In this video, we'll discuss the method of hierarchical clustering.

In hierarchical clustering, the clusters are formed by each data point starting in its own cluster.

As a small example, suppose we have five data points.

Each data point is labeled as belonging in its own cluster.

So this data point's in the red cluster, this one's in the blue cluster, this one's in the purple cluster, this one's in the green cluster, and this one's in the yellow cluster.

Then hierarchical clustering combines the two nearest clusters into one cluster.

We'll use Euclidean and Centroid distances to decide which two clusters are the closest.

In our example, the green and yellow clusters are closest together.

So we would combine these two clusters into one cluster.

So now the green cluster has two points, and the yellow cluster is gone.

Now this process repeats.

We again find the two nearest clusters, which this time are the green cluster and the purple cluster, and we combine them into one cluster.

Now the green cluster has three points, and the purple cluster is gone.

Now the two nearest clusters are the red and blue clusters.

So we would combine these two clusters into one cluster, the red cluster.

So now we have just two clusters, the red one and the green one.

So now the final step is to combine these two clusters into one cluster.

So at the end of hierarchical clustering, all of our data points are in a single cluster.

The hierarchical cluster process can be displayed through what's called a dendrogram.

The data points are listed along the bottom, and the lines show how the clusters were combined.

The height of the lines represents how far apart the clusters were when they were combined.

So points 1 and 4 were pretty close together when they were combined.

But when we combined the two clusters at the end, they were significantly farther apart.

We can use a dendrogram to decide how many clusters we want for our final clustering model.

This dendrogram shows the clustering process with ten data points.

The easiest way to pick the number of clusters you want is to draw a horizontal line across the dendrogram.

The number of vertical lines that line crosses is the number of clusters there will be.

In this case, our line crosses two vertical lines, meaning that we will have two clusters-- one cluster with points 5, 2, and 7, and one cluster with the remaining points.

The farthest this horizontal line can move up and down in the dendrogram without hitting one of the horizontal lines of the dendrogram, the better that choice of the number of clusters is.

If we instead selected three clusters, this line can't move as far up and down without hitting horizontal lines in the dendrogram.

This probably means that the two cluster choice is better.

But when picking the number of clusters, you should also consider how many clusters make sense for the particular application you're working with.

After selecting the number of clusters you want, you should analyze your clusters to see if they're meaningful.

This can be done by looking at basic statistics in each cluster, like the mean, maximum, and minimum values in each cluster and each variable.

You can also check to see if the clusters have a feature in common that was not used in the clustering, like an outcome variable.

This often indicates that your clusters might help improve a predictive model.

In the next video, we'll cluster our movies by genre, and then analyze our clusters to see how they can be used to perform content filtering.