Cluster Analysis - I

George Michailidis
Department of Statistics
The University of Michigan

March 3, 2009

1. Objective

Given a multivariate data set -N objects, p variables- find meaningful groups of the objects.

Some important issues:

- Define “meaningful”
- Homogeneity vs Separation
- Input data: dissimilarities vs profiles

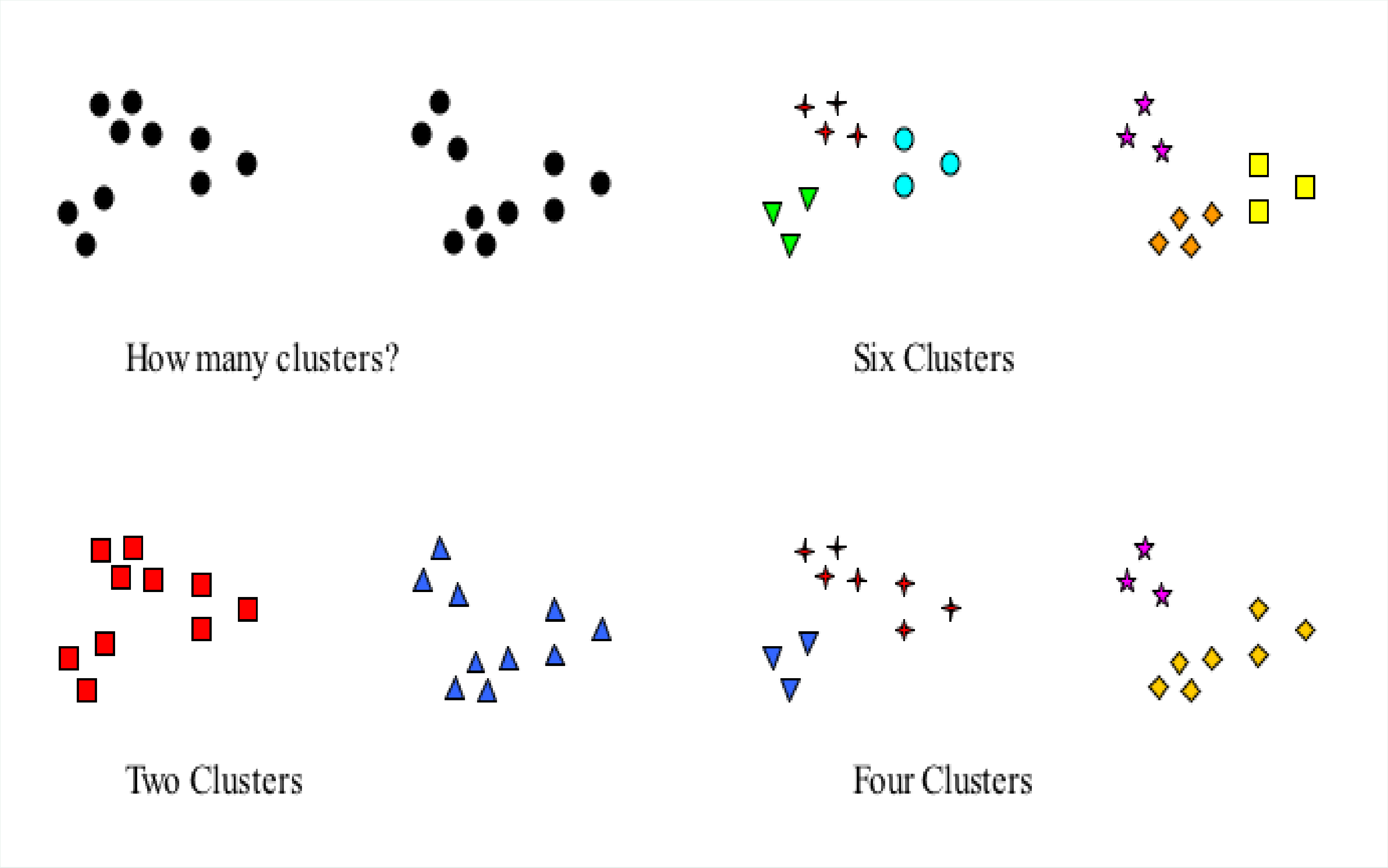
Some applications

- Gaining insight:
  group genes or proteins that have similar functionality
  group stocks with similar price fluctuations
  group document into thematic entities
- Summarization:
  Reduce the size of large data sets

What IS NOT clustering:

- Classification: number of classes predetermined and class labels available
- Simple segmentation (e.g. group people by height)
- Data querying: the groups are the result of an external specification

Difficulties in defining meaningful clusters:



