ECEN 758 Data Mining and Analysis: Lecture 10, Hierarchical Clustering

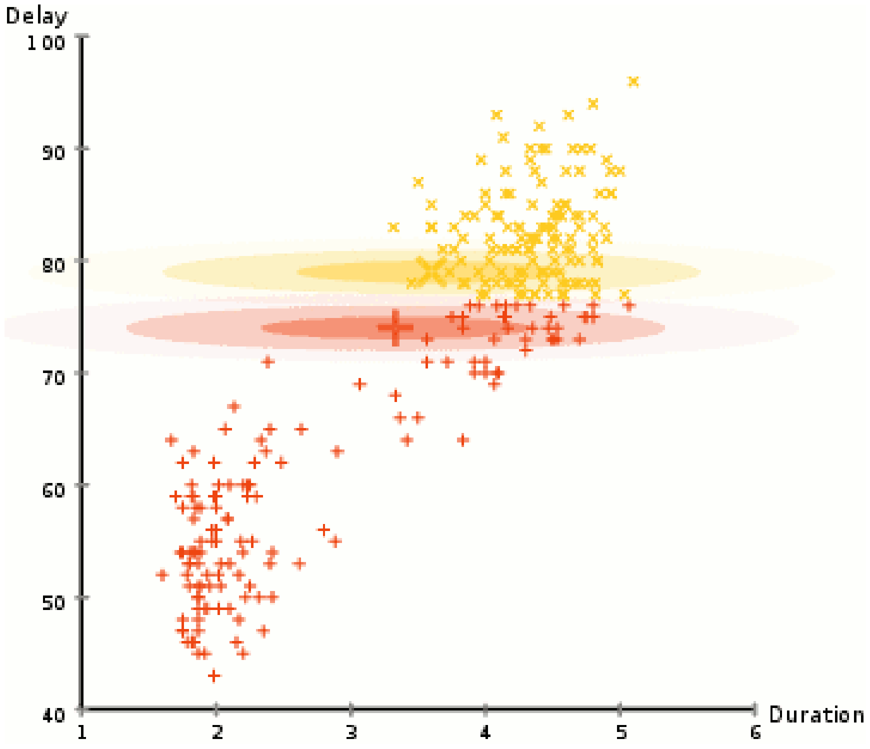
Joshua Peeples, Ph.D.
Assistant Professor
Department of Electrical and Computer Engineering

Announcements

- Assignment #2 due this Friday (09/27)
  - Please upload submission as single PDF
  - Please share Python code (Jupyter Notebooks, Google Colab, etc.)
    - Not submitting your code will result in losing points!

Last Lecture

- Expectation-Maximization Algorithm



Gif from: Expectation-maximization algorithm, Wikipedia

Today

- Hierarchical Clustering
- Reading: ZM Chapter 14
- Supplemental Reading:
  - MMDS Chapter 7
  - Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. Artificial Intelligence Review, 56(8), 8219-8264.

Clustering Overview

- We will discuss several variants of clustering
  - Representative-based Clustering
  - **Hierarchical Clustering**
  - Density-based Clustering

Hierarchical Clustering Overview

Hierarchical Clustering

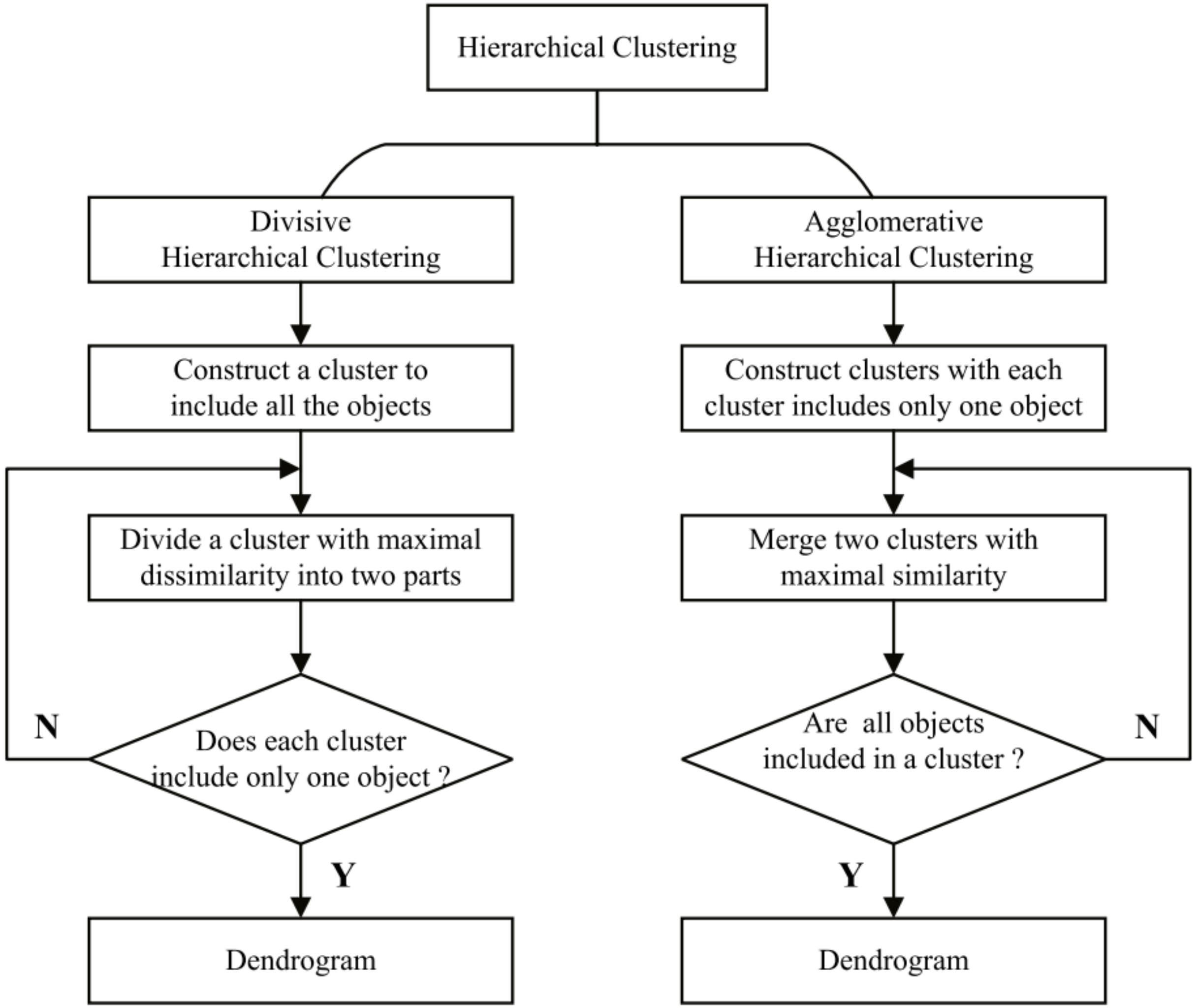


Figure from: X. Ran, Comprehensive survey on hierarchical clustering algorithms and the recent developments

Hierarchical Clustering

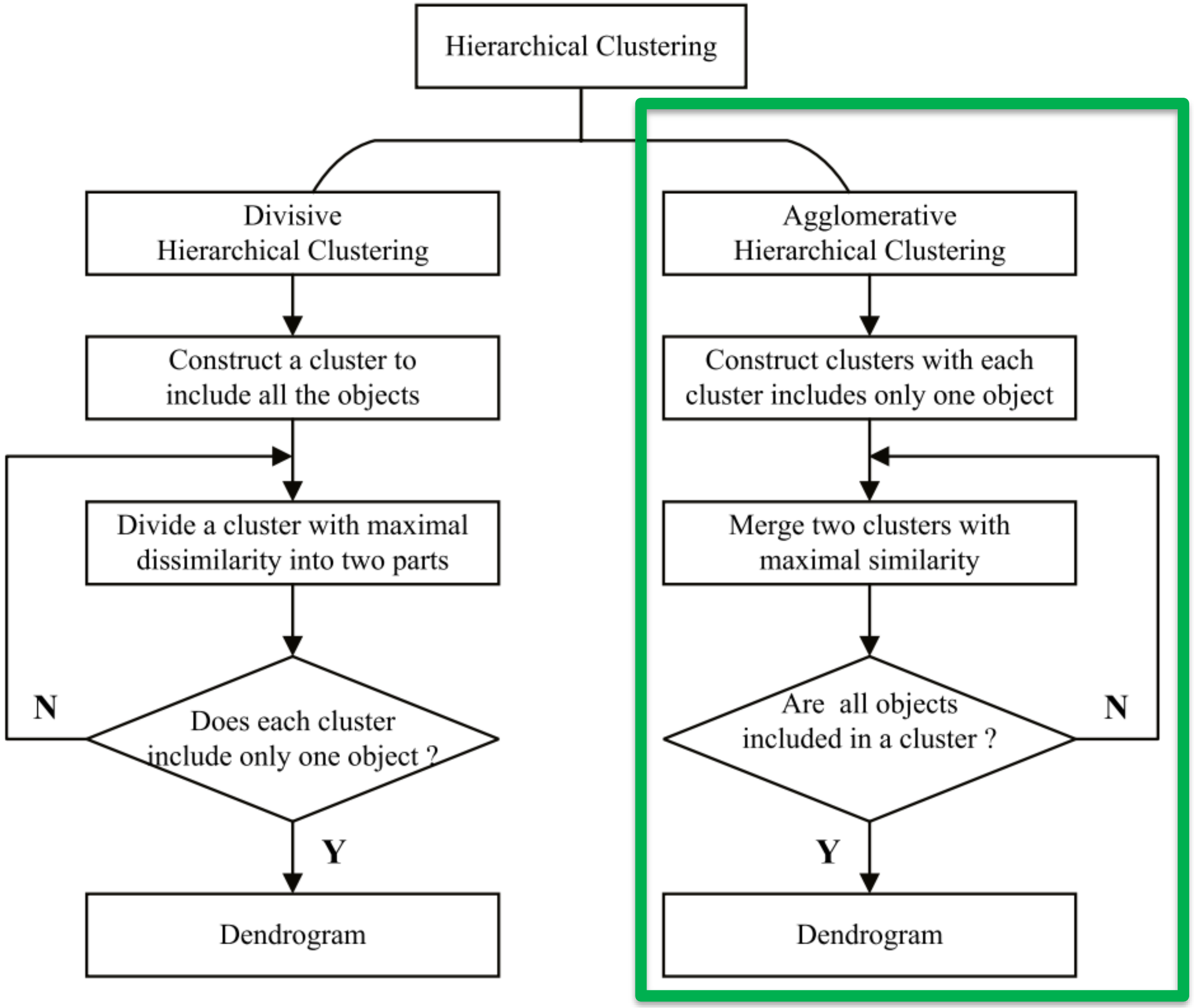


Figure from: X. Ran, Comprehensive survey on hierarchical clustering algorithms and the recent developments

Agglomerative Clustering Overview

Agglomerative Clustering

- “Bottom-up” clustering approach
- Creates sequences of nested partitions that can be viewed with cluster dendrogram
- Each point starts as own cluster and are merged

