



«Human-Centered Data Science»

Exercise 4

Lars Sipos

Human-Centered Computing, Institute of Computer Science Freie Universität Berlin

17.05.2022





After lecture Get-together at Luise

Possible times:

- » 02.06.22
- » 09.06.22
- » 23.06.22

Let's vote!

» 09.06.22







Previously: Assignment 1 Review

Get together in your group

What to consider:

- » Go through the tasks and discuss how you solved them
- » Are there any open questions?
- » Was there anything that particularly caught your eye?

After you reviewed each other, at your group's name to the done.txt file in:

https://github.com/FUB-HCC/hcds-summer-2022/tree/main/assignments/A1_R1_WarmUp/done.txt







Assignment 2 - Peer-review

Assigned groups will review each other

After you checked that the other group has actively done the assignment:

- » Add the other group to the *done.txt* file https://github.com/FUB-HCC/hcds-summer-2022/tree/main/assignments/A2_Data/done.txt
- » Sign with your group name, e.g. group 1 (reviewed by group 6)

We will now decide the groups...





Assignment 2 - Peer-review

Get together in your group and meet up with your partner group

What to consider:

- » Go through the tasks and discuss how you solved them
- » Are there any open questions?
- » Was there anything that particularly caught your eye?

Don't forget to update the done.txt file!





About the dataset...

South German Credit Data: Correcting a Widely Used Data Set

Ulrike Grömping
29 November 2019

1 Abstract

The widely used German credit dataset from the UCI Machine Learning Repository, donated by the German professor Hans Hofmann via the European Statlog project, comes with an incorrect code table. Many variables are wrongly represented, which implies that the data cannot be adequately used for experimenting with methods for interpretable machine learning. This note provides details on data generation and a correction of the code table, derived from German language literature. Correct data have been provided to the UCI Machine Learning Repository under the name "South German Credit" (https://archive.ics.uci.edu/ml/datasets/South+German+Credit).

Grömping, U. (2019). South German Credit Data: Correcting a Widely Used Data Set. Report 4/2019, Reports in Mathematics, Physics and Chemistry, Department II, Beuth University of Applied Sciences Berlin.







«Human-Centered Data Science»

Assignment 3

Reproducibility

https://github.com/FUB-HCC/hcds-summer-2022/wiki/04_exercise_A3





What should you consider in your data science practice?

1. Support using your data

Make your collected data freely available by considering the principles of FAIR data.

2. Support sharing your research

License your data/scripts/software.

