UNIVERSITY OF MANNHEIM

Data Mining

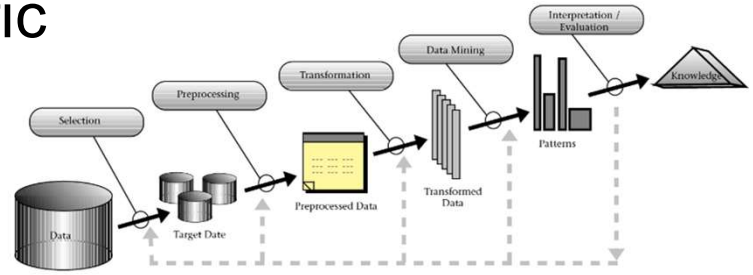
Introduction to the Student Projects

Outline

1. Requirements for the Student Projects
2. Requirements for the Project Reports
3. Final Exam
4. Team Formation

Student Projects

- Goals
  - Gain practical experience with the complete data mining process
  - Get to know additional problem-specific
    - preprocessing methods
    - data mining methods



- Expectation
  - You select an interesting data mining problem of <u>your choice</u>
  - You solve the problem using
    - the data mining methods that we have learned so far, including
      - proper hyperparameter optimization
      - problem-specific pre-processing and smart feature engineering
    - additional data mining methods which might be helpful for solving the problem and build on what we learned in class

Procedure

- Teams of six students
  1. realize a data mining project
  2. write a 10-page summary of the project and the methods employed in the project
  3. present the project results to the other students
     - 10 minutes presentation + 5 minutes discussion

- Final mark for the course
  - 20 % written summary about the project
  - 5 % project presentation
  - 75 % written exam

Schedule

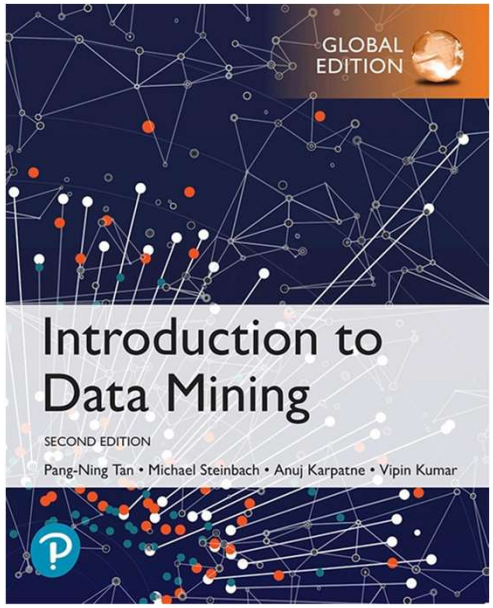
Week	Wednesday	Thursday
Until 18.04.2023	Form and register your team (see Section 4)	
19.04.2023	Introduction to student projects and formation of additional Groups	Preparation of project outline
26.04.2023	Lecture: Association Analysis	Exercise: Association Analysis
Thursday, April 27th 2023, 23:59: Submission of Project Outlines		
03.05.2023	**Feedback on Project Outlines**	Project Work
10.05.2023	Project Work	Feedback on demand
17.05.2023	Project Work	Feedback on demand
24.05.2023	Project Work	Feedback on demand
Sunday, May 28th 2023, 23:59: Submission of Project Reports		
31.05.2023	Presentation of Project Results	
06.06.2023	Final Exam (offline)	

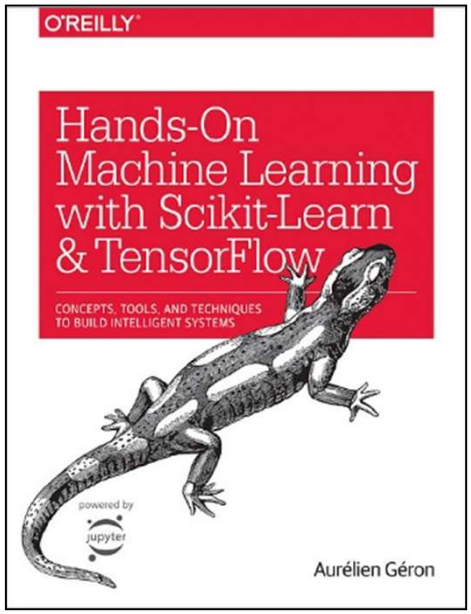
Where to find interesting Data Sets?

