# Gene expression data analysis

Credits: <https://cw.fel.cvut.cz/wiki/courses/bin/tutorials/tutorial10>

This assignment has been slightly adapted to reflect the needs of our course.

* We have a small number of observations (~10^1) and big number of features (~10^3).
* The expected problems with that include false hypotheses and overfitting.
* Interpretability: are the expressed genes the causal ones?
* Decrease the number of hypotheses.
* Analyze in terms of more abstract entities than genes, e.g. principal components.

## Data

7,129 GE profiles of 72 patients

25 samples: acute myeloid leucaemia (AML)

47 samples: acute lymphoblastic leucaemia (ALL)

## The task

Construct a decision model to differentiate these types of cancer.

Golub et al.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.

## Part 1 (max. 2 pts)

1. Train a decision tree on subjected data. Use Python scikit-learn (this is up to you) classification tree fit it in to the data.
2. Show the tree and enumerate its training accuracy.
3. How would you interpret this model? Which gene is crucial for the decision?
4. Is this gene really the one causing the cancer? Look up in the article Golub et al., 1999.
5. Estimate real accuracy of the tree. Use e.g., crossvalidation (alternatively, you can split the data).
6. Compare it with the training accuracy.

## Part 2 (max. 3 pts)

1. Learn a basis-matrix V of the data. Use PCA method.
2. For a range of component numbers K:
   1. project the original data X to the top K components of V. The result are data Z with reduced dimensionality.
   2. Create a tree out of these reduced data. Show it and enumerate its training accuracy.
3. Compare all the trees resulting from the reduced data and pick the “best” according to its accuracy and structure. Follow the Occam razor.
4. Estimate the real accuracy of the “best” chosen tree. Again, by e.g. crossvalidation.

## Part 3 (2 extra pts)

* Extract the genes active in the discriminative components. The discriminative components are those vectors of basis-matrix V, which refer to the features your tree consists of. To extract the active genes from a component use e.g. scikit-learn K-means clustering.
* Resulting gene-sets related to each of the discriminative component shall hopefully refer to some abstract biological processes. Use Gorilla to enrich these gene sets in Gene-ontology terms. http://cbl-gorilla.cs.technion.ac.il/