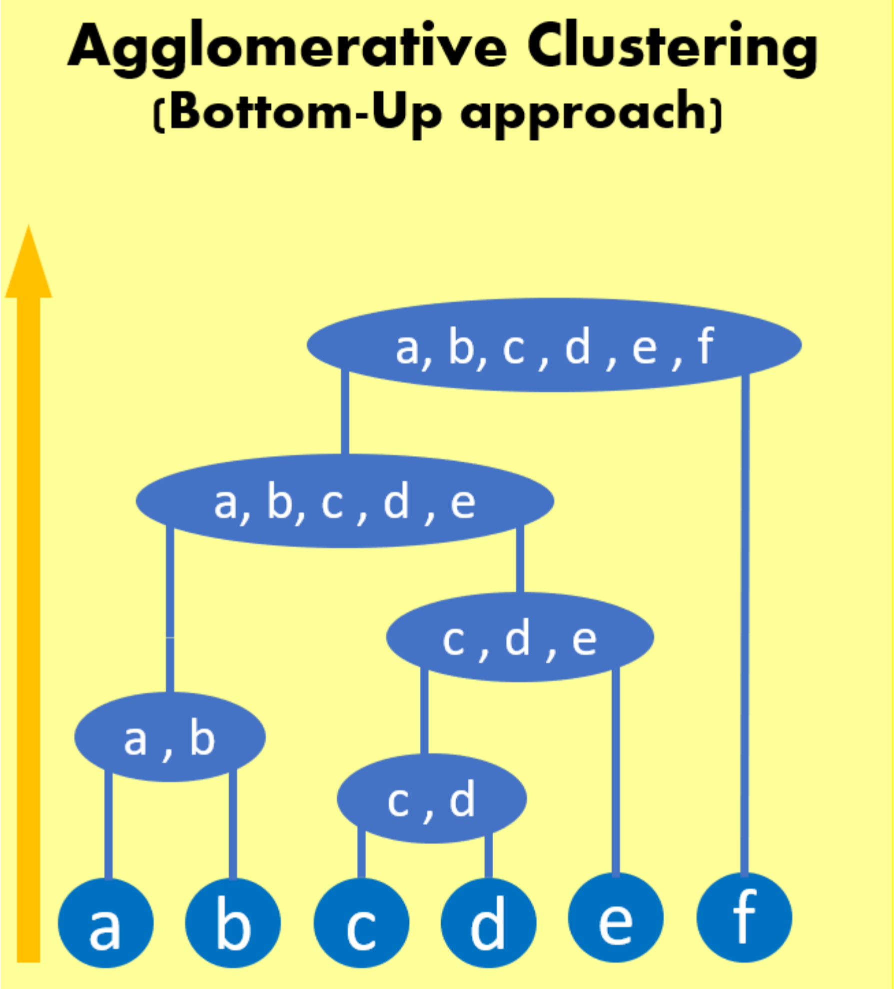
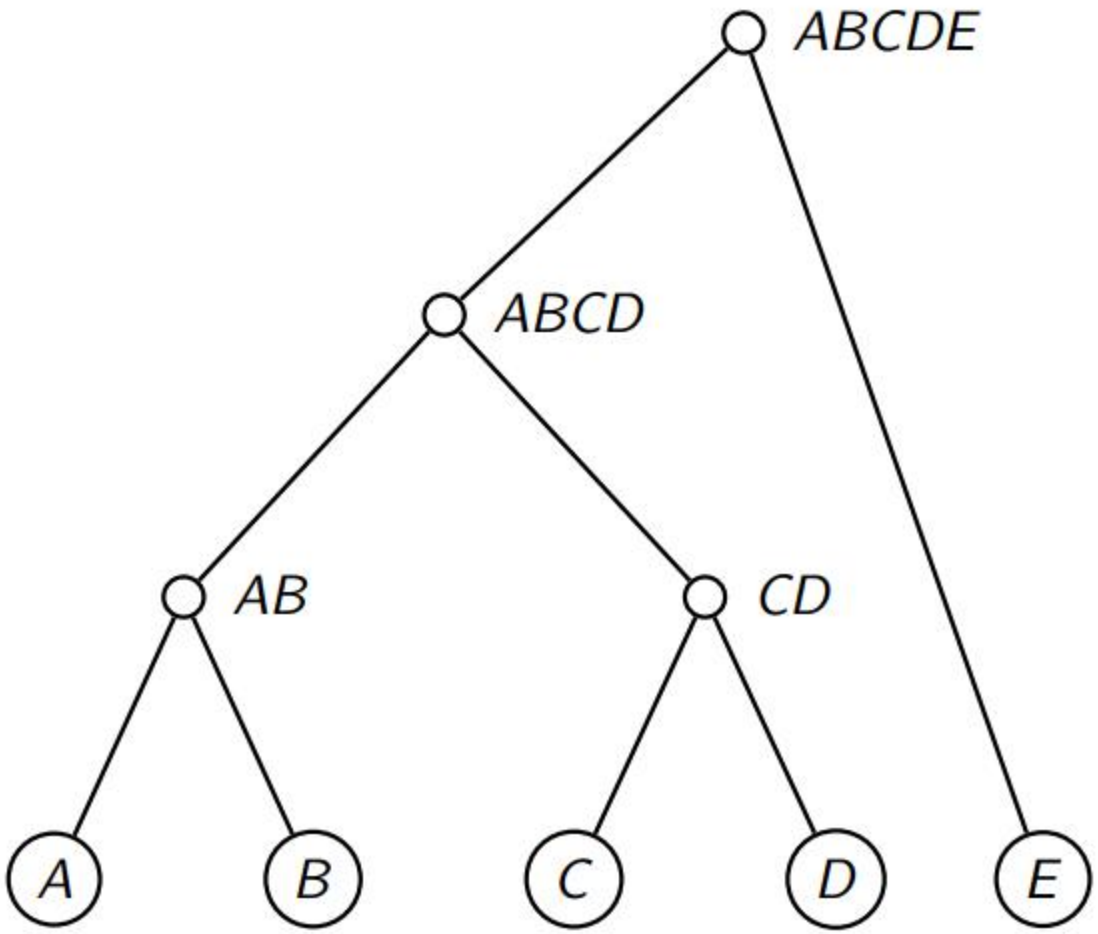


Image from: Dinesh Piyasamara.

Hierarchical Clustering Dendrogram

- Sequences of nested partitions that can be viewed with cluster dendrogram

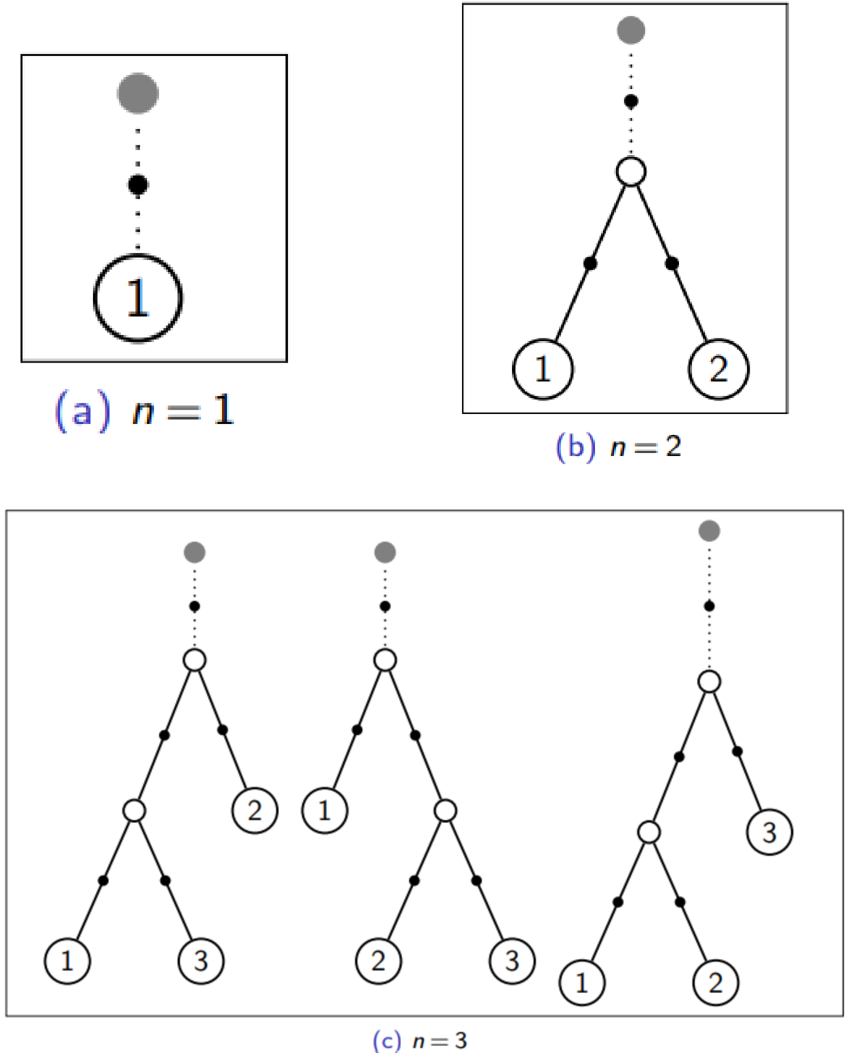
Clustering	Clusters
$\mathcal{C}_1$	$\{A\},\{B\},\{C\},\{D\},\{E\}$
$\mathcal{C}_2$	$\{AB\},\{C\},\{D\},\{E\}$
$\mathcal{C}_3$	$\{AB\},\{CD\},\{E\}$
$\mathcal{C}_4$	$\{ABCD\},\{E\}$
$\mathcal{C}_5$	$\{ABCDE\}$



Hierarchical Clustering Dendrogram

- Can have several different cluster dendrograms
- Number of different dendrograms computed from the following:

$$\prod_{m=1}^{n-1}(2m-1) = 1 \times 3 \times 5 \times 7 \times \cdots \times (2n-3) = (2n-3)!!$$



Hierarchical Clustering Dendrogram

- Do not need to specify number of clusters beforehand
- Can define distance threshold to set number of clusters

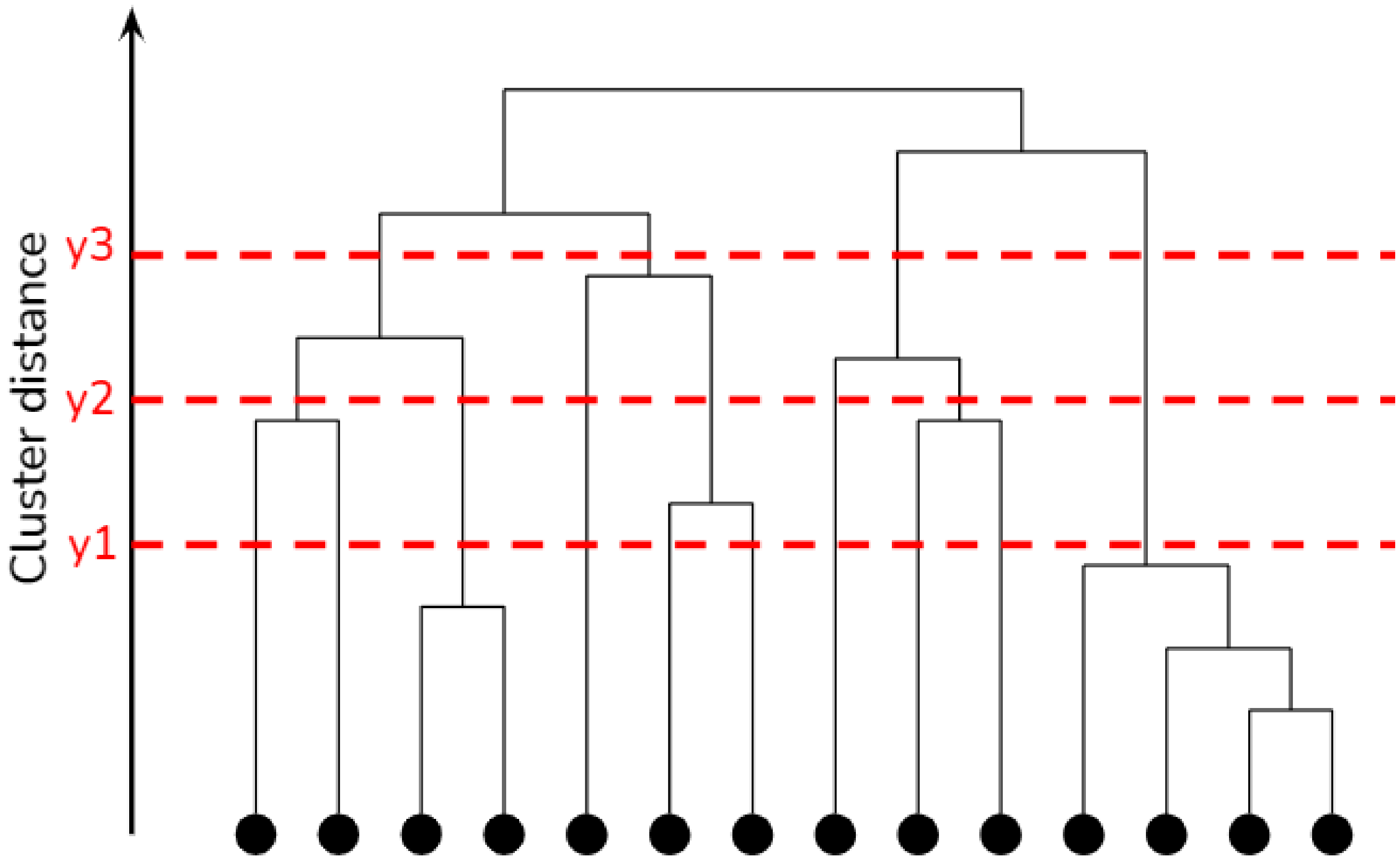


Image from: P. Pai, Hierarchical clustering explained.

Agglomerative Clustering Example

Agglomerative Clustering Example

TEXAS A&M UNIVERSITY Engineering

- Teacher wants to divide class into groups based on scores
- Need to generate proximity matrix

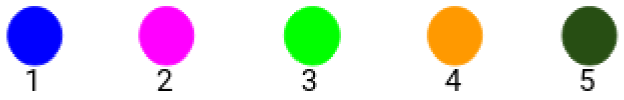
Student_ID	Marks
1	10
2	7
3	28
4	20
5	35

ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

Images from: P. Sharma, What is Hierarchical Clustering in Python?

Agglomerative Clustering Example

- Assign points to individual cluster
- Find smallest distances in proximity matrix to merge points

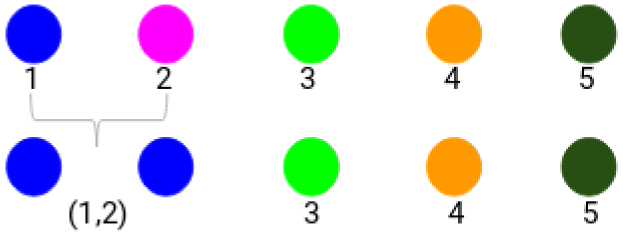


ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

Images from: P. Sharma, What is Hierarchical Clustering in Python?

Agglomerative Clustering Example

- After merge, need to update distance matrix
- Used maximum value to merge clusters but there are other “linkage” options
  - Minimum, average, etc.



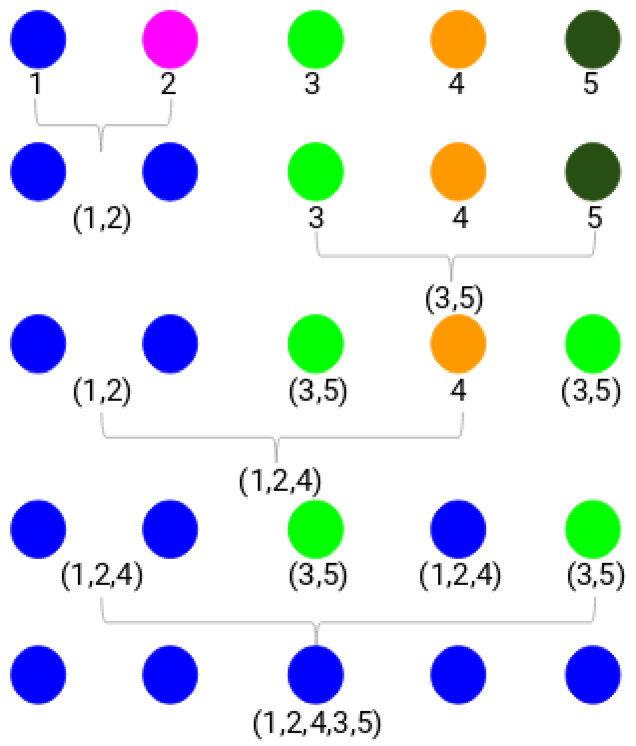
Student_ID	Marks
(1,2)	10
3	28
4	20
5	35

ID	(1,2)	3	4	5
(1,2)	0	18	10	25
3	18	0	8	7
4	10	8	0	15
5	25	7	15	0

Images from: P. Sharma, What is Hierarchical Clustering in Python?

