RELATION EXTRACTION

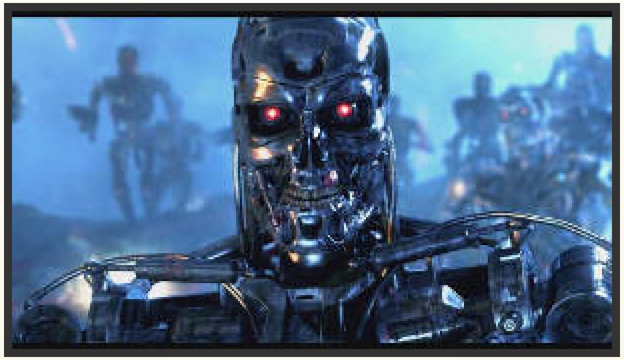
AUTOMATIC KNOWLEDGE BASE CONSTRUCTION

Denis Savenkov

Emory NLP Group Meeting
February 9, 2015

"SCIENTIA POTENTIA EST"

- Solving difficult problems requires a lot of knowledge
- Human spend a significant part of their life learning
- Computers need to have this knowledge as well!

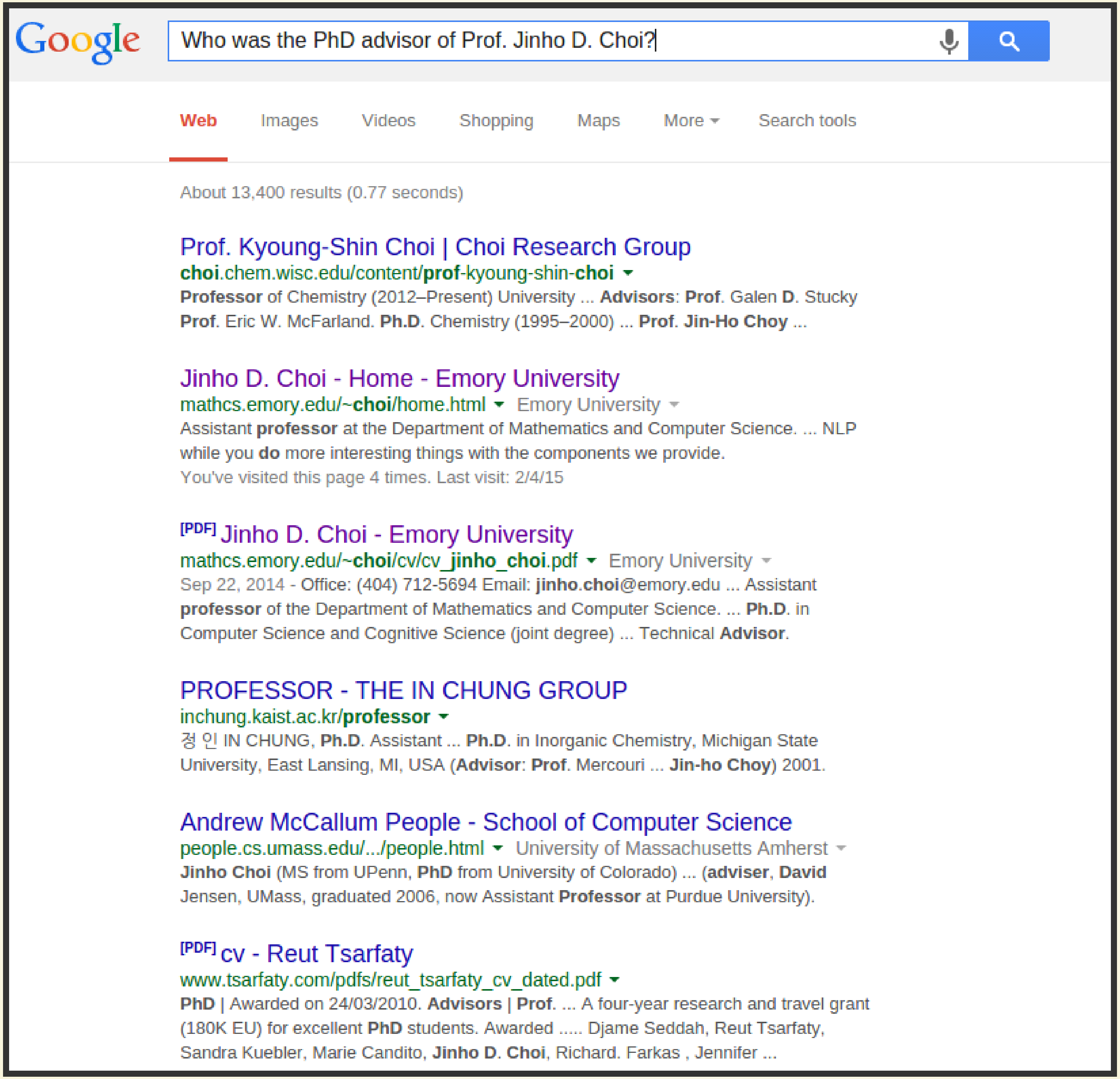


STRUCTURED INFORMATION

- Most of the information in the world is unstructured:
  - `text`
  - `images`
  - `video`
  - `...`
- Unstructured information is hard to work with
- We need to add some structure...

EXAMPLE: QUESTION ANSWERING

Who was the PhD advisor of Prof. Jinho Choi?


Google
Who was the PhD advisor of Prof. Jinho D. Choi?
Web
Images
Videos
Shopping
Maps
More
Search tools
About 13,400 results (0.77 seconds)
Prof. Kyoung-Shin Choi | Choi Research Group
choi.chem.wisc.edu/content/prof-kyoung-shin-choi
Professor of Chemistry (2012–Present) University ... Advisors: Prof. Galen D. Stucky Prof. Eric W. McFarland. Ph.D. Chemistry (1995–2000) ... Prof. Jin-Ho Choy ...
Jinho D. Choi - Home - Emory University
mathcs.emory.edu/~choi/home.html
Emory University
Assistant professor at the Department of Mathematics and Computer Science. ... NLP while you do more interesting things with the components we provide.
You've visited this page 4 times. Last visit: 2/4/15
[PDF] Jinho D. Choi - Emory University
mathcs.emory.edu/~choi/cv/cv_jinho_choi.pdf
Emory University
Sep 22, 2014 - Office: (404) 712-5694 Email: jinho.choi@emory.edu ... Assistant professor of the Department of Mathematics and Computer Science. ... Ph.D. in Computer Science and Cognitive Science (joint degree) ... Technical Advisor.
PROFESSOR - THE IN CHUNG GROUP
inchung.kaist.ac.kr/professor
정 인 IN CHUNG, Ph.D. Assistant ... Ph.D. in Inorganic Chemistry, Michigan State University, East Lansing, MI, USA (Advisor: Prof. Mercouri ... Jin-ho Choy) 2001.
Andrew McCallum People - School of Computer Science
people.cs.umass.edu/.../people.html
University of Massachusetts Amherst
Jinho Choi (MS from UPenn, PhD from University of Colorado) ... (adviser, David Jensen, UMass, graduated 2006, now Assistant Professor at Purdue University).
[PDF] cv - Reut Tsarfaty
www.tsarfaty.com/pdfs/reut_tsarfaty_cv_dated.pdf
PhD | Awarded on 24/03/2010. Advisors | Prof. ... A four-year research and travel grant (180K EU) for excellent PhD students. Awarded ..... Djame Seddah, Reut Tsarfaty, Sandra Kuebler, Marie Candito, Jinho D. Choi, Richard. Farkas , Jennifer ...

