Utilizing Knowledge Graphs for Text-centric Information Retrieval

Laura Dietz (@lauradietz99)
University of New Hampshire

Alexander Kotov (@rusillini)
Wayne State University

Edgar Meij (@edgarmeij)
Bloomberg

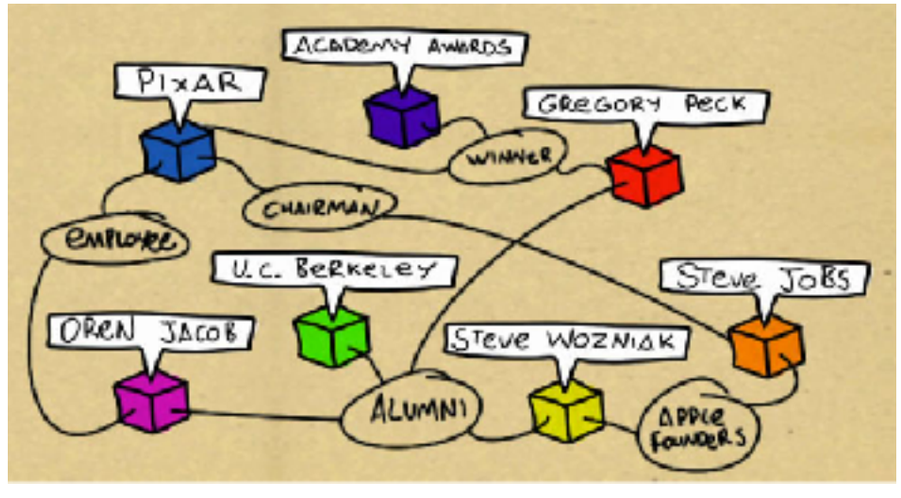
Slides at http://github.com/laura-dietz/tutorial-utilizing-kg
Bibliography at https://www.mendeley.com/community/entity-linking-and-retrieval/

Entity?

- Uniquely identifiable *thing* or *object*
  - “A thing with a distinct and independent existence”
  - people, places, products, companies, etc. etc.

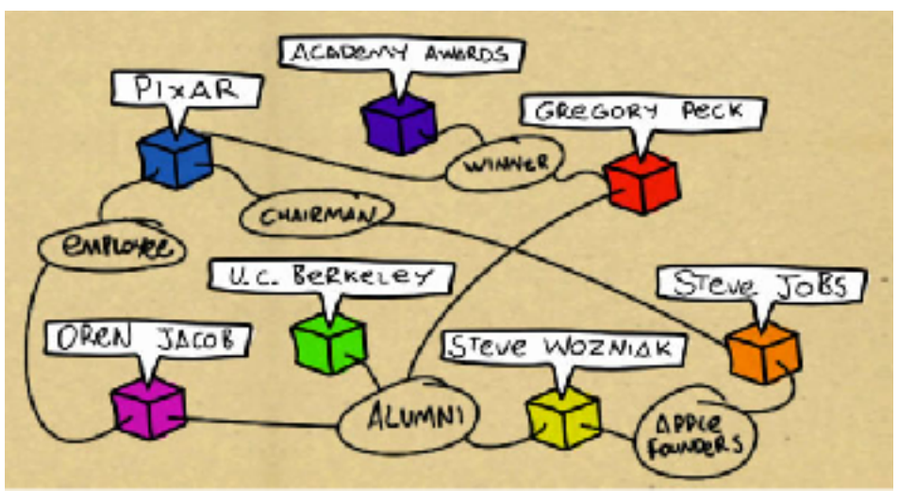
What's so special about entities?

- ID
- Name(s)
- Type(s)
- Attributes (/Descriptions)
- Relationships to other entities



