



# 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...

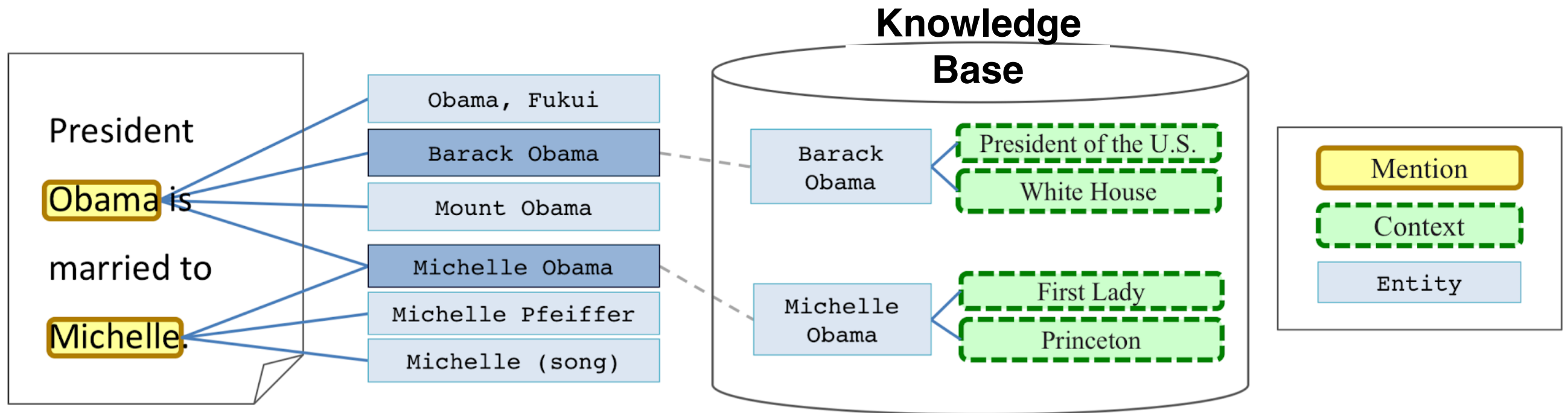
# Task definition

- 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



# NERC + NED = NERD

- Typically it is assumed that the mentions are already detected in text
- Sometimes we do both recognition + disambiguation/linking -> NERD

# 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

[https://en.wikipedia.org/wiki/Mount\\_Everest](https://en.wikipedia.org/wiki/Mount_Everest)

From Wikipedia, the free encyclopedia

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

Coordinates:  27°59'17"N 86°55'31"E

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) <sup>(*Learn how and when to remove this template message*)</sup>

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>

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 [Kalapatthar](#).

**Highest point**

<b>Elevation</b>	8,848 metres (29,029 ft) <sup>[1]</sup>
	<b>Ranked 1st</b>

**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

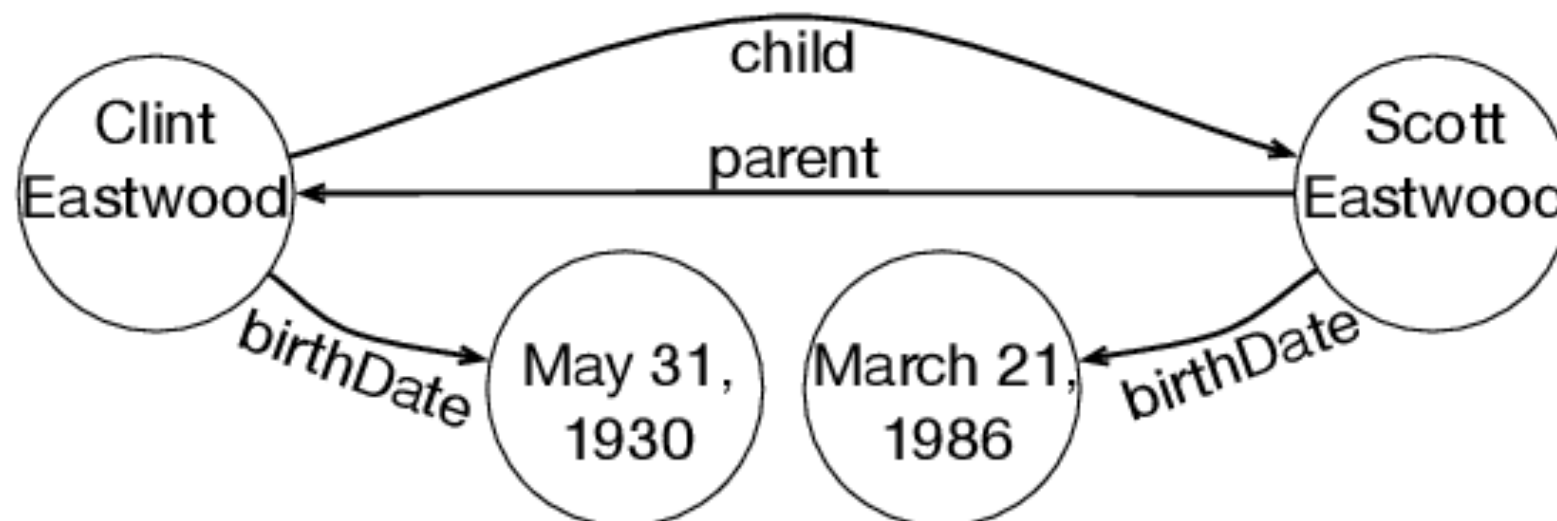
# Structured knowledge bases: DBpedia & Wikidata

