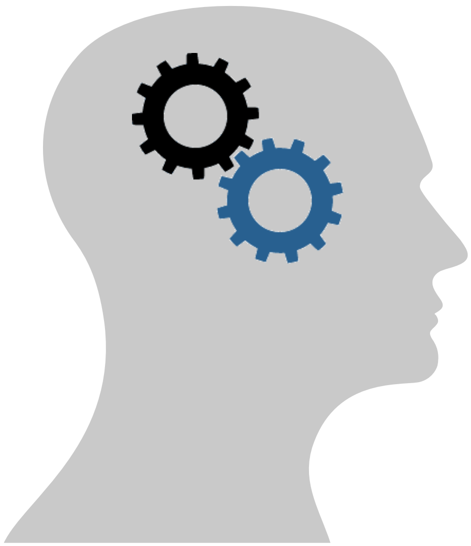
Introduction

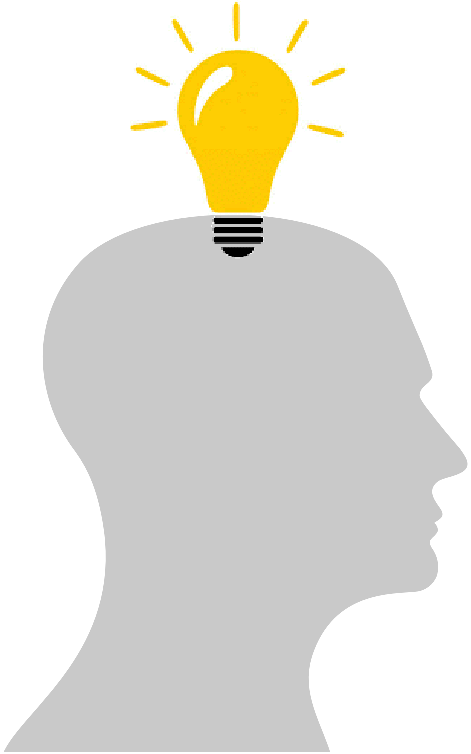


Consider a scenario where a pizza store wants to open a new center.

To choose the best location, it should analyze factors such as distance, accessibility, ease of delivery, population, etc.

Keeping all these factors in mind, how can the best location be predicted?

Introduction



The team should conduct a thorough analysis that would help in understanding how the delivery locations can be grouped, hence reducing the average distance for both people and delivery executives.

This can be done using clustering algorithms.

What Is Clustering?

Cluster analysis or clustering is the most commonly used technique of unsupervised learning used to find data clusters so that each cluster has the most closely matched data.

Unsupervised Learning is a subset of Machine Learning used to extract inferences from datasets that consist of input data without labeled responses.

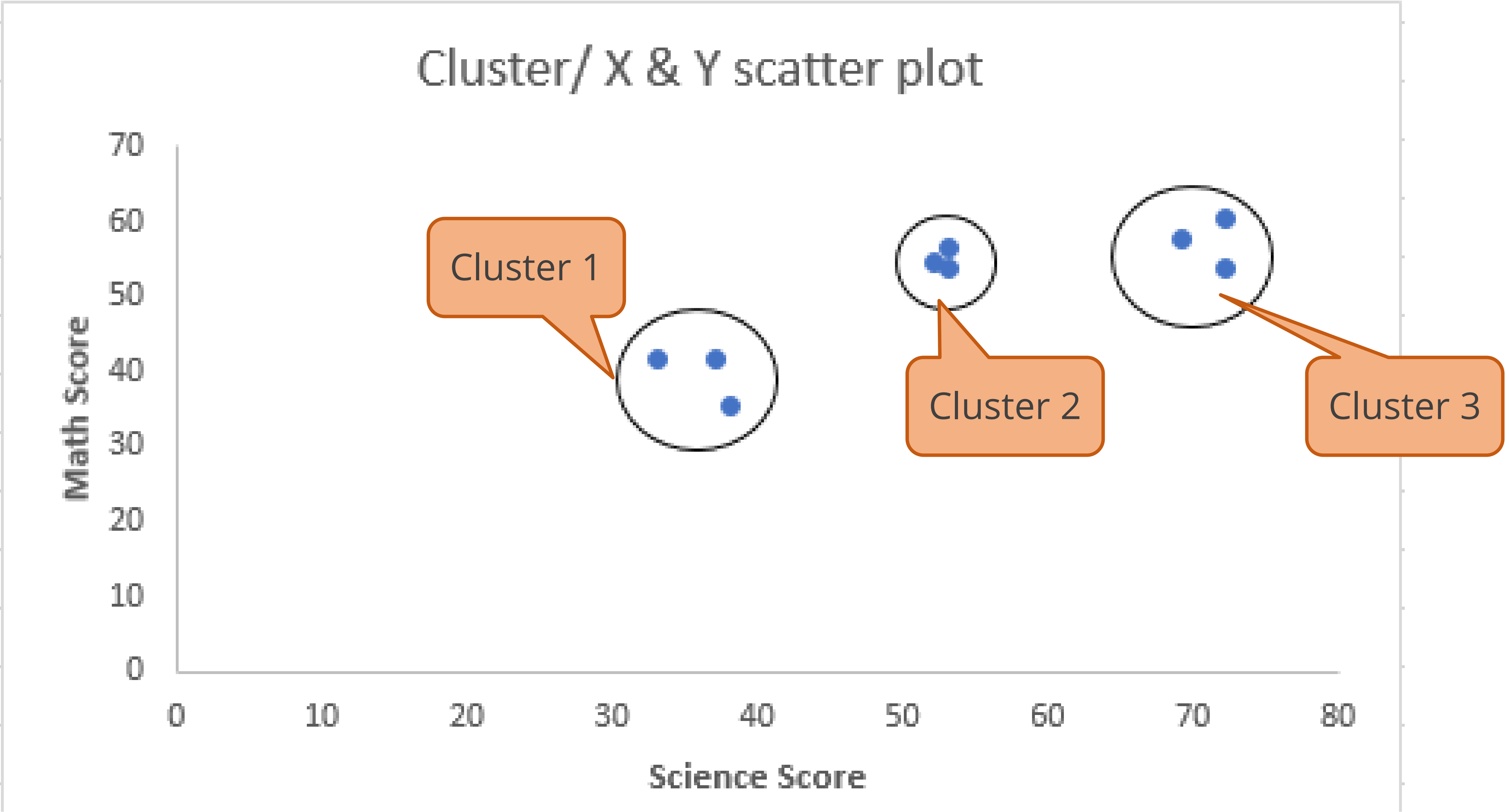
Clustering: Example

Consider a scenario where you need to create a cluster/group of students who are of similar aptitude using clustering. The following data is available.

