Fachhochschule
Dortmund
University of Applied Sciences and Arts

Data Science

03: How to get data?

Klaus Kaiser
WiSe 2024 / 2035

Data Science

Recap: Different data categories

category	**description**
discrete	finite or countable number of possible values
continuous	every value, especially values in between, are possible
nominal	names or categories with no natural order
ordinal	names or categories with natural order, but distances not interpretable
interval	numbers, where you can interpret distances
ratio	given with a natural zero
qualitative	categorical data with quality focus
quantitative	data where intensity is measured

Data Science

Recap: Data interchange formats

Data interchange formats

For the exchange of data, we need a standardized format, such that sender and receiver can understand the data.

- Data interchange format gives syntax and structure to define how data is represented.
- Concrete format is defined by the application.

Common formats for the exchange of tabular data

- CSV (Comma Separated Values)
- JSON (JavaScript Object Notation)

Data Science

Recap: Can we trust data?

6 dimensions of data quality

To measure trust in data, the Data Management Association of the UK defines six dimensions of data quality.

1. **Accuracy:** Degree indicating the level to which the given data describe the situation
2. **Completeness:** Proportion of data compared to completeness
3. **Consistency:** Absence of differences in different representations of values
4. **Timeless:** Degree, which describes how up-to-date the data is for use
5. **Uniqueness:** Absence of duplicates
6. **Validity / conformity:** Degree, which describes how data conforms to its defined syntax

Data Science

Introduction: Motivation

The essence of a data science project is data - the question is, how to get it?

Data Science

Introduction: Motivation

1. What is the question / problem we consider?
   - What is the goal?
   - For what do we need data for?
2. How can we answer this question / address the problem?
   - What type of data is needed?
3. Where do we find the data needed?

Data Science

Introduction: Primary vs. secondary data

Primary vs. secondary data

It is common to distinguish between **primary data** and **secondary data**. While primary data is generated for the investigated purpose, secondary data is obtained from different sources and created for different purposes.

Data Science

Introduction: Primary vs. secondary data

	Primary data	**Secondary data**
Data	Real time data	Past data
Process	Very involved	Quick and easy
Source	Surveys, observations, experiments, questionnaire, personal interview, etc.	Government publications, websites, books, journal articles, internal records etc.
Cost-effectiveness	Expensive	Economical
Collection time	Long	Short
Specific	Always specific to the researcher's needs.	May or may not be specific to the researcher's need.
Available in	Crude form	Refined form
Accuracy and Reliability	More	Relatively less

Table: Comparison of primary and secondary data, adapted from [1].

Data Science

Introduction: Primary vs. secondary data

What about data science?

In a data science project, one often works with secondary data. Often one uses already given data to automate processes or use external sources to enrich the data given.

- Depending on the need of data, a data source could deliver primary and secondary data

Exercise

The goal is to build a tool predicting the need of maintenance of industrial facilities.

Task: What primary and secondary data could be available?

Data Science

Introduction: Motivation

Last lecture: **Can we trust data?**

Box 1: Misconceptions Due to Data Capturing

- A population database contains all individuals in a population.
- The population covered in a database is well defined.
- A database contains complete information for all its records.
- All records in a database are within the scope of interest.
- Each individual in a population is represented by a single record in a database.
- There are no duplicate measurements in a database.
- Records in a population database always refer to real people.
- Errors in personal data are not intentional.
- Certain personal details do not change over time.
- Coding systems do not change over time.
- Data definitions are unambiguous.
- Temporal data aspects do not matter.
- The meaning of data is always known.
- Missing data have no meaning.
- All records in a database were captured using the same process.
- All attribute/field values are correct and valid.
- Data values are in their correct attributes/fields.
- Data validation rules produce correct data.
- All relevant data have been captured.
- Automatically collected data are always correct, complete, and valid.
- Population data provide the same answers as survey data.
- Data are always of value.
- Hardware and software used to capture data are error free.

P. Christen, R. Schnell (2024). When Data Science Goes Wrong: How Misconceptions About Data Capture and Processing Causes Wrong Conclusions . Harvard Data Science Review, 6(1).

