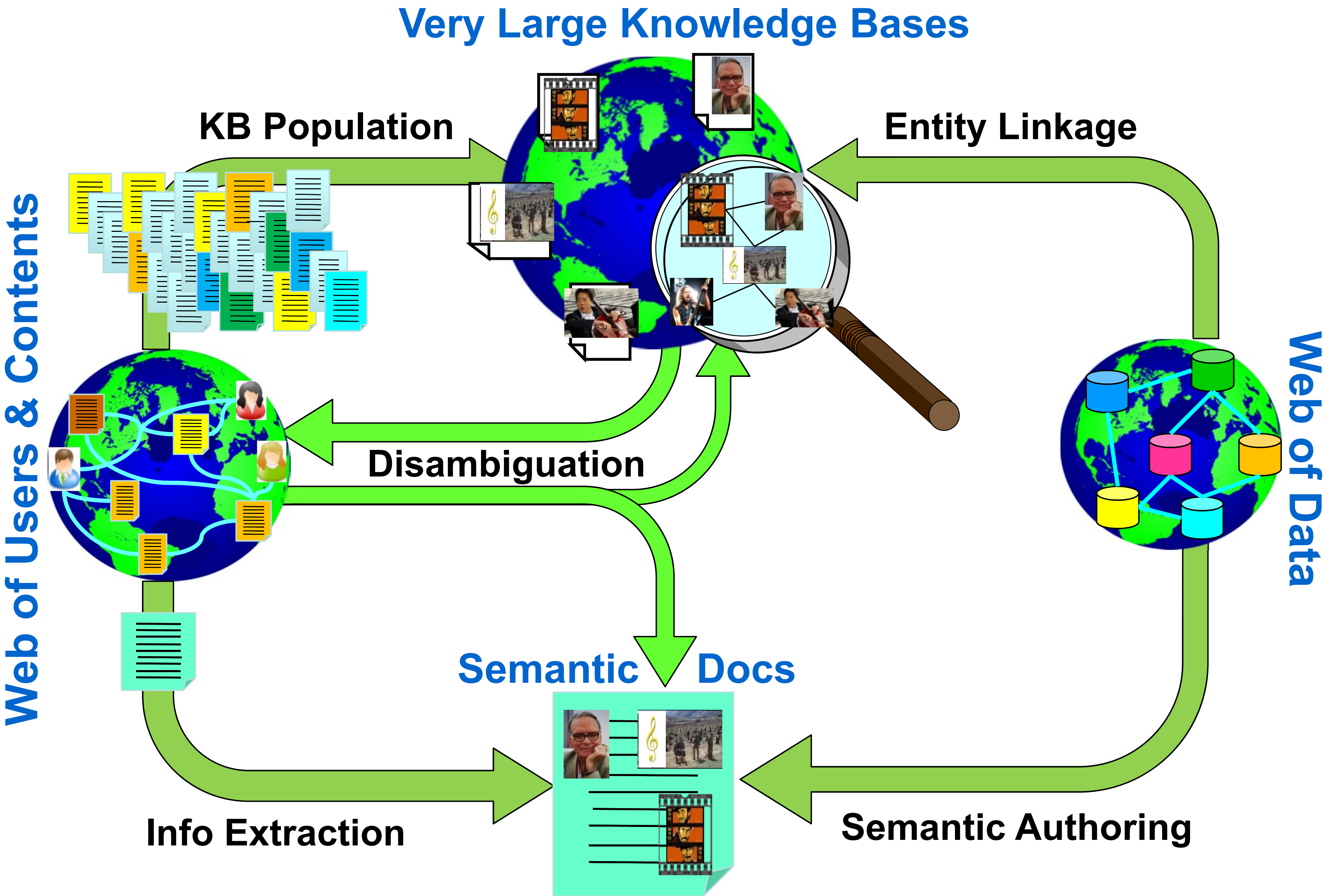
Knowledge Harvesting from Text and Web Sources

Fabian Suchanek & Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany
http://suchanek.name/
http://www.mpi-inf.mpg.de/~weikum/

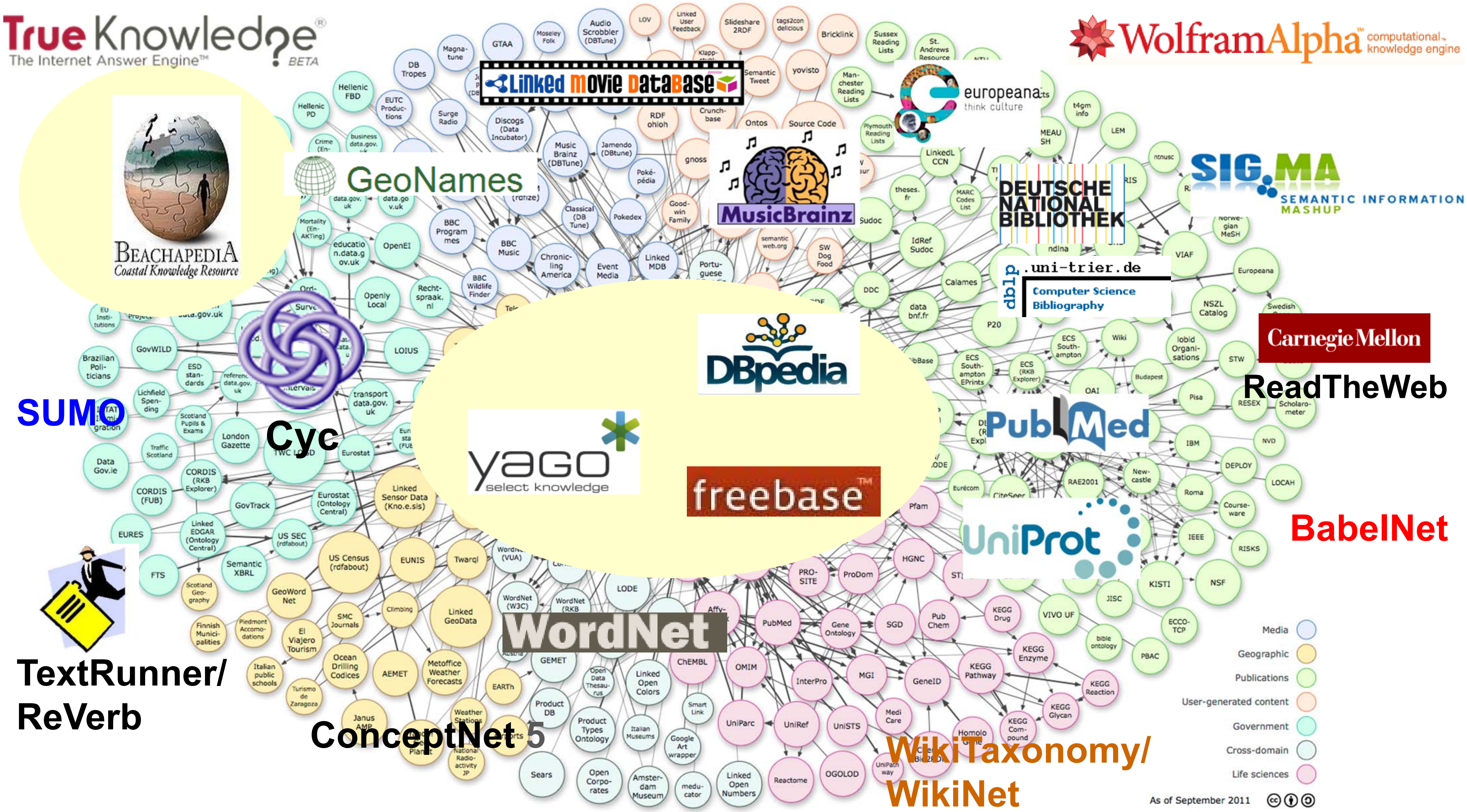
http://www.mpi-inf.mpg.de/yago-naga/icde2013-tutorial/

Turn Web into Knowledge Base



Web of Data: RDF, Tables, Microdata

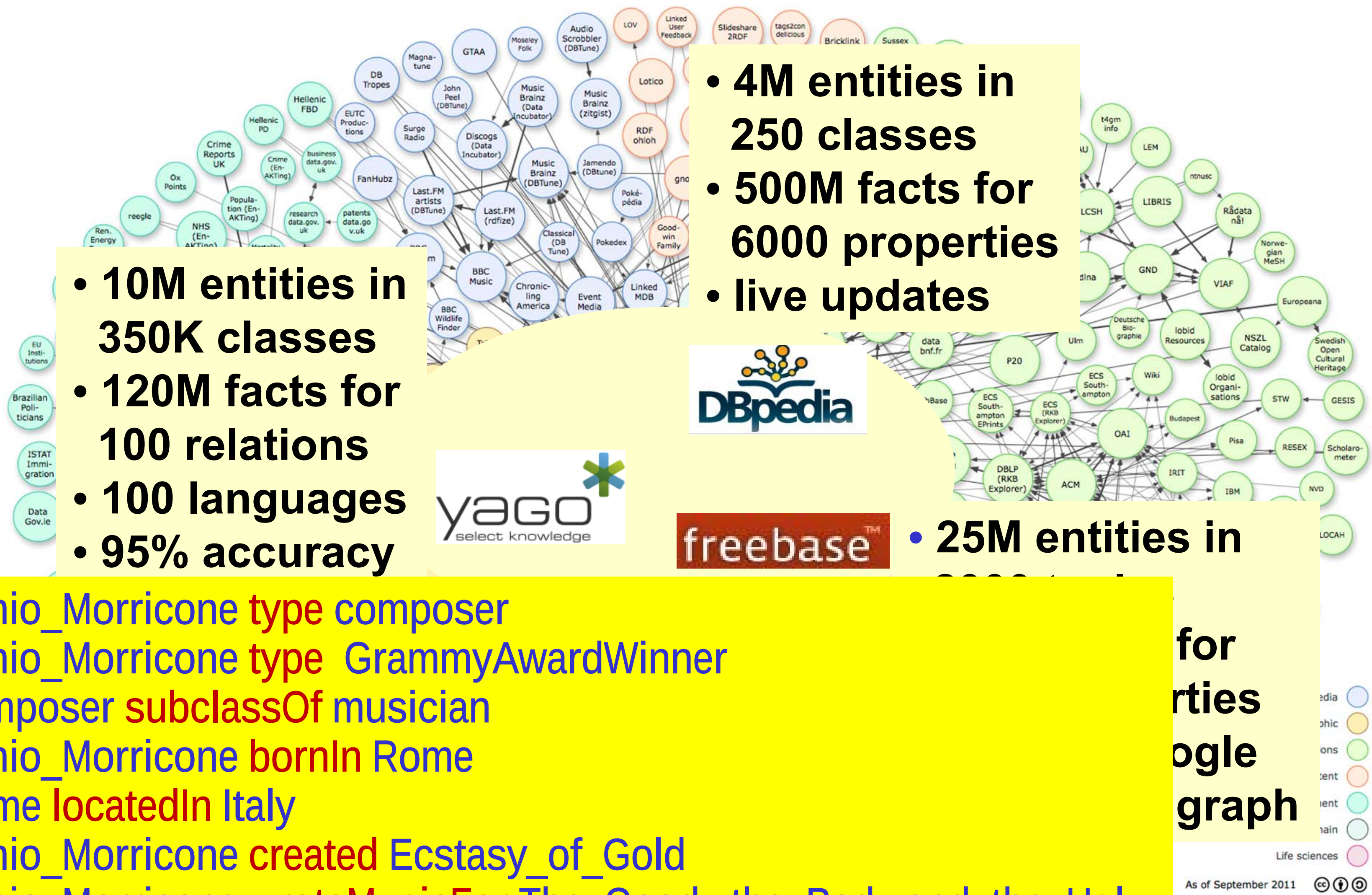
30 Bio. SPO triples (RDF) and growing



http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.png

Web of Data: RDF, Tables, Microdata

30 Bio. SPO triples (RDF) and growing



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

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

- **25M entities in**

for
rties
ogle
graph

Ennio_Morricone type composer
Ennio_Morricone type GrammyAwardWinner
composer subclassOf musician
Ennio_Morricone bornIn Rome
Rome locatedIn Italy
Ennio_Morricone created Ecstasy_of_Gold
Ennio_Morricone wroteMusicFor The_Good,_the_Bad_,and_the_Ugly
Sergio_Leone directed The_Good,_the_Bad_,and_the_Ugly

011-09-19_colored.png

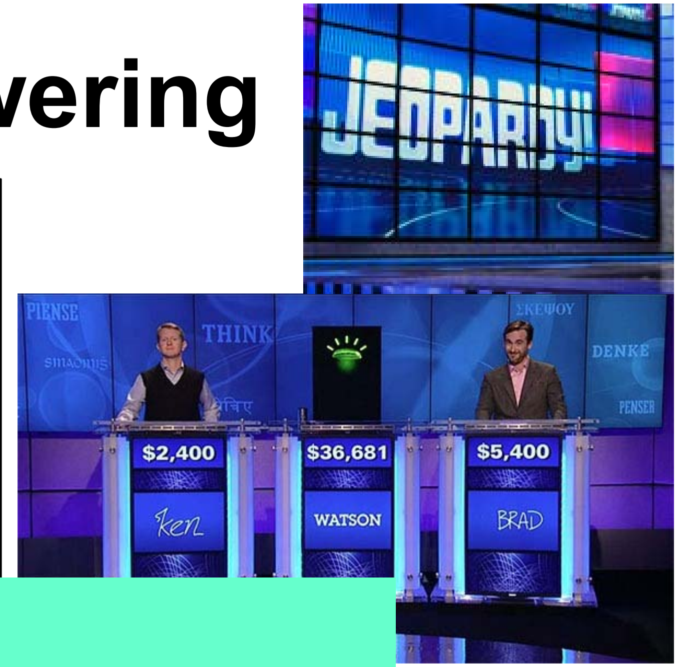
Knowledge for Intelligence

Enabling technology for:

- ✶ **disambiguation** in written & spoken natural language
- ✶ **deep reasoning** (e.g. QA to win quiz game)
- ✶ **machine reading** (e.g. to summarize book or corpus)
- ✶ **semantic search** in terms of entities&relations (not keywords&pages)
- ✶ **entity-level linkage** for the Web of Data

- ★ Politicians who are also scientists?
- ★ European composers who have won film music awards?
- ★ Australian professors who founded Internet companies?
- ★ Relationships between
  John Lennon, Lady Di, Heath Ledger, Steve Irwin?
- ★ Enzymes that inhibit HIV?
  Influenza drugs for teens with high blood pressure?
  ...

Use Case: Question Answering



William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia" inspired this author's most famous novel

This town is known as "Sin City" & its downtown is "Glitter Gulch"

Q: Sin City ?
→ movie, graphical novel, nickname for city, …
A: Vegas ? Strip ?
→ Vega (star), Suzanne Vega, Vincent Vega, Las Vegas, …
→ comic strip, striptease, Las Vegas Strip, …

question classification & decomposition → knowledge back-ends

WIKIPEDIA The Free Encyclopedia

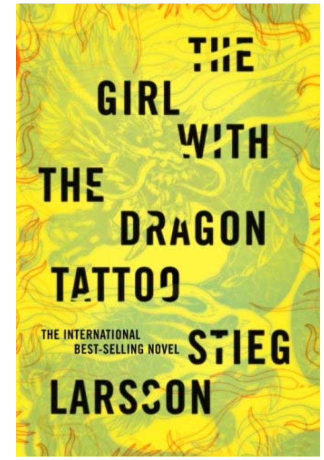
DBpedia

freebase

yago select knowledge

D. Ferrucci et al.: Building Watson. AI Magazine, Fall 2010.
IBM Journal of R&D 56(3/4), 2012: This is Watson.

Use Case: Machine Reading



It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at ... te on th ... and of Hedeby. The old man d... Blomkvist in by promising solid evidence against Wennerström. Blomkvist ag... spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V... niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco... inted with the men ... the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of H...

Af... overing that Salander has hacked into his co... persuade... assist him with research. They event... lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou... gations for Dragan Armansky, who, in tu... ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

uncleOf · same · same · same · owns · uncleOf · hires · same · affairWith · affairWith · enemyOf · same · headOf · same

O. Etzioni, M. Banko, M.J. Cafarella: Machine Reading, AAAI ,06
T. Mitchell et al.: Populating the Semantic Web by Macro-Reading Internet Text, ISWC'09

Outline

✓ Motivation

★ Machine Knowledge

★ Taxonomic Knowledge: Entities and Classes

★ Contextual Knowledge: Entity Disambiguation

★ Linked Knowledge: Entity Resolution

★ Temporal & Commonsense Knowledge

★ Wrap-up

http://www.mpi-inf.mpg.de/yago-naga/icde2013-tutorial/

Spectrum of Machine Knowledge (1)

factual knowledge:
bornIn (SteveJobs, SanFrancisco), hasFounded (SteveJobs, Pixar),
hasWon (SteveJobs, NationalMedalOfTechnology), livedIn (SteveJobs, PaloAlto)

taxonomic knowledge (ontology):
instanceOf (SteveJobs, computerArchitects), instanceOf(SteveJobs, CEOs)
subclassOf (computerArchitects, engineers), subclassOf(CEOs, businesspeople)

lexical knowledge (terminology):
means (“Big Apple“, NewYorkCity), means (“Apple“, AppleComputerCorp)
means (“MS“, Microsoft) , means (“MS“, MultipleSclerosis)

contextual knowledge (entity occurrences, entity-name disambiguation)
maps (“Gates and Allen founded the Evil Empire“,
BillGates, PaulAllen, MicrosoftCorp)

linked knowledge (entity equivalence, entity resolution):
hasFounded (SteveJobs, Apple), isFounderOf (SteveWozniak, AppleCorp)
sameAs (Apple, AppleCorp), sameAs (hasFounded, isFounderOf)

Spectrum of Machine Knowledge (2)

multi-lingual knowledge:
meansInChinese („乔戈里峰“, K2), meansInUrdu („کے ٹو“, K2)
meansInFr („école“, school (institution)), meansInFr („banc“, school (of fish))

temporal knowledge (fluents):
hasWon (SteveJobs, NationalMedalOfTechnology)@1985
marriedTo (AlbertEinstein, MilevaMaric)@[6-Jan-1903, 14-Feb-1919]
presidentOf (NicolasSarkozy, France)@[16-May-2007, 15-May-2012]

spatial knowledge:
locatedIn (YumbillaFalls, Peru), instanceOf (YumbillaFalls, TieredWaterfalls)
hasCoordinates (YumbillaFalls, 5°55‘11.64“S 77°54‘04.32“W),
closestTown (YumbillaFalls, Cuispes), reachedBy (YumbillaFalls, RentALama)

Spectrum of Machine Knowledge (3)

ephemeral knowledge (dynamic services):

wsdl:getSongs (musician ?x, song ?y), wsdl:getWeather (city?x, temp ?y)

common-sense knowledge (properties):

hasAbility (Fish, swim), hasAbility (Human, write),
hasShape (Apple, round), hasProperty (Apple, juicy),
hasMaxHeight (Human, 2.5 m)

common-sense knowledge (rules):

$\forall$ **x: human(x)** $\Rightarrow$ **male(x)** $\vee$ **female(x)**

$\forall$ **x: (male(x)** $\Rightarrow \neg$ **female(x))** $\wedge$ **(female(x))** $\Rightarrow \neg$ **male(x))**

$\forall$ **x: human(x)** $\Rightarrow$ **(**$\exists$ **y: mother(x,y)** $\wedge \exists$ **z: father(x,z))**

$\forall$ **x: animal(x)** $\Rightarrow$ **(hasLegs(x)** $\Rightarrow$ **isEven(numberOfLegs(x))**

Spectrum of Machine Knowledge (4)

free-form knowledge (open IE):

hasWon (MerylStreep, AcademyAward)

occurs („Meryl Streep“, „celebrated for“, „Oscar for Best Actress“)

occurs („Quentin“, „nominated for“, „Oscar“)

multimodal knowledge (photos, videos):

JimGray

JamesBruceFalls

social knowledge (opinions):

admires (maleTeen, LadyGaga), supports (AngelaMerkel, HelpForGreece)

epistemic knowledge ((un-)trusted beliefs):

believe(Ptolemy,hasCenter(world,earth)),

believe(Copernicus,hasCenter(world,sun))

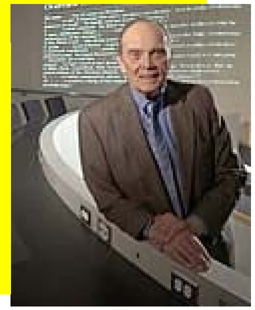
believe (peopleFromTexas, bornIn(BarackObama,Kenya))

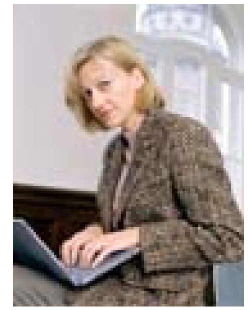
History of Knowledge Bases

Cyc project (1984-1994)
cont'd by Cycorp Inc.

WordNet project (1985-now)

Cyc and WordNet are hand-crafted knowledge bases





Doug Lenat:
„The more you know, the more (and faster) you can learn."

George Miller **Christiane Fellbaum**

$\forall$ x: human(x) $\Rightarrow$ male(x) $\vee$ female(x)
$\forall$ x: (male(x) $\Rightarrow \neg$ female(x)) $\wedge$
(female(x) $\Rightarrow \neg$ male(x))
$\forall$ x: mammal(x) $\Rightarrow$ (hasLegs(x)
$\Rightarrow$ isEven(numberOfLegs(x))
$\forall$x: human(x) $\Rightarrow$
($\exists$ y: mother(x,y) $\wedge$ $\exists$ z: father(x,z))
$\forall$ x $\forall$ e : human(x) $\wedge$ remembers(x,e)
$\Rightarrow$ happened(e) < now


WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: enterprise [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **enterprise**, endeavor, endeavour (a purposeful or industrious undertaking (especially one that requires effort or boldness)) *"he had doubts about the whole enterprise"*
- S: (n) **enterprise** (an organization created for business ventures) *"a growing enterprise must have a bold leader"*
- S: (n) **enterprise**, enterprisingness, initiative, go-ahead (readiness to embark on bold new ventures)

Large-Scale Universal Knowledge Bases

Yago: 10 Mio. entities, 350 000 classes,
180 Mio. facts, 100 properties, 100 languages
high accuracy, no redundancy, limited coverage
http://yago-knowledge.org

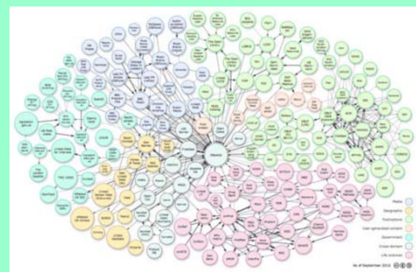
Dbpedia: 4 Mio. entities, 250 classes,
500 Mio. facts, 6000 properties
high coverage, live updates
http://dbpedia.org

Freebase: 25 Mio. entities, 2000 topics,
100 Mio. facts, 4000 properties
interesting relations (e.g., romantic affairs)
http://freebase.com

NELL: 300 000 entity names, 300 classes, 500 properties,
1 Mio. beliefs, 15 Mio. low-confidence beliefs
learned rules
http://rtw.ml.cmu.edu/rtw/

ReadTheWeb

and more … plus Linked Data



Some Publicly Available Knowledge Bases

YAGO:	yago-knowledge.org
Dbpedia:	dbpedia.org
Freebase:	freebase.com
Entitycube:	research.microsoft.com/en-us/projects/entitycube/
NELL:	rtw.ml.cmu.edu
DeepDive:	research.cs.wisc.edu/hazy/demos/deepdive/index.php/Steve_Irwin
Probase:	research.microsoft.com/en-us/projects/probase/
KnowItAll / ReVerb:	openie.cs.washington.edu
	reverb.cs.washington.edu
PATTY:	www.mpi-inf.mpg.de/yago-naga/patty/
BabelNet:	lcl.uniroma1.it/babelnet
WikiNet:	www.h-its.org/english/research/nlp/download/wikinet.php
ConceptNet:	conceptnet5.media.mit.edu
WordNet:	wordnet.princeton.edu

Linked Open Data:	linkeddata.org

Take-Home Lessons

Knowledge bases are real, big, and interesting

Dbpedia, Freebase, Yago, and a lot more
knowledge representation mostly in RDF plus …

Knowledge bases are infrastructure assets
for intelligent applications

semantic search, machine reading, question answering, …

Variety of focuses and approaches
with different strengths and limitations

Open Problems and Opportunities

Rethink knowledge representation

beyond RDF (and OWL ?)
old topic in AI, fresh look towards big KBs

High-quality interlinkage between KBs

at level of entities and classes

High-coverage KBs for vertical domains

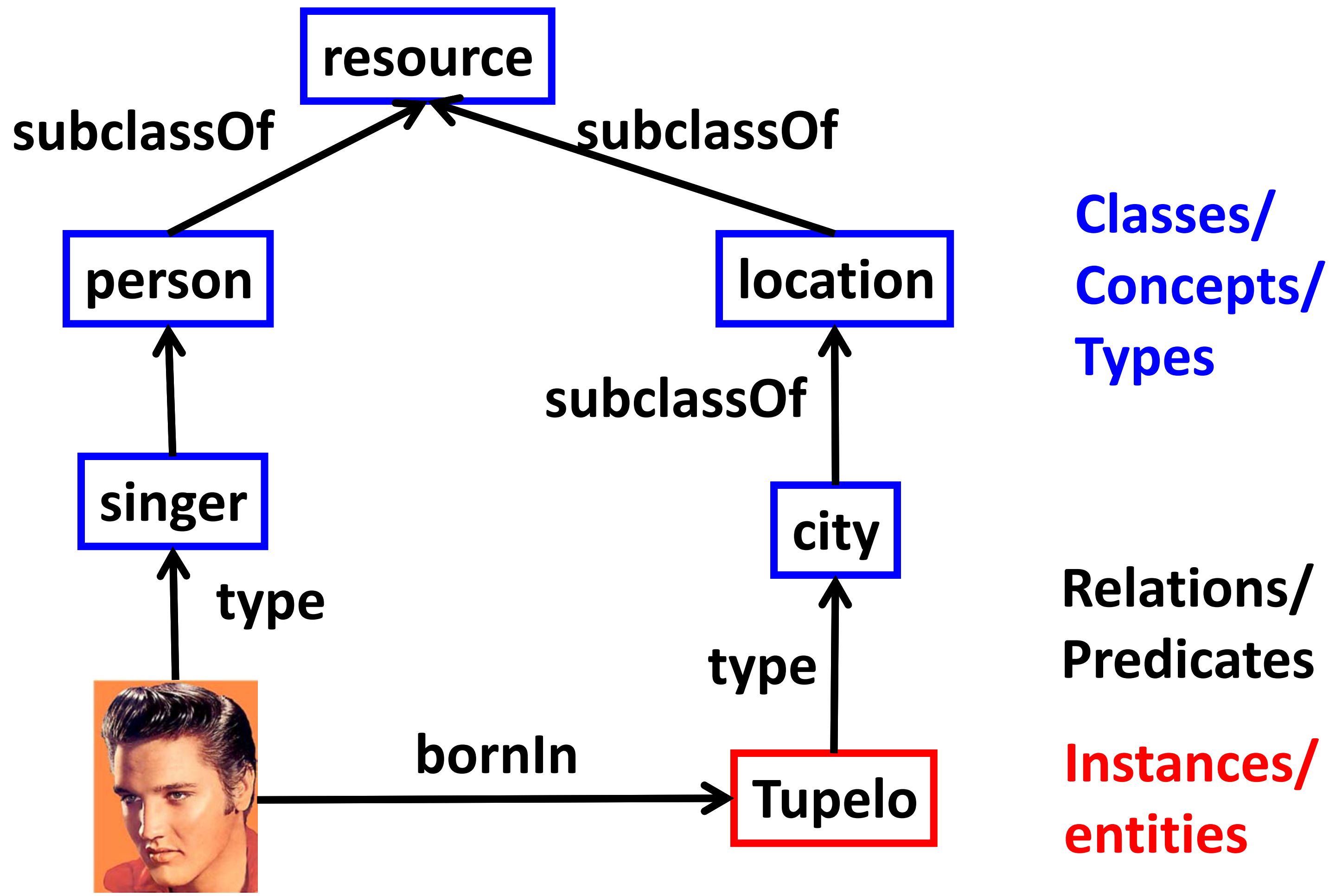
music, literature, health, football, hiking, etc.

