Language Understanding

14 - Information Extraction and Knowledge Graphs

Hossein Zeinali

What is Knowledge Graph?

- *"A huge knowledge graph of interconnected entities and their attributes".* Amit Singhal, Senior Vice President at Google
- *"A knowledge based used by Google to enhance its search engine's results with semantic-search information gathered from a wide variety of sources"* http://en.wikipedia.org/wiki/Knowledge_Graph
- Based on information derived from many sources including *Freebase, CIA World Factbook, Wikipedia*
- By May 2020, contains 5 billion entities and more than 500 billion facts about and relationships between these different objects

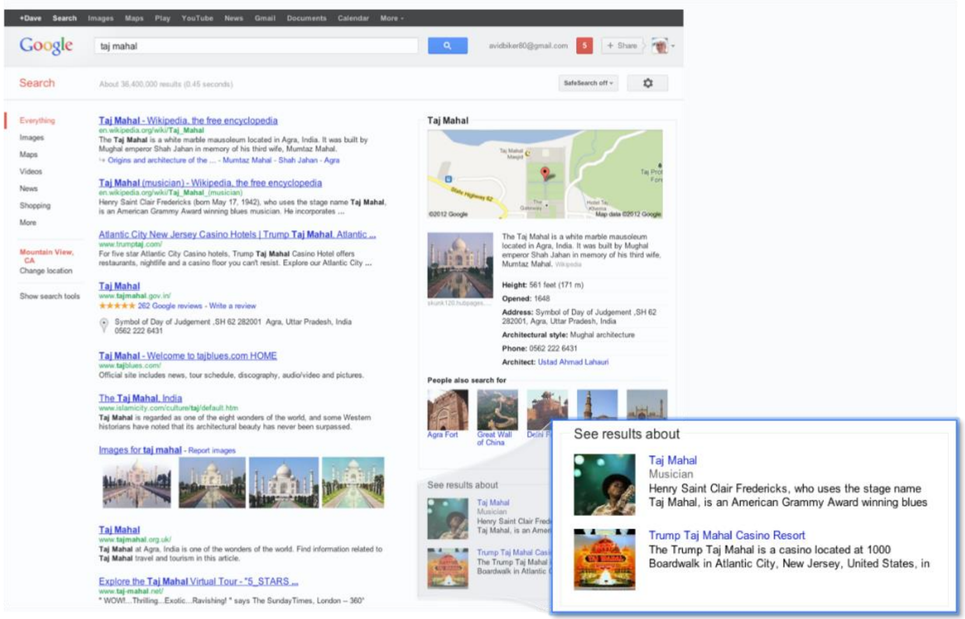
Who Constructs Knowledge Graphs?

- Google
- Amazon
- Microsoft
- Bloomberg (business intelligence)
- National Cancer Institute (part of NIH)
- National Center for Biomedical Ontology (funded by NIH)
- Specialized knowledge graphs:
    - Health / life sciences (health-lifesci.schema.org)
    - Earth science
    - Agriculture
    - NCI Enterprise Vocabulary Services (evs.nci.nih.gov)

Google Knowledge Graph (GKG)

GKG enhances Google Search in three main ways:

- **Find the right thing:** Deals with the ambiguity of the language



taj mahal

Search

About 36,400,000 results (0.45 seconds)

Taj Mahal - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Taj_Mahal
The **Taj Mahal** is a white marble mausoleum located in Agra, India. It was built by Mughal emperor Shah Jahan in memory of his third wife, Mumtaz Mahal.
→ Origins and architecture of the ... - Mumtaz Mahal - Shah Jahan - Agra

Taj Mahal (musician) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Taj_Mahal_(musician)
Henry Saint Clair Fredericks (born May 17, 1942), who uses the stage name **Taj Mahal**, is an American Grammy Award winning blues musician. He incorporates ...

Atlantic City New Jersey Casino Hotels | Trump Taj Mahal, Atlantic ...
www.trumptaj.com/
For five star Atlantic City Casino hotels, Trump **Taj Mahal** Casino Hotel offers restaurants, nightlife and a casino floor you can't resist. Explore our Atlantic City ...

Taj Mahal
www.tajmahal.gov.in/
★★★★★ 262 Google reviews - Write a review
Symbol of Day of Judgement ,SH 62 282001 Agra, Uttar Pradesh, India
0562 222 6431

Taj Mahal - Welcome to tajblues.com HOME
www.tajblues.com/
Official site includes news, tour schedule, discography, audio/video and pictures.

The Taj Mahal, India
www.islamicity.com/culture/taj/default.htm
Taj Mahal is regarded as one of the eight wonders of the world, and some Western historians have noted that its architectural beauty has never been surpassed.

Images for taj mahal - Report images

Taj Mahal
www.tajmahal.org.uk/
Taj Mahal at Agra, India is one of the wonders of the world. Find information related to **Taj Mahal** travel and tourism in this article.

Explore the Taj Mahal Virtual Tour - "5_STARS ...
www.taj-mahal.net/
" WOW!...Thrilling...Exotic...Ravishing! " says The SundayTimes, London – 360°

Taj Mahal

©2012 Google   Map data ©2012 Google

The Taj Mahal is a white marble mausoleum located in Agra, India. It was built by Mughal emperor Shah Jahan in memory of his third wife, Mumtaz Mahal. Wikipedia

skunk120.hubpages....

Height: 561 feet (171 m)

Opened: 1648

Address: Symbol of Day of Judgement ,SH 62 282001, Agra, Uttar Pradesh, India

Architectural style: Mughal architecture

Phone: 0562 222 6431

Architect: Ustad Ahmad Lahauri

People also search for

Agra Fort
Great Wall of China
Delhi F

See results about

Taj Mahal
Musician
Henry Saint Clair Fredericks, who uses the stage name Taj Mahal, is an American Grammy Award winning blues

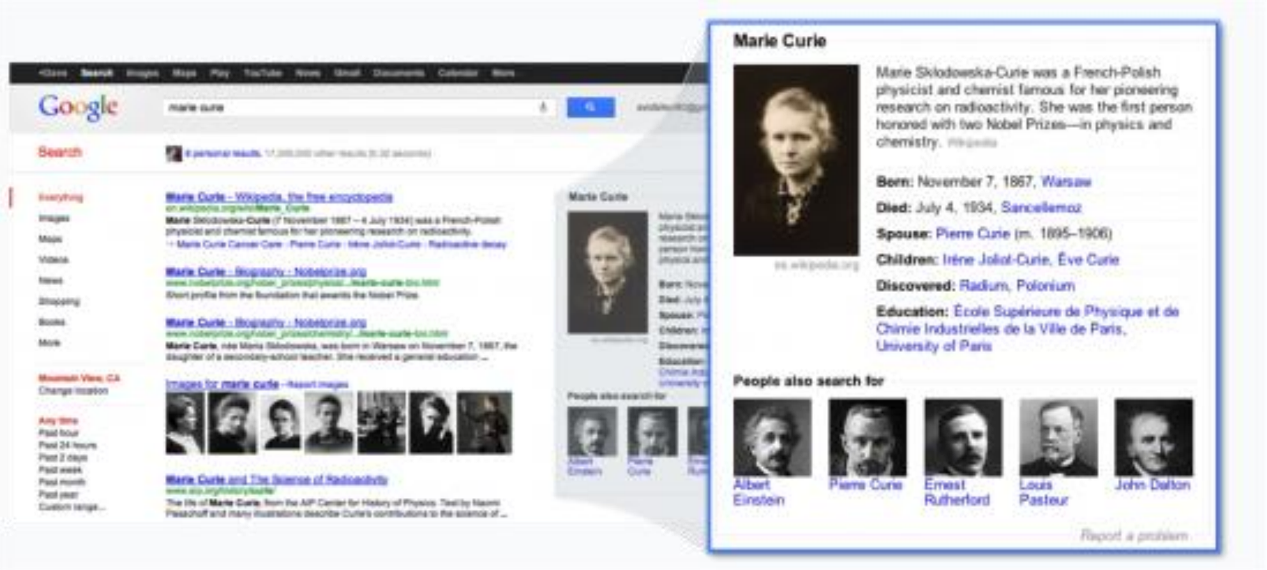
Trump Taj Mahal Casino Resort
The Trump Taj Mahal is a casino located at 1000 Boardwalk in Atlantic City, New Jersey, United States, in

