Utilizing Knowledge Bases in Text-centric Information Retrieval

Laura Dietz (@lauradietz99)
University of New Hampshire

Alexander Kotov (@rusillini)
Wayne State University

Edgar Meij (@edgarmeij)
Bloomberg

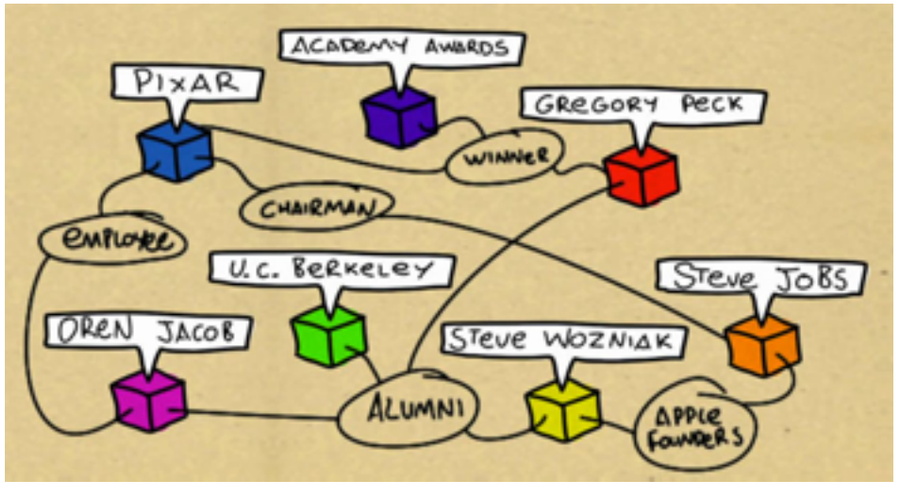
Slides at https://github.com/laura-dietz/tutorial-kb4ir

Entity?

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

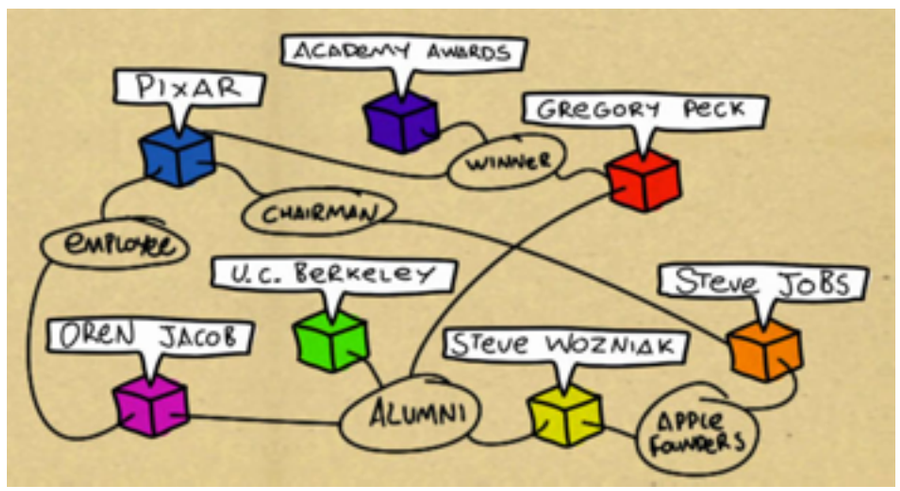
What's so special about entities?

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



Knowledge graphs

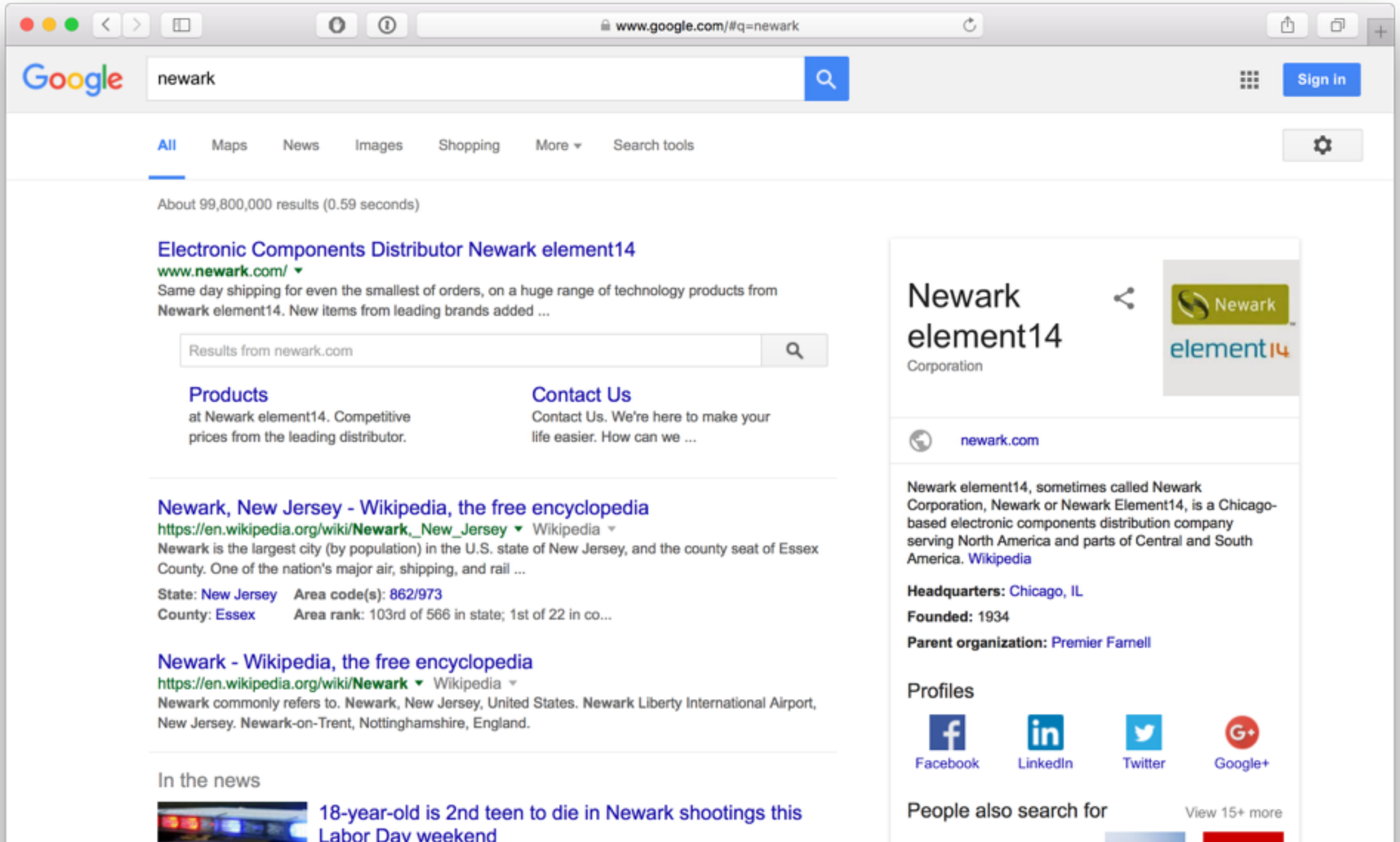
- The “backbone” of semantic search
- They define
  - entities
  - attributes
  - types
  - relations
  - (provenance, sometimes)
  - and more
    - external links, homepages, features, …



