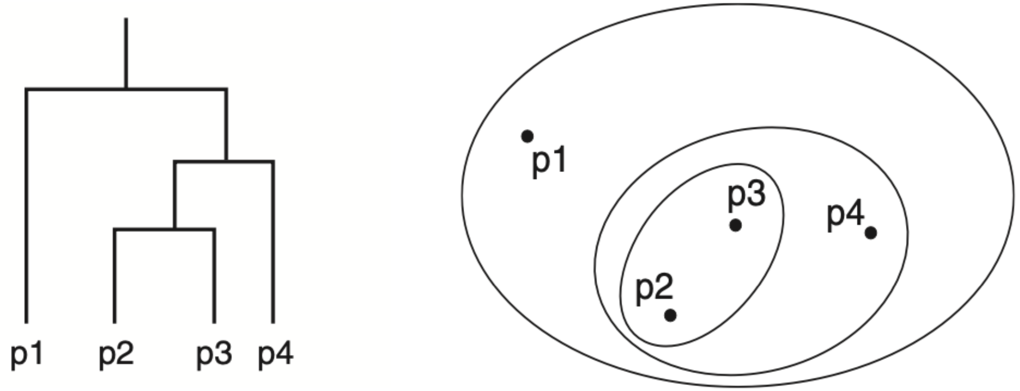


(a) Dendrogram. (b) Nested cluster diagram.

