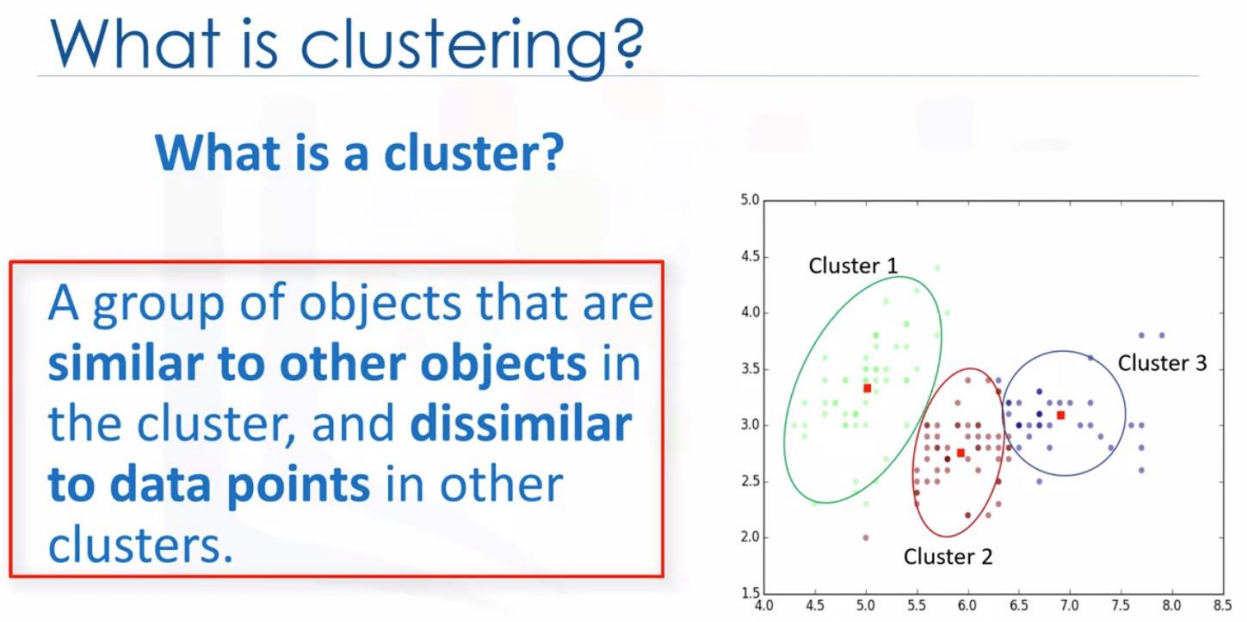
Clustering can group data only unsupervised, based on the similarity of customers to each other.

It will partition your customers into mutually exclusive groups. The customers in each cluster are similar to each other demographically. Now we can create a profile for each group, considering the common characteristics of each cluster.



Classification VS Clustering

1. Classification algorithms predict categorical classed labels. This means assigning instances to predefined classes such as defaulted or not defaulted.

   For example, if an analyst wants to analyze customer data in order to know which customers might default on their payments, she uses a labeled dataset as training data, and uses classification approaches such as decision tree, support vector machines or logistic regression to predict the default value for a new or unknown customer.

   Generally speaking, classification is a supervised learning where training data instance belongs to a particular class.
2. In clustering, data is unlabeled and the process is unsupervised.

   For example, we can use a clustering algorithm, such as k-means to group similar customers as mentioned, and assign them to a cluster, based on whether they share similar attributes.

Why clustering?

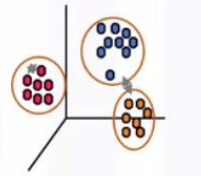
- Exploratory data analysis
- Summary generation
- Outlier detection
- Finding duplicates
- Pre-processing step

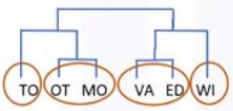
Summary generation or reducing the scale

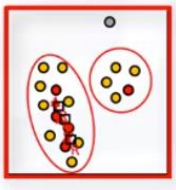
Outlier detection, especially to be used for fraud detection or noise removal

Clustering algorithms

- Partitioned-based Clustering
  - Relatively efficient
  - E.g. k-Means, k-Median, Fuzzy c-Means
- Hierarchical Clustering
  - Produces trees of clusters
  - E.g. Agglomerative, Divisive
- Density-based Clustering
  - Produces arbitrary shaped clusters
  - E.g. DBSCAN







1. Partitioned-based clustering: used for medium and large sized databases
2. Hierarchical clustering: small size datasets
3. Density-based clustering: produce arbitrary shaped clusters. Especially good when dealing with spatial clusters or when there is noise in dataset. E.g. DB scan algorithm