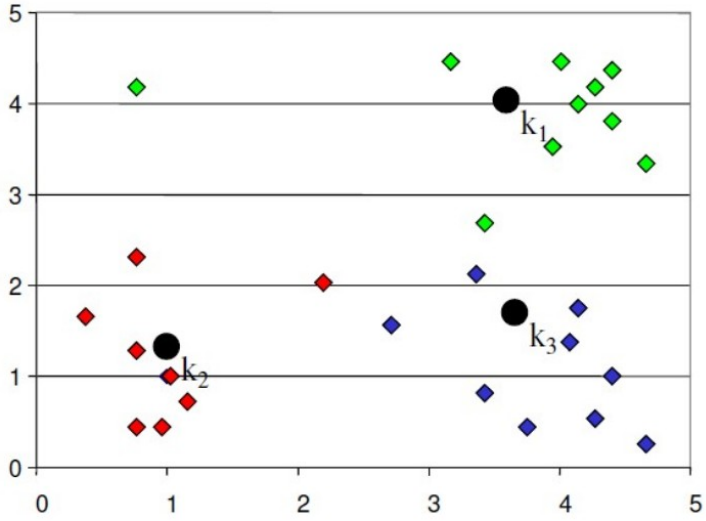
Overview of Clustering

What is Clustering?

So far, everything we have looked at has been **supervised learning** where our aim has been to make a prediction based on numerous factors (e.g., predict home price based on square footage, number of bedrooms, etc). In supervised learning, we have a *label* for our data (e.g., home price) that we use as our response.

Clustering is **unsupervised learning** where we do not have labels (responses) for our data. Instead, the aim is to divide the data into groups such that data in the same groups are more similar to each other than to those in other groups. In other words, we want to create groups with similar traits - these groups are called clusters.



Examples of Clustering

- Customer Segmentation - finding groups of customers with similar behavior
- Geographic - grouping together similar cities based on demographic information
- NLP - finding documents that are similar to each other
- Social media - identifying similar twitter users based on tweets
- Stack overflow - clustering together similar questions to give users suggested questions
- Medical - identifying unknown cancers

Why Use Clustering?

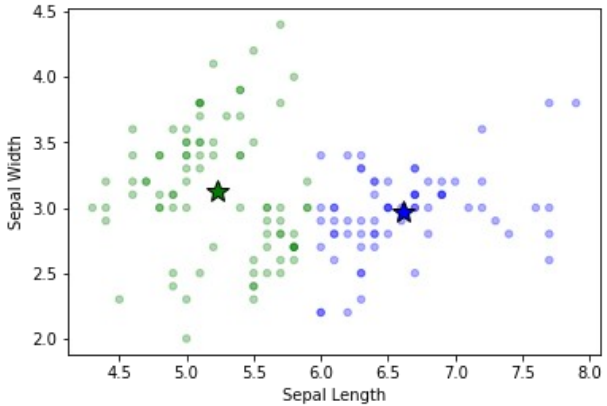
Why would we want to cluster instead of make a prediction? There can be several reasons, including:

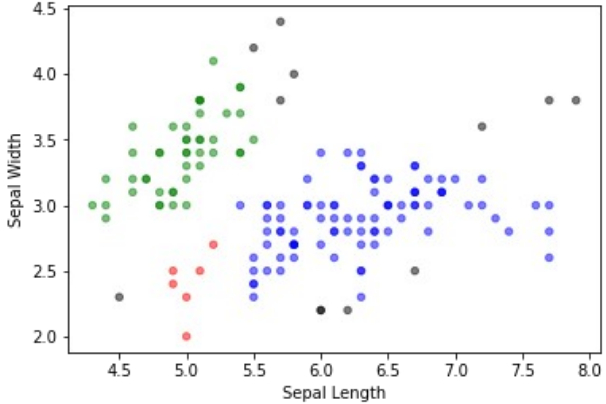
- Labels of data might be unknown
- It may be costly to label large amounts of data
- It can be useful to identify patterns or subclasses in the data

Types of Clustering (1/2)

There are numerous clustering algorithms, but many commonly used methods are broadly divided into the following categories:

- **Centroid-based Clustering** Creates clusters centered around a central point (called the centroid).
- **Density-based Clustering** Creates clusters based on areas of high density. Allows for arbitrarily-shaped clusters. Can work well if you have outlier points (doesn't assign outliers to clusters).

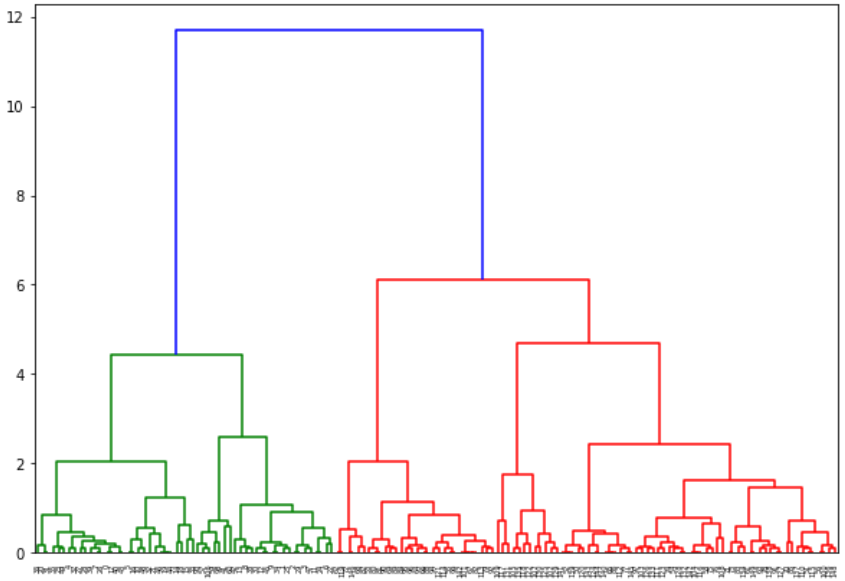




Types of Clustering (2/2)

- **Hierarchical clustering** Creates a tree of clusters. They do not partition the data like other methods, but instead provide a hierarchy of clusters that merge together at certain distances.

- **Distribution-based Clustering** Creates clusters based on probability distribution models



Clustering Algorithms

In this class, we are going to focus on a few commonly used techniques:

- K Nearest Neighbors (not really clustering, but can be used to find similar points. Generally used for classification/regression)
- K Means
- DBSCAN
- Hierarchical Clustering

There are many others. The scikit-learn website has a good overview here.