Day 2: Introduction to Data Science

Samuel Schlenker

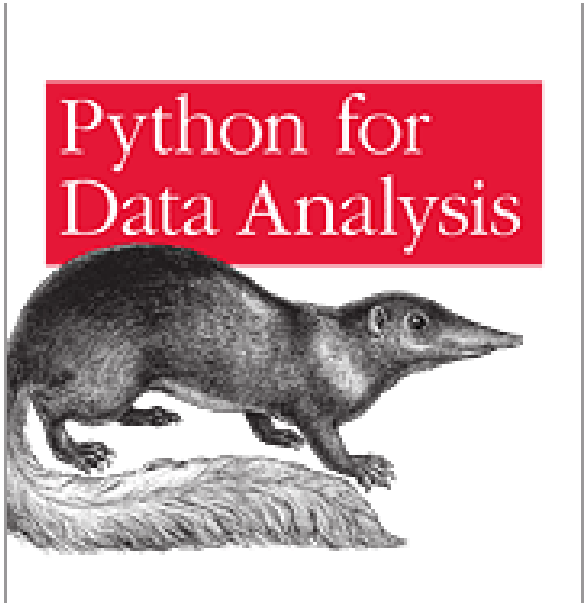
04.11.2025, WWI 2025F

Students should ...

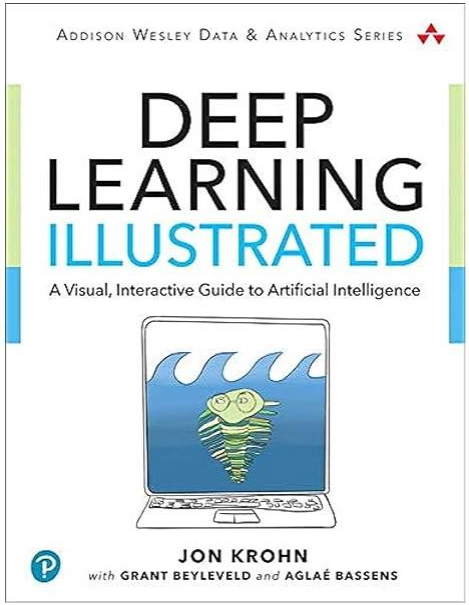
- Understand the fundamental definition and scope of data science as an interdisciplinary field
- Identify the three pillars of data science: domain expertise, statistics/mathematics, and computer science
- Recognize real-world applications of data science across industries (quality control, predictive maintenance, fraud detection, autonomous driving)
- Distinguish between AI, Machine Learning, Deep Learning, and Generative AI
- Understand the characteristics of Big Data (Volume, Velocity, Variety, Veracity, Value)
- Identify different data sources (open-source, private, commercial) and their accessibility
- Apply principles of effective data visualization and storytelling
- Recognize poor or misleading visualizations and understand common pitfalls
- Understand different chart types and their appropriate use cases (line charts, bar charts, scatterplots, heatmaps, etc.)
- Recognize the difference between correlation and causation
- Identify opportunities and risks associated with data science applications

Recommended Reading

Data Science from Scratch, by Joel Grus. O'Reilly



Python for Data Analysis, 2nd ed by Wes McKinney. O'Reilly

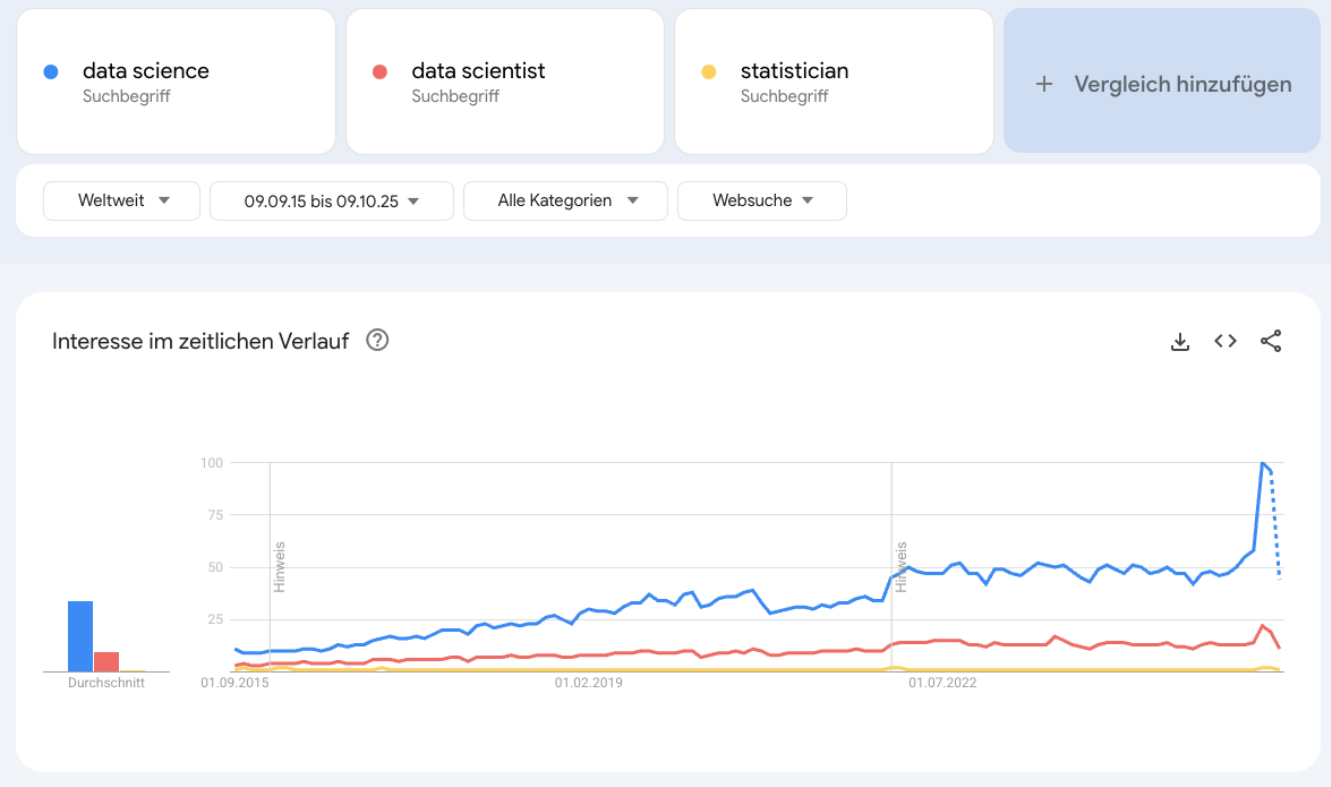


https://www.deeplearningillustrated.com/

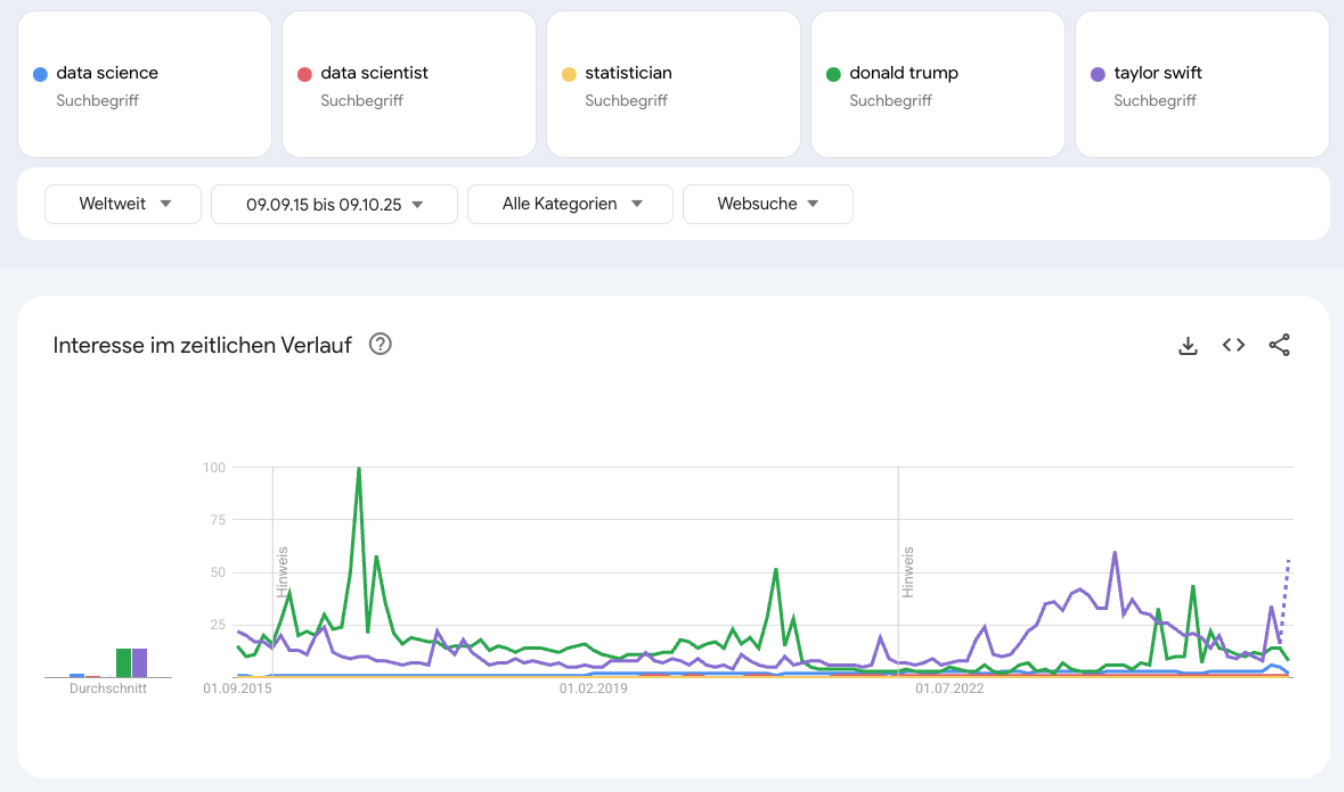
What do you think...

https://www.menti.com/alb5874akgdy

Google Trends



In comparison ...



Facebook knows us better than friends and family

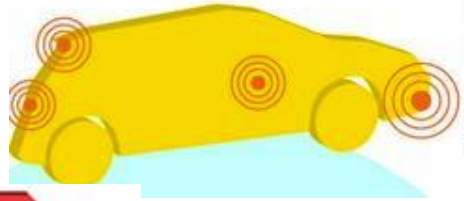
- 2015: Study with 86,220 volunteers
- Collaboration between the University of Cambridge and Stanford University
- Questionnaire with 100 items on the Five Factor Model (FFM) of personality psychology / "Big Five"
- Neuroticism, extraversion, openness to experience, conscientiousness, and agreeableness
- Computer algorithm (linear regression) vs. assessment by individuals
  - From 10 likes: Computer is better than coworkers
  - From 70 likes: Computer is better than friends
  - From 150 likes: Computer is better than family
  - From 300 likes: Computer is better than spouse
- The average Facebook user shares 227 likes

Self-test: https://applymagicsauce.com/demo

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]



Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

40 ZETTABYTES

[43 TRILLION GIGABYTES]

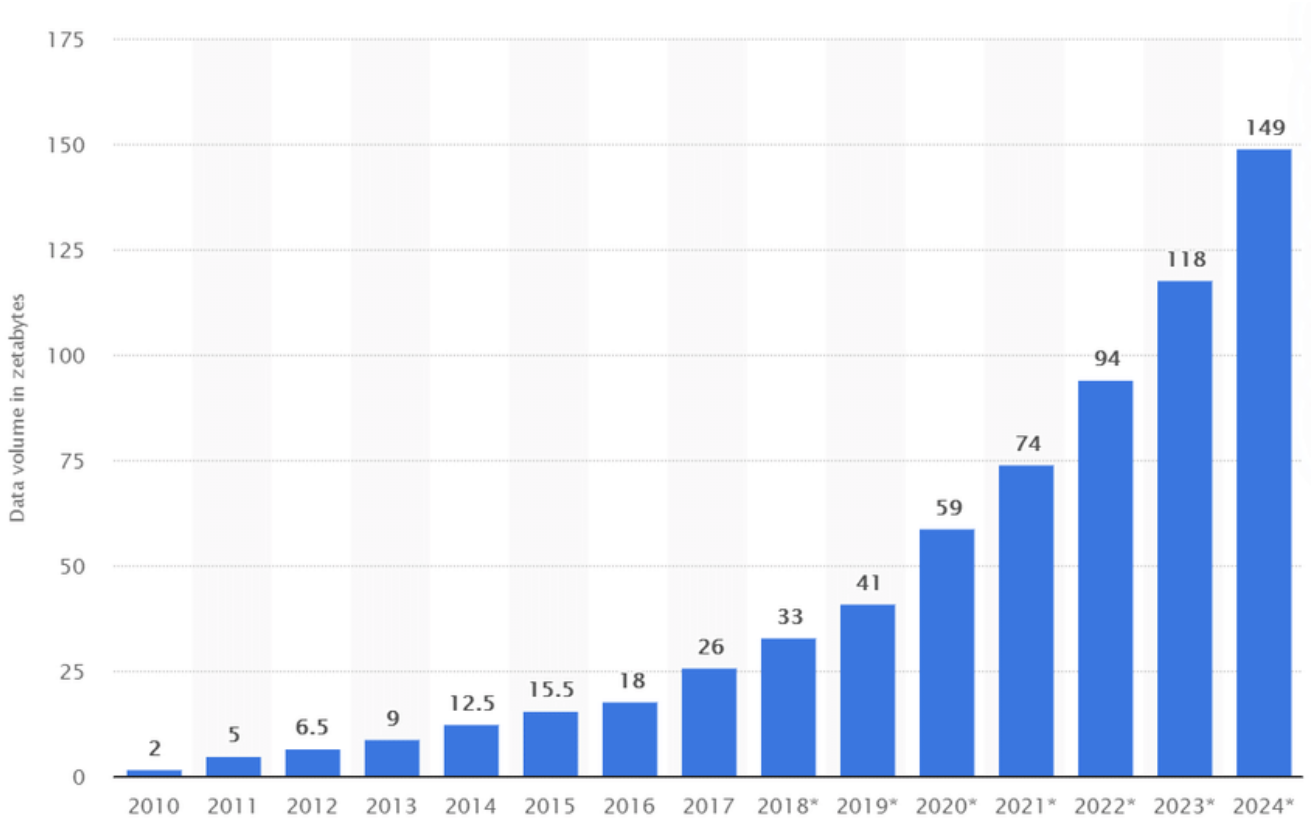
of data will be created by 2020, an increase of 300 times from 2005

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

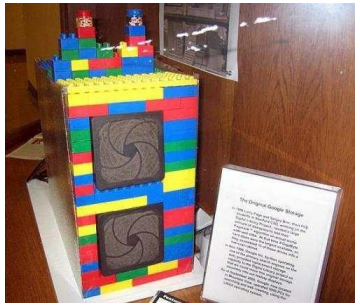
Pure Amount of data



Distributed Computing

This amount of data requires more than one computer.

Google – 1998:



Distributed Computing

This amount of data requires more than one computer.

Google – 2014:


INTRODUCTION TO DATA SCIENCE BY W&S
YOUR TURN!
EXERCISE / QUESTION

Data Science

Needs massive data processing

Distributed computing

Why is this the case?

Data Is Driving Everything

- Modern data acquisition is inexpensive!
    - Smartphones, embedded systems, inexpensive sensors,
    - Medical devices, simulators, …
- Data storage is inexpensive!
- Parallel (compute cluster) computation is inexpensive
    - The Cloud, clusters of computers, GPUs, tensor processors, …
- Science only has explanatory and predictive models in a few (mostly physical sciences-related) domains
- … So: can we use algorithms + data to understand phenomena? Build or augment models? Build detectors? Make diagnoses?

Data Is Driving Everything

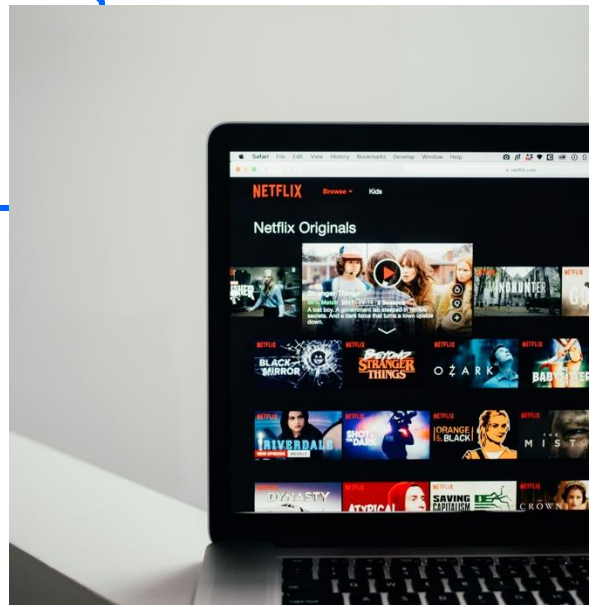
"Big data"
"Data science"
"Data lakes"
"Visual analytics"

"Deep learning"
"Statistical analysis"
"Biomedical informatics"
"Business analytics"

Lots of trends in pursuit of the same goals!
Discovery, models, decision-making, …

Also, new issues -
"Ethical algorithms"
"Reproducibility"

Data Science



Reccomandation engines

Autonomous driving

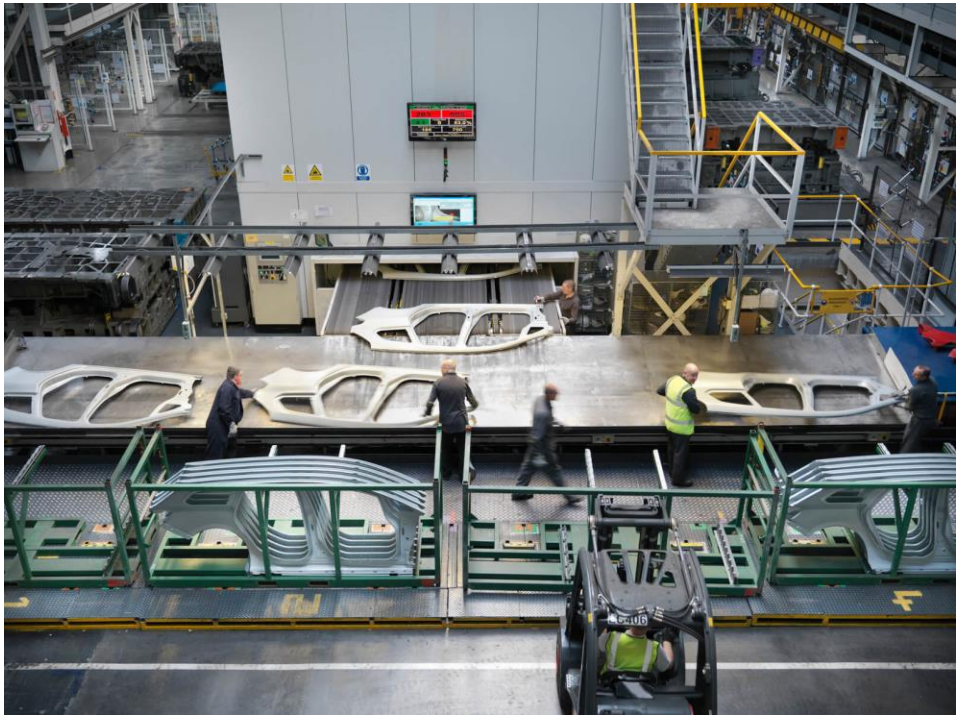
Business intelligence

Applications of data science (Use cases)

Quality control:

- Quantity: Is there the right number of salami slices on the frozen pizza?
- Correct product: Are the right components installed in the server?
- Quality: Are there scratches in the paintwork on the car door? Is the weld seam flawless?
- Results in: Fewer complaints, Cost reduction, Higher customer satisfaction

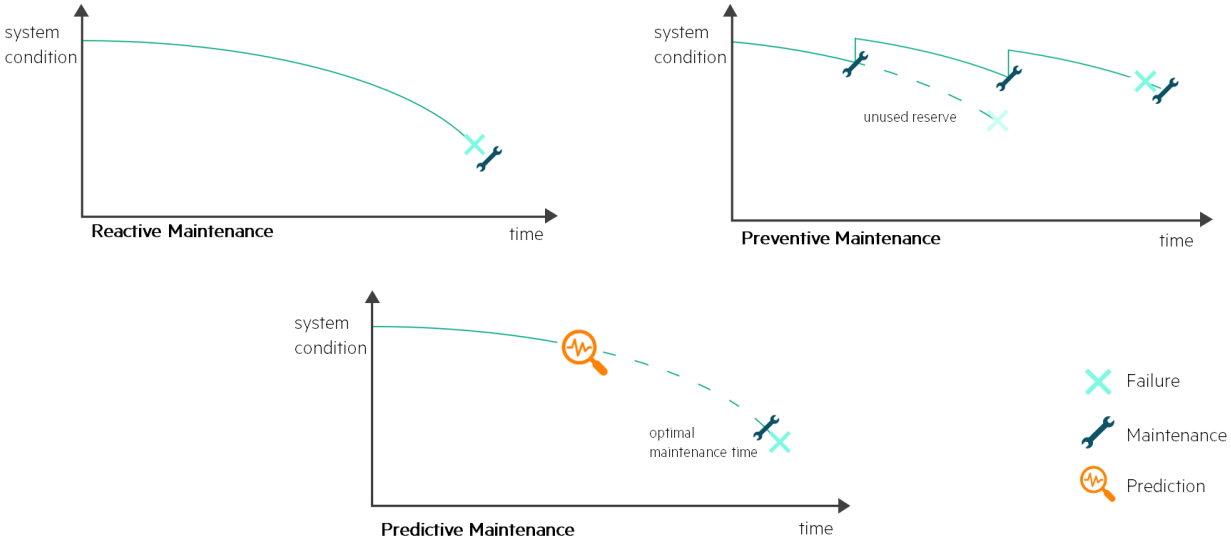
Applications of data science (Use cases)



Predictive Quality: Predicting the quality of the product before the production process is complete.

- Early correction of the product, if possible, or discontinuation of production.
- Reduction of waste.
- Energy, water, and material savings by avoiding finishing steps for poor-quality products.
- Cost reduction.

Applications of data science (Use cases)



Predictive Maintenance:

Predicting the failure of a production machine

- Avoiding unplanned production downtime
- Replacing/repairing the component that is about to fail at the best possible time
- Using the production machine or individual parts for as long as possible

Applications of data science (Use cases)

Fraud Detection:

Detection of credit card fraud, for example, through anomaly detection

- Early account blocking
- Loss minimization

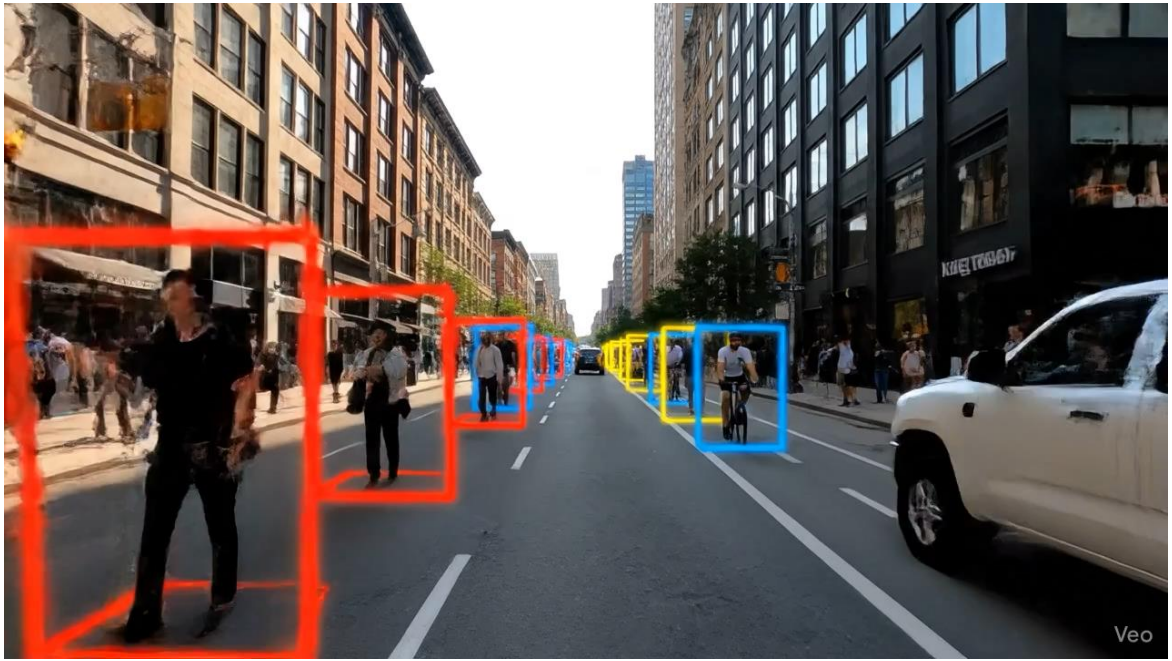
Applications of data science (Use cases)

Personalized recommendations:

For example, shopping cart analysis

- Individualized shopping experience
- Higher customer loyalty
- Higher shopping cart values
- Increased sales

Applications of data science (Use cases)



Autonomous driving:

Detection of pedestrians, their direction of travel, and derivation of the appropriate action

- Traffic sign recognition
- Enables innovation
- Minimization of human error
- Reduction in the number of accidents

Applications by industry

Financial Services

- Fraud Detection
- Risk Assessment
- Algorithmic Trading
- Sentiment Analysis

Healthcare

- Gene Research
- Drug Discovery
- Patient Assistant
- Medical Records

Public Sector

- Urban Planning & Design
- Predictive Policing
- Citizen Engagement
- Document Summarization

Manufacturing

- QC / Defect Detection
- Supply Chain Optimization
- Predictive Maintenance
- Generative Design

Telecommunications

- Customer Support
- Network Optimization
- Predictive Maintenance
- Fraud Detection

Financial Services & Insurance

Public Sector / Defense

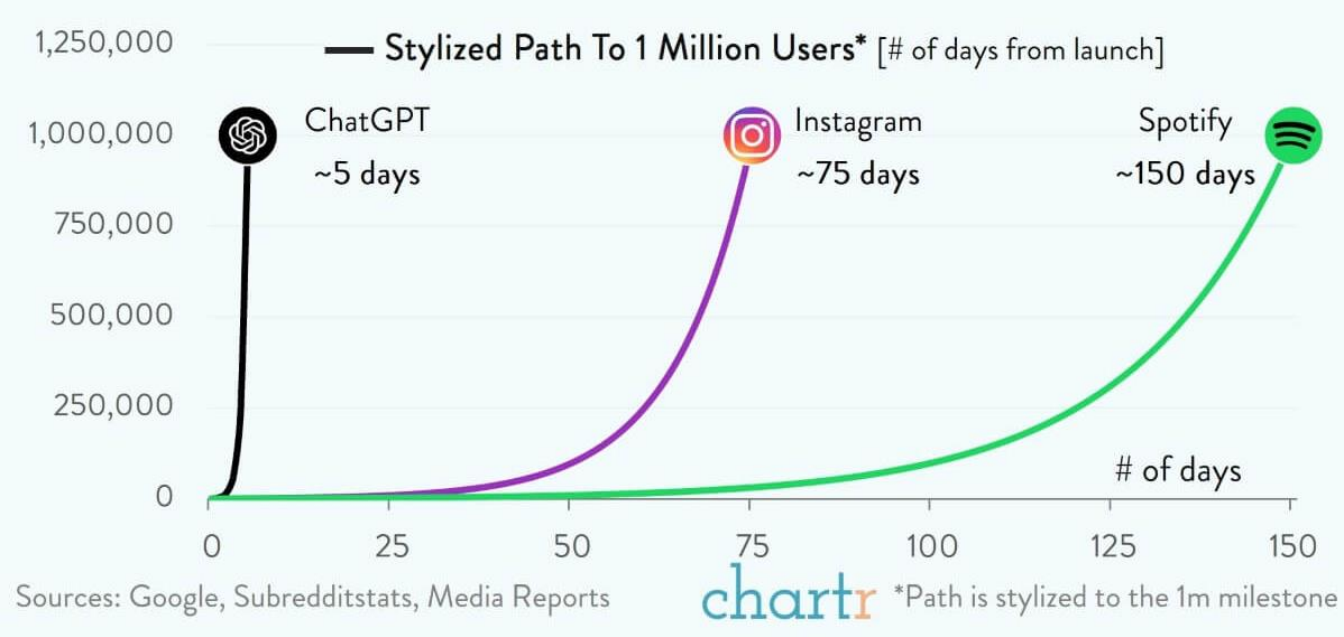
Healthcare / Life Sciences

Autonomous Driving

Manufacturing

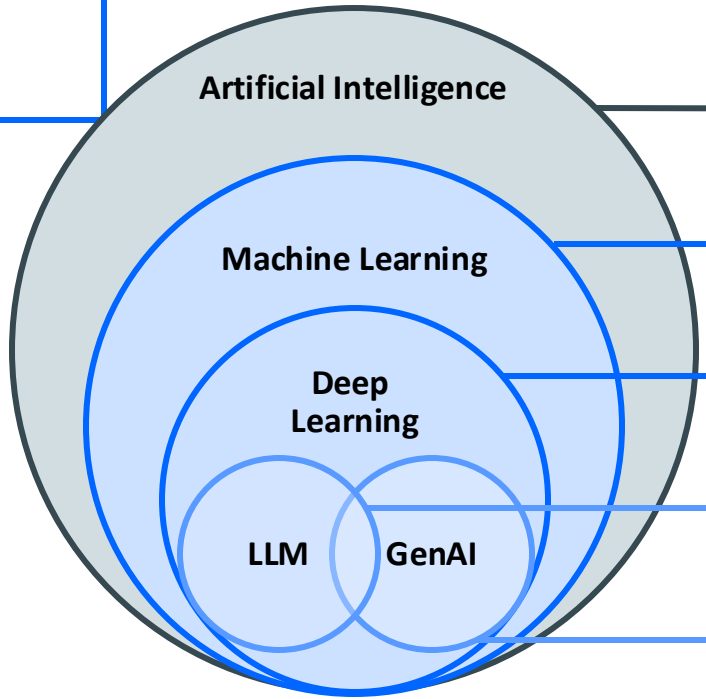
AI goes mainstream with ChatGPT

1M users in 5 days



But how does AI fit into the picture?

Navigating Artificial Intelligence: Understanding its fundamental building blocks



Artificial Intelligence (AI)
Any technology that enables machines to solve tasks in a way like humans do

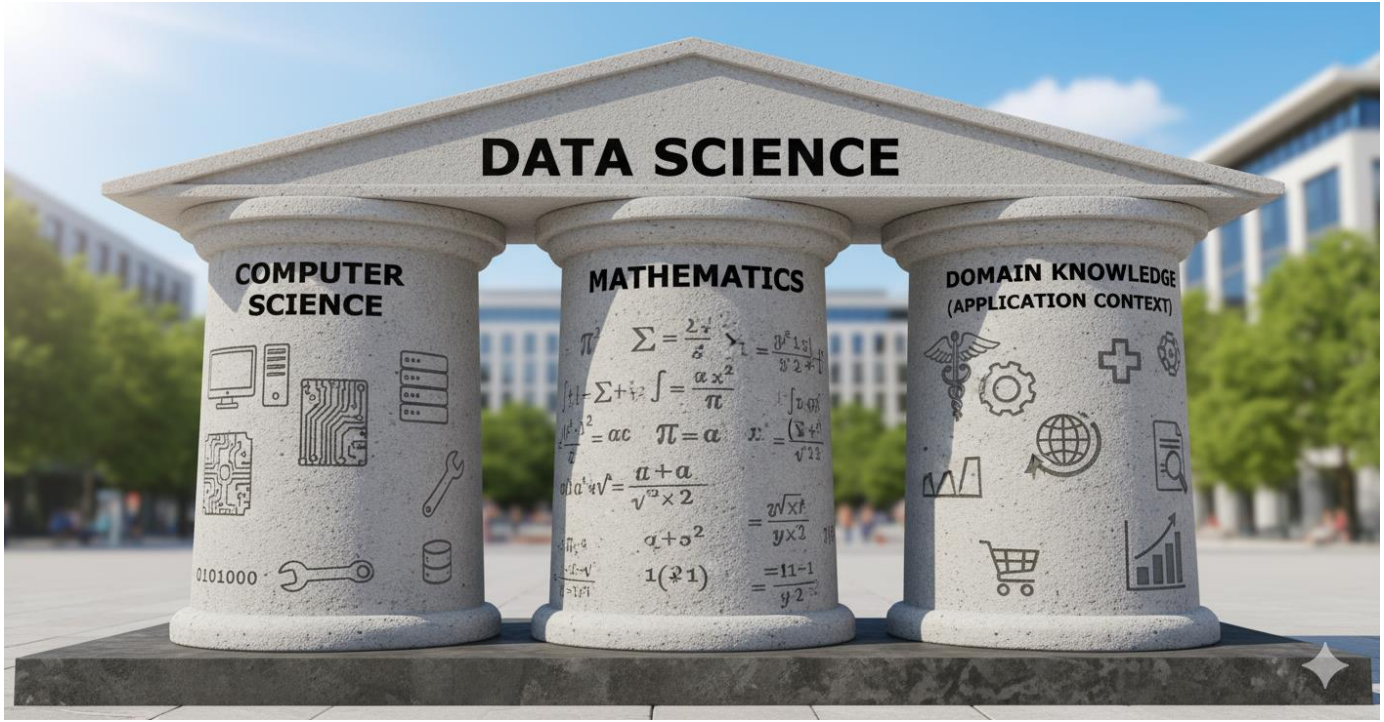
Machine Learning (ML)
Algorithms that allow computers to learn from examples without being explicitly programmed (supervised & unsupervised)

Deep Learning (DL)
Using deep artificial neural networks as models, inspired by the structure and function of the human brain

Large Language Models (LLM)
Models trained on massive datasets to understand and generate human-like text across diverse subjects

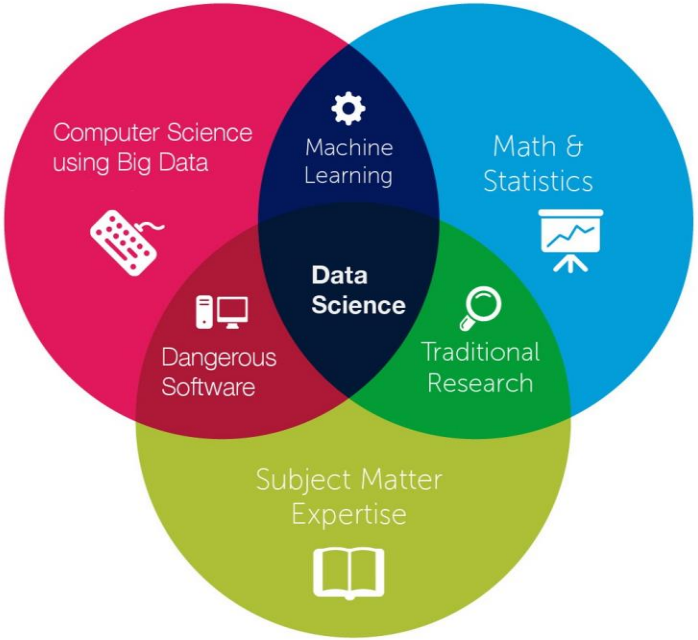
Generative Artificial Intelligence (GenAI)
Refers to technologies that utilize machine learning models to generate human-like text, images, or other content

The three pillars of data science



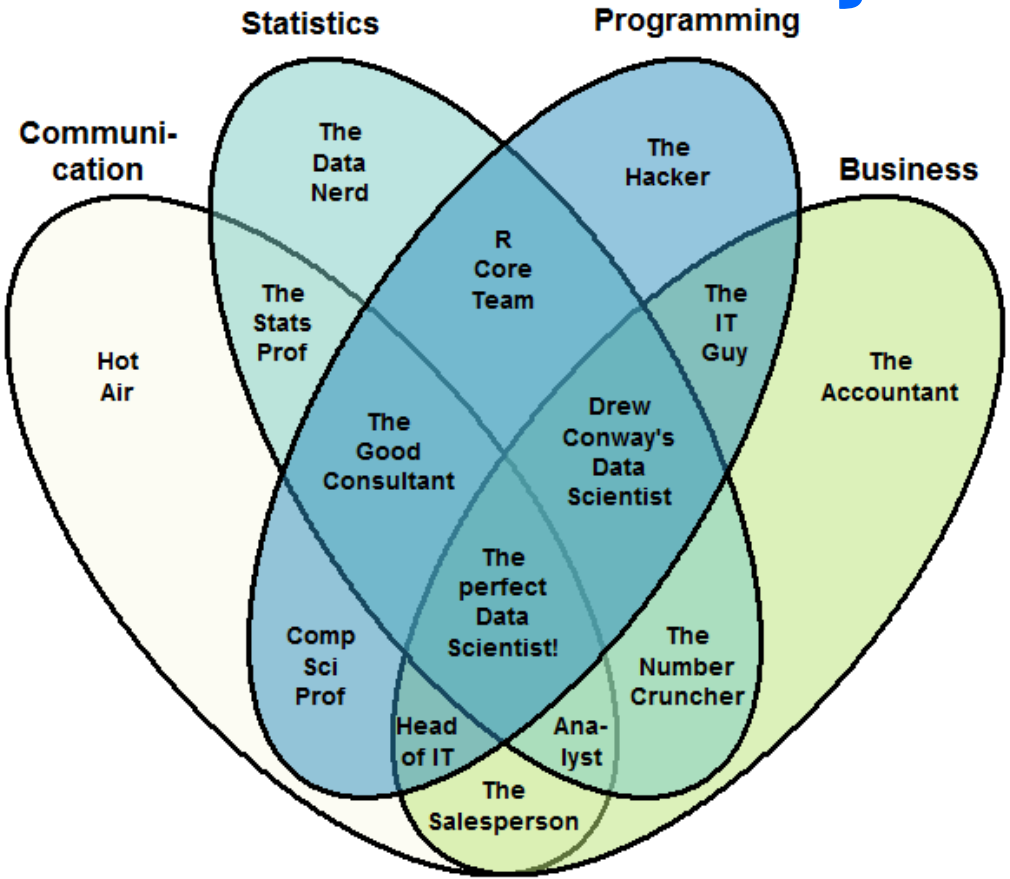
Why so many disciplines?

