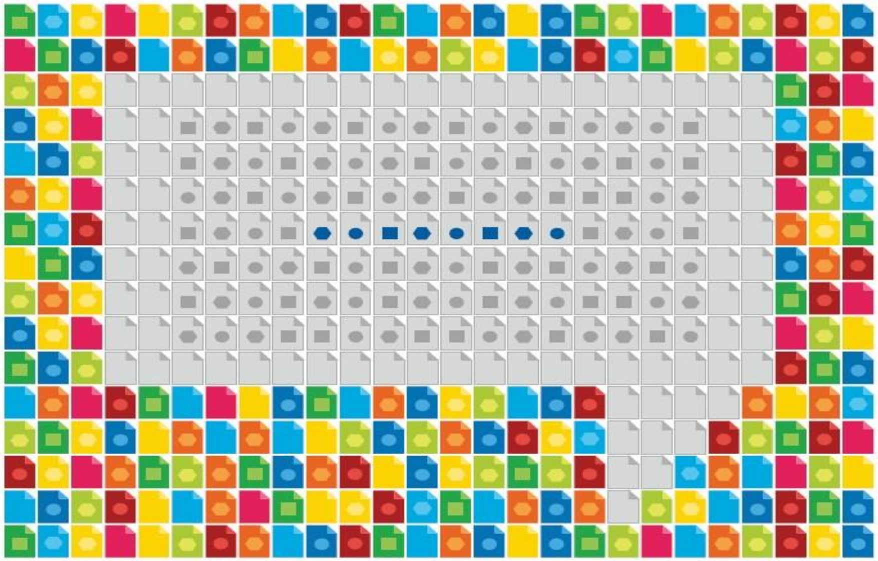
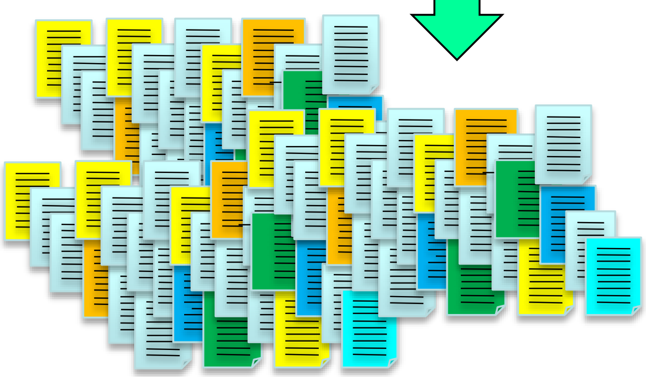
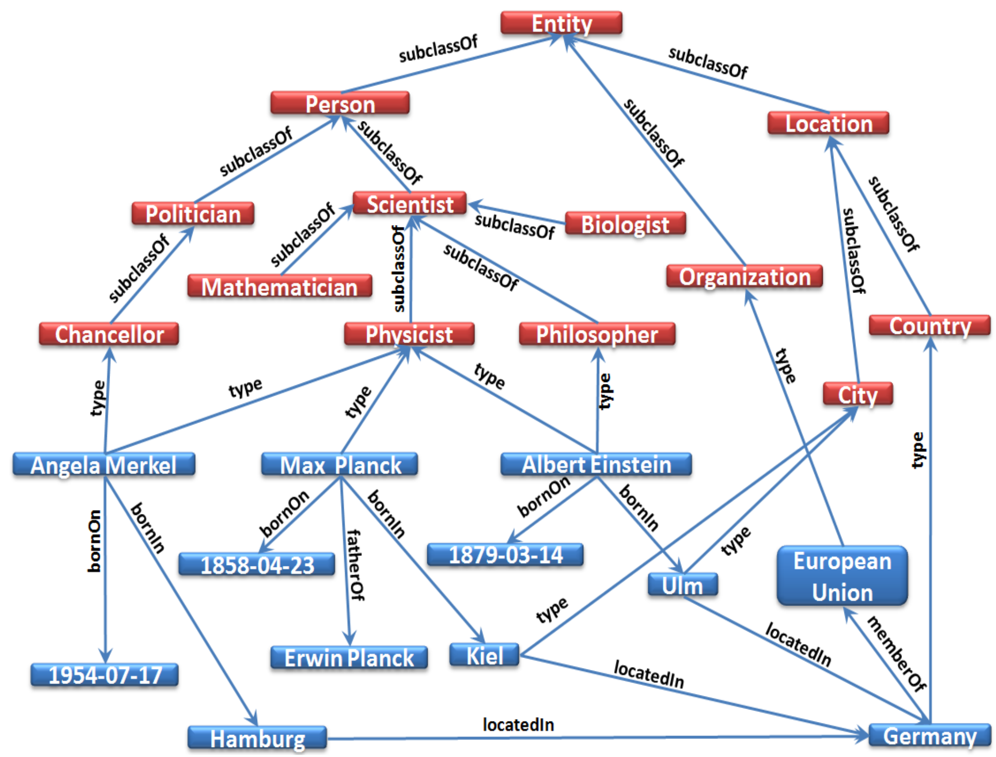
15.3 Knowledge Harvesting



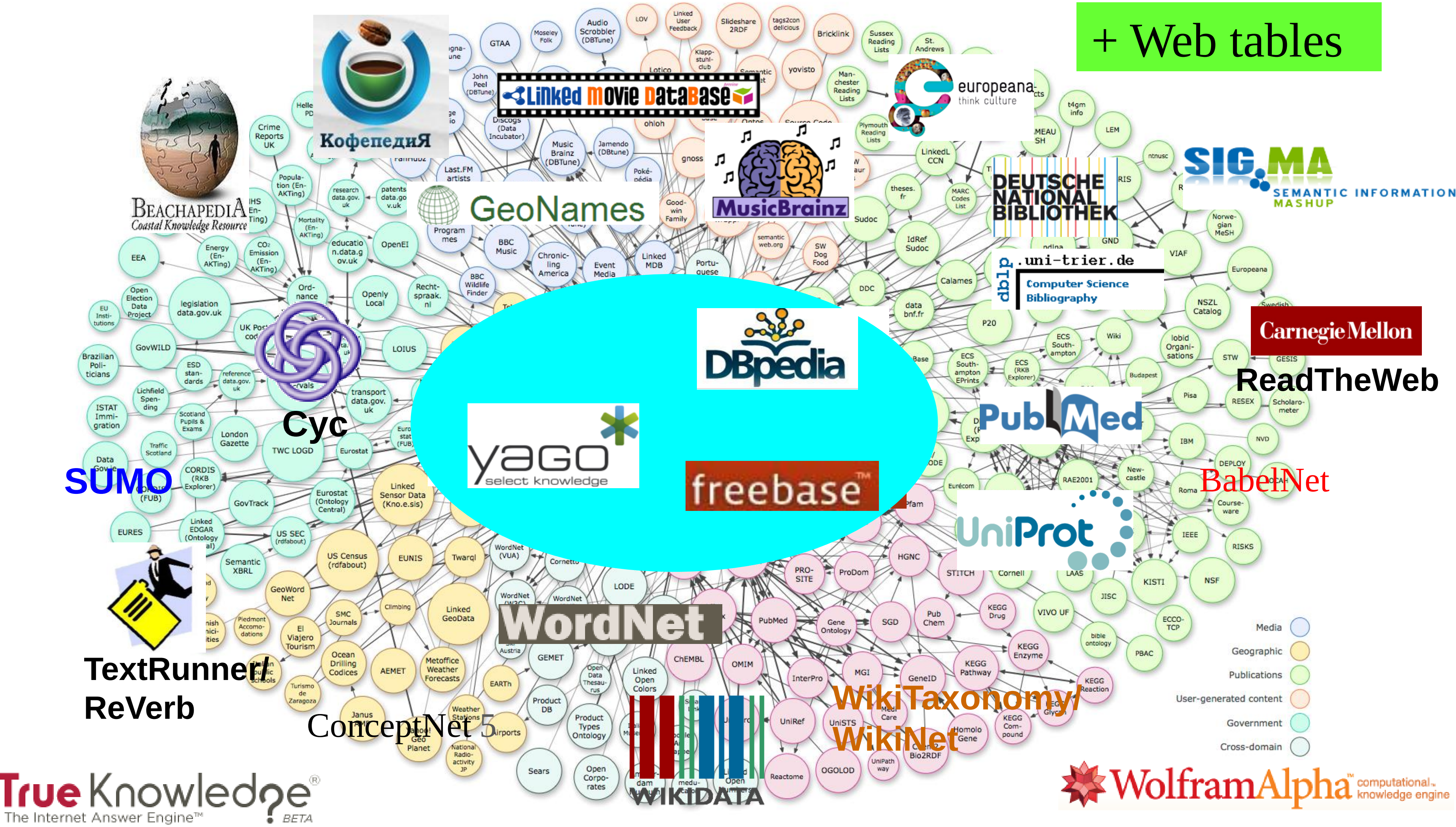
Automatic construction of large knowledge bases about entities, classes, relations from Web sources (incl. Wikipedia) using pattern matching, statistical learning & consistency reasoning





