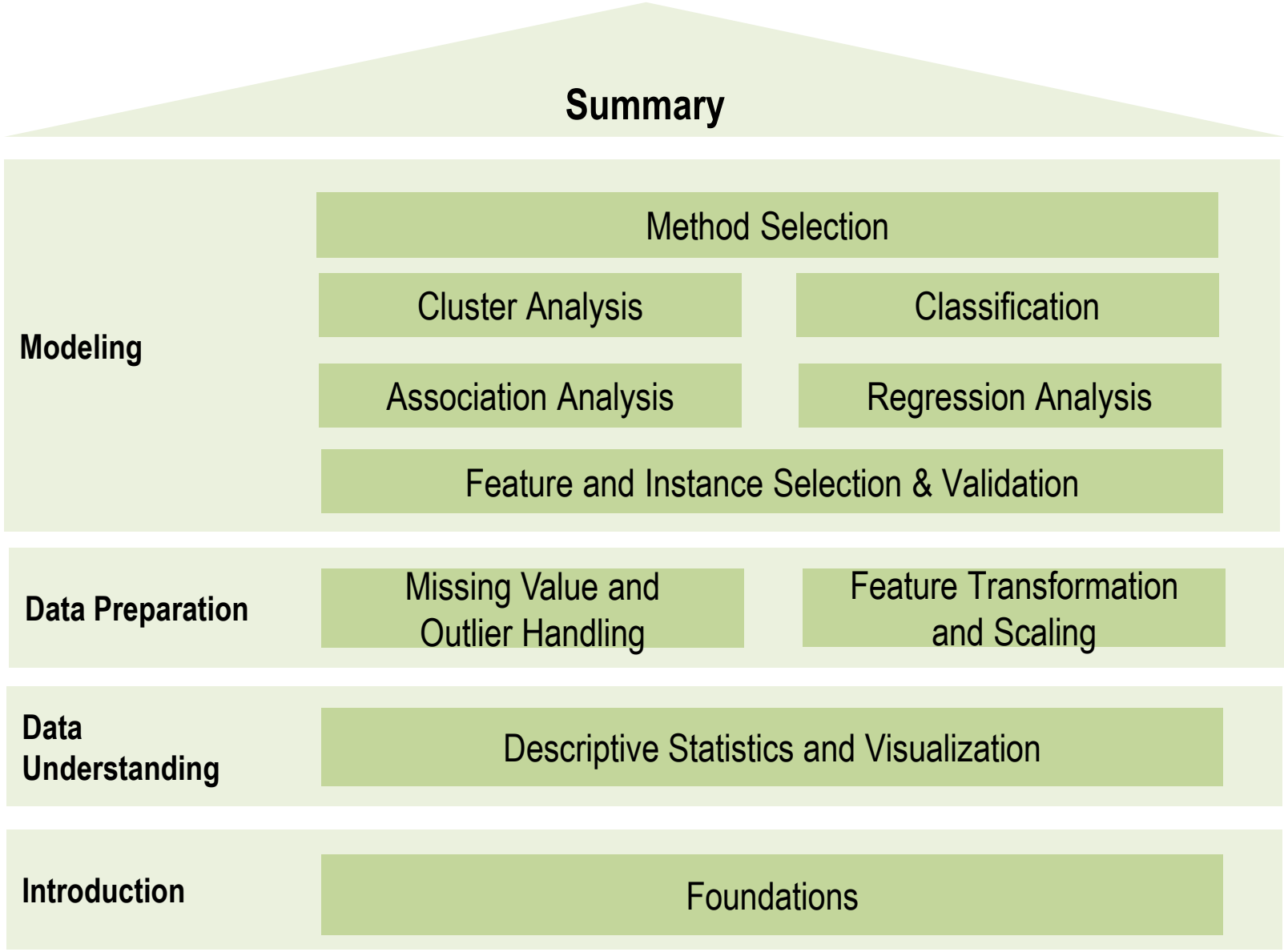
Introduction to Data Science

Lecture Data Science Process

Prof. Dr. Hendrik Meth
HdM, Stuttgart

Lectures: Overview



HOCHSCHULE DER MEDIEN

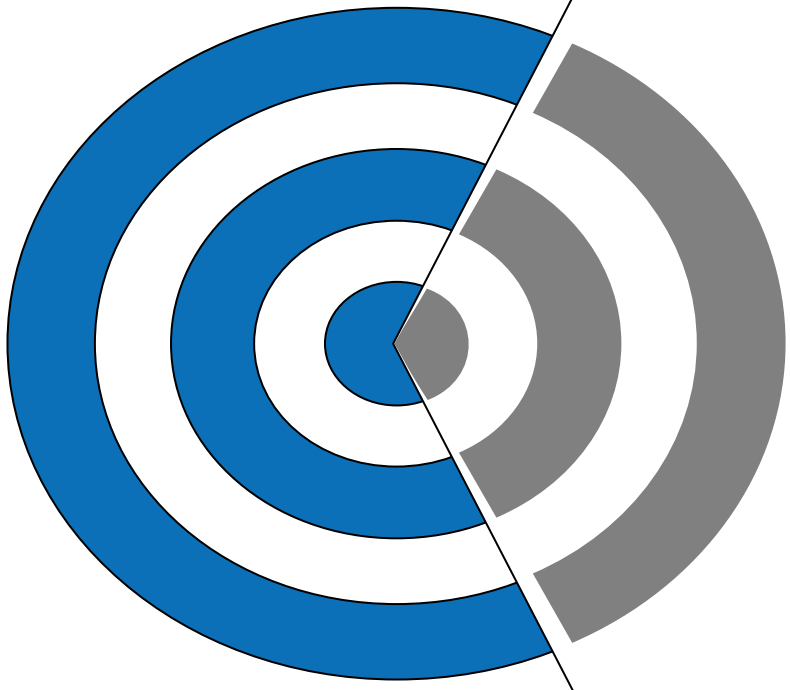
Schedule

	s202	In class	At home
1	13.10.2023	L: Course Organization & Introduction	T: PANDAS1
2	20.10.2023	L: Data Science Process & Case Introduction & PANDAS 2	T: PANDAS 3
3	27.10.2023	L: Data Exploration & PANDAS Exploration	
4	03.11.2023	L: Data Preparation - Transformation & Scaling	T: Data Preparation - Transformation & Scaling
5	10.11.2023	L: Data Preparation - Missing Values & Outliers	T: Data Preparation - Missing Values & Outliers
6	17.11.2023	**P: Data Exploration**	
7	24.11.2023	L: Data Preparation- Feature & Instance Selection	T: Data Preparation- Feature & Instance Selection
8	01.12.2023	L: Regression	T: Regression
9	08.12.2023	L: Clustering	T: Build and Evaluate a Clustering Model
10	15.12.2023	L: Association Rules	T: Derive and Evaluate Association Rules
11	22.12.2023	**P: Data Preparation**	
	29.12.2023	Christmas Break	
	05.01.2024	Christmas Break	
12	12.01.2024	L: Classification – Decision Trees	T: Build and Evaluate a DT (Ensemble) Model
13	19.01.2024	L: Classification – Log. Regression, Naive Bayes, KNN	T: Build and Evaluate a LogRegression and NaiveBayes Model
14	26.01.2024	**P: Model Optimization**	

HOCHSCHULE DER MEDIEN

L = Lecture / T = Technology Lab / P = Presentation

Goals of this (and the next) lecture



- Understand motivation to apply a structured process in data science
- Learn about the process and details of each phase
- Get some first practice with PANDAS

HOCHSCHULE DER MEDIEN

Agenda

	Agenda
1	Project Description
2	Introduction
3	CRISP DM – Phases
4	PANDAS lab

HOCHSCHULE DER MEDIEN

Course Grading

Element	Description	Exam / Due Date

Project (4 teams of 4)	• Analysis case study to be solved with Python in teams