ID	Math	Science
1	37	42
2	33	42
3	38	36
4	53	54
5	52	55
6	53	57
7	69	58
8	72	54
9	72	61

Clustering: Example

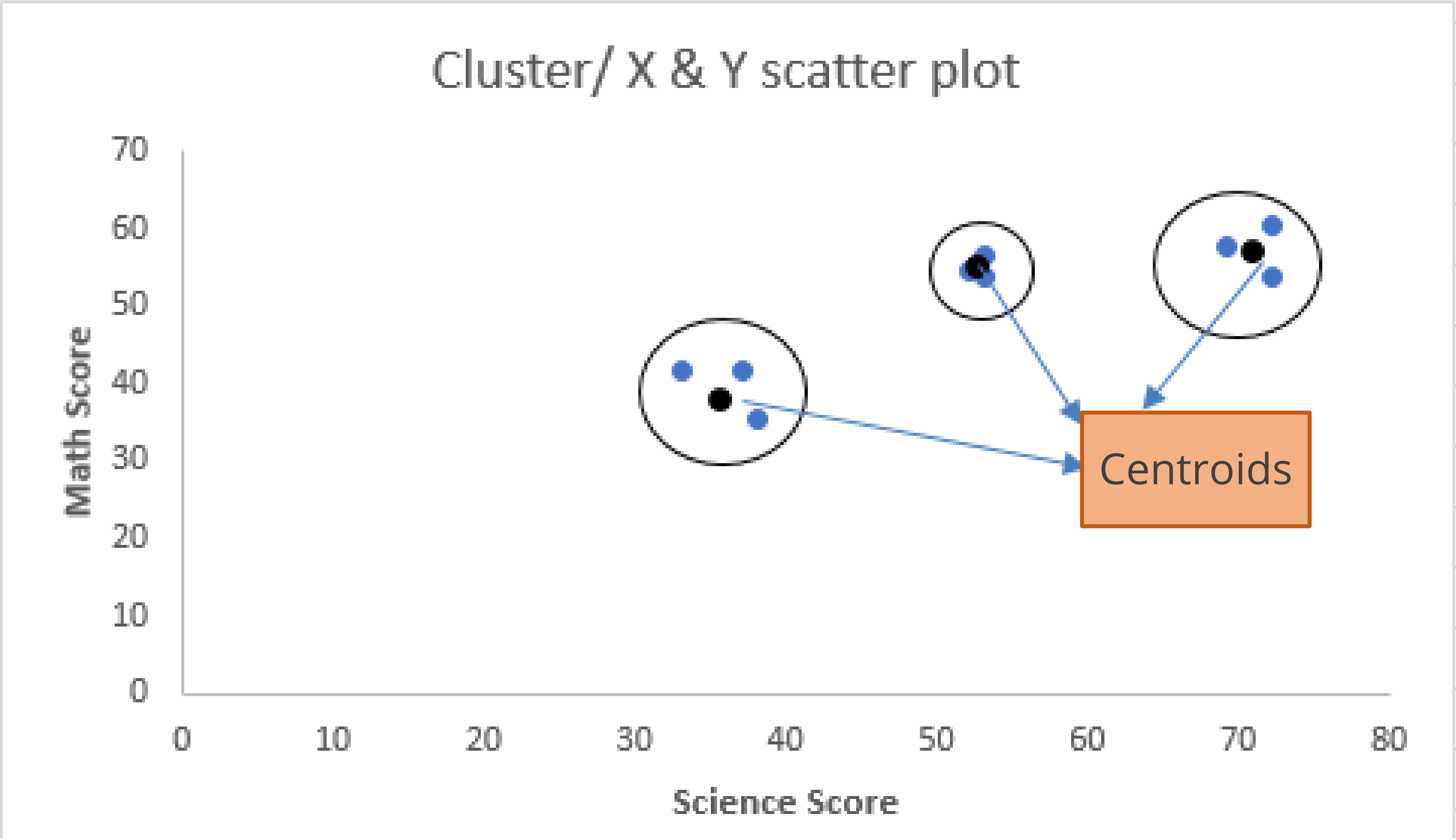
PLOTTING THE OBSERVATION



Clustering: Example

CENTROIDS

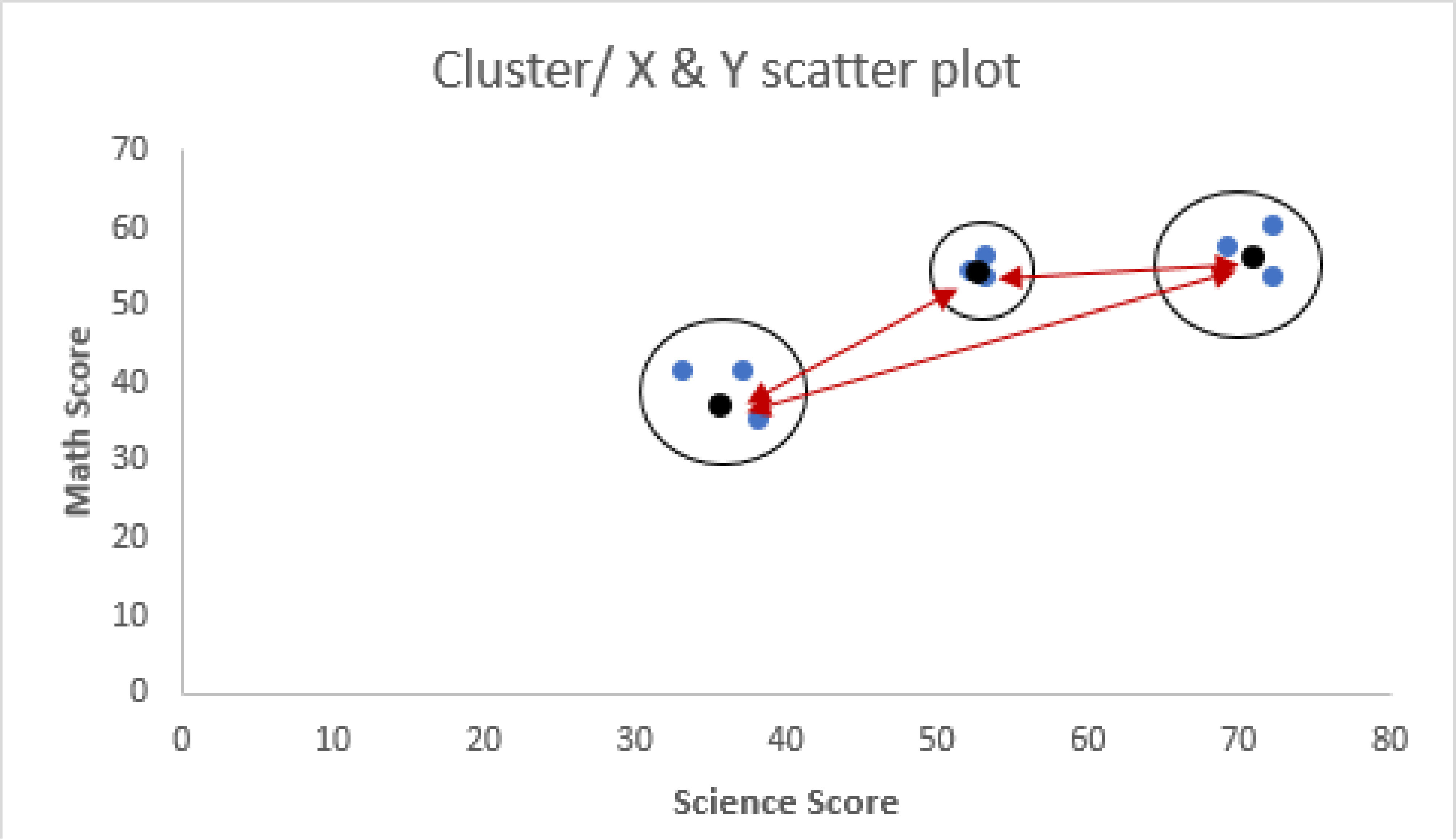
Each of these clusters has center points, called centroids.



