CS290D – Advanced Data Mining

Instructor: Xifeng Yan
Computer Science
University of California at Santa Barbara

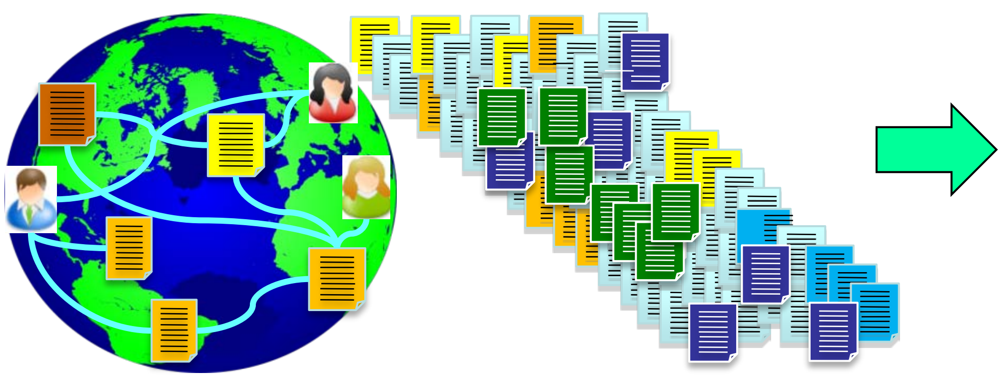
Knowledge Bases

Instructor: Yang Li
Computer Science
University of California at Santa Barbara

Slides adjusted from: http://resources.mpi-inf.mpg.de/yago-naga/vldb2014-tutorial/

What is a Knowledge Base?

- A comprehensive organization of human knowledge
  - Including everything that human knows
  - Absorbing everything machine-readable
  - Capturing rich info of entities, classes and relationships

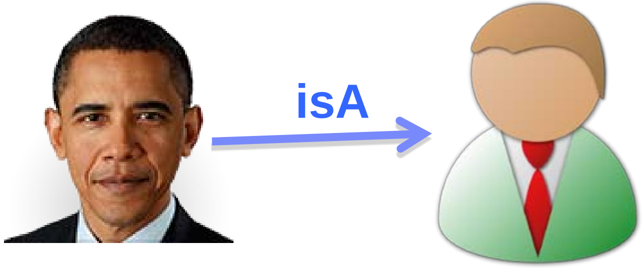


Knowledge Base Components

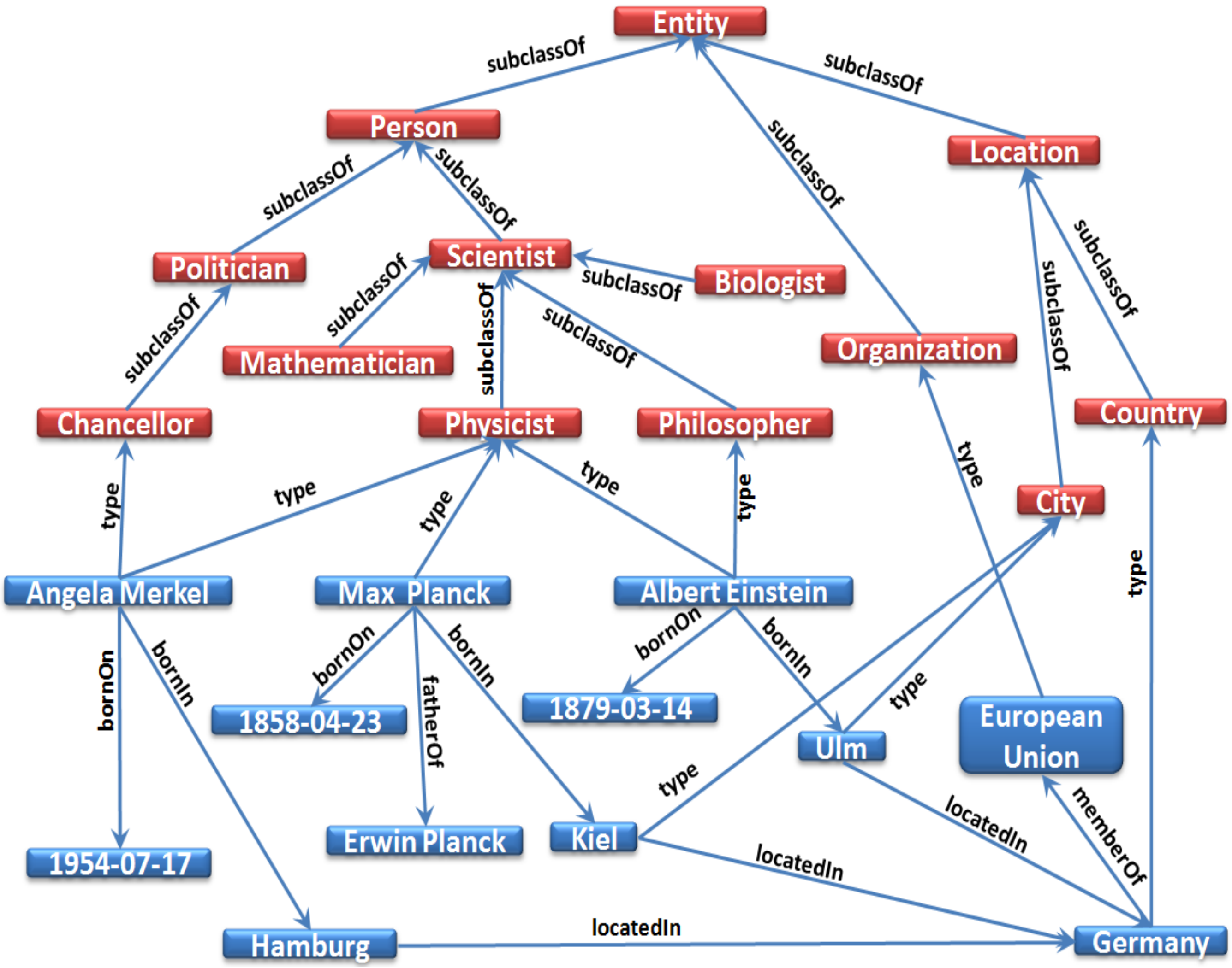
- Entity (individual)
  - Concrete entities like a person or an animal
  - Abstract entities like concepts

- Class (concept)
  - Abstract collection of individuals

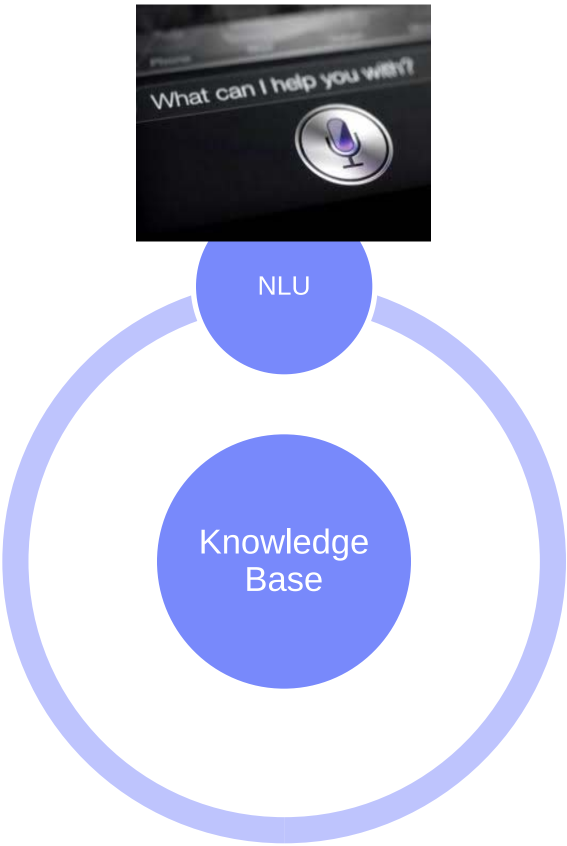
- Relation
  - How classes and individuals can be related
  - isA, subclassOf, etc.



Knowledge Base Example



Why we need Knowledge Bases?

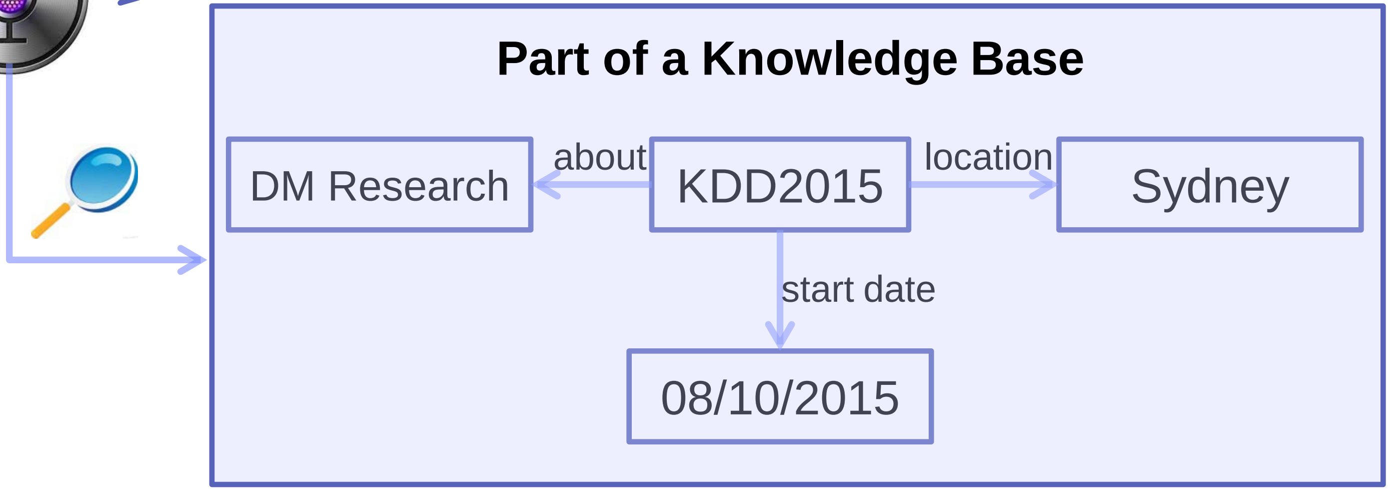


Application 1: Natural Language Understanding

What can I help you with?