- Kaggle
  - website running commercial and educational data science competitions
  - https://www.kaggle.com/
  - If you use a Kaggle task:
    You must compare your results to results from the competition's forum!
- Papers with Code
  - thousands of datasets organized by task together with papers about state-of-the-art methods
  - https://paperswithcode.com/datasets
- Huggingface
  - thousands of datasets organized by task together with deep learning models
  - https://huggingface.co/datasets
- KDD Cup and Data Mining Cup
  - Data mining competitions providing data sets and solutions
  - http://www.kdd.org/kdd-cup
  - https://www.data-mining-cup.com
- Google Dataset Search
  - https://datasetsearch.research.google.com/

Where to Find Information about Additional Methods?

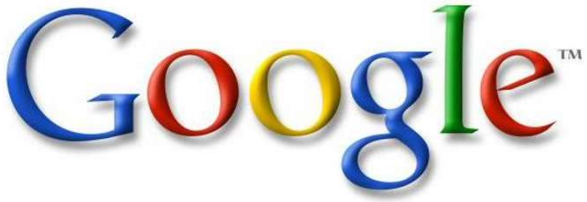
1. Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Pearson / Addison Wesley.
2. Aurélien Géron: Hands-on Machine Learning with Scikit-Learn. O'Reilly.
3. Bing Liu: Web Data Mining, 2nd Edition, Springer.



