FUNDAMENTALS OF MACHINE LEARNING IN DATA SCIENCE

CSIS 3290

CLUSTERING

K-MEANS

FATEMEH AHMADI

Introduction

- In general, clustering is the use of **unsupervised techniques** for grouping similar objects. In machine learning, unsupervised refers to the problem of finding hidden structure within unlabeled data (<u>You can use a labeled dataset for clustering practice if you remove the label column</u>). Clustering techniques are unsupervised in the sense that the data scientist does not determine, in advance, the labels to apply to the clusters.
- For example, based on customers' personal income, it is straightforward to divide the customers into three groups depending on arbitrarily selected values. The customers could be divided into three groups as follows:
  - Earn less than $10,000
  - Earn between $10,000 and $99,999
  - Earn $100,000 or more

Use Cases

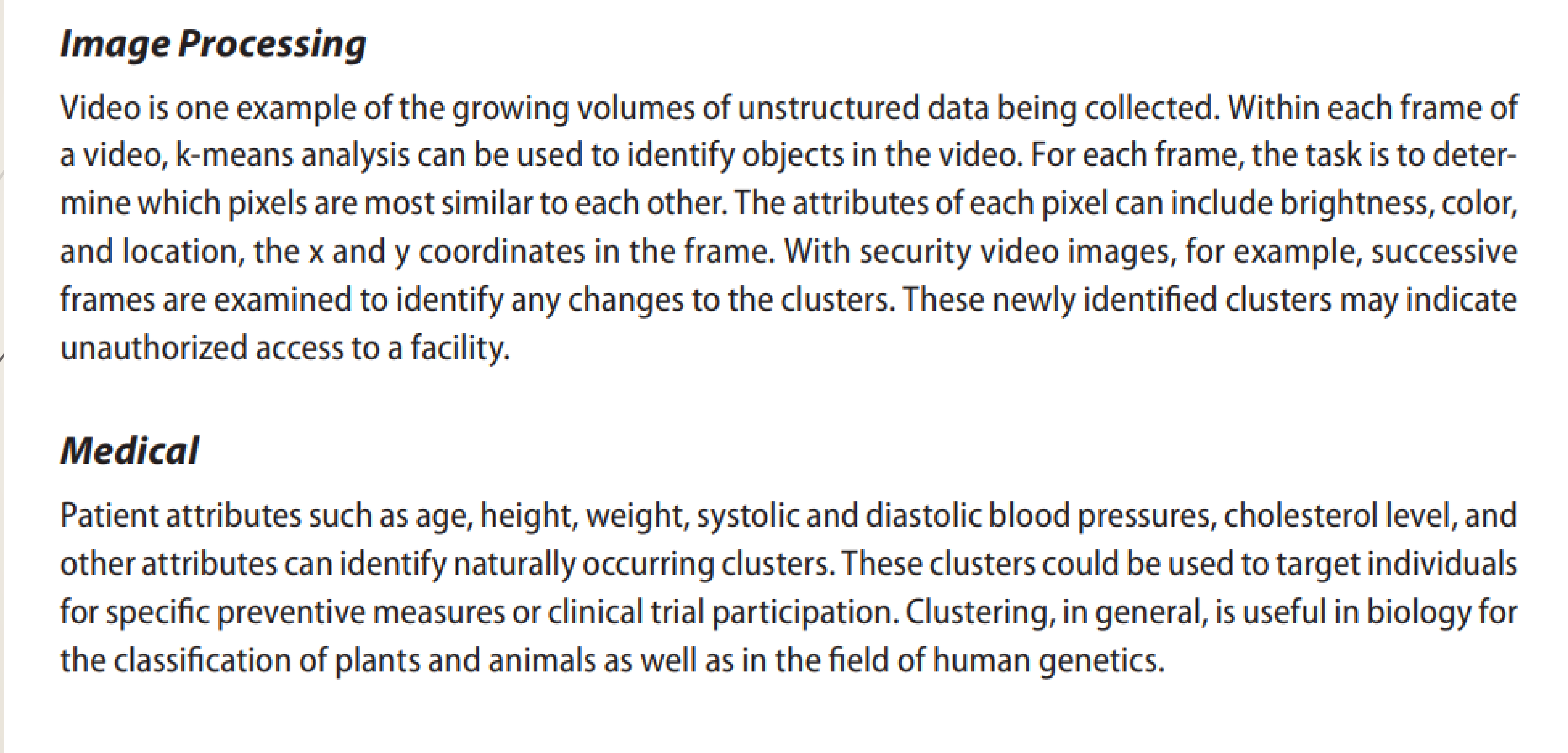


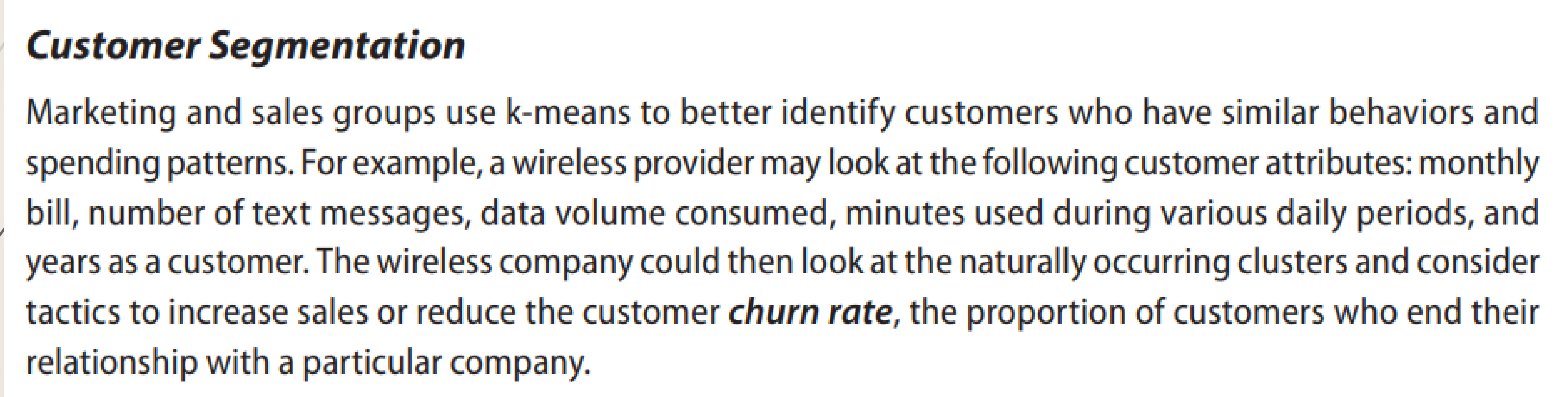
Image Processing

Video is one example of the growing volumes of unstructured data being collected. Within each frame of a video, k-means analysis can be used to identify objects in the video. For each frame, the task is to determine which pixels are most similar to each other. The attributes of each pixel can include brightness, color, and location, the x and y coordinates in the frame. With security video images, for example, successive frames are examined to identify any changes to the clusters. These newly identified clusters may indicate unauthorized access to a facility.

Medical

Patient attributes such as age, height, weight, systolic and diastolic blood pressures, cholesterol level, and other attributes can identify naturally occurring clusters. These clusters could be used to target individuals for specific preventive measures or clinical trial participation. Clustering, in general, is useful in biology for the classification of plants and animals as well as in the field of human genetics.

