What is Clustering?

We know that Unsupervised Learning means finding hidden patterns in the data when there are no labels present in the dataset. There are a lot of different problems that we can solve using Unsupervised Learning.

But the most popular version of Unsupervised Learning is **Clustering**.

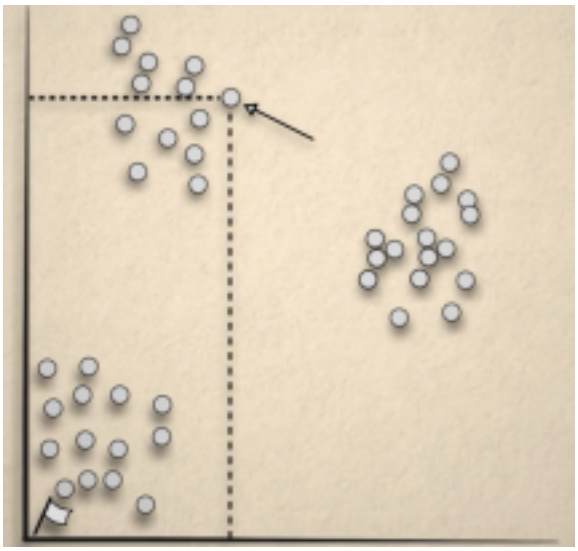
What is Clustering?

In clustering, our goal is to group the data according to similarity.

Let's take an example to understand the keywords **Group**, **Data,** and **Similarity** in more detail.

Data:

Imagine you are leading an archaeological dig. And every time someone on your team unearths an artifact, you have them record the position of the artifact.



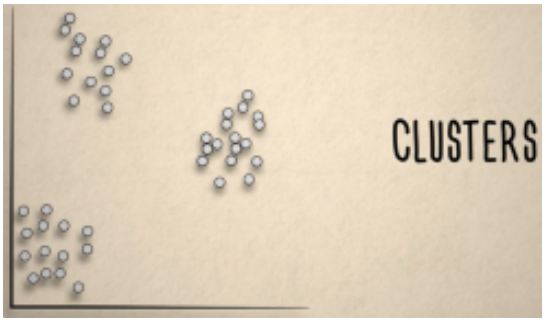
In particular, you can put down some marker near the archaeological site and you can record how far north and how far east each artifact was found relative to your marker. The location for one artifact will be a data point, and when you list all of the locations of the different artifacts, then you get the dataset.

Suppose you consulted with a few historians before your archaeological dig and you come to know that three families lived in the area that you are exploring.
Now, from the first figure, it is easily identifiable where the three families lived and which artifact belongs to which family. To understand the process even better, we need to talk about "**Similarity**".

There is a notion of similarity here. In this archaeology example, we say that two artifacts are similar or close if they are close in terms of physical distance. When we think about where the three families lived, we are grouping the data according to this distance.

Here, we are assigning each data point to only one family group. So, data points that are physically close, tend to belong to the same group.
When we assign each data point to only one group like this, then we call the groups as **Clusters.**

So, Clustering is the Unsupervised Learning. It is the problem of assigning each data point to exactly one group or cluster. Also, we want the clusters to be meaningful, that is we want the clusters to give us some interesting insights into our dataset.

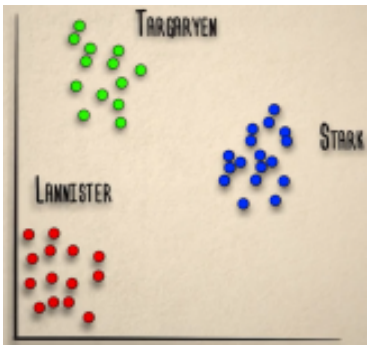


In the example that we have discussed, we imagined the clusters correspond to 3 different families who used to live in the examined area.

The data points in any cluster are for example pottery shards or other items that we Believe that a single-family had discarded when they lived at that spot. Identifying the items that we think belong to one family will let us perform further scientific or archaeological or historical analysis of the artifacts.

Classification vs Clustering:

We already know about Supervised Learning Problems. There is a Supervised Learning problem that is somewhat similar to clustering. The problem is called **Classification.**
In classification, like in Supervised Learning in general, we are given labels.



For instance, consider the archaeology example, if we run classification on this data, we might not only know that there are 3 families, but we might know their names as well. For example, the names may be Lannister, Stark, and Targaryen. Here the labels are Categorical. So, the labels have no order.
For the archaeology example, if we had a classification problem there, then we would have been given a bunch of data points with labels. For instance, we might imagine each family actually made a lot of pots with a sigil or some other identifying mark.

Then our goal would be to find labels for any remaining unlabeled artifacts. So, our goal will be to predict labels for new data points given the labels of all data points.

But in **Clustering,** we don't know the families or their names. We might know that there are 3 families, but we don't know anything at all beyond that number. And none of the artifacts come with labels. So, we have to discover how to group the artifacts from their locations alone.

In many cases, labels may be hard to come by for various reasons. For the archaeology example, we might not be lucky enough for families from thousands of years ago to have carefully labeled their pottery for our benefit.

But on the positive side, Clustering lets us find hidden groupings in data even when we can't run classification. But Clustering can be a more difficult undertaking than Classification, since we don't have the information contained in the labels.

So, the Clustering problem may get complex when we deal with higher dimensional data.

Here for the archaeology example, we have seen that our data was numerical. But our data need not always be numerical. It might take the form of words or pictures or genomes or something else entirely.

And it is not so easy to immediately see the hidden patterns in these types of data. But it turns out that we can still get a lot of information about the data from Clustering.