Web of Open Linked Data and Knowledge

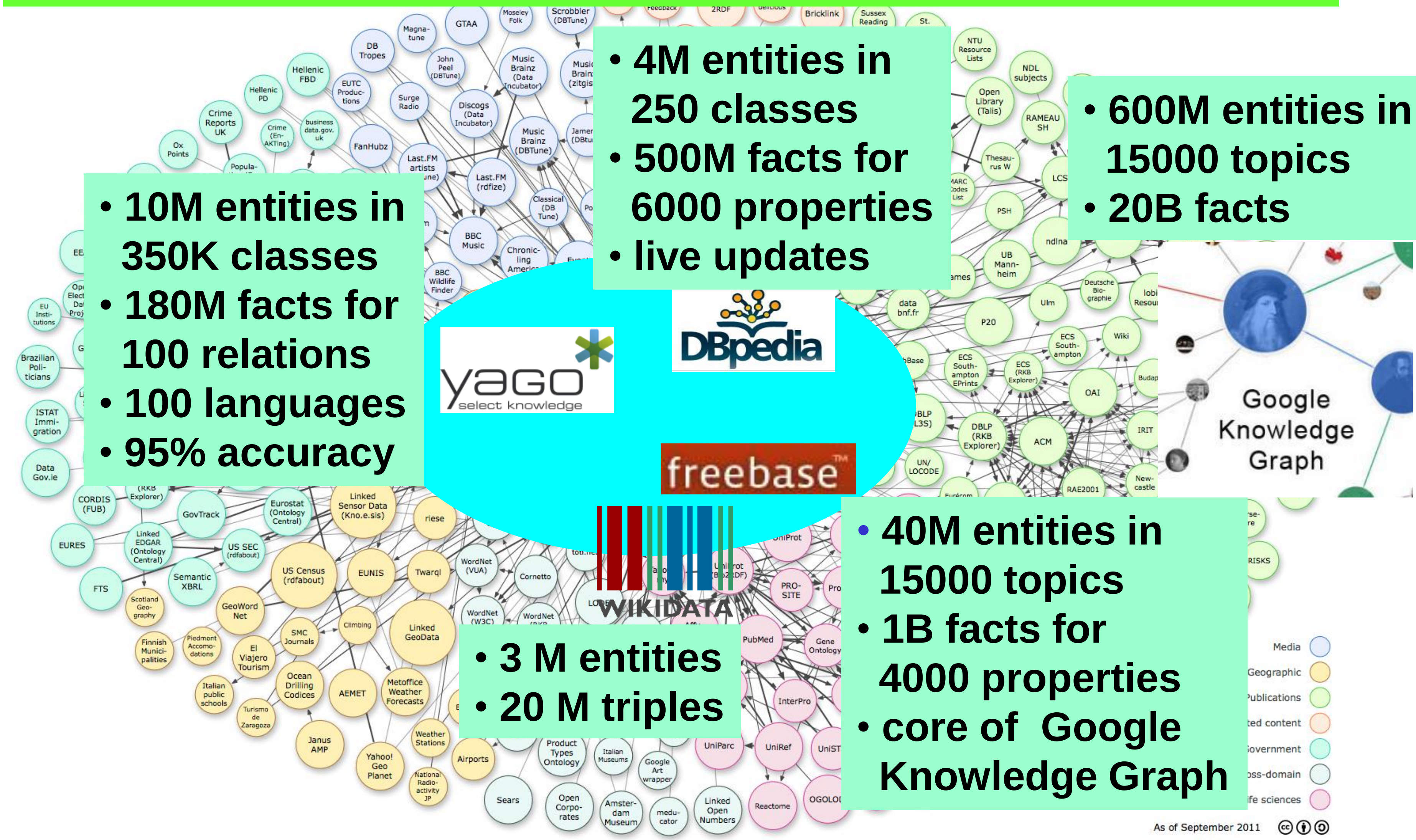
> 50 Bio. subject-predicate-object triples from > 1000 sources



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

Knowlede Bases on the Web

> 50 Bio. subject-predicate-object triples from > 1000 sources



- 10M entities in 350K classes
- 180M facts for 100 relations
- 100 languages
- 95% accuracy

yago select knowledge

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

DBpedia

- 600M entities in 15000 topics
- 20B facts

freebase

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

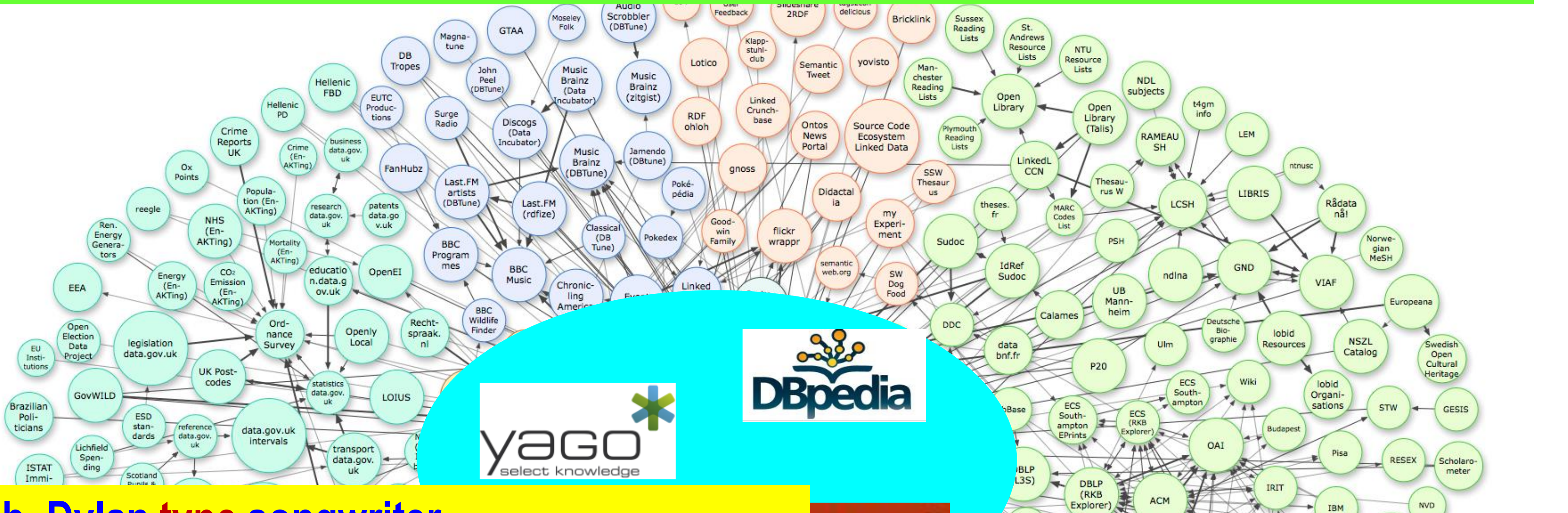
WIKIDATA

- 3 M entities
- 20 M triples

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

Knowlede Bases on the Web

> 50 Bio. subject-predicate-object triples from > 1000 sources



Bob_Dylan type songwriter
Bob_Dylan type civil_rights_activist
songwriter subclassOf artist
Bob_Dylan composed Hurricane
Hurricane isAbout Rubin_Carter
Bob_Dylan marriedTo Sara_Lownds
validDuring [Sep-1965, June-1977]
Bob_Dylan knownAs „voice of a generation“
Steve_Jobs „was big fan of“ Bob_Dylan
Bob_Dylan „briefly dated“ Joan_Baez

