Machine Learning: Clustering

Dr Chris Yeomans

c.m.yeomans@exeter.ac.uk

Camborne School of Mines, CEMPS, Penryn Campus, University of Exeter

UNIVERSITY OF EXETER | DOCTORAL COLLEGE

Researcher Development

Acknowledgements

Dr JJ Valetta

- Now at University of St Andrews

Dr Jiangjiao Xu

- University of Exeter

Additional resources:

https://github.com/GeostatsGuy

https://www.youtube.com/channel/UCLqEr-xV-ceHdXXXrTId5ig

Valletta et al. (2017)

Intended Learning Outcomes

By the end of this lecture you will:

- Know what clustering is
- Understand the basic principles of unsupervised machine learning
- Have a broad overview of different algorithms and what they can be used for
- Appreciate the key drawback of clustering – defining the number of clusters

Overview

- What is clustering?
- Major types of clustering methods
- Clustering algorithms
  - *k*-means clustering
  - Agglomerative hierarchical clustering
  - Gaussian mixture models
  - Self-Organizing Maps
- How do we determine the “correct” number of clusters?

What can clustering do?

- Gene expression: discovering co-regulated genes
- Computer vision: segmenting a digital image for object recognition
- Epidemiology: identifying geographical clusters of diseases
- Geochemistry: similar elements explain geological processes
- Automated mapping: spatial geological or environmental data
- Engineering: predicting rock hardness and drilling requirements
- Market analysis: Amazon, Google, Netflix
- Risk assessment: insurance companies

What is clustering?

- Formal definition: Identifying homogeneous and well separated groups of data points (features) by some similarity measure
- Informal: The process of stereotyping your data
  e.g these are round(ish) faces, these are short(ish) people
- How many clusters: An unsolved problem. Issue lies in the subjectivity of the word similar and its mathematical definition

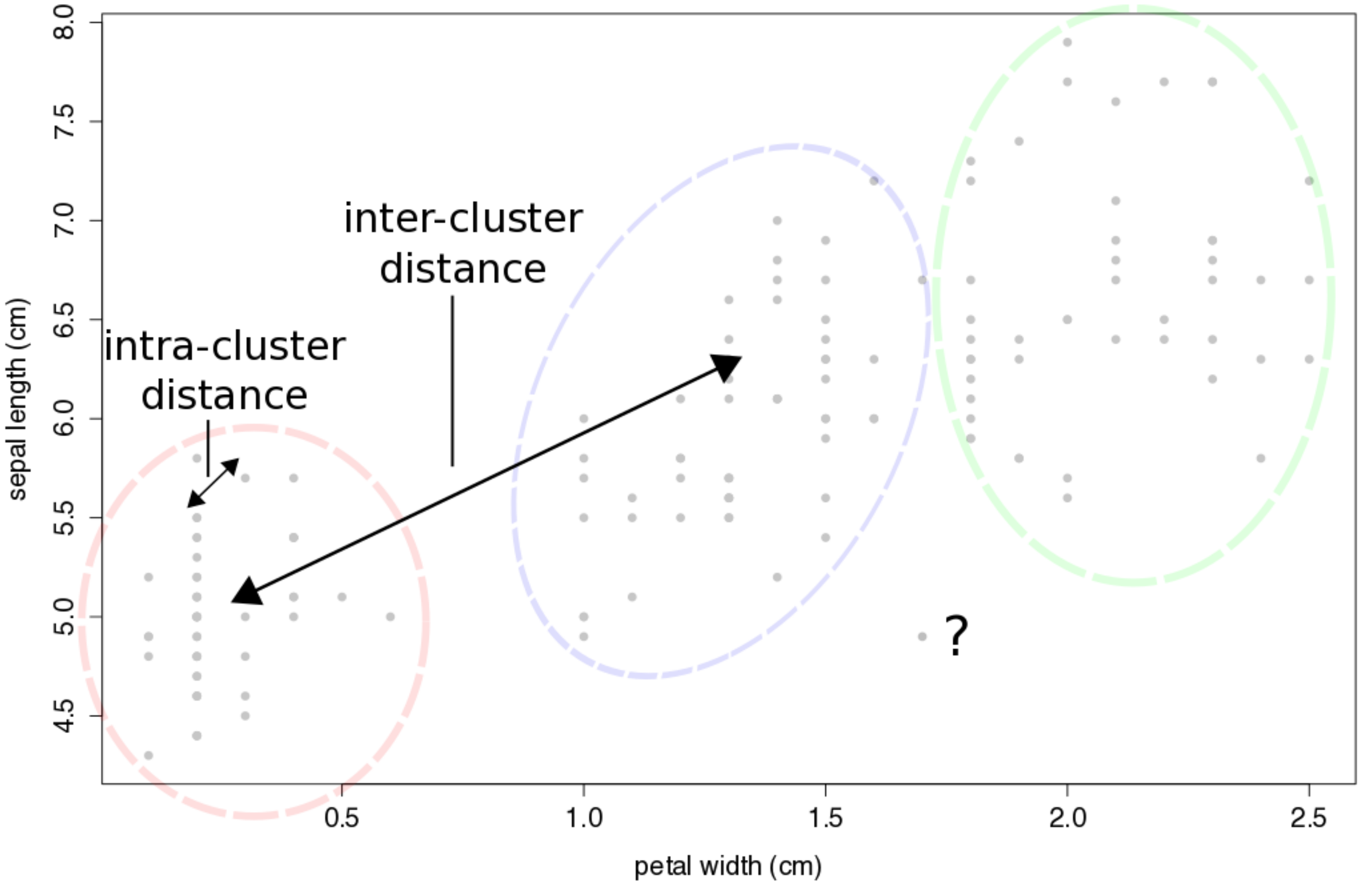
What makes a 'good' clustering?

What makes a ‘good’ clustering?

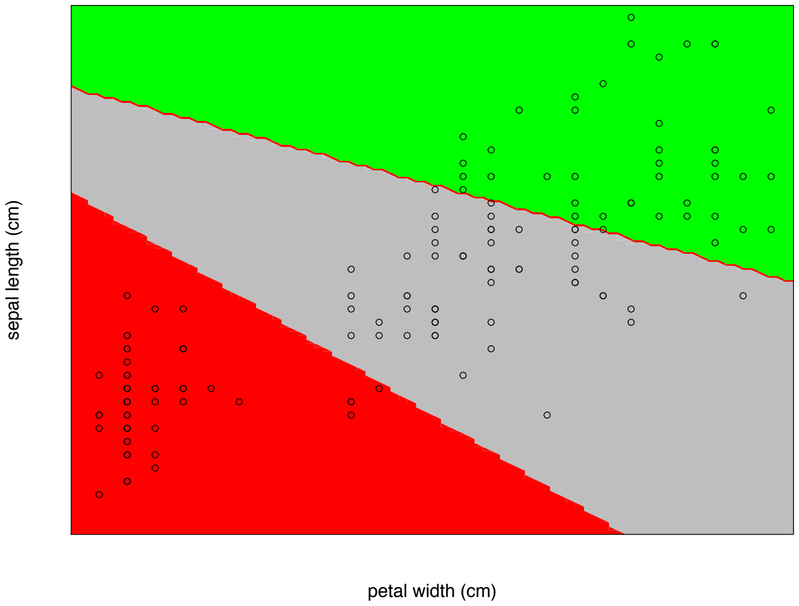
What makes a ‘good’ clustering?

