Machine Learning

Group Project

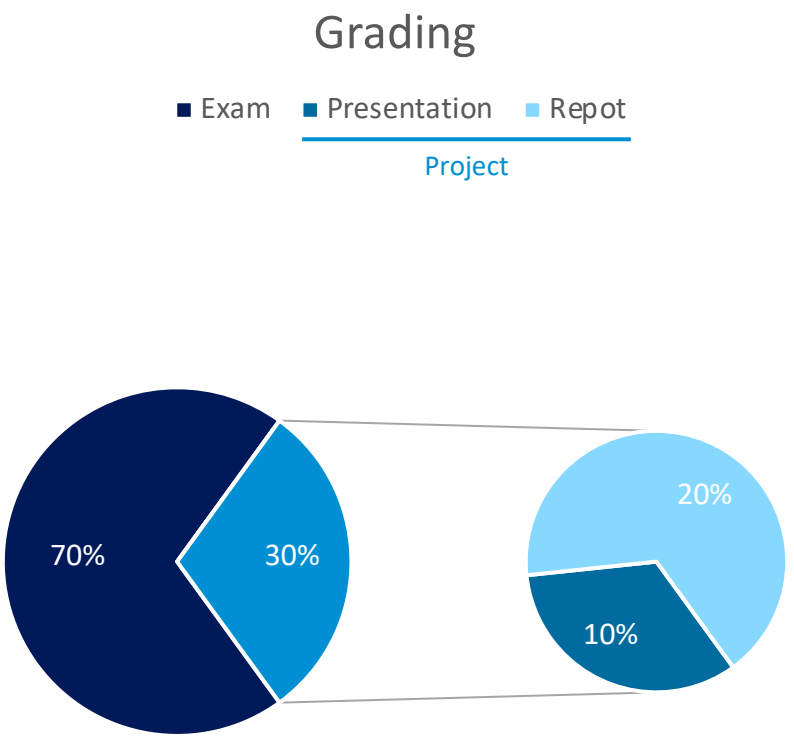
Dr. Jan P. Portisch

Please Mind: Updated Schedule

#	Calendar Week	Date	Weekday	From	To
1	40	Oct. 6th	Friday	02:15 PM	05:30 PM
2	41	Oct. 13th	Friday	08:15 AM	11:30 AM
3	43	Oct. 27th	Friday	08:15 AM	11:30 AM
4	44	Nov. 3rd	Friday	08:15 AM	15:45 PM
PROJECT	50	Dec. 15th	Friday	08:15 AM	11:30 AM
EXAM	51	Dec. 22nd	Friday	08:30 AM	09:30 AM

Grading

- **Written Exam** (70%)
  - individual assignment
  - closed-book
  - tests your concept understanding
- **Group Project** (30%)
  - assigned groups
  - open-book
  - tests your ability to apply learned contents



Group Project

- Teams of four students
  - realize a data mining project
  - present the project result to other students
    10 minutes presentation + 5 minutes Q&A
  - hand in the presentation slides and notebook upfront
- Goals
  - gain practical experience with the complete data mining process
  - apply and learn about preprocessing and data mining methods

