# Clustering

CMT209
Informatics

Cardiff School of **Computer Science & Informatics**

http://www.cs.cf.ac.uk

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Lecture

- in the previous lecture we learnt how to compare objects using distance and similarity measures
- in this lecture, we will learn about clustering, which is used to automatically group similar objects together

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

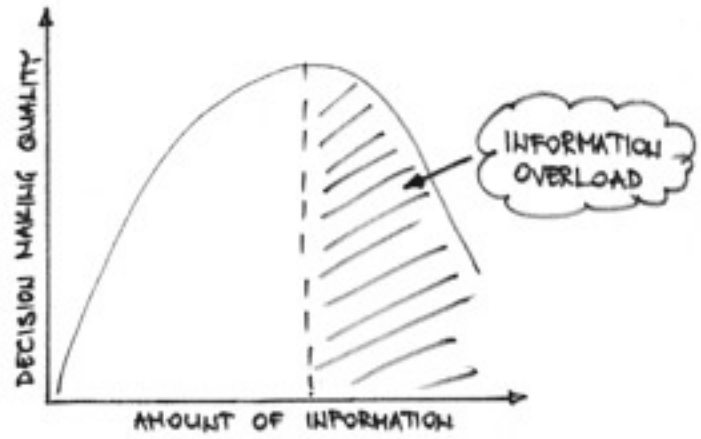
# Information overload


- a difficulty a person may have when trying to deal with more information than they are able to process to make sensible decisions
- results in delaying making decisions or making the wrong decisions
- an increasing problem in the context of big data

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Information overload

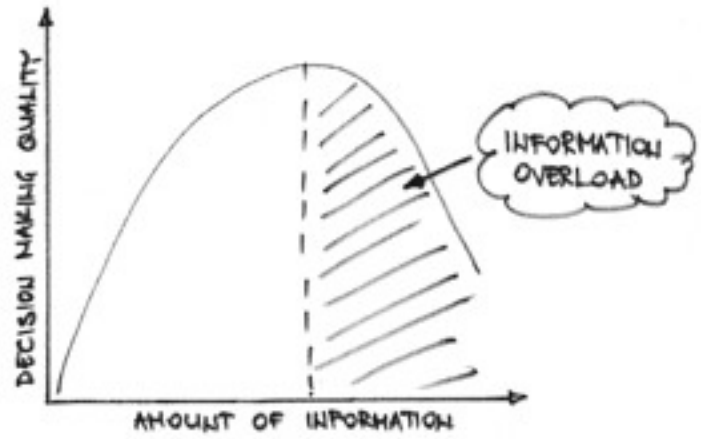

- a difficulty a person may have when trying to deal with more information than they are able to process to make sensible decisions
- results in delaying making decisions or making the wrong decisions
- an increasing problem in the context of big data

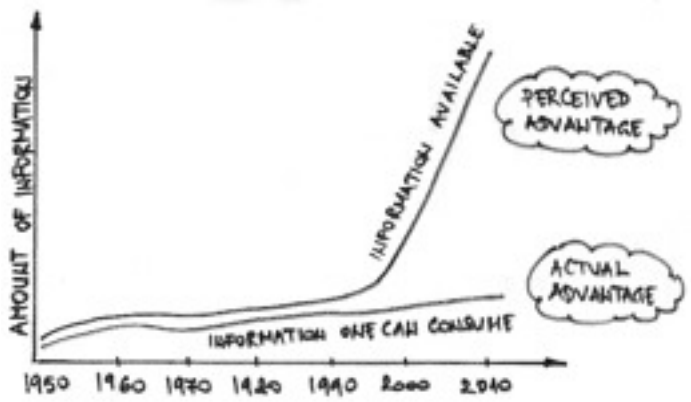



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Information overload


- a difficulty a person may have when trying to deal with more information than they are able to process to make sensible decisions
- results in delaying making decisions or making the wrong decisions
- an increasing problem in the context of big data

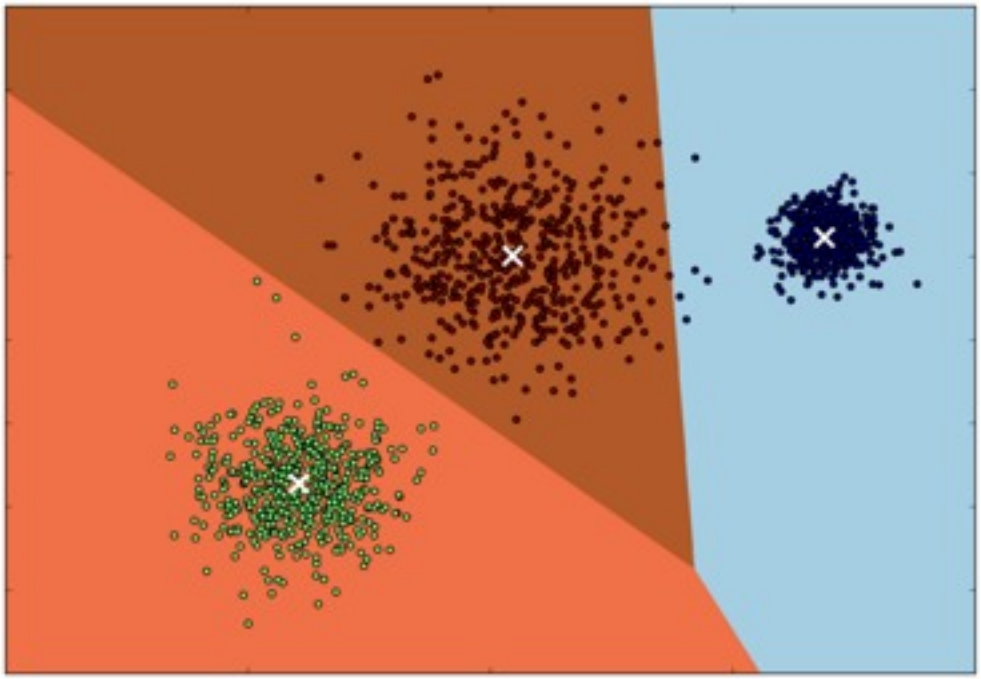



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Clustering

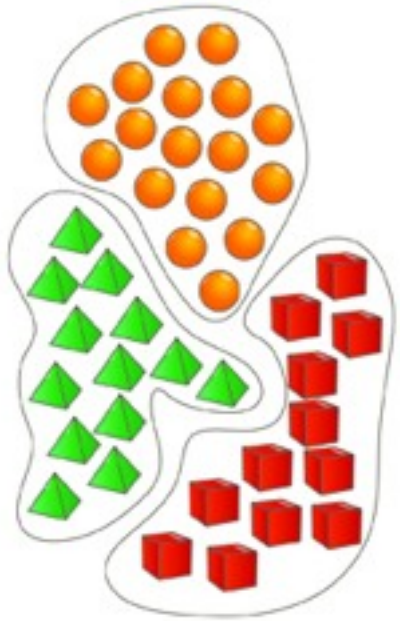
- solution: divide & conquer
- cluster analysis divides data into groups (clusters) that are meaningful, useful or both



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Clustering for understanding

- classes are conceptually meaningful groups of objects that share common characteristics
- classes play an important role in how we analyse and describe the world
- in the context of understanding data, clusters are potential classes



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

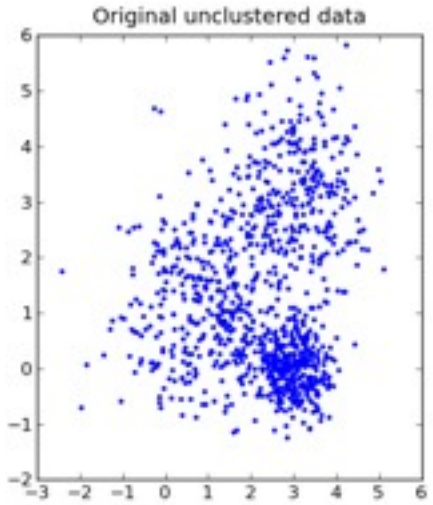
# Clustering for utility

- some clustering techniques characterise each cluster in terms of a cluster prototype
- cluster prototype is a data object that is representative of other objects in the cluster
- cluster prototypes can be used as the basis for a number of data analysis or data processing techniques

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Clustering for utility

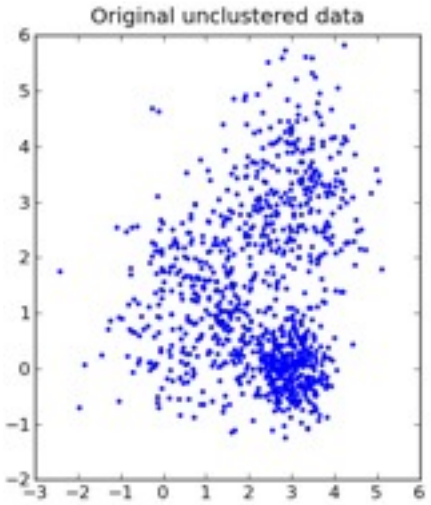
- some clustering techniques characterise each cluster in terms of a cluster prototype
- cluster prototype is a data object that is representative of other objects in the cluster
- cluster prototypes can be used as the basis for a number of data analysis or data processing techniques

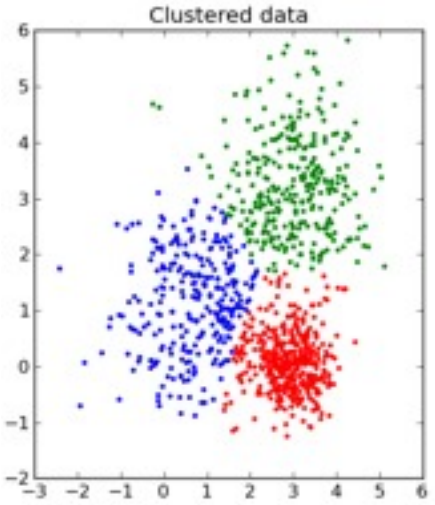



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Clustering for utility

- some clustering techniques characterise each cluster in terms of a cluster prototype
- cluster prototype is a data object that is representative of other objects in the cluster
- cluster prototypes can be used as the basis for a number of data analysis or data processing techniques

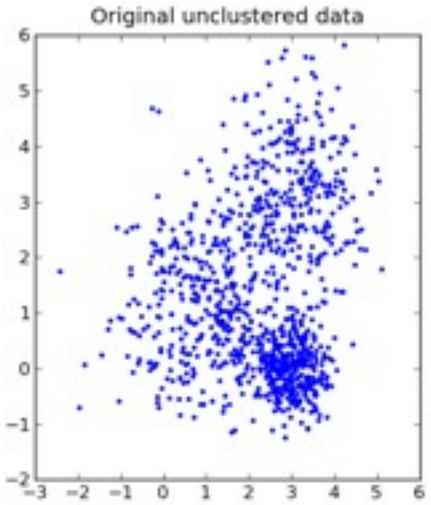


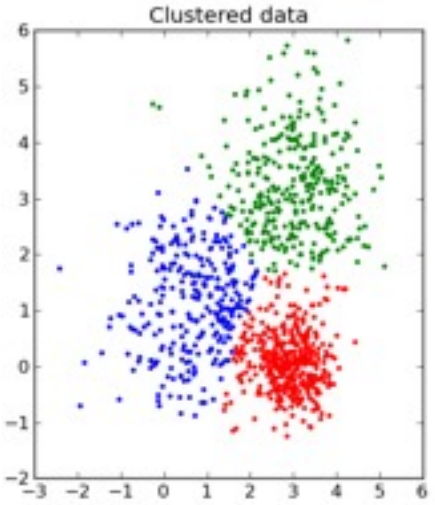




CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Clustering for utility

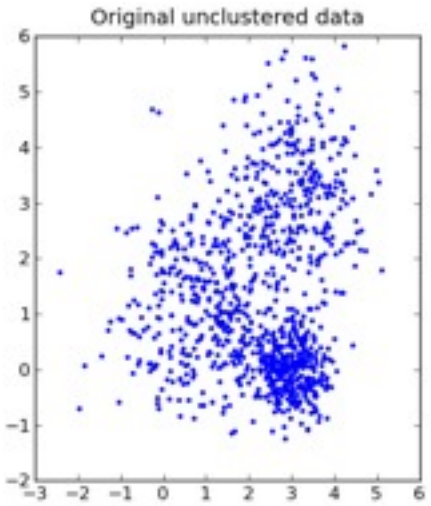
- some clustering techniques characterise each cluster in terms of a cluster prototype
- cluster prototype is a data object that is representative of other objects in the cluster
- cluster prototypes can be used as the basis for a number of data analysis or data processing techniques

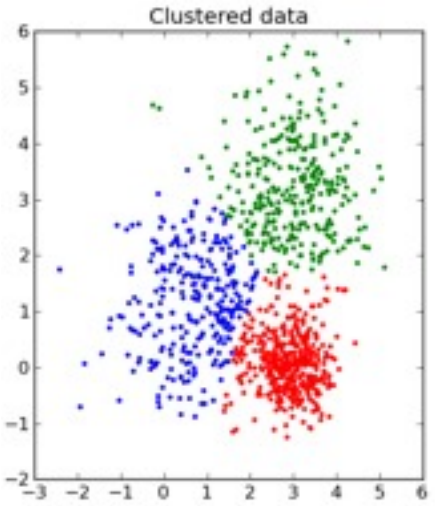



# Clustering for utility

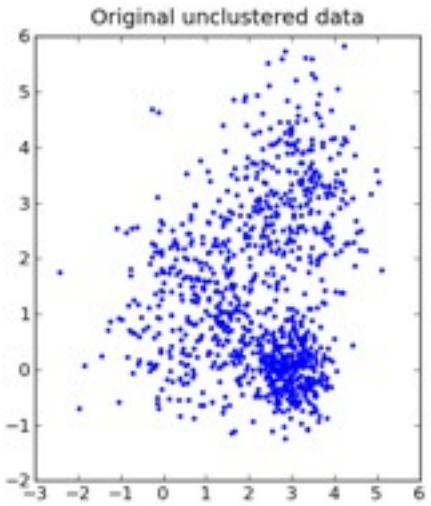
- some clustering techniques characterise each cluster in terms of a cluster prototype
- cluster prototype is a data object that is representative of other objects in the cluster
- cluster prototypes can be used as the basis for a number of data analysis or data processing techniques

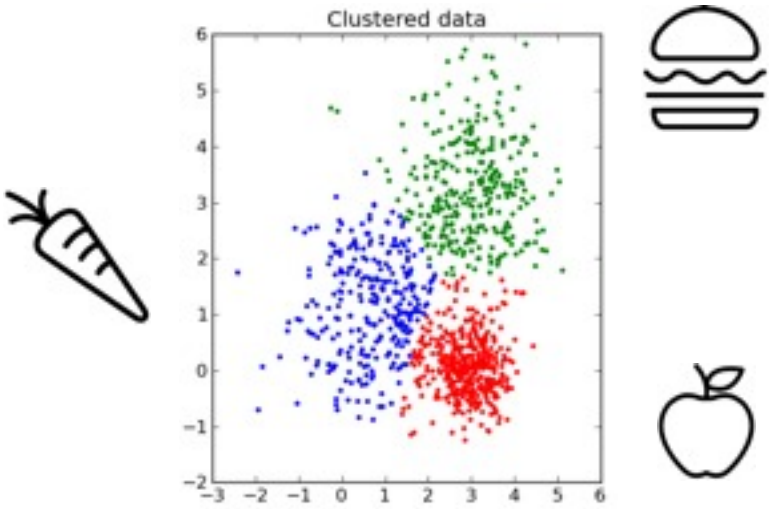



# Clustering for utility

- some clustering techniques characterise each cluster in terms of a cluster prototype
- cluster prototype is a data object that is representative of other objects in the cluster
- cluster prototypes can be used as the basis for a number of data analysis or data processing techniques



# What is clustering?

# Clustering

- **clustering** – a data mining technique that automatically groups objects with similar properties into clusters
- objects within a cluster should be similar (or related) to one another and different from (or unrelated to) the objects in other clusters
- the greater the similarity within a cluster and the greater the difference between clusters, the better the clustering
- unsupervised learning – no annotations/ training required!

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# What are natural clusters?

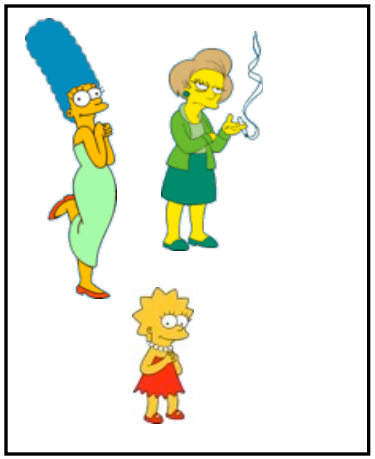
# What are natural clusters?

Clustering is subjective!

# What are natural clusters?

Clustering is subjective!

men



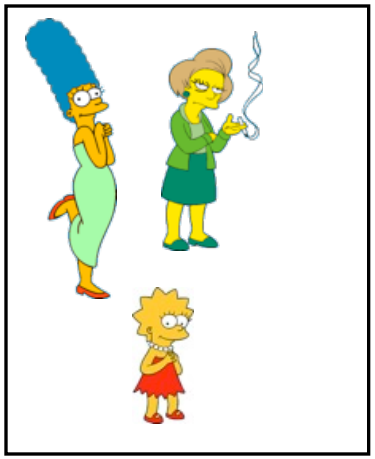
women

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

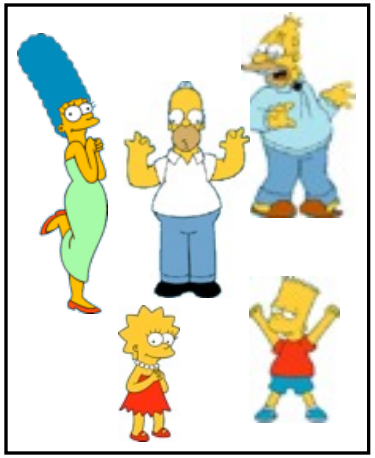
# What are natural clusters?

Clustering is subjective!

men



women



Simpson family

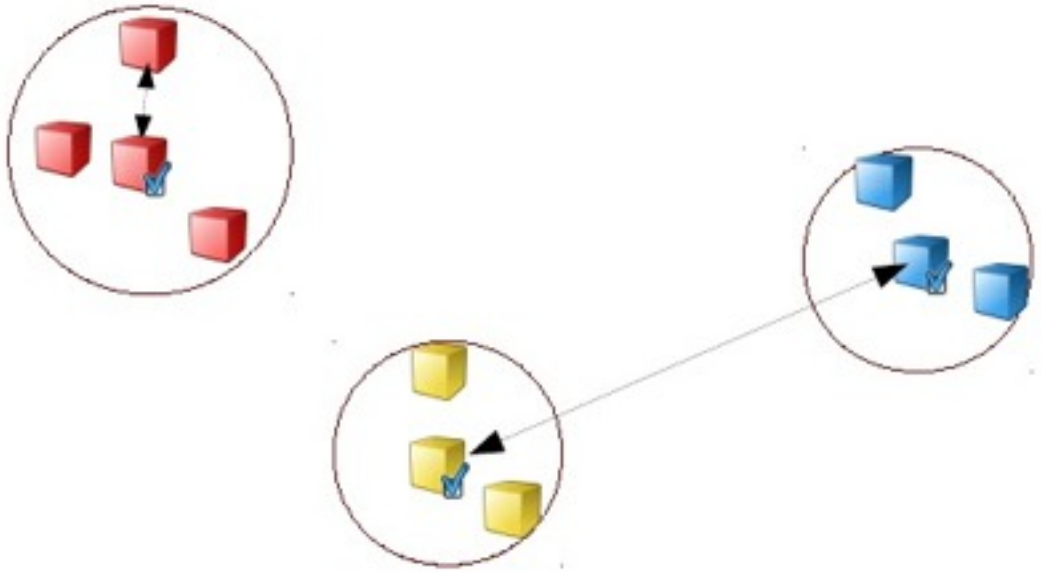
school employees

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Clustering

- the goal is to:

1. minimise intra-cluster distances
2. maximise inter-cluster distances

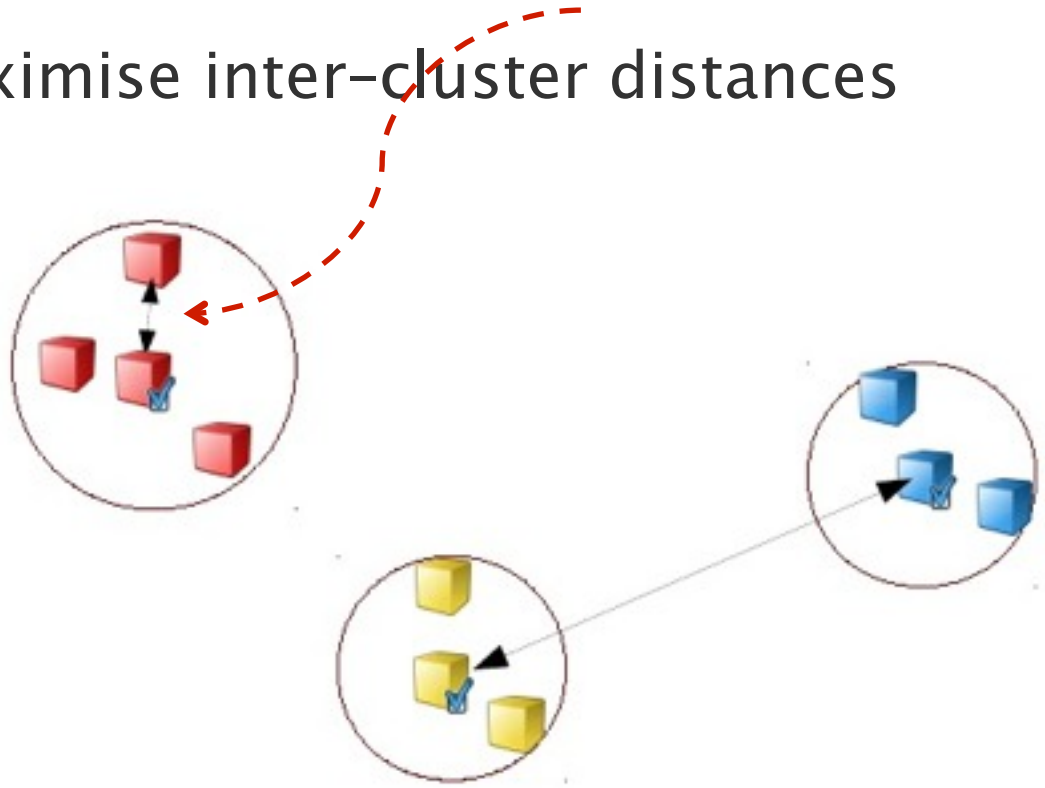


CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Clustering

- the goal is to:

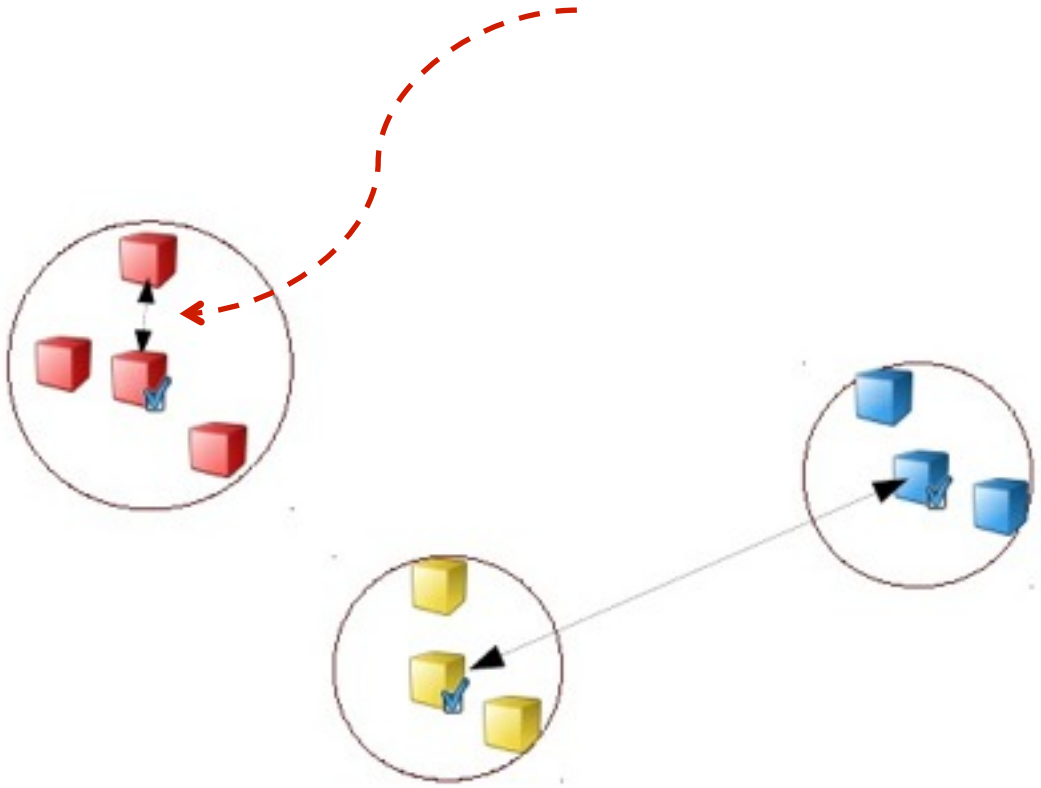
1. minimise intra-cluster distances
2. maximise inter-cluster distances



# Clustering

- the goal is to:

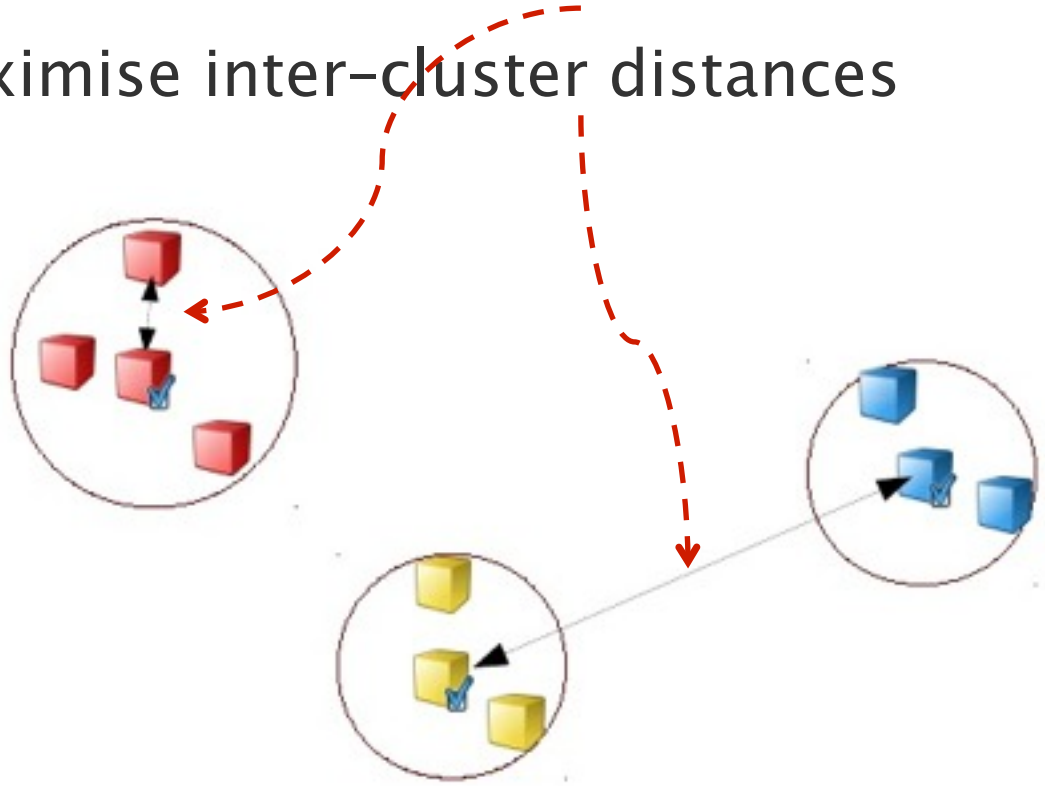
1. minimise intra-cluster distances



# Clustering

- the goal is to:

1. minimise intra-cluster distances
2. maximise inter-cluster distances



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

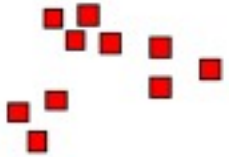
# How many clusters?

