Unsupervised Learning: Clustering

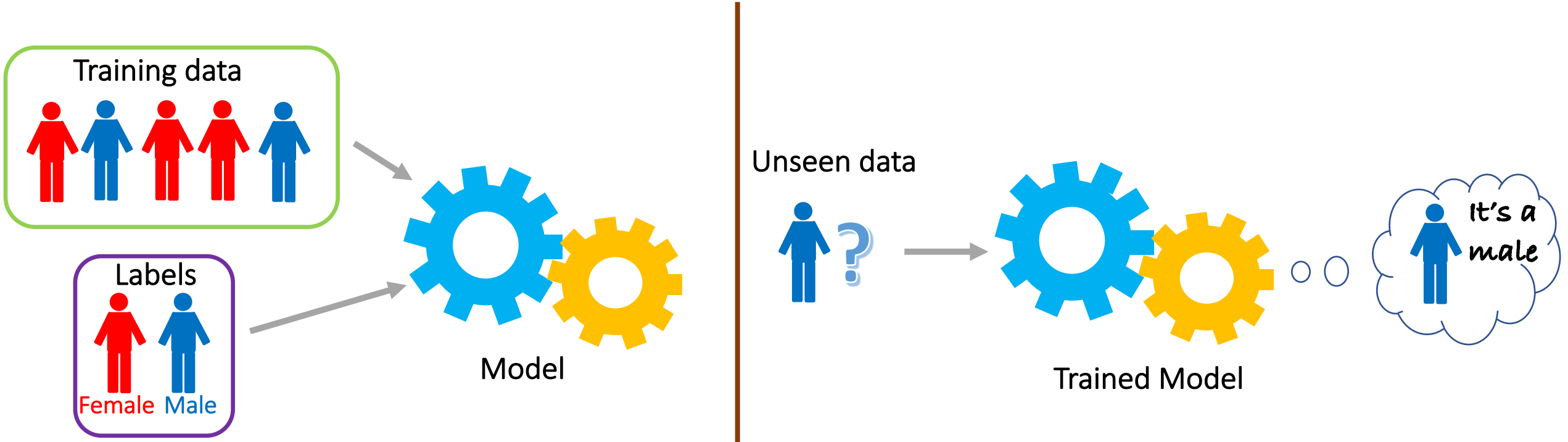
Alfonso Monaco – INFN Bari

Supervised and unsupervised learning

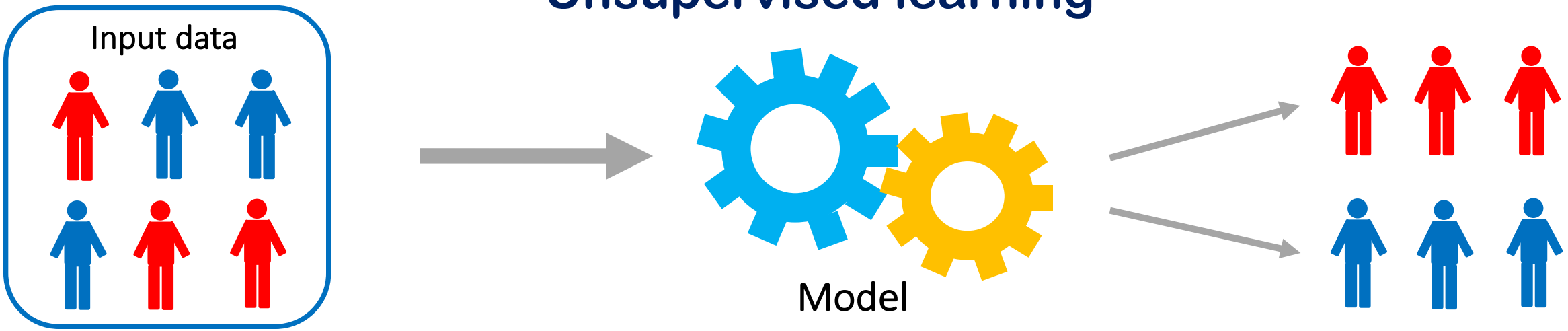
- Supervised learning is done when we get labeled data. We can use this data to train a machine learning model.

- Unsupervised learning is when we just have raw data and are expected to come up with insights without any labels. The system has the ability to understand the data and recognize patterns in it without explicitly being told what patterns to identify.

Supervised learning



Unsupervised learning

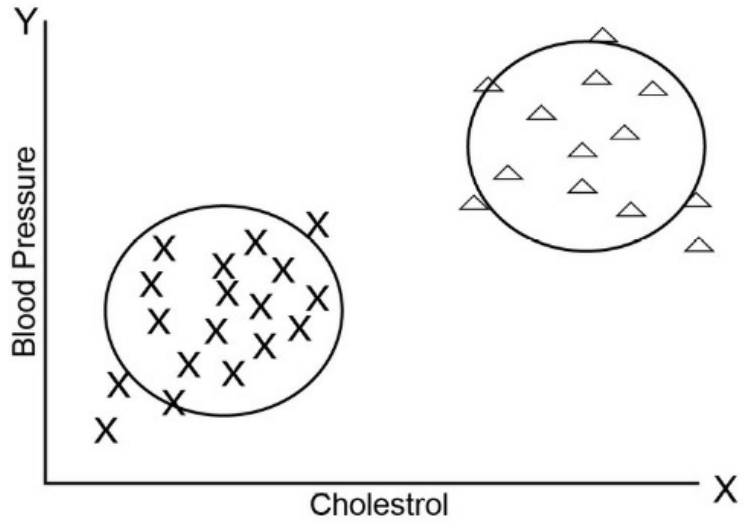


Clustering

- The most basic type of unsupervised learning is clustering.
- Definition: Clustering is a set of methods or algorithms that are used to find **natural** groupings according to predefined properties of variables in a dataset.
- Clustering is mostly used when we don't have labeled data – data with predefined classes. Clustering uses various properties inside the dataset.

Example

- The data points are classified into two clusters, or two bunches, according to the Euclidean distance between them.
- One cluster contains people who are clearly at high risk of heart disease and the other cluster contains people who are at low risk of heart disease.



Some application of clustering

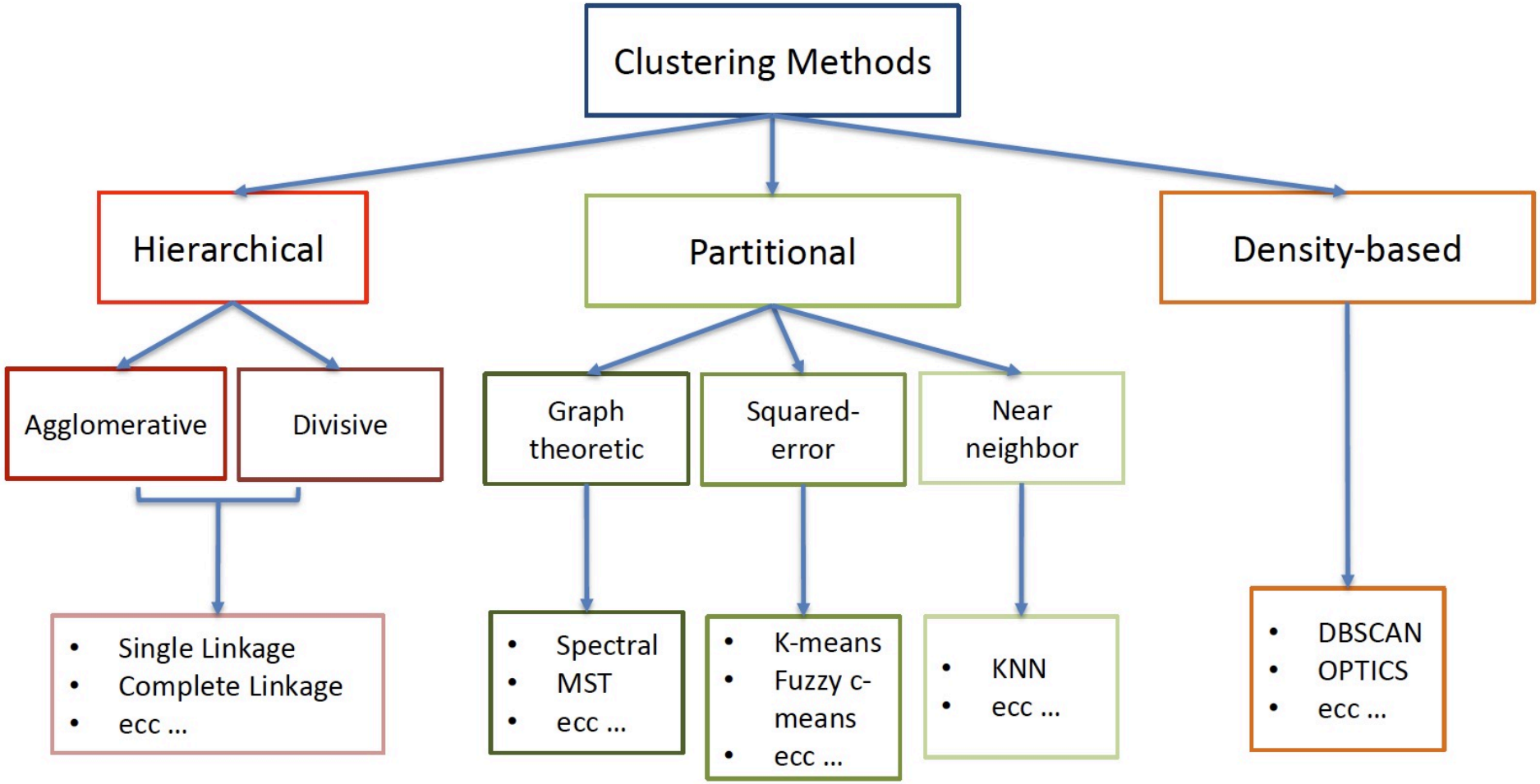
- **Exploratory data analysis:** When we have unlabeled data, we often do clustering to explore the underlying structure and categories of the dataset.

- **Generating training data**: Sometimes, after processing unlabeled data with clustering methods, it can be labeled for further training with supervised learning algorithms.

