max planck institut
informatik

Knowledge Harvesting in the Big Data Era

Fabian Suchanek & Gerhard Weikum

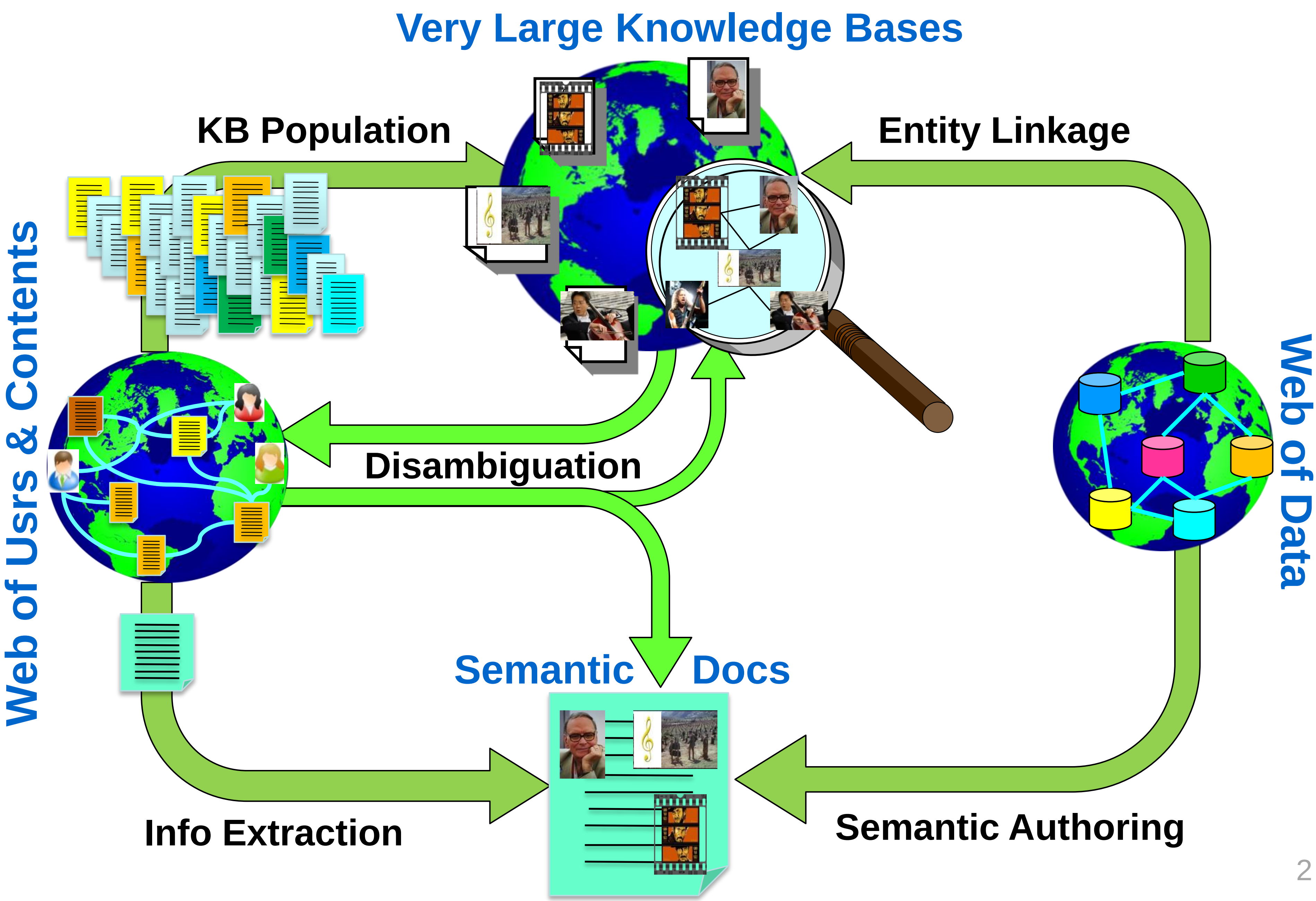
Max Planck Institute for Informatics, Saarbruecken, Germany

http://suchanek.name/

http://www.mpi-inf.mpg.de/~weikum/

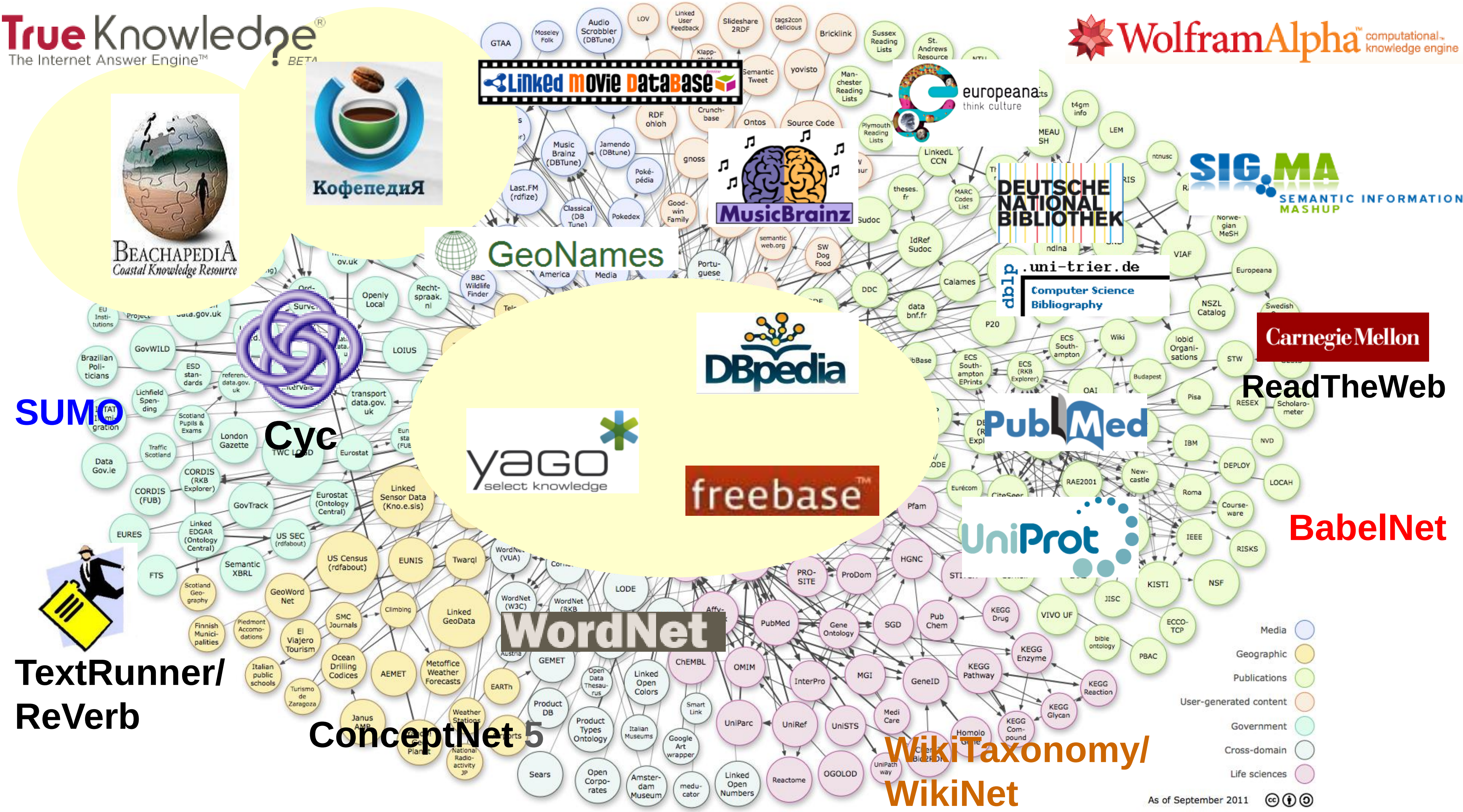
http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Turn Web into Knowledge Base



Web of Data: RDF, Tables, Microdata

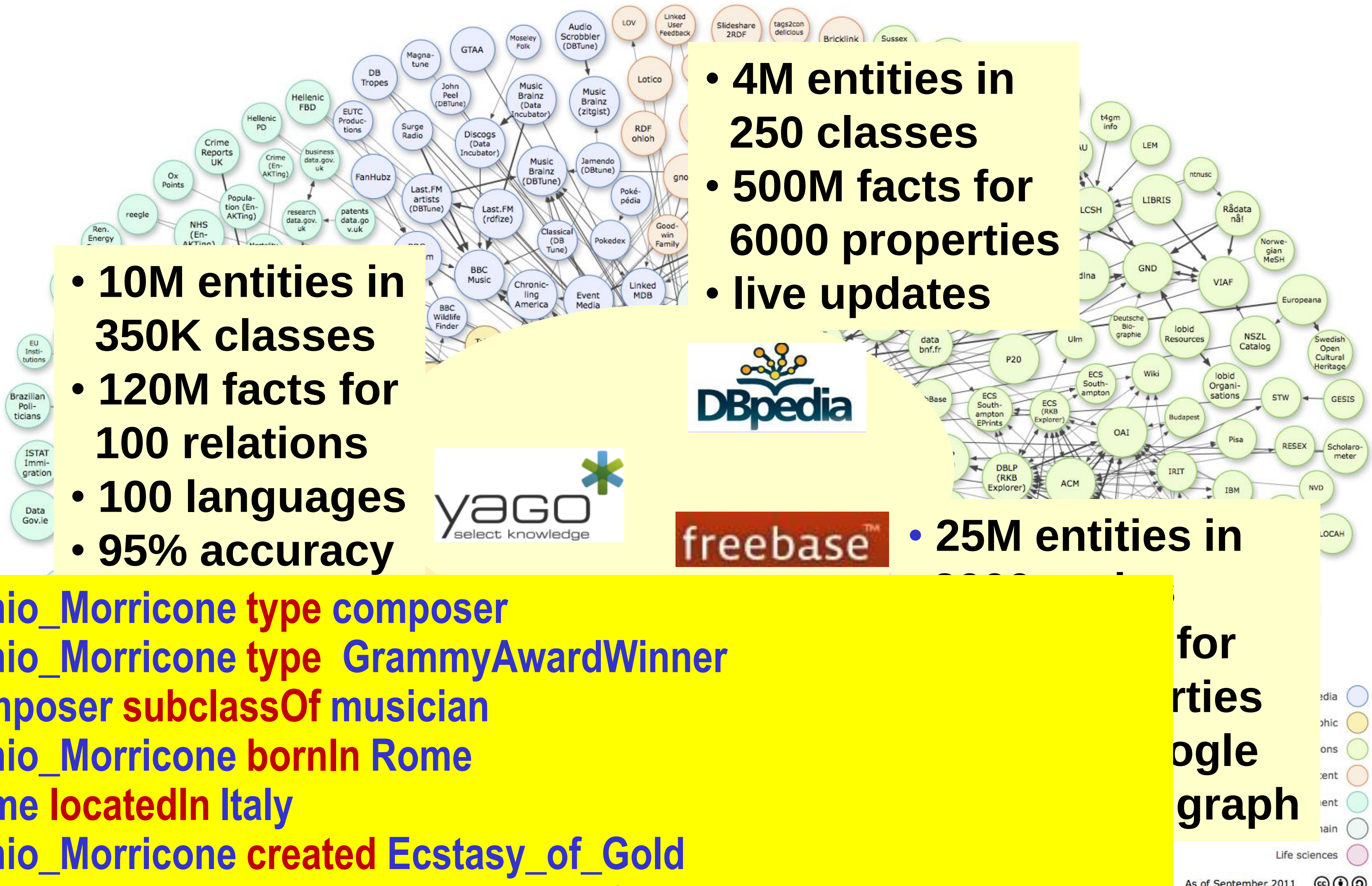
60 Bio. SPO triples (RDF) and growing



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

Web of Data: RDF, Tables, Microdata

60 Bio. SPO triples (RDF) and growing



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

DBpedia

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

yago
select knowledge

freebase

- **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

colored.png

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)

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

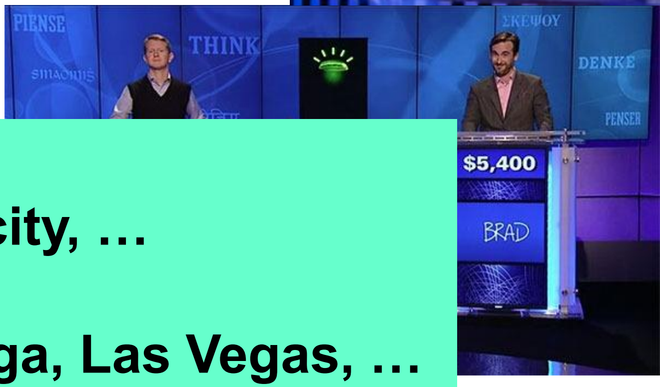
Knowledge for Intelligence

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?
- ★ East coast professors who founded Internet companies?
- ★ Relationships between
  John Lennon, Billie Holiday, Heath Ledger, King Kong?
- ★ Enzymes that inhibit HIV?
  Influenza drugs for teens with high blood pressure?
  ...

Use Case: Question Answering



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, …

This American city has two airports named after a war hero and a WW II battle

question classification & decomposition → knowledge back-ends

WIKIPEDIA The Free Encyclopedia

DBpedia

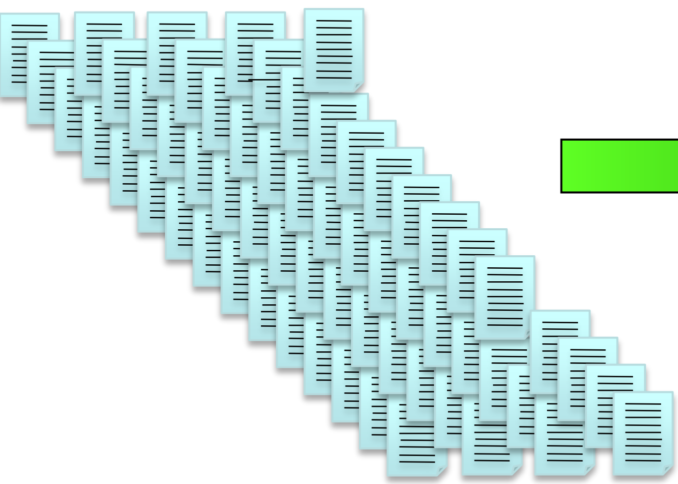
freebase

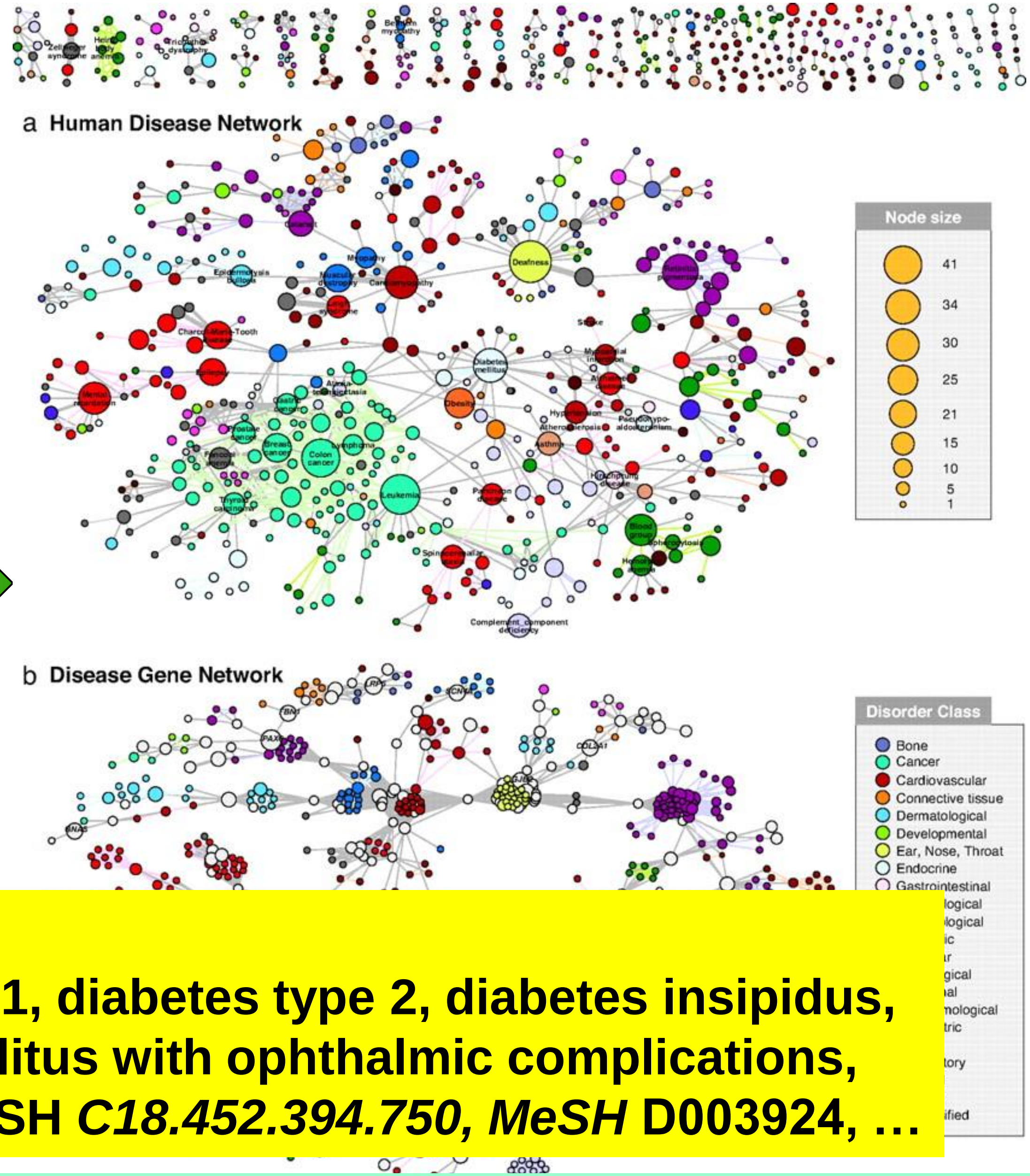
yago select knowledge

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

Use Case: Text Analytics

PubMed





But try this with:

diabetes mellitus, diabetis type 1, diabetes type 2, diabetes insipidus, insulin-dependent diabetes mellitus with ophthalmic complications, ICD-10 E23.2, OMIM 304800, MeSH *C18.452.394.750, MeSH* D003924, …

K.Goh,M.Kusick,D.Valle,B.Childs,M.Vidal,A.Barabasi: The Human Disease Network, PNAS, May 2007

Use Case: Big Data+Text Analytics

Entertainment:

Who covered which other singer?
Who influenced which other musicians?

Health: **Drugs (combinations) and their side effects**

Politics: **Politicians‘ positions on controversial topics and their involvement with industry**

Business: **Customer opinions on small-company products, gathered from social media**

General Design Pattern:

- **Identify relevant contents sources**
- **Identify entities of interest & their relationships**
- **Position in time & space**
- **Group and aggregate**
- **Find insightful patterns & predict trends**

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)

emerging 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))

Knowledge Bases in the Big Data Era

Big Data Analytics

- **Scalable algorithms**
- **Distributed platforms**
- **Discovering data sources**
- **Tapping unstructured data**
- **Connecting structured & unstructured data sources**
- **Making sense of heterogeneous, dirty, or uncertain data**

Knowledge Bases:
entities, **relations**, **time, space, …**

Outline

- ✓ **Motivation and Overview**
- ★ **Taxonomic Knowledge:**
  Entities and Classes
- ★ **Factual Knowledge:**
  Relations between Entities
- ★ **Emerging Knowledge:**
  New Entities & Relations
- ★ **Temporal Knowledge:**
  Validity Times of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

Big Data Methods for Knowledge Harvesting

Knowledge for Big Data Analytics

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

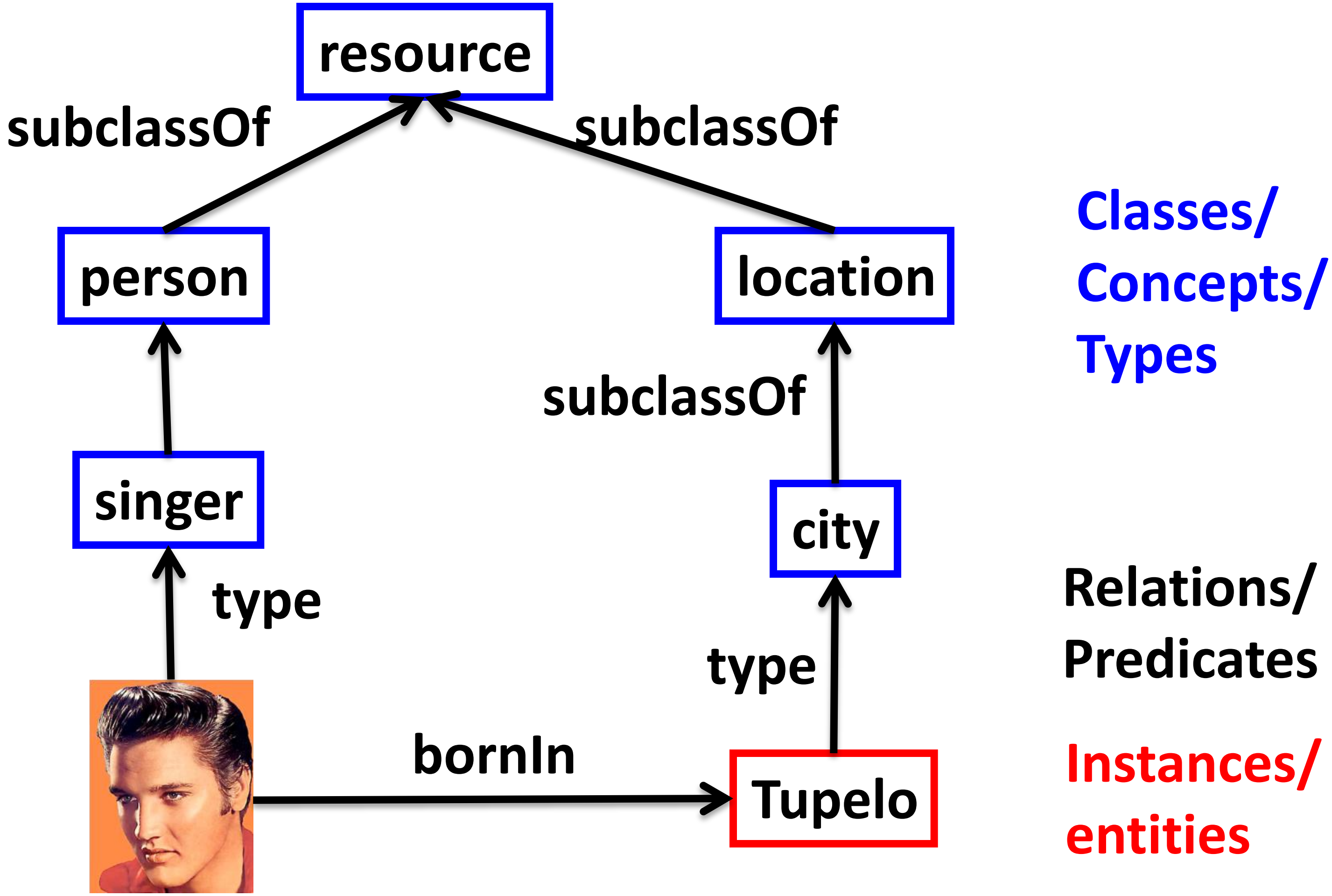
Outline

- ✓ **Motivation and Overview**
- ★ **Taxonomic Knowledge:**
  Entities and Classes
  - ★ **Scope & Goal**
  - ★ **Wikipedia-centric Methods**
  - ★ **Web-based Methods**
- ★ **Factual Knowledge:**
  Relations between Entities

- ★ **Emerging Knowledge:**
  New Entities & Relations
- ★ **Temporal Knowledge:**
  Validity Time of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

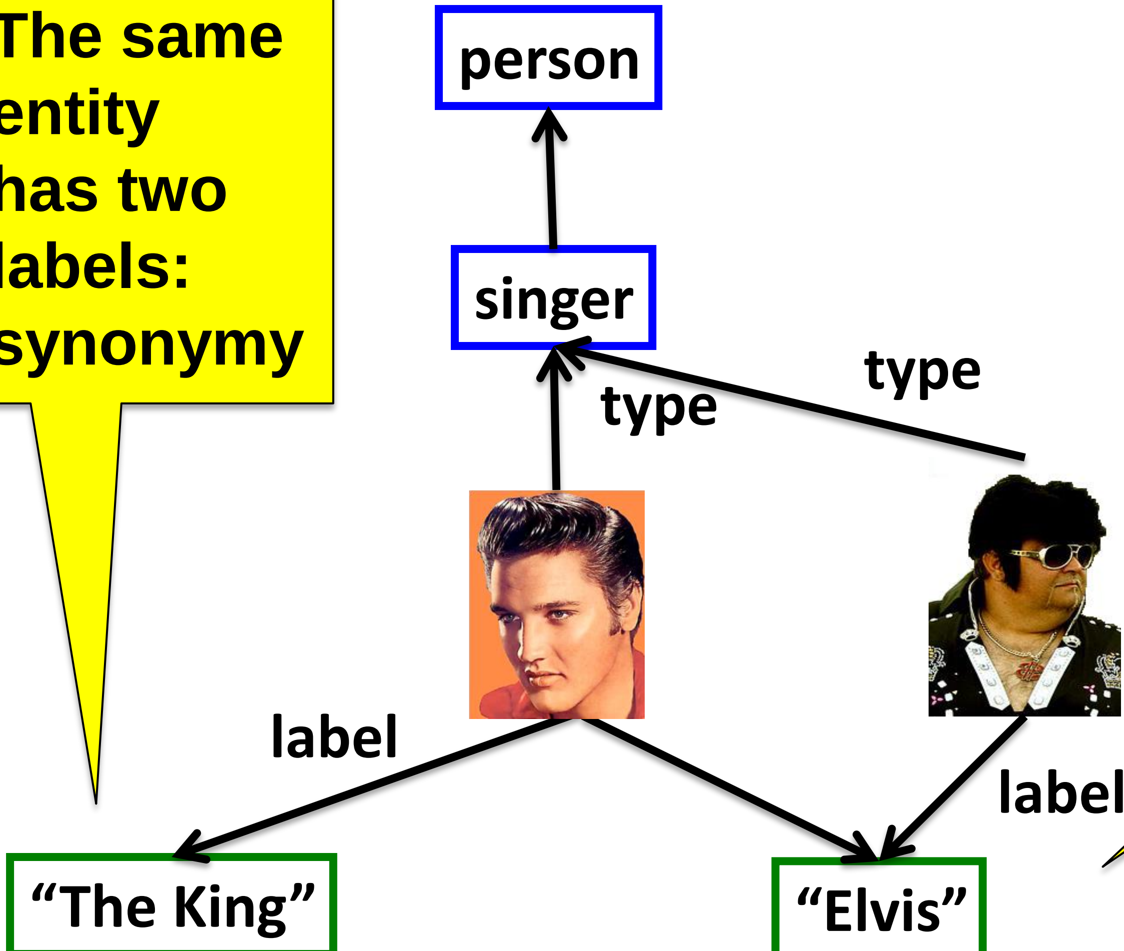
Knowledge Bases are labeled graphs



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

The same entity has two labels: synonymy

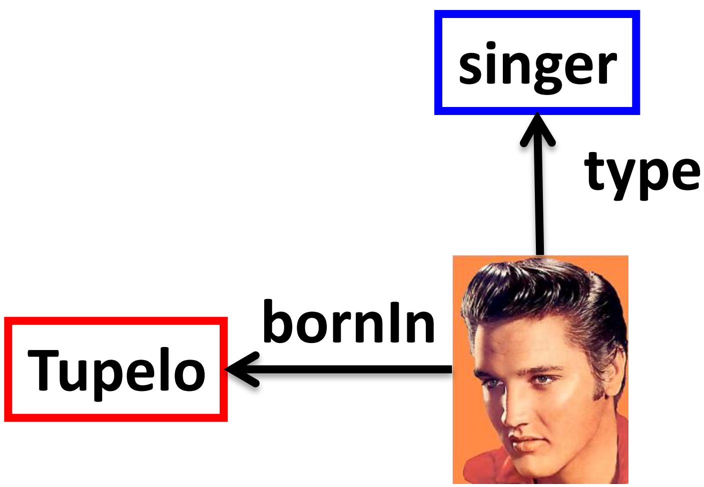


The same label for two entities: ambiguity

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)
...

