Machine Learning

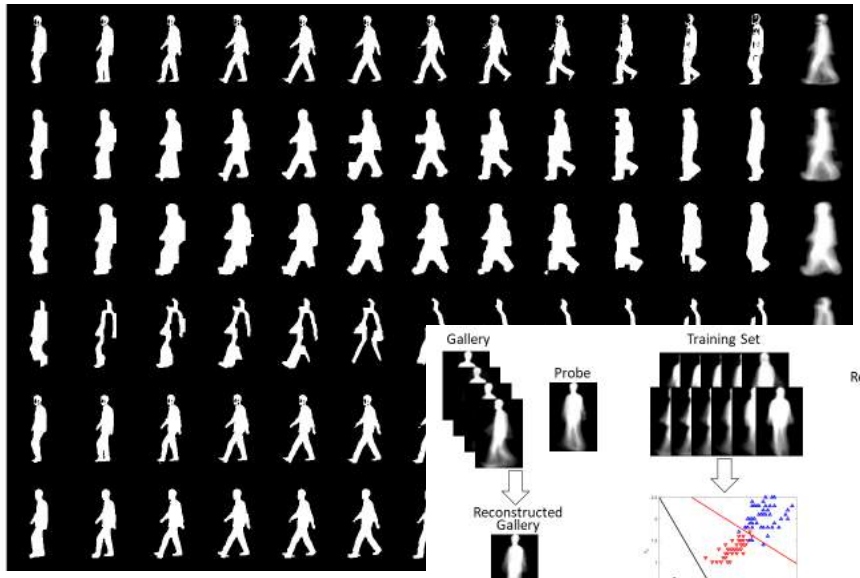
Lecture 8 – Clustering

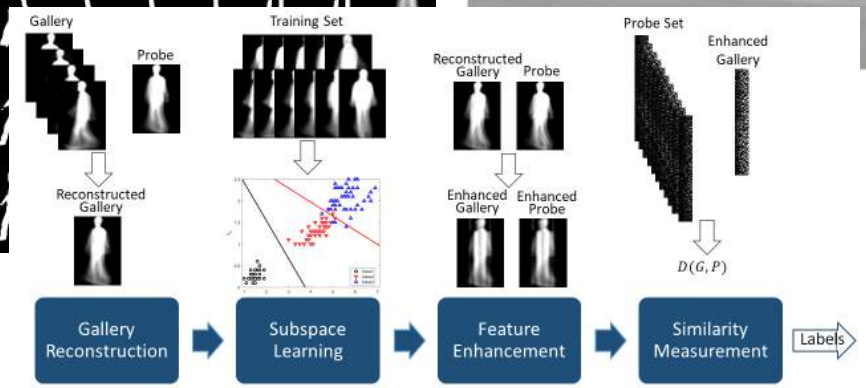
Dr JIA Ning

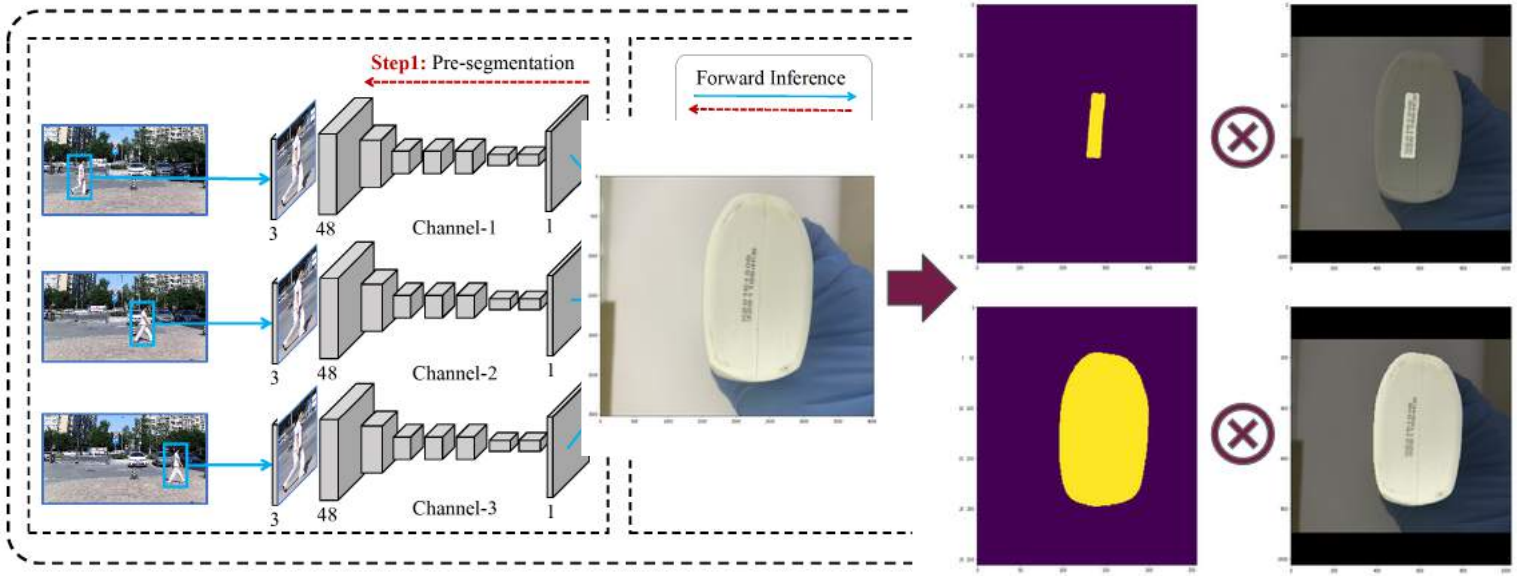
Dr JIA Ning

Dr JIA Ning

- PhD at Warwick
- Post Doc since Oct 2017
- Research interest:
  - Machine Learning
  - Pattern Recognition
  - Image Processing
  - Semantic Segmentation
  - Biometrics
- Models:
  - Convolutional Neural Networks
  - Subspace Learning







