Lecture 10.
Knowledge Base and Linked Data

Knowledge Harvesting in the Big-Data Era

The following materials selected from

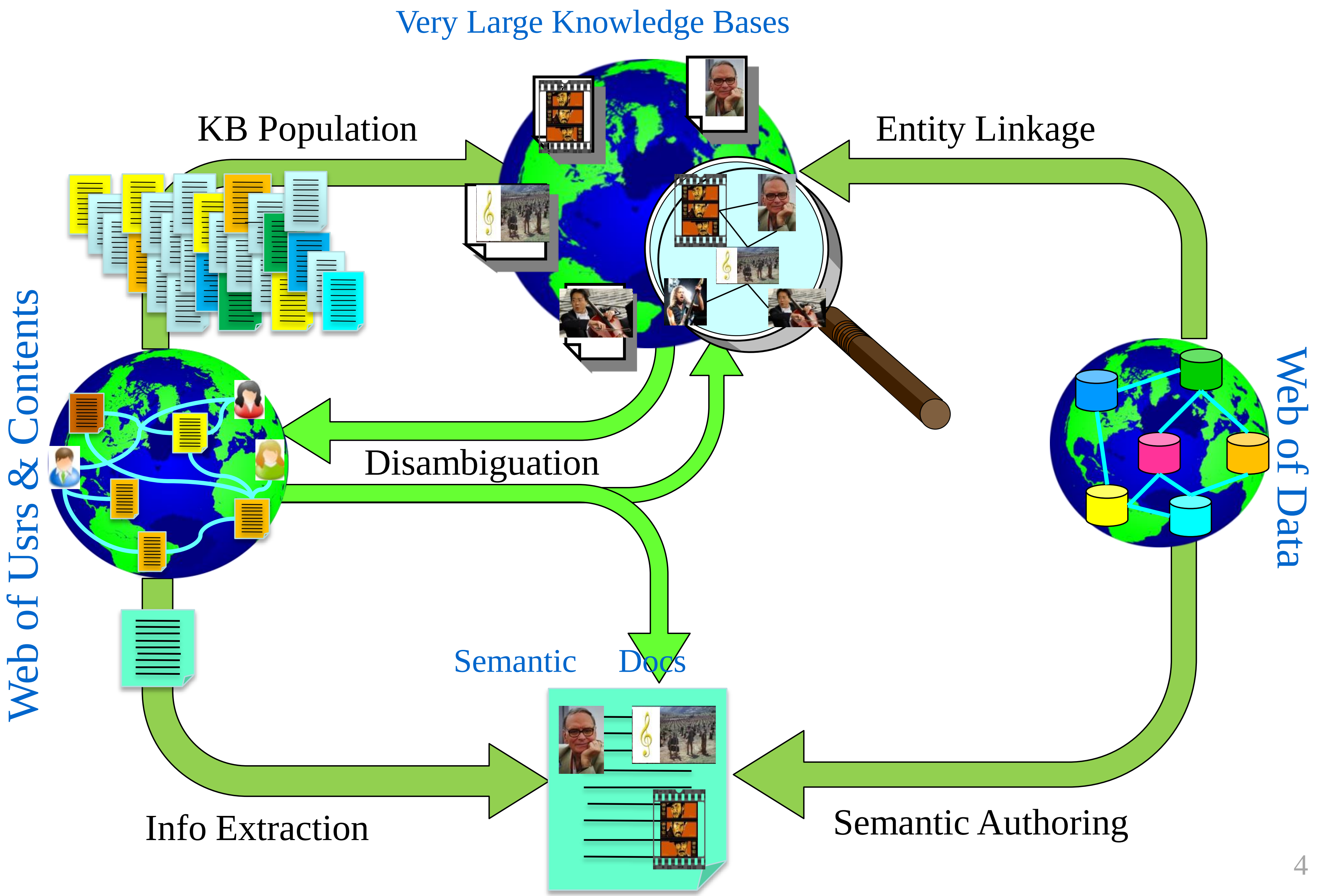
Fabian Suchanek & Gerhard Weikum

SIGMOD'13, June 22–27, 2013

Knowledge Base

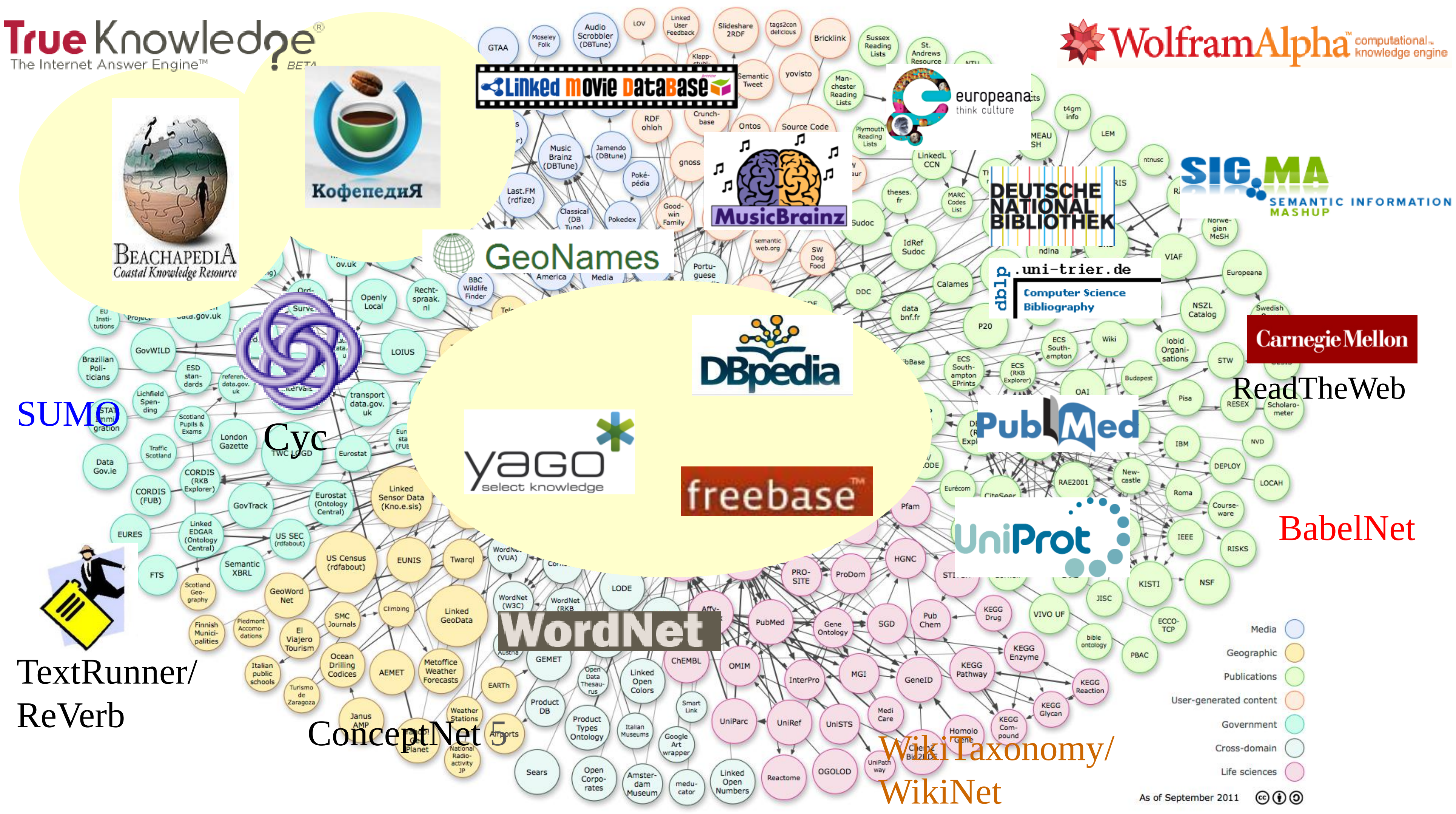
- Manually compiled knowledge collection
  - Cyc, WordNet, a variety of ontologies
- Publicly available resources
  - KnowItAll, ConceptNet, Dbpedia, Freebase, NELL, WikiTaxonomy, and YAGO
- Commercial interest
  - Google Knowledge Graph, EnitityCube/Renlifang at Microsoft Research, knowledge base in IBM’s Watson

Turn Web into Knowledge Base



Web of Data: RDF, Tables, Microdata

60 Bio. SPO (Subject-Predicate-Object) 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 (Subject-Predicate-Object) triples (RDF) and growing

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

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

- 25M entities in

顏尼歐・莫利克奈
義大利作曲家

Ennio_Morricone type composer
Ennio_Morricone type GrammyAwardWinner
composer subclassOf musician
Ennio_Morricone bornIn Rome
Rome locatedIn Italy
Ennio_Morricone created Ecstasy_of_Gold 黃金的狂喜 創作音樂出現於賽爾喬・萊翁內導演的1966年電影黃金三鏢客
Ennio_Morricone wroteMusicFor The_Good,_the_Bad_,and_the_Ugly
Sergio_Leone directed The_Good,_the_Bad_,and_the_Ugly 黃昏三鏢客

History of Knowledge Bases

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

Cyc and WordNet are hand-crafted knowledge bases

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

$\forall$ 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 project (1985-now)

George Miller **Christiane Fellbaum**

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 → Wikidata (www.wikidata.org)
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

Challenging Issues in Knowledge Harvesting

- Covering more entities beyond Wikipedia and discovering newly emerging entities
- Increasing the number of facts about entities and extracting more interesting relationship types in an open manner
- Capturing the temporal scope of relational facts
- Tapping into multilingual inputs such as Wikipedia editions in many different languages
- Extending fact-oriented knowledge bases with commonsense knowledge and (soft) rules
- Detecting and disambiguating entity mentions in natural-language text and other unstructured contents
- Large-scale sameAs linkage across many knowledge and data sources

Enabling Intelligent Applications

- Semantic search and question answering
  - Machine-readable encyclopedia are a rich source of answering expert-level questions in a precise and concise manner.
  - Interpreting users' information needs in terms of entities and relationships yields strong features for informative ranking of search results and entity-level recommendations over Web and enterprise data.
- Deep interpretation of natural language
  - Knowledge is the key to mapping surface phrases of written and spoken languages to their proper meanings.

