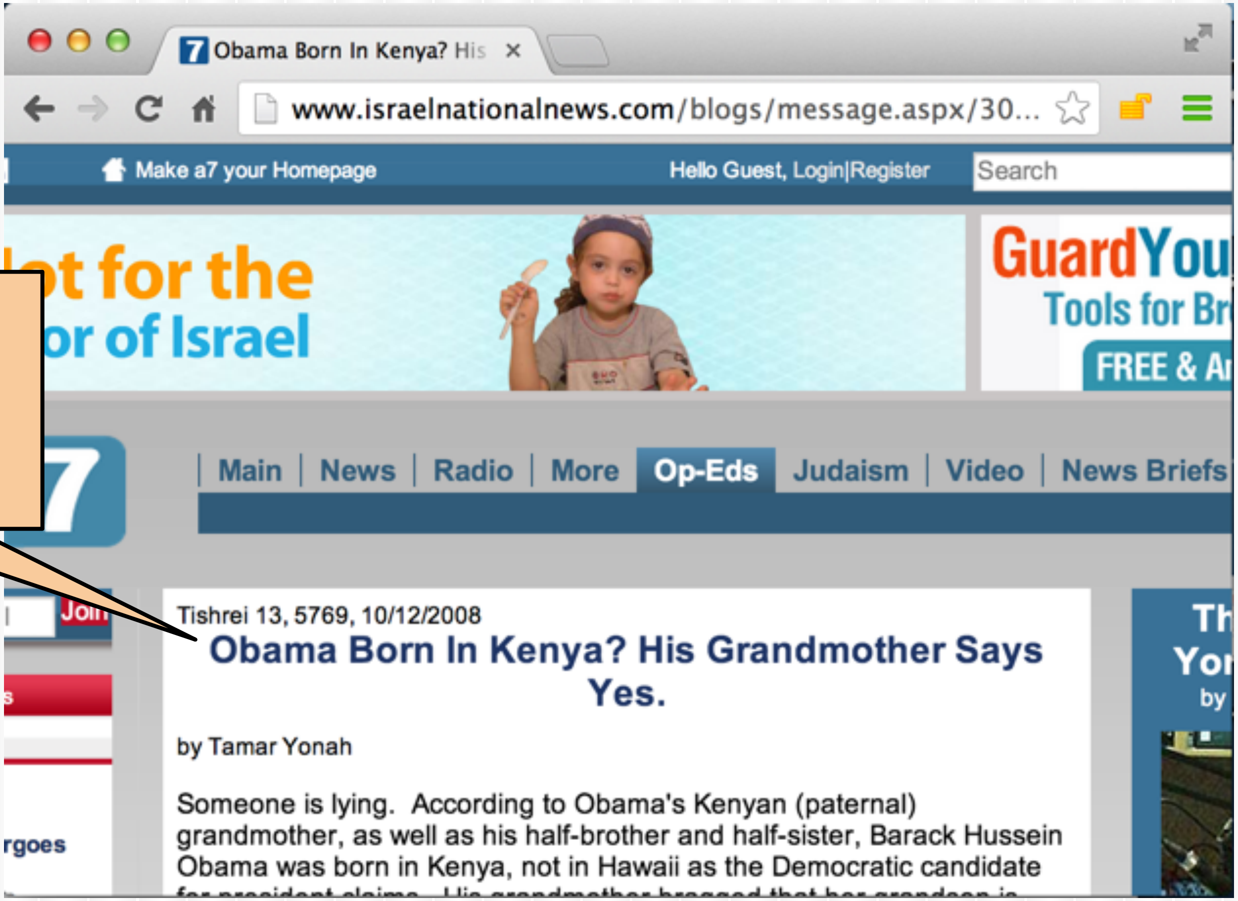
American
President
Barack Obama

9/2013

Errors Can Creep in at Every Stage

Source data error: (Obama, nationality, Kenya)