	• Teams can be chosen by yourself
		Due Dates:
		• WP1 Data Exploration
		  • Moodle upload / GitHub Freeze: 16.11.2023
		  • Presentation: 17.11.2023
		• WP2 Data Preparation
		  • Moodle upload / GitHub Freeze: 21.12.2023
		  • Presentation: 22.12.2023
		• WP3 Modeling & Validation
		  • Moodle upload / GitHub Freeze: 25.1.2024
		  • Presentation: 26.1.2024

16-17 participants

HOCHSCHULE DER MEDIEN

Dataset

- Structured dataset describing bike sharing data
- Method to predict label: Regression
- Format:
  - 732 rows × 18 columns
  - 15 features
  - 3 labels: Main label to be used is “cnt” (number of bike rentals per day)

Picture: pixabay

HOCHSCHULE DER MEDIEN

Dataset

#	Attribute	Description
1	instant	record index
2	dteday	date
3	season	season (1:winter, 2:spring, 3:summer, 4:fall)
4	yr	year
5	mnth	month (1 to 12)
6	holiday	whether day is holiday or not
7	weekday	day of the week
8	workingday	if day is neither weekend nor holiday is 1, otherwise is 0.
9	weathersit	weather situation: 1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
10	temp	Temperature
11	atemp	Feeling temperature
12	hum	Humidity
13	windspeed	Wind speed
14	leaflets	Number of distributed leaflets (for marketing purposes)
15	price reduction	Day with price reduction (yes = 1, no =0)
16	casual	count of casual users
17	registered	count of registered users
18	cnt	count of total rental bikes including both casual and registered

Task 1: Data Exploration

- Explore the provided data sets using descriptive statistics (e.g. mean values, standard deviations, min/max values, missing values) and visualizations (e.g. histograms, boxplots)
- Point out which data quality issues you identified in terms of
  - Missing values
  - Outliers
  - Features to be transformed (e.g. normalization) transformation
  - Features to be removed (feature selection)
  - Other insights which require attention in the following phases

