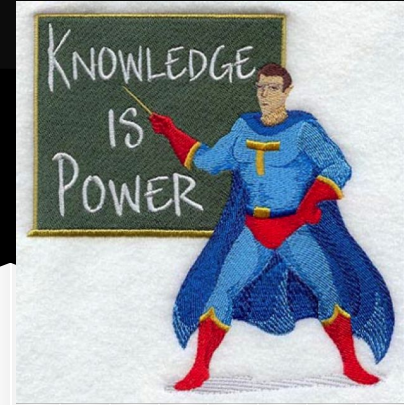


How Far Are We from Collecting the Knowledge in the World

Xin Luna Dong
Google Inc. → Amazon
6/2016

Knowledge Is Power



- Many Knowledge Bases (KB)



NELL: Never-Ending Language Learning

ProBase



facebook



Using KB in Search

downward facing dog



All Images Videos Shopping News More Search tools

About 639,000 results (0.33 seconds)

Downward-Facing Dog | Adho Mukha Svanasana | Yoga Pose

www.yogajournal.com/pose/downward-facing-dog/

Aug 28, 2007 - **Downward-Facing Dog**: Step-by-Step Instructions. ... Then with an exhalation, push your top thighs back and stretch your heels onto or down toward the floor. ... Adho Mukha Svanasana is one of the poses in the traditional Sun Salutation sequence.

How to Do Downward-Facing Dog in Yoga - YogaOutlet.com

www.yogaoutlet.com/guides/how-to-do-downward-facing-dog-in-yoga/

One of the most recognized yoga poses in the West, **Downward-Facing Dog** — Adho Mukha Svanasana (Ah-doh MOO-kuh shvan-AHS-uh-nuh) — is a standing pose and mild inversion that builds strength while stretching the whole body. ... **Downward-Facing Dog** energizes and rejuvenates the ...

How to Perform Downward Facing Dog in Yoga (with quick ...

www.wikihow.com/...>Health>Alternative Health>Yoga

Star yoga

Do
www
I rec
more



Adho Mukha Svanasana

Yoga pose

Adho mukha śvānāsana, adho mukha shvanasana, downward-facing dog Pose, downward dog, or down dog is an asana. [Wikipedia](#)

Note: Consult a doctor before beginning an exercise regime

Strengthens: Leg, Arm

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 KB in Social Media

Past 25 Days

124 TWEETS

Topsy Sentiment Score: 67

Tweets Over Past 30 Days



[View trends on Topsy Analytics](#)



Official ACM @theofficialacm

Michael Stonebraker, the 2014 ACM A.M. Turing Award recipient, see the People of ACM profile bit.ly/1H1GU1a
pic.twitter.com/qvRhG3dX3U

25 days ago Reply Retweet Favorite 24 more

[Interview: Michael Stonebraker, greatest living contributor to database technology](#)

kdnuggets.com/2015/05/interview-michael-stonebraker-greatest-living-contributor-to-database-technology.html



#teamgogetit247 @gogetit247

Turing Award

The ACM A.M. Turing Award is an annual prize given by the Association for Computing Machinery to "an individual selected for contributions of a technical nature made to the computing community". [Wikipedia](#)



Ceremony date (2015): June 20, 2015

People also search for: Fields Medal, Abel Prize, National Medal of Technology and Innovation

Winners

Michael Stonebraker	2014
Leslie Lamport	2013
Silvio Micali	2012



Alan Turing

Computer Scientist

Alan Mathison Turing, OBE, FRS was a British pioneering computer scientist, mathematician, logician, cryptanalyst, philosopher, mathematical biologist, and marathon and ultra distance runner. [Wikipedia](#)

Born: June 23, 1912, Maida Vale, London, United Kingdom
Died: June 7, 1954, Wilmslow, United Kingdom
Education: Princeton University (1936–1938). [More](#)
Parents: Julius Mathison Turing, Ethel Sara Stoney
Siblings: John Turing



Michael Stonebraker

Computer scientist

Michael Ralph Stonebraker is a computer scientist specializing in database research. Through a series of academic prototypes and commercial startups, Stonebraker's research and products are central to ... [Wikipedia](#)

Born: October 11, 1943 (age 72), Milton, NH

Books: [Architecture of a Database System](#), [More](#)

Education: University of Michigan (1971), Princeton University (1965)

Awards: Turing Award, IEEE John von Neumann Medal

Notable awards: IEEE John von Neumann Medal (2005), Turing Award (2014)

Organizations founded: Vertica, StreamBase Systems, Cohera

Using KB in Recommendation

SUMMER CAMP ACTIVITIES EXPO
San Mateo
Today, 11:00 AM – 3:00 PM


The Classic Mission Mural Walk
San Francisco
Today, 1:30 PM

Stories to read

top Washington Post · 12 hours ago

Cruz gains steam with 2 wins on 'Super Saturday'; Trump calls on ...

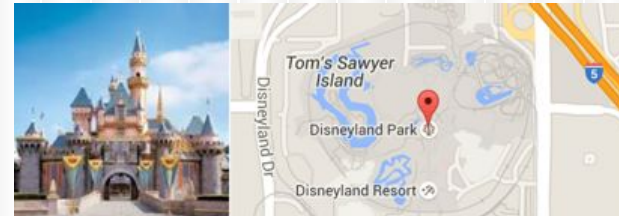

Sen. Bernie Sanders won in Kansas and Nebraska, while Donald Trump took the Louisiana GOP contest, but Ted Cruz secur...



top Washington Post · 6 hours ago

D.C. area forecast: Somewhat cool Sunday gives way to spring ...

We're still on the cool side into this afternoon, but the weather could hardly be better as we head into the work week.



Disneyland ★

Website Directions

4.6 ★★★★★ 3,684 Google reviews

Theme park in Anaheim, California

Disneyland Park, originally Disneyland, is the first of two theme parks built at the Disneyland Resort in Anaheim, California, opened on July 17, 1955. It is the only theme park designed and built under the direct supervision of Walt Disney. [Wikipedia](#)

Address: 1313 Disneyland Dr, Anaheim, CA 92802

Opened: July 17, 1955

Hours: Open today · 8AM–12AM ▾

Founder: Walt Disney

Founded: July 17, 1955, Anaheim, CA

Customer service: 1 (714) 781-7277

Sales: 1 (877) 560-6477



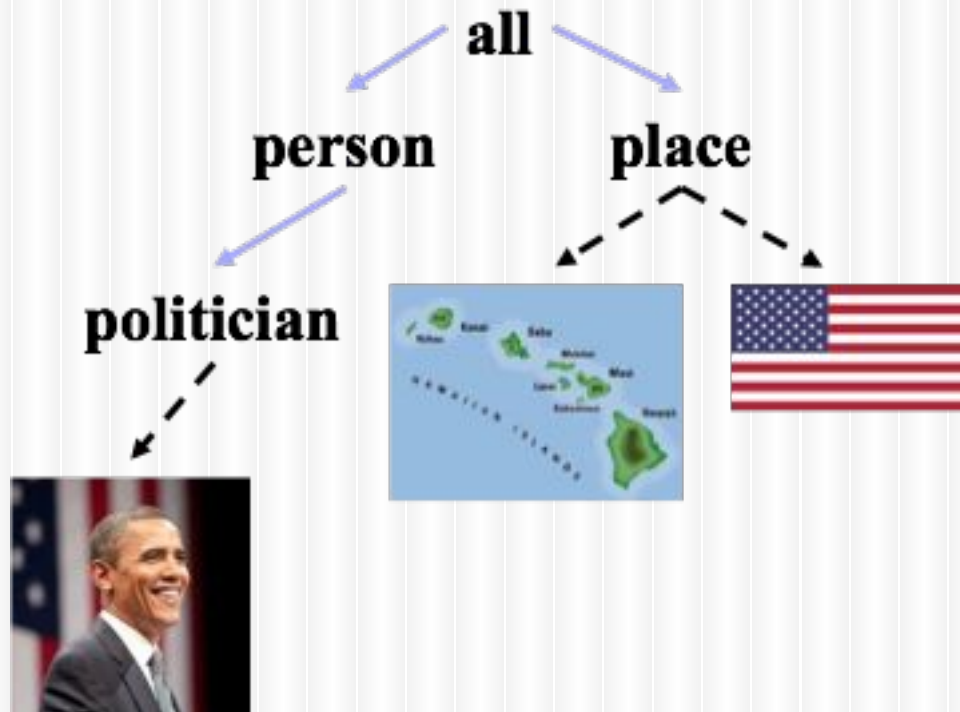
ABC lets it go (big) with 'Disneyland' special - USA Today

www.usatoday.com/.../abc-lets-go-big-disneyland-special/80... ▾ USA Today ▾

Feb 18, 2016 - ANAHEIM, Calif. – ABC will flex some Disney marketing muscle during Sunday's two-hour celebration of Disneyland's 60th anniversary, with ...

