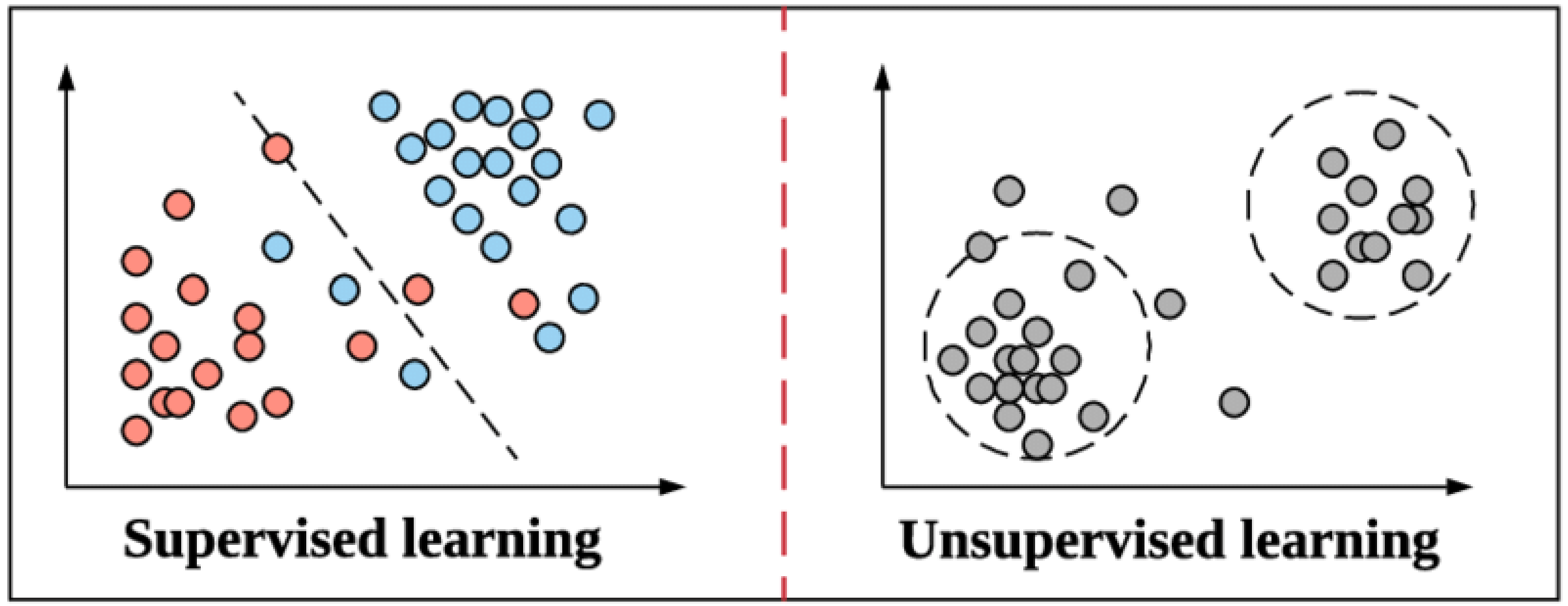
UNSUPERVISED LEARNING

Avni Tonger
A023119820005

UNSUPERVISED LEARNING

Unsupervised Learning Algorithms allow users to perform more complex processing tasks compared to supervised learning.
Although, it can be more unpredictable compared with other natural learning methods.
Unsupervised learning algorithms include clustering, anomaly detection, neural networks, etc.



Goals of Unsupervised Learning

The goal is to discover interesting things about the measurements: is there an informative way to visualize the data? Can we discover subgroups among the variables or among the observations?
We discuss two methods:

- principal components analysis, a tool used for data visualization or data pre-processing before supervised techniques are applied
- clustering, a broad class of methods for discovering unknown subgroups in data.

Challenges of Unsupervised Learning

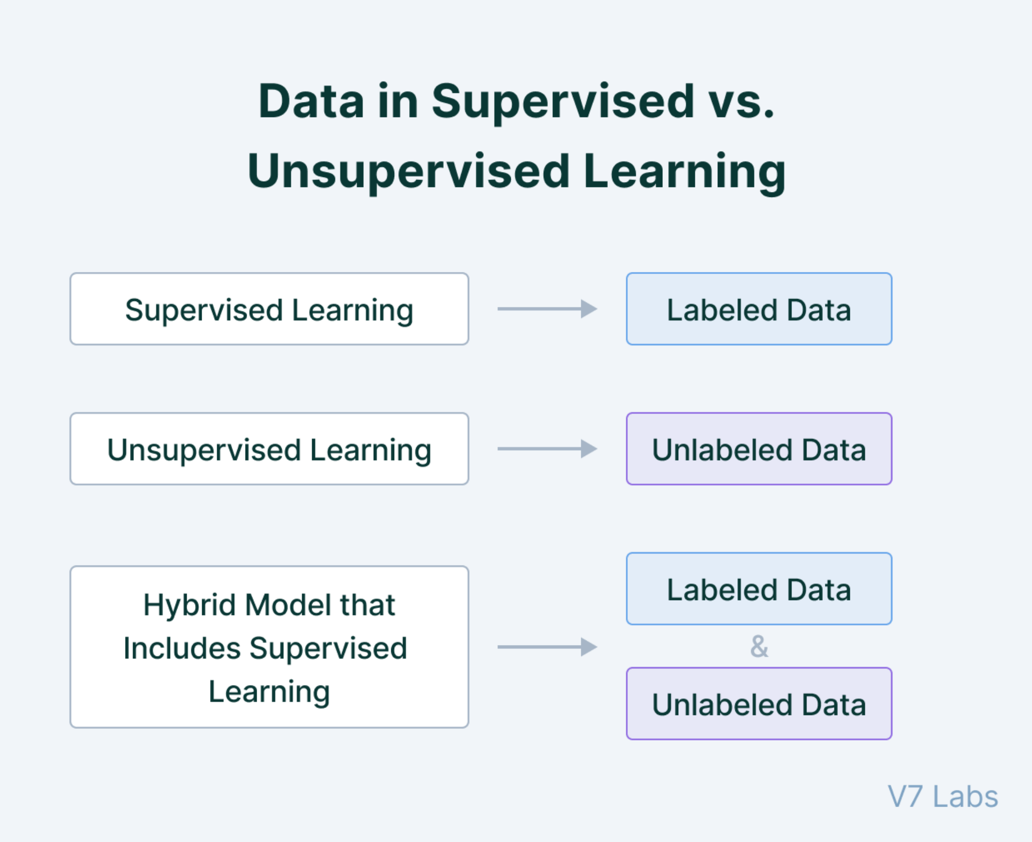
Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response.