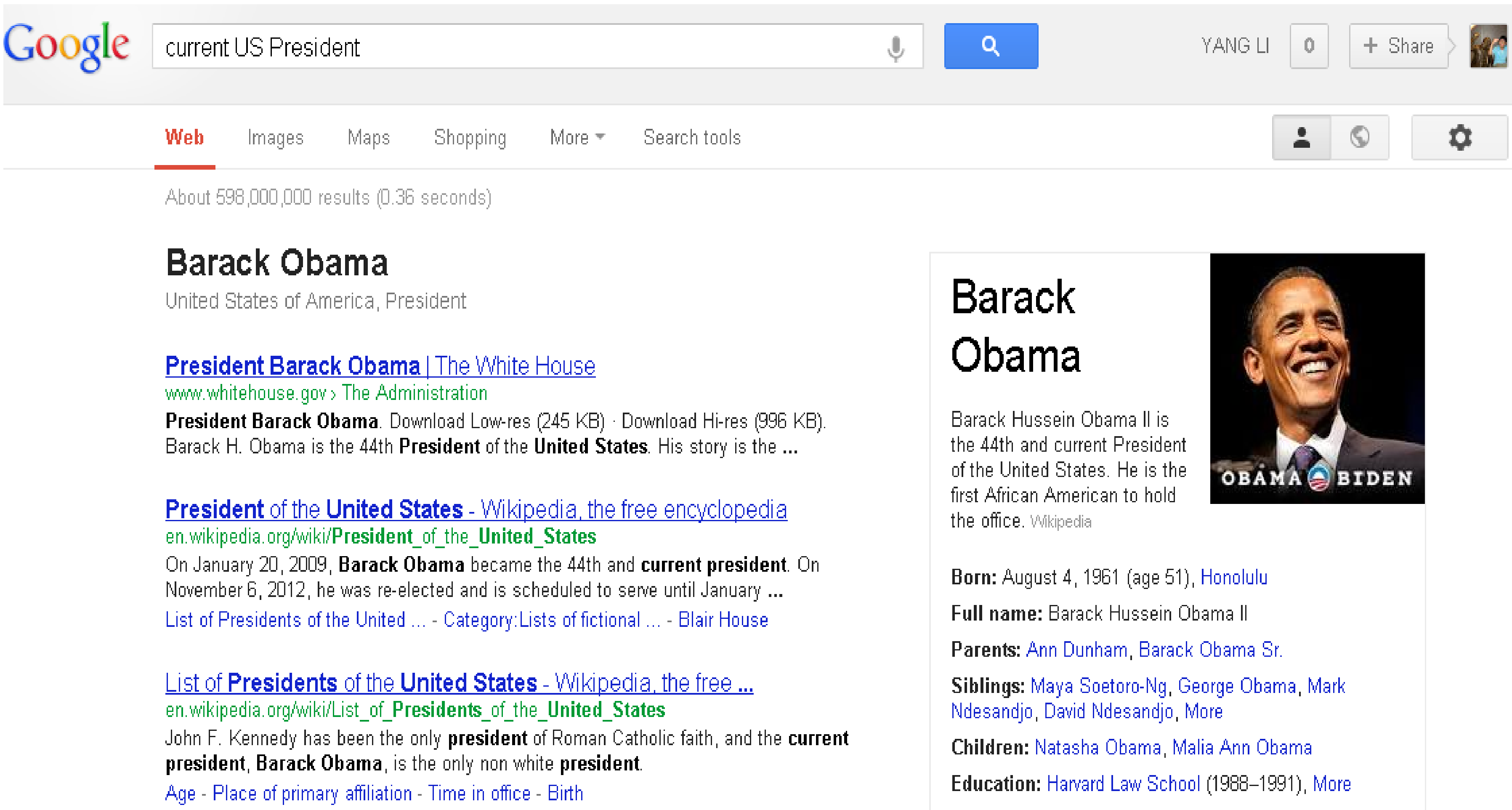
Book me a flight to KDD2015

I am booking you a flight to Istanbul. The conference starts at 08/10/2015, do you want to leave one day earlier?



Why we need Knowledge Bases?

Application 2: Semantic Search



Why we need Knowledge Bases?

NLU

Knowledge Network

Semantic Search

Reasoning

Application 3: Reasoning

Why we need Knowledge Bases?

NLU

deepQA

Knowledge Network

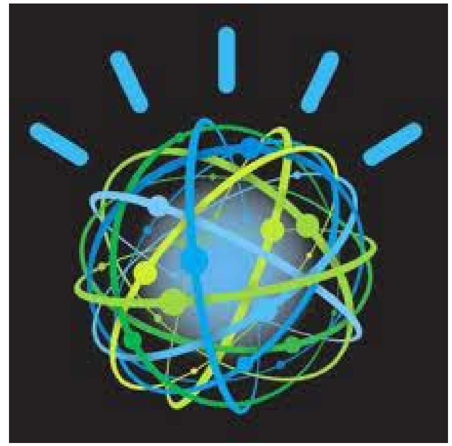
Semantic Search

Reasoning

Application 4: DeepQA

Knowledge Bases

Question Classification & Decomposition



Answers!

History of Knowledge Bases

from humans
for humans

from algorithms
for machines

WordNet

WordNet
A lexical database for English

guitarist ⊂ {player,musician}
⊂ artist

algebraist
⊂mathematician
⊂scientist

Cyc

Wikipedia

4.5 Mio. English articles
20 Mio. contributors

DBpedia

WolframAlpha computational knowledge engine

yago
select knowledge

freebase

1985 1990 2000 2005 2010

Closed-domain KBs vs. Open-domain KBs

- Closed-domain KBs
  - Focus on the important knowledge about a specific domain
  - E.g. DBLP, PubMed, etc.

- Open-domain KBs
  - Try to cover everything that human knows
  - E.g. DBpedia, Freebase, Google Knowledge Graph, YAGO, etc.
  - There are still a lot of uncovered tail knowledge!

Human-curated KBs vs. Auto-constructed KBs

- Human-curated KBs
  - Created by experts or crowdsourcing
  - E.g. Cyc, DBpedia, Freebase *etc*.
  - High accuracy, low coverage, good for small closed domains
- Automatically constructed KBs
  - Extracted by algorithms
  - E.g. YAGO, Reverb, NELL, Probase *etc*.
  - High coverage, low accuracy, good for open domains

DBpedia (S. Auer *et al.* ISWC'07)

About: Mark Zuckerberg

An Entity of Type : agent, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

DBpedia

Mark Elliot Zuckerberg (born May 14, 1984) is an American computer programmer and Internet entrepreneur. He is best known as one of four co-founders of the social networking site Facebook, of which he is chairman and chief executive. Zuckerberg was born and raised in a Jewish household in New York state. While still in middle school in his early teens, he took up writing software programs as a hobby, beginning with BASIC, with help from his father.

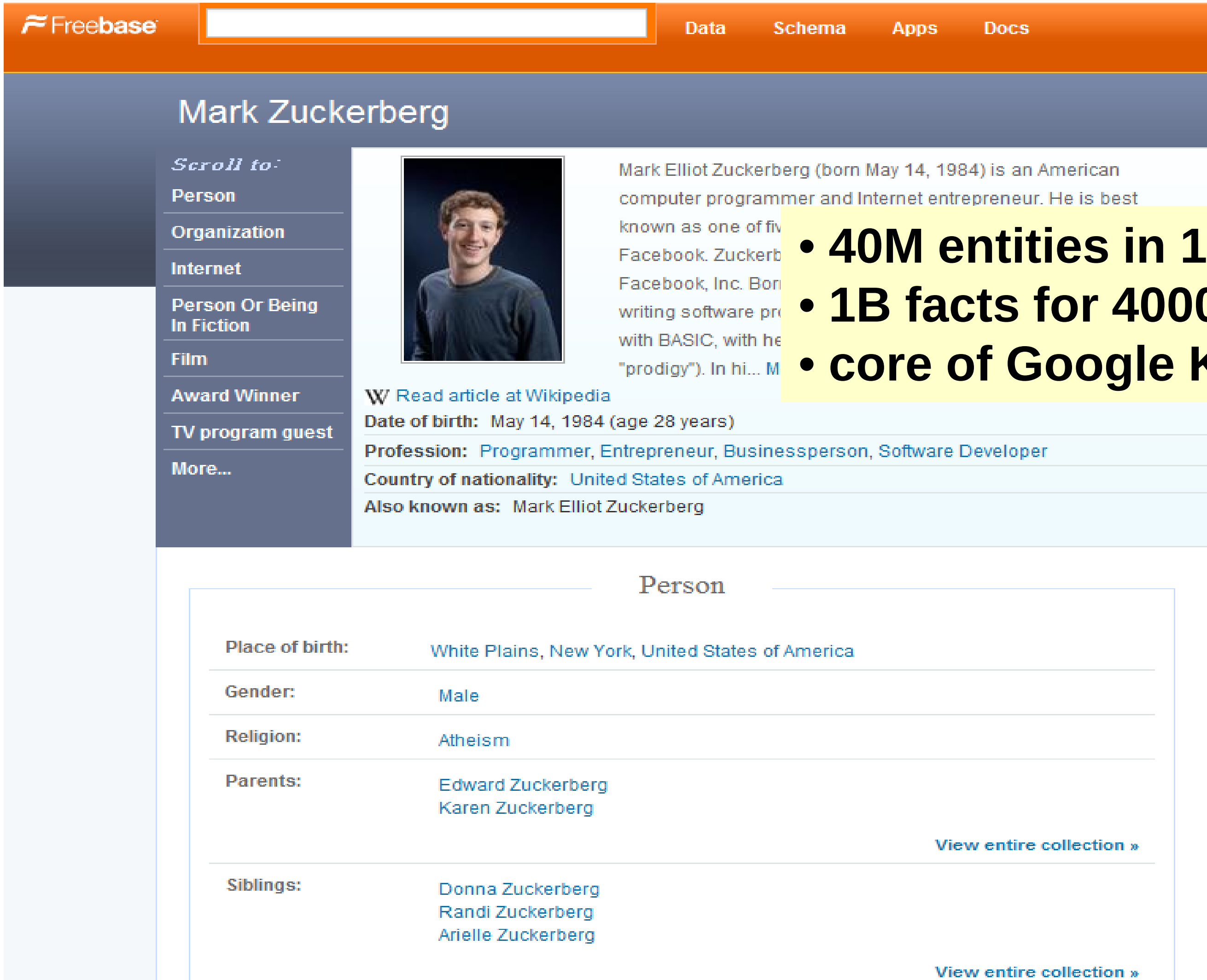
Property	Value
dbpedia-owl:award	dbpedia:Time_Person_of_the_Year
dbpedia-owl:birthDate	1984-05-14 (xsd:date)
dbpedia-owl:birthName	Mark Elliot Zuckerberg
dbpedia-owl:birthPlace	dbpedia:White_Plains,_New_York
dbpedia-owl:birthYear	1984-01-01 00:00:00 (xsd:date)
dbpedia-owl:individualisedPnd	139618171
dbpedia-owl:networth	1.75E10
dbpedia-owl:occupation	dbpedia:Mark_Zuckerberg__1
dbpedia-owl:relative	dbpedia:Randi_Zuckerberg
dbpedia-owl:residence	dbpedia:Palo_Alto,_California
dbpprop:almaMater	Harvard University
dbpprop:awards	TIME Person of the Year 2010
dbpprop:birthDate	1984-05-14 (xsd:date)
dbpprop:birthName	Mark Elliot Zuckerberg
dbpprop:birthPlace	White Plains, New York, U.S.
dbpprop:caption	8.0
dbpprop:dateOfBirth	1984 (xsd:integer)
dbpprop:hasPhotoCollection	http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Mark_Zuckerberg
dbpprop:knownFor	Co-founding Facebook in 2004; world's youngest billionaire as of 2008
dbpprop:name	Mark Zuckerberg Zuckerberg, Mark Elliot
dbpprop:networth	US$ 17.5 billion
dbpprop:occupation	Chairman and CEO of Facebook
dbpprop:placeOfBirth	White Plains, New York, United States
dbpprop:relatives	Randi, Donna and Arielle
dbpprop:residence	Palo Alto, California, U.S.
dbpprop:shortDescription	American computer entrepreneur

- **4M entities in 250 classes**
- **500M facts for 6000 properties**
- **live updates by community**

- **High Quality**

- **Low Coverage**
- **High Cost**

Freebase (K. Bollacker *et al*. SIGMOD'08)