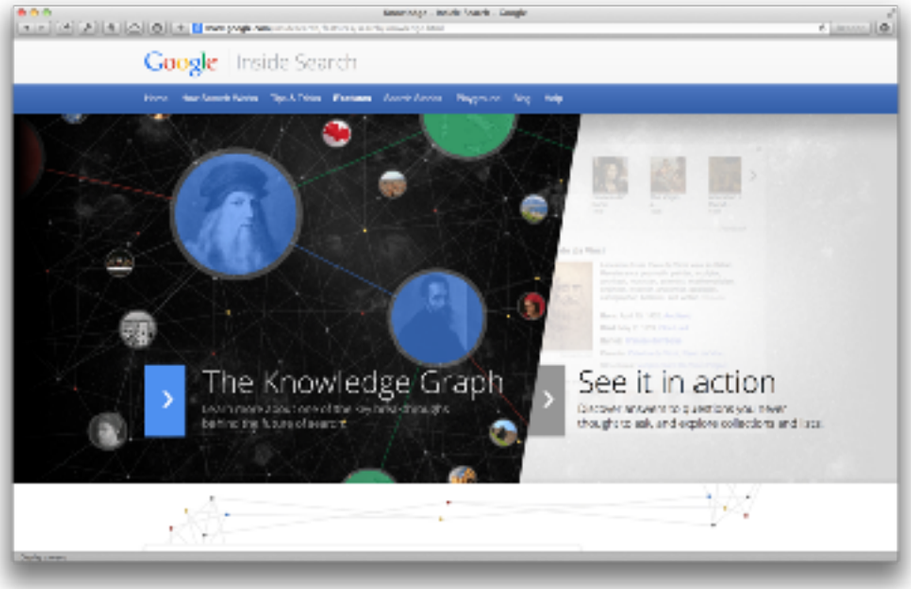
Knowledge graphs

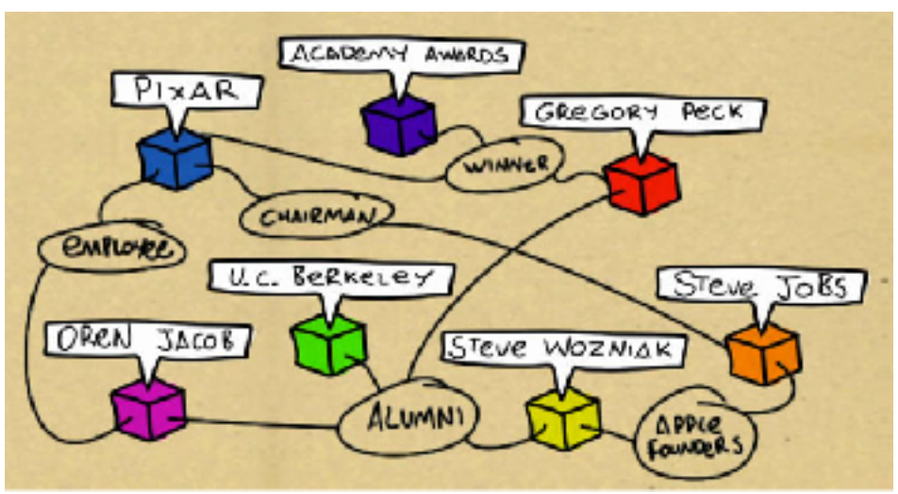
- A knowledge graph (KG) is a repository of entities and their relations and attributes that is represented as a graph



Knowledge graphs

- A KG defines:
  - entities
  - attributes
  - types
  - relations
  - (provenance, sometimes)
  - and more
    - external links
    - homepages
    - features
    - etc.





Knowledge graphs

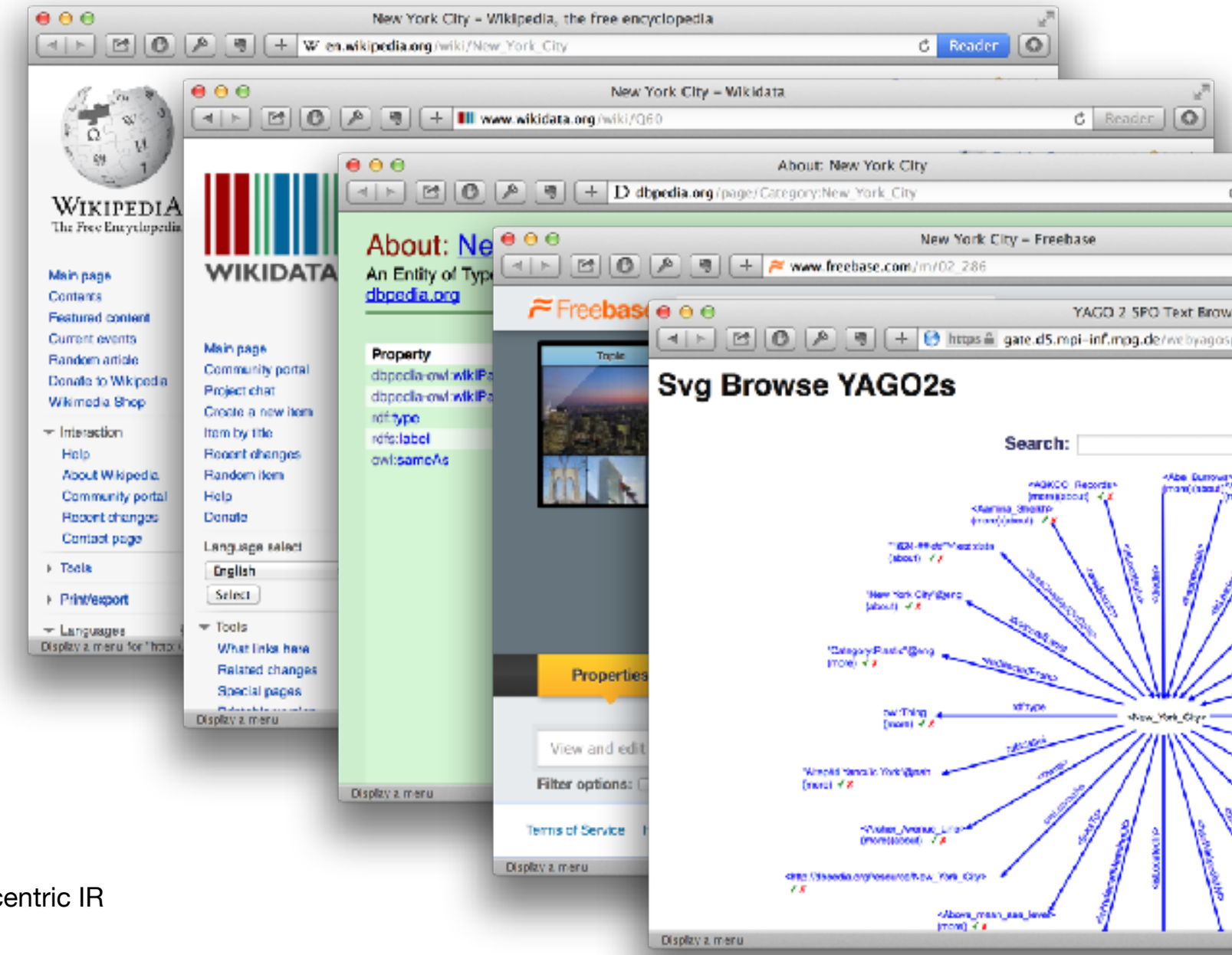
dbpedia:Ann_Arbor,_Michigan

rdfs:label	Ann Arbor, Michigan (en)
rdfs:comment	Ann Arbor is a city in the U.S. state of Michigan and the county seat of Washtenaw County. The 2010 census recorded its population to be 113,934, making it […]
dbpedia-prop:areaCode	734
dbpedia-prop:populationTotal	113934 (xsd:integer)
dbpedia-prop:aprHighF	59.600000 (xsd:double)
dbpedia-prop:isPartOf	**dbpedia:Washtenaw_County,_Michigan**
dbpedia-prop:timezone	**dbpedia:Eastern_Daylight_Time**
owl:sameAs	**nyt:Ann_Arbor,_Michigan** **freebase:Ann_Arbor,_Michigan** **[…]**
rdf:type	**dbpedia-owl:City** **wikidata:Q515** **[…]**
is **dbpedia-owl:birthPlace** of	**dbpedia:Eric_Betzig** **[…]**

[…]

Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO



Freebase

- Initially seeded from high-quality open data
  - then maintained mainly by community
- Harvested from many sources
  - Wikipedia, MusicBrainz, and others.
- Acquired by Google in 2010 (GKG)
  - now in read-only mode
- See http://www.freebase.com/