Obama born in Kenya


Obama Born In Kenya? His
www.israelnationalnews.com/blogs/message.aspx/30...
Make a7 your Homepage
Hello Guest, Login|Register
Search
ot for the
or of Israel
GuardYou
Tools for Br
FREE & Ar
Main | News | Radio | More | Op-Eds | Judaism | Video | News Briefs
Join
rgoes

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

9/2013

Statistics for Triple Correctness

- The gold standard (based on Freebase) contains about 40% of the triples
- Overall accuracy: 30%
- Random sample on 25 false triples
    - Extraction errors: 24 (96%)
    - Source-data errors: 1 (4%)

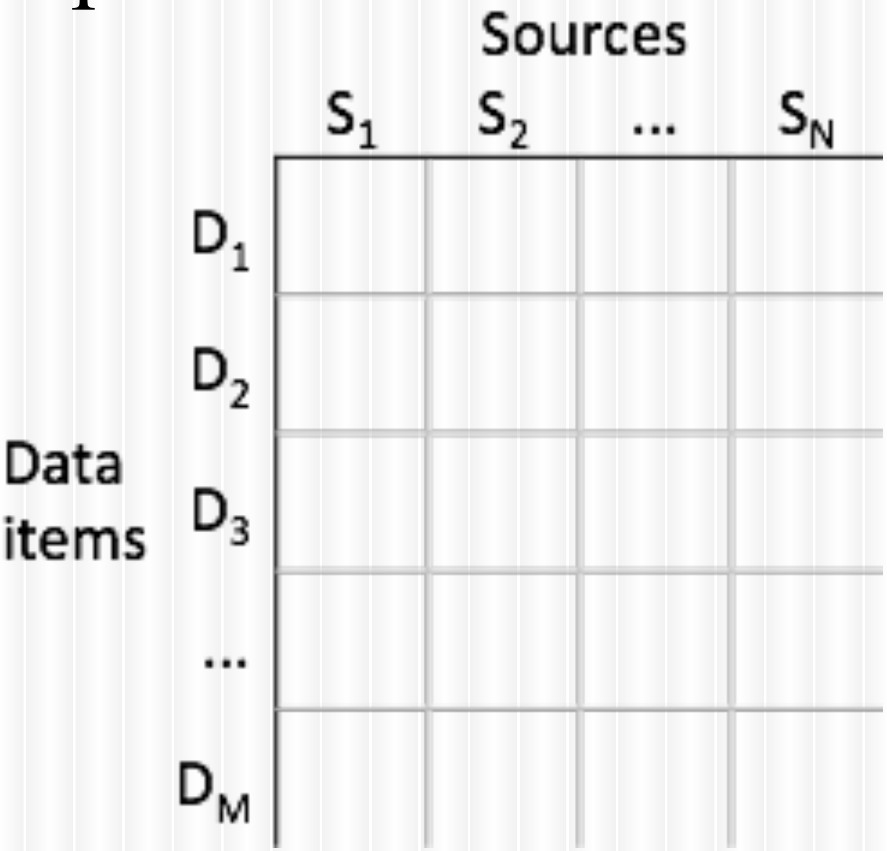
Knowledge Fusion

- Input: Knowledge triples and their provenances (i.e., which extractor extracts from which source)
- Output: a probability in [0,1] for each triple
  - High pr→ search, etc.
  - Medium pr→ active learning, probabilistic inference, etc.
  - Low pr→ Negative training examples

Observation III. *Data Fusion* Techniques Work Fairly Well for *Knowledge Fusion*