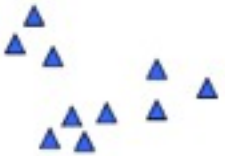
- 2, 4 or 6?
- notion of a cluster is ambiguous
- difficult to decide what constitutes a cluster
- the best definition of a cluster depends on the nature of data and desired outputs

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# How many clusters?

- 2, 4 or 6?
- notion of a cluster is ambiguous
- difficult to decide what constitutes a cluster
- the best definition of a cluster depends on the nature of data and desired outputs

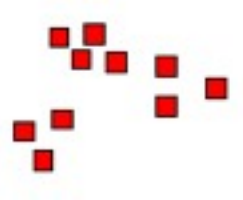


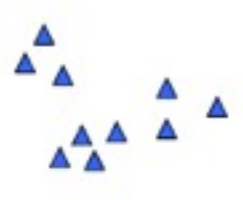


CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# How many clusters?

- 2, 4 or 6?
- notion of a cluster is ambiguous
- difficult to decide what constitutes a cluster
- the best definition of a cluster depends on the nature of data and desired outputs

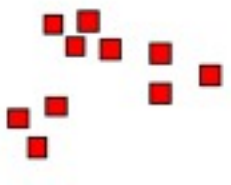




CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# How many clusters?

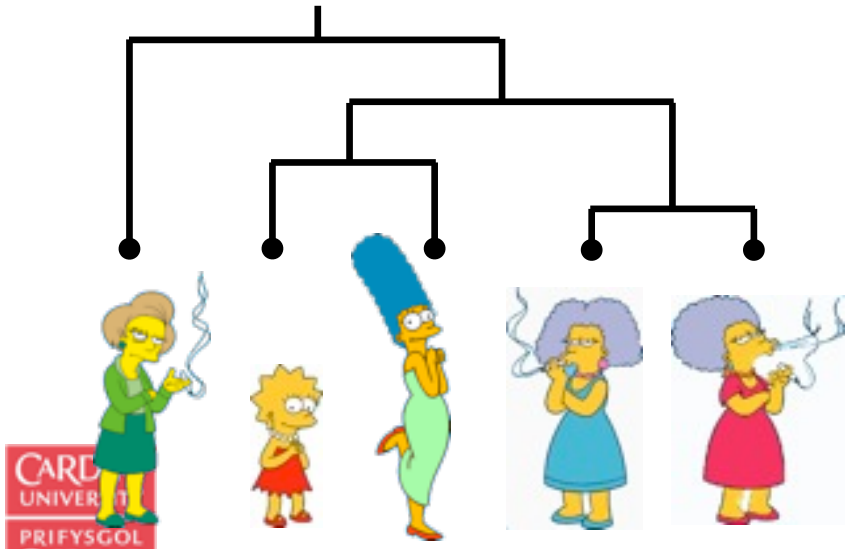
- 2, 4 or 6?
- notion of a cluster is ambiguous
- difficult to decide what constitutes a cluster
- the best definition of a cluster depends on the nature of data and desired outputs

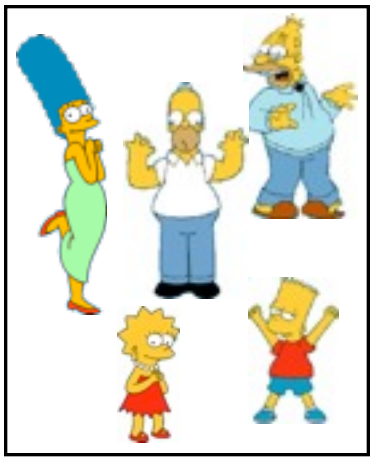


CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Hierarchical vs. partitional clustering

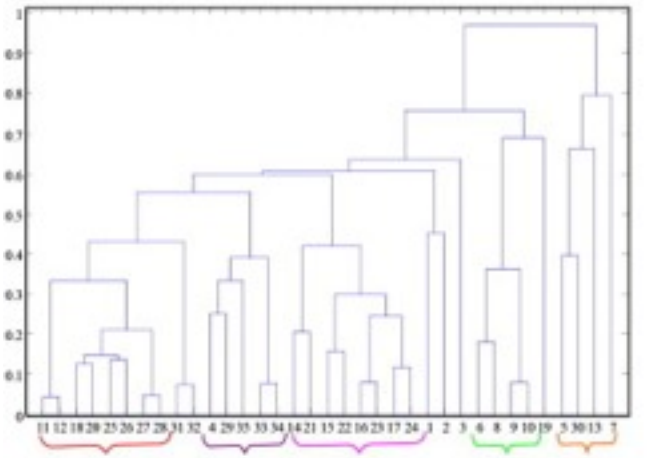
- two basic types of clustering:
  1. hierarchical
  2. non-hierarchical (partitional)

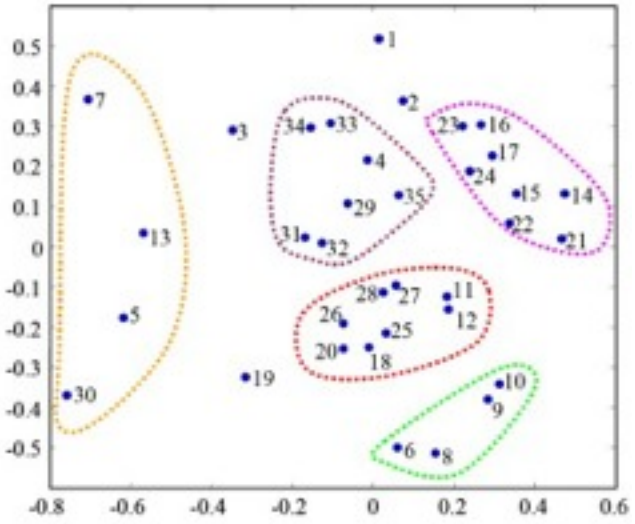




# Hierarchical vs. partitional clustering

- two basic types of clustering:
  1. hierarchical
  2. non-hierarchical (partitional)

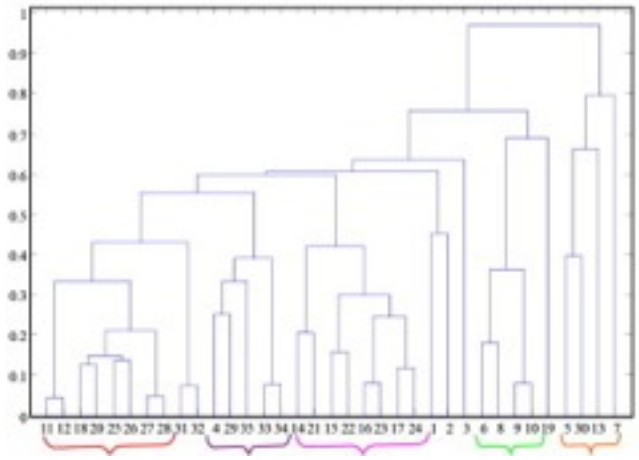





CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Hierarchical clustering

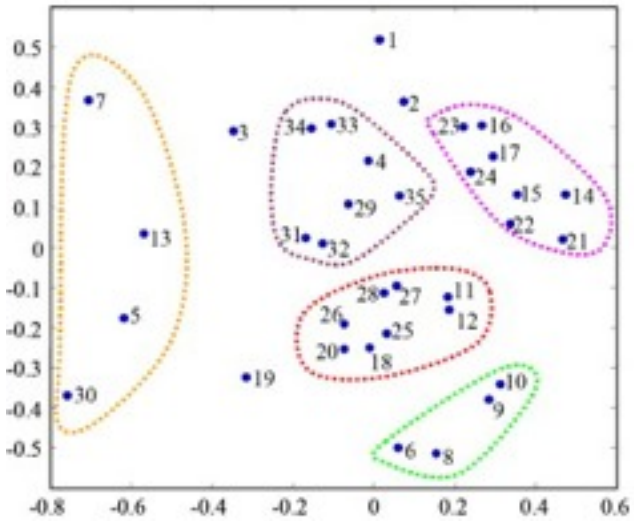
- clusters are permitted to have subclusters
- a set of nested clusters that are organised in a tree
- each node (cluster) in the tree (except the leaves) is the union of its children (subclusters)
- the root of the tree contains all objects

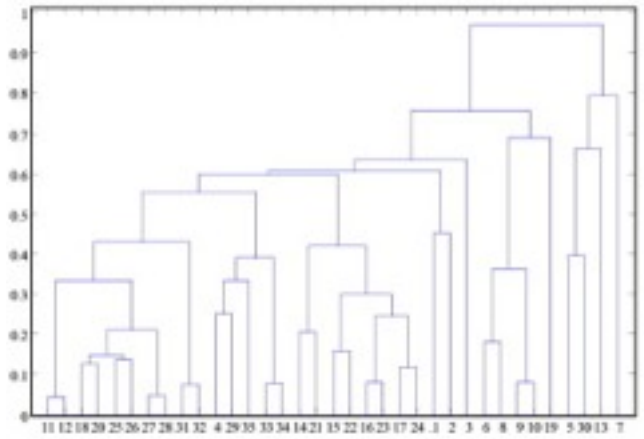


CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Partitional clustering

- partitional clustering is simply a division of the set of objects into non-overlapping subsets (clusters)
- note that hierarchical clustering can be viewed as a sequence of partitional clusterings
- a partitional clustering can be obtained by cutting the tree at a particular level

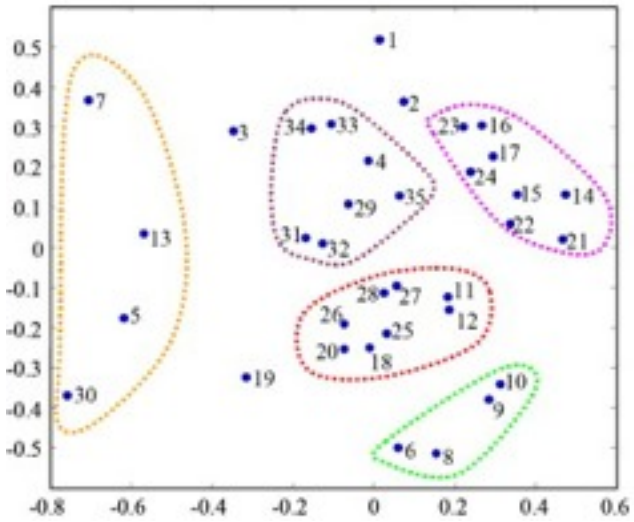


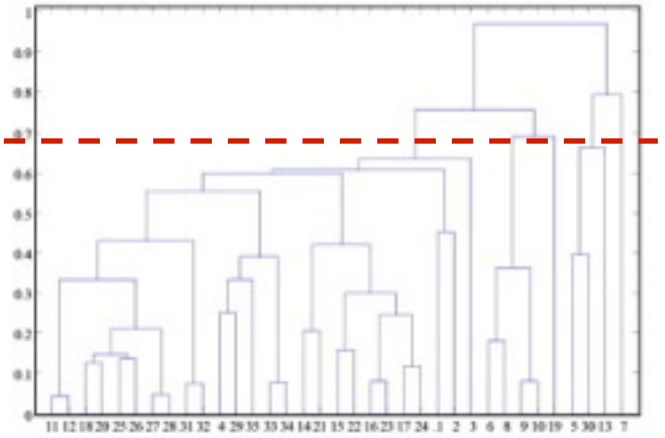



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Partitional clustering

- partitional clustering is simply a division of the set of objects into non-overlapping subsets (clusters)
- note that hierarchical clustering can be viewed as a sequence of partitional clusterings
- a partitional clustering can be obtained by cutting the tree at a particular level

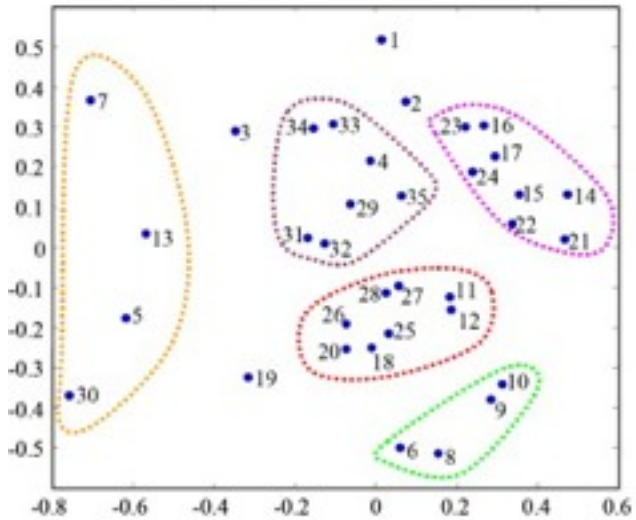


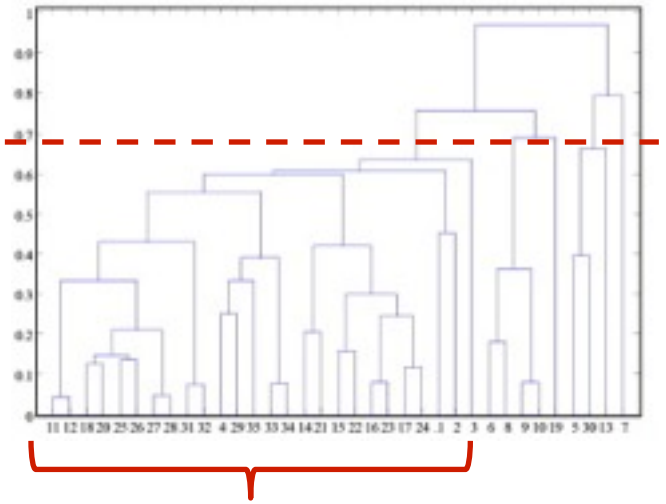



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Partitional clustering

- partitional clustering is simply a division of the set of objects into non-overlapping subsets (clusters)
- note that hierarchical clustering can be viewed as a sequence of partitional clusterings
- a partitional clustering can be obtained by cutting the tree at a particular level

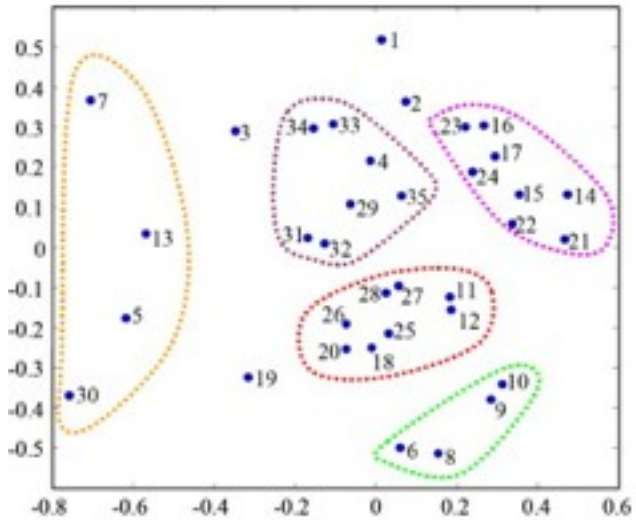


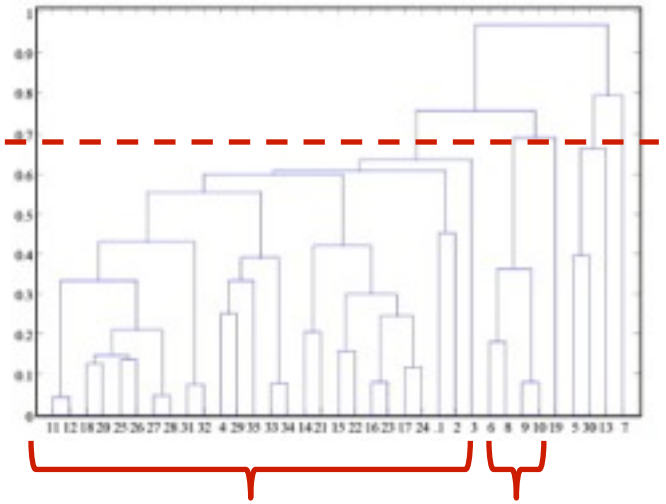


CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Partitional clustering

- partitional clustering is simply a division of the set of objects into non-overlapping subsets (clusters)
- note that hierarchical clustering can be viewed as a sequence of partitional clusterings
- a partitional clustering can be obtained by cutting the tree at a particular level

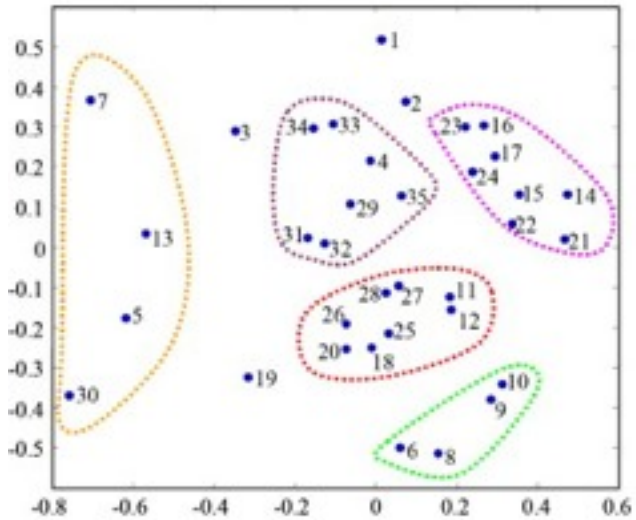


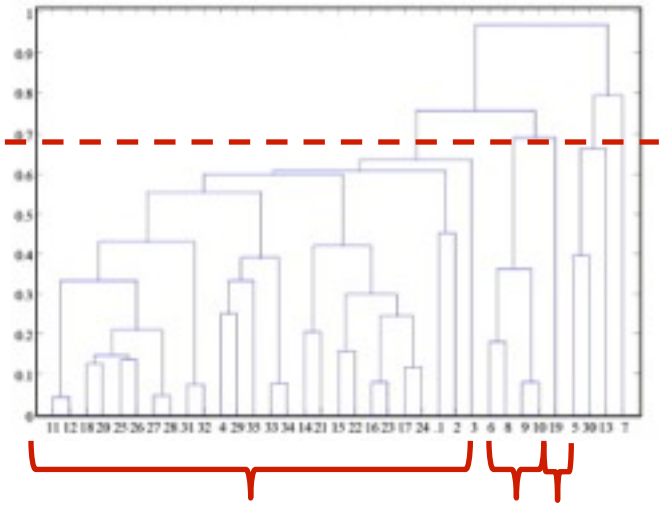



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Partitional clustering

- partitional clustering is simply a division of the set of objects into non-overlapping subsets (clusters)
- note that hierarchical clustering can be viewed as a sequence of partitional clusterings
- a partitional clustering can be obtained by cutting the tree at a particular level

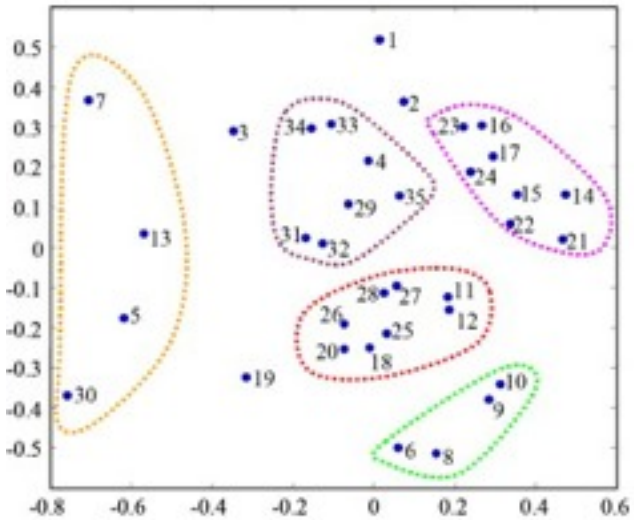


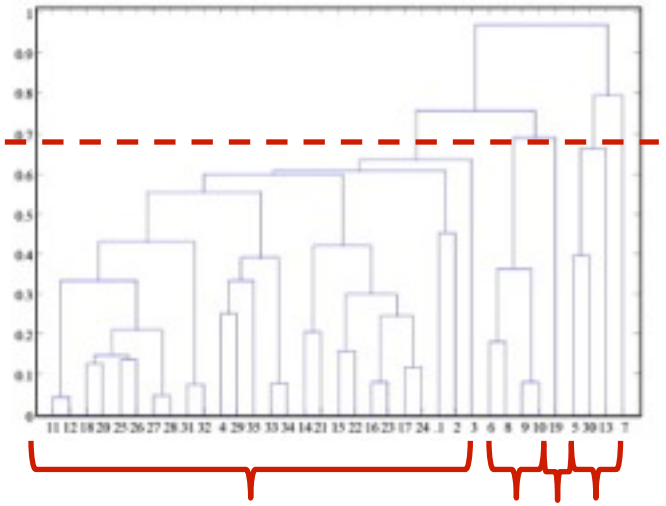


CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Partitional clustering

- partitional clustering is simply a division of the set of objects into non-overlapping subsets (clusters)
- note that hierarchical clustering can be viewed as a sequence of partitional clusterings
- a partitional clustering can be obtained by cutting the tree at a particular level