What is a Knowledge Base

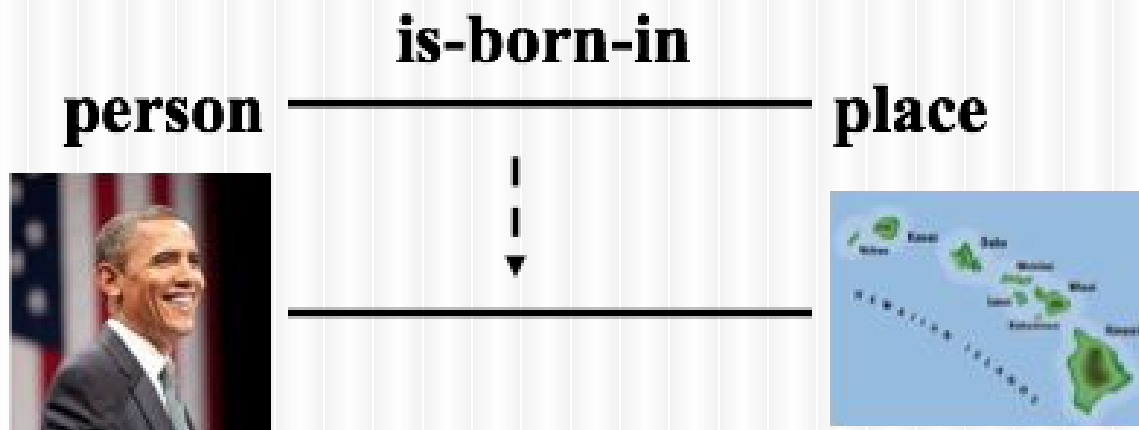
- Entities, entity types
 - An entity is an instance of multiple types
 - Entity types organized in a hierarchy



What is a Knowledge Base

.....

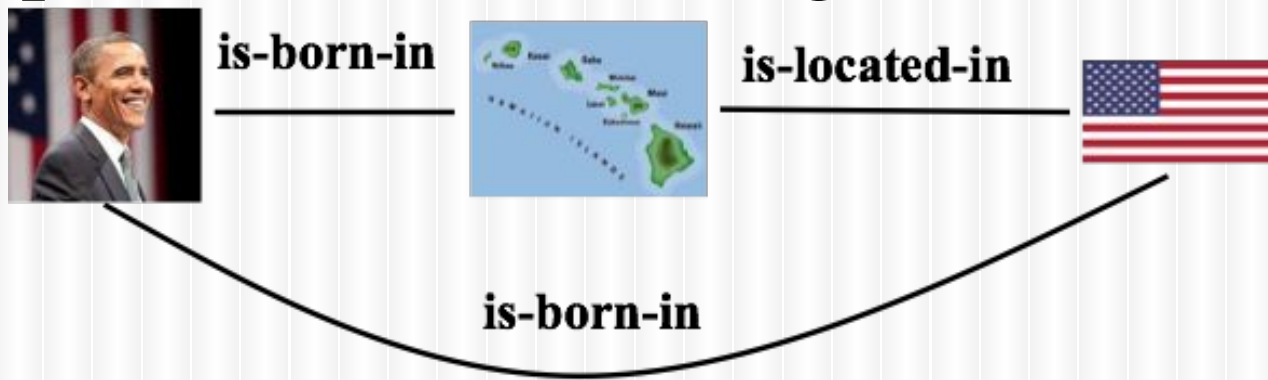
- Entities, entity types
- Predicates, (sub, pred, obj) triples
 - A triple describes an attribute of an entity, or, the relationship between two entities



What is a Knowledge Base

.....

- Entities, entity types
- Predicates, (sub, pred, obj) triples
- Knowledge base: graph with entity nodes and predicate-labeled edges



Advantages over Traditional DBs

- Easy to model complex relationships in the real world
- Easy to extend schema
- Easy to specify rules and make inference

Outline

- I. Gap between existing KBs and knowledge in the world
- II. Efforts to collect tail knowledge
 - Knowledge Vault
 - Lightweight verticals
- III. Key techniques: knowledge fusion and knowledge-based trust
- IV. Conclusions and future directions

Existing Knowledge Bases [DGH+14]

.....

Name	# of Entity Types	# Predicates	# Entities	# Confident Triples
Knowledge Vault (KV)	1100	4469	45M	271M
DeepDive	4	34	2.7M	7M
NELL	271	306	5.1M	0.435M
PROSPERA	11	14	N/A	0.1M
Yago2	350,000	100	9.8M	150M
Freebase	1500	35,000	40M	637M
Knowledge Graph	1500	35,000	570M	18,000M

Freebase Statistics (As of 3/2016)

- 2.3B triples on 130M entities
- Break-down

	#Triples
<i>Total</i>	2.3B
Name/Alias	1.3B
Type	341M
Webpages	88M
Description	31M
Facts	482M (20%)

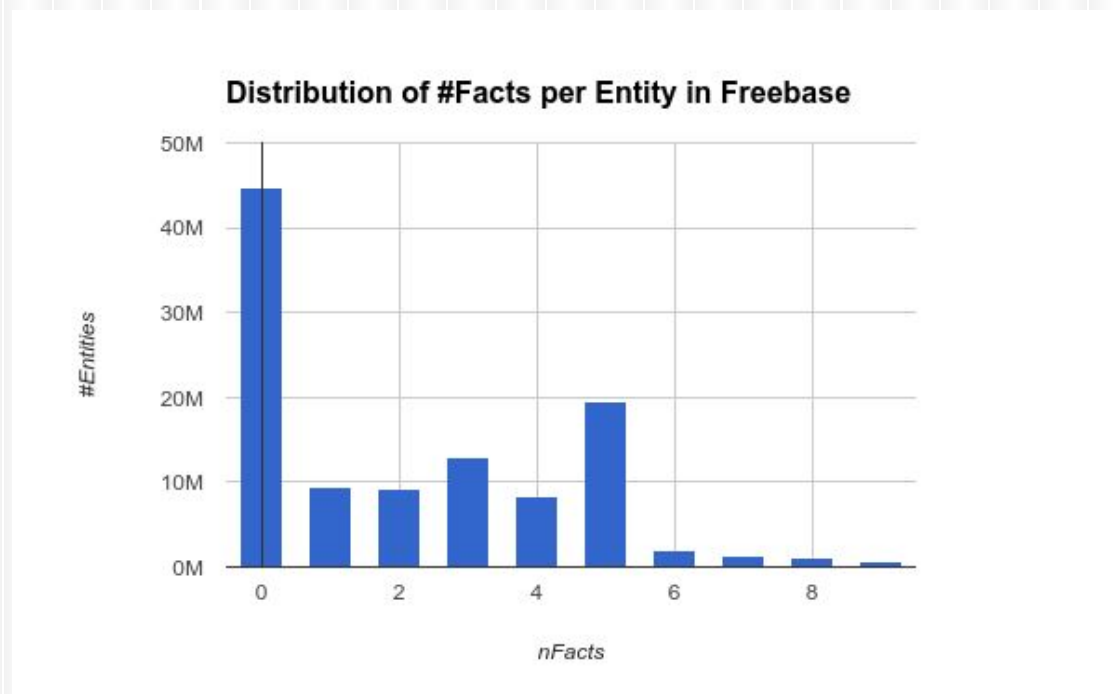
Category I. Head Entities in Head Verticals

- Rich knowledge for head entities in head verticals. E.g., (Freebase as of 3/2016)

Vertical	Percentage ≥ 5 facts	Example entity	#Facts
country	80%	<i>USA</i>	151K
person	43%	<i>Barack Obama</i>	1.5K
business	21%	<i>Google Inc.</i>	1K
film	73%	<i>Frozen</i>	200
album	66%	<i>American Idiot</i>	21

Category II. Tail Entities in Head Verticals

- #Facts/Entity in Freebase (as of 3/2016)
 - 40% entities with no fact
 - 56% entities with <3 facts




Category II. Tail Entities in Head Verticals

● Example 1

[/g/12cq4wlq9](#) copy

Xin Luna Dong (en)

Types: [/base/type_ontology/agent](#), [/base/type_ontology/animate](#), [/base/t](#)



(Missing text description)

Properties Types & Collections Names Keys Triple

Mode: Formatted Raw Show Provenance Show proto.topic


Type	reset invert
<input checked="" type="checkbox"/> /book/author (1)	
<input type="checkbox"/> /common/topic (15)	
<input type="checkbox"/> /freebase/object_profile (1)	
<input type="checkbox"/> /freebase/relevance/scores (1)	

/book	view schema
Author /book/author	
Works Written /book/author/works_written	Big Data Integration

[/m/09pfy0r](#) copy

Xin Luna Dong (en)

Types: [/education/academic](#), [/book/author](#), [/common/topic](#), [VIEW ALL]



(Missing text description)

Equivalent Topic URLs:
www.ams.org/mathscinet/s_genealogy.math.ndsu.nod

Properties Types & Collections Names Keys Triple

Mode: Formatted Raw Show Provenance Show proto.topic

Type	reset invert
<input checked="" type="checkbox"/> /book/author (1)	
<input type="checkbox"/> /common/topic (30)	
<input type="checkbox"/> /freebase/object_profile (1)	
<input type="checkbox"/> /freebase/relevance/scores (1)	
<input type="checkbox"/> /kg/object_profile (1)	
<input checked="" type="checkbox"/> /people/person (1)	
<input type="checkbox"/> /type/object (13)	

