School of Engineering

Data Analytics and Artificial Intelligence Summer School 2025

Dr. Manuel Dömer
Programme Director BSc. Data Science

doem@zhaw.ch
+41 (0) 58 934 43 30

30.06.2025

Manuel Dömer

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

School of Engineering

- Programme Director BSc. Data Science
- Co-Head ZHAW Datalab
- EELISA ZHAW Academic Coordinator

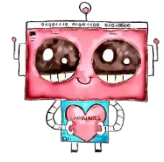
Swiss Re

- Lead Modelling & Analytics in Digital Markets
- Data Scientist
- Credit Portfolio Modelling
- Natural Language Processing
- Junior Corporate Actuary

- Lecturer in Data Science

Lucerne University of Applied Sciences and Arts
HOCHSCHULE LUZERN

- Co-Organiser NLP Zürich-Meetup



EPFL

Universität Zürich[UZH]

- PhD in Computational Chemistry
- BSc and MSc in Chemistry

Research Interests

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

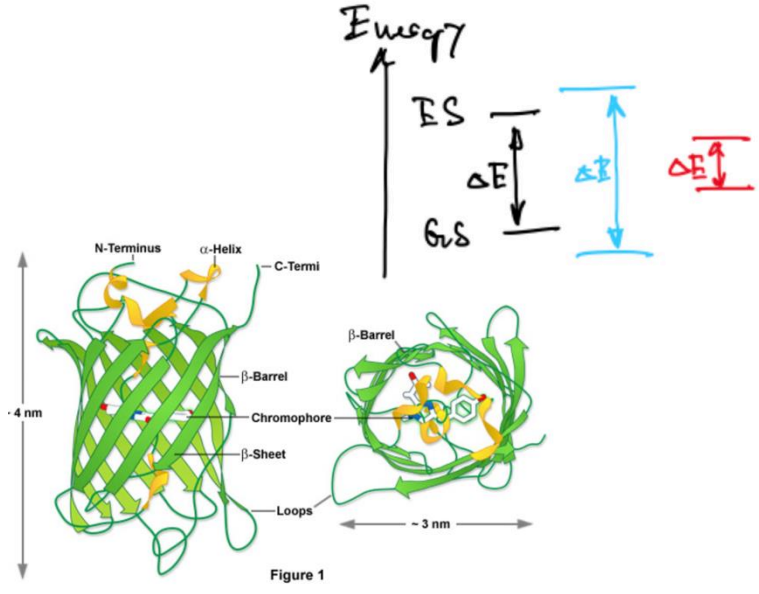


Figure 1

Schedule

Zürcher Hochschule für Angewandte Wissenschaften

zh aw School of Engineering

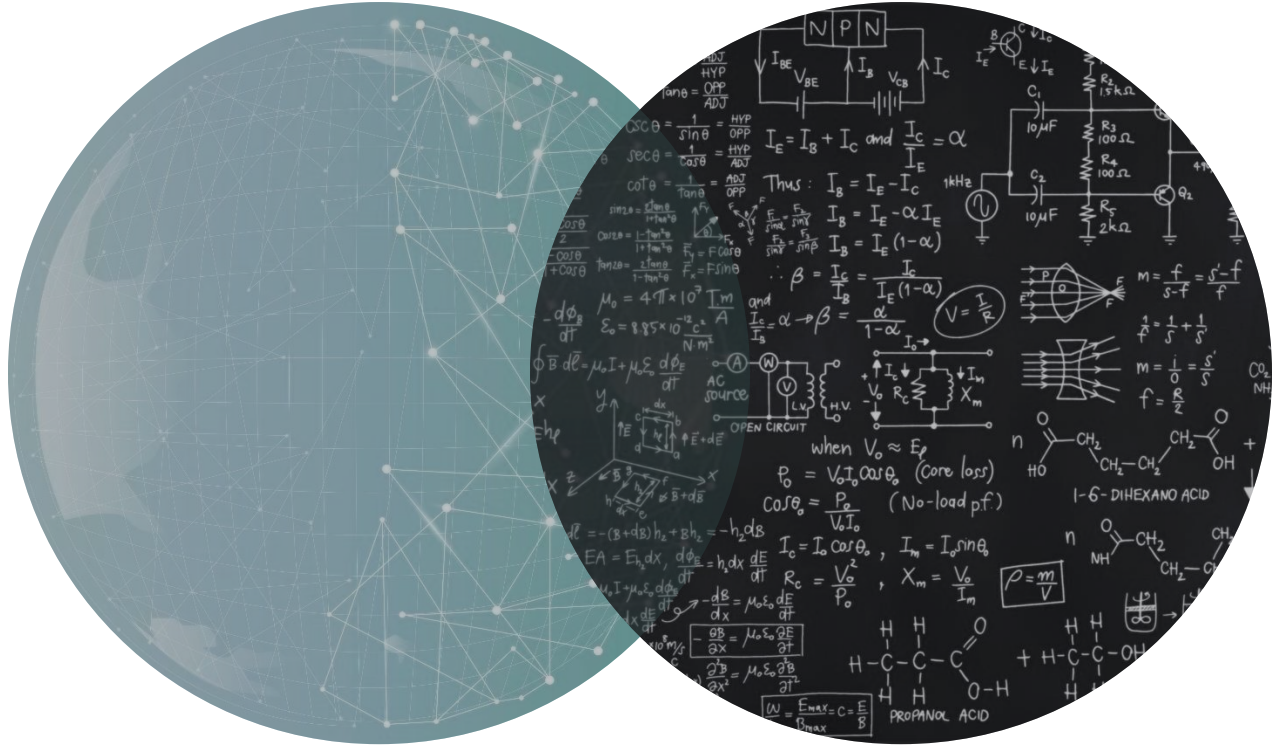
	Mo, June 30	Tue, July 1	Wed, July 2	Thur, July 3
9-12 am	• Introduction • Data Science-Process	• Data Engineering Part 1	• DE Part 2	• DE Part 3
Lunch				
1-5 pm	• Projects	• Projects	• Projects	• Project Plan Presentations

	Mo, July 7	Tue, July 8	Wed, July 9	Thur, July 10
9-12 am	• Supervised Machine Learning	• Association Rules • Recommender Systems	• Projects	• Projects
Lunch				
1-5 pm	• Unsupervised ML	• Projects	• Projects	• Project Status Updates

Data Science

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering



Zurich UAS Racing – Formula Student

Zürcher Hochschule
für Angewandte Wissenschaften

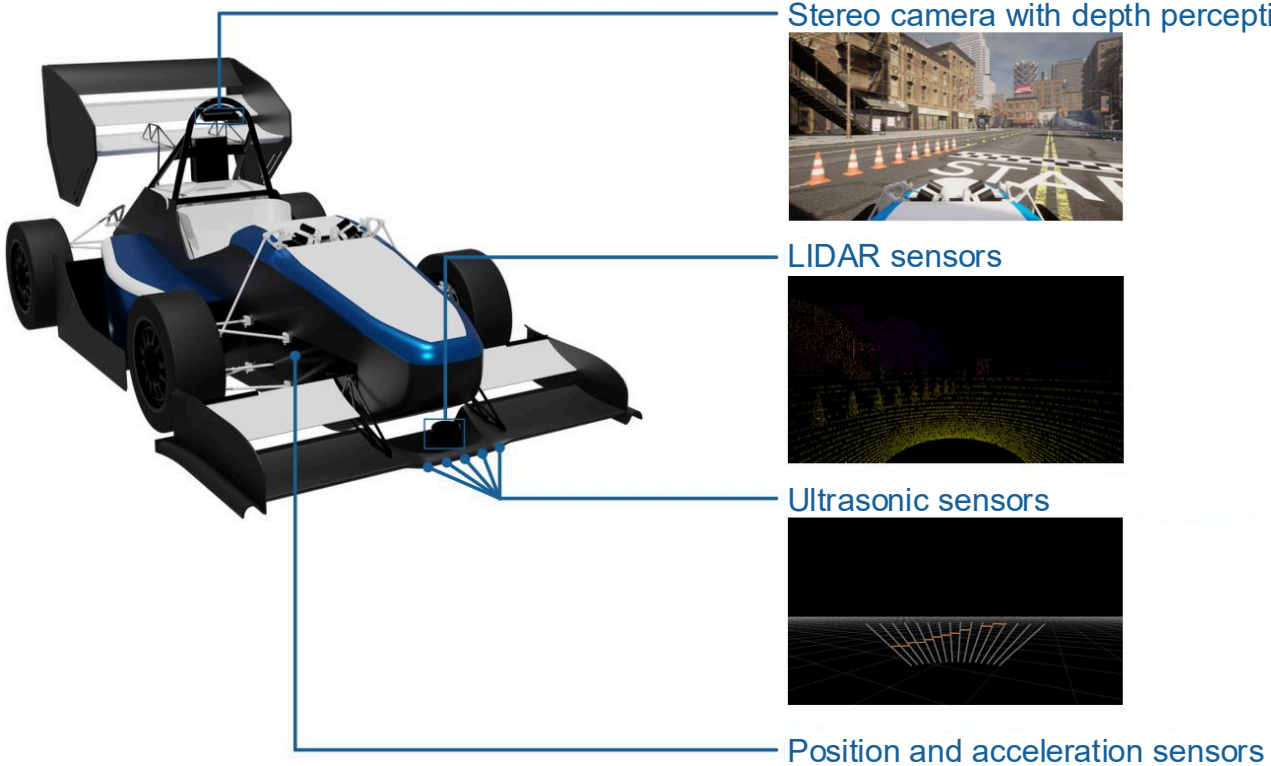
zh aw School of Engineering

👉
https://zurichuasracing.ch

Sensor Data of an Autonomous Vehicle

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering



Data in the context of «Data Science»

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

Data implies digital data in binary form,

which is stored and processed by electronic computer systems

in form of files of given formats that are interpretable by specific software applications, e.g. as text, numbers, tables, images sound or executable programs…

Structured vs. unstructured Data

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

Structured data
- pre-defined data model
- different datasets can follow the same data model

Examples
- tabular data

File formats
- csv
- ods
- xlsx
- HDF (Hierarchical Data Format)
- Apache Parquet

Unstructured data
- not a fixed data model
- no explicit structure, but might feature implicit structure, such as grammatical rules in text

Examples and typical file formats

Content	Formats
Text	txt, odt, docx, html
Images	jpeg, png
Audio	wave, aiff, flac, mp3, aac
Video	mp4, mov, wmv

Structured vs. unstructured Data

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering

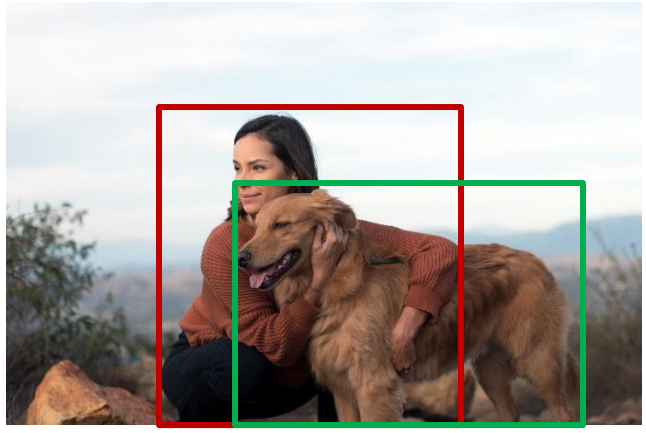
... Roche disappoints analysts and drags the SMI into the red. The daily loss is 4 percentage points higher than that of its competitor Novartis, which closes only just in the red at 0.1%. Financial stocks were able to protect the index from even greater losses, led by Credit Suisse, which ...

