Universität Stuttgart
KI – Institute for Artificial Intelligence
Analytic Computing

Machine Learning
2 CRISP-DM

Prof. Dr. Steffen Staab
Nadeen Fatallah
Daniel Frank
Akram Sadat Hosseini
Jiaxin Pan
Osama Mohamed
Arvindh Arunbabu
Tim Schneider
Yi Wang

https://www.ki.uni-stuttgart.de/

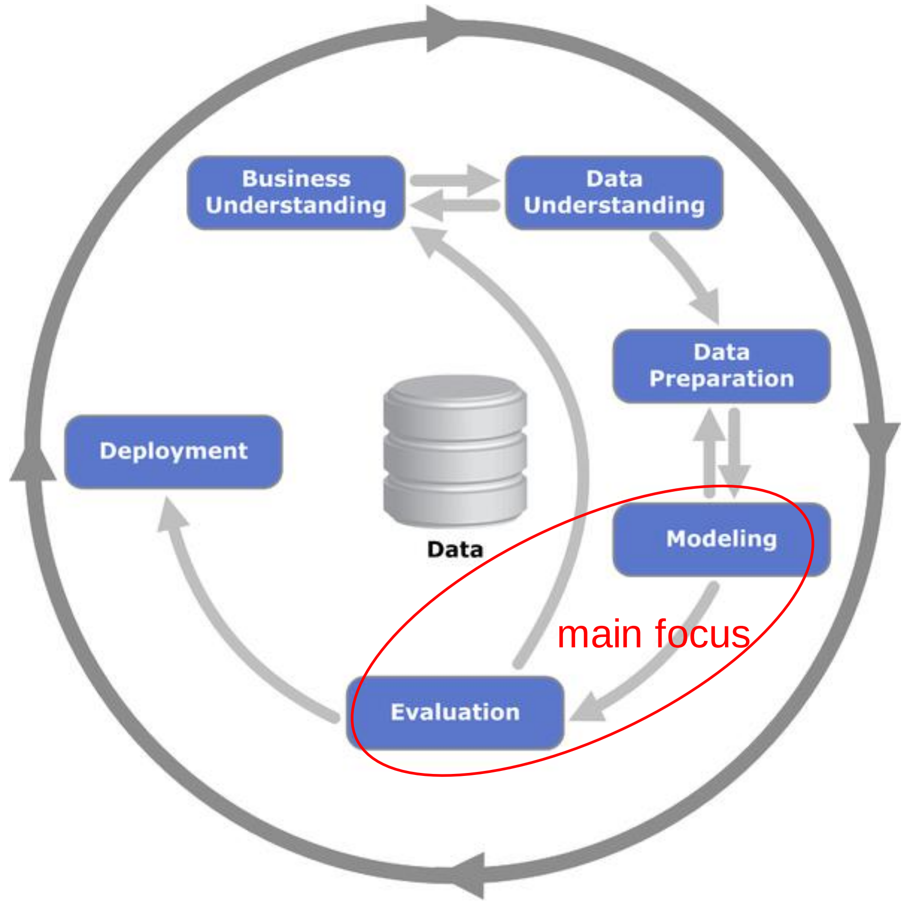
Today‘s objectives (April 7 & April 14, 2025)

Completing this slide deck you should

- Know the steps of the knowledge discovery process
  - meaning, rationale, procedures
- Know their role for ML
- Know techniques underlying these steps

The knowledge discovery process

The Crisp-DM modell



Cross Industry Standard Process for Data Mining

https://www.the-modeling-agency.com/crisp-dm.pdf

https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Overview of the CRISP-DM tasks

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

main focus of
the overall course

Deployment

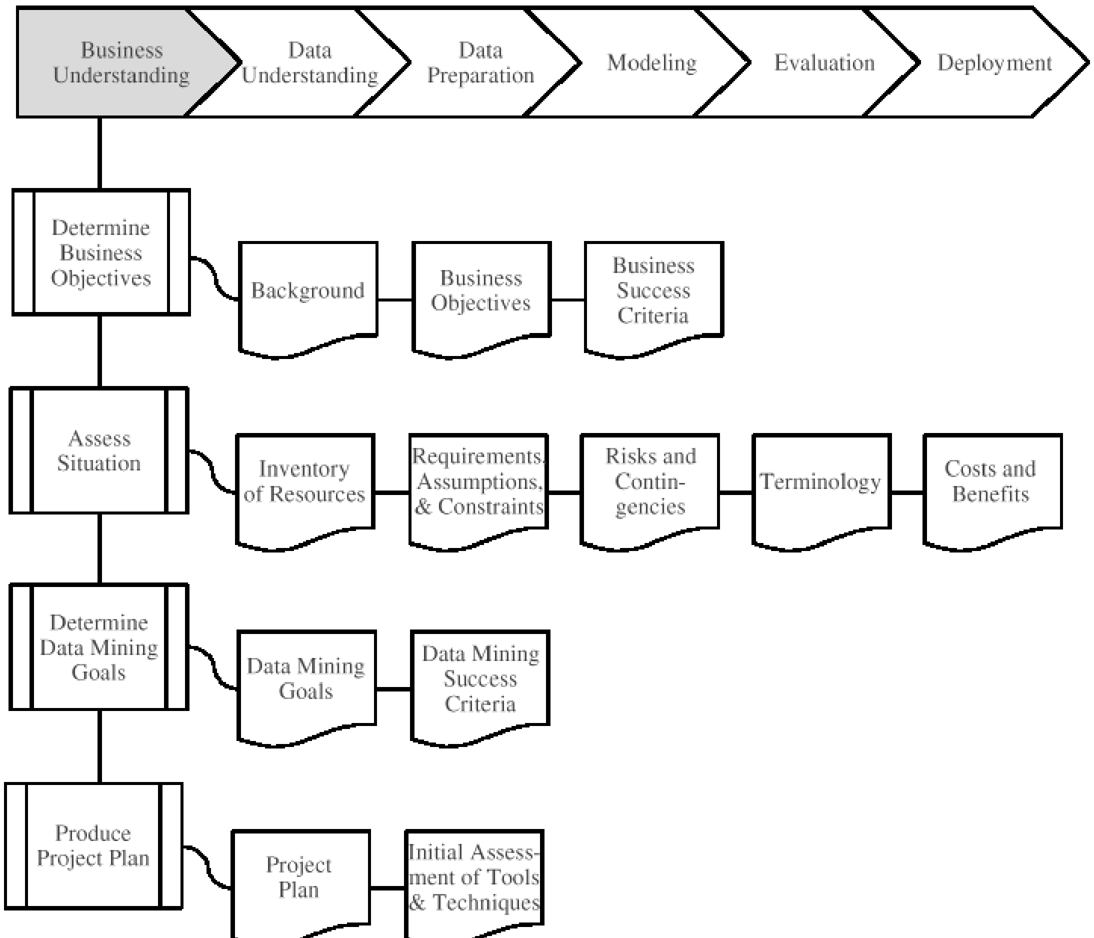
Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Business Understanding



Business Understanding

Determine Business Objectives

- understand from a business perspective what
  the client really wants to accomplish
  ⇔ do not produce the right answers to the wrong question
- identify key persons (management, finance, domain expert, user)
- define **success criteria** - related to business objectives

Assess Situation

- identify available resources as well as constraints and assumptions (e.g. legal issues)
- **identify lack of data/resources**
- **identify/change business process for collecting data/resources**
  - **determine costs for updating and performing new business process**
- identify risks (business, organisational, technical)

Business Understanding

Determine Data Mining Goals

- derive data mining goals from business objectives
- define data mining success criteria (e.g. model accuracy, model performance, ...)

