**Pre-processing**

The data is distributed over several files, as we want only one pandas dataframe with all the observations, the different files will be combined to one data frame. As the pre-processing is already executed, no missing data handling has to performed. No column is categorical and thus no dummy-coding has to be performed. The resulting dataset is cleaned and thus ready to analyze.

The initial dataset contains 130 features but these are not informative. Therefore the choice of algorithms that will be used did not take into account the interpretability.

To start a linear regression model will be created on all the features, this will be used as starting point such that model amandements can be compared.

**Dimension reduction**

* PCA
* Random forest with feature importance
* High correlation filter

**Algorithms that will be used**

1. Elastic net regression model
2. Bagging (decision tree)
3. Boosting (decision tree)

**Model for each gas**

**Model containing all the data without gas feature**

**Model with all the data and gas feature**