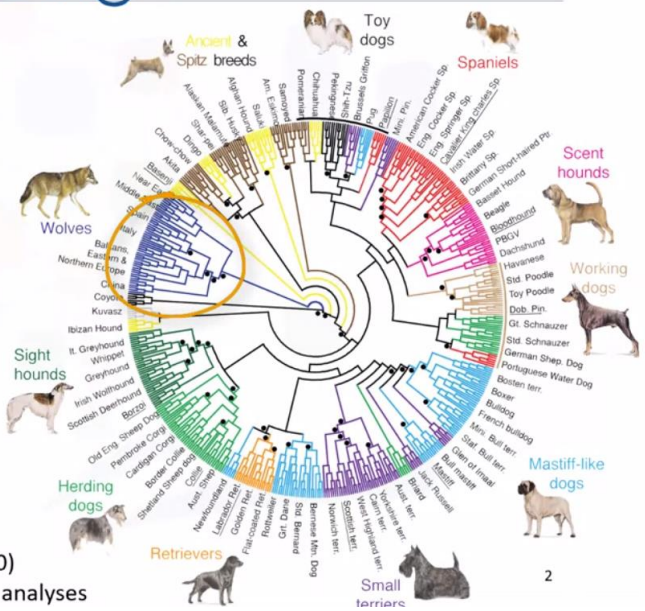
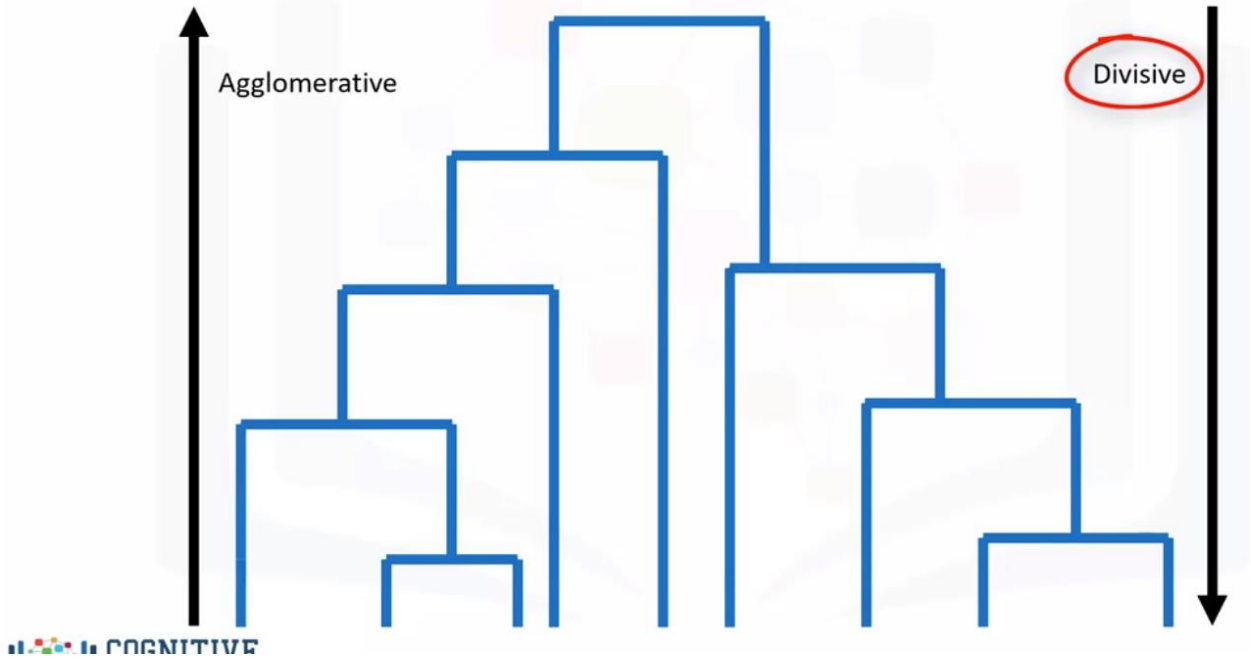
Hierarchical clustering