Clustering: Example

DISTANCE BETWEEN CLUSTERS

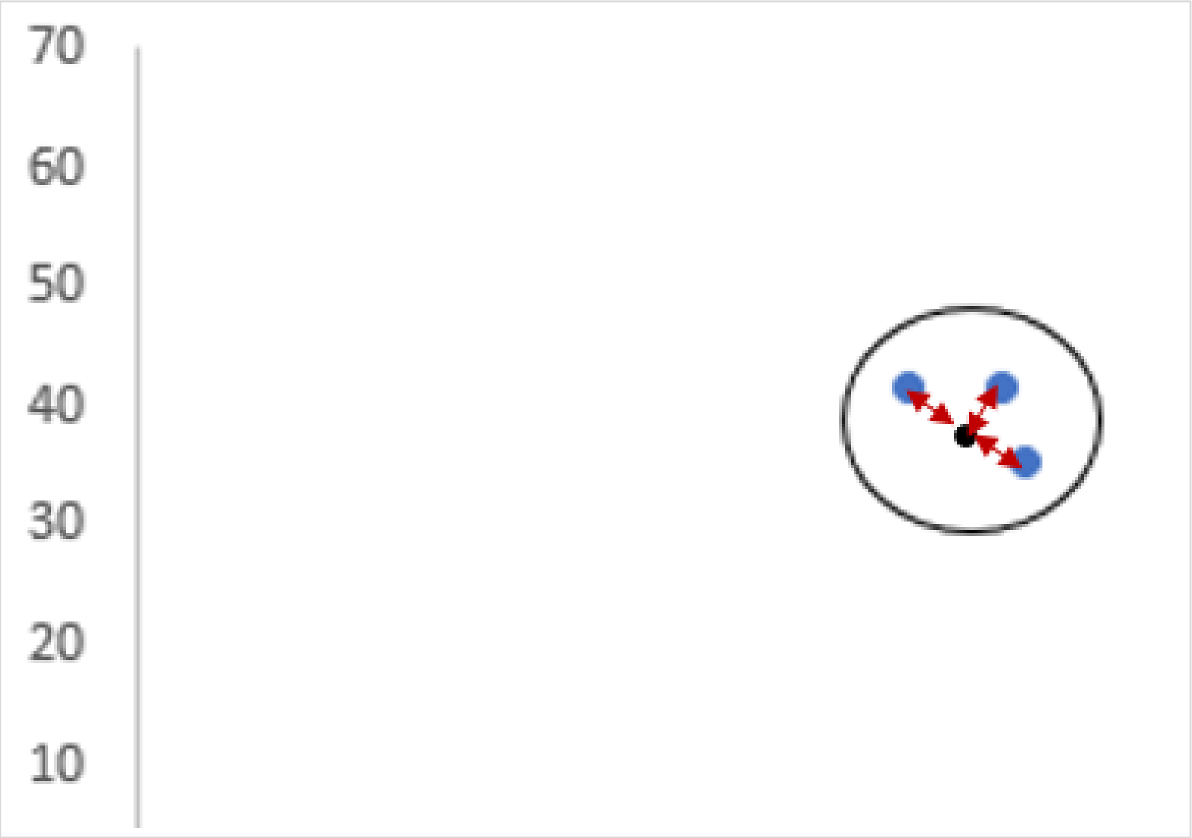
Distance between the cluster centroids is termed distance between clusters.



Clustering: Example

DISTANCE WITHIN CLUSTERS

Average distance of observation in a cluster from its cluster centroid is called distance within cluster.



Other Examples of Clustering

- Grouping the content of a website or product in a retail business
- Segmenting customers or users into different groups on the basis of their metadata and behavioral characteristics
- Segmenting communities in ecology
- Finding clusters of similar genes in DNA analysis
- Creating image segments to be used in image analysis applications

All of this is done using various clustering methods.

Clustering Methods

Prototype-based Clustering

Hierarchical Clustering

Clustering Methods

Prototype-based Clustering

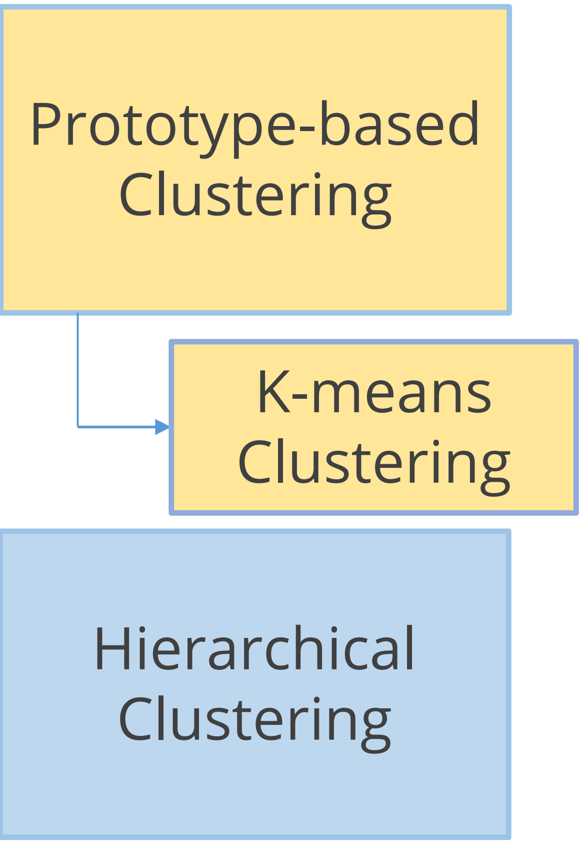
Hierarchical Clustering

Prototype-based clustering assumes that most of the data is located near prototypes (element of data space representing a group of elements).

Example: centroid (average)

It is widely used in banking and sports stat predictions to provide robustifying efforts based on statistics.

Clustering Methods



Steps:

- Decide number of clusters (k)
- Randomly assign k centroids to observations
- Calculate euclidean distance of observations from centroid
- Assign cluster based on min. euclidean distance
- Recalculate the euclidean distance
- Evaluate cluster assignment based on min. euclidean distance
- Repeat 5 and 6 until there is no change in the cluster of observations

Clustering Methods

K-MEANS CLUSTERING: EXAMPLE

Prototype-based Clustering

K-means Clustering

Hierarchical Clustering

The government of California wants to identify high density clusters to build hospitals. (No other ground truth or features are provided apart from the population data). How can the clusters be identified?

