VICTORIA UNIVERSITY OF WELLINGTON
TE HERENGA WAKA

AIML427

Clustering

Dr. Bach Nguyen
Bach.Nguyen@vuw.ac.nz

Online Helpdesk: Friday, 11-11:50 am

In-person Helpdesk: Friday, 2-3 pm

Outline

- Supervised learning and Unsupervised learning
- Clustering analysis

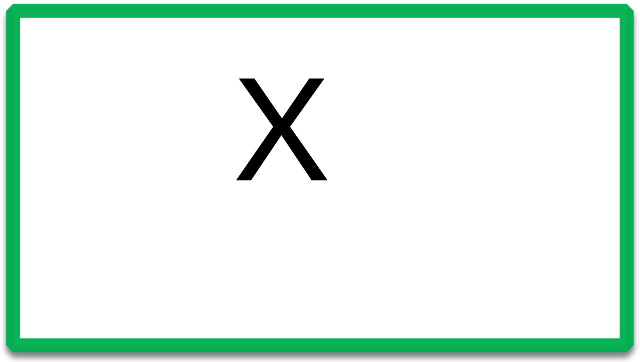
- Clustering Performance
- Clustering Metrics

To understand how to use and interpret:

- K-means clustering
- Hierarchical clustering
- Convex clustering

Unsupervised learning

- In unsupervised learning, we have features $x_1, \dots x_p$ for n observations but there is no associated response y.
- The goal is to find interesting things in the data matrix X itself



n x p

What information can be discovered in X?

Unsupervised learning

- More challenging than supervised learning:
  - no response means no obvious goal for analysis
  - no way to check answers
- More subjective:
  - Part of exploratory data analysis
  - Techniques need to work in high dimensions

Two popular types of unsupervised learning that are a standard starting point are

- Principal components analysis (PCA) and variants
- Clustering, aka cluster analysis

See also ISLR (An Introduction to Statistical Learning: With Application in R): Section 10.1

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earthquake epicenters should be clustered along continent faults

Clustering example

A vast amount of research being done on genetic component of disease.

Suppose we have n patients with melanoma and measurements of the expression levels of p genes. One might like to know:

- Are there clusters within the patients (**observations**)? This might indicate variants of melanoma and suggest different prognoses or treatments
- Are there clusters within the genes (**features**)? Do certain genes work together? Is this the same in individuals without melanoma?

- ***Clustering:***
  - *We will focus on clustering the observations*, i.e. we think of X as representing n points in p-dimensional space. It will be convenient to let x_i denote the ith row of X
  - If we want to cluster features, we just have to take the transpose of X first See also *ISLR 10.3*
  - *Simultaneously* clustering observations and features is also possible. This is known as **biclustering**

Clustering

- Clustering or cluster analysis refers to techniques to find subgroups or clusters in the data.
- The aim is to partition the observations into clusters so that *observations in a cluster are similar (or related or connected)* but *observations in different clusters are not*.
- To do this, need to specify what it means for observations to be similar or different.

