Introduction to Data Science

WS 23/24, RWTH Aachen

October 13, 2023

Contents

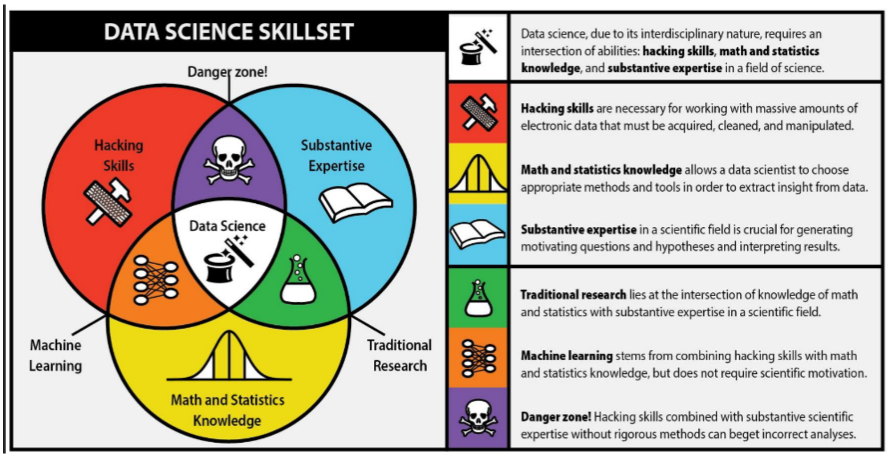
Introduction

Let's first introduce the general term of data science. It is a new and important discipline that can be viewed as:

- An amalgamation of classical disciplines such as statistics, data mining, databases, and distributed systems,
- With additional new challenges constantly emerging an making the field highly dynamic and appealing.

The problems grow in terms of size (”Big data”) and complexity of the questions to be answered. But the basic job can be summarized as:

- Input: data ⇒ Processed by data scientist (with tools) ⇒ Output: value
- Where the skills of a data scientists is the combination of: open mind, human interest, analytical skills, creativity, business-benefiting weighting, . . .
- Or in other terms as can be seen in 0.1



Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills, math and statistics knowledge**, and **substantive expertise** in a field of science.

Hacking skills are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.

Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.

Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

Figure 0.1: Skillset of a data scientist

With the growing importance of data and digitalization, organizations are looking for data scientists, maybe outnumbering computer scientists in the future. Important is the ability to handle data in any form, so basically the need for an all-around skilled ”data wizard”. This importance can be further highlighted when looking at the tech-developement over the past 20 to 30 years. While the hardware got tremendously cheaper, faster, and more compact (20 times faster for MIP = mixed integer programs), also software has progressed in terms of speed (50 times faster for MIP). Interesting to look at is also the aspect of automation.

Dimensions of data science are:

- The different types of data (structured or unstructured, text, images, events, . . .)
- The different types of tasks (supervised or unsupervised, . . .)

- Human versus machine (Who does what?)
- Algorithm versus visualization (What is needed?)
- Flexibility versus usability
- Scalability versus quality (exact versus heuristics)
- Responsibility versus utility (accuracy and precision versus fairness, privacy, transparency, . . .)

Besides raw data science, interesting to look at is also the connection to process science. The interplay between process and data science (PADS) leads to the term of process mining. Imagine the connection as shown in 0.2.

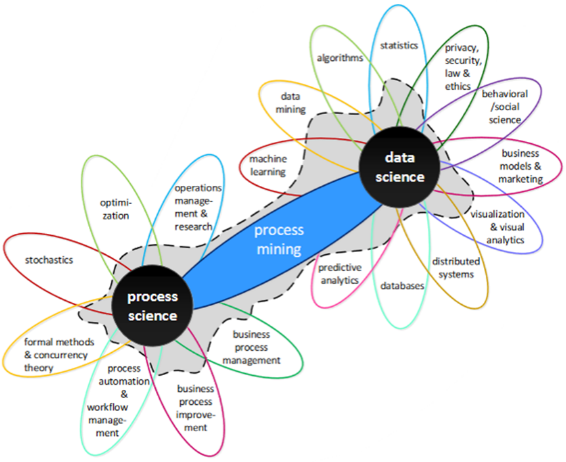

statistics
algorithms
privacy, security, law & ethics
data mining
behavioral /social science
data science
machine learning
business models & marketing
optimi-zation
operations manage-ment & research
process mining
visualization & visual analytics
stochastics
predictive analytics
databases
distributed systems
process science
formal methods & concurrency theory
business process management
process automation & workflow manage-ment
business process improve-ment