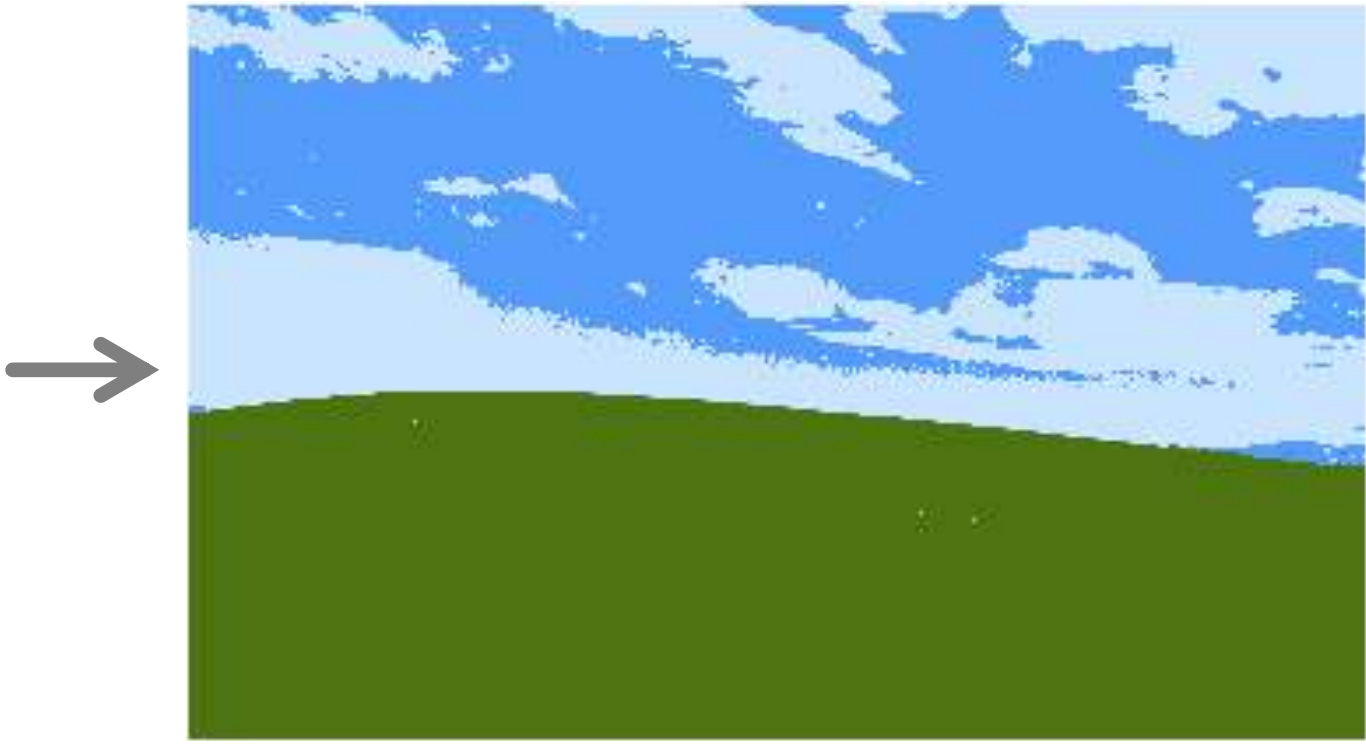
Today

- What is Clustering?
- K-means Algorithm
- Other Clustering Algorithms

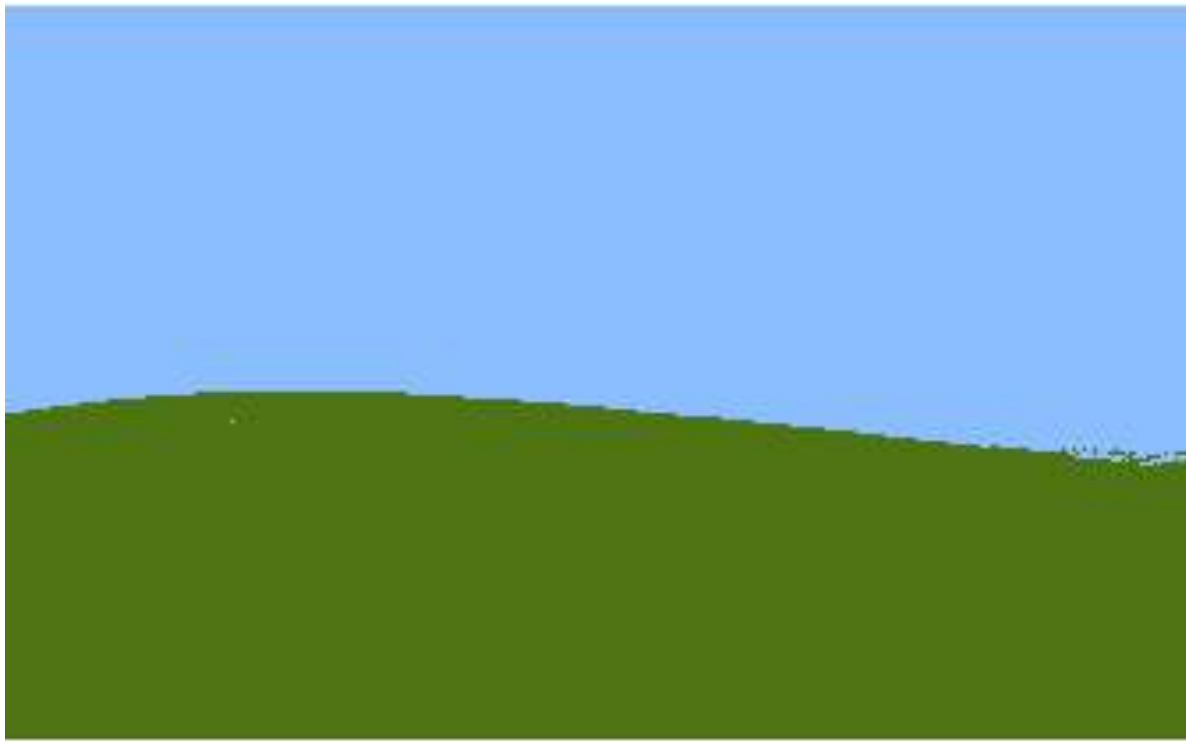
What is Clustering?


How do we recognise this ~~world~~ picture

How do we recognise this ~~world~~ picture

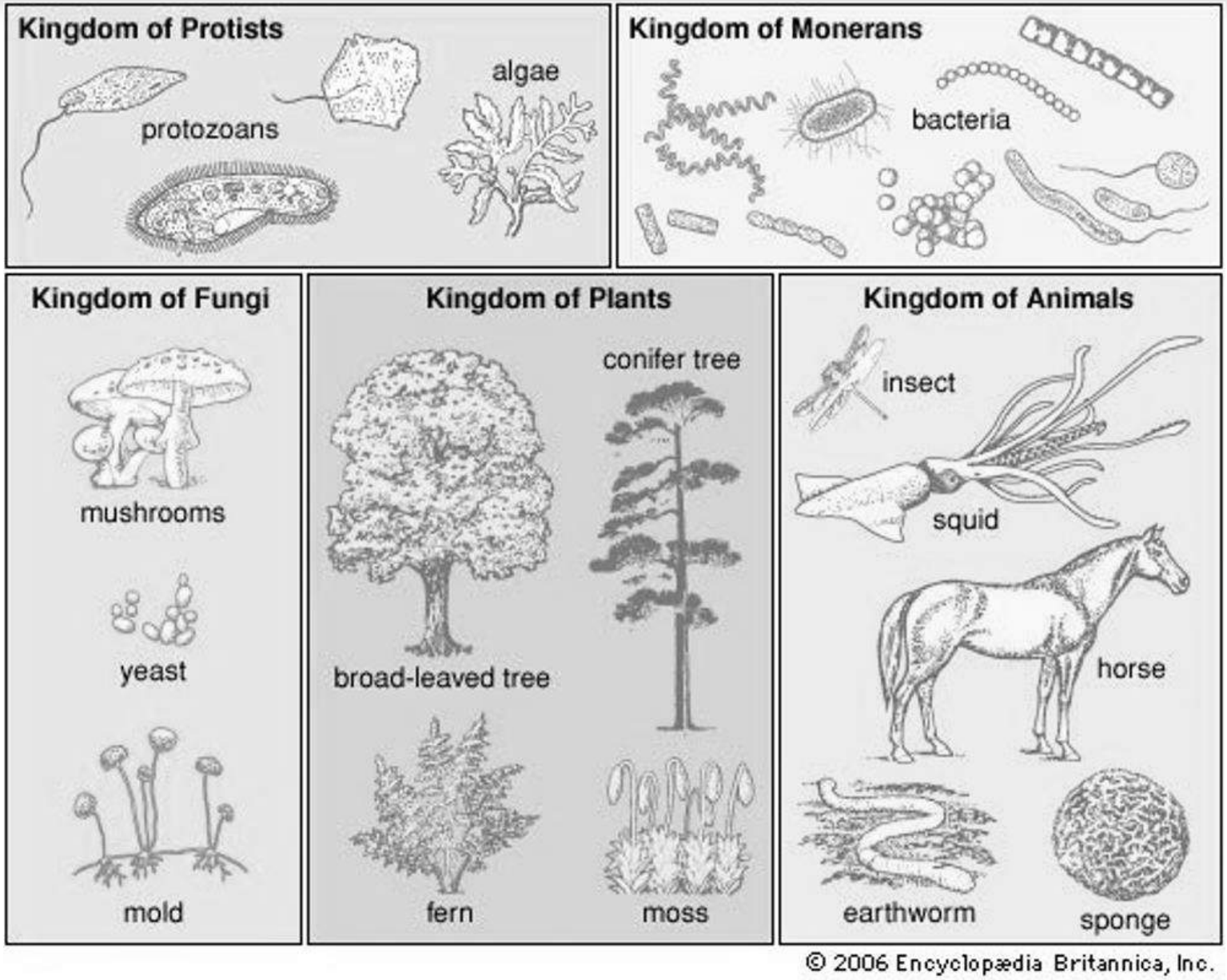


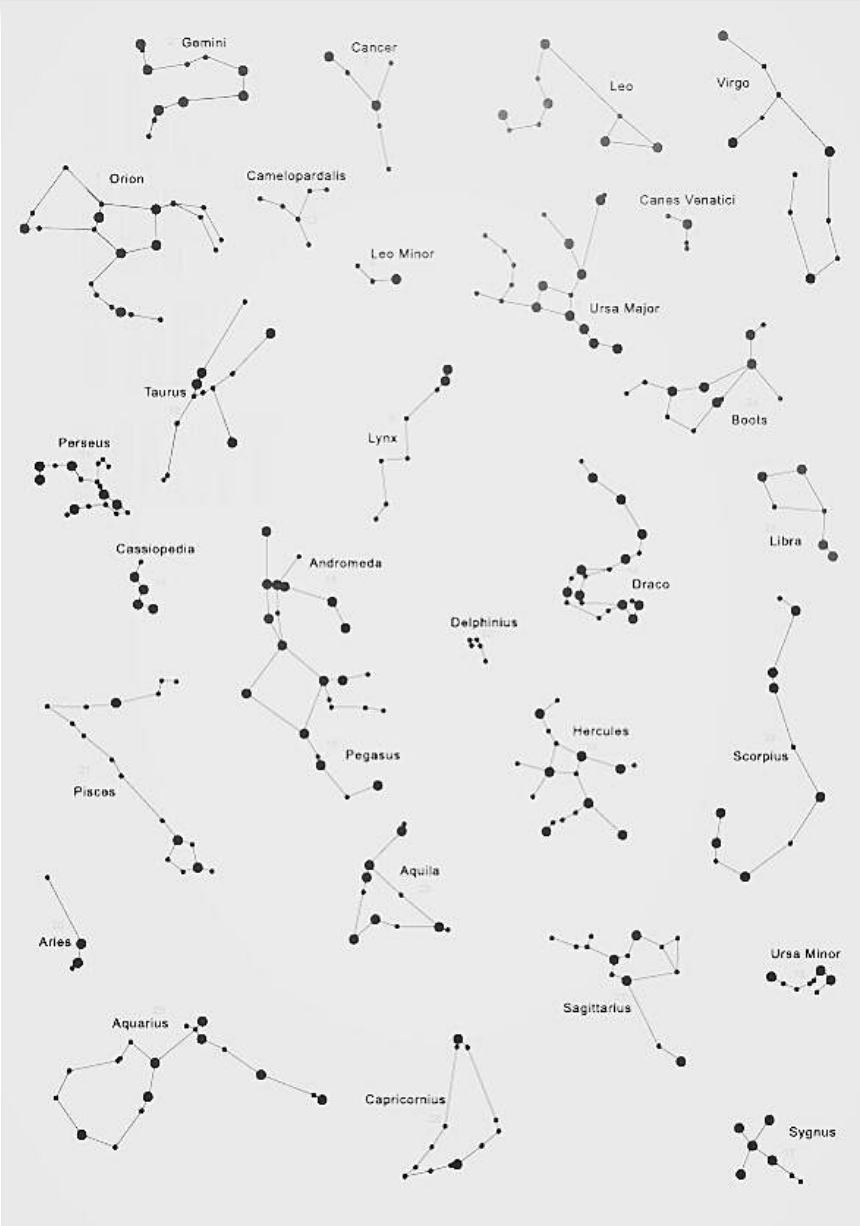
How do we recognise this ~~world~~ picture



How do we recognise this ~~world~~ picture

How do we recognise this world





Reasons for clustering

- Effectively organising a large data set such that it is easy to understand
- We are in the ‘big data era’ – the data is beyond our capability to memorise and analyse
- A well organised and summarised dataset can be used for prediction

> Clustering refers to numerical methods of classification.

