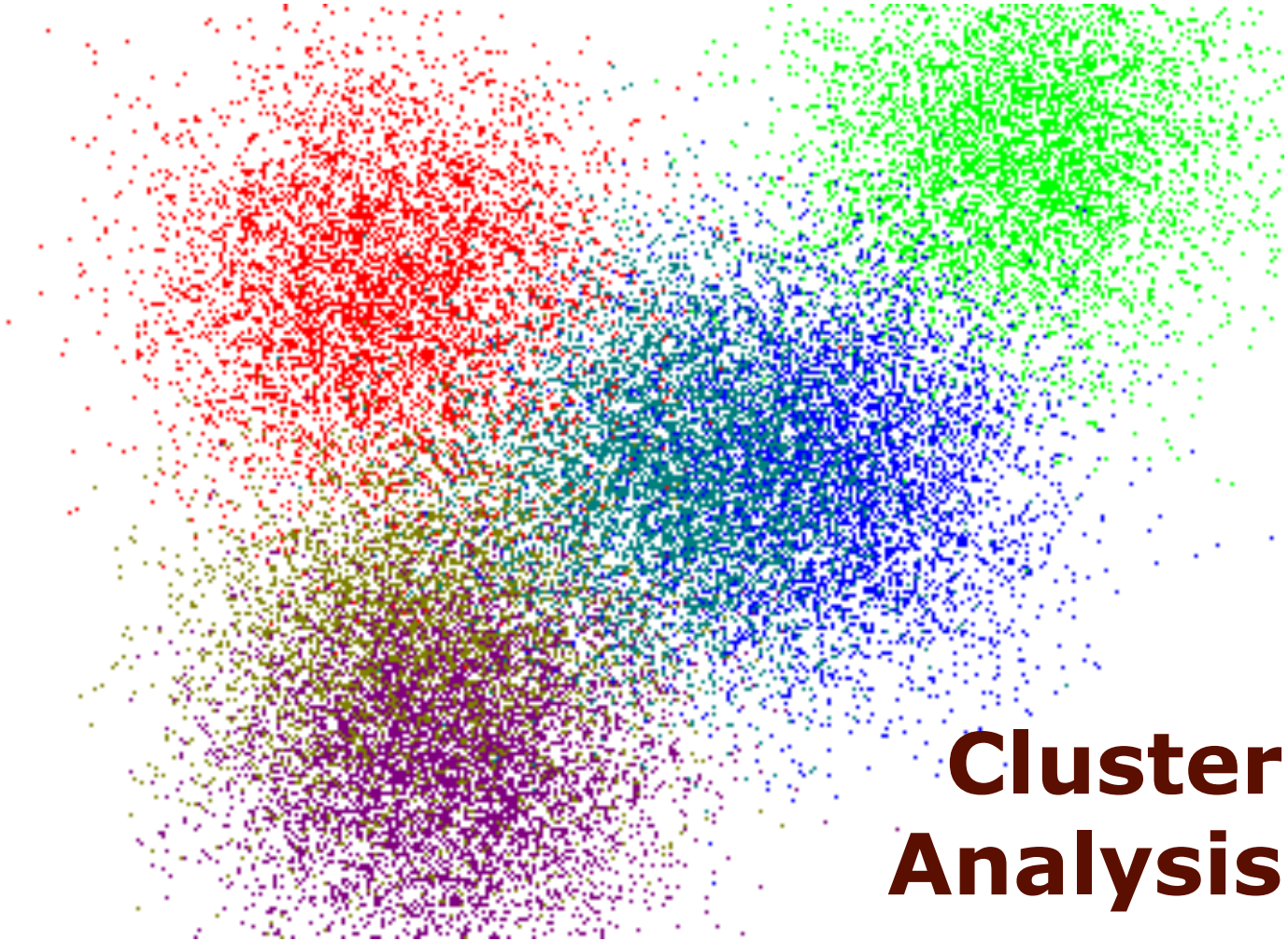


Cluster Analysis

Lucía Santamaría

lucia.santamaria@ymail.com

DATA SCIENCE RETREAT

TODAY

- Clustering and Clustering methods
- Partitioning methods: K-means algorithm*
- Improved seeding: K-means++*
- Finding the K in K-means++ clustering*

(*) Hands-on tutorials on Jupyter [https://github.com/luciasantamaria/geodata]

