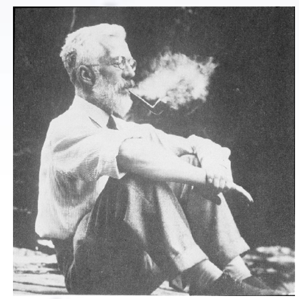
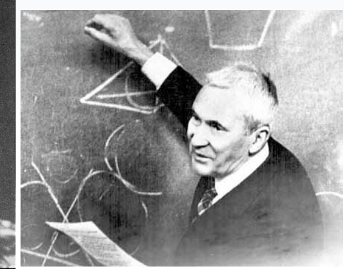
Applied ML

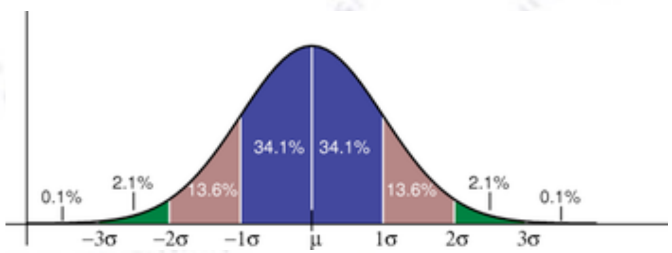
(Kernel) Principle Component Analysis





Troels C. Petersen (NBI)