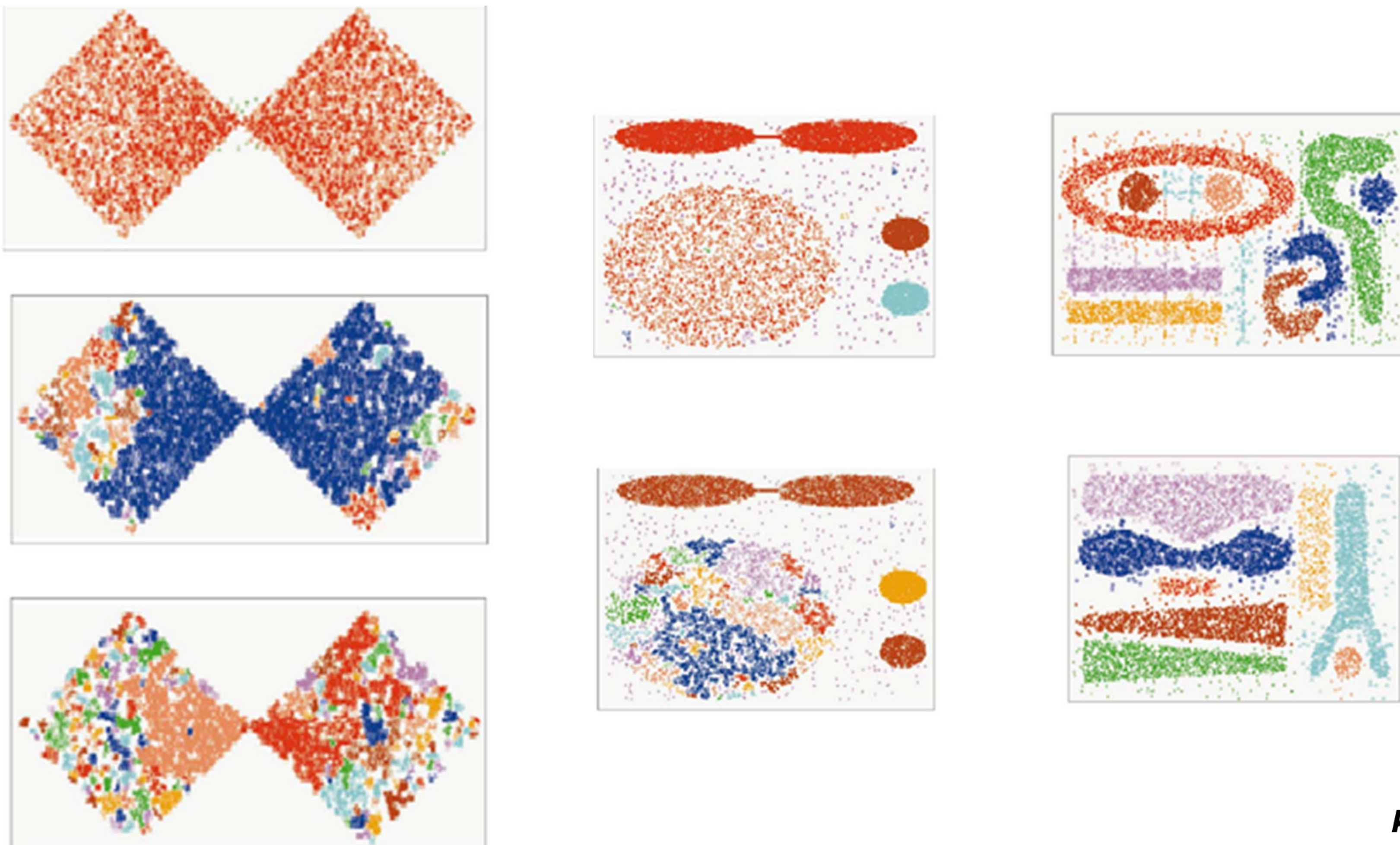
Measure the Quality of Clustering

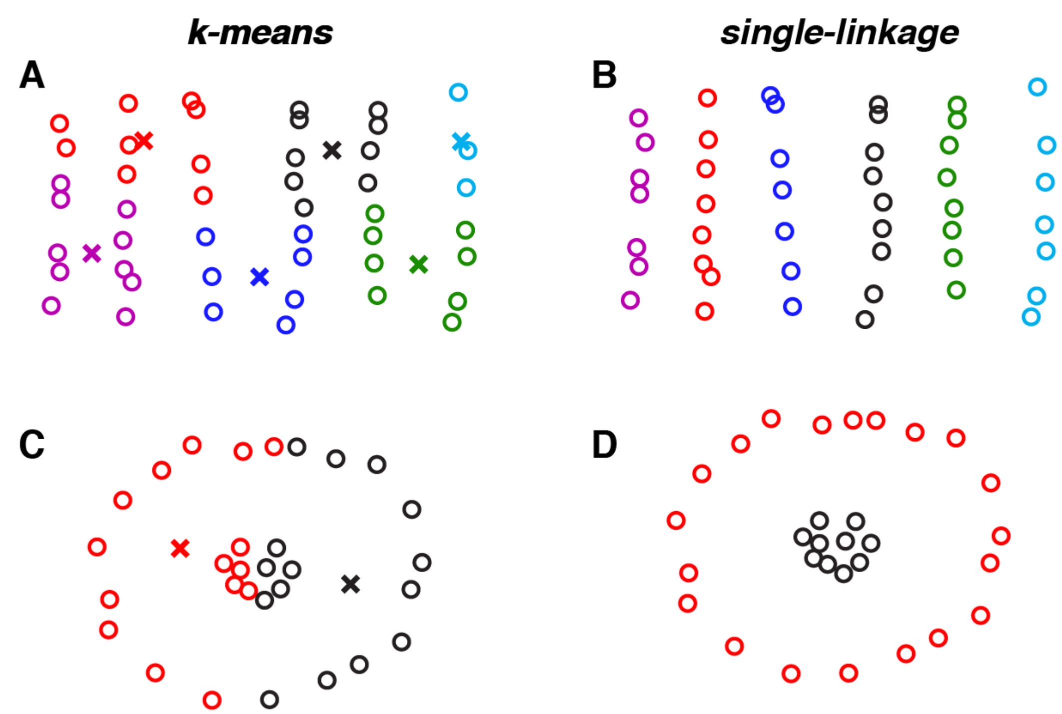
- Dissimilarity/Similarity metric
  - Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$
  - The definitions of distance functions are usually rather different for various types of variables, e.g. real-value, boolean, categorical, ordinal ratio, and vector variables
  - Weights should be associated with different variables based on applications and data semantics
- Quality of clustering:
  - There is usually a separate "quality" function that measures the "goodness" of a cluster.
  - It is hard to define "similar enough" or "good enough"
    - The answer is typically highly **subjective**