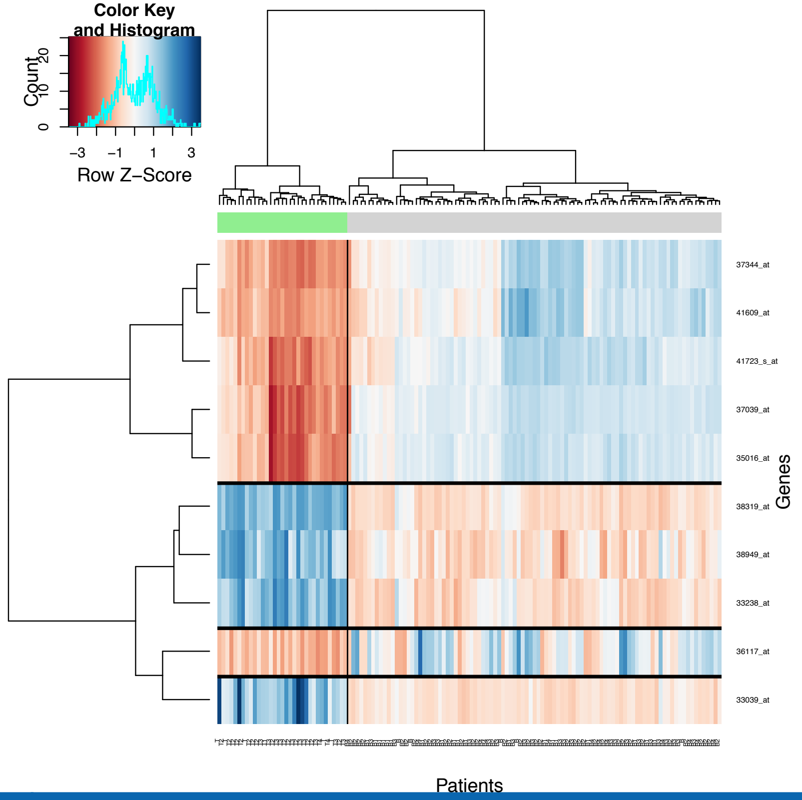
- Aim: Identify data structure and any outliers
- Optimisation: High intra-cluster similarity and low inter-cluster similarity

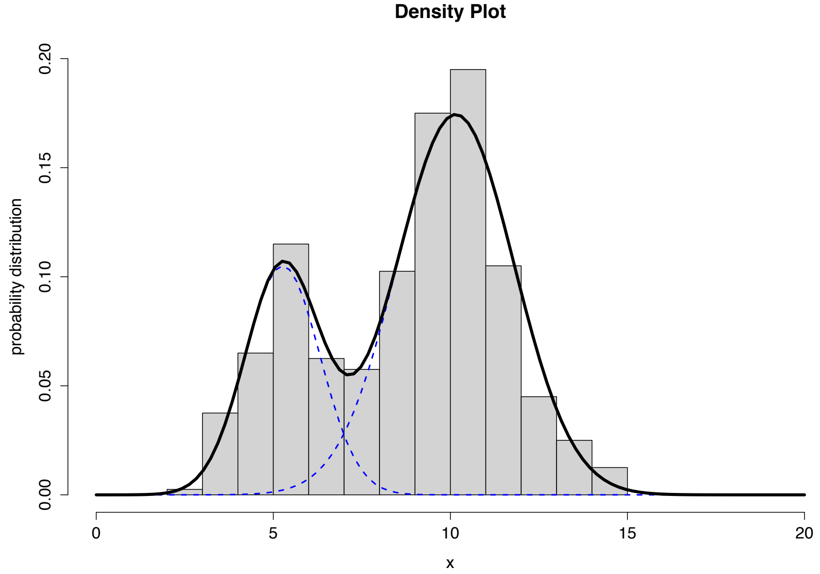


Types of Clustering

- Partitional: separating the feature space into *k* regions
- Hierarchical: Iterative merging (agglomerative) or breaking (divisive) of clusters
- Distribution-based: Fit *k* multivariate statistical distributions

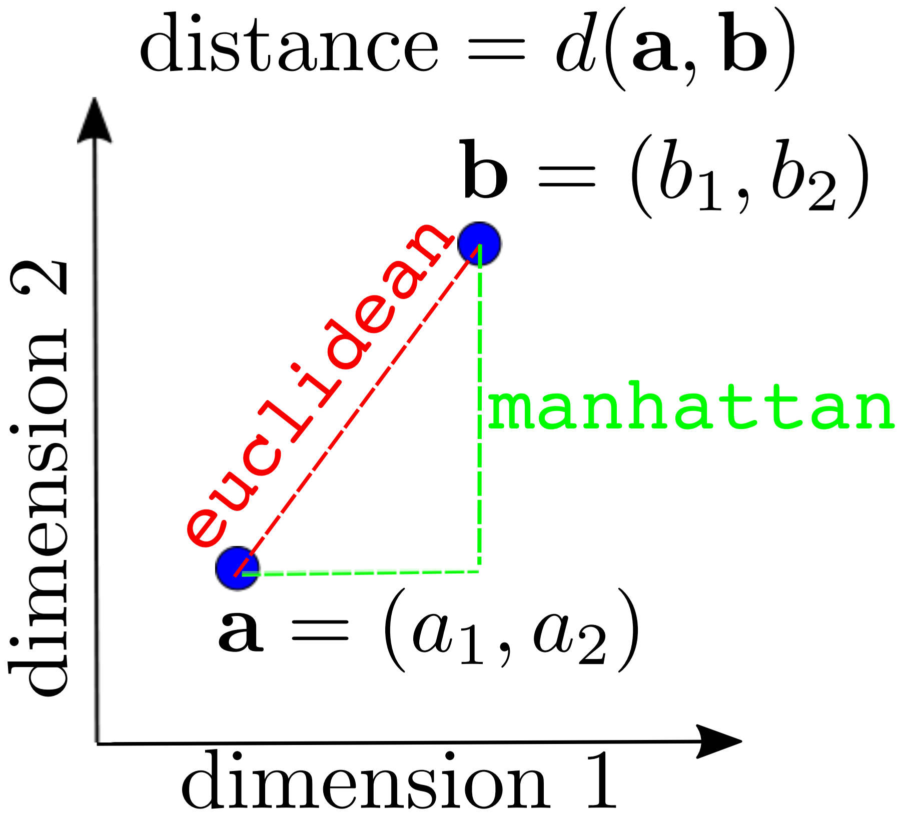


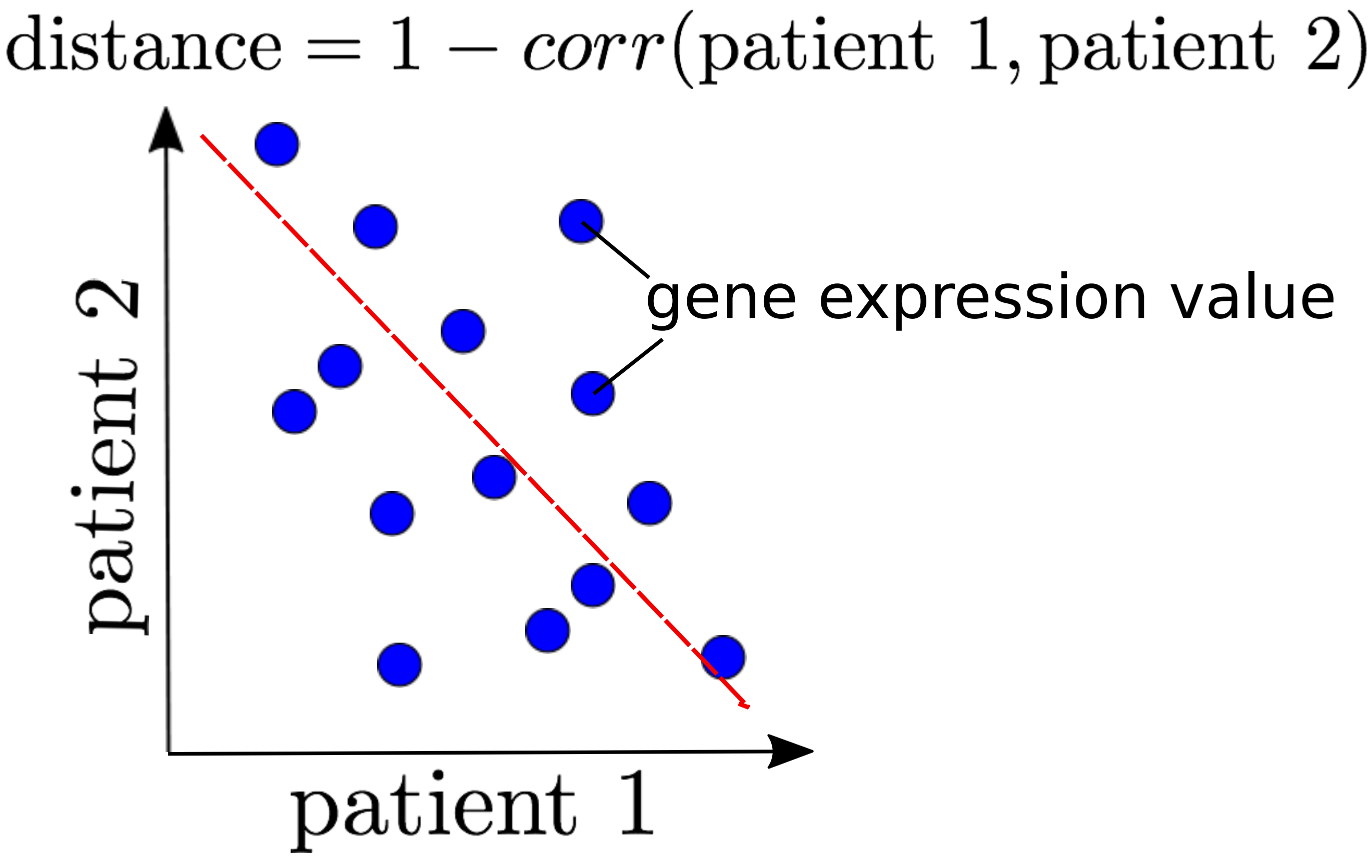




Types of Clustering

- All types of clustering use some form of distance metric to gauge similarity (or dissimilarity)