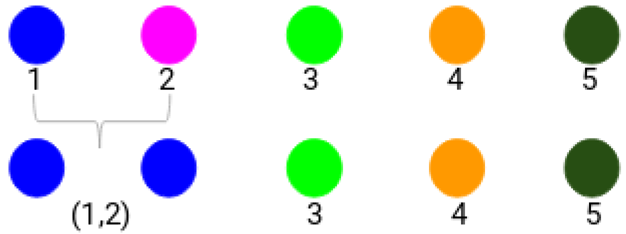
Agglomerative Clustering Example

- Repeat steps until single cluster remains
- Use dendrogram to decide groups

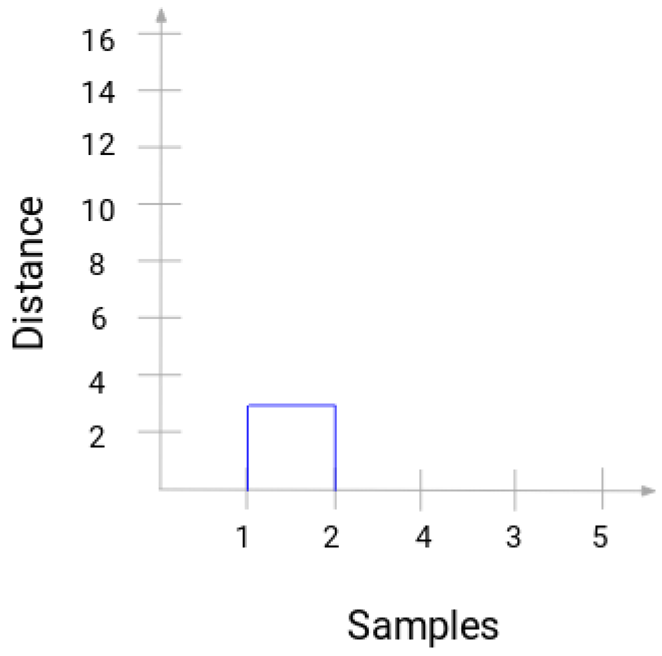


Agglomerative Clustering Example

- Dendrogram plots distances between clusters and shows cluster hierarchy



Student_ID	Marks
1	10
2	7
3	28
4	20
5	35

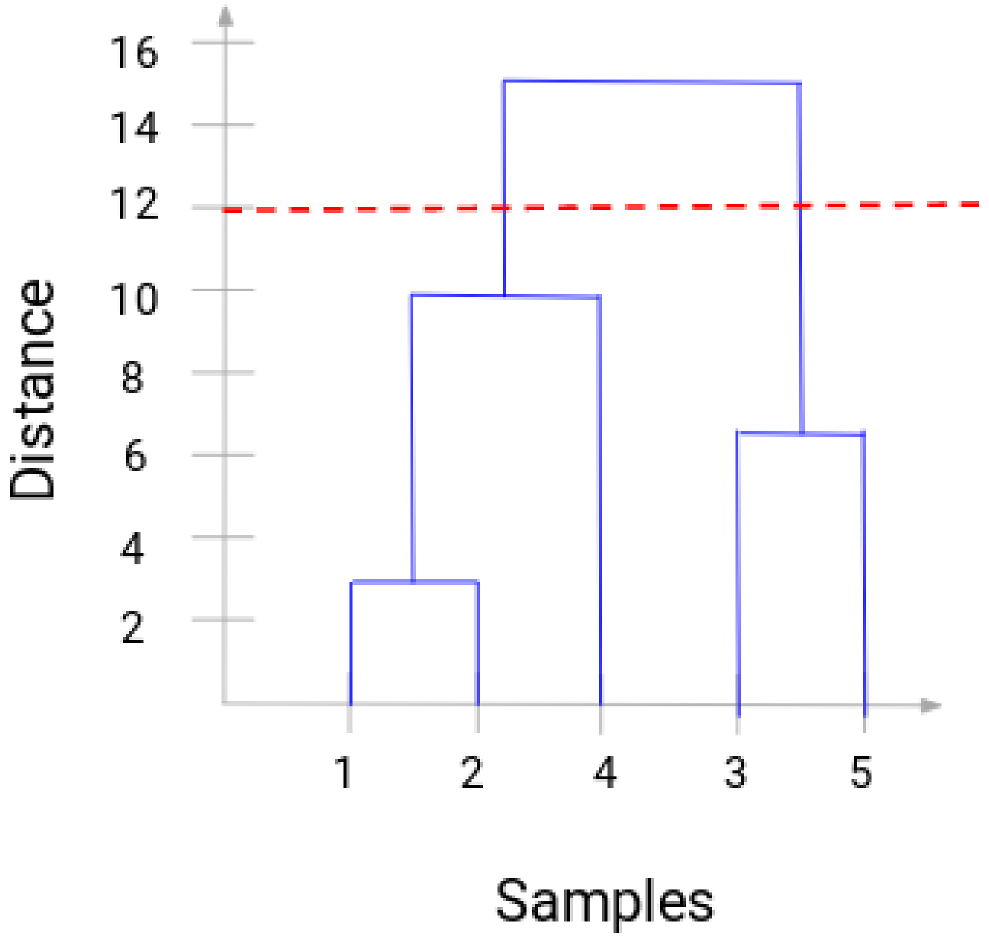


Images from: P. Sharma, What is Hierarchical Clustering in Python?

Agglomerative Clustering Example

TEXAS A&M UNIVERSITY Engineering

- Number of cluster will be number of vertical lines intersected with threshold



Agglomerative Clustering Algorithm

Agglomerative Hierarchal Clustering Pseudocode

AgglomerativeClustering(D, k)**:**

$\mathcal{C} \leftarrow \{C_i = \{\mathbf{x}_i\} \mid \mathbf{x}_i \in \mathbf{D}\}$ `// Each point in separate cluster`

$\Delta \leftarrow \{\delta(\mathbf{x}_i, \mathbf{x}_j) \colon \mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}\}$ `// Compute distance matrix`

repeat

  Find the closest pair of clusters $C_i, C_j \in \mathcal{C}$

  $C_{ij} \leftarrow C_i \cup C_j$ `// Merge the clusters`

  $\mathcal{C} \leftarrow (\mathcal{C} \setminus \{C_i, C_j\}) \cup \{C_{ij}\}$ `// Update the clustering`

  Update distance matrix Δ to reflect new clustering

until $|\mathcal{C}| = k$

Agglomerative Hierarchal Clustering Pseudocode

AgglomerativeClustering$(\boldsymbol{D}, k)$**:**

$\mathcal{C} \leftarrow \{C_i = \{\boldsymbol{x}_i\} \mid \boldsymbol{x}_i \in \boldsymbol{D}\}$ `// Each point in separate cluster`

$\boxed{\Delta \leftarrow \{\delta(\boldsymbol{x}_i, \boldsymbol{x}_j) \colon \boldsymbol{x}_i, \boldsymbol{x}_j \in \boldsymbol{D}\}\ \texttt{// Compute distance matrix}}$

repeat

  Find the closest pair of clusters $C_i, C_j \in \mathcal{C}$

  $C_{ij} \leftarrow C_i \cup C_j$ `// Merge the clusters`

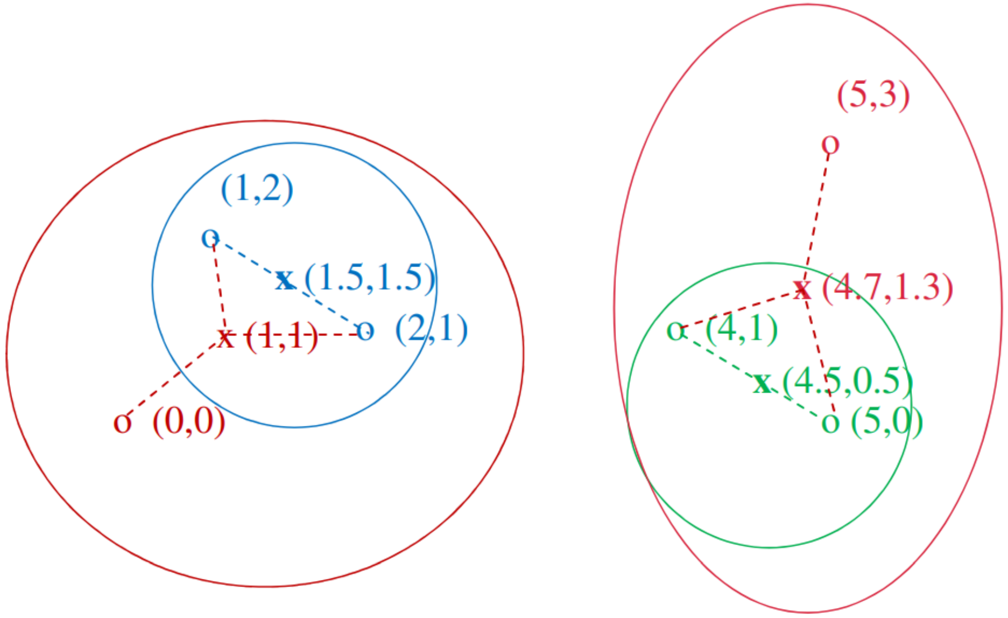
  $\mathcal{C} \leftarrow (\mathcal{C} \setminus \{C_i, C_j\}) \cup \{C_{ij}\}$ `// Update the clustering`

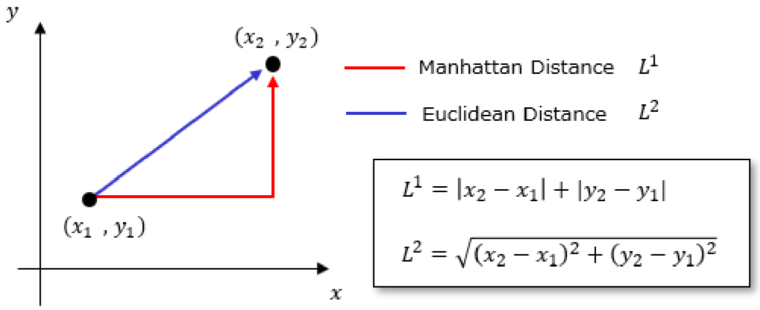
  Update distance matrix Δ to reflect new clustering

until $|\mathcal{C}| = k$

Distances Between Clusters

- Need "linkage" between clusters
- Different distance metrics are used:
  - Euclidean
  - Manhattan

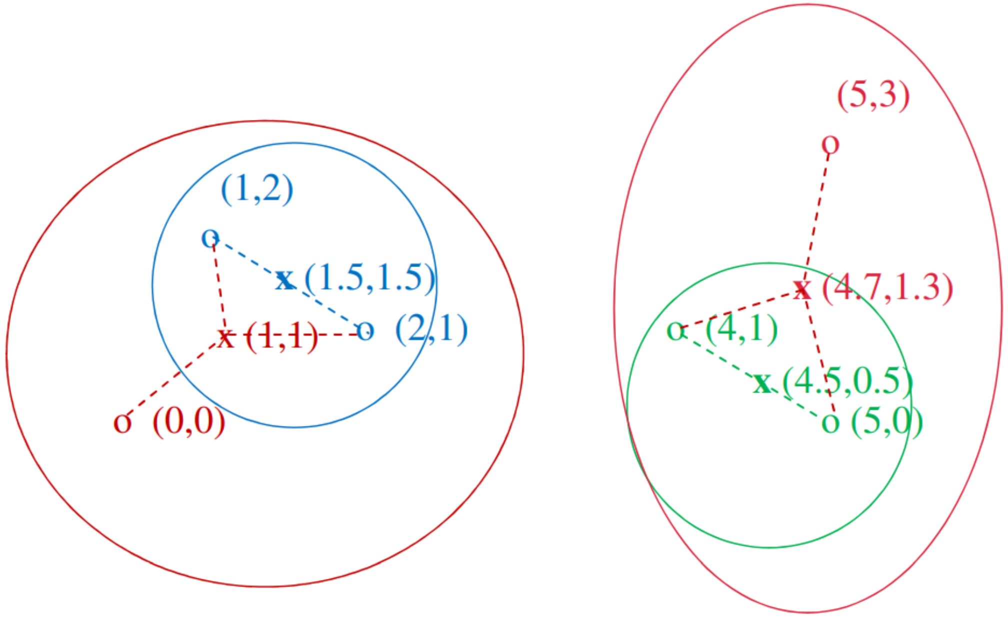




Data:
o … data point
x … centroid

Distances Between Clusters

- Single link
- Complete link
- Group average
- Mean distance
- Ward’s method
  - Minimum variance