Clustering is a division of data into groups of similar objects

Clustering is a division of data into groups of similar objects

- Using clusters to represent data loses some details
- But achieves simplification → data modeled by its clusters

From the ML perspective,

clustering is **unsupervised classification**:

no predefined classes

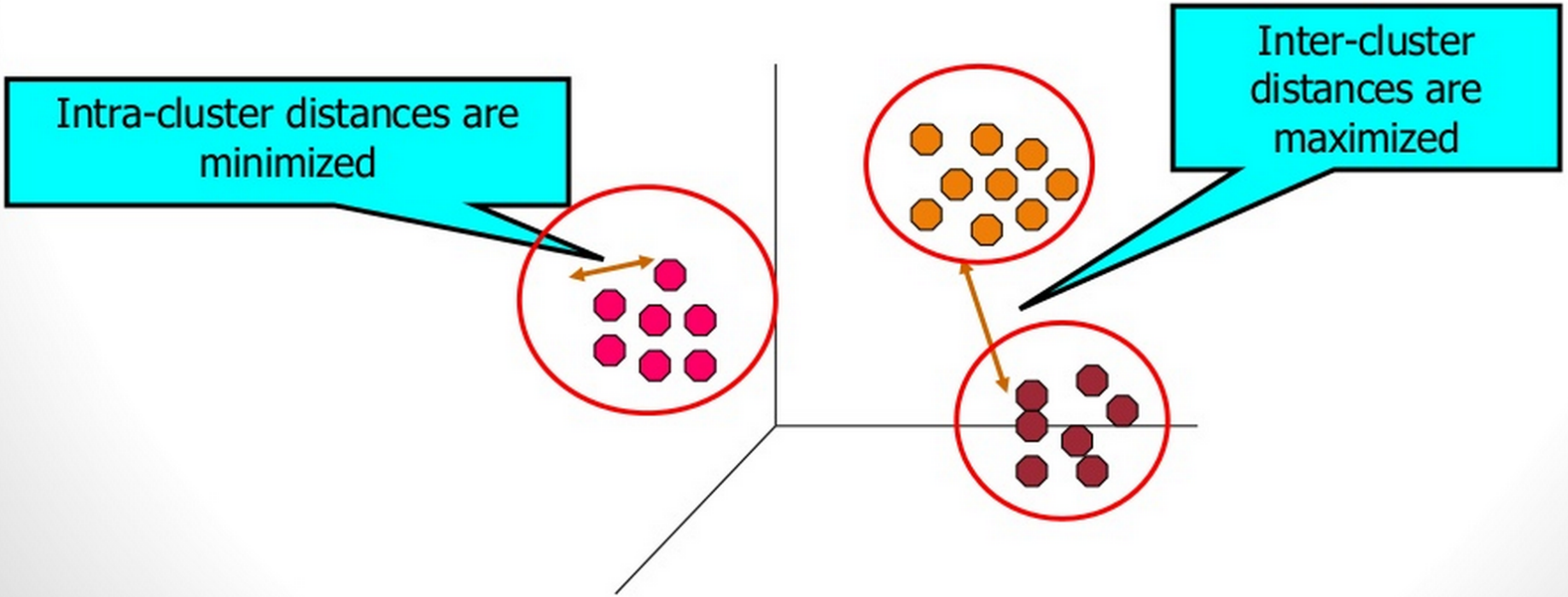
Clusters correspond to *hidden patterns*

→ Clustering is unsupervised learning of a hidden data concept

What does it mean that clustering is good?

1. Closeness between objects inside clusters is essentially more than closeness between clusters themselves (high intra-class similarity and low inter-class similarity)