Some application of clustering

- **Natural language processing:** Clustering can be used for the grouping of similar words, texts, articles, or tweets, without labeled data.
- **Anomaly detection:** You can use clustering to find outliers.
- And so on.

Types of clustering



Clustering methods

- **explicit methods:** such as agglomerative hierarchical clustering and k-means, in which the number of clusters is imposed by the researcher;

- **implicit methods:** such as affinity propagation, where the number of modules is adapted on the dataset analyzed according to other information suggested by the researcher

Types of clustering

- k-means clustering;
- Agglomerative hierarchical clustering;
- Divisive clustering;
- Density-based clustering.

K-means

- K-means clustering is one of the most basic types of unsupervised learning algorithm.
- This algorithm finds natural groupings in accordance with a predefined similarity or distance measure.
- This distance or metric is a measurement of closeness between data point.

K-means

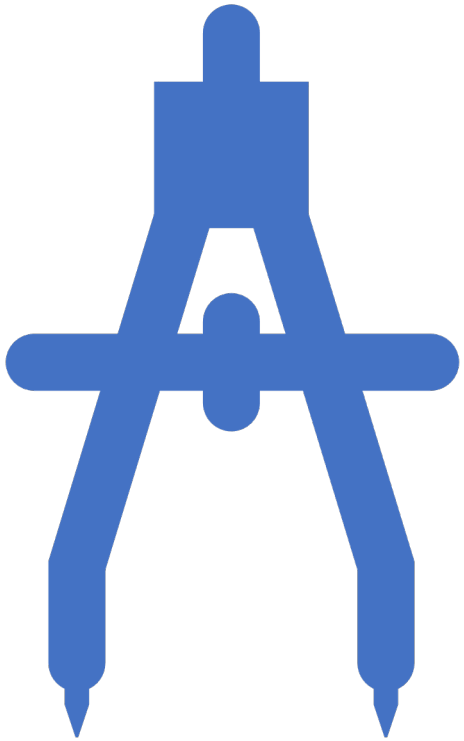
- The k-means algorithm divides a set of n samples X into K clusters C, each described by the mean $\boldsymbol{\mu}_j$ of the samples in the cluster.
- K is chosen by the researcher;
- The means are commonly called the cluster "centroids";
- The K-means algorithm aims to choose centroids that minimise the **inertia**, or **within-cluster sum-of-squares criterion**:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

Metric distances

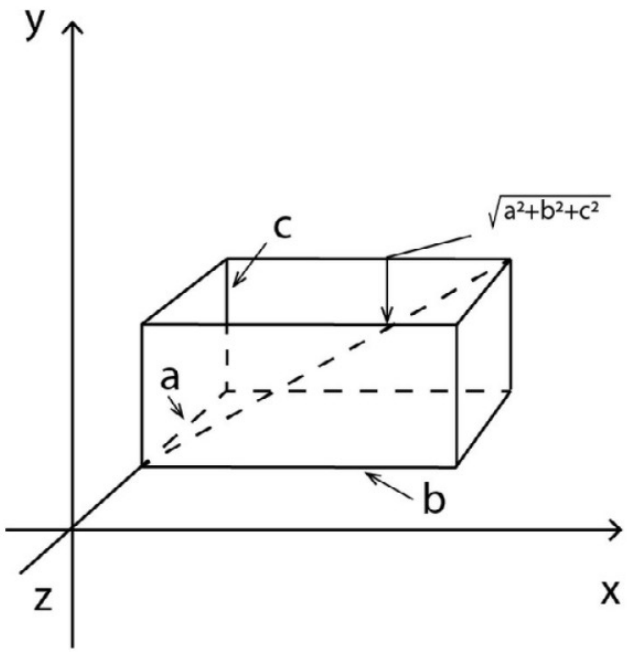
Some types of metric distances used in the k-means algorithm:

- Euclidean distance;
- Manhattan distance;
- Cosine distance.



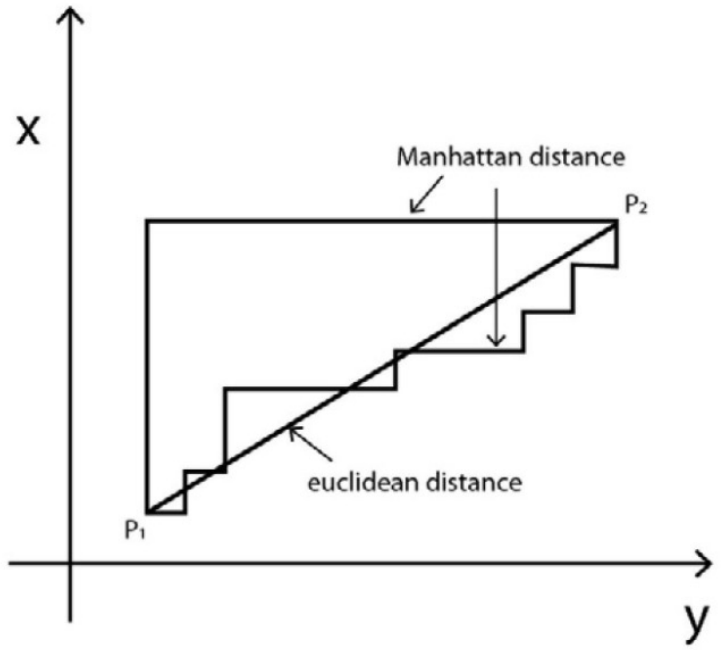
Euclidean distance

- Euclidean distance is the straight-line distance between any two points. Calculation of this distance in two dimensions can be thought of an extension of the Pythagorean theorem.
- But Euclidean distance can be calculated between two points in any n-dimensional space, not just a two-dimensional space.



Manhattan distance

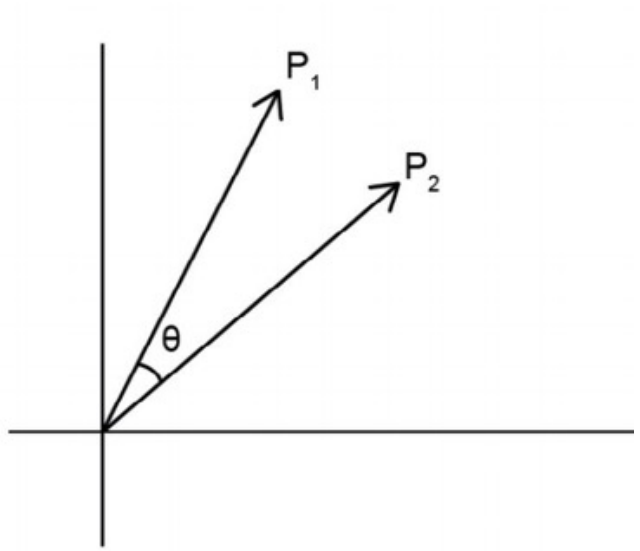
- By definition, Manhattan distance is the distance between two points measured along a right angle to the axes;
- Manhattan distance is simply the sum of the absolute value of the differences between two coordinates.



Cosine distance

- Cosine similarity between any two points is defined as the cosine of the angle between any two points with the origin as its vertex.

- It can be calculated by dividing the dot product of any two vectors by the product of the magnitudes of the vectors.



How K-means works

Firstly the researcher chooses the number of cluster K and the metric distance. Then K-means algorithm:

1. Chooses any K random coordinates, k_1,…, k_K, as initial cluster centroids. Calculate the distance of each data point from coordinates k_i. Assign each data point to a cluster based on whether it is closer to k_i.

How K-means works

Firstly the researcher chooses the number of cluster K and the metric distance. Then K-means algorithm:

- Chooses any K random coordinates, k_1,…, k_K, as initial cluster centroids.
- Calculate the distance of each data point from coordinates k_i.
- Assign each data point to a cluster based on whether it is closer to k_i.
- Find the mean coordinates of all points in each cluster and update the values of centroids to those coordinates respectively.
- Start again from Step 2 until the coordinates of centroids stop moving significantly, or after a certain pre-determined number of iterations of the process.

Deciding the Optimal Number of Clusters

There is more than one way of determining the optimal number of clusters in unsupervised learning. The main ones are:

The Silhouette score;

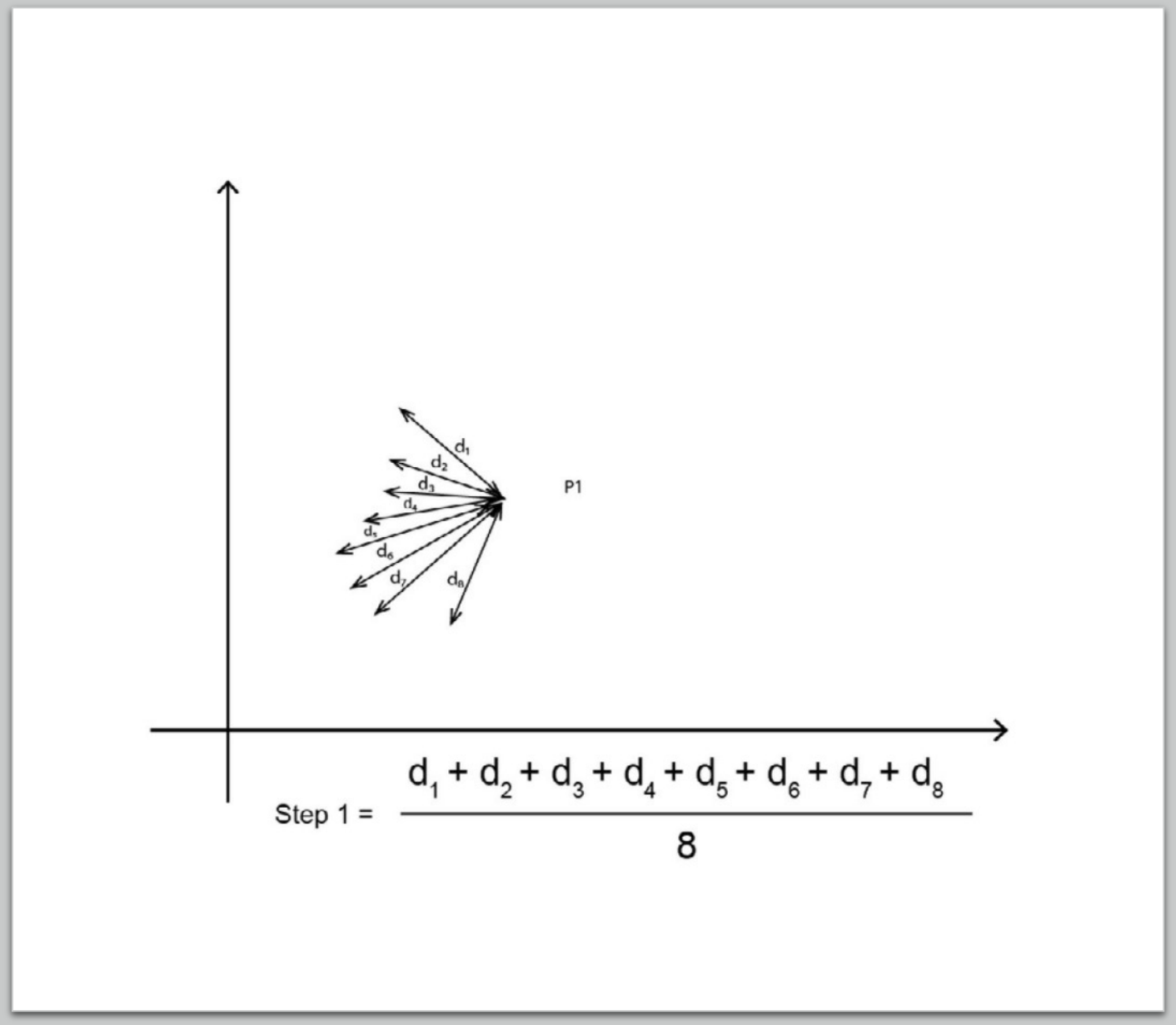
The Elbow method or WSS;

