Introduction to Student Projects

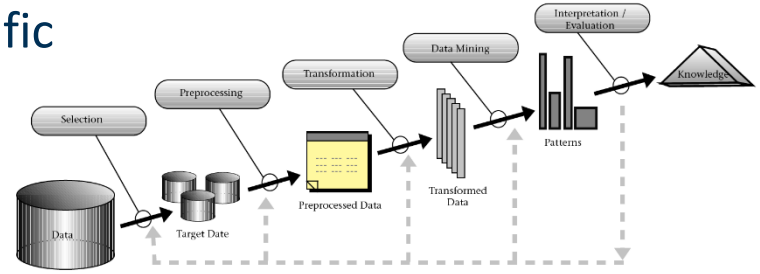
IE500 Data Mining

Outline

1. Requirements for the Student Projects
2. Requirements for the Project Reports
3. Final Exam
4. Team Formation + Start to work!

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 your choice
  - 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Teams of **five** students
  - realize a data mining project
  - write a 12-page summary of the project and the methods employed in the project
  - present the project results to the other students
    - 10 minutes presentation + 5 minutes discussion

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

Schedule

You are here

Week	Monday	Thursday
30.09.2024	Introduction to student projects and formation of groups	Formation of Groups / Preparation of project outline
07.10.2024	Lecture: Regression	Exercise: Regression
14.10.2024	Lecture: Preprocessing	Exercise: Preprocessing
Sunday, October 13th 2024, 23:59: Submission of Project Outlines		
21.10.2024	**Feedback on Project Outlines**	Project Work
28.10.2024	Lecture: Clustering and Anomalies	Exercise: Clustering and Anomalies
04.11.2024	Association Analysis	Exercise: Association Analysis
11.11.2024	Project feedback session	Project Work
18.11.2024	Project feedback session	Project Work
25.11.2024	Project feedback session	Project Work
Sunday, December 1st, 23:59: Submission of Presentation as PDF		
02.12.2024	Presentation of Project Results	Presentation of Project Results
Sunday, December 8th, 23:59: Submission of Project Reports		
18.12.2023	Final Exam	

Where to find interesting Data Sets?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Data registries
  - Datasets hosted on Amazon AWS https://registry.opendata.aws
  - Google's Dataset Search: https://datasetsearch.research.google.com/
  - Microsoft Datasets: https://msropendata.com/
  - Yahoo Webscope Datasets: http://webscope.sandbox.yahoo.com/
  - Dataset collection on Github: https://github.com/awesomedata/awesome-public-datasets
  - Data Hub: http://datahub.io
  - Linked Open Data Cloud: http://lod-cloud.net/
  - Stanford Large Network Dataset Collection: http://snap.stanford.edu/data/index.html
  - Huggingface: https://huggingface.co/datasets

Where to find interesting Data Sets?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Public sector data
  - US government: https://www.data.gov
  - UK government: https://data.gov.uk
  - EU: https://www.europeandataportal.eu
  - CIA World Fact Book: https://www.cia.gov/library/publications/the-world-factbook/
  - Health data (over 125 years): https://www.healthdata.gov/

Where to find interesting Data Sets?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Competitions
  - Kaggle: https://www.kaggle.com/
  - Data Mining Cup: http://www.data-mining-cup.de
  - KDD Cup: https://www.kdd.org/kdd-cup
  - DrivenData: https://www.drivendata.org
  - CrowdAnalytix: https://www.crowdanalytix.com

- If you use a competitions task:
  You **have to** compare your results to results from the competition's forum!

Where to find interesting Data Sets?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

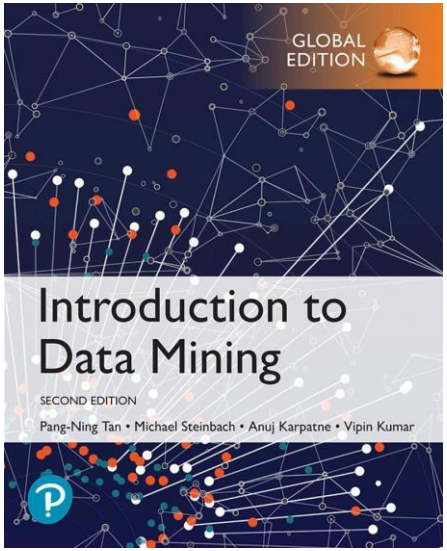
- Language resources
  - WordNet: https://wordnet.princeton.edu
  - EuroWordNet: http://projects.illc.uva.nl/EuroWordNet/
  - Project Gutenberg (36.000 ebooks): http://www.gutenberg.org/
  - New York Times (starts 1851): http://developer.nytimes.com/docs
  - Wikitionary: https://www.wiktionary.org
    as KG: http://kaiko.getalp.org/about-dbnary/