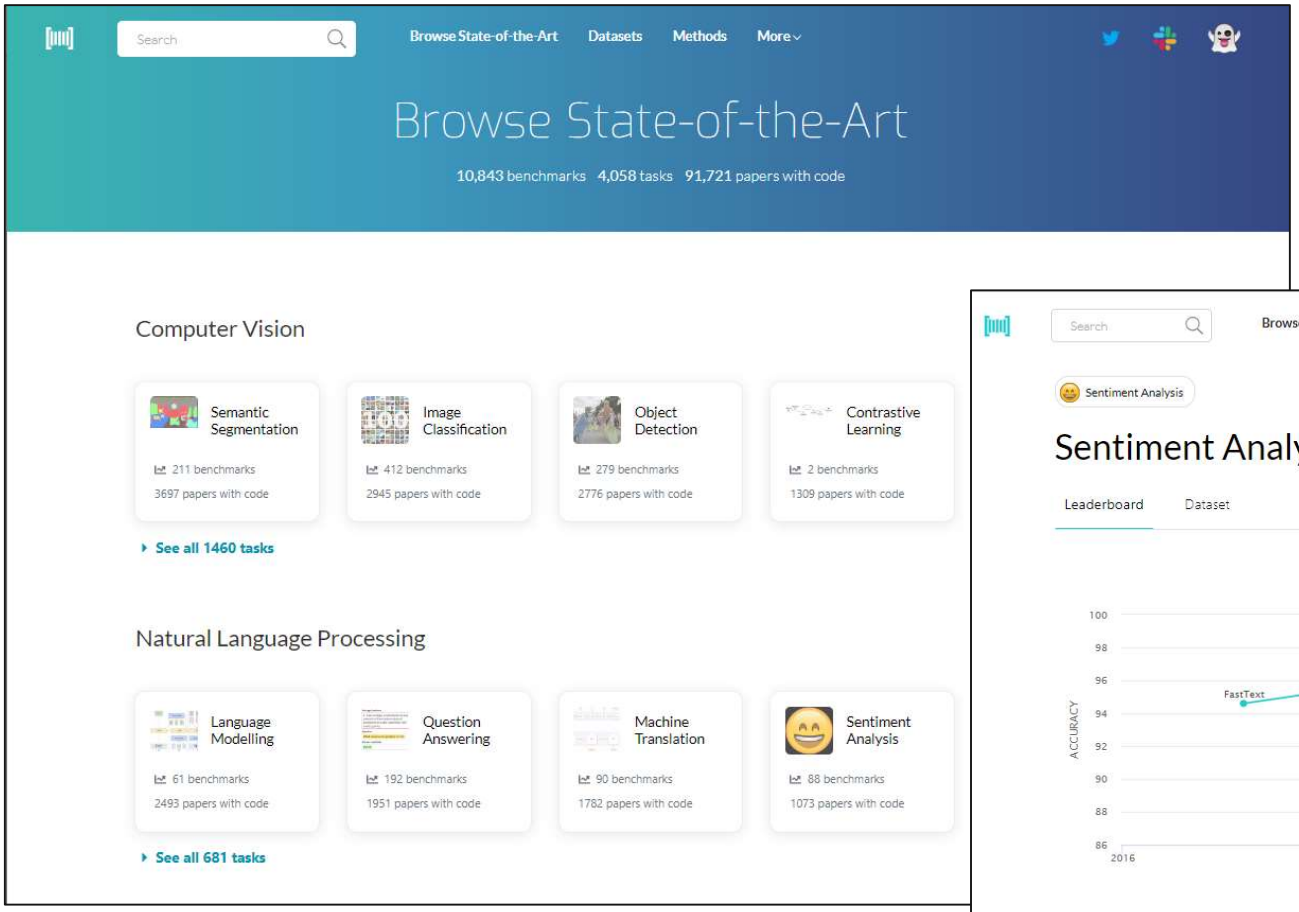


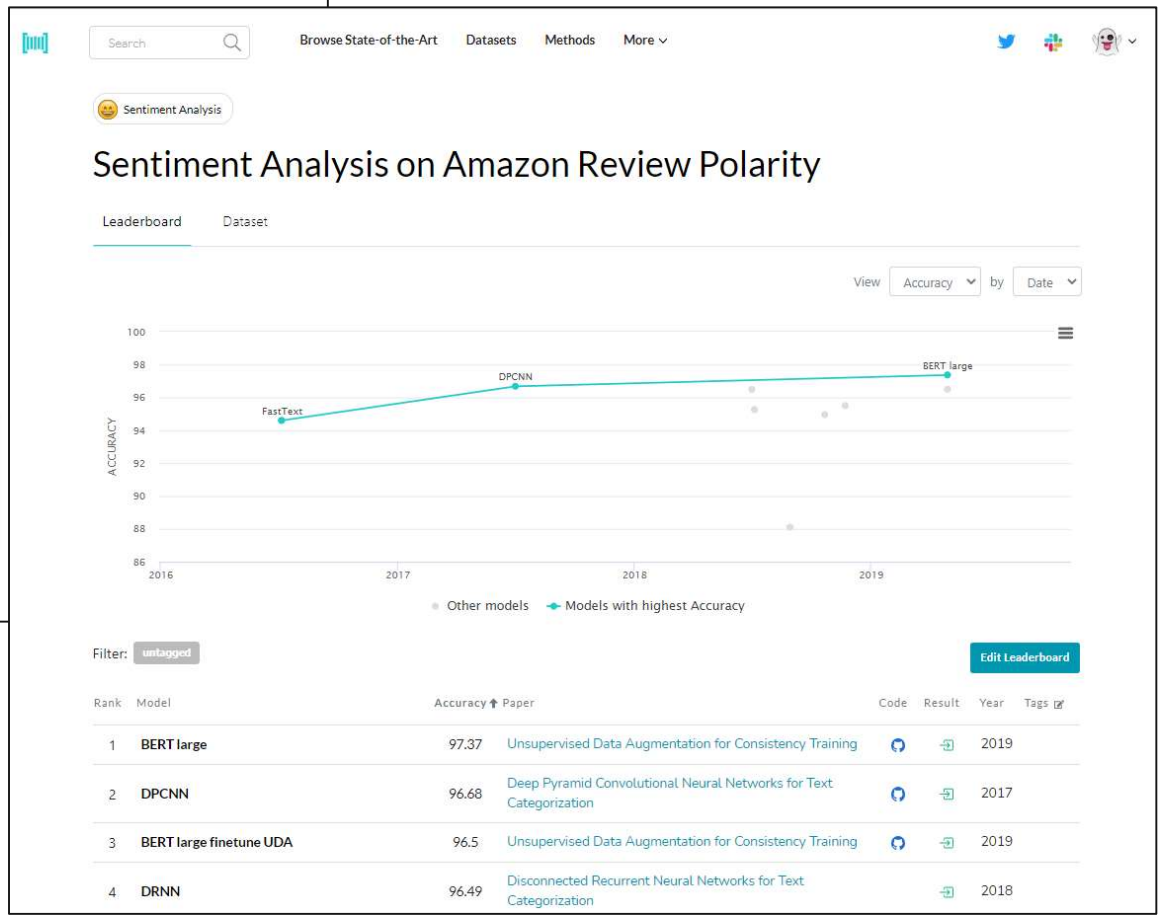
Where to Find Information about Additional Methods?

