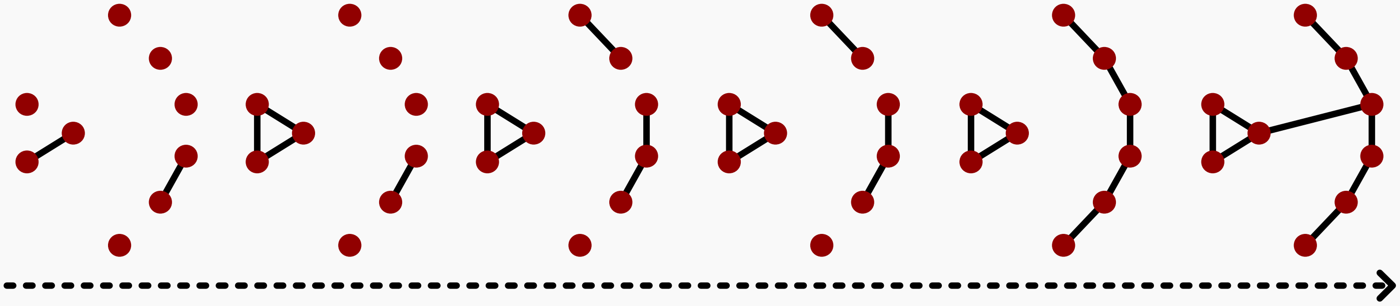
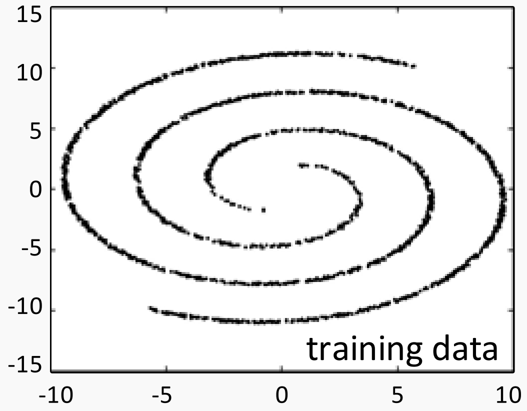
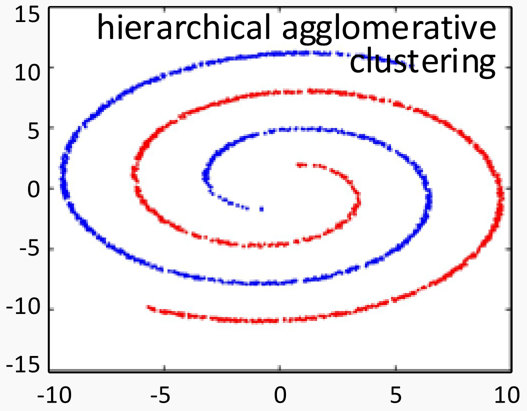
hierarchical agglomerative clustering

- Begins with each datapoint in its own cluster
- Repeatedly merges the clusters of the two closest points

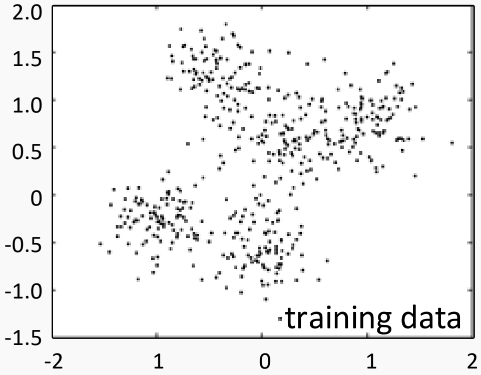


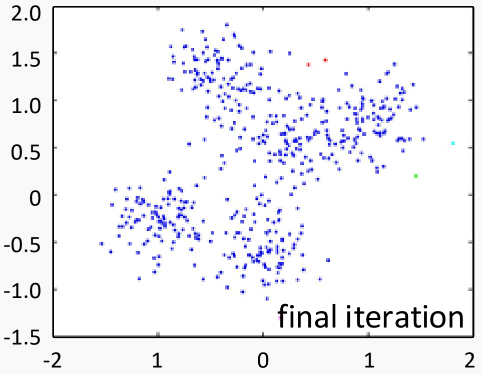
Hierarchical Agglomerative Clustering can handle data that **K-Means** fails to converge effectively:

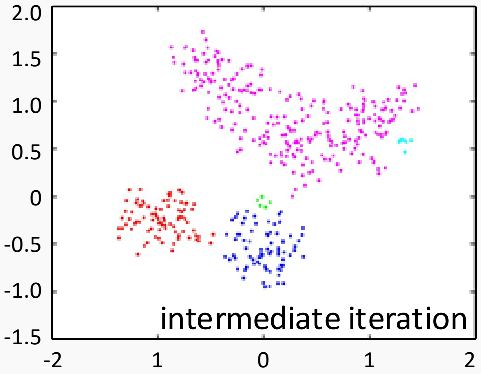




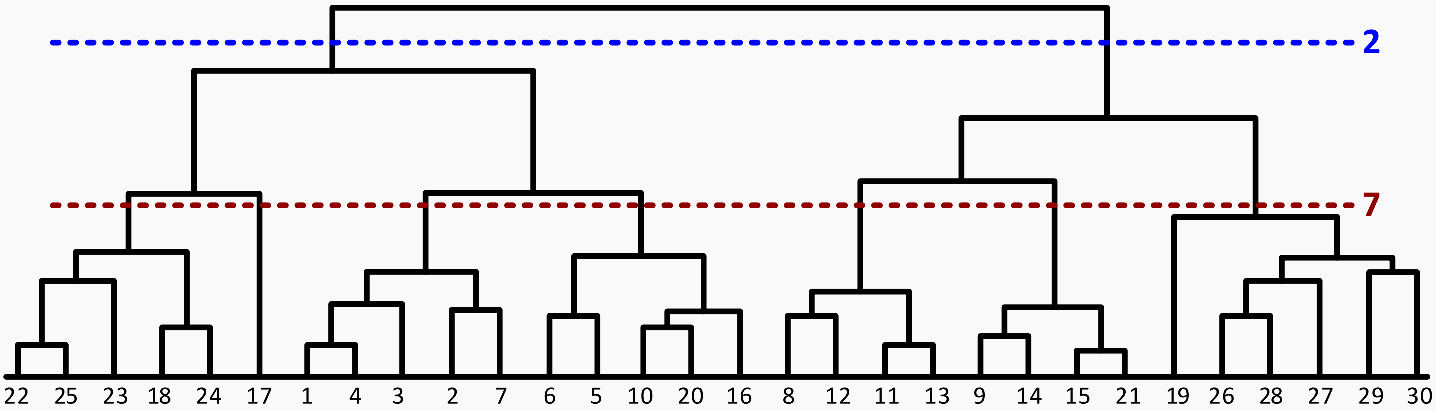
However, **Hierarchical Agglomerative Clustering** can also fail to converge on datasets the **K-Means** performs accurately on:







Dendrogram for Hierarchical Agglomerative Clustering

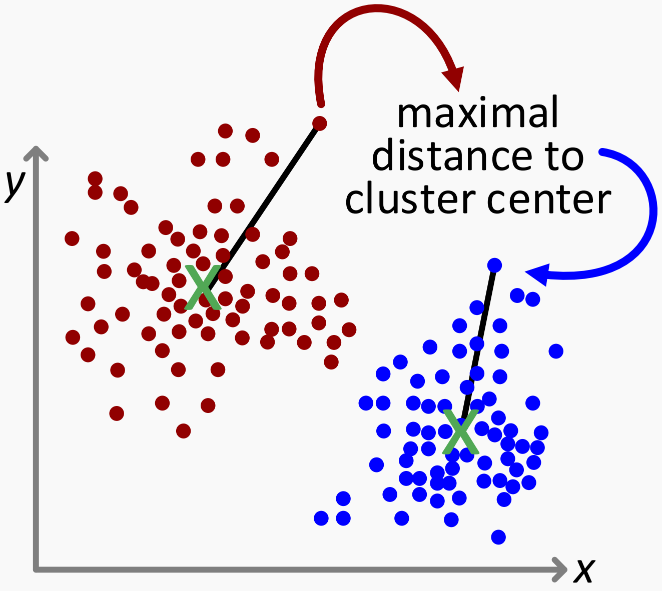


Hierarchical Agglomerative Clustering produces a **dendrogram** that map the nature of clusters.

The **dendrogram** can be pruned to control the number of clusters assigned to the data (seen above).

A problem with Hierarchical Agglomerative Clustering (HAC)

- The number of points in each cluster
    - If the number of points assigned to clusters **<** the total number data points available, some data could fail to be assigned to a cluster with the **Hierarchical Agglomerative Clustering** algorithm.
    - K-Means will not experience the aforementioned problem.
- Maximal distance to the cluster center
    - The maximal distance to cluster center is the largest distance of any point in a cluster
    - If the maximal distance is ≫ **much larger** ≫ then the average distance to the cluster center, the cluster may be overly dispersed.



K-Means Compared to Hierarchical Agglomerative Clustering

- Both are useful for different types of problems. **K-Means** works well with **spherical data**.
- **Hierarchical Agglomerative Clustering** is useful when clusters are **well-separated**. (Meaning data close together should be in the same cluster.)
- For **K-means**, one needs to choose the number of clusters.
  For **hierarchical clustering**, one chooses when to stop merging clusters.
- The **distance metric** is important and can have a large impact on the solution.