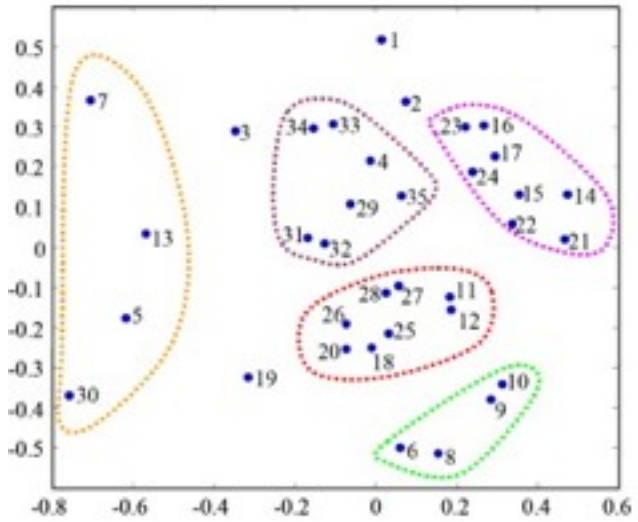


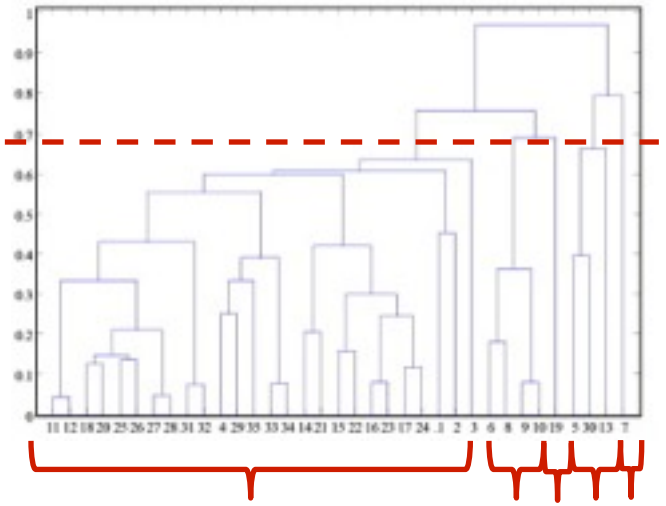


CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Partitional clustering

- partitional clustering is simply a division of the set of objects into non-overlapping subsets (clusters)
- note that hierarchical clustering can be viewed as a sequence of partitional clusterings
- a partitional clustering can be obtained by cutting the tree at a particular level

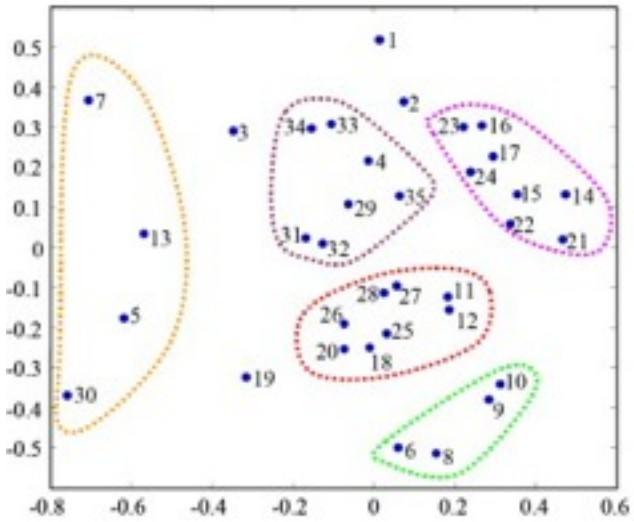


CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Complete vs. partial clustering

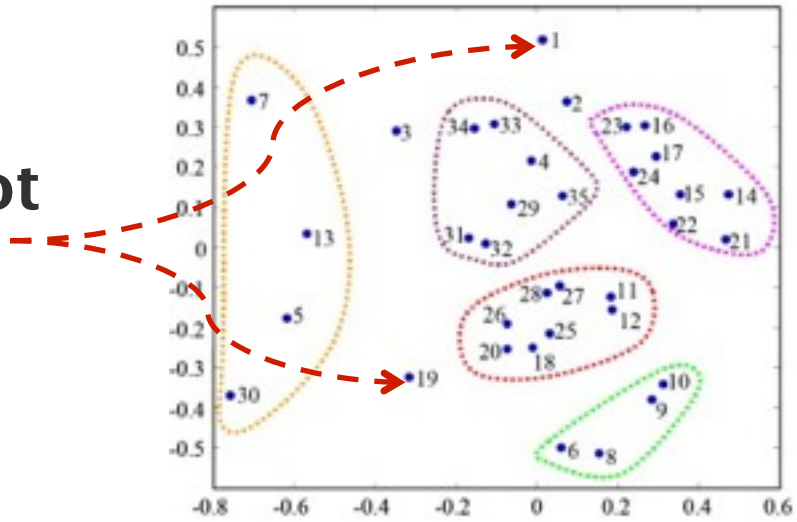
- a complete clustering assigns **every** object to a cluster
- a partial clustering does **not**
- the motivation for partial clustering is that some objects may not belong to well defined groups
- many times data objects may represent noise, outliers of uninteresting background




CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Complete vs. partial clustering

- a complete clustering assigns **every** object to a cluster
- a partial clustering does **not**
- the motivation for partial clustering is that some objects may not belong to well defined groups
- many times data objects may represent noise, outliers of uninteresting background



# Clusters

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Types of clusters

- clustering aims to find useful groups of objects
- usefulness is defined by the goals of data analysis
- there are several different notions of a cluster that prove useful in practice:

1. well-separated
2. prototype-based
3. graph-based
4. density-based
5. shared-property (conceptual)
6. described by an objective function

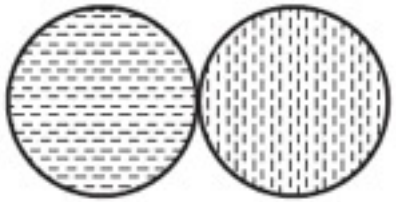
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Well-separated clusters

- each object is closer (or more similar) to every object in the cluster than to any other object not in the cluster
- a threshold can be used to specify that all objects in a cluster must be sufficiently close to one another
- idealistic definition of a cluster
- satisfied only when the data contains natural clusters that are quite far from one another
- well-separated clusters do not need to be spherical, but can have any shape

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

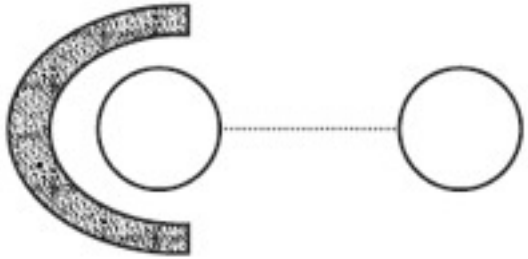
# Prototype-based clusters



- each object is closer (more similar) to the prototype that defines the cluster than the prototype of any other cluster
- for many types of data, the prototype can be regarded as the most central point
- in such instances, we commonly refer to prototype-based clusters as centre-based clusters
- such clusters tend to be spherical

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

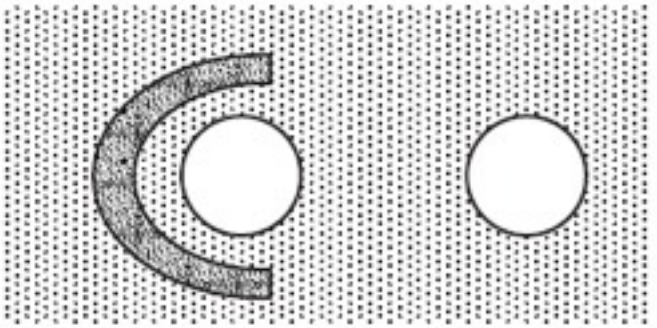
# Graph-based clusters



- if the data is represented as a graph, where the nodes are objects and the links represent connections between objects, then a cluster can be defined as a connected component
- a group of objects that are connected to one another, but are not connected to objects outside of the group
- an important example of graph-based clusters are contiguity-based clusters, where two objects are connected if they are within a specified distance from each other

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Density-based clusters

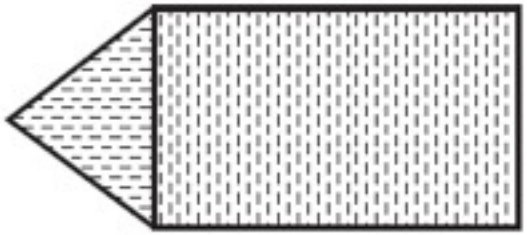


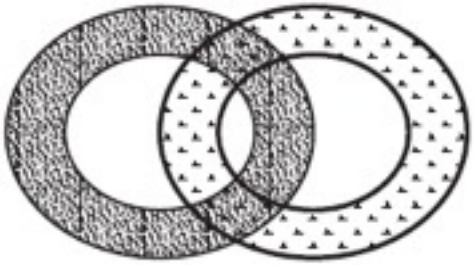
- a cluster is a dense region of objects that is surrounded by a region of low density
- a density-based definition of a cluster is often used when the clusters are irregular or intertwined and when noise and outliers are present
- a contiguity-based definition of a cluster would not work well since the noise would tend to form connections between clusters

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Shared-property (conceptual) clusters

- a group of objects that share some property
- encompasses all previous definitions of clusters
- ... but also covers new types of clusters
- points in a cluster share some general property that derives from the entire set of objects





- too sophisticated
- takes us into the area of pattern recognition

# Clusters described by an objective function

- finds clusters that minimise or maximise an objective function
- enumerate all possible ways of dividing the points into clusters and evaluate them using the objective function (NP hard!)
- can have global or local objectives
  - hierarchical clustering algorithms typically have local objectives
  - partitional clustering algorithms typically have global objectives

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Clustering strategies

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Hierarchical clustering

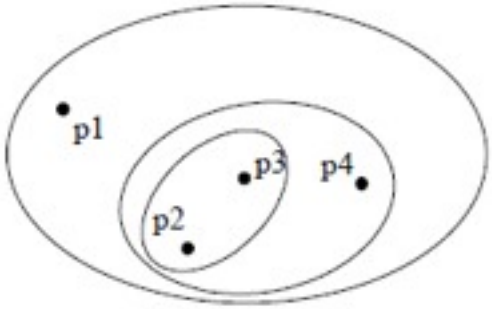
- hierarchical clustering = a method of cluster analysis which seeks to build a hierarchy of clusters
- two strategies:

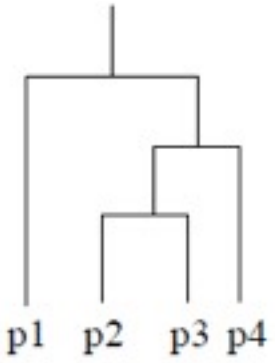
1. agglomerative (bottom-up approach)
   - each instance starts separately in its own cluster
   - pairs of clusters are merged recursively
2. divisive (top-down approach)
   - all instances start together in a single cluster
   - clusters are split performed recursively

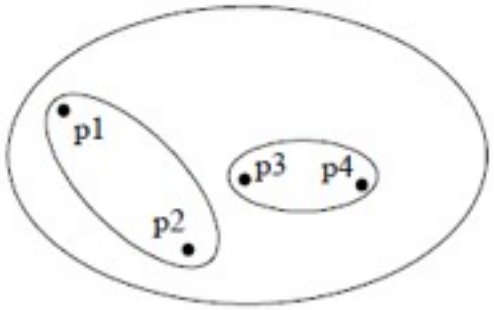
- the result is usually presented as a dendrogram
- … may respond to a meaningful taxonomy

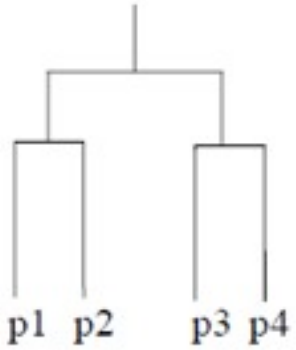
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Hierarchical clustering




nested cluster diagram

dendrogram

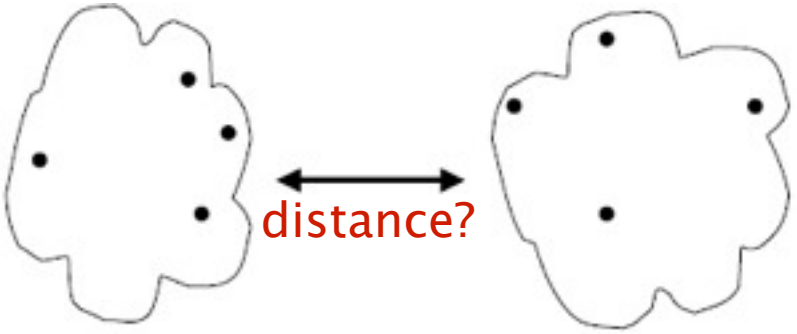
# Agglomerative hierarchical clustering

- starting with individual points as clusters
- successively merge the two closest clusters until only one cluster remains
- basic algorithm:

1. Compute the distance matrix.
2. **repeat**
3. Merge the closest two clusters.
4. Update the distance matrix to reflect the distance between the new cluster and the original clusters.
5. **until** Only one cluster remains.

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Cluster distance

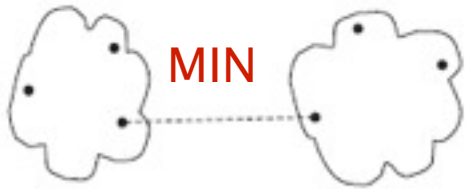


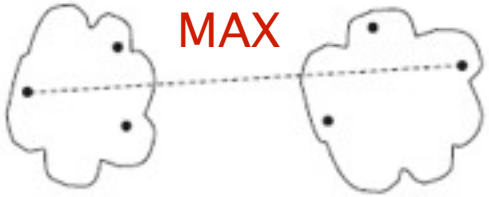

- the key operation of the algorithm is the computation of the distance between clusters
- the definition of cluster distance differentiates the various agglomerative hierarchical techniques
- cluster distance is typically defined with a particular type of cluster in mind
- e.g. many agglomerative hierarchical clustering techniques, such as MIN, MAX & group average are defined with a graph-based clusters in mind

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Cluster distance

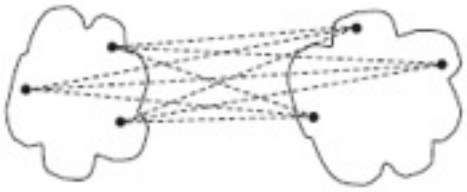
- MIN defines cluster distance as the distance between the closest two points in different clusters
- this yields contiguity-based clusters
- MAX defines cluster distance as the distance between the farthest two points in different clusters
- alternative names:
  - MIN = single link
  - MAX = complete link



# Cluster distance

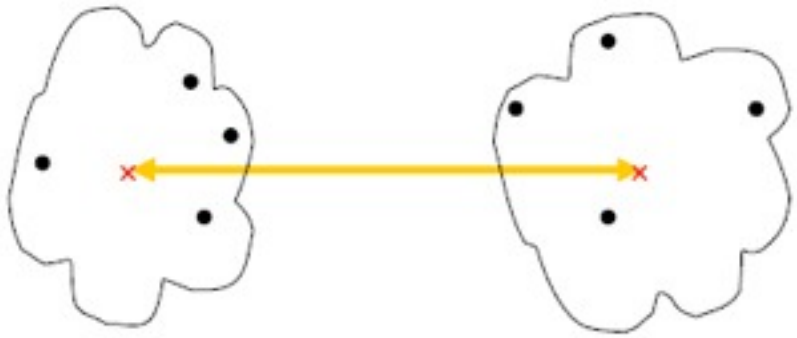
- the group average technique defines cluster distance to be the average pair-wise distances of all pairs of points from different clusters



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Cluster distance

- if, instead of a graph-based view, we take a prototype-based view...
- each cluster is represented by a centroid
- distance is commonly defined as the distance between cluster centroids



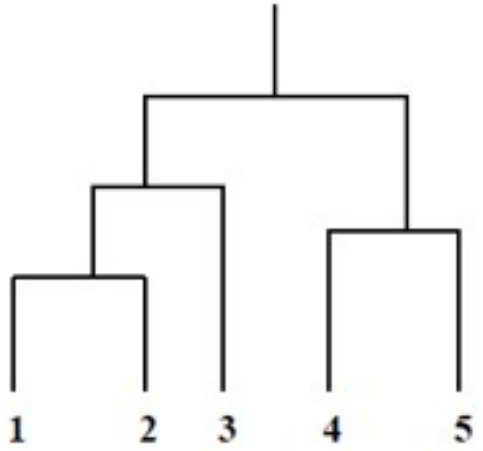
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Single-link (MIN) hierarchical clustering

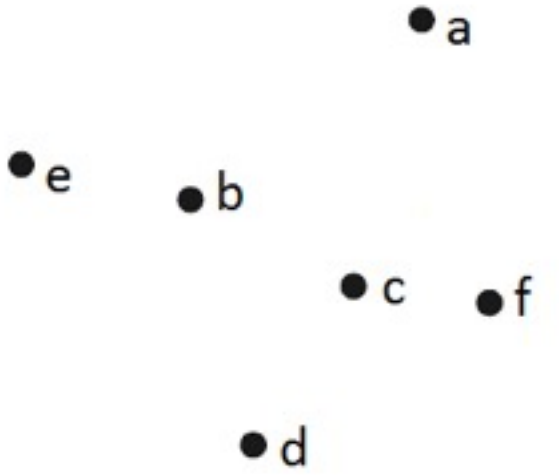
# Single link (MIN) hierarchical clustering

- cluster distance is defined as the distance between the closest two points in different clusters
- distance matrix:

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0.00 | 0.10 | 0.90 | 0.35 | 0.80 |
| 2 | 0.10 | 0.00 | 0.30 | 0.40 | 0.50 |
| 3 | 0.90 | 0.30 | 0.00 | 0.60 | 0.70 |
| 4 | 0.35 | 0.40 | 0.60 | 0.00 | 0.20 |
| 5 | 0.80 | 0.50 | 0.70 | 0.20 | 0.00 |

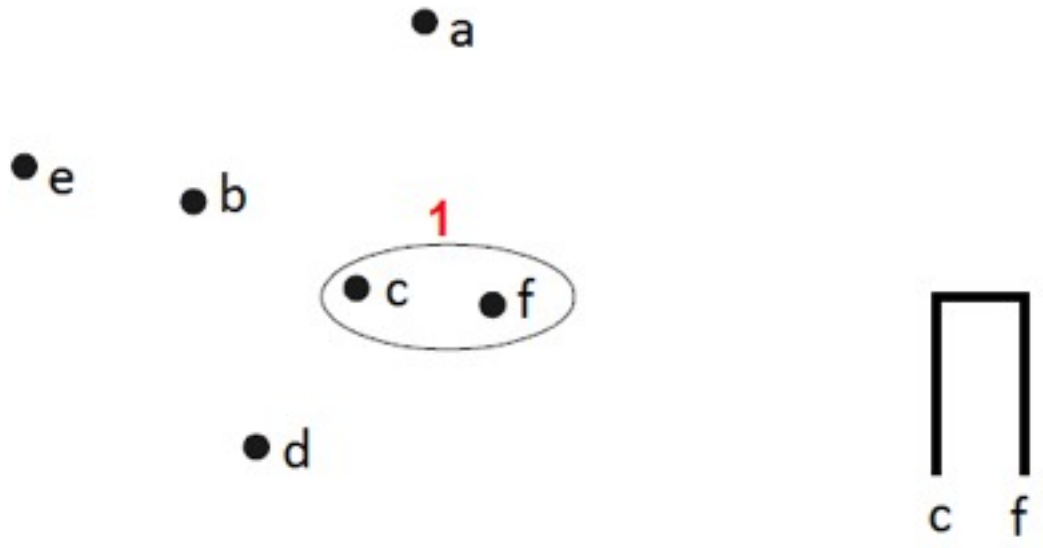


# Single link (MIN) hierarchical clustering



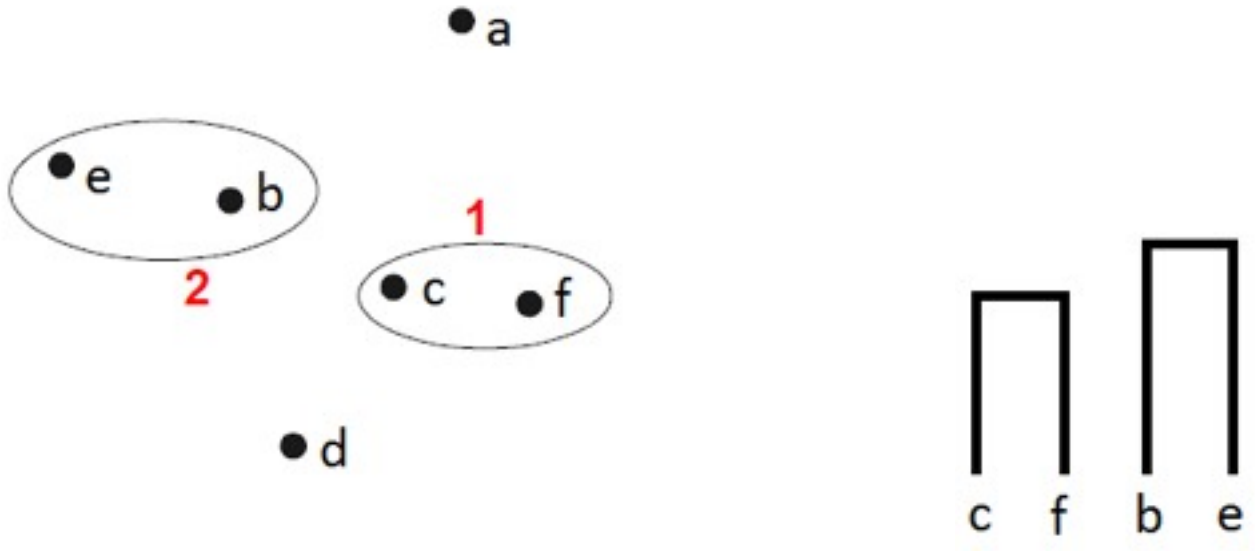

- start with all points as singleton clusters:
  - 6 clusters: {a}, {b}, {c}, {d}, {e}, {f}
- at each step merge two clusters with two closest points

# Single link (MIN) hierarchical clustering



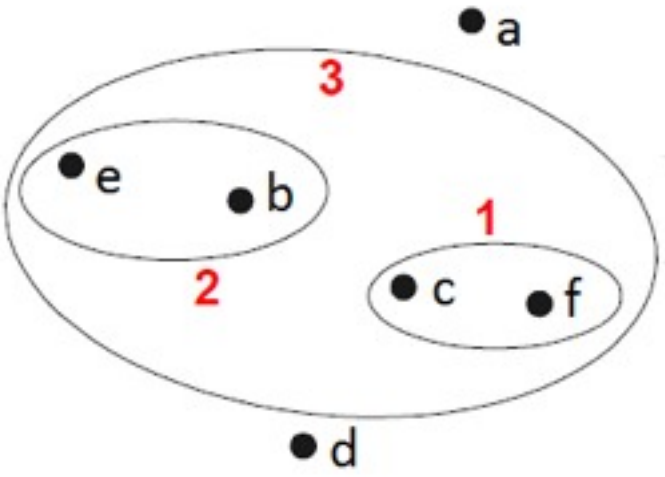

- c and f are the two closest points
- 5 clusters: {a}, {b}, {c, f}, {d}, {e}

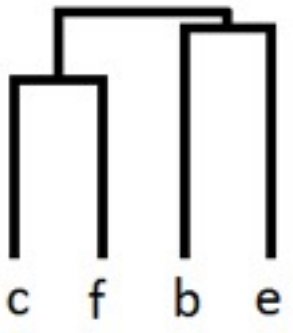
# Single link (MIN) hierarchical clustering




- {b} and {e} are the two closest clusters
- 4 clusters: {a}, {b, e}, {c, f}, {d}

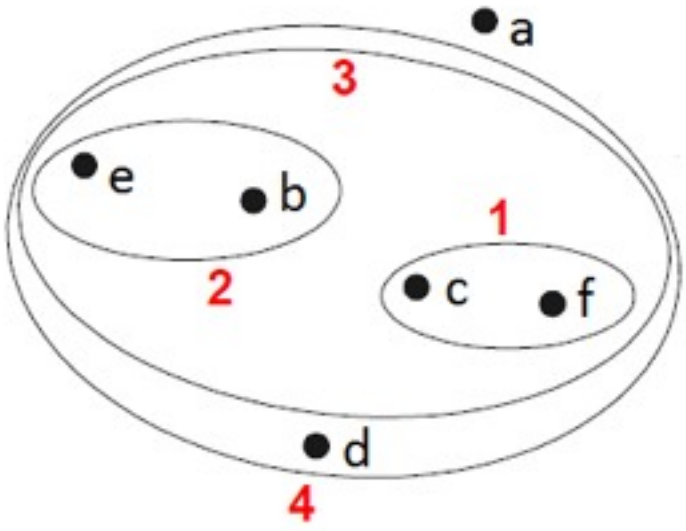
# Single link (MIN) hierarchical clustering

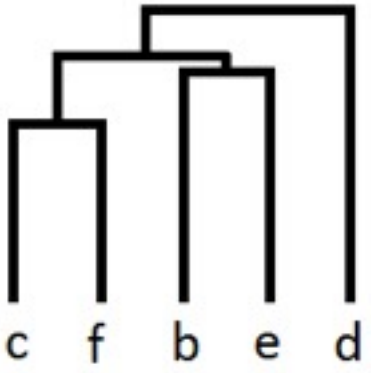



- {c, f} and {b, e} are two closest clusters because of the distance between b and c
- 3 clusters: {a}, {b, e, c, f}, {d}