One way to get around these issues

If possible, data scientists should aim to get involved in the capturing, processing, and linking of any data they plan to use for their work, while those developing data-capturing, processing, and linkage systems should collaborate closely with data scientists. It is crucial for successful data science projects to form multidisciplinary teams with members skilled in data science, statistics, domain expertise, as well as 'business' aspects of research (Jorm, 2015).

P. Christen, R. Schnell (2024). When Data Science Goes Wrong: How Misconceptions About Data Capture and Processing Causes Wrong Conclusions. Harvard Data Science Review, 6(1).

Data Science

Introduction: Ways to obtain data

Data could be ...

- ... captured, e.g. by measures or surveys
- ... retrieved, e.g. by selecting from a database
- ... collected, e.g. by crawling websites or using APIs
- ...

Note that the categorization in capturing, collecting and retrieving is done for presentation purpose only!

Capturing data

Data Science

Capturing data: Motivating example

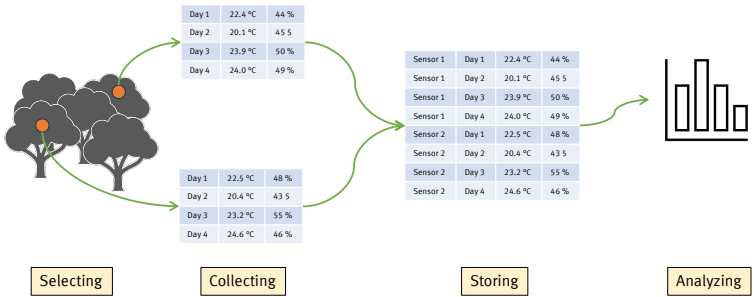


Figure: Example of how a data capturing process could work: Analyzing the climate of a forest by measuring temperature and humidity.

Data Science

Capturing data: Getting data?

Depending on what is investigated, there are different ways to capture the corresponding data.

- Capture data with the help of sensors
- Capture data with the help of cameras, microphones etc.
- Performing observation, interviews, questionnaires, schedules or surveys

Data Science

Capturing data: Example: Sensor data

Figure: A weather station near Sonnenkappe (Harz)

Weather station

Consider a weather station measuring humidity and temperature at its location.

- Sensor are continuously measuring humidity and temperature
- Data is captured in specific intervals, e.g. every second, and stored
- Resulting data are time series

Data Science

Capturing data: Time-series data

Time-series data

Time series data is a series of data points collected or recorded at successive points in time.

- For example, created due to measurements or observations
- It is created rapidly: a lot of data is created
- Two successive data points are often similar
- Actual data is of most interest, old data is of less importance

Data Science

Capturing data: Example

Task: Name different examples for time-series data!

Data Science

Capturing data: Getting data?

Depending on the task it could be useful to perform a **full census** or sampling a **sample**.

- **Full census:** All statistical units of the basic population are considered.
- **Sample:** A subset of the basic population is taken as sample.
  - Sample could be computed randomly, e.g. uniform sample

Data Science

Capturing data: Getting data?

It is common to distinguish between **experiments** and **survey**.

- **Experiment:** Data is captured by creating it new
- **Survey:** Data is captured by asking questions

Data Science

Capturing data: Exercise

Exercise

Consider the example from the first lecture: Creating a tool recognizing monkeys.

Task: What is *basic population* and *sample*?

Retrieve data

Data Science

Retrieve data: Motivating example

In many cases, data is already collected and stored. The question is, how to retrieve this data from its storage.

- Querying data from databases
- Requesting data from a REST-API

In the following, we also take a look on two principles which make retrieving data easier:

- FAIR and open-data

Retrieve data

Databases

Data Science

Databases: Relational databases

Recap: Lecture Datenbanken 1

- Concept of relational databases
- Creating and working with SQL databases
- SQL query language

Attention!

To obtain data from a database it is important to keep this in mind:

- In data science one needs as much data as possible, often taken from many tables. This could result in large and complex queries.
- Working with live data often means, that data is taken from databases used in production.

Data Science

Databases: Data-warehouse

Data-warehouse

A data-warehouse is a central database, which is optimized for analytical purposes. The data-warehouse brings data from different sources condensed together.