The aim is to provide objective and stable classifications for a specific task.

“

A cluster is described in terms of internal homogeneity and external separation.

”

A set of input patterns $\mathrm{X} = \{\boldsymbol{x}_i\}, i = 1, \ldots N$, with n observations of d dimension $\boldsymbol{x}_i \in \mathbb{R}^d$, is to be explained by a fixed number K of homogeneous groups

Proximity Measurement (Continuous Data)

A distance function on a dataset $\boldsymbol{X}$ is called a metric if it satisfies the following conditions:

- Symmetry, $\mathrm{D}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathrm{D}(\boldsymbol{x}_j, \boldsymbol{x}_i)$
- Positivity, $\mathrm{D}(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0 \; for \; all \; \boldsymbol{x}_i \; and \; \boldsymbol{x}_j$
- Triangle equality, $\mathrm{D}(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq \mathrm{D}(\boldsymbol{x}_i, \boldsymbol{x}_k) + \mathrm{D}(\boldsymbol{x}_k, \boldsymbol{x}_j)$
- Reflexivity, $\mathrm{D}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \; if \; \boldsymbol{x}_i = \boldsymbol{x}_j$

A common metric for hyper-spherical clusters is Euclidean distance (L_2 norm):

$$\mathrm{D}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{l=1}^{d} \left(x_{il} - x_{jl}\right)^2}$$

Clustering Criteria

The most commonly used clustering criteria for the $N \times d$ matrix X of continuous data is given as:

$$T = \sum_{k=1}^{K} \sum_{n=1}^{N_k} (\boldsymbol{x}_{kn} - \boldsymbol{\mu})(\boldsymbol{x}_{kn} - \boldsymbol{\mu})^{\mathsf{T}}$$

where $\boldsymbol{x}_{kn}$ is the d-dimensional vector of observations of the nth object in group k, and $\boldsymbol{\mu}$ is the vector of overall sample means.

Partition T into within-group scatter matrix W and between-group scatter matrix B, such that $T = W + B$, W is defined as:

$$W = \sum_{k=1}^{K} \sum_{n=1}^{N_k} (\boldsymbol{x}_{\mathrm{kn}} - \boldsymbol{\mu}_k)(\boldsymbol{x}_{kn} - \boldsymbol{\mu}_k)^{\mathsf{T}}$$

where $\boldsymbol{\mu}_k$ is the mean of all the data points $\boldsymbol{x}_n$ assigned to cluster k.

Objective Function

Decrease of W is equivalent to increase of B.

Thus our **objective function** is given as:

$$J = \sum_{k=1}^{\mathrm{K}} \sum_{n=1}^{N_k} r_{nk} (\boldsymbol{x}_{\mathrm{n}} - \boldsymbol{\mu}_k)^2 \qquad (1)$$

where

$$r_{nk} = \begin{cases} 1 & if\ k = \arg\min_j (\boldsymbol{x}_{\mathrm{n}} - \boldsymbol{\mu}_j)^2 \\ 0 & otherwise \end{cases} \qquad (2)$$

To find values for r_{nk} and $\boldsymbol{\mu}_k$ to minimize $\boldsymbol{J}$, we can perform an iterative procedure in which each iteration involves two successive steps corresponding to **successive optimizations** with respect to r_{nk} and $\boldsymbol{\mu}_k$.

Today

- What is Clustering?
- K-means Algorithm
- Other Clustering Algorithms

K-means Algorithm

K-means Algorithm

Optimisation algorithm:

1. Initialise $\{\boldsymbol{\mu}_k\}$, $\mathrm{k}\epsilon 1, \ldots, K$, then keep $\boldsymbol{\mu}_k$ fixed and minimise E with respect to r_{nj}
2. Assign r_{nj} to each data point $\boldsymbol{x}_\mathrm{n}$ based on equation (2), then keep r_{nj} fixed and minimise $\boldsymbol{J}$ with respect to $\boldsymbol{\mu}_k$
3. Repeat step 1 and 2 until **convergence**

K-means Algorithm

Proof

The derivate of $\boldsymbol{E}$ with respect to $\boldsymbol{\mu}_j$ is:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} J = \sum_{k=1}^{\mathrm{K}} \sum_{n=1}^{N_k} \frac{\partial}{\partial \boldsymbol{\mu}_k} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^2$$

$$= 2 \sum_{n=1}^{\mathrm{N_j}} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)$$

The function is converged when

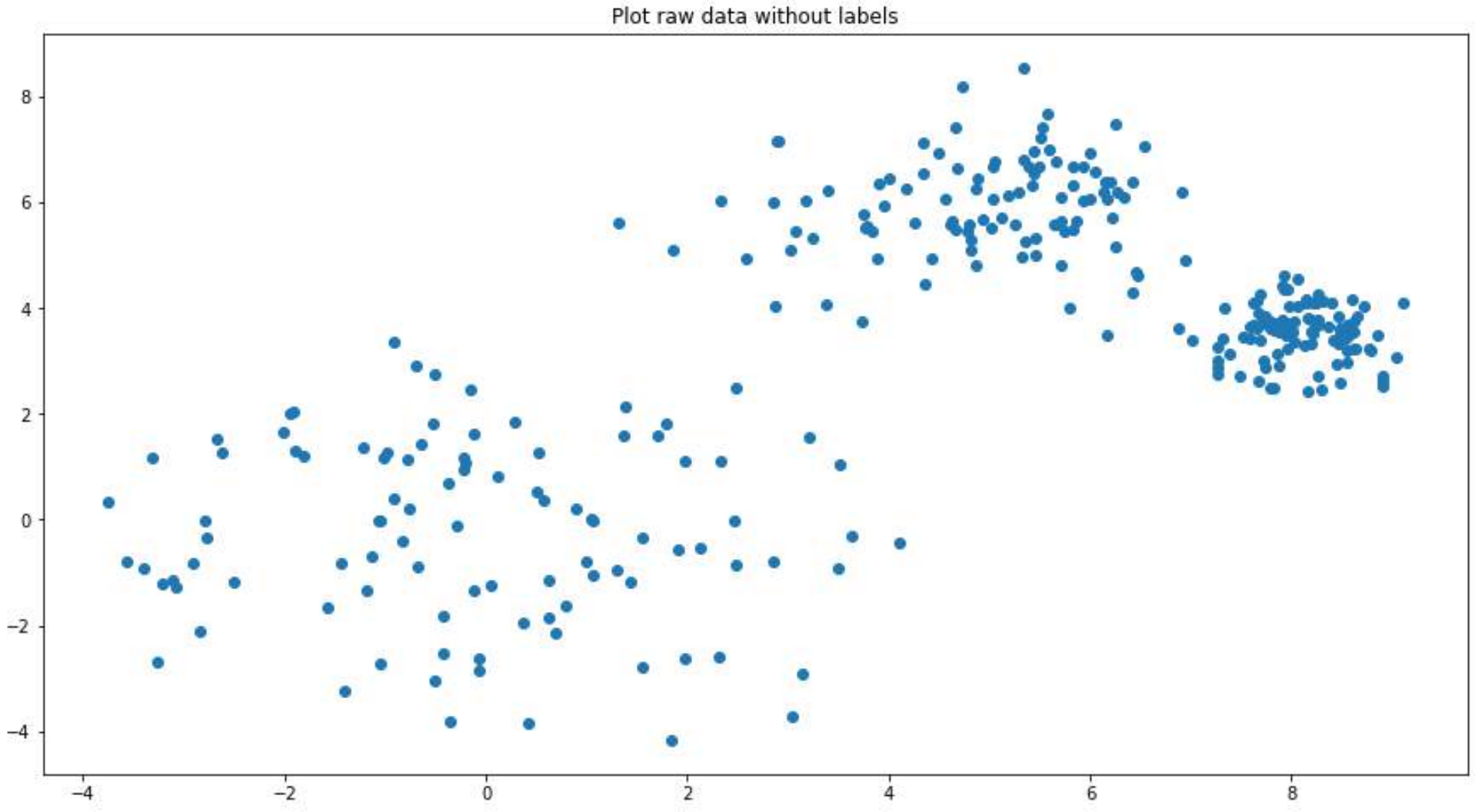
$$2 \sum_{n=1}^{\mathrm{N_j}} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0$$

and we got

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{\mathrm{N}_k} \boldsymbol{x}_n$$

