

GLOW 2024

UNSUPERVISED LEARNING

Clustering



Nur Laila Ab Ghani
Department of informatics
College of Computing and Informatics
Universiti Tenaga Nasional
Laila@uniten.edu.my

Adapted from Dr. Mahmud Dwi Sulistiyo Slides

GLOW 2024

Outline

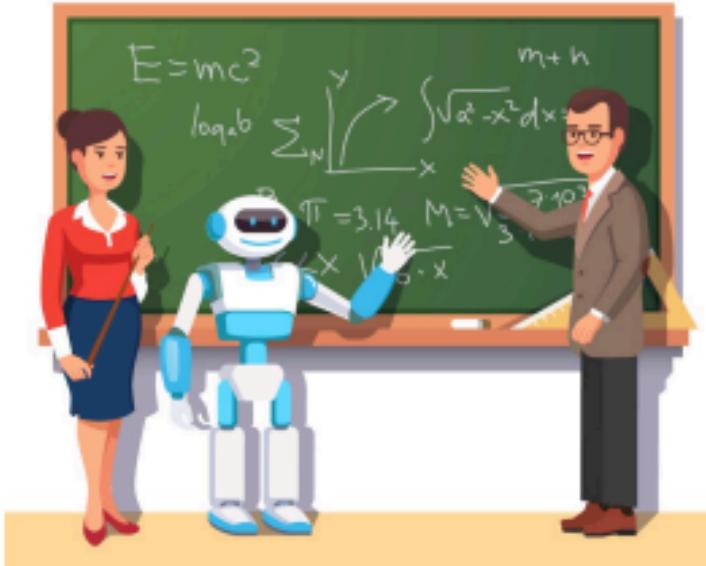
- Supervised vs Unsupervised Learning
- Clustering Methods: K-Means & Hierarchical Clustering
- Evaluating Clustering Results
- Implementing Clustering Methods

Adapted from Dr. Mahmud Dwi Sulistiyo Slides



Supervised Learning

(e.g. Regression, Classification)

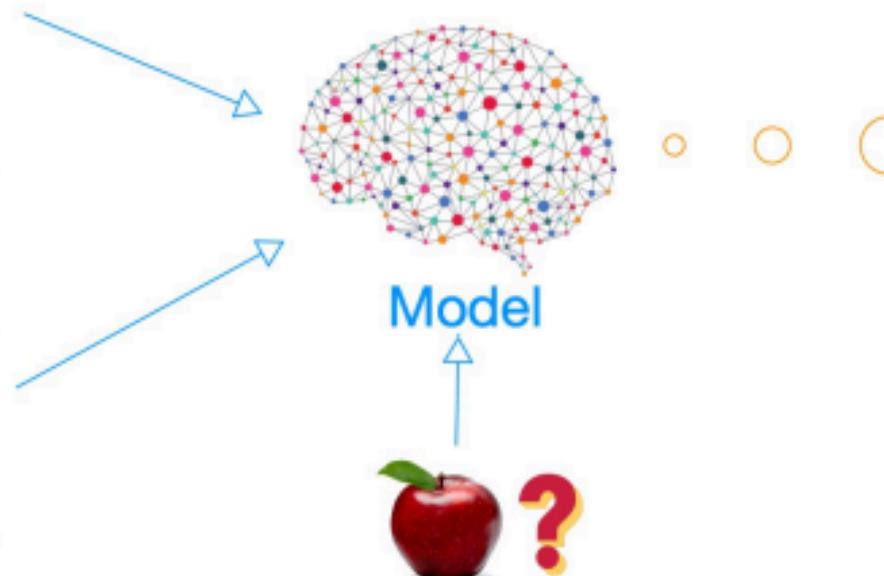


Input data



Annotations

These are
apples

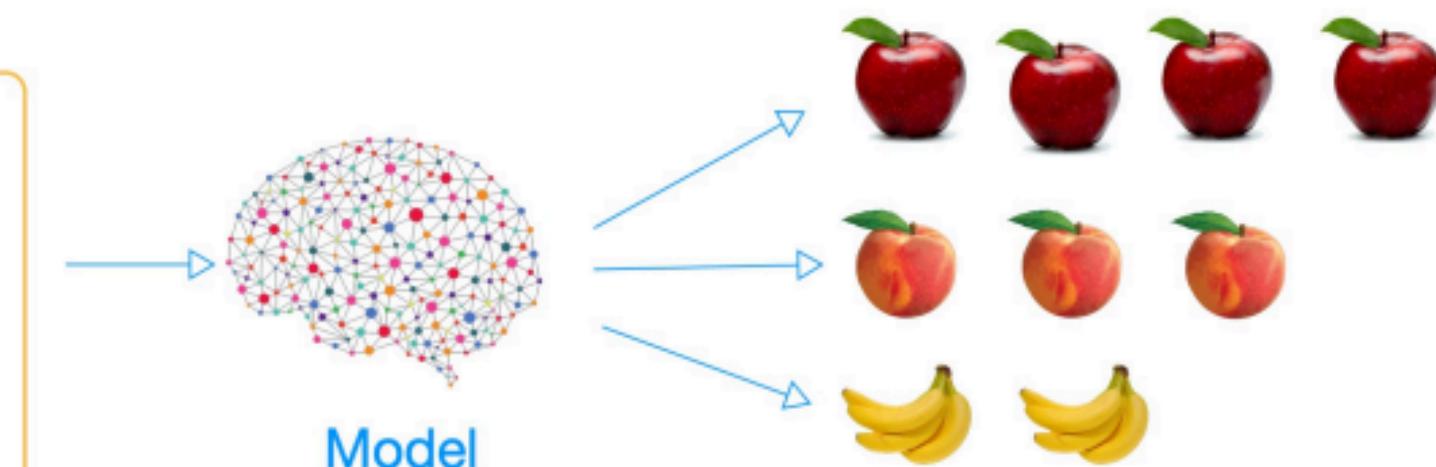
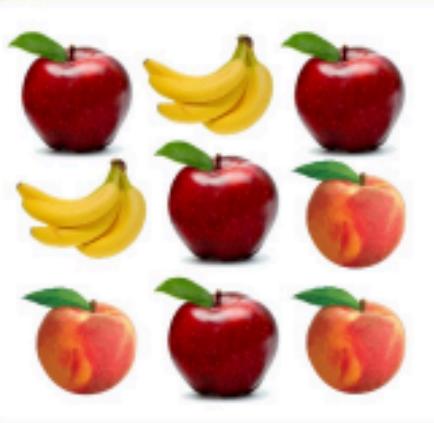


Unsupervised Learning

(e.g. Clustering)



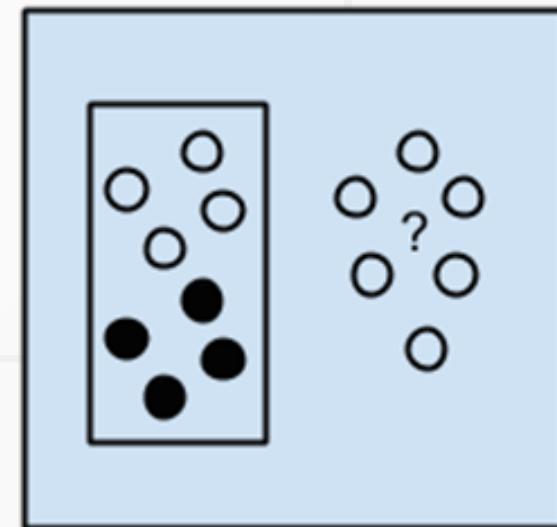
Input data



Supervised VS Unsupervised Learning

- **Supervised**

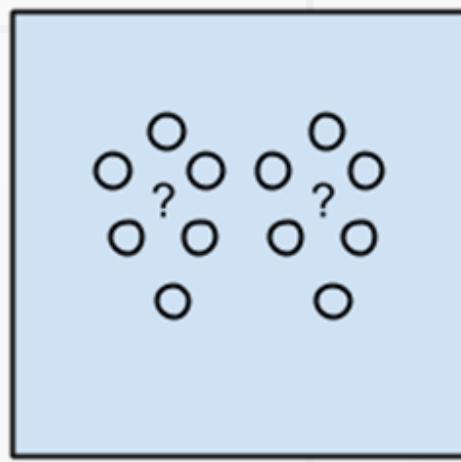
- You have labeled data
- Find a function which can map data to its label
- discover patterns that relate data attributes with a target (class)



Supervised Learning
Algorithms

- **Unsupervised**

- You have unlabeled data
- Discover the underlying structure of the data
- Try to understand the data
- Not predicting anything specific



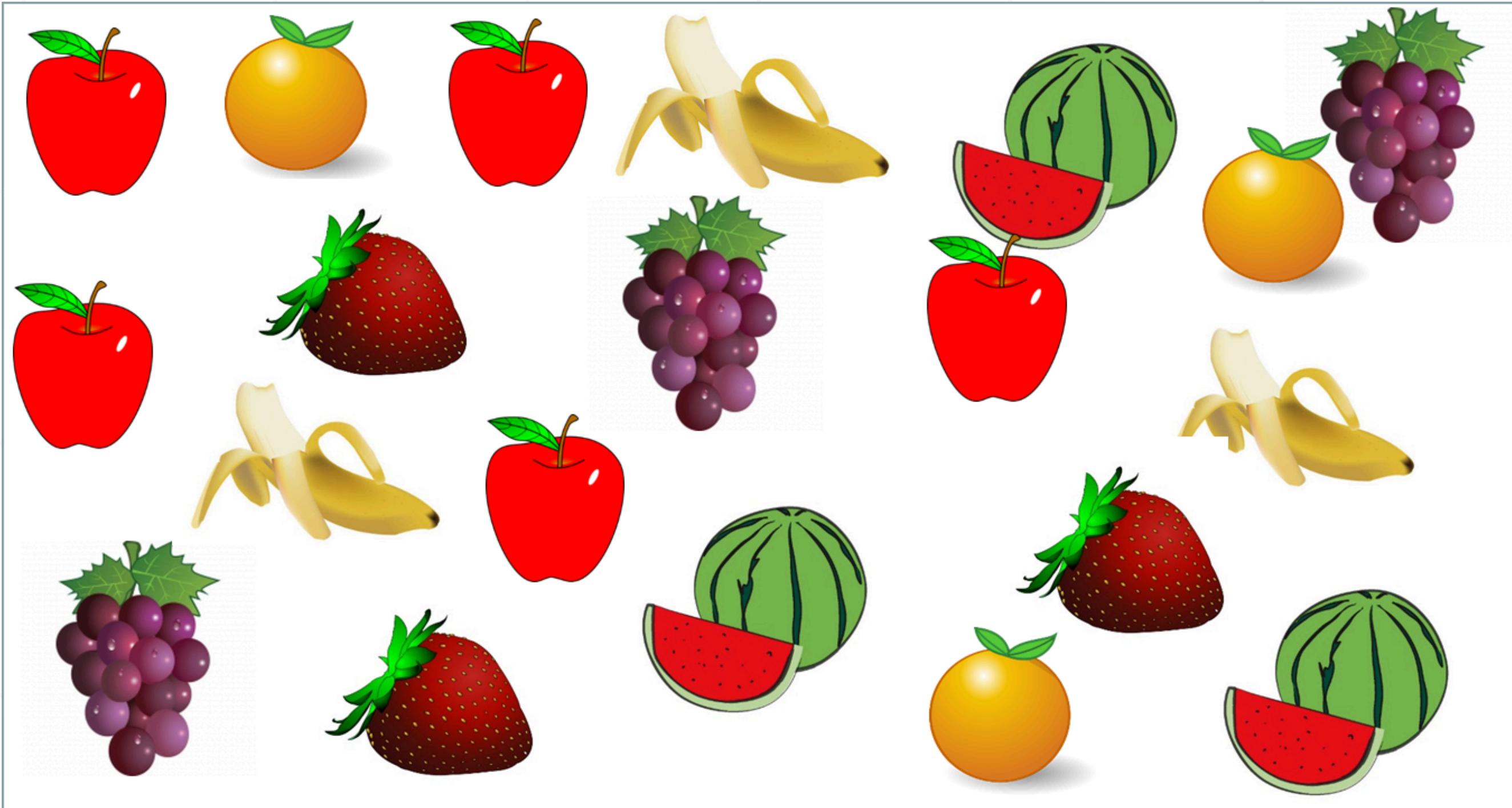
Unsupervised Learning
Algorithms

Clustering

- Grouping data into small groups based on **similarity** such that data in the **same group (cluster)** are as similar as possible and data in different groups are as different as possible.
- Help users **understand the natural grouping or structure in a data set**. Used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms.

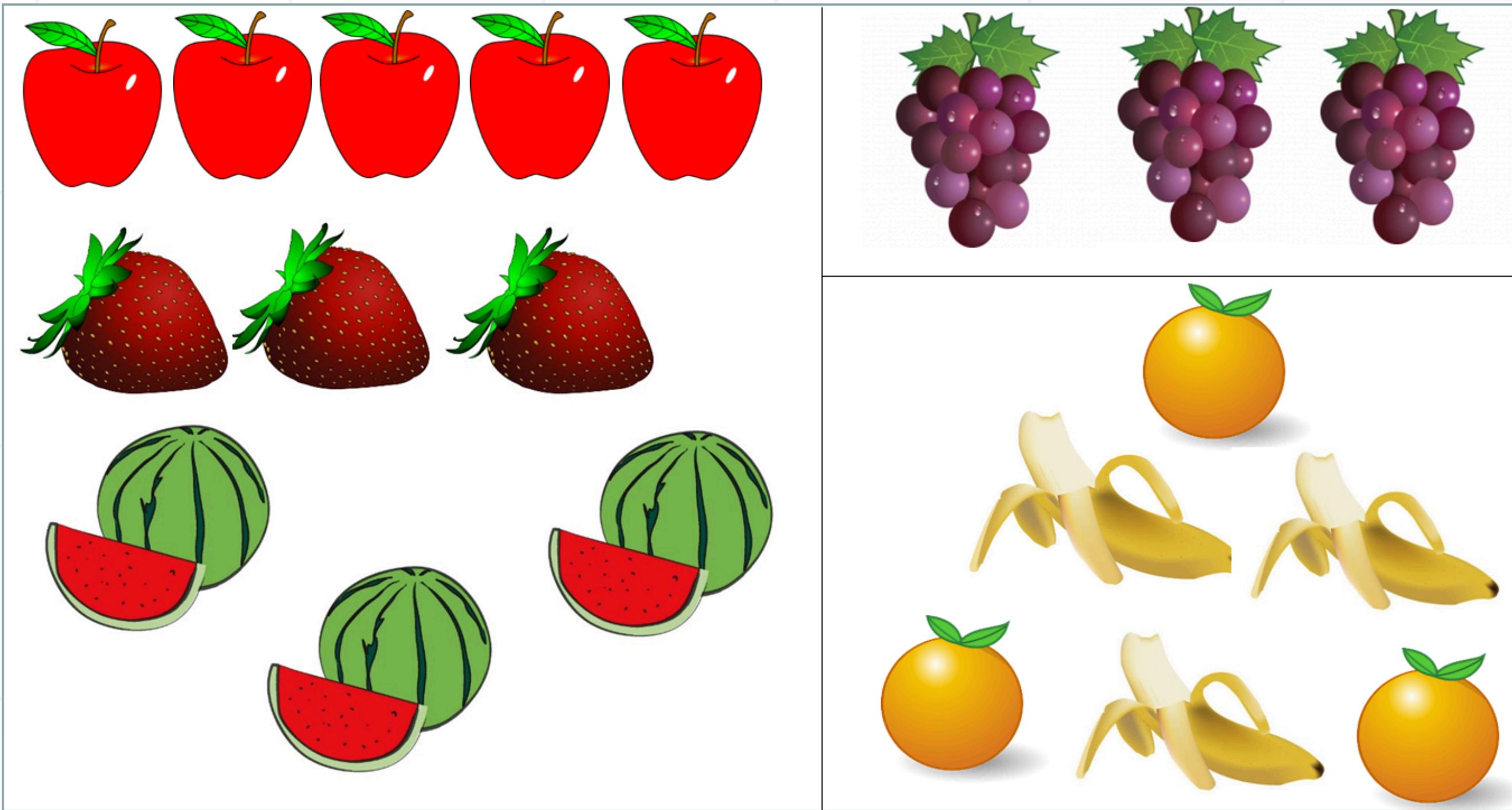
Clustering

- Example: How do we want to group these fruits?



