Hochschule Karlsruhe
University of
Applied Sciences

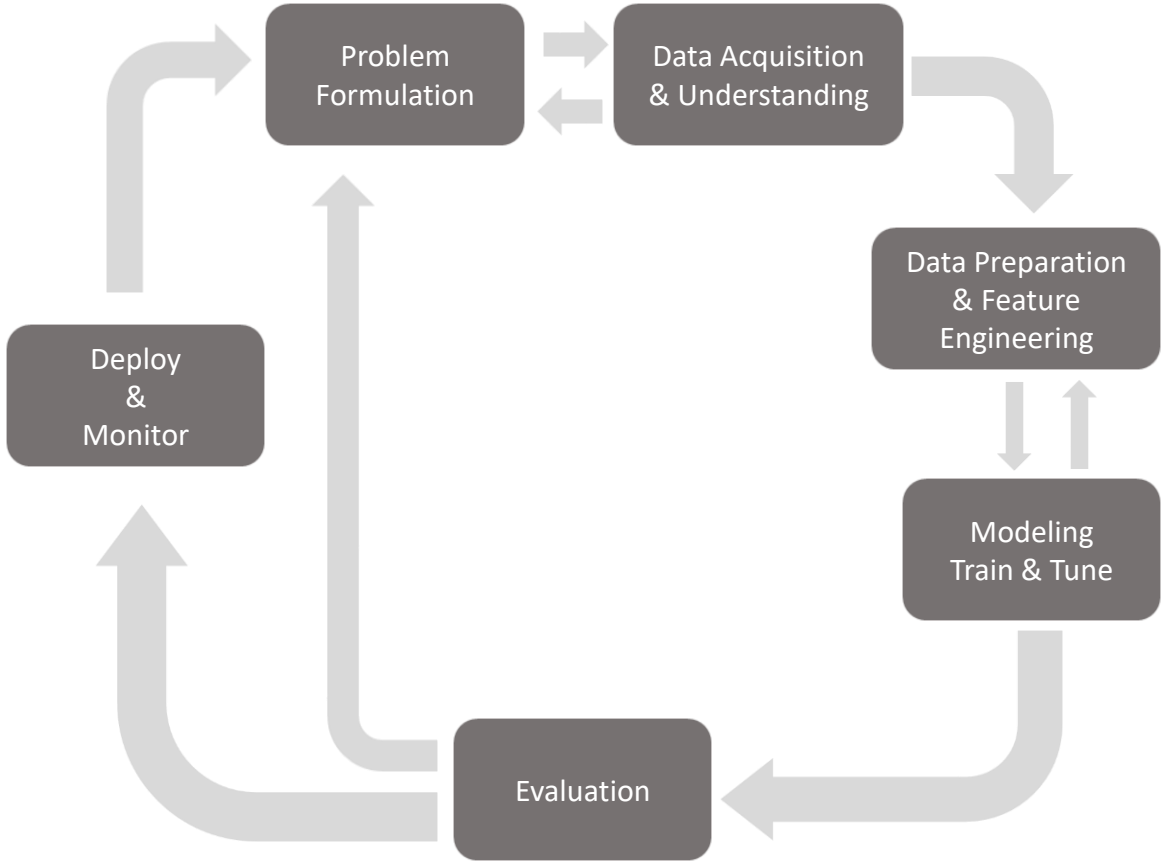
Fakultät für
Elektro- und
Informationstechnik

Data Preparation And Feature Engineering (1)

Source: DALL.E

ML Project Workflow - A Typical Procedure For A Machine Learning Project

+ Projects start with understanding of the business case :
  - What? What for? Where? For whom?
+ Data Acquisition, Data Understanding
+ Data Preparation, Feature Engineering
+ Designing a machine learning model:
  - train, validate, tune, validate , retune…retrain…..
+ Model selection, model evaluation
+ Evaluation:
  - Does the model meet the requirements?
  - Is the business case confirmed?
+ Implementation of the model, model monitoring

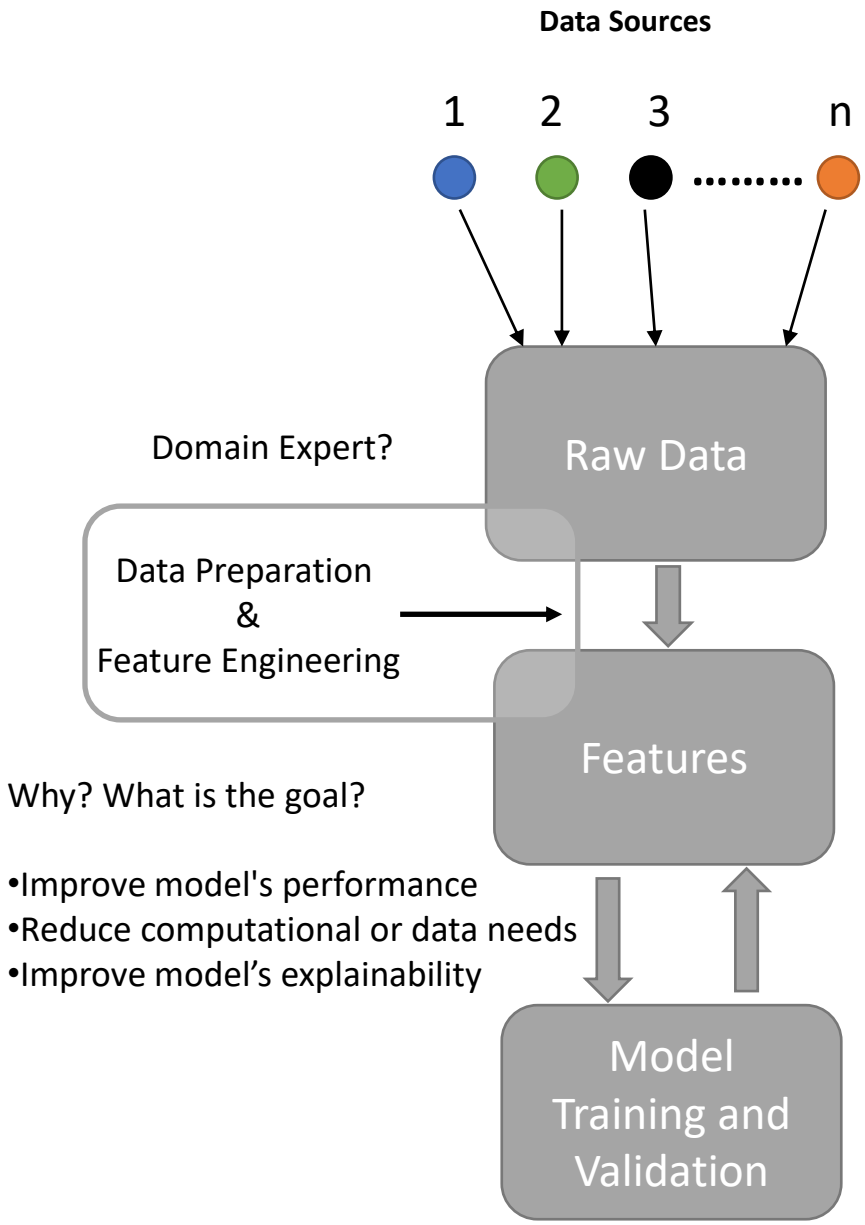


Modified by me: Source: successfactory management coaching gmbh

Data Preparation and Feature Engineering

Data Flow in Machine Learning

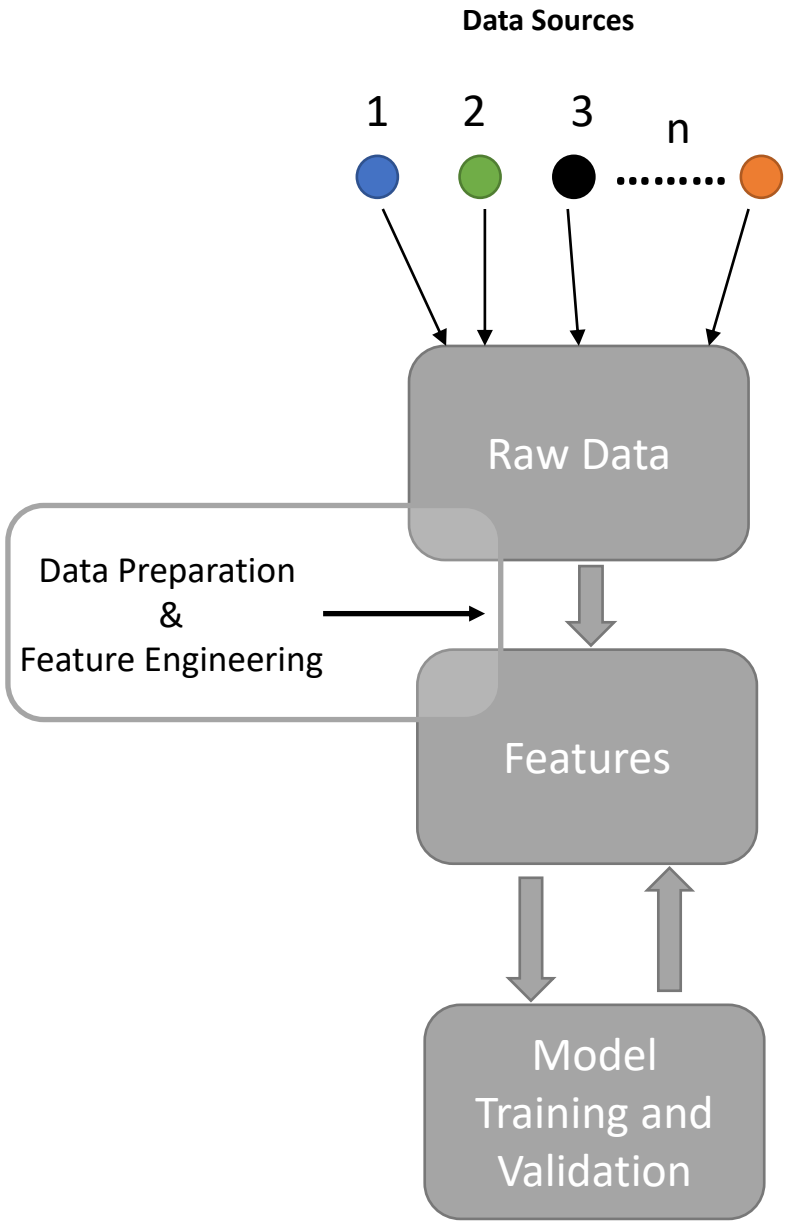
+ Machine learning projects use data generated from different sources, they often have non-uniform formats and structures, requiring preparation before use in machine learning processes.
+ Steps in data preparation:
  - Data exploration
  - Quality assessment/Data cleaning
  - Imputation
  - Feature engineering: (creating the relevant features)
    - Features Selection for ML model
    - Combination, transformation, create a new feature
    - Features encoding
  - After building the models, testing if the selected features achieve the desired outcomes
  - Repeating the preparation process if necessary



Data Preparation and Feature Engineering

Data Flow in Machine Learning

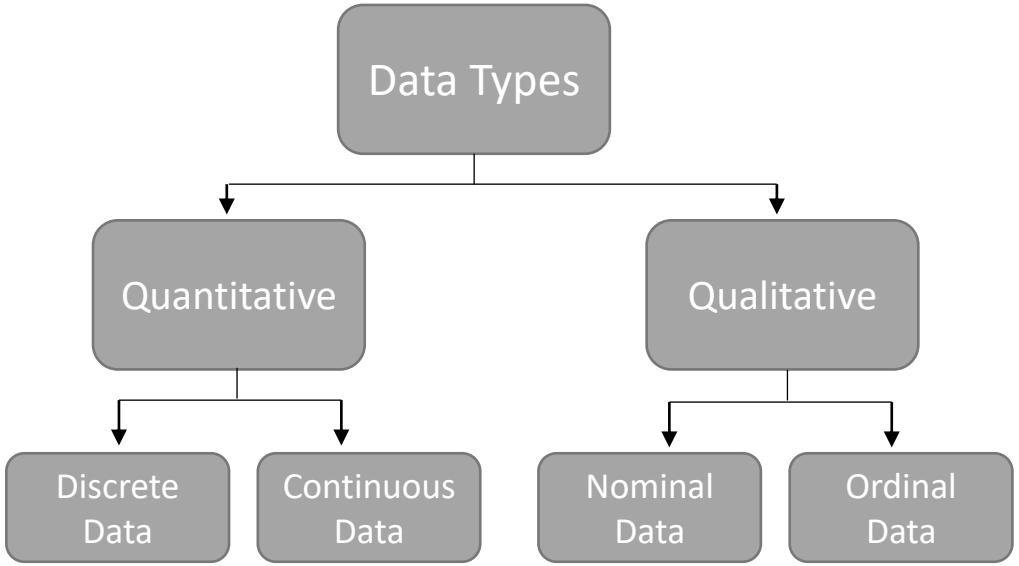
- Machine learning projects use data generated from different sources, they often have non-uniform formats and structures, requiring preparation before use in machine learning processes.
- Steps in data preparation:
  - Data exploration
  - Quality assessment/Data cleaning
  - Imputation
  - Feature engineering: (creating the relevant features)
    - Features Selection for ML model
    - Combination, transformation, create a new feature
    - Features encoding
  - After building the models, testing if the selected features achieve the desired outcomes
  - Repeating the preparation process if necessary



Data Preparation and Feature Engineering

Exploratory Data Analysis (Data Understanding)- Data Types

- Quantitative or numerical data can be expressed in numerical values, making it countable and suitable for statistical analysis.
  - Continuous data: can take any value within a range.
    Examples: temperatures, electrical voltages and currents.
  - Discerete data: can take only a finite number of values.
    Examples: a dice roll (1 2 3 4 5 6), number of students
- Qualitative or Categorical data can't be measured or counted in the form of numbers.
  - Ordinal data: has a natural order or ranking.
    Example: Education Level (Higher, Secondary, Primary)
  - Nominal data: Categories with no specific order.
    Example: Blood Type → (A, B, AB, O)



Data Preparation and Feature Engineering

Exploratory Data Analysis (Data Understanding) – Data Representation

- Exploratory data analysis is like getting to understand the dataset better. This early step will help to gain valuable insights for data cleaning and ideas for feature engineering.
- Interpretation of data in detail is time-consuming. Clear presentation and concise summaries are necessary.
- Data representation
  - Display in tabular form
  - Graphical representation that depends on data type

Heart Failure Prediction Dataset

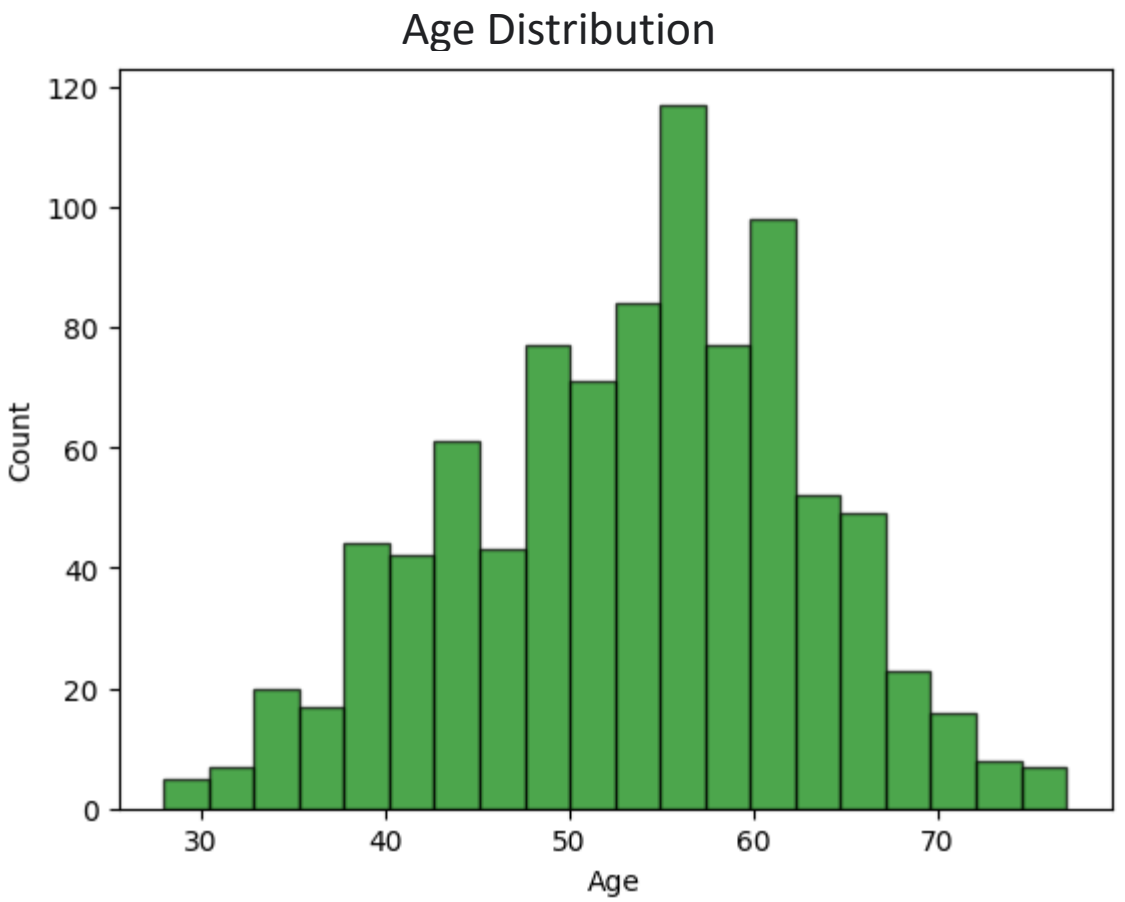
Age	Gender	RestingBP	Cholesterol	HeartDisease
40	M	140	289	0
49	F	160	180	1
37	M	130	283	0
48	F	138	214	1
54	M	150	195	0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Gender          918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

Data Preparation and Feature Engineering

Exploratory Data Analysis (Data Understanding) – Data Representation

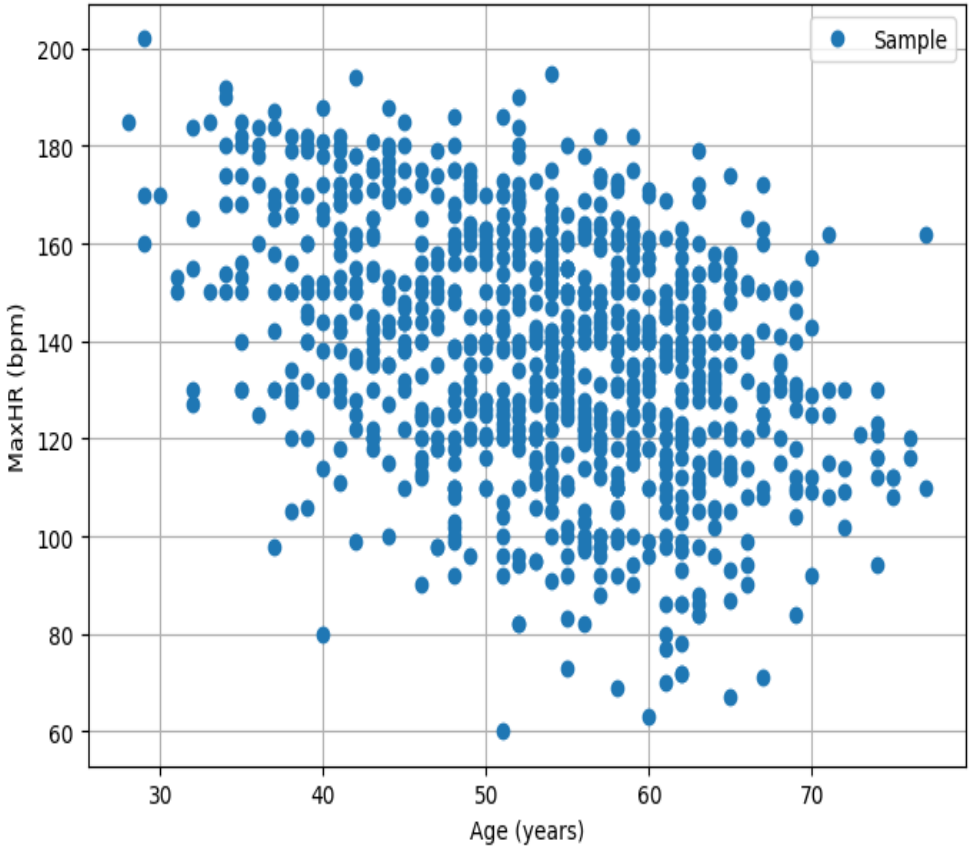
+ Graphical representation of data depends on data types
  - Numerical data
    - Histogramms



Data Preparation and Feature Engineering

Exploratory Data Analysis (Data Understanding) – Data Representation

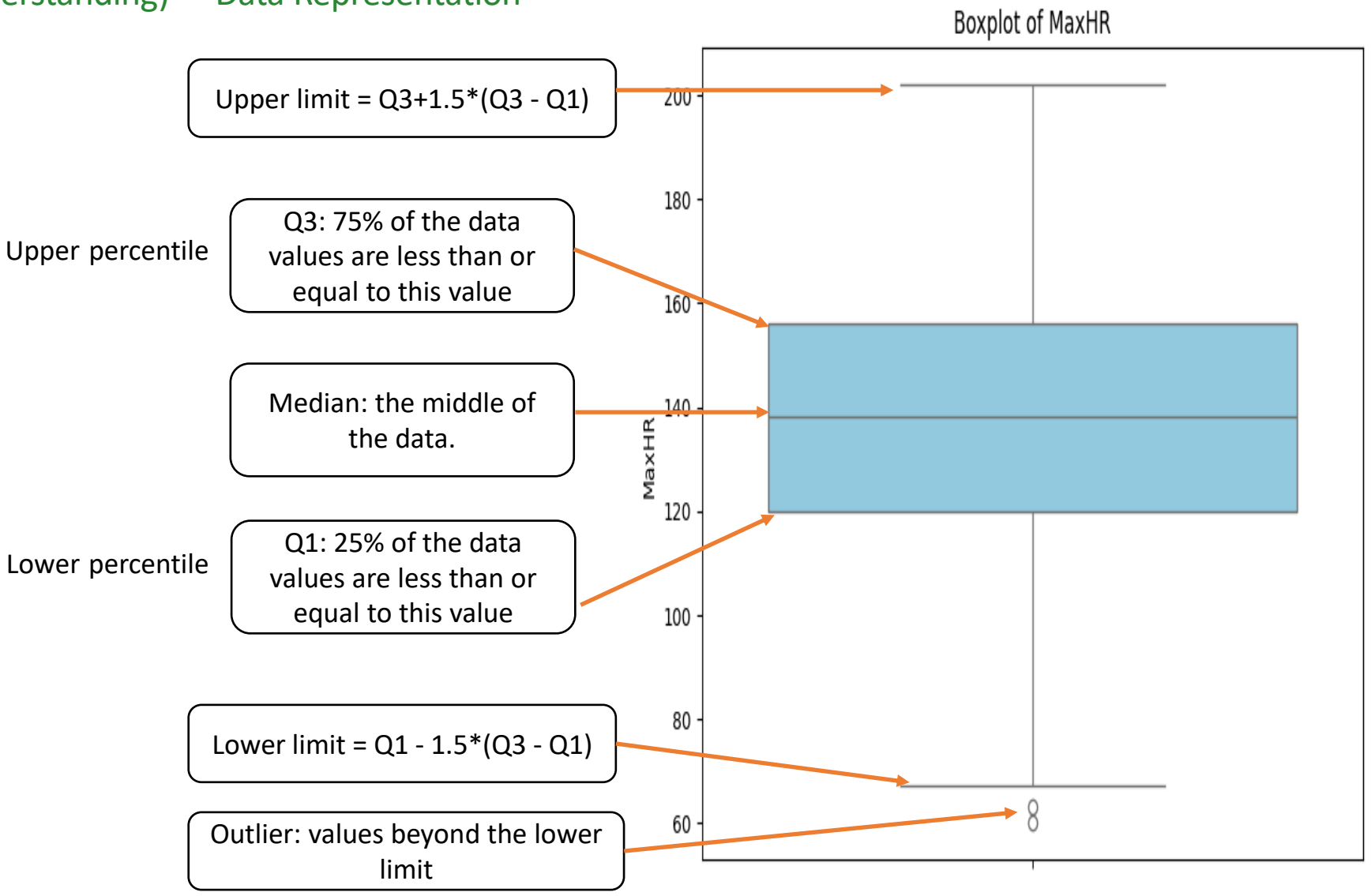
+ Graphical representation of data depends on data types
  - Numerical data
    - Histogramms
    - Scatter plot



Data Preparation and Feature Engineering

Exploratory Data Analysis (Data Understanding) – Data Representation

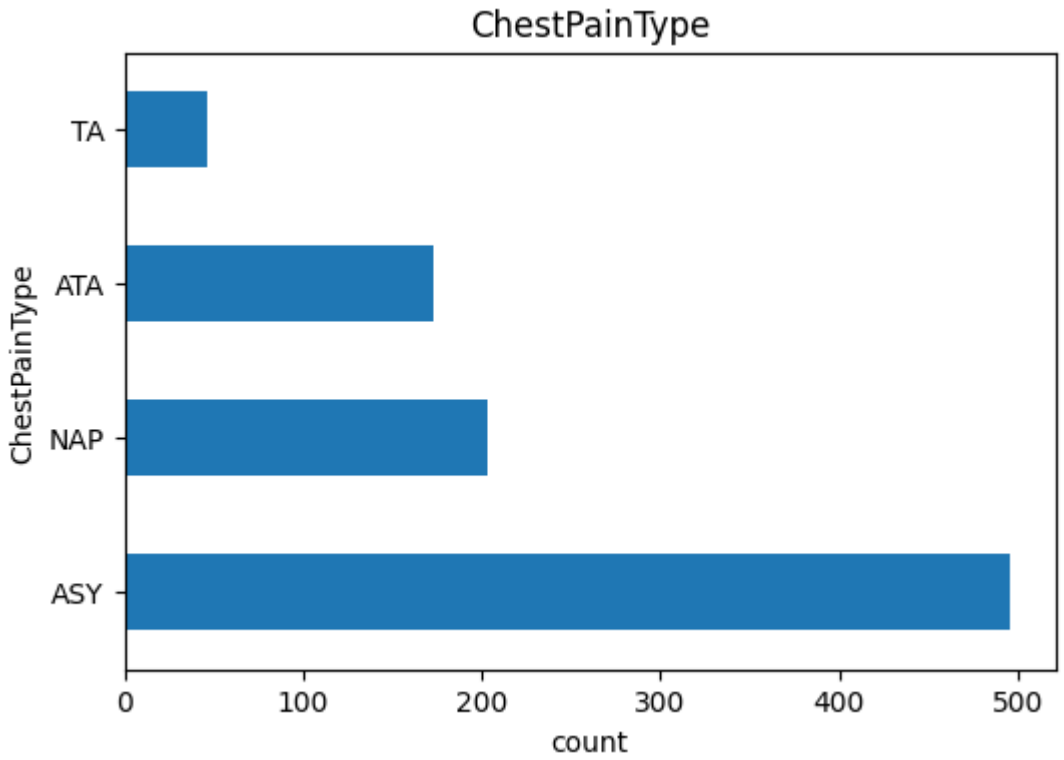
+ Graphical representations
  - Numerical data
    - Histogramms
    - Scatter plot
    - Box-plots



Data Preparation and Feature Engineering

Exploratory Data Analysis (Data Understanding) – Data Representation

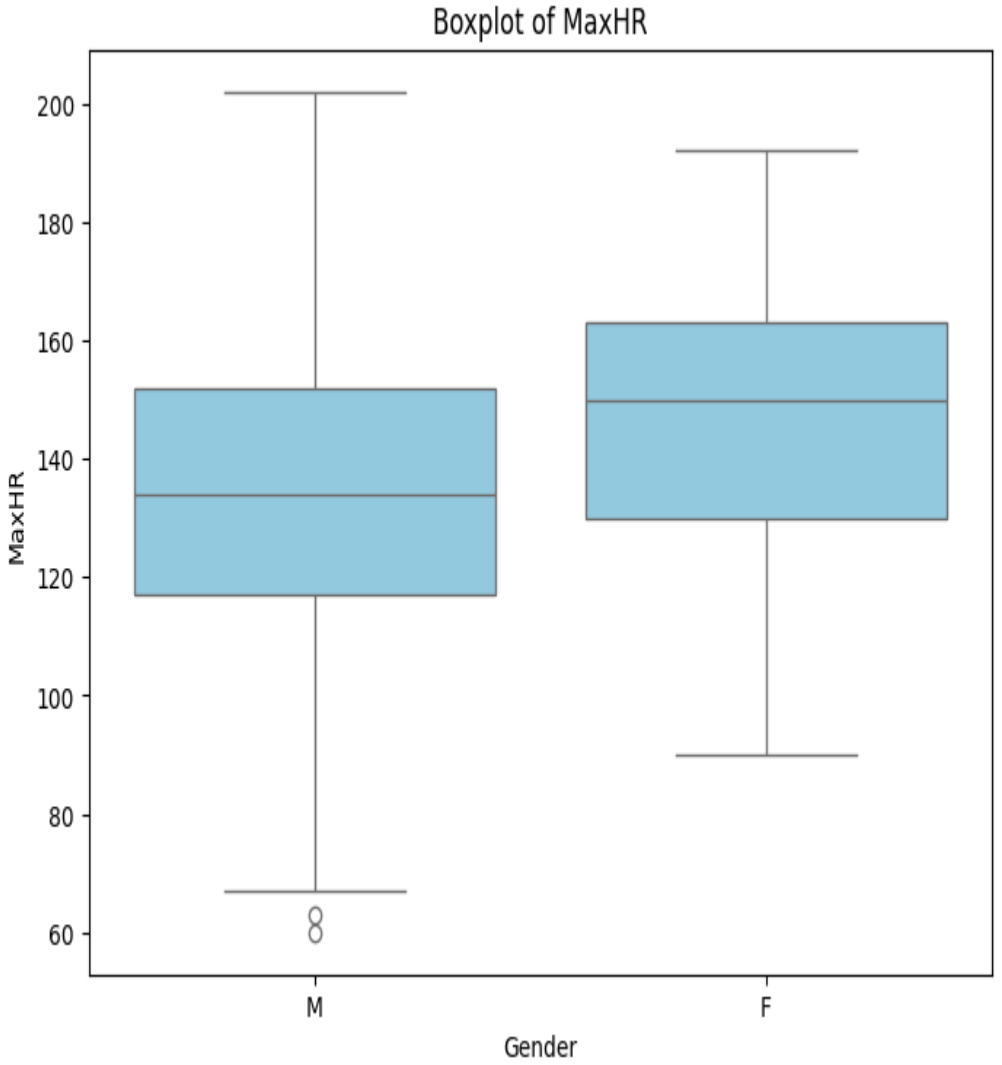
+ Graphical representations
  - Numerical data
    - Histogramms
    - Box-plots
  - Categorical data
    - Bar plots



Data Preparation and Feature Engineering

Exploratory Data Analysis (Data Understanding) – Data Representation

+ Graphical representations
  - Numerical data
    - Histogramms
    - Box-plots
  - Categorical data
    - Bar plots
    - Segmentation : relationship between categorical features and numeric features



Data Preparation and Feature Engineering

Exploratory Data Analysis (Data Understanding) – Descriptive Statistics

+ Descriptive statistics
  - For numerical data
    - Count, mean value and standard deviation
    - minimum and maximum values
    - median, lower percentile and upper percentile
  - For categorical data
    - count
    - Unique, top (the most common category)
    - Frequancy of the most common category

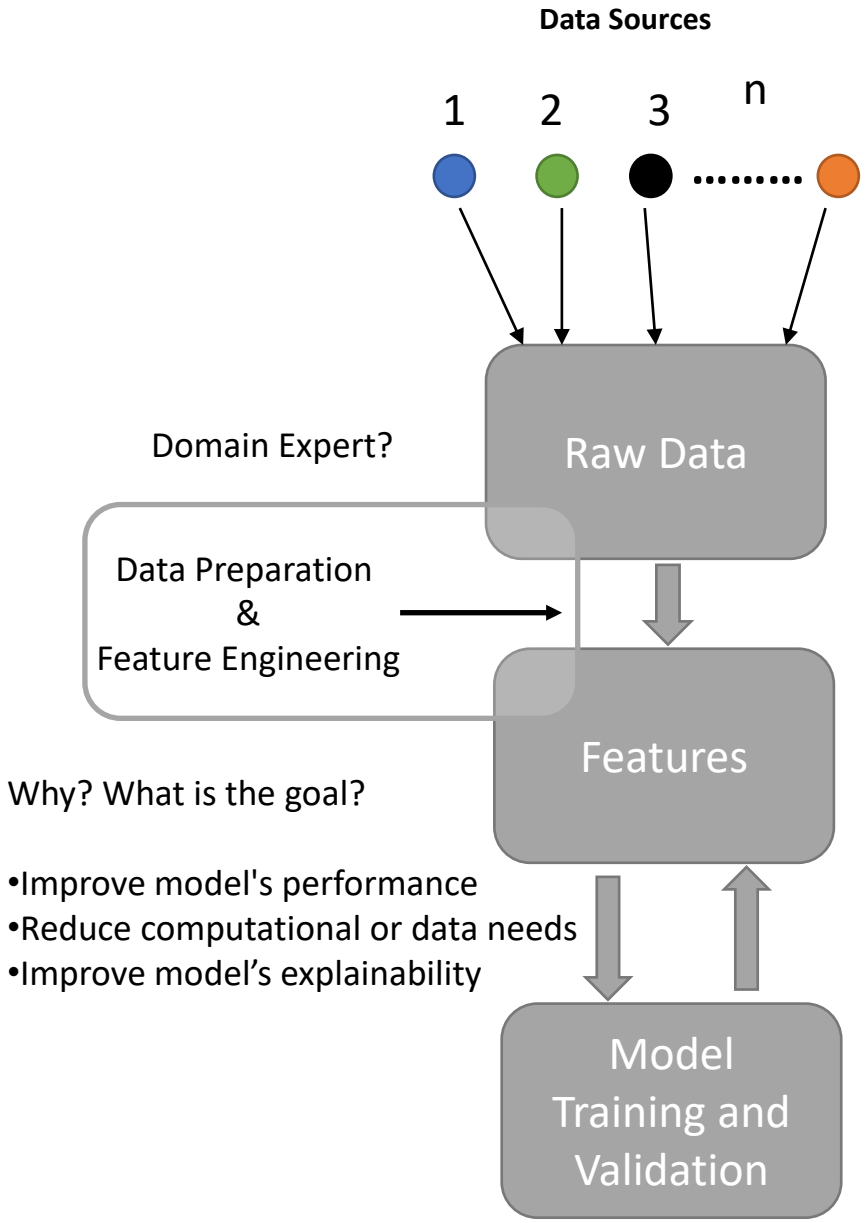
	Age	RestingBP	Cholesterol	FastingBS	MaxHR	HeartDisease
count	918.00	918.00	918.00	918.00	918.00	918.00
mean	53.51	132.40	198.80	136.	136.81	0.55
std	9.430	18.51	109.38	0.42	25.46	0.50
min	28.00	0.00	0.00	0.00	60.00	0.00
25%	47.00	120.00	173.25	0.00	120.00	0.00
50%	54.00	130.00	223.00	0.00	138.00	1.00
75%	60.00	140.00	267.00	0.00	156.00	1.00
max	77.00	200.00	603.00	1.00	202.00	1.00

Name: Gender, dtype: object		
count	918	
unique	2	
top	M	
freq	725	

Data Preparation and Feature Engineering

Data Flow in Machine Learning

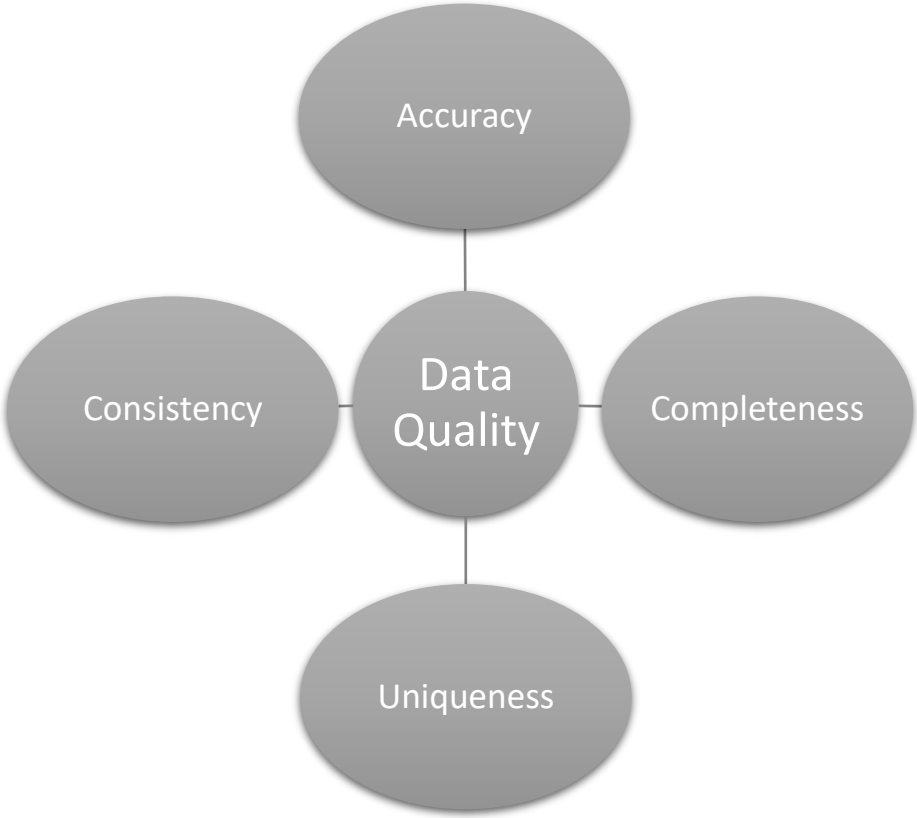
- Machine learning projects use data generated from different sources, they often have non-uniform formats and structures, requiring preparation before use in machine learning processes.
- Steps in data preparation:
  - Data exploration
  - Quality assessment/Data cleaning
  - Imputation
  - Feature engineering: (creating the relevant features)
    - Features Selection for ML model
    - Combination, transformation, create a new feature
    - Features encoding
  - After building the models, testing if the selected features achieve the desired outcomes
  - Repeating the preparation process if necessary



Data Preparation and Feature Engineering

Quality Assessment

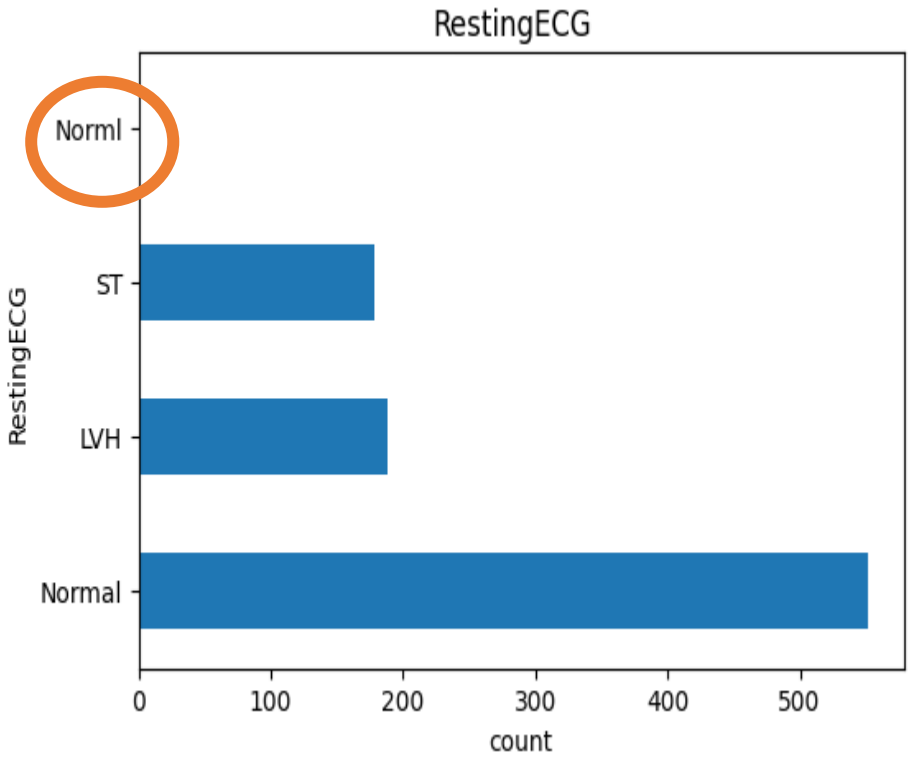
+ Data quality: refers to whether the data is good enough to support its intended use.
+ Data quality main dimensions:
  - Accuracy: the closeness of the values in the data to the real values
  - Completeness
  - Uniqueness
  - Timeliness
  - Consistency



Data Preparation and Feature Engineering

Quality Assessment – Accuracy

+ Data quality: refers to whether the data is good enough to support its intended use.
+ Data quality main dimensions:
  - Accuracy: The closeness of the values in the data to the real values
    - Syntactic Accuracy: Issues with format or structure of the data
    - Semantic Accuracy
  - Completeness
  - Uniqueness
  - Timeliness
  - Consistency

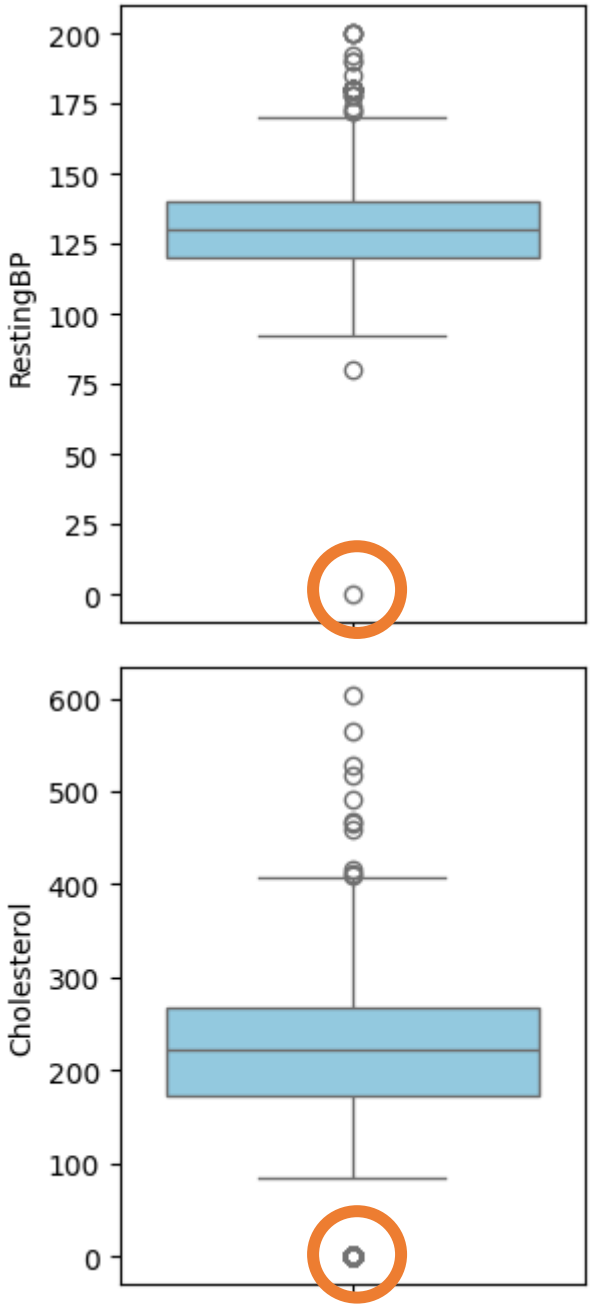


Age	Gender	RestingBP	Cholesterol	HeartDisease
40	M	140	289	0.0
49	F	160	180	1.01
37	M	130	283	0.0
48	F	138	214	1.0
54	M	150	195	0.0

Data Preparation and Feature Engineering

Quality Assessment – Accuracy

- Data quality: refers to whether the data is good enough to support its intended use.
- Data quality main dimensions:
  - Accuracy: The closeness of the values in the data to the real values
    - Syntactic Accuracy: issues with format or structure of the data
    - Semantic Accuracy: Issues with meaning or logic conveyed by the data
  - Completeness
  - Uniqueness
  - Timeliness
  - Consistency

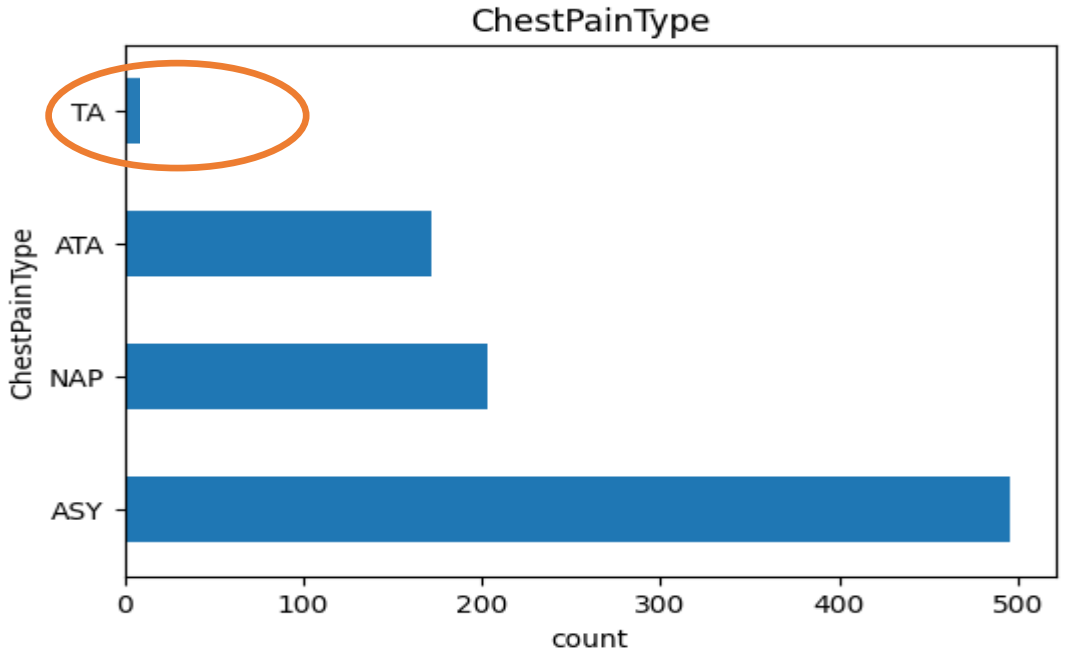


Data Preparation and Feature Engineering

Quality Assessment – Completeness

+ Data quality: refers to whether the data is good enough to support its intended use.
+ Data quality main dimensions:
  - Accuracy: The closeness of the values in the data to the real values
    - Syntactic Accuracy
    - Semantic Accuracy
  - Completeness
    - With respect to attribute values (missing values)
    - With respect to records (are all combinations covered?)
  - Uniqueness
  - Timeliness
  - Consistency

Age	Gender	RestingECG	MaxHR	HeartDisease
54	F	Normal	nan	0.0
38	M	nan	166.0	1.0
43	F	nan	165.0	0.0
60	M	nan	125.0	1.0
36	M	Normal	nan	1.0



Data Preparation and Feature Engineering

Quality Assessment – Uniqueness

- Data quality: refers to whether the data is good enough to support its intended use.
- Data quality main dimensions:
  - Accuracy: The closeness of the values in the data to the real values
    - Syntactic Accuracy
    - Semantic Accuracy
  - Completeness
    - With respect to attribute values (missing values)
    - With respect to records (are all combinations covered?)
  - Uniqueness: Checking for duplicate data entries.
  - Timeliness: Assesses if the data is current enough to reflect the latest information
  - Consistency

Age	Gender	RestingECG	MaxHR	HeartDisease
54	F	Normal	nan	0.0
38	M	nan	166.0	1.0
43	F	nan	165.0	0.0
60	M	nan	125.0	1.0
36	M	Normal	nan	1.0
36	M	Normal	nan	1.0

Data Preparation and Feature Engineering

Quality Assessment – Timeliness

- Data quality: refers to whether the data is good enough to support its intended use.
- Data quality main dimensions:
  - Accuracy: the closeness of the values in the data to the real values
    - Syntactic Accuracy
    - Semantic Accuracy
  - Completeness
    - with respect to attribute values (missing values)
    - with respect to records (are all combinations covered?)
  - Uniqueness: measures the extent of duplication
  - Timeliness: Assesses if the data is current enough to reflect the latest information
  - Consistency

Data Preparation and Feature Engineering

Quality Assessment – Consistency

- Data quality: refers to whether the data is good enough to support its intended use.
- Data quality main dimensions:
  - Accuracy: the closeness of the values in the data to the real values
    - Syntactic Accuracy
    - Semantic Accuracy
  - Completeness
    - with respect to attribute values (missing values)
    - with respect to records (are all combinations covered?)
  - Uniqueness: measures the extent of duplication
  - Timeliness: Assesses if the data is current enough to reflect the latest information
  - Consistency: It is about how uniform your dataset is throughout different data sources. Data sources do not have to be from the same operating system or come from the same countries.

Data Preparation and Feature Engineering

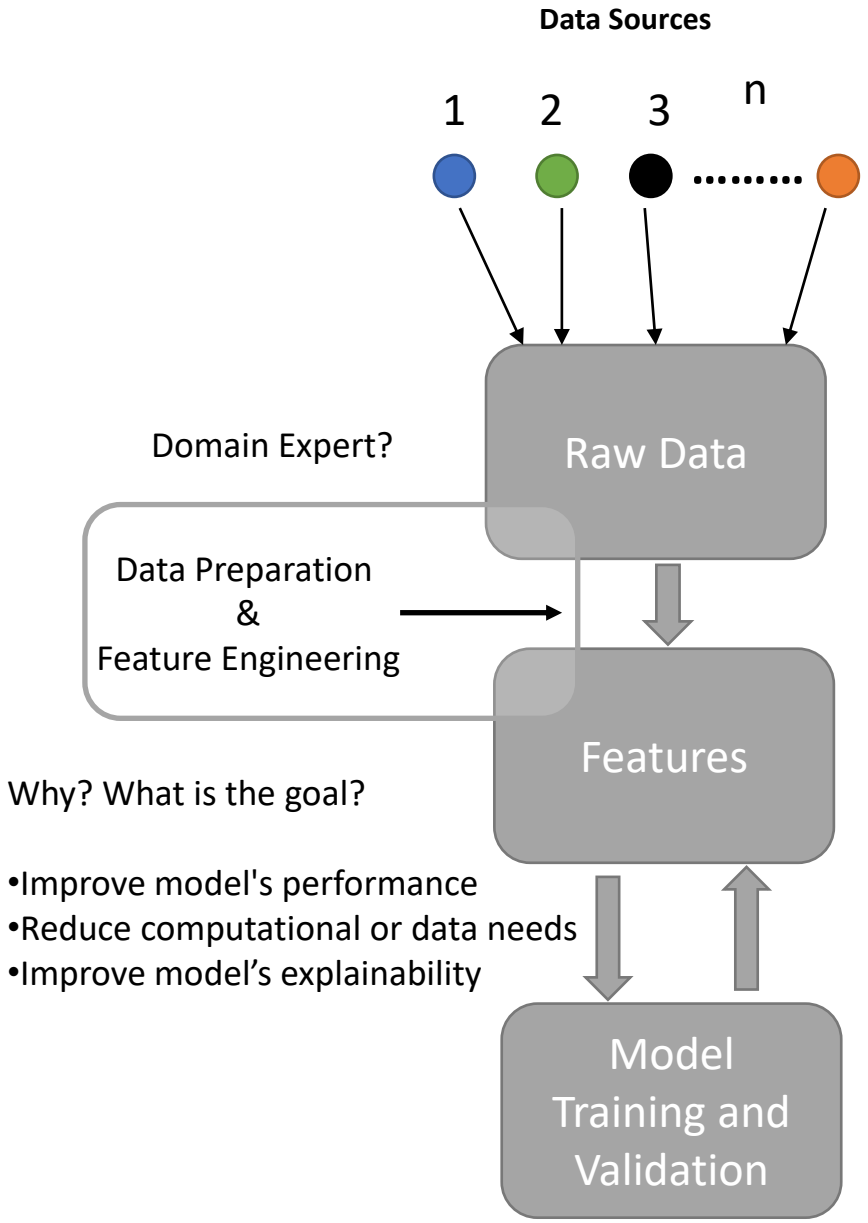
Data Cleaning

- Data cleaning improves data quality by:
  - Remove duplicated records
  - Remove zero variance columns that indicate that all of their values are identical and the irrelevant columns
  - Fix syntactic and semantic errors
  - Remove outliers
  - Unify numbers formats
    - Uniform number formats with identical decimal separator and digit grouping
  - Elimination of detected discrepancies
  - Format string labels consistently
    - Use consistent expressions for the same facts
  - Time and date information
    - Unifying date and time formats across different country specifications
  - Treat missing values (remove or impute)

Data Preparation and Feature Engineering

Data Flow in Machine Learning

- Machine learning projects use data generated from different sources, they often have non-uniform formats and structures, requiring preparation before use in machine learning processes.
- Steps in data preparation:
  - Data exploration
  - Quality assessment/Data cleaning
  - Imputation
  - Feature engineering: (creating the relevant features)
    - Features Selection for ML model
    - Combination, transformation, create a new feature
    - Features encoding
  - After building the models, testing if the selected features achieve the desired outcomes
  - Repeating the preparation process if necessary



Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Elektro- und
Informationstechnik

www.h-ka.de

Manfred Strohrmann