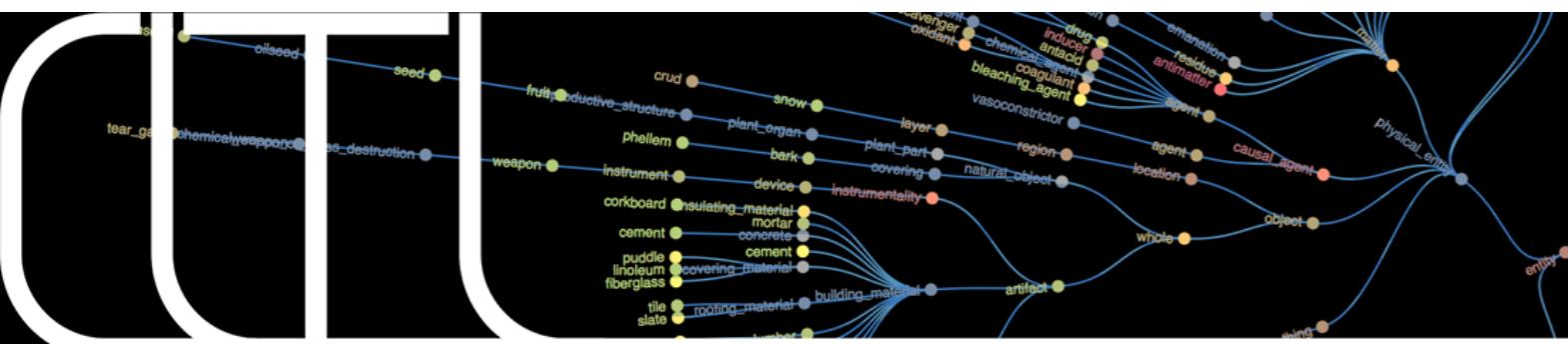


# Text Mining CBS 2019



# Lecture 5: Named entity disambiguation and linking

## Piek Vossen



# Lecture overview

1. Task introduction

a. Definition

b. Knowledge bases

c. Opportunities and challenges

2. Phases of entity linking

3. Entity linkers

a. Approaches

b. Tools

4. Evaluation

a. Aggregation

b. Example

# Entity tasks in NLP

- NER (Recognition): detecting the phrase that is the name of an entity
- NEC (Classification): assigning an entity type to the phrase
- NEL (Linking): establishing the identity of the entity in a given reference database (Wikipedia, DBpedia, YAGO)
- Coreference: any phrase that makes reference to an entity instance, including pronouns, noun phrases, abbreviations, acronyms, etc...

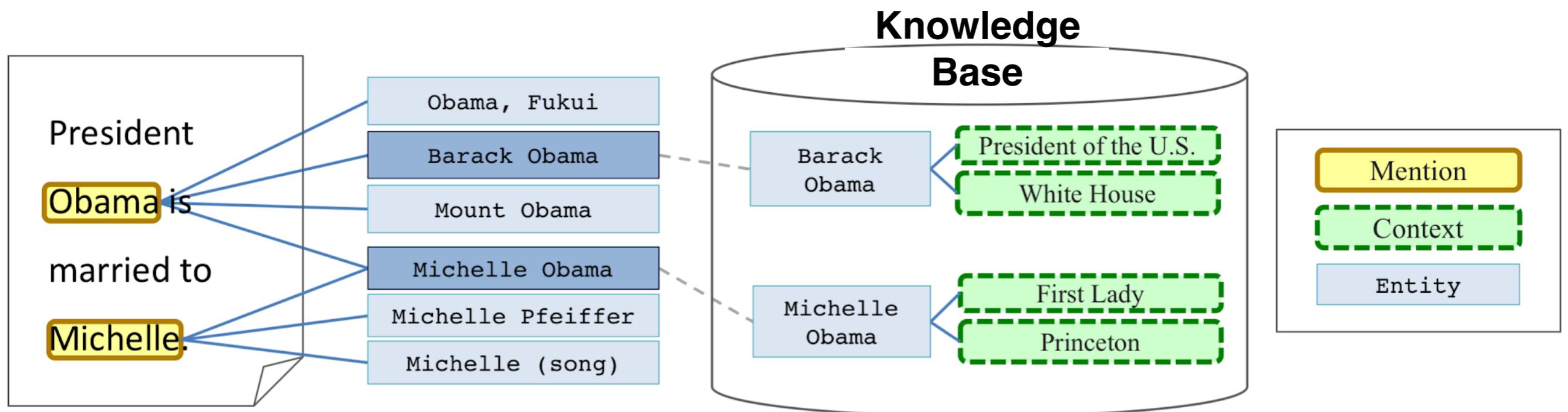
# Task definition

- Potentially ambiguous **entity mention** (“Paris”) needs to be linked to a canonical identifier/**instance** (<http://dbpedia.org/resource/Paris>) that fits the intended referent in the context of the text
- We find these instances in a **knowledge base**.

# Example

“President Obama is married to Michelle.”

# Example



# Knowledge base

A catalog of things, usually entities. Each one has:

- **one or more names**
  - [Mount Everest](#) -> “Mount Everest”, “Everest”, “Mount Qomolangma”, “Mt. Qomolangma”, “Mount Sagarmatha”, “Qomolangma”, “Chomolangma”, “Mt. Everest”, ...
- **other attributes**
  - Elevation: 8,848m
  - Coordinates: 27°59'17"N, 86°55'31"E
- **Connections to other entities**
  - Continent: Asia
  - Country: China, Country: Nepal
- **Textual description**
  - [example](#)

Usually a knowledge base has some of these aspects, but not all.

# Knowledge base types

The knowledge bases can be classified into two types:

- Unstructured (e.g., Wikipedia)
  - Mostly contains a textual (“unstructured”) description
- Structured (e.g., Wikidata, DBpedia, ...)
  - Contains structured description of an entity
  - Property-value pairs

# Unstructured knowledge bases: Wikipedia

https://en.wikipedia.org/wiki/Mount\_Everest

Mount Everest

From Wikipedia, the free encyclopedia

Coordinates: 27°59'17"N 86°55'31"E

"Everest" redirects here. For other uses, see [Everest \(disambiguation\)](#).

This article's tone or style may not reflect the encyclopedic tone used on Wikipedia. See Wikipedia's guide to writing better articles for suggestions. (October 2017) ([Learn how and when to remove this template message](#))

**Mount Everest**, known in [Nepali](#) as [Sagarmatha](#) (सगरमाथा) and in [Tibetan](#) as [Chomolungma](#) (ཇོ་མོ་གླང་མ), is Earth's highest mountain above sea level, located in the [Mahalangur Himal](#) sub-range of the [Himalayas](#). The international border between [Nepal](#) (Province No. 1) and [China](#) (Tibet Autonomous Region) runs across its [summit point](#).

The current official elevation of 8,848 m (29,029 ft), recognized by China and Nepal, was established by a 1955 Indian survey and subsequently confirmed by a Chinese survey in 1975.<sup>[1]</sup> In 2005, China remeasured the rock height of the mountain, with a result of 8844.43 m (29,017 ft). There followed an argument between China and Nepal as to whether the official height should be the rock height (8,844 m., China) or the snow height (8,848 m., Nepal). In 2010, an agreement was reached by both sides that the height of Everest is 8,848 m, and Nepal recognizes China's claim that the rock height of Everest is 8,844 m.<sup>[5]</sup>

In 1865, Everest was given its official English name by the [Royal Geographical Society](#), upon a recommendation by [Andrew Waugh](#), the British [Surveyor General of India](#). As there appeared to be several different local names, Waugh chose to name the mountain after his predecessor in the post, [Sir George Everest](#), despite Everest's objections.<sup>[6]</sup>

Mount Everest attracts many climbers, some of them highly experienced mountaineers. There are two main climbing routes, one approaching the summit from the southeast in Nepal (known as the "standard route") and the other from the north in Tibet. While not posing substantial technical climbing challenges on the standard route, Everest presents dangers such as [altitude sickness](#), weather, and wind, as well as significant hazards from avalanches and the [Khumbu Icefall](#). As of 2017, nearly 300 people have [died on Everest](#), many of whose bodies remain on the mountain.<sup>[7]</sup>

The first recorded efforts to reach Everest's summit were made by British [mountaineers](#). As Nepal did not allow foreigners into the country at the time, the British made several attempts on the north ridge route from the Tibetan side. After the first [reconnaissance expedition](#) by the British in 1921 reached 7,000 m (22,970 ft) on the North Col, the [1922 expedition](#) pushed the north ridge route up to 8,320 m (27,300 ft), marking the first time a human had climbed above 8,000 m (26,247 ft). Seven porters were killed in an avalanche on the descent from the North Col. The [1924 expedition](#) resulted in one of the greatest mysteries on Everest to this day: [George Mallory](#) and [Andrew Irvine](#) made a final summit attempt on 8 June but never returned, sparking debate as to whether or not they were the first to reach the top. They had been spotted high on the mountain that day but disappeared in the clouds, never to be seen again, until Mallory's body was found in 1999 at 8,155 m (26,755 ft) on the north face. [Tenzing Norgay](#) and [Edmund Hillary](#) made the [first official ascent of Everest in 1953](#), using the southeast ridge route. Norgay had reached 8,595 m (28,199 ft) the previous year as a member of the [1952 Swiss expedition](#). The Chinese mountaineering team of [Wang Fuzhou](#), [Gonpo](#), and [Qu Yinhua](#) made the first reported [ascent of the peak from the north ridge](#) on 25 May 1960.<sup>[8][9]</sup>

**Contents** [hide]

- [1 History](#)
- [2 Early surveys](#)
- [3 Name](#)
- [4 Surveys](#)
  - [4.1 Comparisons](#)
- [5 Geology](#)
- [6 Flora and fauna](#)
- [7 Environment](#)

**Mount Everest**



Mount Everest as viewed from Kalapathar.

Highest point	
Elevation	8,848 metres (29,029 ft) <sup>[1]</sup> Ranked 1st
Prominence	Ranked 1st (Notice special definition for Everest)
Listing	Seven Summits Eight-thousander Country high point Ultra
Coordinates	27°59'17"N 86°55'31"E <sup>[2]</sup>
Naming	

# Structured knowledge bases: DBpedia & Wikidata

dbpedia.org/page/Mount\_Everest

**DBpedia** Browse using ▾ Formats ▾

geo:geometry ■ POINT(86.925277709961 27.988056182861)

geo:lat ■ 27.988056 (xsd:float)

geo:long ■ 86.925278 (xsd:float)

prov:wasDerivedFrom ■ wikipedia-en:Mount\_Everest?oldid=744845387

foaf:depiction ■ wiki-commons:Special:FilePath/Mount-Everest.jpg

foaf:isPrimaryTopicOf ■ wikipedia-en:Mount\_Everest

foaf:name ■ Mount Everest (en)

is dbo:deathPlace of

- dbr:Shailendra\_Kumar\_Upadhyaya
- dbr:Mick\_Burke\_(mountaineer)
- dbr:Ray\_Genet
- dbr:Scott\_Fischer
- dbr:Karl\_Gordon\_Henize
- dbr:Hristo\_Prodanov
- dbr:David\_Sharp\_(mountaineer)
- dbr:Rob\_Hall
- dbr:Pasang\_Lhamu\_Sherpa
- dbr:Zygmunt\_Andrzej\_Heinrich
- dbr:Mohammad\_Khaled\_Hossain
- dbr:Andrew\_Irvine\_(mountaineer)
- dbr:Maurice\_Wilson