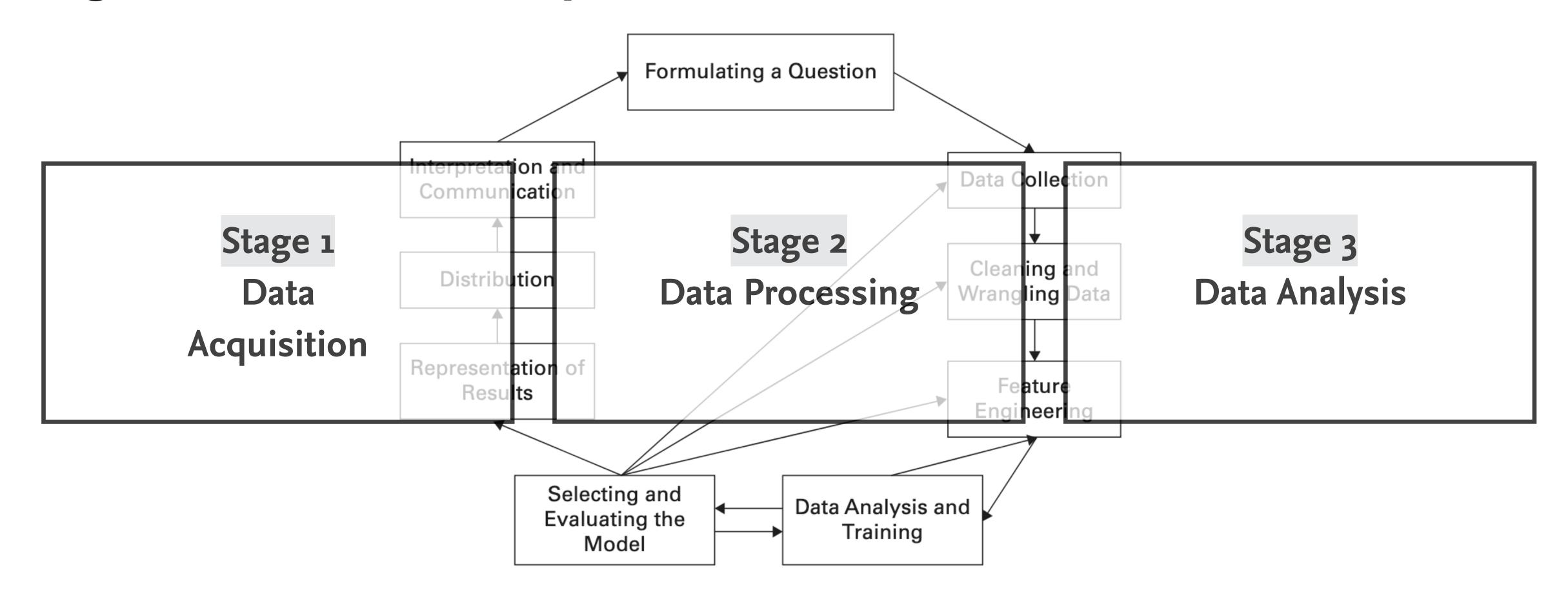
3. Support understanding your research

Prepare your methods and the entire process of data analysis as reproducible as possible and provide relevant documentation.







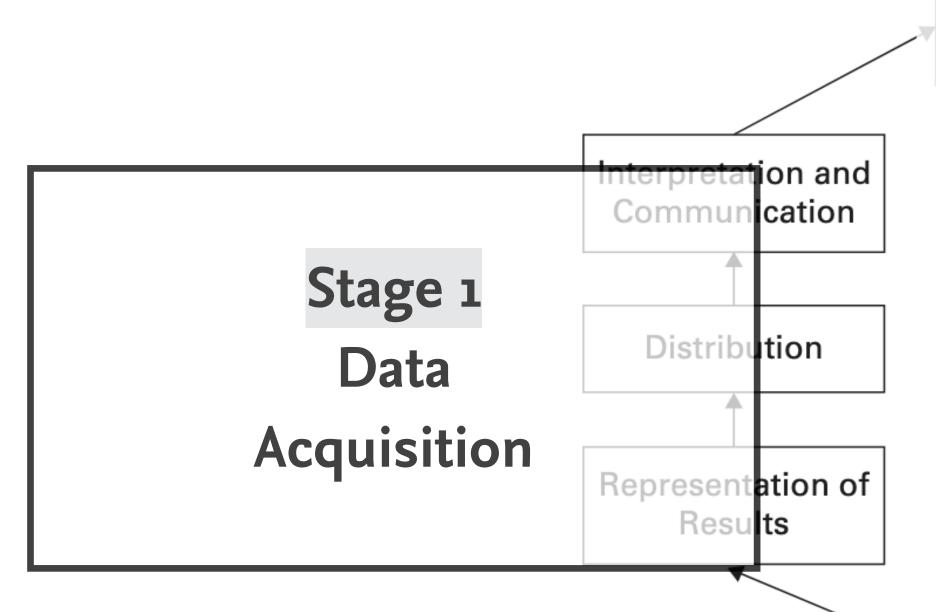


Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2018). The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. Oakland, CA: University of California Press.









- ✓ Where is your data coming from?
- ✓ What licenses apply to the collected data?
- ✓ Who collected your data?

