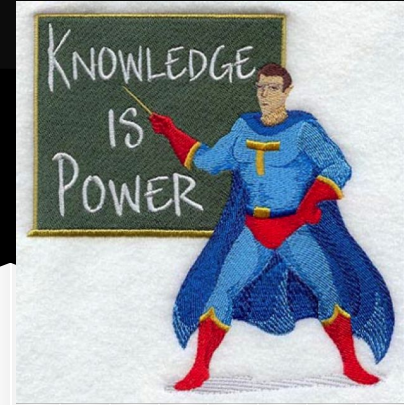


Leaving No Valuable Data Behind: the Crazy Ideas and the Business

Xin Luna Dong
Amazon ML, 2017

Knowledge Is Power



- Many Knowledge Bases (KB)



NELL: Never-Ending Language Learning



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](http://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 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 ...

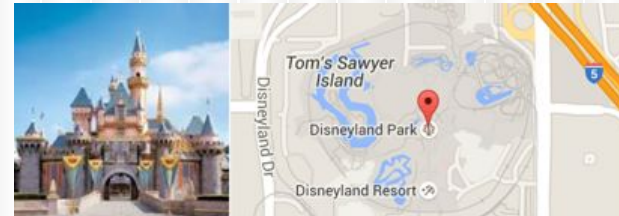

Sen. Bernie Sanders won in Kansas and Nebraska, while Donald Trump took the Louisiana GOP contest, but Ted Cruz secur...



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 ...

Using KB in Personal Assistance

.....

Alexa, play the music by Michael Jackson



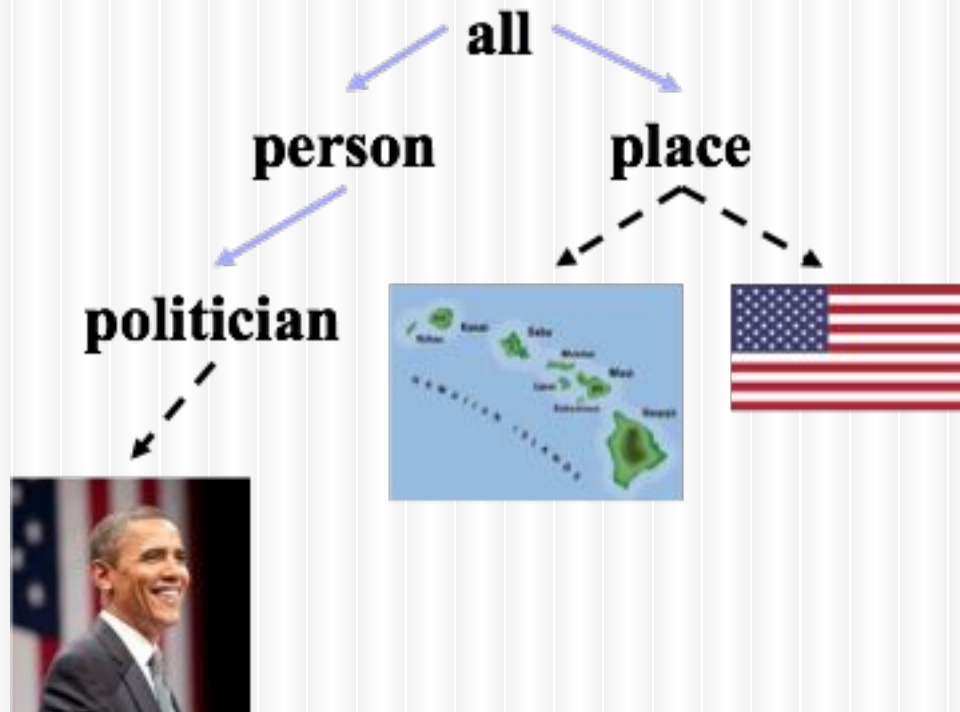
List of officially released compilations and

[\[92\]](#)[\[93\]](#)[\[94\]](#)[\[95\]](#)

- *Portrait of Michael Jackson / Portrait of Jackson 5* (1973)
- *Os Grandes Sucessos, Vol. 2* (1980)
- *Motown Superstar Series, Vol. 7* (1980)
- *Superstar* (1980)
- *Michael Jackson & The Jackson 5* (1983)
- *Ain't No Sunshine* (1984)
- *The Great Love Songs of Michael Jackson* (1984)
- *Ben / Got to Be There* (1986)
- *Looking Back to Yesterday* (1986)
- *The Original Soul of Michael Jackson* (1987)
- *Rockin' Robin* (1993)
- *Dangerous – The Remix Collection* (1993)
- *Michael Jackson Story* (1996)
- *Master Series* (1997)
- *Ghosts – Deluxe Collector Box Set* (1997)
- *Got to Be There / Forever, Michael* (1999)
- *Bad / Thriller* (2000)
- *Forever, Michael / Music & Me / Ben* (2000)
- *Classic – The Universal Masters Collection* (2001)

What is a Knowledge Base

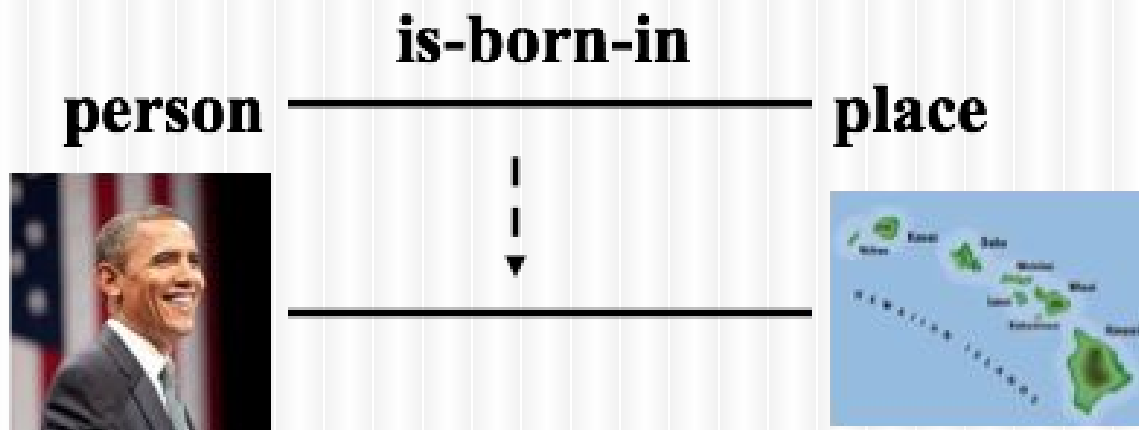
- Entities, entity types
 - An entity is an instance of multiple types
 - Entity types organized in a hierarchy



What is a Knowledge Base

.....