DBpedia

- Extract structured information from Wikipedia
  - infoboxes, categories, and more
  - crowd-sourced community effort
- Open source
  - written in Scala, Java and VSP
  - Virtuoso Universal Server Operating system
- See http://dbpedia.org/About

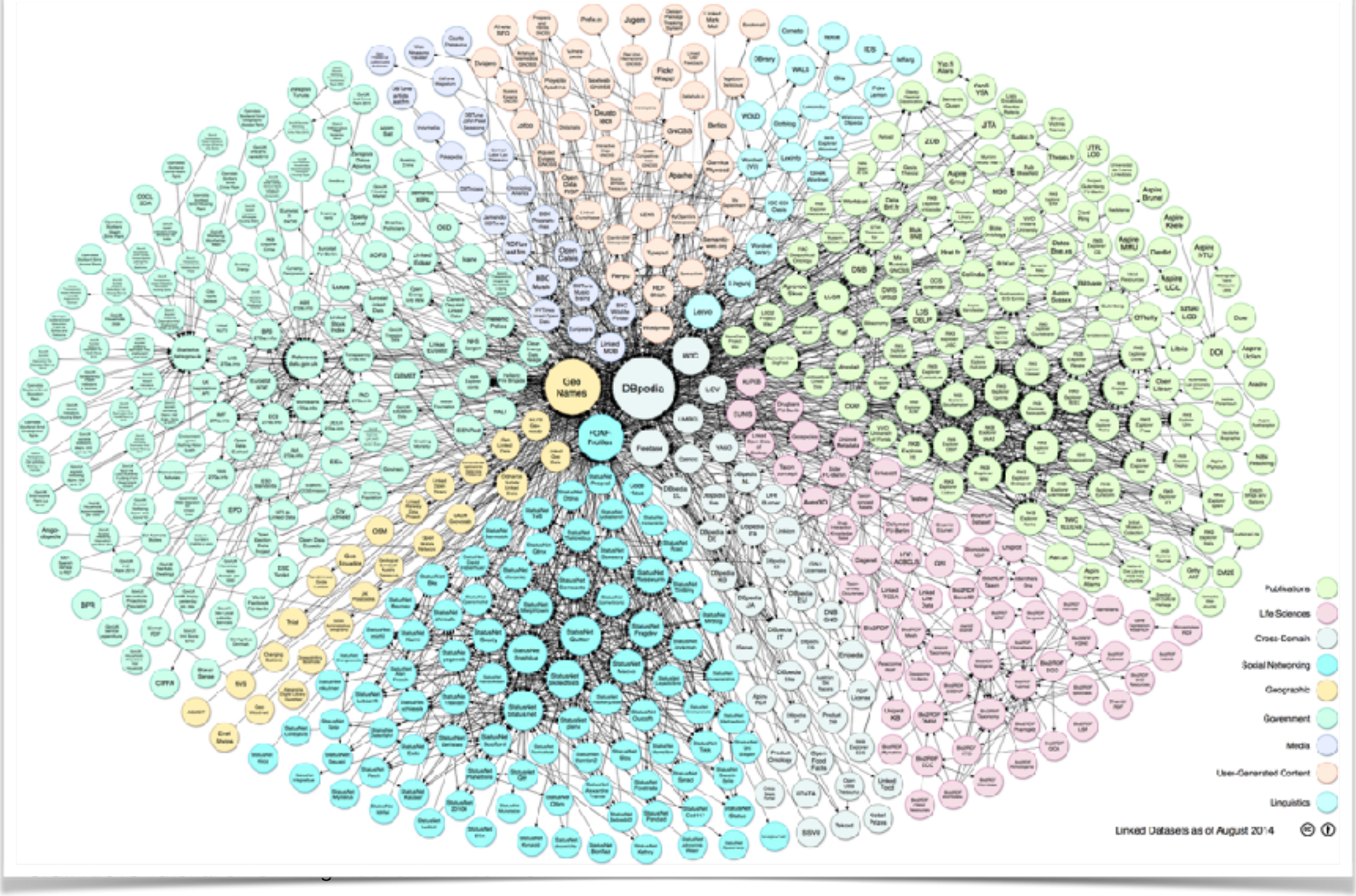
Wikidata

- Structured information from Wikipedia
  - infoboxes, categories, interlanguage links, and more
  - crowd-sourced community effort
    - “As of December 2015, according to Wikimedia statistics, half of the information in Wikidata is unsourced, another 30% is labeled as having come from Wikipedia.”
- See http://wikidata.org/

YAGO

- Accuracy manually evaluated
  - confirmed accuracy of 95%
  - relations annotated with confidence values
- Anchored in Time and Space
  - Thematic domains (e.g. "music" or "science")
- Based on wikipedia, includes WordNet, and GeoNames
- See http://www.mpi-inf.mpg.de/yago-naga