Ticker	Last	Change	Volume
CSGN	14.19	+2.5 %	8’142’755
NOVN	83.06	-0.1 %	4’875’390
PGHN	1433	+1.5 %	98’060’000
ROG	224.15	-4.1 %	2’7680’115
SREN	92.22	+0.7 %	2’298’221

Information extraction

Structured vs. unstructured Data

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering



Object detection

Women	1	0
Men	0	0
Dogs	1	1
Cats	0	1
Environment	Outdoor	Indoor

Semi-structured Data

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering

Does not have a fixed data model, but can contain structural information as **tags** or **markers**, so that the schema can be (partially) (re-)constructed.

Entities of the same class might contain **different attributes** and the **order** is arbitrary.

Popular file formats: **XML** and **JSON** – both machine and human readable

Applications/Technologies:

- Communication in **Web-Applikationen** - JSON (REST) nowadays more common than XML
- **E-Mail**
- Document-based (**No SQL**) databases such as MongoDB and Elasticsearch (fulltext search)

+ Objects can be stored **without pre-defined schema**
+ **Mapping** to a relational data model unnecessary
+ Natural representation of nested objects or hierarchical dependencies

− No performant query language such as **SQL**
− **Error-prone** due to missing data model

Metadata

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

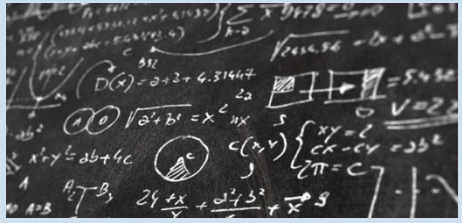
«**Data about data**»: Contains additional information about specific datasets and supports their interpretation - e.g. description of the data model and the origin of the data

File Source	https://commons.wikimedia.org/wiki/File:Cat03.jpg
License	CC BY-NC
Attribution	Fir0002/Flagstaffotos
File Type	JPEG
File Size [kB]	152
X Dimension	1200
Y Dimension	1199
Resolution [pixels/inch]	72
Color Encoding	sRGB

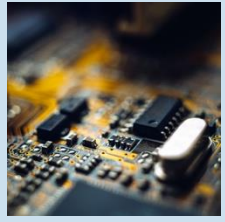
Data Science is data product development

School of Engineering

Data

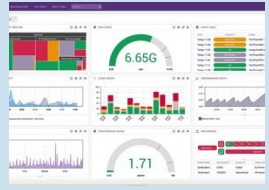


Applied mathematics



Computer science

Domain expertise

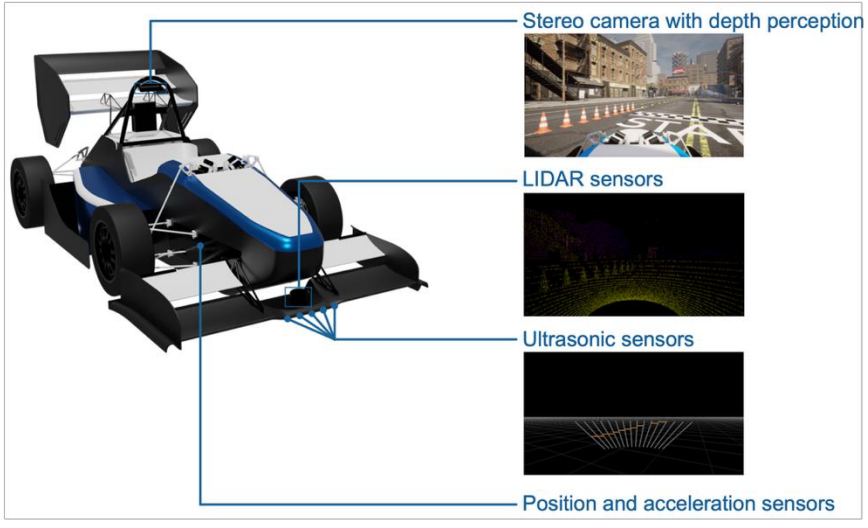


Data product

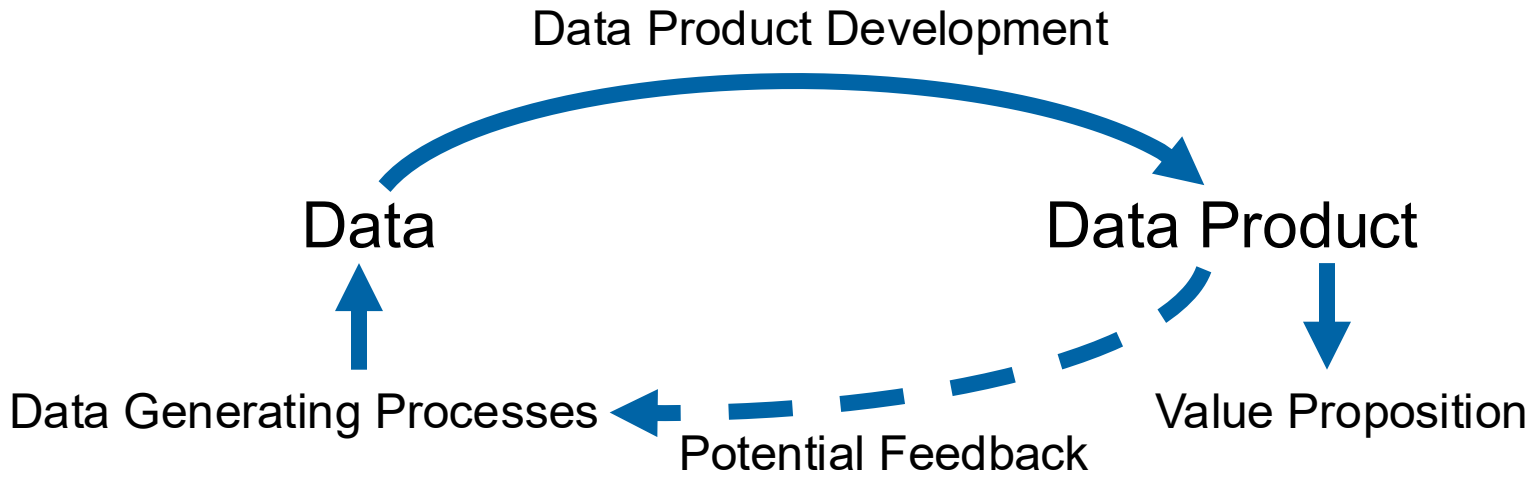
Use Case: Autonomous Vehicle

School of Engineering

Cars can use cameras and other sensors to recognise their surroundings and navigate independently.

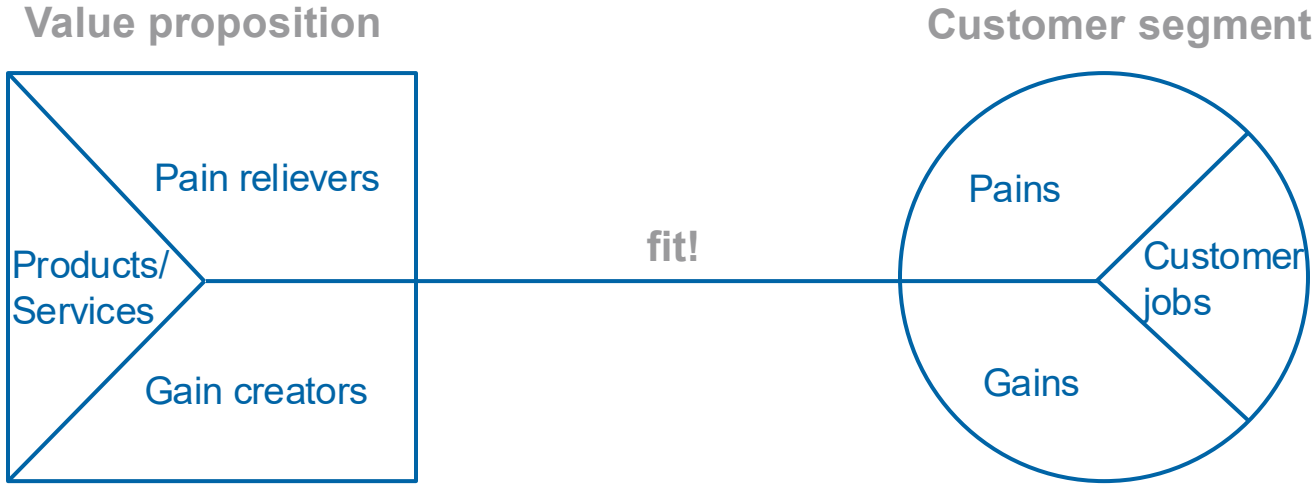


Data product



Value Proposition Canvas

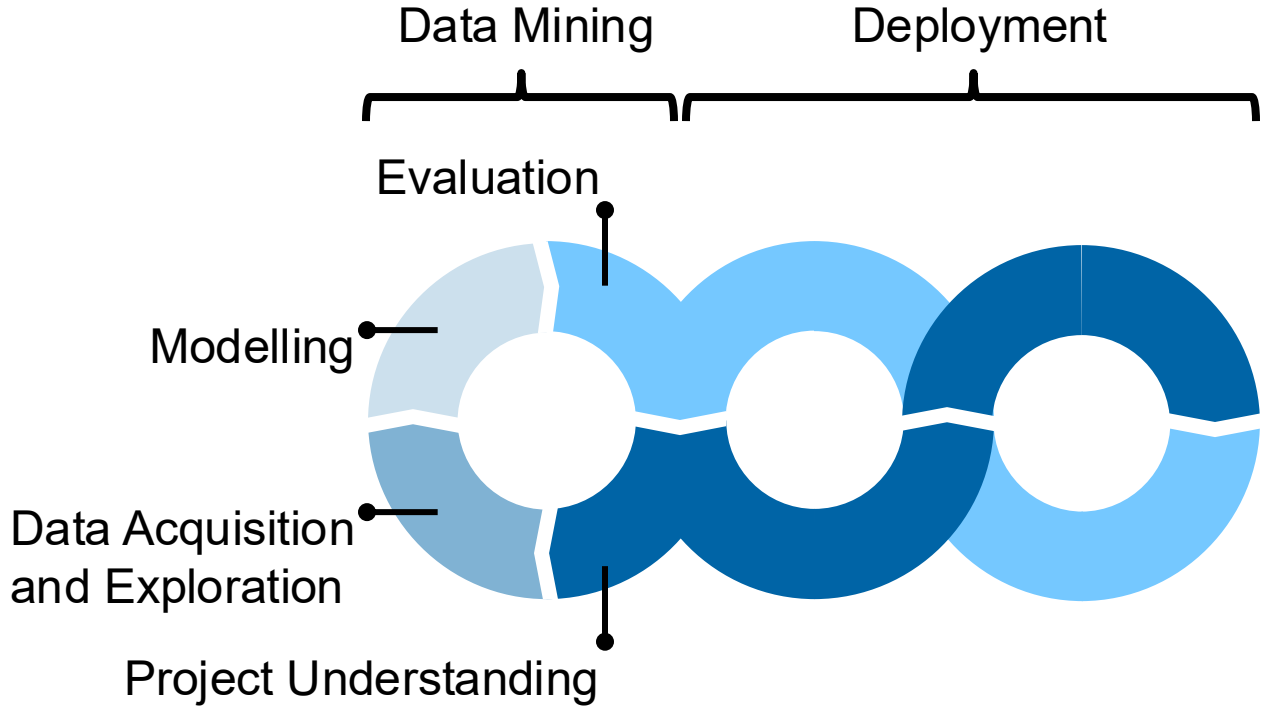
Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering



Osterwalder, A., Pigneur, Y., Bernarda, G., Smith, A., & Wegberg, T. A. A. E. (2015). *Value Proposition Design: Entwickeln Sie Produkte und Services, die Ihre Kunden wirklich wollen.* John Wiley & Sons, Incorporated.

