Hierarchical Clustering

Hierarchical Clustering

- ***Why use Hierarchical Clustering?***
  - Easy to understand and visualize.
  - Helps users decide how many clusters to choose.
  - Not necessary to choose cluster amount **before** running the algorithm.

Hierarchical Clustering

- ***Why use Hierarchical Clustering?***
  - Divides points into ***potential*** clusters:
    - Agglomerative Approach:
      - Each point begins as its own cluster, then clusters are joined.
    - Divisive Approach:
      - All points begin in the same cluster, then clusters are split.

Hierarchical Clustering

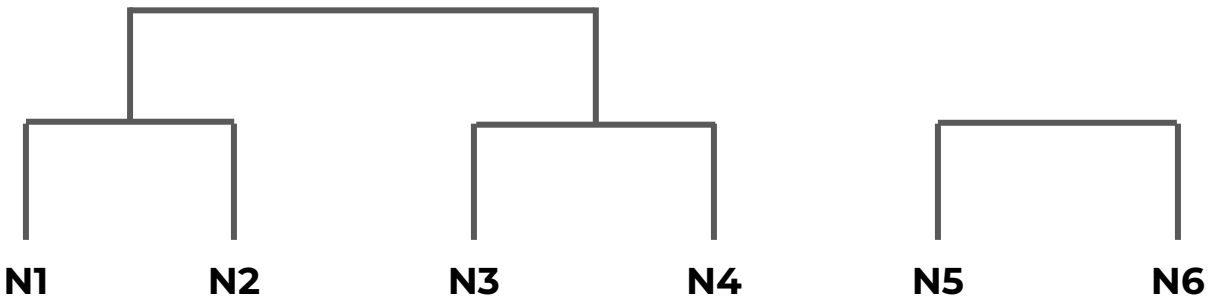
- Hierarchical Clustering
  - Agglomerative:

Hierarchical Clustering

- Hierarchical Clustering
  - Agglomerative:

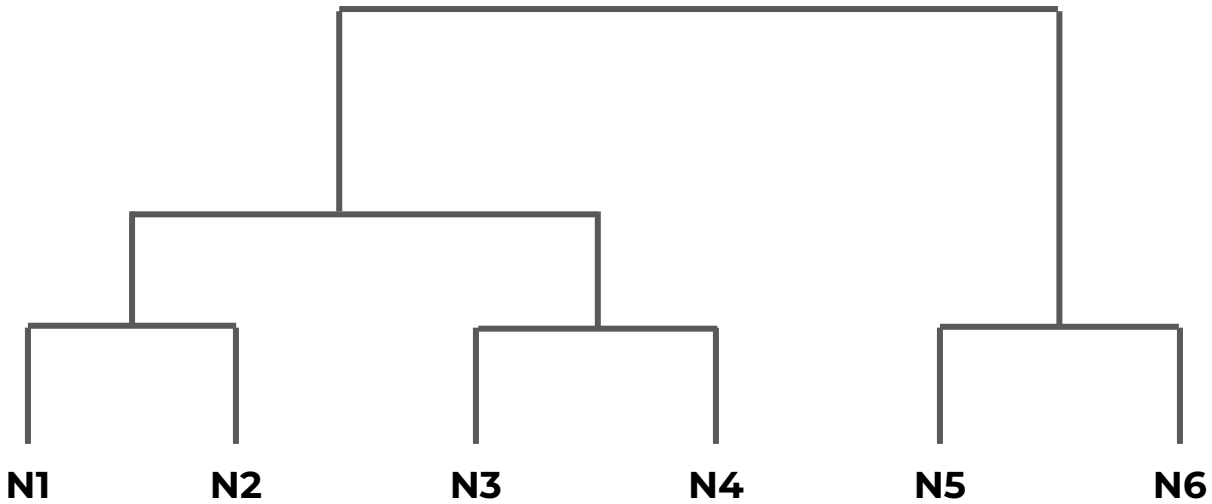
Hierarchical Clustering

- Hierarchical Clustering
  - Agglomerative:



Hierarchical Clustering

- Hierarchical Clustering
  - Agglomerative:

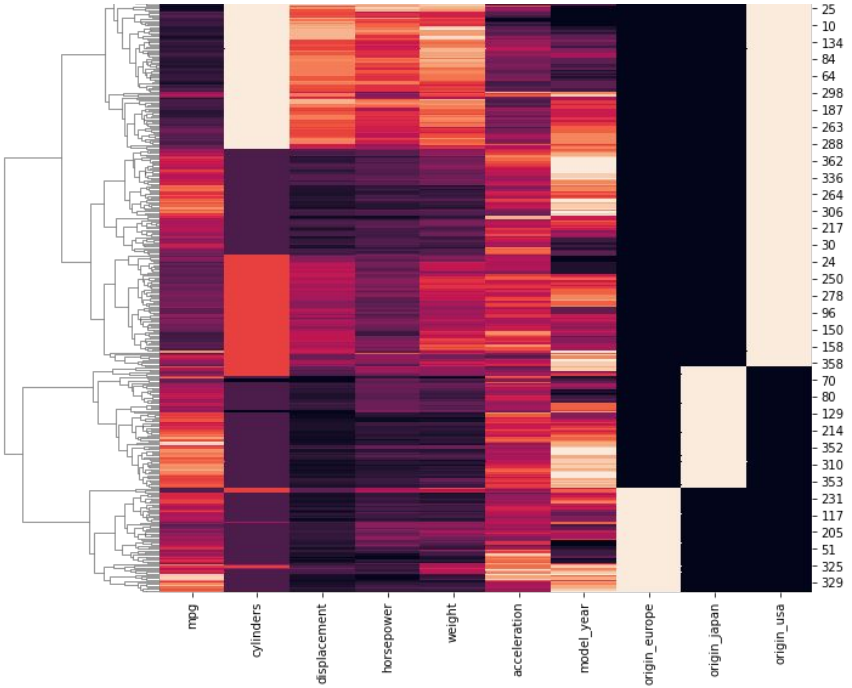


Hierarchical Clustering

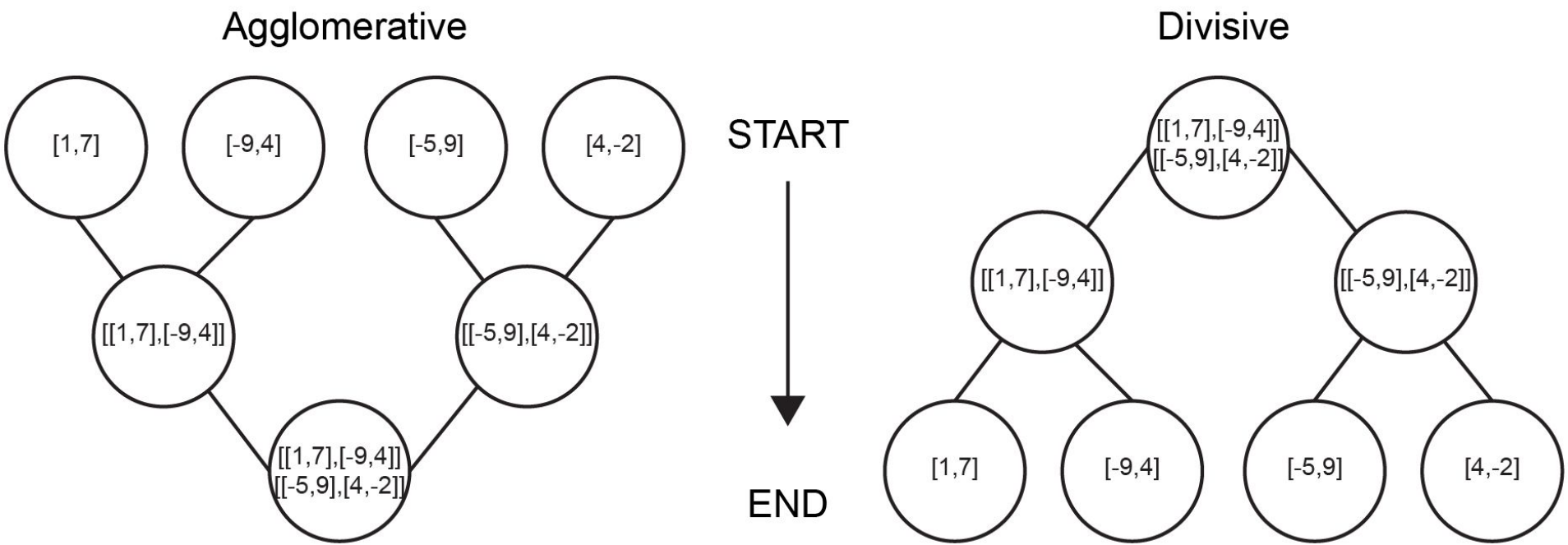
- ***Hierarchical Clustering Process***
    - Compare data points to find most similar data points to each other.
    - Merge these to create a cluster.
    - Compare clusters to find most similar clusters and merge again.
    - Repeat until all points in a single cluster.

Hierarchical Clustering

- ***Hierarchical Clustering Process***
- Dendrogram



Hierarchical Clustering

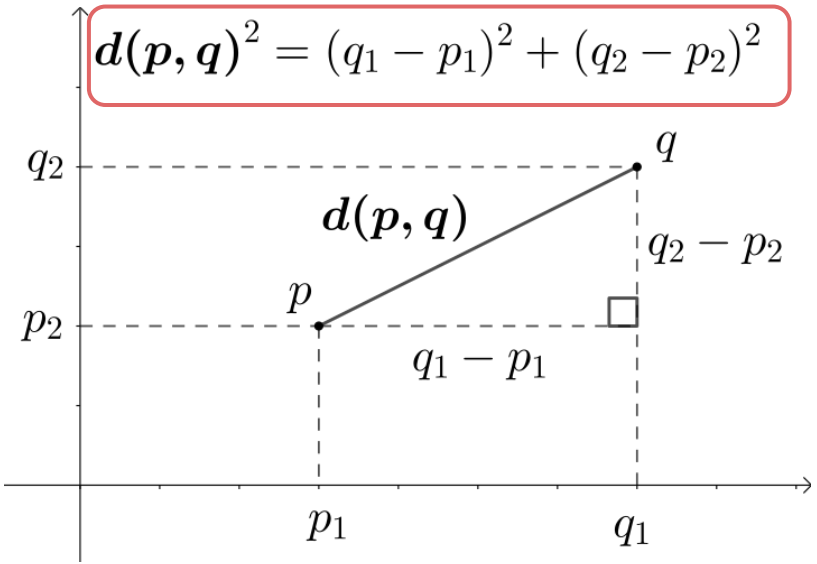


Hierarchical Clustering

- Similarity Metric
  - Measures distance between two points.
  - Many options:
    - Euclidean Distance
    - Manhattan
    - Cosine
    - and many more...

