Human-Centered Computing
Research Group

Freie Universität Berlin

«Human-Centered Data Science»

Exercise 2

Lars Sipos

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

03.05.2022

Changes to the programming assignments

No more mandatory reflection assignments for the exercises!

In order to actively participate in this course, you need to fulfil the following requirements:

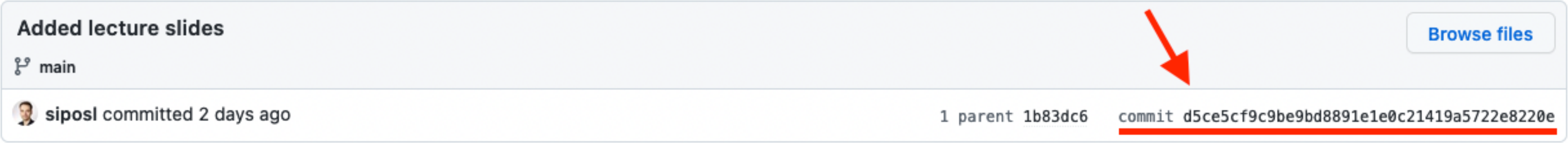
- ~~You need to submit **(n-1) written reflections** and actively do them [planned are 11]~~
- You need to submit **(n-1) scheduled (programming) assignments** and actively work on them [planned are 6]

Each actively ~~done reflection /~~ assignment gives you **1 point**. You need **(n-1) points** for each submission type.

- Please commit a **feedback.txt** file to your assignment folder!

Changes to the programming assignments

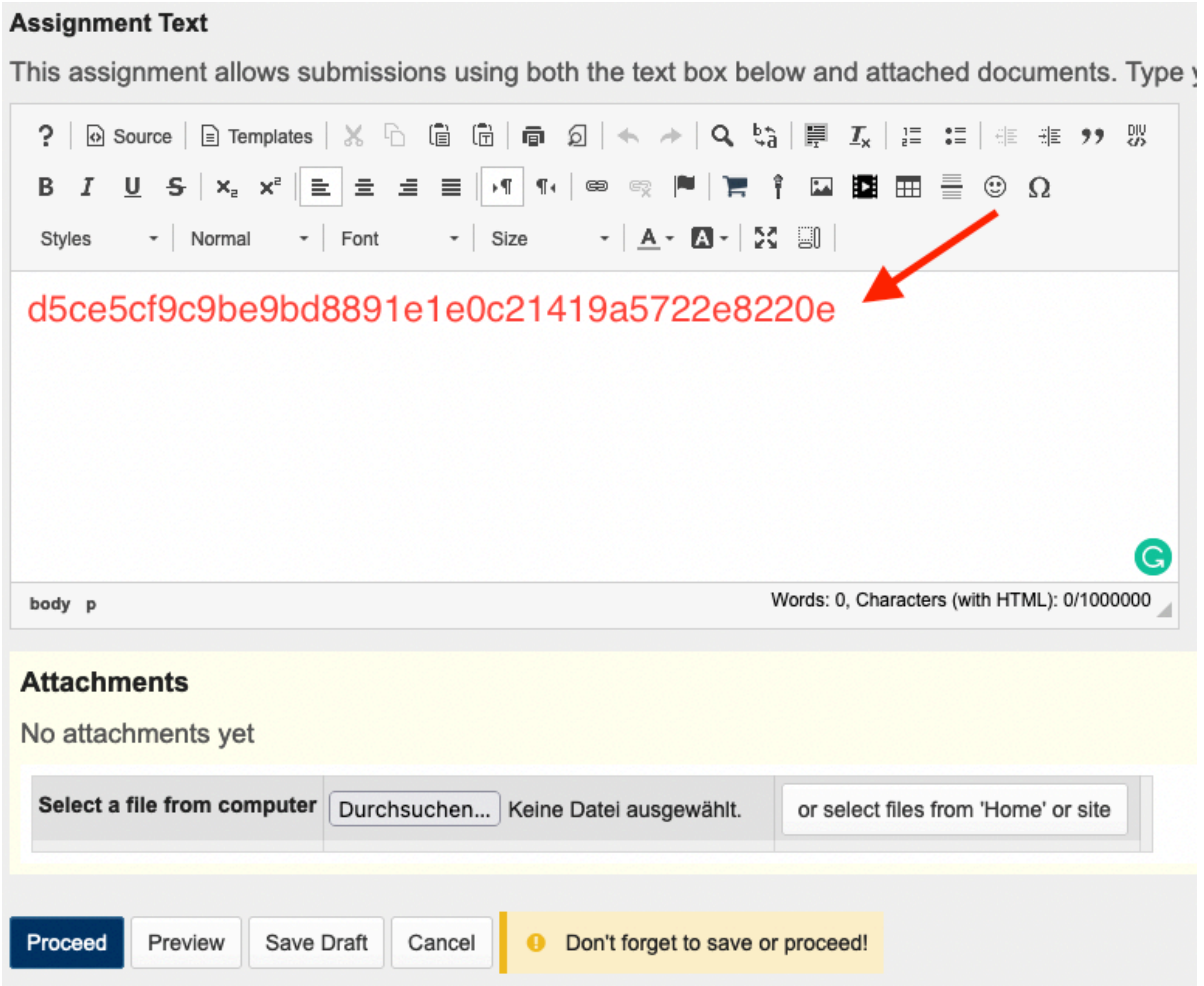
You need to submit your group’s commit hash to Whiteboard!



