Introduction to Data Science

WS 23/24, RWTH Aachen

October 15, 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 and 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 scientist are the combination of open mind, human interest, analytical skills, creativity, business-benefiting weighting, ...
- Or in other terms as can be seen in 0.1

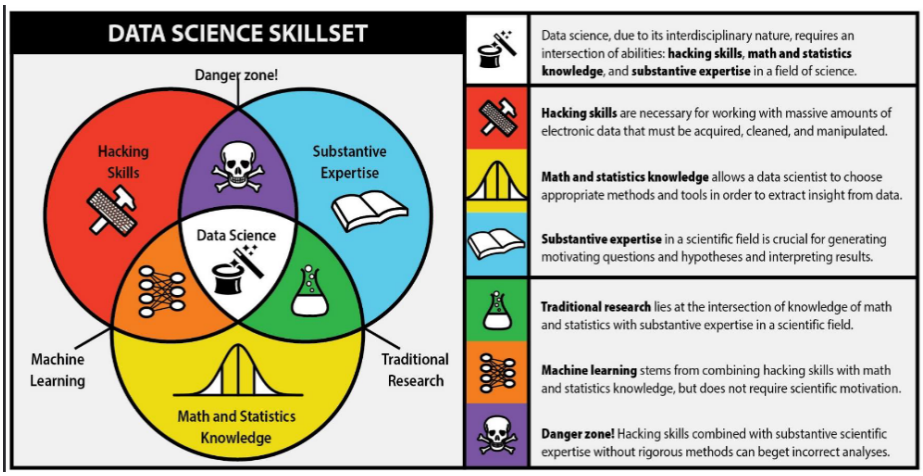


Figure 0.1: Skillset of a data scientist

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 process mining. Imagine the connection as shown in 0.2.

Figure 0.2: Interplay between process science and data science

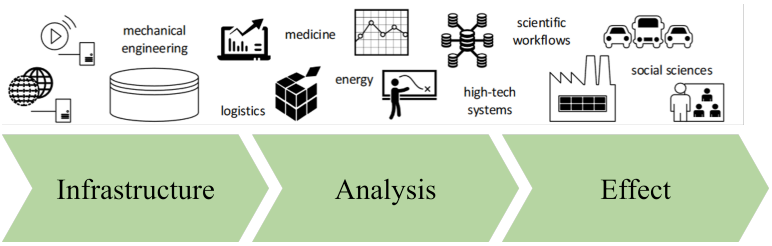
As the final part of the introduction, we will now see the general covered topics in this course:

- 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

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 (responsiveness). Important terms are for example:

Infrastructure

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

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:

Analysis

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

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:

Effects

- Ethics and privacy, and 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
- Structured data like age, time, gender, class, etc., and

Unstructured data
- 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.

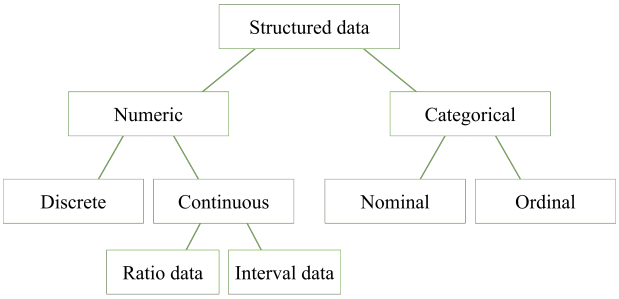


Figure 1.2: Overview structured data types

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)

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

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.

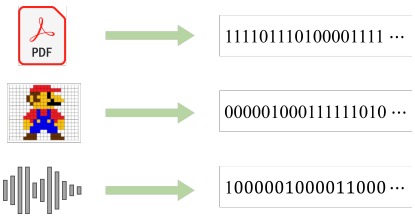


Figure 1.3: Input for unstructured data

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 (also known as individuals, entities, cases, objects, or records). Examples can be seen in 1.4. — Tabular data

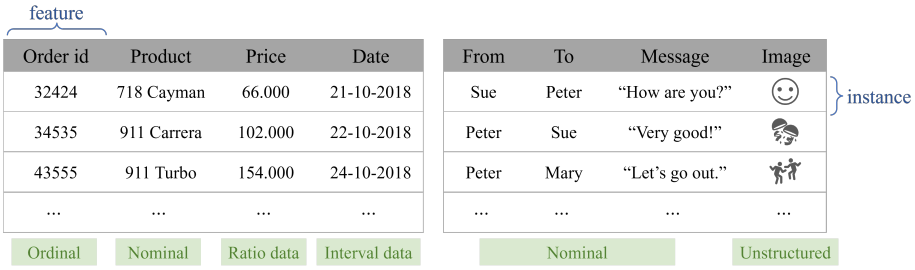


Figure 1.4: Table data with data types

Features can now be raw or derived (e.g. max, min, average, rank, bin, ...). An important aspect is time, as it cannot decrease and we usually want to predict the future based on the past. — Features

An important distinction to be made when it comes to tabular data is whether the items are labeled or not.