Enabling Intelligent Applications

- Machine reading at scale
  - Users wish to obtain overviews of the salient entities and relationships for a week of news, a month of scientific articles, a year of political speeches, or a century of essays on a specific topic.
- Reasoning and smart assistants
  - Rich sets of facts and rules from a knowledge base enable computers to perform logical inferences in application contexts.
- Big-Data analytics over uncertain contents
  - Daily news, social media, scholarly publications, and other Web contents are the raw inputs for analytics to obtain insights on business, politics, health, and more.
  - Knowledge bases are key to discovering and tracking entities and relationships and thus making sense of noisy contents.

Use Case: Question Answering

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

Q: Sin City ? 萬惡城市 (美式漫畫改編的電影)

→ 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

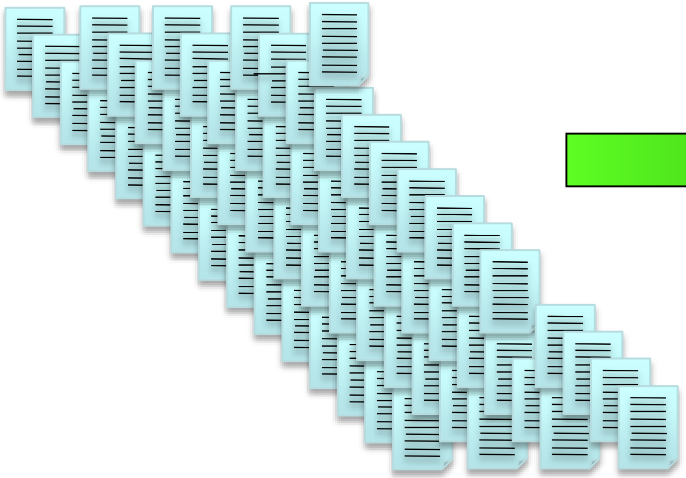
freebase

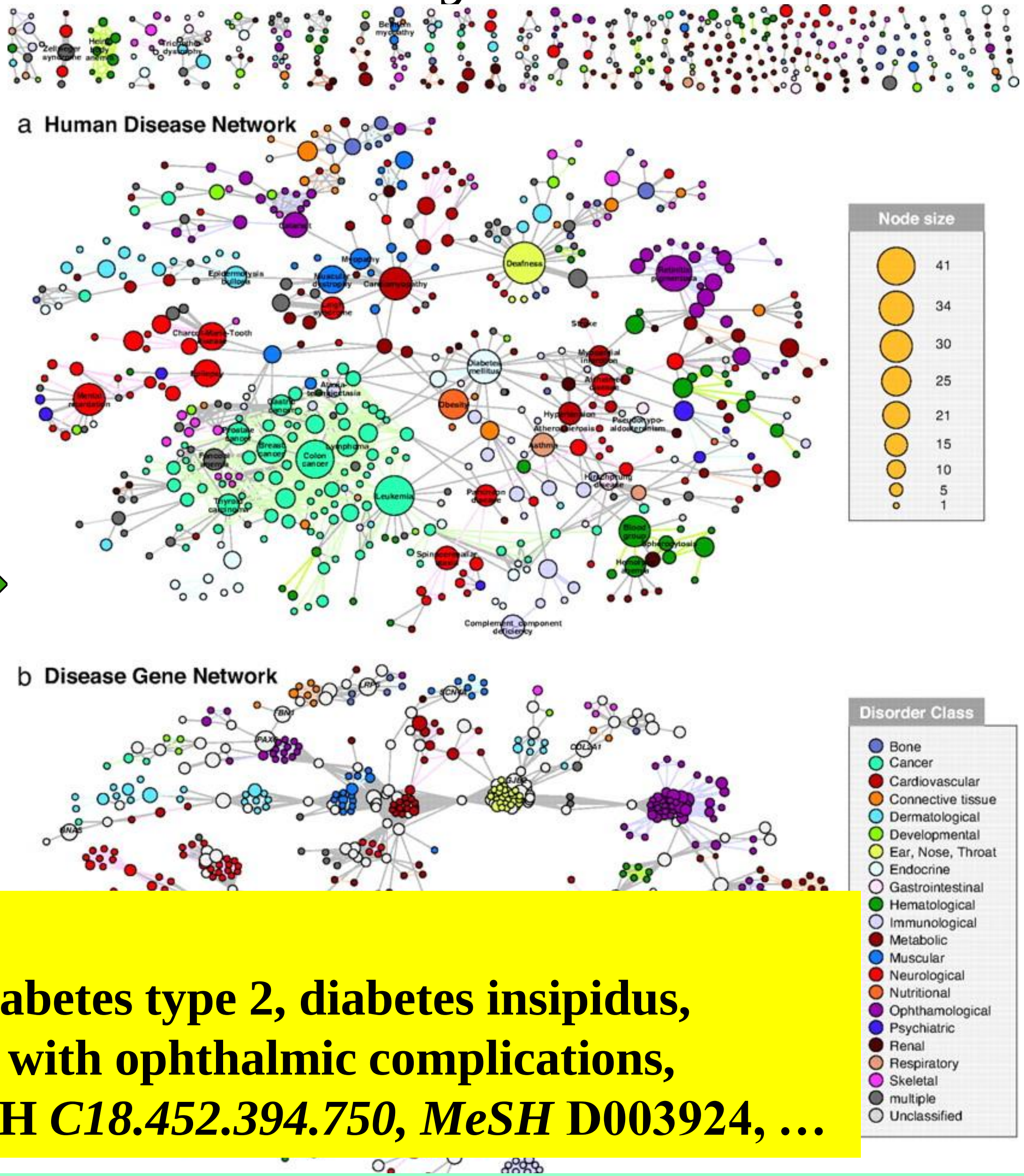
yago
select knowledge

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 Base Construction

- Knowledge Bases in the Big-Data Era
  - News, social media, web sites, and enterprise sources produce huge amounts of valuable contents in the form of text and speech.
  - Knowledge bases are a key asset for lifting unstructured contents into entity-relationship form and making the connection to structured data.
- Harvesting of Entities and Classes
  - Every entity in a knowledge base (such as Steve_Jobs) belongs to one or more classes (such as computer_pioneer).
  - Classes are organized into a taxonomy, where more special classes are subsumed by more general classes (such as person).