Outline

- ✓ Motivation
- ✓ Machine Knowledge
- ★ Taxonomic Knowledge: Entities and Classes
- ★ Contextual Knowledge: Entity Disambiguation
- ★ Linked Knowledge: Entity Resolution
- ★ Temporal & Commonsense Knowledge
- ★ Wrap-up

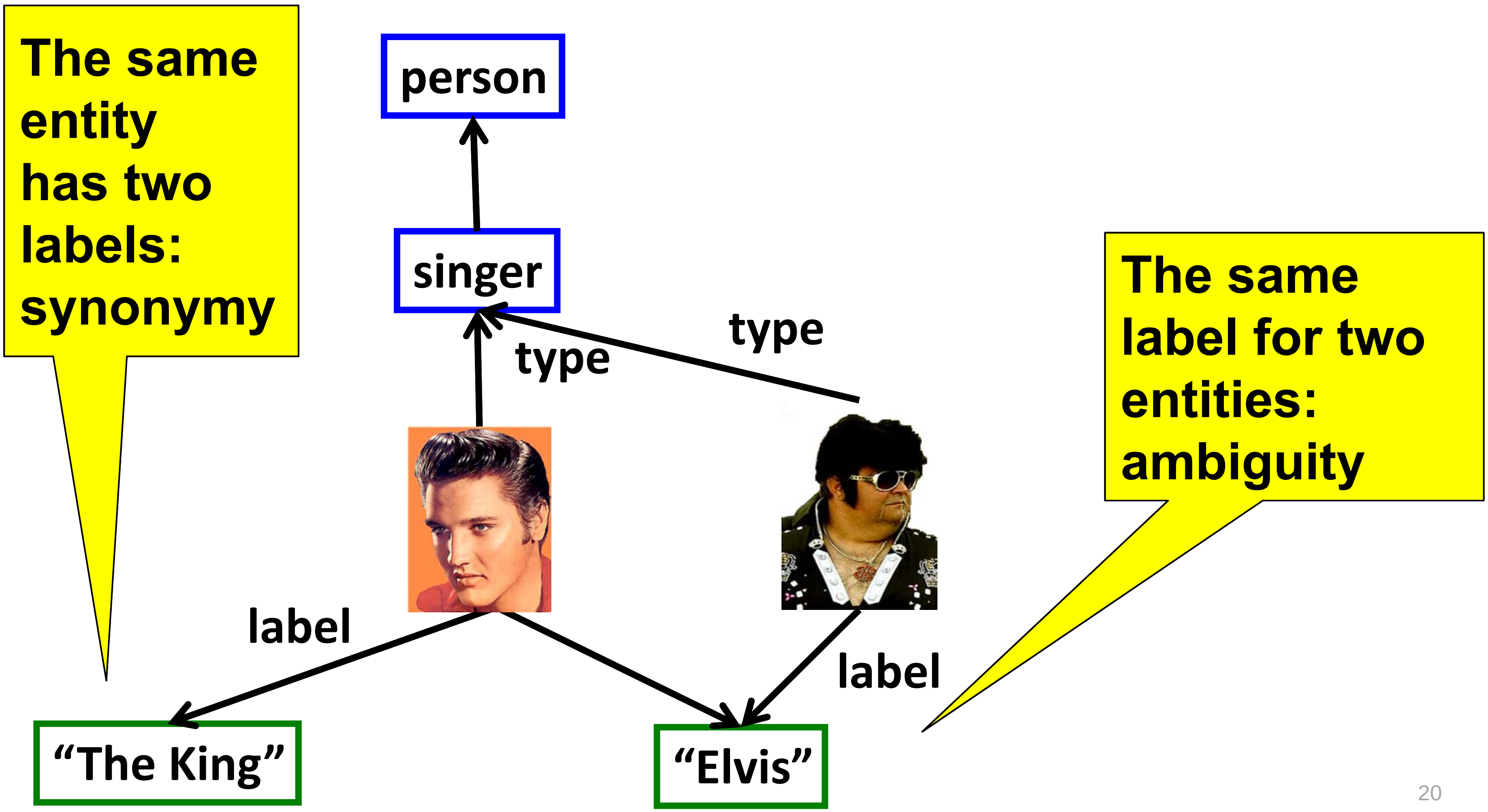
http://www.mpi-inf.mpg.de/yago-naga/icde2013-tutorial/

Knowledge Bases are labeled graphs



A knowledge base can be seen as a directed labeled multi-graph, where the nodes are entities and the edges relations.

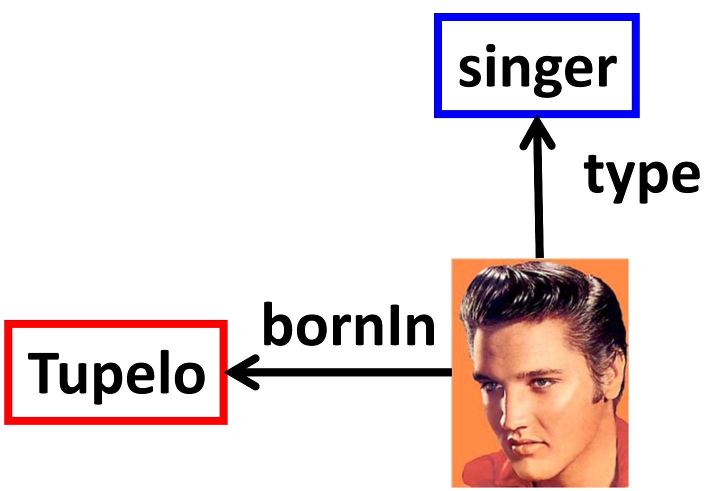
An entity can have different labels



Different views of a knowledge base

We use "RDFS Ontology" and "Knowledge Base (KB)" synonymously.

Graph notation:



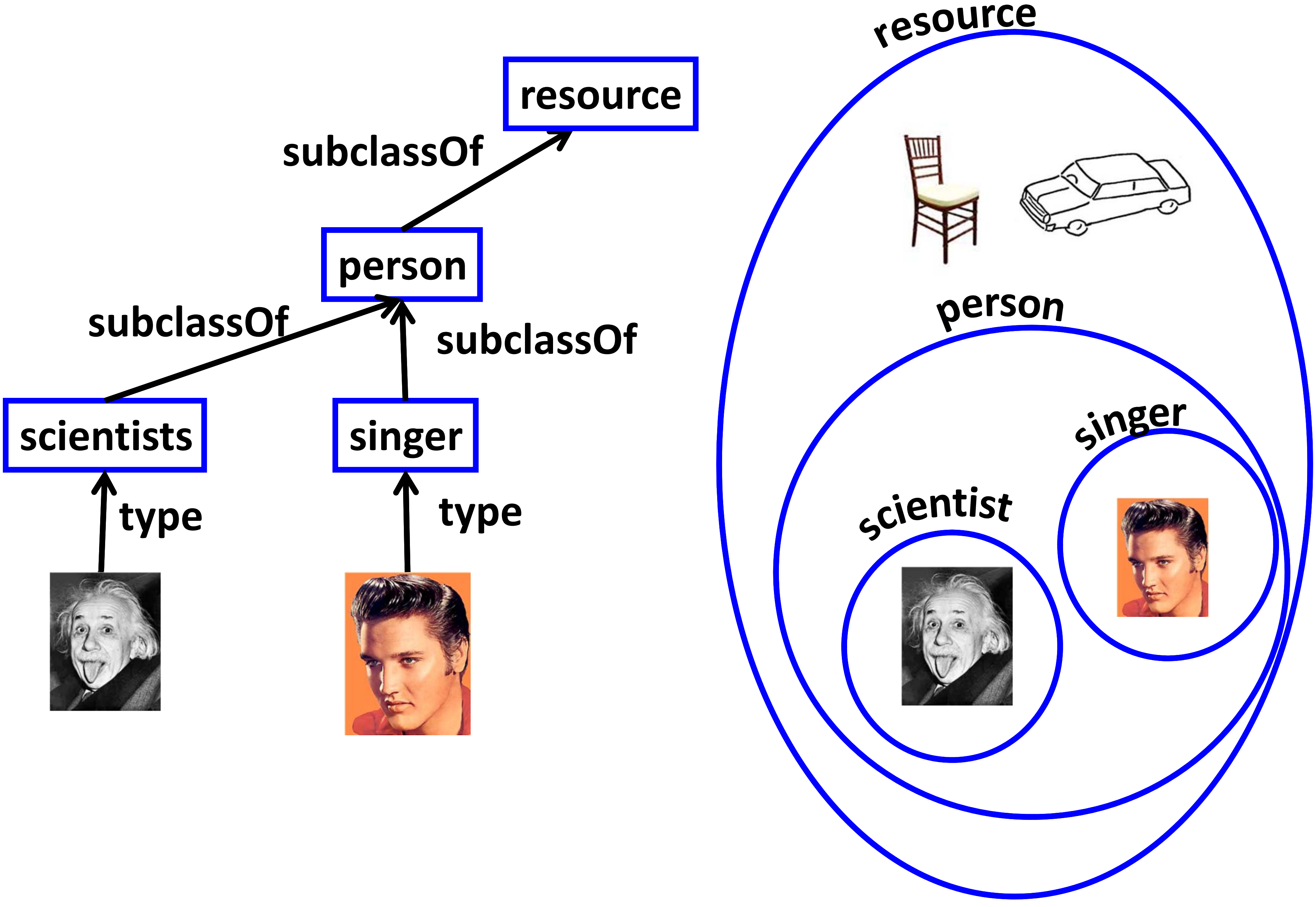
Triple notation:

Subject	Predicate	Object
Elvis	type	singer
Elvis	bornIn	Tupelo
...	...	...

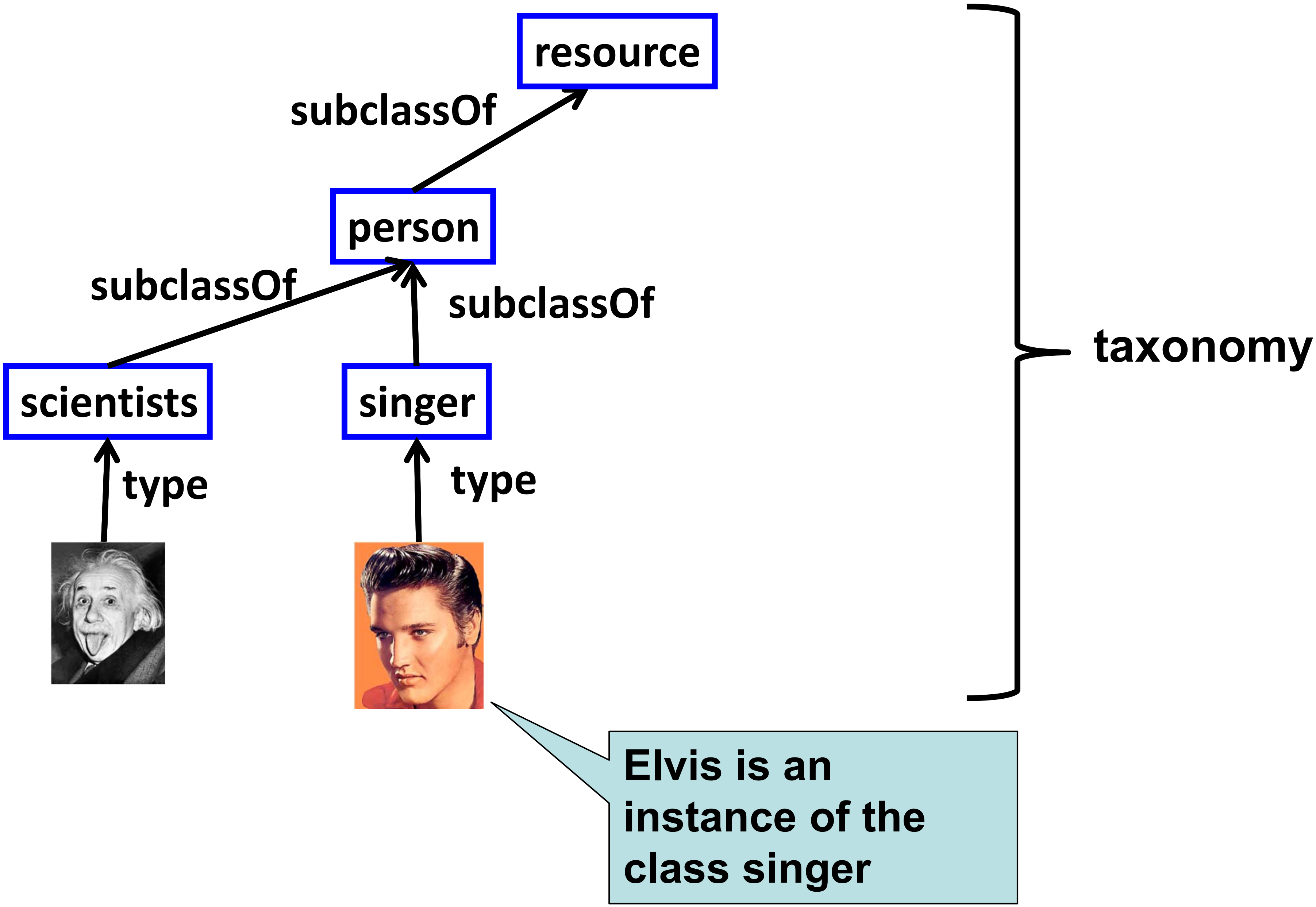
Logical notation:

type(Elvis, singer)
bornIn(Elvis,Tupelo)
...

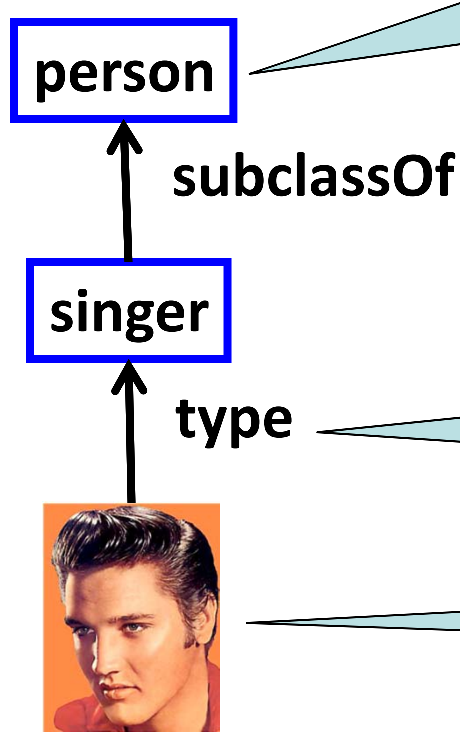
Classes are sets of entities



An instance is a member of a class



Our Goal is finding classes and instances



Which classes exist?
(aka entity types, unary predicates, concepts)

Which subsumptions hold?

Which entities belong to which classes?

Which entities exist?

WordNet is a lexical knowledge base

living being

subclassOf

person

label

subclassOf

"person"

singer

"individual"

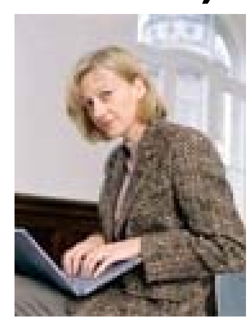
"soul"

WordNet contains 82,000 classes

WordNet contains thousands of subclassOf relationships

WordNet contains 118,000 class labels

WordNet project (1985-now)



WordNet example: superclasses

- S: (n) **singer**, vocalist, vocalizer, vocaliser (a person who sings)
  - *direct hyponym* / *full hyponym*
  - *has instance*
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) musician, instrumentalist, player (someone who plays a musical instrument (as a profession))
      - S: (n) performer, performing artist (an entertainer who performs a dramatic or musical work for an audience)
        - S: (n) entertainer (a person who tries to please or amuse)
          - S: (n) person, individual, someone, somebody, mortal, soul (a human being) *"there was too much for one person to do"*
            - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
              - S: (n) living thing, animate thing (a living (or once living) entity)
                - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
                  - S: (n) object, physical object (a tangible and visible entity; an entity

WordNet example: subclasses

- S: (n) **singer**, vocalist, vocalizer, vocaliser (a person who sings)
  - *direct hyponym* / *full hyponym*
    - S: (n) alto (a singer whose voice lies in the alto clef)
    - S: (n) baritone, barytone (a male singer)
    - S: (n) bass, basso (an adult male singer with the lowest voice)
    - S: (n) canary (a female singer)
    - S: (n) caroler, caroller (a singer of carols)
    - S: (n) castrato (a male singer who was castrated before puberty and retains a soprano or alto voice)
    - S: (n) chorister (a singer in a choir)
    - S: (n) contralto (a woman singer having a contralto voice)
    - S: (n) crooner, balladeer (a singer of popular ballads)
    - S: (n) folk singer, jongleur, minstrel, poet-singer, troubadour (a singer of folk songs)
    - S: (n) hummer (a singer who produces a tune without opening the lips or forming words)
    - S: (n) lieder singer (a singer of lieder)
    - S: (n) madrigalist (a singer of madrigals)
    - S: (n) opera star, operatic star (singer of lead role in an opera)
    - S: (n) rapper (someone who performs rap music)
    - S: (n) rock star (a famous singer of rock music)
    - S: (n) songster (a person who sings)
    - S: (n) soprano (a female singer)

WordNet example: instances

- S: (n) Joplin, Janis Joplin (United States singer who died of a drug overdose at the height of her popularity (1943-1970))
- S: (n) King, B. B. King, Riley B King (United States guitar player and singer of the blues (born in 1925))
- S: (n) Lauder, Harry Lauder, Sir Harry MacLennan Lauder (Scottish ballad singer and music hall comedian (1870-1950))
- S: (n) Ledbetter, Huddie Leadbetter, Leadbelly
  and composer (1885-1949))
- S: (n) Madonna, Madonna Louise Ciccone (Ur
  sex symbol during the 1980s (born in 1958))
- S: (n) Marley, Robert Nesta Marley, Bob Marle
  popularized reggae (1945-1981))
- S: (n) Martin, Dean Martin, Dino Paul Crocetti
  (1917-1995))
- S: (n) Merman, Ethel Merman (United States s
  several musical comedies (1909-1984))
- S: (n) Orbison, Roy Orbison (United States co
  popular in the 1950s (1936-1988))
- S: (n) Piaf, Edith Piaf, Edith Giovanna Gassio
  cabaret singer (1915-1963))
- S: (n) Robeson, Paul Robeson, Paul Bustill Robeson (United States bass singer and an outspoken critic of racism and proponent of socialism (1898-1976))
- S: (n) Russell, Lillian Russell (United States entertainer remembered for her

only 32 singers !?

4 guitarists

5 scientists

0 enterprises

2 entrepreneurs

WordNet classes lack instances

Goal is to go beyond WordNet

WordNet is not perfect:

- **it contains only few instances**
- **it contains only common nouns as classes**
- **it contains only English labels**

... but it contains a wealth of information that can be the starting point for further extraction.