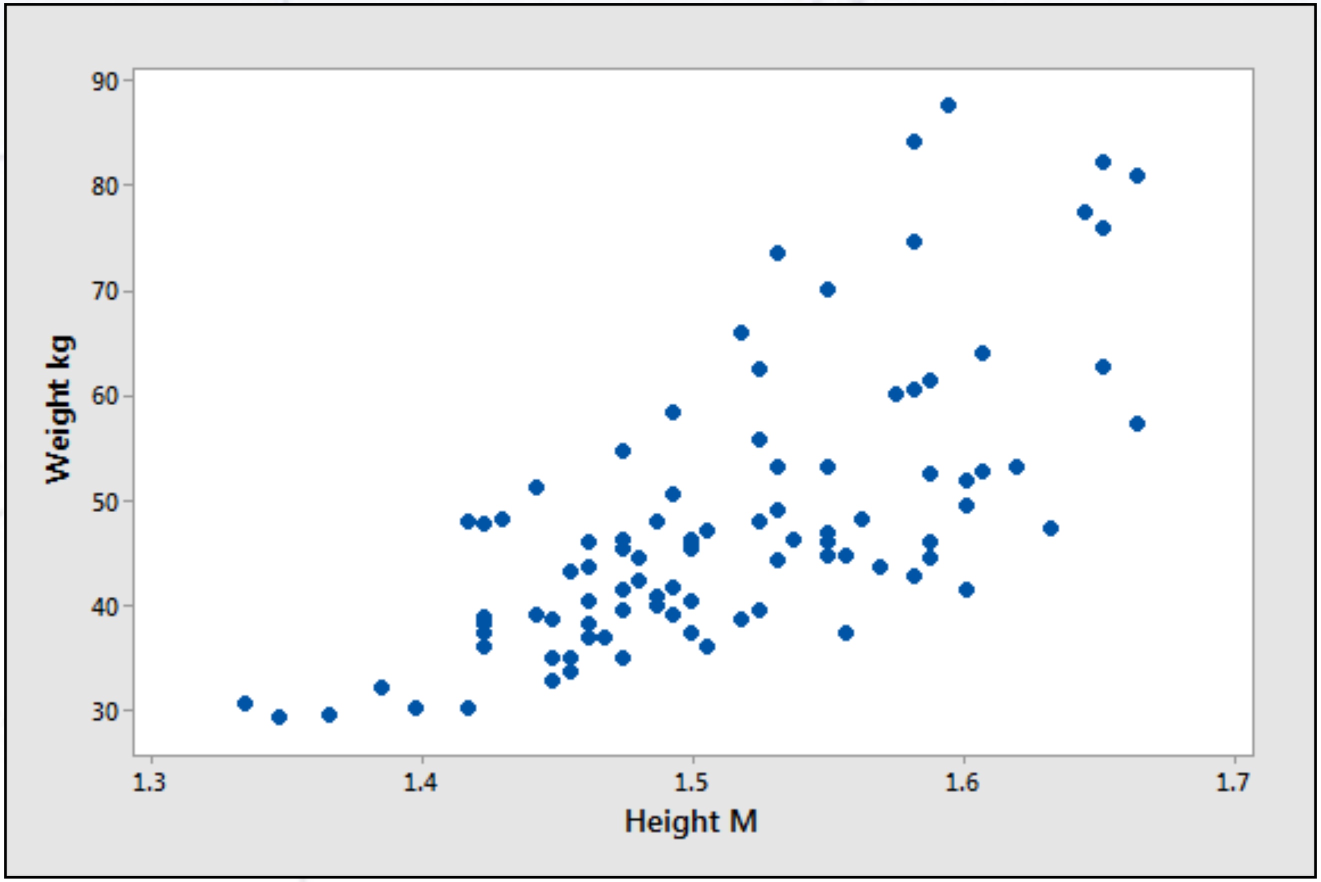


"Statistics is merely a quantisation of common sense - Machine Learning is a sharpening of it!"

PCA overview I

