School of Computer Science

University of St.Gallen

Data Science: Exercise 1

Bernhard Bermeitinger, Thomas Huber

26.09.2023

Overview

- 2x 45 minutes
- Content distribution via the **Lecture** canvas page
- Exercises accompany the lectures
- Discussions about lecture topics
  - In the audience
  - In small groups
- Practical coding
  - Exemplifying abstract lecture content
  - “Real-world” examples

Supported Tools

- There will be some programming tasks
- We recommend and support the use of Jupyter Notebooks or Google Colab (requires a Google account)
- If you are completely new to Python then Google Colab is the easiest way to get started
- Python version: 3.10+ (Colab runs 3.10.2)
- You can use other tools but we may not be able to help with technical issues arising from your choice of IDE

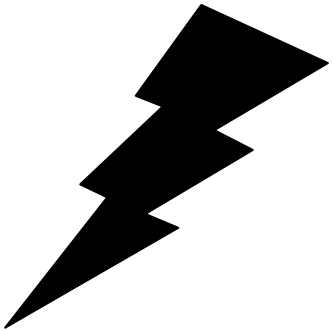
Task 0 - Prior Knowledge Test

- Go to the Canvas page of the course (the lecture, not the exercises) and do the Data Science Prior Knowledge Quiz
- This is not graded and exists only to get an overview of prior knowledge
  - *30 minutes*
- If you finish early: https://aclanthology.org/events/acl-2023/
  - Check some of the papers (briefly) for whether or not they include information about how to reproduce their results (This ties into / is related to Task 2.1)

Task 1.1 - Spurious Correlations

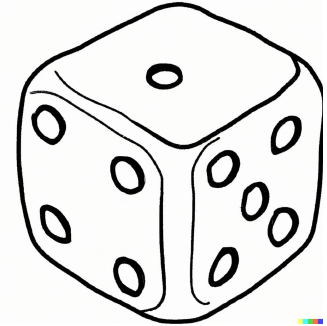
- As a team of 3, go to *Spurious Correlations*
  - Spurious Correlations https://www.tylervigen.com/spurious-correlations
  - Select one graph and discuss what it's about
  - *5 minutes*

Task 1.2 - Spurious Correlations



- Present your graph to the class
  - *1 minute*

Task 1.3 - Spurious Correlations



- Provide your own example of a spurious correlation that could be found when analyzing data. Explain why it is spurious.
  - *2 minutes*

Task 2.1 - “Many Analysts, One Data Set”

- Read the paper (shallow)
  - R. Silberzahn et al. “Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results”, 2017,
  - https://journals.sagepub.com/doi/full/10.1177/2515245917747646
  - *15 minutes*

Task 2.2 - “Many Analysts, One Data Set”

School of Computer Science

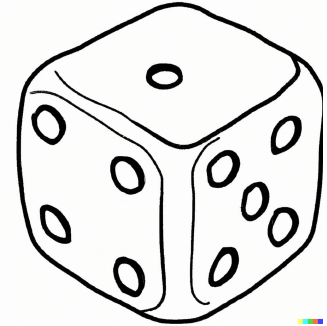
Universität St.Gallen

- In a team of 3: Find a 1-3 papers/articles that cite this paper
  - Why do they cite the paper / Why do they say about it?
  - Make notes
  - *15 minutes*
- Duplicates are allowed (but not actively encouraged)

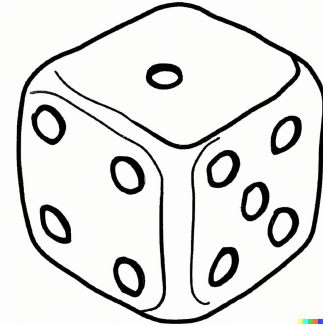
Task 2.3 - “Many Analysts, One Data Set”

School of Computer Science
Universität St.Gallen

- Present your notes about the papers/articles citing the “Many Aspects…” paper
  - *each 5 minutes*



Task 3 - “Many Analysts, One Data Set” - Discussion



- What is correlation between annotators and why do we need it?
  - *1 minute*
- What is p-hacking?
  - *1 minute*
- What is your opinion about the paper and especially the resources they released?
  - *5 minute discussion round*

Task 3.1 - Data Mining

- Split into four groups, one for each of the following data mining techniques:
  - association rule mining
  - clustering
  - classification
  - regression
- Find (or construct) an additional example of that technique being applied in a real-world scenario.
  - *5 minutes*

Task 3.2 - Data Mining

- Split off into new groups
- All new groups should have (at least) one expert of the four techniques
- Briefly present your technique and example in the group
  - *2 minutes per technique*