# LAB 4: differential gene expression analysis and machine learning

# ASSIGNMENTS

First, take a look at the **LAB4_Tips** file and practice with each section described. We strongly believe that they could be helpful with the following assignments.

## Assignment 1: T-test to perform differential gene expression analysis

Download ***Dataset.csv*** from the Teaching Portal and perform a differential gene expression analysis in order to find which genes are able to distinguish between breast cancer Luminal A subtype and breast cancer Luminal B subtype.

Calculate t-value and p-value for each gene and select only genes for which p-value < adjusted Bonferroni p-value.

Finally, convert differentially expressed genes from ENSEMBL notation to the common name (e.g. ENSG00000268889.1 is AC008750.7).

Which genes are differentially expressed between the two populations?

Create now a reduced dataset (from now on referred to as ***reduced_dataset.csv***) made up of all samples with only differentially expressed genes (use common name notation).

## Assignment 2: Use gene expression data to create a classifier for Luminal A / Luminal B breast cancer subtypes

In order to create a Luminal A / Luminal B breast cancer classifier, consider two dataset: ***dataset.csv*** (the one we provided you) and ***reduced_dataset.csv*** (the one you created in Assignment 1). Now, follow the instructions to continue and take a look onto scikit-learn section in LAB4_Tips for some built in classes:

1. Divide both dataset.csv and reduced_dateset.csv into train set and test set
2. Standardize features of both datasets by removing the mean and scaling to unit variance
3. Perform the dimensionality reduction onto dataset.csv with PCA (principal component Analysis) using 80 features.
4. Train a KNN classifier onto the train set of **dataset.csv** with PCA
5. Test the classifier obtained at the previous step onto the test set of **dataset.csv** (remember to apply PCA transformation onto test set)
6. Implement from scratch the following performance metrics: accuracy, precision and recall, F1 score. Compare your results with performance metrics provided by sklearn.metrics
7. Train and test a KNN classifier onto **reduced_dataset.csv**
8. Which are the differences between the performances obtained onto dataset.csv with PCA and these obtained onto reduced_dataset.csv?
9. Use reduced_dataset.csv to train and test (with default parameters) SVM, Random Forest and Naïve Bayes classifiers
10. Which classifier provides you the best result onto reduced_dataset.csv?