- Data is distinguished in facts and dimensions:
  - Facts: Variables which should be measured / investigated (e.g. prices, sales etc.)
  - Dimensions: Variables which describe facts or put them into a context (e.g. country, store etc.)
- Data is mostly given in flat data scheme, e.g. a star- or snowflake schema
- The database-system is optimized for analytics, thus it is optimized on processing complex select queries, and less optimized on insert / update queries
- ETL (extract, transform, load) process brings data from source system to data-warehouse.

Data Science

Databases: Disclaimer

The following is a brief overview on different database concepts. For more details on relational databases and NoSQL databases consider taking the lectures

- Datenbanken 2
- Moderne Datenbanken

Data Science

Databases: Boundaries of relational database

Big data

Large amounts of data are often called *big data*, if they need special treatment (compared to traditional data-processing methods) due to *volume*, *velocity* and *variety*.

- Volume: quantity of the data (size)
- Velocity: speed the data arrives
- Variety: type and nature of the data

Often also:

- Veracity: truthfulness and reliability of the data
- Value: worth for the company
- Validity: correctness of the data

Data Science

Databases: Boundaries of relational databases

Task: Consider your daily life, where does big data occur?

Data Science

Databases: Boundaries of relational database

Due to the properties of *big data*, relational databases are reaching some boundaries.

- Transaction and consistency orientation
  - A database should always give a consistent view on the data
- Relational scheme
  - Database model can not be adjusted in a flexible way
  - It is not possible to use arbitrary data (e.g. images, videos, audio etc.)
  - Complex data-structures are not possible (Object-relational impediment)

Data Science

Databases: Boundaries of relational database

To handle a huge load or lots of data, it could be necessary to scale the database

Vertical scaling

Database is scaled by adding new resources to the server, e.g. new processor or additional memory.

- Limited by memory and processors

Horizontal scaling

Database is scaled by distributing data or requests on multiple servers.

- Complex for highly related data
- See CAP-theorem for more details

Data Science

Databases: NoSQL databases

NoSQL

NoSQL databases do not follow a relational approach. Therefore, they do not require a structured table scheme and try to avoid relational techniques like joins.

- Databases which break with ”traditional” SQL databases
- Relational databases: good all-rounder! NoSQL databases: experts for their special purpose
- Often: focus on horizontal scaling and special types of data

Data Science

Databases: NoSQL databases

Examples

- *Key-Value:* Stores to a key a specific value (object)
- *Document-orientated:* data is organized in documents (often: JSON objects)
- *Graph-oriented:* data is stored as graphs
- *Vector-database:* stores to a vector a specific value (object). Works with distances to find similar vectors
- *Time-series-databases:* optimized to store time/measure pairs which come in rapidly

In the following: One example: document-oriented databases

Data Science

Databases: Document-oriented databases

Document

A document is complete for the purpose it is needed for, i.e. containing all relevant information.

- A document is self-describing for the intended purpose
  - No normalization
- No relations between documents are needed
- Documents are semi-structured
  - They show a structure, but this structure is not uniform
- Documents could contain deeply nested information

Data Science

Databases: Document-oriented databases

Muster Rechnung

Perfekt aussehende Rechnungen und Online- Zahlung für schnelle Begleichungen.
Für kleine Unternehmen entwickelt.

Rechnung

Musterfirechnung GmbH
Emil-Figge-Str. 42
44227 Dortmund

RECHNUNG Nr. 4129444
DATUM 12.07.2024

Alle
Klaus Kaiser
Musterstr. 22
52062 Aachen

LIEFERADRESSE:
Name des Empfängers Klaus Kaiser
Musterstr. 22

Kommentare oder besondere Anweisungen:
Lieferung so schnell wie möglich

VERKÄUFER	P.O. NUMMER	ANFORDERER	VERSANDWEG	FRÜ ion Bord-PUNKT	BEDINGUNGEN
Herr Keiser	1	749	11		Fällig bei Erhalt

MENGE	BESCHREIBUNG	EINZELPREIS	SUMME
1	Rechnung für Vorlesung	12,99 €	12,99 €
		ZWISCHENSUMME	12,99 €
		MEHRWERTSTEUER	0,90€
		VERPACKUNG UND VERSAND	1 €
		GESAMTSUMME	14,90€

Alle Schecks auszahlbar an Firmenname
Wenden Sie sich mit Fragen zu dieser Rechnung an Name, Telefonnummer, E-Mail-Adresse

Vielen Dank für Ihre Bestellung!