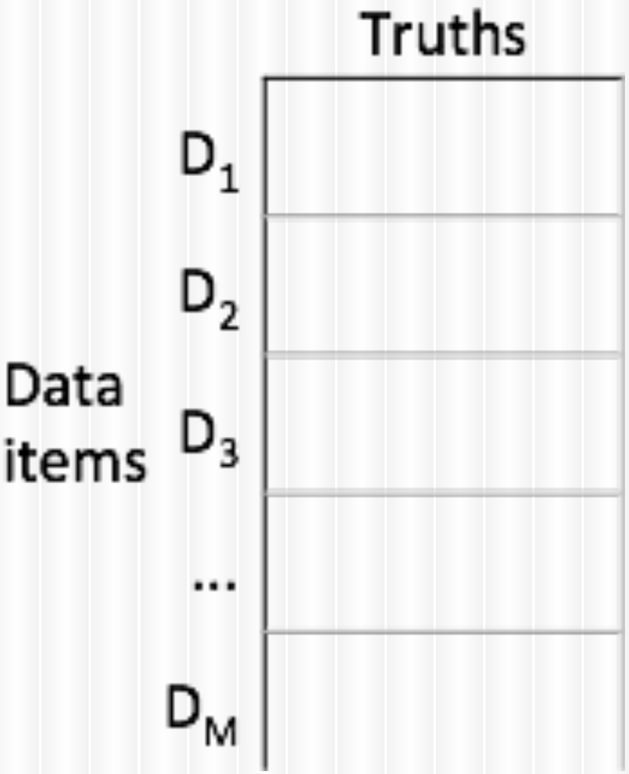
Data Fusion–Definition

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

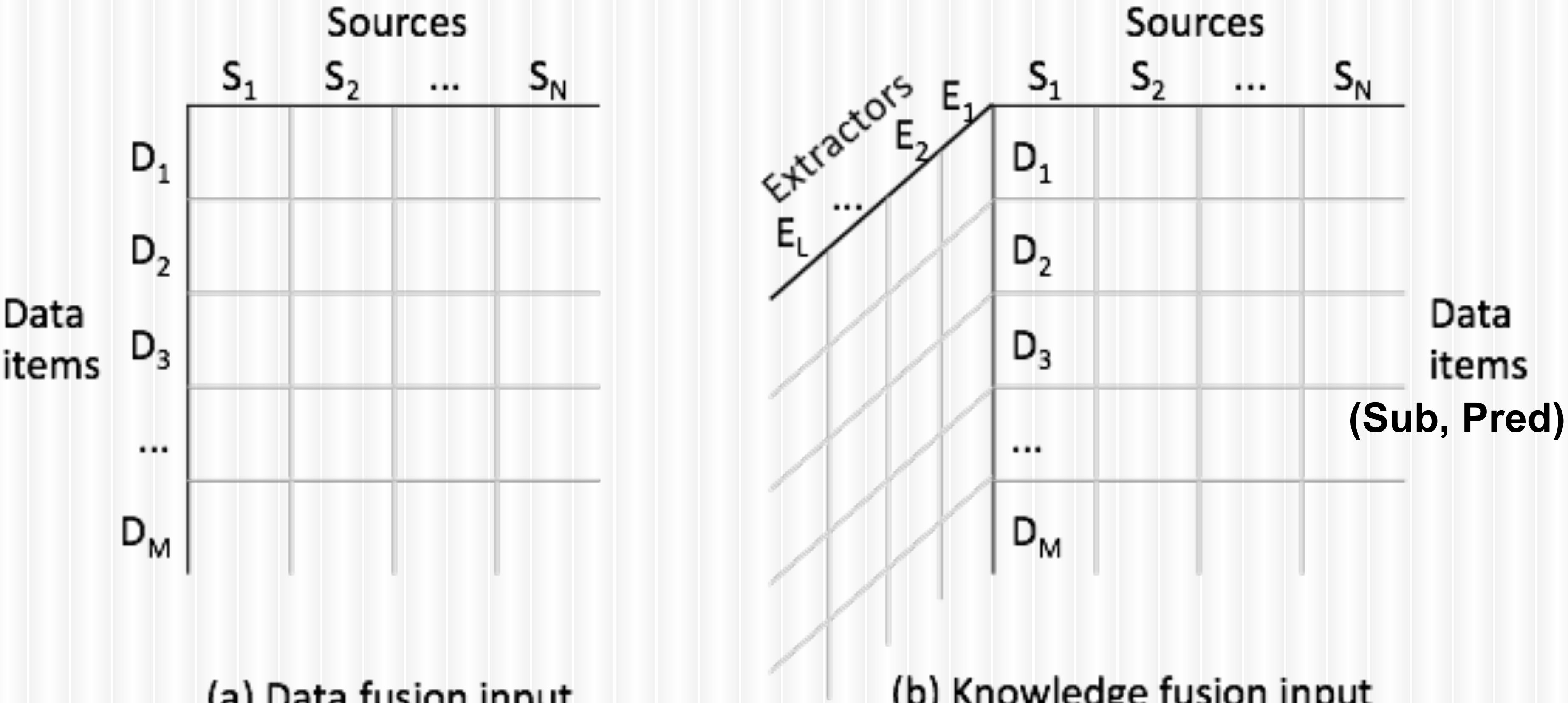
Data Fusion–Intuition

	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.

