Thanks for the introduction Katie.

Hello Everyone. As Katie mentioned, My name is Anirudh Prabhu and today we are rounding up the Data Science for Geosciences Summer 2018 Series with a webinar on Analytics.

The Data Analytics Pipeline involves the following steps. You start by processing the raw data. We then use Descriptive Analytics to explore the data and gain basic information about the data. Next we try to understand the information stored in the data by exploring the various patterns and trends that may be hidden in the data. We can also use predictive analytics to explore potential future scenarios, and later use prescriptive analytics to recommend a course of action based on all the possible future scenarios.

In this webinar, we will explore Descriptive Models, Identification of patterns and relationships, and end with predictive analytics.

People always ask me, “What’s the difference between Analytics and Analysis”. Analysis focuses on transforming and modeling data for discovering useful information. While, Data Analytics is a methodology. In Analytics, we use the results/insights we obtain from the analyses, for example from descriptive or predictive models, to recommend a course of action or as a guide in scientific decision-making. Application of these methods is much art as it is science.

So, lets starts off by loading our datasets. We use the read.csv function for reading in our data from the file. The first table contains Uranium mineral data. As you can see, we have the name, the chemical formula, Age, Locality, latitude and longitude, the unit cell parameters and many more features. Since the previous part of this webinar series focused on Data processing and Cleaning, I have directly loaded a clean version of this dataset for analysis.

The second table is a Uranium Locality dataset where we can have a of geographical locations and the list of uranium minerals found at that locality.

Now that our data is loaded and ready. We begin with the first step, exploring the data. Descriptive or Exploratory Analysis help us understand the features in the data, how they are distributed and what maybe of importance. John tukey’s quote interestingly describes the process.

When we summarize the entire dataset, it can get a little overwhelming with the amount of information thrown on screen. But we can focus on specific features in which we are interested.

Like here, I can see the distribution of the Maximum Age of the minerals, the unit cell parameters of the minerals, and features describing color, and structural complexity of the minerals.

Simple visualizations like histograms and box plots help highlight any descriptive trends we might see in the data. For example the histogram of Max Age of the minerals shows us that most Uranium minerals are young, with the exception of a peak in the 1500 to 2000 million year range.

Similarly, the boxplot shows the variability of the unit cell parameter of Uranium minerals.

Once, we have a general understanding of the data, we then proceed pour over the data to find patterns and trends. Commonly used methods include Clustering and Association Rule Mining. There is a wide variety of algorithms available to perform either task.

Both of these methods are Unsupervised in their approach, since there are no outputs and main task is simply to extract the pattern and relationship in the data.

Clustering or Cluster Analysis is a method to group objects in classes or clusters best on patterns in the data. In this webinar, we are going to be using the “Partitioning Around Medoids” algorithm to perform said clustering.

The first step in the algorithm is to calculate and create a distance for the objects to be clustered.

We then use this distance matrix to run clustering simulations to find the optimum number of clustering. Silhouette width is a measure of how distinct each cluster is. The rule of thumb is, any cluster combination with a silhouette width of 0.35 or greater is considered well clustered. But as with most things in the real world, it is not easy or common to see high silhouette width values. But what this simulation tells us it that 2 clusters would be optimal for our dataset, with a silhouette width of ~0.17.

So next we go ahead, and run the clustering with k=2. The result is fairly well clustered group with a bit of overlap. While this is the result of the clustering algorithm. In the Analytics methodology, we usually explore these groups to find out which features were most important creating these 2 groups and what relationship can be found both between the groups and within each group.

Association Rule Mining, or Market Basket Analysis as it popular known, it an used to discover relationships between the various attributes in the data. These relationships are then encoded as “rules”, which another system can follow for any potential decision-making.

For the association rule mining, we use the second dataset which contain locality information. As you can see, we have locality name, geographical address, and most important a list of uranium minerals found in each locality.

This dataset needs to be restructured and organized to be used for mining association rules. So convert the dataset into a location-mineral binary matrix that indicates presence and absence of mineral at a certain locality.

The next step is run the Mine for association rules. Here we use the apriori algorithm, one of the oldest and most popular algorithms for association rule mining. The results we see are actual rules that give us insights. If we look at rule #3, if a locality has Torbernite, Uranophane-alpha, and Autunite, then Urananite should also be found there. Each rule has 3 metrics, support, confidence, and lift. The higher the lift, the more “interesting” or “novel” the rule.

There are also visualizations that summarize the rulesets created. Like this grouped matrix, Parallel Co-ordinate Plot and Network for the rules generated. These visualizations can also be explored for important inferences.

Now we get to probably the most popular branch of analytics. Predictive Analytics uses statistics, machine learning, data mining, and AI techniques to make assertions about the future.

The 2 main methods for predictive analytics are classification and regression, both supervised methods.

Supervised learning methods use data to train a model on what the correct outputs should be for a given input.

So now onto classification. Classification models are usually used when the predicted variables are categorical, ordinal or discrete. Here, we assign a new data point or observation to a set of known classes.

An example of classification can be seen when we create a model to predict the redox state of uranium minerals. We use the conditional inference tree algorithm to perform this task. We the model is trained we can see that, color, Paragenetic mode, topology and luster play an important part in predicting the redox state.

We now use the trained model to actually predict the redox state and compare it with known values to evaluate the accuracy of the model. We have a 93.82% accuracy, which is very good. But we must be careful before accepting good results at face value. A simple look into the data will tell you that the redox states of the minerals in highly biased towards the redox state 6. So it isn’t surprising that a highly accurate model has been created by mostly guessing redox state 6 as the result. This problem can be overcome by using other statistical techniques to take the bias into account while create a model.

People are at time confused about the difference between clustering and classification. This slide shows us the difference. The left side focused on the clustering where initially no classes are known and we come away to classes defined based on the similarity of features of the data. While on the right side, which focuses on classification, we have predefined groups or classes, and we predict the output classes for new inputs. I like to use the coin analogy while explaining the difference. Lets say you are given a bunch of coins and told to group them. You can group them in many ways, the size, shape, weight, color, year of minting, or the more obvious coin value. This is the clustering process, where you are not told how to group them. On the other hand, if you were told to group the coin only based on the monetary value of the coin. Then it could be a classification task.

Lastly, we cover regression. Regression is commonly used when the predicted variables are numeric and continuous.

Here we use the same conditional inference tree algorithm to perform a regression task by creating a model to predict the maximum age of a uranium mineral. We see the hardness, topology and paragenetic mode play an important in this model.

To validate the predict we look at the difference between the actual values and predicted values. What we can see is that this model performs poorly, since many values stray far away from the zero line.

Well, that about covers the topic for this webinar. The Data Analytics methodology still includes topics like prescriptive analytics, visual analytics, mixed models etc. , but that maybe covered at a later date.

All the slides in this talk along with the code have been created in a Jupyter Notebook. You can access your own Jupyter Deepcarbon account by logging in with your DCO account details.

Thank you so much for attending this webinar.