Knowledge Bases in the Big Data Era

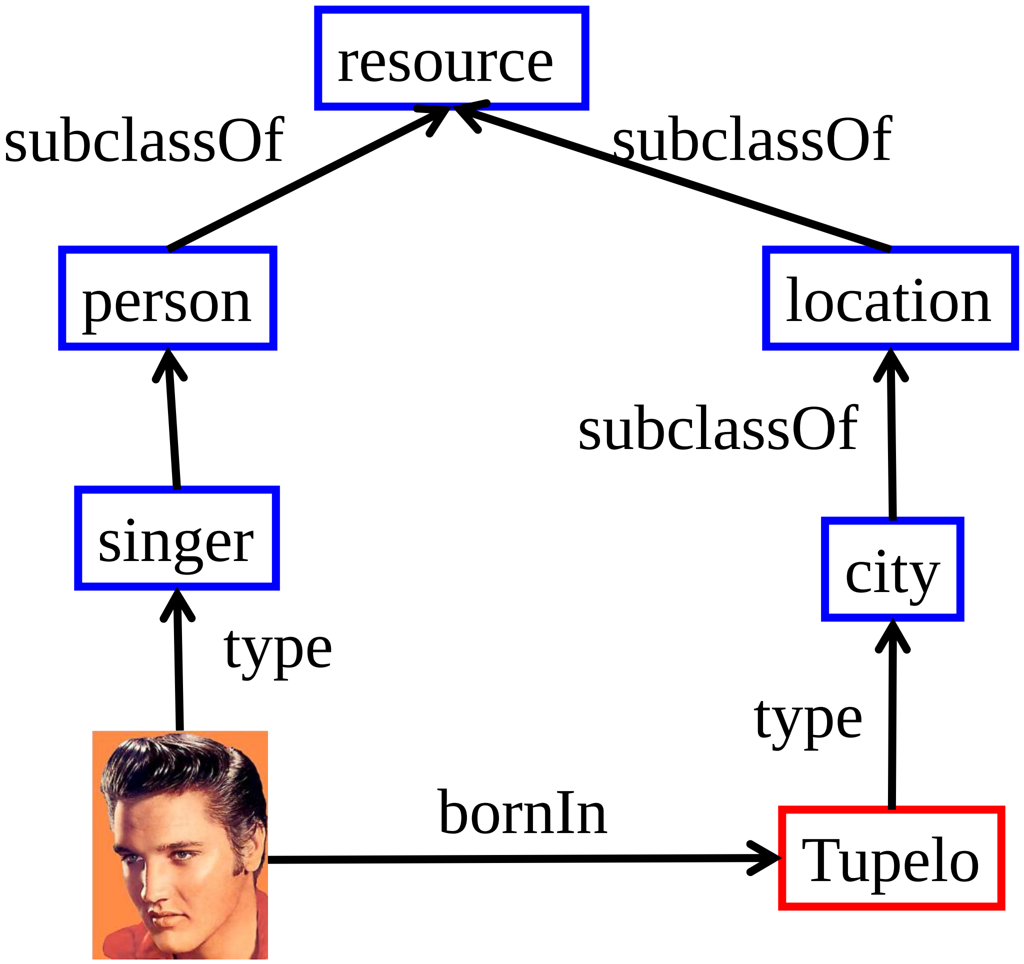
Big Data Analytics

- 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,
…

Taxonomic Knowledge: Entities and Classes

Knowledge Bases are labeled graphs



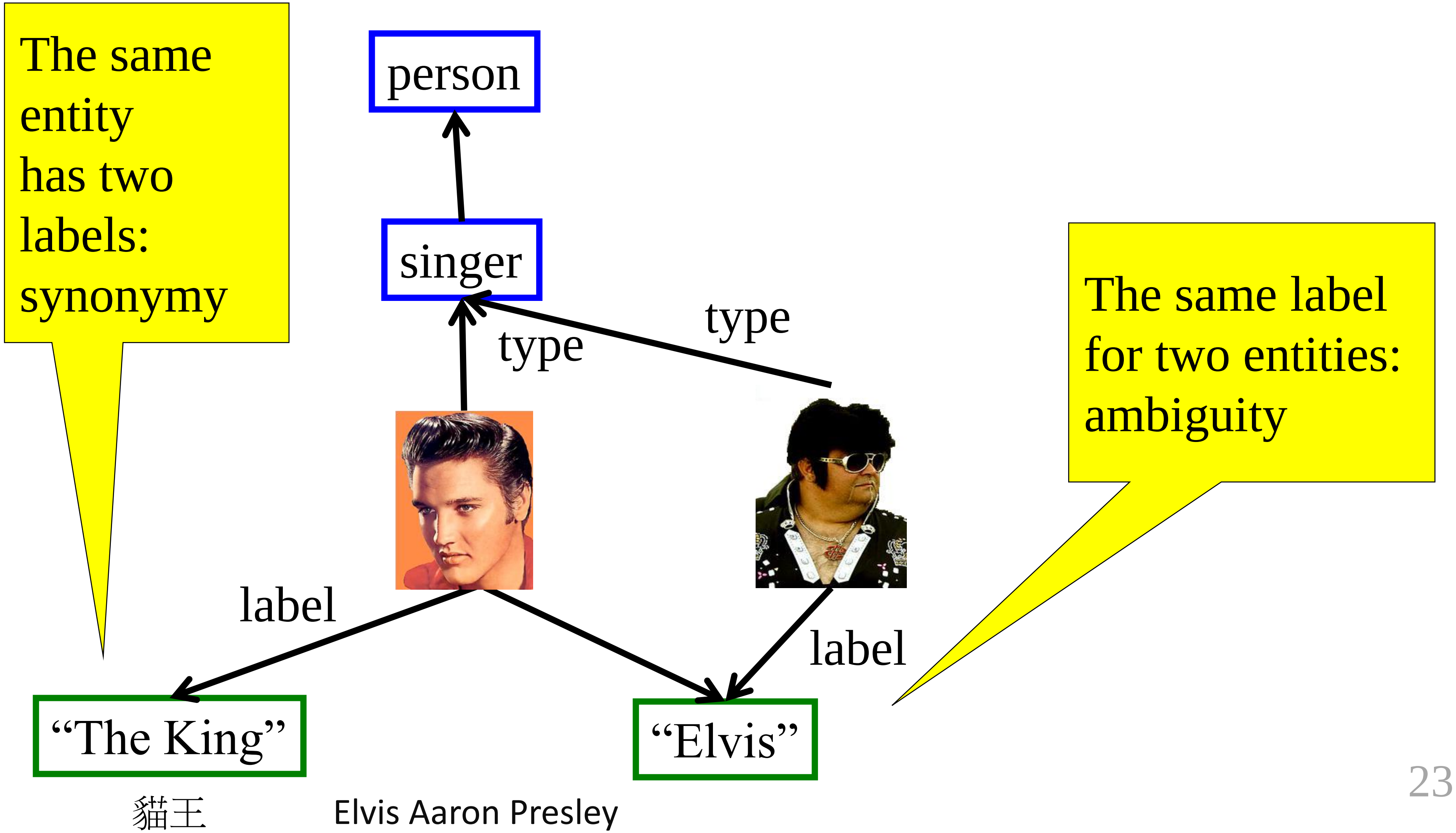
Classes/
Concepts/
Types

Relations/
Predicates

Instances/
entities

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

An entity can have different labels



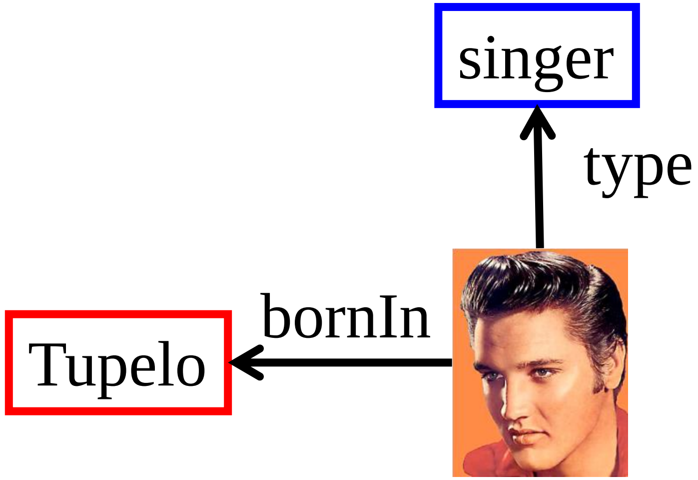
Different views of a knowledge base

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

Triple notation:

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

Graph notation:



Logical notation:

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

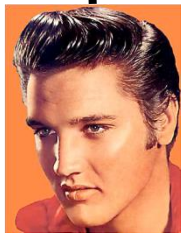
Our Goal is finding classes and instances

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

Noun

- S: (n) **person**, individual, someone, somebody, mortal, soul (a human being) *"there was too much for one person to do"*
- S: (n) **person** (a human body (usually including the clothing)) *"a weapon was hidden on his person"*
- S: (n) **person** (a grammatical category used in the classification of pronouns, possessive determiners, and verb forms according to whether they indicate the speaker, the addressee, or a third party) *"stop talking about yourself in the third person"*

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.

HARVESTING FACTS AT WEB SCALE

- Harvesting Relational Facts
  - Steve Jobs
    - Steve_Jobs founded Apple_Inc.
    - Steve_Jobs was_Board_Member_of Walt_Disney_Company
    - Steve_Jobs died_on 5-Oct-2011
    - Steve_Jobs died_of Pancreas_Cancer
    - Steve_Jobs has_Friend Joan_Baez, and more
  - Information sources
    - Web data, tapping both semistructured sources like Wikipedia infoboxes, lists, and tables
    - natural-language text sources like Wikipedia full-text articles, news and social media
  - Methods
    - pattern matching (e.g., regular expressions)
    - computational linguistics (e.g., dependency parsing)
    - Statistical learning (e.g., factor graphs and Markov Logic Network (MLN’s))
    - logical consistency
    - reasoning

Web-based Methods

Hearst patterns extract instances from text

[M. Hearst 1992]

Goal: find instances of classes

Hearst defined lexico-syntactic patterns for type relationship:

X such as Y; X like Y;
X and other Y; X including Y;
X, especially Y;

Find such patterns in text: //better with POS tagging

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

Derive type(Y,X)

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

Recursively applied patterns increase recall

[Kozareva/Hovy 2010]

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

X such as Y

companies such as Apple
companies such as Google

Y like Z
*, Y and Z

Apple like Microsoft offers
IBM, Google, and Amazon

Y like Z
*, Y and Z

Microsoft like SAP sells
eBay, Amazon, and Facebook

Y like Z
*, Y and Z

Cherry, Apple, and Banana

potential problems with ambiguous words

Doubly-anchored patterns are more robust

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

Goal:

find instances of classes

Start with a set of seeds:

companies = {Microsoft, Google}

Parse Web documents and find the pattern

W, Y and Z

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

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

Cherry, Apple, and Banana → X

Instances can be extracted from tables

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

Goal: find instances of classes

Start with a set of seeds:

cities = {Paris, Shanghai, Brisbane}

Parse Web documents and find tables

Paris	France
Shanghai	China
Berlin	Germany
London	UK

Paris	Iliad
Helena	Iliad
Odysseus	Odysee
Rama	Mahabaratha

If at least two seeds appear in a column, harvest the others:

type(Berlin, city)
type(London, city)

X

Extracting instances from lists & tables

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

State-of-the-Art Approach (e.g. SEAL):

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

Caveats:

Precision drops for classes with sparse statistics (IR profs, …)

Harvested items are names, not entities

Canonicalization (de-duplication) unsolved

Probase builds a taxonomy from the Web

Use Hearst liberally to obtain many instance candidates:

„plants such as trees and grass“
„plants include water turbines“
„western movies such as The Good, the Bad, and the Ugly“

Problem: signal vs. noise

Assess candidate pairs statistically:

$P[X|Y] >> P[X^*|Y] \quad \rightarrow \quad$ 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]

Probase: Using the World as its Model

Knowledge in Probase is harnessed from billions of web pages and years worth of search logs -- these are nothing more than the digitized footprints of human communication. In other words, Probase uses the world as its model.

