

Monitoring

Deployment

Question

- define the research question
- research relevant background

Evaluation

- critically check the outputs
- for prediction models use metrics and look at the models outputs

Modeling

- choose a model/method
- tune its hyperparameters
- train/test split

Explorative Data Analysis (EDA)

- query your data
- visualize your data
- uncover relationships within the data

Data Collection

- identify relevant data sources
- collect your own data
- pay attention to licenses/usage terms

Data Preprocessing

- clean data (filter, missing values)
- enrich data with other data
- feature engineering
- processed data set

Data Sets

- Overnight Stays in Berlin
- Inside Airbnb
- Knowledge Graph
- Allgäu-Walser-Card

Formats

- Excel files
- Relational Databases (SQL); SQLite specifically
- CSV files
- ...

Schemas


- types (categorical, ...)
- cardinality
- structured data (tables) vs unstructured data (images, text, audio)

Queries

- Select
- Filter
- Aggregate/Summarize
- Sort
- Limit
- Join

Charts

- Bar charts
- Scatter Plots
- Line Charts
- Maps



Tools

- Metabase
- Orange
- (Excel)
- (SPSS)

Advanced Methods

- Classification (kNN, Random Forest, Logistic regression)
- Regression (kNN, Random Forest, Linear Regression)
- Clustering (k-means, hierarchical clustering)
- (Image Embedding)
- Sentiment Analysis
- Topic Modeling
- (Document Embedding)

Strategies

- very likely somebody ran into the same problem
- find a working example to understand the method
- principles carry over, tools change