Clustering Analysis Overview

Ke Chen

Department of Computer Science, The University of Manchester

Ke.Chen@manchester.ac.uk

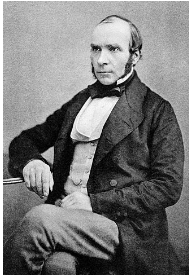
Outline

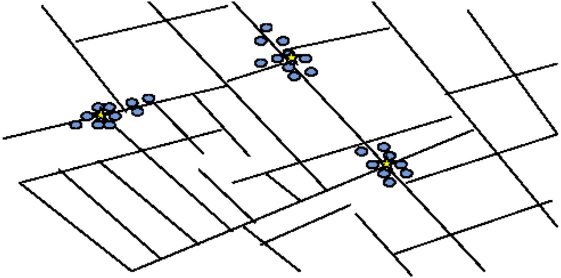
Introduction
History, cluster, and clustering analysis

Fundamental Issue
Illustrative examples, ill-posed nature and clustering axioms

Clustering Methodology
Partitioning, graph-based, model-based, hierarchical, density-based and ensemble clustering approaches

Real Application
Traditional and emerging clustering applications

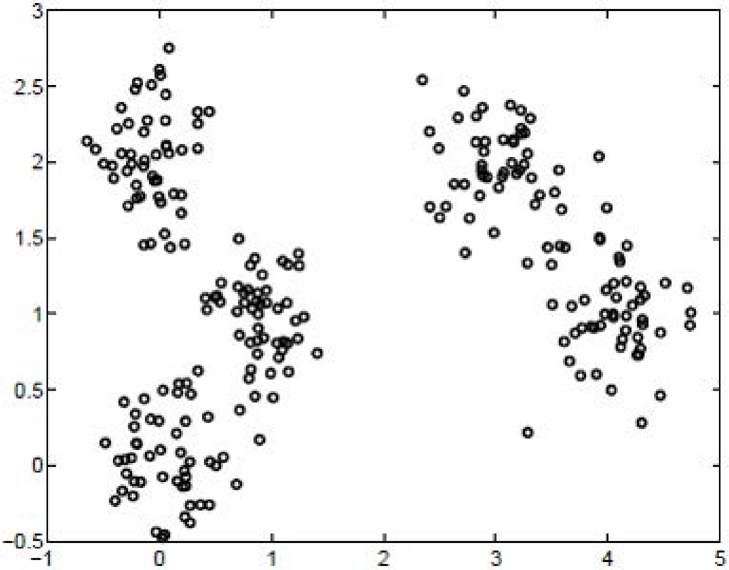




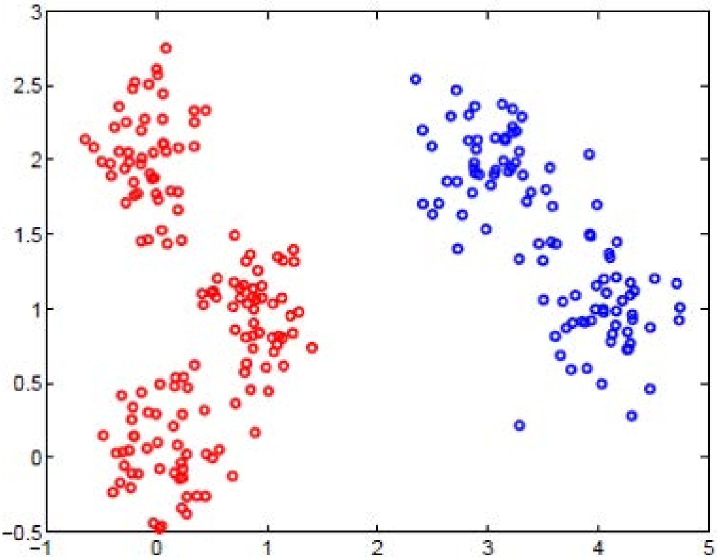
- Clustering analysis was originated by John Snow (1813-1858), a London-based Physician, who used this technique to plot the location of cholera deaths on a city map during an outbreak in the 1850s.
- In his annotated map, the locations indicated that cases were clustered around certain intersections where there were polluted wells, which thus exposed both the problem and the solutions.

