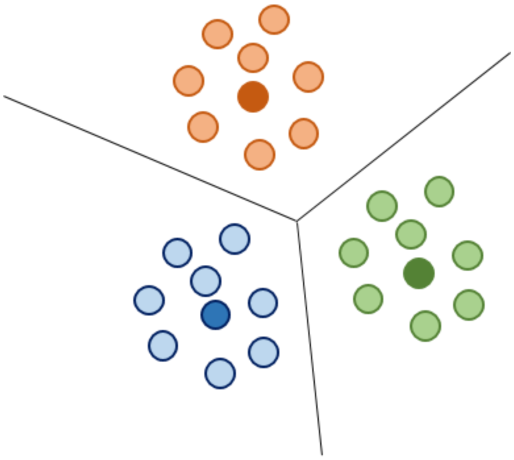
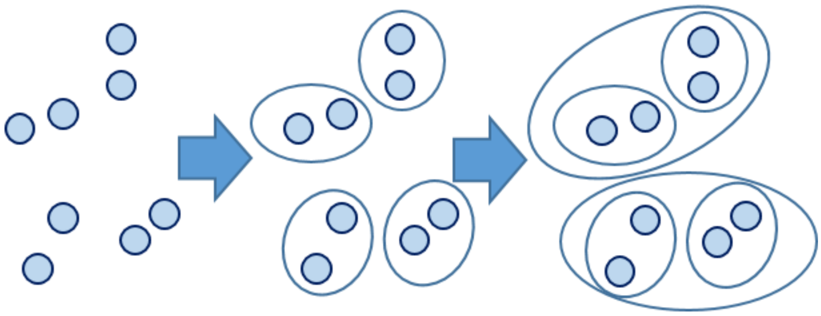
KEY POINTS

- Clustering groups data entities based on their feature values. Each data entity is described as a vector of one or more numeric features (***x***), which can be used to plot the entity as a specific point.
- K-Means clustering is a technique in which the algorithm groups the data entities into a specified number of clusters (***k***). To begin, the algorithm selects ***k*** random *centroid* points, and assigns each data entity to a cluster based on the shortest distance to a centroid. Each centroid is then moved to the central point (i.e. the *mean*) of its cluster, and the distances between the data entities and the new centroids are evaluated, with entities reassigned to another cluster if it is closer. This process continues until every data entity is closer to the mean centroid of its cluster than to the centroid of any other cluster. The following image shows the result of K=Means clustering with a ***k*** value of 3:



- Hierarchical Agglomerative Clustering is an alternative clustering technique in which the point representing each data entity begins as its own cluster. Each cluster is then merged with the cluster closest to it, and this merging process continues iteratively. The following image illustrates the process of Hierarchical Agglomerative Clustering.



- The distance metric used to determine how "close" points are to one another is an important aspect of clustering. The simplest way to conceptualize the entities is as points in Euclidean space (multidimensional coordinates), and measure the simple distance between the points; but other distance metrics can also be used.