- Enormous amounts of data → Big Data
    - New algorithms. Must be scalable and efficient
    - New data management technologies and concepts
- High dimensionality of data
    - Data can have thousands of dimensions/attributes
    - High complexity and variability of data
- New, complex fields of application
- Data science is not (only) out-of-the-box, but domain-specific



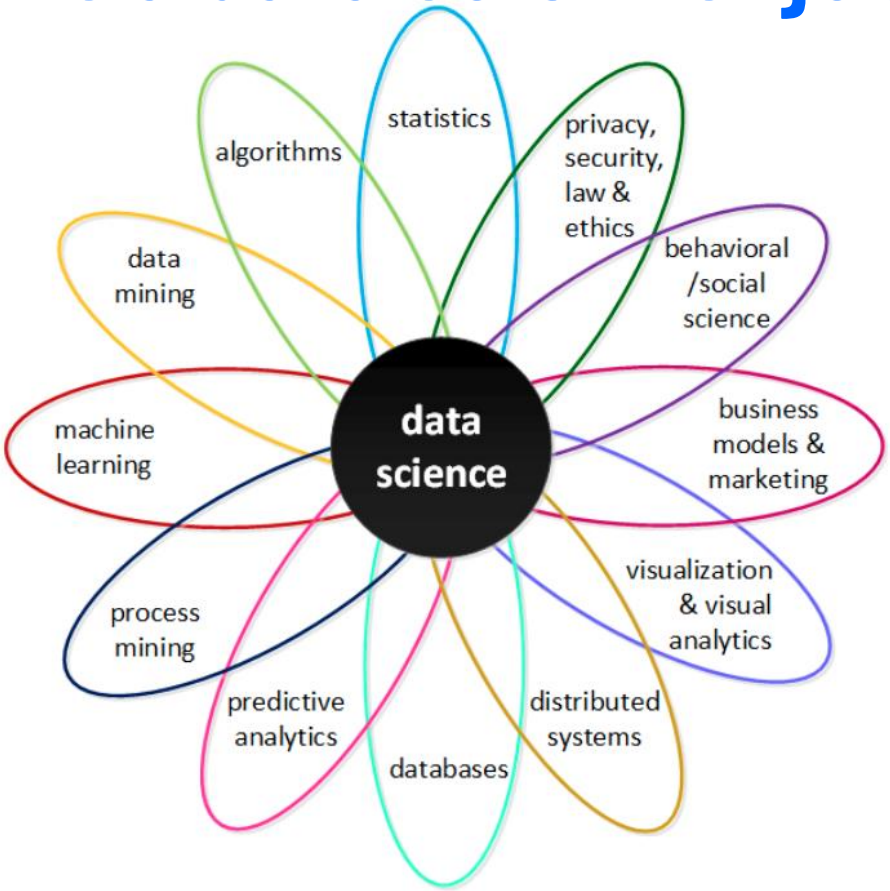
https://miro.medium.com/max/1100/1*aXJWLmf-CYqVTrNiE2pdCw.png

So, what is a data scientist job?



Stephan Kolassa: http://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html/2

So, what is a data scientist job?



https://sda.tech/wp-content/uploads/file-manager/Bonn-SDA-WvdA-2017.pdf

And more concrete ... ?

Data Steward

- Monitoring data quality and integrity
- Responsible for the technical accuracy of data

Tasks:

- Data owner
- Data governance
- Master of data

Data Engineer

- Data supply
- Develop, implement, test, and maintain architecture for data storage

Tasks:

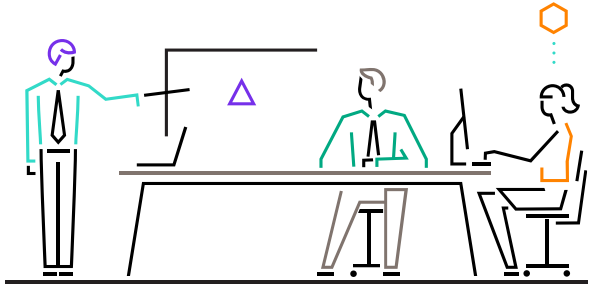
- Data Quality Improvement
- Data Transformation

Data Scientist

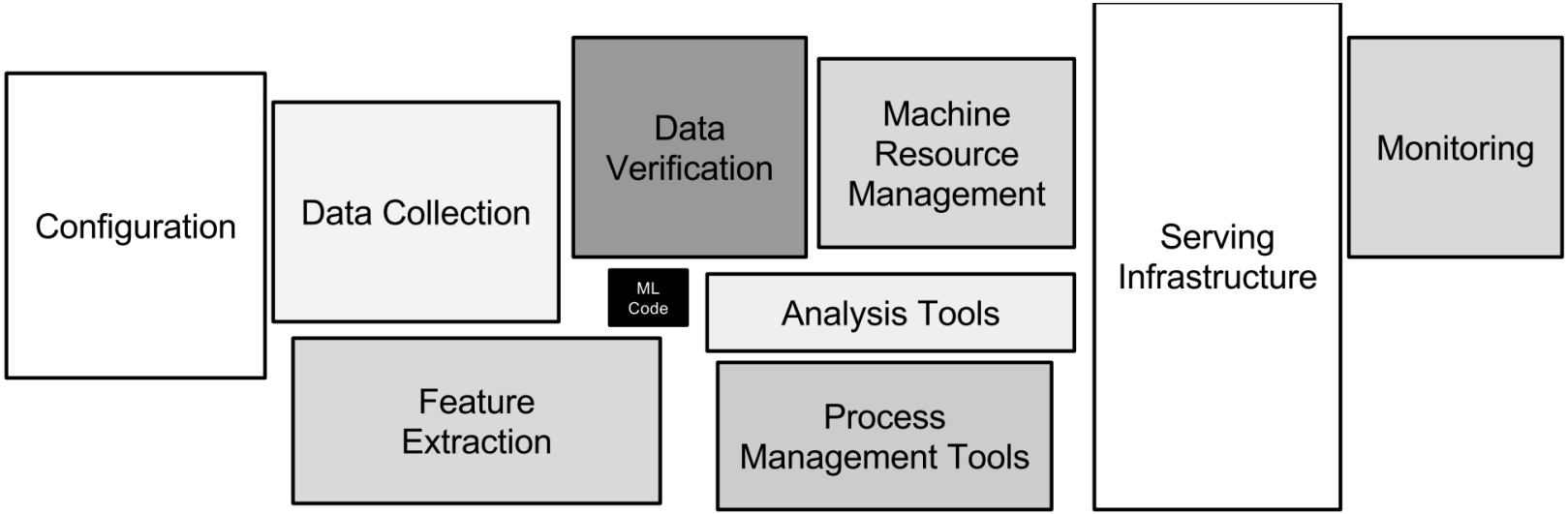
- Provides answers to analytical questions using data

Tasks:

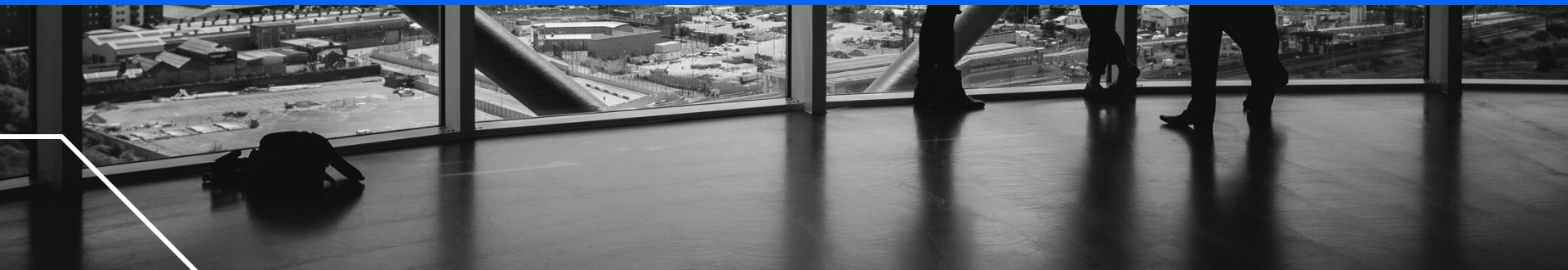
- Data Exploration & Visual Analytics
- Model Deployment & Scoring
- Big Data Handling & Manipulation



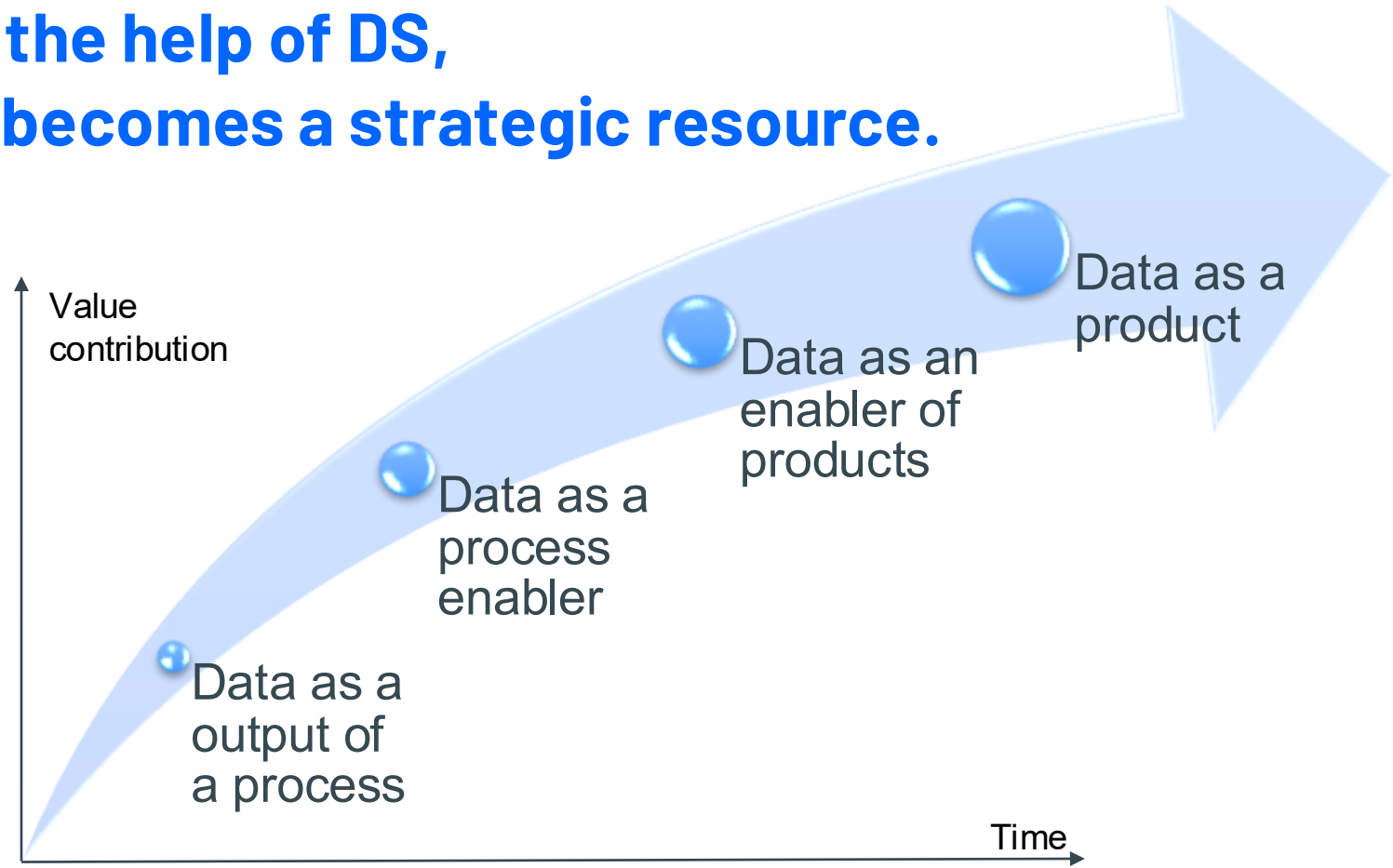
https://www.bertelsmann-university.com/fileadmin/Programs/AI_Curiculum/1_FAQ__%C3%9Cbersicht_Data_Curriculum_DE_final.pdf


Configuration
Data Collection
Data Verification
Machine Resource Management
Monitoring
Serving Infrastructure
ML Code
Analysis Tools
Feature Extraction
Process Management Tools

Big Data



With the help of DS, data becomes a strategic resource.



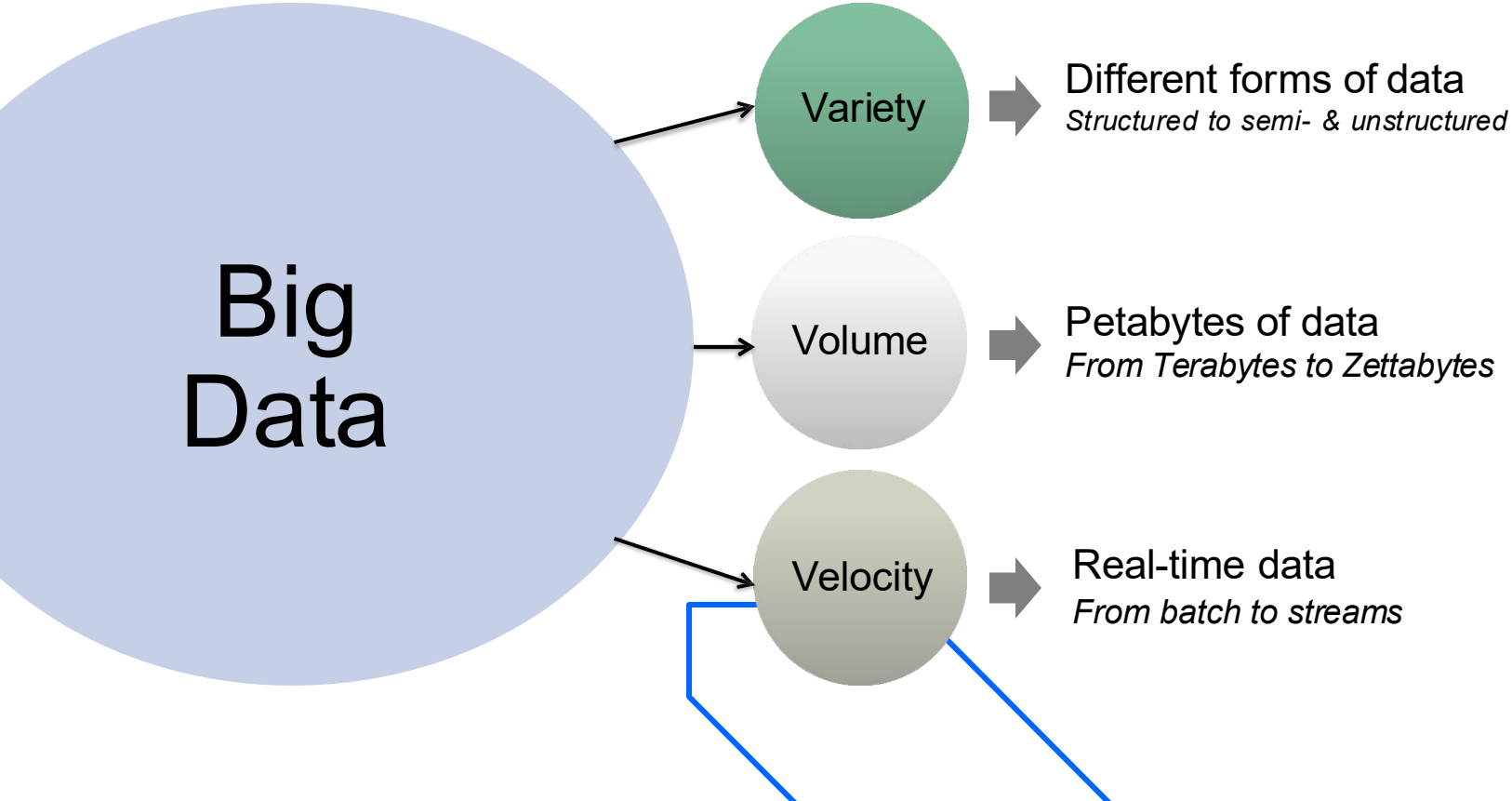
Otto, B.: Industrial Data Space im Überblick, Fraunhofer Institut für Materialfluss und Logistik, Dortmund, 30.10.2015, URL: http://de.slideshare.net/borisotto/berblick- zum-industrial-data-space

Big Data

Refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.

Big does not stand primarily for size, but as an analogy for “overwhelming”

Big can mean “high variety”, “high volume” or “high velocity”


Big Data
Variety
Different forms of data
Structured to semi- & unstructured
Volume
Petabytes of data
From Terabytes to Zettabytes
Velocity
Real-time data
From batch to streams

More Vs

Turning data into knowledge

Variation in meaning in different contexts

Uncertainty of the data

- Not easy to measure
- Depend on context and intended use

Where is the data coming from?

Possible data sources

- Questionnaires, interviews, ...
- Observations & measurements (e.g. in production)
- Analysis of content (documents, web scraping, social media...)

Raw data: Data received or collected

- no variables have been manipulated
- no data removed from the record
- no summary / aggregation

Not prepared
(pre-processed)

Where is the data coming from?

Private data

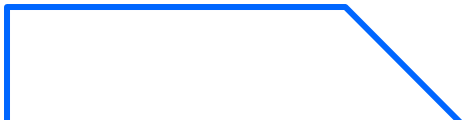
Created by customers
Created during business process execution

Commercial data

Cloud Marketplaces (e.g., AWS Data)
Qlik DataMarket, Statista

Open-source data

Data that is publicly available (check for limits on usage)



Open-Source Data Sources examples

Kaggle
World Health Organization
Our World in Data
Census Bureau (U.S.)
National Oceanic and Atmospheric Administration (U.S.)
UC Irvine Machine Learning Repository
Harvard Dataverse
AWS, Facebook, Google, Microsoft, ...

CAN YOU GET THE DATA OUT OF SILOS?

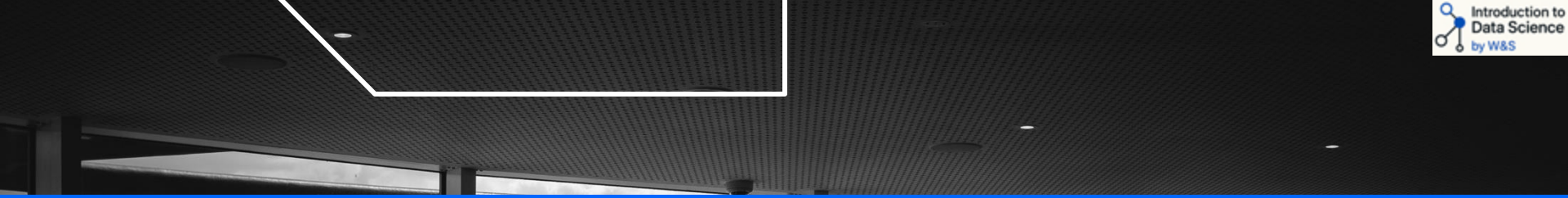
Value ≠ Quality

While value and quality of big data may be correlated, they are conceptually different. For example, one can have high quality data about the names of all the countries in North America, but this list of names may not have much perceived value. In contrast, even relatively incomplete data about the shopping habits of people can be quite valuable to online advertisers.

Xin Luna Dong, Google, 2015

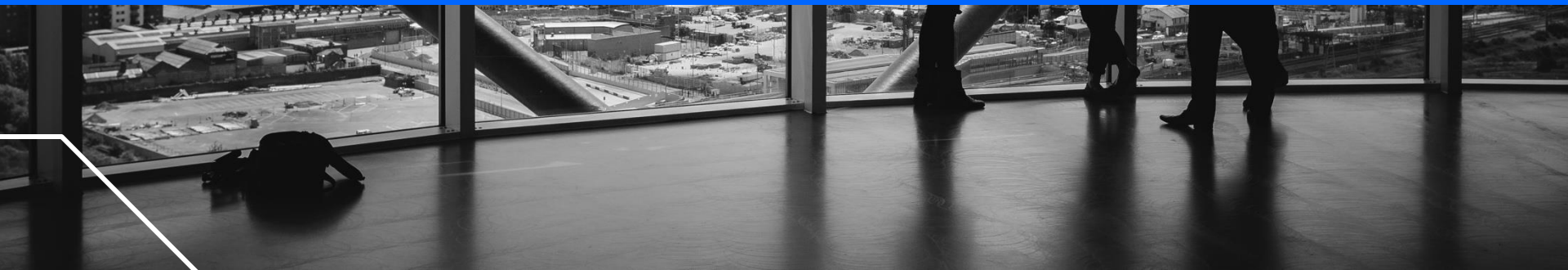
Raw data (including high quality data) itself does not hold any value, unless it is processed in analytical tasks from which humans or downstream applications can derive insight.

Gerhard Weikum, Max Planck Institute for Informatics, 2015



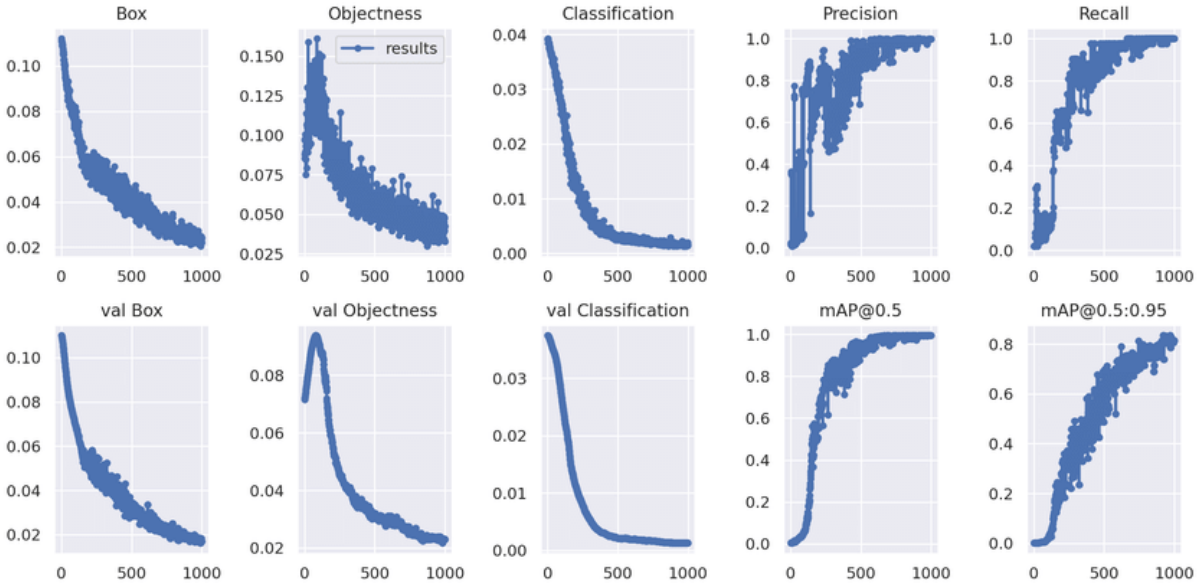
Introduction to Data Science by W&S

Data Storytelling and Visualization



Data story telling

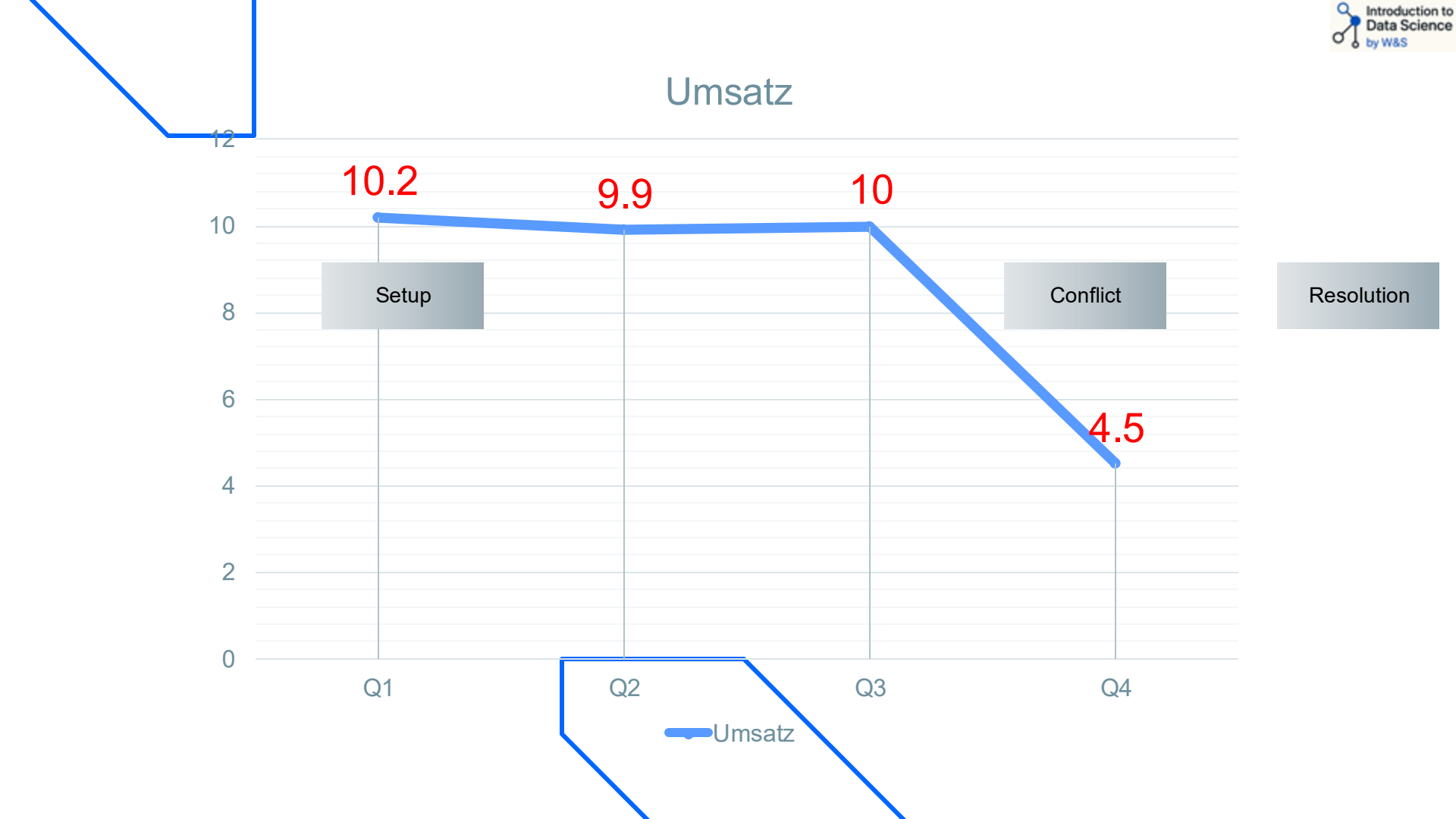
https://twitter.com/wcastillophd/status/1463600788140396550


Box
Objectness
Classification
Precision
Recall
results
val Box
val Objectness
val Classification
mAP@0.5
mAP@0.5:0.95


Introduction to
Data Science
by W&S
30%
Saved marketing budget

https://upload.wikimedia.org/wikipedia/commons/c/c4/Kiss_Logo.svg

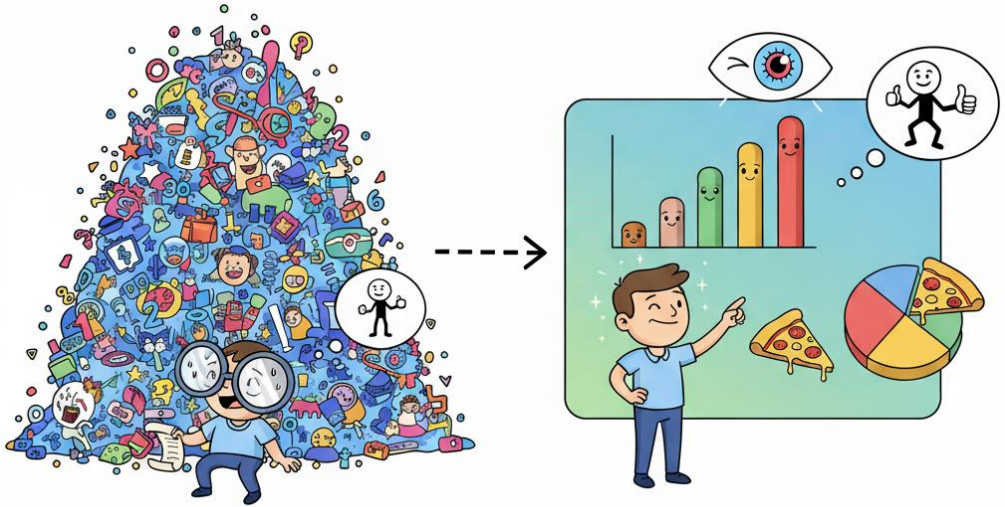
What do you remember


Umsatz
12
10
8
6
4
2
0
10.2
9.9
10
4.5
Setup
Conflict
Resolution
Q1
Q2
Q3
Q4
Umsatz

Why do we visualize?

- Most datasets are far too large to be examined in their raw format.
- Visual analysis uses our **pre-attentive perception** - visual cues that humans process automatically and unconsciously.
- We can perceive and interpret these types of characteristics quickly and without any special effort.

- Example: Using the length of bars to represent sales volumes is an effective choice to indicate differences in sales between categories.