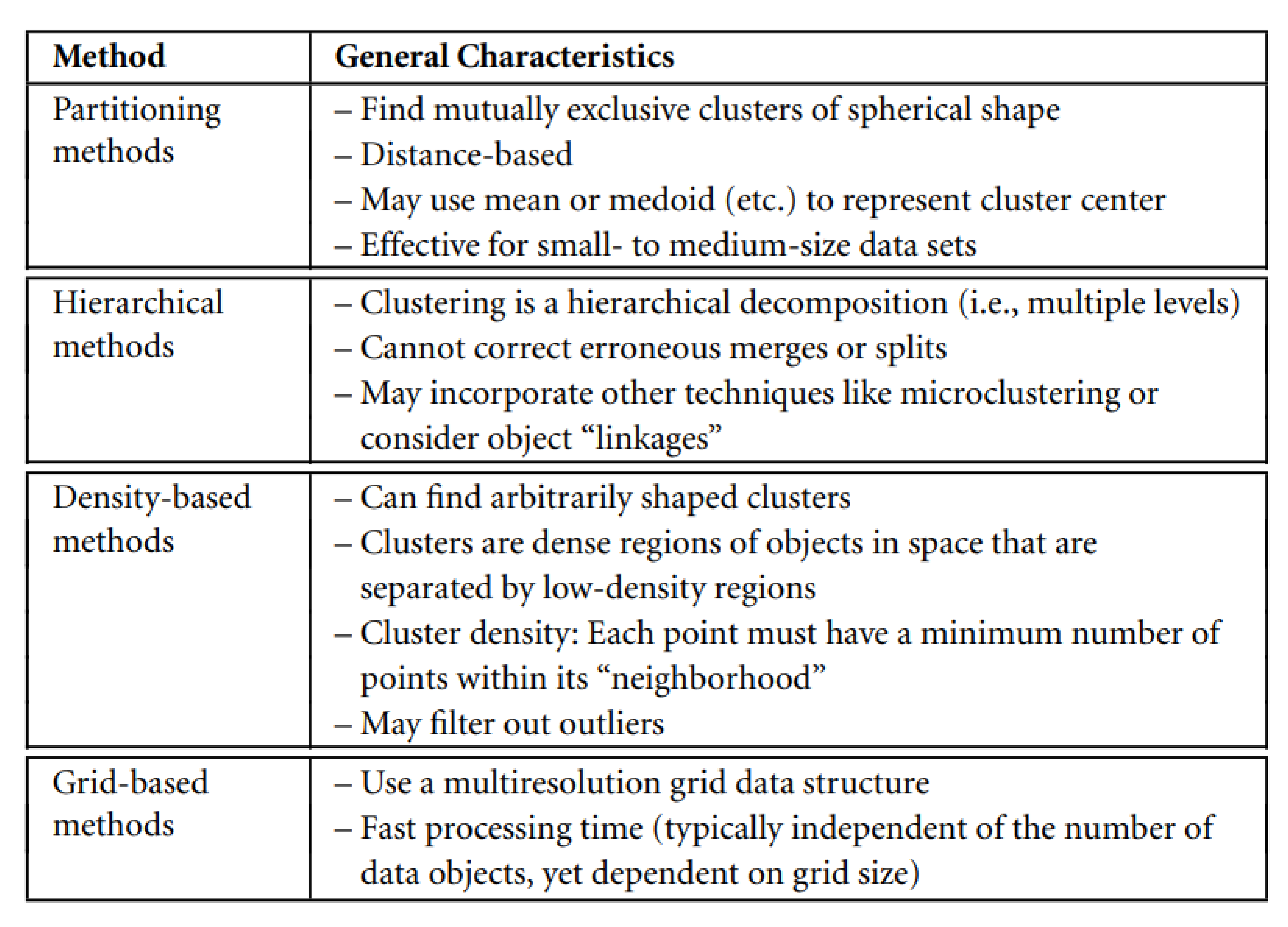
Use Cases



Customer Segmentation

Marketing and sales groups use k-means to better identify customers who have similar behaviors and spending patterns. For example, a wireless provider may look at the following customer attributes: monthly bill, number of text messages, data volume consumed, minutes used during various daily periods, and years as a customer. The wireless company could then look at the naturally occurring clusters and consider tactics to increase sales or reduce the customer ***churn rate***, the proportion of customers who end their relationship with a particular company.

Different Clustering Approaches



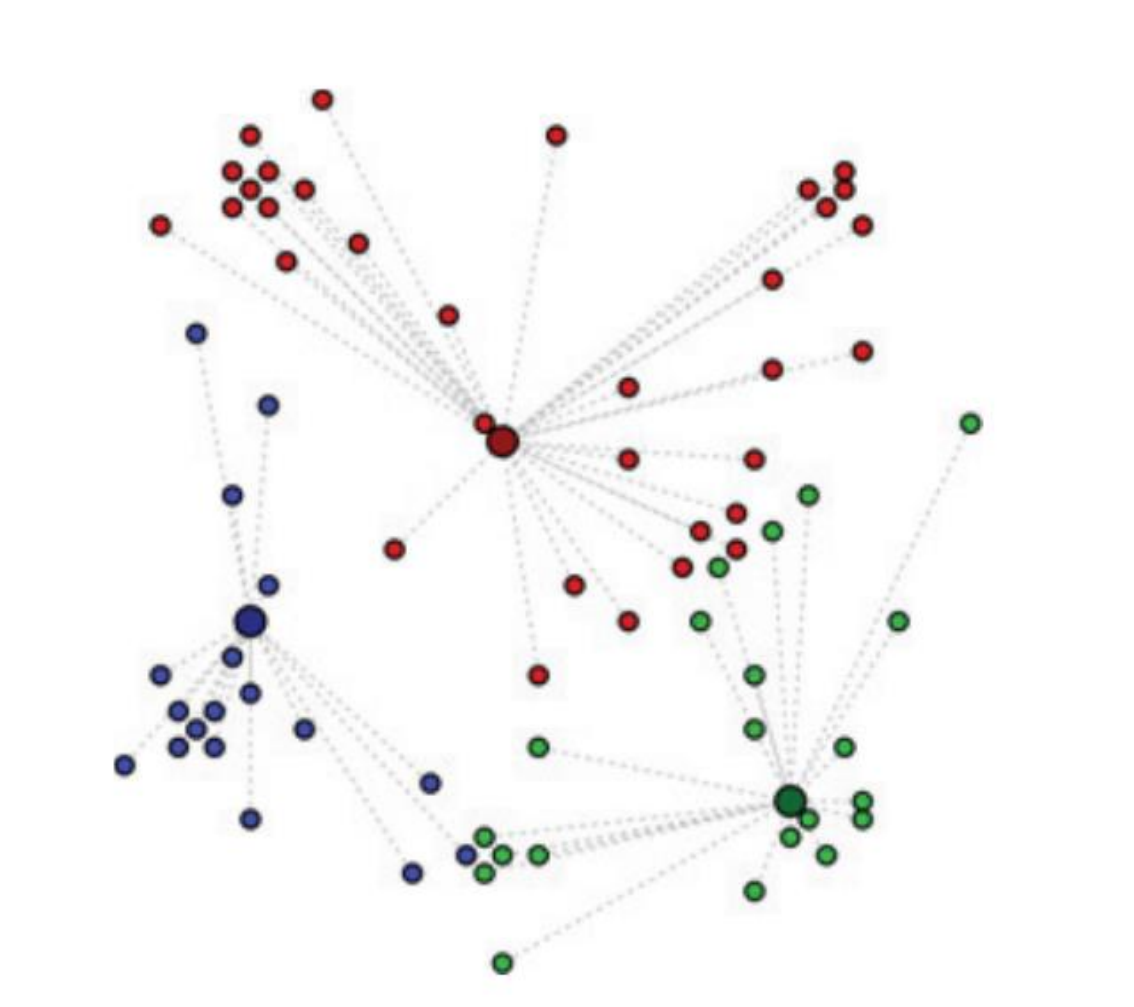
Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

K-MEANS

- Given a collection of objects each with *n* measurable attributes, **k-means** is an analytical technique that, for a chosen value of *k*, identifies *k* clusters of objects based on the objects' proximity to the center of the *k* groups.
- The center is determined as the arithmetic average (mean) of each cluster's n-dimensional vector of attributes. This section describes the algorithm to determine the *k*-means as well as how best to apply this technique to several use cases.
- Next figure illustrates three clusters of objects with two attributes. Each object in the dataset is represented by a small dot color-coded to the closest large dot, the mean of the cluster.

