**Assignment Unsupervised machine learning (due Monday June 13)**

Choose one of the assignments (beginners, intermediate, advanced). Note that each of the assignments can lead to a sufficient mark, you should just choose the one you like the most. If you do not like one of the assignments, you can use your own dataset as well. Just make sure that you explain your choices!

**Beginners level**

Dataset: primates-morphology

The file primates-morphology.csv contains coordinates derived from markers of skull features from various species of primates.

Read the data in and use unsupervised machine learning methods to separate the groups belonging to different species. Do not forget to demonstrate a good working method and a non-working approach and explain why the method works or fails for this type of data. For each method that you try, give a short argument of why the method is worth trying. Make sure you properly inspect and preprocess the data (Check for missing data, skewness, distribution, sample, linearity, correlations etc.).

**Intermediate level**

<http://doi.org/10.7937/K9/TCIA.2015.L4FRET6Z>

clinical data: https://wiki.cancerimagingarchive.net/download/attachments/16056856/Lung3.metadata.xls?version=1&modificationDate=1404237338168&api=v2

gene data

# Gene Expression data available at: https://ftp.ncbi.nlm.nih.gov/geo/series/GSE58nnn/GSE58661/matrix/

Read the two datasets in and conduct proper preprocessing. Combine the clinical data with the genetic data. Use unsupervised learning to develop an accurate machine learning model that can predict lung cancer. Explain your choices. Make sure you properly inspect and preprocess the data. (Check for missing data, skewness, distribution, sample, linearity, correlations etc.). Use an argumentative approach.

**Advanced level**

Dataset: /homes/tsjerk/Datasets/alpha-helix-ends-4-5.dat.gz

This dataset contains helix structure data. The main research question is:

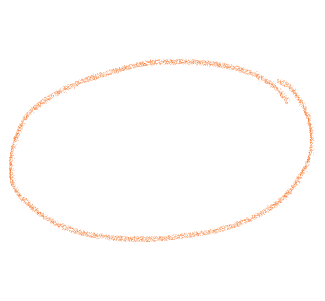
Can we find a pattern in the end of the helix?

If so, what is the pattern

If so, can we couple the pattern with the sequence

Each helix is built of 27 coordinates, 9 times 3 (x, y, z) coordinates. The end of the helix starts after the [0,0,0] point.





**[-x,y,0] [-x,0,0] [0,0,0]. [x,y,z]. [x,y,z] [x,y,z] [x,y,z] [x,y,z]**