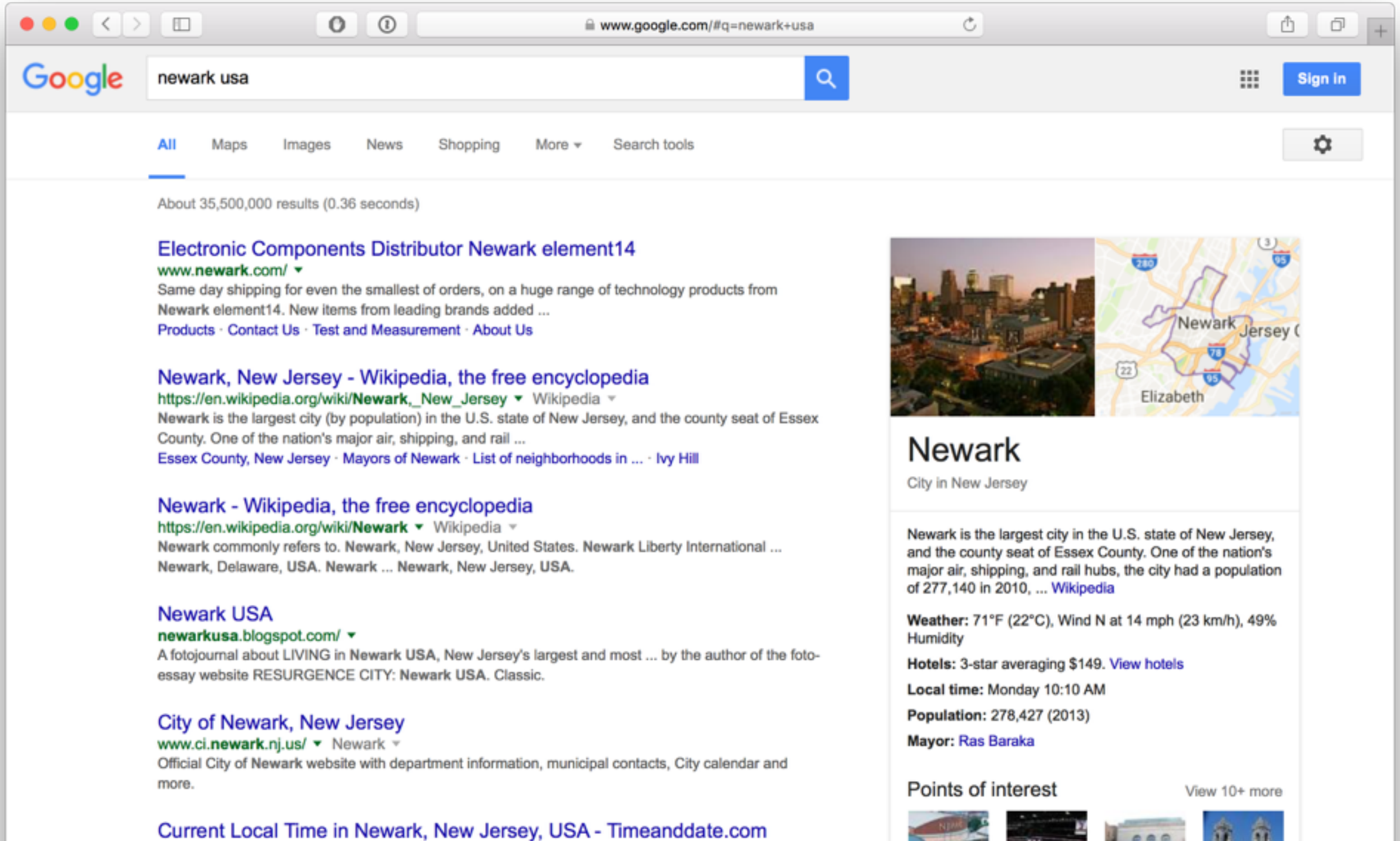
Knowledge graphs

dbpedia:Audi_A4	
foaf:name	Audi A4
rdfs:label	Audi A4
rdfs:comment	The Audi A4 is a compact executive car produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group. The A4 has been built [...]
dbpprop:production	1994 2001 2005 2008
rdf:type	**dbpedia-owl:MeanOfTransportation** **dbpedia-owl:Automobile**
dbpedia-owl:manufacturer	**dbpedia:Audi**
dbpedia-owl:class	**dbpedia:Compact_executive_car**
owl:sameAs	**freebase:Audi A4**
is **dbpedia-owl:predecessor** of	**dbpedia:Audi_A5**
is **dbpprop:similar** of	**dbpedia:Cadillac_BLS**

