Course: Common Sense Reasoning

9. Automatic Generation of Large Scale Data Bases

Martin Molina

The content of large scale data bases can be generated automatically

- The content of large scale data bases can be extracted automatically from different data sources
- Two representative cases are presented:
  - Yago
  - Nell

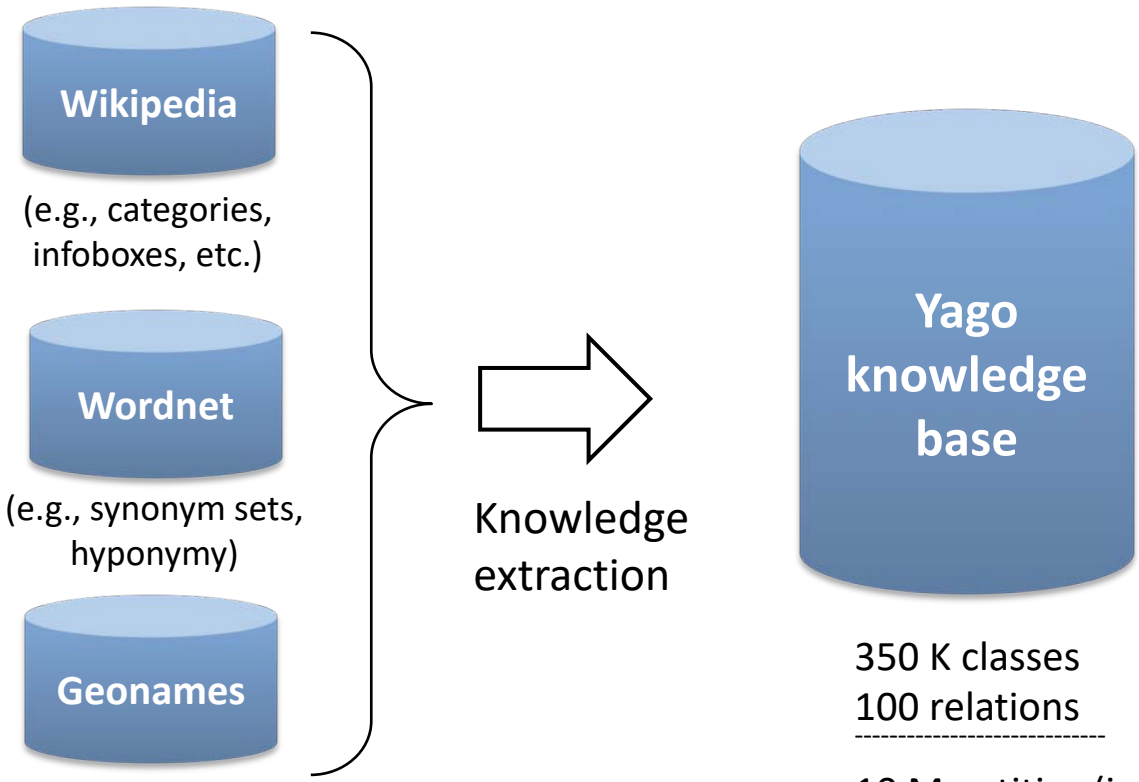
Yago was generated automatically using different information sources

- Yago: Yet Another General Ontology
- Goal of the project: extract knowledge automatically from large information sources (e.g., web data)
- The project was developed in the Max Planck Institute for Computer Science (Germany)

Fabian Suchanek

[Suchanek et al., 2007]

Yago extracted knowledge from three main sources



350 K classes
100 relations

10 M entities (instances)
120 M facts

Yago represents knowledge using triples subject-relation-object

- Turtle format (based on RDF):

```
#@ <id_42>
<Elvis_Presley> rdf:type <person>

<id_42> <occursSince> "1935-01-08"
<id_42> <occursUntil> "1977-08-16"
<id_42> <extractionSource>
<http://en.wikipedia.org/Elvis_Presley>
```

- Includes:
  - Fact identifiers (unique to Yago)
  - Temporal and spatial dimensions

Yago uses a taxonomy with 4 layers

Layer 1: Root:

- *Resource*: `rdfs:Resource`
- *Thing:* `owl:Thing`

Layer 2: WordNet classes:

- *Singer:*
  `<wordnet_singer_110599806> rdfs:subclassOf ...`

Layer 3: Wikipedia categories:

- *American rock singer:*
  `<wikicat_American_rock_singers>`
  `rdfs:subclassOf <wordnet_singer_110599806>`

Layer 4: Instances:

- *Elvis Presley:*
  `<Elvis_Presley> rdf:type  <wikicat_American_rock_singers>`

Website: `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/`

Departments ▸ Databases and Information Systems ▸ Research ▸ YAGO-NAGA ▸ YAGO

- Overview
- Demo
- Downloads
- Statistics
- Publications
- Linking
- Archive
- Acknowledgements
- FAQ

YAGO: A High-Quality Knowledge Base

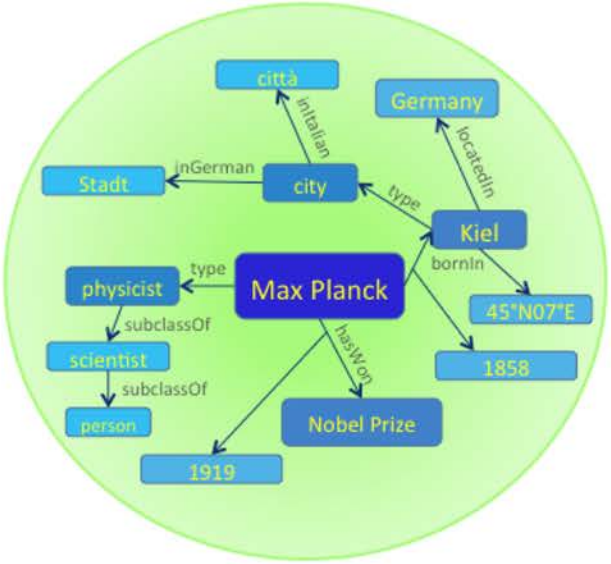
Overview

YAGO is a huge semantic knowledge base, derived from Wikipedia WordNet and GeoNames. Currently, YAGO has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities.