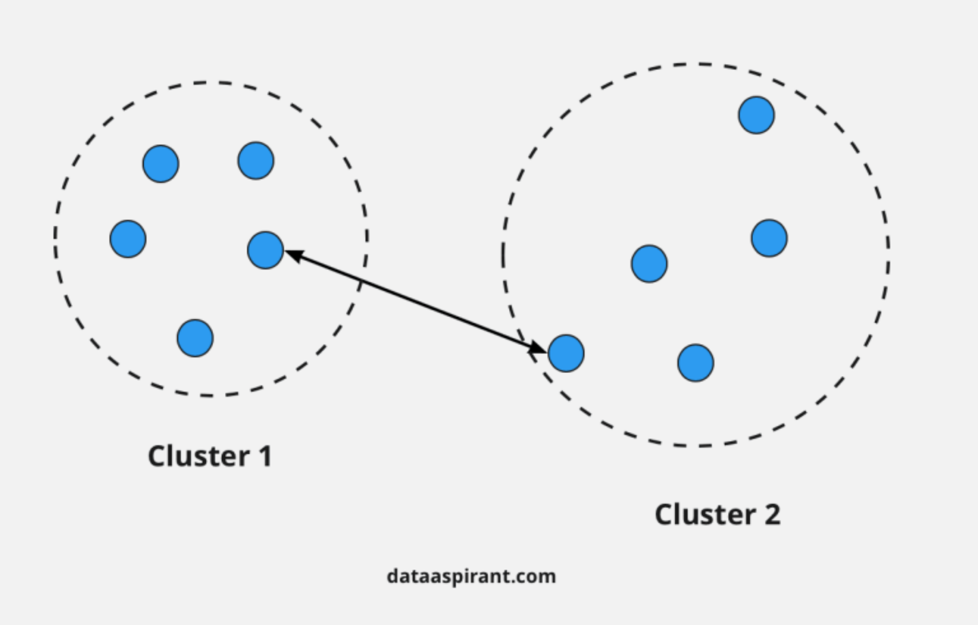
Data:
o … data point
x … centroid

Image from: S. Sultana, How the Hierarchal Clustering Algorithm Works

Distances Between Clusters

TEXAS A&M UNIVERSITY Engineering

- **Single link**
- Complete link
- Group average
- Mean distance
- Ward's method
  - Minimum variance

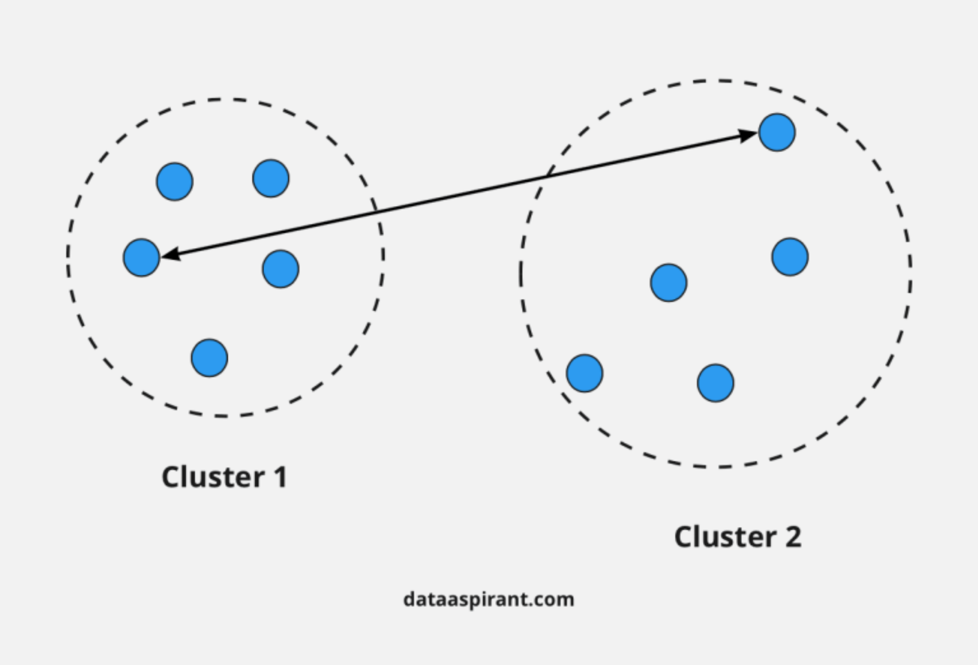


$$\delta(C_i, C_j) = \min\{\|\mathbf{x} - \mathbf{y}\| \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

Image from: S. Sultana, How the Hierarchal Clustering Algorithm Works

Distances Between Clusters

- Single link
- **Complete link**
- Group average
- Mean distance
- Ward’s method
  - Minimum variance

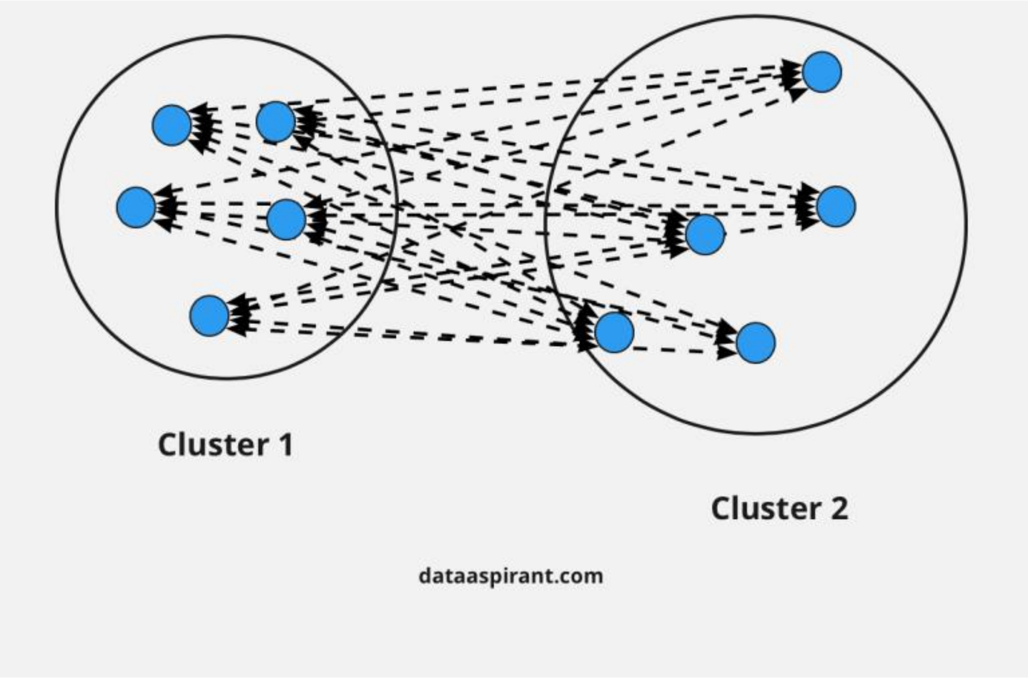


$$\delta(C_i, C_j) = \max\{\|\mathbf{x} - \mathbf{y}\| \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

Image from: S. Sultana, How the Hierarchal Clustering Algorithm Works

Distances Between Clusters

- Single link
- Complete link
- **Group average**
- Mean distance
- Ward's method
  - Minimum variance

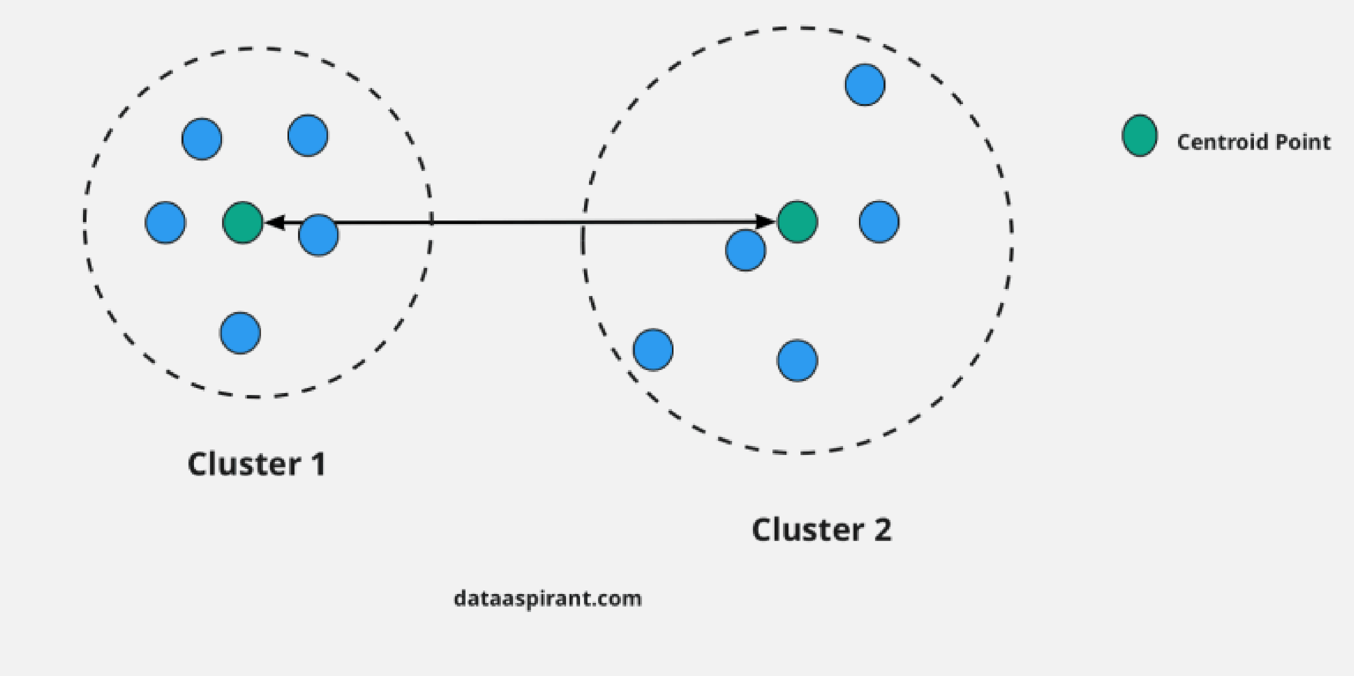


$$\delta(C_i, C_j) = \frac{\sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_j} \|\mathbf{x} - \mathbf{y}\|}{n_i \cdot n_j}$$

Image from: S. Sultana, How the Hierarchal Clustering Algorithm Works

Distances Between Clusters

- Single link
- Complete link
- Group average
- **Mean distance**
- Ward's method
  - Minimum variance

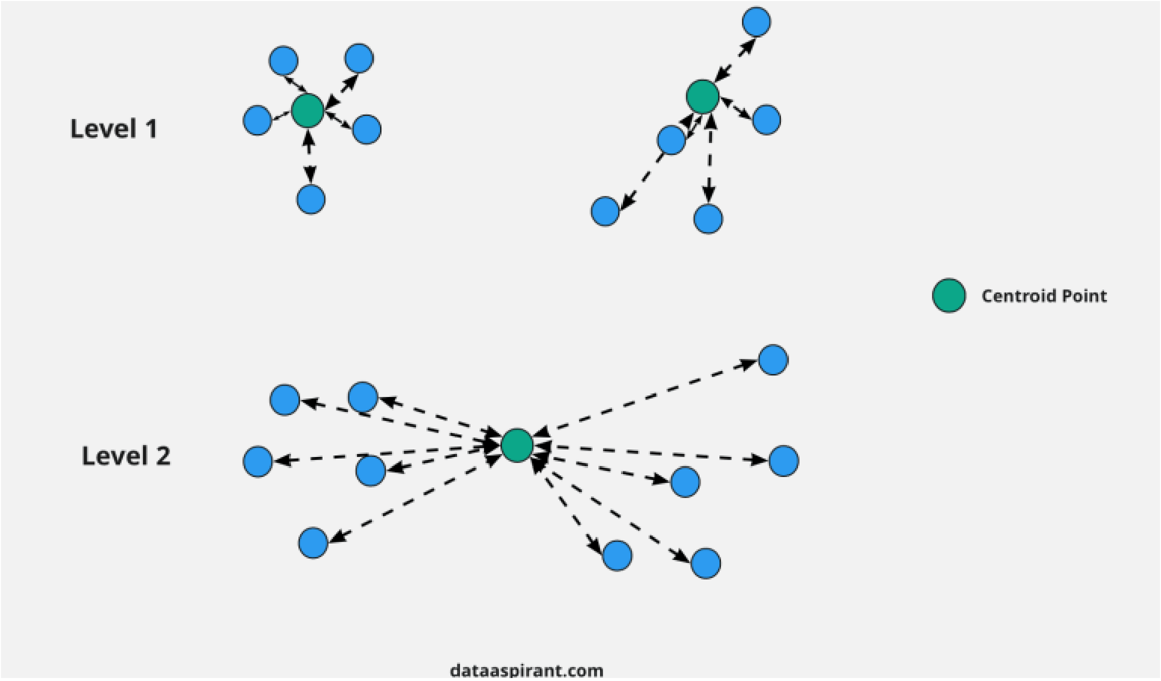


$$\delta(C_i, C_j) = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$$

Image from: S. Sultana, How the Hierarchal Clustering Algorithm Works

Distances Between Clusters

- Single link
- Complete link
- Group average
- Mean distance
- **Ward's method**
  - **Minimum variance**



$$\delta(C_i, C_j) = \left(\frac{n_i n_j}{n_i + n_j}\right) \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2$$

Image from: S. Sultana, How the Hierarchal Clustering Algorithm Works

Agglomerative Hierarchal Clustering Pseudocode

AgglomerativeClustering(D, k)**:**

$\mathcal{C} \leftarrow \{C_i = \{\mathbf{x}_i\} \mid \mathbf{x}_i \in \mathbf{D}\}$ `// Each point in separate cluster`

$\Delta \leftarrow \{\delta(\mathbf{x}_i, \mathbf{x}_j) \colon \mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}\}$ `// Compute distance matrix`

repeat

    Find the closest pair of clusters $C_i, C_j \in \mathcal{C}$

    $C_{ij} \leftarrow C_i \cup C_j$ `// Merge the clusters`

    $\mathcal{C} \leftarrow (\mathcal{C} \setminus \{C_i, C_j\}) \cup \{C_{ij}\}$ `// Update the clustering`

    Update distance matrix Δ to reflect new clustering

until $|\mathcal{C}| = k$

Lance-Williams Formula

- After merging, need to update distance matrix
- Lance-Williams formula provides general equation to recompute distances

$$\delta(C_{ij}, C_r) = \alpha_i \cdot \delta(C_i, C_r) + \alpha_j \cdot \delta(C_j, C_r) + \beta \cdot \delta(C_i, C_j) + \gamma \cdot \left|\delta(C_i, C_r) - \delta(C_j, C_r)\right|$$