EMORY

Jinho D. Choi

DEPARTMENT OF MATHEMATICS & COMPUTER SCIENCE
EMORY UNIVERSITY

Biography

Jinho Choi is an assistant professor in the Department of Mathematics and Computer Science as well as the Institute of Quantitative Theory and Methods at Emory University. He obtained a B.A. in Computer Science and Mathematics (dual degree) from Coe College in 2002, a M.S.E. in Computer and Information Science from the University of Pennsylvania in 2003 with Mitchell Marcus, a Ph.D. in Computer Science and Cognitive Science (joint degree) from the University of Colorado Boulder in 2012 with Martha Palmer, and did his postdoctoral work at the University of Massachusetts Amherst in 2014 with Andrew McCallum. He was a full-time lecturer in the Department of Computer Science at the Korea Military Academy from 2004 to 2007 while he was serving his military duty in South Korea. He was a R&D team lead of the Amelia project, the next generation machine reading system developed at IPsoft Inc.

Jinho Choi has been active in research on natural language processing; especially, on the optimization of low-level NLP (e.g., part-of-speech tagging, dependency parsing, semantic role labeling) for robustness on various data and scalability on large data. He has developed an open source project called ClearNLP, providing NLP components with state-of-the-art accuracy and speed, which has been widely used for both academic and industrial research. His current research focuses on the development of NLP components for different domains (e.g., social media, medical data) and the applications of these NLP components for end-user systems such as question-answering, information extraction, dialog management, etc. He is also interested in interdisciplinary research where NLP can enhance researches in other areas.

Documents

- Curriculum vitae (mostly up-to-date).
- Research statement (written in 2013).
- Teaching statement (written in 2013).

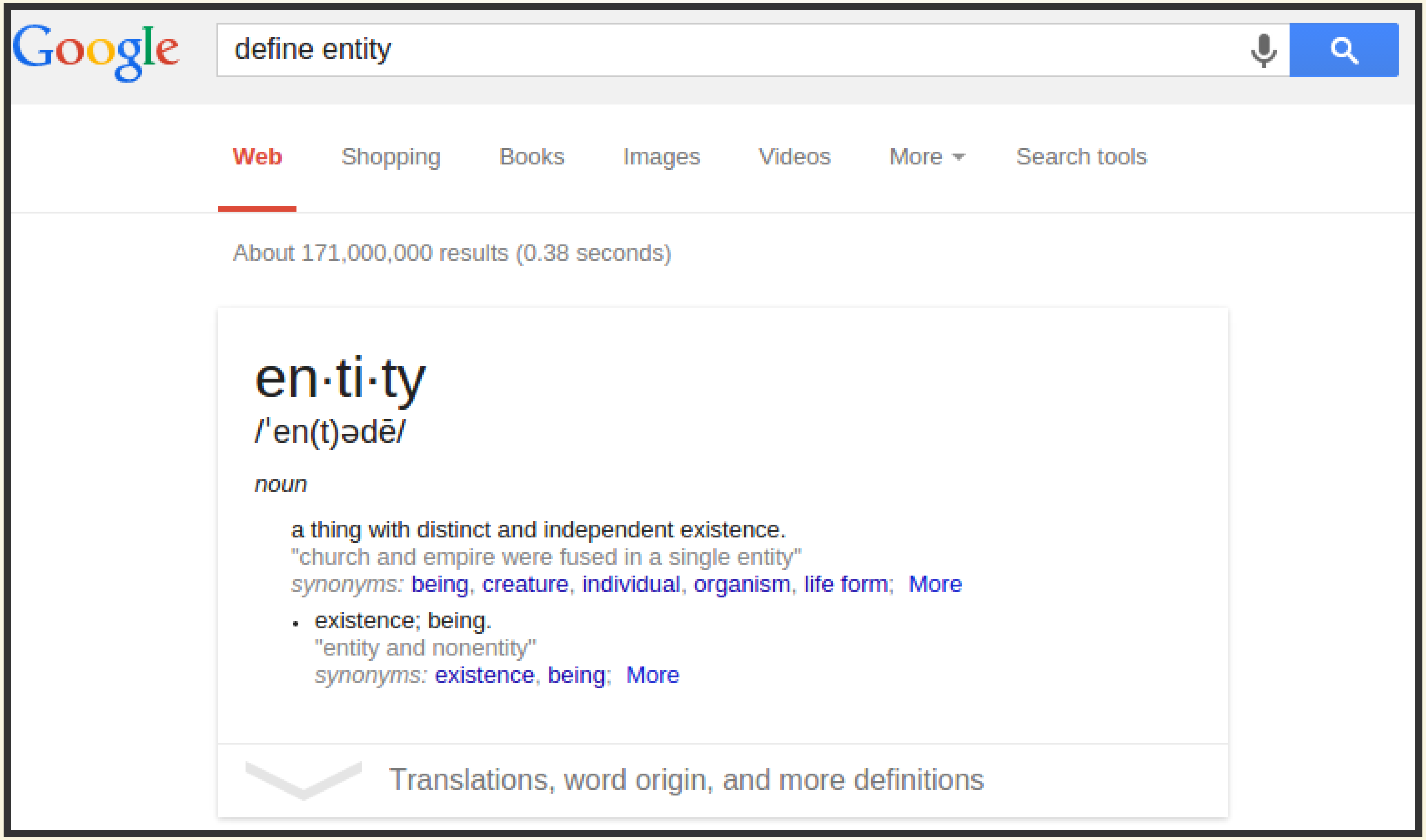
WHAT IF WE HAD A DATABASE WITH FACTS?

We could just query the database and get the answer:

Who was the PhD advisor of Prof. Jinho Choi?

```
SELECT advisor
FROM phd_advisors
WHERE student = "Jinho D. Choi"
```

FROM WORDS TO ENTITIES



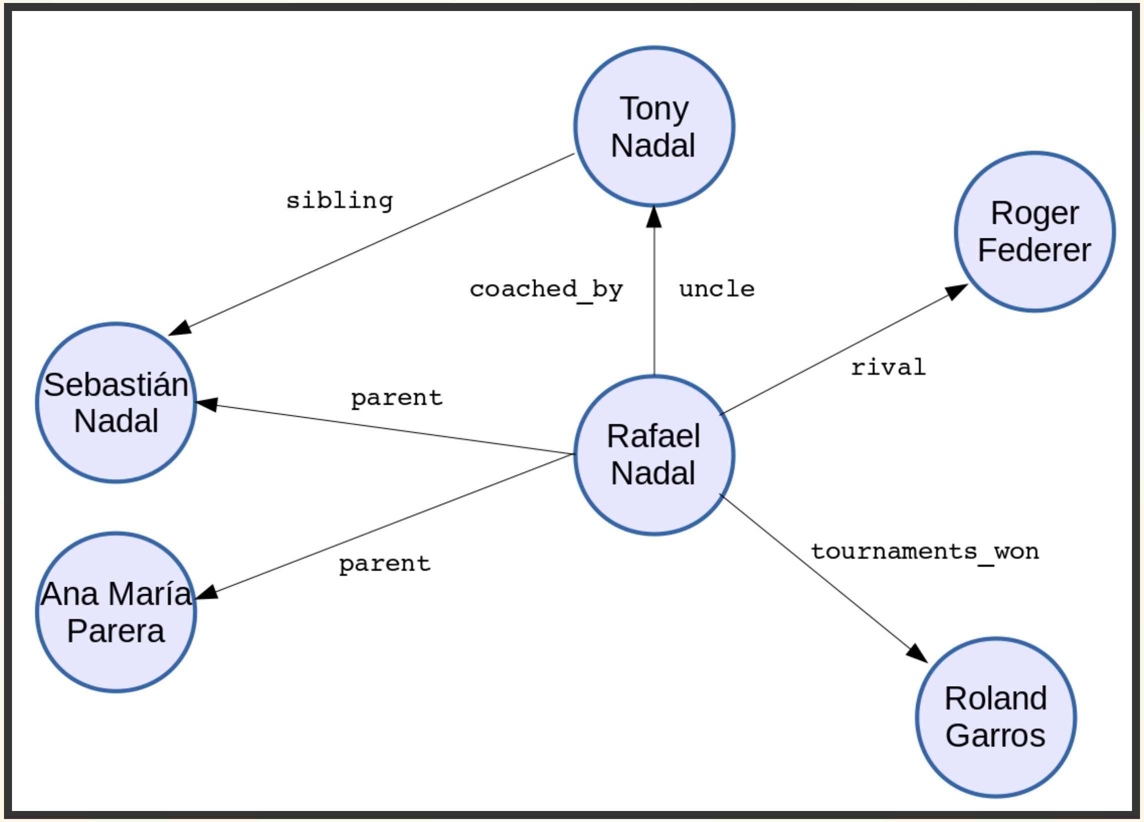
ENTITIES

- We may use different words to refer to the same thing
  - `Rafa, The King of Clay, Rafael Nadal`
- Attributes of entities
  - `types: e.g. tennis player`
  - `characteristics: e.g. height, weight, birth date`
- Some entities are related
  - `[Tony Nadal] <coach> [Rafael Nadal]`