The Gap statistic.

The Silhouette score

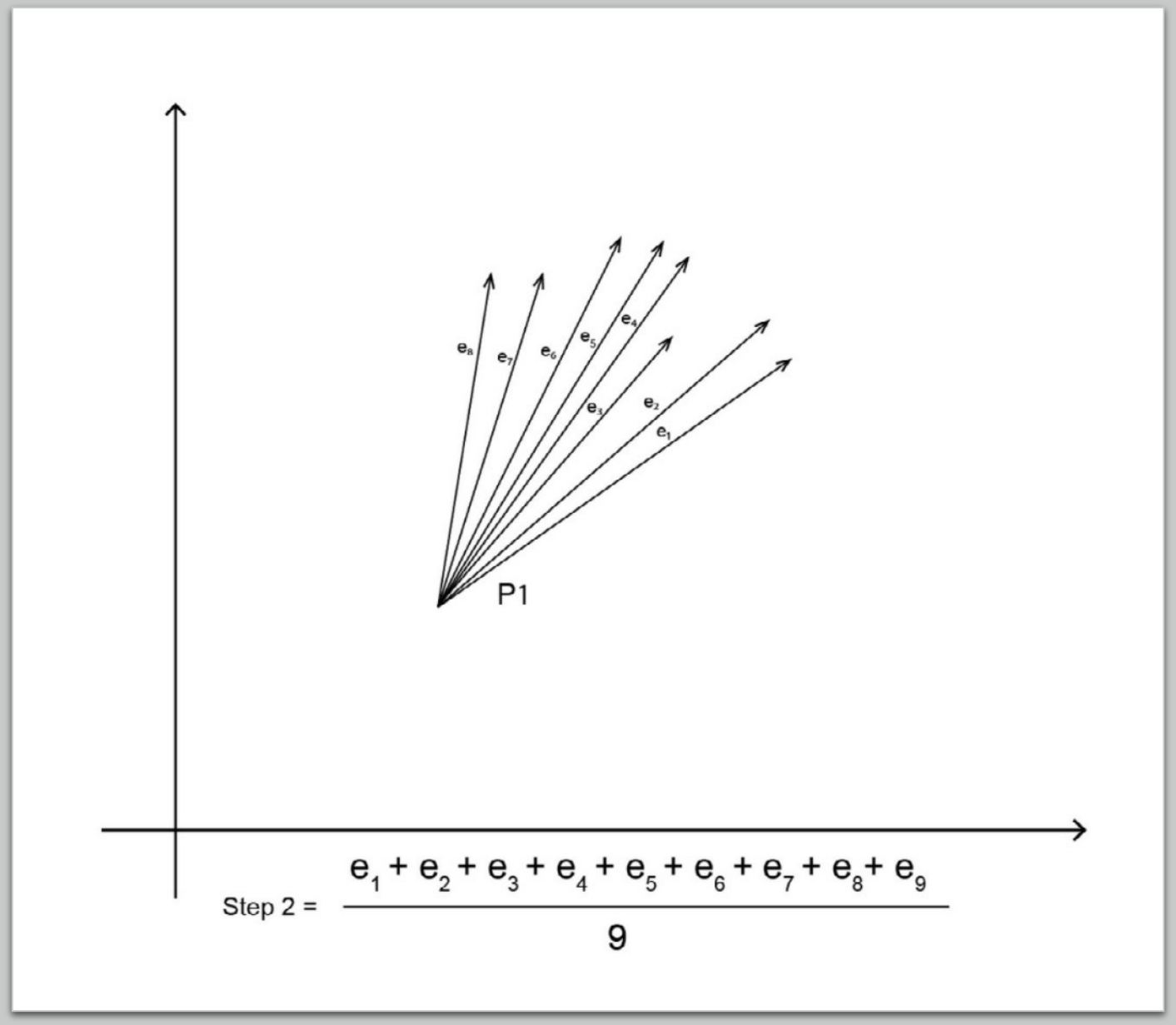
- The silhouette score or average silhouette score calculation is used to quantify the quality of clusters achieved by a clustering algorithm. Let's take a point, **a**, in a cluster, **x**:

1. Calculate the average distance between point **a** and all the points in cluster **x** (denoted by dxa):



The Silhouette score

2. Calculate the average distance between point **a** and all the points in another cluster (**y**) nearest to a (dya):



The Silhouette score

3. Calculate the silhouette score for that point by dividing the difference of the result of Step 1 from the result of Step 2 by the max of the result of Step 1 and Step 2:

$$(dya-dxa)/max(dxa,dya)$$

4. Repeat the first three steps for all points in the cluster.

5. After getting the silhouette score for every point in the cluster, the average of all those scores is the silhouette score of that cluster.

The Silhouette score

6. Repeat the preceding steps for all the clusters in the dataset.

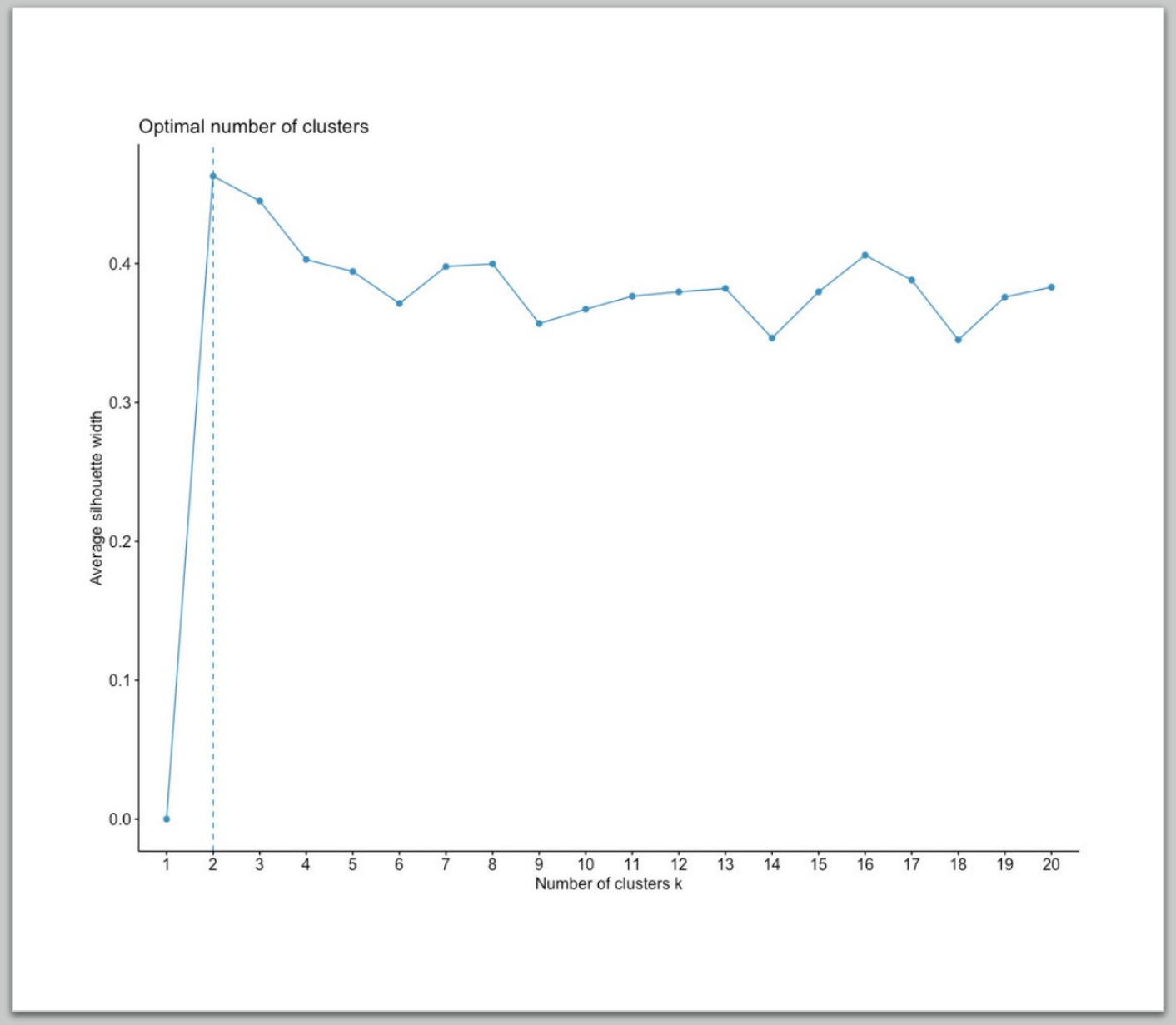
7. After getting the silhouette score for all the clusters in the dataset, the average of all those scores is the silhouette score of that dataset.