Wikipedia is a rich source of instances

Steve Jobs

From Wikipedia, the free encyclopedia

Jimmy Wales

Larry Sanger

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

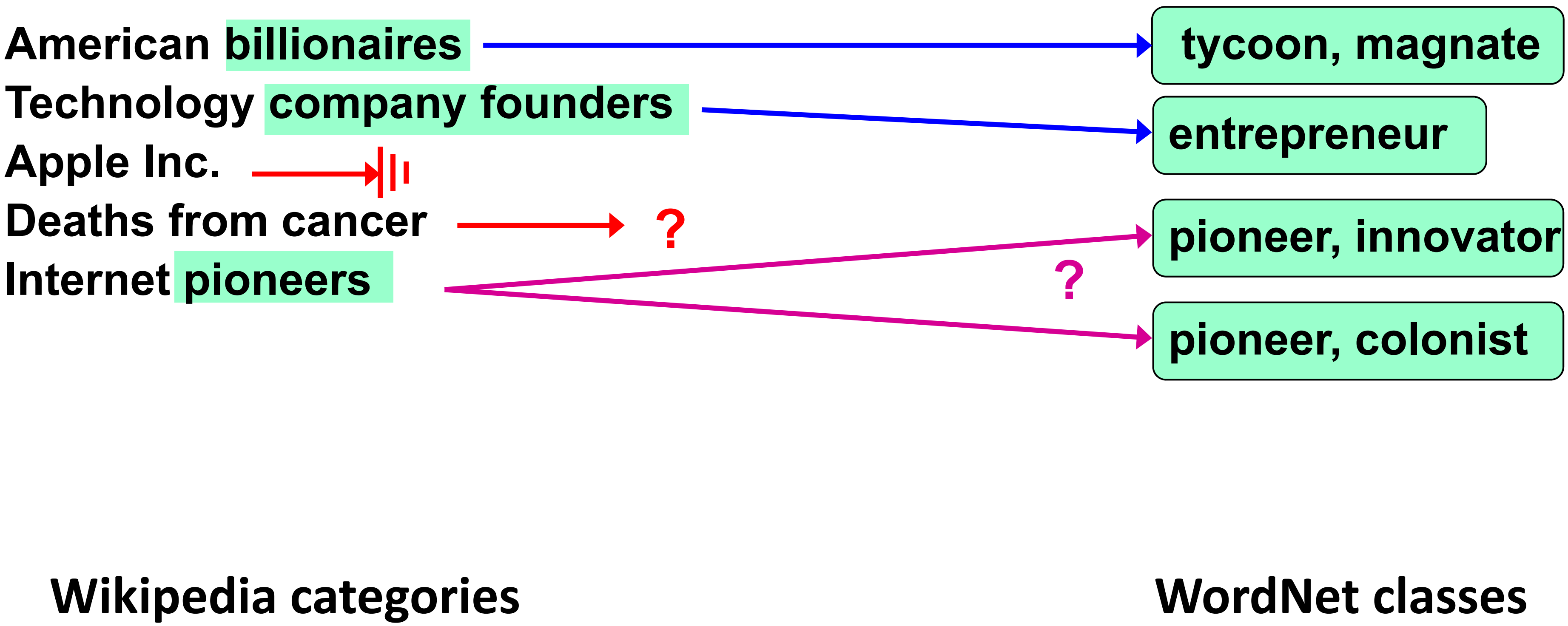
Born	Steven Paul Jobs February 24, 1955[1][2] San Francisco, California, U.S.[1][2]
Died	October 5, 2011 (aged 56)[2] Palo Alto, California, U.S.
Nationality	American
Alma mater	Reed College (dropped out)

Wikipedia's categories contain classes

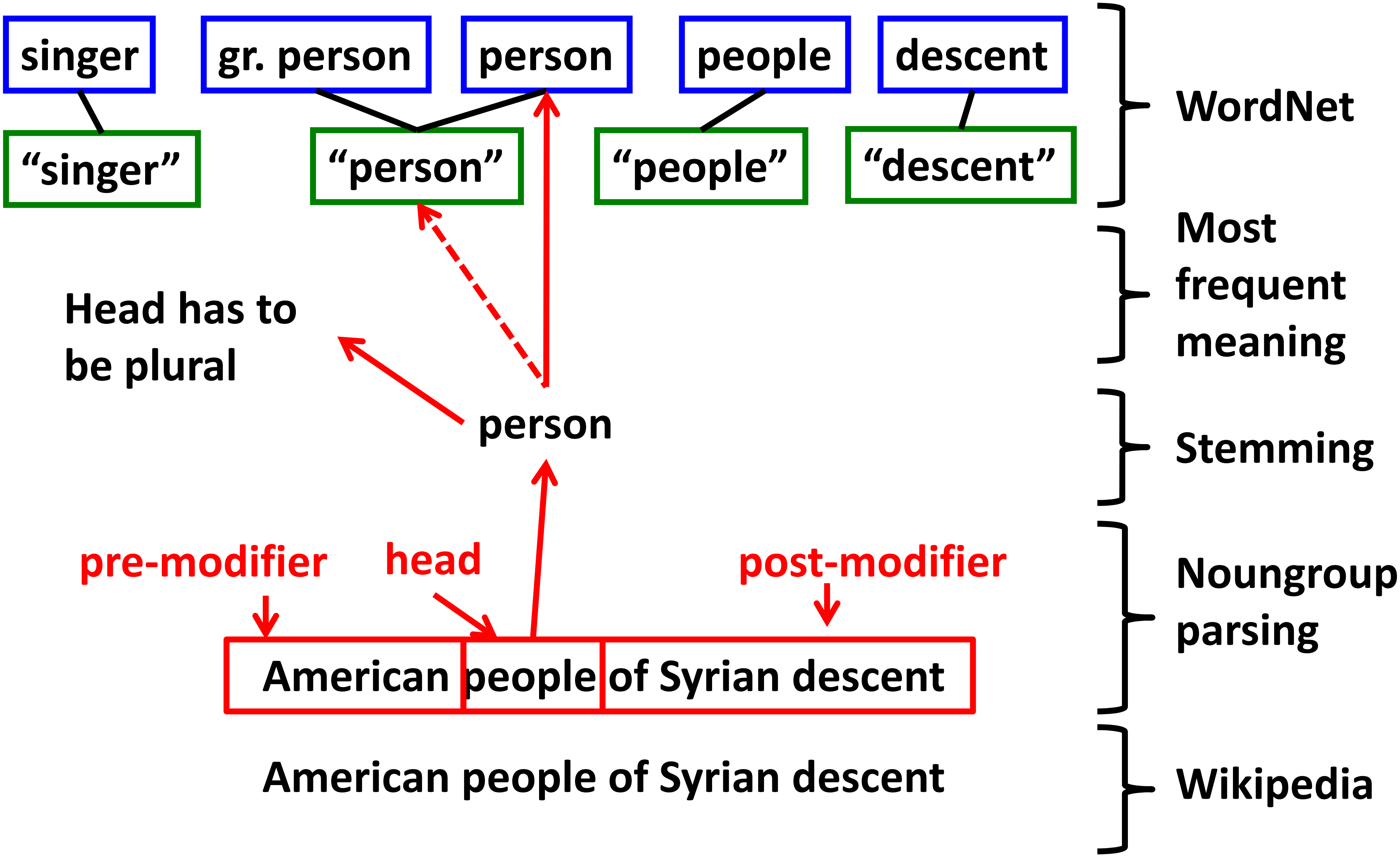
Categories: Steve Jobs | 1955 births | 2011 deaths | American adoptees | American billionaires | American chief executives | American computer businesspeople | American industrial designers | American inventors | American people of German descent | American people of Swiss descent | American people of 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

Link Wikipedia categories to WordNet?



Categories can be linked to WordNet

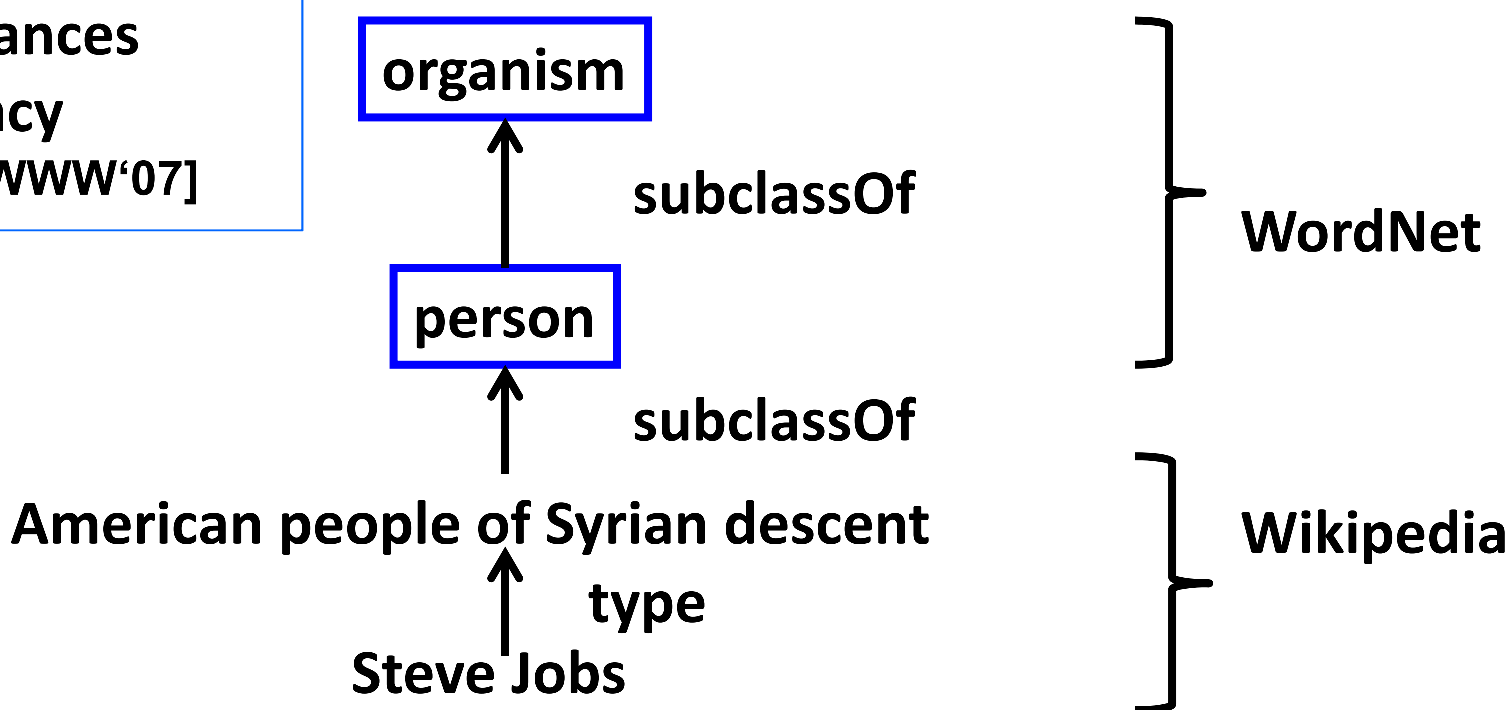


YAGO = WordNet+Wikipedia

200,000 classes
460,000 subclassOf
3 Mio. instances
96% accuracy
[Suchanek: WWW'07]

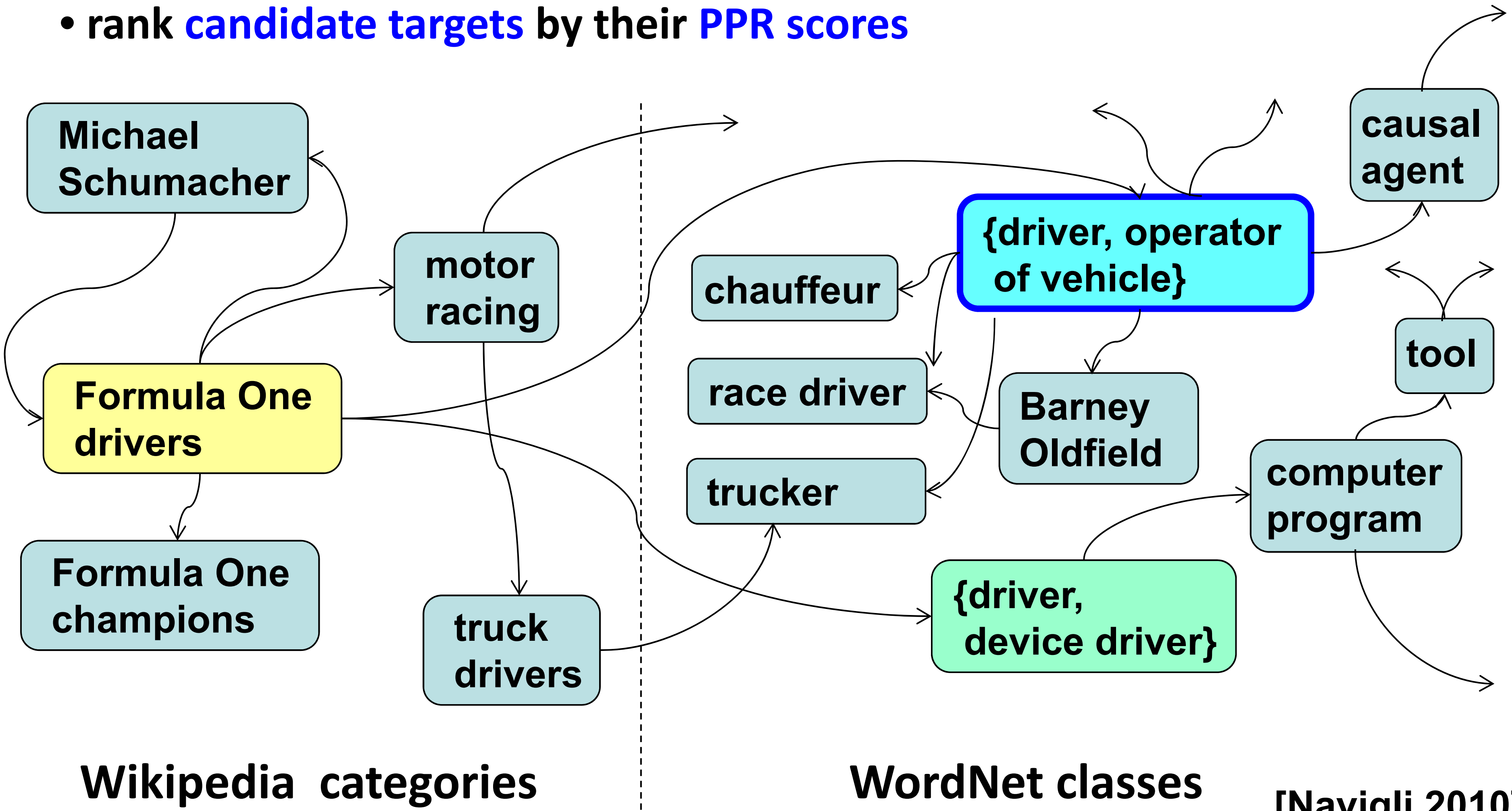
Related project:
WikiTaxonomy

105,000 subclassOf links
88% accuracy
[Ponzetto & Strube: AAAI'07]



Link Wikipedia & WordNet by Random Walks

- construct neighborhood around source and target nodes
- use contextual similarity (glosses etc.) as edge weights
- compute personalized PR (PPR) with source as start node
- rank candidate targets by their PPR scores



[Navigli 2010]

Categories yield more than classes

[Nastase/Strube 2012]

Examples for "rich" categories:

- Chancellors of Germany
- Capitals of Europe
- Deaths from Cancer
- People Emigrated to America
- Bob Dylan Albums

Generate candidates from pattern templates:

$e \in$ NP1 IN NP2	$\longrightarrow$	e type NP1, e spatialRel NP2
$e \in$ NP1 VB NP2		e type NP1, e VB NP2
$e \in$ NP1 NP2		e createdBy NP1

Validate and infer relation names via infoboxes:

check for infobox attribute with value NP2 for e
for all/most articles in category c

http://www.h-its.org/english/research/nlp/download/wikinet.php

Which Wikipedia articles are classes?

European_Union	instance
Eurovision_Song_Contest	instance
Central_European_Countries	class
Rocky_Mountains	instance
European_history	?
Culture_of_Europe	?

Heuristics:

1) Head word singular → entity
2) Head word or entire phrase mostly capitalized in corpus → entity
3) Head word plural → class
4) otherwise → general concept (neither class nor individual entity)

[Bunescu/Pasca 2006, Nastase/Strube 2012]

Alternative features:

- time-series of phrase freq. etc.

[Lin: EMNLP 2012]

Hearst patterns extract instances from text

[M. Hearst 1992]

Goal: find instances of classes

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: //better with POS tagging

companies such as Apple
Google, Microsoft and other companies
Internet companies like Amazon and Facebook
Chinese cities including Kunming and Shangri-La
computer pioneers like the late Steve Jobs
computer pioneers and other scientists
lakes in the vicinity of Brisbane

Derive type(Y,X)

type(Apple, company), type(Google, company), ...

Recursively applied patterns increase recall

[Kozareva/Hovy 2010]

use results from Hearst patterns as seeds
then use „parallel-instances“ patterns

X such as Y → companies such as Apple
companies such as Google

Y like Z
*, Y and Z → Apple like Microsoft offers
IBM, Google, and Amazon

Y like Z
*, Y and Z → Microsoft like SAP sells
eBay, Amazon, and Facebook

Y like Z
*, Y and Z → Cherry, Apple, and Banana

potential problems with ambiguous words

Doubly-anchored patterns are more robust

[Kozareva/Hovy 2010, Dalvi et al. 2012]

Goal:

find instances of classes

Start with a set of seeds:

companies = {Microsoft, Google}

Parse Web documents and find the pattern

W, Y and Z

If two of three placeholders match seeds, harvest the third:

Google, Microsoft and Amazon → type(Amazon, company)

Cherry, Apple, and Banana → X

Instances can be extracted from tables

[Kozareva/Hovy 2010, Dalvi et al. 2012]

Goal: find instances of classes

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)

Extracting instances from lists & tables

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

State-of-the-Art Approach (e.g. SEAL):
- Start with seeds: a few class instances
- Find lists, tables, text snippets ("for example: …"), … that contain one or more seeds
- Extract candidates: noun phrases from vicinity
- Gather co-occurrence stats (seed&cand, cand&className pairs)
- Rank candidates
  - point-wise mutual information, …
  - random walk (PR-style) on seed-cand graph

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

Probase builds a taxonomy from the Web

Use Hearst liberally to obtain many instance candidates:

„plants such as trees and grass“

„plants include water turbines“

„western movies such as The Good, the Bad, and the Ugly“

Problem: signal vs. noise

Assess candidate pairs statistically:

$P[X|Y] >> P[X^*|Y] \quad \rightarrow \quad \text{subclassOf}(Y\ X)$

Problem: ambiguity of labels

Merge labels of same class:

X such as Y_1 and Y_2 $\rightarrow$ same sense of X

ProBase

2.7 Mio. classes from 1.7 Bio. Web pages

[Wu et al.: SIGMOD 2012]

Use query logs to refine taxonomy

[Pasca 2011]

Input:

type(Y, X_1), type(Y, X_2), type(Y, X_3), e.g, extracted from Web

Goal: rank candidate classes X_1, X_2, X_3

Combine the following scores to rank candidate classes:

H1: X and Y should co-occur frequently in queries

$\rightarrow$ score1(X) $\sim$ freq(X,Y) * #distinctPatterns(X,Y)

H2: If Y is ambiguous, then users will query X Y:

$\rightarrow$ score2(X) $\sim$ $(\prod_{i=1..N}$ term-score$(t_i \in X))^{1/N}$

example query: "Michael Jordan computer scientist"

H3: If Y is ambiguous, then users will query first X, then X Y:

$\rightarrow$ score3(X) $\sim$ $(\prod_{i=1..N}$ term-session-score$(t_i \in X))^{1/N}$

Take-Home Lessons

Semantic classes for entities

> 10 Mio. entities in 100,000's of classes
backbone for other kinds of knowledge harvesting
great mileage for semantic search
e.g. politicians who are scientists,
French professors who founded Internet companies, …

Variety of methods

noun phrase analysis, random walks, extraction from tables, …

Still room for improvement

higher coverage, deeper in long tail, …

Open Problems and Grand Challenges

Wikipedia categories reloaded: larger coverage

comprehensive & consistent instanceOf and subClassOf across Wikipedia and WordNet

e.g. people lost at sea, ACM Fellow,
Jewish physicists emigrating from Germany to USA, …

Long tail of entities

beyond Wikipedia: domain-specific entity catalogs

e.g. music, books, book characters, electronic products, restaurants, …

New name for known entity vs. new entity?

e.g. Lady Gaga vs. Radio Gaga vs. Stefani Joanne Angelina Germanotta

Universal solution for taxonomy alignment

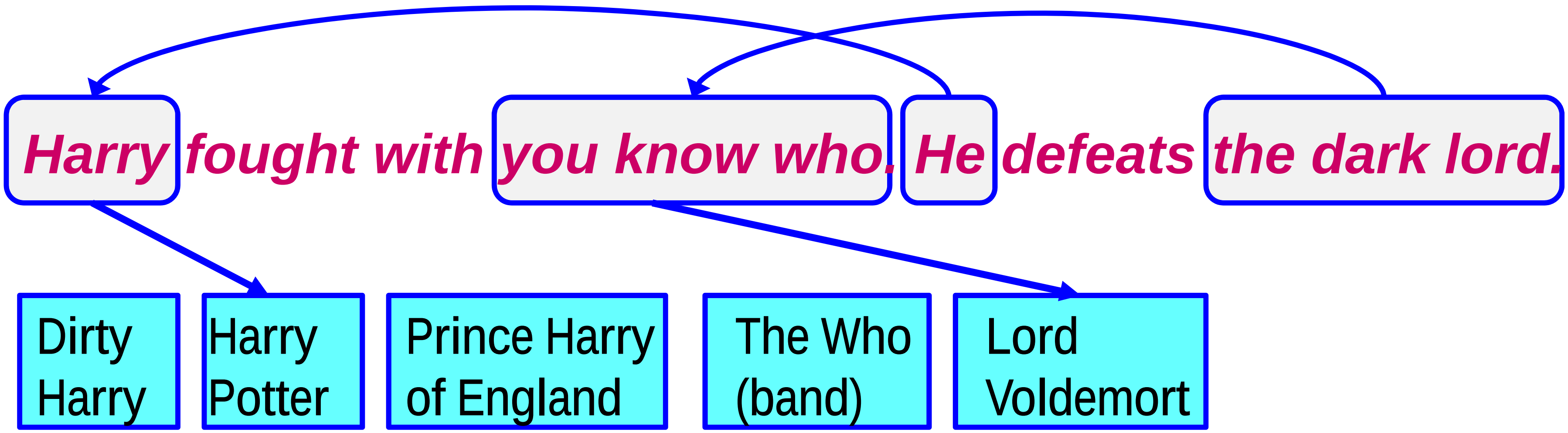
e.g. Wikipedia's, dmoz.org, baike.baidu.com, amazon, librarything tags, …

Outline

- ✓ Motivation
- ★ Machine Knowledge
- ★ Taxonomic Knowledge: Entities and Classes
- ★ Contextual Knowledge: Entity Disambiguation
- ★ Linked Knowledge: Entity Resolution
- ★ Temporal & Commonsense Knowledge
- ★ Wrap-up

http://www.mpi-inf.mpg.de/yago-naga/icde2013-tutorial/

Three Different Problems



Three NLP tasks:

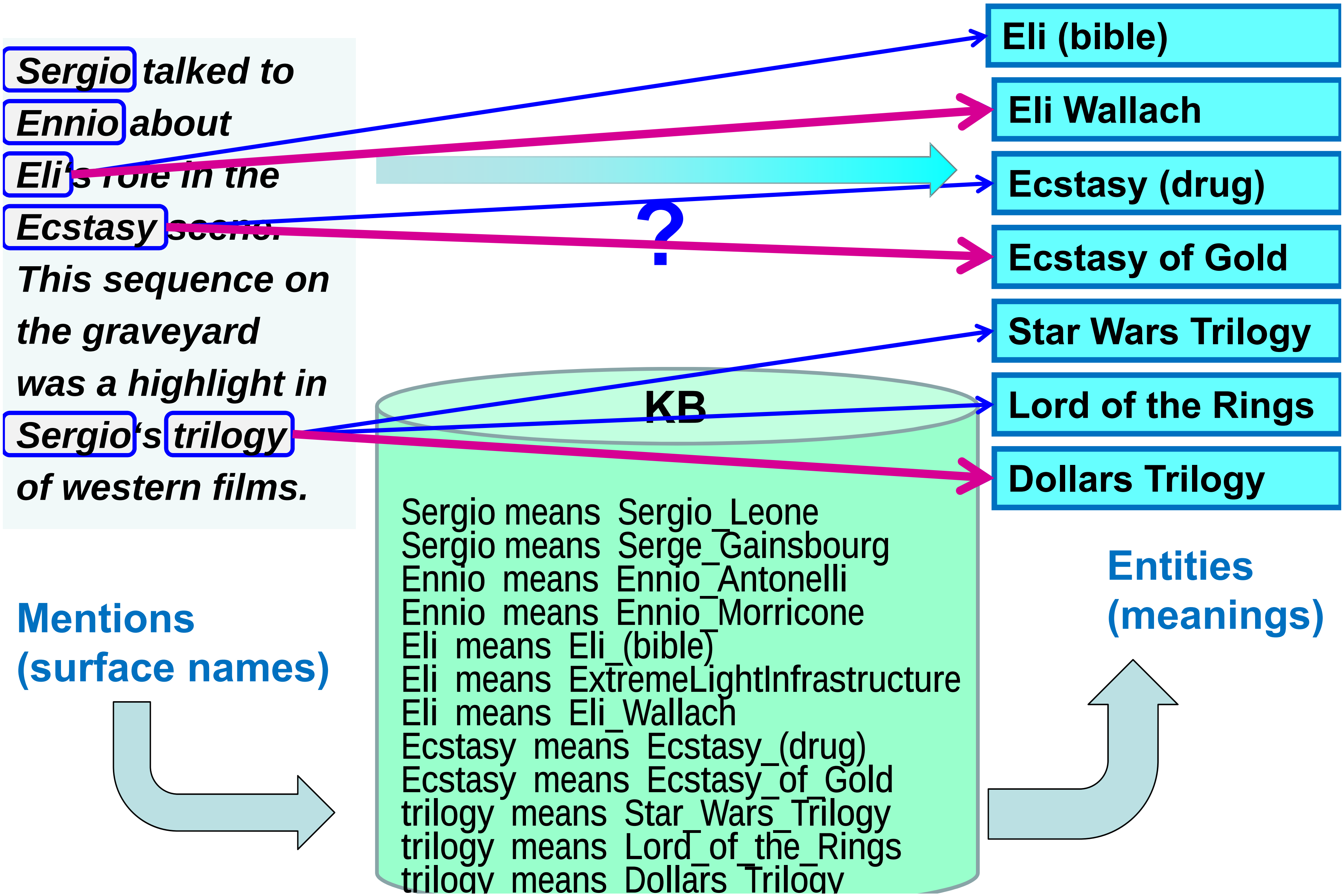
1) named-entity recognition (NER): segment & label by CRF (e.g. Stanford NER tagger)

2) co-reference resolution: link to preceding NP (trained classifier over linguistic features)

3) named-entity disambiguation (NED): map each mention (name) to canonical entity (entry in KB)

tasks 1 and 3 together: NERD

Named Entity Disambiguation



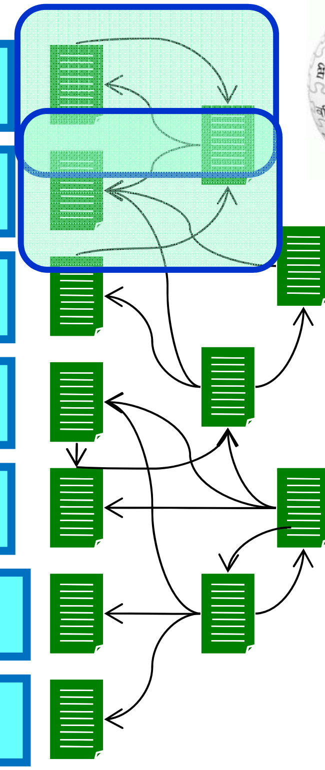
Mention-Entity Graph

weighted undirected graph with two types of nodes

Sergio talked to Ennio about Eli's role in the Ecstasy scene. This sequence on the graveyard was a highlight in Sergio's trilogy of western films.