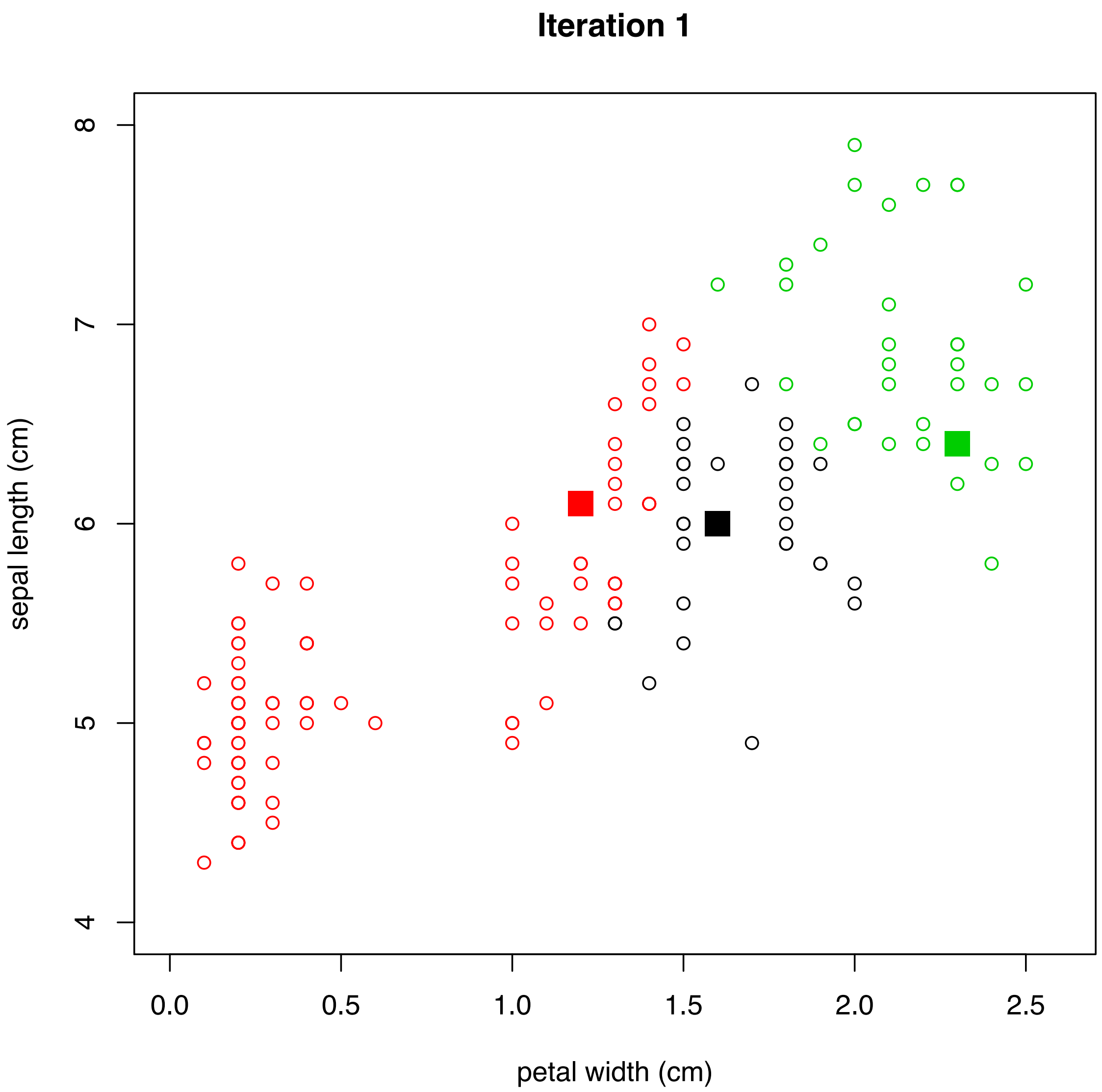




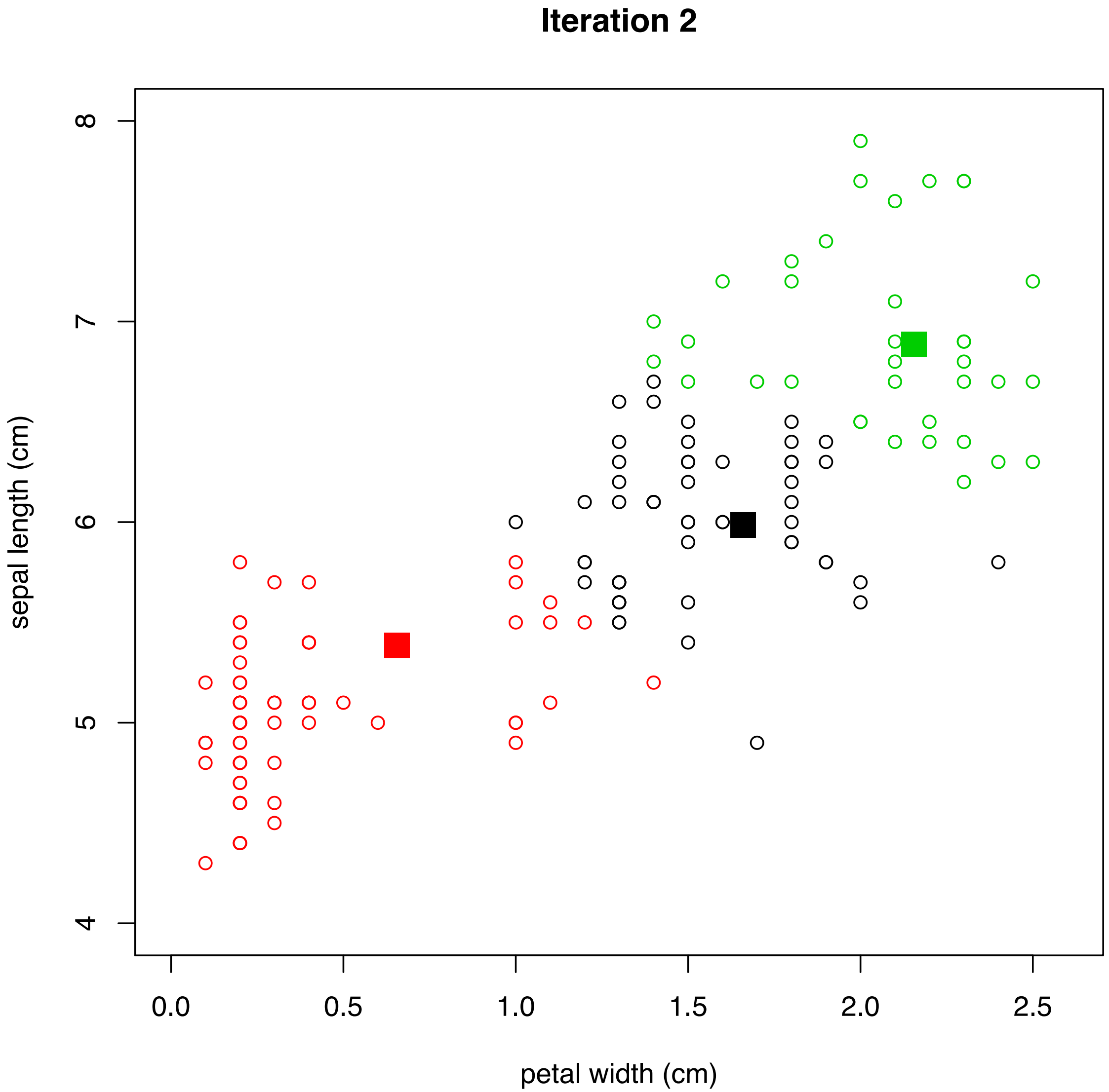
k-means algorithm

1. Define parameters: *k*, the number of cluster centroids
2. Randomise: randomly select starting point for *k* centroids
3. Similarity: calculate distance between points
4. Cluster: assign points to nearest centroid
5. Optimise: re-calculate centroid based on new point assignment
6. Iterate: repeat steps 3-5 until change threshold or maximum iterations are reached

