- Registration in QIS: everyone who wants to participate in the exams has to register in QIS
- Exam Date 1: Monday 18th July 2022 13:00 to 15:00 Hörsaaltrakt Bockenheim - H IV
- Exam Date 2: Tuesday 23rd August 2022 13:00 to 15:00 Hörsaaltrakt Bockenheim - H IV
- Present your outcomes of the following programming task in class (choose a time slot in Moodle)

TOPIC 4
STREAM MINING: One-hot encoding and DGIM

Prepare some presentation slides to present the following steps:

1) Briefly describe the concept of One-hot encoding in your own words (1 slide).

2) Briefly describe the DGIM algorithm (see Chapter 4.6 of MMDS book) (1 slide).

3) From a public repository (see list below), choose a dataset for a medical use case with a large amount of instances that you can apply stream mining on and give a description of the dataset (1 slide).

4) Choose a column of the dataset chosen under 3) such that there are many repetitions of values in the column. Apply one-hot encoding to the column (thus, generating several new columns).

5) Choose a public implementation[1] of the DGIM algorithm and apply it on one of the columns generated under 4) such that the column and the window length *N* can be flexibly set as parameters. Report on the implementation (1 slide).

6) Implement a simple web frontend (e.g. using streamlit or svelte) to visualize the data set, set parameters (like the column to execute DGIM on and the window length *N*) and display the 1-count for the selected column (from 5). Compare the 1-count obtained by DGIM to the exact 1-count in the selected column. Show some sample screenshots (1 to 3 slides).

7) Create your own repository at github.com and commit your source code. Put your github link and any other references you have used on a References slide (1 slide). Submit your programming task slides in Moodle.

8) Choose a date in Moodle for your slide presentation in class.

Datasets:

1. UCI ML Repository Subject Area Life Science and large instance count https://archive-beta.ics.uci.edu/ml/datasets?f%5Barea%5D%5B0%5D=life-sciences&f%5Binstances%5D=greater-than-thousand&p%5Boffset%5D=10&p%5Blimit%5D=10&p%5BorderBy%5D=NumHits&p%5Border%5D=desc

1 You can also decide to implement the chosen algorithm on your own