Added lecture slides

siposl committed 2 days ago

1 parent 1b83dc6 commit d5ce5cf9c9be9bd8891e1e0c21419a5722e8220e



Assignment Text

This assignment allows submissions using both the text box below and attached documents. Type y

d5ce5cf9c9be9bd8891e1e0c21419a5722e8220e

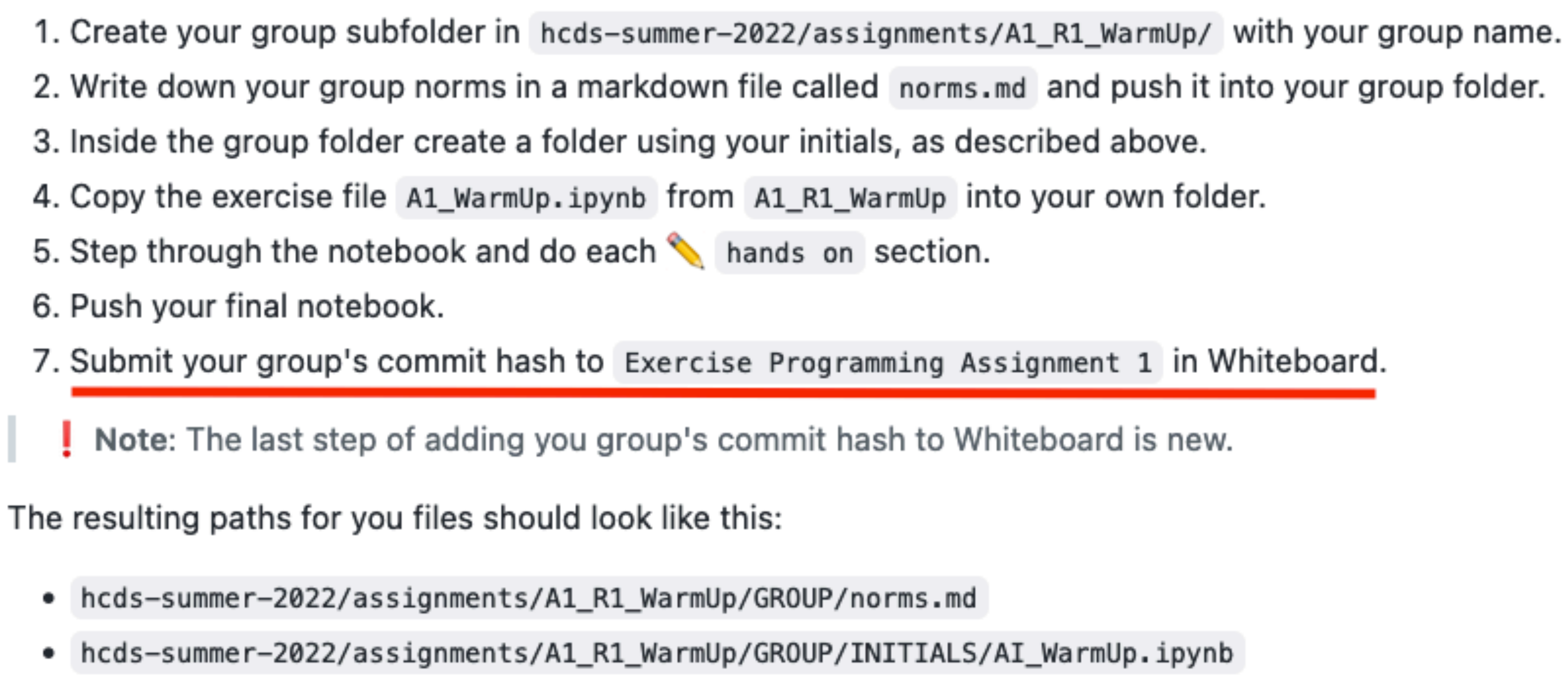
Words: 0, Characters (with HTML): 0/1000000