- Check out the solutions to your problem that other people have tried.
  - by looking into the Kaggle discussion groups and code
  - by investigating the state-of-the-art for your your task on Papers with Code
  - by looking at submissions of the KDD Cup or Data Mining Cup
  - or search for relevant scientific papers using Google Scholar, search term: “task name + survey”



State of the Art for Specific Tasks





https://paperswithcode.com/sota

Some Project Ideas (not binding)

- Web Log Mining
  - Learn a classifier for categorizing the visitors of your website.
  - Which features matter? Number of pages visited, time on site, .. (Bing Liu Chapter 12.x)
  - Preprocess some web log data
  - Learn and evaluate classifier
- Sentiment Analysis for Discussion Forum / Rating Site / Tweets
  - Are people positive, neutral, or negative about topic / product? (Bing Liu 11.x)
- Estimate House or Car Prices
  - using different regression methods or transfer learning to localize method
- Wikipedia Contributors / Hoax Articles
  - Examine the edit history of Wikipedia contributors
  - Cluster users by different attributes (no of edits, edits/day, topic, ...)
  - Or learn a classifier for categorizing Wikipedia contributors

Some Projects Realized in Previous Semesters

- Mannheim Police Reports
  - Learn classifiers for police reports
  - Identify type of incident, severity of incident, location of incident
- last.fm Playlist Analysis
  - Cluster last.fm users according to the style of the songs they are listening to
  - Find commons sets of songs for the different clusters
- Analysis of Training Data of a Fitness Center
  - Identify different customer groups by clustering exercise data
  - Find frequent combinations of exercises
- Bundesliga Betting Rules
  - Find rules that help you to predict the outcome of a Bundesliga game
- Transfer Learning for Sentiment Analysis of Tweets about Movies
  - Learned classifier from IMDB movie reviews
  - Applied and tested with tweets afterwards
- Classifying a Document‘s Perspective
  - using the example of Israeli – Palestinian Essays

Project Outlines

- maximum 4 pages using Springer Computer Science Proceedings layout or Word
  - Include a project name and your team number on the first page!
- due Thursday, April 27th 2023, 23:59
- send by eMail to Chris, Alex, Keti, Ralph
- answer the following questions:

1. What is the problem you are solving?
2. What data will you use?
   - Where will you get it?
   - How will you gather it?
3. How will you solve the problem?
   - What preprocessing steps will be required?
   - Which algorithms do you plan to use? Be as specific as you can!
4. How will you measure success? (Evaluation method)
5. What do you expect your results to look like? (Model/Clusters/Patterns)

- Feedback about your project outlines **if required**: Wednesday, 3.05.2023, 10:15-11:45
- We will inform you Tuesday 02.05.2022 in the afternoon if feedback is required

Coaching Sessions

- We will give you tips and answer questions concerning your project.
- Registration via email to Keti, Alex & Ralph is mandatory!
  - until Tuesday night!
  - including the questions that you like to discuss
  - including which session you prefer (Thursday B2/B3)
- We will assign you a time slot afterwards and inform you about the slot via email.
- Coaching sessions will take place in Room B6 A 1.04
- Every team must attend at least one coaching session!

Some Project Management Hints

- Organize your project in multiple iterations
  - Every artefact will be improved over time!
- Get a simple process running early on to have a baseline
- Parallelize tasks while keeping centrally track of results
  - e.g. one central document with results plus reference to exact version of the notebooks/datasets that produced these results
  - sub-groups should explore specific ideas for a specified amount of time
- Define concrete milestones: When should what be finished?
  - e.g. 10.5.23 Data exploration results collected in single document
  - e.g. 14.5.23 Subgroup on sentiment lexica adds results to central document
- Infrastructure
  - use shared folder for result document, versions of data, processes, slideset (e.g. MS Teams, Google Drive, github)
  - use ChatGPT for inspiration about additional methods as well as coding

Tasks within the Iterations of the Project

1. Data Exploration and Visualization
2. Data Preprocessing: value normalization, deal with outliers, deal with missing values, feature generation, balance training data if necessary
3. Establish/update baseline (majority class, predict mean value)
4. Try different learning methods using different feature creation methods and feature combinations
5. Perform error analysis in order to understand what is going on!
6. Later iteration:
    1. run automatic hyperparameter optimization and attribute selection
    2. employ more sophisticated evaluation setup: x-val + holdout vs. nested x-val

Project Report

- 10 pages (exactly!) plus references page, no appendix ➔ document length: 11 pages
- Each extra page and each day of late submission downgrades your mark by 0.3!
- due Sunday, May 28th 2023, 23:59
- send by email to Chris, Keti, Alex & Ralph
- Outline for project report:
  1. Application area and goals (0.5 pages)
  2. Profile (structure and size) of your data set (minimum 1 page)
  3. Preprocessing and Mining
     - describe different approaches and parameter settings (**parameter optimization**) that you tried
     - including description of **evaluation setup** (split, x-val, nested-x-val?) and evaluation results
     - including an **analysis of the errors** still made by the best method, a discussion of the results, and a comparison to state-of-the-art results (together: minimum 2 pages)
- Requirements
  1. You **must use** the latex template of the Springer Computer Science Proceedings
  2. Please cite sources properly and use your references page
  3. Also submit your **Python code** and (a subset) of **your data**
  4. Include your names and your team number on the first page!

Template: Springer Computer Science Proceedings

my title

author1 name[1] and author2 name[1,2]

[1] University 1
email1@gmail.com
[2] University 2
email2@gmail.com

Abstract. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

1 Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

http://www.springer.com/de/it-informatik/lncs/conference-proceedings-guidelines

Checklist for Project Reports

- Business Understanding
  - What is the actual problem (in the domain)?
  - What is the target variable?
    - Classification/Regression/Cluster Analysis?
- Data Understanding
  - What is the distribution of labels / target variable?
  - Are all attributes and their types listed and important attributes explained?
  - What is the quality of the data? Wrong values? Outdated?
  - What does correlation analysis reveal about attribute importance?
- Preprocessing
  - Are missing values replaced (in case needed)?
  - Checked for outliers (and handled them)?
  - Validity tests of attributes (Height above sea level < 9000)?
  - Check for inconsistencies (age=42, birthday=03/07/1997)
  - Check for duplicates
  - Performed data normalization (e.g. US vs United States)
  - Additional features generated?
  - Has binning been tried out?
  - Feature subset selection necessary?

