Data Analytics with R

Clustering

Business Scenario

- Philip works for an entertainment company. The company wants to introduce a new show. However, before launching the show, they would like to gauge the possible target audience of the show. Philip has been asked to identify the target audience for the show.

Approach: To identify the target audience, Philip needs to create customer segments based on the customer behavior. To perform the task, he needs to understand how to create segments of the customer and how to perform clustering, which is the process used to create segments.

Learning Objectives

By the end of this lesson, you will be able to:

- Perform clustering
- Differentiate classification and clustering algorithms
- List the types of clustering algorithms
- Use dimensionality reduction technique

Clustering Vs. Classification

Clustering

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

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

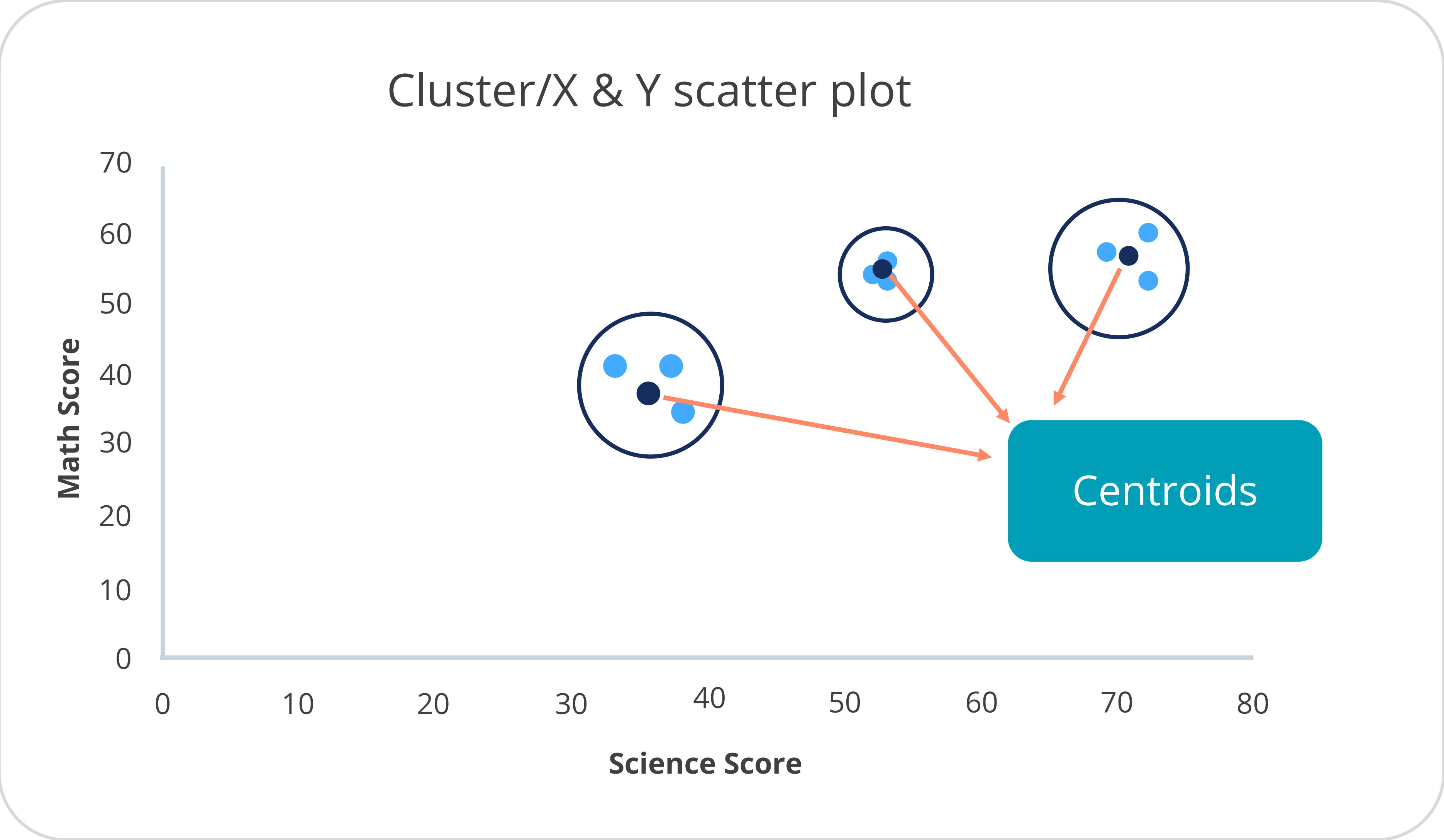
Clustering: Example

Consider a scenario where a cluster or group of students with similar aptitude needs to be created. 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

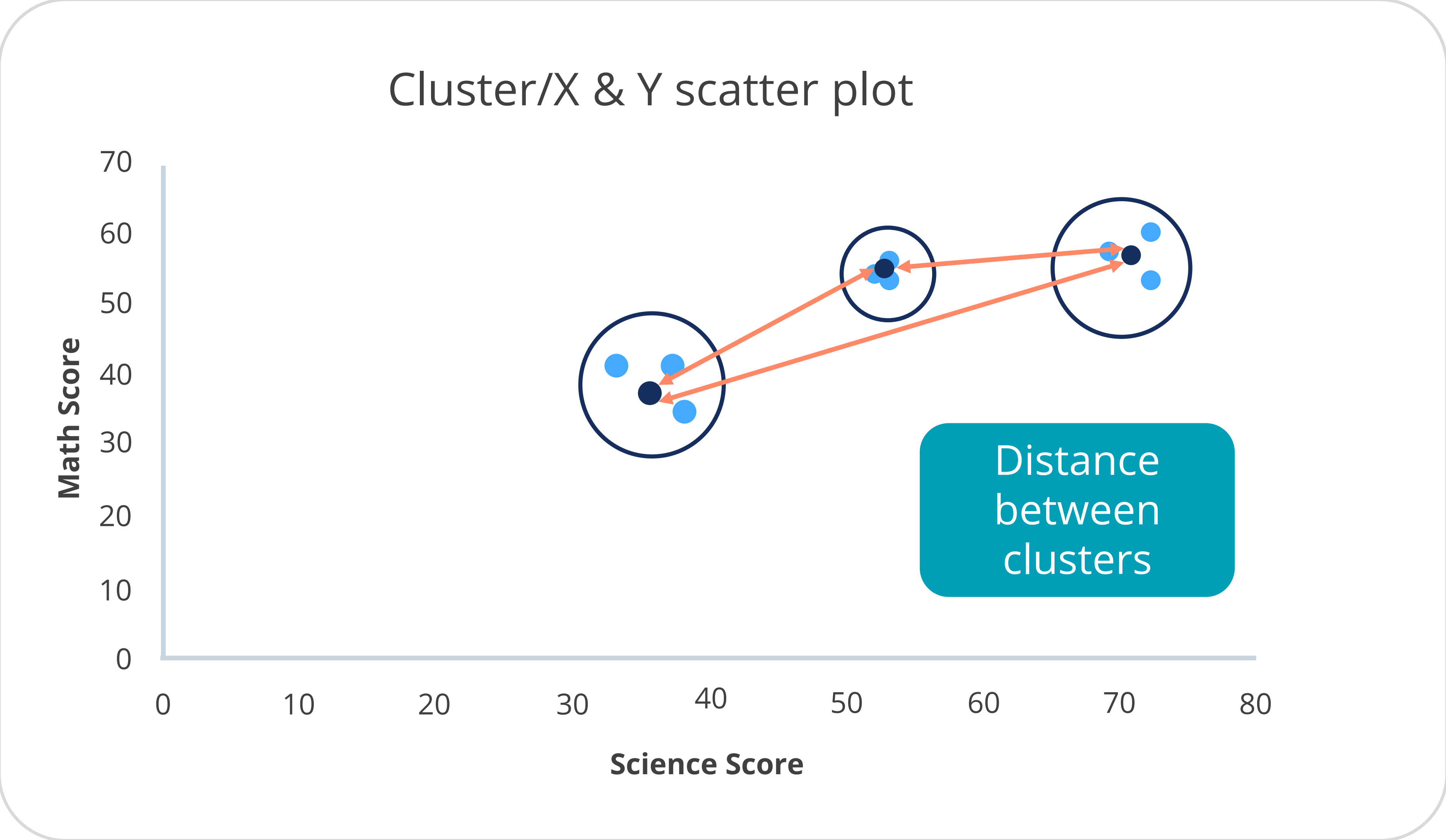
Clustering: Example

Each of these clusters has center points, called centroids.