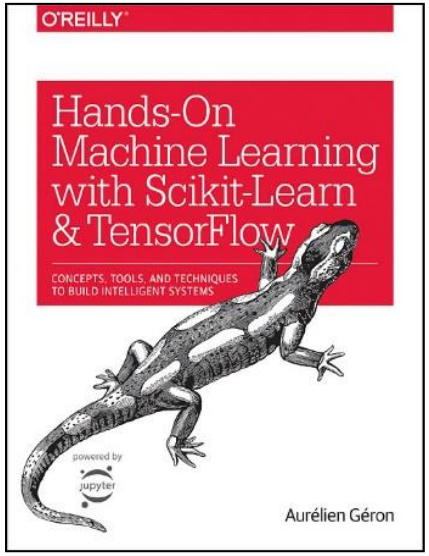
- Knowledge graphs
  - Wikidata: https://www.wikidata.org
  - BabelNet: https://babelnet.org
  - DBpedia: http://wiki.dbpedia.org

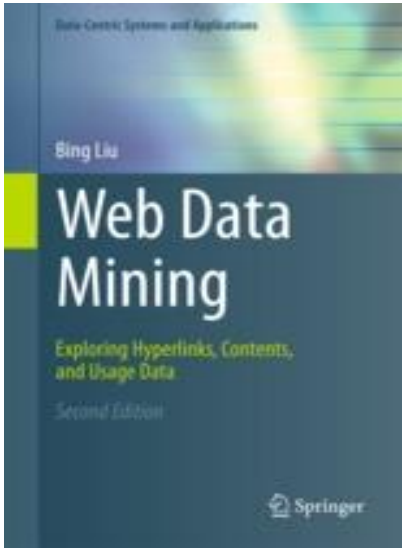
Where to Find Additional Information

UNIVERSITY OF MANNHEIM
Data and Web Science Group

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







Where to Find Additional Information

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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"