Hierarchical Clustering

- Similarity Metric
  - Default choice is Euclidean



Hierarchical Clustering

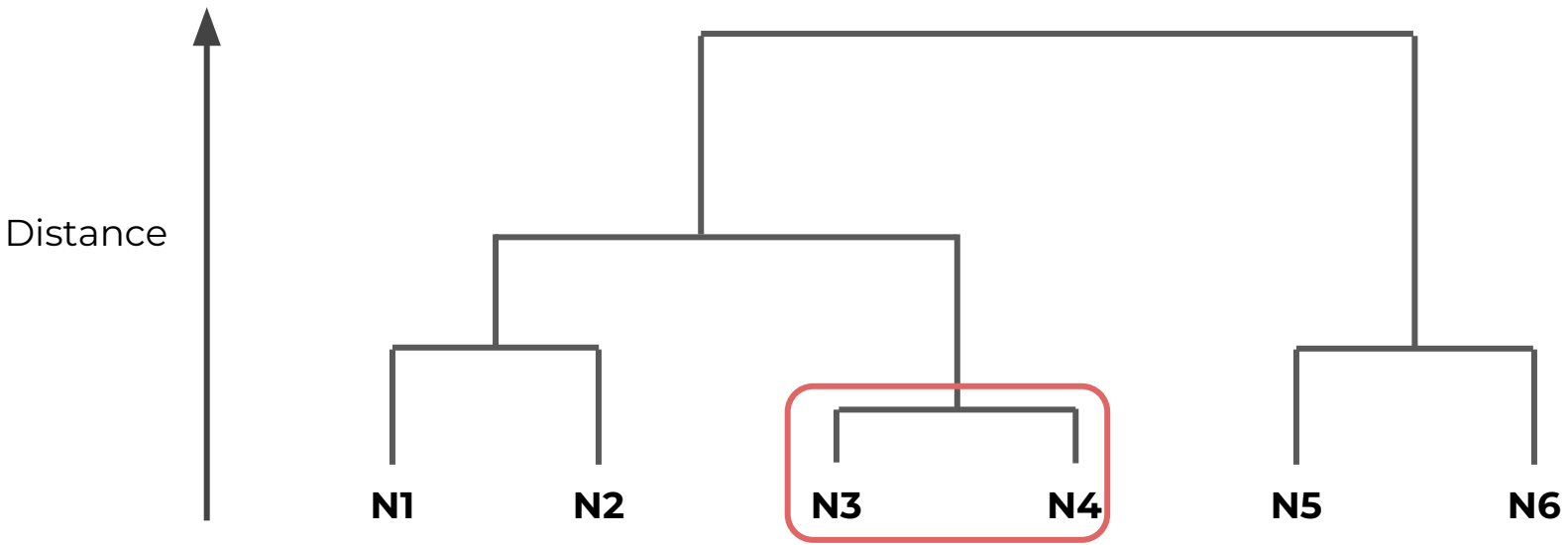
- Similarity Metric
  - Each dimension would be a feature
  - For **n** data points and **p** features:
    - $D^2 = (x_{11} - x_{12})^2 + \ldots + (x_{n-1p-1} - x_{np})^2$
  - Using MinMaxScaler we can scale all features to be between 0 and 1.
  - This allows for maximum distance between a feature to be 1.

Hierarchical Clustering

- Dendrogram:
  - Plot displaying all potential clusters.
  - Very computationally expensive to compute and display for larger data sets.
  - Very useful for deciding on number of clusters.