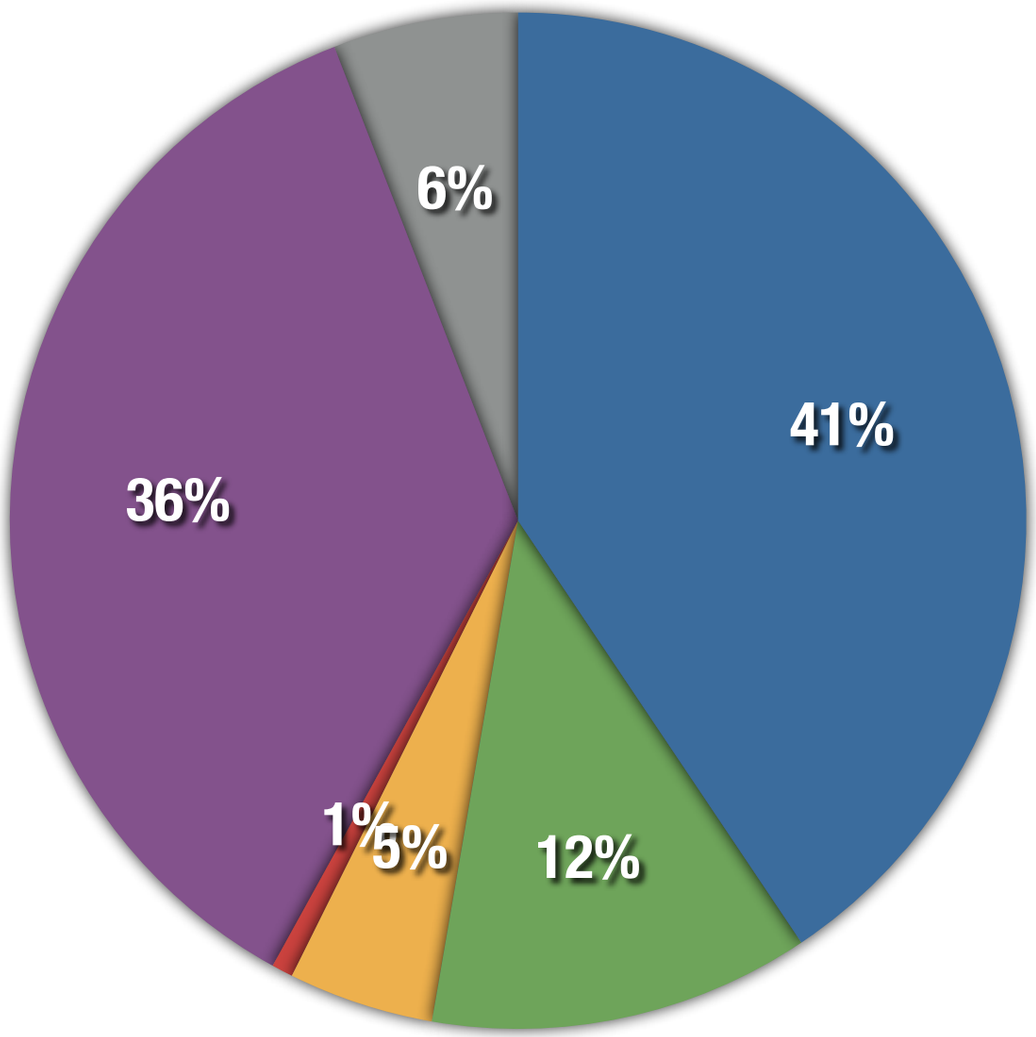
Linking Open Data (LOD)?



RDFa, microdata

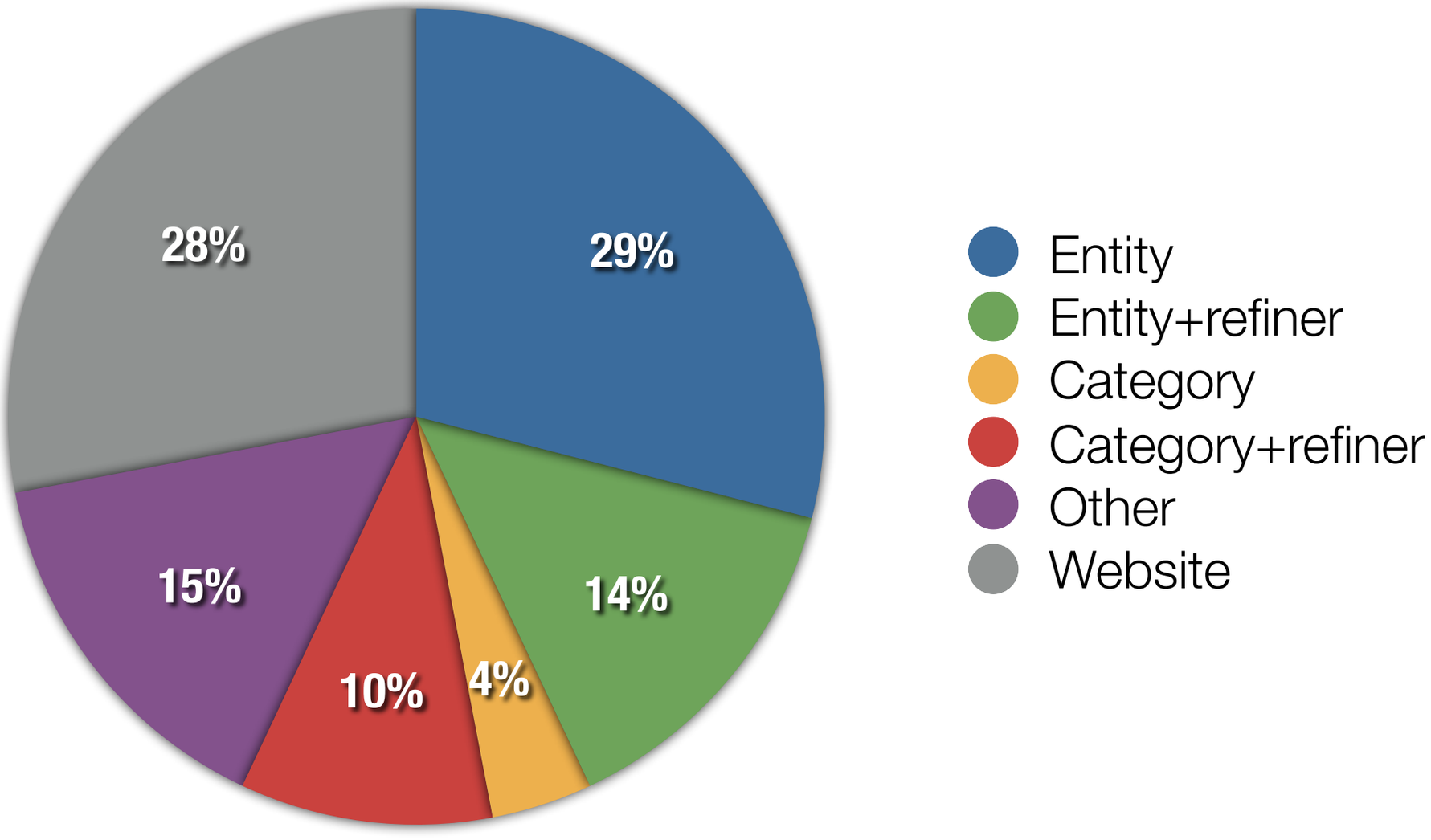
- schema.org, sitemaps.org
  - used by Google, Bing, Yandex, Yahoo!, IPTC, etc.