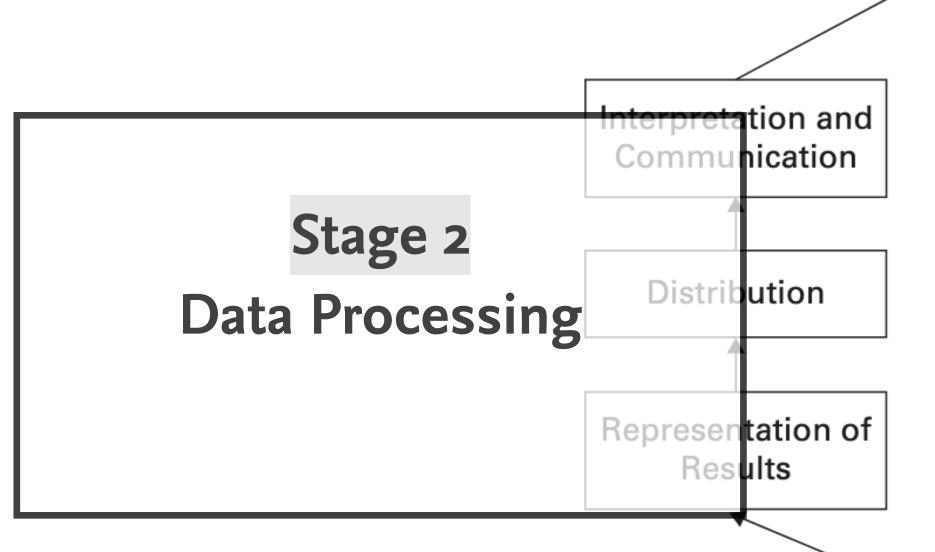
Formulating a Question

- ✓ What client-side tools/environments were used to collect your data?
- ✓ If your data is a sample, what were the parameters?
- √ What features are described in your data?
- ✓ Is a local copy of your source data available?
- Selecting Are there known errors, inconsistencies, or incompleteness in your Model source data?









- ✓ What tools/libraries/environments were used in the processing of your Fordata?ng a Question
- ✓ What sub-sampling, filtering, aggregation, or transformation steps were performed?
 Data Collection
- ✓ What order were processing steps performed in?
- ✓ Why were these processing steps performed?
- ✓ How were errors, inconsistencies, or incompletes discovered, and how were they addressed?
 Feature
 Engineering
- ✓ Did your data processing involve any manual (i.e. non-programmatic)

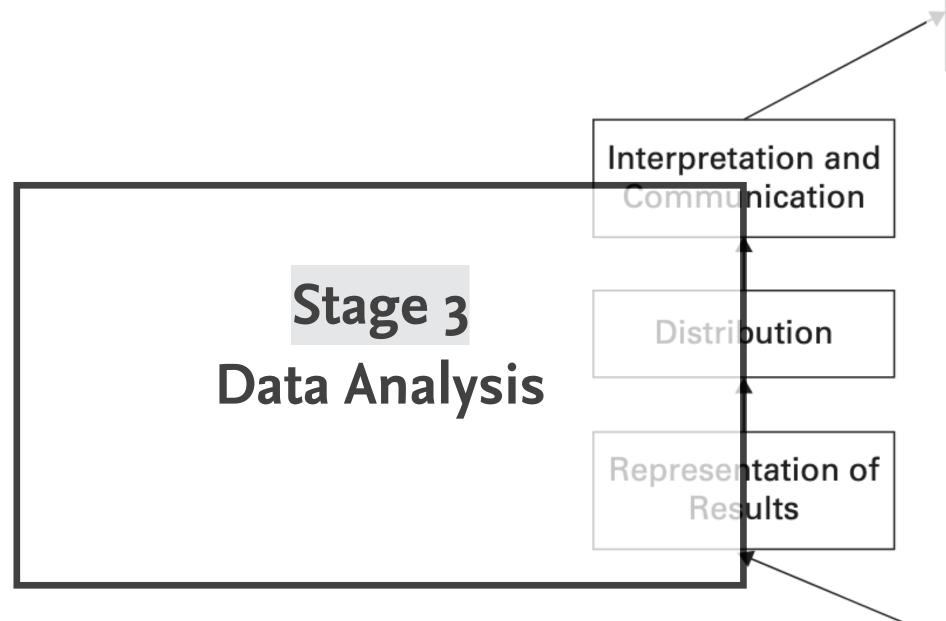
Selecting and the Data Analysis and Evaluating the Training Model Are you making incremental datasets available?

- ✓ Are you making a final processed dataset available?









- ✓ What are the goals of your analysis?
- ✓ What is the nature of your analysis?
- ✓ What assumptions about your data are required for your analysis?
- ✓ What tools/libraries/environments were used in the analysis of your data?

 Cleaning and Wrangling Data
- ✓ What order were analysis steps performed in?
- √ Why and how was each analytical step performed?
- ✓ Are you making samples, demos, or test sets available?
- Selecting and Evaluat of How are the results of your analysis presented?
 - ✓ Are you making a final analyzed dataset available?







Three Key Practices for your Reproducible Workflow

- 1. Clearly separate, label, and document all data, files, and operations that occur on data and files.
- 2. Document all operations fully, automating them as much as possible, and avoiding manual intervention in the workflow when feasible.
- 3. Design a workflow as a sequence of small steps that are glued together, with intermediate outputs from one step feeding into the next step as inputs.





Next Time

you will have ...

- 1. actively participated in the lecture
- 2. worked on the third programming assignment

Have fun!

