# Computing at Scale in Machine Learning: Distributed computing and algorithmic approaches (14038)

B-TU Cottbus-Senftenberg · WiSe 2023
Prof. Dr. Alexander Schliep, Nathalie Gocht, M.Sc. and Aleksandra Khatova, M.Sc.

**Problem set 5 from November 22, 2023 · Due on December 1, 2023**

**Problem 1** (Microsoft BitFunnel, 6 points)**.** Let's analyze how the number of set bits in Microsoft's BitFunnel Signatures increases when going from rank $i$ to rank $i+1$ rows. Assume that in a row of $b = 2r$ bits, there are $2n$ bits set: $n$ bits in the first half $b_1$ and $n$ bits in the second half $b_2$. These set bits are the result of inserting some number of items or prior `OR` operations to arrive at the rank i row. Assume that for each of the 2n bits, the position in the row was chosen randomly with uniform probability. Compute the expected number of set bits in $b_1$ `OR` $b_2$. Attach a sketch of the problem setting to your solutions.

Imagine you have a BitFunnel with a rank $i$ row size of $b = 128 \times 512$, a total $30000$ bits are set in the row. What number of bits do you expect to be set in rank $i+1$ for the given row?

**Problem 2** (Bloom filter, 8 Points)**.** You want build a quick look up search for the BTU library using Bloom filters. Sending the book entries to a central server proofs difficult since each branch of the library has a lot of books. The BTU library has 3 branches: Central Campus, Sachsendorf and Senftenberg.

a) Suggest a simplified solution to index the data entries at each branch and combine them to a central look up search. The answer should state how the book IDs are stored, what is queried and how data is processed. Use sketches where applicable.

b) Program this solution using a total of 1 billion unique string identifiers.

**Note:** Please hand in:

1. A single PDF file containing:
   - the answers to Problem 1.
   - the answers to Problem 2a.
2. A Python script named `problem2b.py`

Python code must be handed in as '.py'. Code submission in PDF will be graded with 0 points. Please submit only the files specified above.

You are allowed to use the example solutions provided where they might be helpful for answering question on this problem set.

Please feel free to discuss the problems and approaches to solving them with other students. However *you* should understand results of discussions well enough to write up the solution by yourself and possibly explain the solution. When you do collaborate, please make other's contributions clear. Present your solutions concisely.