Knowledge Fusion Challenges

I. Input is ***three-dimensional***

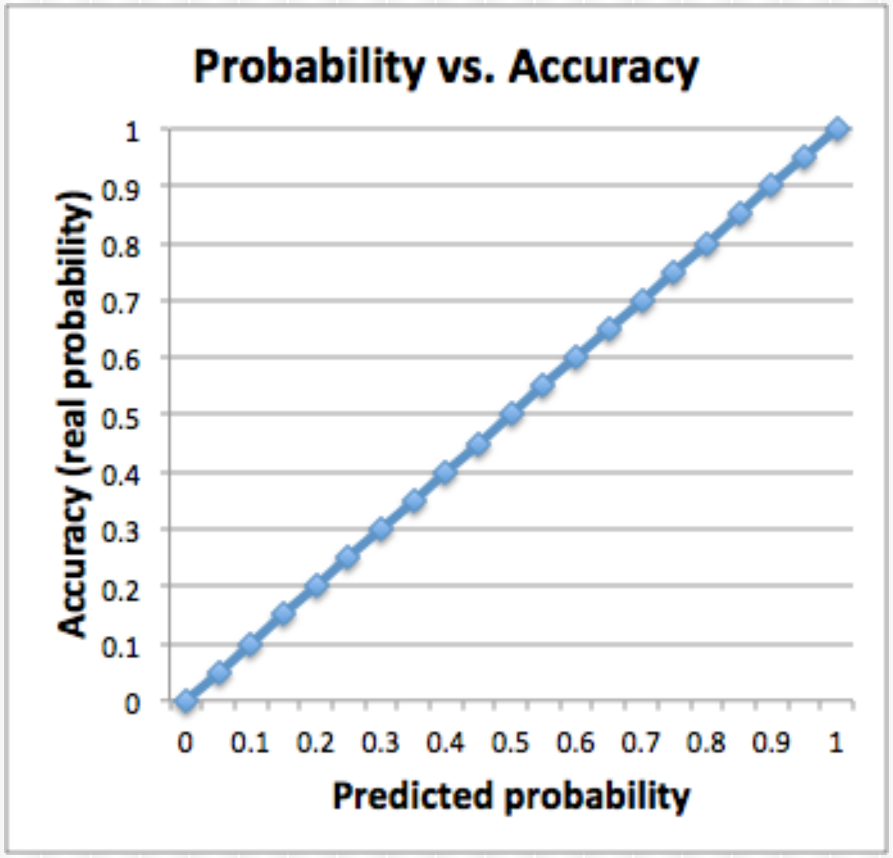


(a) Data fusion input

(b) Knowledge fusion input

Knowledge Fusion Challenges

II. Output prs should be ***well-calibrated***



Knowledge Fusion Challenges