taxonomic knowledge

factual knowledge

temporal knowledge
terminological knowledge

evidence & belief knowledge

Knowledge Base (aka. Knowledge Graph): a Pragmatic Definition

Comprehensive and semantically organized

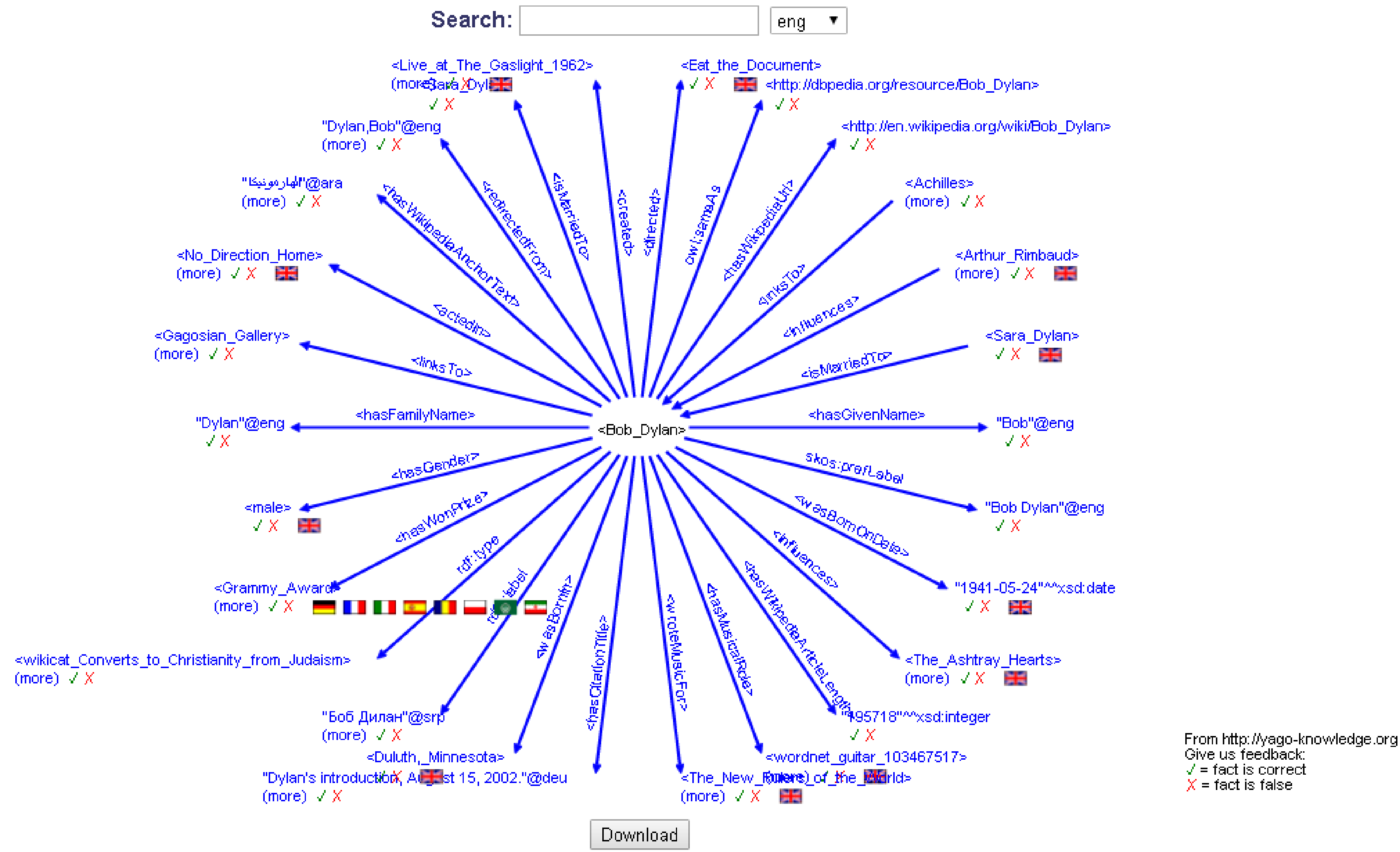
machine-readable collection of
universally relevant or domain-specific
entities, **classes**, and
SPO facts (attributes, relations)

plus spatial and temporal dimensions
plus commonsense properties and rules
plus contexts of entities and facts
(textual & visual witnesses, descriptors, statistics)
plus …..

Some Publicly Available Knowledge Bases

YAGO:	yago-knowledge.org
Dbpedia:	dbpedia.org
Freebase:	freebase.com
Wikidata:	www.wikidata.org
Entitycube:	entitycube.research.microsoft.com
	renlifang.msra.cn
NELL:	rtw.ml.cmu.edu
DeepDive:	deepdive.stanford.edu
Probase:	research.microsoft.com/en-us/projects/probase/
KnowItAll / ReVerb:	openie.cs.washington.edu
	reverb.cs.washington.edu
BabelNet:	babelnet.org
WikiNet:	www.h-its.org/english/research/nlp/download/
ConceptNet:	conceptnet5.media.mit.edu
WordNet:	wordnet.princeton.edu
Linked Open Data:	linkeddata.org

Example: YAGO



http://yago-knowledge.org

Example: DBpedia

About: Steve Jobs

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

Steven Paul Jobs (/ˈdʒɒbz/; February 24, 1955 – October 5, 2011) was an American entrepreneur, marketer, and inventor, who was the cofounder, chairman, and CEO of Apple Inc.

Property	Value
dbo:activeYearsEndYear	▪ 2011-01-01 (xsd:date)
dbo:activeYearsStartYear	▪ 1974-01-01 (xsd:date)
dbo:alias	▪ Jobs, Steven Paul
dbo:almaMater	▪ dbr:Reed_College
dbo:birthDate	▪ 1955-02-24 (xsd:date)
dbo:birthName	▪ Steven Paul Jobs
dbo:birthPlace	▪ dbr:California ▪ dbr:San_Francisco
dbo:birthYear	▪ 1955-01-01 (xsd:date)
dbo:board	▪ dbr:Apple_Inc. ▪ dbr:The_Walt_Disney_Company
dbo:child	▪ dbr:Lisa_Brennan-Jobs
dbo:deathDate	▪ 2011-10-05 (xsd:date)
dbo:deathPlace	▪ dbr:California
dbo:deathYear	▪ 2011-01-01 (xsd:date)
dbo:networth	▪ 8.3E9
dbo:occupation	▪ dbr:Pixar ▪ dbr:Apple_Inc. ▪ dbr:NeXT ▪ dbr:Steve_Jobs__1 ▪ dbr:Steve_Jobs__2 ▪ dbr:Steve_Jobs__3 ▪ dbr:Steve_Jobs__4 ▪ dbr:Steve_Jobs__5 ▪ dbr:Steve_Jobs__6
dbo:partner	▪ dbr:Chrisann_Brennan
dbo:relative	▪ dbr:Mona_Simpson
dbo:religion	▪ dbr:Lutheranism ▪ dbr:Zen
dbo:residence	▪ dbr:California

http://dbpedia.org/page/Steve_Jobs

Example: Wikidata

David Bowie (Q5383)