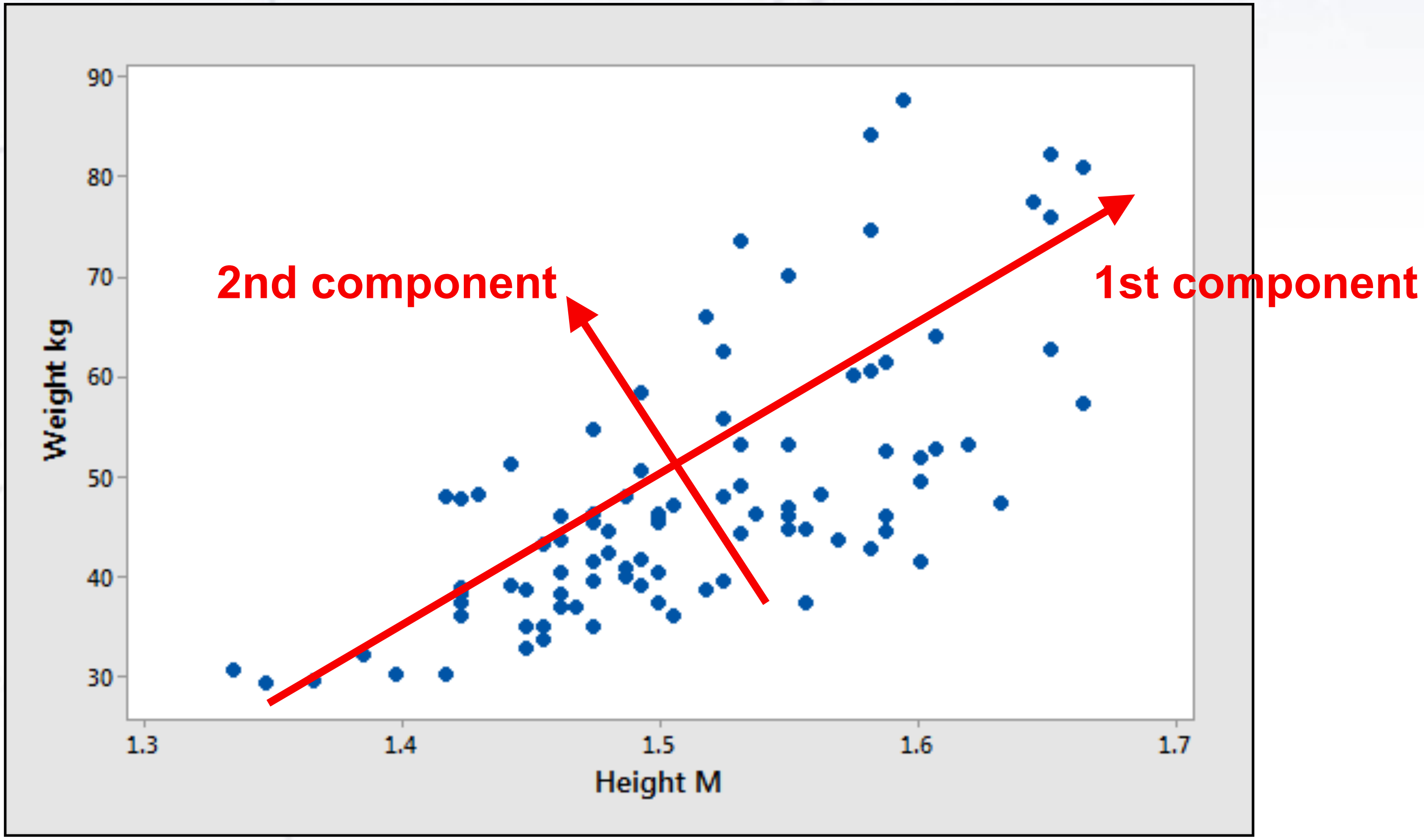
Consider data which have correlations, here in 2D (for visualisation), but potentially in (very) high dimension.