Possibilities of Visual Presentation

Length

- Very good for quantitative variables
- Poorly suited for qualitative variables

Possibilities of Visual Presentation

Width

- Limited to quantitative variables
- Poorly suited for qualitative variablest

Possibilities of Visual Presentation

Orientation

- Limited to quantitative variables
- Poorly suited for qualitative variables

Possibilities of Visual Presentation

Size

- Limited to quantitative variables
- Poorly suited for qualitative variables

Possibilities of Visual Presentation

Form

- Poor for quantitative variables
- Very suitable for qualitative variables

Possibilities of Visual Presentation

Position

- Very good for quantitative variables
- Poorly suited for qualitative variables

Possibilities of Visual Presentation

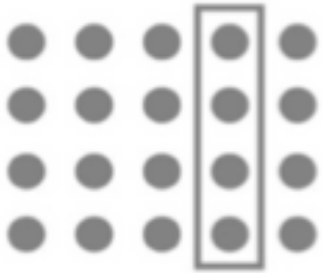
Grouping

- Limited to quantitative variables
- Very suitable for qualitative variables

Possibilities of Visual Presentation

Containment

- Poor for quantitative variables
- Limited suitable for qualitative variables



Possibilities of Visual Presentation

Hue

- Poor for quantitative variables
- Very suitable for qualitative variables

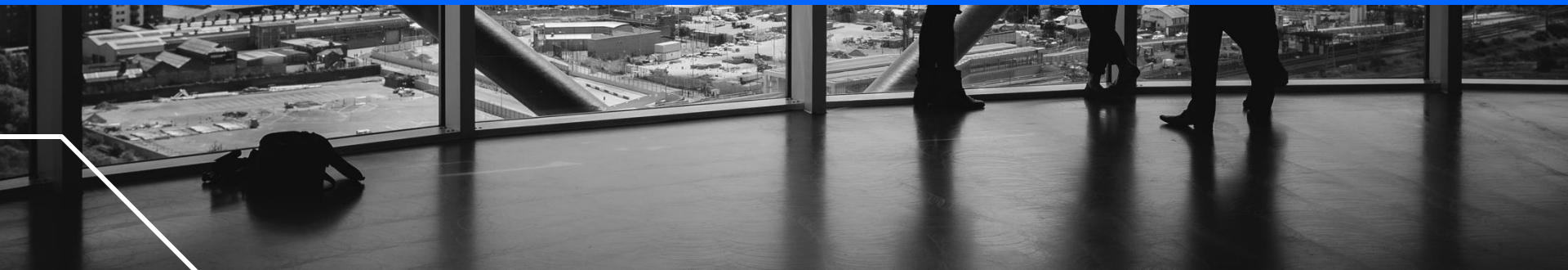
Possibilities of Visual Presentation

Colour intensity

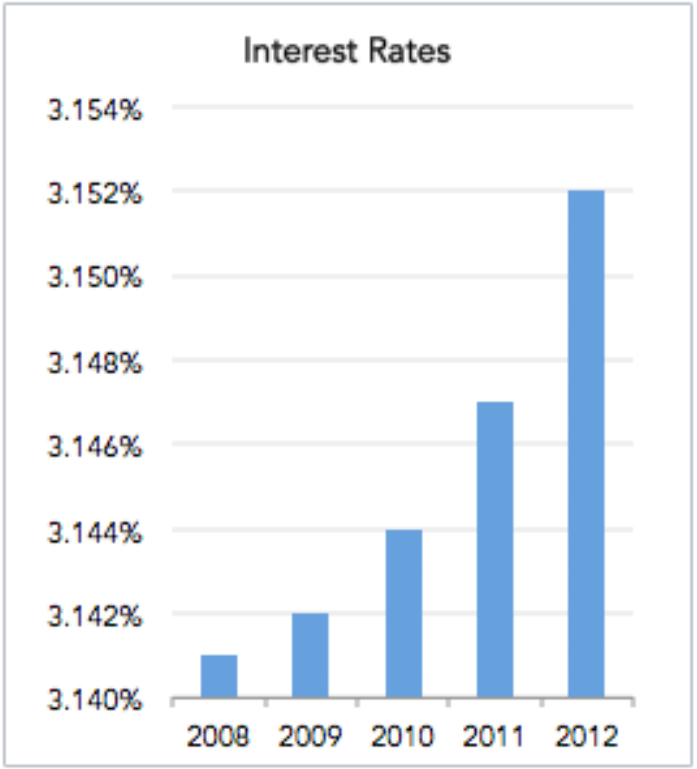
- Limited to quantitative variables
- Poorly suited for qualitative variables

Introduction to Data Science by W&S

Poor or misleading visualizations

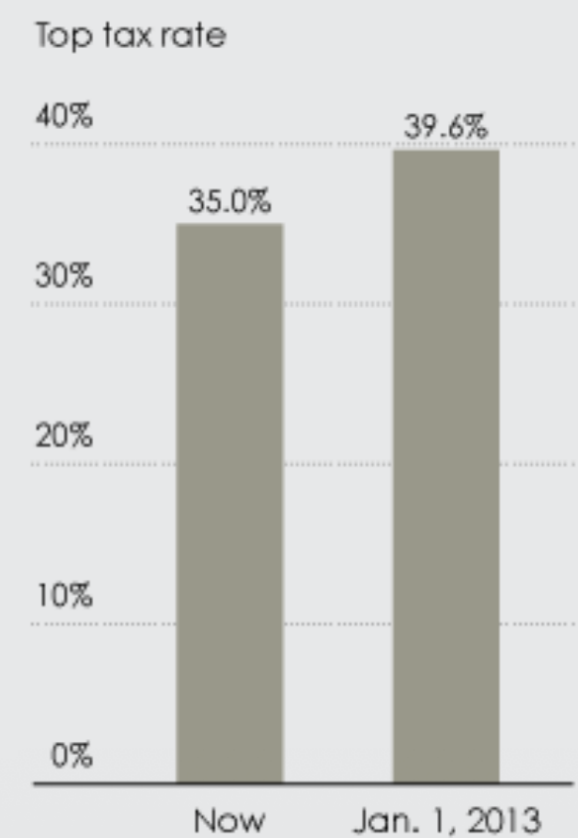


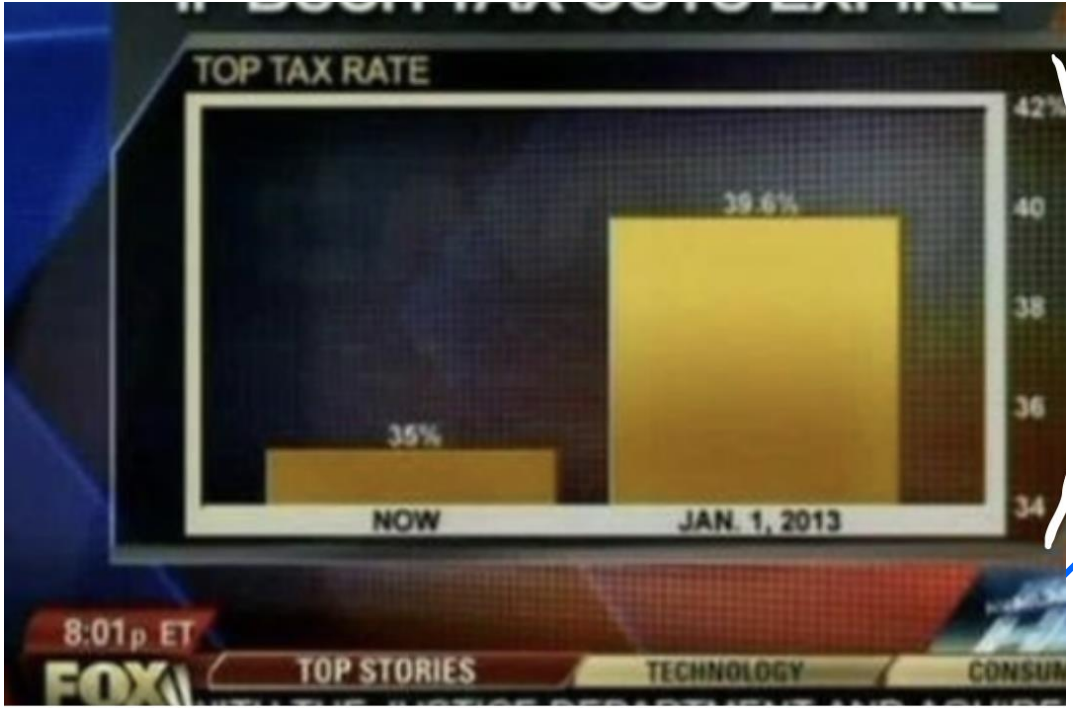
Poor or misleading visualizations



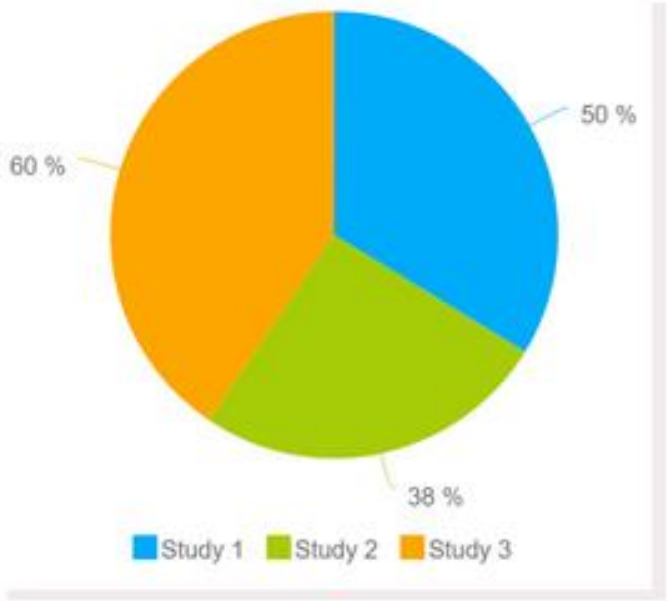
https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

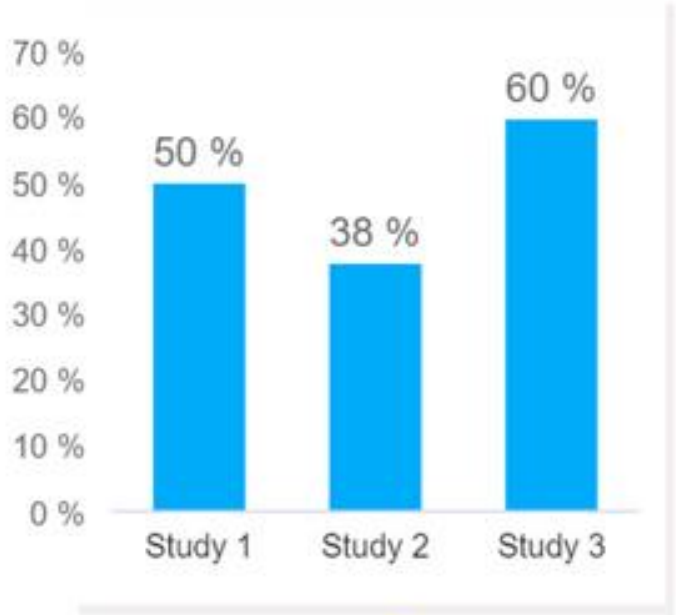
If Bush's tax rate reduction expires ...





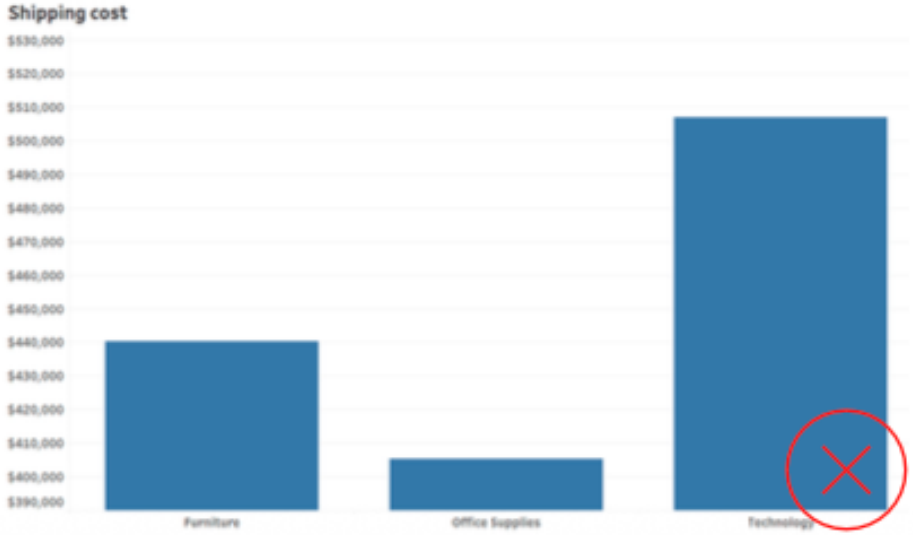
Poor or misleading visualizations

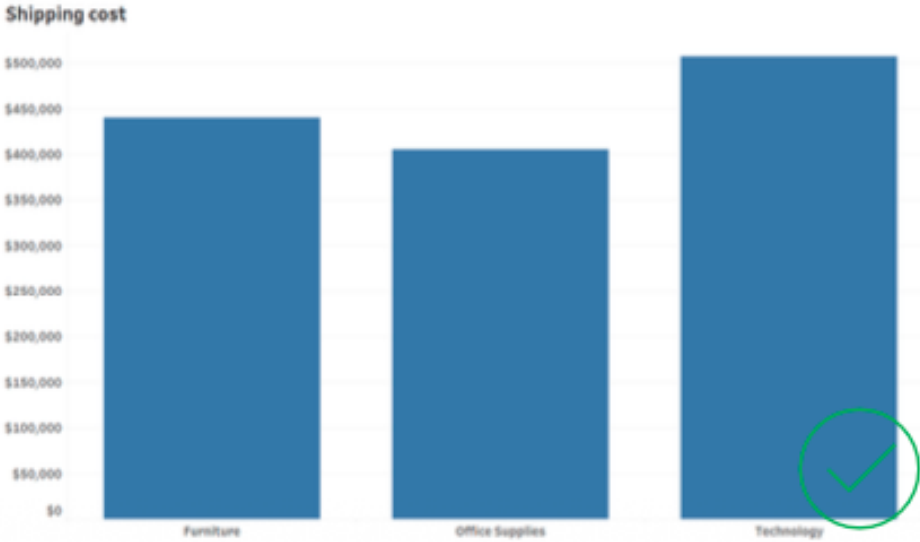




https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

Poor or misleading visualizations

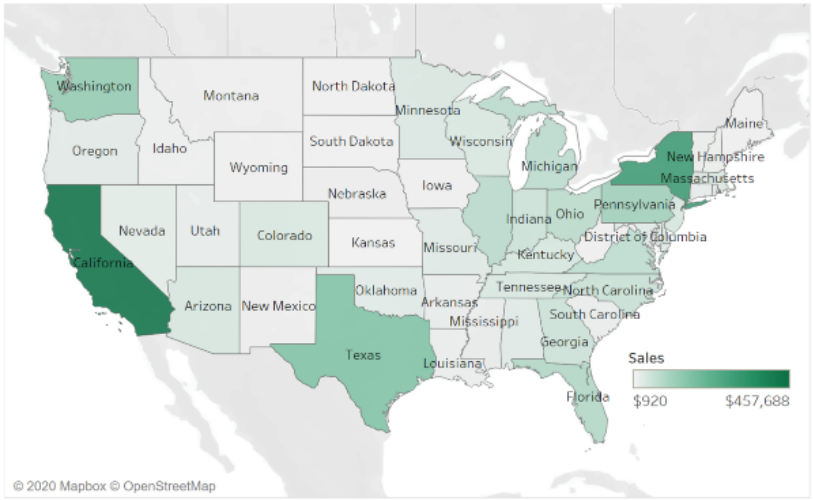




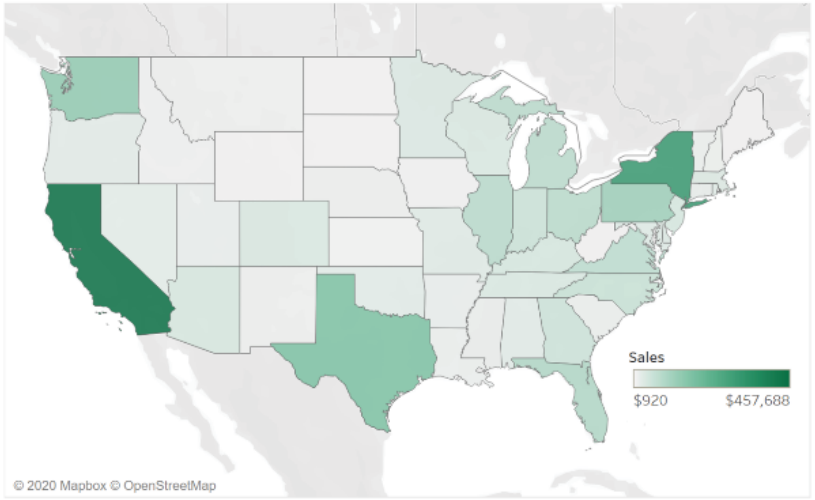
https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

Poor or misleading visualizations

total sales map



total sales map



ineffective

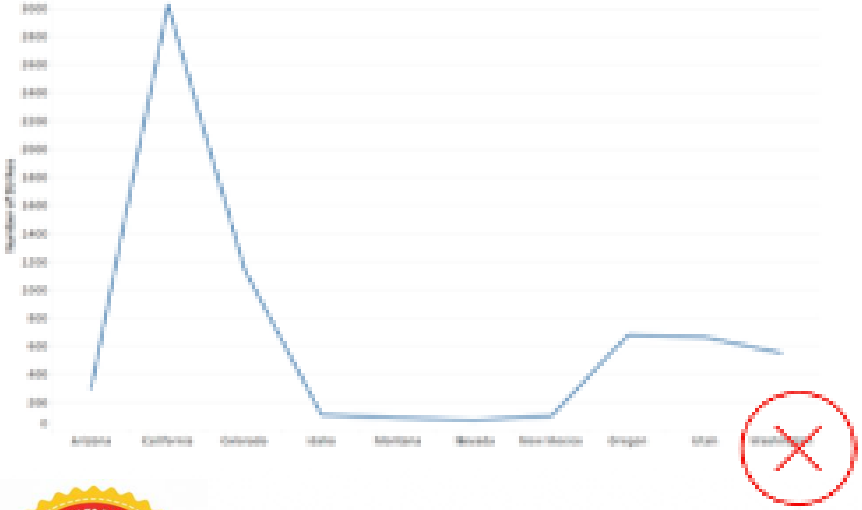
effective

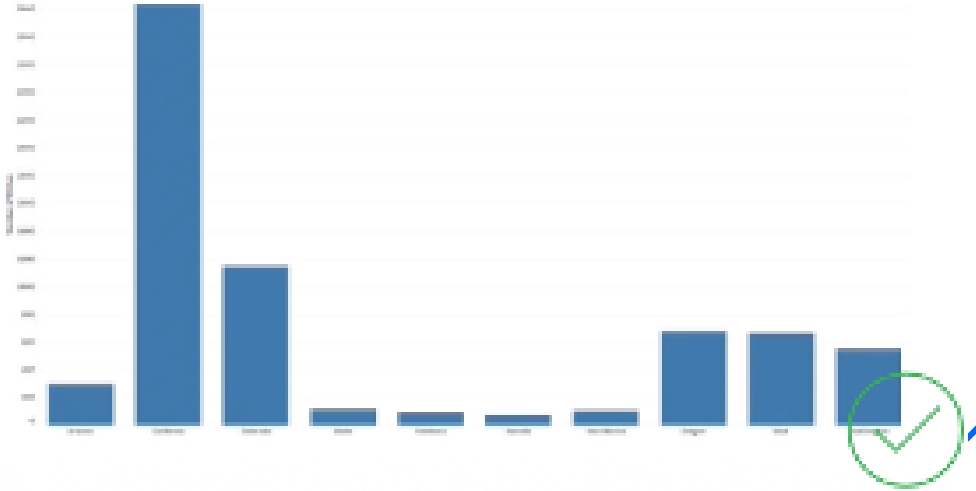
https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

Poor or misleading visualizations

https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

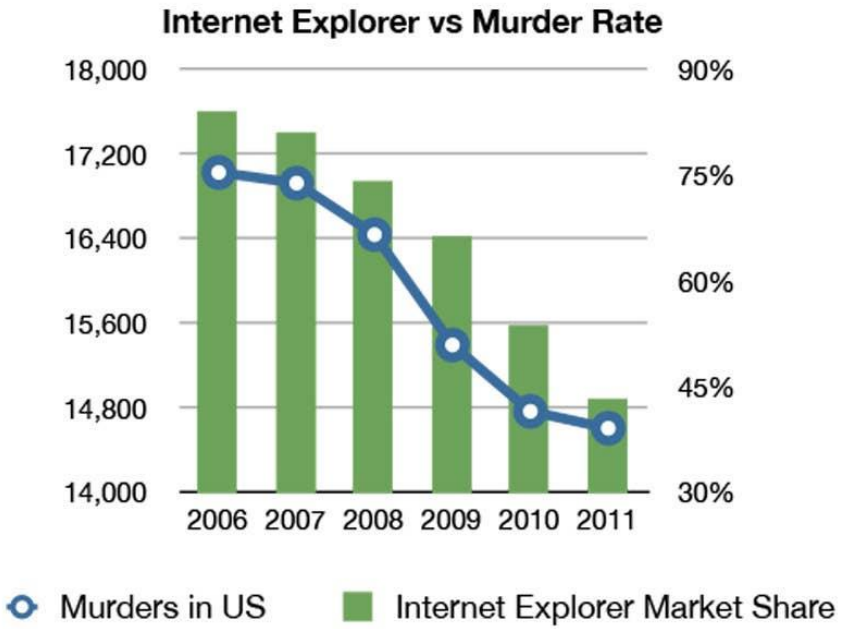
Poor or misleading visualizations

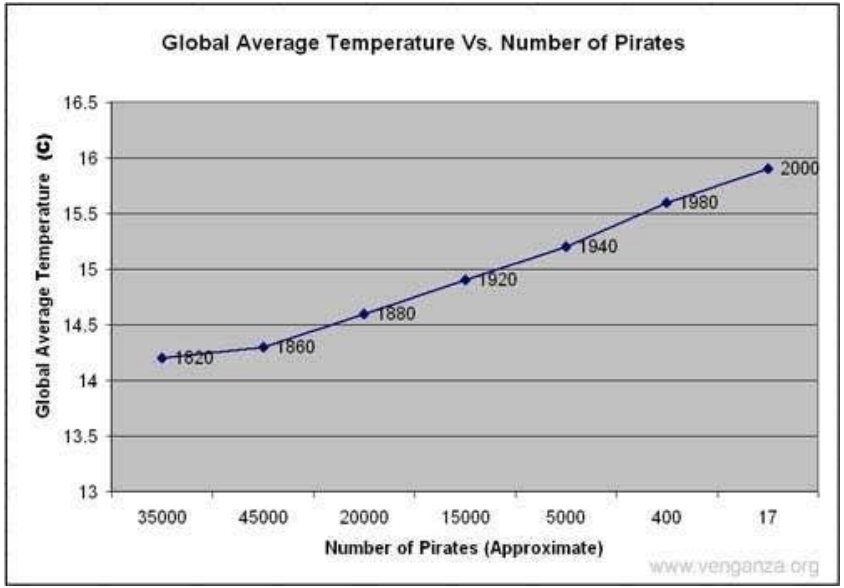




https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

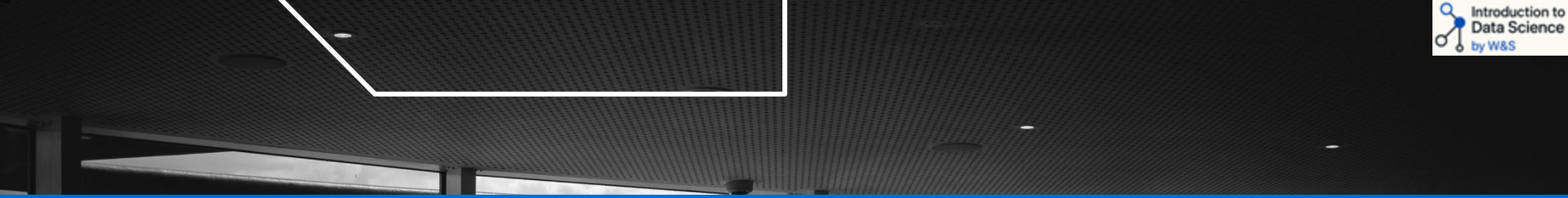
Correlation = causality?





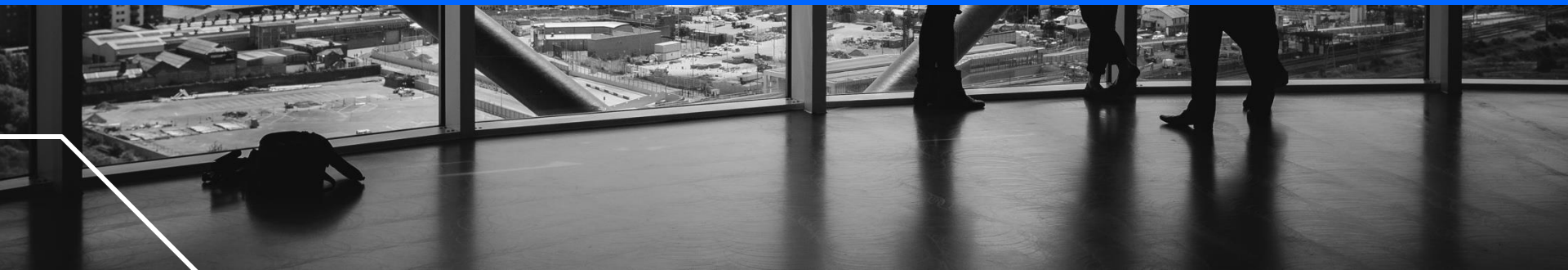
https://www.buzzfeednews.com/article/kjh2110/the-10-most-bizarre-correlations

https://www.tylervigen.com/spurious-correlations



Introduction to Data Science
by W&S

CHART TYPES



Line Chart

- Displays the change over time for a measure
- What kind of questions does this chart answer?
  - How has this variable changed over the past period?
  - When did this variable change?
  - How quickly did this variable change?
  - What are the trends? Can future trends be derived?
- Example: Stock market prices of the last five years

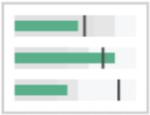
Bar Chart (horizontal | vertical)

- Comparison of data of different categories (dimensions)
- What kind of questions does this chart answer?
  - Which of these categories shows the highest/lowest value?
  - Are there any extraordinary categories?
  - What is the gap (deviation) between the lowest and highest values of different categories?

- Example: Sales per department

Bullet Diagram (horizontal | vertical)

- Modification of a bar chart that shows the performance of a primary measure of achievement of key figures.
  - What kind of questions does this chart answer?
  - As with the bar chart
  - Additional: Comparison per bar against a key figure



- Example: What is the actual turnover compared to the expected turnover?

Histogram

- Representation of the distribution of values
- What kind of questions does this chart answer?
  - Are events grouped around a certain probability?
  - Which group shows the highest values?
  - Which area covers the most observations?
- Example: Students' performance in an exam

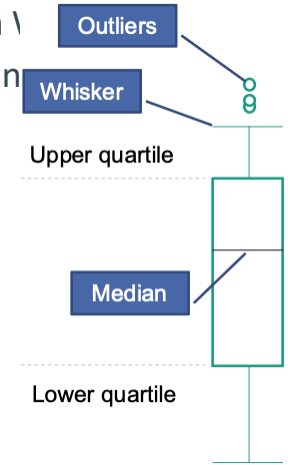
Boxplot

- Representation of the distribution within categories (dimensions)
- What kind of questions does this chart answer?
  - In which range are the values of most of the data in a category located?
  - Are there outliers in the data?
  - What is the median of values in a category?

- Example: Distribution of discounts in different product groups

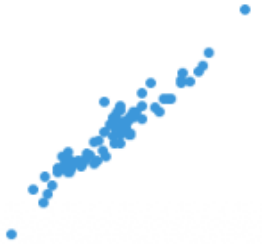
Boxplot

- Inside the box is the middle 50 percent of the data
- Whiskers (antennas) describe boundaries outside of which we speak of outliers
    - No uniform definition
    - Definition according to John \ of the whiskers to a maximum of 1.5 times the interquartile distan



Scatterplot

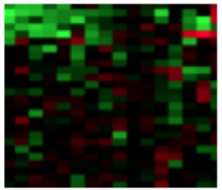
- Displays relationships between two numeric variables
- What kind of questions does this chart answer?
  - Are there any patterns when looking at the data points?
  - Are there correlation relationships between the variables?
  - Are there exits in the data?



- Example: Relationship between daily calorie intake and a person's body weight.

Heat Map

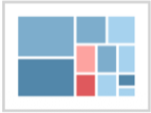
- Comparisons between two variables
- What kind of questions does this chart answer?
  - Is there a relationship between two variables?
  - Are certain areas particularly prominent?
  - Do two variables correlate?



- Example: Which nations won medals at the Olympics (divided into gold, silver,
- bronze)?

Tile Chart (Tree Map)

- Displays the proportion of an overall distribution.
- What kind of questions does this chart answer?
  - How much does this value contribute to the total?
  - How does the distribution of a variable change over time?



- Example: What proportion of total sales do an item's sales by sales region
- provide?
- Alternative chart types: pie charts, area charts, stacked bar charts

Pie Chart

- Displays the proportion of an overall distribution.
- What kind of questions does this chart answer?
  - How much does this value contribute to the total?
  - How does the distribution of a variable change over time?

- Example: What proportion of the votes did a party receive in the last elections?
- Alternative chart types: tree map, area charts, stacked bar charts

Map Charts

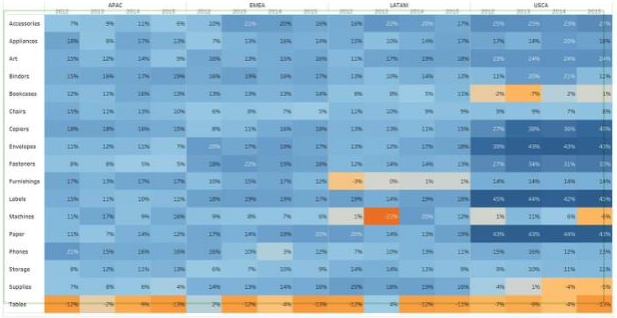
- Map charts represent positions and geographic patterns in the data.
- Variants: filled cards, point distribution cards, symbol cards, density cards...

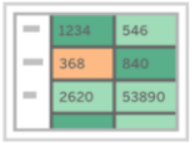
- What types of questions can this diagram answer?
  - Which place has the largest/smallest values?
  - Which region has the largest/smallest values?
  - How do you represent deviations geographically?

- Example: How is COVID-19 spreading worldwide?

Highlight Table

- Data table with color coding
- What kind of questions does this chart answer?
  - Shows all the details of the data points
  - Colors can be assigned to specific dimensions and categories or highlight key figures (quantiles, max/min values, ...)
  - Apply color markers from diagrams and thus display their details.





Tipps and Guidelines

Order your data

When displaying the value of several entities, ordering them makes the graph much more insightful.

To cut or not to cut?

Cutting the Y-axis is one of the most controversial practice in data viz. See why.

The spaghetti chart

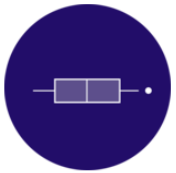
A line graph with too many lines becomes unreadable: it is called a spaghetti graph.

Pie chart

The human eye is bad at reading angles. See how to replace the most criticized chart ever.

Play with histogram bin size

Always try different bin sizes when you build a histogram, it can lead to different insights.



Do boxplots hide information?

Boxplots are a great way to summarize a distribution but hide the sample size and their distribution.

The problem with error bars

Barplots with error bars must be used with great care. See why and how to replace them.

Too many distributions.

If you need to compare the distributions of many variables, don't clutter your graphic.

https://www.data-to-viz.com/caveats.html

Presenting results

Promises

- Enhanced Analyses and Predictions: Leveraging large volumes of data can enable deeper insights and more accurate forecasts than ever before.
- Uncertainty at the Micro Level, Accuracy at the Macro Level: Big Data allows for precise identification of patterns and trends at the macro level, despite uncertainty at the individual level.
- Quality is Crucial: Despite the vast amounts of data, ensuring data quality is essential for drawing valid conclusions.

Risks

- Incorrect Modeling and Conclusions: The complexity and volume of Big Data can lead to errors in data modeling, resulting in false or misleading outcomes.
- Privacy and Individual Rights: Processing large datasets poses risks to privacy and can conflict with individual rights.

Recommendations for Mitigating Risks

- Rigorous Data Verification: Implement strict data quality verification processes to ensure the integrity of analyses.
- Transparent Modeling: Promote transparency in your modeling processes to build trust and avoid misinterpretations.
- Preserving Privacy: Adhere to privacy regulations and practices to protect individuals' rights and foster trust in Big Data initiatives.
- Reason about results: be able to explain; why have predictions been made?
- Reason about input: data set; metrics

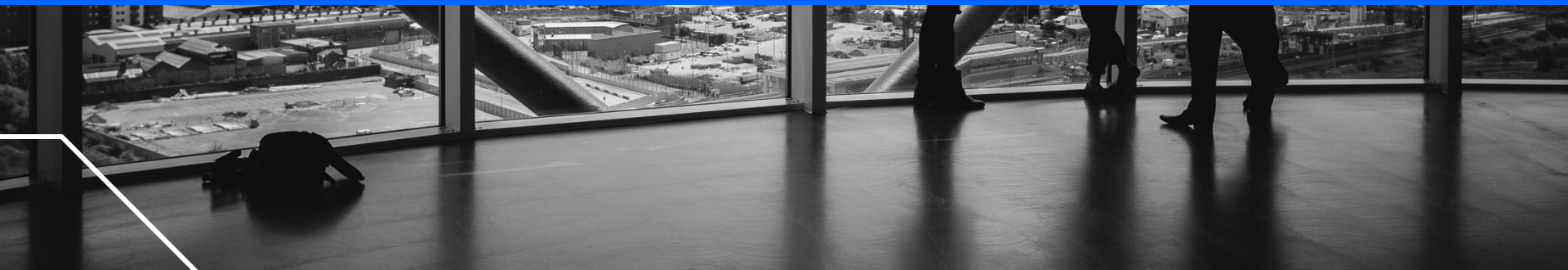
Use of Diagrams

- Usually there are data sets with many variables (characteristics) or
- we want to determine the utility value of other variables

- Application of diagrams
  - Often the same / different diagrams are used one after the other / combined
  - Decision on the type of visualization based on
    - variables,
    - dimensions in our data, and
    - the question asked


HANDS-ON
TIME
INTRODUCTION TO DATA SCIENCE
BY W&S
YOUR TURN!
EXERCISE / QUESTION

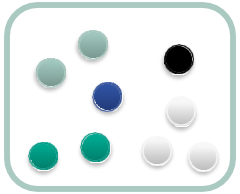
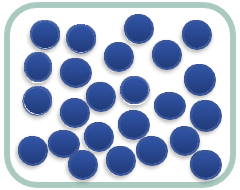
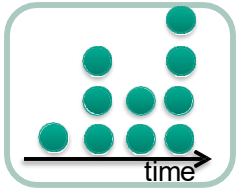
Challenges



Whats hard about data science

- Getting the data (usually)
- Overcoming assumptions
- Communication
    - With domain experts
    - Expectation management for client
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype to Production transition
- Data pipeline complexity (team boundary)

Data Challenges (1)

	Variety	Volume	Velocity
			
Challenge	Handling multiplicity of types, sources and formats	Dealing with large volumes of data	Streams, sensors, near real-time data
Impacted Tasks	Data integration	Storage, processing & analytics	Processing & analytics
Solution	Semantic technologies are a good fit	Distributed storage & parallel processing	Real-time technologies

Data Challenges (2)

- **Data veracity:** coping with uncertainty, imprecision, missing values, misstatements or untruths.

- **Data quality:** determining the quality of datasets and relevance to particular issues. Depends on the use case:
  - How broad/complete is the data?
  - How fine is the sample resolution? How timely are the readings?
  - Does the data contain any “noise” (errors)? Is it representative?

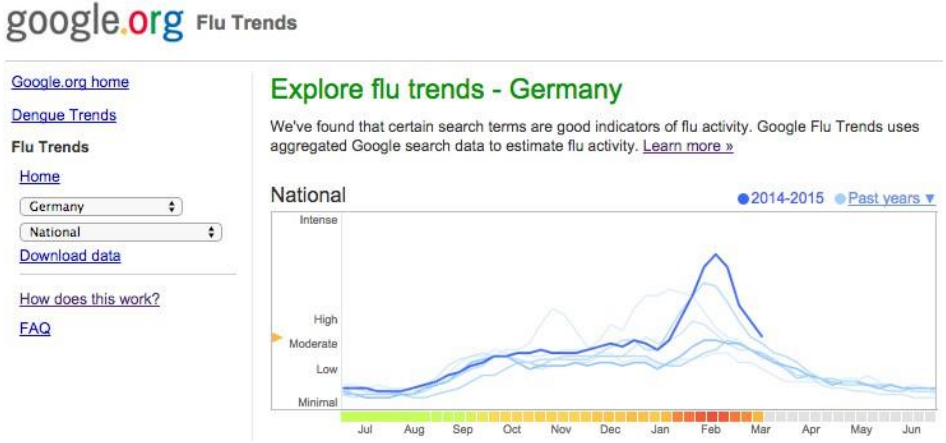
Data Challenges (3)

- **Data discovery:** finding relevant data from enormous amount of data available on the Web.

- **Data dogmatism:** analysis of Big Data can offer remarkable insights. However, data analysis should not entirely replace domain expert knowledge, but act as a tool to support /confirm facts.
    - E.g., Google Flu Trends

Data Challenges (3)

Example: Data dogmatism



Source: https://www.google.org/flutrends/de/#DE

"By counting how often we see these **[flu-related topics] search queries**, we can estimate **how much flu is circulating** in different countries and regions around the world."

Data Challenges (3)

Example: Data dogmatism

Comparison of Google Flu Trends model against medical reports.

SCIENCE BIG DATA

Google's Flu Project Shows the Failings of Big Data

Bryan Walsh @bryanrwalsh | March 13, 2014

A new study shows that using big data to predict the future isn't as easy as it looks—and that raises questions about how Internet companies gather and use information

Source: http://time.com/23782/google-flu-trends-big-data-problems/

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer[1,2,*], Ryan Kennedy[1,3,4], Gary King[3], Alessandro Vespignani[5,6,3]

+ Author Affiliations

*Corresponding author. E-mail: d.lazer@neu.edu.

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

Read the Full Text

Source: http://www.sciencemag.org/content/343/6176/1203

Process Challenges (1)

Big-Data Scientists *"Janitor Work"*

50% - of data scientists' time
80% - spent in "Data wrangling"

Data acquisition:

Data availability

Data permissions

Aligning/integrating data from different sources:

Syntactic challenge: Data in different formats

Semantic challenges: Resolving when two objects are the same, describing relationships between data points, resolving inconsistencies

Transforming, cleaning and organizing the data into a form suitable for analysis

Source: "For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights" in The New York Times. 2014.

Process Challenges (2)

Modeling data:

Mathematically: mathematical models to describe the data Statistical methods or Machine Learning

Simulations

Knowledge representation: ontologies and rules

Understanding the output:

Interpreting the results

Visualizing

Sharing the results

Data privacy, security, and Governance:

- Ensuring that data is used correctly (abiding by its intended uses and relevant laws).
- Tracking how the data is used, transformed, derived, …
- Managing data lifecycle.

*"Many data warehouses contain sensitive data such as personal data. There are **legal** and **ethical concerns** with accessing such data. So the data must be secured and access controlled as well as logged for audits".*

Michael Blaha, Modelsoft Consulting Corporation, 2012

Source: http://www.odbms.org/blog/2012/03/data-modeling-for-analytical-data-warehouses-interview-with-michael-blaha/

Data science – two sides of the same coin

Opportunities	Risks
• Discover potential • Develop new services • Informed decisions based on forecasts • Service and product improvement	• Surveillance • Manipulation • Uselessness • Data protection

Bis Donnerstag

samuel.schlenker@hpe.com

Tag 4: Datenaufbereitung & Feature Engineering

Samuel Schlenker

06.11.2025, WWI 2025F

Students should ...

