Technische Hochschule Rosenheim

Chapter 06
Data Science

Lecture A2I2

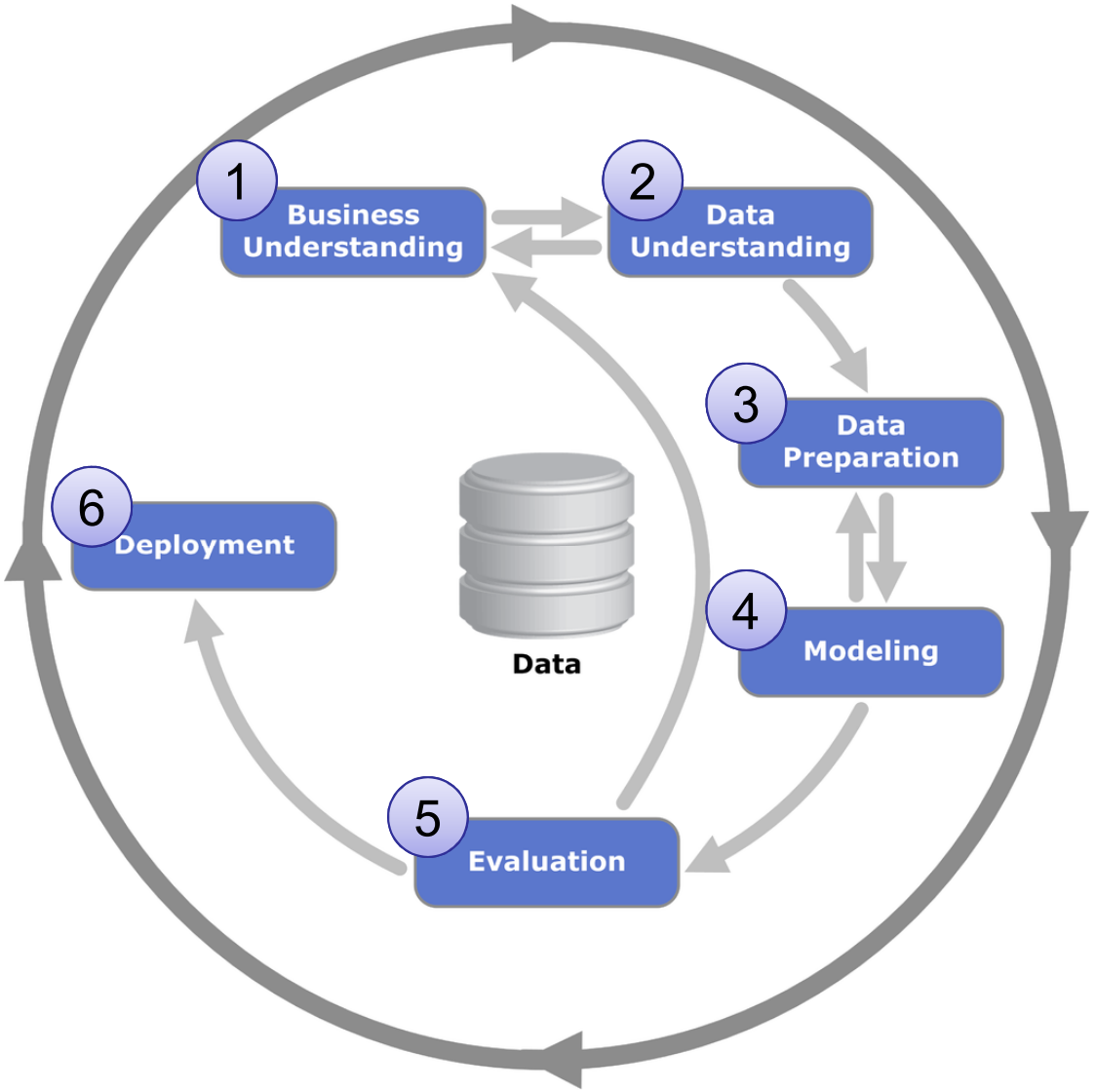
Kai Höfig & Markus Breunig

CRISP-DM

Cross-industry standard process for data mining

- Extracting useful knowledge from data to solve business problems

1. Usecase and problem to be solved
2. What are the cost of data required to solve the problem? How good is the existing data?
3. Converting, normalizing, completing data
4. Capture patterns in the data
5. Gain confidence about the quality of the models
6. Implement the model in order to solve the business problem and get a return of investment.



Task

- Play with the CRISP-DM Cycle

- Part 1: Business Understanding and Mapping to Data Science Problem
- Part 2: Data Understanding and Data Preparation
- Part 3: Modeling and Evaluation (Cost-Benefit-Matrix)

Assignment Part 1

- **Business Problem A (Room 1 and 2)**
- You are working for a railway company operating a railway line from Stuttgart via Ulm to Bieberach. A considerable amount of money is spent on maintenance of your trains (both locomotives and wagons - gears are worn down, brakes need to be maintained etc.). You want to optimize this budget by using machine learning.
- **Business Problem B (Room 3 and 4)**
- You are working for the marketing departments of a company selling luxury cars to affluent customers. Apart from online marketing, direct mailing campaigns still account for a large part of the marketing budget and continue to drive revenue. You want to optimize the impact of your direct mailing budget by using machine learning.
- ***Assignment***
- Understand the business problem in detail (use google!)
- Decompose the assigned problem into at least two different subproblems and decide on the machine learning method you will try for each subproblem
- *40min, Result: short presentation of results.*

Assignment Part 2

- **Business Problem A (Room 1 and 2)**
- Optimize wine revenue. Wine can have different quality - better wines are more expensive than average or below average wines. Predict wine quality based on measured attributes (classification). Data preparation and outlier removal.
- **Business Problem B (Room 3 and 4)**
- Cardiotocography Evaluation. During pregnancy, many doctors perform "fetal cardiotocograms", recording the heartbeat and other measurements of the fetus in order to assess fetal wellbeing. Automatically interpret cardiotocography data to assess fetal health (classification). Data preparation and outlier removal.
- ***Assignment***
- Data preparation and outlier removal.
- *40min, Result: short presentation of results.*

Assignment Part 3

- **Business Problem A (Room 1 and 2)**
- Optimize wine revenue. Wine can have different quality - better wines are more expensive than average or below average wines. Predict wine quality based on measured attributes (classification). Evaluate and create classifiers.
- **Business Problem B (Room 3 and 4)**
- Cardiotocography Evaluation. During pregnancy, many doctors perform "fetal cardiotocograms", recording the heartbeat and other measurements of the fetus in order to assess fetal wellbeing. Automatically interpret cardiotocography data to assess fetal health (classification). Evaluate and create classifiers.
- ***Assignment***
- Evaluate and create classifiers.
- *40min, Result: short presentation of results.*