YAGO is special in several ways:

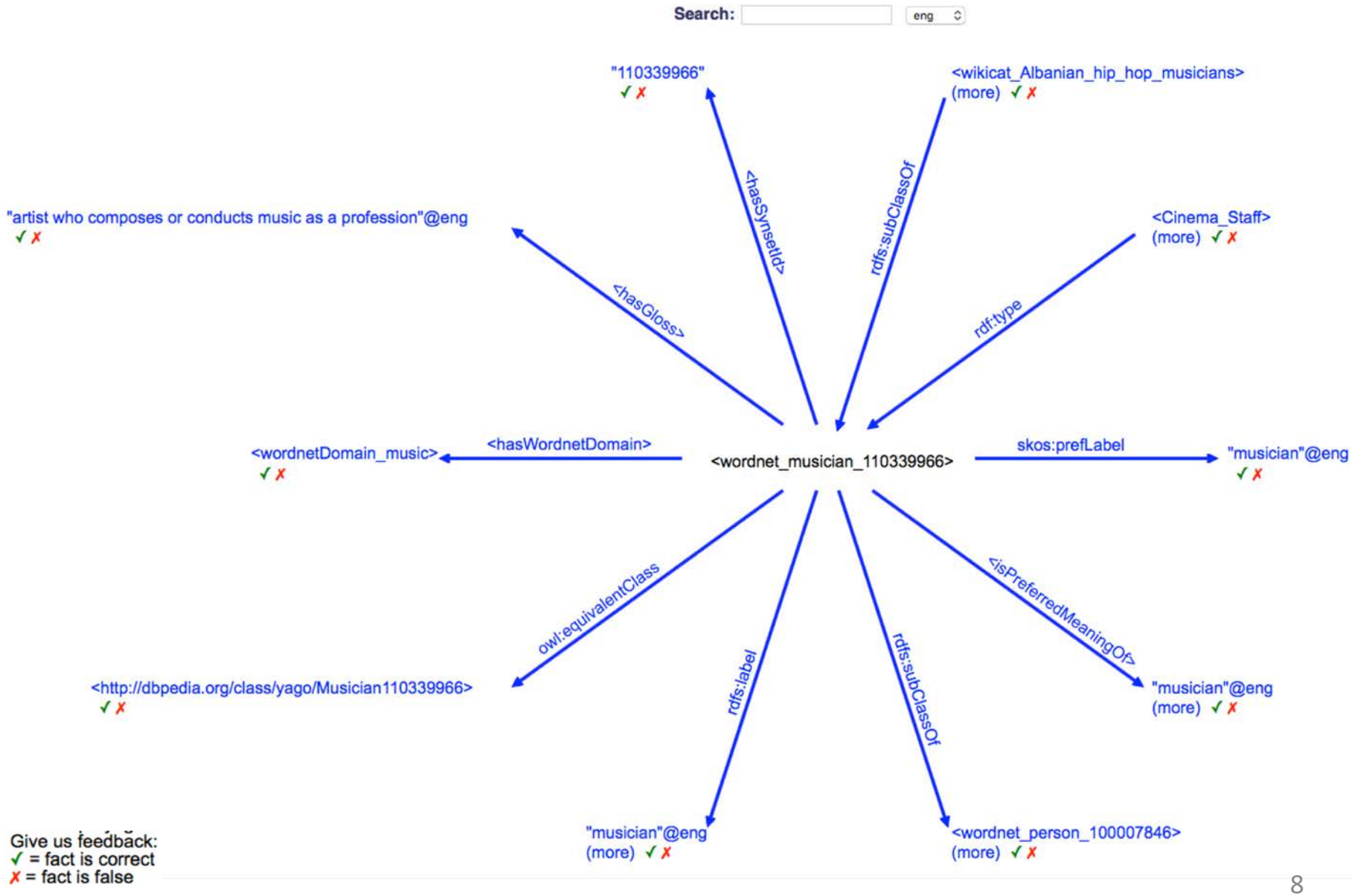
1. The accuracy of YAGO has been manually evaluated, proving a confirmed accuracy of 95%. Every relation is annotated with its confidence value.
2. YAGO combines the clean taxonomy of WordNet with the richness of the Wikipedia category system, assigning the entities to more than 350,000 classes.
3. YAGO is an ontology that is anchored in time and space. YAGO attaches a temporal dimension and a spacial dimension to many of its facts and entities.
4. In addition to a taxonomy, YAGO has thematic domains such as "music" or "science" from WordNet Domains.
5. YAGO extracts and combines entities and facts from 10 Wikipedias in different languages.

YAGO is developed jointly with the DBWeb group at Télécom ParisTech University.