Attachments

No attachments yet

Select a file from computer Durchsuchen... Keine Datei ausgewählt. or select files from 'Home' or site

Don't forget to save or proceed!



1. Create your group subfolder in `hcds-summer-2022/assignments/A1_R1_WarmUp/` with your group name.
2. Write down your group norms in a markdown file called `norms.md` and push it into your group folder.
3. Inside the group folder create a folder using your initials, as described above.
4. Copy the exercise file `A1_WarmUp.ipynb` from `A1_R1_WarmUp` into your own folder.
5. Step through the notebook and do each ✏️ `hands on` section.
6. Push your final notebook.
7. Submit your group's commit hash to `Exercise Programming Assignment 1` in Whiteboard.

> ❗ **Note:** The last step of adding you group's commit hash to Whiteboard is new.

The resulting paths for you files should look like this:

- `hcds-summer-2022/assignments/A1_R1_WarmUp/GROUP/norms.md`
- `hcds-summer-2022/assignments/A1_R1_WarmUp/GROUP/INITIALS/AI_WarmUp.ipynb`

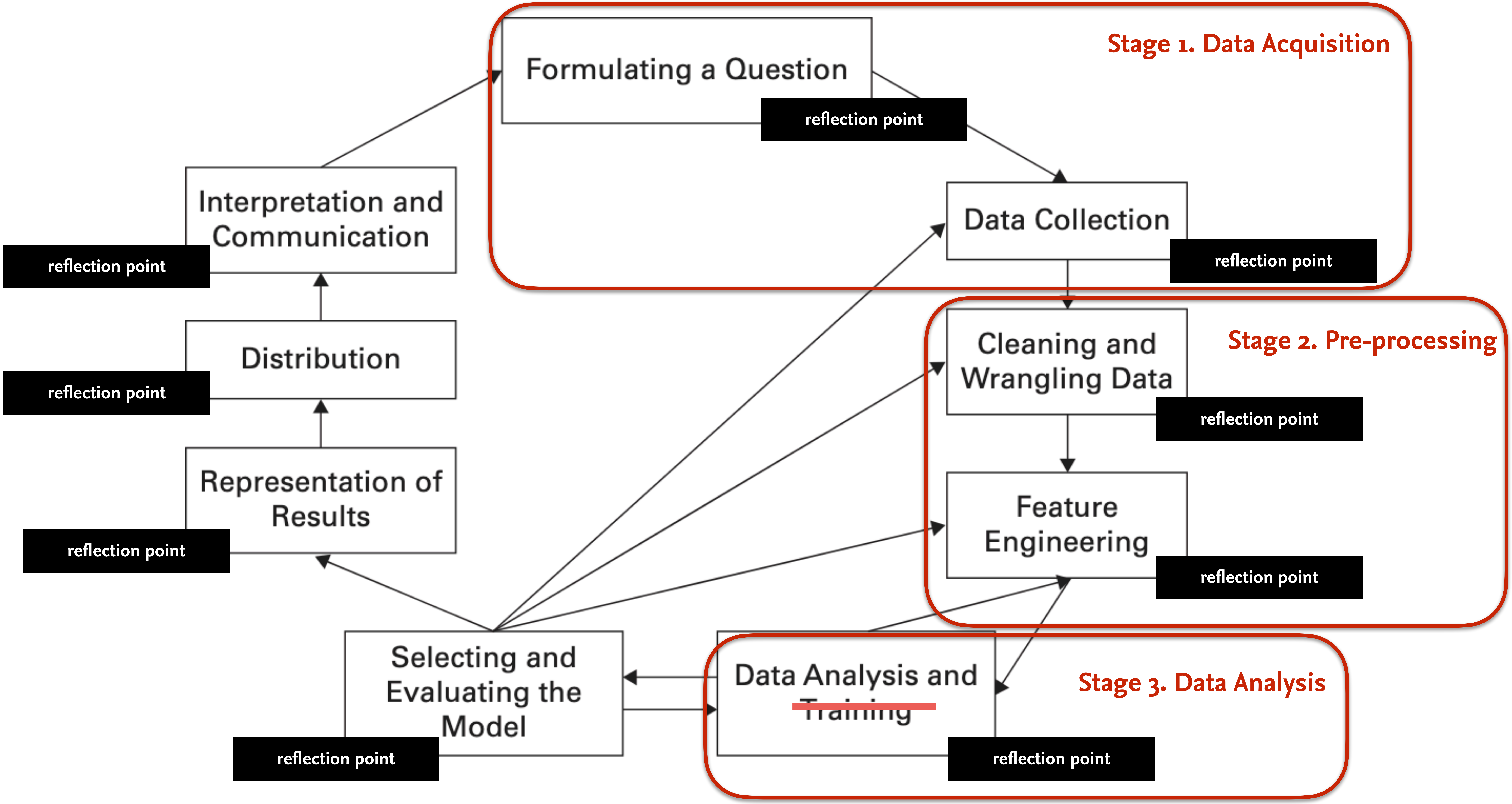
(Your view may differ)