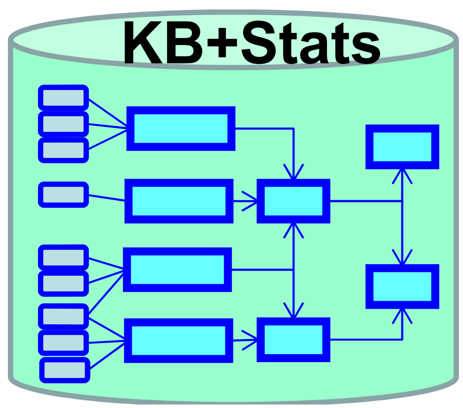
bag-of-words or language model: words, bigrams, phrases

- Eli (bible)
- Eli Wallach
- Ecstasy (drug)
- Ecstasy of Gold
- Star Wars
- Lord of the Rings
- Dollars Trilogy



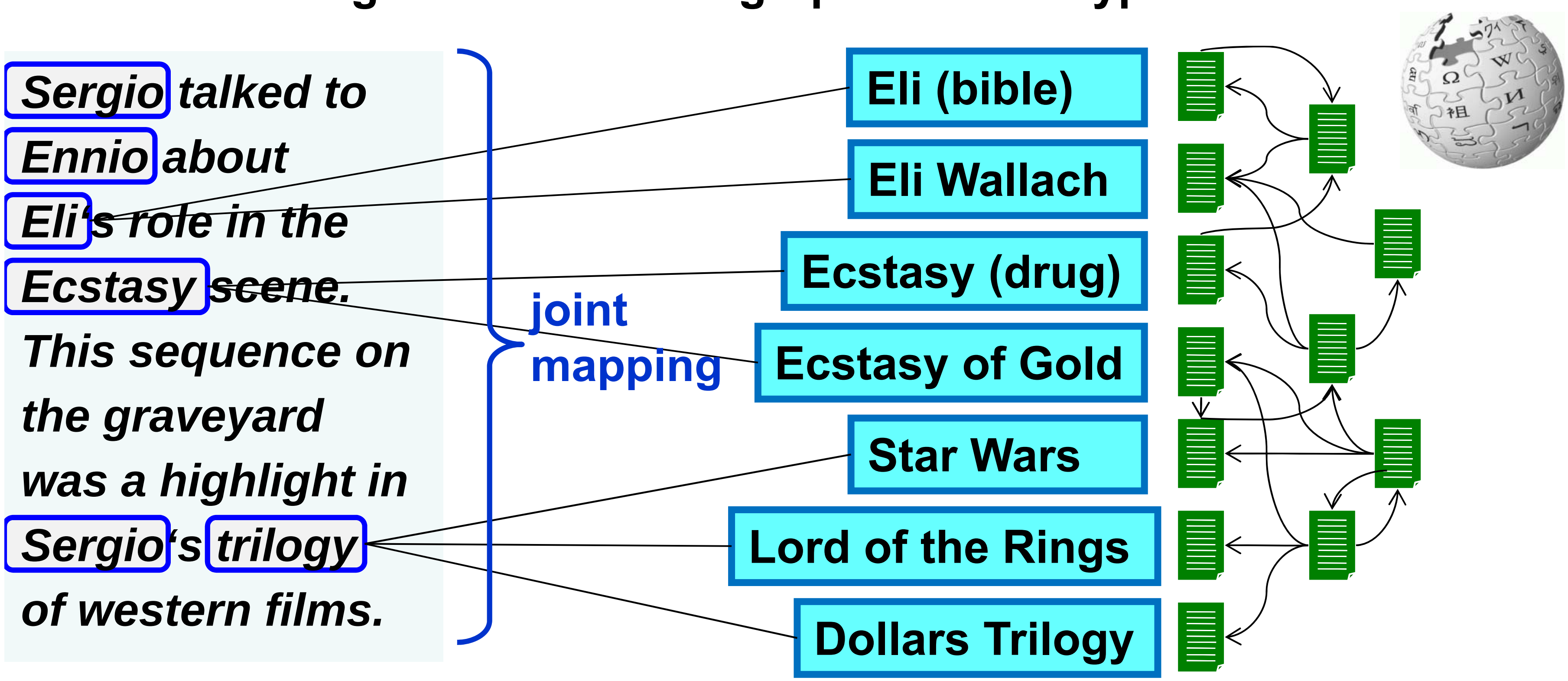
Popularity (m,e):
- freq(e|m)
- length(e)
- #links(e)

Similarity (m,e):
- cos/Dice/KL (context(m), context(e))



Mention-Entity Graph

weighted undirected graph with two types of nodes

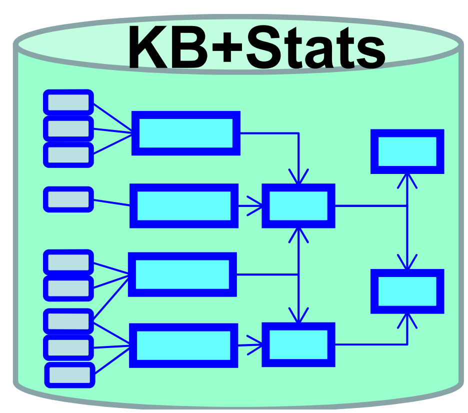


Popularity (m,e):

- freq(e|m)
- length(e)
- #links(e)

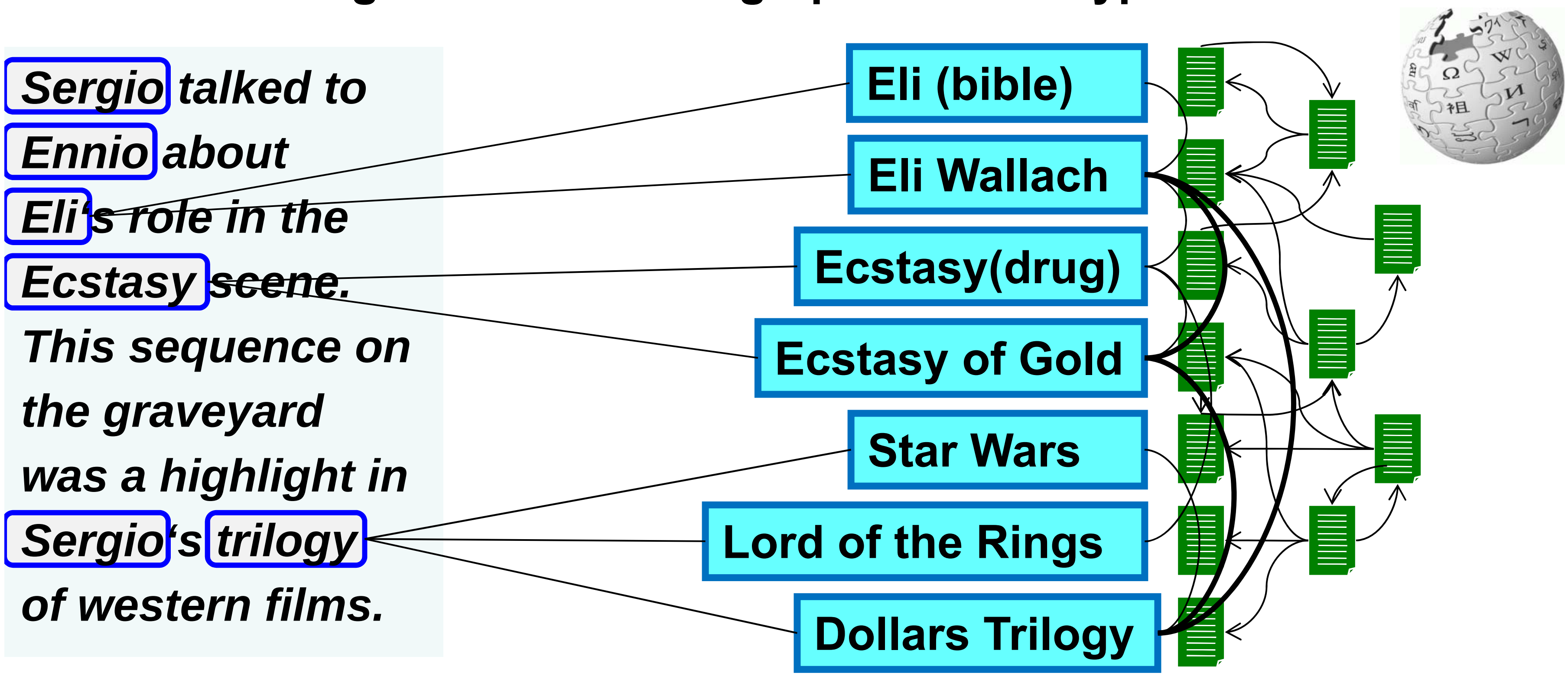
Similarity (m,e):

- cos/Dice/KL (context(m), context(e))



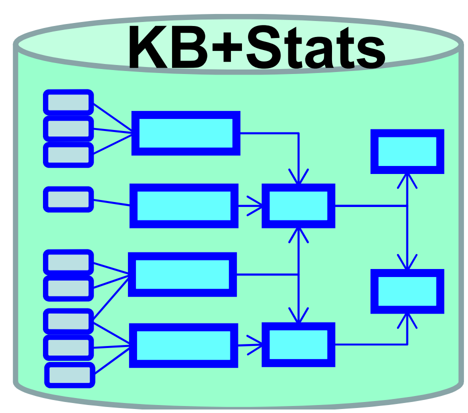
Mention-Entity Graph

weighted undirected graph with two types of nodes



Popularity (m,e):
- freq(m,e|m)
- length(e)
- #links(e)

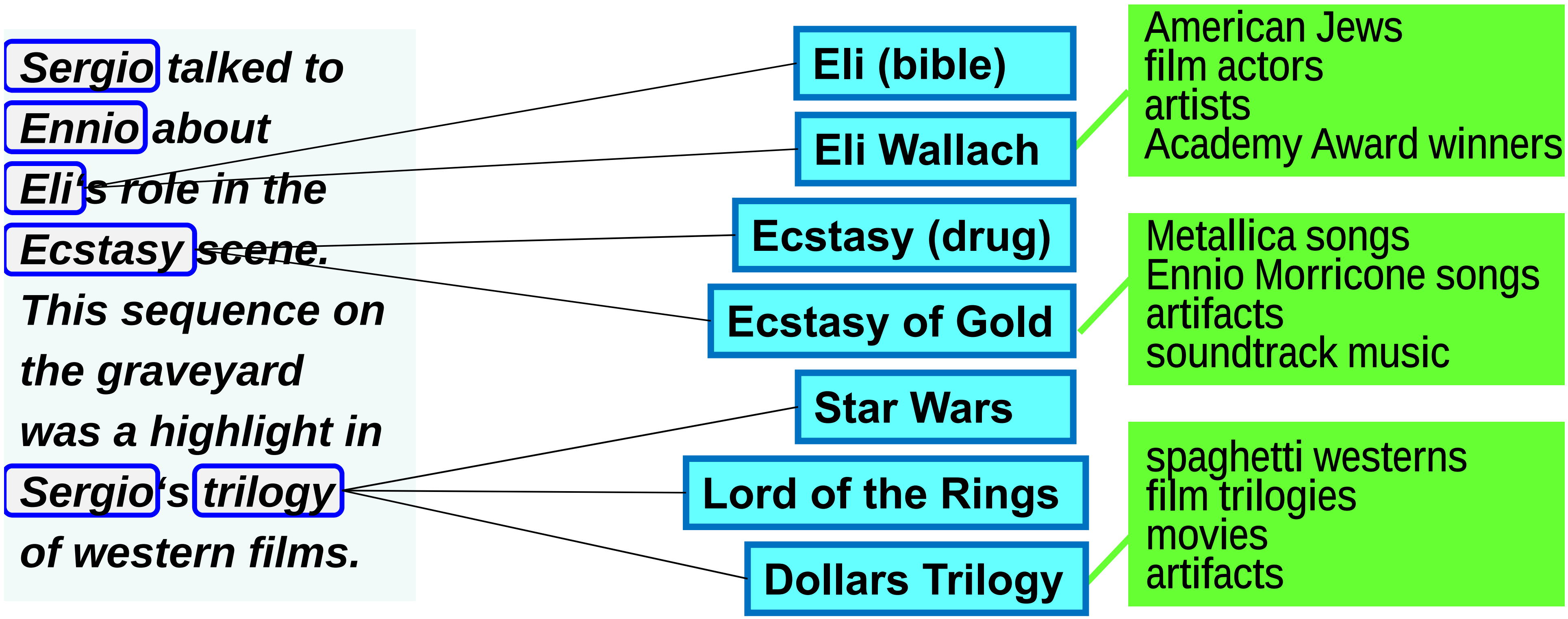
Similarity (m,e):
- cos/Dice/KL (context(m), context(e))



Coherence (e,e'):
- dist(types)
- overlap(links)
- overlap (anchor words)

Mention-Entity Graph

weighted undirected graph with two types of nodes

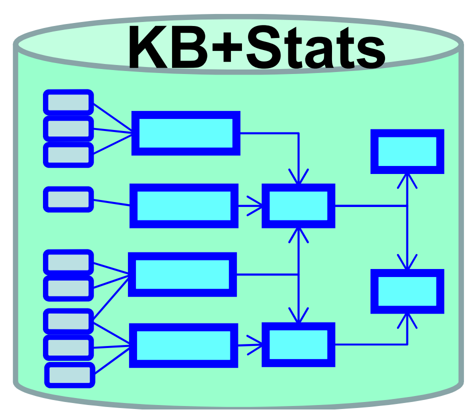


Popularity (m,e):

- freq(m,e|m)
- length(e)
- #links(e)

Similarity (m,e):

- cos/Dice/KL (context(m), context(e))

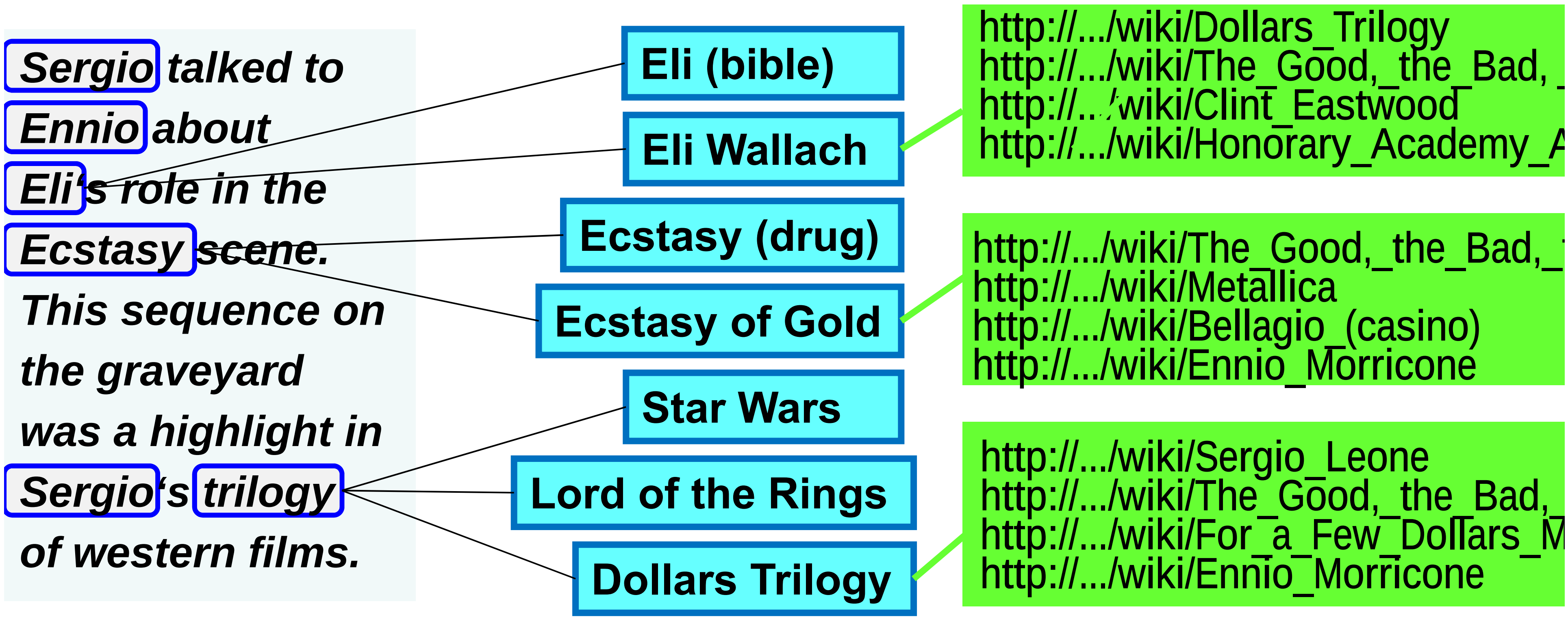


Coherence (e,e'):

- dist(types)
- overlap(links)
- overlap (anchor words)

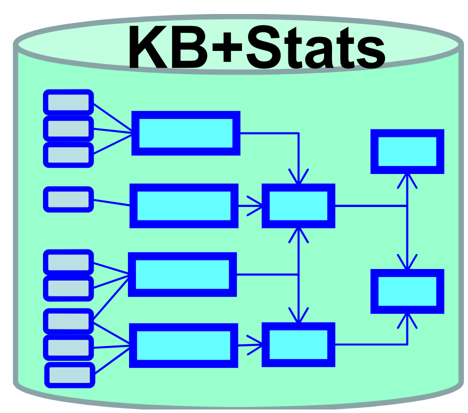
Mention-Entity Graph

weighted undirected graph with two types of nodes



Popularity (m,e):
- freq(m,e|m)
- length(e)
- #links(e)

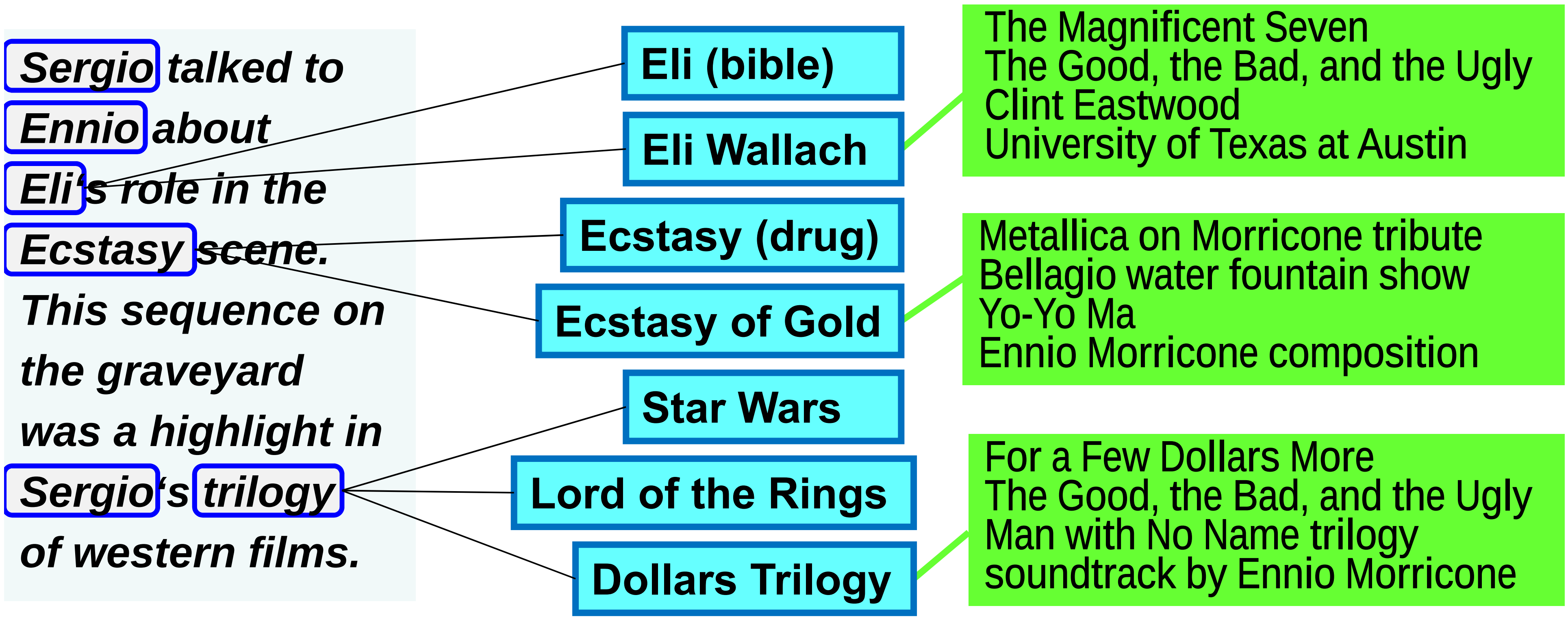
Similarity (m,e):
- cos/Dice/KL (context(m), context(e))



Coherence (e,e'):
- dist(types)
- overlap(links)
- overlap (anchor words)

Mention-Entity Graph

weighted undirected graph with two types of nodes

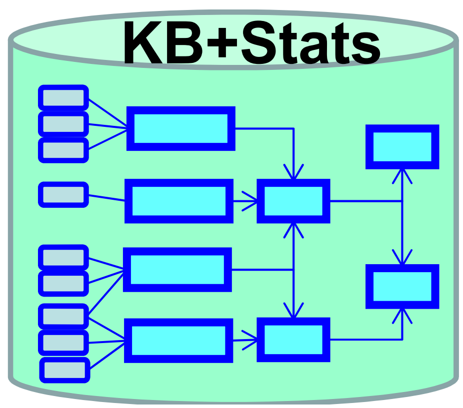


Popularity (m,e):

- freq(m,e|m)
- length(e)
- #links(e)

Similarity (m,e):

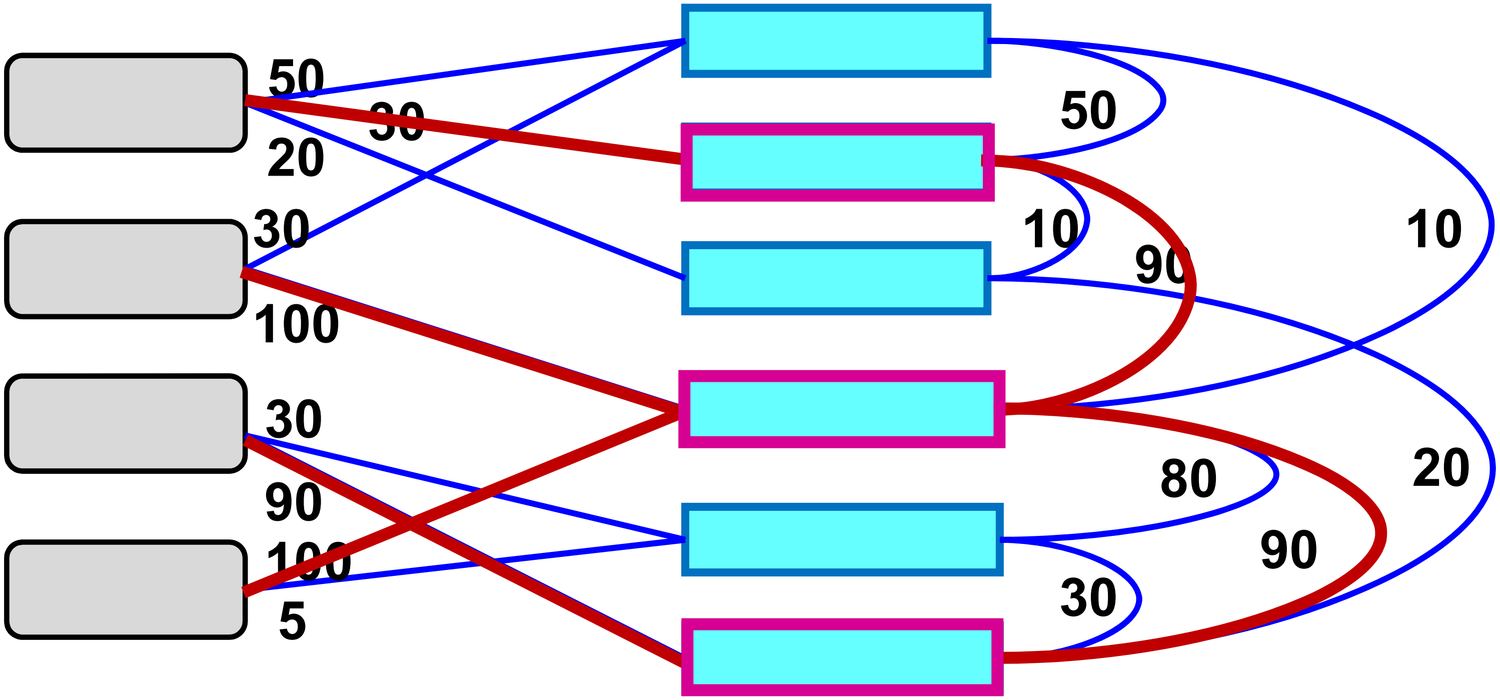
- cos/Dice/KL (context(m), context(e))