Download YAGO

Example: Musician



Example: Musician

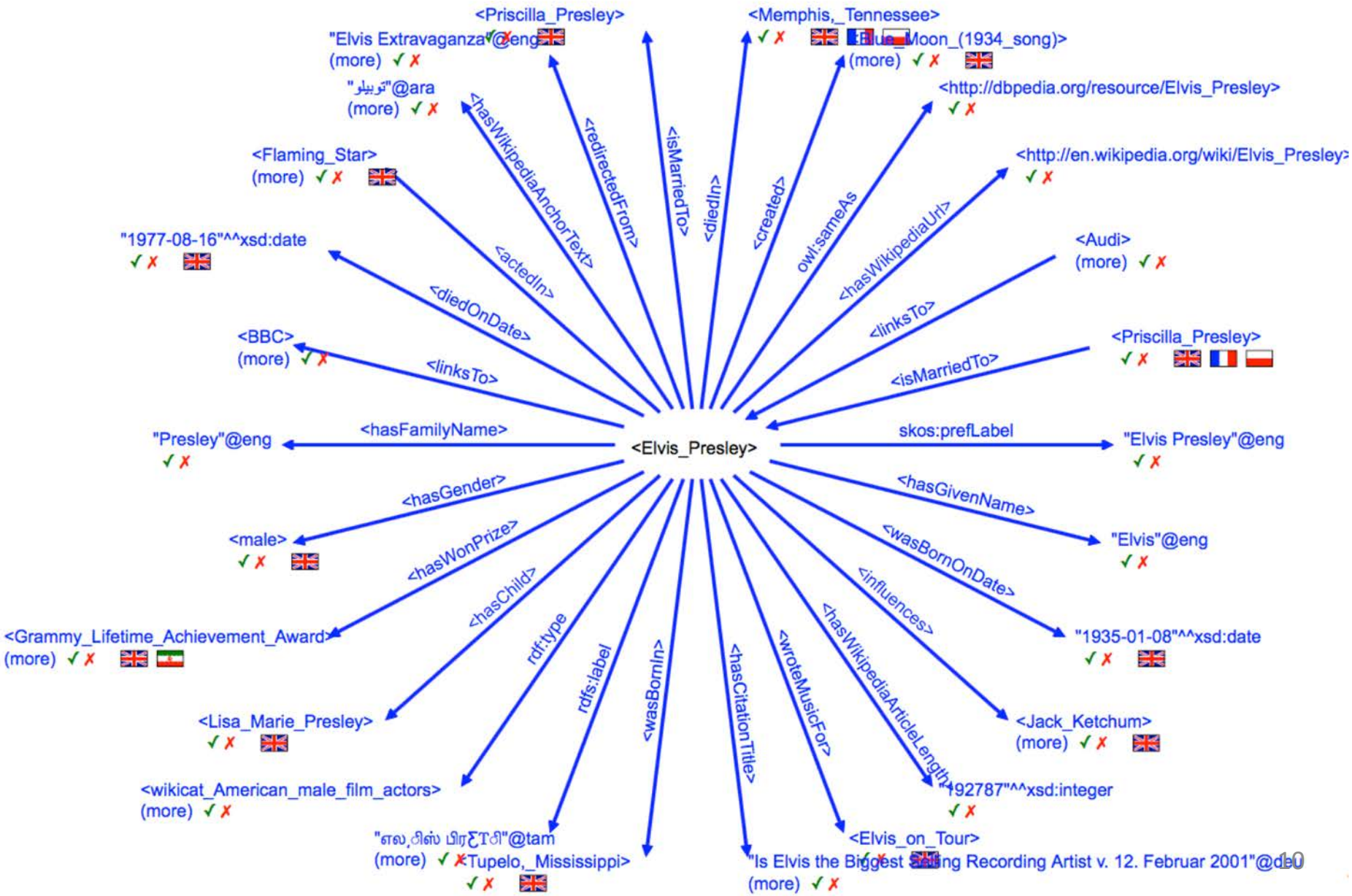
Search: eng

<wordnet_musician_110339966>

← <Joe_Doherty_(singer)> ← <Juan_Formell>	<hasMusicalRole>
← <wikicategory_14th-century_musicians> ← <wikicategory_15th-century_musicians> ← <wikicategory_16th-century_musicians> ← <wikicategory_17th-century_musicians> ← <wikicategory_18th-century_musicians> ← <wikicategory_19th-century_musicians> ← <wikicategory_20th-century_French_musicians> ← <wikicategory_20th-century_musicians> ← <wikicategory_21st-century_British_musicians> ← <wikicategory_21st-century_musicians> ← <wikicategory_Acid_house_musicians> ← <wikicategory_Acoustic_blues_musicians> ← <wikicategory_Afghan_musicians> ← <wikicategory_African-American_brass_musicians> ← <wikicategory_African-American_classical_musicians> ← <wikicategory_African-American_female_musicians> ← <wikicategory_African_American_musicians> ← <wikicategory_African-American_musicians> ← <wikicategory_African-American_Musicians> ← <wikicategory_African-American_rock_musicians> ...	rdfs:subClassOf
<A._A._Bondy> <Aadesh_Shrivastava> <Aafje_Heynis> <Aage_Emborg> <Aage_Fønss> <Aage_Haugland> <Aage_Kvalbein> <Aage_Oxenvad> <Aage_Samuelsen> <Aage_Stentoft> <Aage_Tanggaard> <Aakanksha_Jachak> <Aalap_Raju> <Aaliyah> <Aaltje_Noordewier-Reddingius> <Aamir_Saleem> <Aamir_Zaki> <Aapo_Häkkinen> <Aapo_Ilves> <Aaradhna> ...	rdf:type

rdfs:label	"ceoltóir"@gle "glasbenik"@slv "hudebník"@ces "mpiga muziki"@swh "musicant"@pms "músic"@cat "musician"@eng "musicista"@ita "musicista"@pms "músico"@glg "musico"@ita "músico"@por "Músico"@por "músico"@spa "músicu"@ast "musicus"@nld "musigan"@vol "Musikant"@deu "musikari"@eus "musiker"@dan ...
<hasSynsetId>	"110339966"
rdfs:subClassOf	<wordnet_artist_109812338> <wordnet_person_100007846>
<isPreferredMeaningOf>	"musician"@eng
skos:prefLabel	"musician"@eng
<hasGloss>	"artist who composes or conducts music as a profession"@eng