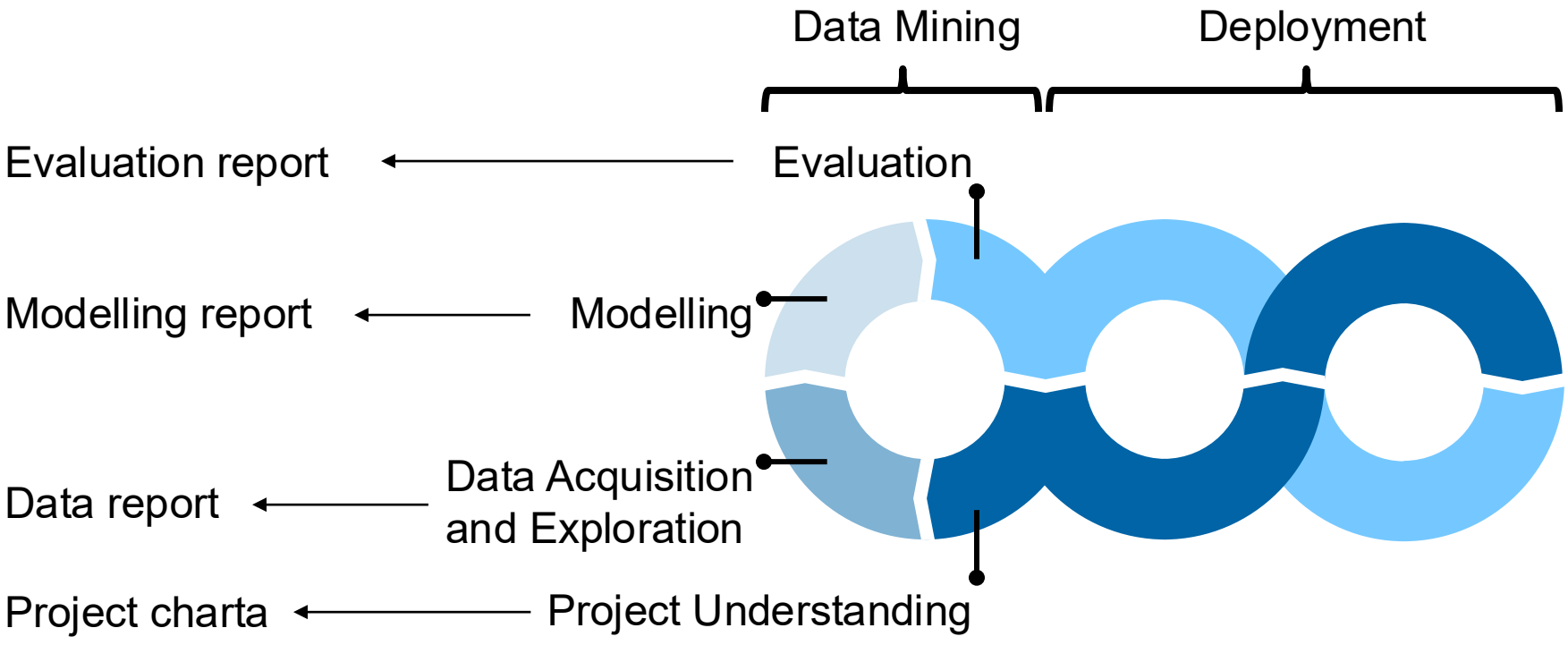
Data Product-Development

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering



Documentation Artefacts

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering



Tooling

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

Local Development Environment

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

Suggestion:

https://code.visualstudio.com

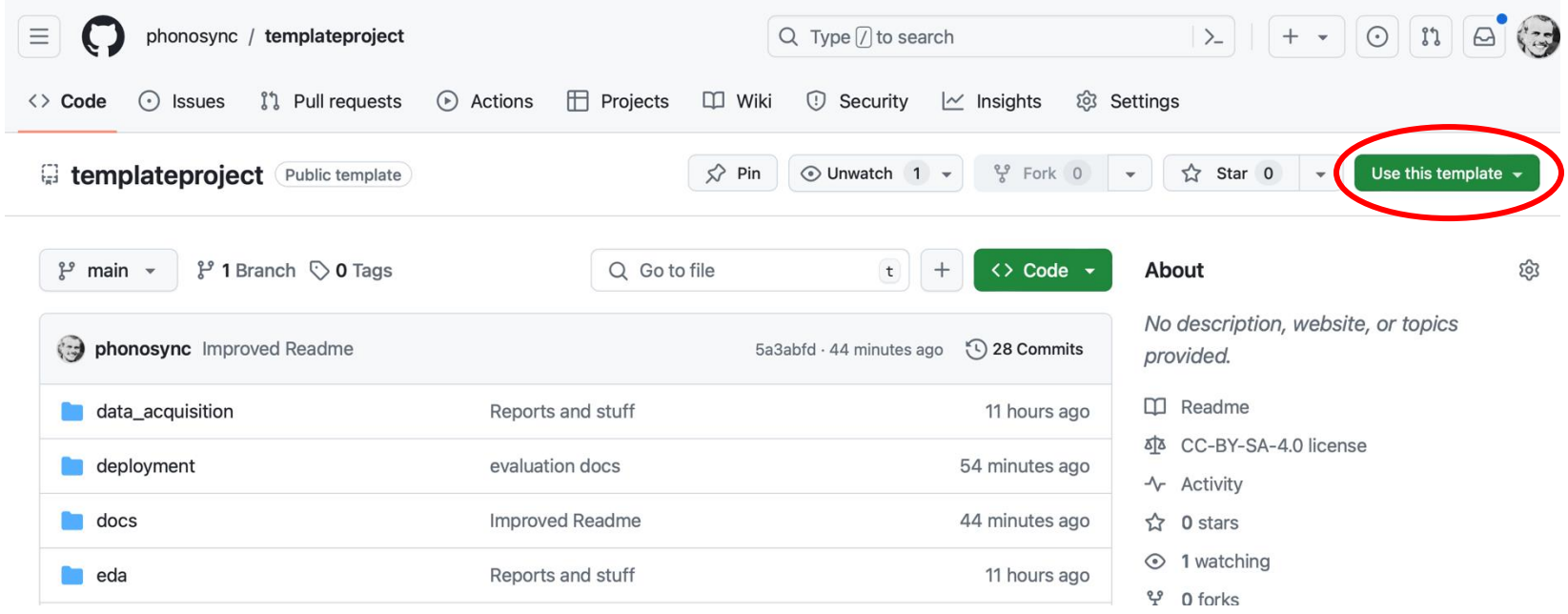
Extensions:

- Python (Microsoft)
- Jupyter (Microsoft)
- GitHub Copilot (Microsoft)
- Live Preview (Microsoft)
- Quarto

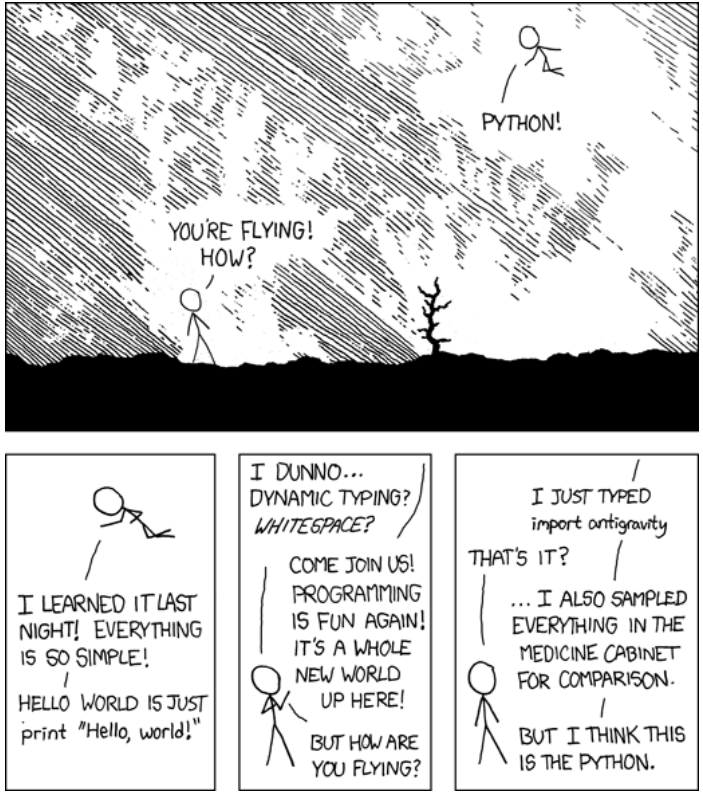
Structure and Version Control in a Data Science Project

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

https://github.com/phonosync/templateproject



Python - The power comes with the packages



Popular registries:

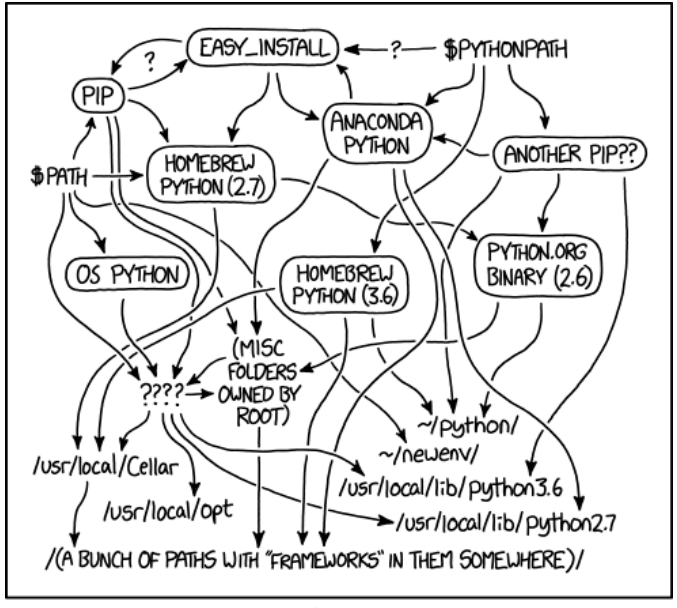
- https://pypi.org/
- https://anaconda.org/anaconda/repo

Package Management

Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

Different projects require different version of packages and dependencies…



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

Python Environments

Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

… let you isolate project specific interpreter and package installations.