Lance-Williams Formula

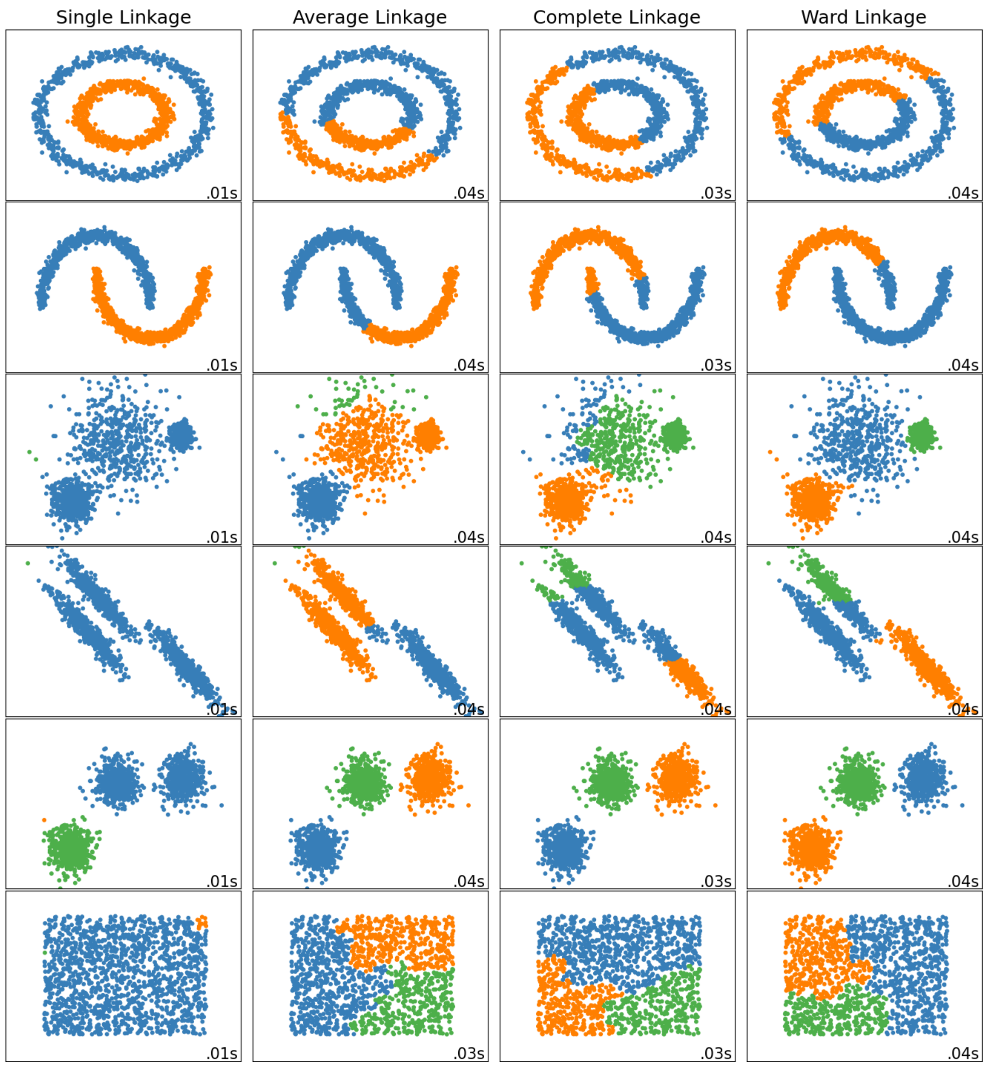
- Equations for each measure are derived in ZM Ch. 14 slides!

$$\delta(C_{ij}, C_r) = \alpha_i \cdot \delta(C_i, C_r) + \alpha_j \cdot \delta(C_j, C_r) + \beta \cdot \delta(C_i, C_j) + \gamma \cdot \left|\delta(C_i, C_r) - \delta(C_j, C_r)\right|$$

Measure	α_i	α_j	β	γ
Single link	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete link	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Group average	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
Mean distance	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$\frac{-n_i \cdot n_j}{(n_i+n_j)^2}$	0
Ward's measure	$\frac{n_i+n_r}{n_i+n_j+n_r}$	$\frac{n_j+n_r}{n_i+n_j+n_r}$	$\frac{-n_r}{n_i+n_j+n_r}$	0

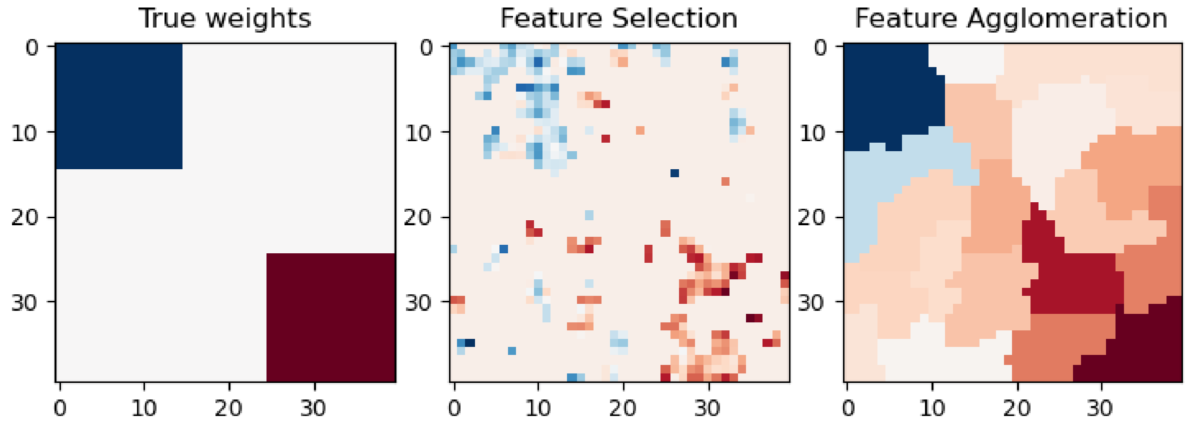
Agglomerative Hierarchal Clustering Implementation

- Available in Sklearn
- Four linkage strategies:
  - Single
  - Complete
  - Average
  - Ward



Agglomerative Hierarchal Clustering Implementation

- Can also use for dimensionality reduction
- Instead of clustering samples, you can cluster features
  - [Feature Agglomeration]



Divisive Clustering Overview

Hierarchical Clustering

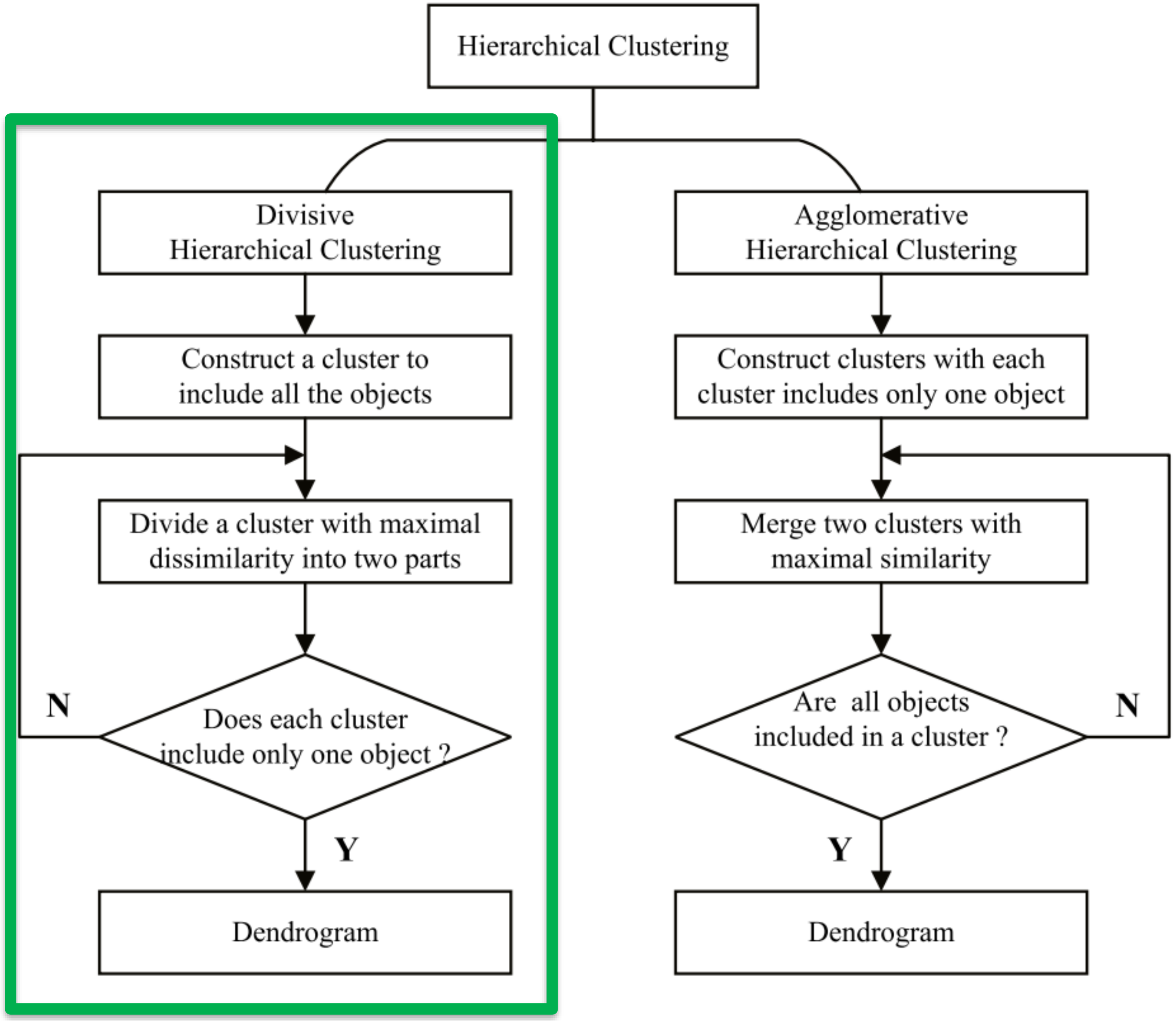


Figure from: X. Ran, Comprehensive survey on hierarchical clustering algorithms and the recent developments

Divisive Clustering

TEXAS A&M UNIVERSITY Engineering

- “Top-down” clustering approach
- Creates sequences of nested partitions that can be viewed with cluster dendrogram
- All points start in single cluster and are split

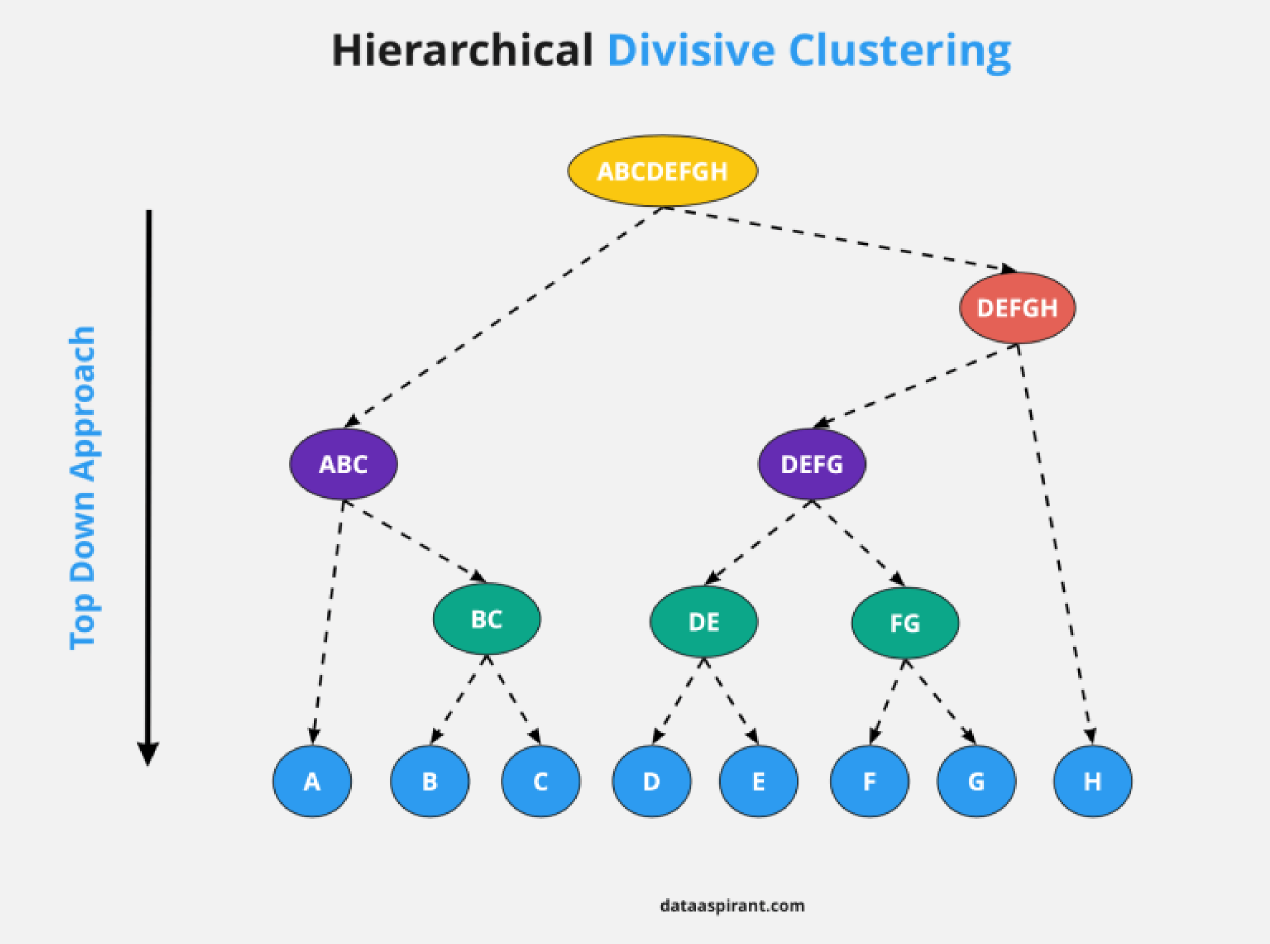


Image from: S. Sultana, How the Hierarchal Clustering Algorithm Works

Divisive Clustering: Bisecting k-Means

- Steps:
  - Initialization
  - Bisection
  - Update SSE
  - Repeat bisection and SSE updates
  - Split until desired number of clusters are reached