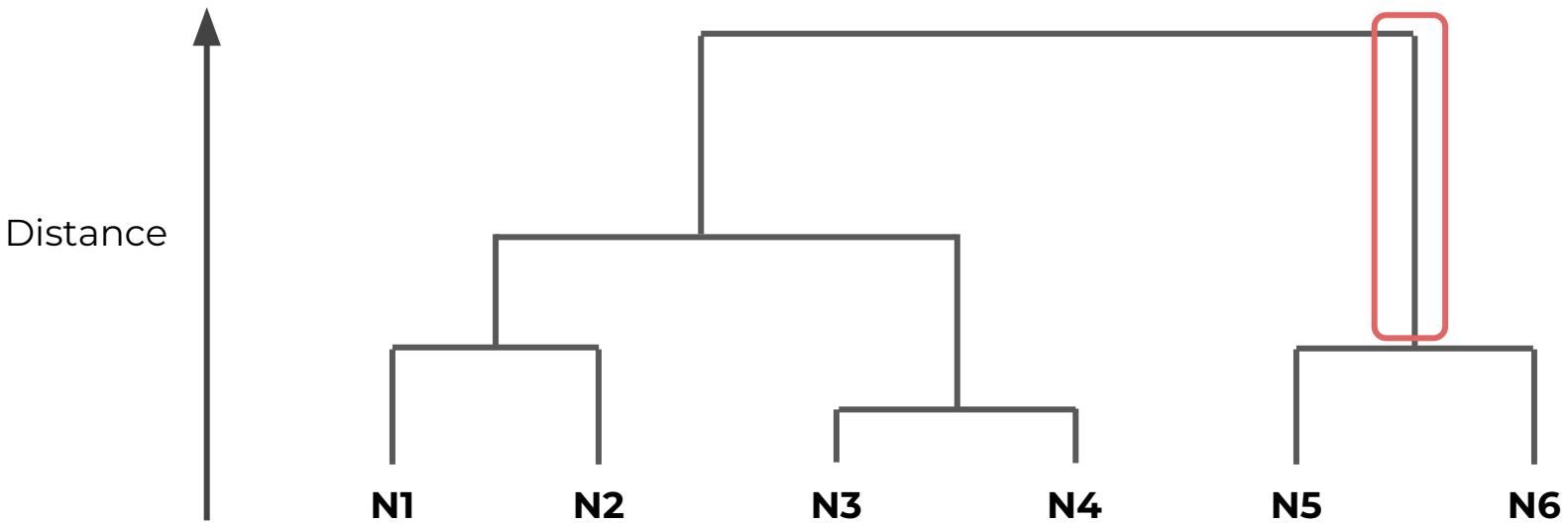
Hierarchical Clustering

- Dendrogram:



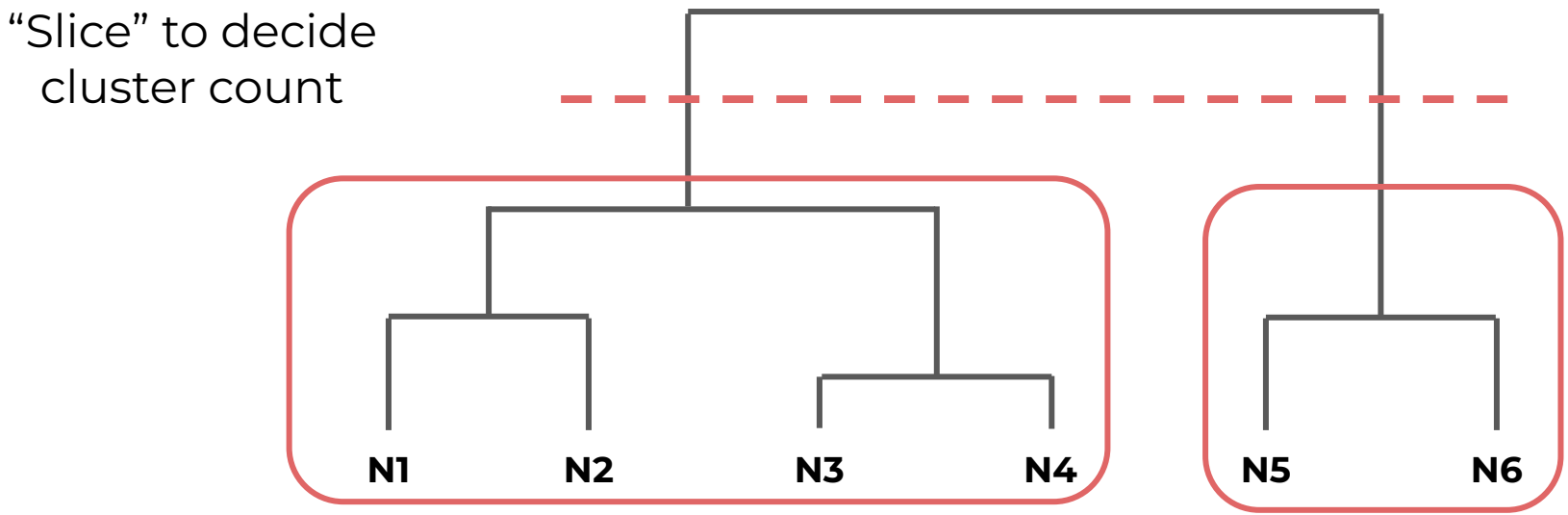
Hierarchical Clustering

- Dendrogram:



Hierarchical Clustering

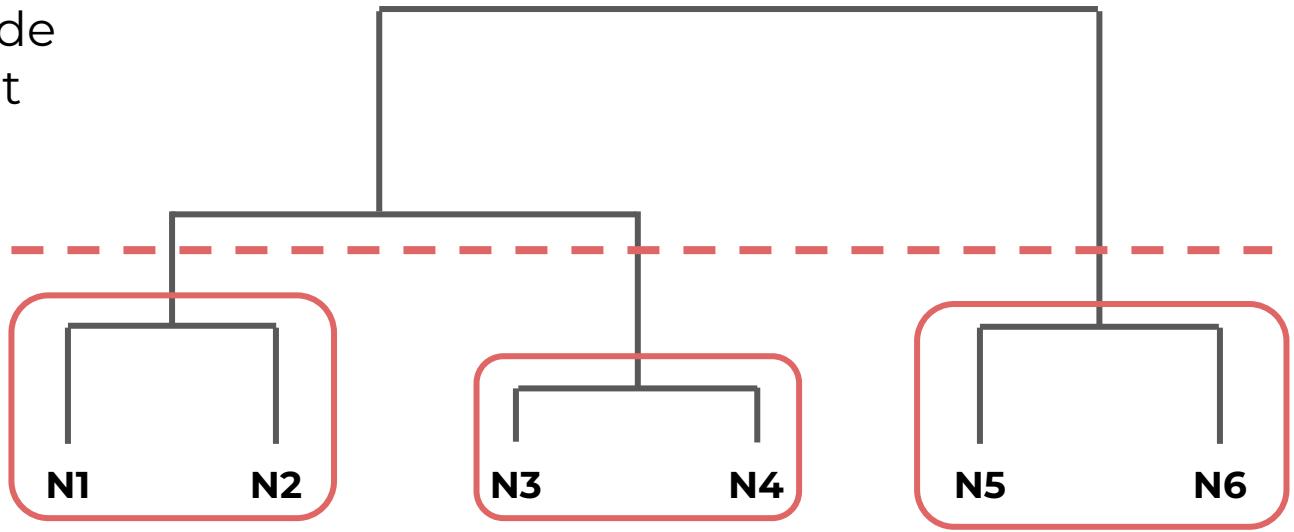
- Dendrogram:



Hierarchical Clustering

- Dendrogram:

"Slice" to decide cluster count

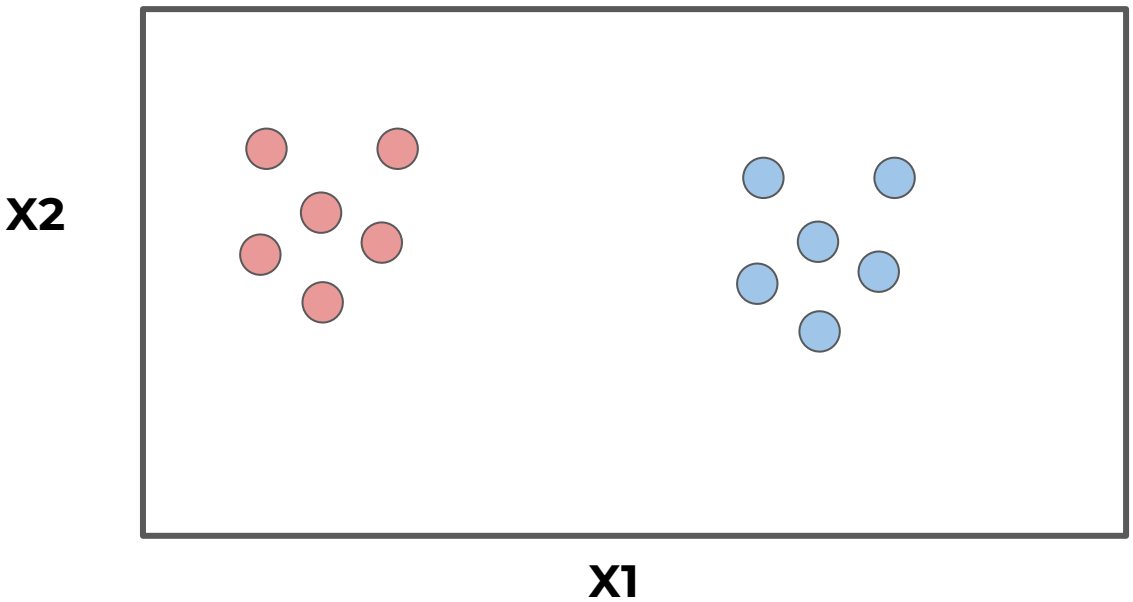


Hierarchical Clustering

- *Linkage* parameter
  - How do we measure distance from a point to an entire cluster?
  - How do we measure distance from a cluster to another cluster?