Task: Which information could be found on the invoice?

Figure: Example of an invoice.

Data Science

Databases: Document-oriented databases

Muster Rechnung

Perfekt aussehende Rechnungen und Online- Zahlung für schnelle Begleichungen.
Für kleine Unternehmen entwickelt.

Rechnung

Musterfirechnung GmbH
Emil-Figge-Str. 42
44227 Dortmund

RECHNUNG Nr. 4129441
DATUM 12.07.2024

Alle
Klaus Kaiser
Musterstr. 22
52062 Aachen

LIEFERADRESSE:
Name des Empfängers Klaus Kaiser
Musterstr. 22

Kommentare oder besondere Anweisungen:
Lieferung so schnell wie möglich

VERKÄUFER	P.O. NUMMER	ANFORDERER	VERSANDWEG	FREI bis Bord-PUNKT	BEDINGUNGEN
Herr Kaiser	1	749	11		Fällig bei Erhalt

MENGE	BEZEICHNUNG	EINZELPREIS	SUMME
1	Rechnung für Vorlesung	12,99 €	12,99 €
		ZWISCHENSUMME	12,99 €
		MEHRWERTSTEUER	0,90€
		VERPACKUNG UND VERSAND	1 €
		GESAMTSUMME	14,90€

Alle Schecks ausstellen an Firmenname
Wenden Sie sich mit Fragen zu dieser Rechnung an Name, Telefonnummer, E-Mail-Adresse

Vielen Dank für Ihre Bestellung!

Figure: Example of an invoice.

Task: Which information could be found on the invoice?

- Receiver
- Shipper
- Shipping address
- Payment information
- Invoice items
- Prices, …

The invoice can be uniquely identified with the invoice number. For the purpose ”paying”, all relevant information are given.

Data Science

Databases: Document-oriented databases

- A key is assigned to a document (semi-structured object)
- Documents contain nested attribute-value pairs without referential integrity
- Structure of documents is schema-free, i.e. any attributes can be used in any document

Often used formats for storing a document: JSON, XML, BSON, ...

Examples:

MongoDB, Terrastore, Amazon DynamoDB, Microsoft Azure Cosmos DB, Couchbase, CouchDB and some relational databases offer possibilities to store documents.

Document-oriented databases like MongoDB use their own query language, which differ from SQL.

Data Science

Databases: Boundaries of relational databases

Task: Where could document orientated databases be used?

Document-oriented databases are well scalable, if all relevant information is stored in a single document. They are often used in a field such as e-commerce, IoT, social media etc.

Documents are structured or semi-structured data, **what about unstructured data?**

Blob-storage

Blob-storage or object-storage is a system to store large amounts of data. Files are seen as objects, including additional meta-data and a unique identifier.

Retrieve data

API

Data Science

API Overview

Often:

- Direct database access is not given
  - Direct access could be a security issue
  - Data model could change
- Needed data is available at different institution
  - **Example:** Weather information available at weather services

Many companies or data holder offer a service to access their data. Mostly, this is done by offering a **REST-API**. Note that these APIs often cost or have a rate limit.

REST-API

A Representational State Transfer - Application Programming Interface (REST-API) is a software interface, enabling two different systems to communicate via a network connection.

- The client calls the server, and the server responds with a representation of a resource
- Communication is stateless, meaning every call is a new situation and must contain all necessary information
- Data is often sent and returned in JSON or XML format

HTTP-methods

A REST-API offers HTTP-methods for communication. Some of them are

- **GET:** Get should return a representation of the requested data. It should not induce any change on the data.
- **POST:** Post delivers a resource to the server and requests that this resource is processed.
- **PUT:** Put requests the server to set or update a state with the state given in the body.
- **DELETE:** Delete requests that the server delete its state.

A request is sent to a server by calling a URL. The request contains a header with some information and may include a request body with data.