Figure 0.2: Interplay between process science and data science

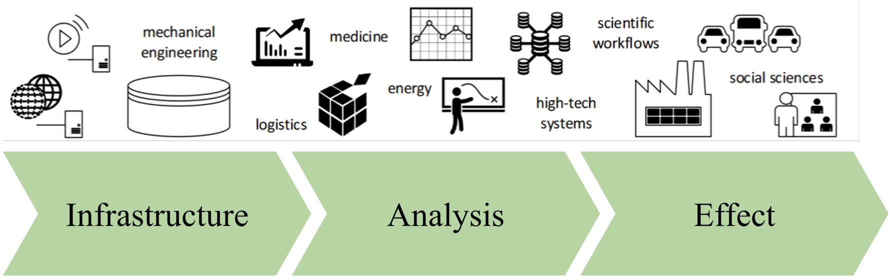
And as the final part of the introduction, we will now see the general covered topics in this course:

- Basic data exploration and visualization
- Decision trees, regression, support vector machines
- Neural networks, evaluation of supervised learning problems, clustering
- Frequent items sets, association rules, sequence mining, process mining, text mining
- Data preprocessing, data quality and binning, visual analytics and information visualization
- Responsible data science
- Big data technologies

1 Basics of data science

1.1 Data science pipeline

First, we are going to look at how data is processed in terms of the **data science pipeline** as it can be seen in 1.1.



Data science pipeline

Figure 1.1: Pipeline of data science

Infrastructure

Let's look at the individual components. The first step to pay attention to when wanting to handle data is the **infrastructure** with the keywords **"volume and velocity"**. The main challenge is making things scalable and instant. Important terms are for example:

- Instrumentation
- Big data infrastructures, distributed systems
- Data engineering (databases and data management)
- Programming
- Security
- Scalability, responsiveness

Analysis

Next, we have the step of the actual **analysis** concerned with **extracting knowledge** from data. The core challenge can be put as providing answers to known and unknown unknowns. Important terms are for example:

- Statistics
- Data and process mining
- Machine learning, artificial intelligence
- Operations research
- Algorithms
- Visualization

Effects

Finally, we also need to be concerned with the **effect** of our results on people, organizations, and society. The main challenge of this pipeline step is to do **responsibly** perform data handling. Important terms are for example:

- Ethics and privacy

- IT law
- Human-technology interaction
- Operations management
- Business models
- Entrepreneurship

This course will look into all the steps of the pipeline, but the main focus lies on the data analysis.

1.2 Four generic data science questions

Important to answering all these questions is to keep attention to all three pipeline steps, so not only what analysis we need to perform to answer them, but also how we collect our input (data) and how to deal with our output (result).

Nonetheless, here are the four generic data science questions, with variety in terms of difficulty and predicting the future:

1. **What** happened?
2. **Why** did it happen?
3. What will happen in the **future**?
4. What is the **best** that can happen?

1.3 Types of data

Now that we know that we have some kind of data as our input, we need to take a look at what this data can look like. Generally speaking, there are two types:

Structured data

Unstructured data

- Structured data like age, time, gender, class, etc., and
- Unstructured data like text, audio, video, etc.

For **structured data** we have a further subdivision into structured data types. The data types depicted in 1.2 will be described in detail.

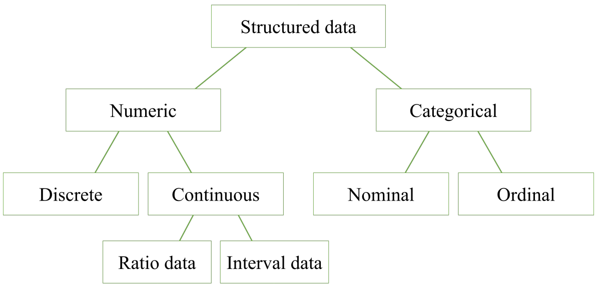


Figure 1.2: Overview structured data types

Categorical data

- **Categorical** data can be stored and identified based on names or labels given to them and is also known as ”qualitative” data. Matching can be applied, where data is grouped based on similarities.

