Unsupervised Learning

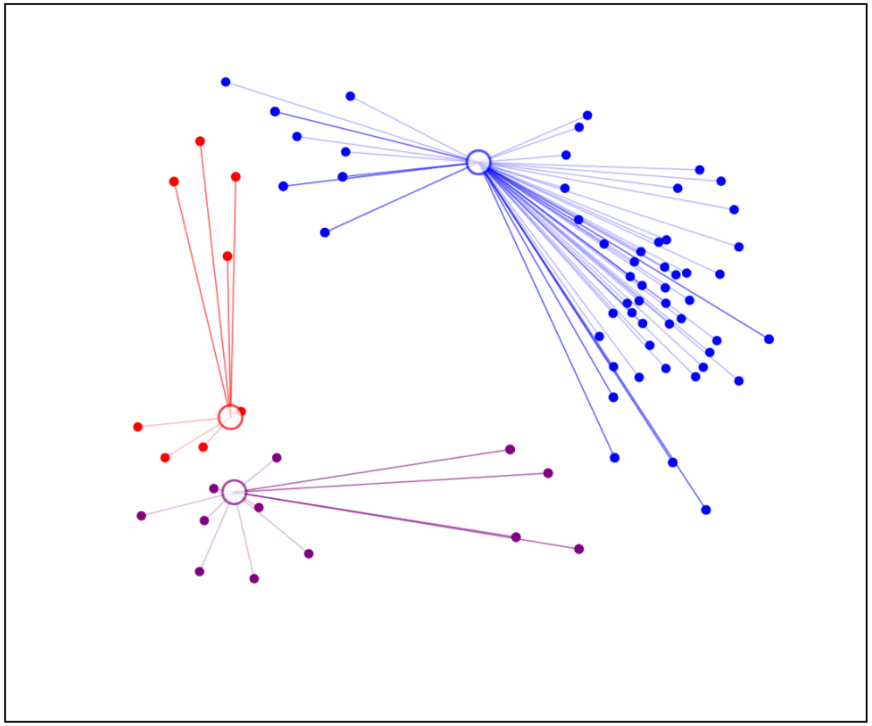


Image source: deepclimate.org

Brunel University London

Learning

- Remember the early chatbots?
- No ability to learn / adapt
- Learning is Key to AI
- Appears in different forms

Brunel University London

Learning

- Different Classes of Learning
- Unsupervised learning: learning without the desired output (‘teacher’ signals).
- Supervised learning: learning with the desired output.
- Reinforcement Learning – reward / punishment signals

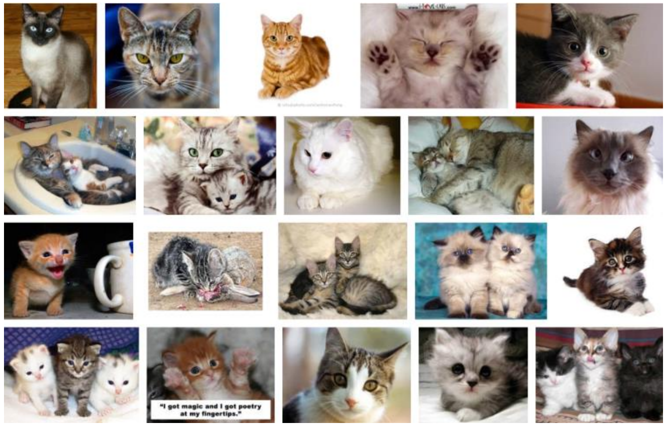
Learning

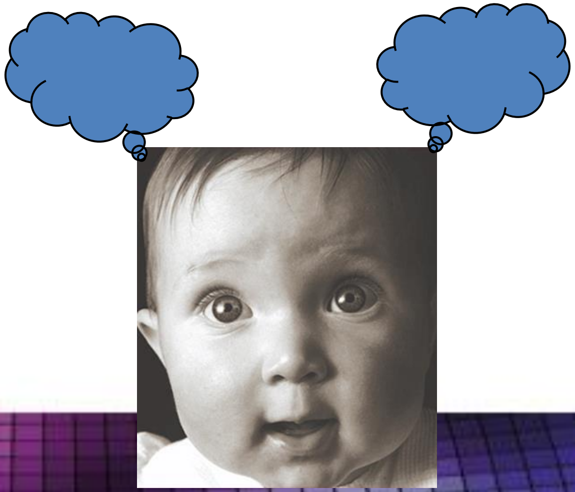
- Clustering is one of the widely-used unsupervised learning methods.
- Other Unsupervised learning:
  - Dimensionality reduction (e.g. PCA)
  - Association Rules / Recommender Systems
- Supervised learning:
  - Classification (see next lecture)
  - Regression
- Reinforcement Learning commonly seen in Neural network training (see in 2 weeks)

Brunel University London

Clustering – in everyday learning

For example:



