Human-Centered Computing
Research Group

Freie Universität Berlin

«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!

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!

Assignment 2 Insights

Assignment 2 Insights ctd.

«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**

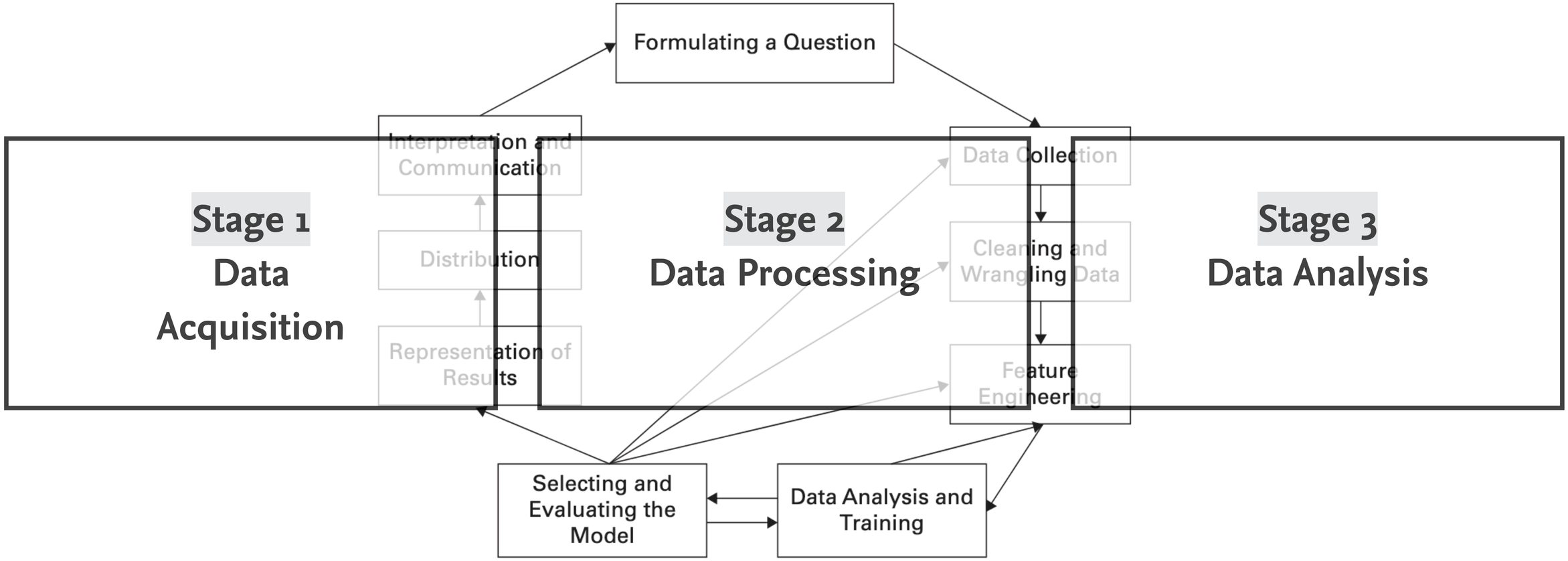
   License your data/scripts/software.

3. **Support understanding your research**

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

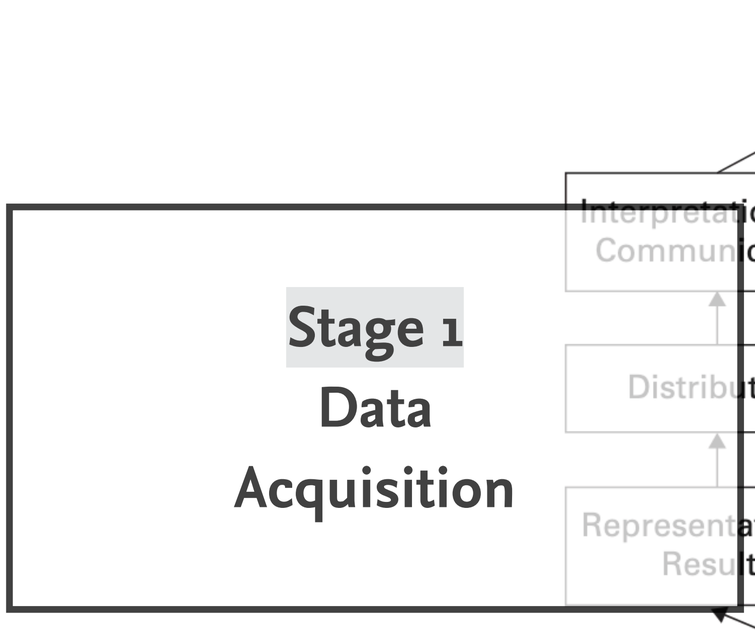
Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512)

Stages of the Basic Reproducible Workflow



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.