Task 2: Model Baseline and Data Preparation

- Create a baseline linear regression model using the originally provided training dataset with minimal preprocessing and evaluate it with your test dataset based on accuracies (MAE) and *coefficient of determination* (R^2)
- Preprocess the original datasets to address the identified data quality issues
  - Missing values
  - Outliers
  - Features to be transformed (e.g. normalization)
  - …
- Within the **test** dataset, you are not allowed to:
  - Remove examples / rows
  - Change label values
- Create a new linear regression model using the preprocessed training dataset and evaluate it with your preprocessed test dataset based on *accuracies* (MAE) and *coefficient of determination* (R^2)
- Export the pre-processed training and test dataset to a CSV

Task 3: Modeling Optimization

- Create a new baseline linear regression model using the preprocessed training dataset and evaluate it with your preprocessed test dataset based on *accuracies* (MAE) and *coefficient of determination* (R^2)
- Optimize your model / create further model versions
  - Algorithm Selection: Experiment with different regression algorithms, e.g. linear regression, polynomial regression, regression trees etc.
  - Hyper-parameter Tuning: Change the hyper-parameters of your algorithms (e.g. „degree“ in case of polynomial regression)
  - Feature Selection: Remove features, which are not helpful for your model
- Evaluate each model version based on accuracies (MAE) and *coefficient of determination* (R^2) using the test data and store evaluation results in a data frame
- Create an overview of your evaluation data frame…
  - By generating an overview table
  - By generating an appropriate visualization

Task Presentations

For each task:

- 10-15 mins presentation
- Structure your presentation based on the bullet points of the task description
- Upload one zip in Moodle containing:
  - One Jupyter Notebook per presentation
  - The data files used in your Jupyter notebook
  - One Powerpoint file per presentation
- Every team member presents his/her part

Task Presentations

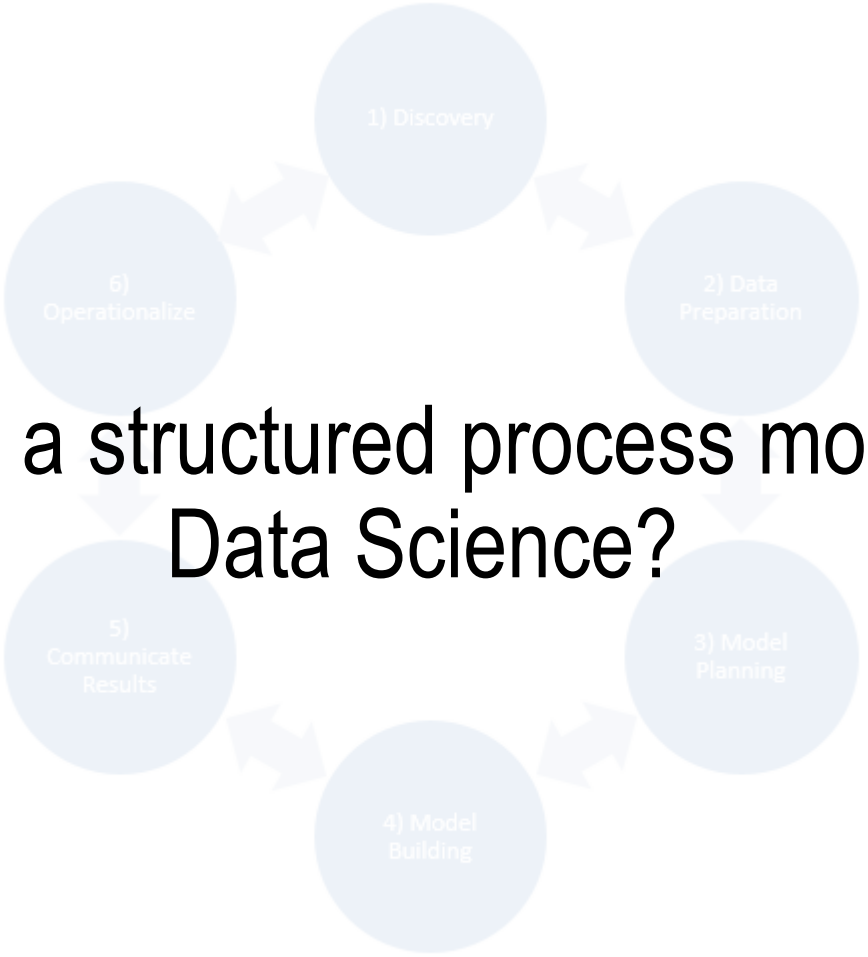
- Jupyter Notebook
  - Should contain extensive comments and markup to structure and explain the coding
  - Should be uploaded in executed state (all result cells visible)
  - Should be uploaded together with the data files you used
- Powerpoint presentation
  - Should not contain coding
  - Should focus on result cells of your Jupyter notebook (tables, diagrams etc.)
  - Focus on interpretation and analysis of results

