MACHINE LEARNING

Shoaib Farooq

Department of Computer Science

Instructor Information

Instructor	Shoaib Farooq	E-mail	m.shoaib1050@gmail.com
GitHub	https://github.com/SHOAIB1050		
YouTube	https://www.youtube.com/c/pakithub		
LinkedIn	https://www.linkedin.com/in/shoaib-farooq-b5b190105/		

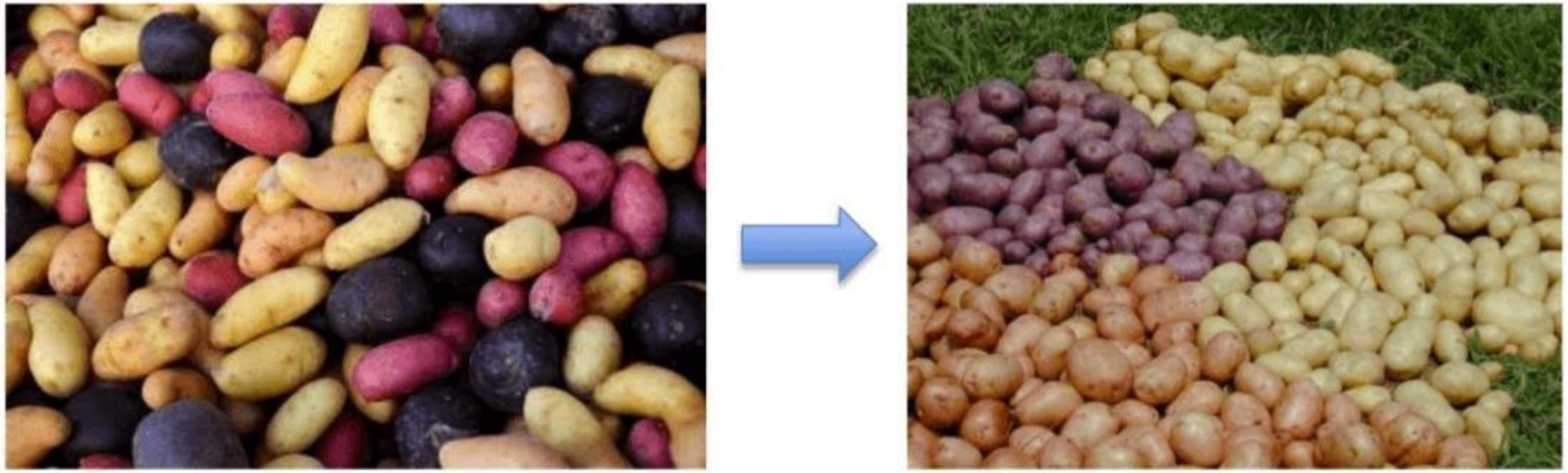
Let’s Start …..

Lecture #14

GOALS

- This Lecture Will Cover:
  - Unsupervised Learning
  - Clustering
  - K Means Clustering
  - Hierarchical Clustering