The Silhouette score

The silhouette score ranges between 1 and -1 where:

- ST close to 1 it means that the clusters are well defined and the distance between the points of a cluster is low and their distance from points of other clusters is high.

- ST between 0 and -1 means that the cluster is spread out or the distance between the points of that cluster is high.



The Elbow method or WSS

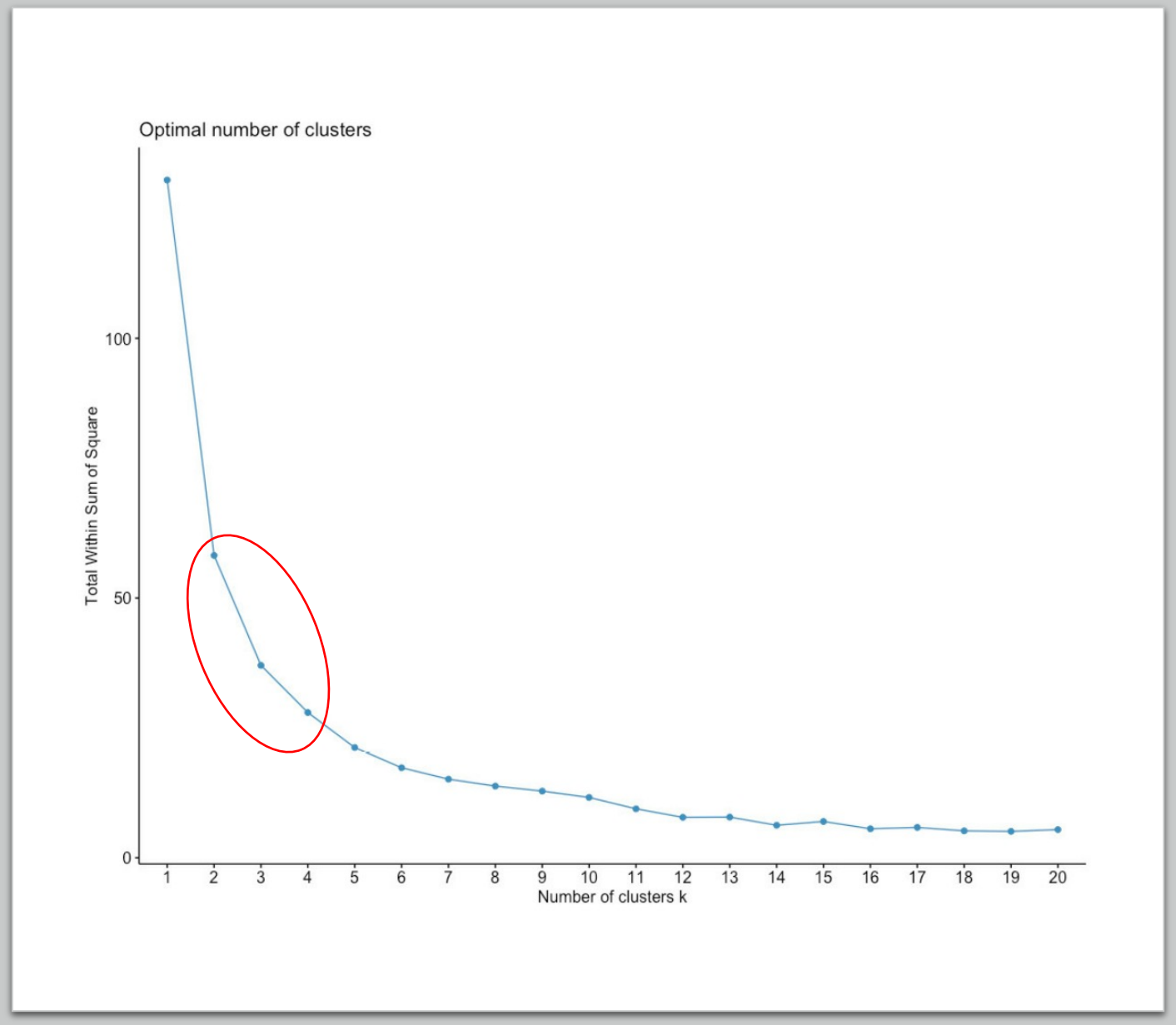
The Within-Sum-of-Squares WSS score is the sum of the squares of the distances between each point of a cluster and the centroid of the cluster. It performs the following steps:

1. Calculate clusters by using different values of k.
2. For every value of k, calculate WSS using the following formula:

$$\mathrm{WSS} = \sum_{i=1}^{k} \sum_{x \in C_i} d(\mathbf{x}, \bar{\mathbf{x}}_{\mathbf{C_i}})^2$$

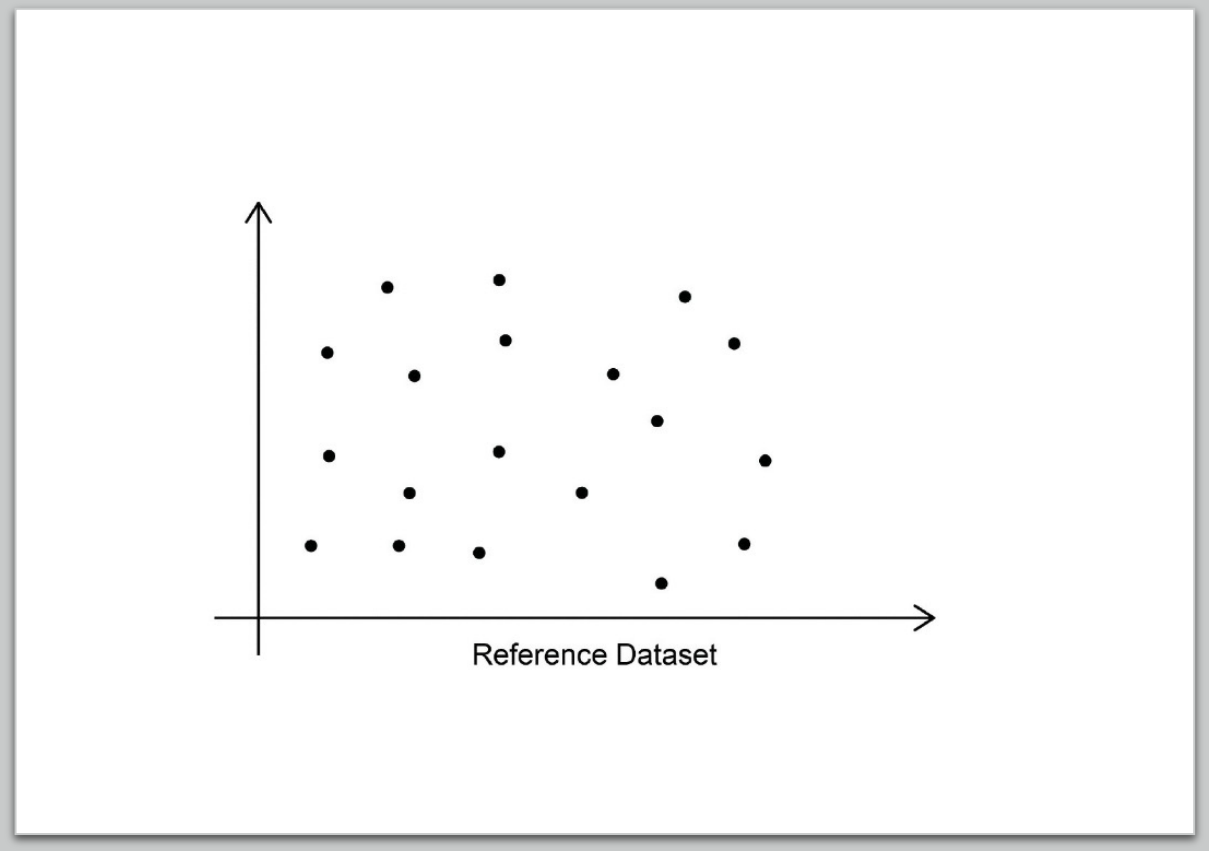
The Elbow method or WSS

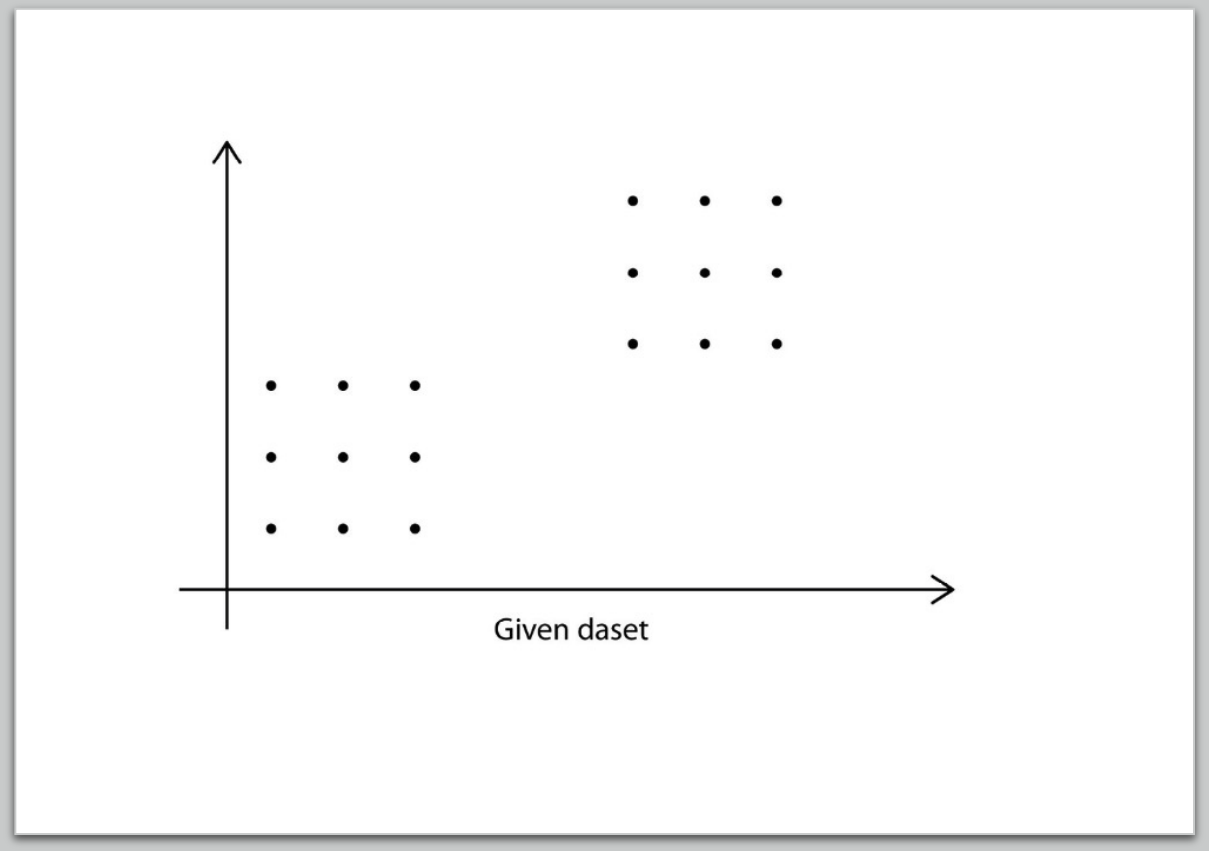
- Plot number of clusters k versus WSS score.
- Identify the k value after which the WSS score doesn't decrease significantly and choose this k as the ideal number of clusters. This point is also known as the elbow of the graph, hence the name "elbow method".