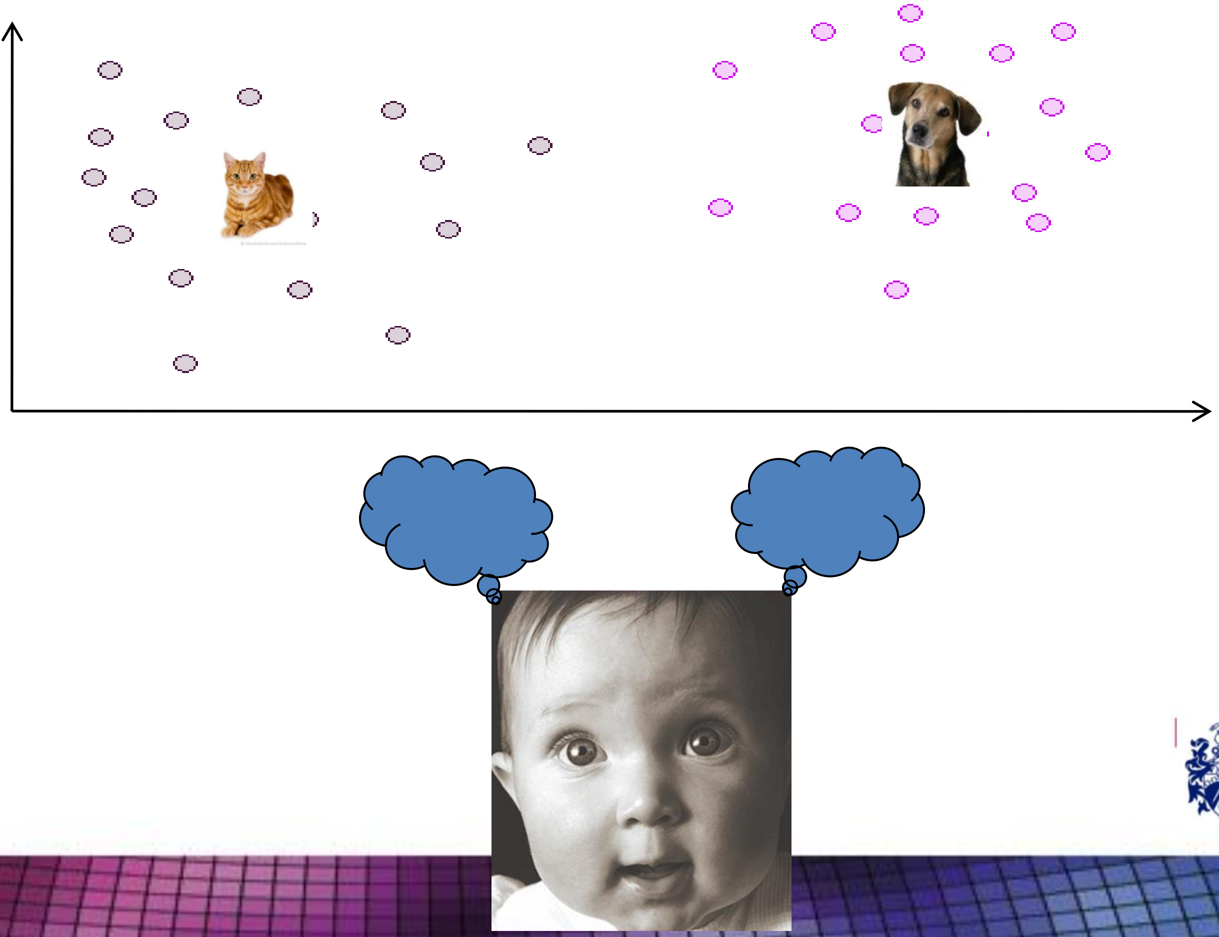


Amazingbaby.net

Brunel University London

Clustering – more formally

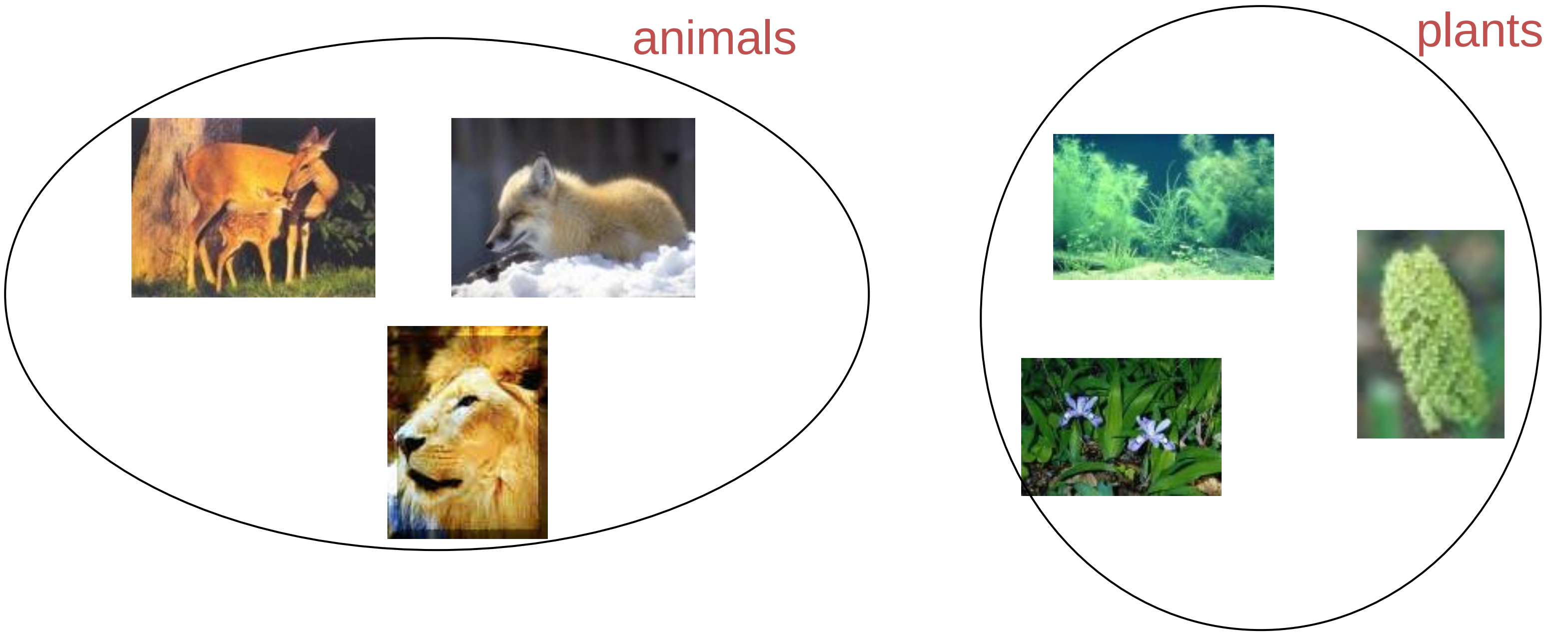
For example:



Amazingbaby.net

Clustering

Already defined in biology…



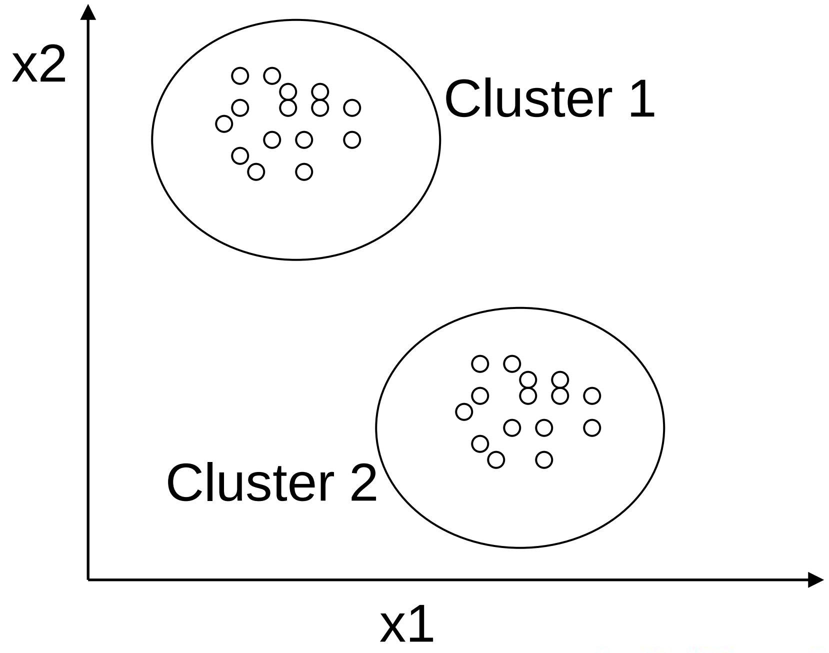
Things that are brown and run away

Things that are green and don't run away

Clustering

Clustering: to partition a data set into subsets (clusters), so that the data in each subset share some common trait - often similarity or proximity for some defined distance measure.

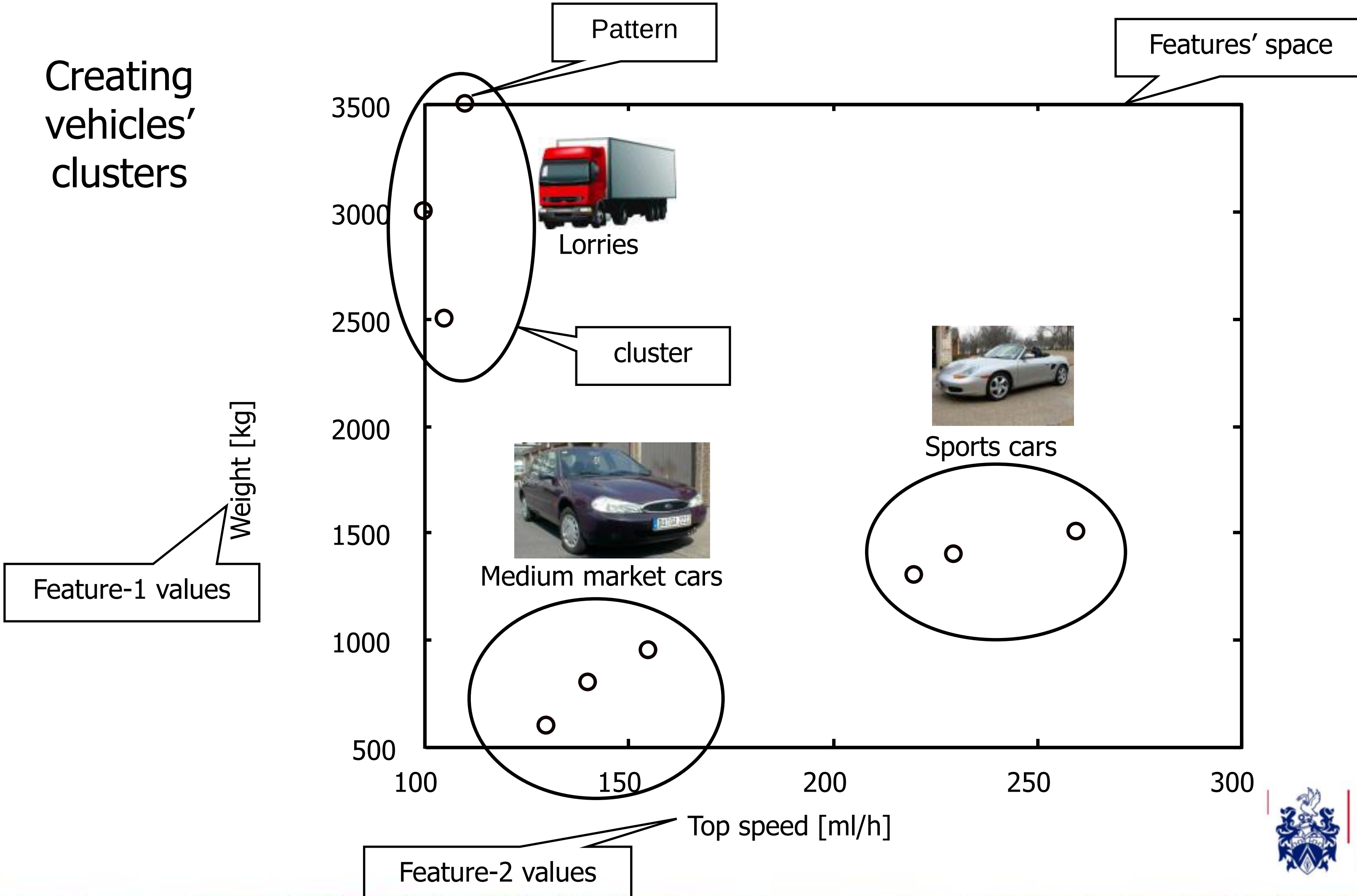
- The process of organizing objects into groups whose members are similar in some way.
- A *cluster* is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.
- Unsupervised: No need for the ‘teacher’ signals, i.e. the desired output.



