

Building, Maintaining, and Using Knowledge Bases: A Report from the Trenches

Omkar Deshpande^{1,2}, Digvijay S. Lamba^{1,2}, Michel Tourn¹,
Sanjib Das³, Sri Subramaniam^{1,2}, Anand Rajaraman¹, Venky Harinarayan¹, AnHai Doan^{1,2,3}

¹ Kosmix, ² @WalmartLabs, ³ University of Wisconsin-Madison

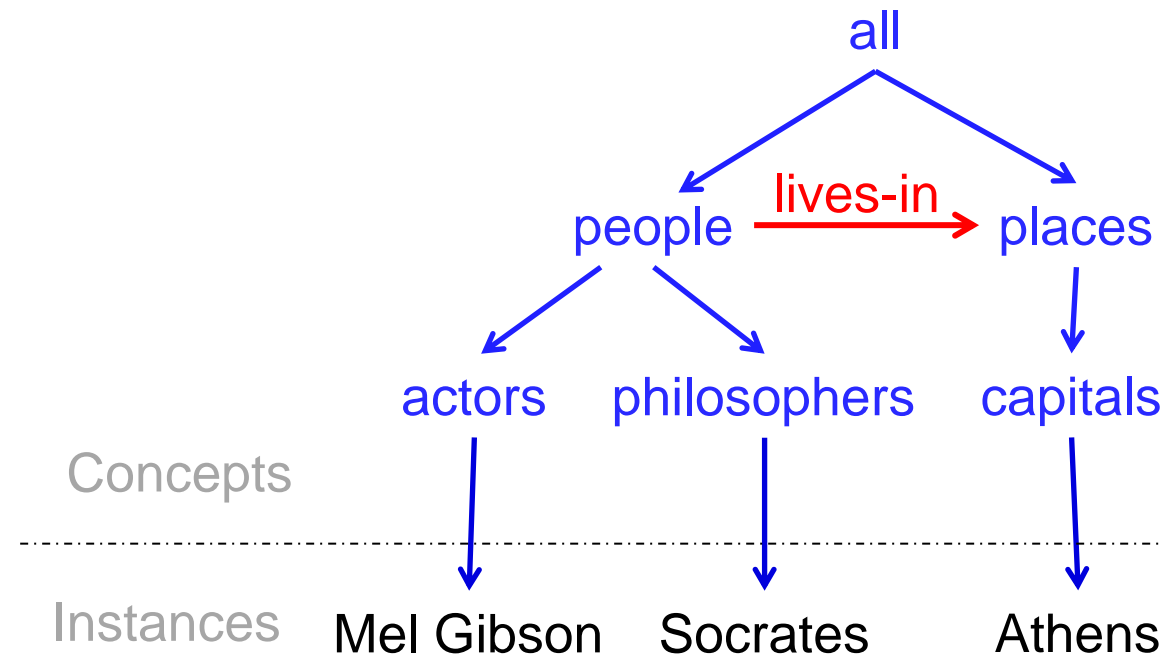
Kosmix

@WalmartLabs



Knowledge Bases (KBs)

- Concept taxonomy
- Instances
- Relationships



Increasingly Critical to a Wide Variety of Applications

- **General search**
 - Google search using Knowledge Graph
- **Product search**
 - Walmart.com, Amazon.com
- **Question answering**
 - IBM Watson, Apple Siri
- **Advertising**
- **Information extraction**
- **Recommendation, playlisting, fingerprinting music (e.g., echonest.com)**
- **Biomedical expert finding (e.g., knode.com)**
- **Data mining in heating and cooling (e.g., Johnson Control)**
- **Deep Web search**
- **Social media analysis (e.g., event discovery, event monitoring)**
- **Social commerce (e.g., social gifting), and many more ...**


Ads related to **las vegas**


Las Vegas | VEGAS.com™ - Las Vegas Shows, Hotels and more
www.vegas.com/
The Official VEGAS Travel Site™.
297 people +1'd this page

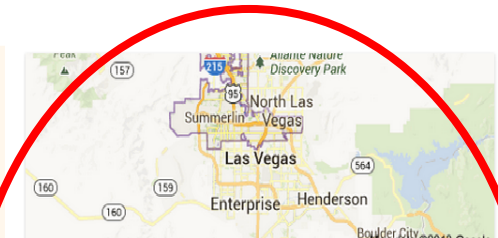
Vegas Show Deals Vegas Air-Hotel Packages
Vegas Hotel Discounts Vegas Weekly Deals and Promotions

Vegas Visitor Guide - Plan your Perfect Vacation Today
lasvegas.travelnevada.com/
Order your Free Visitors Package.
Travel Nevada has 396 followers on Google+
Travel Deals - Events and Shows - View Online Visitor's Guide

Las Vegas - Find the Best Vegas Travel Deals - KAYAK.com
www.kayak.com/
Compare 100s of Sites At Once.

(Official City of Las Vegas Web Site)
www.lasvegasnevada.gov/ 
City government, services, neighborhoods, and businesses.

Las Vegas Hotels, Shows, Casinos, Restaurants, Maps and Things ...
www.lasvegas.com/ 
Your official What happens in Vegas, stays in Vegas® resource. Find details on hotels, restaurants, casinos, events, golf and things to do in Las Vegas.
Shows & Events - Hotels - Special Offers & Deals - Know the Code




Las Vegas

3,365 followers on Google+

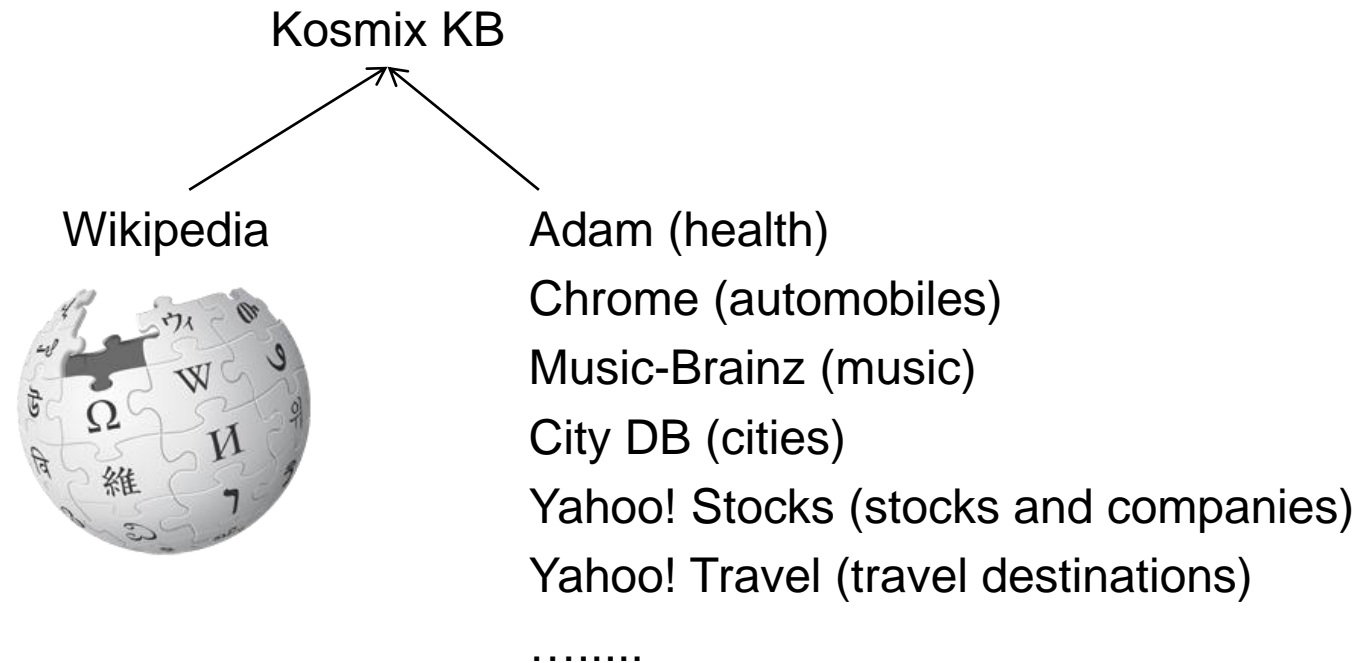
Las Vegas is the most populous city in the U.S. state of Nevada and the county seat of Clark County. Wikipedia