https://www.wikidata.org/wiki/Q5383

occupation

painter
edit
0 references
+ add reference

singer-songwriter
edit
0 references
+ add reference

guitarist
edit
0 references
+ add reference

saxophonist
edit
1 reference

composer
edit
0 references
+ add reference

film actor
edit
0 references
+ add reference

Example: NELL

50 Mio. SPO assertions, 2.5 Mio high confidence

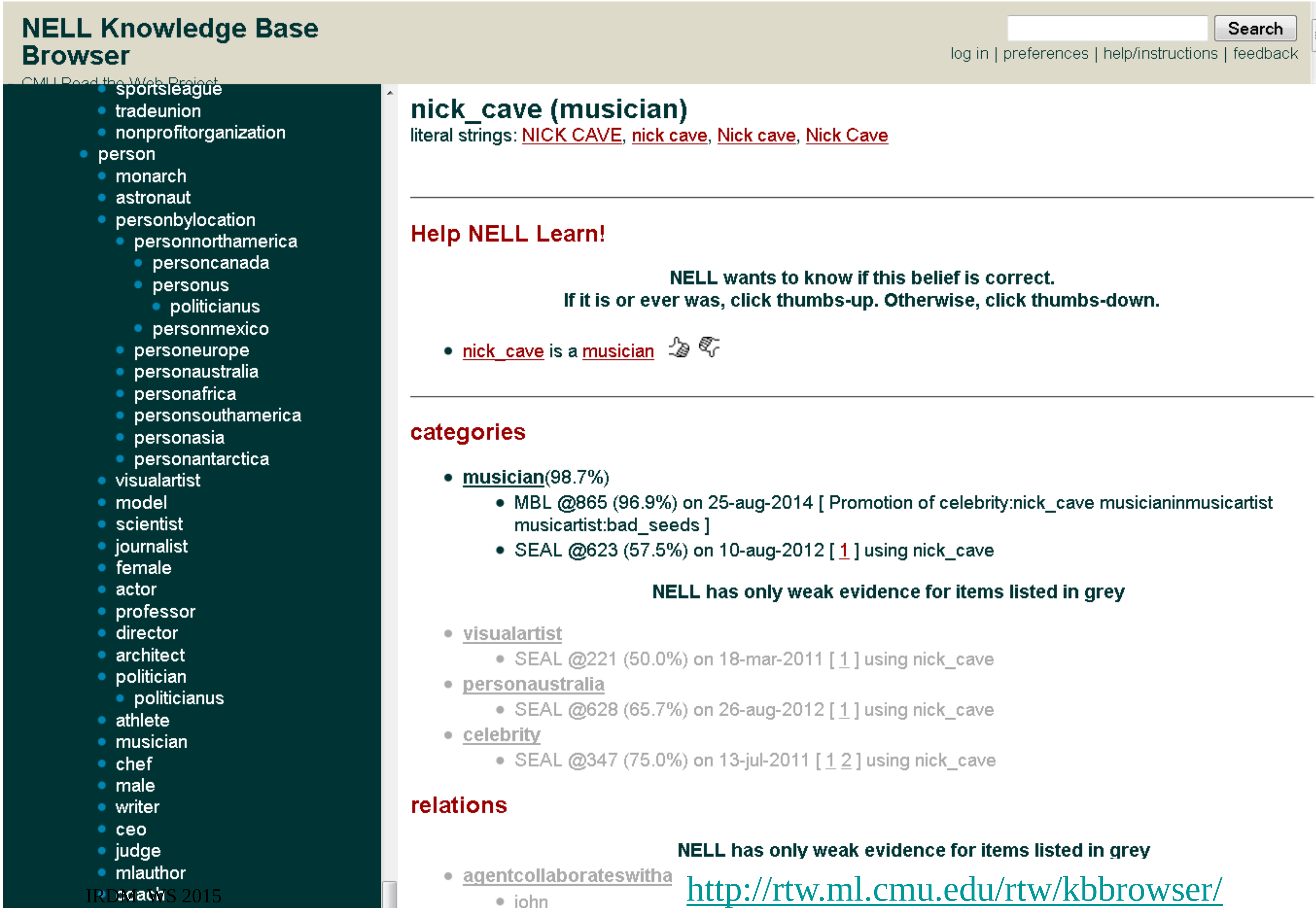
Recently-Learned Facts

instance	iteration	date learned	confidence
bresaola is a visualizable thing	922	05-may-2015	96.4
francis_derwent_wood is a visual artist	922	05-may-2015	99.9
frank_g is an Australian person	922	05-may-2015	92.2
g_protein_coupled_receptor_124 is a protein	922	05-may-2015	100.0
n_butyl_benzyl_phthalate is a chemical	922	05-may-2015	100.0
chicken001 eat potatoes	926	20-may-2015	100.0
bioinformatics is an academic program at the university college	922	05-may-2015	93.8
samuel_j__palmisano is the CEO of ibm	926	20-may-2015	100.0
national001 is a company that has an office in the country czech_republic	922	05-may-2015	99.2
the companies dc and fox_news_channel compete with eachother	922	05-may-2015	98.4

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

Example: NELL

50 Mio. SPO assertions, 2.5 Mio high confidence



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

Example: NELL

50 Mio. SPO assertions, 2.5 Mio high confidence

NELL Knowledge Base Browser

CMU Read the Web Project

Search

log in | preferences | help/instructions | feedback

- touristattractionsuchastouristattraction
- celebritysuchascelebrity
- archaeasuchasarchaea
- agriculturalproductincludingagricultura
- plantincludeplant
- actorsuchasactor
- arachnidssuchasarachnids
- mammalsuchasmammal
- architectssuchasarchitects
- companyeconomicsector
- trophywonbycoaches
- agentinvolvedwithitem
  - wineryproduceswine
  - agentworkedondrug
  - producesproducttype
  - producesproduct
    - automakerproducesmodel
  - mlauthorofsoftware
  - musicianplaysinstrument
- animaldevelopdisease
- countrylanguage
- issueofpoliticsbill
- universityoperatesinlanguage
- iteminvolvedwithagent
  - drugworkedonbyagent
  - wineproducedbywinery
  - mlsoftwareauthor
  - producedby
    - automodelproducedbymaker
  - typeproducedby
  - instrumentplayedbymusician
- dateof
  - dateatwhichexistsitem
  - dateevent
    - dateofsportsgame
    - dateofmeetingeventtitle

musicianplaysinstrument

(relation: domain musician, range musicinstrument)

Specifies that a musical instrument is played by a paritcular musician

See metadata for musicianplaysinstrument

712 instances, 1 page

instance	iteration	date learned	confidence
adam, drums	799	27-dec-2013	100.0
adam, guitar	799	27-dec-2013	100.0
bach, piano	551	19-apr-2012	100.0
bach, violin	551	19-apr-2012	100.0
barber, violin	598	21-jun-2012	100.0
bb_king, guitar	680	09-jan-2013	100.0
beethoven, piano	853	11-jul-2014	100.0
beethoven, violin	853	11-jul-2014	100.0
ben_harper, guitar	820	08-mar-2014	100.0
billie_joe_armstrong, guitar	818	03-mar-2014	100.0
brahms, piano	592	13-jun-2012	100.0
brahms, violin	503	06-feb-2012	100.0
buddy_guy, guitar	664	01-dec-2012	100.0
b_b__king, guitar	406	08-sep-2011	(Seed) 100.0
charlie, guitar	724	12-apr-2013	100.0
chopin, piano	683	15-jan-2013	100.0
copland, piano	665	05-dec-2012	100.0
david, bass	904	20-feb-2015	100.0
david, drums	904	20-feb-2015	100.0
david, guitar	904	20-feb-2015	100.0
david, keyboards	904	20-feb-2015	100.0
earl_scruggs, banjo			
eddie, guitar			

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

Knowledge for Intelligent Applications

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 Big Data & Deep Text analytics

15.3.1 Harvesting Unary Predicates with Patterns

Which **entity types (classes, unary predicates)** are there?

scientists, doctoral students, computer scientists, …
female humans, male humans, married humans, …

Which **subsumptions** should hold
(subclass/superclass, hyponym/hypernym, inclusion dependencies)?

subclassOf (computer scientists, scientists)
subclassOf (physicists, scientists),
subclassOf (scientists, humans), …

Which **individual entities** belong to which classes?

instanceOf (Jim Gray computer scientists),
instanceOf (Barbara Liskov, computer scientists),
instanceOf (Barbara Liskov, female humans),
instanceOf (Steve Jobs, male humans),
instanceOf (Steve Jobs, entrepreneurs), … …

Hearst Patterns

[M. Hearst 1992]

Goal: find **instances of classes** (and/or: find subclasses of classes)

Hearst specified **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 including the surrounding Hangzhou hills

occurrence statistics for better precision (e.g. #occurrences w/ different patterns)

Derive type(Y,X)

type(Apple, company), type(Google, company), ...
or as unary predicates: company(Apple), …

Doubly-anchored patterns

[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 → --- (no output)

Set Completion 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)

Set Completion Example 1

sets labs

Automatically create sets of items from a few examples.

Enter a few items from a set of things. (example)
Next, press *Large Set* or *Small Set* and we'll try to predict other items in the set.

- Stanford
- Princeton
- Penn State
-
-

Set Completion Example 1

Predicted Items	
penn state	georgetown
stanford	michigan
princeton	arizona
ucla	washington
harvard	dartmouth
mit	oregon
usc	nyu
yale	california
columbia	brown
cornell	chicago
berkeley	northwestern
duke	caltech
	virginia
	penn

Set Completion Example 2

Discuss Terms of Use

Google sets labs

Automatically create sets of items from a few examples.

Enter a few items from a set of things. (example)
Next, press *Large Set* or *Small Set* and we'll try to predict other items in the set.

- Pushkin
- Tolstoy
- Pasternak
-
-

(clear all)

Large Set | Small Set (15 items or fewer)

Examples:
green, purple, red chicken dance, macarena, ymca alexander, gladiator, troy hilary duff, kelly clarkson more...

labs.google.com - All About Google

http://labs.google.com/sets

Set Completion Example 2

Predicted Items
tolstoy
pushkin
leo tolstoy
anna karenina
gogol
drama
danielle steel
dostoevsky
maxim gorky
russia
fyodor dostoevsky
anton chekhov
ivan turgenev
paulo coelho
dan brown
ernest hemingway
dostojevski
alexander pushkin

john steinbeck
russian literature
lermontov
stephen king
cs lewis
madame bovary
bible
the idiot
mark twain
mikhail bulgakov
fyodor dostoyevsky
nikolai gogol
susanna tamaro
edward said
dirty dancing
albert camus
shakespeare
romance novel
jack london
george orwell
fiction
authors

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**

$$\text{PMI}(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

Caveats:

Precision drops for classes with sparse statistics

Harvested items are names, not entities

Canonicalization (de-duplication) unsolved

15.3.2 Harvesting Binary Predicates with Seeds and Constraints

Which **instances** (pairs of individual entities) are there for given **binary relations** with specific **type signatures**?

hasAdvisor (JimGray, MikeHarrison)
graduatedAt (JimGray, Berkeley)
graduatedAt (Chris Manning, Stanford)
hasWonPrize (JimGray, TuringAward)
hasWonPrize (VintCerf, TuringAward)
bornOn (JohnLennon, 9-Oct-1940)
diedOn (JohnLennon, 8-Dec-1980)
marriedTo (JohnLennon, YokoOno)

Which additional & interesting **relation types** are there between given classes of entities?