Example: Elvis Presley



Example query

Companies founded in the last 3 decades, together with their founders

Query

Id	Subject	Property	Object	Time		Location		Keywords
?id0:	?x	<created>	?c	during	1980,2010			
?id1:	?c	rdf:type	company					
?id2:								
?id3:								
?id4:								

query

Results

>>

	Id	Subject	Property	Object	Time	Location	Keywords
1	<id_jt2wvk_1gi_pb0jkw>	<Victor Scheinman>	<created>	<Automatix>	1980-01-01 ↓↑, 1980-01-31 ↓↑	-	closed form | ...
	null	<Automatix>	rdf:type	<wordnet company 108058098>	-	-	-
	null	<wordnet company 108058098>	rdfs:label	"company"@eng	-	-	-
2	<id_1enefsn_1gi_s6ldnn>	<Quincy Jones>	<created>	<Qwest Records>	1980-01-01 ↓↑, 1980-12-31 ↓↑	-	soul | big ...
	null	<Qwest Records>	rdf:type	<wordnet company 108058098>	-	-	-
	null	<wordnet company 108058098>	rdfs:label	"company"@eng	-	-	-
3	<id_my5ze3_1gi_1s81322>	<Jim O'Neal>	<created>	<Rooster Blues>	1980-01-01 ↓↑, 1980-12-31 ↓↑	-	American | Kansas ...
	null	<Rooster Blues>	rdf:type	<wordnet company 108058098>	-	-	-
	null	<wordnet company 108058098>	rdfs:label	"company"@eng	-	-	-
4	<id_1owh2x_1gi_1765v47>	<Art Evans>	<created>	<Dixie Chopper>	1980-01-01 ↓↑, 1980-12-31 ↓↑	-	U.S. | Claudine ...
	null	<Dixie Chopper>	rdf:type	<wordnet company 108058098>	-	-	-
	null	<wordnet company 108058098>	rdfs:label	"company"@eng	-	-	-

What are the strengths and the weaknesses of Yago?

- Strengths
  - Large number of instances (10M) and large number of facts about these instances (120M)
  - Spatial and temporal references
  - High accuracy 95% (manually evaluated of a sample of facts)
- Weaknesses
  - Few relations (130 relations)
  - Limited representation and limited inference (triples, taxonomic)
  - Medium number of general knowledge (350K classes)
  - Limited knowledge sources (Wikipedia, WordNet, ...)

Nell generates the content of the knowledge base in an iterative loop

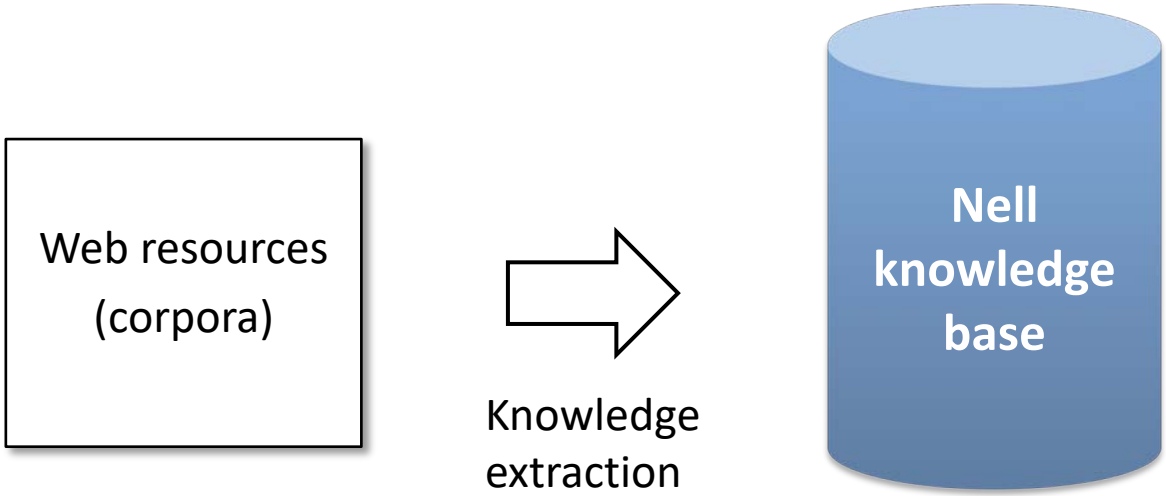
[Mitchell et al., 2015]

- Nell: Never-Ending Language Learning
- Nell extracts knowledge automatically from web data using what it has been learned to learn new things
- The project was developed in the Carnegie Mellon University