Issues in Clustering

- Many applications operate in a very high-dimensional space
  - Almost all pairs of points are at about the same distance!
- For the small number of dimensions and small amount of data, its “easy” but
  - Number of clusters is typically not known
  - Exclusive vs non-exclusive clustering
  - Clusters may be of arbitrary shapes and sizes
  - Quality of clustering result
    - Depends on the similarity measure used and the method and its implementation
    - Measured by its ability to discover some or all of the hidden patterns

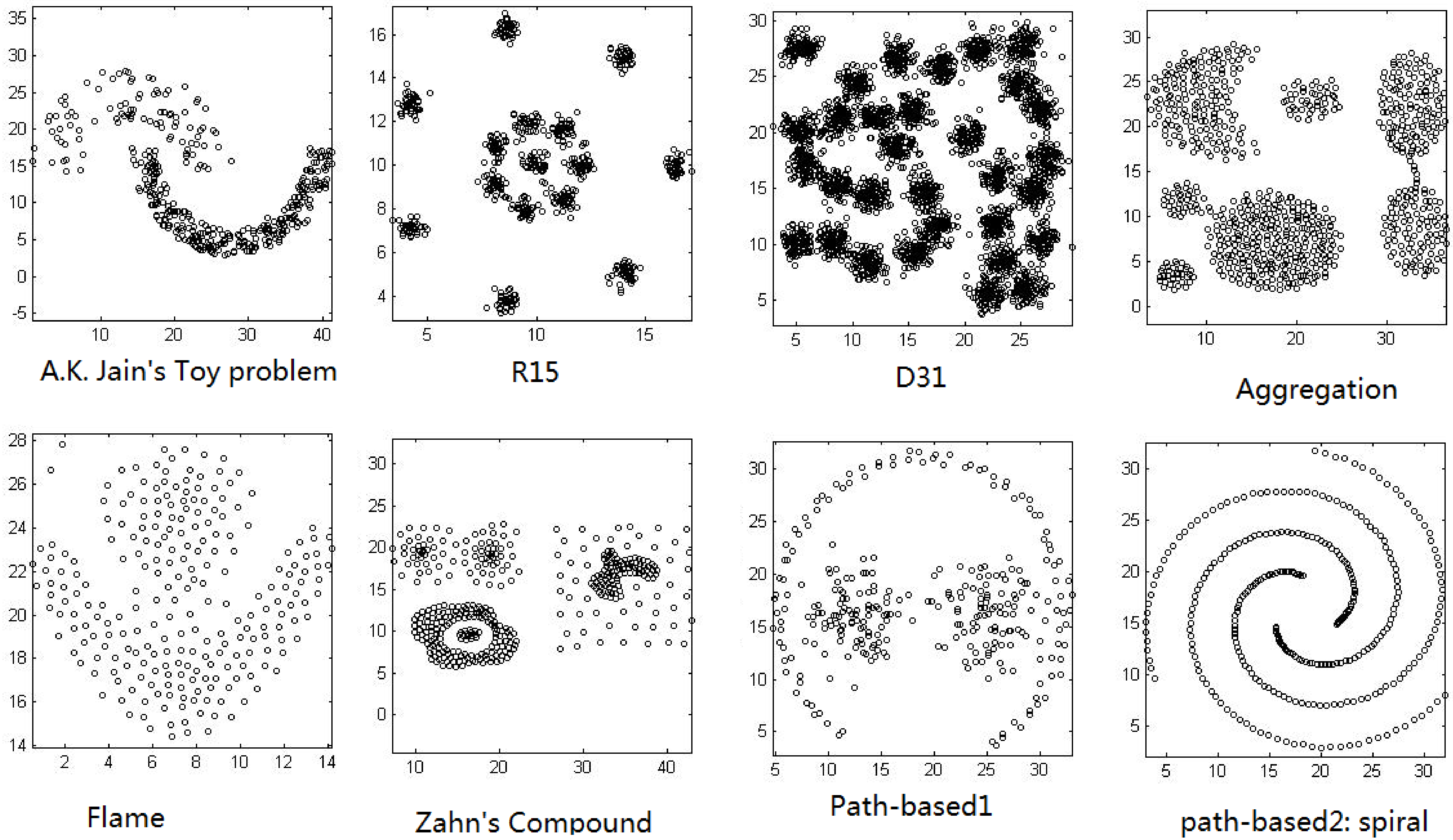
Clustering





How many clusters?