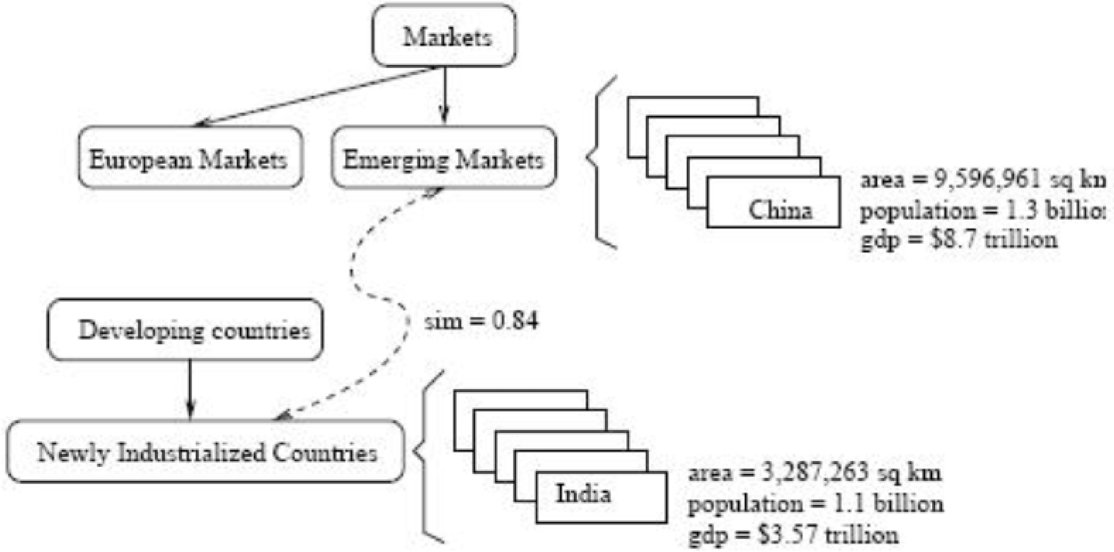


Figure 1: A snippet of Probase's core taxonomy

Figure 1 shows what is inside Probase. The knowledgebase consists of ***concepts*** (e.g. emerging markets), ***instances*** (e.g., China), ***attributes*** and ***values*** (e.g., China's population is 1.3 billion), and ***relationships*** (e.g., emerging markets, as a concept, is closely related to newly industrialized countries), all of which are automatically derived in an unsupervised manner.

But Probase is much more than a traditional ontology/taxonomy. Probase is unique in two aspects. First, **Probase has an extremely large concept/category space (2.7 million categories).** As these concepts are automatically acquired from web pages authored by millions of users, it is probably true that they cover most concepts in our mental world (about worldly facts). Second, **data in Probase, as knowledge in our mind, is not black or white. Probase quantifies the**

http://research.microsoft.com/en-us/projects/probase/

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

Y: instance X_1, X_2, X_3: type

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

using documents   → score1(X) ~ freq(X,Y) * #distinctPatterns(X,Y)

3   2

Tree, plant
Tree, plant, water
Tree, plant

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

using queries   → score2(X) ~ $(\prod_{i=1..N} \text{term-score}(t_i \in X))^{1/N}$

example query: "Michael Jordan computer scientist"

Y t_1 t_2 X

computer scientist
…
Michael Jordan
…

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

using query sessions   → score3(X) ~ $(\prod_{i=1..N} \text{term-session-score}(t_i \in X))^{1/N}$

computer: 1 scientist: 2

scientist
…
Michael Jordan
…

Open Information Extraction

- Open IE harvests arbitrary subject-predicate-object triples from natural-language documents
  - Noun phrases: entity candidates
  - Verbal phrases: prototypic patterns for relations
- hasWonPrize relation
  - “candidate for . . . prize”
  - “expected to win . . . prize”

Factual Knowledge:
Relations between Entities

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

- **Semi-structured data**

  "Low-Hanging Fruit" 唾手可得

  - Wikipedia infoboxes & categories
  - HTML lists & tables, etc.

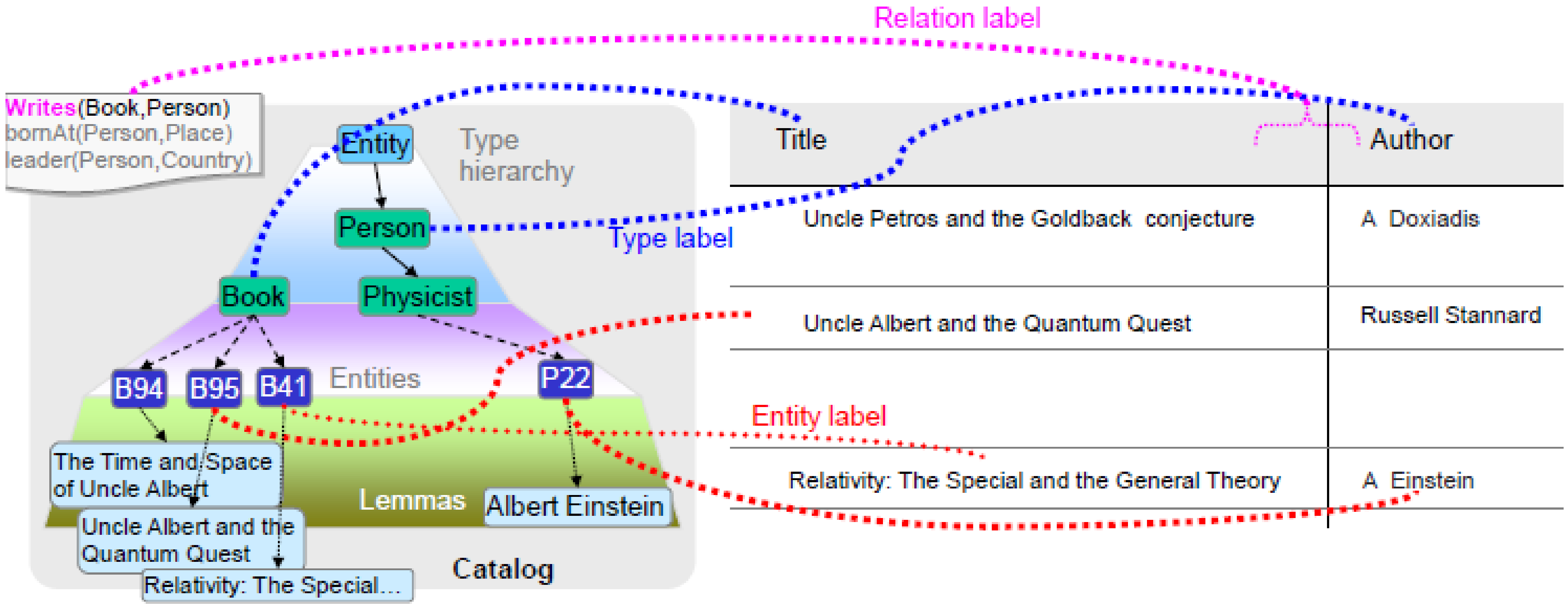
- **Free text**

  "Cherrypicking" 最佳選擇

  - Hearst patterns & other shallow NLP
  - Iterative pattern-based harvesting
  - Consistency reasoning

- **Web tables**

Annotating Tables with Entity, Type, and Relation links



from Sunita Sarawagi

Source-centric IE vs. Yield-centric IE

Source-centric IE

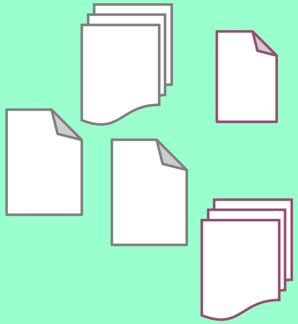
Surajit obtained his PhD in CS from Stanford ...

one source

1) recall !
2) precision

Document 1:
instanceOf (Surajit, scientist)
inField (Surajit, c.science)
almaMater (Surajit, Stanford U)
…

Yield-centric IE
量



many sources

+ (optional) targeted relations

1) precision !
2) recall

hasAdvisor

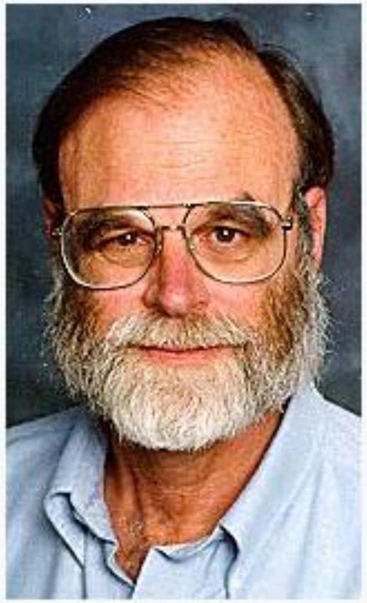
Student	Advisor
Surajit Chaudhuri	Jeffrey Ullman
Jim Gray	Mike Harrison
…	…

worksAt

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

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	

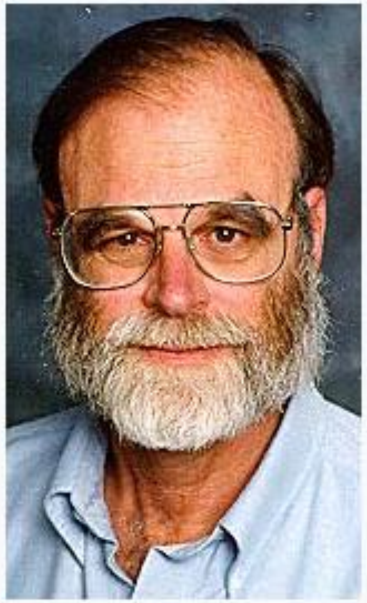
Joseph M. Hellerstein

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



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

```
{{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}}
```