→ 15.3.3

attendedSchool(x,y), competedWith(x,y), nominatedForPrize(x,y), …
divorcedFrom(x,y), affairWith(x,y), …
assassinated(x,y), rescued(x,y), admired(x,y), …

Relational Facts from Text

composed (<musician>, <song>) **appearedIn (<song>, <film>)**

Bob Dylan wrote the song Knockin' on Heaven's Door
Lisa Gerrard wrote many haunting pieces, including Now You Are Free
Morricone's masterpieces include the Ecstasy of Gold
Dylan's song Hurricane was covered by Ani DiFranco
Strauss's famous work was used in 2001, titled Also sprach Zarathustra
Frank Zappa performed a jazz version of Rota's Godfather Waltz
Hallelujah, originally by Cohen, was covered in many movies, including Shrek

composed (Bob Dylan, Knockin' on Heaven's Door)
composed (Lisa Gerrard, Now You Are Free)
…
appearedIn (Knockin' on Heaven's Door, Billy the Kid)
appearedIn (Now You Are Free, Gladiator)
…

Pattern-based Gathering (statistical evidence) **+** **Constraint-aware Reasoning (logical consistency)**

Pattern-based Harvesting: Fact-Pattern Duality

Task populate relation *composed*
starting with seed facts

[Brin 1998, Etzioni 2004,
Agichtein/Gravano 2000]

Facts **&** ***Fact Candidates***

(Dylan, Knockin)
(Gerrard, Now)

(Dylan, Hurricane)
(Morricone, Ecstasy)
(Zappa, Godfather)
(Mann, Buddenbrooks)

(Gabriel, Biko)
(Puebla, Che Guevara)
(Mezrich, Zuckerberg)
(Jobs, Apple)
(Newton, Gravity)

Patterns

X wrote the song Y
X wrote … including Y

X covered the story of Y
X has favorite movie Y
X is famous for Y

…

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

Improving Pattern Precision or Recall

- Statistics for confidence:

  occurrence frequency with seed pairs

  distinct number of pairs seen

- Negative seeds for confusable relations:

  capitalOf(city,country) → X is the largest city of Y

  pos. seeds: (Paris, France), (Rome, Italy), (New Delhi, India), ...

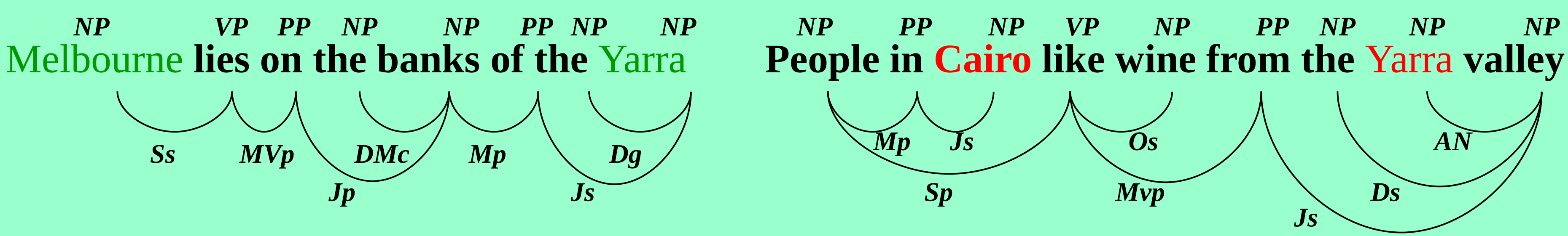
  neg. seeds: (Sydney, Australia), (Istanbul, Turkey), ...

- Generalized patterns with wildcards and POS tags:

  hasAdvisor(student,prof) → X met his celebrated advisor Y

  → X * PRP ADJ advisor Y

- Dependency parsing for complex sentences:



Statistics for Pattern Quality 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 for Improved 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 for Improved 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

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

Frequent sequence mining

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

Constrained Reasoning for Logical Consistency

Use knowledge (consistency constraints) for joint reasoning on hypotheses and pruning of false candidates

Hypotheses:

composed (Dylan, Hurricane)
composed (Morricone, Ecstasy)
~~composed (Zappa, Godfather)~~
composed (Rota, Godfather)
composed (Gabriel, Biko)
~~composed (Mann, Buddenbrooks)~~
~~composed (Jobs, Apple)~~
~~composed (Newton, Gravity)~~

Constraints:

$\forall$ x, y: composed (x,y) $\Rightarrow$ type(x)=musician
$\forall$ x, y: composed (x,y) $\Rightarrow$ type(y)=song
$\forall$ x, y, z: composed (x,y) $\wedge$ appearedIn(y,z) $\Rightarrow$ wroteSoundtrackFor (x,z)
$\forall$ x,y,t,b,e: composed (x,y) $\wedge$ composedInYear (y, t) $\wedge$ bornInYear (x, b) $\wedge$ diedInYear (x,e) $\Rightarrow b < t \leq e$
$\forall$ x, y, w: composed (x,y) $\wedge$ composed(w,y) $\Rightarrow$ x = w
$\forall$ x, y: sings(x,y) $\wedge$ type(x,singer-songwriter) $\Rightarrow$ composed(x,y)

consistent subset(s) of hypotheses ("possible world(s)", "truth")
$\rightarrow$ **Weighted MaxSat solver for set of logical clauses**
$\rightarrow$ **max a posteriori (MAP) for probabilistic factor graph**

Weighted Max-Sat Reasoning

- **Grounding** of formulas produces **clauses** (propositional logic: disjunctions of positive or negative literals) connecting patterns, facts, hypotheses, constraints
  Ex.: **composed(Gabriel,Biko); ¬composed(Gabriel,Biko) ∨ type(Gabriel,musician);**
  composed(Mandela,Biko); ¬composed(Mandela,Biko) ∨ type(Mandela,musician);
  ¬ composed(Gabriel,Biko) ∨ ¬ appearedIn(Biko,CryForFreedom) ∨ wroteSoundtrack(Gabriel,CryForFreedom);
  ¬ composed(Gabriel,Biko) ∨ ¬ composed(Mandela,Biko) ∨ False; …..
- Treat **hypotheses** (literals) **as variables**, facts as constants:
  A; ¬A∨B; C; ¬C∨D; ¬A∨¬E∨F; ¬A∨¬C; …..
- Clauses are weighted by pattern statistics and rule confidence
- Solve **weighted Max-Sat** problem:
  assign truth values to variables s.t.
  total weight of satisfied clauses is max!
  → NP-hard, but good approximation algorithms

Markov Logic Networks (MLN's)

Map logical constraints & fact candidates (M. Richardson / P. Domingos 2006)

into probabilistic graph model: Markov Random Field (MRF)

spouse(x,y) $\wedge$ diff(y,z) $\Rightarrow$ $\neg$spouse(x,z)

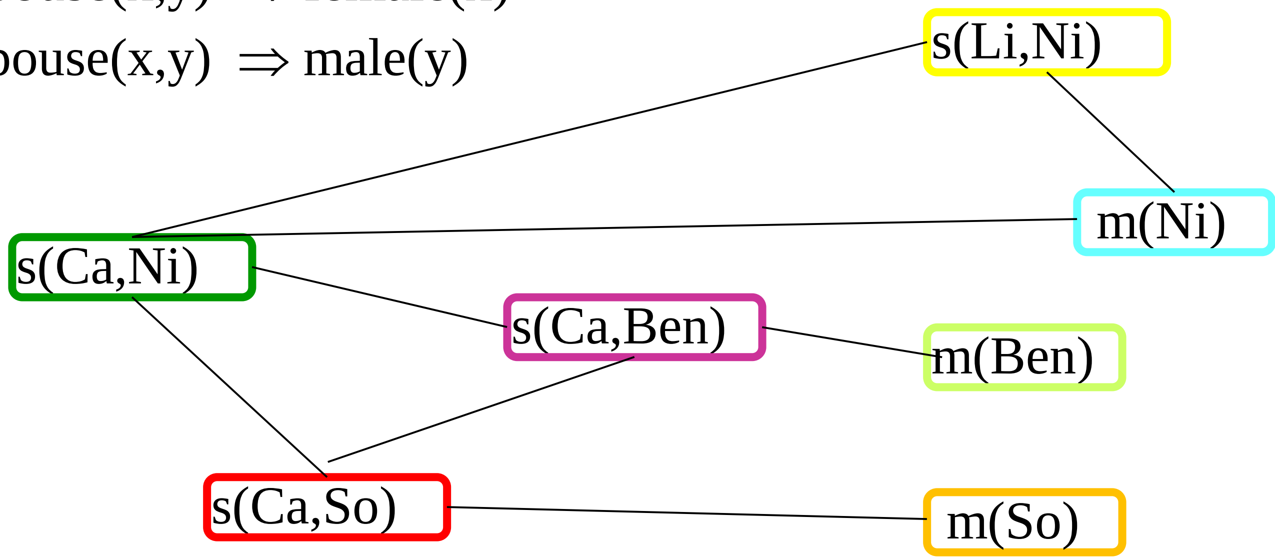
spouse(x,y) $\wedge$ diff(w,y) $\Rightarrow$ $\neg$spouse(w,y)

spouse(x,y) $\Rightarrow$ female(x)

spouse(x,y) $\Rightarrow$ male(y)