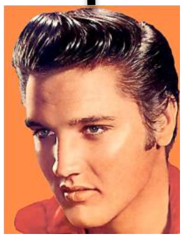
Our Goal is finding classes and instances

person

subclassOf

singer

type



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

singer

"person"

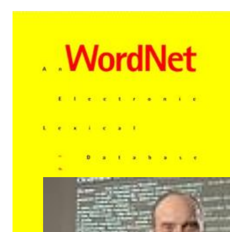
"individual"

"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 (United States folk singer 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 Gassion
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.

Outline

- ✓ **Motivation and Overview**
- ★ **Taxonomic Knowledge:**
  Entities and Classes
  - ✓ **Basics & Goal**
  - ✶ **Wikipedia-centric Methods**
  - ✶ **Web-based Methods**
- ★ **Factual Knowledge:**
  Relations between Entities
- ★ **Emerging Knowledge:**
  New Entities & Relations
- ★ **Temporal Knowledge:**
  Validity Times of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Wikipedia is a rich source of instances

Jimmy Wales

Larry Sanger

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'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?

American billionaires → tycoon, magnate

Technology company founders → entrepreneur

Apple Inc. →

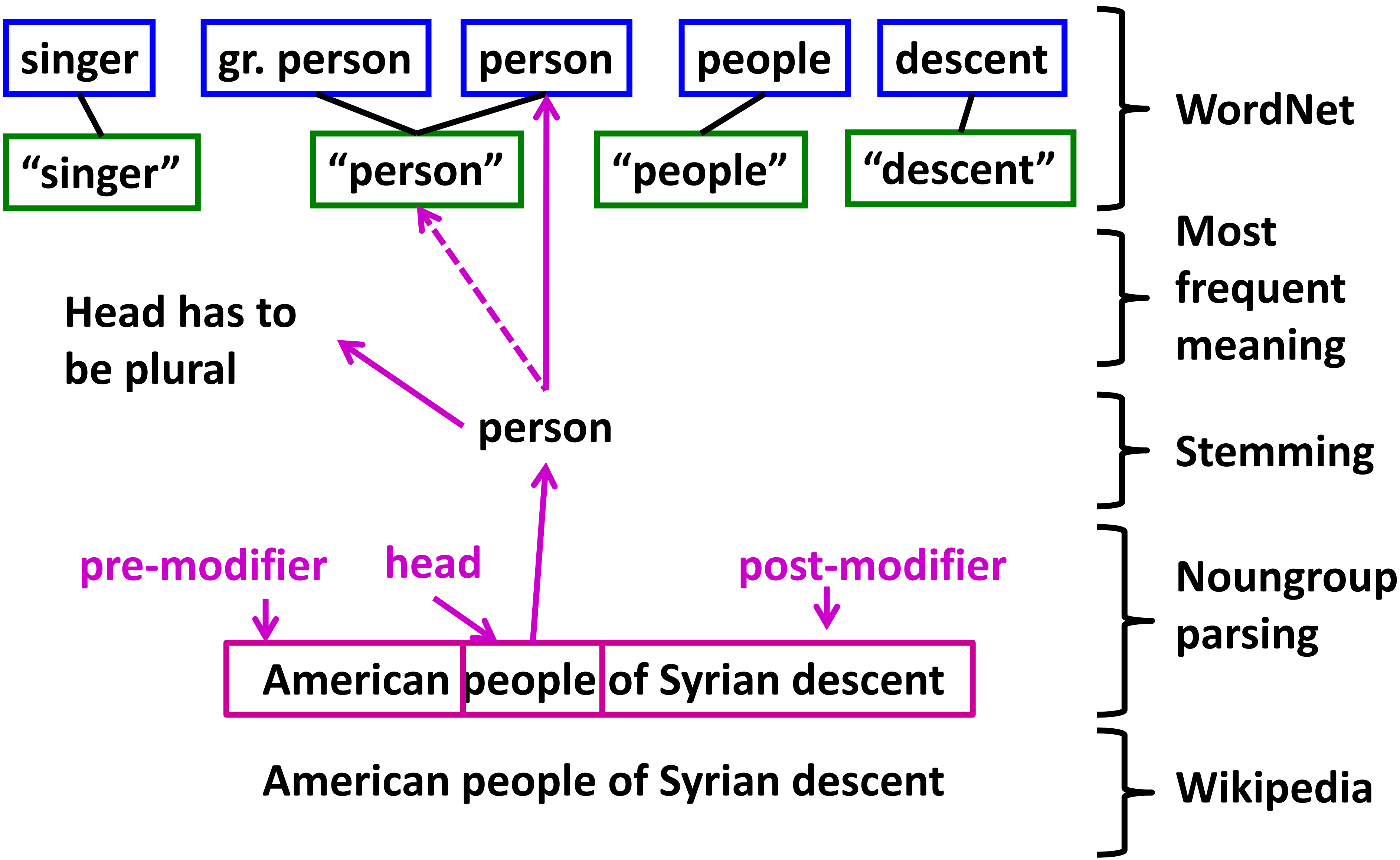
Deaths from cancer → ?

Internet pioneers → pioneer, innovator ?

Internet pioneers → pioneer, colonist

Wikipedia categories | **WordNet classes**

Categories can be linked to WordNet



YAGO = WordNet+Wikipedia

yago
select knowledge

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]

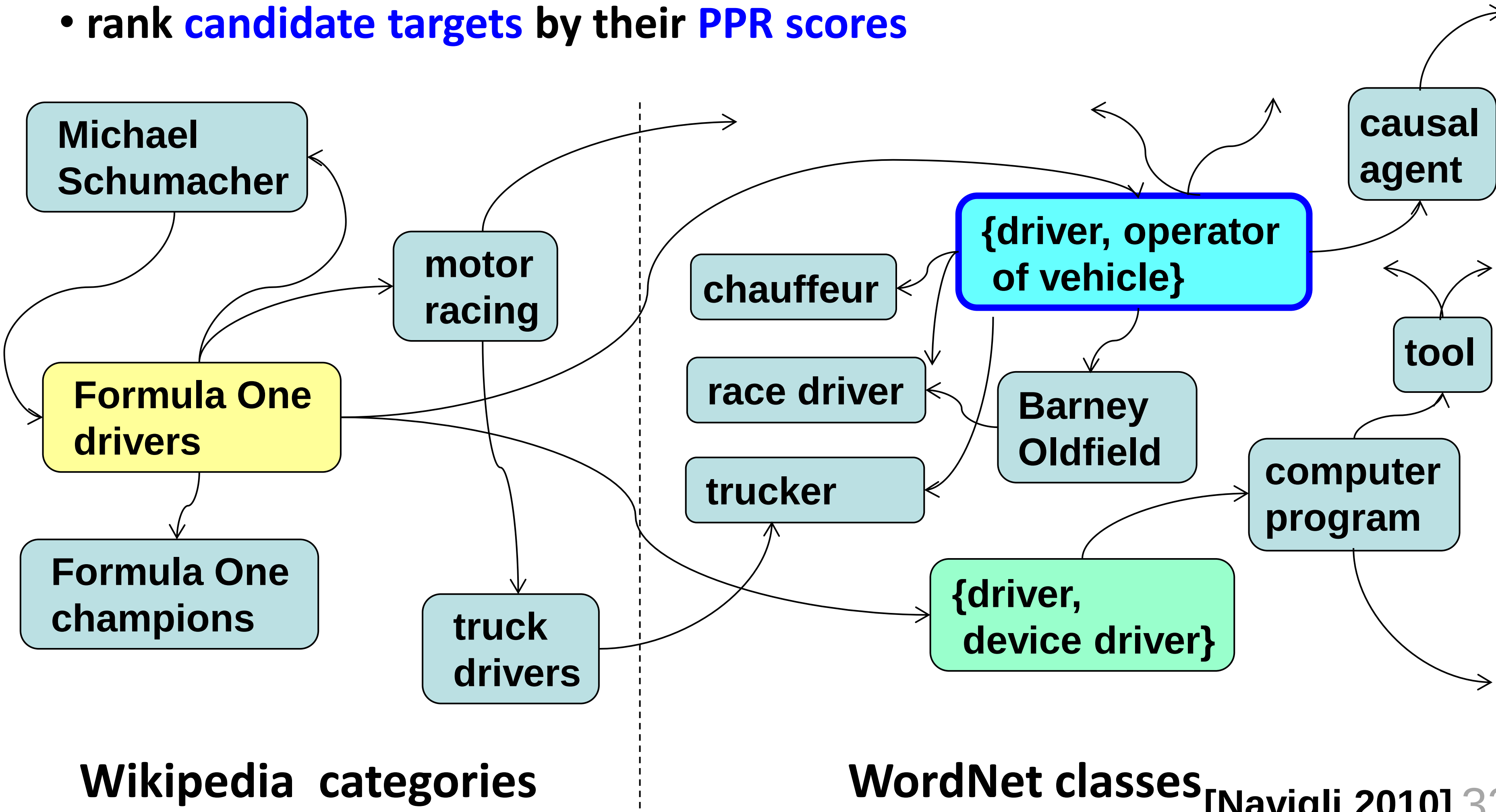
organism
↑ subclassOf
person
↑ subclassOf
American people of Syrian descent
↑ type
Steve Jobs

WordNet

Wikipedia

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]

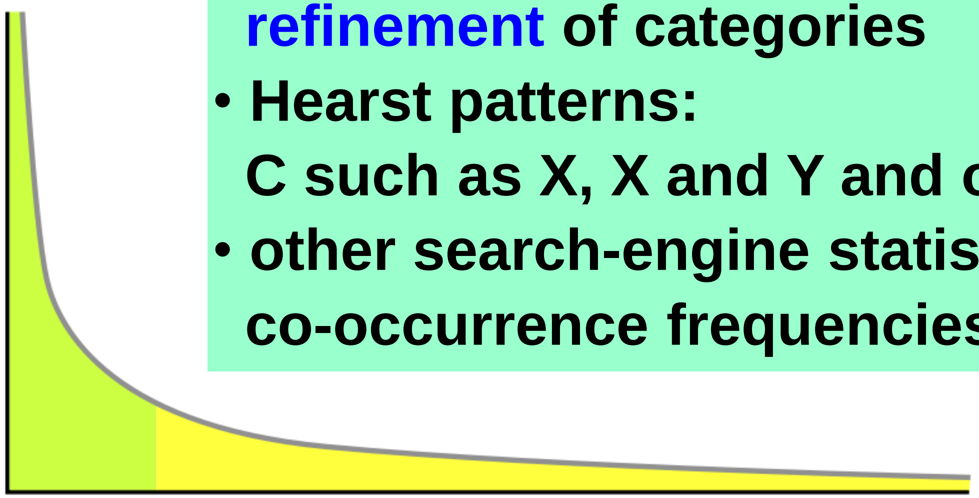
Learning More Mappings [Wu & Weld: WWW‘08]

Kylin Ontology Generator (KOG):

learn classifier for subclassOf across Wikipedia & WordNet using

- YAGO as training data
- advanced ML methods (SVM‘s, MLN‘s)
- rich features from various sources

- category/class name similarity measures
- category instances and their infobox templates: template names, attribute names (e.g. knownFor)
- Wikipedia edit history: refinement of categories
- Hearst patterns: C such as X, X and Y and other C‘s, …
- other search-engine statistics: co-occurrence frequencies



> 3 Mio. entities
> 1 Mio. w/ infoboxes
> 500 000 categories

Outline

✓ **Motivation and Overview**

★ **Taxonomic Knowledge:**
Entities and Classes

- ✓ **Basics & Goal**
- ✓ **Wikipedia-centric Methods**
- ✶ **Web-based Methods**

★ **Factual Knowledge:**
Relations between Entities

★ **Emerging Knowledge:**
New Entities & Relations

★ **Temporal Knowledge:**
Validity Times of Facts

★ **Contextual Knowledge:**
Entity Name Disambiguation

★ **Linked Knowledge:**
Entity Matching

★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

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 → ✗

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] $\rightarrow$ 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) ~ freq(X,Y) * #distinctPatterns(X,Y)

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

$\rightarrow$ score2(X) ~ $(\prod_{i=1..N}$ 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) ~ $(\prod_{i=1..N}$ 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 and Overview**

✓ **Taxonomic Knowledge:**
Entities and Classes

★ **Factual Knowledge:**
Relations between Entities

- **Scope & Goal**
- **Regex-based Extraction**
- **Pattern-based Harvesting**
- **Consistency Reasoning**
- **Probabilistic Methods**
- **Web-Table Methods**

★ **Emerging Knowledge:**
New Entities & Relations

★ **Temporal Knowledge:**
Validity Times of Facts

★ **Contextual Knowledge:**
Entity Name Disambiguation

★ **Linked Knowledge:**
Entity Matching

★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

We focus on given binary relations

Given binary relations with type signature

- hasAdvisor: Person × Person
- graduatedAt: Person × University
- hasWonPrize: Person × Award
- bornOn: Person × Date

...find instances of these relations

- hasAdvisor (JimGray, MikeHarrison)
- hasAdvisor (HectorGarcia-Molina, Gio Wiederhold)
- hasAdvisor (Susan Davidson, Hector Garcia-Molina)
- graduatedAt (JimGray, Berkeley)
- graduatedAt (HectorGarcia-Molina, Stanford)
- hasWonPrize (JimGray, TuringAward)
- bornOn (JohnLennon, 9-Oct-1940)

IE can tap into different sources

Information Extraction (IE) from:

- **Semi-structured data**

  "Low-Hanging Fruit"
  - Wikipedia infoboxes & categories
  - HTML lists & tables, etc.
- **Free text**

  "Cherrypicking"
  - Hearst patterns & other shallow NLP
  - Iterative pattern-based harvesting
  - Consistency reasoning
- **Web tables**

Source-centric IE vs. Yield-centric IE

Source-centric IE

Surajit obtained his PhD in CS from Stanford ...

one source

1) recall !

2) precision

Document 1:

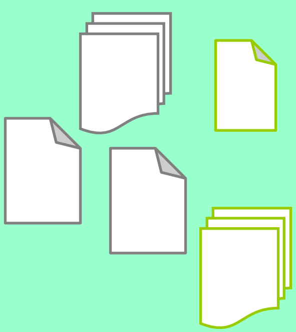
instanceOf (Surajit, scientist)

inField (Surajit, c.science)

almaMater (Surajit, Stanford U)

…

Yield-centric IE



many sources

+ (optional) targeted relations

1) precision !

2) recall

hasAdvisor

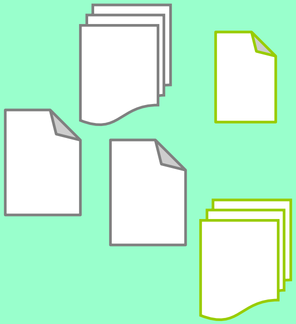
Student	Advisor
Surajit Chaudhuri	Jeffrey Ullman
Jim Gray	Mike Harrison
…	…

worksAt

Student	University
Surajit Chaudhuri	Stanford U
Jim Gray	UC Berkeley
…	…

We focus on yield-centric IE

Yield-centric IE



many sources

+ (optional) targeted relations

1) precision !

2) recall

hasAdvisor

Student	Advisor
Surajit Chaudhuri	Jeffrey Ullman
Jim Gray	Mike Harrison
…	…

worksAt

Student	University
Surajit Chaudhuri	Stanford U
Jim Gray	UC Berkeley
…	…

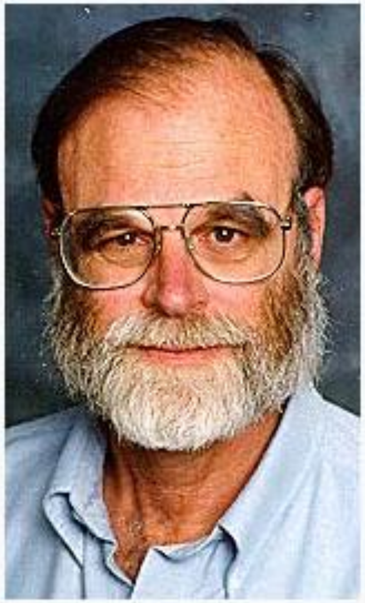
Outline

- ✓ **Motivation and Overview**
- ✓ **Taxonomic Knowledge:**
  Entities and Classes
- ★ **Factual Knowledge:**
  Relations between Entities
  - ✓ **Scope & Goal**
  - ✶ **Regex-based Extraction**
  - ✶ **Pattern-based Harvesting**
  - ✶ **Consistency Reasoning**
  - ✶ **Probabilistic Methods**
  - ✶ **Web-Table Methods**
- ★ **Emerging Knowledge:**
  New Entities & Relations
- ★ **Temporal Knowledge:**
  Validity Times of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Wikipedia provides data in infoboxes

James Nicholas "Jim" Gray



Born	January 12, 1944[1] San Francisco, California[2]
Died	(lost at sea) January 28, 2007
Nationality	American
Fields	Computer Science
Institutions	IBM, Tandem Computers, DEC, Microsoft
Alma mater	University of California, Berkeley
Doctoral advisor	Michael Harrison[2]
Known for	Work on database and transaction processing systems
Notable awards	Turing Award

Barbara Liskov

Born	1939 (age 70–71)
Nationality	American
Fields	Computer Science
Institutions	Massachusetts Institute of Technology
Alma mater	University of California, Berkeley Stanford University
Doctoral advisor	John McCarthy[1]
Notable awards	IEEE John von Neumann Medal, A. M. Turing Award

Serge Abiteboul

Citizenship	French
Nationality	French
Fields	Computer Science
Institutions	INRIA
Alma mater	University of Southern California
Doctoral	

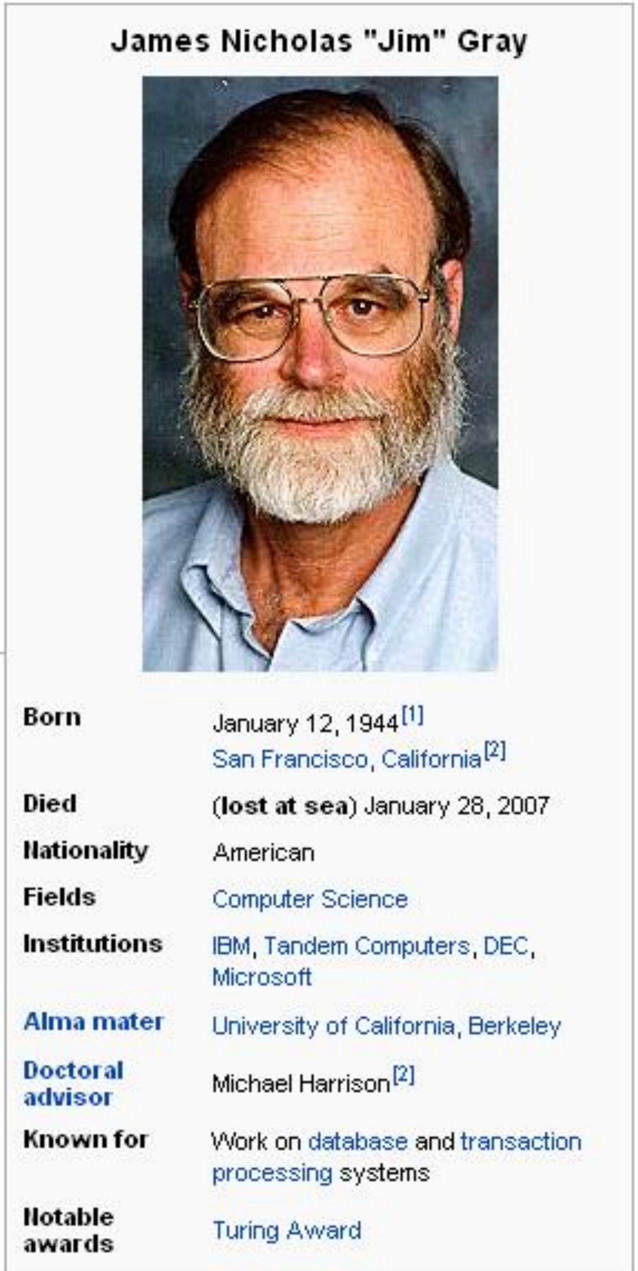
Joseph M. Hellerstein

Fields	Computer Science
Institutions	University of California, Berkeley
Alma mater	University of Wisconsin–Madison
Doctoral advisor	Jeffrey Naughton, Michael Stonebraker

Jeffrey Ullman

Born	November 22, 1942 (age 67)
Citizenship	American
Nationality	American
Alma mater	Columbia University, Princeton University
Doctoral advisor	Arthur Bernstein, Archie McKellar
Doctoral students	Alexander Birman, Surajit Chaudhuri, Evan Cohn, Alan Demers, Marcia Derr, Nahed El Djabri, Amelia Fong Lochovsky, Deepak Goyal, Ashish Gupta, Himanshu Gupta, Udaiprakash Gupta, Venkatesh Harinarayan, Taher Haveliwala, Matthew Hecht, Daniel Hirschberg, Peter Hochschild, Peter Honeyman, Edward Horvath, Gregory Hunter, Nam (Pierre) Huyn, Hakan Jakobsson, John Kam, Marc

Wikipedia uses a Markup Language




```
{{Infobox scientist
| name           = James Nicholas "Jim" Gray
| birth_date     = {{birth date|1944|1|12}}
| birth_place  = [[San Francisco, California]]
| death_date  = ('''lost at sea''')
        {{death date|2007|1|28|1944|1|12}}
| nationality   = American
| field              = [[Computer Science]]
| alma_mater = [[University of California,
                           Berkeley]]
| advisor         = Michael Harrison
...
```

Infoboxes are harvested by RegEx

```
{{Infobox scientist
| name          = James Nicholas "Jim" Gray
| birth_date    = {{birth date|1944|1|12}}
```

Use regular expressions

- **to detect dates**

  `\{\{birth date \|(\d+)\|(\d+)\|(\d+)\}\}`

- **to detect links**

  `\[\[([^\|\]]+)`

- **to detect numeric expressions**

  `(\d+)(\.\d+)?(in|inches|")`

Infoboxes are harvested by RegEx

```
{{Infobox scientist
| name          = James Nicholas "Jim" Gray
| birth_date    = {{birth date|1944|1|12}}
```

Map attribute to canoncial, predefined relation (manually or crowd-sourced)

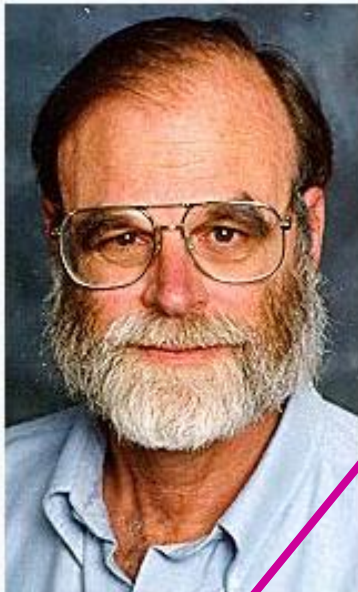
Extract data item by regular expression

wasBorn   1944-01-12

wasBorn(Jim_Gray, "1944-01-12")

Learn how articles express facts

James Nicholas "Jim" Gray



Born	January 12, 1944[1] San Francisco, California[2]
Died	(**lost at sea**) January 28, 2007
Nationality	American
Fields	Computer Science
Institutions	IBM, Tandem Computers, DEC, Microsoft
Alma mater	University of California, Berkeley
Doctoral advisor	Michael Harrison[2]
Known for	Work on database and transaction processing systems
Notable awards	Turing Award

James "Jim" Gray (born January 12, 1944

find attribute value in full text

learn pattern

XYZ (born MONTH DAY, YEAR

Extract from articles w/o infobox

Name: R.Agrawal
Birth date: ?

Rakesh Agrawal (born April 31, 1965) ...

propose attribute value...

apply pattern

XYZ (born MONTH DAY, YEAR

... and/or build fact

bornOnDate(R.Agrawal,1965-04-31)

[Wu et al. 2008: "KYLIN"]

Use CRF to express patterns

$\vec{x}$ = James "Jim" Gray (born January 12, 1944

$\vec{x}$ = James "Jim" Gray (born in January, 1944

$\vec{y}$ = OTH OTH OTH OTH OTH VAL VAL

$$P(\vec{Y} = \vec{y} | \vec{X} = \vec{x}) = \frac{1}{Z} \exp \sum_t \sum_k w_k f_k(y_{t-1}, y_t, \vec{x}, t)$$

Features can take into account

- token types (numeric, capitalization, etc.)
- word windows preceding and following position
- deep-parsing dependencies
- first sentence of article
- membership in relation-specific lexicons

[R. Hoffmann et al. 2010: "Learning 5000 Relational Extractors]

Outline

- ✓ **Motivation and Overview**
- ✓ **Taxonomic Knowledge:**
  Entities and Classes
- ★ **Factual Knowledge:**
  Relations between Entities
  - ✓ Scope & Goal
  - ✓ Regex-based Extraction
  - ✶ Pattern-based Harvesting
  - ✶ Consistency Reasoning
  - ✶ Probabilistic Methods
  - ✶ Web-Table Methods
- ★ **Emerging Knowledge:**
  New Entities & Relations
- ★ **Temporal Knowledge:**
  Validity Times of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Facts yield patterns – and vice versa

<u>Facts</u> & ***<u>Fact Candidates</u>***

(JimGray, MikeHarrison)

(BarbaraLiskov, JohnMcCarthy)

(Surajit, Jeff)
(Alon, Jeff)
(Sunita, Mike)
(Renee, Yannis)

(Sunita, Soumen)
(Soumen, Sunita)
(Surajit, Moshe)
(Alon, Larry)
(Surajit, Microsoft)

<u>Patterns</u>

X and his advisor Y

X under the guidance of Y

X and Y in their paper

X co-authored with Y

X rarely met his advisor Y

…

- **good for recall**
- **noisy, drifting**
- **not robust enough for high precision**

Statistics yield pattern assessment

Support of pattern p:

$$\frac{\#\ \text{occurrences of p with seeds (e1,e2)}}{\#\ \text{occurrences of all patterns with seeds}}$$

Confidence of pattern p:

$$\frac{\#\ \text{occurrences of p with seeds (e1,e2)}}{\#\ \text{occurrences of p}}$$

Confidence of fact candidate (e1,e2):

$$\Sigma_p\ \text{freq(e1,p,e2)*conf(p)} \ /\ \Sigma_p\ \text{freq(e1,p,e2)}$$

$$\text{or: PMI (e1,e2)} = \log \frac{\text{freq(e1,e2)}}{\text{freq(e1) freq(e2)}}$$

- **gathering can be iterated,**
- **can promote best facts to additional seeds for next round**

Negative Seeds increase precision

(Ravichandran 2002; Suchanek 2006; ...)