Research team

Tom Mitchell

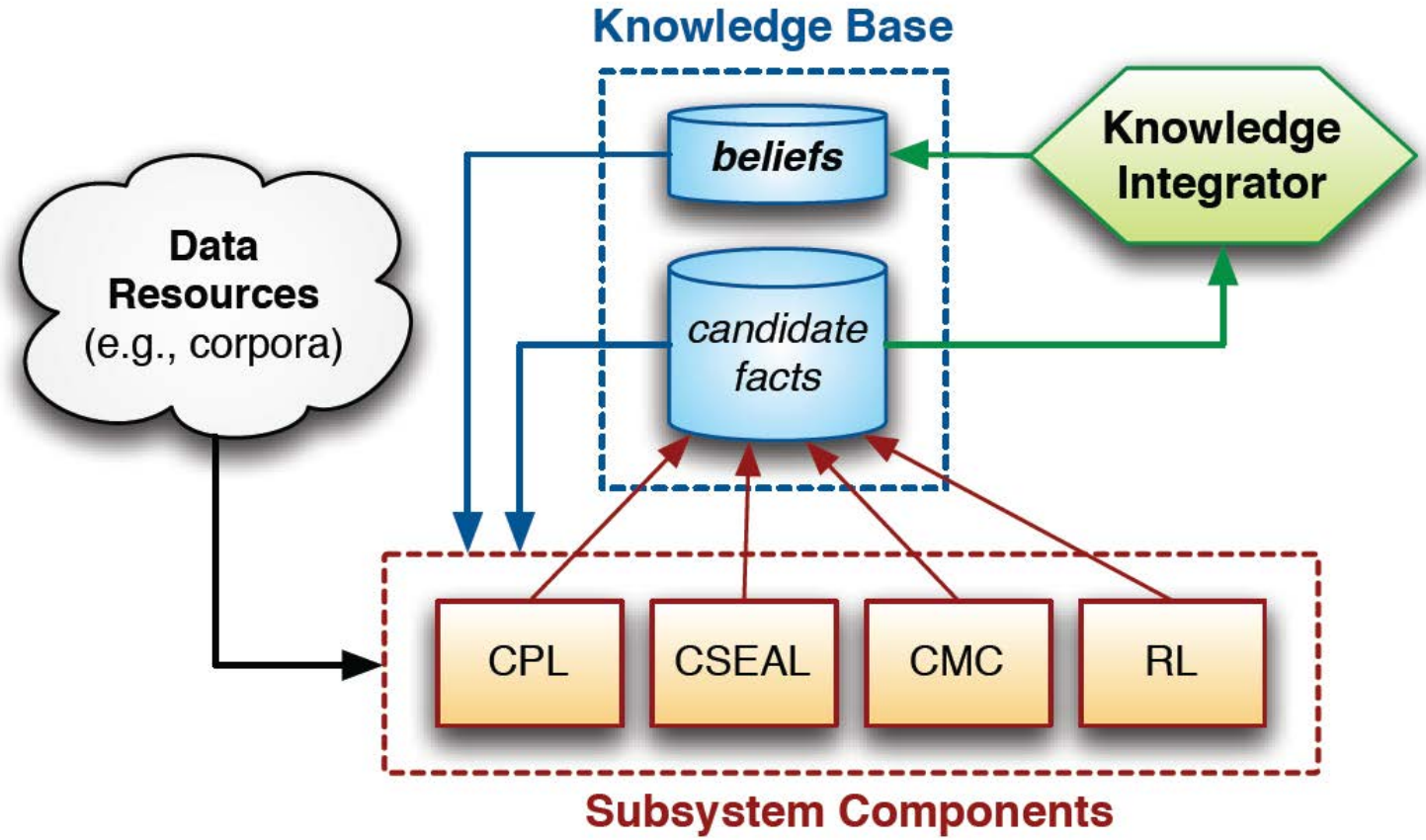
Nell extracts knowledge from web resources



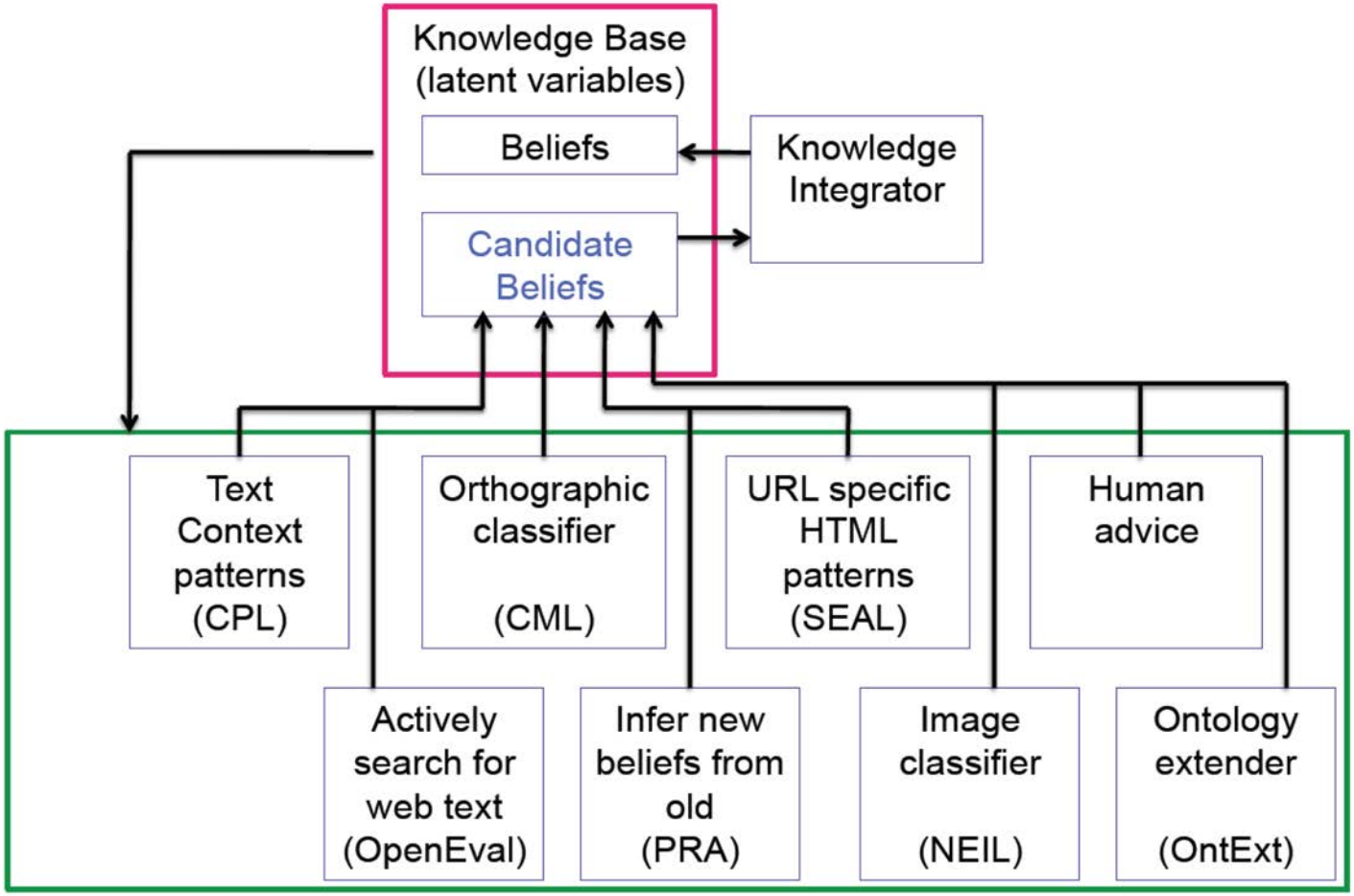
500 million
web pages

300 categories
900 relations
2M beliefs
(size in 2015)

Nell uses several specialized components to extract knowledge



The architecture in 2015 included eight components to extract knowledge



Nell represents knowledge with categories and relations

- Categories (330):
  - person, musician, sportsTeam, fruit, emotion
- Relations (934):
  - playsInstrument(musician, instrument)
  - playsOnTeam(athlete, sportsTeam)
- Instances (2M):
  - Barack Obama is a person
  - Barack Obama is a politician
  - <George Harrison, guitar> is an instance of playsInstrument().
  - <Jason Giambi, Yankees> is an instance of playsOnTeam()

Web site: `http://rtw.ml.cmu.edu/rtw/`

Read the Web

Research Project at Carnegie Mellon University

Home | Project Overview | Resources & Data | Publications | People

NELL: Never-Ending Language Learning

Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:

- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.