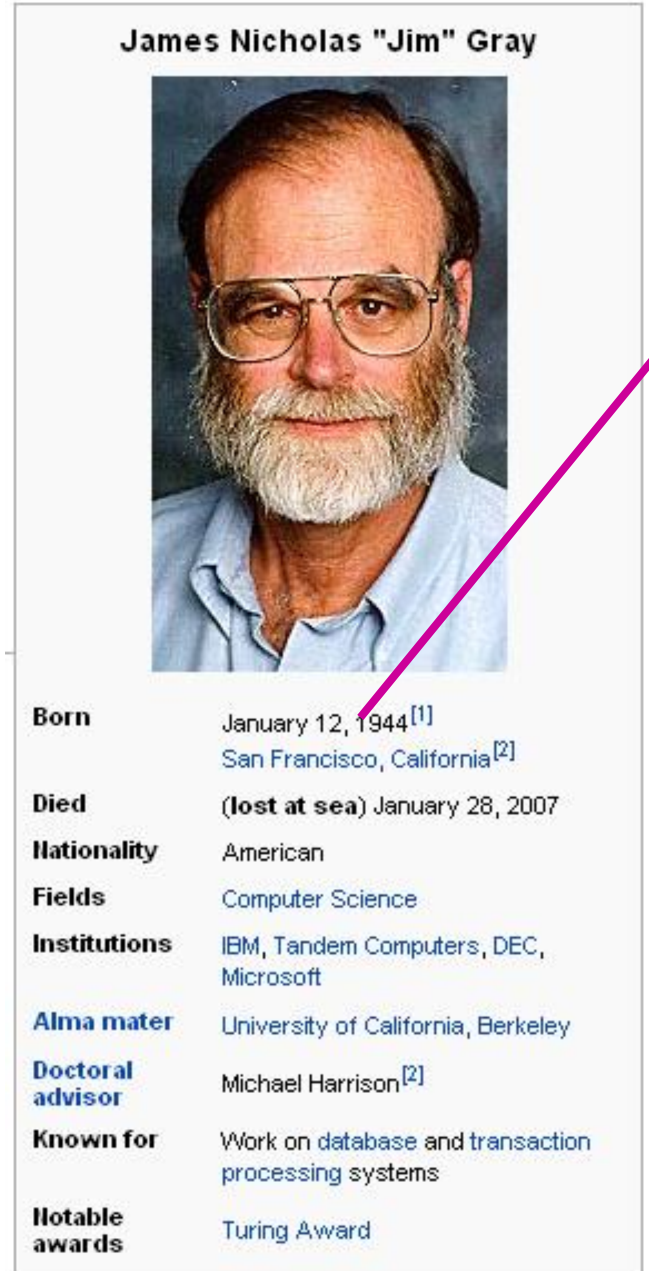
Map attribute to canoncial, predefined relation (manually or crowd-sourced)

Extract data item by regular expression

wasBorn 1944-01-12

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

Learn how articles express facts



James "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

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

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]

Facts yield patterns – and vice versa

Facts & *Fact Candidates*

(JimGray, MikeHarrison)

(BarbaraLiskov, JohnMcCarthy)

(Surajit, Jeff)

(Alon, Jeff)

(Sunita, Mike)

(Renee, Yannis)

(Sunita, Soumen)

(Soumen, Sunita)

(Surajit, Moshe)

(Alon, Larry)

(Surajit, Microsoft)

Patterns

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, …)

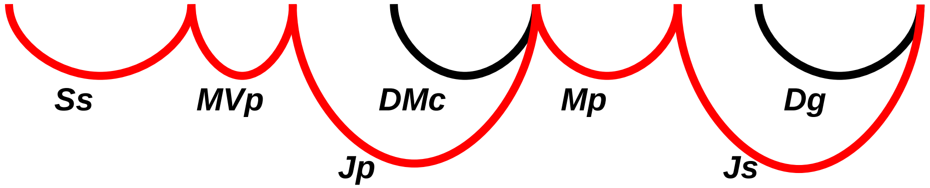
Problem: Surface patterns fail if the text shows variations

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

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] • *Maléna* – 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

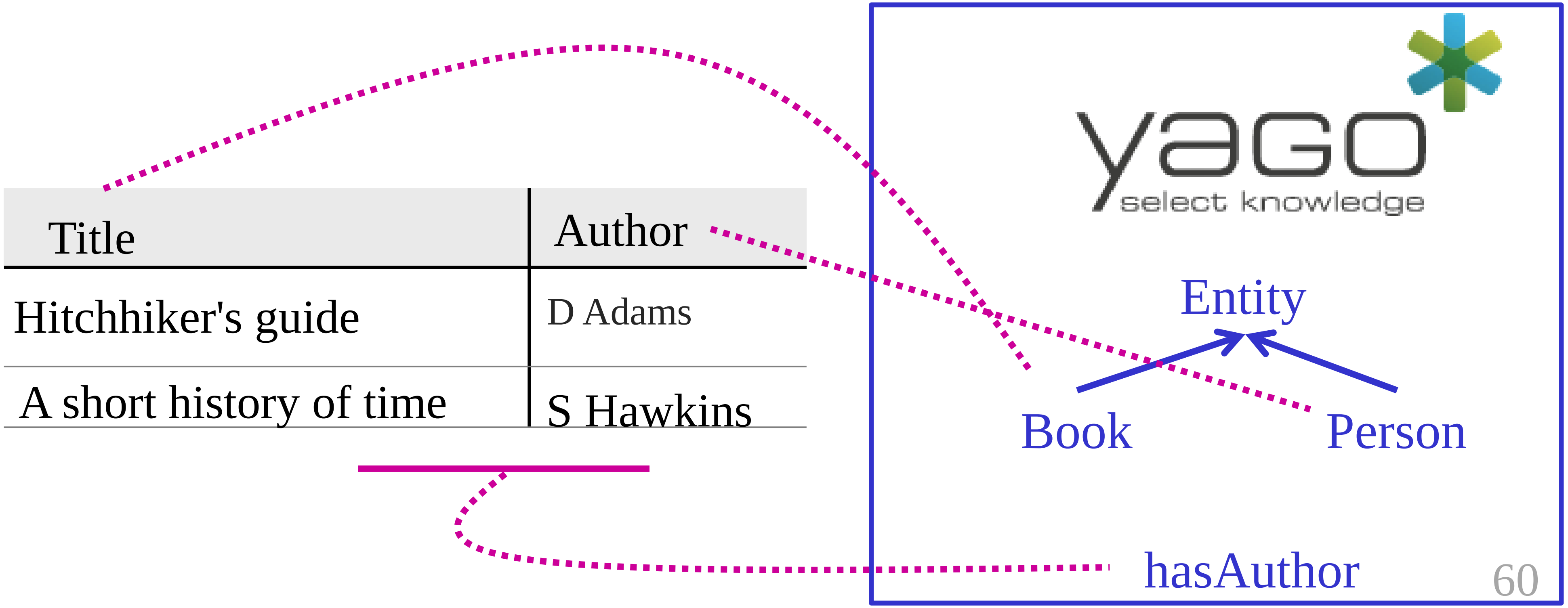
Web Tables can be annotated with YAGO

[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



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

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

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, …

KNOWLEDGE FOR BIG DATA

Emerging Knowledge: New Entities & Relations

Discovering "Unknown" Knowledge

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

< CarlaBruni marriedTo NicolasSarkozy> $\in$ **Person × R × Person**
< NataliePortman wonAward AcademyAward > $\in$ **Person × R × 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 …

Temporal Knowledge:
Validity Times of Facts

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]

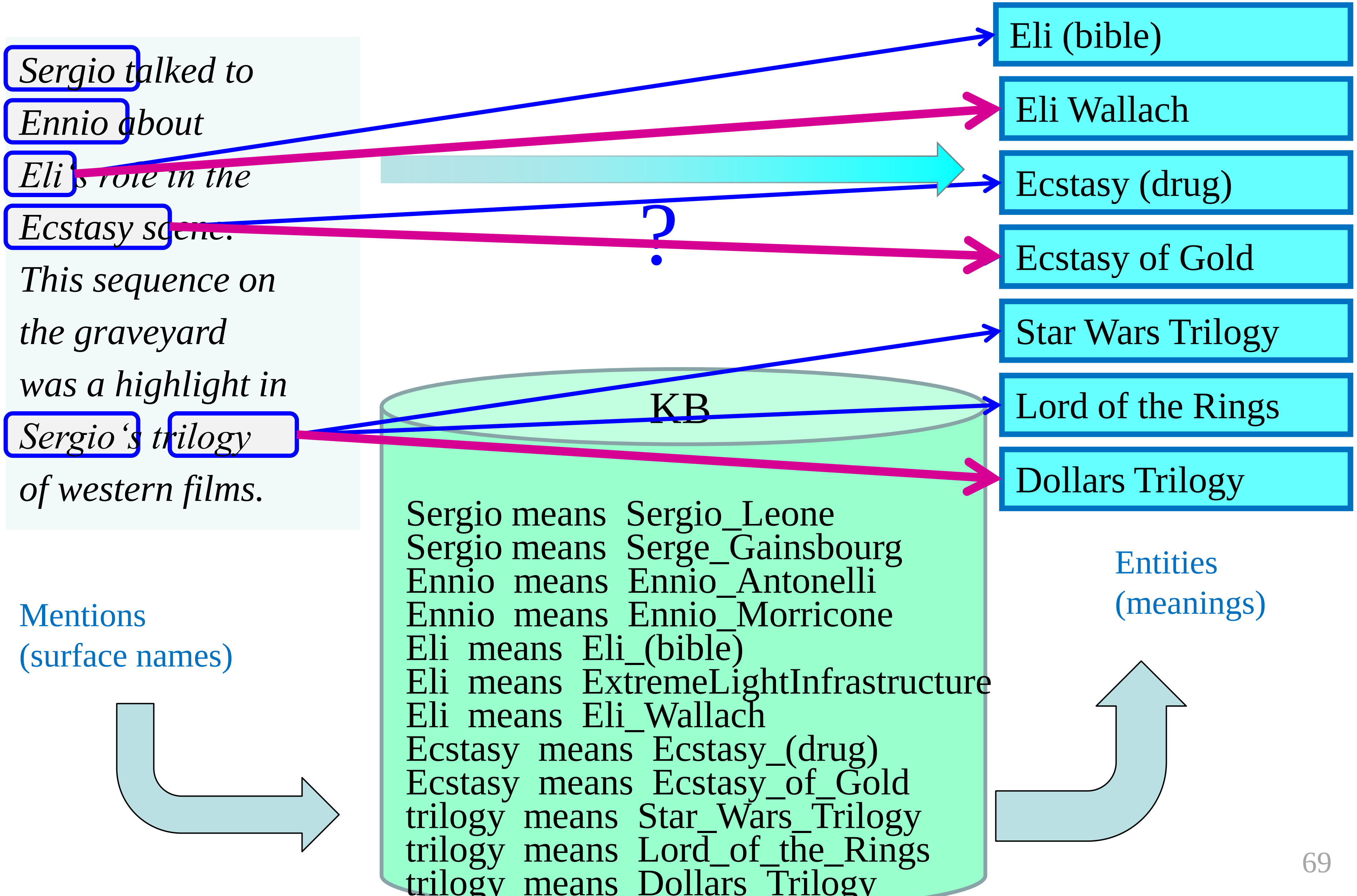
How can we query & reason on entity-relationship facts
in a "time-travel" manner - with uncertain/incomplete KB ?