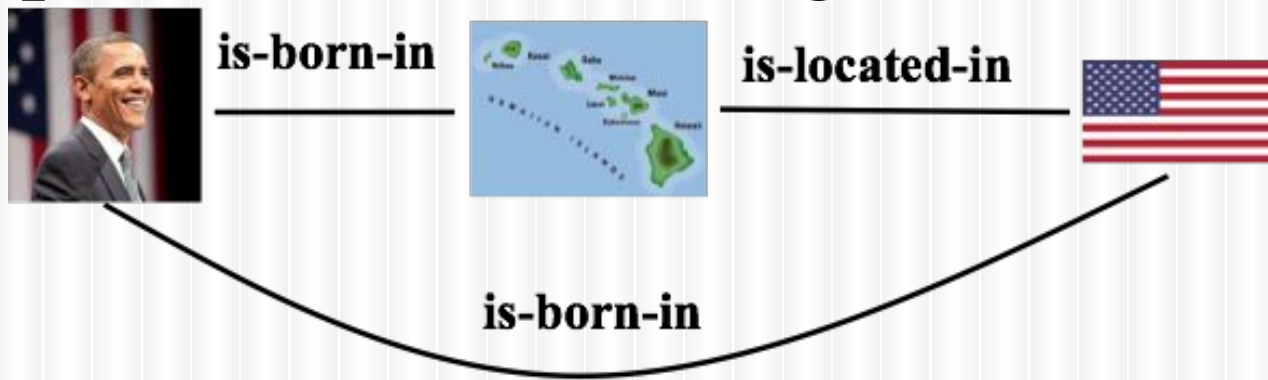
- Entities, entity types
- Predicates, (sub, pred, obj) triples
 - A triple describes an attribute of an entity, or, the relationship between two entities



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

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	70B

Freebase Statistics

- 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%)

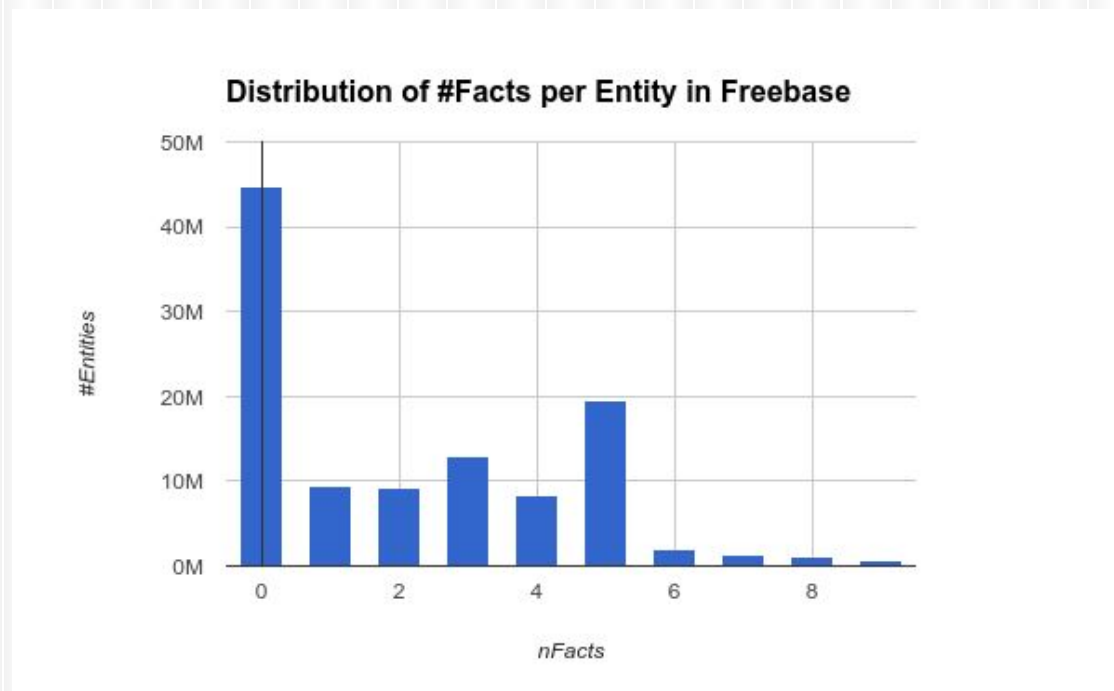
Category I. Head Entities in Head Verticals

- Rich knowledge for head entities in head verticals. E.g., (Freebase)

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
 - 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

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	Big Data Integration
/book/author/works_written	

[/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.node

Properties Types & Collections Names Keys Triple

Mode: Formatted Raw

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	Providing Best Effort Services in Dataspace Systems
/book/author/works_written	

/people [view schema](#)

Person [/people/person](#)

Profession	Mathematician
/people/person/profession	

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	Publication City/Country: London, United Kingdom
Dimensions: 157.48 x 233.68 x 15.24mm 317.51g	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)
 - 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,[[]



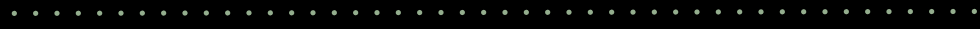
Faceted aquamarine, 13.24ct, Brazil

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?

Mission

A horizontal dotted line is positioned below the title. Below the dotted line is a decorative scalloped border that separates the title area from the main content area.