- Differentiate between objects, data, databases, and information using the knowledge hierarchy
- Understand key properties of quality datasets (volume, history, consistency, purity, level of detail, clarity, transparent origin)
- Distinguish between structured and unstructured data formats
- Apply dimensional data structures (dimensions vs. facts) for analysis
- Perform data operations: slicing, dicing, roll-up, and drill-down
- Understand the importance of data preparation and cleansing (representing 45-80% of data scientists' time)
- Identify and handle missing data using different strategies (MCAR, MAR, MNAR)
- Apply imputation techniques for missing values
- Detect and handle outliers using methods like IQR, Z-score, and Isolation Forest
- Apply normalization techniques (Min-Max, Z-Score standardization)
- Perform categorical encoding using one-hot encoding
- Conduct feature selection and extraction to improve model performance
- Understand the data preparation pipeline from raw data to analysis-ready datasets

Differentiation of data

- **Differentiation between Object**, **Data**, **Database** and **Information**
- Differentation by Prof. Klaus North
- Differentation by various properties

Differentiation of data

- **Object** (also data set): Object of consideration, has various interesting features (which are generic to its class).
- **Data**: Value of a single (characteristic) property of an object
- **Database**: is the group of objects in which we are interested.
- **Information**: Created by interpreting the data (e.g., assignment of meanings)

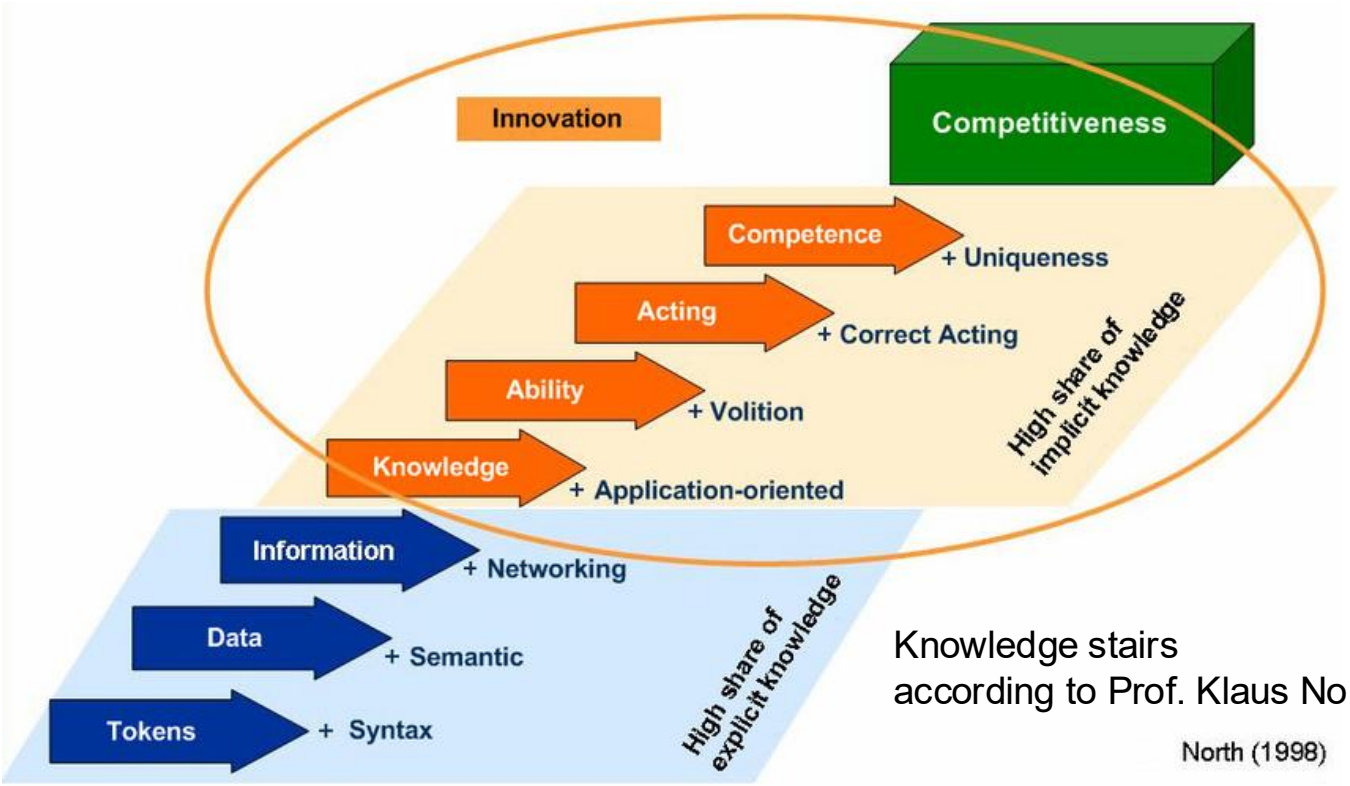
Object

Data

Database

Information

Differentiation of data



Knowledge stairs
according to Prof. Klaus North

North (1998)

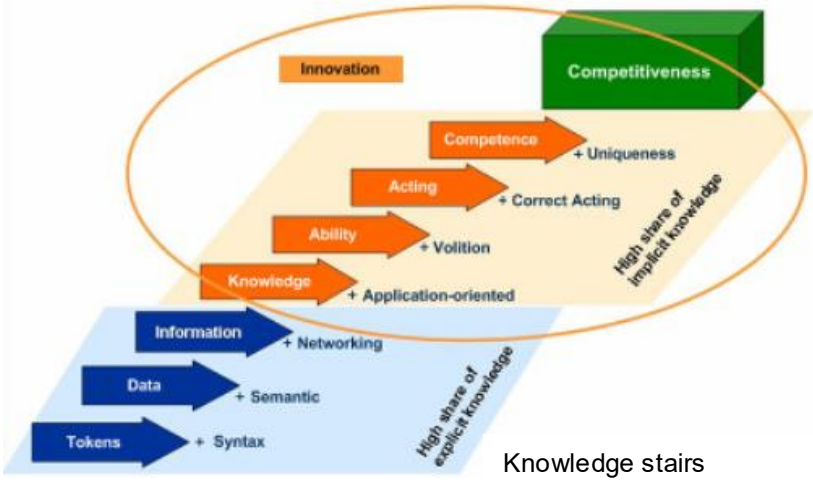
Differentiation of data

Data: Syntactically correct, symbolic representations consisting of individual data elements and values.
T = 16, P = 928, R = CEU
(Data type: Syntactic structure of data elements)

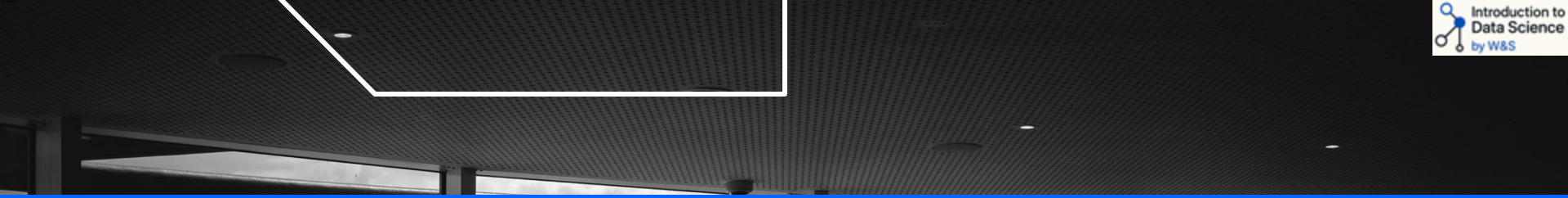
Information: Places data in a semantic context.
T = 16 °C, P = 928 mbar, R = CEU (Central Europe)

Knowledge: Systematic linking of information
T = 16 C°, P = 928 mbar, R = CEU (Central Europe), probable rain

Skills and actions: (Automated) application of knowledge
Mountain guide, farmer
Weather report, weather app => Models and algorithms

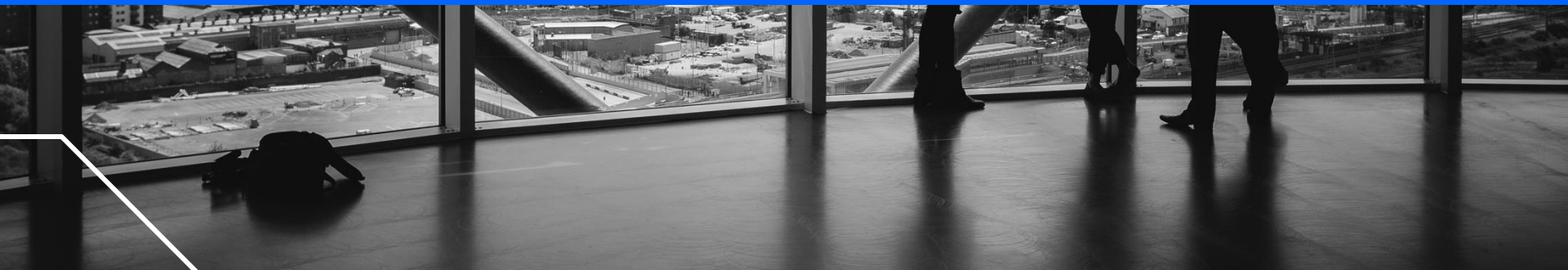


Knowledge stairs
according to Prof. Klaus North



Introduction to Data Science by W&S

Properties of Data and Datasets



Properties of Data and Datasets

Volume
Amount of data
A large amount of relevant, available data means there's a better chance you have what you need to answer your questions.

Note: There is no need to collect data simply for its own sake. Relevance is important!

Properties of Data and Datasets

History / Temporal Development

Data obtained at different points in time allows us to see how the current situation developed (based on patterns).

Example: the sales trends of the last 10 years make it possible to detect increases or decreases.

Properties of Data and Datasets

Consistency
If the facts change, the data should

- reflect this (the data must be adjusted)
- do not create inconsistencies
- continue to be compliant with the data model

Example: Inflation and price adjusted salary and price data.

Properties of Data and Datasets

Purity (cleanliness)
For data to be meaningful, it should not be inaccurate or incomplete and should not contain errors.

Note:

- Data that is inaccurate or misleading (high degree of impurity) is also said to be corrupt.
- Often related to consistency.

Properties of Data and Datasets

Level of **Detail**
The more detailed the data, the better we can examine it at different levels.
Operations:

- **Aggregation** (Roll-Up)
- **Disaggregation** (Drill-Down)

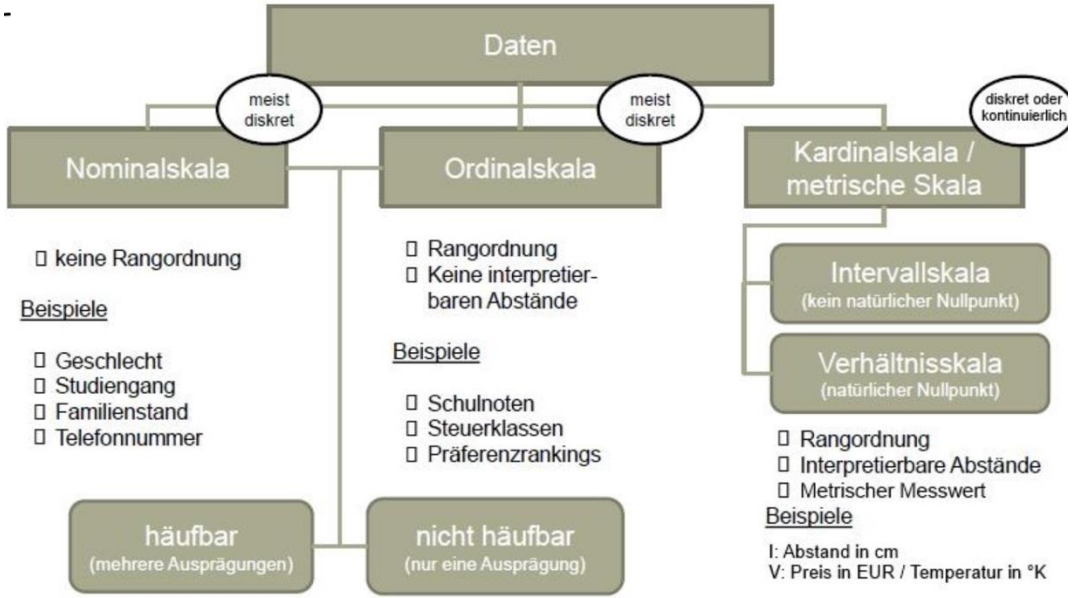
Example: We want to understand cycling trends in Baden-Württemberg. Then it would be helpful t

Properties of Data and Datasets

Variables (also: field, column)
Data usually includes **quantitative** (numerically measurable) as well as **qualitative** (characteristic, non-numerically measurable) variables.

Note: More variables usually let us discover more!

Properties of Data and Datasets



Properties of Data and Datasets

Segments

Grouping variables based on similar characteristics can be integrated into the data for easier analysis.

Example: Data about movies can be grouped by genre (e.g., action, science fiction, romance, comedy).

Properties of Data and Datasets

Clarity (comprehensibility)
Data should be described in terms that are easy to understand (not in code/encodings)

Example: Description of apartment types such as "single-family house", "two-family conversion" and "terraced house" are much easier to understand than "1Fam", "2fmCon" and "TwnhsE".

Properties of Data and Datasets

Transparent Origin

To trust the data, we need to know if it comes from a reliable source and if it has been treated in a trustworthy way.

Properties of Data and Datasets

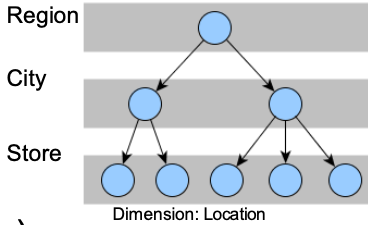
Dimensional Structure

Structuring of data into two types:

- dimensions (usually qualitative values) and
- facts or measures (quantitative values)

Note: data can be examined in terms of dimensions. In many cases, individual (operational) data sets/values are not of interest, but aggregated values/quantities provide statements about data in its entirety.

Dimensions



Are **descriptive** in nature

- are scaled ordinal or nominal
- can be hierarchically structured and
- index facts

A dimension forms a view of the facts (side of a data cube)
The hierarchy set H represents the set of hierarchical attributes of a dimension

- granularity and levels of compaction
- irreflexive half-order $(H, \prec)$

Examples: Product, Time or Country, ...

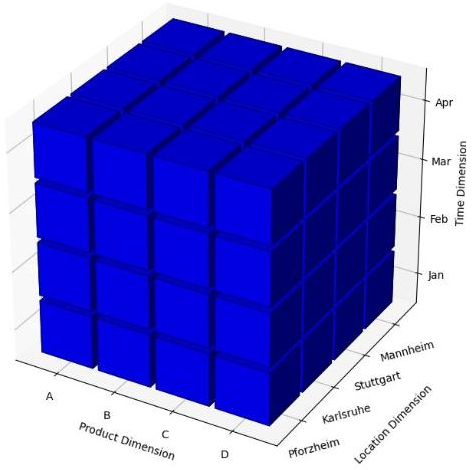
Facts

Represent key performance indicators

- also called numerical or summary attributes
- can be calculated by aggregate functions and formulas
- for example: min, max, count, sum, avg

Facts always have a **quantitative** (numeric) data type
Examples of facts:

- sum(turnover)
- count(employee-number)
- min(contribution-margin)

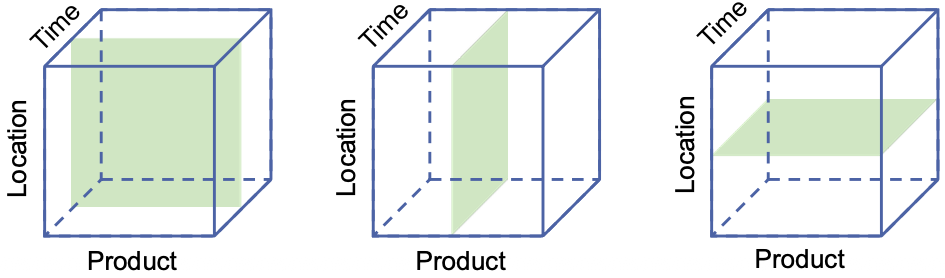


Slicing

Slicing or selection $\sigma T\,(cond)$

- selects tuples from the cube
- according to selection conditions ($cond$)

Conditioning (Statistics)

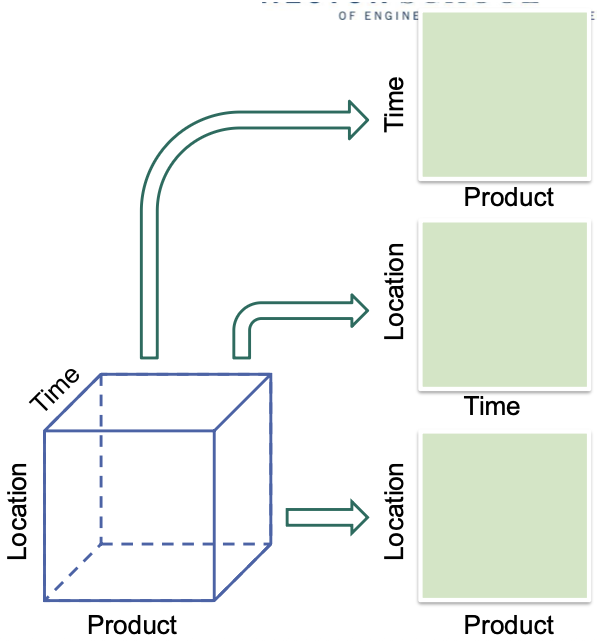


Dicing

Dicing or projection $\pi T\,(D'\ )$

- sets dimensions for projection
- creates a sub cube

Marginalizing (multivariate statistics)



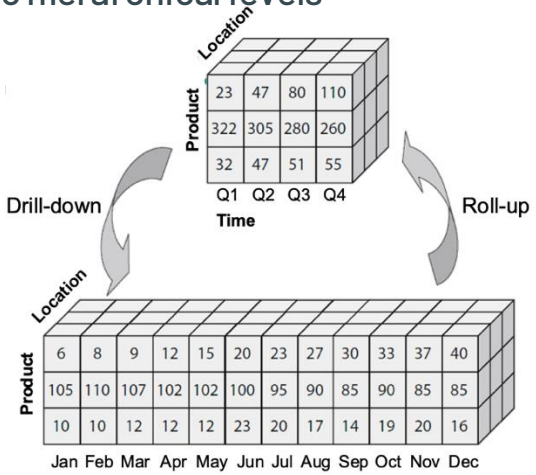
Roll-Up & Drill-Down

Roll-up

- corresponds to aggregation
- dimensional hierarchy from specific to general
- aggregation can be calculated from the more specific hierarchical levels

Drill-down

- disaggregation
- reversal of the roll-up (inverse)
- $\delta T = \rho T - 1$



Database (relational model)

- Data is structured in rows and columns
- Each column represents a different variable (characteristic, field).
  - a variable is the measurement of a property that can vary or change. each variable is (in) a column with a column header.
  - any other observation of a variable (value) is in a different row.
- Each row represents a record.

elevation	latitude	longitude	population	country	province	name
247	9	49	585.890	D	Baden-Württemberg	Stuttgart
97	8	49	290.117	D	Baden-Württemberg	Mannheim
115	8	49	289.173	D	Baden-Württemberg	Karlsruhe
278	8	48	209.628	D	Baden-Württemberg	Freiburg
114	9	49	146.751	D	Baden-Württemberg	Heidelberg
261	9	49	114.411	D	Baden-Württemberg	Pforzheim
478	10	48	116.761	D	Baden-Württemberg	Ulm

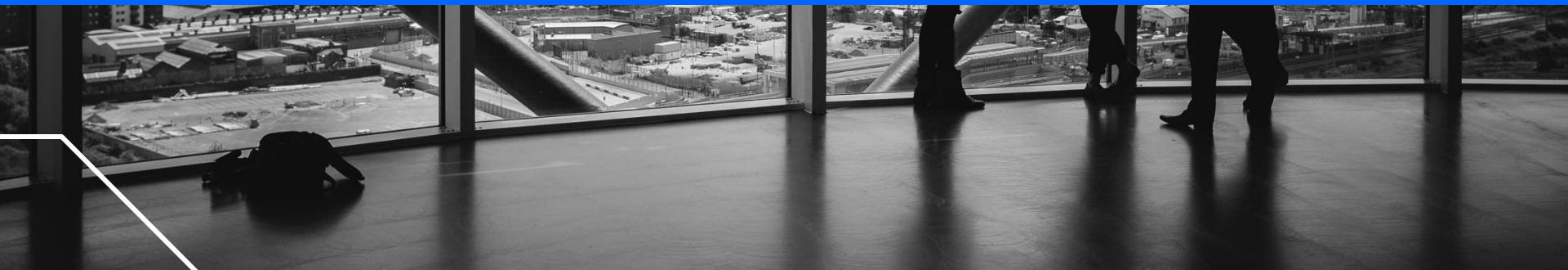
Database

- Data can also be poorly structured.
- Variables (fields) are not in a column with a column heading at a time. Not every observation is on a different line.
- Labels (headings) are inserted several times as rows above the column headings or as
- additional columns.

- What should you do if your data is not well structured?
- change the underlying database
- use of a programming language (e.g., Python) for modification
- use ETL (Extract, Transform, Load) tools

	Structured Data	Unstructured Data
Data Definition	• Has clearly defined data types	• Undefined and stored in its native format
	• Stored in rows and columns, can be mapped to fields	• No predefined model
Data Analysis	• Easy to search and process by humans and algorithms	• Difficult to search and process
Data Nature	• Quantitative in nature	• Qualitative in nature
	• Processing methods include clustering, regression, relationships, and classification	• Not processed and analyzed using conventional tools
		• Processing methods used include data mining and stacking
Data Storage	• Stored in data warehouses in a relational database	• Stored in data lakes in non-relational (NoSQL) databases
	• Require little storage space	• Requires more storage space
Data Format	• Format: numbers and text	• Wide variety of data sizes and shapes, from imagery to email, audio, video, etc.
	• The data format is defined beforehand	• It has no data model and requires no transformation

Data preparation & cleansing

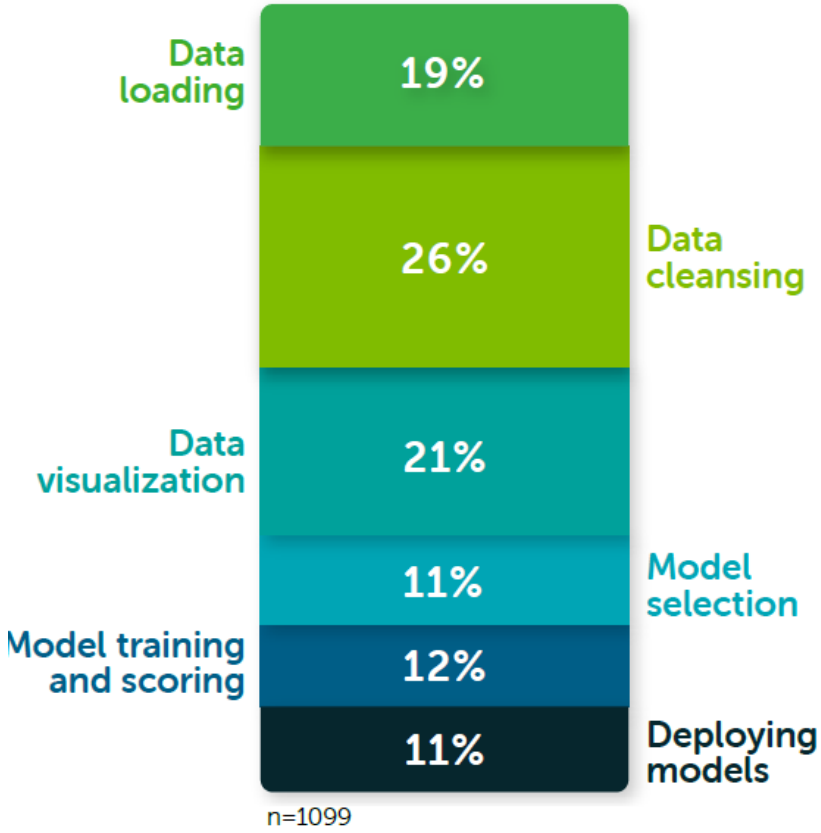


Introduction to Data Preparation and Cleansing

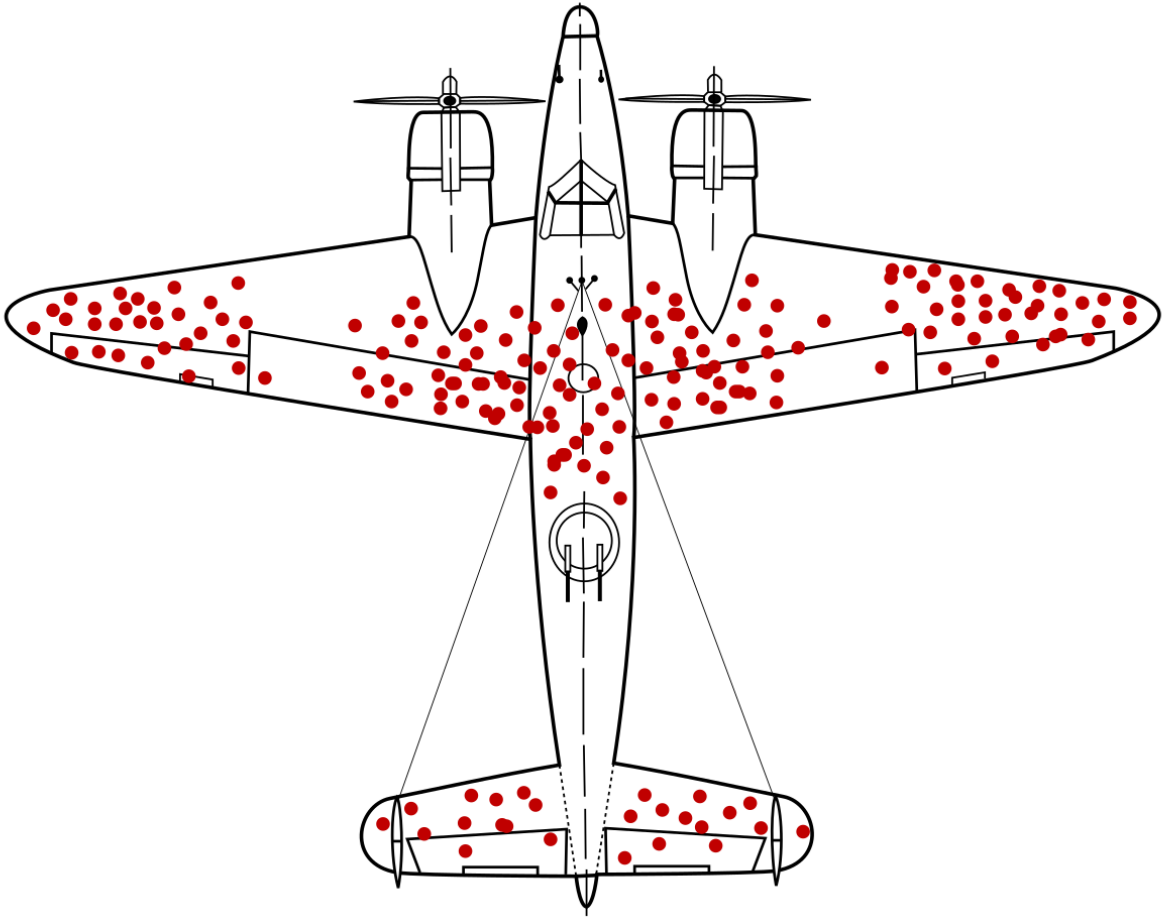
- **Data Preparation** and **Cleansing** is a fundamental step in the machine learning workflow that involves getting the data ready for analysis and model building.

- This step is crucial because the quality and format of your data can determine the performance of your machine learning algorithms.

45% of the time for preparation



Anaconda, 2020, State of Data Science: Moving From Hype Toward Maturity




The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data. – John Turkey (1986)

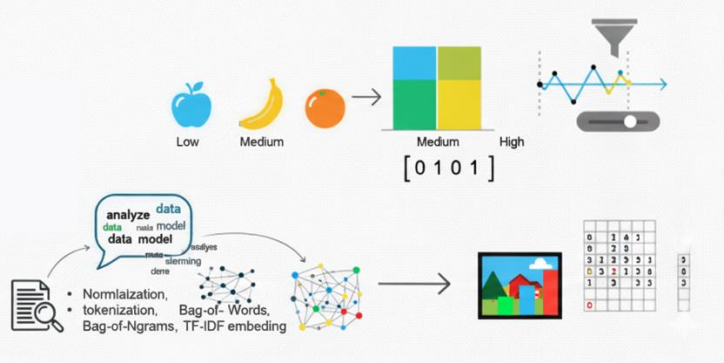
Importance of Data Preparation and Cleansing

- **Accuracy**: Clean data is representative of the problem at hand and leads to more accurate analysis and predictions.
- **Efficiency**: Algorithms work more efficiently with data in a consistent format.
- **Reliability**: Decision-making based on analysis of clean data is more reliable.

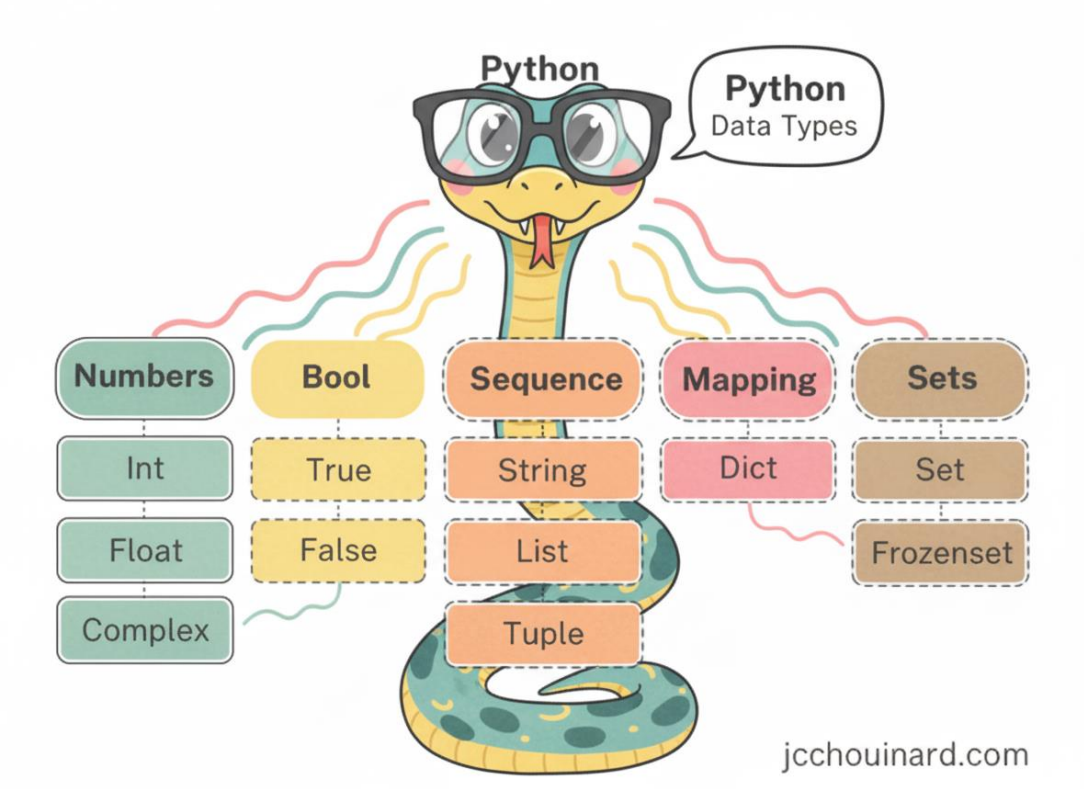
Data Preprocessing

Transform extracted attributes first. For example:

- Number to normalized number, e.g., min-max normalization
- Number to category, e.g., binning (quantile, equidistant) Category to numeric vector, e.g., one-hot encoding Text to numeric vector, e.g.:
  - Normalization, tokenization, stemming Bag-of-Words, Bag-of-Ngrams, TF-IDF Latent word embedding
  - Image to numeric value, e.g., color histogram



Data Type Conversion



Data Preparation

The process of converting raw data into a clean dataset.

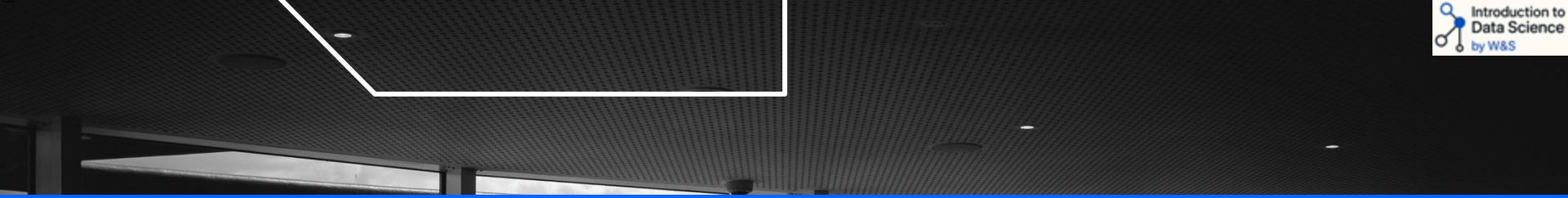
Data preparation might include:

1. **Data Collection**: Gathering data from multiple sources, which could be databases, files, sensors, or online repositories.
2. **Data Integration**: Combining data from different sources, which may require resolving data conflicts and inconsistencies.
3. **Data Transformation**: Converting data into a suitable format or structure for analysis. This might involve normalizing data (scaling data within a range, typically 0 to 1, or -1 to 1), encoding categorical variables (changing text-based categories into numerical values), or creating new variables through feature engineering.

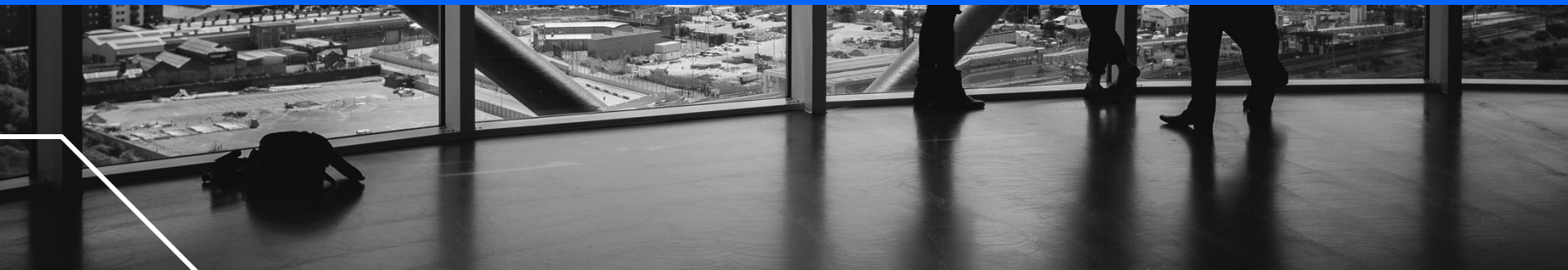
Data Cleansing

Once the data is in a preliminary structured format, data cleansing takes place, which involves:

1. **Handling Missing Data**: Data can have missing values due to various reasons, and dealing with them is essential. Techniques include imputing the missing values using statistical methods (like mean, median, or mode) or predictive modeling, or discarding the rows or columns with missing data altogether.
2. **Identifying and Correcting Errors**: This includes spotting and rectifying mistakes or inconsistencies in the data, such as typos, incorrect entries, or mislabeled classes.
3. **Removing Duplicates**: Duplicate data can bias the analysis, so it's important to identify and remove any duplicates from the dataset.
4. **Detecting and Filtering Outliers**: Outliers are data points that deviate significantly from the rest of the dataset. They can be due to variability in the data or experimental errors, and they can affect the results of the analysis.