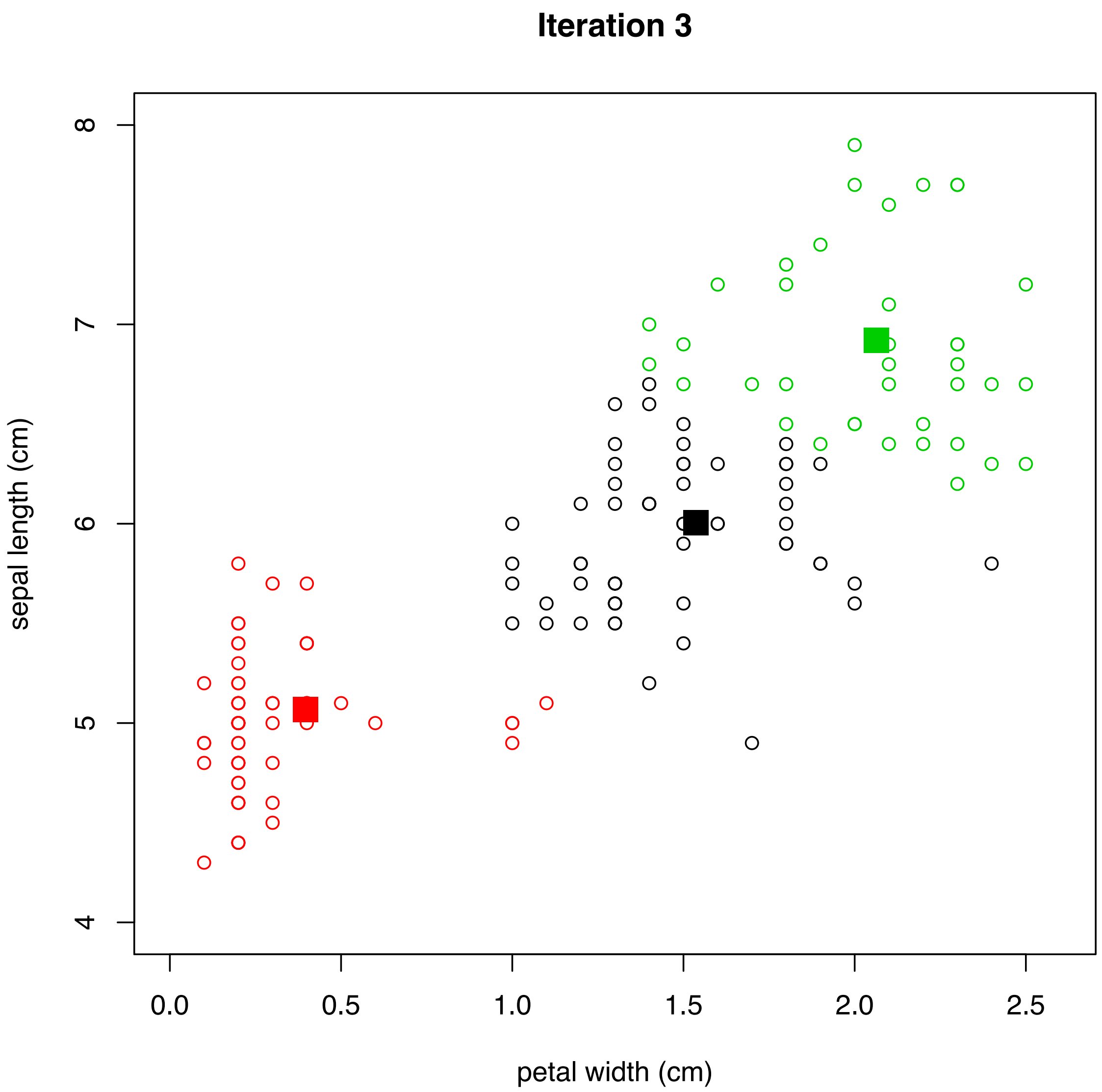
k-means algorithm



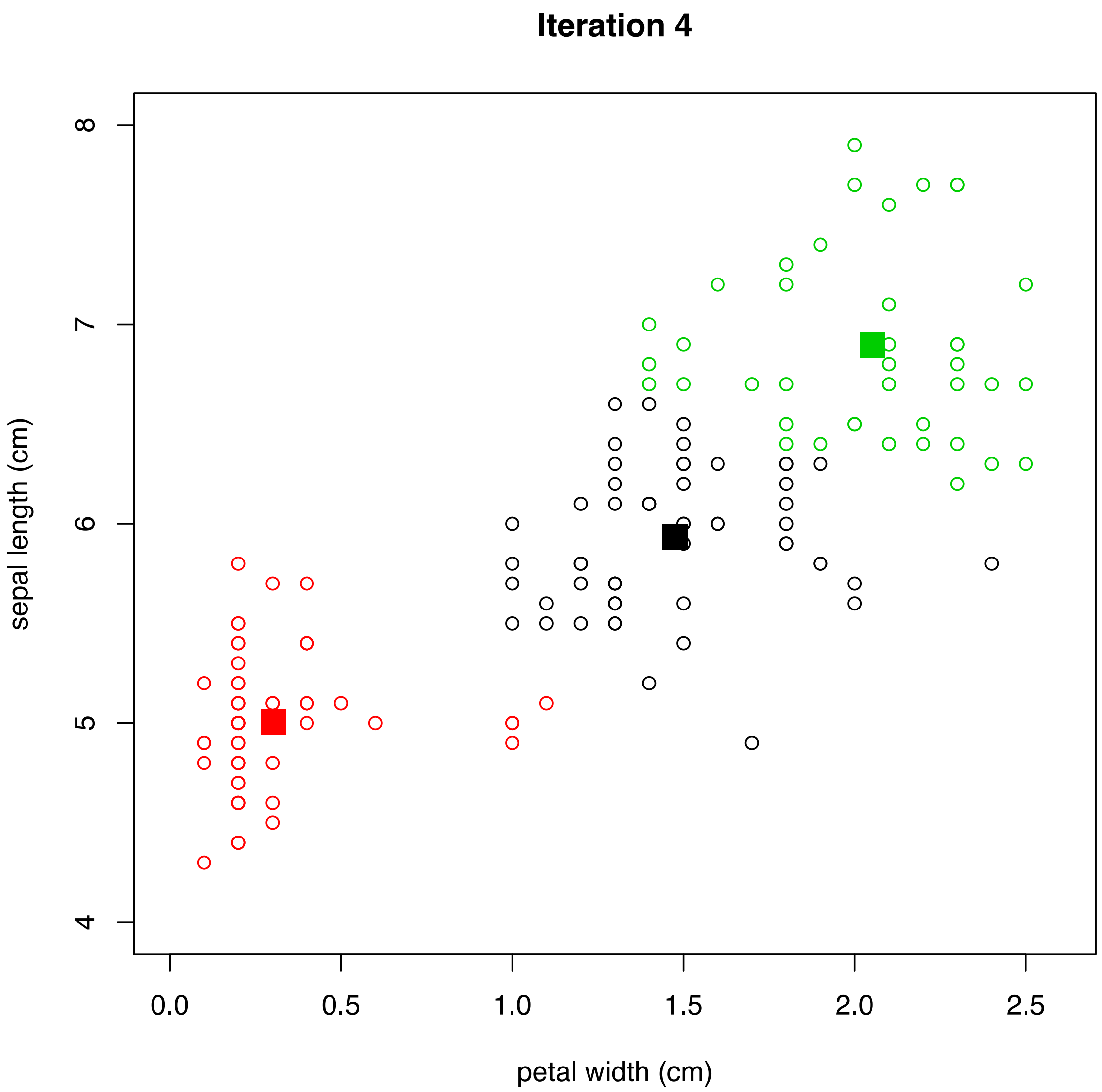
k-means algorithm



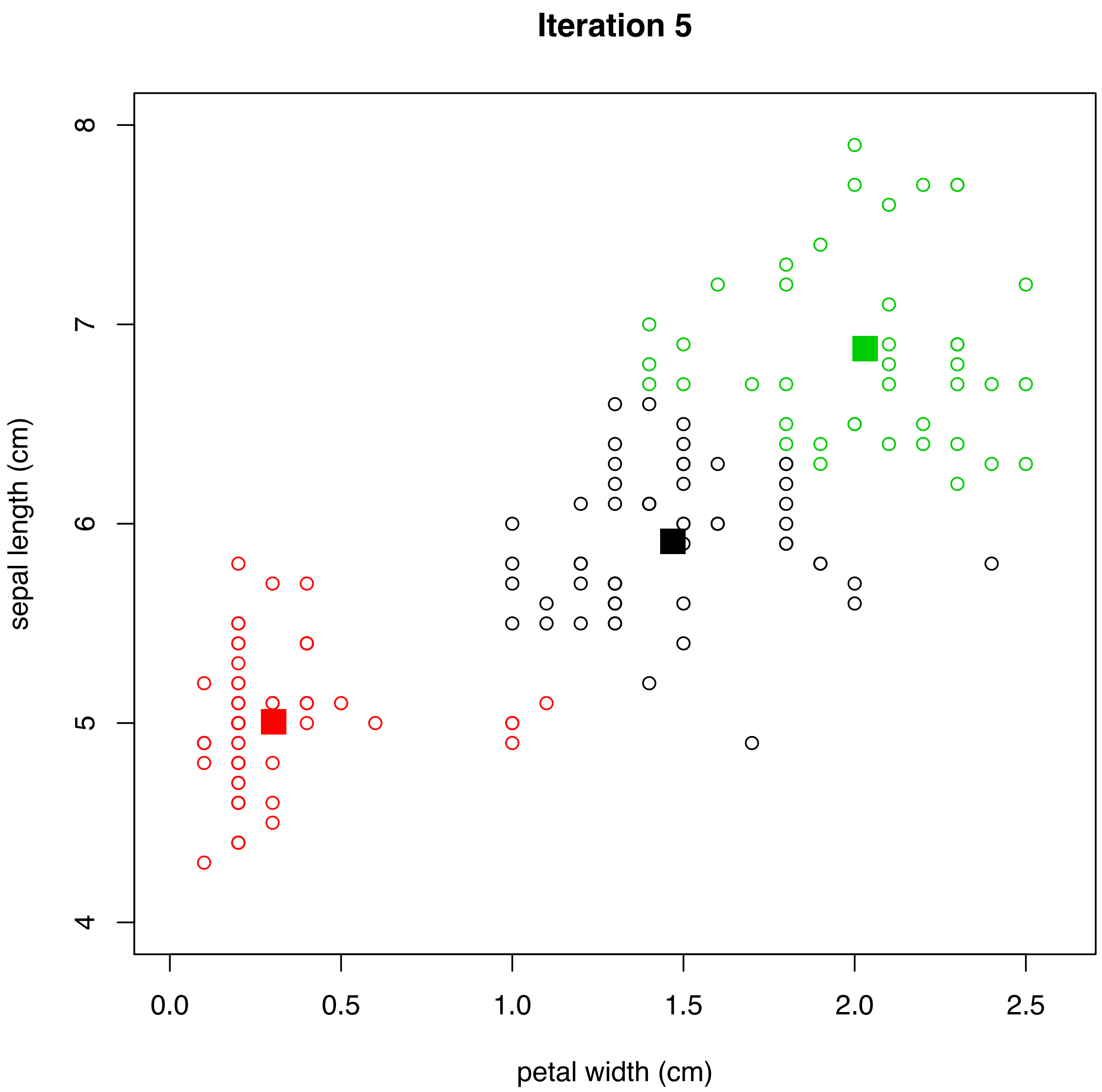
k-means algorithm



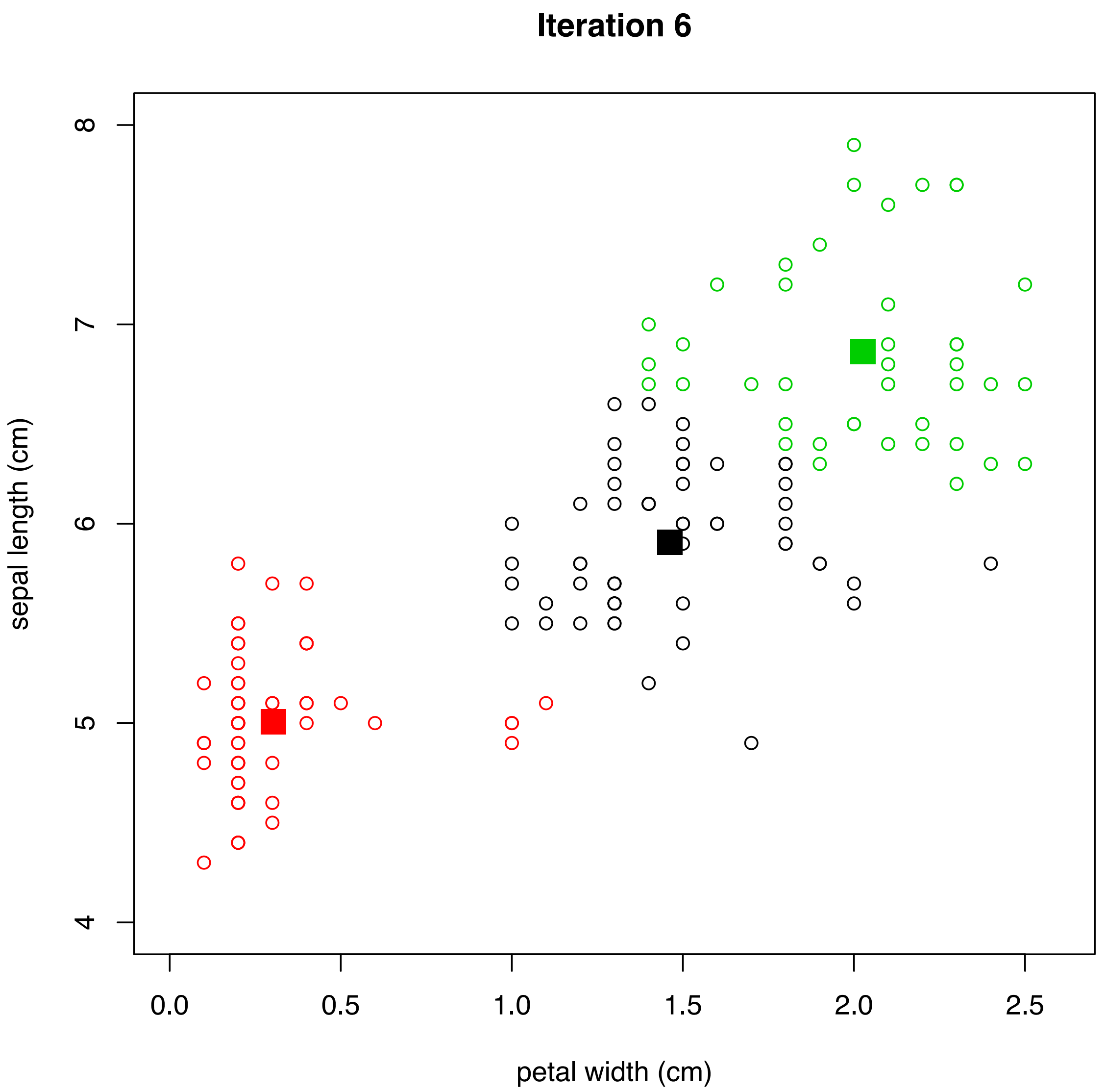
k-means algorithm



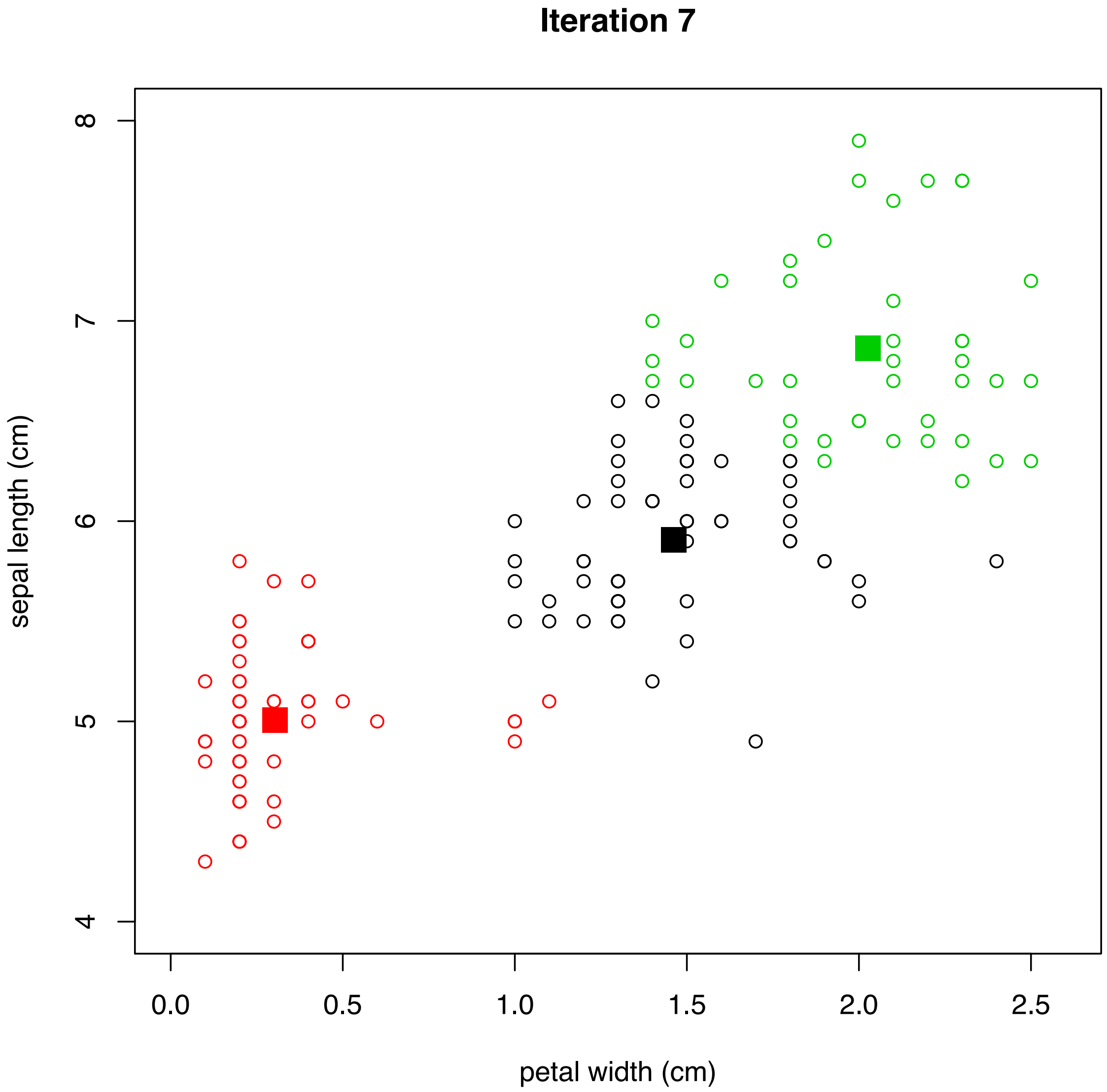