- venv
- virtualenv
- Anaconda
- uv

We go with **Anaconda** - it provides:

- a package index: https://anaconda.org/anaconda/repo
- a package manager: > conda install pandas
- an environment manager: > conda create --name myproject

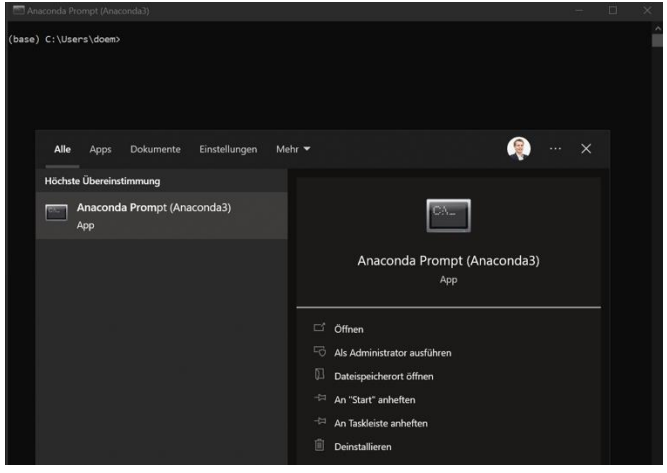
Download and install Anaconda for your system: installation instructions

Using conda

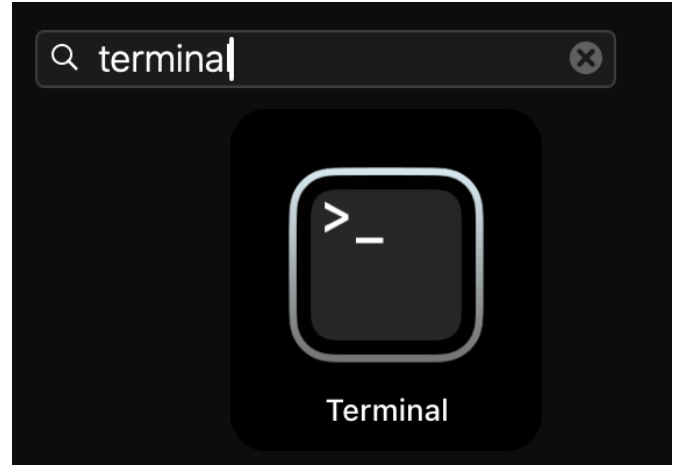
Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

Windows



Mac OS X



2.In the console navigate to your project folder, e.g. cd C:\Users\doem\projects\demo
3.execute conda env create --file conda.yml
4.confirm installation

Using conda

Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

To activate the environment execute in the console:

```
conda activate demo
```

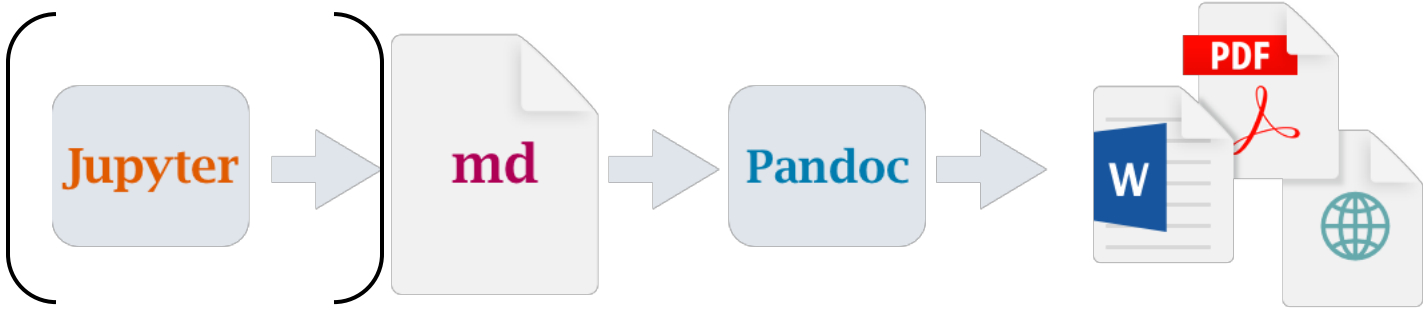
Remember to activate whenever you work on the project!

Mor information on managing the conda environments in the template project Readme

Project Documentation with Quarto

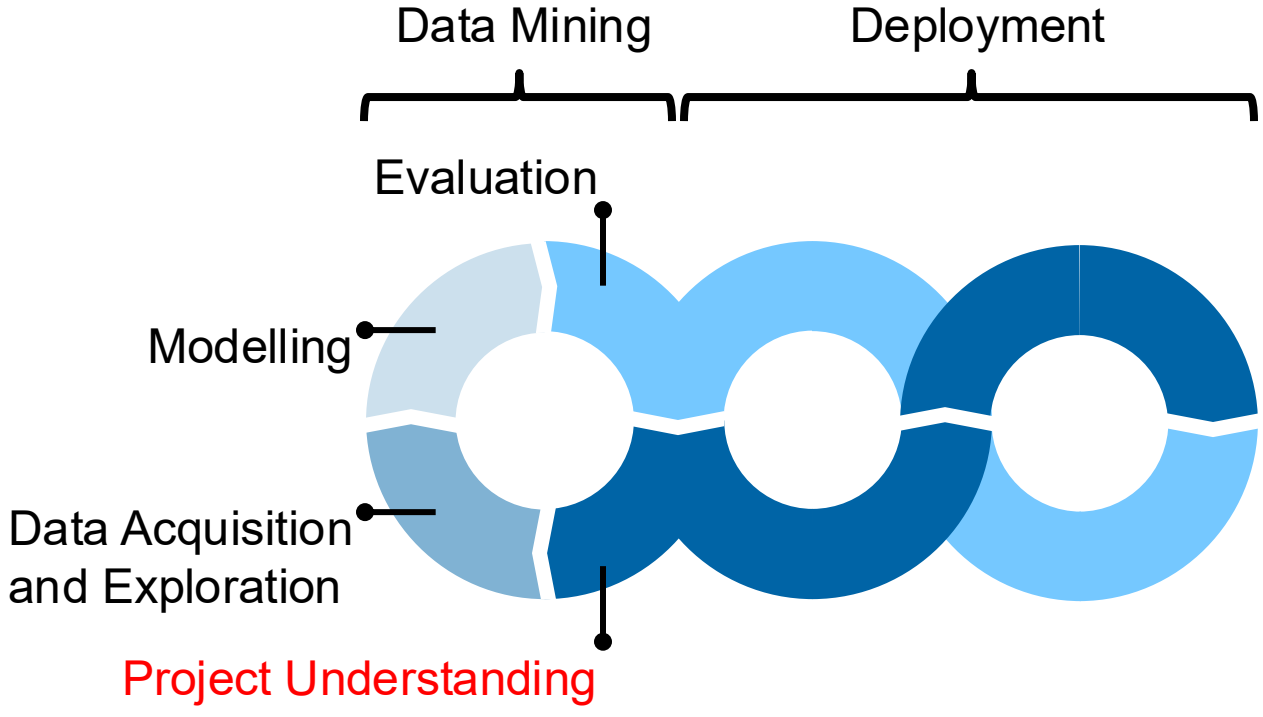
Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

- Download from https://quarto.org and -> https://quarto.org/docs/get-started/



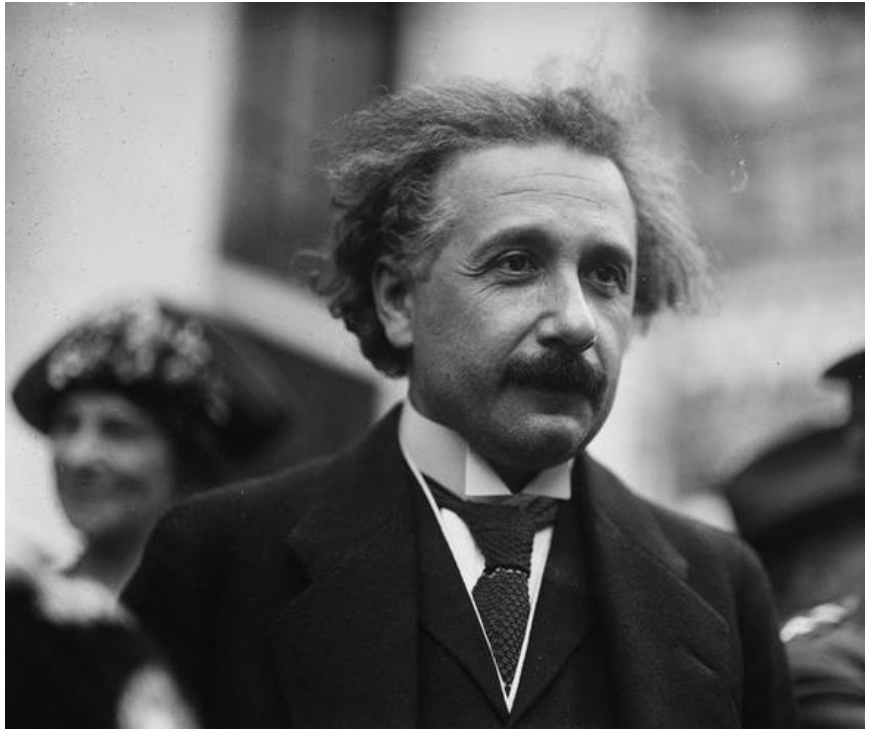
Data Product-Development

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering



Project Understanding with Albert Einstein

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering



„If I had an hour to solve a problem, I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions“

Project Understanding Overview

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

	Activities	Output/Artefact
Project Understanding	• Formulate problem statement • Analyse initial situation • Set project goals and success criteria • Derive data analysis goals • Make project plan	• Project charta

The target values of the key metrics provide transparency on progress and a quantitative basis for decision-making at project milestones.

Map the Business Problem onto a Data Mining Task

Zürcher Hochschule
für Angewandte Wissenschaften

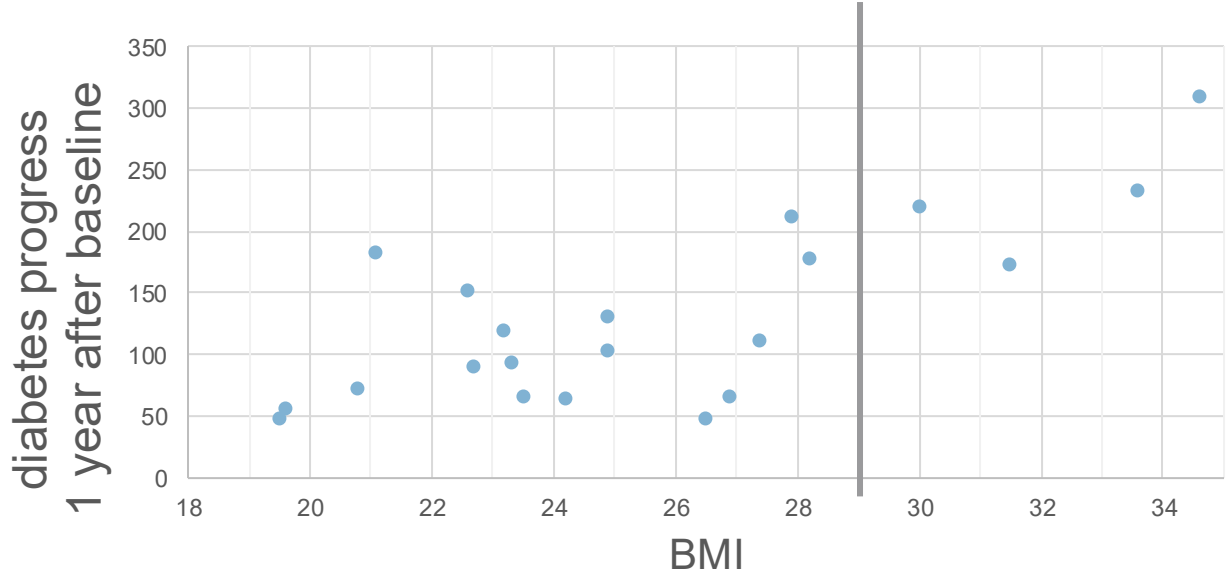
zh aw School of Engineering

- Regression
- Classification
- Clustering
- Outlier Detection
- Association rule learning (market basket analysis)
- Recommender System

Regression
Example: Predict diabetes progress based on BMI

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering



Regression aims at predicting a continuous target variable based on a set of dependent variables.