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

Named-Entity Disambiguation

- Entity mentions are just noun phrases and still ambiguous.
- Mapping mentions to canonicalized entities registered in a knowledge base is the task of named-entity disambiguation (NED).
- NED is a special case of the general word-sense disambiguation problem.
- State-of-the-art NED methods combine context similarity between the surroundings of a mention and salient phrases associated with an entity.

Named Entity Disambiguation



Entity Linkage

- More and more structured data on theWeb, in the form of (HTML) tables, microdata embedded inWeb pages (using, e.g., the schema.org vocabulary), and Linked Open Data.
- For knowledge bases and Linked Open Data, it is of particular interest because of the need for generating and maintaining owl:sameAs linkage across knowledge resources.

MicroData

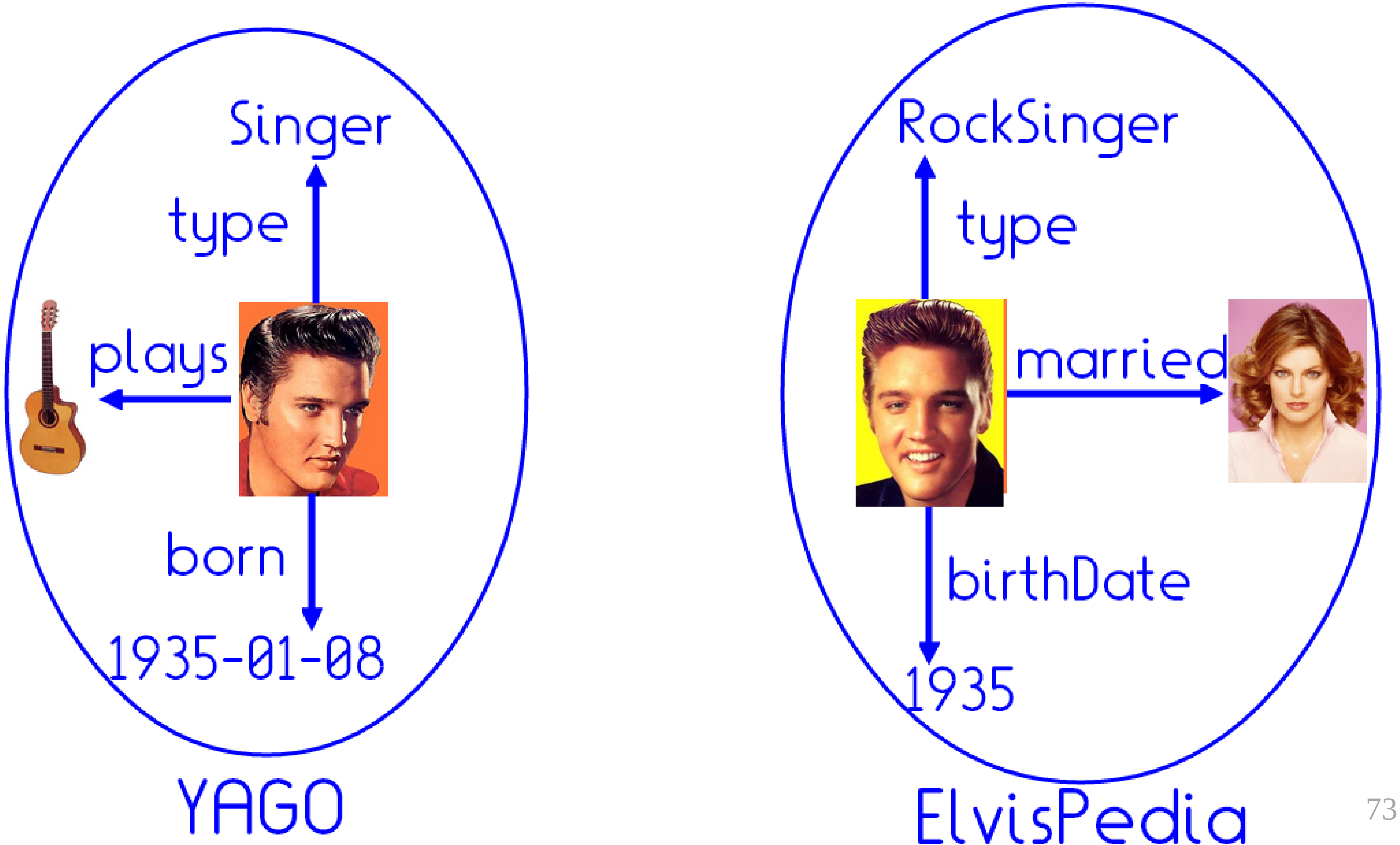
```
<div>
 <h1>Avatar</h1>
 <span>Director: James Cameron (born August 16, 1954)</span>
 <span>Science fiction</span>
 <a href="../movies/avatar-theatrical-trailer.html">Trailer</a>
</div>
```

```
<div itemscope itemtype="http://schema.org/Movie">
  <h1>Avatar</h1>
  <span>Director: James Cameron (born August 16, 1954)</span>
  <span>Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html">Trailer</a>
</div>
```

https://schema.org/docs/gs.html

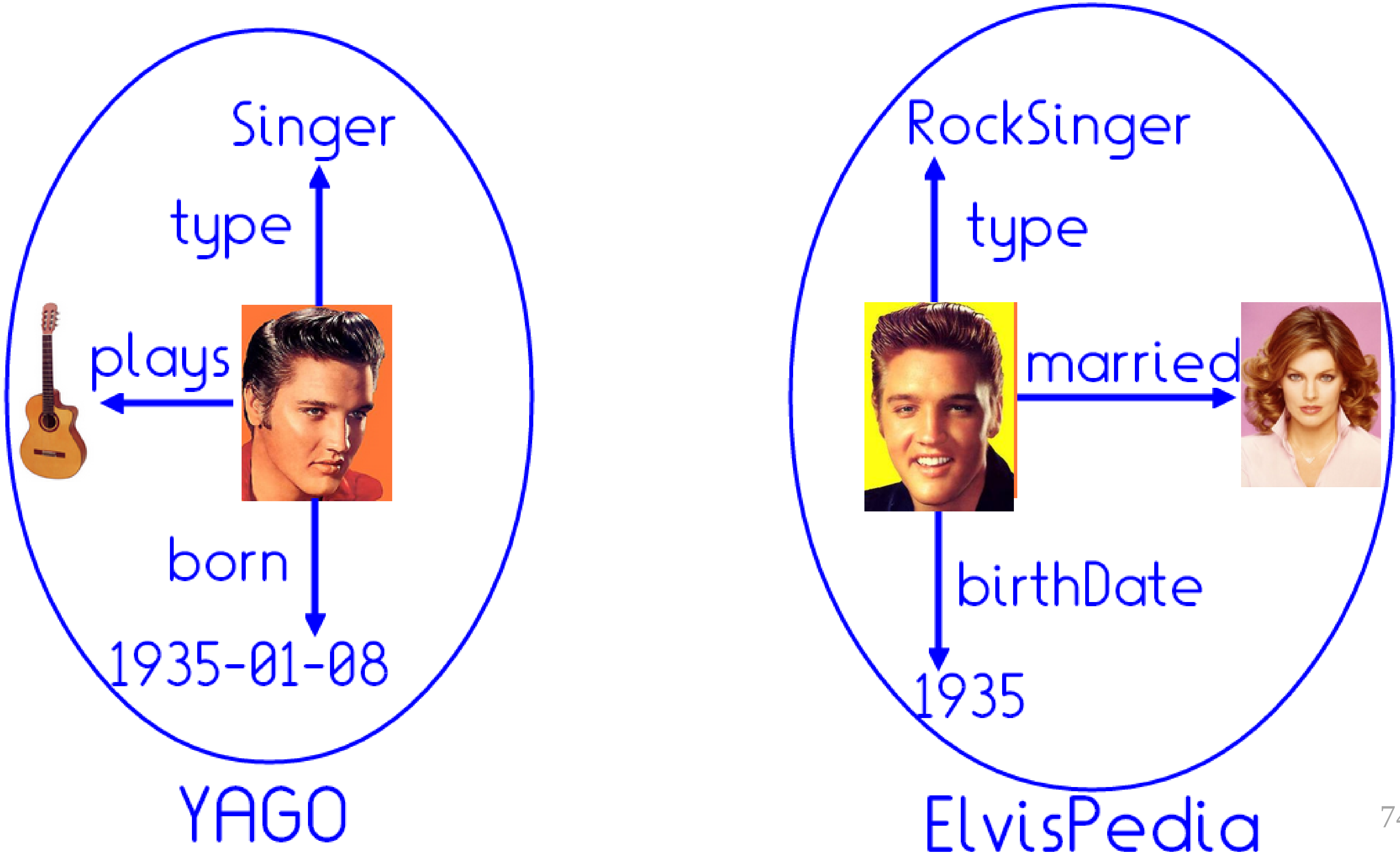
```
<div itemscope itemtype ="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <span>Director: <span itemprop="director">James Cameron</span>
                                    (born August 16, 1954)</span>
  <span itemprop="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" itemprop="trailer">Trailer</a>
</div>
```