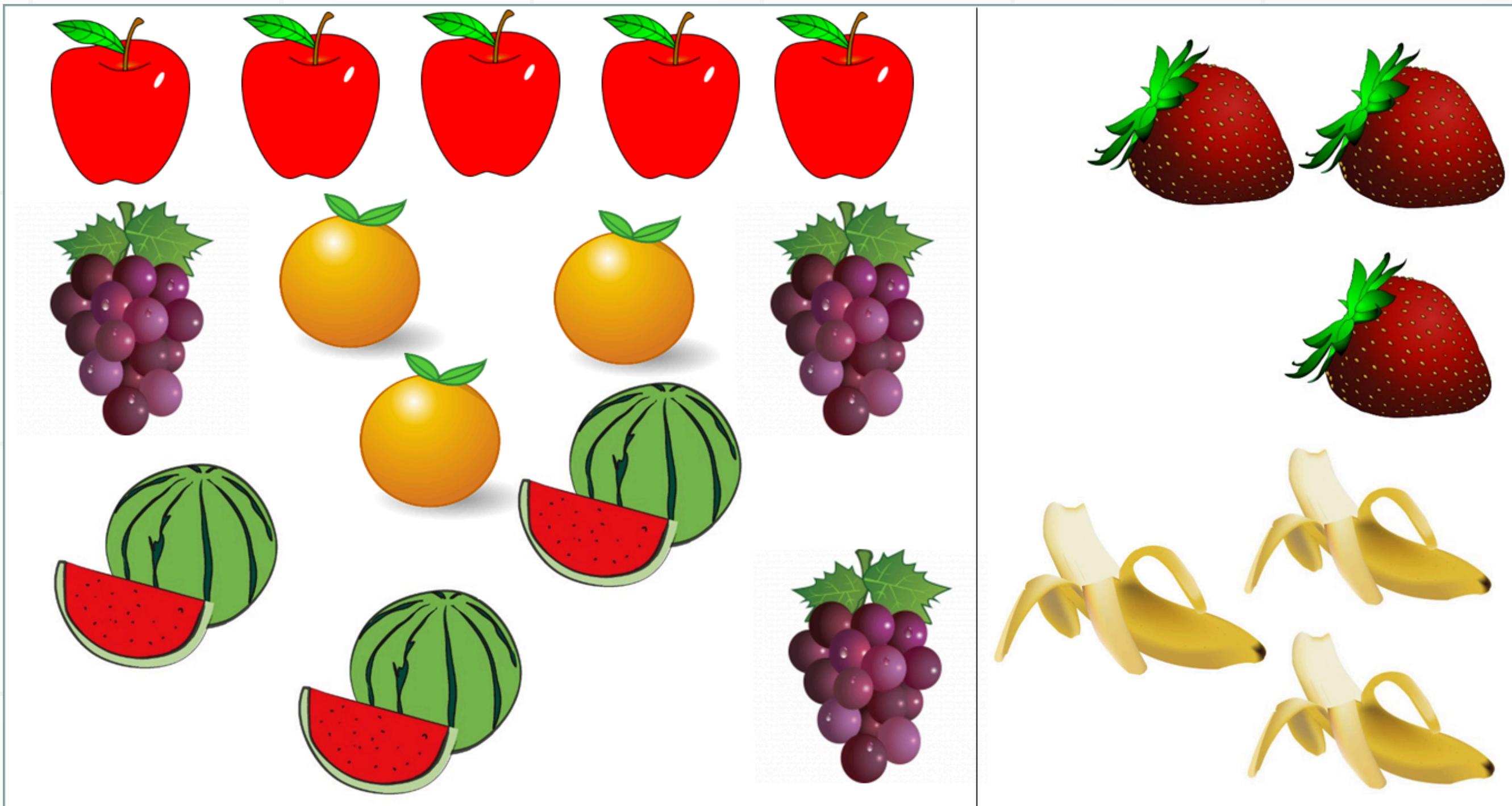
Clustering

- a) Grouping based on colour



Clustering

- a) Grouping based on shape



Clustering

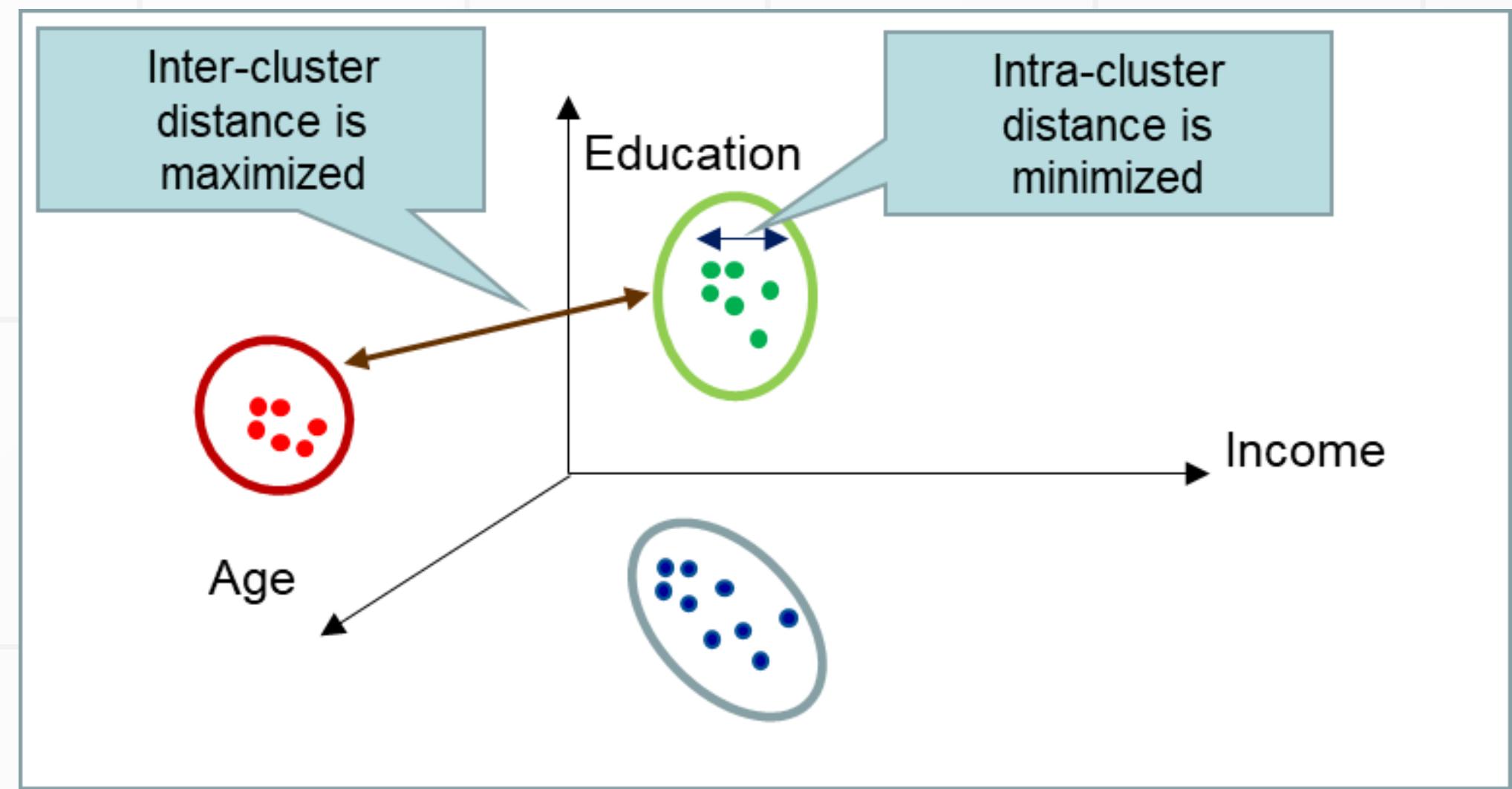
- Any other ways/approaches to group these fruits?
 - Based on size
 - Based on weight
 - Based on locality/origin

Clustering

- Other examples:
 - Cluster customers based on their purchase histories, so that a targeted marketing program can be developed
 - Cluster products based on the sets of customers who purchased them
 - Cluster documents based on similar words
 - Cluster DNA sequences based on edit distance

Clustering

- Good clustering method will produce high quality clusters with:
 - High intra-cluster similarity
 - Low inter-cluster similarity



Clustering

- The quality of a clustering result depends on
 - the similarity measure used
 - implementation of the similarity measure
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

Steps to Perform Clustering

Step 1:

Formulate the problem – Decide on the clustering variable

Step 2:

Decide on the clustering procedure

Step 3:

Decide on the number of clusters

Step 4:

Validate and Interpret cluster solution

Steps to Perform Clustering

Step 1: Formulate the problem – Decide on the clustering variable

- The objective of this step is:
 - To select variables that could provide a clear-cut **differentiation between segments/groups** regarding a specific managerial objective
- Types and examples of clustering variables:

	General	Specific
Observable (directly measurable)	Cultural, geographic, demographic, socio-economic	User status, usage frequency, store and brand loyalty
Unobservable (inferred)	Psychographics, values, personality, lifestyle	Benefits, perceptions, attitudes, intentions, preferences

Steps to Perform Clustering

Step 2:
Decide on the clustering procedure

Option 1:
Hierarchical methods

Option 2:
Partitioning methods

Steps to Perform Clustering

1. Hierarchical methods

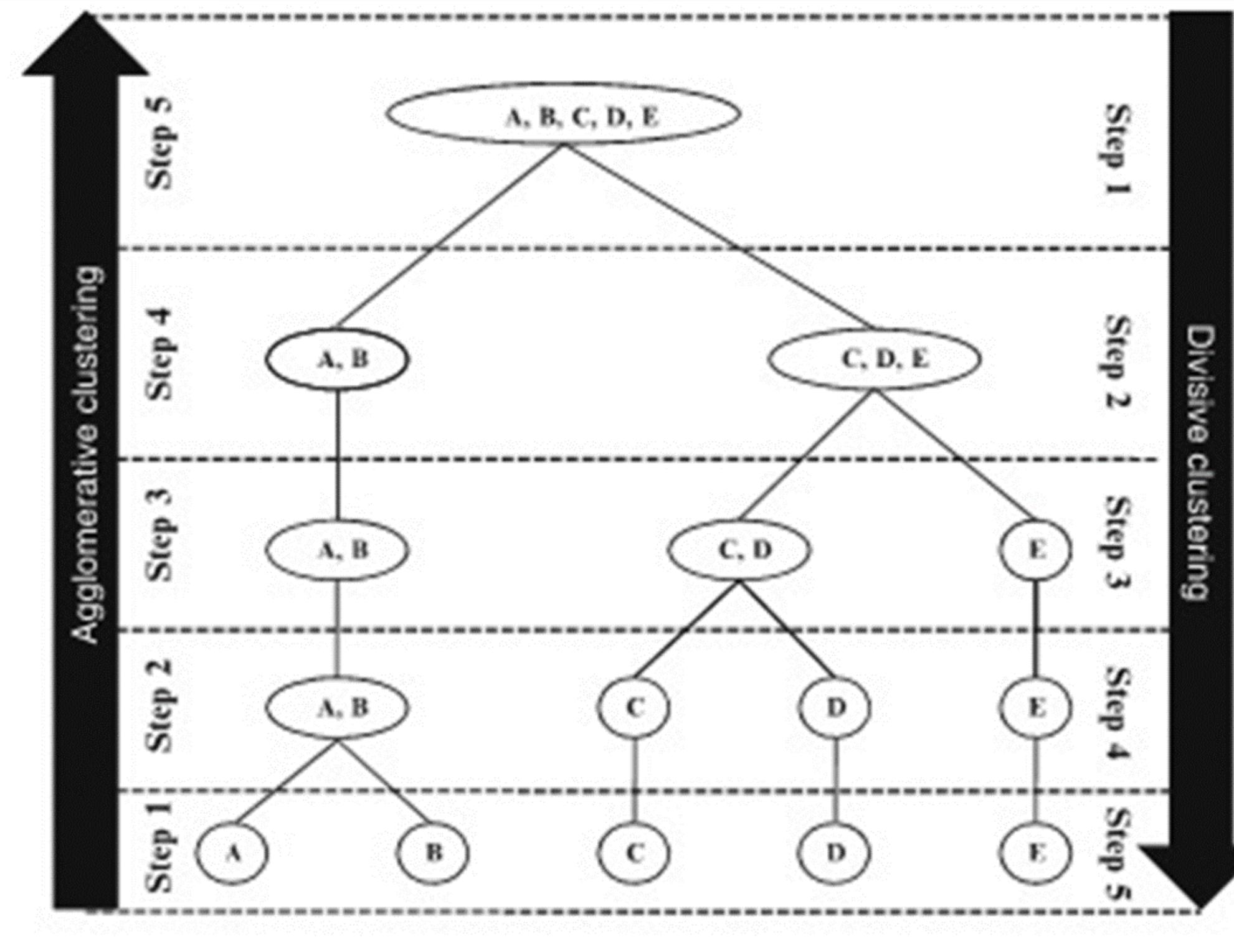
Agglomerative

Each object starts with their own separate cluster. Then, two ‘closest’ (most similar) cluster is combined and repeatedly performed until all objects become in one cluster

Divisive

All objects start in the same cluster, and gradually split up until each object becomes in individual cluster