/book	view schema
Author /book/author	
Works Written /book/author/works_written	Providing Best Effort Services in Dataspace Systems

/people	view schema
Person /people/person	
Profession /people/person/profession	Mathematician

Freebase

Category II. Tail Entities in Head Verticals

● Example 2

Topic base | Topic Diff

[/m/0c5225n](#) copy

From the beginnings to 1945 (en)

Types: [/book/book](#), [/book/written_work](#), [/common/topic](#), [\[VIEW ALL\]](#)

(Missing text description)

Equivalen

Properties Types & Collections Names Keys Triple

Mode: Formatted Raw Show Provenance Show proto.topic

Type reset invert

- [/book/book](#) (1)
- [/book/written_work](#) (1)
- [/common/topic](#) (6)
- [/freebase/object_profile](#) (1)
- [/freebase/relevance/scores](#) (1)
- [/kg/object_profile](#) (1)

/book view schema
Book /book/book
Editions /book/book/editions From the beginnings to 1945
Written Work /book/written_work
Author /book/written_work/author Mary Ann. Dimand

Freebase

Book Depository.com

Search for books by keyword / title / author / ISBN Advanced search

Bestsellers Coming soon Highlights Bargain Shop

Economics / Economics / Economic Theory & Philosophy / Economic Theory & Philosophy / Game Theory

The History Of Game Theory, Volume 1 : From the Beginnings to 1945

Paperback | Routledge Studies in the History of Economics | English

By (author) Mary Ann Dimand , By (author) Robert W. Dimand

Share [x](#) [f](#) [t](#) [p](#)

Game Theory - the formal modelling of conflict and cooperation - first emerged as a recognized field with a publication of John von Neumann and Oskar Morgenstern's Theory of Games and Economic Behaviour in 1944. Since then, game-theoretic thinking about choice of strategies and the interdependence of people's actions has influenced all the social sciences. However, little is known about the history of the theory of strategic games prior to this publication. In this volume, the history of strategic games - from its origins up to 1945 - is traced through the work of: * 19th Century economists such as Cournot and Edgeworth * Voting theorists - including Lewis Carroll * Conflict theorists - Richardson and Lanchester * Probabilists such as Bertrand, Borel and Ville * Later economists - notably Stackelberg and Zeuthen This authoritative account of the history of game theory concludes with a historical perspective on the achievement of von Neumann and Morgenstern, and an appraisal of the reception of their book.

Product details

Format: Paperback | 200 pages

Dimensions: 157.48 x 233.68 x 15.24mm | 317.51g

Publication City/Country: London, United Kingdom

Language: English

Publication date: 15 Aug 2014

ISBN10: 1138006602

Publisher: Taylor & Francis Ltd

ISBN13: 9781138006607

Imprint: ROUTLEDGE

www.bookdepository.com

Category III. Head Entities in Tail Verticals

- 100 sample tail verticals
(Freebase as of 3/2016)
 - Example verticals: philosopher, profession, yoga_poses, pokemon_characters
 - Entities collected from 1-3 authoritative sources for the vertical
 - In total 17K entities; **6.5K (40%) entities** not in Freebase
 - No vertical-related attributes (**~1K in total**)

Category III. Head Entities in Tail Verticals

- Example: Aquamarine (March gemstone)
 - No entity in Freebase
 - No gemstone-related attributes

Aquamarine and maxixe [\[edit\]](#)



Aquamarine

Aquamarine (from Latin: *aqua marina*, "water of the sea") is a blue or cyan variety of beryl. It occurs at most localities which yield ordinary beryl. The gem-gravel placer deposits of Sri Lanka contain aquamarine. Clear yellow beryl, such as that occurring in Brazil, is sometimes called *aquamarine chrysolite*.^[*citation needed*] The deep blue version of aquamarine is called *maxixe*. Maxixe is commonly found in the country of Madagascar. Its color fades to white when exposed to sunlight or is subjected to heat treatment, though the color returns with irradiation.

The pale blue color of aquamarine is attributed to Fe^{2+} . Fe^{3+} ions produce golden-yellow color, and when both Fe^{2+} and Fe^{3+} are present, the color is a darker blue as in maxixe. Decoloration of maxixe by light or heat thus may be due to the charge transfer between Fe^{3+} and Fe^{2+} .^{[8]^{[9]^{[10]^[11]}} Dark-blue maxixe color can be produced in green, pink or yellow beryl by irradiating it with high-energy particles (gamma rays, neutrons or even X-rays).^[12]}

In the United States, aquamarines can be found at the summit of Mt. Antero in the Sawatch Range in central Colorado. In Wyoming, aquamarine has been discovered in the Big Horn Mountains, near Powder River Pass. Another location within the United States is the Sawtooth Range near Stanley, Idaho. Although the minerals are within a wilderness area which prevents collecting. In Brazil, there are mines in the states of Minas Gerais, Espírito Santo, and Bahia, and minorly in Rio Grande do Norte. The mines of Colombia, Zambia, Madagascar, Malawi, Tanzania and Kenya also produce aquamarine.

The largest aquamarine of gemstone quality ever mined was found in Marambaia, Minas Gerais, Brazil, in 1910. It weighed over 110 kg (240 lb), and its dimensions were 48.5 cm (19 in) long and 42 cm (17 in) in diam. The Dom Pedro aquamarine, now housed in the Smithsonian Institution's National Museum of Natural History,[[]



Faceted aquamarine, 13.24ct, Brazil

Wikipedia

Aquamarine Gemological Properties: [Back to Top](#)

Chemical Formula:	Al ₂ Be ₃ Si ₆ O ₁₈ , Aluminum beryllium silicate
Crystal Structure:	Hexagonal, hexagonal prisms
Color:	Light-blue to dark-blue, blue-green
Hardness:	7.5 - 8 on the Mohs scale
Refractive Index:	1.564 - 1.596
Density:	2.68 - 2.74
Cleavage:	Indistinct
Transparency:	Transparent to opaque
Double Refraction / Birefringence:	-0.004 to -0.005
Luster:	Vitreous
Fluorescence:	None

Gap Between KBs and World Knowledge



	A_1	A_2	A_3	A_4	A_5	A_6	A_n	UNKNOWN ATTRIBUTES						
E_1																
E_2					EXISTING KNOWLEDGE											
E_3																
E_4																
E_5																
E_6																
...																
E_m			UNKNOWN VALUES													
UNKNOWN ENTITIES																

Head knowledge mainly collected by manual curation or importing large data sets

How to collect long-tail knowledge in a scalable way?

Challenges in Collecting Long-Tail Knowledge

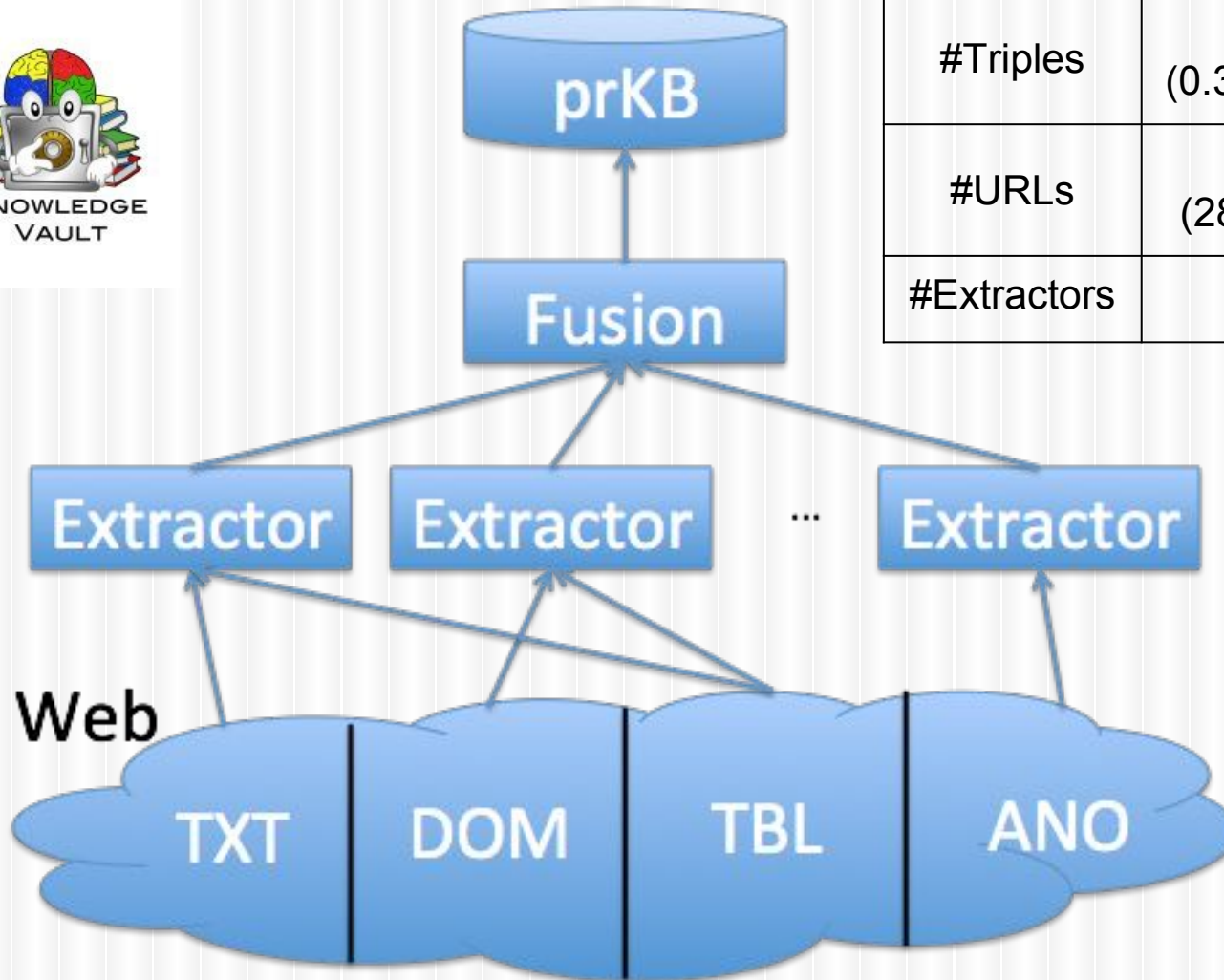
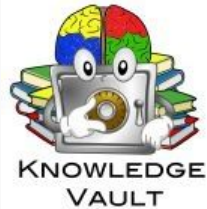
- Curation does not scale
- Automatic extraction
 - Information sparse on the Web
 - Extractors are often trained on head entities/verticals
 - Reconciliation for long-tail entities is error-prone