Coherence (e,e‘):

- dist(types)
- overlap(links)
- overlap (anchor words)

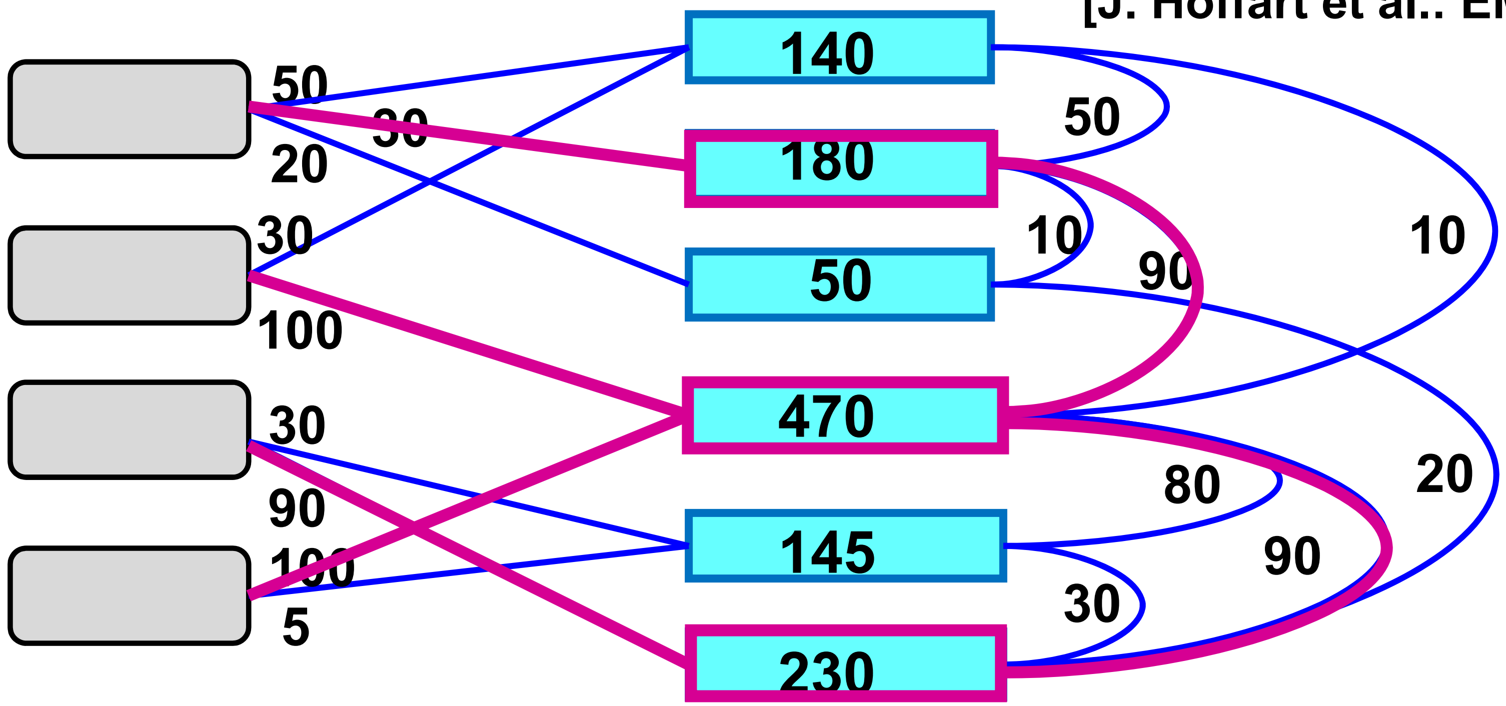
Joint Mapping



- Build mention-entity graph or joint-inference factor graph from knowledge and statistics in KB
- Compute high-likelihood mapping (ML or MAP) or dense subgraph such that: each m is connected to exactly one e (or at most one e)

Coherence Graph Algorithm

[J. Hoffart et al.: EMNLP'11]



- Compute dense subgraph to
  maximize min weighted degree among entity nodes
  such that:
  each m is connected to exactly one e (or at most one e)
- Greedy approximation:
  iteratively remove weakest entity and its edges
- Keep alternative solutions, then use local/randomized search

Mention-Entity Popularity Weights

[Milne/Witten 2008, Spitkovsky/Chang 2012]

- Need dictionary with entities' names:
  - full names: Arnold Alois Schwarzenegger, Los Angeles, Microsoft Corp.
  - short names: Arnold, Arnie, Mr. Schwarzenegger, New York, Microsoft, …
  - nicknames & aliases: Terminator, City of Angels, Evil Empire, …
  - acronyms: LA, UCLA, MS, MSFT
  - role names: the Austrian action hero, Californian governor, CEO of MS, …

  …

  plus gender info (useful for resolving pronouns in context):

  Bill and Melinda met at MS. They fell in love and <u>he</u> kissed <u>her</u>.

- Collect hyperlink anchor-text / link-target pairs from
  - Wikipedia redirects
  - Wikipedia links between articles and Interwiki links
  - Web links pointing to Wikipedia articles
  - query-and-click logs

  …
- Build statistics to estimate P[entity | name]

Mention-Entity Similarity Edges

Precompute characteristic keyphrases q for each entity e:
anchor texts or noun phrases in e page with high PMI:

$$weight(q,e) = \log \frac{freq(q,e)}{freq(q)\, freq(e)}$$

„Metallica tribute to Ennio Morricone“

Match keyphrase q of candidate e in context of mention m

$$score(q\,|\,e) \sim \frac{\#\,matching\ \ words}{length\ of\ \ cover(q)} \left(\frac{\sum_{w \in cover(q)} weight\,(w\,|\,e)}{\sum_{w \in q} weight(w\ \,|\,e)} \right)^{1+\gamma}$$

Extent of partial matches **Weight of matched words**

The Ecstasy piece was covered by Metallica on the Morricone tribute album.

Compute overall similarity of context(m) and candidate e

$$score\,(e\,|\,m) \sim \sum_{\substack{q \in keyphrases\ (e) \\ in\ \ context\ (m)}} score\,(q)\, dist\,(cover(q)\,,m)^{-\alpha}$$

Entity-Entity Coherence Edges

Precompute overlap of incoming links for entities e1 and e2

$$mw\text{-}coh(e1,e2) \sim 1 - \frac{\log \max(in(e1,e2)) - \log(in(e1) \cap in(e2))}{\log |E| - \log \min(in(e1), in(e2))}$$

Alternatively compute overlap of anchor texts for e1 and e2

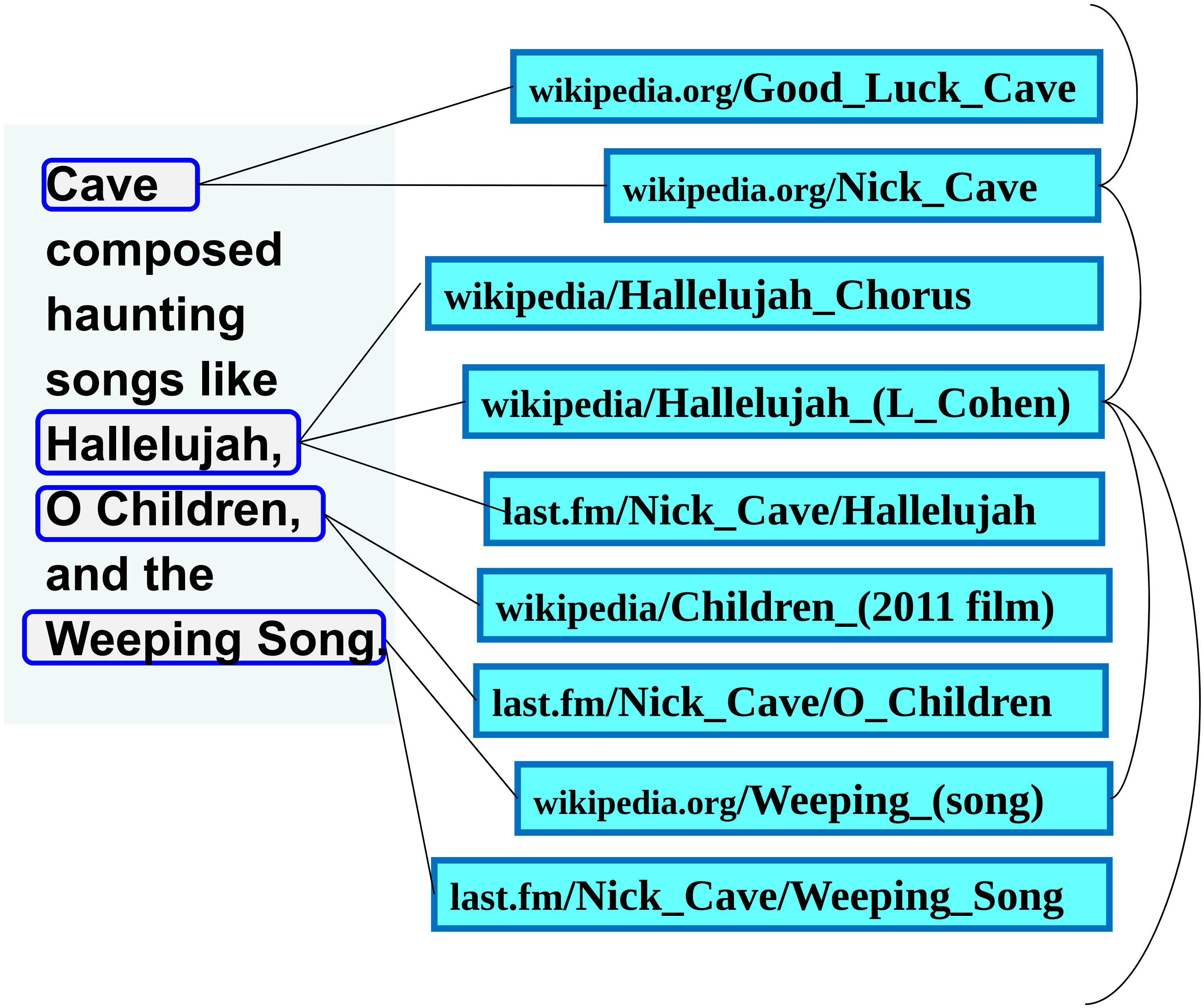
$$ngram\text{-}coh(e1,e2) \sim \frac{|ngrams(e1) \cap ngrams(e2)|}{|ngrams(e1) \cup ngrams(e2)|}$$

or overlap of keyphrases, or similarity of bag-of-words, or …

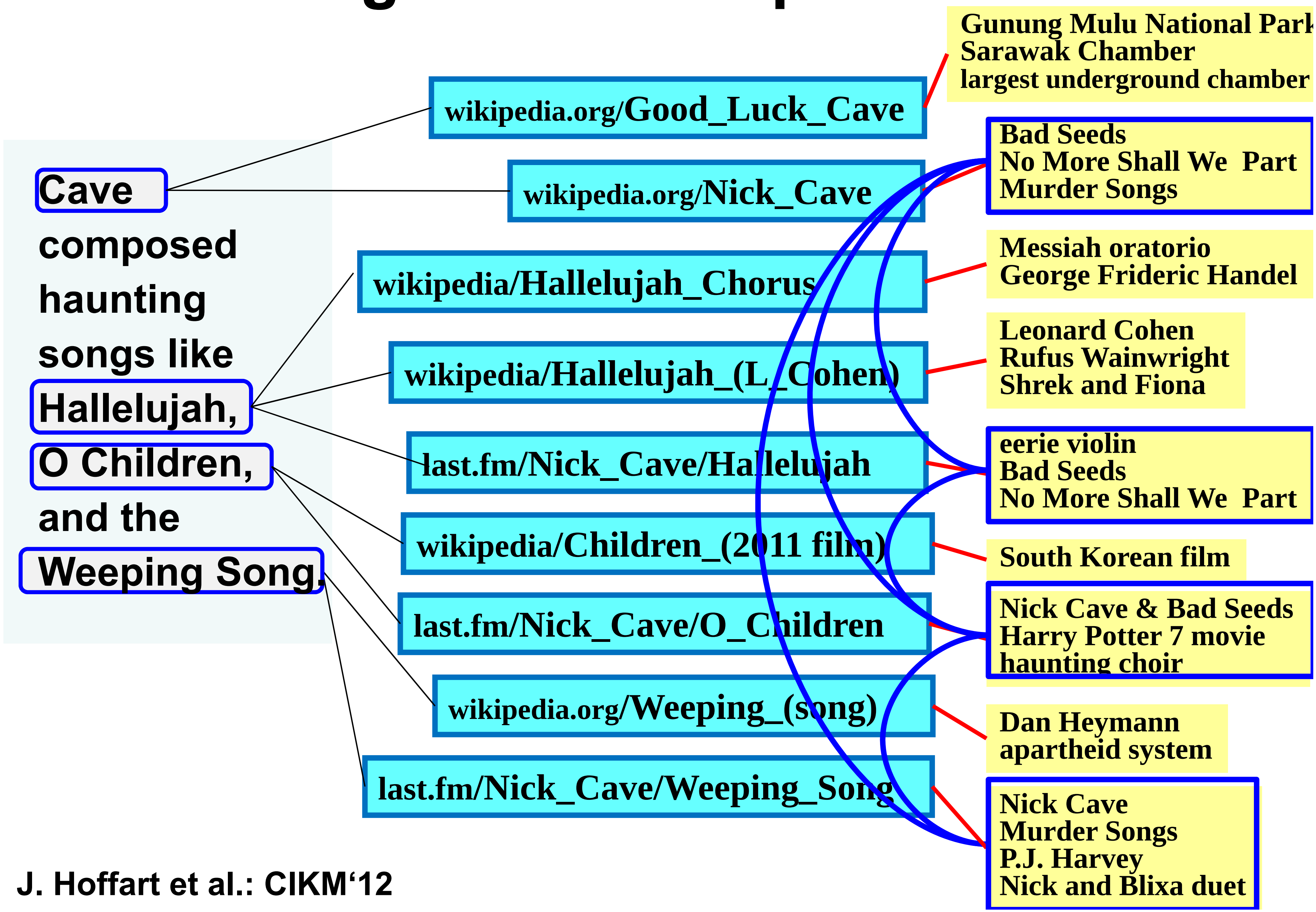
Optionally combine with type distance of e1 and e2
(e.g., Jaccard index for type instances)

For special types of e1 and e2 (locations, people, etc.)
use spatial or temporal distance

Handling Out-of-Wikipedia Entities

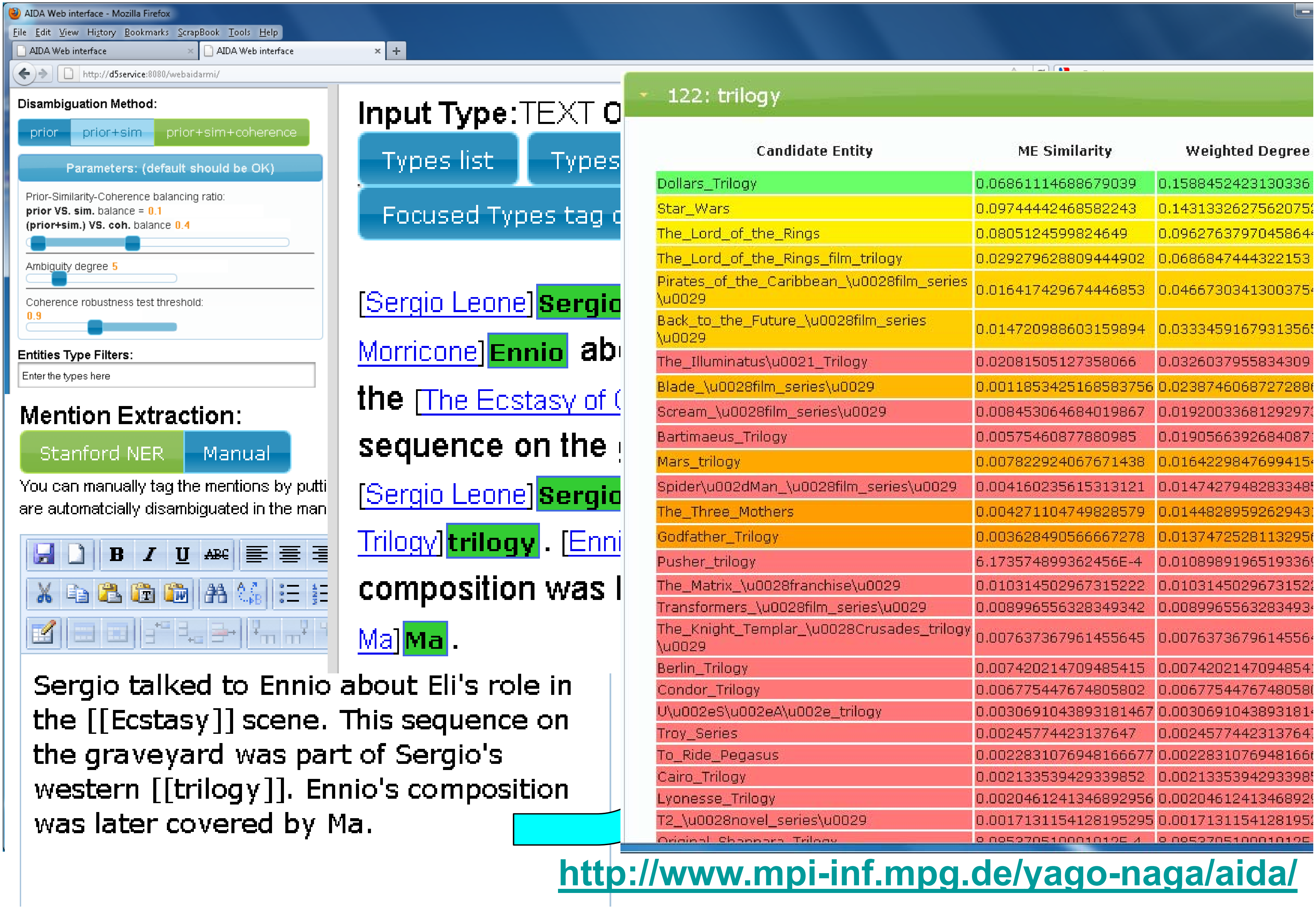


Handling Out-of-Wikipedia Entities



J. Hoffart et al.: CIKM'12

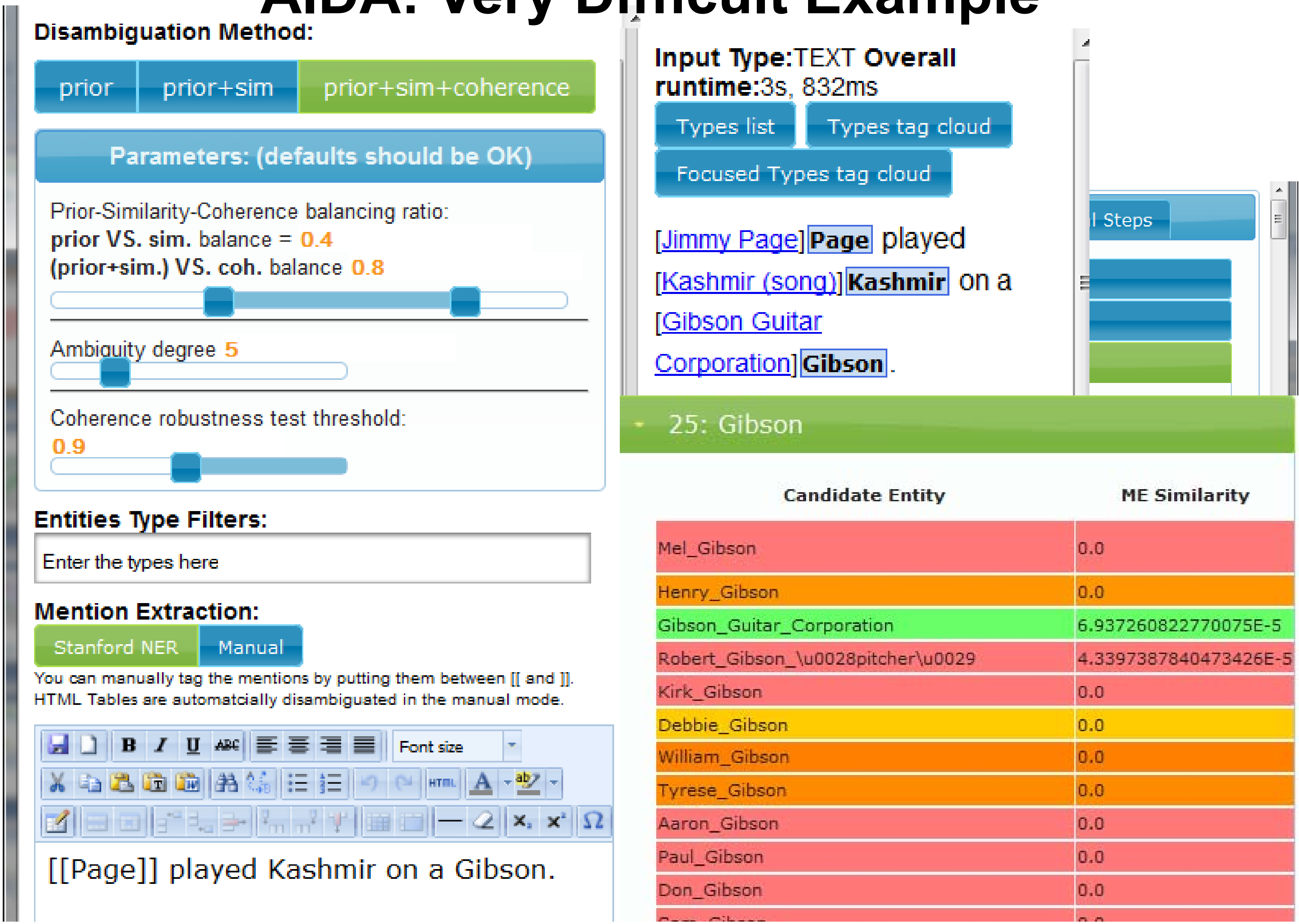
AIDA: Accurate Online Disambiguation



Candidate Entity	ME Similarity	Weighted Degree
Dollars_Trilogy	0.06861114688679039	0.1588452423130336
Star_Wars	0.09744442468582243	0.14313326275620752
The_Lord_of_the_Rings	0.0805124599824649	0.0962763797045864
The_Lord_of_the_Rings_film_trilogy	0.029279628809444902	0.0686847444322153
Pirates_of_the_Caribbean_\u0028film_series \u0029	0.016417429674446853	0.0466730341300375
Back_to_the_Future_\u0028film_series \u0029	0.014720988603159894	0.0333459167931356
The_Illuminatus\u0021_Trilogy	0.02081505127358066	0.0326037955834309
Blade_\u0028film_series\u0029	0.0011853425168583756	0.0238746068727288
Scream_\u0028film_series\u0029	0.008453064684019867	0.0192003368129297
Bartimaeus_Trilogy	0.00575460877880985	0.0190566392684087
Mars_trilogy	0.007822924067671438	0.0164229847699415
Spider\u002dMan_\u0028film_series\u0029	0.004160235615313121	0.0147427948283348
The_Three_Mothers	0.004271104749828579	0.0144828959262943
Godfather_Trilogy	0.003628490566667278	0.0137472528113295
Pusher_trilogy	6.173574899362456E-4	0.0108989196519336
The_Matrix_\u0028franchise\u0029	0.010314502967315222	0.0103145029673152
Transformers_\u0028film_series\u0029	0.008996556328349342	0.0089965563283493
The_Knight_Templar_\u0028Crusades_trilogy \u0029	0.007637367961455645	0.0076373679614556
Berlin_Trilogy	0.007420214709485415	0.0074202147094854
Condor_Trilogy	0.006775447674805802	0.0067754476748058
U\u002eS\u002eA\u002e_trilogy	0.0030691043893181467	0.0030691043893181
Troy_Series	0.00245774423137647	0.0024577442313764
To_Ride_Pegasus	0.0022831076948166677	0.0022831076948166
Cairo_Trilogy	0.002133539429339852	0.0021335394293398
Lyonesse_Trilogy	0.0020461241346892956	0.0020461241346892
T2_\u0028novel_series\u0029	0.0017131154128195295	0.0017131154128195

http://www.mpi-inf.mpg.de/yago-naga/aida/

AIDA: Very Difficult Example



NED: Experimental Evaluation

Benchmark:

- Extended CoNLL 2003 dataset: 1400 newswire articles
- originally annotated with mention markup (NER), now with NED mappings to Yago and Freebase
- difficult texts:

… Australia beats India … → Australian_Cricket_Team

… White House talks to Kreml … → President_of_the_USA

… EDS made a contract with … → HP_Enterprise_Services

Results:

Best: AIDA method with prior+sim+coh + robustness test
82% precision @100% recall, 87% mean average precision
Comparison to other methods, see [Hoffart et al.: EMNLP'11]

see also [P. Ferragina et al.: WWW'13] for NERD benchmarks

NERD Online Tools

J. Hoffart et al.: EMNLP 2011, VLDB 2011
https://d5gate.ag5.mpi-sb.mpg.de/webaida/

P. Ferragina, U. Scaella: CIKM 2010
http://tagme.di.unipi.it/

R. Isele, C. Bizer: VLDB 2012
http://spotlight.dbpedia.org/demo/index.html

Reuters Open Calais: http://viewer.opencalais.com/

Alchemy API: http://www.alchemyapi.com/api/demo.html

S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti: KDD 2009
http://www.cse.iitb.ac.in/soumen/doc/CSAW/

