In this lecture, we'll be using data from MovieLens to explain clustering and perform content filtering.

movielens.org is a movie recommendation website run by the GroupLens research lab at the University of Minnesota.

They collect user preferences about movies and do collaborative filtering to make recommendations to users, based on the similarities between users.

We'll use their movie database to do content filtering using a technique called clustering.

First, let's discuss what data we have.

Movies in the MovieLens data set are categorized as belonging to different genres.

There are 18 different genres as well as an unknown category.

The genres include crime, musical, mystery, and children's.

Each movie may belong to many different genres.

So a movie could be classified as drama, adventure, and sci-fi.

The question we want to answer is, can we systematically find groups of movies with similar sets of genres?

To answer this question, we'll use a method called clustering.

Clustering is different from the other analytics methods we've covered so far.

It's called an unsupervised learning method.

This means that we're just trying to segment the data into similar groups, instead of trying to predict an outcome.

In this image on the slide, based on the locations of points, we've divided them into three clusters-- a blue cluster, a red cluster, and a yellow cluster.

This is the goal of clustering-- to put each data point into a group with similar values in the data.

A clustering algorithm does not predict anything.

However, clustering can be used to improve predictive methods.

You can cluster the data into similar groups and then build a predictive model for each group.

This can often improve the accuracy of predictive methods.

But as a warning, be careful not to over-fit your model to the training set.

This works best for large data sets.

There are many different algorithms for clustering.

They differ in what makes a cluster and how the clusters are found.

In this class, we'll cover hierarchical clustering and K-means clustering.

In this lecture, we'll discuss hierarchical clustering.

And in the next lecture, we'll discuss K-means clustering.

You'll learn how to create clusters using either method in R. There are other clustering methods also, but hierarchical and K-means are two of the most popular methods.

To cluster data points, we need to compute how similar the points are.

This is done by computing the distance between points, which we'll discuss in the next video.