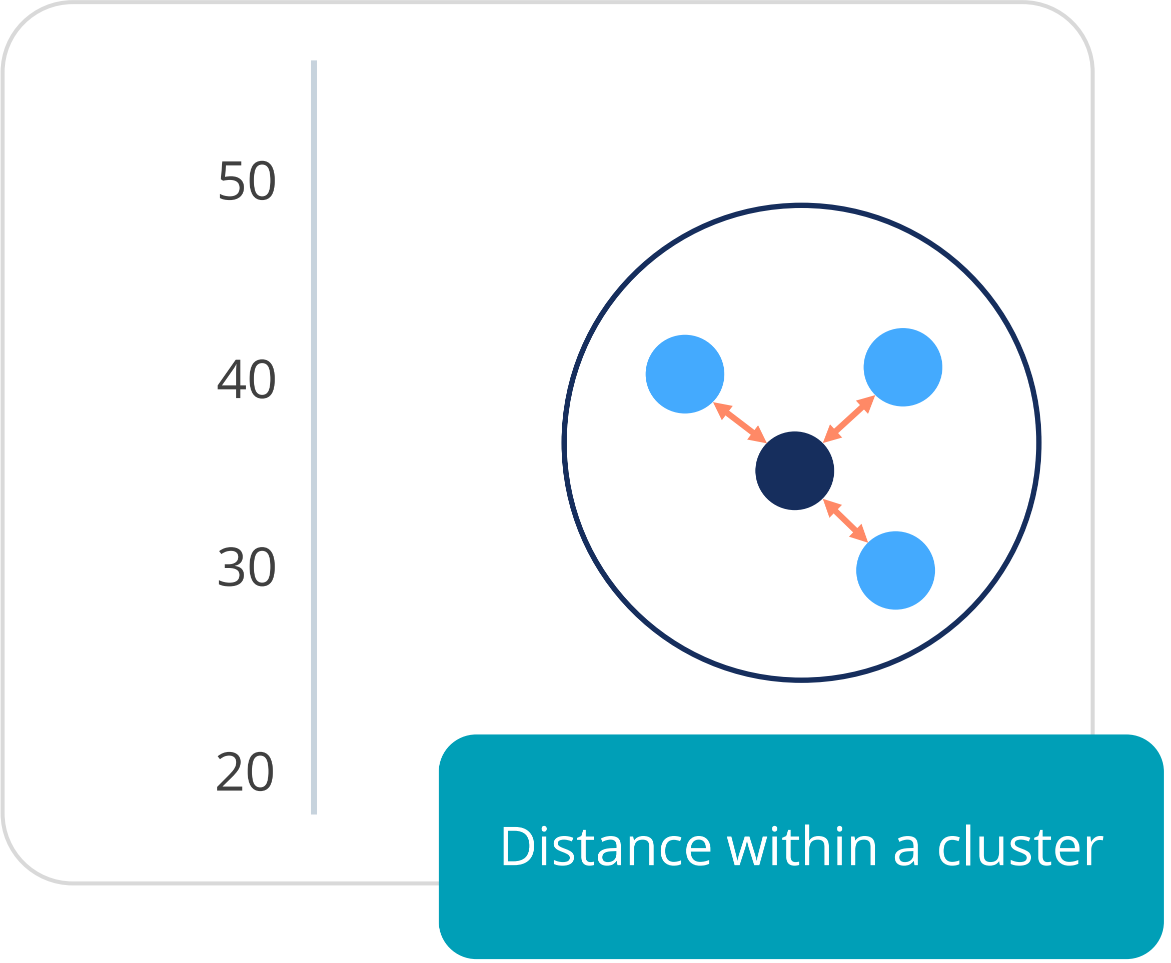
Clustering: Example

The distance between the cluster centroids is called the distance between clusters.



Clustering: Example

The average distance of observation in a cluster from its cluster centroid is called the distance within the cluster.



Distance within a cluster

Other Examples of Clustering

- Grouping the content of a website or product in a retail business.
- Segmenting customers or users into different groups based on 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.

Clustering vs. Classification

Clustering	Classification
This is an unsupervised learning technique, that is, the class label must be learned.	This a supervised learning technique, that is, the class label is known.
The goal of a clustering algorithm is to identify similar groups of observations using attribute data based on similarity and density of these data points.	The goal of classification is to predict labels for a set of attribute data based on the rules estimated from a training dataset consisting of attributes and a class label.
Clustering algorithms includes K-means, hierarchical, and DBSCAN.	Classification algorithms include K-nearest neighbors, decision trees, Naïve Bayes, and SVM.

Clustering Methods

Clustering Methods

01 Prototype-Based Clustering

02 Hierarchical Clustering

03 Density-Based Clustering (DBSCAN)

K-Means Clustering

Prototype-Based Clustering

Prototype-based clustering

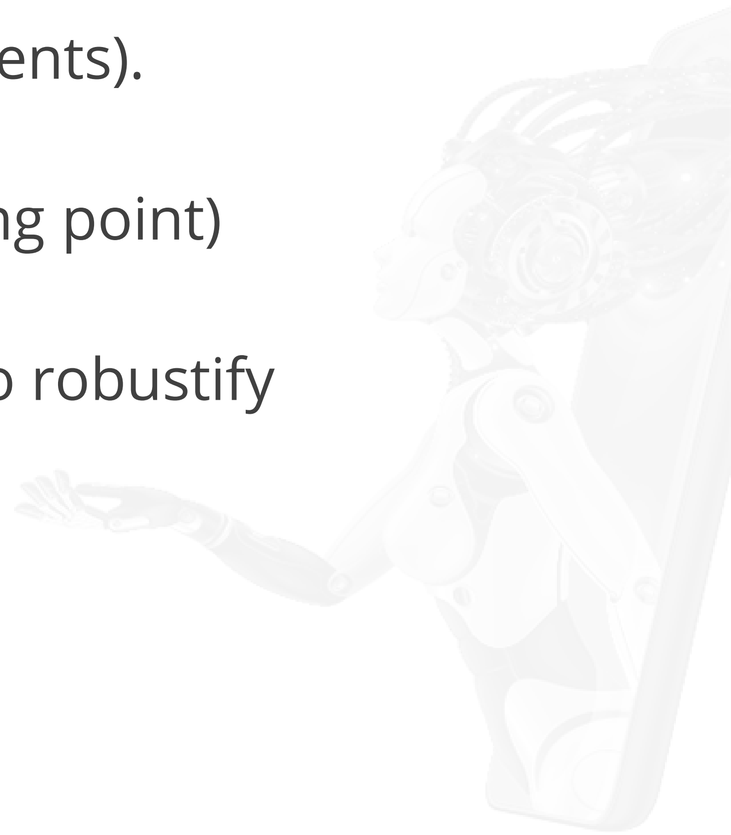
Hierarchical clustering

Density-based clustering (DBSCAN)

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

Example: Centroid (average) or medoid (most frequently occurring point)

It is widely used in banking and to predict sport-based statistics to robustify one's efforts based on statistics.



Prototype-Based Clustering

- **Prototype-based clustering**
- Hierarchical clustering
- Density-based clustering (DBSCAN)

K-means is a prototype-based method. It is the most popular method for clustering. It involves:

Assigning training data to match clusters based on similarity.

Iterative process to get data points in the best clusters possible.

Steps in K-Means Clustering

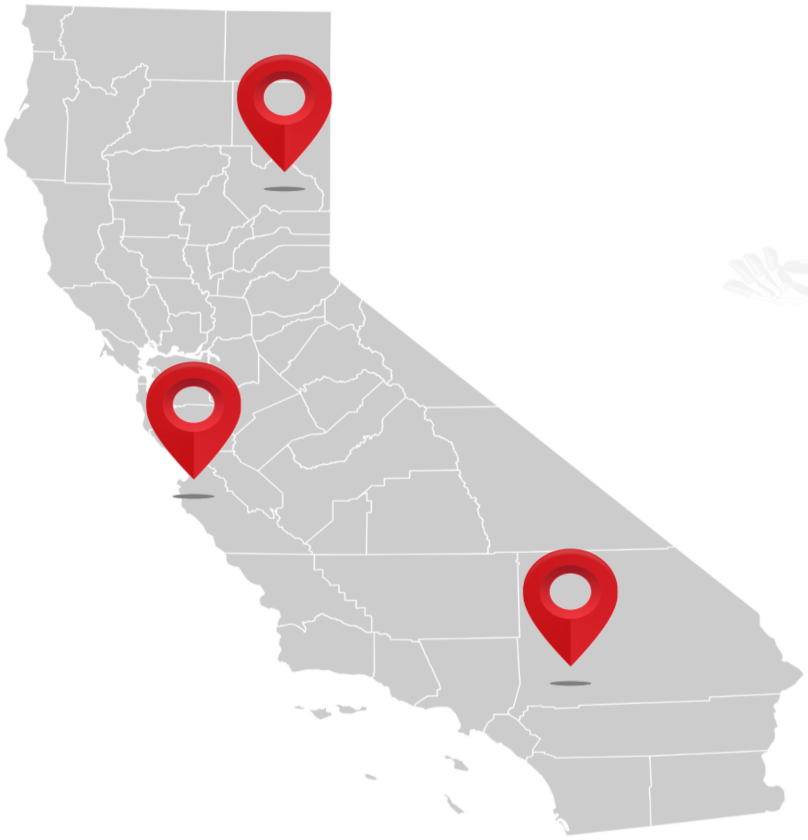
Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

K-means clustering example: The government of California wants to identify high density clusters of people to build hospitals. No other ground truth or features are provided apart from the population data.