Mission of a data-integration researcher:

Leaving NO Valuable Data Behind



Science is to test crazy ideas;
Engineering is to bring these
ideas into Business

–Andreas Holzinger

The Crazy Ideas

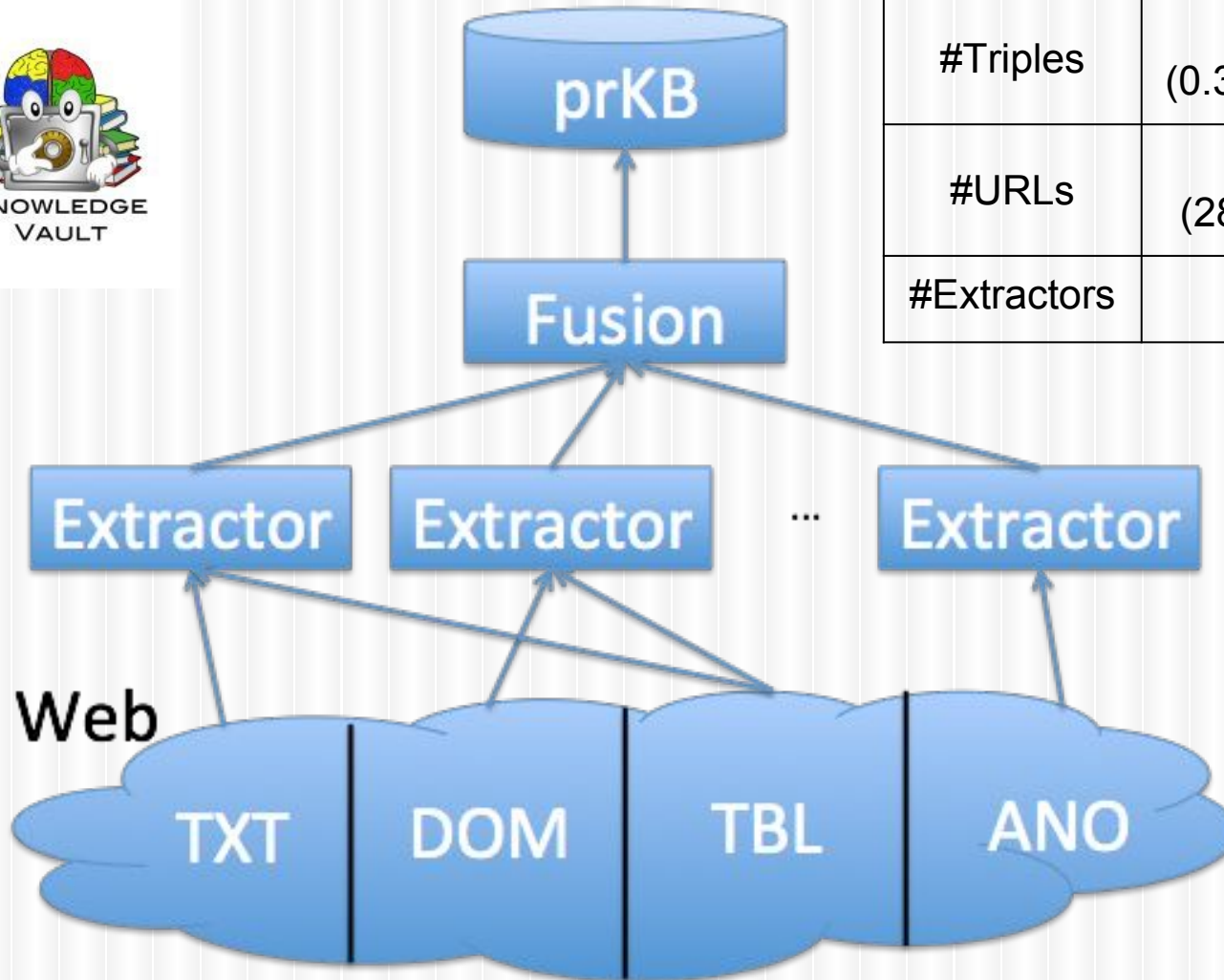
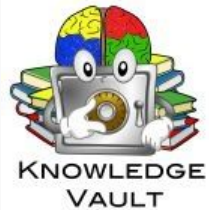
Science is to test crazy ideas

Crazy Idea I. Knowledge Vault

- Conventional approach
Manually curate knowledge from a few major data sources; e.g., Google Knowledge Graph
- Knowledge Vault
Automatically extract knowledge from the Web

Knowledge Vault: Automatically extracting knowledge from Web

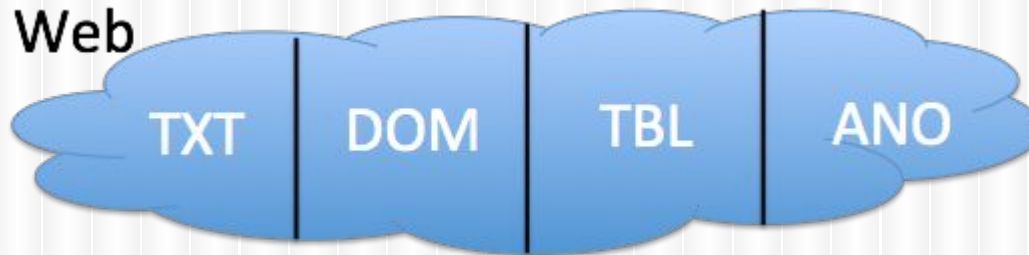
[SIGKDD, 2014]
[VLDB, 2014]



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

Four Types of Web Sources

Web



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 include *The Last Supper*, *Leda and the Swan* and many others -- influenced countless artists and made da Vinci a key figure of the Italian Renaissance.

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	6/27/1836

Annotations

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

schema.org

DOM Tr

yelp

Search for (e.g. 'pizza') in (e.g. 'New York')

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

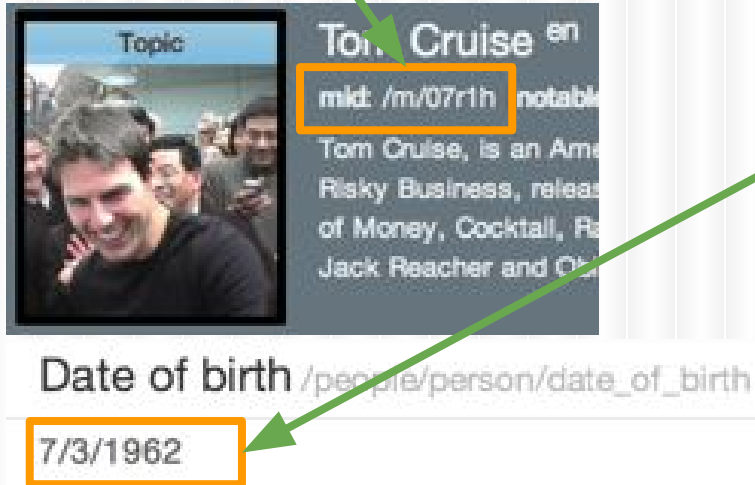
Alcohol: Full Bar
Noise Level: Average
Ambiance: Trendy, Casual
Has TV: No
Caters: No
Wheelchair Accessible: Yes

