**Freelive**

In the paper of Shin Li, a new variable selection approach is described that uses double cross validation and feature selection to make sure the correct variables are selected. For the regression models PLS regression was used and the Variable Importance in Projection (VIP) is used to make the selection.

Three approaches were used to find the minimum set of variables, a maximum set, and a set in between needed for good predictions.

One of the data sets used in the paper is the Freelive data set in which individuals take note of what they eat during a period. The diet forms are used to calculate a Rye Intake. Urine samples were taken and LCMS was used to measure metabolite levels.

Detailed information on study design and metabolomics data acquisition is provided elsewhere ([Hanhineva et al., 2015](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6419897/" \l "bty710-B18)). In brief, free-living participants with no diagnosed or perceived gastrointestinal diseases or symptoms were invited to participate and instructed to adhere to their habitual diet. Untargeted LC-qTOF-MS metabolomics was performed on urine samples. The dataset consisted of reported wholegrain rye consumption (continuous Y variable) from 112 observations (58 unique participants; two individuals had samples from one occasion available), codes for individual (numerical ID variable) and 1, 147 features as X matrix (a molecular entity with a unique m/z and retention time as measured by an LC-MS instrument)

The goal of this project is to perform variable selection approaches from these data. Which metabolites are necessary to provide good prediction. In the MUVR approach mentioned in the paper the VIP is used, but another good candidate is the Selectivity Ratio.

Make a strategy using PLS or PCR regression models to select a set of variables. Compare your results with the ones from the paper in terms of selected variables, and quality of the regression models (in RMSEP or Q2).