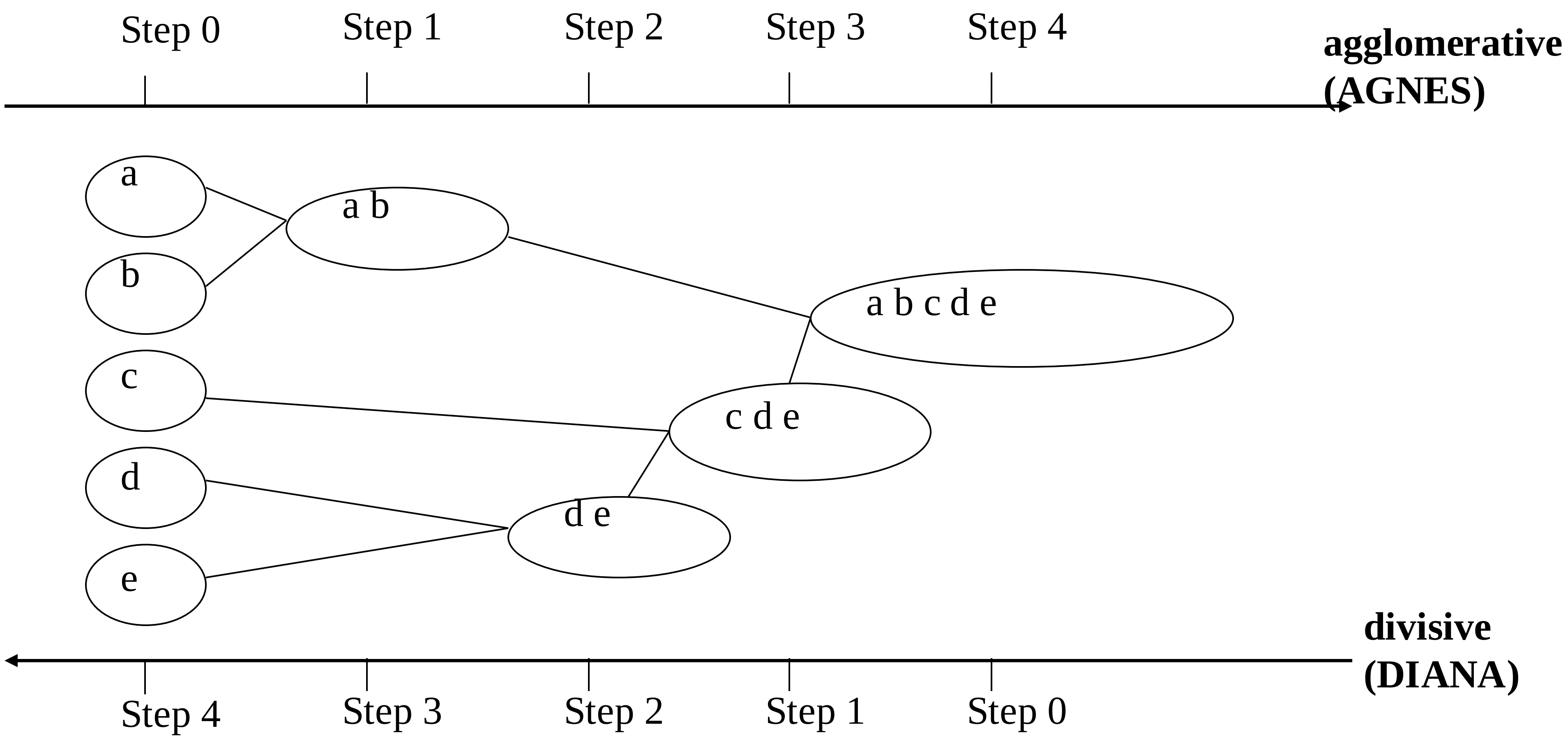
Hierarchical Clustering Methods

A hierarchical clustering method works by grouping data objects into a hierarchy or “tree” of clusters.

Hierarchical Clustering: Basic Concepts

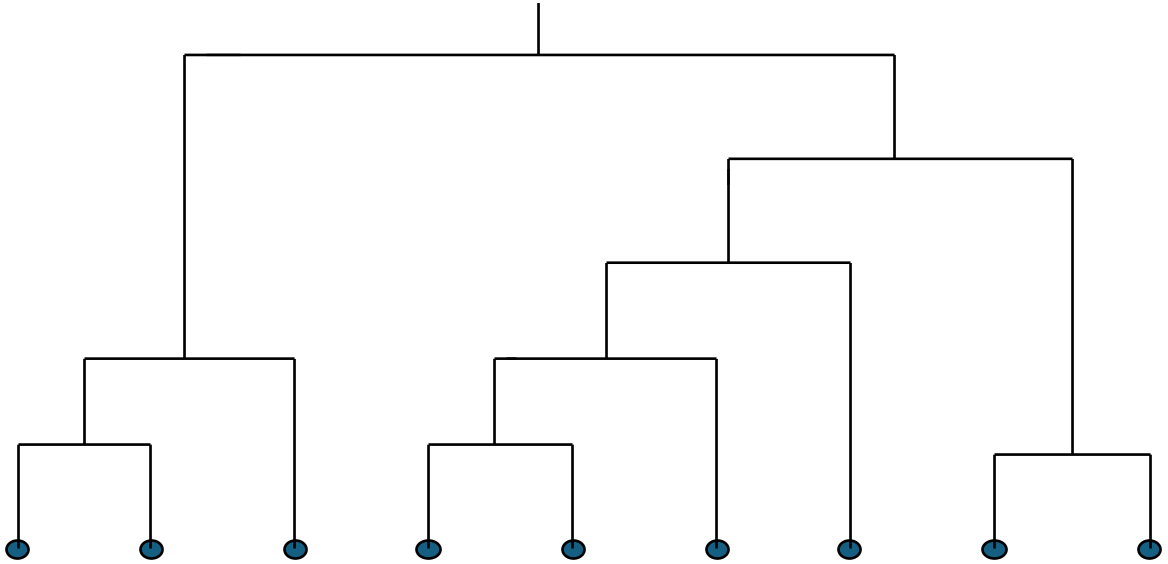
- Hierarchical clustering
  - Generate a clustering hierarchy (drawn as a dendrogram)
  - Not required to specify K, the number of clusters
  - More deterministic
  - No iterative refinement
- Two categories of algorithms
  - Agglomerative: Start with singleton clusters, continuously merge two clusters at a time to build a bottom-up hierarchy of clusters
  - Divisive: Start with a huge macro-cluster, split it continuously into two groups, generating a top-down hierarchy of clusters
  - Agglomerative far more common.