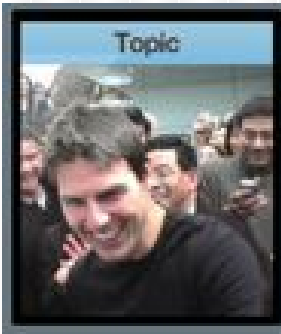
Food Photos

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

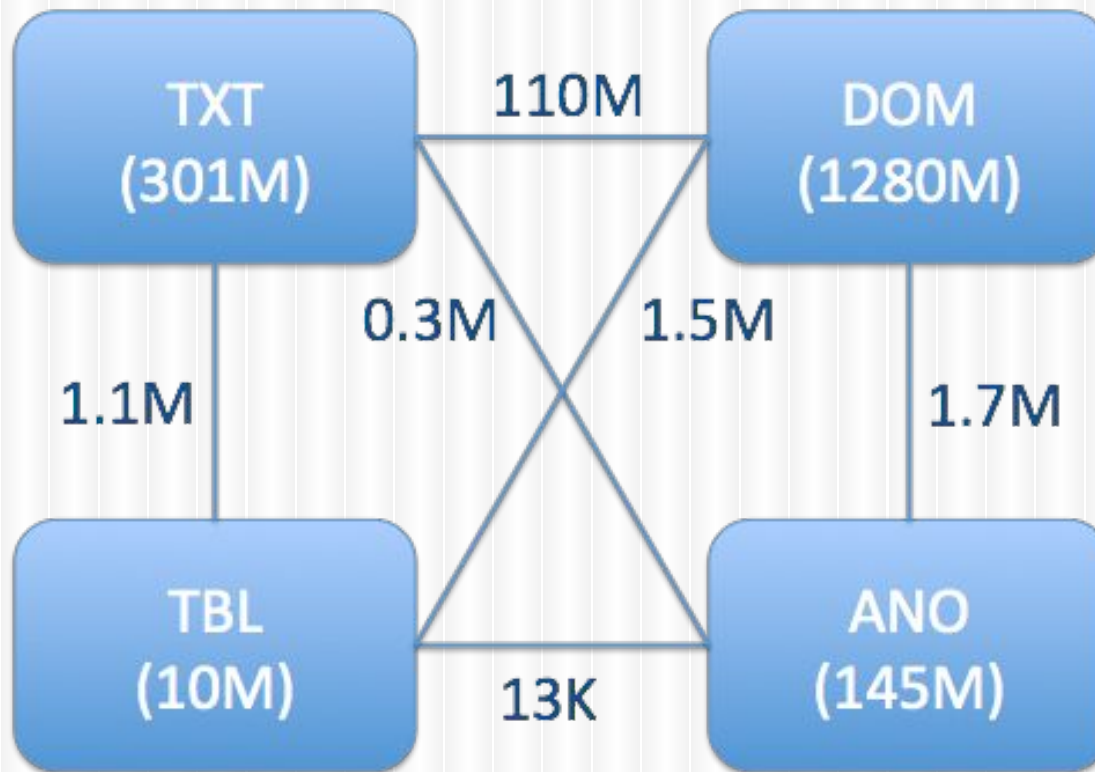


Topic	Tom Cruise ^{en}
	Tom Cruise ^{en} mkt: /m/07r1h notable Tom Cruise, is an Ame Risky Business, releas of Money, Cocktail, Ra Jack Reacher and Oth
Date of birth	/people/person/date_of_birth
	7/3/1962

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

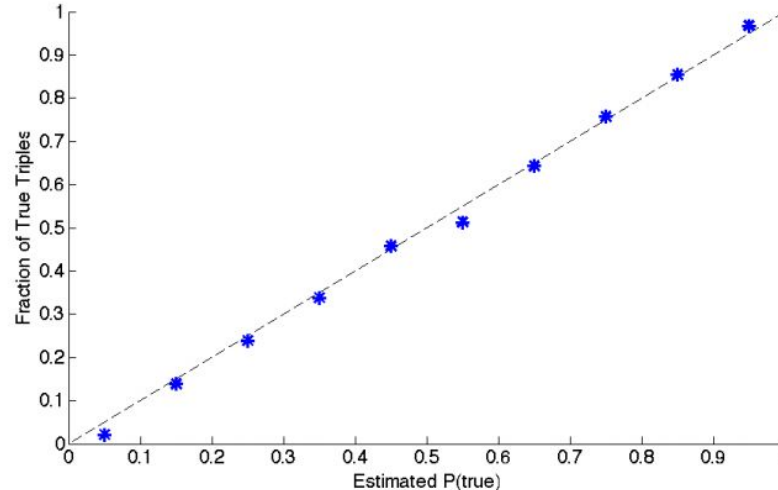


Figure 1: True probability vs estimated probability for each triple in KV.

Crazy Idea II. Knowledge-Based Trust

- Conventional approaches
Evaluate trustworthiness of sources by **exogenous signals**: hyperlinks, click-rate, etc.; e.g., PageRank
- Knowledge-based trust
Evaluate by **endogenous signals**: the correctness of its factual information

Crazy Idea II: Knowledge-Based Trust: Evaluating Trustworthiness of Factual Info



WIKIPEDIA
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia
- Wikimedia Shop
- Interaction
 - Help
 - About Wikipedia
 - Community portal
 - Recent changes
 - Contact page
- Tools
- Print/export
- Languages
 - Acèh
 - Адыгэбзэ
 - Afrikaans
 - Alemannisch
 - አማርኛ
 - Ænglisc
 - Англис
 - Англис
 - العربية
 - Aragonés

Create account Log in

Article **Talk** Read **View source** Search

United States

From Wikipedia, the free encyclopedia
(Redirected from USA)

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

The **United States of America** (**USA**), commonly referred to as the **United States (US)**, **America** or simply **the States**, is a federal republic^{[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"






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

Crazy Idea II: Knowledge-Based Trust: Evaluating Trustworthiness of Factual Info



WIKIPEDIA
The Free Encyclopedia

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

- Interaction
 - Help
 - About Wikipedia
 - Community portal
 - Recent changes
 - Contact page

Tools

- Print/export

Languages

- Acèh
- Адыгэбзэ
- Afrikaans
- Alemannisch
- አማርኛ
- Ænglisc
- Аңсәуә
- العربية
- Aragonés
- ᱯᱟᱨᱱᱟᱲ