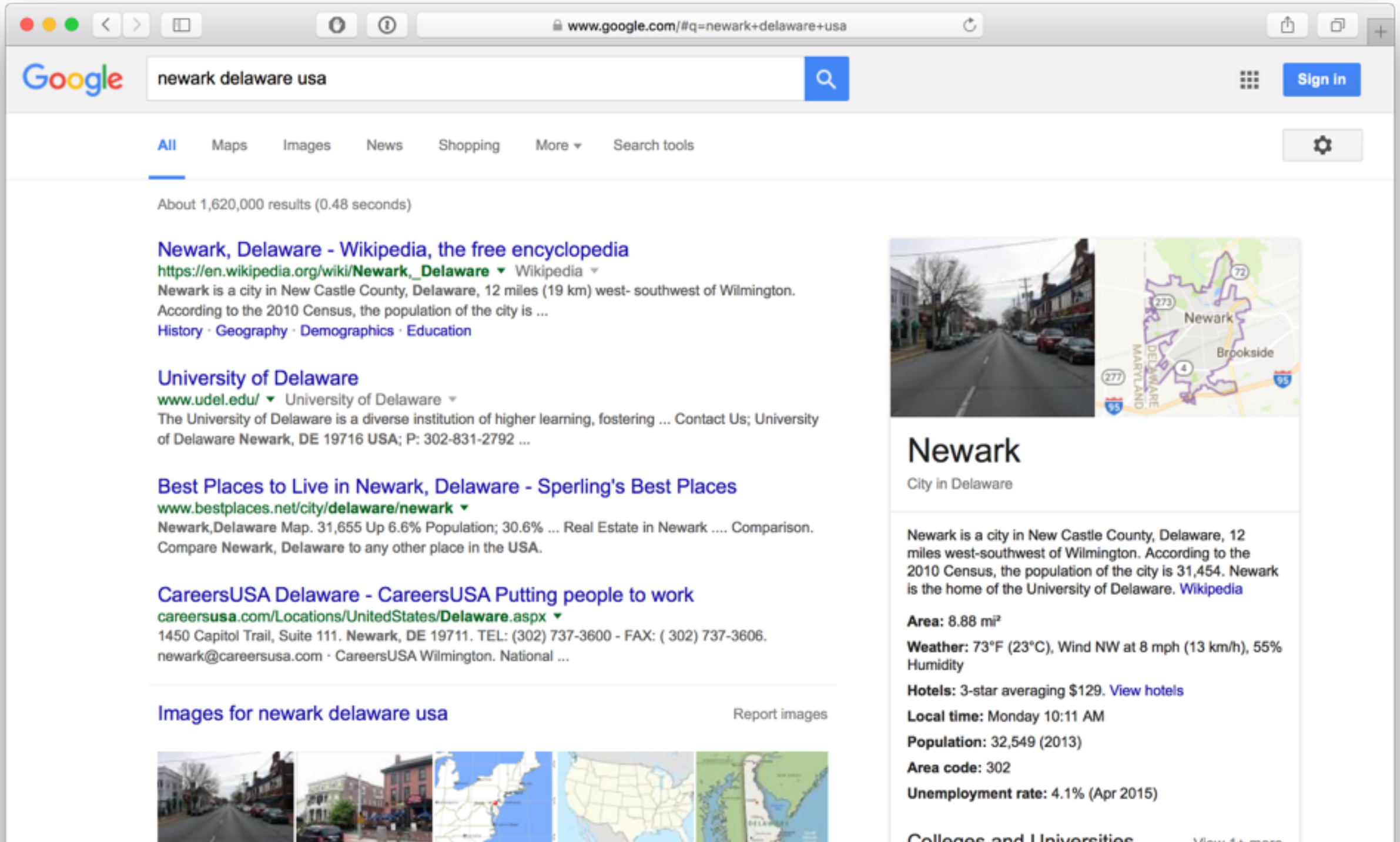
Entity Linking/Retrieval



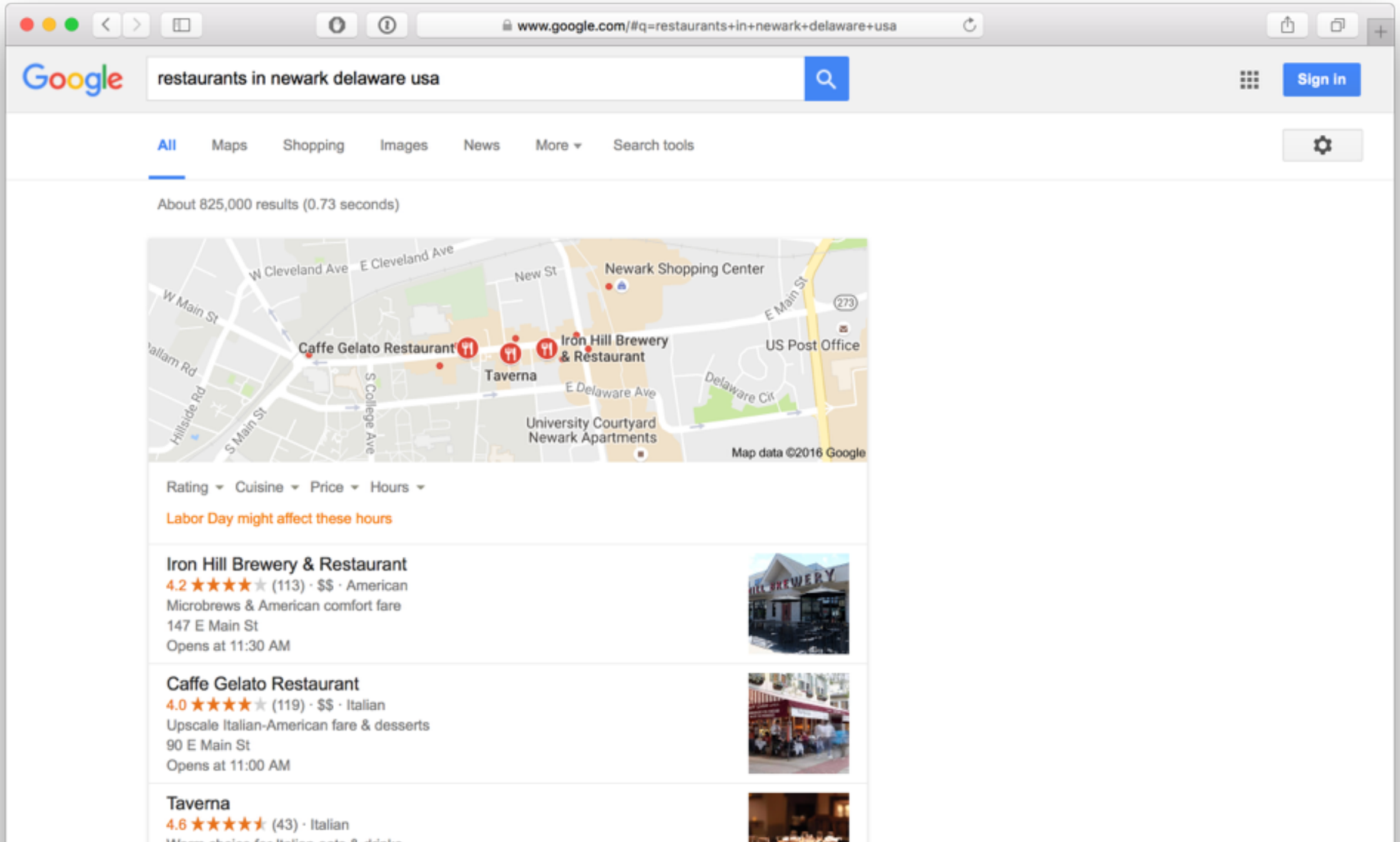
Entity Linking/Retrieval



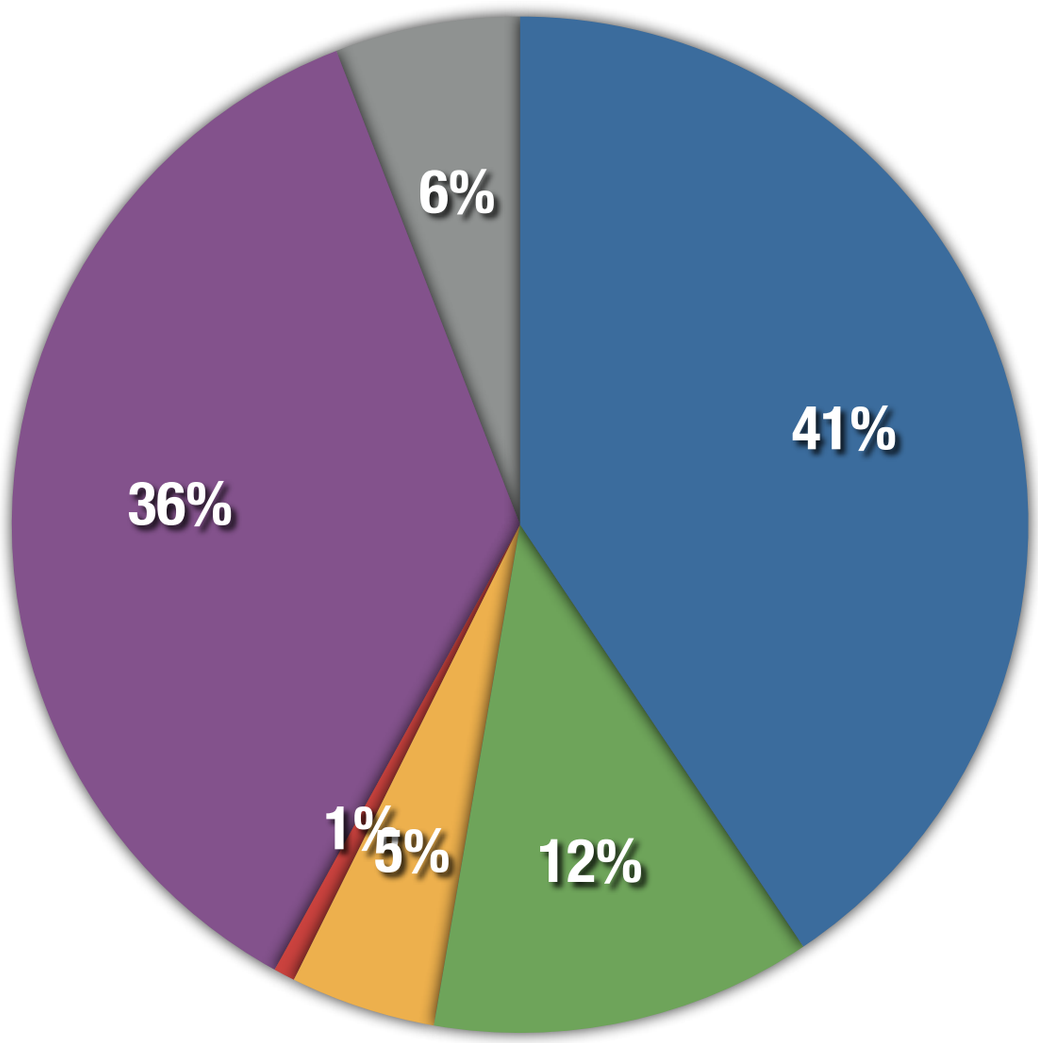
Entity Linking/Retrieval



Entity Retrieval