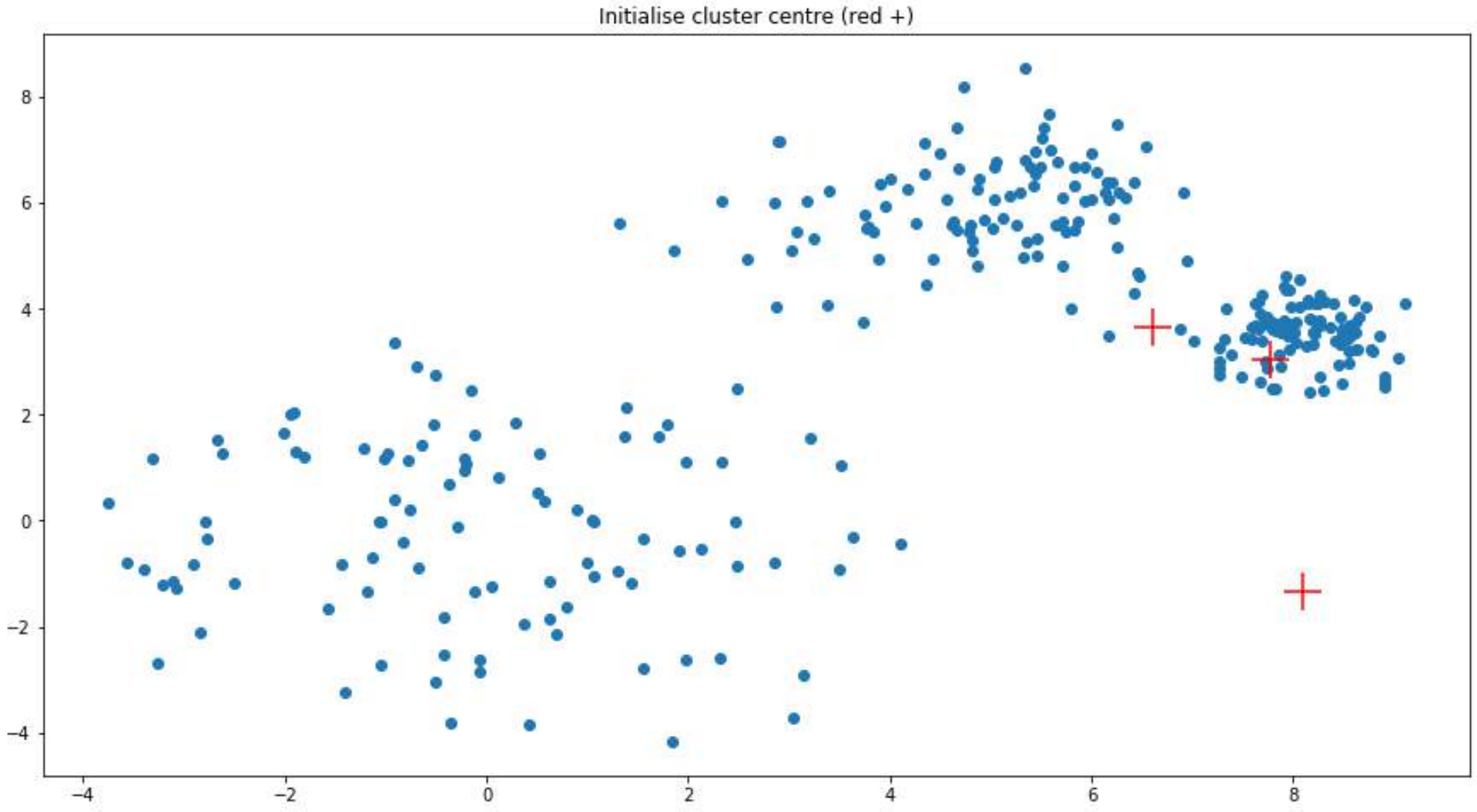
K-means Algorithm

Demonstration by Steps - Visualise data distribution



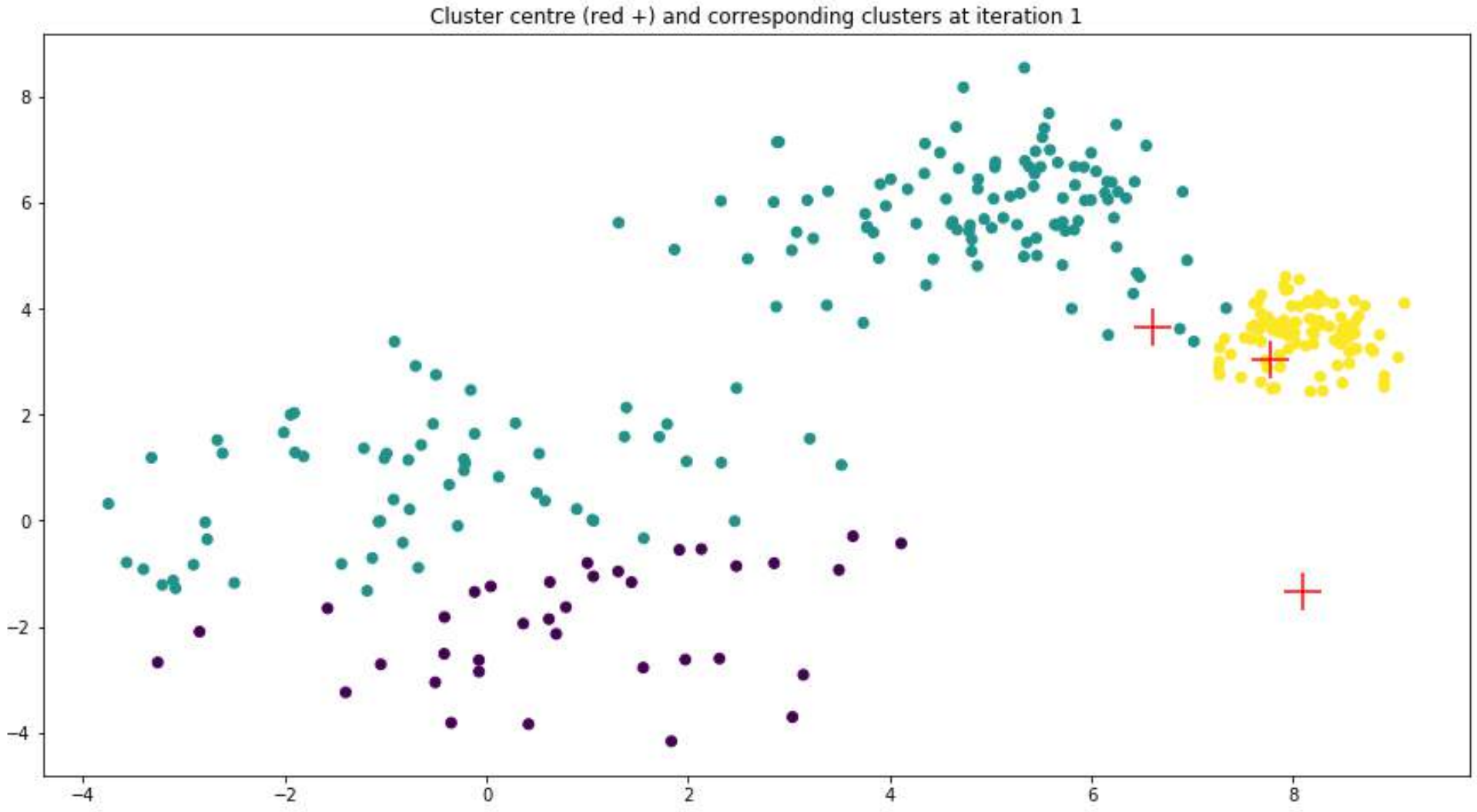
K-means Algorithm

Demonstration by Steps - Initialise cluster centres μk (randomly)