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"

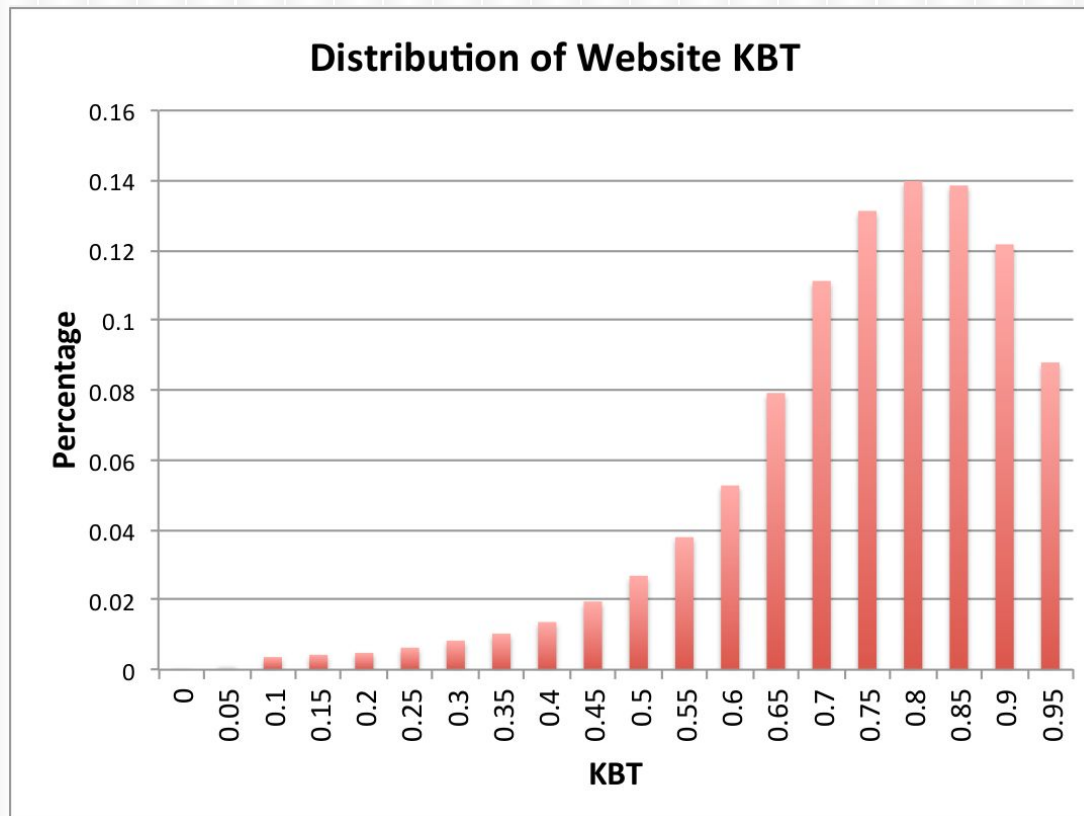





	Triple Corr	Extraction Corr
Triple 1	1.0	1.0
Triple 2	0.9	1.0
Triple 3	0.3	1.0
Triple 4	0.8	1.0
Triple 5	0.4	0.9
Triple 6	0.8	0.9
Triple 7	0.9	0.8
Triple 8	1.0	0.2
Triple 9	0.7	0.1
Triple 10	0.2	0.1
...
Accu	0.73	

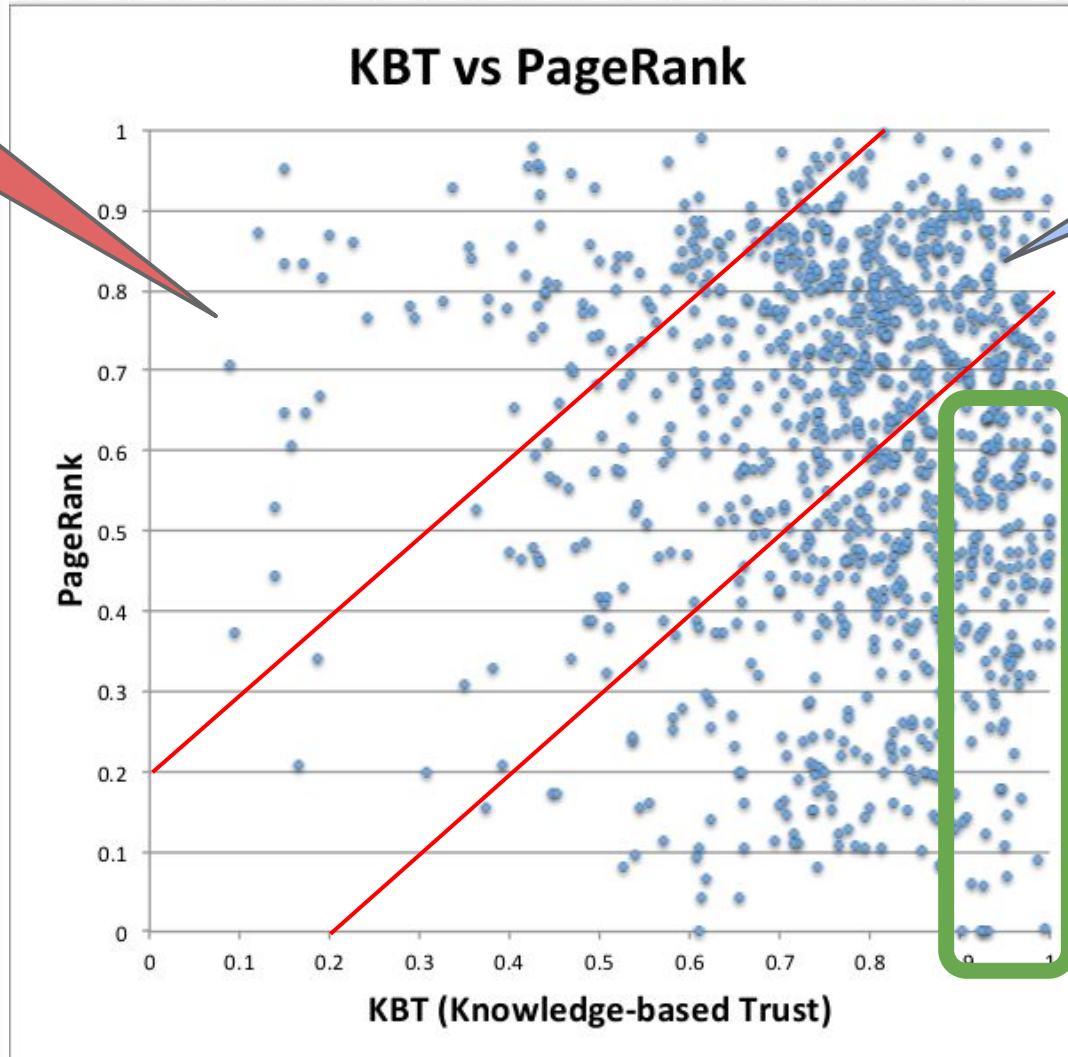
Knowledge-Based Trust (KBT) [VLDB, 2015]

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



Knowledge-Based Trust vs. PageRank

Often sources w. low accuracy



Correlated scores

Often unpopular 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, at their best points, and they would be gorgeous.

