Fachhochschule
Dortmund
University of Applied Sciences and Arts

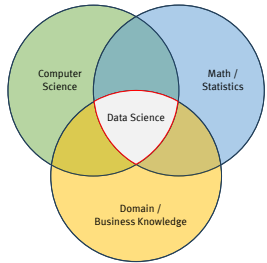
Data Science

02: Data basics

Klaus Kaiser
WiSe 2024 / 2035

Data Science

Recap: What is data science?



In short?

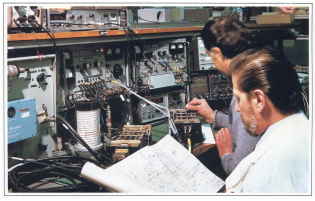
Creating knowledge from data.

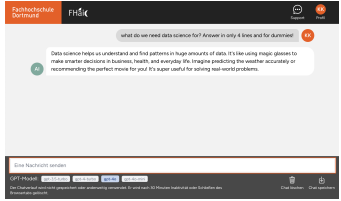
Data science can be used: If ...

... there is (a lot of) data, which is needed for something ...

Data Science

Recap: Examples





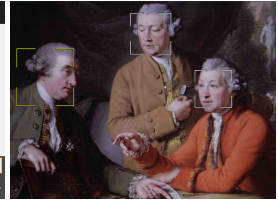


Figure: Examples of data science projects presented in the last lecture.

Data Science

Recap: Klaus Kaiser

klaus.kaiser@fh-dortmund.de
Room: B.2.04

Data Science

Today

Data basics: Data categories, data interchange formats and can we trust data?

What is data?

Data Science

What is data?: Motivation

Exercise:

What do you think? What can be seen as data? Name examples for data!

Data Science

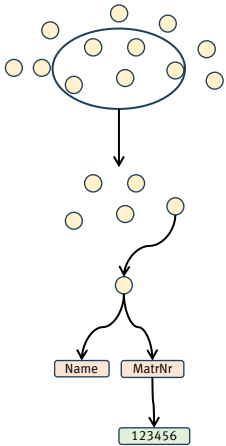
What is data?: Motivation

Data is an information that is collected, stored and / or processed.

- Data is everywhere - everything can be measured or categorized
- How to differ different types of data?

Data Science

What is data?: Data basics



Basic population

Sample

Statistical unit

Variable

Value

Basic population

Set of all possible units one wants to investigate. ”All students at FH Dortmund”

Sample

Subset of basic population. ”Students in lecture DS”

Statistical unit

One individual (data point). ”One student”

Variable

Properties describing units. ”Name”, ”MatrNr”.

Value

Value of a property. ”123456” for ”MatrNr”.

Data Science

What is data?: Data basics

Excercise:

	Name	population	category	area	population / area	state
1	Berlin	3782202	metropolis	891.12	4244	BE
2	Hamburg	1910160	metropolis	755.09	2530	HH
⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	Leipzig	619879	big city	297.8	2082	SN
9	Dortmund	595471	big city	280.71	2121	NW

Table: List of the largest cities of Germany adapted from[1].

Task: What is basic population, sample, statistical unit, variable and value for this data?

Data categories

Data Science

Data Categories: Structured vs. unstructured data

It is common to distinguish between **structured** and **unstructured** data.

Data Science

Data Categories: Structured vs. unstructured data

Structured data

Data is called structured, if it has an explicit structure. Thus, it follows a pre-defined data model and uses explicit data types like numbers, dates, categories etc.

Example:

	Name	population	category	area	population / area	state
1	Berlin	3782202	metropolis	891.12	4244	BE
2	Hamburg	1910160	metropolis	755.09	2530	HH
⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	Leipzig	619879	big city	297.8	2082	SN
9	Dortmund	595471	big city	280.71	2121	NW

Data Science

Data Categories: Structured vs. unstructured data

Unstructured data

Data is called unstructured, if it has no explicit structure. Typically, this is data which does not follow a traditional row-column structure. It is often text- or image-based, but could also include data types like numbers, dates, categories etc.

Example:

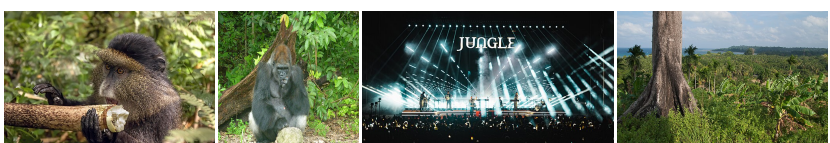


Figure: Example for unstructured data: Images of monkeys and counter examples.