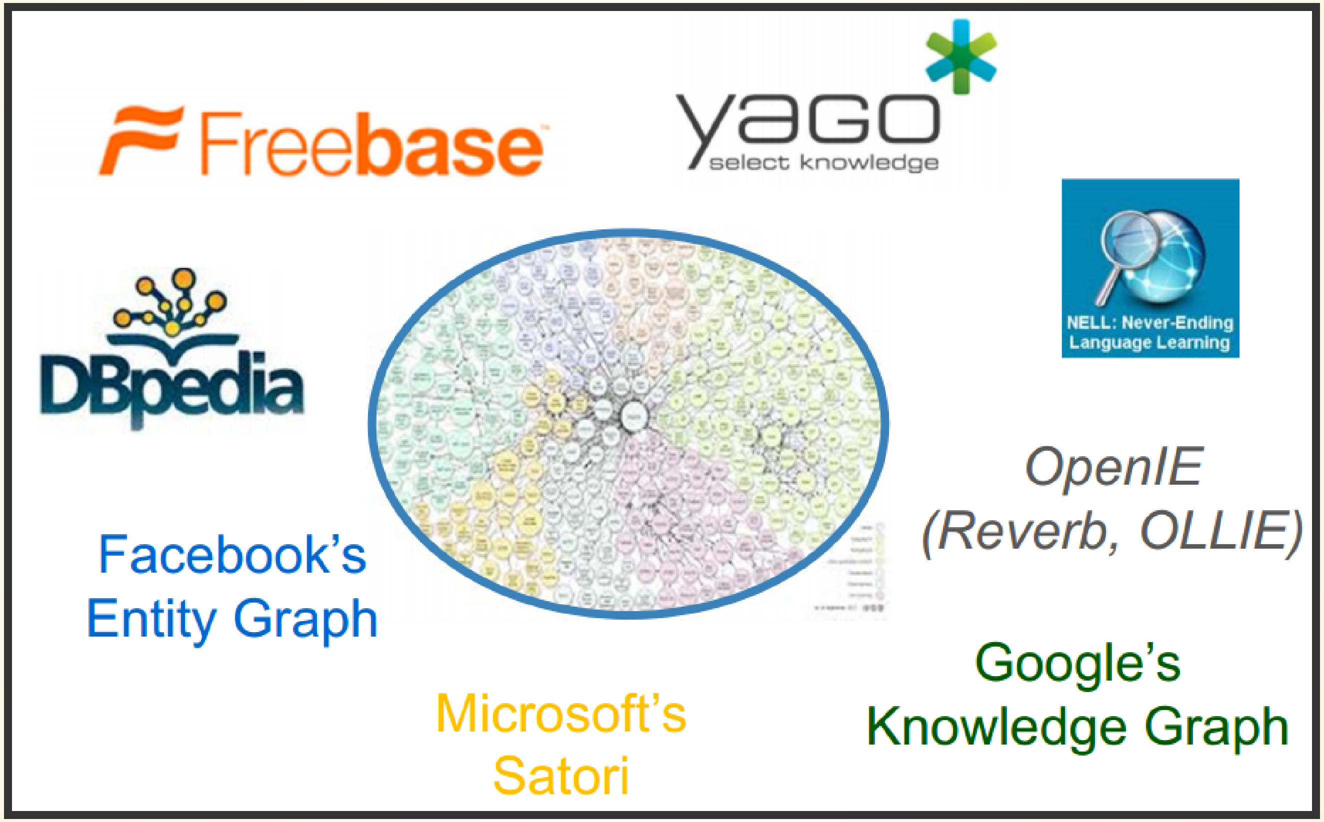
PREDICATES

- Entities are related in different ways:
  - [Tony Nadal] <coach-of> [Rafael Nadal]
  - [Tony Nadal] <uncle-of> [Rafael Nadal]
  - [Rafael Nadal] <parents> [Sebastián Nadal]
  - [Rafael Nadal] <parents> [Ana María Parera]
- Entities with their relations constitute a **knowledge base**
- We can represent a knowledge base as a graph

KNOWLEDGE BASE



KNOWLEDGE GRAPHS



* image from KDD'14 "Constructing and Mining Web-scale Knowledge Graphs" workshop slides

SOME CHALLENGES

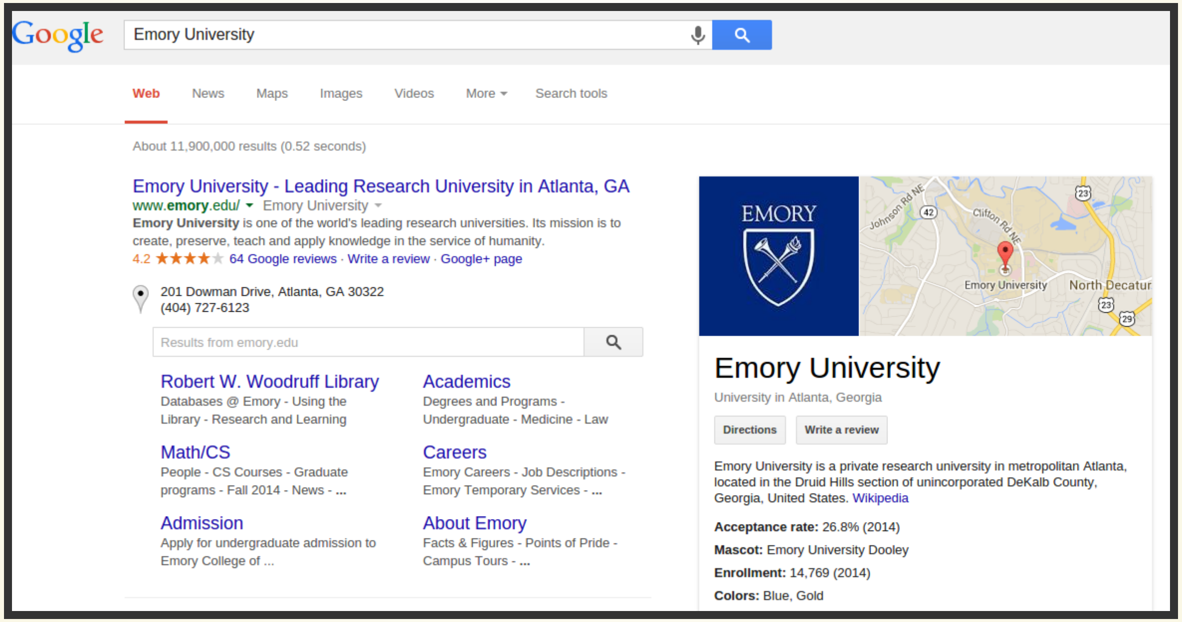
- **Validation**: knowledge graphs are not always correct
- **Interface**: how to make it easier to access the knowledge?
- **Intelligence**: how to create AI fom knowledge graphs?
- **Growth**: knowledge graphs are incomplete
    - link prediction
    - ontology matching
    - knowledge extraction (this presentation)