Brunel University London

Patterns, Clusters and Features (2)

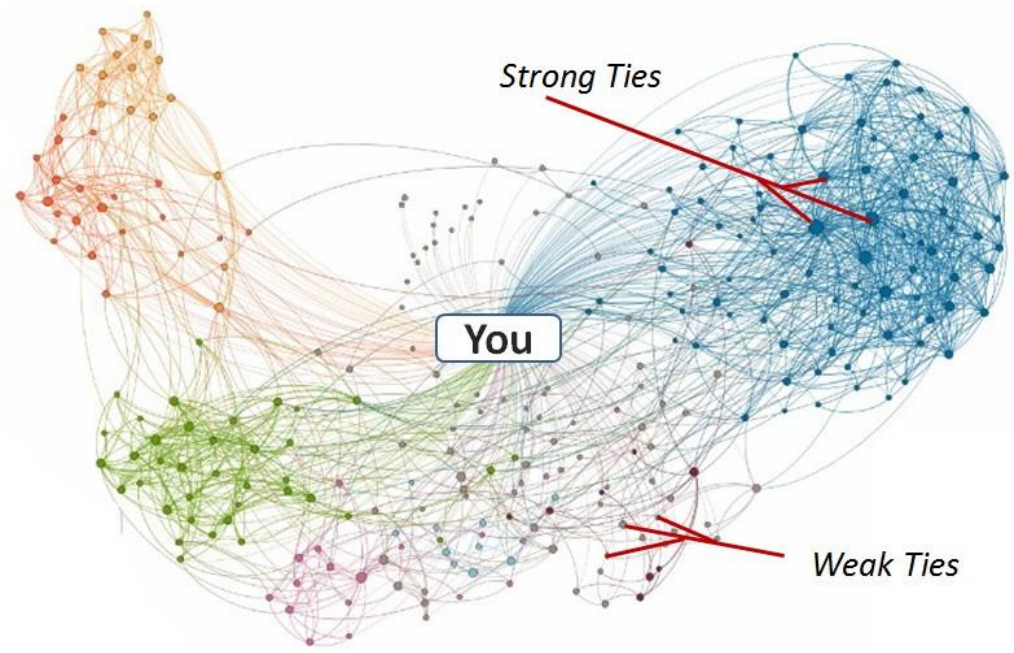
Creating vehicles' clusters



Brunel University London

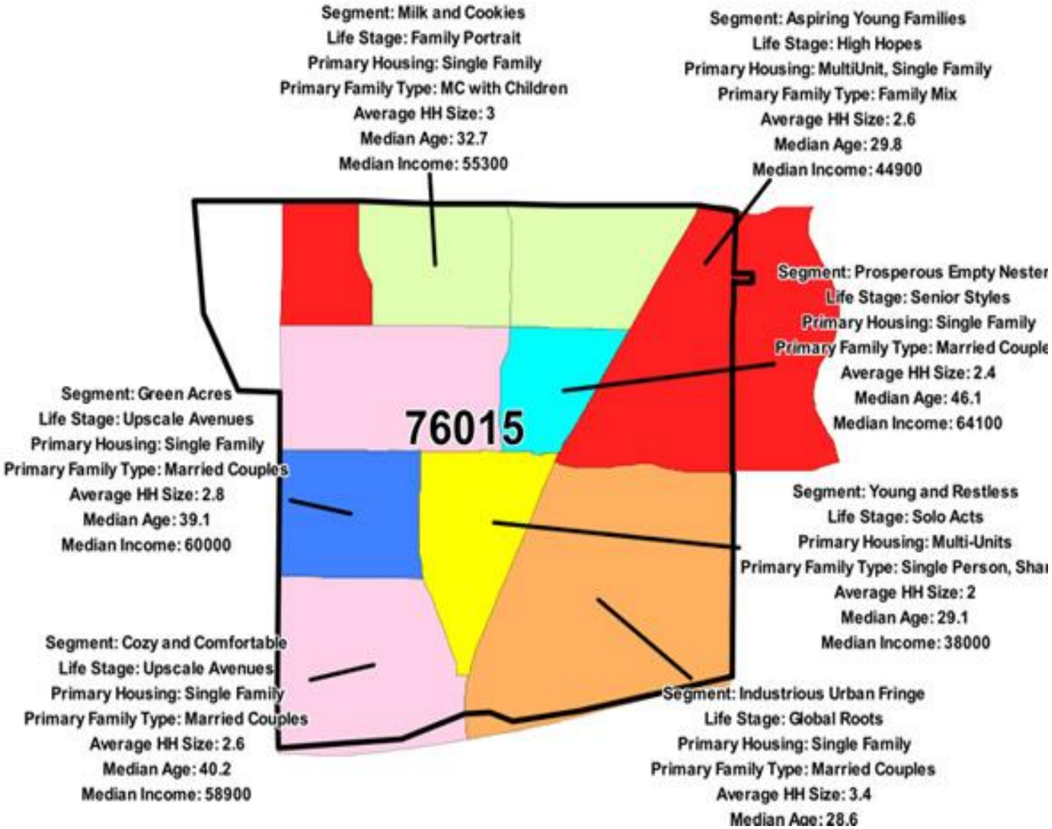
Uses: Social networks

- Marketing
- Terror networks
- Allocation of resources in a company / university



