# BMCS 4480
# Spring 2024
# Assignment 2

Submission Due: **Feb 13th, 1pm**

*Submit your work in the form of a Python Jupyter notebook (preferred), or MATLAB/R code and writeup containing figures to Courseworks. The last question can be submitted in any format. In your notebook or writeup, please include the names of all collaborators you may have discussed these problems with.*

Download single-cell gene expression data for ~6K PBMCs (peripheral blood cells) from a healthy donor from here. Note the data is already filtered and contains only the 500 most variable genes. Rows and columns represent genes and cell IDs respectively.

**Q1a.** Normalize the data by scaling to median library size and log transform the normalized data and perform PCA, followed by t-SNE or UMAP on the top 20 PCs. Plot the 2D or 3D embedding [15pt].
**Q1b.** Is 20 a good choice for the number of PCs? [bonus 5 pt]

**Q2.** Cluster cells using the K-means method and color the embedding from Q1 with cluster IDs. Justify your choice of K [20 pts].

**Q3.** Compute a 30-NN (nearest neighbor) graph between cells. Plot a heatmap of the adjacency matrix for the graph [10 pts]. Justify your distance metric [bonus 5 pts].

**Q4.** Cluster cells using a graph-based algorithm such as Louvain with the kNN graph from Q3. Color the embedding with cluster IDs. How does it compare to K-means? [20 pts]

**Q5a.** Perform a t-test to find differentially expressed genes (DEGs) in a cluster of your choice. Color the embedding with the expression of the top 5 DEGs [20 pts].
**Q5b.** How would you characterize the cluster [5 pts]?
**Q5c.** Which other differential expression method might be appropriate for the expression distribution in this data? [bonus 5pt]

**Q6.** Solve problem 9.4 of the PRML textbook (pg 456) [10 pts].