Knowledge bases are complementary

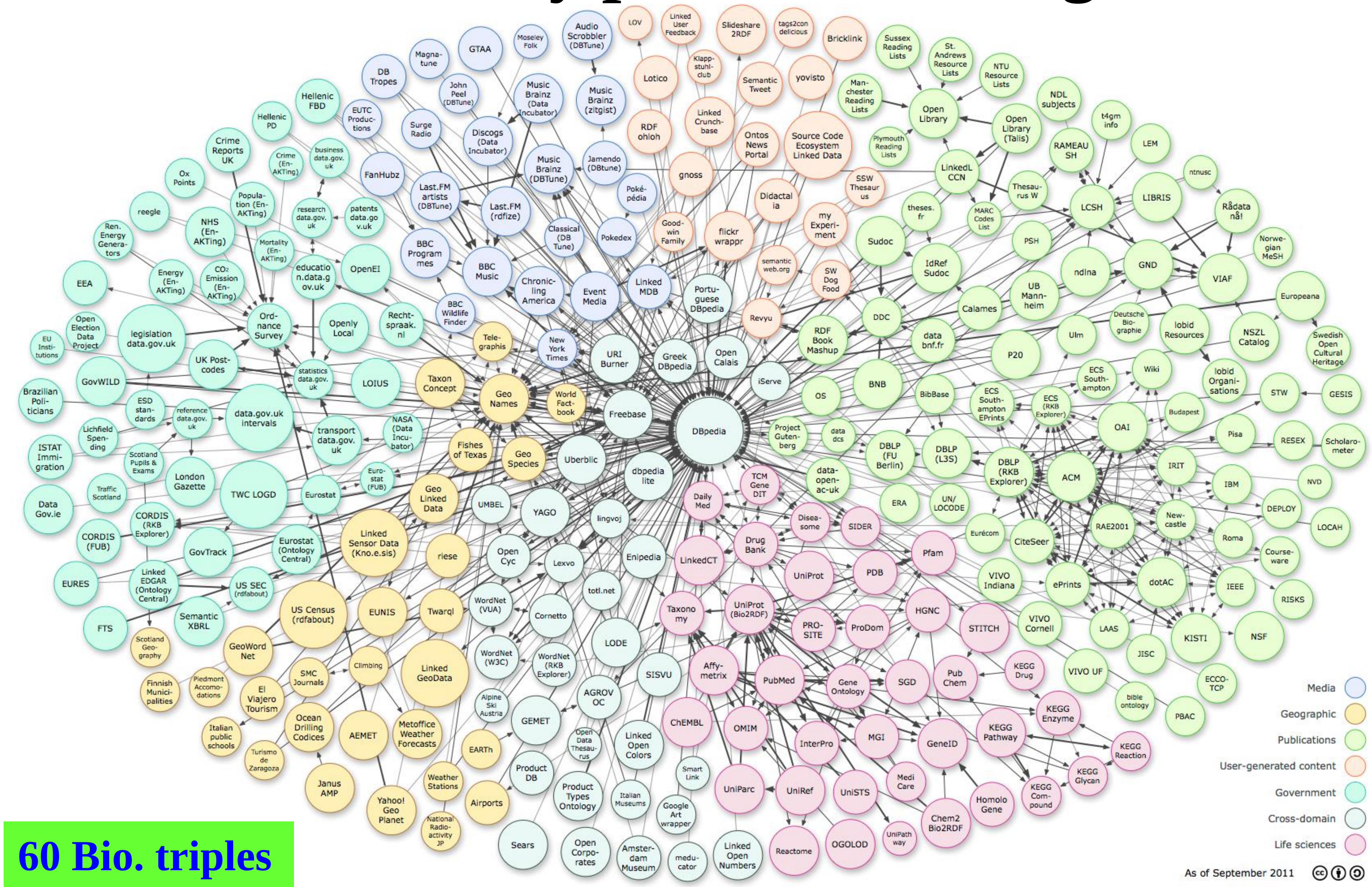


No Links ⇒ No Use

Who is the spouse of the guitar player?



