

# Introduction to Knowledge bases

# The limits of text

Information Retrieval focuses on the retrieval of documents from large corpora. Initially, the retrieval was keyword-based

Web Images Video More ▾ Anytime ▾

[jrbn ~# Jacopo Urbani](#)

[www.jacopourbani.it](http://www.jacopourbani.it) ▾

jrbn ~# **Jacopo Urbani** "You must believe in spring" - Bill Evans. Menu Skip to content. Home; ...  
Email: [jacopo AT cs.vu.nl](#) [jurbani AT mpi-inf.mpg.de](#) (Social) Networks ...

[Jacopo Urbani | LinkedIn](#)

[www.linkedin.com/in/jacopo-urbani-78635623](http://www.linkedin.com/in/jacopo-urbani-78635623)

View **Jacopo Urbani**'s professional profile on LinkedIn. LinkedIn is the world's largest business network, helping professionals like **Jacopo Urbani** discover inside ...

[Jacopo Urbani - Image Results](#)



[More Jacopo Urbani images](#)

[Jacopo Urbani Profiles | Facebook](#)

[www.facebook.com/public/Jacopo-Urbani](http://www.facebook.com/public/Jacopo-Urbani) ▾

View the profiles of people named **Jacopo Urbani**. Join Facebook to connect with **Jacopo Urbani** and others you may know. Facebook gives people the power to...

Good ranking  
criterion?

How can we  
diversify the  
results?

Good string  
similarities?

# The limits of text

Nowadays, web search is *entity-based*

Larry Page (2012): “The perfect search engine would really understand everything in the world deeply, give you back kind of exactly what you need” [1]

Machine learning and KR are parts of core business in many Web companies [2]

[1] <http://fortune.com/2012/12/11/fortune-exclusive-larry-page-on-google/>

[2] <https://backchannel.com/how-google-is-remaking-itself-as-a-machine-learning-first-company-ada63defcb70#.nljh17nb5>

# The limits of text

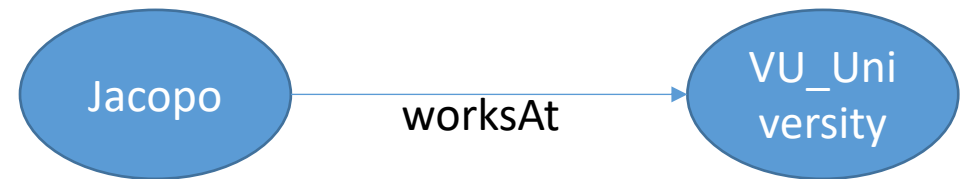
One approach to overcome the limits of text (data) is to build *knowledge repositories* from the Web

- Symbolic Knowledge Bases (KBs)
  - Meaning accessible to humans
  - Typically constructed manually or from unstructured sources
- Latent models
  - Meaning is hidden to us
  - Typically learned using machine learning techniques

# Symbolic Knowledge Bases

- Logic is the language that humans designed to express knowledge
- **Knowledge bases:** crystallization of factual knowledge in the form of associations between entities and relations
  - Can be expressed as first-order logic
  - Recently Google re-branded knowledge bases as ***knowledge-graphs***

worksAt(Jacopo,VU\_University)



# Latent models

- Recently, latent models became very popular due to the rise of *deep learning*
- They achieved impressive results!
- One of the most prominent examples is Google's word2vec [1]

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

# Outline

- We will have one lecture to discuss latent models next week
- We will discuss the theory behind knowledge bases when we focus on logic-based reasoning
- In the remaining, I will introduce some of the largest symbolic knowledge bases available on the Web

# Knowledge bases on the Web

**WordNet**  
A lexical database for English

**yago**   
select knowledge

 **Freebase**<sup>TM</sup>

  
**DBpedia**

  
**WIKIDATA**

 **BabelNet**<sup>2.0</sup>  
A very large multilingual **encyclopedic dictionary** and **ontology**



# WordNet

- Wordnet is the most famous lexical database for English [1]. Project started in 1985 at Princeton by George Armitage Miller
- Groups words into sets of synonyms called *synsets*

## Noun

- [S:](#) (n) **plant**, [works](#), [industrial plant](#) (buildings for carrying on industrial labor) *"they built a large plant to manufacture automobiles"*
- [S:](#) (n) **plant**, [flora](#), [plant life](#) ((botany) a living organism lacking the power of locomotion)
- [S:](#) (n) **plant** (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- [S:](#) (n) **plant** (something planted secretly for discovery by another) *"the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"*

## Verb

- [S:](#) (v) **plant**, [set](#) (put or set (seeds, seedlings, or plants) into the ground) *"Let's plant flowers in the garden"*
- [S:](#) (v) **plant**, [implant](#), [engraft](#), [embed](#), [imbed](#), **plant** (fix or set securely or deeply) *"He planted a knee in the back of his opponent"; "The dentist implanted a tooth in the gum"*
- [S:](#) (v) **plant**, [establish](#), [found](#), **plant**, [constitute](#), [institute](#) (set up or lay the groundwork for) *"establish a new department"*
- [S:](#) (v) **plant** (place into a river) *"plant fish"*
- [S:](#) (v) **plant** (place something or someone in a certain position in order to secretly observe or deceive) *"Plant a spy in Moscow"; "plant bugs in the dissident's apartment"*
- [S:](#) (v) **plant**, [implant](#) (put firmly in the mind) *"Plant a thought in the students' minds"*

[1] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

# WordNet

- Words can be *monosemous* (one meaning) or *polysemous* (multiple meanings)
- Each synset has a gloss (short description) and is connected to other synsets with different linguistic relations. Most important ones are
  - *Hypernyms/Hyponyms (isA)*
  - *Meronym/Holonyms (partOf)*

# RDF

- The World Wide Web Consortium (W3C) has standardized a number of languages to exchange knowledge on the Web [1]
- The **Resource Description Framework (RDF)** is a standard used to report statements that describe properties of resources
  - Properties are represented by IRIs while resources are either special IRIs, labels or special placeholders called *blank nodes*
- The statements can be represented as *triples* of the form  $\langle s \ p \ o \rangle$  (*subject predicate object*) and serialized with different formats: RDF/XML, N3, Turtle
- A RDF dataset can be represented as a directed graph

[1] <https://www.w3.org/2001/sw/wiki/RDF>



# RDF

## Example

IRIs: <http://www.vu.nl>, <http://www.vu.nl#studies>

Literal: “VU University”

blank node: \_:x

*RDF dataset  
(serialized with  
Turtle)*

```
<http://www.vu.nl> <rdf:type> <wikipedia/University> .  
<http://www.vu.nl> <rdf:label> "VU University" .  
_:x <http://www.vu.nl#studies> http://www.vu.nl .
```

# SPARQL

RDF datasets can be stored on a variety of engines

- Native RDF stores (Virtuoso, Sesame, Jena, etc.)
- Relational databases (IBM DB2, Oracle, etc.)

## How do we query a RDF dataset?

The W3C has standardized a specific language called SPARQL [1]

SPARQL is a query language which has a SQL-inspired syntax. Finding answers to a SPARQL query corresponds to find all possible *graph homomorphisms* between the query and the graph

[1] <https://www.w3.org/TR/sparql11-query/>



# SPARQL

## Example SPARQL query

```
SELECT ?X,?Y FROM {  
  ?X <rdf:type> <wikipedia/University> .  
  ?X <rdf:label> ?Y . } }
```

## Answers SPARQL query

```
{?X-> <http://www.vu.nl>  
  ?Y-> "VU University" }
```

## Example Input

```
<http://www.vu.nl> <rdf:type> <wikipedia/University> .  
<http://www.vu.nl> <rdf:label> "VU University" .  
_:x <http://www.vu.nl#studies> <http://www.vu.nl> .
```

# DBpedia

- Project to convert Wikipedia pages into RDF
- Leverages structured content contained in the pages
  - Infoboxes
  - Labels
  - Categories
  - Redirects
  - etc.

## Vrije Universiteit Amsterdam

Vrije Universiteit Amsterdam



[Seal: Maiden in the Garden](#) <sup>[1]</sup>

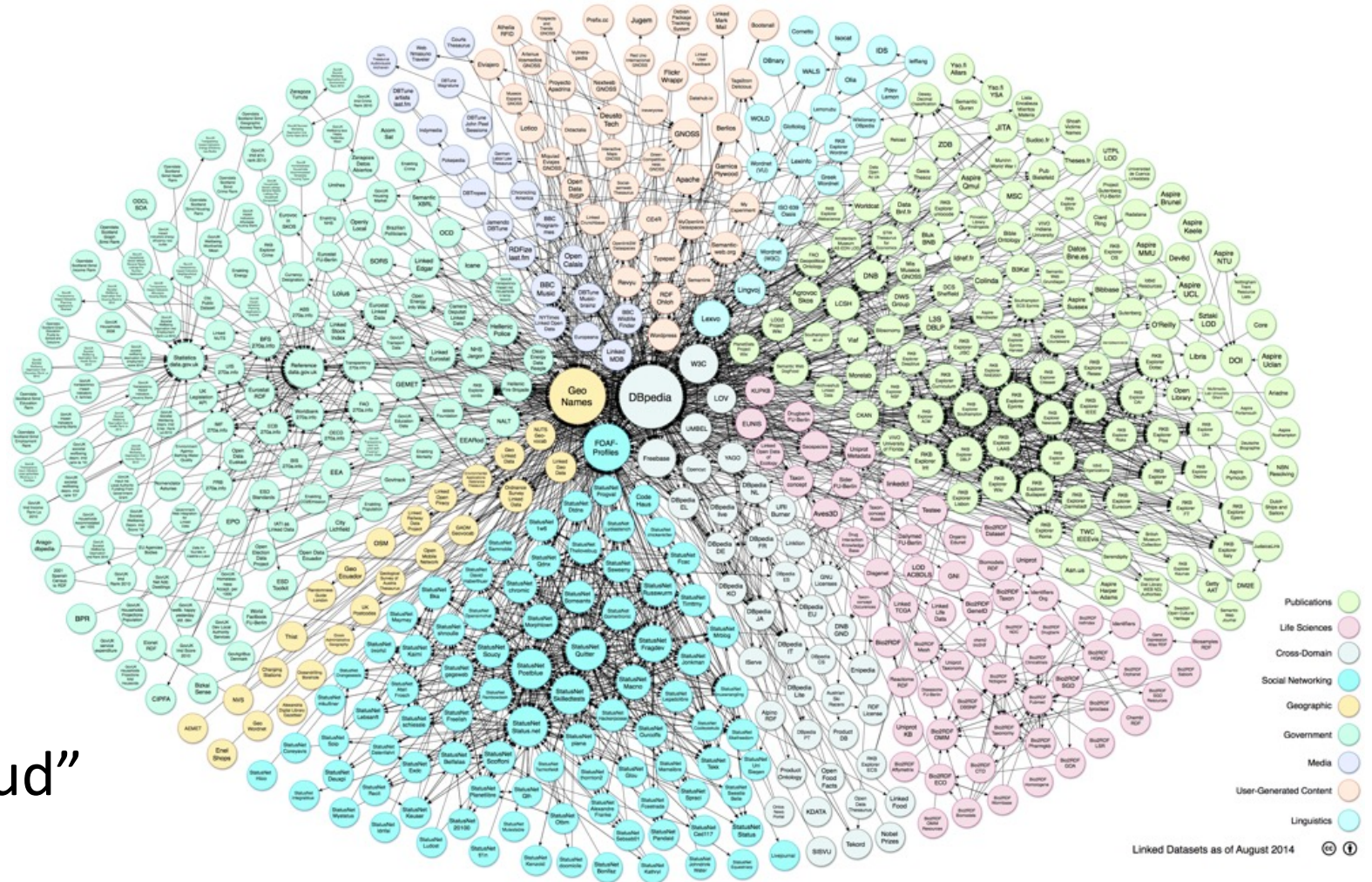
[Latin: Universitas Libera](#)  
(*Reformata Amstelodamensis*)

<b>Motto</b>	Auxilium nostrum in nomine Domini ( <a href="#">Latin</a> )
<b>Motto in English</b>	Our help is in the name of the Lord
<b>Type</b>	<a href="#">Private</a> (publicly funded)
<b>Established</b>	1880 <sup>[2]</sup>



# DBpedia

- Started in 2007, 4.5M entities, 3B RDF triples [1]
- Contains links to other KBs
- Widely popular in the “linked-data-cloud”



[1] C. Bizer *et al.*, “DBpedia-A crystallization point for the Web of Data,” *Web Semantics: science, services and agents on the world wide web*, vol. 7, no. 3, pp. 154–165, 2009.



# DBpedia

- Fairly large ontology but not rich in terms of expressiveness [1]
  - 320 classes
  - 1650 properties
- Alignment between infoboxes and ontologies is done via community-provided mappings

Ontology Class	Instances	Example Properties
Person	198,056	name, birthdate, birthplace, employer, spouse
Artist	54,262	activeyears, awards, occupation, genre
Actor	26,009	academyaward, goldenglobeaward, activeyears
MusicalArtist	19,535	genre, instrument, label, voiceType
Athlete	74,832	currentTeam, currentPosition, currentNumber
Politician	12,874	predecessor, successor, party
Place	247,507	lat, long
Building	23,304	architect, location, openingdate, style
Airport	7,971	location, owner, IATA, lat, long
Bridge	1,420	crosses, mainspan, openingdate, length
Skyscraper	2,028	developer, engineer, height, architect, cost
PopulatedPlace	181,847	foundingdate, language, area, population
River	10,797	sourceMountain, length, mouth, maxDepth
Organisation	91,275	location, foundationdate, keyperson
Band	14,952	currentMembers, foundation, homeTown, label
Company	20,173	industry, products, netincome, revenue
Educ.Institution	21,052	dean, director, graduates, staff, students
Work	189,620	author, genre, language
Book	15,677	isbn, publisher, pages, author, mediatype

[1] J. Lehmann *et al.*, “DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia,” *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

# YAGO

- YAGO (Yet another great ontology) is a project started at Max-Planck Institute for Informatics (2006) [1]
- Goals
  - Unify Wikipedia and Wordnet
  - Exploit Wikipedia Infoboxes to extract clean facts
  - Check the plausibility of facts via type checking
- Last version is YAGO4
  - 67M+ entities
  - 343M RDF facts
- High standards in terms of quality [2]


[1] T. Rebele, F. Suchanek, J. Hoffart, J. Biega, E. Kuzey, and G. Weikum, "YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames," in *The Semantic Web – ISWC 2016*, 2016, pp. 177–185.


[2] <http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/statistics/>

# Freebase

- Freebase was a collaborative Knowledge Base created by its community members
- Initially created by a company called Metaweb, it was later acquired by Google in 2010
- In 2014 Google decided to shutdown Freebase. Part of the data was moved to Wikidata
- 1.9B RDF facts


# Freebase

 Find... [Browse](#) [Query](#) [Help](#) [Sign In or Sign Up](#) [English](#)



Topic

## Eric Clapton<sup>en</sup>

mid: /m/02qwg notable type: /music/artist on the web:  [wikipedia.org](#)

Eric Patrick Clapton, CBE, is an English musician, singer-songwriter and guitarist. He is the only three-time inductee to the Rock and Roll Hall of Fame: once as a solo artist and separately as a member of the Yardbirds and Cream. Clapton has been referred to as one of the most important and influential guitarists of all time. Clapton ranked second in Rolling Stone magazine's list of the "100 Greatest Guitarists of All Time" and fourth in Gibson's "Top 50 Guitarists of All Time". He was also named number five in Time magazine's list of "The 10 Best Electric Guitar Players" in 2009. In the mid-1960s, Clapton left the Yardbirds to play blues with John Mayall & the Bluesbreakers. Immediately after leaving Mayall, Clapton joined Cream, a power trio with drummer Ginger Baker and bassist Jack Bruce in which Clapton played sustained blues improvisations and "arty, blues-based psychedelic pop". For most of the 1970s, Clapton's output bore the influence of the mellow style of JJ Cale and the reggae of Bob Marley. His version of Marley's "I Shot the Sheriff" helped reggae reach a mass market. [-]


Created by mwcl\_musicbrainz on 3/23/2013

Properties

118n

Keys

Links



View and edit specific domains, types, or properties..

Filter options: ☐ Show all domains and properties

Common /common

Topic /common/topic

Also known as /common/topic/alias

Also known as

Clapton, Eric

Eric Clapton

Eric Patrick Clapton

eric\_clapton

Slow Hand

God

Derek and the Dominoes

Blind Faith

The Yardbirds

Types:

Common

Topic

Film

Film music contributor

Film actor

Person or entity appearing in film

Music

Musical Artist

Musician

Record Producer

Composer

Guitarist

Songwriter

Lyricist

# Wikidata

- Wikipedia is mainly text. It's hard to verify and keep consistency
  - e.g., population of Rome is reported both in the Italian and English articles: Numbers are different
- Wikidata is the "data-version" of Wikipedia [1]
  - Data is validated by the community
  - Keeps provenance of the data
  - Multilingual by design
  - Supports plurality



[1] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledge base," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

# Wikidata

- High-quality knowledge!
- Use *qualifiers* to express information that requires arity > 3 [1]

IronLady castMember Meryl\_Streep  
role (qualifier) Margaret Thatcher

[1] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić, “Introducing Wikidata to the linked data web,” in *International Semantic Web Conference*, 2014, pp. 50–65.

# Wikidata

- 1B+ triples
- 54K registered contributors
- 15M+ statements

## Vrije Universiteit (Q1065414)

university in Amsterdam, The Netherlands

 [edit](#)

Free University of Amsterdam | VU

[▼ In more languages](#) [Configure](#)

Language	Label	Description	Also known as
English	Vrije Universiteit	university in Amsterdam, The Netherlands	Free University of Amsterdam VU
Dutch	Vrije Universiteit Amsterdam	universiteit in Nederland	Vrije Universiteit VUA Amsterdamse Universiteit VU University Amsterdam Vrije Universiteit van Amsterdam VU Amsterdam
German	Freie Universität Amsterdam	Universität in den Niederlanden	Vrije Universiteit Vrije Universiteit Amsterdam
French	université libre d'Amsterdam	université des Pays-Bas	Université libre d'Amsterdam Vrije Universiteit Amsterdam

[All entered languages](#)

## Statements