- In case of labelled data we have **descriptive features** and a **target feature**. — Labelled data
  - The descriptive features are also known as predictor variables or independent variables. — Descriptive features
  - Alternative names for target features are response variable, dependent variable, or also label. — Target feature
- Unlabelled data on the other hand doesn't have a selected target feature. — Unlabelled data

1.4 Supervised and unsupervised learning

Derived from the different kinds of tabular data, we have two fundamental learning paradigms. Exemplary input data and possible results for both paradigms can be seen

in 1.5.

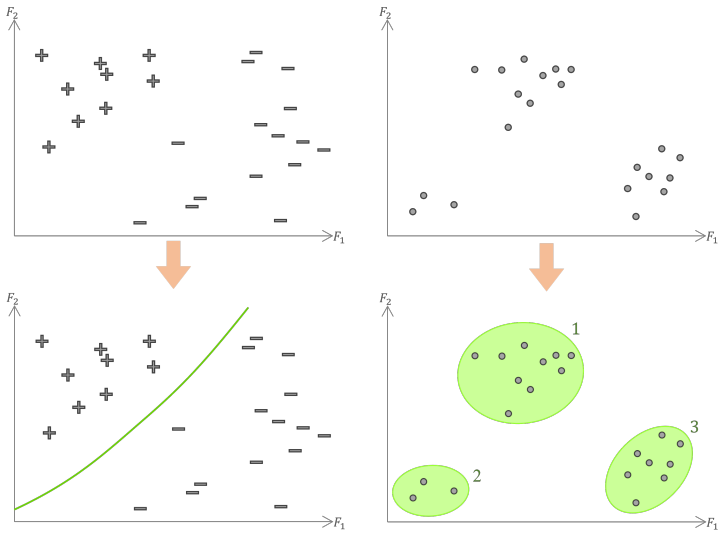


Figure 1.5: Comparing supervised (left) and unsupervised (right) learning

Supervised learning

In the case of labeled data, we can apply **supervised learning** . The goal is to find a "rule" in terms of descriptive features explaining the target feature as well as possible. Examples include:

- Hospital environments where:
  - The target variable can be recover (yes or no), and
  - The descriptive variables can be age, gender, smoking, ...
- University environments where:
  - The target variable can be drops out (yes or no), and
  - The descriptive variables can be mentor, prior education, ...
- Production environments where:
  - The target variable can be order is delivered in time (yes or no), and
  - The descriptive variables can be product, agent, ...

Unsupervised learning

In contrast to labeled data, we can also have instances without target labels, where we can only apply techniques of **unsupervised learning** . The goal is to find clusters or patterns.

Cluster

- **Clusters** are homogeneous sets of instances. Examples include finding similar groups of patients, students, customers, orders, cars, companies, and so on.

Pattern

- **Patterns** on the other hand reveal hidden structures in the data, so basically the unknown unknowns. Rules of some form can be found in many environments and can for example look like this:
  - Customers who buy bread and butter typically pay by phone.
  - Patients who drink and smoke typically pay the hospital bill earlier than others.
  - Products produced by team A on Monday tend to be returned more frequently by customers.

Interesting to regard is process discovery as a form of unsupervised learning in the way that a process model is just a very sophisticated rule. Important to mention, that this

task can get very complex very quickly.

Terminology

Important to see for all of data science: many different names are used to refer to the key disciplines contributing to data science.

- This includes statistics, data analytics, data mining, machine learning, artificial intelligence, predictive analytics, process mining, generative AI, etc.
- Since frequently the same name is used for different concepts (names describe heavily overlapping areas), they really need to be put in context and interpreted accordingly.

The point can be highlighted when looking at scoping machine learning. Here are examples of confusions:

- Sometimes "machine learning" is used as a synonym for "deep neural networks" and sometimes they cover the entire spectrum of learning techniques.
- Neural networks can be used as classifiers. But this doesn't imply that numerous classification techniques developed in data mining are part of machine learning in the narrow sense.