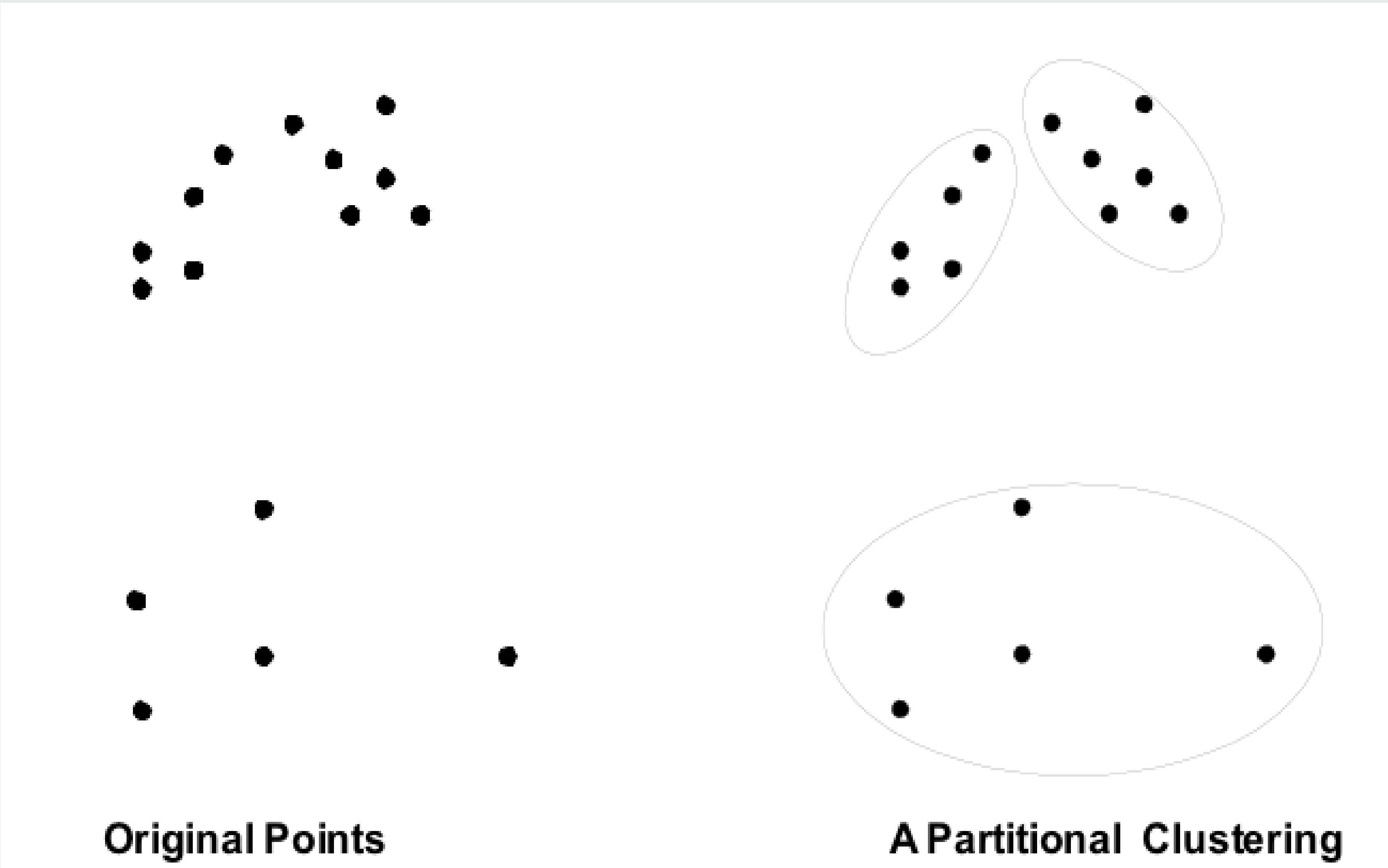
Types of cluster analysis:

- Partition methods:

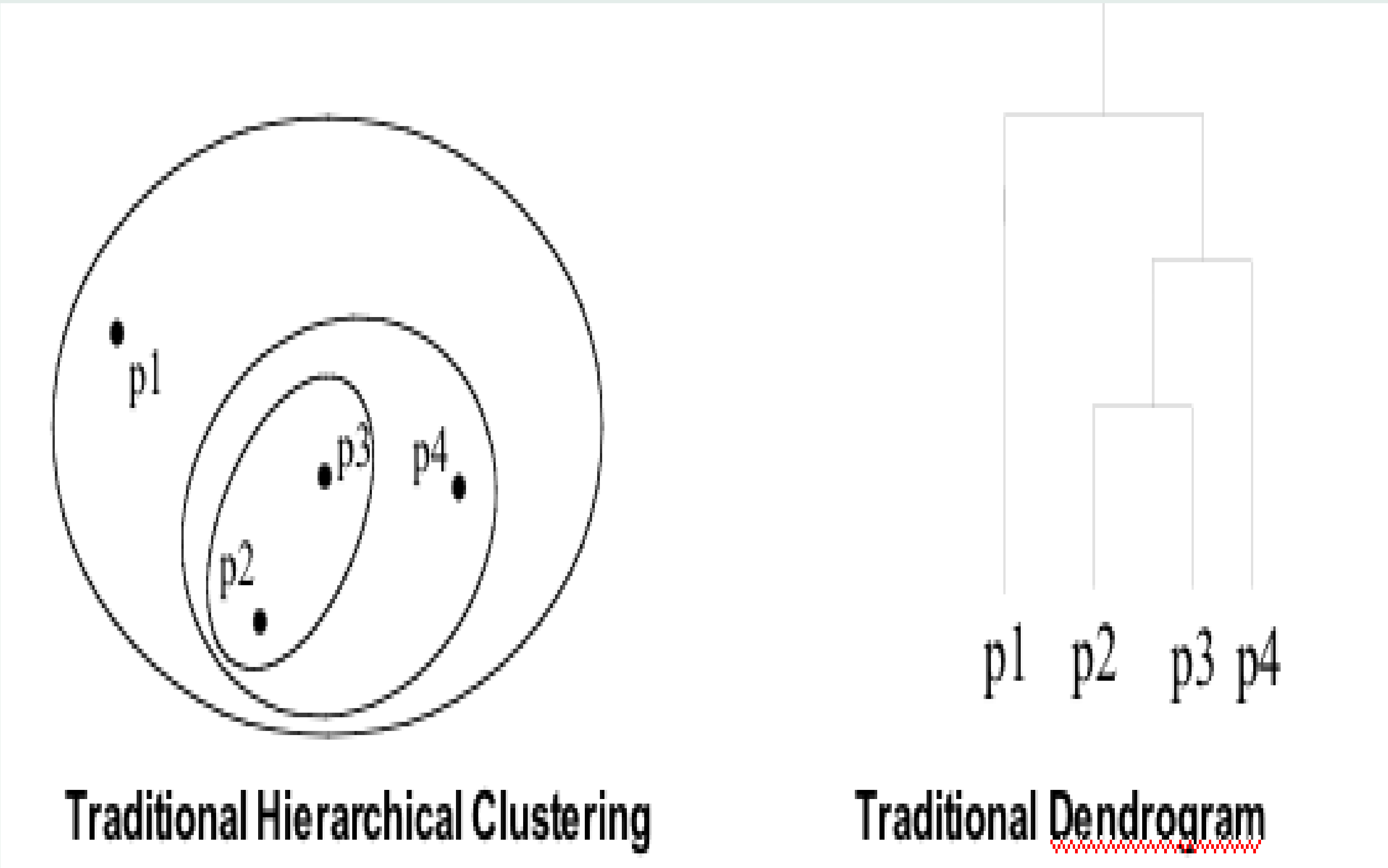
  Objects are partitioned into non-overlapping groups and each object belongs to one group only
- Hierarchical methods:

  Objects are partitioned into nested groups that are organized as a hierarchical tree

Some toy examples



Some toy examples (ctd)



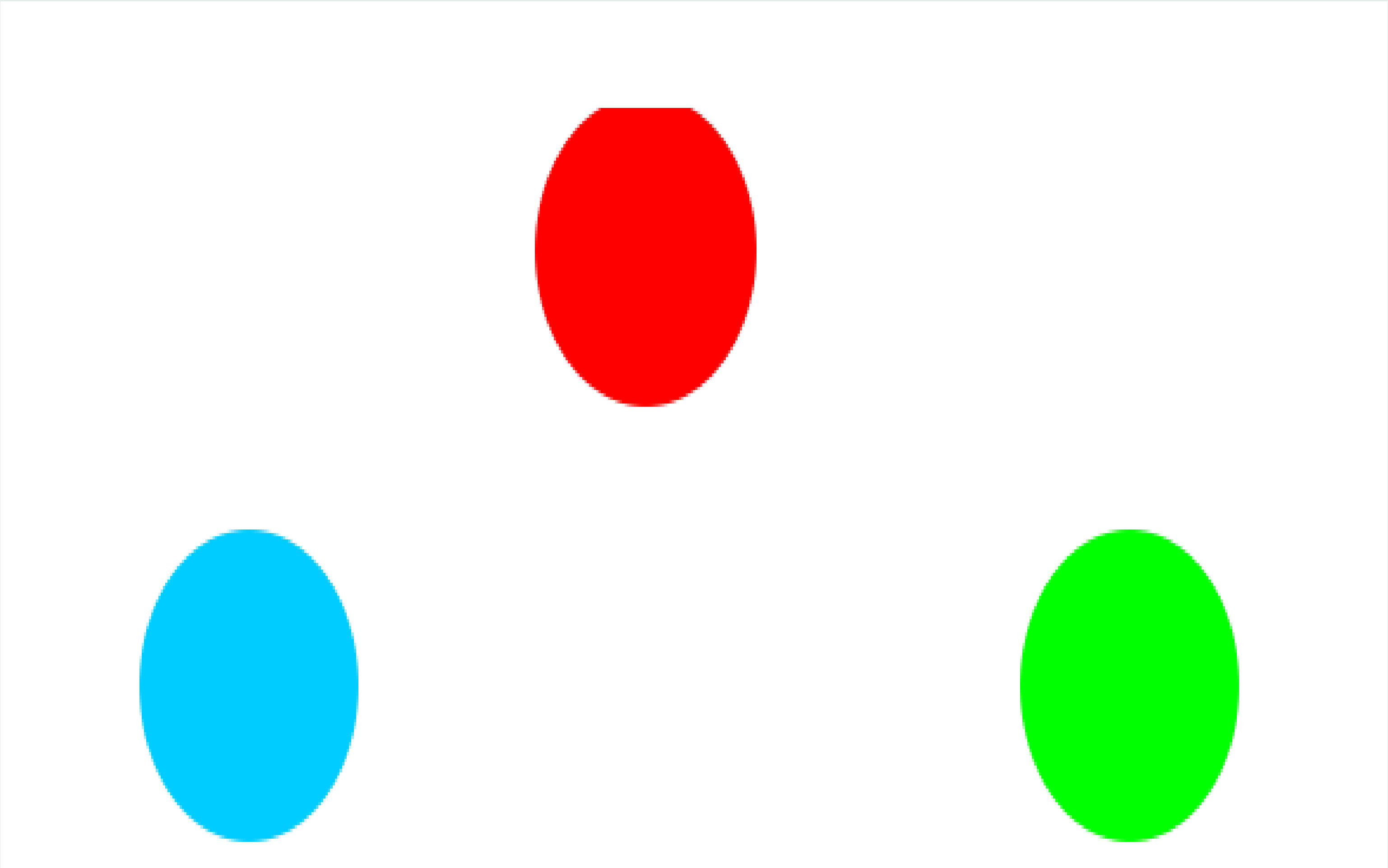
Some other distinctions in cluster analysis