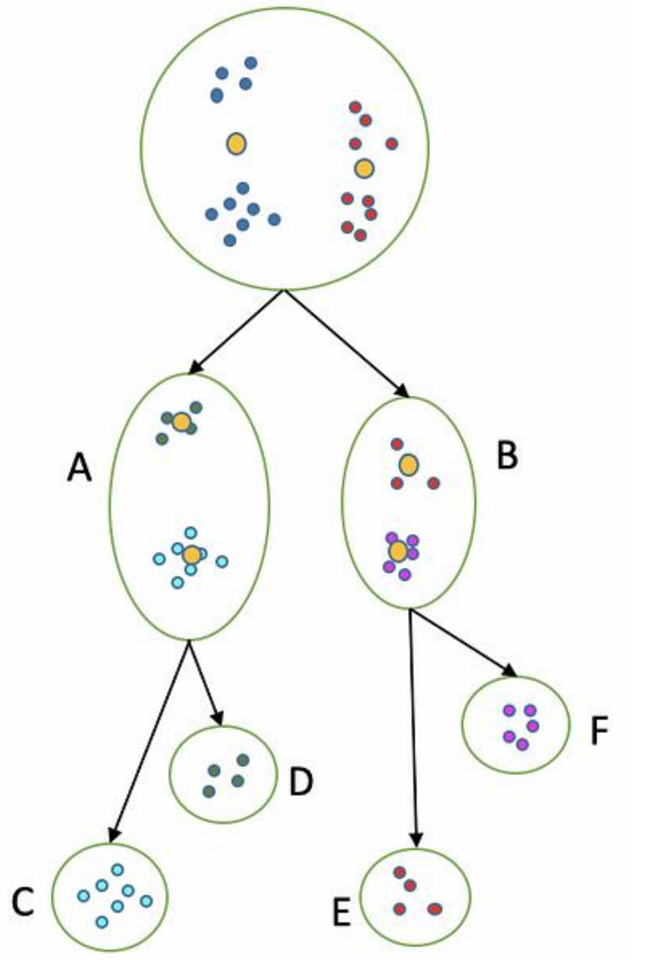


Image from: A. Firdaus, Bisecting k-Means Clustering.

Divisive Clustering: Bisecting k-Means

- Advantages over k-Means:
  - More robust to outliers and complex data structures
  - Hierarchal organization
  - Faster convergence

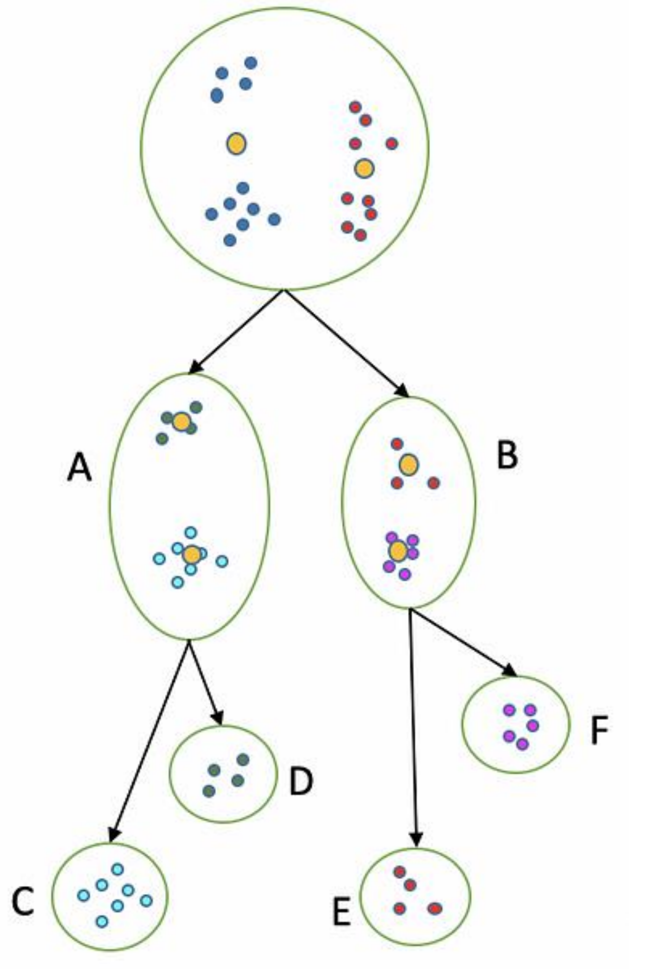
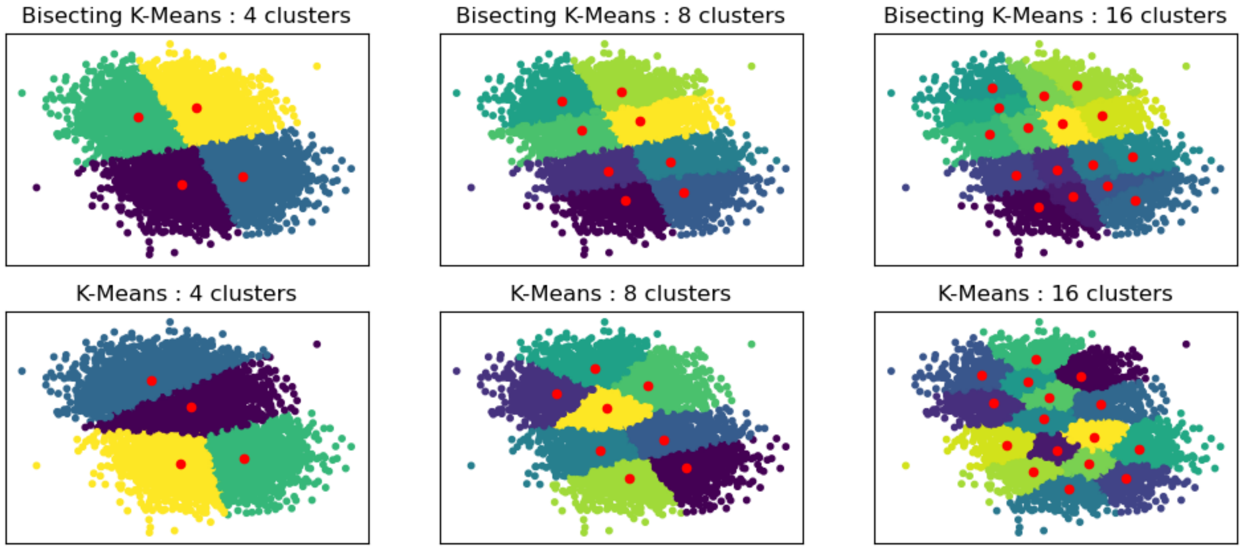


Image from: A. Firdaus, Bisecting k-Means Clustering.

Bisecting k-Means Implementation

- Available in Sklearn
- Two splitting strategies:
  - "Biggest inertia"
    - Split cluster with largest SSE
  - "Largest cluster"
    - Split cluster with largest number of data points



Hierarchical Clustering Summary

Hierarchical Clustering Summary

- Advantages
  - No assumptions of cluster shapes
  - No need to set number of clusters beforehand
  - Interpretable
- Disadvantages
  - Costly for large datasets
  - Difficult to visualize dendrograms for large datasets
  - Results depend on linkage
  - Prediction of new points not straightforward

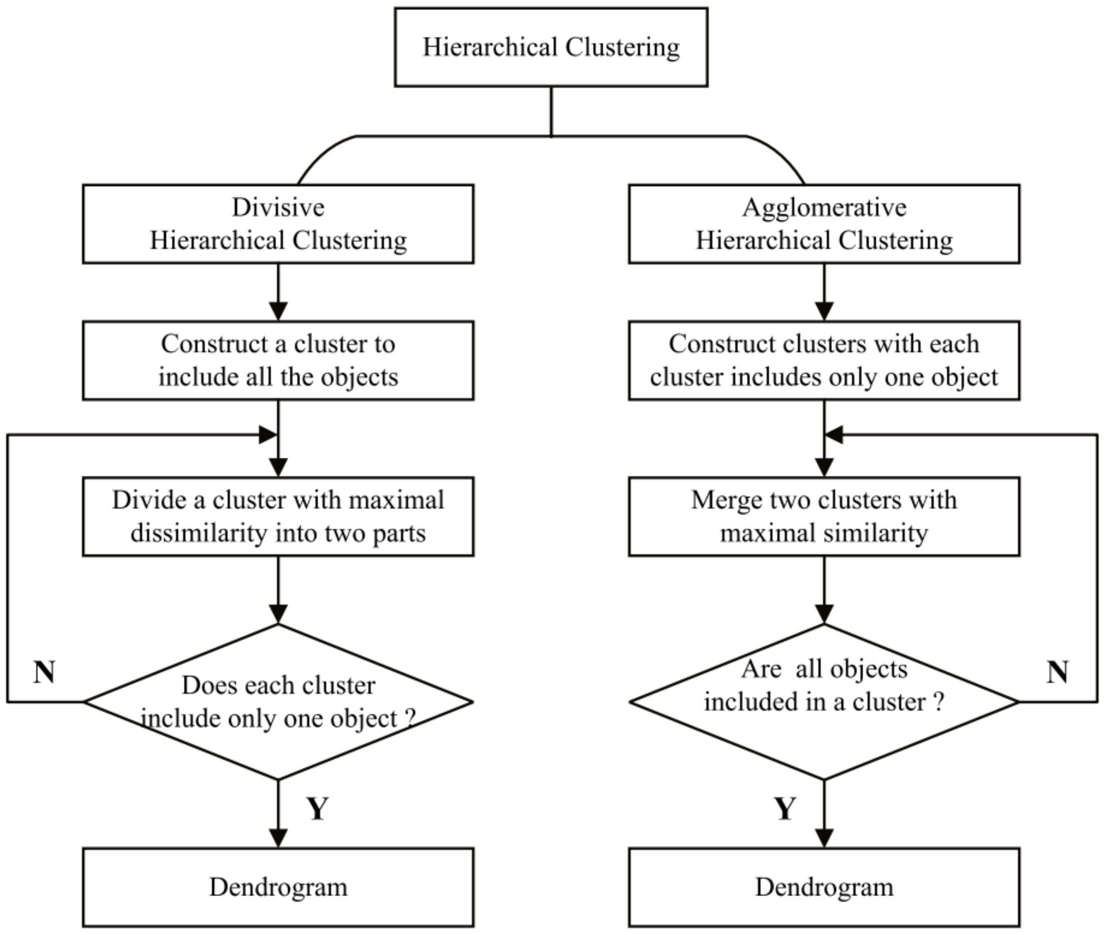
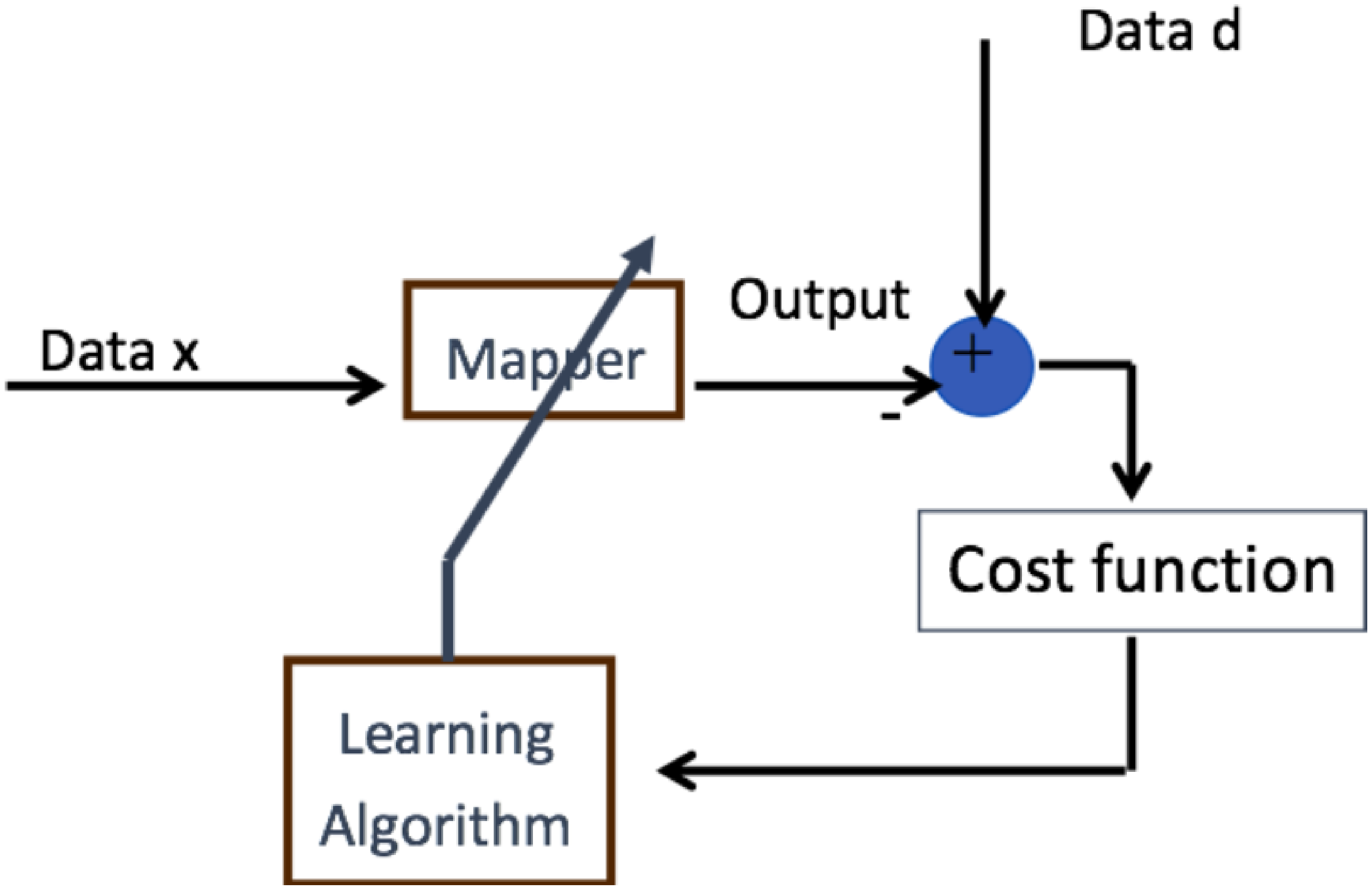