So far, NELL has accumulated over 50 million candidate beliefs by reading the web, and it is considering these at different levels of confidence. NELL has high confidence in 2,810,379 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or @cmunell on Twitter, browse and download its knowledge base, read more about our technical approach, or join the discussion group.

Recently-Learned Facts

twitter

Refresh

instance	iteration	date learned	confidence
james_finlay is a company	1111	06-jul-2018	95.6
hideki_okajima is a Mexican person	1111	06-jul-2018	100.0
beetroot_juice is a beverage	1111	06-jul-2018	100.0
development_education is a political issue	1111	06-jul-2018	100.0

Example: music instrument

NELL Knowledge Base Browser
CMU Read the Web Project

Search
log in | preferences | help/instructions | feedback

categories | relations

- everypromotedthing
  - abstractthing
    - event
      - convention
      - musicfestival
      - protestevent
      - meetingeventtitle
      - conference
        - mlconference
      - weatherphenomenon
      - sportsevent
        - sportsgame
        - race
        - olympics
        - grandprix
      - crimeorcharge
      - earthquakeevent
      - election
      - bombingevent
      - militaryeventtype
        - militaryconflict
      - productlaunchevent
      - filmfestival
      - roadaccidentevent
      - meetingeventtype
      - eventoutcome
    - mlalgorithm
    - physiologicalcondition
      - disease
      - nondiseasecondition
    - religion
    - creativework
      - musicalbum
      - book
      - poem
      - lyrics

musicinstrument
(category)

View list | map | metadata
3,599 instances, 1 page

A Musical instrument is a device constructed or modified with the purpose of making music. In principle, anything that produces sound, and can somehow be controlled by a musician, can serve as a musical instrument. (See e.g. the hardart.)

instance	iteration	date learned	confidence
acoustic_and_electronic_percussion	249	13-may-2011	100.0
acoustic_archtop_guitar	249	13-may-2011	100.0
acoustic_guitar	883	02-nov-2014 (Seed)	100.0
acoustic_upright_bass	249	13-may-2011	100.0
aeolian_harp	249	13-may-2011	100.0
african_djembe	249	13-may-2011	100.0
african_drums	249	13-may-2011	100.0
african_guitar	249	13-may-2011	100.0
alto_clarinet	249	13-may-2011	100.0
alto_flute	229	08-apr-2011	100.0
alto_recorder	249	13-may-2011	100.0
alto_sax	199	08-feb-2011	100.0
antique_cymbals	249	13-may-2011	100.0
autoharp	249	13-may-2011	100.0
baby_grand_pianos	429	06-oct-2011	100.0
bamboo_flute	102	22-may-2010	100.0
banjo	120	19-jun-2010 (Seed)	100.0
banjo_ukulele	249	13-may-2011	100.0
baritone_guitar	125	21-jun-2010	100.0
baritone_saxophone	116	04-jun-2010	100.0
baroque_guitar	490	21-jan-2012	100.0
bass	601	25-jun-2012 (Seed)	100.0
basset_horn	233	13-apr-2011	100.0
bass_clarinet	111	03-jun-2010	100.0
bass_drum	114	04-jun-2010	100.0

Example: guitar

NELL Knowledge Base Browser
CMU Read the Web Project

Search
log in | preferences | help/instructions | feedback

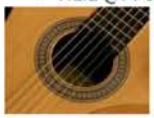
categories | relations

- everypromotedthing
  - abstractthing
    - event
      - convention
      - musicfestival
      - protestevent
      - meetingeventtitle
      - conference
        - mlconference
      - weatherphenomenon
      - sportsevent
        - sportsgame
        - race
        - olympics
        - grandprix
      - crimeorcharge
      - earthquakeevent
      - election
      - bombingevent
      - militaryeventtype
        - militaryconflict
      - productlaunchevent
      - filmfestival
      - roadaccidentevent
      - meetingeventtype
      - eventoutcome
    - mlalgorithm
    - physiologicalcondition
      - disease
      - nondiseasecondition
    - religion
    - creativework
      - musicalbum
      - book
      - poem
      - lyrics
      - visualartform
      - movie
      - musicsong
      - televisionshow
    - chemical
    - date
      - dayofweek
      - year
      - month
      - dateliteral
    - hobby
    - game