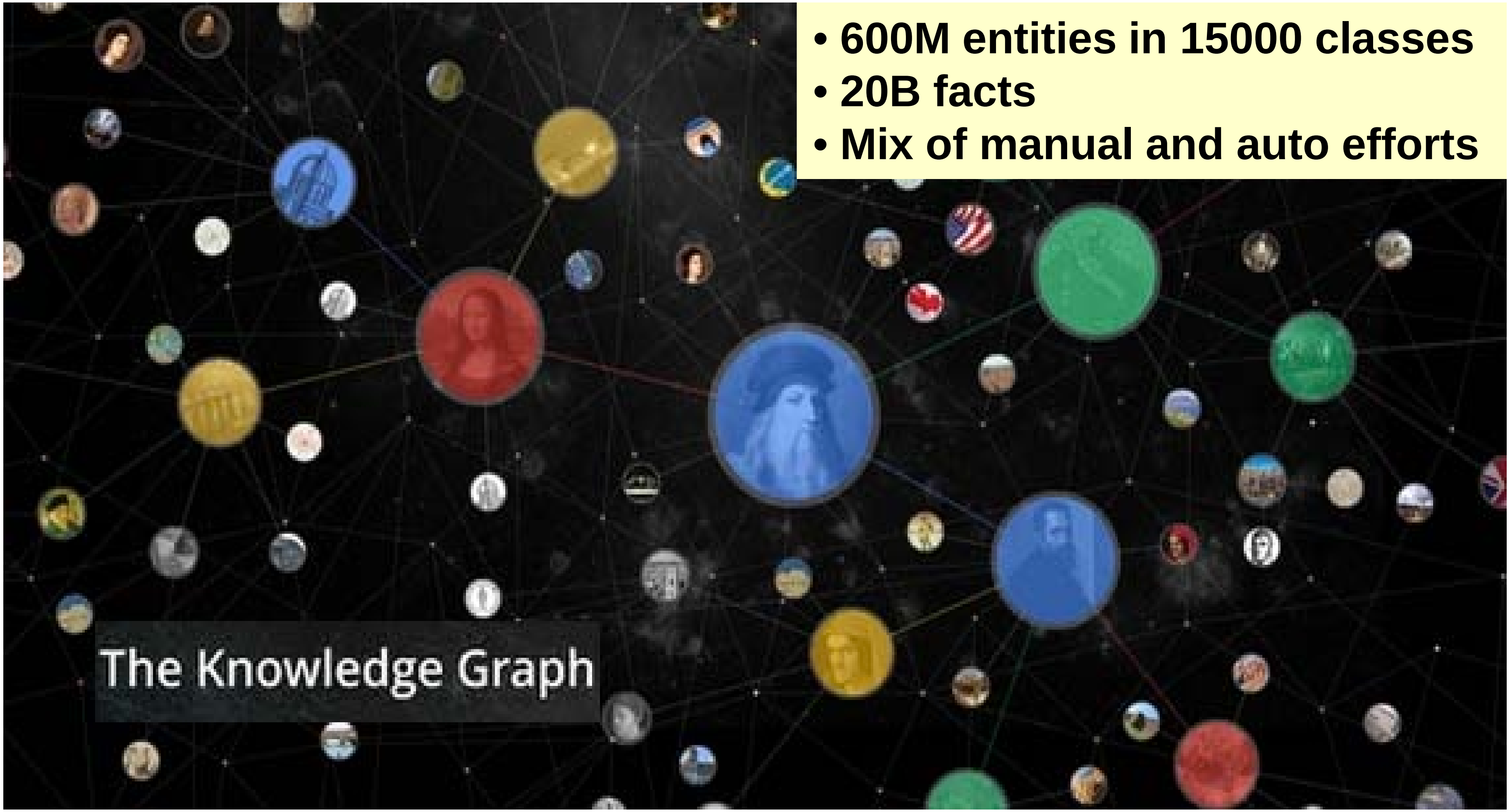


- **40M entities in 15000 topics/classes**
- **1B facts for 4000 properties**
- **core of Google Knowledge Graph**

- **High Quality**

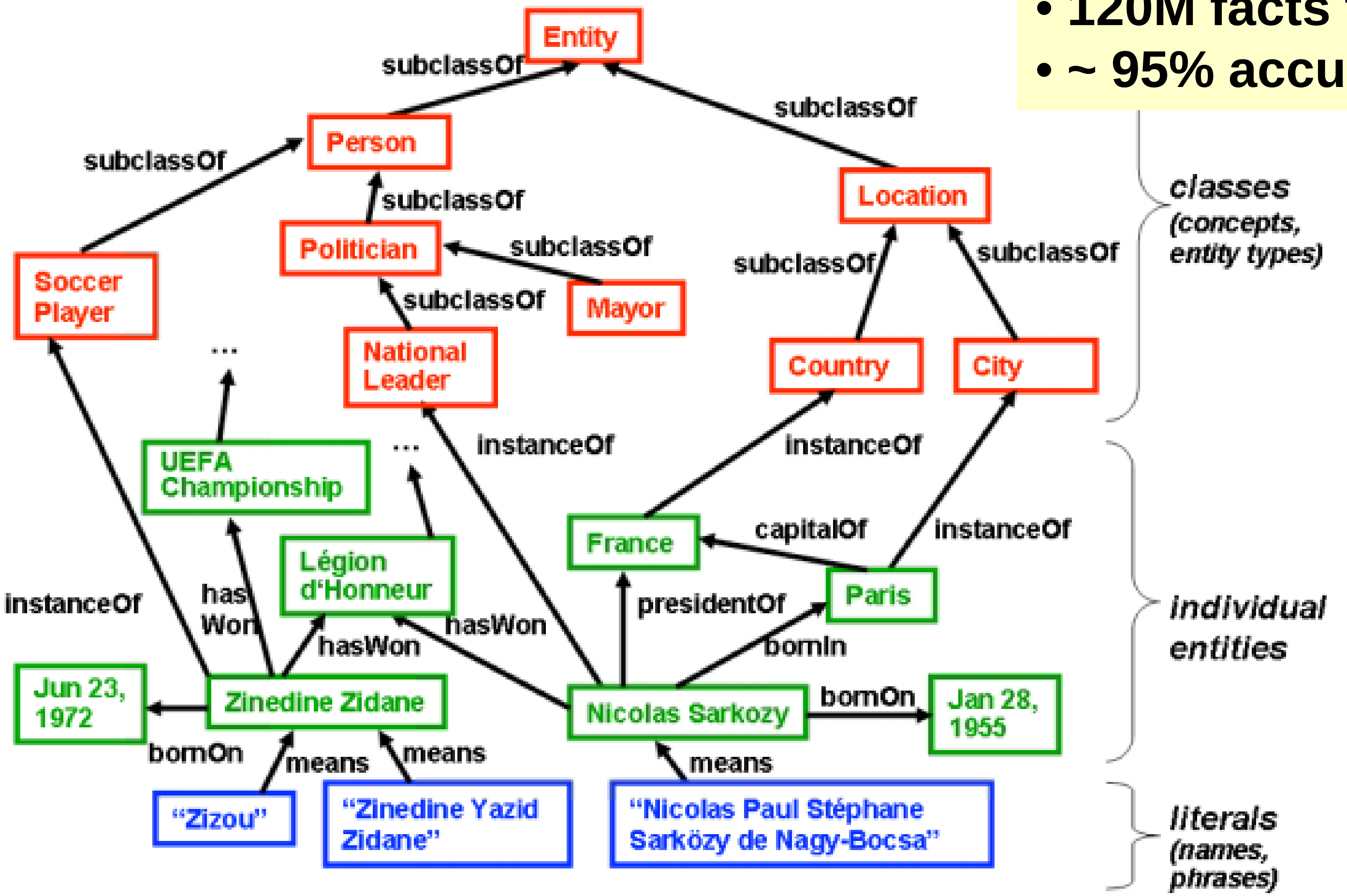
- **Low Coverage**
- **High Cost**

Google Knowledge Graph



- **600M entities in 15000 classes**
- **20B facts**
- **Mix of manual and auto efforts**

YAGO (F. Suchanek *et al.* WWW'07)



- 10M entities in 350K classes
- 120M facts for 100 relations
- ~ 95% accuracy

pros

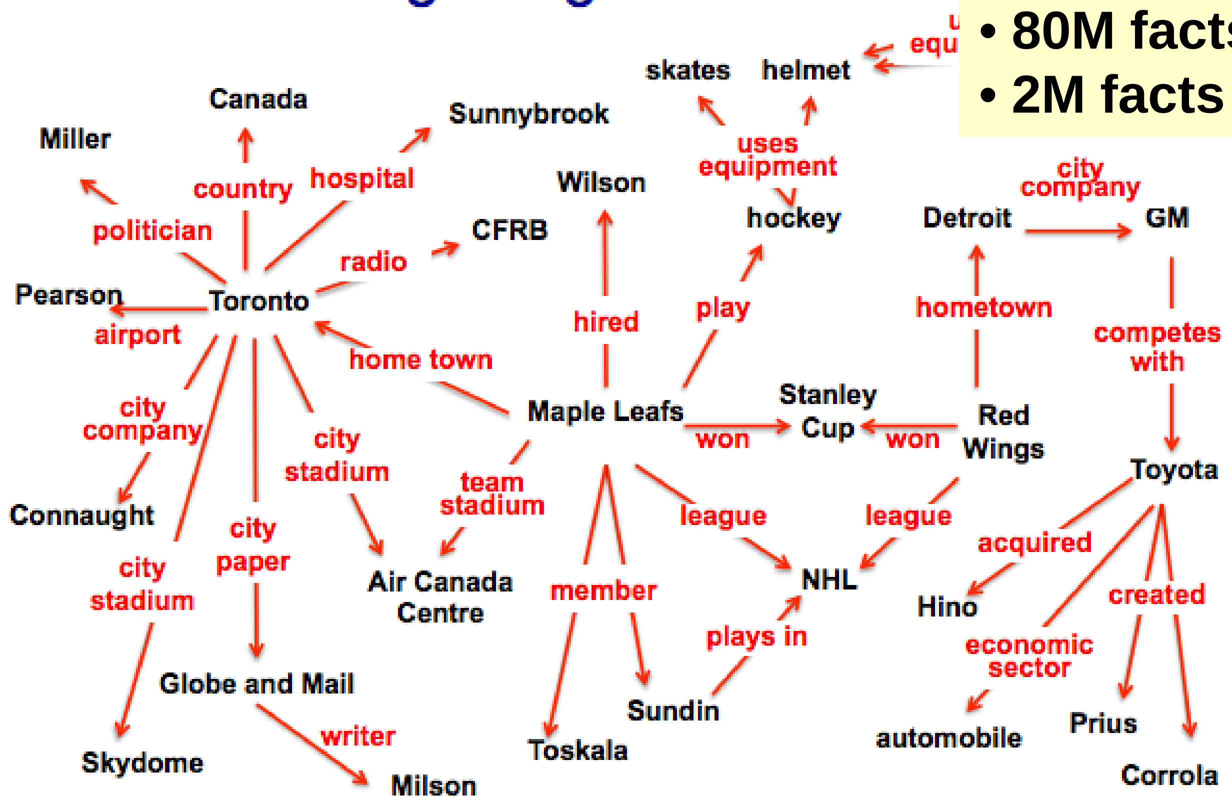
- High Quality
- Low Cost

cons

- Low Coverage

NELL (A. Carlson *et al.* AAAI'10)