Agenda

	Agenda
1	Project Description
2	Introduction
3	CRISP DM – Phases
4	PANDAS lab

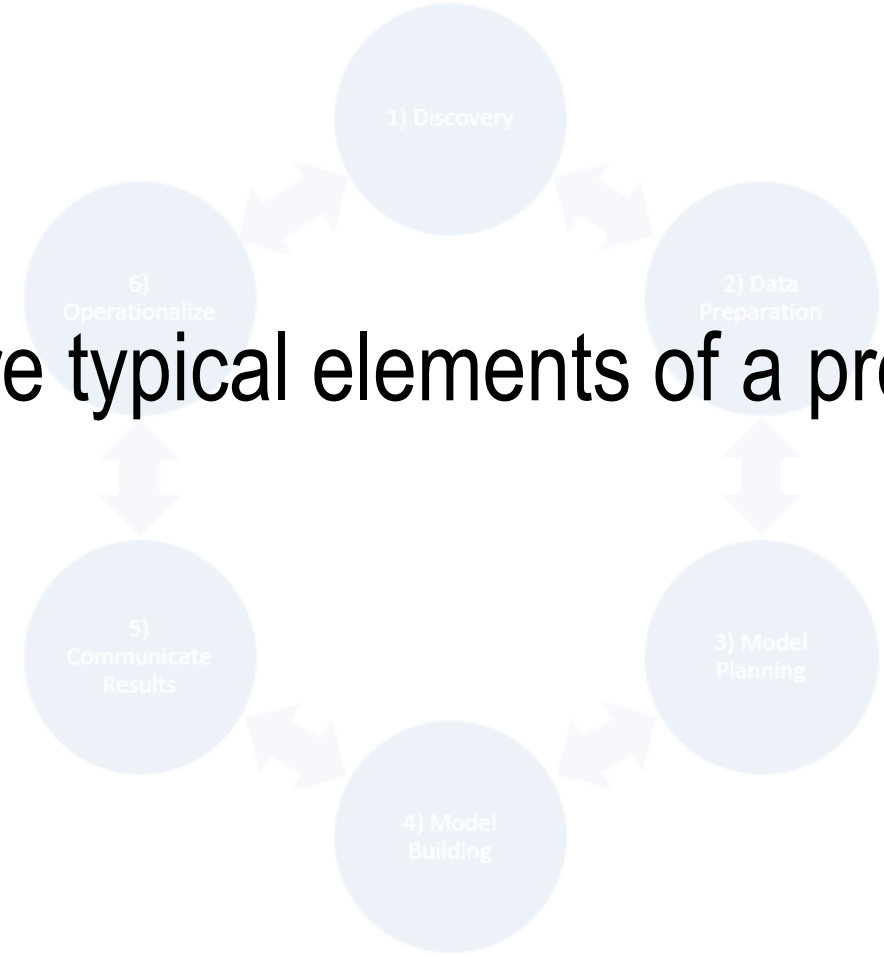
HOCHSCHULE DER MEDIEN

Introduction



Why work with a structured process model to perform Data Science?

Introduction

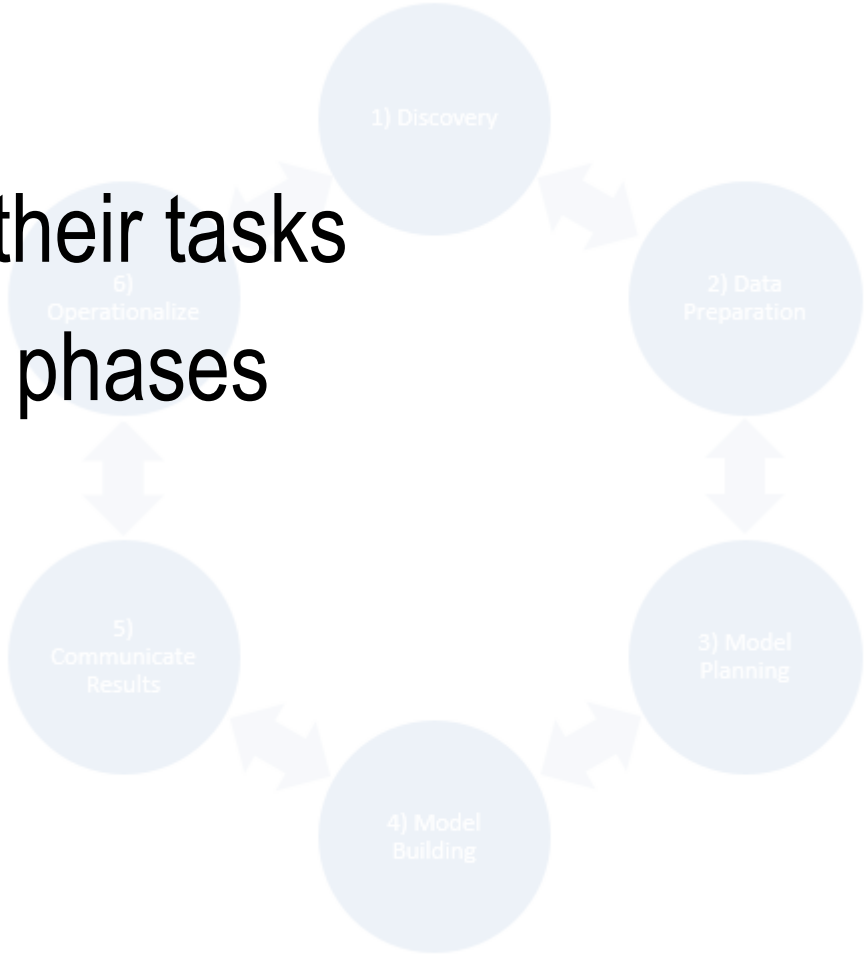


What are typical elements of a process model ?