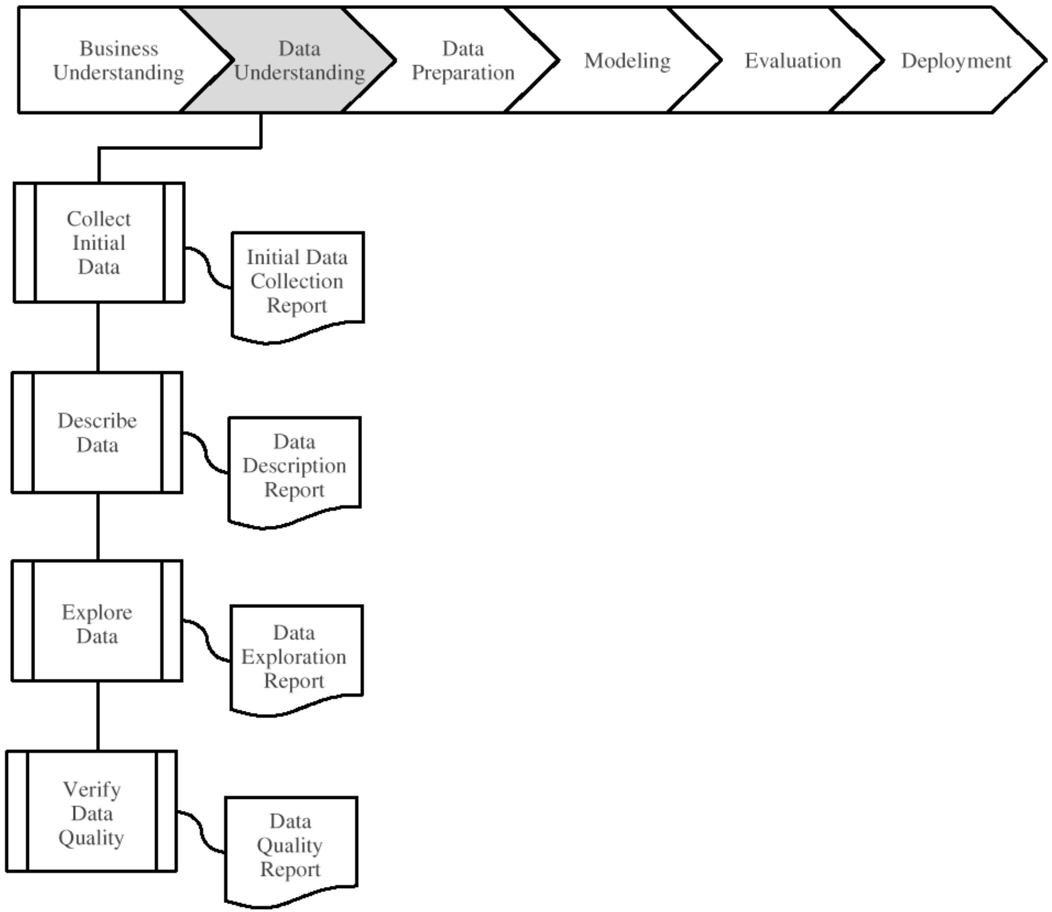
Produce Project Plan

- take iterations into account
- typical effort distribution:
  - 50% - 70% in Data Preparation Phase
  - 20% - 30% in Data Understanding Phase
  - 10% - 20% in Modeling, Evaluation and Business Understanding Phase
  - 5% - 10% in Deployment Phase

Business Understanding: Predicting DAX index value

- No predictions reasonable for Saturdays and Sundays
- Limited predictability
  - E.g. politically motivated announcement of tariffs cannot be predicted from the data
    - people try to include social media data for predictions!

Data Understanding



Data Understanding

- **Collect Initial Data**
  - identify relevant attributes
  - identify inconsistencies
    between sources (record linkage!)
- **Describe Data**
  - characterize attributes
    (relevance, statistical characteristics, ...)
    - mean, median, skewness, outliers, …
    - correlations between attributes
- **Explore Data**
  - querying, visualization
- **Verify Data Quality**
  - identify errors in data
  - number of missing values
    - identify false encodings of missing values
      (e.g. May 20, 1875, data reference point in Cobol)

Two people not understanding data

Elon Musk claims there are 150-year-olds on Social Security

Data Understanding

- sufficient number of examples?
  - overall? for all target classes / for all target value areas?
  - varies depending on degrees of freedom / no of parameters
- examples contain all/sufficiently many relevant attributes?
- sufficient quality of data?
  - small number of errors in values
  - small number of missing values
- possibility to score quality of learned knowledge?

Data Understanding: Biases

Is the data biased?

- Class sizes:
  - Gay population: 1-16% depending on definition & survey context
  - Wang, Yilun, and Michal Kosinski. "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images." Journal of personality and social psychology 114.2 (2018): 246.
    - roughly 50:50 split
- Claudia Wagner et al. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia.
- Ntoutsi, Eirini, et al. "Bias in data-driven artificial intelligence systems—An introductory survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10.3 (2020): e1356.

Data Understanding for Learning a Classifier

Thinking about data

id	size	price	color	target
1	3	3,29	white	cheap
2	4	24,99	black	exp
3	4	23,59	red	exp
4	3	4,59	black	cheap
5	3	6,99	black	exp
6	5	19,99	red	exp
7	3	24,99	black	exp
8	4	2,99		cheap
9	4	21,99	blue	cheap
10	5	29,99	red	exp
11	6	23,99	red	exp
12	5	8,99	black	cheap
13	5	3389,99	black	exp

All data given in one table?
Very often in machine learning:
Universal Relation

Which attribute is a target variable?

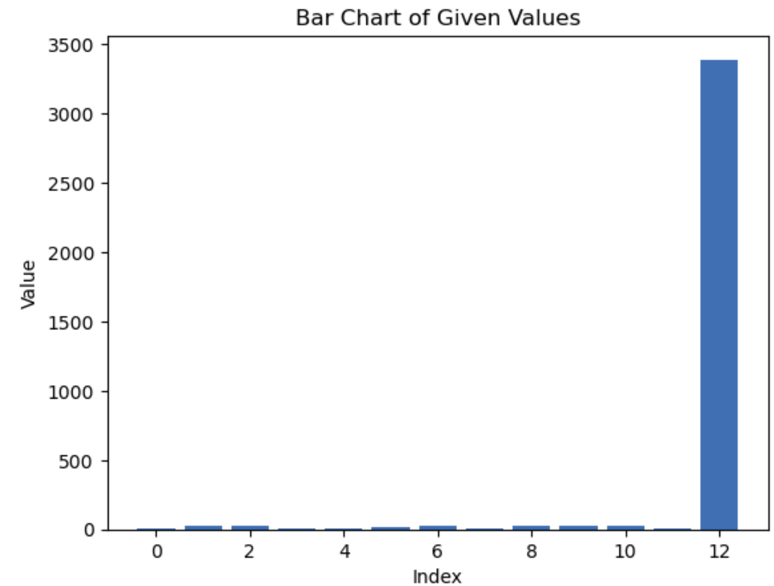
Which attribute is useless? (e.g. id)

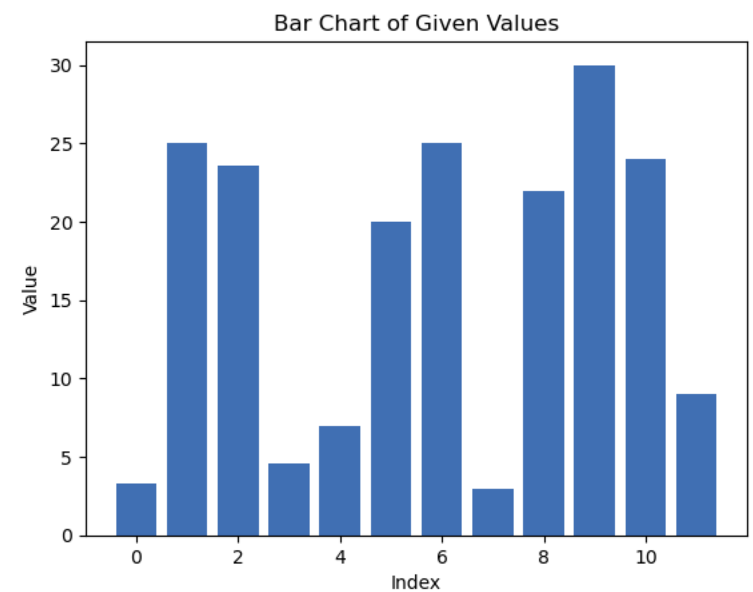
Which attribute is nominal (color), ordinal (size, target), or from an interval scale, ratio (price), cardinal

Missing values?

Bar chart with and without outlier

id	size	price	color	target
1	3	3,29	white	cheap
2	4	24,99	black	exp
3	4	23,59	red	exp
4	3	4,59	black	cheap
5	3	6,99	black	exp
6	5	19,99	red	exp
7	3	24,99	black	exp
8	4	2,99		cheap
9	4	21,99	blue	cheap
10	5	29,99	red	exp
11	6	23,99	red	exp
12	5	8,99	black	cheap
13	5	3389,99	black	exp





Data Understanding for Learning a Classifier

Thinking about data

id	size	price	color	target
1	3	3,29	white	cheap
2	4	24,99	black	exp
3	4	23,59	red	exp
4	3	4,59	black	cheap
5	3	6,99	black	exp
6	5	19,99	red	exp
7	3	24,99	black	exp
8	4	2,99		cheap
9	4	21,99	blue	cheap
10	5	29,99	red	exp
11	6	23,99	red	exp
12	5	8,99	black	cheap
13	5	3389,99	black	exp

All data given in one table?
Very often in machine learning:
Universal Relation

Which attribute is a target variable?

