QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Data Analysis & Visualisation

CSC3062

BEng (CS & SE), MEng (CS & SE), BIT & CIT

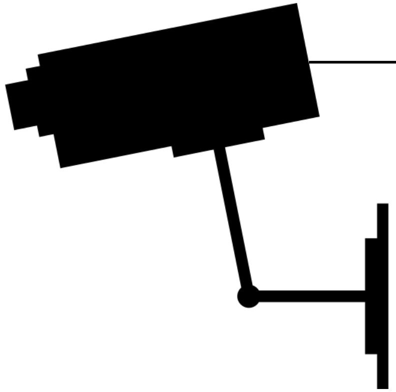
Dr Reza Rafiee

Semester 1 2019

Pattern recognition systems

Object(s)
(samples)

Set up a camera
(or transducer)



Measurement

Pre-processing → Feature extraction → Categorisation → **Salmon** / **Sea bass**

Pre-processing: Filtering, removing noise and so on

Feature extraction: Choosing the best features

Categorisation: Clustering
Classification

x_1: lightness
x_2: width

The aim is to partition the feature space into two regions

Feature space: two dimensions $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

Suppose that we measure the feature vectors for our samples

Reference: Pattern Classification, R. O. Duda, P. E. Hart, D. G. Stork; Wiley

Classification vs. clustering

Classification

- Objects/samples with classes
- Pre-processing
- Feature selection
- Find a decision boundary
- Final model

It's called model training or classifier training stage

Clustering

- Objects/samples with no classes
- Pre-processing
- Feature selection
- Feature space partitioning

