ML Lab Project Description:

Our project aims to build a machine learning model that predicts if a dataset contains fabricated data or real data. To accomplish this, we will feed the model various fake and real data sets to train on patterns. Some predetermined patterns may be unrealistic distributions (too close to the Normal Curve), suspicious digit patterns (violating Benford's Law), lack of variation, and other patterns that seem forced. We will also look at the difficulty of replicating results as a measure of potential data inaccuracies. Specifically, we will collect data from sites such as Data Colada (Fake Data section) and kaggle as well as more reputable sites such as Data.Gov, CDC, or Census.Gov data. These will be our training and testing data sets for the model. We plan to do a clustering model that sorts the data sets into fake or real clusters based on the predefined characteristics. We may explore other models as well such as KNN and Trees.

We will visualize the distributions, potentially comparing histograms, scatterplots, etc– to assess whether there are any initial noticeable differences. We will apply Benford's Law to both the fabricated and real data sets to see if the distribution of the first digits follows the expected pattern. We also plan to calculate standard deviations to determine if fake data sets tend to have more or less variation. Finally, we will investigate how fake vs. real data sets handle missing values– the proportions of present vs. missing data, how it is imputed or noted, etc. These tests and experiments will set up the basis of our predictive model.

Our datasets will likely focus on numeric, rather than categorical data, as many of our tests require numeric patterns. We will also try to find "related" fake vs. data sets (similar sector, problem, research question, etc.) to improve the accuracy of our predictive model. We have yet to determine how many datasets we will include, as this will depend on the difficulty of finding

relevant data sets and conducting the analysis. However, we plan to include as many as we possibly can.