The Gap statistic

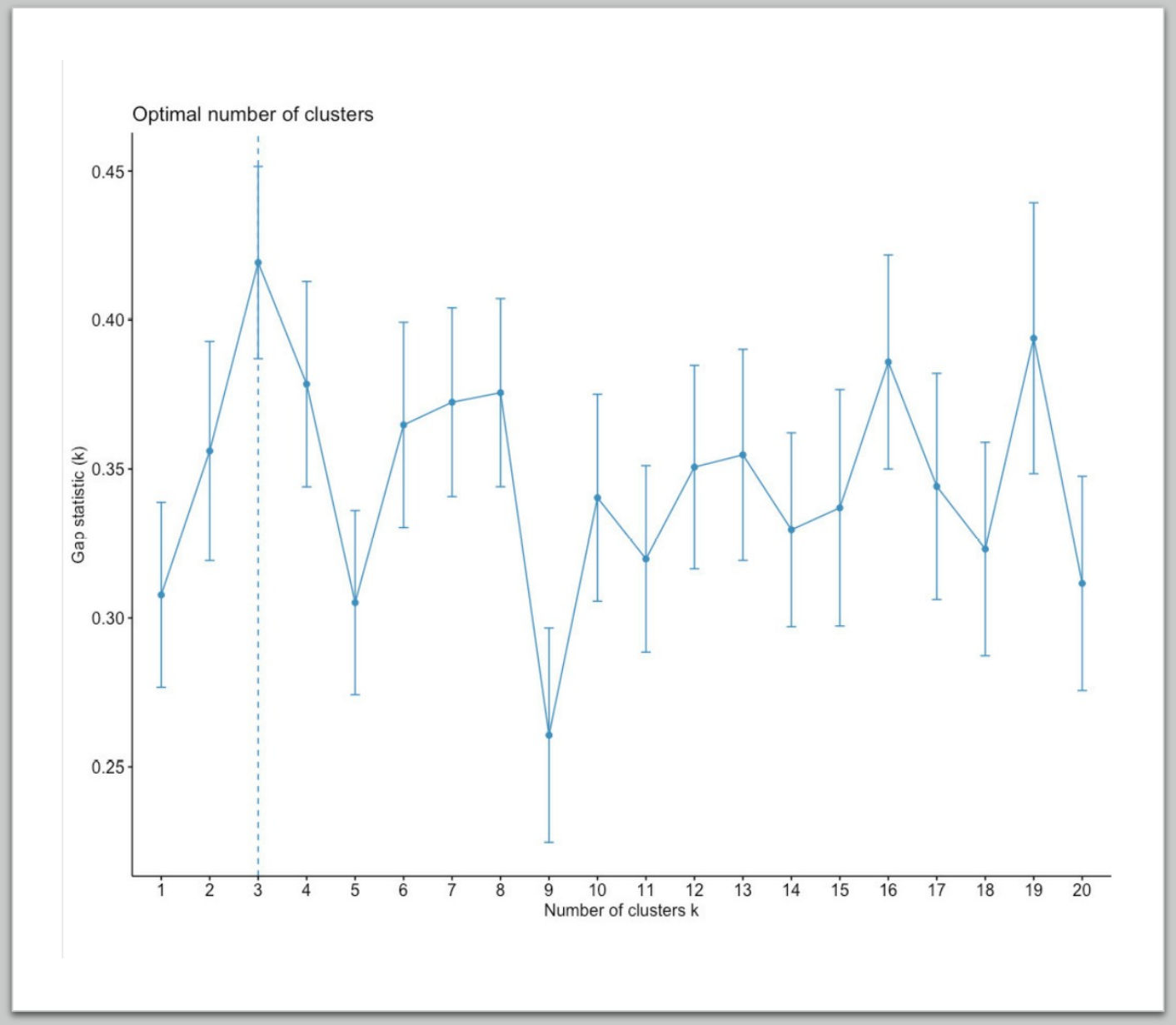
- The Gap statistic is calculated by comparing the WSS value for the clusters generated on our observed dataset versus a reference dataset in which there are no apparent clusters.
- The reference dataset is a random distribution of data points between the minimum and maximum values of our dataset on which we want to calculate the Gap statistic.





The Gap statistic

- So, in short, the Gap statistic measures the WSS values for both observed and random datasets and finds the deviation of the observed dataset from the random dataset.
- To find the ideal number of clusters, we choose a value of k that gives us the maximum value of the Gap statistic.



Hierarchical Clustering

- A hierarchy is defined as "a system in which people or things are placed in a series of levels with different importance or status”;
- Hierarchical clustering merges clusters sequentially. This sequence of merged clusters is called a hierarchy;
- The output of a hierarchical clustering algorithm is called dendrogram.

Types of hierarchical Clustering

Hierarchical clustering comes in two types:

- Agglomerative
- Divisive

Agglomerative

- Agglomerative clustering is also known as the bottom-up approach to hierarchical clustering.
- In this method, each data point is assumed to be a single cluster at the outset.
- From there, we start merging the most similar clusters according to a similarity or distance metric until all the data points are merged in a single cluster.

Divisive

- It is a top-down approach to hierarchical clustering;
- In this method, all the data points are assumed to be in a single cluster initially;
- From there on, we start splitting the cluster into multiple clusters until each data point is a cluster on its own;

Advantages of hierarchical Clustering

- The final output of hierarchical clustering, dendrograms, can help us visualize the clustering results;

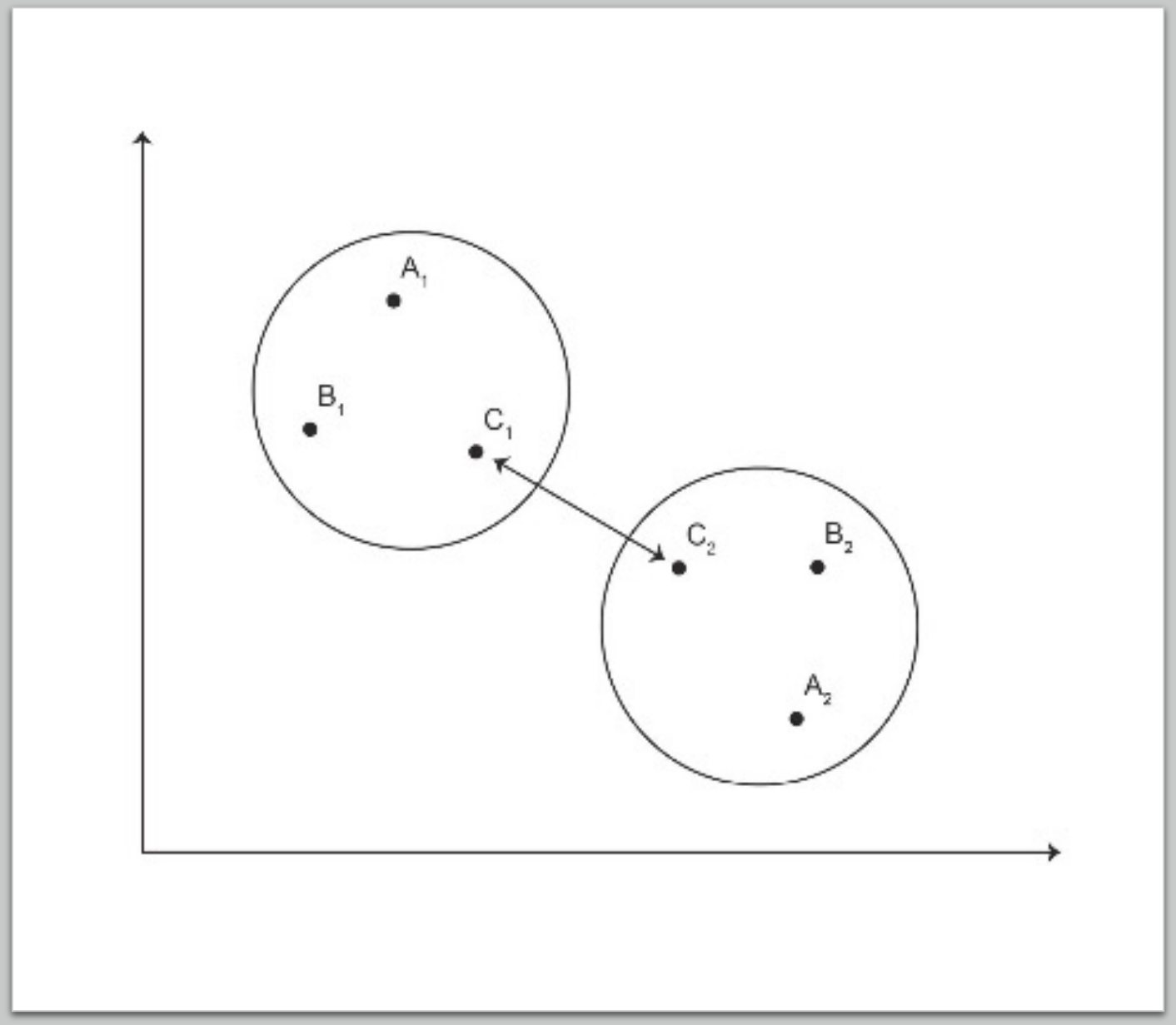
- Unlike k-means, any type of distance metric can be used in hierarchical clustering.