But techniques for unsupervised learning are of growing importance in a number of fields:

- subgroups of breast cancer patients grouped by their gene expression measurements,
- groups of shoppers characterized by their browsing and purchase histories,
- movies grouped by the ratings assigned by movie viewers

Advantage of Unsupervised Learning

Easier to obtain unlabelled data — from a lab instrument or a computer — than labelled data, which can require human intervention.



Data in Supervised vs. Unsupervised Learning

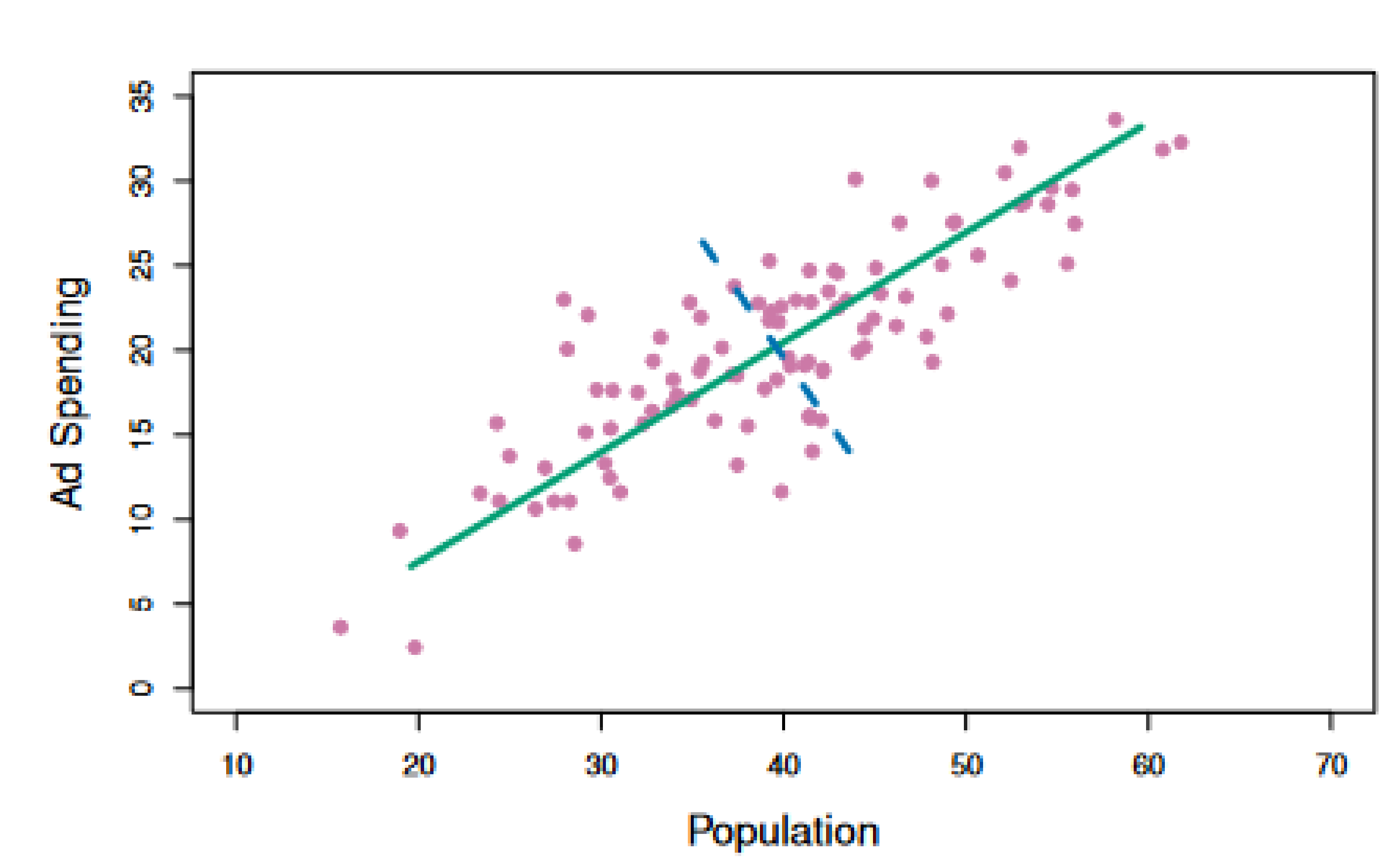
Supervised Learning → Labeled Data

Unsupervised Learning → Unlabeled Data

Hybrid Model that Includes Supervised Learning → Labeled Data & Unlabeled Data

V7 Labs

