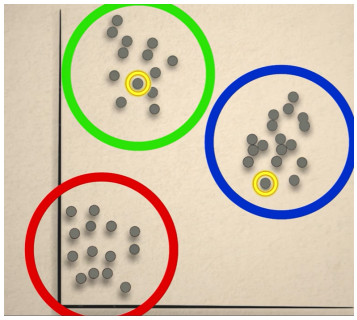
How to evaluate clustering?

We already know how to develop a K-Means algorithm. Now the important thing is to evaluate the output of a clustering, that is we want to know how good the output of the algorithm is.

Let us see, how to evaluate a clustering of a data,
Let's first talk back about the Supervised Learning problem of classification.
There we were given a set of data with labels and we wanted to predict the labels for some new data. So, for classification, the evaluation is straightforward if we have enough data.
There, we typically set aside some data where we know the labels. The data points that we feed to the algorithm together with their labels is called **Training Data**. And the data points that we set aside for evaluation purposes are called **Test Data**.
We use our classification algorithms to predict the labels on the test data and compare them to the true labels. Then we have an absolute universal scale to check what percentage of the test, did the algorithm predicted correctly.
We can easily compare other algorithms. A better algorithm gets more labels right. But we can also see whether an algorithm on its own is good or not. Did the algorithm get a significant proportion of the labels correct or not?

Now sometimes in Clustering, one may have access to a ground truth clustering of the data. In the archaeology example, perhaps in some cases, the artifacts clusters were known. So, we can compare to the known clusters or perhaps someone was paid to cluster some images manually by hand.



It's still not quite as easy to evaluate clustering as evaluating a classification.
Since, any permutations of the labels leads to an equally good clustering.
The only thing that defines the clustering is whether two data points belong to the same clusters or to different clusters in both our algorithm and ground truth.

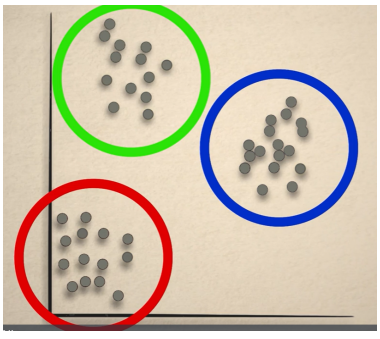
For example,
In figure, the two marked data points (in yellow) belong to different clusters.
There are some measurements like **Rand Index** and **Adjusted Rand Index** that help us measure whether the clustering found by our algorithm really captures the information in the ground truth clustering.

They calculate it by counting the pairs of data points that belong to the same or different clusters according to the algorithm and the same or different clusters according to the ground truth. Then the value is normalized.

Measures like the Rand Index are called **External Evaluation**, because they require **outside information** about a **ground truth clustering.**
The problem of evaluation is still even trickier in Clustering, because there is rarely any ground truth information that we can use for testing.

Then, how to evaluate whether a clustering is good or whether one cluster is better than another?
The answer is, we can check K-Means Objective function value across different clusters. It lets us compare two clustering. But it does not tell us whether a particular clustering is good or not on its own. It also might not capture exactly the patterns which we are looking for in the data.

So, the evaluation of Clustering is Hard! But, researchers have developed a number of useful heuristics.
The first and most useful observation is that, we are often trying to find some particular meaning, latent pattern in our data via clustering.
We might be able to check the pattern by visualizing the result of the clustering.

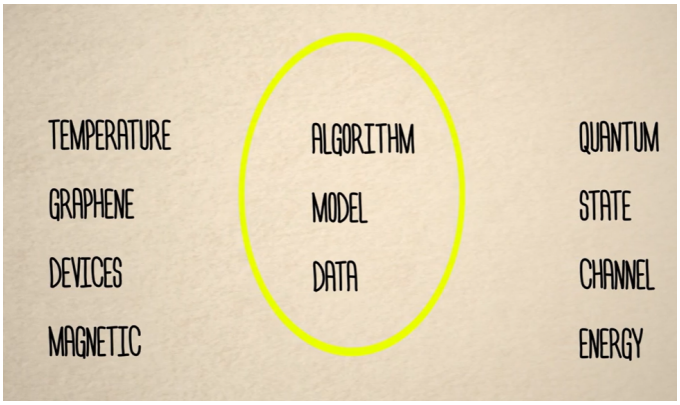


For instance:

- For the archaeology example, it was obvious what the cluster should be.
- Similarly for the image segmentation example, we would like our clustering to distinguish the three

different types of objects in the image. And by visualizing the clustering of points we can easily check this.

- Consider the topic modeling example, where the words were clustered together. We can list out the clusters of words and ask ourselves do these words go together or do the words have any meaningful association ?



Another option we have for evaluation is to use special tools for visualization. Like the GGOBI tool.

We might have a dataset where D (= Number of dimensions) is much higher than 2.
For example we might have collected data on a large number of different health metrics like age, daily calories consumed, daily water consumed, weekly alcohol consumed, weekly miles driven, weekly exercise in minutes and so on.

But this data is not an image or a text data set. So, it might be hard to plot meaningful pictures of the data.
GGOBI tools help us find meaningful visualizations of high dimensional data for evaluating the clustering.

While these methods are often more qualitative than quantitative on their own. They can be made more quantitative by incorporating some form of crowdsourcing. For example, Amazon Mechanical Turk workers could be asked to weigh in on the quality of the clustering. But even with the help of Amazon Mechanical Turk or other platforms human evaluation can be expensive in terms of both time and money.

So, we may consider some other automated forms of evaluation.
By contrast to External Evaluation, there are also a number of measures for internal evaluation that depend only on the data at hand.

The basic idea of this measure is to typically make sure that data within a cluster are relatively close to each other and data in different clusters are relatively far from each other. An example of this is the Silhouette Coefficient.

Another example is to split the data into two data sets and then applying clustering to each and then computing the clustering found across the two sub data sets.

So, in this article, we have discussed some of the ways to evaluate the output of the K-means Algorithm and clustering algorithm in general.