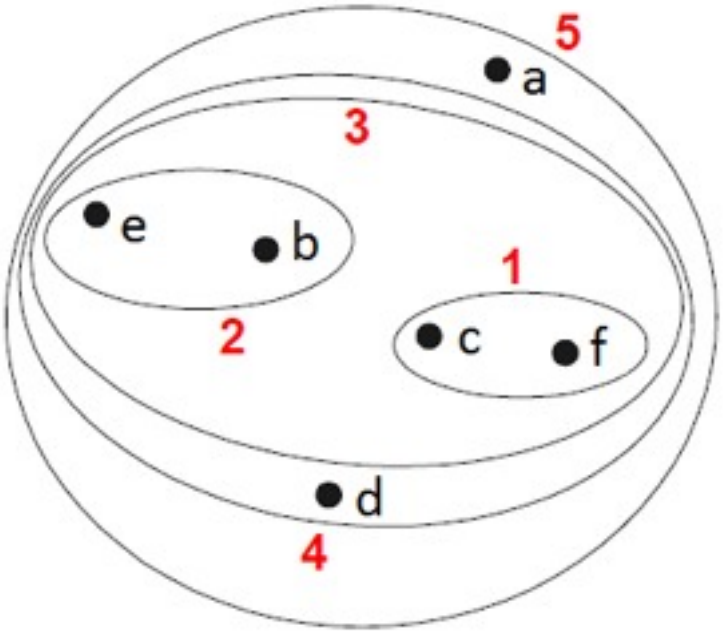
# Single link (MIN) hierarchical clustering

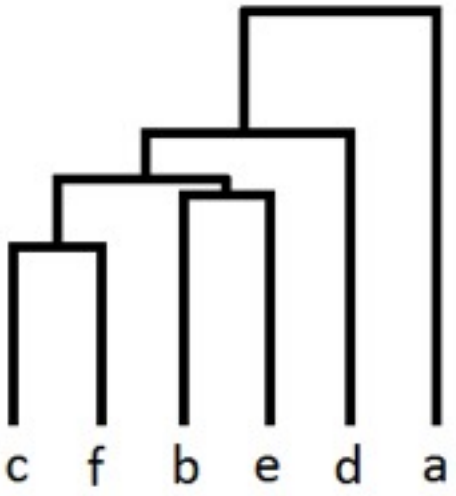



- d is closer to c than a is to b
- 2 clusters: {a}, {b, c, d, e, f}

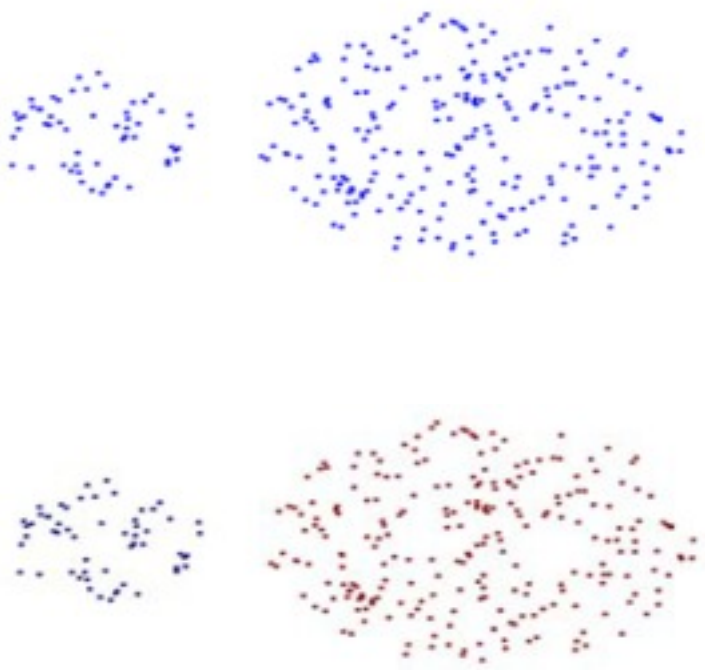
# Single link (MIN) hierarchical clustering



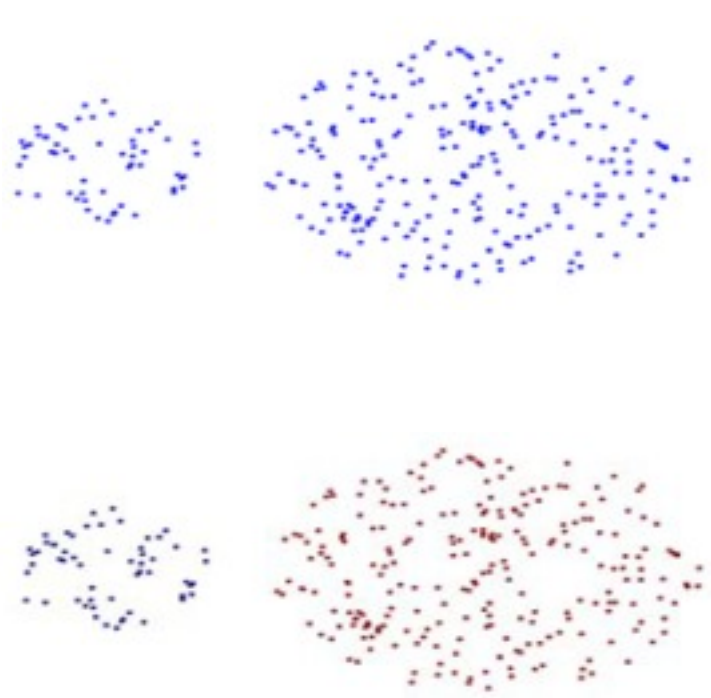

- merge the only two remaining clusters
- 1 cluster: {a, b, c, d, e, f}

# Strength/limitation of MIN



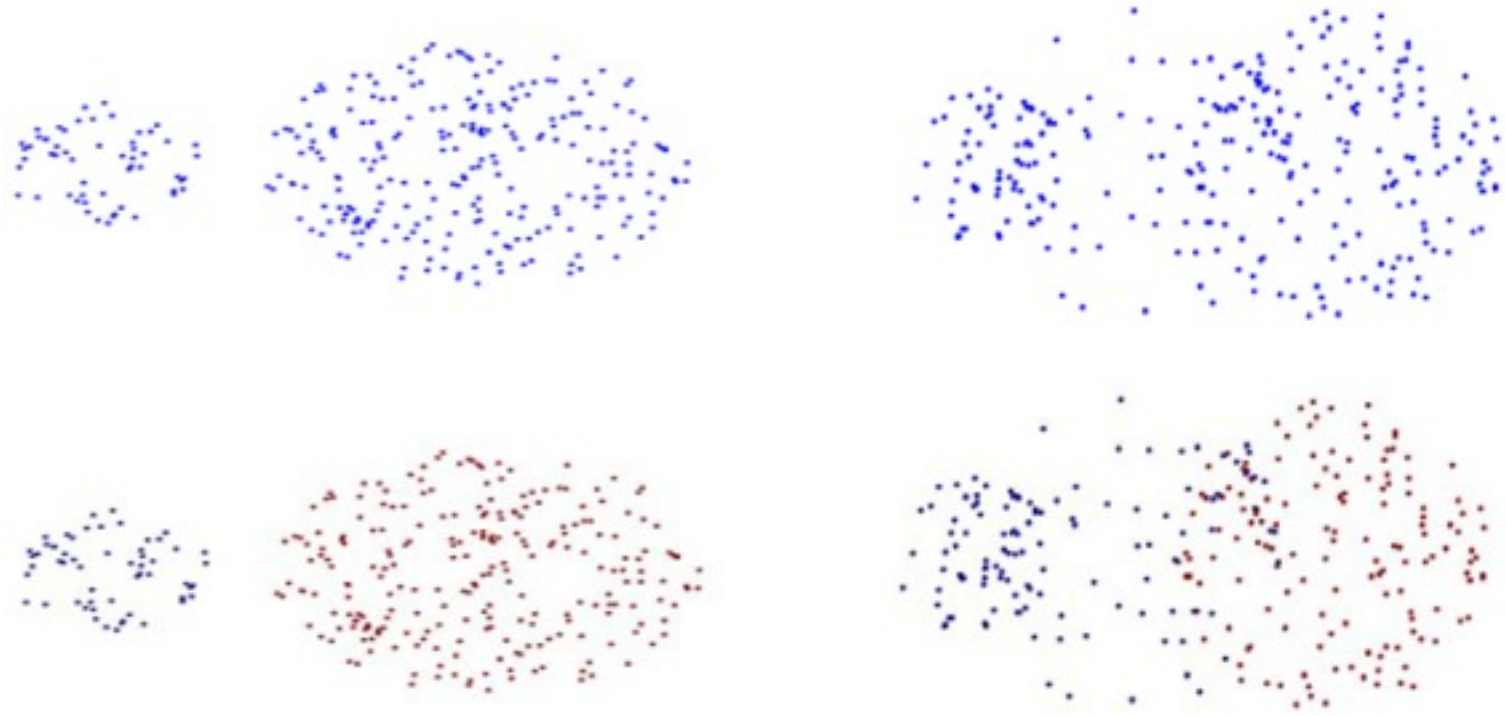
- strength: can handle non-elliptical shapes

# Strength/limitation of MIN



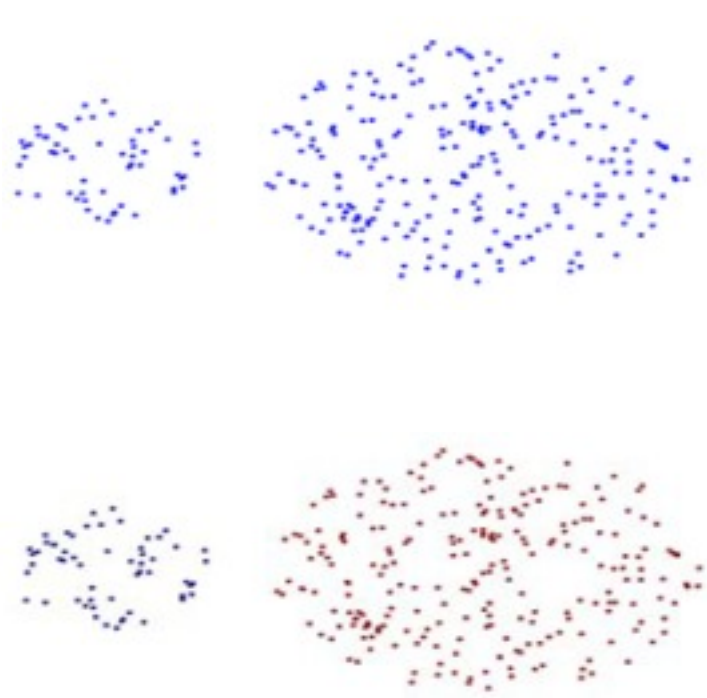
- strength: can handle non-elliptical shapes

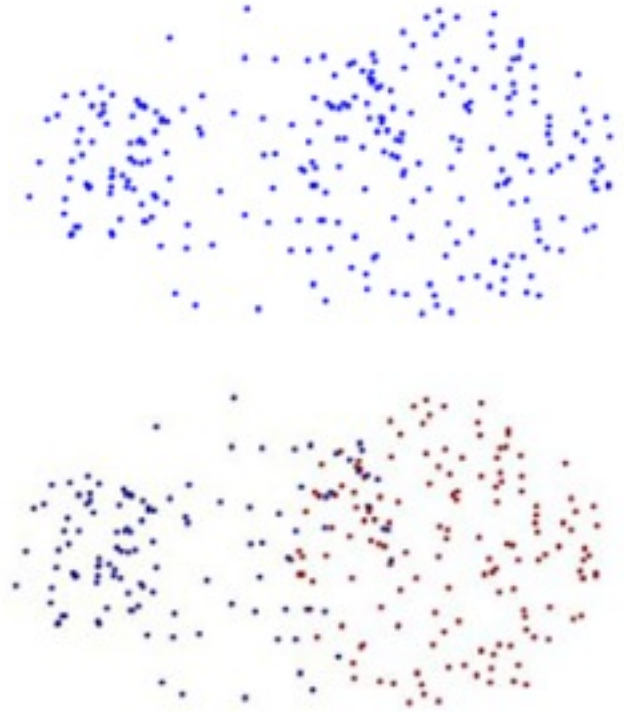
# Strength/limitation of MIN



- strength: can handle non-elliptical shapes

# Strength/limitation of MIN

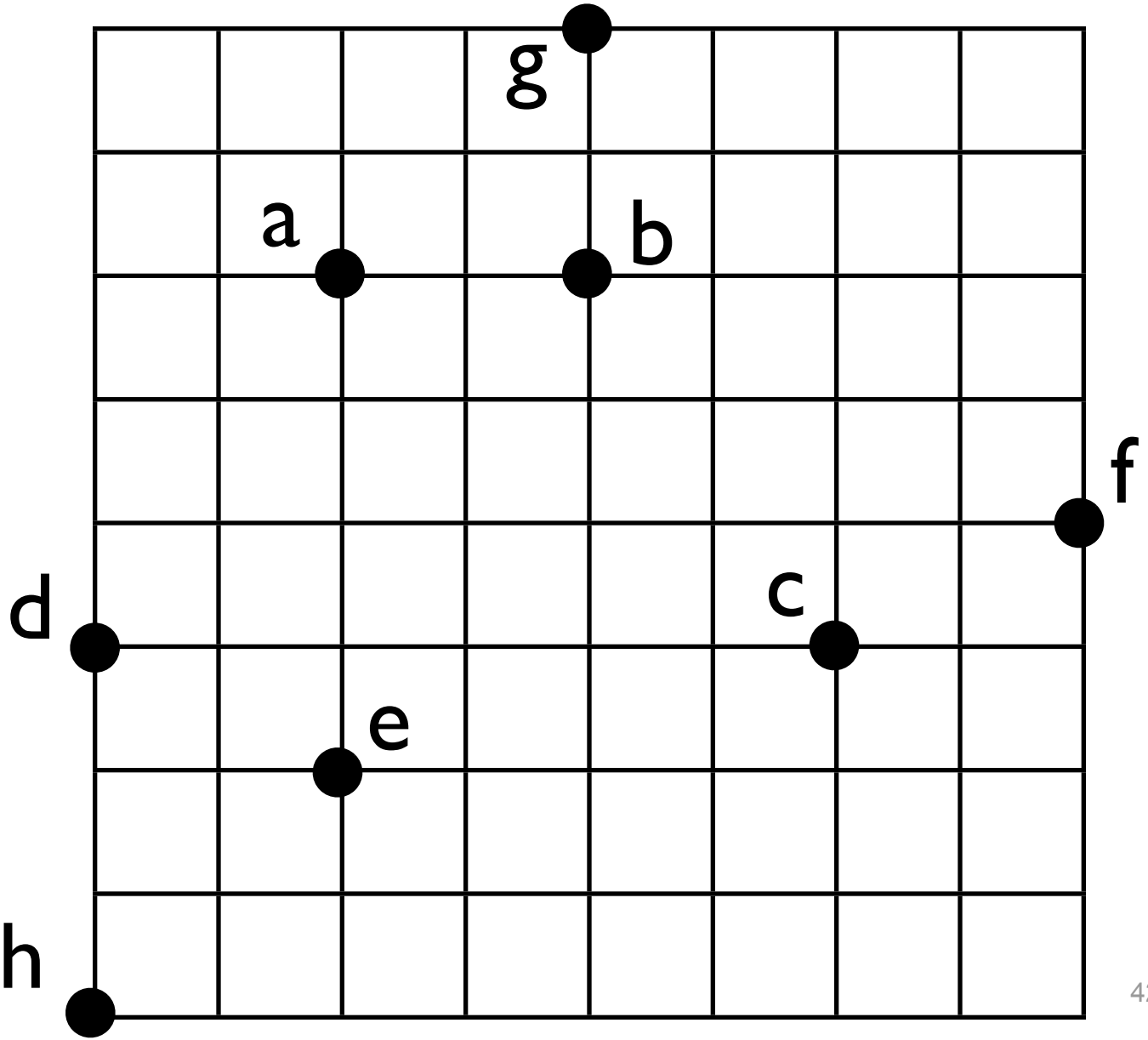




- strength: can handle non-elliptical shapes
- limitation: sensitive to noise and outliers

# Exercise: use Euclidean distance

(hint: $\sqrt{}$ is monotone – ignore it)



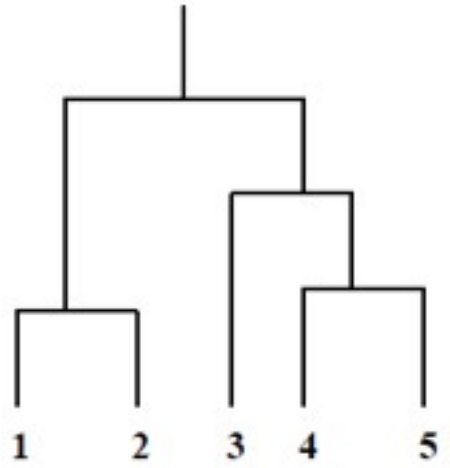

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Complete-link (MAX) hierarchical clustering

# Complete link (MAX) hierarchical clustering

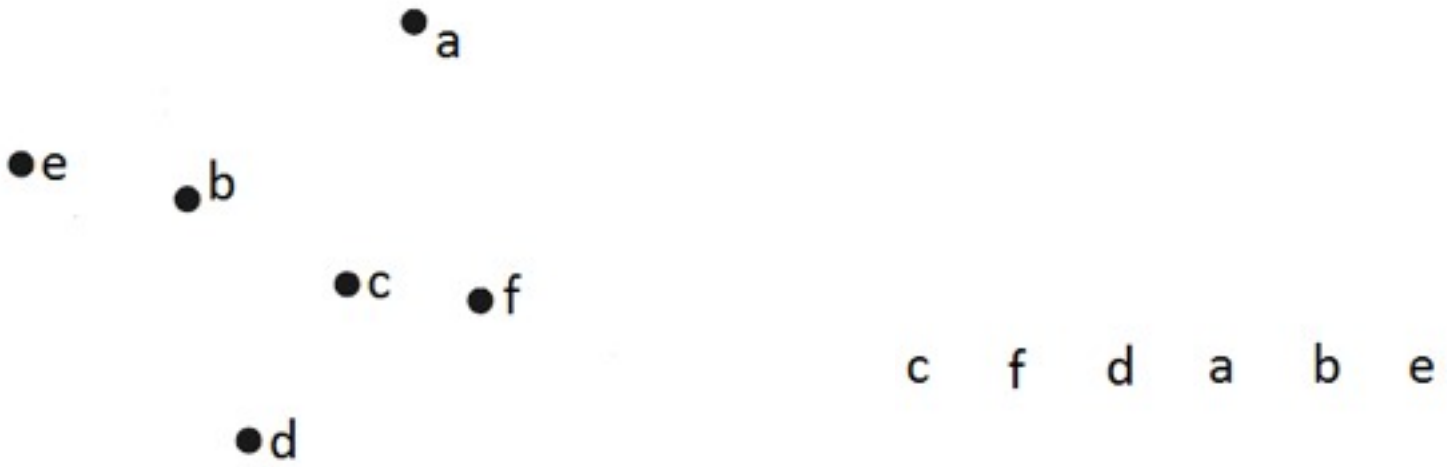
- cluster distance is defined as the distance between the farthest two points in different clusters
- distance matrix:

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0.00 | 0.10 | 0.90 | 0.35 | 0.80 |
| 2 | 0.10 | 0.00 | 0.30 | 0.40 | 0.50 |
| 3 | 0.90 | 0.30 | 0.00 | 0.60 | 0.70 |
| 4 | 0.35 | 0.40 | 0.60 | 0.00 | 0.20 |
| 5 | 0.80 | 0.50 | 0.70 | 0.20 | 0.00 |

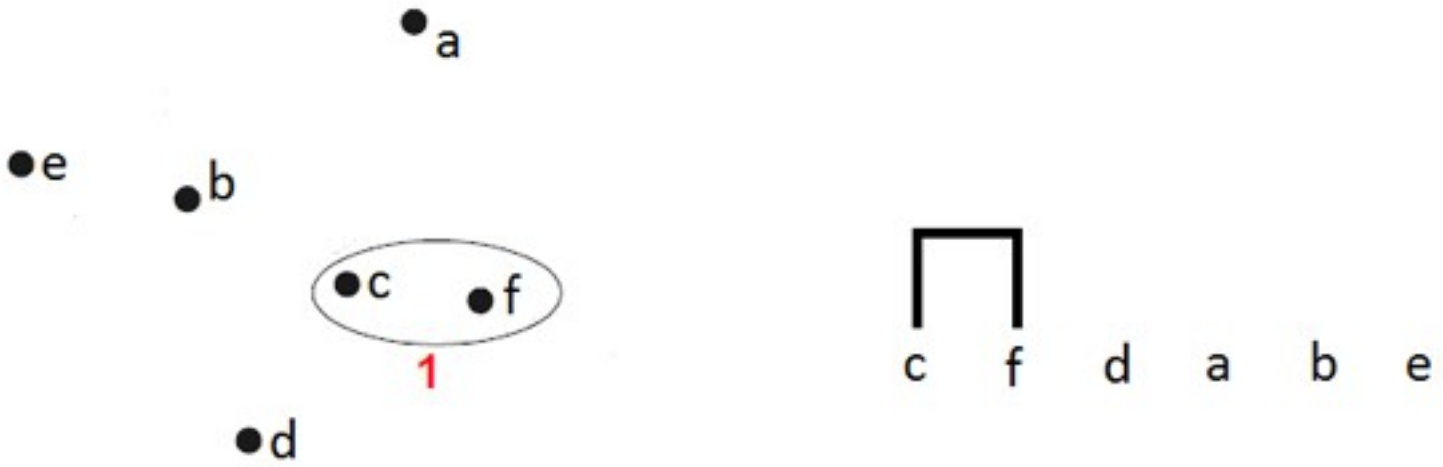



- dist({3}, {1, 2}) = dist(3, 1) = 0.90
- dist({3}, {4, 5}) = dist(3, 5) = 0.70
- dist({1, 2}, {4, 5}) = dist(1, 5) = 0.80

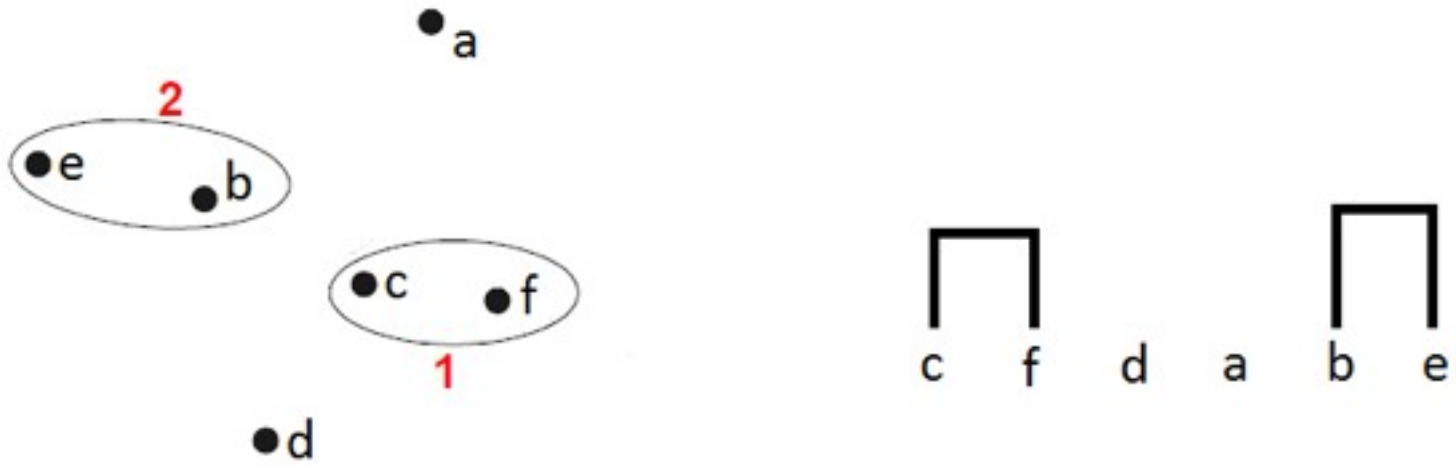
# Complete link (MAX) hierarchical clustering