2. Final clusters correspond to intuitive segmentation of data (they are natural clusters)


Intra-cluster distances are minimized
Inter-cluster distances are maximized

Applications of clustering

Banking

ATM Fraud Detection

Anomalies, outliers

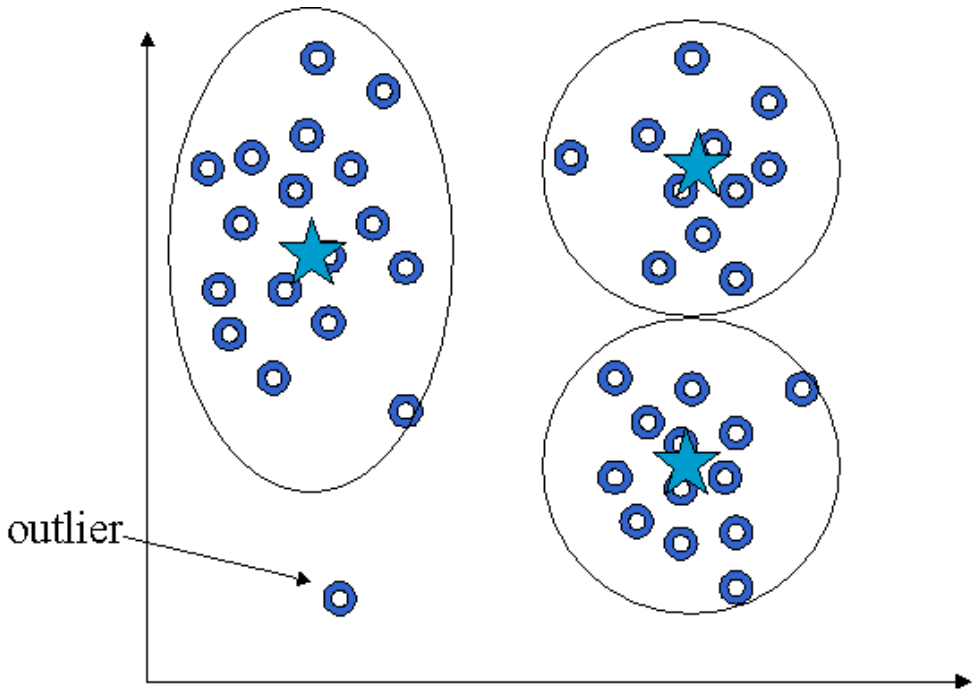


Image processing

Image segmentation

Image recognition

Source image.

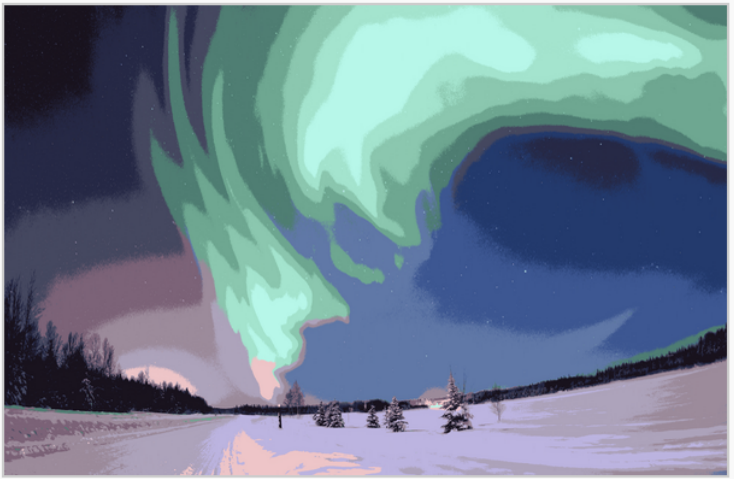
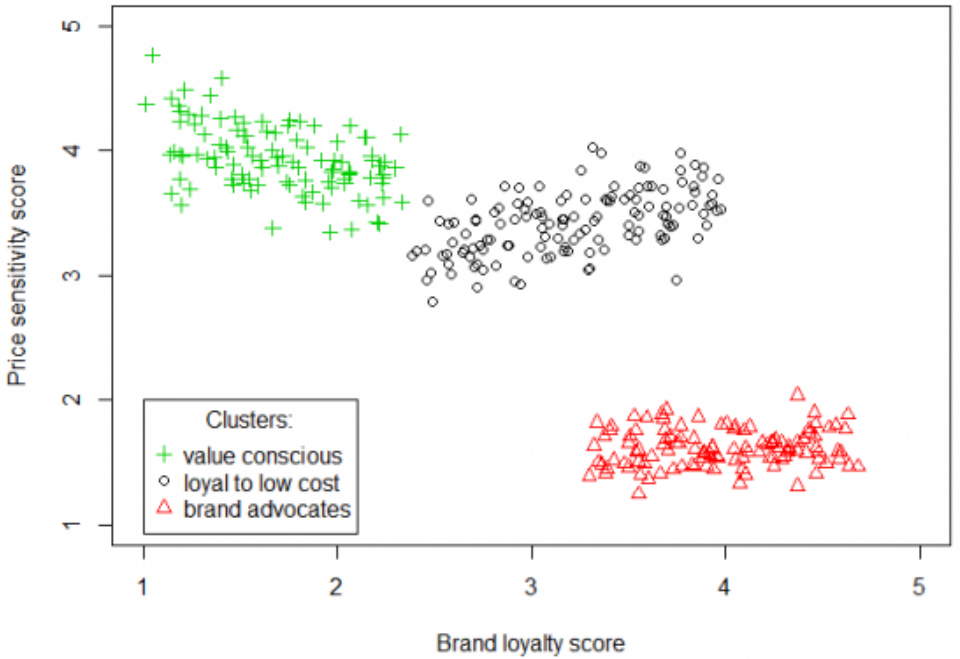