- Cluster refers to a collection or group of data items or objects that meet the following properties:
  - data items in the same collection are similar/related to each other
  - data items in different collections are dissimilar or unrelated
- Clustering analysis: a process of organising unlabelled data items into groups named clusters with either similarity or dissimilarity criteria.
- As an unsupervised representation learning methodology, clustering analysis acts for high-level data summarisation and can be applied as
  - stand-alone tool to gain an insight into data distribution
  - pre-processing step of other learning algorithms

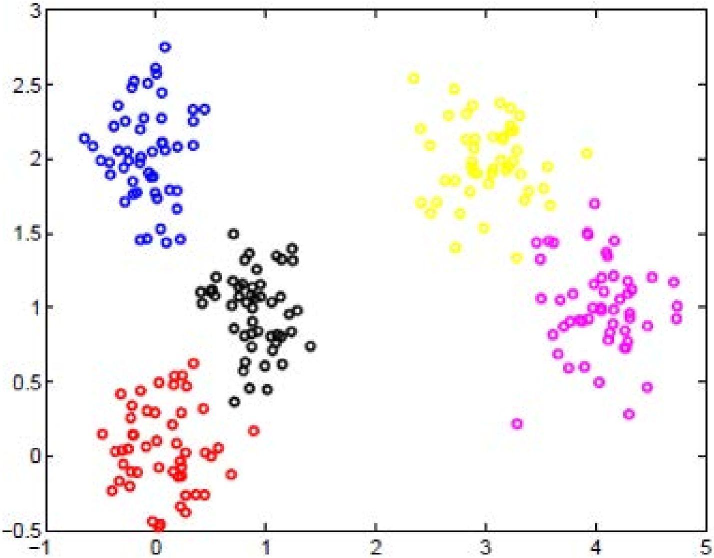
- **Illustrative example**: how many clusters in a dataset?



- **Illustrative example**: how many clusters in a dataset? (2 clusters)



- **Illustrative example**: how many clusters in a dataset? (5 clusters)



- **Illustrative example**: are they always in same clusters?

sheep, cat, red mullet, frog, dog, lizard, sparrow, viper, seagull, blue shark, gold fish

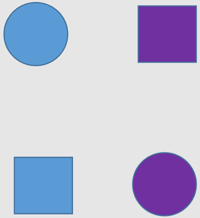
Criterion 1: the birth-giving method

- blue shark, sheep, cat, dog
- lizard, sparrow, viper, seagull, gold fish, frog, red mullet

Criterion 2: the existence of lung

- gold fish, red mullet, blue shark
- sheep, sparrow, dog, cat, seagull, lizard, frog, viper

Shape versus Colour



Fundamental Issue

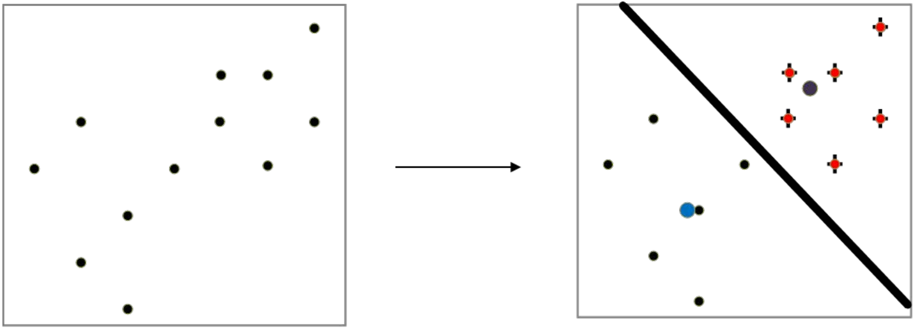
- In general, clustering analysis is well known as an ill-posed problem without ground-truth!
- Key issues involved in clustering analysis
  - choose an appropriate similarity or distance (dissimilarity) measure
  - discover the number of intrinsic or natural clusters underlying data
  - find out a manner to group data items into sensible or “wanted” clusters via a proper clustering algorithm
- In addition, clustering validity indexes encoding prior knowledge or expected “gold standard” could be used to evaluate the clustering results. However, such indexes reflect only some aspects; none of an individual index works for all real applications.