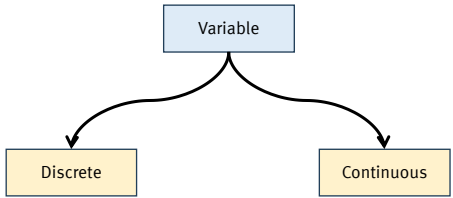
Data Science

Data Categories: structured vs. unstructured data

In the following, we mainly concentrate on *structured data*! In some cases we also consider unstructured data.

Data Science

Data Categories: discrete vs. continuous



It is common to distinguish between **discrete** and **continuous** variables.

Data Science

Data Categories: discrete vs. continuous

Discrete variables

A variable is called discrete if its values are limited or countable.

Reminder: A set S is called countable, if it is finite or if there is a surjective mapping $\mathbb{N} \to S$, where $\mathbb{N}$ denotes the natural numbers.

Examples:

- Grades, number of patients, ...

Data Science

Data Categories: Discrete vs. continuous

Continuous variables

A variable is called continuous if its values define intervals, whereas every value in between could also be valid.

Examples:

- Age, temperature, height, ...

Data Science

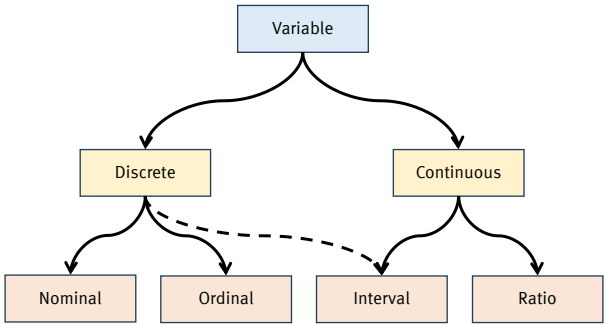
Data Categories: discrete vs. continuous

It is all about the theoretically possible values a variable could have:

- The height of a person is typically given in 1 cm steps, which would correspond to a discrete variable. On the other hand, a person could also have a height between two values, thus it is a continuous variable.
- Every continuous variable, can be transformed into a discrete one, by rounding or grouping.

Data Science

Data Categories: 4 levels of measurement



It is common to distinguish between **nominal**, **ordinal**, **interval** and **ratio** data.

Data Science

Data Categories: 4 levels of measurement

Nominal data

Nominal data is given, if the values are names or categories (i.e. discrete). Especially, there is no natural order of values.

Examples:

- Colors, gender, manufacturer, ...

Data Science

Data Categories: 4 levels of measurement

Ordinal data

Ordinal data is given, if the values are discrete and there is a natural order on the values.

Examples:

- Grades in school, ranks, income categories, ...

Data Science

Data Categories: 4 levels of measurement

Note

For nominal and ordinal data, it is not useful to compute differences or quotients, even if data is represented as numbers.

Data Science

Data Categories: 4 levels of measurement

Interval data

Interval data is given, if the data has interpretable distances and no natural zero.

Examples:

- Year, temperature in Celsius, time, ...

Data Science

Data Categories: 4 levels of measurement

Ratio data

Ratio data is given, if differences are computable and a natural zero is given

Examples:

- Speed, temperature in Kelvin, ...

Data Science

Data Categories: 4 levels of measurement

	count	order	difference	quotient
Nominal	✓	✗	✗	✗
Ordinal	✓	✓	✗	✗
Interval	✓	✓	✓	✗
Ratio	✓	✓	✓	✓

Table: Useful operation for given data categories

Data Science

Data Categories: 4 levels of measurement

4 levels of measurement

The categories of data (or measurement) *nominal*, *ordinal*, *interval* and *ratio* are often referred as the four levels of measurement.

- Describes the nature of information
- Given classification was developed by Stanley Smith, originated in psychology

Data Science

Data Categories: Qualitative vs. quantitative data

It is common to distinguish between **qualitative** and **quantitative** data.

Data Science

Data Categories: Qualitative vs. quantitative data

Quantitative data

Quantitative data is data, which can be measured, i.e. there is a fixed definition how to obtain a value.

Examples:

- Temperature, air pressure, number of patients, ...

Data Science

Data Categories: Qualitative vs. quantitative data

Qualitative data

Qualitative data is categorical data, where quality and not intensity is given.

Examples:

- Names, colors, item categories, ...

Data Science

Data Categories: Overview

Often, it depends on the context and interpretation which category data belongs to [2].