NELL knowledge fragment

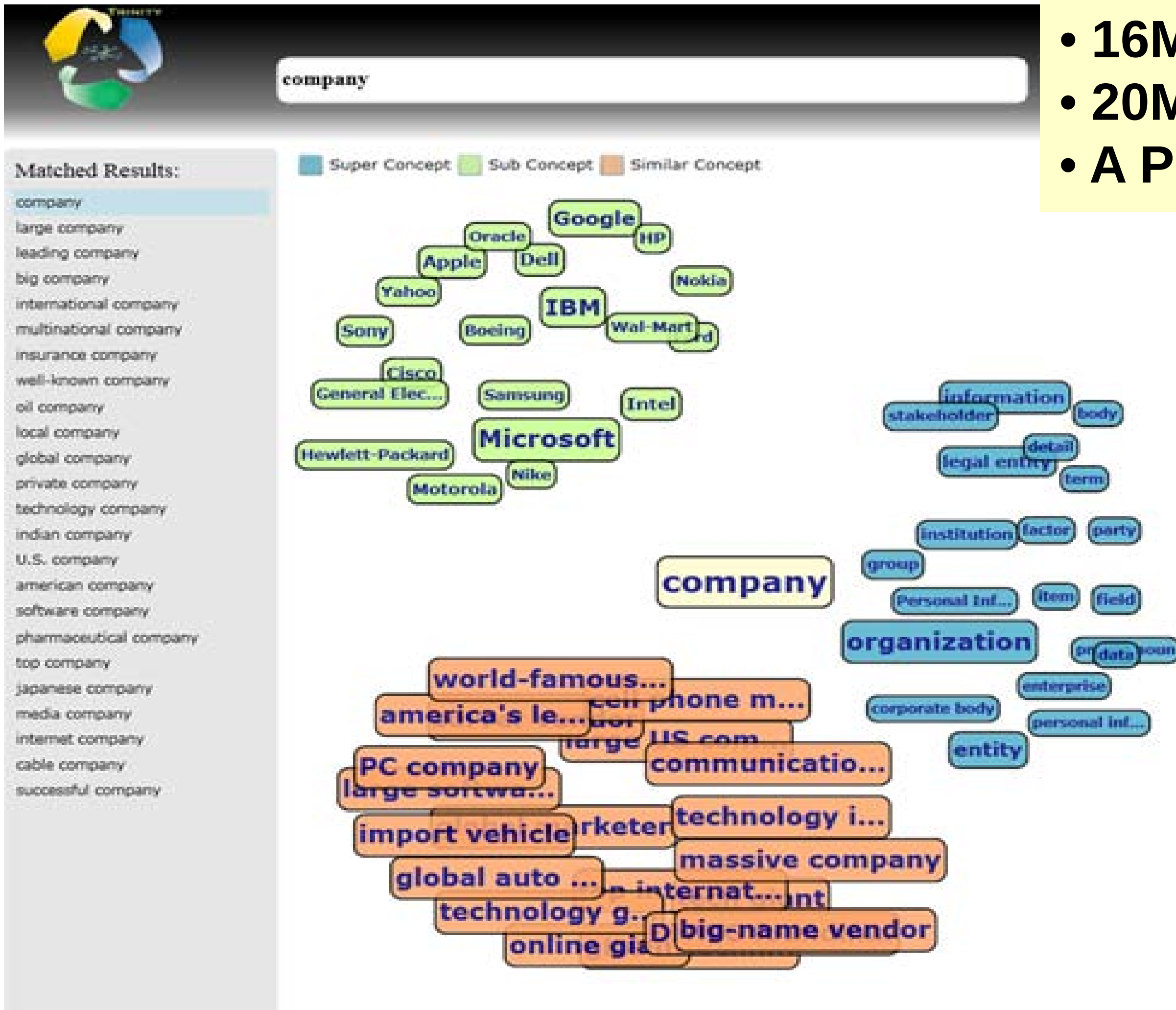


- **covers entities in 280 classes**
- **80M facts for 327 relations**
- **2M facts are highly confident**

- **High Quality**
- **Low Cost**

- **Low Coverage**

Probase (W. Wu *et al.* SIGMOD'12)



- **16M entities in 2.7M classes**
- **20M *isA* relationship pairs**
- **A Probabilistic Taxonomy**

pros

- **Large Concept Space**
- **Handling Noisy Data**
- **Modeling Uncertainty**
- **Low Cost**

cons

- **Only *isA* Relationship**
- **No Entity-level Info**

Automatic Knowledge Harvesting

- Taxonomic Knowledge:
  - Entities and Classes
- Factual Knowledge:
  - Relations between Entities
- Emerging Knowledge:
  - New Entities & Relations
- Temporal Knowledge:
  - Validity Times of Facts
- Common-sense Knowledge

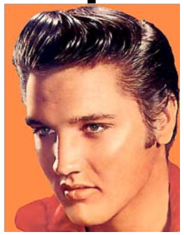
Taxonomic Knowledge

person

subclassOf

singer

type



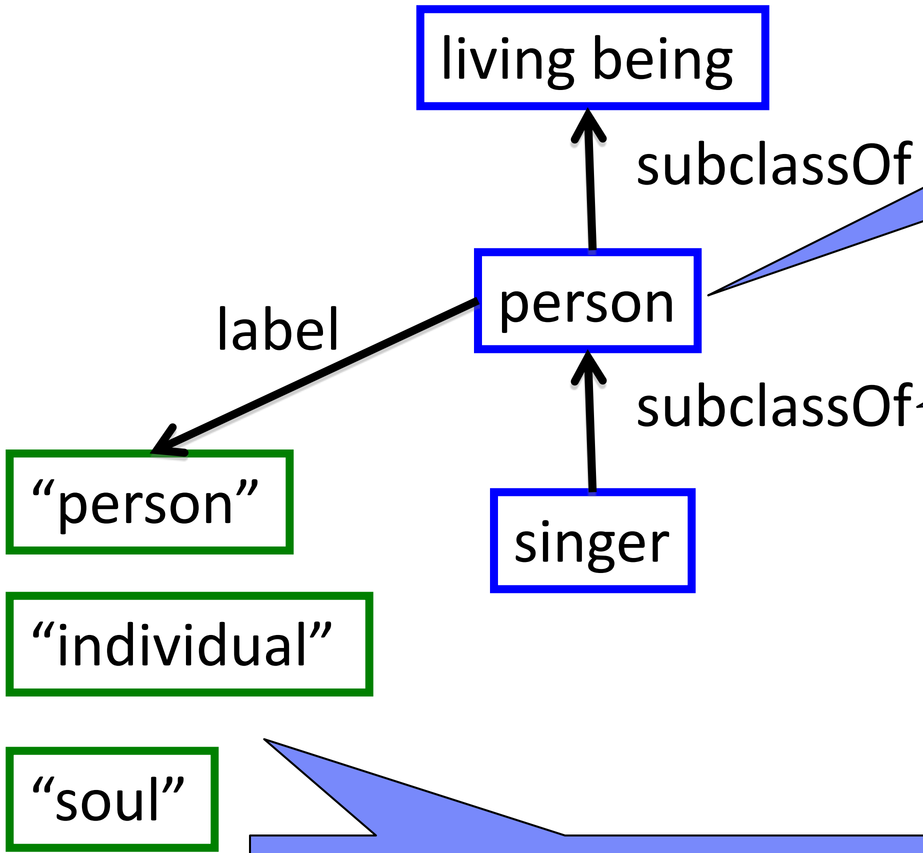
Which classes exist?

Which subsumptions hold?

Which entities belong to which classes?

Which entities exist?

WordNet contains rich taxonomic knowledge



WordNet contains 82,000 classes

WordNet contains thousands of subclassOf relationships

WordNet contains 118,000 class labels

only 32 singers !?
4 guitarists
5 scientists
0 enterprises
WordNet classes lack instances

Wikipedia is rich in instances

Steve Jobs

From Wikipedia, the free encyclopedia

For the biography, see Steve Jobs (biography).

Steven Paul Jobs (/ˈdʒɒbz/; February 24, 1955 – October 5, 2011)[4][5] was an American businessman and inventor widely recognized as a charismatic pioneer of the personal computer revolution.[6][7] He was co-founder, chairman, and chief executive officer of Apple Inc. Jobs also co-founded and served as chief executive of Pixar Animation Studios; he became a member of the board of directors of The Walt Disney Company in 2006, following the acquisition of Pixar by Disney.

In the late 1970s, Apple co-founder Steve Wozniak engineered one of the first commercially successful lines of personal computers, the Apple II series. Jobs directed its aesthetic design and marketing along with A.C. "Mike" Markkula, Jr. and others. In the early 1980s, Jobs was among the first to see the commercial potential of Xerox PARC's mouse-driven graphical user interface, which led to the creation of the Apple Lisa (engineered by Ken Rothmuller and John Couch) and, one year later, creation of Apple employee Jef Raskin's Macintosh.

After losing a power struggle with the board of directors in 1985, Jobs left Apple and founded NeXT, a computer platform development company specializing in the higher-education and business markets. NeXT was eventually acquired by Apple in 1996, which brought Jobs back to the company he co-founded, and provided Apple with the NeXTSTEP codebase, from which the Mac OS X was developed."[8] Jobs was named Apple advisor in 1996, interim CEO in 1997, and CEO from 2000 until his resignation. He oversaw the development of the iMac, iTunes, iPod, iPhone, and iPad and the company's Apple Retail Stores.[9] In 1986, he acquired the computer graphics division of Lucasfilm Ltd, which was spun off as Pixar Animation Studios.[10] He was credited in *Toy Story* (1995) as an executive producer. He remained CEO and majority shareholder at 50.1 percent until its acquisition by The Walt Disney Company in 2006,[11] making Jobs Disney's largest individual shareholder at seven percent and a member of Disney's Board of Directors.[12][13]

In 2003, Jobs was diagnosed with a pancreas neuroendocrine tumor. Though it was initially treated, he reported a hormone imbalance, underwent a liver transplant in 2009, and appeared progressively thinner as his health declined.[14] On medical leave for most of 2011, Jobs resigned as Apple CEO in August that year and was elected Chairman of the Board. On October 5, 2011, Jobs died of respiratory arrest related to his metastatic tumor. He

Steve Jobs

Jobs holding a white iPhone 4 at Worldwide Developers Conference 2010

Born	Steven Paul Jobs February 24, 1955[1][2] San Francisco, California, U.S.[1][2]
Died	October 5, 2011 (aged 56)[2] Palo Alto, California, U.S.
Nationality	American
Alma mater	Reed College (dropped out)

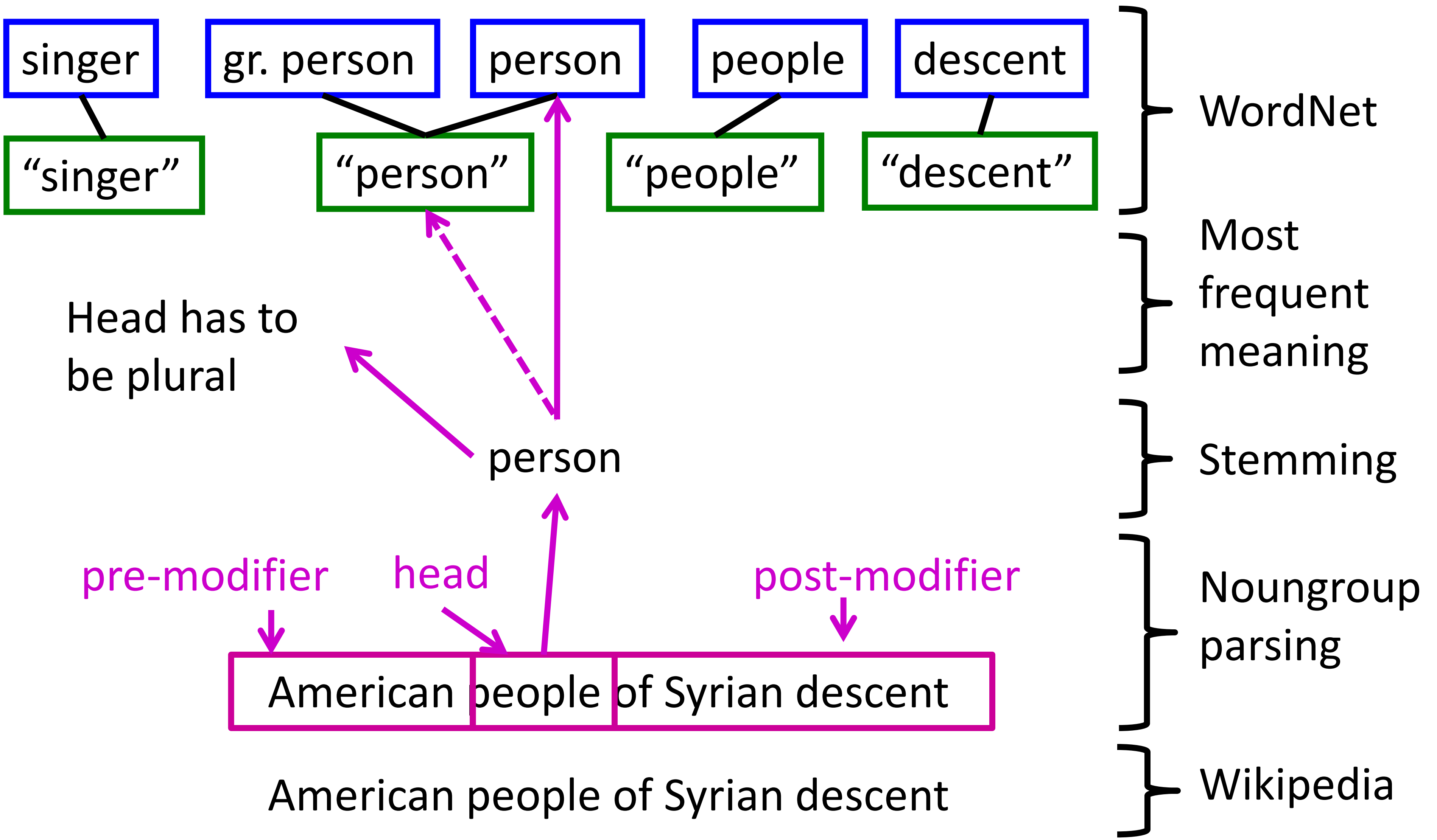
Wikipedia is rich in instances

- Wikipedia's categories contain classes

Categories: Steve Jobs | 1955 births | 2011 deaths | American adoptees | American billionaires | American chief executives | American computer businesspeople | American industrial designers | American inventors | American people of German descent | American people of Swiss descent | American people of Syrian descent | American technology company founders | American Zen Buddhists | Apple Inc. | Apple Inc. employees | Businesspeople from California | Businesspeople in software | Cancer deaths in California | Computer designers | Computer pioneers | Deaths from pancreatic cancer | Disney people | Internet pioneers | National Medal of Technology recipients | NeXT | Organ transplant recipients | People from the San Francisco Bay Area | Pescetarians | Reed College alumni

- But categories do not form a taxonomic hierarchy

YAGO: Link Wiki categories to WordNet



Extract instances from text: Hearst Patterns [M. Hearst 1992]

- Hearst defined lexico-syntactic patterns for type relationship:
  - *X such as Y; X like Y;*
  - *X and other Y; X including Y;*
  - *X, especially Y;*
- Find such patterns in text:
  - companies such as Apple
  - Google, Microsoft and other companies
  - Internet companies like Google and Facebook
- Derive type(Y,X)
  - type(Apple, company), type(Google, company), ...

