Why and When to use Clustering?

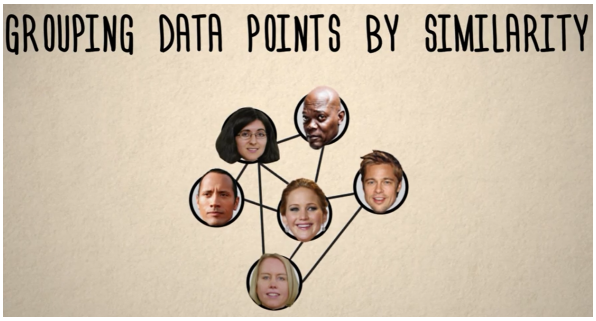
We already know that clustering is a particular form of unsupervised learning, and generally, in unsupervised learning, we try to find any hidden pattern in a set of data points without labels.

In clustering, we want to find a latent grouping such that each data point belongs to exactly one group called a cluster.

We have already seen an example of clustering in action, which is how we might use clustering to find the arrangement of artifacts at an archaeological dig.

Now we will look into some more examples of clustering and understand why and when we should use clustering in practice.

Sometimes we have a collection of data, but we are not quite sure what to do with it yet. We might want to explore the data without a particular end goal in mind. Perhaps the data will suggest some interesting approaches for further analysis.



In this case, we say that we are performing Exploratory Data Analysis.

For instance, suppose we have some data points of users on meetup.com. Roughly each data point corresponds to a person. In order to apply clustering, we have to define a **notion of similarity** too.

Suppose, the similarity is the number of common interests between two people. We can't actually feed a user or person into our computer. So, we need to be more precise about what constitutes a data point in this case.

For the data points, perhaps what we have actually collected for each person is a list or vector of whether that person is interested in each meetup.com group on the meetup website.

With this dataset, it is easy to calculate the similarity between two users. **It's the total number of groups those two people are both interested in.**

After we apply clustering, we might hope to find groups of users with broadly similar interests. Maybe after running clustering, it may turn out that, for example, some of the users are interested in musical performance, some are interested in outdoor sports. Also, some users are interested in hackathons and so on.

So, we have found some patterns or groupings from our dataset.

Now the question is,

What are some other times when we might use clustering?

So basically, we can use clustering for problems that kind of look like classification, but we don't have labels in advance for the dataset.

***Example**:*

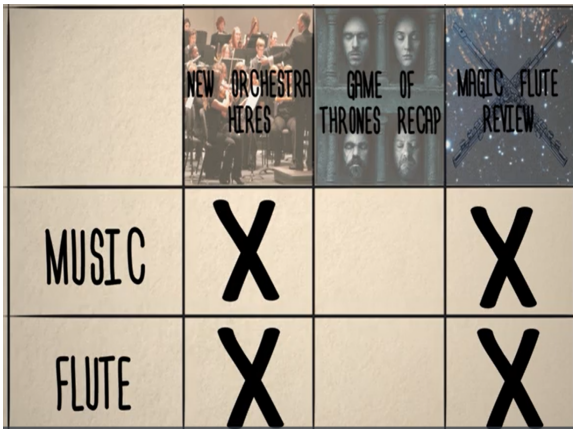
Suppose we have all the documents on Wikipedia or all the past articles from WIRED magazine. Now we would like to find out topics or themes that are represented in these documents.

In general, we don't know the possible topics in advance. So, we might guess at some topics. But we can't be sure in advance that we have got all the topics that would really appear in the corpus.

So, to discover the topics we can apply clusters. We can say each data point corresponds to a word and **we can measure the similarity of two words by counting how many documents they both appear in. And we can group together similar words to form the cluster.**

Here the data points that correspond to any particular word could be a list of vectors of whether this word occurs in each document, by labeling '**Yes**' or '**No**'.

Then we can actually calculate the overall similarity of the two words.

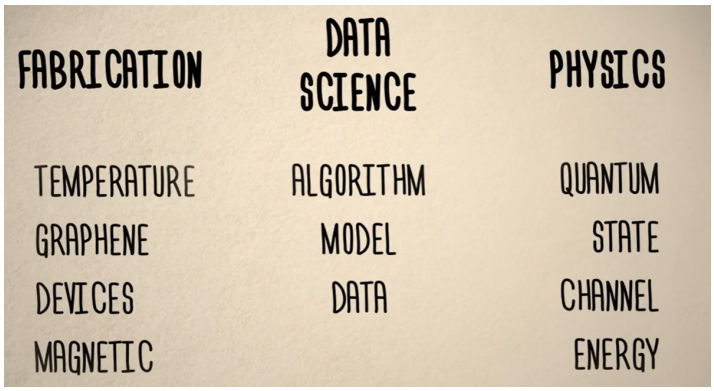


Let's see another example:

Some students at MIT looked at research abstracts from Professors of the Electrical and Computer Science department. And they found some patterns of some words which tend to co-occur.

Words like temperature, graphene, devices, and magnetic go together. Words like algorithm, model, and data go together. And words like quantum, state, channel, and energy go together. It turns out that there are some groupings among these words. The first set of words are related to fabrication, the second set of words are related to Data Science and similarly, the last set of words are related to physics.

The interesting thing here is, the students didn't use any knowledge in advance about what MIT professors are working on.



The different topics in the data came out in an automated way.

So, this is one of the examples where we don't know the labels of our clusters in advance.

But sometimes we don't have labels on the data because it's expensive to label data. It might take a long time to manually label each data point. And the labels might be changing too quickly, so it's tough to keep up with it.

Take the example of Google News. We might want to put news documents about similar topics in a cluster. But the topics in the news are changing every day. For instance, just hours after a new Game of Thrones episode comes out, there are many great articles available about that episode.

We would like to detect all those articles and group them together. And also we would like to do this even when a surprise wild event happens, which we didn't anticipate in advance.

In this case, our data points might be a list of the topics which occur in a given document, and similarity measures between two documents might be accounted for by looking at how many topics are shared by the two documents or how similar the topic proportions are between the two documents?

Alternatively, it might also be expensive to label data simply because there is way too much data available.

So in summary, here in this article we have seen a lot of different reasons to use clustering. We have also seen a lot of different types of data where we might apply clustering.