Classification
Example: Muffin/Chihuahua

Classification groups objects into distinct categories (classes).

Clustering

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering







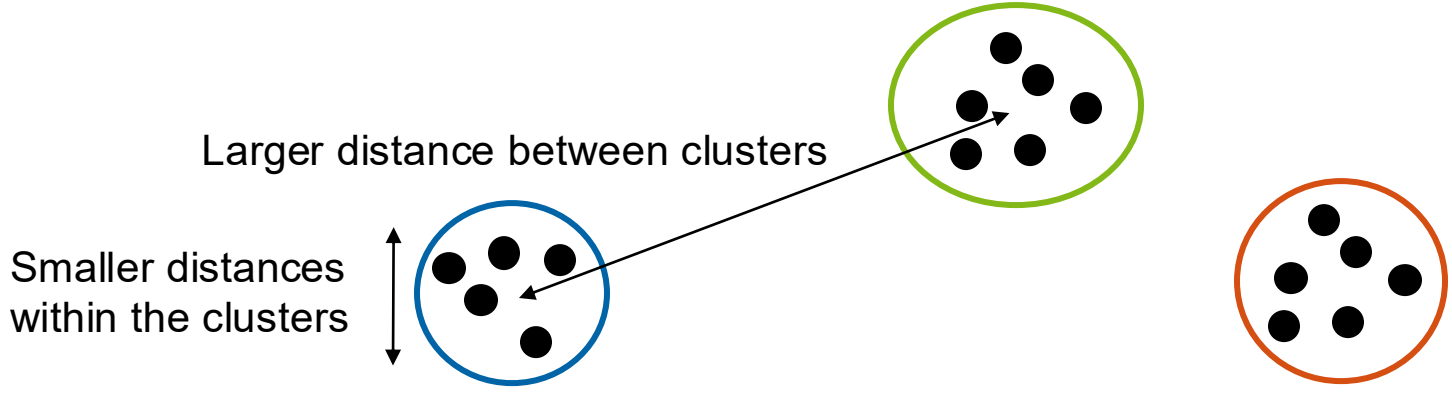



Clustering identifies subgroups of datapoints that are more similar to each other than to the elements in other subgroups.

Clustering

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

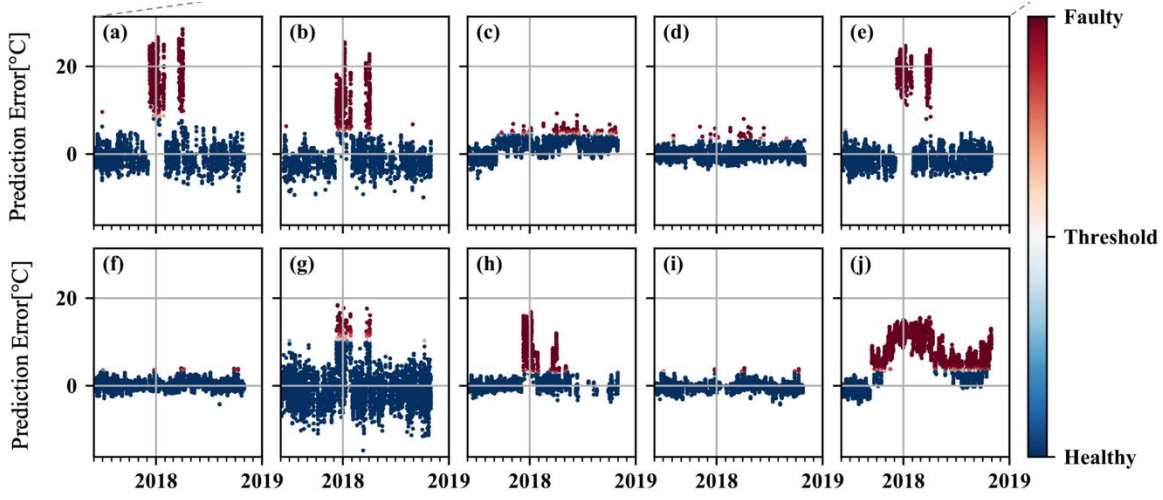
Goal**: Identify subgroups** of datapoints that are more similar to each other than to the elements in other subgroups.



→ Needs metric to quantify similarities.

Outlier Detection
Example: Detection of malfunctions from the sensor data of a wind turbine

Zürcher Hochschule für Angewandte Wissenschaften
zh aw School of Engineering



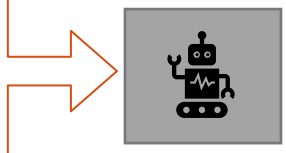
Outlier detection identifies datapoints that do not belong to the same distribution as „normal“ observations.

Recommender Systems
Example: Online shop – personalised recommendations based on shopping behaviour

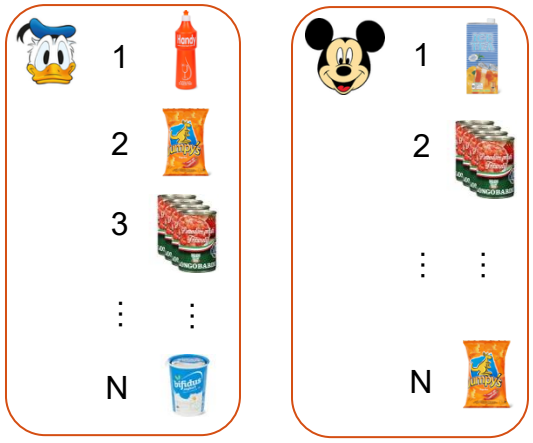
Zürcher Hochschule für Angewandte Wissenschaften
zh aw School of Engineering

MIGROS | Operations

	Bought	Sales	Feature XYZ
	True	CHF 12	23.23
	True	CHF 5	56.23
	False	CHF 0	12.34
	True	CHF 65	45.89
	True	CHF 12	23.54
	True	CHF 34	63.55
	False	CHF 0	28.23



Personalised recommendations



Recommender systems provides suggestions for items that are most pertinent to a particular user.

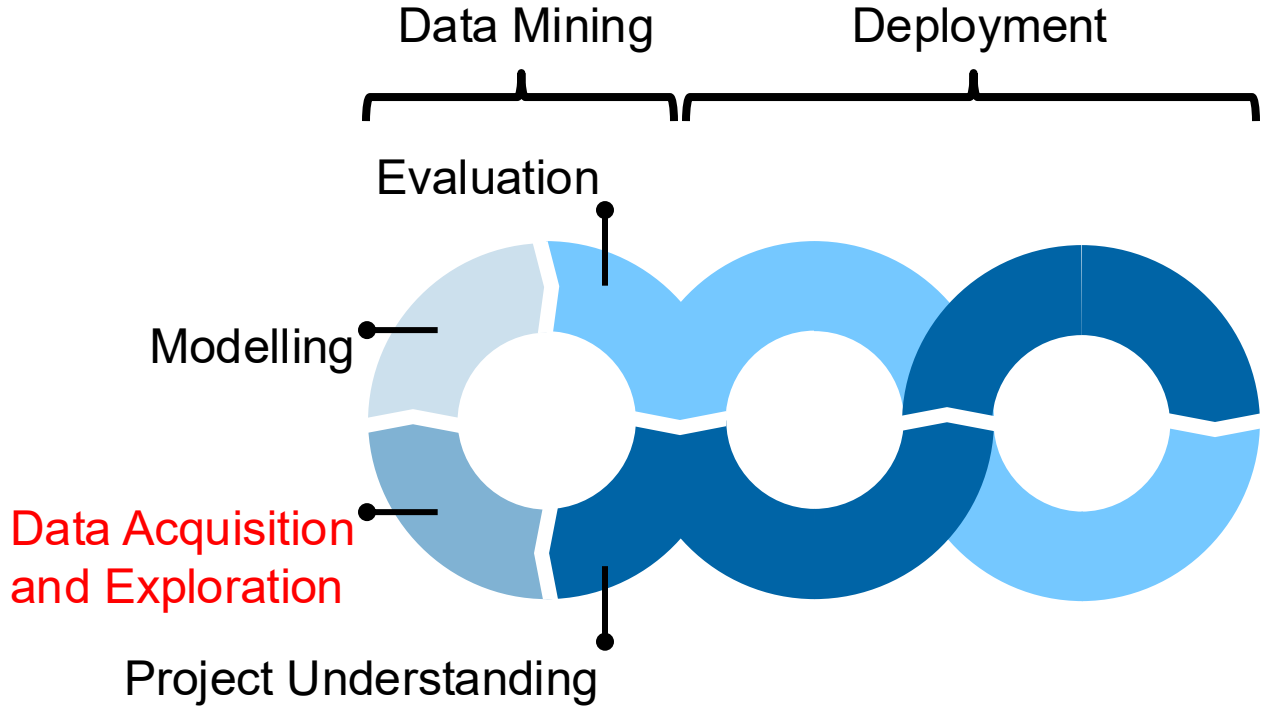
Identify the Data Mining Task

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering

Develop an automatic system for a real estate agent that determines the sales price of a flat based on data (location, size, year of construction, type of construction).

Data Product-Development

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering



Data Acquisition and Exploration

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

	Activities	Output/Artefact
Data Acquisition	• Identify data sources • Acquire data • Store and organize data • Compile data catalogue	• Data report
Data Exploration	• Inject data in analytics environment • Explore data, identify main characteristics	

Setup programatic data pipelines and document them to support traceability, reproducibility and allow for a systematic expansion of the data pool.

Exploratory data analysis

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

Data science-projects often start with an **incomplete knowledge** of the data creation process and different **data sources** are **combined**.

- Data Science «in the wild»
- «Observational Data»

A thorough understanding of the characteristics and data quality issues is required for the subsequent analysis and modelling steps.

Objectives of an exploratory data analysis

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

Identify…

- the main characteristics of the dataset
- data quality issues
  - missing attributes and/or values
  - duplicates
  - noise: erroneous entries and outliers
  - inconsistencies
- necessary pre-processing steps
  - cleaning
  - transformations/feature engineering, e.g. scaling, normalisation,…
- the need for more or different data
- properties that impact the subsequent modelling steps

Which issues can you find in the data?