See results about
Taj Mahal
Musician
Henry Saint Clair Fred
Taj Mahal, is an Amer
Trump Taj Mahal Casi
The Trump Taj Mahal
Boardwalk in Atlantic

Google Knowledge Graph (GKG)

GKG enhances Google Search in three main ways:

- **Summaries:** summarize relevant content around that topic, including key facts about the entity



4 personal results. 17,200,000 other results (0.32 seconds)

Marie Curie - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Marie_Curie
Marie Skłodowska-Curie (7 November 1867 – 4 July 1934) was a French-Polish physicist and chemist famous for her pioneering research on radioactivity.
+ Marie Curie Cancer Care - Pierre Curie - Irène Joliot-Curie - Radioactive decay

Marie Curie - Biography - Nobelprize.org
www.nobelprize.org/nobel_prizes/physics/...marie-curie-bio.html
Short profile from the foundation that awards the Nobel Prize.

Marie Curie - Biography - Nobelprize.org
www.nobelprize.org/nobel_prizes/chemistry/...marie-curie-bio.html
Marie Curie, née Maria Sklodowska, was born in Warsaw on November 7, 1867, the daughter of a secondary-school teacher. She received a general education ...

Images for marie curie - Report images

Marie Curie and The Science of Radioactivity
www.aip.org/history/curie/
The life of Marie Curie, from the AIP Center for History of Physics. Text by Naomi Pasachoff and many illustrations describe Curie's contributions to the science of ...

Marie Curie

Marie Sklodowska-Curie was a French-Polish physicist and chemist famous for her pioneering research on radioactivity. She was the first person honored with two Nobel Prizes—in physics and chemistry. Wikipedia

en.wikipedia.org

Born: November 7, 1867, Warsaw

Died: July 4, 1934, Sancellemoz

Spouse: Pierre Curie (m. 1895–1906)

Children: Irène Joliot-Curie, Ève Curie

Discovered: Radium, Polonium

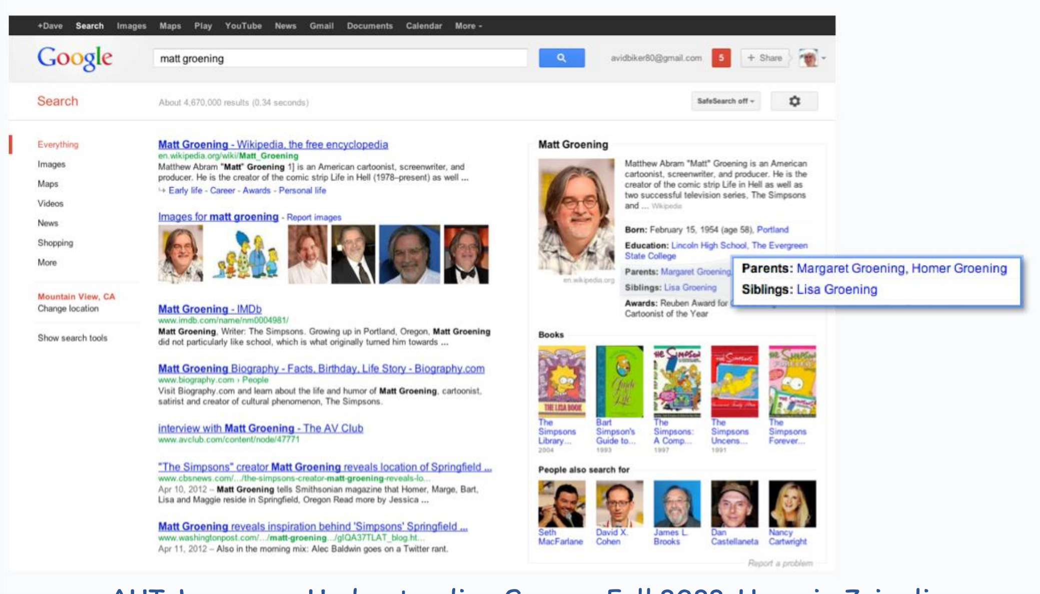
Education: École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris, University of Paris

People also search for

Google Knowledge Graph (GKG)

GKG enhances Google Search in three main ways:

- **Deeper and broader information:** reveal new facts, anticipate what the next questions and provide the information beforehand



matt groening

About 4,670,000 results (0.34 seconds)

Matt Groening - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Matt_Groening**
Matthew Abram "**Matt**" **Groening** 1] is an American cartoonist, screenwriter, and producer. He is the creator of the comic strip Life in Hell (1978–present) as well ...
↳ Early life - Career - Awards - Personal life

Images for **matt groening** - Report images

Matt Groening - IMDb
www.imdb.com/name/nm0004981/
Matt Groening, Writer: The Simpsons. Growing up in Portland, Oregon, **Matt Groening** did not particularly like school, which is what originally turned him towards ...

Matt Groening Biography - Facts, Birthday, Life Story - Biography.com
www.biography.com › People
Visit Biography.com and learn about the life and humor of **Matt Groening**, cartoonist, satirist and creator of cultural phenomenon, The Simpsons.

interview with **Matt Groening** - The AV Club
www.avclub.com/content/node/47771

"The Simpsons" creator **Matt Groening** reveals location of Springfield ...
www.cbsnews.com/.../the-simpsons-creator-**matt-groening**-reveals-lo...
Apr 10, 2012 – **Matt Groening** tells Smithsonian magazine that Homer, Marge, Bart, Lisa and Maggie reside in Springfield, Oregon Read more by Jessica ...

Matt Groening reveals inspiration behind 'Simpsons' Springfield ...
www.washingtonpost.com/.../**matt-groening**.../glQA37TLAT_blog.ht...
Apr 11, 2012 – Also in the morning mix: Alec Baldwin goes on a Twitter rant.