dbpedia.org/page/Mount_Everest	
 Browse using  Formats 	
geo:geometry	▪ POINT(86.925277709961 27.988056182861)
geo:lat	▪ 27.988056 (xsd:float)
geo:long	▪ 86.925278 (xsd:float)
prov:wasDerivedFrom	▪ <a href="https://en.wikipedia.org/wiki/Mount_Everest?oldid=744845387">wikipedia-en:Mount_Everest?oldid=744845387</a>
foaf:depiction	▪ <a href="https://commons.wikimedia.org/wiki/File:Mount-Everest.jpg">wiki-commons:Special:FilePath/Mount-Everest.jpg</a>
foaf:isPrimaryTopicOf	▪ <a href="https://en.wikipedia.org/wiki/Mount_Everest">wikipedia-en:Mount_Everest</a>
foaf:name	▪ Mount Everest (en)
is dbo:deathPlace of	▪ <a href="#">dbr:Shailendra_Kumar_Upadhyaya</a> ▪ <a href="#">dbr:Mick_Burke_(mountaineer)</a> ▪ <a href="#">dbr:Ray_Genet</a> ▪ <a href="#">dbr:Scott_Fischer</a> ▪ <a href="#">dbr:Karl_Gordon_Henize</a> ▪ <a href="#">dbr:Hristo_Prodanov</a> ▪ <a href="#">dbr:David_Sharp_(mountaineer)</a> ▪ <a href="#">dbr:Rob_Hall</a> ▪ <a href="#">dbr:Pasang_Lhamu_Sherpa</a> ▪ <a href="#">dbr:Zygmunt_Andrzej_Heinrich</a> ▪ <a href="#">dbr:Mohammad_Khaled_Hossain</a> ▪ <a href="#">dbr:Andrew_Irvine_(mountaineer)</a> ▪ <a href="#">dbr:Maurice_Wilson</a>

https://www.wikidata.org/wiki/Q513	
instance of	<div><div></div>mountain<div> 0 references</div></div>
part of	<div><div></div>Seven Summits<div> 1 reference</div></div> <div><div></div>Himalayas<div> 1 reference</div></div>
image	<div><div></div><div>Mount Everest from Rongbuk may 2005.JPG 3,008 × 2,000; 1.12 MB<div> 0 references</div></div></div>

# Structured KBs are essentially graph networks

... with billions (!) of such facts



# Reflection

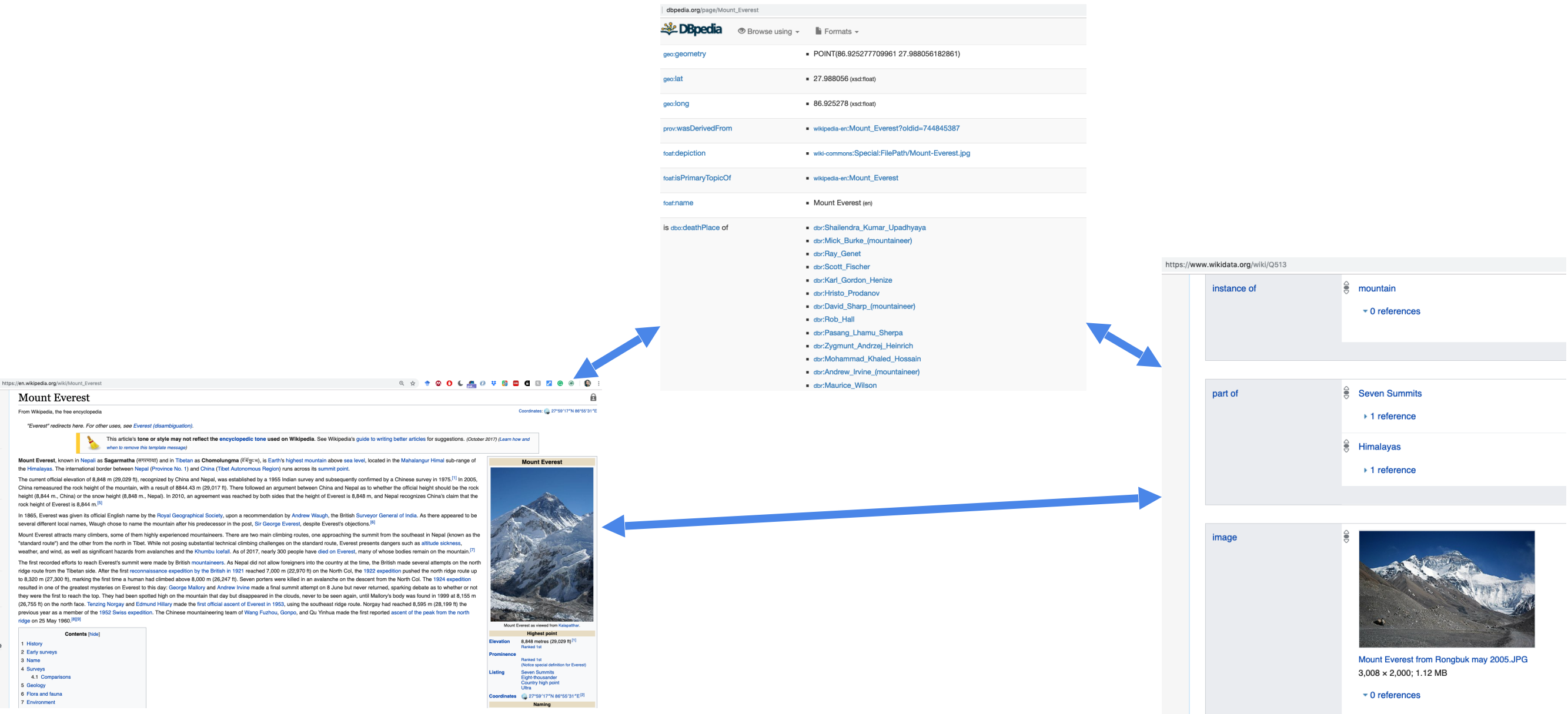
Q: What is the value we could get from:

- unstructured knowledge bases?
- structured knowledge bases?