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering

id	title	date	rating	episode	season	length
1	live free or die	01.11.2012	9.3	1	5	43
2	madrigal	06.12.2013	89	2	5	48
3	sunset	25.10.2011	9.3	6	3	47
4	grilled	20.12.2009	9.3	2	2	46
5	down	42.12.2009	8.3	4	2	333
6	cancer man	19.10.2008	8.3	4	1	48
7	live fre or die	01.11.2012	9.3	1	5	43

firstname	lastname	bdate	age	gender	phone
bryan	cranston	03-07-1956	65	1	999-9999
aaron	paul	27-08-1979	44	m	777-53474
anna, gunn	gunn	08-11-1968	52	0	040-15627

(Near) duplicate detection

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

id	title	date	rating	episode	season	length
1	live free or die	01.11.2012	9.3	1	5	43
2	madrigal	06.12.2013	89	2	5	48
3	sunset	25.10.2011	9.3	6	3	47
4	grilled	20.12.2009	9.3	2	2	46
5	episode title	25.10.2011	9.3	6	3	**46**
6	cancer man	19.10.2008	8.3	4	1	48
7	**live fre or die**	01.11.2012	9.3	1	5	43

- For numeric values: Distance/similarity of feature vectors
- For text: pairwise Levensthein distance between the texts

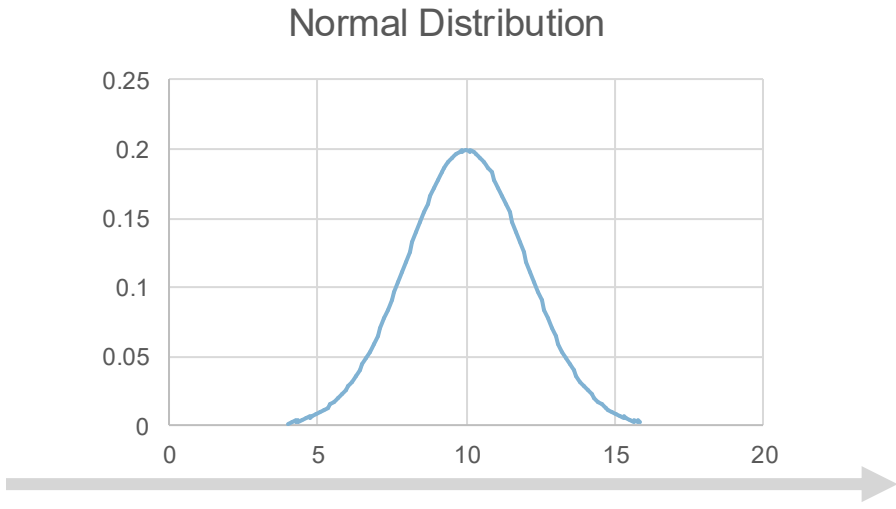
The tools in EDA: **Descriptive statistics** and **visualisation**

Zürcher Hochschule für Angewandte Wissenschaften

zh aw School of Engineering

The **data**:

...
7.6 0.09709303
7.72 0.1041539
7.84 0.11132675
7.96 0.11856598
8.08 0.12582217
8.2 0.13304262
8.32 0.14017191
8.44 0.14715251
8.56 0.15392563
8.68 0.1604319
8.8 0.1666123
8.92 0.172409
9.04 0.17776626
9.16 0.18263134
9.28 0.1869553
9.4 0.19069391
9.52 0.19380831
9.64 0.19626574
9.76 0.19804011
9.88 0.19911242
10 0.19947114
10.12 0.19911242
10.24 0.19804011
10.36 0.19626574
10.48 0.19380831
10.6 0.19069391
10.72 0.1869553
10.84 0.18263134
10.96 0.17776626
11.08 0.172409
11.2 0.1666123
11.32 0.1604319
11.44 0.15392563
11.56 0.14715251
11.68 0.14017191
...



The probability density of the normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

is determined by 2 parameters:

- Expectation value (mean) μ
- Standard deviation σ

The **description**:

- $\mu = 10.0$
- $\sigma = 2.0$

Statistical variable types

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

Variable Type	Level of Measurement (Scale)	Allowed Operations	Allowed Statistics	Examples
Categorical	nominal	count	mode, Chi-squared	profession, sex
	ordinal	count, sort	+ rank statistics (Spearman)	ratings, school marks
Numerical • discrete • continuous	interval	count, sort, subtract	+ mean, median, standard dev., correlation	time, temperature in Celsius
	ratio	count, sort, subtract, divide	+ geometric mean, harmonic mean	cost, age, size, temperature in Kelvin

No strict connection between statistical variable types and data types / encoding schemes on the computer system.

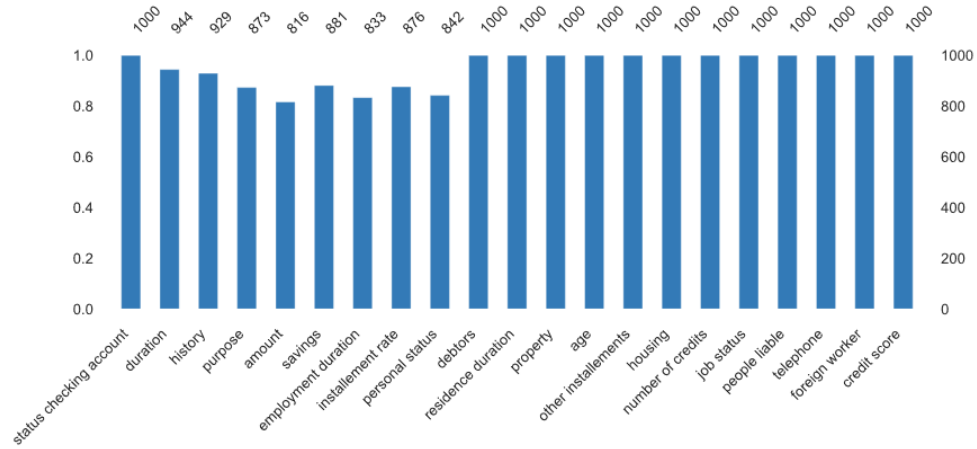
Dataset overview

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

- number of variables per type, i.e. numerical/categorical
- number of observations
- missing values
- duplicates

Dataset statistics

Number of variables	21
Number of observations	1000
Missing cells	1006
Missing cells (%)	4.8%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	813.6 KiB
Average record size in memory	833.2 B



A simple visualization of nullity by column.

Univariate description of categorical variables

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

- number of unique values
- missing values
- frequency distribution of the values

purpose
Categorical

HIGH CORRELATION
MISSING

Distinct	10
Distinct (%)	1.1%
Missing	127
Missing (%)	12.7%
Memory size	55.3 KiB

number of missing values divided by number of observations

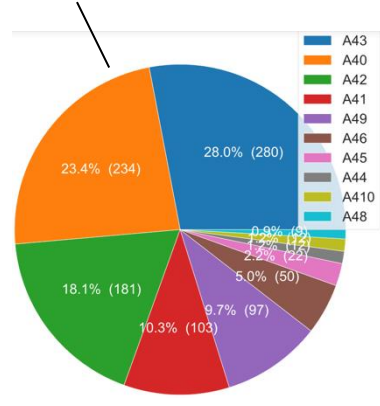
relative frequency: number (Count) divided by total number of observations

bar chart of the relative frequencies

pie chart: relative frequencies encode the area of the circle segments

Common Values

Value	Count	Frequency (%)
A43	248	24.8%
A40	201	20.1%
A42	160	16.0%
A41	90	9.0%
A49	80	8.0%
A46	45	4.5%
A45	19	1.9%
A44	11	1.1%
A410	11	1.1%
A48	8	0.8%
(Missing)	127	12.7%

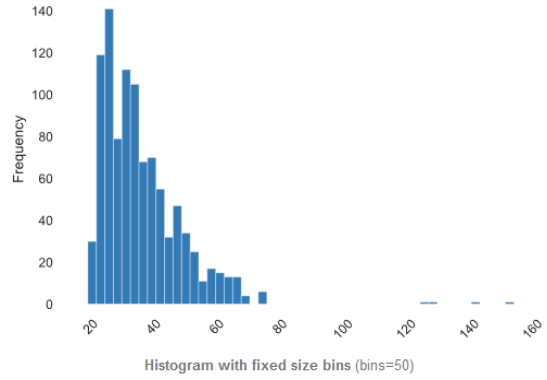


Univariate description of numerical variables

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

- number of unique values
- frequencies
- missing values
- extreme values: minimum and maximum
- fraction of zero-values
- negative values
- histogramm



age

Distinct	57	Minimum	19
Distinct (%)	5.7%	Maximum	154
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	35.946	Memory size	7.9 KiB

Statistics | Histogram | Common values | Extreme values

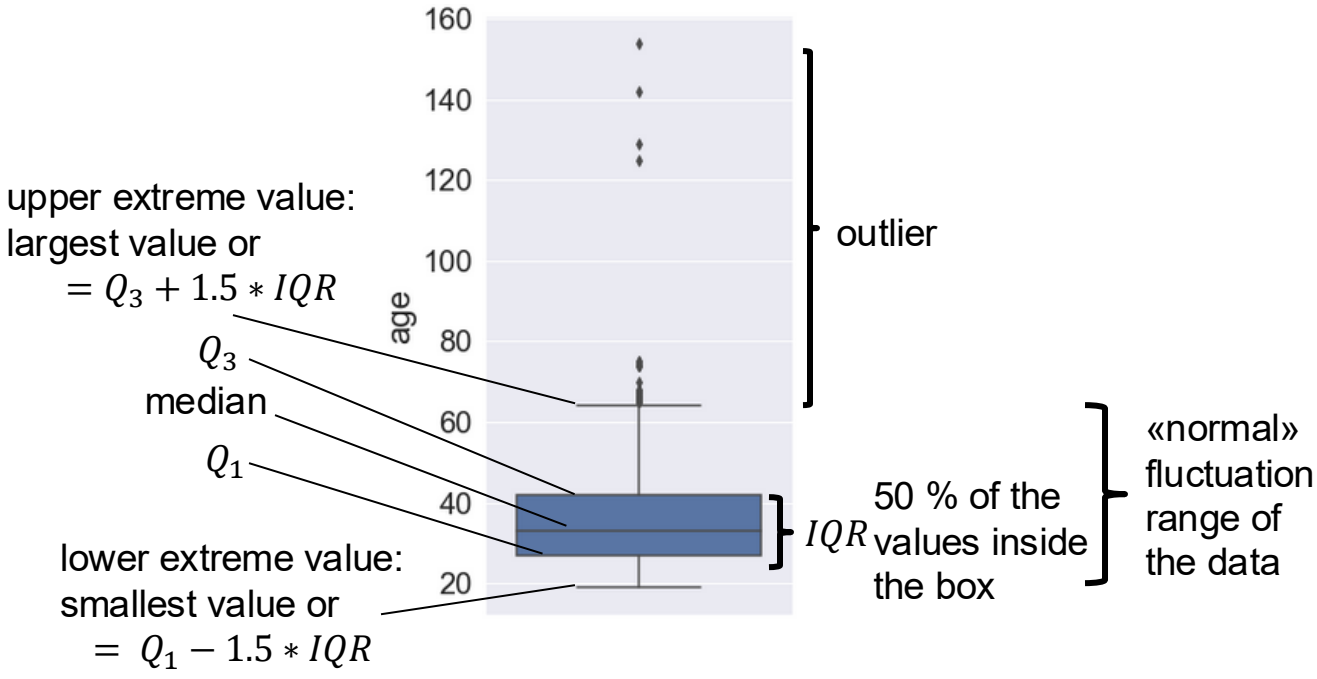
Minimum 5 values | Maximum 5 values

Value	Count	Frequency (%)
154	1	0.1%
142	1	0.1%
129	1	0.1%
125	1	0.1%
75	2	0.2%
74	4	0.4%
70	1	0.1%
68	3	0.3%
67	3	0.3%
66	5	0.5%

Univariate description of numerical variables

- Central tendency:
  - arithmetic mean
  - median
  - mode
  - quantiles
- Dispersion metrics:
  - interquartile range IQR
  - variance
  - standard deviation

