1 Machine Learning Basics

1.1 Training, Validation, Testing

For building a model, three data sets are needed:

1. **Training set:** A data set for optimizing parameters (e.g. orientation of separating hyperplanes)
2. **Validation set:** A data set for optimizing hyper parameters (e.g. Number of layers, trees in a forest)
3. **Test set:** A data set for evaluating the performance of your model

These sets should (exceptions confirm the rule) be independent of each other (i.e. there must be no information leaking from one set to the other).

1.2 Cost Functions

To analyze how close to the true values the values from our algorithm are, we use cost functions.

Residual Sum of Squares: $\sum_{i=1}^{n}(y_i - f(x_i))^2$ (n = total number of samples, y_i = label of the i^{th} sample, $f(x_i)$ = prediction of the i^{th} sample)