Clustering Methods

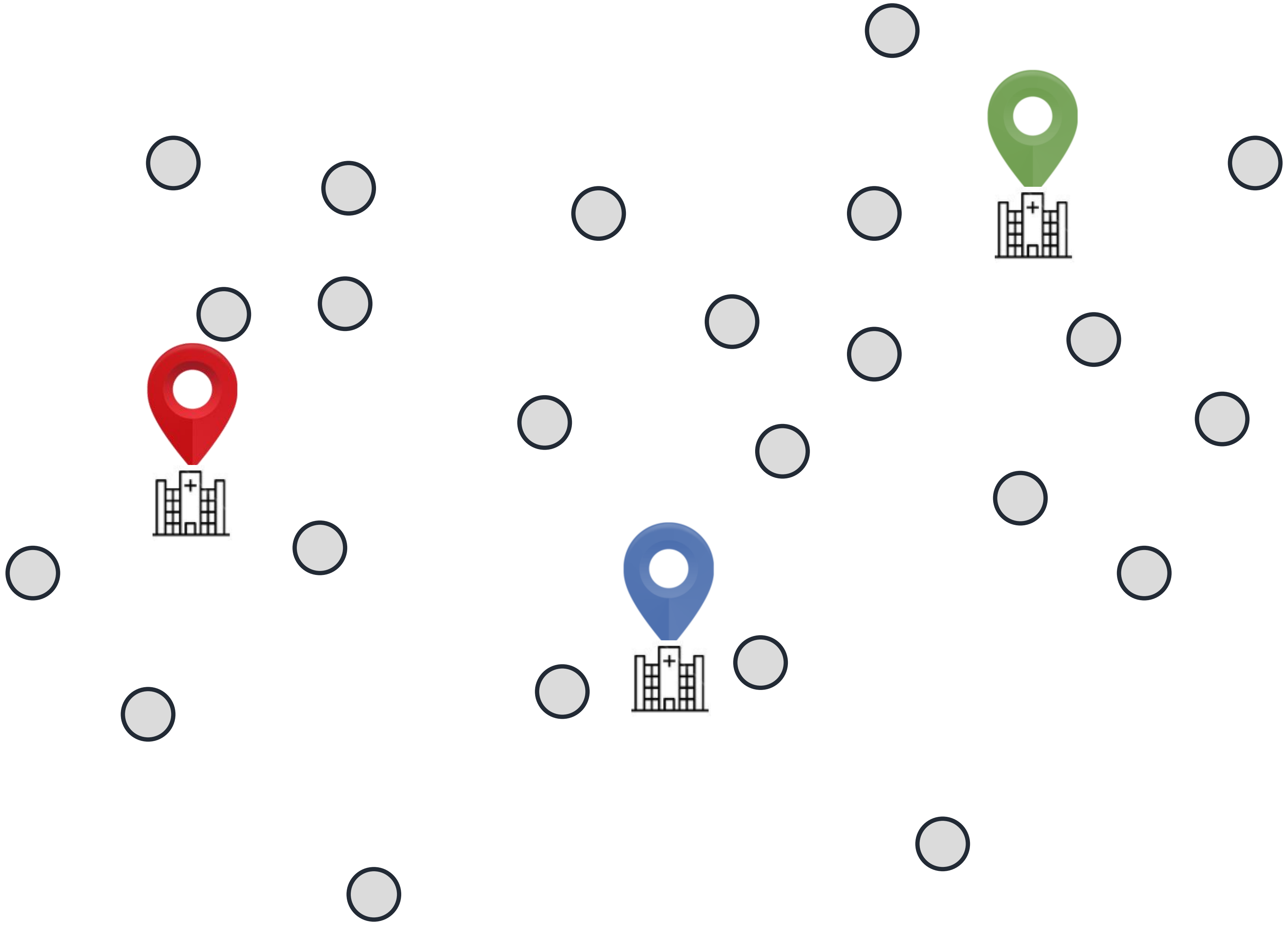
K-MEANS CLUSTERING: EXAMPLE

Start by picking k random centroids. Assume, k = 3.

Prototype-based Clustering

K-means Clustering

Hierarchical Clustering



Clustering Methods

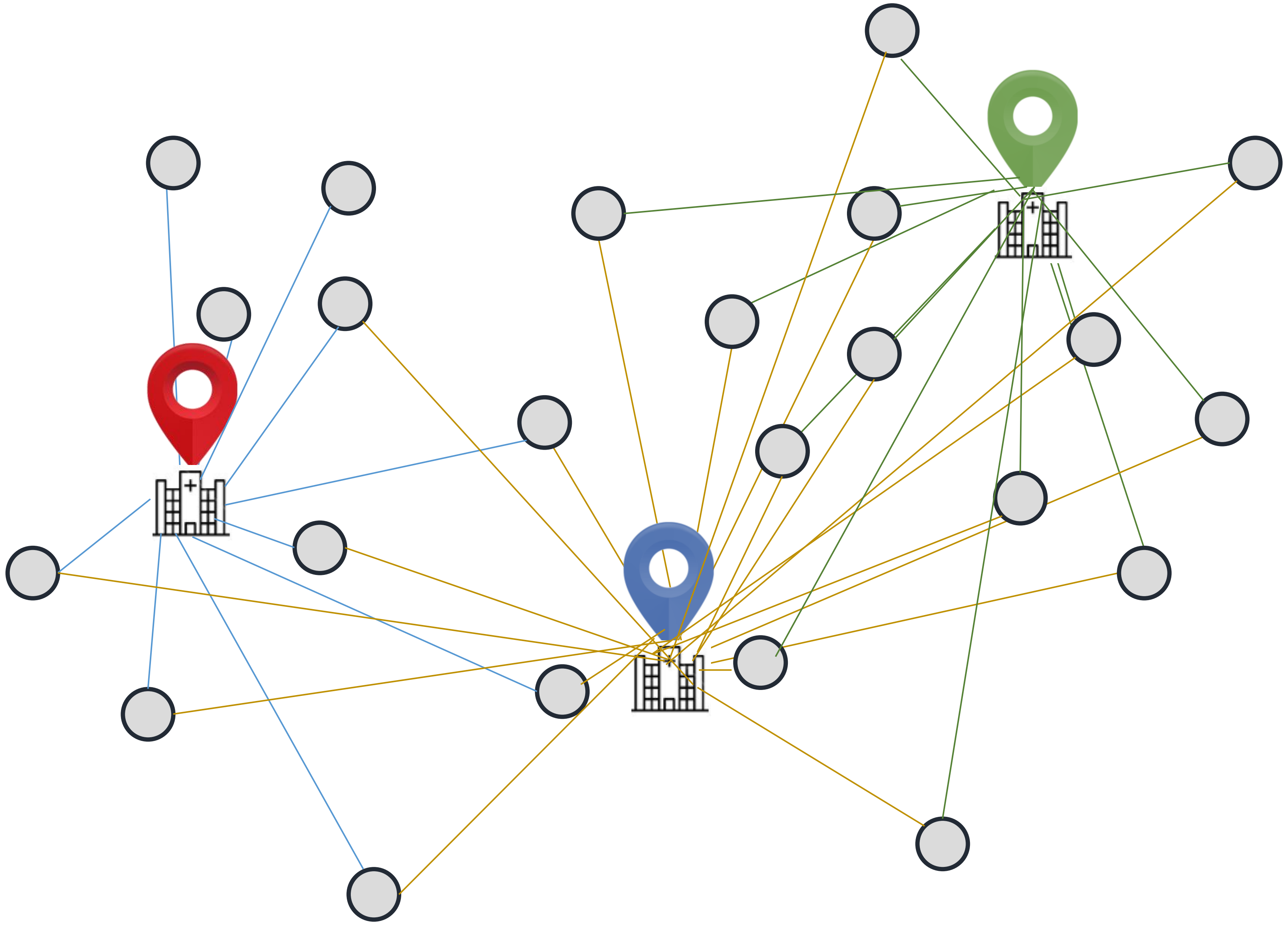
K-MEANS CLUSTERING: EXAMPLE

Assign each point to the nearest centroid.