Clustering Datasets

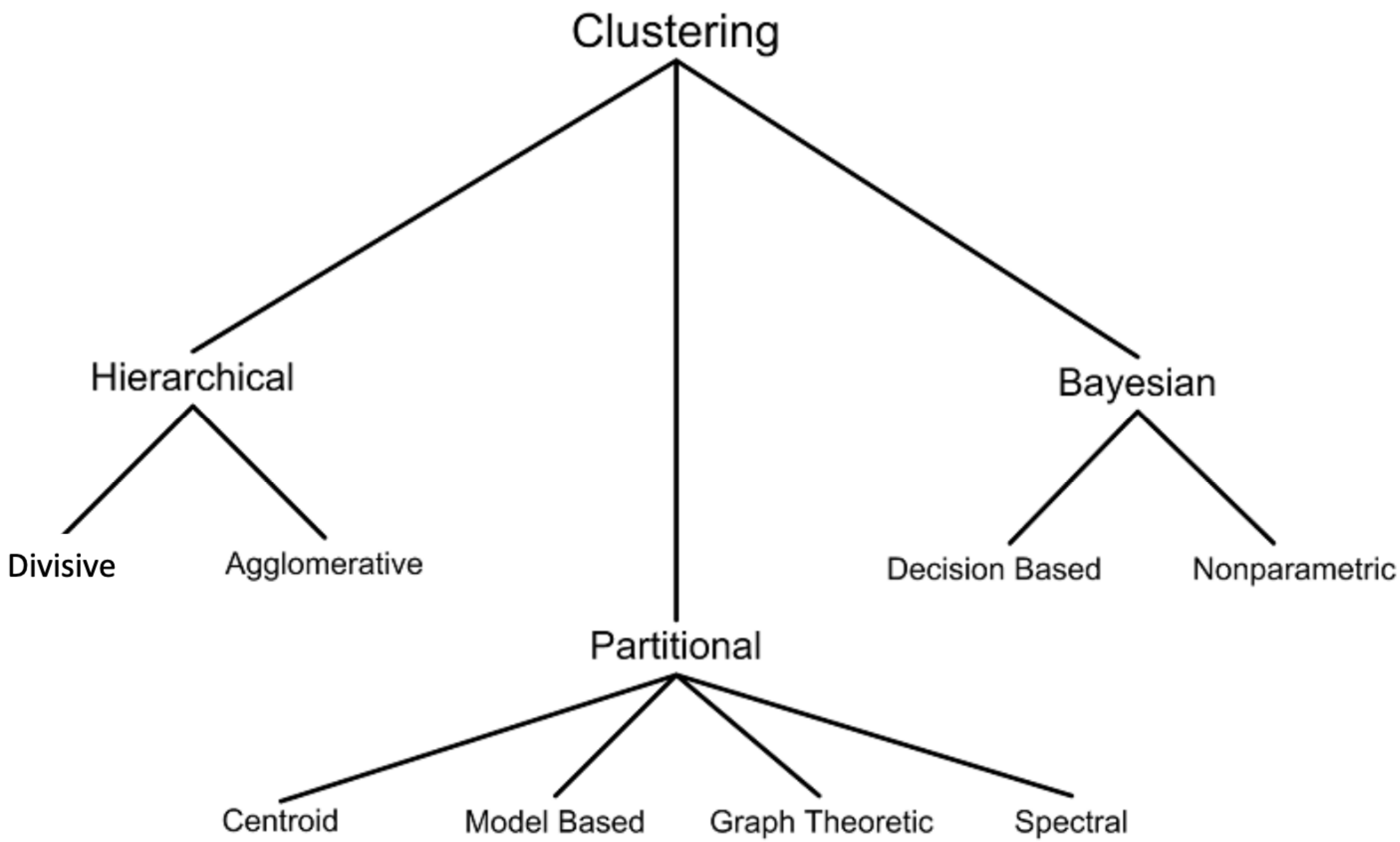


- Hand-crafted datasets exhibiting a range of geometries and densities

Clustering Approaches

- There are many clustering approaches. These include
  - K-means clustering
  - Hierarchical clustering
  - Convex clustering
  - Gaussian mixture models
  - DBSCAN (Density-based spatial clustering of applications with noise) and variants

Clustering Approaches



K-means algorithm

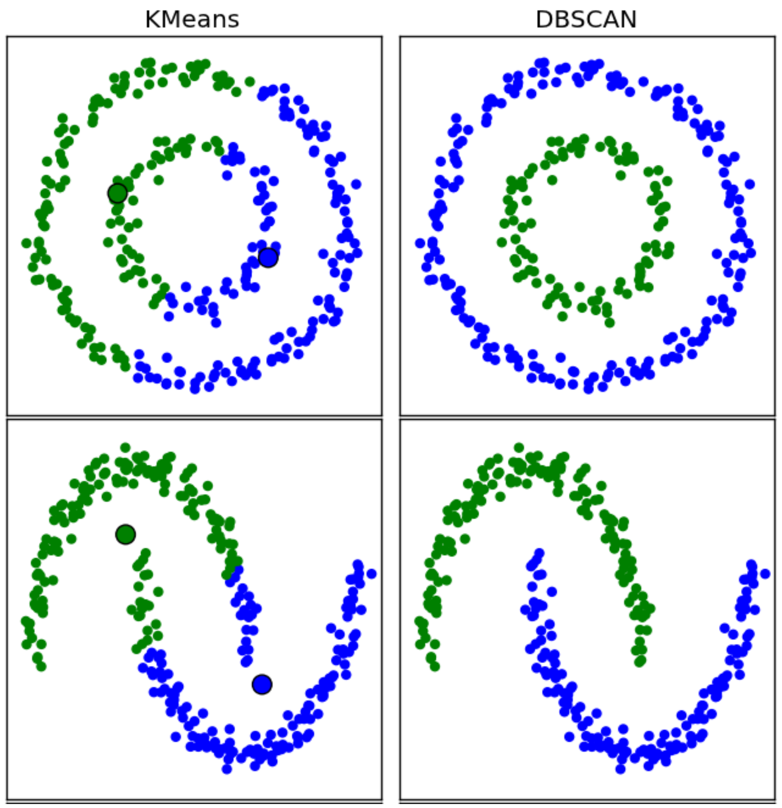
Main steps of K-means:

- Initialise $C_1, \ldots, C_K$ by randomly assigning each observation a number from 1 to K
- Repeat until the the cluster assignments don't change:
  - (a) Compute the *centroid* for each cluster
  - (b) Assign each observation to the cluster whose *centroid* is *closest* in Euclidean distance
- Algorithm 10.1 of ISLR

- The algorithm finds a *local minimum* of the objective function $\sum_{k=1}^{K} W(C_k)$.

Comments on K-means

- Have to predefine K: no guidance on how to choose K
- K-means is based on *spherical clusters*, which might not always be appropriate.
- Sensitive to initial seeds, local minima
- Sensitive to outliers
- Generalising the distance function is possible, e.g. K-medians clustering defines centroids via *component-wise median* and assignment to a cluster is in terms of the *Manhattan* distance (aka taxicab geometry, l_1-norm)
- Care needs to be taken in *high dimensions; irrelevant* features can conceal information about clusters. Idea of distance also breaks down – curse of dimensionality again.
    - Dimension reduction prior to clustering is a good idea