---

from KDD 2014 Tutorial on Constructing and Mining Web-scale Knowledge Graphs, New York, August 24, 2014

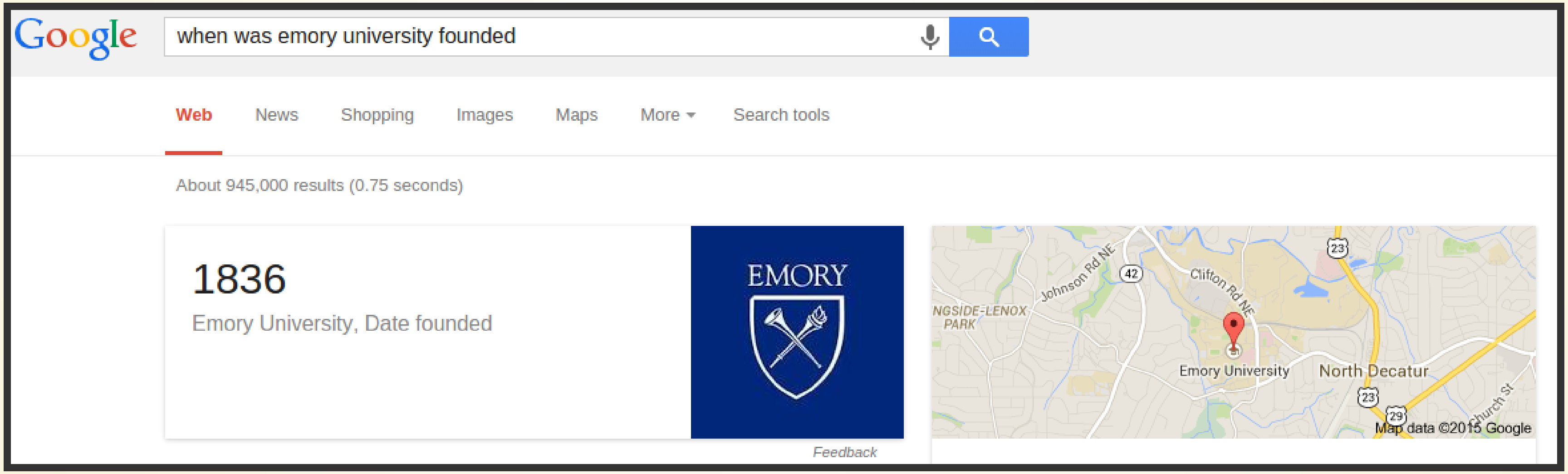
APPLICATIONS

Entity summarization



APPLICATIONS

Question Answering



Freebase

- 47M entities and 2.5B facts
- fully structured (entities and relations come from a fixed lexicon rather than free text)
- constructed by community members
- Built by MetaWeb and acquired by Google in 2010
- Data is publicly available
- Will be shut down in 2015 and data transitioned to WikiData
- Tuple: [**/m/0jcx, /m/04m8, /m/019xz9**] means Albert Einstein was born in Ulm

INCOMPLETENESS

*71% of people in Freebase have no information on place of birth and 75% have no known nationality **

- Long-tail distribution: we know a lot about popular entities, but there is a heavy tail of less known entities
- How to increase coverage?
  - Ask people: crowdsourcing
  - Merge with other knowledge bases: ontology matching
  - Extract from the available data

---

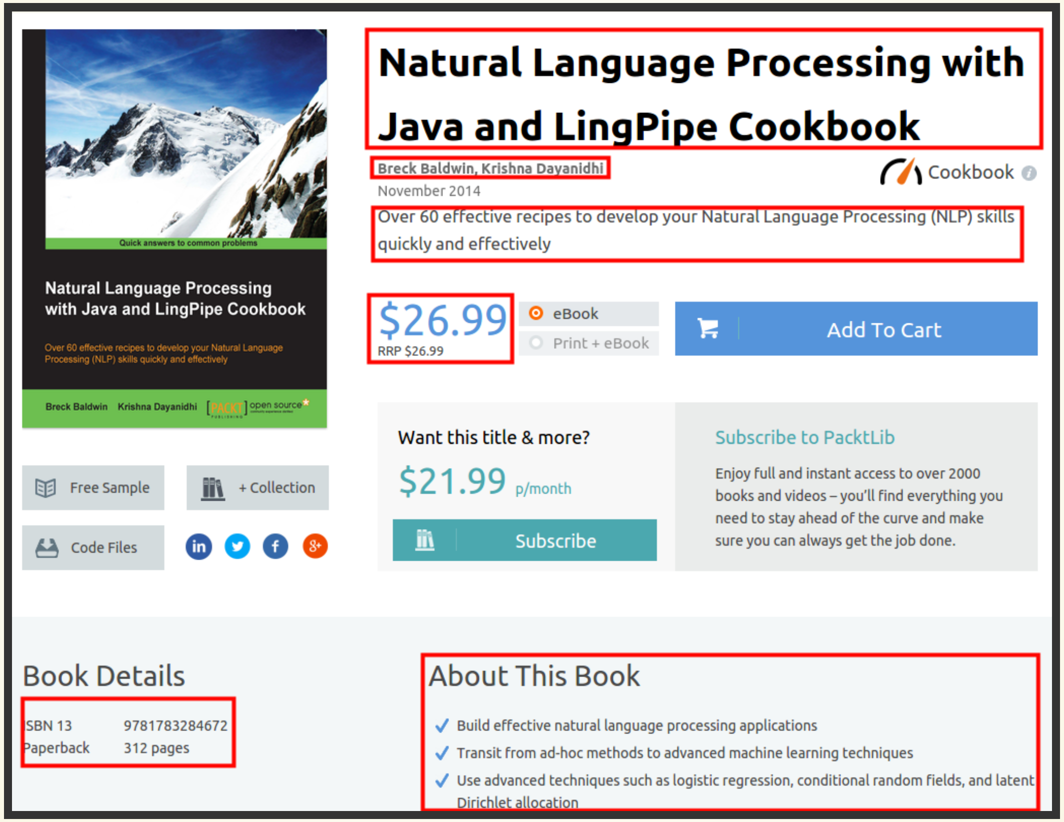
* from "Knowledge Vault : A Web-Scale Approach to Probabilistic Knowledge Fusion" by X.Dong et al. 2014