Papers With Code

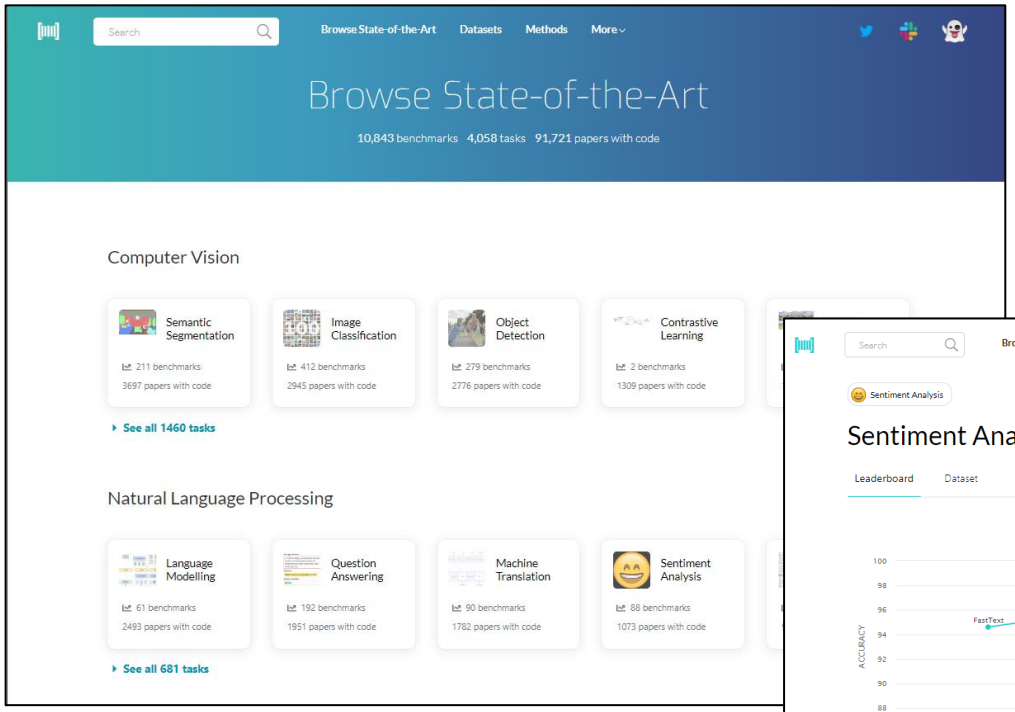
kaggle

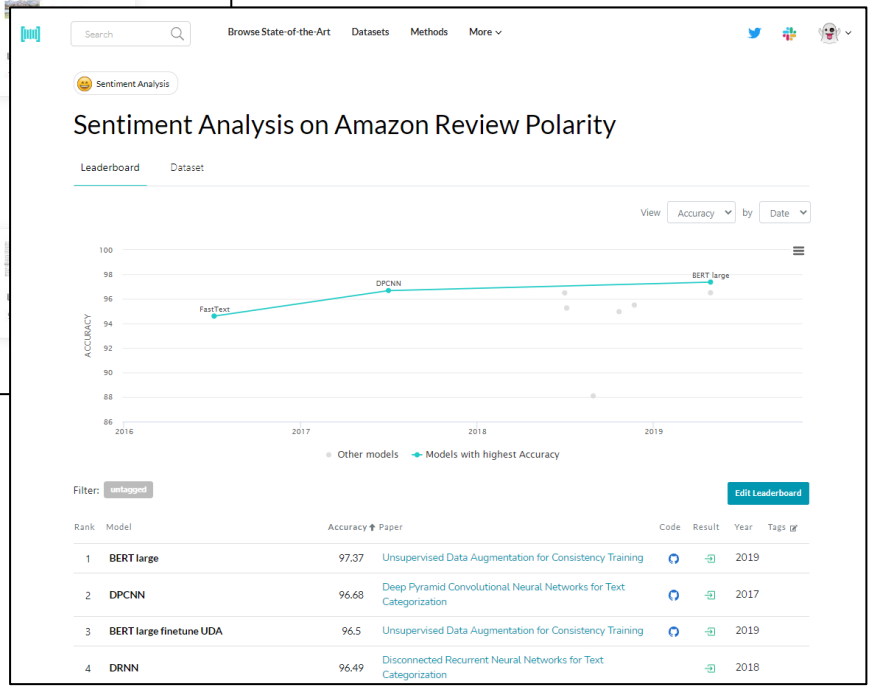
Google

CHAT GPT

DATA MINING CUP
International Student Competition

State of the Art for Specific Tasks





https://paperswithcode.com/sota

Some Project Ideas (not binding)

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Web Log Mining
  - Learn a classifier for the categorizing the visitors of your website.
  - Which features matter? Number of pages visited, time on site, ..
  - Learn and evaluate classifier
- 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
- Sentiment Analysis for Discussion Forum / Rating Site / Tweets
  - Are people positive or negative about topic / product? (Bing Liu 11.x)
- SPAM Detection
  - eMail, blog or discussion forum (Bing Liu 6.10, 11.9)
  - You Tube comments

Some Projects realized in previous Semesters

- Twitter data
  - humor / hate speech detection
  - Sentiment Analysis of Tweets about Movies
    - Learned classifier from IMDB movie reviews
    - Applied and tested with tweets afterwards
- Airbnb (done very often)
  - predict the prices of new apartments
- Bundesliga Betting Rules
  - Find rules that help you to predict the outcome of a Bundesliga game
- 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
  - Find different customer groups by clustering exercise data
  - Find frequent combinations of exercises
- Sentiment Analysis of Tweets about Movies

Some Projects realized in previous Semesters

- Twitter data
  - humor / hate speech detection
  - Sentiment Analysis of Tweets about Movies
    - Learned classifier from IMDB movie reviews
    - Applied and tested with tweets afterwards
- Airbnb (done very often)
  - pred
- Bundes
  - Find rules that help you
- 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
  - Find different customer groups by clustering exercise data
  - Find frequent combinations of exercises
- Sentiment Analysis of Tweets about Movies

Choose a task/dataset where you have a ground truth (or can easily generate one)

Team Formation

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- You are allowed to form teams of 5 students as you like!
  - You enter your team consisting of 5 students into the Group Formation Google spreadsheet (see last slide) until Sunday, October 6th 23:59
  - If you are still looking for a team, enter yourself to the respective section of the spreadsheet also until Sunday, October 6th 23:59
    - Ilias message board can also be used to find teams (see corresponding channel)
  - We will form teams out of the remaining students who did not find a team by themselves on Monday, October 7th
    - We send an information in Ilias message board once the assignment is done
  - If you already formed a team, you can start writing the project outline