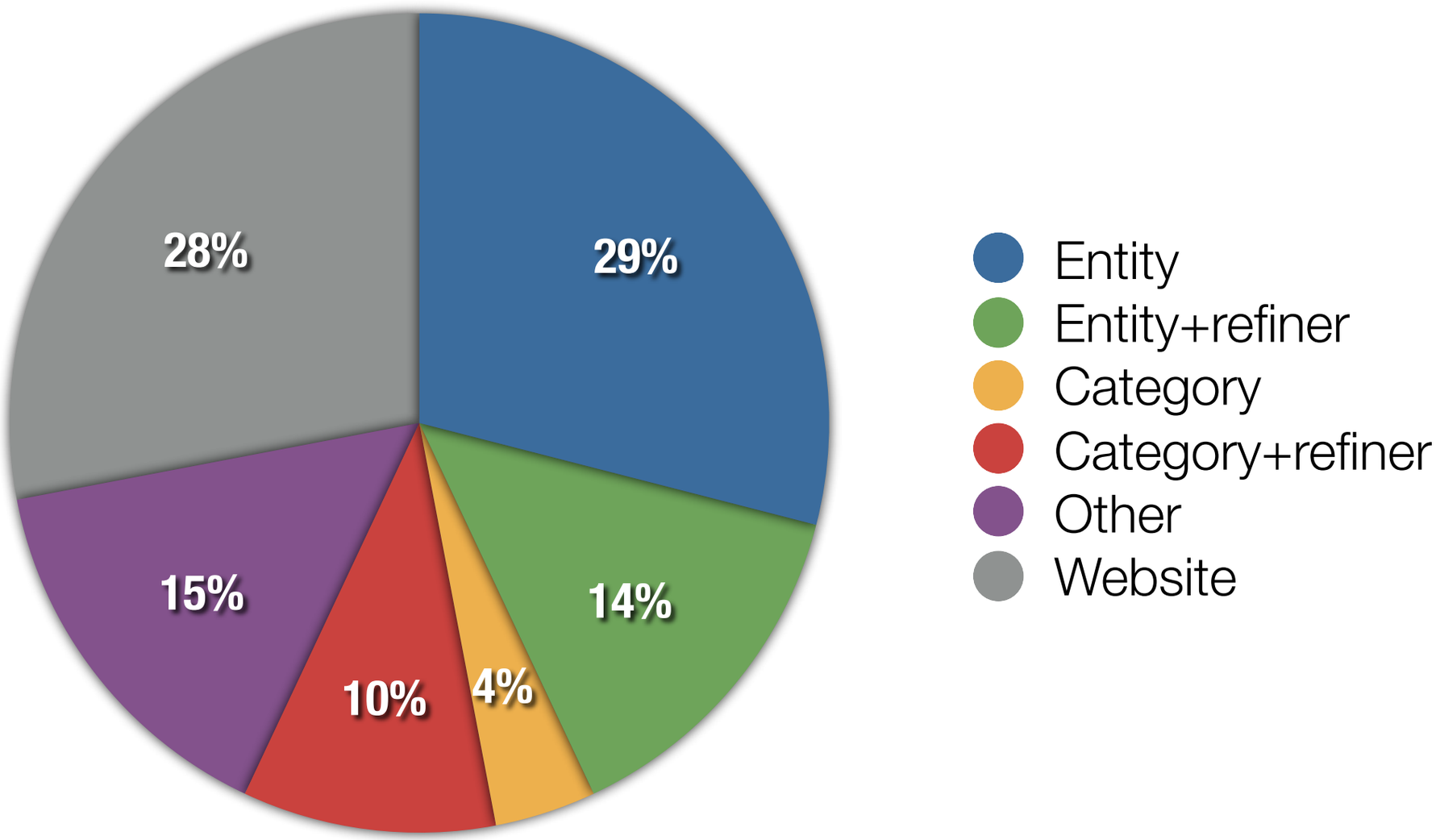


Distribution of web search queries [Pound et al. 2010]



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

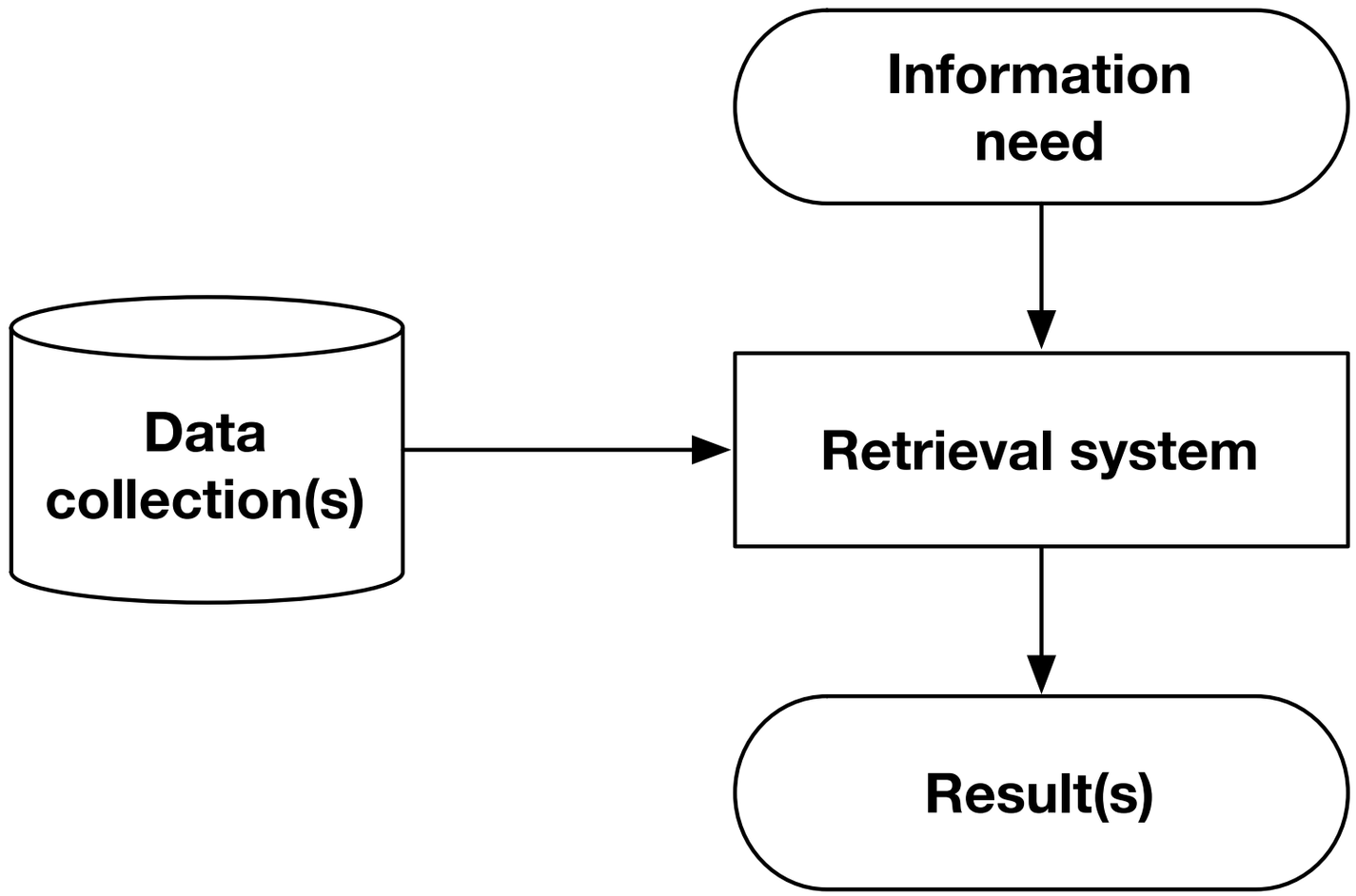
Distribution of web search queries [Lin et al. 2011]