s(Carla,Nick)
s(Lisa,Nick)
s(Carla,Ben)
s(Carla,Sofie)
…

m(Nick)
m(Ben)
m(Sofie)
…



RVs coupled by MRF edge if they appear in same clause

MRF assumption:
$\mathbf{P[X_i|X_1..X_n]=P[X_i|N(X_i)]}$

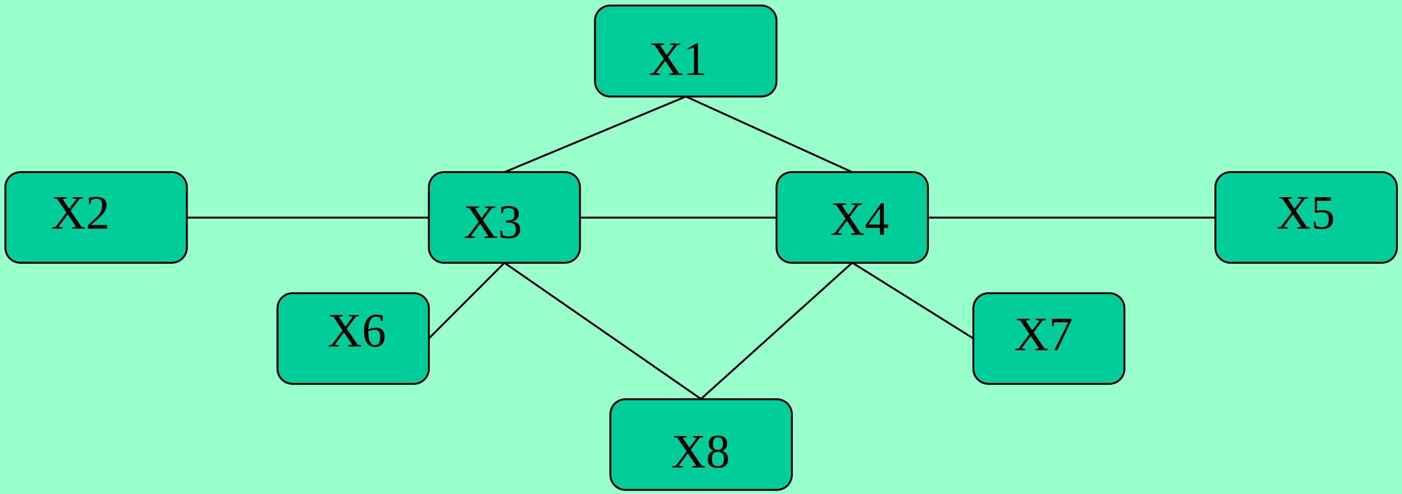
joint distribution has product form over all cliques

Variety of algorithms for joint inference:
- Gibbs sampling, other MCMC, belief propagation, …

MAP inference equivalent to Weighted MaxSat

MRF: Markovian Probabilistic Graphical Model

Network of discrete random variables (often binary)



Markov assumption: $\boldsymbol{P[X_1|X_2,X_3 \dots X_n] = P[X_1|Neighbors(X_1)]}$

Hammersley-Clifford Theorem:

$$P[X_1X_2\dots] = 1/Z \prod_c \Phi_c(X_iX_j\dots \in c)$$

over all cliques c

or as log-linear model:

$$P[X_1X_2\dots] = 1/Z \exp\left(\sum_c \underbrace{w_c}_{\text{weights}} \underbrace{f_c}_{\text{features}}(X_iX_j\dots \in c)\right)$$

Inference for Xi's by Monte Carlo sampling, belief propagation, etc.

Parameter learning by non-convex optimization

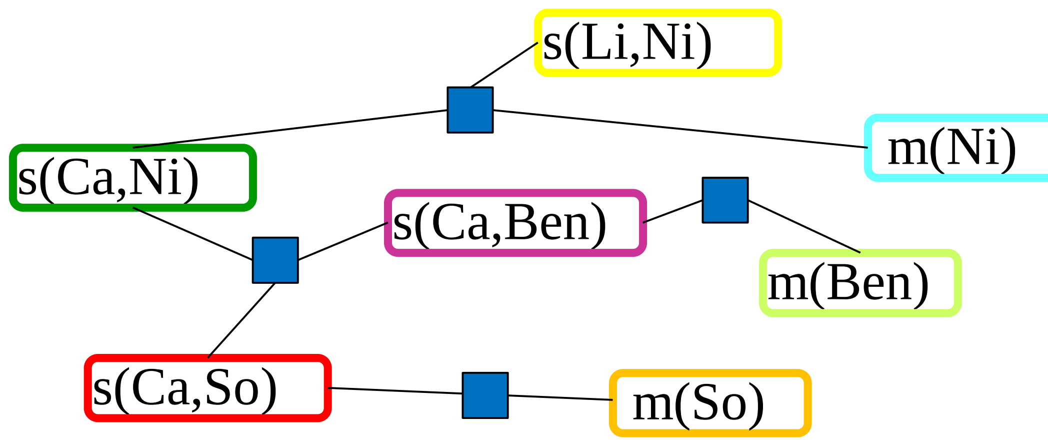
Related Alternative Probabilistic Models

Constrained Conditional Models [Roth et al.]

log-linear classifiers with constraint-violation penalty
mapped into Integer Linear Programs

Factor Graphs with Imperative Variable Coordination [A. McCallum et al.]

RV‘s share “factors“ (joint feature functions)
generalizes MRF, BN, CRF; inference via advanced MCMC
flexible coupling & constraining of RV‘s



Probabilistic Soft Logic (PSL) [L. Getoor et al.]

gains MAP efficiency by continuous RV‘s (degree of truth)

15.3.3 Harvesting SPO Triples by Open Information Extraction

so far KB has explicit model:

- **canonicalized entities**
- **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 …

Example: ReVerb

Open Information Extraction

?x „an affair with" ?y

Argument 1: [ ] **Relation:** affair with **Argument 2:** [ ] All Search

307 answers from **1015 sentences** (cached)

all | person (54) | author (35) | tv actor (33) | person or entity appearing in film (31) | actor (29) | misc. | more types

Whitney Houston , **Jermaine Jackson** (7)

John McCain , a lobbyist (5)

Bill Clinton , **Monica Lewinsky** (5)

Jesus , **Mary Magdalene** (5)

Suzanne Coleman , **Bill Clinton** (3)

her mother , **Tiger Woods** (3)

the medias , **Barack Obama** (3)

Newt Gingrich , House (3)

Thomas Jefferson , **Sally Hemings** (3)

Saddam Hussein , **Samira Shahbandar** (3)

Suzanne Coleman Reportedly , **Bill Clinton** (3)

his wife , **George Foreman** (2)

Clementine Churchill, Baroness Spencer-Churchill , Terence Phillip (2)

the extraterrestrial , **Hillary Rodham Clinton** (2)

an unnamed intern , **John F. Kennedy** (2)

http://openie.cs.washington.edu

http://openie.allenai.org

Open IE with ReVerb

[A. Fader et al. 2011, T. Lin 2012, Mausam 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**

Mining Paraphrases of Relations

composed (<musician>, <song>) **covered (<musician>, <song>)**

Dylan wrote his song Knockin‘ on Heaven‘s Door, a cover song by the Dead
Morricone ‘s masterpiece is the Ecstasy of Gold, covered by Yo-Yo Ma
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 sa in her version of Don‘t Explain
Cale performed Hallel ah written by L. Cohen

covered by: (Amy,Cupid), (Ma, Ecstasy), (Nina, Do
(Cat, Don‘t), (Cale, Hallelujah), …

voice in version of: (Amy,Cupid), (Sam, Cupid), (Nina, Do
(Cat, Don‘t), (Cale, Hallelujah), …

performed: (Amy,Cupid), (Amy, Black), (Nina, Do
(Col , Hallelujah), (Dylan, Knockin)

frequent sequence mining
for relational phrases
support sets of entity pairs
for paraphrases
clustering for “synsets“

covered (<musician>, <song>):
cover song, interpretation of, singing of, voice in … version , …

composed (<musician>, <song>):
wrote song, classic piece of, ‘s old song, written by, composition of, …

PATTY: Pattern Taxonomy for Relations

[Nakashole et al.: EMNLP-CoNLL'12, VLDB'12,
Moro et al.: CIKM'12, Grycner et al.: COLING'14]