How can the clusters be identified?



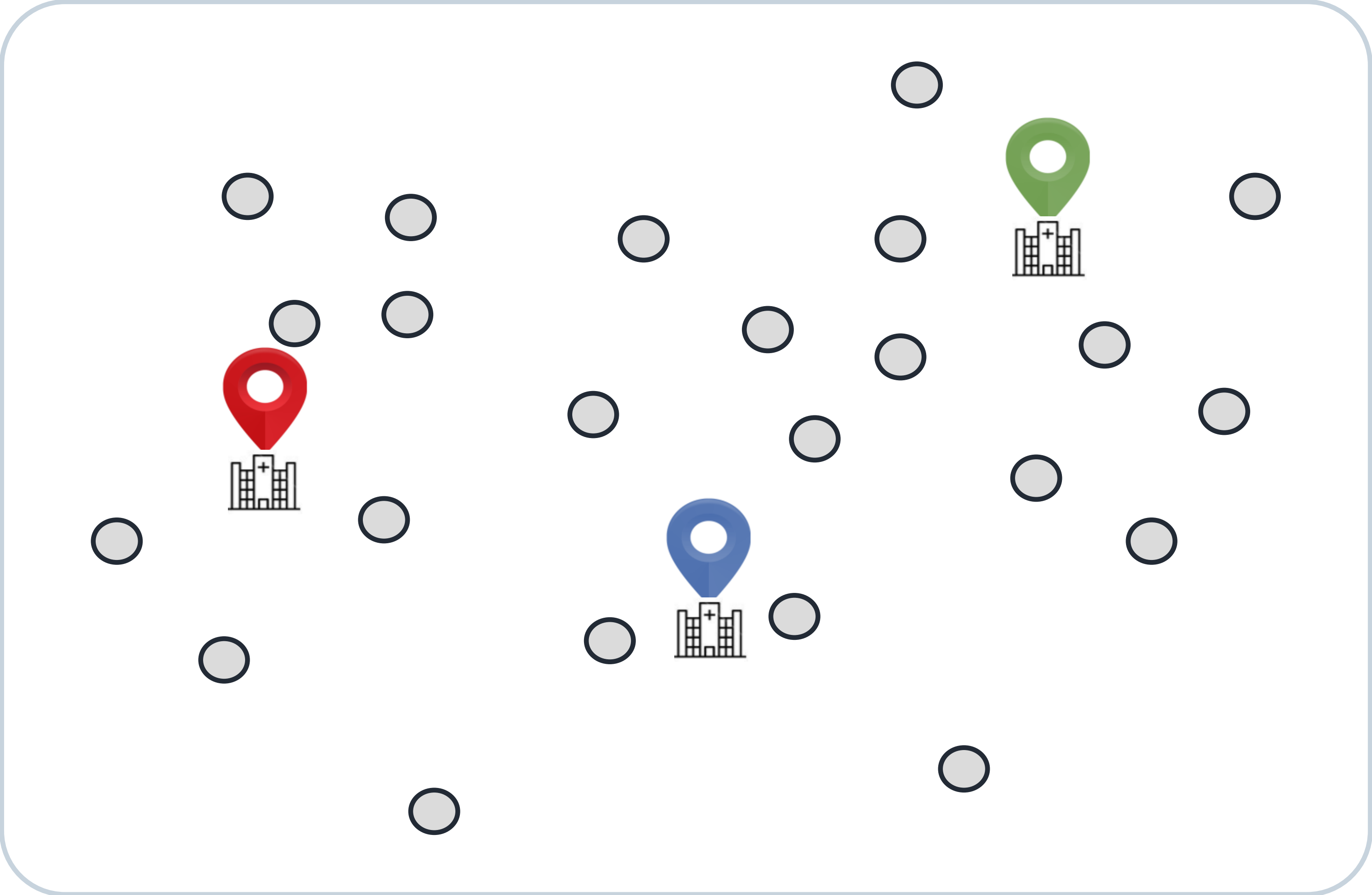
Steps in K-Means Clustering

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

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



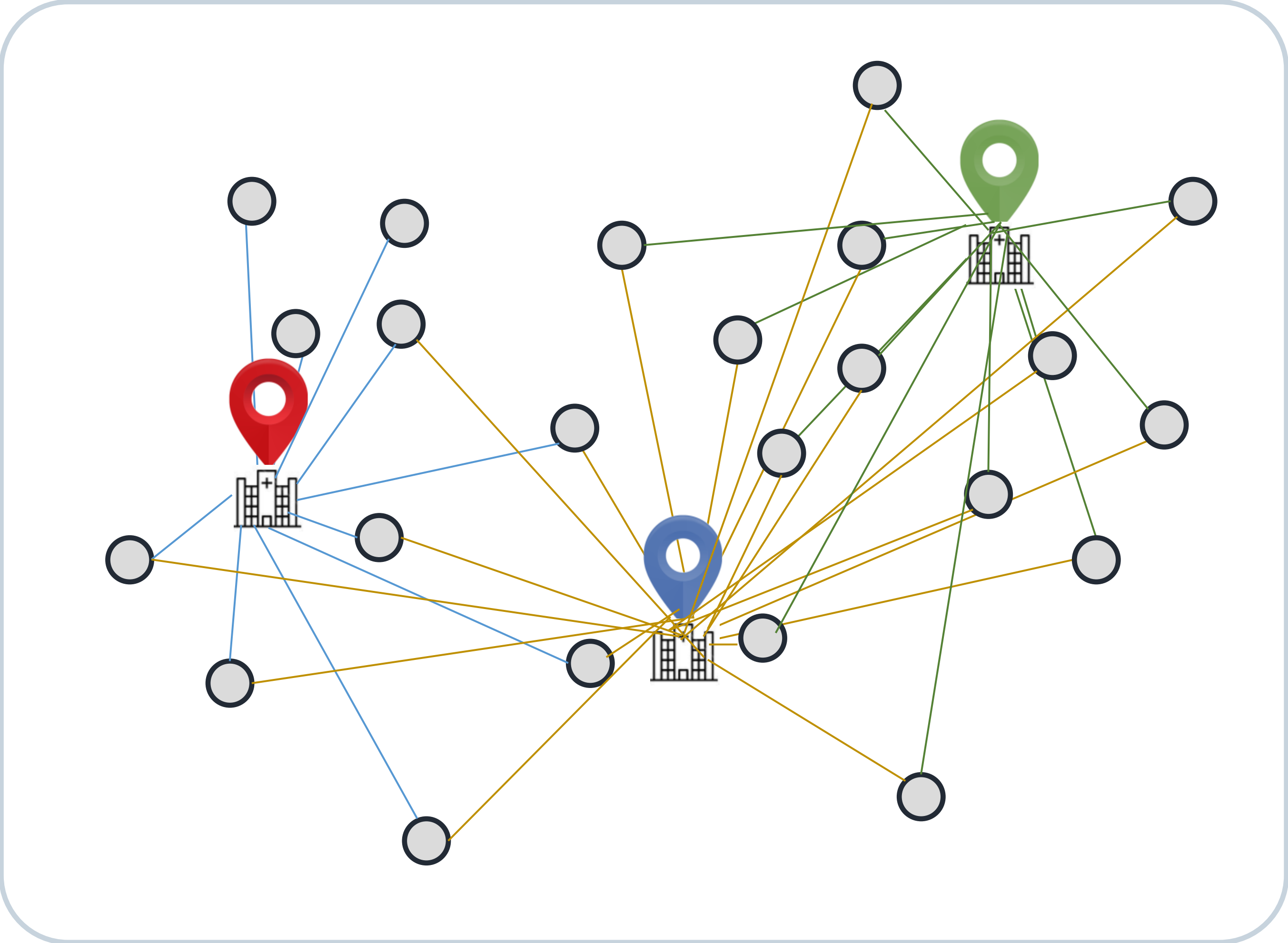
Steps in K-Means Clustering

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

Assign each point to the nearest centroid.