Problem: Some patterns have high support, but poor precision:

X is the largest city of Y for isCapitalOf (X,Y)
joint work of X and Y for hasAdvisor (X,Y)

Idea: Use positive and negative seeds:

pos. seeds: (Paris, France), (Rome, Italy), (New Delhi, India), ...
neg. seeds: (Sydney, Australia), (Istanbul, Turkey), ...

Compute the confidence of a pattern as:

$$\frac{\text{\# occurrences of p with pos. seeds}}{\text{\# occurrences of p with pos. seeds or neg. seeds}}$$

- can promote best facts to additional seeds for next round
- can promote rejected facts to additional counter-seeds
- works more robustly with few seeds & counter-seeds

Generalized patterns increase recall

(N. Nakashole 2011)

Problem: Some patterns are too narrow and thus have small recall:

X and his celebrated advisor Y
X carried out his doctoral research in math under the supervision of Y
X received his PhD degree in the CS dept at Y
X obtained his PhD degree in math at Y

Idea: generalize patterns to n-grams, allow POS tags

Compute n-gram-sets by frequent sequence mining

X { his doctoral research, under the supervision of} Y
X { PRP ADJ advisor } Y
X { PRP doctoral research, IN DET supervision of} Y

Compute match quality of pattern p with sentence q by Jaccard:

$$\frac{|\{\text{n-grams} \in p\} \cap \{\text{n-grams} \in q\}|}{|\{\text{n-grams} \in p\} \cup \{\text{n-grams} \in q\}|}$$

=> Covers more sentences, increases recall

Deep Parsing makes patterns robust

(Bunescu 2005 , Suchanek 2006, …)

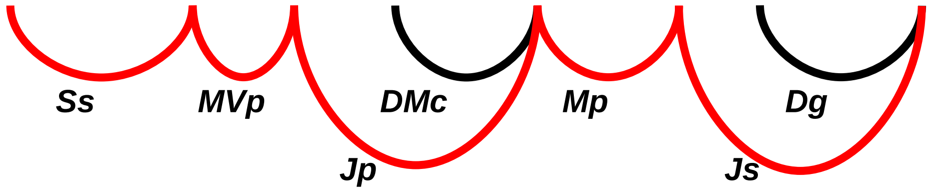
Problem: Surface patterns fail if the text shows variations

Cologne lies on the banks of the Rhine.

Paris, the French capital, lies on the beautiful banks of the Seine

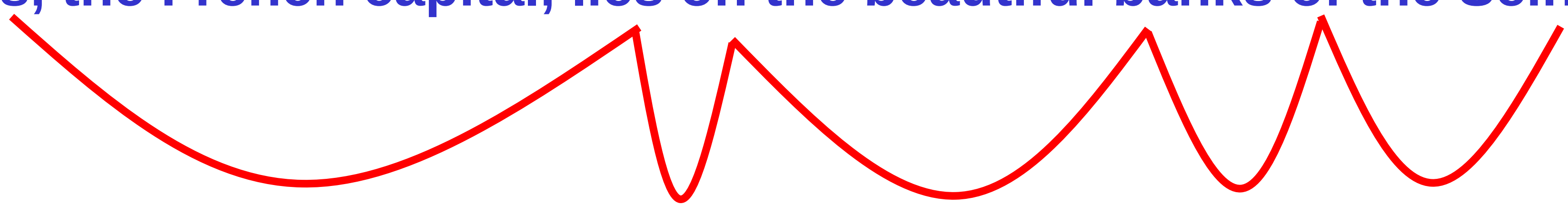
Idea: Use deep linguistic parsing to define patterns

Cologne lies on the banks of the Rhine



Deep linguistic patterns work even on sentences with variations

Paris, the French capital, lies on the beautiful banks of the Seine



Outline

✓ **Motivation and Overview**

✓ **Taxonomic Knowledge:**
Entities and Classes

★ **Factual Knowledge:**
Relations between Entities

- ✓ **Scope & Goal**
- ✓ **Regex-based Extraction**
- ✓ **Pattern-based Harvesting**
- ✶ **Consistency Reasoning**
- ✶ **Probabilistic Methods**
- ✶ **Web-Table Methods**

★ **Emerging Knowledge:**
New Entities & Relations

★ **Temporal Knowledge:**
Validity Times of Facts

★ **Contextual Knowledge:**
Entity Name Disambiguation

★ **Linked Knowledge:**
Entity Matching

★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Extending a KB faces 3+ challenges

(F. Suchanek et al.: WWW'09)

Problem: If we want to extend a KB, we face (at least) 3 challenges

1. Understand which relations are expressed by patterns

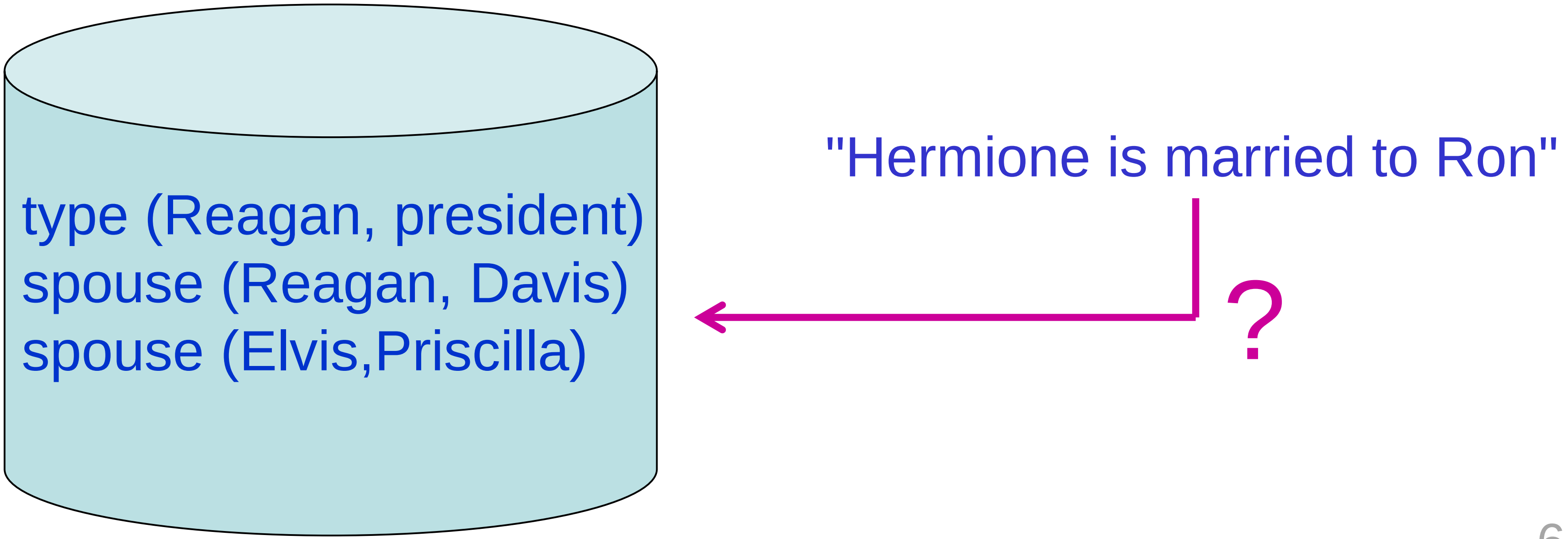
"x is married to y" ~ spouse(x,y)

2. Disambiguate entities

"Hermione is married to Ron": "Ron" = RonaldReagan?

3. Resolve inconsistencies

spouse(Hermione, Reagan) & spouse(Reagan,Davis) ?



SOFIE transforms IE to logical rules

(F. Suchanek et al.: WWW'09)

Idea: Transform corpus to surface statements

"Hermione is married to Ron"

occurs("Hermione", "is married to", "Ron")

Add possible meanings for all words from the KB

means("Ron", RonaldReagan)

means("Ron", RonWeasley)

Only one of these can be true

means("Hermione", HermioneGranger)

means(X,Y) & means(X,Z) $\Rightarrow$ Y=Z

Add pattern deduction rules

occurs(X,P,Y) & means(X,X') & means(Y,Y') & R(X',Y') $\Rightarrow$ P~R

occurs(X,P,Y) & means(X,X') & means(Y,Y') & P~R $\Rightarrow$ R(X',Y')

Add semantic constraints (manually)

spouse(X,Y) & spouse(X,Z) $\Rightarrow$ Y=Z

The rules deduce meanings of patterns

(F. Suchanek et al.: WWW‘09)

type(Reagan, president)
spouse(Reagan, Davis)
spouse(Elvis,Priscilla)

"Elvis is married to Priscilla"

"is married to“ ~ spouse

Add pattern deduction rules

occurs(X,P,Y) & means(X,X') & means(Y,Y') & R(X',Y') ⇒ P~R

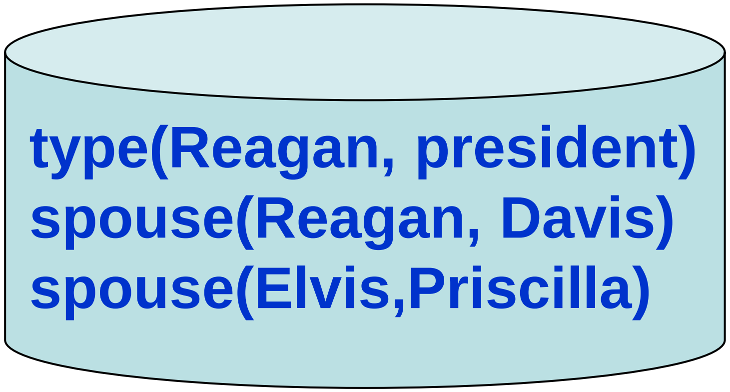
occurs(X,P,Y) & means(X,X') & means(Y,Y') & P~R ⇒ R(X',Y')

Add semantic constraints (manually)

spouse(X,Y) & spouse(X,Z) ⇒ Y=Z

The rules deduce facts from patterns

(F. Suchanek et al.: WWW'09)



"Hermione is married to Ron"

"is married to" ~ married

spouse(Hermione,RonaldReagan)
spouse(Hermione,RonWeasley)

Add pattern deduction rules

occurs(X,P,Y) & means(X,X') & means(Y,Y') & R(X',Y') $\Rightarrow$ P~R

occurs(X,P,Y) & means(X,X') & means(Y,Y') & P~R $\Rightarrow$ R(X',Y')

Add semantic constraints (manually)

spouse(X,Y) & spouse(X,Z) $\Rightarrow$ Y=Z

The rules remove inconsistencies

(F. Suchanek et al.: WWW'09)

type(Reagan, president)
spouse(Reagan, Davis)
spouse(Elvis,Priscilla)

~~spouse(Hermione,RonaldReagan)~~
spouse(Hermione,RonWeasley)

Add pattern deduction rules

occurs(X,P,Y) & means(X,X') & means(Y,Y') & R(X',Y') ⇒ P~R

occurs(X,P,Y) & means(X,X') & means(Y,Y') & P~R ⇒ R(X',Y')

Add semantic constraints (manually)

spouse(X,Y) & spouse(X,Z) ⇒ Y=Z

The rules pose a weighted MaxSat problem

(F. Suchanek et al.: WWW‘09)

type(Reagan, president) [10]
married(Reagan, Davis) [10]
married(Elvis,Priscilla) [10]

We are given a set of rules/facts, and wish to find the most plausible possible world.

spouse(X,Y) & spouse(X,Z) => Y=Z [10]
occurs("Hermione","loves","Harry") [3]
means("Ron",RonaldReagan) [3]
means("Ron",RonaldWeasley) [2]
...

Possible World 1:

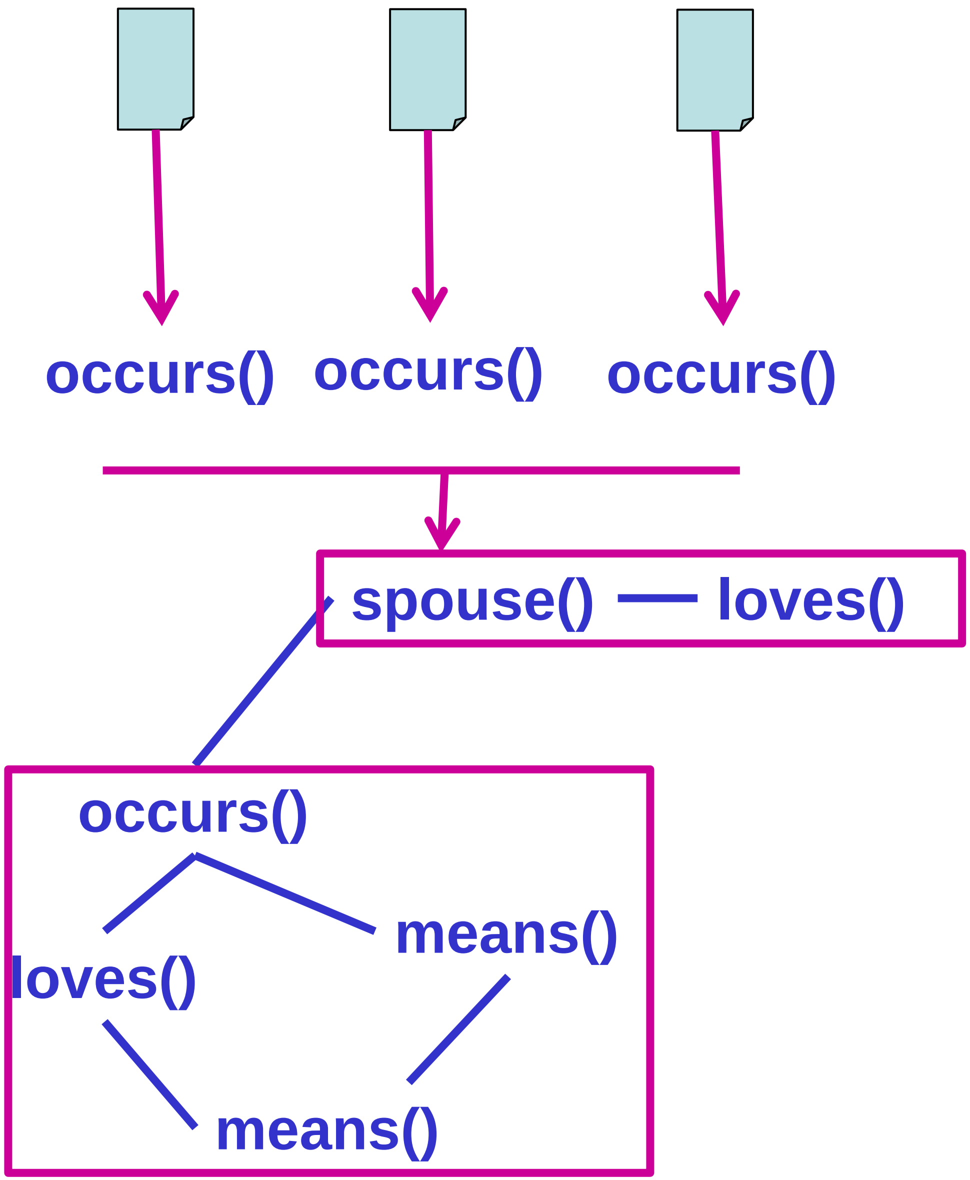
Weight of satisfied rules: 30

Possible World 2:

Weight of satisfied rules: 39

PROSPERA parallelizes the extraction

(N. Nakashole et al.: WSDM‘11)



Mining the pattern occurrences is embarassingly parallel

Reasoning is hard to parallelize as atoms depends on other atoms

Idea: parallelize along min-cuts

Outline

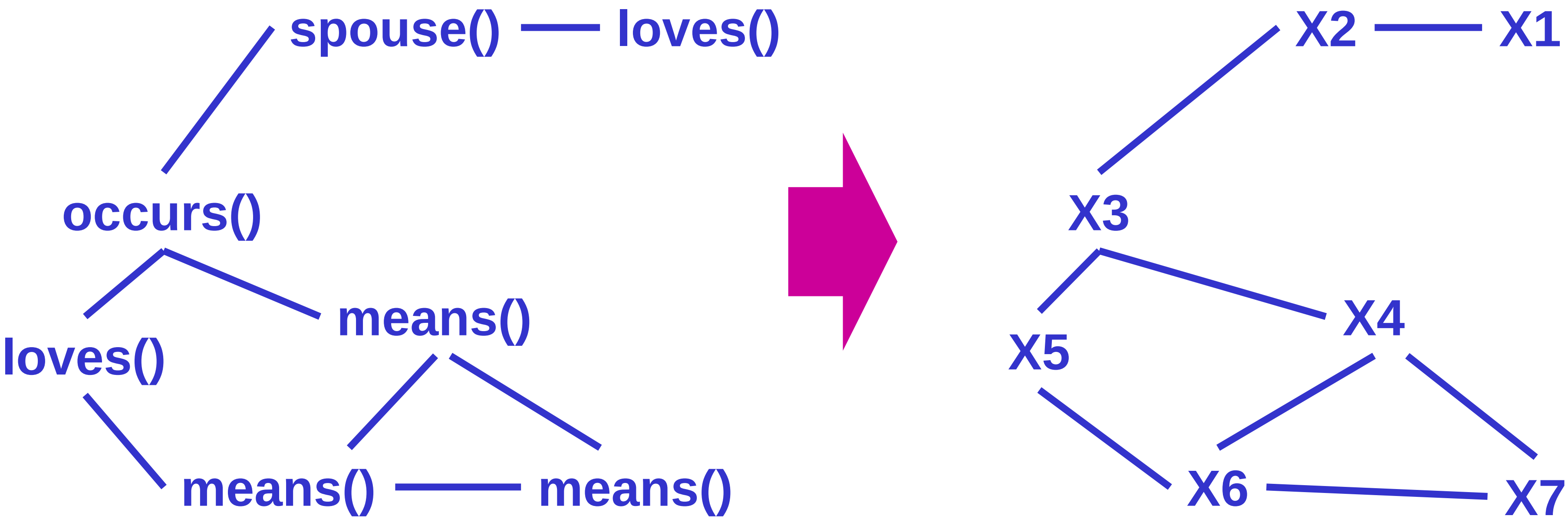
- ✓ **Motivation and Overview**
- ✓ **Taxonomic Knowledge:**
  Entities and Classes
- ★ **Factual Knowledge:**
  Relations between Entities
  - ✓ **Scope & Goal**
  - ✓ **Regex-based Extraction**
  - ✓ **Pattern-based Harvesting**
  - ✓ **Consistency Reasoning**
  - ✶ **Probabilistic Methods**
  - ✶ **Web-Table Methods**
- ★ **Emerging Knowledge:**
  New Entities & Relations
- ★ **Temporal Knowledge:**
  Validity Times of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Markov Logic generalizes MaxSat reasoning

(M. Richardson / P. Domingos 2006)

In a Markov Logic Network (MLN), every atom is represented by a Boolean random variable.



Dependencies in an MLN are limited

The value of a random variable X_i depends only on its neighbors:

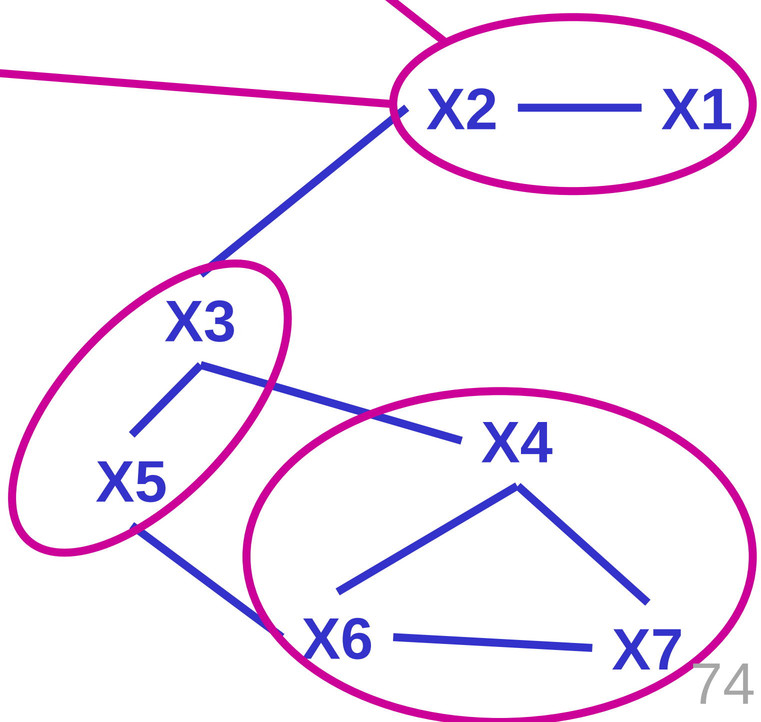
$$P(X_i|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n) = P(X_i|N(X_i))$$

The Hammersley-Clifford Theorem tells us:

$$P(\vec{X}=\vec{x}) = \frac{1}{Z}\prod \varphi_i(\pi_{Ci}(\vec{x}))$$

We choose φ_i so as to satisfy all formulas in the the i-th clique:

$$\varphi_i(\vec{z}) =$$
$$\exp(w_i \times [formulas\ i\ sat.with\ \vec{z}])$$



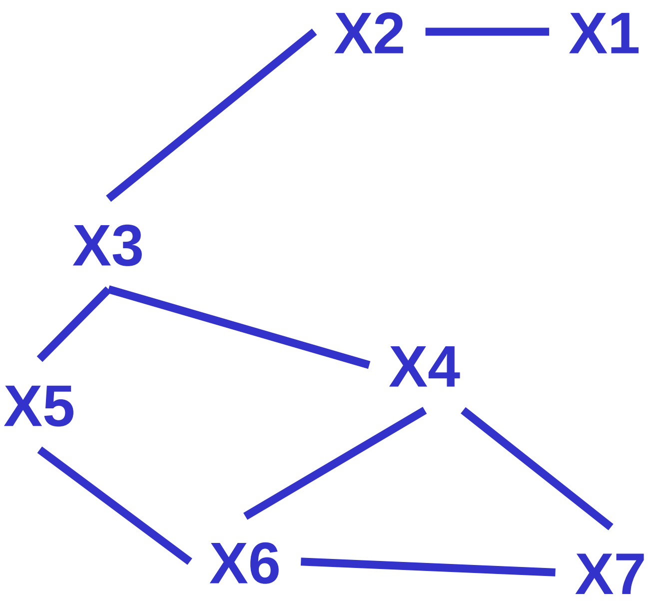
There are many methods for MLN inference

To compute the values that maximize the joint probability (MAP = maximum a posteriori) we can use a variety of methods:

Gibbs sampling, other MCMC, belief propagation, randomized MaxSat, …

In addition, the MLN can model/compute

- **marginal probabilities**
- **the joint distribution**



Large-Scale Fact Extraction with MLNs

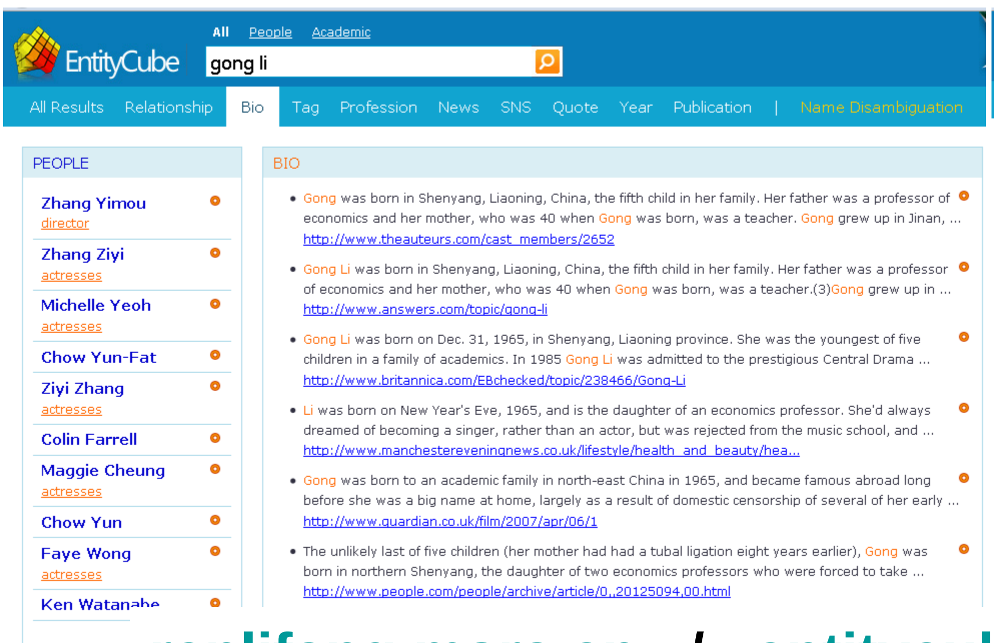
[J. Zhu et al.: WWW'09]

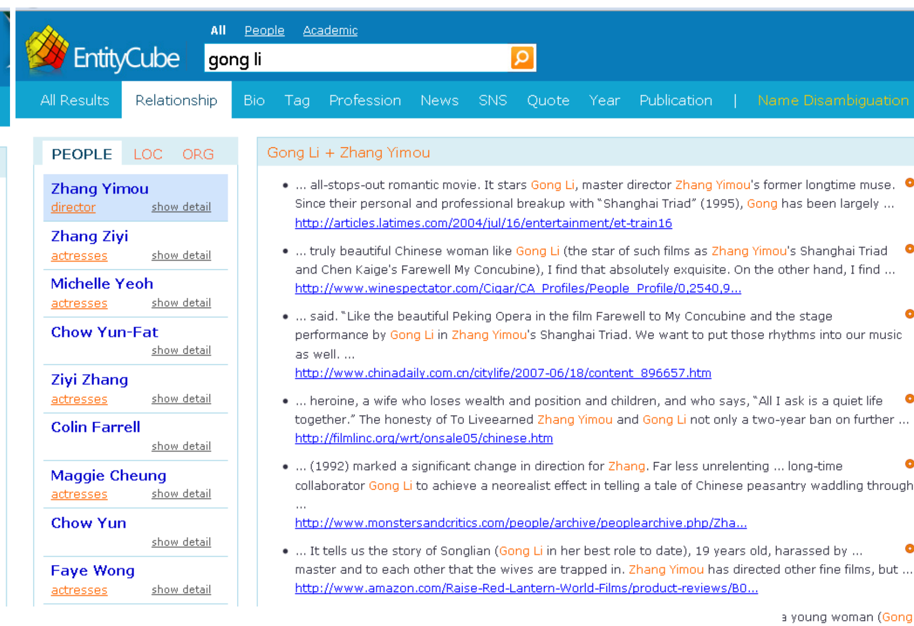
StatSnowball:

- start with seed facts and initial MLN model
- iterate:
  - extract facts
  - generate and select patterns
  - refine and re-train MLN model (plus CRFs plus …)

BioSnowball:

- automatically creating biographical summaries

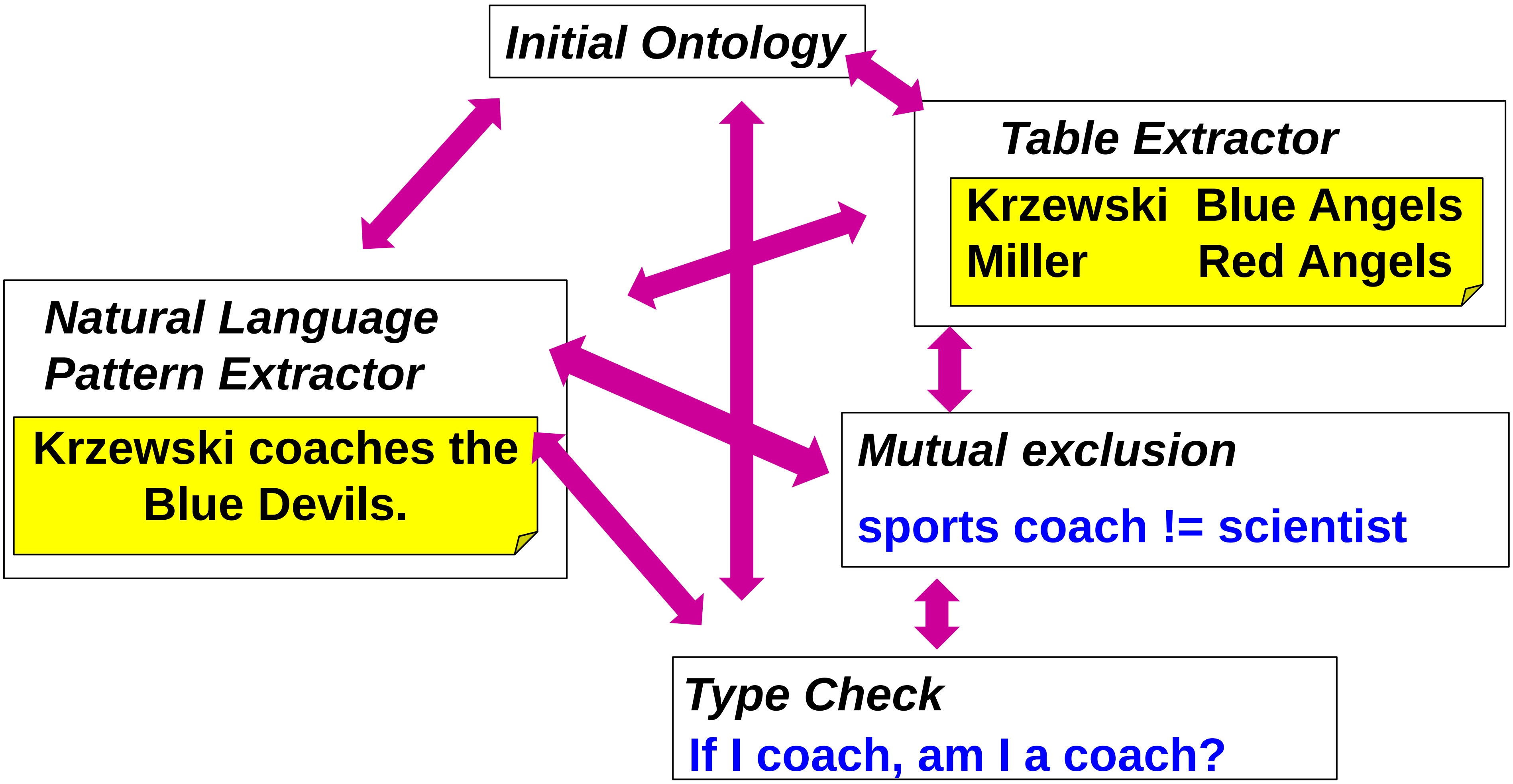




renlifang.msra.cn / entitycube.research.microsoft.com

NELL couples different learners

[Carlson et al. 2010]



http://rtw.ml.cmu.edu/rtw/

Outline

- ✓ **Motivation and Overview**
- ✓ **Taxonomic Knowledge:**
  Entities and Classes
- ★ **Factual Knowledge:**
  Relations between Entities
  - ✓ **Scope & Goal**
  - ✓ **Regex-based Extraction**
  - ✓ **Pattern-based Harvesting**
  - ✓ **Consistency Reasoning**
  - ✓ **Probabilistic Methods**
  - ★ **Web-Table Methods**
- ★ **Emerging Knowledge:**
  New Entities & Relations
- ★ **Temporal Knowledge:**
  Validity Times of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Web Tables provide relational information

[Cafarella et al: PVLDB 08; Sarawagi et al: PVLDB 09]

Academy Awards

(Reference:[1])

Year	Nominated work	Category	Result
1978	*The Deer Hunter*	Best Supporting Actress	Nominated
1979	*Kramer vs. Kramer*	Best Supporting Actress	Won
1981	*The*		
1982			

Academy Awards

Year	Category	Film	Result
	Academy Award for Best Actor	*Sweeney Todd: The Demon Barber of Fleet Street*	Nominated
	Academy Award for Best Actor	*Finding Neverland*	Nominated
	Academy Award for Best Actor	*Pirates of the Caribbean: The Curse of the Black Pearl*	Nominated

Academy Awards

Winner

- Best Art Direction
- Best Cinematography
- Best Makeup

Nominated

- Best Original Score
- Best Original Screenplay
- Best Foreign Language Film

Year	Winner Composer	Nominees
2000	***Crouching Tiger, Hidden Dragon*** – Tan Dun	• *Chocolat* – Rachel Portman • *Gladiator* – Hans Zimmer [3] • *Malèna* – Ennio Morricone • *The Patriot* – John Williams

Year	Image	Recipient	Category	Film
2010		Sandra Bullock	Worst Actress Worst Screen Couple	*All About Steve*

Academy Awards (2009): Nominees and Winners

NOMINATIONS		AWARDS	
9	**Avatar**	6	**The Hurt Locker**
9	**The Hurt Locker**	3	**Avatar**
8	**Inglourious Basterds**	2	Crazy Heart
6	**Precious**	2	**Precious**
6	**Up in the Air**	2	**Up**
5	**Up**	1	**The Blind Side**
4	**District 9**	1	The Cove
4	Nine	1	**Inglourious Basterds**
4	Star Trek	1	Logorama

Web Tables can be annotated with YAGO

[Limaye, Sarawagi, Chakrabarti: PVLDB 10]

Goal: enable semantic search over Web tables

Idea:

- **Map column headers to Yago classes,**
- **Map cell values to Yago entities**
- **Using joint inference for factor-graph learning model**

Title	Author
Hitchhiker's guide	D Adams
A short history of time	S Hawkins

Statistics yield semantics of Web tables

Conference	City	
description	**location**	**deadline**
Third Workshop on Large-scale Data Mining: Theory and Applications (LDMTA 2011)	San Diego, CA, USA	May 21st, 2011
Mining Data Semantics (MDS2011) Workshop	San Diego, CA, USA	May 10th, 2011

Idea: Infer classes from co-occurrences, headers are class names

$$P(class|val_1, \ldots, val_n) = \prod \frac{P(class|val_i)}{P(class)}$$

Result from 12 Mio. Web tables:

- **1.5 Mio. labeled columns (=classes)**
- **155 Mio. instances (=values)**

[Venetis,Halevy et al: PVLDB 11]

Statistics yield semantics of Web tables

description	location	deadline
Third Workshop on Large-scale Data Mining: Theory and Applications (LDMTA 2011)	San Diego, CA, USA	May 21st, 2011
Mining Data Semantics (MDS2011) Workshop	San Diego, CA, USA	May 10th, 2011

Idea: Infer facts from table rows, header identifies relation name

hasLocation(ThirdWorkshop, SanDiego)

but: classes&entities not canonicalized. Instances may include:
Google Inc., Google, NASDAQ GOOG, Google search engine, …
Jet Li, Li Lianjie, Ley Lin Git, Li Yangzhong, Nameless hero, …

Take-Home Lessons

Bootstrapping works well for recall

but details matter: seeds, counter-seeds, pattern language, statistical confidence, etc.

For high precision, consistency reasoning is crucial:

various methods incl. MaxSat, MLN/factor-graph MCMC, etc.

Harness initial KB for distant supervision & efficiency:

seeds from KB, canonicalized entities with type contraints

Hand-crafted domain models are assets:

expressive constraints are vital, modeling is not a bottleneck, but no out-of-model discovery

Open Problems and Grand Challenges

Robust fact extraction with both high precision & recall
as highly automated (self-tuning) as possible

Efficiency and scalability of best methods for (probabilistic) reasoning without losing accuracy

Extensions to ternary & higher-arity relations
events in context: who did what to/with whom when where why …?

Large-scale studies for vertical domains
e.g. academia: researchers, publications, organizations, collaborations, projects, funding, software, datasets, …

Real-time & incremental fact extraction for continuous KB growth & maintenance
(life-cycle management over years and decades)

Outline

- ✓ **Motivation and Overview**
- ★ **Taxonomic Knowledge:**
  Entities and Classes
- ★ **Factual Knowledge:**
  Relations between Entities
- ★ **Emerging Knowledge:**
  New Entities & Relations
  - ✶ **Open Information Extraction**
  - ✶ **Relation Paraphrases**
  - ✶ **Big Data Algorithms**
- ★ **Temporal Knowledge:**
  Validity Times of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

Big Data Methods for

Knowledge for Big Data Analytics

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Discovering "Unknown" Knowledge

so far KB has relations with type signatures
<entity1, relation, entity2>

< CarlaBruni marriedTo NicolasSarkozy> $\in$ Person $\times$ R $\times$ Person

< NataliePortman wonAward AcademyAward > $\in$ Person $\times$ R $\times$ Prize

Open and Dynamic Knowledge Harvesting:
would like to discover new entities and new relation types
<name1, phrase, name2>

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 …
Bonham Carter was disappointed that her nomination for the Oscar …

Open IE with ReVerb [A. Fader et al. 2011, T. Lin 2012]

Consider all verbal phrases as potential relations and all noun phrases as arguments

Problem 1: incoherent extractions

“New York City has a population of 8 Mio” → <New York City, has, 8 Mio>
“Hero is a movie by Zhang Yimou” → <Hero, is, Zhang Yimou>

Problem 2: uninformative extractions

“Gold has an atomic weight of 196” → <Gold, has, atomic weight>
“Faust made a deal with the devil” → <Faust, made, a deal>

Problem 3: over-specific extractions

“Hero is the most colorful movie by Zhang Yimou”
→ <..., is the most colorful movie by, …>

Solution:

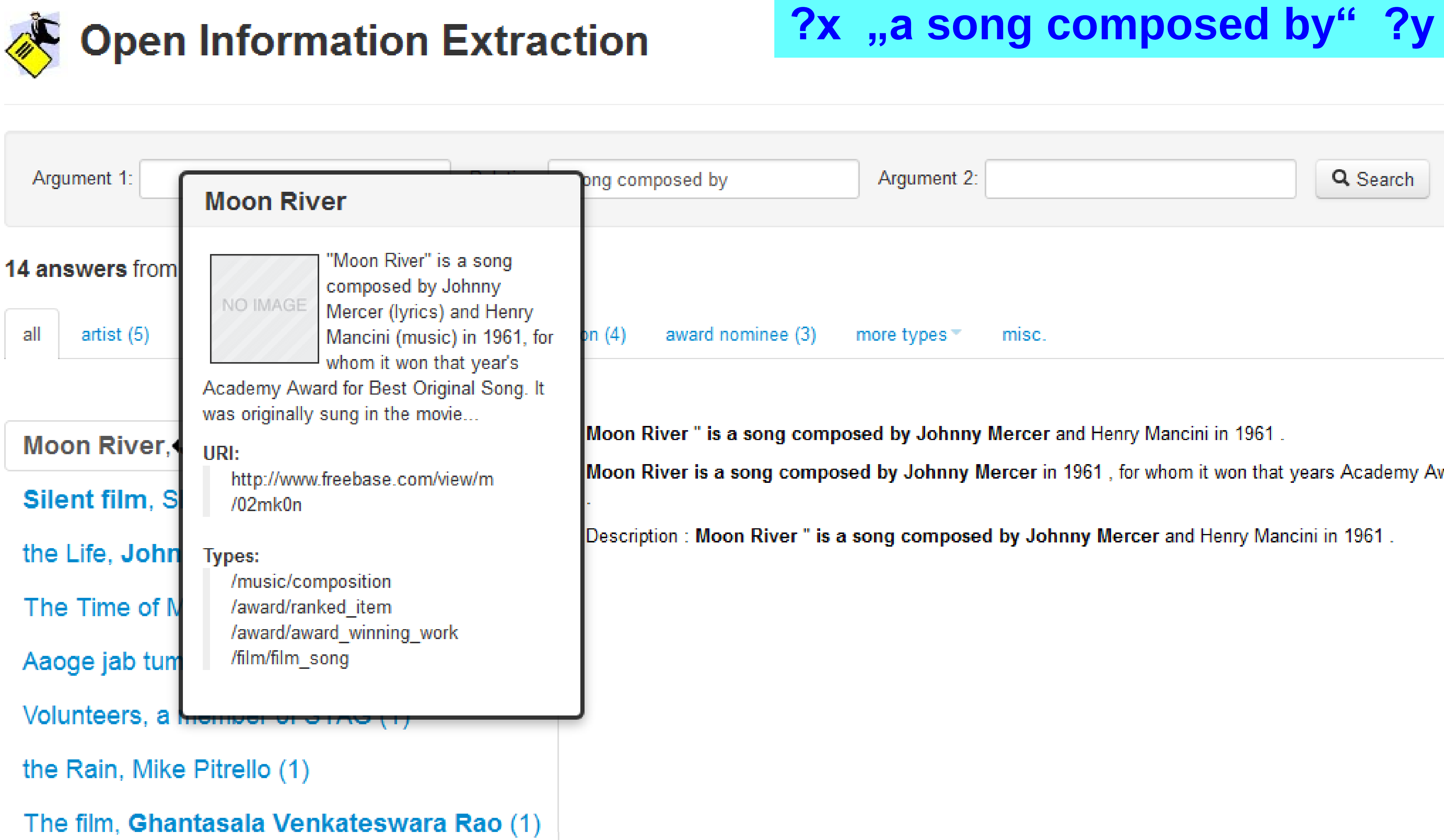
- regular expressions over POS tags:
  VB DET N PREP; VB (N | ADJ | ADV | PRN | DET)* PREP; etc.
- relation phrase must have # distinct arg pairs > threshold

http://ai.cs.washington.edu/demos

Open IE Example: ReVerb

http://openie.cs.washington.edu/

?x „a song composed by“ ?y



Open IE Example: ReVerb

http://openie.cs.washington.edu/

?x „a piece written by“ ?y

Open Information Extraction

Argument 1: | Relation: a piece written by | Argument 2:

13 answers from **14 sentences**

all | author (3) | person (3) | misc.

The link, Bill Maxwell (2)

Secondary sources, someone (1)

The first section, prisoners (1)

the concert, Karl (1)

The real standouts, veterans and others (1)

This website, Charlie (1)

The fun-filled songs, **Bob Dylan** (1)

their parents, Isioma Daniel (1)

Diversity and Ambiguity of Relational Phrases

Who covered whom?

Amy Winehouse's concert included cover songs by the Shangri-Las

Amy's souly interpretation of Cupid, a classic piece of Sam Cooke

Nina Simone's singing of Don't Explain revived Holiday's old song

Cat Power's voice is sad in her version of Don't Explain

16 Horsepower played Sinnerman, a Nina Simone original

Cale performed Hallelujah written by L. Cohen

Cave sang Hallelujah, his own song unrelated to Cohen's

{cover songs, interpretation of, singing of, voice in, …} ⇔ **SingerCoversSong**

{classic piece of, 's old song, written by, composition of, …} ⇔ **MusicianCreatesSong**

Scalable Mining of SOL Patterns

[N. Nakashole et al.: EMNLP-CoNLL’12, VLDB‘12]

Syntactic-Lexical-Ontological (SOL) **patterns**

- **Syntactic-Lexical:** **surface words, wildcards, POS tags**
- **Ontological:** **semantic classes as entity placeholders**
  <singer>, <musician>, <song>, …
- **Type signature of pattern: <singer> × <song>, <person> × <song>**
- **Support set of pattern: set of entity-pairs for placeholders**
  → support and confidence of patterns

SOL pattern: **<singer> ’s ADJECTIVE voice * in <song>**

Matching sentences:

Amy Winehouse’s soul voice in her song ‘Rehab’
Jim Morrison’s haunting voice and charisma in ‘The End’
Joan Baez’s angel-like voice in ‘Farewell Angelina’

Support set:

(Amy Winehouse, Rehab)
(Jim Morrison, The End)
(Joan Baez, Farewell Angelina)

Pattern Dictionary for Relations

[N. Nakashole et al.: EMNLP-CoNLL'12, VLDB'12]

WordNet-style dictionary/taxonomy for relational phrases based on SOL patterns (syntactic-lexical-ontological)

Relational phrases are typed

<*person*> graduated from <*university*>
<*singer*> covered <*song*> **<*book*> covered <*event*>**

Relational phrases can be synonymous

"graduated from" ⇔ "obtained degree in * from"
"and PRONOUN ADJECTIVE advisor" ⇔ "under the supervision of"

One relational phrase can subsume another

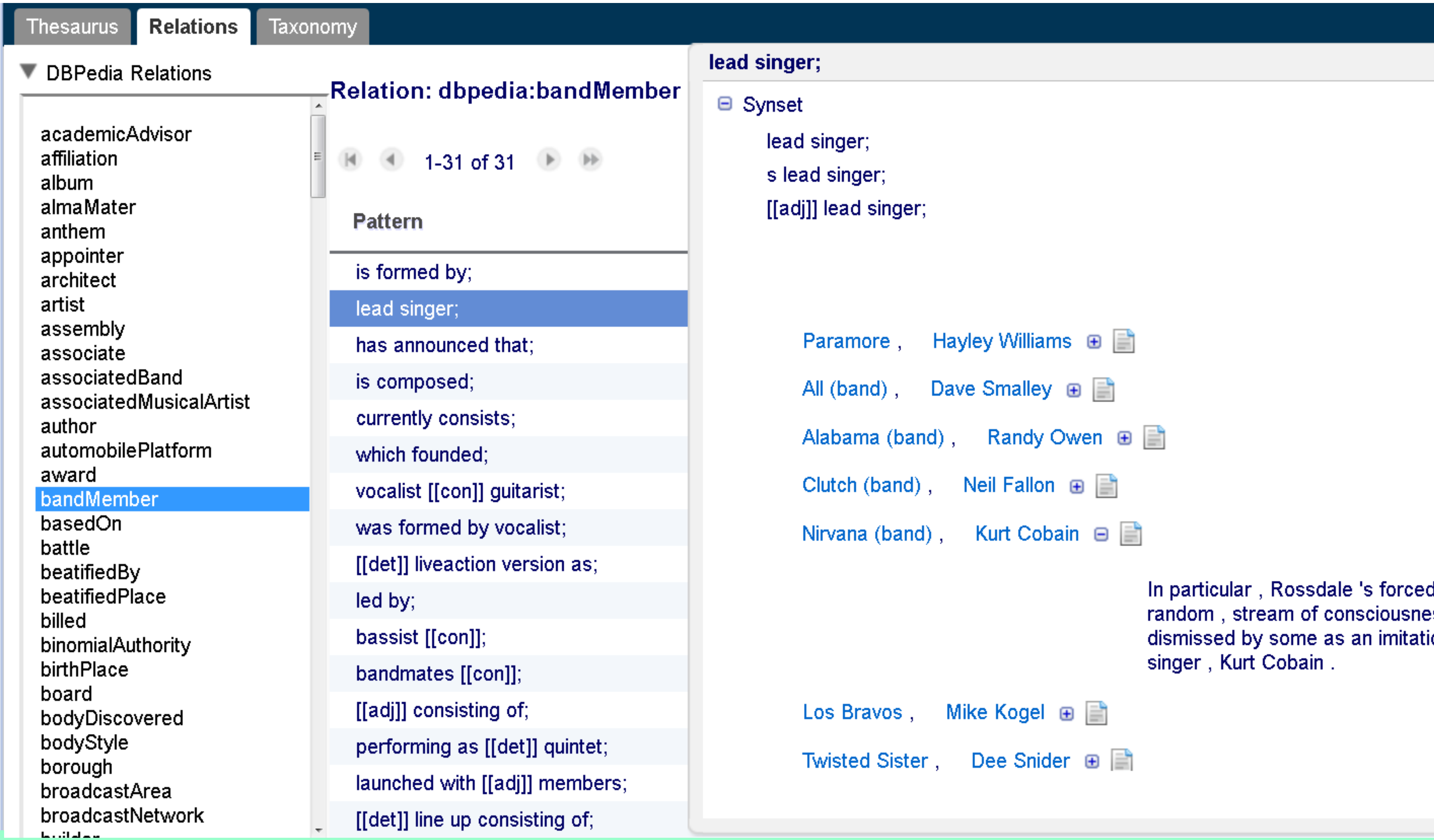
"wife of" ⇒ " spouse of"

350 000 SOL patterns from Wikipedia, NYT archive, ClueWeb

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

PATTY: Pattern Taxonomy for Relations

[N. Nakashole et al.: EMNLP 2012, demo at VLDB 2012]



350 000 SOL patterns with 4 Mio. instances
accessible at: www.mpi-inf.mpg.de/yago-naga/patty

Big Data Algorithms at Work

Frequent sequence mining

with generalization hierarchy for tokens

Examples: famous → ADJECTIVE → *

her → PRONOUN → *

<singer> → <musician> → <artist> → <person>

Map-Reduce-parallelized on Hadoop:

- identify entity-phrase-entity occurrences in corpus
- compute frequent sequences
- repeat for generalizations

text pre-processing → n-gram mining → pattern lifting → taxonomy construction

Take-Home Lessons

Triples of the form <name, phrase, name>
can be mined at scale and
are beneficial for entity discovery

Scalable algorithms for extraction & mining
have been leveraged – but more work needed

Semantic typing of relational patterns
and pattern taxonomies are vital assets

Open Problems and Grand Challenges

Overcoming sparseness in input corpora and coping with even larger scale inputs

tap social media, query logs, web tables & lists, microdata, etc. for richer & cleaner taxonomy of relational patterns

Cost-efficient crowdsourcing for higher coverage & accuracy

Exploit relational patterns for question answering over structured data

Integrate canonicalized KB with emerging knowledge

KB life-cycle: today‘s long tail may be tomorrow‘s mainstream

Outline

- ✓ **Motivation and Overview**
- ★ **Taxonomic Knowledge:**
  Entities and Classes
- ★ **Factual Knowledge:**
  Relations between Entities

- ★ **Emerging Knowledge:**
  New Entities & Relations
- ★ **Temporal Knowledge:**
  Validity Times of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-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

consistency constraints are potentially helpful:

- functional dependencies: *husband, time → wife*
- inclusion dependencies: *marriedPerson ⊆ 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

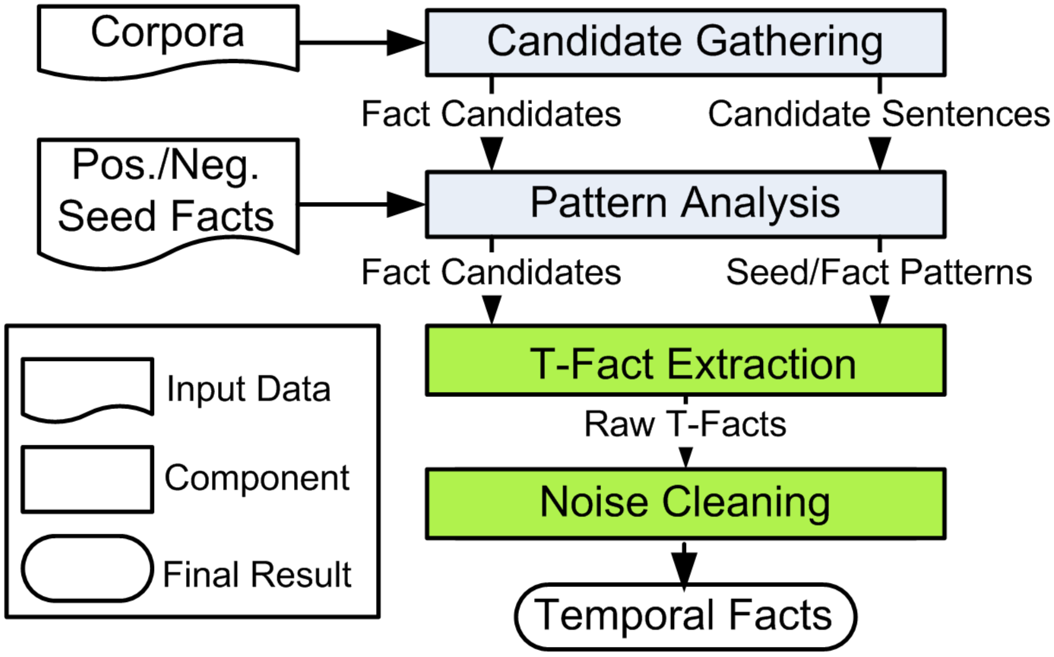
Education

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]

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

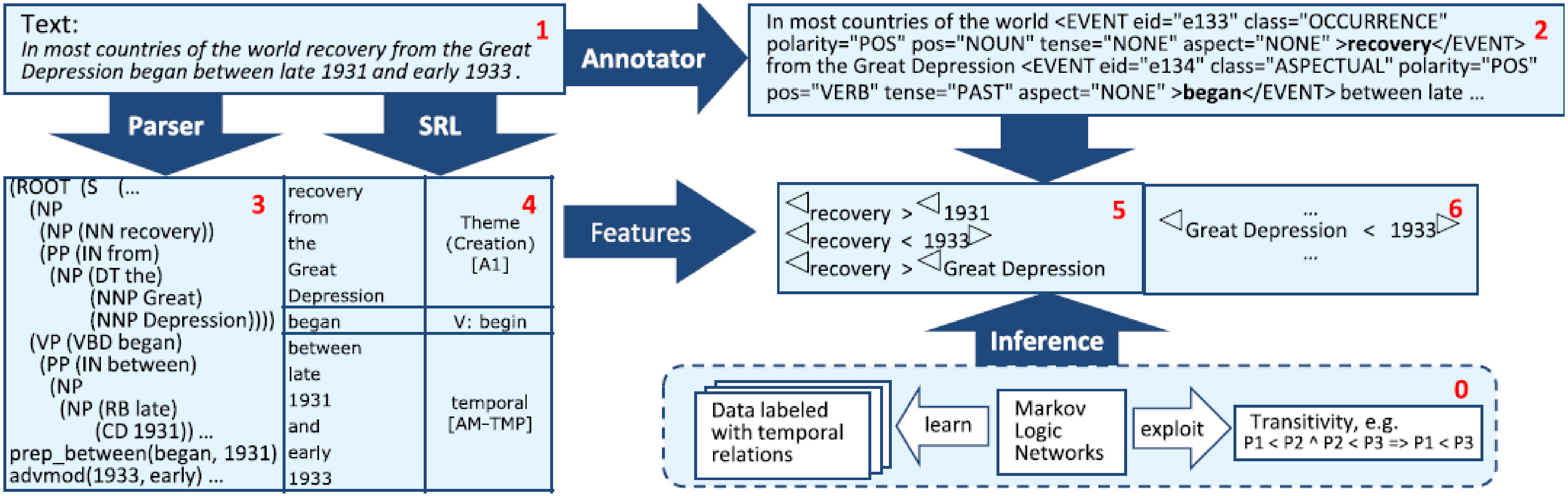
Efficient ILP solvers:
www.gurobi.com
IBM Cplex
…