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

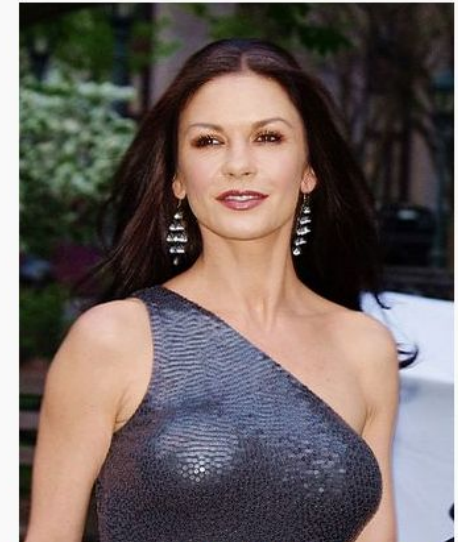
☆ Follow 37 answers

[Are you getting your biggest tax refund?](#)
Get your taxes done right, and your biggest refund, guaranteed. Start for free today!
TurboTax Sponsored

[California Programs Contribute to STEM Careers](#)
California's public school system contributes to STEM careers by offering science-centric activities.
CBS Local Sponsored



Catherine Zeta-Jones CBE



Zeta-Jones at the 2012 Tribeca Film Festival

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

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

Knowledge Vault in Media

News

Google's Knowledge Vault

1.6 billion facts

FELICITY

Google's fact-checking bots build vast knowledge bank

ains

Knowledge Vault, a massive database of the world's facts

ection of all human



The search that could

Database could be the foundation for array of

To Power Future

Good Bye Knowledge Graph,

Hello Google Knowledge Vault?

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

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

Limitations of the Crazy Ideas



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

UNKNOWN
ENTITIES

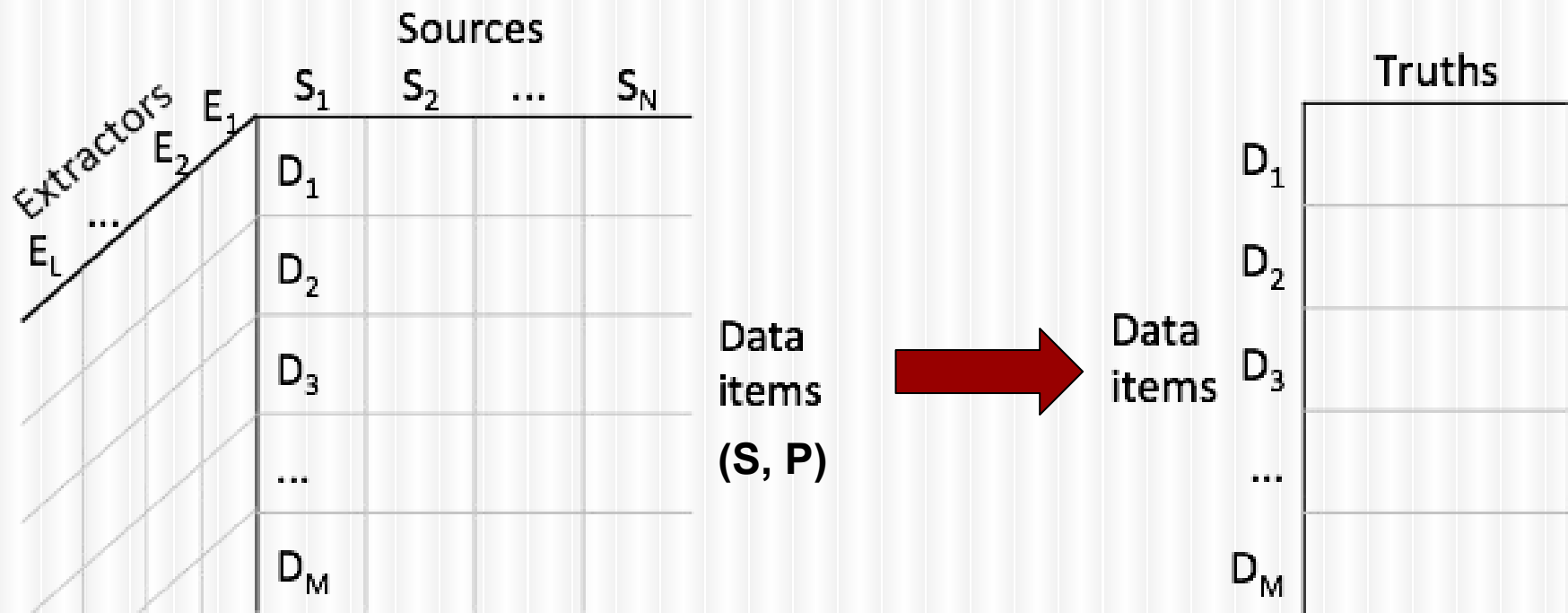
- Focusing on existing entities and attributes
 - Training data contain only FB predicates
 - Entities need to be annotated as FB entities
- KV: Among the 0.3B high-confidence triples
 - 0.18B triples not in KG
 - KG contains 18B triples (100X) [KDD'14]
- KBT: Compute reliable KBT for <20% websites and <<5% webpages

The Business

Engineering is to bring these
ideas into Business

The Model Behind the Crazy Ideas: 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



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

Researcher affiliation

	Prov1	Prov2	Prov3
Jagadish	UM	ATT	UM
Dewitt	FB	FB	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	Chicago	Chicago	UCB

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

Researcher affiliation

	Prov1	Prov2	Prov3
Jagadish	UM	ATT	UM
Dewitt	FB	FB	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	Chicago	Chicago	UCB

Voting--Trust the majority.