Agglomerative vs. Divisive Clustering



Dendrogram: How Clusters are Merged

- Dendrogram: Decompose a set of data objects into a tree of clusters by multi-level nested partitioning
- A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster

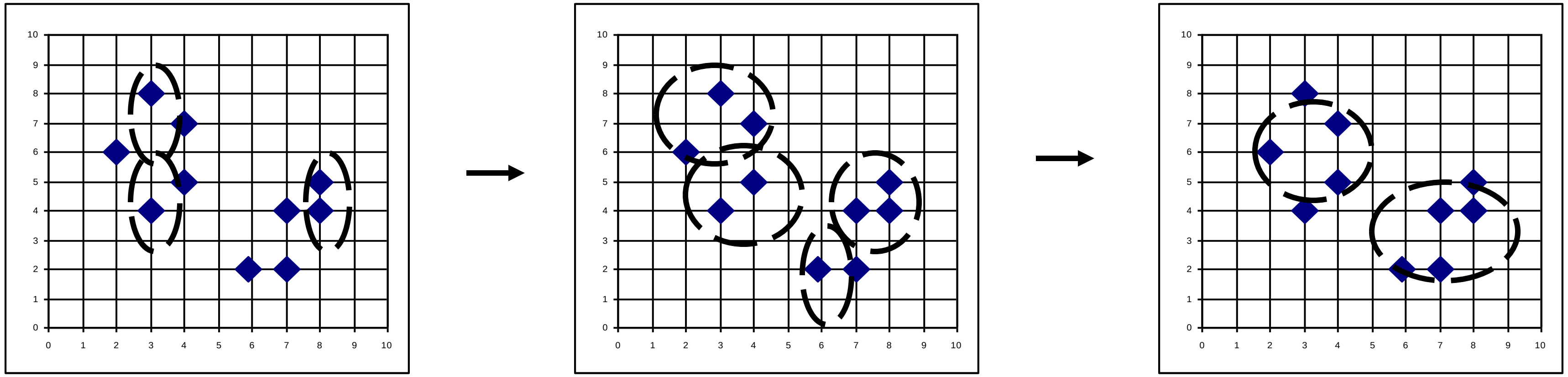


Hierarchical clustering generates a dendrogram (a hierarchy of clusters)

Agglomerative Clustering Algorithm

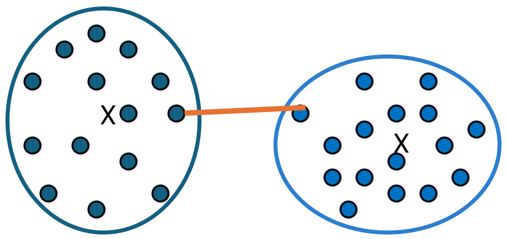
- AGNES (AGglomerative NESting) (Kaufmann and Rousseeuw, 1990)
  - Use the single-link method and the dissimilarity matrix
  - Continuously merge nodes that have the least dissimilarity
  - Eventually all nodes belong to the same cluster
- Agglomerative clustering varies on different similarity measures among clusters
  - Single link (nearest neighbor)
  - Complete link (diameter)
  - Average link (group average)
  - Centroid link (centroid similarity)

Agglomerative Clustering Algorithm



Single Link vs. Complete Link in Hierarchical Clustering

- Single link (nearest neighbor)
  - The similarity between two clusters is the similarity between their most similar (nearest neighbor) members
  - Local similarity-based: Emphasizing more on close regions, ignoring the overall structure of the cluster
  - Capable of clustering non-elliptical shaped group of objects
  - Sensitive to noise and outliers

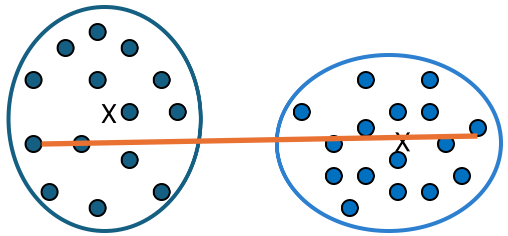


Minimum distance: $$dist_{min}(C_i, C_j) = \min_{\boldsymbol{p} \in C_i, \boldsymbol{p'} \in C_j} \{\|\boldsymbol{p} - \boldsymbol{p'}\|\}$$

|p− p'| is the distance between two objects or points, p and p';
m_i is the mean for cluster C_i; and
n_i is the number of objects in C_i.