UNSUPERVISED LEARNING

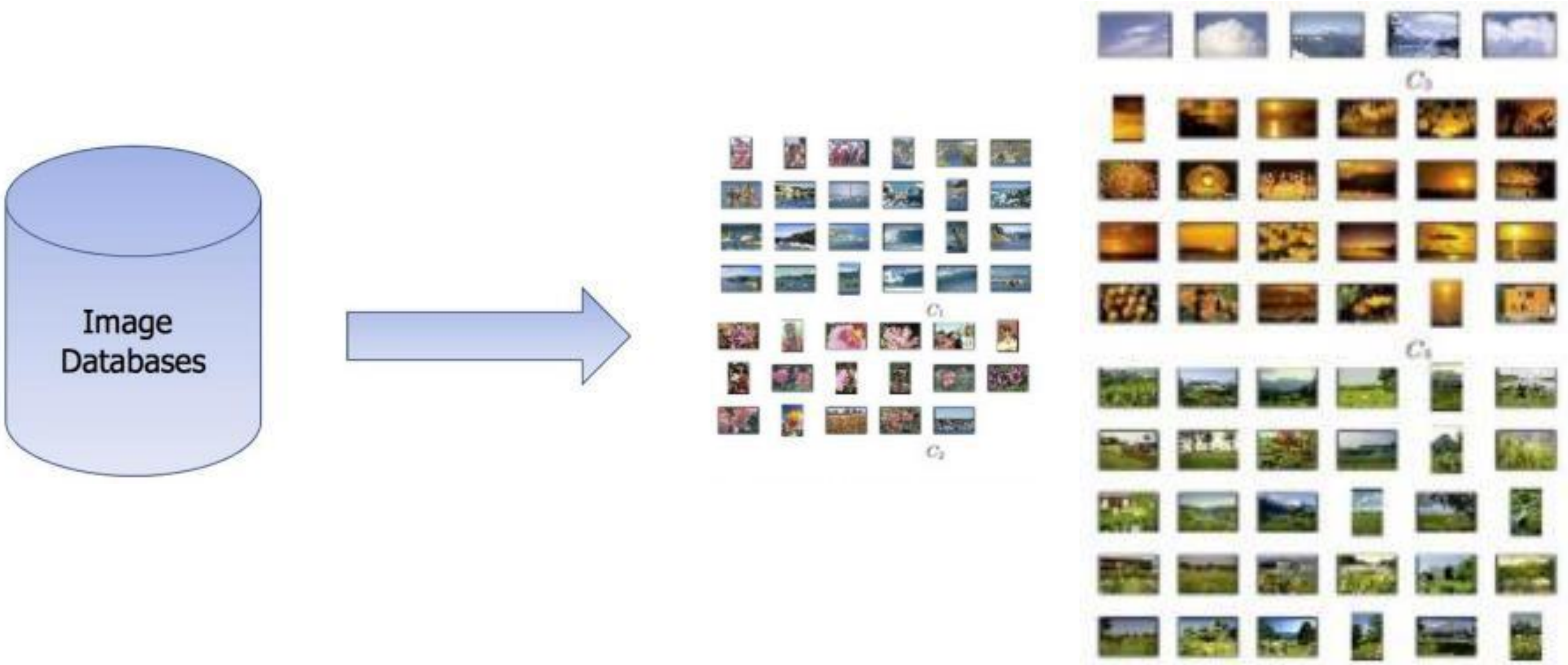


sample

Cluster/group

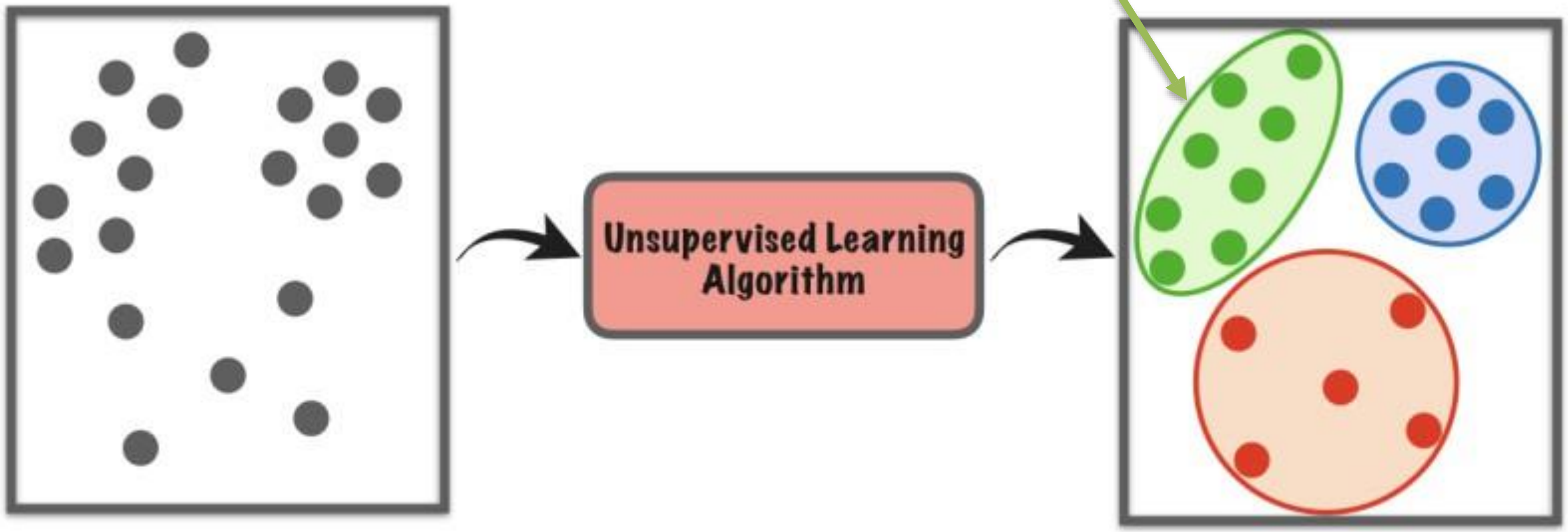
Presented by Shoaib Farooq

UNSUPERVISED LEARNING



UNSUPERVISED LEARNING

Why they are in one group?