Nominal data

- Concretely, **nominal** data or naming data has a label and its characteristic similar to a noun and doesn’t imply an order.
  - Example: name, color=red, country=NL

Ordinal data

- **Ordinal** data on the other hand is ranked, ordered, or used on a rating scale. This means, you can count and order ordinal data but are not able to measure it.
  - Example: risk=medium, score=good

Numerical data

- In contrast to categorical data, we also have **numerical** data referring to data in the form of numbers instead of another language or descriptive form. It is also known as ”quantitative” data. Important is the ability to be statistically and arithmetically calculated (allowing for $+, -, >, =, \dots$).

Discrete data

- One subtype of numerical data is **discrete** data representing countable items, that are collected in a list (finite or infinite).
  - Example: number of items=5, age=18

Continuous data

- Then, there’s also **continuous** data in the form of intervals or ranges. The data represents measurements with their intervals falling on a number line (so counting isn’t involved).

Interval data

- Continuous data can now be further distinguished. One subtype is **interval** data where the data can be measured only along a scale at equal distances from each other, so only addition and subtraction operations are allowed. There is no true zero (and hence no $\cdot, /$).
  - Example: data=11-11-2018, temp=18.5°C

Ratio data

- And finally, we have **ratio** data describing measurement with a defined (true) zero point.
  - Example: dropout=33%, speed=128.34km/h

For **unstructured data**, we just take the raw data and interpret it as a stream of bits. This goes for text, audio, images, signals, and videos exactly the same. Examples can be seen in 1.3.

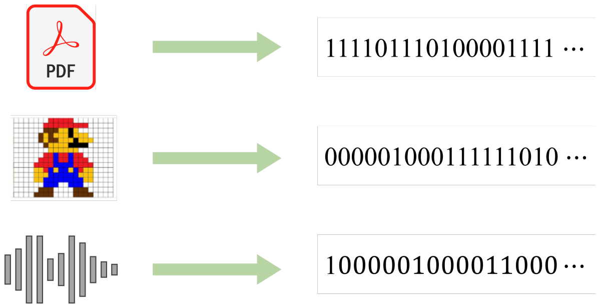
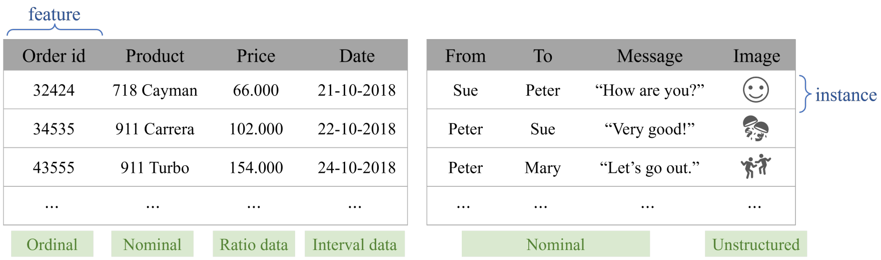


Figure 1.3: Input for unstructured data

Tabular data

Data can now be stored and ordered together by putting it into **tables** . Concretely, columns represent different features (can be different kinds of data types) whereas rows describe data instances. Examples can be seen in 1.4.



feature
instance
Ordinal
Nominal
Ratio data
Interval data
Nominal
Unstructured

Order id	Product	Price	Date
32424	718 Cayman	66.000	21-10-2018
34535	911 Carrera	102.000	22-10-2018
43555	911 Turbo	154.000	24-10-2018
...	...	...	...

From	To	Message	Image
Sue	Peter	“How are you?”	☺
Peter	Sue	“Very good!”	👥
Peter	Mary	“Let’s go out.”	👫
...	...	...	...

Figure 1.4: Table data with data types