Team Members: Andrew Wei, nowei@cs.washington.edu
(I couldn't find a partner :c)

Note - I ended up choosing a project from the example topics given. From the website:

Project description: Understanding how gene expression patterns are different in different subtypes of cancer

> **Goal**: Our goal is to understand how differently genes regulate each others' expression levels in each subtype. One way is to learn the regulatory network in each subtype and interpret how they similar/ different. A different approach is to build a classifier that can predict the subtype of leukemia based on the expression data from a patient, which will enable molecular diagnosis of leukemia.

What scientific question will I address?

I will try to understand gene expression patterns in different types of leukemia based on expression data by looking for patterns in the data and applying various machine learning techniques to try and classify the data. If there is time, I will explore trying to make networks on this data.

What computational methods do I plan to use?

I wanted to do some graph learning approach, but this isn't graph data, and it seems like it would be computationally intensive to compute gene networks of size 20,000 for 2096 patients. As a result, I plan to take a more ad-hoc approach to this problem by analyzing statistics between gene expression data as features and using simple ML techniques like multi-class logistic regression to identify potential trends in the data before moving onto DL models to try to maintain the interpretability of the results.

How I will implement things?

1. Split the dataset up into 80% training, 10% validation data, and 10% test data by sub-type of leukemia
2. Analyze the RNA levels of `Non-leukemia and healthy bone marrow` patients. Calculate mean and standard deviation for each gene. Then standardize the RNA levels for other patients based on the mean and standard deviation of the non-leukemia patients.
3. Feed these as features to a multi-class logistic regression model and see what results I get
4. Analyze the trends between how specific types of leukemia differ from the non-leukemia and other leukemia patients.
5. Stretch goal: employ other features/statistics/ML techniques to get better results

What data sets will I analyze?

Data: We have an expression dataset measuring RNA levels of 20,000 genes from 2096 patients suffering from leukemia. There are 18 sub-types of leukemia and it is important to understand the expression signature that characterizes each subtype of leukemia. [Gene Expression Data] [Classification Labels] [Batch corrected Gene Expression Data]

The last link is broken though, so I'm not sure what was there.