k-means algorithm



k-means algorithm



k-means algorithm



k-means algorithm

Pros

- Simple and intuitive
- Computationally inexpensive/fast

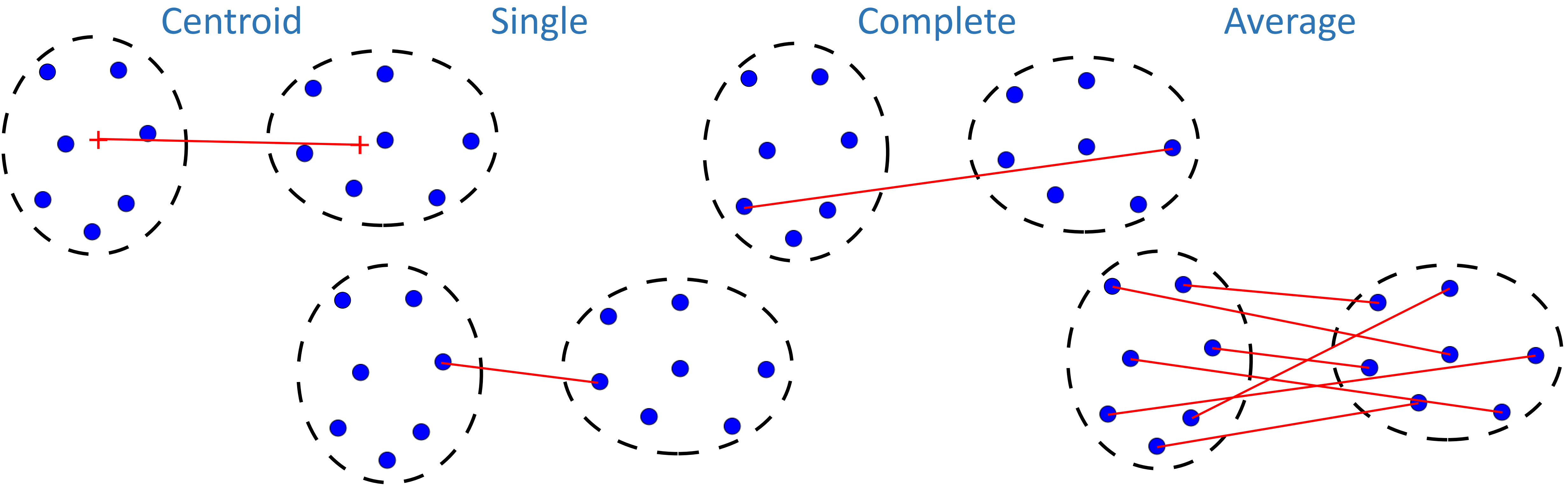
Cons

- What is *k*?
- Only applicable to continuous data where a mean is defined