K-MEANS

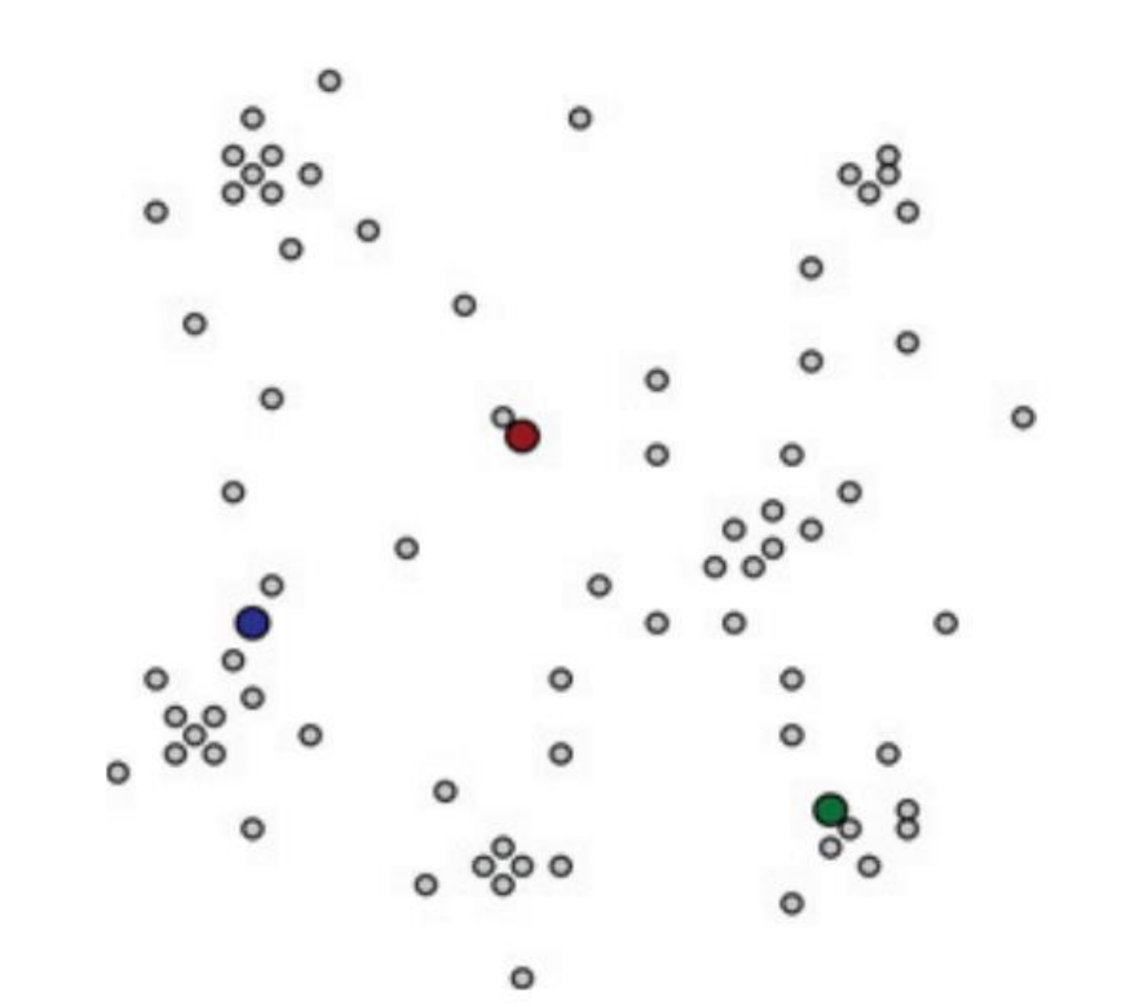
K-Means for K=3



K-MEANS

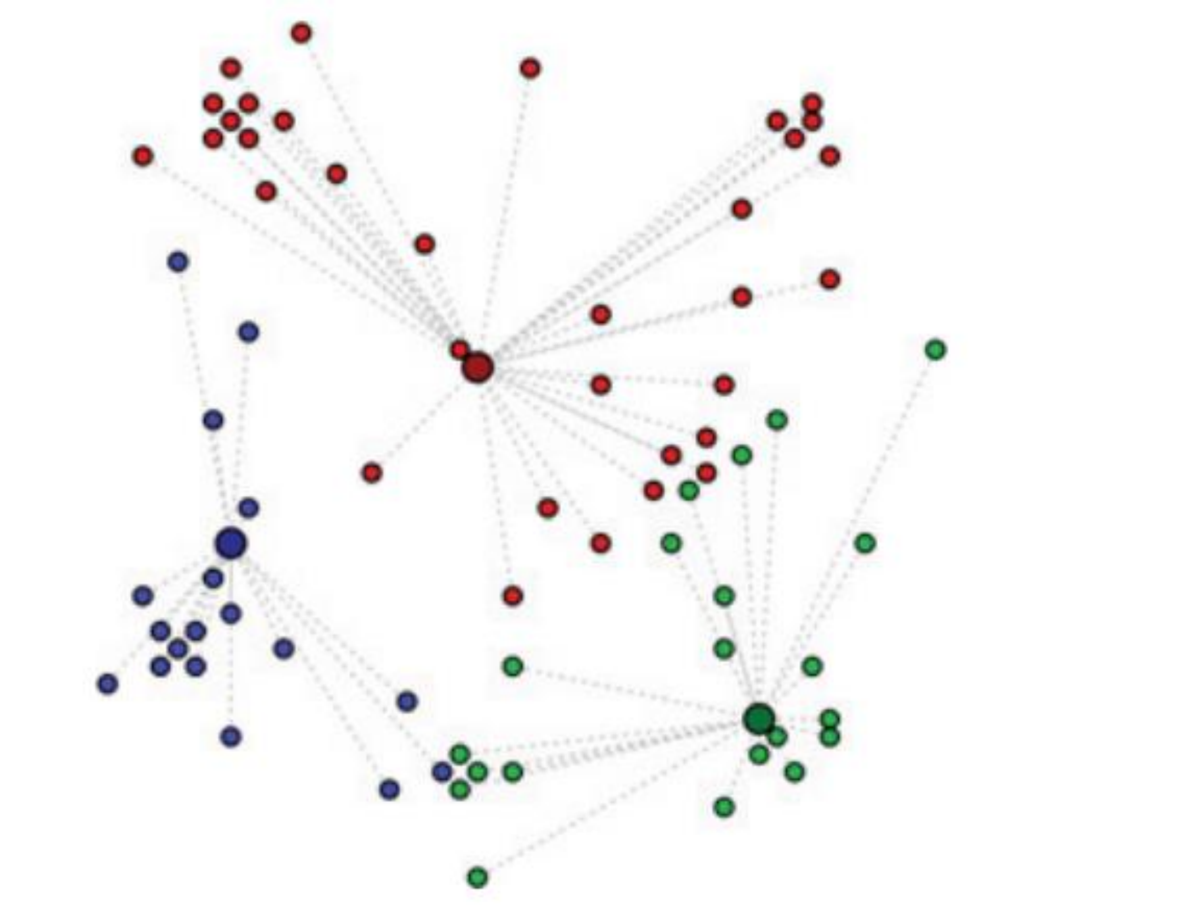
Step 1:

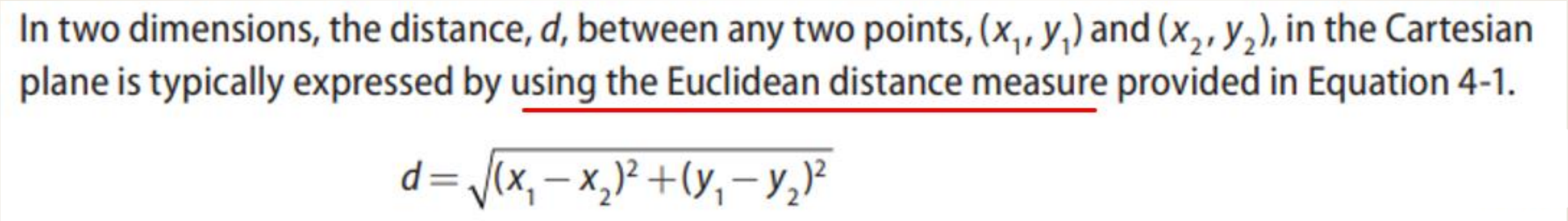
https://www.bing.com/videos/search?ptag=ICO-a37ed4490a84afc3&pc=1MSC&q=kmeans+algorithm+video&ru=%2fsearch%3fptag%3dICO-a37ed4490a84afc3%26form%3dINCOH1%26pc%3d1MSC%26q%3dkmeans%2520algorithm%2520video&view=detail&mmscn=vwrc&mid=FFB271DF90375FC9E55CFFB271DF90375FC9E55C&FORM=WRVORC



K-MEANS

Step 2:



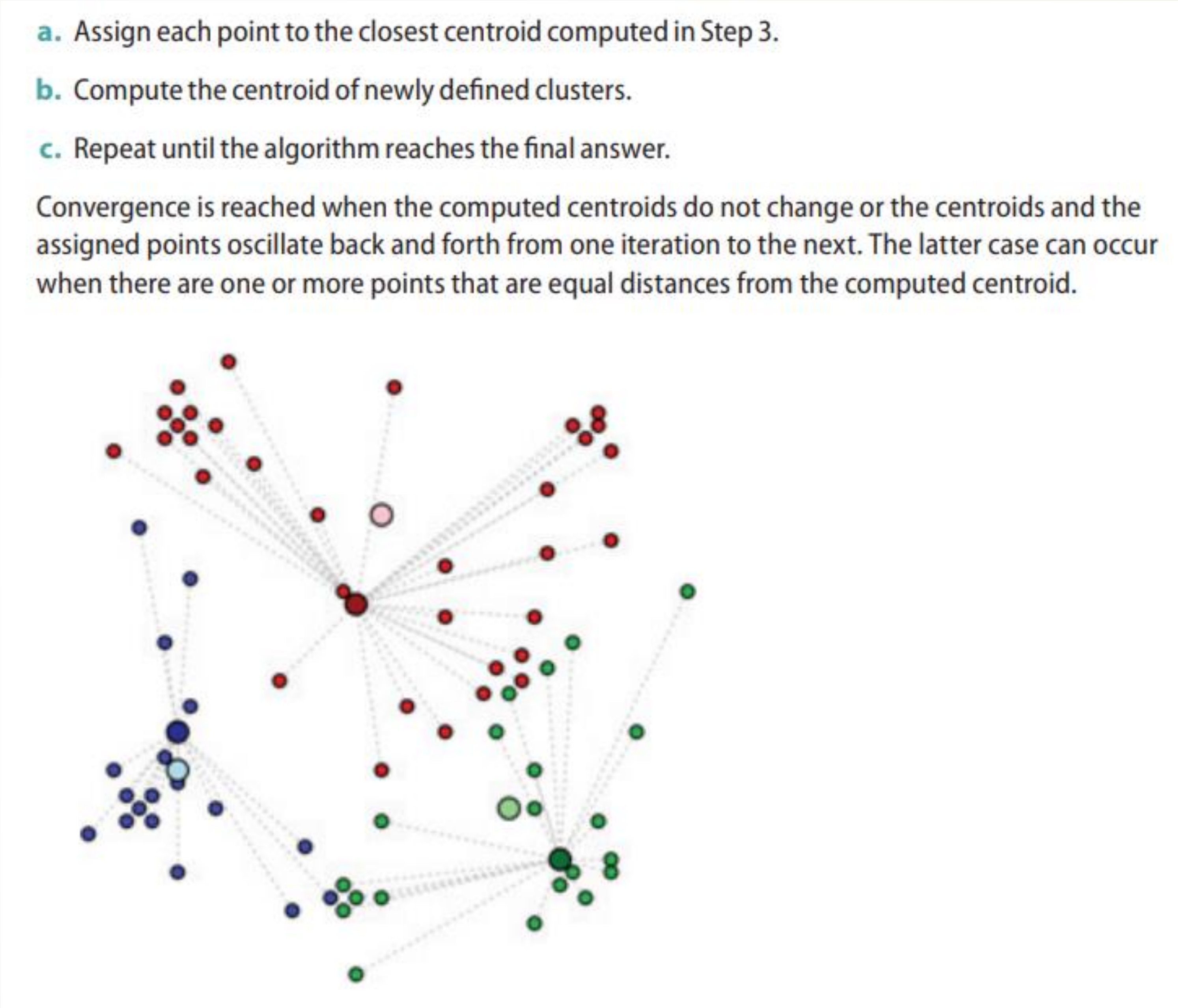


In two dimensions, the distance, *d*, between any two points, (x_1, y_1) and (x_2, y_2), in the Cartesian plane is typically expressed by using the Euclidean distance measure provided in Equation 4-1.

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

K-MEANS

Step 3:



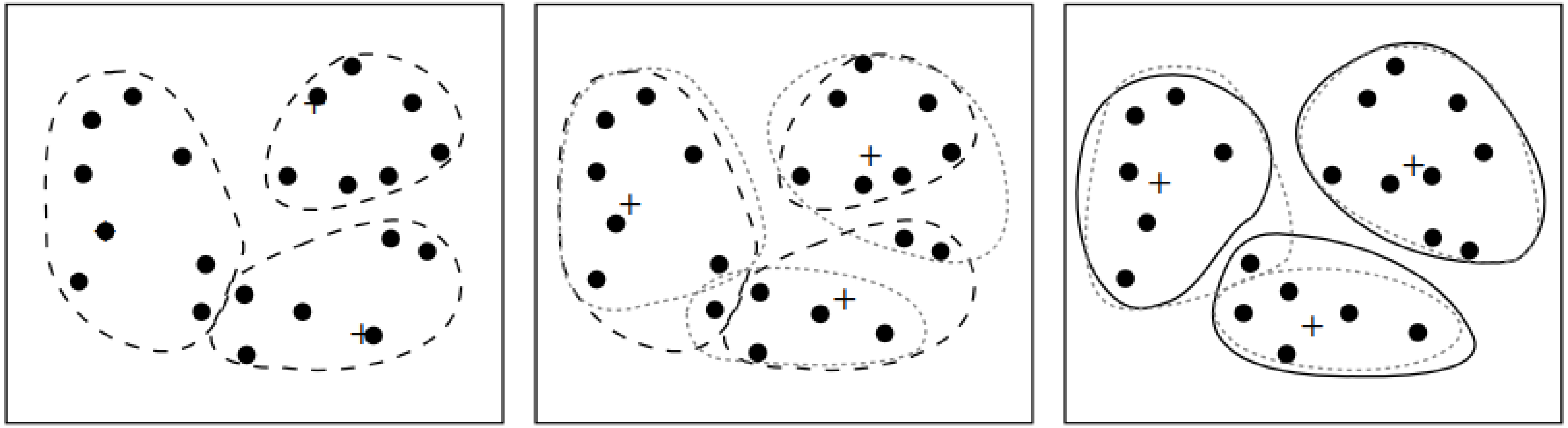
a. Assign each point to the closest centroid computed in Step 3.

b. Compute the centroid of newly defined clusters.

c. Repeat until the algorithm reaches the final answer.

Convergence is reached when the computed centroids do not change or the centroids and the assigned points oscillate back and forth from one iteration to the next. The latter case can occur when there are one or more points that are equal distances from the computed centroid.

K-Means



(a) Initial clustering

(b) Iterate

(c) Final clustering

Additional Consideration About K-Means

- K-means clustering is applicable to objects that can be described by attributes that are **numerical** with a **meaningful** distance measure. However, k-means does not handle categorical variables well.

- For example, suppose a clustering analysis is to be conducted on new car sales. Among other attributes, such as the sale price, the color of the car is considered important. Although one could assign numerical values to the color, such as red = 1, yellow = 2, and green = 3, it is not useful to consider that yellow is as close to red as yellow is to green from a clustering perspective. In such cases, it may be necessary to use an alternative clustering methodology.