WEB OF DATA

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

see http://schema.org

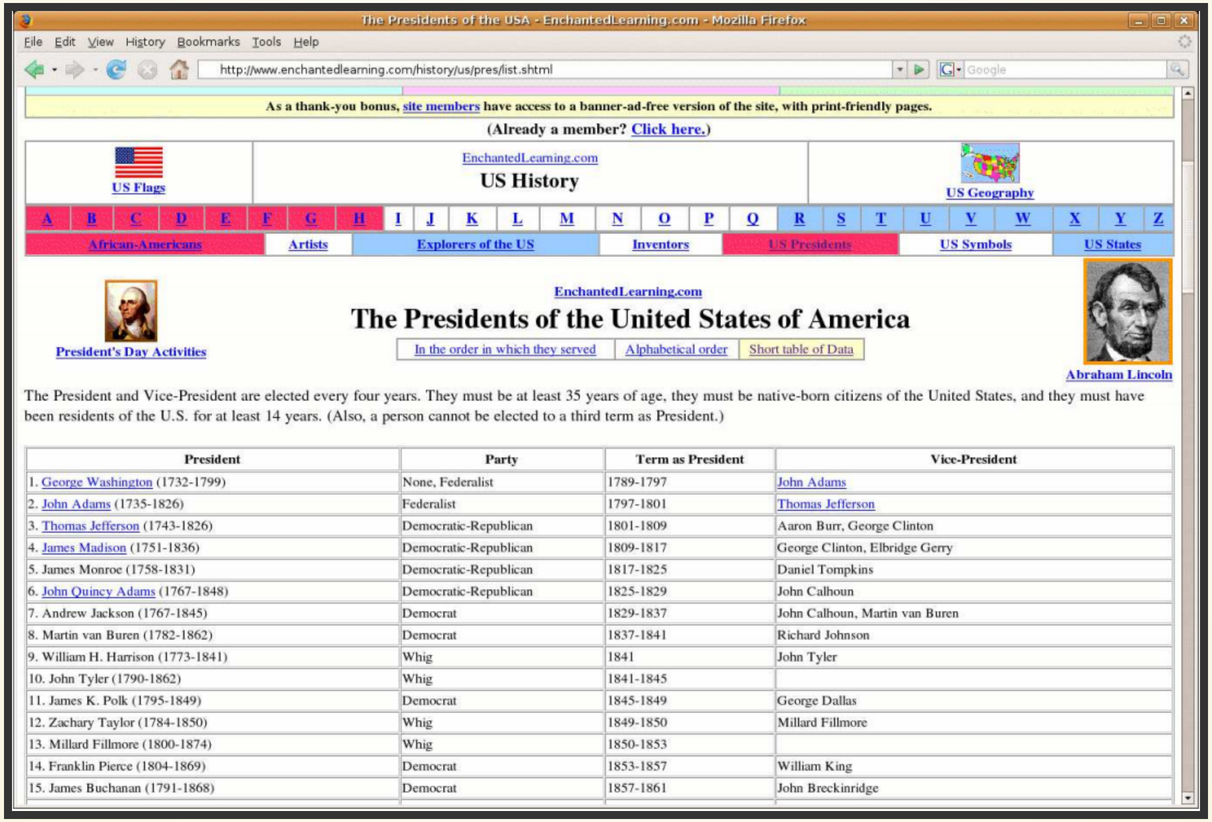
WRAPPER INDUCTION



"Wrapper Induction for Information Extraction" by N.Kushmerick et al. 1997

TABLES ON THE WEB

- Relational data on the web is often represented as tables and it is possible to extract this data (e.g. [1])


The Presidents of the USA - EnchantedLearning.com - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.enchantedlearning.com/history/us/pres/list.shtml

As a thank-you bonus, site members have access to a banner-ad-free version of the site, with print-friendly pages.

(Already a member? Click here.)

US Flags | EnchantedLearning.com US History | US Geography

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

African-Americans | Artists | Explorers of the US | Inventors | US Presidents | US Symbols | US States

President's Day Activities

EnchantedLearning.com

The Presidents of the United States of America

In the order in which they served | Alphabetical order | Short table of Data

Abraham Lincoln

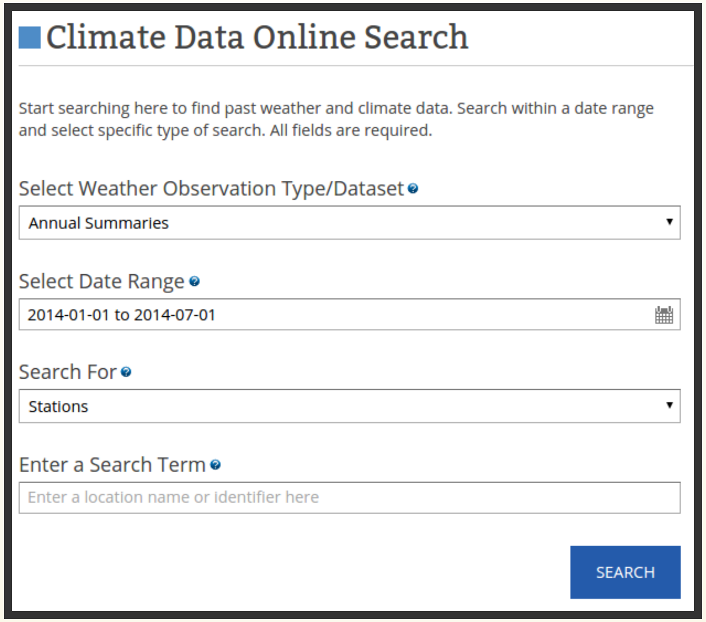
The President and Vice-President are elected every four years. They must be at least 35 years of age, they must be native-born citizens of the United States, and they must have been residents of the U.S. for at least 14 years. (Also, a person cannot be elected to a third term as President.)

President	Party	Term as President	Vice-President
1. George Washington (1732-1799)	None, Federalist	1789-1797	John Adams
2. John Adams (1735-1826)	Federalist	1797-1801	Thomas Jefferson
3. Thomas Jefferson (1743-1826)	Democratic-Republican	1801-1809	Aaron Burr, George Clinton
4. James Madison (1751-1836)	Democratic-Republican	1809-1817	George Clinton, Elbridge Gerry
5. James Monroe (1758-1831)	Democratic-Republican	1817-1825	Daniel Tompkins
6. John Quincy Adams (1767-1848)	Democratic-Republican	1825-1829	John Calhoun
7. Andrew Jackson (1767-1845)	Democrat	1829-1837	John Calhoun, Martin van Buren
8. Martin van Buren (1782-1862)	Democrat	1837-1841	Richard Johnson
9. William H. Harrison (1773-1841)	Whig	1841	John Tyler
10. John Tyler (1790-1862)	Whig	1841-1845	
11. James K. Polk (1795-1849)	Democrat	1845-1849	George Dallas
12. Zachary Taylor (1784-1850)	Whig	1849-1850	Millard Fillmore
13. Millard Fillmore (1800-1874)	Whig	1850-1853	
14. Franklin Pierce (1804-1869)	Democrat	1853-1857	William King
15. James Buchanan (1791-1868)	Democrat	1857-1861	John Breckinridge

[1] "WebTables: Exploring the Power of Tables on the Web", M.Cafarella et al. 2008

DEEPWEB

- Large volumes of data is accessible only through HTML form interfaces
- We can automatically make queries and extract the hidden knowledge e.g. [1]



[1] "Web-Scale Extraction of Structured Data" by M.Cafarella et al. 2008

RELATION EXTRACTION FROM TEXT

[Emory College] was founded in [1836] in [Oxford, Georgia] by the [Methodist Episcopal Church].

- **Focused extraction**: need to find a particular attribute of a particular entity (slot-filling)
- **Unfocused extraction**: process text and extract everything we can