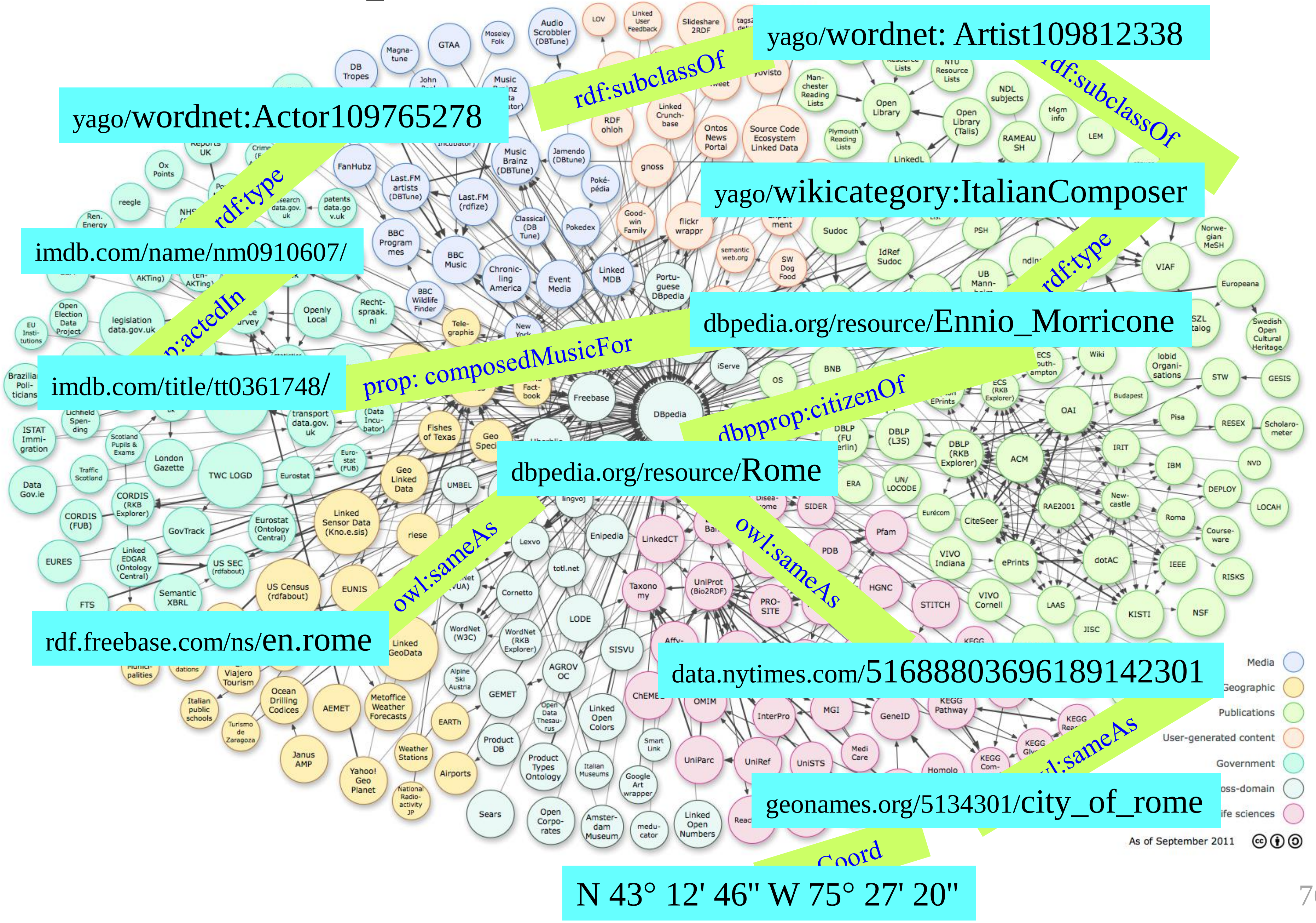
There are many public knowledge bases



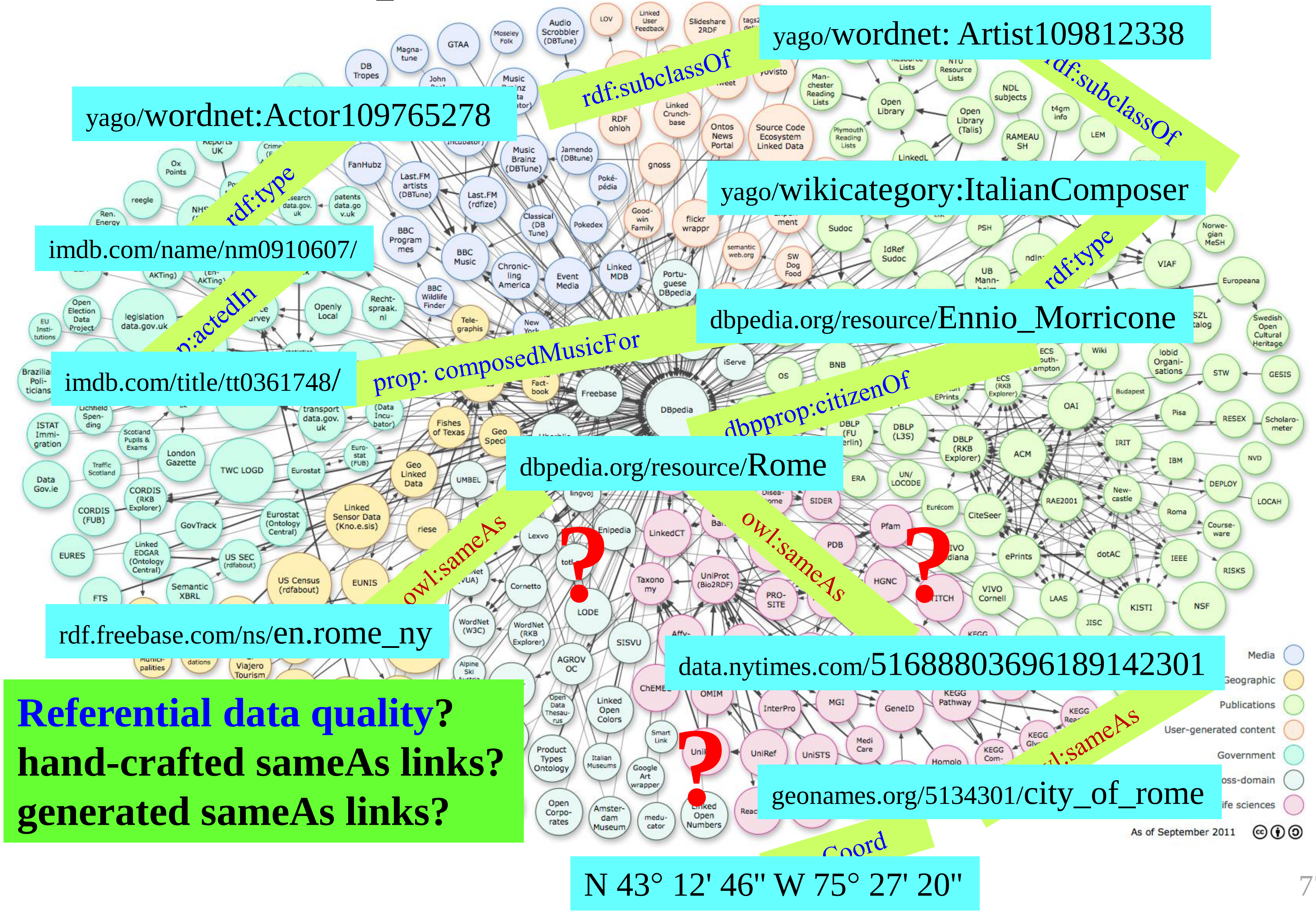
60 Bio. triples
500 Mio. links

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

Link equivalent entities across KBs

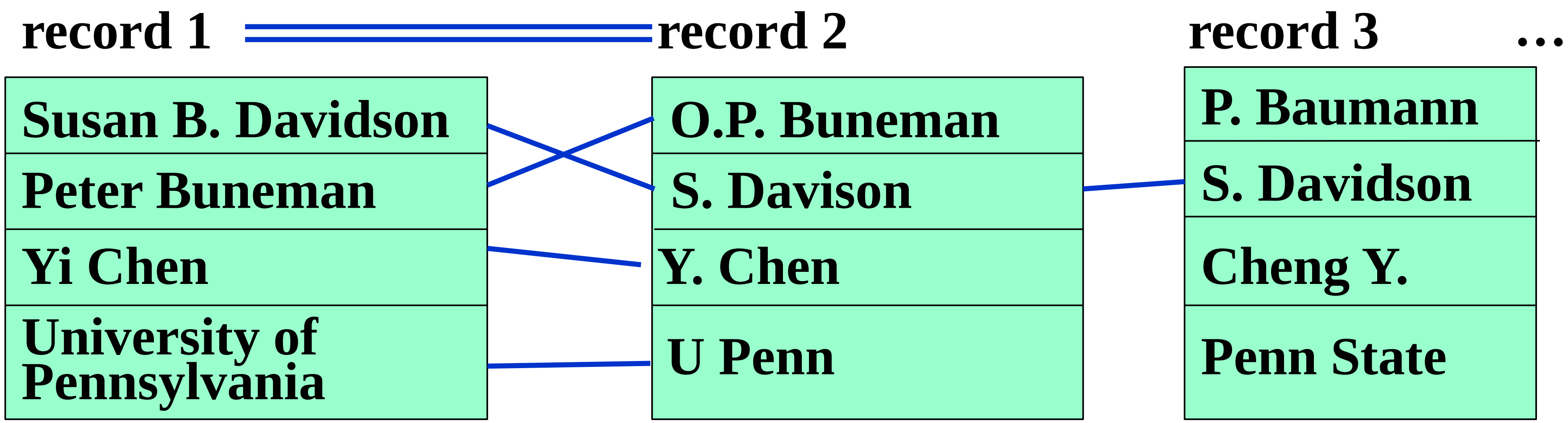


Link equivalent entities across KBs



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.)

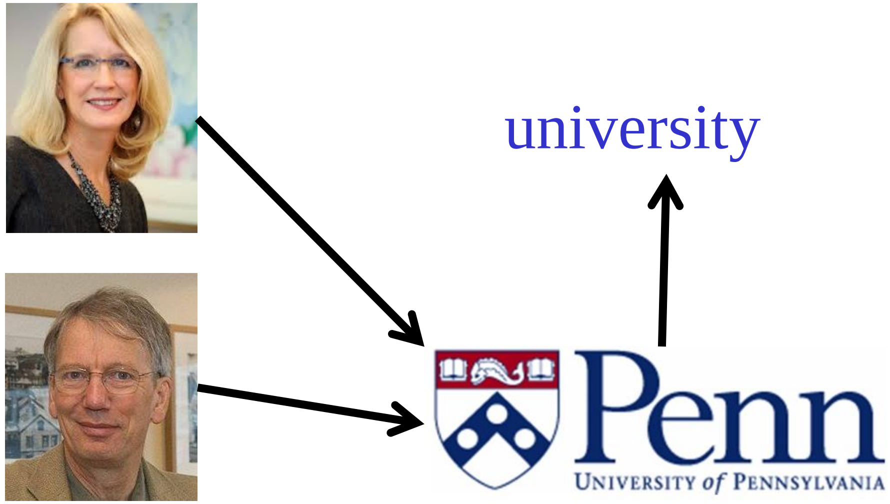
Halbert L. Dunn: Record Linkage. American Journal of Public Health. 1946
H.B. Newcombe et al.: Automatic Linkage of Vital Records. Science, 1959.
I.P. Fellegi, A.B. Sunter: A Theory of Record Linkage. J. of American Statist. Soc., 1969.

Linking Records vs. Linking Knowledge

record

Susan B. Davidson
Peter Buneman
Yi Chen
University of Pennsylvania

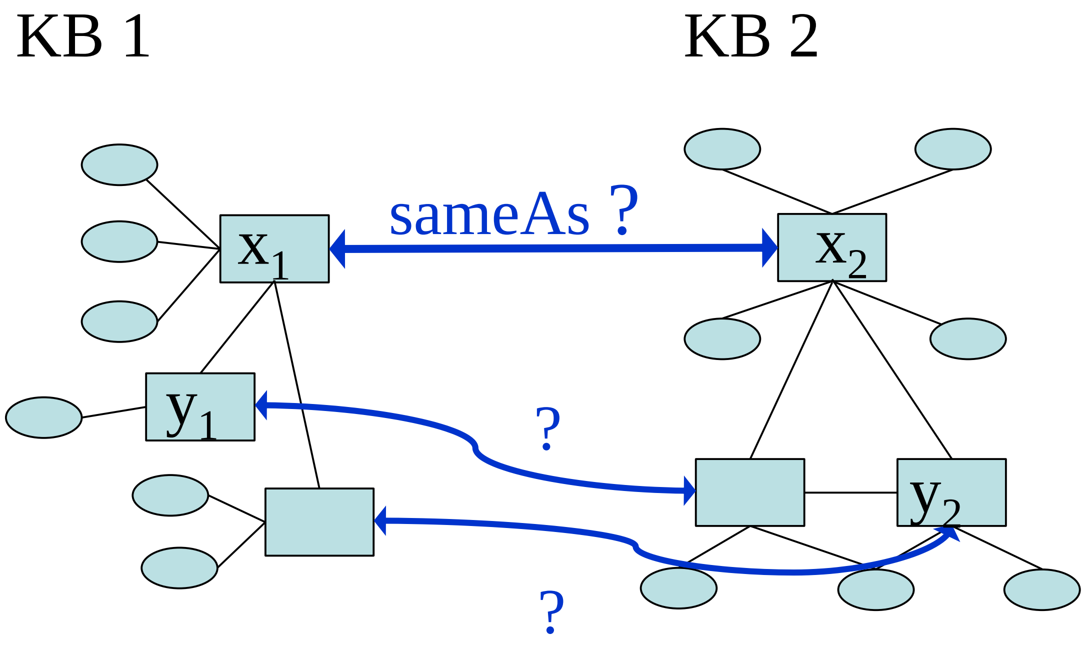
KB / Ontology



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)