Area: 135.9 sq miles (352 km²)
Founded: May 15, 1905
Weather: 93°F (34°C), Wind S at 4 mph (6 km/h), 5% Humidity
Local time: Monday 11:34 AM
Population: 589,317 (2011)
Points of interest: Las Vegas Strip, Bellagio, Wynn Las Vegas, More

Upcoming events



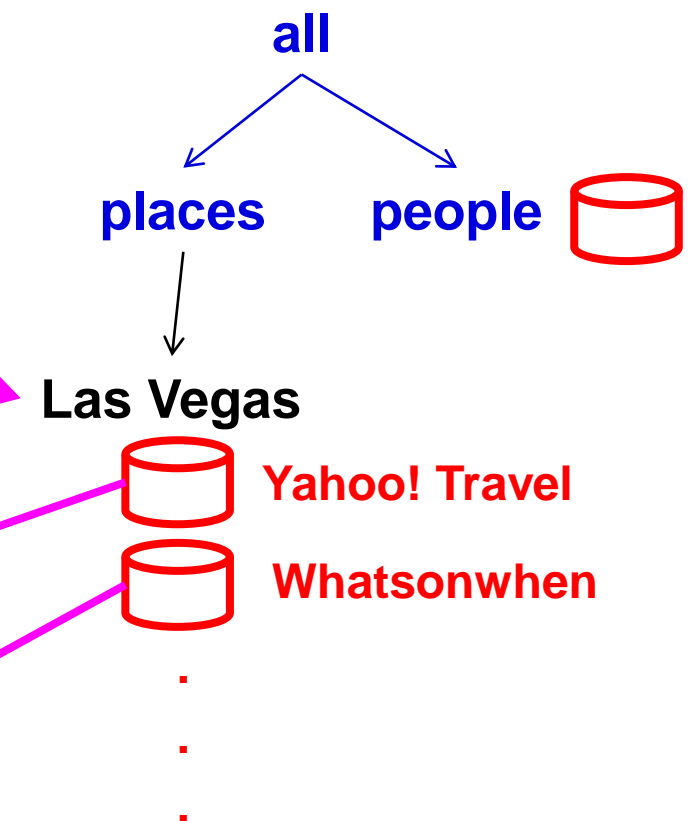
Example Knowledge Base: Kosmix KB



- **6.5M concepts, 6.7M concept instances, 165M relationship instances**
- **23 verticals, 30G of disk space**
- **First built around 2005 at Kosmix**
 - for Deep Web search, advertising, social media analysis
- **Has been significantly expanded at WalmartLabs since 2011**
 - for product search, social commerce, mining of social media, understanding Web data

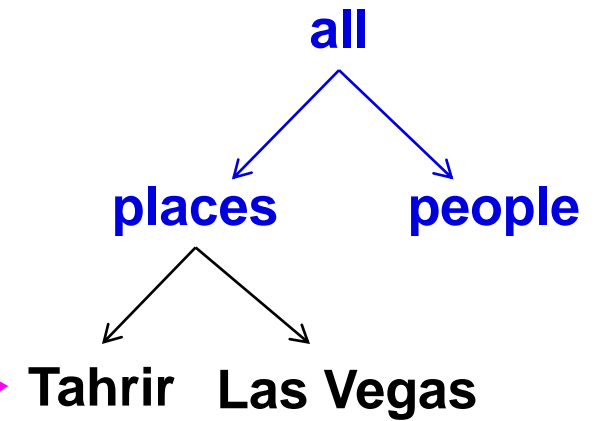
Example Application: Deep Web Search at Kosmix

The screenshot shows the Kosmix website interface. At the top left, the 'Kosmix' logo is visible. A search bar contains the text 'Las Vegas' and is highlighted with a pink circle. To the right of the search bar is a green 'Explore' button. Below the search bar, the page title is 'Las Vegas metropolitan area'. A navigation menu includes 'Overview', 'Reference', 'Hotels & Lodging', 'Images', 'Video', 'Travel Guides', 'Travel Forums', and 'News'. The main content area is divided into several sections: 'Articles' (with a sub-section for 'Kosmix Staff'), 'Hotels & Lodging' (listing 'Plaza Hotel Las Vegas' and 'Trump International Hotel Las Vegas'), and 'Events & Activities' (listing 'La Vega Carnival' and 'Las Vegas International Mariachi Festival'). On the right side, there are advertisements for 'Babson Fast Track MBA', 'Driving in California?', 'Shocking: \$9 Car Insurance', 'Las Vegas Daily Deals', '50%-65% Off Vegas Hotels', and 'Las Vegas' travel deals. At the bottom right, there is an 'Images' section showing two license plates with the numbers '564' and '147'.



Example Application: Event Monitoring in Social Media

The image shows two overlapping screenshots of the ABC News website. The top screenshot displays the main news page with a navigation bar and a featured article titled "EXCLUSIVE: Suleiman: 'Egypt Will Not Be Anything Like Tunisia'". The bottom screenshot shows a "Twitter: Egypt in Real-Time" feed with several tweets. A pink arrow points from the tweet mentioning "Tahrir Square" to a diagram on the right.



Example Application: Social Gifting at WalmartLabs

Walmart Shopycat
Your holiday gift finder

Like 1k
Invite Friends

Home Friends My Info

Type the name of a friend or an interest

Jim K. Alan C. Craig D. Benjamin W. Glen E. Brett J. Sue Z. Waleed A.

Sue Zann Toh

Interests:
DONNIE DARKO FAMILY GUY YOGA MUSIC
CHRISTINA AGUILERA KATY PERRY COMICS PEANUT LABS
RIHANNA FARGO (FILM)

Donnie Darko (Blu-ray) (Widescreen)
★★★★★ Released Feb 10, 2009
\$10.00

Family Guy Clue
★★★★★
\$39.95

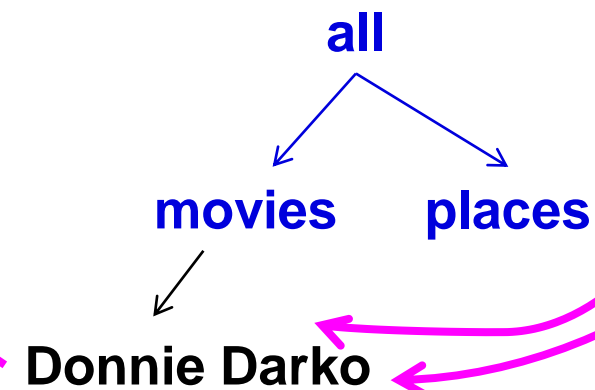
Everything Fits Gym Bag
★★★★★
\$60.00

Apple iTunes Silhouette \$15 Gift Card
★★★★★
\$15.00

Is it bad that I want a Donnie Darko tattoo?