Q: Let's say you want to do entity linking and you can choose a KB, which one would you pick?

In practice, depends on the approach.

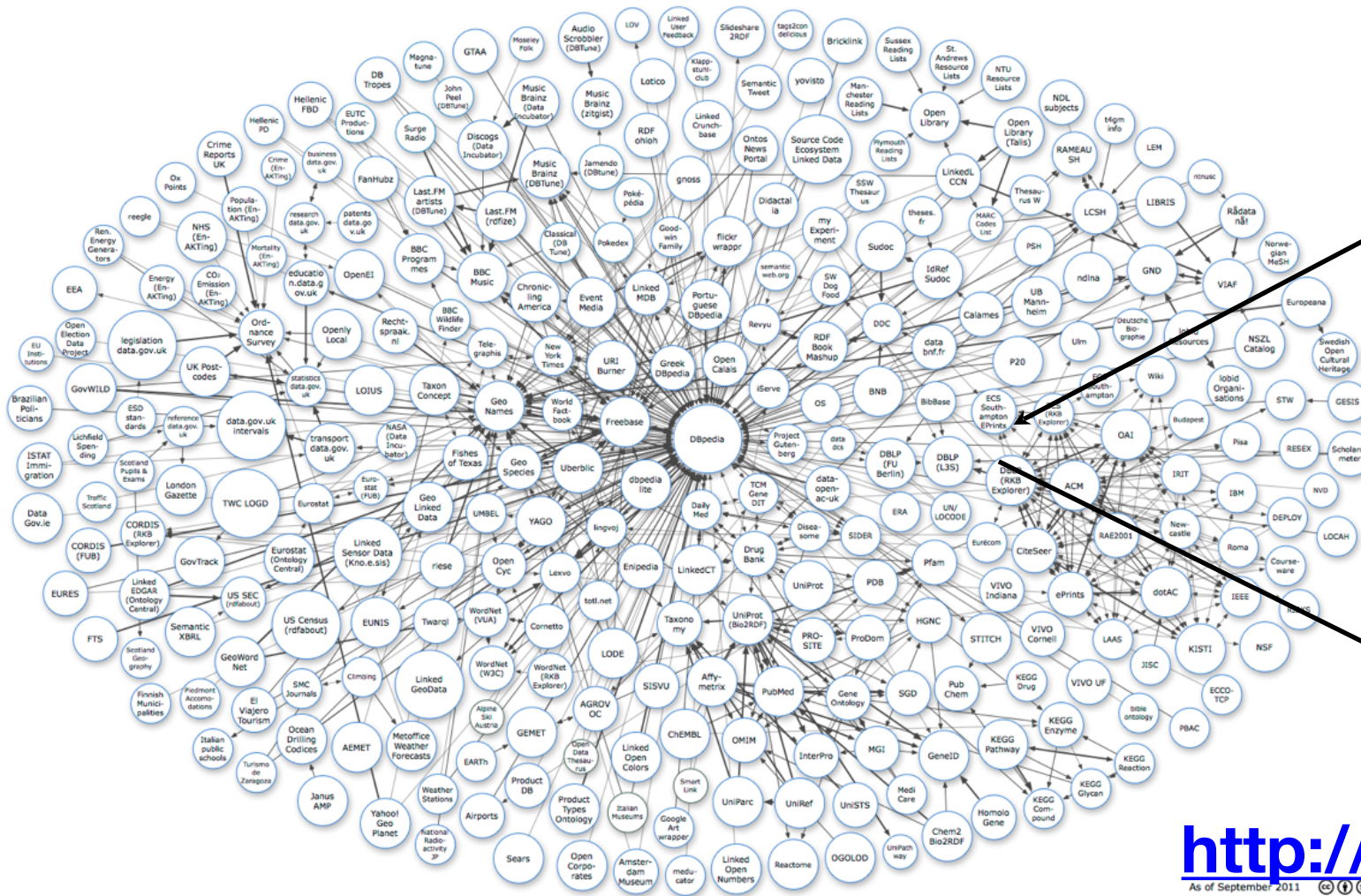
# Knowledge bases are also connected to each other





