WELCOME TO

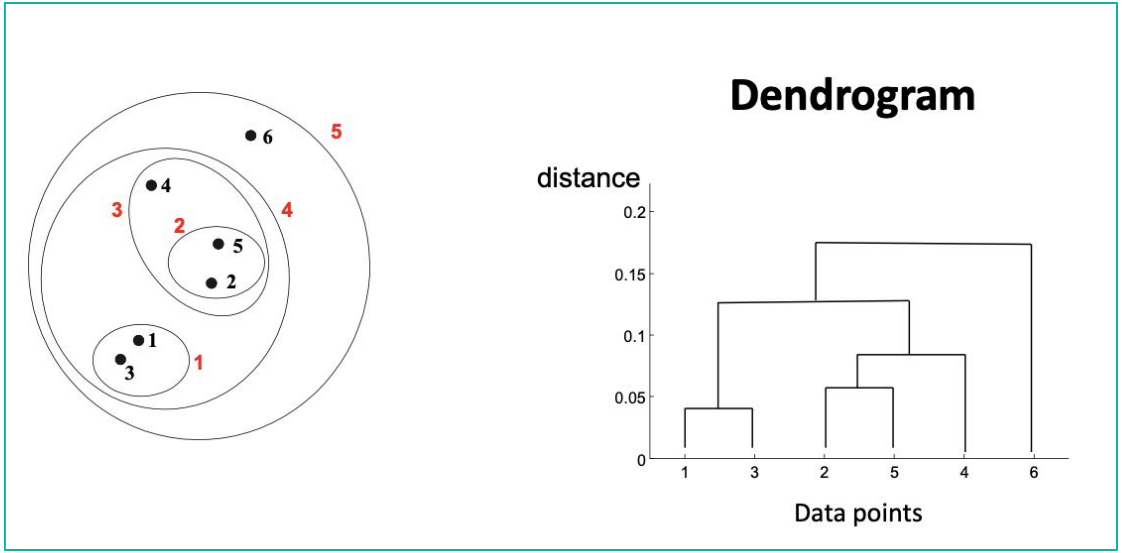
TECH I.S.

Hierarchical Clustering

Hierarchical Clustering

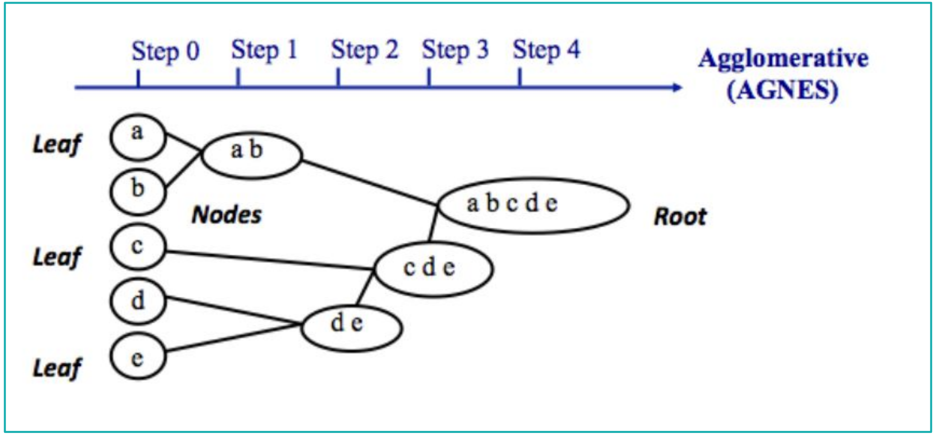
Produces a set of nested clusters organized as a hierarchical tree called a dendrogram.

The dendrogram shows at what distance points join into a cluster.



Agglomerative Clustering - Overview

Start with the points as individual clusters - At each step, merge the closest pair of clusters until only one cluster (or k clusters) remaining