Today’s focus

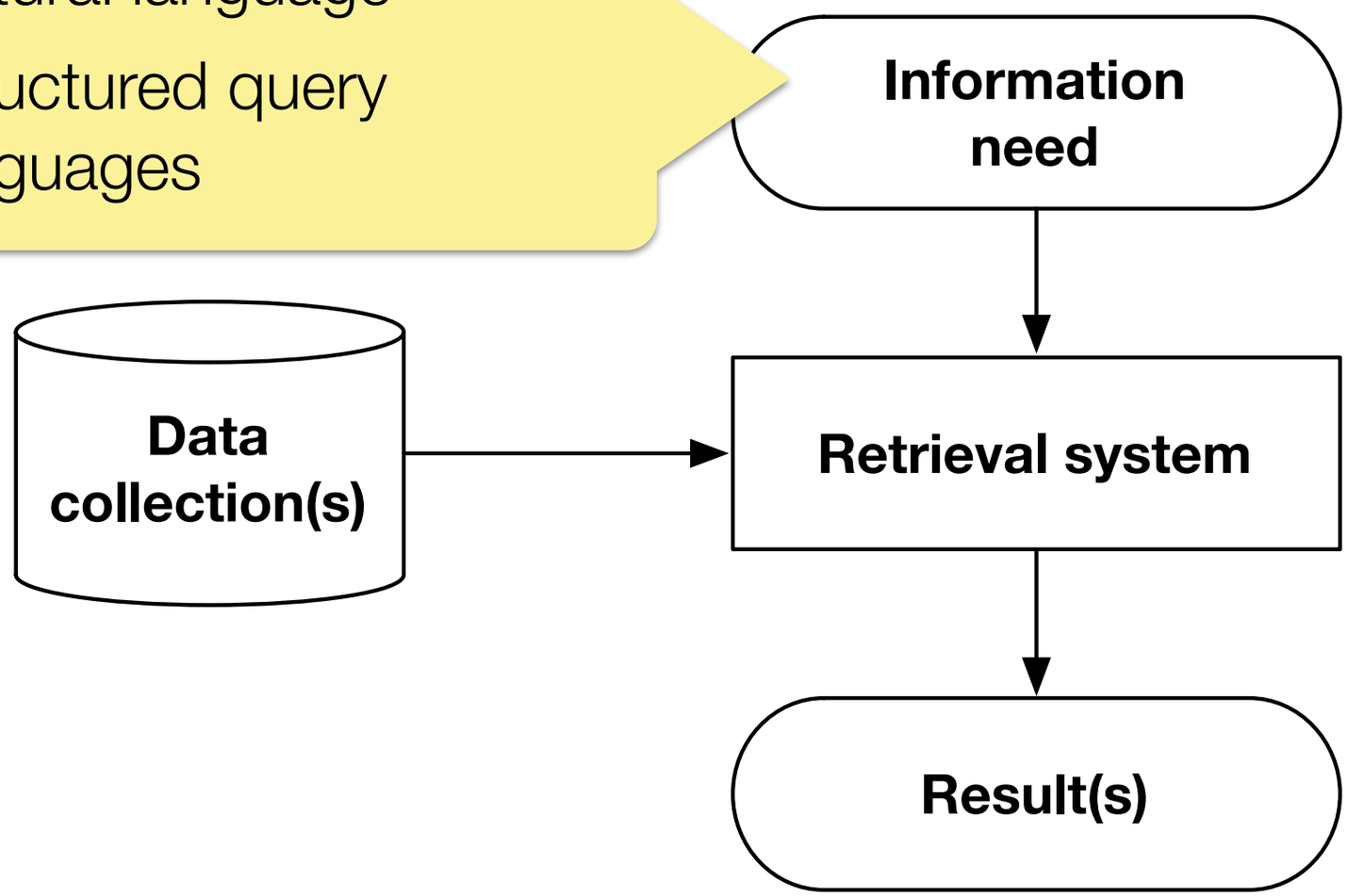
How to use KGs to improve information access.