Checklist for Project Reports

- External Knowledge
  - Are additional datasets used?
- ML approaches
  - Which ML approaches were tried out?
  - How did you optimize hyperparameters (which attributes/ in which range / nested-x-val) ?
  - Do you have at least one baseline (majority class / mean value / domain specific …)?
- Evaluation
  - Do you use fix train/test split, x-validation, or nested x-validation?
  - Is eval stratified?
  - Cost matrix or not?
  - Analyze a symbolic model (how does the decision tree / rules /… look like?)
  - What features do have a high impact on the result?
  - What types of errors are done by the best model? (error analysis)
- Result
  - Is the result is critically evaluated?
  - Is the best result compared to the baseline? Compared to the state-of-the-art?
  - What does the result mean given the problem? Could you use the model in practice?

Get Additional Advice from a Stanford Professor

Christopher Potts

- How to evaluate your model?
  - https://www.youtube.com/watch?v=TxTbIROT9lY
- How to structure your project report?
  - https://www.youtube.com/watch?v=DZNwO-p5PGY
- How to present the results of your project?
  - https://www.youtube.com/watch?v=GGx7klcahzY

Severe Errors to Avoid

1. Normalize numeric data before calculating any similarity scores

```
# import min-max scaler
from sklearn import preprocessing.MinMaxScaler()

# create scaler
scaler = MinMaxScaler()

# normalize the relevant attributes
dataset[['Att1', 'Att2']] = scaler.fit_transform(datas
```

2. If your data is unbalanced
   - balance your training data
   - do NOT balance your test data
   - report P/R/F1, not accuracy

```
from imblearn.over_sampling import RandomOverSampler

# Up-sample positive class
sampler = RandomOverSampler()
balanced_training_data, balanced_target = sampler.fit_re
```

3. Implement the recommendations concerning model evaluation, hyperparameter selection and feature selection given on the summary slides

```
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import cross_val_score
from sklearn.svm import SVC

# Specify hyperparameter combinations for search
parameter_grid = {"C": [1, 10, 100, 1000], "gamma": [.001, .01, .1, 1]}

# Create SVM
estimator_svm = SVC(kernel='rbf')

# Create the grid search for model selection
estimator_gs = GridSearchCV(estimator_svm, parameter_grid, scoring='accu

# Run nested cross-validation for model evaluation
accuracy_cv = cross_val_score(estimator_gs, dataset, labels, cv=5, scor
```

Questions?

3. Final Exam

- Date: June 6th, time tba Be at room 15 minutes before the start.
- Duration: 60 minutes
- Location: tba
- Structure: 6 open questions that
  - check whether you have understood the content of the lecture
    - we try to cover all major chapters of the lecture: cluster analysis, classification, evaluation, regression, association analysis, and text mining
  - require you to describe the ideas behind algorithms or apply the methods
    - What is the advantage or problem of X compared to Y?
    - How do methods react to this special pattern in the data?
    - Given the following data. What happens?
  - might require you to do some simple calculations
    - you need to be able to use the most relevant formulas
    - you do not need to use a calculator

Questions?

4. Team Formation

- You are allowed to form teams of 6 students as you like!
  - You enter your team consisting of 6 students into the Group Formation Google spreadsheet until Tuesday, April 18th
  - If you are still looking for a team, enter yourself to the respective section of the spreadsheet
  - The URL of the spreadsheet is shared in the exercises on Thursday 30.3.
- We will form teams out of the remaining students who did not find a team by themselves
- We will send you the contact information of all your team members after the kickoff session via email.
- Meet with your team after the group formation session on April 19th to organize your work!
  - decide project topic
  - organize writing of project proposal