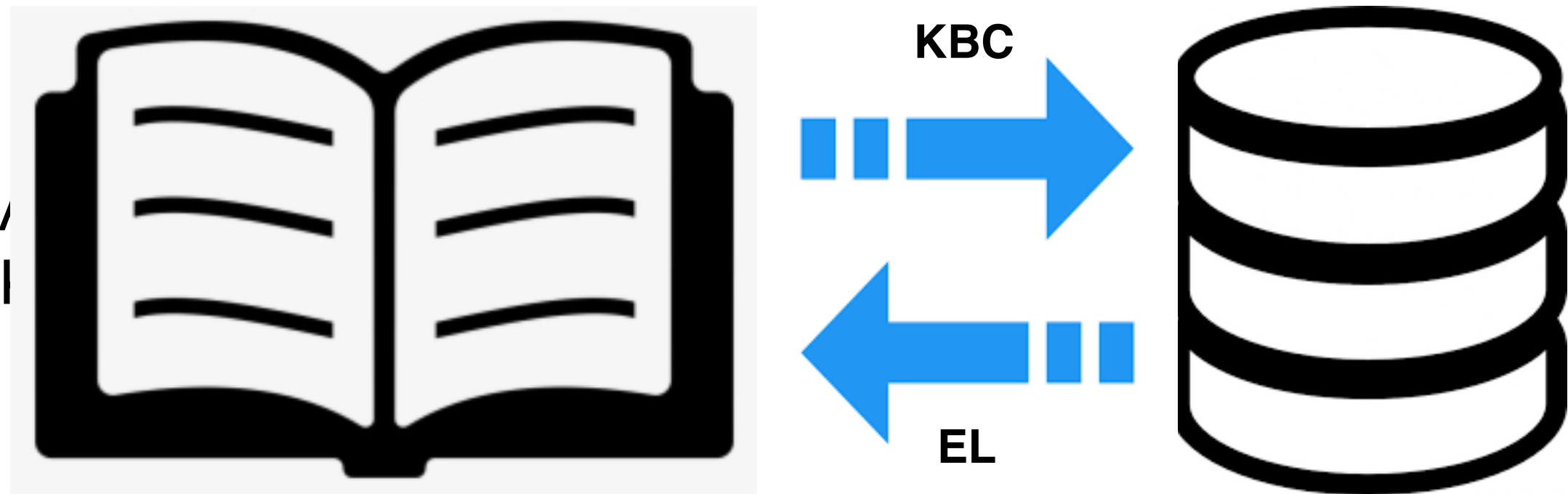
**connected ->  
the LOD Cloud**

# “Abraham Lincoln”



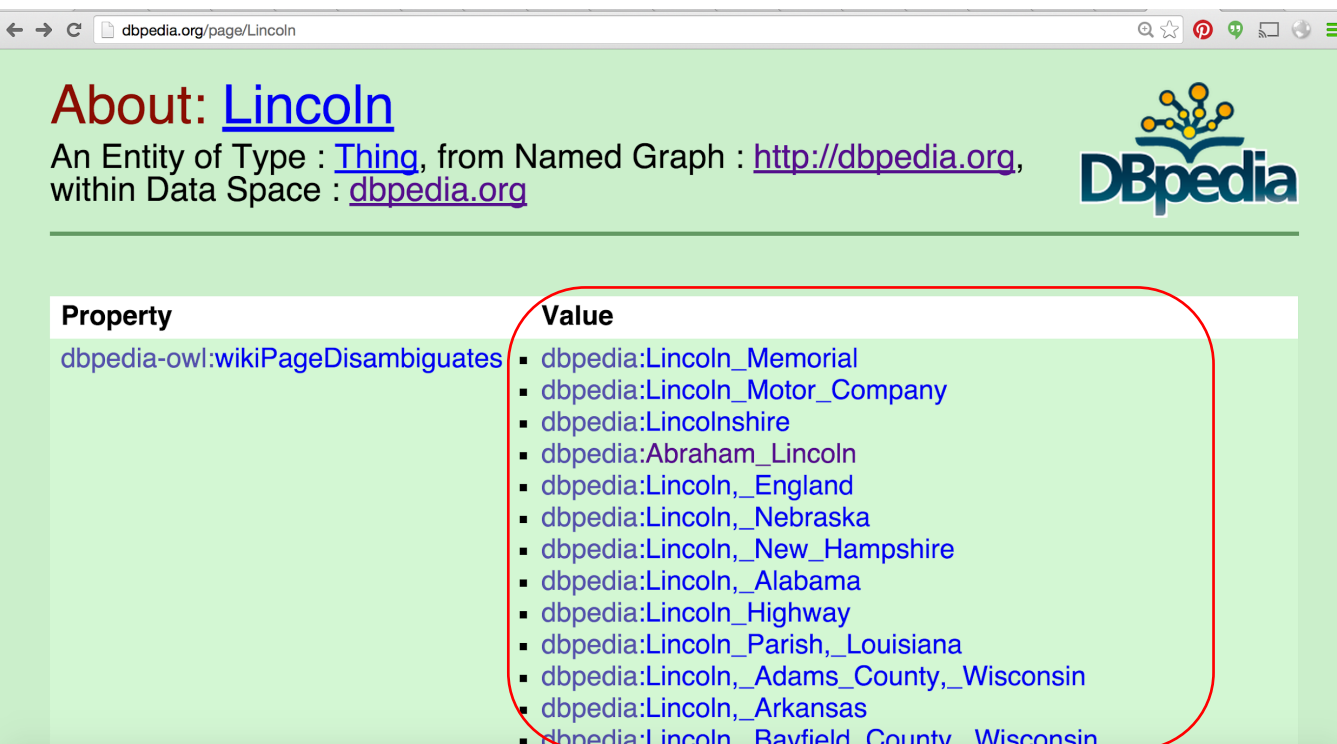
[http://dbpedia.org/resource/Abraham\\_Lincoln](http://dbpedia.org/resource/Abraham_Lincoln)