Group Project Timeline

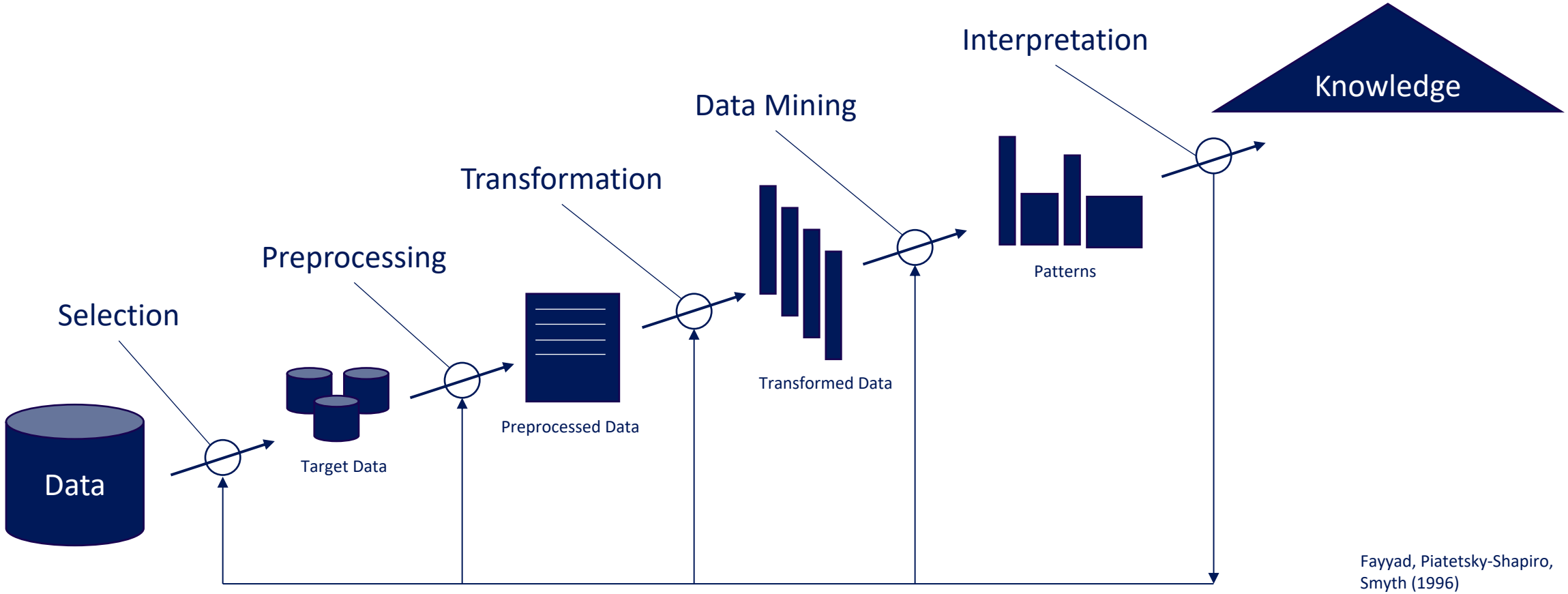
There are two deliverables to be provided by December 10th.

- One notebook together with your data (zipped) that can be run easily on any PC.
- One presentation slide deck (PDF or PPTX) that is used for the presentation. You are allowed to create “backup slides” taht are not submitted.

You present on December 15th.

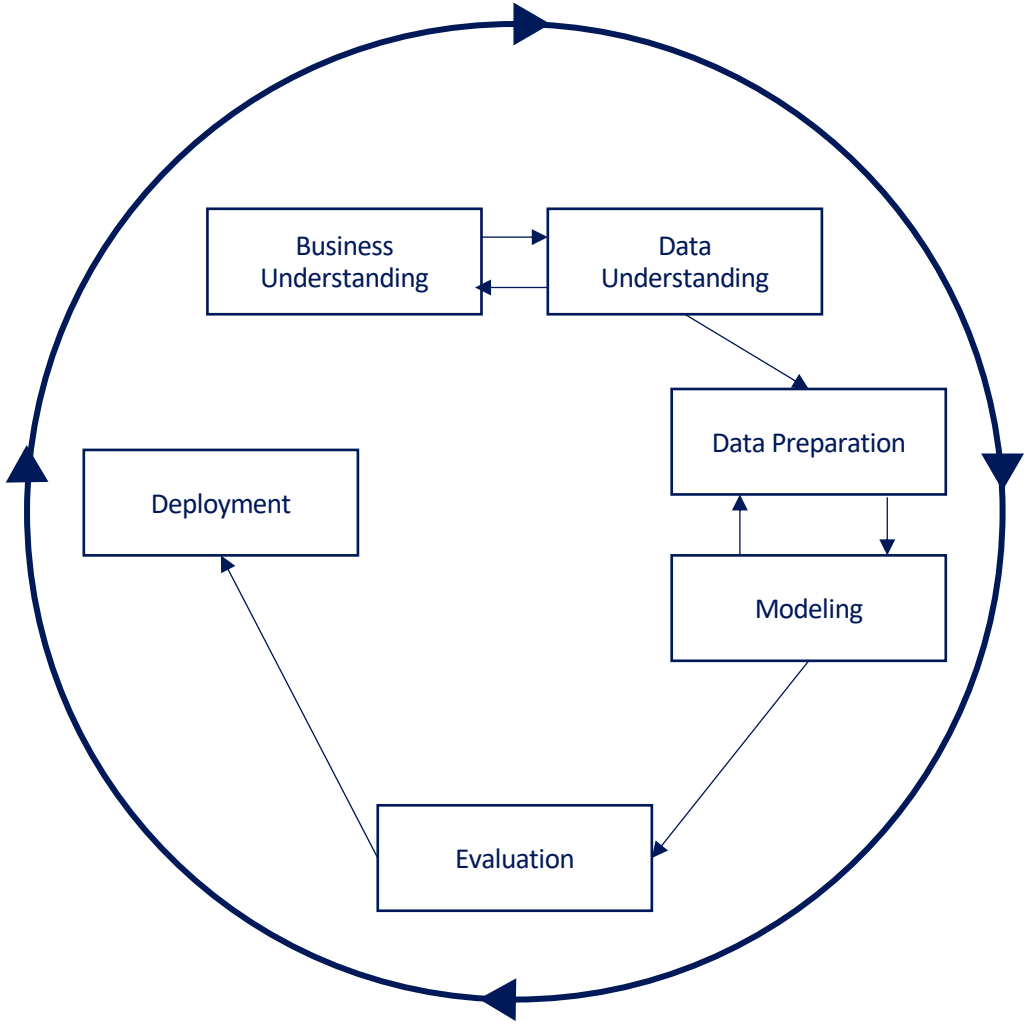
- Everybody must present; everybody must understand the full project.
- You may be asked to show and run (parts of) your submitted notebook.
- Everybody must be able to explain the notebook.