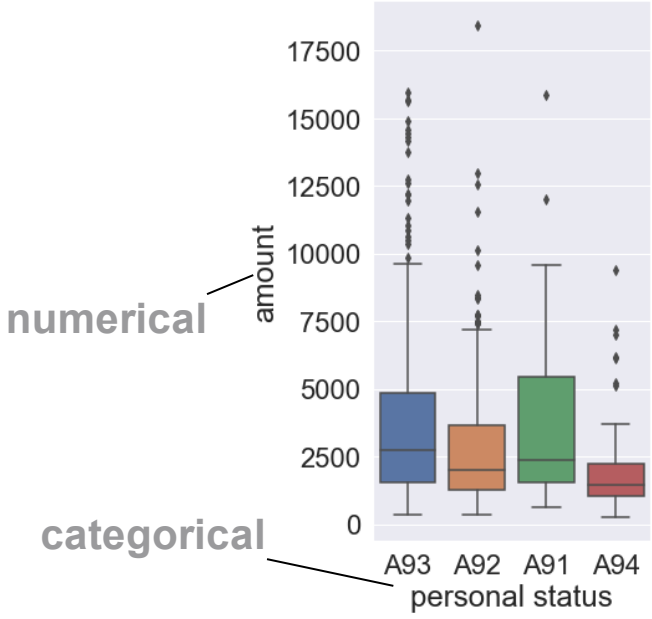
Five number-summary and boxplot:

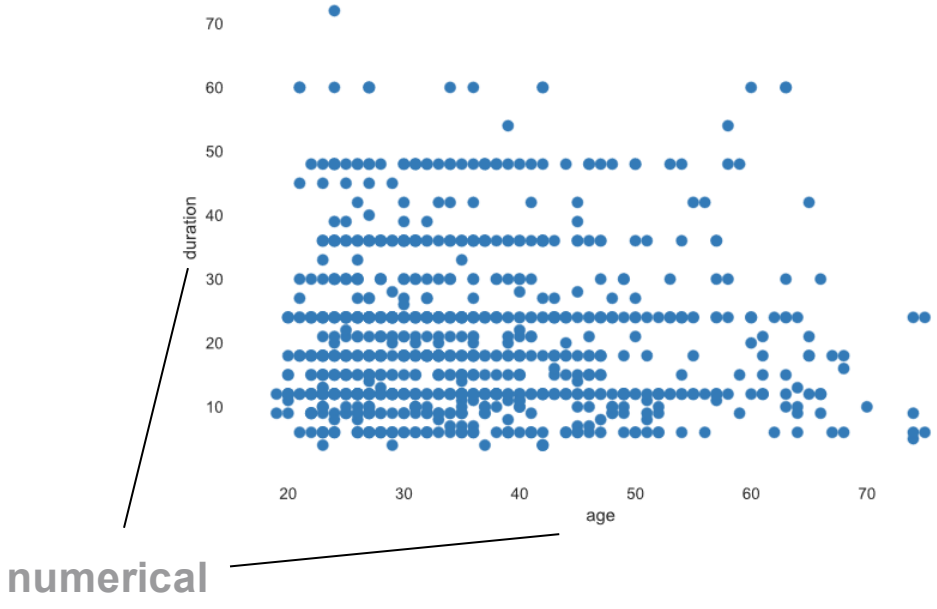


Outlier detection

title	Date	…	length
live free or die	01.11.2012		43
madrigal	06.12.2013		48
sunset	25.10.2011		47
grilled	20.12.2009		46
down	22.12.2009		**91**
cancer man	19.10.2008		48
shotgun	09.11.2012		47

Assessing relationships between variables

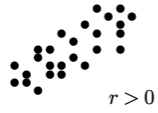


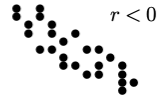


Measuring the strength of relationships between variables

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

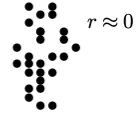
positive correlation,
values (x_i, y_i) group around
a line with positive slope





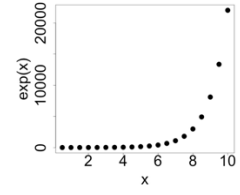
negative correlation,
inverse relationship,
values (x_i, y_i) group around
a line with negative slope

uncorrelated,
or no linear (or monotonic)
relationship

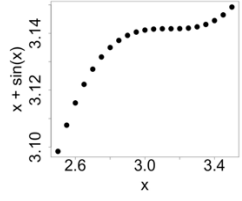


$r \approx 0$

$r = 0.7 \quad r_{SP} = 1.0$

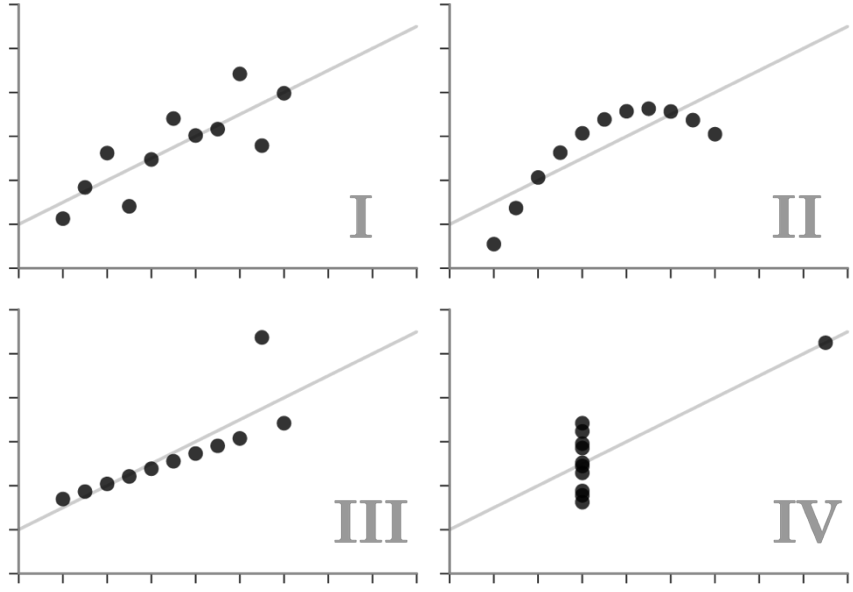


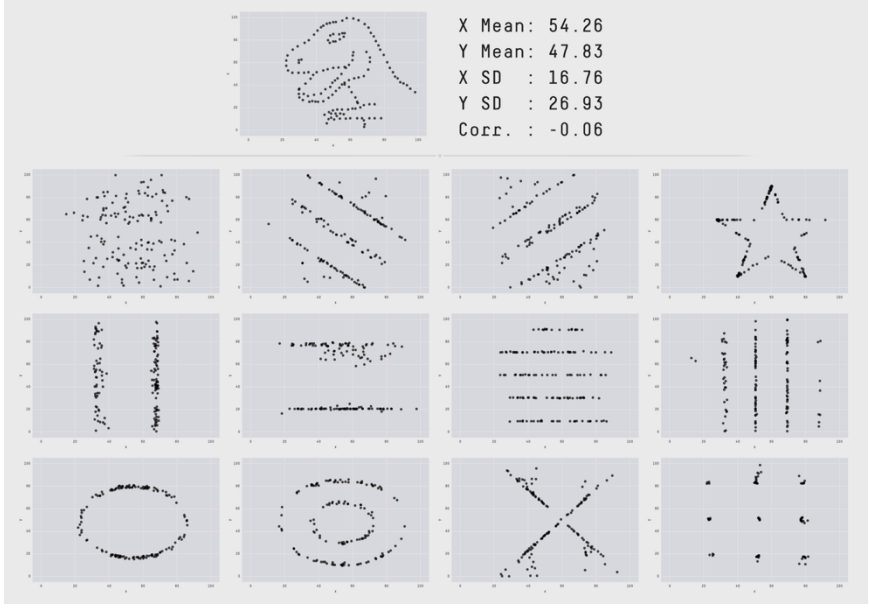
$r = 0.86 \quad r_{SP} = 1.0$



EDA: Combination of statistical metrics and visualisation

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering





Matejka, J., & Fitzmaurice, G. (n.d.). *Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing*. https://doi.org/10.1145/3025453.3025912
https://www.autodesk.com/research/publications/same-stats-different-graphs

How do we deal with missing data?

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

title	date	rating	episode	season	length
live free or die	01.11.2012	9.3	1	5	43
madrigal	06.12.2013	8.9	2	5	48
sunset	25.10.2011		6	3	47
grilled	20.12.2009	9.3	2	2	46
down	22.12.2009	8.3	4	2	
cancer man	19.10.2008	8.3	4	1	48
shotgun	09.11.2012	8.7	5	4	47

How do we deal with missing or incorrect data?