Matt Groening

Matthew Abram "Matt" Groening is an American cartoonist, screenwriter, and producer. He is the creator of the comic strip Life in Hell as well as two successful television series, The Simpsons and ... Wikipedia

en.wikipedia.org

Born: February 15, 1954 (age 58), Portland

Education: Lincoln High School, The Evergreen State College

Parents: Margaret Groening

Siblings: Lisa Groening

Awards: Reuben Award for
Cartoonist of the Year

Parents: Margaret Groening, Homer Groening
Siblings: Lisa Groening

Books

The Simpsons Library...	Bart Simpson's Guide to...	The Simpsons: A Comp...	The Simpsons Uncens...	The Simpsons Forever...
2004	1993	1997	1991	

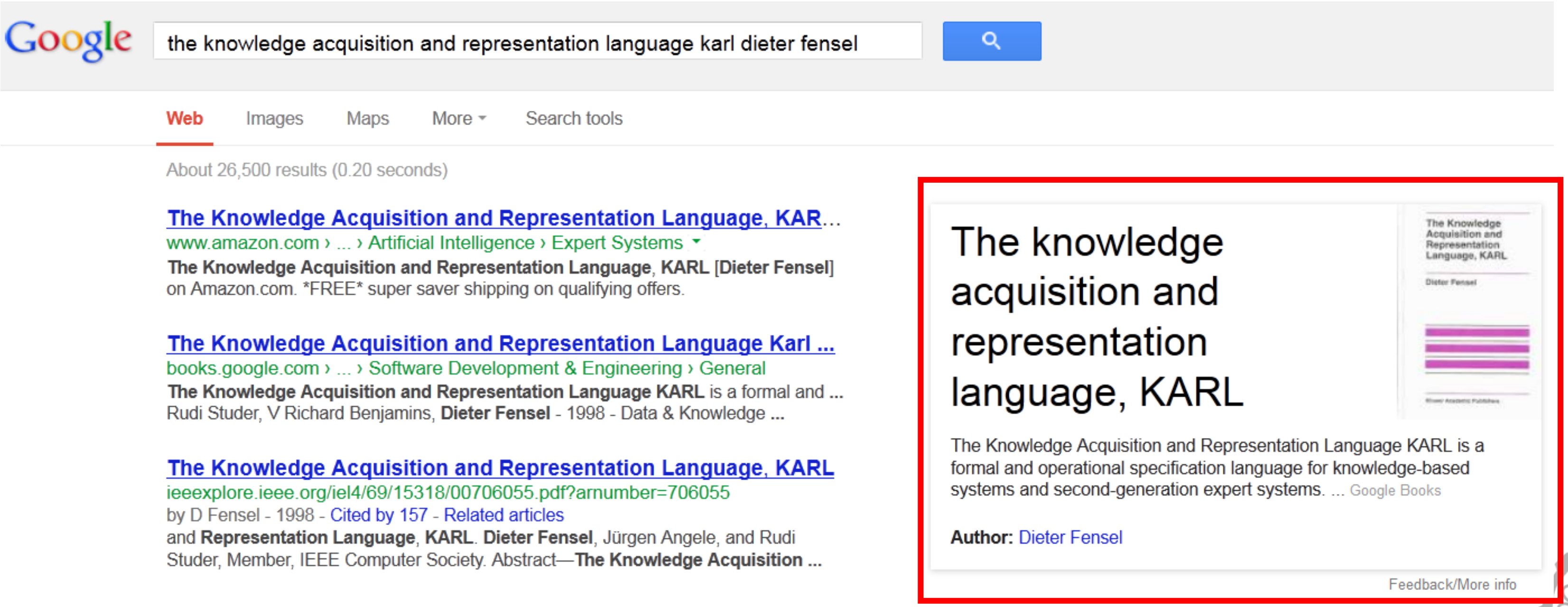
People also search for

Seth MacFarlane · David X. Cohen · James L. Brooks · Dan Castellaneta · Nancy Cartwright

Report a problem

How GKG is used?

- Explore your search



the knowledge acquisition and representation language karl dieter fensel

About 26,500 results (0.20 seconds)

The Knowledge Acquisition and Representation Language, KAR...
www.amazon.com › ... › Artificial Intelligence › Expert Systems
The Knowledge Acquisition and Representation Language, **KARL** [**Dieter Fensel**] on Amazon.com. *FREE* super saver shipping on qualifying offers.

The Knowledge Acquisition and Representation Language Karl ...
books.google.com › ... › Software Development & Engineering › General
The Knowledge Acquisition and Representation Language KARL is a formal and ...
Rudi Studer, V Richard Benjamins, **Dieter Fensel** - 1998 - Data & Knowledge ...

The Knowledge Acquisition and Representation Language, KARL
ieeexplore.ieee.org/iel4/69/15318/00706055.pdf?arnumber=706055
by D Fensel - 1998 - Cited by 157 - Related articles
and **Representation Language**, **KARL**. **Dieter Fensel**, Jürgen Angele, and Rudi Studer, Member, IEEE Computer Society. Abstract—**The Knowledge Acquisition** ...

The knowledge acquisition and representation language, KARL

The Knowledge Acquisition and Representation Language KARL is a formal and operational specification language for knowledge-based systems and second-generation expert systems. ... Google Books

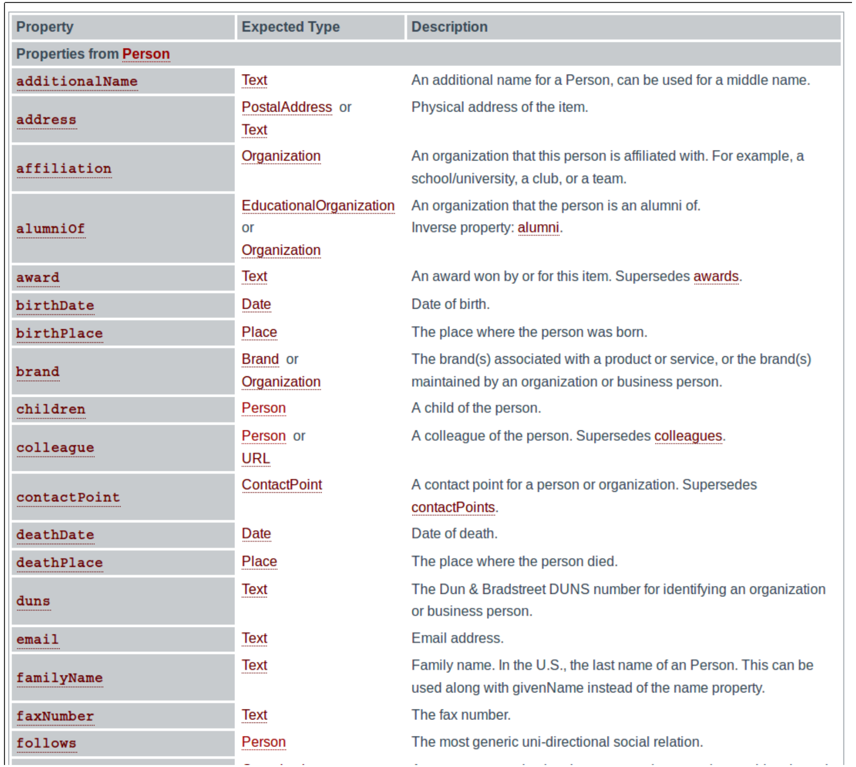
Author: Dieter Fensel

Feedback/More info

Sampling of GKG Entities

- Book
- BookSeries
- EducationalOrganization
- Event
- GovernmentOrganization
- LocalBusiness
- Movie
- MovieSeries
- MusicAlbum
- MusicGroup
- MusicRecording
- Organization
- Periodical
- Person
- Place
- SportsTeam
- TVEpisode
- TVSeries
- VideoGame
- VideoGameSeries
- WebSite

