DAT630

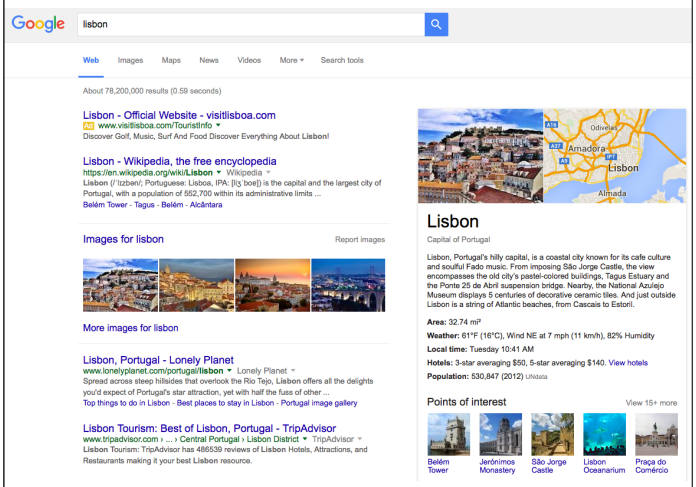
# Entity Retrieval I.

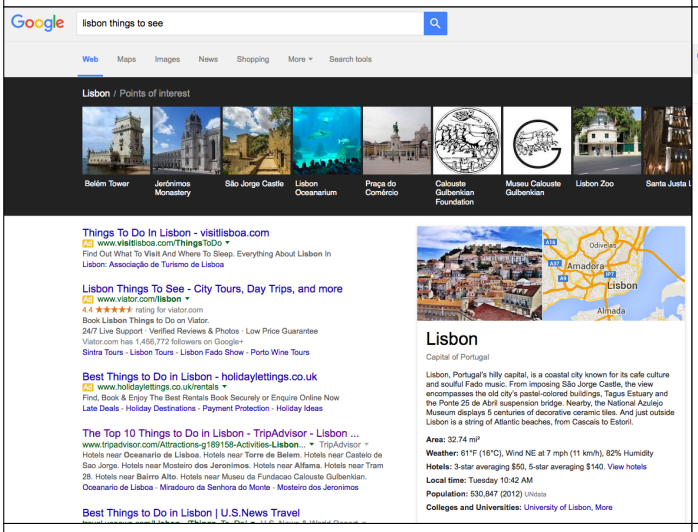
01/11/2016

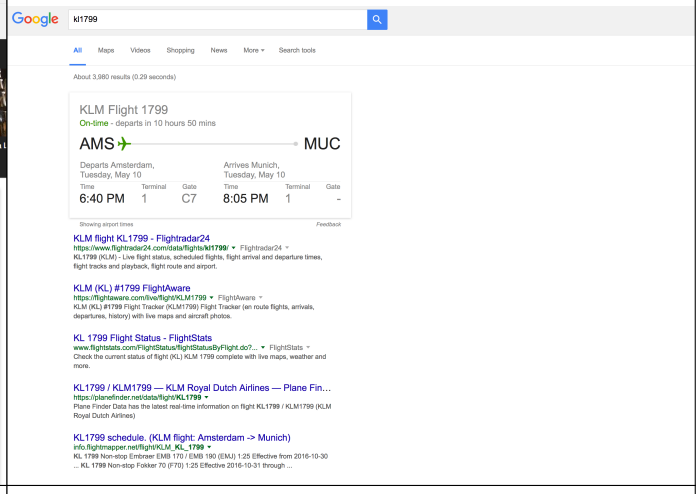
**Krisztian Balog** | University of Stavanger

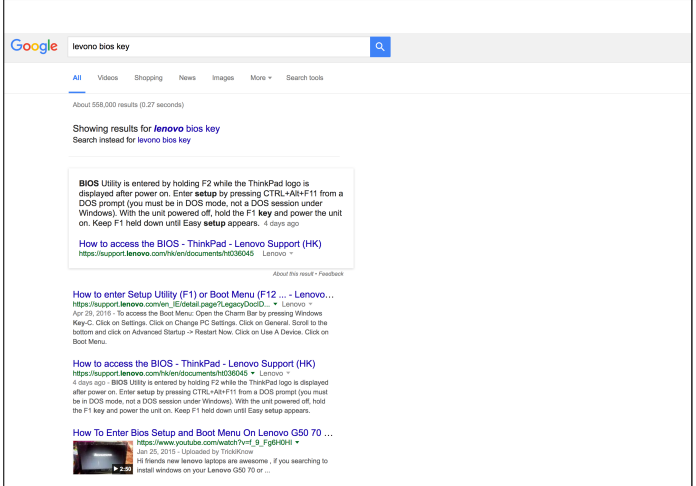
# Semantic Search

# What is semantic search?




# What is semantic search?

- "Search with meaning"
- Improve search accuracy by **understanding** searcher intent and the **contextual meaning** of terms/documents/…
- Move beyond "ten blue links" (towards actually answering information needs) using rich context