https://www.wikidata.org/wiki/Q513

instance of mountain ▾ 0 references

part of Seven Summits ▾ 1 reference

Himalayas ▾ 1 reference

image

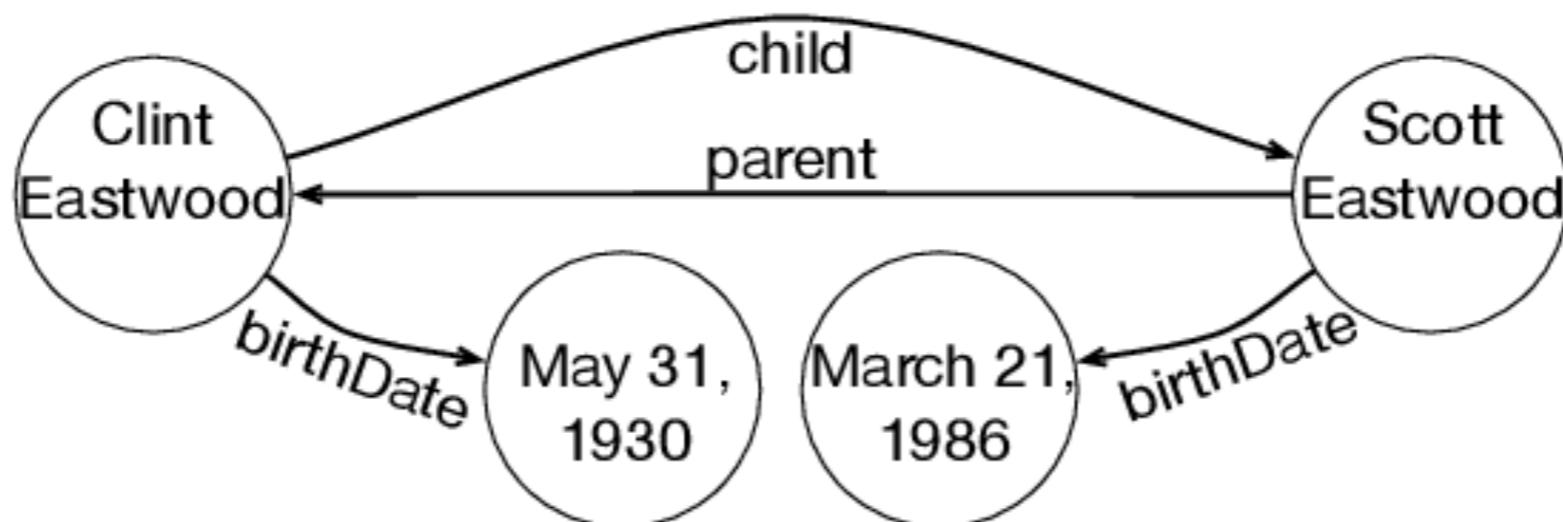


Mount Everest from Rongbuk may 2005.JPG  
3,008 × 2,000; 1.12 MB

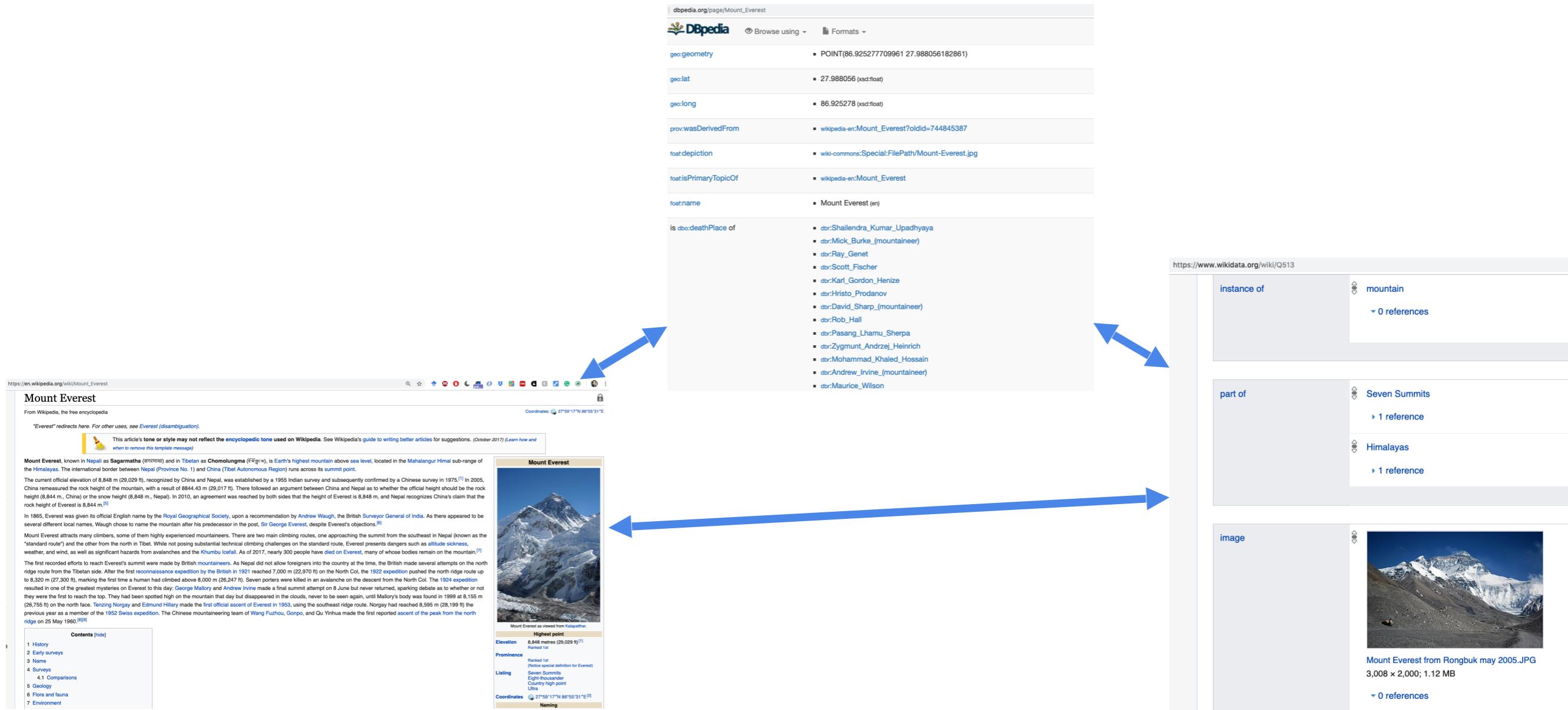
▼ 0 references

# Structured KBs are essentially graph networks

... with billions (!) of such facts

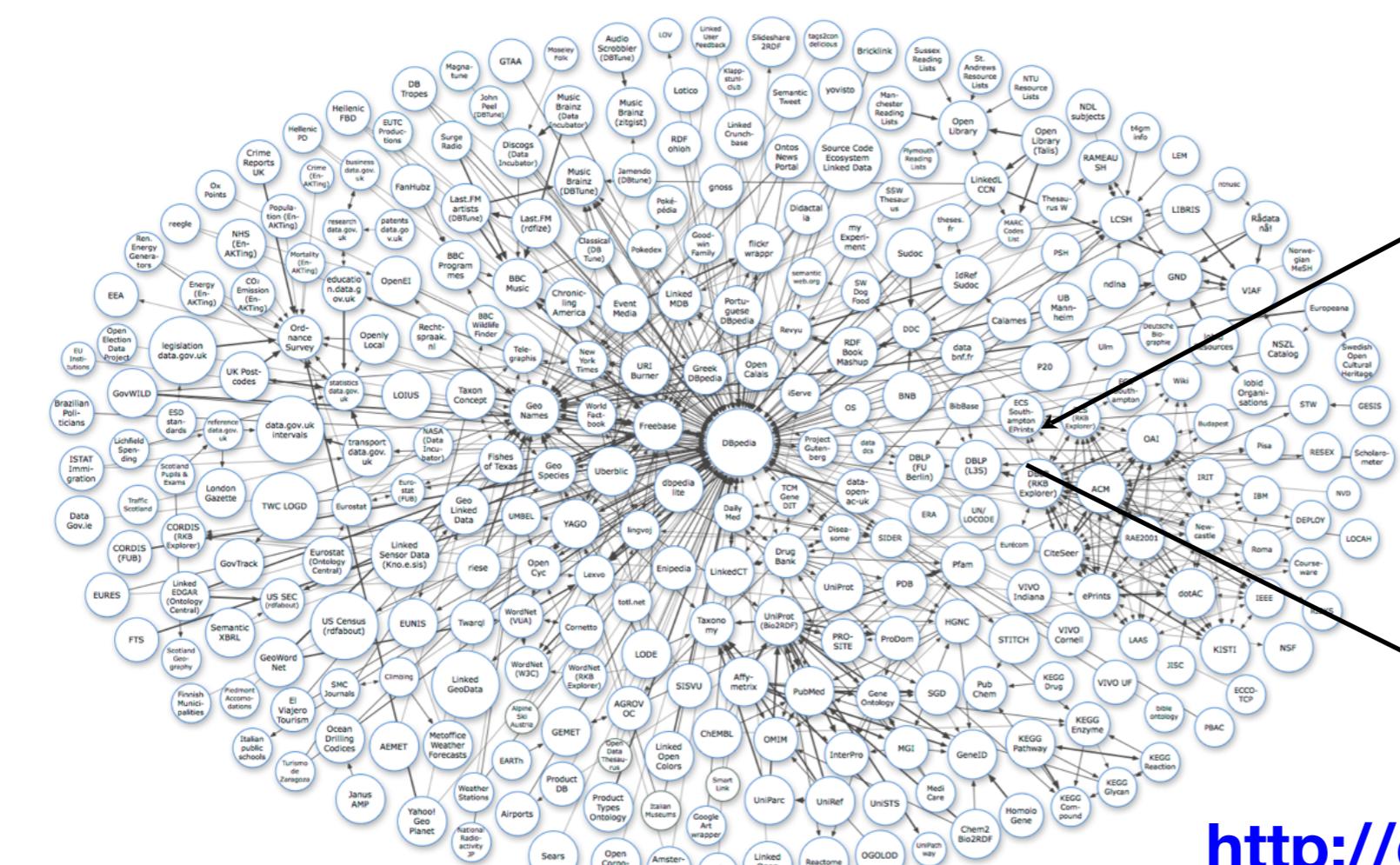


# Knowledge bases are also connected to each other



# Many more KBs exist, and they are connected -> the LOD Cloud

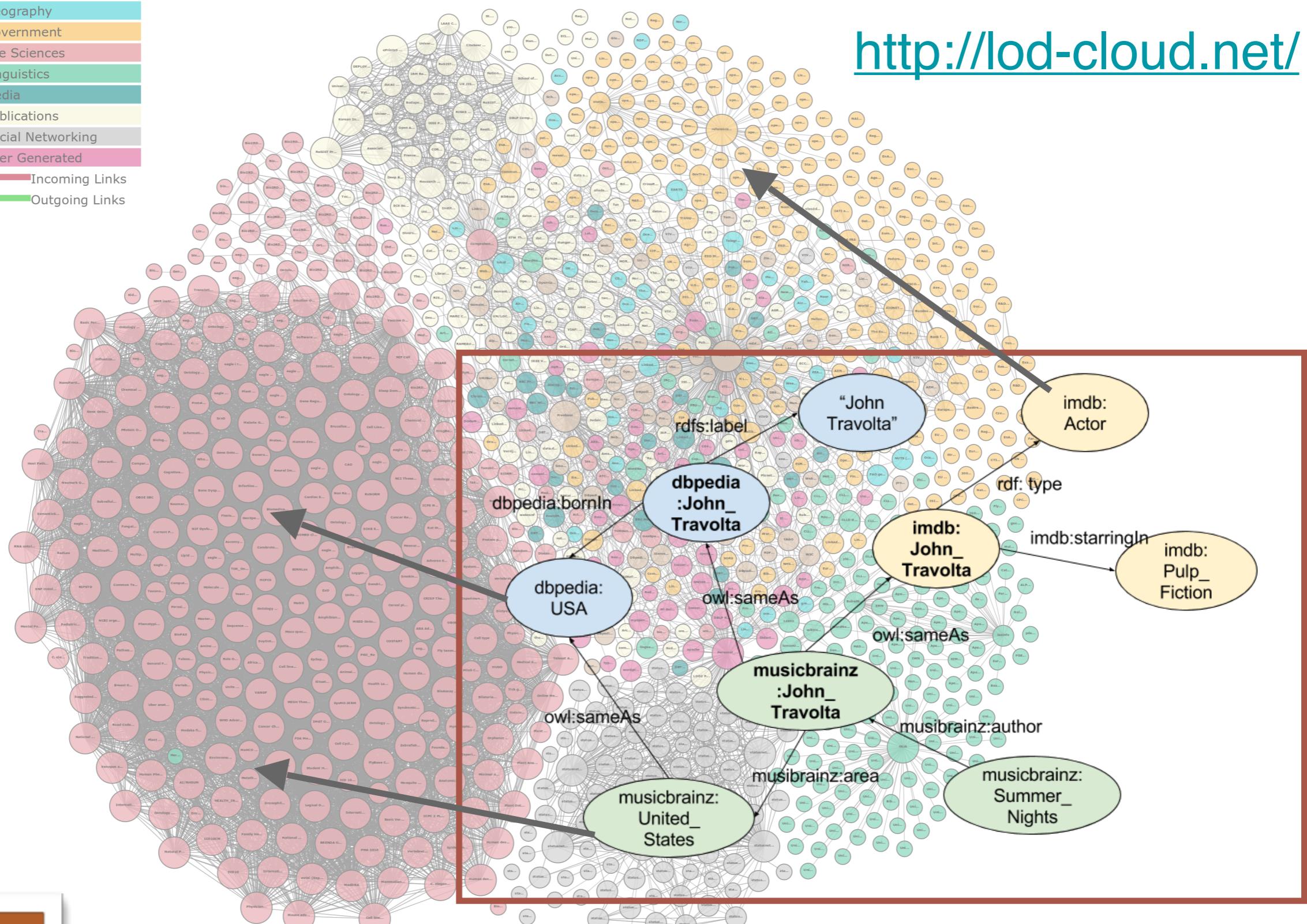
“John Travolta”