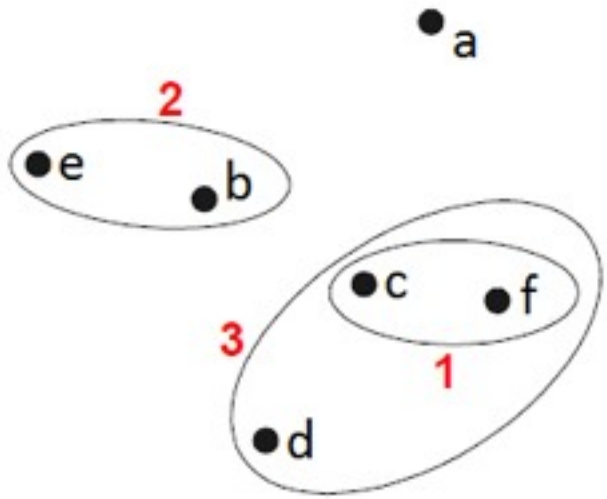
# Complete link (MAX) hierarchical clustering

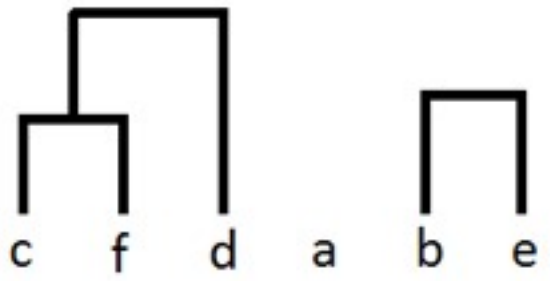


# Complete link (MAX) hierarchical clustering

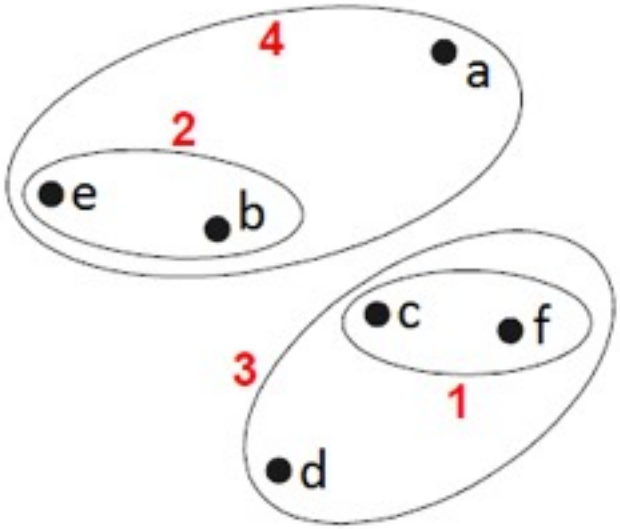


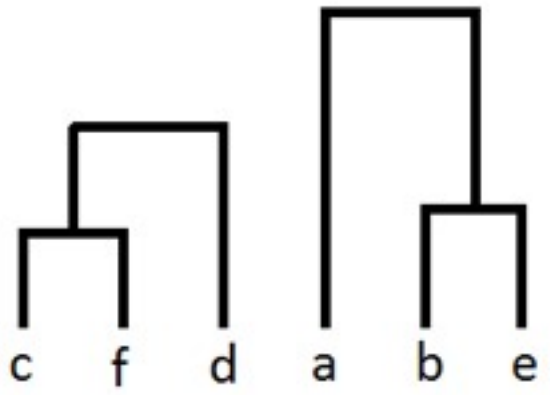
# Complete link (MAX) hierarchical clustering



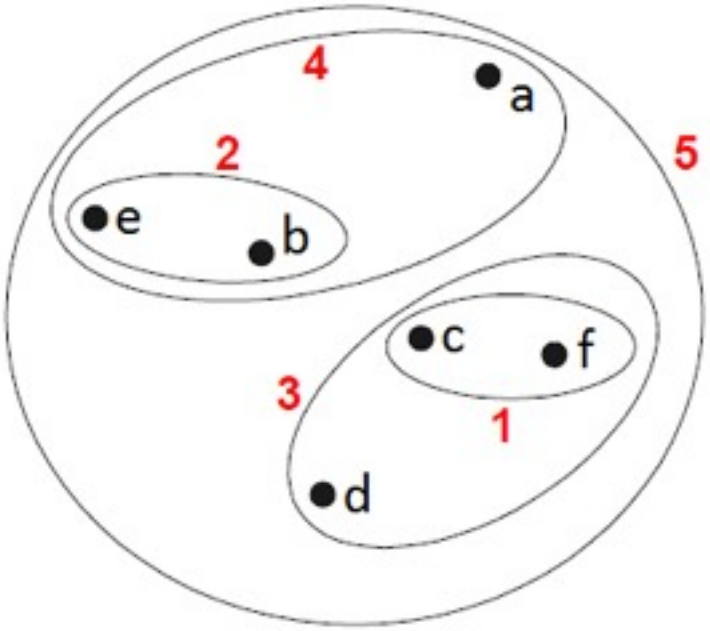



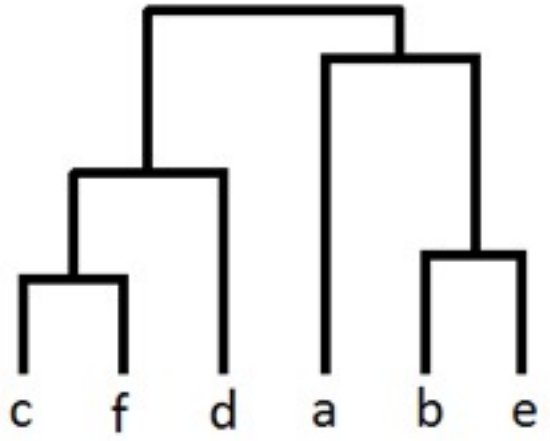
# Complete link (MAX) hierarchical clustering



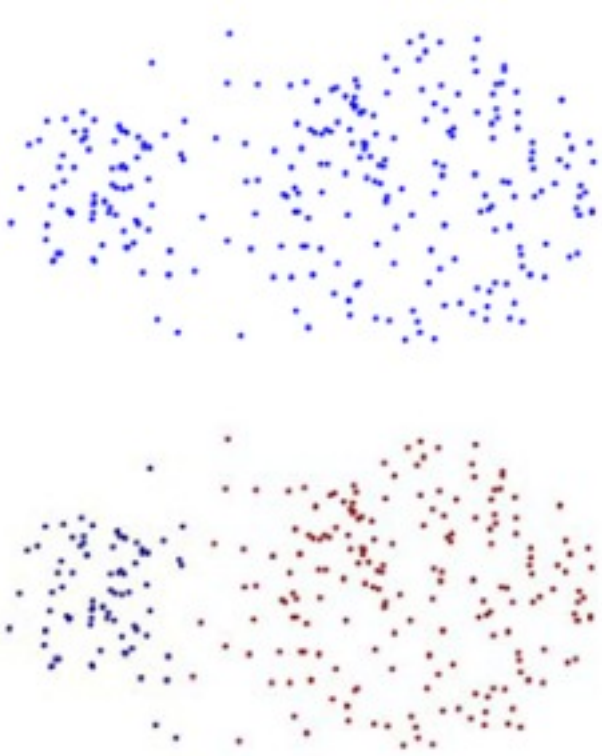



# Complete link (MAX) hierarchical clustering



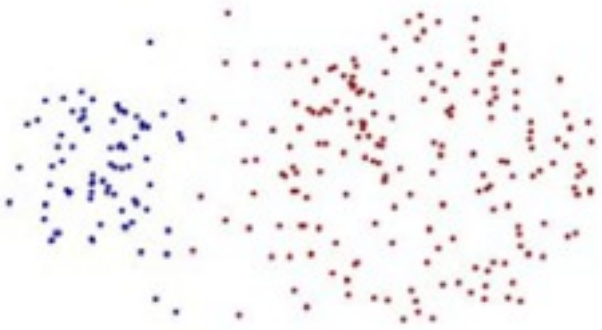



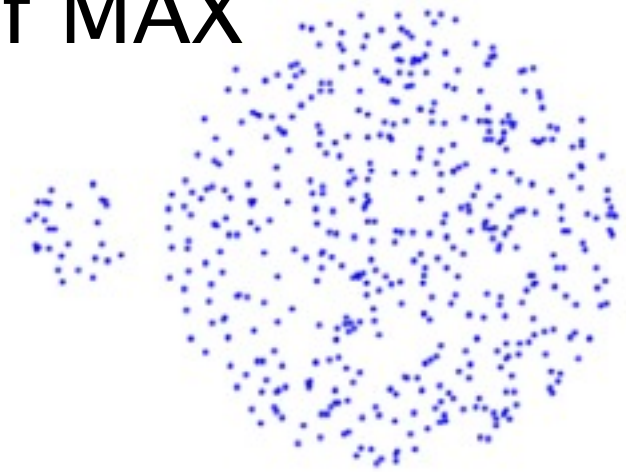
# Strength/limitation of MAX



- strength: less susceptible to noise and outliers

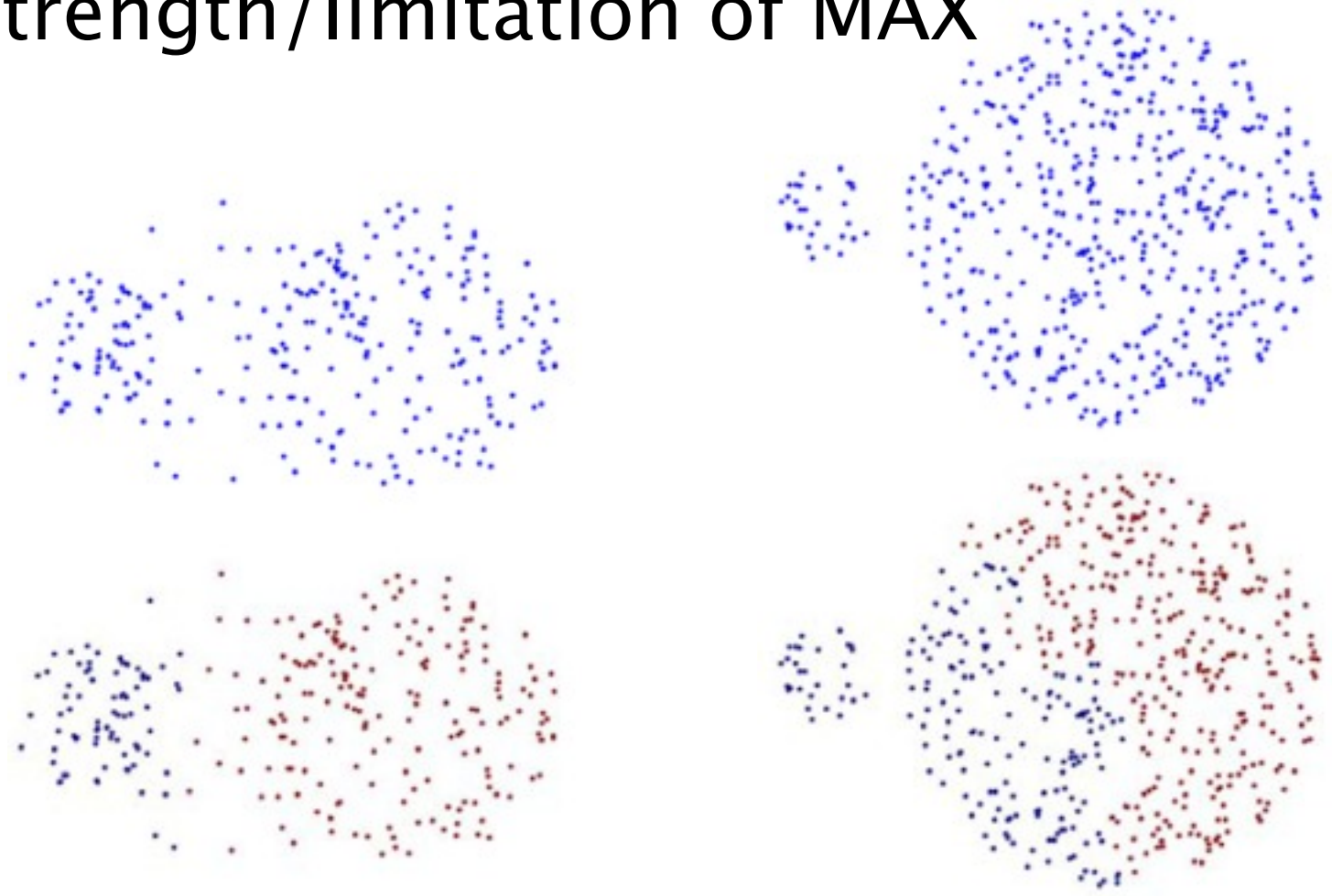
# Strength/limitation of MAX





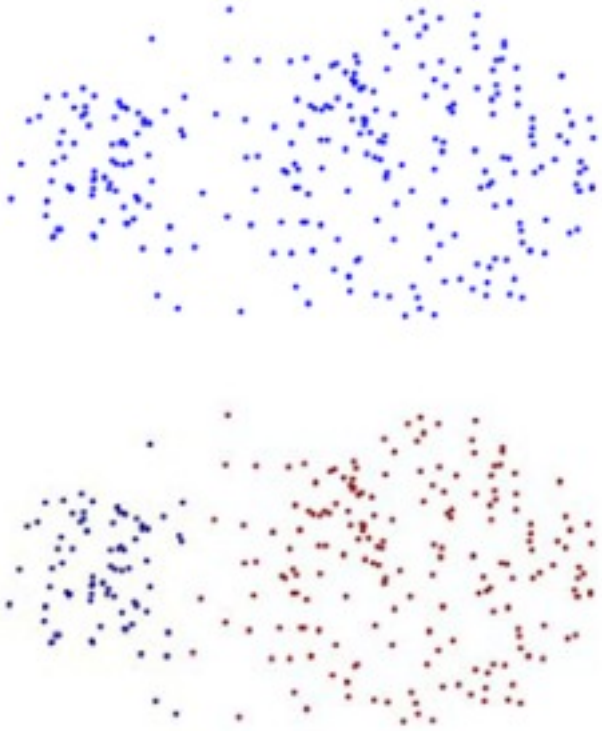
- strength: less susceptible to noise and outliers

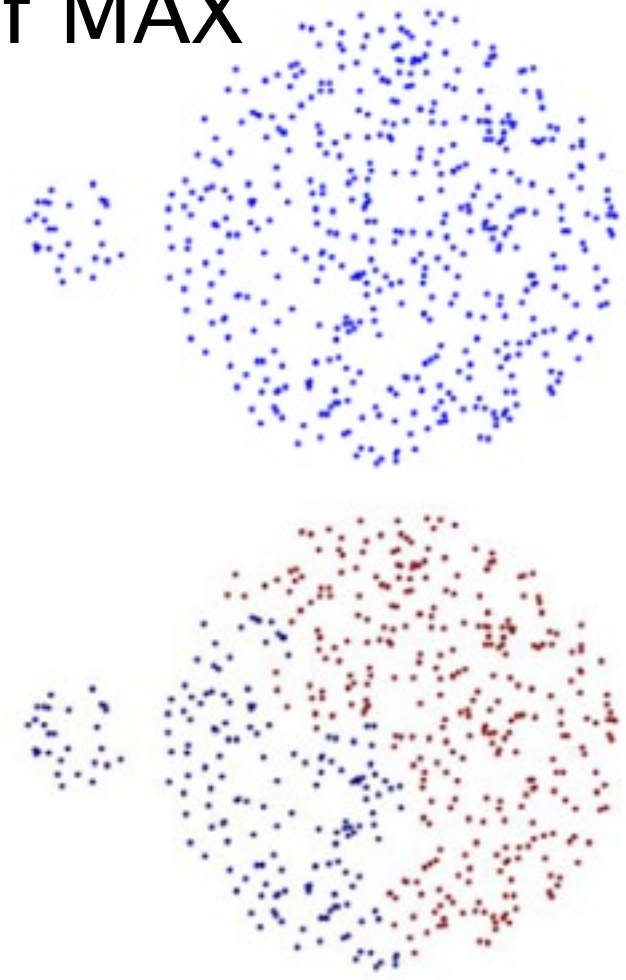
# Strength/limitation of MAX



- strength: less susceptible to noise and outliers

# Strength/limitation of MAX

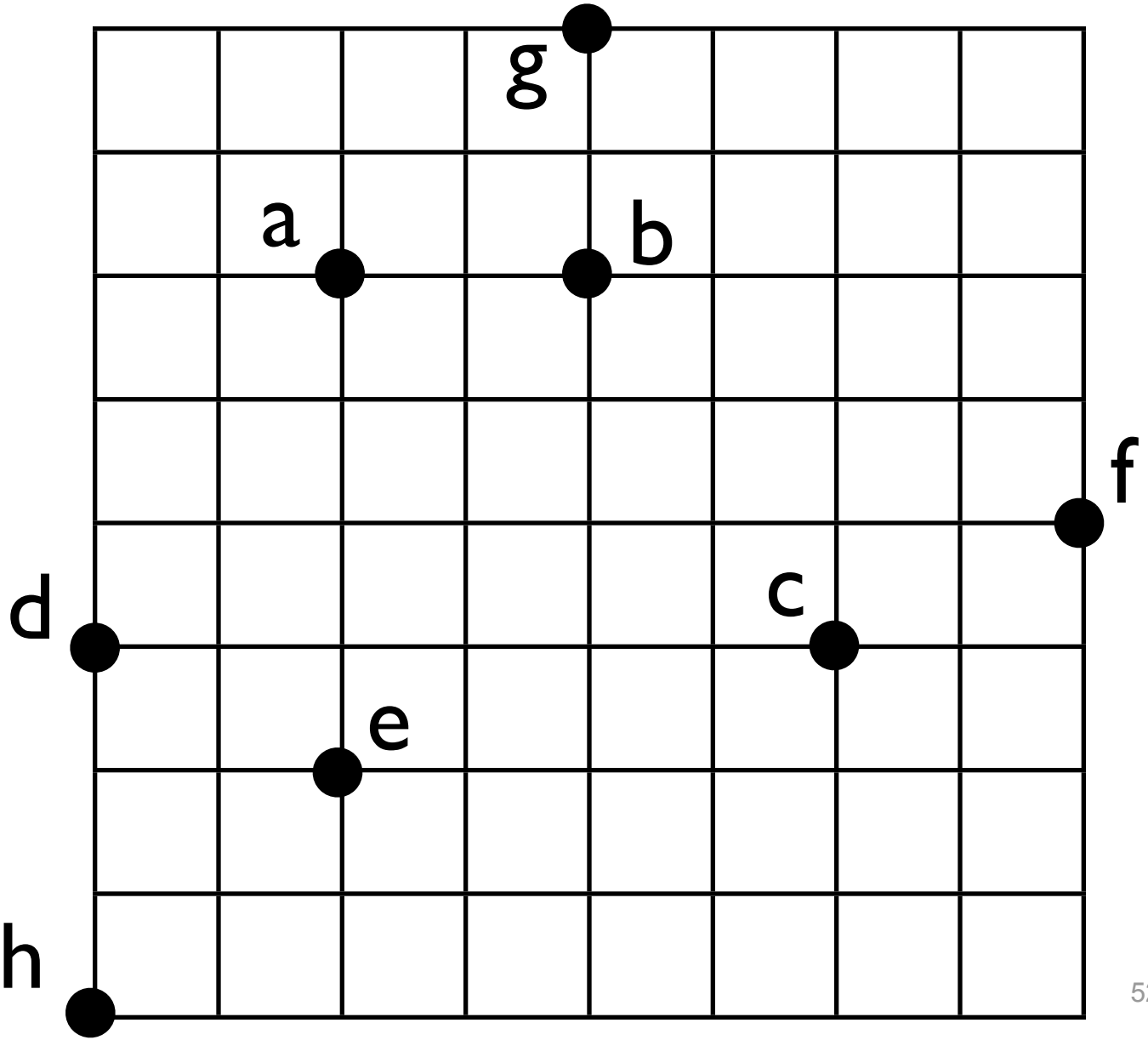




- strength: less susceptible to noise and outliers
- limitation: tends to break large clusters

# Exercise: use Euclidean distance

(reuse values from previous exercise)



# Hierarchical clustering: group average

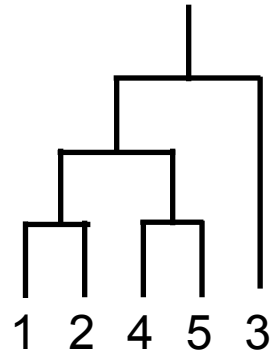
# Hierarchical clustering: group average

- cluster distance is defined as the average of pair-wise distance between points in the two clusters

$$\text{dist}(C_1, C_2) = \frac{\sum_{p_i \in C_1, p_j \in C_2} \text{dist}(p_i, p_j)}{|C_1| \cdot |C_2|}$$

- distance matrix:

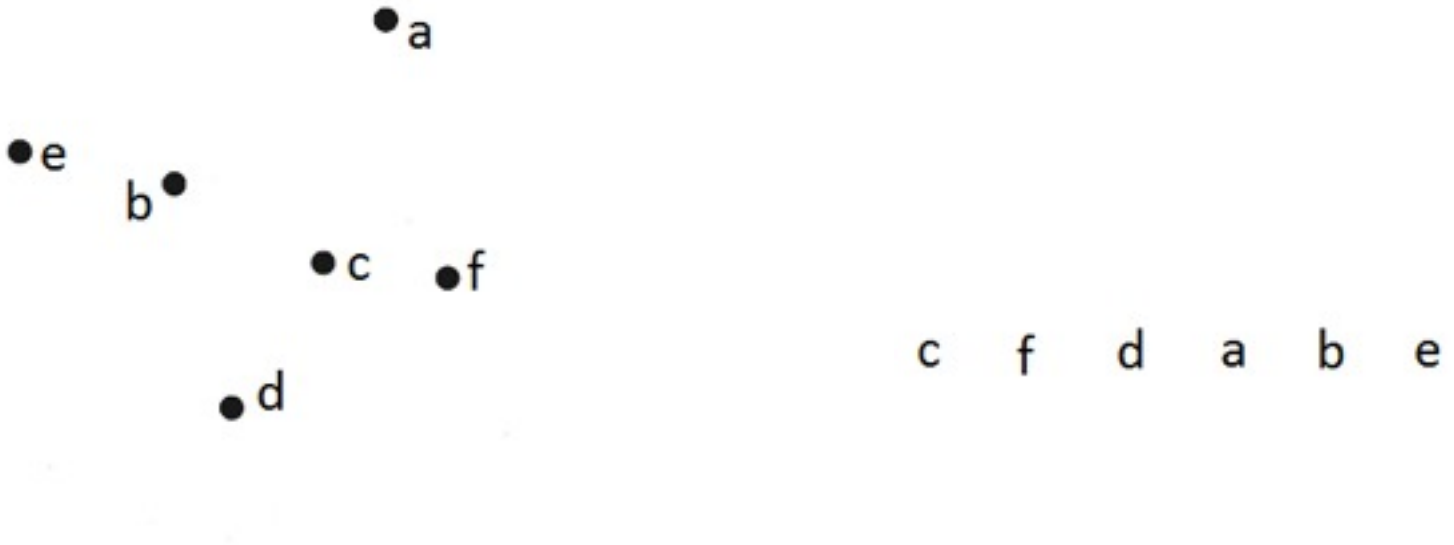
| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0.00 | 0.10 | 0.90 | 0.35 | 0.80 |
| 2 | 0.10 | 0.00 | 0.30 | 0.40 | 0.50 |
| 3 | 0.90 | 0.30 | 0.00 | 0.60 | 0.70 |
| 4 | 0.35 | 0.40 | 0.60 | 0.00 | 0.20 |
| 5 | 0.80 | 0.50 | 0.70 | 0.20 | 0.00 |



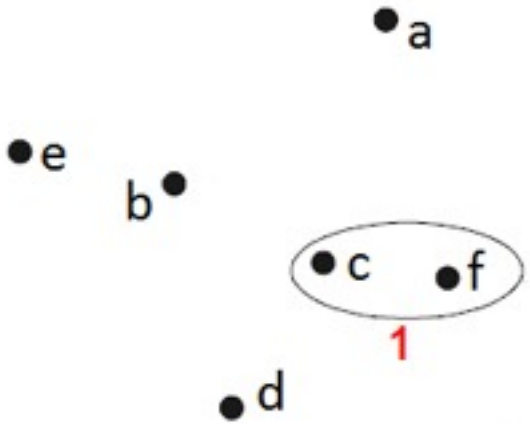

d({1,2},3)=(0.9+0.3)/2=0.6
d(3,{4,5})=(0.6+0.7)/2=0.65
d({1,2},{4,5})=(0.35+0.8+0.4+0.5)/4=0.5125

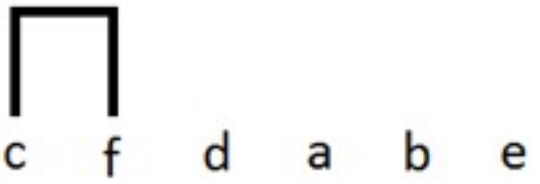
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Hierarchical clustering: group average