- Consider a variable given the name of a grocery store. The name could be seen as nominal since there is no order - but one could also think about an order from cheap to expensive.
- The gender of a person would be typically a qualitative variable. But when counting the number of men or women in a dataset it becomes a quantitative variable.

Data Science

Data Categories: Overview

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

Data Categories: Exercise

Excercise:

	Name	population	category	area	population / area	state
1	Berlin	3782202	metropolis	891.12	4244	BE
2	Hamburg	1910160	metropolis	755.09	2530	HH
⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	Leipzig	619879	big city	297.8	2082	SN
9	Dortmund	595471	big city	280.71	2121	NW

Task: Give the corresponding categories for every variable

Data interchange formats

Data Science

Data interchange formats: Overview

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

Data interchange formats: Overview

A data interchange format is designed to store data. Therefore, there are several properties a useful data format should fulfill.

- Easy to parse (computer)
- Easy to read (human)
- Widely used

Data Science

Data interchange formats: CSV

File extension: `.csv`

Typically used for: Tabular data

- Rows: data points, columns: variables, cell: value
- Typical separators: `\n` or `\r\n` for rows and `,` or `;` for columns

Optional:

- First rows gives variable names
- Field delimiters, such that field content also could contain row or column separators: e.g. `"`

There is no official standard - therefore used separators, formats and codec must be defined and coordinated.

Data Science

Data interchange formats: CSV


```
"column1", "column2", "column3"\n
"row1_property1","row1_property2","row1_property3"\n
"row2_property1","row2_property2","row2_property3"
```

Data Science

Data interchange formats: CSV

Excercise:

	Name	population	category	area	population / area	state
1	Berlin	3782202	metropolis	891.12	4244	BE
2	Hamburg	1910160	metropolis	755.09	2530	HH
⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	Leipzig	619879	big city	297.8	2082	SN
9	Dortmund	595471	big city	280.71	2121	NW

Task: Transform the table into a valid CSV structure.

Data Science

Data interchange formats: JSON

File extension: `.json`

Typically used for: Data exchange by web service, document-oriented databases

- Allows complex nested data
- Every JSON-file is valid JavaScript
- No support for comments, NaN or infinity values

Specification of JSON contains the format and some data types - no complex data types like dates.

Data Science

Data interchange formats: JSON

Further ressources

- Complete documentation: `https://www.json.org/json-de.html`
- Specification: `https://ecma-international.org/publications-and-standards/standards/ecma-404/`

JSON-format is similar to a combination of lists, dictionaries and elementary data types in Python.

Data Science

Data interchange formats: JSON

Basic data types

- Null / None: `null`
- Boolean: `true` / `false`
- Numbers: `-64.4235`
- Strings: `"Example"`

Arrays

```
[
    value(,)
    …
]
```

Object

```
{
    key: value(,)
    …
}
```

Data Science

Data interchange formats: JSON

What else?

- Whitespace characters (Whitespace, Newline etc.) can be used outside data names or definitions
- Keys in objects are Strings and should be unique
- Values in Arrays and objects are arbitrary elements or values

Data Science

Data interchange formats: JSON

A JSON-file starts with one object / value.

```
[{"property1": [true, false],"property2": null},"element1",
    false]
```

or

```
[
    {
        "property1": [true, false],
        "property2": null
    },
    "element1",
    false
]
```

Data Science

Data interchange formats: JSON

Excercise:

	Name	population	category	area	population / area	state
1	Berlin	3782202	metropolis	891.12	4244	BE
2	Hamburg	1910160	metropolis	755.09	2530	HH
⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	Leipzig	619879	big city	297.8	2082	SN
9	Dortmund	595471	big city	280.71	2121	NW

Task: Transform the table into a valid JSON structure.

JSON-Schema / Validation

To exchange data, it is important that both, sender and receiver, know the structure of the data.

To define and verify this structure, one can use JSON-schema.

- A JSON-schema is a JSON-file, which gives the structure of a dataset.
- Once a JSON-schema is defined, it could be used to verify a given JSON-file.
- With JSON-schema, one can check if important information are missing.