Introduction

Elements of the Data Science process model

- Phases and their tasks
- Sequence of phases
- Roles



CRISP-DM

Image Source: Wikimedia

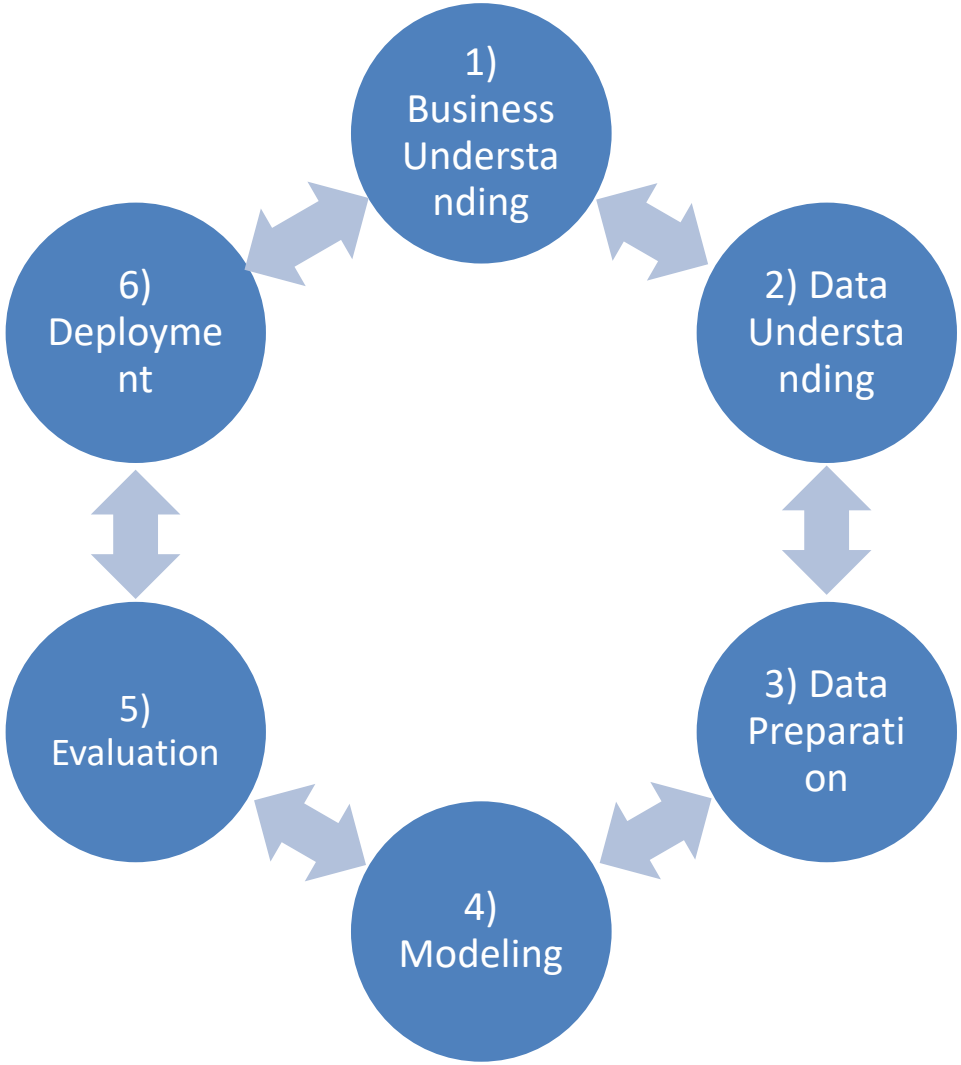
- Cross-industry standard process for data mining (CRISP-DM)
- Open standard process model describing common approaches used by data mining / data science experts
- Defined in 1996 through an industry project, became most widely-used analytics model
- Successor called Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM) which refines and extends CRISP-DM was released in 2015 by IBM