# Other benefits of connecting text and knowledge bases





# Named Entity Disambiguation isn't that easy though



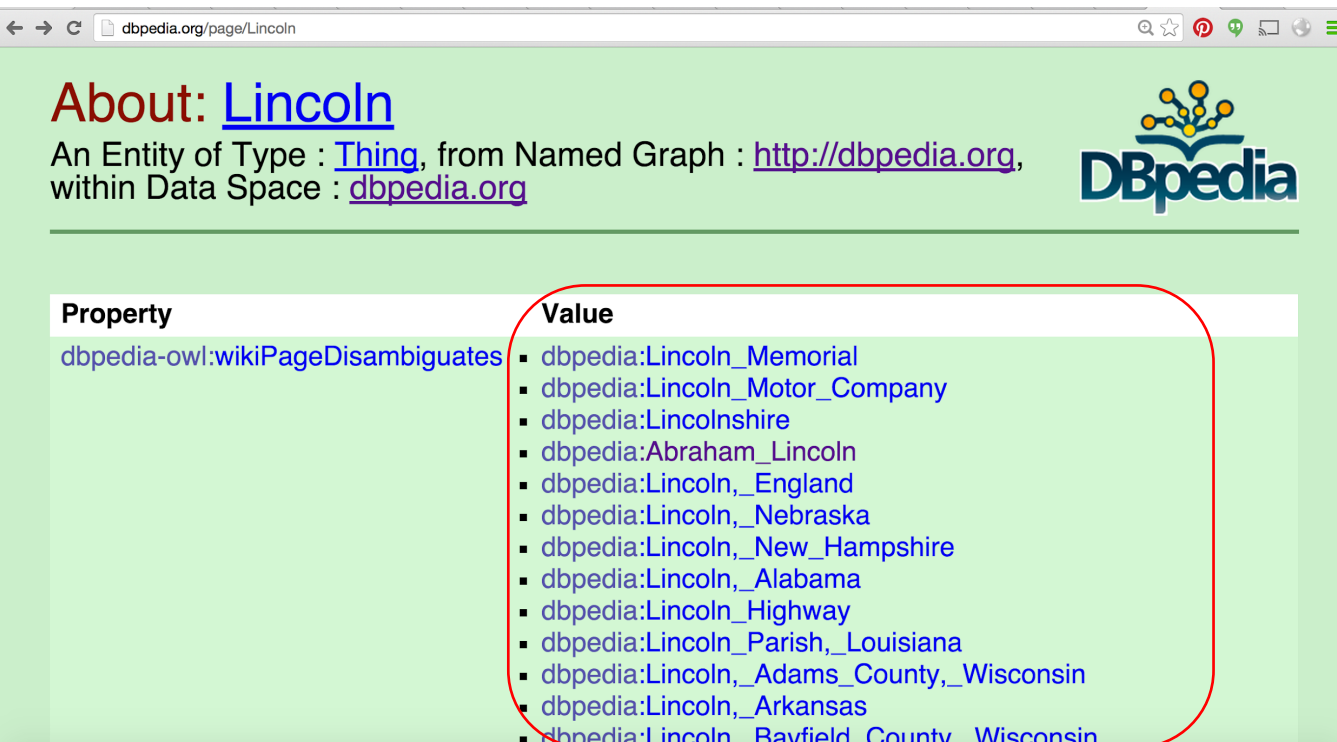
The screenshot shows a web browser window with the address bar displaying 'dbpedia.org/page/Lincoln'. The page title is 'About: Lincoln'. Below the title, it says 'An Entity of Type : [Thing](#), from Named Graph : [http://dbpedia.org](\"http://dbpedia.org\"), within Data Space : [dbpedia.org](\"dbpedia.org\")'. The DBpedia logo is in the top right corner. The main content area is a table with two columns: 'Property' and 'Value'. The 'Property' column contains 'dbpedia-owl:wikiPageDisambiguates'. The 'Value' column contains a list of 14 entities, each preceded by a bullet point and a DBpedia URI. A red rounded rectangle highlights the list of values.

Property	Value
dbpedia-owl:wikiPageDisambiguates	<ul style="list-style-type: none"><li>▪ <a href="#">dbpedia:Lincoln_Memorial</a></li><li>▪ <a href="#">dbpedia:Lincoln_Motor_Company</a></li><li>▪ <a href="#">dbpedia:Lincolnshire</a></li><li>▪ <a href="#">dbpedia:Abraham_Lincoln</a></li><li>▪ <a href="#">dbpedia:Lincoln,_England</a></li><li>▪ <a href="#">dbpedia:Lincoln,_Nebraska</a></li><li>▪ <a href="#">dbpedia:Lincoln,_New_Hampshire</a></li><li>▪ <a href="#">dbpedia:Lincoln,_Alabama</a></li><li>▪ <a href="#">dbpedia:Lincoln_Highway</a></li><li>▪ <a href="#">dbpedia:Lincoln_Parish,_Louisiana</a></li><li>▪ <a href="#">dbpedia:Lincoln,_Adams_County,_Wisconsin</a></li><li>▪ <a href="#">dbpedia:Lincoln,_Arkansas</a></li><li>▪ <a href="#">dbpedia:Lincoln,_Bayfield_County,_Wisconsin</a></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 the DBpedia page for 'Lincoln'. The page title is 'About: Lincoln'. Below the title, it says 'An Entity of Type : [Thing](#), from Named Graph : [http://dbpedia.org](\"http://dbpedia.org\"), within Data Space : [dbpedia.org](\"dbpedia.org\")'. The DBpedia logo is in the top right. Below this, there is a table with two columns: 'Property' and 'Value'. The 'Property' column contains 'dbpedia-owl:wikiPageDisambiguates'. The 'Value' column contains a list of entities, each preceded by a bullet point and a DBpedia URI. 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. A red rounded rectangle highlights the entire list of values.