TIE for T-Fact Extraction & Ordering

[Ling/Weld : AAAI 2010]

TIE (Temporal IE) architectures builds on:

- **TARSQI (Verhagen et al. 2005)**
  for event extraction, using linguistic analyses
- **Markov Logic Networks**
  for temporal ordering of events



Take-Home Lessons

Temporal knowledge harvesting:

crucial for machine-reading news, social media, opinions

Combine linguistics, statistics, and logical reasoning:

harder than for „ordinary" relations

Open Problems and Grand Challenges

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

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

Understand temporal relationships in biographies and narratives

machine-reading of news, bios, novels, …

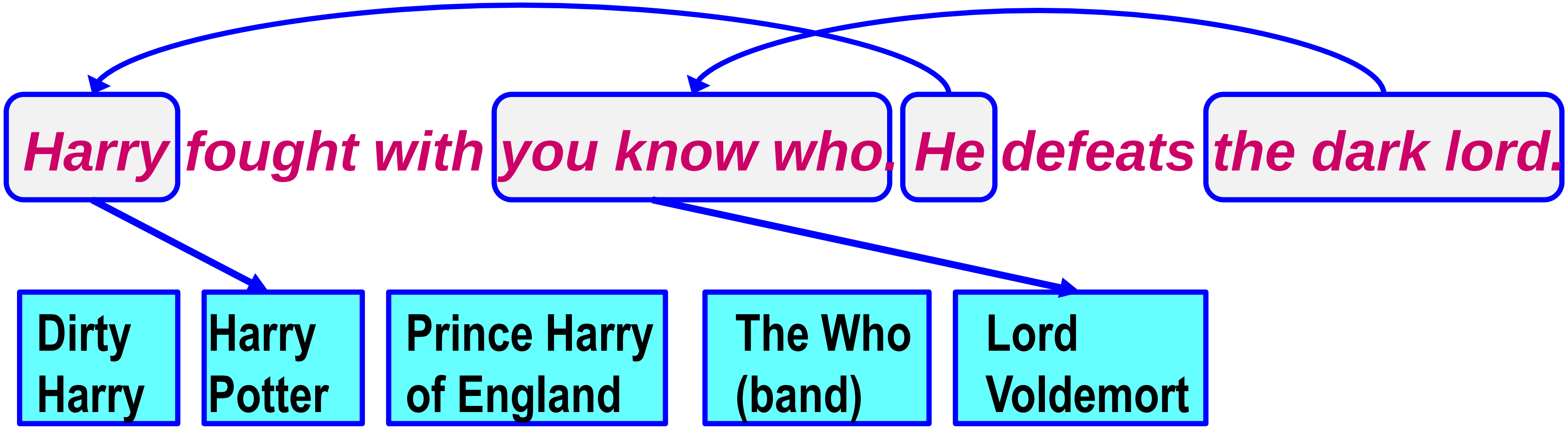
Outline

- ✓ **Motivation and Overview**
- ★ **Taxonomic Knowledge:**
  Entities and Classes
- ★ **Factual Knowledge:**
  Relations between Entities

- ★ **Emerging Knowledge:**
  New Entities & Relations
- ★ **Temporal Knowledge:**
  Validity Times of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
  - **NERD Problem**
  - **NED Principles**
  - **Coherence-based Methods**
  - **Rare & Emerging Entities**
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-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

Outline

✓ **Motivation and Overview**

★ **Taxonomic Knowledge:**
Entities and Classes

★ **Factual Knowledge:**
Relations between Entities

★ **Emerging Knowledge:**
New Entities & Relations

★ **Temporal Knowledge:**
Validity Times of Facts

★ **Contextual Knowledge:**
Entity Name Disambiguation

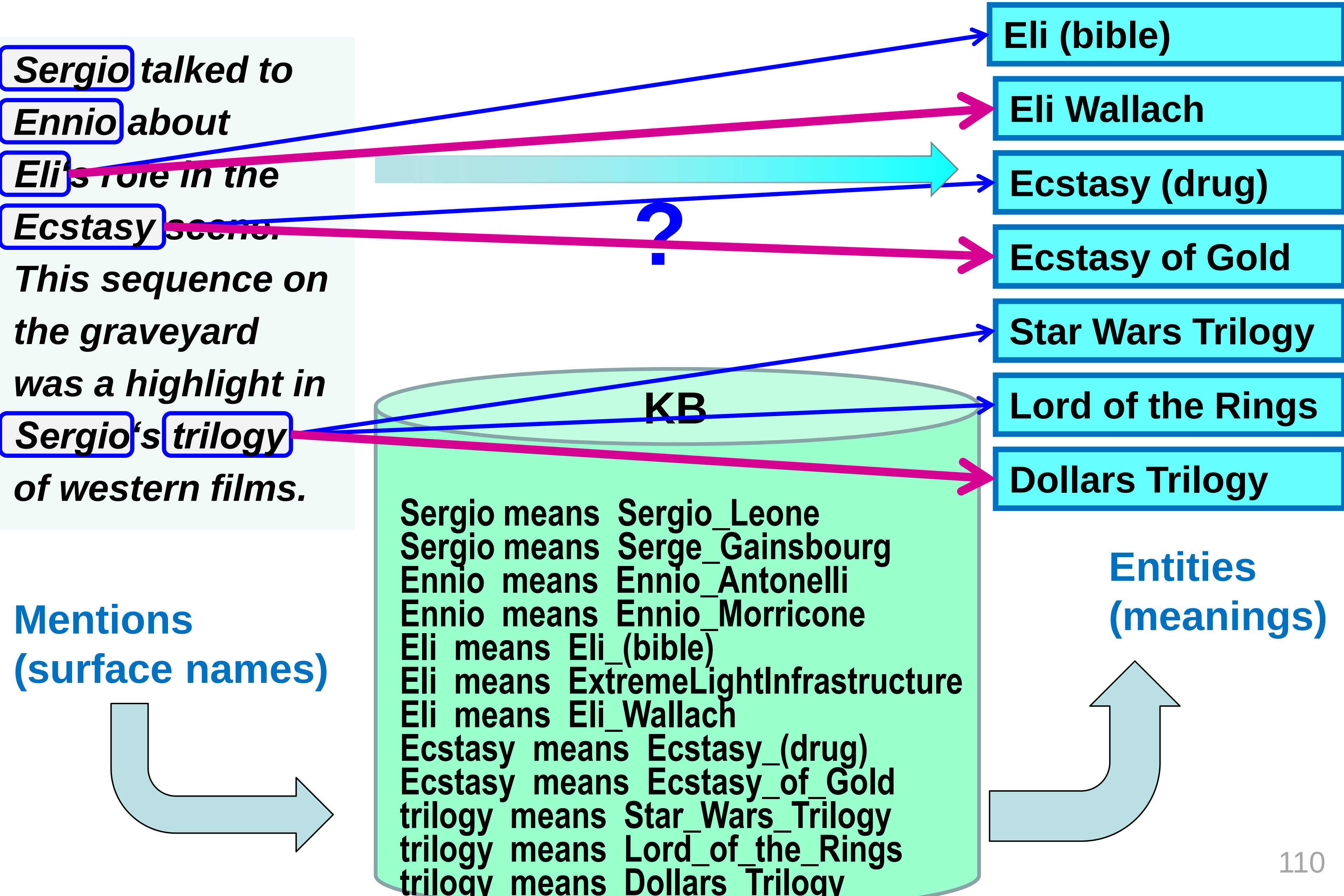
- ✓ **NERD Problem**
- ✶ **NED Principles**
- ✶ **Coherence-based Methods**
- ✶ **Rare & Emerging Entities**

★ **Linked Knowledge:**
Entity Matching

★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

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

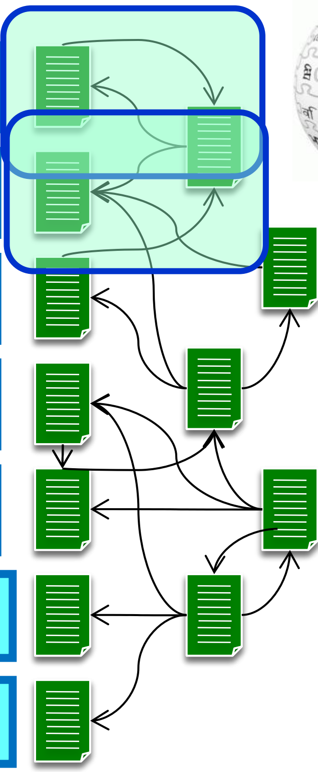
Ecstasy (drug)

Ecstasy of Gold

Star Wars

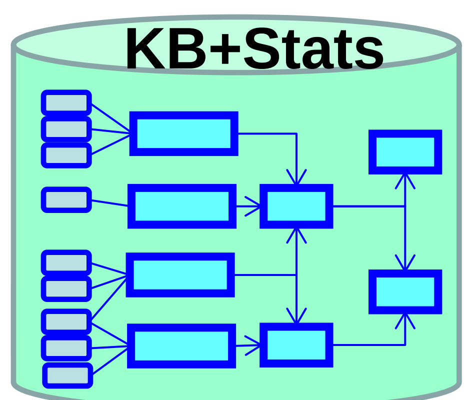
Lord of the Rings

Dollars Trilogy



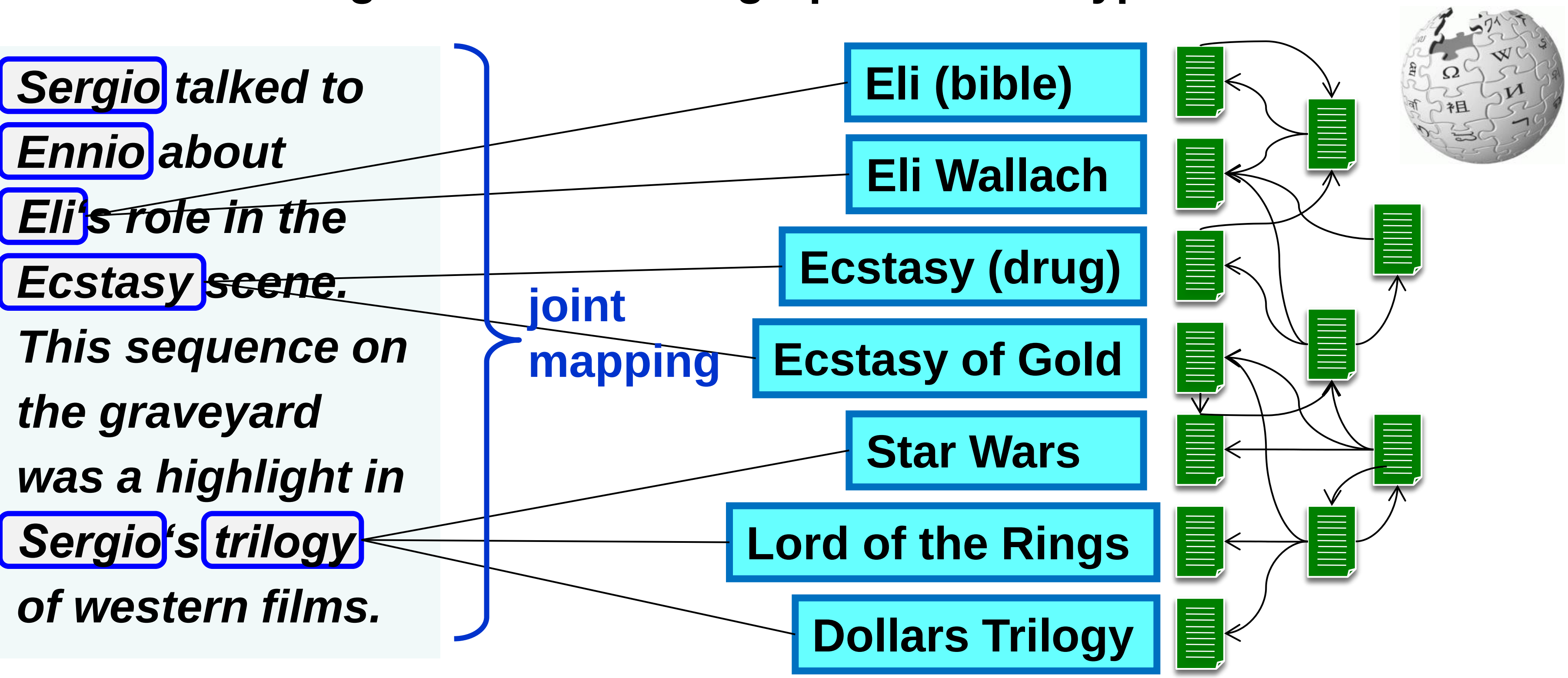
Popularity (m,e):
- freq(e|m)
- length(e)
- #links(e)

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



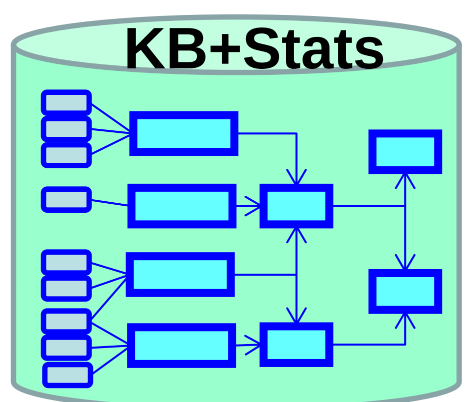
Mention-Entity Graph

weighted undirected graph with two types of nodes



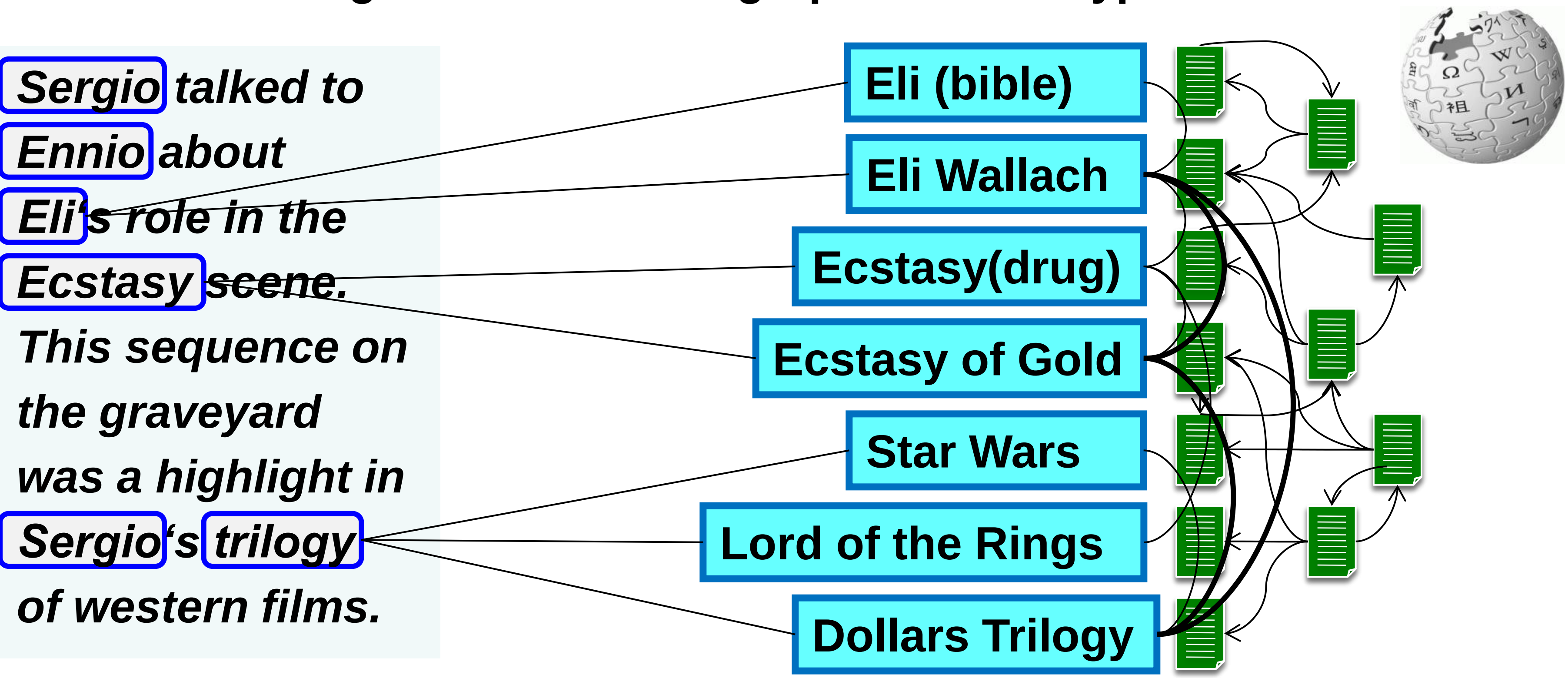
Popularity (m,e):
- freq(e|m)
- length(e)
- #links(e)

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



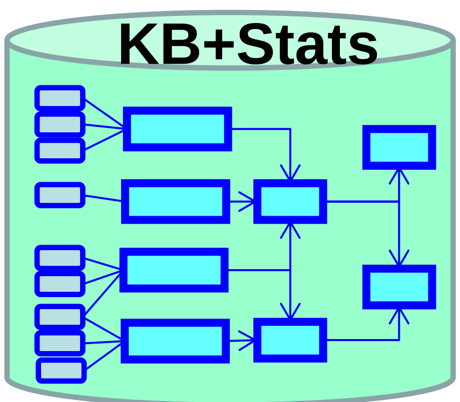
Mention-Entity Graph

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 (keyphrases)

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.

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

American Jews
film actors
artists
Academy Award winners

Metallica songs
Ennio Morricone songs
artifacts
soundtrack music

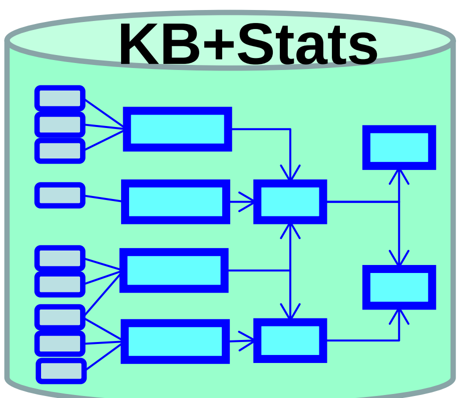
spaghetti westerns
film trilogies
movies
artifacts

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 (keyphrases)

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.

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

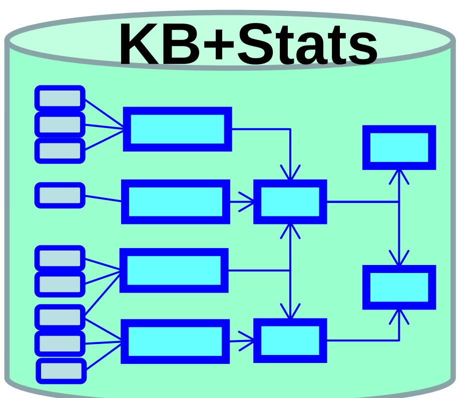
http://.../wiki/Dollars_Trilogy
http://.../wiki/The_Good,_the_Bad,_
http://.../wiki/Clint_Eastwood
http://.../wiki/Honorary_Academy_A

http://.../wiki/The_Good,_the_Bad,_t
http://.../wiki/Metallica
http://.../wiki/Bellagio_(casino)
http://.../wiki/Ennio_Morricone

http://.../wiki/Sergio_Leone
http://.../wiki/The_Good,_the_Bad,_
http://.../wiki/For_a_Few_Dollars_M
http://.../wiki/Ennio_Morricone

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 (keyphrases)

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.

Eli (bible)

Eli Wallach

Ecstasy (drug)

Ecstasy of Gold

Star Wars

Lord of the Rings

Dollars Trilogy

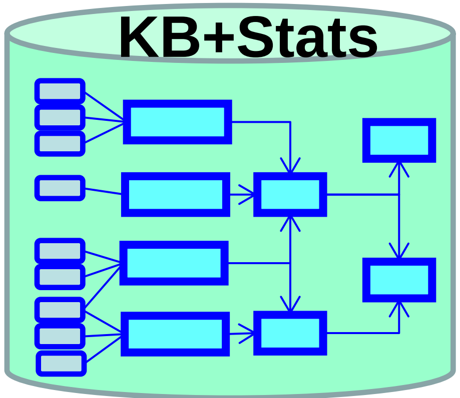
The Magnificent Seven
The Good, the Bad, and the Ugly
Clint Eastwood
University of Texas at Austin

Metallica on Morricone tribute
Bellagio water fountain show
Yo-Yo Ma
Ennio Morricone composition

For a Few Dollars More
The Good, the Bad, and the Ugly
Man with No Name trilogy
soundtrack by Ennio Morricone

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 (keyphrases)

Outline

✓ **Motivation and Overview**

★ **Taxonomic Knowledge:**
Entities and Classes

★ **Factual Knowledge:**
Relations between Entities

★ **Emerging Knowledge:**
New Entities & Relations

★ **Temporal Knowledge:**
Validity Times of Facts

★ **Contextual Knowledge:**
Entity Name Disambiguation

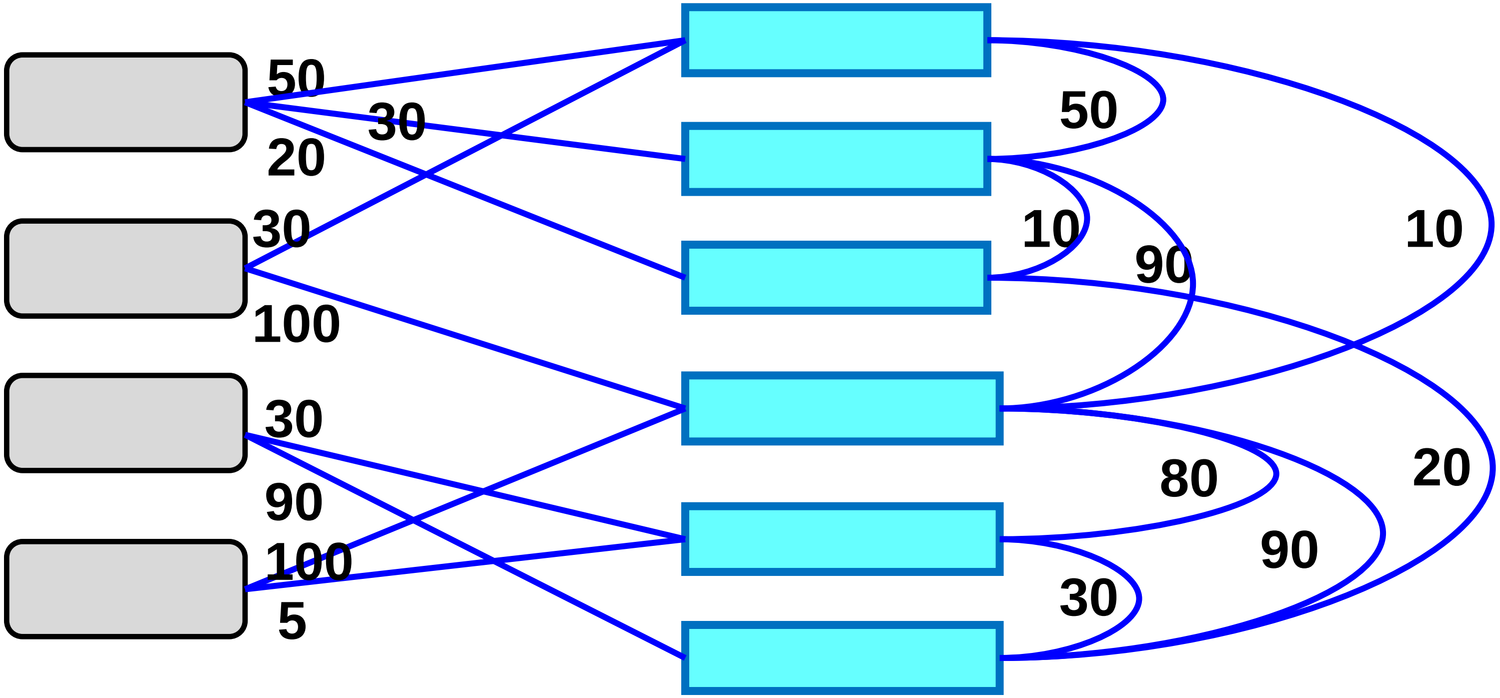
- ✓ **NERD Problem**
- ✓ **NED Principles**
- ✶ **Coherence-based Methods**
- ✶ **Rare & Emerging Entities**

★ **Linked Knowledge:**
Entity Matching

★ **Wrap-up**

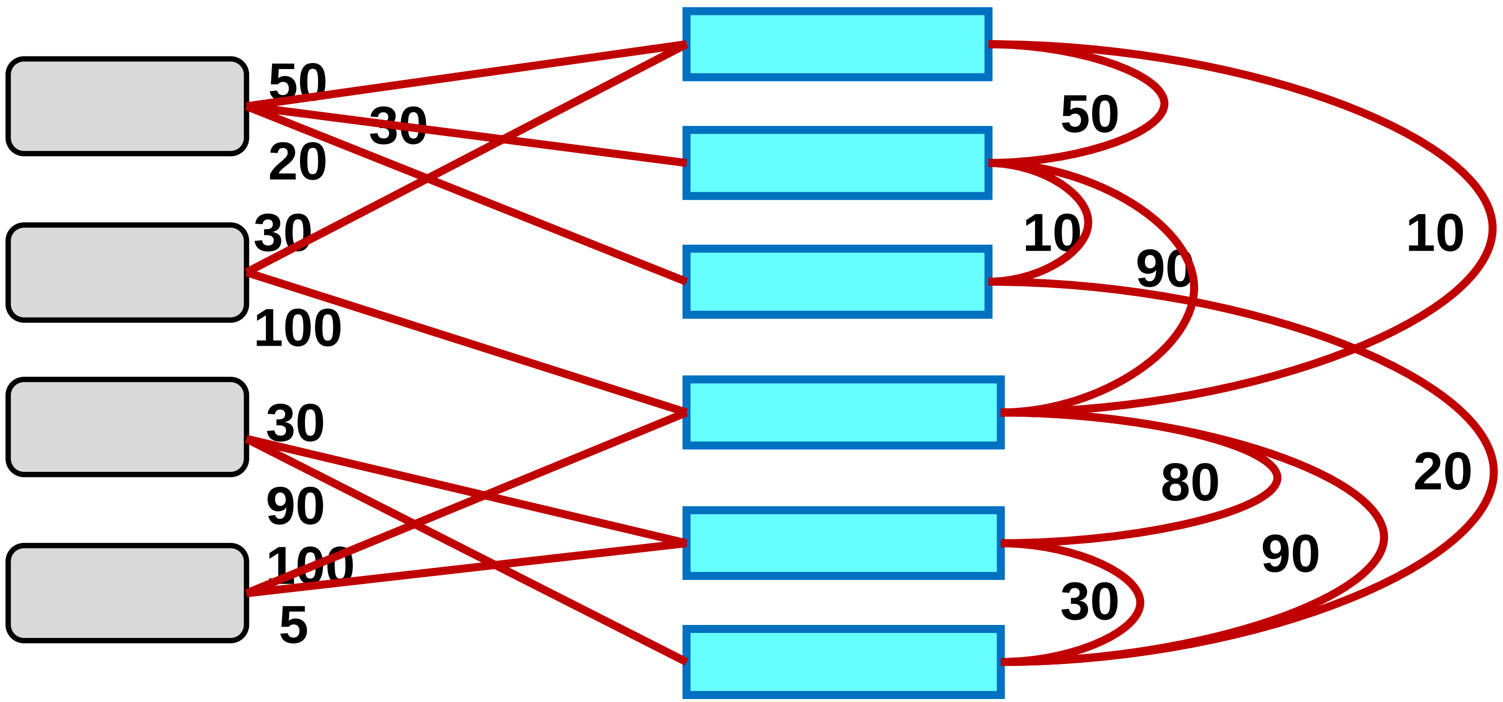
http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