[http://dbpedia.org/resource/  
John Travolta](http://dbpedia.org/resource/John_Travolta)

### Legend

Cross Domain
Geography
Government
Life Sciences
Linguistics
Media
Publications
Social Networking
User Generated
Incoming Links
Outgoing Links



Power of the  
division of labour

12 data sets  
2 billion triples

2007

295 data sets  
31 billion triples

2011

2017

<http://lod-cloud.net/>

# Other benefits of connecting text and knowledge bases



# Named Entity Disambiguation isn't that easy though

The screenshot shows a web browser window with the URL dbpedia.org/page/Lincoln. The page title is "About: Lincoln". Below it, text states "An Entity of Type : Thing, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org". The DBpedia logo is visible. A table lists properties and their values. The "dbpedia-owl:wikiPageDisambiguates" property is highlighted with a red oval, and its value is listed below:

Property	Value
dbpedia-owl:wikiPageDisambiguates	<ul style="list-style-type: none"><li>▪ dbpedia:Lincoln_Memorial</li><li>▪ dbpedia:Lincoln_Motor_Company</li><li>▪ dbpedia:Lincolnshire</li><li>▪ dbpedia:Abraham_Lincoln</li><li>▪ dbpedia:Lincoln,_England</li><li>▪ dbpedia:Lincoln,_Nebraska</li><li>▪ dbpedia:Lincoln,_New_Hampshire</li><li>▪ dbpedia:Lincoln,_Alabama</li><li>▪ dbpedia:Lincoln_Highway</li><li>▪ dbpedia:Lincoln_Parish,_Louisiana</li><li>▪ dbpedia:Lincoln,_Adams_County,_Wisconsin</li><li>▪ dbpedia:Lincoln,_Arkansas</li><li>▪ dbpedia:Lincoln,_Bayfield_County,_Wisconsin</li></ul>

## a. name **ambiguity**

Very frequent => Wikipedia and DBpedia have special **disambiguation** pages that list the entities that are referred to by a mention.

# Named Entity Disambiguation isn't that easy though

The screenshot shows a web browser window with the URL dbpedia.org/page/Lincoln. The page title is "About: Lincoln". Below it, a sub-header states: "An Entity of Type : Thing, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org". The DBpedia logo is visible. A red circle highlights a list of entity values under the "Value" column for the property "dbpedia-owl:wikiPageDisambiguates". The list includes: dbpedia:Lincoln\_Memorial, dbpedia:Lincoln\_Motor\_Company, dbpedia:Lincolnshire, dbpedia:Abraham\_Lincoln, dbpedia:Lincoln,\_England, dbpedia:Lincoln,\_Nebraska, dbpedia:Lincoln,\_New\_Hampshire, dbpedia:Lincoln,\_Alabama, dbpedia:Lincoln\_Highway, dbpedia:Lincoln\_Parish,\_Louisiana, dbpedia:Lincoln,\_Adams\_County,\_Wisconsin, dbpedia:Lincoln,\_Arkansas, and dbpedia:Lincoln,\_Bayfield\_County,\_Wisconsin.

Property	Value
dbpedia-owl:wikiPageDisambiguates	<ul style="list-style-type: none"><li>▪ dbpedia:Lincoln_Memorial</li><li>▪ dbpedia:Lincoln_Motor_Company</li><li>▪ dbpedia:Lincolnshire</li><li>▪ dbpedia:Abraham_Lincoln</li><li>▪ dbpedia:Lincoln,_England</li><li>▪ dbpedia:Lincoln,_Nebraska</li><li>▪ dbpedia:Lincoln,_New_Hampshire</li><li>▪ dbpedia:Lincoln,_Alabama</li><li>▪ dbpedia:Lincoln_Highway</li><li>▪ dbpedia:Lincoln_Parish,_Louisiana</li><li>▪ dbpedia:Lincoln,_Adams_County,_Wisconsin</li><li>▪ dbpedia:Lincoln,_Arkansas</li><li>▪ dbpedia:Lincoln,_Bayfield_County,_Wisconsin</li></ul>

## a. name ambiguity

Very frequent => Wikipedia and DBpedia have special **disambiguation** pages that list the entities that are referred to by a mention.

## Abraham Lincoln (Q91)

16th President of the United States

Honest Abe | A. Lincoln | President Lincoln | Abe Lincoln | Lincoln

## b. name variation

(+ “Mr. Lincoln”, “Lincoln”, “Abraham”, ...)

# The meaning puzzle: Who is Ford?

---

- ❖ President Woodrow Wilson asked *Ford*[?] to run as a Democrat for the United States Senate from Michigan in 1918.
- ❖ Gerald Ford
- ❖ Ford, the motor company
- ❖ Henry Ford

---

# The meaning puzzle: Who is Ford?

---

President [6] Woodrow Wilson [10] asked [7] Ford [8] to run [41] as a Democrat [2] for the United States [4] Senate [2] from Michigan [3] in 1918.

$6 * 10 * 7 * 8 * 41 * 2 * 4 * 2 * 3 = 6,612,480$  combinations of word senses and entities

# The meaning puzzle: Who is Ford?

---

- ❖ President Woodrow Wilson asked *Ford[?] to run as* a Democrat for the United States Senate from Michigan in 1918.
- ❖ Semantic knowledge:
  - ❖ meaning of *to run as*
- ❖ World knowledge:
  - ❖ run as senator: +human, >18 years old, US citizen
  - ❖ Gerald Ford born in 1917, so 1 year old
  - ❖ Henry Ford the founder of Ford Motor Company, born in 1863, died in 1947, so 56 years old

# c. Missing (NIL) entities

SAN BENITO – Police have released more information connected to a murder investigation in San Benito.

Edgar Gonzalez, 30, was shot and killed last Thursday.

It happened on Buena Vida Street near the expressway and Sam Houston.

Police released a photo of a white Chevrolet Tahoe they say was used as a getaway vehicle.

Police are asking nearby businesses to check their surveillance video for any images of this vehicle.

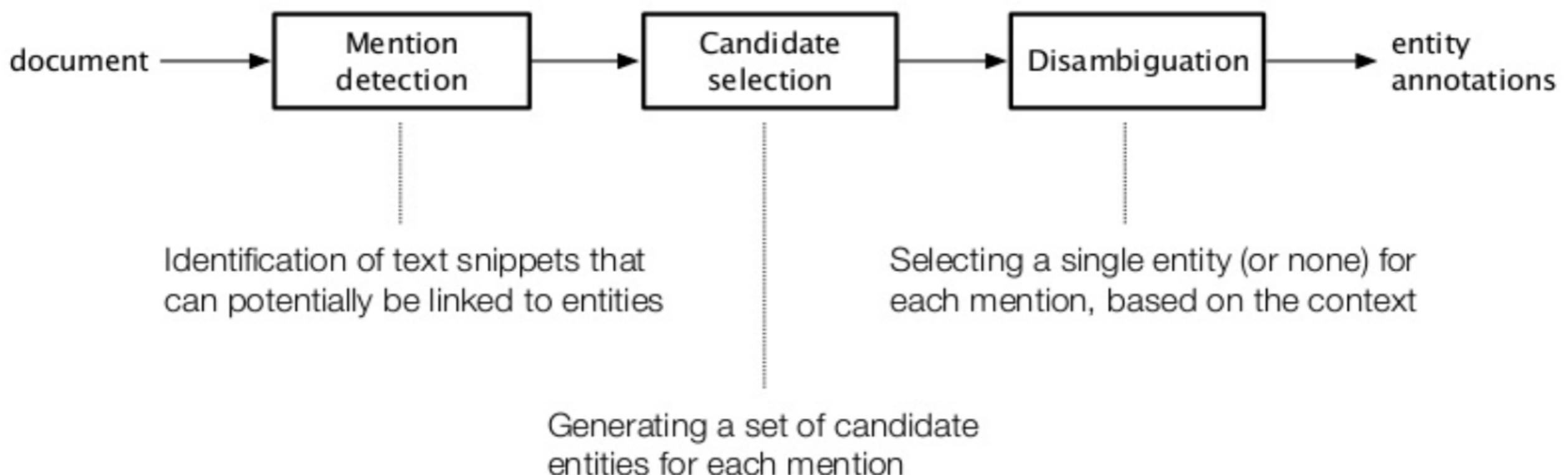
If you have any information, contact the San Benito Police Department at 361-3880.

What to link “Edgar Gonzalez” to???

<https://www.krgv.com/news/police-seeking-surveillance-footage-to-aid-in-san-benito-murder-investigation>

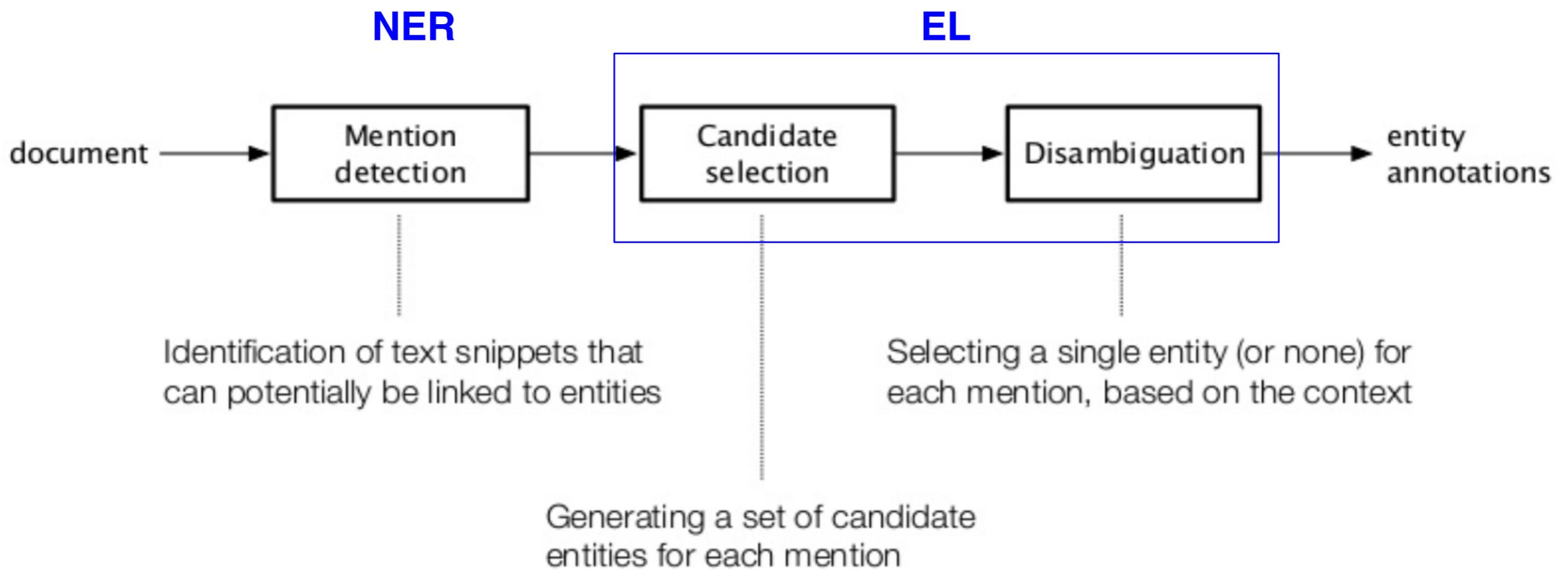
## 2. Components of an entity linking system

### Anatomy of an entity linking system



**Source:** <https://www.slideshare.net/krisztianbalog/entity-linking-65308055>

# Anatomy of an entity linking system



**Source:** <https://www.slideshare.net/krisztianbalog/entity-linking-65308055>

# Phase I: Recognition

Same as we saw in the previous lecture

Detect entity mentions in text

# Phase II: Candidate generation/selection

- For each of the recognised mentions in text, get the potential referents (instances) in a knowledge base (KB), following the “closed world assumption” (=the world is in the KB).
- The goal is to balance between generating too many candidates (too much ‘noise’) and generating too little candidates (missing the correct one)
- Trade-off between precision and recall
- Candidate generation is an art by itself!

In practice, about 30 candidates per mention is enough.