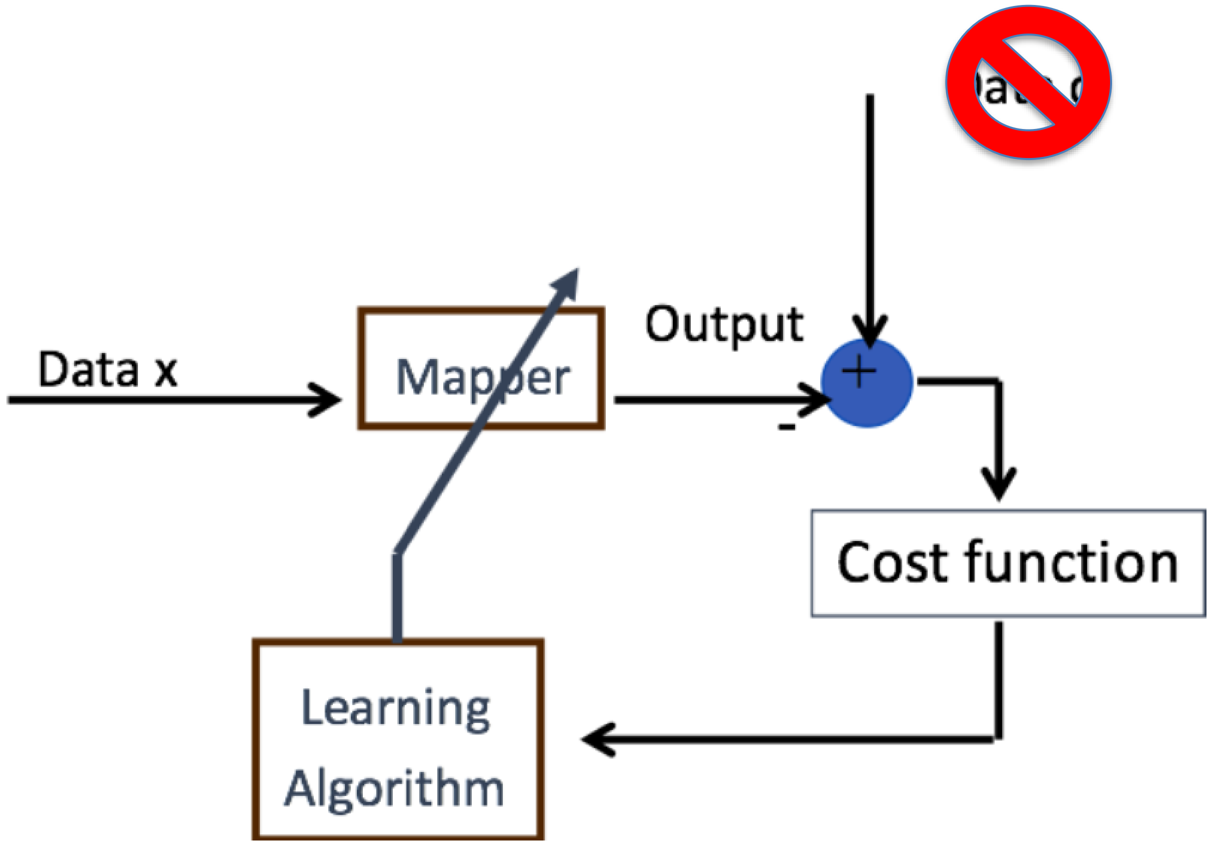
Figure from: X. Ran, Comprehensive survey on hierarchical clustering algorithms and the recent developments

Hierarchical Clustering Machine Learning Model



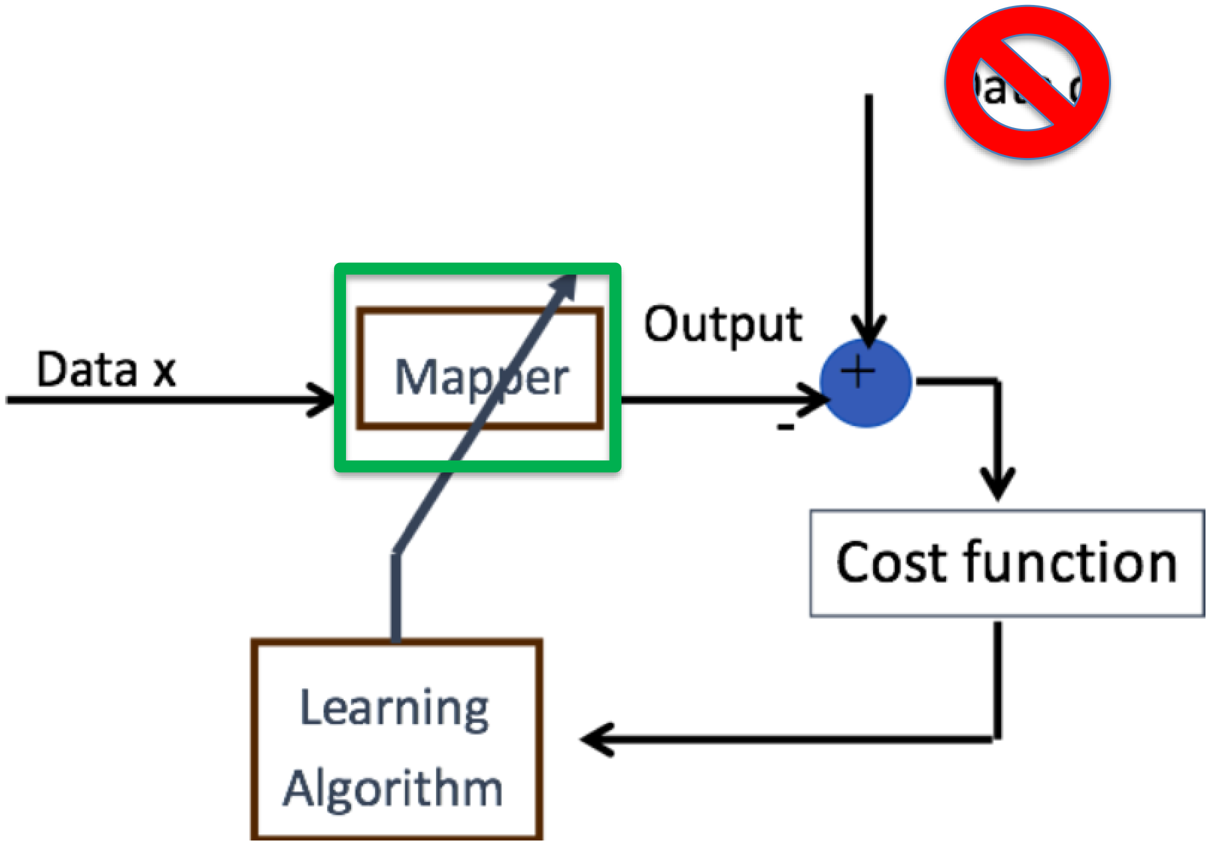
Hierarchical Clustering Machine Learning Model

- Unsupervised: No labels, *d*



Hierarchical Clustering Machine Learning Model

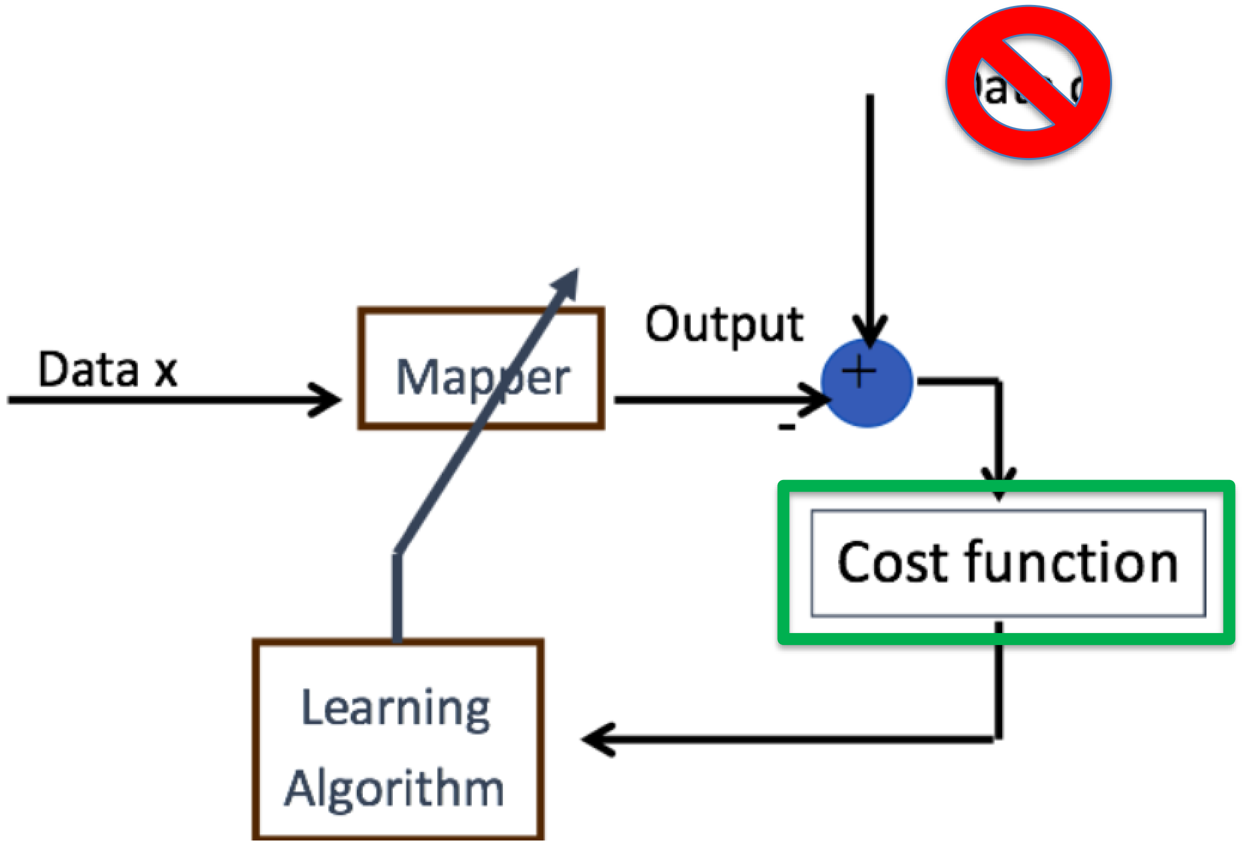
- Unsupervised: No labels, *d*
- **Mapper:**
  - Hierarchical Clustering
  - Takes input data and groups into *k* clusters



Hierarchical Clustering Machine Learning Model

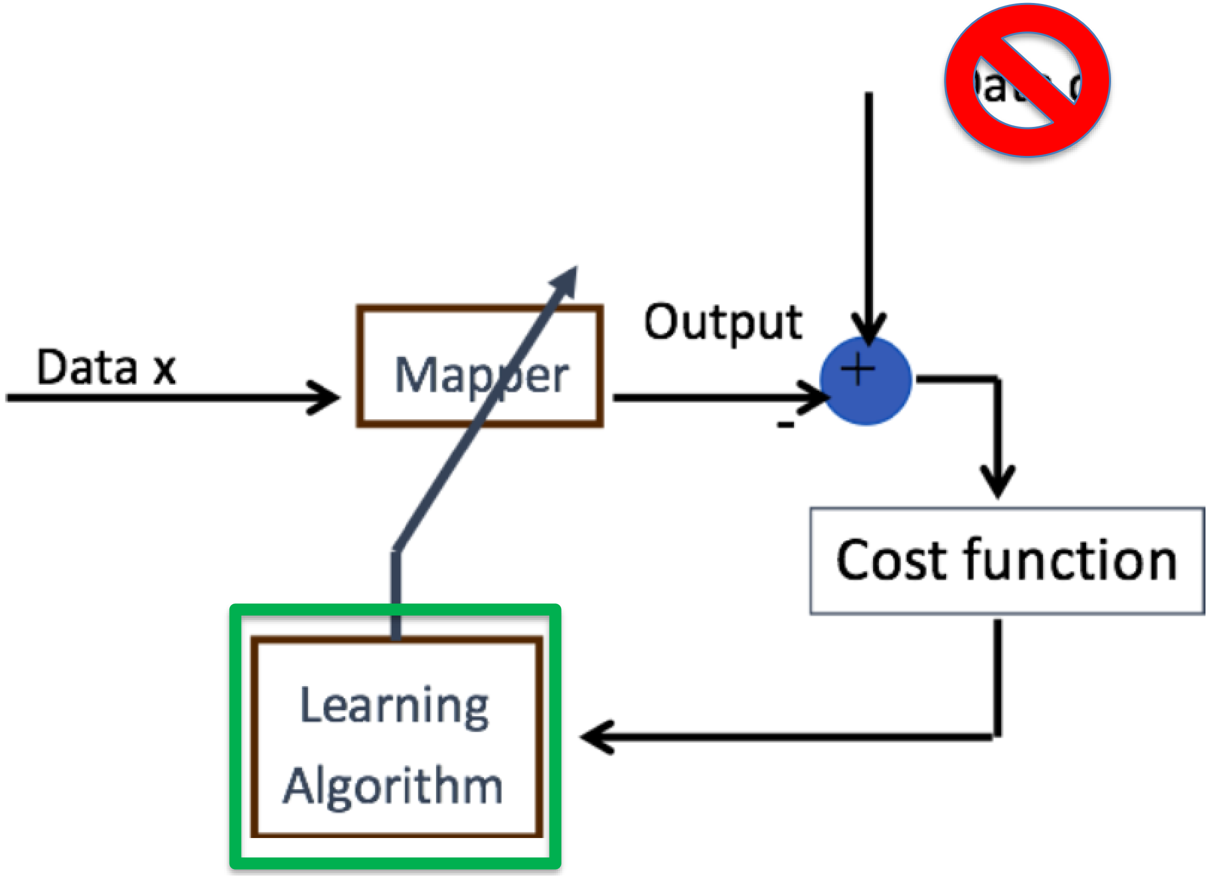
- Unsupervised: No labels, *d*
- Mapper:
  - Hierarchical Clustering
  - Takes input data and groups into *k* clusters
- **Cost function:**
  - Linkage

$$\delta(\mathbf{x}_i, \mathbf{x}_j): \mathbf{x}_i, \mathbf{x}_j \in \boldsymbol{D}$$