Distribution of web search queries [Pound et al. 2010]

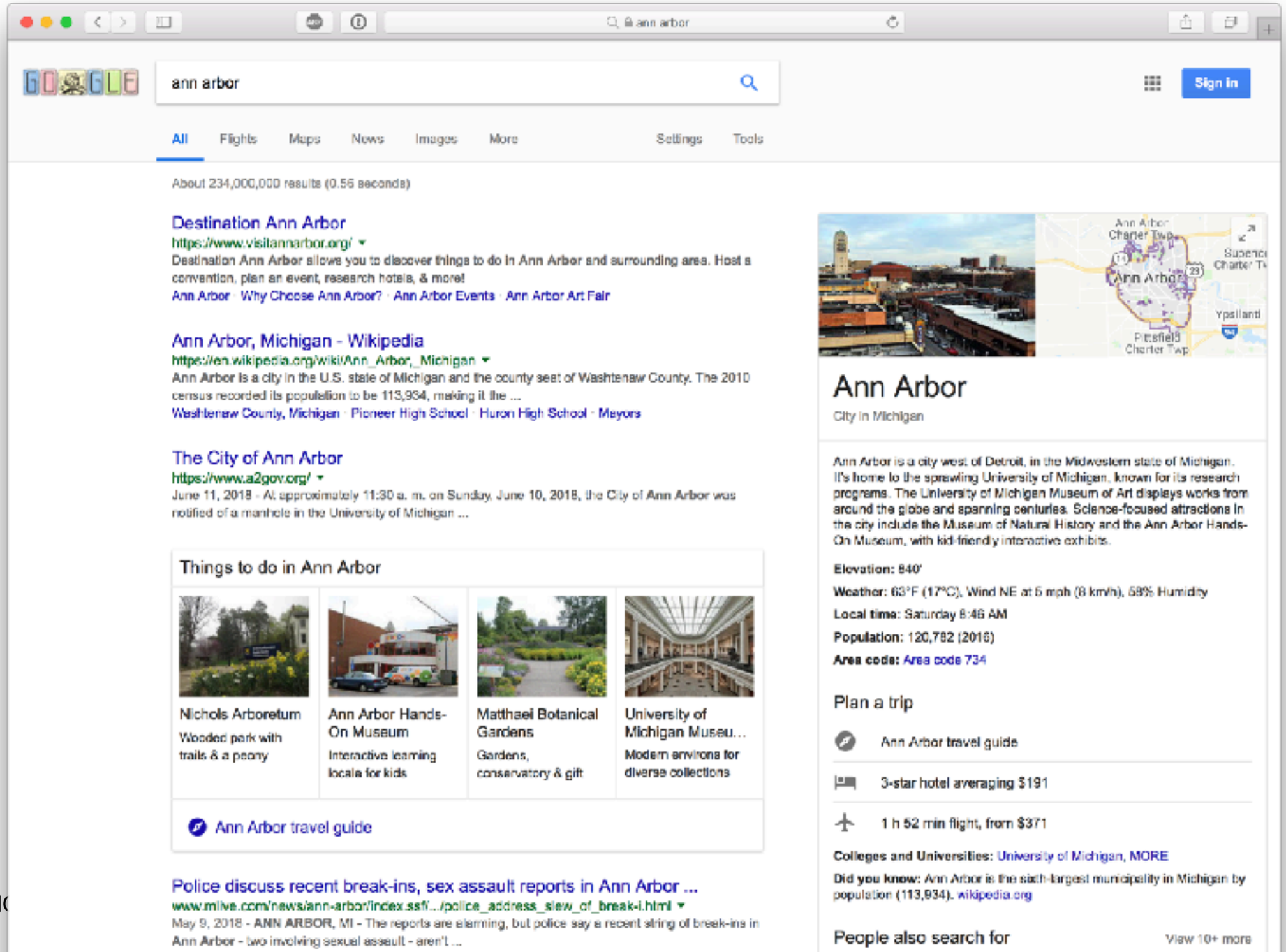


- Entity (“1978 cj5 jeep”)
- Type (“doctors in barcelona”)
- Attribute (“zip code waterville Maine”)
- Relation (“tom cruise katie holmes”)
- Other (“nightlife in Barcelona”)
- Uninterpretable