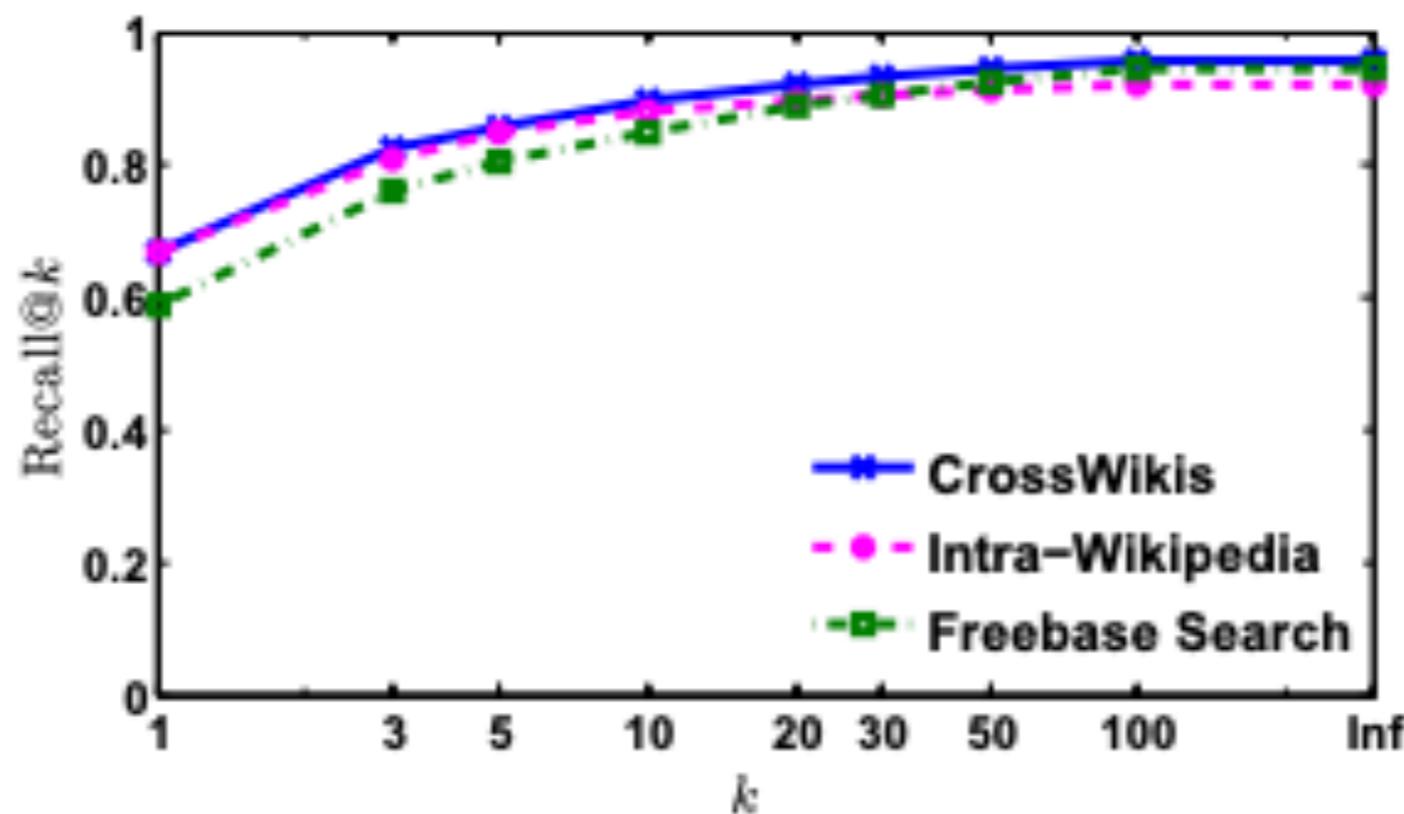


Figure 3: Recall@ $k$  on an aggregate of nine data sets, comparing three **candidate generation** methods.

Source: [https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacl\\_a\\_00141](https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacl_a_00141)

# But... how do you choose the top X (or 30) candidates?

- We need a way to rank them somehow.
  - A common ranking criteria is **commonness**: for a given mention, how relatively often it refers to some instance in Wikipedia.
  - For example, of all the mentions of “Germany” in Wikipedia, what is the percentage that refers to the country vs the football club vs the handball club vs the government vs etc.
- Perform the ranking of candidate entities based on their overall popularity, i.e., "most common sense"

$$P(e|m) = \frac{n(m, e)}{\sum_{e'} n(m, e')}$$

→ the number of times entity  $e$  is the link destination of mention  $m$

→ total number of times mention  $m$  appears as a link

# Example

Bulgaria's best **World Cup** performance was in the **1994 World Cup** where they beat **Germany**, to reach the semi-finals, losing to Italy, and finishing in fourth ...

Entity	Commonness
FIFA_World_Cup	0.2358
FIS_Apline_Ski_World_Cup	0.0682
2009_FINNA_Swimming_World_Cup	0.0633
World_Cup_(men's_golf)	0.0622
...	

Entity	Commonness
1998_FIFA_World_Cup	0.9556
1998_IAAF_World_Cup	0.0296
1998_Alpine_Skiing_World_Cup	0.0059
...	

Entity	Commonness
Germany	0.9417
Germany_national_football_team	0.0139
Nazi_Germany	0.0081
German_Empire	0.0065
...	

Also, observe:

- Dominance within a form
- Topical bias

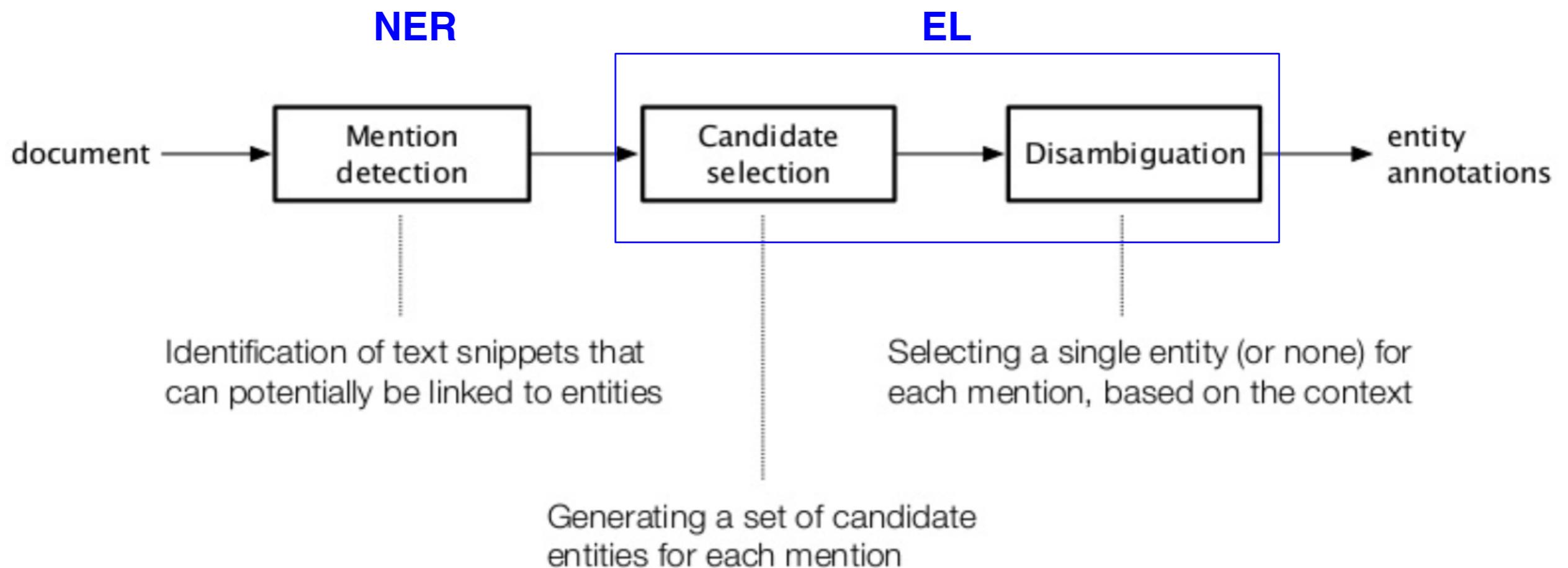
# Phase III: Disambiguation

Goal: decide which of the candidates (or none) is the correct referent.

When would this phase be easy and when difficult?

# 3. Tools/systems

## Anatomy of an entity linking system



# Entity linking methods

	<b>Word-based</b>	<b>Graph-based</b>
Main idea	Find the candidate with the most similar description to the one of a mention in text	Find candidates that are coherent with each other according to connections in the KB
Scoring example	measure text similarity, combine with <b>TF/IDF</b> weighting to measure relevance of a word	Put all candidates with their facts in a graph network and prune until only one candidate per mention is left
Decision unit	individual/local	collective/global
KB	unstructured (Wikipedia)	Structured (DBpedia, etc.)
Example	<u><a href="#">DBpedia Spotlight</a></u>	<u><a href="#">AIDA/AGDISTIS</a></u>

TF/IDF = term frequency \* inverse document frequency

Measures the degree to which terms are important for a document, based on the frequency in the document but normalised by checking if it occurs in all documents or just a few

# 3a. Word-based methods: DBpedia Spotlight

- Compute cosine similarity between the text paragraph with an entity mention and Wikipedia descriptions of each candidate.
- Decide for one mention at a time.
- The linking can be restricted to certain types or even to a custom set of entities.

The screenshot shows the DBpedia Spotlight interface. At the top, there's a logo with a blue and yellow sunburst graphic and the text "DBpediaSpotlight". Below the logo are several input fields and dropdown menus:

- Confidence: A slider set to 0.0.
- Contextual score: A slider set to 0.0.
- Prominence (support): A slider set to 0.
- Buttons: "No 'common words'", "Default Disambiguation", "Show best candidate", "SELECT TYPES...", and "ANNOTATE".

A large text area displays the following text about Ry Cooder:

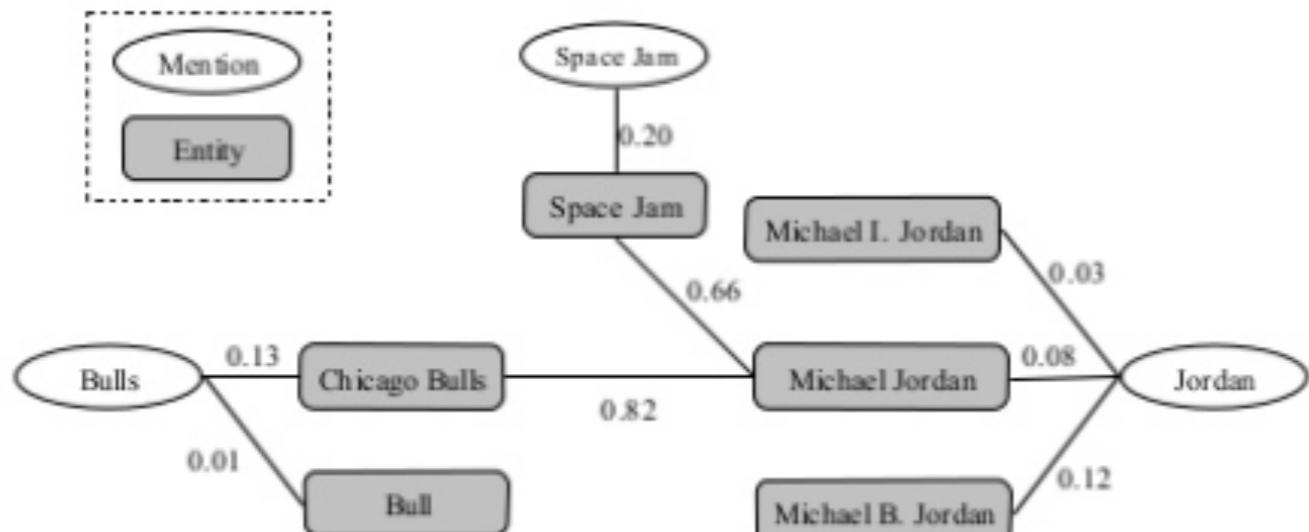
Ryland Peter "Ry" Cooder (born [March 15, 1947](#)) is an [American](#) guitarist, [singer](#) and [composer](#). He is known for his [slide guitar](#) work, his interest in [roots music](#) from the [United States](#), and, more recently, his collaborations with traditional musicians from many [countries](#).  
[Ry Cooder](#) grew up in [Santa Monica, California](#), and attended [Santa Monica High School](#). His [solo](#) work has been eclectic, encompassing [folk](#), [blues](#), Tex-Mex, [soul](#), [gospel](#), [rock](#), and much else. He has collaborated with many [musicians](#), including [Larry Blackmon](#), [Eric Clapton](#), [The Rolling Stones](#), [Van Morrison](#), [Neil Young & Crazy Horse](#), [Randy Newman](#), [Taj Mahal](#), [Earl Hines](#), [Little Feat](#), [Captain Beefheart](#), [The Doobie Brothers](#), [The Chieftains](#), [John Lee Hooker](#), [Pops](#) and [Mavis Staples](#), [Flaco Jiménez](#), [Ibrahim Ferrer](#), [Terry Evans](#), [Bobby King](#), [Freddy Fender](#), [Vishwa Mohan Bhatt](#) and [Ali Farka Touré](#). He formed the [band](#) Little Village with [Nick Lowe](#), [John Hiatt](#), and [Jim Keltner](#).