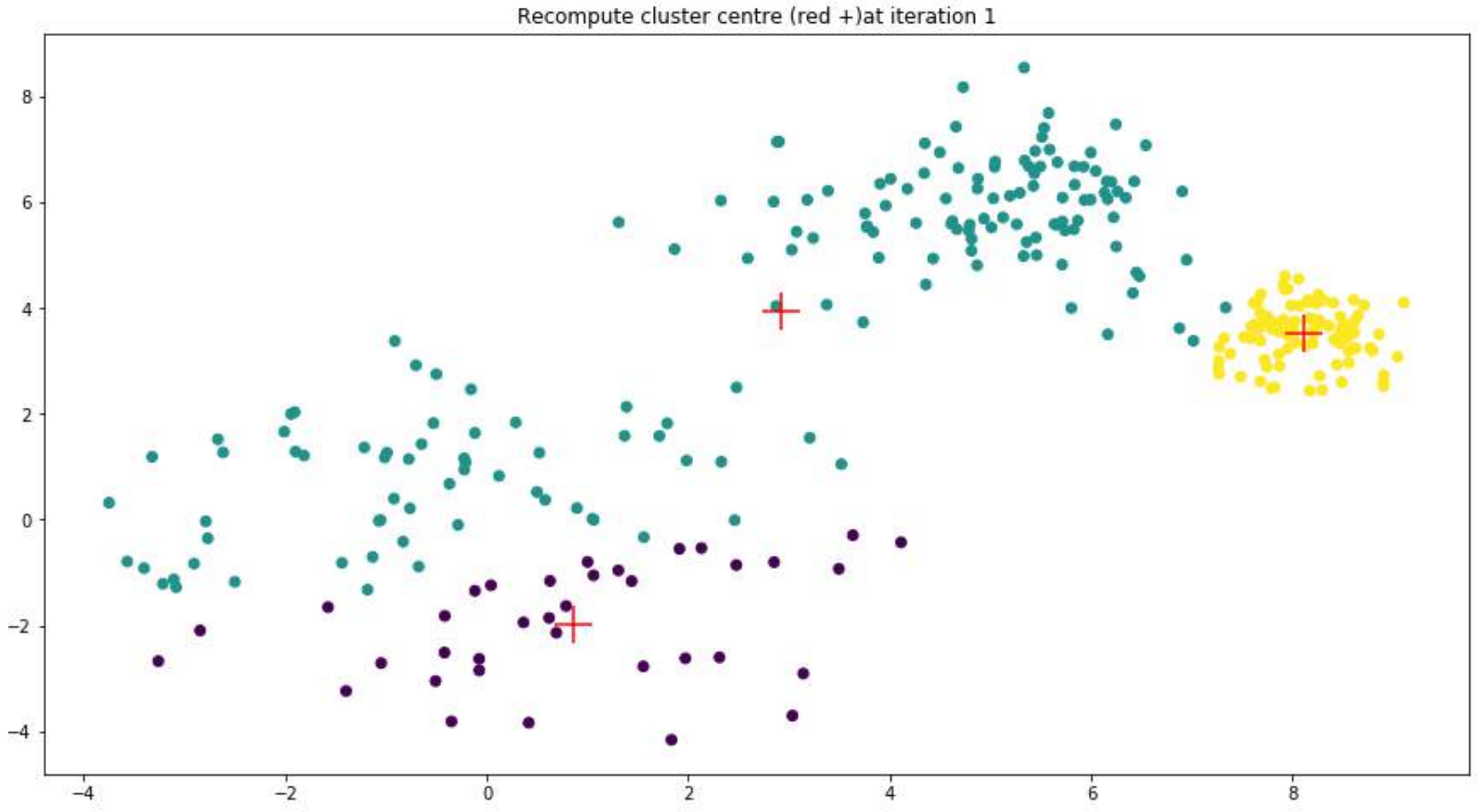
K-means Algorithm

Demonstration by Steps - Assign xn to its nearest cluster centre μk



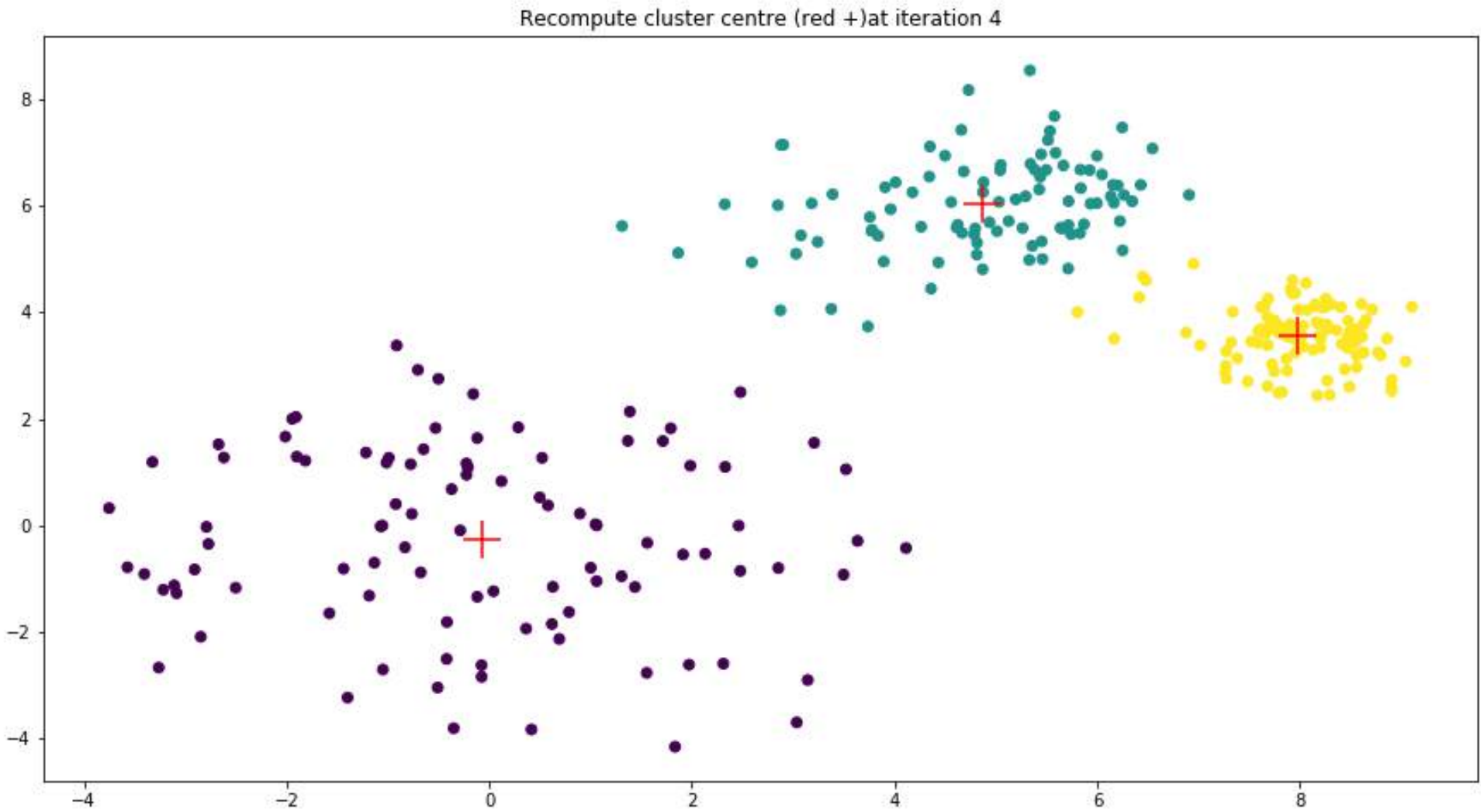
K-means Algorithm

Demonstration by Steps - Recompute centre μk



K-means Algorithm

Demonstration by Steps - Until convergence



K-means Algorithm

K-means++ Initialisation

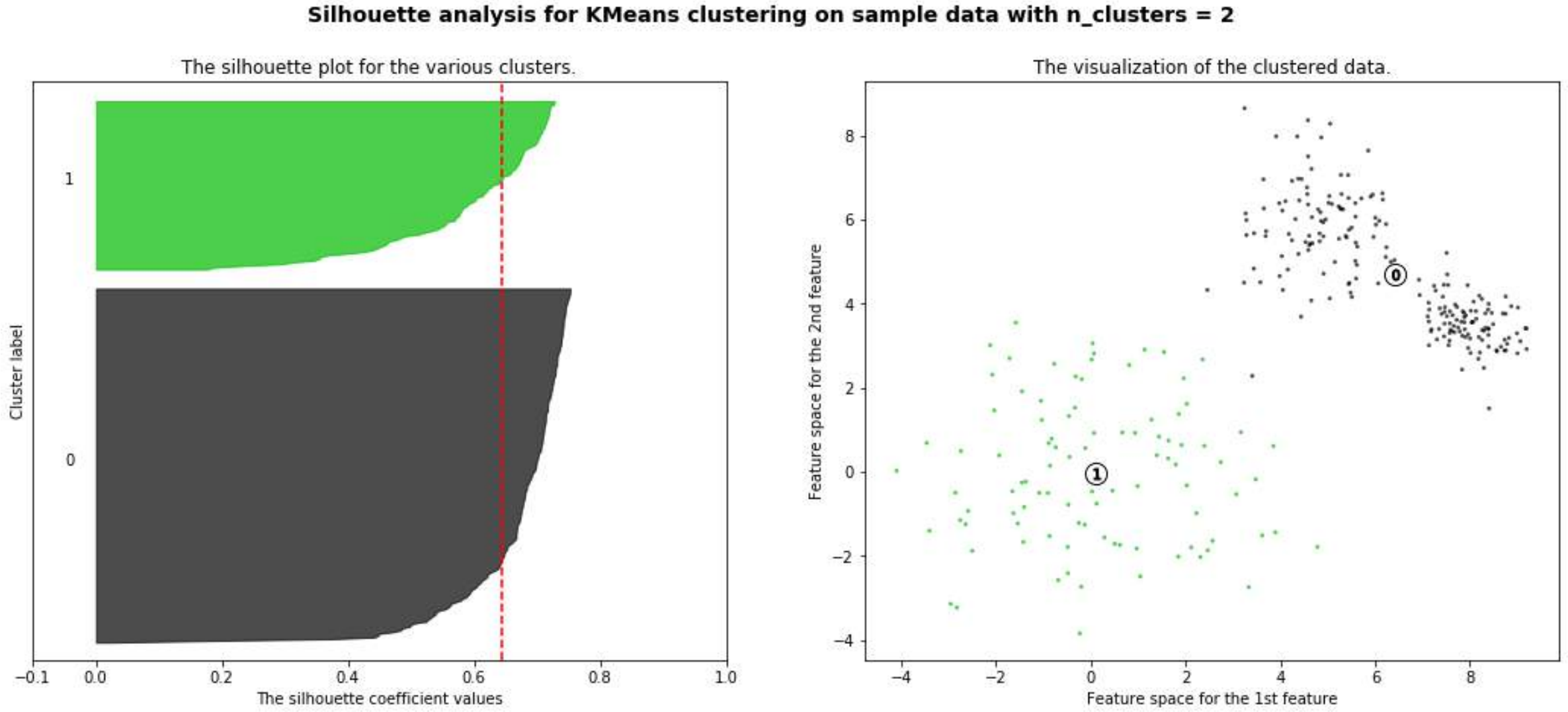
1. Choose one centre uniformly at random from among the data points
2. For each data point $\boldsymbol{x}$, compute $\mathrm{D}(\boldsymbol{x})$, the distance between x and the nearest centre that has already been chosen
3. Choose one new data point at random as a new centre, using a weighted probability distribution where a point $\boldsymbol{x}$ is chosen with probability proportional to $\mathrm{D}(\boldsymbol{x})^2$
4. Repeat Steps 2 and 3 until k centres have been chosen