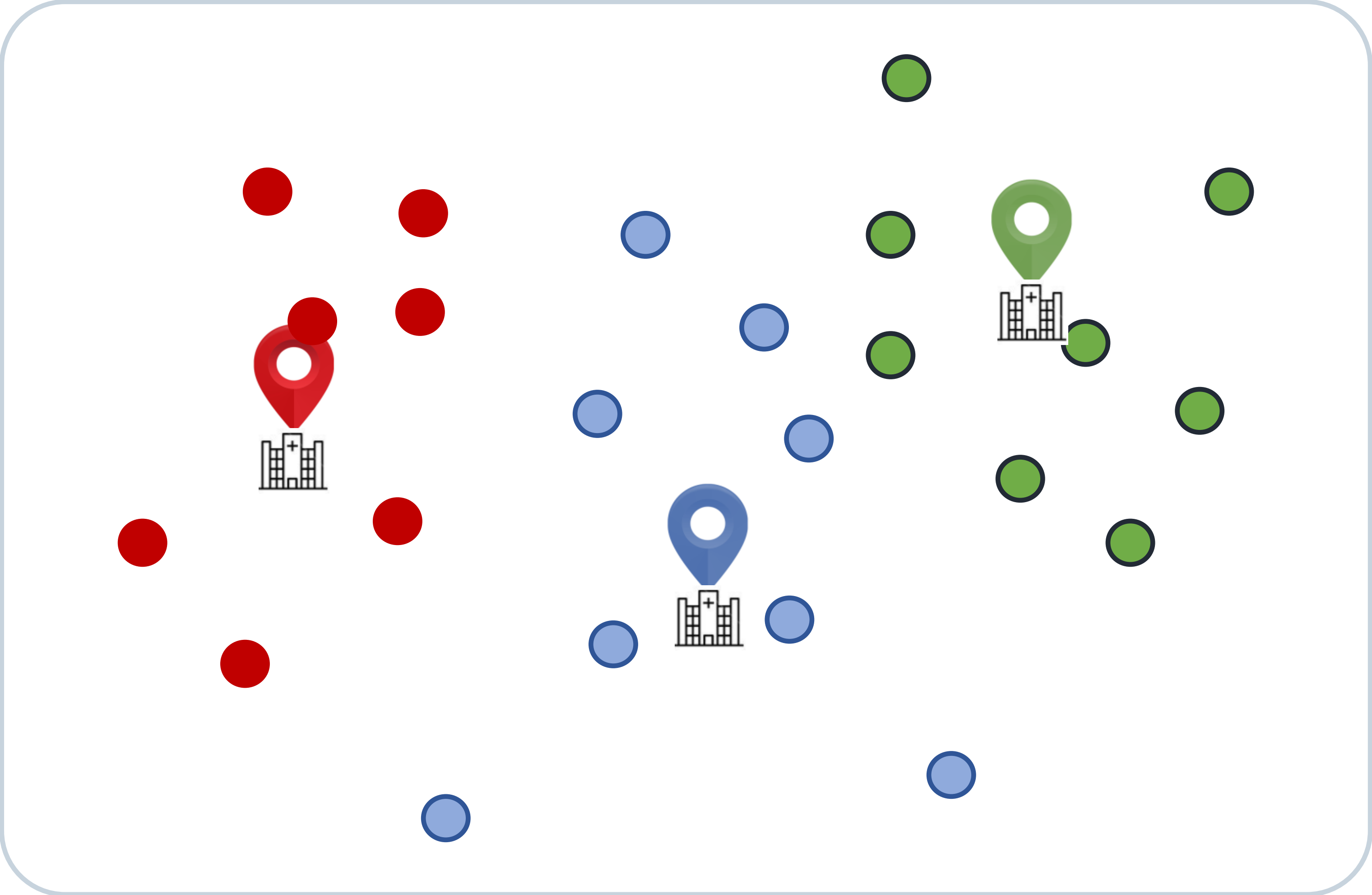
Steps in K-Means Clustering

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

Move each centroid to the center of the respective cluster.



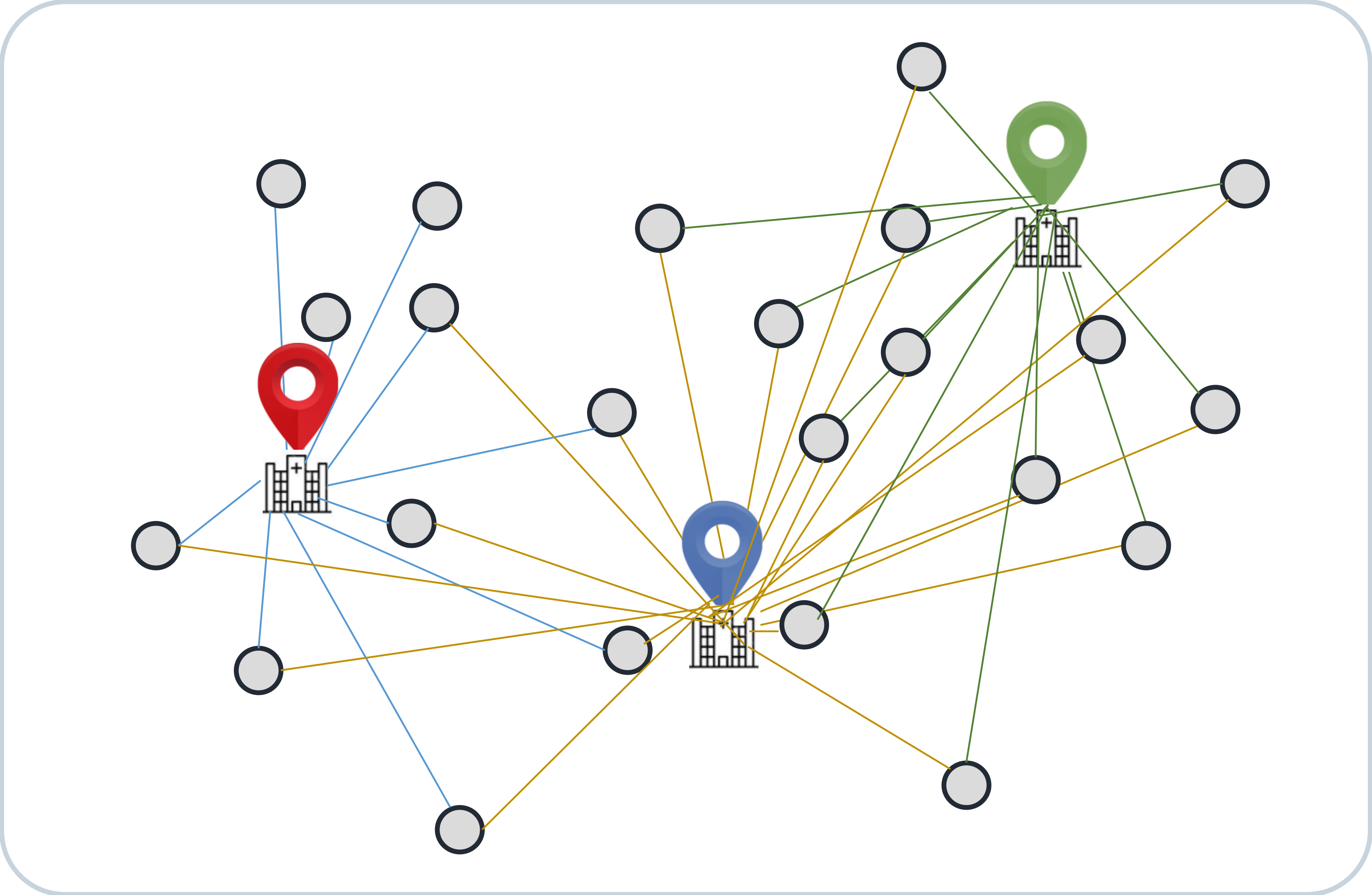
Steps in K-Means Clustering

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

Calculate the distance of the centroids from each point again.