“Person” Schema



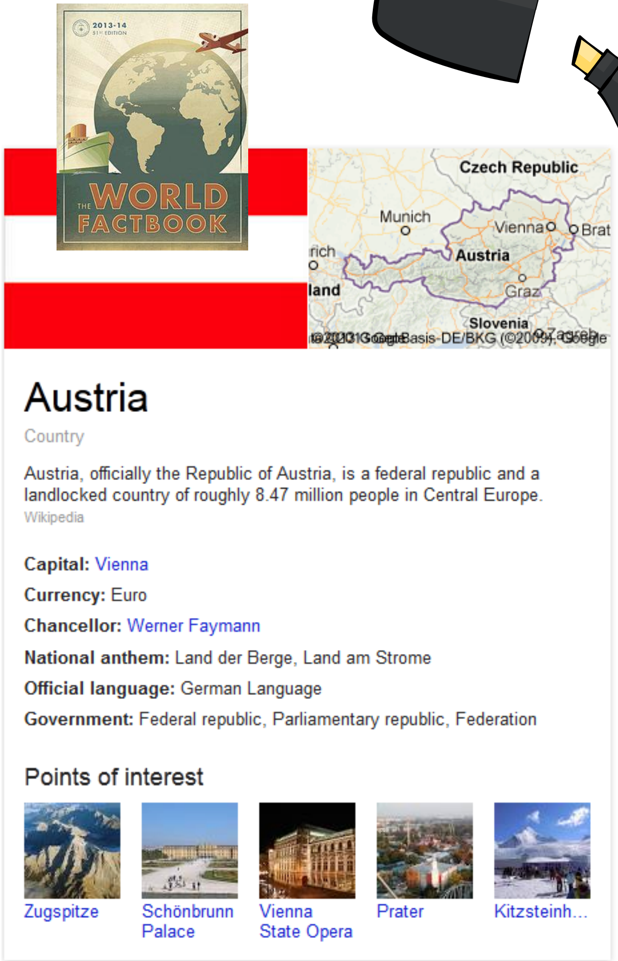
Property	Expected Type	Description
Properties from Person		
additionalName	Text	An additional name for a Person, can be used for a middle name.
address	PostalAddress or Text	Physical address of the item.
affiliation	Organization	An organization that this person is affiliated with. For example, a school/university, a club, or a team.
alumniOf	EducationalOrganization or Organization	An organization that the person is an alumni of. Inverse property: alumni.
award	Text	An award won by or for this item. Supersedes awards.
birthDate	Date	Date of birth.
birthPlace	Place	The place where the person was born.
brand	Brand or Organization	The brand(s) associated with a product or service, or the brand(s) maintained by an organization or business person.
children	Person	A child of the person.
colleague	Person or URL	A colleague of the person. Supersedes colleagues.
contactPoint	ContactPoint	A contact point for a person or organization. Supersedes contactPoints.
deathDate	Date	Date of death.
deathPlace	Place	The place where the person died.
duns	Text	The Dun & Bradstreet DUNS number for identifying an organization or business person.
email	Text	Email address.
familyName	Text	Family name. In the U.S., the last name of an Person. This can be used along with givenName instead of the name property.
faxNumber	Text	The fax number.
follows	Person	The most generic uni-directional social relation.

Data Sources Used for GKG

CIA World Factbook

- CIA World Factbook is a reference resource produced by the Central Intelligence Agency of the United States with almanac-style information about the countries of the world.

- GKG integrates information about geography, government, economy, etc. from CIA World Factbook.



Austria

Country

Austria, officially the Republic of Austria, is a federal republic and a landlocked country of roughly 8.47 million people in Central Europe.
Wikipedia

Capital: Vienna

Currency: Euro

Chancellor: Werner Faymann

National anthem: Land der Berge, Land am Strome

Official language: German Language

Government: Federal republic, Parliamentary republic, Federation

Points of interest

Zugspitze
Schönbrunn Palace
Vienna State Opera
Prater
Kitzsteinh...

Freebase

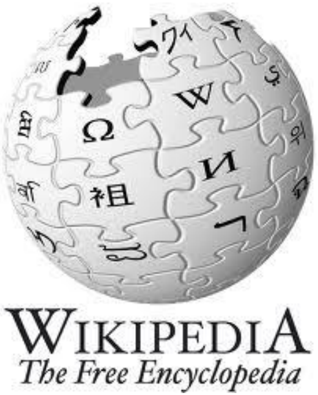
- Freebase is large collaborative knowledge base, developed by Metaweb and acquired by Google in 2010.

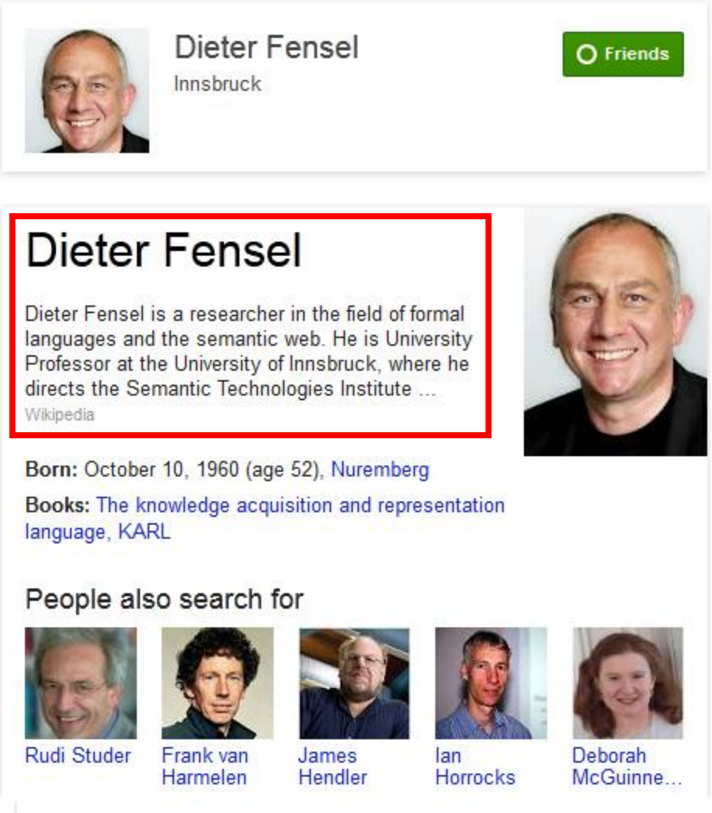
- GKG uses UIDs directly from the Freebase; detective work of Andreas Thalhammer showing how to get from GKG UIDs to Freebased UIDs using base64 and gzip

- Check the “Knowledge Graph links to Freebase” thread on w3c mailing list http://lists.w3.org/Archives/Public/semantic-web/2012Jun/0028.html

Wikipedia

- For most search results first sentences come from Wikipedia





Dieter Fensel

Innsbruck

○ Friends

Dieter Fensel

Dieter Fensel is a researcher in the field of formal languages and the semantic web. He is University Professor at the University of Innsbruck, where he directs the Semantic Technologies Institute ...