Image after running *k*-means with *k = 16*.

Marketing

Customer segmentation

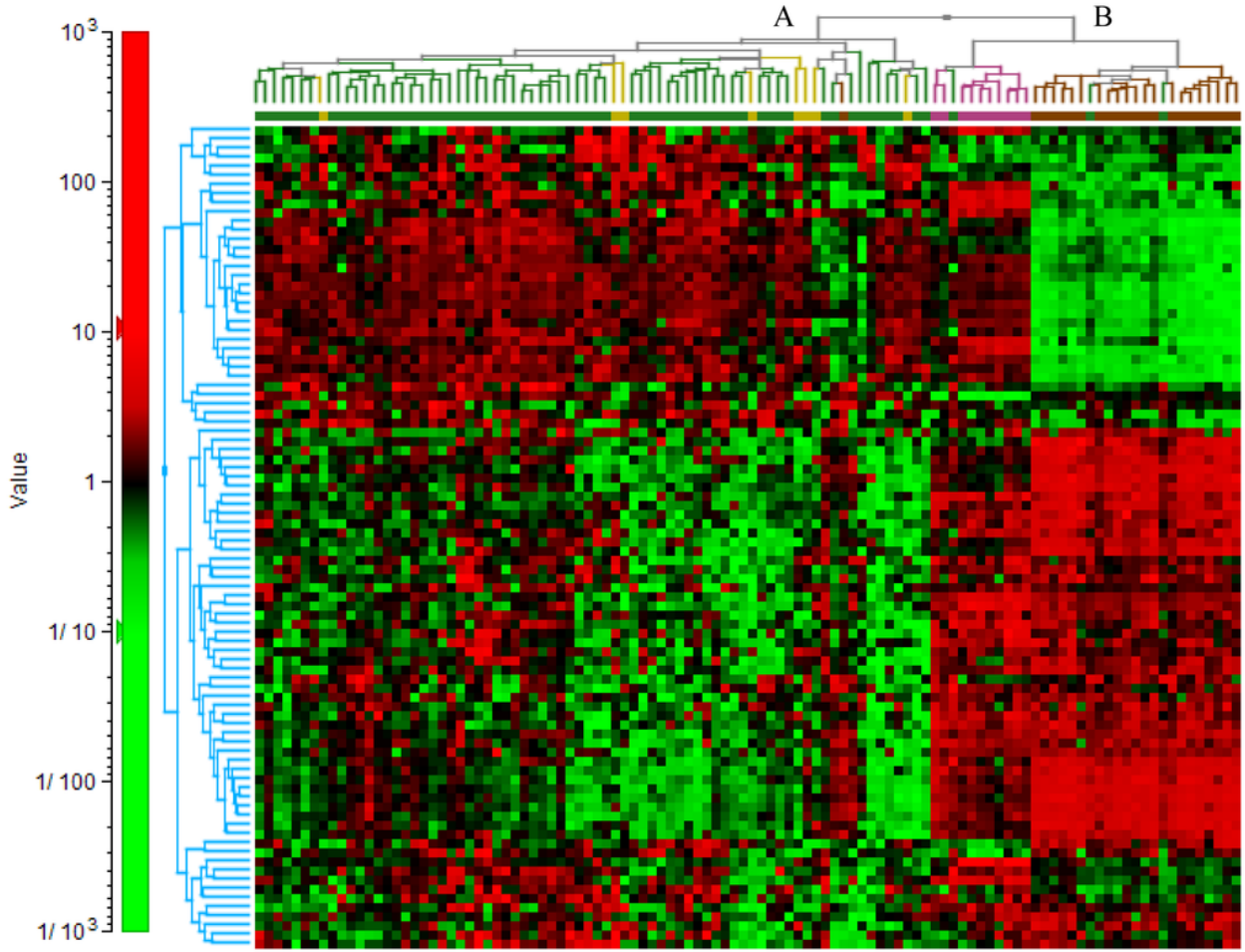
[Targeted marketing]