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

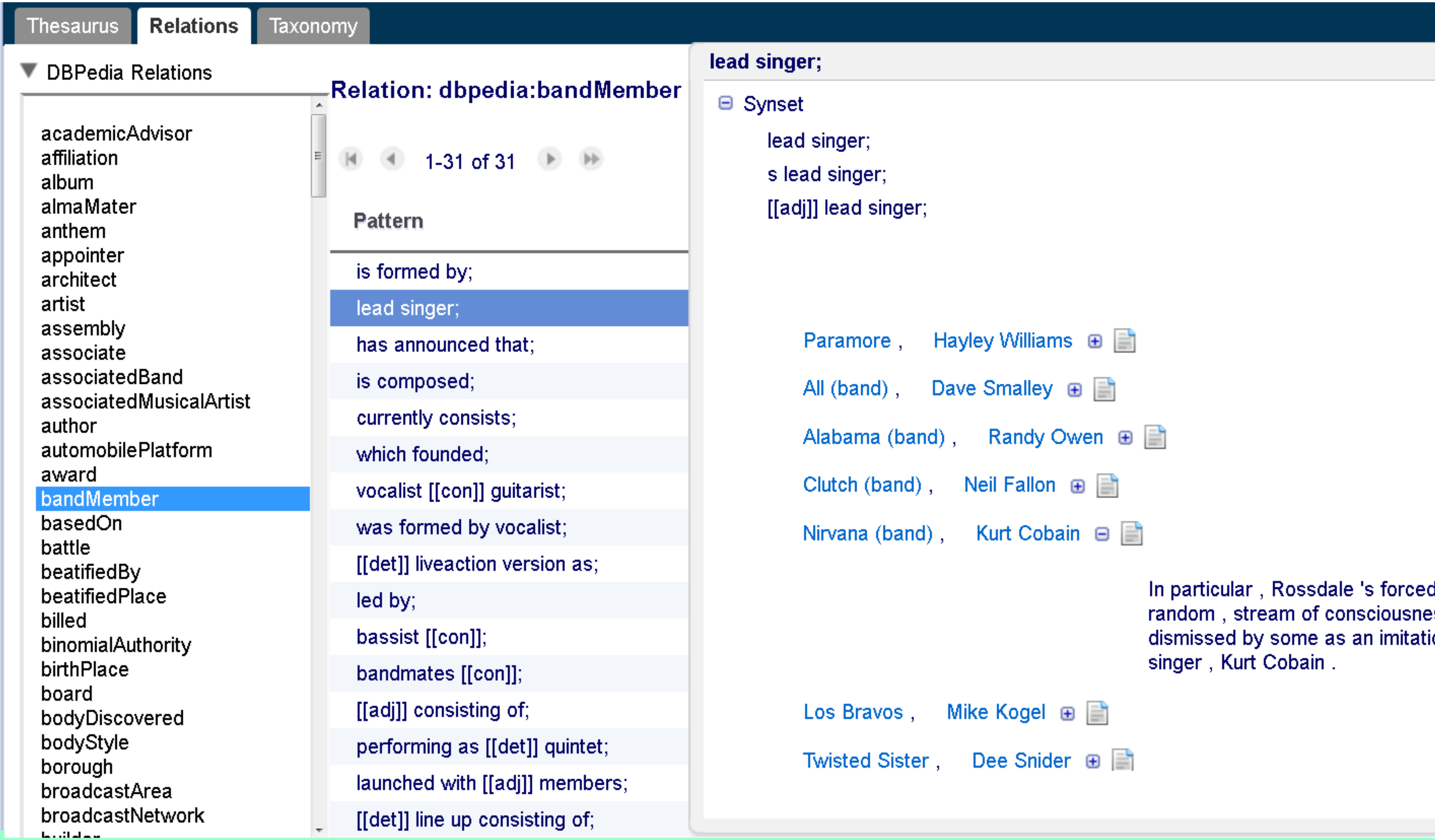
- **"graduated from" ⇔ "obtained degree in * from"**
- **"and PRP ADJ advisor" ⇔ "under the supervision of"**

One relational phrase can **subsume** another

- **"wife of" ⇒ " spouse of"**

PATTY: Pattern Taxonomy for Relations

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



350 000 SOL patterns with 4 Mio. instances

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

15.3.4 Harvesting Commonsense by Patterns and Logical & Statistical Inference

Assertions **about general concepts** (not individual entities) and their attributes and relations

hasProperty (circle, round), hasProperty (lake, round)

hasProperty (coffee, strong)

hasAbility (bird, fly), hasAbility (human, make jokes)

hasColor (cherry, red), hasTaste (cherry, juicy), hasShape (cherry, round)

smallerThan (cherry, apple), largerThan (cherry, pea)

partOf (pedal, bike), partOf (nose, human), visualPartOf (nose, human)

locatedAt (bike, park), locatedAt (coffee, cup),

usedFor (cherry, ice cream), usedFor (book, learn),

happensAtTime (traffic jam, rush hour), happensAtLocation (traffic jam, street)

Commonsense Acquisition: Not So Easy

Every child knows that

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

pots and pans are in the kitchen or cupboard, on the stove, …
but not in in the bedroom, in your pocket, in the sky, …

children usually live with their parents

But: commonsense is rarely stated explicitly
Plus: web and social media have reporting bias

rich family: 27.8 Mio on Bing
poor family: 3.5 Mio on Bing

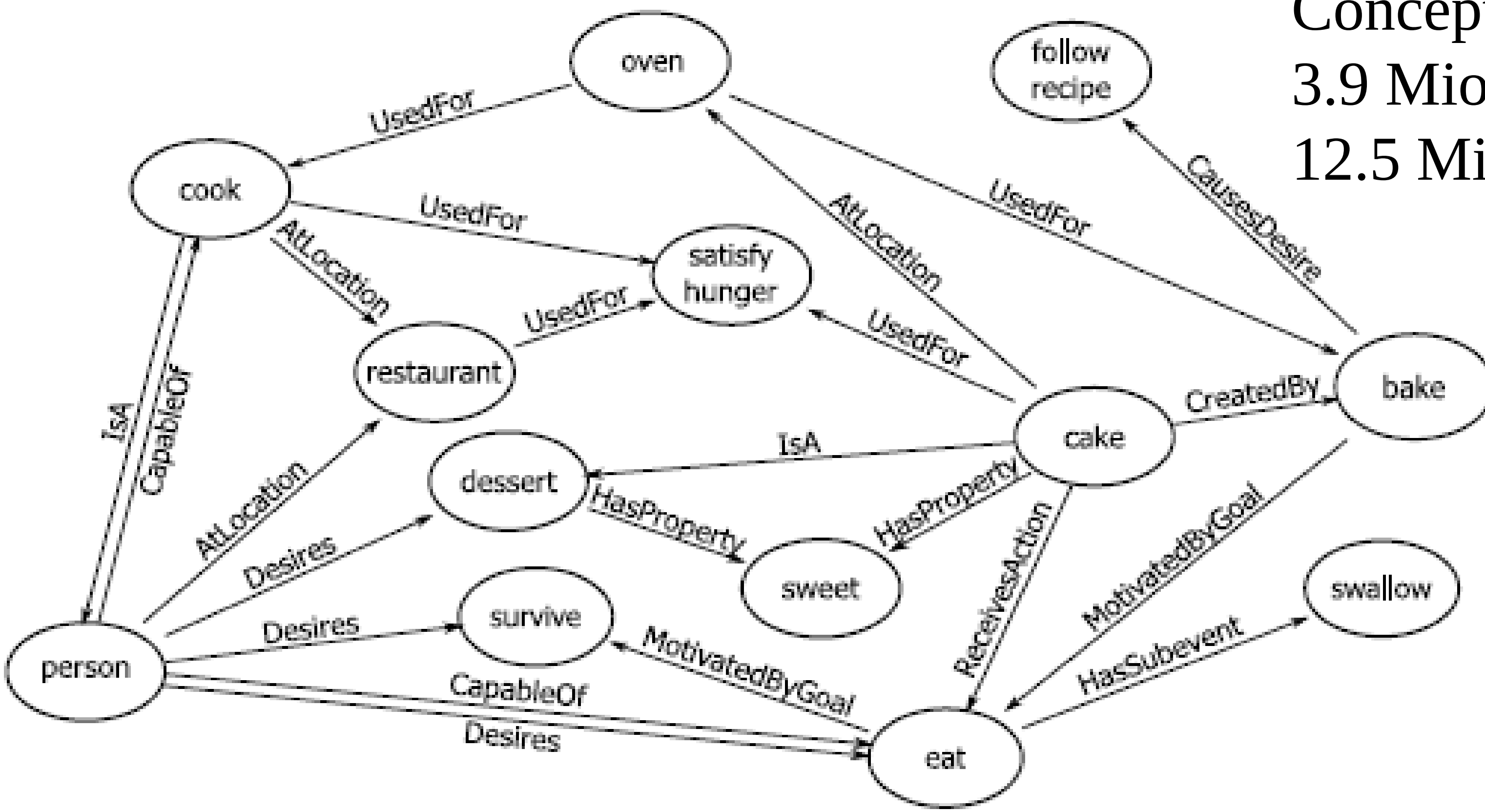
singers: 22.8 Mio on Bing
workers: 14.5 Mio on Bing

Example: ConceptNet

[Speer & Havasi 2012]

many inputs incl. WordNet, Verbosity game, etc.

ConceptNet 5:
3.9 Mio concepts
12.5 Mio. edges



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

Example: WebChild

WEBCHILD Commonsense Browser

beer

Guess the concept

- Domain
- Comparable
- Physical Part
- Activity
- Property
- Location

Ask me!

beer

https://gate.d5.mpi-inf.mpg.de/webchild

a general name for alcoholic beverages made by fermenting a cereal (or mixture of cereals) flavored with hops

TYPE OF	brew
	Related to **food**, under the category of **food**
COMPARABLES	beer,wine · cider,beer · coffee,beer · ale,beer · beer,liquor · More
ACTIVITIES	drink beer · buy beer · make beer · order beer · finish beer
HAS PHYSICAL PARTS	food
HAS SUBSTANCE	beverage · silica · glass
IS SUBSTANCE OF	brewpub · microbrewery · brewery
IN SPATIAL PROXIMITY WITH	pub · house · bar · store · beach · More

Pattern-Based Harvesting of Commonsense Properties

(N. Tandon et al.: AAAI 2011)

Approach: Start with seed facts for

apple hasProperty round
dog hasAbility bark
plate hasLocation table

Find patterns that express these relations, such as

X is very Y, X can Y, X put in/on Y, …

Apply these patterns to find more facts.

