

# Big Open Research Data

---

UNIVERSITY OF MANNHEIM

DATA AND WEB SCIENCE GROUP

ANNA PRIMPELI

# Outline

---

- Requirements of a data scientist
- Open access datasets
  - DBpedia
  - Common Crawl
  - The Web Data Commons project
- Why should we share data?
- How should we share data?

# Data and Web Science group

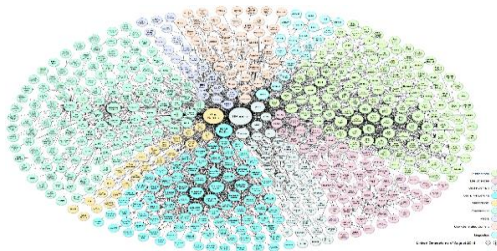
Data profiling

Data  
integration

Data mining

Extract knowledge from data

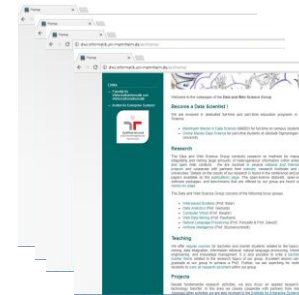
What kind of **data**?



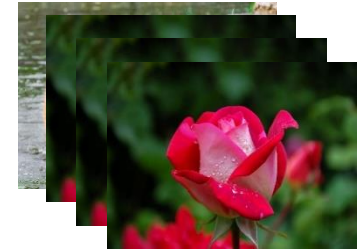
Graphs



Text documents



Webpages

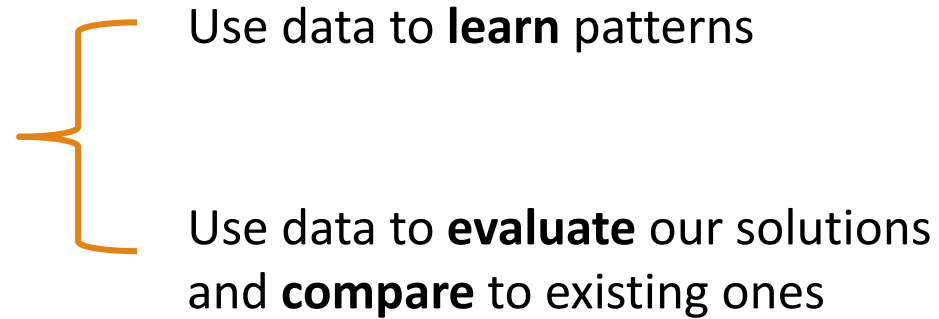


Images

# Data is all we need

---

We need structured data  
to **extract** knowledge



## BUT

- we need a different data corpus for every task
- creating a big data corpus can be too expensive for a single researcher

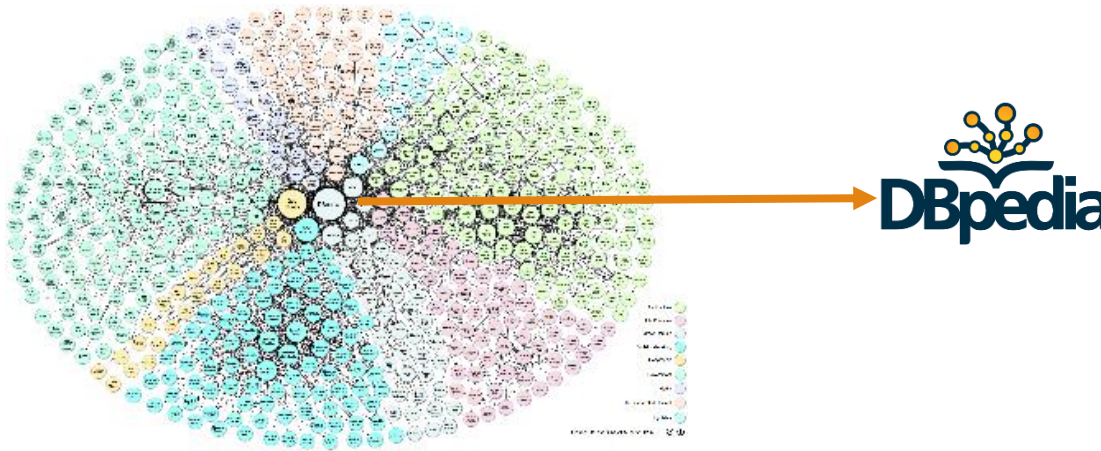
**Data scientists spend 80% of their time on data preparation** (Zhang et al., 2013)