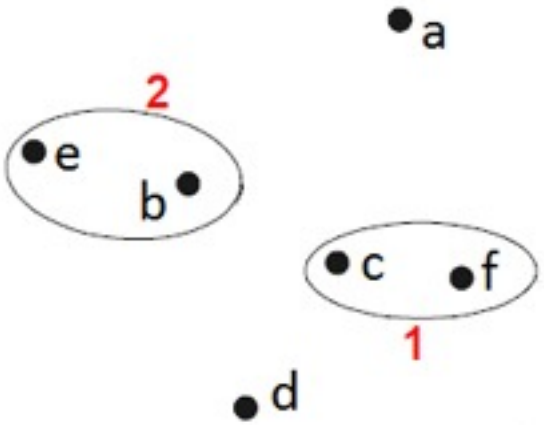


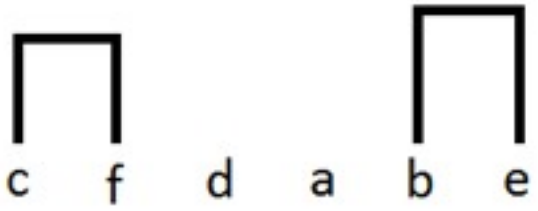
# Hierarchical clustering: group average



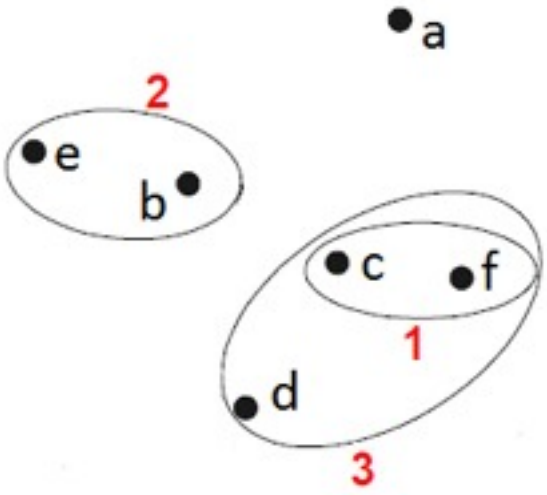



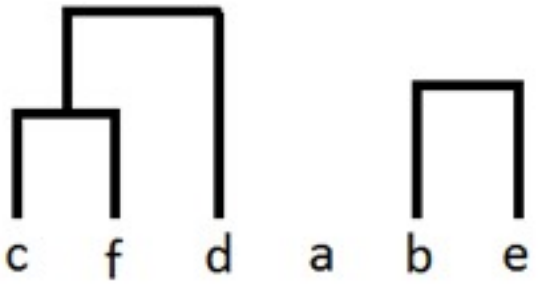
# Hierarchical clustering: group average



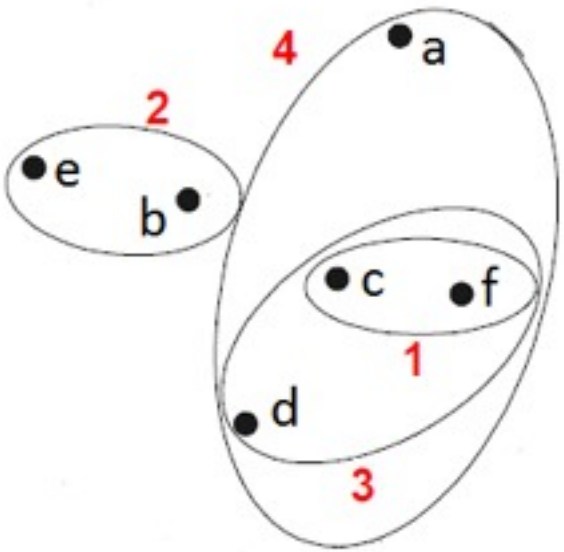



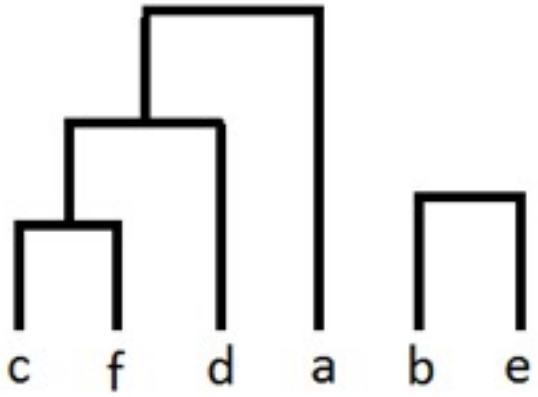
# Hierarchical clustering: group average



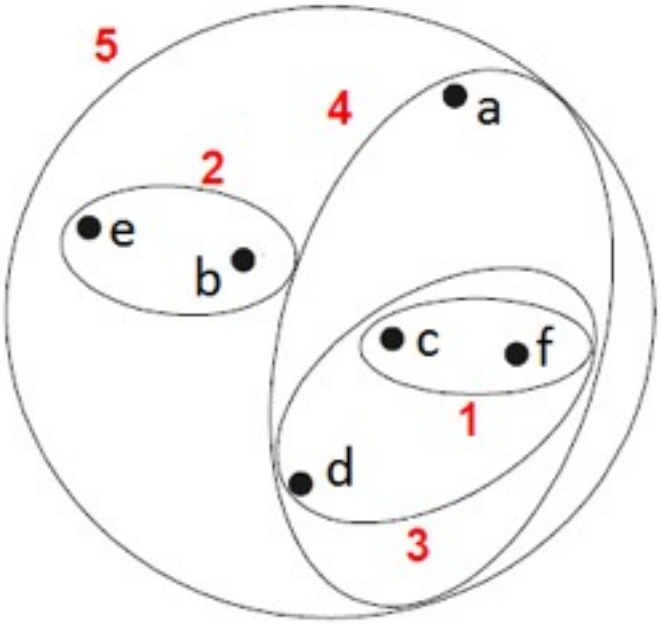



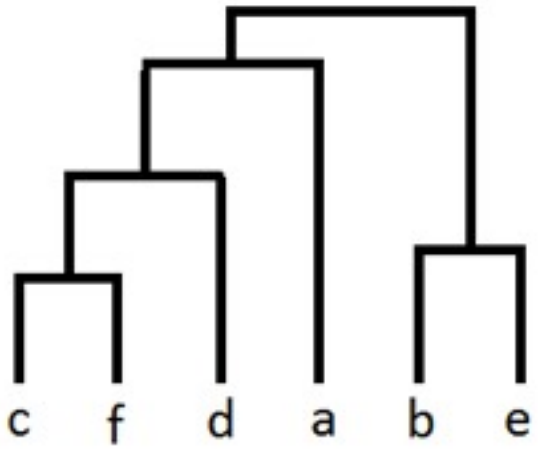
# Hierarchical clustering: group average



# Hierarchical clustering: group average



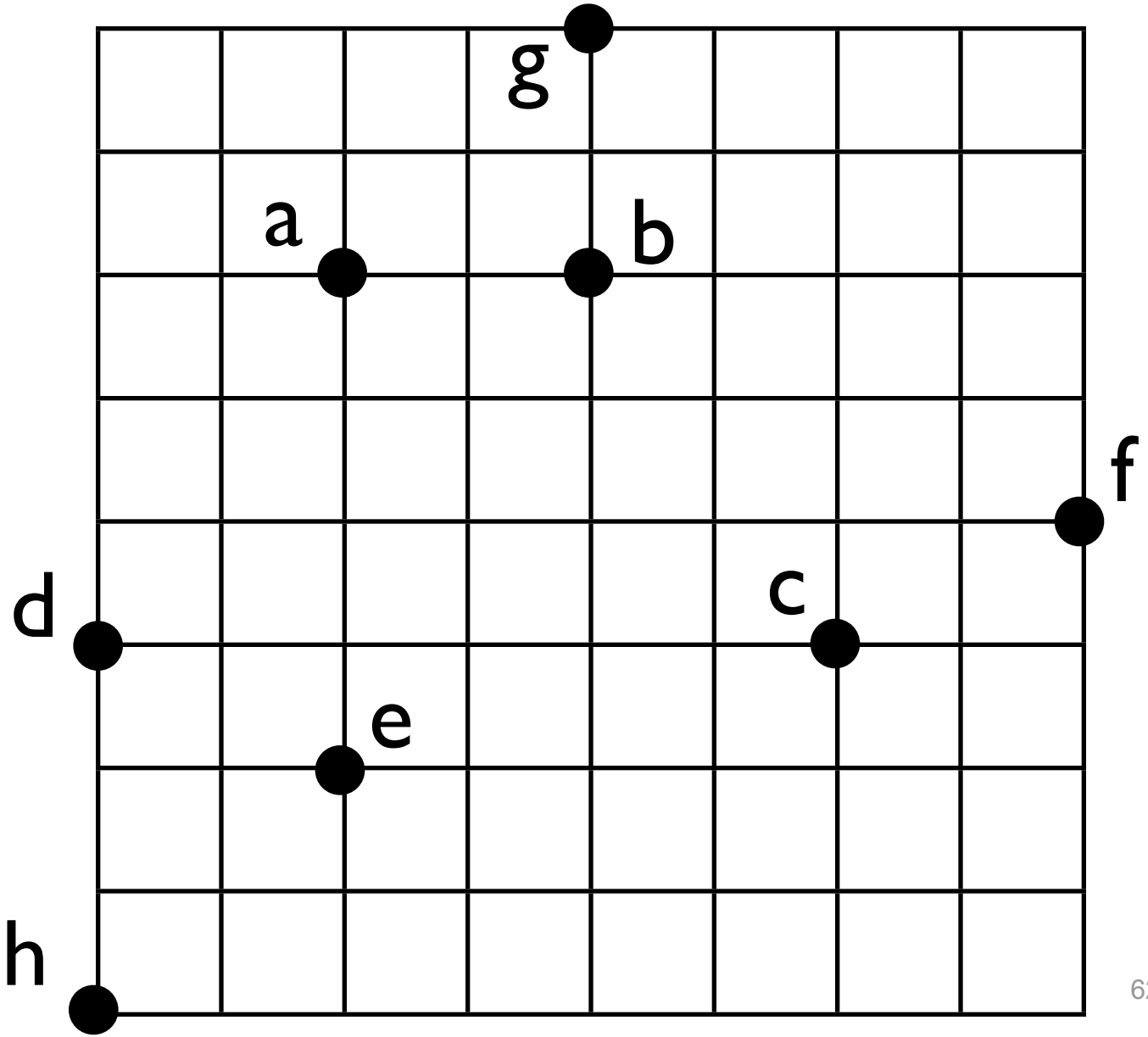
# Strength/limitation of group average

- group average is a compromise between MIN and MAX
- strength: less susceptible to noise and outliers
- limitation: biased towards spherical clusters

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Exercise: use Euclidean distance

(reuse values from previous exercise)

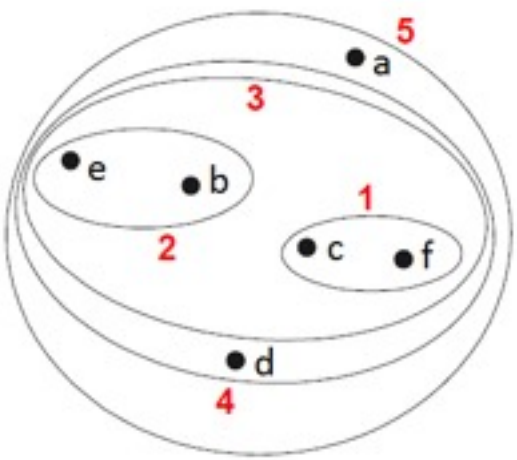



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Hierarchical clustering: Summary

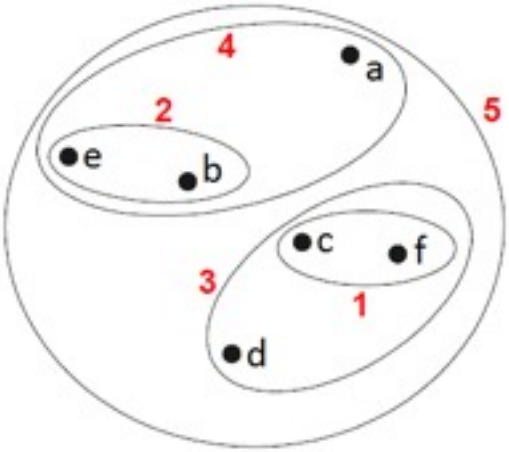
# Comparison

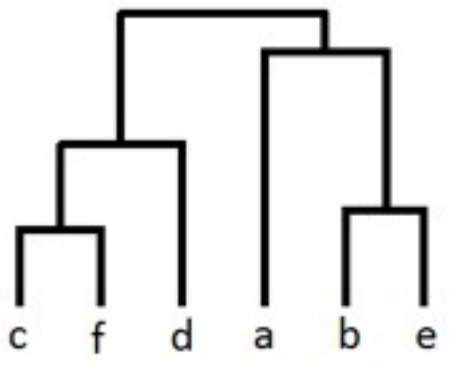
MIN



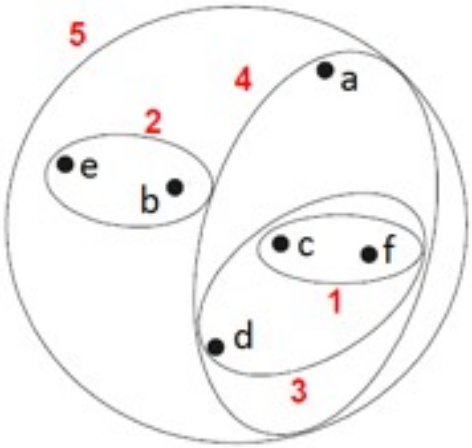


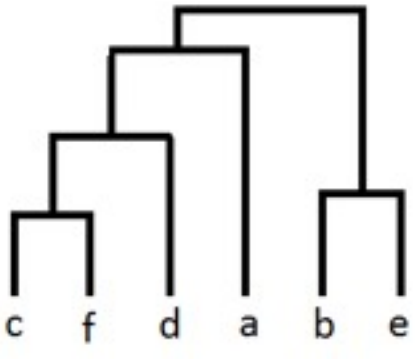
MAX




group average




CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Time and space requirements

- n = number of points
- space: $O(n^2)$
  - it uses a distance matrix n x n
- time: $O(n^3)$
  - there are n steps
  - at each step the distance matrix of size n x n needs to be searched and updated
  - can be reduced to $O(n^2 \log(n))$ in some approaches

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

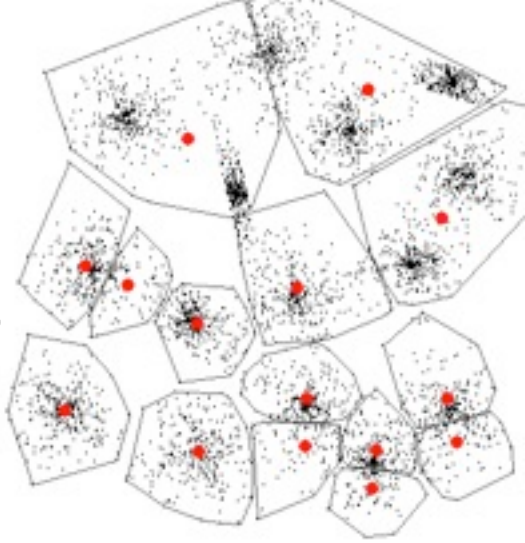
# Limitations

- once a decision has been made to combine two clusters, it cannot be undone
- no objective function is directly optimised
- different schemes have problems with one or more of the following:
  - sensitivity to noise and outliers
  - difficulty handling different sized clusters and convex shapes
  - breaking large clusters

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# K-means clustering

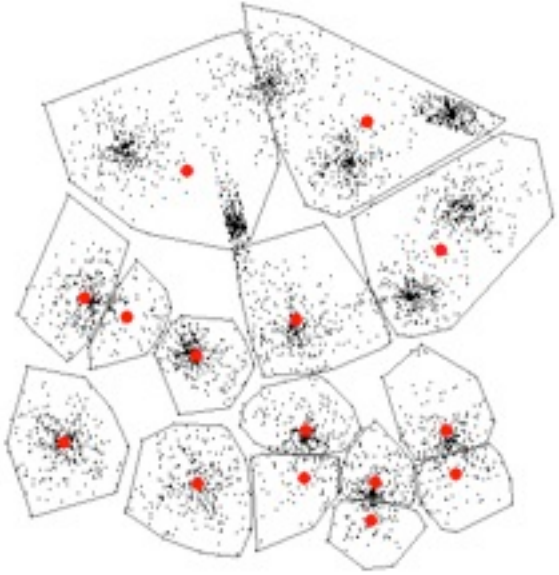
# Non-hierarchical clustering



- non-hierarchical clustering attempts to directly partition the dataset into a set of disjoint clusters
- k-means clustering aims to partition dataset into k clusters in which each instance belongs to the cluster with the nearest mean
- basic principles:
  - each cluster is associated with its centre (centroid)
  - each point is assigned to the cluster with the closest centroid

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# K-means algorithm

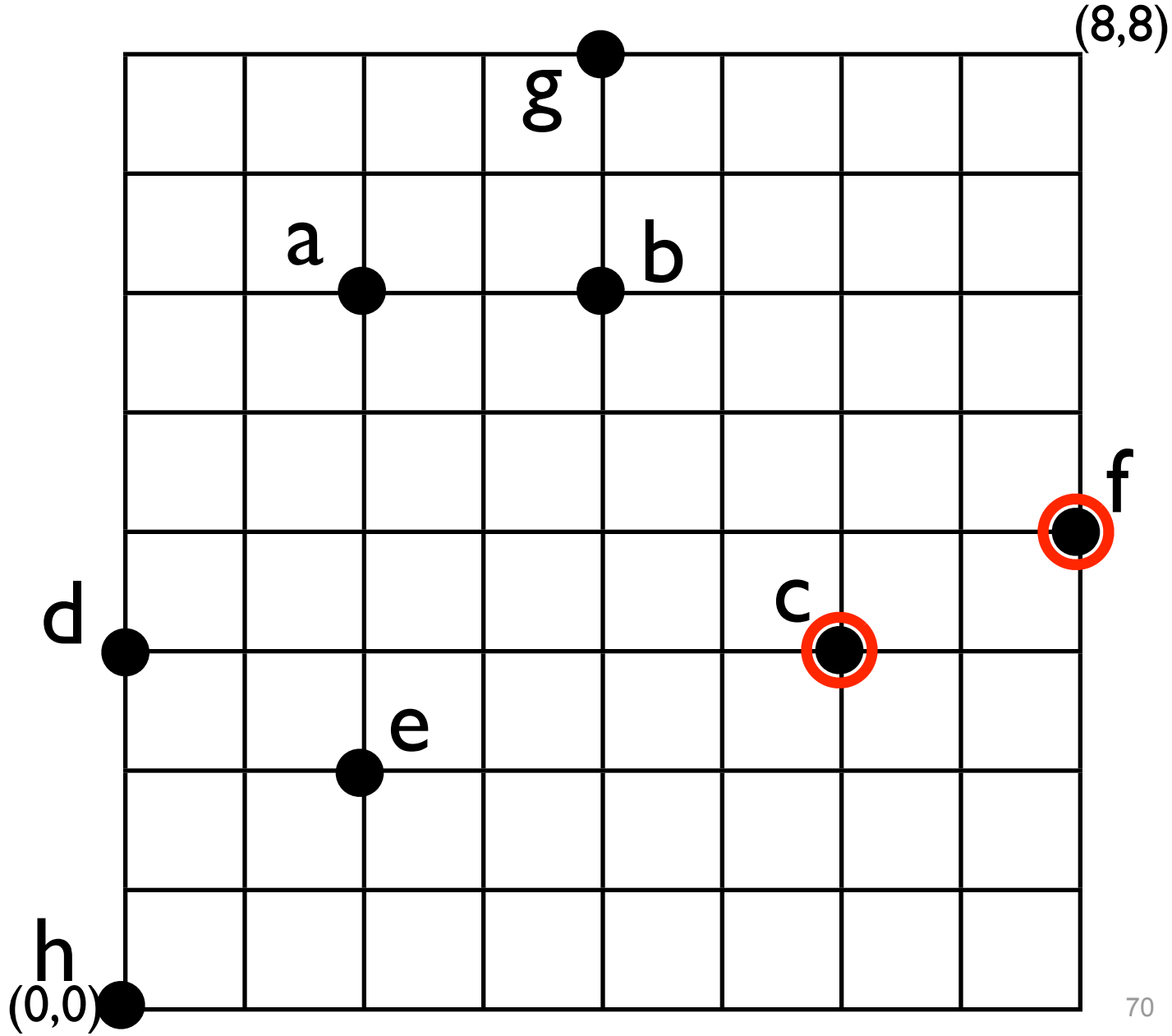


1. Select k points as the initial centroids.
2. **repeat**
3. Form k clusters by assigning all points to the closest centroid.
4. Recompute the centroid for each cluster.
5. **until** The centroids don't change.

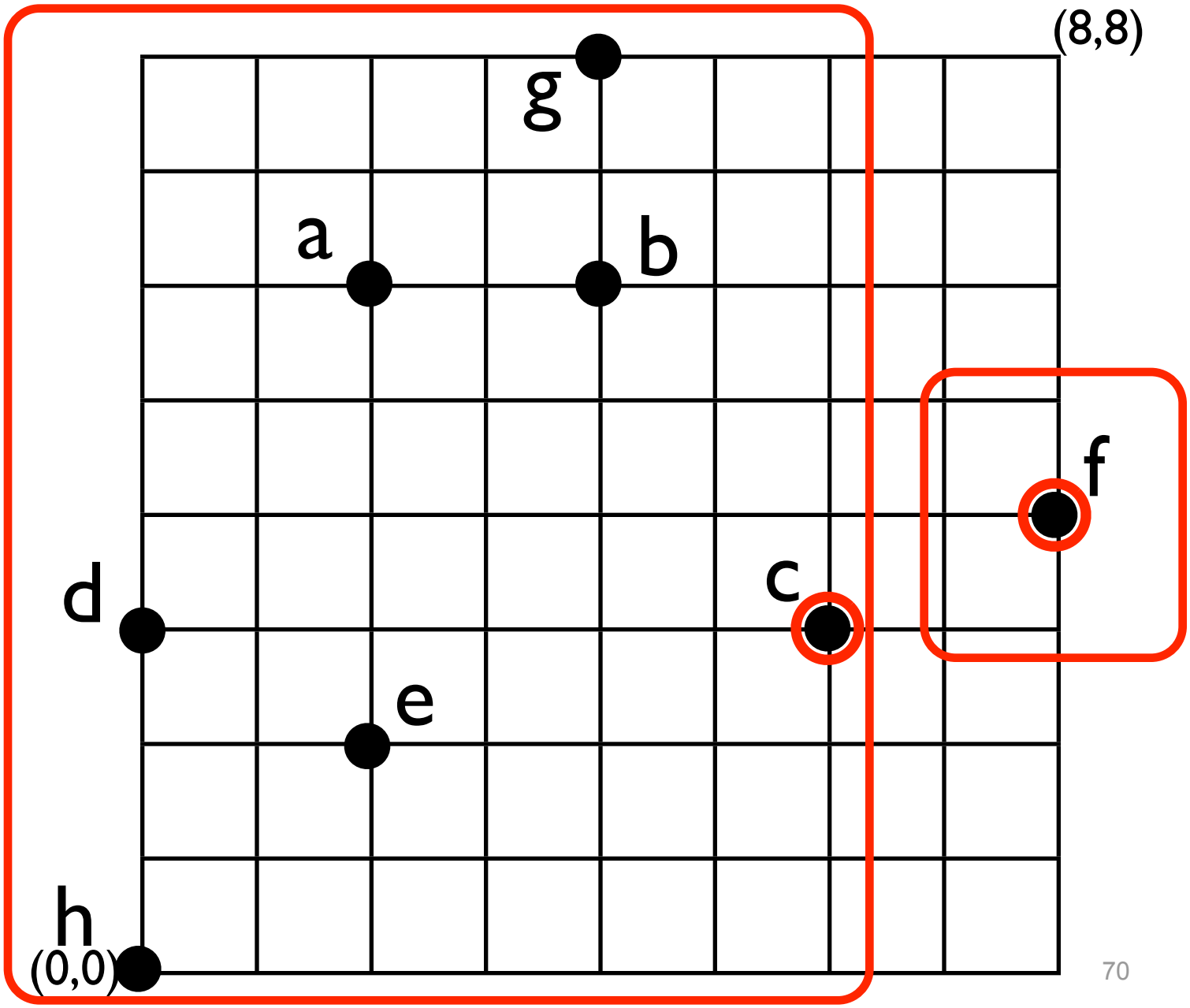
- initial centroids are often chosen randomly
- stopping condition is often changed to "until relatively few points change clusters"