Missing values

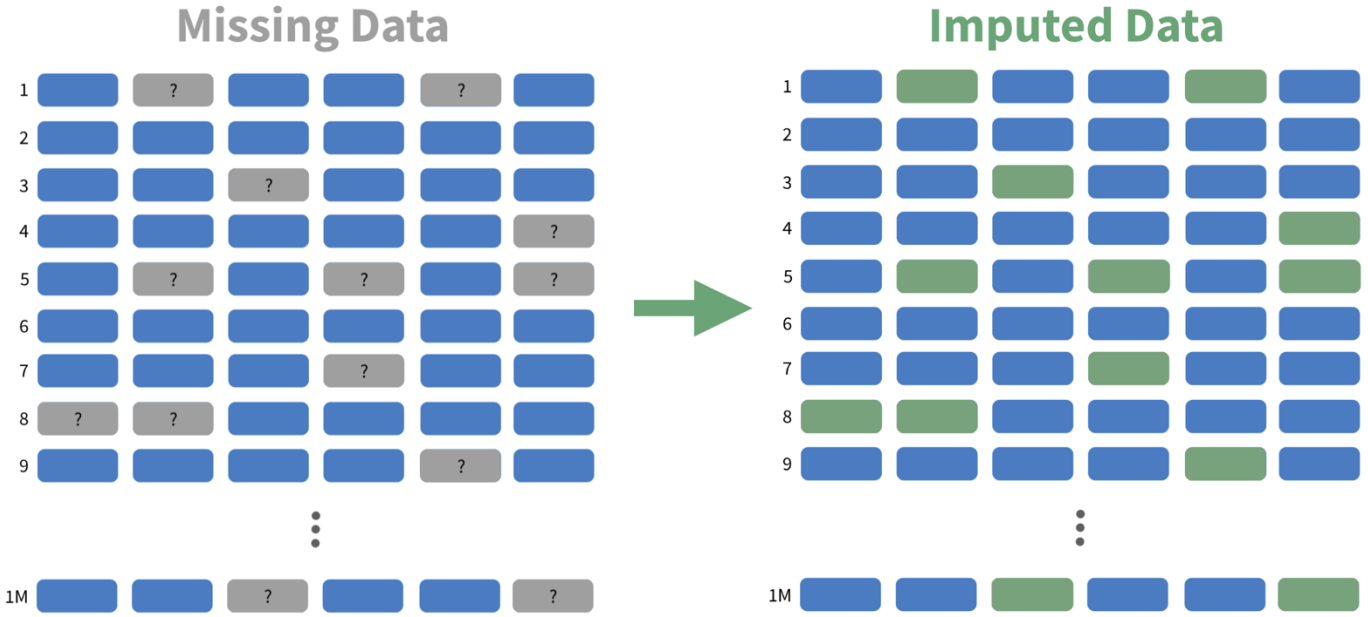


Understanding Missing Values

What are Missing Values? Missing values (data) occur(s) when no data value is stored for a variable in an observation.

Implications of Missing Values: The absence of data can lead to biased estimates, loss of efficiency, and complications in analyzing and interpreting results.

Understanding Missing Values



htt ps://www.google. com/ur l? sa=i&url=h ttps%3A%2F %2F toward sdatascience .com%2Fh ow-to-han dle-missing-dat a-b55 7c9e82f a0%2 F&psig=AOvVaw0ez.lig_9o 1N8 8kVT h1pt2N&ust= 176021 044669 2000&sour ce=imag es&cd=vfe&opi= 89 784498,ved= 0CBg QjhxqFwoT CQjomY WImpA DFQAAAAAdAAAAABB H

Types of Missing Data

- **Missing Completely at Random (MCAR)**
  - Definition: The probability of a data point being missing is the same for all cases.
  - Implication: Analyses can be unbiased, but power is reduced due to a smaller sample size.

- **Missing at Random (MAR)**
  - Definition: The likelihood of a data point being missing is related to the observed data but not the missing data.
  - Implication: Statistical techniques can be employed to handle MAR data, provided the model is correctly specified.

- **Missing Not at Random (MNAR)**
  - Definition: The probability of a data point being missing is related to the unobserved data, i.e., the missingness is related to the missing value itself.
  - Implication: MNAR can lead to biased analyses if not properly addressed.

Identifying Missing Data Patterns

Exploratory Analysis

- Analyzing patterns of missingness.
- Investigating the relationship between missing and observed values.

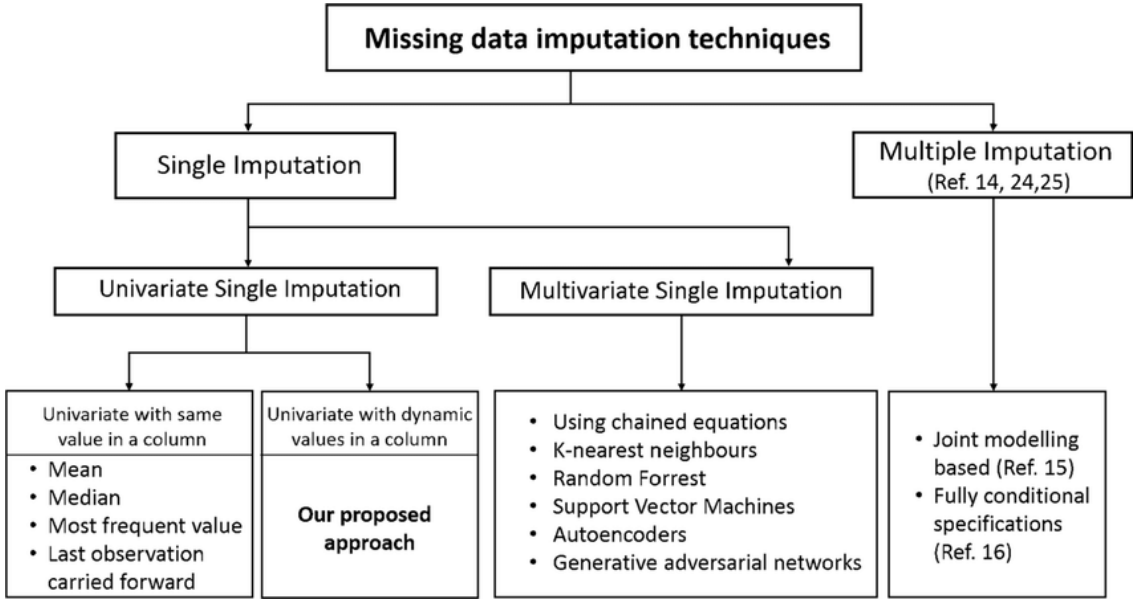
Sensitivity Analysis

- Assessing how different assumptions about the missing data affect the results.

Challenges in determining missing data

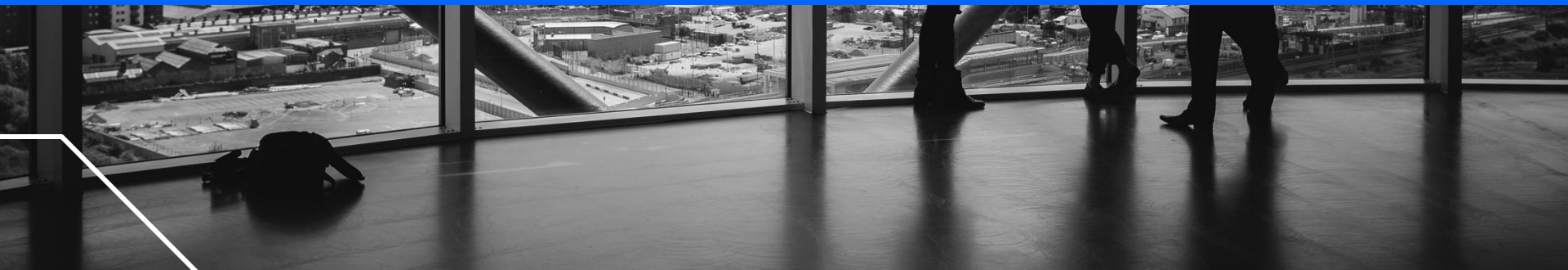
- Often, the true mechanism of missingness is unknown and must be inferred.
- The analysis is contingent on assumptions which should be tested for robustness.

Imputing Missing Data

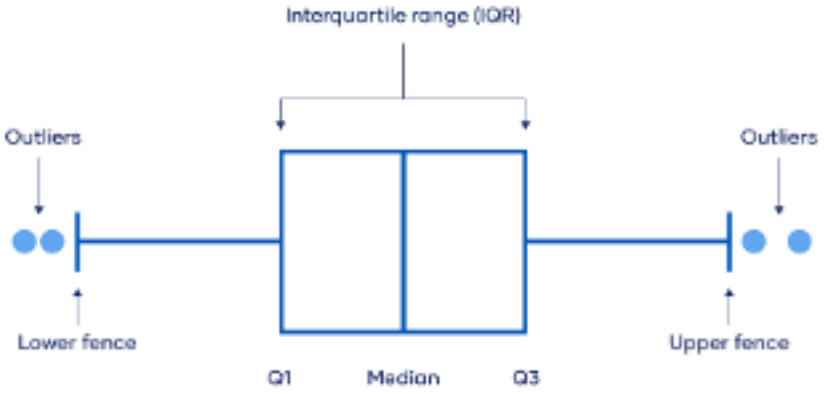


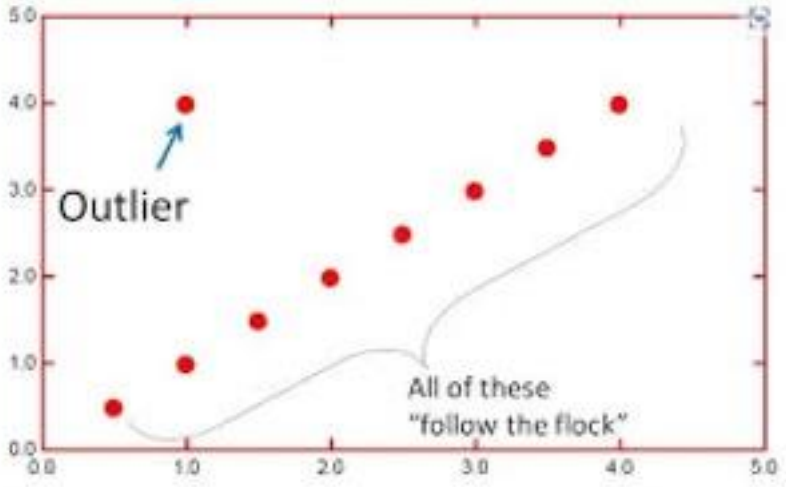
https://www.researchgate.net/figure/A-broad-categorization-of-missing-data-imputation-techniques_fig1_358529185

Data Cleansing



Understanding outliers

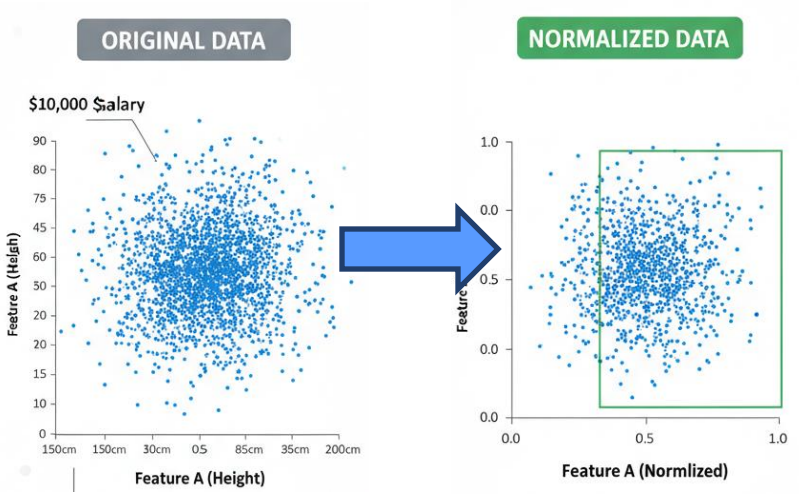


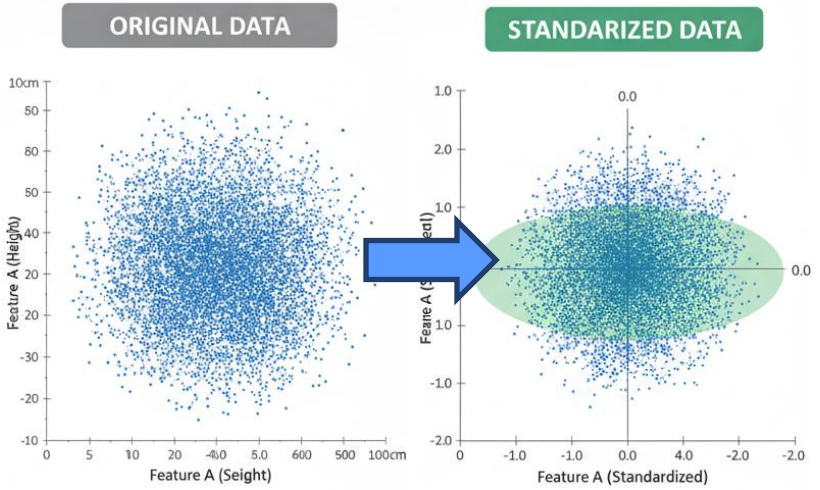


Outlier Detection Methods:
IQR, Z-score, Isolation Forest with formulas/examples

Outlier Handling:
Removal, capping, transformation strategies

Further pre-processing





Normalization scales numerical data to a fixed range, typically between 0 and 1. This is useful when features have different ranges and you want to prevent features with larger values from dominating the learning process.

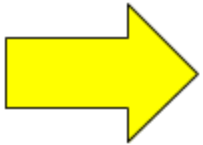
Standardization, on the other hand, transforms data to have a mean of 0 and a standard deviation of 1. This is particularly useful when your data has a Gaussian distribution or when algorithms assume normally distributed data.

Normalization Formulas

Purpose: Scale data to a standard range for better comparison and model performance		
	Min-Max Normalization	**Z-Score (Standardization)**
Formula	(x - min) / (max - min)	(x - μ) / σ μ = mean, σ = standard deviation
Result	Scales data to range [0, 1]	Centers data around 0 with unit variance
Example	Dataset: [10, 20, 30, 40, 50] Normalize x = 30: (30 - 10) / (50 - 10) = 20/40 = **0.5** Results: [0, 0.25, 0.5, 0.75, 1.0]	Dataset: [10, 20, 30, 40, 50] Mean (μ) = 30, Std Dev (σ) = 14.14 Normalize x = 30: (30 - 30) / 14.14 = **0** Results: [-1.41, -0.71, 0, 0.71, 1.41]
When to use	When you need bounded range; sensitive to outliers	When data is normally distributed; handles outliers better

One hot encoding

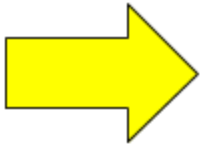
Color
Red
Red
Yellow
Green
Yellow



https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding

One hot encoding

Color
Red
Red
Yellow
Green
Yellow



Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

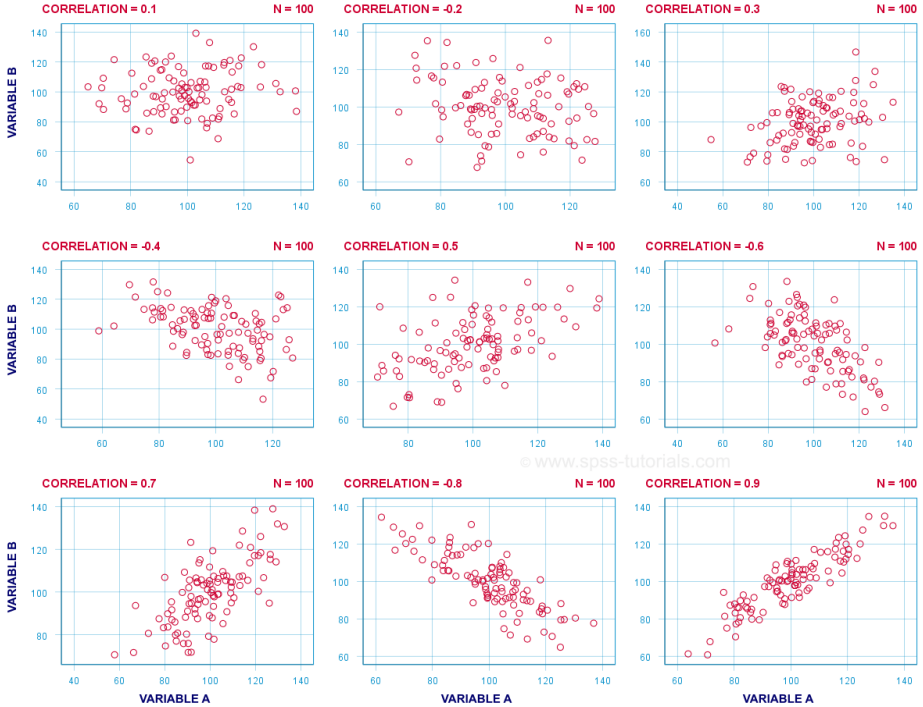
https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding

Featuring

Features = dimensions that describe each data point.

- Large number of features may have negative impact on performance of Machine Learning approaches.
- In the featuring step, the **number of features can be reduced** by deleting non-relevant data.
  - Challenge: How to identify non-relevant data in each use case?
- **Feature extraction** can be performed in this step which consists on deriving new dimensions to describe the data (=> getting more/ new features).

Featuring



SPSS Correlation Analysis - The Ultimate Guide


HANDS-ON
TIME
INTRODUCTION TO DATA SCIENCE
BY W&S
YOUR TURN!
EXERCISE / QUESTION

Bis Dienstag!

samuel.schlenker@hpe.com

Tag 6: Einführung in maschinelles Lernen

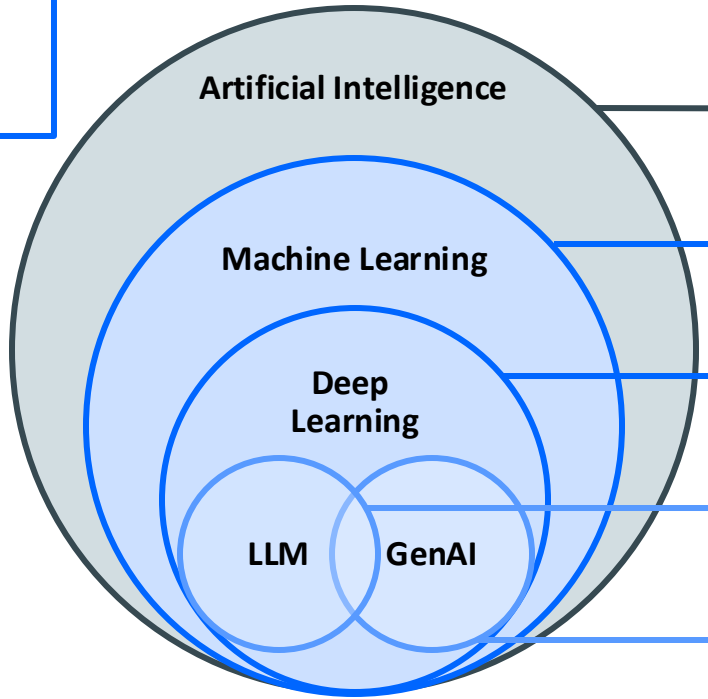
Samuel Schlenker

11.11.2025, WWI 2025F

Students should ...

- Define machine learning as a subset of artificial intelligence focused on learning from data
- Understand machine learning as an optimization task and parametric programming approach
- Differentiate between supervised, unsupervised, and reinforcement learning paradigms
- Apply the train-test split methodology and understand cross-validation
- Understand classification tasks and algorithms (K-Nearest Neighbors, Decision Trees, SVM)
- Understand regression tasks for predicting continuous variables
- Apply unsupervised learning techniques like K-means clustering for pattern discovery
- Recognize the differences between supervised and unsupervised learning approaches
- Understand the basic architecture of neural networks (input layer, hidden layers, output layer)
- Identify key hyperparameters (learning rate, batch size, epochs, activation functions)
- Understand the role of activation functions in neural networks
- Recognize the model training process and the importance of loss/cost functions

Machine Learning



Artificial Intelligence (AI)
Any technology that enables machines to solve tasks in a way like humans do

Machine Learning (ML)
Algorithms that allow computers to learn from examples without being explicitly programmed (supervised & unsupervised)

Deep Learning (DL)
Using deep artificial neural networks as models, inspired by the structure and function of the human brain

Large Language Models (LLM)
Models trained on massive datasets to understand and generate human-like text across diverse subjects

Generative Artificial Intelligence (GenAI)
Refers to technologies that utilize machine learning models to generate human-like text, images, or other content

Machine Learning

Machine Learning is a **subset** of **AI** focused on developing algorithms that enable computers to learn from and make predictions on data.
The transition from AI to ML represents a shift from rule-based systems to **data-driven approach**es.

It's motivated by the goal of **reducing the need for explicit programming**, allowing machines to adapt to new scenarios independently.

Machine Learning

Generic term for the artificial generation of knowledge based on experience.

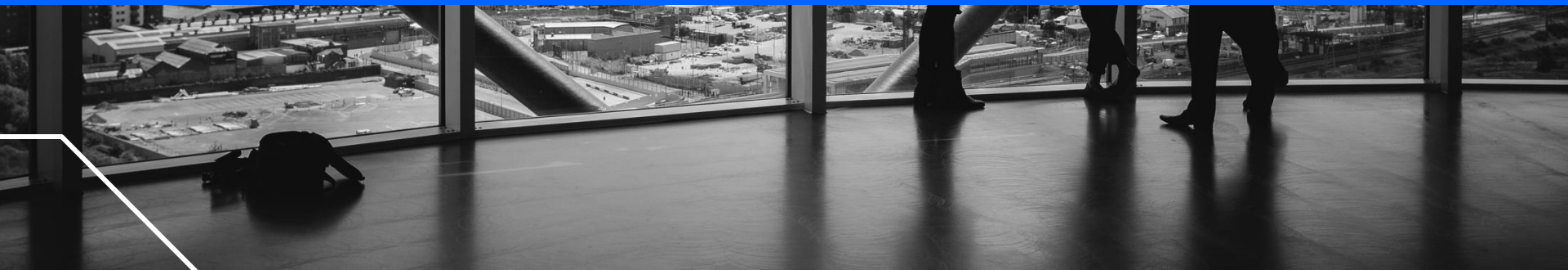
In machine learning, a system

- learns based on examples
- generalizes knowledge (after the learning phase)
- recognizes patterns and laws (based on learning data)
- can evaluate unknown data (learning transfer) or
- fail to learn unknown data (e.g. under- or overfitting)

Machine Learning

Machine learning as an optimization task
Machine learning as probabilistic inference
Machine learning as parametric programming
Machine learning as an evolutionary search
...

Modeling through learning



Differentiation

Model

- Representation of a situation (reality)
- Representation using a modeling language

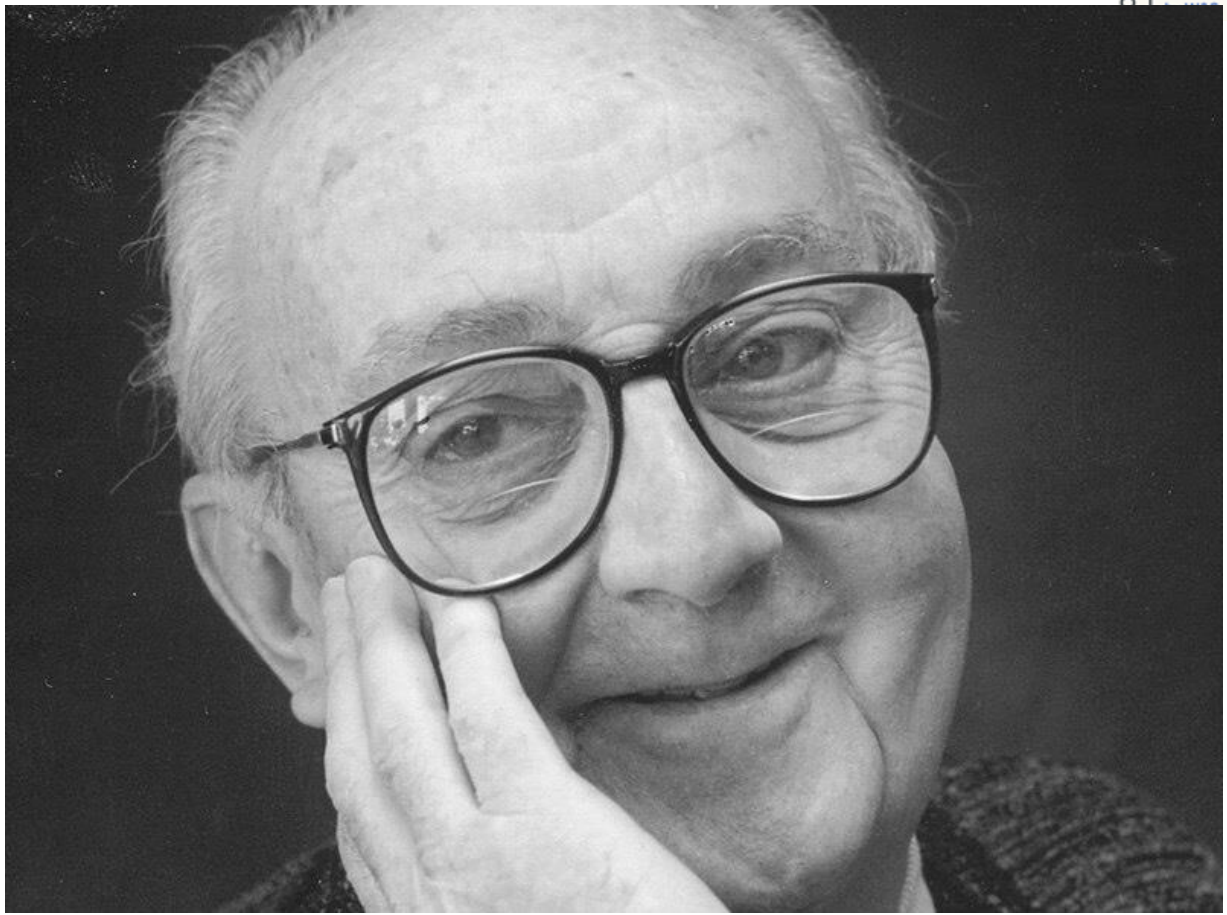
Modeling language

- Consists of a class of artifacts
- Allows these artifacts to be characterized in the form of descriptive properties and relationships between the characterized artifacts
- Describes the models belonging to this modeling language as artifact characteristics

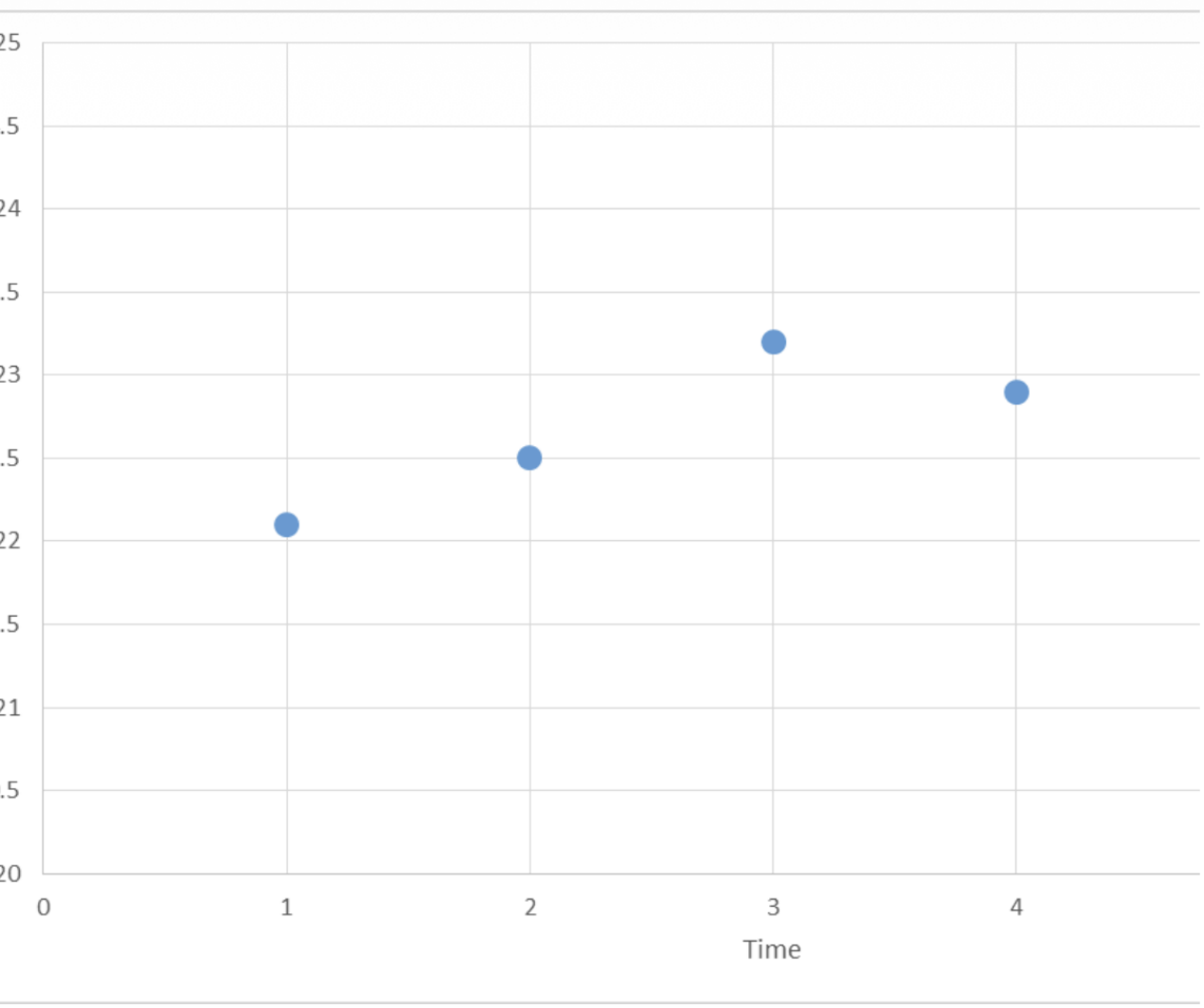
Metamodel

- A modeling language of a model (M1) can itself be regarded as a model (M2) that can be represented by a modeling language.
- This model M2 is called the metamodel for M1 ... (!)

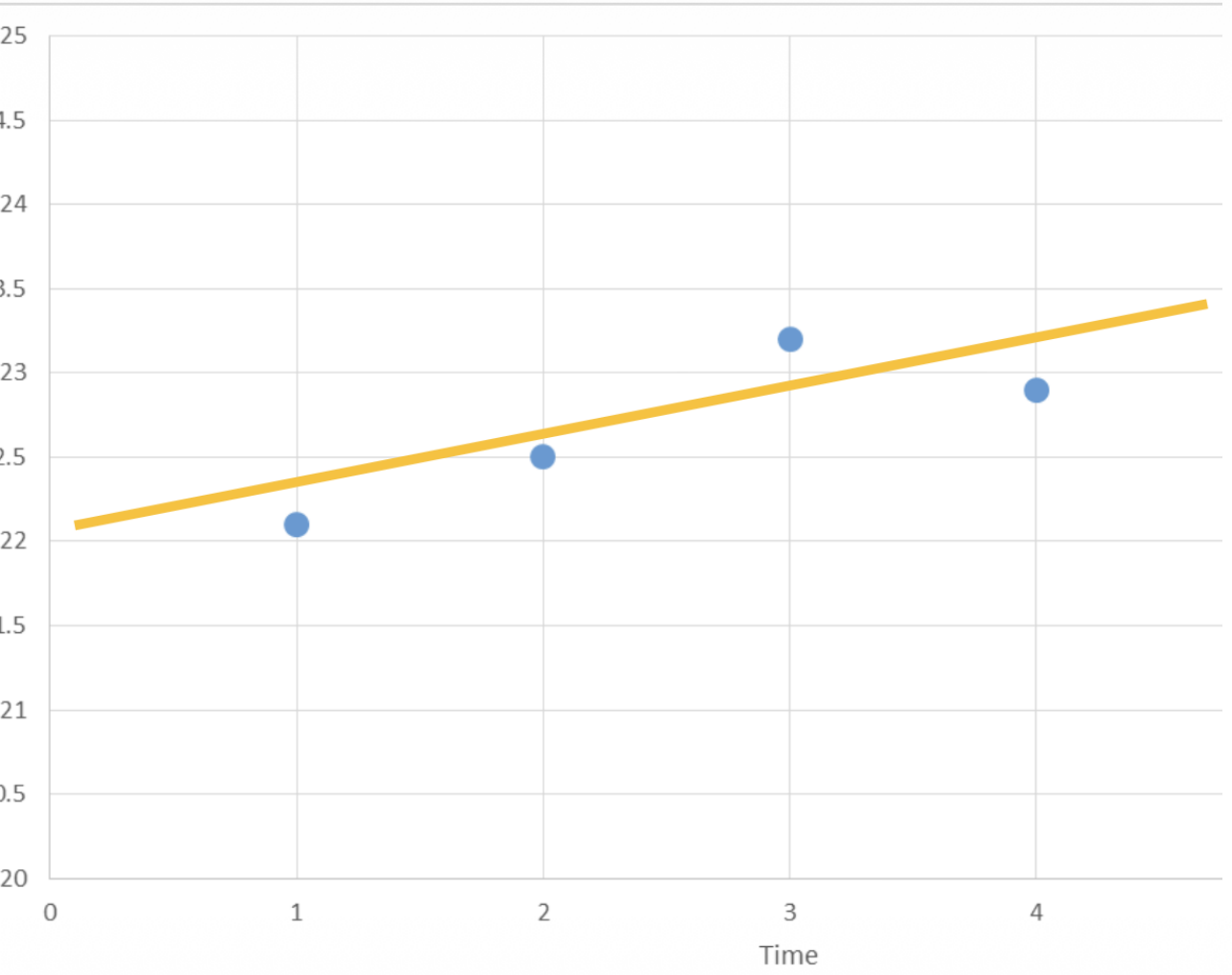
All Models are wrong but some are useful



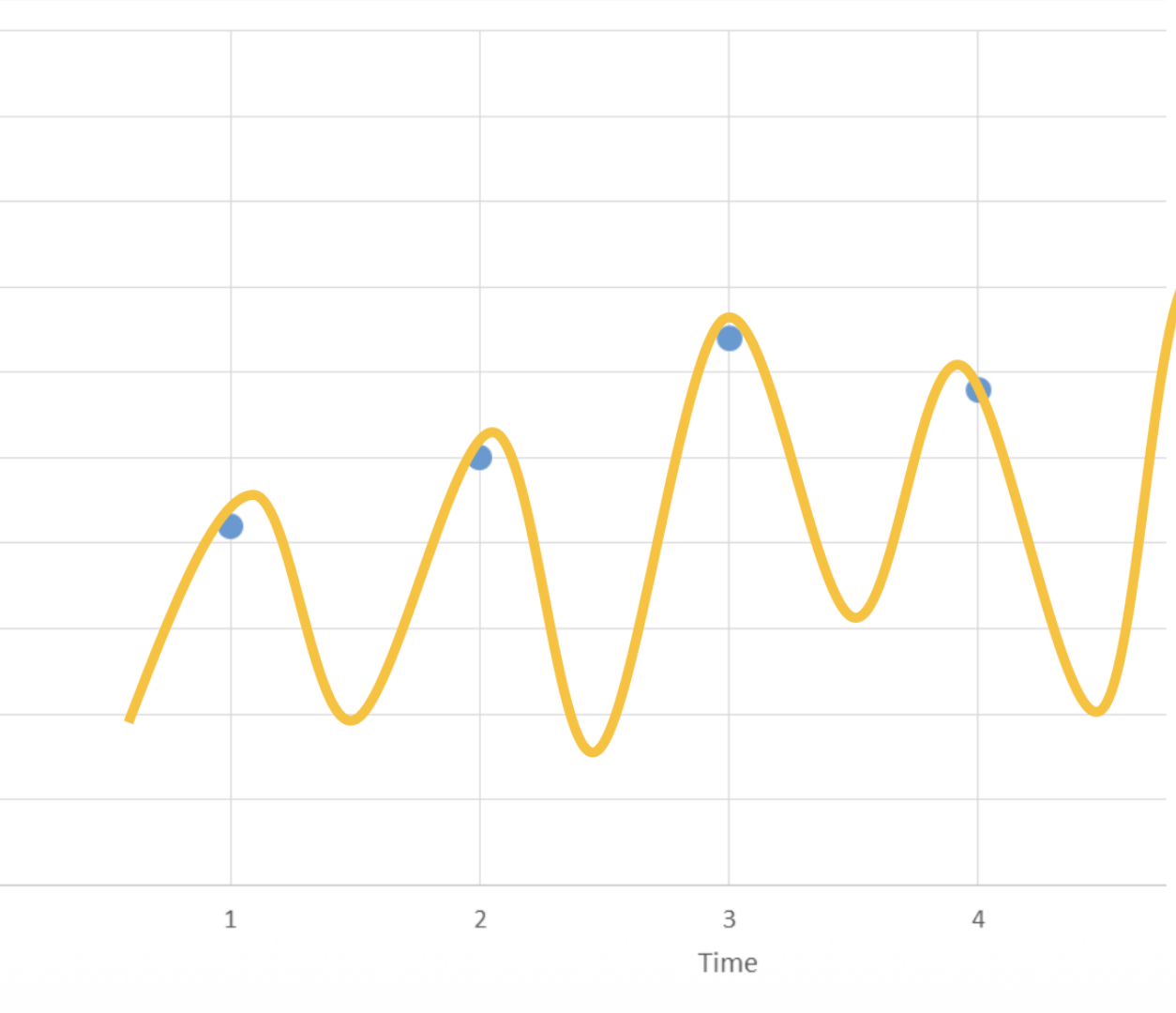
George Box



DO YOU UNDERSTAND THE DATA FULLY?