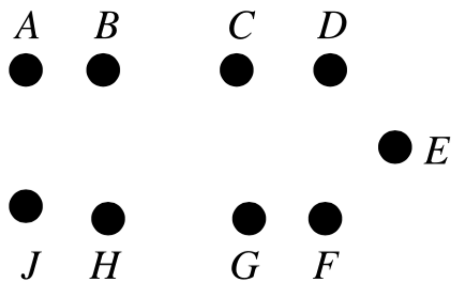
Single Link vs. Complete Link in Hierarchical Clustering

- Complete link (diameter)
  - The similarity between two clusters is the similarity between their most dissimilar members
  - Merge two clusters to form one with the smallest diameter
  - Nonlocal in behavior, obtaining compact shaped clusters
  - Sensitive to outliers

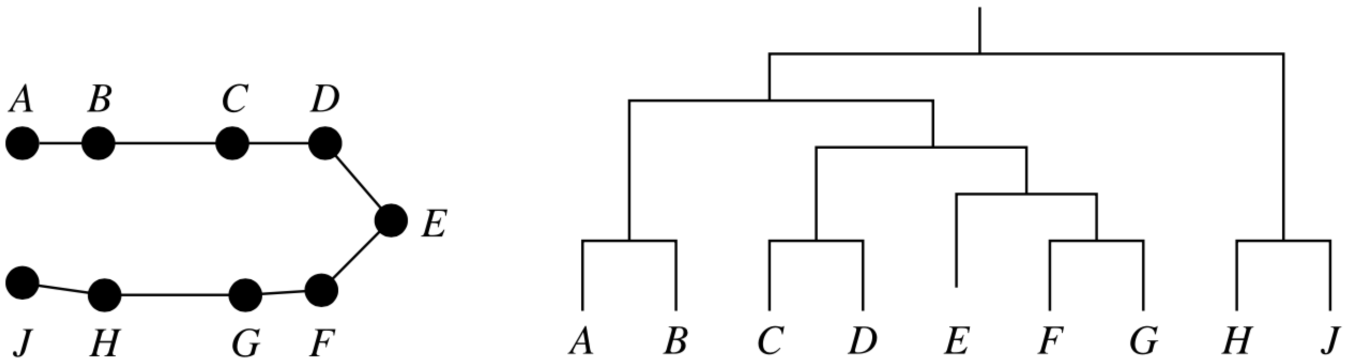


Maximum distance: $$dist_{max}(C_i, C_j) = \max_{\boldsymbol{p} \in C_i, \boldsymbol{p'} \in C_j} \{\|\boldsymbol{p} - \boldsymbol{p'}\|\}$$

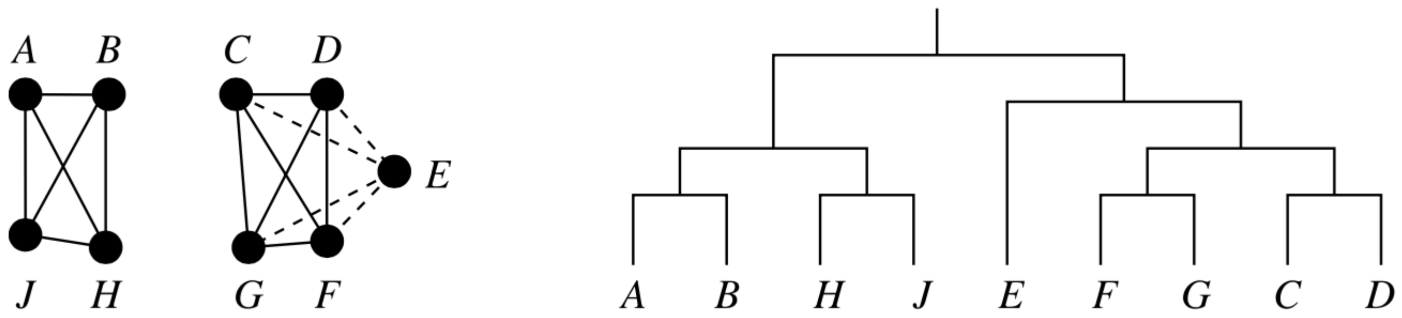
|p− p'| is the distance between two objects or points, p and p';
m_i is the mean for cluster C_i; and
n_i is the number of objects in C_i.