Measuring Clustering Performance

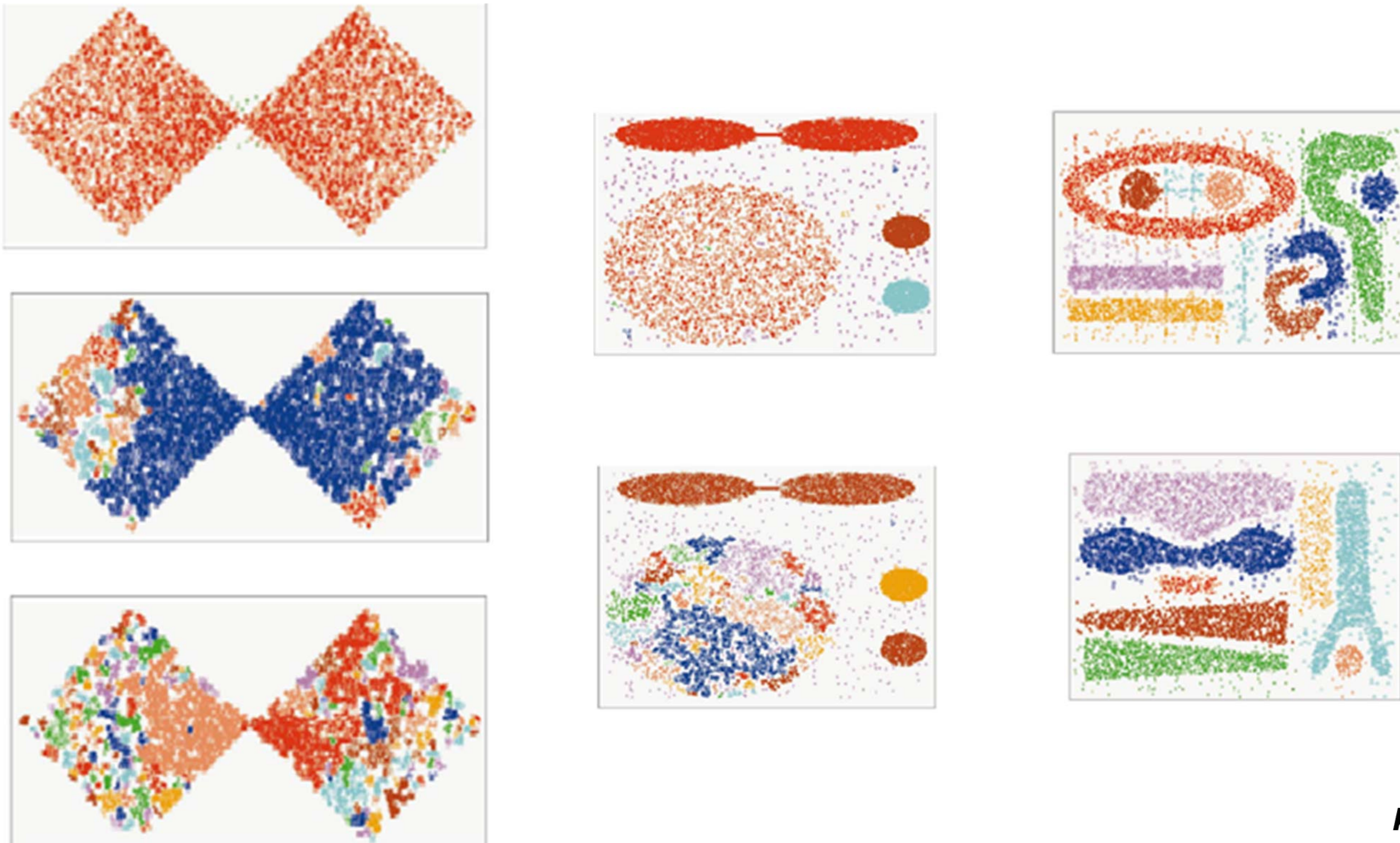
- **Compactness**: how tightly-packed a cluster is.
  - Clusters should be as compact as possible, so as to ensure that only the most related/similar instances have been grouped together.