- Fuzzy clustering:
  every object can belong to several clusters with a certain probability (probabilities over all clusters should sum to one)

- Partial clustering:
  find groups for a subset of the objects; ignore the rest
- Cluster heterogeneity:
  shape, size and density of different clusters

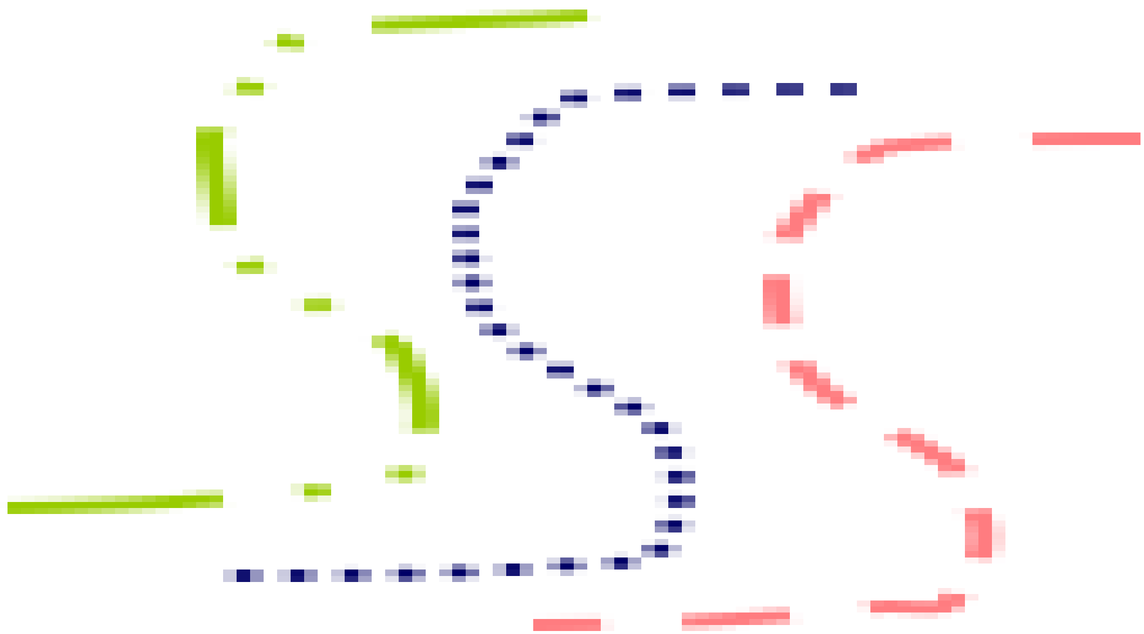
Some desirable properties: Separation

Objects in one cluster are closer (more similar) to every other object in the same cluster and further apart than all objects in other clusters.