Outline

- I. Gap between existing KBs and knowledge in the world
- II. Efforts to collect tail knowledge
 - Knowledge Vault
 - Lightweight verticals
- III. Key techniques: knowledge fusion and knowledge-based trust
- IV. Conclusions and future directions

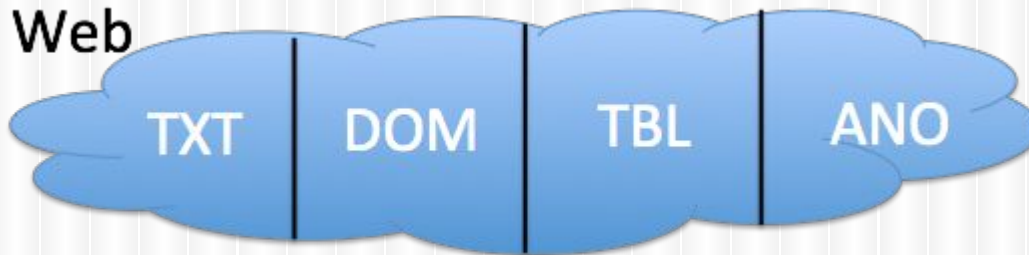
Knowledge Vault– Building a Probabilistic KB

[SIGKDD, 2014]
[VLDB, 2014]



#Triples	3.2B (0.3B w. $pr \geq 0.7$)
#URLs	2.5B (28M Websites)
#Extractors	16

Four Types of Web Sources



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 inventor. His ideas and body of work -- which included *The Last Supper*, *Leda and the Swan* and many others -- influenced countless artists and made the Italian Renaissance.

DOM Tr

yelp Search for (e.g. 'sushi near me')

Welcome About Me Write a Review Find Friends

Shana Thai Restaurant
★★★★ 140 reviews Rating Details

Category: Thai (14)

311 Moffett Blvd
Ste A
Mountain View, CA 94043
(855) 940-9990
http://www.shanathai.com

Explore the menu

Hours:
Mon-Sun 11 am - 2 pm
Mon-Sun 5 pm - 10 pm
Good for Kids: Yes
Accepts Credit Cards: Yes
Parking: Private Lot
Atmos: Casual
Good for Groups: Yes

Price Range: \$\$
Take Reservations: Yes
Delivery: No
Takeout: Yes
Water Service: Yes
Outdoor Seating: No
Wi-Fi: No
Good For: Dinner

Alcohol: Full Bar
Noise Level: Average
Ambiance: Trendy, Casual
Has TV: No
Caters: No
Wheelchair Accessible: Yes

Web tables & Lists

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


Annotations

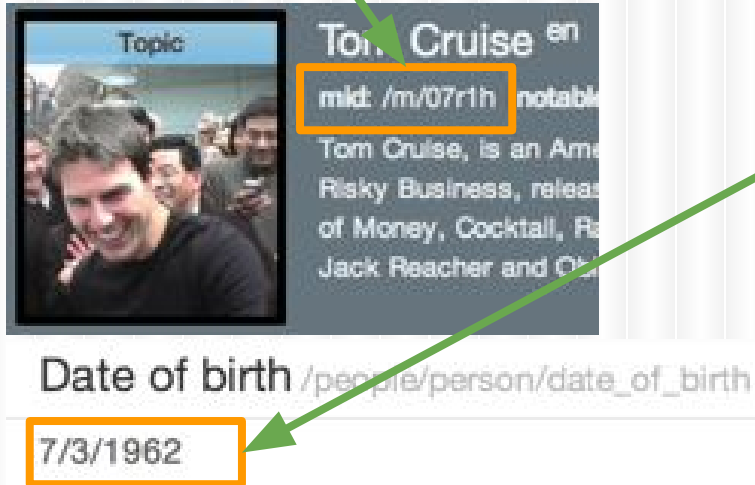
```
<h1 itemprop="name">
Tom Cruise </h1>
<span itemprop="birthDate">
7/3/1962 </span>
<span itemprop="gender">
Male </span>
```

schema.org

Knowledge Extraction

- 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}
	mild : /m/07r1h notable
	Tom Cruise, is an Ame Risky Business, releas of Money, Cocktail, Ra Jack Reacher and Oth

Date of birth /people/person/date_of_birth

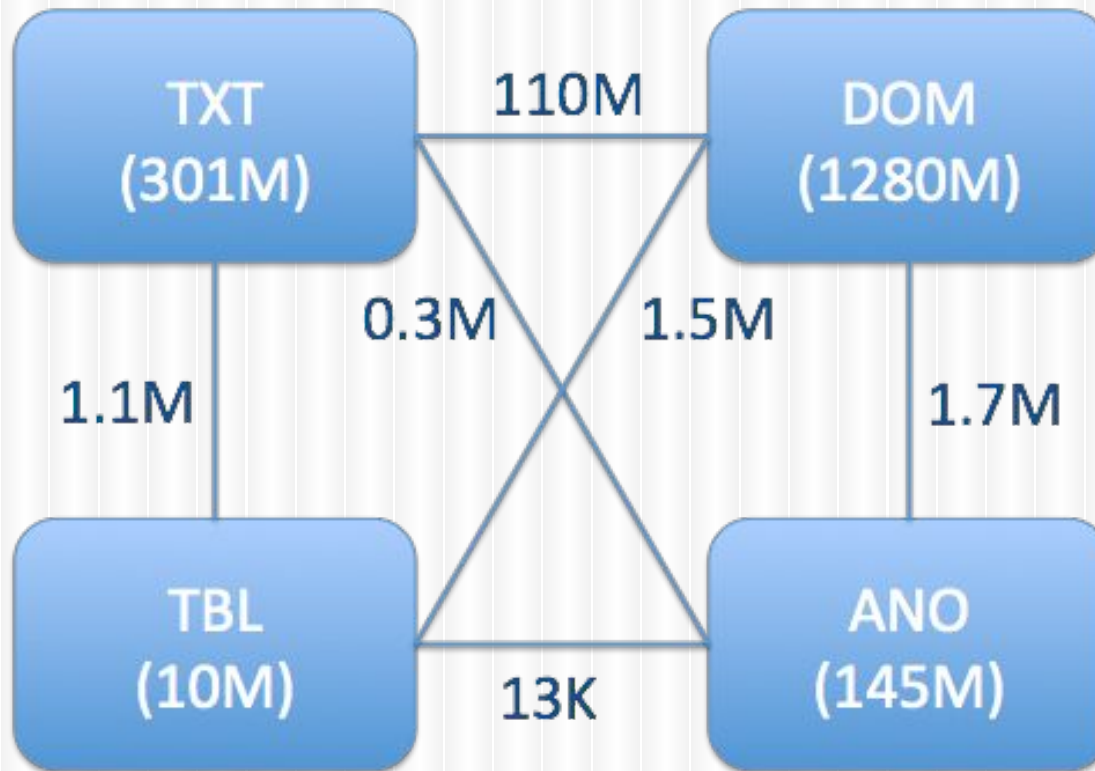
7/3/1962

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

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

Statistics for Data Sources

.....



