**Directions**

Please answer the following questions as thoroughly as possible by showing your work via code and/or equations (markdown/latex) and plots as often as possible in a jupyter lab or google colab. If the latter, please download and send me a copy of your final answers as well as share the original colab with me. If you do not use google colab, it is worth checking out :)

Most Important: You may use any resource at your disposal to answer the below questions, including coding libraries, instructional resources, etc. However, **AI Therapeutics trusts that all work will be 100% your own. No one else may be consulted or contribute to answering the below questions. Discovery that any amount of work is not your own will result in termination of your candidacy.**

Some of the questions below are technically challenging. They were selected to mimic typical challenges you will face as a data scientist at AI Therapeutics. Remember: science is difficult and requires creativity to solve problems. The questions were designed to be challenging so please relax, think carefully and have fun!

The first 3 questions refer to the data in distributions.pickle

1. Characterize/describe the 3 distributions in file distributions.pickle
2. Measure the density of each distribution empirically or analytically. Significant bonus for analytical solutions.
3. Which family[ies] of models would be expected to perform well predicting data shaped like each distribution? Why? What are some of the expected error distribution shapes for each dataset and model type (please give a minimum of 1 example for each distribution)?
4. Write a script to scrape drugbank.com. Use the script to retrieve drug name, drug bank accession numbers, chemical SMILES and any indications listed by drug bank for 100 drugs. Please share the script and data on your 100 drugs as a pandas dataframe. Hint: lists of drugs available in the drugbank database can be found at https://go.drugbank.com/drugs
5. You are training a model with data from the Gene Expression Omnibus to predict what disease state samples from a gene expression profiling experiment came from. Given dataset labels (disease identifier and experiment ID) found in disease\_ID\_GSE.pickle, devise a training strategy that will yield a model expected to generalize well on new data. Note, the actual data is not provided, I am strictly asking about training strategy. Please respond with code and verbal description/explanation. The disease identifier is the first array in disease\_ID\_GSE.pickle and the experiment ID/GSE is the second array in disease\_ID\_GSE.pickle. The *i*th element of disease identifier corresponds to the *i*th element of GSE. So, index *i* corresponds to an example of disease[*i*] that was produced from experiment GSE[*i*]
6. AI Therapeutics performed gene expression profiling experiments on a number of candidate drugs. A subset of this data has been processed into fold change values from untreated per gene and can be found in the dictionary saved in gene\_expression.pickle. Each key of the dictionary is a drug ID. The items of each key contain a matrix with gene expression values. Each row is 1 sample and each column contains the expression level of a given gene. For example, drug ‘AIT\_0’ contains 50 samples, each sample contains measurements from 978 genes. The column ordering is identical across all items and their identities (i.e., gene label) is the second item in gene\_expression.pickle. What can you learn about the dataset? This question is open ended and you may respond however you wish. A few example inquiries are: which genes are affected by most by each drug, how noisy are the samples, do any drugs exhibit similar behavior, etc. Feel free to analyze and report on the data in any way you wish, but please include visualizations in your response(s) wherever possible.
7. What recent development in machine learning excites you most?