D. Milne, I. Witten: CIKM 2008
http://wikipedia-miner.cms.waikato.ac.nz/demos/annotate/

L. Ratinov, D. Roth, D. Downey, M. Anderson: ACL 2011
http://cogcomp.cs.illinois.edu/page/demo_view/Wikifier

some use Stanford NER tagger for detecting mentions
http://nlp.stanford.edu/software/CRF-NER.shtml

Take-Home Lessons

NERD is key for contextual knowledge

High-quality NERD uses joint inference over various features:
popularity + similarity + coherence

State-of-the-art tools available

Maturing now, but still room for improvement,
especially on efficiency, scalability & robustness

Handling out-of-KB entities & long-tail NERD

Still a difficult research issue

Open Problems and Grand Challenges

Entity name disambiguation in difficult situations

Short and noisy texts about long-tail entities in social media

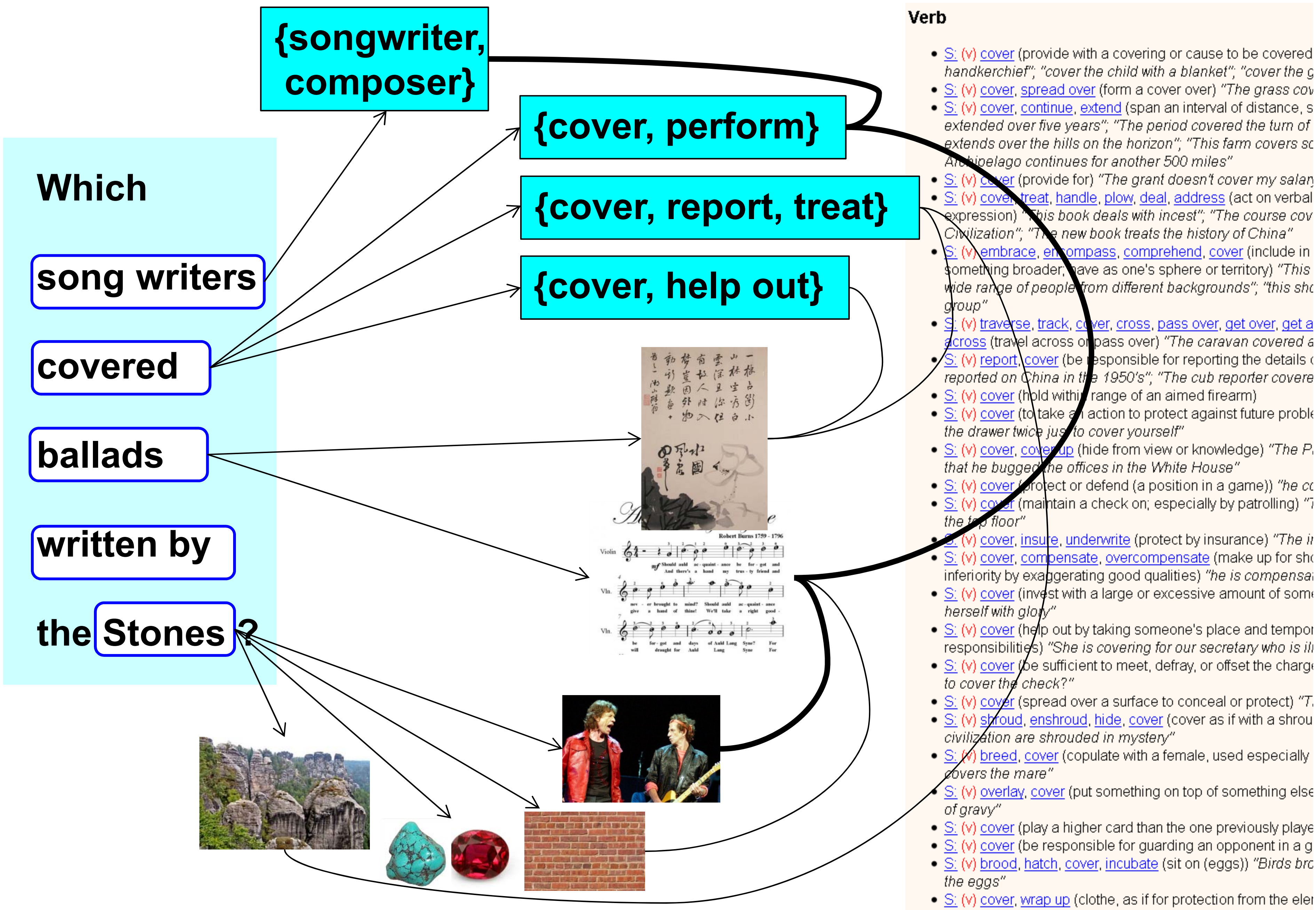
Robust disambiguation of entities, relations and classes

Relevant for question answering & question-to-query translation
Key building block for KB building and maintenance

Word sense disambiguation in natural-language dialogs

Relevant for multimodal human-computer interactions
(speech, gestures, immersive environments)

General Word Sense Disambiguation



Outline

✓ Motivation

★ Machine Knowledge

★ Taxonomic Knowledge: Entities and Classes

★ Contextual Knowledge: Entity Disambiguation

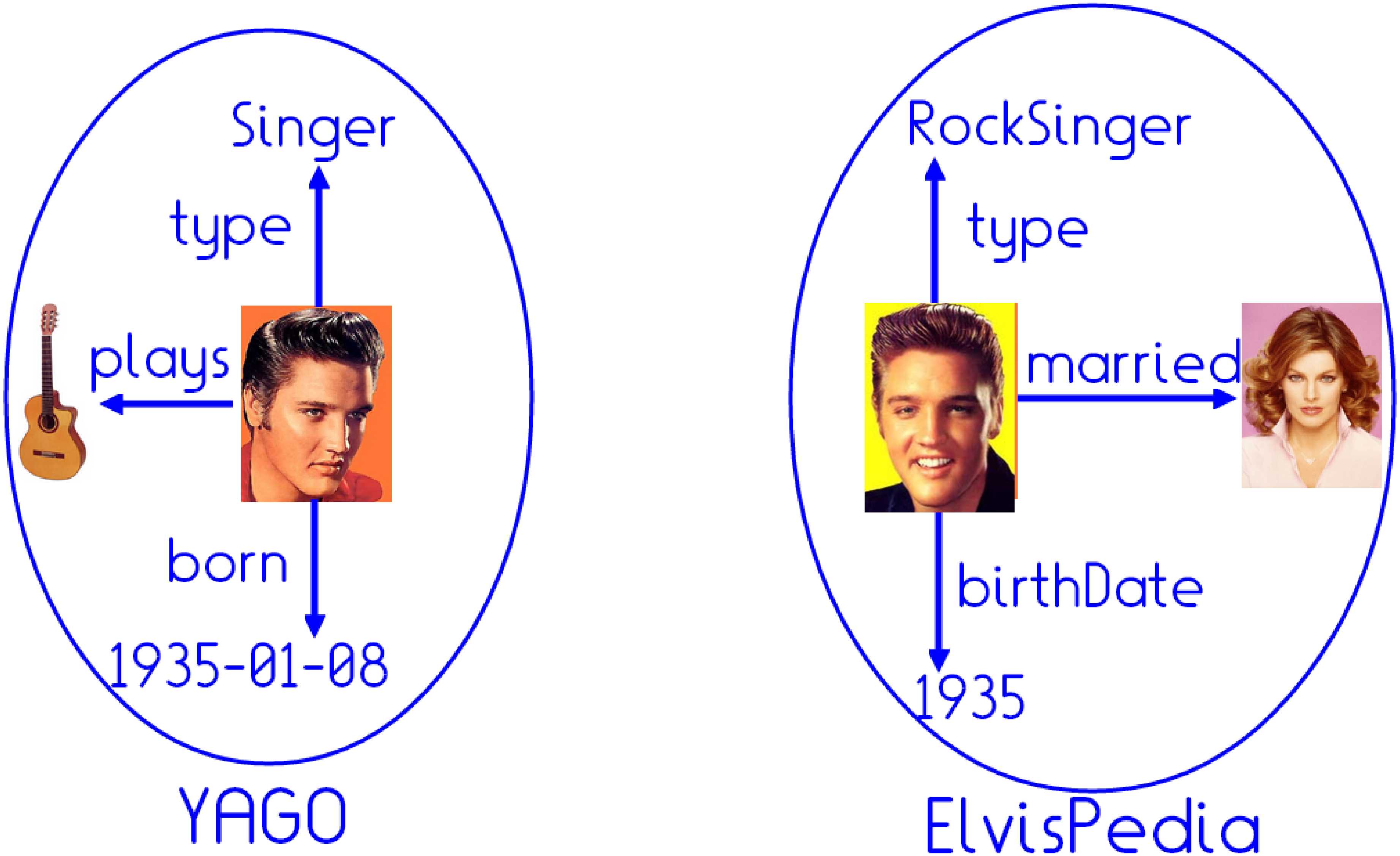
★ Linked Knowledge: Entity Resolution

★ Temporal & Commonsense Knowledge

★ Wrap-up

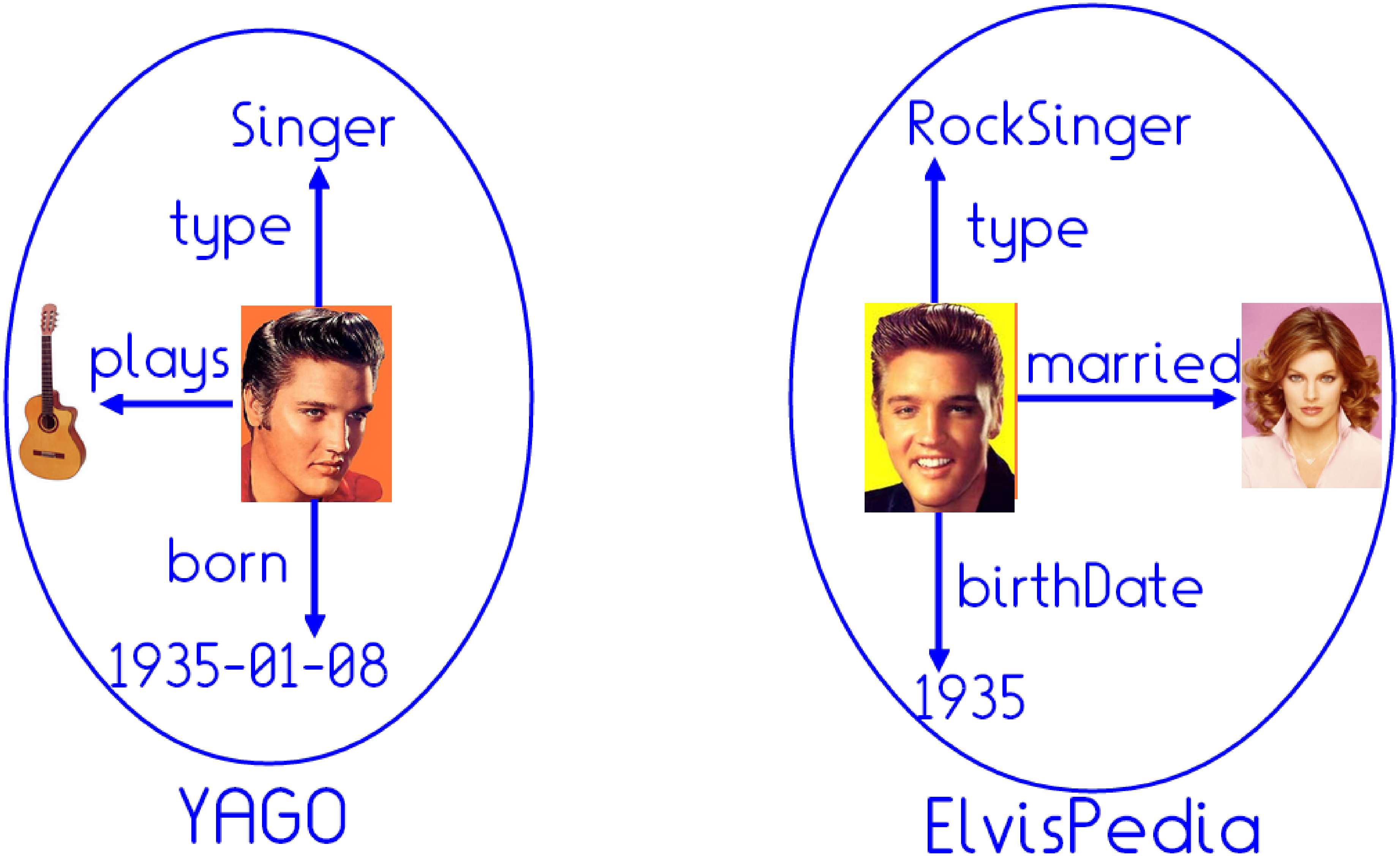
http://www.mpi-inf.mpg.de/yago-naga/icde2013-tutorial/

Knowledge bases are complementary

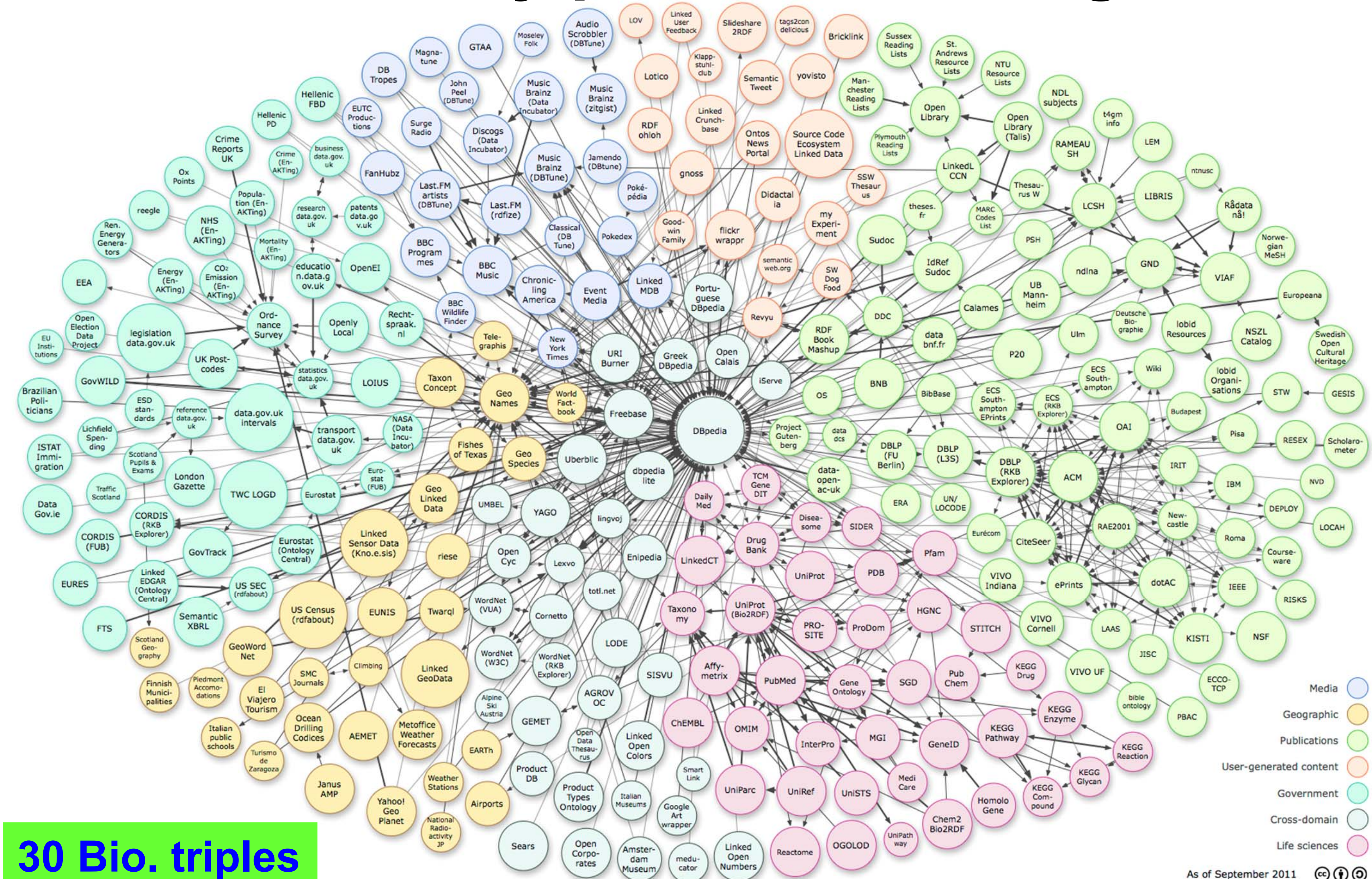


No Links ⇒ No Use

Who is the spouse of the guitar player?



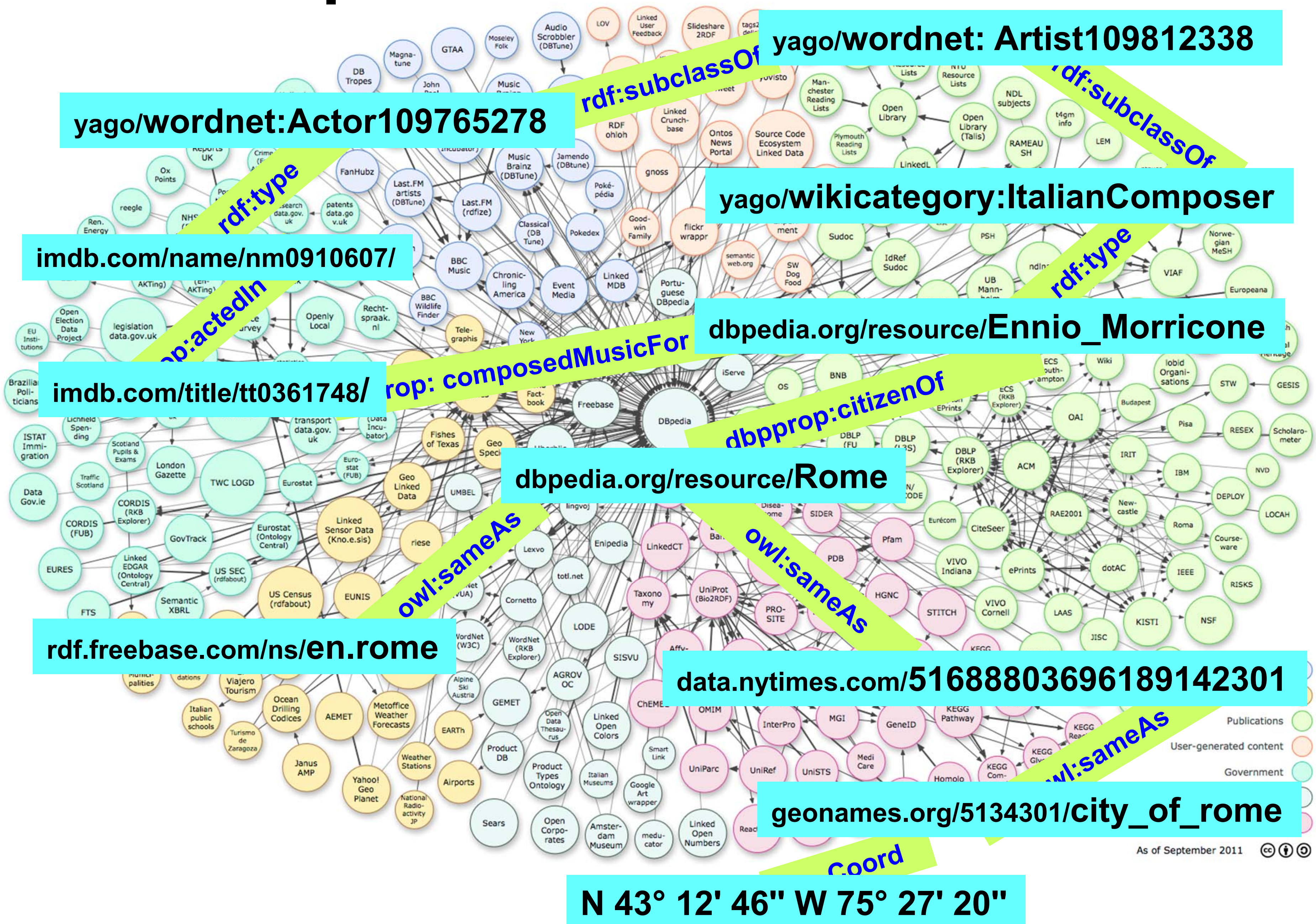
There are many public knowledge bases



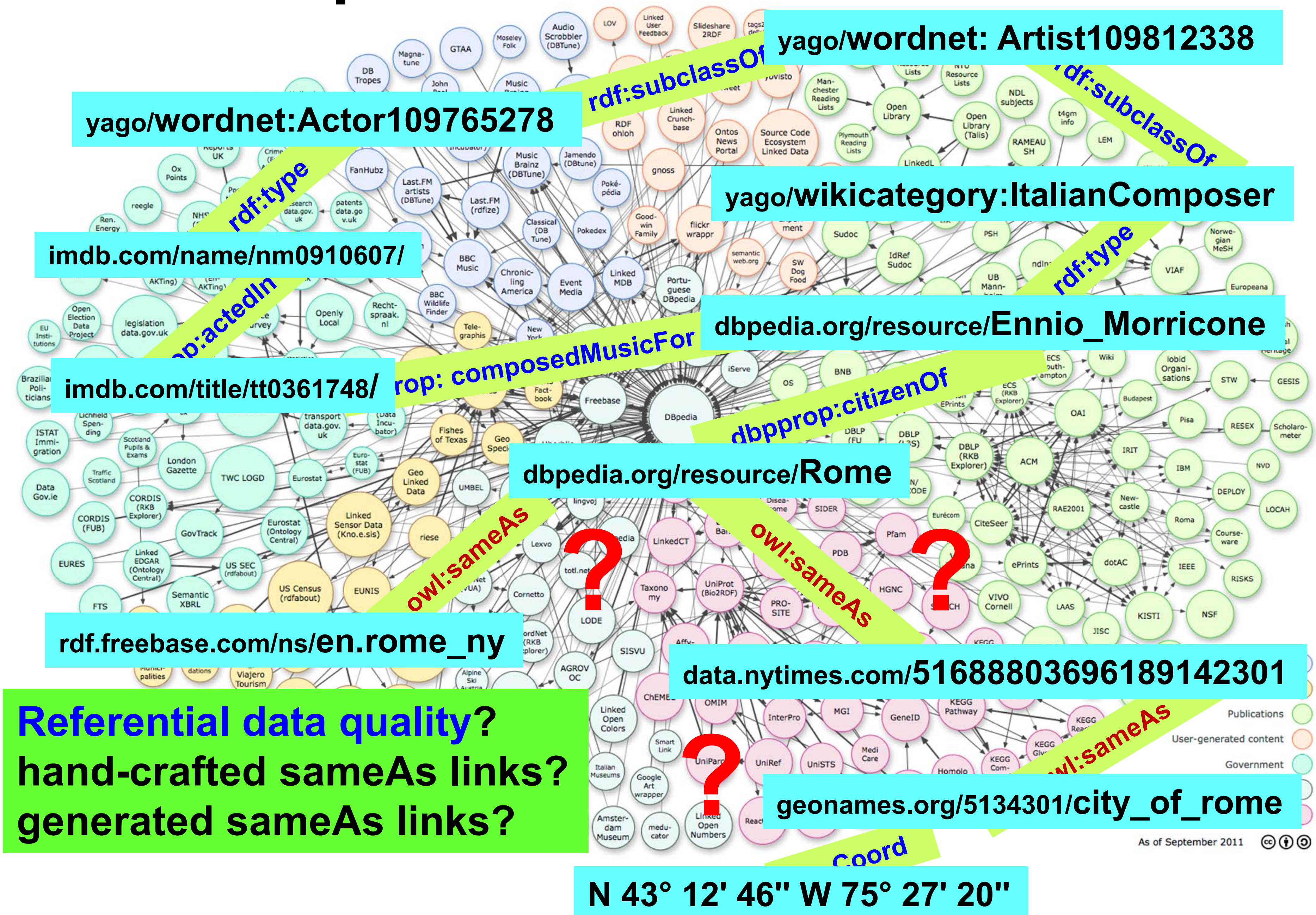
30 Bio. triples
500 Mio. links

http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.png

Link equivalent entities across KBs

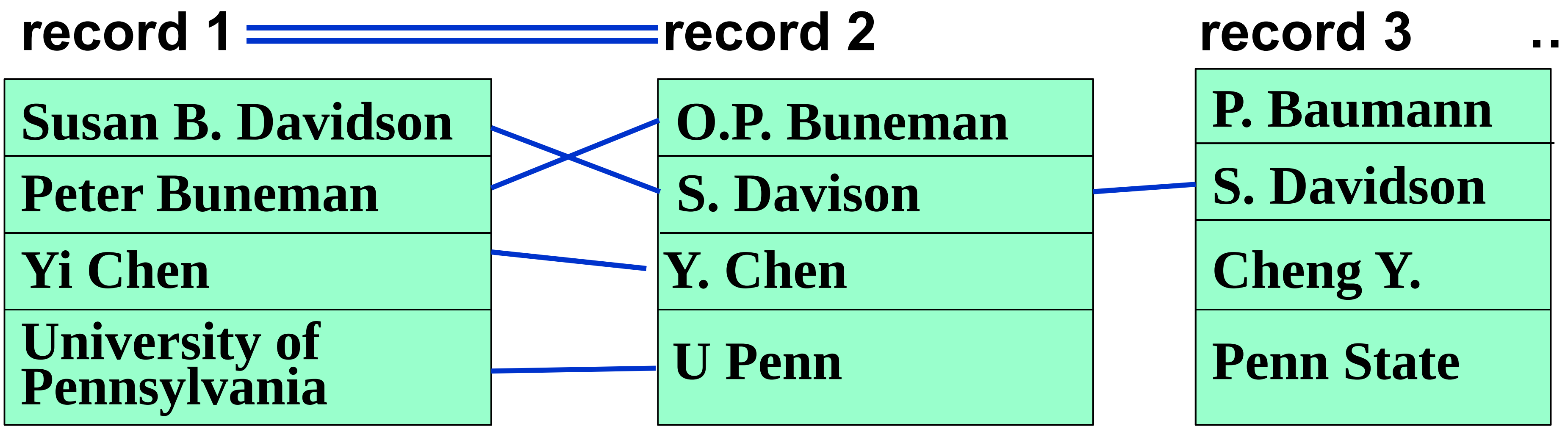


Link equivalent entities across KBs



Referential data quality?
hand-crafted sameAs links?
generated sameAs links?