- No guarantee of a global optimum solution

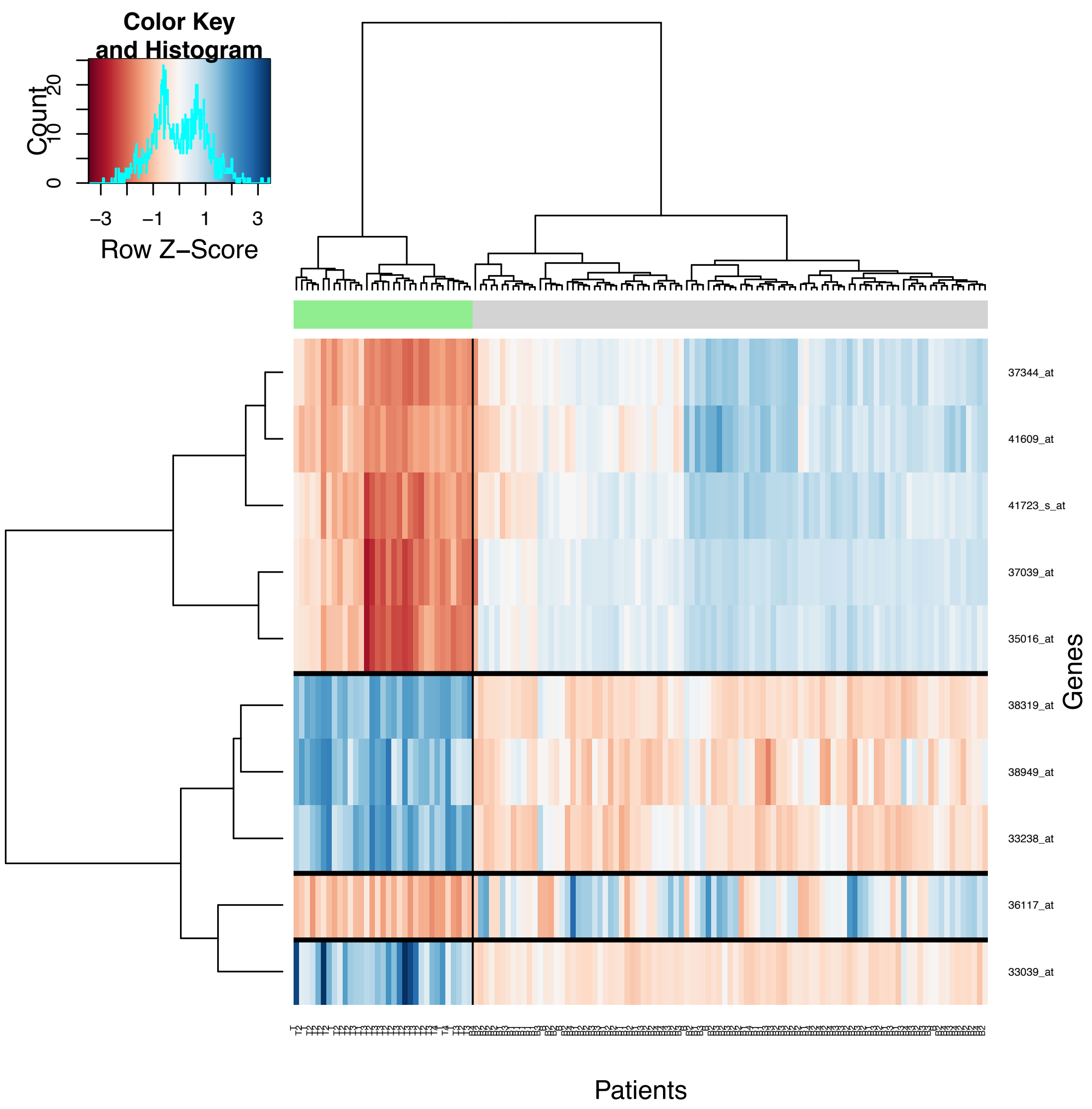
See also, fuzzy c-means algorithm

Hierarchical clustering algorithms

- Agglomerative clustering is the most common
- Measures distance through a linkage function