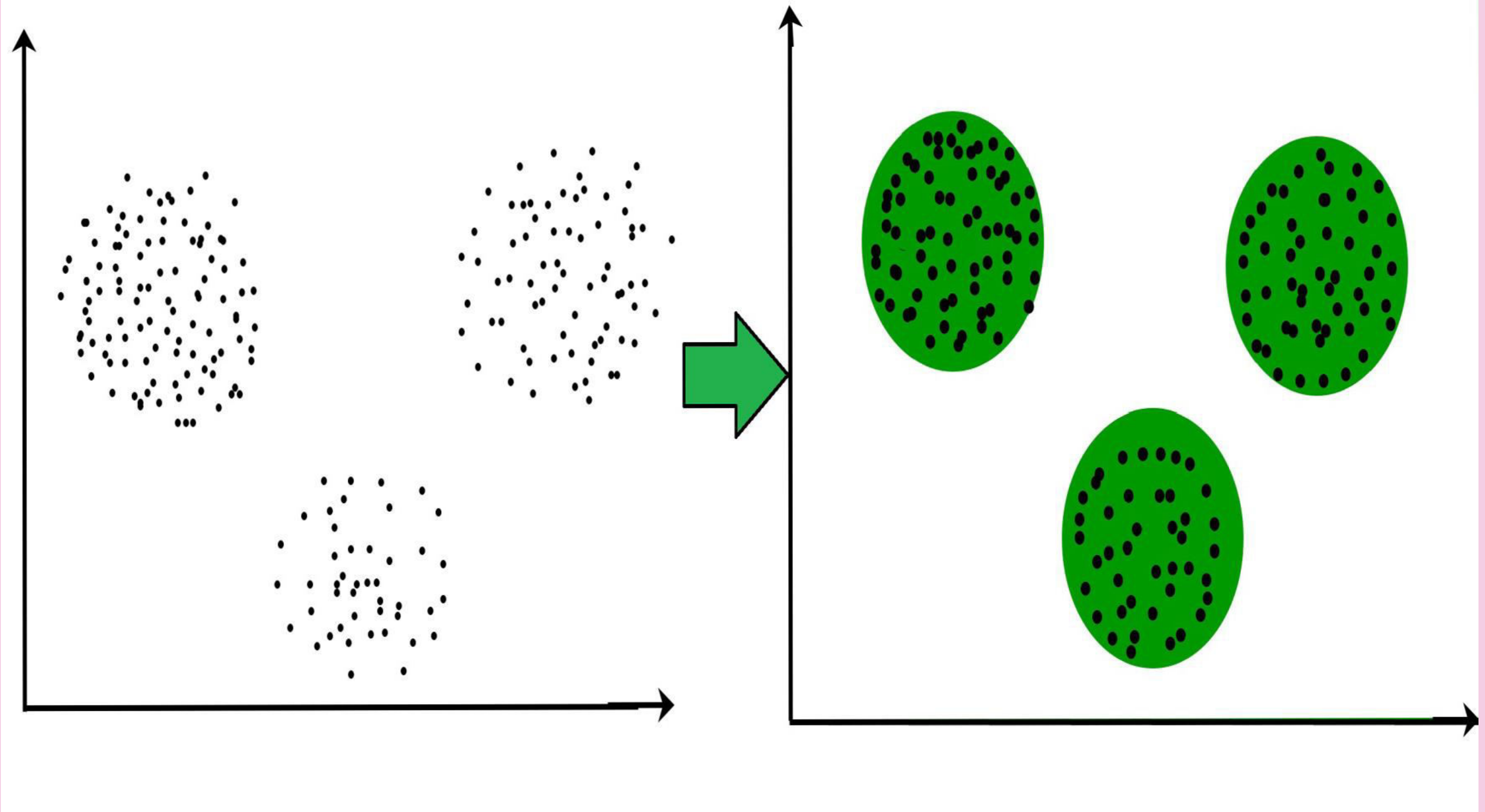
PRINCIPAL COMPONENT ANALYSIS



Principal Components Analysis

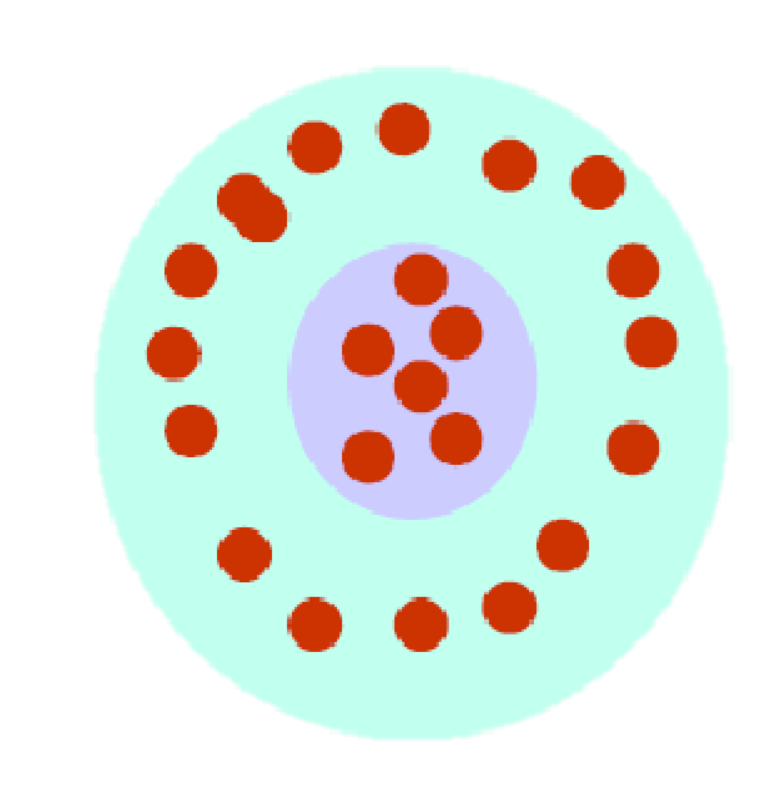
- PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

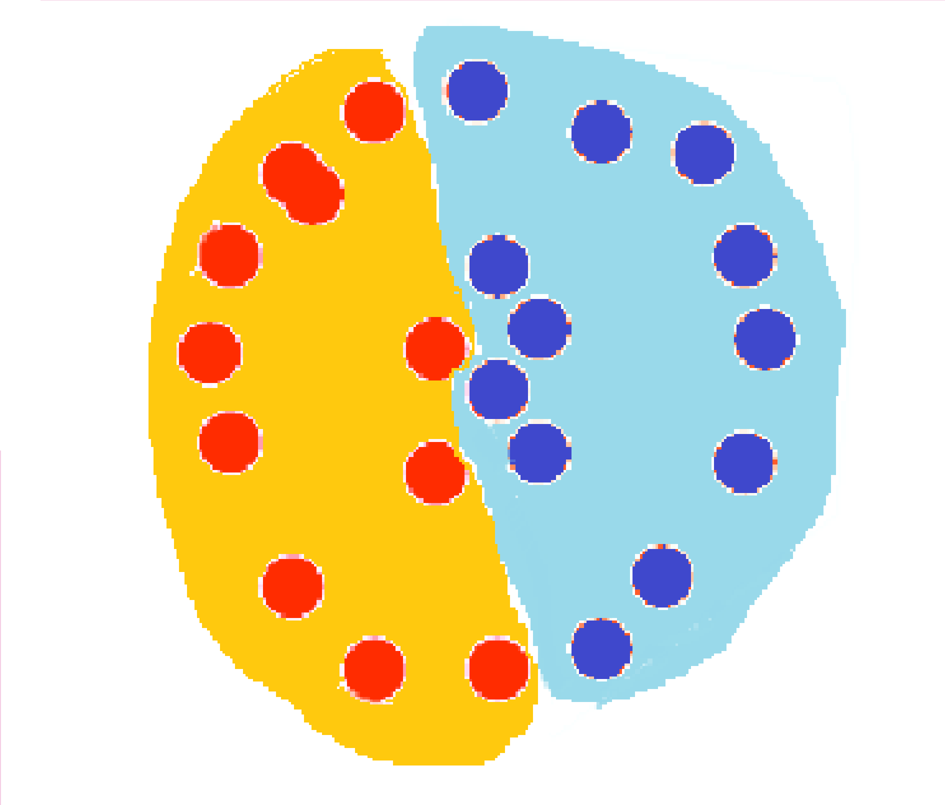
CLUSTERING



Clustering algorithms are used to group data points based on certain similarities. There’s no criterion for good clustering. You can also modify how many clusters your algorithms should identify. It allows you to adjust the granularity of these groups. Clustering determines the grouping with unlabelled data. It mainly depends on the specific user and the scenario. In this clustering method, Data are grouped in such a way that one data can belong to one cluster only.