Data Science

API Overview

The API defines how the data is structured, and which data interchange format is used.
This is often done in a so-called swagger documentation.

Example: `https://petstore.swagger.io/`

Authentication

There are some freely available APIs which can be used, but most APIs require some kind of authentication.

Examples

- HTTP-basic authentication: Username and password is sent in the request header
- API-key: An API-key, generated by the API itself, is sent in the request header
- ...

Data Science

API Examples

There are plenty examples of APIs given on the internet to retrieve information. Here are some examples:

- Sport results API: `https://www.openligadb.de`
- Weather API: `https://openweathermap.org`
- Bahn API: `https://developers.deutschebahn.com`
- Google search: `https://developers.google.com/custom-search/v1/overview?hl=de`
- Spotify API: `https://developer.spotify.com/documentation/web-api`
- ISS location API: `http://open-notify.org/Open-Notify-API/ISS-Location-Now/`
- ...

Retrieve data

FAIR and Open Data

Data Science

FAIR and Open Data: FAIR data

The **FAIR** principles are defined to make data usage sustainable [2]

- **Findability:** Metadata and data should be easy to find for both humans and computers.
- **Accessibility:** It should be clear how data can be accessed, possibly including authentication.
- **Interoperability:** Metadata and data use names and vocabulary, which follow common principles.
- **Reusability:** Metadata and data should be well-described so that they can be replicated and/or combined in different settings

With **FAIR/O data,** FAIR data is meant which are also open available.

The concept of open source is well known in computer science: Source code is published under a license that allows the user to use, study, change and distribute it. A similar concept for data is **open data**.

Open data

Data which is set under a license allowing the data to be *accessible*, *exploitable*, *editable* and *shareable* is called **open data**.

Example

The **OpenStreetMap (OSM)** project creates maps of the world. The resulting maps / data is available under an open data license and can be freely downloaded and used[3].

Data Science

FAIR and Open Data: Why is open data important?

Open data is especially important in science: The availability of data allows for the reproduction and verification of results.

Data Science

FAIR and Open Data: Open data Dortmund

Some governments publish open data due to an open government doctrine. In Germany this initiative started in 2018.

Figure: Logo of the open data portal of Dortmund.

- EU: `https://data.europa.eu`
- Germany: `https://www.govdata.de`
- NRW: `https://open.nrw/open-data`
- Dortmund: `https://open-data.dortmund.de`

Data Science

FAIR and Open Data: Available datasets

As mentioned before, open data is important to reproduce results. Additionally, available datasets provide opportunities to learn and train models for initial tests. Therefore, hundreds of datasets are available for this purpose.

Some famous datasets

- iris[4]: Features of different flowers for distinguish theorem
- MNIST[5]: Images of handwritten numbers for character recognition
- Titanic[6]: Information of passenger of the titanic to predict the survivors
- CIFAR-10 / CIFA-100[7]: Thousands of small images for image classification
- imagenet[8]: Millions of images for image classification.
- coco[9]: Dataset of thousands of images for detecting objects on.

Data Science

FAIR and Open Data: Available datasets

Where to find datasets?

1. Reference in a paper / product: Often, if a freely available dataset was used, it is referenced in the paper or application.
2. Dataset search (by google): `https://datasetsearch.research.google.com`
3. Dataset databases:
   - `https://datahub.io/collections`
   - `http://kaggle.com`

Collecting data

Data Science

Collecting data: Motivating example

Sometimes, data is already available, but it cannot be retrieved directly.

Examples

- Needed data could be given on a website, but the website does not offer an API to retrieve the data.
- Needed data is hidden in log files or other output of programs.

Data Science

Collecting data: Boundaries of relational databases

Task: Which data could be crawled from websites?

Data Science

Collecting data: Data scraping

Data scraping refers to the technique of extracting data from the outputs of other programs.

- Input for scraping is output which is intended to be presented to an end user
- Output is mostly structured, meaning the same information is always given at the same position

Scraping should only be used if no other technique to obtain the data, e.g. APIs, are available.

- Structure of the output could change, thus information is located at a different position
- Scraping could be restricted by the source, e.g. limiting page calls etc.