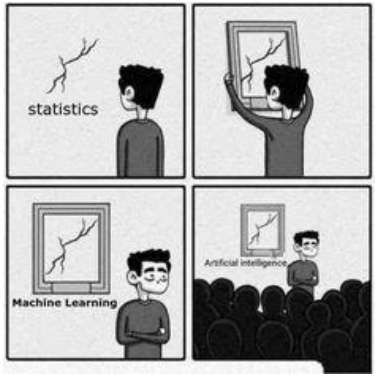


DO YOU UNDERSTAND THE DATA FULLY?



DO YOU UNDERSTAND THE DATA FULLY?

Machine Learning

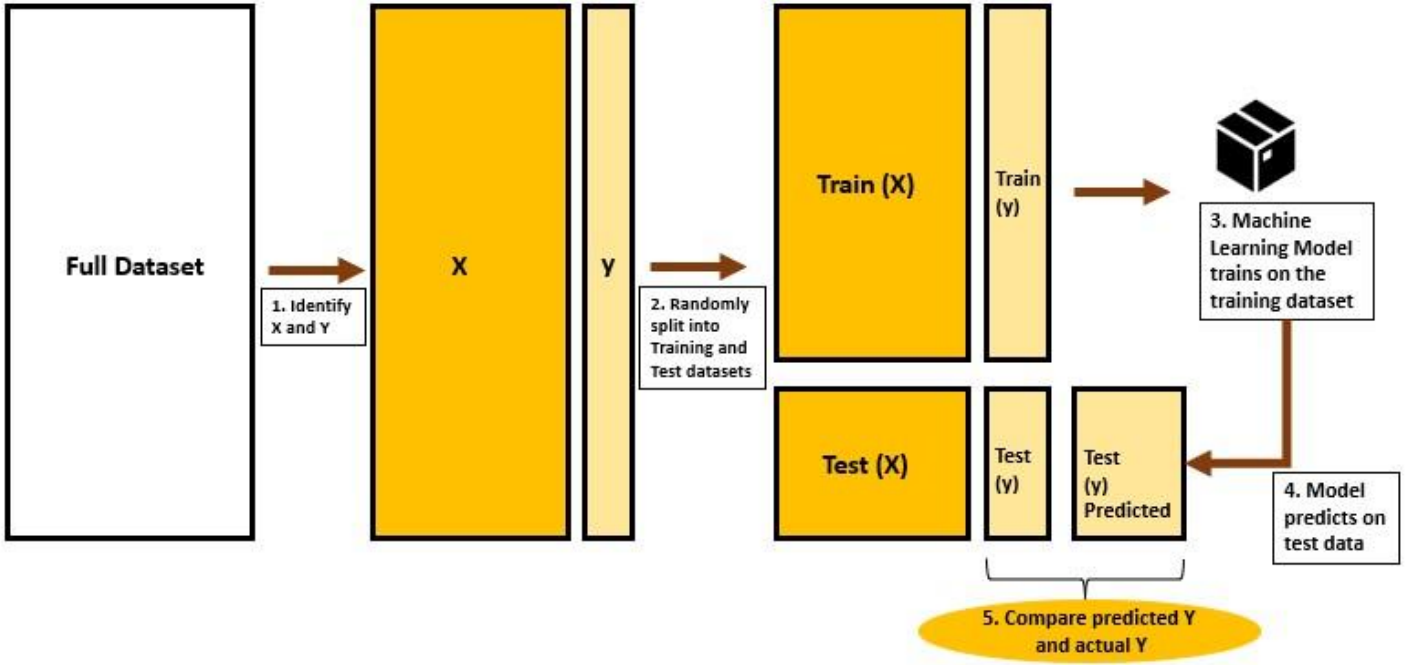


Training: the model is trained using a training data set where the output/target variable is known

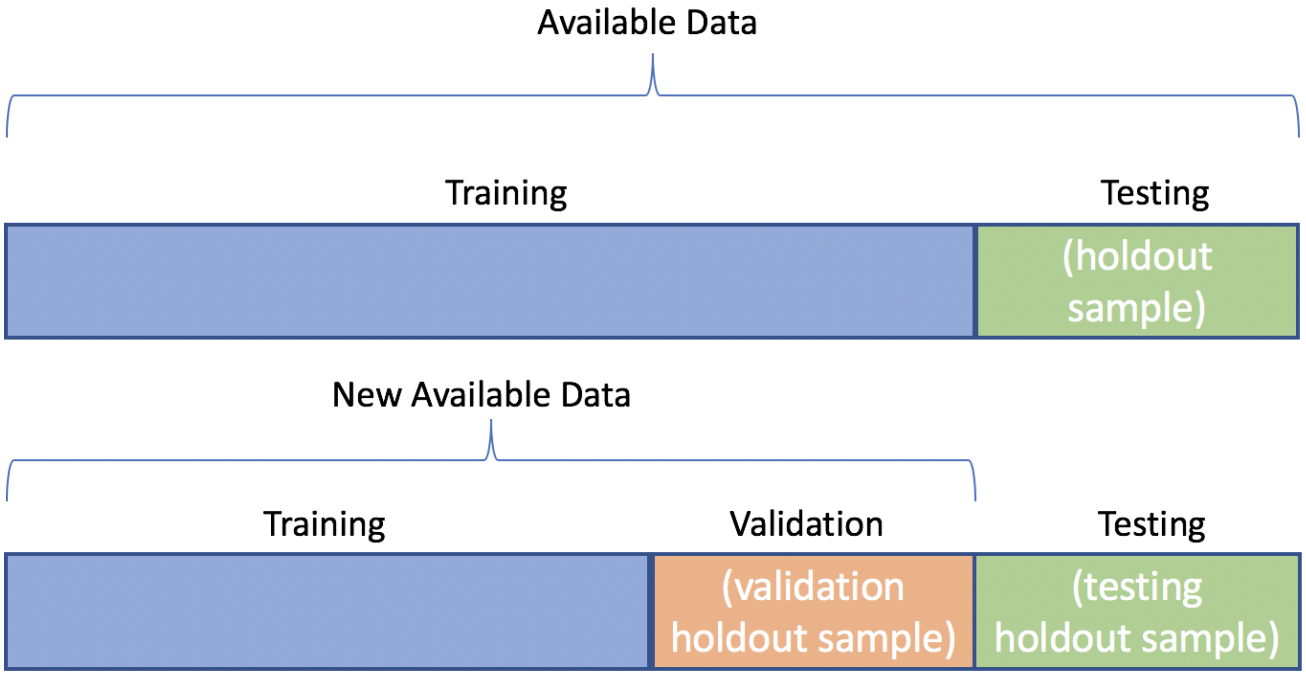
Testing: the model is tested using a test data set where the output/target variable is known in order to check the accuracy of the model

Application: the model delivers results for a new data set with an unknown target variable

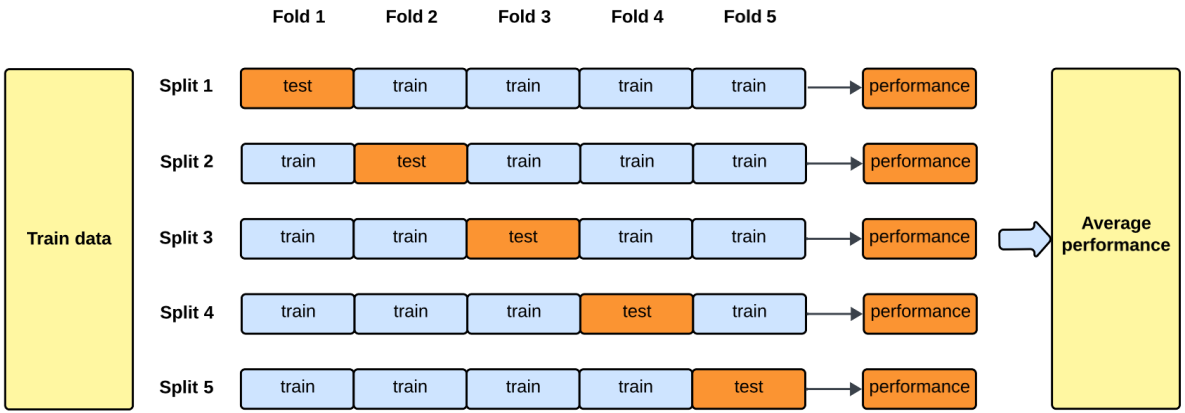
Train test split



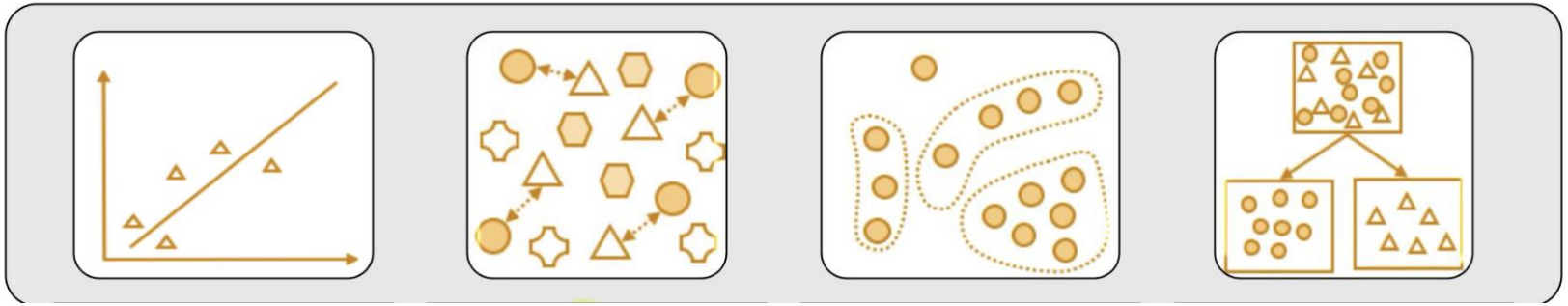
Validation data



Cross validation

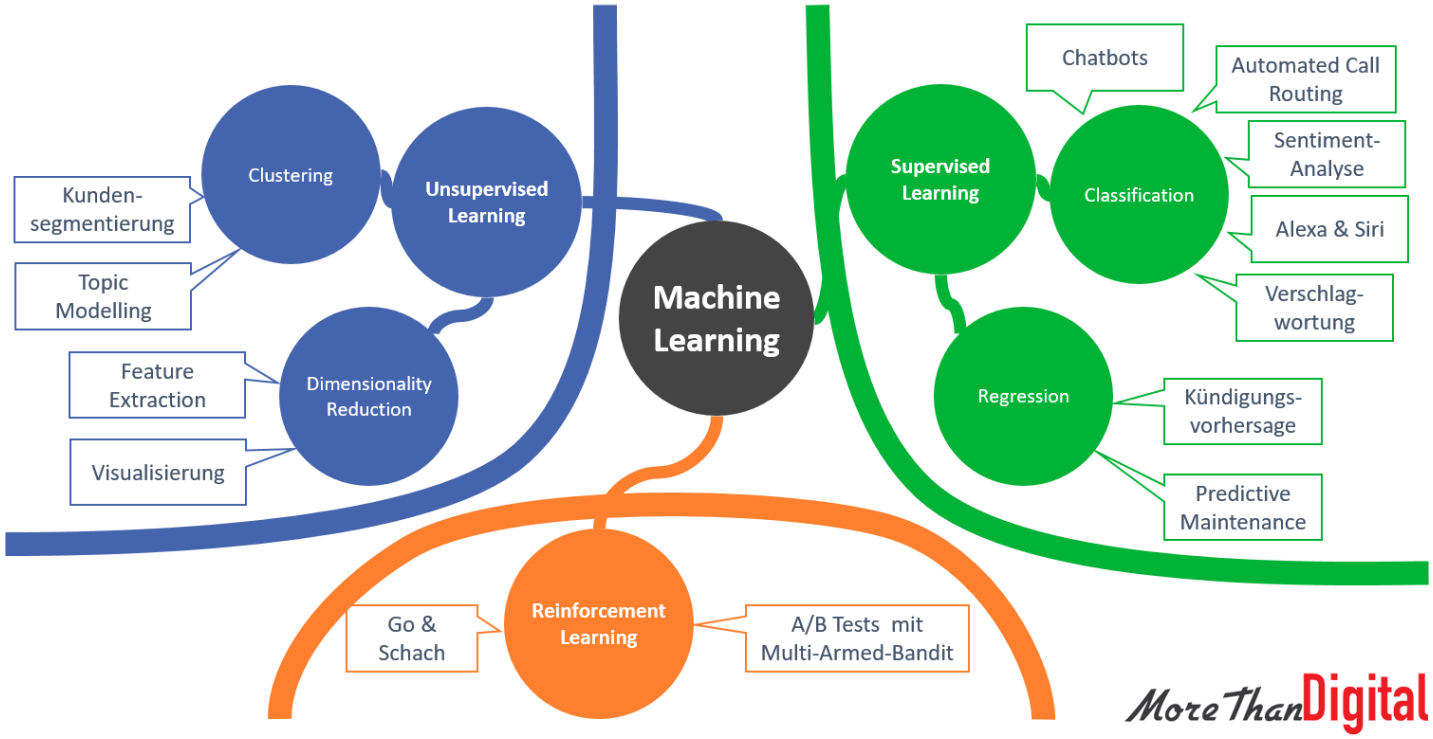


Modeling in general



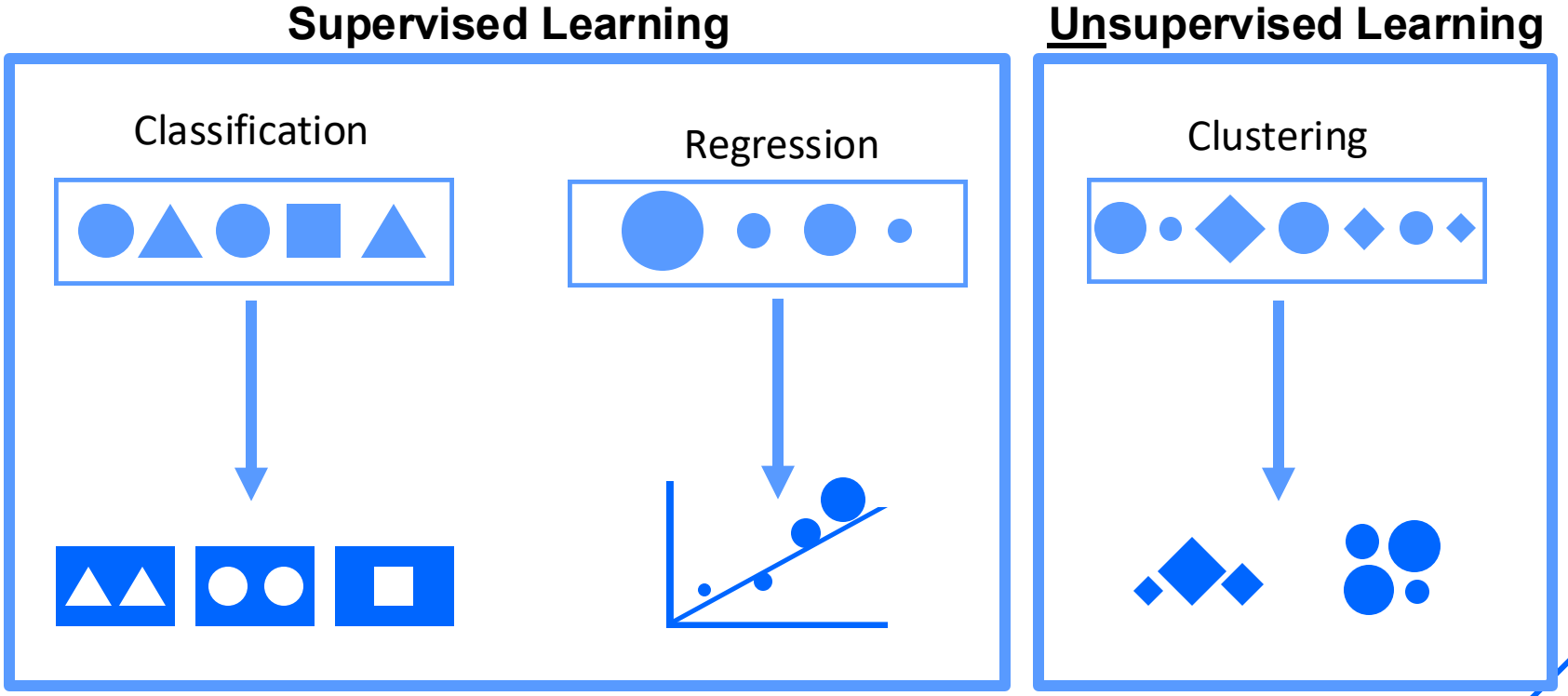
Forecasts	Association	Segmentation	Classification
Identification of trends	Search for dependencies	Finding homogeneous subsets	Division into predefined classes
Sales and revenue forecasts (e.g., sales/production planning)	Analysis of shopping baskets (product recommendations)	Creating customer portfolios (differentiated marketing)	Churn analysis (customer loyalty measures)

Machine Learning

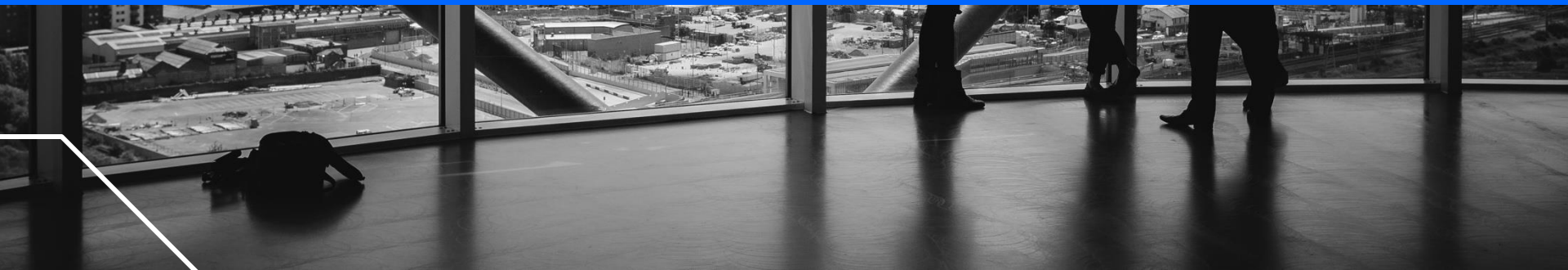


https://morethandigital.info/wp-content/uploads/2017/12/Machine-Learning-Übersicht-der-Möglichkeiten-und-Teilbereiche.png

Common ML Tasks



Supervised learning



Supervised learning

Models learn to predict outcomes based on **labeled training data**.

For a given a data set in which

$xi \in \mathbb{R}n$ is a vector (that contains descriptions), and

$yi \in \mathbb{R}$ is an observable result for xi, where

$xi, yi \propto p\ x, y$ independent identically distributed

Goal: find function that approximates y by a given x

$$f\ xi \approx yi$$

Important:

$$f\ xnew \approx ynew$$

Classification

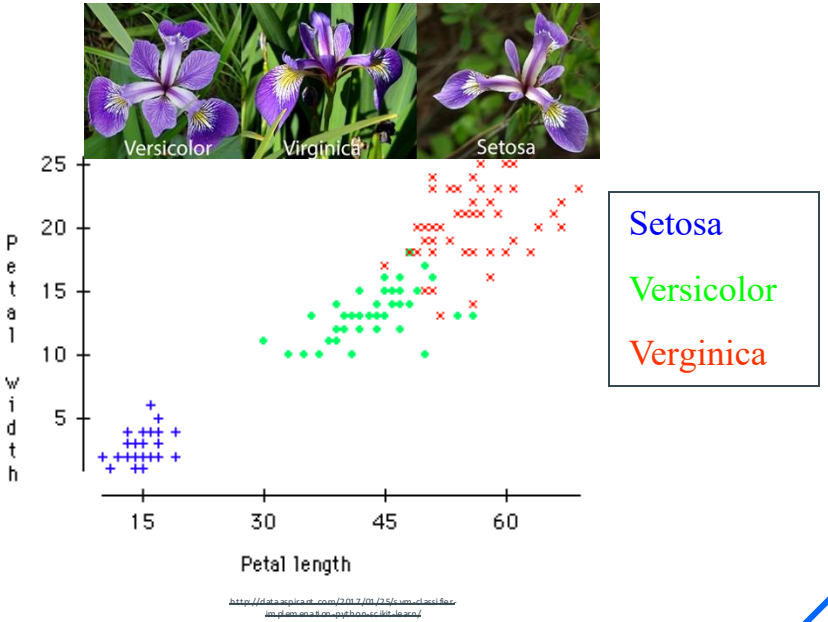
Objective: Assign input to a specified class.

To do this, parameters are determined using a training data set that has already been classified (assigned to classes), which can then be used to determine which category the input belongs to.

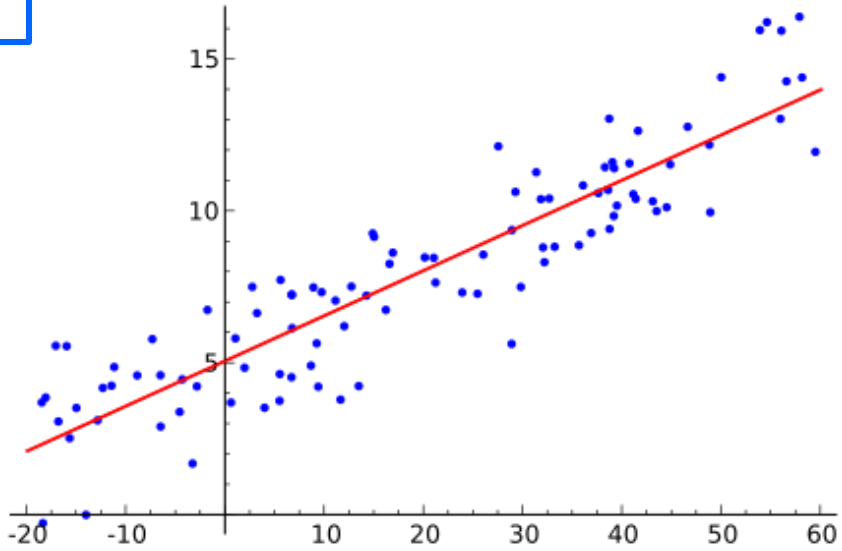
Example: Spam email detection, digit recognition, image recognition

Types of classification algorithms:

- K-Nearest Neighbors
- Decision Tree Classification
- Random Forest Classification
- ...



Regression



Goal: Make predictions for a variable.

To do this, dependencies between variables are determined, the dependent variable (input to be predicted) and independent variables (predictors)

Example: Weather forecasting, predicting market trends

Types of regression algorithms:
Simple linear regression
Multiple linear regression
Decision tree regression
Random forest regression
...

Supervised learning - Example

Imagine you have a dataset containing information about houses such as their size (in square feet), number of bedrooms, number of bathrooms, and distance from the city center.

Each house in the dataset has a corresponding price (the observable result) associated with it.

xi would be a vector containing the features of a house, such as size, number of bedrooms, number of bathrooms, and distance from the city center.

yi would be the price of the house.

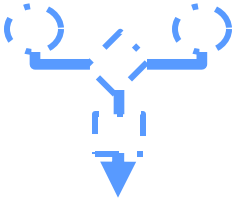
xi, yi represents a pair of features and its corresponding price.

The goal in this scenario is to find a function $f\ xi$ that can approximate the price yi given the features xi .

Essentially, we want to find a function that accurately predicts the price of a house based on its features.

Regression

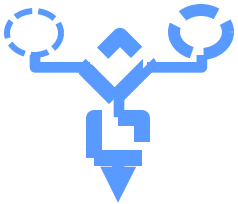
Simple example: ML model predicts the selling price of a house



Algorithm
Mathematical function

Algorithm

Result = weight1*var1 + weight2*var2 + weight0

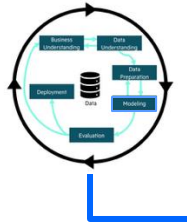


Model
Algorithm with weighted variables

Model

Price in $US = 100*(square_feet) + 10000*(bedrooms) + 100,000

How do you find the right weights for the model?
Training



Supervised training

1st pass

Record 1
sqft: 1500
BR: 3
…
Label:
314,000

Untrained model

Record 1
sqft: 1500
BR: 3
…
Label:
280,000

1. Assess distance from correct label
2. Adjust model

2nd pass

Record 2
sqft: 1750
BR: 4
…
Label:
380,000

Model—new weights

Record 2
sqft: 1750
BR: 4
…
Label:
346,500

nth pass

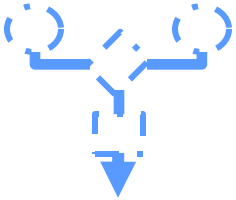
Record 9999
sqft: 1800
BR: 2
…
Label:
400,000

Trained model—final weights

Record 9999
sqft: 1800
BR: 2
…
Label:
400,000

Supervised training

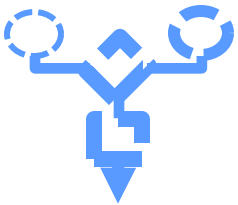
Simple example: ML model predicts the selling price of a house



Algorithm
Mathematical function

Algorithm

Result = weight1*var1 + weight2*var2 + weight0



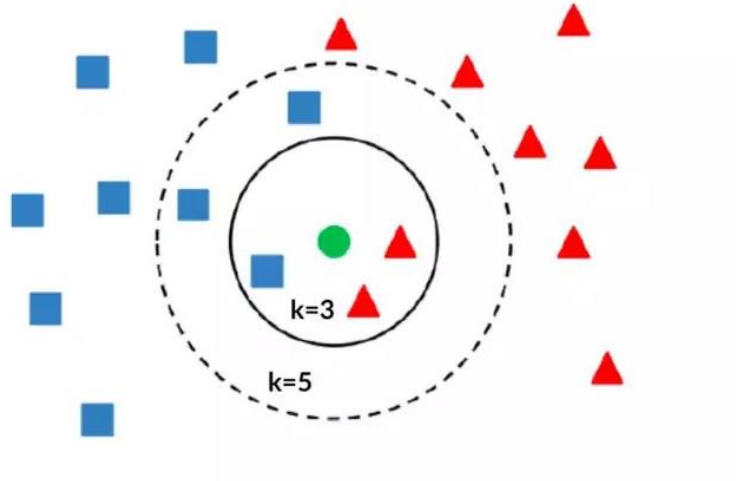
Model
Algorithm with weighted variables

Model

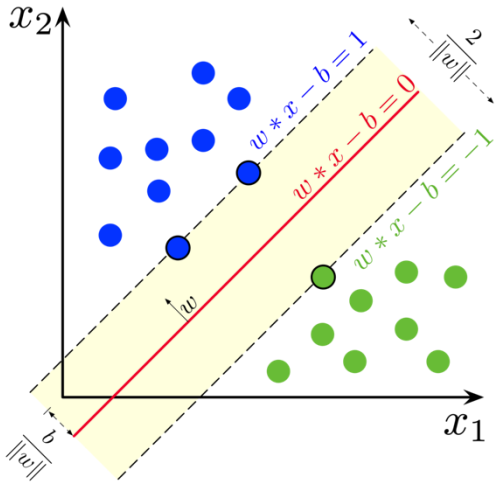
Price in $US = 100*(square_feet) + 10000*(bedrooms) + 100,000

How do you find the right weights for the model?
Training

Supervised Learning – More example Algorithms

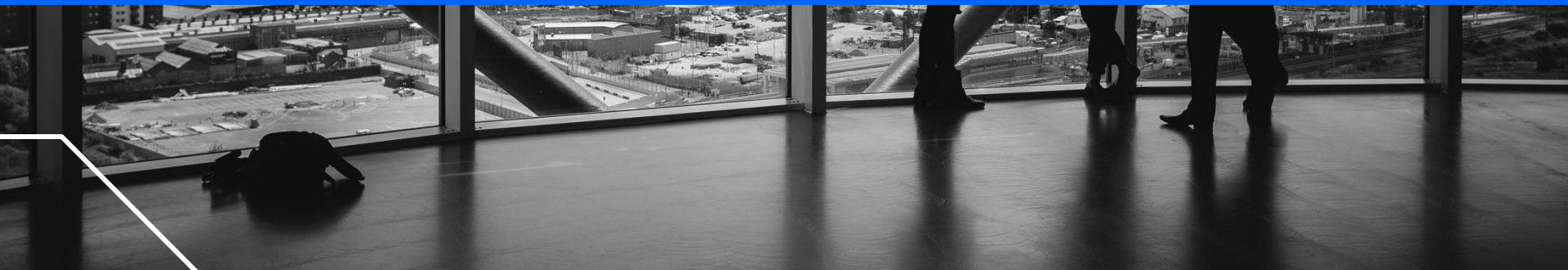


K-nearest neighbor (KNN)



Support Vector machine (SVM)

Unsupervised learning



Unsupervised learning

Models identify patterns and relationships in unlabeled data.

Model examples:

- **K-Means Clustering**: Partitions data into distinct groups based on feature similarity.
- **Principal Component Analysis** (PCA): Reduces dimensionality while preserving data
- variance.
- **Autoencoders**: Neural networks designed for unsupervised learning tasks.

Unsupervised learning

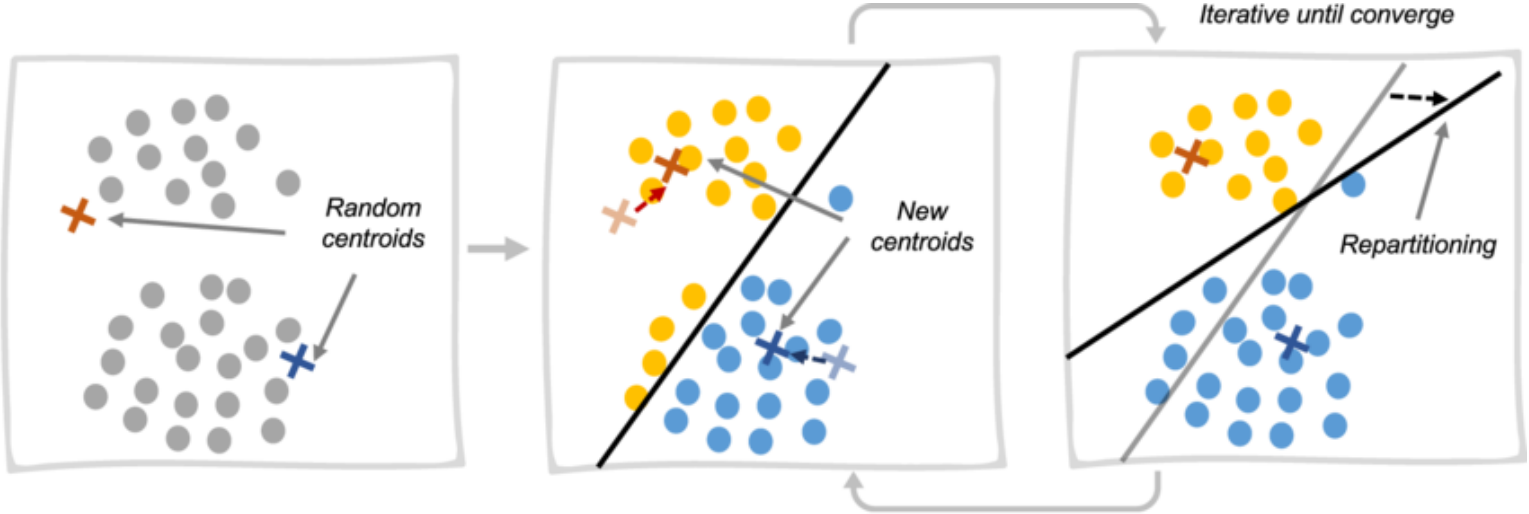
For a given a data set in which

$xi \in \mathbb{R}n$ is a vector (that contains data descriptions), and

$xi \propto p\ x$ independent identically distributed

Goal: learn about p

K-means clustering



Input: distance matrix *D* & number of clusters *k*

https://www.researchgate.net/figure/Illustration-of-K-means-Clustering-Note-The-estimation-routine-of-K-means-involves-i_fig2_371115662

Unsupervised learning - Example

Imagine analyzing customer transaction data from an online retail store. We explore features like items purchased, purchase amount, and time of purchase. Our aim is to uncover hidden patterns in customer behavior without predefined labels.

Goal:

Understand customer behavior distribution $p\ x$ without labels.
Discover item associations and customer segments.

Approaches:

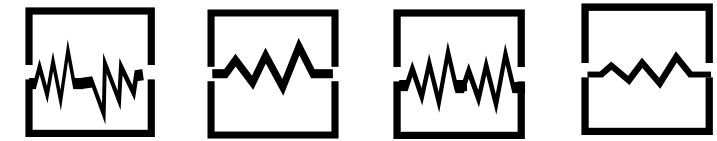
Market Basket Analysis: Identify frequently co-occurring items.
Customer Segmentation: Group customers with similar behaviors.

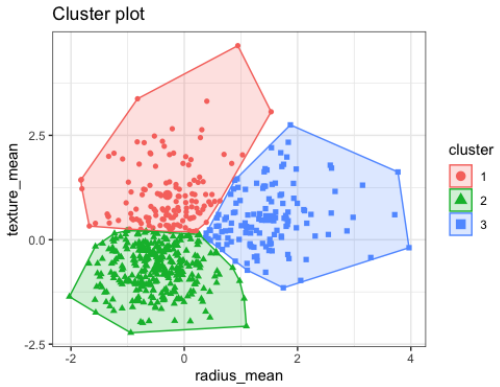
Benefits:

Tailor marketing strategies and product recommendations.
Optimize engagement without manual labeling.