CLUSTERING





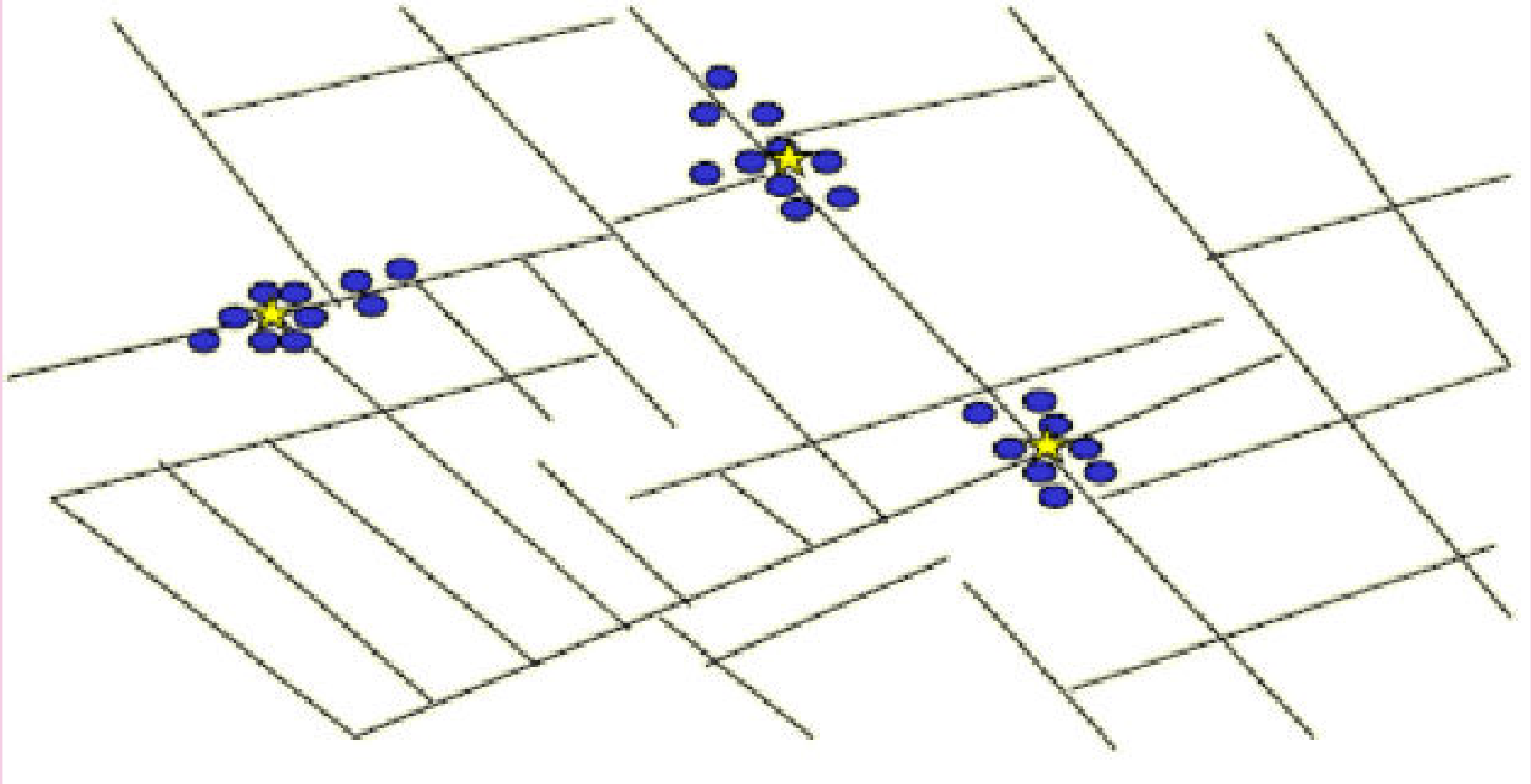
Explaining it in simple terms:

- The organization of unlabelled data into similarity groups called clusters.

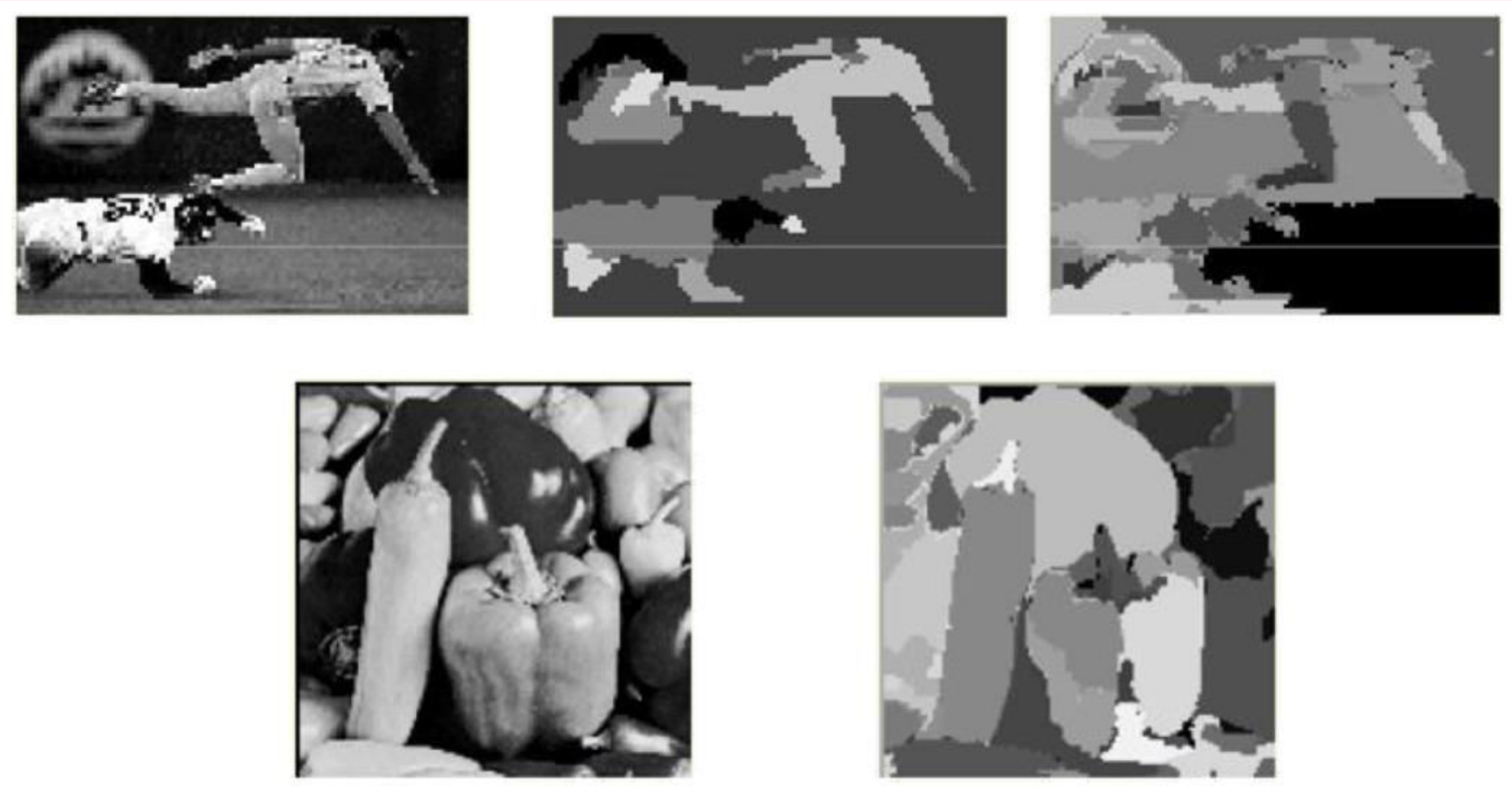
- A cluster is a collection of data items which are “similar” between them, and “dissimilar“ to data items in other clusters.