K-means Algorithm

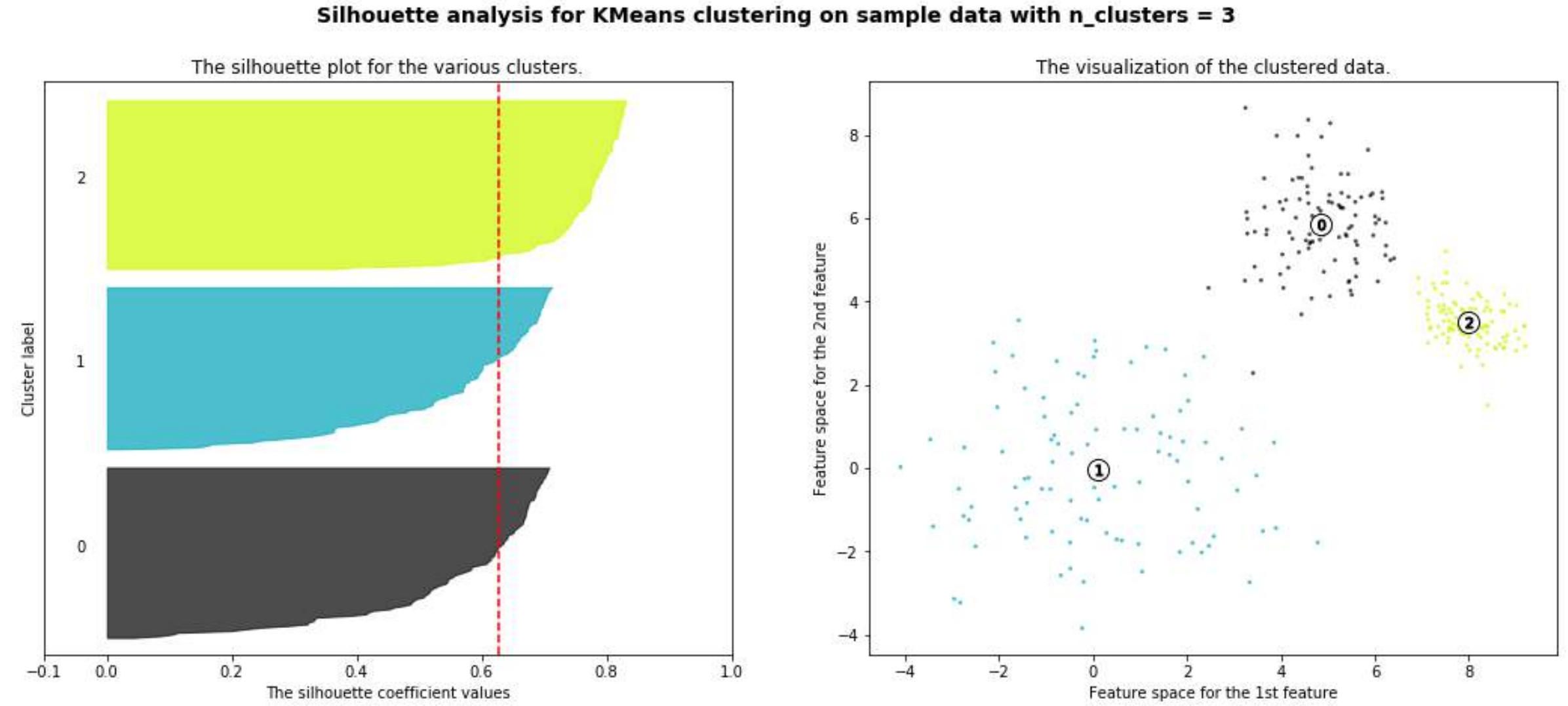
Estimating the Number of Clusters - Silhouette analysis

- Study the separation distance between the resulting clusters
- The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters
- This measure has a range of [-1, 1].