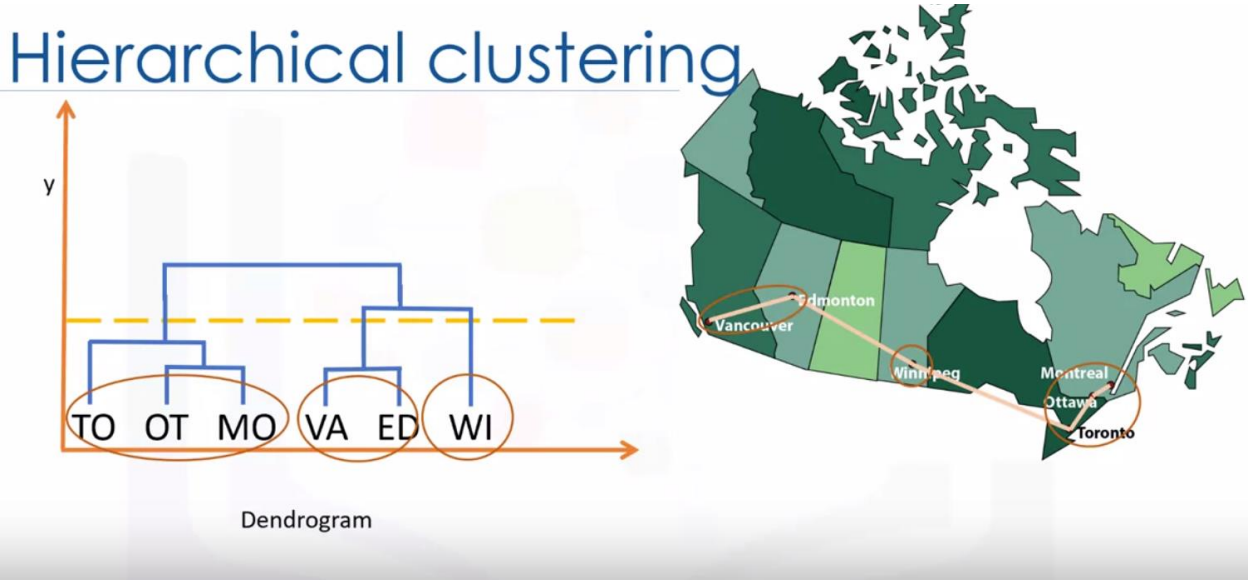
Hierarchical clustering algorithms build a hierarchy of clusters where each node is a cluster consists of the clusters of its daughter nodes.