John Snow and how he implemented clustering for cholera

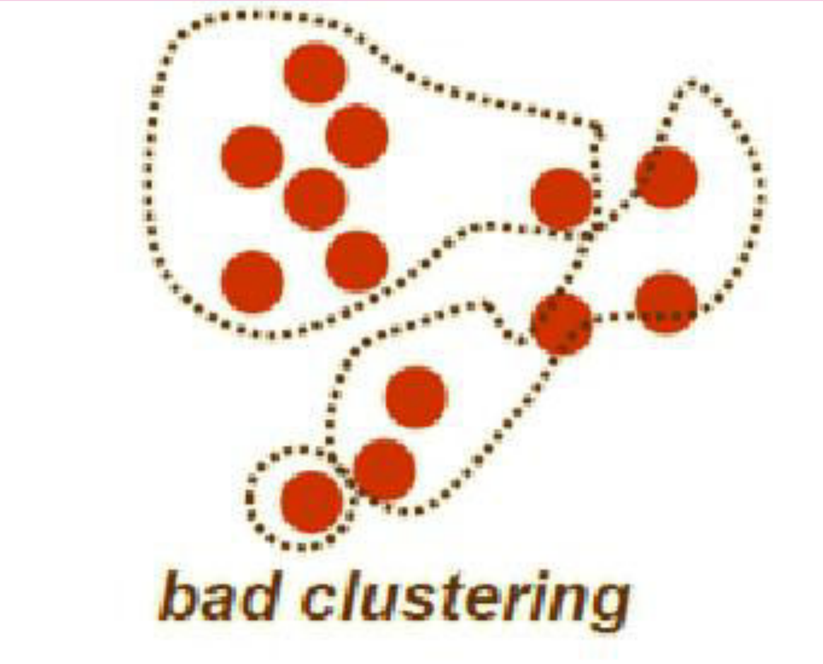
- Plotted the cholera deaths on a map around the 1850s in London.
- The locations indicated the cholera patients were centered around polluted wells, which exposed the problem and the cause.



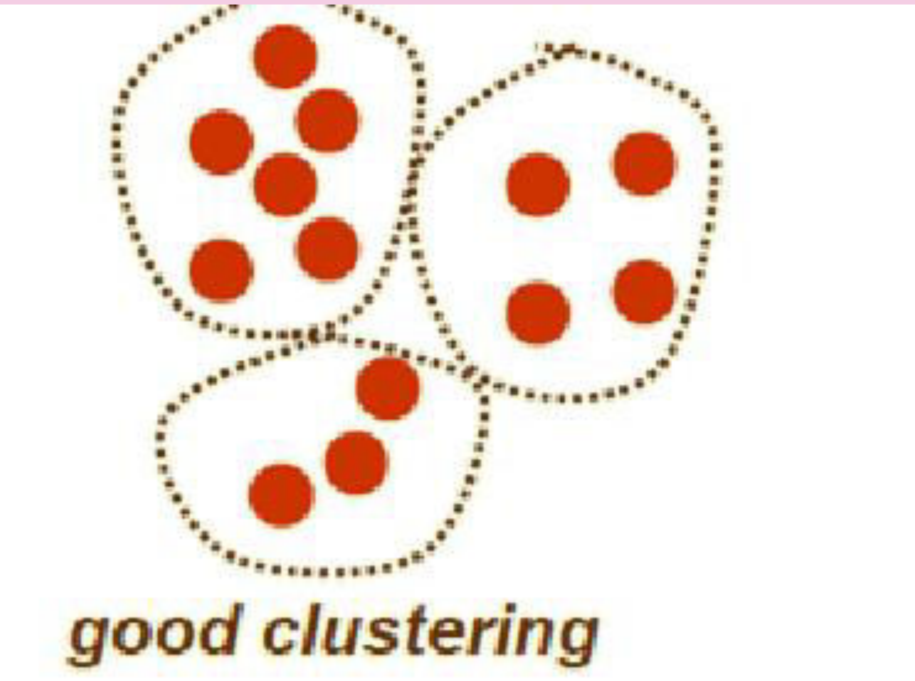
Computer Vision Application: Image Segmentation