Prototype-based Clustering

K-means Clustering

Hierarchical Clustering



Clustering Methods

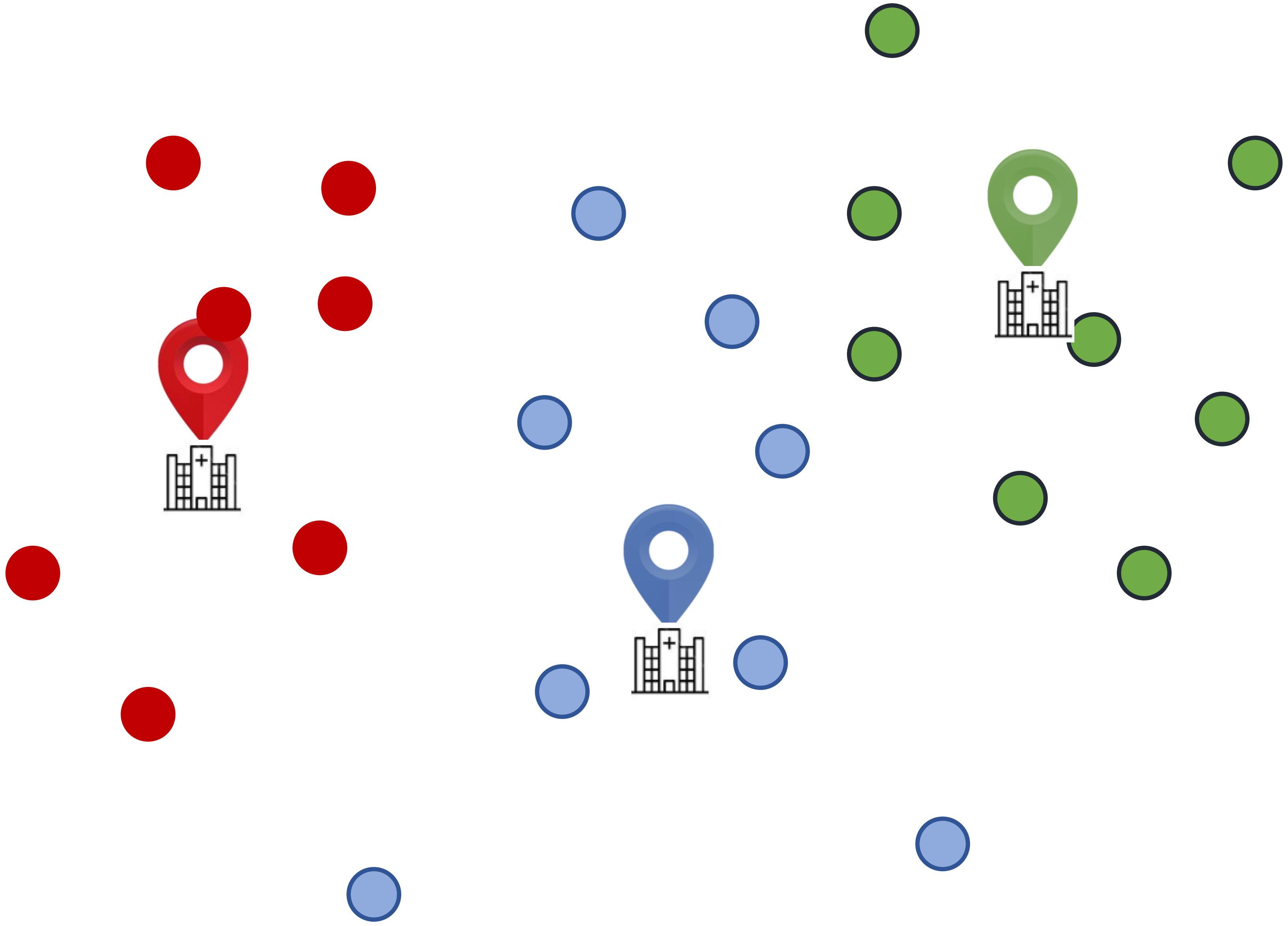
K-MEANS CLUSTERING: EXAMPLE

Prototype-based Clustering

K-means Clustering

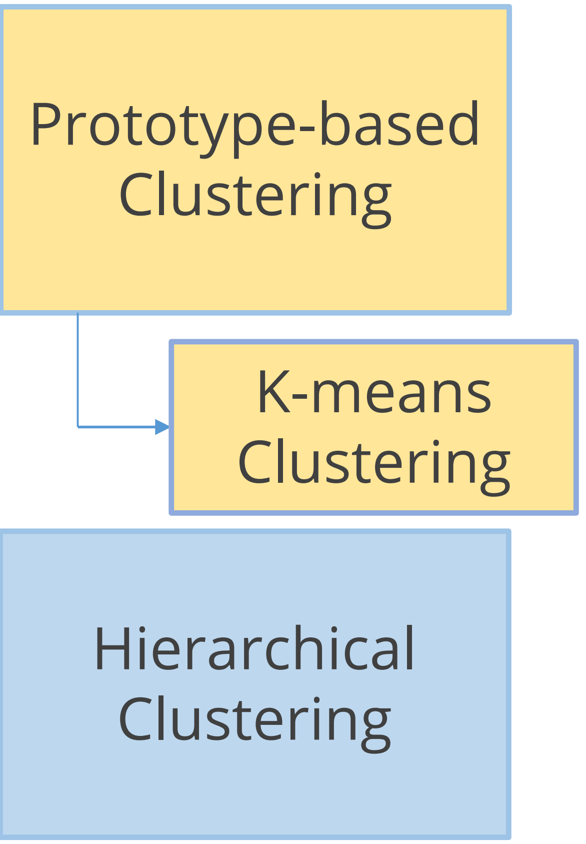
Hierarchical Clustering

Move each centroid to the center of the respective cluster.