How confusing specifically the arrangement of terms around machine learning is and how fluent and unclear the actual terms are, is depicted in 1.6.

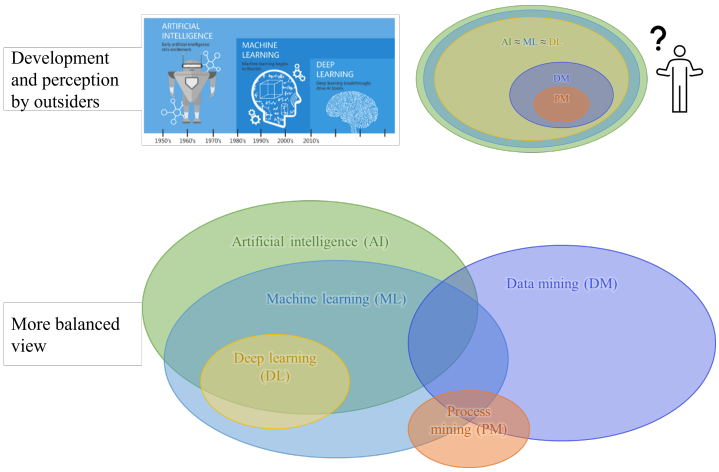


Figure 1.6: Terms around machine learning

1.5 Data science process

There are many different lifecycle models to describe phases in a data science project. This section will give a quick overview of some important ones.

We'll start with **CRISP-DM** which stands for "Cross-industry standard process for data mining". It was developed in the late 1990s by different involved companies (SPSS, Teradata, Daimler AG, NCR Corporation, Ohra). The process consists of multiple steps playing together as visualized in 1.7

CRISP-DM

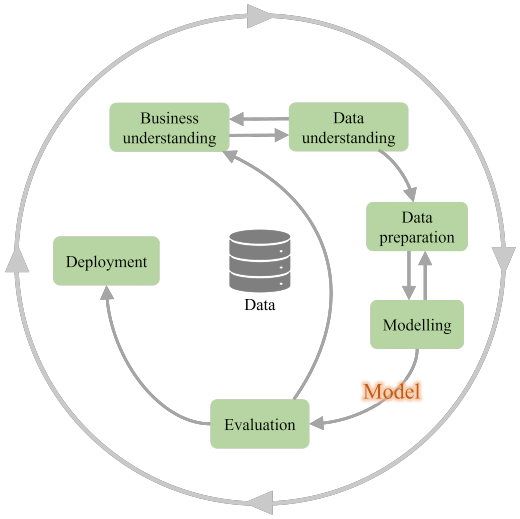


Figure 1.7: CRISP-DM process

Business understanding	
Determine business objective	Background, business objective, business success criteria
Situation assessment	Inventory of resources, requirements, assumptions, constraints, risks, contingencies, terminology, costs, benefits
Determine data mining goal	Data mining goals, data mining success criteria
Produce project plan	Project plan, initial assessment of tools and techniques
Data understanding	
Collect initial data	Initial data collection report
Describe and explore data	Data description, exploration report
Verify data quality	Data quality report
Data preparation	
Starting point: data set	Data set, data set description
Select data	Rationale for inclusion and exclusion
Clean data	Data cleaning report
Construct data	Derived attributes, generated records
Integrate and format data	Merged/reformatted data
Modeling[1]	
Select modeling technique	Modeling technique, modeling assumptions
Generate test design	Test design
Build model	Parameter settings, models, model description
Assess model	Model assessment, revised parameter settings
Evaluation	
Evaluate results	Assessment of data mining results w.r.t. business success criteria, approved models
Review process	Review of process
Determine next steps	List of possible actions settings
Deployment	
Plan deployment	Deployment plan
Plan monitoring and maintenance	Monitoring and maintenance plan

[1]The term "modeling" can be misleading, meant is the selection and assumptions (human) or automated learning by a tool or algorithm

Produce final report	Final report and final presentation
Review project	Experience documentation

Next, we have the **KDD** (Knowledge Discovery in Databases) process as shown in 1.8. Another process model also developed by SAS institute is called **SEMMA** consisting of the phases Sample, Explore, Modify, Model, and Assess.