- Meet with your team after the group formation session to organize your work!
  - Decide project topic
  - Organize writing of project outline

Project Outlines

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Maximum 4 pages (sharp!) including title page
  - Using DWS master thesis layout (PDF!)
  - Include a project name, your team number and name on the first page!
- Due **Sunday, October 13th, 23:59**
- Send by eMail to Andreea **&** Franz (together, not separate)

- Feedback about your project outlines if required: Monday, 21.10.2024, lecture time (13:45-15:15)
  - We will inform you Friday, 18.10.2024 with some feedback via mail and let you know if you need to show up on Monday, 21.10.2024

Project Outlines

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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)

Coaching Sessions

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- We will give you tips and answer questions concerning your project.
- At the time of the lecture (Mondays)
- **Registration via email** is mandatory!
  - Via mail to Andreea **&** Franz (together, not separate)
  - Until Thursday night (23:59)!
  - Including the questions that you like to discuss
- We will assign you a time slot afterwards and inform you about the slot via email

- **Every team has to attend at least one coaching session!**

Some Project Management Hints

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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

Some Project Management Hints

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Define concrete milestones**: When should what be finished?
  - e.g. 18.10.24 Data exploration results collected in single document
  - e.g. 01.11.24 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

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:
   - run automatic hyperparameter optimization and attribute selection
   - employ more sophisticated evaluation setup: x-val + holdout vs. nested x-val

Project Presentation

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Present the project results to the other students
  - 10 minutes presentation + 5 minutes discussion
  - During lecture/exercise slot
  - Everyone
- Send your presentation in **PDF format**
  - Via mail to Andreea **&** Franz (together, not separate)
  - Until **Sunday, December 1st, 23:59**

Project Report

- Max. 12 pages including title/toc page and reference page
  - max. 10 pages content, no appendix
  - Each extra page and each day of late submission downgrades your mark by 0.3!
- due Sunday, December 8th, 23:59
- send by email to Andreea & Franz (together, not separate)

Project Report

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Outline for project report:
  - Application area and goals (0.5 pages)
  - Profile (structure and size) of your data set (minimum 1 page)
  - Preprocessing
  - Data Mining
    - Describe different approaches and parameter settings/optimizations that you tried
  - Evaluation
    - Including description of evaluation setup (split, x-val, nested-x-val?)
    - 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)
  - Results

Project Report

- Requirements
  - You have to use the latex template of the DWS Thesis
  - Please cite sources properly and use your references page
  - Also submit your Python code and (a subset) of your data
  - Include your names and your team number on the first page!

- Usage of AI Tools needs to be declared

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT	Rephrasing	Throughout	+
DeepL	Translation	Throughout	+
ResearchGPT	Summarization of related work	Sec. 2.2	-
Dall-E	Image generation	Figs. 2, 3	++
GPT-4	Code generation	functions.py	+
ChatGPT	Related work hallucination	Most of bibliography	++

Checklist for Project Reports

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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?

Checklist for Project Reports

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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?

- External Knowledge:
  - Are additional datasets used?

Checklist for Project Reports

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- ML approaches
  - How many different ML approaches were tried out?
  - Do you have at least one symbolic and one non symbolic approach?
  - Do you have at least one baseline (majority class / mean value / domain specific …)?

- Evaluation
  - Is there a train test split or 10-fold cross validation implemented
  - Is the evaluation stratified?
  - Cost matrix or not?
  - Are the hyper parameters tuned (in which range / which attributes) ?
  - Are the tests systematic?
  - Analyse a symbolic model (how does the decision tree / rules /… looks like)
  - What features do have a high impact on the result?