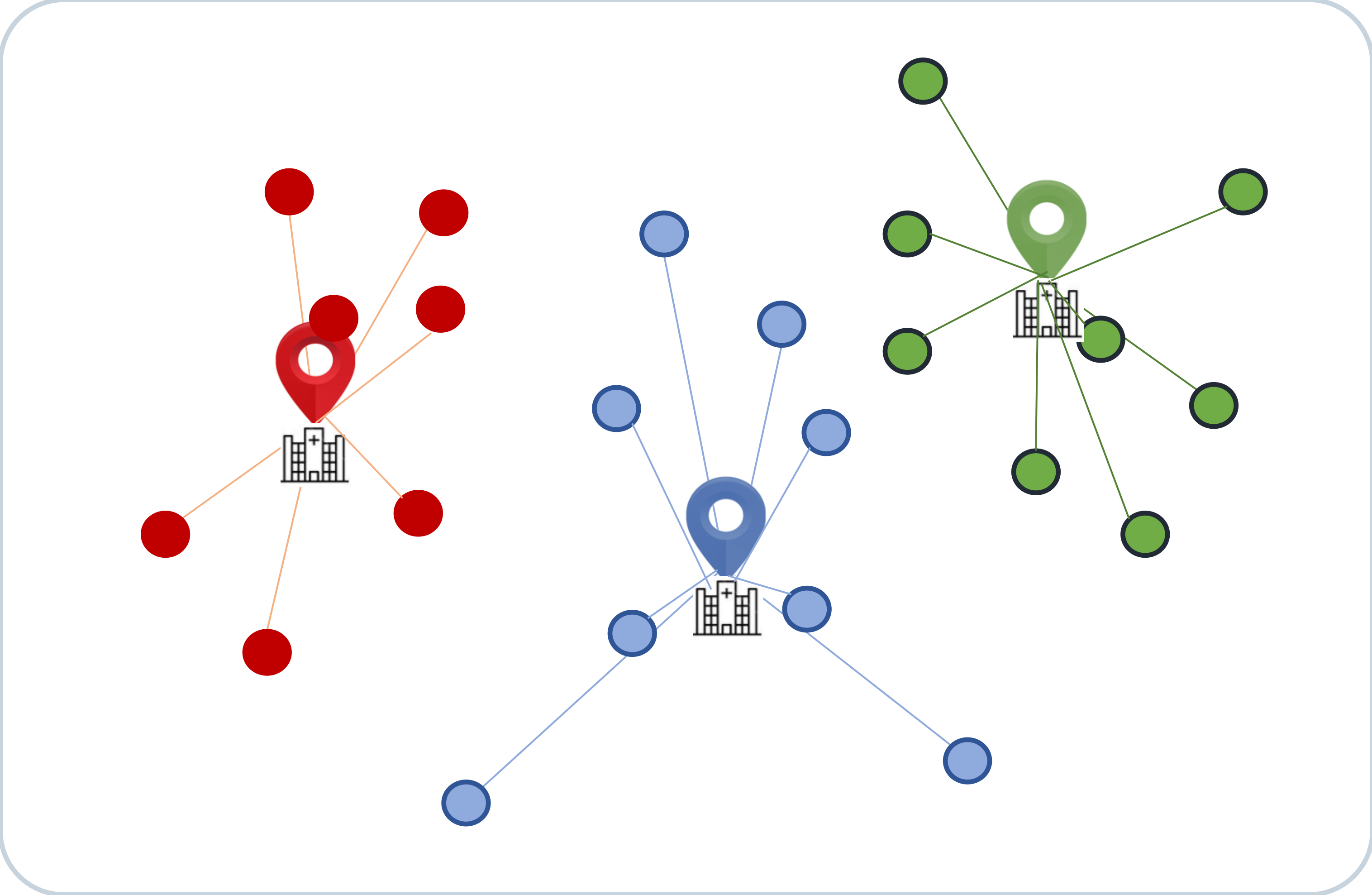
Steps in K-Means Clustering

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

Move points across clusters and recalculate the distance from the centroid.



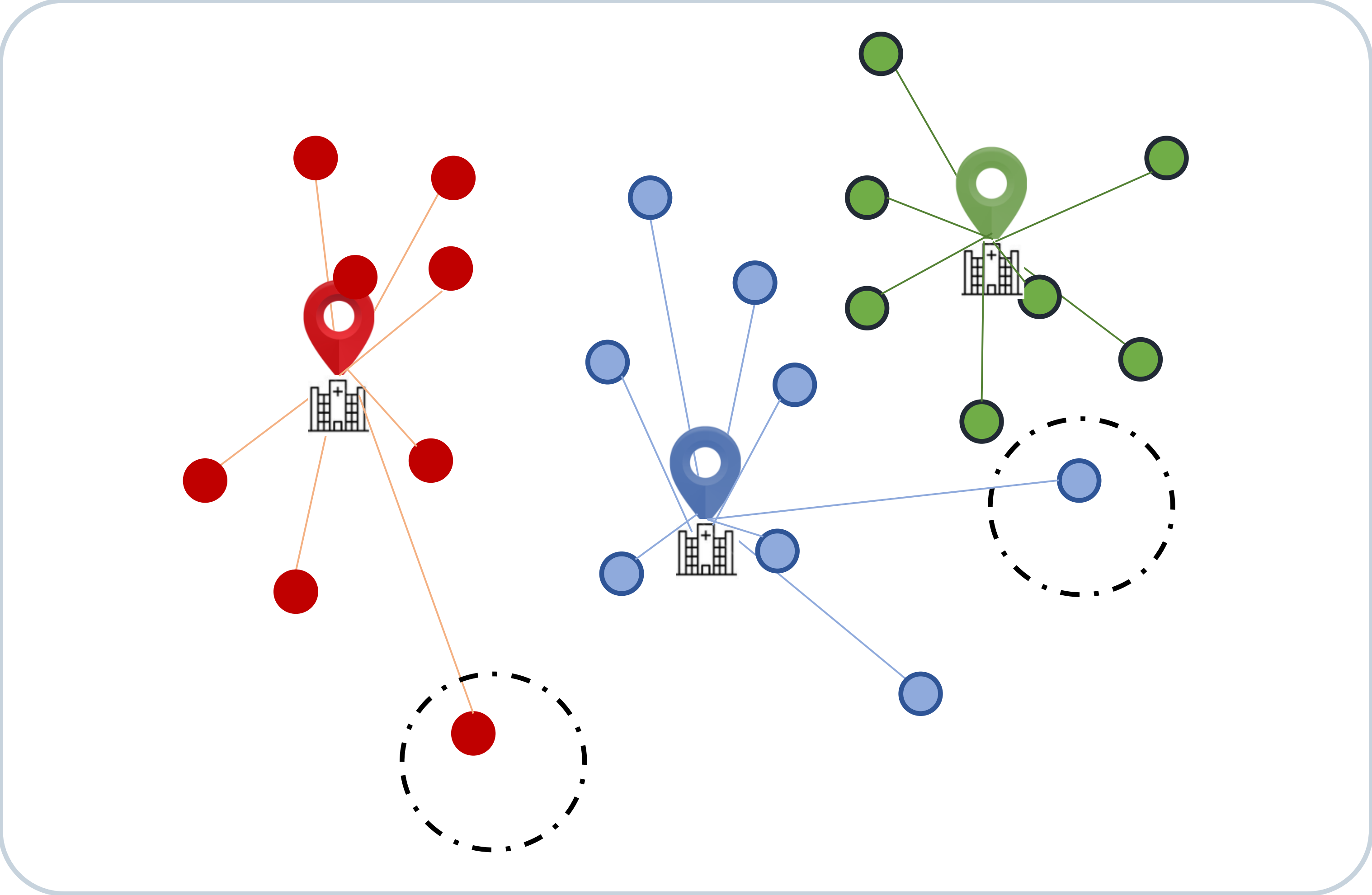
Steps in K-Means Clustering

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

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



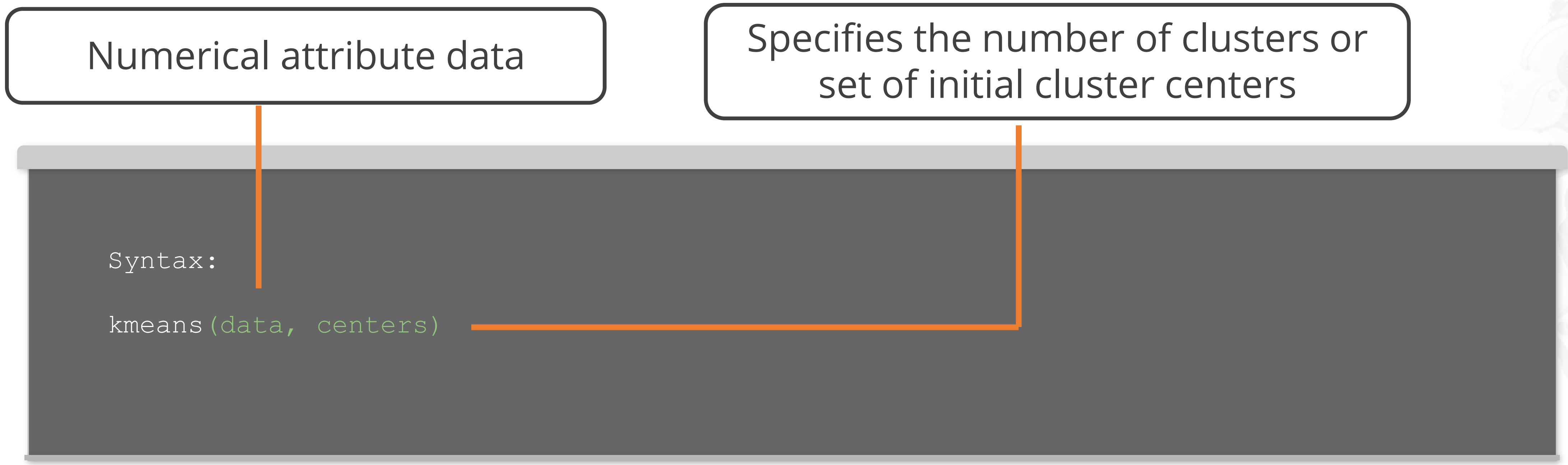
Steps in K-Means Clustering

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

In R, K-Means clustering is done by using the kmeans() function.



Numerical attribute data

Specifies the number of clusters or set of initial cluster centers

```
Syntax:

kmeans(data, centers)
```

The function returns the cluster label vector corresponding to each data observation, a matrix for all cluster centers, and the squared deviations.