Brunel University London

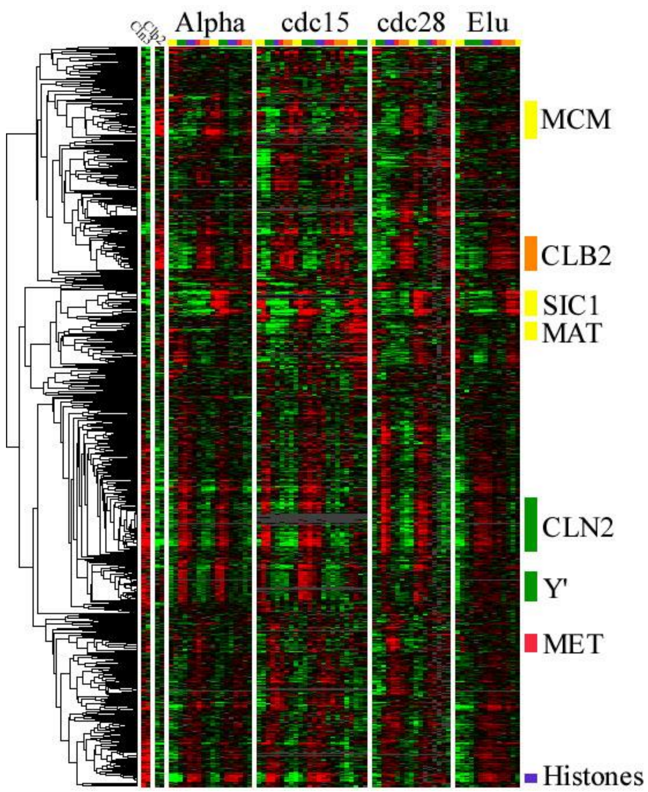
Uses: Customer Segmentation

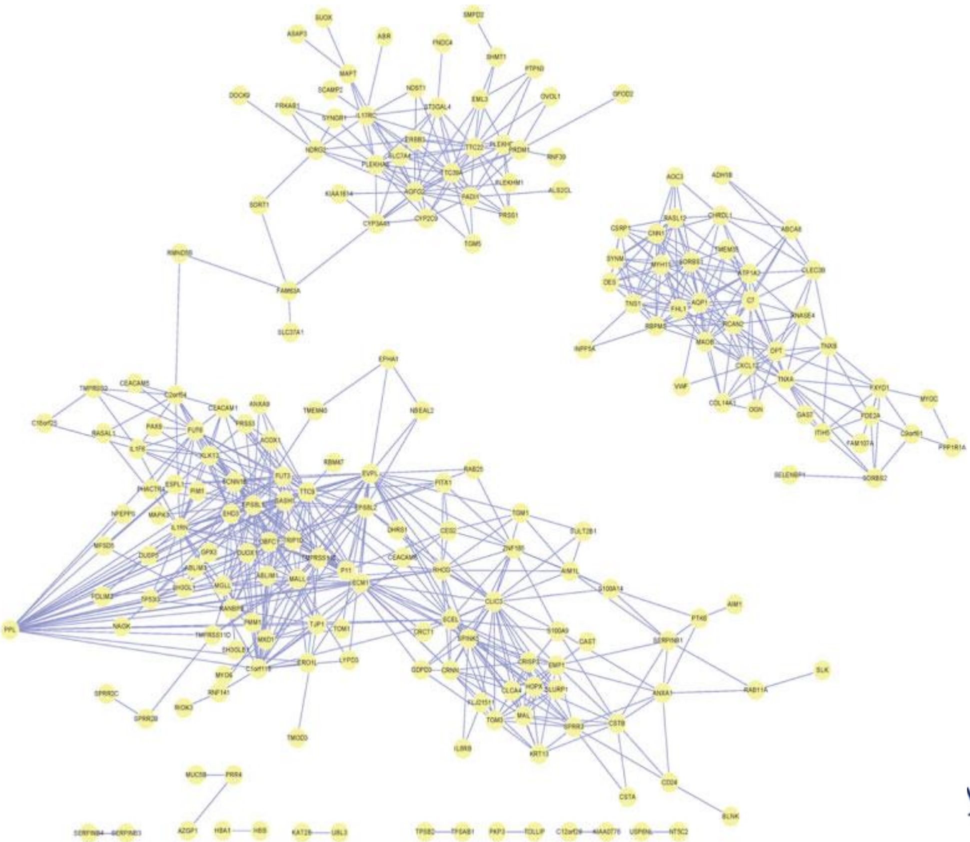


Brunel University London

Uses: Gene networks

- Understanding gene interactions
- Identifying important genes linked to disease



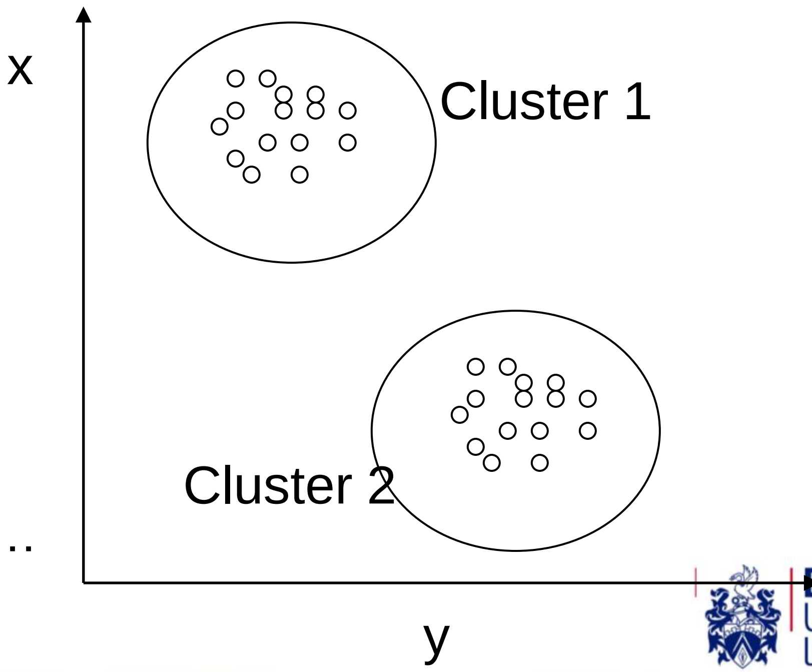


How to do clustering?

Given some data x (e.g. with two features):

e.g.

x	*y*
1.34	0.42
0.23	1.59
…	…
3.21	1.12



But can generalise to more features …

Pattern Similarity

- Clusters are formed by similar patterns.
- In computer science, we need to define some metric to measure similarity. One of the commonly adopted similarity metrics is distance.
- Works on any number of features in a dataset (here *N*):

Euclidean: $d(x,y) = \sqrt{}((x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_N - y_N)^2)$

Manhattan: $d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_N - y_N|$

The shorter the distance, the more similar the two patterns.

Brunel University London

Euclidean / Manhattan Distance

x	5.5	2.9	4.8	6.7	0.6
y	0.2	1.0	4.8	3.8	9.2

- Euclidean d(x,y)=

$\sqrt{((5.5-0.2)^2 + (2.9-1.0)^2 + (4.8-4.8)^2 + (6.7-3.8)^2 + (0.6-9.2)^2)}$

$= \sqrt{((5.3)^2 + (1.9)^2 + (0.0)^2 + (2.9)^2 + (-8.6)^2)}$

$= \sqrt{(28.09 + 3.61 + 0.0 + 8.41 + 73.96)}$

$= \sqrt{(114.07)} = 10.68$

Euclidean / Manhattan Distance

x	5.5	2.9	4.8	6.7	0.6
y	0.2	1.0	4.8	3.8	9.2

- Manhattan d(x,y)=

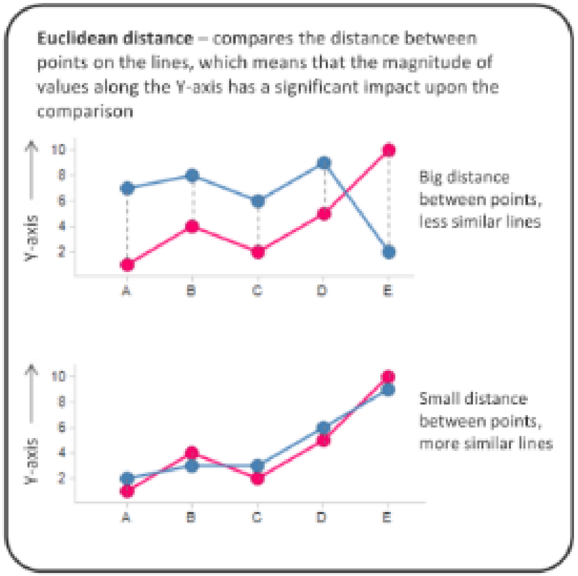
  $(|5.5-0.2| + |2.9-1.0| + |4.8-4.8| + |6.7-3.8| + |0.6-9.2|)$

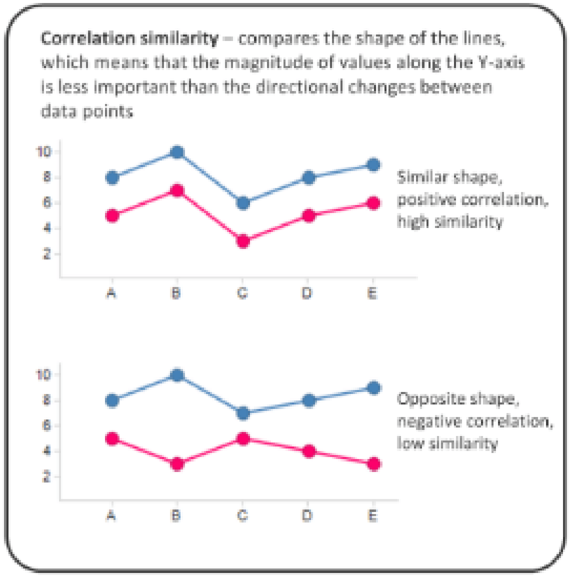
  $= 5.3 + 1.9 + 0.0 + 2.9 + 8.6$

  $= 18.7$

Pattern Similarity & Distance Metrics

- Distance Metrics
  - Euclidean
  - Correlation
  - Minkowski
  - Manhattan
  - Mahalanobis





- How to choose?
  - Often Application Dependant
  - What is important? Shape / Distance / Scale