TREC KBA (knowledge base acceleration) http://trec-kba.org/

FOCUSED EXTRACTION

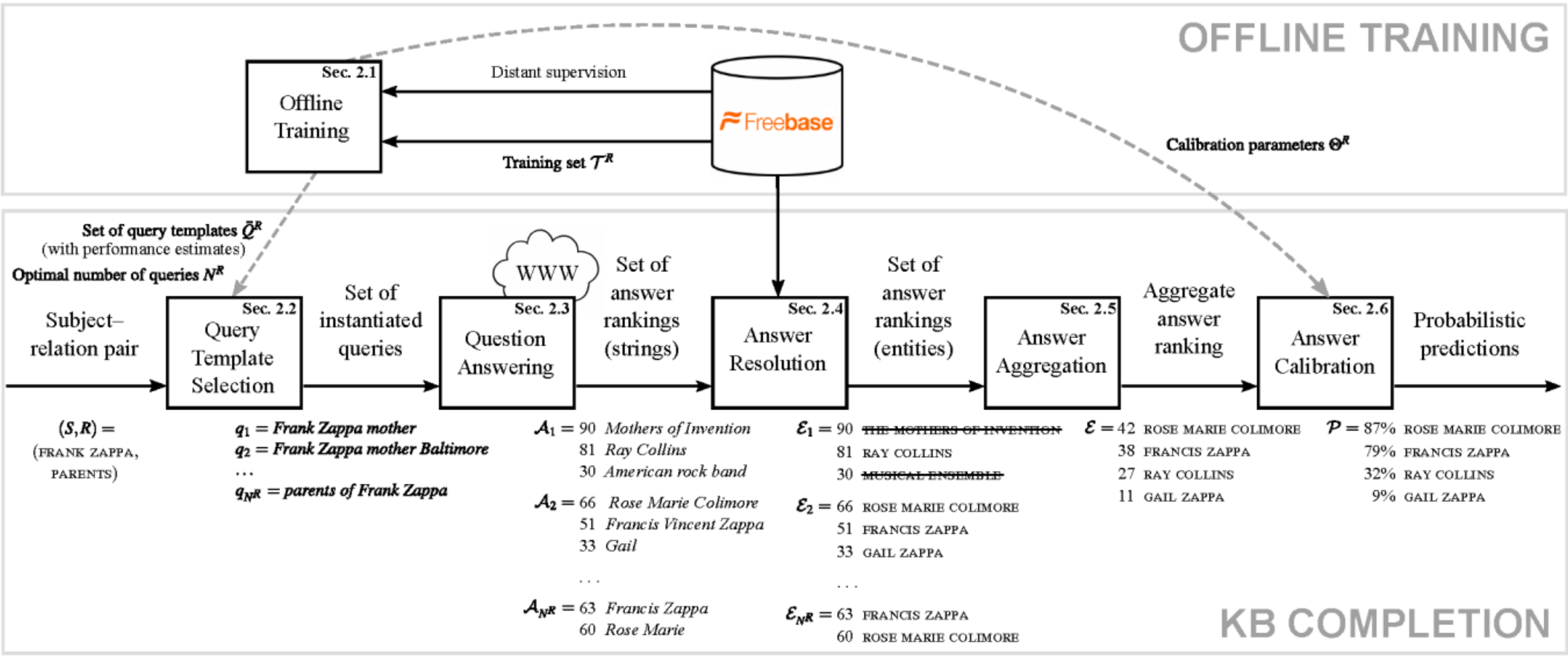


Figure 1: Overview of the pipeline for knowledge base completion, illustrated by the Frank Zappa example of Table 2. The set $\bar{Q}^R$ of query templates and the optimal number N^R of queries to use are relation-specific. The square boxes contain references to the sections in which the respective stages of the pipeline are described.

"Knowledge Base Completion via Search-Based Question Answering" by B.West et al 2014 (WWW)

RELATION EXTRACTION FROM NL

- Structured extractions (fixed entity/relations lexicon)
  1. Supervised relation extraction
  2. Semi-supervised relation extraction
  3. Distant supervision for relation extraction
- Open information extraction (entities and relations expressed in natural language)

RELATION EXTRACTION FROM NL

- Today, computers can't understant natural language text
- How do we teach them to extract knowledge then?
- M.Hearst* proposed to extract hyponyms using simple patterns (Hearst patterns)
  - `Bruises, wounds, broken bones or other` *`injuries`*`...`
  - `temples, treasuries,and other important` *`civic buildings`*
  - `All common-law` *`countries`*`, including Canada and England...`
  - `...`

---

* "Automatic Acquisition of Hyponyms from Large Text Corpora" by Marti Hearst, 1992

SUPERVISED RELATION EXTRACTION

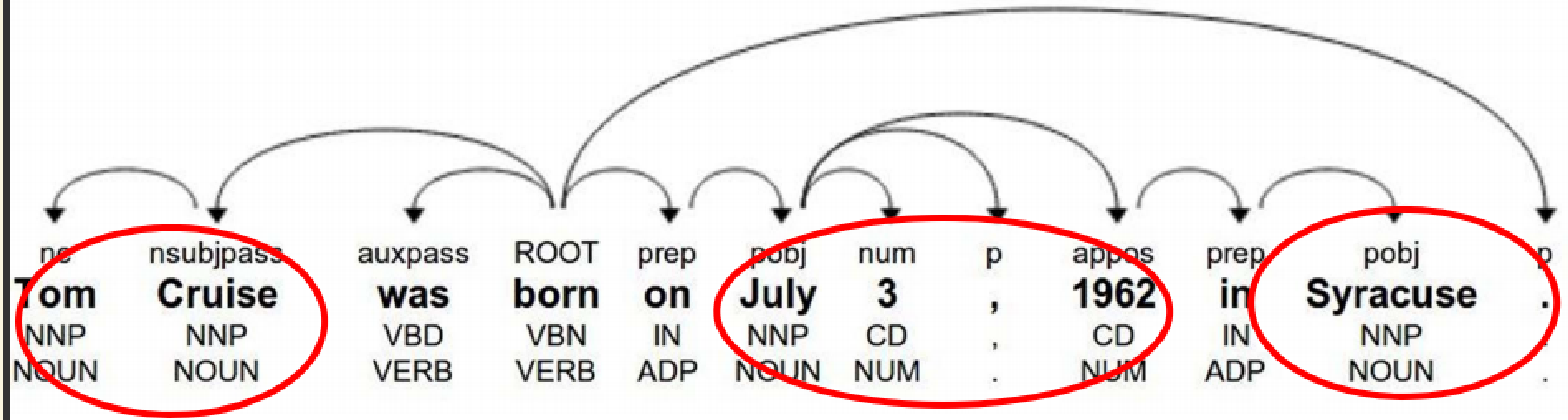
- Training dataset with sentence-level labels for each relation
  - Emory College was founded in 1836 (+)
  - Founded in 1836, Emory College ... (+)
  - Emory College opened in 1838 (-)
- Datasets: ACE 2004 (Automatic Content Extraction), MUC-7 (Message Understanding Conference), BioNLP challenges
- Solves relation extraction as **binary classification problem**
- Research studied various features* and training methods

---

* "Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations" by N.Kambhatla 2004

FEATURES

- words between entities
- types of entities (person, location, organizaton, etc)
- # of words between entities
- path between entities in a parse tree
- ...



X was born on DDDD in Y

- DEP: X <*nsubjpass* / born *prep*> on *pobj*> DATE *prep*> in *pobj*> Y
- NER: X = PER, Y = LOC
- POS: X = NOUN, NNP; Y = NOUN, NNP
- Context: born, on, in , "born on"

KERNEL-BASED METHODS

Alternatively, one can define a kernel (think similarity measure) between text fragments and apply kernel-based ML method (e.g. SVM or anything else)

- **Parse-tree kernels**: similarity of parse trees of 2 text fragments
- Each node can have multiple attributes (word, POS, NER type, etc), which are than used to compute the kernel