Fundamental Issue

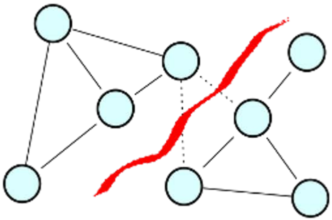
- Impossibility theorem for clustering (Kleinberg, 2002) shows none of clustering algorithms can meet all 3 axioms required by all-purpose clustering analysis:
    - Scale-invariance: for any distance measure d and any $\alpha > 0$ so as to have scaled distance $d' = \alpha d$, the clustering function, $f(\cdot)$, produces the same clustering results on a dataset, $f(d) = f(d')$.
    - Richness: the clustering function should be "rich" to produce all possible partitions of a dataset.
    - Consistency: suppose the clustering result Γ arises from the distance measure d. If another distance d' is made by reducing distances within the clusters and enlarging distance between the clusters, the use of the distance d' should lead to the same clustering result Γ.

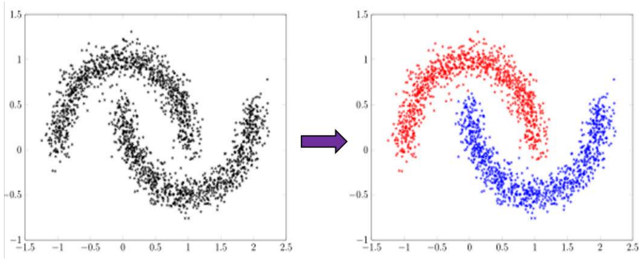
Clustering Methodology

- Partitioning clustering
  - find out an optimal partition of a dataset from different candidates in terms of some criteria, e.g. minimising the sum of squared distances within clusters.
  - Algorithms: K-means, K-medians, K-medoids, CLARANS, $\cdots$

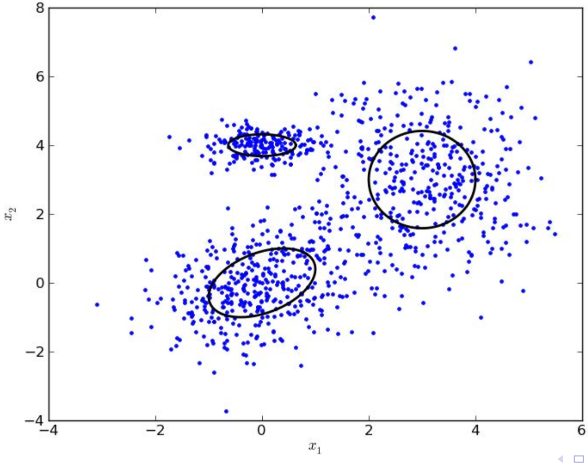


- Graph-based or spectral clustering
  - Convert dataset into weighted graph, then conduct spectral analysis on the weighted "similarity" matrix to produce an optimal partition.
  - Often, the spectral analysis is treated as feature extraction and a partitioning method is further applied in the new feature space for clustering analysis.
  - Algorithms: Normalised min-cut, unnormalised/normalised spectral clustering, $\cdots$

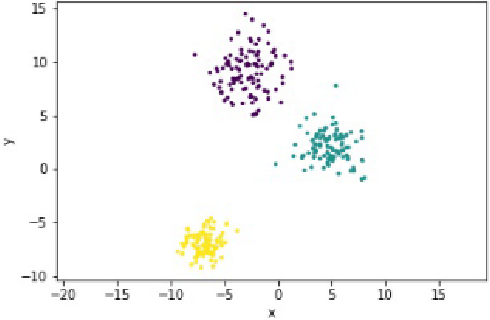


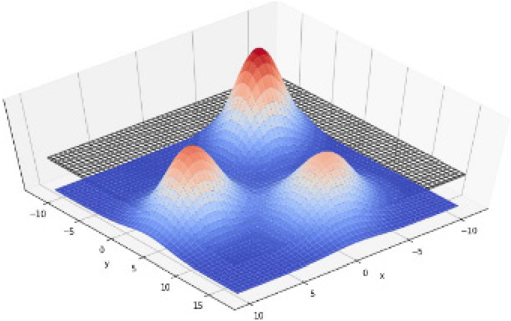


- Model-based clustering
  - A generative (probabilistic) model is hypothesised for each of the clusters and tries to find out the best fit of that model to each other.
  - Algorithms: Gaussian mixture model (GMM), finite mixture models, $\cdots$