Biology

Gene analysis

Medical image processing



2D hierarchical clustering of the 90 gene set. Cluster A contained all cancer-prone samples, Cluster B contained all cancer-free clusters

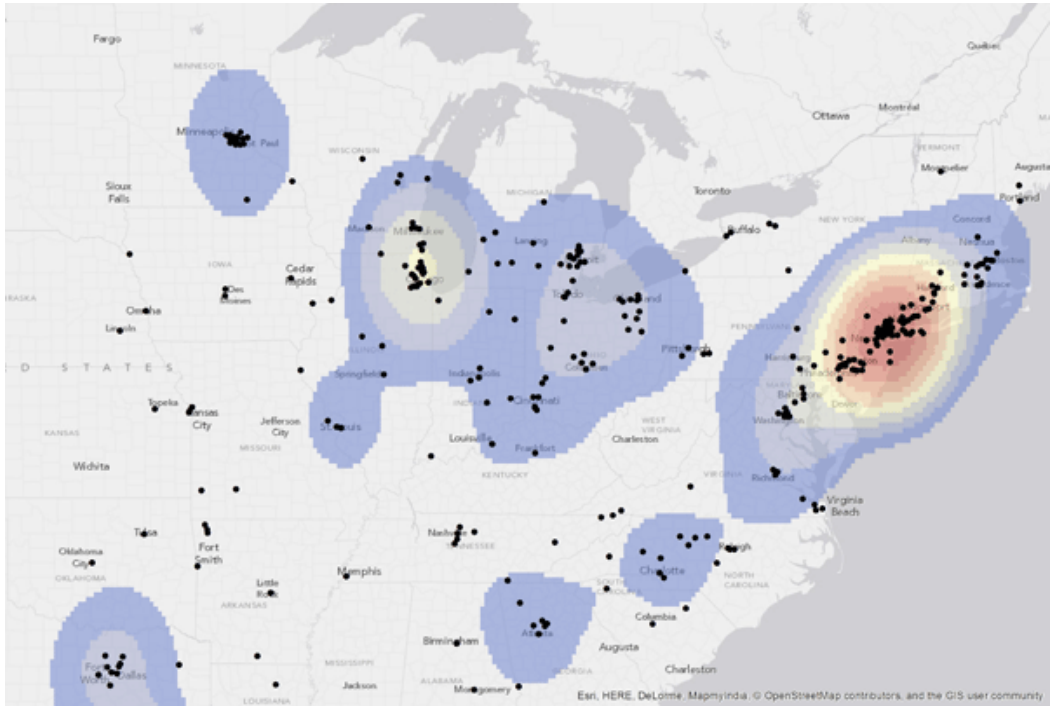
GIS

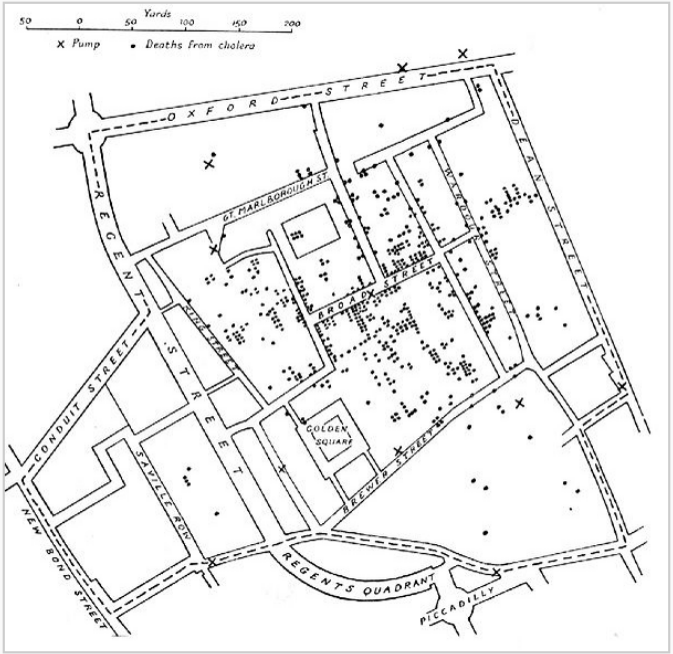
Land use

City planning

Earthquake studies

…





E. W. Gilbert's version (1958) of John Snow's 1855 map of the Soho cholera outbreak showing the clusters of cholera cases in the London epidemic of 1854