Steps to Perform Clustering



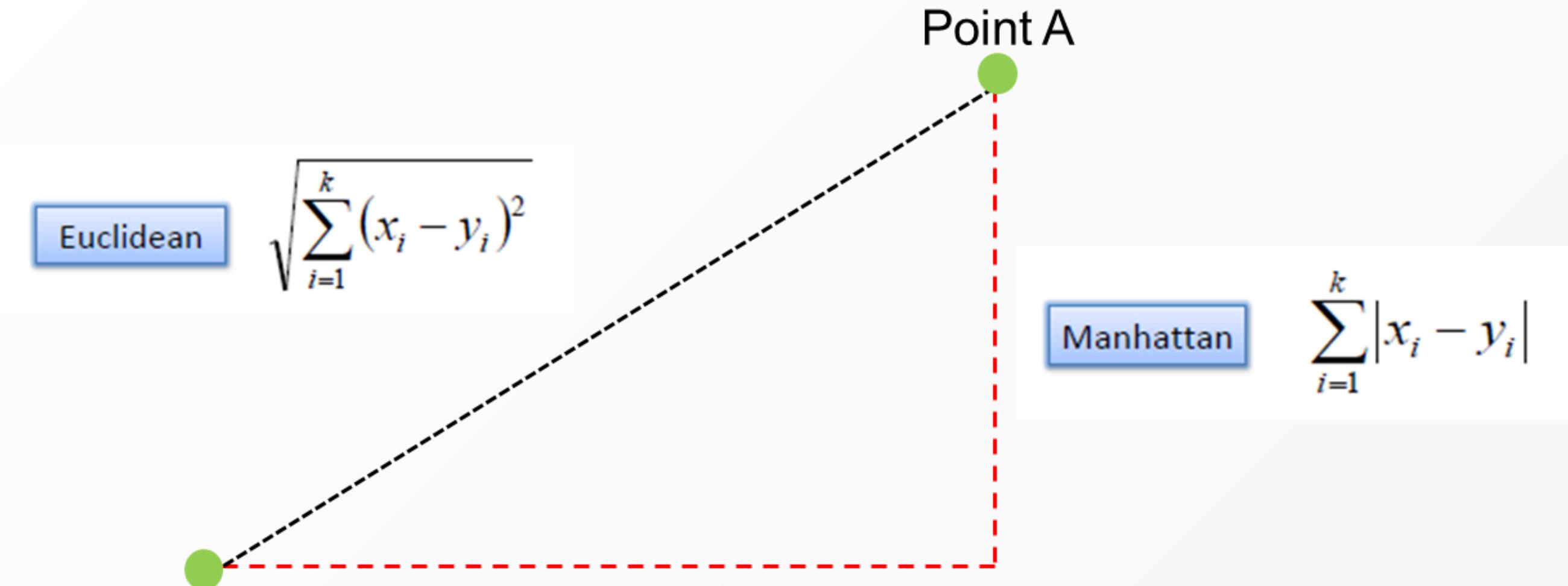
Adapted from Dr. Mahmud Dwi Sulistiyo Slides

Steps to Perform Clustering

- Steps in **hierarchical methods**:

1. Measure of similarity/dissimilarity

- Euclidean distance
- Manhattan distance



Steps to Perform Clustering

- Let's do some exercises:

Calculate the Euclidean distance for the following data sets:

$$A1 = (2,10)$$

$$A2 = (2,5)$$

$$A3 = (8,4)$$

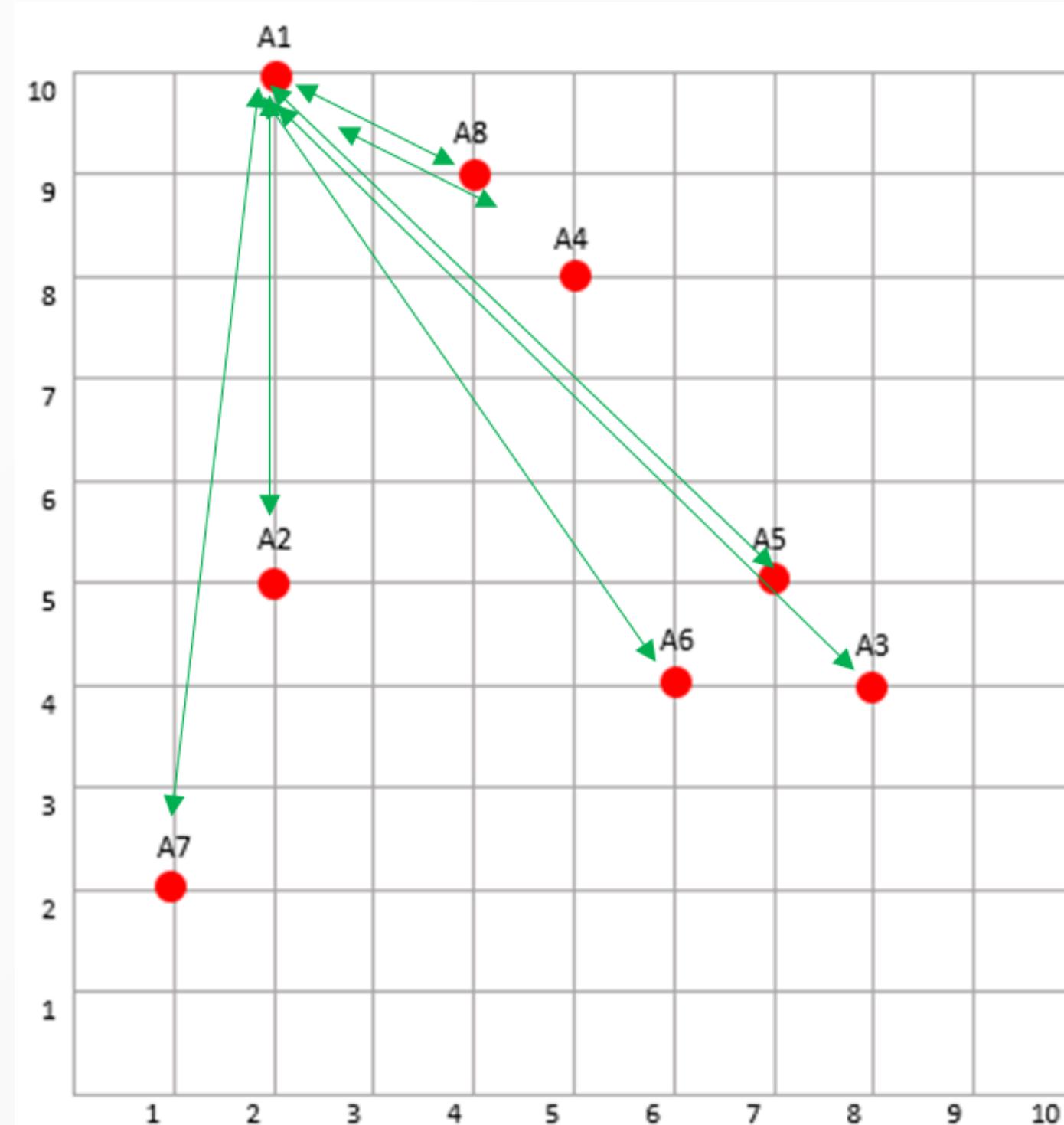
$$A4 = (5,8)$$

$$A5 = (7,5)$$

$$A6 = (6,4)$$

$$A7 = (1,2)$$

$$A8 = (4,9)$$



Steps to Perform Clustering

- Euclidean distance matrix

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	5	6	3.6	7.1	7.2	8.1	2.2
A2		0	6.1	4.2	5	4.1	3.1	4.5
A3			0	5	1.4	1.4	7.2	6.4
A4				0	3.6	4.1	7.2	1.4
A5					0	1.4	6.7	5
A6						0	5.4	5.4
A7							0	7.6
A8								0

Steps to Perform Clustering

2. Choose clustering algorithm

- **Single linkage**

- The distance between two objects is defined to be the smallest distance possible between them.
- If both objects are clusters, the distance between the two closest members are used.

- **Complete linkage**

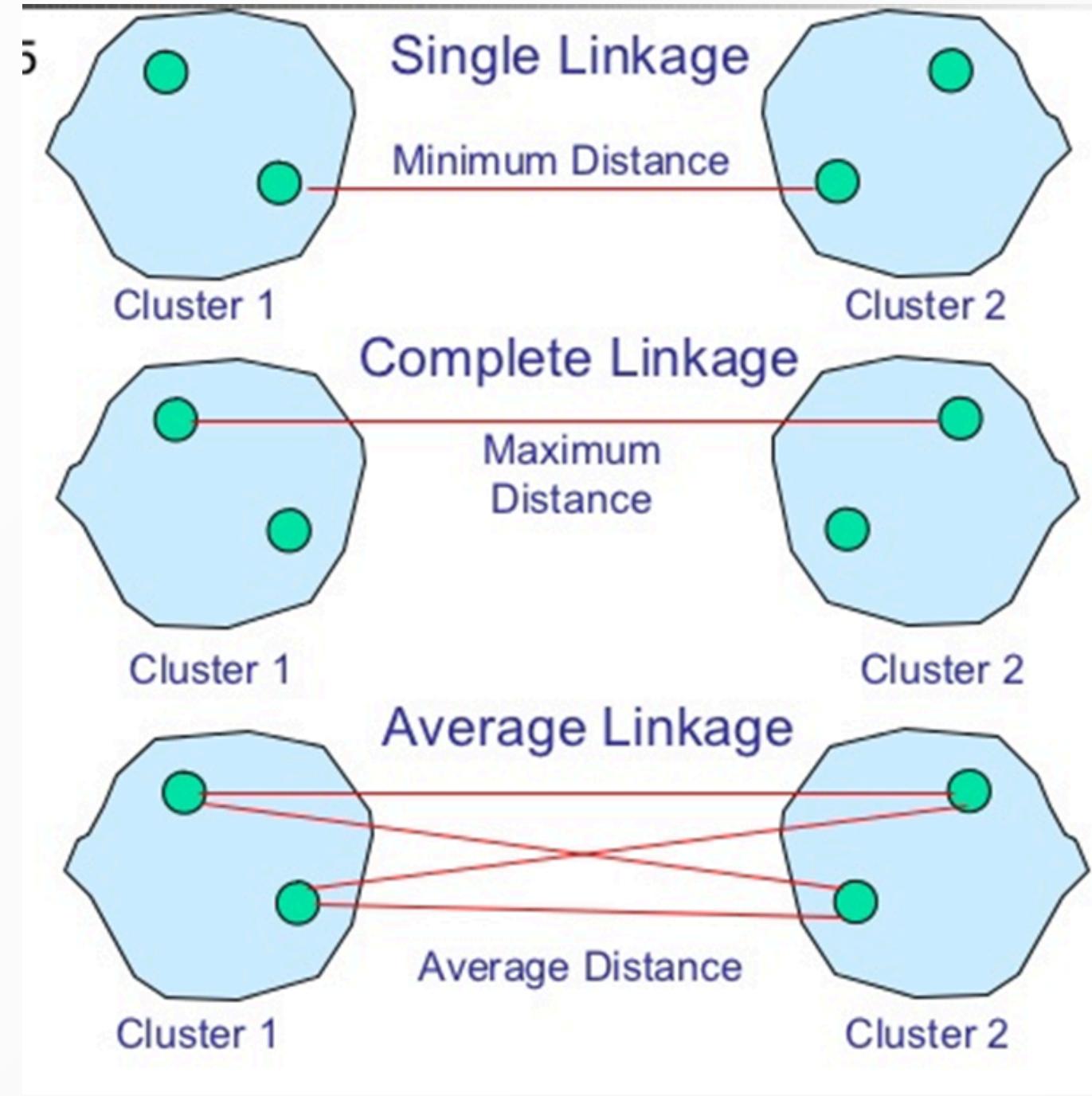
- This method is much like the single linkage, but instead of using the minimum of the distances, we use the maximum distance

- **Average linkage**

- The distance between two clusters is defined as the average distance between all pairs of the two clusters' members

Steps to Perform Clustering

2. Choose clustering algorithm



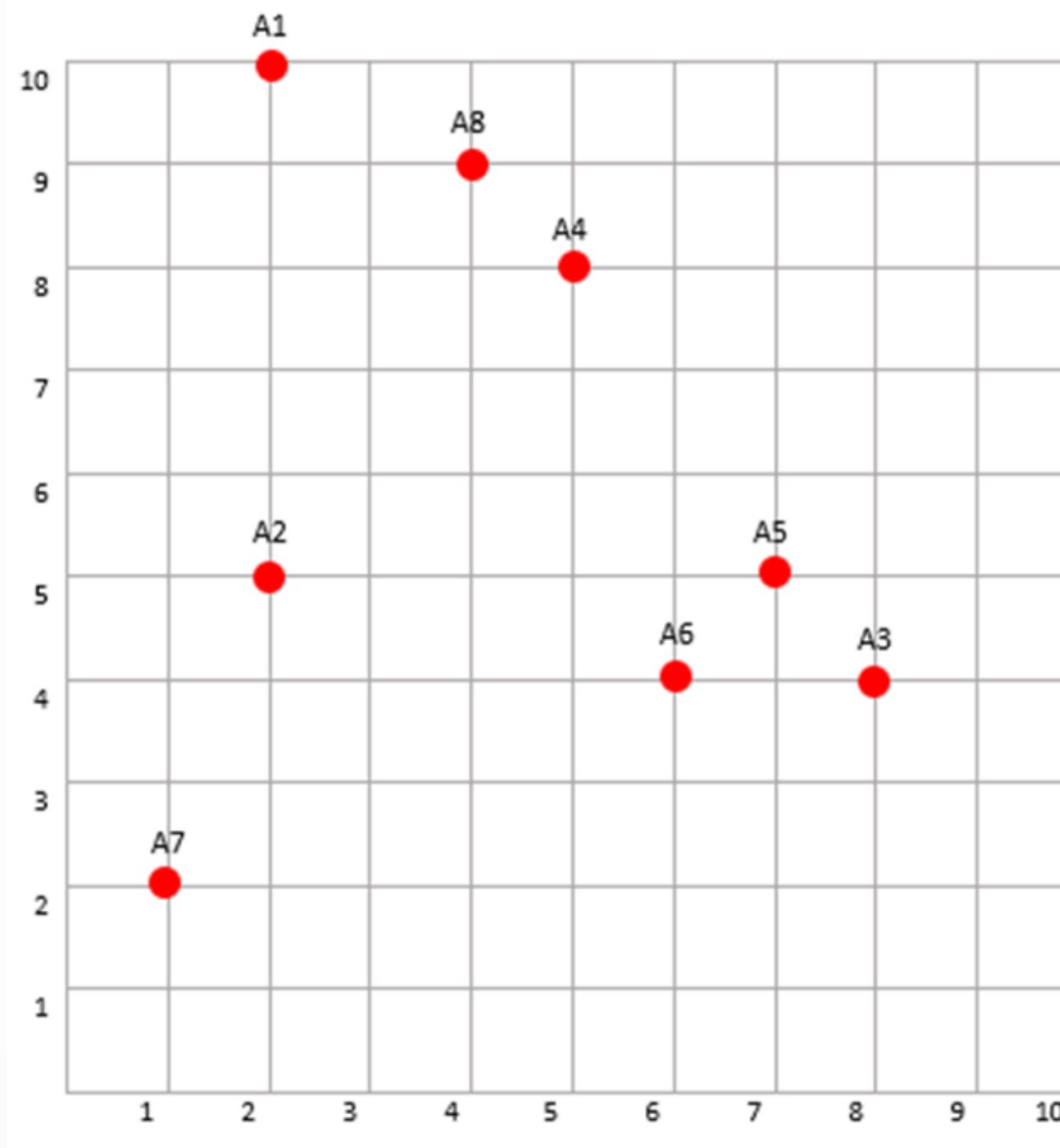
Steps to Perform Clustering

2. Perform hierarchical clustering using single link algorithms for the sample datasets

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	5	6	3.6	7.1	7.2	8.1	2.2
A2		0	6.1	4.2	5	4.1	3.1	4.5
A3			0	5	1.4	1.4	7.2	6.4
A4				0	3.6	4.1	7.2	1.4
A5					0	1.4	6.7	5
A6						0	5.4	5.4
A7							0	7.6
A8								0

Steps to Perform Clustering

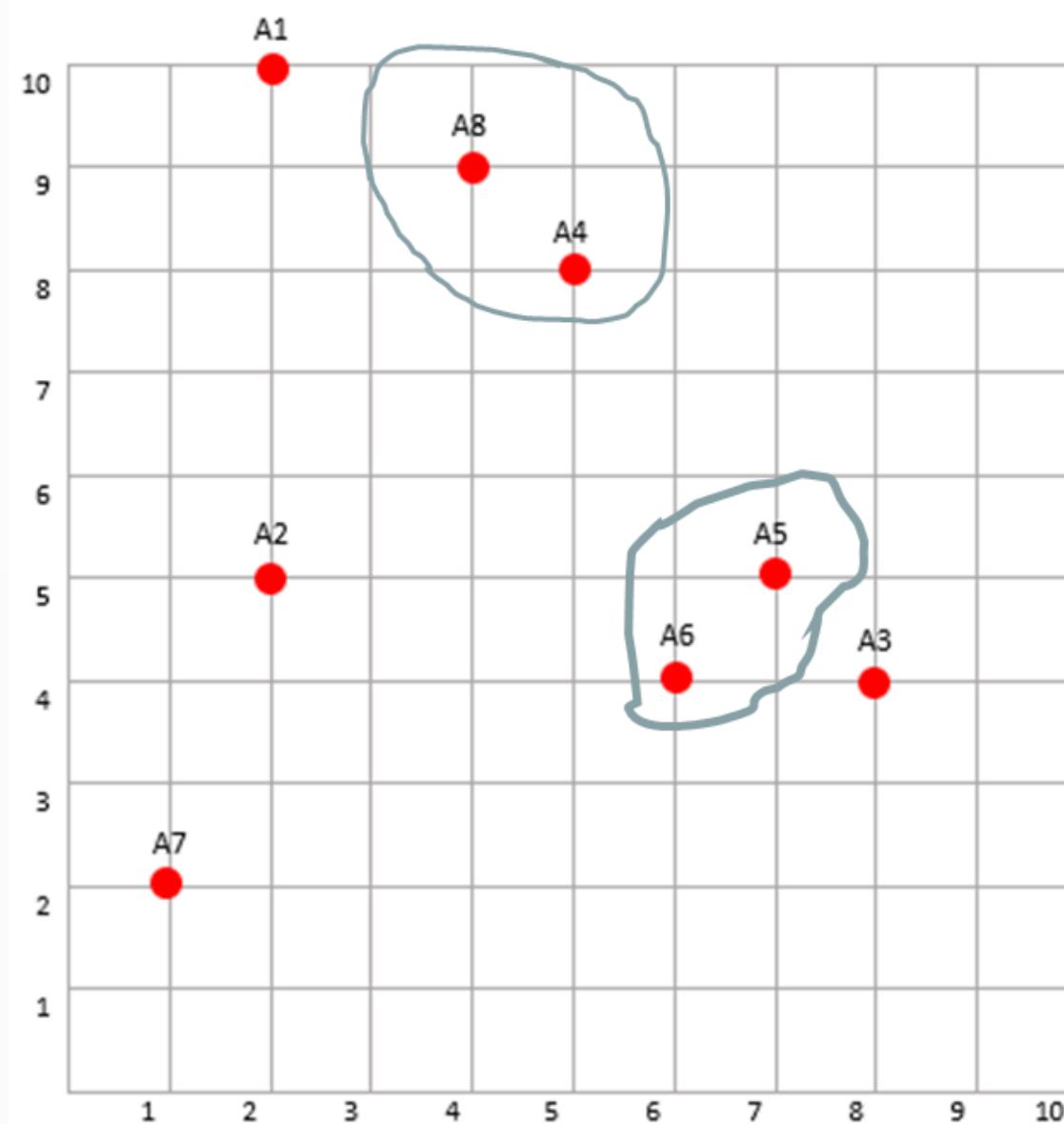
Initial state – 8 clusters



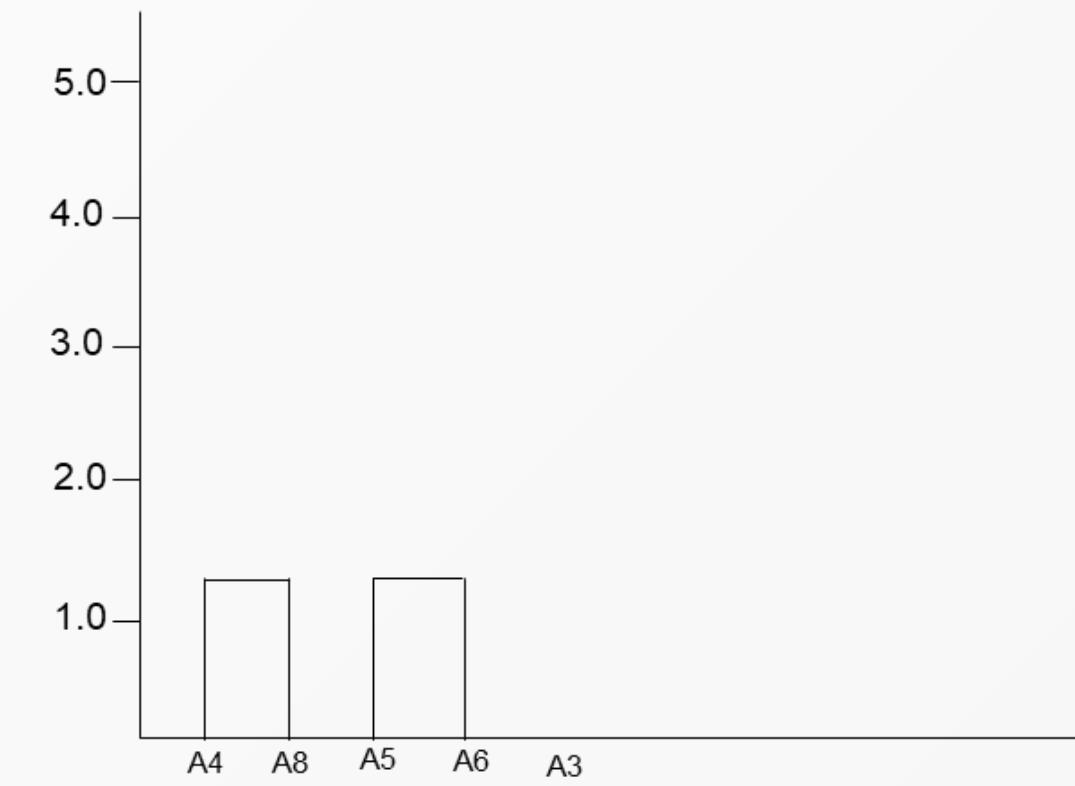
	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	5	6	3.6	7.1	7.2	8.1	2.2
A2		0	6.1	4.2	5	4.1	3.1	4.5
A3			0	5	1.4	1.4	7.2	6.4
A4				0	3.6	4.1	7.2	1.4
A5					0	1.4	6.7	5
A6						0	5.4	5.4
A7							0	7.6
A8								0

Steps to Perform Clustering

1st clustering cycle

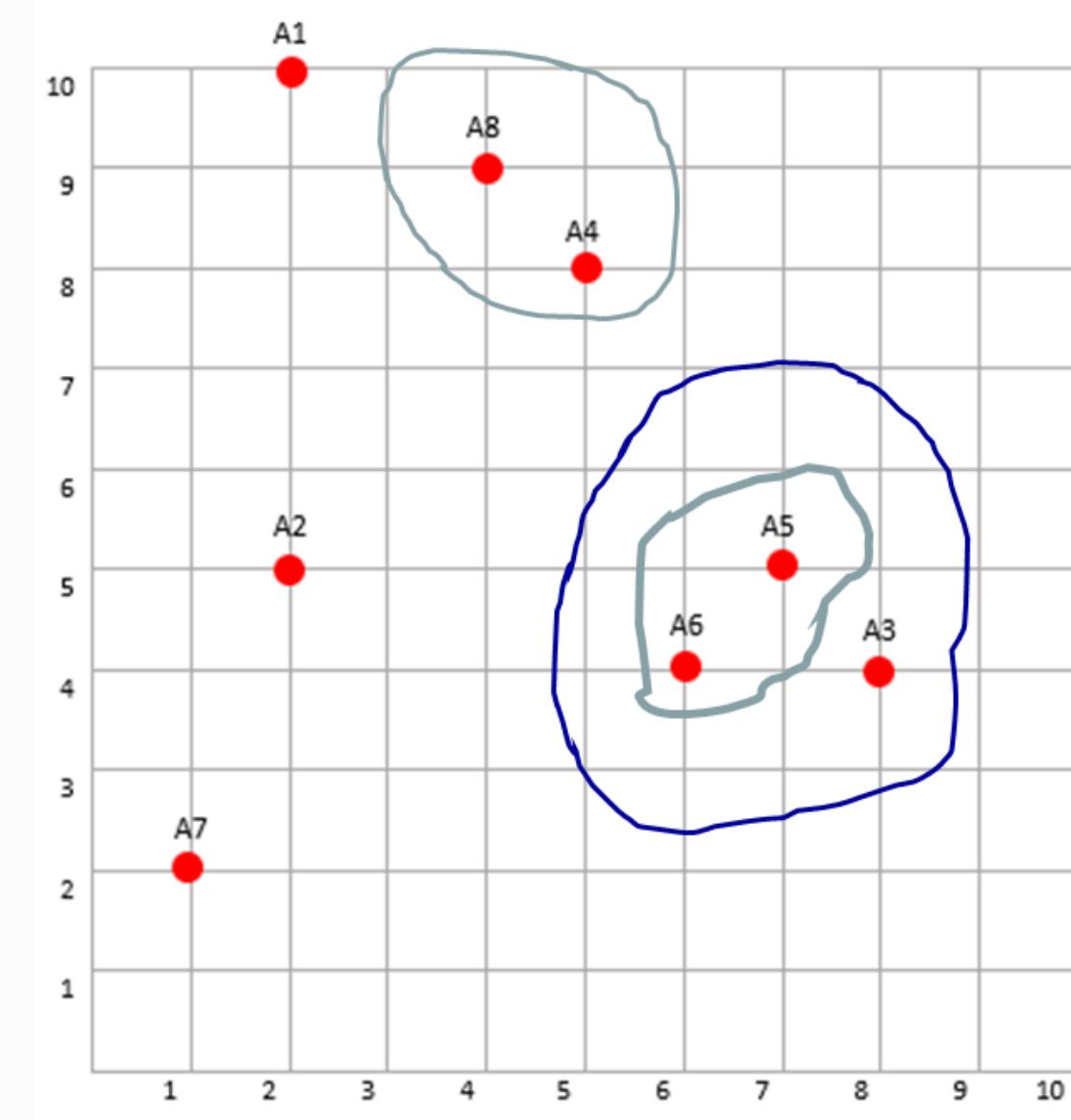


	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	5	6	3.6	7.1	7.2	8.1	2.2
A2		0	6.1	4.2	5	4.1	3.1	4.5
A3			0	5	1.4	1.4	7.2	6.4
A4				0	3.6	4.1	7.2	1.4
A5					0	1.4	6.7	5
A6						0	5.4	5.4
A7							0	7.6
A8								0

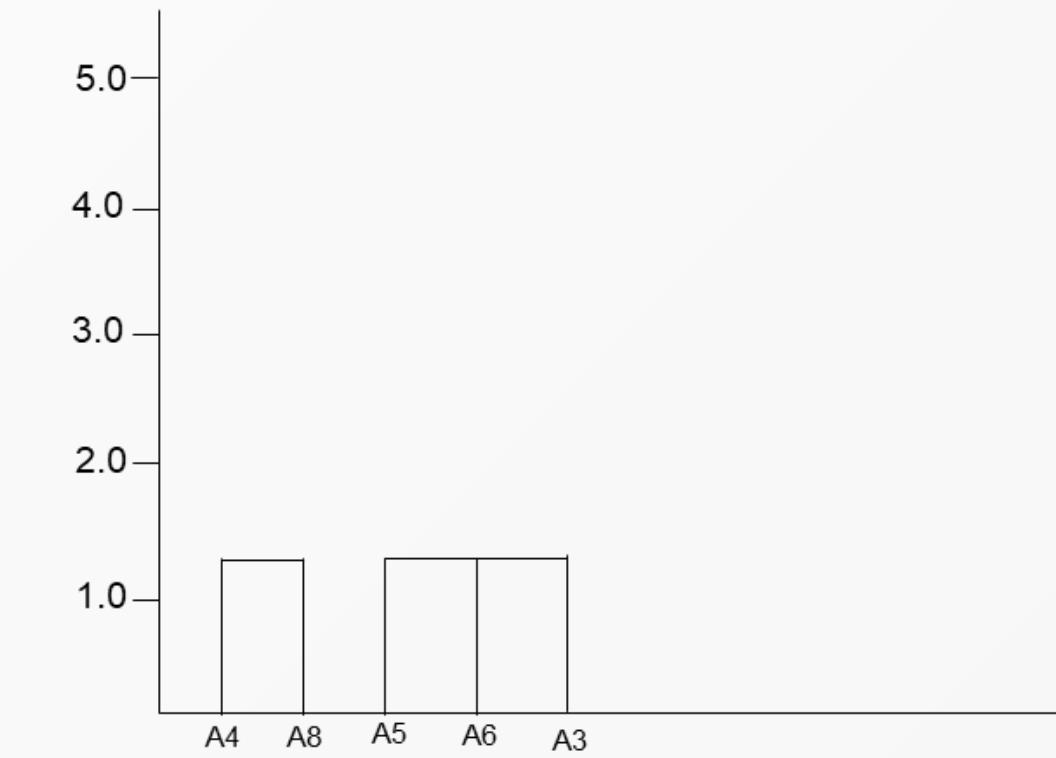


Steps to Perform Clustering

2nd clustering cycle

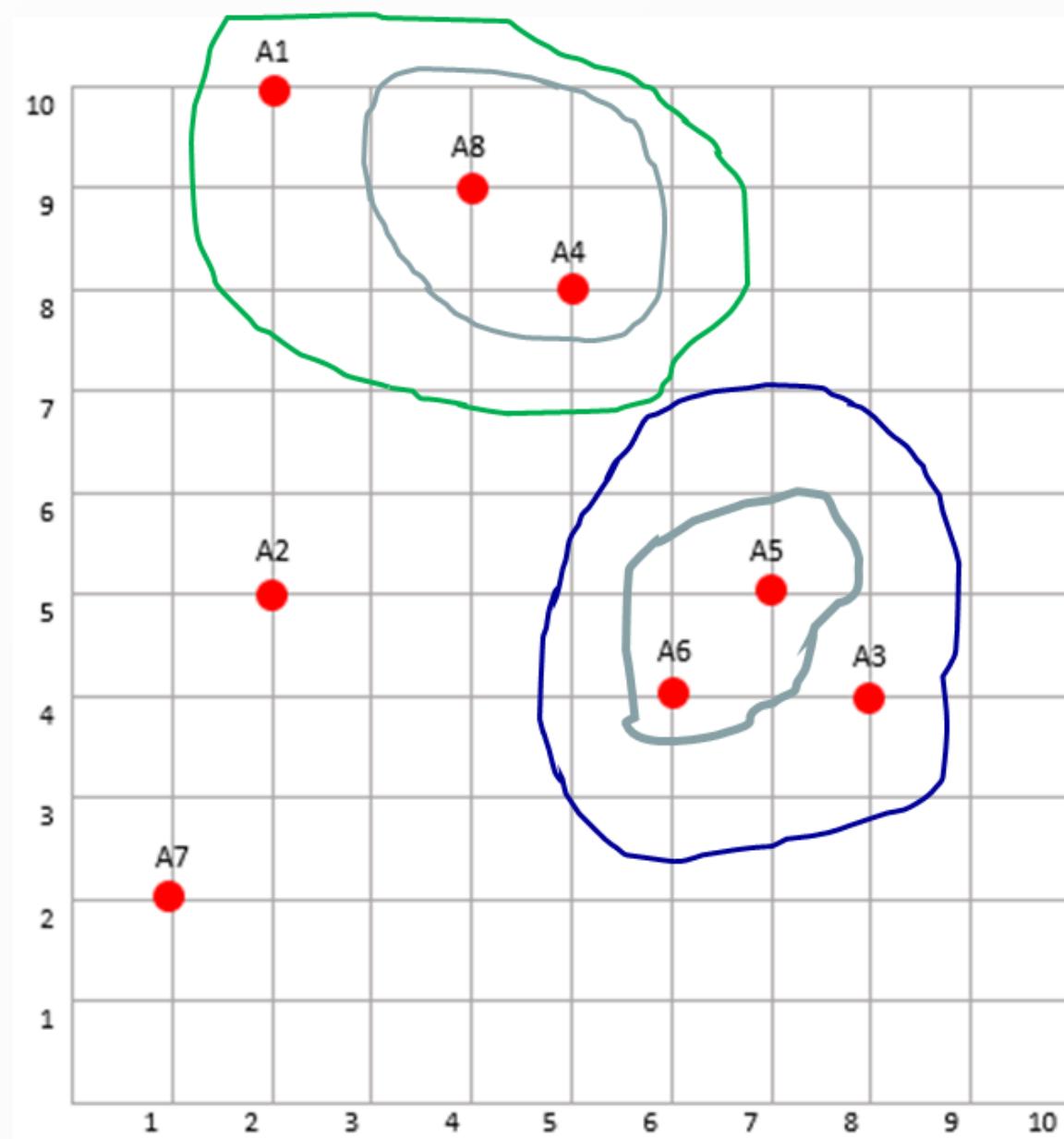


	A1	A2	A3	A4, A8	A5, A6	A7
A1	0	5	6	2.2	7.1	8.1
A2		0	5	4.2	4.1	3.1
A3			0	5	1.4	7.2
A4, A8				0	3.6	7.2
A5, A6					0	5.4
A7						0

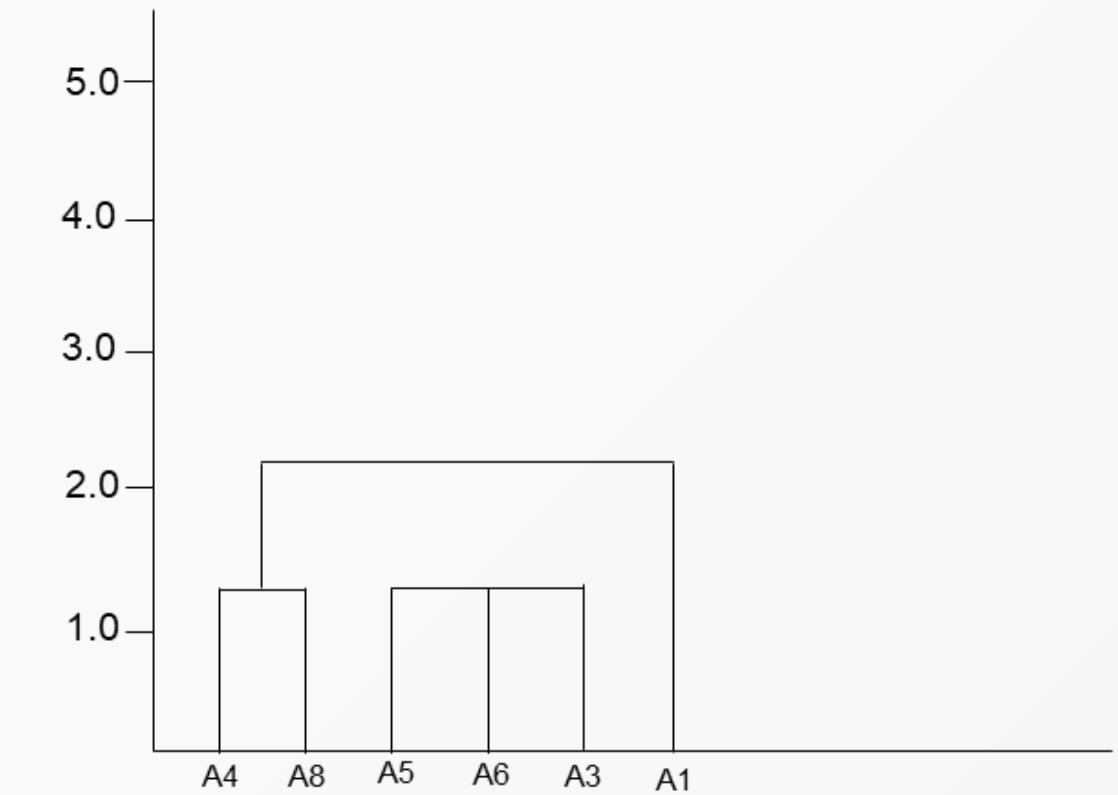


Steps to Perform Clustering

3rd clustering cycle

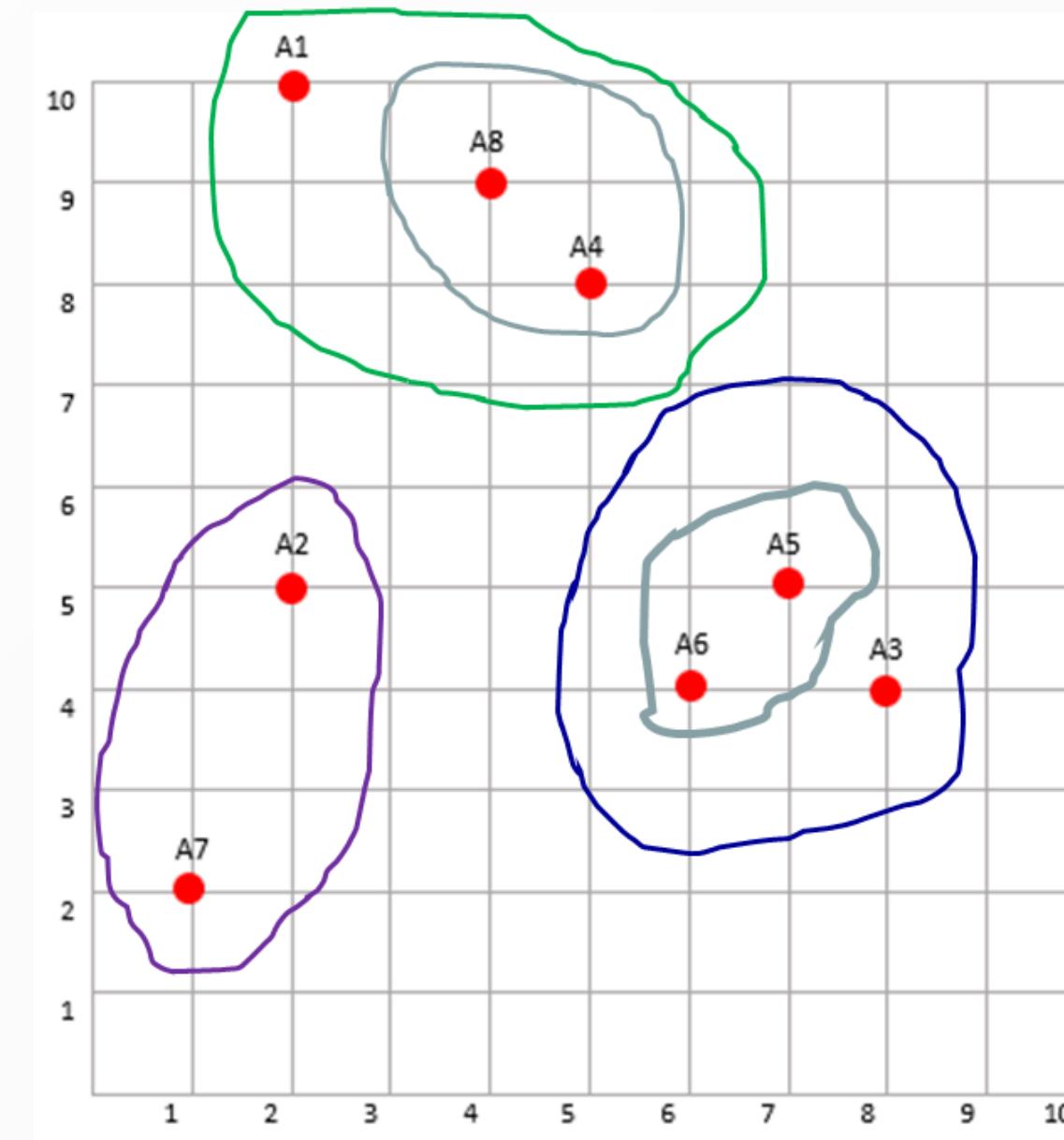


	A1	A2	A4, A8	A3, A5, A6	A7
A1	0	5	2.2	6	8.1
A2		0	4.2	4.1	3.1
A4, A8			0	3.6	7.2
A3, A5, A6				0	5.4
A7					0

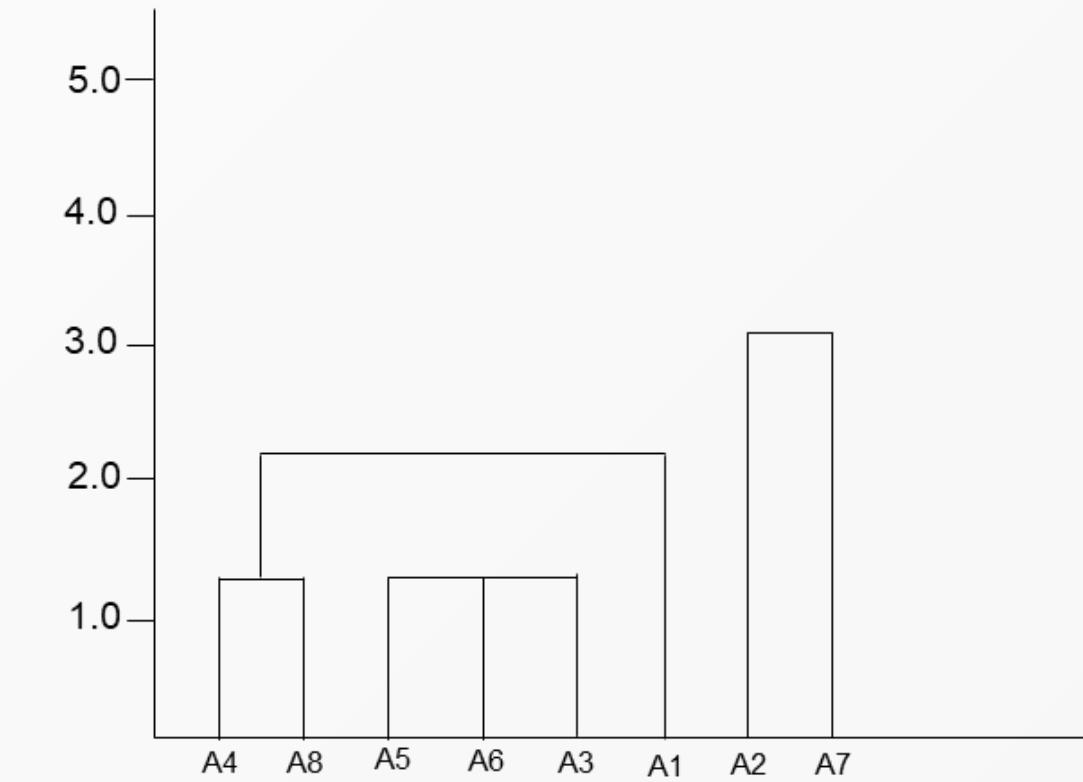


Steps to Perform Clustering

4th clustering cycle

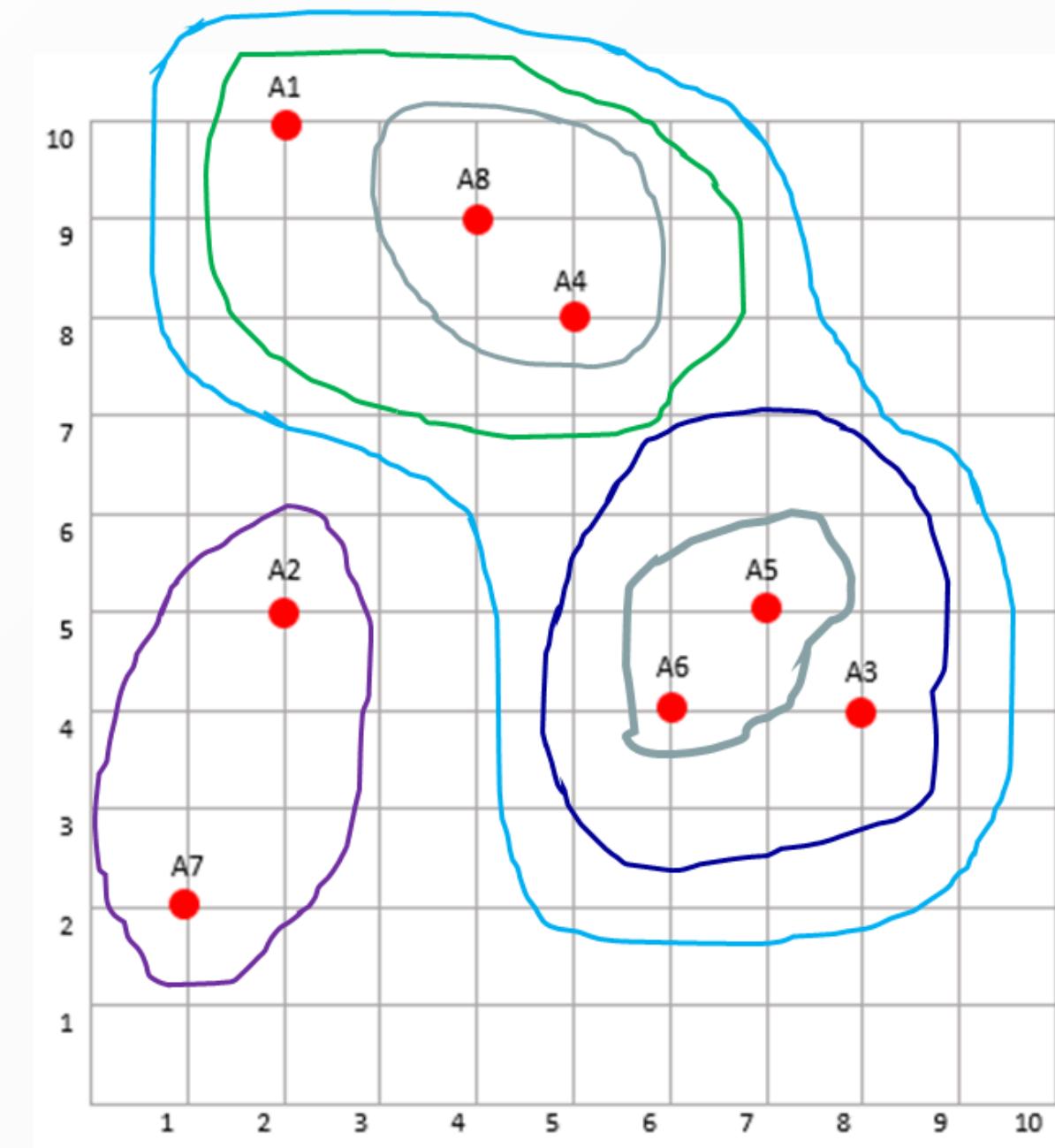


	A2	A1, A4, A8	A3, A5, A6	A7
A2	0	4.2	4.1	3.1
A1, A4, A8		0	3.6	7.2
A3, A5, A6			0	5.4
A7				0

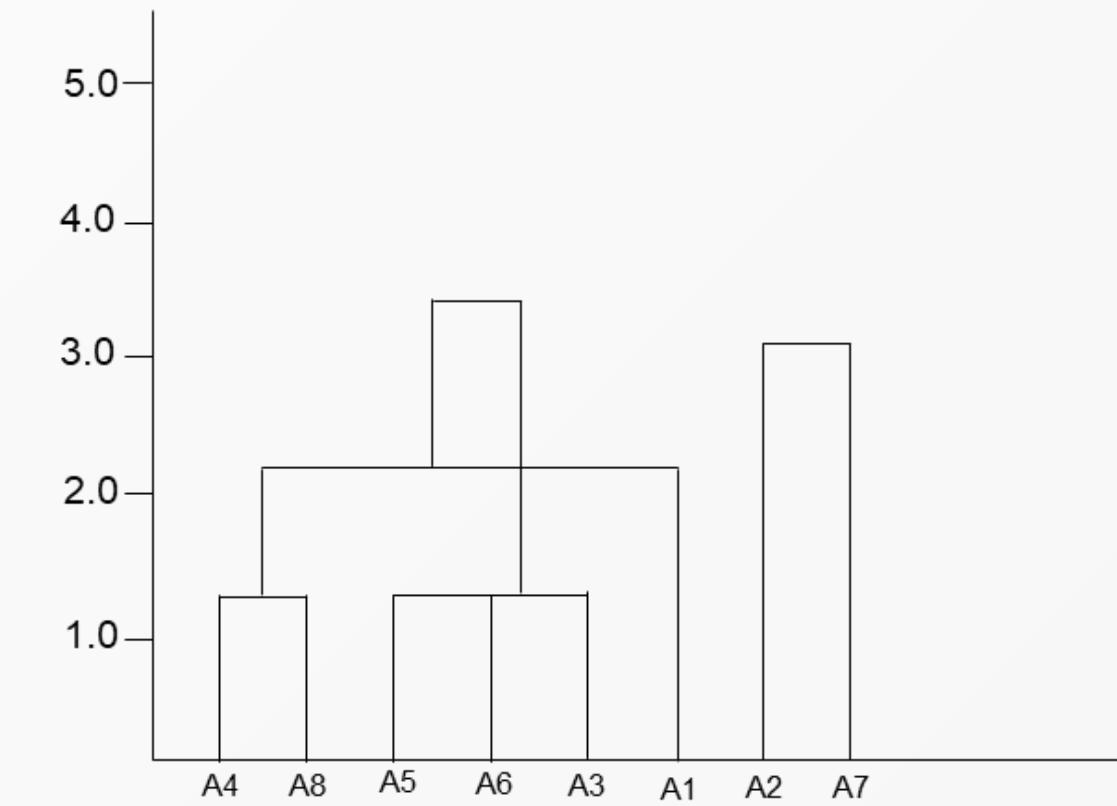


Steps to Perform Clustering

5th clustering cycle

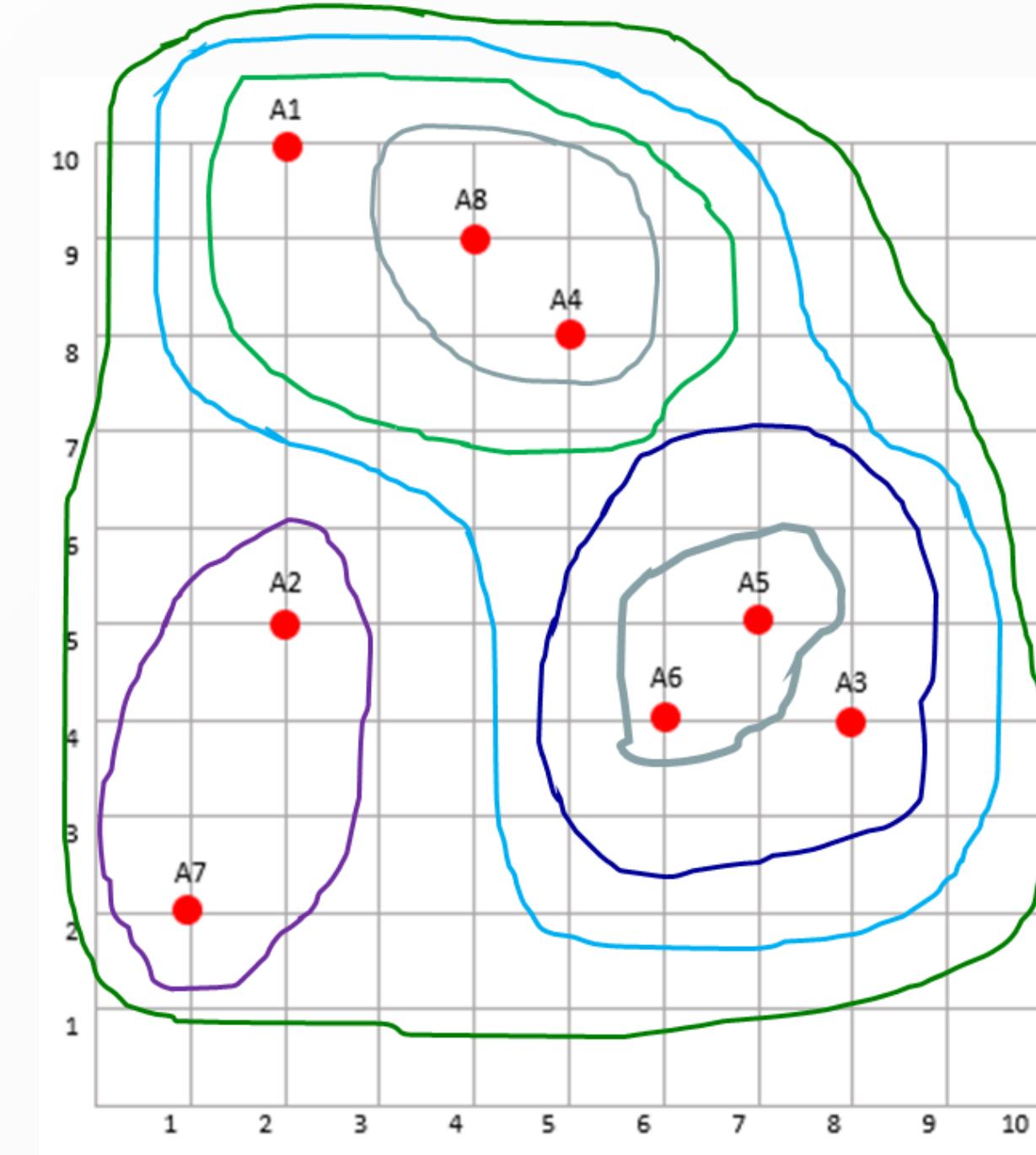


	A2, A7	A1, A4, A8	A3, A5, A6
A2, A7	0	4.2	4.1
A1, A4, A8		0	3.6
A3, A5, A6			0

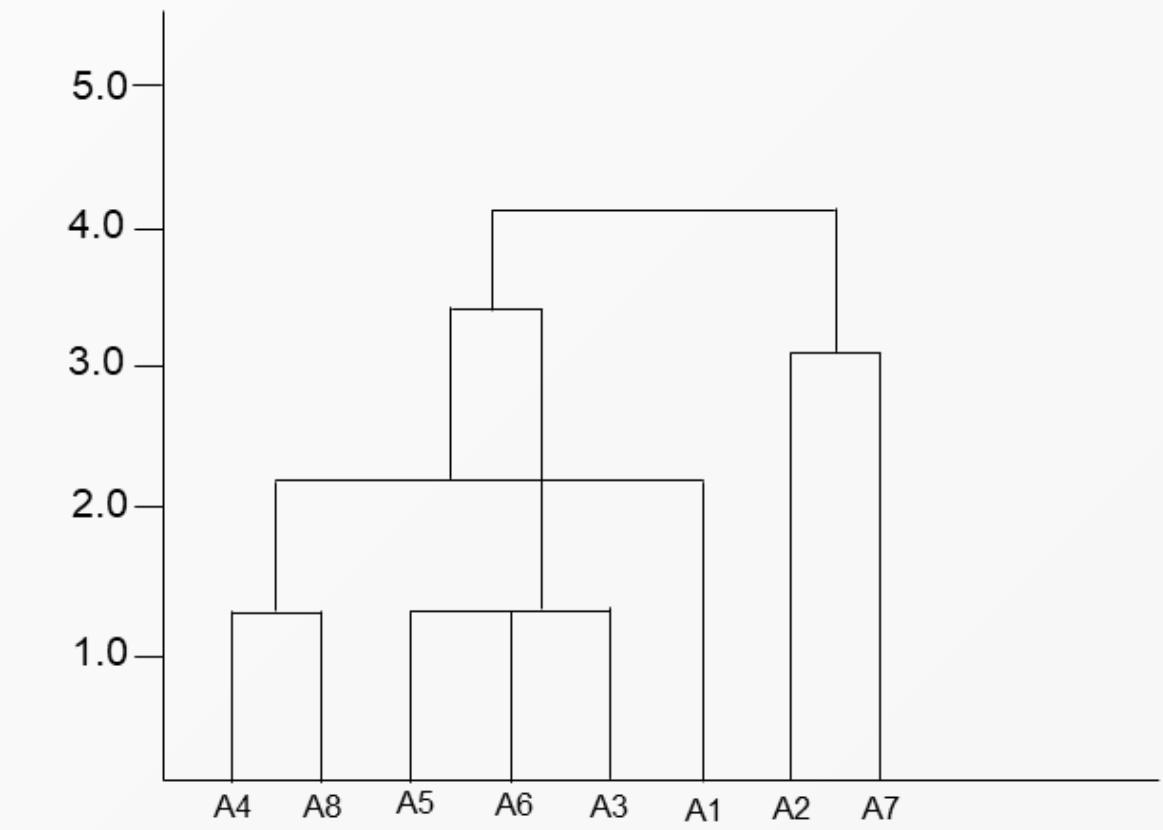


Steps to Perform Clustering

6th clustering cycle

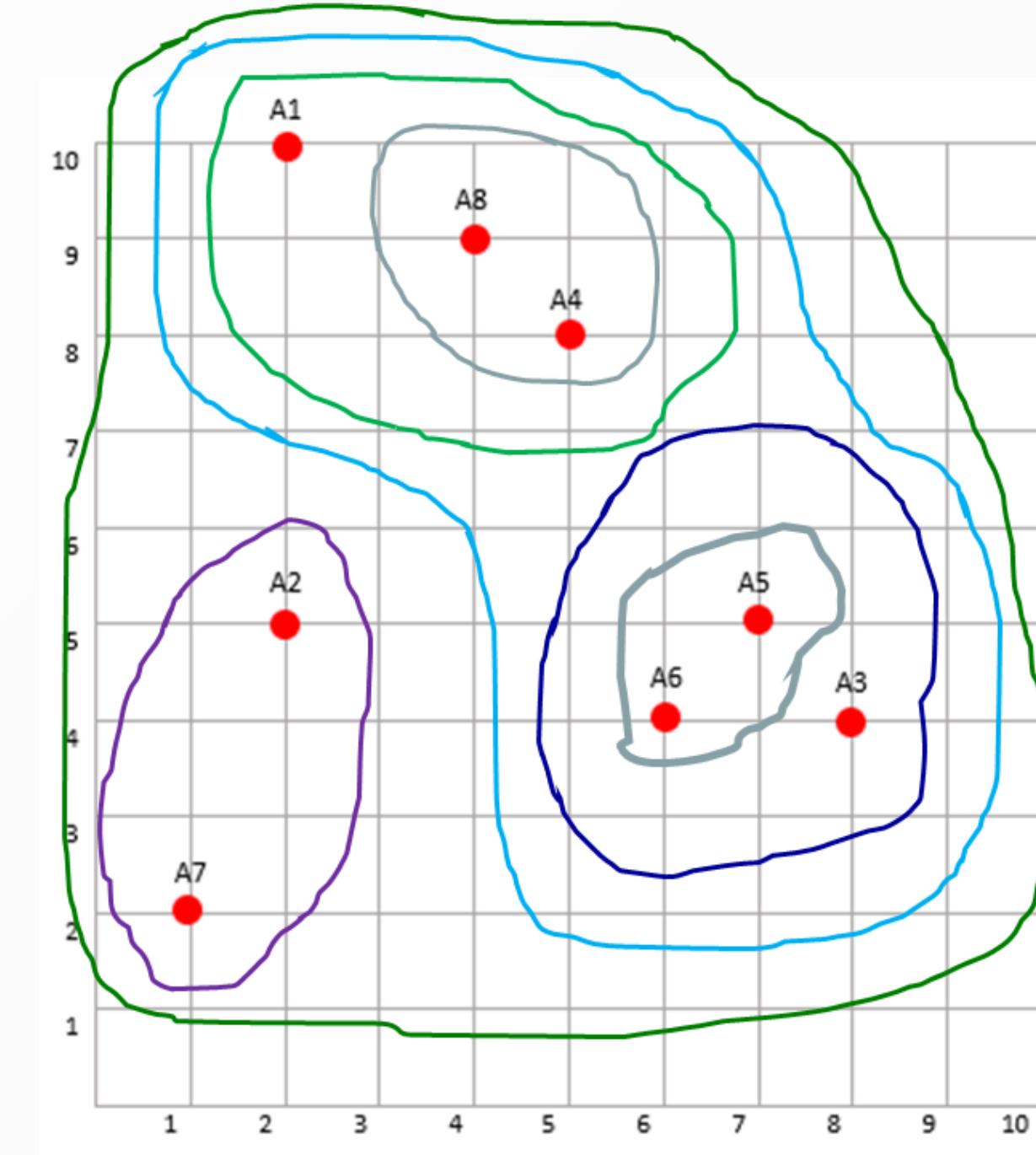


	A2, A7	A1, A4, A8, A3, A5, A6
A2, A7	0	4.1
A1, A4, A8, A3, A5, A6	0	

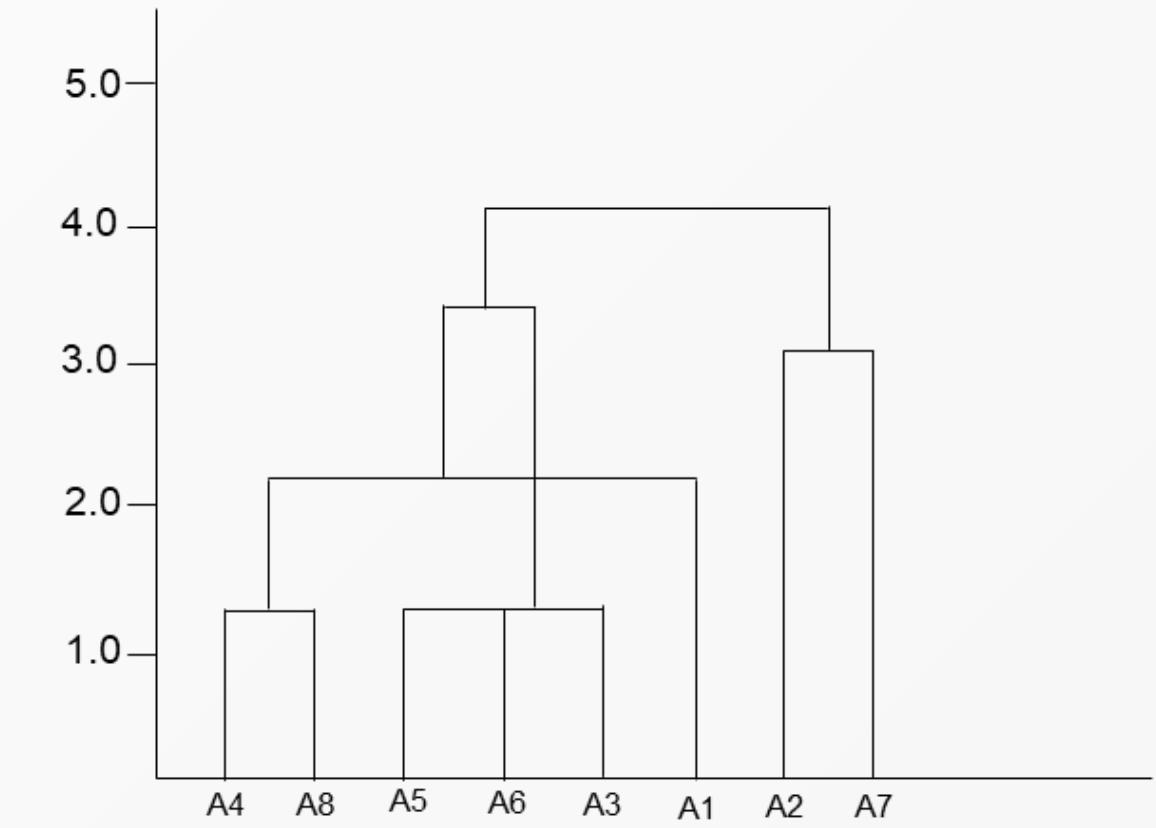


Steps to Perform Clustering

6th clustering cycle



	A2, A7	A1, A4, A8, A3, A5, A6
A2, A7	0	4.1
A1, A4, A8, A3, A5, A6	0	



Steps to Perform Clustering

2. Partitioning methods

- The most common one is k-means
- However, for partitioning methods, we need to determine our initial cluster.
- We will get back to k-means later

Steps to Perform Clustering

Step 3: Decide on the number of clusters

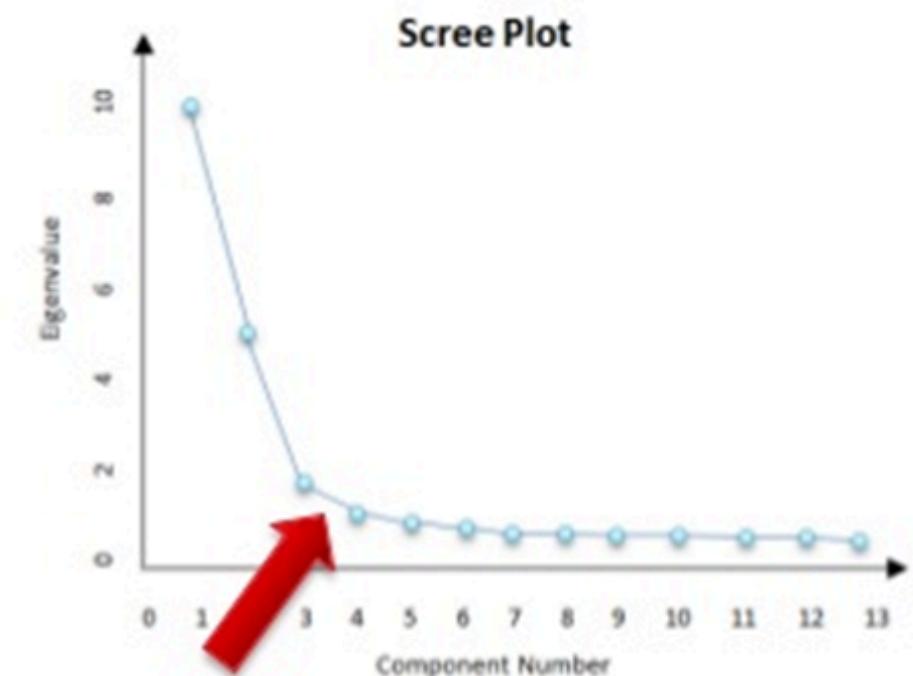
- The main objective is:
- To achieve **maximum inter-variance** and **minimum intra-variance**
- One of the method to determine number of clusters is by using **elbow plot**

Steps to Perform Clustering

- Elbow criterion:

Choose a number of clusters so that by adding another cluster does not add sufficient additional information

- ❖ A scree plot is sometimes referred to as an "elbow" plot.



of Clusters should be 3 or 4

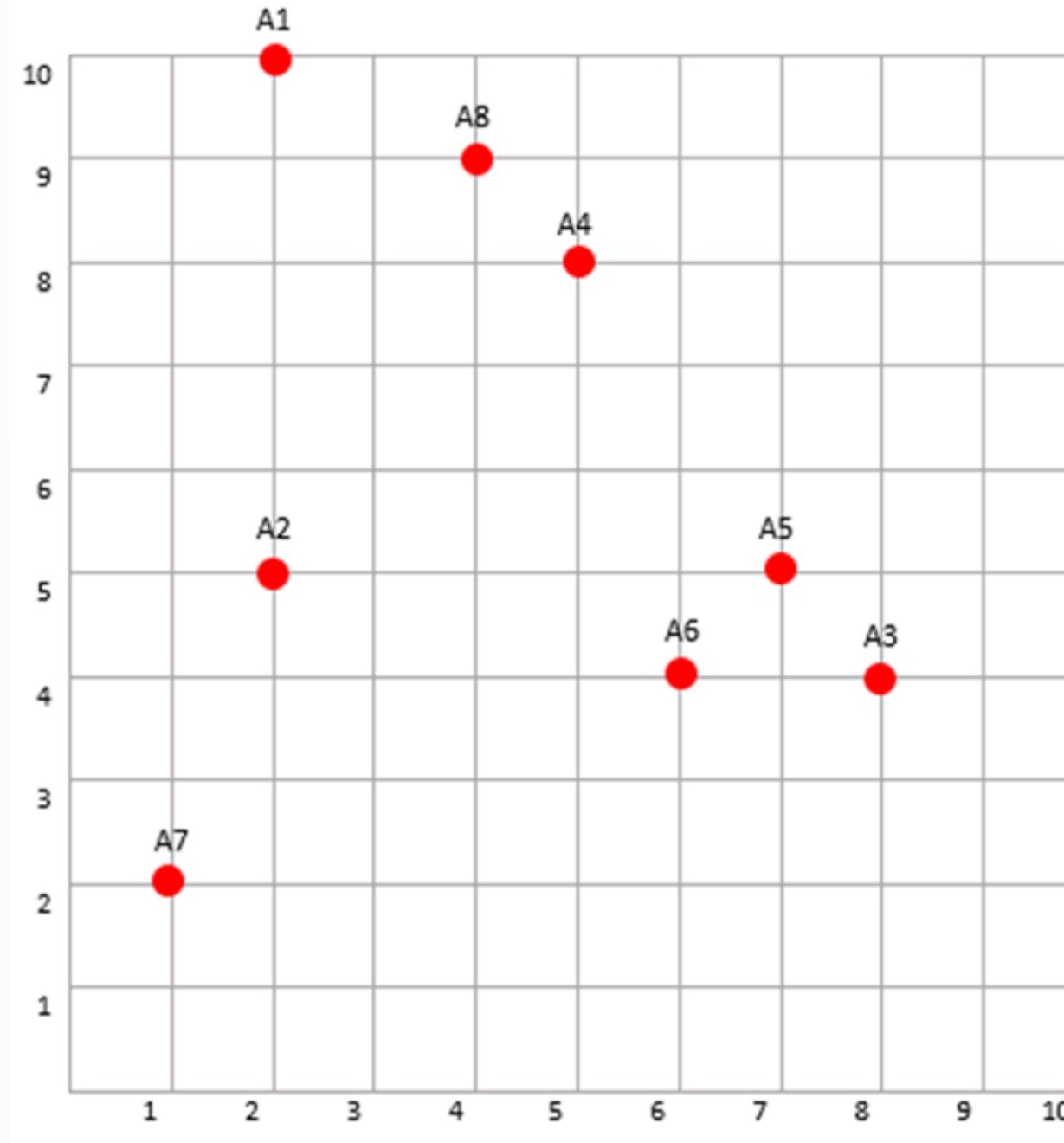
- ❖ In order to identify the optimal number of clusters for further analysis, we need to look for the "bend" in the graph at the elbow.

Steps to Perform Clustering

- Steps to perform k-means cluster analysis:
 - Choose the number of cluster, k
 - Generate k random points as cluster centroids
 - Assign each point to the nearest cluster centroid
 - Recompute the new cluster centroid
 - Repeat the two previous steps until the convergence criterion is met

Steps to Perform Clustering

- Let's say we use the same data set:



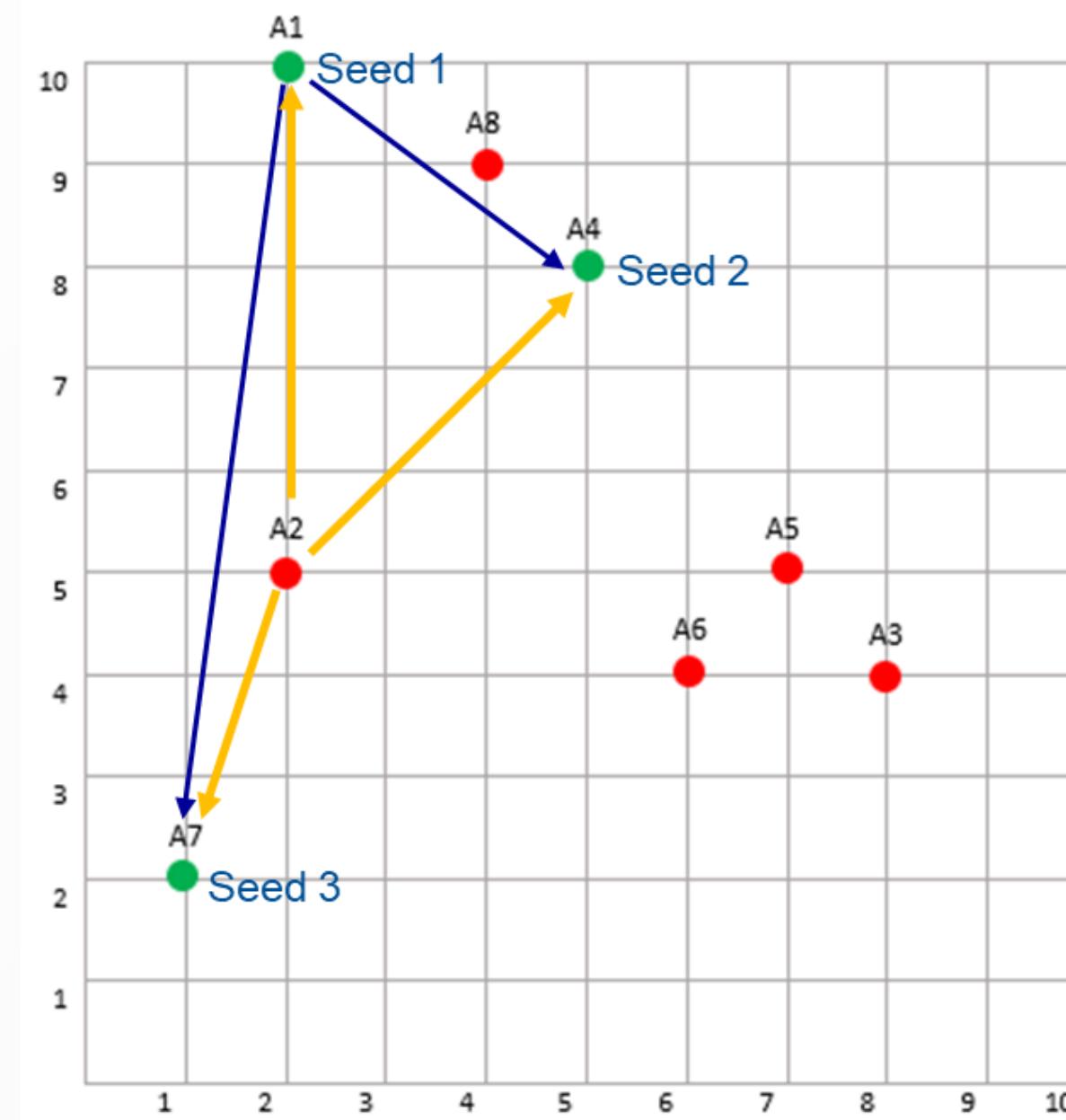
	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	5	6	3.6	7.1	7.2	8.1	2.2
A2		0	6.1	4.2	5	4.1	3.1	4.5
A3			0	5	1.4	1.4	7.2	6.4
A4				0	3.6	4.1	7.2	1.4
A5					0	1.4	6.7	5
A6						0	5.4	5.4
A7							0	7.6
A8								0

Steps to Perform Clustering

- Steps to perform k-means cluster analysis:
 1. Choose the number of cluster, k
 - Let's say we start with $k=3$
 2. Generate k random points as cluster centroids
 - And we choose the following points as the cluster centroids
 - A1, A4, A7

Steps to Perform Clustering

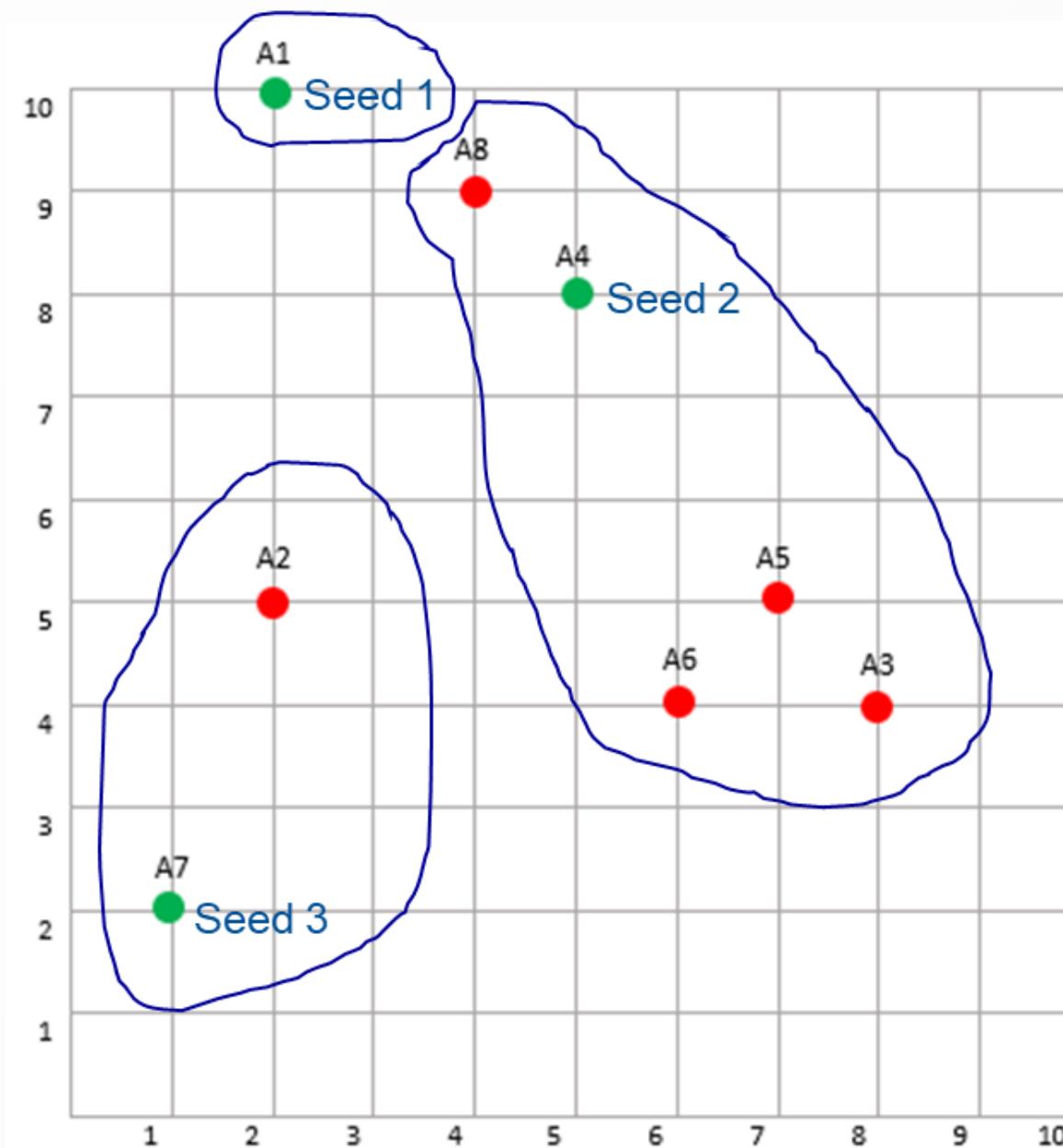
- Then, we need to calculate the distance for each point to each centroid (seed)



Distance from A1 to:		Distance from A2 to:	
Seed 1	0	Seed 1	5
Seed 2	3.6	Seed 2	4.2
Seed 3	8.1	Seed 3	3.1
Distance from A3 to:		Distance from A4 to:	
Seed 1	6	Seed 1	3.6
Seed 2	5	Seed 2	0
Seed 3	7.2	Seed 3	7.2
Distance from A5 to:		Distance from A6 to:	
Seed 1	7.1	Seed 1	7.2
Seed 2	3.6	Seed 2	4.1
Seed 3	6.7	Seed 3	5.4

Steps to Perform Clustering

- Assign each point to the nearest cluster centroid



Distance from A1 to:		Distance from A2 to:	
Seed 1	0	Seed 1	5
Seed 2	3.6	Seed 2	4.2
Seed 3	8.1	Seed 3	3.1

Distance from A3 to:		Distance from A4 to:	
Seed 1	6	Seed 1	3.6
Seed 2	5	Seed 2	0
Seed 3	7.2	Seed 3	7.2

Distance from A5 to:		Distance from A6 to:	
Seed 1	7.1	Seed 1	7.2
Seed 2	3.6	Seed 2	4.1
Seed 3	6.7	Seed 3	5.4

Distance from A7 to:		Distance from A8 to:	
Seed 1	8.1	Seed 1	2.2
Seed 2	7.2	Seed 2	1.4
Seed 3	0	Seed 3	7.6

Steps to Perform Clustering

- Recompute the new cluster centroid

Seed 1 = (2,10)

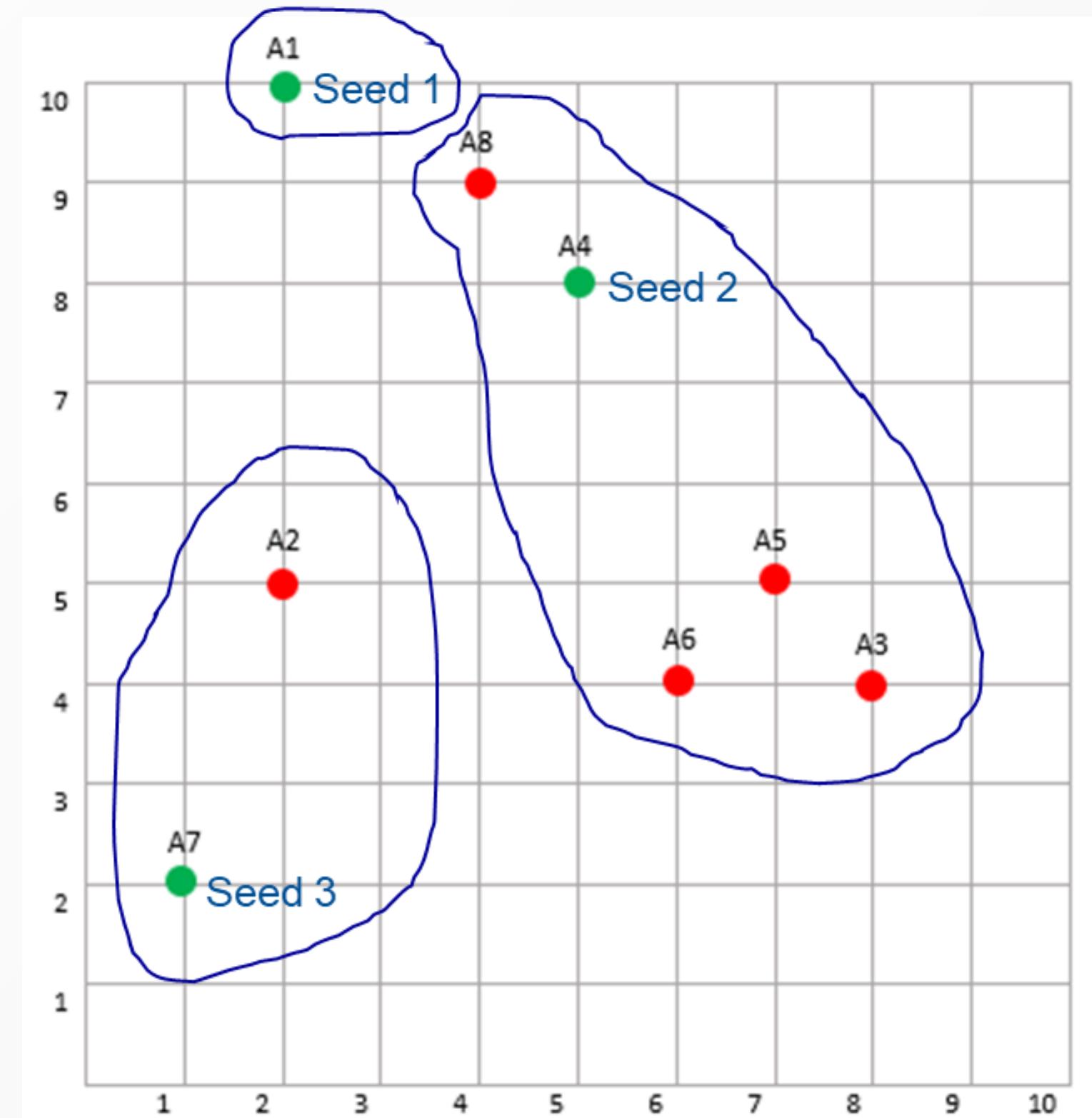
$$\text{Seed 2} = (8+5+7+6+4)/5,$$

$$= (4+8+5+4+9)/5$$

$$= (6, 6)$$

$$\text{Seed 3} = (2+1)/2, (5+2)/2$$

$$= (1.5, 3.5)$$



Steps to Perform Clustering

- Recompute the new cluster centroid

Seed 1 = (2,10)

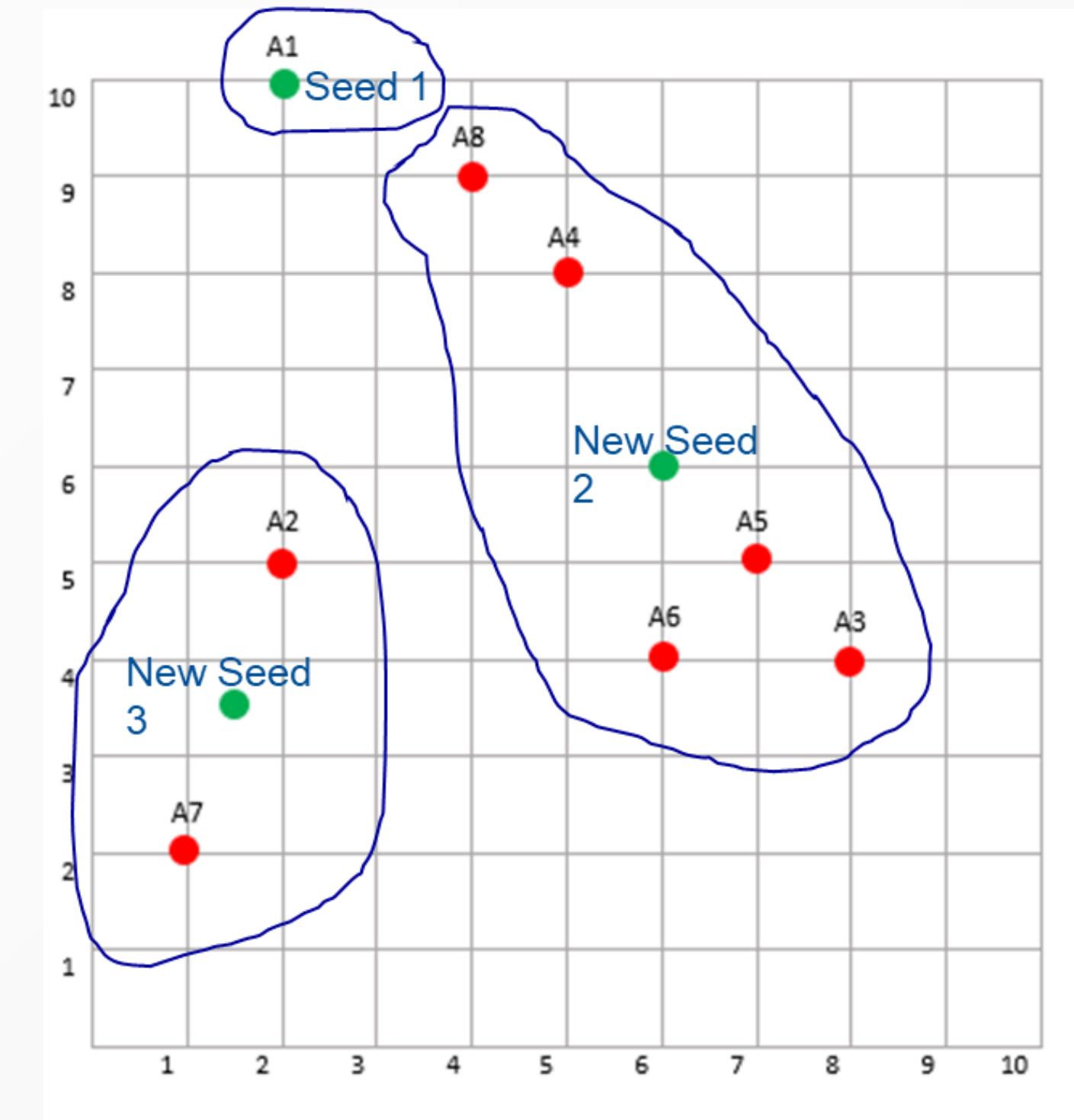
$$\text{Seed 2} = (8+5+7+6+4)/5,$$

$$= (4+8+5+4+9)/5$$

$$= (6, 6)$$

$$\text{Seed 3} = (2+1)/2, (5+2)/2$$

$$= (1.5, 3.5)$$



Steps to Perform Clustering

- Repeat the two previous steps until the convergence criterion is met
- Convergence criterion – when the assignment of points in clusters do not change over multiple iterations

Steps to Perform Clustering

Step 4: Validate and Interpret cluster solution

- Stability and Validity
- Stability is evaluated using different clustering procedures on the same data and testing whether these yield the same results.
- Eg: for hierarchical clustering, use different distance measures
- Validity can be evaluated using criterion validity

Steps to Perform Clustering

Step 4: Validate and Interpret cluster solution

- Profiling of cluster
- Interpreting the clusters by examining the cluster centroids
- It helps to shed light on whether the segments are conceptually distinguishable
- This information will also help to find the meaningful label or name for the cluster to adequately reflects the objects in the cluster

GLOW 2024

lailaghani/ **GLOW2024**



1
Contributor

0
Issues

0
Stars

0
Forks



lailaghani/GLOW2024

Contribute to lailaghani/GLOW2024 development by creating an account on GitHub.

 GitHub