Birds-eye view



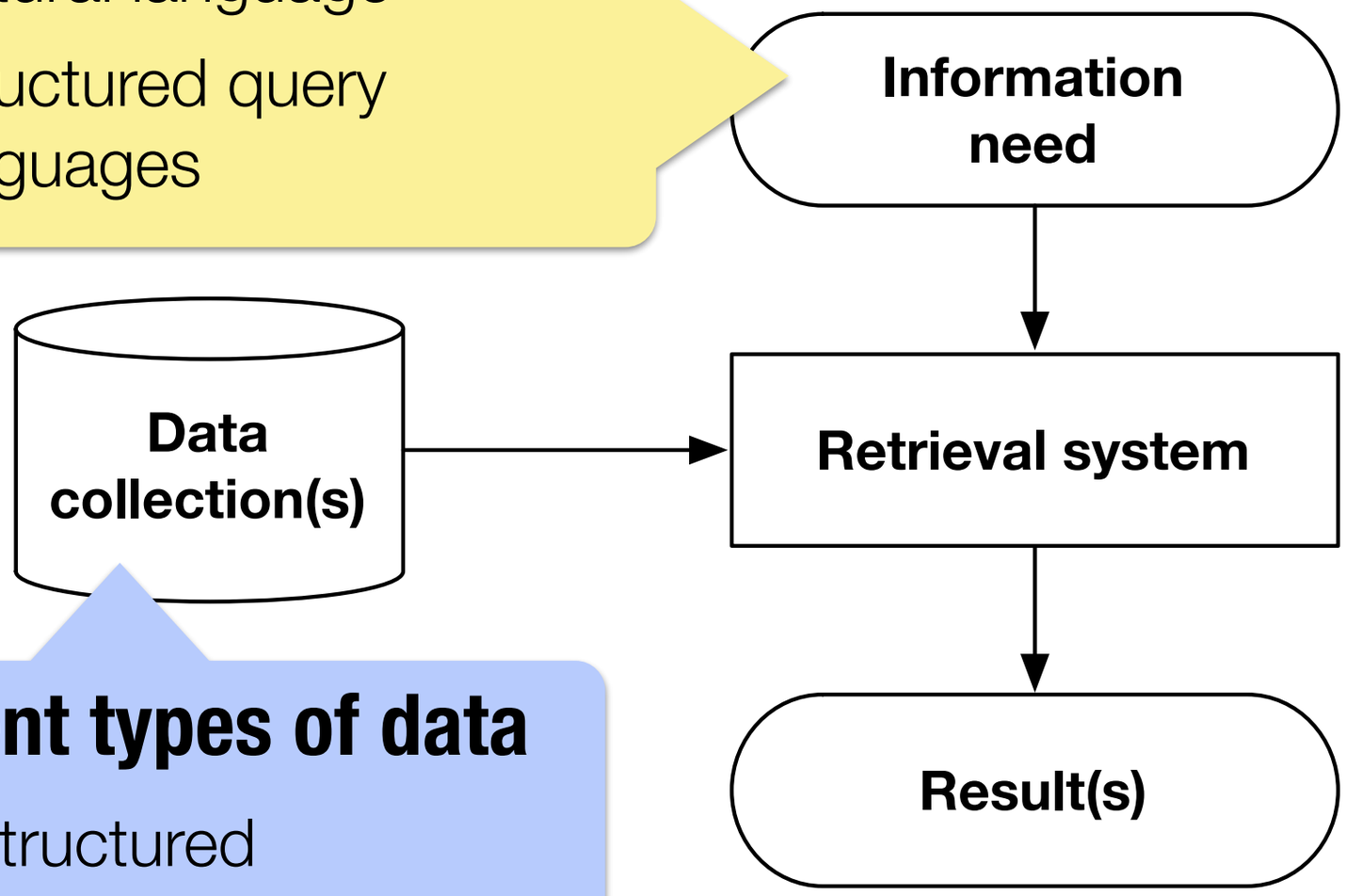
Many ways to express

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



Many ways to express

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

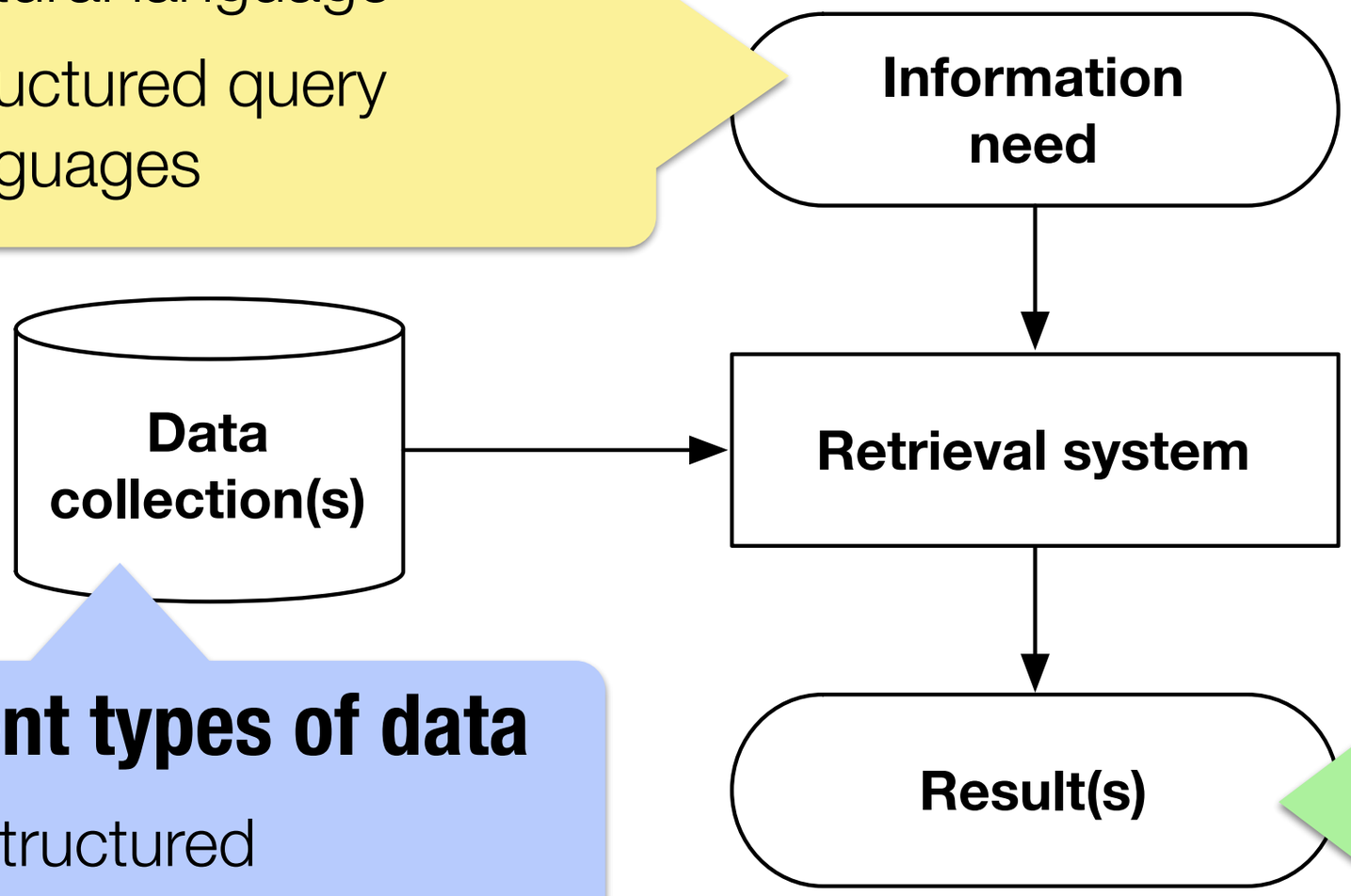


Different types of data

- Unstructured
- Semistructured
- Structured

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

Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO

Popular (semi)structured data sources

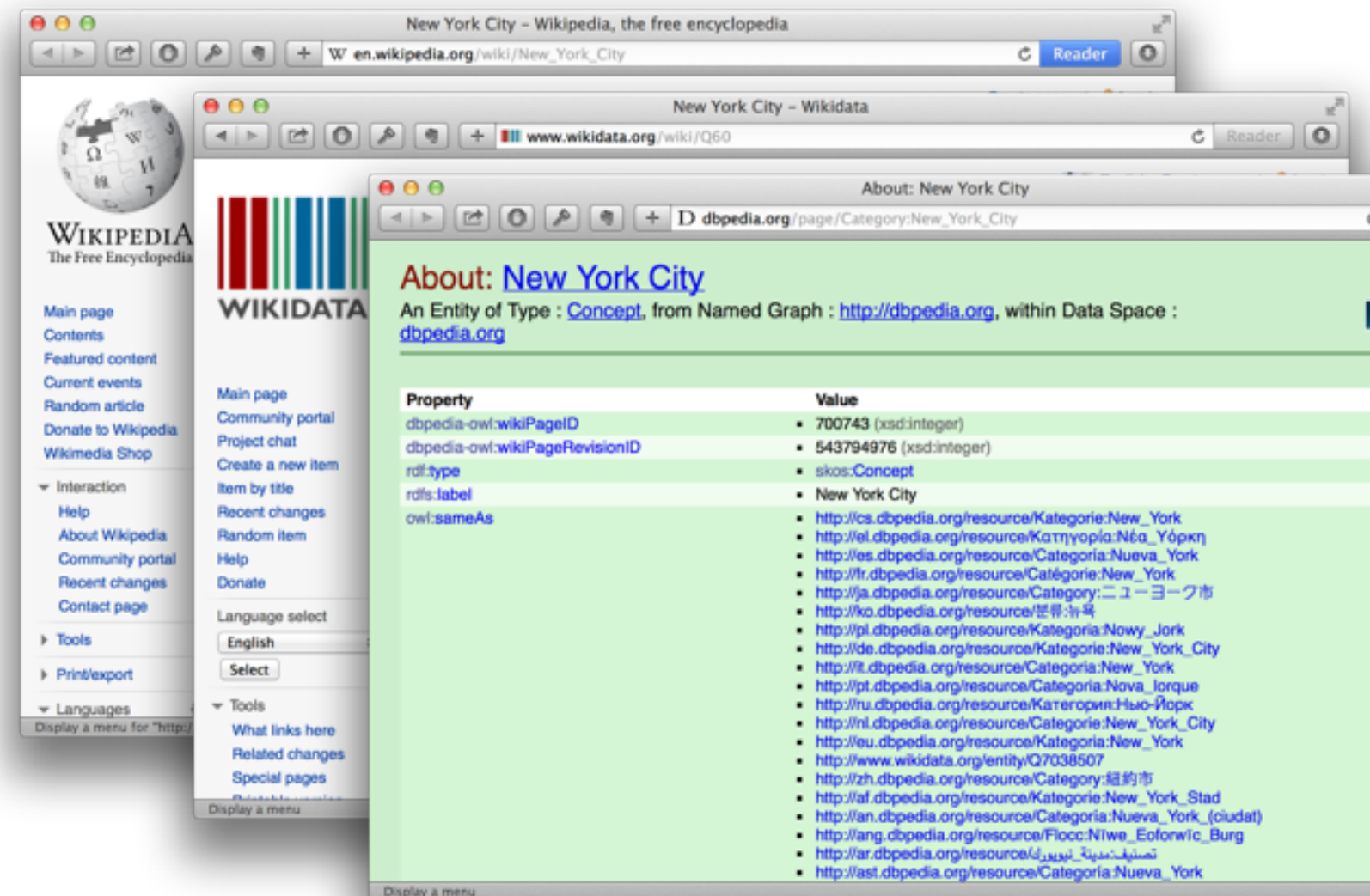
- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO

Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO

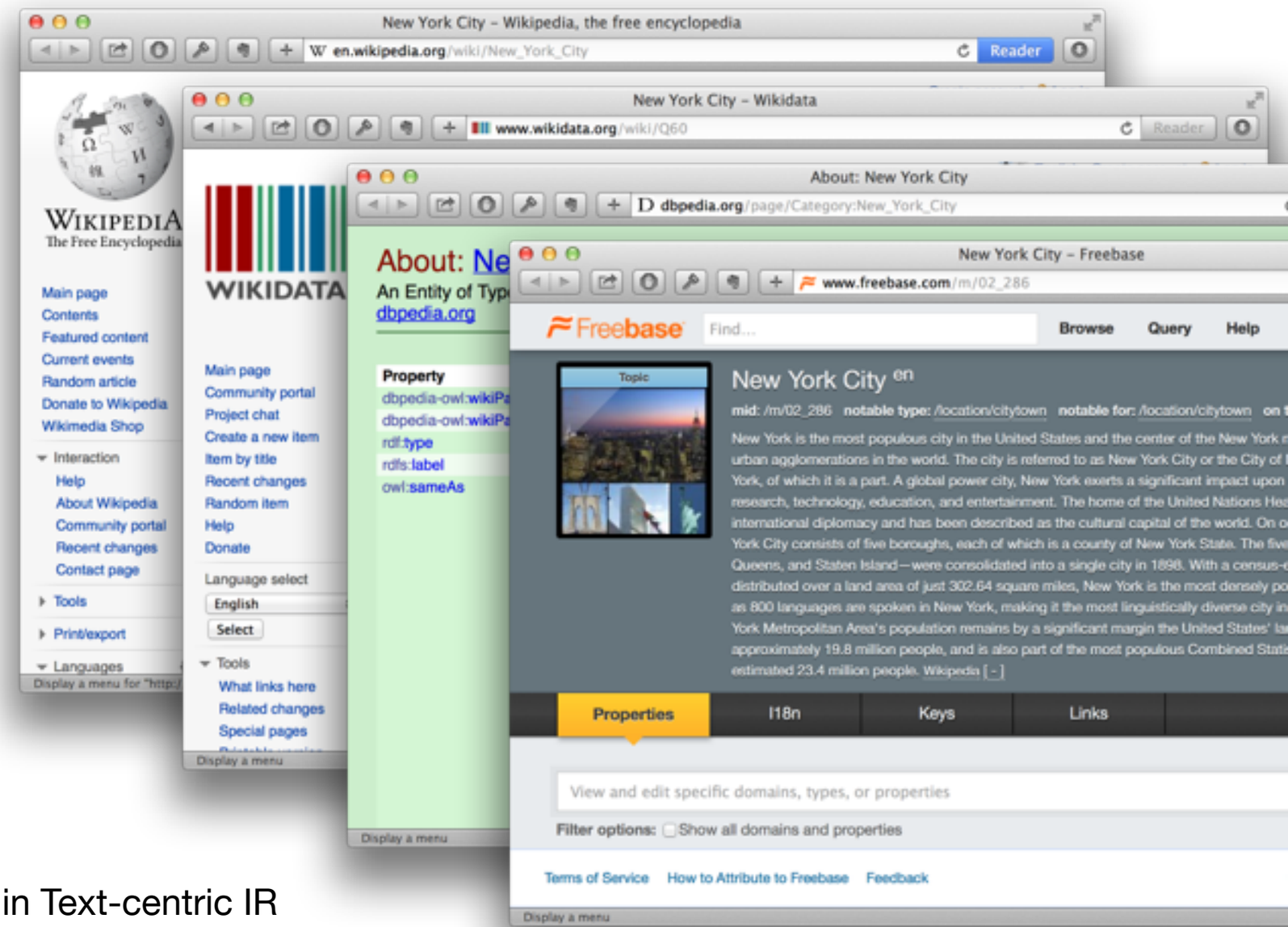
Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO



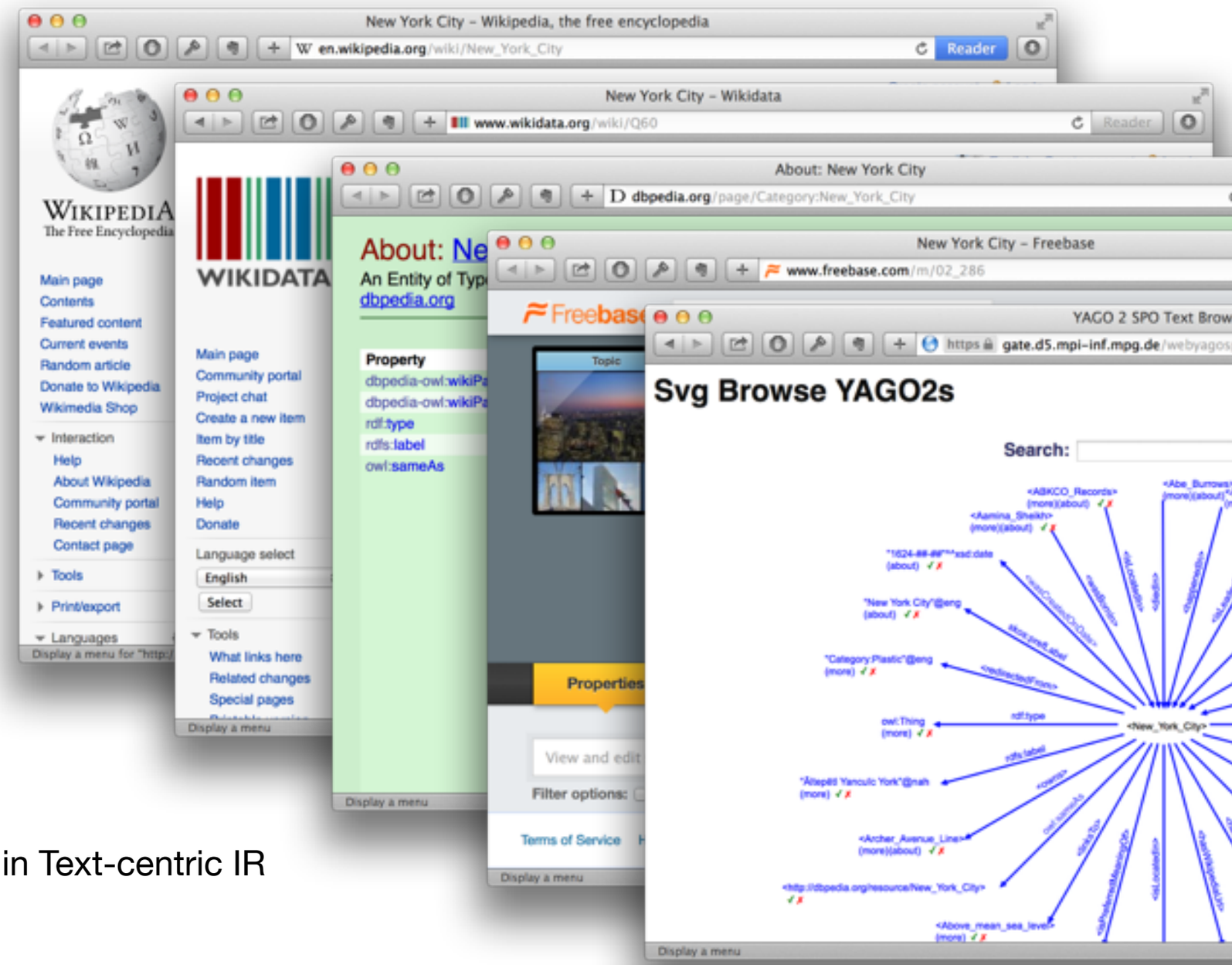
Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO



Popular (semi)structured data sources

- Wikipedia
- Wikidata
- DBpedia
- Freebase
- YAGO



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

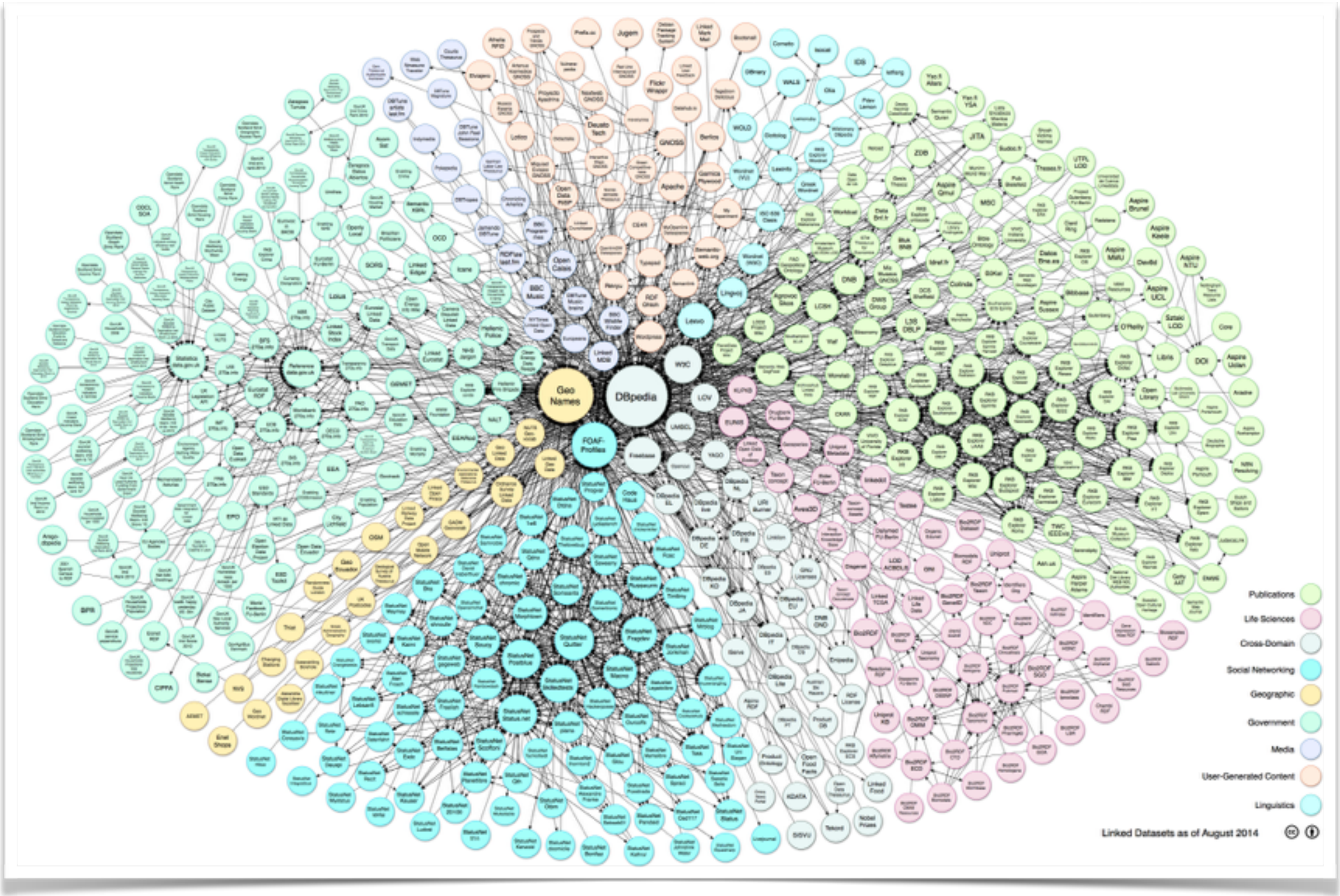
Freebase

- 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/

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
- See http://www.mpi-inf.mpg.de/yago-naga/yago/

Linking Open Data (LOD)?

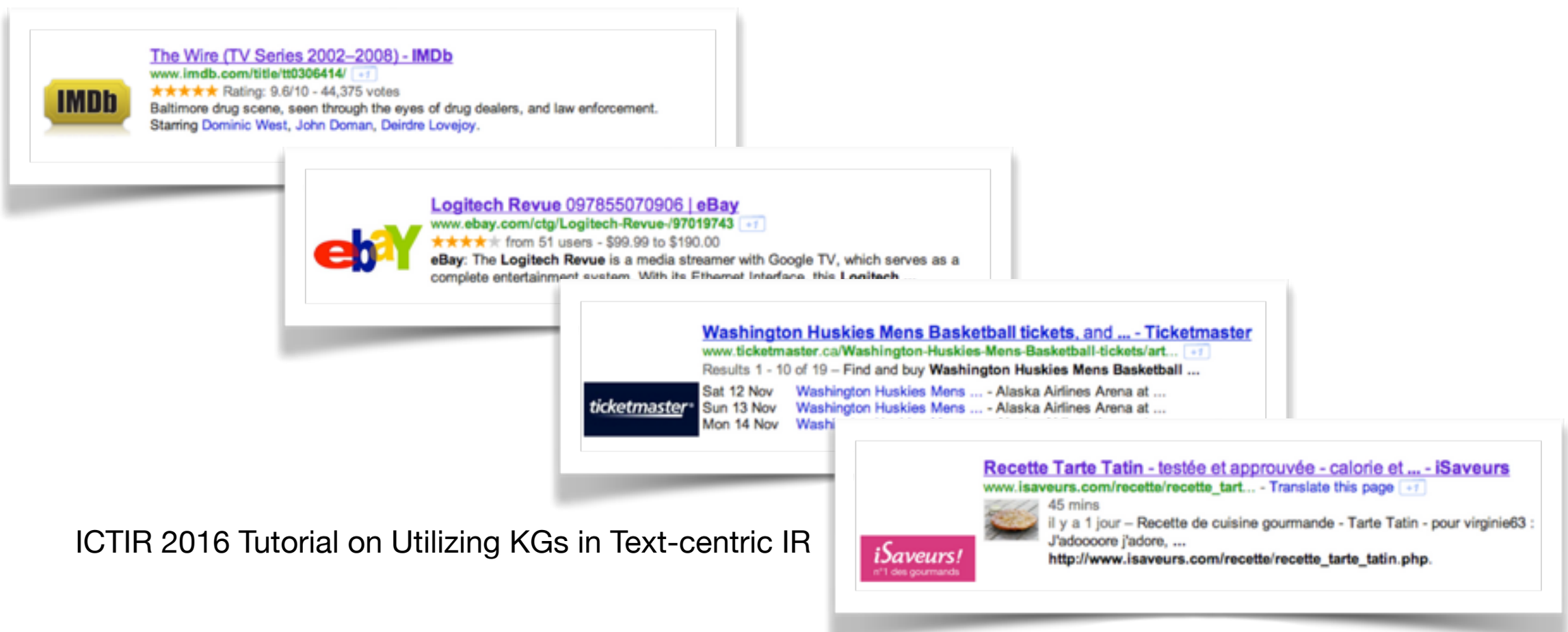


RDFa

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

RDFa

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



Menu

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