**Microsoft:** DAT210x Programming with Python for Data Science

Help

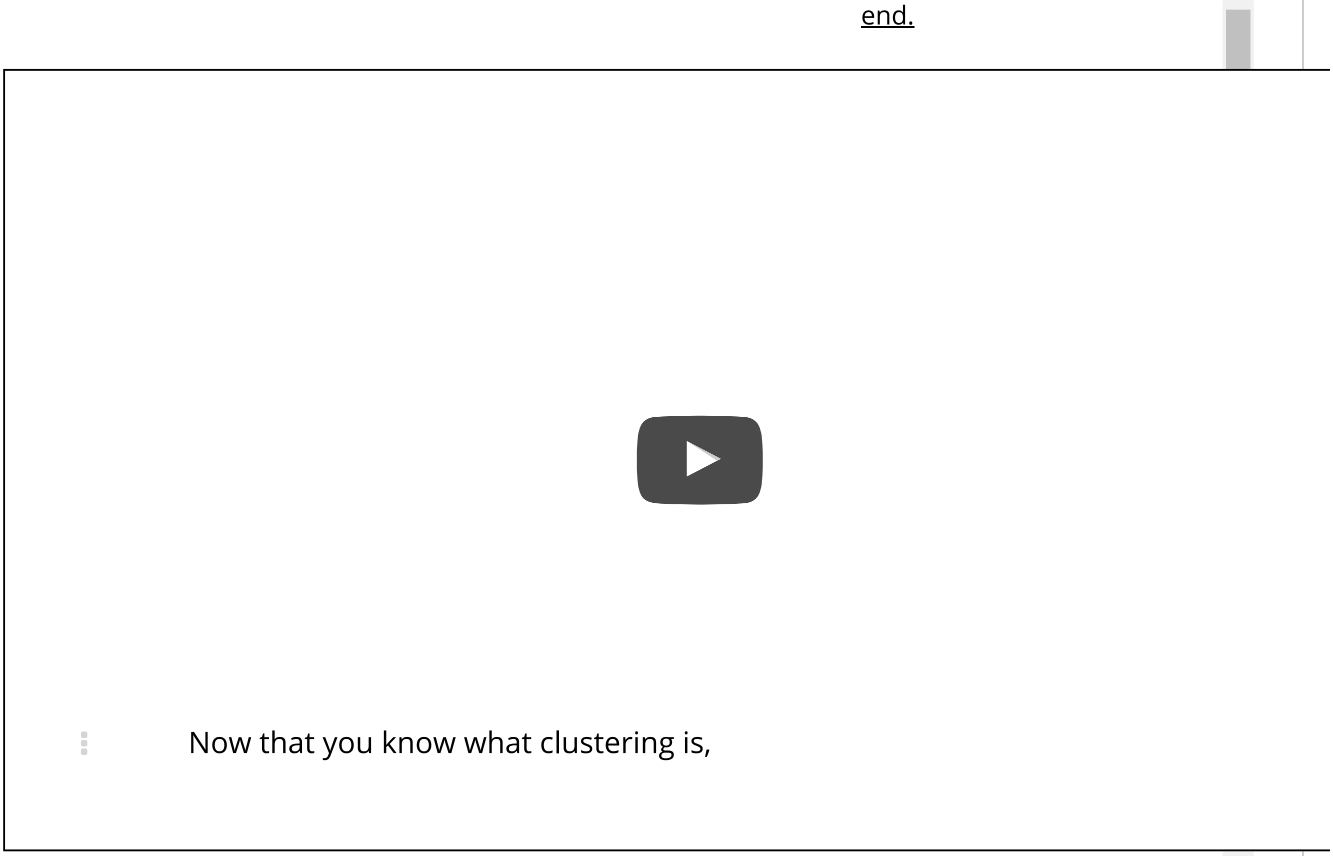
Course › 5. Data Modeling › Lecture: Clustering › Video

# Video

Bookmark this page

## When Should I Use K-Means?

Start of transcript. Skip to the end.



Now that you know what clustering is,

I like to buy a lot of science fiction books, history books,

zombie books and the like thereof.

### Video

Download video file

### Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Clustering is a natural action we do even as children, by arranging similar shaped blocks and colors. K-Means clustering is best suited when you have a good idea of the number of distinct clusters your unlabeled dataset should be segmented into. Generally, the output of K-Means is used in two ways. To separate your unlabeled data into K groups, which is the clear use case, or to find and use the resulting centroids.

## Separate Your Data

Astronomers use clustering to group different star types, classes of planets, and galaxies. Biologists use it to group every living thing by species, genus, and kingdom. In business, clustering is used to segment likely and unlikely prospects, for location assignment, factor endowment, and the assignment and deployment of remote services.

## Centroid Usage

Besides divvying up samples, clustering can also provide a layer of abstraction, by directing attention to the cluster and its attributes and not each samples. In the climate change case study from the previous module, you saw how climate divisions were used as a cluster abstraction over individual ground stations for various mentioned reasons. Another example of centroid usage would be a company looking for ideal locations to open a limited number of branches, based on the location of their customers.

You can use the centroid to 'compress' your data. By referring to the centroid rather than the data sample, the number of unique values is reduced, which optimizes the execution speed of other algorithms. Isomap, for instance, uses a nearest neighbors algorithm to calculate the distance from the record you want to transform to every sample in the training dataset. By using the record-to-cluster distance approximation in replacement of the individual record-to-sample distances, since there are far fewer clusters than records, you can achieve unprecedented orders of optimization.

English ▼

POWERED BY OPENedX