Should 'family guy' end? NEVER

How can you not like Darko?! :o



Paper Overview

- **Important for Big Data**
 - Big Data => Big Semantics => Big KBs
- **Prior works have addressed only isolated aspects:**
 - Initial construction, data representation, storage format, query APIs, ...
- **No work has addressed the end-to-end process**
- **This work: end-to-end process of building, maintaining, using Kosmix KB**
 - How to maintain the KB over time?
 - How to handle human feedback?
 - How to integrate various data sources?
 - What kinds of applications is a not-so-accurate KB good for?
 - How big of a team is required to build such a KB? What should the team do?

Key Distinguishing Aspects of Kosmix KB

● Building the KB

- started with Wikipedia, added many more data sources
- extracting a KB from Wikipedia is non-trivial, use Web and social data / curation to guide the process
- adding a lot of social/Web metadata to KB nodes

● Updating the KB

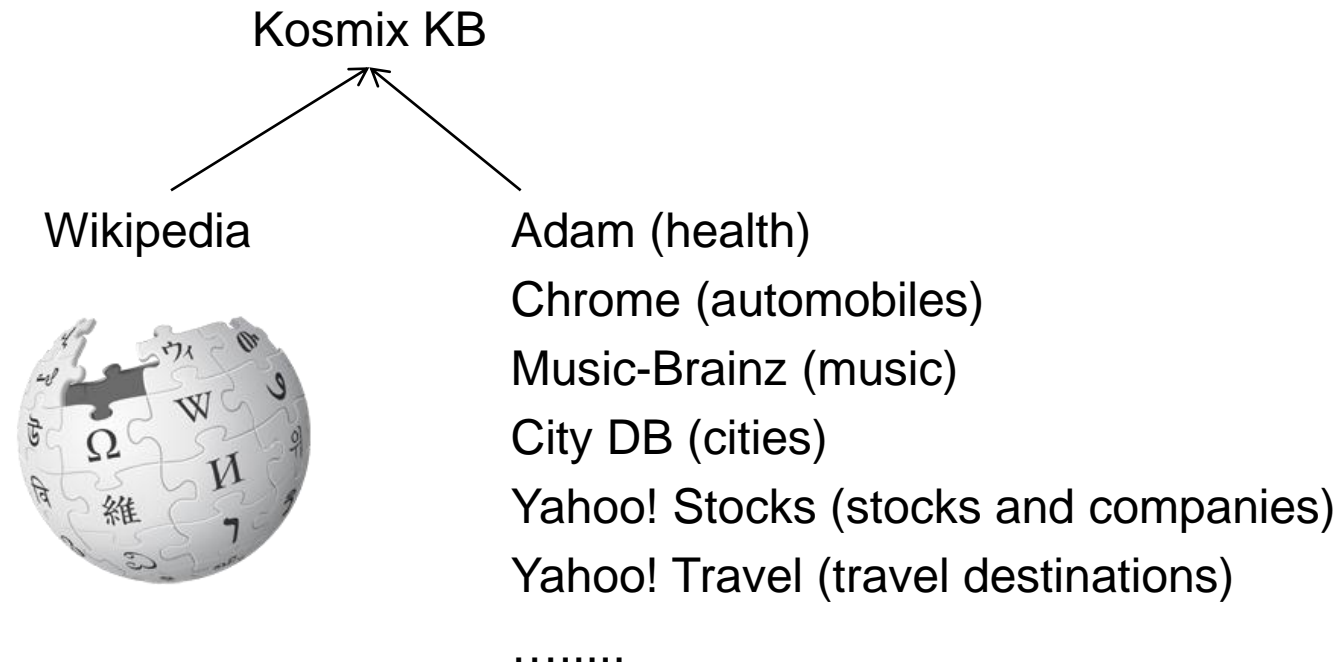
- rerun from scratch instead of incremental updating
- must reuse human curation

● Curating the KB

- ongoing process, regularly evaluate the KB
- add curations in form of commands → enable reusing of human curation
can curate multiple errors all at once

Building the Kosmix KB

- **Convert Wikipedia into a KB, then add more data sources**

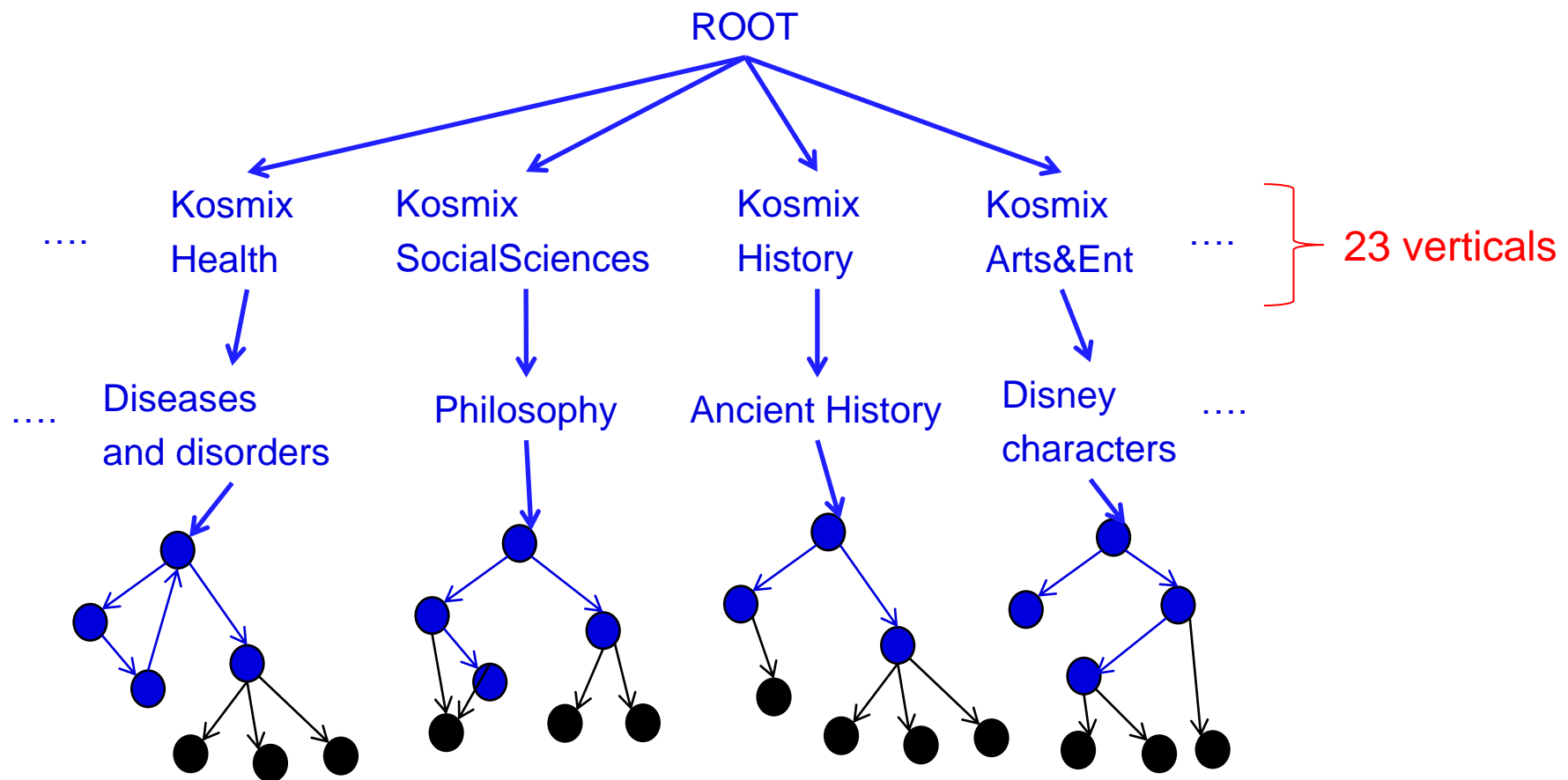


- **Why starting with Wikipedia?**

- “global” and “real-time”
- social media often mentions latest events/persons/... → need them to be in our KB asap
- Wikipedia is ideal for this
 - e.g., Nadal won US Open, Wikipedia updated within minutes