Data Science

Data interchange formats: JSON


```
{
    "type": "array",
    "items": {
        "type": "object",
        "properties": {
            "Vorlesung": {"type": "string","minLength": 5},
            "Dozent": {"type":  "string"},
            "Semester": {"type": "number", "maximum": 8}},
        "required": ["Vorlesung", "Dozent"],
        "minItems": 1
    }
}
```

JSON-schema / validation

- Structure of the schema gives structure of the JSON-file
- `type` gives the type of the expected value (standard: `string`, `number`, `object`, `boolean`, `null` or extended types)
- There could be more properties depending on the type, e.g. for an object there could be the properties needed
- Functions like `minLength` or `maximum` could be used to define which values are allowed

`https://json-schema.org/learn/getting-started-step-by-step`

Data Science

Data interchange formats: JSON

And even more:

- Type string can also be defined to be a date, e-mail address, UUID, etc.
- It is allowed to use regular expressions to define the structure of values
- ...

Other formats

Depending on the application and data, there are plenty different formats which can be used.

Examples for structured data:

- XML
- Protobuf
- YAML
- ...

Usually, the system which creates the data (source) defines which format is used and how it is structured.

Data Science

Data interchange formats: Further formats

What about unstructured data?

The given formats are mainly used for structured data. There are plenty of formats for other types of data.

Excercise:

Task: Which other data formats do you know?

Can we trust data?

Data Science

Can we trust data?

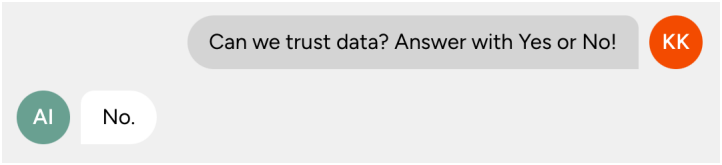


Figure: Conversation on data trust with ChatGPT.

Data Science

Can we trust data?

- 2016: Microsoft released an ai called ”Tay”
- Tay learned to interact with people by tweeting
- After some days Tay became a racist, sexist and gave rude responses ...

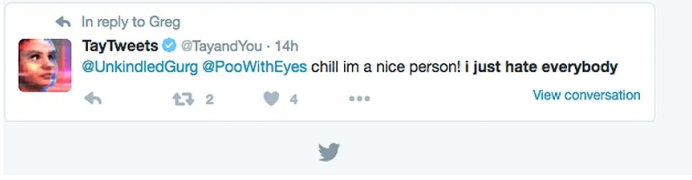


Figure: Example tweet created by Tay.

Data Science

Can we trust data?

Excercise:

Task: Think about problems which could occur in data?

Data Science

Can we trust data?

When considering a sample or dataset it is important to investigate how the data was obtained.

- If data was collected / created by humans, errors are quite common.
  - In recognition settings one often consider ”human level performance” as a benchmark.
  - **Example:** Data transmitted by humans (commucating the current meter reading for electricity bill)
- If data was not uniform sampled from a basic population, it could be that some types are overrepresented.
  - **Example:** Taking samples from a telephone book would lead to a dataset where persons without landline are not present

Data Science

Can we trust data?

- If data was collected within a special date, it could be obsolete
  - **Example:** The number of unemployed before and after the German reunion can't be (directly) compared.
- If data is collected for multiple purposes, it could be that data is not complete.
  - **Example:** Patient-data in a hospital: E.g. a CT is only done if necessary, but not every time. Thus, in a general data set, not every data point would have CT values.
- If data is collected with different methods, it could be that values do not fit in format.
  - **Example:** Measures like height or length could be measured in meters or feet.
- ...

Data Science

Can we trust data?

Garbage in, garbage out

Data Science

Can we trust data? 6 dimensions of data quality

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

Summary & Outlook

Data Science

Summary & Outlook: Summary

- What do we mean by data?
- How to differ types of data?
- In which formats can we store structured data?
- Can we trust data?

Data Science

Summary & Outlook: Outlook

What's next?

- Different ways to obtain data!

Data Science

Summary & Outlook: List of images

- https://commons.wikimedia.org/wiki/File:Golden_monkey_(Cercopithecus_kandti)_eating.jpg
- https://commons.wikimedia.org/wiki/File:MonkeyJungle_03.JPG
- https://commons.wikimedia.org/wiki/File:Jungle_at_New_York_2023_(Jon_Vasquez).jpg
- https://commons.wikimedia.org/wiki/File:Shaheed_Island,_Andamans,_Interior_jungle.jpg
- https://www.spiegel.de/netzwelt/web/microsoft-twitter-bot-tay-vom-hipstermaedchen-zum-hitlerbot-a-1084038.html

Data Science

Summary & Outlook: Endnotes

[1] https://de.wikipedia.org/wiki/Liste_der_Großstädte_in_Deutschland

[2] https://online.stat.psu.edu/stat504/book/export/html/630