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)

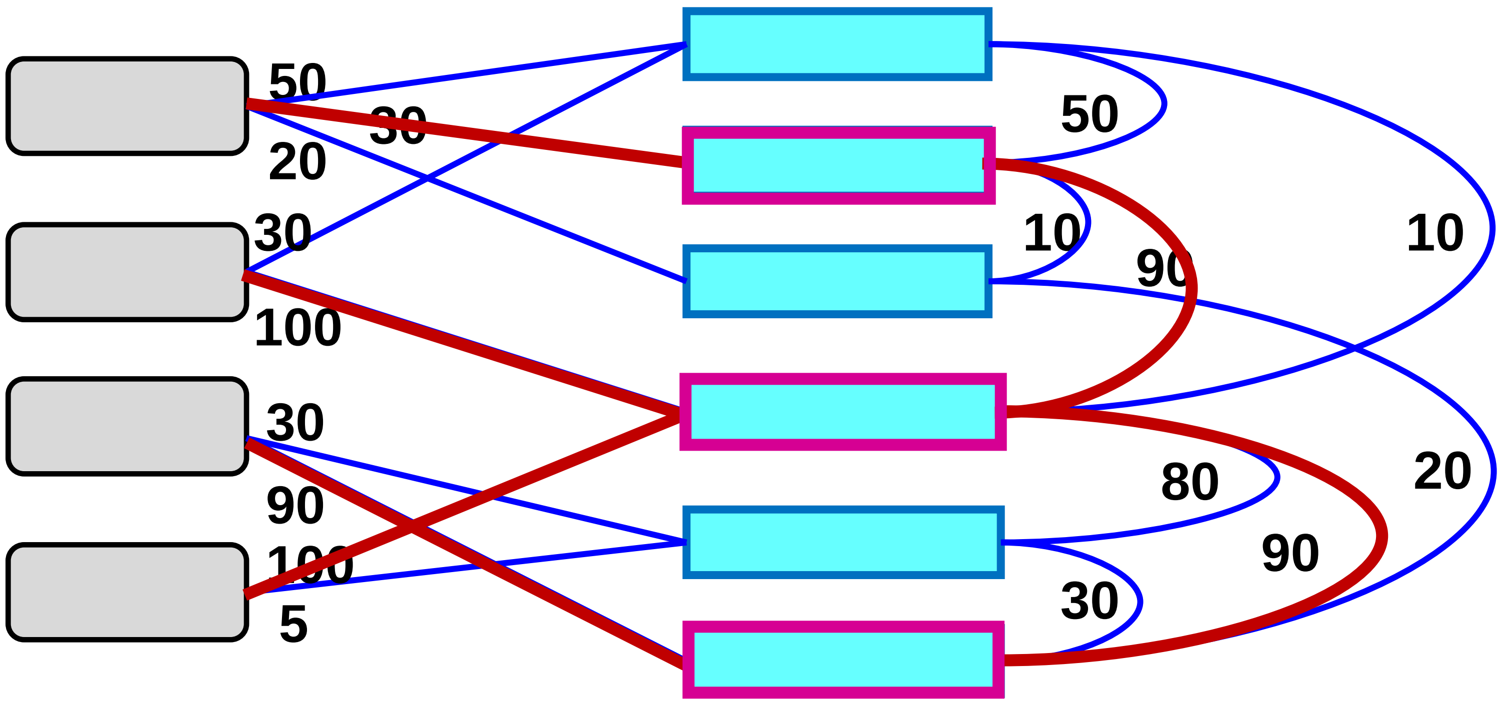
Joint Mapping: Prob. Factor Graph



Collective Learning with Probabilistic Factor Graphs
[Chakrabarti et al.: KDD'09]:

- model **P[m|e]** by similarity and **P[e1|e2]** by coherence
- consider **likelihood** of **P[m1 … mk | e1 … ek]**
- **factorize** by all **m-e pairs** and **e1-e2 pairs**
- use MCMC, hill-climbing, LP etc. for solution

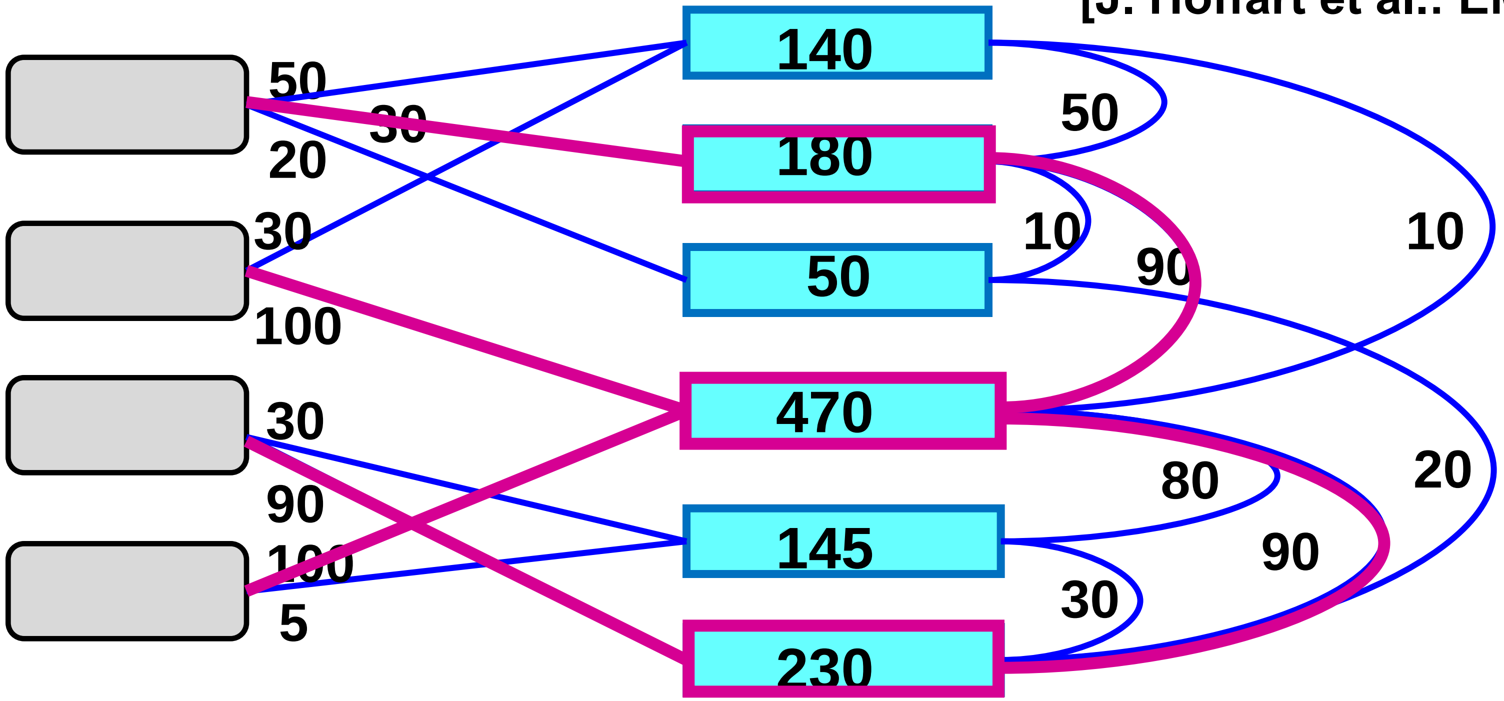
Joint Mapping: Dense Subgraph



- Compute dense subgraph such that:
  each m is connected to exactly one e (or at most one e)
- NP-hard $\rightarrow$ approximation algorithms
- Alt.: feature engineering for similarity-only method
  [Bunescu/Pasca 2006, Cucerzan 2007, Milne/Witten 2008, …]

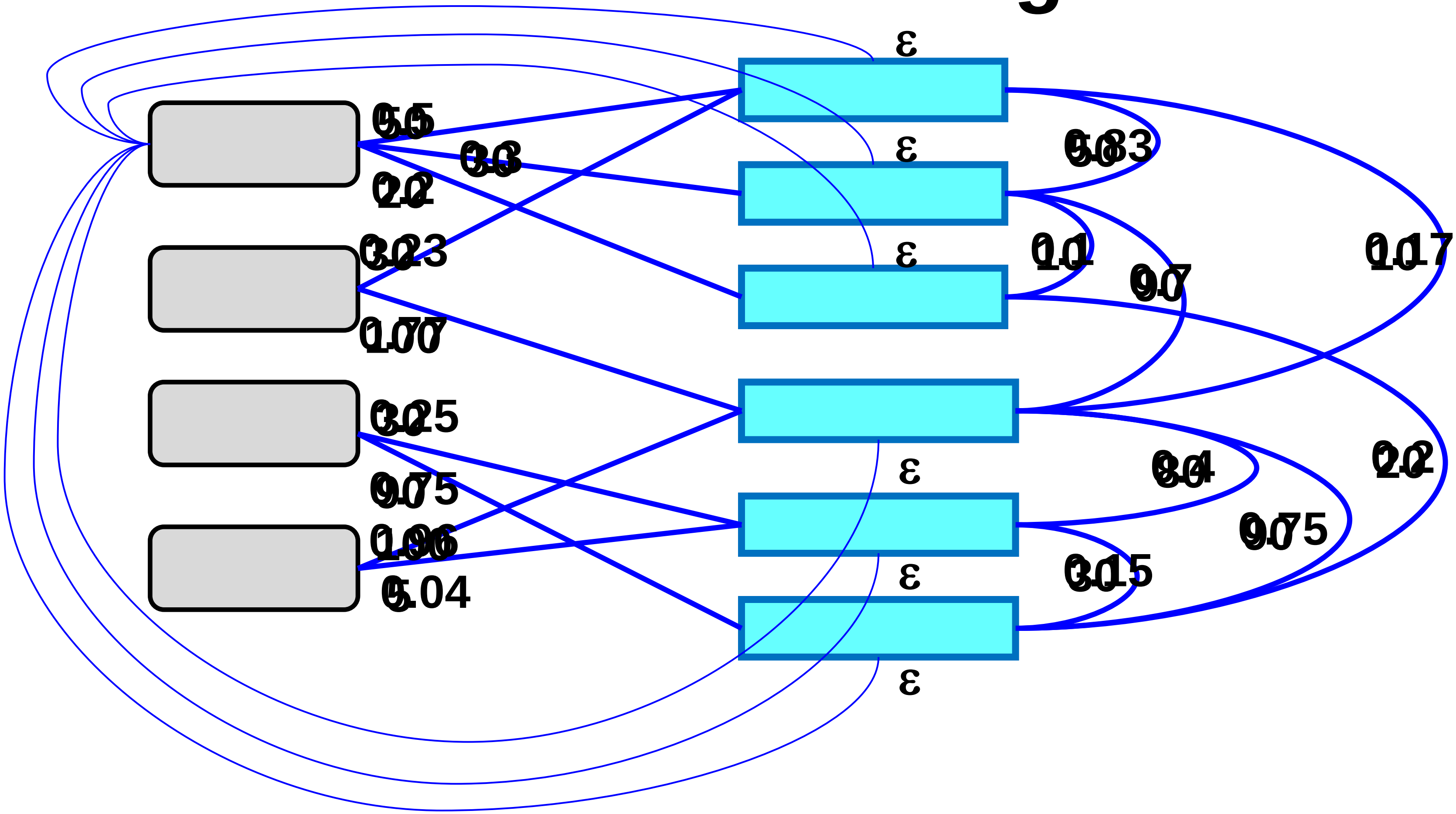
Coherence Graph Algorithm

[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

Random Walks Algorithm



- for each mention run random walks with restart
  (like personalized PageRank with jumps to start mention(s))
- rank candidate entities by stationary visiting probability
- very efficient, decent accuracy

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

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

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

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

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 keyphrases for e1 and e2

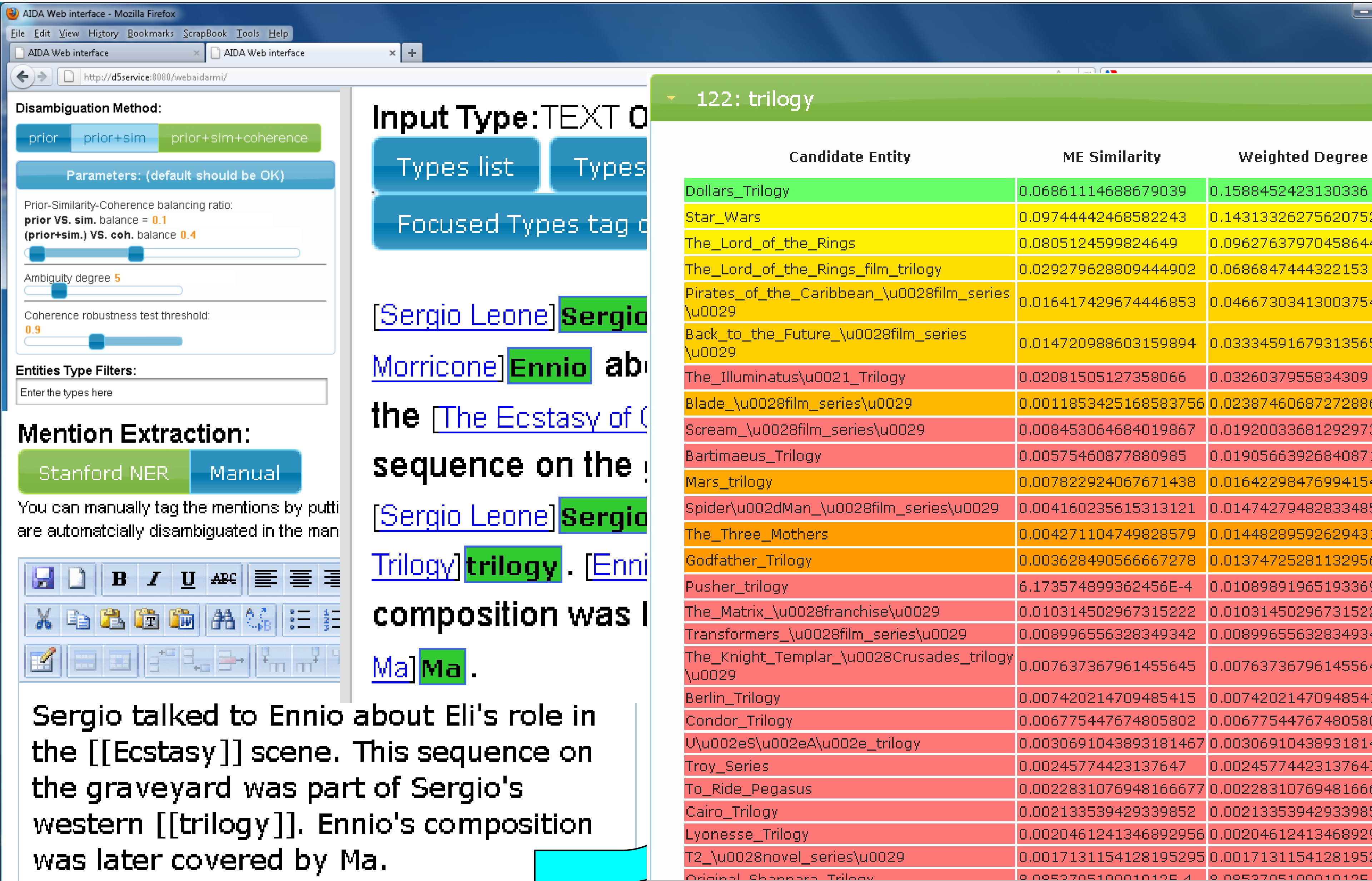
$$ngram\text{-}coh(e1, e2) \sim \frac{|ngrams(e1) \cap ngrams(e2)|}{|ngrams(e1) \cup ngrams(e2)|}$$

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

AIDA: Accurate Online Disambiguation



Candidate Entity	ME Similarity	Weighted Degree
Dollars_Trilogy	0.0686114688679039	0.1588452423130336
Star_Wars	0.0974442468582243	0.14313326275620752
The_Lord_of_the_Rings	0.0805124599824649	0.09627637970458644
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
Original_Shannara_Trilogy	[illegible]	[illegible]

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

AIDA: Very Difficult Example

Disambiguation Method:

prior | prior+sim | prior+sim+coherence

Parameters: (defaults should be OK)

Prior-Similarity-Coherence balancing ratio:
prior VS. sim. balance = 0.4
(prior+sim.) VS. coh. balance 0.8

Ambiguity degree 5

Coherence robustness test threshold:
0.9

Entities Type Filters:

Enter the types here

Mention Extraction:

Stanford NER | Manual

You can manually tag the mentions by putting them between [[and]].
HTML Tables are automatcially disambiguated in the manual mode.

Font size

[[Page]] played Kashmir on a Gibson.

Input Type:TEXT **Overall runtime:**3s, 832ms

Types list | Types tag cloud | Focused Types tag cloud

[Jimmy Page] **Page** played [Kashmir (song)] **Kashmir** on a [Gibson Guitar Corporation] **Gibson**.

l Steps

25: Gibson

Candidate Entity	ME Similarity
Mel_Gibson	0.0
Henry_Gibson	0.0
Gibson_Guitar_Corporation	6.937260822770075E-5
Robert_Gibson_\u0028pitcher\u0029	4.3397387840473426E-5
Kirk_Gibson	0.0
Debbie_Gibson	0.0
William_Gibson	0.0
Tyrese_Gibson	0.0
Aaron_Gibson	0.0
Paul_Gibson	0.0
Don_Gibson	0.0
[illegible]	[illegible]

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/

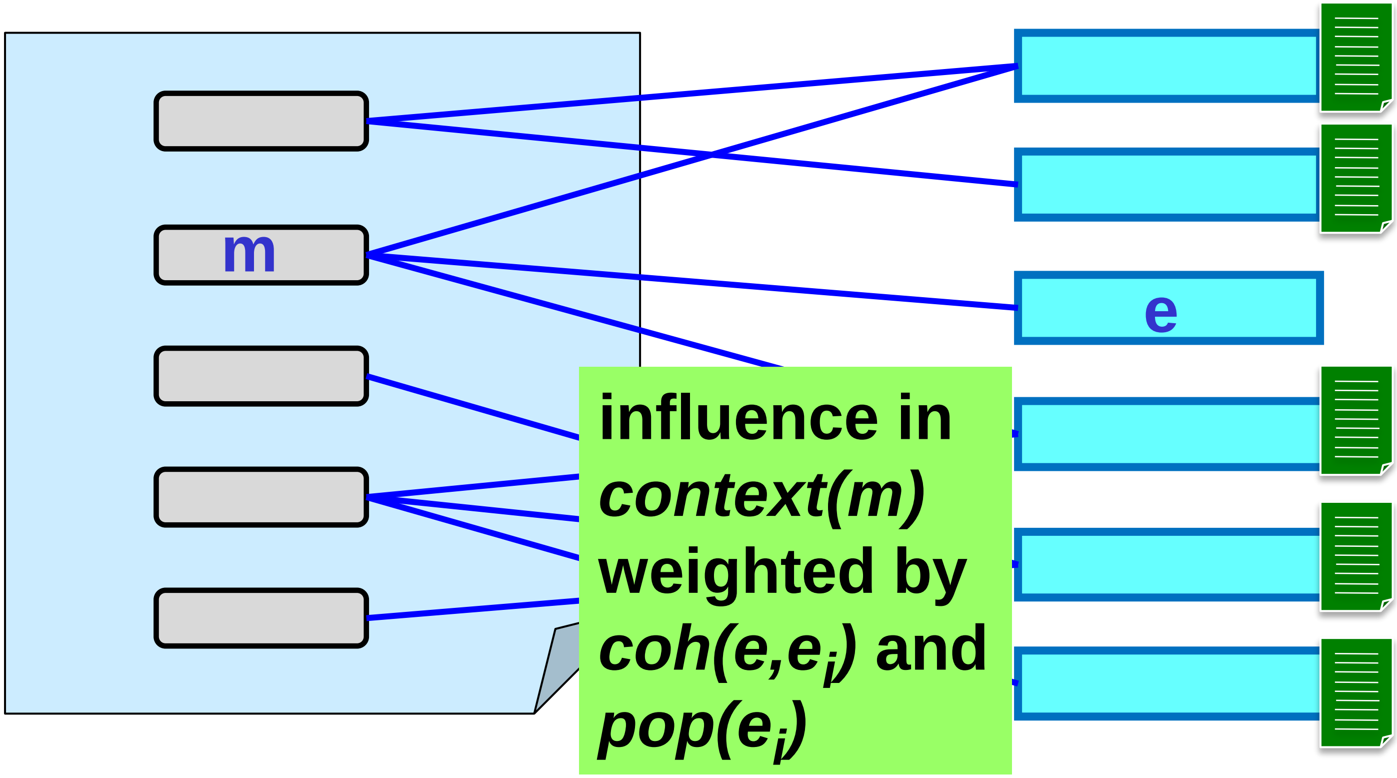
D. Milne, I. Witten: CIKM 2008
http://wikipedia-miner.cms.waikato.ac.nz/demos/annotate/

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

Coherence-aware Feature Engineering

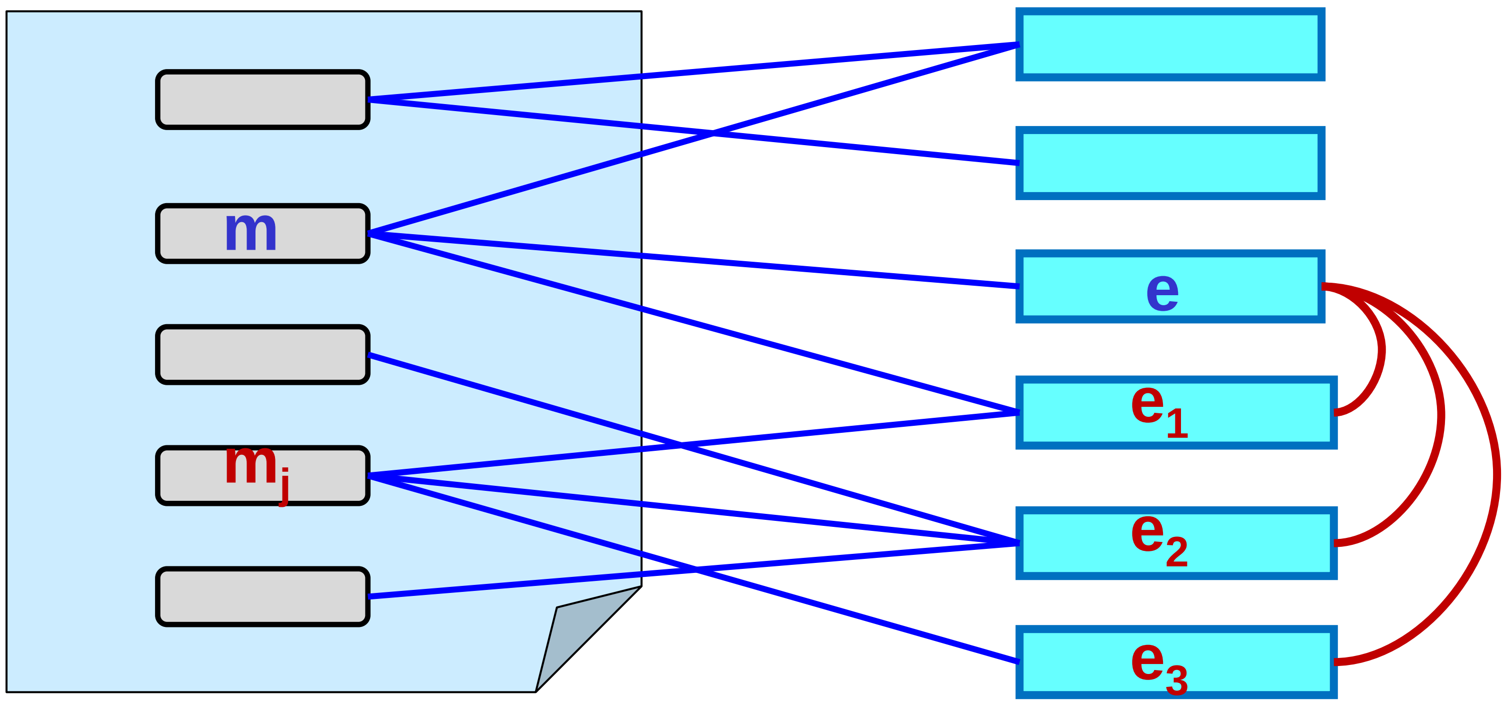
[Cucerzan: EMNLP 2007; Milne/Witten: CIKM 2008, Art.Int. 2013]



- Avoid explicit coherence computation by turning other mentions‘ candidate entities into features
- sim(m,e) uses these features in context(m)
- special case: consider only unambiguous mentions or high-confidence entities (in proximity of m)

TagMe: NED with Light-Weight Coherence

[P. Ferragina et al.: CIKM'10, WWW'13]



- **Reduce combinatorial complexity by using avg. coherence of other mentions' candidate entities**
- **for score(m,e) compute**

$$\text{avg}_{e_i \in cand(m_j)} \; coherence(e_i, e) \cdot popularity(e_i \mid m_j)$$

then sum up over all $m_j \neq m$ („voting")

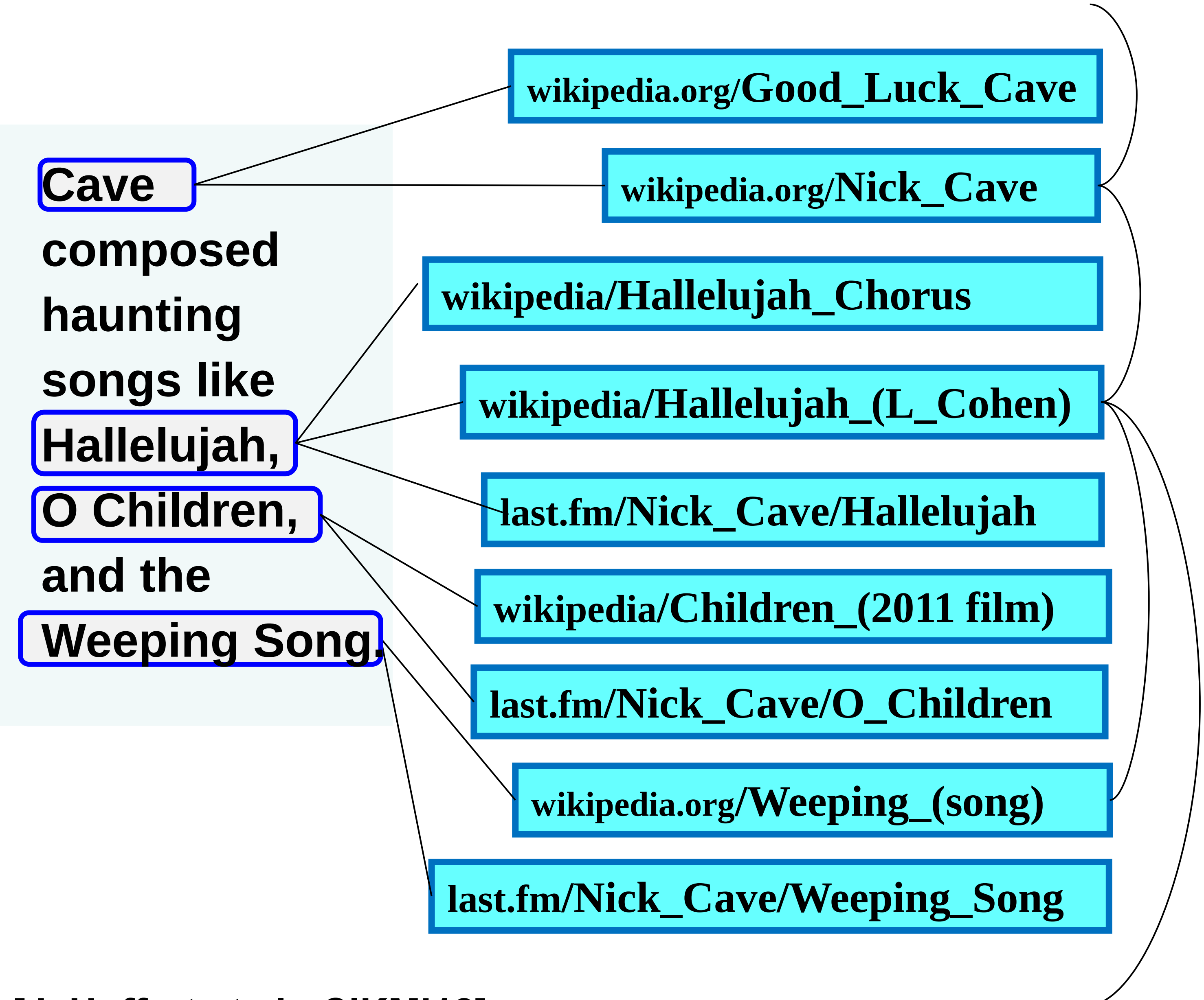
Outline

- ✓ **Motivation and Overview**
- ★ **Taxonomic Knowledge:**
  Entities and Classes
- ★ **Factual Knowledge:**
  Relations between Entities

- ★ **Emerging Knowledge:**
  New Entities & Relations
- ★ **Temporal Knowledge:**
  Validity Times of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
  - ✓ **NERD Problem**
  - ✓ **NED Principles**
  - ✓ **Coherence-based Methods**
  - ★ **Rare & Emerging Entities**
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Long-Tail and Emerging Entities



[J. Hoffart et al.: CIKM'12]

Long-Tail and Emerging Entities

Cave composed haunting songs like **Hallelujah**, **O Children**, and the **Weeping Song**.

