## Data:

File *cell-count.csv* contains cell count information for various immune cell populations of each patient sample. There are five populations: *b_cell*, *cd8_t_cell*, *cd4_t_cell*, *nk_cell*, and *monocyte*. Each row in the file corresponds to a biological sample.

## Python:

1. **Please write a python program to convert cell count in *cell-count.csv* to relative frequency (in percentage) of total cell count for each sample.** Total cell count of each sample is the sum of cells in the five populations of that sample. Please return an output file in csv format with cell count and relative frequency of each population of each sample per line. The output file should have the following columns:

   ***sample***: the sample id as in column *sample* in *cell-count.csv*
   ***total_count***: total cell count of *sample*
   ***population***: name of the immune cell population (e.g. *b_cell, cd8_t_cell, etc.*)
   ***count***: cell count
   ***percentage***: relative frequency in percentage

2. Among patients who have treatment *tr1*, we are interested in comparing the differences in cell population relative frequencies of melanoma patients who respond (responders) to *tr1* versus those who do not (non-responders), with the overarching aim of predicting response to treatment *tr1*. Response information can be found in column *response*, with value *y* for responding and value *n* for non-responding. Please only include PBMC (blood) samples.
   a. **For each immune cell population, please generate a boxplot of the population relative frequencies comparing responders versus non-responders.**
   b. **Which cell populations are significantly different in relative frequencies between responders and non-responders? Please include statistics to support your conclusion.**

Please return both the code and the outputs. Return all files in a .zip file of the form *FirstName_LastName_Teiknical.zip*. Answers that do not have a .zip file in this form will be considered incomplete. Please also specify any dependencies that you use and instructions on how to run your code to reproduce the outputs.

## Database:

1. **How would you design a database to capture the type of information and data in *cell-count.csv*?** Imagine that you'd have hundreds of projects, thousands of samples and various types of analytics you'd want to perform, including the example analysis of responders versus non-responders comparisons above. Please provide a rough prototype schema.

2. **What would be some advantages in capturing this information in a database?**

3. Based on the schema you provide in (1), **please write a query to summarize the number of subjects available for each condition.**

4. **Please write a query that returns all melanoma PBMC samples at baseline (*time_from_treatment_start* is 0) from patients who have treatment *tr1*.**

5. **Please write queries to provide these following further breakdowns for the samples in (4):**

    a. How many samples from each project
    b. How many responders/non-responders
    c. How many males, females