Hierarchical clustering algorithms



Hierarchical clustering algorithms

Pros

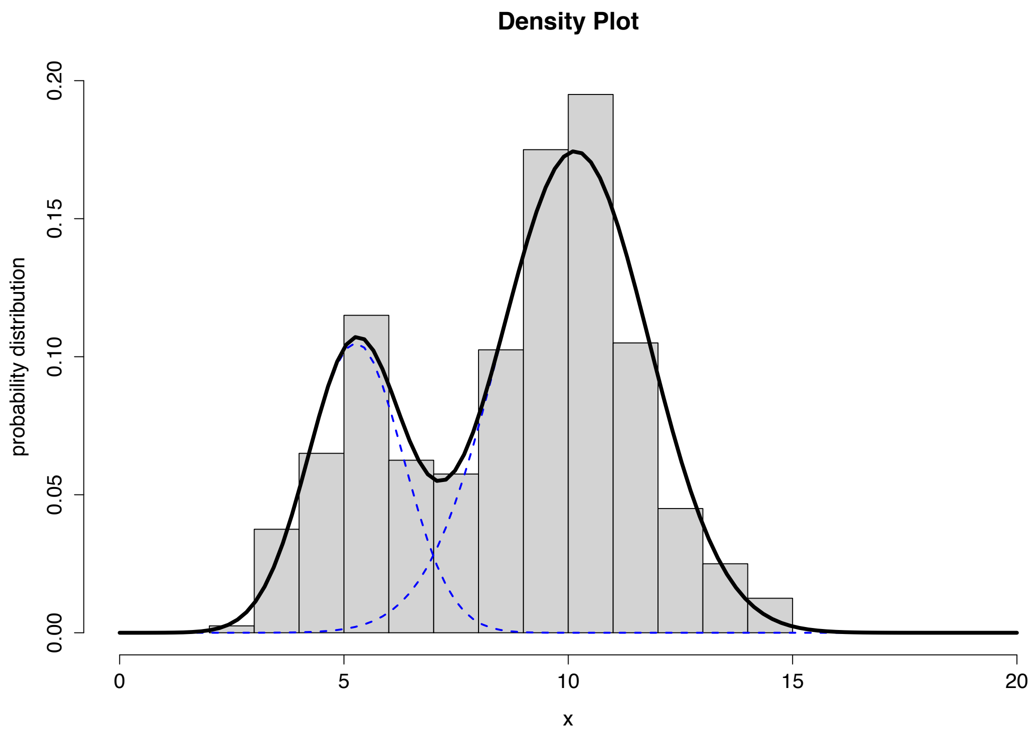
- No need to specify *k*
- Results can be visualised nicely irrespective of number of dimensions
- Sub-groups within larger clusters can be easily identified

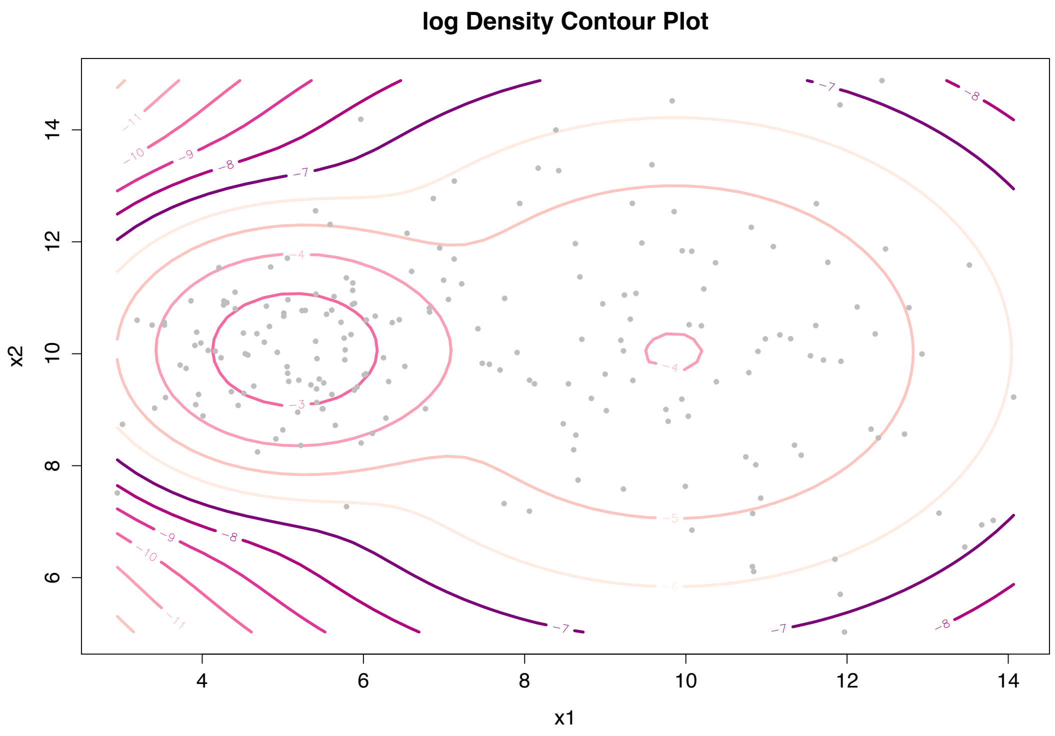
Cons

- Can be computationally expensive
- Interpretation is subjective. Where should we draw the line (to separate clusters)?
- Choice of distance method and linkage function can significantly change the result

Gaussian Mixture Models

- Fitting *k* multivariate Gaussian distributions to explain clusters
- Distance is calculated with Expectation-Maximisation (EM) algorithm
- Every point is part of every cluster with varying levels of membership