Agenda

	Agenda
1	Project Description
2	Introduction
3	CRISP DM – Phases
4	PANDAS lab

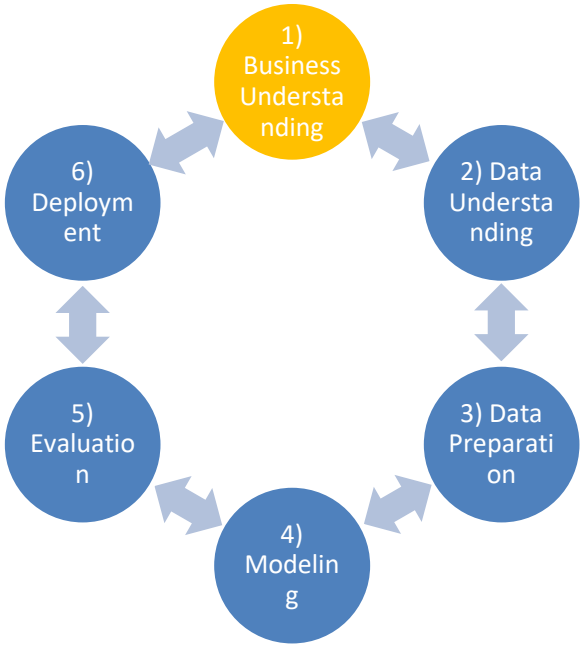
HOCHSCHULE DER MEDIEN

CRISP-DM



HOCHSCHULE DER MEDIEN

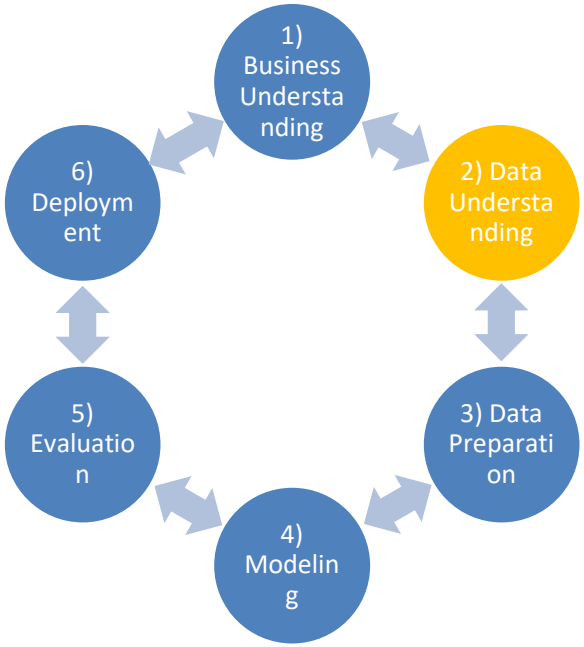
CRISP DM – Business Understanding



- Ideation
  - Identify data science use case
- Assess current situation
  - resources, requirements, risks, terminology, costs and benefits
- Determine project goals
  - business and data science success criteria
- Create project plan
  - including initial assessment of tools and methods needed

