Implementing efficient statistical tests using distributed parallelism in Dask

Supervisor: Fabian Woller

The goal of this project is to implement an efficient version of standard statistical tests in the presence of missing values. Existing tools make use of shared-memory parallelization based on a C++ or Numba backend – in this project the focus lies on evaluating the use of the Python framework Dask. The implemented tests should then be analyzed on efficiency, benchmarked on simulated and/or real-world data and compared to existing tools.

The steps of this project can roughly be summarized as follows:

- Become familiar with Dask
- Read into mathematical background of respective statistical tests
- Implement parallelized version of statistical test in Dask
- Analyze and benchmark runtime of implementation against competitors

Requirements:

- Programming skills in Python
- Basic understanding of statistical testing

This project can also be done by a group of two people.