In this homework you will analyze gene expression and drug sensitivity data from cancer cell lines. The files are uploaded in the **data** directory. The files' contents are the following:

- **sample_info.csv:** general metadata about the used cancer cell lines. The first column (*DepMap_ID*) is an unique ID of the cell line. Following columns give information about the cell lines like tissue type etc..
- **CCLE_expression.csv**: gene expression data of the used cancer cell line. The first columns is the *DepMap_ID* of the cell line, while the first row contains the gene ids (in *X (Y)* format, where *X* is gene symbol, *Y* is Gene ID). Gene expression is displayed in log2(TPM+1) units.
- **GDSC2_fitted_dose_response_25Feb20.xlsx**: drug sensitivity data of the cancer cell lines. Each row contains the drug sensitivity metric (*LN_IC50* column, which represents natural logarithm of the half maximal inhibitory concentration) for a given cell line (*SANGER_MODEL_ID* column) and a given drug (*DRUG_NAME* column). Hint: you can match the cell lines with the help of sample_info.csv, which contains not only *DepMap_ID*, but also *Sanger_Model_ID*.

These datasets are coming from:

- https://depmap.org/ (1st and 2nd)
- https://www.cancerrxgene.org/ (3rd) , where you can find also some additional literature regarding these datasets.

The goals are the following:

- Use some dimension reduction / clustering / visualization methods to show the general clustering of cell lines (based on their gene expression values)!
- Try to identify gene expression-based biomarkers for the drug **Lapatinib** using some statistical methodologies, and give some brief biological interpretation of these biomarkers.
- Try to use some predictive model (statistical or machine learning) to predict a cell line's Lapatinib sensitivity based on its gene expression values. Try to give some estimate about the prediction performance of this model.
- What other data type could be used to improve the prediction performance of the model? Could you suggest some database to find this type of data?

You must submit your summary, code and generated figures. You should use Python. You can choose between notebooks or scripts.