**Chemistry of Mars Project Summary**

**Participants:** Michele Myong, Sara Mezuri, Nikolas Eptaminitakis, Katherine Martin

**Description of Data Sets:** For each set of data there is metadata.csv, submission_format.csv, train_labels.csv, validation_labels.csv and folders that contain the training features and validation features.

**Problem description:** Help NASA scientists determine the presence of certain minerals in Mars rock and soil samples. The goal is to automate the analysis of mass spectrometry and/or gas chromatography data so that scientists can more quickly make informed decisions about data collection while on time-limited missions.

**Data Cleaning:** We dropped the non-integer values of mass of charge ratio, selected a range of the relevant values, namely (0-100), and finally ignored the value 4, which corresponds to the carrier gas helium. After that, we normalized the abundance curves to be into the range 0 to 1.

**Feature Engineering:** Our features engineering mainly involved two strategies. Our first strategy consisted in choosing the five most prominent ones and recording these, the corresponding peak abundance, the temperature at which it occurs, and the sum of the mean and the standard deviation of the abundance curve. Our second and more successful strategy involved considering 16 temperature intervals and for each of them computing the area under the abundance curve for the temperatures within this interval, for each mass to charge ratio.

**Basic Classification Models:** Because this is a multi-label classification problem, we were able to implement several basic classification models. Each model was computed using the temperature bin data. For this data, the most accurate basic classification model was K Nearest Neighbors with 1 neighbor. The accuracy of each model was first taken over each sample and then also calculated by taking the average accuracy over each label, each mineral. However this is a bit misleading because for each sample, each mineral is unlikely to be present.

**Best Model:** Our data is imbalanced, as each mineral is unlikely to be present in a given sample. We selected Binary Relevance with a Random Forest Classifier as our best model. Binary Relevance breaks a multi-label classification problem with L labels into L separate binary classification problems, each handled by a base classifier (we tested four different classifiers). The final prediction combines the results of all individual label classifiers. This approach achieved the highest accuracy and F1 score using micro-averaging.

We also tested neural network and ML-KNN models, but despite their high accuracy, we rejected them. They relied on macro averages instead of sample averages and performed poorly on our imbalanced data, resulting in a low F1 score.

**Future Directions:** We hope to talk with NASA scientists about getting recently collected data from Mars rock in a nice format that we can try our highest performing model on. We could also try to engineer new features using wavelet decompositions of the abundance curves for the various mass to charge ratios and use them to train models.