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Example: k=2

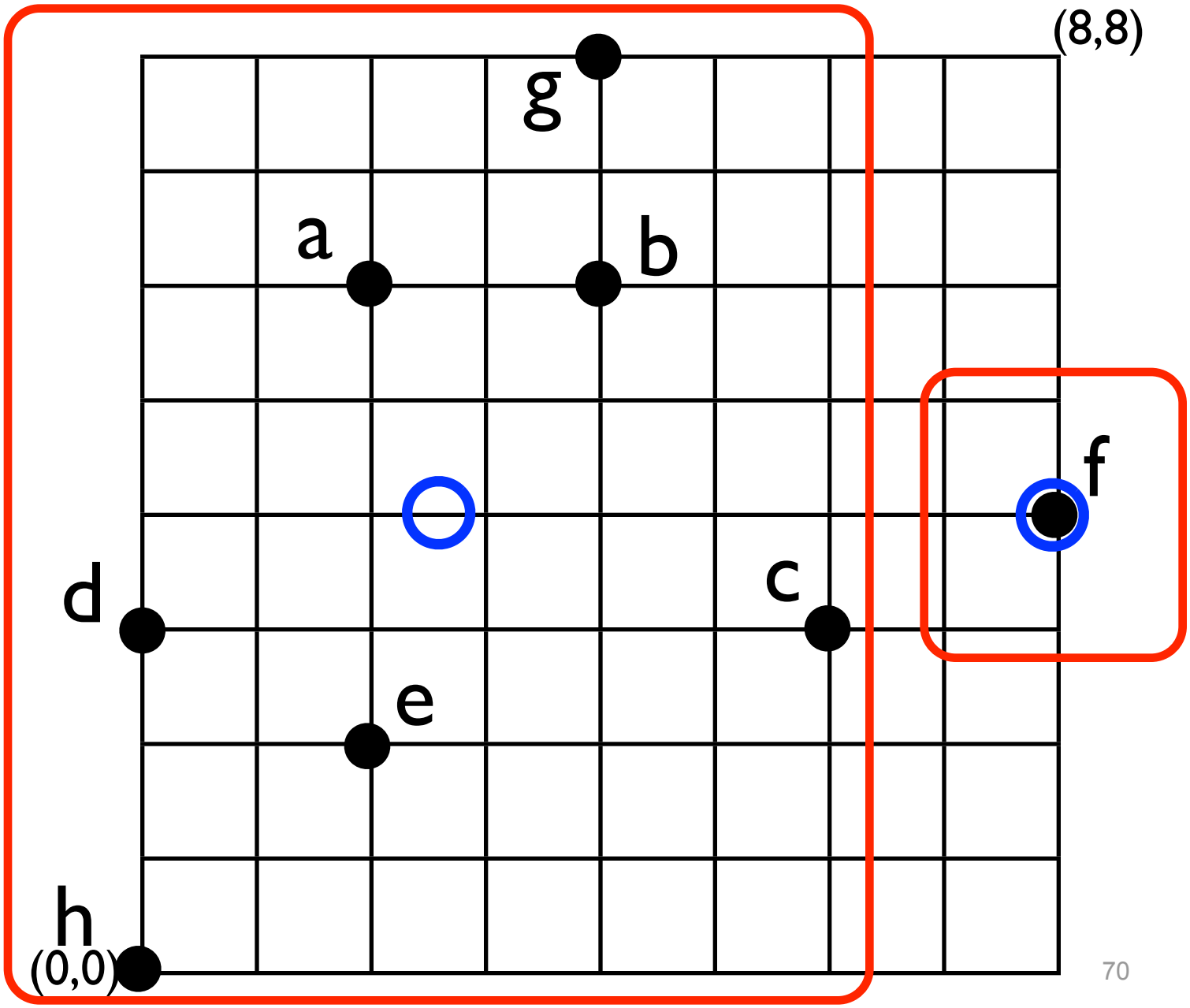


# Example: k=2



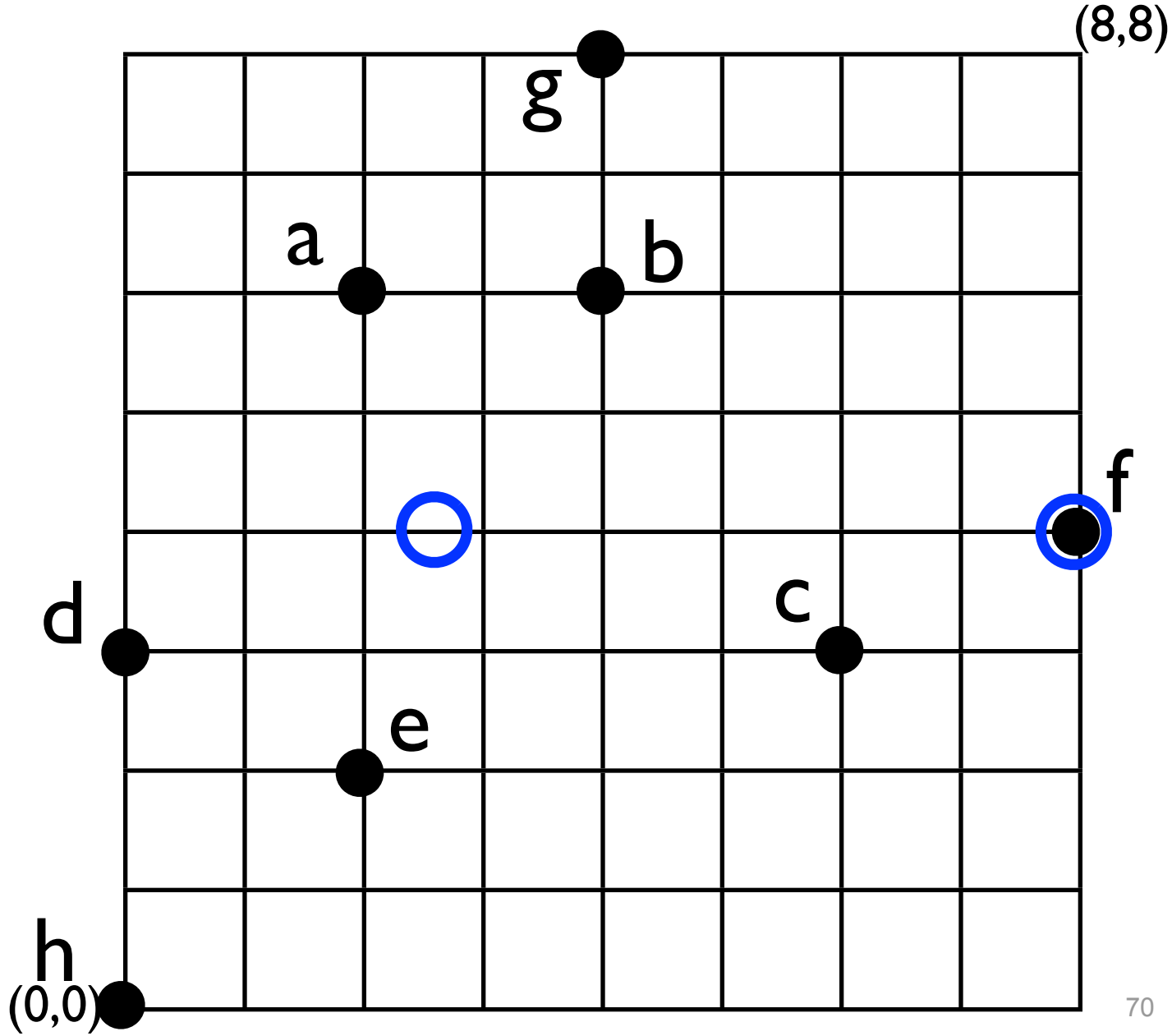
# Example: k=2

(2.57,4)
(8,4)

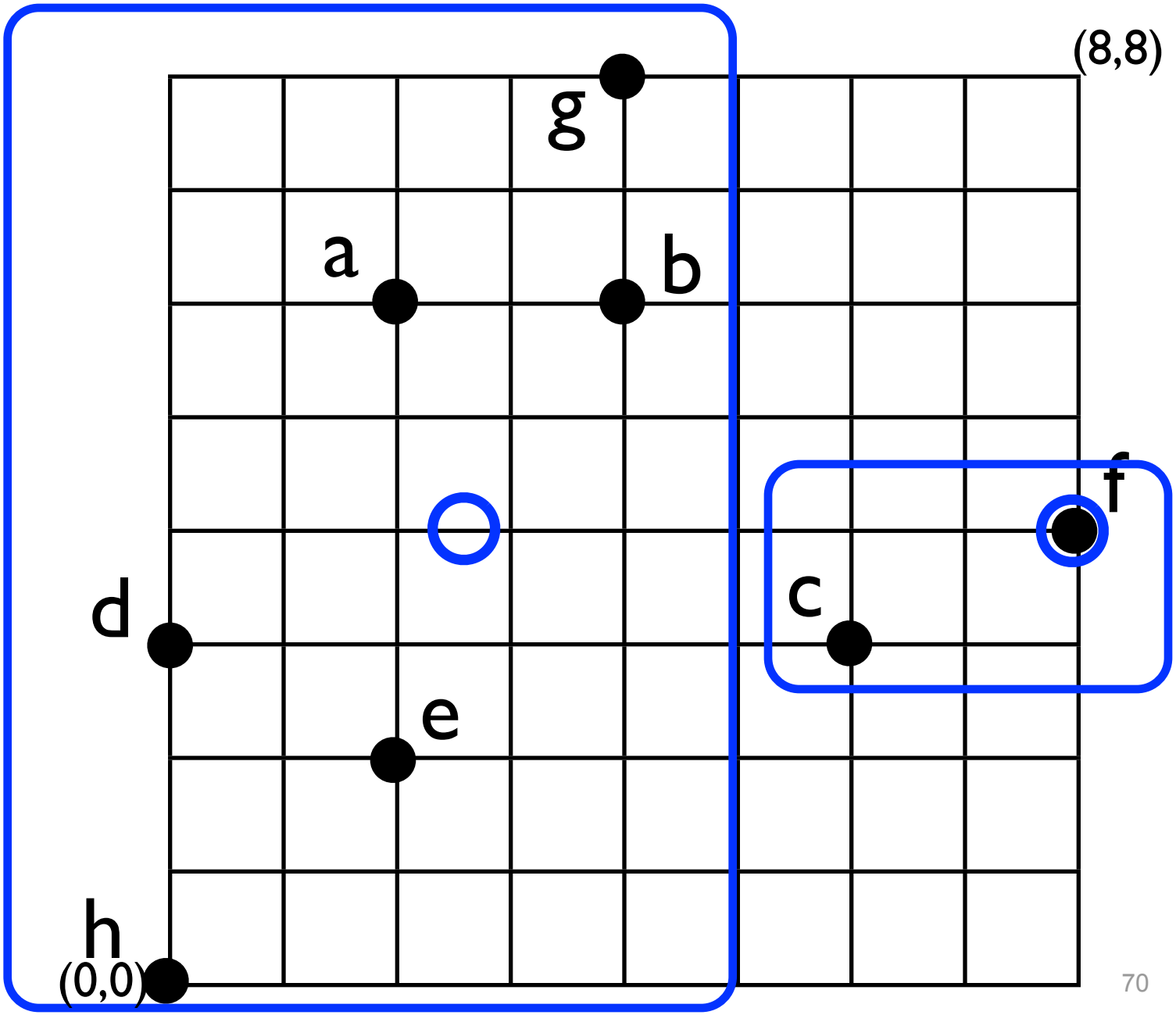


# Example: k=2

(2.57,4)
(8,4)

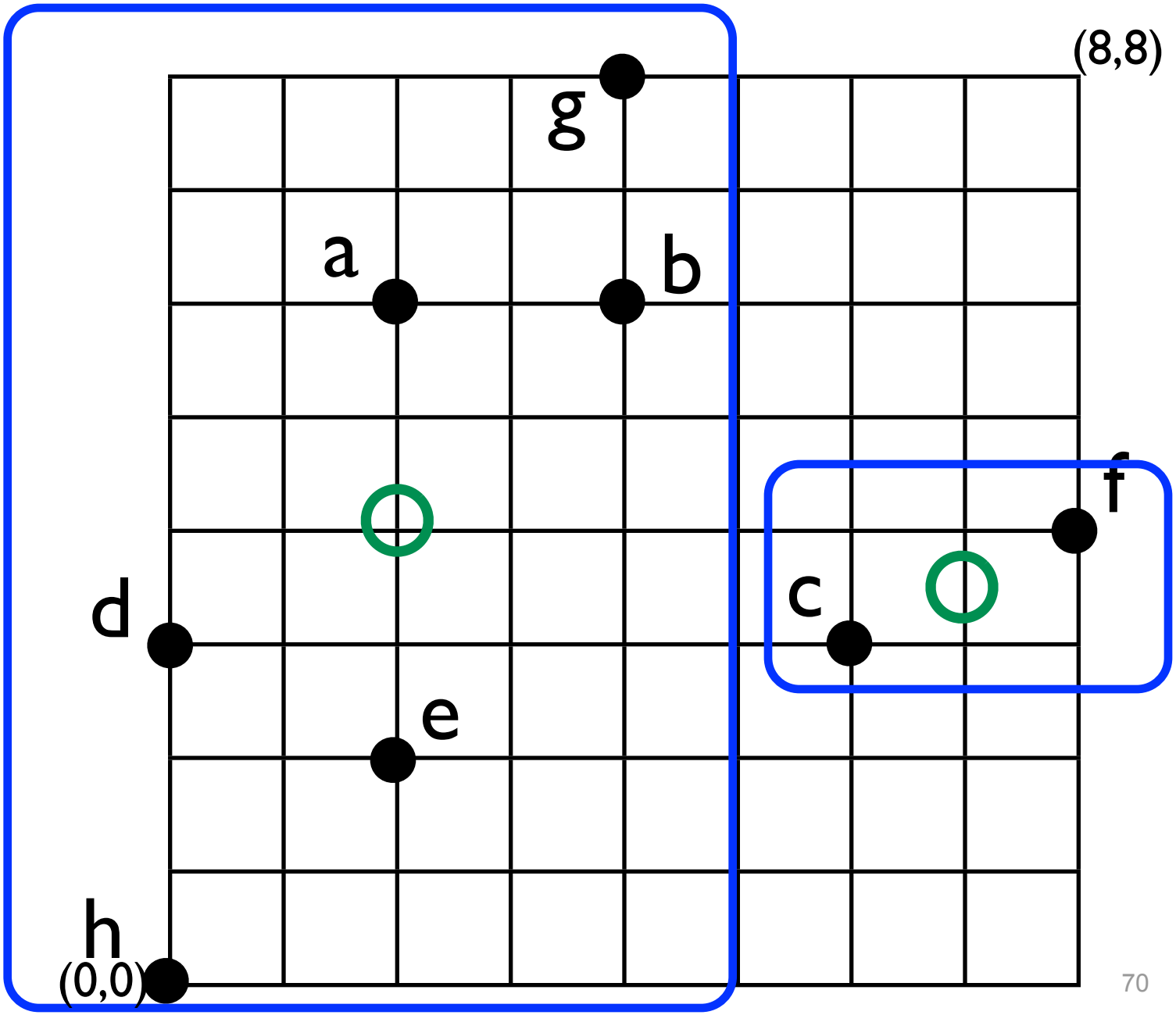


# Example: k=2



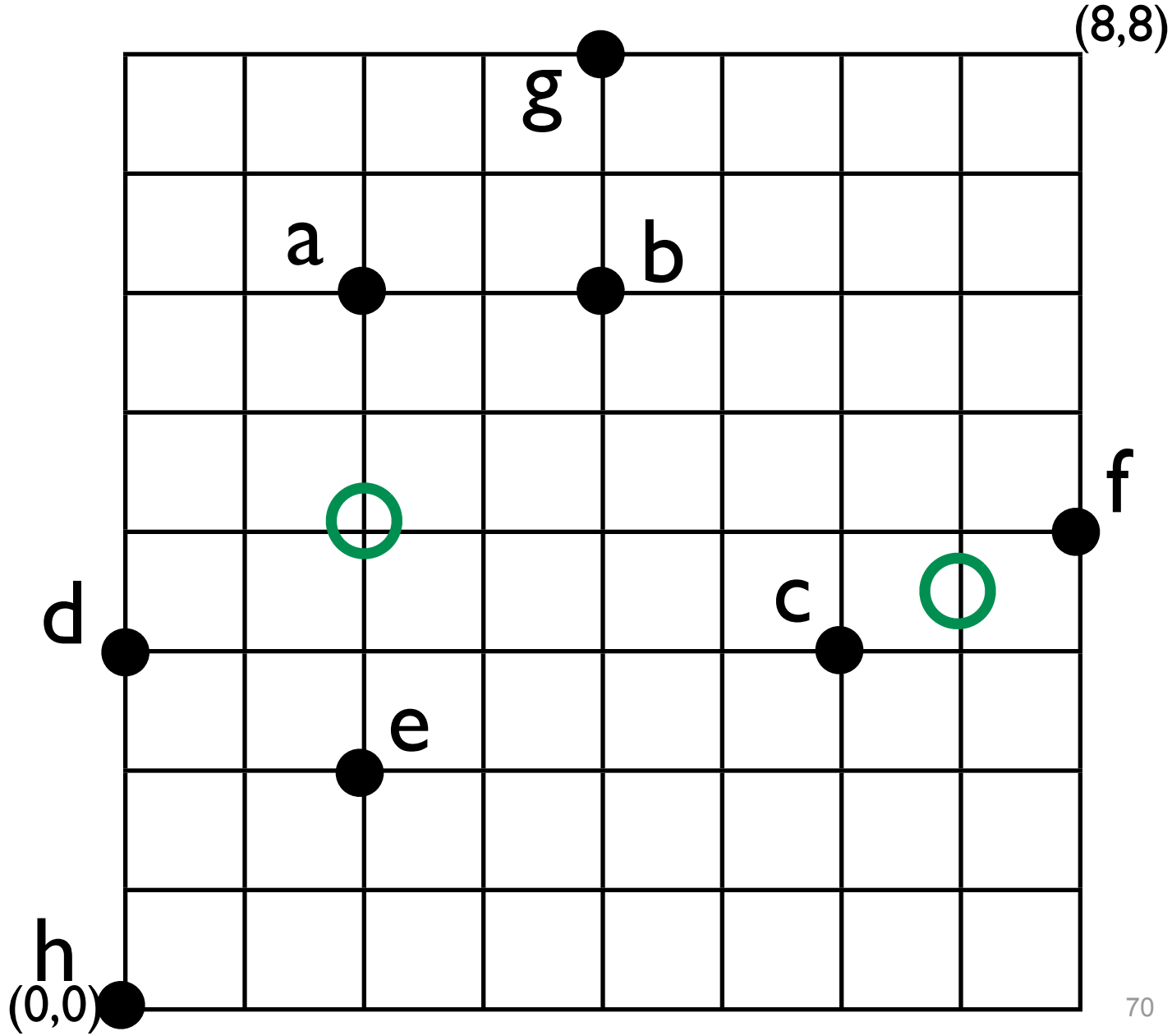

(2.57,4)
(8,4)

# Example: k=2

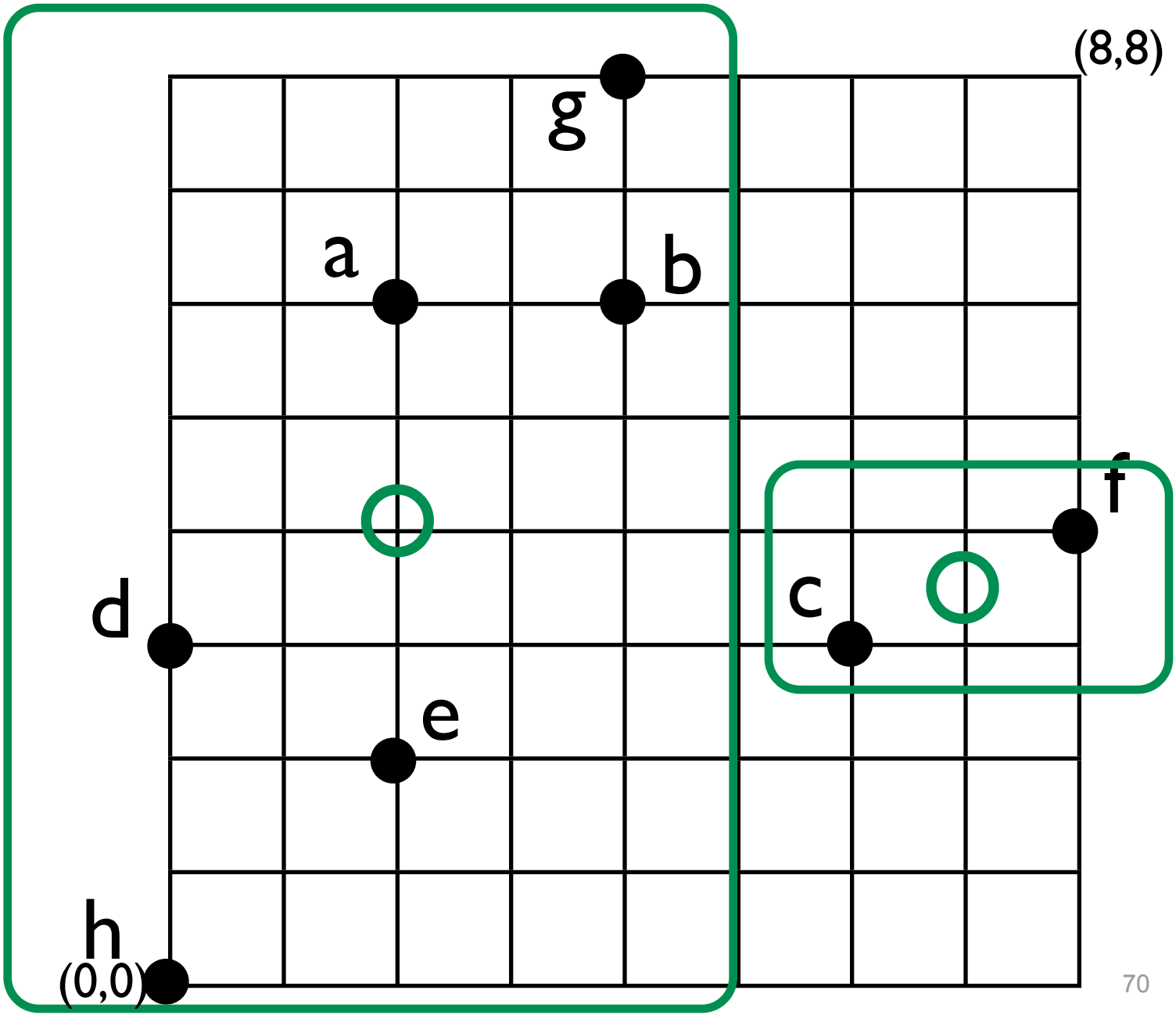


# Example: k=2

(2,4.17)
(7,3.5)



# Example: k=2



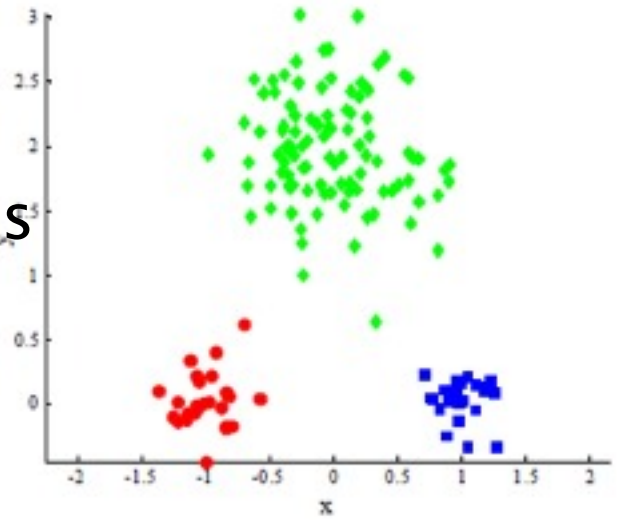
# K-means clustering

- works well when a given set is composed of a number of distinct classes
- the number of clusters k **must** be specified
- problem: How to choose k objectively?
- K-means will converge for most common similarity measures in the first few iterations
- less computationally intensive than hierarchical methods: $O(n \cdot k \cdot i \cdot a)$, where n, k, i and a are the numbers of points, clusters, iterations and attributes
- modest space requirements because only points and centroids are stored: $O((n + k) \cdot a)$

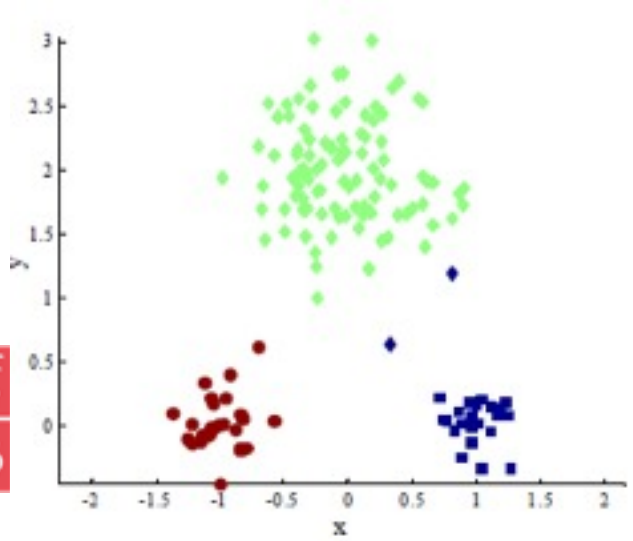
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Two different k-means clusterings