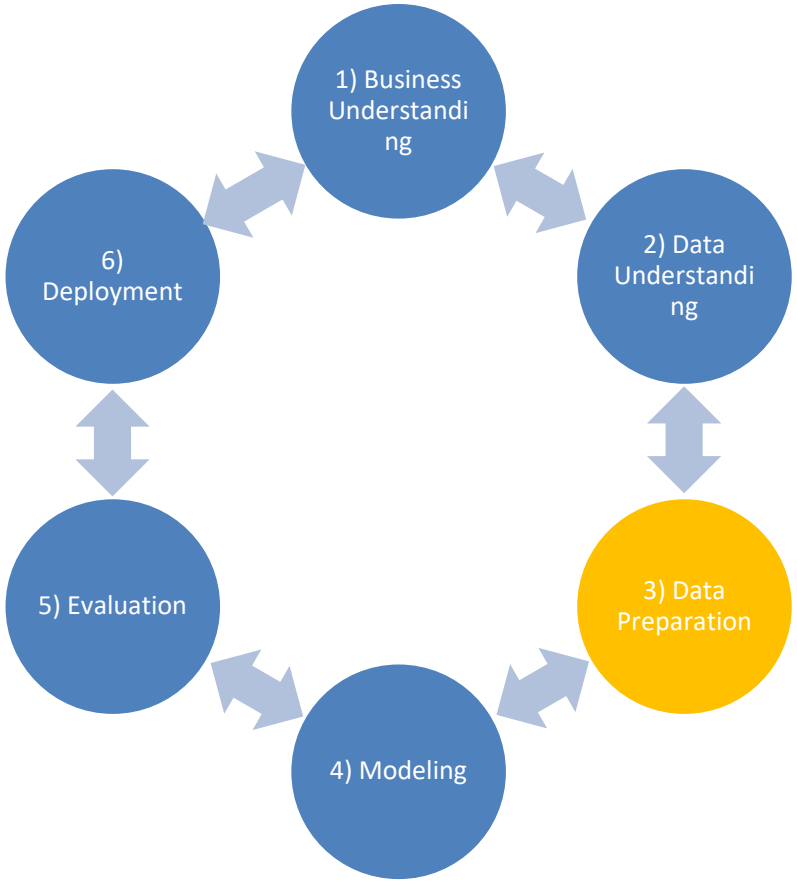
HOCHSCHULE DER MEDIEN

CRISP DM – Data Understanding



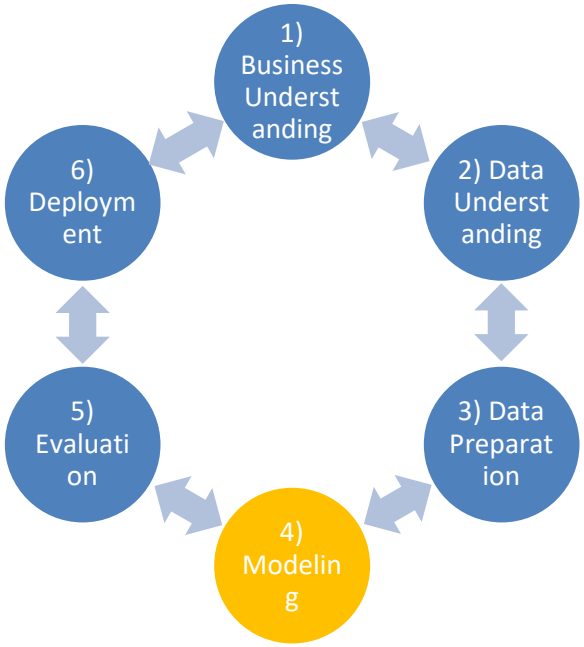
- Explore Data / Verify Data Quality
  - descriptive statistics and visualizations
  - distributions of key attributes, relationships between attributes
  - aggregations, descriptive measures,..

CRISP DM – Data Preparation



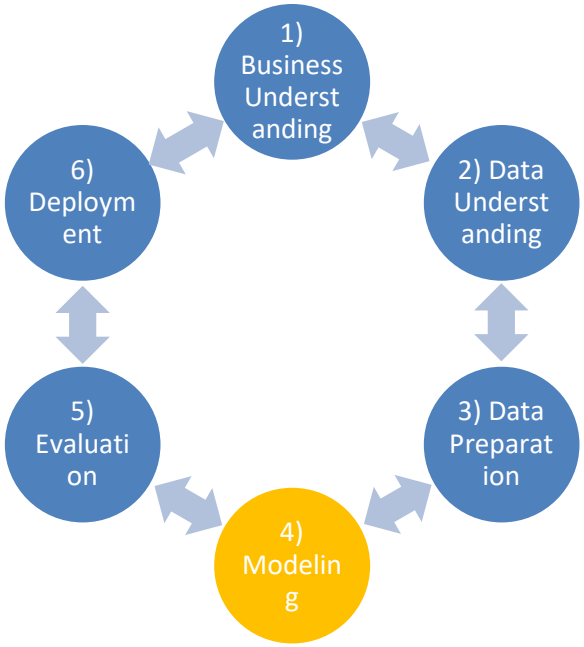
- Select data
- Clean data
- Construct required data
- Integrate data

CRISP DM – Modeling



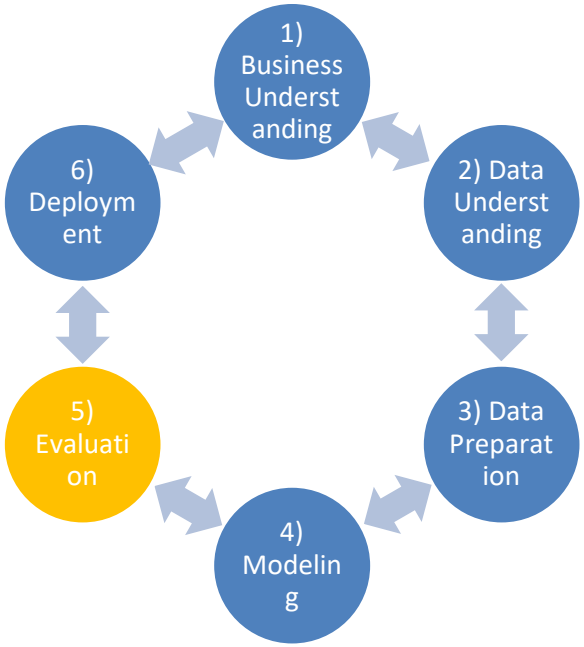
- Select method / algorithm
  - Document choice including corresponding assumptions about data
  - Also consider ensembles (combination of multiple models)
- Generate test design
  - Define evaluation criteria (e.g. accuracy, recall, etc.)
  - Define split strategy (training vs. test data)

CRISP DM – Modeling - cont.