As of 11/2013

Knowledge Quality

- Gold standard: Freebase under 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

Knowledge Quality

- Gold standard: Freebase under LCWA (Local Closed-World Assumption)
- Well-calibrated probabilities

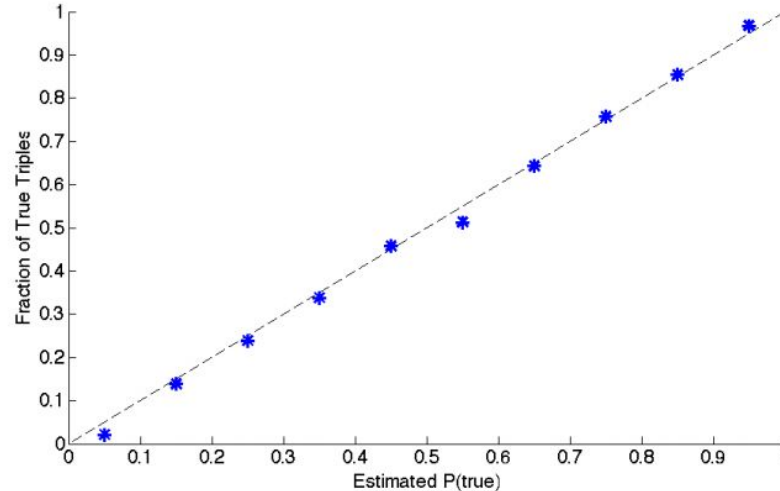


Figure 1: True probability vs estimated probability for each triple in KV.

Challenges in Building KV

- Errors can creep in at every stage
 - Overall accuracy: 20%
 - 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%)
- Challenge: Predict triple correctness

News

Google's Knowledge Vault gains 1.6 billion facts

FELICITY NELSON
SATURDAY, 23 AUGUST



Google's fact-checking bots build vast Knowledge bank

The search that could

Google "Knowledge Vault" To Power Future
Database could be the foundation for array of new

**Good Bye Knowledge Graph,
Hello Google Knowledge Vault?**

Limitations of KV



	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A _n	UNKNOWN ATTRIBUTES										
E ₁																				
E ₂					EXISTING KNOWLEDGE															
E ₃																				
E ₄																				
E ₅																				
E ₆																				
...																				
E _m																				

UNKNOWN
ENTITIES

- KV focuses on existing entities and attributes
 - Training data contain only FB predicates
 - Entities need to be annotated as FB entities
- Among the 0.3B high-confidence triples
 - 0.18B triples not in KG
 - KG contains 18B triples (100X) [KDD'14]
- Oftentimes still regarding head entities

Outline

- I. Gap between existing KBs and knowledge in the world
- II. Efforts to collect tail knowledge
 - Knowledge Vault
 - Lightweight verticals
- III. Key techniques: knowledge fusion and knowledge-based trust
- IV. Conclusions and future directions

Lightweight Vertical Project

- Collecting Tail-Vertical Knowledge by Crowd-Sourcing

Step 1. Decide interesting tail verticals and up to 3 sources for each vertical

Step 2. Have the crowd collect triples from the given sources through annotation tools

Lightweight Vertical Project

- Collecting Tail-Vertical Knowledge by Crowd-Sourcing

Step 1. Decide interesting tail verticals and up to 3 sources for each vertical

Step 2. Have the **crowd** collect triples from the given sources through annotation tools

Step 3. Heavy curation to reach 99.9% precision

Knowledge Collected on Tail Verticals

- Knowledge in 100+ tail verticals
 - 2.2M triples
 - 10K entities, ~700 predicates
 - millions of daily registered users
- Most vs. least popular vertical



Pikachu

Pokemon

Pikachu are a species of Pokémon, fictional creatures that appear in an assortment of video games, animated television shows and movies, trading card games, and comic books licensed by The Pokémon Company, a Japanese corporation. [Wikipedia](#)

Species: Mouse

Type: Electric

Abilities: Static

Weaknesses: Ground

Evolves from: [Pichu](#)

Evolves to: [Raichu](#)

Ability (hidden): [Lightning Rod](#)

VLDB 2014

VLDB conference

City: [Hangzhou](#)

Country: [China](#)

Conference number: 40

Conference date: September 1, 2014 – September 5, 2014

PC chairs: [Aoying Zhou](#), [H. V. Jagadish](#)

Vice chairs: [Divesh Srivastava \(Tutorial\)](#), [Xiaoyong Du \(Tutorial\)](#), [More](#)

PC members industrial: [Sailesh Krishnamurthy](#), [Ashok Joshi](#), [More](#)

Challenges in Lightweight Verticals

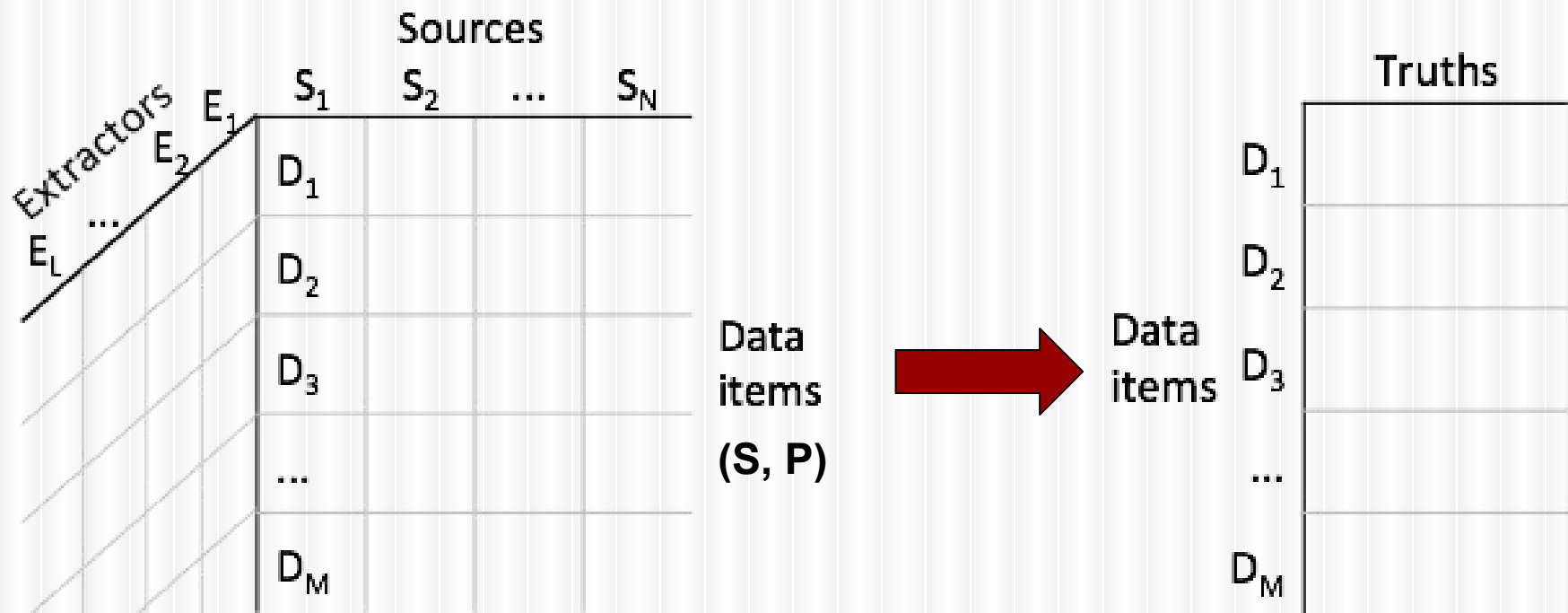
- Challenge 1. Find interesting verticals and relevant high-quality sources
- Challenge 2: Again, how to detect errors?

Outline

- I. Gap between existing KBs and knowledge in the world
- II. Efforts to collect tail knowledge
 - Knowledge Vault
 - Lightweight verticals
- III. Key techniques: knowledge fusion and knowledge-based trust
- IV. Conclusions and future directions

Knowledge Fusion: Decide 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



Model I. Single-Truth Model [VLDB, 2009]

Researcher affiliation

	Prov1	Prov2	Prov3
Jagadish	UM	ATT	UM
Dewitt	MS	MS	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

Model I. Single-Truth Model [VLDB, 2009]

Researcher affiliation


	Prov1	Prov2	Prov3
Jagadish	UM	ATT	UM
Dewitt	MS	MS	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

Voting--Trust the majority.

Model I. Single-Truth Model [VLDB, 2009]

.....