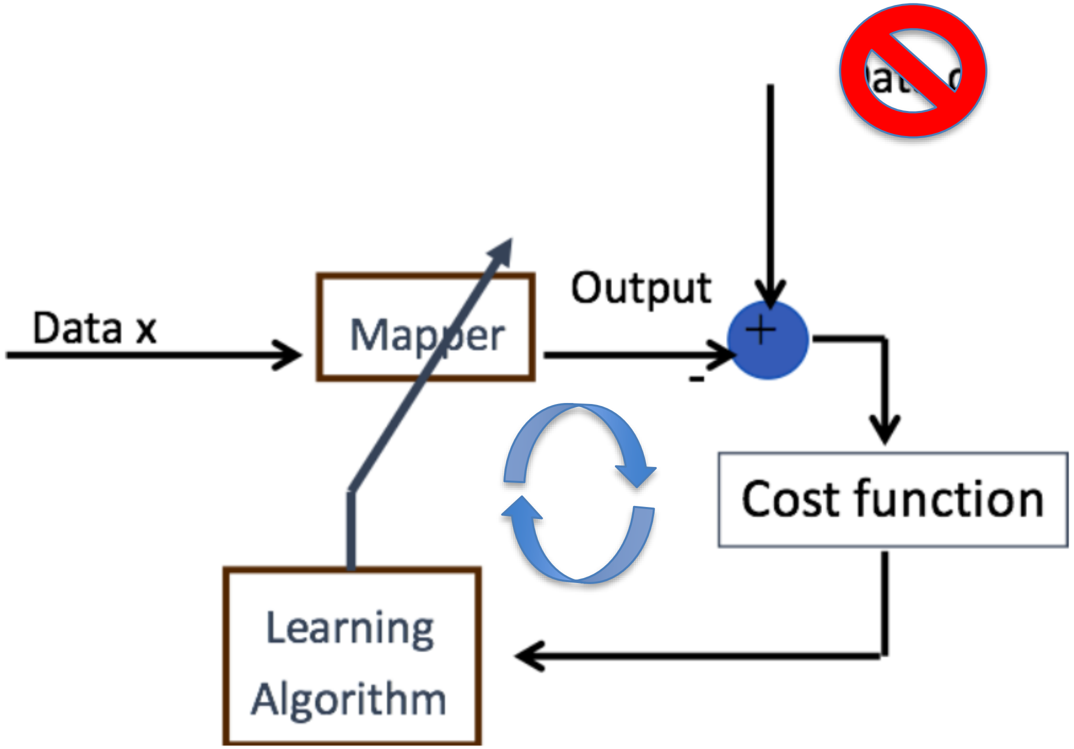
Hierarchical Clustering Machine Learning Model

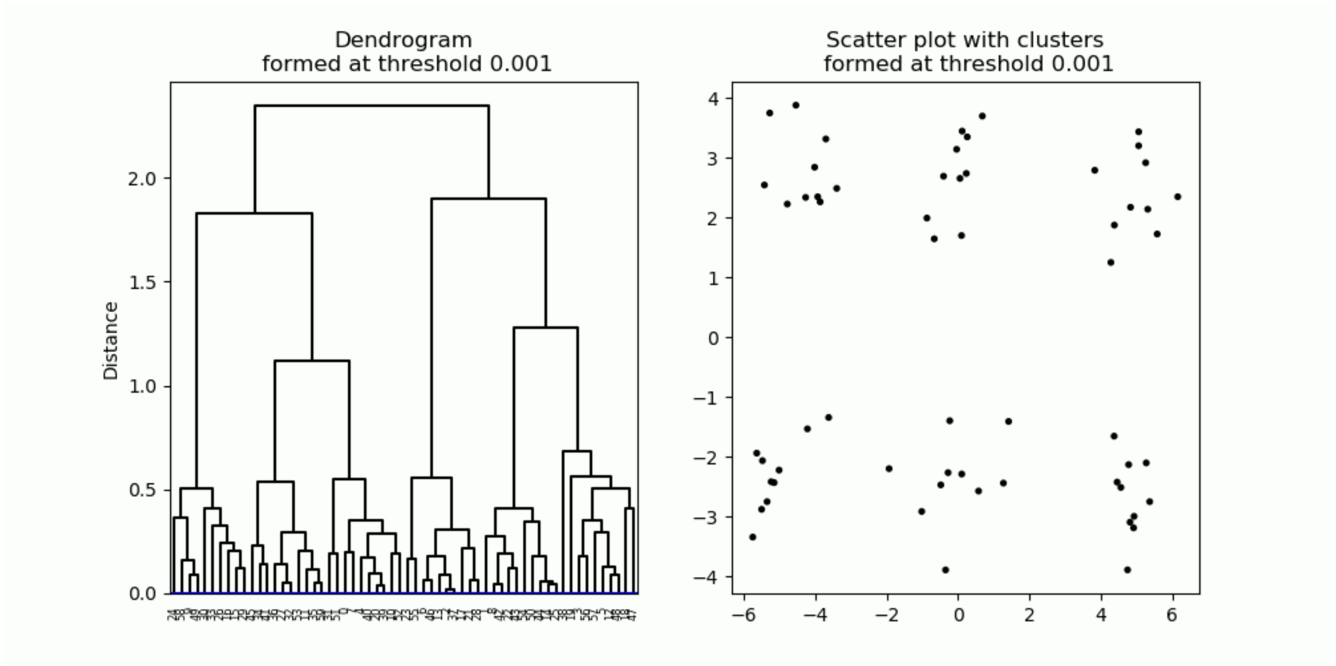
- Unsupervised: No labels, *d*
- Mapper:
  - Hierarchical Clustering
  - Takes input data and groups into *k* clusters
- Cost function:
  - Linkage
- **Learning algorithm**
  - Lance-Williams algorithm



$$\delta(C_{ij}, C_r) = \alpha_i \cdot \delta(C_i, C_r) + \alpha_j \cdot \delta(C_j, C_r) + \beta \cdot \delta(C_i, C_j) + \gamma \cdot |\delta(C_i, C_r) - \delta(C_j, C_r)|$$

Hierarchical Clustering Machine Learning Model



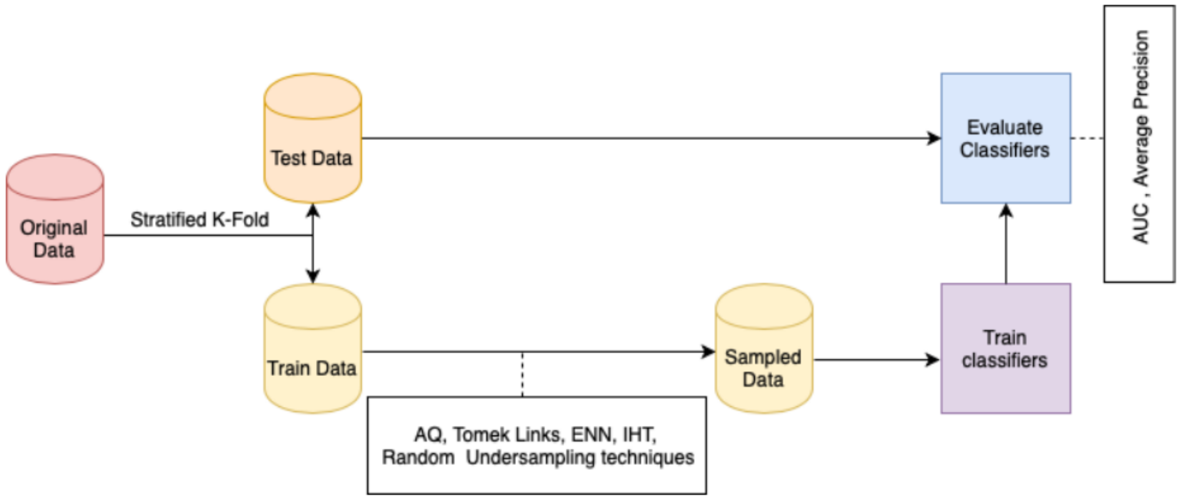


Gif from: A. Makkar, Vast Data Deduction: insight on how a Computer Pinpoints Data

Hierarchical Clustering Application

Bias in Machine Learning

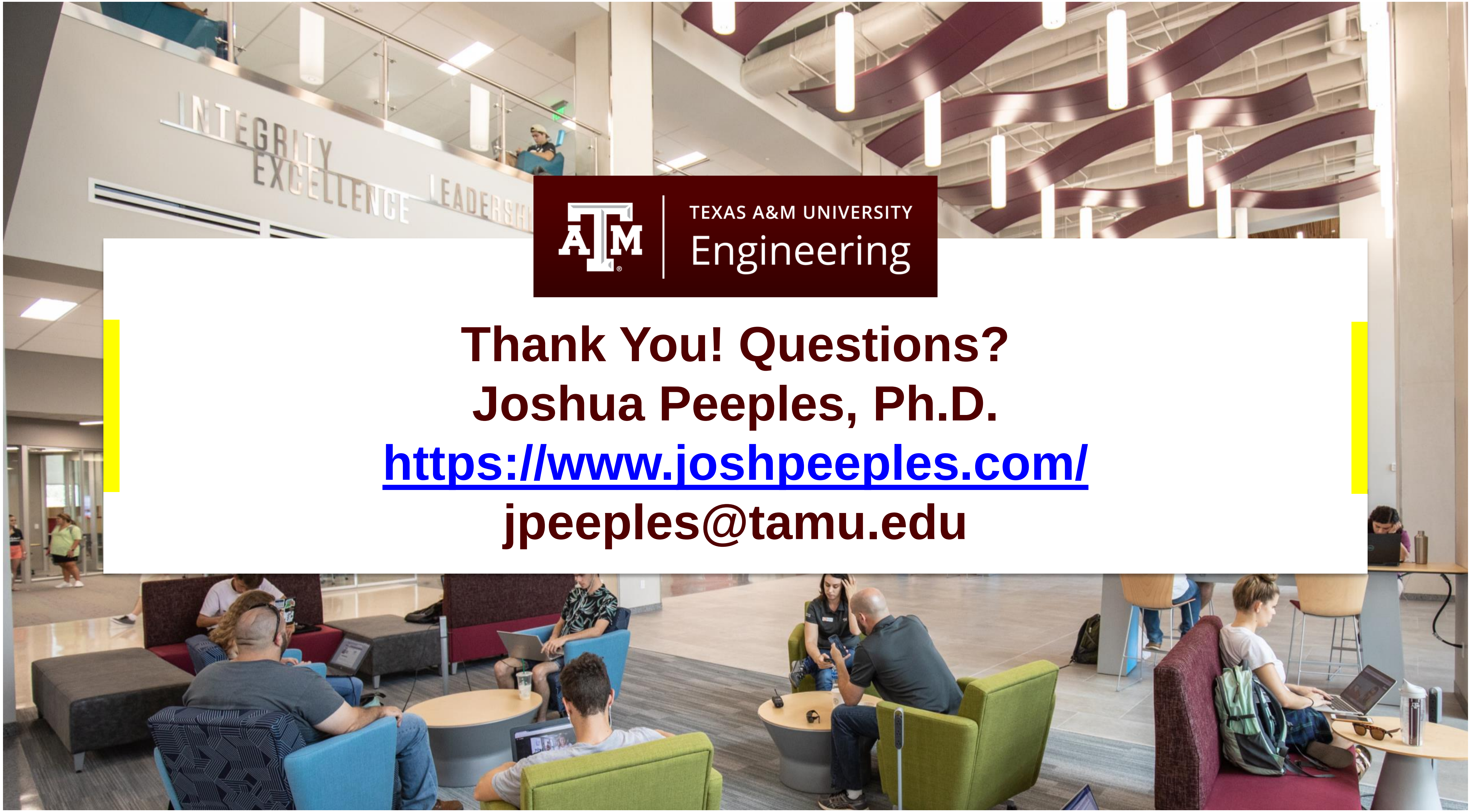
- Lack of diversity and representativeness in data can cause several issues
- Used a divisive clustering approach to improve the diversity in training data without reducing accuracy



D. Prioleau, et al., Application of Divisive Clustering for Reducing Bias in Imbalanced Data, Machine Learning and Data Mining in Pattern Recognition, 2021.

Next class

- Density estimation and density-based clustering



TEXAS A&M UNIVERSITY
Engineering

Thank You! Questions?

Joshua Peeples, Ph.D.
https://www.joshpeeples.com/
jpeeples@tamu.edu

TEXAS A&M UNIVERSITY
Engineering

Supplemental Slides

Useful Links

TEXAS A&M UNIVERSITY Engineering

- StatQuest: Hierarchical Clustering
- Hierarchical Clustering Cluster Distances
- How the Hierarchical Clustering Algorithm Works
- Lance-Williams Algorithm