Which attribute is useless? (e.g. id)

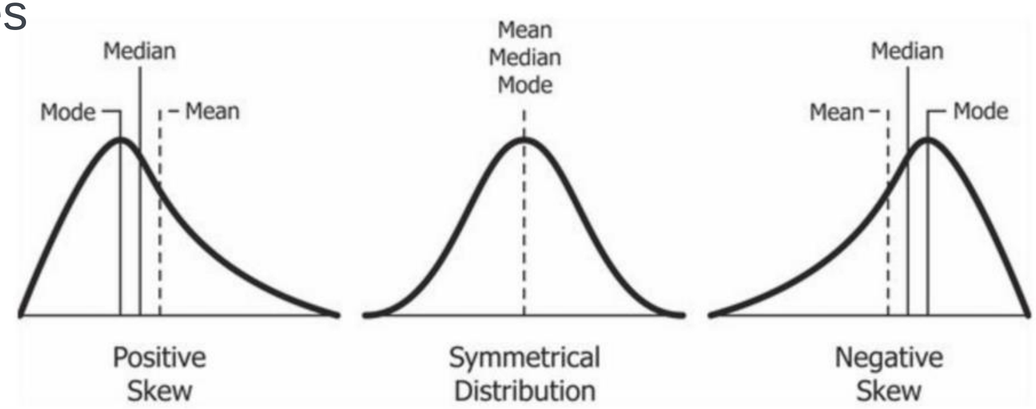
Which attribute is nominal (color), ordinal (size, target), or from an interval scale, ratio (price), cardinal

Missing values?

Outliers? (invalidate statistics)

Statistical descriptions of data: mode, median, mean

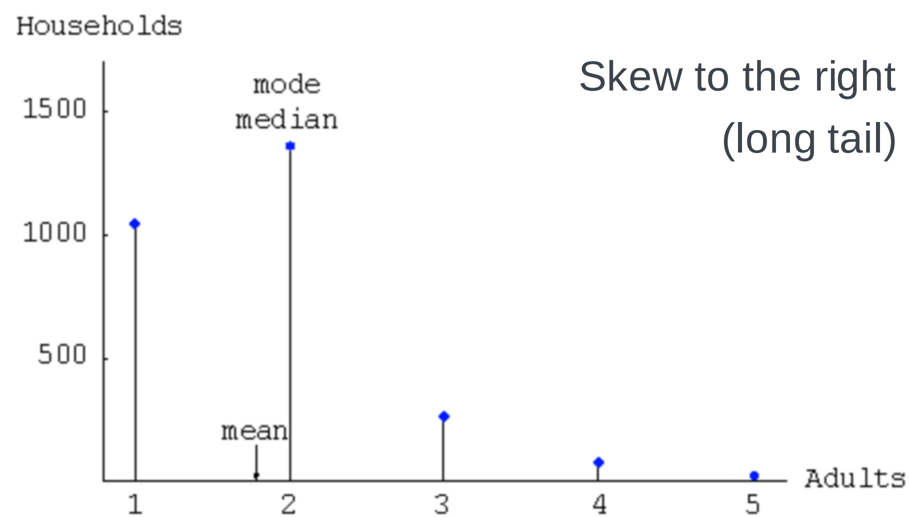
- Mode: Which attribute value appears most often?
  - May 20, 1875; black
  - always applicable
- Median $\tilde{x}$
  - given $x_1 \leq x_2 \leq \cdots \leq x_n$, $\tilde{x}$ is $x_{\frac{n+1}{2}}$ if n is odd, else $\frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$
  - applicable to ordinal, interval and ratio scales
- Mean $\mu = \mathbb{E}_{x \in X}[x] = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - applicable to interval and ratio scales
- Compare mode, median and mean!



https://en.wikipedia.org/wiki/File:Relationship_between_mean_and_median_under_different_skewness.png

Statistical descriptions of data: deviation from mean

- standard deviation σ / variance σ^2
  - $\mathbb{E}_{x \in X}[(x-\mu)^2]$
  - σ^2 = *Sample variance* for n data points: $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2$
- Fisher-Pearson skewness
  - $\mathbb{E}_{x \in X}[(\frac{x-\mu}{\sigma})^3]$
  - $\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^3$



Skew to the right
(long tail)

id	size	price	color	target
1	3	3,29	white	cheap
2	4	24,99	black	exp
3	4	23,59	red	exp
4	3	4,59	black	cheap
5	3	6,99	black	exp
6	5	19,99	red	exp
7	3	24,99	black	exp
8	4	2,99		cheap
9	4	21,99	blue	cheap
10	5	29,99	red	exp
11	6	23,99	red	exp
12	5	8,99	black	cheap
13	5	3389,99	black	exp

Mode: 24.99

Median: 21.99

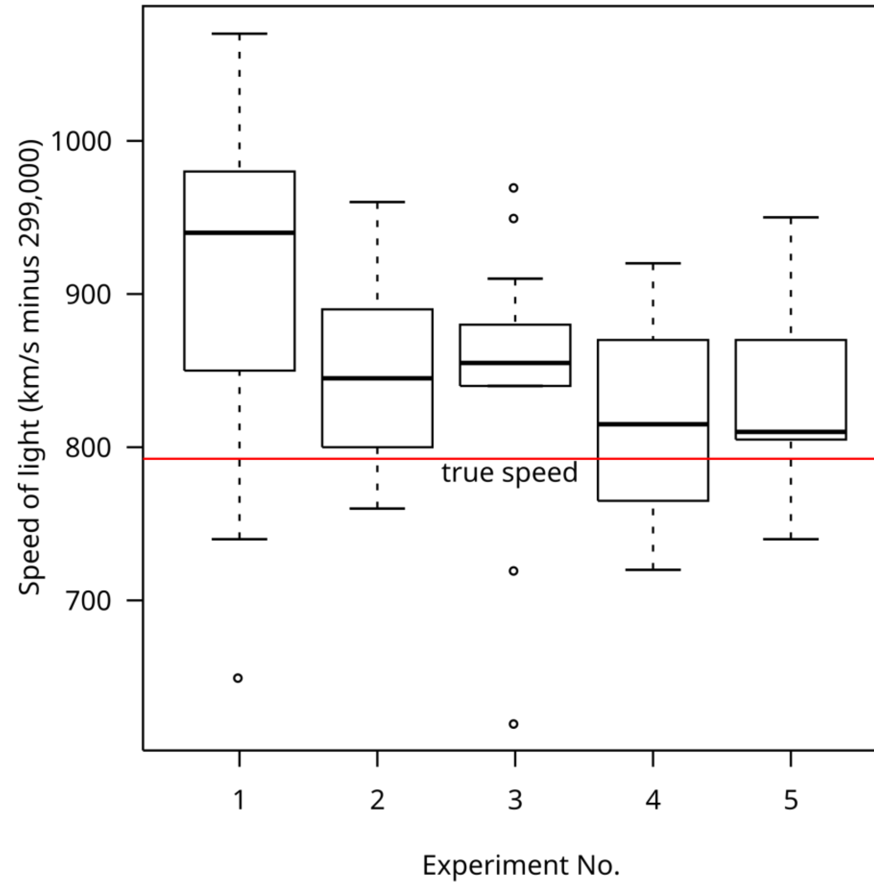
Mean: 275.87461538461537

Standard Deviation: 935.7251126320466

Fisher-Pearson Skew: 3.604869460277469

Quartiles and Outliers

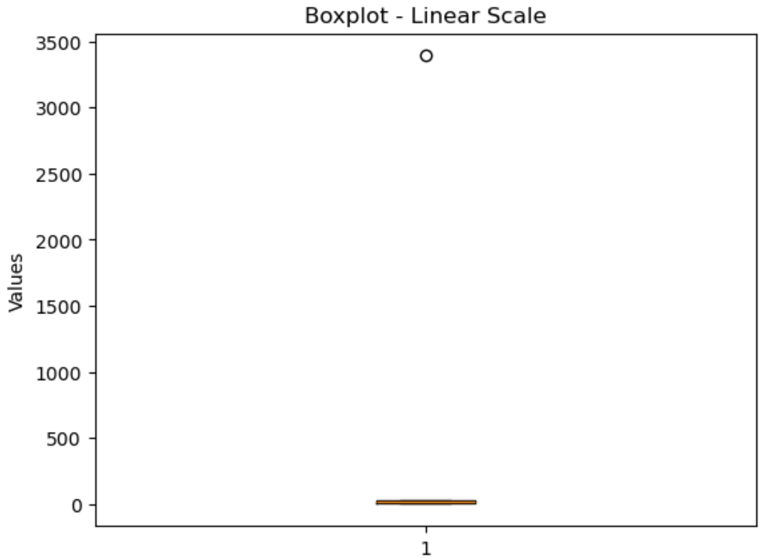
- 1st, 2nd, 3rd, 4th Quartile
- $\mathrm{IQR} = Q_3 - Q_1$
- Outliers
  - Beyond Tukey's fences $[Q_1 - k\ \mathrm{IQR},\quad Q_3 + k\ \mathrm{IQR}]$
    - $k = 1.5$ or $k = 3$ (far out)
  - Alternative: more than 3 standard deviations from mean (z-score rule)

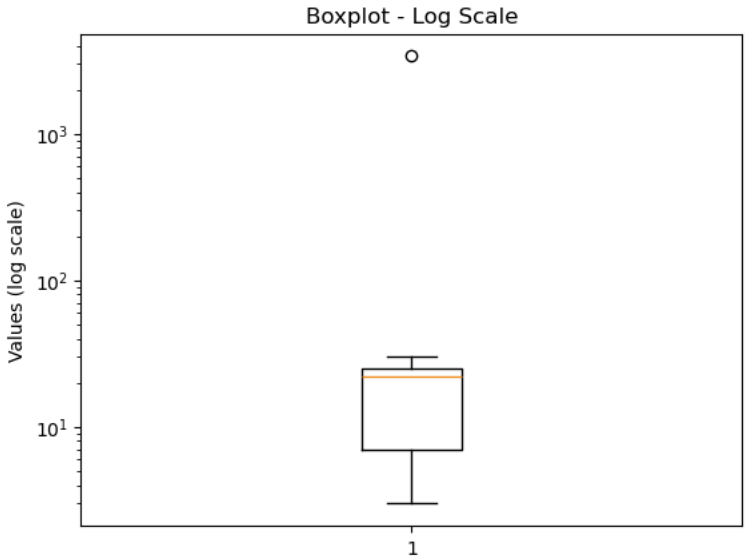


Box plots with $Q_1, Q_2 = \mu, Q_3$ boundaries and outliers as dots

Boxplot of price data

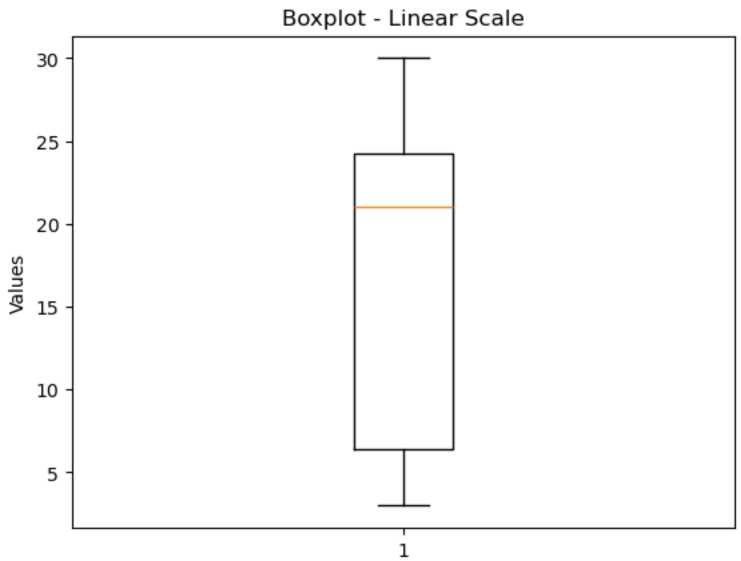
id	size	price	color	target
1	3	3,29	white	cheap
2	4	24,99	black	exp
3	4	23,59	red	exp
4	3	4,59	black	cheap
5	3	6,99	black	exp
6	5	19,99	red	exp
7	3	24,99	black	exp
8	4	2,99		cheap
9	4	21,99	blue	cheap
10	5	29,99	red	exp
11	6	23,99	red	exp
12	5	8,99	black	cheap
13	5	3389,99	black	exp



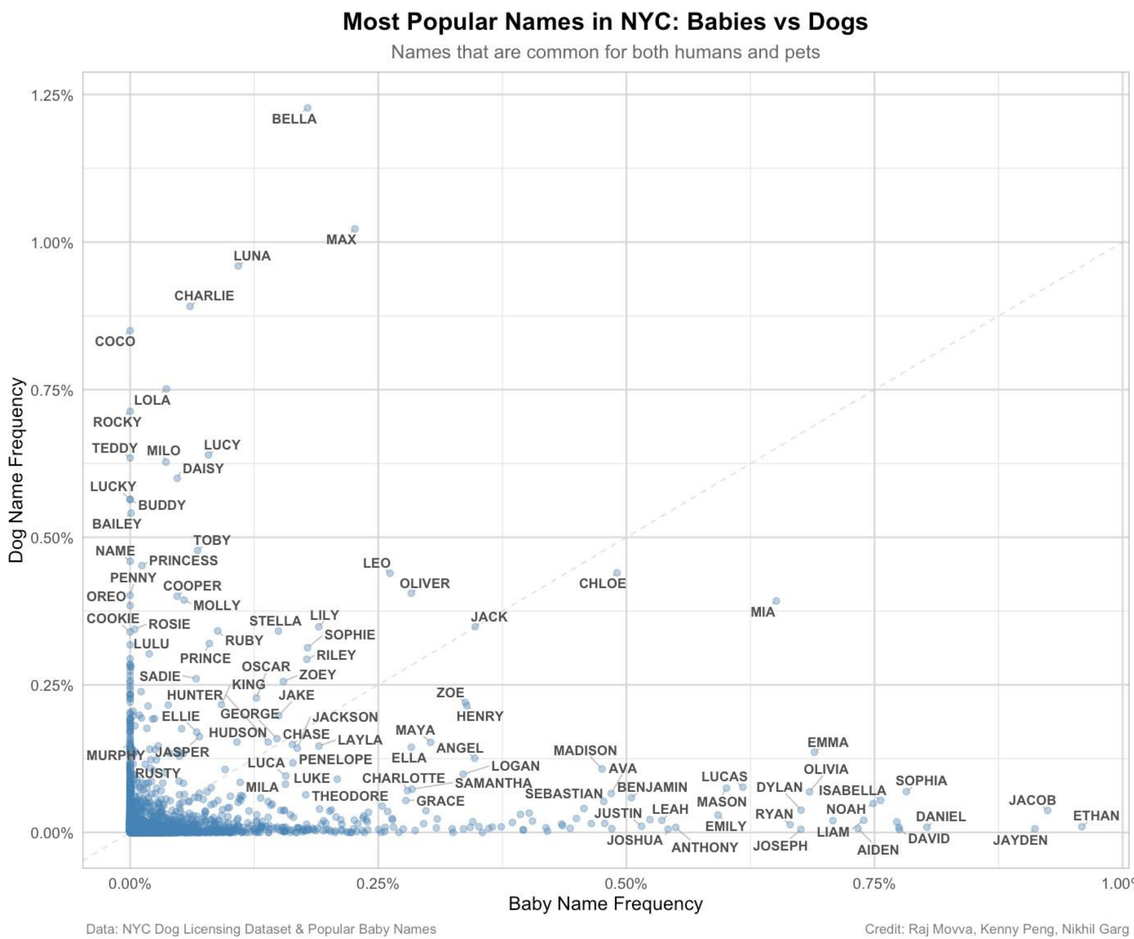


Without the outlier

id	size	price	color	target
1	3	3,29	white	cheap
2	4	24,99	black	exp
3	4	23,59	red	exp
4	3	4,59	black	cheap
5	3	6,99	black	exp
6	5	19,99	red	exp
7	3	24,99	black	exp
8	4	2,99		cheap
9	4	21,99	blue	cheap
10	5	29,99	red	exp
11	6	23,99	red	exp
12	5	8,99	black	cheap
13	5	3389,99	black	exp



Think about creative ways to inspect/visualize your data



Last week

PhD student:
"My method has performance p"

I: "Have you looked at the data?"

PhD student: "No."

I: "Then you do not know whether or why your method behaves poorly or well."

Look at your data!

More data understanding

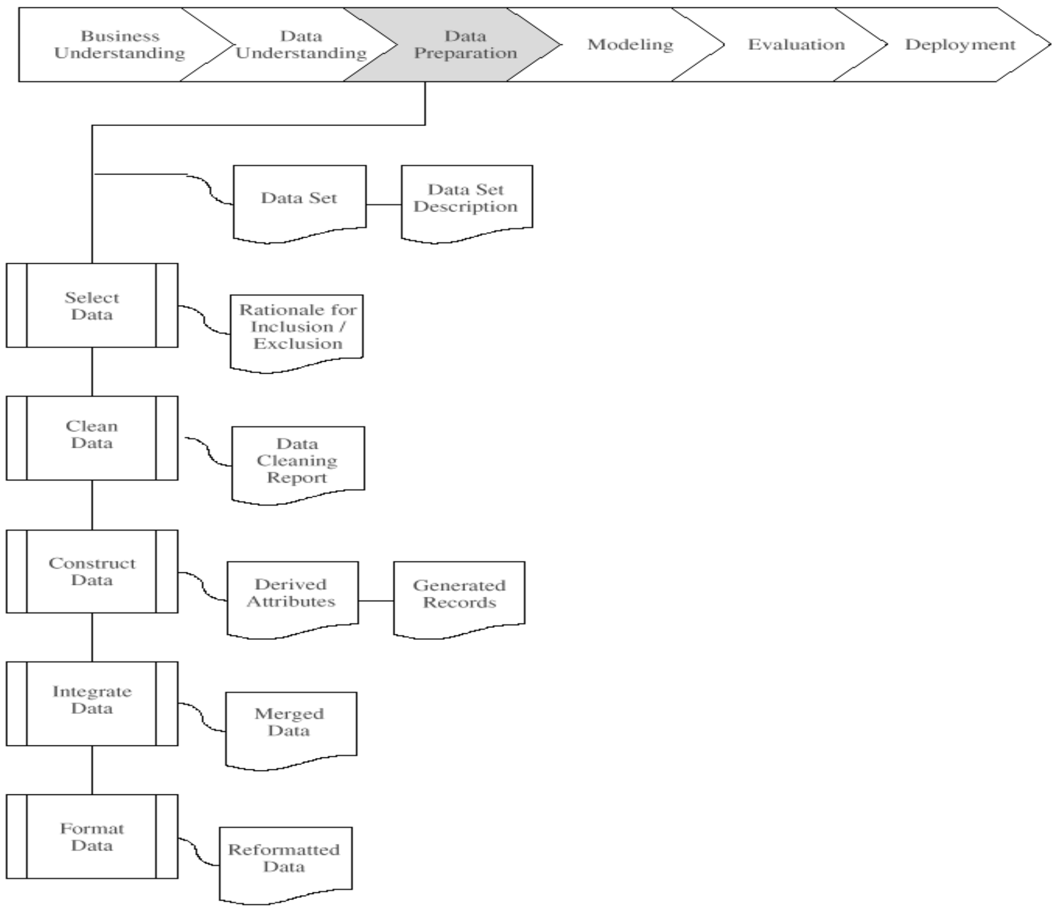
The next slide deck will look at

- Dimensionality reduction
- Clustering

These are pivotal means for data understanding

If you do not understand the data,
you are fully replaceable
by AutoML and/or Large Language Models

Data Preparation



Data Preparation (1/3)

- Select data (feature selection)
  - focus on useful attributes/features
  - not all attriibutes are useful (e.g. IDs)
  - discard a highly correlated attribute,
    e.g., to avoid poor linear regression
- Integrate Data
  - combine data from different sources (record/entity linkage)
  - be aware of syntactic / semantic inconsistencies
    - which units of measurement?
      (e.g. NASA lost spacecraft in 1999 because of metric conversion problem)
  - which reference system (e.g. geocoordinates WGS-84 vs WGS-72)

Data Preparation (2/3)

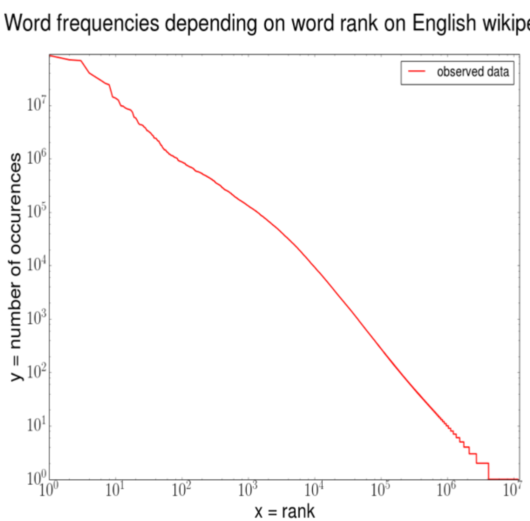
- Clean data
  - correct false values
    - e.g. birthdate field that encodes gender in one bit
    - insert suitable defaults if needed
  - remove outliers
  - estimate missing values: data imputation
    - some ML algorithms can handle missing data (and use the missingness as useful hint) others cannot
- Format data

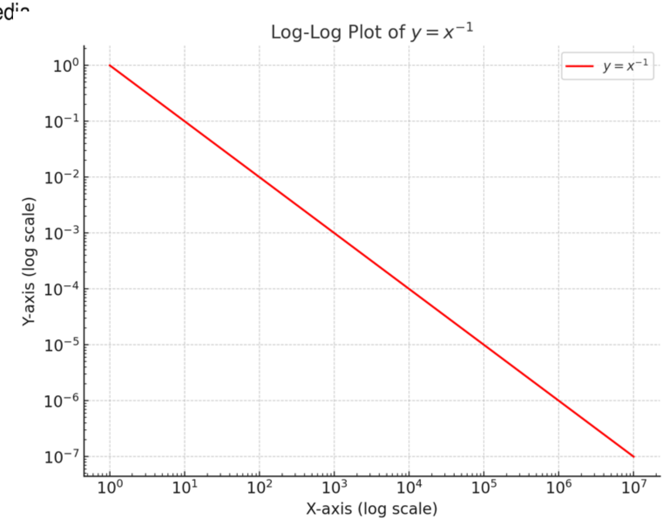
Data Preparation (3/3): Feature engineering

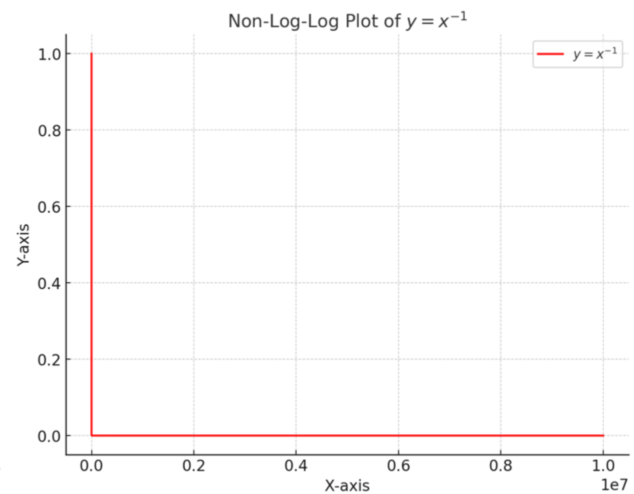
- define features from given attributes
  - e.g. feature distance instead of coordinates $\mathrm{x_d} = \sqrt{x_1^2 + x_2^2}$
  - learn representations
    - representation learning; e.g. what is a fur vs what are scales
- normalize / transform single attributes (if needed)
  - Use domain knowledge to rescale features
    - $x_1 := \log x_1$ or $x_1 := \sqrt{x_1}$
  - convert categorical data to numerical data (e.g. for linear regression)

Assumptions about f

- Machine learning algorithms make assumptions about f and $\mathcal{D}$
  - e.g. linearity
  - to fulfill them, transformation may be necessary, e.g. logarithmize



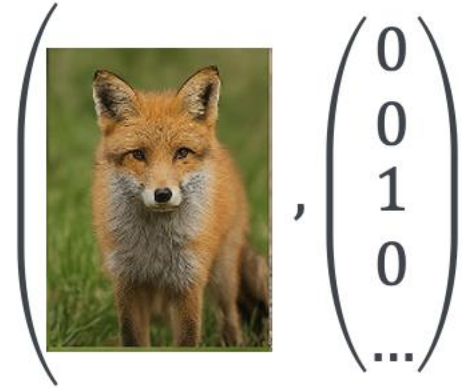




Data Preparation (3/3): Feature engineering

- define features from given attributes
  - e.g. feature distance instead of coordinates $\mathrm{x_d} = \sqrt{x_1^2 + x_2^2}$
  - learn representations
    - representation learning; e.g. what is a fur vs what are scales
- normalize / transform single attributes (if needed)
  - Use domain knowledge to rescale features
    - $x_1 := \log x_1$ or $x_1 := \sqrt{x_1}$
  - convert e.g. numeric data to categorical data (e.g. for decision trees)
  - convert categorical data to numerical data (e.g. for linear regression)

Categorial to numeric: 1-hot encoding



- Given a dataset $\mathcal{D} = \{x_1, \ldots, x_n\} \subseteq X_1 \times X_2 \times \cdots \times X_m$
- Categorical attribute $X_j, 1 \leq j \leq m$ with categories

$$\mathrm{dom}(X_j) = \{a^{(1)}, a^{(2)}, \ldots, a^{(k)}\}$$

- Define $\widehat{X_j} \subseteq \{0,1\}^k$ and

$$x_{\widehat{i,j,l}} = \begin{cases} 1, & \text{if } j = l \\ 0, & \text{else} \end{cases}$$

- Replace all $x_{i,j}$ in $\mathcal{D}$ by $x_{\widehat{i,j}}$

- **issues**: lack of scalability and sparsity issues due to the creation of many orthogonal dimensions

Mougan, Carlos, et al. "Fairness implications of encoding protected categorical attributes."
In: AAAI/ACM Conference on AI, Ethics, and Society. 2023.

Categorial to numeric: Target encoding

- Categorical attribute A with categories $\text{dom}(A) = \{a^{(1)}, a^{(2)}, \dots, a^{(k)}\}$
- Given a target attribute Y with $\text{dom}(Y) = \{0,1\}$
  - (can be generalized to multiple categories)
- Given a dataset $\mathcal{D} \subseteq A \times X \times Y$
- Given observed occurrences

$$\widehat{a^{(i)}} = \frac{|\{(a^{(i)}, .\,, 1) \in \mathcal{D}\}|}{|\{(a^{(i)}, .\,, .) \in \mathcal{D}\}|} \in [0,1]$$

- replace occurrences of $a^{(i)}$ in $\mathcal{D}$ by $\widehat{a^{(i)}}$

- **issues**: overfitting, bias

Mougan, Carlos, et al. "Fairness implications of encoding protected categorical attributes." *In: AAAI/ACM Conference on AI, Ethics, and Society*. 2023.

Data Preparation (3/3): Feature engineering

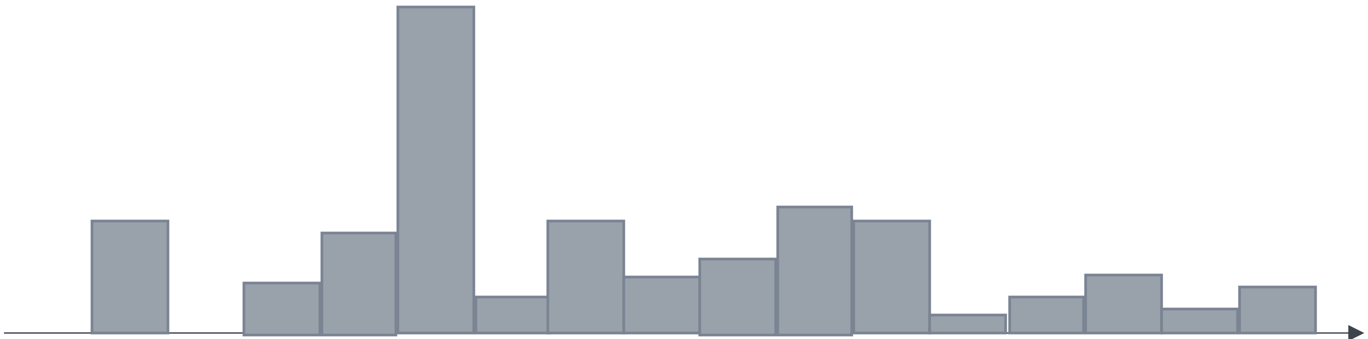
- define features from given attributes
  - e.g. feature distance instead of coordinates $\mathrm{x_d} = \sqrt{x_1^2 + x_2^2}$
  - learn representations
    - representation learning; e.g. what is a fur vs what are scales
- normalize / transform single attributes (if needed)
  - Use domain knowledge to rescale features
    - $x_1 := \log x_1$ or $x_1 := \sqrt{x_1}$
  - convert categorical data to numerical data (e.g. for linear regression)
  - convert e.g. numeric data to categorical data (e.g. for decision trees)

Convert numerical data to nominal data

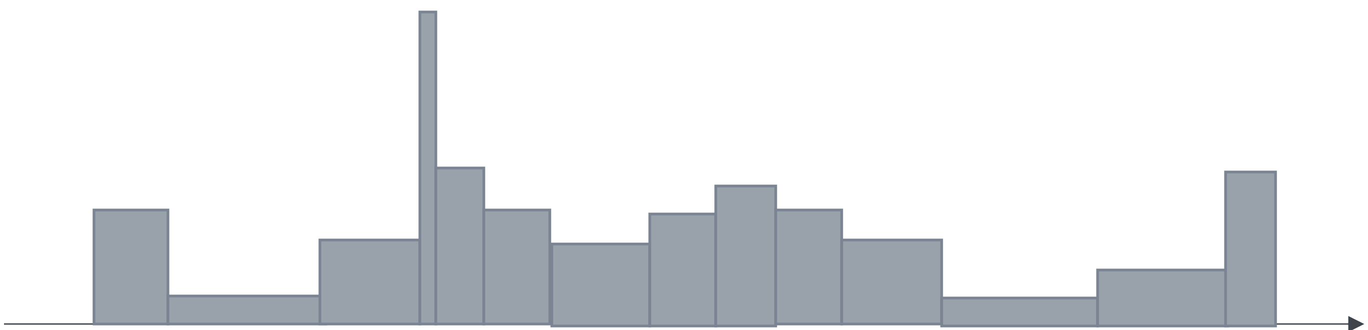
- Given a dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\} \subseteq X \times Y$
- Given $X = \mathbb{R}$
- Define $\hat{X} = \{1, \dots, k\}, k \geq 2$
- Problem
  - Define $\widehat{\mathcal{D}} = \{(\widehat{x_1}, y_1), \dots, (\widehat{x_1}, y_n)\} \subseteq \hat{X} \times Y$ to represent the original learning problem well
- For $k = 2$ take median to decide between bin 1 or 2