Data Science

Collecting data: Alert

Attention!

In some cases, scraping a website could be illegal - consider data protection and copyright laws, especially, if the scraped data is intended to be published or used commercially!

Data Science

Collecting data: Example

Largest cities of Germany

For a project we want to extract the *Katasterfläche* from the largest cities in Germany.

Task: Which steps would you perform to obtain the information?

Data Science

Collecting data: Example

Largest cities of Germany

1. Information is available in the Wikipedia
2. Collecting a list of all relevant Wikipedia pages - by hand or automatic:

Die größten deutschen Städte 2020

Die Einwohnerzahlen und der Gebietsstand beziehen sich auf den 31. Dezember 2020.

Rang	Stadt	Einwohnerzahl	Land
1.	Berlin	3.664.088	Berlin
2.	Hamburg	1.852.478	Hamburg
3.	München	1.488.202	Bayern
4.	Köln	1.083.498	Nordrhein-Westfalen
5.	Frankfurt am Main	764.104	Hessen

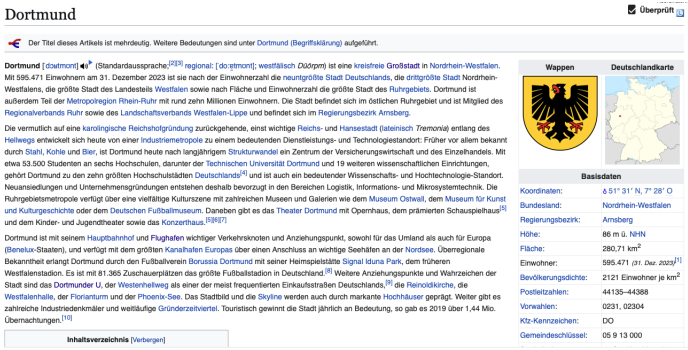
Figure: List of the largest cities in Germany, taken from [10]

Data Science

Collecting data: Example

Largest cities of Germany

3. Pages of the cities have a similar structure, e.g. on the right side is a box with some facts of the city:



Dortmund

Überprüft

Der Titel dieses Artikels ist mehrdeutig. Weitere Bedeutungen sind unter Dortmund (Begriffsklärung) aufgeführt.

Dortmund [ˈdɔʁtmʊnt] (Standardaussprache;[2][3] regional: [ˈdoːɐ̯tmʊnt]; westfälisch *Düörpm*) ist eine kreisfreie **Großstadt** in Nordrhein-Westfalen. Mit 595.471 Einwohnern am 31. Dezember 2023 ist sie nach der Einwohnerzahl die neuntgrößte Stadt Deutschlands, die drittgrößte Stadt Nordrhein-Westfalens, die größte Stadt des Landesteils Westfalen sowie nach Fläche und Einwohnerzahl die größte Stadt des Ruhrgebiets. Dortmund ist außerdem Teil der Metropolregion Rhein-Ruhr mit rund zehn Millionen Einwohnern. Die Stadt befindet sich im östlichen Ruhrgebiet und ist Mitglied des Regionalverbands Ruhr sowie des Landschaftsverbands Westfalen-Lippe und befindet sich im Regierungsbezirk Arnsberg.

Die vermutlich auf eine karolingische Reichshofgründung zurückgehende, einst wichtige Reichs- und Hansestadt (lateinisch *Tremonia*) entlang des Hellwegs entwickelt sich heute von einer Industriemetropole zu einem bedeutenden Dienstleistungs- und Technologiestandort: Früher vor allem bekannt durch Stahl, Kohle und Bier, ist Dortmund heute nach langjährigem Strukturwandel ein Zentrum der Versicherungswirtschaft und des Einzelhandels. Mit etwa 53.500 Studenten an sechs Hochschulen, darunter der Technischen Universität Dortmund und 19 weiteren wissenschaftlichen Einrichtungen, gehört Dortmund zu den zehn größten Hochschulstädten Deutschlands[4] und ist auch ein bedeutender Wissenschafts- und Hochtechnologie-Standort. Neuansiedlungen und Unternehmensgründungen entstehen deshalb bevorzugt in den Bereichen Logistik, Informations- und Mikrosystemtechnik. Die Ruhrgebietsmetropole verfügt über eine vielfältige Kulturszene mit zahlreichen Museen und Galerien wie dem Museum Ostwall, dem Museum für Kunst und Kulturgeschichte oder dem Deutschen Fußballmuseum. Daneben gibt es das Theater Dortmund mit Opernhaus, dem prämierten Schauspielhaus[5] und dem Kinder- und Jugendtheater sowie das Konzerthaus.[5][6][7]