- wikipedia.org/Good_Luck_Cave
  - Gunung Mulu National Park; Sarawak Chamber; largest underground chamber
- wikipedia.org/Nick_Cave
  - Bad Seeds; No More Shall We Part; Murder Songs
- wikipedia/Hallelujah_Chorus
  - Messiah oratorio; George Frideric Handel
- wikipedia/Hallelujah_(L_Cohen)
  - Leonard Cohen; Rufus Wainwright; Shrek and Fiona
- last.fm/Nick_Cave/Hallelujah
  - eerie violin; Bad Seeds; No More Shall We Part
- wikipedia/Children_(2011 film)
  - South Korean film
- last.fm/Nick_Cave/O_Children
  - Nick Cave & Bad Seeds; Harry Potter 7 movie; haunting choir

em

duet

KO (p,q) = $\frac{\sum_t min(weight(t\ in\ p), weight(t\ in\ q))}{\sum_t max(weight(t\ in\ p), weight(t\ in\ q))}$

KORE (e,f) ~ $\sum_{p \in e, q \in f}) KO(p,q)^2 \times min(weight(p\ in\ e), weight(q\ in\ f))$

implementation uses min-hash and LSH

[J. Hoffart et al.: CIKM'12]

Long-Tail and Emerging Entities

Cave's brand-new album contains masterpieces like Water's Edge and Mermaids.

wikipedia.org/Good_Luck_Cave

Gunung Mulu National Park
Sarawak Chamber
largest underground chamber

wikipedia.org/Nick_Cave

Bad Seeds
No More Shall We Part
Murder Songs

…/Water's Edge Restaurant

excellent seafood
clam chowder
Maine lobster

…/Water's Edge (2003 film)

Nathan Fillion
horrible acting

any OTHER „Water's Edge“

all phrases minus keyphrases of known candidate entities

…/Mermaid's Song

Pirates of the Caribbean 4
My Jolly Sailor Bold
Johnny Depp

…/The Little Mermaid

Walt Disney
Hans Chrisitan Andersen
Kiss the Girl

any OTHER „Mermaids“

all phrases minus keyphrases of known candidate entities

Semantic Typing of Emerging Entities

[N. Nakashole et al.: ACL 2013, T. Lin et al.: EMNLP 2012]

Problem: what to do with newly emerging entities

Idea: infer their semantic types using PATTY patterns

Sandy *threatens to hit* New York
Nive Nielsen *and her band performing* Good for You
Nive Nielsen*'s warm voice in* Good for You

Given triples (x, p, y) with new x,y
and all type triples (t1, p, t2) for known entities:

- score (x,t) ~ $\Sigma_{p:(x,p,y)}$ P [t | p,y] + $\Sigma_{p:(y,p,x)}$ P [t | p,y]
- corr(t_1,t_2) ~ Pearson coefficient $\in$ [-1,+1]

For each new e and all candidate types t_i:

$$\max \alpha \Sigma_i \text{score}(e,t_i) X_i + \beta \Sigma_{ij} \text{corr}(t_i,t_j) Y_{ij}$$

$$\text{s.t. } X_i, Y_{ij} \in \{0,1\} \text{ and } Y_{ij} \le X_i \text{ and } Y_{ij} \le X_j \text{ and } X_i + X_j - 1 \le Y_{ij}$$

Big Data Algorithms at Work

Web-scale keyphrase mining

Web-scale entity-entity statistics

MAP on large factor graph or dense subgraphs in large graph

data+text queries on huge KB or LOD

Applications to large-scale input batches:

- **discover all musicians in a week‘s social media postings**
- **identify all diseases & drugs in a month‘s publications**
- **track a (set of) politician(s) in a decade‘s news archive**

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

Good approaches, more work needed

Open Problems and Grand Challenges

Efficient interactive & high-throughput batch NERD

a day's news, a month's publications, a decade's archive

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)

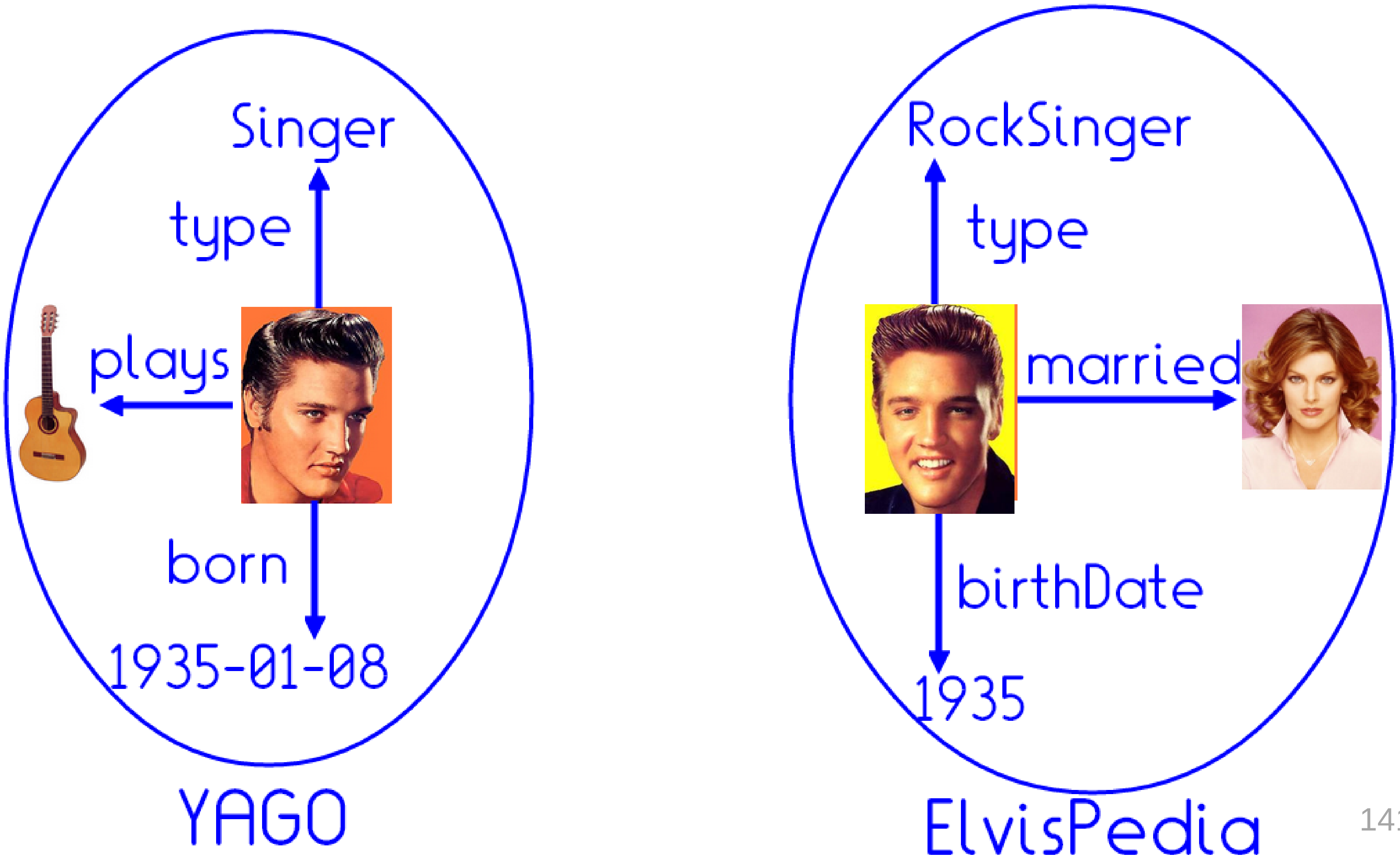
Outline

- ✓ **Motivation and Overview**
- ★ **Taxonomic Knowledge:**
  Entities and Classes
- ★ **Factual Knowledge:**
  Relations between Entities

- ★ **Emerging Knowledge:**
  New Entities & Relations
- ★ **Temporal Knowledge:**
  Validity Times of Facts
- ★ **Contextual Knowledge:**
  Entity Name Disambiguation
- ★ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

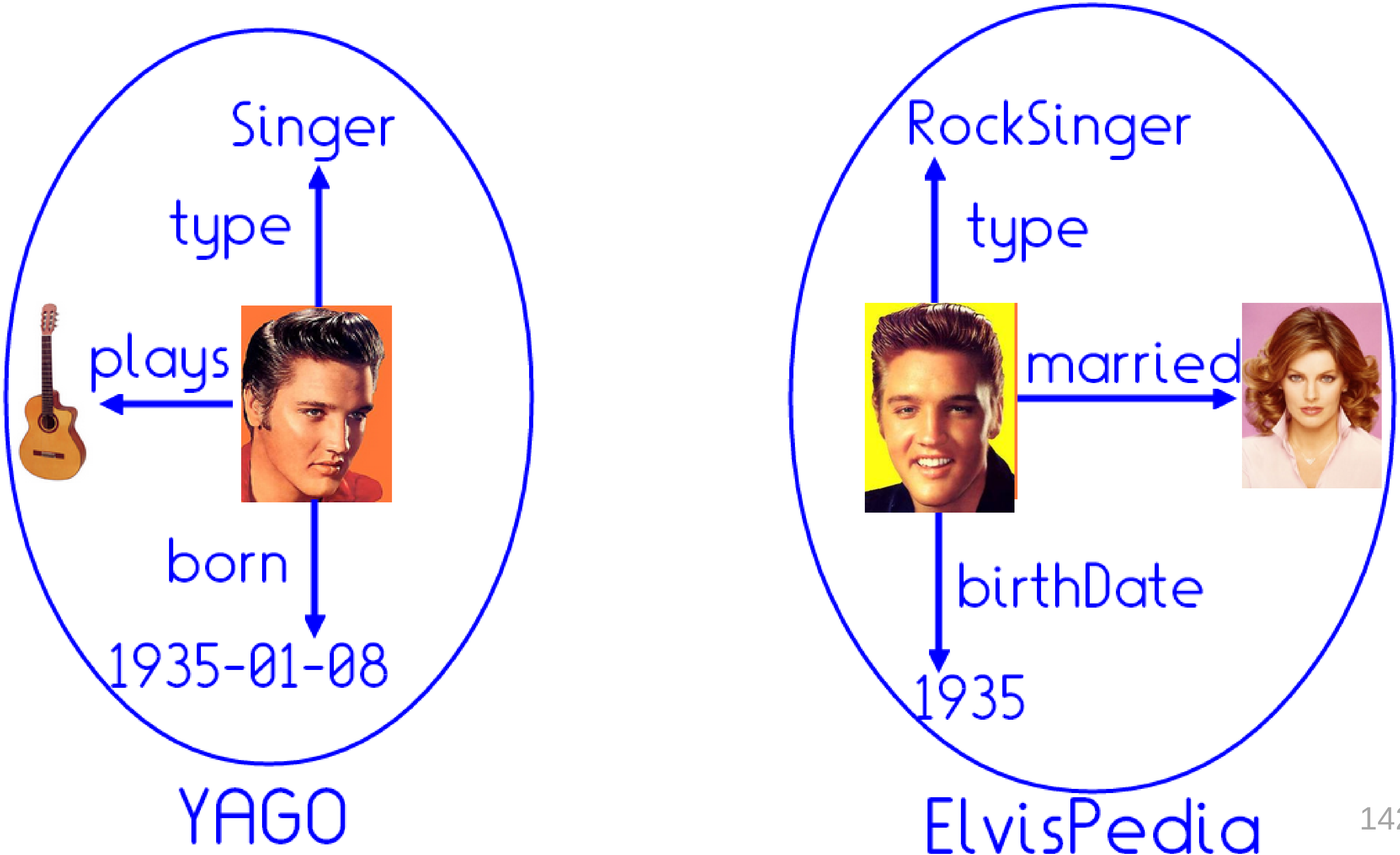
http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Knowledge bases are complementary

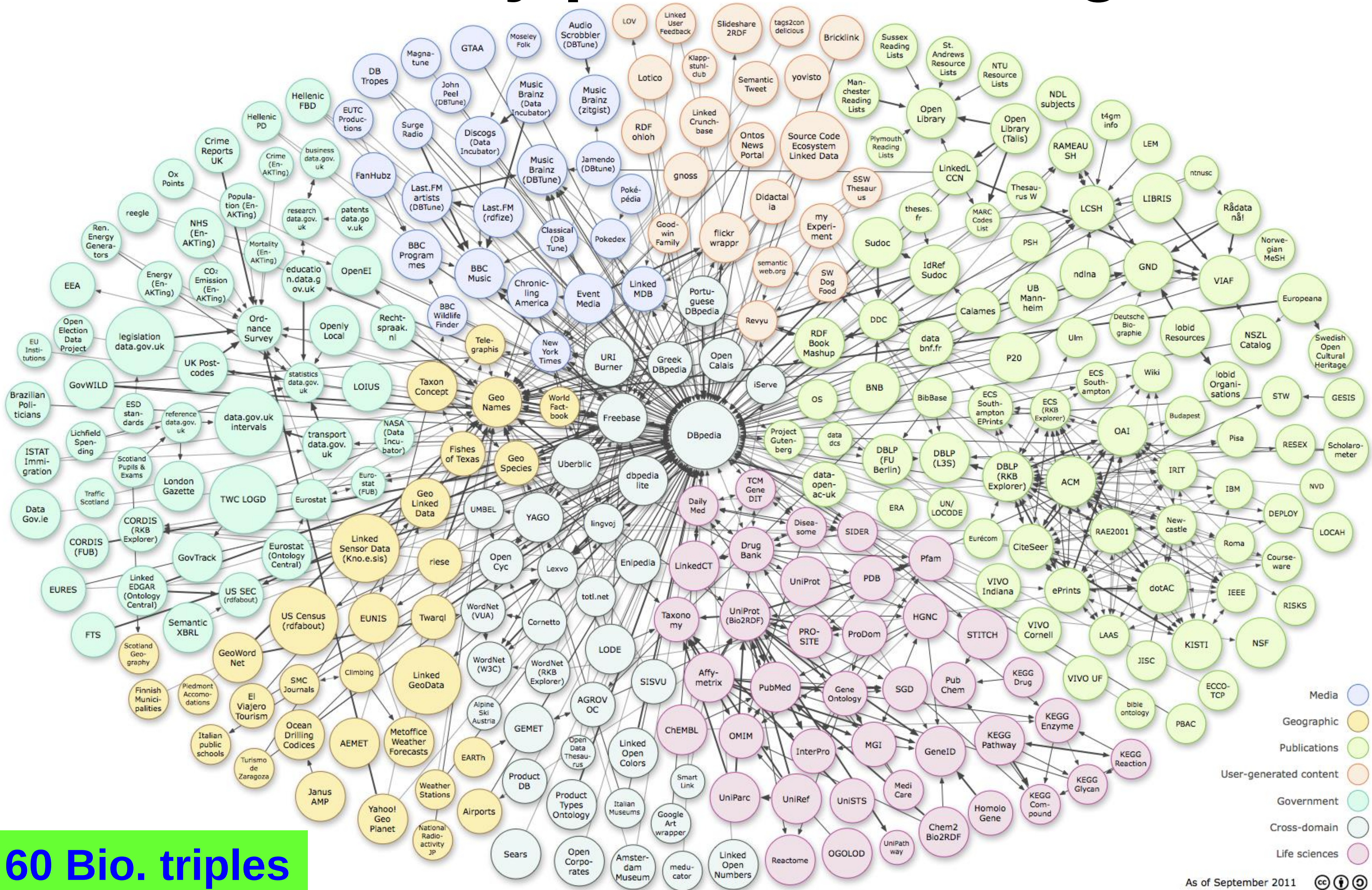


No Links ⇒ No Use

Who is the spouse of the guitar player?



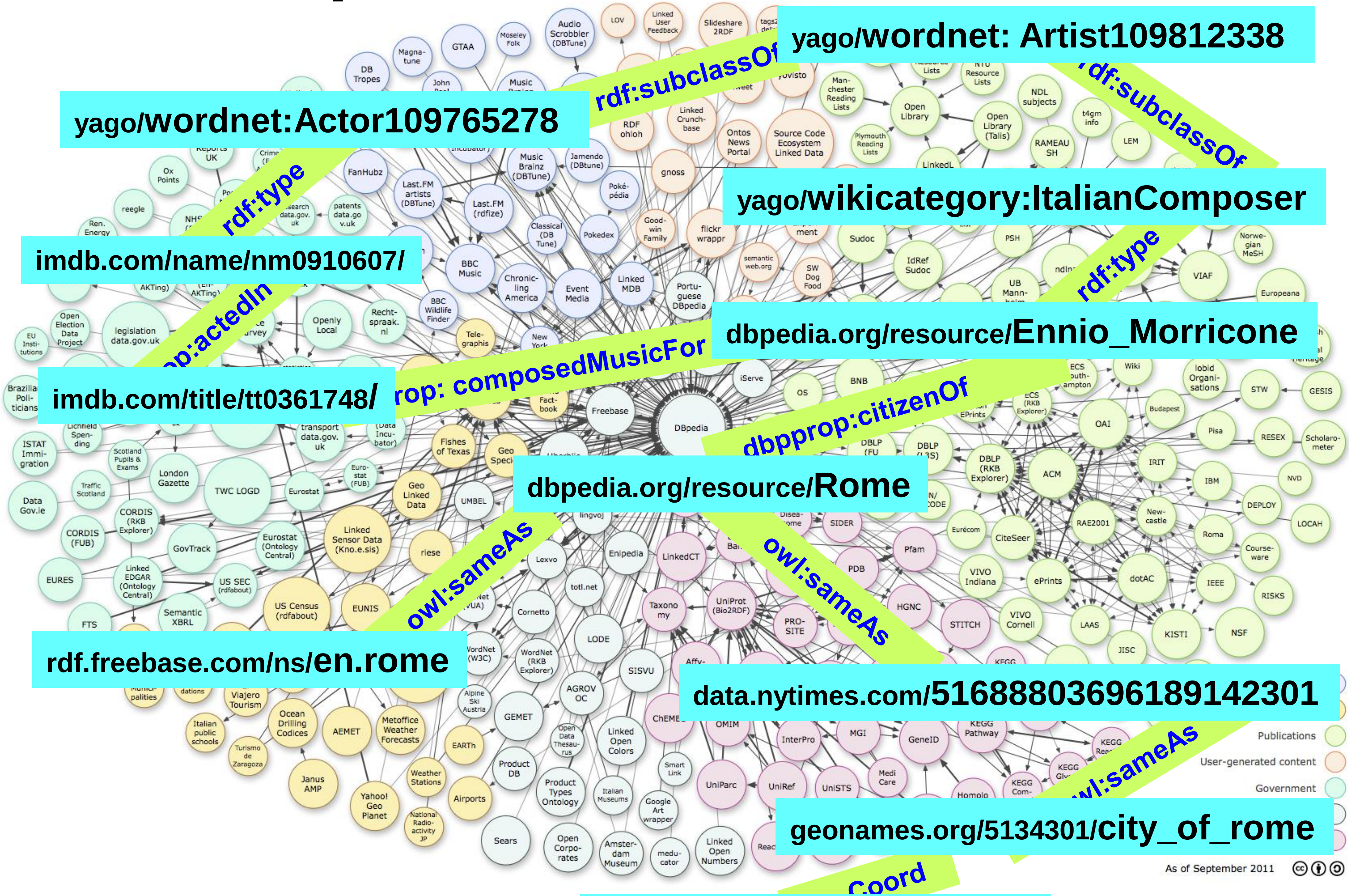
There are many public knowledge bases



60 Bio. triples
500 Mio. links

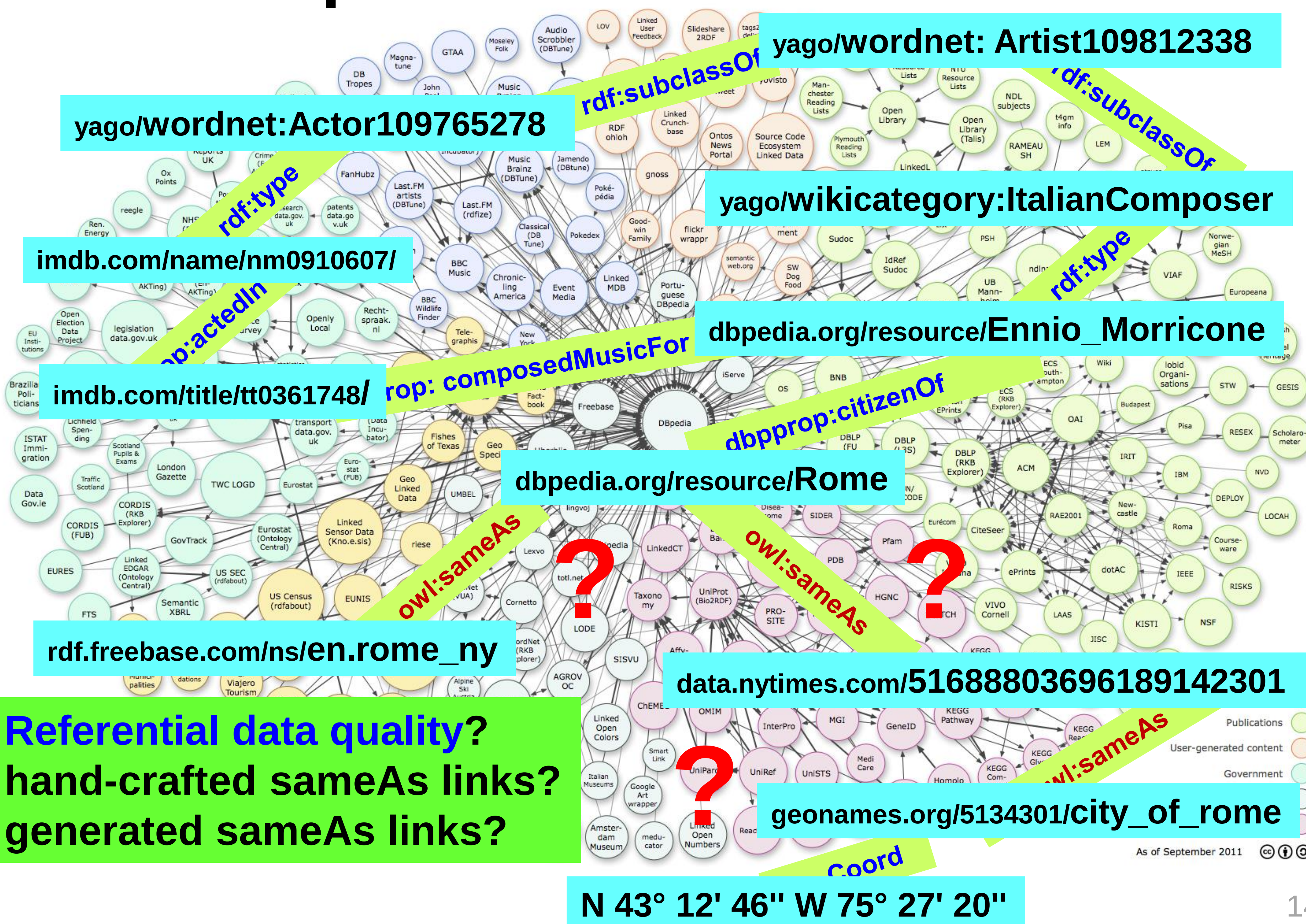
http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.png

Link equivalent entities across KBs



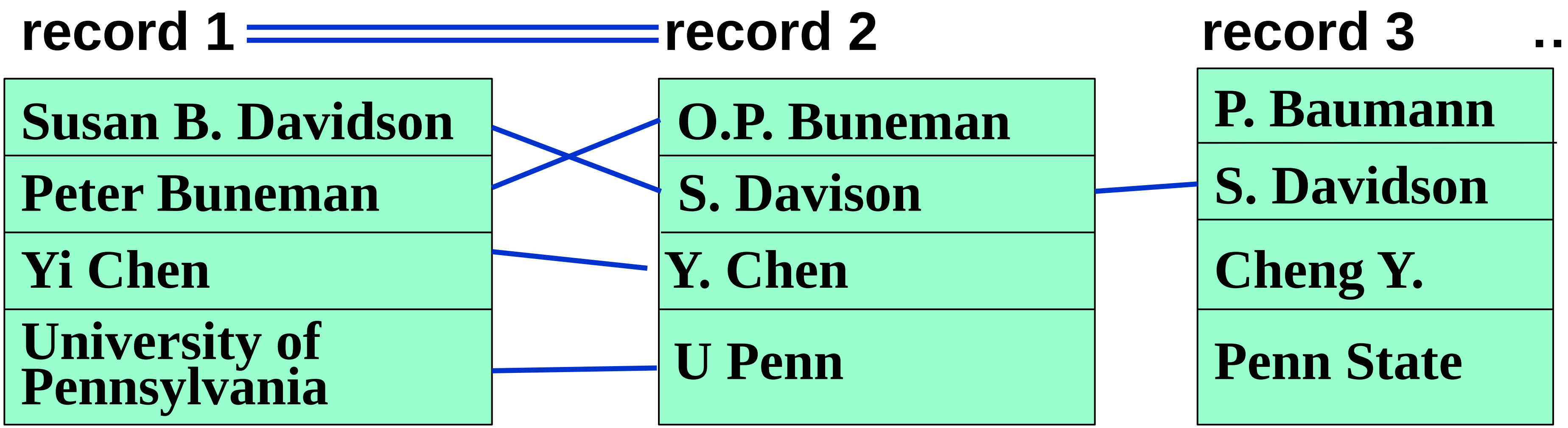
N 43° 12' 46'' W 75° 27' 20''

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.)**

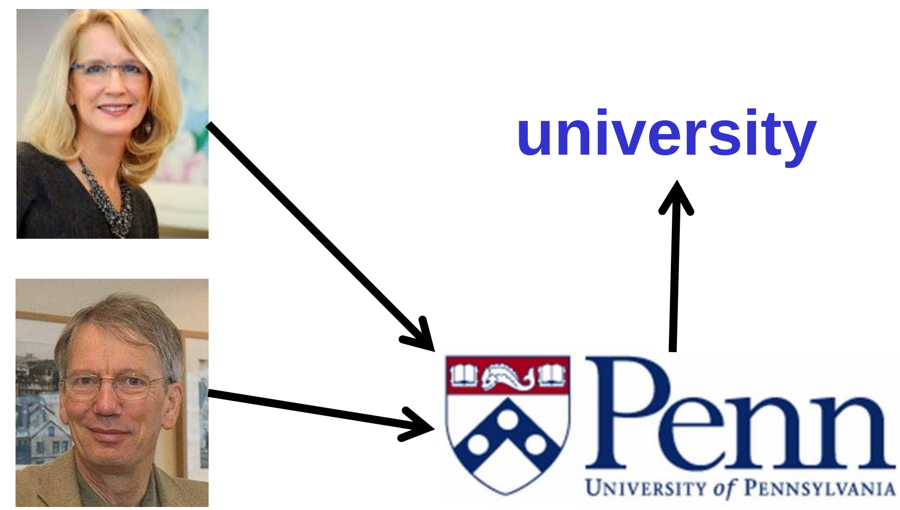
Halbert L. Dunn: Record Linkage. American Journal of Public Health. 1946
H.B. Newcombe et al.: Automatic Linkage of Vital Records. Science, 1959.
I.P. Fellegi, A.B. Sunter: A Theory of Record Linkage. J. of American Statist. Soc., 1969.

