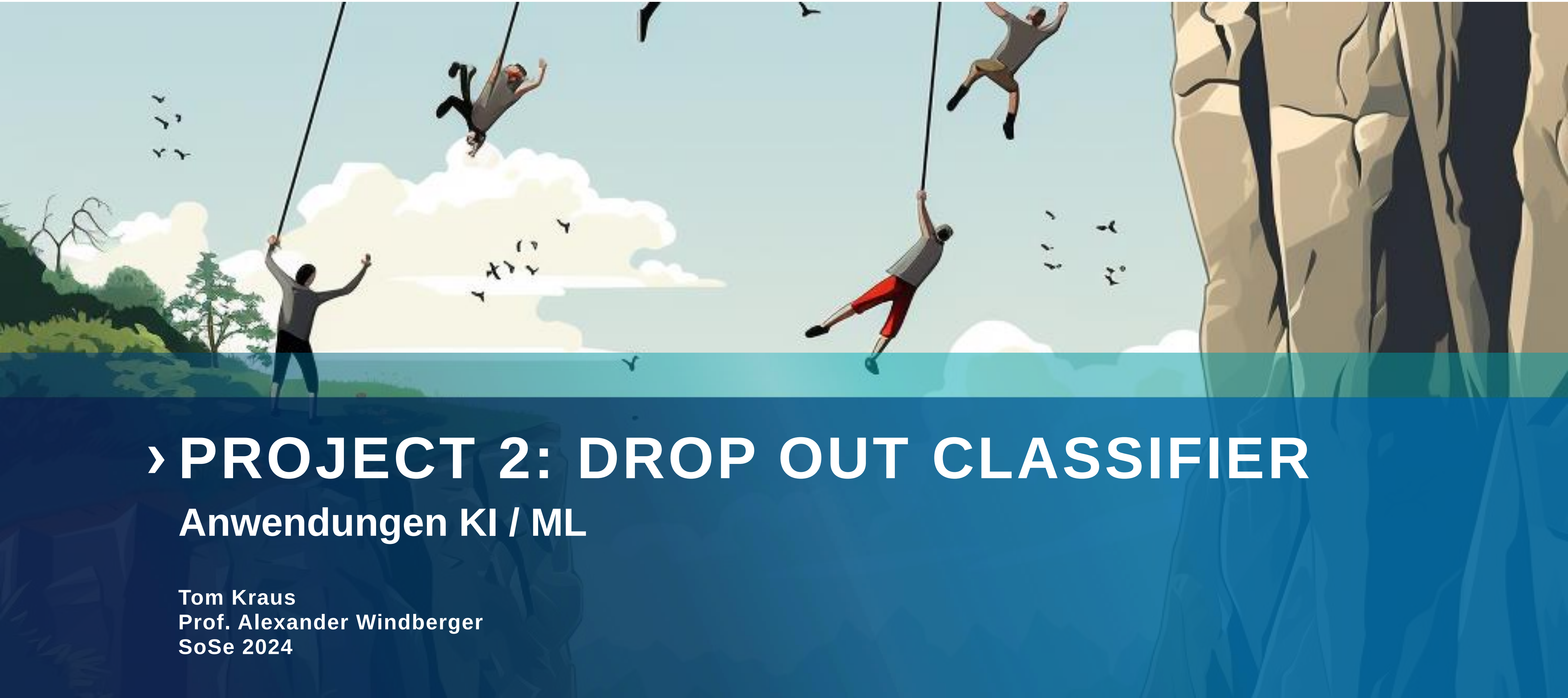
HEILBRONN UNIVERSITY
OF APPLIED SCIENCES



› PROJECT 2: DROP OUT CLASSIFIER

Anwendungen KI / ML

Tom Kraus
Prof. Alexander Windberger
SoSe 2024

PROJECT 2: DROP OUT CLASSIFIER

Develop and evaluate a model that identifies students who are at risk of not graduating to support them accordingly. For this purpose, you will be provided with characteristics that describe different backgrounds of the students. You may use the Python packages Pandas, MatPlotLib, Scikit-Learn and Numpy to solve the tasks.

The dataset has been created through the following research work:

Martins, M.V., *et al*. (2021). Early Prediction of student's Performance in Higher Education: A Case Study. In: Trends and Applications in Information Systems and Technologies. WorldCIST 2021. Advances in Intelligent Systems and Computing, vol 1365. Springer, Cham. https://doi.org/10.1007/978-3-030-72657-7_16

Siehe: https://link.springer.com/chapter/10.1007/978-3-030-72657-7_16

PROJECT 2: DROP OUT CLASSIFIER

Tasks (**7 Points**):

1. Perform data pre-processing and clean, analyse and explore the data set.
2. Calculate and visualise how the features correlate with each other and with the labels. Pick an interesting correlation and discuss it.
3. Train at least four machine learning algorithms: One probabilistic, one tree-based, and one distance-based, and one ensemble method. Are all models equally well suited for this task? Discuss your conclusion.
4. Evaluate the four models using k-fold cross validation and give at least accuracy (mean and standard deviation) and confusion matrix for the trained models. Is one of the models significantly better than the others?
5. Pick your favorite model. Which features were most relevant for the for the students' success?
6. Save your favorite model as pickle-file with https://scikit-learn.org/stable/model_persistence.html. Call the file "best_model.pkl".