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>

## Abraham Lincoln (Q91)

16th President of the United States

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

### a. name **ambiguity**

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

### b. name variation

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

# 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.

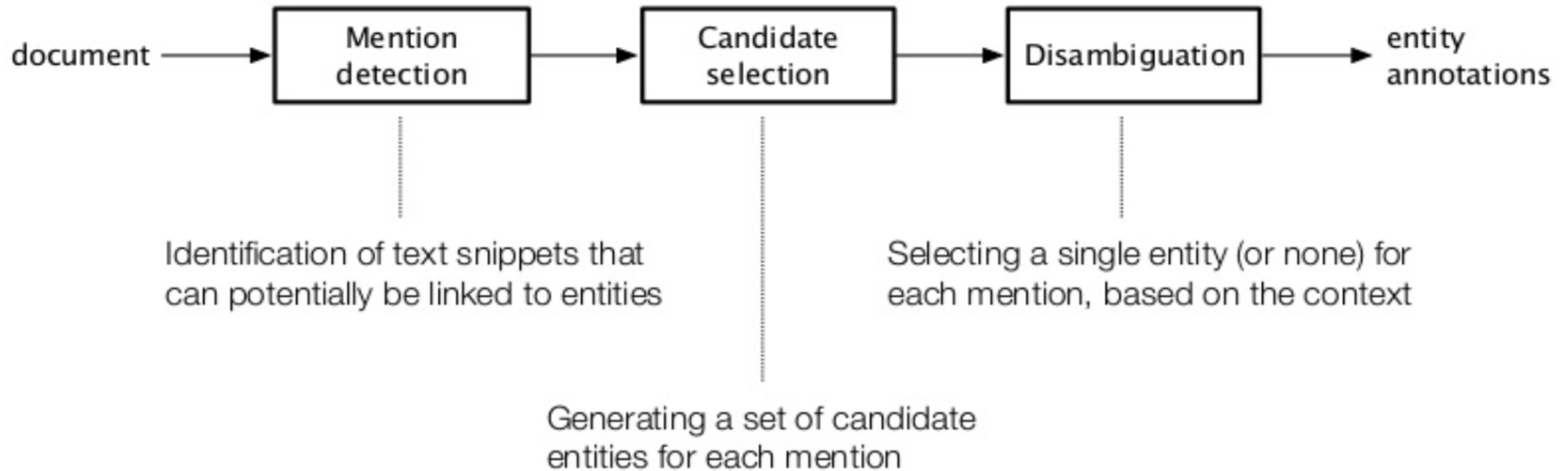
If you have any information, contact the **San Benito Police Department** at 361-3880.

## What to link “Edgar Gonzalez” to???

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

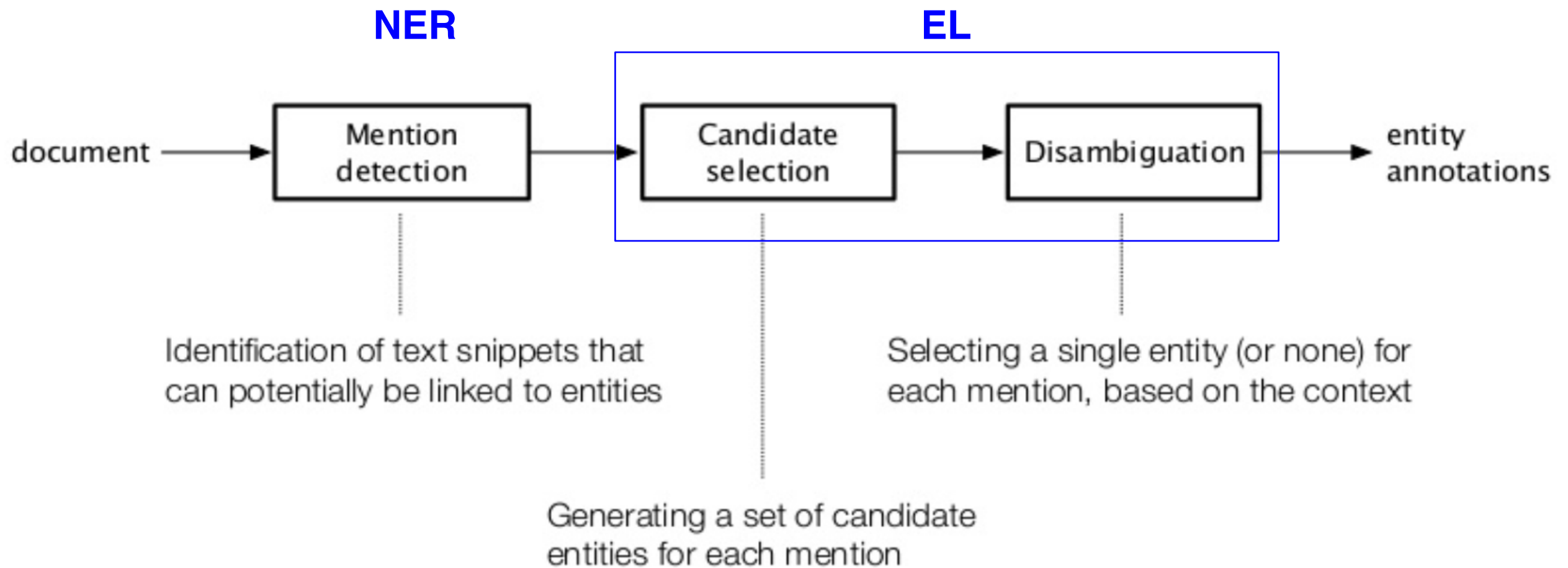
## **2. Components of an entity linking system**

# 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 recognized mentions in text, get the potential referents (instances) in a knowledge base, following the “closed world assumption”.
- 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.