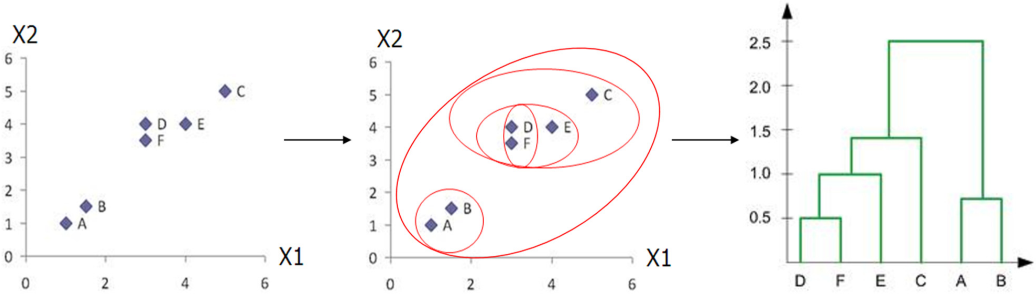


- Density-based clustering
  - Clustering is based on density function underlying data distribution.
  - Regions of high density form clusters and regions of low density separate clusters.
  - Algorithms: DBSCAN, OPTICS, DenClue, $\cdots$

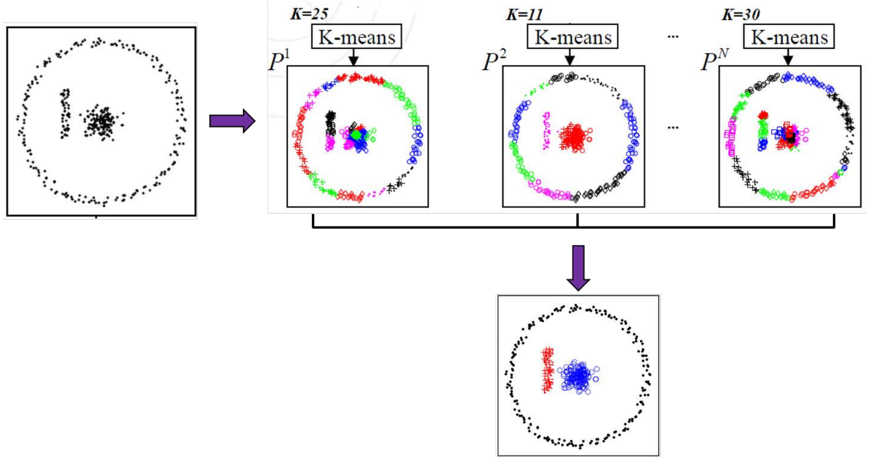




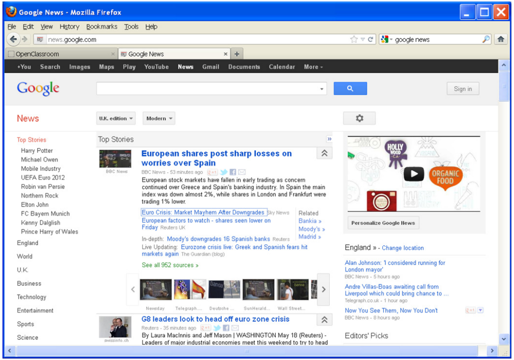
- Hierarchical clustering
  - Clustering is done by creating hierarchical decomposition of data items or objects based on certain grouping/splitting criteria for all possible clusters.
  - Algorithms: Agglomerative, Diana, BIRCH, ROCK, $\cdots$

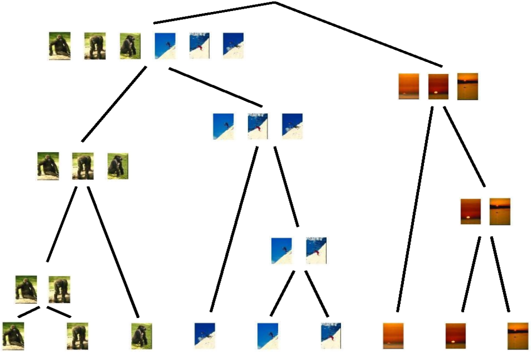


- Clustering ensemble
  - Combine multiple clustering partitions resulting from different clustering analyses to generate a consensus partition better than individual partitions.
  - Algorithms: Evidence-accumulation, Information-theoretic, Hyper-graph, $\cdots$