[BACK TO TEXT](#)

# 3b. Graph-based methods: AIDA and AGDISTIS

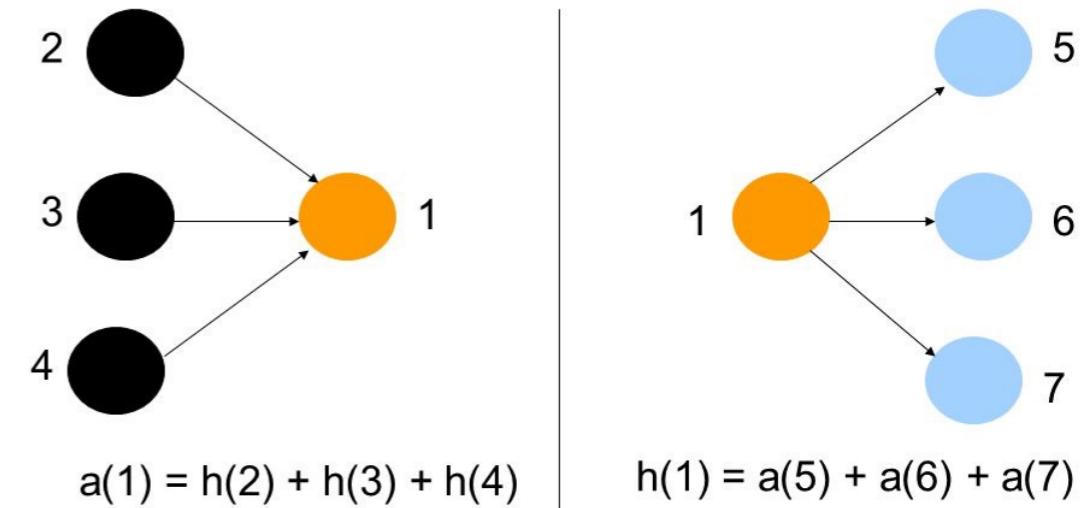
1. Construct a subgraph that contains all entity candidates with some facts from a KB.
2. Find the best connected candidates per mention:

## Example



Compute relatedness between the candidates (AIDA)

## Authority and Hubness



Find the “hubs” in the graph (AGDISTIS)

# Local vs global disambiguation

- Note that the idea in the graph-based approaches is to make the optimal global decision (we disambiguate all entities together).
- This is different than in DBpedia Spotlight, where we disambiguate entities one by one.

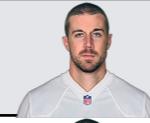
# 4. Evaluation

- The correctness of an entity linking system is measured in terms of precision, recall, and F1-score.
- You already know these metrics, but... here we aggregate the scores differently.
- In Sentiment classification, we compute a score per class (positive, neutral, negative).

# 4. Evaluation

- We could do the same in entity linking, but here we have far too many classes (millions).
- For this reason, we usually evaluate entity linking by aggregating **per mention occurrence.**
- Instead of computing confusion between the classes, here we:
  1. Assign a true positive (TP), false positive (FP), and/or false negative (FN) per mention occurrence
  2. Count the TPs, FPs, and FNs across all mentions
  3. Compute precision, recall, and F1-scores once on top of these

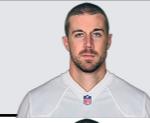
# 4. EL evaluation example

	<b>Chiefs</b>	<b>Alex Smith</b>	<b>Washington</b>	<b>Smith</b>
GOLD				
SYSTEM				NIL
TP, FP, FN	TP	TP	FP, FN	FN

*“It seems like months ago that the Chiefs traded Alex Smith to Washington...”  
Smith, 33, originally entered ...”*

(<https://profootballtalk.nbcsports.com/2018/03/14/washington-announces-alex-smith-trade/>)

# 4. EL evaluation example

	<b>Chiefs</b>	<b>Alex Smith</b>	<b>Washington</b>	<b>Smith</b>
GOLD				
SYSTEM				NIL
TP, FP, FN	TP	TP	FP, FN	FN

*“It seems like months ago that the Chiefs traded Alex Smith to Washington...”  
Smith, 33, originally entered ...”*

$$\text{precision} = \text{TP}/(\text{TP}+\text{FP}) = 2/3 \sim 0.67$$

$$\text{recall} = \text{TP}/(\text{TP}+\text{FN}) = 2/4 = 0.5$$

$$\text{f1} = 2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall}) = 0.67 / 1.17 = 0.57$$

# EL data sets

Group	Data Set	# of Mentions	Entity Types	KB	# of NILs	Eval. Metric
UIUC	ACE	244	Any Wikipedia Topic	Wikipedia	0	BOC F1
	MSNBC	654	Any Wikipedia Topic	Wikipedia	0	BOC F1
AIDA	AIDA-dev	5917	PER,ORG,LOC,MISC	Yago	1126	Accuracy
	AIDA-test	5616	PER,ORG,LOC,MISC	Yago	1131	Accuracy
TAC KBP	TAC09	3904	PER <sup>T</sup> ,ORG <sup>T</sup> ,GPE	TAC ⊂ Wiki	2229	Accuracy
	TAC10	2250	PER <sup>T</sup> ,ORG <sup>T</sup> ,GPE	TAC ⊂ Wiki	1230	Accuracy
	TAC10T	1500	PER <sup>T</sup> ,ORG <sup>T</sup> ,GPE	TAC ⊂ Wiki	426	Accuracy
	TAC11	2250	PER <sup>T</sup> ,ORG <sup>T</sup> ,GPE	TAC ⊂ Wiki	1126	B <sup>3</sup> +F1
	TAC12	2226	PER <sup>T</sup> ,ORG <sup>T</sup> ,GPE	TAC ⊂ Wiki	1049	B <sup>3</sup> +F1

- ACE = Automatic content extraction evaluation (follow-up of Message Understanding Competition, MUC), <http://www.itl.nist.gov/iad/mig/tests/ace/>
- MSNBC = news [https://cogcomp.cs.illinois.edu/page/resource\\_view/4](https://cogcomp.cs.illinois.edu/page/resource_view/4)
- AIDA = Accurate Online Disambiguation of Named Entities in Text and Tables, <https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/>
- TAC = Text Analysis Conference, <http://www.nist.gov/tac/>
- KBP = Knowledge Base Population, [www.nist.gov/tac/2016/KBP/](http://www.nist.gov/tac/2016/KBP/)

# End-to-end performance

Approach	TAC09	TAC10	TAC10T	TAC11	TAC12	AIDA-dev	AIDA-test	ACE	MSNBC	Overall
CrossWikis	80.4	85.6	86.9	78.5	62.4	62.6	62.4	87.7	70.3	75.0
+FIGER	81.0	86.1	86.9	78.8	63.5	66.7	64.5	87.7	75.4	76.7
+Coref	80.9	86.2	87.0	78.6	59.9	68.9	66.3	87.7	86.6	78.0
+Coherence =VINCULUM	<b>81.4</b>	<b>86.8</b>	<b>87.0</b>	79.9	63.7	69.4	67.5	<b>88.5</b>	86.9	79.0
AIDA	73.2	78.6	77.5	68.4	52.0	71.9	<b>74.8</b>	77.8	75.4	72.2
WIKIFIER	79.7	86.2	86.3	<b>82.4</b>	<b>64.7</b>	<b>72.1</b>	69.8	85.1	<b>90.1</b>	<b>79.6</b>

Table 6: **End-to-end performance (%)**: We compare VINCULUM in different stages with two state-of-the-art systems, AIDA and WIKIFIER. The column “Overall” lists the average performance of nine data sets for each approach. CrossWikis appears to be a strong baseline. VINCULUM is 0.6% shy from WIKIFIER, each winning in four data sets; AIDA tops both VINCULUM and WIKIFIER on AIDA-test.

Low performance due to extended ambiguity  
Compare most-frequent and less-frequent sense observation

# System comparison

	VINCULUM	AIDA	WIKIFIER
Mention Extraction	NER	NER	NER, noun phrases
Candidate Generation	<b>CrossWikis</b>	an intra-Wikipedia dictionary	an intra-Wikipedia dictionary
Entity Types	<b>FIGER</b>	NER	NER
Coreference	find the representative mention	-	re-rank the candidates
Coherence	link-based similarity, relation triples	link-based similarity	link-based similarity, relation triples
Learning	unsupervised	trained on AIDA	trained on a Wikipedia sample

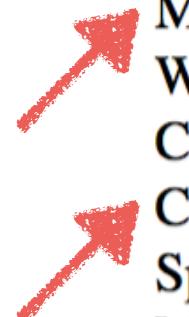
- VINCULUM unsupervised, AIDA & WIKIFIER supervised —> the latter may suffer from **overfitting!**
- Coherence and coreference small gains —> principle or implementation?

# Error analysis

Category	Example	Gold Label	Prediction
Metonymy	<u>South Africa</u> managed to avoid a fifth successive defeat in 1996 at the hands of the All Blacks ...	South Africa national rugby union team	South Africa
Wrong Entity Types	Instead of Los Angeles International, for example, consider flying into <u>Burbank</u> or <u>John Wayne Airport</u> ...	Bob Hope Airport	Burbank, California
Coreference	It is about his mysterious father, <u>Barack Hussein Obama</u> , an imperious if alluring voice gone distant and then missing.	Barack Obama Sr.	Barack Obama
Context	<u>Scott Walker</u> removed himself from the race, but Green never really stirred the passions of former Walker supporters, nor did he garner outsized support “outstate”.	Scott Walker (politician)	Scott Walker (singer)
Specific Labels	What we like would be Seles , ( <u>Olympic</u> champion Lindsay ) Davenport and Mary Joe Fernandez .	1996 Summer Olympics	Olympic Games
Misc	<u>NEW YORK</u> 1996-12-07	New York City	New York

Table 8: We divide linking errors into **six error categories** and provide an example for each class.

# Error analysis

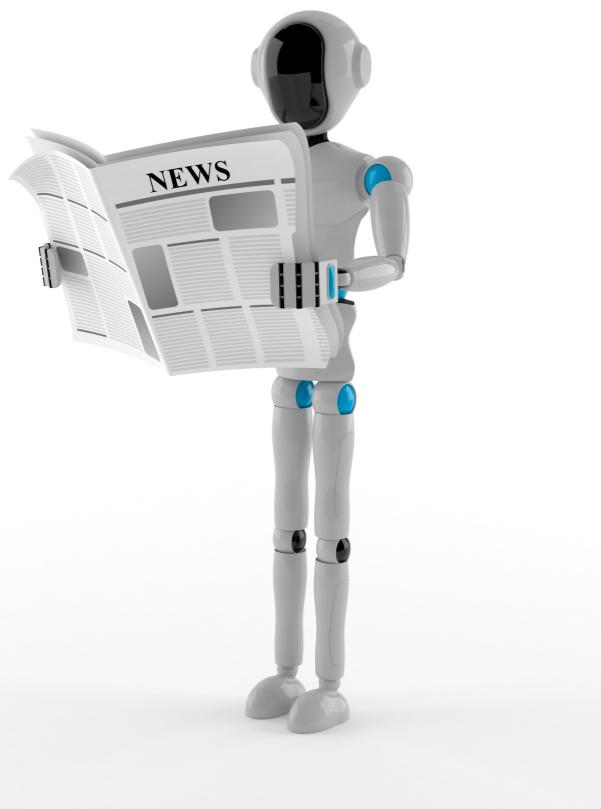


Error Category	TAC09	TAC10	TAC10T	TAC11	TAC12	AIDA-dev	AIDA-test	ACE	MSNBC
Metonymy	16.7%	0.0%	3.3%	0.0%	0.0%	60.0%	60.0%	5.3%	20.0%
Wrong Entity Types	13.3%	23.3%	20.0%	6.7%	10.0%	6.7%	10.0%	31.6%	5.0%
Coreference	30.0%	6.7%	20.0%	6.7%	3.3%	0.0%	0.0%	0.0%	20.0%
Context	30.0%	26.7%	26.7%	70.0%	70.0%	13.3%	16.7%	15.8%	15.0%
Specific Labels	6.7%	36.7%	16.7%	10.0%	3.3%	3.3%	3.3%	36.9%	25.0%
Misc	3.3%	6.7%	13.3%	6.7%	13.3%	16.7%	10.0%	10.5%	15.0%
# of examined errors	30	30	30	30	30	30	30	19	20

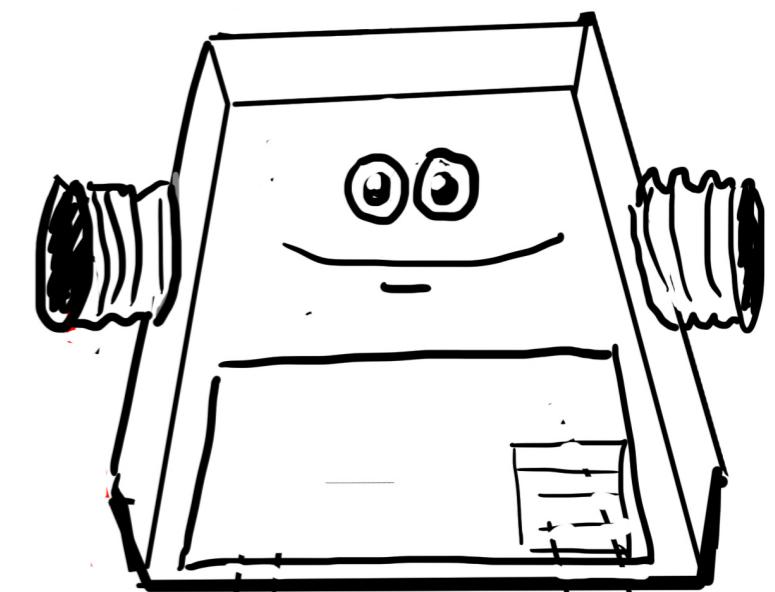
Table 9: **Error analysis:** We analyze a random sample of 250 of VINCULUM’s errors, categorize the errors into six classes, and display the frequencies of each type across the nine datasets.