Wikipedia

Born: October 10, 1960 (age 52), Nuremberg

Books: The knowledge acquisition and representation language, KARL

People also search for

Rudi Studer

Frank van Harmelen

James Hendler

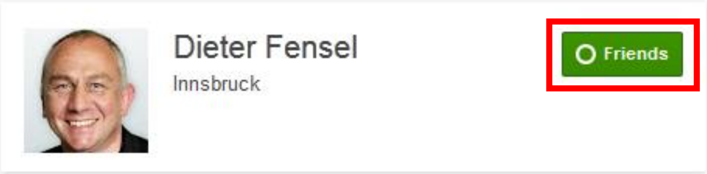
Ian Horrocks

Deborah McGuinne...

Other Sources

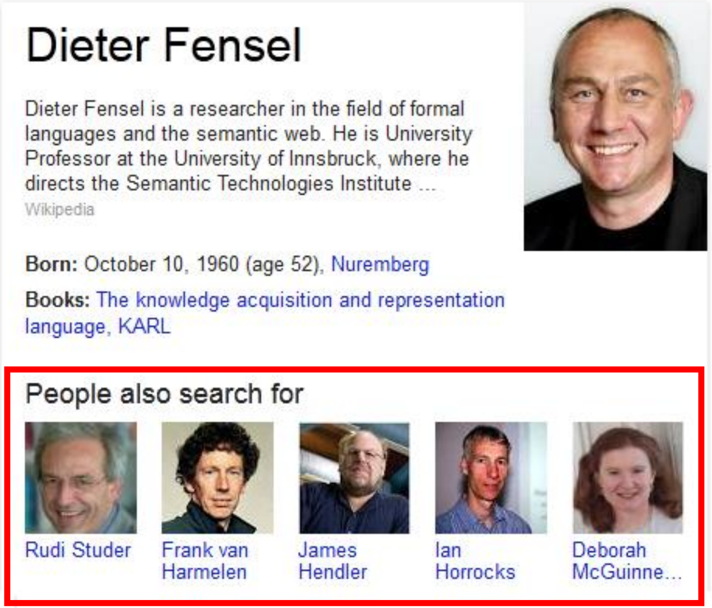
- GKG also considers the information Google retrieves from the volume of queries done by the users and the links those users have clicked on the results presented for those queries.

- GKG is also integrated with other Google products e.g. Google+



Dieter Fensel

Innsbruck



Dieter Fensel

Dieter Fensel is a researcher in the field of formal languages and the semantic web. He is University Professor at the University of Innsbruck, where he directs the Semantic Technologies Institute ...

Wikipedia

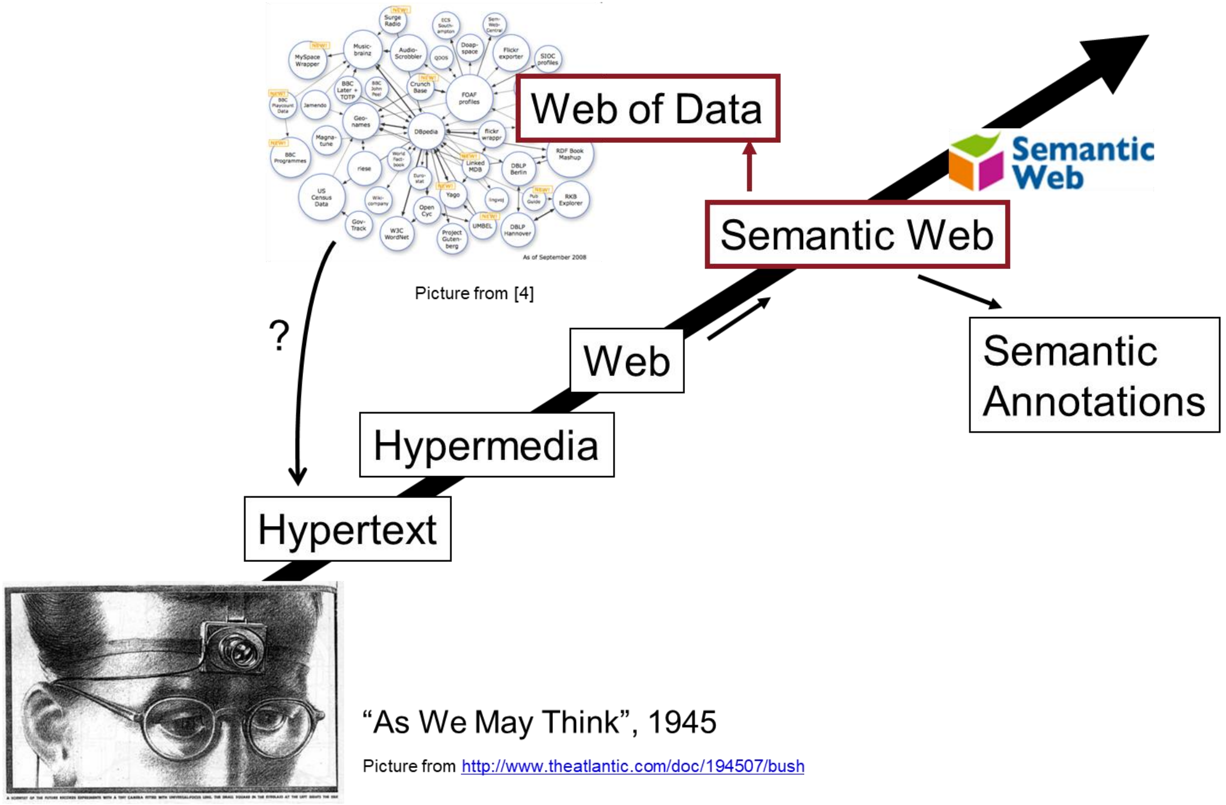
Born: October 10, 1960 (age 52), Nuremberg

Books: The knowledge acquisition and representation language, KARL

People also search for

Rudi Studer, Frank van Harmelen, James Hendler, Ian Horrocks, Deborah McGuinne...

Web of Data



Picture from [4]

“As We May Think”, 1945

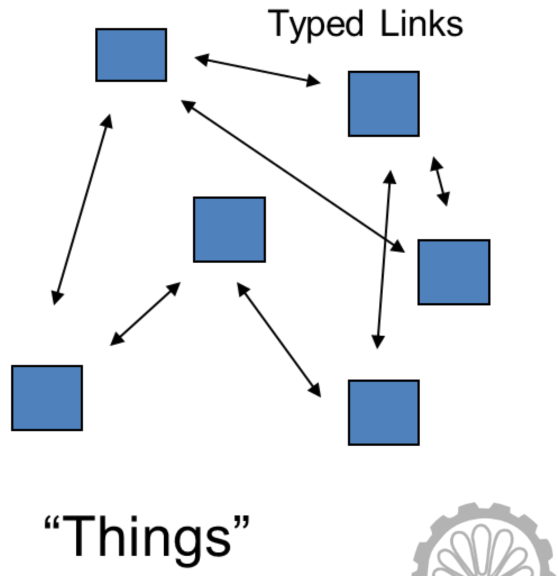
Picture from http://www.theatlantic.com/doc/194507/bush

Web of Data

- Characteristics:
  - Links between arbitrary things (e.g., persons, locations, events, buildings)
  - Structure of data on Web pages is made explicit
  - Things described on Web pages are named and get URIs
  - Links between things are made explicit and are typed

- HTML defines how text should look when presented to humans. Semantic web markup defines how information should be organized to be interpretable by machines.