Some desirable properties: Homogeneity

Clusters have a small 'diameter'



Example of lack of homogeneity

2. Data structures:

- Data matrix (N objects, p variables)
  most suitable for partition methods
- Similarity/dissimilarity matrix ($N \times N$ calculated from the data matrix)
  most suitable for agglomerative methods Gap between support vectors is called the margin

3. Similarity - Dissimilarity:

- Similarity measure:

  A numerical measure that indicates how similar two objects are; the higher its value the higher the similarity

  In many cases it is normalized to have a [0,1] range

  Formally, a similarity measure satisfies:

  $s(i, j) \geq 0$ and $s(i, j) = s(j, i)$

- Dissimilarity measure:

  A numerical measure that indicates how different two objects are; the lower its value the more similar the objects are

  The lowest value is 0 (all diagonal elements are 0)

  Upper bound varies

  Formally, a dissimilarity measure satisfies:

  $\delta(i, j) \geq 0$ and $\delta(i, j) = s(j, i)$

  A distance (d) is automatically a dissimilarity measure

Some common dissimilarity (distance) measures:

- Euclidean distance: $\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}$ (p variables)
- Manhattan distance: $\sum_{k=1}^{p} x_{ik}-x_{jk}$ (p variables)
- Minkowski distance: $(\sum_{k=1}^{p}(x_{ik}-x_{jk})^q)^{1/q}$ (p variables)
  When $q \to \infty$ it corresponds to the ℓ_∞ distance
- Mahalanobis distance: consider p-dimensional vectors x_i and x_j;
  $(x_i - x_j)'\Sigma^{-1}(x_- x_j)$, where Σ is the covariance matrix of all x_k vectors

Some common similarity measures:

- Cosine measure: $(\sum_{k=1}^{p}(x_{ik}x_{jk}))/((\sum_{k=1}^{p} x_{ik}^{2})(\sum k = 1$
  $(< x_i, x_j > / x_{i2}x_{j2})$ (p variables)
- Jaccard-Tanimoto coefficient: $< x_i, x_j > / {x_i}_2^2 + {x_j}_2^2 - < x_i, x_j >$ (p variables)
- Correlation coefficient

Combining similarity/dissimilarity measures:

Many data sets have mixed measurement variables (nominal, ordinal, numerical).

One approach is to use an appropriate (dis)similarity measure for each variable and then combining them, provided that they take values in $[0, 1]$

$$\frac{\sum_{k=1}^{p} w_k \delta_k(i,j)(s_k(i,j))}{\sum_{k=1}^{p} w_k}$$

Notice that if $w_k = 0$, this approach can also handle missing data

4. Hierarchical clustering:

- Produces a sequence of solutions (nested clusters), organized in a hierarchical tree structure - For not enormously large data sets, the solution can be visualized by a dendrogram

Advantages of hierarchical clustering

- Gives a family of possible solutions - Computationally fast - In many cases, results in meaningful taxonomies

Disadvantages of hierarchical clustering

- No optimization criterion - Final solution chosen by the data analyst - Different merging (splitting) criteria give rise to different solutions

Types of hierarchical clustering

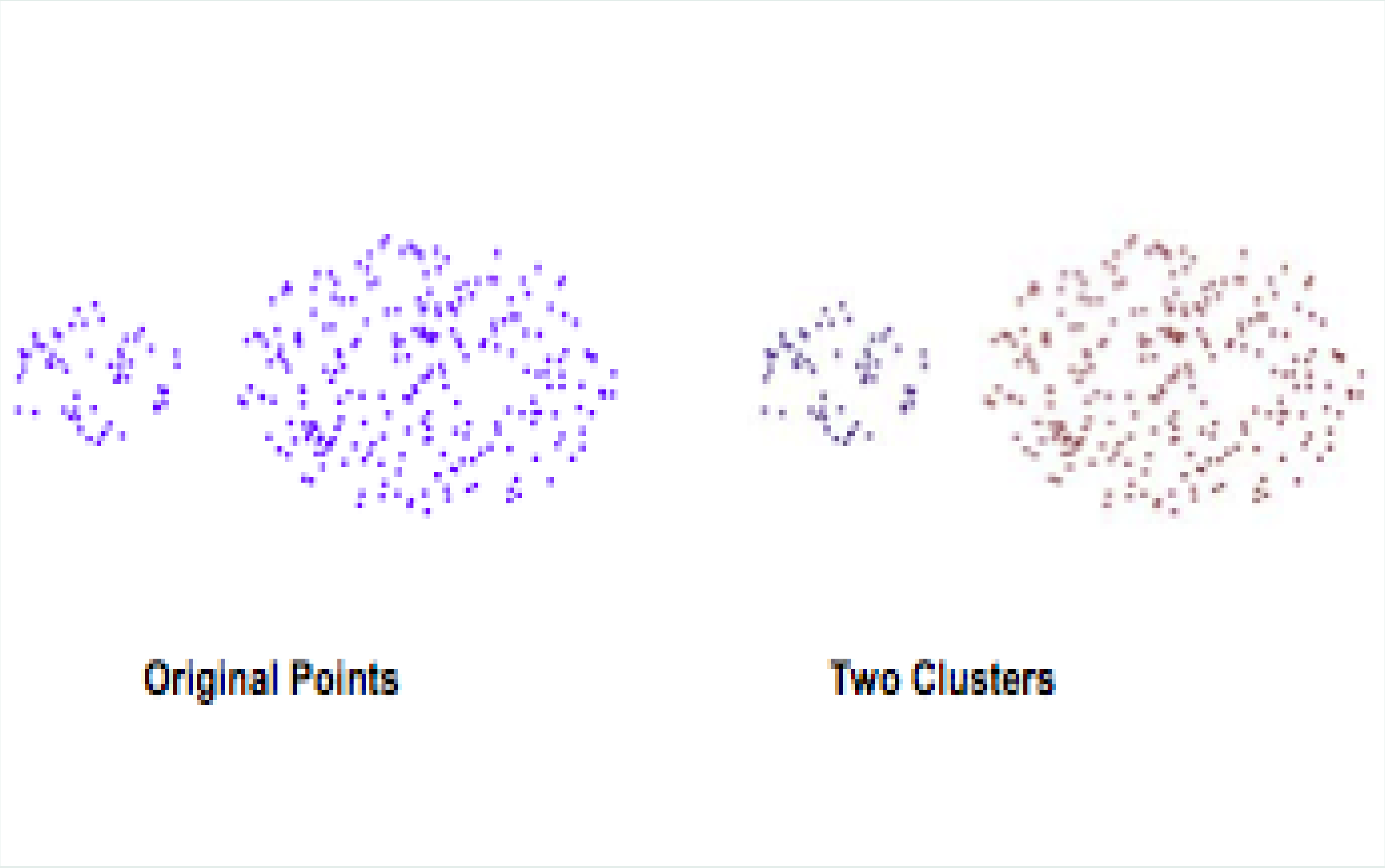
- Agglomerative (bottom-up approach):
  - Start with N clusters (number of objects in the data)
  - At each subsequent step, merge the closest pair of clusters until all objects form a single cluster
- Divisive (top-down approach):
  - Start with single cluster (all objects in the data form 1 cluster)
  - At each subsequent step, split the most 'heterogeneous' cluster until all objects form their own cluster

