Project Concepts of Data Science  
2023-2024

**Description**  
  
1. The development of the software should be done using version control hosted on GitHub or Gitlab. If your repository is private, do not forget to give me access. You should also follow up on whether the invitation has been accepted, since such invitations easily get lost. Since you are  
working in teams of two, GitHub/GitLab is also your tool to collaborate within the team. You are expected to collaborate while developing the code, so that should be reflected by the commit messages for the repository. If there are no or few commit messages by a team member, I  
will conclude that this person contributed little to the code and that will be reflected by that person’s grade. The repository should also contain a README file that documents the content of the repository, as well as a summary of your conclusions.

2. The Bloom filter should be either implemented using an object-oriented approach, or a functional approach. Clearly, code quality is important. Your code should be easy to read and documented clearly. You have seen several examples in the course. The data structure is implemented in Python as a module, so that it can be used in a Jupyter notebook for demonstration purposes and testing, but also from a Python script for benchmarking on the HPC infrastructure.

3. The implementation must be tested thoroughly for correctness as explained several times in the course.

4. To implement a Bloom filter, you must define a family of hash functions. These hash functions should be tested to verify that they produce appropriate values. Note that these hash functions may work well for certain data (e.g., natural language words) but not so well for other data (e.g., random strings or DNA). Test with at least two data types.

5. Discuss the expected time and space complexity of your implementation.

6. The performance of the implementation has to be tested as well using a large data sample. Time the insert and search functions for an increasing number of words and create plots. These benchmarks should be performed on the HPC infrastructure. Include the job script and the Python test script in your repository, as well as the output of the benchmark runs.  
Consult the VSC documentation if necessary.

7. Check how the false positive rate changes as a function of the number of words inserted in the Bloom filter. Also check this if the number of words exceeds the expected number of words for which the Bloom filter was designed.

8. Check the compression rate of a Bloom filter as a function of the expected number of and the rate of false positives.