Perform Clustering Using K-Means

Duration: 10 minutes

Problem Scenario: Jenny wants to apply for college, but she does not want to apply for all colleges. Instead, she wants to find colleges that would suit her requirements

The ***University.csv*** file lists the top 25 universities in the US along with their accepted SAT scores, SF ratio, and other parameters.

Help Jenny identify two groups of college clusters that fit her requirements using K-means clustering.

Note: Please download the data set and the solution document from the **Course Resources** section and follow the steps given in the document

Hierarchical Clustering

Hierarchical Clustering

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

It clusters n units or 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.

Hierarchical Clustering

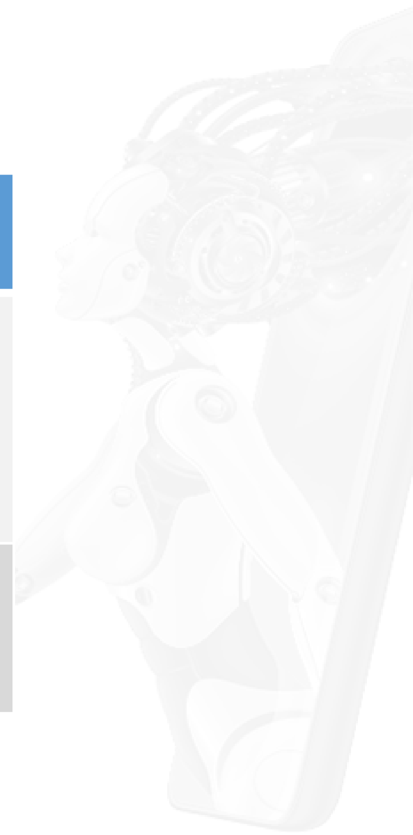
Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

There are 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



Hierarchical Clustering

Prototype-based clustering

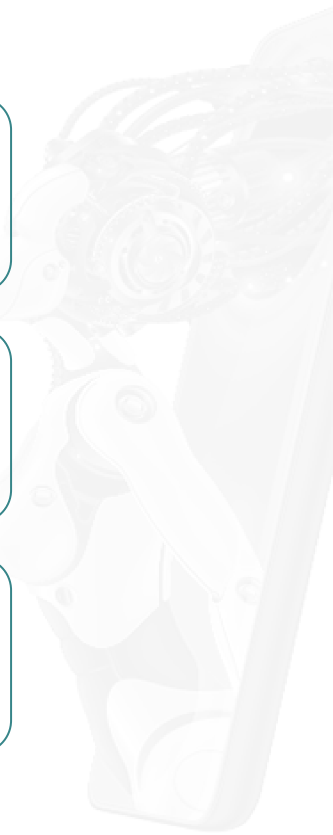
Hierarchical clustering

Density-based clustering (DBSCAN)

Agglomerative clustering is a process where:

- An n × 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.

For R package cluster, use the *agnes* function.
For stats package, use the *hclust* function.



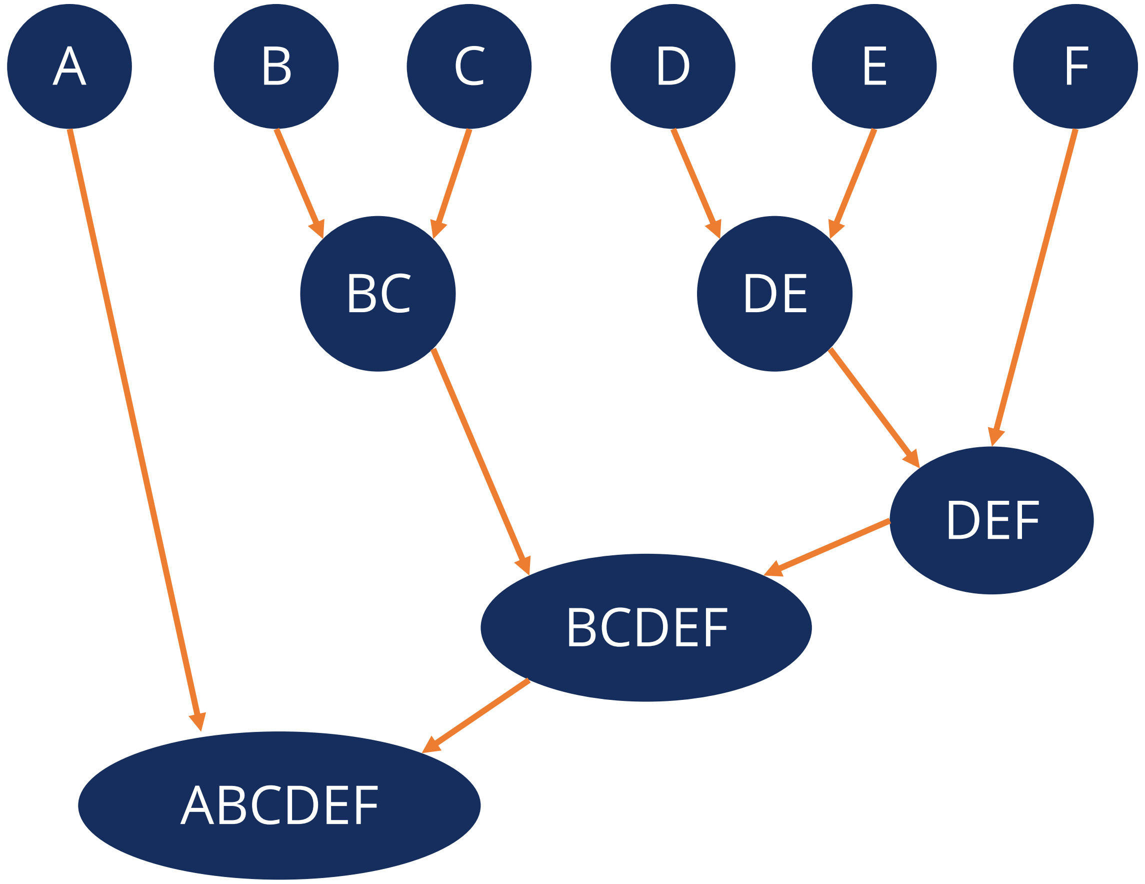
Hierarchical Clustering

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

Divisive clustering is a top-down clustering approach and is used in practical applications of image retrieval.



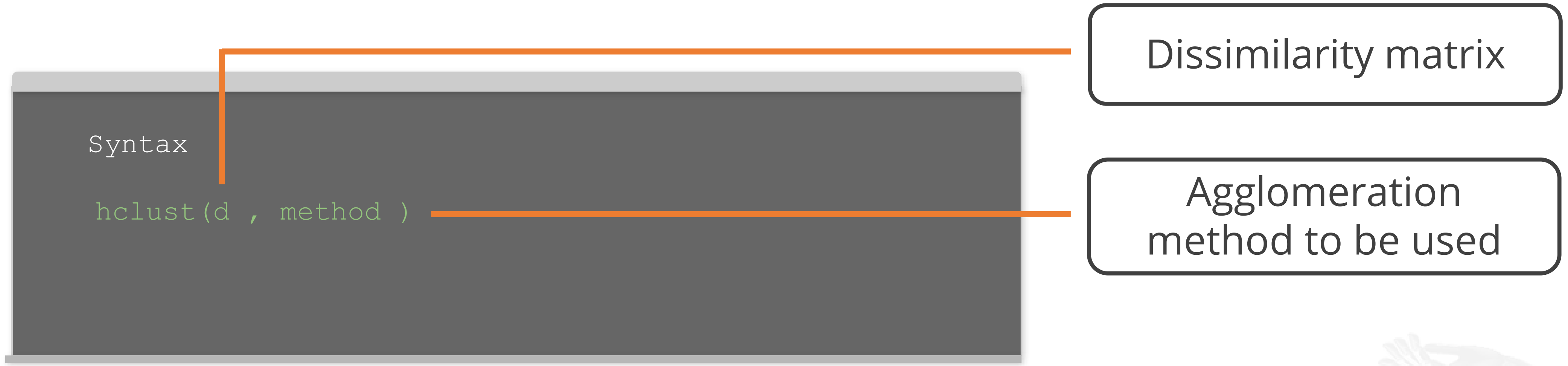
Hierarchical Clustering

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

Hierarchical clustering in R can be performed using hclust() function.



Syntax

```
hclust(d , method )
```

Dissimilarity matrix

Agglomeration method to be used

A dissimilarity matrix can be created by using dist() function.

- The output of hclust function is an object that describes the tree produced by the clustering process.
- The hclust object can be used with the plot function to create the dendrogram.

Perform Hierarchical Clustering

Duration: 10 minutes

Problem Scenario: Jenny wants to apply for college, but she does not want to apply for all colleges. Instead, she wants to find colleges that would suit her requirements.