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

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

Commonsense with SPO Properties

[N. Tandon et al.: WSDM'14]

Who **looks hot** ?

What **tastes hot** ?

What **is hot** ?

What **feels hot** ?

→ 4 Mio **sense-disambiguated SPO triples** for predicates:
hasProperty, hasColor, hasShape, hasTaste, hasAppearance, isPartOf, hasAbility, hasEmotion, …

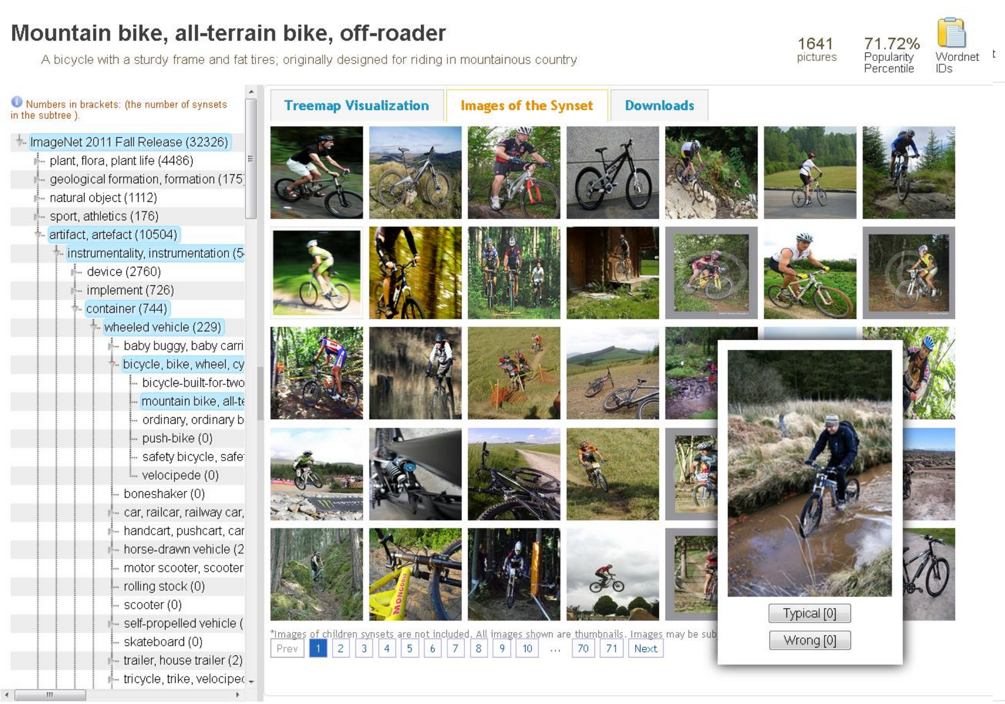
- **pattern learning with seeds: high recall**
- **semisupervised label propagation: good precision**
- **integer linear program: sense disambiguation, high precision**

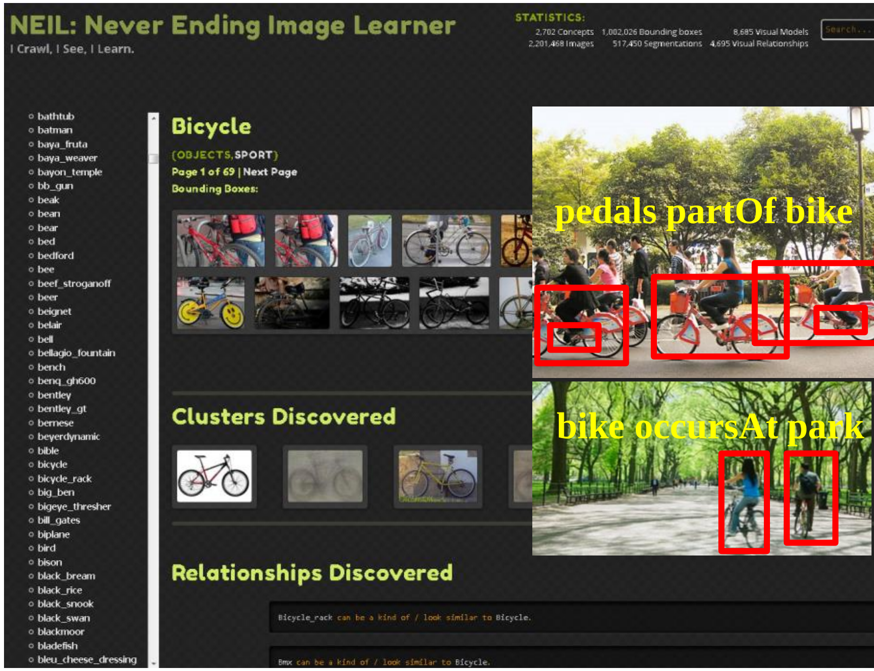
https://gate.d5.mpi-inf.mpg.de/webchild/

Visual Commonsense

ImageNet: populate WordNet classes with many photos
[J. Deng et al.: CVPR‘09]
http://www.image-net.org

NEIL: infer instances of partOf occursAt, inScene relations
[X. Chen et al.: ICCV‘13]
http://www.neil-kb.com/





How:
crowdsourcing for seeds, distantly supervised classifiers, object recognition (bounding boxes) in computer vision

Commonsense for Visual Scenes

[N. Tandon et al.: CIKM‘15, AAAI‘16]

Activity knowledge from movie&TV scripts, aligned with visual scenes

→ 0.5 Mio activity types with attributes: location, time, participants, prev/next

Refined part-whole relations from web&books text and image tags

→ 6.7 Mio sense-disambiguated triples for physicalPartOf, visualPartOf, hasCardinality, memberOf, substanceOf

Challenge: Commonsense Rules

Horn clauses:
can be learned by Inductive Logic Programming

$\forall$ x,m,c: type(x,child) $\wedge$ mother(x,m) $\wedge$ livesIn(m,t)) $\Rightarrow$ livesIn(x,t)
$\forall$ x,m,f: type(x,child) $\wedge$ mother(x,m) $\wedge$ spouse(m,f) $\Rightarrow$ father(x,f)

Advance rules beyond Horn clauses:
specified by human experts

$\forall$ x: type(x,spider) $\Rightarrow$ numLegs(x)=8
$\forall$ x: type(x,animal) $\wedge$ hasLegs(x) $\Rightarrow$ even(numLegs(x))
$\forall$ x: human(x) $\Rightarrow$ ($\exists$ y: mother(x,y) $\wedge$ $\exists$ z: father(x,z))
$\forall$ x: human(x) $\Rightarrow$ (male(x) $\vee$ female(x))

Additional Literature for 15.3

- F.M. Suchanek, G. Weikum: Knowledge harvesting in the big-data era, SIGMOD 2013
- M. Hearst: Automatic Acquisition of Hyponyms from Large Text Corpora. COLING 1992
- S Brin: Extracting Patterns and Relations from the World Wide Web. WebDB 1998:
- E. Agichtein. Snowball: extracting relations from large plain-text collections, ACM DL 2000
- O. Etzioni et al.:Unsupervised named-entity extraction from the Web, Art. Intell. 2005
- R.C. Wang, W. Cohen:Iterative Set Expansion of Named Entities Using the Web, ICDM 2008
- F. Suchanek et al.: SOFIE: a self-organizing framework for information extraction, WWW '09
- M. Mintz et al.: Distant supervision for relation extraction without labeled data. ACL 2009
- N. Nakashole : Scalable knowledge harvesting with high precision and high recall, WSDM '11
- Z. Nie, J.-R. Wen, W.-Y. Ma:: Statistical Entity Extraction From the Web. Proc. IEEE 2012
- T. Mitchell et al.: Never-Ending Learning. AAAI 2015:
- A. Fader et al.: Identifying Relations for Open Information Extraction, EMNLP 2011
- Mausam et al.:Open Language Learning for Information Extraction, EMNLP 2012
- N. Nakashole: PATTY: A Taxonomy of Relational Patterns with Semantic Types, EMNLP'12
- R. Speer, C. Havasi: Representing General Relational Knowledge in ConceptNet 5, LREC'12
- N. Tandon et al.: Deriving a Web-Scale Common Sense Fact Database, AAAI 2011
- N. Tandon et al.: WebChild: harvesting and organizing commonsense knowledge from the web, WSDM 2014.

Summary of Chapter 15

- Information Extraction lifts **text&Web contents** into **structured data**: entities, attributes, relations, facts and opinions
- **Regex-centric rules and patterns** good for **homogenous** Web sites
- **Statistical learning** of patterns (HMM, CRF/MRF, classifiers, etc.) crucial for **heterogenous** sources and **natural-language** text
- **Knowledge harvesting** exploits Web-scale redundancy & statistics