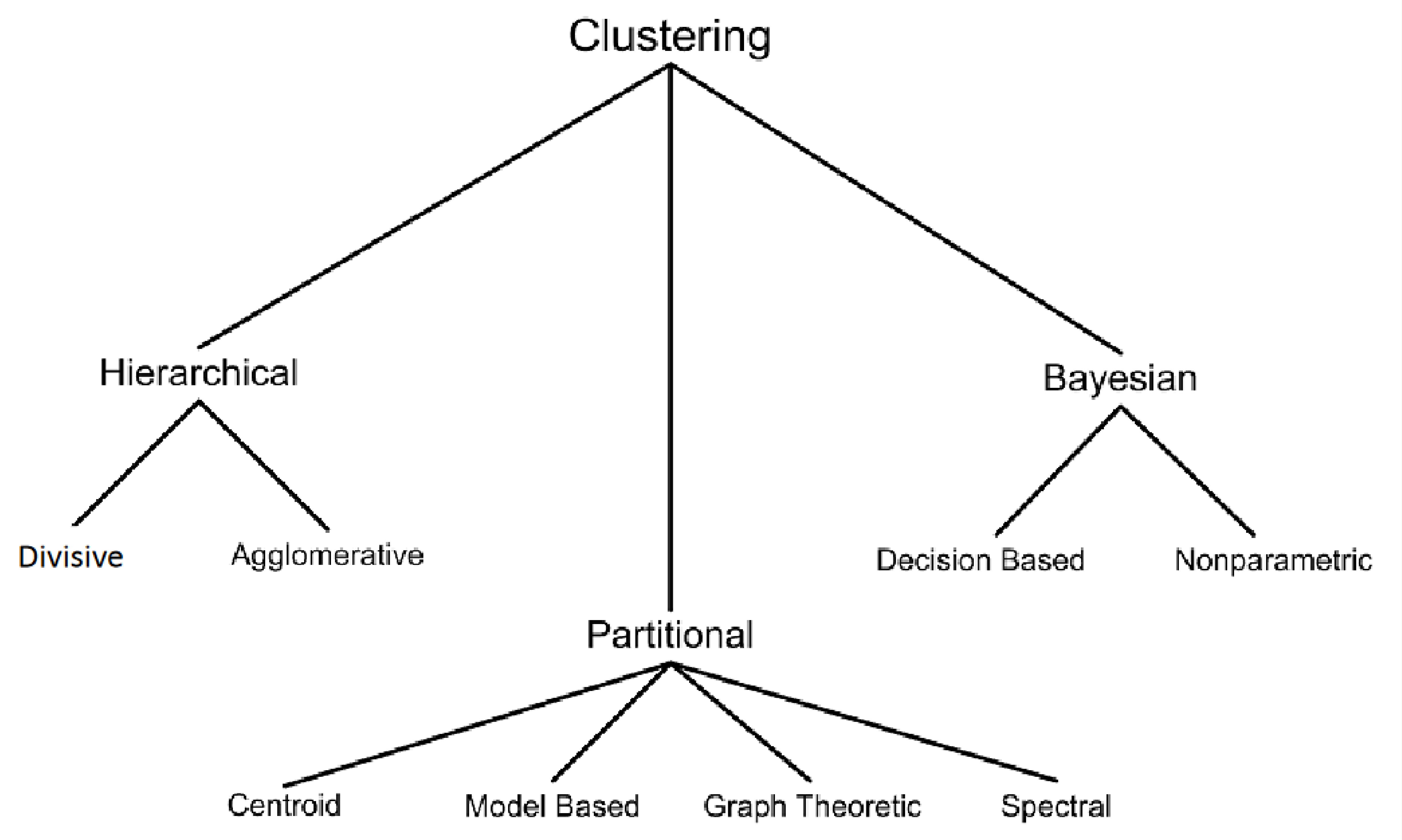
REQUIREMENTS FOR CLUSTERING



- ❑ **Proximity Measure**
  - **based on dissimilarity or similarity measure**
- ❑ **Criterion Function**
- ❑ **Algorithm to compute clustering**