Similarity metrics

- Single link;
- Complete link;
- Group average;
- Centroid similarity.

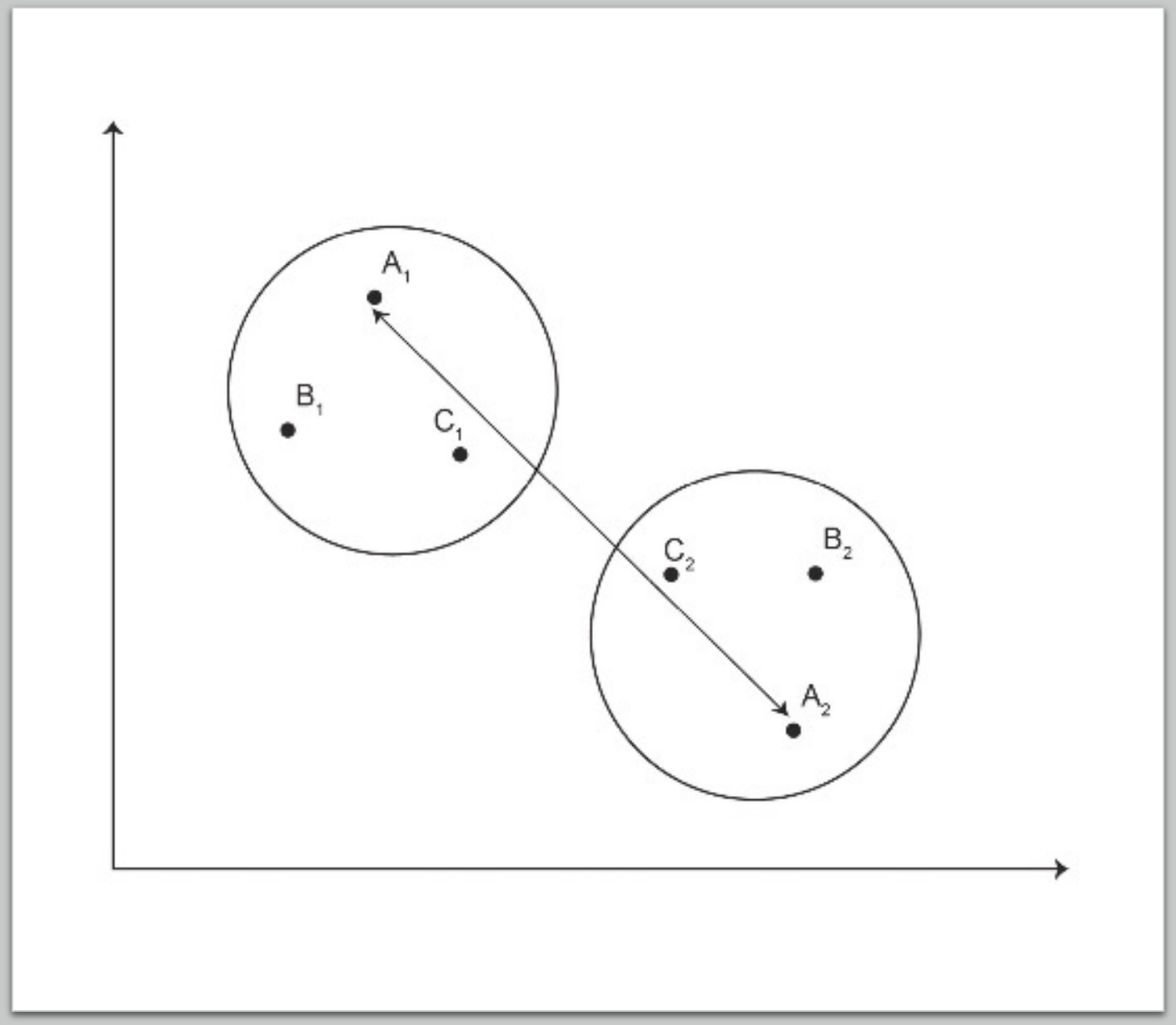
Single link

- In single-link similarity, we measure the distance or similarity between the two closest points of two clusters.



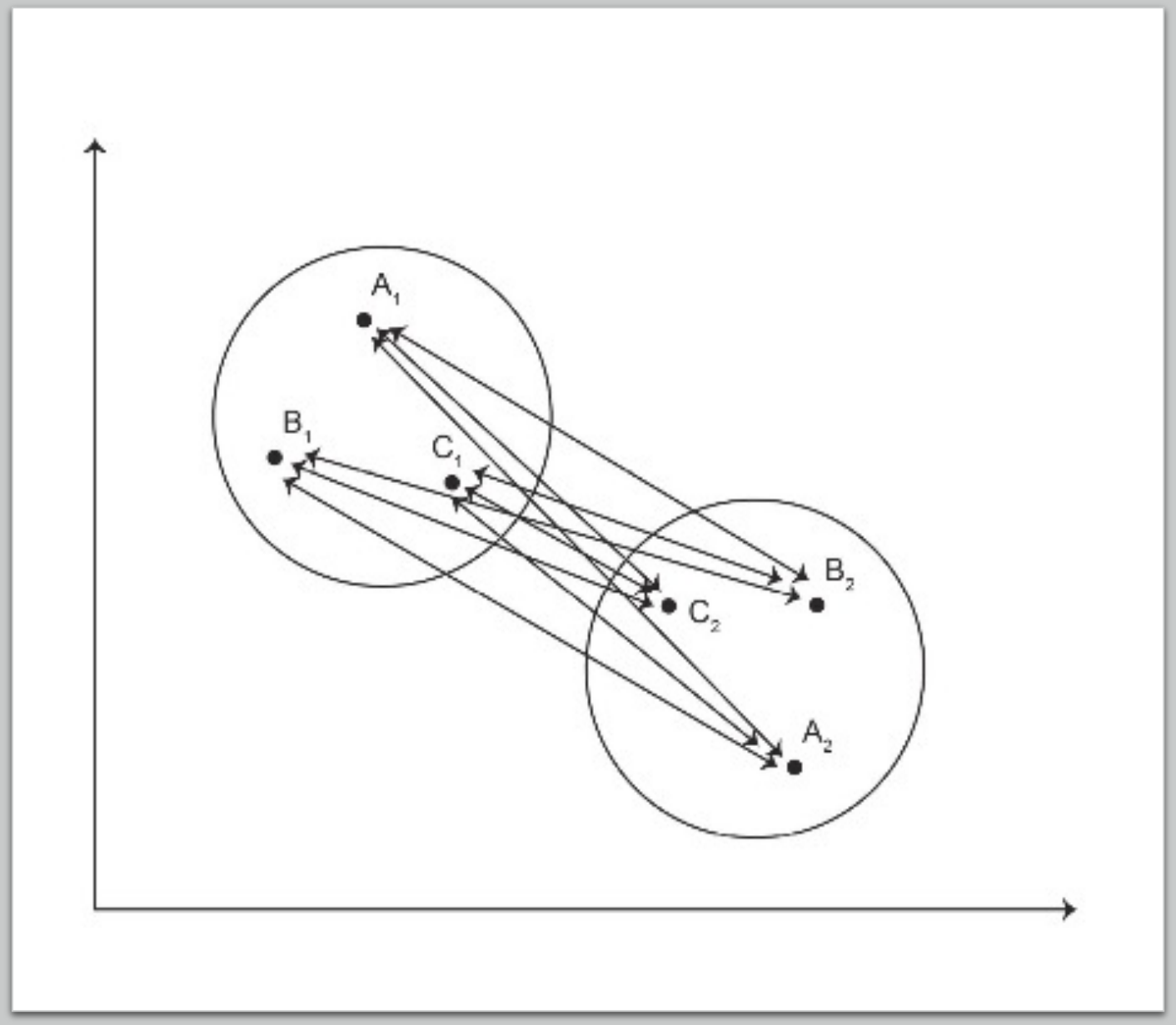
Complete link

- In this type of metric, we measure the distance or similarity between the two most distant points of a cluster.



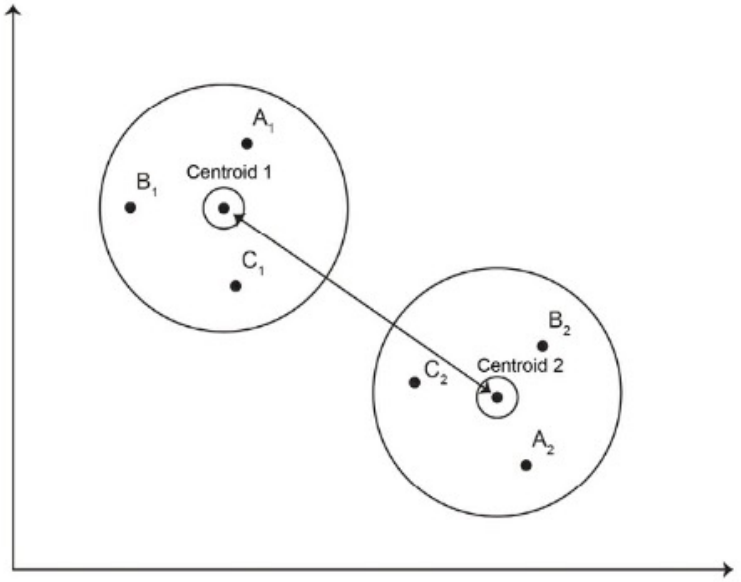
Group average

- In this metric, we measure the average distance between all members of one cluster and any members of a second cluster.