# Semantic search

- Centers around entities
  - “Who was the first human in outer space?”
  - “How tall is the Eiffel tower?”
  - “Who is Brad Pitt married to?”
  - “Where is the closest Starbucks?”
  - “Which airlines fly the Airbus A380?”
  - “What is the best Chinese restaurant in Montreal?”
- Entity/Attribute/Relationship retrieval
  - + social, + personal
  - + (hyper)local

# Semantic search

- Combination of entity-related techniques, from various fields
  - Information Retrieval (IR)
  - Natural Language Processing (NLP)
  - Databases (DB)
  - Semantic Web (SW)

# What is an entity?

people


organizations

locations

products

# What is an entity?

- Uniquely identifiable *thing* or *object*
  - “A thing with a distinct and independent existence”
- Characterized by having:
  - Unique ID
  - Name(s)
  - Type(s)
  - Attributes (/Descriptions)
  - Relationships to other entities

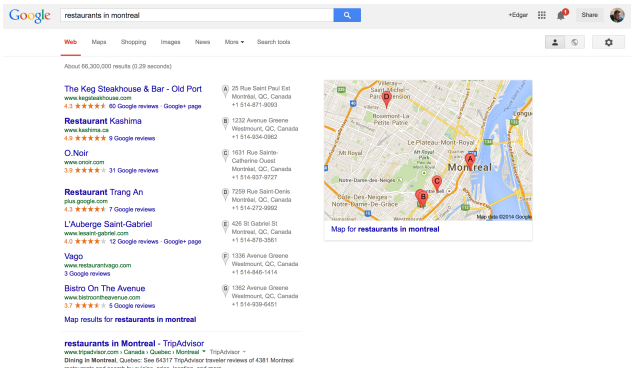
# Entities…

- are meaningful units for organizing information
- are a key enabling component in semantic search

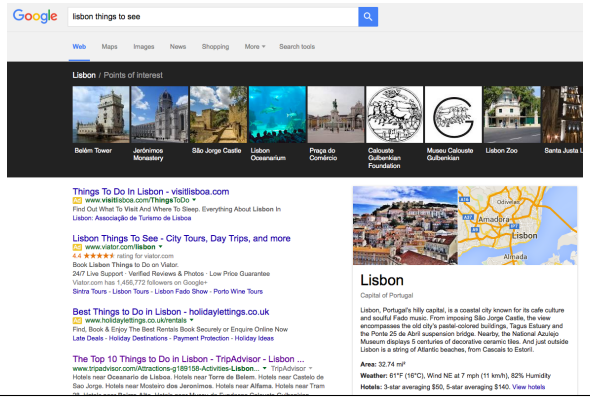
# Entity Linking


Image taken from Milne and Witten (2008b). Learning to Link with Wikipedia. In CIKM '08.

# Entity Retrieval



# Entity Linking/Retrieval




# Meet the Data

# Data collection

- Unstructured
  - Documents, web pages, snippets, …
- Semistructured
  - XML, RDF, …
- Structured
  - Relational DBs, RDF, …

} Often organized around entities

# Single most popular semistructured data source

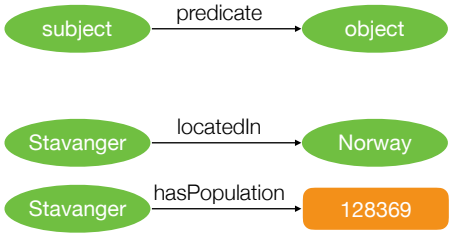

# Knowledge Bases

- Aimed at *machine understanding*
- Comprese a large set of large set of assertions about the world
  - Describe (specific) entities and their relationships

# RDF Data Model

- Resource Description Framework
  - Each resource is identified by a URI (Unique Resource Identifier)
  - (Entities = resources)
- Assertions are represented as triples
  - **Subject** (resource)
  - **Predicate** (relation)
  - **Object** (resource or literal)


# RDF Data Model




# Example

- How can this information be represented using RDF?