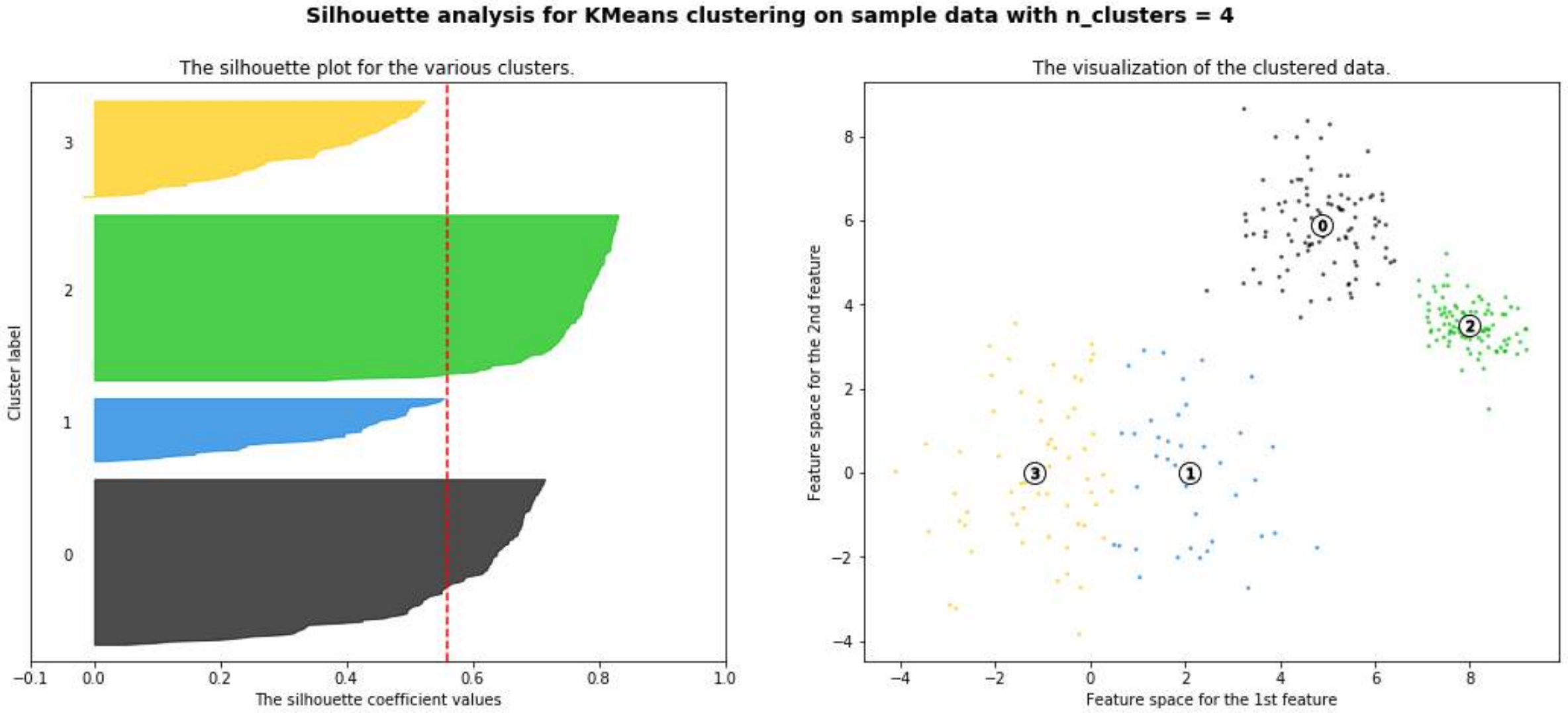
Illustrate Silhouette Analysis



Illustrate Silhouette Analysis

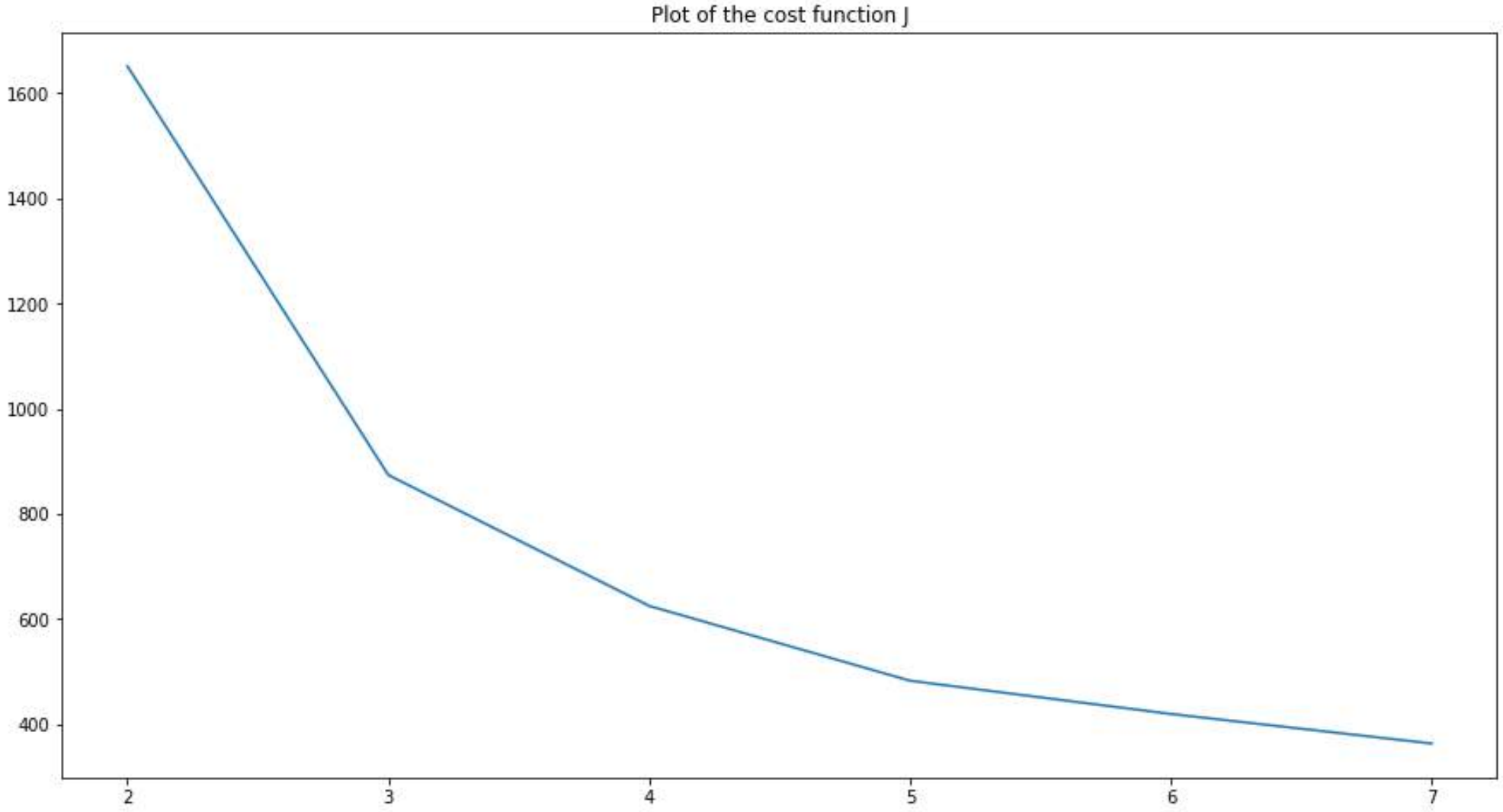


Illustrate Silhouette Analysis



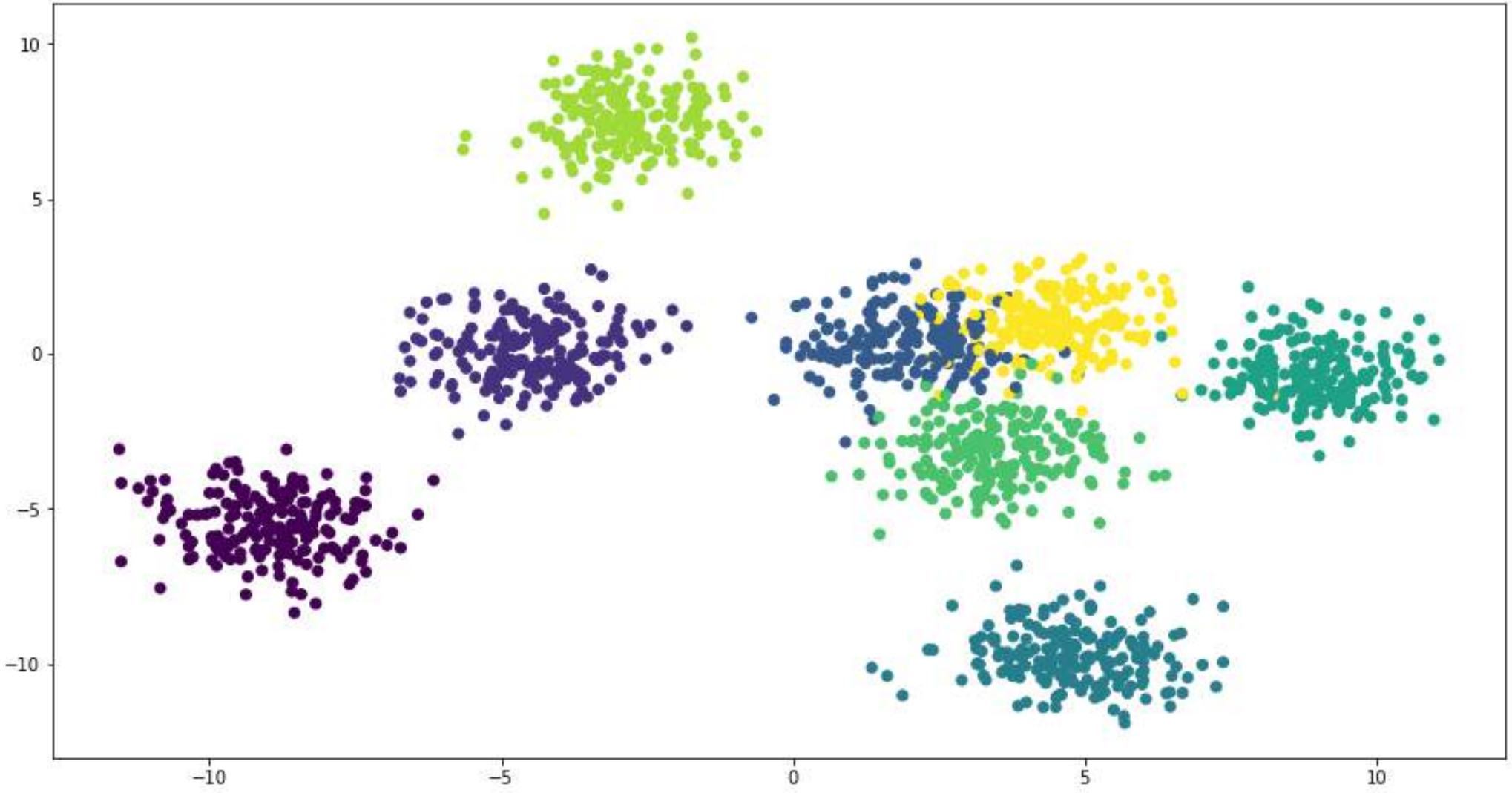
K-means Algorithm

Plot of the Cost Function J Over Different K Values



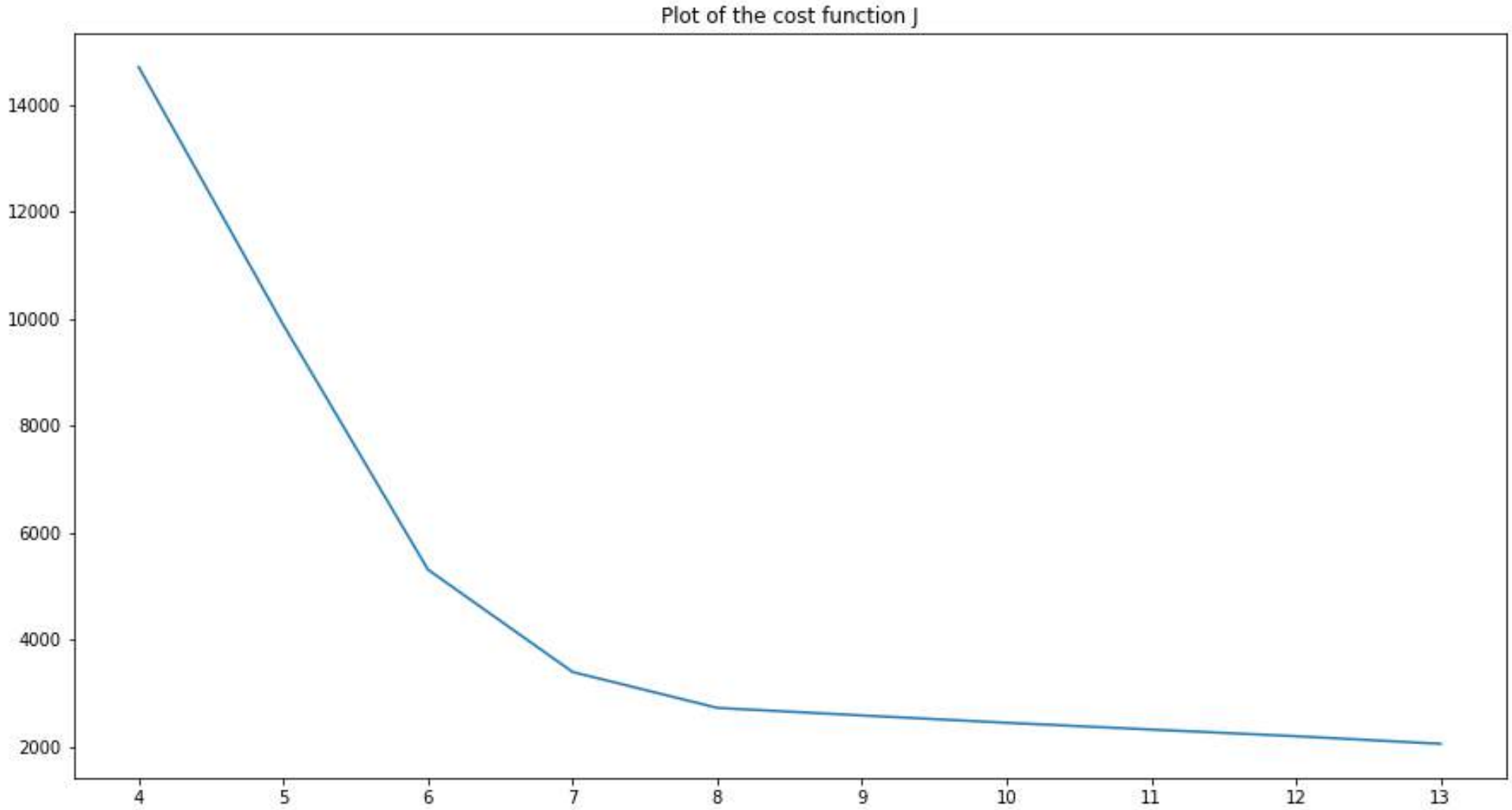
K-means Algorithm

Plot of the Cost Function J Over Different K Values



K-means Algorithm

Plot of the Cost Function J



K-means Algorithm

E.g. Colour Image Compression

- Using less colour while preserving the overall appearance quality

Original image (96,615 colours)

Quantised image (64 colours, K-Means)

K-means Algorithm

E.g. Colour Image Compression

- Using less colour while preserving the overall appearance quality

Original image (96,615 colours)

Quantised image (64 colours, K-Means)

K-means Algorithm

E.g. Colour Image Compression

- Using less colour while preserving the overall appearance quality

Quantised image (64 colours, K-Means)

Quantised image (64 colours, Random)

K-means Algorithm

Data Preparation (Pre-processing)

- Normalize the data to make each feature contribute equally to the distance (feature scaling)
  - Standardisation (Z-score normalisation):

$$x_i' = \frac{x_i - \bar{x}}{\sigma}$$

  - Rescaling

$$x_i' = \frac{x_i - min(x)}{max(x) - min(x)}$$

K-means Algorithm

Summary

- Initialise cluster means and assign data points to clusters
- Iteratively repeat the two phases computation:
  - Re-assigning data points to clusters
  - Re-computing the cluster means
- Until there is no further change in the assignments (or until some maximum number of iterations is exceeded)

K-means Algorithm

Issues

- Convergence
- The number of K
- Robustness
- Extension of the Definition of Means

Today

- What is Clustering?
- K-means Algorithm
- Other Clustering Algorithms

Other Clustering Algorithms

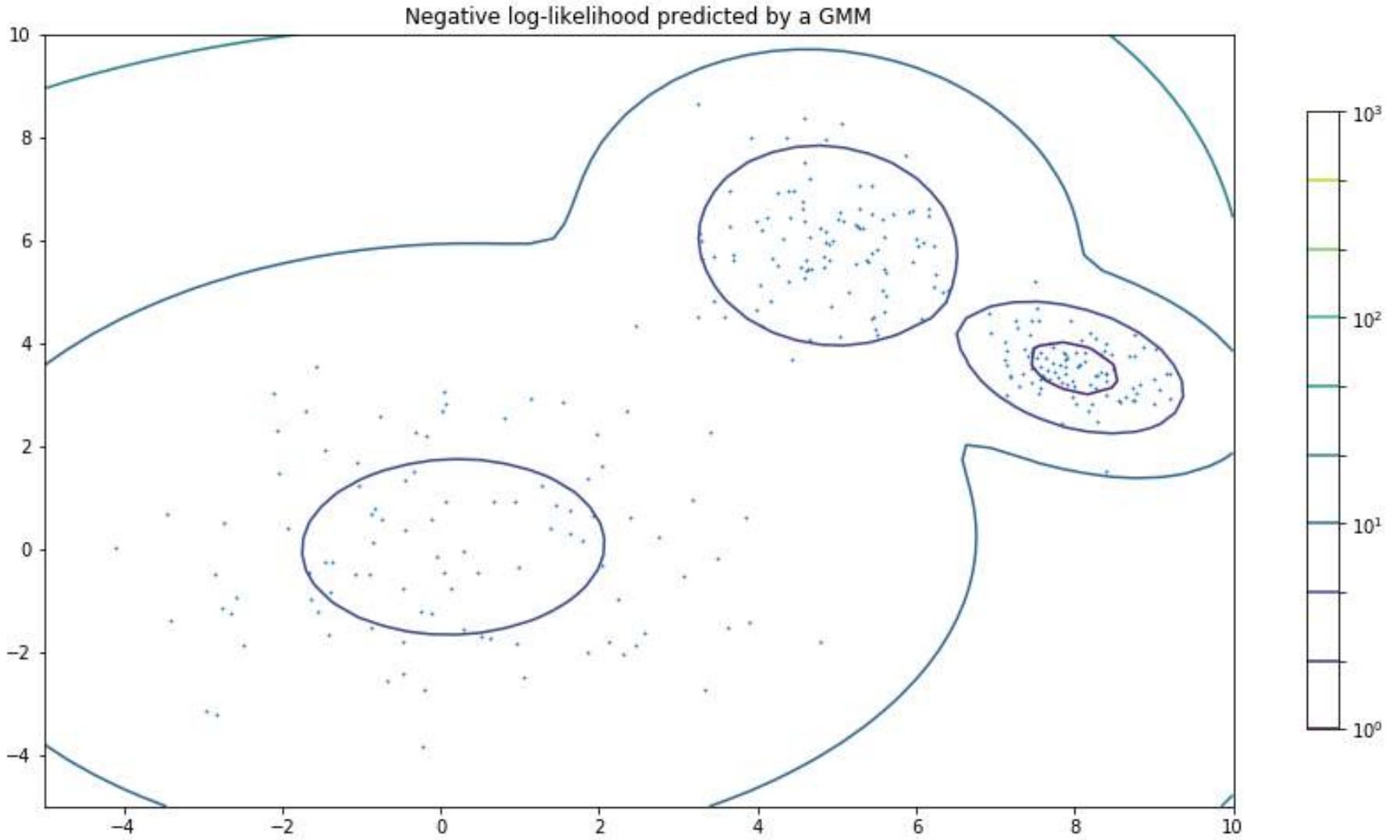
Other Clustering Algorithms

Gaussian Mixture Model (GMM)

- Each data point follows one of the K underlying probability distributions
- These probability sources can take different functional forms
- Assume the form of the mixture densities are known
- The objective of finding clusters is equivalent to estimating the parameters of the K underlying models

Other Clustering Algorithms

Gaussian Mixture Model (GMM)



Other Clustering Algorithms

Clustering the Clustering Algorithms

- **Partitional clustering:**
  - directly divides data points into some prespecified number of clusters without the hierarchical structure
- **Hierarchical clustering:**
  - groups data with a sequence of nested partitions, either from singleton clusters to a cluster including all individuals (agglomerative hierarchical clustering) or vice versa (divisive hierarchical clustering).