Feature selection for high separability

Classification – training vs. prediction

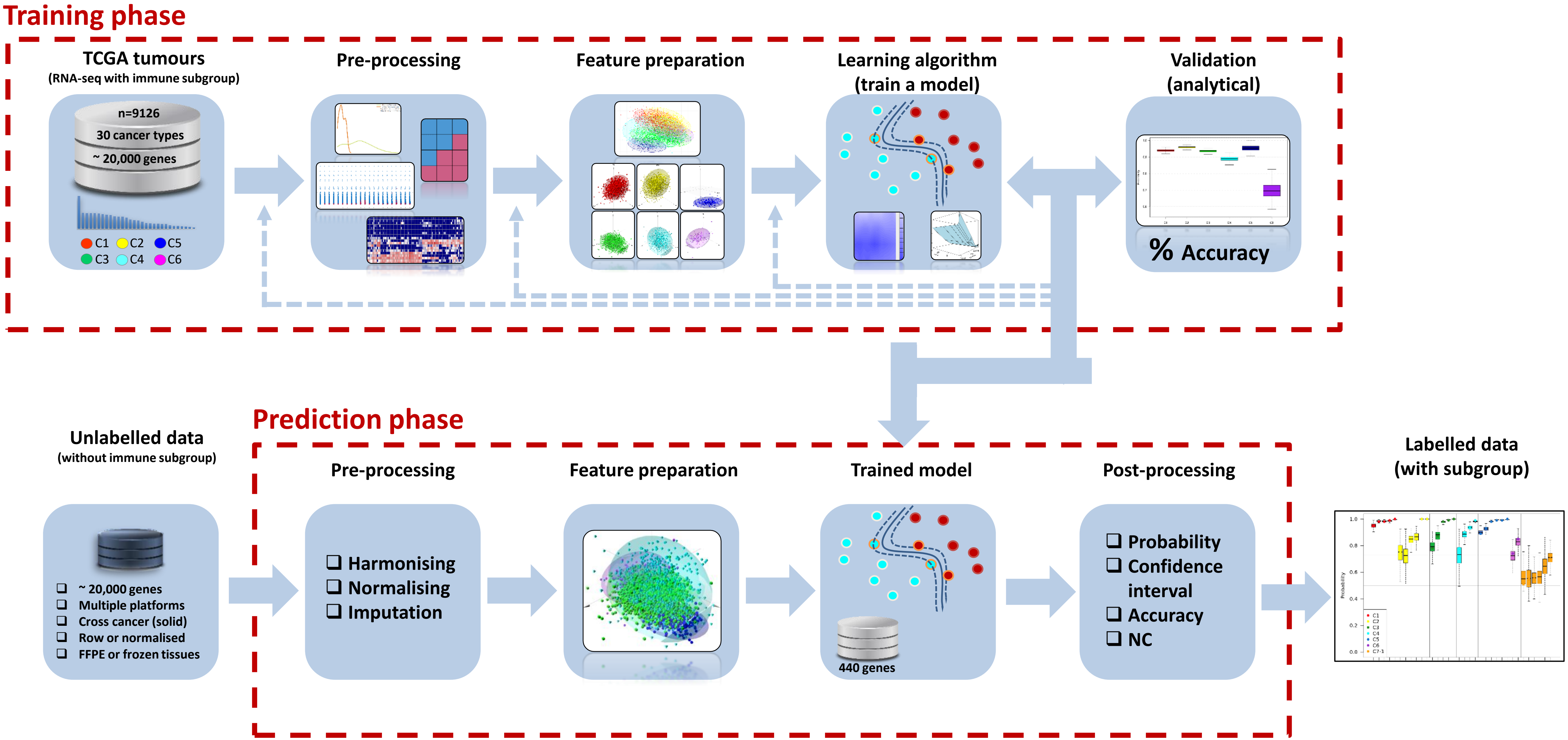


Figure 1.11 |Data pre-processing stage. a, Training phase. **b**, Prediction phase.

Unsupervised learning

Unsupervised clustering

- What is clustering?
- Why would we want to cluster?
- How would you determine clusters?
- How can you do this efficiently?

Clustering - concept

Basic idea: group together similar objects/samples/data

Organising unlabelled data into similar groups called clusters

Clustering or grouping

Cluster analysis or clustering is the task of grouping/partitioning a set of instances/objects/data points in such a way that data points in the same group are more similar to each other than to those in other groups

Clustering - concept

What could "similar" mean?

Clustering - concept

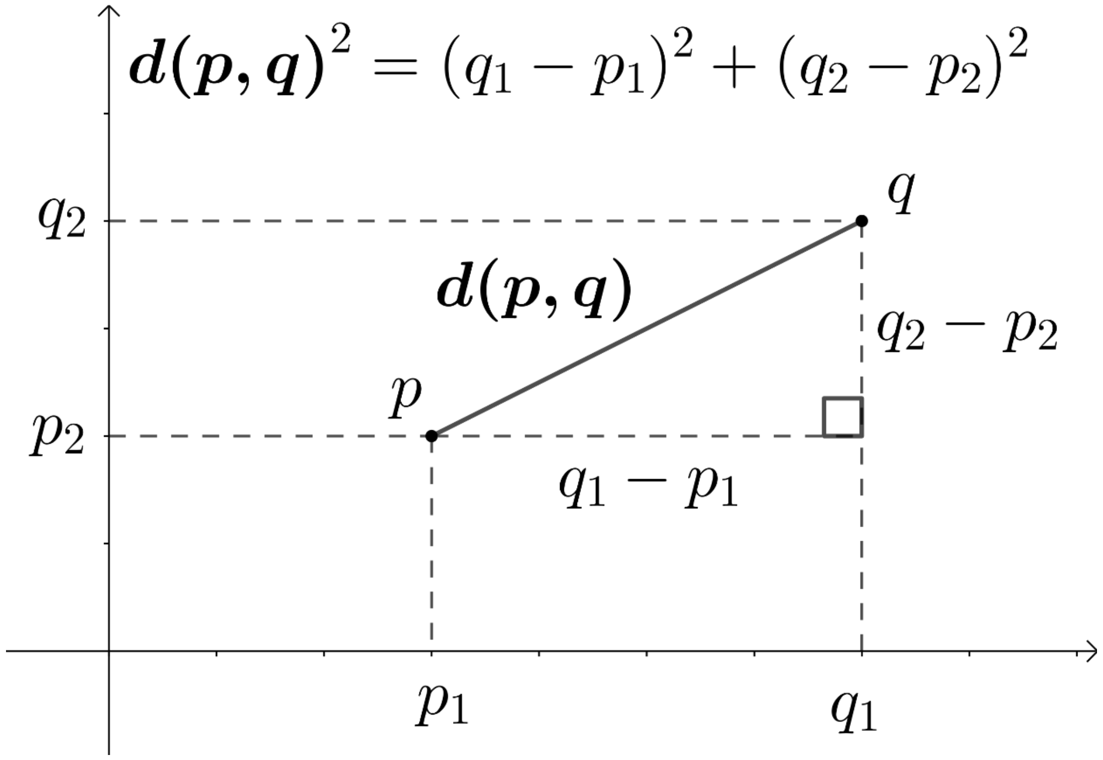
What could "similar" mean?