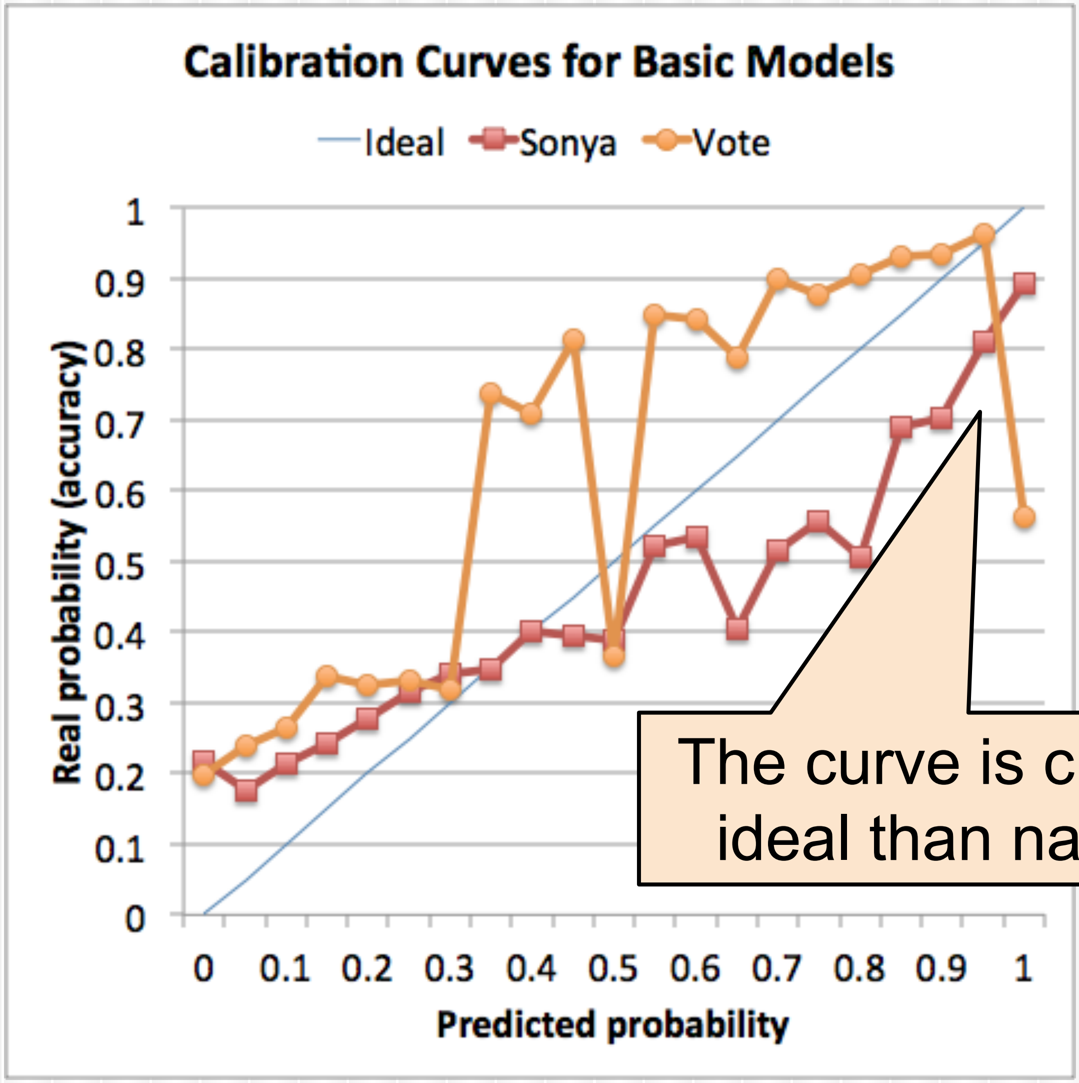
III. Data are of ***Web-scale***

- 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 a Bayesian model to improve probability calibration
- MapReduce Based Framework
    - Sample for *too big* data items or provenances

Basic Sonya Solution vs. Voting



The curve is closer to the ideal than naive voting

Refinements

I. Ignore low-coverage provenances

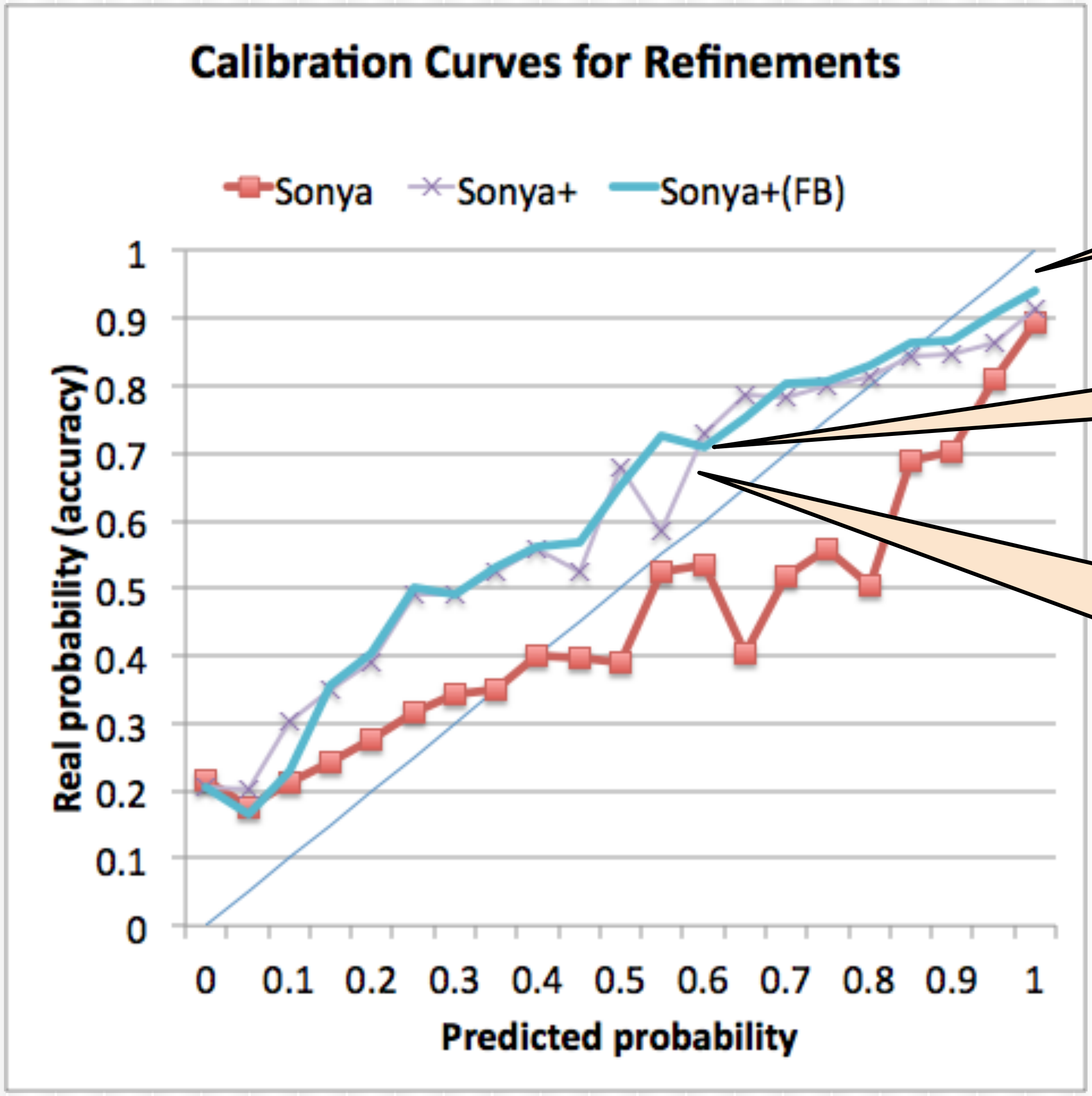
II. Granularity (URL->Site, Extractor->Pattern, Predicate)