KDD

SEMMA

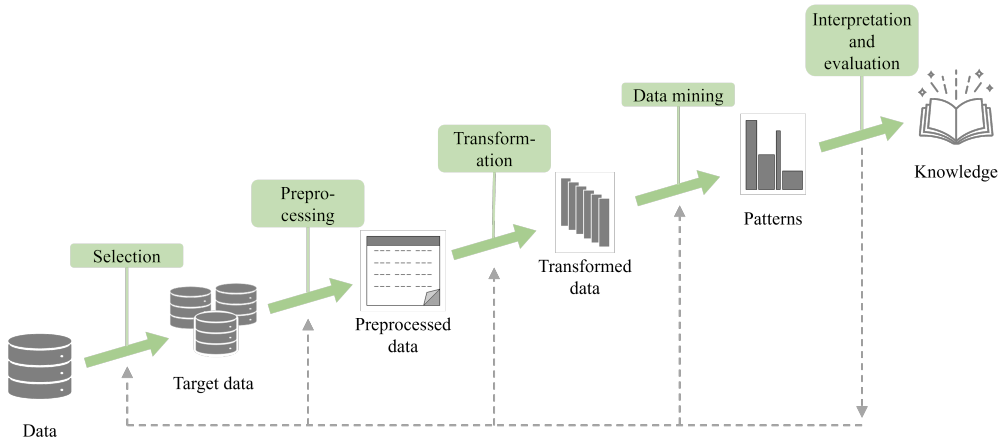


Figure 1.8: KDD process

The next process model is specifically developed for **L* lifecycle model** with multiple stages as shown in 1.9.

L* lifecycle model

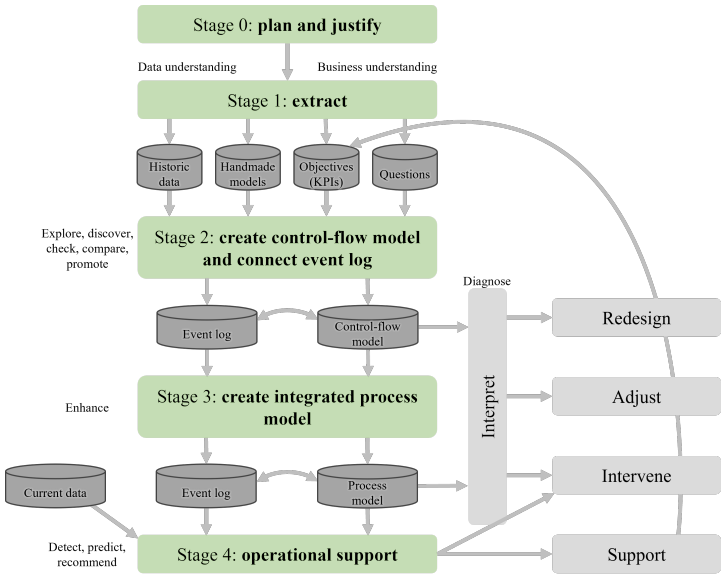


Figure 1.9: L* lifecycle model

Furthermore, we have two methodologies the process model can be related to. Important to implement and solidify are improvements in both.

PDCA

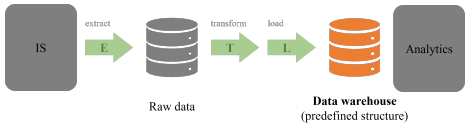
- **PDCA** stands for Plan-Do-Check-Act and is a never-ending cycle with exactly these steps.

DMAIC

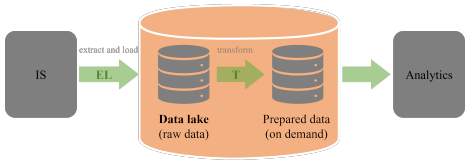
- The other one **DMAIC** stands for Define-Measure-Analyze-Improve-Control, with the following subtasks:
  - Define: launch team, establish charter, plan project, gather VOC/VOB, plan for change
  - Measure: document process, collect baseline data, narrow project focus
  - Analyze: analyze data, identify root causes, identify and remove waste
  - Improve: generate, evaluate, and optimize solutions, pilot, plan and implement
  - Control: control the process, validate project benefits

ETL
ELT

Finally, we have two processes with the same components, but different ordering of the steps as can be seen in 1.10. The short terms for the processes are **ETL** (extract, transform, load) and **ELT** (extract, load, transform).



(a) ETL with data warehouse



(b) ELT with data lake

Figure 1.10: Processes with extraction, transform, and load steps

As a final note on which steps are usually the most time-expensive: there is a so-called "80/20 rule" stating:

- 80% of a data scientist's time is spent on finding, cleaning, preprocessing, and organizing data. This leaves only 20% to actually perform an analysis.
- On the other hand, we have 20% effort determining 80% of the final result.