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

.....
Researcher affiliation



	Prov1	Prov2	Prov3
Jagadish	UM	ATT	UM
Dewitt	FB	FB	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	Chicago	Chicago	UCB

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

.....
Researcher affiliation

	Prov1	Prov2	Prov3
Jagadish	UM	ATT	UM
Dewitt	FB	FB	UW
Bernstein	MSR	MSR	MSR
Carey	UCI	ATT	BEA
Franklin	Chicago	Chicago	UCB

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

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

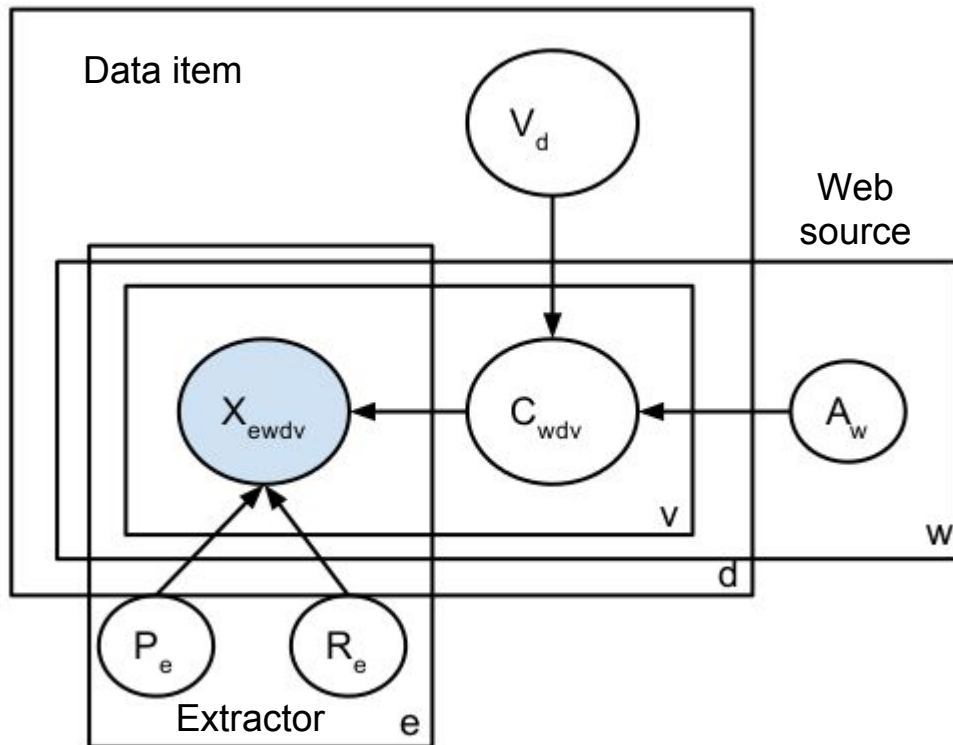
[Sigmod, 2014]

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-Layer Graphical Model



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

Library of Fusion Models

- Application 1: Google Now email extraction
 - Single-truth model
 - $\text{Prec} = 0.999$, $\text{Rec} = 0.993$
 - Remove 84% errors by rule-based fusion
- Application 2: Entity type identification
 - Multi-truth model
 - $\text{Prec} = 0.91$, $\text{Rec} = 0.98$

Lightweight Vertical Project

- Goal: Collecting knowledge for tail verticals (*e.g.*, yoga pose, hindu deity)

Lightweight Vertical Project

- Goal: Collecting knowledge for tail verticals (*e.g.*, yoga pose, hindu deity)
- Method
 - 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

- Goal: Collecting knowledge for tail verticals (*e.g.*, yoga pose, hindu deity)
- Method
 - 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**

Challenges in Lightweight Verticals

- Challenge 1. Find interesting verticals and relevant high-quality sources
- Challenge 2: Detect errors from curation and from sources

Challenges in Lightweight Verticals

- Challenge 1. Find interesting verticals and relevant high-quality sources
- Challenge 2: Detect errors from curation and from sources
 - Solution: Triangulate from 3 web sources
 - Evidence quality: Prec = 0.2, Rec = 0.65
 - Fusion quality: Prec = 0.85, Rec = 0.5

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 tail vertical



Pikachu

Pokémon

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](#)

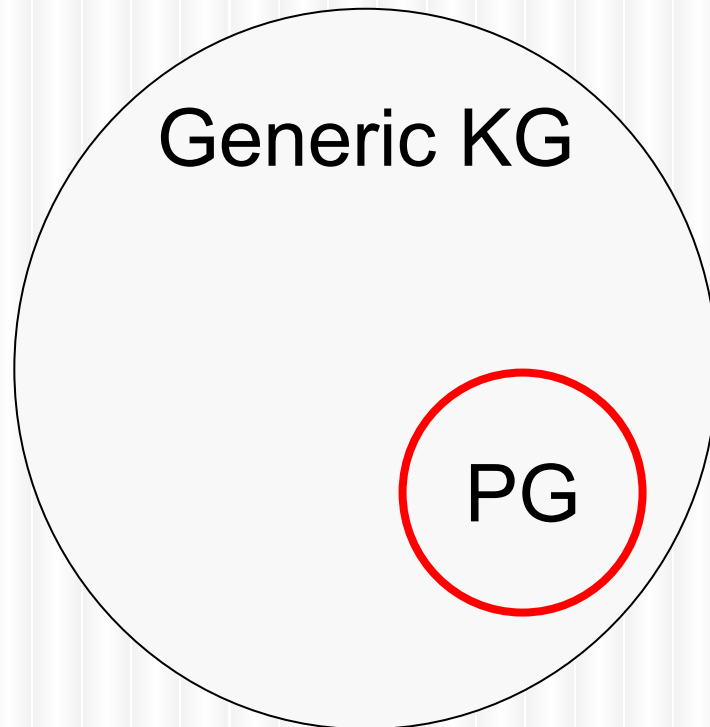
PC members industrial: [Sailesh Krishnamurthy](#), [Ashok Joshi](#), [More](#)

.....