(a) Data set

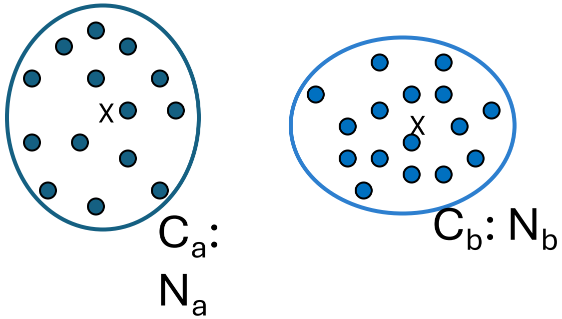


(b) Clustering using single linkage



(c) Clustering using complete linkage

Agglomerative Clustering: Average vs. Centroid Links



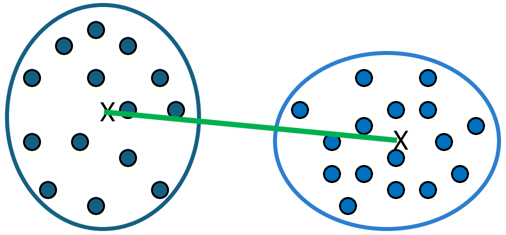
- Agglomerative clustering with average link
  - Average link: The average distance between an element in one cluster and an element in the other (i.e., all pairs in two clusters)
    - Expensive to compute

Average distance: $$dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{\boldsymbol{p} \in C_i, \boldsymbol{p'} \in C_j} \|\boldsymbol{p} - \boldsymbol{p'}\|$$

|p− p'| is the distance between two objects or points, p and p';
m_i is the mean for cluster C_i; and
n_i is the number of objects in C_i.

Agglomerative Clustering: Average vs. Centroid Links

- Agglomerative clustering with centroid link
  - Centroid link: The distance between the centroids of two clusters



Mean distance: $$dist_{mean}(C_i, C_j) = \|\boldsymbol{m}_i - \boldsymbol{m}_j\|$$

|p− p'| is the distance between two objects or points, p and p';
m_i is the mean for cluster C_i; and
n_i is the number of objects in C_i.

Agglomerative Clustering with Ward's Criterion

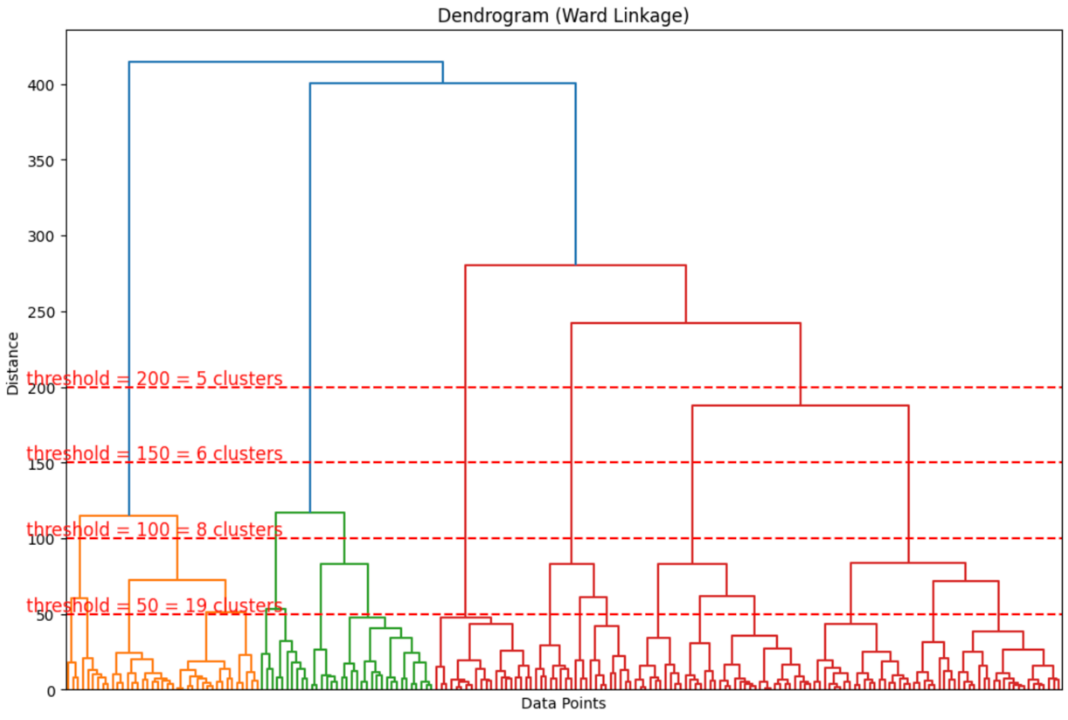
- Suppose two disjoint clusters C_i and C_j are merged, and m_{ij} is the mean of the new cluster
- Ward's criterion: $W(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \left| m_i - m_j \right|^2$
- Minimize the increase in total within-cluster **variance** when merging two clusters. This results in clusters that are more compact and homogeneous.
- At each step, Ward's method merges the two clusters that result in the **smallest increase in the total within-cluster variance** after merging.

