Introduction to Student Projects

IE500 Data Mining

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Outline

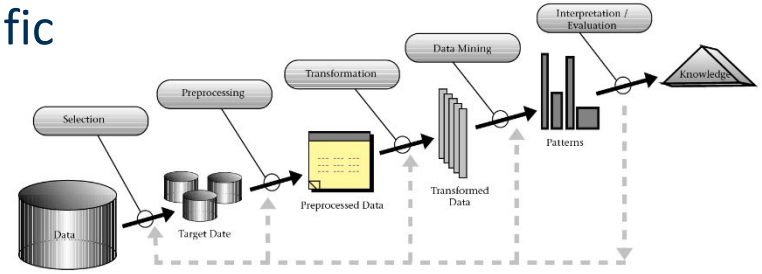
UNIVERSITY OF MANNHEIM
Data and Web Science Group

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

Student Projects

UNIVERSITY OF MANNHEIM
Data and Web Science Group

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

Some Projects Realized in Previous Semesters

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Bundesliga Betting
  - learn classifier to predict outcome of soccer games
- Airbnb (done very often)
  - predict the prices of new apartments
- Analysis of Training Data of a Fitness Center
  - Find different customer groups by clustering exercise data
  - Find frequent combinations of exercises
- Mannheim Police Reports
  - Analyze public police reports to predict the severity of a crime/incident
- Twitter data
  - humor / hate speech detection
  - Sentiment Analysis of Tweets about Movies
    - Learned classifier from IMDB movie reviews
    - Applied and tested with tweets afterwards
- Sentiment Analysis of Tweets about Movies

Some Projects Realized in Previous Semesters

- Bundesliga Betting
  - learn classifier to predict outcome of soccer games
- Airbnb (done very often)
  - predict the prices of new apartments
- Analysis of Training Data of a Fitness Center
  - Find… …ns by clustering exercise data
  - Find…
- Mannheim Police Repo…
  - Analyze public police reports to predict the severity of…
- Twitter data
  - humor / hate speech detection
  - Sentiment Analysis of Tweets about Movies
    - Learned classifier from IMDB movie reviews
    - Applied and tested with tweets afterwards
- Sentiment Analysis of Tweets about Movies

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

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
- Predict Prices of Used Cars
  - use existing benchmark datasets, learn which features matter most
- SPAM Detection
  - eMail, blog or discussion forum (Bing Liu 6.10, 11.9)
  - You Tube comments

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 competition 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

- 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/
  - 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?

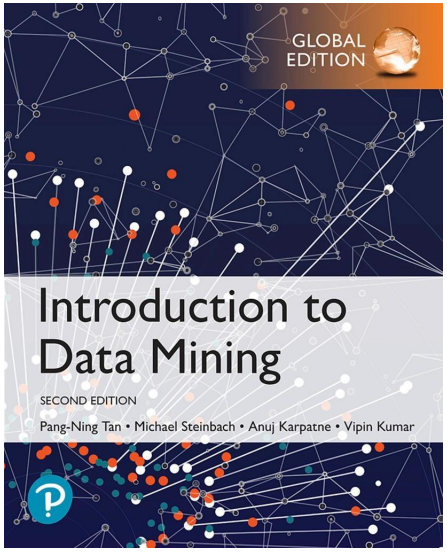
UNIVERSITY OF MANNHEIM
Data and Web Science Group

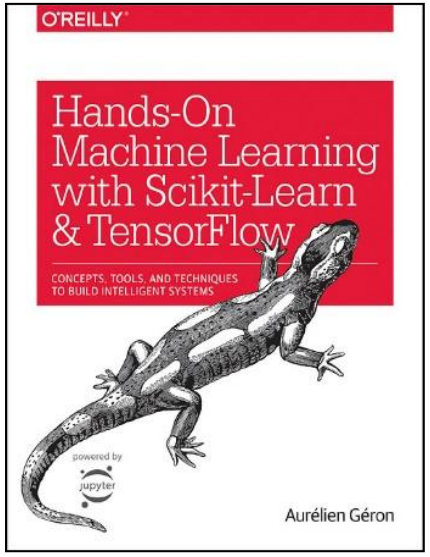
- Public sector data
  - US government: https://www.data.gov
  - UK government: https://data.gov.uk
  - EU: https://www.europeandataportal.eu
  - CIA World Fact Book: https://www.cia.gov/library/publications/the-world-factbook/
  - Health data (over 125 years): https://www.healthdata.gov/