Researcher affiliation



	Prov1	Prov2	Prov3
Jagadish	UM	ATT	UM
Dewitt	MS	MS	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

Model I. Single-Truth Model [VLDB, 2009]

.....
Researcher affiliation



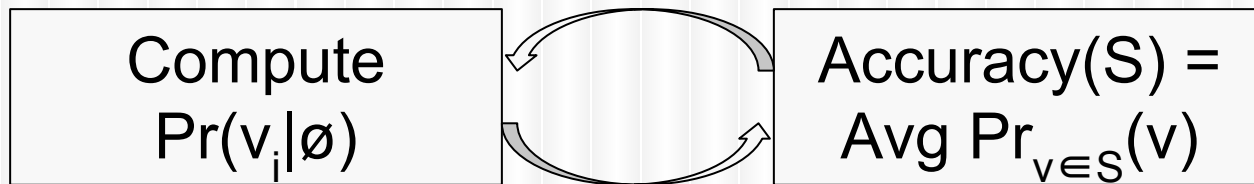
	Prov1	Prov2	Prov3
Jagadish	UM	ATT	UM
Dewitt	MS	MS	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	UCB	UCB	UMD

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

Single-Truth Model and Applications

.....

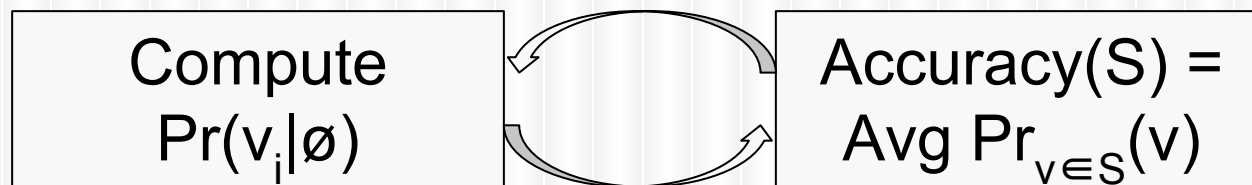
- Single-truth model: For each data item D with values v_1, \dots, v_n , $\Pr(v_1) + \dots + \Pr(v_n) = 1$



Single-Truth Model and Applications

.....

- Single-truth model: For each data item D with values v_1, \dots, v_n , $\Pr(v_1) + \dots + \Pr(v_n) = 1$



- Application: *Personal knowledge extraction from emails*
 - Prec = 0.999, Rec = 0.993
 - Remove 84% errors by rule-based fusion

Model II. Multi-Truth Model

[Sigmod, 2014]

Harry Potter actors/actresses

Harry Potter	Prov1	Prov2	Prov3
Daniel	✓	✓	✓
Emma	✓		✓
Rupert	✓	✓	
Jonny	✓		
Eric			✓

Model II. Multi-Truth Model

[Sigmod, 2014]

Harry Potter actors/actresses

Harry Potter	Prov1	Prov2	Prov3
Daniel	✓	✓	✓
Emma	✓		✓
Rupert	✓	✓	
Jonny	✓		
Eric			✓

Voting--Trust the majority.

Model II. Multi-Truth Model

[Sigmod, 2014]

Harry Potter actors/actresses

Harry Potter	Prov1 (high rec)	Prov2 (high prec)	Prov3 (med prec/rec)
Daniel	✓	✓	✓
Emma	✓		✓
Rupert	✓	✓	
Jonny	✓		
Eric			✓

Model II. Multi-Truth Model

[Sigmod, 2014]

Harry Potter actors/actresses

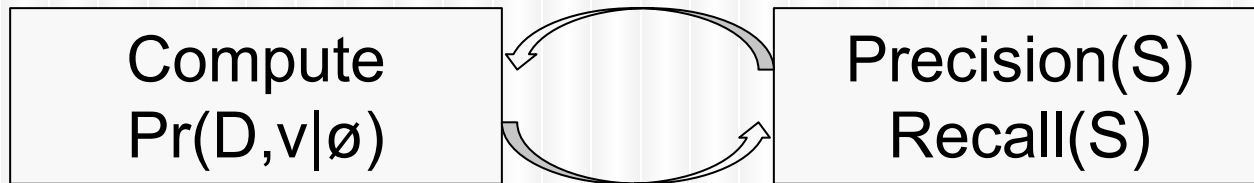
Harry Potter	Prov1 (high rec)	Prov2 (high prec)	Prov3 (med prec/rec)
Daniel	✓	✓	✓
Emma	✓		✓
Rupert	✓	✓	
Jonny	✓		
Eric			✓

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

Multi-Truth Model and Applications

.....

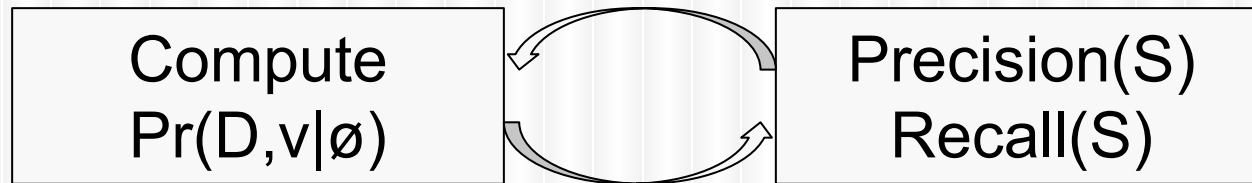
- Multi-truth model: For each data item D with values v_1, \dots, v_n , compute $\Pr(v_i)$
($\Pr(v_1), \dots, \Pr(v_n)$ are independent)



Multi-Truth Model and Applications

.....

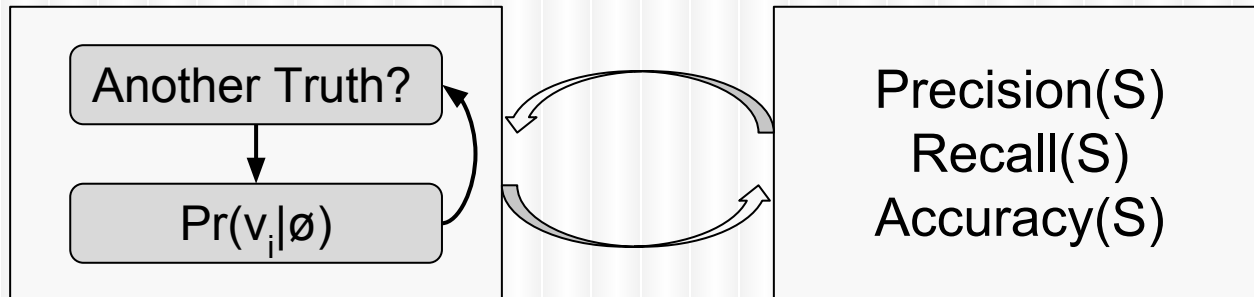
- Multi-truth model: For each data item D with values v_1, \dots, v_n , compute $\Pr(v_i)$
($\Pr(v_1), \dots, \Pr(v_n)$ are independent)



- Application: *Entity type identification*
 - $\text{Prec} = 0.91, \text{Rec} = 0.98$

Model III. Hybrid Model

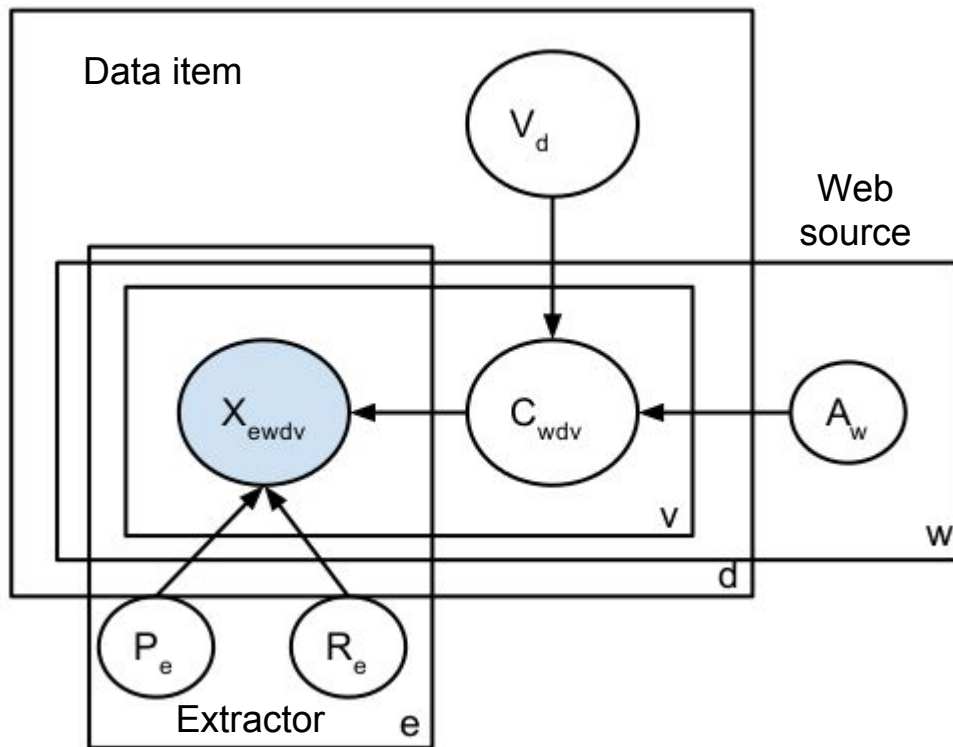
- Hybrid model: for each non-functional predicate, the number of truths is often in a range



- Application: *Fact check for lightweight verticals*
 - Evidence quality: Prec = 0.2, Rec = 0.65
 - Final result quality: Prec = 0.85, Rec = 0.5

Model IV. Multi-Layer Model [VLDB, 2015]

- Multi-layer model
 - Separate source quality and extractor quality
 - Distinguish data errors and extraction errors



Observations

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

Latent variables

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

Parameters

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

Model IV. Multi-Layer Model [VLDB, 2015]


- Multi-layer model
 - Separate source quality and extractor quality
 - Distinguish data errors and extraction errors
- Application 1: *Knowledge Vault*

	PR Area-under-curve	ROC Area-under-curve
Logistic Regression	0.795	0.886
LR + KF-models	0.886	0.937

Multi-Layer Model Application II.