Gaussian Mixture Models

Pros

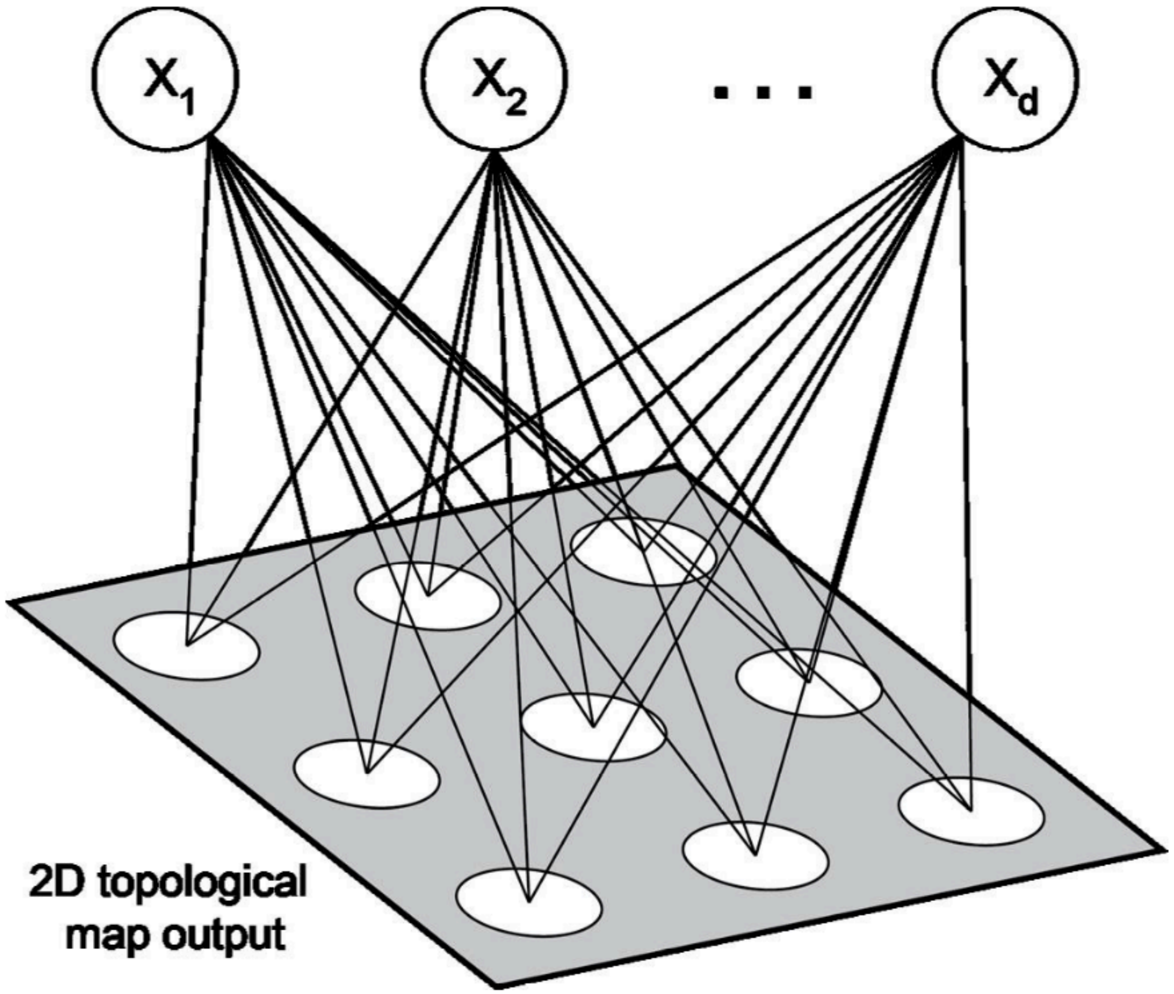
- Intuitive interpretation
- Computationally inexpensive

Cons

- Unknown *k*
- Assumption of normality
- No guarantee of a global optimum solution
- Fails when number of features is much greater than observations

Self-Organizing Maps

- Akin to an unsupervised artificial neural network
- Constrained version of *k*-means using topology to create a 2D map
- Topology uses vector quantisation and vector similarity to map multi-dimensional space
- Each resulting node has a membership for each input feature



Self-Organizing Maps

Pros

- Allows non-linear clustering
- Easy visualisation
- Acts to reduce dimensionality

Cons

- User defined topologies are highly subjective
- Number of nodes in 2D map can be iterative
- Not the most computationally efficient method

How many clusters are required?

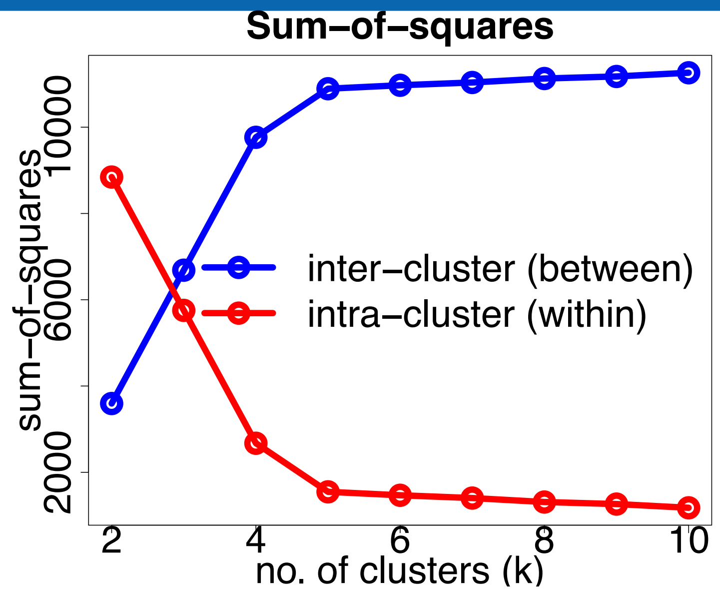
Determining the ‘correct’ number of clusters is essentially impossible

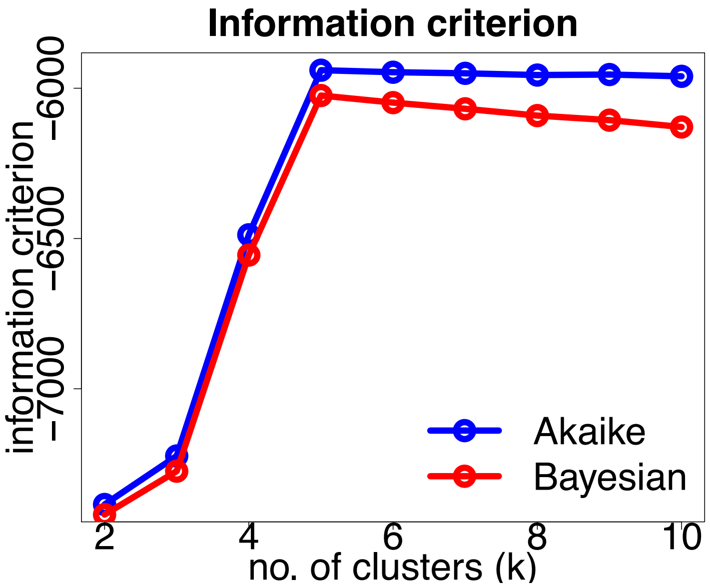
With unlabelled data, *k* is ambiguous

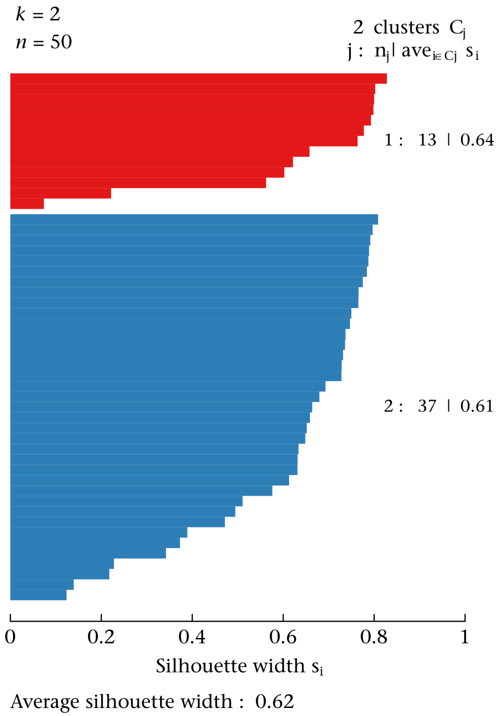
Some statistical methods are available to guide the decision:

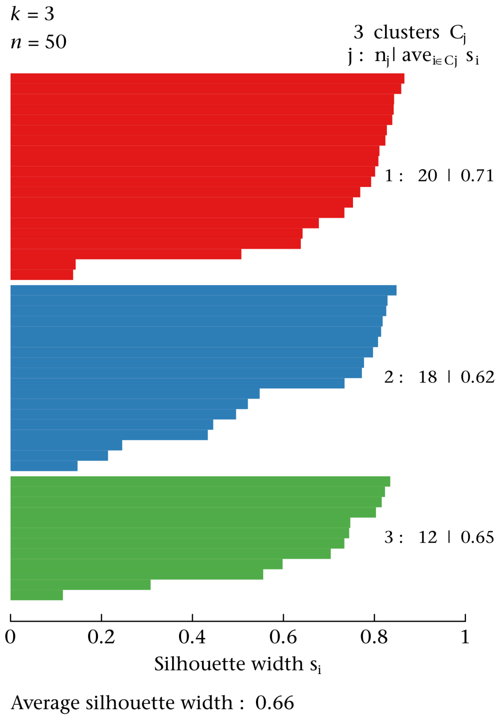
- Sum-of-squares distances
- Akaike’s Information Criterion / Bayesian Information Criterion
- Silhouette plots

How many clusters are required?









How many clusters are required?

- These methods are a guide
- Are the clusters practically relevant?
- Do they make sense?
- Using prior knowledge is not just acceptable, it is necessary

E.g how many different phenotypes are you expecting in your population?

E.g can you separate intrusive rocks from their extrusive equivalents?