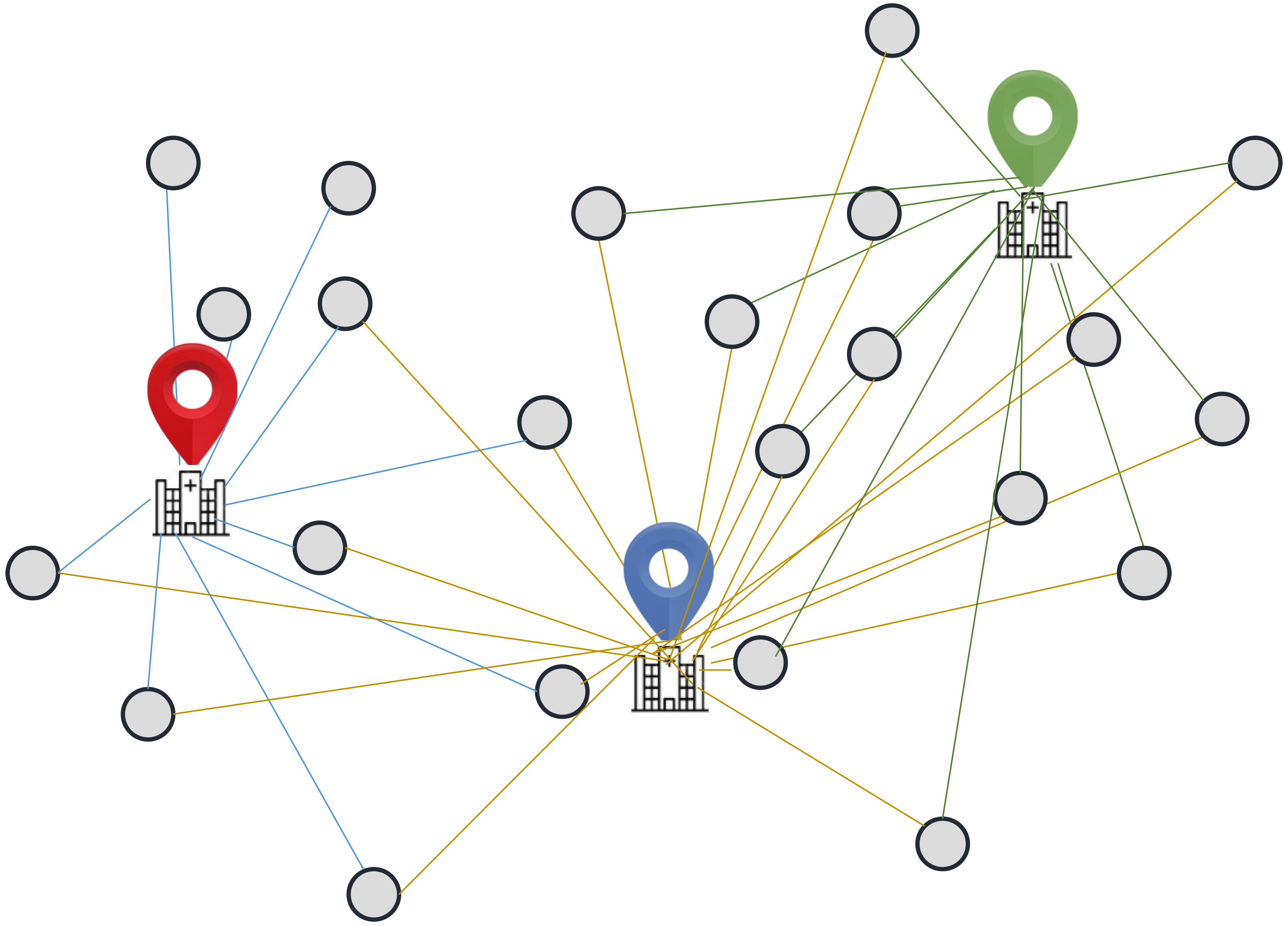


Clustering Methods

K-MEANS CLUSTERING: EXAMPLE



Calculate the distance of the centroids from each point again.



Clustering Methods

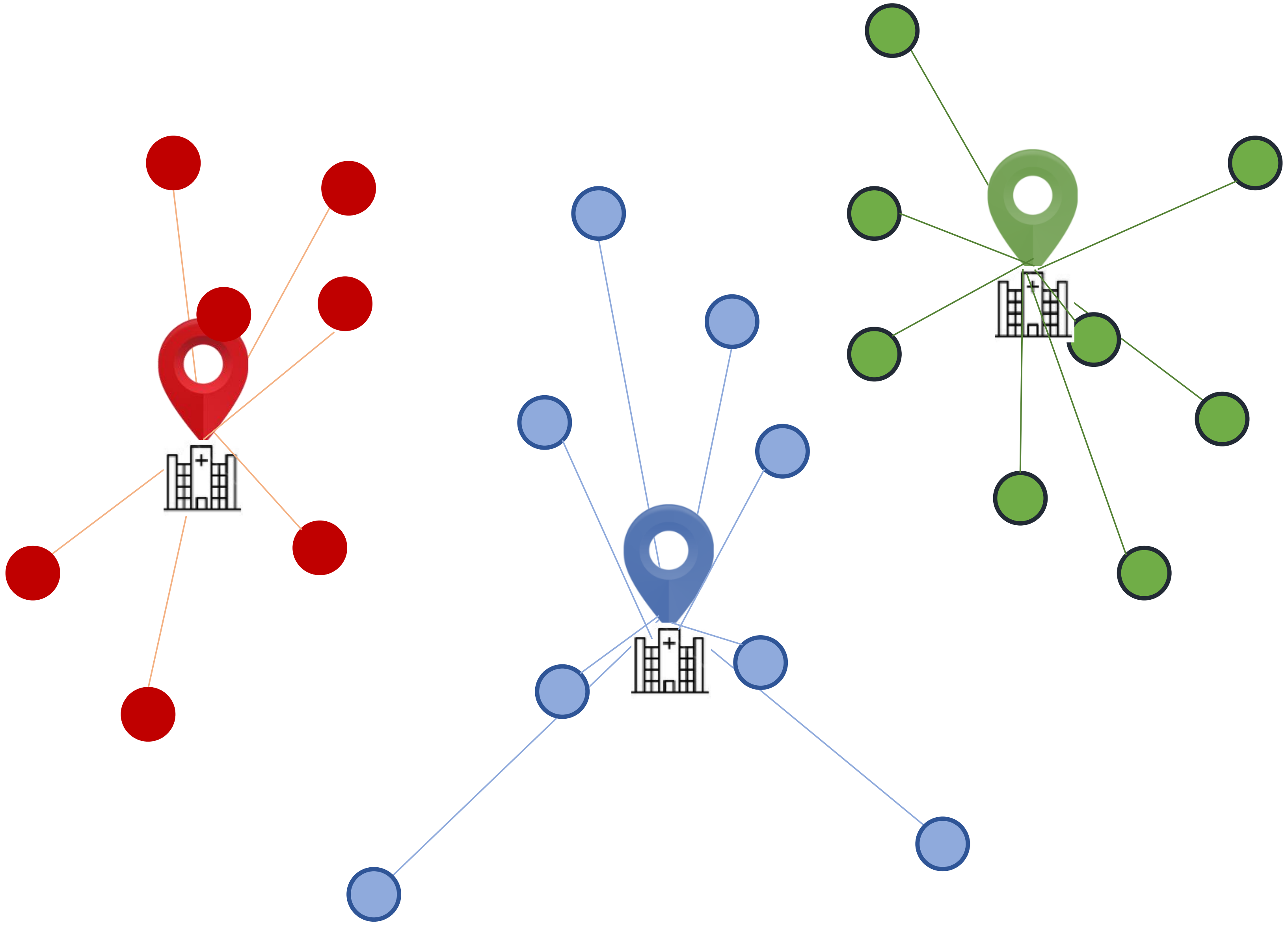
K-MEANS CLUSTERING: EXAMPLE

Prototype-based Clustering

K-means Clustering

Hierarchical Clustering

Move points across clusters and re-calculate the distance from the centroid.