- Web of Data

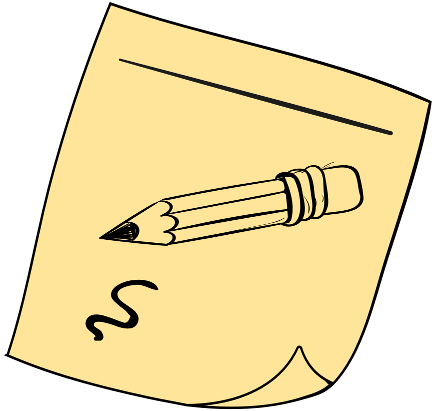


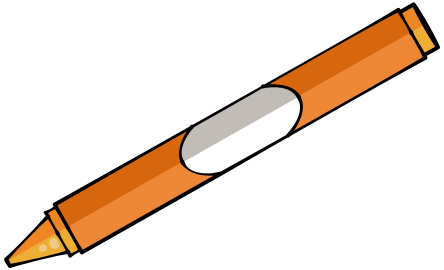
Web of Data

- A **closed** implementation of ***Web of Data*** principles
    - is not about documents, but objects such as people, places and things
    - objects are interlinked in the GKG
    - objects have structured information which is obtained from the web

- The Google Knowledge Graph is the basis for transforming Google' core search product from an *information engine* to a *knowledge engine (entity search engine)*

Building a KG Through ML



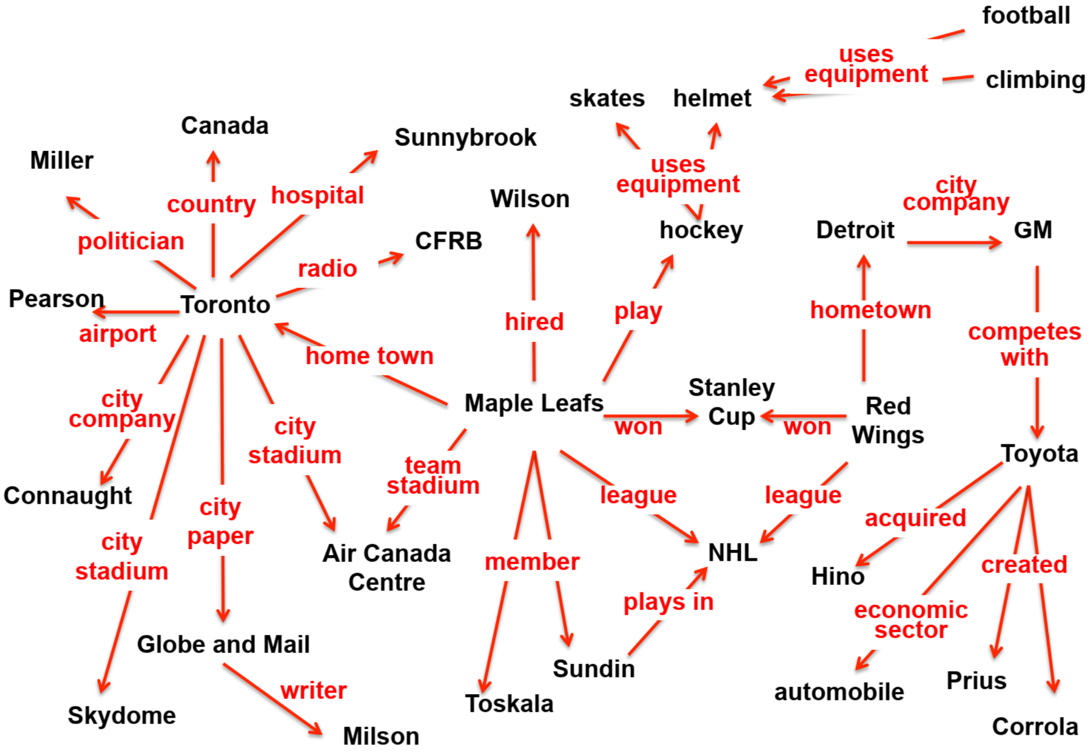


NELL: Never-Ending Language Learning

- Tom Mitchell et al. (CMU), 2010 to present.
- Learning to “read the web” 24 hours/day.
- Training data includes a collection of 1.2 billion web pages.
- Access to additional data through search engine APIs (100K calls/day).
- KB has 2.8 million instances over 1186 different categories.
- KB is freely available for download.
- You can help train NELL via Twitter.

Source: Never-Ending Learning (2018)

NELL Knowledge Fragment



Motivation for NELL

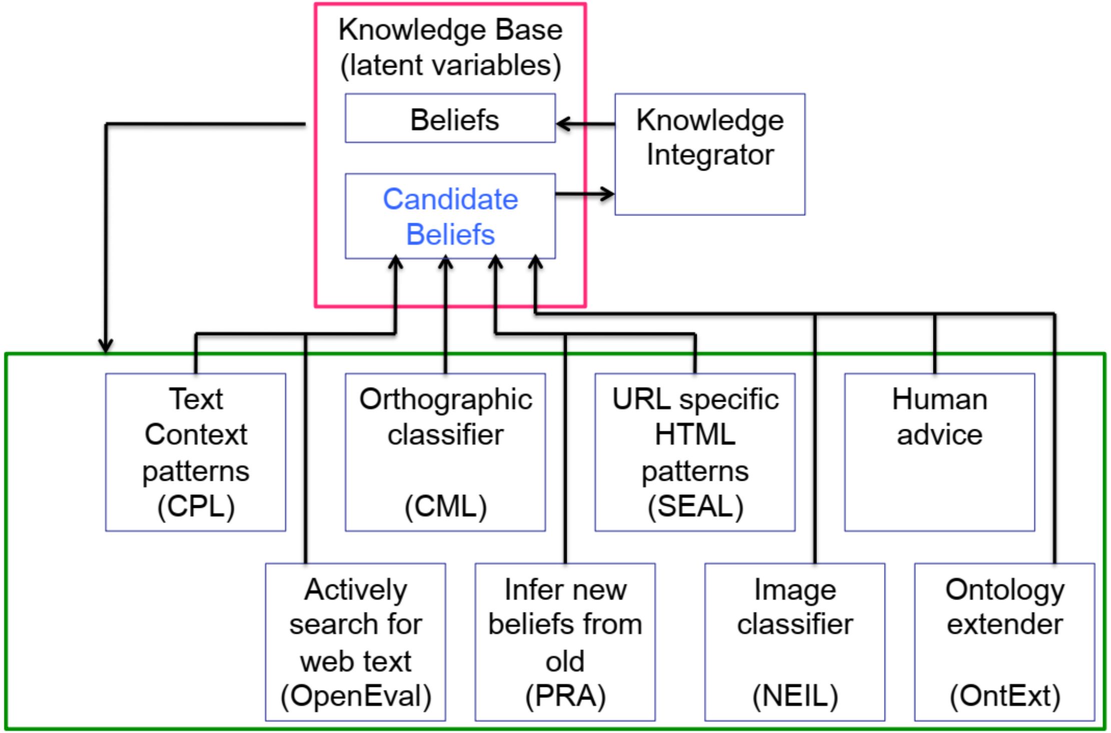
Thesis: “we will never truly understand human or machine learning until we can build computer programs that, like people,

- Learn many different types of knowledge or functions
- From years of diverse, mostly self-supervised experience
- In a staged curricular fashion, where previously learned knowledge enables learning further types of knowledge
- Where self-reflection and the ability to formulate new representations and new learning tasks enable the learning to avoid stagnation and performance plateaus.”