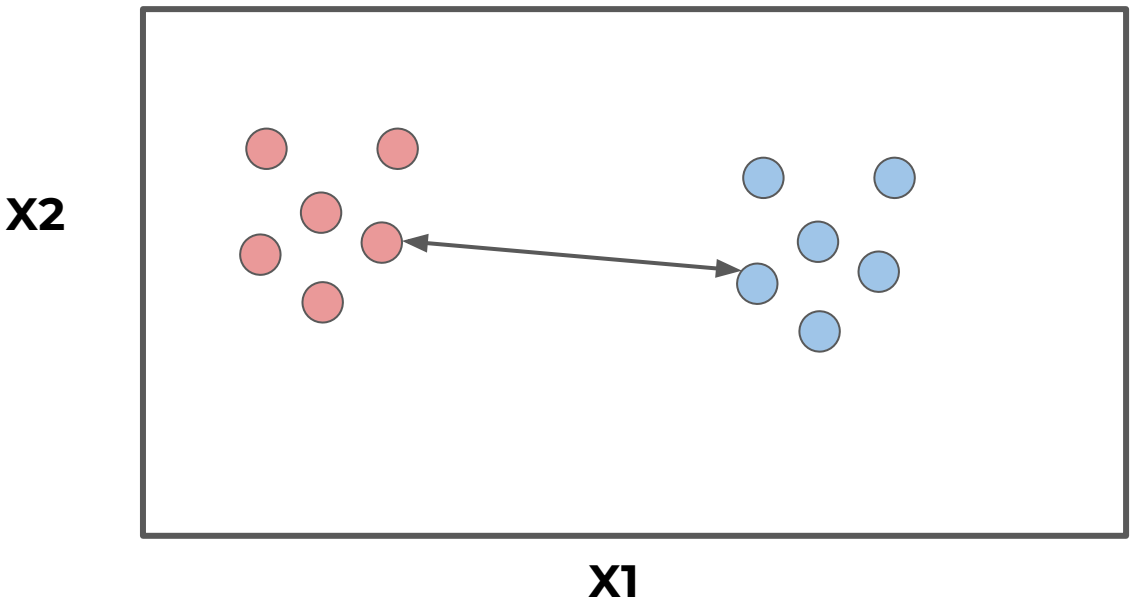
Hierarchical Clustering

- Linkage



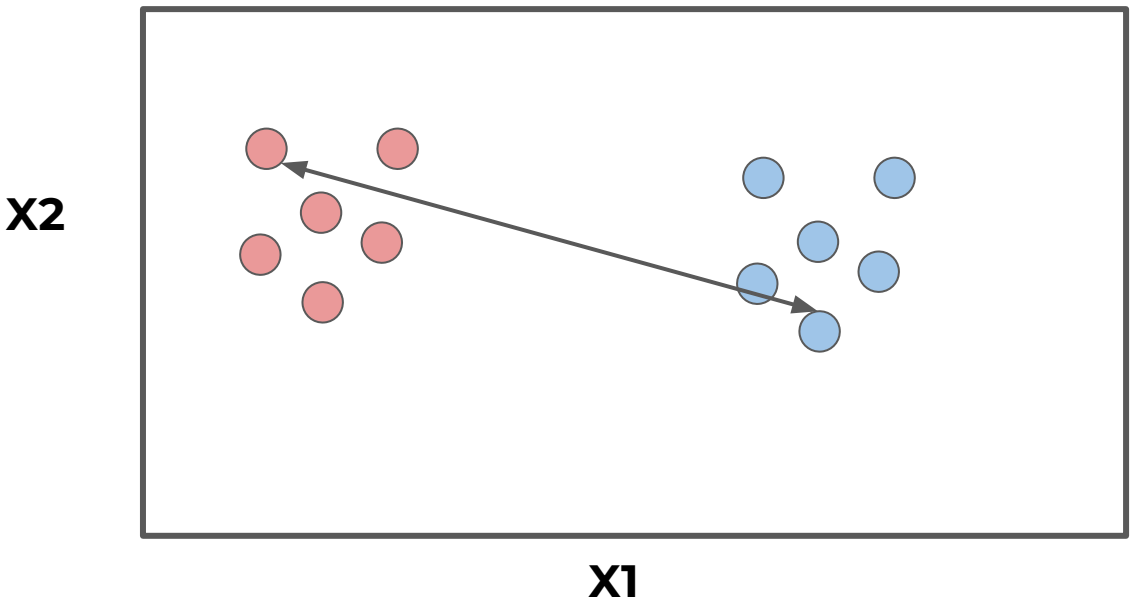
Hierarchical Clustering

- Linkage



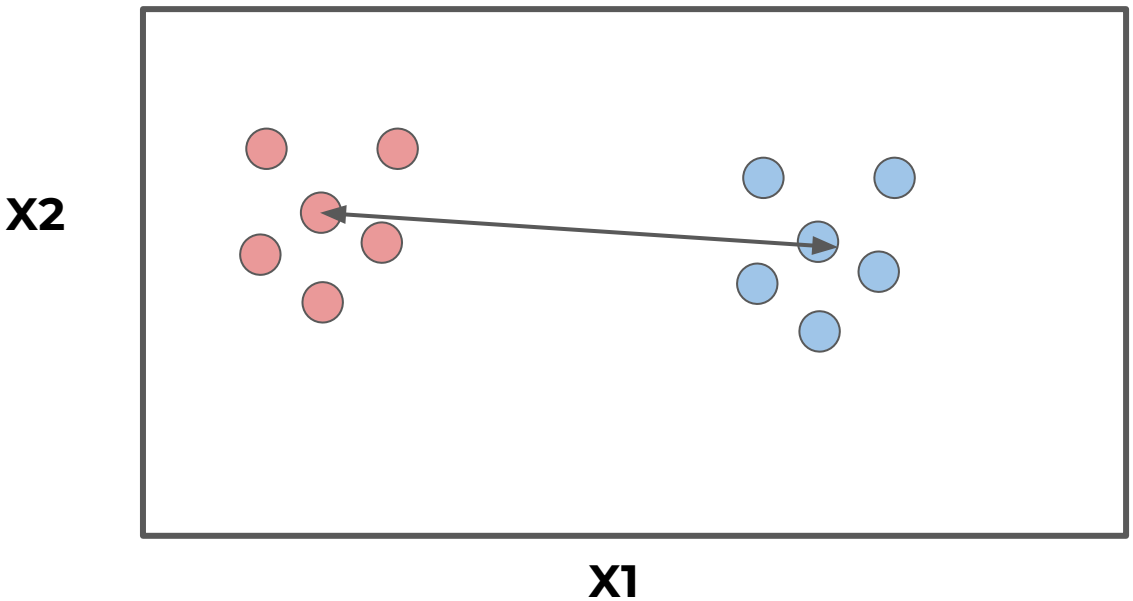
Hierarchical Clustering

- Linkage



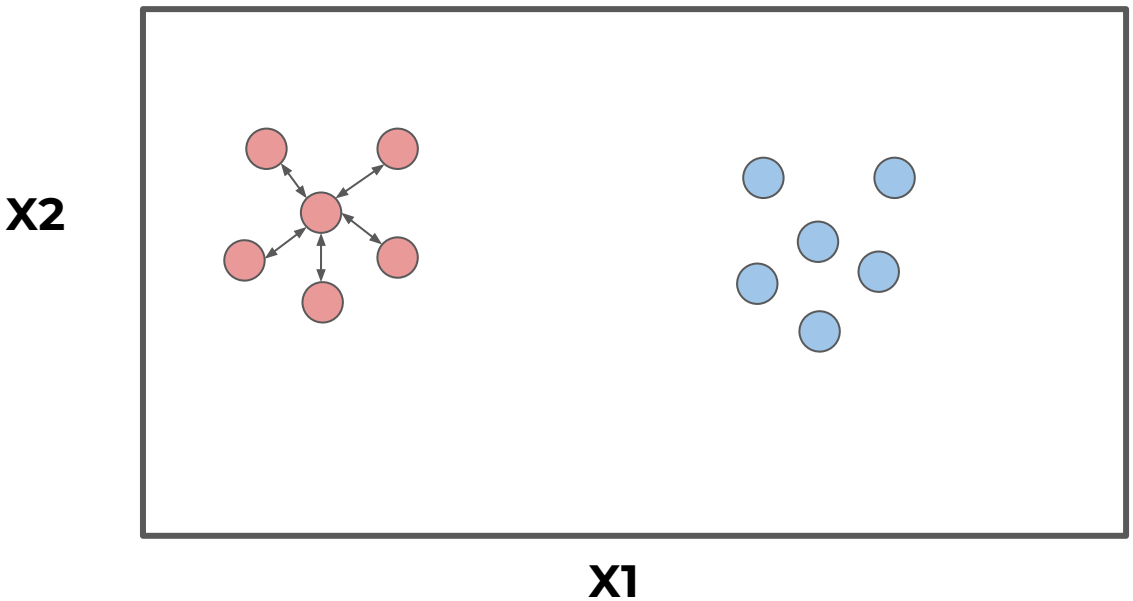
Hierarchical Clustering