---

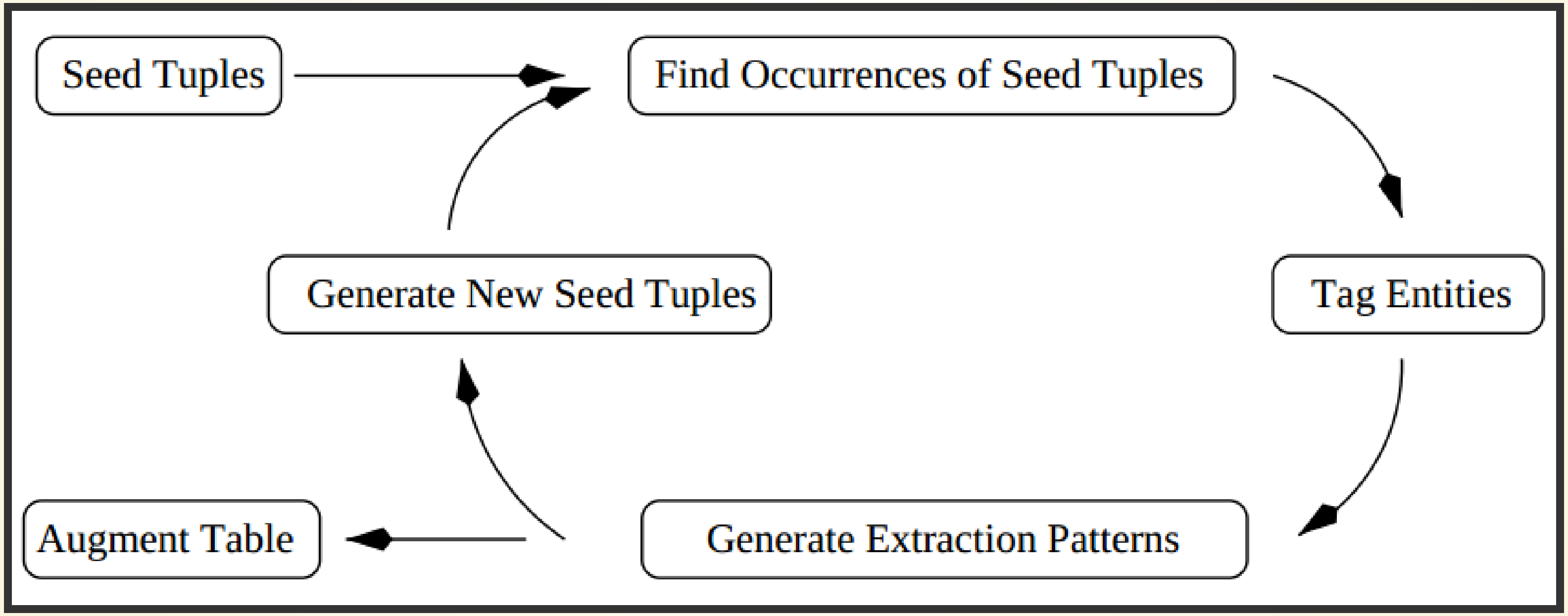
- "A shortest path dependency kernel for relation extraction" by R. Bunescu & R. Mooney, 2005
- "Tree Kernel-based Relation Extraction with Context-Sensitive Structured Parse Tree Information" by GuoDong Zhou et al, 2007

SUPERVISED RELATION EXTRACTION

- Cons:
  1. Expensive to obtain the data!
  2. Adding new relation requires labelling
- Pros:
  1. High quality training data
  2. Explicit negative examples

SEMI-SUPERVISED RELATION EXTRACTION

- Reduce the amount of supervision required
- Examples: DIPRE, Snowball, KnowItAll
- Based on bootstrapping (iteratively improving the system)



From "Snowball: Extracting Relations from Large Plain-Text Collections" by E.Agichtein & L.Gravano, 2000

SEMI-SUPERVISED RELATION EXTRACTION

- Pros:
  1. Less supervision required
  2. Can extract more knowledge triples thanks to bootstrapping
- Cons:
  1. Semantic drift: as we iterate the system extracts more and more incorrect patterns/triples
  2. Extending to new relations still requires seed data

DISTANT SUPERVISION

Utilize existing knowledge base to label data and train a model

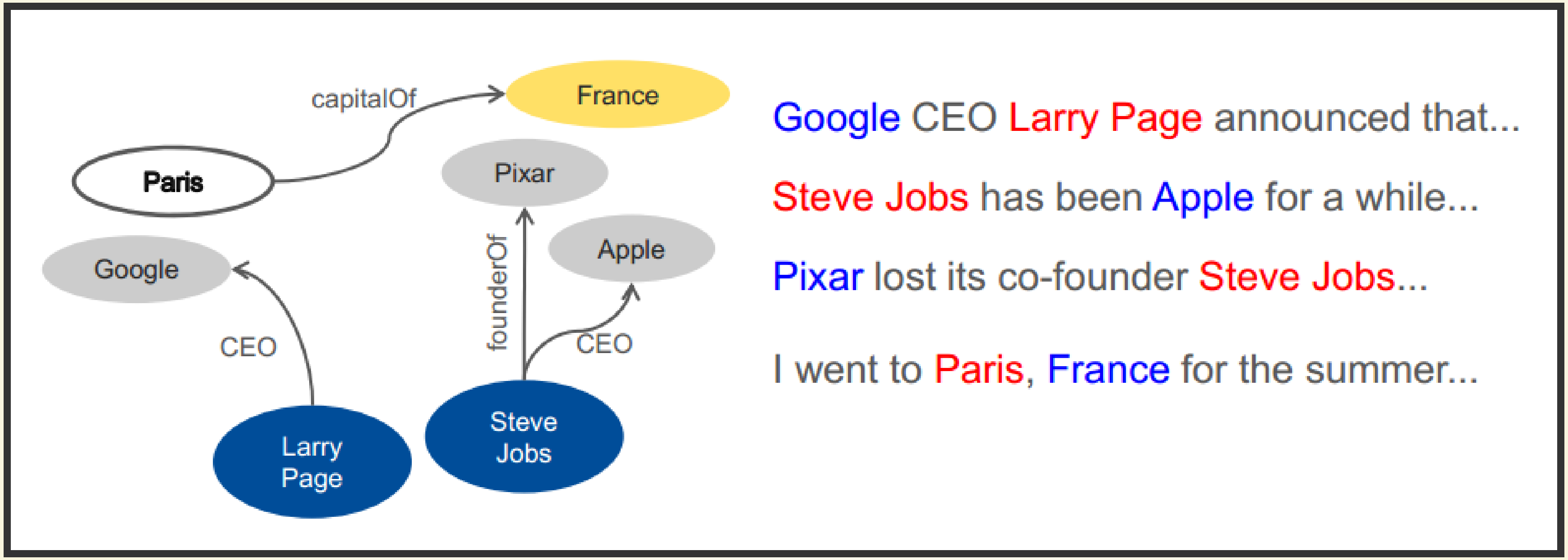


Image from KDD 2014 Tutorial on Constructing and Mining Web-scale Knowledge Graphs, New York, August 24, 2014

DISTANT SUPERVISION ASSUMPTIONS

Assume we have a knowledge triple

$$(e_1, p, e_2)$$

1. All sentences that mentions e_1 and e_2 together expresses the predicate p
2. At least one sentence that mentions e_1 and e_2 together expresses the predicate p (multi-instance setting)
3. A sentence that mentions e_1 and e_2 together might express the predicate p and a pair of entities can be related with different predicates (multi-instance multi-label setting)

1. "Distant supervision for relation extraction without labeled data" by M.Mintz et al 2009
2. "Modeling Relations and Their Mentions without Labeled Text" by S.Riedel et al 2010
3. "Multi-instance Multi-label Learning for Relation Extraction" by M.Surdeanu et al 2012

DISTANT SUPERVISION TRAINING

- Extract features for all sentences that mention a related pair of entities
- Randomly sample sentences with non-related entities as negative examples
- Train a multiclass classification model

Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[,]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]

DISTANT SUPERVISION

- Pros:
  1. Scalable!
  2. Can be applied in different languages
- Cons:
  1. Training data is noisy!
  2. No explicit negative examples

OPEN INFORMATION EXTRACTION

- Introduced in [1]
- Extracts natural language triples from text:
  - `Apple announced a new iPhone 6. => (Apple, announced, iPhone 6)`
- Extracts noun phrases as entities and verb phrases as predicates
- A trained classifier is used to predict whether an extraction is good

```
TextRunner [trained extractor] -> ReVerb [chunking] -> Ollie [dependency tree] -> OpenIE 4 (Srlie + RelNoun) [semantic roles]
```

---

[1]. "Open Information Extraction from the Web" by M.Banko et al. 2007

OPEN INFORMATION EXTRACTION

- Pros:
  1. Even more scalable! ($O(N)$ vs $O(N|R|)$)
  2. Do not require any training data
- Cons:
  1. Lack of structure: need to cluster predicates