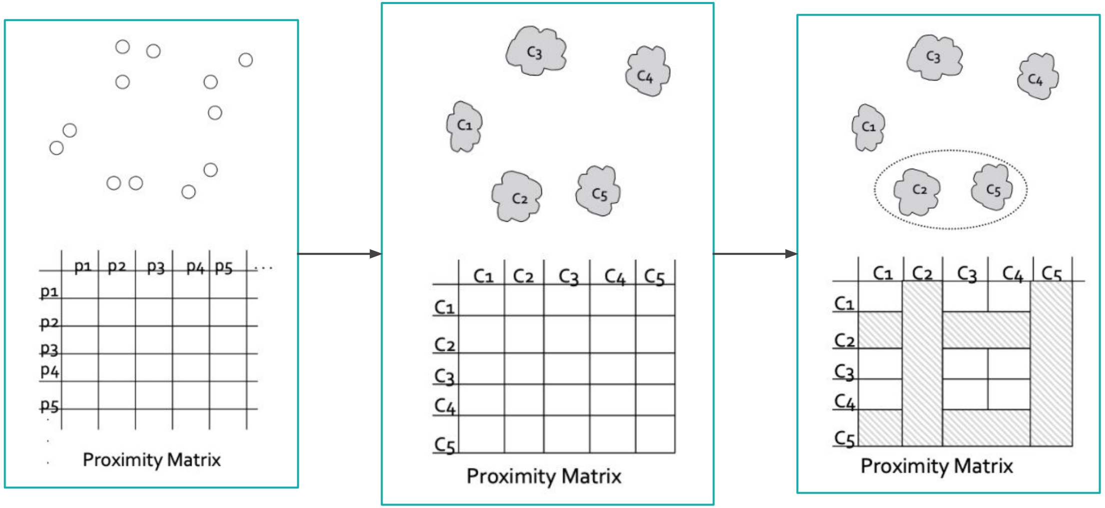


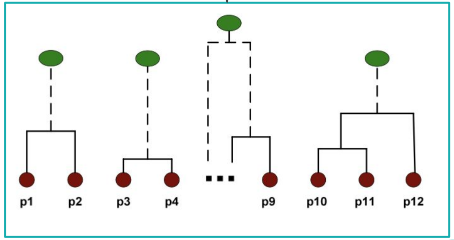
Key operation is the computation of the proximity of two clusters.

Agglomerative Clustering Algorithm

1. Start with clusters of individual points and a proximity matrix
2. After some merging steps, we have some clusters
3. We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.

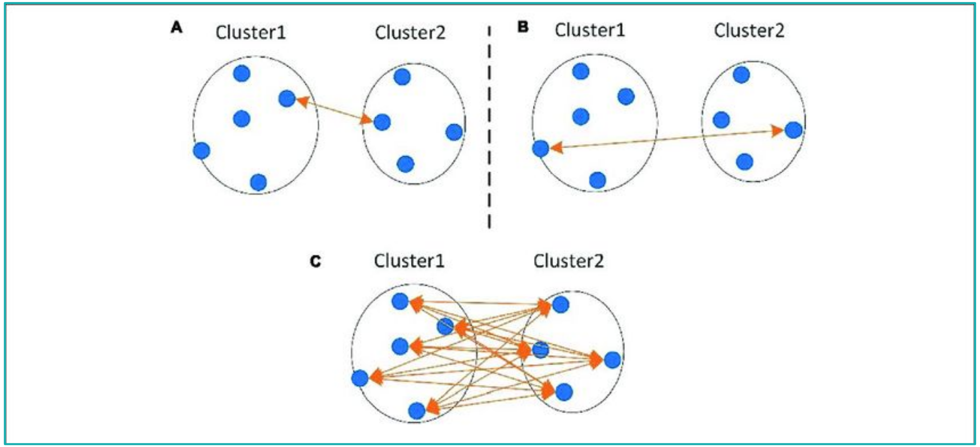
```
Compute the proximity matrix
Let each data point be a cluster
Repeat
        Merge the two closest clusters
        Update the proximity matrix
Until only a single cluster remains
```





Measuring the distance of two clusters

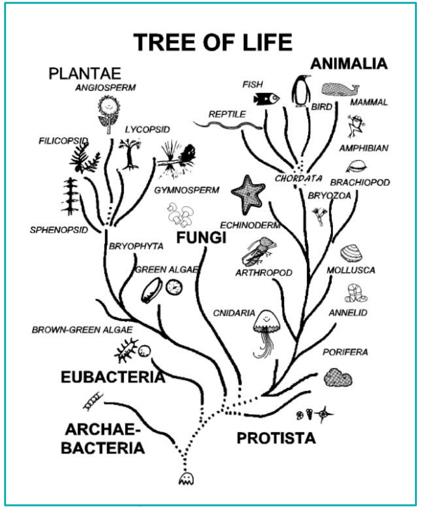
1. Single link method: Distance between two closest data points in the two clusters.
2. Complete link method: Distance of two furthest data points in the two clusters.
3. Average link: Average distance of all pairwise distances between the data points in two clusters.



Strengths of Hierarchical Clustering

We do not have to assume any particular number of clusters:

- Any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level.
    - They may correspond to meaningful taxonomies: Example in biological sciences.



Limitations of Hierarchical Clustering

- Greedy: Once a decision is made to combine two clusters, it cannot be undone
- No global objective function is directly minimized
- Sensitivity to noise and outliers
- Difficulty handling different sized clusters and convex shapes
- Chaining, breaking large clusters

Much obliged.

TECH I.S.