- NEIL @770 (100%) on 21-sep-2013

- NEIL @770 (100%) on 21-sep-2013

- NEIL @770 (100%) on 21-sep-2013



- NEIL @770 (100%) on 21-sep-2013

- instrumentplayedbymusician
  - bb_king (100.0%)
    - CPL @749 (75.0%) on 06-jul-2013 ["arg1 tune his arg2" "arg1 plays blues arg2"] using (bb_king, guitar)
    - SEAL @174 (100.0%) on 08-dec-2010 [1 2 3] using (bb_king, guitar)
  - ben_harper (100.0%)
    - OE @803 (94.8%) on 13-jan-2014 [ ] using (ben_harper, guitar)
    - CPL @215 (99.2%) on 26-feb-2011 ["arg1 on lead arg2" "arg1 guesting on arg2" "arg2 licks of arg1" "arg1 plays slide arg2" "arg1 received his first arg2" "arg1 was given his first arg2" "arg1 plays bass arg2"] using (ben_harper, guitar)
    - SEAL @628 (100.0%) on 26-aug-2012 [1 2] using (ben_harper, guitar)
  - billie_joe_armstrong (100.0%)
    - CPL @668 (93.8%) on 12-dec-2012 ["arg1 sing and play arg2" "arg1 model Gibson arg2" "arg1 on lead vocals and arg2" "arg1 on rhythm arg2"] using (billie_joe_armstrong, guitar)
    - OE @799 (100.0%) on 25-dec-2013 [http://www.uberproaudio.com/who-plays-what/130-green-day-billie-joe-armstrongs-guitar-gear-rig-and-equipment http://www2.gibson.com/Products/Electric-Guitars/Les-Paul/Gibson-USA/Billie-Joe-Armstrong-Les-Paul-Jr.aspx http://www.uberproaudio.com/who-plays-what/130-green-day-billie-joe-armstrongs-guitar-gear-rig-and-equipment http://www.premierguitar.com/articles/Rig_Rundown_Green_Day http://articles.latimes.com/2012/sep/24/entertainment/la-et-ms-green-day-billie-joe-armstrong-rant-las-vegas-rehab-20120923 http://www.uberproaudio.com/who-plays-what/130-green-day-billie-joe-armstrongs-guitar-gear-rig-and-equipment] using (billie_joe_armstrong, guitar)
    - SEAL @230 (100.0%) on 08-apr-2011 [1 2] using (billie_joe_armstrong, guitar)
  - bob_dylan (100.0%)
    - Seed
    - CPL @215 (100.0%) on 26-feb-2011 ["arg1 played an electric arg2" "arg1 strumming his arg2" "arg2 tabs or chords from arg1" "arg2 tabs or chords for arg1" "arg1 plays his arg2" "arg2 and harmonica to arg1" "arg1 on acoustic arg2" "arg1 sang and played arg2" "arg1 singing and playing arg2" "arg1 songs on his arg2" "arg2 and sounding like arg1" "arg1 play electric arg2" "arg1 songs on acoustic arg2" "arg1 with his acoustic arg2" "arg1 is strumming his arg2" "arg2 on albums by arg1" "arg1 enjoys playing arg2"] using (bob_dylan, guitar)
    - SEAL @179 (87.5%) on 15-dec-2010 [1 2 3] using (bob_dylan, guitar)
  - buddy_guy (100.0%)
    - CPL @664 (96.9%) on 30-nov-2012 ["arg1 on acoustic arg2" "arg2 solo by arg1" "arg1 on rhythm arg2" "arg1 play his arg2" "arg2 like jimi Hendrix or arg1"] using (buddy_guy, guitar)
    - SEAL @179 (100.0%) on 15-dec-2010 [1 2 3 4 5 6 7 8] using (buddy_guy, guitar)