HEILBRONN UNIVERSITY
OF APPLIED SCIENCES

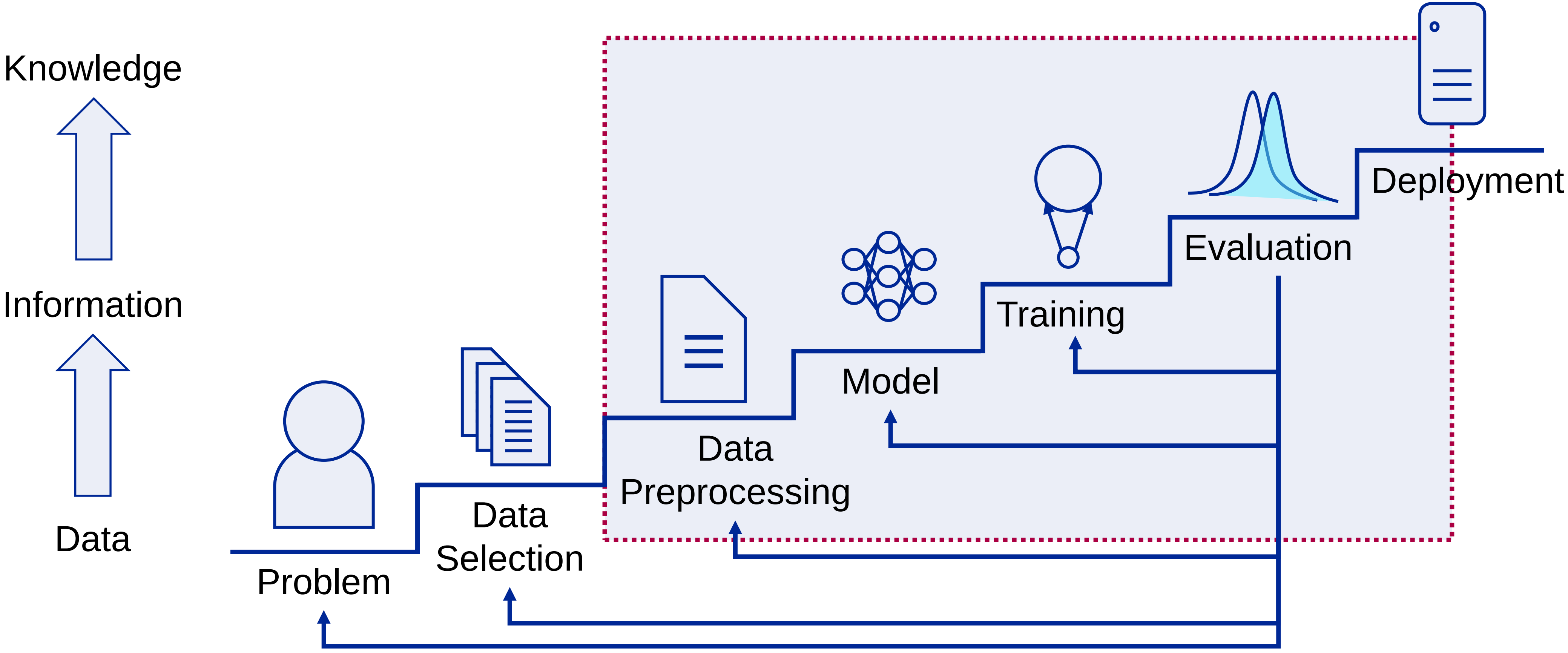
PROJECT 2: DROP OUT CLASSIFIER

- The submission consists of two files:

  1. A Jupyter Notebook containing the preprocessing, the training, and the evaluation of your models.
  2. A pickle-file “best_model.pkl”.

- Due Date is **Mai 29th**
- There are **3 Points** for presentation, style, and creativity
- There will be **1 bonus Point** for the top-5 accuracy models on a retained test data set

WHAT'S TO BE LEARNED
THE MACHINE LEARNING PROCESS



WHAT‘S TO BE LEARNED

- **Learn how to extract information from data**
  - Deal with different data types (structured and unstructured)
  - ML algorithms
  - Explore
  - Statistics
  - Linear algebra (why live in 3 dimensions, when you can master 1000s?)
  - Use analysis to find the sweet spots

WHAT‘S TO BE LEARNED

- **Learn how to interpret your findings**
    - What are the methods and metrics?
    - What correlations can be found?
    - What‘s significant and what‘s not?

WHAT‘S TO BE LEARNED

- **Communicate and present your results**
  - Make pretty graphs
  - Your data contains a story. Tell it!
  - Make mistaces, harness criticism, defend your theses
  - Convince your peers, boss, customer, client, referee …