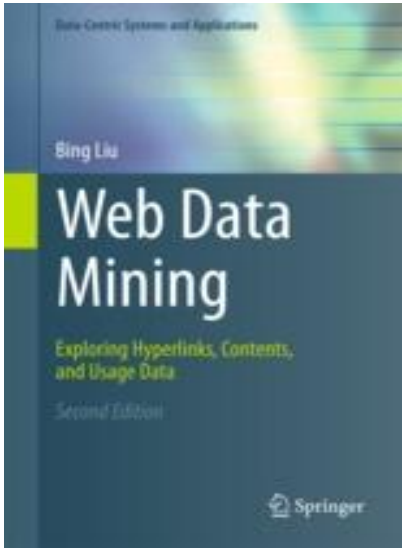
Where to Find Relevant Methods

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 Relevant Methods

UNIVERSITY OF MANNHEIM
Data and Web Science Group

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

Google

Google Scholar

kaggle

DATA MINING CUP
International Student Competition

Where to Find Relevant Methods

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Discuss your topic/task with ChatGPT or Claude
  - Ask for benchmark datasets for this specific task
  - Ask which methods are successfully used in literature
  - Ask for data preparation steps that are often used for the task
  - Ask for references to current papers on the topic

Dataset Selection: Key Considerations - Pros

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **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 Completeness:** 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **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.

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, April, 12th, 23:59**
- Submission via Ilias

- On **Tuesday, April, 14th** you will receive feedback about your project via mail
  - Including if you need to show up for the feedback session on April, 15th (lecture time slot)

Project Outlines

UNIVERSITY OF MANNHEIM
Data and Web Science Group

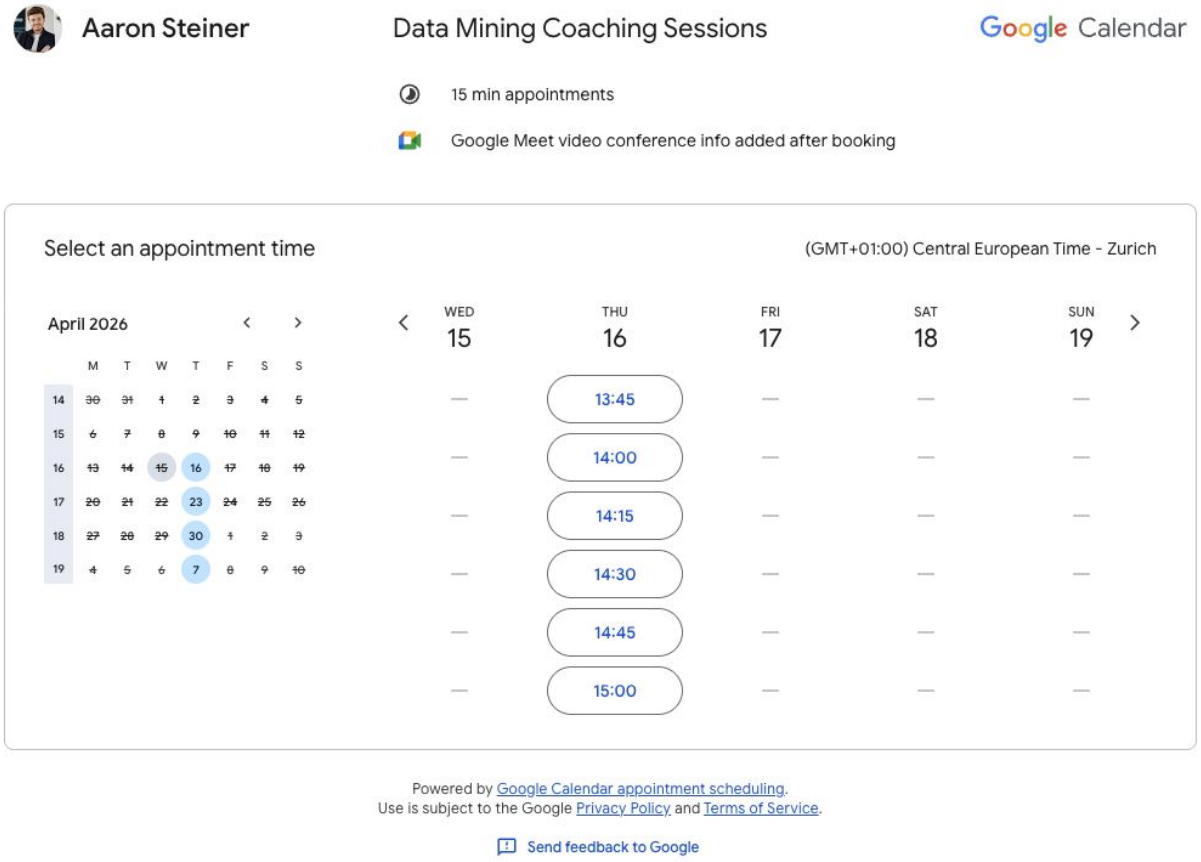
- Answer the following questions:
  1. What is the problem you are solving?
  2. What data will you use?
     - Where will you get it?
     - How will you gather it?
  3. How will you solve the problem?
     - What preprocessing steps will be required?
     - Which algorithms do you plan to use? Be as 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 give you tips and answer questions about your project.
- At the time of the exercise (Thursday 13:45-15:15)
- **Registration** is mandatory if you want coaching!
- **Every team has to attend at least one coaching session!**
- Make sure to register until Wednesday (23:59) of the week you want to attend the coaching session
- Each coaching session lasts for 15 minutes
  - Include your questions when booking the session, so we can prepare!
  - Most time efficient use of the session
  - We will also answer any question you pose directly in the session