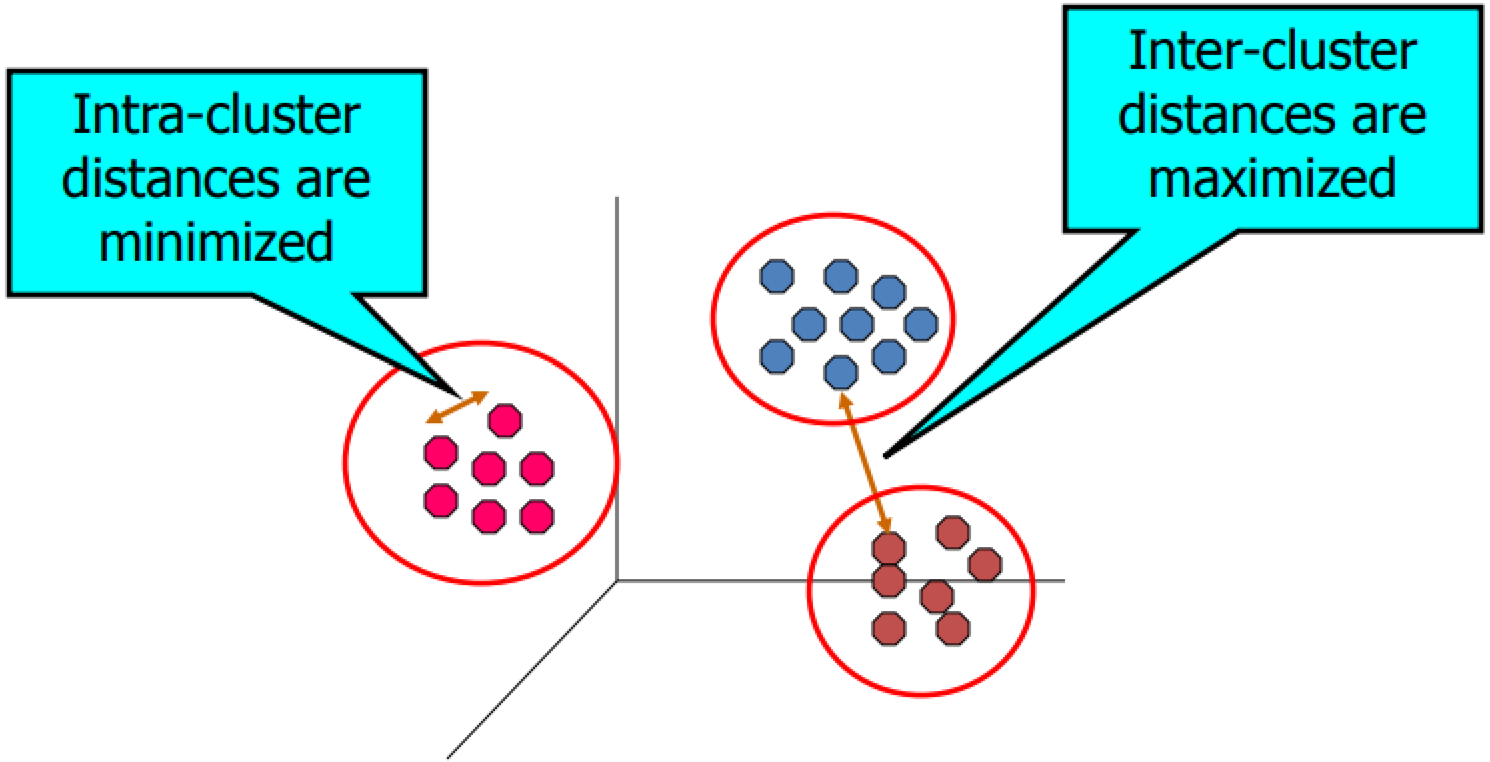
CLUSTERING

- Partition unlabeled examples into disjoint subsets of *clusters (Groups)*, such that:
  - Samples within a cluster are very similar
  - Samples in different clusters are very different
- Discover new categories in an *unsupervised* manner
  - No labels provided.