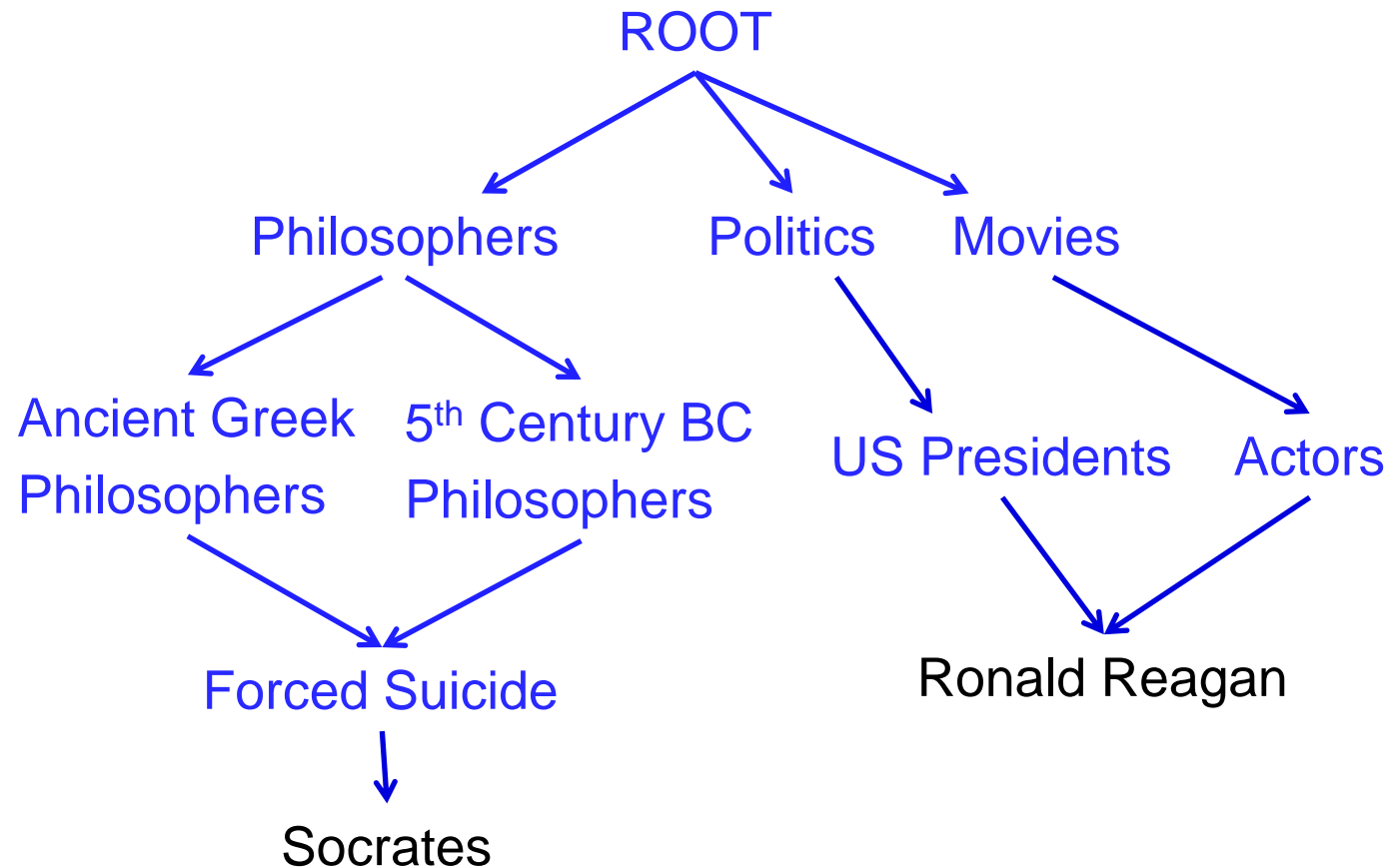
1. Convert Wikipedia into a Graph

- **Crawl Wikipedia, parse & construct a graph**
 - nodes = Wikipedia pages, edges = links among Wikipedia pages
- **Remove irrelevant parts of graph**
 - administration, help, discussion, ...
- **Glue remaining parts into a new graph with a ROOT node**



2. Extract Taxonomy of Concepts from Graph

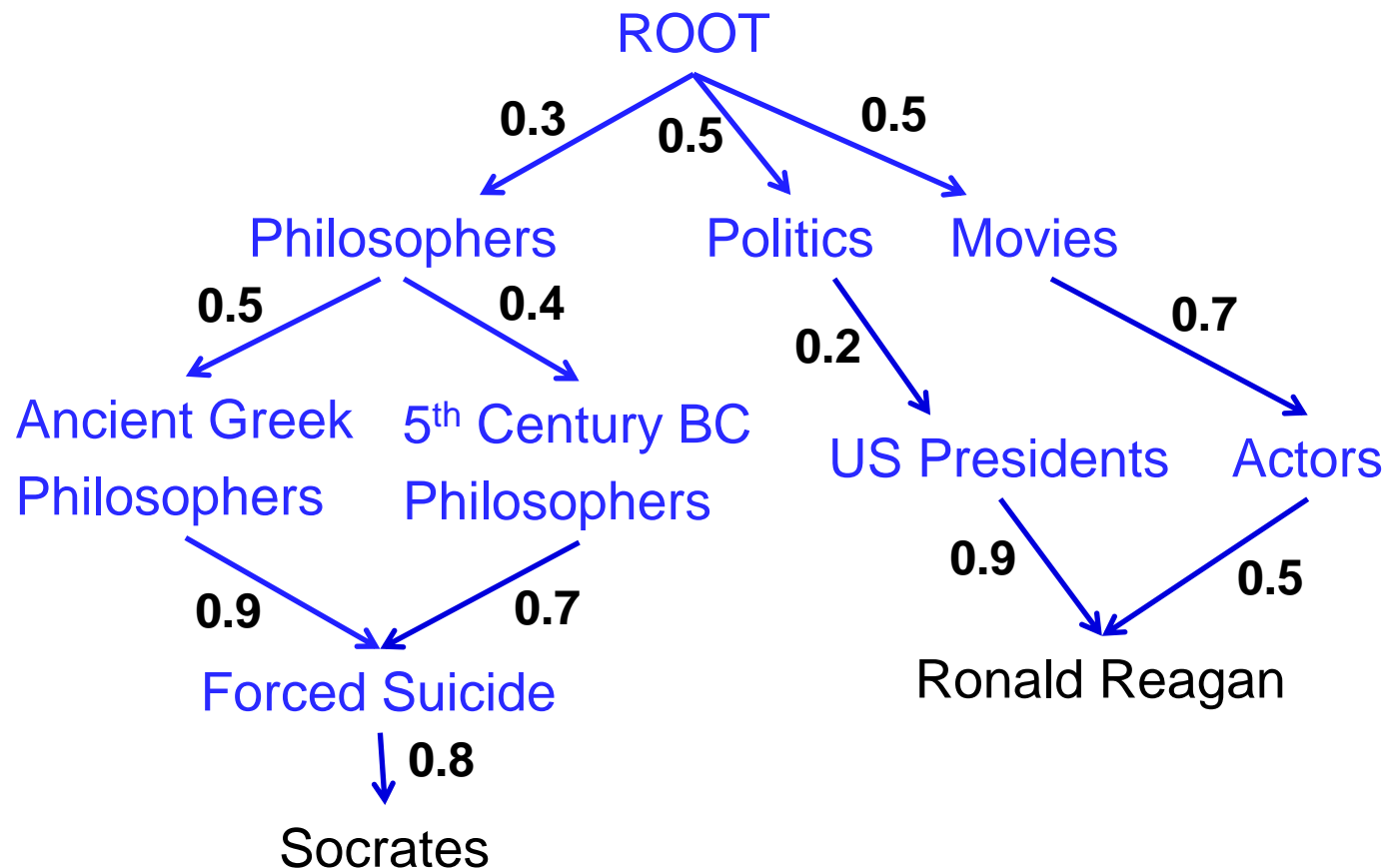
- To obtain taxonomic tree → for each node, find a single path to ROOT
- But nodes can have multiple paths to ROOT; which one to pick?