- Metonymy errors due to wrong FIGER type prediction —> better FIGER typing
- Context errors due to graph approach —> hybrid method including machine learning from training set with context features.

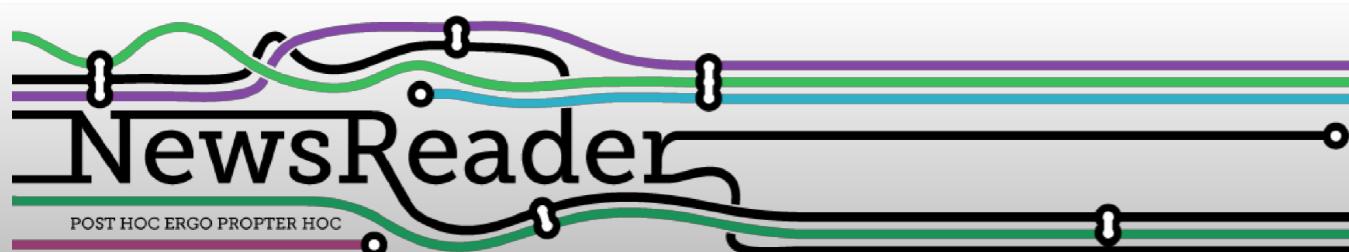
# The case of NewsReader



ICT 316404, FP7-ICT-2011-8  
[www.newsreader-project.eu](http://www.newsreader-project.eu)



<https://youtu.be/rYLaVN3oqLI>

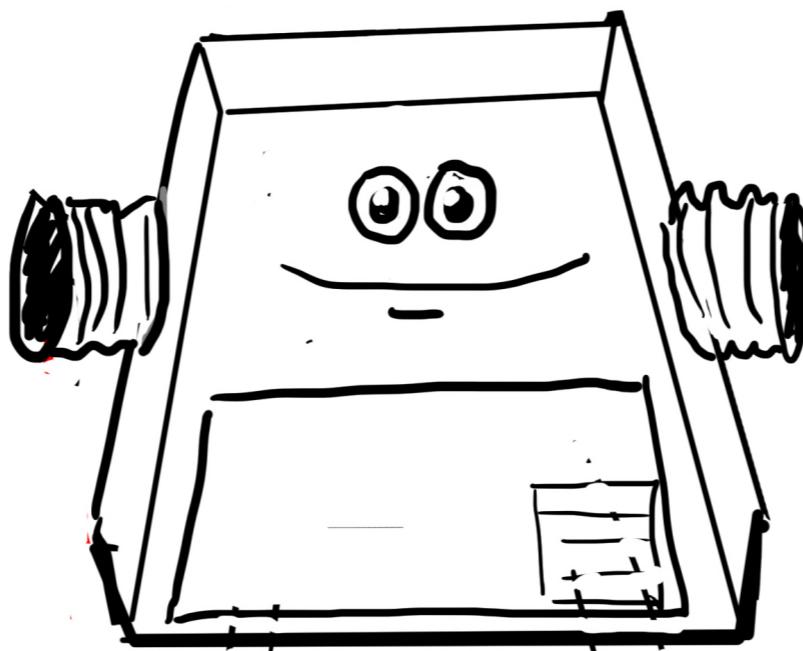


# The case of NewsReader *a reading machine*

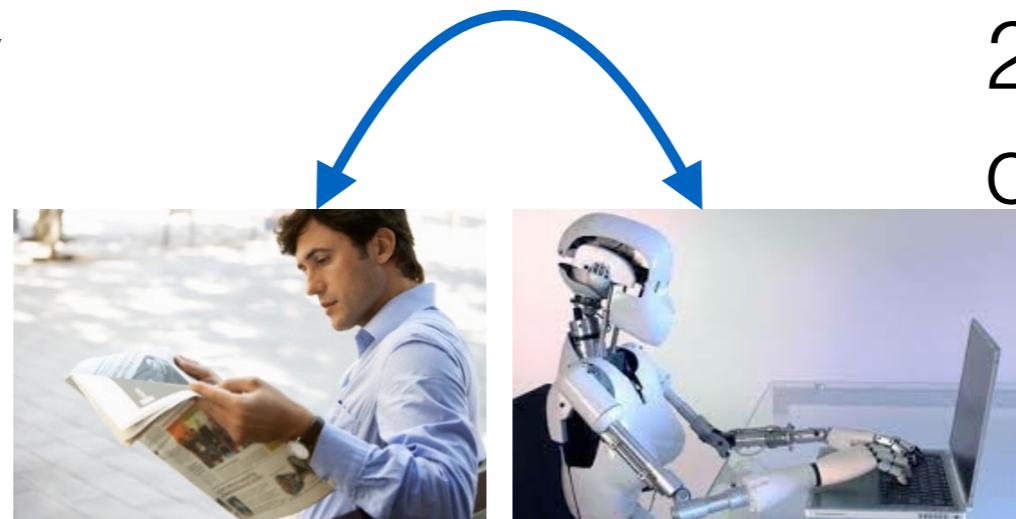
automotive industry  
2003 - 2015



2.3 million articles  
50 million entity  
mentions

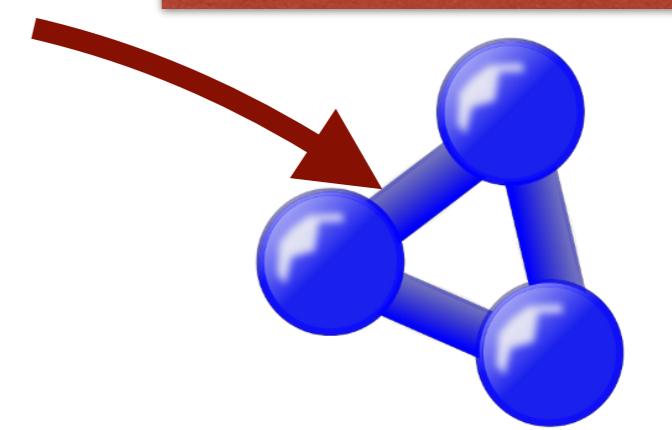


*Ask instead of Read*



Semantic Web  
RDF-TRIPLES

*what - who - where - when*



1.2 billion statements  
2.2 million people,  
organisations, places

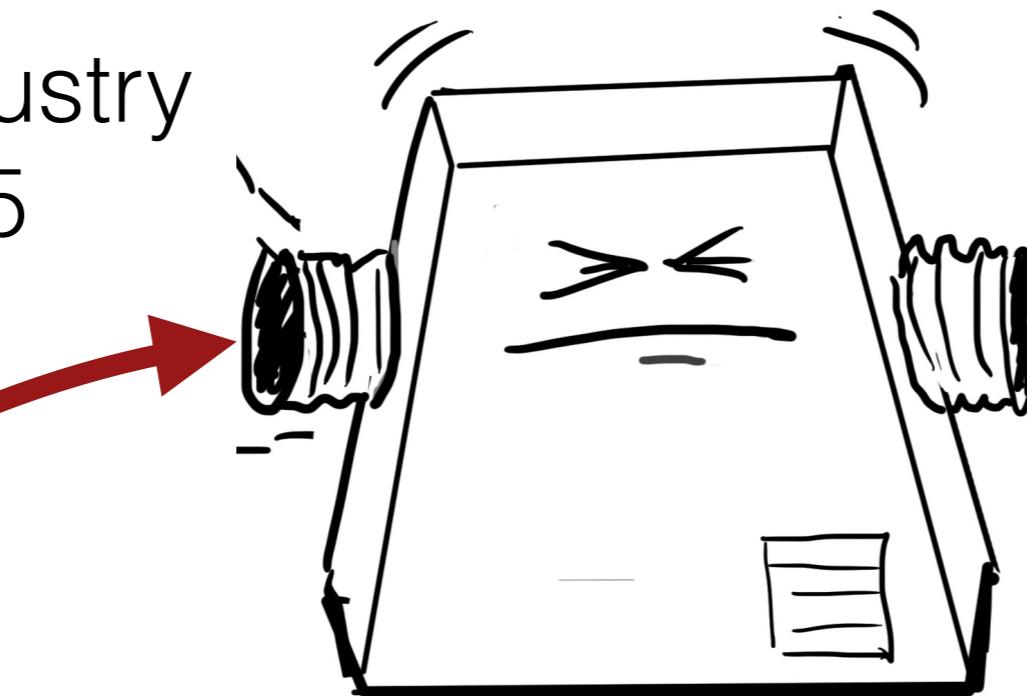
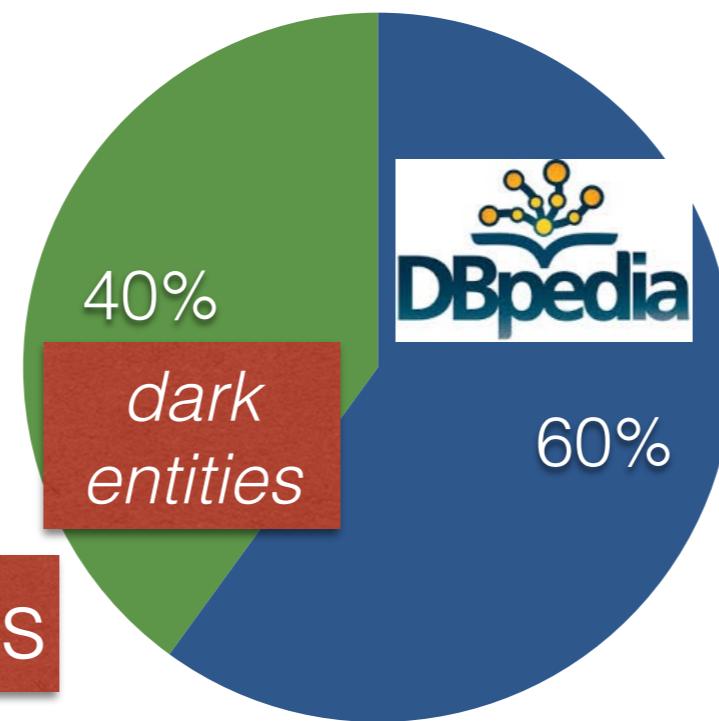
# Dark entities and false identities

automotive industry  
2003 - 2015



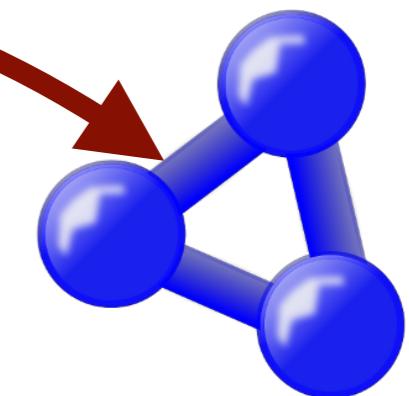
2.3 million articles  
50 million entity  
mentions

40% 'dark' entities



Semantic Web  
RDF-TRIPLES

*what - who - where - when*



1.2 billion statements  
2.2 million people,  
organisations, places

60% 'known' entities

# Entities

	in DBpedia	generated	Total	% DBpedia
linked Persons	121,834	530,257	652,091	18.68%
linked Organizations	59,032	814,088	873,120	6.76%
linked Locations	86,236	188,130	274,366	31.43%