The New Commission

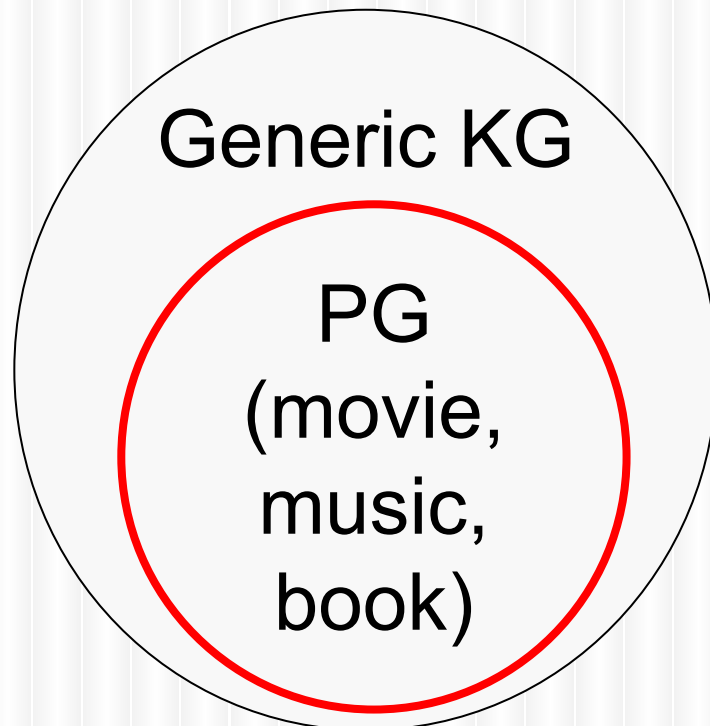
Building a Product Graph at Amazon

- Goal: Build the authoritative knowledge base for every product in the world



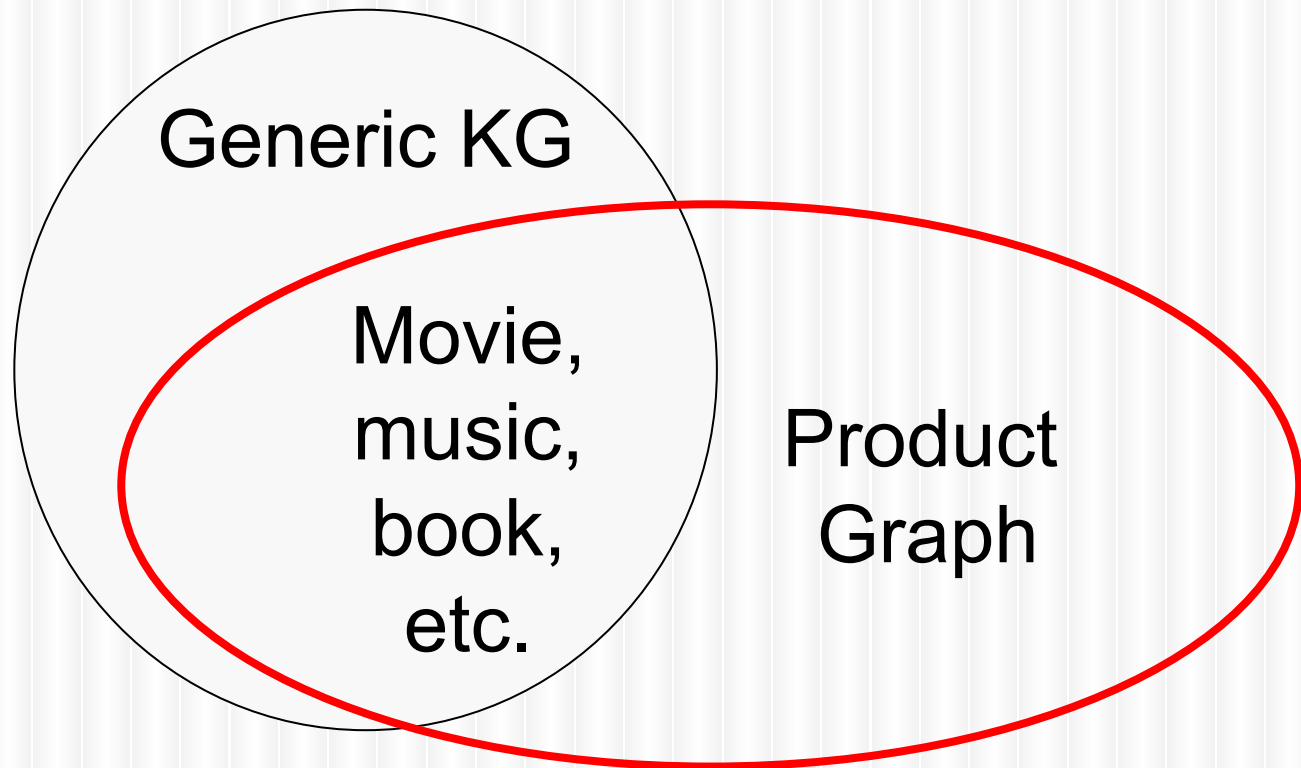
Building a Product Graph at Amazon

- Goal: Build the authoritative knowledge base for every product in the world



Building a Product Graph at Amazon

- Goal: Build the authoritative knowledge base for every product in the world



Challenges in Building Product Graph I

- No major sources to curate product data from
 - Wikipedia does not help too much
 - A lot of structured data buried in text descriptions in Catalog
 - Retailers gaming with the system so even more noisy data

Challenges in Building Product Graph II

- Large number of new products everyday
 - Curation is impossible
 - Freshness is a big challenge

Challenges in Building Product Graph II

- Large number of product categories
 - A lot of work to manually define ontology
 - Hard to catch the trend of new product categories and properties

Our Solution

- Human-in-the-loop knowledge learning
 - Crowd-sourcing evaluation
 - Active learning
 - Close IE vs. Open IE
- Hands-off-the-wheel data integration
 - Machine learning for data quality
 - Building generic systems for data cleaning and data integration

We are HIRING!!

lunadong@amazon.com

THANK YOU!

Questions?

Acknowledgement

● Google

Fan Bu

Van Dang

Evgeniy Gabrilovich

Jeremy Heitz

Wilko Horn

Kevin Lerman

Camillo Lugaresi

Kevin Murphy

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

Xiao Cheng

Anish Das Sarma

Alon Halevy

Furong Li

Xian Li

Kenneth Lyons

Alexandra Meliou

Weiyi Meng

Ravali Pochampally

Barna Saha

Divesh Srivastava

Xiaolan Wang