The Data Mining Process
Knowledge Discovery in Databases



Fayyad, Piatetsky-Shapiro, Smyth (1996)

CRISP-DM Process Model
Cross Industry Standard Process for Data Mining



- Use model in business context
- keep iterating in order to maintain an improve model

About Your Presentation

- Your colleagues are your audience – choose an appropriate level of complexity and language.
- Your colleagues do neither know your data nor its context. Introduce your tasks and decisions accordingly.
- Don’t overwhelm your audience with complexity.
- Present your task in an appealing manner, and use suitable visualizations.
- Your slides and notebooks must be in English but you can present in German *or* English language.
- All group members must present.
- There may be questions targeted to specific group members.
- All group members must understand their code and be able to provide a “walk-through”.

Aspects to Cover in the Presentation

Use/Business Case

- What is the context?
- What is to be predicted?
- How can such an ML algorithm be helpful

Aspects to Cover in the Presentation

Use/Business Case

- What is the context?
- What is to be predicted?
- How can such an ML algorithm be helpful

Aspects to Cover in the Presentation

Nature of the Data

- What attributes are available and what do they mean?
- What is the Data Type of the attributes?
- What are the dataset statistics?
- What are (interesting) statistical patterns in the data?
  - Unbalanced data
  - Missing values
  - Outliers
  - etc.
- Are the class labels equally distributed?

Aspects to Cover in the Presentation

Preprocessing and Transformation

- What transformations did you apply?
  - e.g., binning, normalization
- Why did you settle for this kind of transformation?

Aspects to Cover in the Presentation

Data Mining

- How did you model the problem?
- What algorithms did you try out?
- Why?

Aspects to Cover in the Presentation

Parameter Tuning

- What was your setup to determine optimal parameters?
- What were the optimal parameters for the algorithms you chose?

Aspects to Cover in the Presentation

Evaluation

- What was your evaluation setup?
- What is a baseline solution?
- How did your algorithms perform?
- Which algorithm performed best (and why)?
- Did others also work on the data? How was their performance?

Aspects to Cover in the Presentation

Discussion of the Results

- How do you judge the results?
- How hard was the task?
- Are there any recommendations to improve the data?
- What could be done to improve the results?

Group Project

Group 1
Béla Gallin
Joscha Stähle
Samira Kuklinski
Maximilian Knapczyk

Group 2
Louis Hefter
Philipp Strauss
Samuel Sonnenwald
Lucette Kohl

Group 3
Paul Linus Klarer
Yonis Teubner
Lucas Guttensohn
Daria Ermantraut

Group 4
Lean Henriques Fürst
Tobias Tronicek
Naja Pia Lehmann
Lars Christian Gauch

Group 5
Ana Margaride dos Santos Teixeira
Jeremias Matthies
Lukas Strickler
Adrian von Auenmüller

Group 6
Hakon Rosenberger
Lisa Sterner
Tim Strohmenger
Leo Waigel

Group 1: Bank Marketing

Dataset
https://archive.ics.uci.edu/dataset/222/bank+marketing

Group 2: Finding Rich Americans

Dataset
http://archive.ics.uci.edu/dataset/2/adult

Group 3: Determining the Quality of Wine

Dataset
http://archive.ics.uci.edu/dataset/186/wine+quality

Group 4: Predicting the Chance of a Heart Attack

Dataset
https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

Group 5: Predicting the Salaries of Data Scientists

Dataset
https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023/data

Group 6: Predicting Airline Delays

Dataset
https://www.kaggle.com/datasets/jimschacko/airlines-dataset-to-predict-a-delay/discussion