Extract instances from tables [Kozareva/Hovy 2010, Dalvi et al. 2012]

- Start with a set of seeds:
  - cities = {Paris, Shanghai, Brisbane}

- Parse Web documents and find tables

Paris	France
Shanghai	China
Berlin	Germany
London	UK

Paris	Iliad
Helena	Iliad
Odysseus	Odysee
Rama	Mahabaratha

- If at least two seeds appear in a column, harvest the others:
  - type(Berlin, city), type(London, city)

Extract instances from lists, tables and text

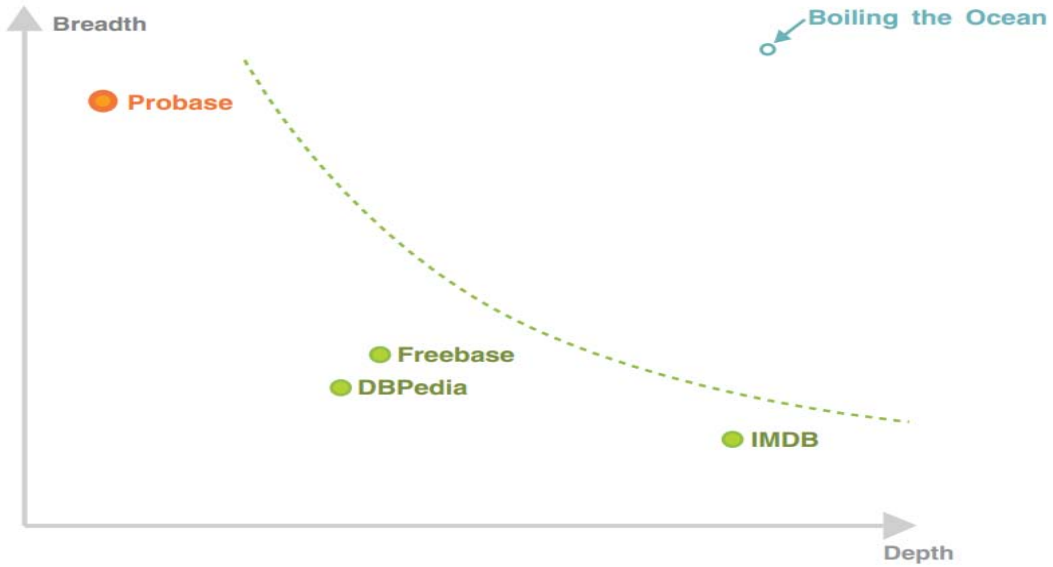
[Etzioni et al. 2004, Cohen et al. 2008, Mitchell et al. 2010]

- Start with seeds: a few class instances
  - Find lists, tables, text snippets that contain one or more seeds
- Extract candidates: noun phrases from vicinity
- Gather co-occurrence statistics
  - E.g. seed & candidate, candidate & className
- Rank candidates w.r.t. Pointwise Mutual Information

Precision drops for classes with sparse statistics
Harvested items are names, not entities
Canonicalization (de-duplication) unsolved

Probase: builds a taxonomy from the Web [Wu et al.2012]

- Extract with Hearst Patterns and isA Patterns
- Iterative Extraction
- Various kinds of scores supporting inferences
  - Typicality
    - "robin" is a more *typical* bird than a "penguin"
  - Vagueness
    - factors, items, things…
  - Ambiguity
    - apple, python…
  - Similarity
    - microsoft & google



Automatic Knowledge Harvesting

- Taxonomic Knowledge:
  - Entities and Classes
- Factual Knowledge:
  - Relations between Entities
- Emerging Knowledge:
  - New Entities & Relations
- Temporal Knowledge:
  - Validity Times of Facts
  - Common-sense Knowledge

Supervised Relation Extraction

- Pros
  - High quality supervision

- Cons
  - Very expensive to generate supervision
  - Not easy to add more relations
  - Cannot generalize to text from different domains

Semi-Supervised Relation Extraction

- Pros
  - Do not require manually labeling
- Cons
  - Sensitive to original set of seeds
  - Semantic drift at each iteration
  - Hard to measure confidences of patterns/extractions

Distant-Supervised Relation Extraction

- Pros
  - Can scale to the web, as no supervision required
  - Generalizes to text from different domains
  - Generates a lot more supervision than patterns

- Cons
  - False positive/negative training data (open research problem!)
  - Poor coverage for tail relations

Open Relation Extraction

- Pros
  - Require no assumptions about domain knowledge
  - Require no prior information of relation types
  - Extract a large number of relations with high coverage

- Cons
  - Result is noisy
  - Extracted relations are not canonicalized
  - Far from high quality

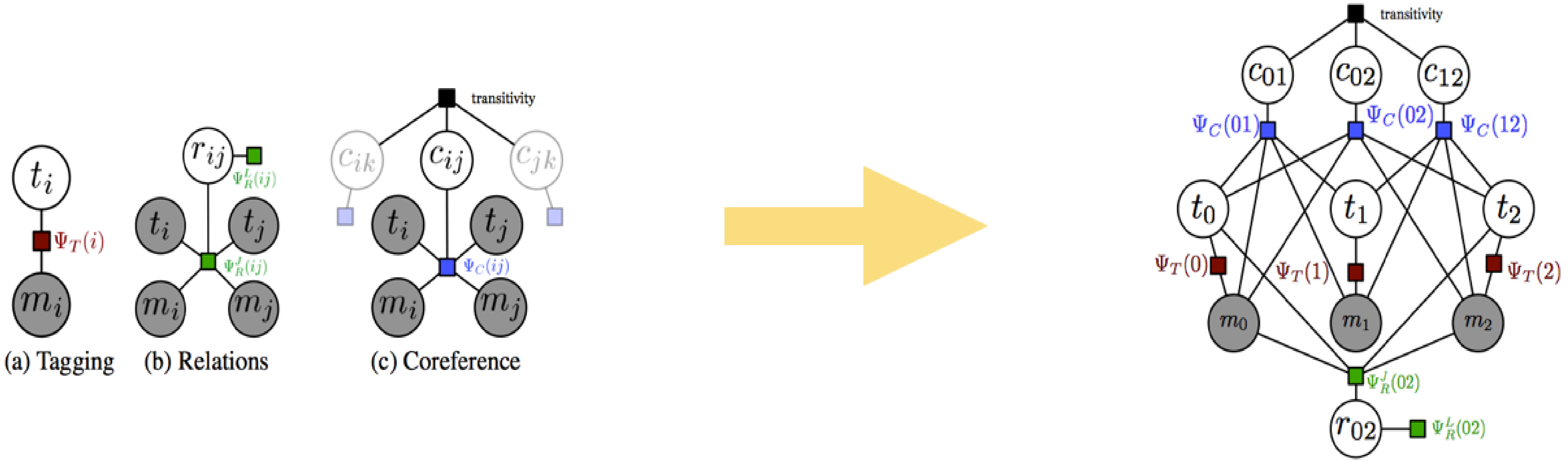
More Relation Extraction

- Joint relation extraction + other NLP tasks
  - Named Entity Tagging [Yao et al., 10]
  - Entity linking [Chen et al., 14]
  - Coreference Resolution [Singh et al., 13]
- Jointly perform relation extraction and link prediction [Bordes et al., 12; Weston et al., 13; Riedel et al., 13]

Joint Inference of Entities, Relations & Coreference

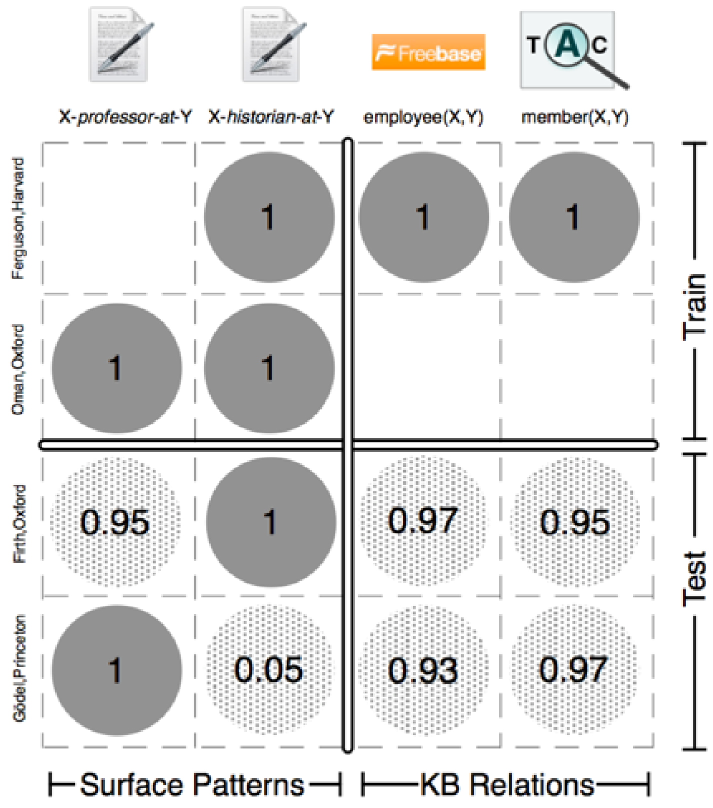
[Singh et al., 2013]

- Entities tagging, relation extraction, coreference resolution can mutually facilitate each other.
- Joint inference is effective to avoid cascading of errors.