«Human-Centered Data Science»

Assignment 2

Data Acquisition, Pre-processing and Data Analysis

https://github.com/FUB-HCC/hcds-summer-2022/wiki/02_exercise



Stage 1. Data Acquisition

Formulating a Question

reflection point

Data Collection

reflection point

Stage 2. Pre-processing

Cleaning and Wrangling Data

reflection point

Feature Engineering

reflection point

Stage 3. Data Analysis

Data Analysis and ~~Training~~

reflection point

Selecting and Evaluating the Model

reflection point

Representation of Results

reflection point

Distribution

reflection point

Interpretation and Communication

reflection point

First stages of the data science workflow

Consider the first three stages of a typical DS workflow (data acquisition, processing, analysis)

For each stage, think about:

» Your prior experiences with it

» What are you doing there?

» How do you do it?

» What are you not supposed to do?

Get into your groups!

We get back together at: 11:05 a.m.

Data Acquisition

What to do:

- Precise formulation of your question
- Look at the web (Google, Kaggle, Github, Iris Privacy Concern)
- Understanding how the data was generated
- Reputation of data source
- Properly design your survey

What *not* to do:

- Don’t forget to acquire what have previously decided on
- No mishandling of data (e.g. be mindful with who you share it with)
- Don’t overload the source with traffic (i.e. when you are crawling data)

(Result of the group discussion)

Pre-processing

What to do:

- Clean up missing values, outliers
- Think about dropping rows or not
- Think about introducing protected attributes
- Splitting data into training and test
- Scale, normalize the data properly
- Think about your data structures (e.g. sparse matrices)
- Think about feature selection and dimensional reduction
- Packages: Numpy, Pandas, Scikit-learn, Pyspark

What *not* to do:

- Don't just use the mean of numerical data (e.g. replacing missing values with it)
- Don't introduce more biases
- Don't give up
- Don't rush through it (spend about 70-80% of your time on it)
- Don't obfuscate your data

(Result of the group discussion)

Data Analysis

What to do:

- Look at rows with missing values; is there are pattern?
- Think about your axes
- Look at the outliers
- Visualize your dataset
- Plotting: Matplotlib, Plotly
- Try to understand your data (e.g. patterns, relationships)

What *not* to do:

- Axes: Don't cut them off
- Don't fool yourself or others (e.g. don't present your data to suite your question)

(Result of the group discussion)

«Human-Centered Data Science»

Assignment 2

Data Acquisition, Pre-processing and Data Analysis

https://github.com/FUB-HCC/hcds-summer-2022/wiki/02_exercise

Next Time

you will have …

1. actively participated in the lecture
2. submitted the first lecture reflection (Due 05.05.22 4 p.m.)
3. submitted the first programming assignment (Due 10.05.22 10 a.m.)
4. survived last week (and hopefully enjoyed it)

Have fun!