NELL Basic Idea

- NELL learns several things:
  - Categories
  - Triples: noun phrase 1 - relation - noun phrase 2
  - New relations
- Multiple inference algorithms propose triples and gather evidence for them.
  - Linguistic information
  - Word co-occurrence
  - Image labeling
  - Etc.
- Categories and triples supported by multiple sources of evidence grow in confidence.

NELL Architecture



Knowledge Base (latent variables)

Beliefs

Candidate Beliefs

Knowledge Integrator

Text Context patterns (CPL)

Orthographic classifier (CML)

URL specific HTML patterns (SEAL)

Human advice

Actively search for web text (OpenEval)

Infer new beliefs from old (PRA)

Image classifier (NEIL)

Ontology extender (OntExt)

Never-Ending Learning

- Set of learning tasks $L = \{L_i\}$
- Task $L_i = < T_i, P_i, E_i >$
  - ○ T_i is a task $< X_i, Y_i >$ specifying the domain of a function $f_i^* : X_i \rightarrow \ Y_i$
  - ○ Pi is a performance metric P_i: $f \ \rightarrow \mathbb{R}$
  - ○ E_i is an experience
- Coupling constraints $C \ = \ \{< \phi_k, V_k >\}$
  - ○ ϕ_k specifies degree of satisfaction of the coupling constraint among tasks
  - ○ V_k is a vector of indices over learning tasks specifying the arguments to ϕ_k
- $f_i^* = \min\limits_{f \in F_i} P_i(f)$

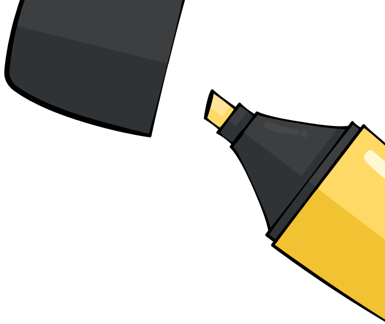
✓Goal: improve the quality of the task functions f_i as measured by the P_i.

✓NELL faces over 4100 distinct learning tasks.

Category Classification Tasks

1. Character string features of the noun phrase: Coupled Morphological Classifier system (CMC)
2. Distribution of text contexts found around this noun phrase in the 1.2 billion page database: Coupled Pattern Learner system (CPL)
3. Distribution of text contexts found through active web search (OpenEval).
4. HTML structure of web pages that mention the noun phrase: Set Expander for Any Language system (SEAL)
5. Visual images associated with the noun phrase: Never Ending Image Learner (NEIL)
6. Learned vector embeddings (feature vectors) of the noun phrase: LE (Learned Embeddings)

Relation Classification



Does “Pittsburgh” + “US” satisfy the relation CityLocatedInCountry(x,y) ?

There are 461 relations in the ontology.

Four methods are used for relation classification:

1. Distribution of text contexts from CPL
2. Distribution of text context from OpenEval
3. HTML structure from SEAL
4. Learned vector embeddings from LE

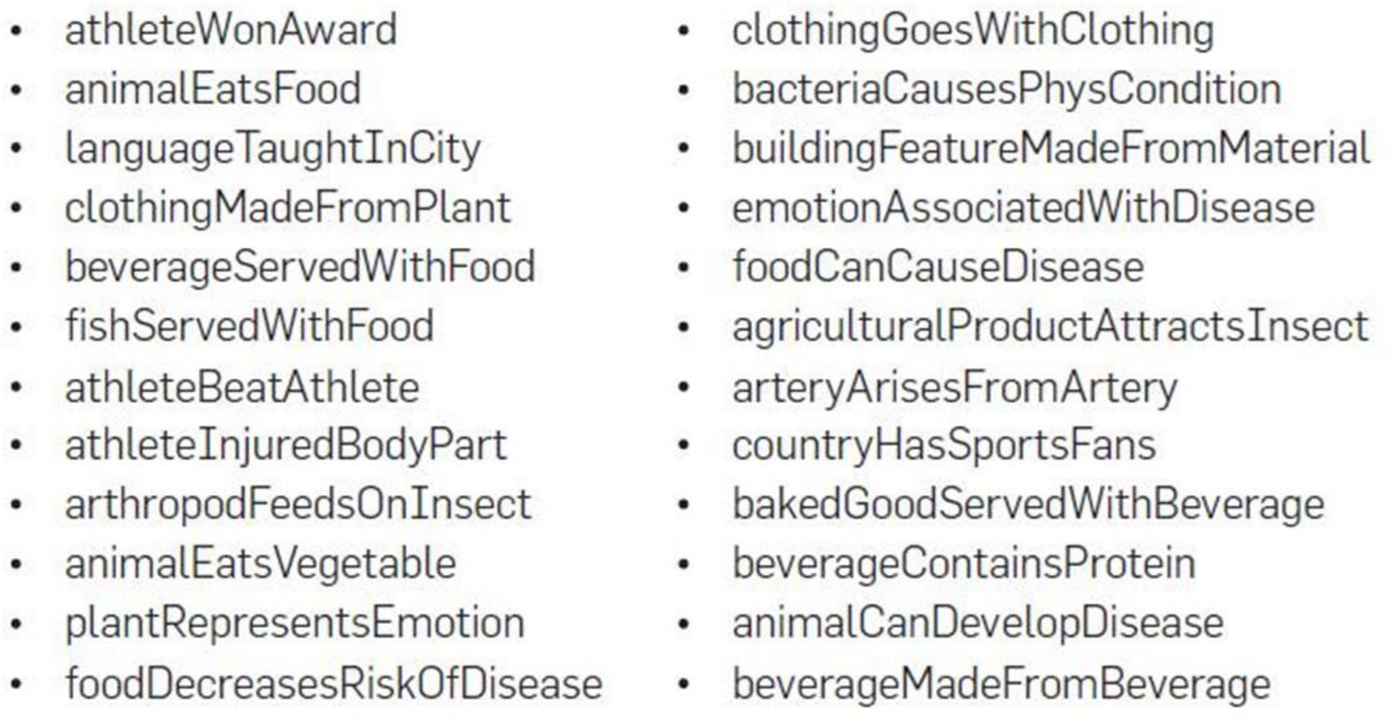
Entity Resolution

- Functions to classify whether pairs of noun phrases are synonyms.
- Noun phrases are kept distinct from the entities to which they refer.
- Necessary to deal with polysemy.
  - “Coach” can be either a person or a vehicle.
- Two methods are used:
  - String similarity
  - Similarities in beliefs about the entities
- NELL learns for each category what are the good types of knowledge to take as evidence for synonymy.