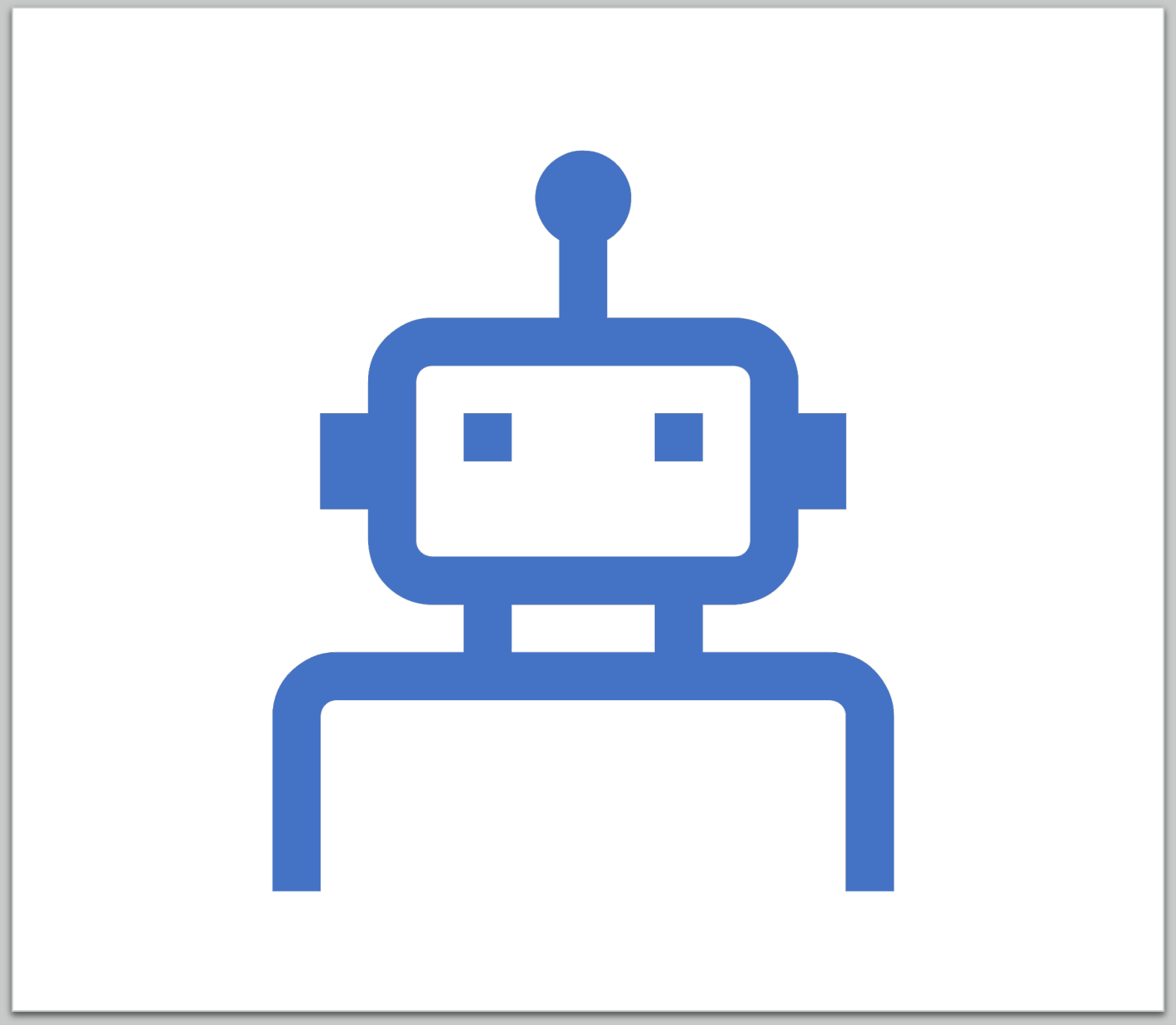
Centroid similarity

- In this type of similarity, the similarity between two clusters is defined as the similarity between the centroids of both clusters.



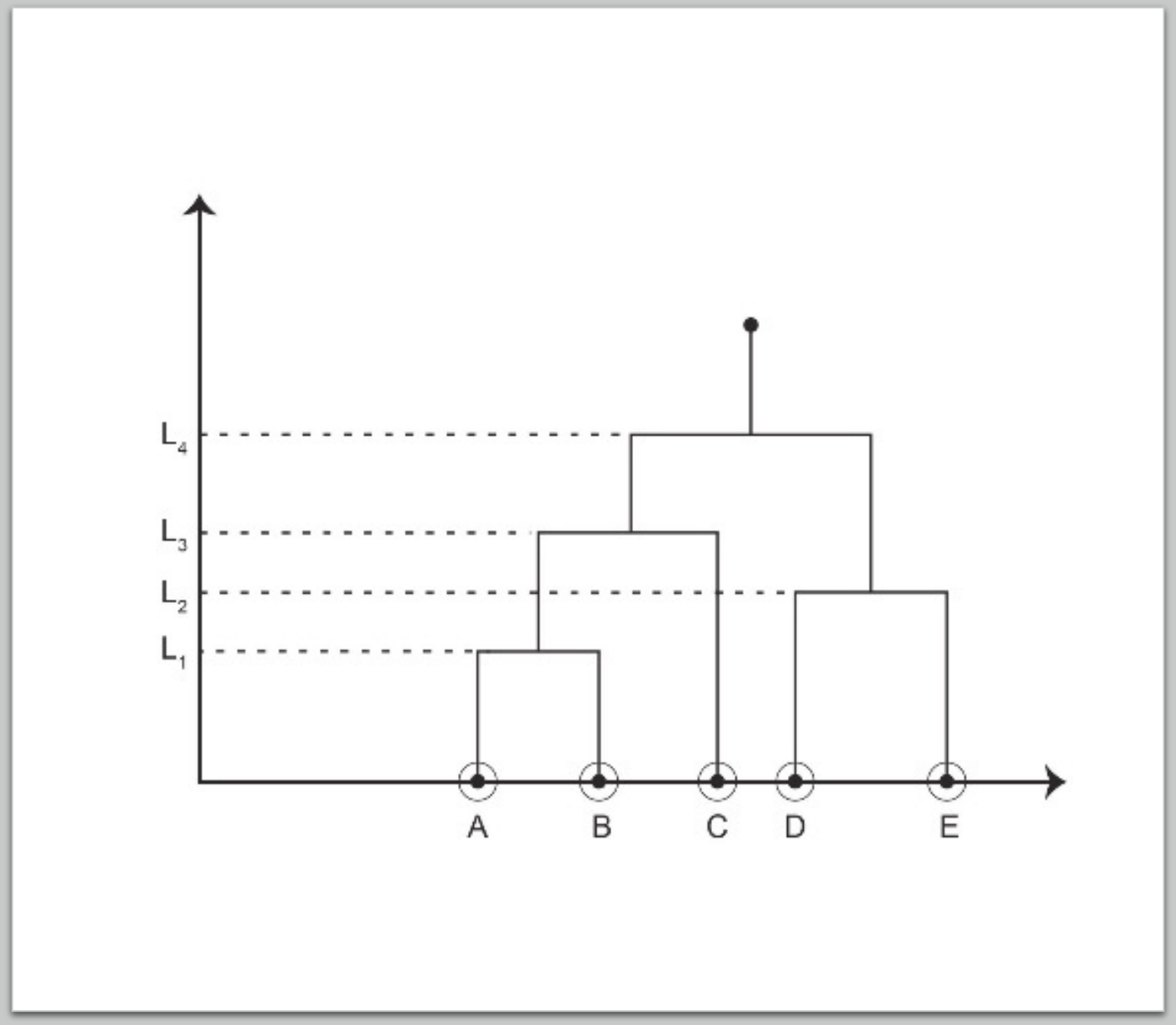
Steps to perform an agglomerative algorithm

- Initialize each point as a single cluster;
- Calculate the similarity metric between every pair of clusters. The similarity metric can be any of the four metrics we just read about;
- Merge the two most similar clusters according to the similarity metric metric selected in step 2;
- Repeat the process from step 2 until we have only one cluster left.