- Linkage



Hierarchical Clustering

- Linkage



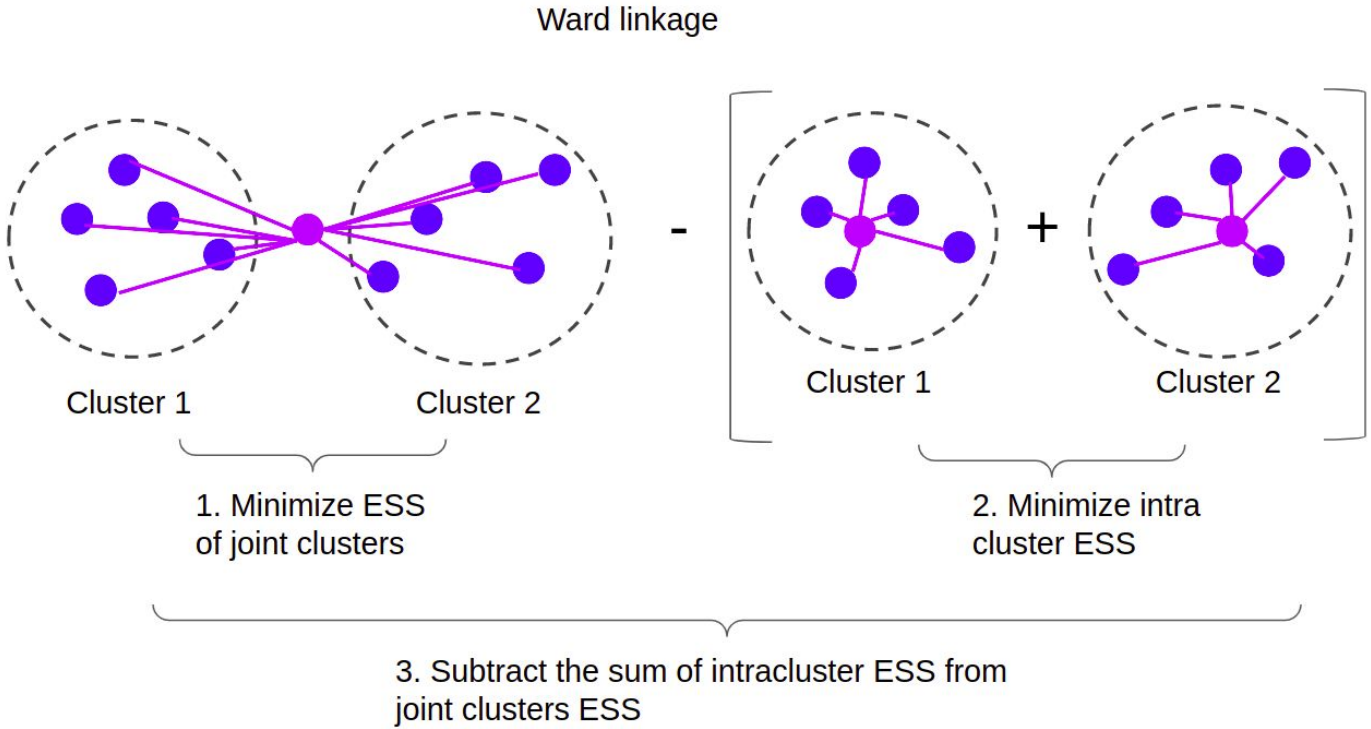
Hierarchical Clustering

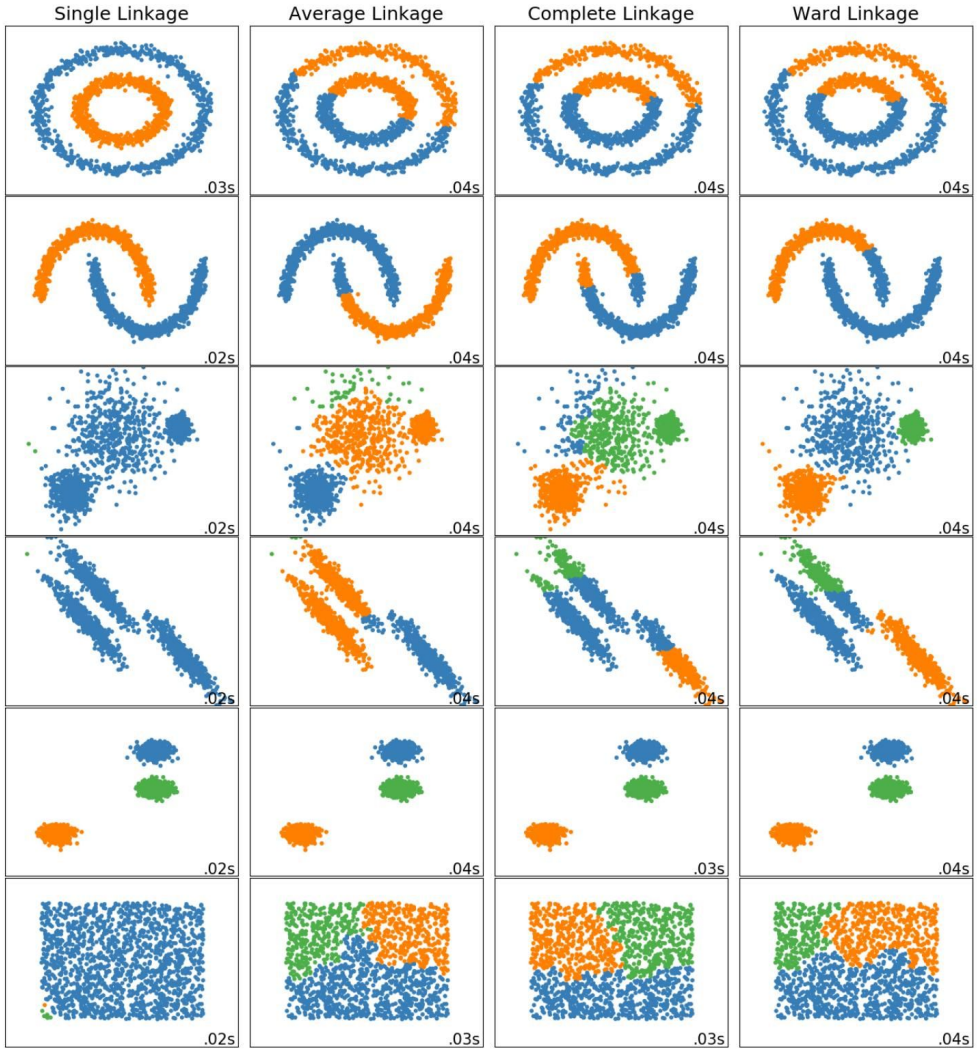
- Linkage
  - Criterion determining which distance to use between sets of observations.
  - Algorithm will merge pairs of clusters that minimizes the criterion.

Hierarchical Clustering

- Linkage:
  - **Ward:** minimizes variance of clusters being merged.
  - **Average:** uses average distances between two sets.
  - **Minimum** or **Maximum** distances between all observations of the two sets.

Hierarchical Clustering





PIERIAN DATA