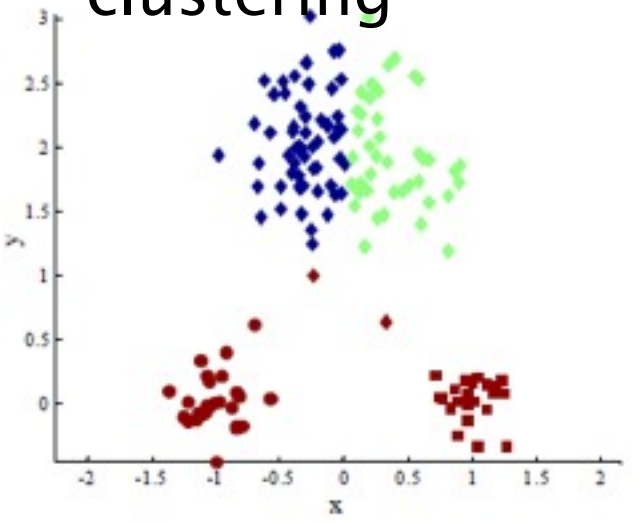
original points




optimal clustering

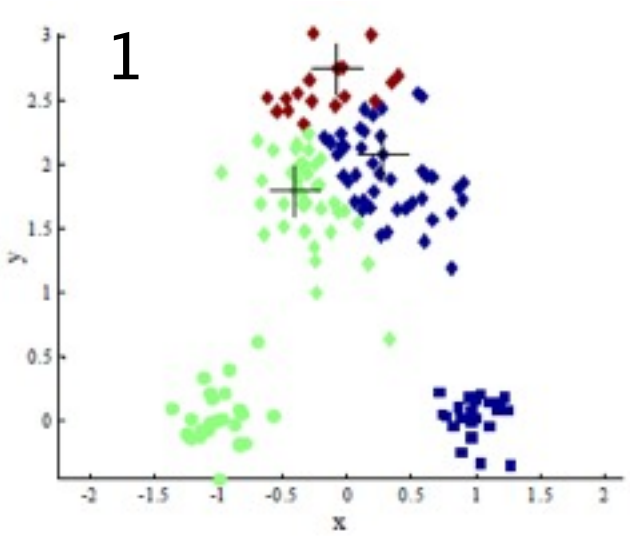



sub-optimal clustering



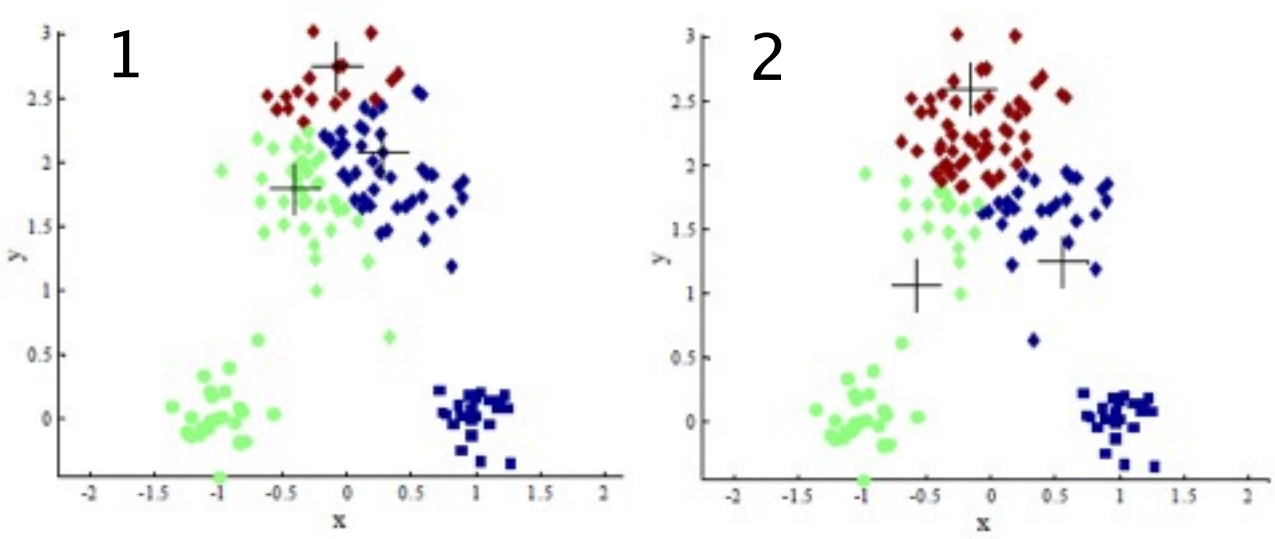

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Importance of choosing initial centroids

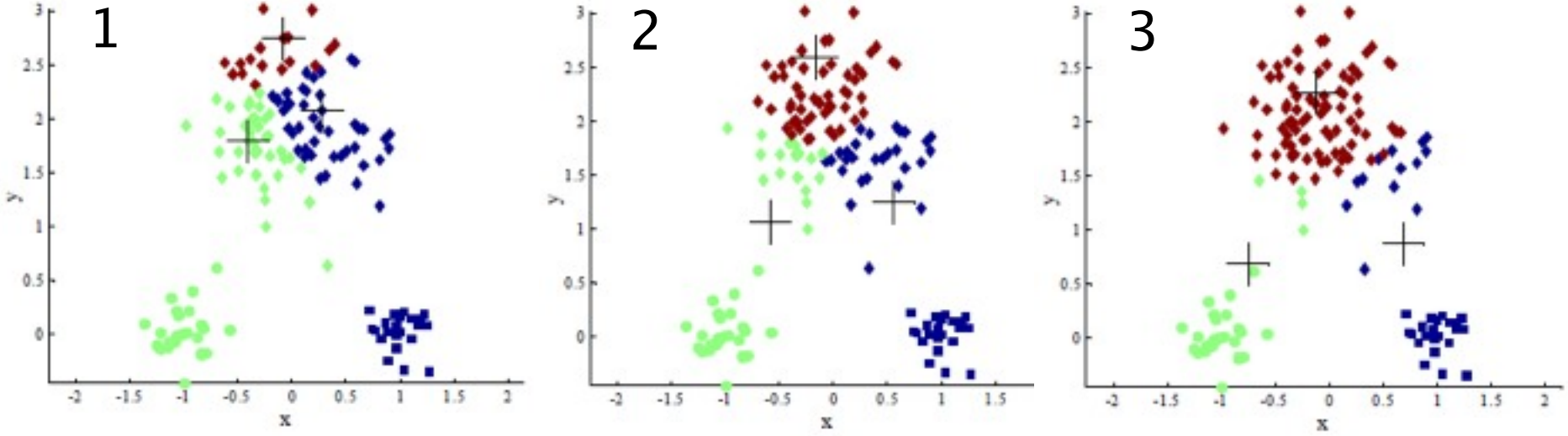



CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

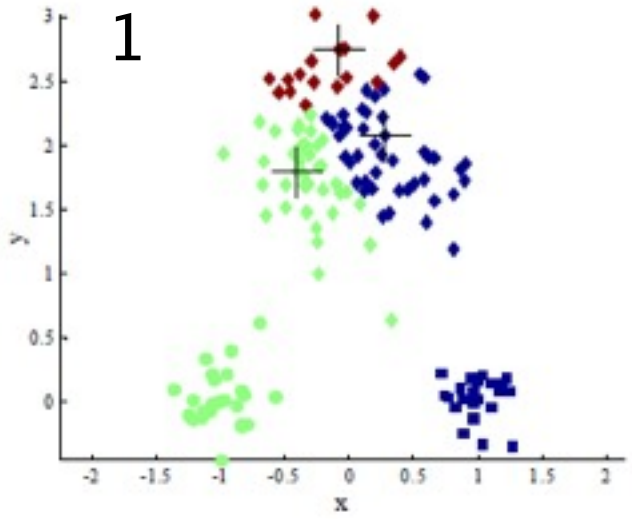
# Importance of choosing initial centroids

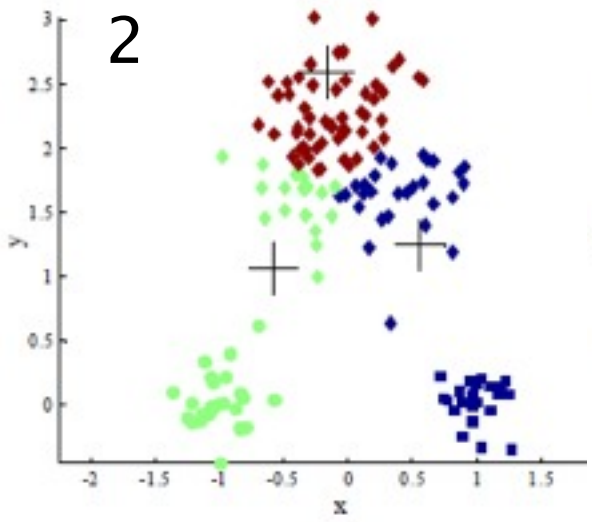


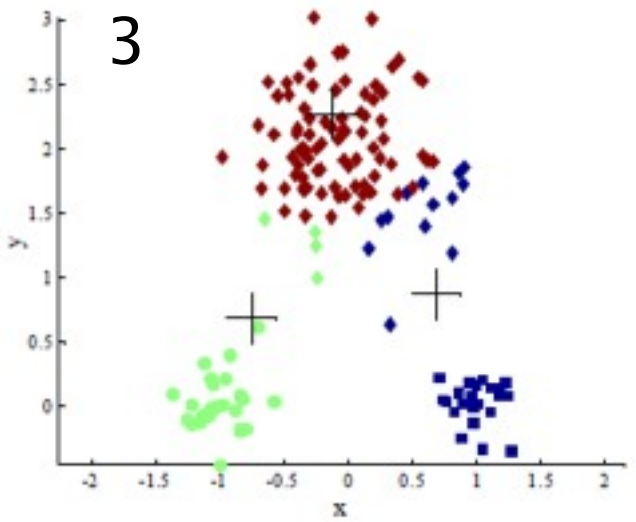
# Importance of choosing initial centroids

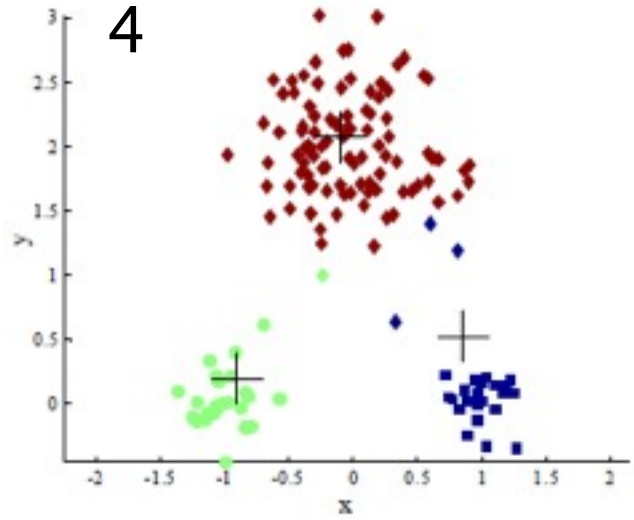


# Importance of choosing initial centroids

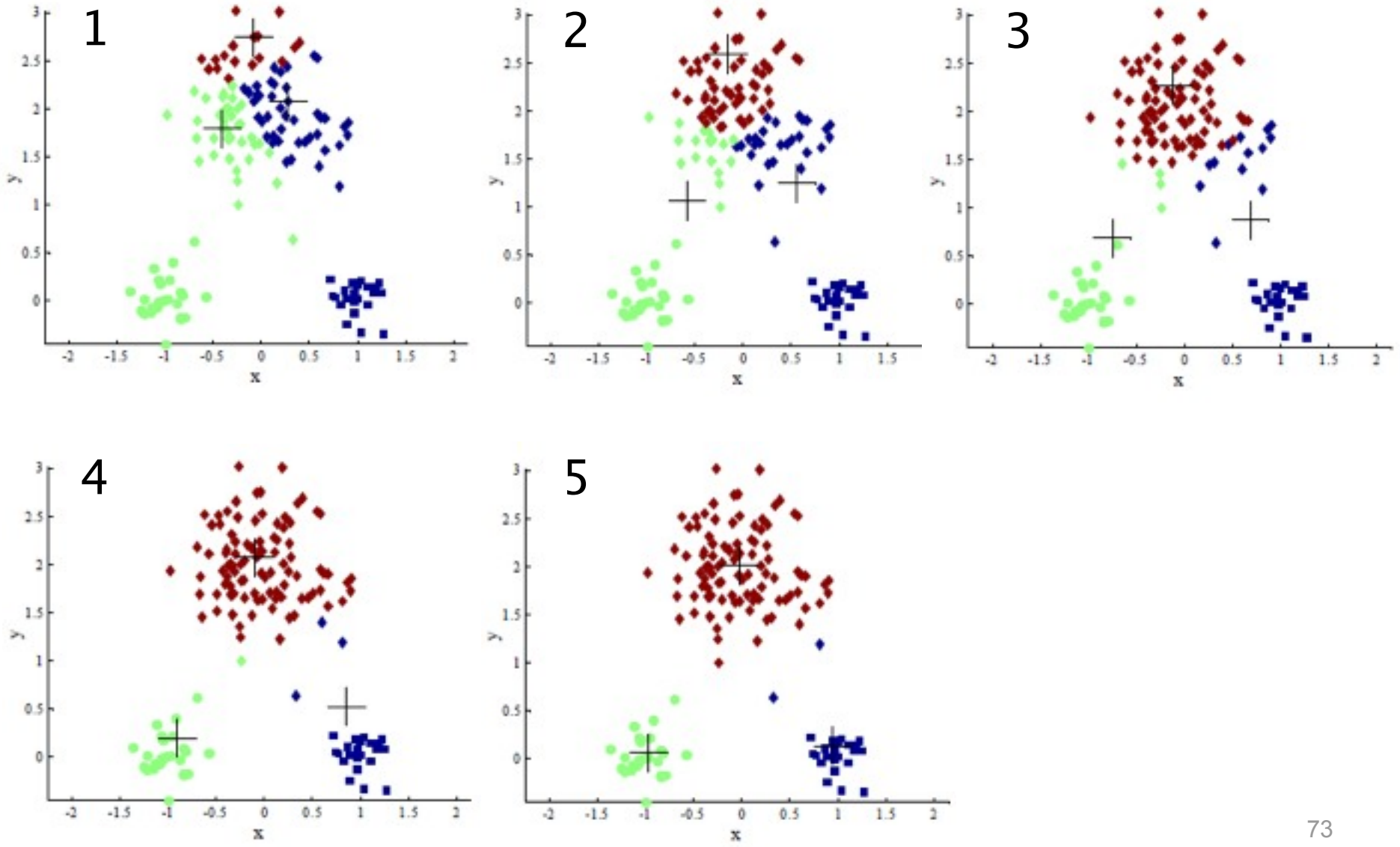


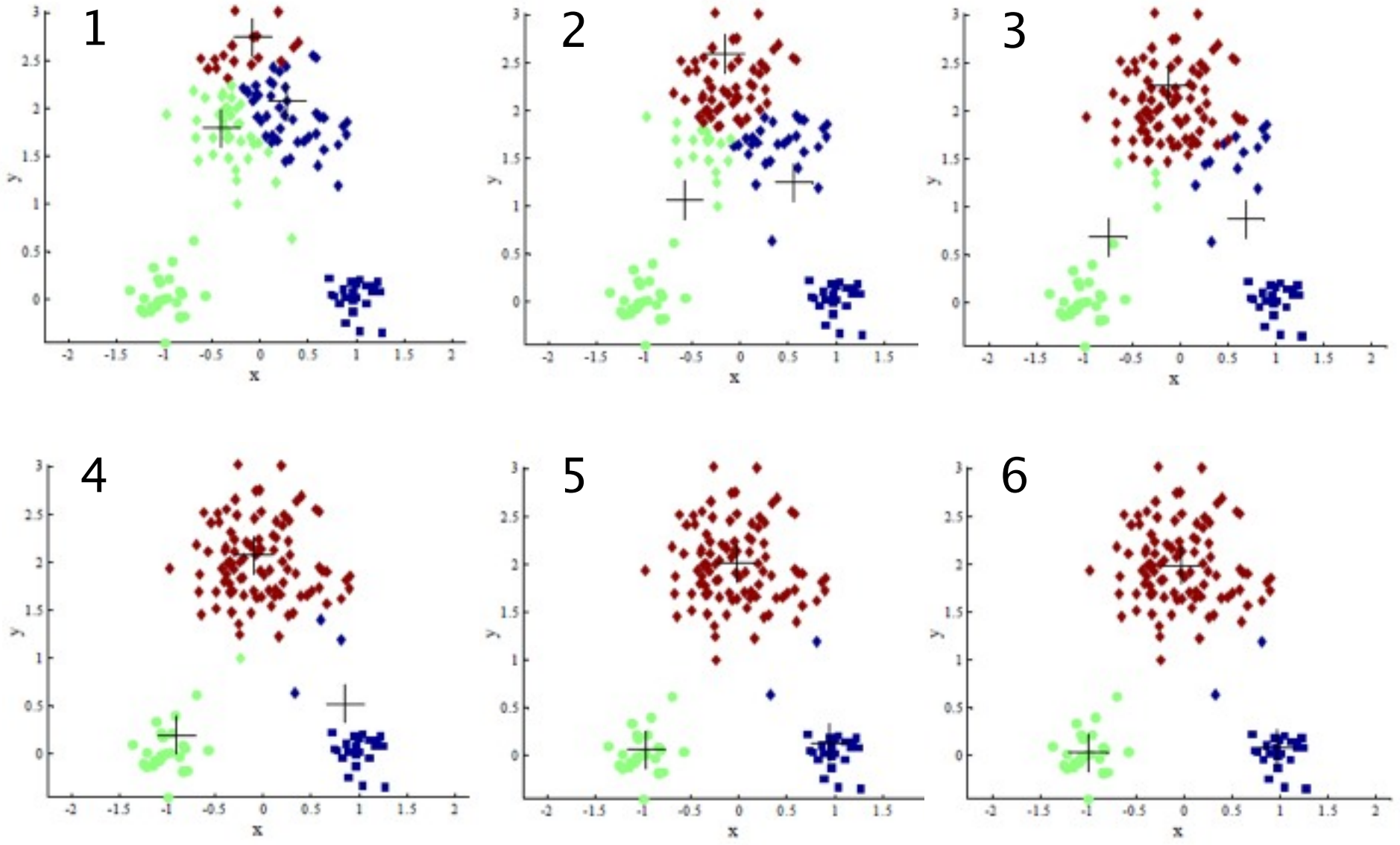



# Importance of choosing initial centroids



# Importance of choosing initial centroids



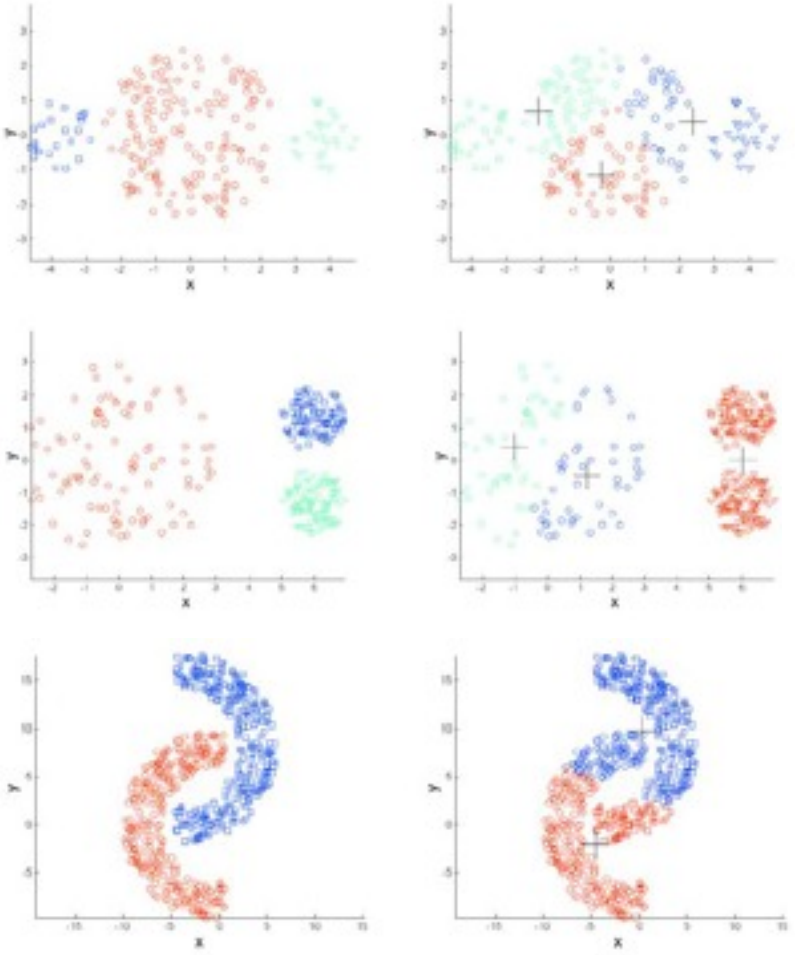
# Problems with choosing initial centroids

- if there are k "real" clusters then the chance of selecting one centroid from each cluster is small
- ... very small when k is large!
- sometimes the initial centroids will readjust themselves in the "right" way and sometimes they will not
- possible "solutions":
  - multiple runs may help
  - sample and use hierarchical clustering to determine initial centroids
  - select most widely separated
  - ...

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Limitations of k-means

- has problems when clusters are of differing:
  - sizes
  - densities
  - non-spherical shapes



# Evaluation

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Evaluation

- evaluating clustering algorithms is complex because it is difficult to find an objective measure of quality of clusters
- typical objective functions in clustering formalise the goal of attaining:
  1. cohesion – high intra-cluster similarity (instances within a cluster are similar)
  2. separation – low inter-cluster similarity (instances from different clusters are dissimilar)

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

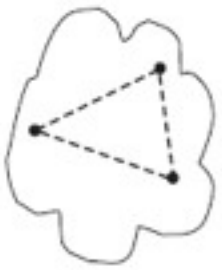
# Graph-based view of cohesion and separation

- mathematically, cohesion and separation for a graph-based cluster can be expressed as follows:

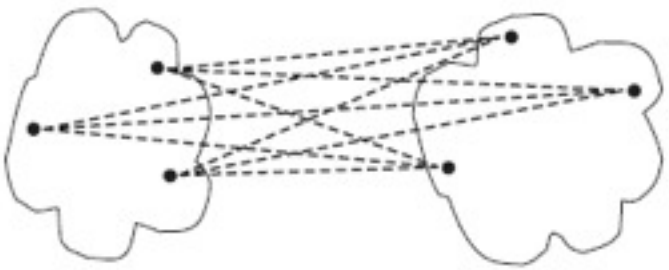
$$cohesion(C_i) = \sum_{\substack{\mathbf{x} \in C_i \\ \mathbf{y} \in C_i}} proximity(\mathbf{x}, \mathbf{y})$$

$$separation(C_i, C_j) = \sum_{\substack{\mathbf{x} \in C_i \\ \mathbf{y} \in C_j}} proximity(\mathbf{x}, \mathbf{y})$$

cohesion



separation



# Prototype-based view of cohesion and separation

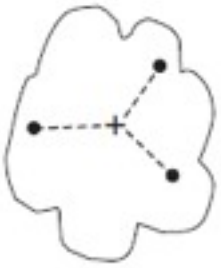
- mathematically, cohesion and separation for a prototype-based cluster can be expressed as follows:

$$cohesion(C_i) = \sum_{\mathbf{x} \in C_i} proximity(\mathbf{x}, \mathbf{c}_i)$$

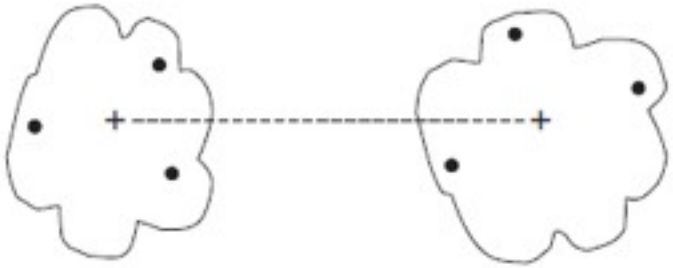
$$separation(C_i, C_j) = proximity(\mathbf{c}_i, \mathbf{c}_j)$$

centroids

cohesion



separation



# Cluster validity

- previous definitions of cohesion and separation give us some simple and well-defined measures of cluster validity
- they can be combined into an overall validity measure by using a weighted sum:

$$overall\ validity = \sum_{i=1}^{K} w_i\ validity(C_i)$$

- the weights will vary depending on the cluster validity measure
- in some cases, the weights are simply 1 or the cluster sizes

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# Summary

- Clustering to analyze data
- Different types of clusters
- Hierarchical clustering
- k-means

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD