

Introduction to Student Projects

IE500 Data Mining



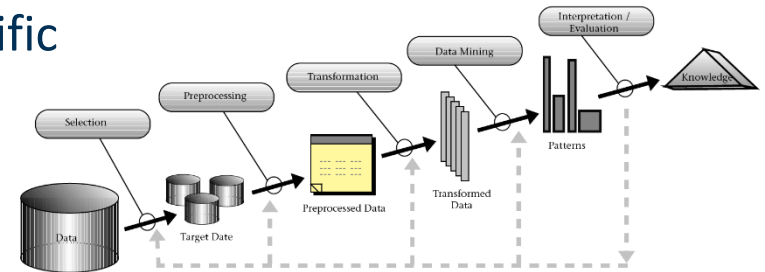
Outline

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

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

- Teams of **five to six** 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

Where to find interesting Data Sets?

- 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?

- 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?

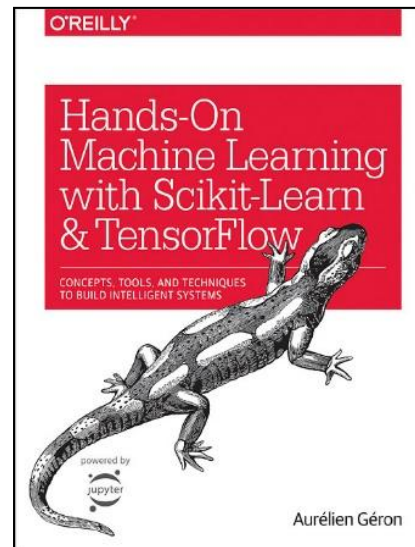
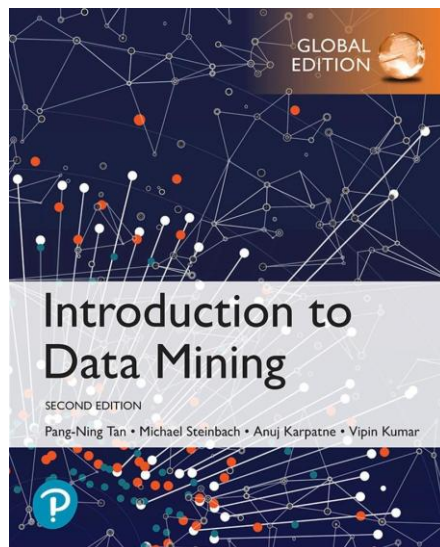
- Competitions
 - Kaggle: <https://www.kaggle.com/>
 - Data Mining Cup: <http://www.data-mining-cup.de>
 - KDD Cup: <https://www.kdd.org/kdd-cup>
 - DrivenData: <https://www.drivendata.org>
 - CrowdAnalytix: <https://www.crowdanalytix.com>
- If you use a competitions task:
You **have to** compare your results to results from the competition's forum!

Where to find interesting Data Sets?

- 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>
 - Wiktionary: <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

- 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

- 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

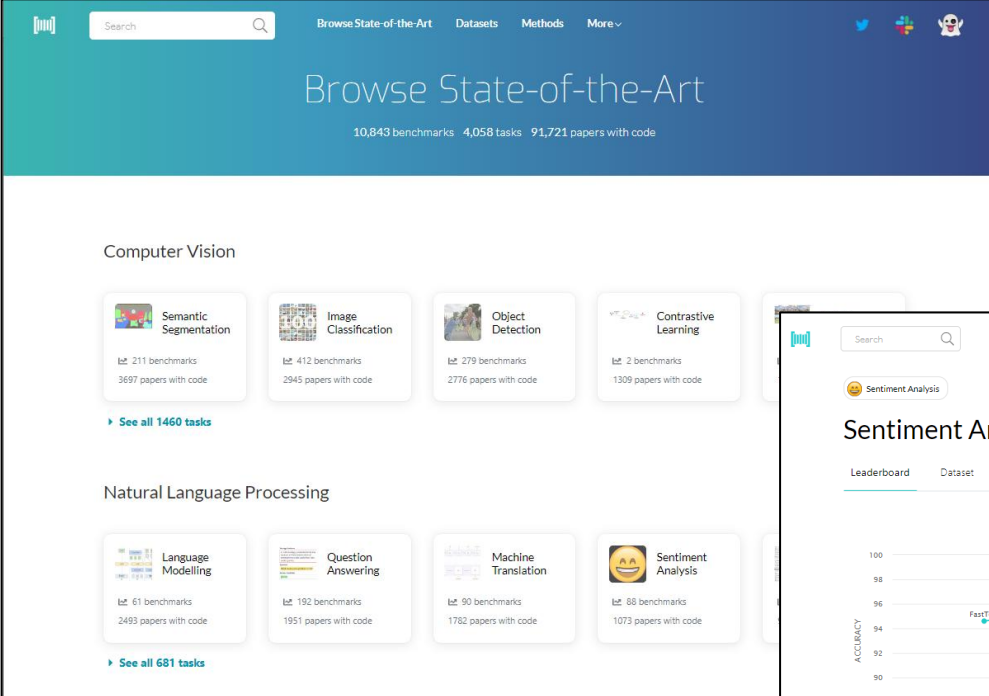
kaggle

Google™



DATA MINING CUP
International Student Competition

State of the Art for Specific Tasks



Browse State-of-the-Art
10,843 benchmarks 4,058 tasks 91,721 papers with code

Computer Vision

- Semantic Segmentation**: 211 benchmarks, 3697 papers with code
- Image Classification**: 412 benchmarks, 2945 papers with code
- Object Detection**: 279 benchmarks, 2776 papers with code
- Contrastive Learning**: 2 benchmarks, 1309 papers with code

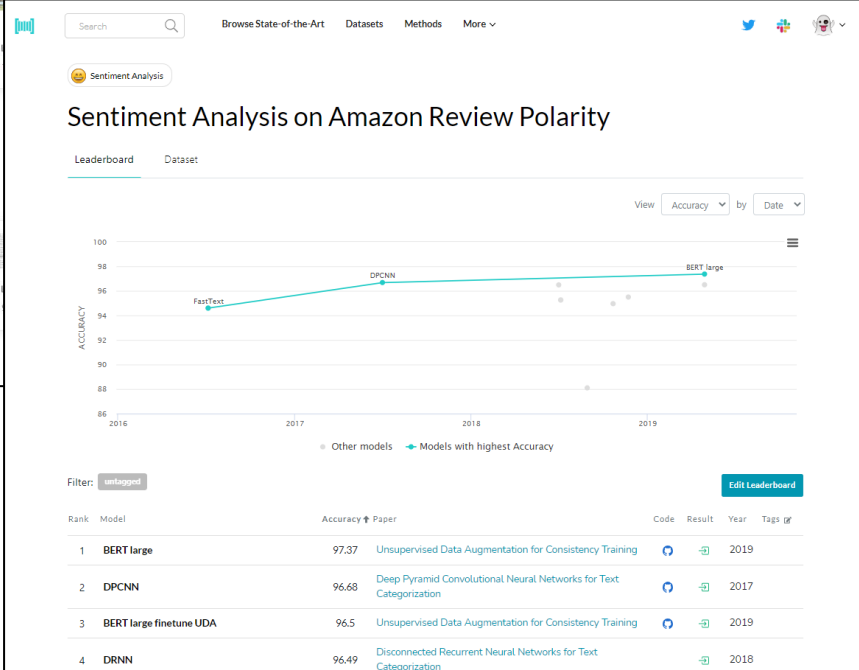
[See all 1460 tasks](#)

Natural Language Processing

- Language Modelling**: 61 benchmarks, 2493 papers with code
- Question Answering**: 192 benchmarks, 1951 papers with code
- Machine Translation**: 90 benchmarks, 1782 papers with code
- Sentiment Analysis**: 88 benchmarks, 1073 papers with code

[See all 681 tasks](#)

<https://paperswithcode.com/sota>



Sentiment Analysis on Amazon Review Polarity

Leaderboard Dataset

View Accuracy by Date

ACCURACY

2016 2017 2018 2019

Other models Models with highest Accuracy

Filter: sentiment Edit Leaderboard

Rank	Model	Accuracy	Paper	Code	Result	Year	Tags
1	BERT large	97.37	Unsupervised Data Augmentation for Consistency Training	Code	Result	2019	
2	DPCNN	96.68	Deep Pyramid Convolutional Neural Networks for Text Categorization	Code	Result	2017	
3	BERT large finetune UDA	96.5	Unsupervised Data Augmentation for Consistency Training	Code	Result	2019	
4	DRNN	96.49	Disconnected Recurrent Neural Networks for Text Categorization	Code	Result	2018	

Some Project Ideas (not binding)

- 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 common 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)
 - predict ratings
- Bundeswahl (done very often)
 - Find rules that help you predict the results
- last.fm Playlist Analysis
 - Cluster last.fm users according to the style of the songs they are listening to
 - Find common 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)*

Dataset Selection: Key Considerations

- Pros

- **Rich Feature Space:** Datasets should have multiple, diverse features that allow for creative feature engineering.
- **Adequate Sample Size:** Aim for datasets with at least 10,000 examples to ensure robust modeling.
- **Balanced Complexity:** A dataset should be complex enough to challenge students without being computationally prohibitive.
- **High Data Quality:** Ensure key columns are well-populated (e.g., <5% missing values) so that the data can be effectively used.
- **Novelty:** Prefer datasets that haven't been overused in existing challenges, offering room for innovative approaches.

Dataset Selection: Key Considerations

- Cons

- **Overly Simple:** Avoid datasets with too few features (< 5) or a too-basic topic, as this limits feature engineering.
- **Excessively Large:** Datasets with over 1 million records (e.g., huge product datasets) can be too compute-intensive.
- **Over-Saturated:** Datasets with clear guidelines and abundant available code (e.g., well-established challenges).
- **Poor Data Usability:** Be wary of datasets where important columns are empty more than 5% of the time, or where the ground truth is ambiguous.
- *Additional Tip:* Check prior usage—if you're the first to work on a dataset, verify that the dataset is practically usable and that data quality issues won't undermine your project.

Team Formation

- You are allowed to form teams of five to six students as you like!
 - You enter your team into the Group Formation Google spreadsheet (see last slide) until Sunday, October 5th 23:59
 - If you are less than five you can still enter your team (but you will be assigned new team members)
 - If you are still looking for a team, enter yourself to the respective section of the spreadsheet also until Sunday, October 5th 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 6th

Team Formation

- Once your team is formed:
 - Meet with your team to organize your work
 - Decide project topic
 - Organize writing of project outline
 - You can start writing the project outline
- For further communication, we will create groups in Ilias
 - Check that you enabled notifications for the message board and your group



Data Mining [V] [1. PG] (HWS 2025)

Inhalt Info Einstellungen Mitglieder Rechte Zum Portal² Portal²-Funktionen

Zeigen Verwalten

Neues Objekt hinzufügen Voransicht als Mitglied aktivieren Seite bearbeiten

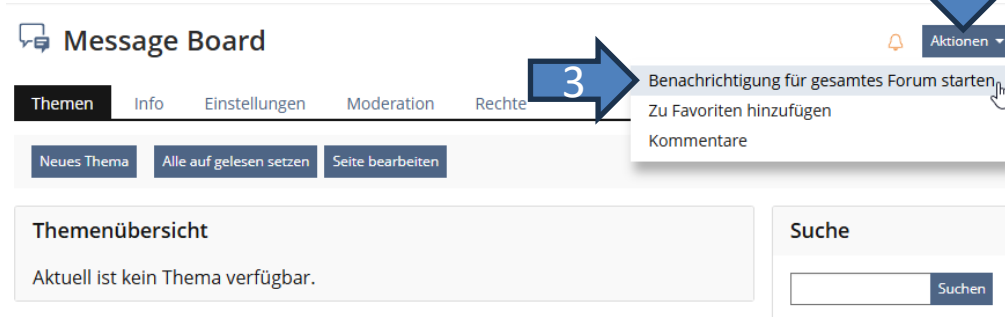
Ordner

Dateien

Foren

Message Board

Beiträge (Ungelesen): 0 (0)



Message Board

Themen Info Einstellungen Moderation Rechte

Neues Thema Alle auf gelesen setzen Seite bearbeiten

Themenübersicht

Aktuell ist kein Thema verfügbar.

Suche

Suchen

Benachrichtigung für gesamtes Forum starten

Zu Favoriten hinzufügen

Kommentare

Project Outlines

- 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 Tuesday, October, 14th, 23:59**
- Submission via Ilias

- On Friday, October, 17th you will receive feedback about your project
 - Including if you need to show up for the first feedback session on October, 20th (lecture time slot)

Project Outlines

- 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

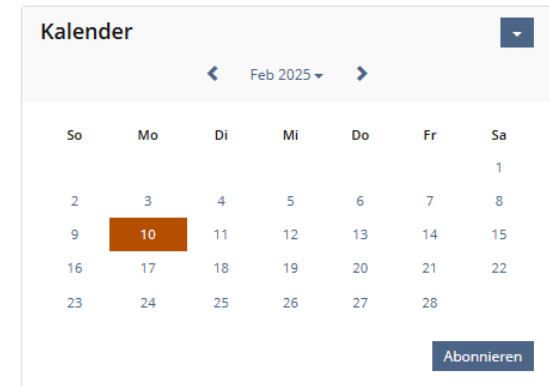
- We will give you tips and answer questions concerning your project
- At the time of the lecture (Mondays)
- **Every team has to attend at least one coaching session!**

Course Organization - Calendar

This semester, we will use the calendar feature in Ilias to schedule **project feedback sessions**.

Project Feedback Sessions:

- You register via the calendar in Ilias for your coaching session
- A few days before your session, we will **inform you via the main forum** about your **exact time slot** within the 90 min window.
- **Only one person per group** should book a slot on behalf of the group.
- When booking, you must include your **group number** and a few **questions or topics** you want to discuss. **Blank requests will be ignored!**
- The registration opens one week before it (Monday, 16:00)



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

Some Project Management Hints

- **Define concrete milestones:** When should what be finished?
 - e.g. 05.11.25 Data exploration results collected in single document
 - e.g. 10.11.25 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 LLMs 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:
 - run automatic hyperparameter optimization and attribute selection
 - employ more sophisticated evaluation setup: x-val + holdout vs. nested x-val

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!
- Reports and additional material need to be uploaded in Ilias within the respective Ilias groups
 - **Deadline: Sunday, November 30th, 23:59**

Project Report

- 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

- 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

- 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

- 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)

Project Presentation

- Present the project results to the other students
 - 10 minutes presentation + 5 minutes discussion
 - During exercise slot
 - Everyone
- Presentations need to be uploaded in Ilias within the respective Ilias groups
 - **Deadline: Wednesday, December 3rd, 23:59**
- Three **90-minute sessions** will be available.
- For **presentations, attendance is mandatory per session** for all group members, so the exact timing within the session does not matter.
- Keep an eye on the **general forum**—we will announce the exact time when slots become available **at least one week in advance**.

Get Additional Advice from a Stanford Professor

- How to evaluate your model?
 - <https://www.youtube.com/watch?v=TxTbIROt9IY>
- 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>



Christopher Potts

Final Exam

- Date: **Monday, 15th December 2025**, time tba.
 - Duration: 60 minutes
 - Location: tba
- Structure: 6 open questions that
 - Check whether you have understood the lecture content
 - We try to cover all major chapters of the lecture
 - Require you to describe the ideas behind algorithms and methods
 - Often: How do methods react to special patterns in the data?
 - Might require you to do some simple calculations for which
 - You need to know the most relevant formulas
 - You do **not** need a calculator
 - There will be at most 1 question containing Python content
 - Should be solvable without a lot of Python knowledge
 - You do not need to know specialized Python functions by heart

Deadlines - Overview

- Team formation until **Sunday, October 5th 23:59**
 - Either enter your whole team or
 - Enter your name if you are looking for a team (team assignment on Monday, October 6th)
- Project outline until **Tuesday, October 14th, 23:59**
- Coaching Sessions
 - Every team has to attend at least one coaching session
- Project report until **Sunday, November 30th, 23:59**
- Project presentation in PDF until **Wednesday, December 3rd, 23:59**

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 Sunday October 5th 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)				
27					
28					
29					
30					
31					
32					

Link in Ilias

Questions?