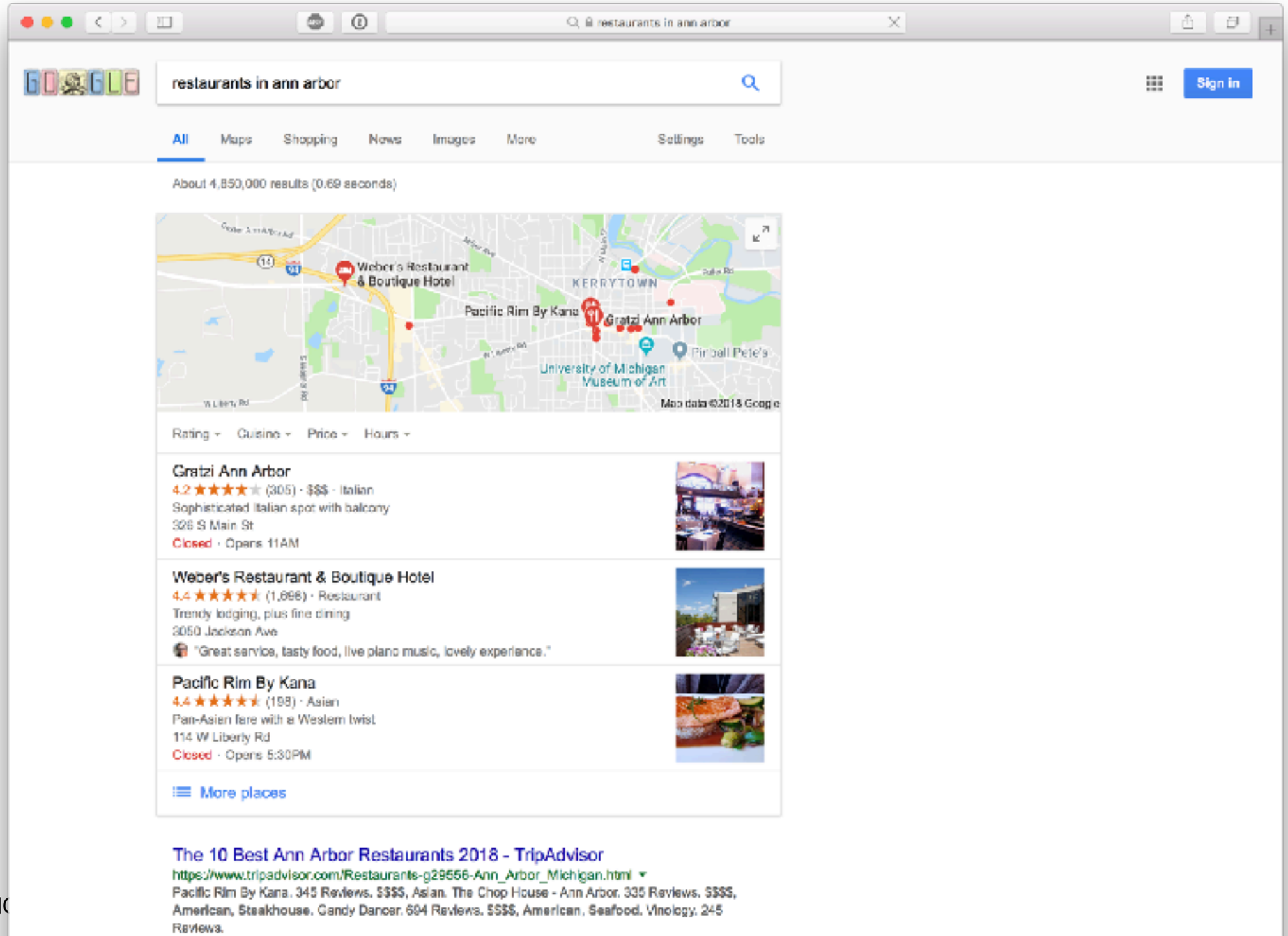
Distribution of web search queries [Lin et al. 2011]