The ***University.csv*** file lists the top 25 universities in the US along with their accepted SAT scores, SF ratio, and other parameters. Help Jenny identify colleges that fit her requirements using hierarchical clustering.

Note: Please download the data set and the solution document from the **Course Resources** section and follow the steps given in the document

DBSCAN Clustering

DBSCAN

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

DBSCAN (Density-Based Spatial Clustering and Application with Noise) is used to identify clusters of any shape in a data set containing noise and outliers.

DBSCAN algorithms use density to create clusters, that is, it groups closely-packed data observations together to form clusters.

DBSCAN

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

DBSCAN algorithm creates clusters using two important parameters.

Epsilon or eps is the maximum distance between two points belonging to the same cluster.

Minimum points or minPts is the minimum number of data points to form a cluster.

DBSCAN

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

DBSCAN algorithm creates a radius around every data point to classify them as:

Board point: It is the data point that has at least one core point at a distance of epsilon from it.

Core point: It is the data point which has the minimum number of points (minPts) or more within epsilon distance from it.

Noise point: These are outliers. These are neither core nor border points and have less than the minPts within the epsilon distance from it.

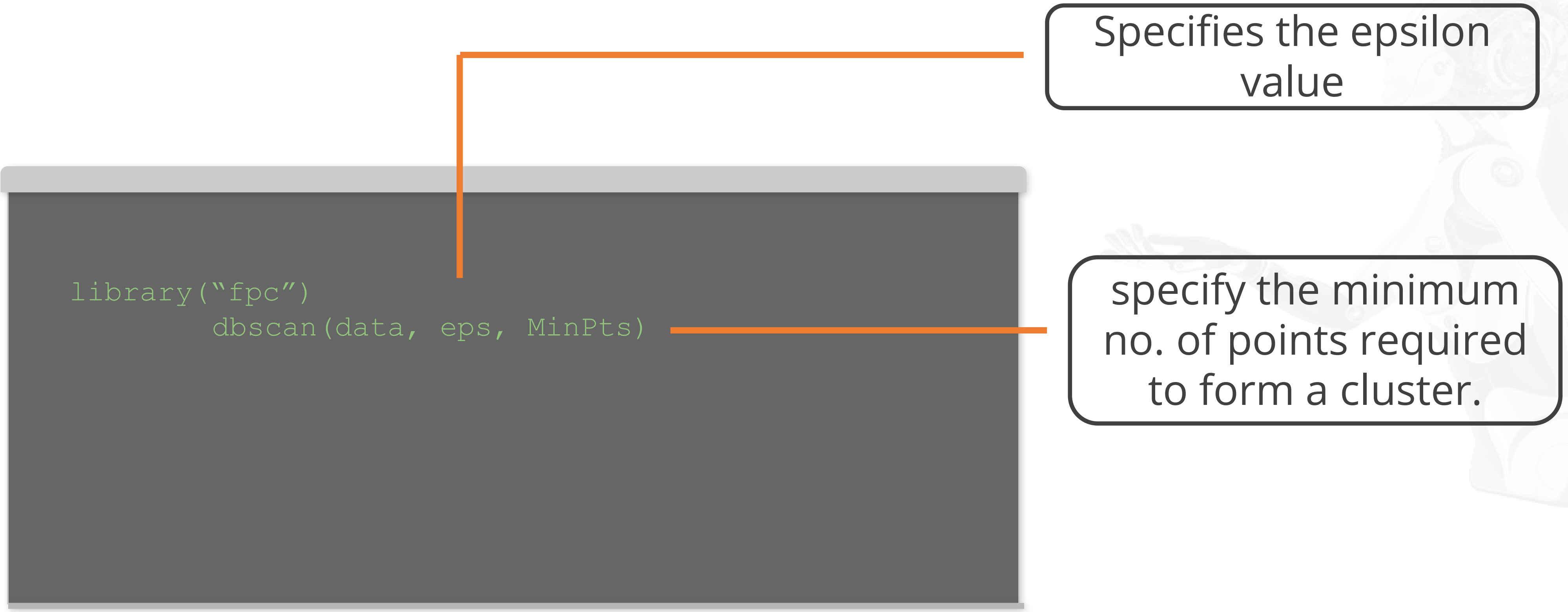
DBSCAN

Prototype-based clustering

Hierarchical clustering

Density-based clustering (DBSCAN)

In R, DBSCAN clustering is performed using dbscan() function from the fpc package.



```
library(“fpc”)
        dbscan(data, eps, MinPts)
```

Specifies the epsilon value

specify the minimum no. of points required to form a cluster.