---

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17(5-6), 375-381.

# DBpedia: A big “open access” knowledge graph

---



Extracts infobox data from Wikipedia in 125 languages. The data is linked producing a single graph which provides knowledge of the world.

- Initial release in 2007
- Joint effort of research communities of companies and universities
- Central in the Linked Open Data cloud
- **>20,000** publications listed in Google Scholar that include in their title the term DBpedia
- **50,000** data downloads per month

# Common Crawl: an open large-scale Web corpus

Only big search engine companies had access to realistic snapshots of the Web.

Common Crawl extracts snapshots of the Web on a regular basis and provides the data to the public free of charge.

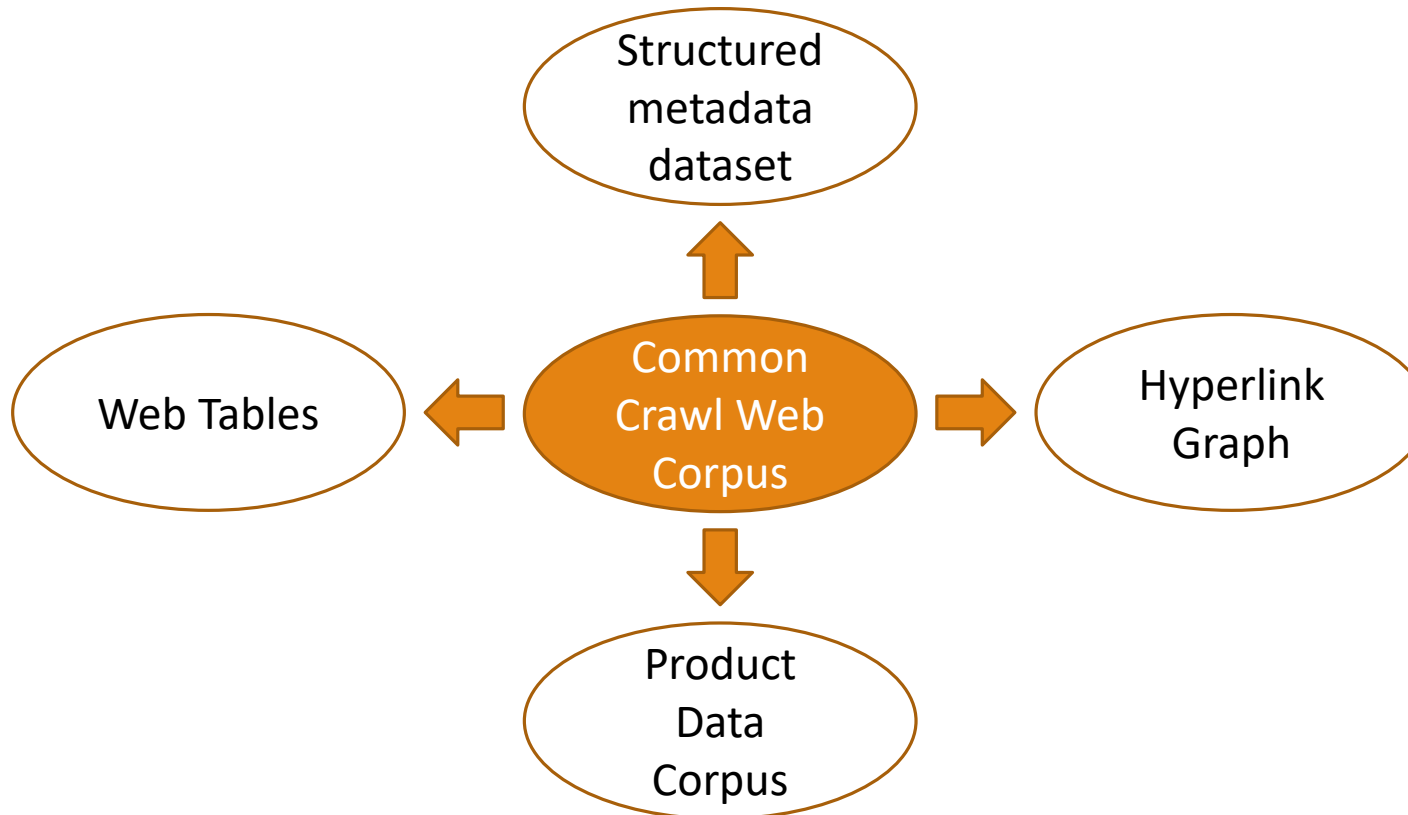


Data offered by Common Crawl the last five years

	2012	2013	2014	2015	2016
Size of Data	40.1 TB	44 TB	64 TB	45 TB	54 TB
#Pages	3 billion	2.2 billion	2 billion	1.7 billion	3.1 billion
#Domains	40 million	12 million	15 million	14 million	34 million

# Creating more „open access“ data

## The Web Data Commons project (2012 - today)



- ✓ Available for download
- ✓ Free of charge
- ✓ Enhanced with metadata
- ✓ Listed in **MADATA**



Web Data Commons  
<http://webdatacommons.org>



# The Web Data Commons structured data corpus

---

## What is it?

A large collection of structured data, like product offers, found on the Web

## What for?

Supports researchers and companies in exploiting the wealth of structured information found on the Web

## Required Resources

Last release:

- 1 TB of structured data
- 650 \$ total cost
- 50 hours of calculations
- 200 working hours

## Who uses our dataset?

- Members of the Web and Data Science group
