Rule-based learning for generating semantic descriptions of patient subgroups

Supervisor: Fabian Woller

For our work with population cohort data we want to retrieve semantic descriptions of patient subgroups resulting from clusterings. On such large multi-omics datasets, our workflow consists of clustering patients based on one data dimension (e.g. gene expression) and then trying to find descriptions of the clusters on phenotype level (e.g. one cluster could consist only of men over 60 that already smoke for more than 20 years). Rule-based learning methods for extracting such descriptions already exist, with the goal of this project being to examine their quality and efficiency on large multi-omics datasets, also in combination with missing data.

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

- Retrieve and preprocess suitable cohort-level multi-omics dataset (e.g. TCGA PanCanAtlas)
- Implement clustering method(s) for generating patient subgroups
- Apply existing rule-based learning methods (e.g. RuleKit) for retrieving semantic descriptions of patient subgroups
- Analyze and report clustering time, rule extraction time and the overall usability of results

Requirements:

- Programming skills in Python
- Familiarity with UNIX systems and command line usage