Input is in the form of a (dis)similarity $N \times N$ matrix

Defining inter-cluster similarities

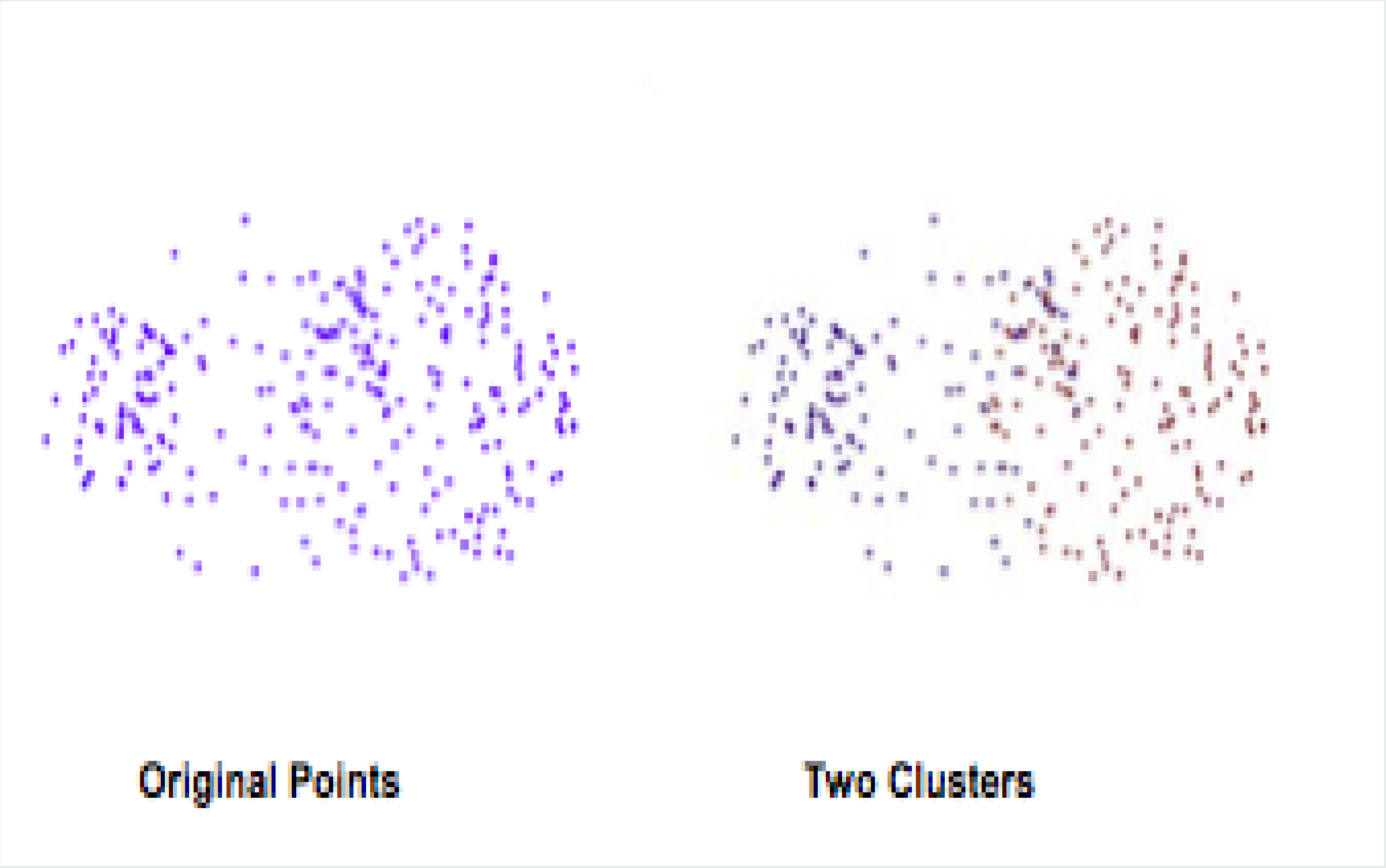
- Single linkage (min)
- Complete linkage (max)
- Average linkage
- Distance between centroids
- Ward's method