- One option: Euclidean distance (squared)

Clustering - similarity

What could "similar" mean?

- One option: Euclidean distance (squared)



Euclidean distance in $\mathcal{R}^2$

Two dimensions

if **p** = (p_1, p_2) and **q** = (q_1, q_2) then the distance is given by

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}$$

Clustering - similarity

What could "similar" mean?

- One option: Euclidean distance (squared)

Euclidean distance in $\mathcal{R}^n$

n dimensions

$$\boldsymbol{X} = (x_1, x_2, \dots, x_n) \quad \boldsymbol{Y} = (y_1, y_2, \dots, y_n)$$

Then the Euclidean distance is given by

$$d(p,q) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_n - x_n)^2}$$

$$d(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Clustering - similarity

What could "similar" mean?

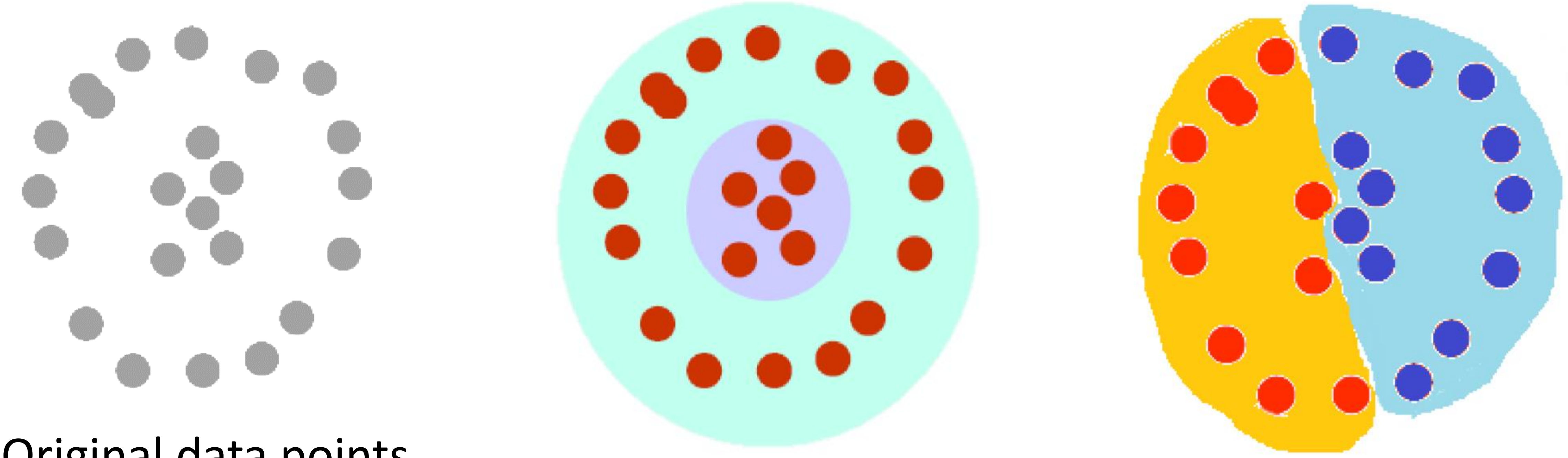
- One option: Euclidean distance (squared)
- Clustering results are remarkably dependent on the measure of similarity (or distance) between data points to be clustered