- **Picking wrong path causes many problems**
 - e.g., ROOT → Movies → Actors → Ronald Reagan
“Reagan left a mixed legacy”: will be classified incorrectly under “Movies”

2. Extract Taxonomy of Concepts from Graph

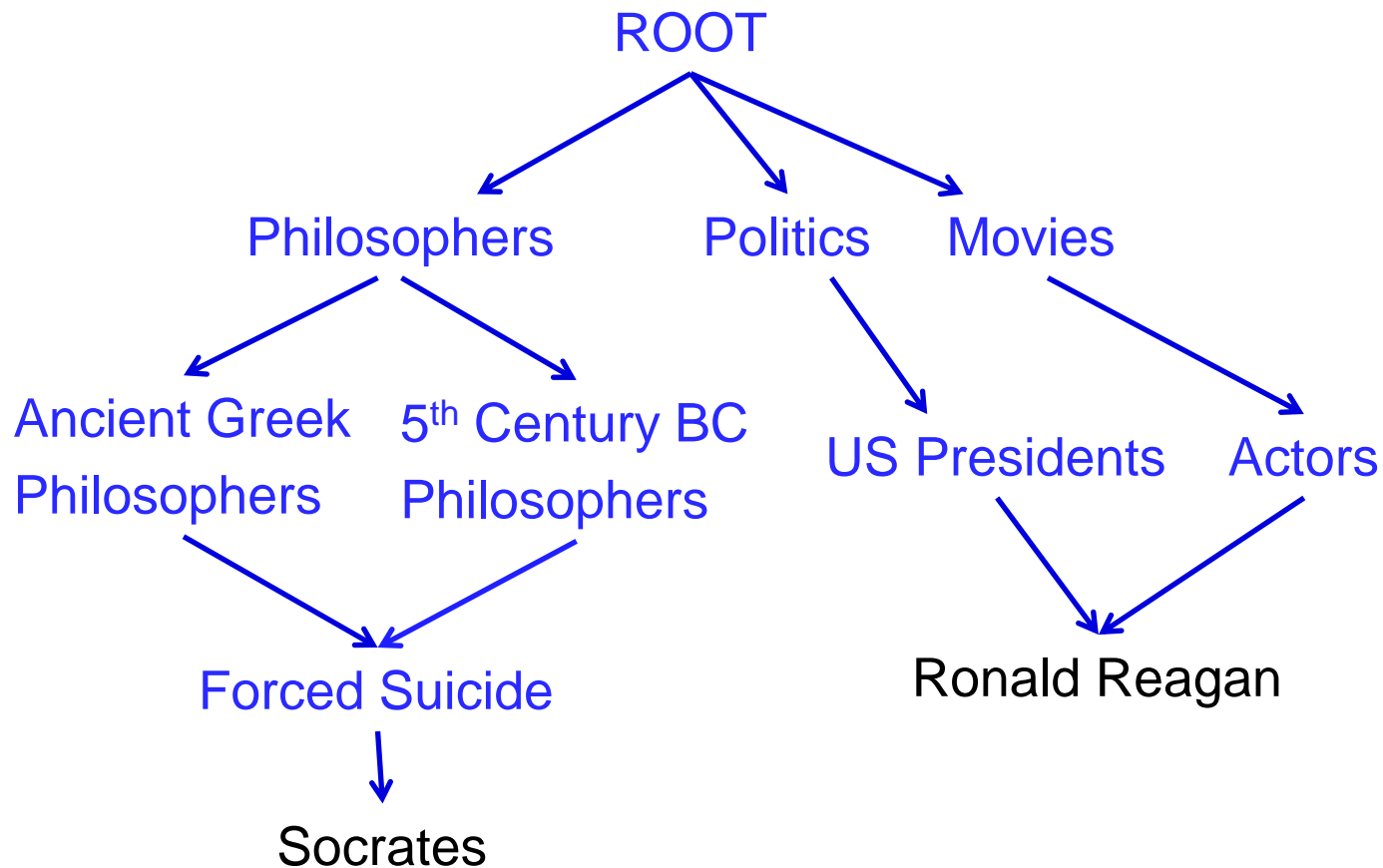
- Intuitively, pick **most popular/important/relevant path**
 - e.g., most people know Reagan as a president, not as an actor
- **Solution:**
 - assign to each edge $A \rightarrow B$ a weight to capture its popularity/importance/relevance
 - run a spanning tree discovery algorithm using these weights
 - output a maximum spanning tree



2. Extract Taxonomy of Concepts from Graph

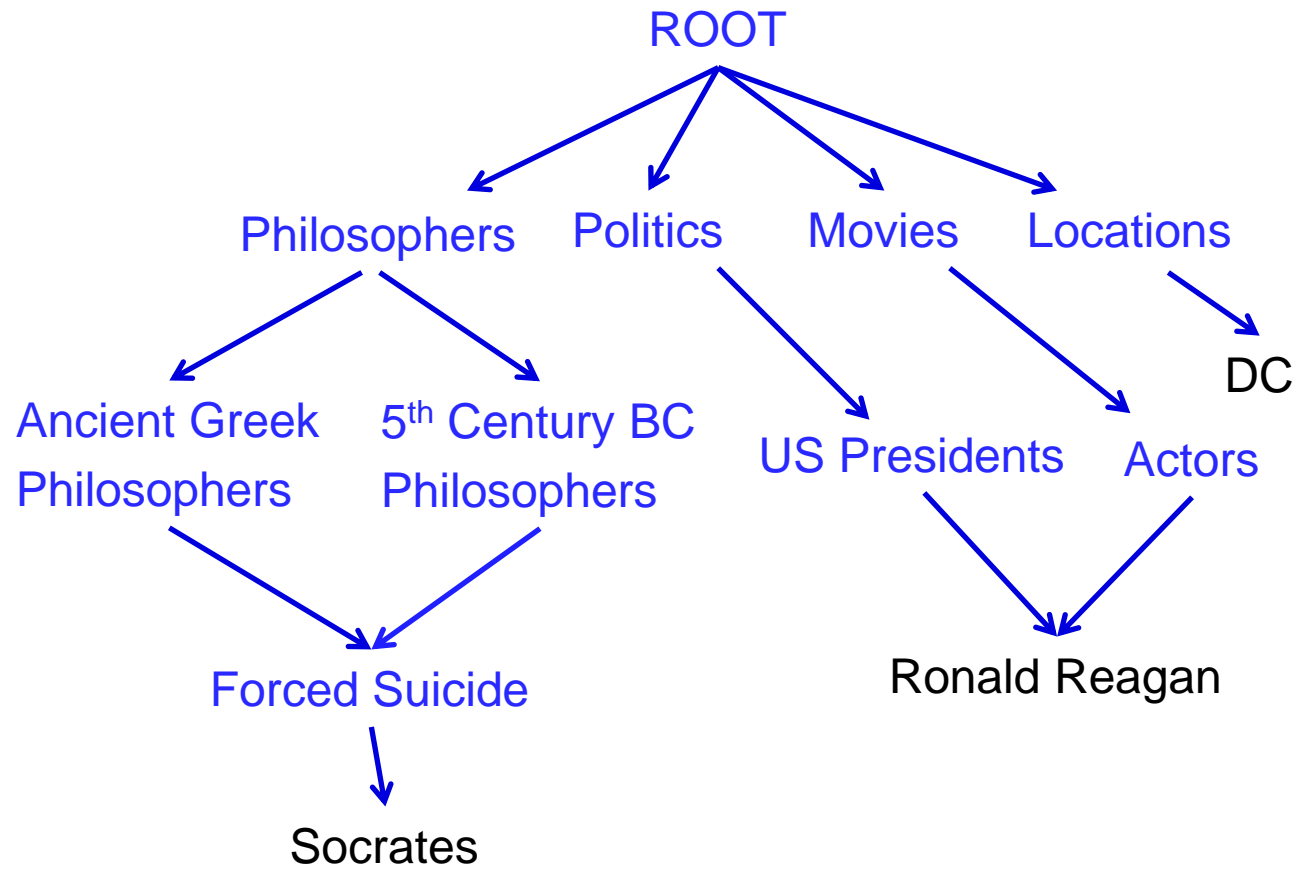
- **How to assign weights to edge $A \rightarrow B$?**
 - assign multiple weights, they form a weight vector
- **Examples**
 - **Web signal:** co-occurrence count of A and B on the Web
 - e.g., how many times “Ronald Reagan” and “President” co-occur in same Web page?
 - **Social signal:** same as Web signal, but measure co-occurrence in social media
 - **List signal:** how many times A and B co-occur in the same Wikipedia list?
 - ...
 - analyst can also assign weights to the edges