Record Linkage between Databases



Goal: Find equivalence classes of entities, and of records

Techniques:

- similarity of values (edit distance, n-gram overlap, etc.)
- joint agreement of linkage
- similarity joins, grouping/clustering, collective learning, etc.
- often domain-specific customization (similarity measures etc.)

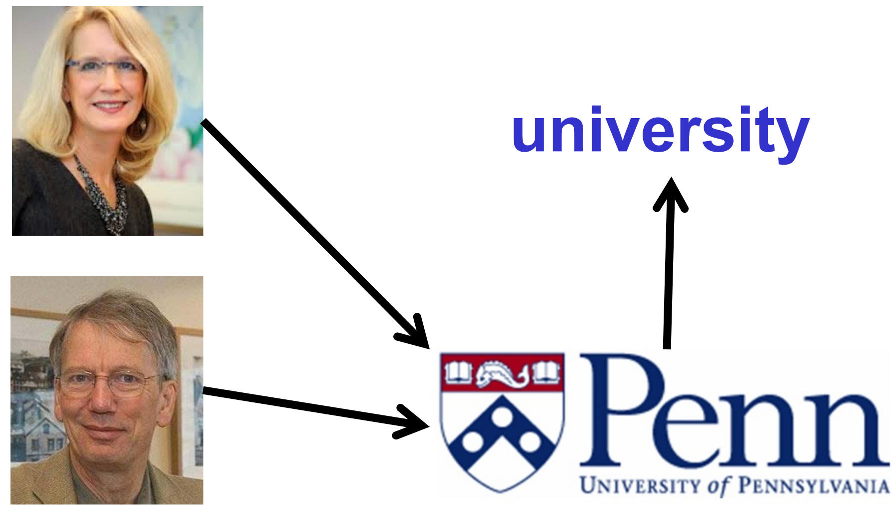
Halbert L. Dunn: Record Linkage. American Journal of Public Health. 1946
H.B. Newcombe et al.: Automatic Linkage of Vital Records. Science, 1959.
I.P. Fellegi, A.B. Sunter: A Theory of Record Linkage. J. of American Statistical Soc., 1969.

Linking Records vs. Linking Knowledge

record

Susan B. Davidson
Peter Buneman
Yi Chen
University of Pennsylvania

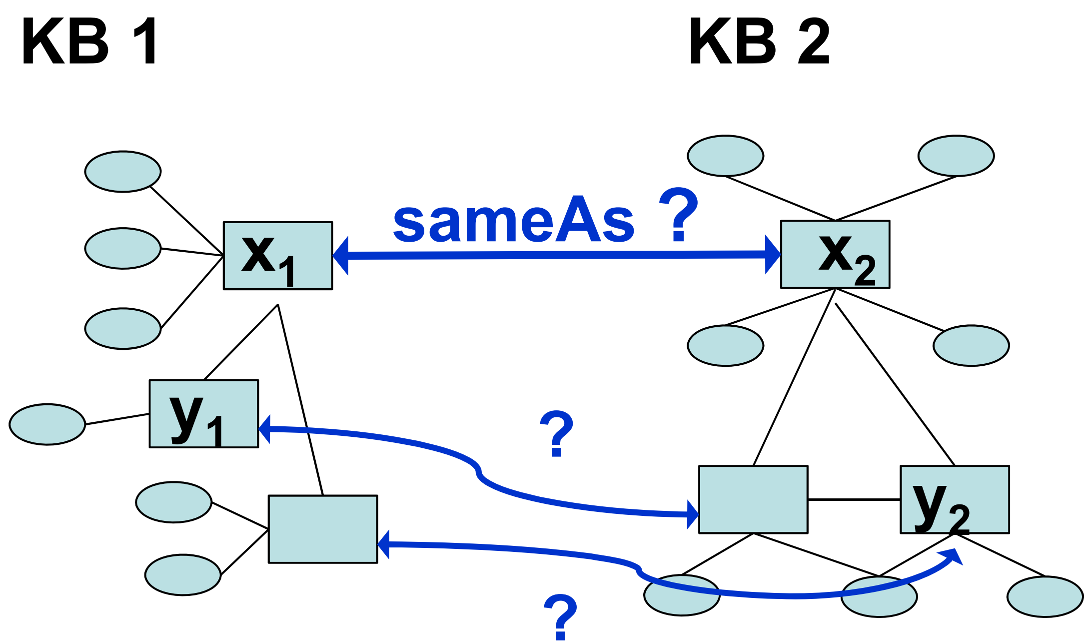
KB / Ontology



Differences between DB records and KB entities:

- Ontological links have rich semantics (e.g. subclassOf)
- Ontologies have only binary predicates
- Ontologies have no schema
- Match not just entities,
  but also classes & predicates (relations)

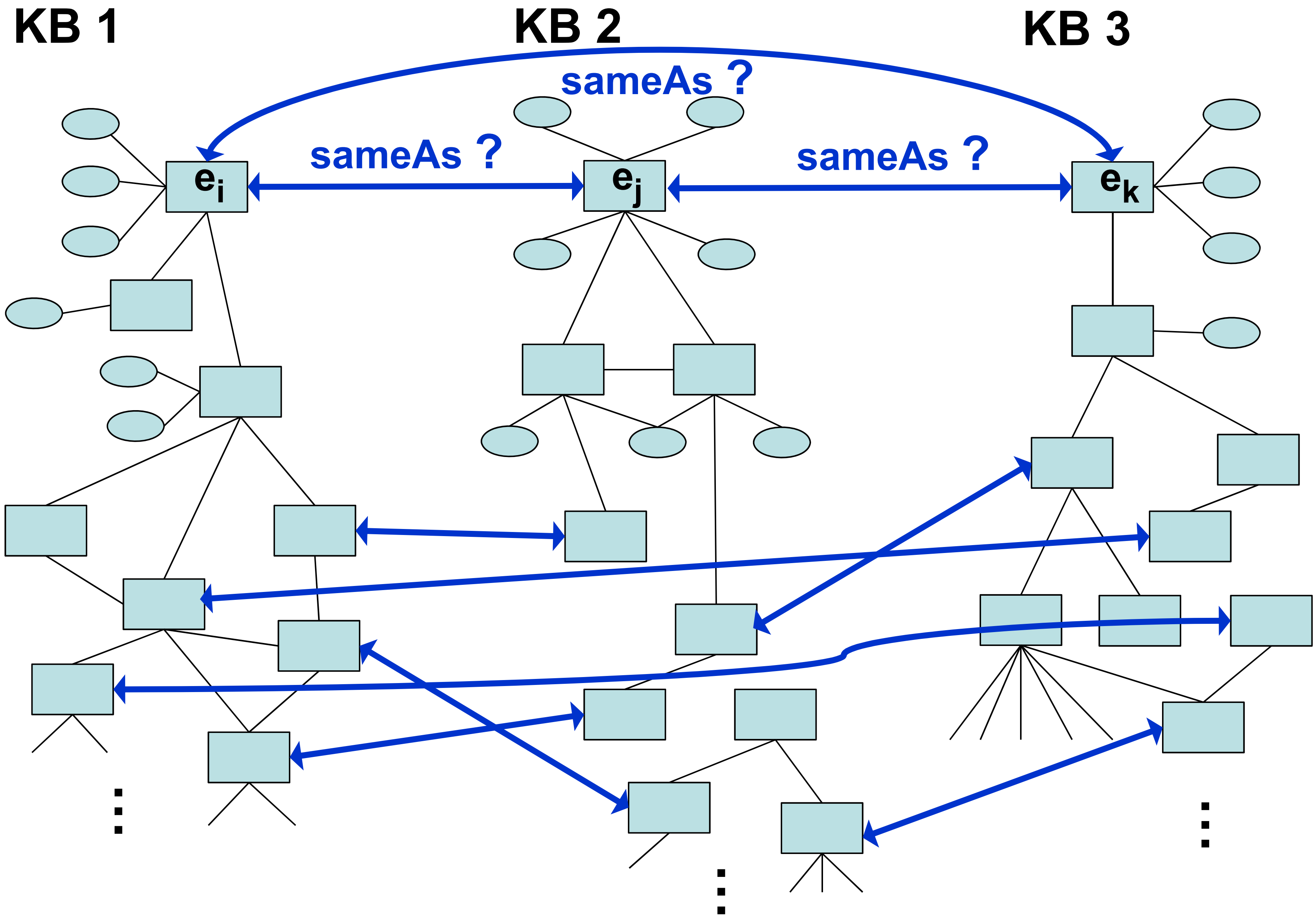
Similarity of entities depends on similarity of neighborhoods



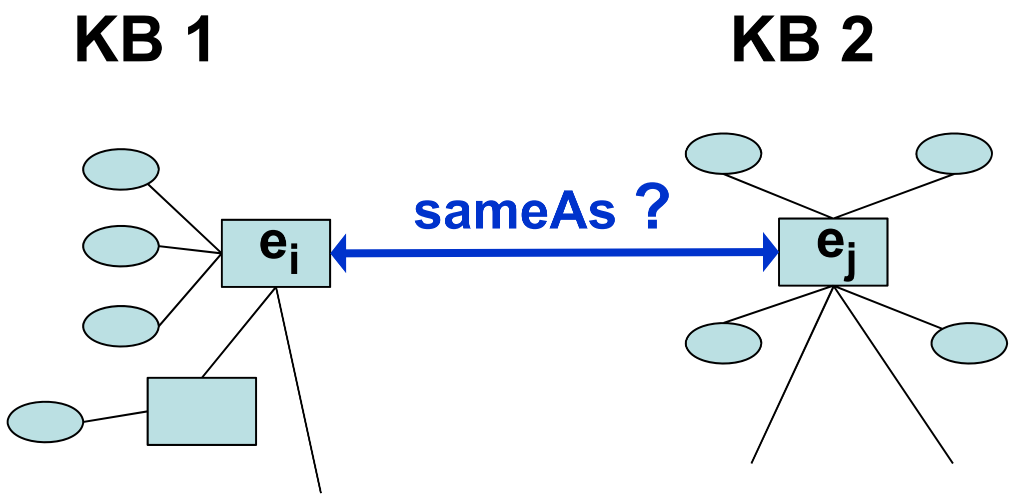
sameAs(x1, x2) depends on sameAs(y1, y2)

which depends on sameAs(x1, x2)

Equivalence of entities is transitive



Matching is an optimization problem



Define:

$sim(e_i, e_j) \in$ [-1,1]: Similarity of two entities

$coh(x, y) \in$ [-1,1]: likelihood of being mentioned together

decision variables $X_{ij} = 1$ if sameAs(x_i, x_j), else 0

Maximize

$\Sigma_{ij} X_{ij}$ (sim(e_i,e_j) + $\Sigma_{x \in N_i,\, y \in N_j}$ coh(x,y))
+ Σ_{jk} (…) + Σ_{ik} (…)

... under constraints:

$\forall i \sum_j X_{ij} < 1$

$\forall\, i, j, k$: $(1 - X_{ij}) + (1 - X_{jk}) \geq (1 - X_{ik})$

Problem cannot be solved at Web scale

KB 1 **KB 2**

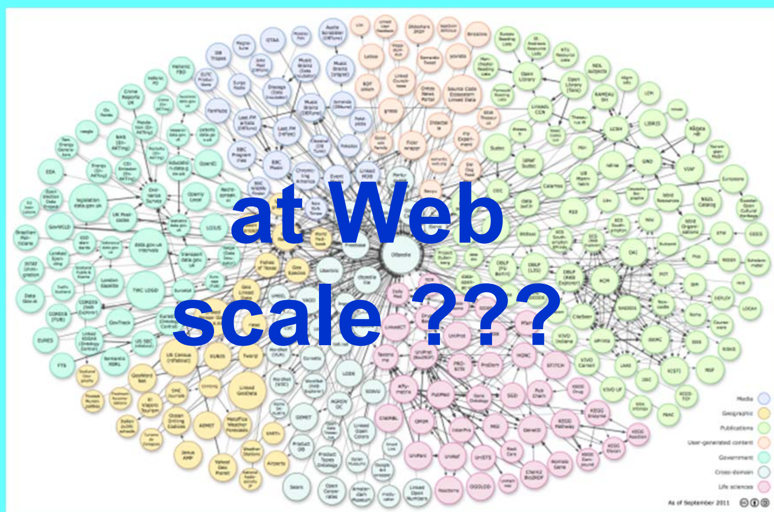
- **Joint Mapping**
- **ILP model**
  or prob. factor graph or …
- **Use your favorite solver**
- **How?**

Define

sim

coh

deci



at Web scale ???

Maxin

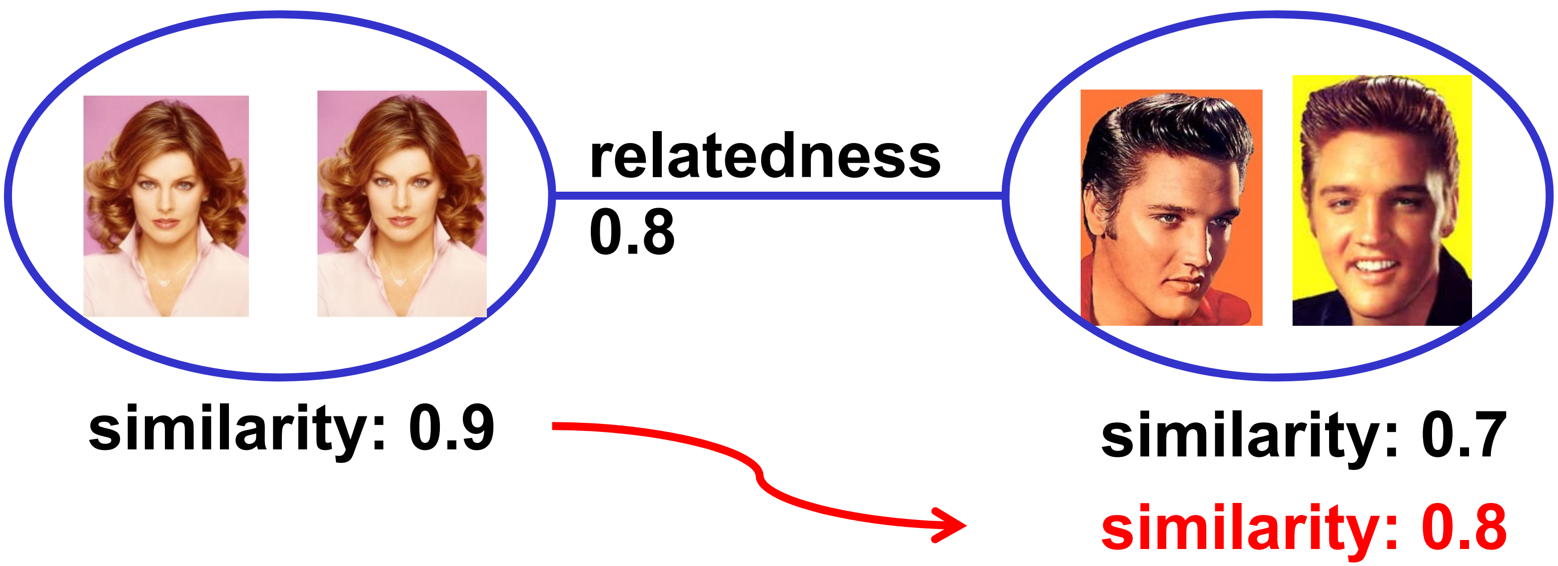
$\Sigma_{ij} X_{ij}$

$+ \Sigma_{jk} (\ldots) + \Sigma_{ik} (\ldots)$

$\forall\, i,j,k$: $(1-X_{ij}) + (1-X_{jk}) \geq (1-X_{ik})$

Similarity Flooding matches entities at scale

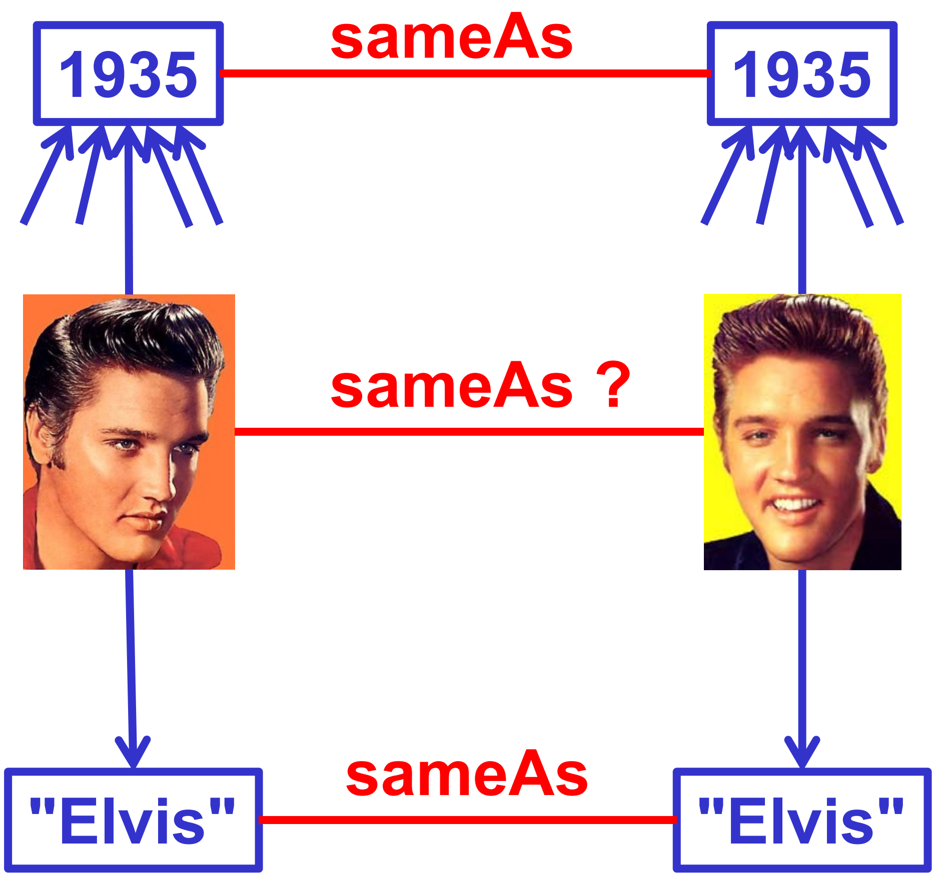
Build a graph:
  nodes: pairs of entities, weighted with similarity
  edges: weighted with degree of relatedness



Iterate until convergence:
  similarity := weighted sum of neighbor similarities

many variants (belief propagation, label propagation, etc.)

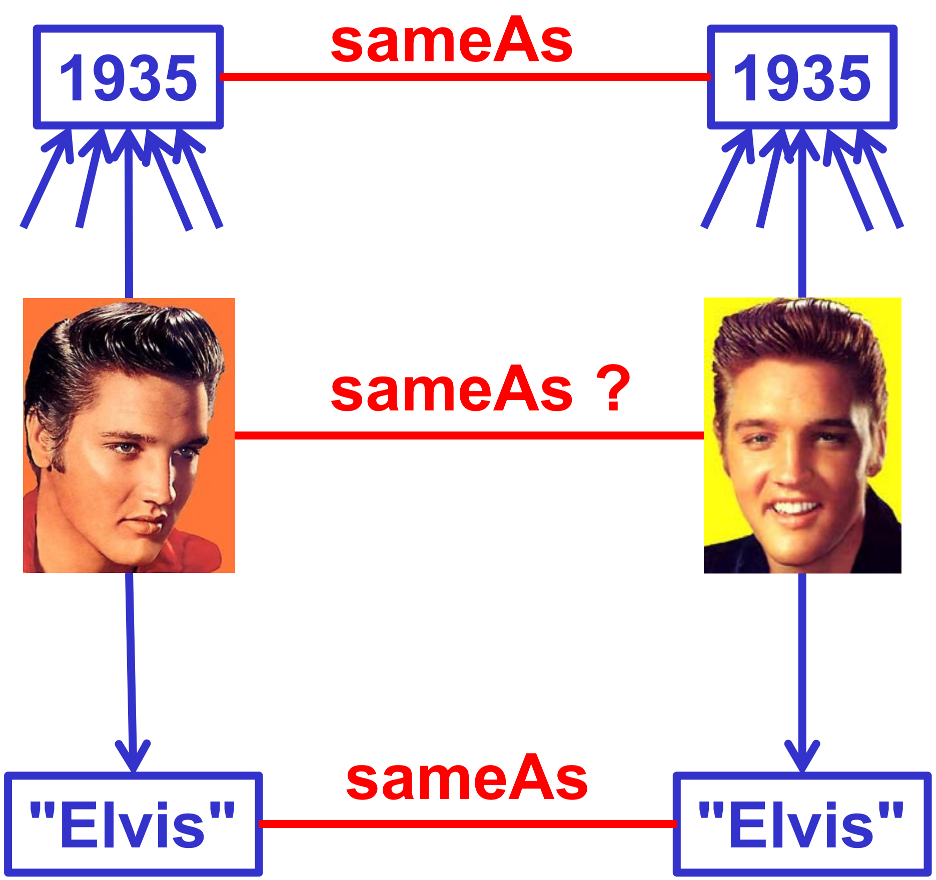
Some neighborhoods are more indicative



Many people born in 1935
⇒ not indicative

Few people called "Elvis"
⇒ highly indicative

Inverse functionality as indicativeness



$$ifun(r,y) = \frac{1}{|\{x: r(x,y)\}|}$$

$$ifun(born, 1935) = \frac{1}{5}$$

$$ifun(r) = HM_y\, ifun(r,y)$$

$$ifun(born) = 0.01$$

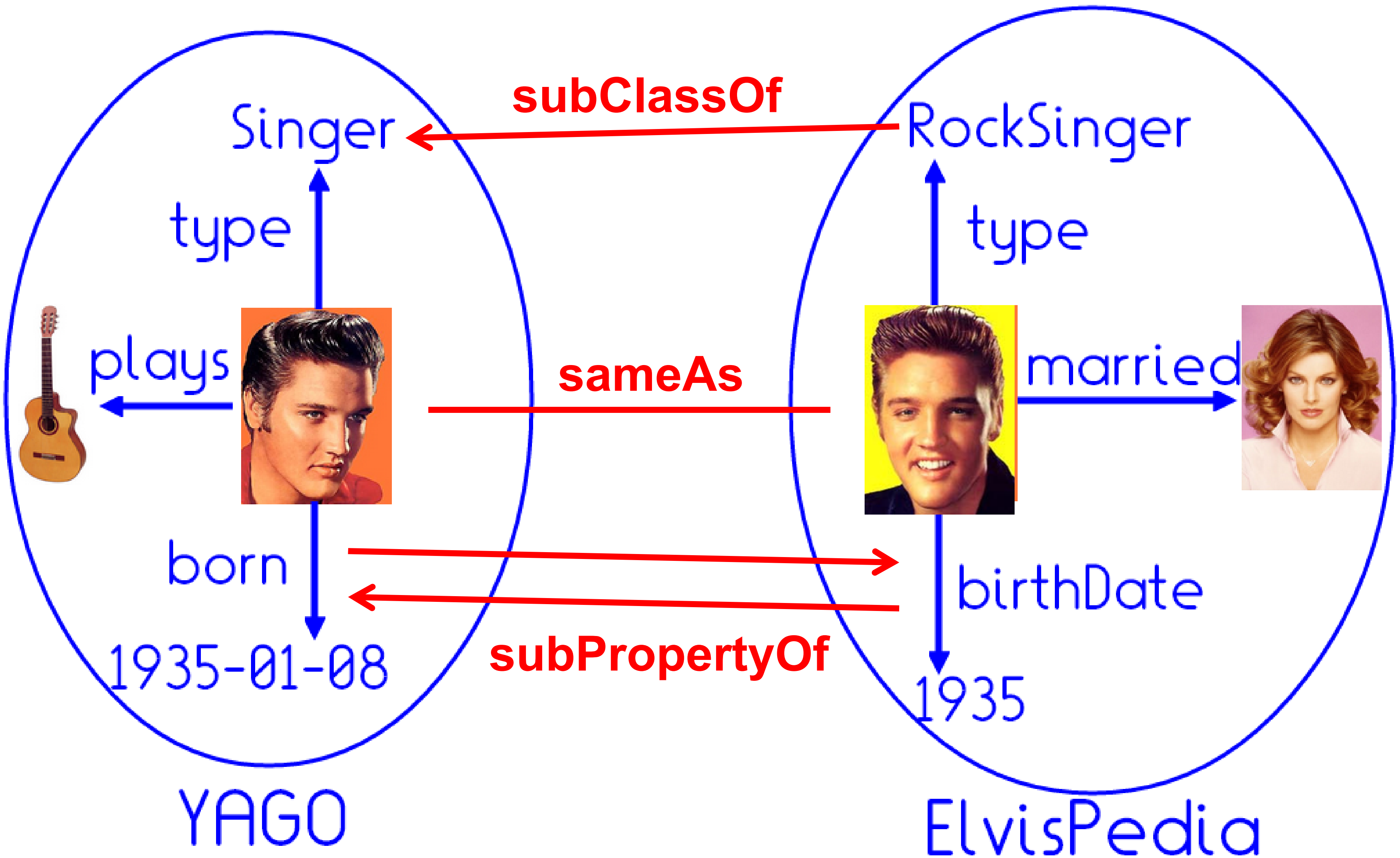
$$ifun(label) = 0.9$$

The higher the inverse functionality of r for r(x,y), r(x',y), the higher the likelihood that x=x'.