Actually, the clustering of cholera cases in London was possibly the first application of GIS!

Clustering methods

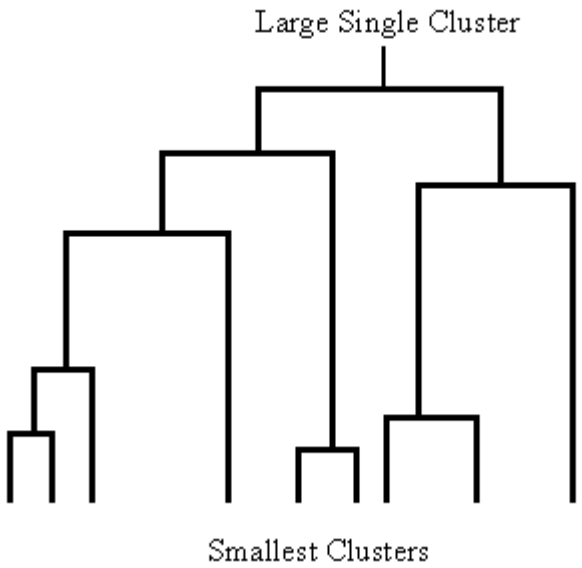
Hierarchical

Set of nested clusters organized as a hierarchical tree

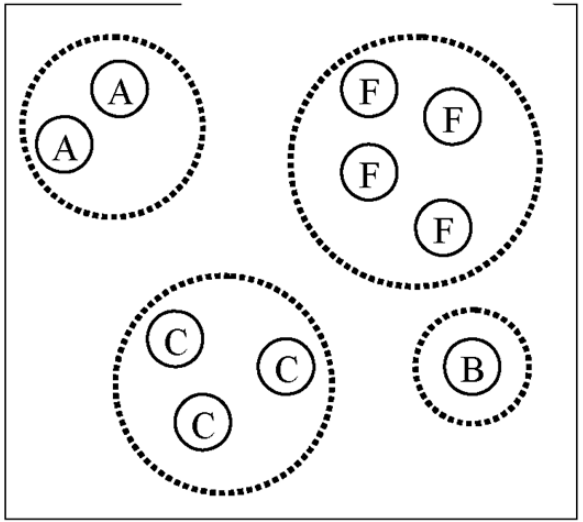
Partitional

Division of data into non-overlapping subsets such that each object is in exactly one subset

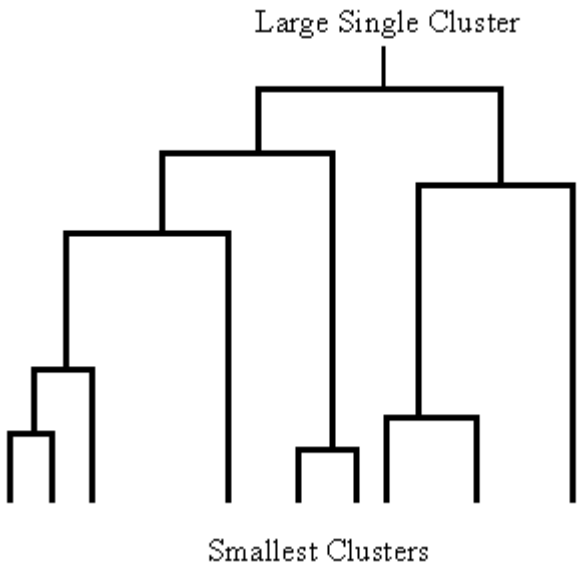
Hierarchical



Partitional



Hierarchical



Partitional