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

Ignore the Tuple

+ can be easily done
+ no computational effort

- loss of information
- unnecessary if the attribute is not needed

Enter Value Manually

+ for small datasets effective
+ “real” value

- not for large datasets
- time consuming
- error-prone

Use Attribute Mean

+ simple to implement

- ignores correlations between attributes

Use a Global Constant

+ can be easily done
+ missing values are marked

- values can not be used in algorithms

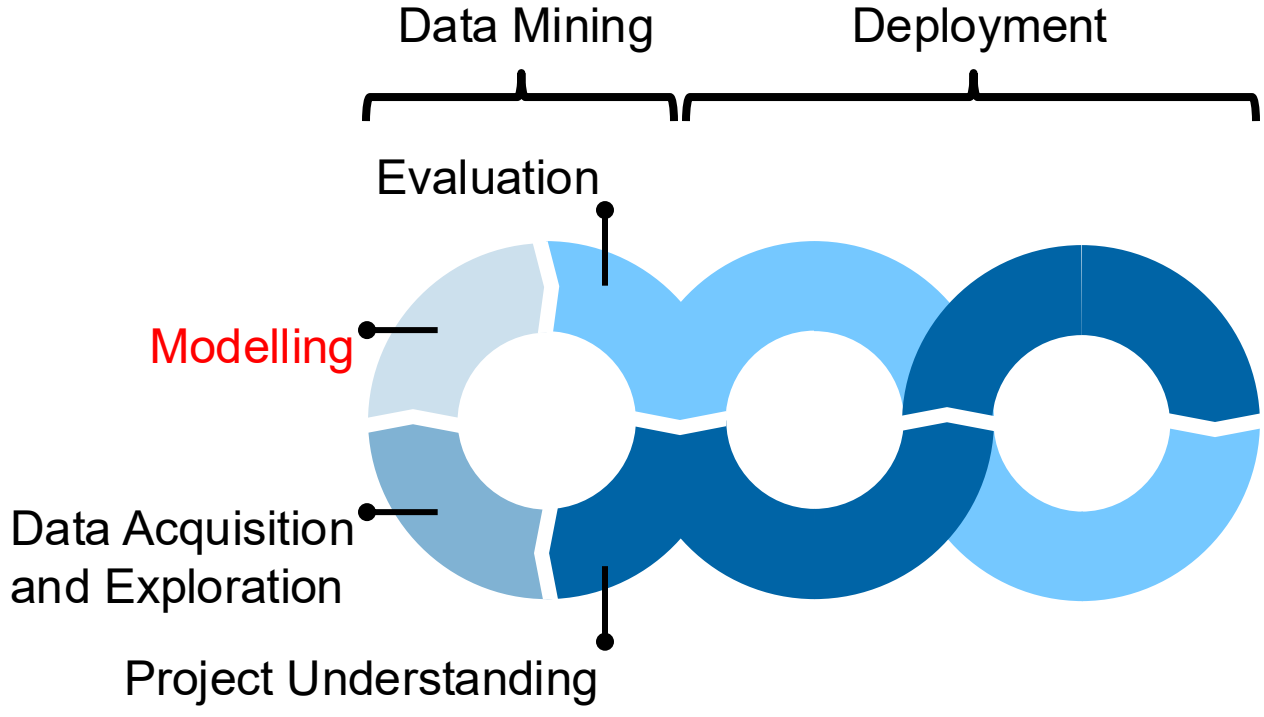
Use Most Probable Value

+ most accurate approximation of the value

- most computational effort

Data Product-Development

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering



Modelling

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering

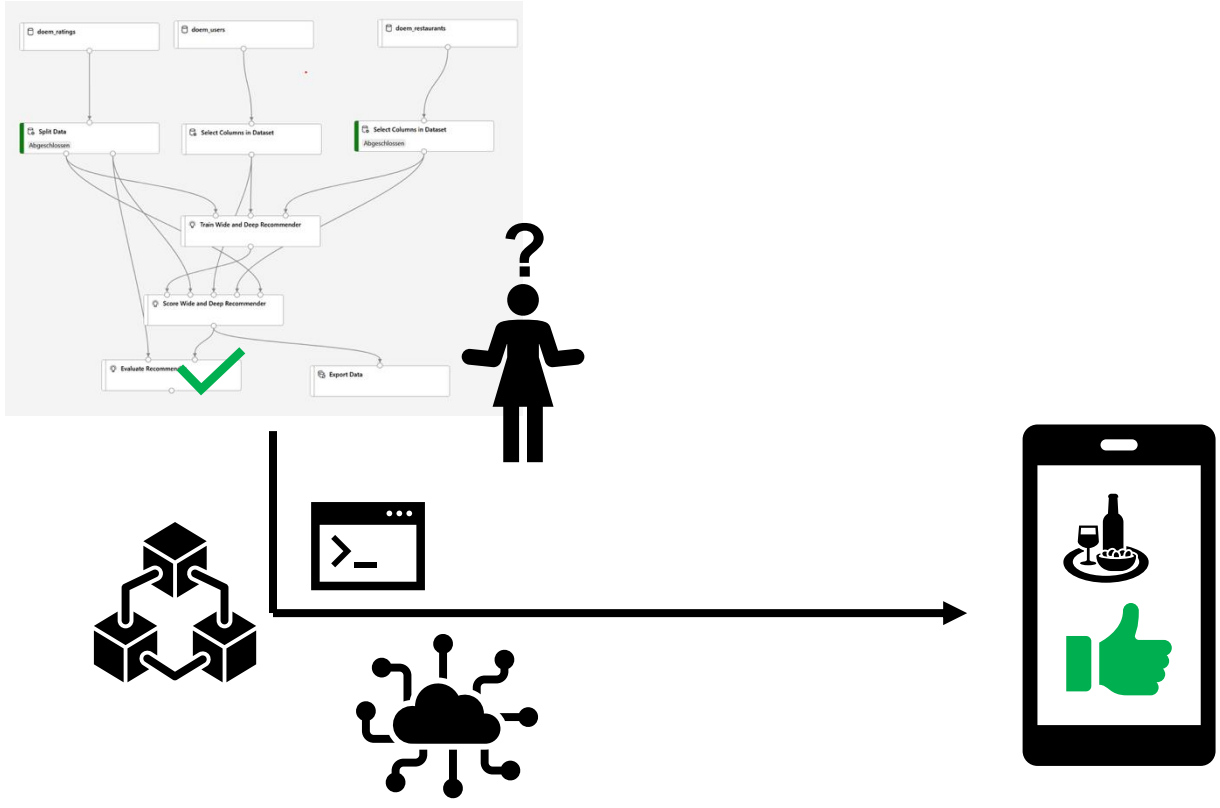
	Activities	Output/Artefact
Modelling	• Construct modelling pipeline • Transform data, engineer features • Train model • Evaluate model • Interpret results (predictions and model)	• Modelling report

Evaluation

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering

	Activities	Output/Artefact
Evaluation	• Decide: Do results meet user needs? → planning of the deployment → project stop → Additional data mining iteration	• Evaluation decision log

Deployment

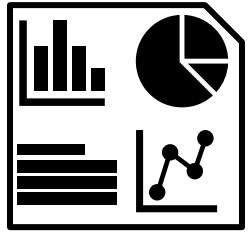

doem_ratings
doem_users
doem_restaurants
Split Data
Abgeschlossen
Select Columns in Dataset
Select Columns in Dataset
Abgeschlossen
Train Wide and Deep Recommender
Score Wide and Deep Recommender
Export Data

Drivers of Complexity

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

simple — complex

e.g. one –off report

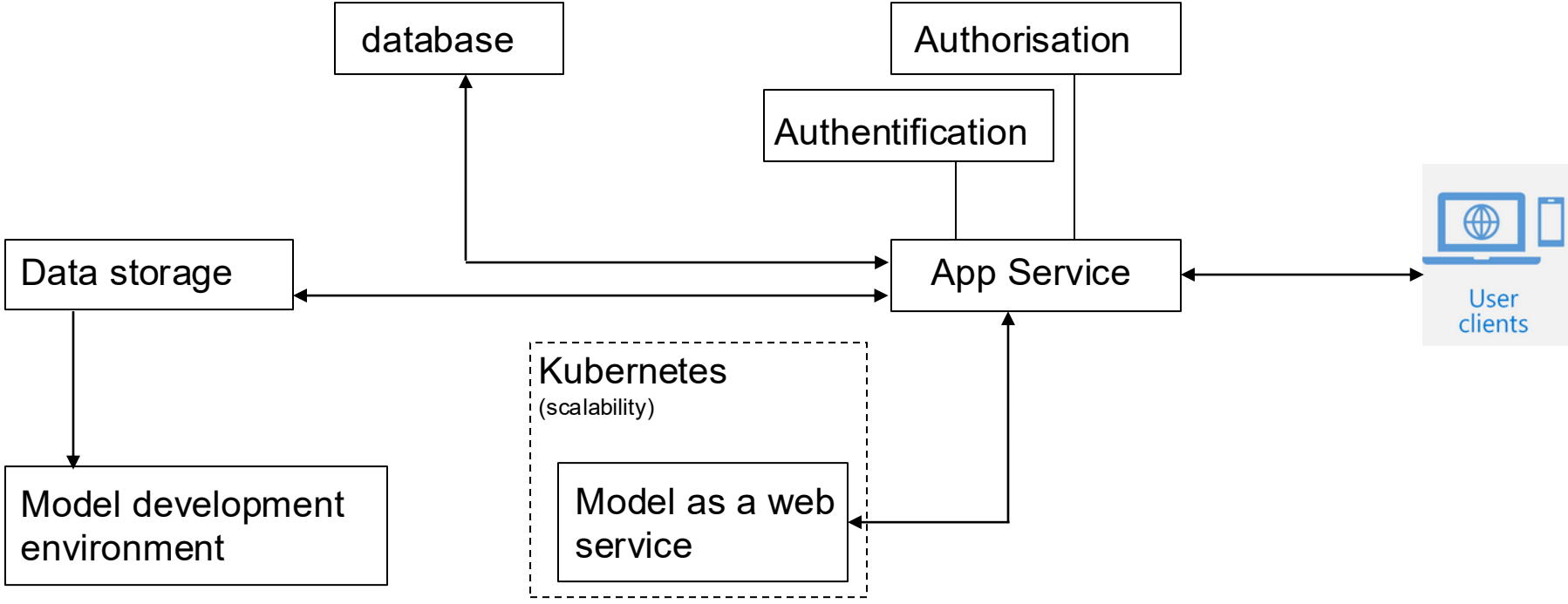


Drivers of complexity

- **Automated training** and deployment on changing data
- **Interactivity** and personalisation
- Integration in larger **software application**
- **External** users
- **Legal/regulatory** requirements

«continuously adapting» recommender system in online shop

Architecture Blueprint



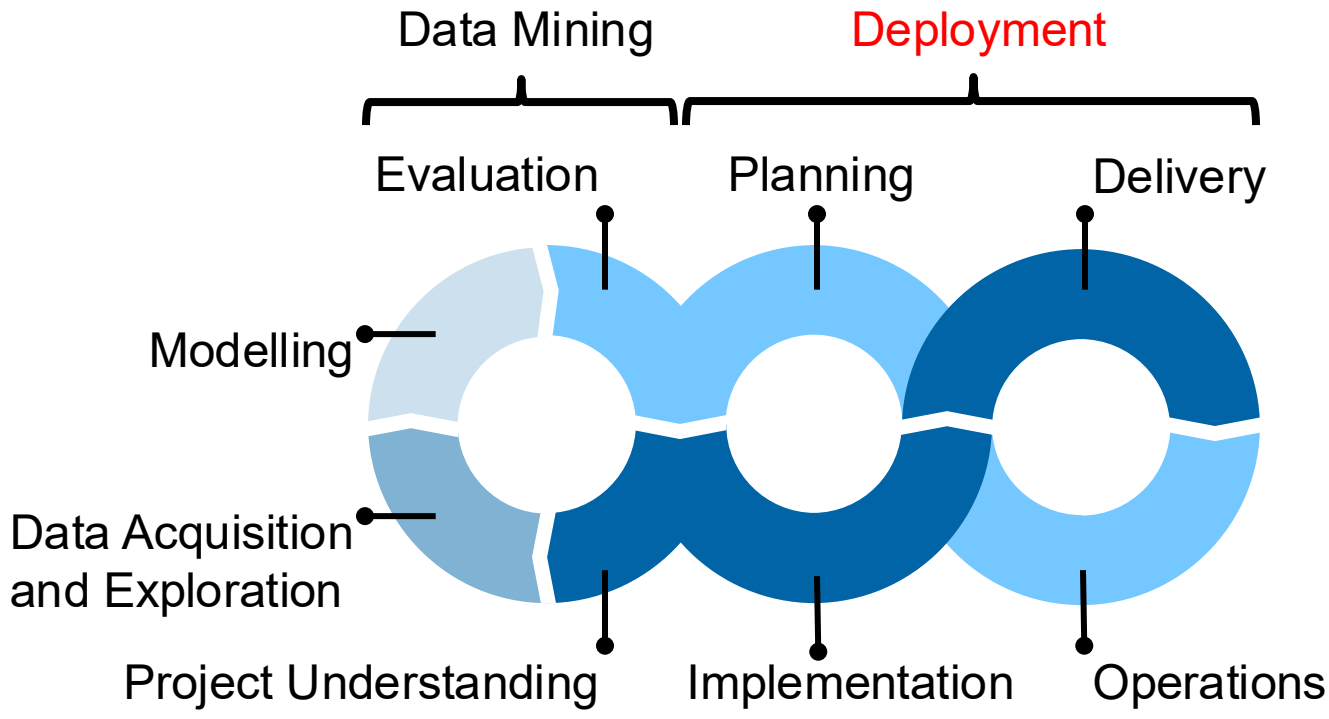
Deployment Beyond Technical Aspects

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering

The data product only adds value if it is accepted by the user base. Aspects to consider are for example:

- Training and Change Management
- Adjustments to business processes
- Connection to other products and offers (platform)
- Continuous collection of user feedback for iterative refinements

Data Product-Development



Data Product-Development
References and Further Reading

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering

- Kempf D., Dömer M. (**2022**). Is It Ops That Make Data Science Scientific? *Archives of Data Science*, Series A, vol 8, p. 12.
- Microsoft **Team-Data-Science-Process**:
  https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview
  https://github.com/Azure/Azure-TDSP-ProjectTemplate
- The “Cross-Industry Standard Process for Data Mining”: Shearer, C. (**2000**). The **CRISP-DM** Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5*(4), 13–22.
  (Datei auf Moodle)
- **KDD**: Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (**1996**). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–54. https://doi.org/10.1609/aimag.v17i3.1230
  (Datei auf Moodle)