Dortmund ist mit seinem Hauptbahnhof und Flughafen wichtiger Verkehrsknoten und Anziehungspunkt, sowohl für das Umland als auch für Europa (Benelux-Staaten), und verfügt mit dem größten Kanalhafen Europas über einen Anschluss an wichtige Seehäfen an der Nordsee. Überregionale Bekanntheit erlangt Dortmund durch den Fußballverein Borussia Dortmund mit seiner Heimspielstätte Signal Iduna Park, dem früheren Westfalenstadion. Es ist mit 81.365 Zuschauerplätzen das größte Fußballstadion in Deutschland.[8] Weitere Anziehungspunkte und Wahrzeichen der Stadt sind das Dortmunder U, der Westenhellweg als einer der meist frequentierten Einkaufsstraßen Deutschlands,[9] die Reinoldikirche, die Westfalenhalle, der Florianturm und der Phoenix-See. Das Stadtbild und die Skyline werden auch durch markante Hochhäuser geprägt. Weiter gibt es zahlreiche Industriedenkmäler und weitläufige Gründerzeitviertel. Touristisch gewinnt die Stadt jährlich an Bedeutung, so gab es 2019 über 1,44 Mio. Übernachtungen.[10]

Inhaltsverzeichnis [Verbergen]

Wappen	Deutschlandkarte

Basisdaten	
Koordinaten:	51° 31′ N, 7° 28′ O
Bundesland:	Nordrhein-Westfalen
Regierungsbezirk:	Arnsberg
Höhe:	86 m ü. NHN
Fläche:	280,71 km²
Einwohner:	595.471 (31. Dez. 2023)[1]
Bevölkerungsdichte:	2121 Einwohner je km²
Postleitzahlen:	44135–44388
Vorwahlen:	0231, 02304
Kfz-Kennzeichen:	DO
Gemeindeschlüssel:	05 9 13 000

Figure: Screenshot of the Wikipedia article of Dortmund [11]

Largest cities of Germany

4. Underling, the source of the website is HTML, which can be parsed to extract the information needed

```
<tr class="hintergrundfarbe-basis">
    <td>
        <a href="/wiki/Katasterfl" title="Katasterfläche">
    Fläche</a>:
    </td>
    <td>
        280,71 km <sup>2</sup>
    </td>
</tr>
```

Data Science

Collecting data: Data Scraping

- Alternatives to web scraping: listening to data feeds of web servers
- Log scraping: Extracting information from log files
- Screen scraping: Extracting information from screen (e.g. from screenshots)

Summary & Outlook

Data Science

Summary & Outlook: Summary

- You can differ primary and secondary data
- You know what time series data is
- You know ways to get data like APIs, databases (document), crawler
- You know what open data & FAIR data is

Data Science

Summary & Outlook: Outlook

- Some words on data protection
- Starting with statistics!

References

Data Science

Summary & Outlook: List of images

- https://commons.wikimedia.org/wiki/File:Weather_station_near_Sonnenkappe_08.jpg

Data Science

Summary & Outlook: Endnotes

[1] https://researchguides.ben.edu/c.php?g=282050&p=4036581
[2] http://www.go-fair.org/fair-principles/.
[3] https://www.openstreetmap.org/copyright
[4] https://en.wikipedia.org/wiki/Iris_flower_data_set
[5] https://en.wikipedia.org/wiki/MNIST_database
[6] https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html
[7] https://www.cs.toronto.edu/~kriz/cifar.html
[8] https://www.image-net.org
[9] https://cocodataset.org
[10] https://de.wikipedia.org/wiki/Liste_der_größten_deutschen_Städte
[11] https://de.wikipedia.org/wiki/Dortmund