III. Ignore low-accuracy provenances

IV. Initiate provenance accuracy by FB

+I, II, III. Sonya+ : unsupervised
+IV. Sonya+(FB) : semi-supervised

Calibration Curve



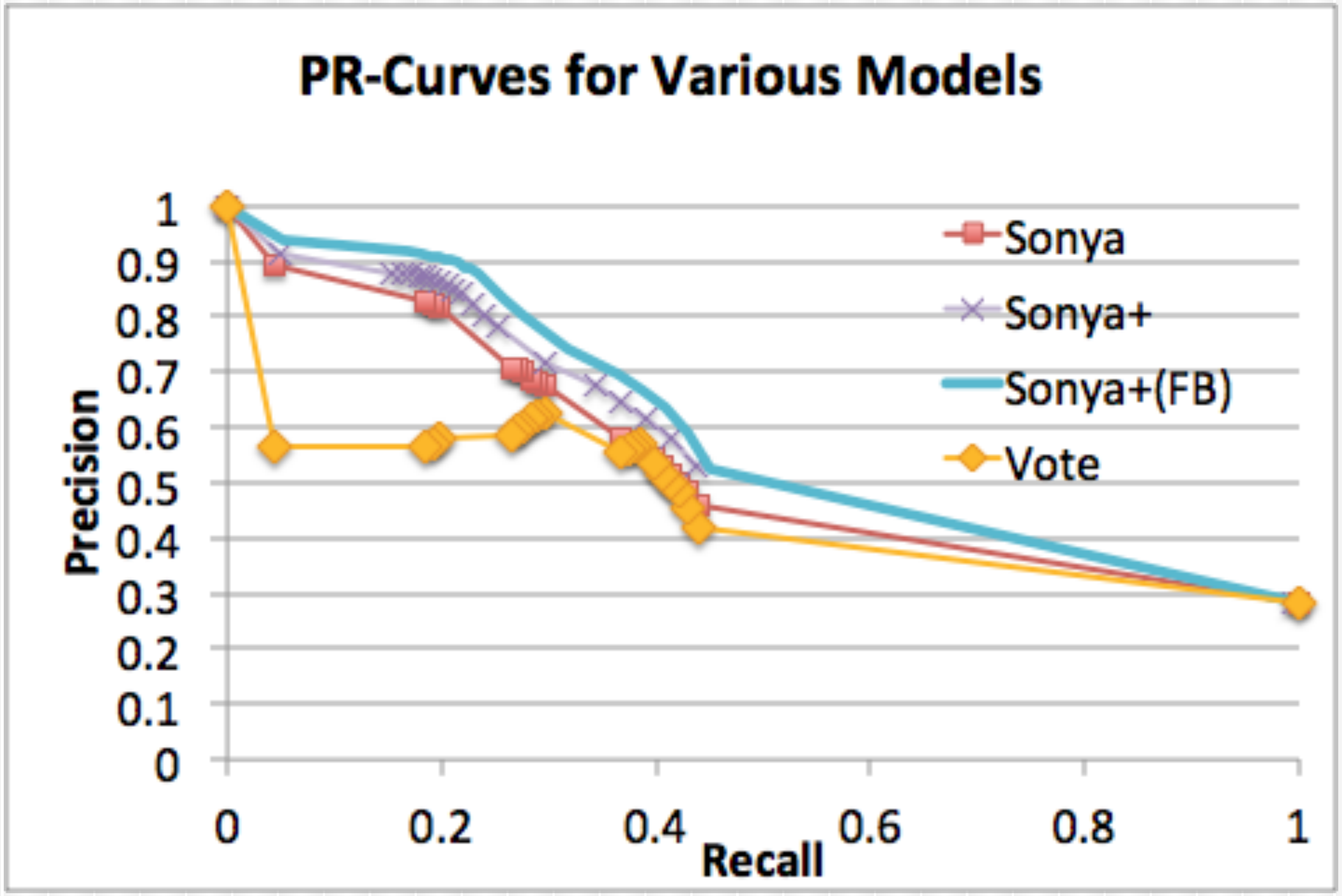
Higher accuracy

Smoother curve

Unsupervised methods very effective

Weighted deviation:
0.037→0.035→0.032

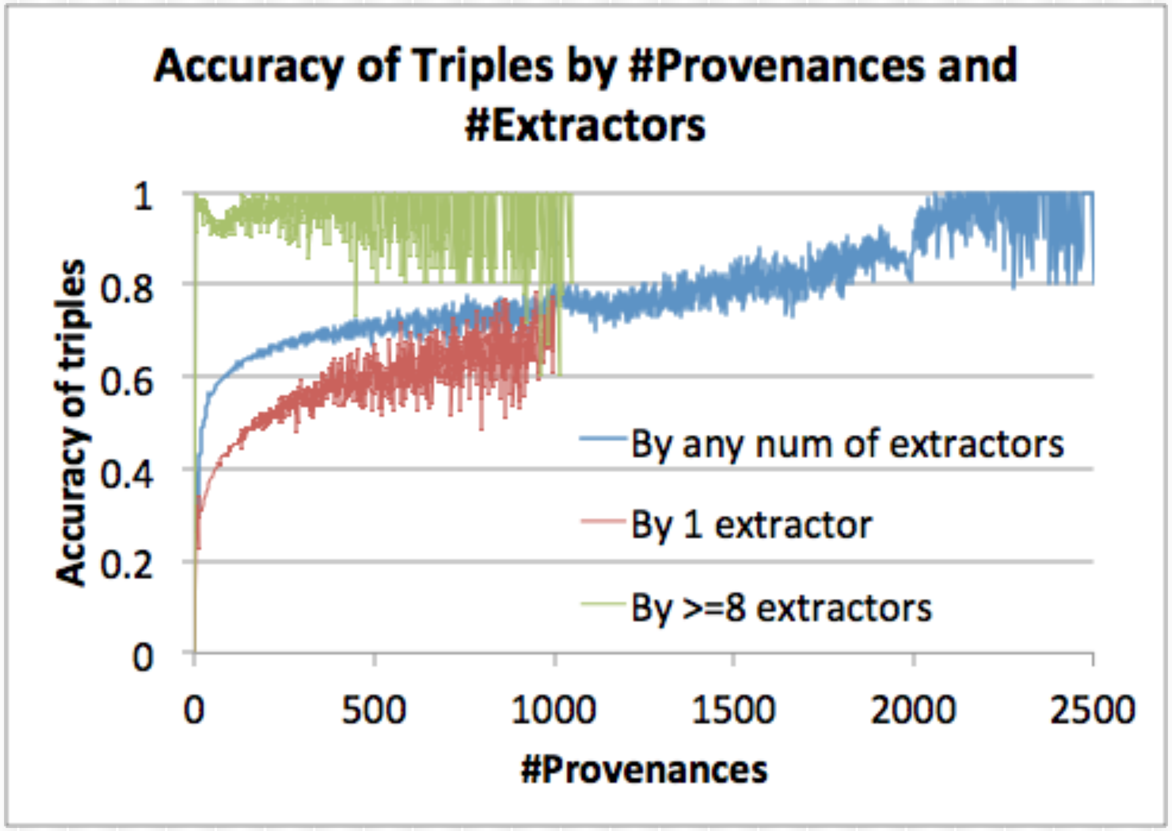
Precision-Recall Curve



Observation IV.
Still Many Places to Improve!

One Inherent Limitation

Cannot distinguish errors from extractors and from sources



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!!

THANK YOU!

Questions?