Website Recommendation by Vertical


Sula

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

Multi-Layer Model Application II.

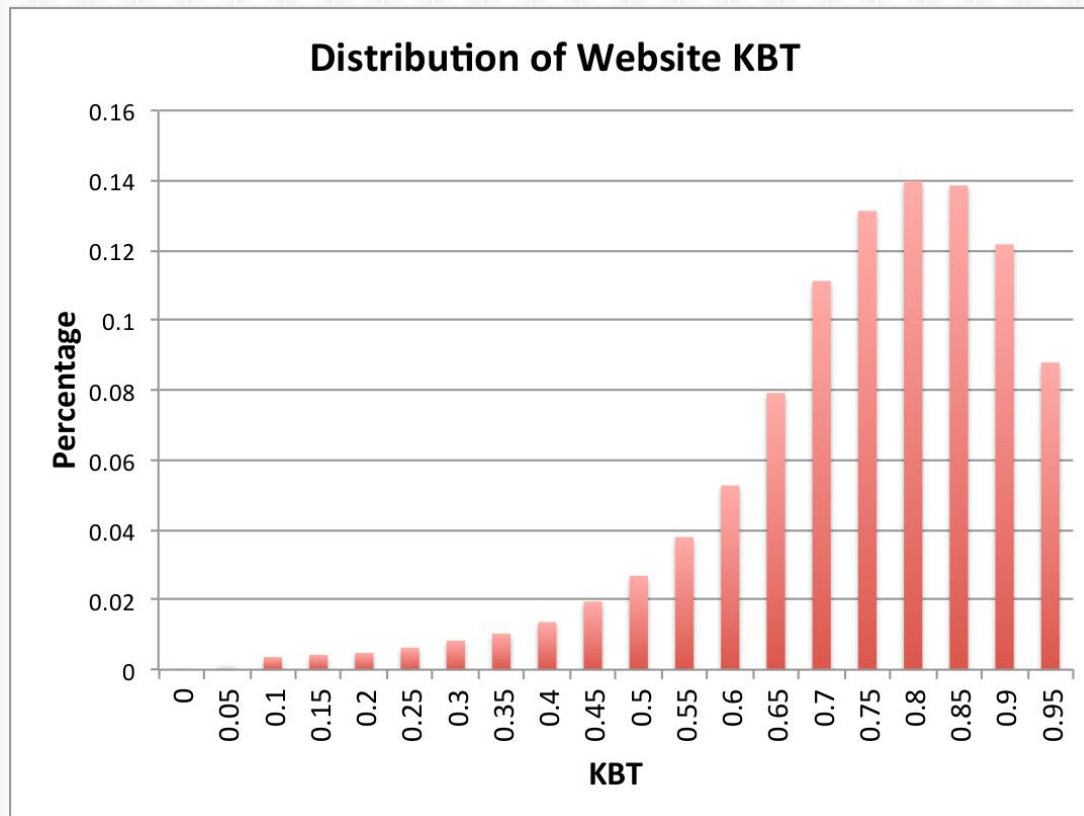
Website Recommendation by Vertical

Sula

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

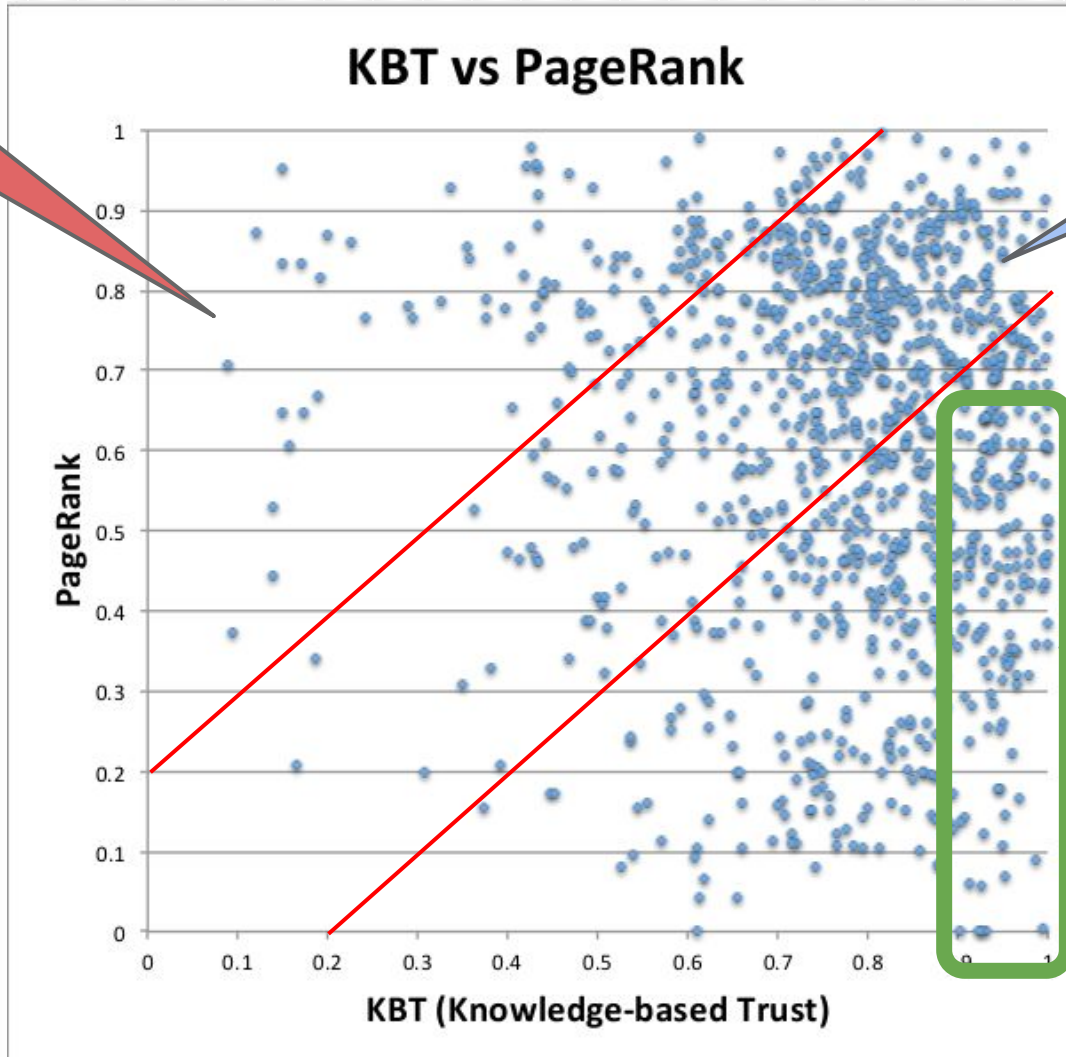
Multi-Layer Model Application II. Knowledge-Based Trust (KBT)

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



Knowledge-Based Trust vs. PageRank

Often sources w. low accuracy



Correlated scores

Often tail sources w. high trustworthiness

KBT for Gossip Websites

Gossip Websites

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

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

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

All have knowledge-based trust in bottom 50%

KBT for Social-Media Webpages

YAHOO!
ANSWERS

Entertainment & Music > Celebrities



Why are British women so unattractive?

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

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

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

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

☆ Follow 37 answers

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

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

TurboTax Sponsored

[California Programs Contribute to STEM Careers](#)

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

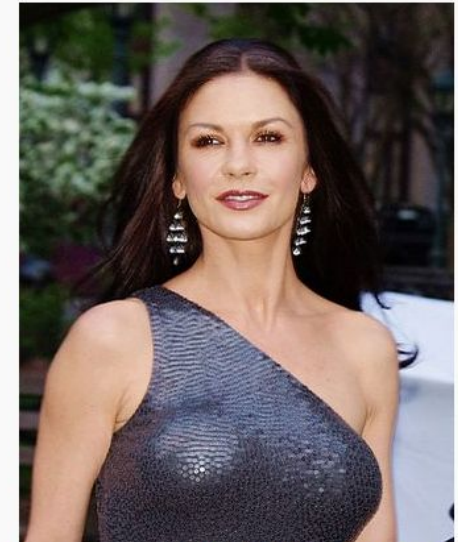
CBS Local Sponsored



WIKIPEDIA
The Free Encyclopedia

Catherine Zeta-Jones

CBE



Zeta-Jones at the 2012 Tribeca Film Festival

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

Knowledge Fusion & Knowledge-Based Trust in Media

Google wants to rank websites based on facts not links

The trustworthiness of a web page might help it rise in search engine rankings. Google wants to measure quality by facts, not just links.