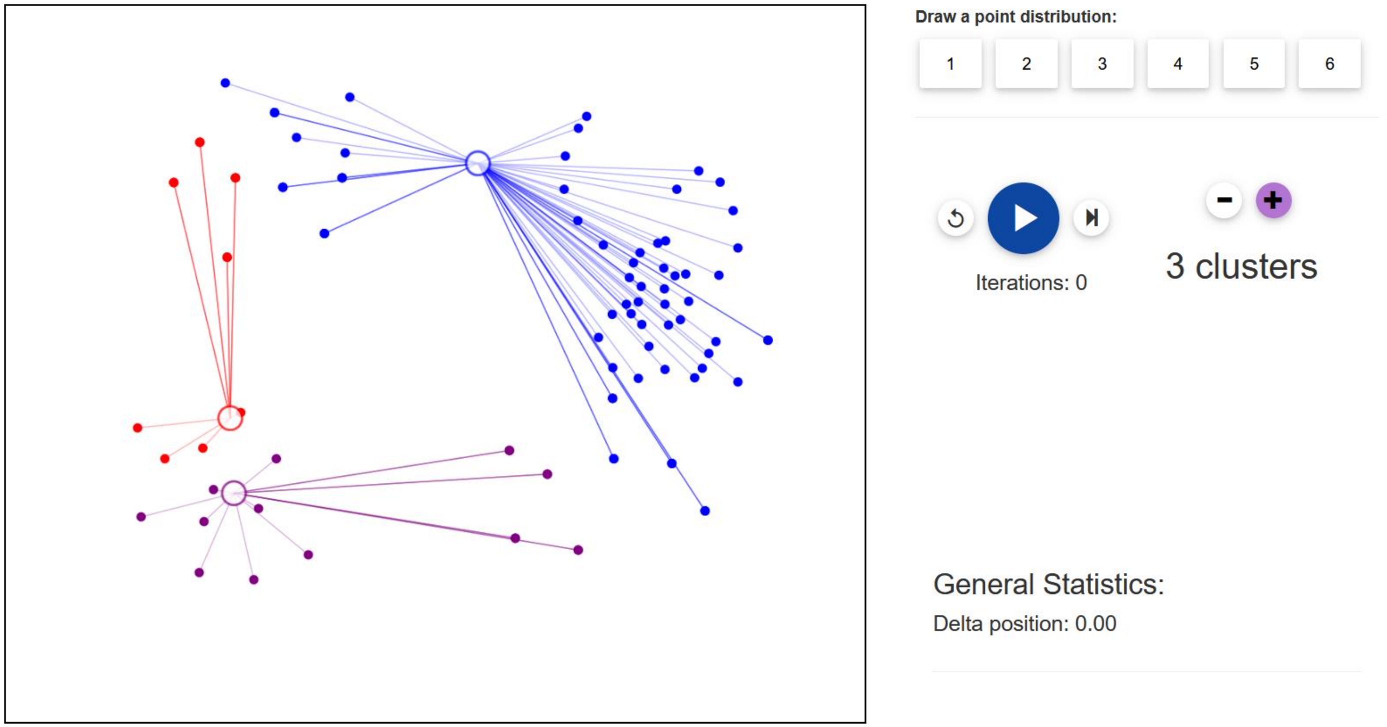
Algorithm 1: K-Means Clustering

1. Place K points into the feature space. These points represent initial cluster centroids.
2. Assign each pattern to the closest cluster centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the assignments do not change.

K-Means Clustering

Interactive Demo:

https://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/k-means/



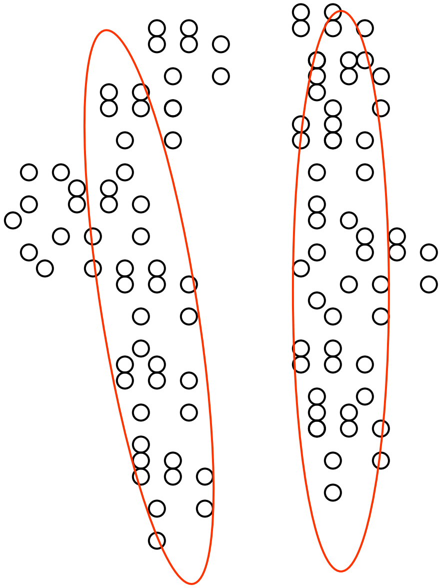
Discussions

1. How to determine *k*, the number of clusters?
2. Any alternative ways of choosing the initial cluster centroids?
3. Does the algorithm converge to the same results with different selections of initial cluster centroids? If not, what should we do in practice?

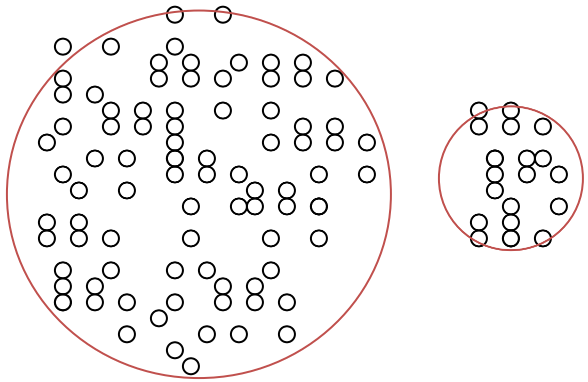
Brunel University London

Discussions

4. Intuitively, what is the ideal partition of the following problem? Can K-means Clustering algorithm give a satisfactory answer to this problem?



Elongated clusters



Clusters with different variation

Pros and Cons of KM

Advantages

- May be computationally faster than hierarchical clustering (if K is small).
- May produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

Pros and Cons of KM

Advantages

May be computationally faster than hierarchical clustering (if K is small).

May produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

Disadvantages

Fixed number of clusters can make it difficult to predict what K should be.

Different initial partitions can result in different final clusters.

Potential empty clusters (not always bad)

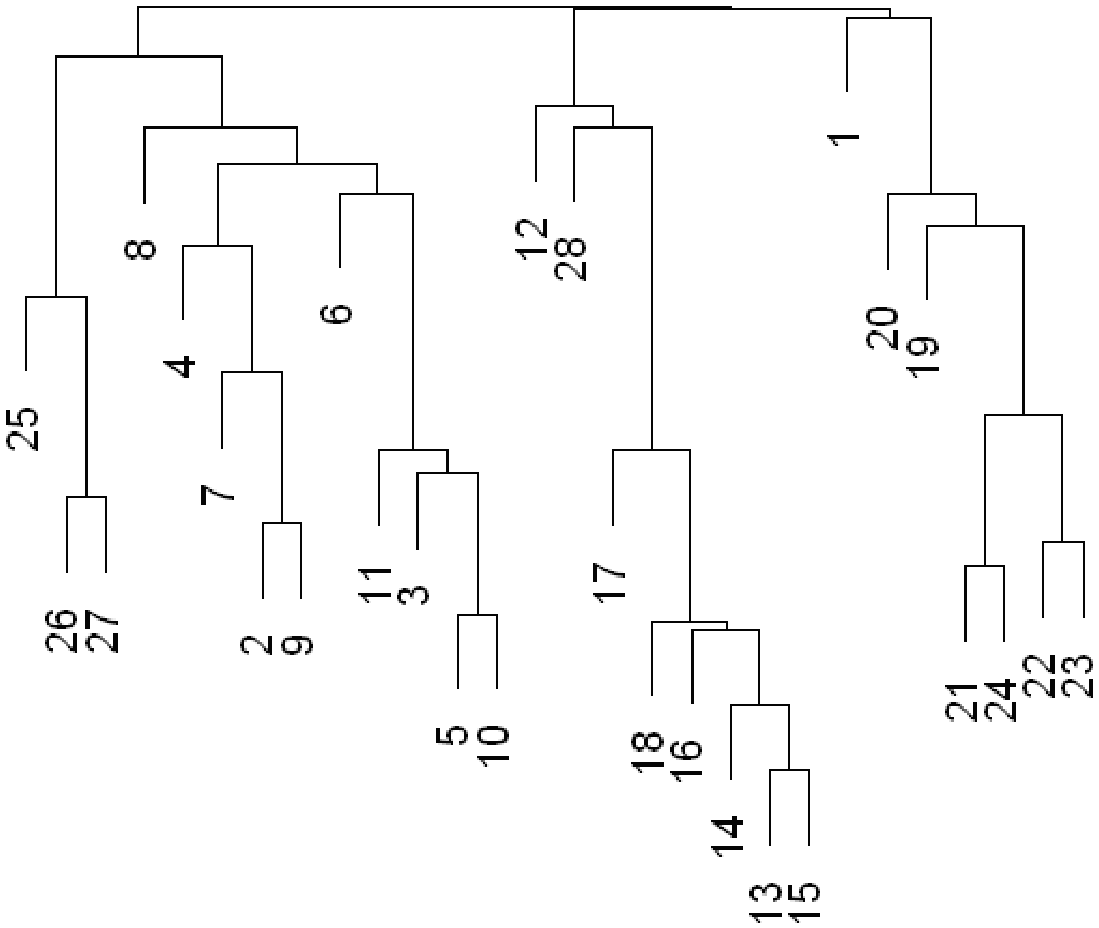
Does not work well with *non-globular* clusters.