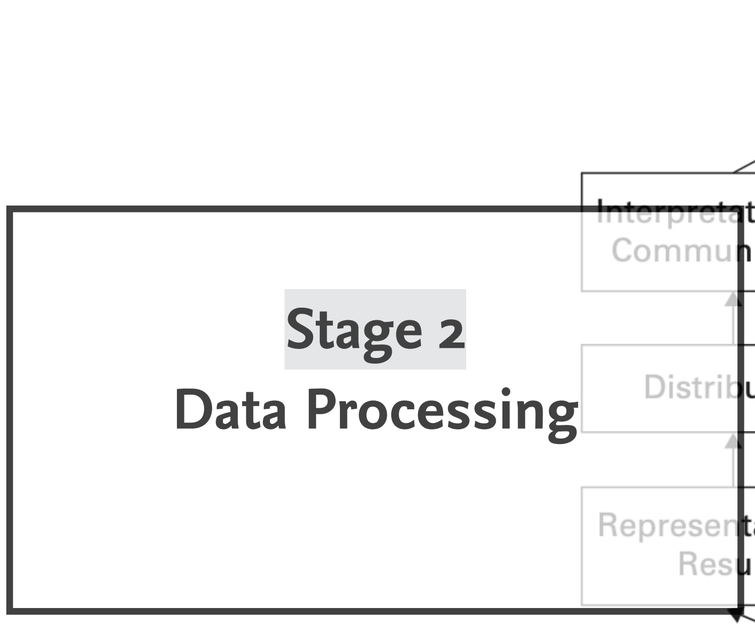
Stages of the Basic Reproducible Workflow



- ✓ Where is your data coming from?
- ✓ What licenses apply to the collected data?
- ✓ Who collected your data?
- ✓ 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?
- ✓ Are there known errors, inconsistencies, or incompleteness in your source data?

Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512)

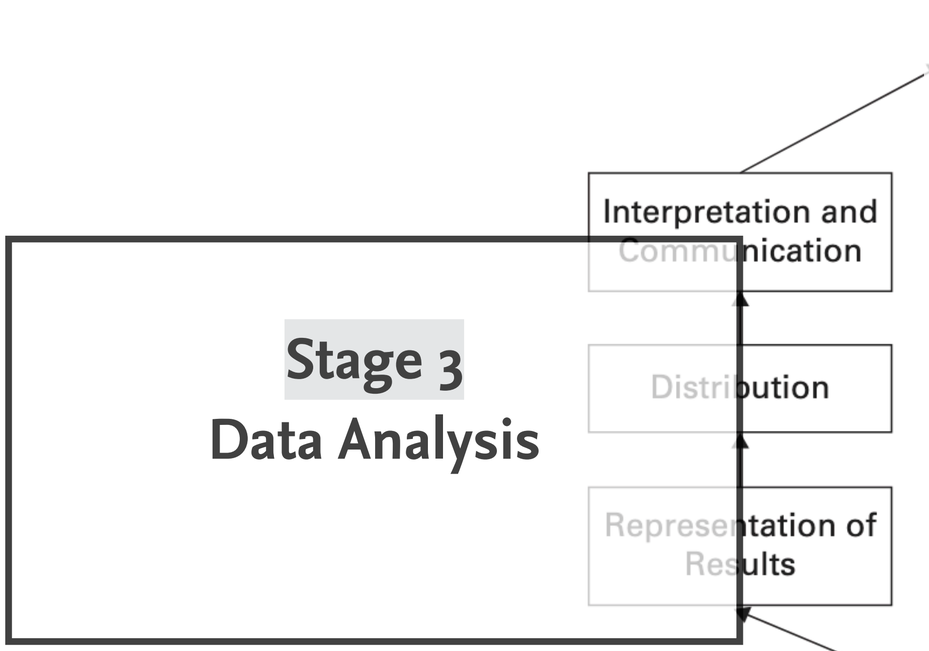
Stages of the Basic Reproducible Workflow



- ✓ What tools/libraries/environments were used in the processing of your data?
- ✓ What sub-sampling, filtering, aggregation, or transformation steps were performed?
- ✓ 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?
- ✓ Did your data processing involve any manual (i.e. non-programmatic) steps?
- ✓ Are you making incremental datasets available?
- ✓ Are you making a final processed dataset available?

Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512)

Stages of the Basic Reproducible Workflow



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

Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512)

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.

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.

Next Time

you will have …

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

Have fun!