Checklist for Project Reports

- Result
  - Is the result critically evaluated
  - Is the result analyzed against the baseline
  - What does the result mean given the problem (could you use it)

Get Additional Advice from a Stanford Professor

Christopher Potts

- How to evaluate your model?
  - https://www.youtube.com/watch?v=TxTblROT9lY
- 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Normalize numeric data before calculating any similarity metrics

Python

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

# create scaler
scaler = MinMaxScaler()

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

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

```python
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='acc

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

Final Exam

- Date: **Wednesday, 18th December 2024**, time tba.
  - Duration: 60 minutes
  - Location: tba
- Structure: 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 are not allowed to use a calculator (calculations are simple)

Deadlines - Overview

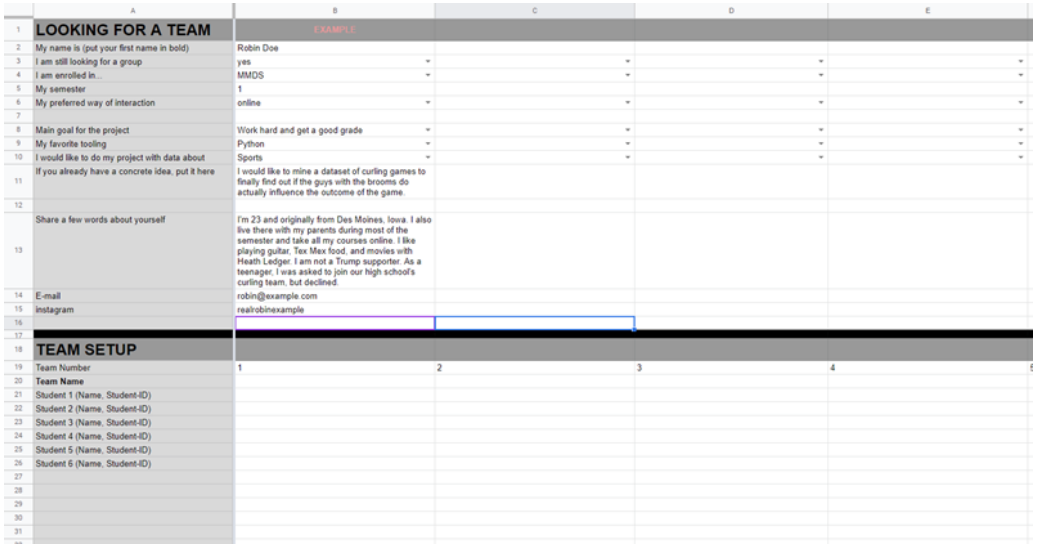
- Team formation until **Sunday, October 6th 23:59**
  - Either enter your whole team or
  - Enter your name if you are looking for a team (team assignment on Monday, October 7th)
- Project outline until **Sunday, October 13th, 23:59**
- Coaching Sessions
  - Every team has to attend at least one coaching session
- Project presentation in PDF until **Sunday, December 1st, 23:59**
- Project report until **Sunday, December 8th, 23:59**

Questions?

Team Assignment

- Find your team now!
- Enter your group in “Team Setup” in Google Sheet
  - In case you do not have a team, fill in your details in “Looking for a team”

  => then you will be assigned to a team after the registration period
- Do so until the end of week (Sunday October 6th 23:59)



	A	B	C	D	E
1	**LOOKING FOR A TEAM**	EXAMPLE			
2	My name is (put your first name in bold)	Robin Doe			
3	I am still looking for a group	yes			
4	I am enrolled in...	MMDS			
5	My semester	1			
6	My preferred way of interaction	online			
7					
8	Main goal for the project	Work hard and get a good grade			
9	My favorite tooling	Python			
10	I would like to do my project with data about	Sports			
11	If you already have a concrete idea, put it here	I would like to mine a dataset of curling games to finally find out if the guys with the brooms do actually influence the outcome of the game.			
12					
13	Share a few words about yourself	I'm 23 and originally from Des Moines, Iowa. I also live there with my parents during most of the semester and take all my courses online. I like playing guitar, Tex Mex food, and movies with Heath Ledger. I am not a Trump supporter. As a teenager, I was asked to join our high school's curling team, but declined.			
14	E-mail	robin@example.com			
15	instagram	realrobinexample			
16					
17					
18	**TEAM SETUP**				
19	Team Number	1	2	3	4
20	**Team Name**				
21	Student 1 (Name, Student-ID)				
22	Student 2 (Name, Student-ID)				
23	Student 3 (Name, Student-ID)				
24	Student 4 (Name, Student-ID)				
25	Student 5 (Name, Student-ID)				
26	Student 6 (Name, Student-ID)				

https://docs.google.com/spreadsheets/d/1Luy7aV8FIRu_nyf6mh8LAyZHo7IJ_CkCcvWtVh75MYY/edit?usp=sharing

Thank you