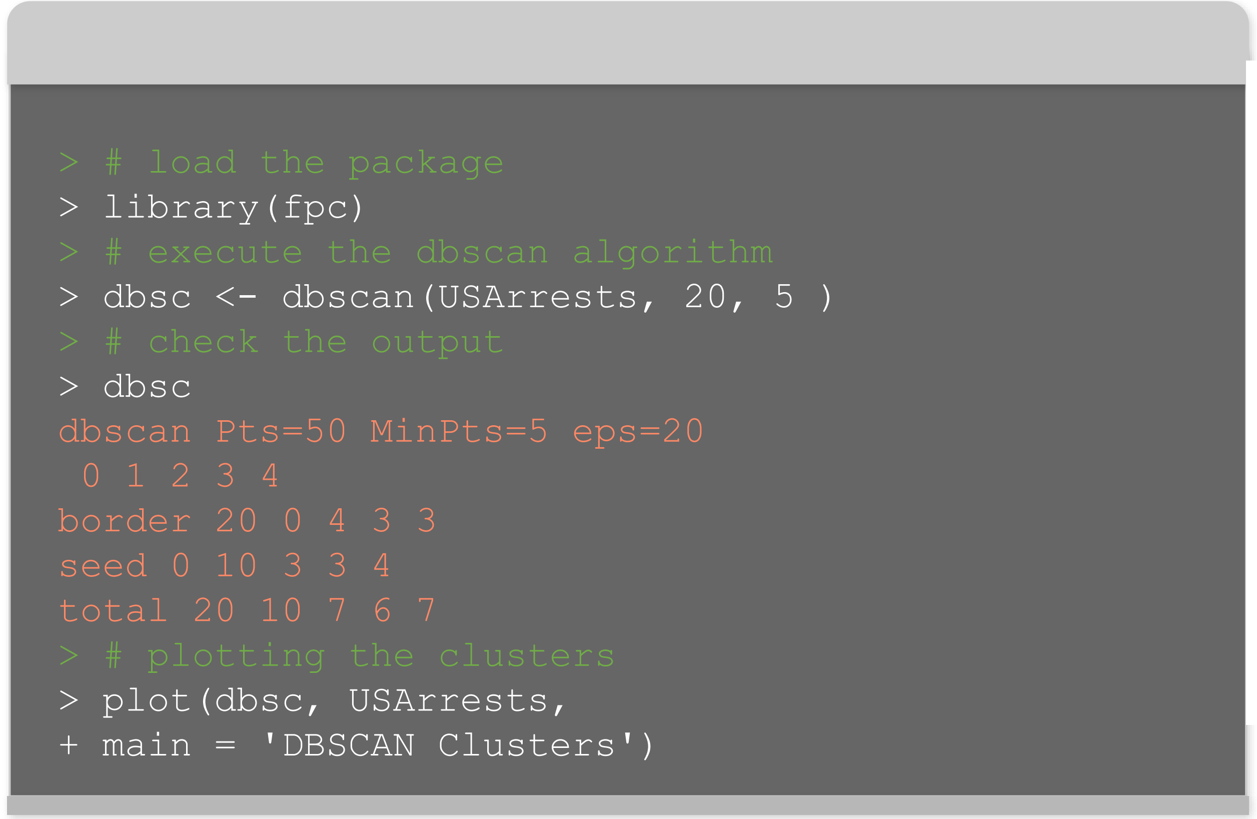
DBSCAN

Prototype-based clustering

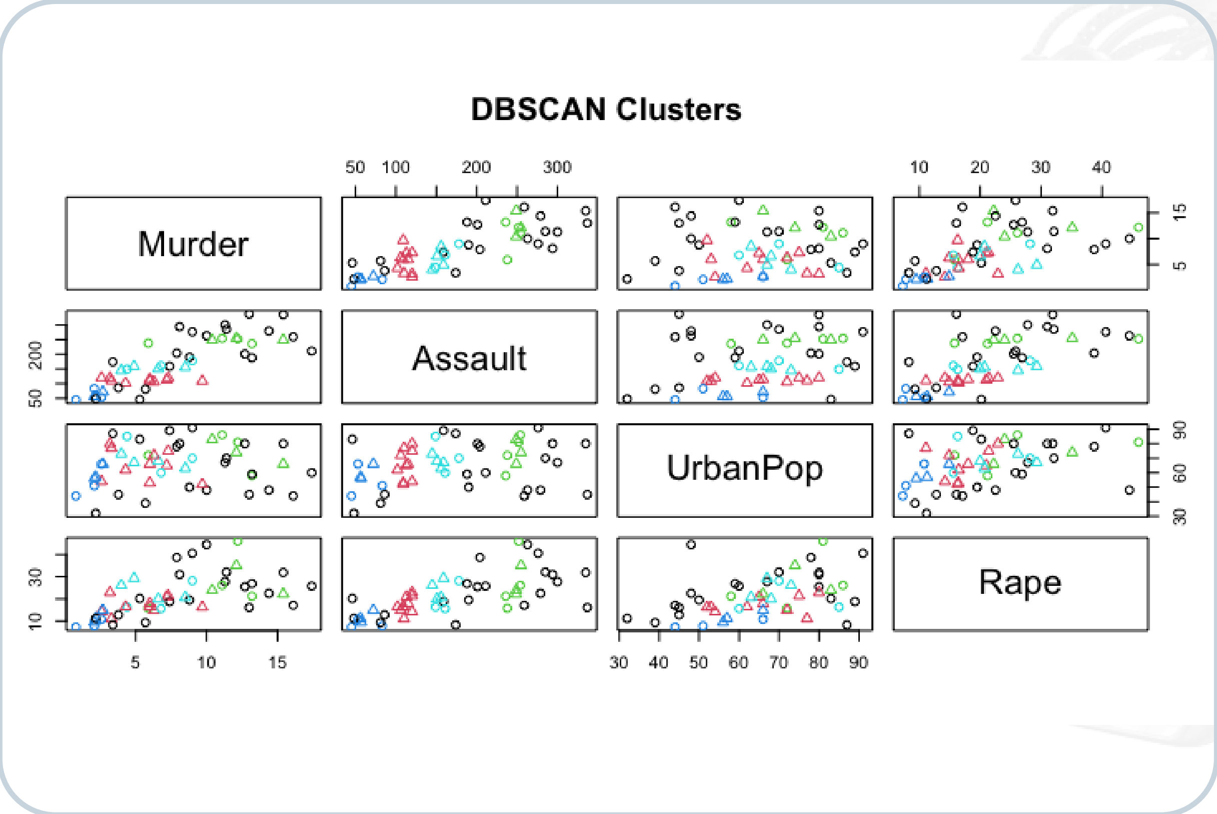
Hierarchical clustering

Density-based clustering (DBSCAN)

DBSCAN clustering performed on USArrests data from R datasets.



```
> # load the package
> library(fpc)
> # execute the dbscan algorithm
> dbsc <- dbscan(USArrests, 20, 5 )
> # check the output
> dbsc
dbscan Pts=50 MinPts=5 eps=20
  0  1  2  3  4
border 20  0  4  3  3
seed   0  10  3  3  4
total  20  10  7  6  7
> # plotting the clusters
> plot(dbsc, USArrests,
+ main = 'DBSCAN Clusters')
```



Principal Component Analysis

Principal Component Analysis: Scenario

A medical student wants to use a classification algorithm to predict the probability of occurrence of a disease using the features extracted from a scan. The feature extraction algorithm identifies 150 numerical features from the scan.

Using all the 150 attributes will increase the complexity level of the model and might also introduce redundancy in the final model.

Reducing the number of attributes is important.

Principal Component Analysis

Principal component analysis (PCA) is a dimensionality reduction technique.

Dimensionality reduction algorithms are used to reduce the number of variables being considered in a classification model.

PCA achieves dimensionality reduction by transforming the original data to a smaller set of principal components such that the variance captured is maximized.

Steps for Extracting Principal Component

Step 1: Select the numerical attributes (predictor) columns and create a standardized set

Step 2: Compute the covariance or correlation matrix to identify associations

Step 3: Use eigen value decomposition method to extract eigen values and corresponding eigen vectors

Step 4: Identify the optimum number of components to be used from a scree plot (Scree plot plots number of PCs vs. percentage of variance captured)

Step 5: Re-create the data using the selected principal component

Principal Components

Principal components are linear components of the original variables. They tend to capture as much variance as possible in a dataset.

$$PC_1 = \varnothing_1' * X_1 + \varnothing_2' * X_2 + \varnothing_3' * X_3 + \varnothing_4' * X_4 + \ldots.\ \varnothing_n' * X_n$$

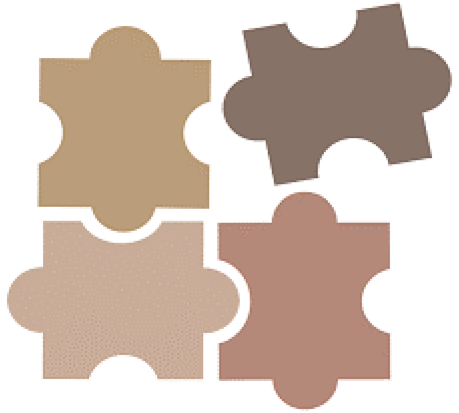
$$PC_2 = \varnothing_1'' * X_1 + \varnothing_2'' * X_2 + \varnothing_3'' * X_3 + \varnothing_4'' * X_4 + \ldots.\ \varnothing_n'' * X_n \text{ and so on.}$$

where $\varnothing_1'$ is the factor loading of variable X1 in principal component 1 and $\varnothing_1''$ is the factor loading of variable X1 in principal component 2.

The total number of PCs created is same as the number of original variables.

Principal Components

Each of the principal components is orthogonal to the other and has zero correlation, which solves the problem of multicollinearity.



The first principal component captures the maximum variance determining the direction of higher variance.

The remaining extracted principal components capture the variance in a decreasing order.

PCA in R

In R, PCA is performed using the prcomp() function. The function returns an object containing these attributes:

```
Syntax : prcomp(x)
where x is the scaled data.
```

- ✓ **x:** Contains the transformed data or the principal components
- ✓ **sdev**: Standard deviation of each PC, that is the square root of corresponding eigen values
- ✓ **rotation**: Matrix of factor loadings

Finding Principal Components

Duration: 10 minutes

Problem Scenario: The ***cancer.csv*** data from Wisconsin contains 30 attributes describing the features from a digitized image of a breast mass classified as benign and malignant.

For effective analysis, apply the PCA technique to reduce the number of dimensions.

Note: Please download the data set and the solution document from the **Course Resources** section and follow the steps given in the document

ASSISTED PRACTICE

Key Takeaways

- Cluster analysis (clustering) is the most used technique of unsupervised learning to find data clusters such that each cluster has the most similar data.

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

- K-means clustering assigns training data to matching clusters based on similarity and uses an iterative process to get data points in the best clusters possible.

Key Takeaways

- Hierarchical clustering clusters n units or objects, each with p features, into smaller groups and creates a hierarchy of clusters as a dendrogram.
- DBSCAN (density-based spatial clustering and application with noise) is used to identify clusters of any shape in a dataset containing noise and outliers.
- PCA is a linear transformation technique for dimensionality reduction such that the variance captured by a smaller set of PC is maximum.

Knowledge Check

Knowledge Check
1

Identify the correct statement(s) about k-means clustering.

A. It attempts to partition a set of data points into k distinct clusters.

B. It allows the use of k-means in the R package named “stats.”

C. It yields single points scattered around the datasets as outliers.

D. It presents a hierarchy of clusters as a dendrogram.

Knowledge Check

1

Identify the correct statement(s) about k-means clustering.

A. It attempts to partition a set of data points into k distinct clusters.

B. It allows the use of k-means in the R package named “stats.”

C. It yields single points scattered around the datasets as outliers.

D. It presents a hierarchy of clusters as a dendrogram.

The correct answers are **A and B**

The k-means clustering technique attempts to partition a set of data points into k distinct clusters and allows the use of k-means in the R package named “stats.”

Knowledge Check

2

Which of the following statements is/are true for hierarchical clustering?

A. It is inspired by the natural clustering approach.

B. Its algorithms are only of the agglomerative type.

C. It yields single points scattered around the datasets as outliers.

D. It presents a hierarchy of clusters as a dendrogram.

Knowledge Check

2

Which of the following statements is/are true for hierarchical clustering?

A. It is inspired by the natural clustering approach.

B. Its algorithms are only of the agglomerative type.

C. It yields single points scattered around the datasets as outliers.

D. It presents a hierarchy of clusters as a dendrogram.

The correct answer is **D**

Hierarchical clustering presents a hierarchy of clusters as a dendrogram.

Knowledge Check

3

Which of the following statement is/are True?

A. PCA is a linear transformation technique.

B. The total number of PC extracted is always less than the total number of original variables.

C. There is zero correlation between the PC.

D. The variance captured by the first PC is maximum.

Knowledge Check

3

Which of the following statement is/are True?

A. PCA is a linear transformation technique.

B. The total number of PC extracted is always less than the total number of original variables.

C. There is zero correlation between the PC.

D. The variance captured by the first PC is maximum.

The correct answer is **C and D**

There is zero correlation between the PCs. Hence, it eliminates the issue of multicollinearity. The first PC always captures the maximum variance.