Justin Valentine
Ephraim Romesberg
Ryan Foley

1. GOALS OF THE PROJECT
   About 50 lava flows from 12-16 million years ago formed much of the landmass surrounding the Columbia River Gorge. Not all of the lava flows were continuous which has made it difficult to correlate lava flows from place to place. In order to classify rock samples geologists need to rely on geochemical analysis as well as geological domain knowledge. This reliance on geochemical analysis leads to several questions:
   - Knowing the MapUnit for which a sample is classified:
     - How much of the classification could be accomplished solely from the geochemical analysis?
     - Which of the available predictors distinguish MapUnit classes?
     - What would be the best approach to classify a sample to its MapUnit on the basis of its geochemistry, and how can we assess the confidence of that assessment?
   - Disregarding the MapUnit classification:
     - What are the natural groupings of the samples, and what are the primary factors distinguishing the groups? How do these groupings differ from the MapUnit classification?

2. SUMMARY OF AVAILABLE DATA

The data is a csv file with a little over 1000 rows. Each row is a rock sample from the Columbia River Gorge. The columns are geochemical data from the rock samples. Most of the columns contain information on the amount of a specific element or compound contained in the rock sample, some are parts per million and others are percent mass. Some of the columns containing info on the rare earth elements have missing values. The data is

grouped by the MapUnit column which is a categorical variable. In many of the MapUnit groups the rock samples come from the same lava flow, some groups may contain samples from more than one lava flow though and some of the data may be inaccurately classified. The date the geochemical analysis was completed is almost always included as a column, and some samples are analyzed twice as a way to calculate the random error in the analysis process. These repeat columns will be removed for the analysis. We will collaborate with our research partners to come up with a method to handle the thresholded and missing data.

3. THE METHODOLOGY

Different methodologies will be used for the different questions that need to be answered. We may employ discriminant analysis and regression models such as multinomial logistic regression to find significant predictors of MapUnit category. Discriminant analysis will give us an intuitive formula to use for classification and allow us to determine which variables play the biggest role in MapUnit classification. Multinomial logistic regression is another technique that deals with data with a categorical response so this may be useful as well. We will also consider using a random forest algorithm for classification. We may split the data into a test/train set or use cross validation to determine how accurately these methods classify the data by MapUnit. Since some of the variables are percentages, we may have issues with multicollinearity. We may need to reduce the dimensionality of the data to deal with this, methods such as principal component analysis may be useful for dealing with this.

For the clustering portion we plan on using algorithms similar to k-means clustering and hierarchical clustering to determine the most natural grouping of the data (ignoring the map unit groups). We will likely use hierarchical clustering to obtain a useful visualization of the clusters and determine the appropriate amount of clusters to use. We will then use k-means clustering after we have determined the appropriate k. After we

Justin Valentine
Ephraim Romesberg
Ryan Foley
have found the optimal grouping of the data using these methods we will compare our clusters to the MapUnit groups given in the dataset. Other methods such as fuzzy c-means might be used.

Another method we would like to try in order to classify samples solely from the geochemical analysis is a machine learning technique known as transfer learning. We do not have enough samples in the data to create a new neural network, but what we can do is take a neural network that was trained on a related task (identifying some type of objects based on their chemical analysis) and retrain this neural network to identify rock samples using the data that we do have. This is contingent upon our ability to find a neural network that was pre-trained on a similar task. If this method works well on our training data we can then compare it with our discriminant analysis to see which method is better suited to the clients needs.

These methods are subject to change as our analysis progresses.

4. TIMELINE OF PROJECT MILESTONES

The analysis for this project will be completed this term. Jim and Charlie would like the team to present its findings in a presentation to their group. The date of the presentation is still to be determined. We plan on working on this project in the following order:

- Cleaning up the data (imputation, dimension reduction, etc)
- Exploratory data analysis (expected to be finished by 2/7/23)
- Classification based on MapUnit using various methods (expected to be finished by 2/21/23)
- Determination on if there is a suitable neural network available for transfer learning (expected to be finished by 2/7/23)
- Cross validation to test/compare accuracy of our methods (expected to be finished by 2/28/23)
- Comparing and analyzing our results and deciding which predictors are significant

Justin Valentine
Ephraim Romesberg
Ryan Foley

- Clustering the data (disregarding the MapUnit variable) and comparing clusters to MapUnit groups
- Midterm Update (2/21/23)
- Working out any issues in our analysis and consolidating our results
- Final Presentation (3/16/2023)