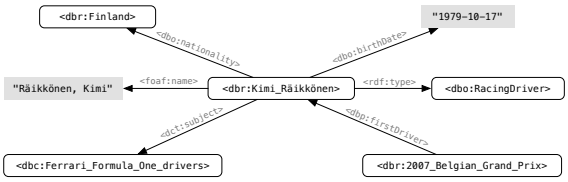
Kimi Räikkönen is a Finnish racing driver, born on October 17, 1979, currently driving for Ferrari in Formula One.

# Example

| Subject | Predicate | Object |
|---|---|---|
| <dbr:Kimi_Räikkönen> | <foaf:name> | "Räikkönen, Kimi" |
| <dbr:Kimi_Räikkönen> | <dbo:nationality> | <dbr:Finland> |
| <dbr:Kimi_Räikkönen> | <dbo:birthDate> | "1979-10-17" |
| <dbr:Kimi_Räikkönen> | <rdf:type> | <dbo:RacingDriver> |
| <dbr:Kimi_Räikkönen> | <dct:subject> | <dbc:Ferrari_Formula_One_drivers> |

# Knowledge bases

- Conceptually form a large, directed graph
- Also called ***knowledge graphs*** when the emphasis is on relationships between entities




# SPARQL

- Structured query language for RDF

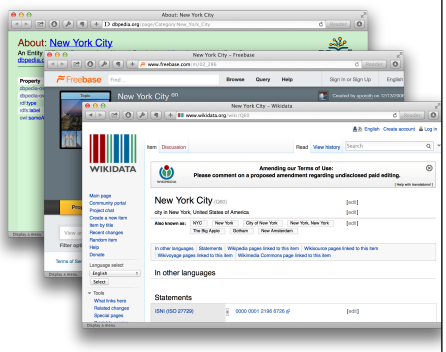
```
SELECT ?p WHERE {
?p has-profession Computer_Scientist .
?p has-gender Female .
?p occurs-with "semantic search"}
```

# Early Attempt: Cyc

- Started in 1984 with the goal to manually build a knowledge base of everyday common knowledge
- … still building and far from complete
- "one of the most controversial endeavors of the artificial intelligence history"

# Popular (Public) Knowledge Bases

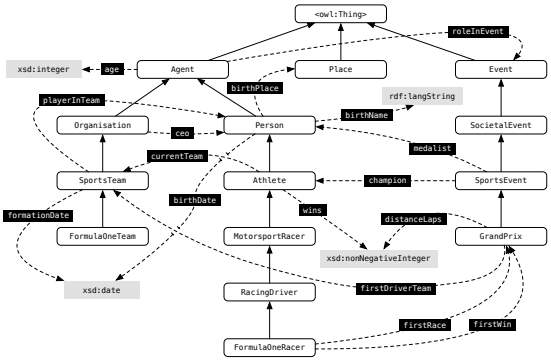
- DBpedia
- Freebase
- Wikidata



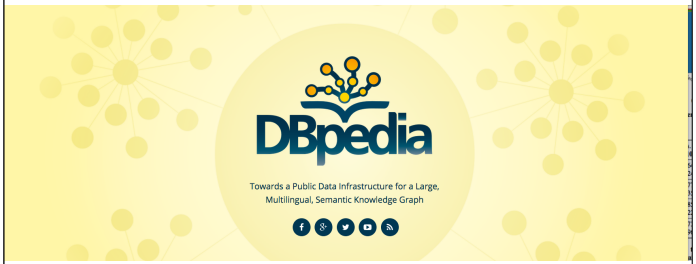

# DBpedia

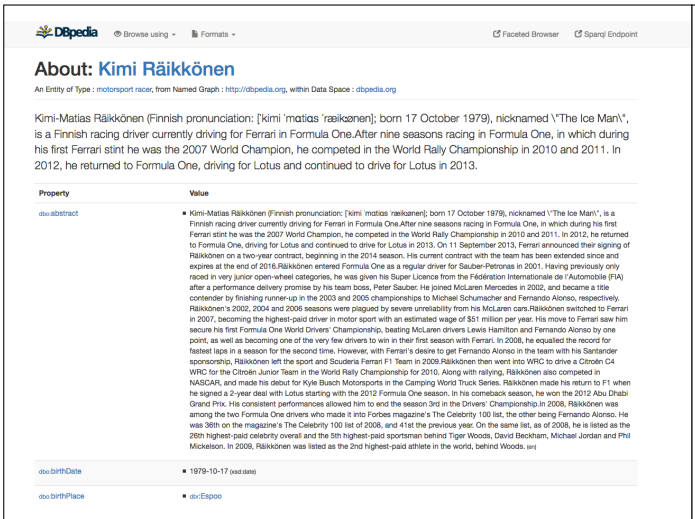
- "A database version of Wikipedia"
- Extracts RDF statements from Wikipedia articles
  - Mostly relies on infoboxes
  - Further homogenization or "normalization" is performed to achieve high data quality
    - Using manual mappings against the DBpedia Ontology