Additional Consideration About K-Means

- The k-means clustering method is easily applied to numeric data where the concept of distance can naturally be applied. However, it may be necessary or desirable to use an alternative clustering algorithm.
- As discussed at the end of the previous section, ***k-means* does not handle categorical data**. In such cases, ***k-modes* is a commonly used method for clustering categorical data** based on the number of differences in the respective components of the attributes.
- For example, if each object has four attributes, the distance from *(a, b, e, d)* to *(d, d, d, d)* is 3. In R, the function *kmode*() is implemented in the *klaR* package.

Other Algorithms

- Other clustering methods include **hierarchical agglomerative clustering and density clustering** methods. In hierarchical agglomerative clustering, each object is initially placed in its own cluster. The clusters are then combined with the most similar cluster. This process is repeated until one cluster, which includes all the objects, exists.

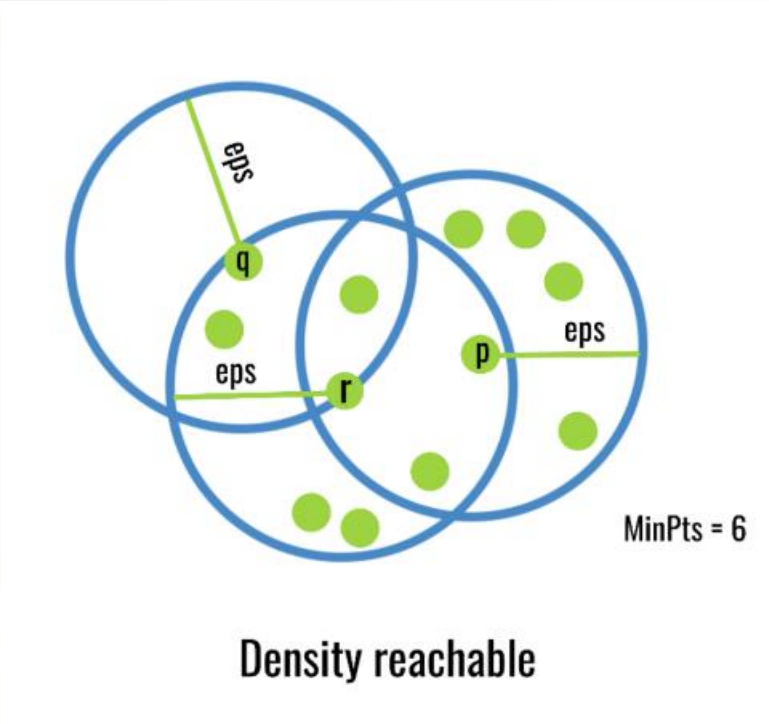
- In **density-based clustering** methods, the clusters are identified by the **concentration** of points. **Density-based** clustering can be useful to identify **irregularly shaped clusters**.

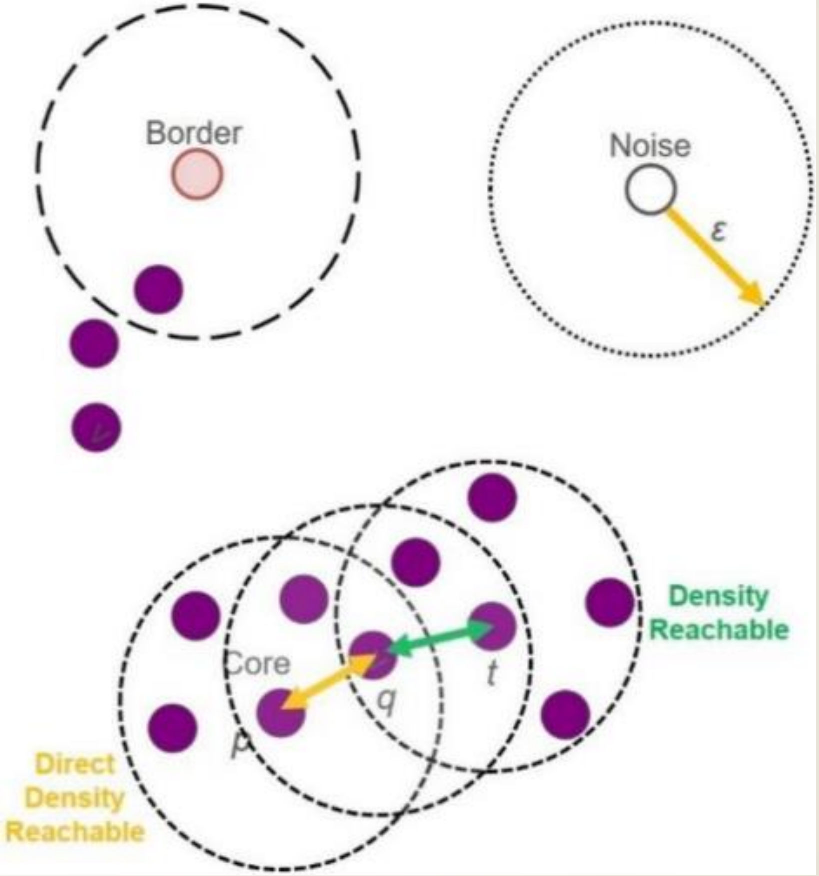
DBScan

- "How can we find dense regions in density-based clustering?" The density of an object **o** can be measured by the number of objects close to ***o***.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** finds core objects, that is, objects that have dense neighborhoods. It connects core objects and their neighborhoods to form dense regions as clusters.
- "How does DBSCAN quantify the neighborhood of an object?" A user-specified parameter ***eps*** > 0 is used to specify the **radius** of a neighborhood we consider for every object. The neighborhood of an object ***o*** is the space within a radius centered at ***o***. Due to the fixed neighborhood size parameterized by ***eps***, the density of a neighborhood can be measured simply by the number of objects in the neighborhood (**min points**).