Strengths of single linkage



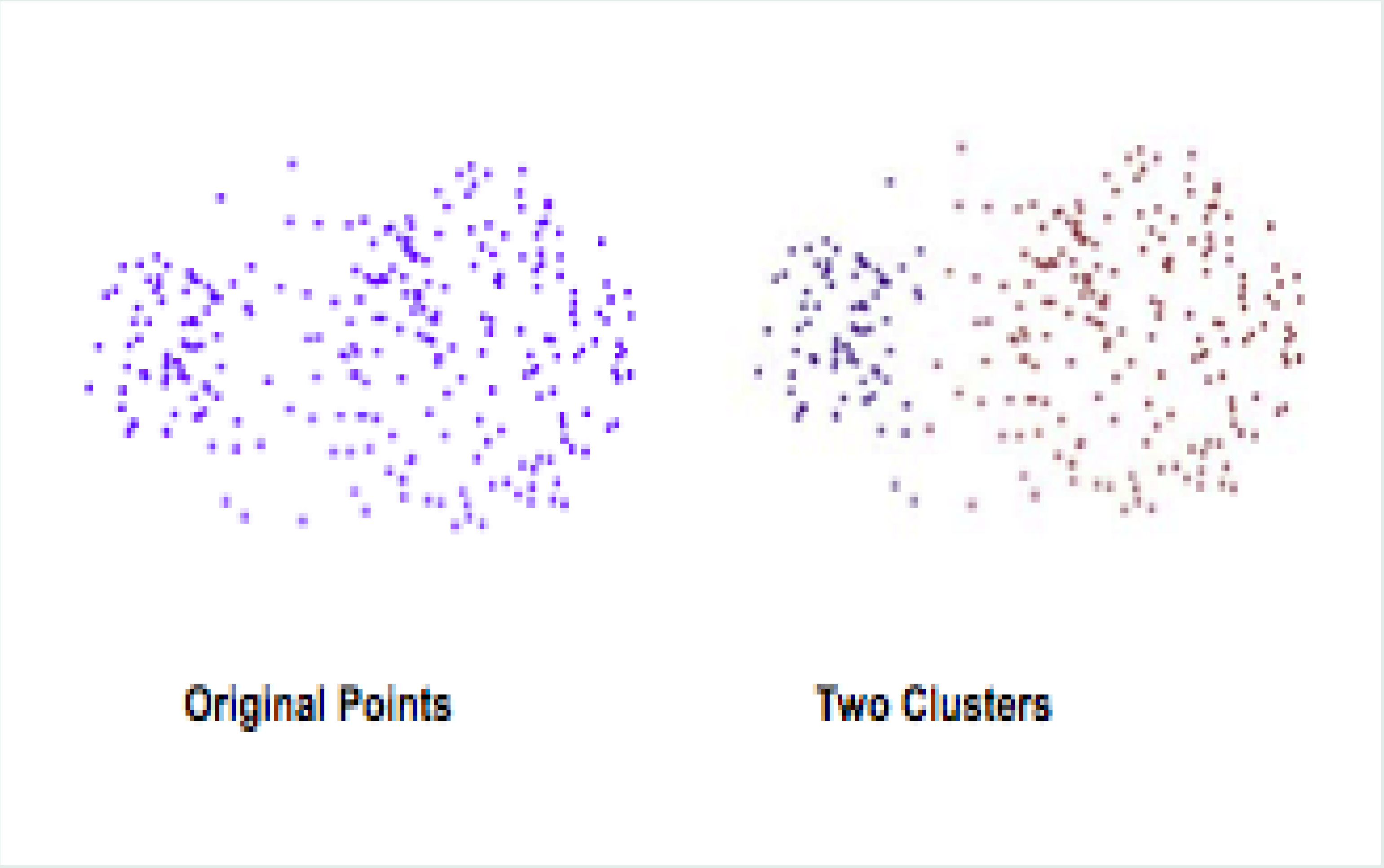
Can handle diverse shapes

Weaknesses of single linkage



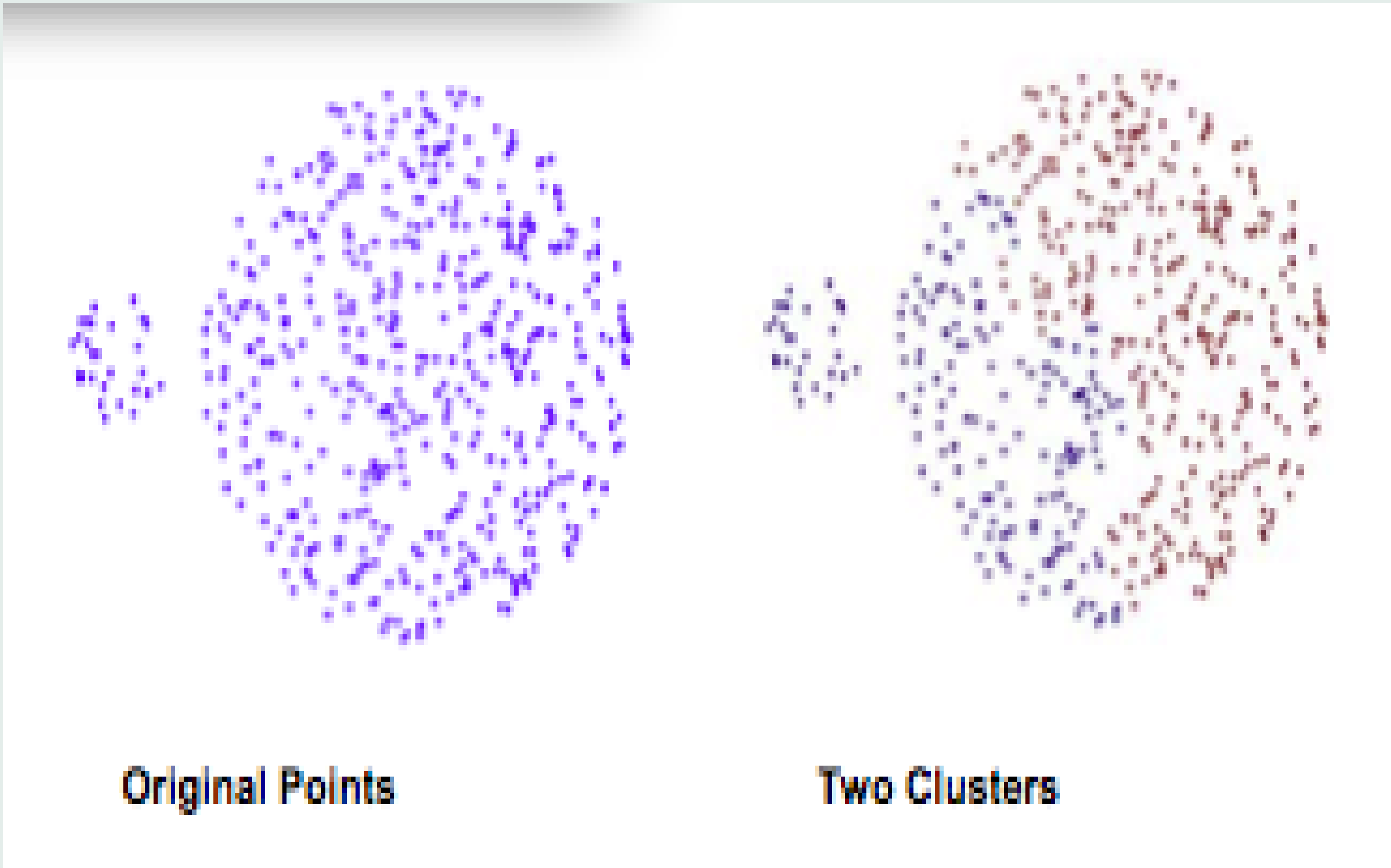
Sensitive to noise and outliers

Strengths of complete linkage



Robust to noise and outliers

Strengths of complete linkage



Tendency towards breaking large clusters; also prefers spherical clusters

Average linkage clustering

- Compromise between single and complete linkage clustering
- Advantages: less susceptible to noise and outliers
- Disadvantages: preference towards spherical clusters

Ward's method for hierarchical clustering

- Similarity of two clusters is based on the increase in squared error when two clusters are merged;
  if the distance between objects is given by squared Euclidean distance, it reduces to average linkage clustering
- Advantages: less susceptible to noise and outliers
- Disadvantages: preference towards spherical clusters

Hierarchical clustering: computational issues

- Space expensive ($\mathcal{O}(N^2)$); requires to compute and store in memory an $N \times N$ (dis)similarity matrix
- Time intensive (best case scenario ($\mathcal{O}(N^2 \log(N))$); requires searching an $N \times N$ (dis)similarity matrix