# DBpedia Ontology

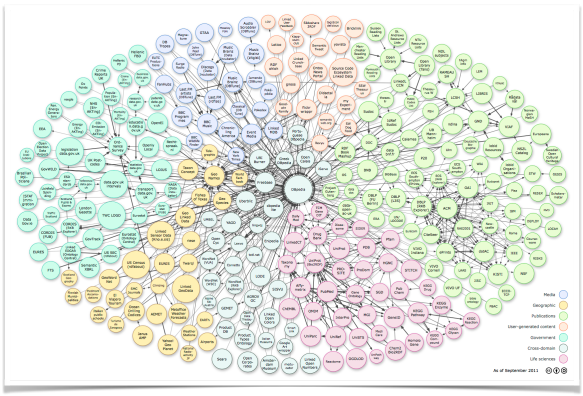



# dbpedia.org






# Linking Open Data (LOD)



# (re)Branding

- Semantic Web data
- Linking Open Data
- Web of Data

# Proprietary Knowledge Bases

Google Knowledge Graph

facebook Entity Graph

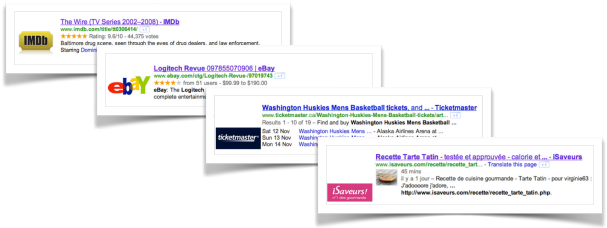
Microsoft Satori

> … the knowledge graph is one of Google's biggest search milestones of the last decade…
>
> —Amit Singhal, Google's director of search

See: https://www.youtube.com/watch?v=mmQl6VGvX-c

# RDFa, Microdata, …

- Different protocols for marking up web pages
- schema.org
  - shared vocabulary
  - used by Google, Bing, Yandex, etc.
  - powers rich result snippets

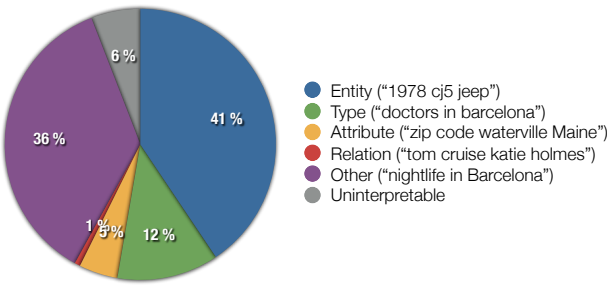



# Entity Retrieval

# Entity retrieval

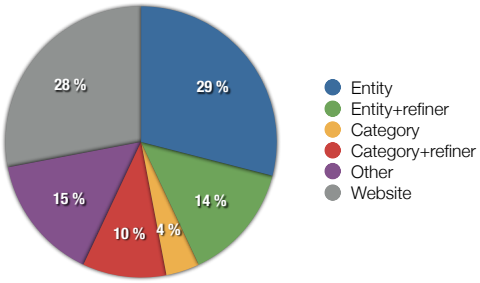
*Addressing information needs that are better answered by **returning specific objects** (entities) instead of just any type of documents.*

# Distribution of web search queries (Pound et al., 2010)




Pound, Mika, and Zaragoza (2010). Ad-hoc object retrieval in the web of data. In WWW '10.

# Distribution of web search queries (Lin et al., 2011)




Lin, Pantel, Gamon, Kannan, and Fuxman (2012). Active objects. In WWW '12.

# Entities

- Objects (or "things") with
  - Unique identifier
  - Name(s)
  - Attributes and/or description
  - Type(s)
  - Relationships to other entities

# Ranking Entities with Ready-made Descriptions

# "Entity homepages"

# Document-based entity representations

- Most entities have a “home page”
- I.e., each entity is described by a document
- In this scenario, ranking entities is much like ranking documents
  - unstructured
  - semi-structured

# Using Language Models

- Standard document retrieval methods applied on entity description documents
  - Just replacing *d* with *e*

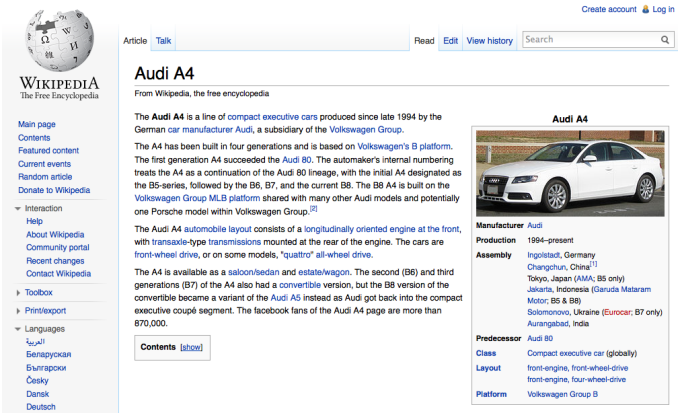
$$P(e|q) \propto P(e)P(q|\theta_e) = \underbrace{P(e)}_{\text{Entity prior}} \prod_{t \in q} \underbrace{P(t|\theta_e)}_{\text{Entity language model}}^{n(t,q)}$$

**Entity prior**
Probability of the entity being relevant to *any* query

**Entity language model**
Multinomial probability distribution over the vocabulary of terms

# Semi-structured entity representation

- Entity description documents are rarely unstructured
  - Different sections, fields, etc.



# How to rank entities in knowledge bases?

**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<br>2001<br>2005<br>2008 |
| **rdf:type** | **dbpedia-owl:MeanOfTransportation**<br>**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** |

# How to rank entities in knowledge bases?

- Represent entities as **fielded documents**

# Fielded models

- Fielded extensions of document retrieval methods
- E.g., Mixture of Language Models (MLM)

$$P(t|\theta_d) = \sum_{j=1}^{m} \mu_j P(t|\theta_{d_j})$$

**Field weights**

$$\sum_{j=1}^{m} \mu_j = 1$$

**Field language model**
Smoothed with a collection model built from all document representations of the same type in the collection

# Setting field weights

- Heuristically
  - Proportional to the length of text content in that field, to the field's individual performance, etc.
- Empirically (using training queries)
- Problems
  - Number of possible fields is huge
    - It is not possible to optimise their weights directly
- Entities are sparse w.r.t. different fields
  - Most entities have only a handful of predicates

# Predicate folding

- **Idea**: reduce the number of fields by grouping them together
- Grouping based on
  - type
  - manually determined importance

# Predicate Folding

| | | |
|---|---|---|
| **Name** | **foaf:name** | Audi A4 |
| | **rdfs:label** | Audi A4 |
| **Attributes** | **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<br>2001<br>2005<br>2008 |
| **Out-relations** | **rdf:type** | **dbpedia-owl:MeanOfTransportation**<br>**dbpedia-owl:Automobile** |
| | **dbpedia-owl:manufacturer** | **dbpedia:Audi** |
| | **dbpedia-owl:class** | **dbpedia:Compact_executive_car** |
| | **owl:sameAs** | **freebase:Audi_A4** |
| **In-relations** | is **dbpedia-owl:predecessor** of | **dbpedia:Audi_A5** |
| | is **dbpprop:similar** of | **dbpedia:Cadillac_BLS** |

# Entity Resolution

- Need to replace entity URIs with their names
  - so that they become "searchable" terms

| | | |
|---|---|---|
| **Name** | **foaf:name** | Audi A4 |
| | **rdfs:label** | Audi A4 |
| **Attributes** | **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<br>2001 **Mean of transportation**<br>2005<br>2008 |
| **Out-relations** | **rdf:type** | **dbpedia-owl:MeanOfTransportation**<br>**dbpedia-owl:Automobile** |
| | **dbpedia-owl:manufacturer** | **dbpedia:Audi** |
| | **dbpedia-owl:class** | **dbpedia:Compact_executive_car** |
| | **owl:sameAs** | **freebase:Audi_A4** |
| **In-relations** | is **dbpedia-owl:predecessor** of | **dbpedia:Audi_A5** → **Audi A5** |
| | is **dbpprop:similar** of | **dbpedia:Cadillac_BLS** |