Organizations	Number of Mentions	Persons	Number of Mentions	Locations	Number of Mentions
dbpedia:Toyota	130,783	dbpedia:Abraham_Lincoln	68,904	dbpedia:England	105,342
dbpedia:General_Motors	77,420	dbpedia:Andrew_Johnson	26,306	dbpedia:Japan	102,530
dbpedia:Ford_Brasil	69,222	dbpedia:Barack_Obama	19,761	dbpedia:United_States	98,872
dbpedia:Republican_Party_(United_States)	53,106	dbpedia:Presidency_of_Barack_Obama	16,037	dbpedia:Europe	97,154
dbpedia:Volvo	51,897	dbpedia:United_Nations	12,819	dbpedia:Taiwan	86,450
dbpedia:Democratic_Party_(United_States)	50,768	dbpedia:Zachary_Taylor	11,713	dbpedia:India	84,941
dbpedia:Volkswagen	48,748	dbpedia:George_W._Bush	10,891	dbpedia:Australia	72,135
dbpedia:Saab_Group	45,346	dbpedia:Gray_Davis	10,670	dbpedia:Germany	56,223
dbpedia:UEFA	40,692	dbpedia:Ron_Paul	10,140	dbpedia:Canada	51,164
dbpedia:Chrysler	39,572	dbpedia>List_of_Heroes_characters	9,968	dbpedia:Chinese_people	50,120
dbpedia:Renault	39,443	dbpedia:Gerry_Anderson	9,888	dbpedia:Americas	48,707
dbpedia:UD_Trucks	37,719	dbpedia:John_Howard	9,206	dbpedia:Russia	45,643
dbpedia:Porsche	32,863	dbpedia:Walter_Scott	9,198	dbpedia:Empire_of_Japan	36,517
dbpedia:Daimler_AG	30,555	dbpedia:George_Washington	8,814	dbpedia:France	36,024
dbpedia:Honda	29,650	dbpedia:Ralf_Schumacher	8,765	dbpedia:Billboard_200	30,834
dbpedia:Toyota_Motorsport_GmbH	28,513	dbpedia:Carlos_Ghosn	8,497	dbpedia:London	28,553
dbpedia:Opel	28,164	dbpedia:John_G._Robinson	7,866	dbpedia:China	27,342
dbpedia:Volkswagen_Group	26,953	dbpedia:Lewis_Hamilton	7,742	dbpedia:Nebraska	27,141
dbpedia:BMW	26,854	dbpedia:Trecia-Kaye_Smith	7,728	dbpedia:Detroit	26,778
dbpedia:Suzuki	25,260	dbpedia:Fernando_Alonso	7,702	dbpedia>New_York	26,555
dbpedia:Ford_Germany	25,030	dbpedia:John_Adams	7,091	dbpedia:Second_Spanish_Republic	25,175
dbpedia:Fiat	24,280	dbpedia:Robert_E._Lee	7,059	dbpedia:Spain	23,984
dbpedia:Jaguar_Cars	23,415	dbpedia:James_Baker	7,031	dbpedia:Malaysia	23,972
dbpedia:Toyota_Manufacturing_UK	22,996	dbpedia:Celtic_art	6,868	dbpedia:Tsardom_of_Russia	23,820
dbpedia:Proton_(automobile)	20,994	dbpedia:Don_Murphy	6,433	dbpedia:Italy	22,569
dbpedia:Mazda	20,438	dbpedia:David_Cameron	6,219	dbpedia:Chicago	21,880
dbpedia:Audi	20,330	dbpedia>List_of_Sons_of_Anarchy_characters	5,946	dbpedia:Michigan	21,785
dbpedia:Mercedes-Benz	19,510	dbpedia:Lyle_Campbell	5,911	dbpedia:Australia_national_women's_cricket_team	21,617
dbpedia:Airbus	18,481	dbpedia:Jock_Young	5,749	dbpedia:North_America	21,566

# Missing entities: NIL or Dark entities

- Dark Entities: domain entities may not be present

## Ford Names Paul Bellew It's First Chief Data and Analytics Officer



By [andreaharrison](#)

[FOLLOW](#)

| 22 December 2014

| [f Like](#) 0



Ford Motor Co.'s hiring last week of a new senior officer illustrates the auto industry's accelerating dive into big data.

[Ford taps for General Motors exec to help it find its way in big data](#)

Automotive News reports that the company named Paul Ballew, a former top sales analyst at General Motors, to the new post of chief data and analytics officer. Mr. Ballew, 50, will oversee the automaker's global analytics efforts

# Missing entities: NIL or Dark entities

- Dark Entities: domain entities may not be present



**WIKIPEDIA**  
The Free Encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikipedia store

Interaction  
[Help](#)  
[About Wikipedia](#)  
[Community portal](#)  
[Recent changes](#)  
[Contact page](#)

Tools  
[Upload file](#)  
[Special pages](#)  
[Printable version](#)

Languages

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

Special page

## Search results

Results 1 - 20 of 125

[Content pages](#) [Multimedia](#) [Everything](#) [Advanced](#)

**Did you mean: paul *bailey***

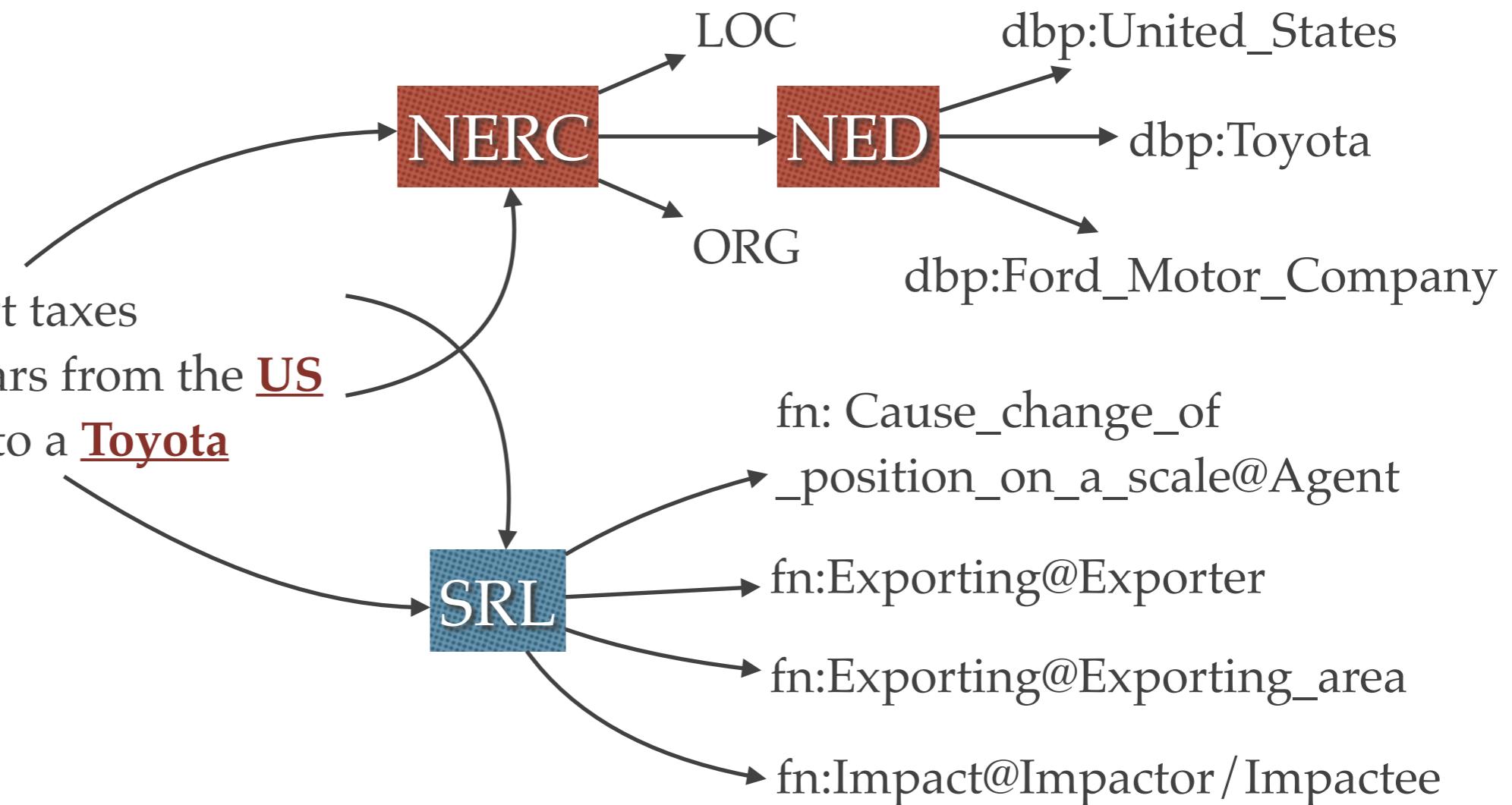
*The page "**Paul ballew**" does not exist. You can [ask for it to be created](#), but consider checking the search results below to see whether the topic is already covered.*

[Kurt Knoff](#) (category People from the Minneapolis–Saint Paul metropolitan area)  
post-football career as a commercial real estate broker in Minneapolis – Saint Paul  
**Ballew, Bill.** Tough Enough to be Vikings: Minnesota's Purple Pride from A to  
**5 KB (503 words) - 15:10, 12 January 2016**

**Billy Ballew Motorsports**  
Billy Ballew Motorsports is a team that competes in the NASCAR Camping World Truck Series. They were formed in 1996 by Georgia businessman **Billy Ballew**.

# Identity crisis

- The US raises import taxes
- Ford exports most cars from the US
- The Ford crashed into a Toyota



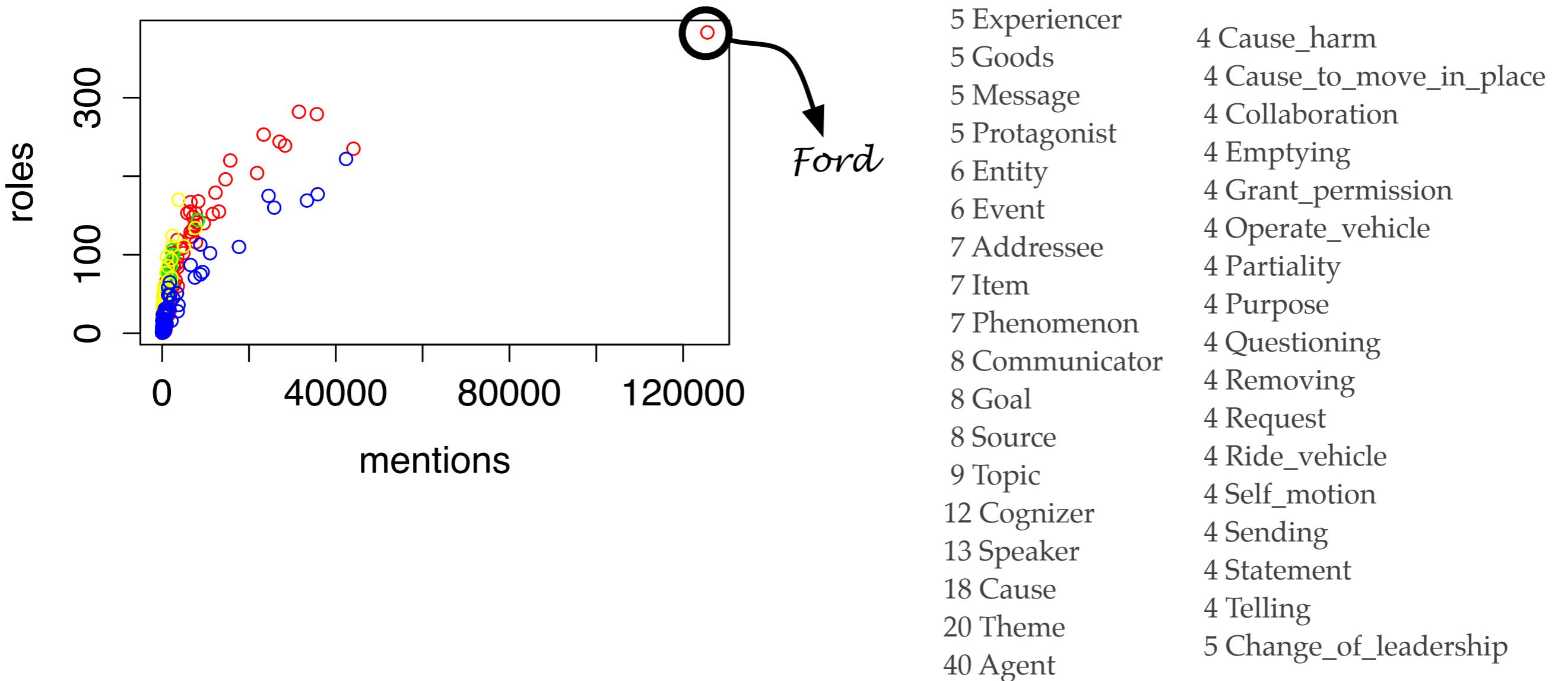
# Identity crisis

**Table 2.** Occurrences as Actor or Places per Class

Class	Actor (total)	Place (total)	Actor (%)	Standard Deviation
Country	152,708	132,341	53,57%	40.93%
Person	106,318	9,447	91.84%	21.56%
Company	609,971	84,378	87.85%	29.63%
Motor Company	431,619	63,917	87,10%	28.46%
Automobile	116,893	5,804	95.27%	14.45%