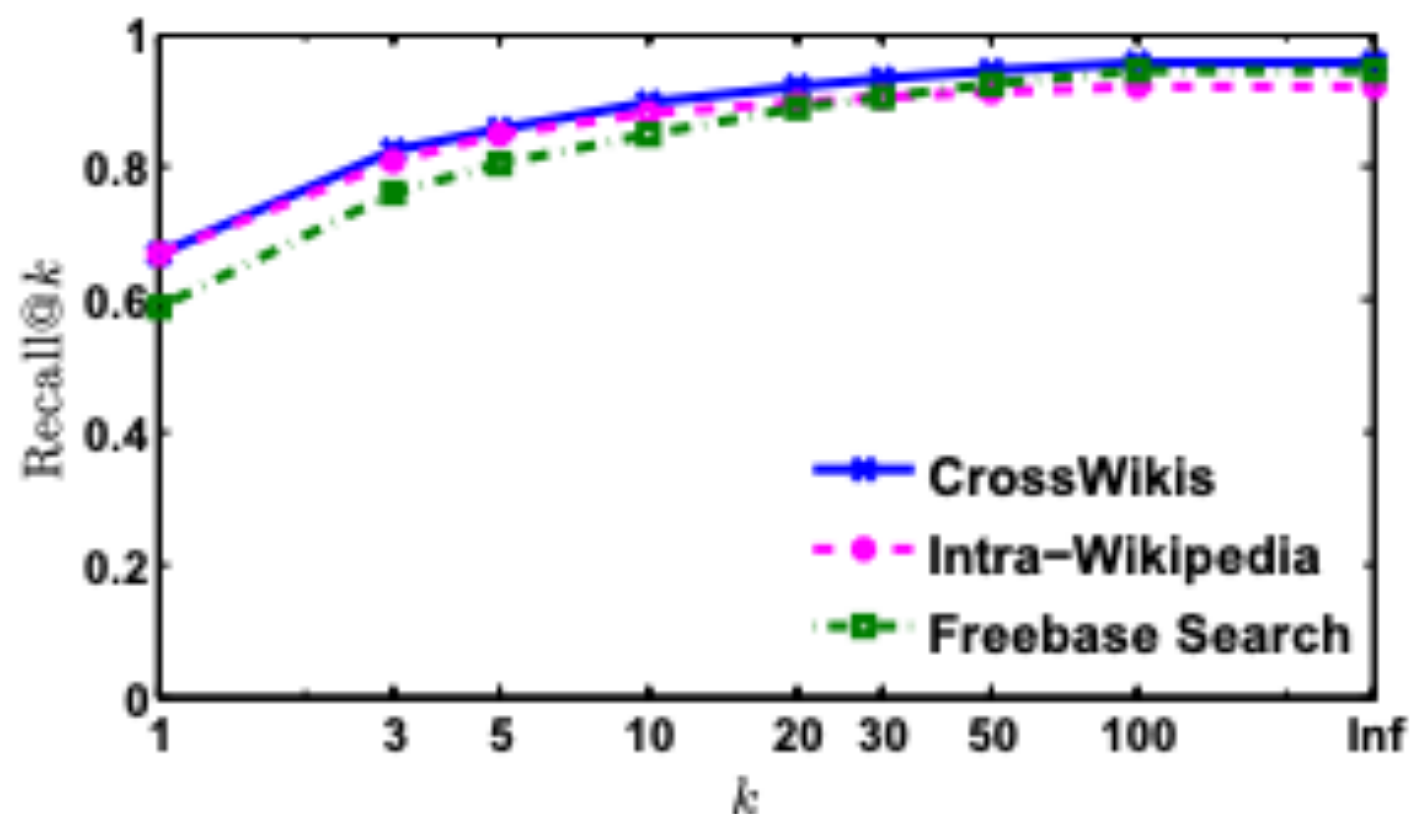


Figure 3: Recall@ $k$  on an aggregate of nine data sets, comparing three **candidate generation** methods.

Source: [https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacI\\_a\\_00141](https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacI_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

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

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
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

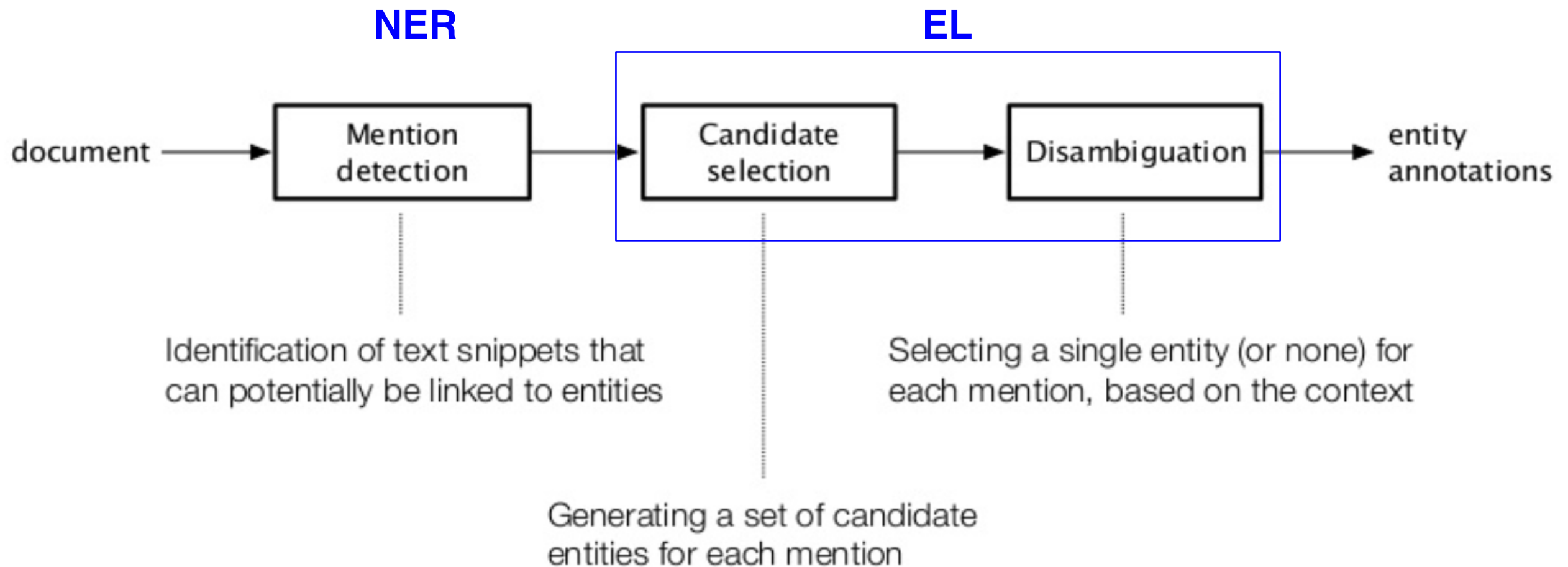
# Phase III: Disambiguation

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

	Word-based	Graph-based
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 TF/IDF 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>DBpedia Spotlight</u>	<u>AIDA/AGDISTIS</u>