Booking Coaching Sessions: How-to



https://calendar.google.com/calendar/u/0/appointments/schedules/AcZssZ2VIQTrXZOcLM5YXCbmcM54kLrbEM39dAgUt-_nyW3g7kdaxlzxru0JFeMhHOPVghA4-UdBY084

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. 25.04.26 Data exploration results collected in single document
  - e.g. 30.04.26 Subgroup on feature creation 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 (e.g. in VS Code)

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

2. Project Report

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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, May 17th, 23:59**

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 and Data Mining
    - describe different approaches and parameter settings/optimizations that you tried
  - Evaluation
    - description of evaluation setup (split, x-val, nested-x-val?)
    - including an analysis of the errors still made by the best method
  - Results
    - presentation and discussion of the results
    - comparison to state-of-the-art results (minimum 0.5 pages)

Project Report

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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: Thursday, May 21st, 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

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

3. Final Exam

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Date: Thursday**, 11th June 2026**, time tba.
  - Duration: 60 minutes, Location: tba
- Content:
  - content of all slide sets of the offline lecture
  - the following online lectures: Ensembles, Comparing Classifiers, Time Series, Hierarchical Clustering
- 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

Deadlines - Overview

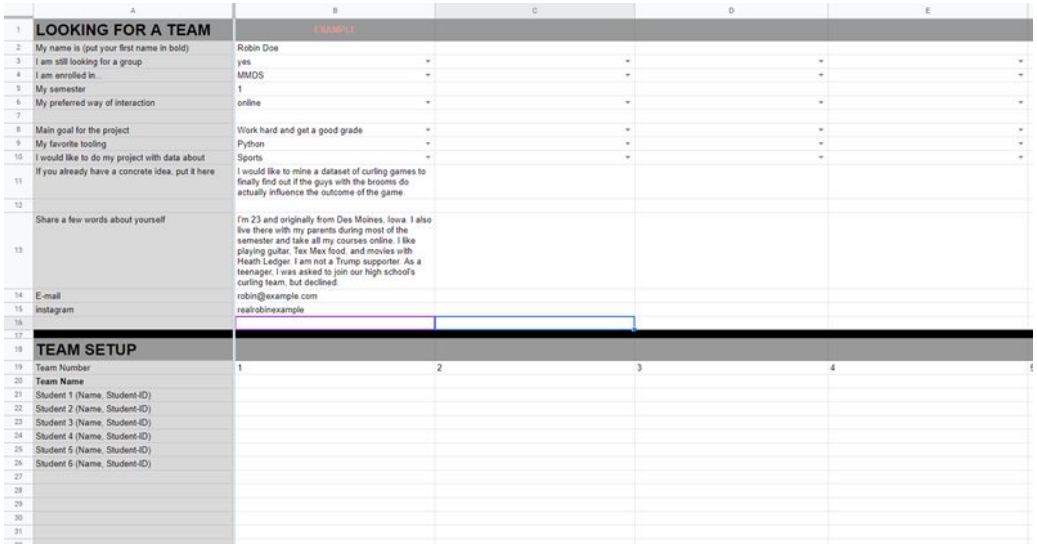
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Team formation until **Sunday, March 22nd 23:59**
  - Either enter your whole team or
  - Enter your name if you are looking for a team (team assignment on Tuesday, March 24th)
- Project outline until **Sunday, April 12th, 23:59**
- Coaching Sessions
  - Every team has to attend at least one coaching session
- Project report until **Sunday, May 17th, 23:59**
- Project presentation in PDF until **Thursday, May 21st, 23:59**

4. Team Formation

- 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



	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					

https://docs.google.com/spreadsheets/d/1ecSkGDGqZoFaRx8DlDENQvKa3ORQpuXVCvr4mW-63cw/edit?usp=sharing

Team Formation

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- You are allowed to form teams of **six students** as you like!
  - You enter your team into the Group Formation Google spreadsheet (see last slide) until **Sunday, March 22nd 23:59**
  - If you are less than six 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, March 22nd 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 **Tuesday, March 24th**

Thank you!

UNIVERSITY OF MANNHEIM
Data and Web Science Group