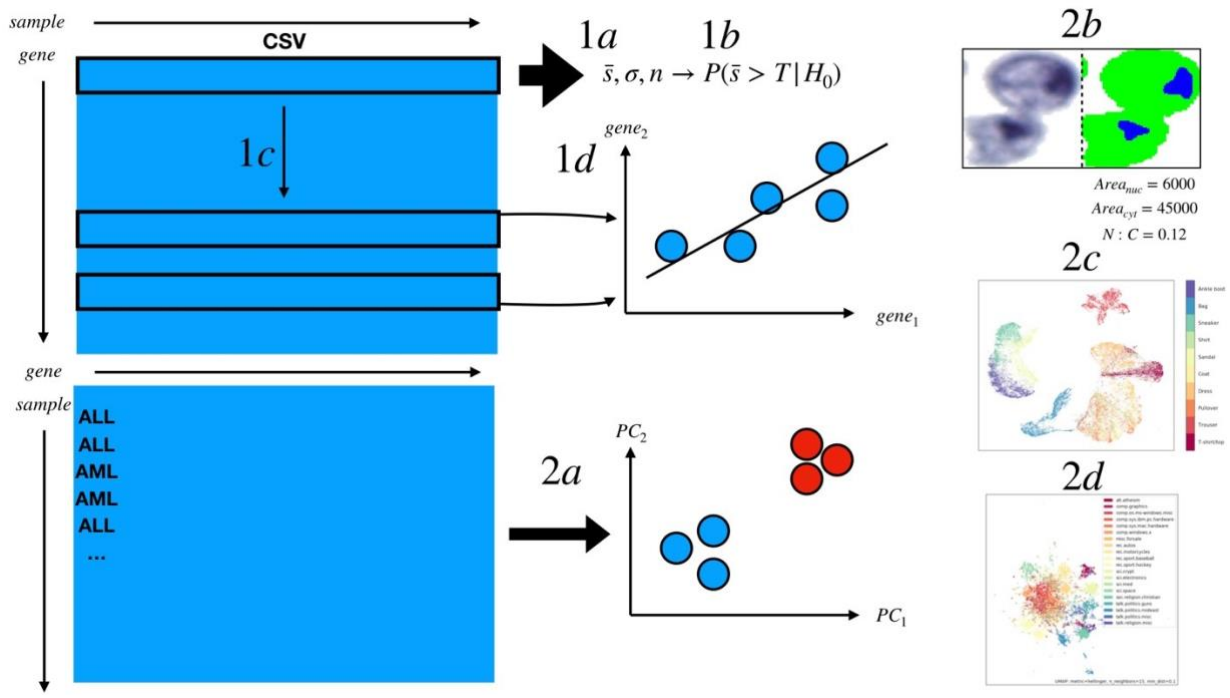
Technical Interview Questions:



Please try to answer some of the following questions to the best of your ability while explaining all of your logic. It is okay if you do not know the answer to these questions, feel free to email Joshua for hints if you are stuck. If you are unable to answer the question, please explain how you would go about solving the question. Popular tools for displaying answers to these questions include jupyter notebook and Rstudio, though any printout is fine as long as you can walk us through it.

We have included test data that matches each question number and letter in the zipped file attached to this email:

1. Programming/Statistics:
    a. Functional programming:
        Write a python/R function that returns the mean, standard deviation and number of elements of a python/R list/vector. As an example, you can utilize this python list: [5.99342831, 4.7234714 , 6.29537708, 8.04605971, 4.53169325, 4.53172609, 8.15842563, 6.53486946, 4.06105123, 6.08512009]. b. Statistics:
        Given a float array (use the one from the previous problem) and a supplied “threshold” value, using the function from the functional programming problem, devise a one sample z-test that tests to see whether the mean value of the sample exceeds the supplied threshold value (you can use scipy or the package pingouin for help). Assume that the sample level variance matches the population-level variance (feel free to conduct a t-test if you want to violate this assumption). Suppose the threshold here is 4 if using the previous dataset. The analogy here in bioinformatics is that imagine that as a subcomponent of some experiment, we want to see whether the expression of certain genes surpasses a

threshold that carries some previously established diagnostic or prognostic significance. Let's call a gene with mean expression greater than the threshold a "bad" gene. Be prepared to speak to what the p-value means in the context of your test.

c. Object Oriented Programming:
   We're going to generate a python class that is able to perform these tests across many such "genes". Generate a python/R data class that loads a csv file in its init method. The csv file will be comprised of a header line containing patient names, then each subsequent line will first start with a gene name, followed by "expression" values for each of the patients (this is a gene-by-sample matrix). We want to see if each of the genes is a "good/bad" gene using the previously established threshold (this is an unrealistic situation, but designed this way for simplicity). Define two methods for this class, one that re-implements your hypothesis test function from the previous section, and the other that calls this method on each of the lines of the csv, outputting a list of p-values and some indication of whether the gene was "good" or "bad". We'll use the same expression threshold of 4.
d. Plotting:
   We think that some of the expression of these genes may be correlated, but want to plot a scatter plot of the two genes expression values across the cohort to see if this is the case (each axis is each gene's expression, each point is an individual). Please write a function or class method that takes as input two lists of floats, or two of the gene lists from the previous problem, and outputs a scatterplot of the two gene's expression across the cohort. Libraries like matplotlib, seaborn in python and ggplot in R are good for this task.