Relation Extraction with Matrix Factorization [Riedal et al., 2013]

- Jointly infer relations across text (like OpenIE-style) and existing KBs, by writing everything down in a matrix and doing matrix completion.
- Universal Schema
  - union of all inputs: NL & KBs
- Learn implicit relations
  - "fill in" unobserved relations
  - via matrix factorization



Relation Taxonomy [N. Nakashole et al.: EMNLP2012, VLDB2012]

- Relations can be synonymous
  - E.g: “graduated from” ⇔ “obtained degree from”

- One relation can subsume another
  - E.g: “wife of” ⇔ “spouse of”

- Leveraging syntactic-lexical-ontological patterns
  - Pattern Extraction
  - Patten Typing
  - Synset Generation
  - Subsumption Mining

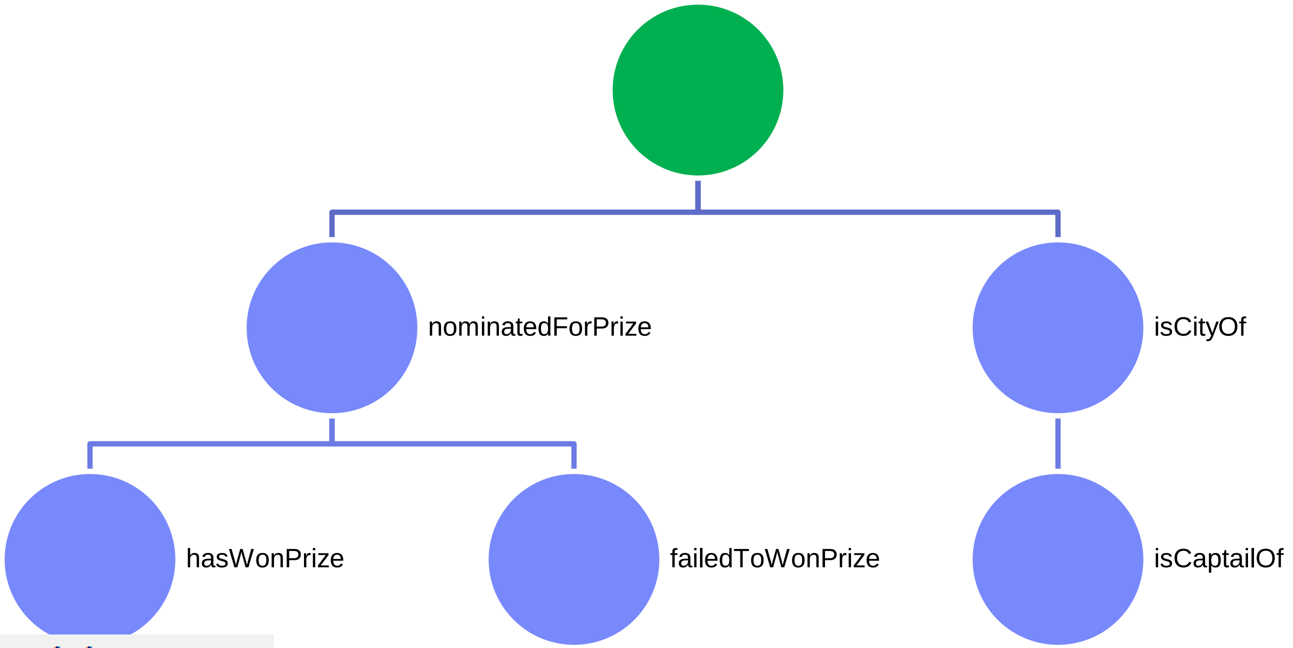
Relation Taxonomy [N. Nakashole et al.: EMNLP2012, VLDB2012]

- SOL: Syntactic-Ontologic-Lexical Patterns
  - A sequence of words, POS-tags, * and ontological types
- Example: <person>'s [adj] voice * <song>
  - Matches “Amy’s soft voice in ‘Rehab’”
  - Matches “Presley’s solid voice in his song ‘All shook up’”
  - Type signature: <person> X <song>
  - Support set: {(Amy, Rehab), (Presley, All Shook Up)}
- To generate SOL pattern from textual pattern:
  - Replace entities with their types
  - Decompose pattern into n-grams
  - Infrequent n-grams are replaced by wildcards

Relation Taxonomy [N. Nakashole et al.: EMNLP2012, VLDB2012]

- Pattern B is "syntactically more general" than Pattern A
  - Every string matches A also matches B
- Pattern B is "semantically more general" than Pattern A
  - A's support set ⊆ B's support set
- Synonymous: Pattern A ⊆sem Pattern B ∧ B ⊆sem A
  - A set of synonymous patterns form a pattern synset

Relation Taxonomy [N. Nakashole et al.: EMNLP2012, VLDB2012]



been awarded;

⊟ Synset

been awarded;

was awarded after;

was awarded on;

be awarded during;

[[con]] awarded [[det]];

350 000 SOL Patterns from Wikipedia, NYT Archive, ClueWeb

Automatic Knowledge Harvesting

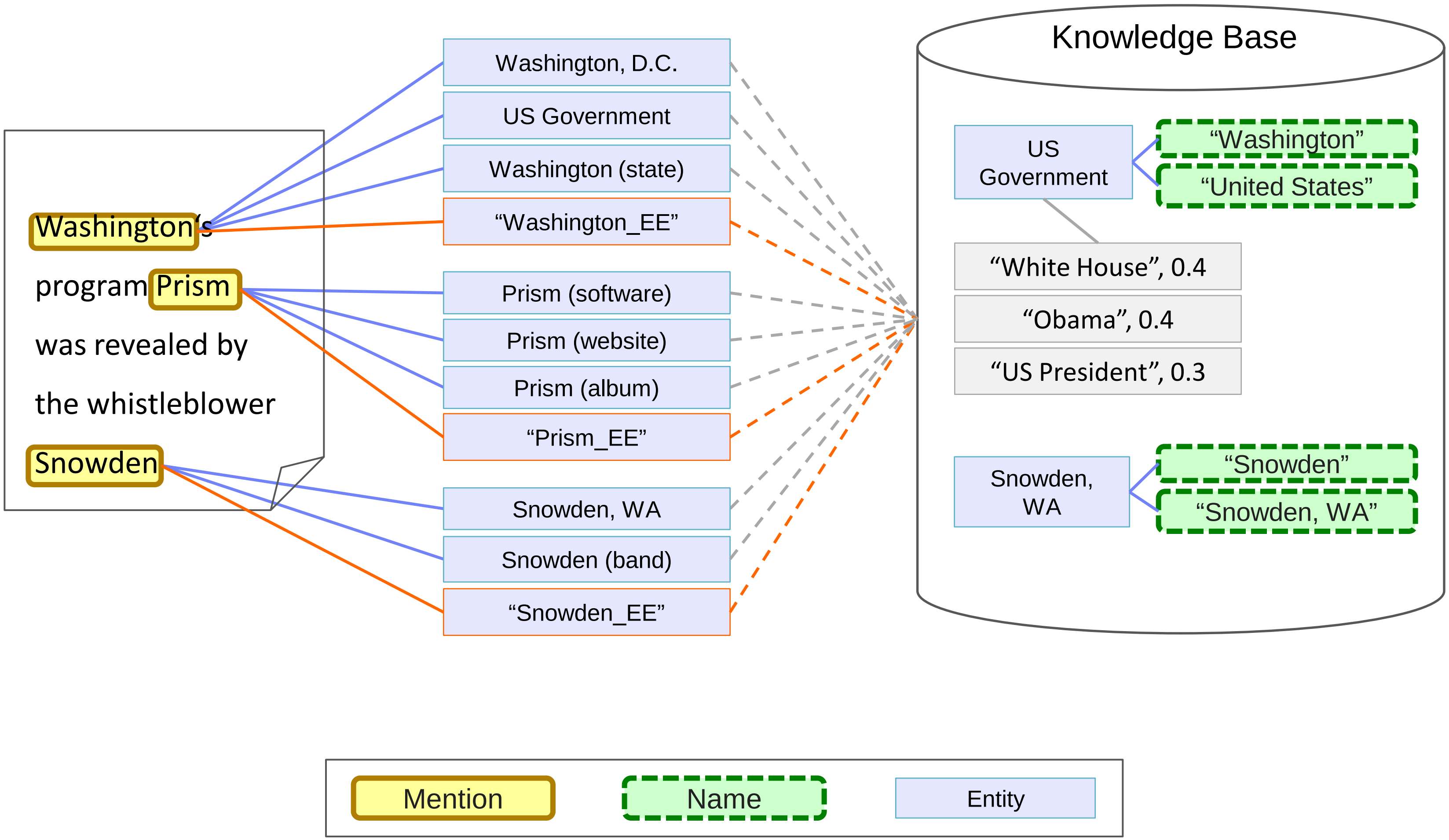
- Taxonomic Knowledge:
  - Entities and Classes
- Factual Knowledge:
  - Relations between Entities
- Emerging Knowledge:
  - New Entities & Relations
- Temporal Knowledge:
  - Validity Times of Facts
- Common-sense Knowledge

Emerging Entities [Hoffart et. al, WWW2014]

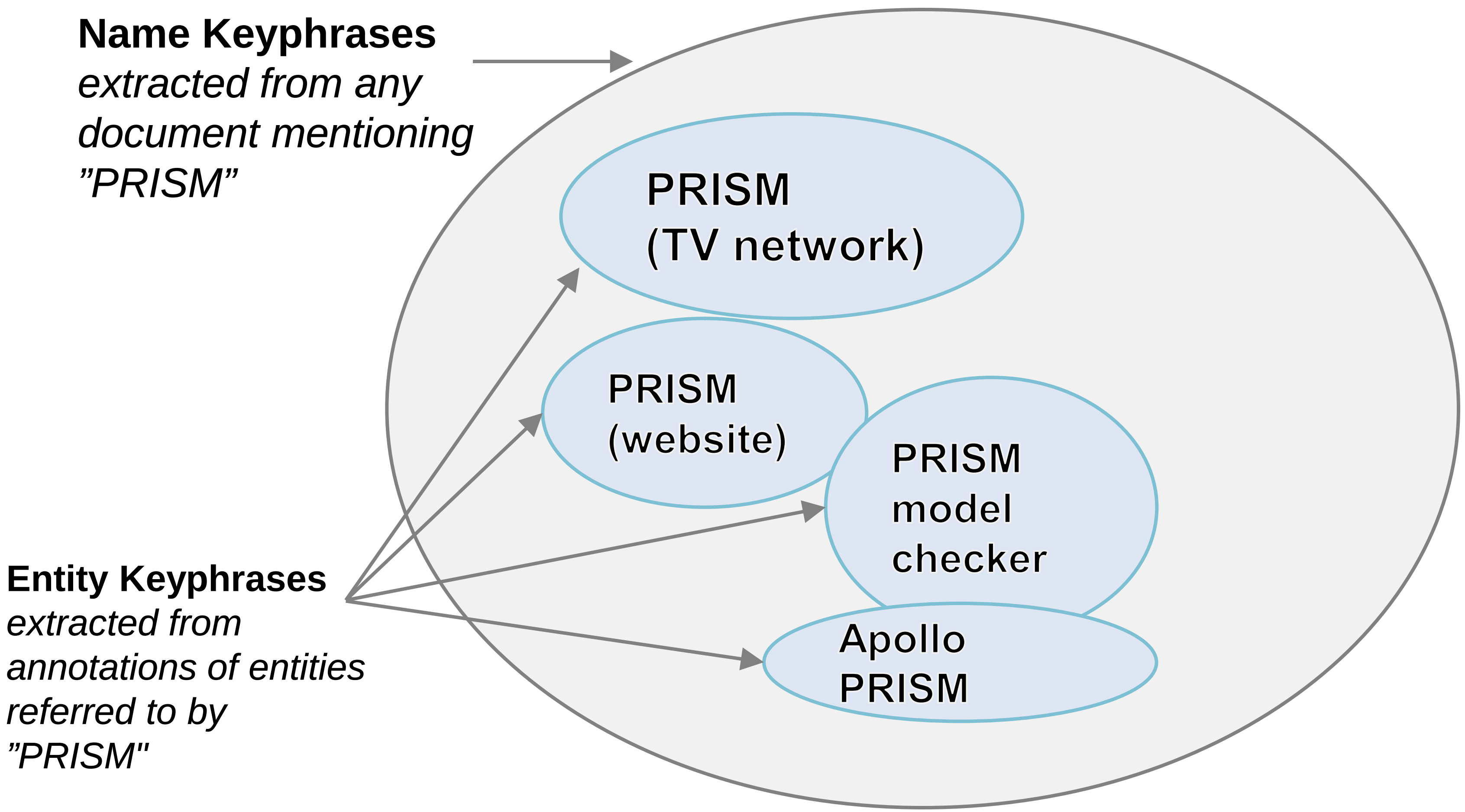
- Emerging entities are usually not covered by KB
  - E.g. "**Prism** program was revealed by the whistleblower **Snowden**."
- Identifying emerging entities using existing KB and Web text
  - KBs have lexicon of (name, entity) pairs
  - Key idea: profile emerging entities from the Web
  - Assumption: for one name, one emerging entity

	New Entity	Existing Entity
New Name	assumption	✖
Existing Name	?	disambiguation

Emerging Entities [Hoffart et. al, WWW2014]