Hierarchical (agglomerative) Clustering

- Hierarchical clustering results in a series of clustering results
- The results start off with each object in their own cluster and end with all of the objects in the same cluster
- The intermediate clusters are created by a series of merges
- The resultant tree like structure is called a dendrogram

Brunel University London

Dendrogram



The Hierarchical Clustering Algorithm

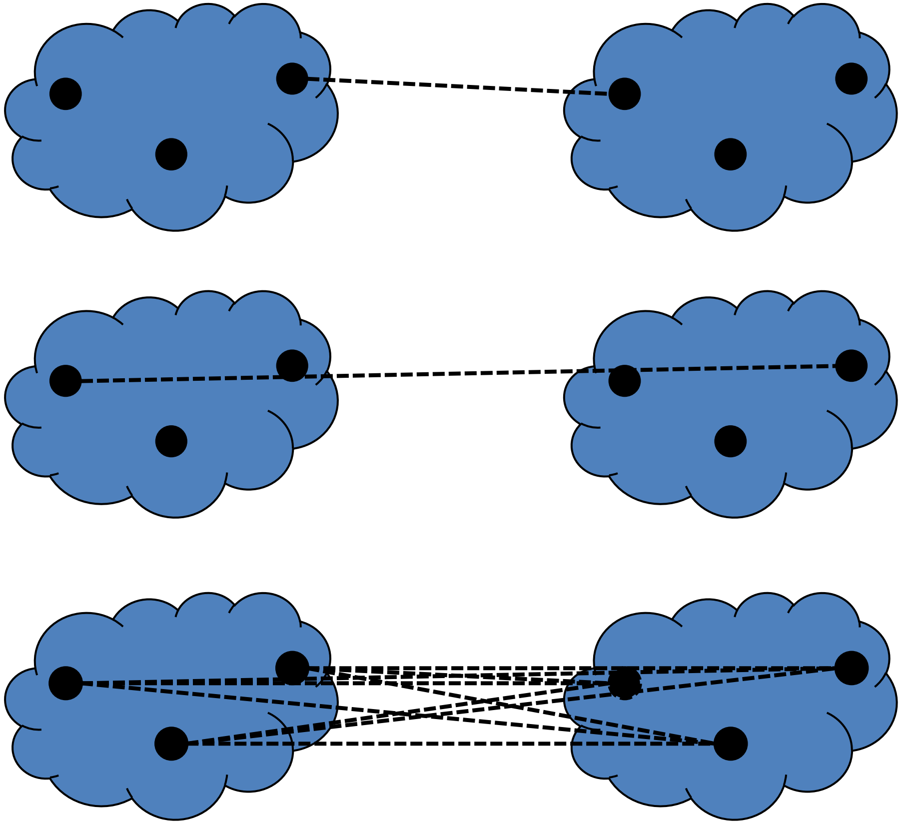
1) Each item is assigned to its own cluster (n clusters of size one)
2) Let the distances between the clusters equal the distances between the objects they contain
3) Find the closest pair of clusters and merge them into a single cluster (one less cluster)
4) Re-compute the distances between the new cluster and each of the old clusters
5) Repeat steps 3 and 4 until there is only one cluster left

Re-computing Distances

Single Linkage

Complete Linkage

Average Linkage



Re-computing Distances

Linkage	Description
Single	The smallest distance between any two pairs from the two clusters (one from each) being compared/measured
Average	The average distance between pairs
Complete	The largest distance between any two pairs from the two clusters (one from each) being compared/measured

Other methods include Ward, McQuitty, Median and Centroid

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

Pros and Cons of HC

Advantages

Can produce an ordering of the objects, which may be informative for data display.

Smaller clusters are generated, which may be helpful for discovery.

Pros and Cons of HC

Advantages

Can produce an ordering of the objects, which may be informative for data display.

Smaller clusters are generated, which may be helpful for discovery.

Disadvantages

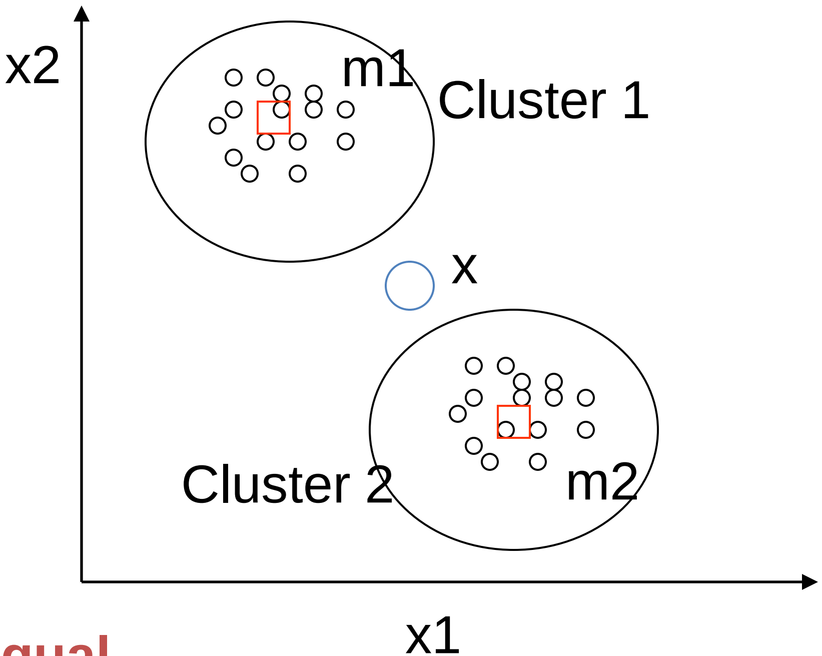
No provision can be made for a relocation of objects that may have been 'incorrectly' grouped at an early stage.

Use of different distance metrics for measuring distances between clusters may generate different results.

Limitations of K-Means (and Hierarchical)

At each iteration of the K-Means Clustering algorithm, a pattern can be assigned to one cluster only, i.e. the assignment is '**hard**'.

Observe an extra pattern x:
It locates in the middle of the two cluster centroids. So with K-Means Clustering algorithm, it will either
(1) drag m1 down, or
(2) drag m2 up.



But intuitively, it should have equal contributions to both clusters…

Other clustering methods
Fuzzy Clustering

- For example: Fuzzy c-means
- In real applications often no sharp boundary between clusters
- Fuzzy clustering is often better suited
- Fuzzy c-means is a fuzzification of k-Means and the most well-known

