High-dimensional data, where the number of predictors (features, variables) is much greater than the number of observations, are very common in different fields of science. Often the objective of data analysis is the choice of predictive model, which is finding a simple but strong dependence between a small group of predictors and the dependent variable. Such a model can be later used for identification of causal link or for prediction using low-dimensional data obtained from a different, much cheaper technology.

In the project we will construct, analyze, make publicly available and apply to genomics SOSnet -- a new model selection and parameters estimation method for the main classes of regression models. Moreover, we will use SOSnet for a parsimonious, that is based on a low number of variables, prediction of phenotype using data from next generation sequencing. Based on preliminary results, we believe that SOSnet will successfully compete with the best algorithms in the field.

Recently, the prof. Ploski's research team prepared next generation sequencing data based on blood samples of several hundred patients. Last year, using Lasso -- a popular method for model selection, we found new age markers, which are positions on the chromosomes, where cytosine methylation levels are correlated with age. These markers can be useful in criminology to assess the age of the offender. Prediction of age was accepted very well, although its prediction accuracy (sd ~ 7 years) can definitely be improved. Therefore, the first task in this part of the project will be applying SOSnet to improve prediction accuracy of the age based on genome-wide DNA methylation data.

The research methodology will be mathematical inference and computer simulations, mainly in an open-source environment R. We will analyse cost and error of the SOSnet method. All computations will be performed on computers in MIMUW.