Agglomerative Clustering Steps

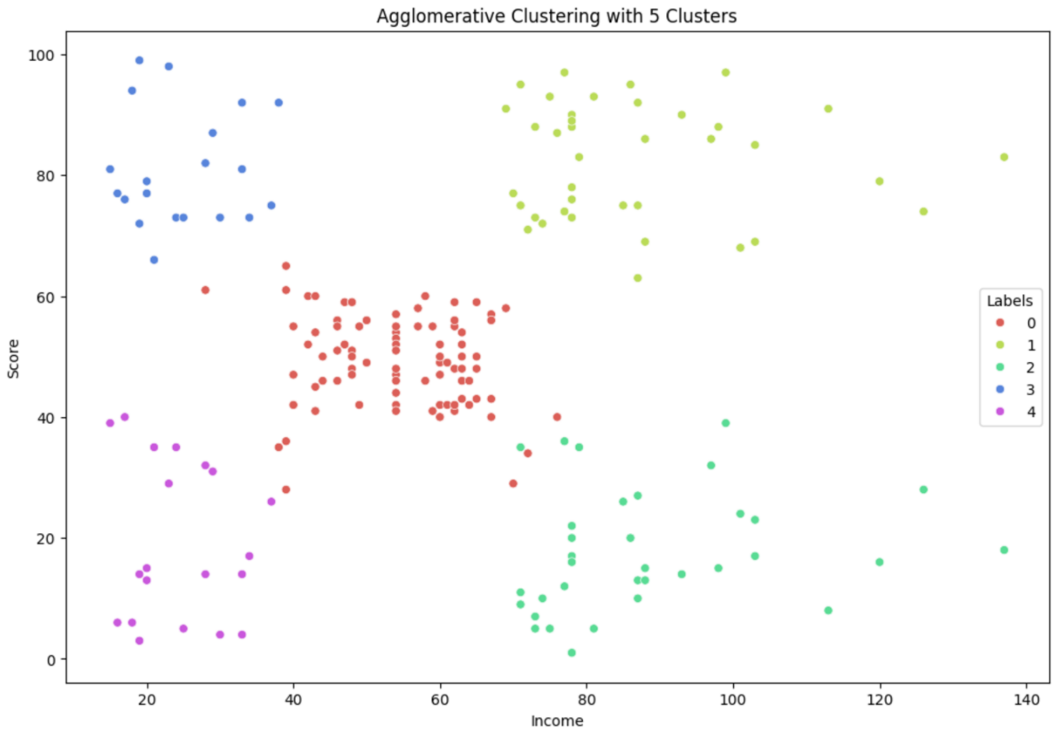
- Step 1: Create a dendrogram
  - Use a suitable distance metric and linkage method here
- Step 2: Identify Largest Gaps
  - Look for the longest vertical lines that are not interrupted by merges.
  - These large gaps suggest significant dissimilarity between clusters, making them a good place to “cut” the dendrogram.
- Step 3: Generate clusters based on the identified threshold.