Chebyshev distance measures distance assuming only the most significant dimension is relevant.
Manhattan distance measures distance following only axis-aligned directions.
Minkowski distance is a generalization that unifies Euclidean distance, Manhattan distance, and Chebyshev distance

A cluster is a collection of data points which are "similar" between them, and "dissimilar" to data points in other clusters.

Clustering – cluster/group

Two different clustering results (i.e., clusters)



Original data points

Clustering - some applications

- Social network analysis
  - the discovery of clusters or communities, target marketing schemes, etc.
- Market segmentation
- Search result grouping
- Medical imaging
- Image segmentation and image concept extraction
- Anomaly detection
- …

QUEEN'S UNIVERSITY BELFAST

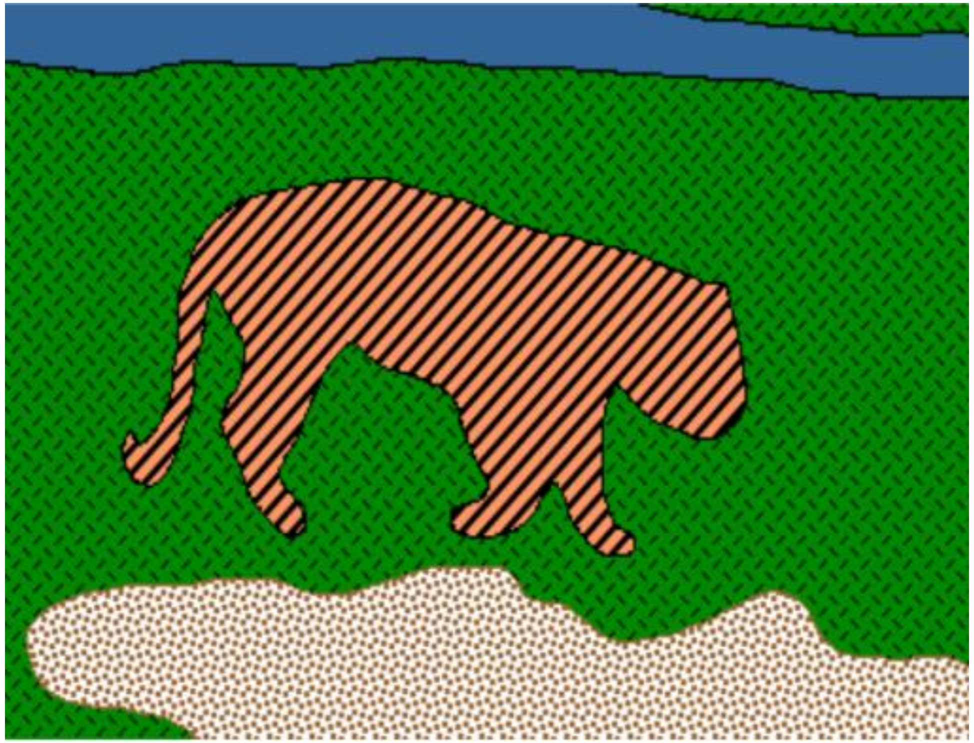
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Clustering image pixels

Image segmentation

Goal: identify groups of pixels that are similar and meaningfully connected

?