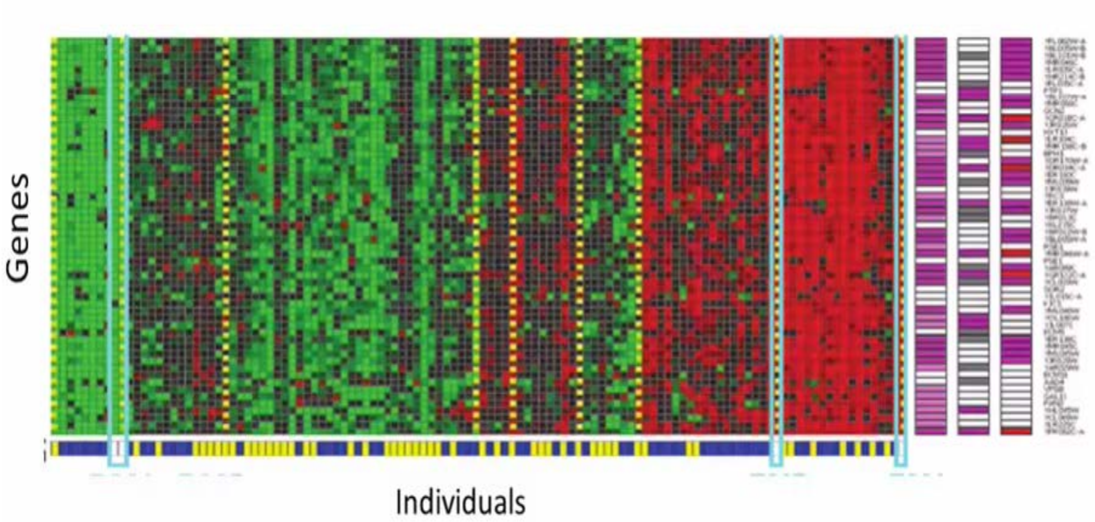


- News article and picture organisation on website/database



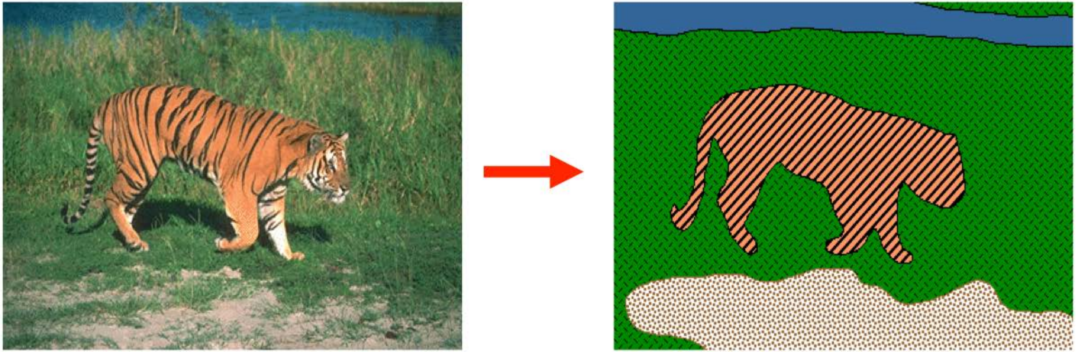


- Protein/gene sequence analysis according to expression profile



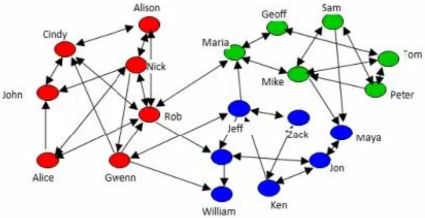
Application

- Computer vision: image segmentation

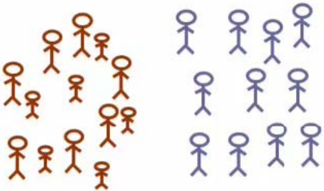


- Miscellaneous real applications

Organize computing clusters



Social network analysis



Market segmentation.

Astronomical data analysis

Reference

If you want to deepen your understanding and learn something beyond this lecture, you can self-study the optional references below.

[Kleinberg, 2002] Kleinberg J. (2002): An impossibility theorem for clustering. In *Advance in Neural Information Processing Systems* (NIPS'02).

[Jain et al., 1999] Jain A.K., Murty M.N. and Flynn P.J. (1999): Data clustering: A review. *ACM Computing Survey*, Vol. 31, No. 3, pp. 264-323.

[Xu & Wunsch II, 2005] Xu R. and Wunsch II D. (2005): Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, Vol. 16, No. 3, pp. 645-678.

[Xu & Tian, 2015] Xu D. and Tian Y. (2015): A comprehensive survey of clustering algorithms. *Annals of Data Science*, Vol. 2, pp. 165-193.