Google has developed a new meaning to ranking sites based on

Breakthrough: whether 'fact-based' or 'Orwellian' knowledge
The huge implications of Google's idea to rank sites based on their accuracy

Why some people are so terrified by the idea of a Google truth machine

Direction I. Open IE for DOM Extraction

.....

- Huge volume of knowledge (often new entities and attributes) in Web DOM trees

Dumbbell Lying Pronation

Type: Strength

Main Muscle Worked: Forearms

Other Muscles:

Equipment: Dumbbell

Mechanics Type: Isolation

Level: Intermediate

Sport: No

Force: Pull

Barbell Lying Cambered Row

1



Main Muscle Group : Back

Detailed Muscle Group : Lats

Other Muscle Groups : Biceps

Type : Strength

Mechanics : Isolation

Equipment : Barbell , Bench

Difficulty : Beginner

Direction II. Keep Knowledge Up-to-date

Lee Sedol



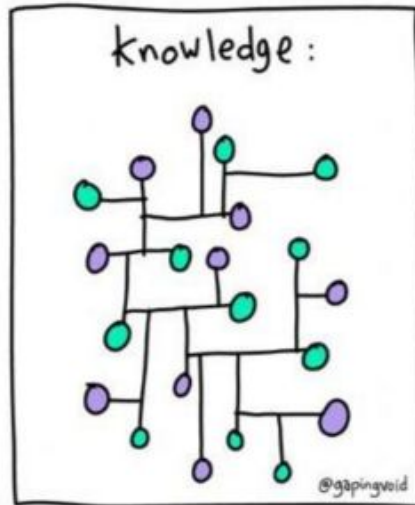
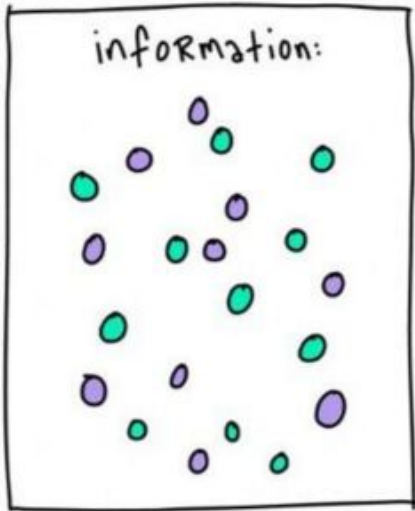
Lee Sedol during a game in 2012

Hangul 이세돌
Hanja 李世돌
Revised Romanization I Sedol
McCune–Reischauer Ri Sedol
Born 2 March 1983
 (age 33)^[1]
 Sinan County, South
 Jeolla Province,
 South Korea
Residence  South Korea
Teacher Kweon Kab-yong^[1]
Turned pro 1996
Rank 9 dan^[1]
Affiliation Hanguk Kiwon^[1]

International		
Asian TV Cup	4 (2007, 2008, 2014, 2015)	1 (2009)
LG Cup	2 (2003, 2008)	2 (2001, 2009)
BC Card Cup	2 (2010, 2011)	
Samsung Cup	4 (2004, 2007, 2008, 2012)	1 (2013)
Chunlan Cup	1 (2011)	1 (2013)
Fujitsu Cup	3 (2002, 2003, 2005)	1 (2010)
World Oza	2 (2004, 2006)	
Zhonghuan Cup		1 (2005)
Mlily Cup (梦百合杯)		1 (2015/16)

As of 3/12/2016

Direction III. Finding Connections in Knowledge



John Adams

2nd U.S. President

John Adams, Jr. was an American lawyer, author, statesman, and diplomat. He served as the second President of the United States, the first Vice President, and as a Founding Father was a leader of American independence from Great Britain. [Wikipedia](#)

Born: October 30, 1735, Braintree, MA

Died: July 4, 1826, Quincy, MA

Vice president: Thomas Jefferson (1797–1801)

Predecessor: George Washington

Presidential term: March 4, 1797 – March 4, 1801

Children: John Quincy Adams, Abigail Adams Smith, Charles Adams, Thomas Boylston Adams, Susanna Adams



Thomas Jefferson

3rd U.S. President

Thomas Jefferson was an American Founding Father who was principal author of the Declaration of Independence. He was elected the second Vice President of the United States, serving under John Adams and in 1800 was elected third President. [Wikipedia](#)

Born: April 13, 1743, Shadwell, VA

Died: July 4, 1826, Charlottesville, VA

Predecessor: John Adams

Children: Martha Jefferson Randolph, Estlin Hemings, More

Quotes

[View 7+ more](#)

The tree of liberty must be refreshed from time to time with the blood of

Direction IV. Complex Knowledge

- Simple fact: *The X (nationality) of Y (Barack Obama) is Z (American)*
- Complex fact: *Most of the global warming in the last 50 years was caused by human beings, to a high degree of certainty.*
- Common sense: Birds can fly

TAKE AWAYS

- We have collected rich knowledge for head entities in head verticals by curation and importing from big data sets
- Collecting tail knowledge is challenging
 - Collecting new entities/attributes is hard
 - Knowledge Vault & Lightweight Verticals
 - Knowledge fusion is a KEY
- There is still a long way to go

Acknowledgement

● Google

Fan Bu

Van Dang

Evgeniy Gabrilovich

Jeremy Heitz

Wilko Horn

Kevin Murphy

Kevin Lerman

Camillo Lugaresi

Shaohua Sun

Ali Tamur

Wei Zhang

Sreeram Balakrishnan

Shawn Jeffrey

Anno Langen

Yang Li

Rod McChesney

Crystal Sno Riley

Mike Shwe

Anna Wolferman

● DB Comm

Laure Berti-Equille

Anish Das Sarma

Furong Li

Xian Li

Kenneth Lyons

Alexandra Meliou

Weiyi Meng

Ravali Pochampally

Barna Saha

Divesh Srivastava

Xiaolan Wang

Bo Zhao

THANK YOU!

Questions?

vldb 2014



[All](#) [Shopping](#) [Videos](#) [News](#) [Images](#) [More ▾](#) [Search tools](#)

About 328,000 results (0.52 seconds)

VLDB2014 - Conference Overview

www.vldb.org/2014/ ▾ VLDB Endowment Inc. ▾

VLDB2014 will take place at Hangzhou, which is one of the best tourism cities in China. Hangzhou is also one of the eight ancient capitals in Chinese history ...

VLDB2014 - Accepted Papers

Accepted Papers (Proceedings of the VLDB Endowment, Volume ...

Demo Papers

Accepted Demo Papers. X-LiSA: Cross-lingual Semantic ...

Detailed Program

VLDB 2014 Program. At a Glance. Monday AT A GLANCE, SEP 1 ...

Research Track

Accepted papers will form the Research Track for the 2014 ...

Program at a Glance

VLDB 2014 Program Overview ... Monday AT A GLANCE, SEP ...

Demonstrations

Choose the "Demonstrations" track for your demo paper submission ...

[More results from vldb.org »](#)

VLDB 2014

VLDB conference

City: Hangzhou

Country: China

Conference number: 40

Conference date: September 1, 2014 – September 5, 2014

PC chairs: Aoying Zhou, H. V. Jagadish

Vice chairs: Divesh Srivastava (Tutorial), Xiaoyong Du (Tutorial), [More](#)

PC members industrial: Saalesh Krishnamurthy, Ashok Joshi, [More](#)

[Feedback](#)

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



The screenshot shows a web browser window with the URL `en.wikipedia.org/wiki/Donna_Summer`. The article text includes a paragraph about her influences and a paragraph about her performance at the Nobel Peace Prize Concert in Oslo, Norway, in honor of American President Barack Obama. A blue highlight is placed over the text "American President Barack Obama". A callout box on the left points to this text with the text "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)

Obama born
in Kenya



Predicting Extraction and Triple Correctness

- (Obama, nationality, Kenya)

2087 extractions:

- Example of a correct extraction ($Pr_extCorr=0.792$)

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

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

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

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

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

- Example of a wrong extraction ($Pr_extCorr=0.130$)

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

US will respect winner of
Kenya election, Obama says

[SHARE](#) [BOOKMARK](#) [PRINT](#) [RATING](#) ☆☆☆☆☆

- $Pr_tripleCorr=0$ (not enough support)

Predicting Extraction and Triple Correctness

- (Obama, nationality, USA)

2481 extractions:

- Example of a correct extraction ($Pr_extCorr=0.999$)

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

- Example of a wrong extraction ($Pr_extCorr=0.261$)

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

Tim Stanley

Dr Tim Stanley is a historian of the United States. His new book about Hollywood politics is out in May. His personal website is www.timothystanley.co.uk and you can follow him on Twitter @timothy_stanley.

 Follow 15.6K followers



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

- $Pr_tripleCorr=1$ (higher support)

Predicting Extraction and Triple Correctness

Distribution of providers for Kenya and USA