Source: von Holdt B.M. et al. (2010)
Genome-wide SNP and haplotype analyses

Hierarchical clustering



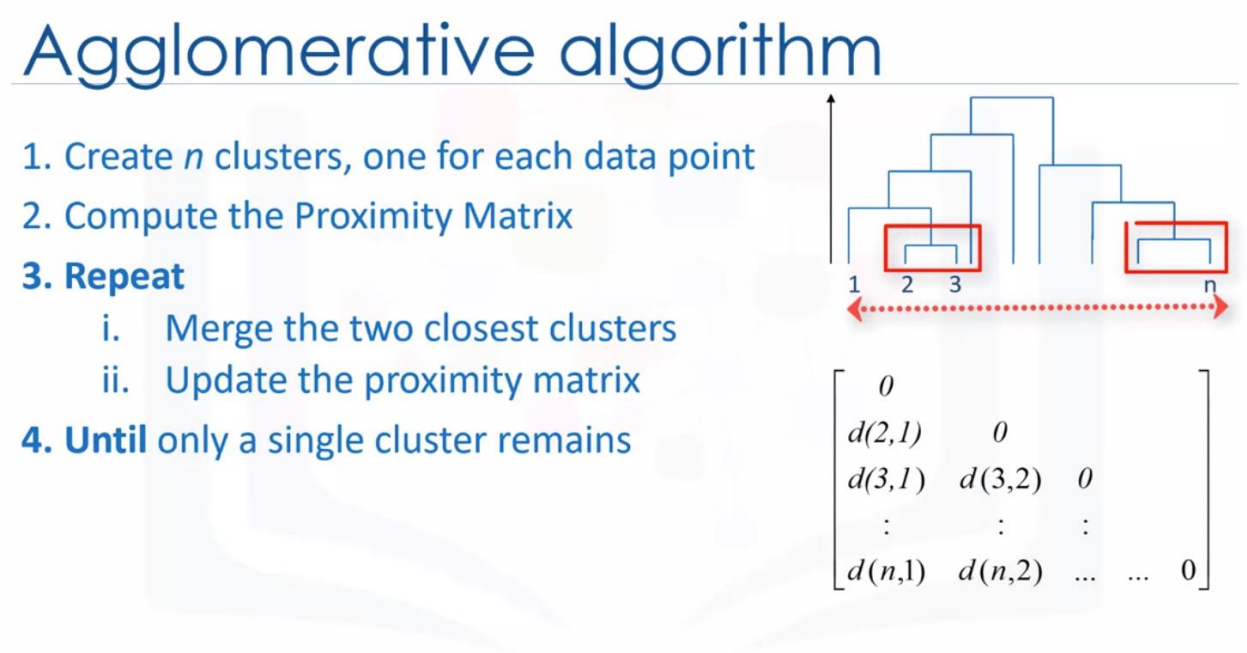


Hierarchical clustering doesn't require a prespecified number of clusters.

However, in some applications, we want a partition of disjoint clusters just as in flat clustering.

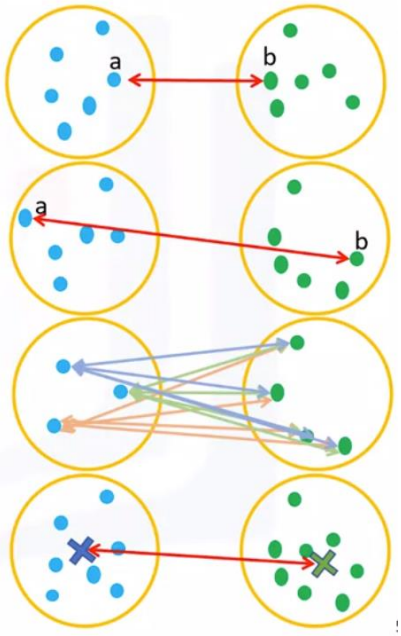
In those cases, the hierarchy needs to be cut at some point. For example here, cutting in a specific level of similarity, we create 3 clusters of similar cities.

How do we calculate the distance from Winnipeg to the Ottawa/Montreal cluster? Let's assume for example, we just select the distance from the center of the Ottawa/Montreal cluster to Winnipeg.



Distance between clusters

- Single-Linkage Clustering
  - Minimum distance between clusters
- Complete-Linkage Clustering
  - Maximum distance between clusters
- Average Linkage Clustering
  - Average distance between clusters
- Centroid Linkage Clustering
  - Distance between cluster centroids



COGNITIVE CLASS

Centroid is the average of the feature sets of points in a cluster.

Advantages vs. disadvantages

Advantages	Disadvantages
Doesn't required number of clusters to be specified.	Can never undo any previous steps throughout the algorithm.
Easy to implement.	Generally has long runtimes.
Produces a dendrogram, which helps with understanding the data.	Sometimes difficult to identify the number of clusters by the dendrogram.

Hierarchical clustering Vs. *K*-means

K-means	Hierarchical Clustering
1. Much more efficient	1. Can be slow for large datasets
2. Requires the number of clusters to be specified	2. Does not require the number of clusters to run
3. Gives only one partitioning of the data based on the predefined number of clusters	3. Gives more than one partitioning depending on the resolution
4. Potentially returns different clusters each time it is run due to random initialization of centroids	4. Always generates the same clusters