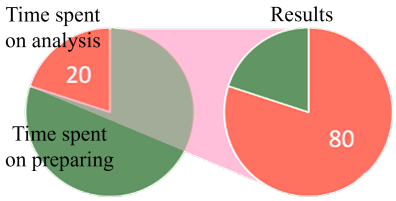


Figure 1.11: 80-20 rule

1.6 Challenges

To finalize the overview and basics of data science, let's look at the typical challenges.

First, we have the challenge of **finding data**. There may be hundreds or thousands of tables, for example in the case of SAP the numbers can easily for up to 800′000. But,

different entities differ in their relevance, meaning some are less relevant than others.

The next challenge is the **transformation of data**, meaning reorganization of data, filtering, extraction of relevant features, and so on. Not only for transformations, but also in general other challenges are **dealing with big data and streaming data**. The challenge of big data evolved over the last few decades, meaning typical stochastic methods try to solve the problem of saying something about entities given only a small amount of samples, whereas now we have a very high load of data, and need to solve the problem of dealing with these large amounts in a correct way. Also for streaming data, new approaches need to be thought of. Additionally, we also need to **deal with a concept shift**.

Another huge challenge is ensuring **data quality**. This goes especially, since our provided data may be incomplete, invalid, inconsistent, imprecise, and/or outdated. Consider for example timestamps. They might be incomplete (event is missing), invalid (e.g. 14-14-2018), inconsistent (14-07-2018 in contrast to 7-14-2018), or imprecise (only regard part of available data: 2018-09-21~~'T'13:00:10~~).

A very typical problem is **overfitting and underfitting** as it can be seen 1.12.

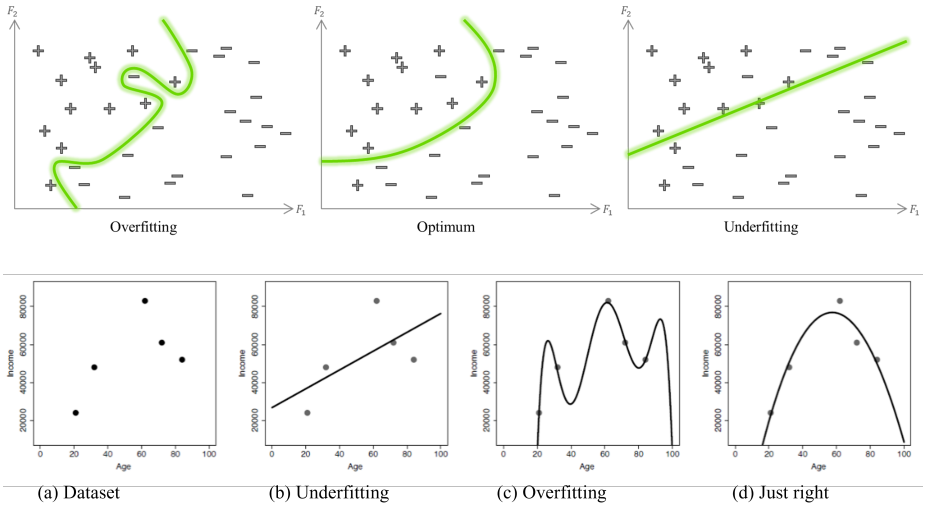


Figure 1.12: Over- and underfitting visualized

The next challenge is the distinction of **correlation and causation**, explicitly that correlation does not imply causation. Consider this example:

- Sunburn and ice cream have a strong correlation. When only these two features are considered, one might derive that either ice cream causes sunburn, or the other way around.
- We know of course, that this is not correct and instead an additional factor causes both phenomena: if the sun is shining, it's warm and people eat ice cream, and also sun directly causes sunburn.

Besides the accuracy of our results, we also need to look into whether our results are valuable. Concretely, **results** should be **made actionable**. This means, that analysis results should be relevant, specific, timely, novel, and clear. Our goal is to go from "data" to "insight" and finally "action". Consider these examples:

- Warning about a traffic jam should come before entering said traffic jam.

- That it's currently raining is not too helpful information. Preferably is a notice ahead of time.

Responsible data science

The last, but very important challenge is **responsible data science** (RDS) . This includes ensuring of:

- **Fairness**, meaning data science should exclude prejudice (How to avoid unfair conclusions even if they are true?)
- **Accuracy**, so data science without guesswork (How to answer questions with a guaranteed level of accuracy?)
- **Confidentiality** (How to answer questions without revealing secrets?)
- **Transparency** (How to clarify answers such that they become indisputable?)