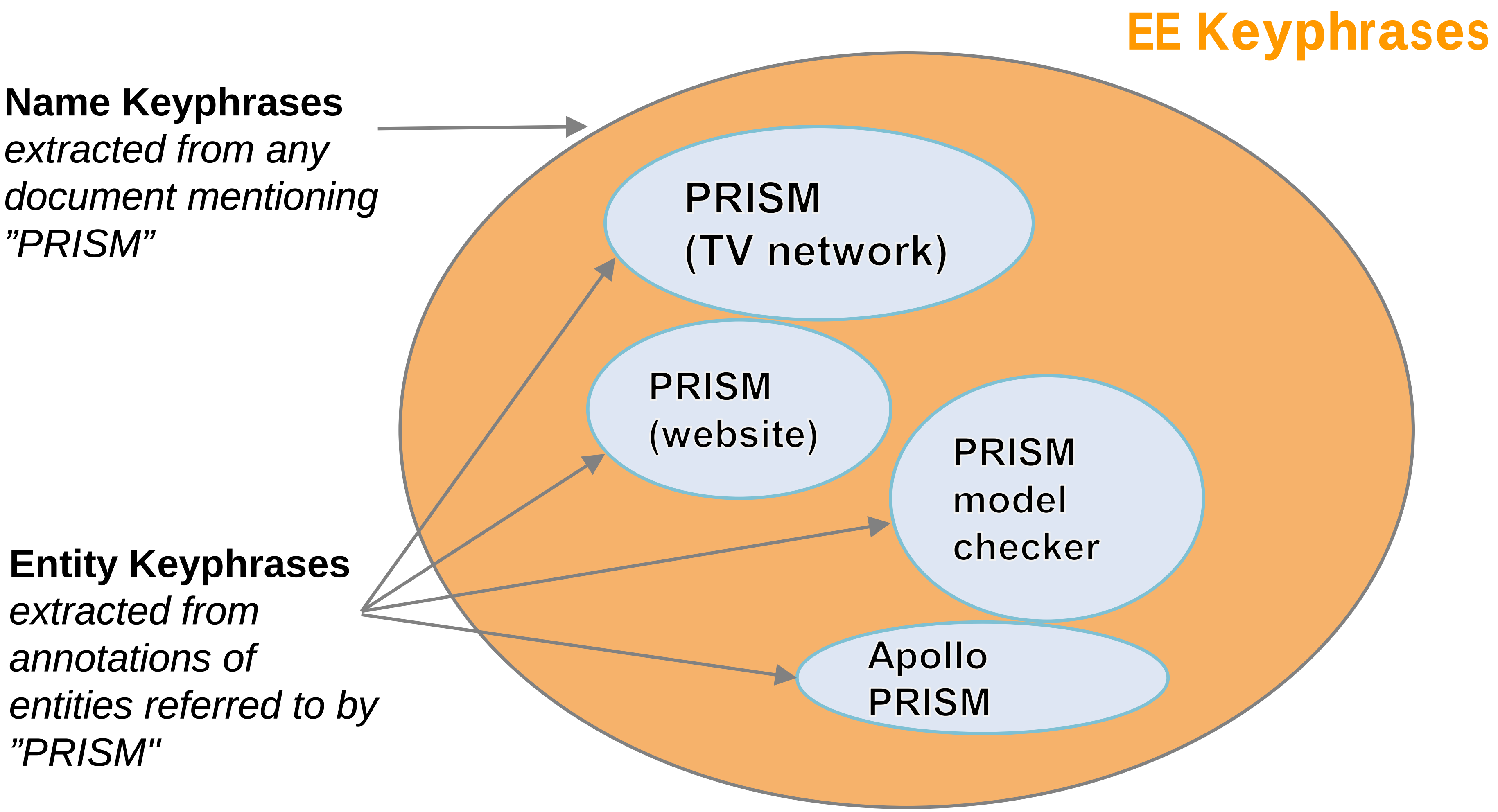


Emerging Entities [Hoffart et. al, WWW2014]



keyphrases defined by POS pattern filters

Emerging Entities [Hoffart et. al, WWW2014]



keyphrases defined by POS pattern filters

Emerging Relations: OpenIE [Fader et al. 2011, Lin et al. 2012, Mausam et al. 2012]

- Extracting phrases as relations
  - *Madame Bruni in her happy marriage with the French president …*
  - *The first lady had a passionate affair with Stones singer Mick …*
  - *Natalie was honored by the Oscar …*

- Grouping phrases with relation taxonomy (e.g. Patty)
  - *{cover songs, interpretation of, singing of, voice in, …}* ⇔ *SingerCoversSong*
  - *{classic piece of, 's old song, written by, composition of, …}* ⇔ *MusicianCreatesSong*

- Grouping phrases with paraphrase DBs

Automatic Knowledge Harvesting

- Taxonomic Knowledge:
  - Entities and Classes
- Factual Knowledge:
  - Relations between Entities
- Emerging Knowledge:
  - New Entities & Relations
- Temporal Knowledge:
  - Validity Times of Facts
- Common-sense Knowledge

Knowledge is temporal

- Which facts for given relations hold at what ***time point*** or during which ***time intervals*** ?
  - hasWonPrize (JimGray, TuringAward) [1998]
  - capitalOf (Berlin, Germany) [1990, now]
  - capitalOf (Bonn, Germany) [1949, 1989]
- How can we query & reason on entity-relationship facts in a “time-travel“ manner?
  - US president‘s wife when Steve Jobs died?
  - Students of David Blei while he was at Princeton?

Explicit vs. Implicit Temporal Information

Nicolas Sarkozy

From Wikipedia, the free encyclopedia

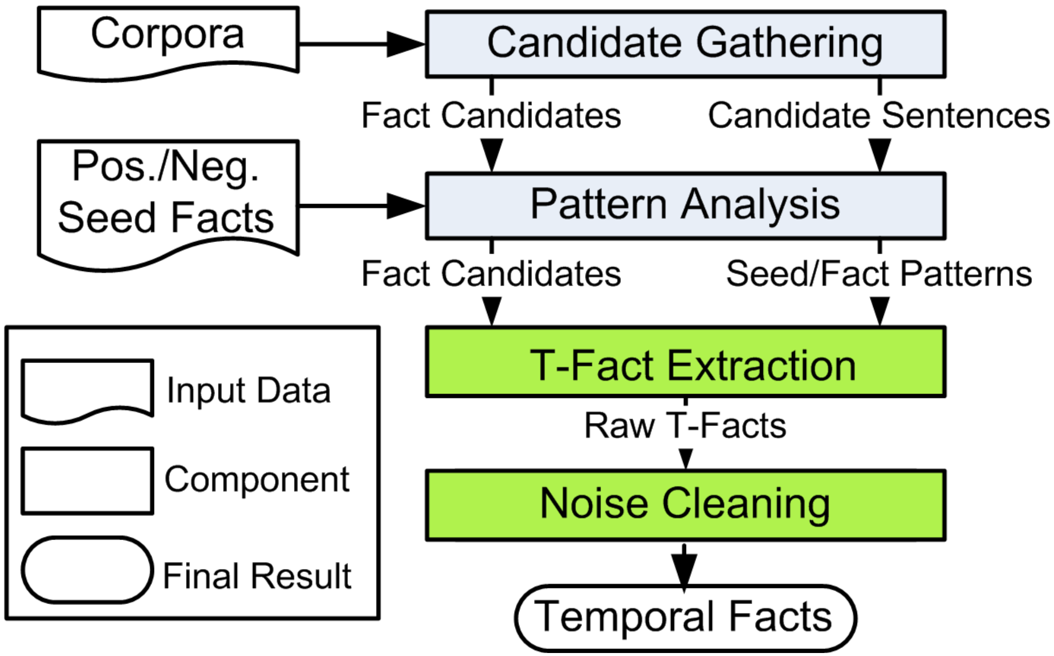
Nicolas Sarkozy (pronounced [ni.kɔ.la saʁ.kɔ.zi] (listen), born **Nicolas Paul Stéphane Sarközy de Nagy-Bocsa**; 28 January 1955) is the 23rd and current President of the French Republic and *ex officio* Co-Prince of Andorra. He assumed the office on 16 May 2007 after defeating the Socialist Party candidate Ségolène Royal 10 days earlier.

Before his presidency, he was leader of the Union for a Popular Movement (UMP). Under Jacques Chirac's presidency he served as Minister of the Interior in Jean-Pierre Raffarin's (UMP) first two governments (from May 2002 to March 2004), then was appointed Minister of Finances in Raffarin's last government (March 2004 to May 2005) and again Minister of the Interior in Dominique de Villepin's government (2005–2007).

Sarkozy was also president of the General council of the Hauts-de-Seine department from 2004 to 2007 and mayor of Neuilly-sur-Seine, one of the wealthiest communes of France from 1983 to 2002. He was Minister of the Budget in the government of Édouard Balladur (RPR, predecessor of the UMP) during François Mitterrand's last term.

Extract Temporal Facts from Text [Y. Wang et al. 2011]

1) Candidate gathering:
   extract pattern & entities of basic facts and time expression
2) Pattern analysis:
   use seeds to quantify strength of candidates
3) Label propagation:
   construct weighted graph of hypotheses and minimize loss function
4) Constraint reasoning:
   use ILP for temporal consistency



Automatic Knowledge Harvesting

- Taxonomic Knowledge:
  - Entities and Classes
- Factual Knowledge:
  - Relations between Entities
- Emerging Knowledge:
  - New Entities & Relations
- Temporal Knowledge:
  - Validity Times of Facts
  - Common-sense Knowledge

Common-sense Knowledge

- About general concepts instead of specific entities

Apples are green, red, round, juicy, …
but not fast, funny, verbose, …

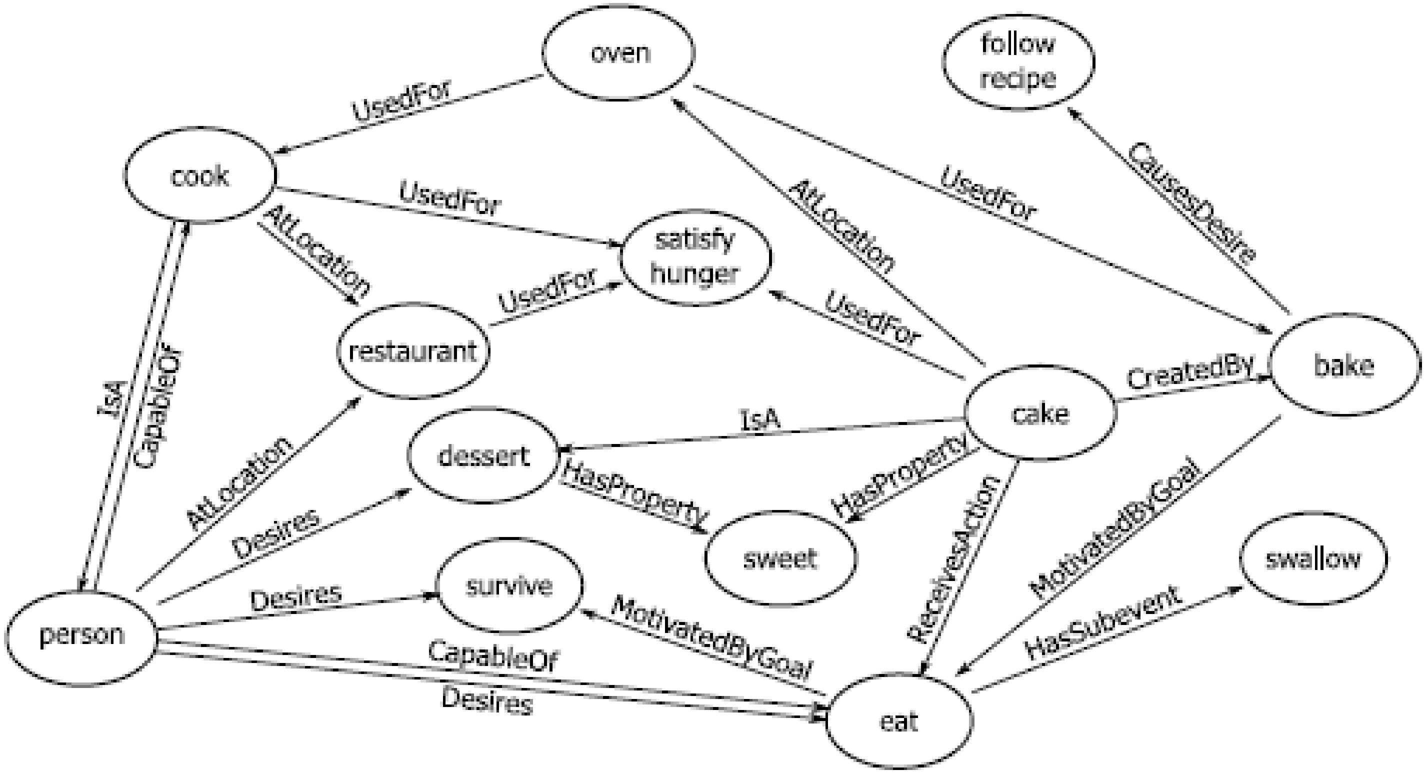
Snakes can crawl, doze, bite, hiss, …
but not run, fly, laugh, write, …