CLUSTERING

- Select a Similarity / Dissimilarity function
- Group samples based on this function
  - Similar ones be in one cluster
  - Different put in other cluster

CLUSTERING



Intra-cluster distances are minimized
Inter-cluster distances are maximized

CLUSTERING

- There are many types of clustering algorithms.
- Many algorithms use similarity or dissimilarity criteria
- The way they find the grouping is different.
- Some popular clustering algorithms are:
  - K-Means Clustering
  - Hierarchical Clustering
  - Mean Shift
  - Spectral Clustering
  - Mixture of Gaussians

K-MEANS CLUSTERING

- K-means clustering is a very famous, simple and powerful unsupervised machine learning algorithm.
- Has been applied to many complex unsupervised machine learning problems.
- *A K-means clustering algorithm tries to group similar items in the form of clusters.*
  - *The number of groups is represented by K. It should be given beforehand.*
  - *K should be predetermined*

K-MEANS CLUSTERING

- Each cluster is associated with a centroid(center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The objective is to minimize the sum of distances of the points to their respective centroid

K-MEANS CLUSTERING

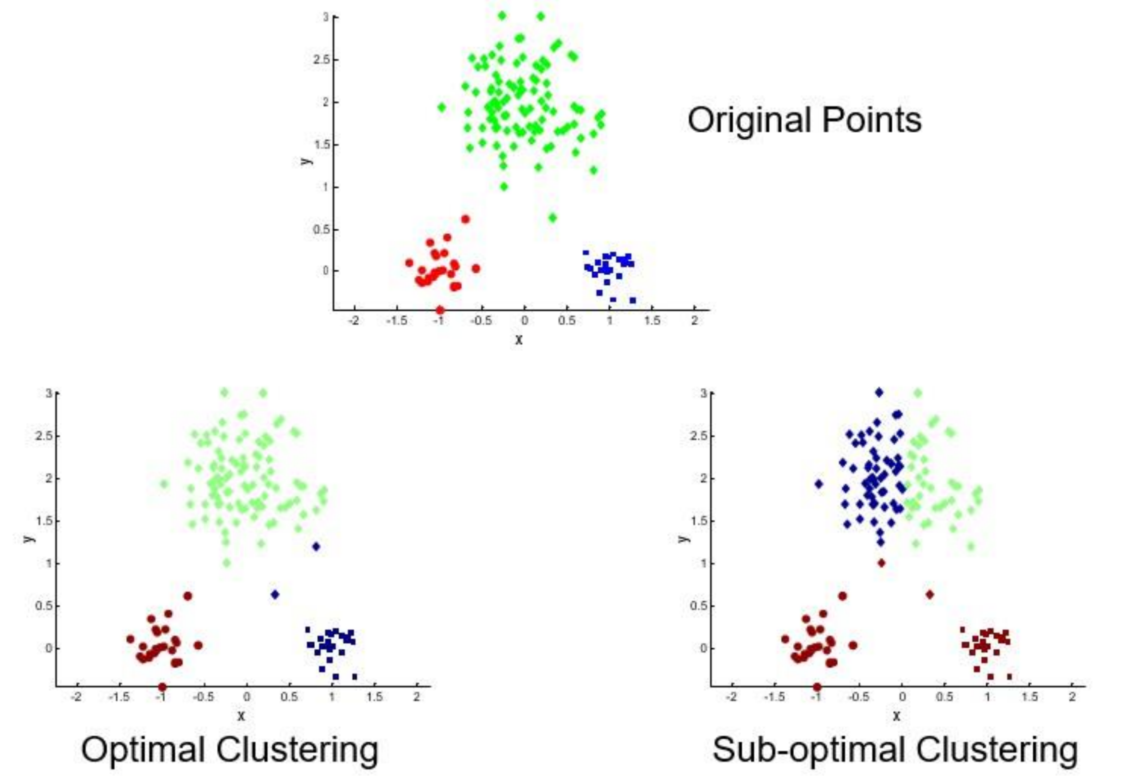
- **Problem:** Given a set X of n points in a d- dimensional space and an integer K. Group the points into K clusters C= {C1, C2,…,Ck} such that

$$\mathrm{Cost}(C) = \sum_{i=1} \sum_{x \in C_i} \mathrm{dist}(x, c)$$

- is minimized, where c_i is the centroid of the points in cluster C_i
- Most common definition is with Euclidean distance, minimizing the Sum of Squares Error (SSE) function

K-MEANS CLUSTERING

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.