Other Clustering Algorithms

Agglomerative Clustering

- Start with N singleton clusters. Calculate the proximity matrix (usually based on the distance function) for the N clusters
- In the proximity matrix, search the minimal distance D, where $D(\cdot,\cdot)$ is the distance function discussed later in the section, and combine cluster C_i and C_j to form a new cluster C_{ij}
- Update the proximity matrix by computing the distances between the cluster C_{ij} and the other clusters
- Repeat steps 2 and 3 until only one cluster remains.

Other Clustering Algorithms

Divisive Clustering

- Compared to agglomerative hierarchical clustering, divisive clustering proceeds in the opposite way.
- In the beginning, the entire data set belongs to a cluster, and a procedure successively divides it until all clusters are singletons.
- For a dataset with N objects, a divisive hierarchical algorithm would start by considering $2^{\mathrm{N}-1}-1$ possible divisions of the data into 2 nonempty subsets.
- However, the divisive clustering algorithms do provide clearer insights into the main structure of the data, since the larger clusters are generated at the early stage the less likely it suffers from the accumulated erroneous decisions, which cannot be corrected by the successive process

Today

- What is Clustering?
- K-means Algorithm
- Other Clustering Algorithms

Clustering in Practice

Clustering in Practice

Market research

Astronomy

Psychiatry

Weather classification

Archaeology

Bioinformatics and genetics

Clustering - Summary

- The motivation of developing clustering algorithms:
  - Explores the unknown natures of the data that are integrated with little or no prior information
  - Saves effort of data labelling, which can be extremely expensive and time consuming
  - Provides a compressed representation of the data and is useful in large-scale data analysis

> In cluster analysis a group of objects is split up into a number of more or less homogeneous subgroups on the basis of an often subjectively chosen measure of similarity (i.e. chosen subjectively based on its ability to create "interesting" clusters), such that the similarity between objects within a subgroup is larger than the similarity between objects belonging to different subgroups.

Moreover, a different clustering criterion or clustering algorithm, even for the same algorithm but with different selection of parameters, may cause completely different clustering results.

> There is no universal clustering algorithm to solve all problems.

Therefore, it is important to carefully investigate the characteristics of a problem in order to select or design an appropriate clustering strategy.

Questions?