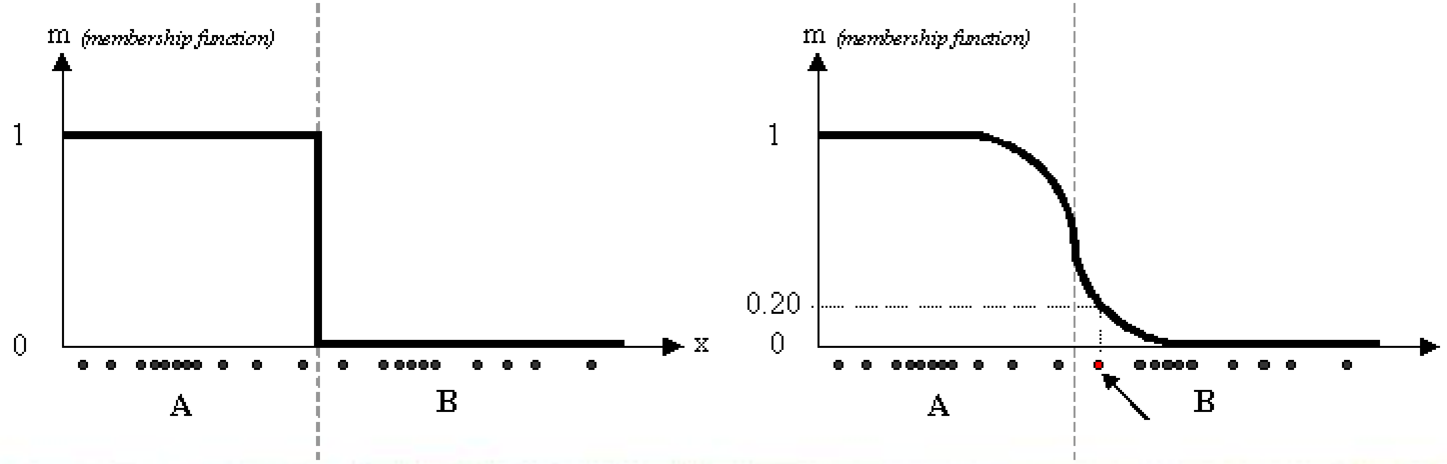
Brunel University London

Fuzzy Clustering

- http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletFCM.html

Cluster membership is now a weight between zero and one

The distance to a centroid is multiplied by the membership weight



Brunel University London

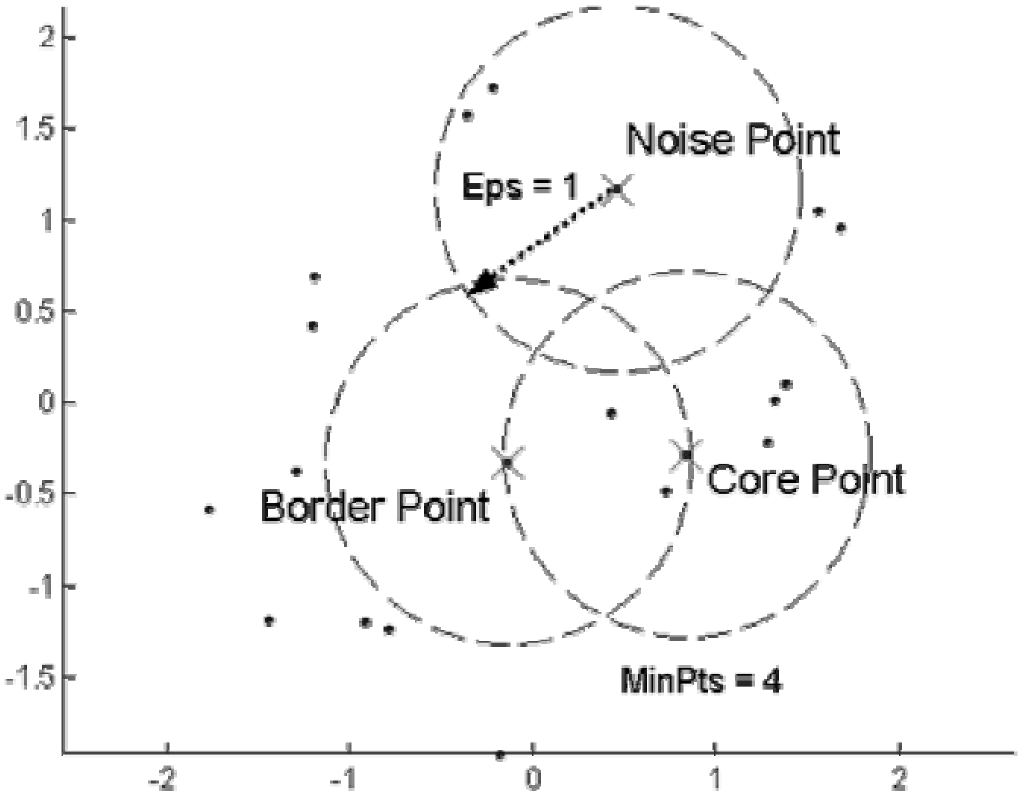
DBSCAN

(from Ranjay Sankar notes, University of Florida)

- DBSCAN is a density based clustering algorithm
- Density = number of points within a specified radius (*Eps)*
- A point is a *core point if it has more than* specified number of points (*MinPts) within Eps*
- (Core point is in the interior of a cluster)

DBSCAN

(from Ranjay Sankar notes, University of Florida)



http://www.cise.ufl.edu/class/cis4930sp09dm/notes/dm5part4.pdf

Brunel University London

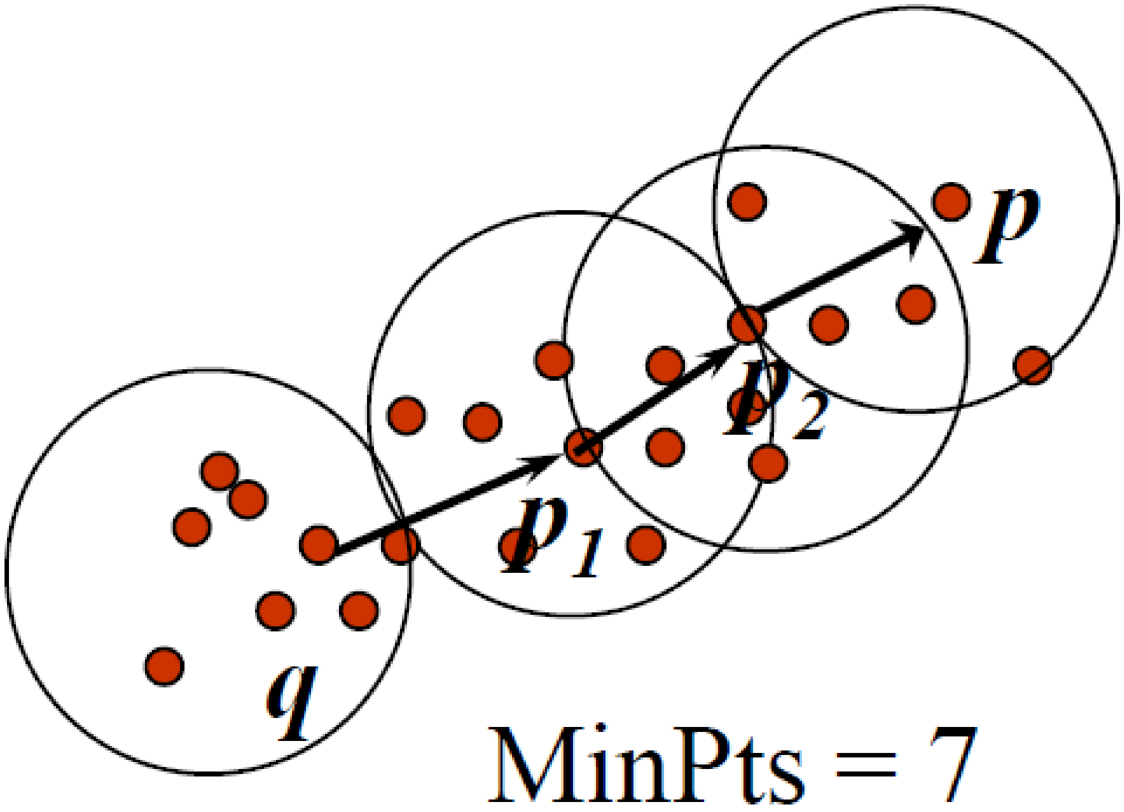
DBSCAN

- Density-Reachable (directly and indirectly):
- A point *p is directly density-reachable from p2*
- *p2 is directly density-reachable from p1*
- *p1 is directly density-reachable from q*
- *p <- p2 <- p1 <- q form a chain*

DBSCAN

(University of Buffalo)

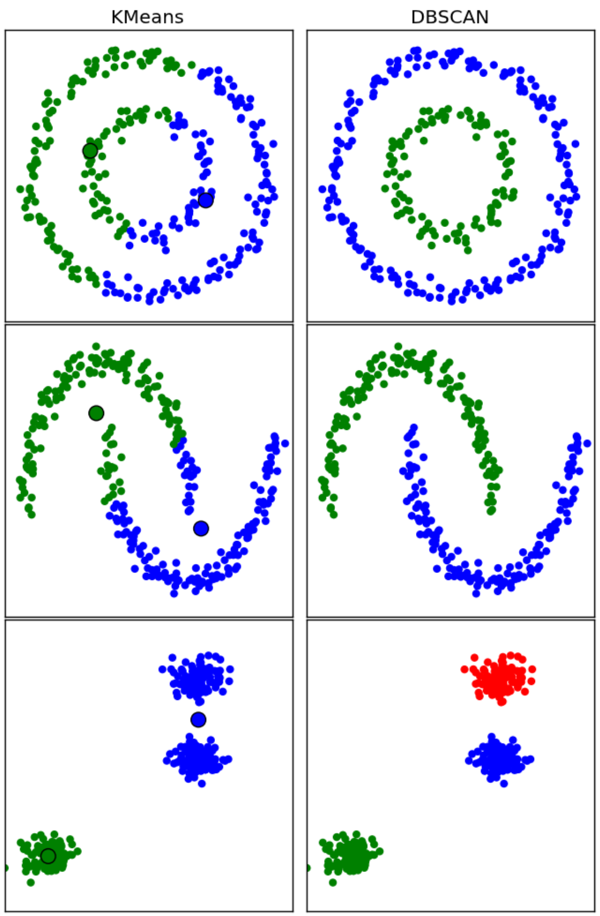
http://www.cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_density.pdf