Based on 63K news articles on the automotive industry

# 383 Roles of Ford



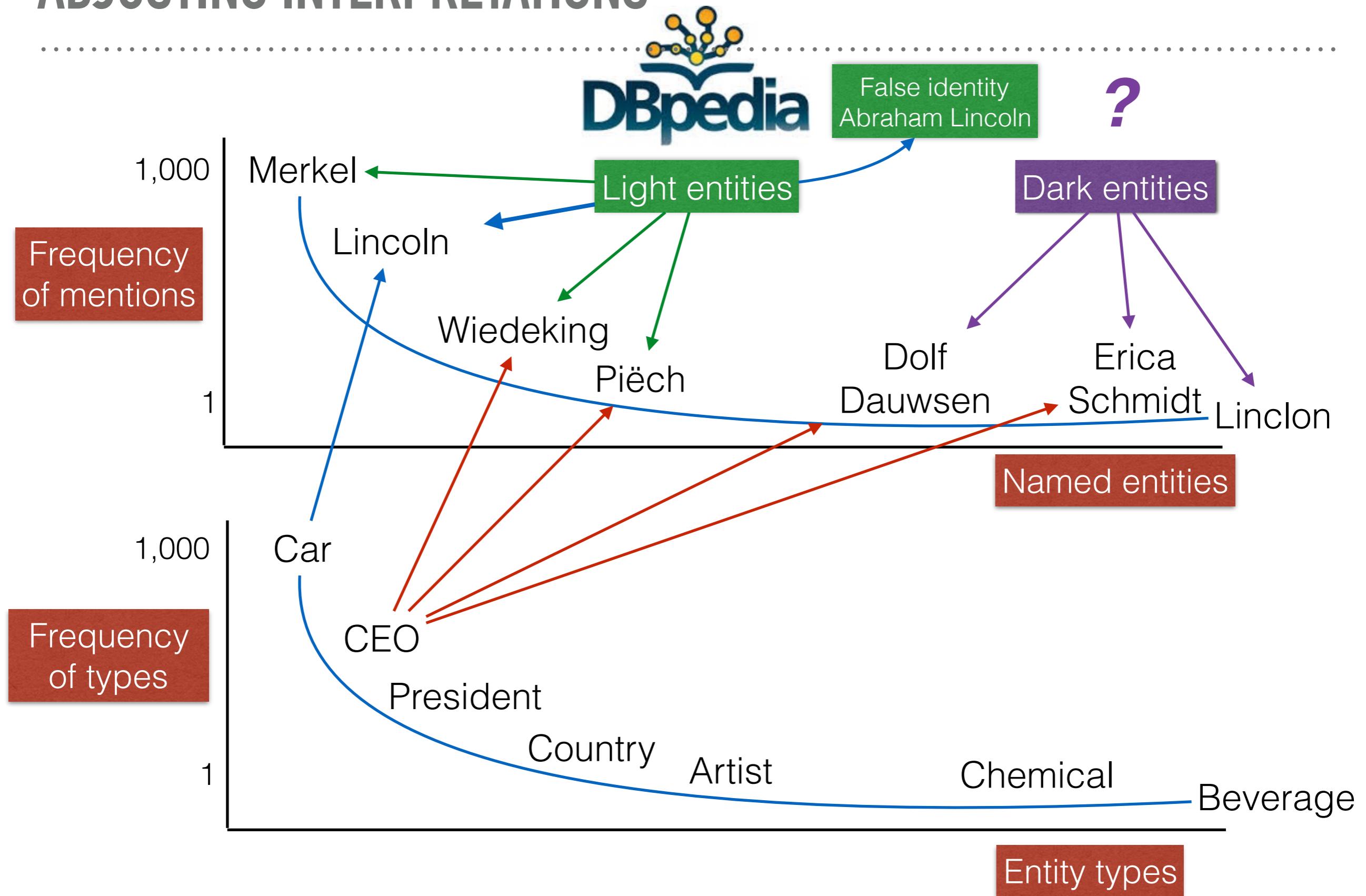
# Entities resolved?

- 36% dark entities in news
- Many disambiguation errors
- Many dark entities linked to light entities
- Within the same document
- Within the same domain
- Why link Lincoln to the president in the context of automotive industry?

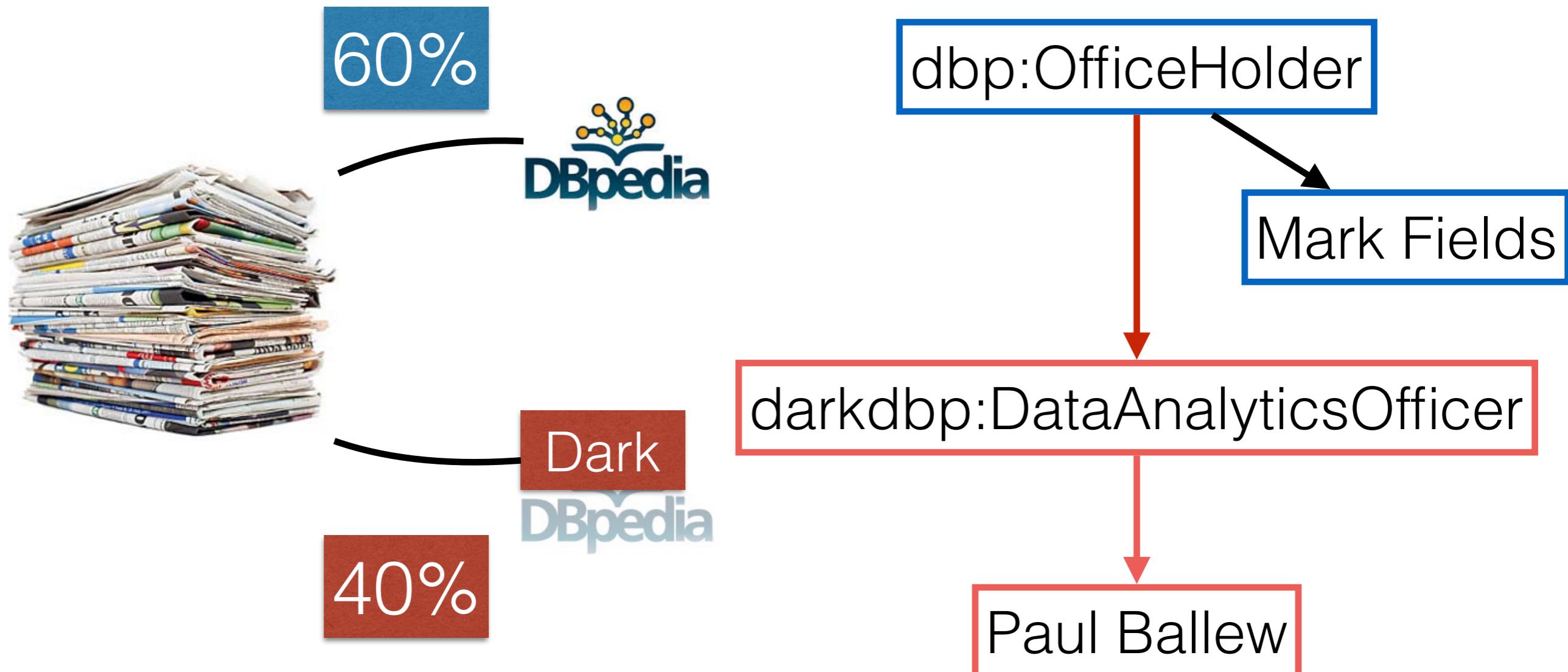
dominant domain resolver:  
Peter  
Schreyer  
Peter Schreyer

relevant time resolver  
Abraham Lincoln's  
relevance for 2003-2015

# ADJUSTING INTERPRETATIONS

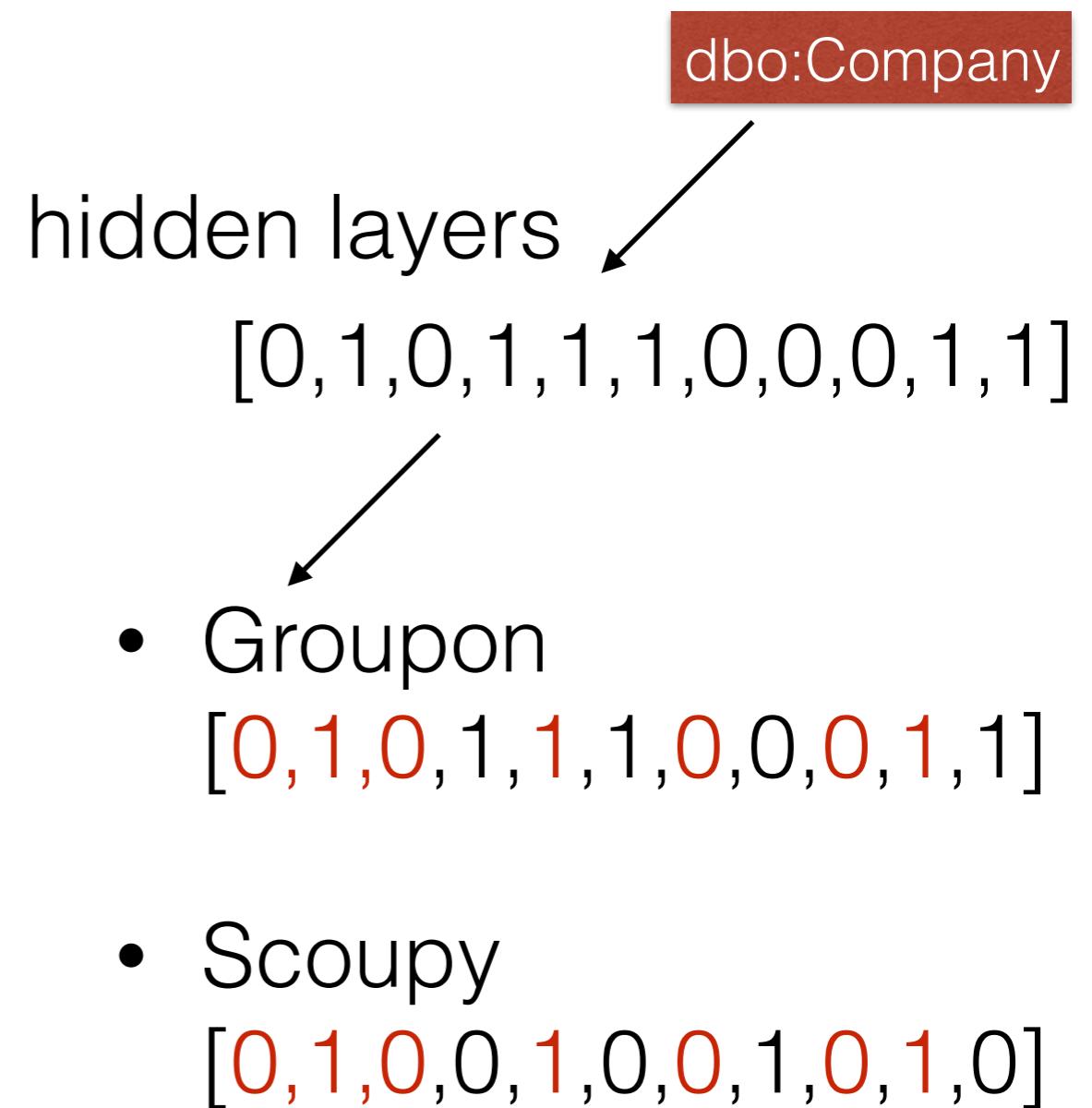


# Learning about the world



# Deep learning

- Light entity: Groupon (dbo:Company)
  - **Met de korting**sbonnen van Groupon kunt u tegen hoge **korting kennismaken** met de diensten van een professionele fotograaf.
  - **Met de** Groupon **app** kan je nu ook onderweg de deals bekijken, kopen en inwisselen.
- Dark Entity: Scoupy
  - Via de gratis **app** voor iPhone of Android of via de website [www.scoupy.nl](http://www.scoupy.nl) kun je met **hoge korting** of zelfs helemaal gratis **kennismaken** met allerlei producten.
  - **Met de** Scoupy **app** kan je op zoek naar **korting**scoupons voor winkels en restaurants bij jou in de buurt.



# Further reading

- Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011, September). *DBpedia spotlight: shedding light on the web of documents*. In *Proceedings of the 7th international conference on semantic systems* (pp. 1-8). ACM.
- Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., ... & Weikum, G. (2011, July). *Robust disambiguation of named entities in text*. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 782-792). Association for Computational Linguistics.
- Ling, X., Singh, S., & Weld, D. S. (2015). *Design challenges for entity linking*. *Transactions of the Association for Computational Linguistics*, 3, 315-328.
- Van Erp, M., Mendes, P., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., Waitelonis, J. (2016). *Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job*. LREC