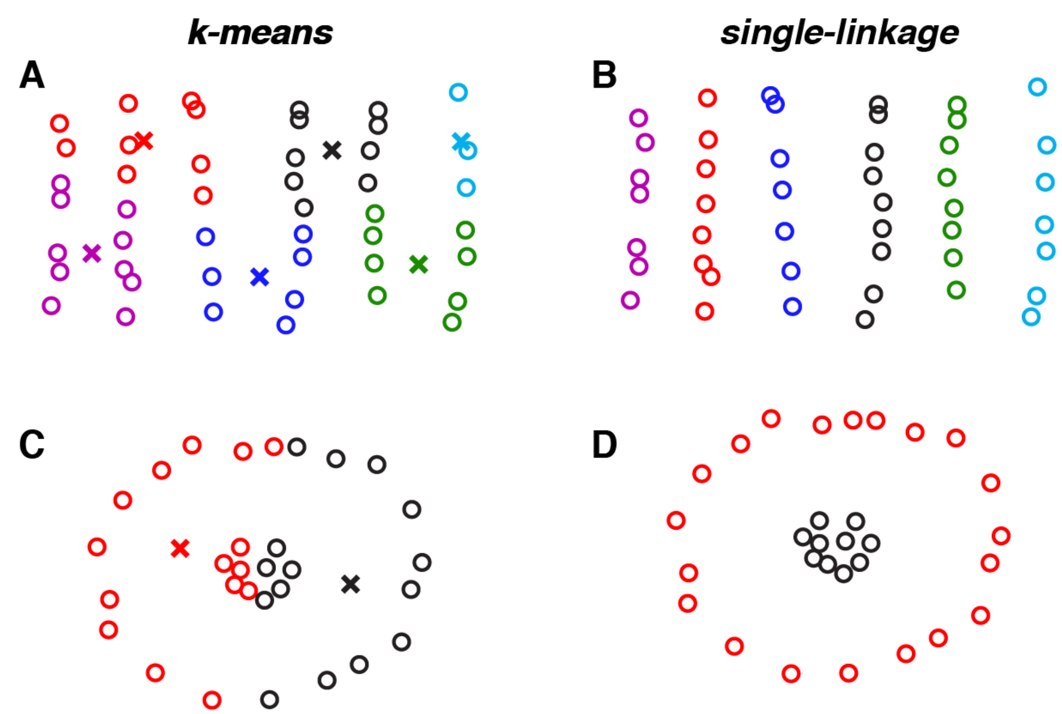
- **Separability**: how well neighbouring clusters are separated in the feature space.

- **Connectedness**: instances that are close together should generally be allocated to the same cluster as they have similar characteristics.
  - Connectedness is generally measured per-instance rather than per-cluster. The most common approach used is to find the mean distance from each instance to its n-nearest neighbours.

- K-means' the clustering performance?