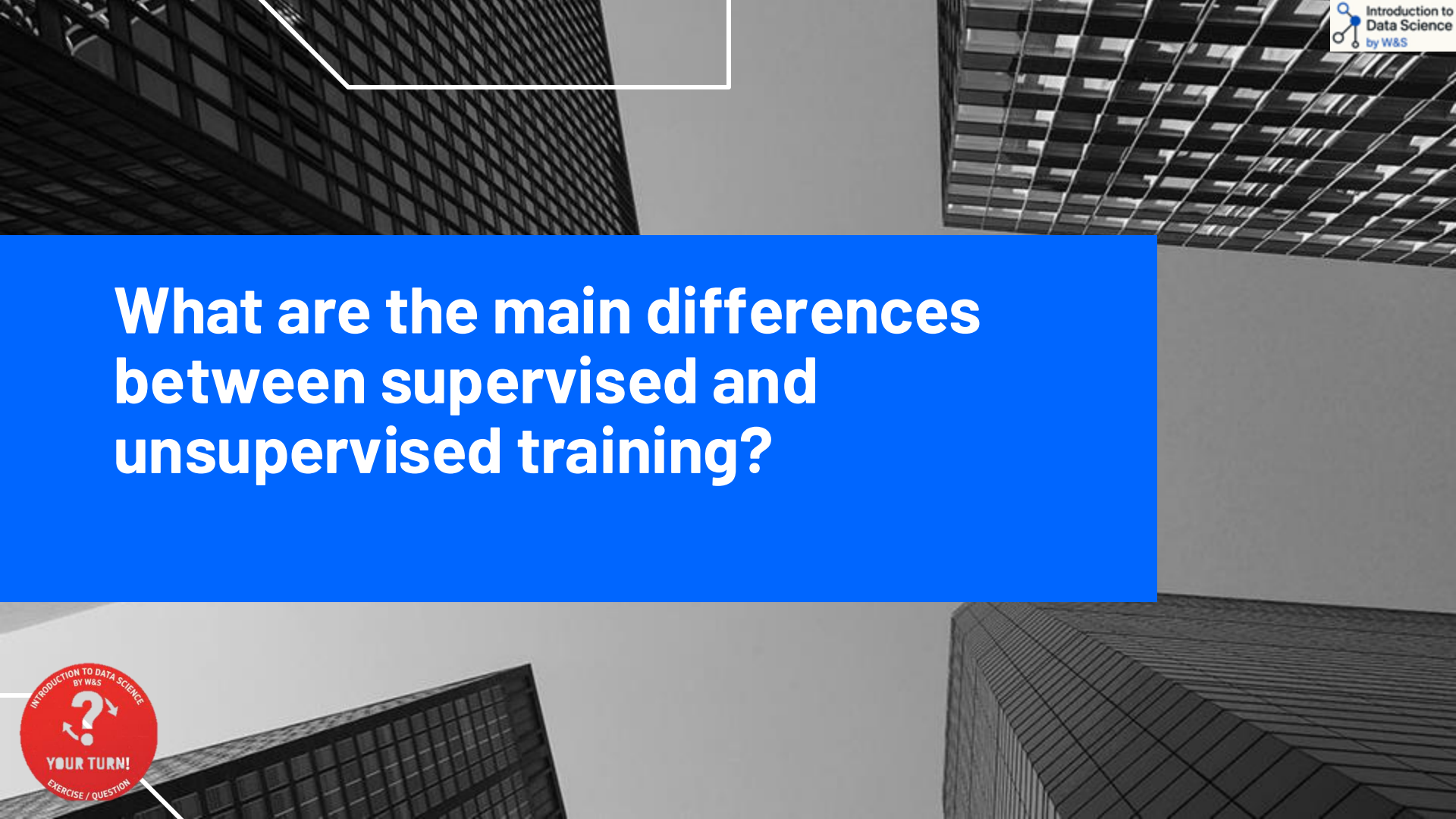
Unsupervised learning

No labels





Learns to cluster or detect anomalies


Introduction to Data Science
by W&S
What are the main differences between supervised and unsupervised training?
INTRODUCTION TO DATA SCIENCE BY W&S
YOUR TURN!
EXERCISE / QUESTION

Reinforcement Learning

Reinforcement Learning (RL) is a type of machine learning where an agent learns to make decisions by interacting with an environment.

Key Components:

- **Agent:** Makes decisions and takes actions.
- **Environment:** The external system with which the agent interacts.
- **Rewards:** Feedback signal indicating the quality of actions.
- **Policy:** Strategy or algorithm used by the agent to make decisions.

Objective:

Maximize cumulative reward over time by learning optimal policies.

Reinforcement Learning - example

Imagine training a self-driving car using reinforcement learning. The scenario involves the self-driving car (agent) navigating a road network and responding to traffic conditions.

Objective:

Teach the car to navigate roads safely and efficiently.

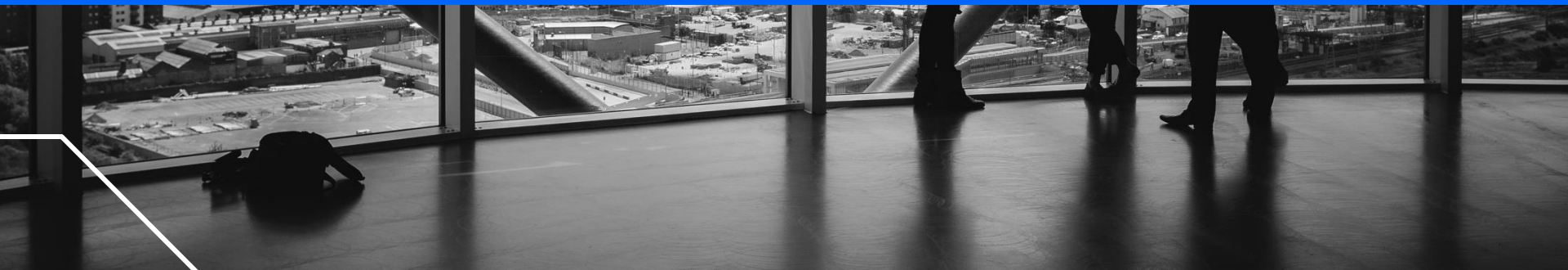
Benefits:

Enables adaptive and intelligent driving behavior.
Improves safety and efficiency on the roads.
Reduces human intervention and enhances autonomy.

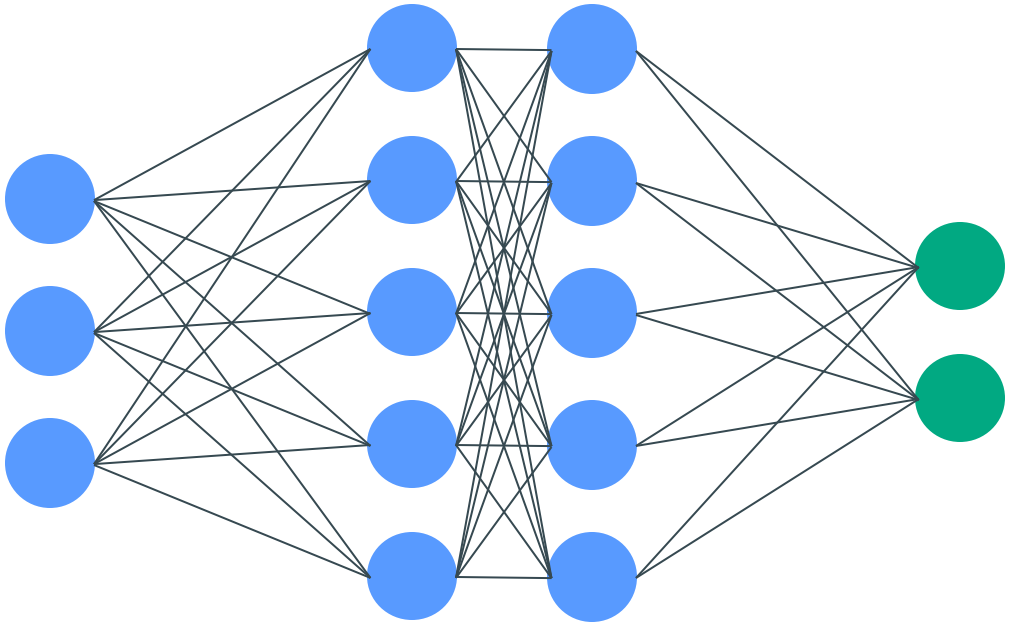
Further Classes of ML

- **Semi-Supervised Learning**: This is an approach to machine learning that involves a small amount of labeled data and a large amount of unlabeled data during training. Semi-supervised learning falls between supervised learning (with completely labeled data) and unsupervised learning (with no labeled data).
- **Self-Supervised Learning**: A type of unsupervised learning where the data provides the supervision. Here, the model is trained to predict part of the input from other parts of the input.
- **Transfer Learning**: This technique involves taking a pre-trained model (usually on a large dataset) and fine-tuning it for a specific task. It is very useful when you have a limited amount of data for your task.
- **Ensemble Learning**: This approach combines the predictions from multiple machine learning algorithms to make more accurate predictions than any individual model. Common ensemble methods include bagging, boosting, and stacking.
- **Multi-Instance Learning**: In this setting, labels are associated with groups of instances (bags), rather than individual instances. The task is to predict the labels of unseen bags based on the learned patterns from the labeled bags.
- **Multi-Label Learning**: Unlike traditional classification tasks where each instance is assigned to only one label from a set of disjoint labels, multi-label learning allows for the assignment of multiple labels to each instance.
- **Multi-Task Learning**: This is an approach to inductive transfer that improves learning for one task by using the information contained in the training signals of other related tasks.
- **Meta-Learning**: Sometimes called "learning to learn", it aims to design models that can learn new skills or adapt to new environments rapidly with a few training examples.
- **Active Learning**: A special case of machine learning where the learning algorithm can interactively query the user (or some other information source) to label new data points with the desired outputs.
- **Few-Shot Learning**: The goal here is to learn information about object categories from one or just a few training samples/images.

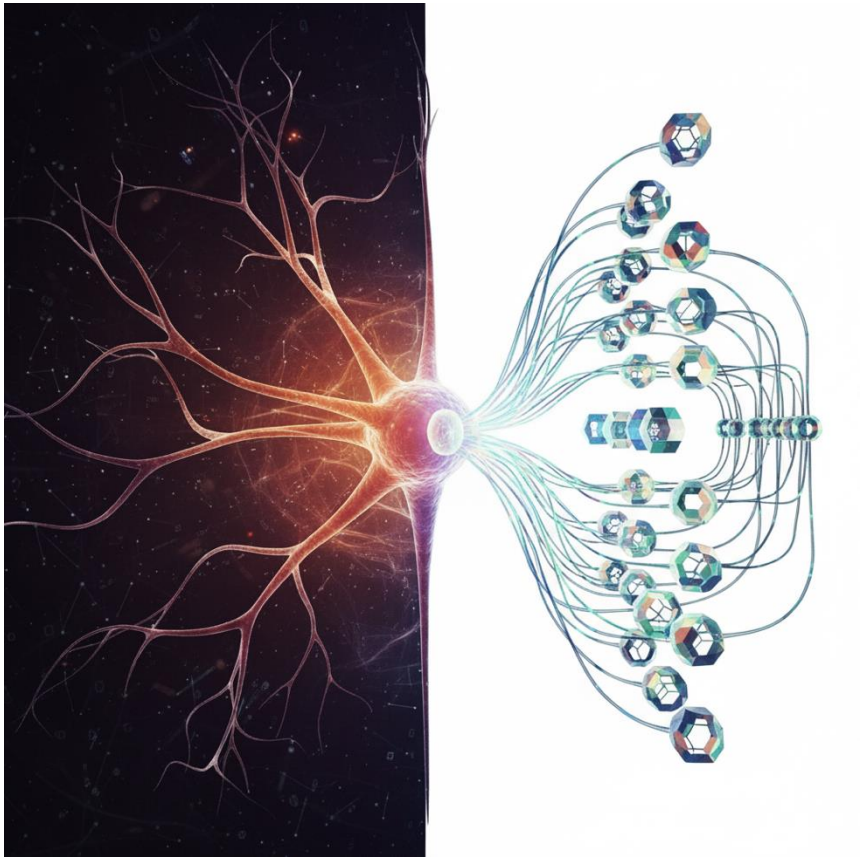
Neural networks



Neural networks



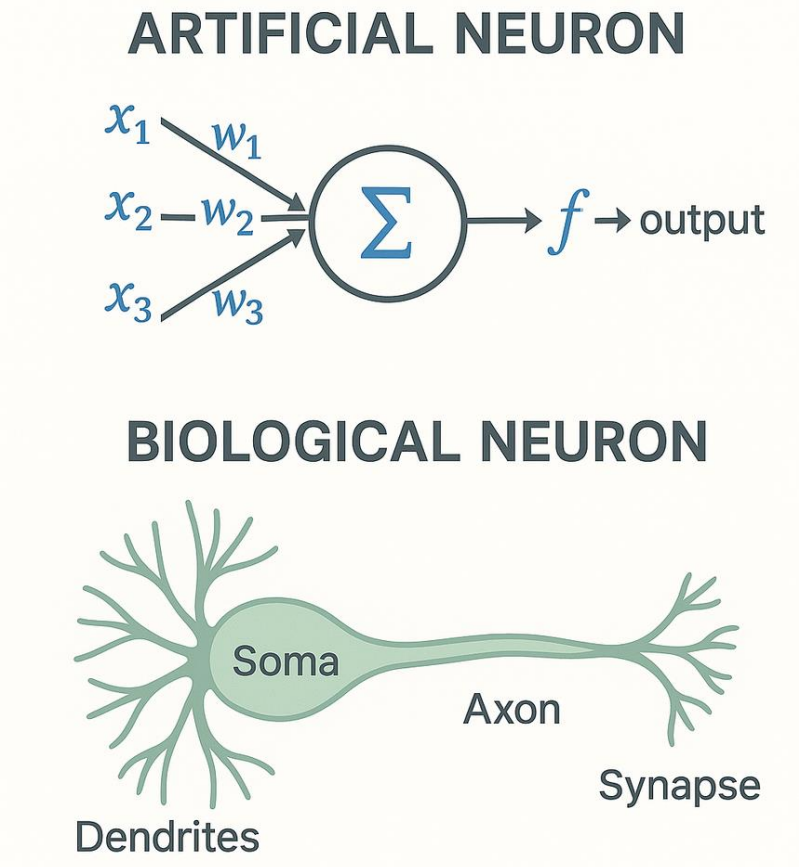
Artificial neurons



Model training

Train/Test Ratio:	What percentage of the available data is training data and what percentage is test data?
Batch Size:	How large is a data unit according to which the model is adjusted?
Epoch:	How often do I train the model (on the entire training data set)?
Learning Rate:	How strongly is weighting adjusted after an error is detected?
Loss/Cost:	How well does the algorithm perform?

Artificial neurons

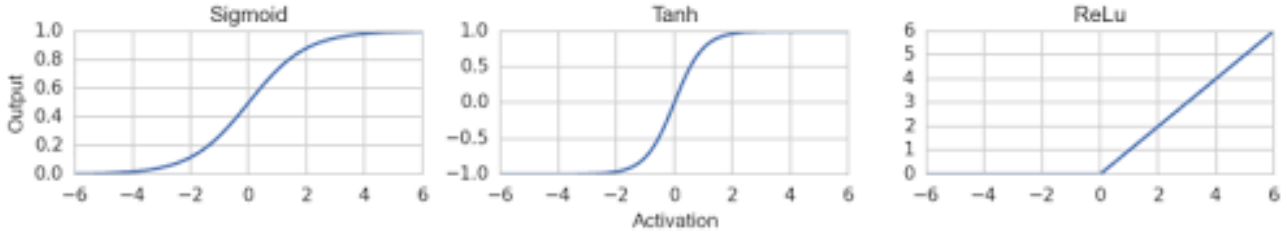


Activation function

Decides whether and how a neuron is “activated.”
There are various activation functions; the choice depends on the problem and the resulting requirements for the model.

Well-known activation functions:

- Sigmoid (output value 0 or 1) binary classification
- Tanh (output values between -1 and 1) if negative outputs are possible
- ReLU Rectified Linear Unit (0 if negative, if positive exactly the value it received), often used in classification

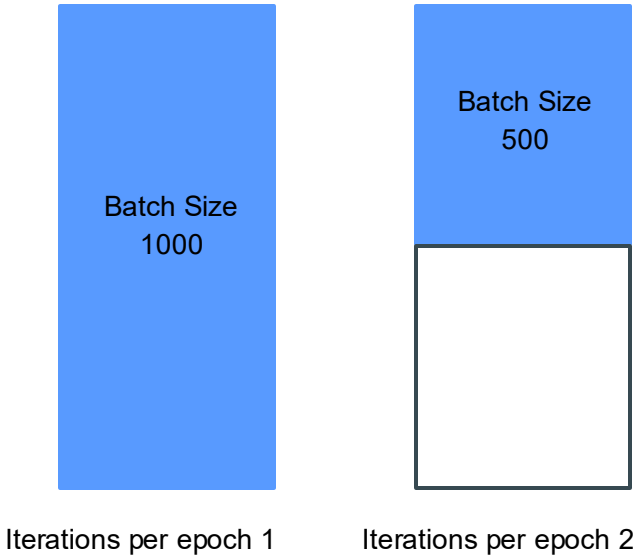


Hyperparameter

- Affects the performance of the model
- Tuning requires a lot of effort, but there are tools that automate this process

Hyperparameters:

- Train/test split ratio
- Epoch
- Batch size
- Learning rate
- Number of hidden layers
- Activation function
- Loss function / cost function
- Optimization algorithm




HANDS-ON
TIME
INTRODUCTION TO DATA SCIENCE
BY W&S
YOUR TURN!
EXERCISE / QUESTION

Bis Donnerstag

samuel.schlenker@hpe.com

Tag 8: Modellbewertung & Prüfungsleistung

Samuel Schlenker

11.11.2025, WWI 2025F

Students should ...

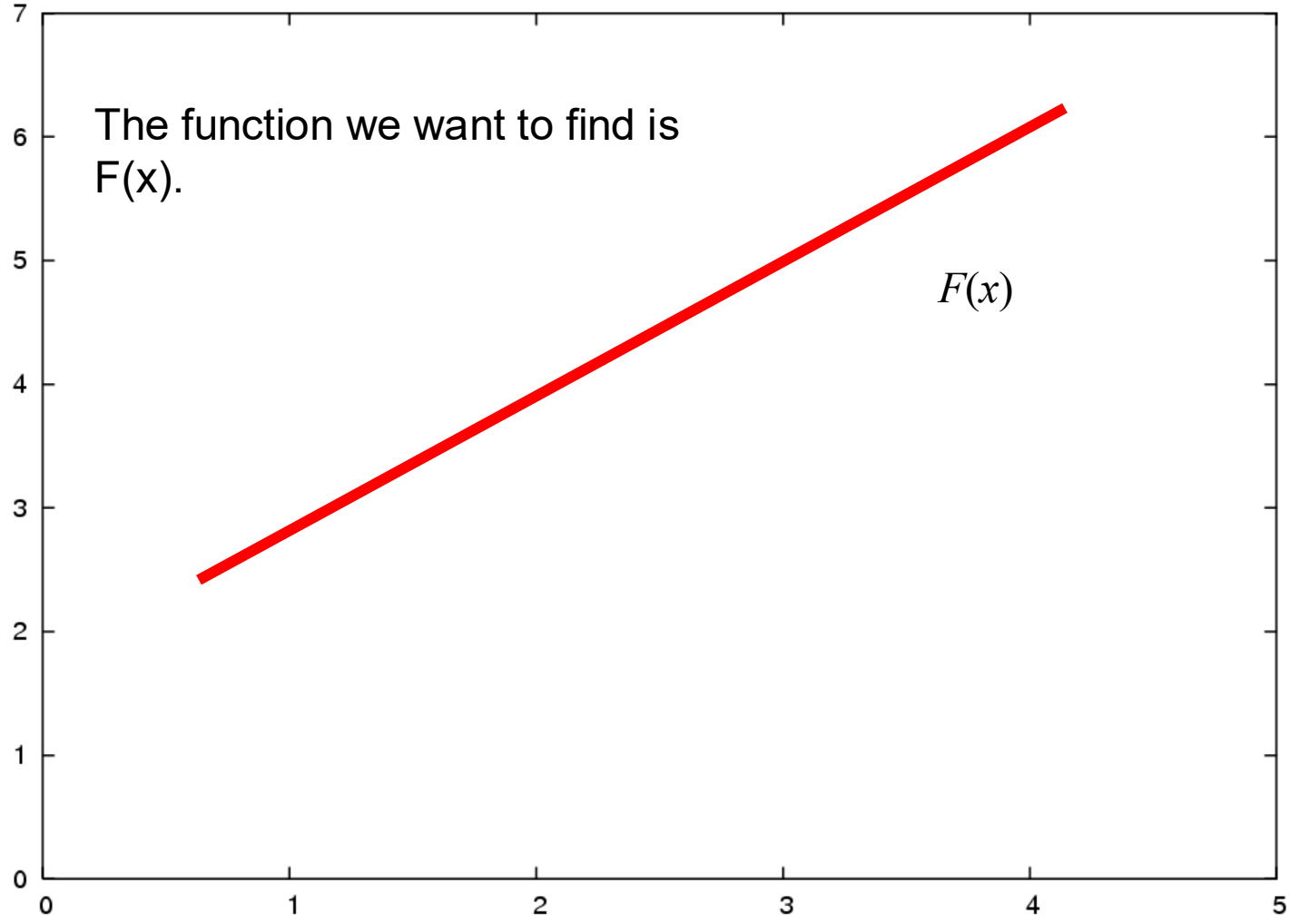
- Understand the concepts of bias and variance in machine learning models
- Identify overfitting (high variance) and underfitting (high bias) scenarios
- Apply performance metrics for regression models (MSE, RMSE, MAE, R^2)
- Apply performance metrics for classification models (Precision, Recall, Accuracy, F1-Score)
- Interpret confusion matrices (True Positives, False Positives, True Negatives, False Negatives)
- Understand and interpret ROC/AUC curves for model evaluation
- Differentiate between Type 1 errors (false positives) and Type 2 errors (false negatives)
- Understand deployment considerations for machine learning models in production
- Recognize ethical considerations and risks in AI systems (fairness, bias, privacy)
- Understand the EU AI Act and its risk-based approach to AI regulation
- Recognize the importance of Trustworthy AI principles (explainability, fairness, robustness, privacy, accountability)
- Understand the concept of Explainable AI (XAI) and why "how" matters as much as "what"
- Identify the timeline and requirements of AI governance and compliance

How well the trained model fits reality depends on:

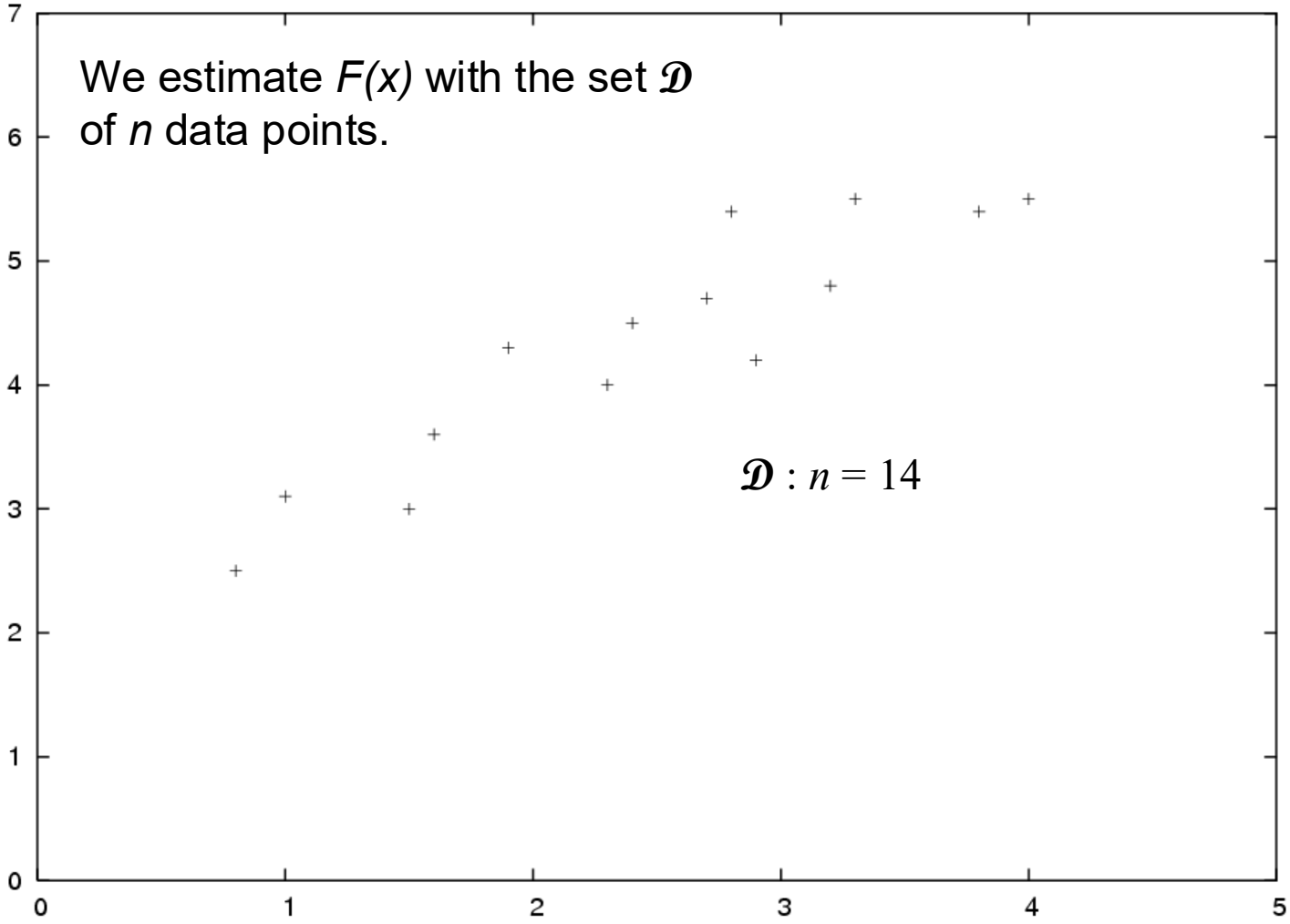
Model | Training data

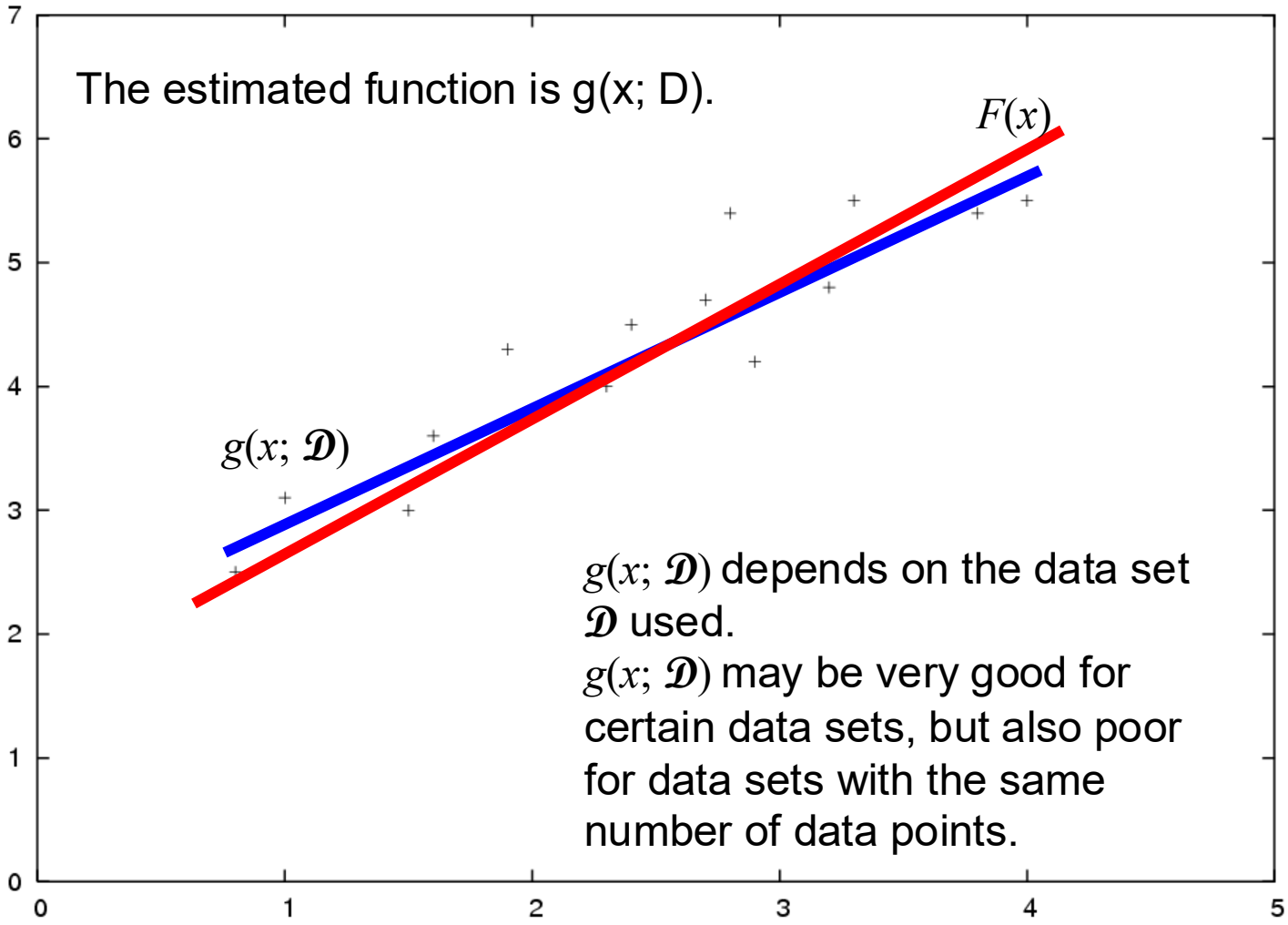
Interdependent

Also known as bias and variance

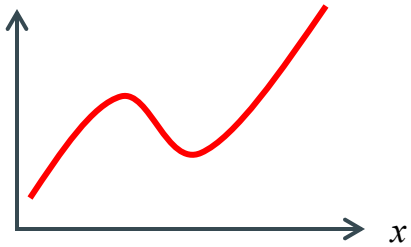

The function we want to find is F(x).
$F(x)$
7
6
5
4
3
2
1
0
0
1
2
3
4
5

We estimate $F(x)$ with the set $\mathcal{D}$ of n data points.

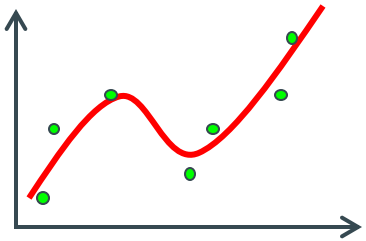



The estimated function is g(x; D).
$F(x)$
$g(x; \mathcal{D})$
$g(x; \mathcal{D})$ depends on the data set $\mathcal{D}$ used.
$g(x; \mathcal{D})$ may be very good for certain data sets, but also poor for data sets with the same number of data points.
0
1
2
3
4
5
6
7

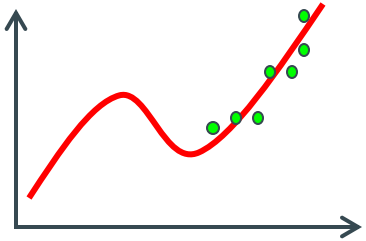
- *F(x)*



- With this data set, $g(x; \mathcal{D})$ would be close

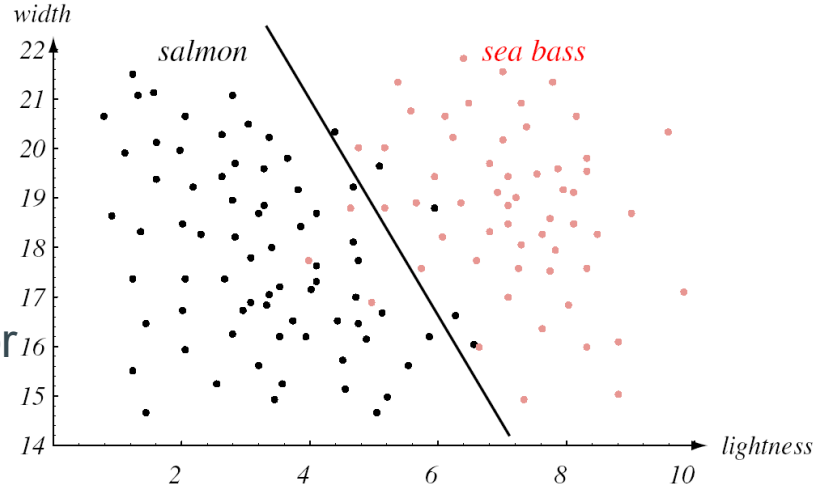


- Not with this data set



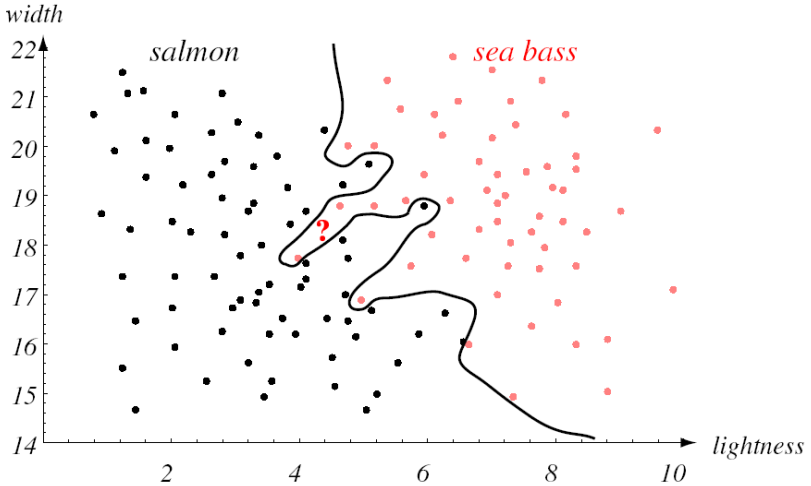
High bias (underfitting)

- The model is too simple or inflexible



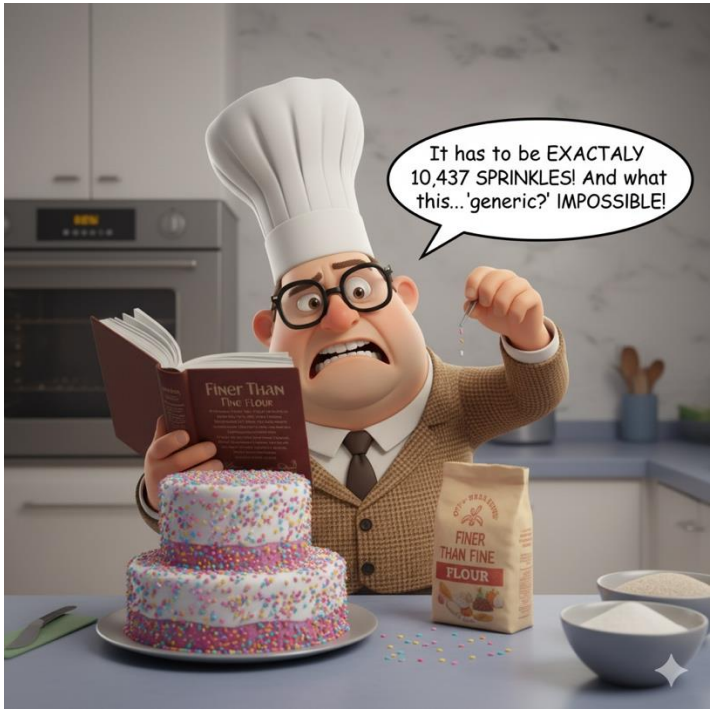
High variance (overfitting)

- The model is too flexible or too closely adapted to the training data set
- "Memorization" of the training data

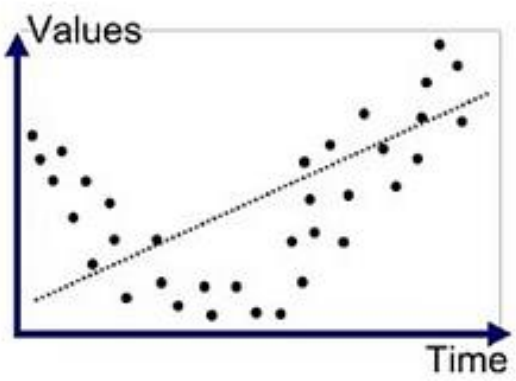


(Aus Duda, Hart, Stork: Pattern Classification, 2nd ed.)