Convert numerical data to nominal data

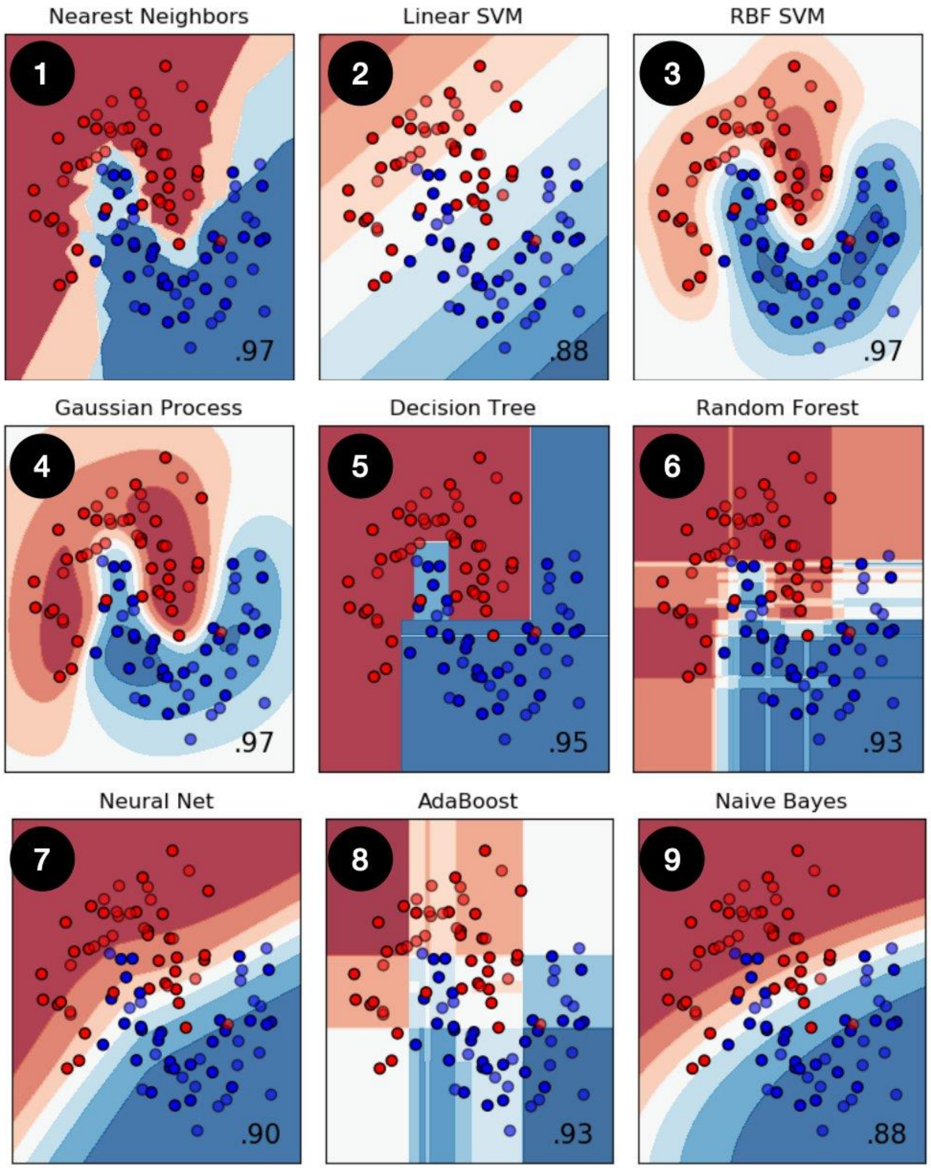
Equi-width vs. Equi-depth Bins



Fixed width for the bins



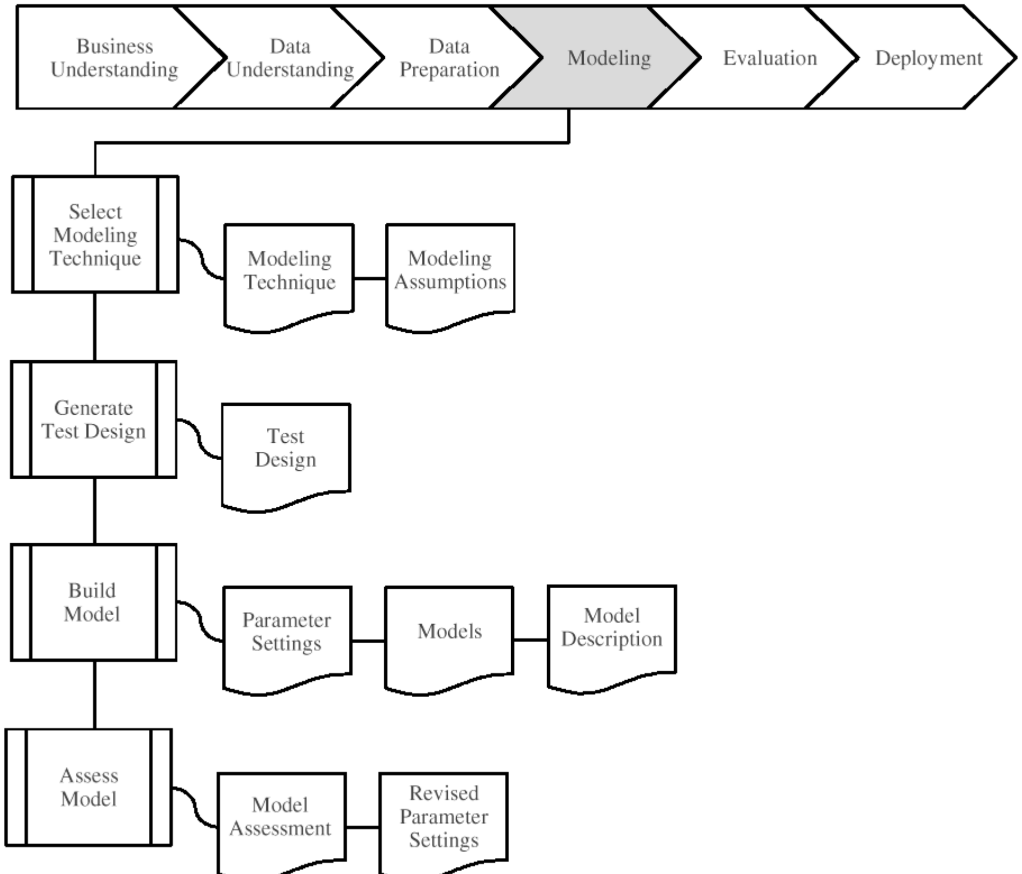
Fixed number of elements per bin (a.k.a. frequency binning)



Different machine learning algorithms have different capabilities and therefore may need different preprocessing

https://twitter.com/seanjtaylor/status/1251043814715711489

Modeling



Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Select Modeling Technique
Modeling Technique
Modeling Assumptions
Generate Test Design
Test Design
Build Model
Parameter Settings
Models
Model Description
Assess Model
Model Assessment
Revised Parameter Settings

Select Modeling technique

Take into account:

- experience with specific techniques
- experience with specific tools
- „political requirements“ (e.g. how explainable; e.g. Schufa must not use neural networks)

Generate **Test Design**

- divide data sets into **training data**, **validation data** and **test data**

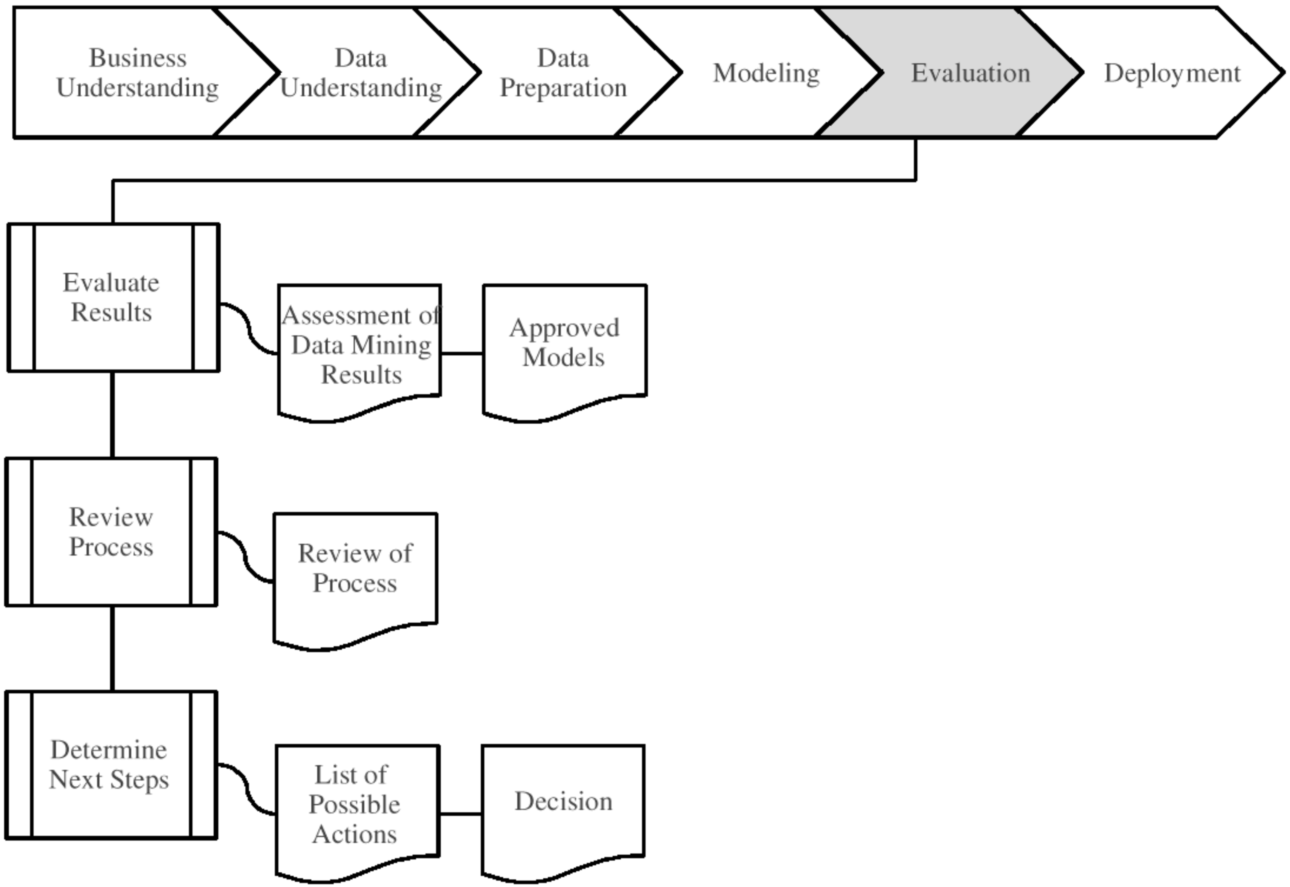
Build Model

- select/explore appropriate **(hyper-)parameter** settings
  (typically, several iterations are needed)

Assess Model

- evaluate results with respect to data mining/machine learning success criteria
- check model against already known knowledge
- revise parameter settings (if needed) and go back to „Build Model“
- rank the generated models with respect to success criteria

Evaluation


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Evaluate Results
Assessment of Data Mining Results
Approved Models
Review Process
Review of Process
Determine Next Steps
List of Possible Actions
Decision

Evaluation

- **Define your evaluation criteria during business understanding**

- **Evaluate Results**
  - evaluate results with respect to business objectives
  - what are other findings of the project
    (e.g. quality of available data should be improved)
  - **your algorithm must not determine your business evaluation**

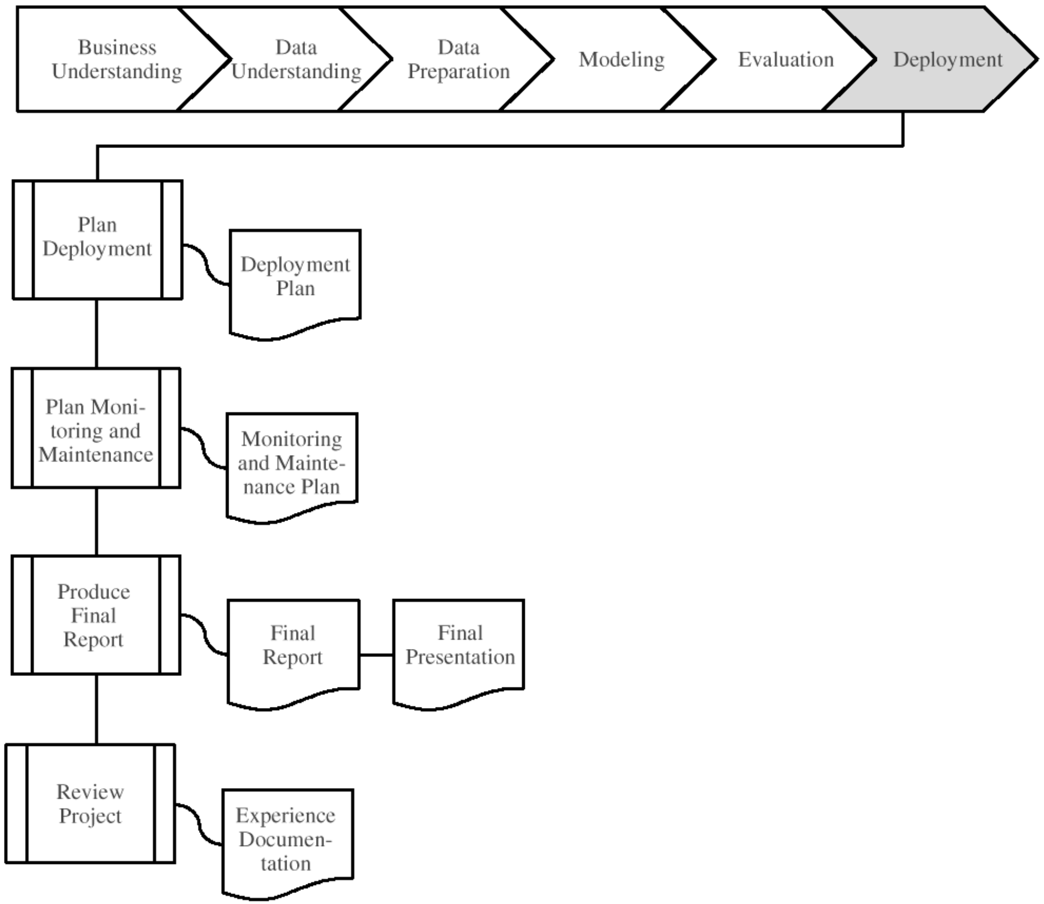
- **Review Process**
  - identify failures

- **Determine Next Steps**
  - analyse potential for „Deployment“

- **Andrew Ng: Machine Learning Yearning**
  - **How to avoid evaluation-deployment mistakes**

  https://home-wordpress.deeplearning.ai/wp-content/uploads/2022/03/andrew-ng-machine-learning-yearning.pdf

Deployment


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Plan Deployment
Deployment Plan
Plan Moni-toring and Maintenance
Monitoring and Mainte-nance Plan
Produce Final Report
Final Report
Final Presentation
Review Project
Experience Documen-tation

Deployment

- **Plan Deployment**
  - set up deployment plan
    - target platform? mobile? sensor quality?
- **Plan Monitoring and Maintenance**
  - when should the model not be used any more?
    - **models become stale much, much faster than you would ever imagine (often weeks or few months!)**
  - will the business objectives change over time?
- **Produce Final Report**
  - what are target groups for final presentations?
- **Review Project**
  - summarize important insights and experiences
  - integrate review results into knowledge management strategy
    - turnaround time matters
    - **if board of directors wants to know today a number, it may not care about it next week**

Additional aspects: data privacy and security

The Application of knowledge discovery/machine learning must not break laws

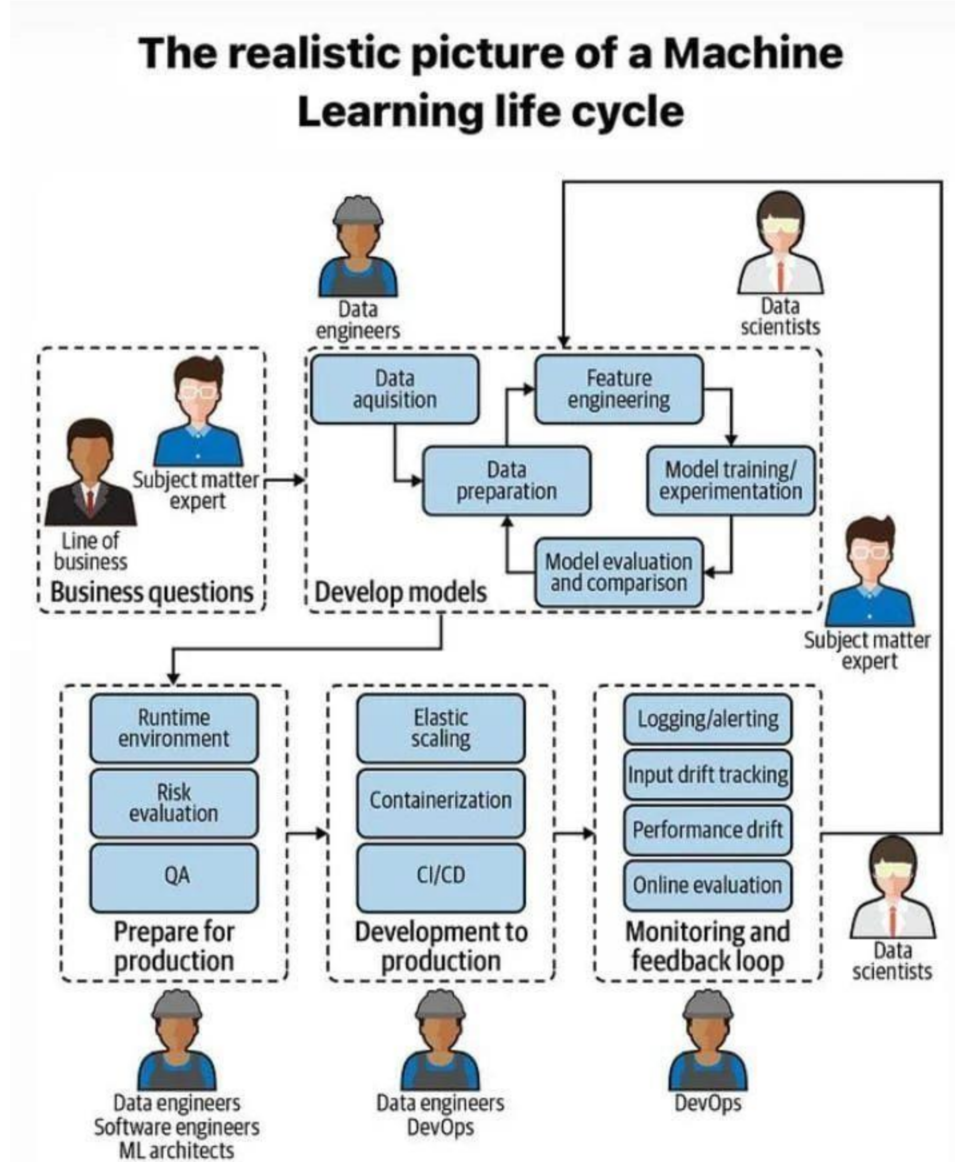
- **GDPR – general data protection regulation**
  - e.g. no usage other than the one needed for the purpose
  - e.g. medical data cannot just be mined because it is available
  - **data privacy is very important while focussing**

    $\Longrightarrow$ the reduction of examples must not allow to draw conclusions on single persons or small groups of persons
      - data must be made anonymous
      - use sufficient number of examples

- **EU AI Act (2024)**
  - Applications with unacceptable risks are banned (e.g. facial recognition in public spaces)
  - High-risk applications must comply with security, transparency and quality obligations, and undergo conformity assessments.
    - e.g. AI used in health or management of critical infrastructure
  - Limited-risk applications only have transparency obligations
    - e.g. inform users about video manipulation software
  - Minimal-risk applications are not regulated
  - General-purpose AI (Foundation Models) have specific rules

Teams Involved in MLOps Process

- **Business Team:**-They give the problem.
- **Domain Expert**:-These people are also from the business but they do have expertise in the domain of that specific problem they are handling now and can provide feedback on the ML system.
- **Data Engineer**:-This team is mostly engaged in extracting, transforming and loading the data. Storing and managing data is their top priority.
- **Data Scientist:**-To create models, preprocess and reengineer the data as required by the model.
- **Devops**:-To deploy models on various platforms.
- **ML Architect**:- The ML Architect works closely with DevOps team to streamline the ML model.
- **Software Engineers**:-To create various integration APIS with another system, front-end design.



Universität Stuttgart
KI

Thank you!

Steffen Staab

E-Mail Steffen.staab@ki.uni-stuttgart.de

Telefon +49 (0) 711 685-88100

www.ki.uni-stuttgart.de/

Universität Stuttgart
Analytic Computing, Institut für Künstliche Intelligenz
Universitätsstraße 32, 50569 Stuttgart