2. Extract Taxonomy of Concepts from Graph



- **We keep all paths for the nodes**
 - very useful for applications
- **To keep all paths, must detect and break cycles (see paper)**
- **End result: DAG of concepts + taxonomic tree imposed on the DAG**

3. Extract Relations for the KB



Typical solution:

- **Define a set of relations**

- livesIn, birthYear

- **Write extractors for them**

- using rules, machine learning

- **Apply extractors**

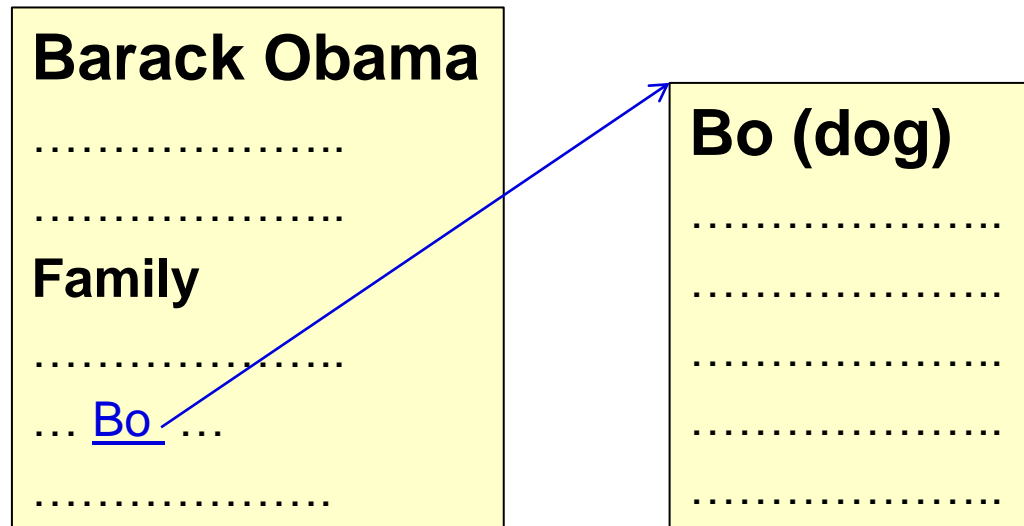
- livesIn(Reagan, DC),
birthYear(Reagan, 1911)

- **Problems:**

- Wikipedia has 10,000+ interesting relations →
can't manually define and extract all
- difficult to obtain high accuracy

3. Extract Relations for the KB

- **Our solution: extract fuzzy relations**



- **Extract <Barack Obama, Bo (dog), Family> as a relation**
 - a relation exists between “Barack Obama” and “Bo (dog)”, encoded by string “Family”
 - but we don’t know anything more precise
- **Yet this is already quite useful**
- **Example: querying “Obama family” on a search engine**
 - search query contains “family”, above relation also contains “family”
 - can return “Bo (dog)” as an answer
 - even though word “family” never appears in the page “Bo (dog)”

Socrates

From Wikipedia, the free encyclopedia
(Redirected from Sokrat)

This article is about the classical Greek philosopher. For other uses of Socrates, see [Socrates \(disambiguation\)](#).

Socrates (ⁱ/sɒˈkrɑːti/; **Z**; **Greek**: Σωκράτης, Ancient Greek pronunciation: [soːkrátɛːs], *Sōkrátēs*; c. 469 BC – 399 BC)^[1] was a **classical Greek Athenian philosopher**. Credited as one of the founders of **Western philosophy**, he is an enigmatic figure known chiefly through the accounts of later classical writers, especially the writings of his students **Plato** and **Xenophon**, and the plays of his contemporary **Aristophanes**. Many would claim that Plato's dialogues are the most comprehensive accounts of Socrates to survive from antiquity.^[2]

Through his portrayal in Plato's dialogues, Socrates has become renowned for his contribution to the field of **ethics**, and it is this Platonic Socrates who also lends his name to the concepts of Socratic irony and the **Socratic method**, or *elenchus*. The latter remains a commonly used tool in a wide range of discussions, and is a type of **pedagogy** in which a series of questions are asked not only to draw individual answers, but also to encourage fundamental insight into the issue at hand. It is Plato's Socrates that also made important and lasting contributions to the fields of **epistemology** and **logic**, and the influence of his ideas and approach remains strong in providing a foundation for much western philosophy that followed.

Contents [hide]

- 1 Biography
 - 1.1 The Socratic problem
 - 1.2 Life
 - 1.3 Trial and death
- 2 Philosophy
 - 2.1 Socratic method
 - 2.2 Philosophical beliefs
 - 2.2.1 Socratic paradoxes
 - 2.2.2 Knowledge
 - 2.2.3 Virtue
 - 2.2.4 Politics
 - 2.2.5 Covertness
- 3 Satirical playwrights
- 4 Prose sources
 - 4.1 The Socratic dialogues
- 5 Legacy
 - 5.1 Immediate influence
 - 5.2 Later historical effects
 - 5.3 Criticism

Relation from infobox:
<Socrates, Greek, Nationality>

Relation from template:
<Socrates, Plato, Disciples>


Relation from text:
<Socrates, Thucydides, The Socratic Problem>

Socrates (Σωκράτης)



Socrates

Born	c. 469 / 470 BC ^[1] <i>Deme</i> Alopece, Athens
Died	399 BC (age approx. 71) Athens
Nationality	Greek
Era	Ancient philosophy
Region	Western philosophy
School	Classical Greek
Main interests	Epistemology, ethics
Notable ideas	Socratic method, Socratic irony
Influenced	[show]



Part of a series on
Socrates

"I know that I know nothing"
Social gadfly · Trial of Socrates

Eponymous concepts
Socratic dialogue · Socratic method
Socratic questioning · Socratic paradox
Socratic problem