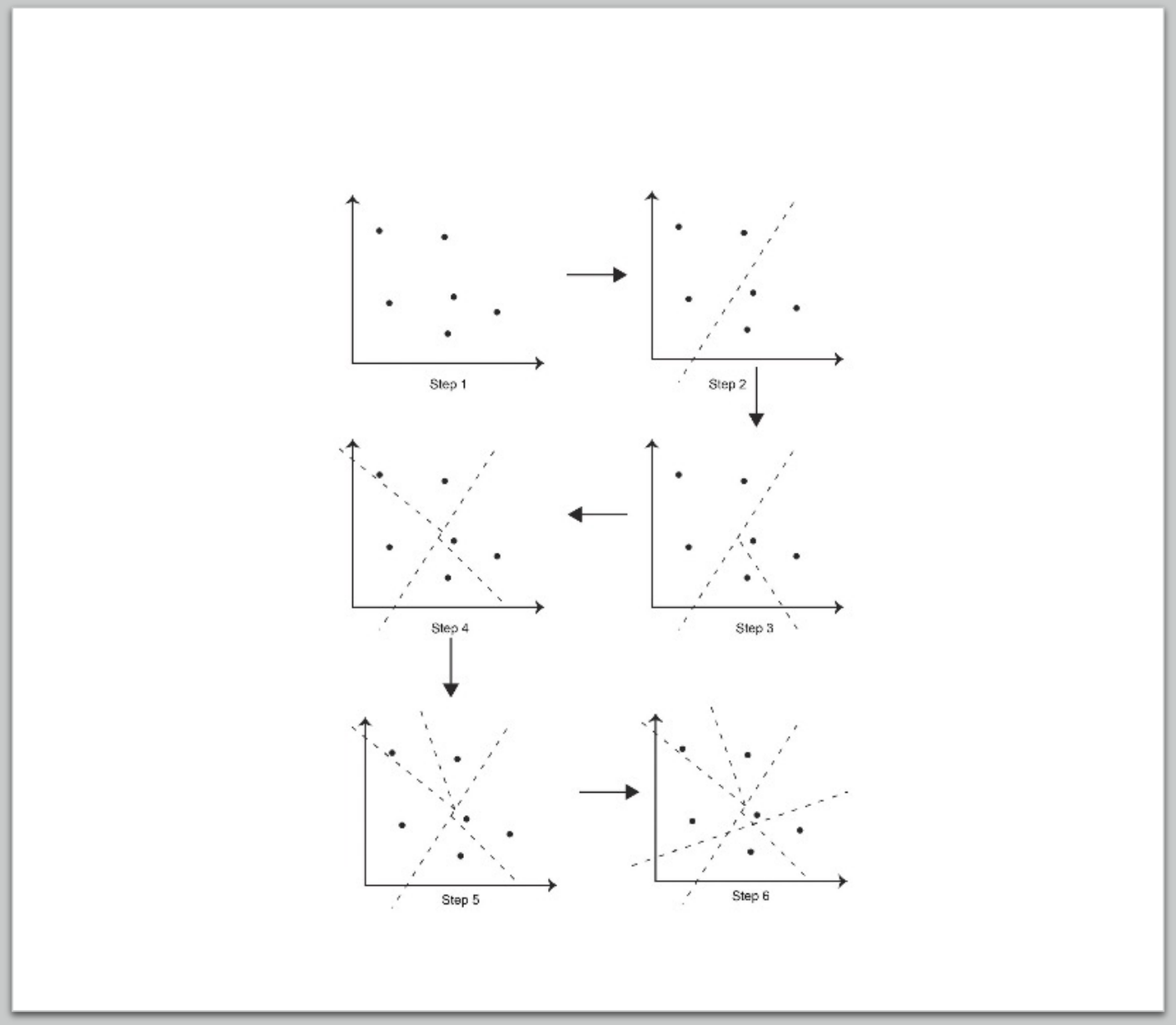
Agglomerative algorithm output

- This whole process will produce a graph called a dendrogram.

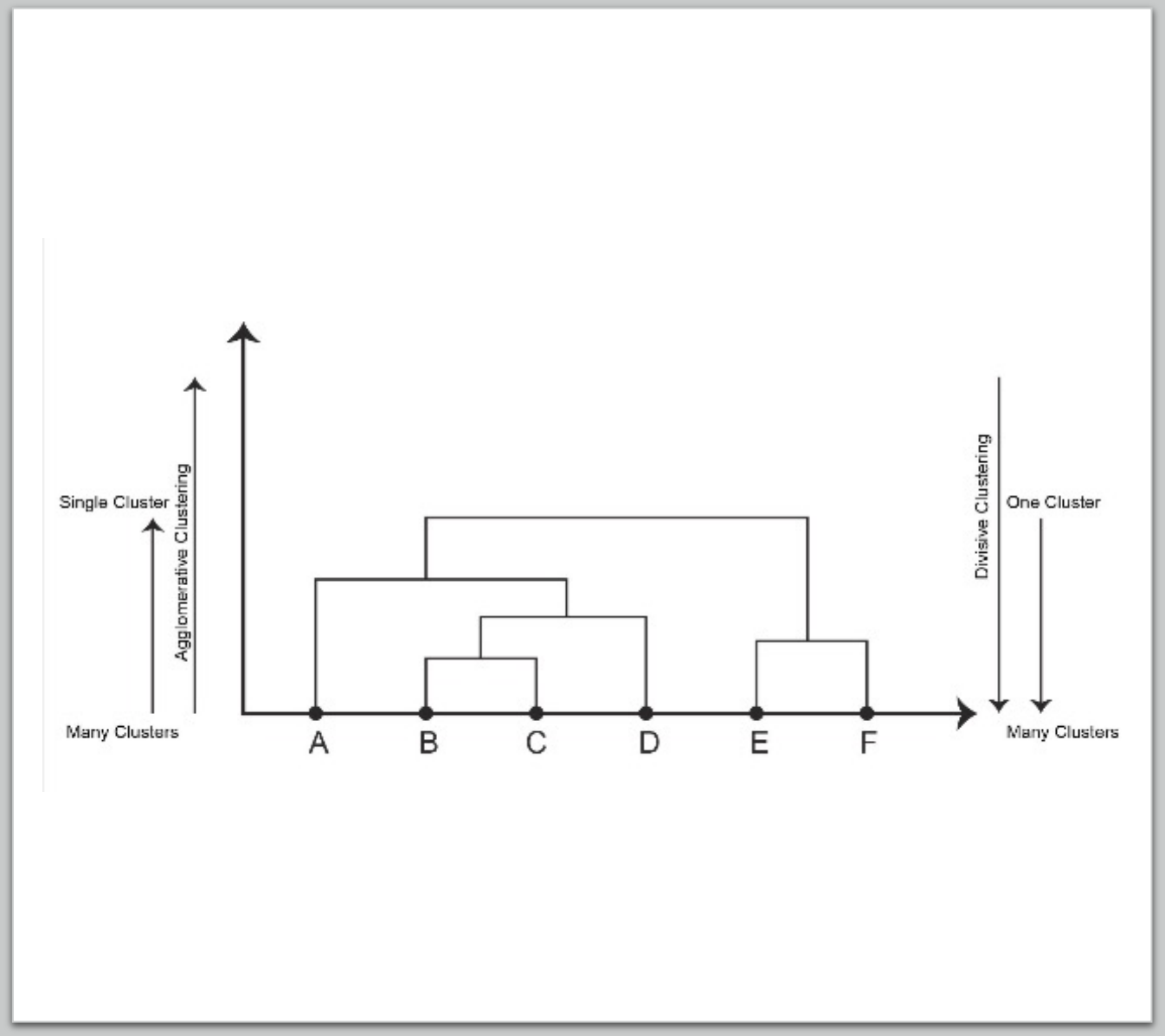


Steps to perform a divisive algorithm

- Start with all the points in the dataset in one single cluster.
- Choose two of the most dissimilar clusters of all the possible clusters in the dataset according to any distance metric you like.
- Repeat step 2 until all the points in the dataset are clustered on their own.



Divisive algorithm output



Hints on other types of clustering algorithms

K-Modes clustering

Density-based clutering

K-modes clustering

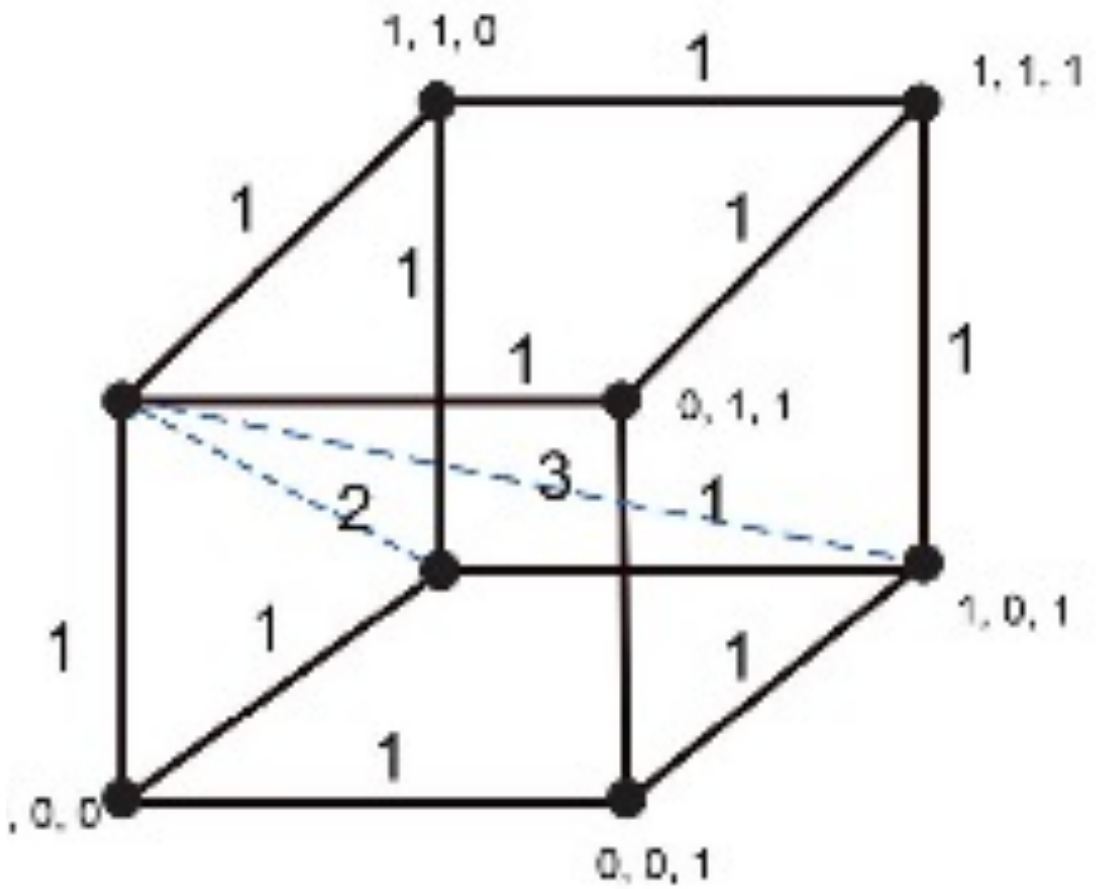
- k-modes clustering is an extension of k-means clustering, dealing with modes instead of means.
- One of the major applications of k-modes clustering is analyzing categorical data such as survey results.
- Mode is defined as the most frequently occurring value. So, for k-modes clustering, we're going to calculate the mode of categorical values to choose centers.

K-modes clustering

- Steps are the same of k-means, but in this case we use the Hamming distance;
- The Hamming distance distance is a special type of distance that is used for categorical variables.
- Given two points of equal dimensions, the Hamming distance is defined as the number of coordinates differing from one another.

The Hamming distance

- For example, let's take two points, (0, 1, 1) and (0, 1, 0). Only one value, which is the last value, is different between these two variables. As such, the Hamming distance between them is 1.



Density based of clustering

- Density-based clustering (DBSCAN) is very easy to find naturally occurring clusters and outliers in data with this type of clustering.
- it doesn't require you to define a number of clusters;
- It finds regions of high density separated by regions of low density in a scatter plot.