- Other researchers: Google groups community
- 54 citations of the paper describing the dataset (Meusel et al., 2014)

---

Meusel, R., Petrovski, P., & Bizer, C. (2014, October). The webdatacommons microdata, rdfa and microformat dataset series. In *International Semantic Web Conference* (pp. 277-292). Springer, Cham.



# Big Open Data - Motivation

---

## Effort sharing

- Support the research community
- Different researchers have different areas of expertise

## Collaboration enhancement

- Create better data
- Create better algorithmic methods

## Benchmarking

- Compare different solutions of the same task

## „Honest research“


- Published knowledge relies on data. Your results need to be reproducible.

## Increase of your publications' popularity


- Publishing your data will drive up your citation count

# How should we share the data?


- Publish on the Web
- Enhance with metadata
- Include the relevant data sources in the publications
- Inform the related research community with announcements in relevant mailing lists
- Add in data repositories

 **UNIVERSITÄT  
MANNHEIM**

**MADATA**  
Mannheim Research Data Repository



1. Bizer, Christian and Meusel, Robert and Primpeli, Anna (2016) *Web Data Commons - RDFa, Microdata, and Microformat October 2016 Data Set*. [Dataset]
2. Bizer, Christian and Meusel, Robert and Primpeli, Anna (2015) *Web Data Commons - RDFa, Microdata, and Microformat November 2015 Data Set*. [Dataset]
3. Bizer, Christian and Petrovski, Petar and Meusel, Robert and Primpeli, Anna (2015) *Web Data Commons - Product Data Corpus*. [Dataset]
4. Petrovski, Petar and Primpeli, Anna and Meusel, Robert and Bizer, Christian (2015) *Web Data Commons - Gold Standard for Feature Extraction*. [Dataset]



**Thank you very much for your attention!**