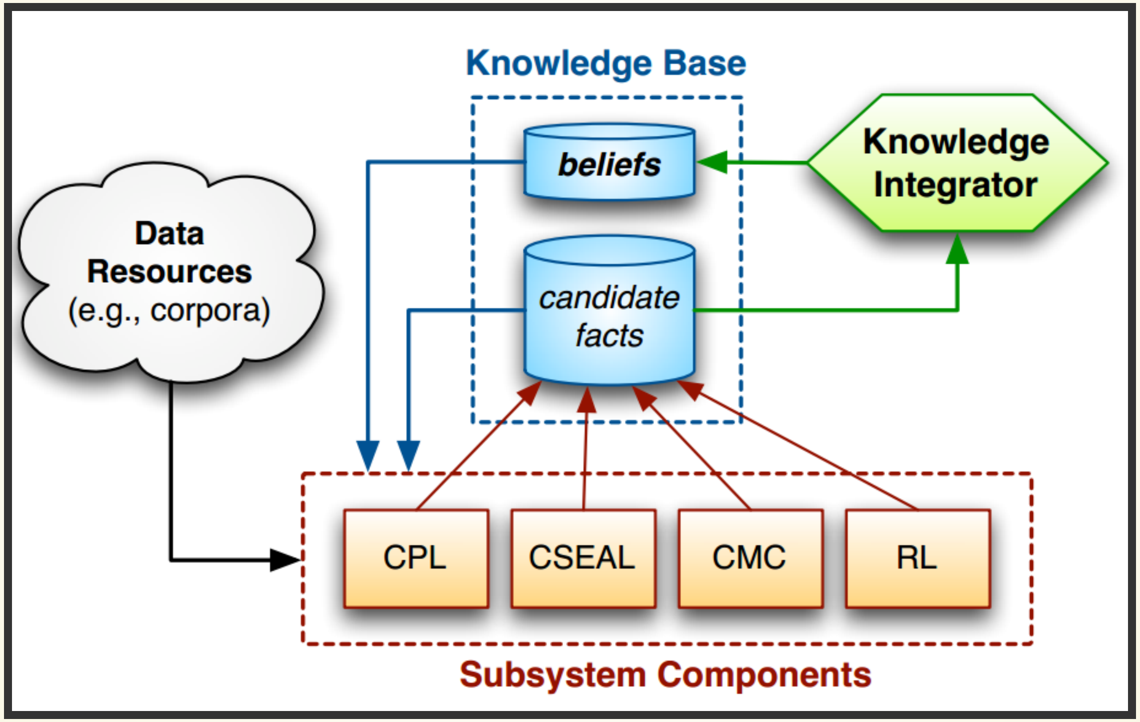
LINK PREDICTION

- Some knowledge can be inferred from already acquired knowledge
  - ```
    [Kyle Korver] plays_for [Atlanta Hawks]
    + [Atlanta Hawks] league [NBA]
    = Means that: [Kyle Korver] is [basketball player]
    ```

---

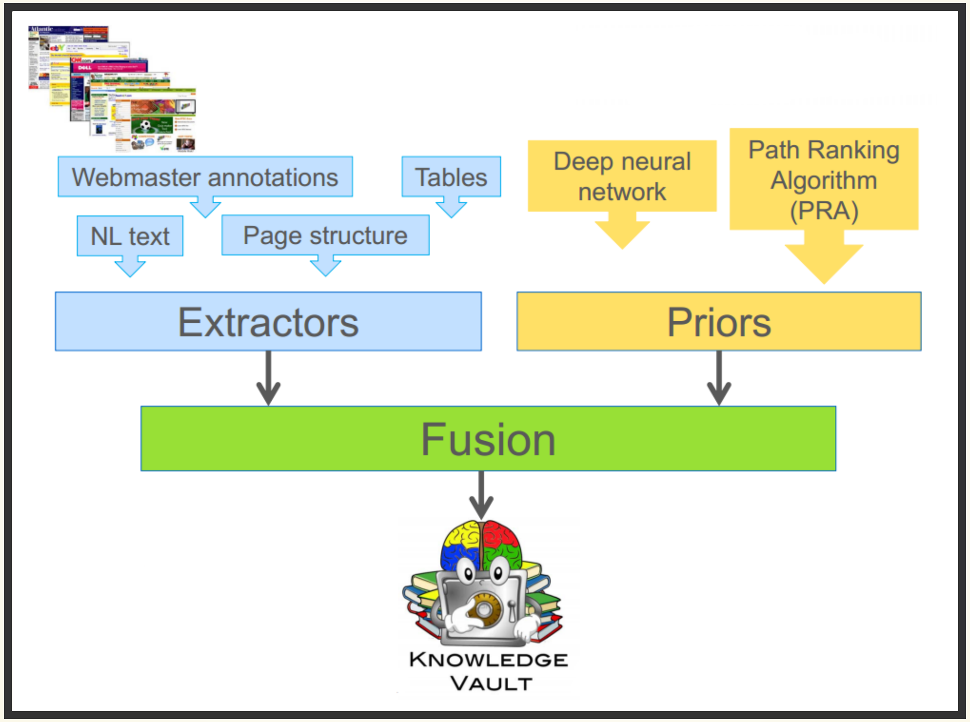
1. "Random Walk Inference and Learning in A Large Scale Knowledge Base" by N.Lao et al, 2011
2. "Logistic Tensor Factorization for Multi-Relational Data" by M.Nickel and B.Tresp, 2013
    ```

NEVER ENDING LANGUAGE LEARNING



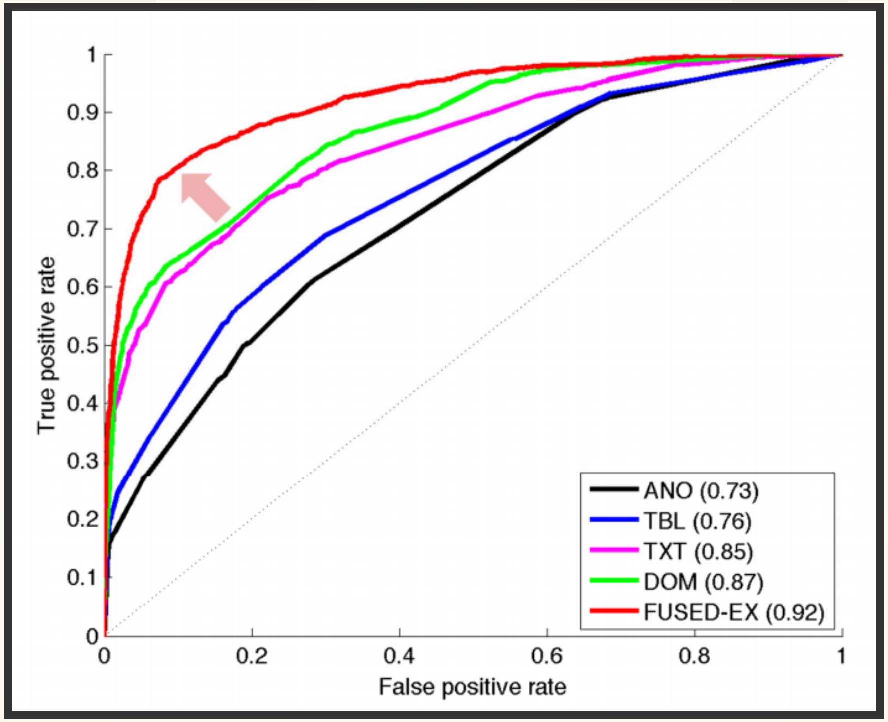
"Toward an Architecture for Never-Ending Language Learning" by A. Carlson et al 2010

GOOGLE KNOWLEDGE VAULT



"Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion" by X.Dong et al 2014

GOOGLE KNOWLEDGE VAULT



from KDD 2014 Tutorial on Constructing and Mining Web-scale Knowledge Graphs, New York, August 24, 2014

SUMMARY

- Computers need data structures
- Knowledge graphs can be used to structure knowledge: entities and relations (RDF graphs)
- Knowledge can be effectively acquired from unstructured data, e.g. natural language text

THANKS!

Questions?