K-MEANS CLUSTERING

Algorithm 1 (K-means clustering)

```
begin initialize n, c, μ_1, μ_2, ..., μ_c
        do classify n samples according to nearest μ_i
            recompute μ_i
        until no change in μ_i
    return μ_1, μ_2, ..., μ_c
end
```

K-MEANS CLUSTERING

- **Initialization of cluster centers**
- Do multiple runs and select the clustering with the smallest error
- Select the original set of points by methods other than random .
- E.g., pick the most distant (from each other) points as cluster centers (K-means++algorithm)

K-MEANS CLUSTERING

- The centroid depends on the distance function
  - The minimizer for the distance function
- Generally, we sue Euclidean or Minkowski, cosine similarity for clustering.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-MEANS CLUSTERING

- New Centroid:

$$Centroid(k_i) = \frac{1}{N_i} \sum_{j=1 \in N_i} x_j$$

- N_i is number of samples in i^{th} cluster

K-MEANS CLUSTERING

➢ Data Preprocessing

Step	Status
Data is numeric	✅
Scaled/Standardized data	✅
Missing values handled	✅
Outliers managed	✅
Optimal `k` determined	✅
Features selected/reduced	✅

EXAMPLE

- Suppose that the data mining task is to cluster points into three clusters,
- where the points are
- $A1(2, 10)$, $A2(2, 5)$, $A3(8, 4)$, $B1(5, 8)$, $B2(7, 5)$, $B3(6, 4)$, $C1(1, 2)$, $C2(4, 9)$.
- The distance function is Euclidean distance.
- Suppose initially we assign $A1$, $B1$, and $C1$ as the center of each cluster, respectively.

EXAMPLE

Initial Centroids:

A1: (2, 10)

B1: (5, 8)

C1: (1, 2)

Data Points			Distance to			Cluster	New Cluster
			2 10	5 8	1 2		
A1	2	10					
A2	2	5					
A3	8	4					
B1	5	8					
B2	7	5					
B3	6	4					
C1	1	2					
C2	4	9					

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

EXAMPLE

Initial Centroids:

A1: (2, 10)

B1: (5, 8)

C1: (1, 2)

Data Points			Distance to			Cluster	New Cluster
			2 10	5 8	1 2		
A1	2	10	0.00	3.61	8.06		
A2	2	5	5.00	4.24	3.16		
A3	8	4	8.49	5.00	7.28		
B1	5	8	3.61	0.00	7.21		
B2	7	5	7.07	3.61	6.71		
B3	6	4	7.21	4.12	5.39		
C1	1	2	8.06	7.21	0.00		
C2	4	9	2.24	1.41	7.62		

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

EXAMPLE

Initial Centroids:

A1: (2, 10)

B1: (5, 8)

C1: (1, 2)

Data Points			Distance to						Cluster	New Cluster
			2	10	5	8	1	2		
A1	2	10	0.00		3.61		8.06		1	
A2	2	5	5.00		4.24		3.16		3	
A3	8	4	8.49		5.00		7.28		2	
B1	5	8	3.61		0.00		7.21		2	
B2	7	5	7.07		3.61		6.71		2	
B3	6	4	7.21		4.12		5.39		2	
C1	1	2	8.06		7.21		0.00		3	
C2	4	9	2.24		1.41		7.62		2	

EXAMPLE

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			2	10	6	6	1.5	1.5		
A1	2	10							1	
A2	2	5							3	
A3	8	4							2	
B1	5	8							2	
B2	7	5							2	
B3	6	4							2	
C1	1	2							3	
C2	4	9							2	

EXAMPLE

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			2	10	6	6	1.5	1.5		
A1	2	10	0.00		5.66		6.52		1	1
A2	2	5	5.00		4.12		1.58		3	3
A3	8	4	8.49		2.83		6.52		2	2
B1	5	8	3.61		2.24		5.70		2	2
B2	7	5	7.07		1.41		5.70		2	2
B3	6	4	7.21		2.00		4.53		2	2
C1	1	2	8.06		6.40		1.58		3	3
C2	4	9	2.24		3.61		6.04		2	1

EXAMPLE

Current Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
A1	2	10				1	
A2	2	5				3	
A3	8	4				2	
B1	5	8				2	
B2	7	5				2	
B3	6	4				2	
C1	1	2				3	
C2	4	9				1	

EXAMPLE

Current Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3	9.5	6.5	5.25	1.5	3.5		
A1	2	10	1.12		6.54		6.52		1	
A2	2	5	4.61		4.51		1.58		3	
A3	8	4	7.43		1.95		6.52		2	
B1	5	8	2.50		3.13		5.70		2	
B2	7	5	6.02		0.56		5.70		2	
B3	6	4	6.26		1.35		4.53		2	
C1	1	2	7.76		6.39		1.58		3	
C2	4	9	1.12		4.51		6.04		1	

EXAMPLE

Current Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

New Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
			3, 9.5	6.5, 5.25	1.5, 3.5		
A1	2	10	1.12	6.54	6.52	1	1
A2	2	5	4.61	4.51	1.58	3	3
A3	8	4	7.43	1.95	6.52	2	2
B1	5	8	2.50	3.13	5.70	2	1
B2	7	5	6.02	0.56	5.70	2	2
B3	6	4	6.26	1.35	4.53	2	2
C1	1	2	7.76	6.39	1.58	3	3
C2	4	9	1.12	4.51	6.04	1	1

EXAMPLE

Current Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3.67	9	7	4.33	1.5	3.5		
A1	2	10	1.94		7.56		6.52		1	
A2	2	5	4.33		5.04		1.58		3	
A3	8	4	6.62		1.05		6.52		2	
B1	5	8	1.67		4.18		5.70		1	
B2	7	5	5.21		0.67		5.70		2	
B3	6	4	5.52		1.05		4.53		2	
C1	1	2	7.49		6.44		1.58		3	
C2	4	9	0.33		5.55		6.04		1	

EXAMPLE

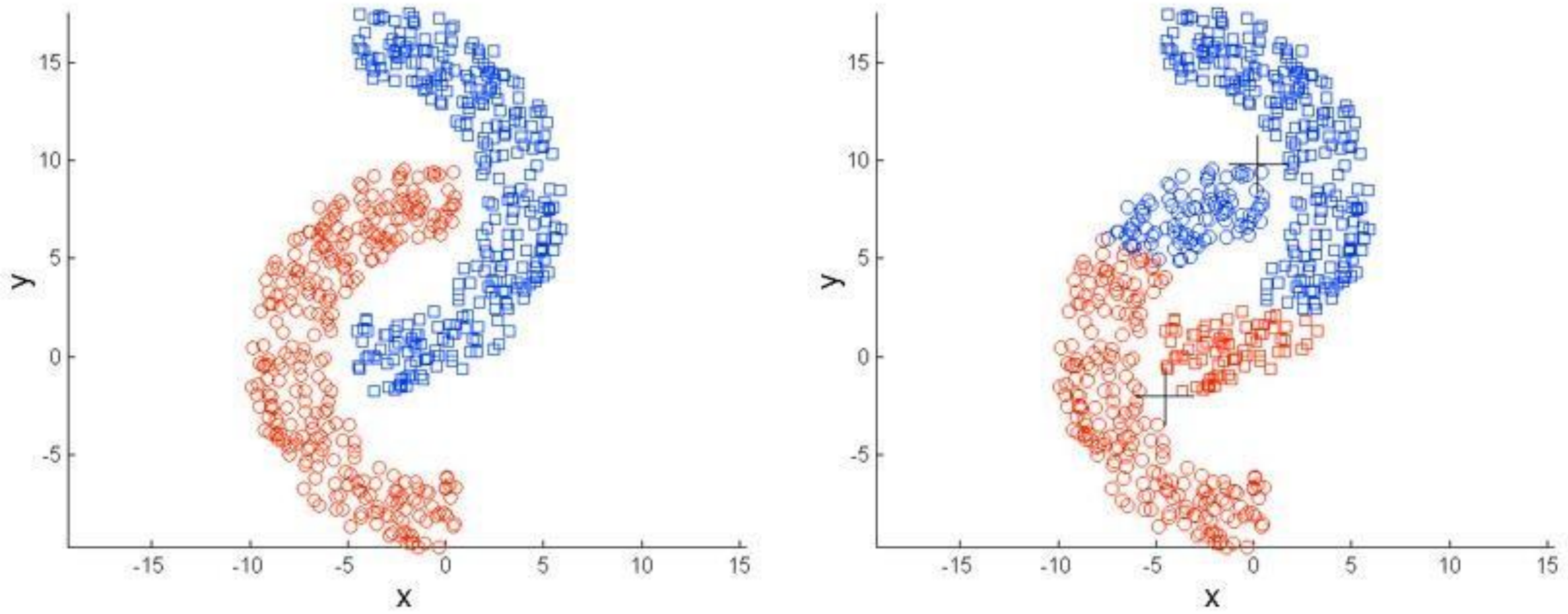
Current Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3.67	9	7	4.33	1.5	3.5		
A1	2	10	1.94		7.56		6.52		1	1
A2	2	5	4.33		5.04		1.58		3	3
A3	8	4	6.62		1.05		6.52		2	2
B1	5	8	1.67		4.18		5.70		1	1
B2	7	5	5.21		0.67		5.70		2	2
B3	6	4	5.52		1.05		4.53		2	2
C1	1	2	7.49		6.44		1.58		3	3
C2	4	9	0.33		5.55		6.04		1	1

DISADVANTAGES

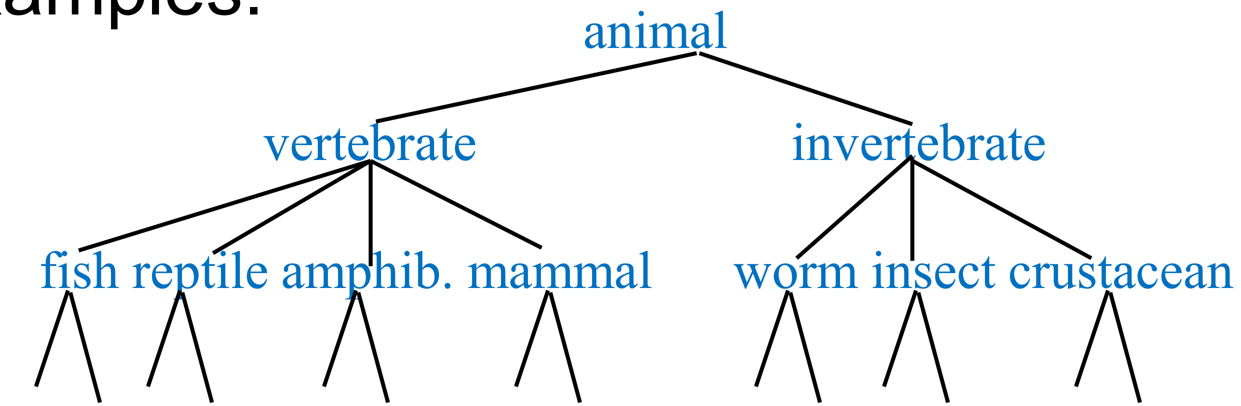
- K-means has problems when clusters are of different
  - Sizes
  - Densities
  - Complex shapes
- K-means has problems when the data contains outliers.

DISADVANTAGES



HIERARCHICAL CLUSTERING

Build a tree-based hierarchical taxonomy (*dendrogram*) from a set of unlabeled examples.



Recursive application of a standard clustering algorithm can produce hierarchical clustering.

HIERARCHICAL CLUSTERING

Hierarchical clustering is a method of cluster analysis that is used to cluster similar data points together.

Hierarchical clustering follows either the top-down or bottom-up method of clustering.

There are two types of hierarchical clustering methods:

1. Divisive Clustering
2. Agglomerative Clustering

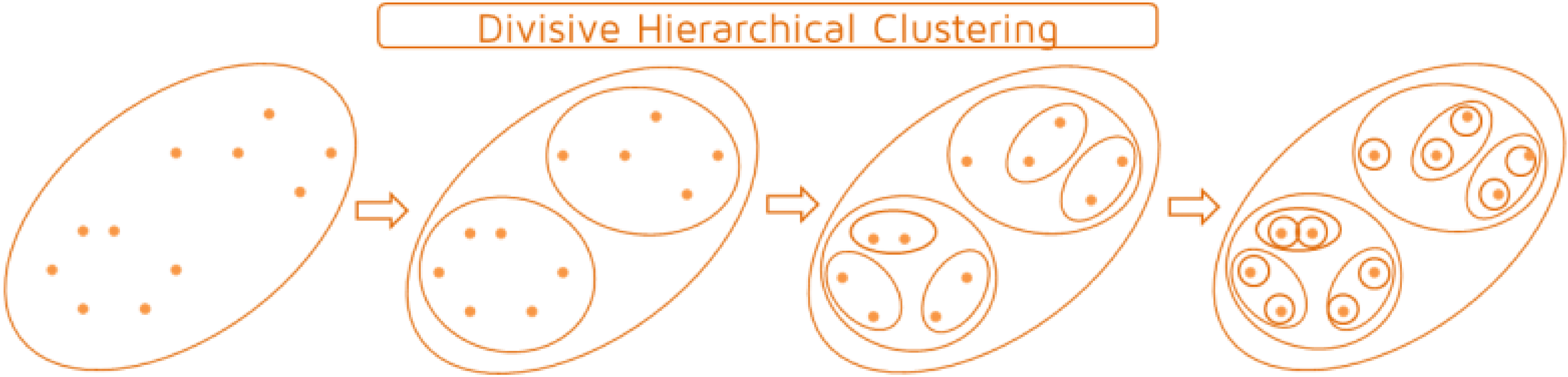
HIERARCHICAL CLUSTERING

Divisive Clustering

•**Also Known As**: Top-Down Clustering

•**Approach**:

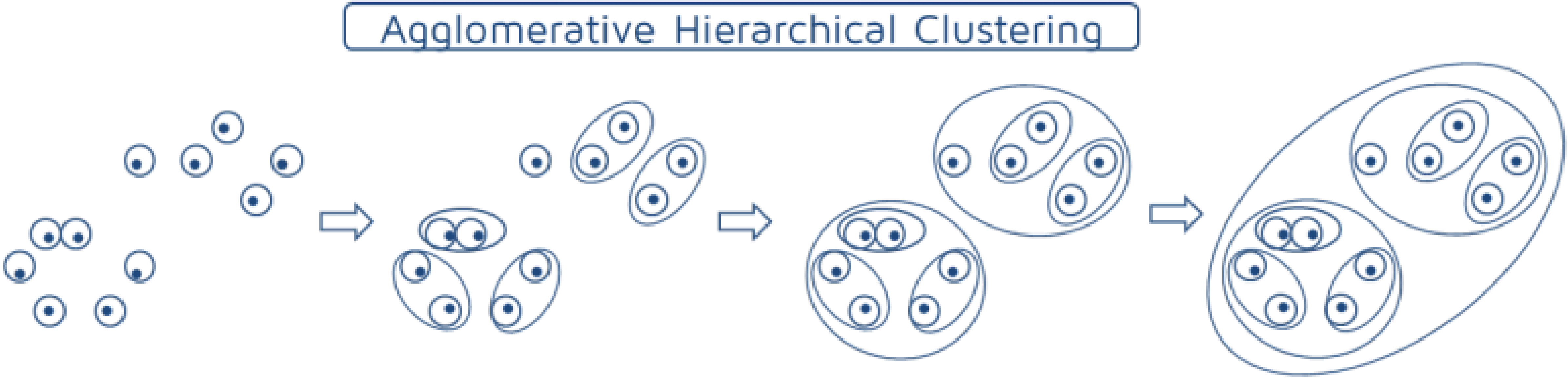
- Starts with **all data points in one single cluster**.
- Gradually splits the cluster into smaller sub-clusters.
- The process continues recursively until each data point becomes its own cluster or the desired number of clusters is reached.



HIERARCHICAL CLUSTERING

Agglomerative Clustering

- **Also Known As**: Bottom-Up Clustering
- **Approach**:
  - Starts with **each data point as its own cluster**.
  - Gradually merges the closest clusters until all points are in a single cluster or the desired number of clusters is reached.



HIERARCHICAL CLUSTERING

Comparison of Divisive and Agglomerative Clustering

Feature	Divisive Clustering	Agglomerative Clustering
Approach	Top-Down	Bottom-Up
Initial State	One large cluster	Individual data points
Merging/Splitting	Splits clusters	Merges clusters
Computational Cost	Higher (due to global splits)	Lower (local merges)
Common Usage	Rare	Widely used

HIERARCHICAL CLUSTERING

Applications of Hierarchical Clustering

1. **Bioinformatics**: Grouping genes or proteins with similar functions.

2. **Social Network Analysis**: Identifying communities in networks.

3. **Market Segmentation**: Clustering customers based on purchasing behavior.

4. **Document Clustering**: Grouping similar documents for topic modeling.

Thank You ☺