# 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.



Confidence:  0.0

Contextual score:  0.0

Prominence (support):  0

No 'common words'

Default Disambiguation

Show best candidate

**SELECT TYPES...** **ANNOTATE**

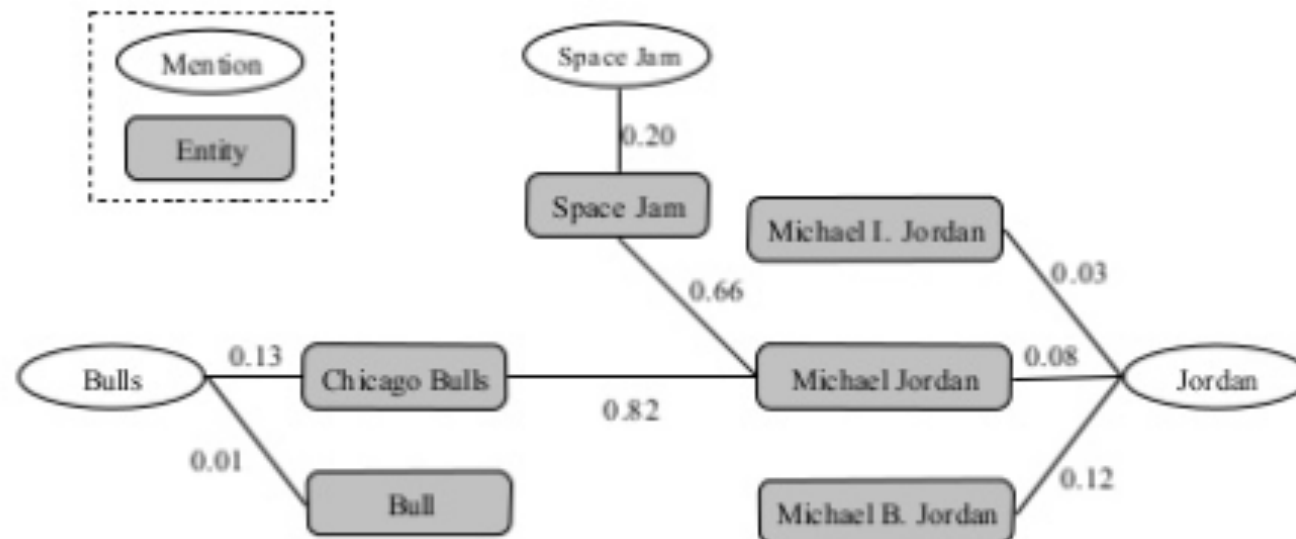
Ryland Peter "Ry" Cooder (born [March 15, 1947](#)) is an [American](#) guitarist, [singer](#) and [composer](#). He is known for his [slide guitar](#) work, his interest in [roots music](#) from the [United States](#), and, more recently, his collaborations with traditional musicians from many [countries](#).  
[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, [Freddie 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

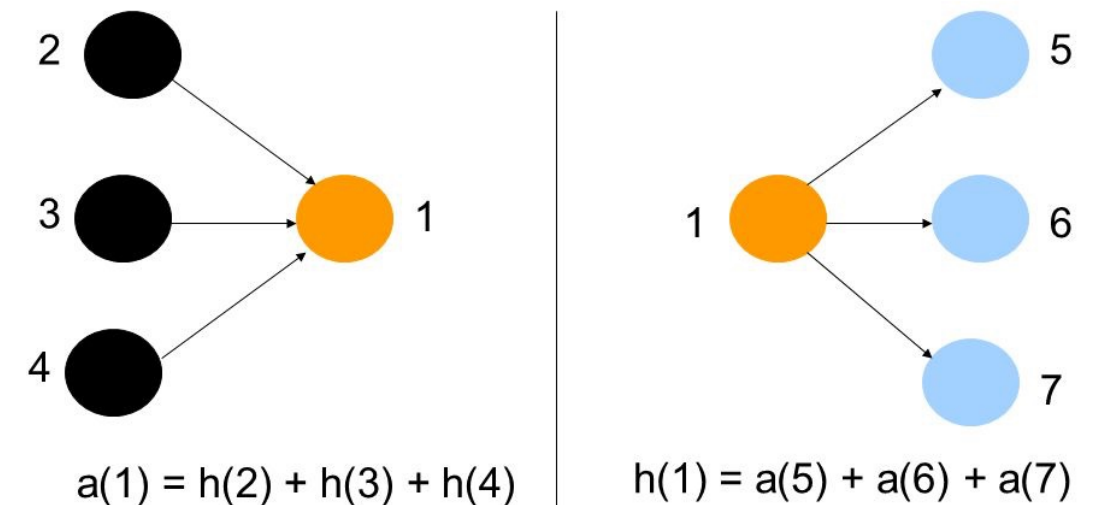
1. Construct a subgraph that contains all entity candidates with some facts from a KB.
2. Find the best connected candidates per mention:

## 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

	Chiefs	Alex Smith	Washington	Smith
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

	Chiefs	Alex Smith	Washington	Smith
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 ...”*

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

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

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

# 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*