Agglomerative Clustering Steps

Step 1 and 2: Create a dendrogram and Identify a threshold to use

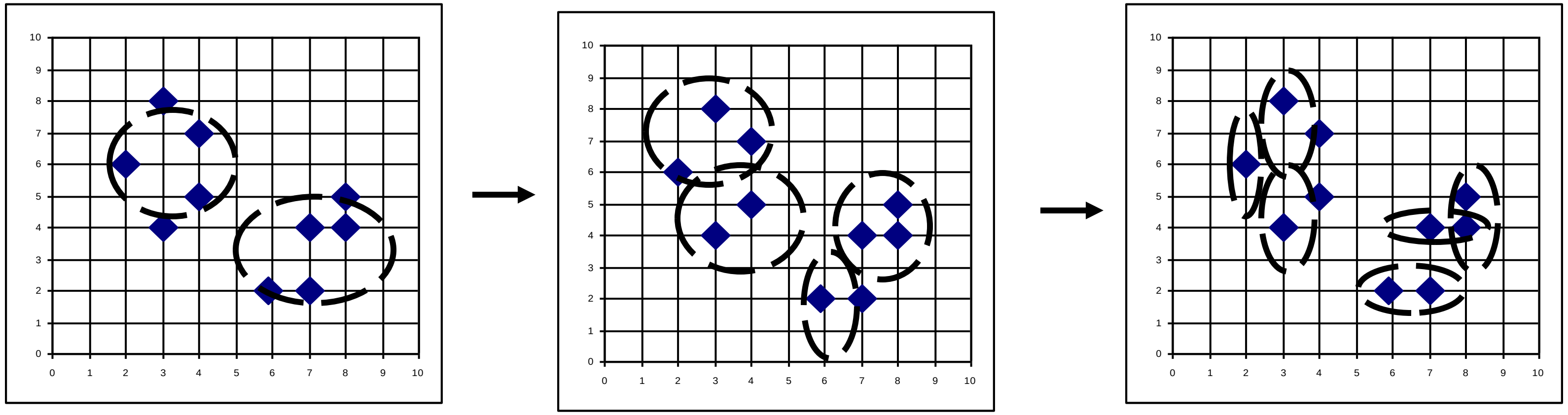


Step 3: Create clusters based on identified threshold



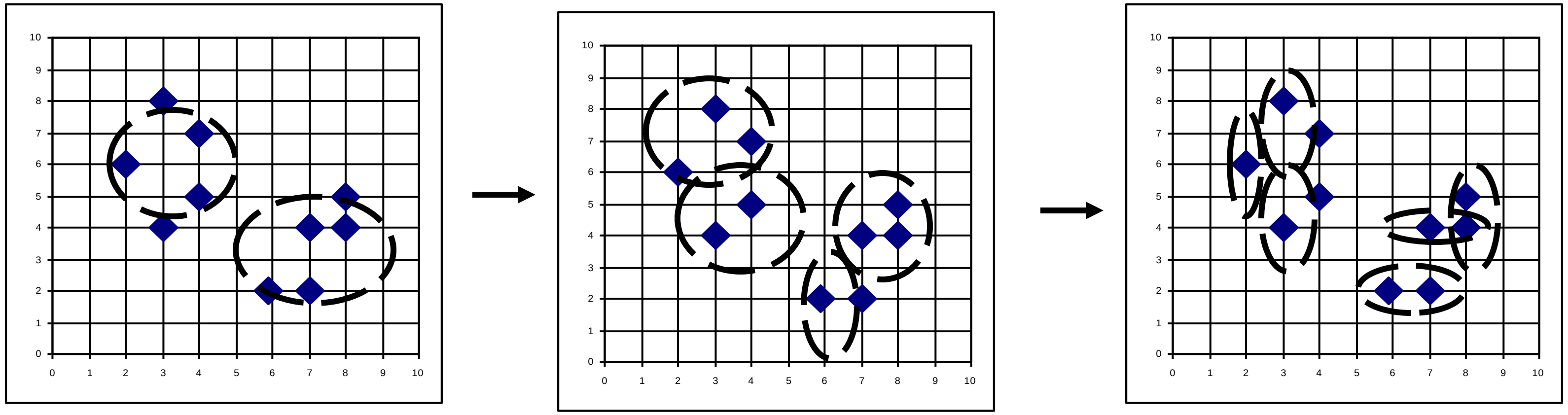
Divisive Clustering

- DIANA (Divisive Analysis) (Kaufmann and Rousseeuw,1990)
  - Implemented in some statistical analysis packages, e.g., Splus
- Inverse order of AGNES: Eventually each node forms a cluster on its own



Divisive Clustering Is a Top-down Approach

- The process starts at the root with all the points as one cluster
- It recursively splits the higher level clusters to build the dendrogram
- Can be considered as a global approach
- More efficient when compared with agglomerative clustering



More on Algorithm Design for Divisive Clustering

- Choosing which cluster to split
  - Check the sums of squared errors of the clusters and choose the one with the largest value
- Splitting criterion: Determining how to split
  - One may use Ward’s criterion to chase for greater reduction in the difference in the SSE criterion as a result of a split
  - For categorical data, Gini-index can be used
- Handling the noise
  - Use a threshold to determine the termination criterion (do not generate clusters that are too small because they contain mainly noises)