# Clusetering, data mining and machine learning

Unsupervised learning

* We look for patterns in data
* We create clusters
* We determine how many clusters the algorithm should be looking for. We can try different numbers of clusters to see, what works

Supervised Learning

* We know the labels beforehand and train the models on these, so the algorithms is capable of determining key features.
* Then we test the algorithm on thest data to see, how effective it is
* W3 can utilize more data for the model to train on and improve in precision
  + Regression and classification

## Dimensionality reduction

* High dimensionality means to a have a lot of input features. This makes the algorithm more complicated and difficult to compute, while increasing the likelihood of the model to perform overfitting. For this reason, we try to limit the amount of input variables, so we ´focus on those who captures the essence of the data.
* Techniques
  + Feature selection
    - Simply choose relevant features
    - Often used in scoring or statistical methods
    - The selection is done by the model based on
      * Wrapper methods: Here we fit and evaluate a model with different subsets of data
      * Filter: Here we look at for instance corraletions values and base core features on this
  + Matrix factorization
    - Look this up – something about the eigenvectors
  + Principal component analysis
    - This is the most common one
    - We standardize data
    - We compute a covariance matrix to identify correlation between variables
    - Compute eigenvector and eigenvalues. We need these to compute principal components
    - Principal components are different linear combinations of the initial variables. They are new variables. Principal components are uncorrelated. You try to save as much information in the first variable as possible and then the residual information in the next and so on.
    - A three dimensional dataset has three eigenvectors and three eigenvalues
    - We rank the principal components by, which holds the bigget amount of information. From thereon we can choose, what to keep and what features to let before
  + Finally we can make a feature matrix to use for the model. This matrix underlines the finalization of out dimnesionaly reduction.