Example of relation

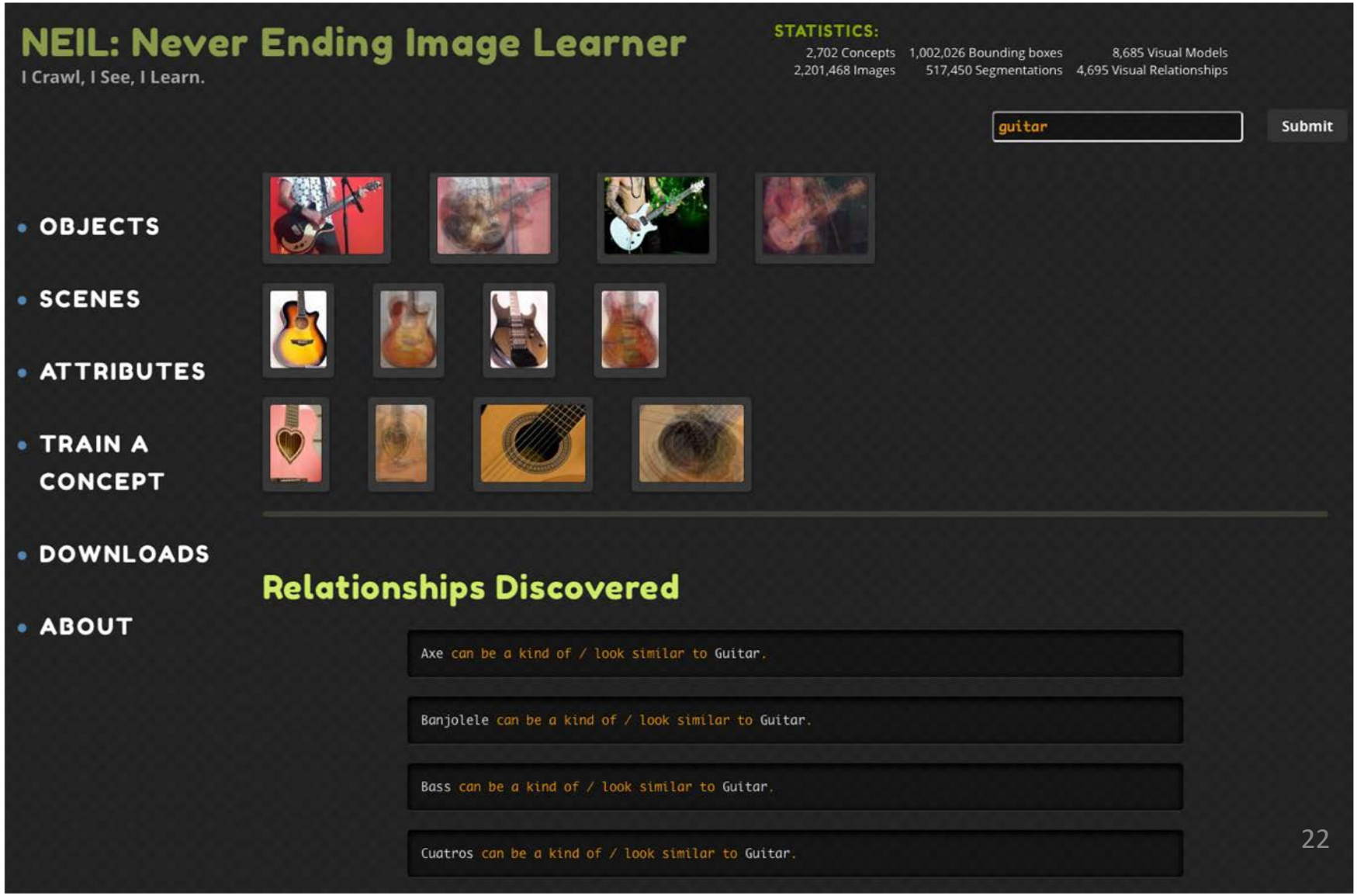
fooddecreasestheriskofdisease

(relation: domain food, range disease) *Describes the foods that can reduce the risks of having a disease*

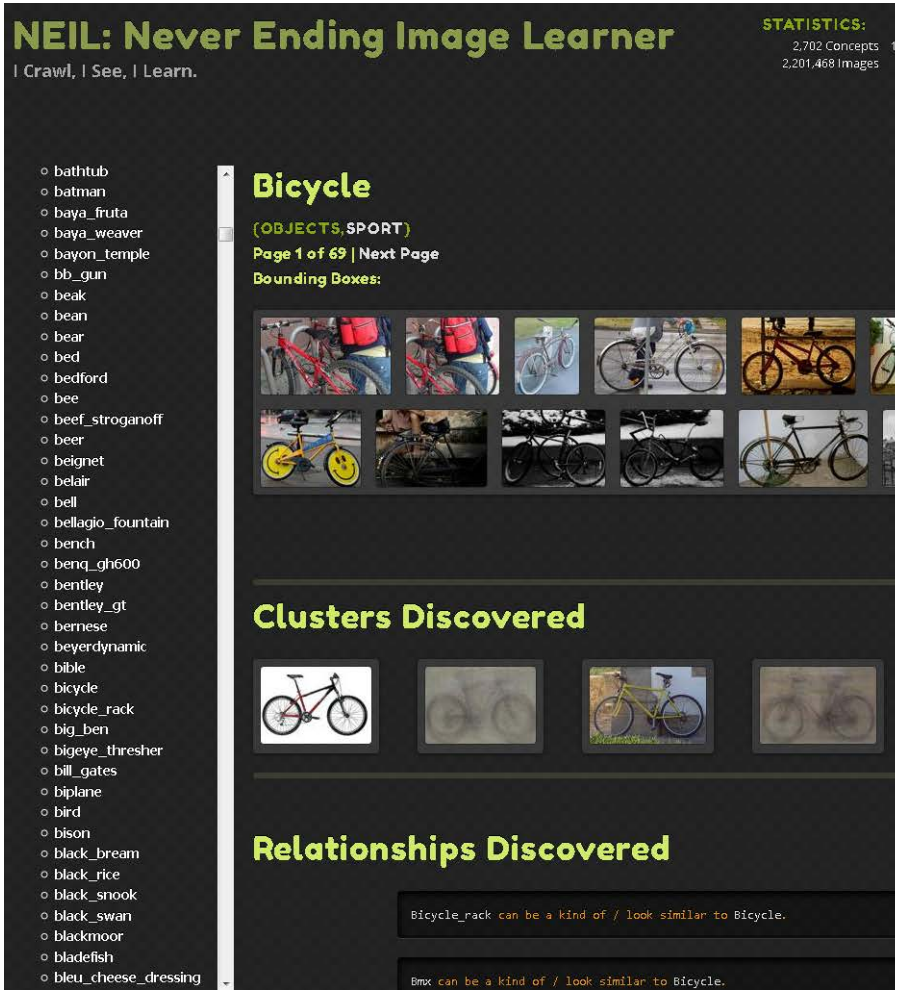
See metadata for
fooddecreasestheriskofdisease
245 instances, 1 page

instance	iteration	date learned	confidence
mushroom, breast_cancer	785	04-nov-2013	(Seed) 100.0
alcohol, heart_disease	670	17-dec-2012	100.0
dairy_foods, cancers	631	09-sep-2012	(Seed) 100.0
fish_oil, heart_disease	728	29-apr-2013	100.0
fruits, breast_cancer	670	17-dec-2012	100.0
fruits, cancer	670	17-dec-2012	(Seed) 100.0
fruits, heart_disease	670	17-dec-2012	100.0
peanuts, heart_disease	670	17-dec-2012	(Seed) 100.0
antioxidants, cancer	557	29-apr-2012	(Seed) 100.0
antioxidants, damage	557	29-apr-2012	100.0
calcium, cancer	670	17-dec-2012	100.0
calcium, osteoporosis	670	17-dec-2012	(Seed) 100.0
fats, cancer	655	02-nov-2012	(Seed) 100.0
fats, heart_disease	655	02-nov-2012	(Seed) 100.0
fiber, cancer	435	18-oct-2011	(Seed) 100.0
fiber, chd	435	18-oct-2011	(Seed) 100.0
fiber, colon	551	19-apr-2012	(Seed) 100.0

Neil (Never Ending Image Learning) extracts knowledge from images



Neil examples



What are the strengths and the weaknesses of Nell?

- Strengths
  - Learned things help to learn new things
  - Uses a combination of methods for knowledge extraction
  - Includes knowledge extraction from images (Neil)
  - More than 2M beliefs (in 2015)
- Weaknesses
  - Limited representation based on few categories (300) and few relations (900)
  - Limited inference

Work citation in APA style:

Molina, M. (2019). Common sense reasoning [Lecture slides]. OpenCourseWare, Universidad Politécnica de Madrid. Retrieved from http://ocw.upm.es/course