Disciples
Plato · Xenophon
Antisthenes · Antistippus

Biography

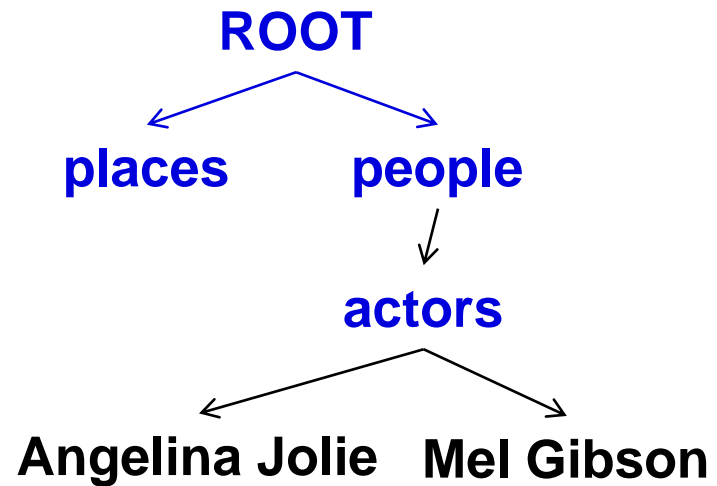
The Socratic problem

An accurate picture of the historical Socrates and his philosophical viewpoints is problematic, known as the **Socratic problem**.

As Socrates did not write philosophical texts, the knowledge of the man, his life, and his philosophy is only based on writings by his students and contemporaries. Foremost among them is **Plato**; however, works by **Xenophon**, **Aristotle**, and **Aristophanes** also provide important insights.^[3] The difficulty of finding the "real" Socrates arises because these works are often philosophical or dramatic texts rather than straightforward histories. **Aside from Thucydides** (who makes no mention of Socrates or philosophers in general) and Xenophon, there are in fact no straightforward histories contemporary with Socrates that dealt with his own time and place. A corollary of this is that sources that do mention Socrates do not necessarily claim to be historically accurate, and are often partisan (those who prosecuted and convicted Socrates have left no testament). Historians therefore face the challenge of reconciling the various texts that come from these men to create an accurate and consistent account of Socrates' life and work. The result of such an effort is not necessarily realistic, merely consistent.

Plato is frequently viewed as the most informative source about Socrates' life and philosophy.^[4] At the same time, however, many scholars believe that in some works Plato, being a literary artist, pushed his avowedly brightened-up version of "Socrates" far beyond anything the historical Socrates was likely to have done or said; and that Xenophon,

4. Extract Metadata for KB Instances



Web URLs

- en.wikipedia.org/wiki/Mel_Gibson
- movies.yahoo.com/person/mel-gibson/
- imdb.com/name/nm0000154/

Twitter ID

- @melgibson

Wikipedia page visits (last day, last week,..)

- 7, 33, ...

Web signature

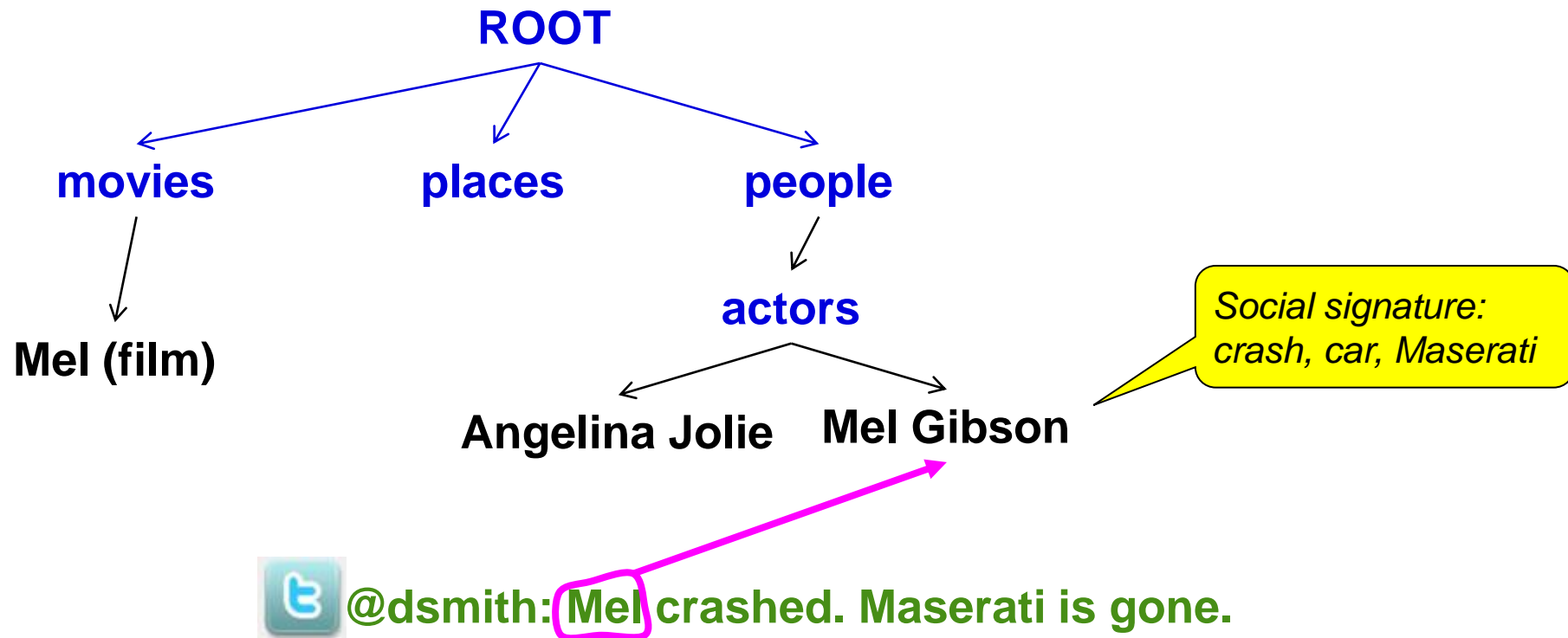
- “actor”, “Hollywood”, “Oscar”, ...

Social signature (last 3 hours)

- “car”, “crash”, “Maserati”, ...

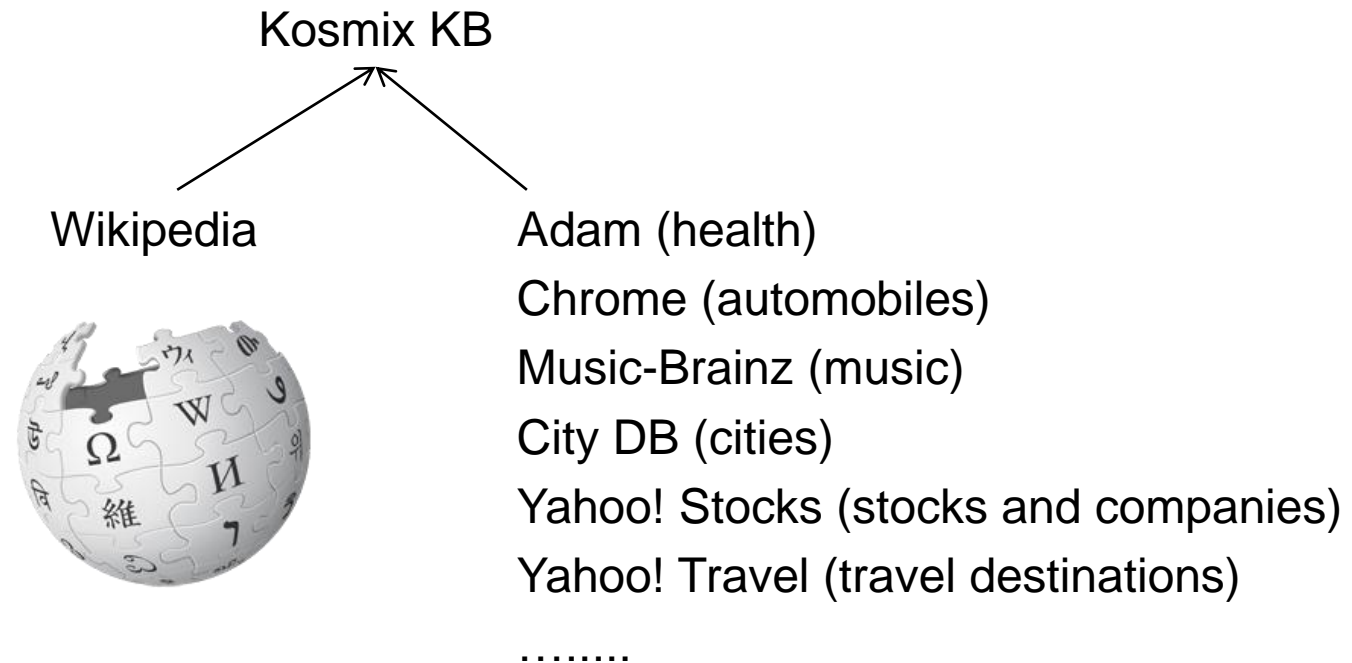
...