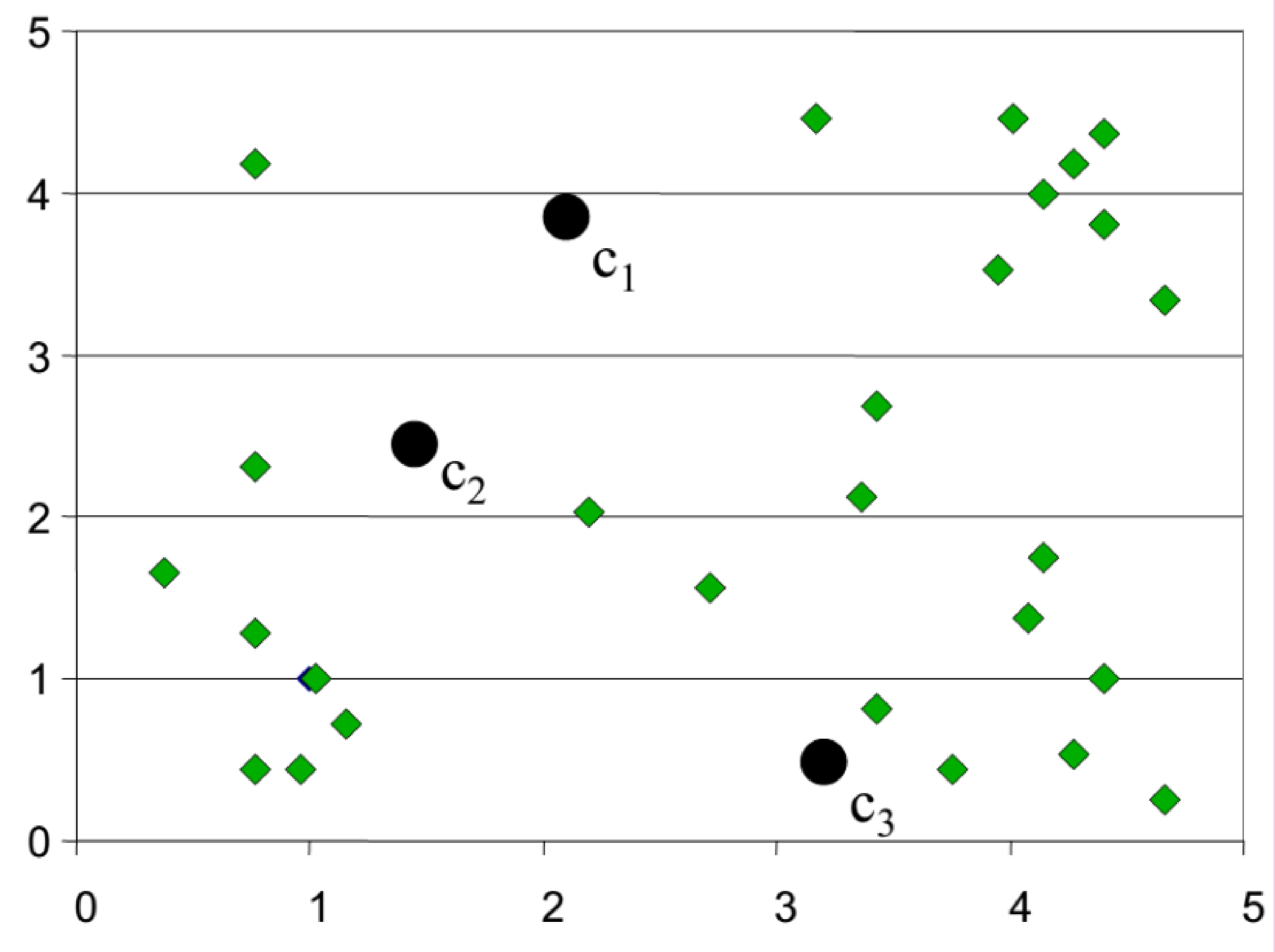
Clustering Techniques



K-Means clustering

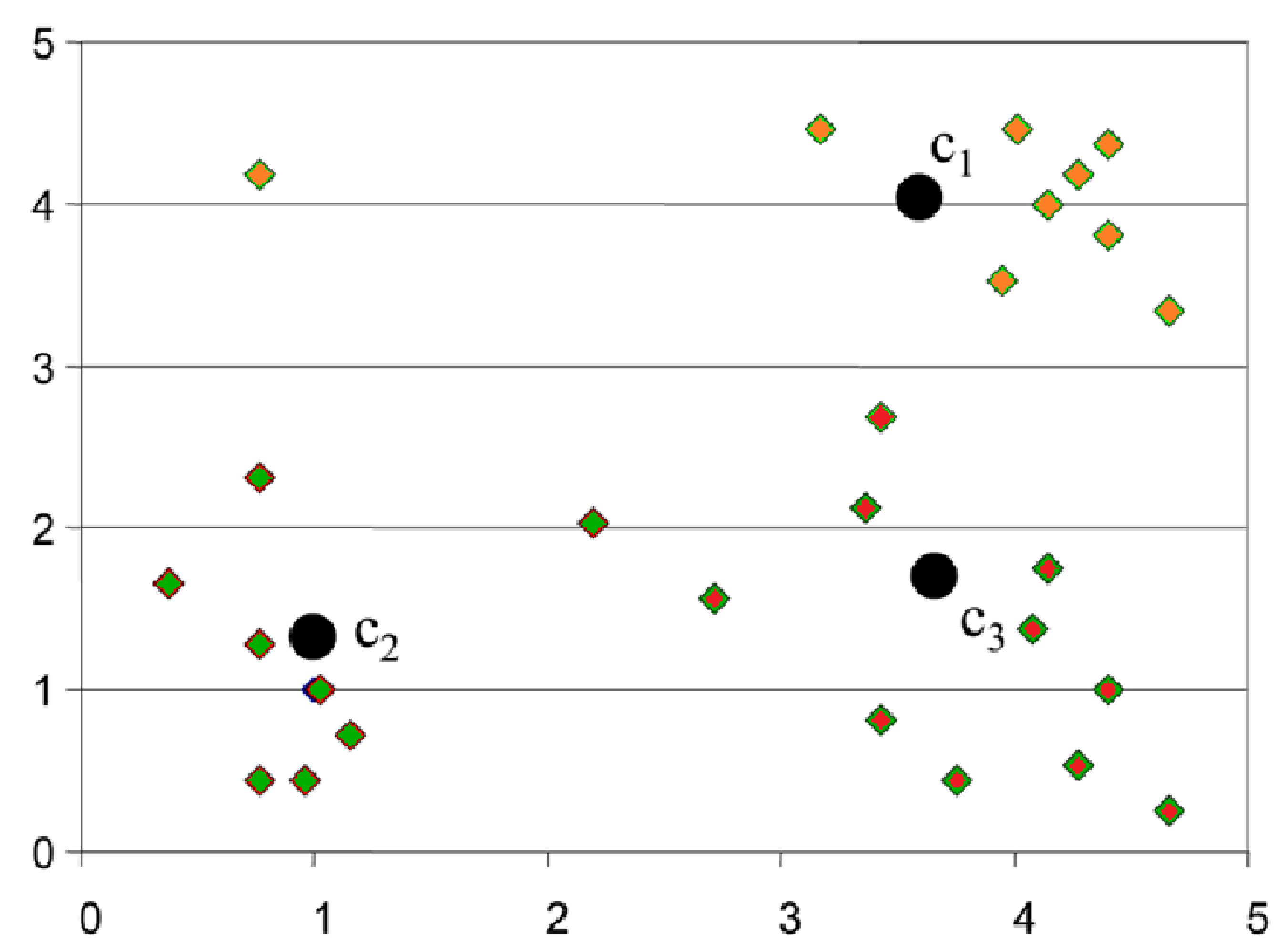
Given k, the k-means algorithm works as follows:

1. Choose k (random) data points (seeds) to be the initial centroids, cluster centres
2. Assign each data point to the closest centroid
3. Re-compute the centroids using the current cluster memberships
4. If a convergence criterion is not met, repeat steps 2 and 3



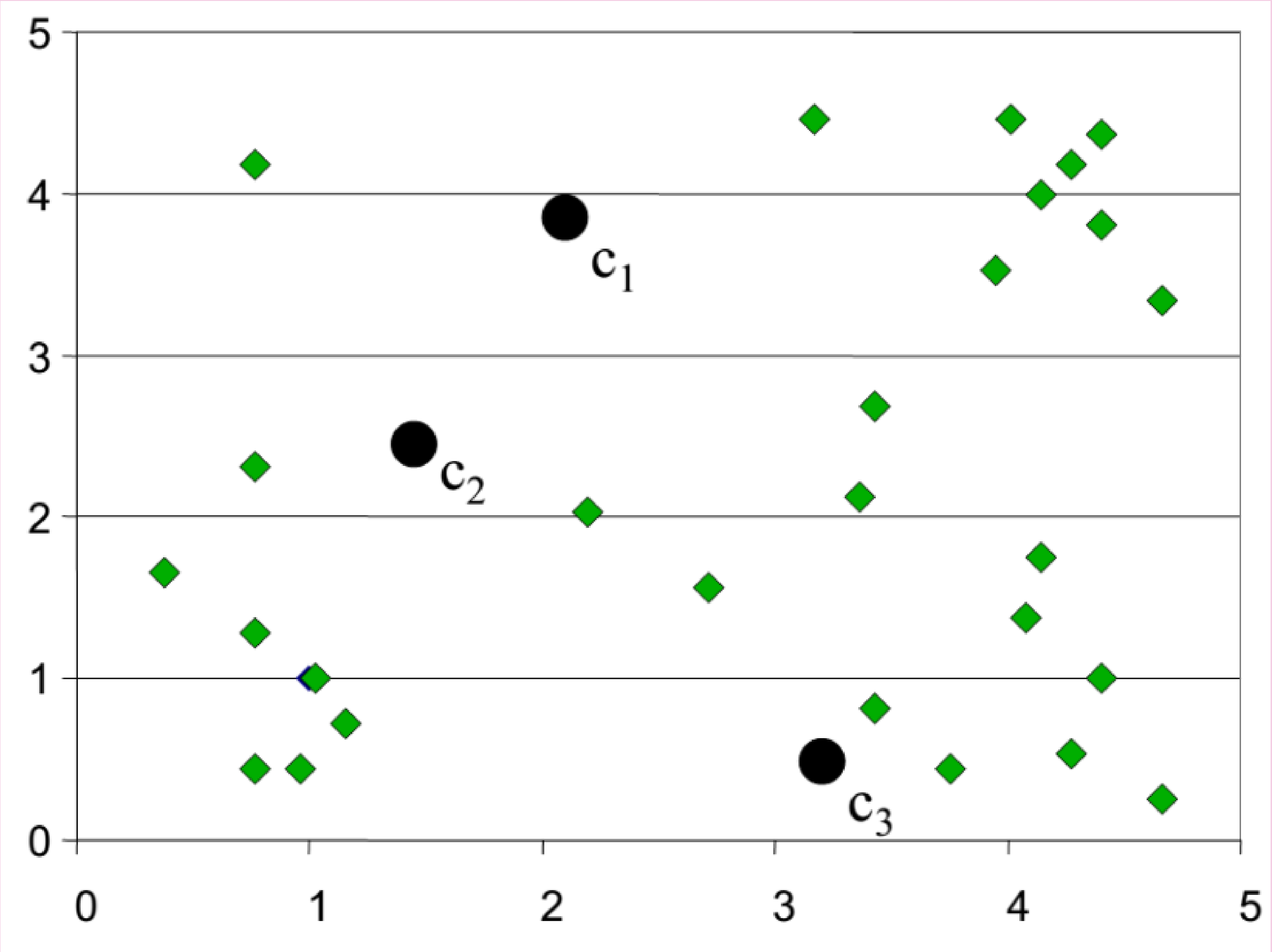
Why use K-means?