$ifun(r) = 1 \;\Rightarrow x = x'$

[Suchanek et al.: VLDB'12]

Match entities, classes and relations



PARIS matches entities, classes & relations

[Suchanek et al.: VLDB'12]

Goal:
given 2 ontologies, match entities, relations, and classes

Define
P(x ≡ y) := probability that entities x and y are the same
P(p ⊇ r) := probability that relation p subsumes r
P(c ⊇ d) := probability that class c subsumes d

Initialize
P(x ≡ y) := similarity if x and y are literals, else 0
P(p ⊇ r) := 0.001

Iterate until convergence
P(x ≡ y) := $\int 42\nabla e^{-i\omega t}$ ... $P(p \supseteq r)$
P(p ⊇ r) := $\vartheta\aleph + {}_{1}^{n}Y$ ... $P(x \equiv y)$

Recursive dependency

Compute
P(c ⊇ d) := ratio of instances of d that are in c

PARIS matches entities, classes & relations

[Suchanek et al.: VLDB'12]

Goal:

given 2 ontologies, match entities, relations, and classes

Defin

P(x

P(p

P(c

Initial

P(x

P(p

Iterat

P(x

P(p

PARIS matches YAGO and DBpedia

- time: 1:30 hours
- precision for instances: 90%
- precision for classes: 74%
- precision for relations: 96%

Compute

P(c $\supseteq$ d) := ratio of instances of d that are in c

Many challenges remain

Entity linkage is at the heart of semantic data integration.
More than 50 years of research, still some way to go!

- **Highly related entities with ambiguous names**
  George W. Bush (jun.) vs. George H.W. Bush (sen.)
- **Long-tail entities with sparse context**
- **Enterprise data (perhaps combined with Web2.0 data)**
- **Records with complex DB / XML / OWL schemas**
- **Entities with very noisy context (in social media)**
- **Ontologies with non-isomorphic structures**

Benchmarks:

- OAEI Ontology Alignment & Instance Matching: oaei.ontologymatching.org
- TAC KBP Entity Linking: www.nist.gov/tac/2012/KBP/
- TREC Knowledge Base Acceleration: trec-kba.org

Take-Home Lessons

Web of Linked Data is great

100‘s of KB‘s with 30 Bio. triples and 500 Mio. links
mostly reference data, dynamic maintenance is bottleneck
connection with Web of Contents needs improvement

Entity resolution & linkage is key

for creating sameAs links in text (RDFa, microdata)
for machine reading, semantic authoring,
knowledge base acceleration, …

Linking entities across KB‘s is advancing

Integrated methods for aligning entities, classes and relations

Open Problems and Grand Challenges

Web-scale, robust ER with high quality

Handle huge amounts of linked-data sources, Web tables, …

Combine algorithms and crowdsourcing

with active learning, minimizing human effort or cost/accuracy

Automatic and continuously maintained sameAs links for Web of Linked Data with high accuracy & coverage

Outline

- ✓ Motivation
- ★ Machine Knowledge
- ★ Taxonomic Knowledge: Entities and Classes
- ★ Contextual Knowledge: Entity Disambiguation
- ★ Linked Knowledge: Entity Resolution
- ★ Temporal & Commonsense Knowledge
- ★ Wrap-up

http://www.mpi-inf.mpg.de/yago-naga/icde2013-tutorial/

As Time Goes By: Temporal Knowledge

Which facts for given relations hold at what time point or during which time intervals ?

marriedTo (Madonna, GuyRitchie) [22Dec2000, Dec2008]
capitalOf (Berlin, Germany) [1990, now]
capitalOf (Bonn, Germany) [1949, 1989]
hasWonPrize (JimGray, TuringAward) [1998]
graduatedAt (HectorGarcia-Molina, Stanford) [1979]
graduatedAt (SusanDavidson, Princeton) [Oct 1982]
hasAdvisor (SusanDavidson, HectorGarcia-Molina) [Oct 1982, forever]

How can we query & reason on entity-relationship facts in a “time-travel“ manner - with uncertain/incomplete KB ?

US president‘s wife when Steve Jobs died?
students of Hector Garcia-Molina while he was at Princeton?

Temporal Knowledge

for all people in Wikipedia (300 000) gather all spouses, incl. divorced & widowed, and corresponding time periods!

>95% accuracy, >95% coverage, in one night

1) **recall: gather temporal scopes for base facts**
2) **precision: reason on mutual consistency**

	28 January 1955 (age 53)
	Paris, France
	Nicolas Paul Stéphane Sarközy
Political party	RR (?–2002) UMP (2002–)
Spouse	Marie-Dominique Culioli (div.) Cécilia Ciganer-Albéniz (div.) Carla Bruni
Children	Pierre (by Culioli) Jean (by Culioli) Louis (by Ciganer-Albéniz)
Residence	Élysée Palace
Alma mater	University of Paris X: Nanterre
Occupation	Lawyer
Religion	Roman Catholic

consistency constraints are potentially helpful:

- **functional dependencies:** *husband, time $\rightarrow$ wife*
- **inclusion dependencies:** *marriedPerson $\subseteq$ adultPerson*
- **age/time/gender restrictions:** *birthdate + Δ < marriage < divorce*

Dating Considered Harmful

explicit dates vs. implicit dates

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.

Machine-Reading Biographies

vague dates
relative dates

Early life

During Sarkozy's childhood, his father allegedly refused to give his wife
help, even though he had founded his own advertising agency and had become wealthy. The family lived in a mansion owned by Sarkozy's grandfather, Benedict Mallah, in the 17th Arrondissement of Paris. The family later moved to Neuilly-sur-Seine, one of the wealthiest

narrative text
relative order

Education

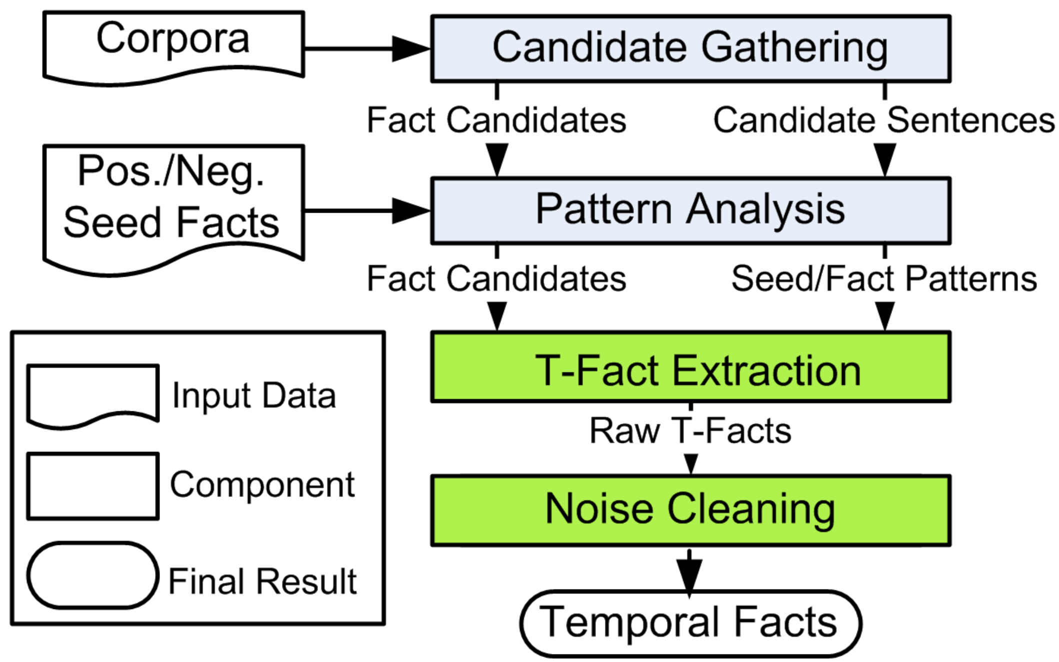
Sarkozy was enrolled in the *Lycée Chaptal*, a well regarded public midd
Paris's 8th arrondissement, where he failed his *sixième*. His family then sent him to the *Cours Saint-Louis de Monceau*, a private Catholic school in the 17th arrondissement, where he was reportedly a mediocre student,[9] but where he nonetheless obtained his *baccalauréat* in 1973. He enrolled at the *Université Paris X Nanterre*, where he graduated with an MA in Private law, and later with a DEA degree in Business law. Paris X Nanterre had been the starting place for the May '68 student movement and was still a stronghold of leftist students. Described as a quiet student, Sarkozy soon joined the right-wing student organization, in which he was very active. He completed his military service as a part time Air Force cleaner.[10] After graduating, he entered the *Institut d'Études Politiques de Paris*, better known as Sciences Po, (1979–1981) but failed to graduate[11] due to an insufficient

PRAVDA for T-Facts from Text

[Y. Wang et al. 2011]

Variation of the 4-stage framework with enhanced stages 3 and 4:

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



Reasoning on T-Fact Hypotheses

[Y. Wang et al. 2012, P. Talukdar et al. 2012]

Temporal-fact hypotheses:

m(Ca,Nic)@[2008,2012]{0.7}, m(Ca,Ben)@[2010]{0.8}, m(Ca,Mi)@[2007,2008]{0.2}, m(Cec,Nic)@[1996,2004]{0.9}, m(Cec,Nic)@[2006,2008]{0.8}, m(Nic,Ma){0.9}, …

Cast into evidence-weighted logic program or integer linear program with 0-1 variables:

for temporal-fact hypotheses X_i
and pair-wise ordering hypotheses P_{ij}
maximize $\Sigma\, w_i X_i$ with constraints

- $X_i + X_j \leq 1$
  if X_i, X_j overlap in time & conflict
- $P_{ij} + P_{ji} \leq 1$
- $(1 - P_{ij}) + (1 - P_{jk}) \geq (1 - P_{ik})$
  if X_i, X_j, X_k must be totally ordered
- $(1 - X_i) + (1 - X_j) + 1 \geq (1 - P_{ij}) + (1 - P_{ji})$
  if X_i, X_j must be totally ordered

Efficient ILP solvers:
www.gurobi.com
IBM Cplex
…

Commonsense Knowledge

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 or cupboard, on the stove, …
but not in in the bedroom, in your pocket, in the sky, …

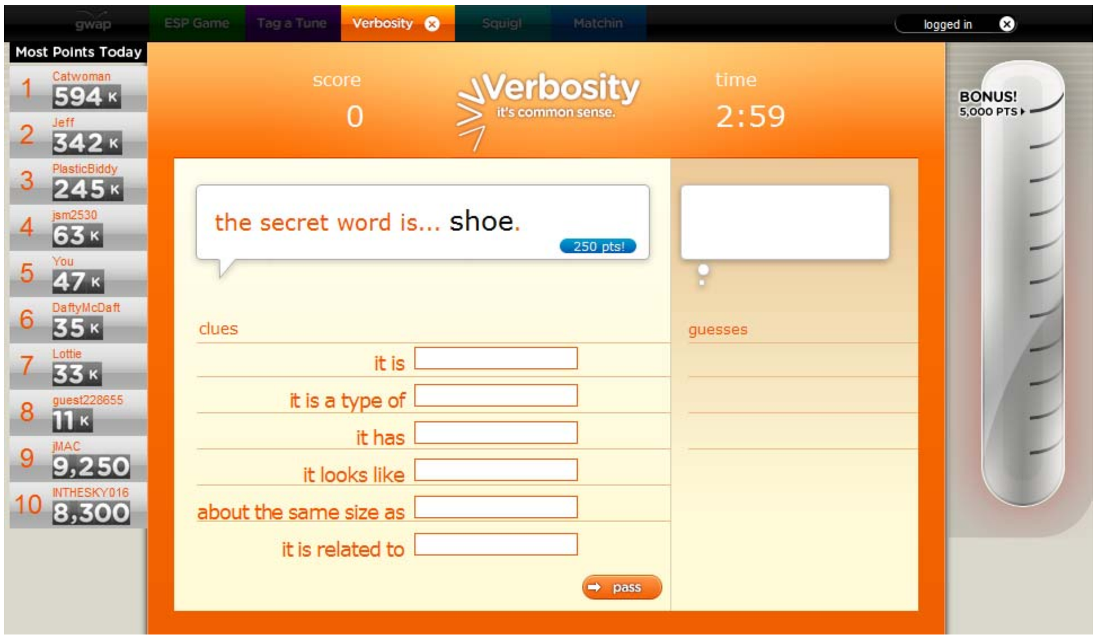
Approach 1: Crowdsourcing
→ ConceptNet (Speer/Havasi)
Problem: coverage and scale

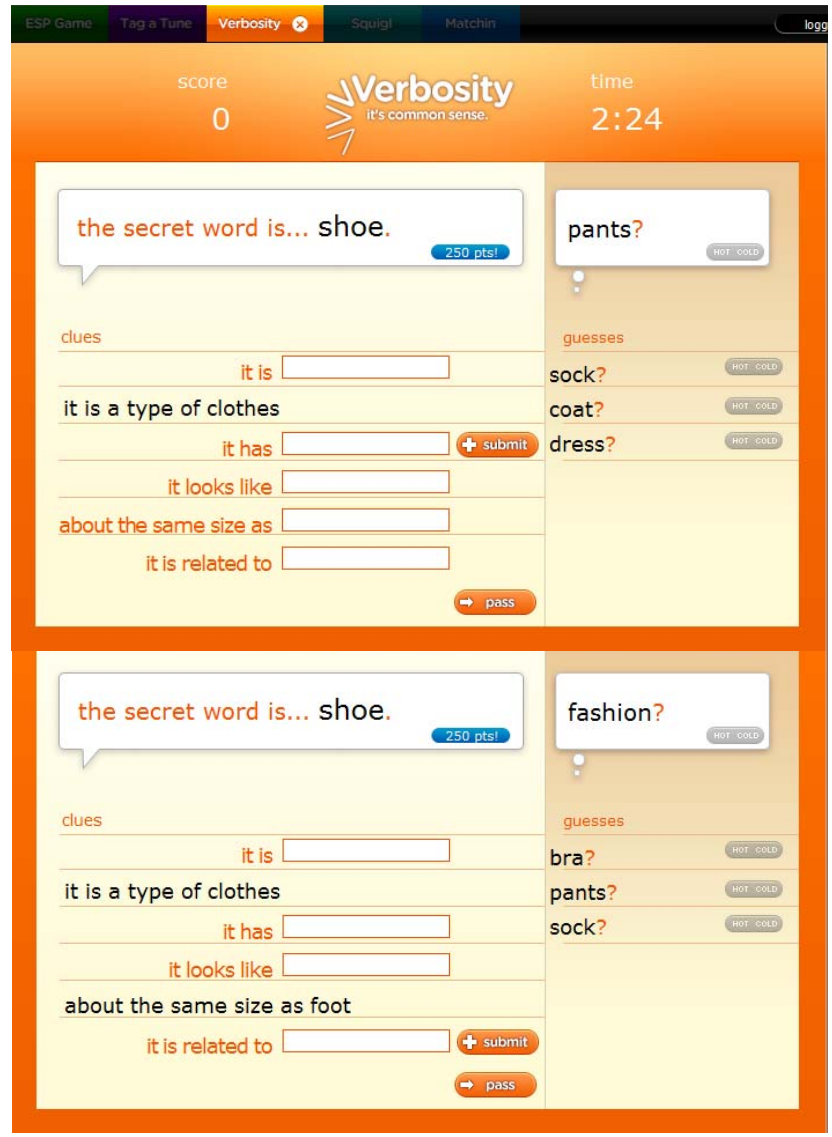
Approach 2: Pattern-based harvesting
→ CSK (Tandon et al., part of Yago-Naga project)
Problem: noise and robustness

Crowdsourcing for Commonsense Knowledge

[Speer & Havasi 2012]

many inputs incl. WordNet, Verbosity game, etc.





http://www.gwap.com/gwap/

Pattern-Based Harvesting of Commonsense Knowledge

(N. Tandon et al.: AAAI 2011)

Approach 2: Use Seeds for Pattern-Based Harvesting

Gather and analyze patterns and occurrences for

<common noun> hasProperty <adjective>
<common noun> hasAbility <verb>
<common noun> hasLocation <common noun>

→ Patterns: X is very Y, X can Y, X put in/on Y, …

Problem: noise and sparseness of data
Solution: harness Web-scale n-gram corpora
→ 5-grams + frequencies

Confidence score: PMI (X,Y), PMI (p,(XY)), support(X,Y), …
are features for regression model

Patterns indicate commonsense rules

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

Rule mining builds conjunctions

[L. Galarraga et al.: WWW'13]

inductive logic programming / association rule mining
but: with open world assumption (OWA)

$\boldsymbol{motherOf(x,z) \wedge marriedTo(x,y)}$ **#y,z: 1000**

$\boldsymbol{motherOf(x,z) \wedge marriedTo(x,y) \wedge fatherOf(y,z)}$ **#y,z: 600**

$\exists w$: $\boldsymbol{motherOf(x,z) \wedge marriedTo(x,y) \wedge fatherOf(w,z)}$ **#y,z: 800**

$\boldsymbol{motherOf(x,z) \wedge marriedTo(x,y) \Rightarrow fatherOf(y,z)}$

std. conf.: 600/1000

OWA conf.: 600/800

AMIE inferred 1000's of commonsense rules from YAGO2

$$\boldsymbol{marriedTo(x,y) \wedge livesIn(x,z) \Rightarrow livesIn(y,z)}$$

$$\boldsymbol{bornIn(x,y) \wedge locatedIn(y,z) \Rightarrow citizenOf(x,z)}$$

$$\boldsymbol{hasWonPrize(x, LeibnizPreis) \Rightarrow livesIn(x, Germany)}$$

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

Take-Home Lessons

Temporal knowledge harvesting:
crucial for machine-reading news, social media, opinions
statistical patterns and logical consistency are key,
harder than for „ordinary“ relations

Commonsense knowledge is cool & open topic:
can combine rule mining, patterns, crowdsourcing, AI, …

Open Problems and Grand Challenges


Robust and broadly applicable methods for temporal (and spatial) knowledge

populate time-sensitive relations comprehensively: marriedTo, isCEOof, participatedInEvent, …

Comprehensive commonsense knowledge organized in ontologically clean manner

especially for emotions and visually relevant aspects

Outline

✓ Motivation

★ Machine Knowledge

★ Taxonomic Knowledge: Entities and Classes

★ Contextual Knowledge: Entity Disambiguation

★ Linked Knowledge: Entity Resolution

★ Temporal & Commonsense Knowledge

★ Wrap-up

http://www.mpi-inf.mpg.de/yago-naga/icde2013-tutorial/

Summary

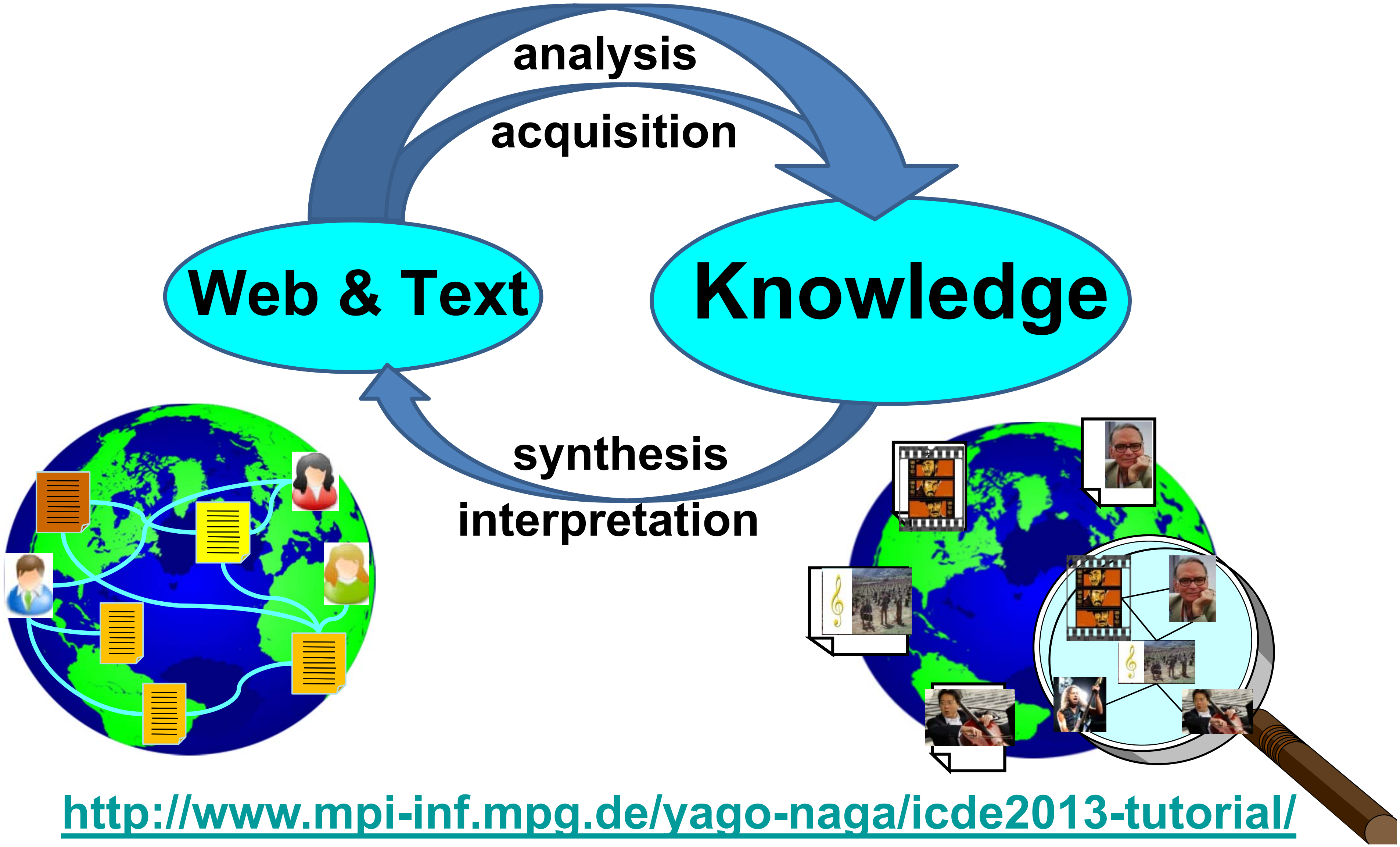
- **Knowledge Bases from Web are Real, Big & Useful:**
  Entities, Classes & Relations
- **Key Asset for Intelligent Applications:**
  Semantic Search, Question Answering, Machine Reading, Digital Humanities, Text&Data Analytics, Summarization, Reasoning, Smart Recommendations, …
- **Harvesting Methods for Entities & Classes Taxonomies**
- **Methods for Relational Facts Not Covered Here**
- **NERD & ER: Methods for Contextual & Linked Knowledge**
- **Rich Research Challenges & Opportunities:**
  scale & robustness; temporal, multimodal, commonsense; open & real-time knowledge discovery; …
- **Models & Methods from Different Communities:**
  DB, Web, AI, IR, NLP

References

see comprehensive list in

Fabian Suchanek and Gerhard Weikum:
Knowledge Harvesting from Text and Web Sources,
Proceedings of the 29th IEEE International
Conference on Data Engineering,
Brisbane, Australia, April 8-11, 2013,
IEEE Computer Society, 2013.

Take-Home Message:
From Web & Text to Knowledge



http://www.mpi-inf.mpg.de/yago-naga/icde2013-tutorial/