We want to apply a PCA to this data, to reduce dimensionality!



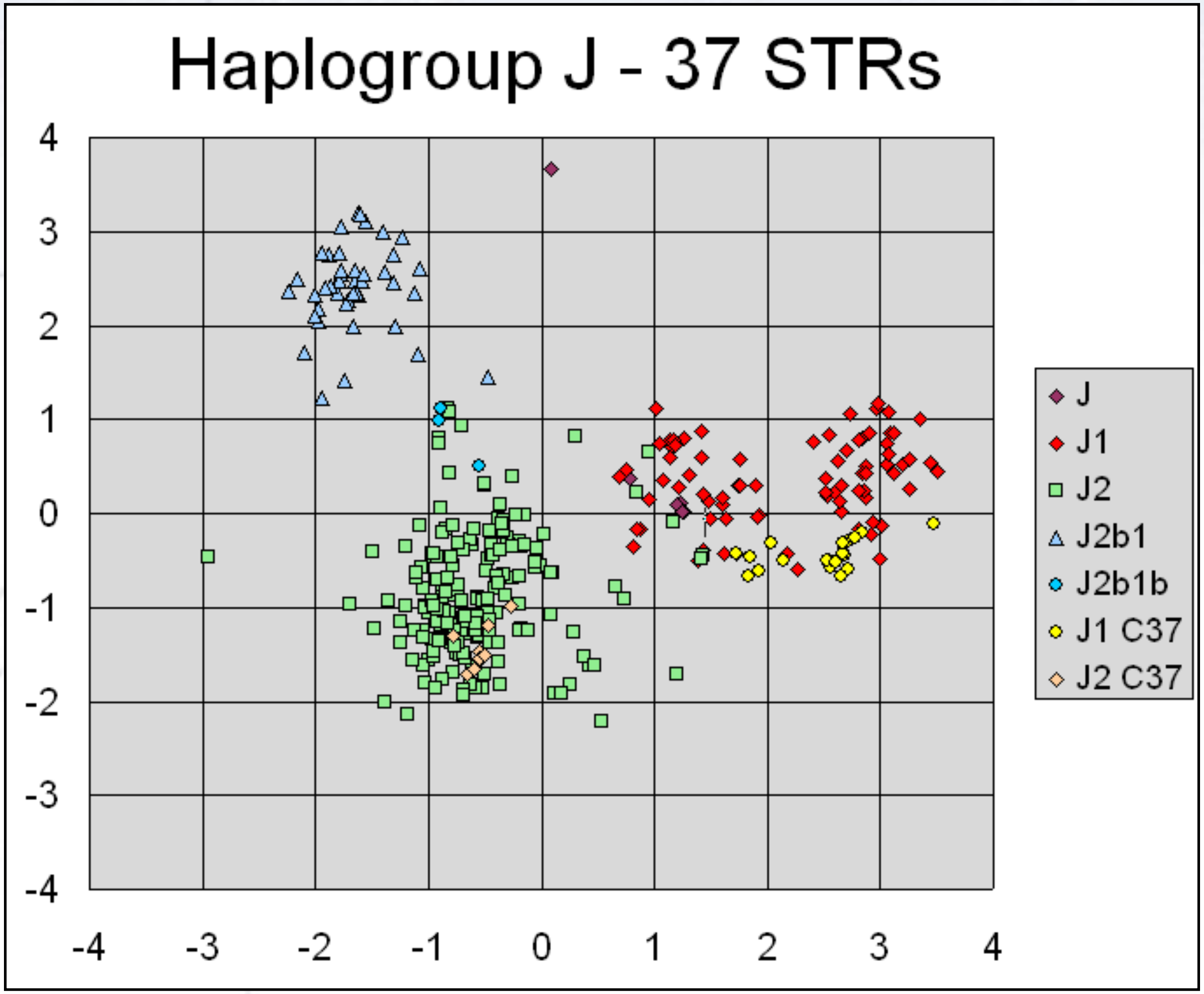
PCA overview II

Find the direction, which has the maximum variance, i.e. “best along the direction of the data”.
The effective way to do this, is to find the eigenvectors and eigenvalues, and rank the eigenvectors (i.e. directions) according to eigenvalues.



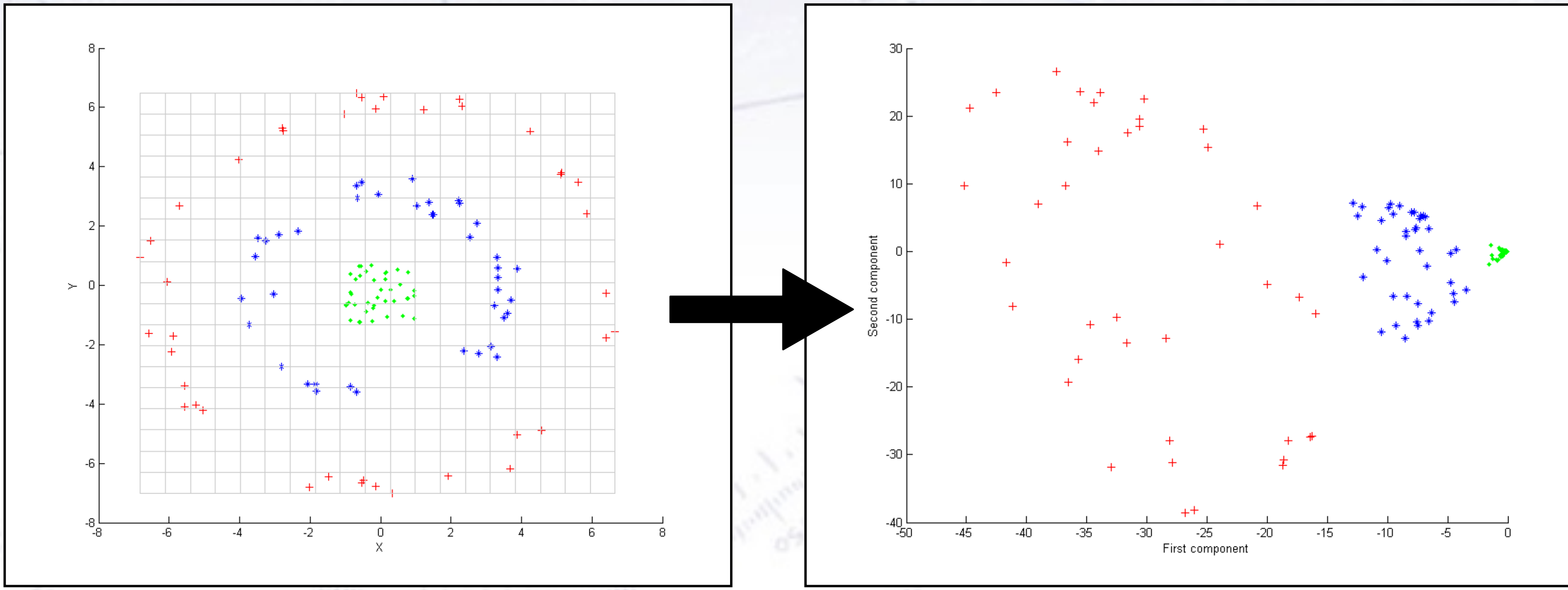
PCA example

Using PCA on genetic data is quite common, as there is no “labels” and the dimensionality is very high (millions), here boiled down to two dimensions:



Kernel PCA

For non-linear problems, the kernel PCA might be the solution. Here, a (non-linear) kernel is applied before the PCA transformation. This is computationally heavy, but often works well as shown below:



There are other non-linear unsupervised methods, in particular t-SNE and UMAP have gained popularity from their performance.