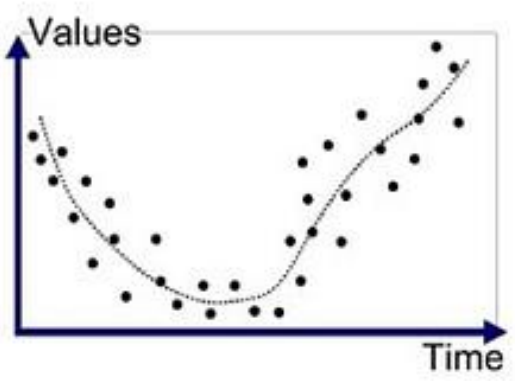
Over- and underfitting



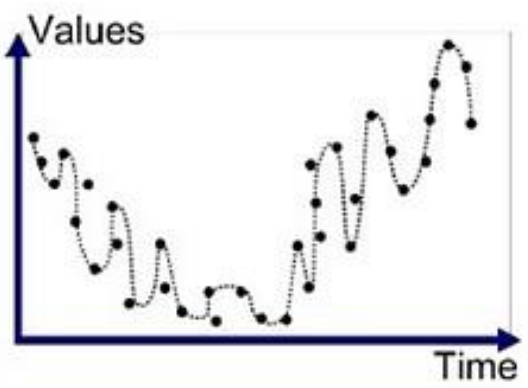
Over- and underfitting



Underfitted



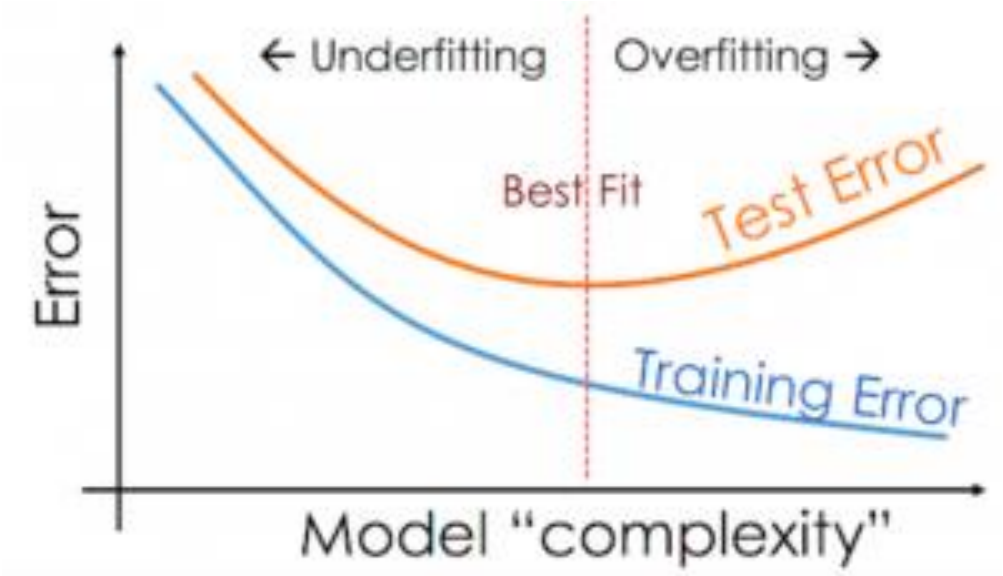
Good Fit/Robust



Overfitted

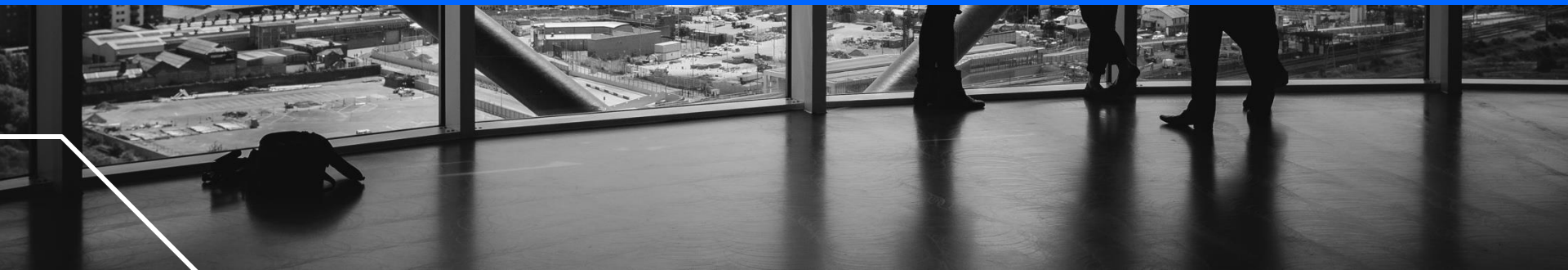
https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Over- and underfitting

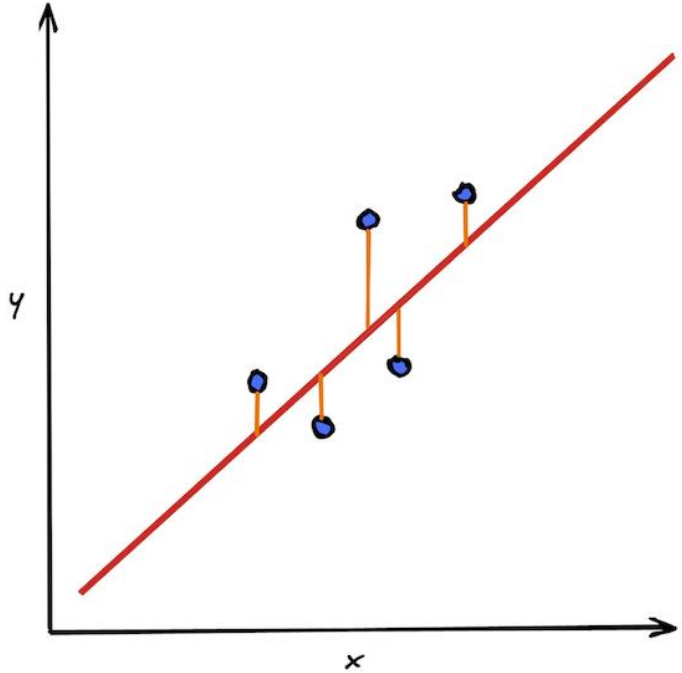


https://vitalflux.com/overfitting-underfitting-concepts-interview-questions/

Model evaluation

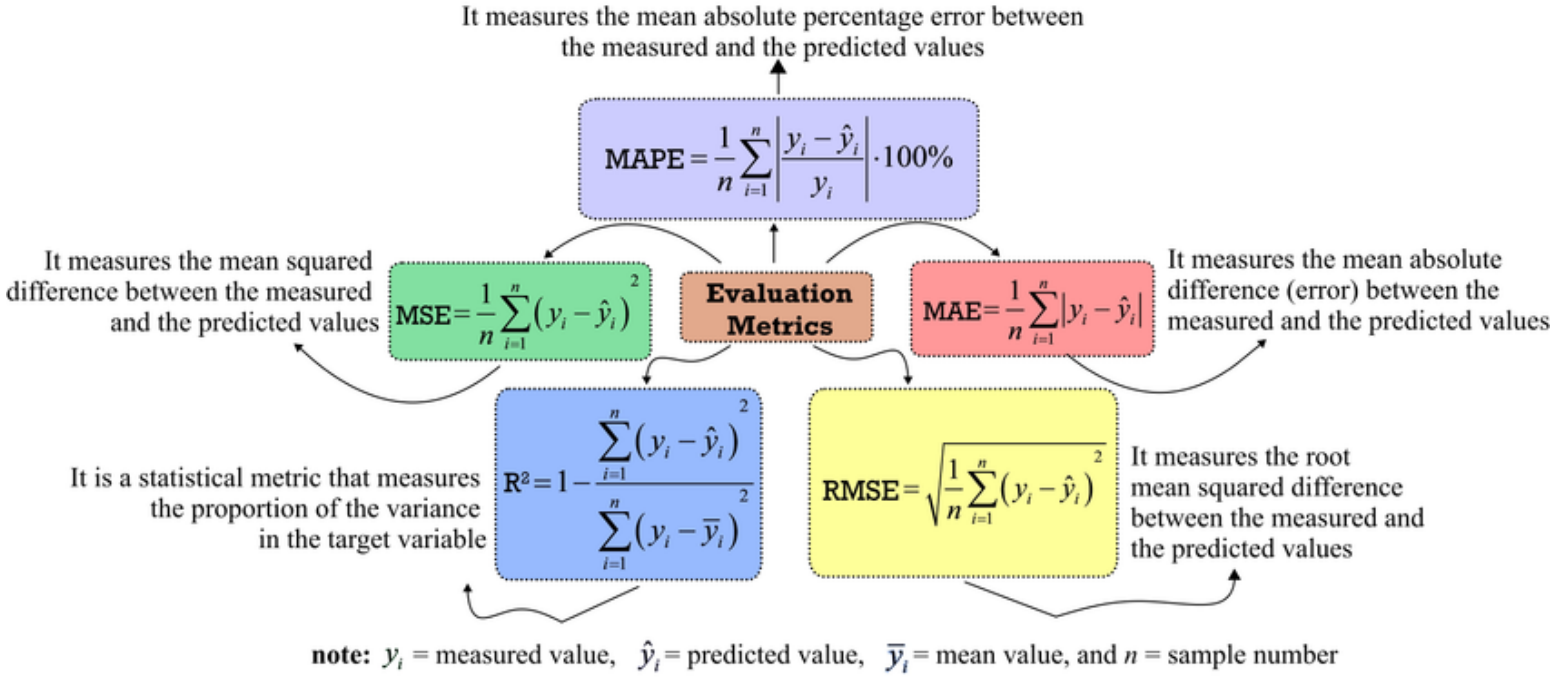


Measuring Performance for Regressions



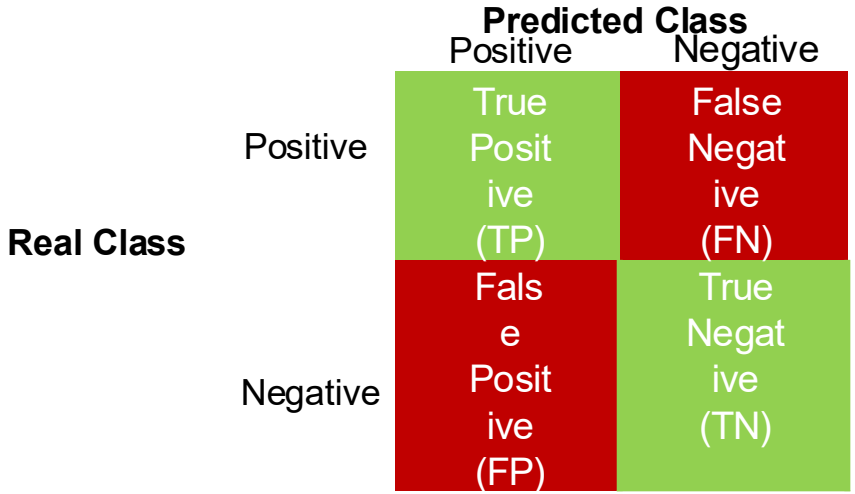
https://towardsdatascience.com/a-comprehensive-overview-of-regression-evaluation-metrics-6264af0926db/

Measuring Performance for Regressions



https://www.researchgate.net/figure/Performance-evaluation-metrics-for-regression-models_fig5_381401914

Measuring Performance for Classifications



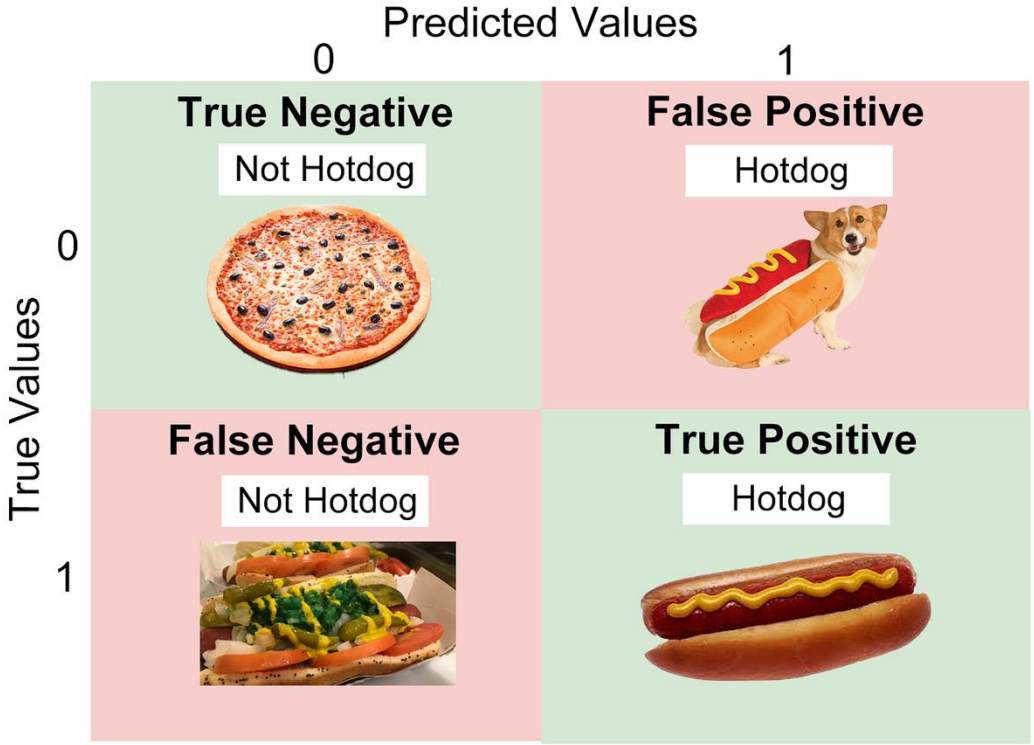
$$Precision = \frac{\Sigma TP}{\Sigma TP + FP}$$

$$Recall = \frac{\Sigma TP}{\Sigma TP + FN}$$

$$Accuracy = \frac{\Sigma TP + TN}{\Sigma TP + FP + FN + TN}$$

- **Precision:** From the predictions of the system, how many the system predicted correctly.
- **Recall:** From the real classes in the dataset, how many the system predicted correctly.

Measuring Performance for Classifications

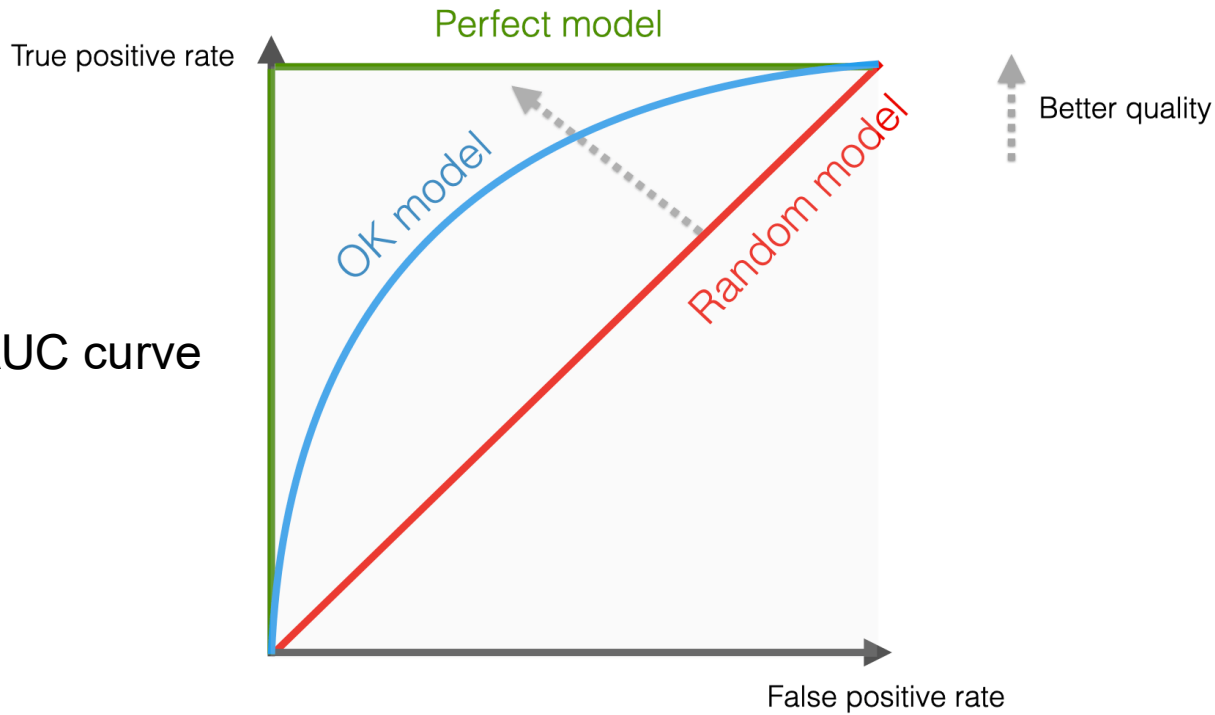


Measuring Performance for Classifications

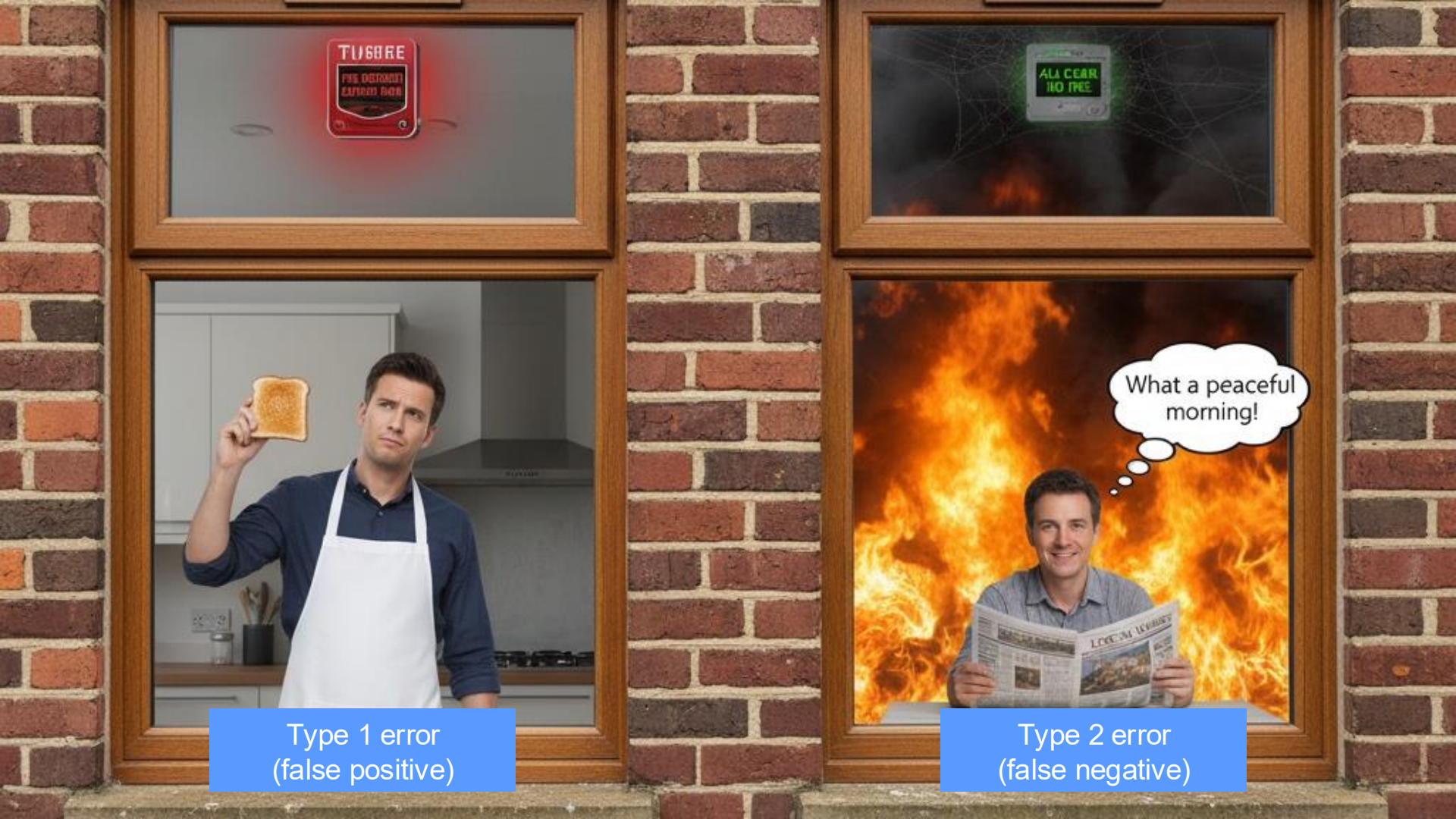
Exercise: *We have a dataset with 500 bank transactions, from which we know that 90 are fraudulent. The system predicts that 50 are fraudulent. From those 50 predictions, the system got 40 correct predictions. What is the precision and recall of the system?*

Measuring Performance for Classifications

ROC/AUC curve



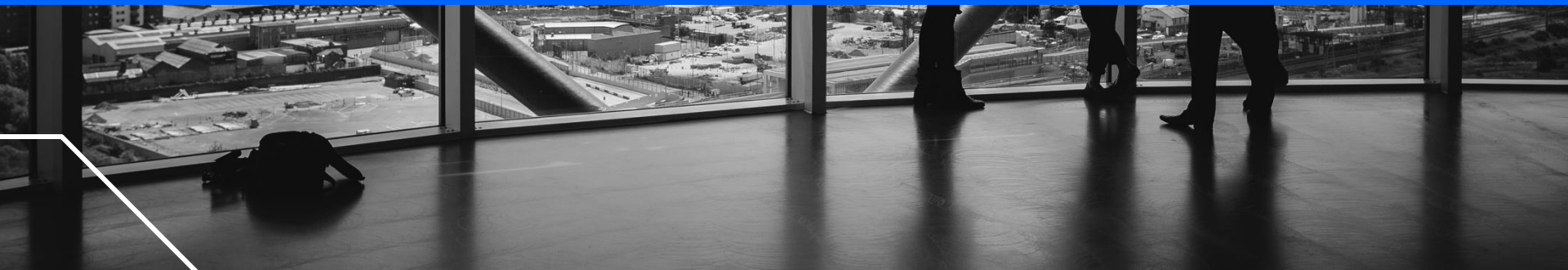
https://www.evidentlyai.com/classification-metrics/explain-roc-curve



Type 1 error
(false positive)

Type 2 error
(false negative)

Deployment considerations

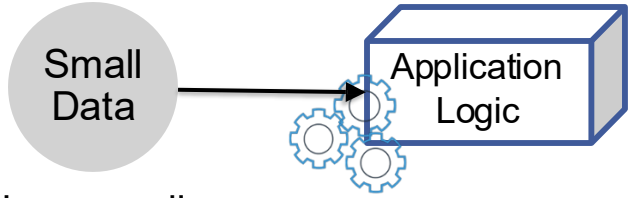


Deployment consideration

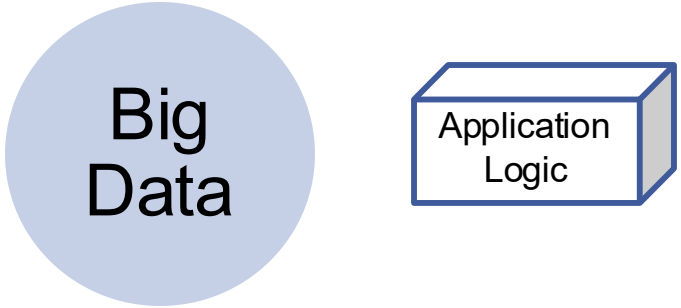
Deployment consideration

Now the hard work begins

Old paradigm: Bring data to computation



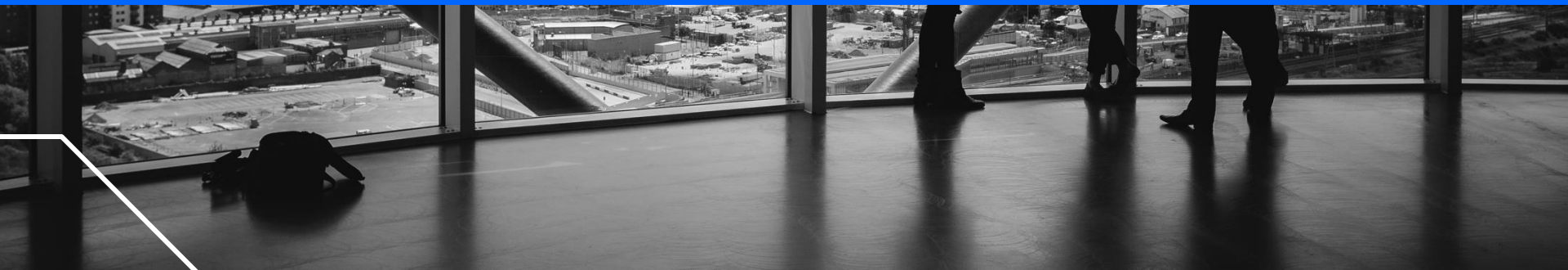
New paradigm: Bring computation to data



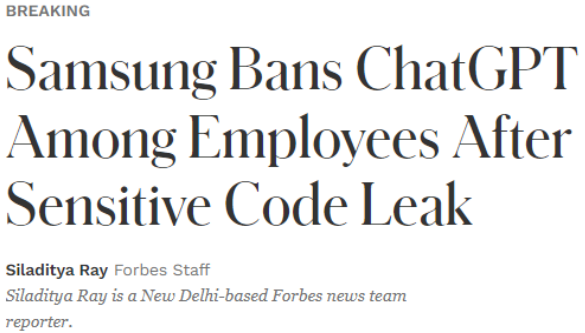
- Data Management & Database Areas
- Scalable data infrastructures;
- Coping with data diversity;
- End-to-end processing and understanding of data;
- Cloud services; and
- Managing the diverse roles of people in the data life cycle.

Source: "The Beckman Database Research Self-Assessment" Meeting Report. 2013.

Ehtical considerations



Why is there a lack of trust?



BREAKING

Samsung Bans ChatGPT Among Employees After Sensitive Code Leak

Siladitya Ray Forbes Staff

Siladitya Ray is a New Delhi-based Forbes news team reporter.

LLMs collect and share confidential information*

https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/

Tesla Autopilot feature was involved in 13 fatal crashes, US regulator says

Federal transportation agency finds Tesla's claims about feature don't match their findings and opens second investigation

Autonomous cars cause accidents

https://www.theguardian.com/technology/2024/apr/26/tesla-autopilot-fatal-crash

A.I. has a discrimination problem. In banking, the consequences can be severe

PUBLISHED FRI, JUN 23 2023·1:45 AM EDT

Unfair AI credit decisions lead to legal and financial damage

https://www.cnbc.com/2023/06/23/ai-has-a-discrimination-problem-in-banking-that-can-be-devastating.html

... the risks must be addressed

LLMs collect and share confidential information*

https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/

Autonomous cars cause accidents

https://www.theguardian.com/technology/2024/apr/26/tesla-autopilot-fatal-crash

Unfair AI credit decisions lead to legal and financial damage

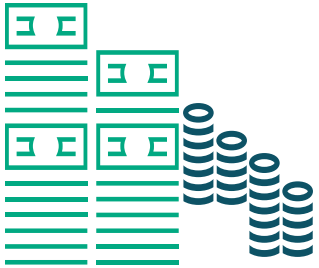
https://www.cnbc.com/2023/06/23/ai-has-a-discrimination-problem-in-banking-that-can-be-devastating.html

Regulation

The EU AI Act at a glance

Risk-based approach
Unacceptable, high, limited, minimal

Requirements
Data management, accuracy, cybersecurity, transparency, monitoring, robustness



Penalty
up to €35 million / 7% of global annual turnover

Regulation: The EU AI Act

AI requirements are categorized according to risk, and ignoring requirements is very costly.

Unacceptable risk applications
- Pose a clear threat to people's safety and livelihoods
- Expressly prohibited by AI law
- Example: social scoring, real-time biometric law enforcement

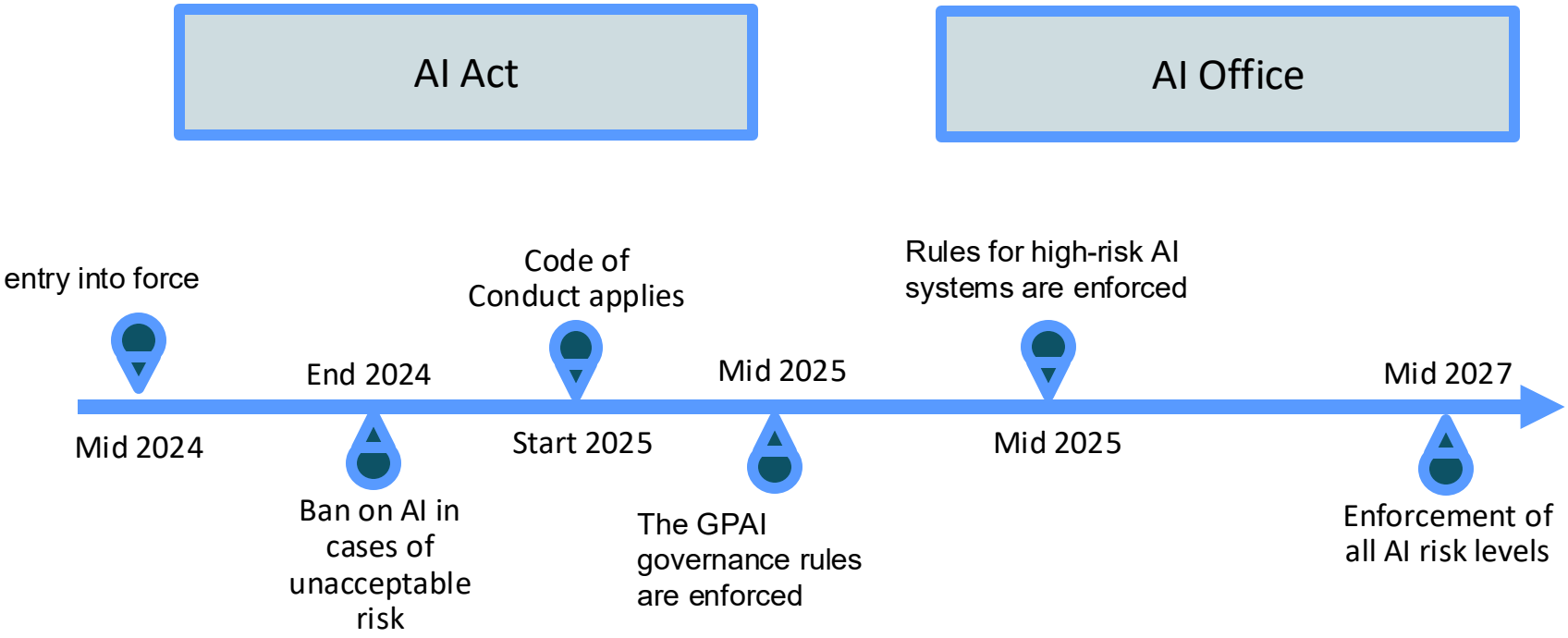
High-risk applications
- Have the potential to cause physical or financial harm to people
- Regulation: must be of good software quality, transparent, and fair
- Example: credit checks, personal employment

Limited-risk applications
- Low potential for harm
- Regulations: Transparency and cooperation
- Example: Chatbot, deepfakes

Minimal risk applications
- (Almost) no risk to humans
- Regulations: Voluntary code of conduct
- Example: Targeted marketing, spam filters

Penalty of up to €35 million / 7% of global annual turnover

EU AI Act Timeline



1 https://www.alexanderthamm.com/en/blog/eu-ai-act-timeline/
2 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

“What?”

https://www.autoscout24.be/fr/voiture/voiture-sportive/

Image z

“What?”

https://www.autoscout24.be/fr/voiture/voiture-sportive/

Image z

“Car”

Label z

"What?" is not sufficient

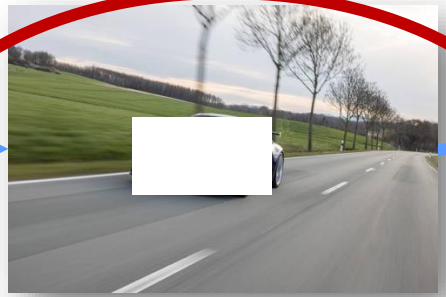
Causal, robust indicator

Car body

https://www.autoscout24.be/fr/voiture/voiture-sportive/

Image z

"Car"



Street

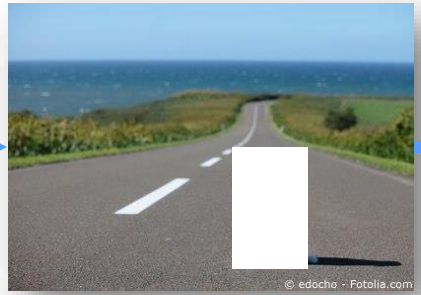
Label z

Non-causal, false clue

“What?” is not sufficient

https://www.hunde-info.de/hundehaftpflicht-versicherung-2115.html

Image x



Street

“Auto”

Label x

- It is not enough to know “whether” you can solve the problem.
- It also depends on “how” you solve it.
- Robustness and explainability of AI systems.

It needs a 'how'

Nothing prevents the model from using only non-causal, false clues for recognition.

https://www.hunde-info.de/hundehaftpflicht-versicherung-2115.html

Image x

Street

“Car”

Label x

- It is not enough to know “whether” you can solve the problem.
- It also depends on “how” you solve it.
- Robustness and explainability of AI systems.

It needs a 'how'

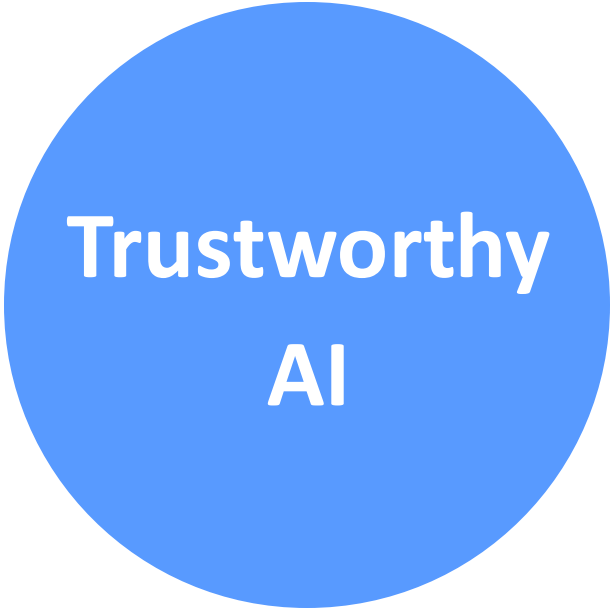
Nothing prevents the model from using only non-causal, false clues for recognition.

“ML 1.0”: Learn the prediction p(x,y) with the data (x,y)
“ML 2.0”: Learn the prediction p(x,z,y) with the data (x,y)

AI quality is more than just performance

AI quality is more than just performance

- Comprehensible & Explainable
- Fair & Inclusive
- Maintaining privacy
- Robust & performant
- Responsible
- Secure



VIELEN Dank und VIEL ERFOLG

samuel.schlenker@hpe.com
joel-weiss@outlook.de
j.weiss@vfb-stuttgart.de