- Build model
  - Execute selected method with defined parameters and prepared data
  - Adjust parameters and data preparation and document different model versions
- Assess model
  - Rank implemented models according to evaluation criteria (e.g. accuracy)
  - Interpret models, judge success from analytics / statistics perspective
  - Tune model parameters according to evaluation criteria

CRISP DM – Evaluation

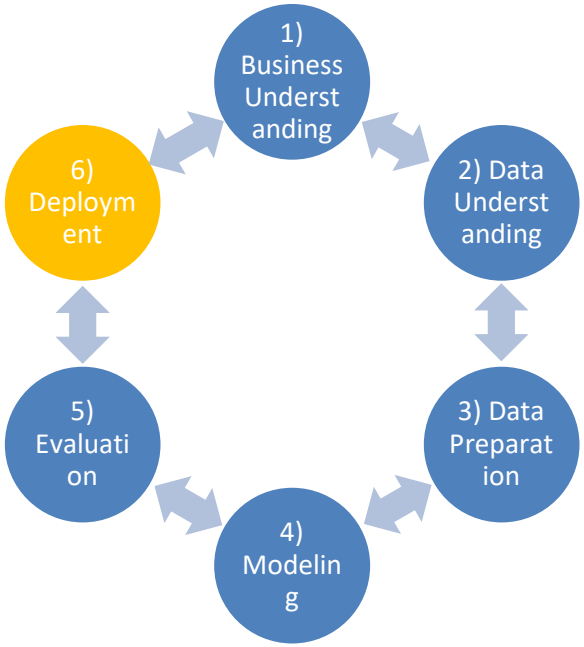


- Evaluate results from business perspective (reflecting business goals)
- Review process (critical reflection of selected methods, parameters, data, tools etc.)
- Determine next steps (deployment vs. further iterations vs. follow-up project)

HOCHSCHULE DER MEDIEN

Foto Source: Pixabay

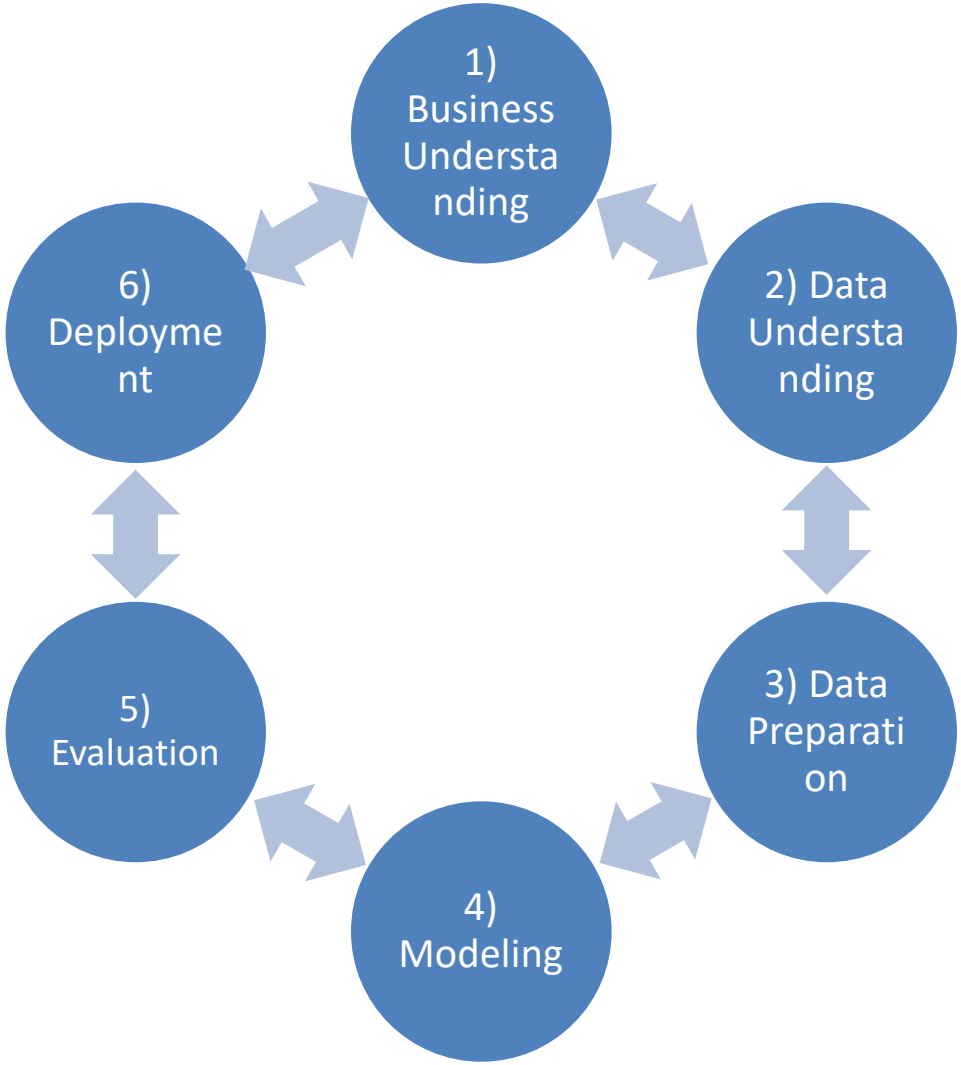
CRISP DM – Deployment



- Plan deployment
- Plan monitoring and maintenance
- Produce final report
- Review project

Foto Source: Pixabay

CRISP-DM



HOCHSCHULE DER MEDIEN

Agenda

	Agenda
1	Project Description
2	Introduction
3	CRISP DM – Phases
4	PANDAS lab

HOCHSCHULE DER MEDIEN