- 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 3
STREAM MINING: Bloom Filter and Count-distinct

Prepare some presentation slides to present the following steps:

1) Briefly describe the concept of Bloom filter in your own words (1 slide).

2) Briefly describe the Flajolet-Martin algorithm or a more recent alternative method (see hint below) (1 slide).

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

4) Generate a large stream of unique values (for example the instance ID or row ID from the dataset chosen under 3)) and implement a Bloom filter such that the length *n* can be flexibly set as a parameter. Implement a query module such that for any query key the Bloom filter reports a "miss" or a "match". Report on the implementation (1 slide).

5) Choose a public implementation[1] of the algorithm chosen under 2); choose a column of the dataset chosen under 3) such that there are many repetitions of values in the column. Apply the chosen count-distinct implementation on it. Report on the implementation (1 slide).

6) Implement a simple web frontend (e.g. using streamlit or svelte) to visualize the data set and set parameters (like Bloom filter length *n* and the amount of key values to insert into the Bloom filter). For the Bloom filter, the web frontend should accept as input an arbitrary query key and display whether this is a true of false positive or a true negative. For the count-distinct problem, the web frontend should display the count obtained by 5) and compare it to the real amount of distinct values. 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.

Hint:
Probabilistic Algorithms: Approximate Counting, *LogLog and Bloom Filters by Lucas Schmidt
https://github.com/lucasschmidtc/Probabilistic-Algorithms/blob/master/Probabilistic%20Algorithms.ipynb

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