Clustering

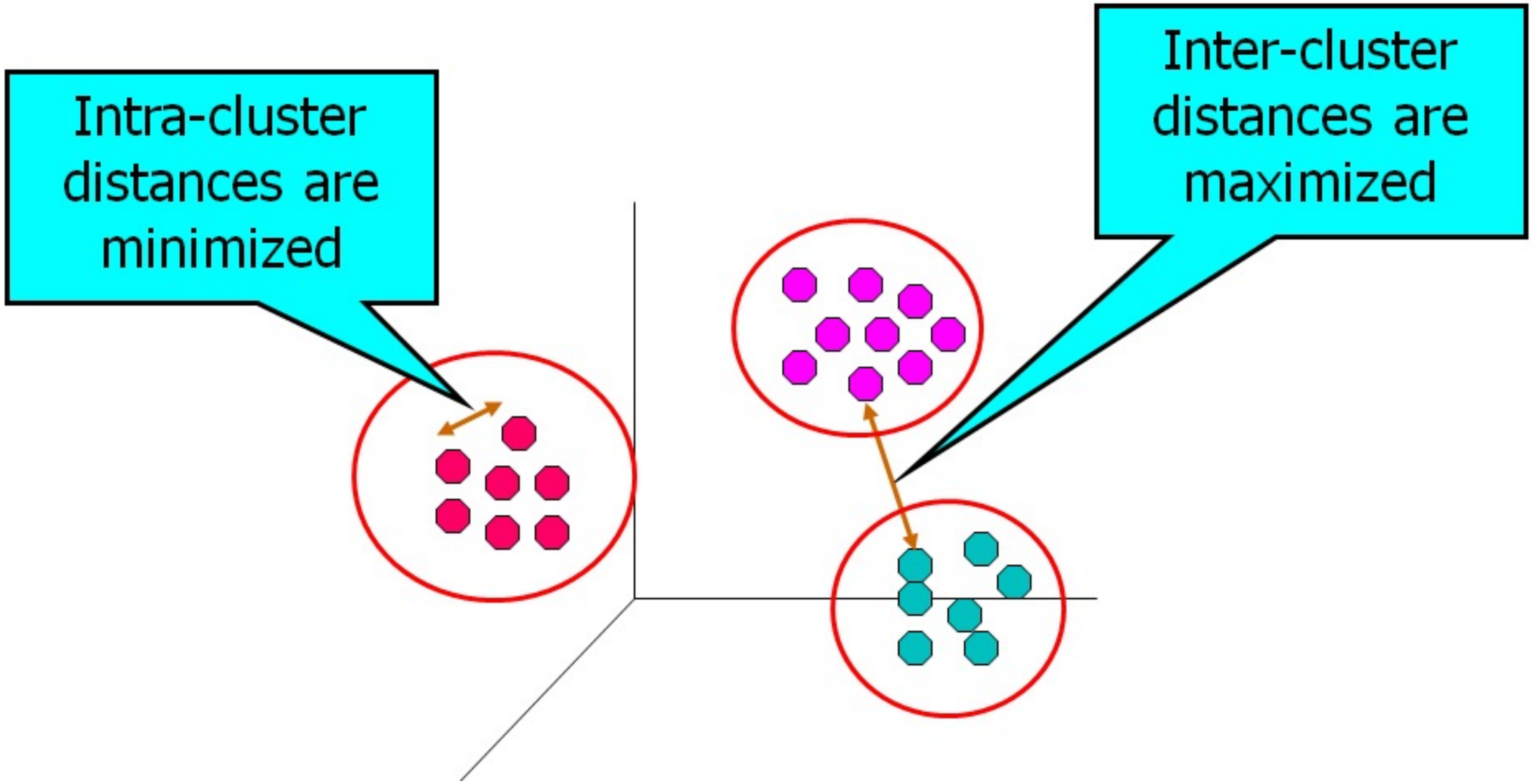




How many clusters?

Clustering

- How to measure/represent Intra-cluster and inter-cluster distances?



Clustering Metrics

- Sum intra-cluster distance:

$$\text{Intra}_{\text{Sum}} \downarrow = \sum_{i=1}^{K} \sum_{a \in C_i} d(a, Z_i)$$

  - z_i represent the mean of the *i*th cluster

- Root Mean Squared Error:

$$\text{RMSE} \downarrow = \sqrt{\frac{1}{K} \sum_{i=1}^{K} CSE_i^2} \qquad \text{CSE} = \sqrt{\frac{1}{|C_i|} \sum_{a \in C_i} d(a, Z_i)^2}$$

- Sum *inter*-cluster distance ($i \neq j$):

$$\text{Inter}_{\text{Sum}} \uparrow = \sum_{i=1}^{K} \sum_{j=1}^{K} d(Z_i, Z_j)$$

$$\text{Inter}_{\text{minDistSum}} \uparrow = \sum_{i=1}^{K} \sum_{j=1}^{K} \min_{a \in C_i, b \in C_j} dist(a, b)$$

Clustering Metrics

- Davis-Bouldin index:

$$\text{Davies-Bouldin} \downarrow = \frac{1}{K} \max_{1 \le i < j \le K} \frac{S_{C_i} + S_{C_j}}{dist(Z_i, Z_j)}$$

$$S_{C_i} = \frac{1}{|C_i|} \sum_{a \in C_i} d(a, Z_i)$$

  - The Davies-Bouldin index measures the ratio of intra-cluster distance (i.e. within-cluster scatter) to inter-cluster separability.
  - The two clusters which have the highest ratio give the output of the Davies-Bouldin index.

- overly pessimistic or optimistic?

Clustering Metrics

- Dunn index:

$$\text{Dunn Index} \uparrow = \frac{\min_{1 \leq i < j \leq K} dist(Z_i, Z_j)}{\max_{1 \leq i \leq K} \max_{a,b \in C_i} dist(a, b)}$$

- The numerator finds the minimum distance between any two clusters.
- The denominator finds the maximum distance between any two instances which are in the same cluster.

- Similar to the Davies-Bouldin index in that it considers the inter-cluster distance of the two closest clusters.

Clustering Metrics

- Silhouette

$$Silhouette(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

- $a(i)$ is the *average* distance between instance i and all other instances in its cluster;
- $b(i)$ is the *minimum* average distance between instance i and the instances in each other cluster.

- Measures how well a given instance is matched to its cluster
  - The average silhouette computed across all instances in a partition gives a measure of how good the partition is,
  - implicitly balances both the intra- and inter-cluster metrics.

- 1 indicates an instance is perfectly clustered
- -1 indicates it should be in a neighbouring cluster;
- 0 indicates it is on the border of two clusters