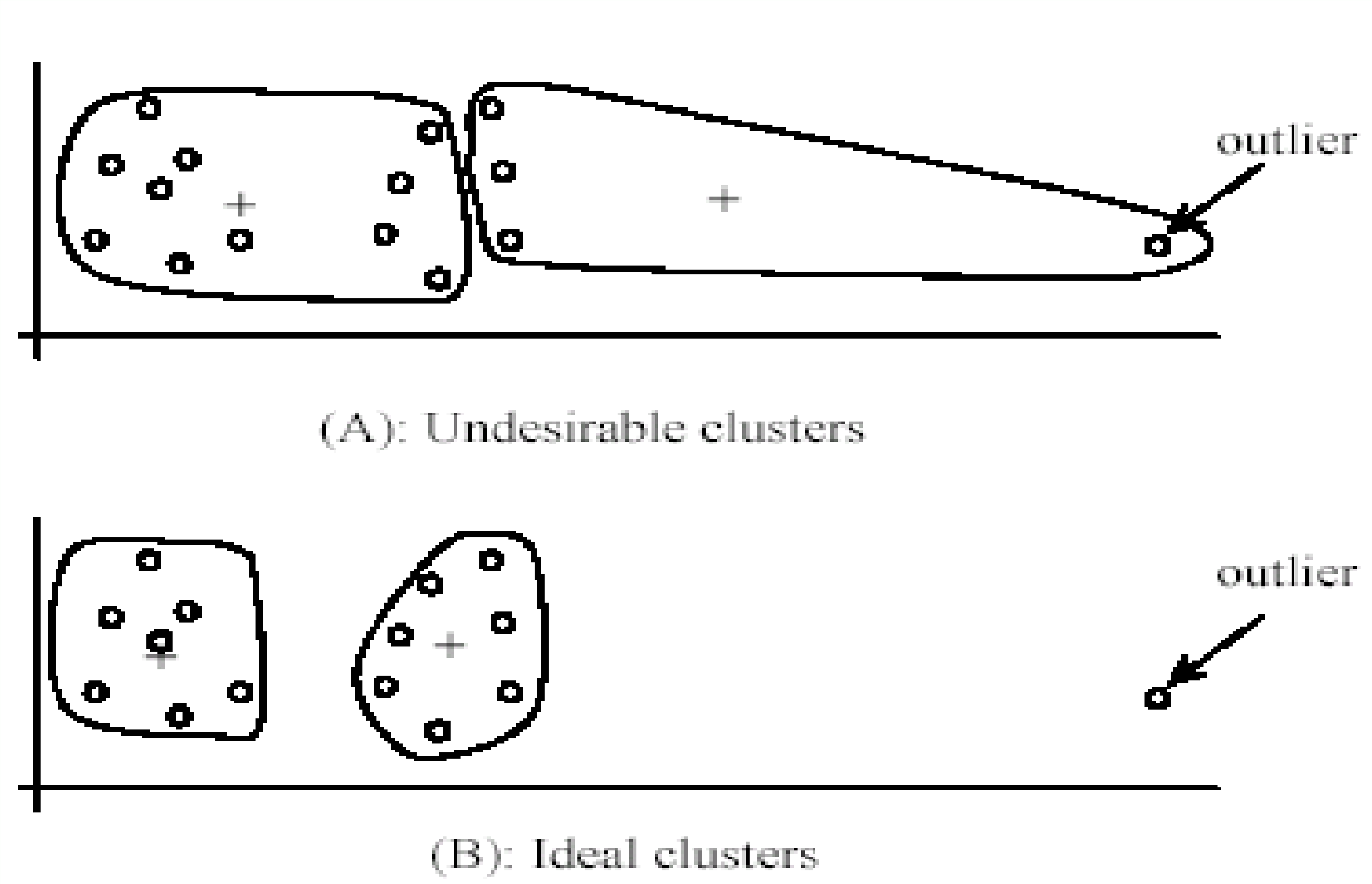
Strengths:
– Simple: easy to understand and to implement
– Efficient: Time complexity: O(tkn),
where n is the number of data points,
k is the number of clusters, and
t is the number of iterations.
– Since both k and t are small. k-means is considered a linear algorithm.
• K-means is the most popular clustering algorithm.
• Note that: it terminates at a local optimum if SSE is used. The global optimum is hard to find due to complexity.



Weaknesses of K-means

The algorithm is only applicable if the mean is defined.
– For categorical data, k-mode - the centroid is represented by most frequent values.
• The user needs to specify k.
• The algorithm is sensitive to outliers
– Outliers are data points that are very far away from other data points.
– Outliers could be errors in the data recording or some special data points with very different values





OUTLIERS

Thanking You