Linking Records vs. Linking Knowledge

record

Susan B. Davidson
Peter Buneman
Yi Chen
University of Pennsylvania

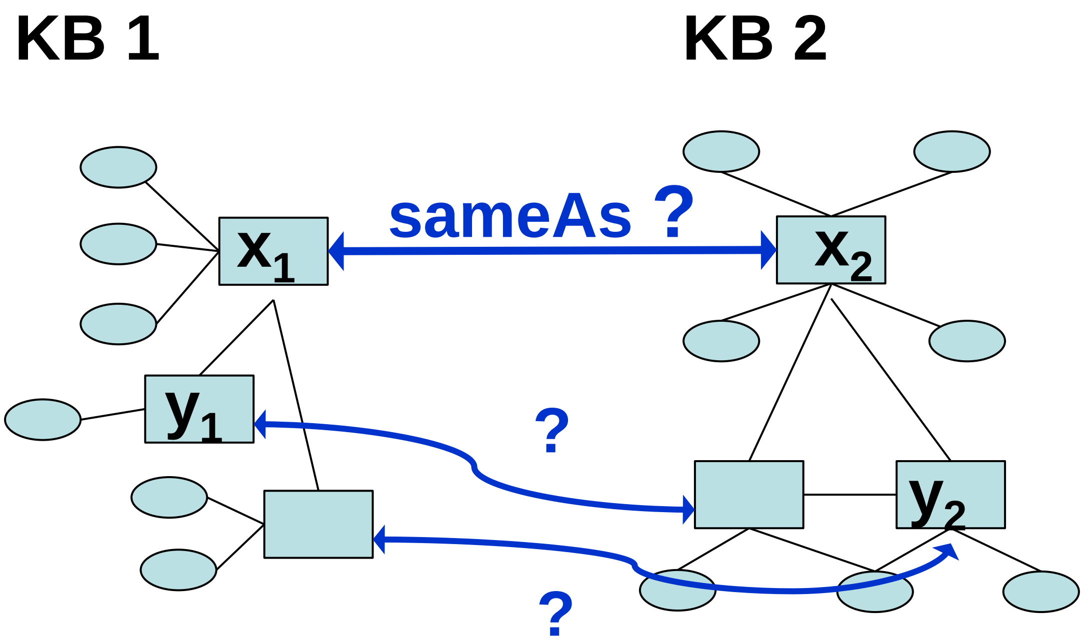
KB / Ontology



Differences between DB records and KB entities:

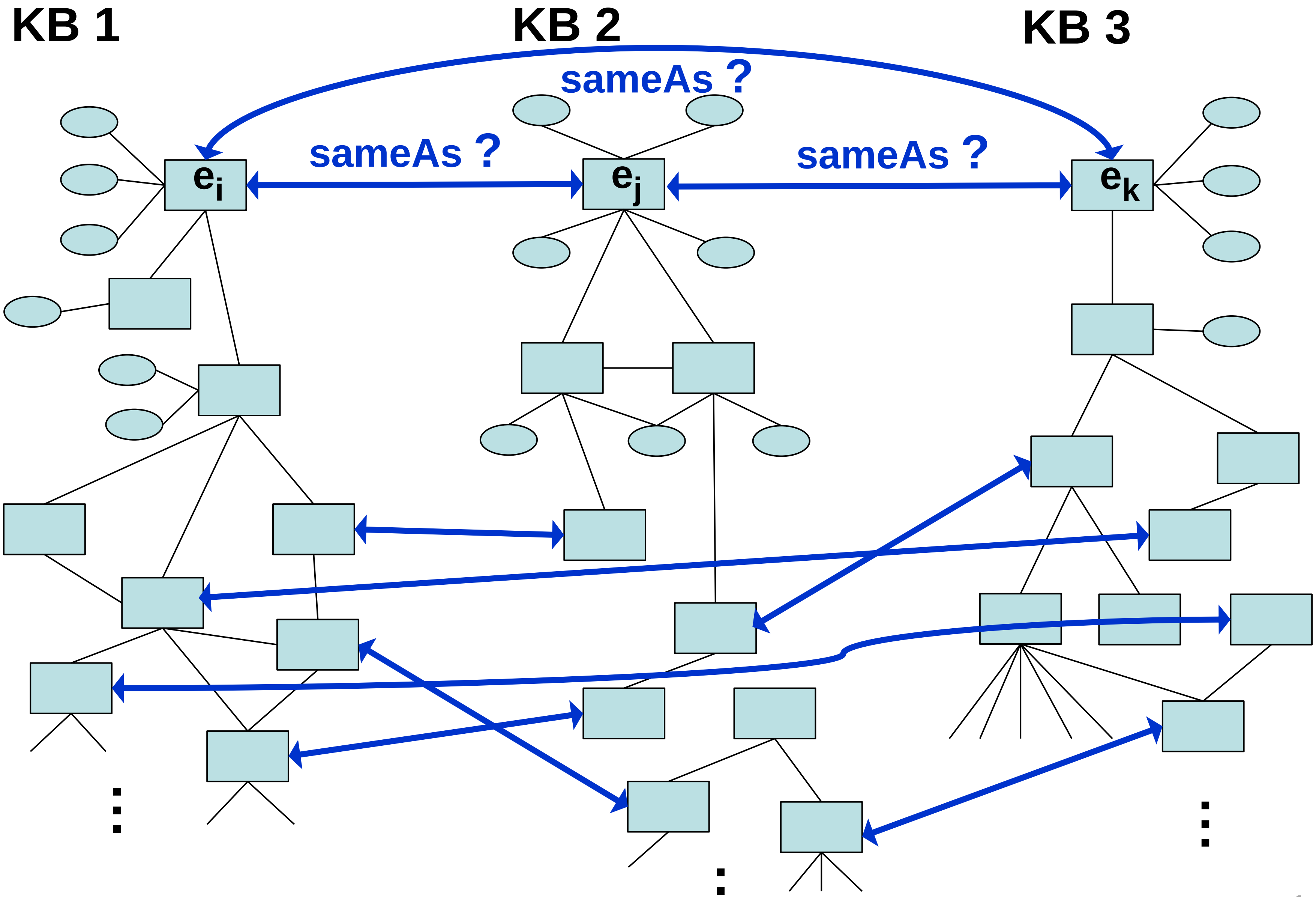
- **Ontological links have rich semantics (e.g. subclassOf)**
- **Ontologies have only binary predicates**
- **Ontologies have no schema**
- **Match not just entities,**
  but also classes & predicates (relations)

Similarity of entities depends on similarity of neighborhoods



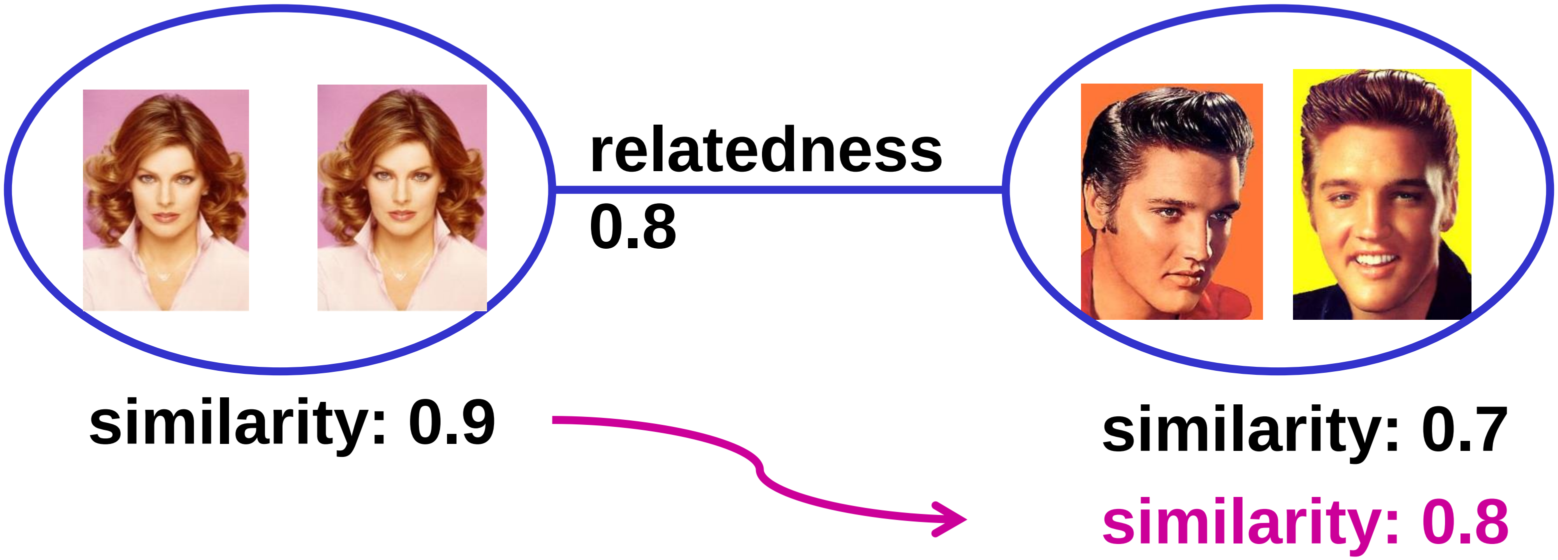
sameAs(x1, x2) **depends on** **sameAs(y1, y2)**
which depends on **sameAs(x1, x2)**

Equivalence of entities is transitive



Similarity Flooding matches entities at scale

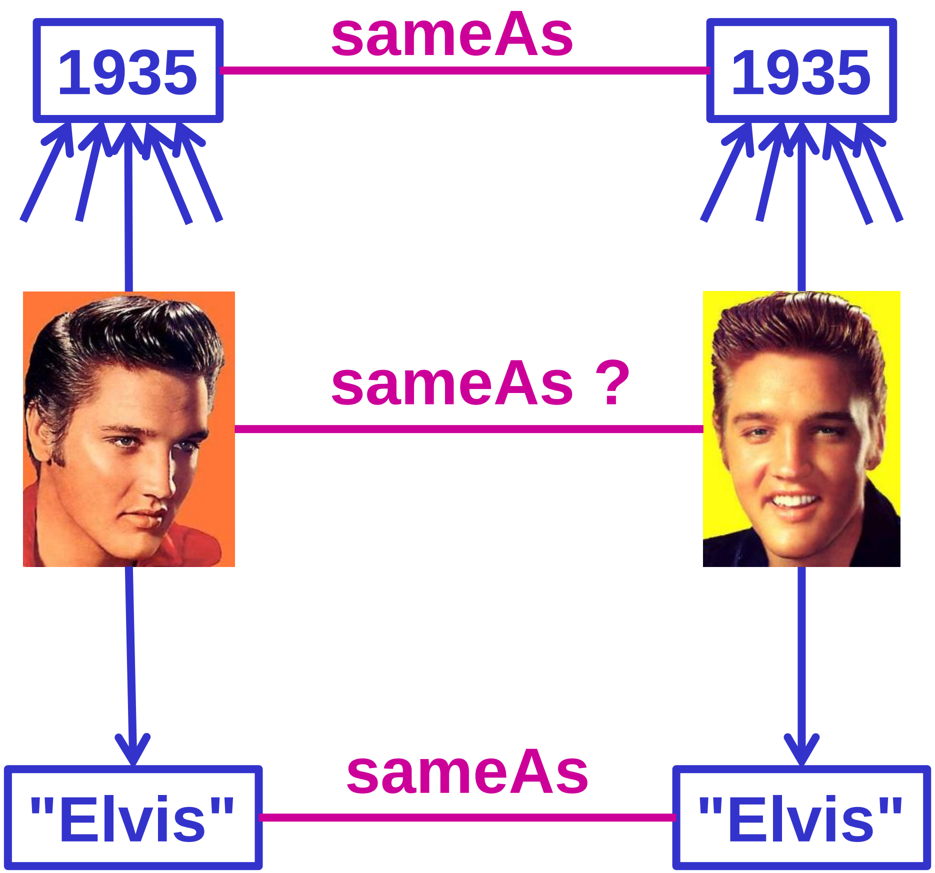
Build a graph:
nodes: pairs of entities, weighted with similarity
edges: weighted with degree of relatedness



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

many variants (belief propagation, label propagation, etc.), e.g. SigMa

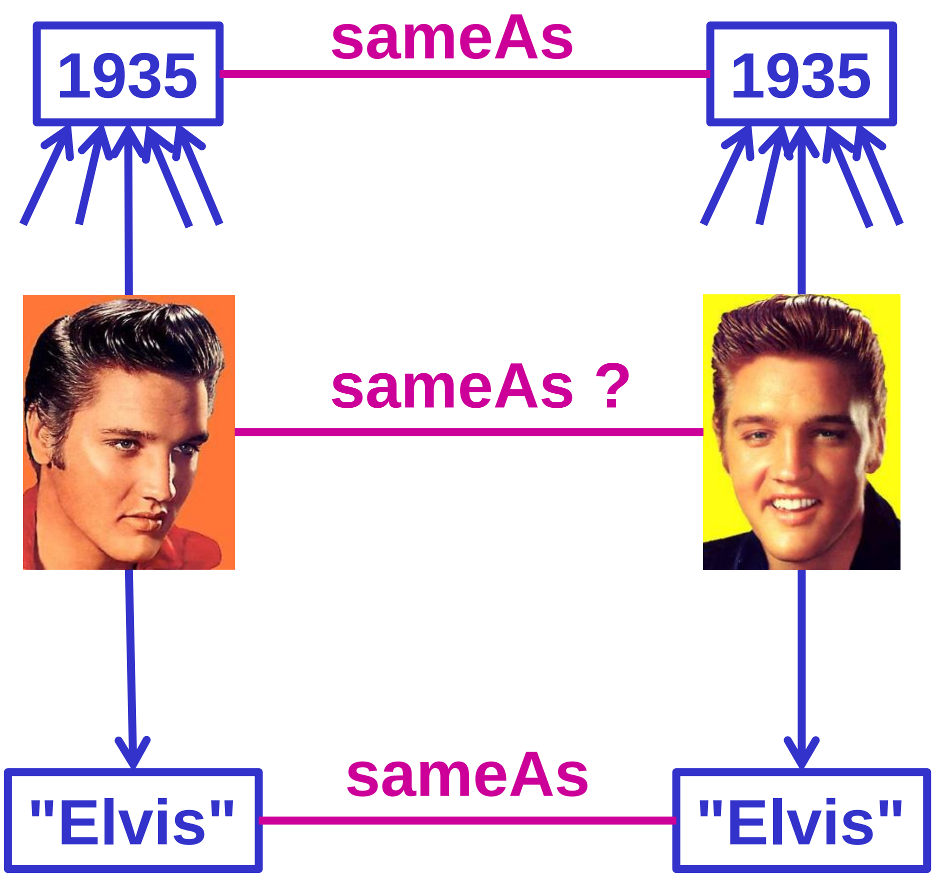
Some neighborhoods are more indicative



Many people born in 1935
⇒ not indicative

Few people called "Elvis"
⇒ highly indicative

Inverse functionality as indicativeness



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

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

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

$$ifun(born) = 0.01$$

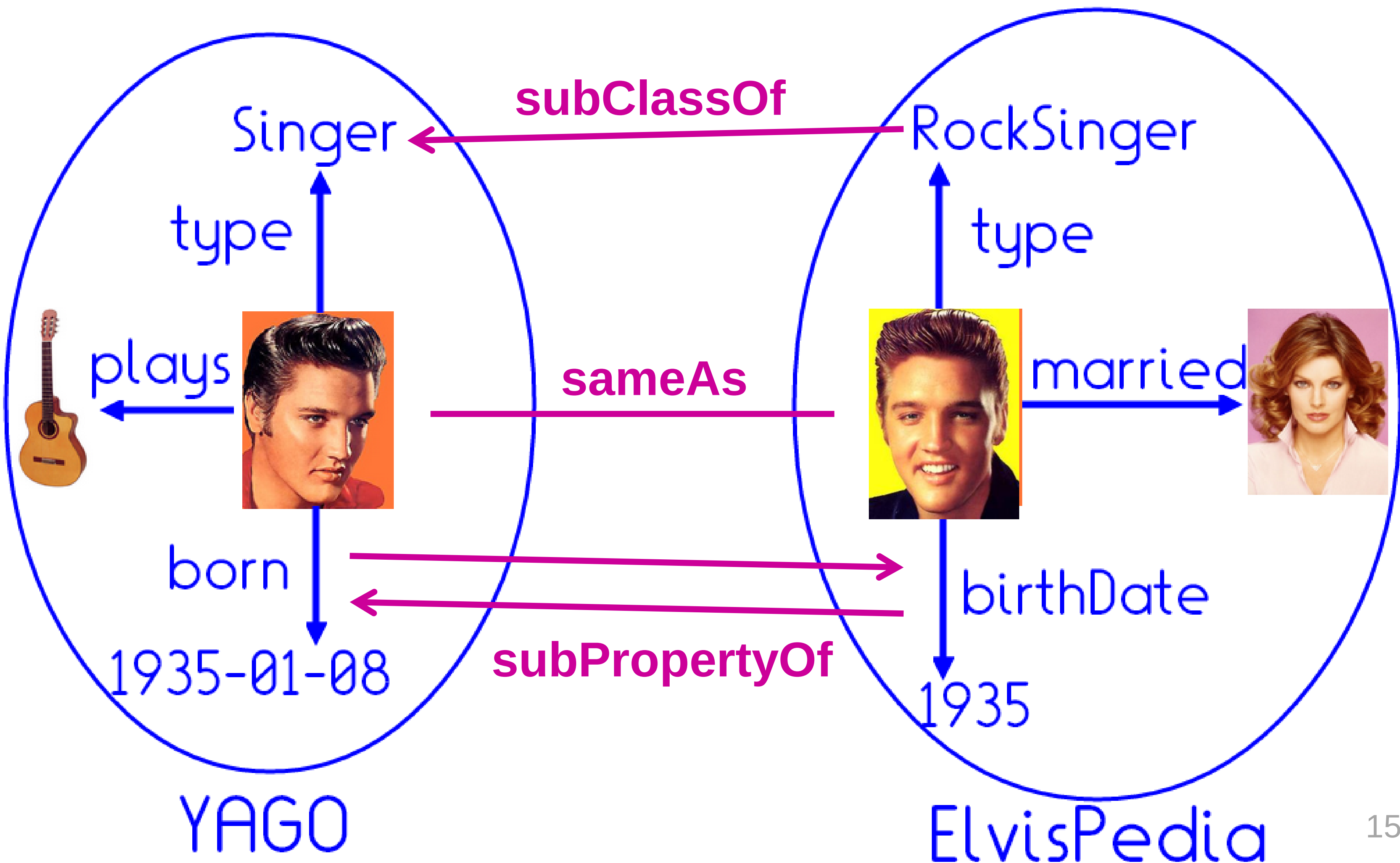
$$ifun(label) = 0.9$$

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 and Overview**
- ✓ **Taxonomic Knowledge:**
  Entities and Classes
- ✓ **Factual Knowledge:**
  Relations between Entities

- ✓ **Emerging Knowledge:**
  New Entities & Relations
- ✓ **Temporal Knowledge:**
  Validity Times of Facts
- ✓ **Contextual Knowledge:**
  Entity Name Disambiguation
- ✓ **Linked Knowledge:**
  Entity Matching
- ★ **Wrap-up**

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-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 extracting Relational Facts**

- **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

Knowledge Bases in the Big Data Era

Big Data Analytics

- Scalable algorithms
- Distributed platforms
- Discovering data sources
- Tapping unstructured data
- Connecting structured & unstructured data sources
- Making sense of heterogeneous, dirty, or uncertain data

Knowledge Bases: **entities**, **relations**, **time, space,** **…**

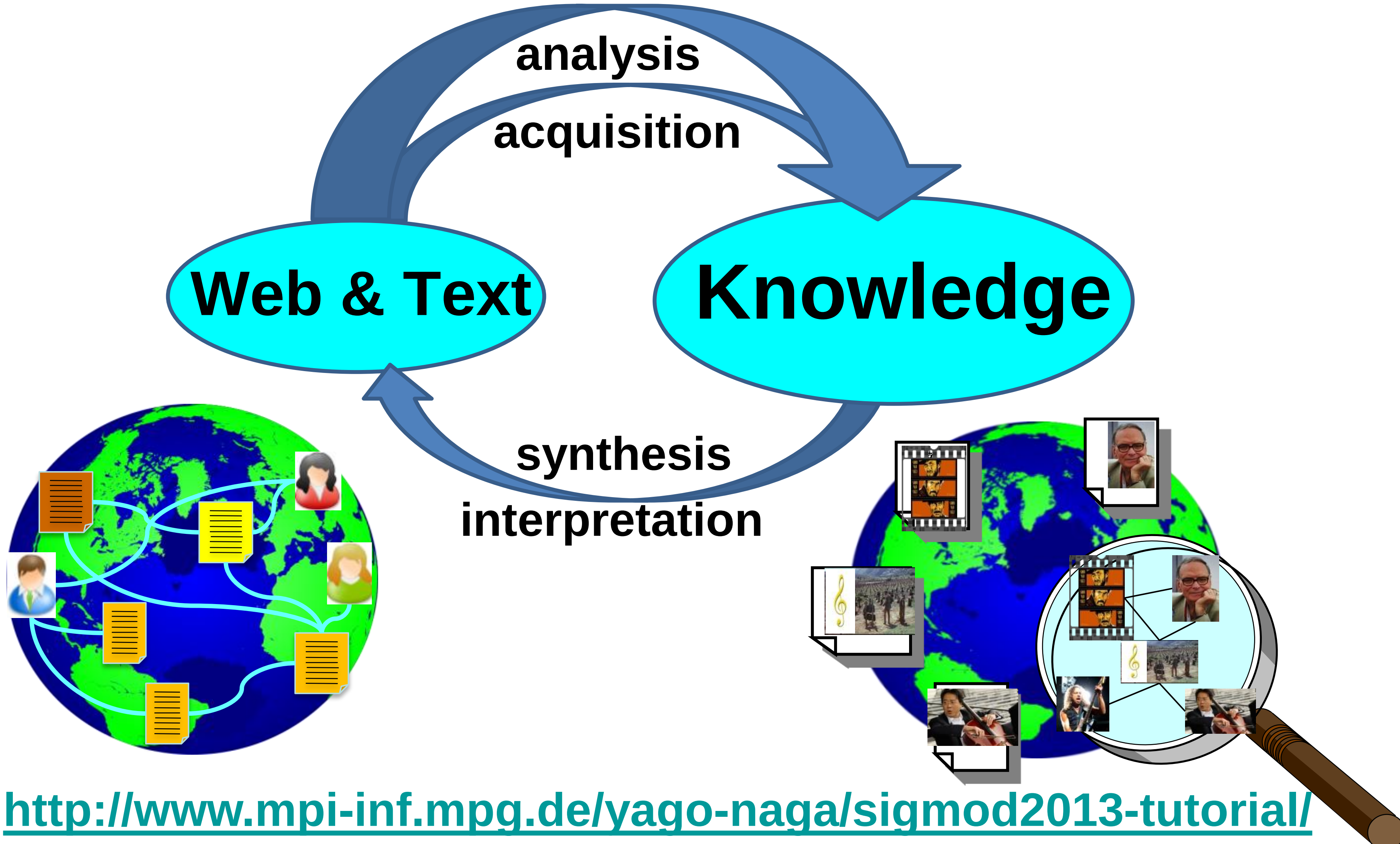
References

see comprehensive list in

Fabian Suchanek and Gerhard Weikum:
Knowledge Harvesting in the Big-Data Era,
Proceedings of the ACM SIGMOD
International Conference on Management of Data,
New York, USA, June 22-27, 2013,
Association for Computing Machinery, 2013.

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Take-Home Message: From Web & Text to Knowledge



http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/

Thank You !

http://www.mpi-inf.mpg.de/yago-naga/sigmod2013-tutorial/