DBSCAN

(from Ranjay Sankar notes, University of Florida)

https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

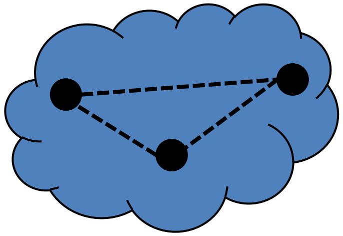


Evaluating Cluster Quality

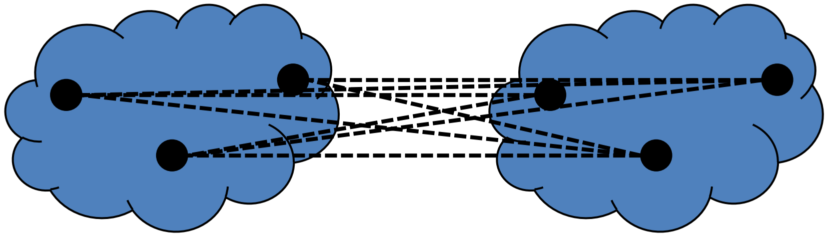
- How do we know if the discovered clusters are any good?
- The choice of correct metric for judging the worth of a clustering arrangement is vital for success
- There are as many metrics as methods!
- Each has their own merits and drawbacks

Cohesion & Separation

- Cohesion



- Separation



Supervised

- But what if we know something about the “true clusters”
- Can we use this to test the effectiveness of different clustering algorithms?

Comparing Clusters

- Metrics exist to measure how similar two clustering arrangements are
- Thus if a method produces a set of similar clustering arrangements (according to the metric) then the method is consistent
- We will consider the Weighted-Kappa metric which has been adapted from Medical Statistics

Weighted-Kappa

Weighted Kappa (WK)	Agreement Strength
$-1 \leq WK \leq 0$	Very Poor
$0 < WK \leq 0.2$	Poor
$0.2 < WK \leq 0.4$	Fair
$0.4 < WK \leq 0.6$	Moderate
$0.6 < WK \leq 0.8$	Good
$0.8 < WK \leq 1.0$	Very Good

The Weighted Kappa Guideline

Brunel University London

Association Rules

Another form of unsupervised learning

Works with “basket data”

Market Basket Example

? Where should detergents be placed in the Store to maximize their sales?

? Are window cleaning products purchased when detergents and orange juice are bought together?

? Is soda typically purchased with bananas? Does the brand of soda make a difference?

? How are the demographics of the neighborhood affecting what customers are buying?

Image source: deepclimate.org

Support, Confidence & Lift

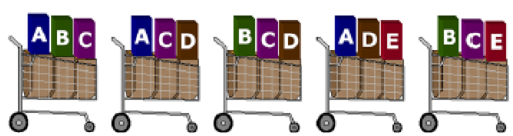
$$Support = \frac{frq(X,Y)}{N}$$

$$Rule: \; X \Rightarrow Y$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Example:



Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \,\&\, C \Rightarrow D$	1/5	1/3	5/9

Brunel University London

Unsupervised Learning: Association Rules

- Supermarkets / Online Sites use this all the time
- Given a large amount of basket data, generate rules:

<Set of items>
<Set of items> → If <items> then <items>
(confidence / support)

<Set of items>

⋮

Unsupervised Learning: Association Rules

- Supermarkets / Online Sites use this all the time
- Given a large amount of basket data, generate rules:

<Set of items>

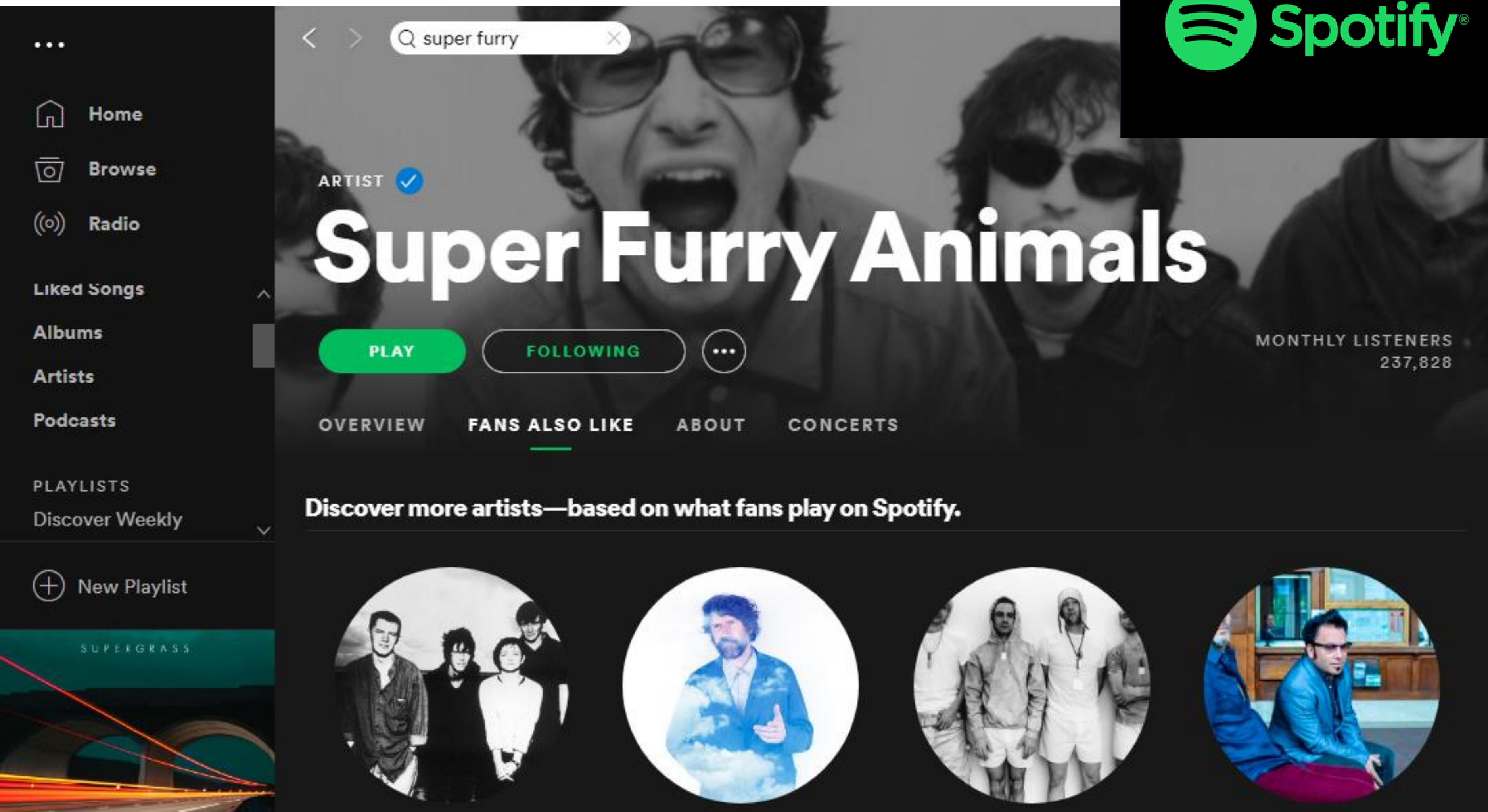
<Set of items> ➡ If <items> then <items>

(confidence / support)

<Set of items>

⋮

Unsupervised Learning: Association Rules



In the Lab

Cluster some toy datasets

Cluster some real datasets

Visualise with scatterplots and compare with WK

THIS IS ASSESSED

Brunel University London

Reading

- Chapter 9, Section 9.3: David Hand "Principles of Data Mining", MIT Press

- Pang-Ning Tan "Introduction to Data Mining" (Chapter 8): http://www-users.cs.umn.edu/~kumar/dmbook/index.php

- Anil Jain: "Data Clustering: 50 Years Beyond K-Means", Pattern Recognition Letters

Tang et al., Kumar, Introduction to Data Mining (Chapter 6): https://www-users.cs.umn.edu/~kumar001/dmbook/index.php