Clustering Methods

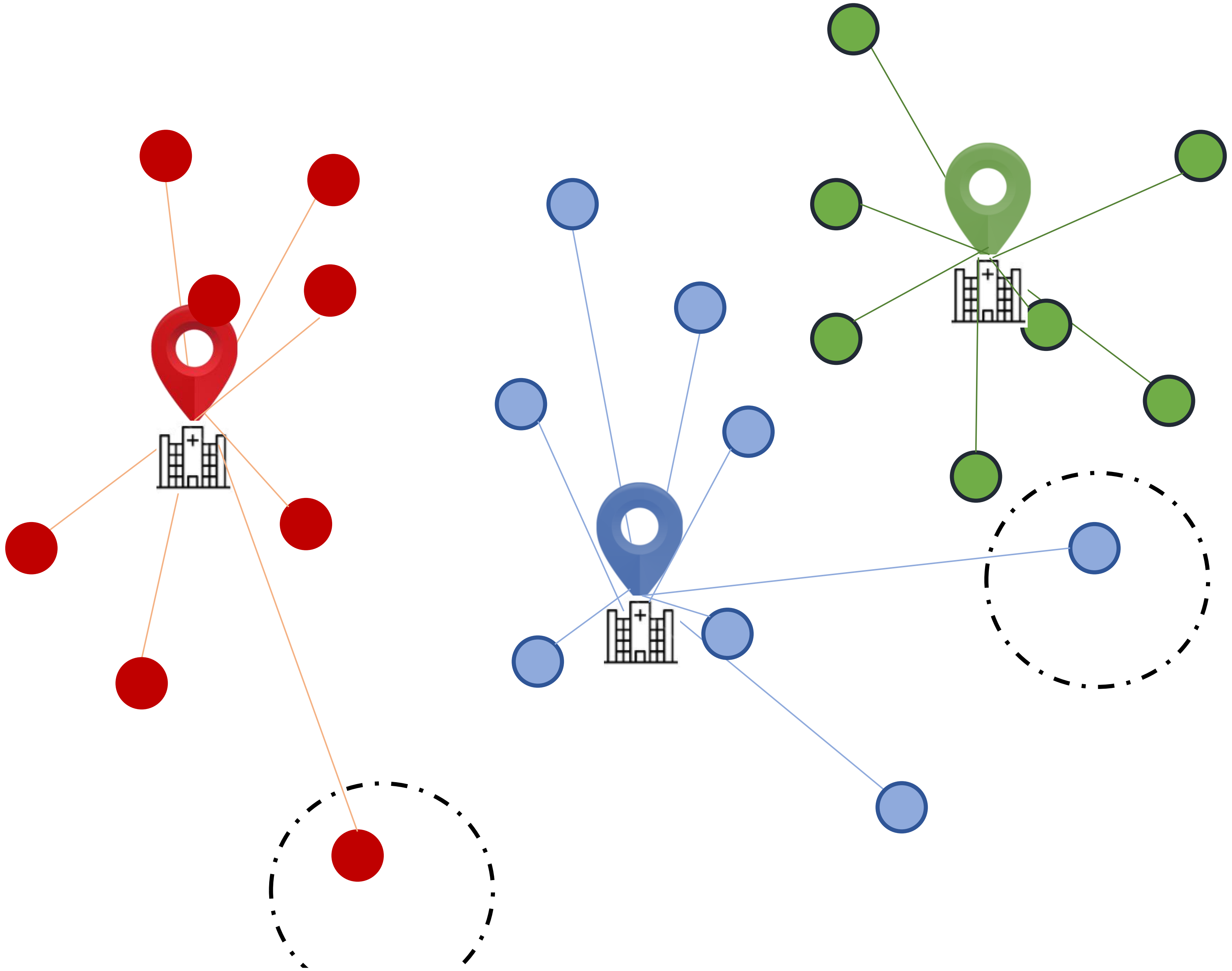
K-MEANS CLUSTERING: EXAMPLE

Prototype-based Clustering

K-means Clustering

Hierarchical Clustering

Keep moving the points across clusters until the distance from the center is minimized.



Clustering Methods

K-MEANS CLUSTERING: EXAMPLE

Prototype-based Clustering

K-means Clustering

Hierarchical Clustering

medicine	weight	pH index
A	1	1
B	2	1
D	4	3
E	5	4

First Iteration

No. of Clusters, k = 2

	weight	pH
C1	1	1
C2	2	1

d1 & d2 values are Euclidean distance C1 and C2

Iter-1	d1	d2	Cluster
A	0	1	C1
B	1	0	C2
D	13	8	C2
E	25	18	C2

medicine	d1	d2	Cluster
A	0	1	C1
B	1	0	C2
D	13	8	C2
E	25	18	C2

Medicine with smallest distance is assigned to the corresponding cluster

Clustering Methods

K-MEANS CLUSTERING: EXAMPLE

Prototype-based Clustering

K-means Clustering

Hierarchical Clustering

medicine	weight	pH index	Cluster
A	1	1	C1
B	2	1	C2
D	4	3	C2
E	5	4	C2

Second Iteration

No. of Clusters, k = 2

	Weight	pH
C1	1	1
C2	3	2

d1 & d2 values are Euclidean distance C1 and C2

med	d1	d2	Cluster
A	0	9.89	C1
B	1	5.56	C1
D	13	0.22	C2
E	25	3.56	C2

med	d1	d2	Cluster
A	0	9.89	C1
B	1	5.56	C1
D	13	0.22	C2
E	25	3.56	C2

Medicine with smallest distance is assigned to the corresponding cluster

Clustering Methods

K-MEANS CLUSTERING: EXAMPLE

Prototype-based Clustering

K-means Clustering

Hierarchical Clustering

medicine	weight	pH index	Cluster
A	1	1	C1
B	2	1	C1
D	4	3	C2
E	5	4	C2

Second Iteration

No. of Clusters, k = 2

	Weight	pH
C1	1.5	1
C2	4.5	3.5

d1 & d2 values are Euclidean distance C1 and C2

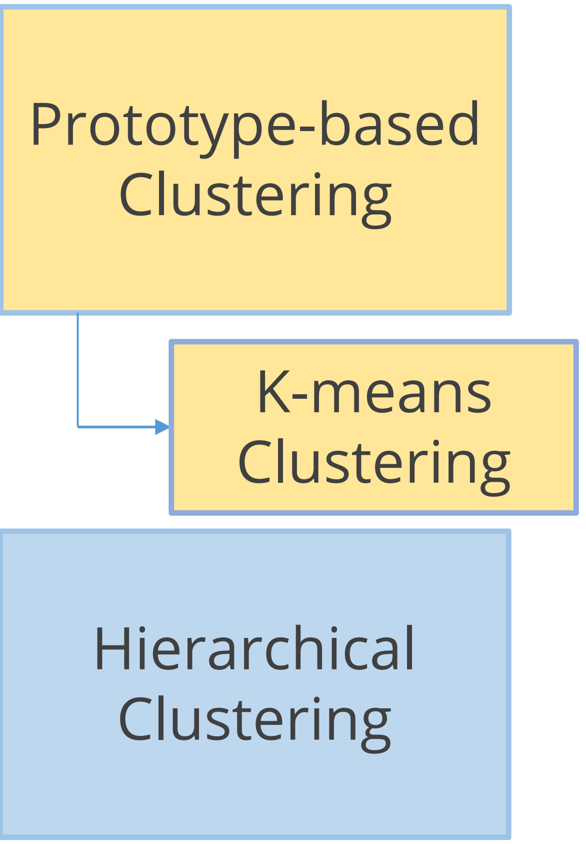
med	d1	d2	Cluster
A	0.25	18.5	C1
B	0.25	12.5	C1
D	10.25	0.5	C2
E	21.25	0.5	C2

med	d1	d2	Cluster
A	0.25	18.5	C1
B	0.25	12.5	C1
D	10.25	0.5	C2
E	21.25	0.5	C2

Medicine with smallest distance is assigned to the corresponding cluster

Clustering Methods

K-MEANS CLUSTERING: EXAMPLE



Steps:

- Decide number of clusters (k)
- Randomly assign k centroids to observations
- Calculate euclidean distance of observations from centroid
- Assign cluster based on min. euclidean distance
- Recalculate the euclidean distance
- Evaluate cluster assignment based on min. euclidean distance
- Repeat 5 and 6 until there is no change in the cluster of observations

Clustering Methods

Prototype-based Clustering

Hierarchical Clustering

It clusters n units/objects, each with p features, into smaller groups and creates a hierarchy of clusters as a dendrogram.

Dendrograms are units in the same cluster joined by a horizontal line. They provide a visual representation of clusters.

Clustering Methods

Prototype-based Clustering

Hierarchical Clustering

They are of two types of Hierarchical clustering:

Type	Method	Approach
Agglomerative clustering	Starts at the individual leaves and successively merges clusters together	Bottom-up
Divisive clustering	Starts at the root and recursively splits the clusters	Top-down

Clustering Methods

Prototype-based Clustering

Hierarchical Clustering

Agglomerative clustering is a process where:

- An $n \times n$ distance matrix is considered, where the number in the i^{th} row and j^{th} column is the distance between the i^{th} and j^{th} units.
- The distance matrix is symmetric with zeros in the diagonal.
- Rows and columns are merged as clusters and the distances between them are updated.

Clustering Methods

Prototype-based Clustering

Hierarchical Clustering

Agglomerative clustering

Consider the distance matrix:

	a	b	c	d	e
a	0				
b	9	0			
c	3	7	0		
d	6	5	9	0	
e	11	10	2	8	0

Clustering Methods

Prototype-based Clustering

Hierarchical Clustering

Consider min distance

	a	b	c	d	e
a	0				
b	9	0			
c	3	7	0		
d	6	5	9	0	
e	11	10	2	8	0

	ce	a	b	d
ce	0			
a	3	0		
b	7	9	0	
d	8	6	5	0

Min (a c , a e)
Min (b c , b e)
Min (d c , d e)

Clustering Methods

Prototype-based Clustering

Hierarchical Clustering

Consider min distance

	ce	a	b	d
ce	0			
a	3	0		
b	7	9	0	
d	8	6	5	0

	cea	b	d
cea	0		
b	7	0	
d	6	5	0

Min (b ce, b a)
Min (d ce, d a)

Clustering Methods

Prototype-based Clustering

Hierarchical Clustering

Consider min distance

	cea	b	d
cea	0		
b	7	0	
d	6	5	0

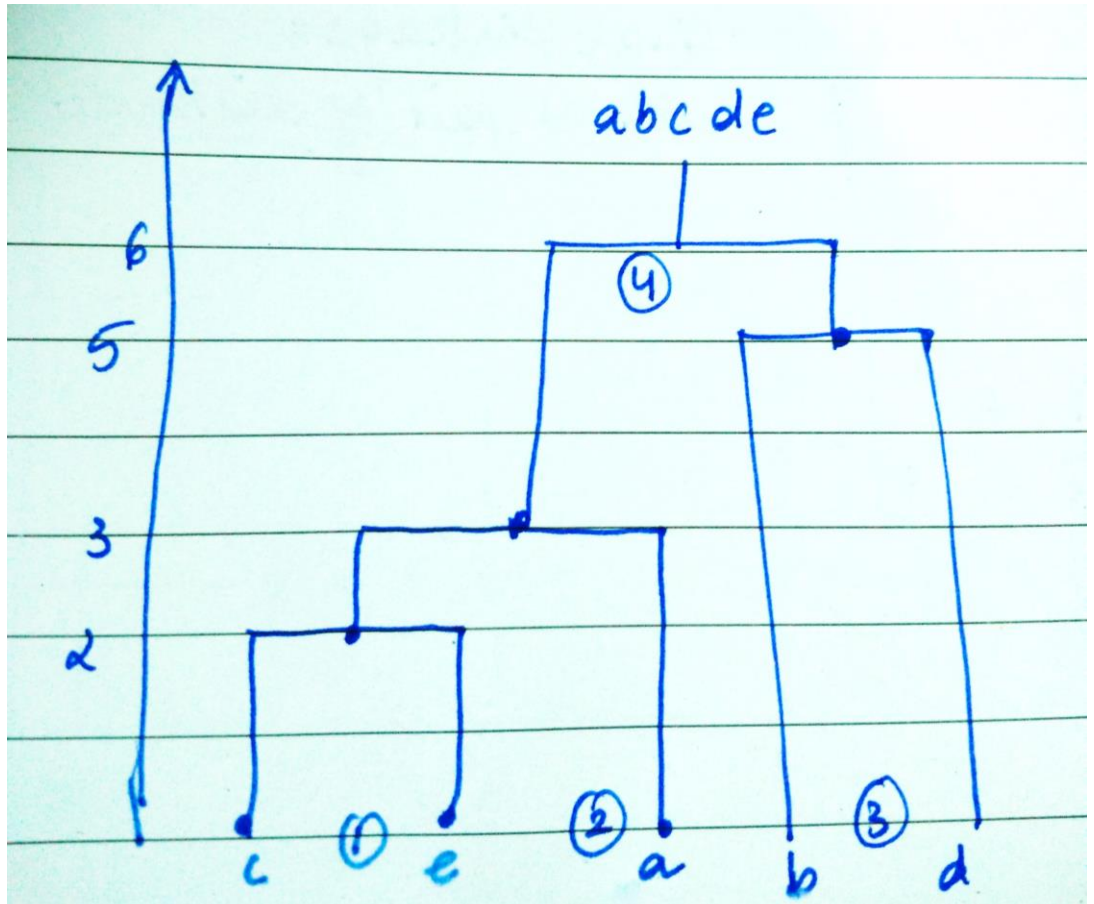
	cea	bd
cea	0	
bd	6	0

Min (b cea, d cea)

Clustering Methods

Prototype-based Clustering

Hierarchical Clustering



ce	2
ace	3
bd	5
abcde	6

Clustering Methods

Prototype-based Clustering

Hierarchical Clustering

Agglomerative clustering

Single Linkage – look for min distance – consider min distance for creation next matrix

Average Linkage – look for min distance – consider average distance for creation next matrix

Complete Linkage – look for min distance – consider max distance for creation next matrix