Example: Using Metadata in Social Media Analysis



For more detail, see “Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-Based Approach”, VLDB-13

5. Add More Data Sources to the KB



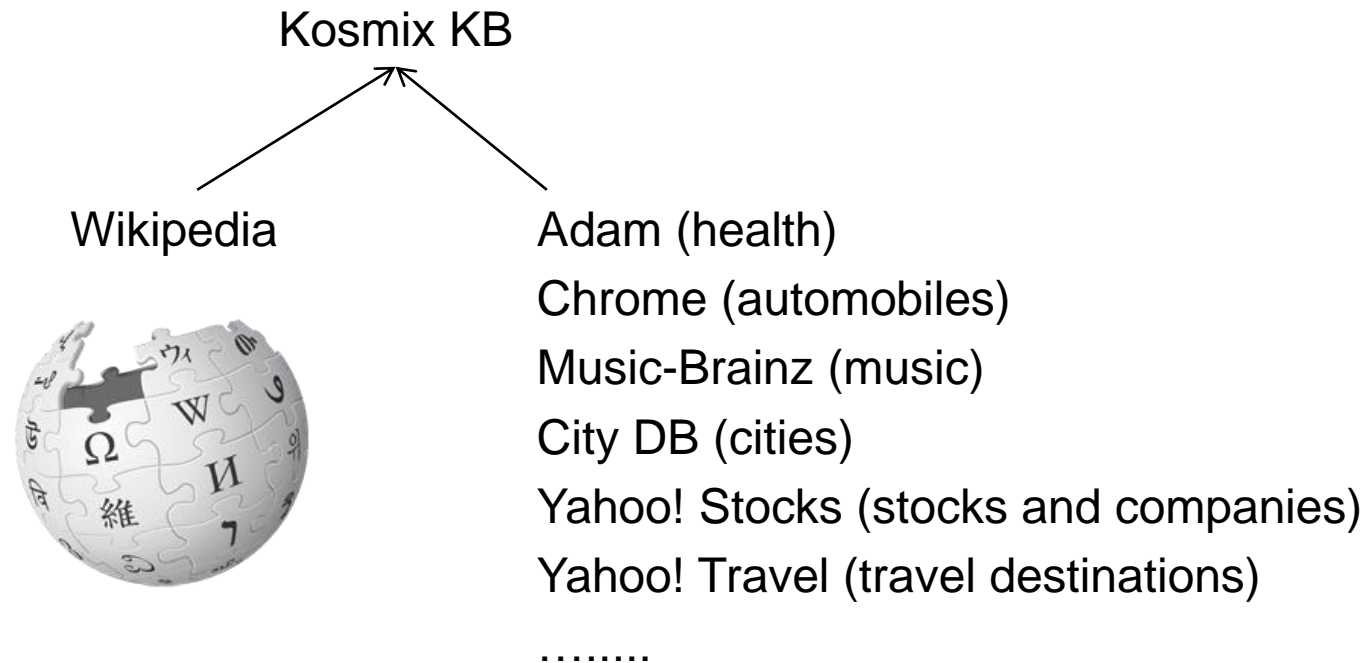
- **Challenges**

1. Match source taxonomy (if any) to KB taxonomy
2. Match source instances to KB instances

- **Key innovations (see paper)**

1. Interleave taxonomy matching and instance matching
2. Heavily use node metadata to match instances

Updating the KB



- **Typical solution : Incremental updates**

- fast, relatively easy to preserve human curations

- **But difficult in our case**

- we use “global” algorithms (e.g., spanning tree discovery) during KB construction

- **Our solution**

- run the pipeline from the scratch daily

- challenge: how to preserve human curation?

Human Curation

- **Automatically constructed KB often contains errors**

- automatic version of Kosmix KB is about 70% accurate

➔ **need human curation**

- **A human analyst**

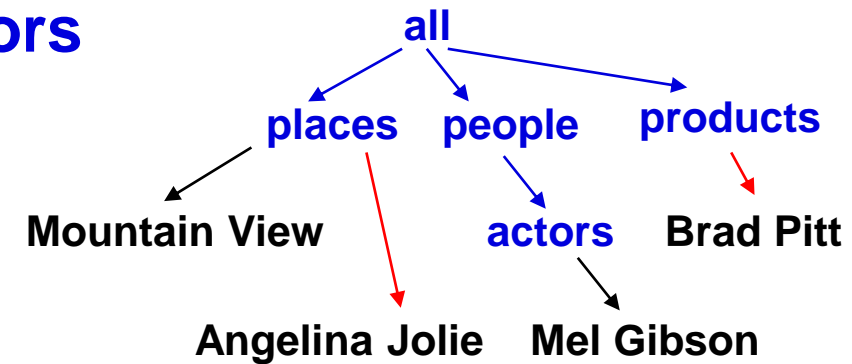
- evaluates the quality of our KB
- writes curations

- **Curate by writing commands**

- e.g. **Angelina Jolie | actors | 0.9**, or even better: **infobox:actors | actors | 0.9**

- **Current KB contains several thousand commands (written over 3-4 years)**

- **Raises the accuracy of the KB to well above 90%**



Team Organization

- **A core team of 4 people (in 2010-2011)**

- **1 data analyst**
 - performed quality evaluation and curation
- **1 developer**
 - wrote code, developed new features, added new signals on edges, etc.
- **0.5 systems expert**
 - crawled data sources, maintained in-house Wikipedia mirror and Web corpus
- **0.5 UI specialist**
 - worked on the look-and-feel of the tools
- **1 team lead**
 - designed, supervised and coordinated the work



Concluding Remarks

- **Possible to build relatively large KBs with modest hardware and team size**
- **Human curation is important**
 - raises the accuracy of our KB from 70% to well above 90%
 - possible to make a lot of curation with just 1-2 persons, using commands
- **An imperfect KB is still very useful for a variety of real world applications**
 - search, advertising, social media analysis, product search, user query understanding, social gifting, social mining, ...
 - often, these apps use KB internally and do not need to show KB data to end users
- **Imperfect relationships still quite useful**
 - provide contexts for KB nodes, show how they relate to one another
- **Capturing contexts is critical for processing social media**
 - especially social contexts
- **Important to have clear & proven methodologies to build & maintain KBs**
 - as multiple teams try to build their own KBs