DBScan

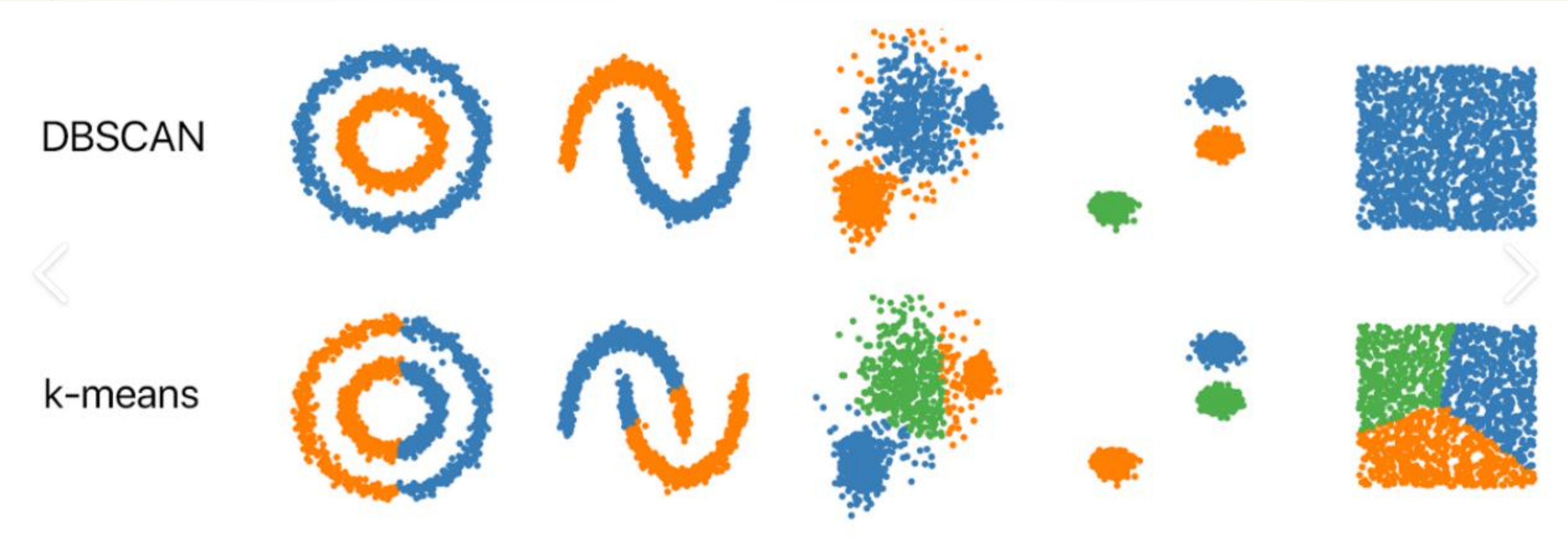
https://www.bing.com/videos/search?q=dbscan+algorithm+video&qs=n&sp=-1&lq=0&pq=dbscan+algorithm+video&sc=6-22&sk=&cvid=F20866D7EA994B10A858FF746FB4DD29&ghsh=0&ghacc=0&ghpl=&ru=%2fsearch%3fq%3ddbscan%2balgorithm%2bvideo%26qs%3dn%26form%3dQBRE%26sp%3d-1%26lq%3d0%26pq%3ddbscan%2balgorithm%2bvideo%26sc%3d6-22%26sk%3d%26cvid%3dF20866D7EA994B10A858FF746FB4DD29%26ghsh%3d0%26ghacc%3d0%26ghpl%3d&view=detail&mmscn=vwrc&mid=4F8A4D63F261C954ABF64F8A4D63F261C954ABF6&FORM=WRVORC





https://www.bing.com/videos/riverview/relatedvideo?&q=dbscan&&mid=35884E6B4BFFDAB2F00035884E6B4BFFDAB2F000&&FORM=VRDGAR

Output of DBScan and K-Means Algorithms



https://www.bing.com/images/search?view=detailV2&ccid=3LdJGuOj&id=8CC2488C0F8B17E22DD76E9D59F2ACDFDA158EEE&thid=OIP.3LdJGuOjolHFG7m8W2NmXQHaCw&mediaurl=https%3a%2f%2fmiro.medium.com%2fmax%2f1200%2f1*KqWIl7sFp1JL0EXwJGpqFw.png&cdnurl=https%3a%2f%2fth.bing.com%2fth%2fid%2fR.dcb7491ae3a3a251c51bb9bc5b63665d%3frik%3d7o4V2t%252bs8lmdbg%26pid%3dImgRaw%26r%3d0&exph=448&expw=1200&q=dbscan&simid=607992534376195191&FORM=IRPRST&ck=D05FFCD01B189C12EA922B280124E7E8&selectedIndex=3&ajaxhist=0&ajaxserp=0

Reference

Data Mining, Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei. MK. Chapter 10.