Inference Rules Among Belief Triples

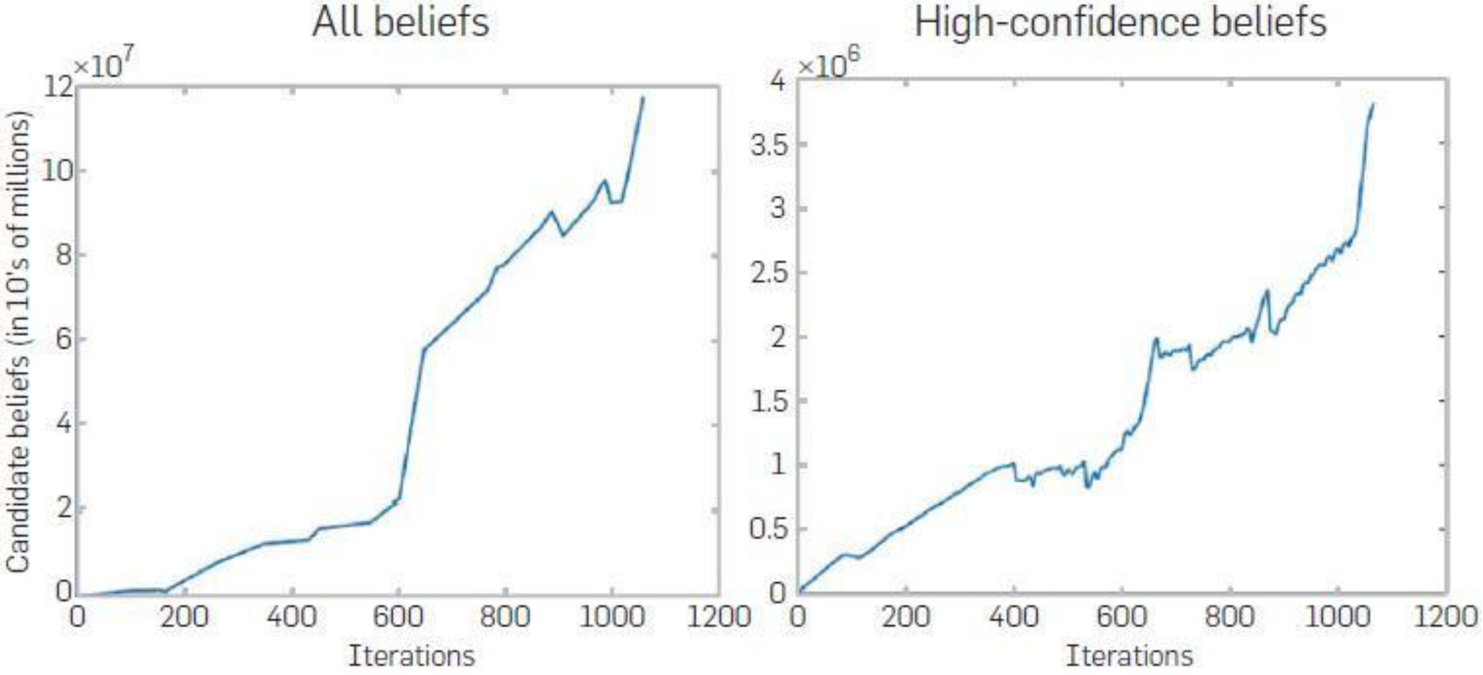
- Functions that propose new beliefs to be added to the KB.
- For each relation, the corresponding function is represented by a collection of restricted Horn Clause rules learned by the Path Ranking Algorithm (PRA)

Samples of Self-discovered Relations

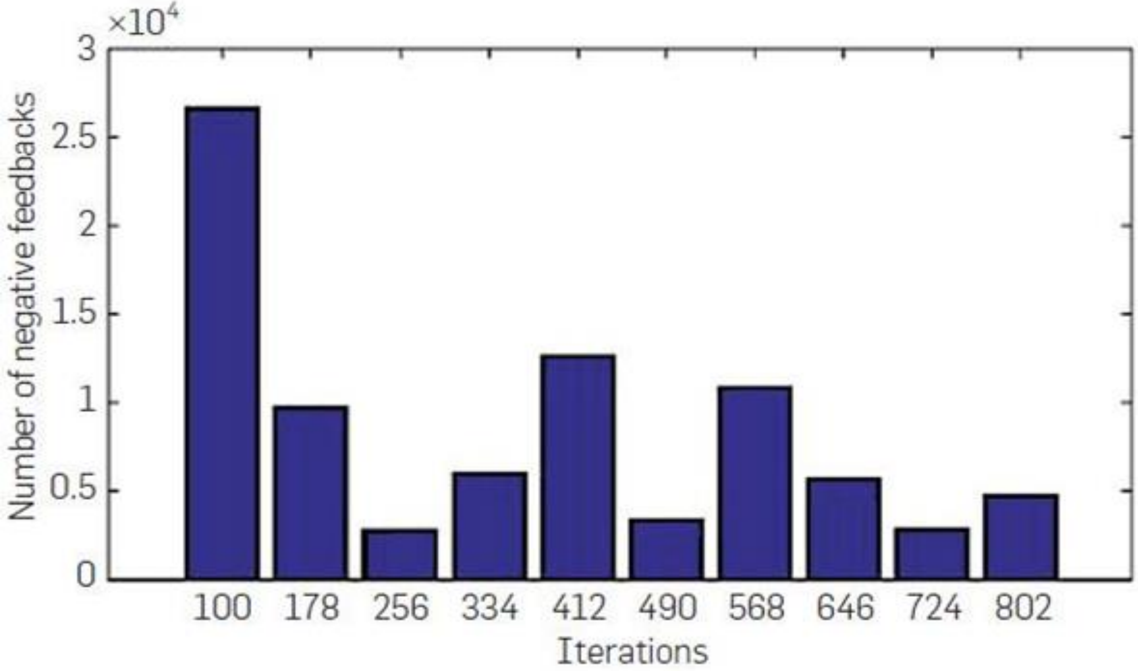


- athleteWonAward
- animalEatsFood
- languageTaughtInCity
- clothingMadeFromPlant
- beverageServedWithFood
- fishServedWithFood
- athleteBeatAthlete
- athleteInjuredBodyPart
- arthropodFeedsOnInsect
- animalEatsVegetable
- plantRepresentsEmotion
- foodDecreasesRiskOfDisease
- clothingGoesWithClothing
- bacteriaCausesPhysCondition
- buildingFeatureMadeFromMaterial
- emotionAssociatedWithDisease
- foodCanCauseDisease
- agriculturalProductAttractsInsect
- arteryArisesFromArtery
- countryHasSportsFans
- bakedGoodServedWithBeverage
- beverageContainsProtein
- animalCanDevelopDisease
- beverageMadeFromBeverage

Growth of the KB Over Time



Human Correction of NELL



Average 2.4 negative feedback items per month per predicate.

Thanks for your attention

References and IP Notice

- Some slides from Dave Touretzky's slides on knowledge graph.
- The begging slides are mostly based on Ioan Toma’s slides.
- Some graphics from Slidesgo online template.