Entity Linking/Retrieval

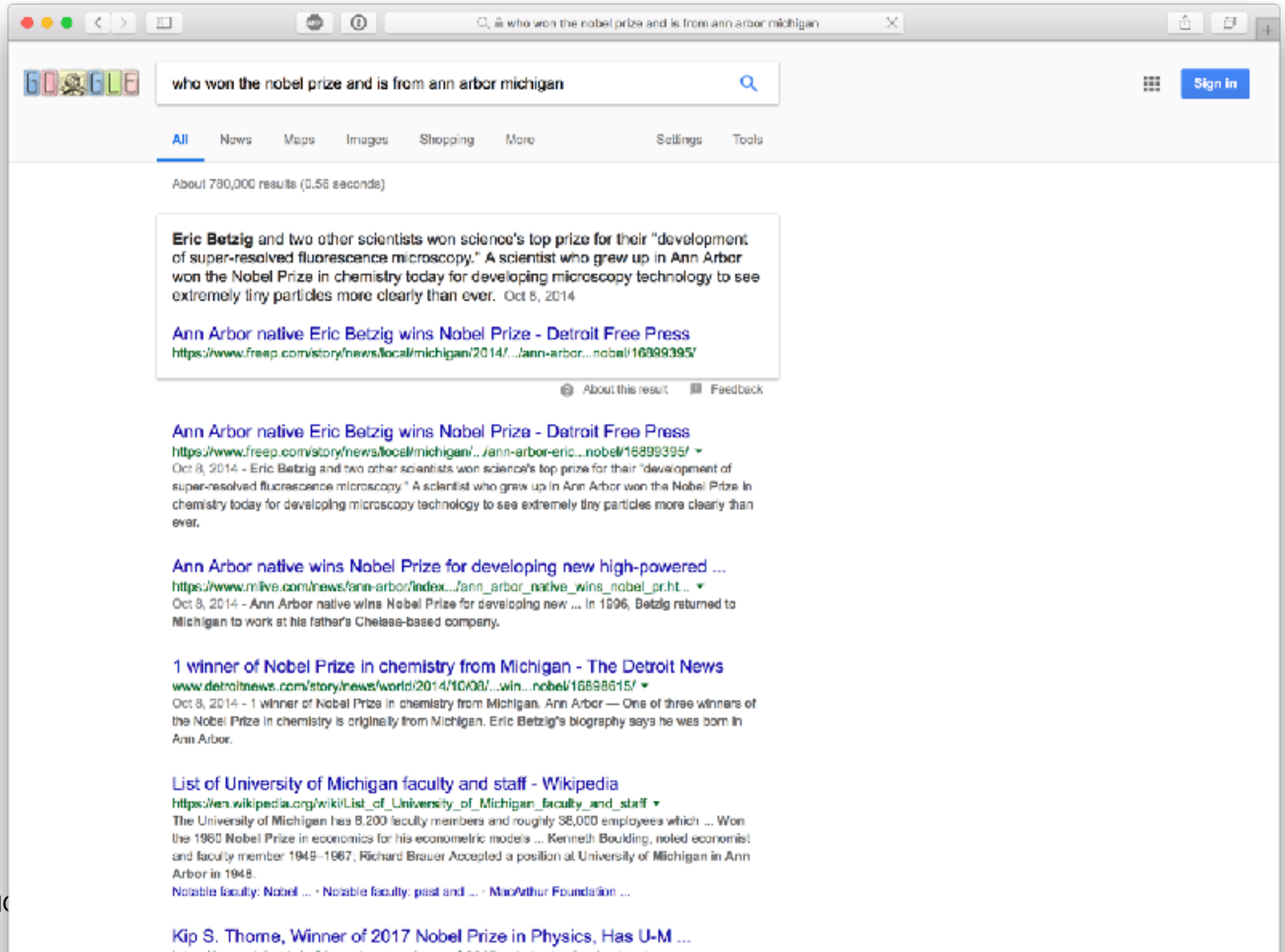


Entity Linking/Retrieval



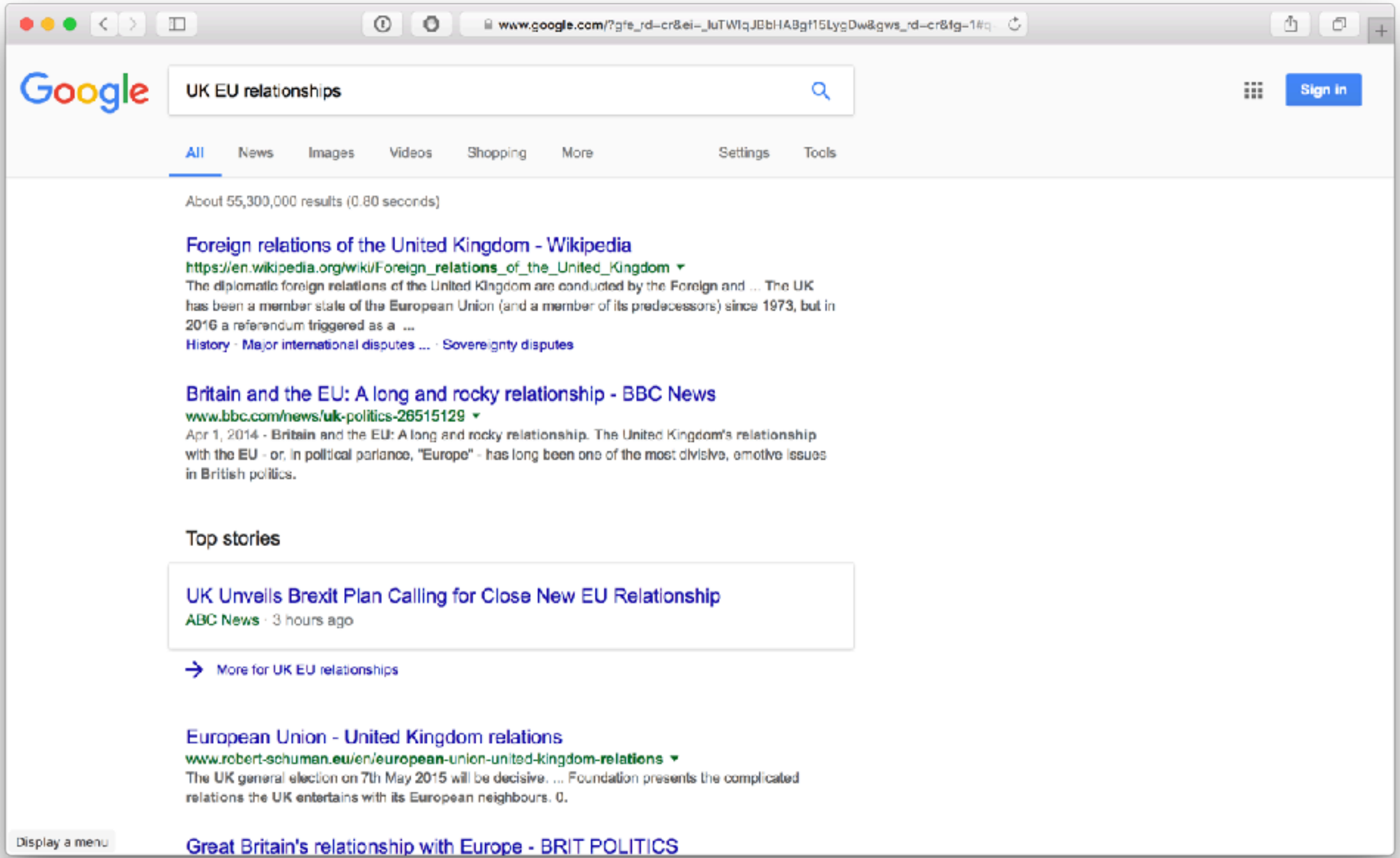
SIG

Entity Linking/Retrieval



SI

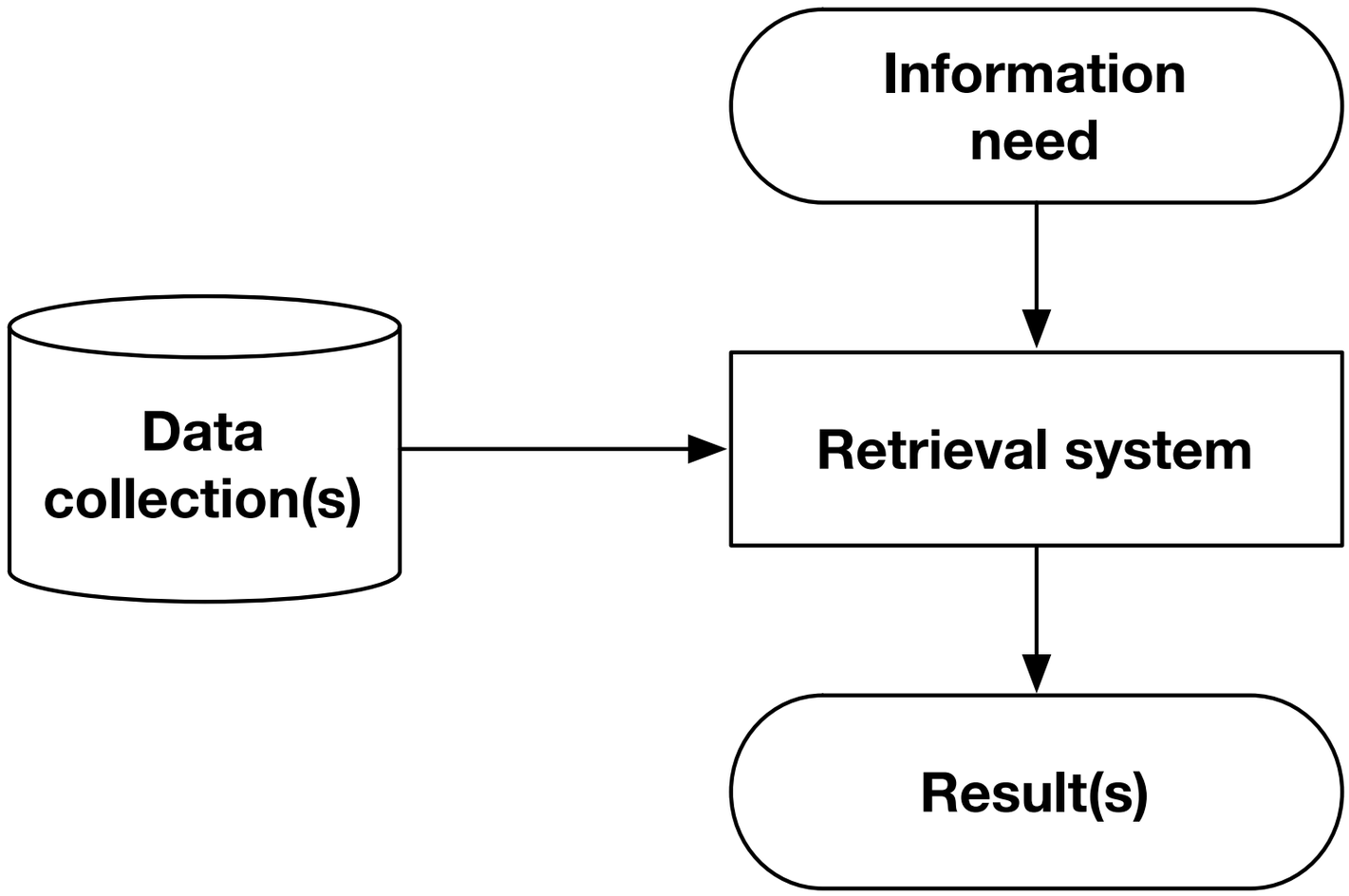
Entity Linking/Retrieval



Today's focus

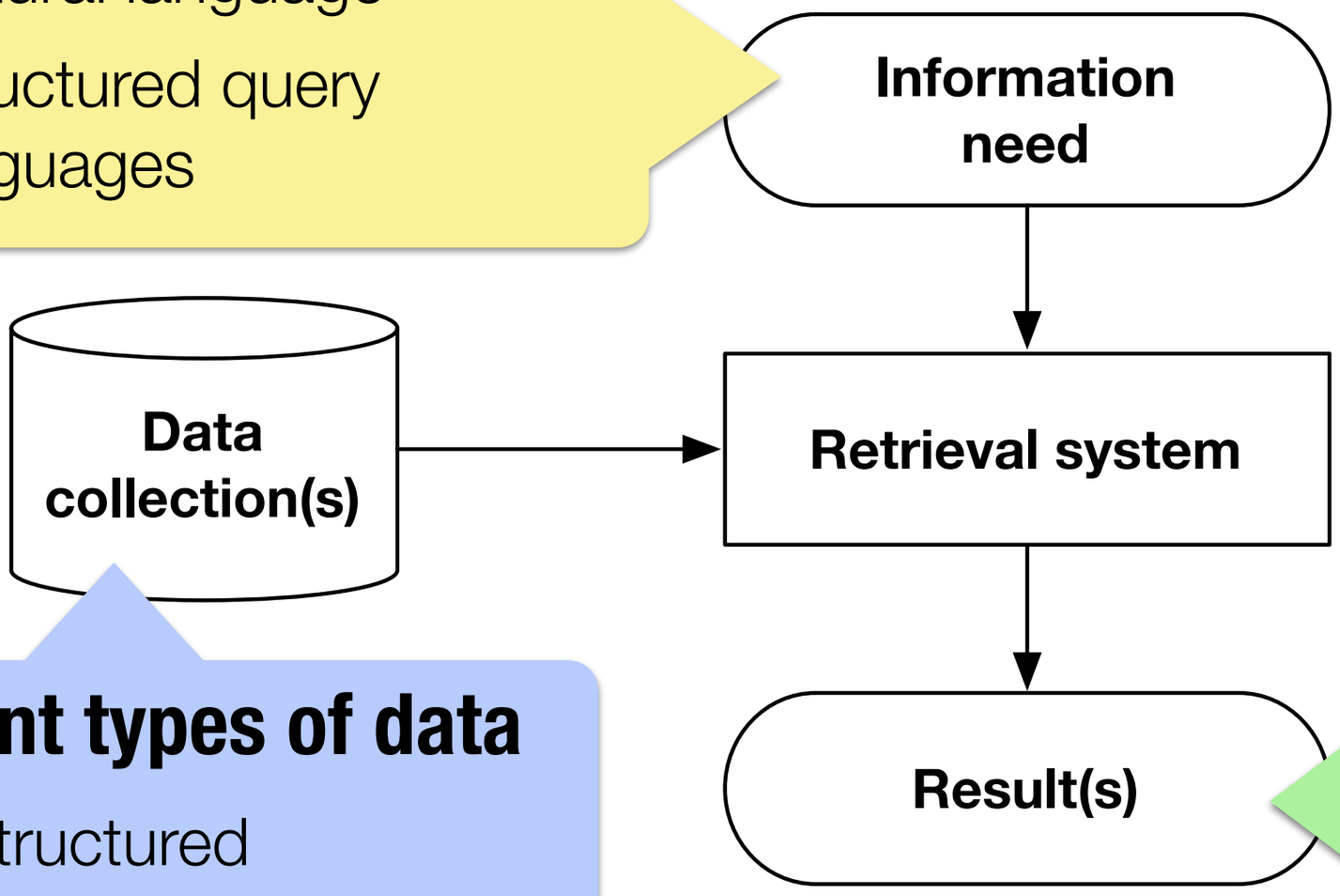
How to use KGs to improve information access.

Birds-eye view



Many ways to express

- Keyword
- Keyword++
- Natural language
- Structured query languages



Different types of data

- Unstructured
- Semistructured
- Structured

Result format

- Ranked list
- Tuples
- (Sub)graphs
- Natural language

Menu

13:30 - 13:45	Introduction
13:45 - 14:15	Part 1 – Entity linking
14:15 - 15:00	Part 2 – Entity Representation and Retrieval
15:00 - 15:30	Coffee break
15:30 - 16:30	Part 3 – Utilizing KGs in Text-centric IR
16:30 - 17:00	Discussion and wrap-up