Pots and pans are in the kitchen, cupboard, …
but not in in the bedroom, in your pocket, in the sky, …

- Critical for reasonings/inferences

Crowdsourcing for Common-sense Knowledge

ConceptNet 5: 3.9M concepts & 12.5M edges



http://conceptnet5.media.mit.edu/

Harvesting Common-sense Knowledge from Text

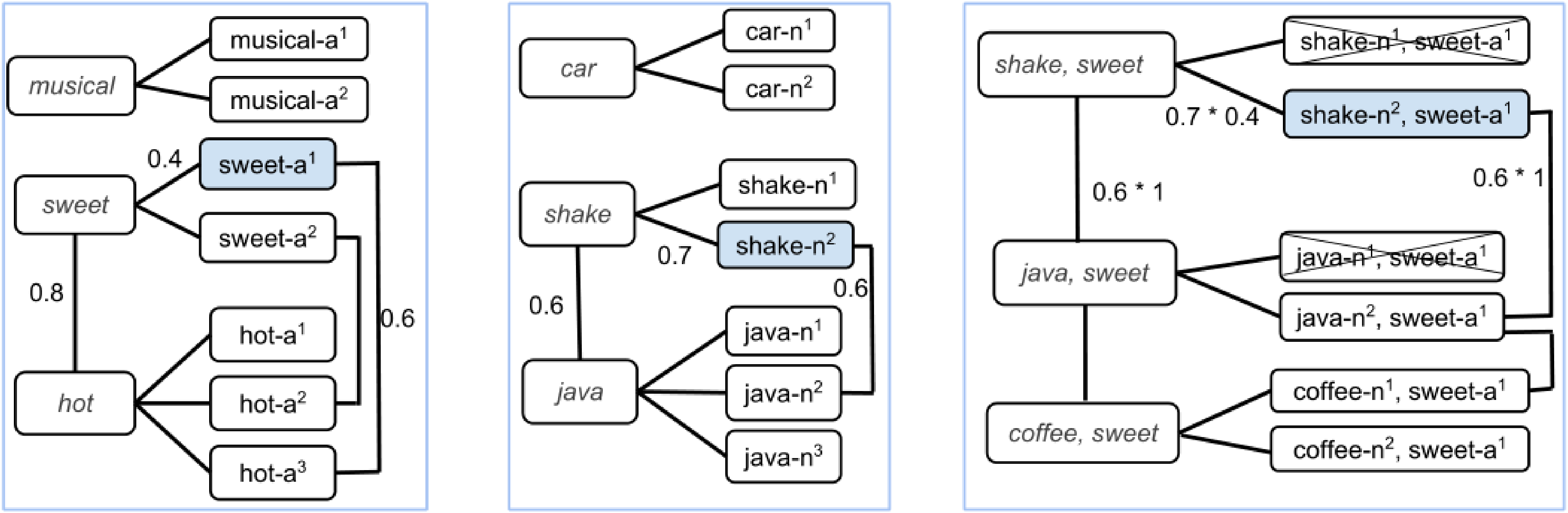
[Tandon et al. WSDM 2014]

- Compute the ranges for common-sense relations
  - hasTaste: sweet, sour, spicy, …
- Compute the domains for common-sense relations
  - hasTaste: shake (milk shake), juice…
- Compute assertions
  - hasTaste: { shake/sweet, … }

- For all 3 tasks, use label propagation on a graph with few seeds from WordNet and with edges from n-gram corpus.

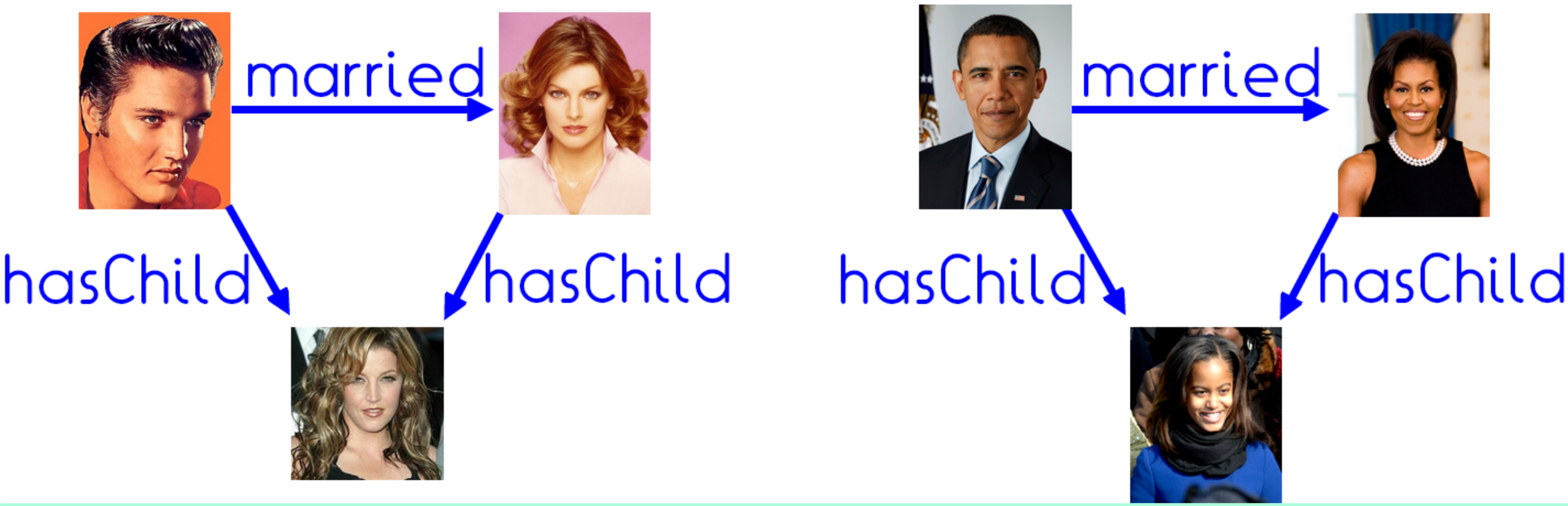
Harvesting Common-sense Knowledge from Text

[Tandon et al. WSDM 2014]



WebChild: 4M triples for 19 common-sense relations

Common-sense Rules Mining on KBs [Galarraga et al. WWW'13]



AMIE inferred commonsense rules from YAGO, such as

$$marriedTo(x,y) \wedge livesIn(x,z) \Rightarrow livesIn(y,z)$$
$$bornIn(x,y) \wedge locatedIn(y,z) \Rightarrow citizenOf(x,z)$$
$$hasWonPrize(x, LeibnizPreis) \Rightarrow livesIn(x, Germany)$$

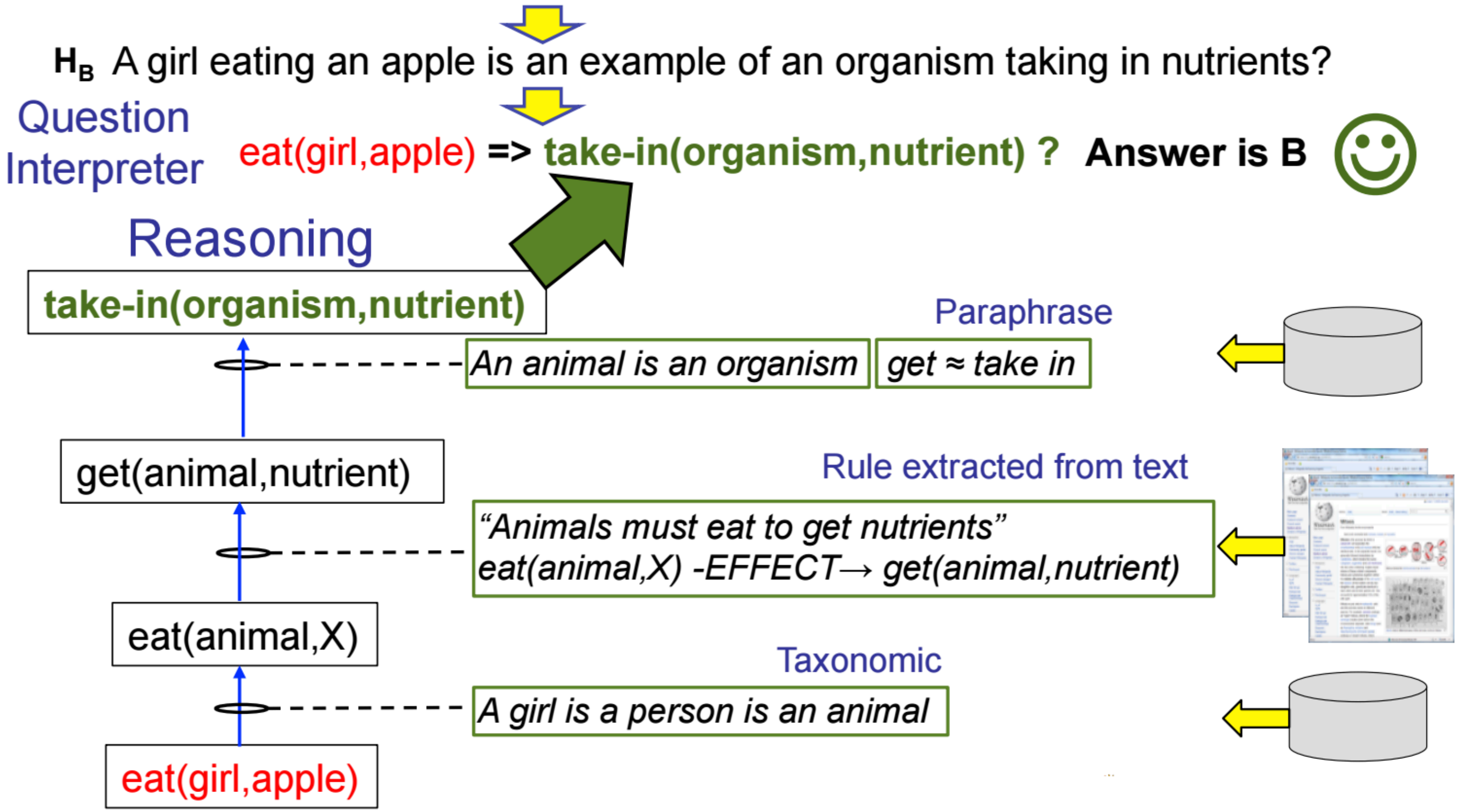
http://www.mpi-inf.mpg.de/departments/ontologies/projects/amie/

$$married(x,y) \wedge hasChild(x,z) \Rightarrow hasChild(y,z)$$

Common-sense Question Answering

1 Which example describes an organism taking in nutrients?

A a dog burying a bone
B a girl eating an apple
C an insect crawling on a leaf
D a boy planting tomatoes in a garden



Question dataset: http://allenai.org/content/data/Ariscienceexams.txt