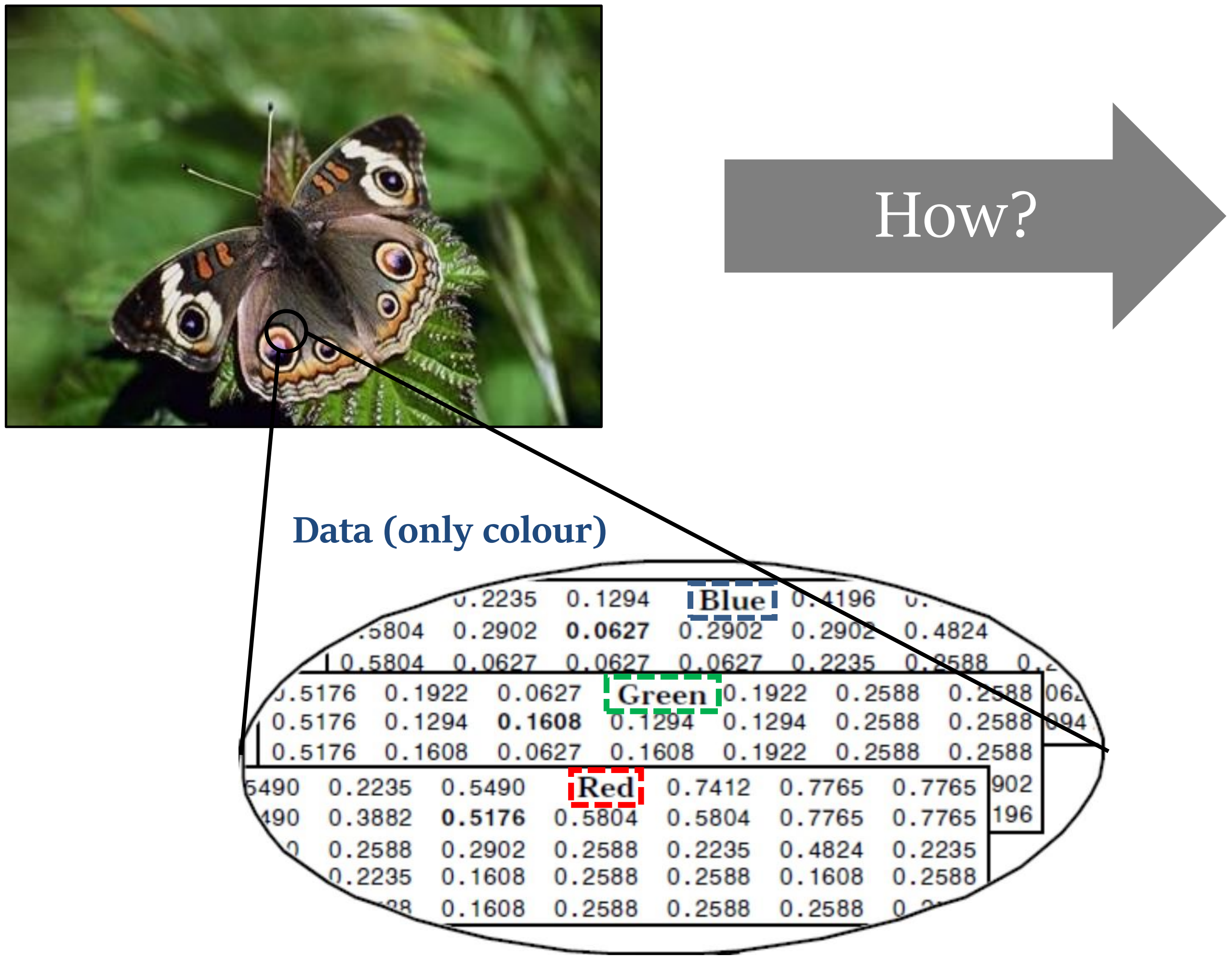
Discuss about data points, feature types for this clustering example

Clustering image pixels

Aim: detecting and extracting interest regions from an image

Identify groups of pixels that are similar and meaningfully connected

a



b

a) A colour image including an interest region (i.e., butterfly). **b)** Interest region extracted by a computer program

https://www.sciencedirect.com/science/article/pii/S0031320313001179

QUEEN'S UNIVERSITY BELFAST

SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

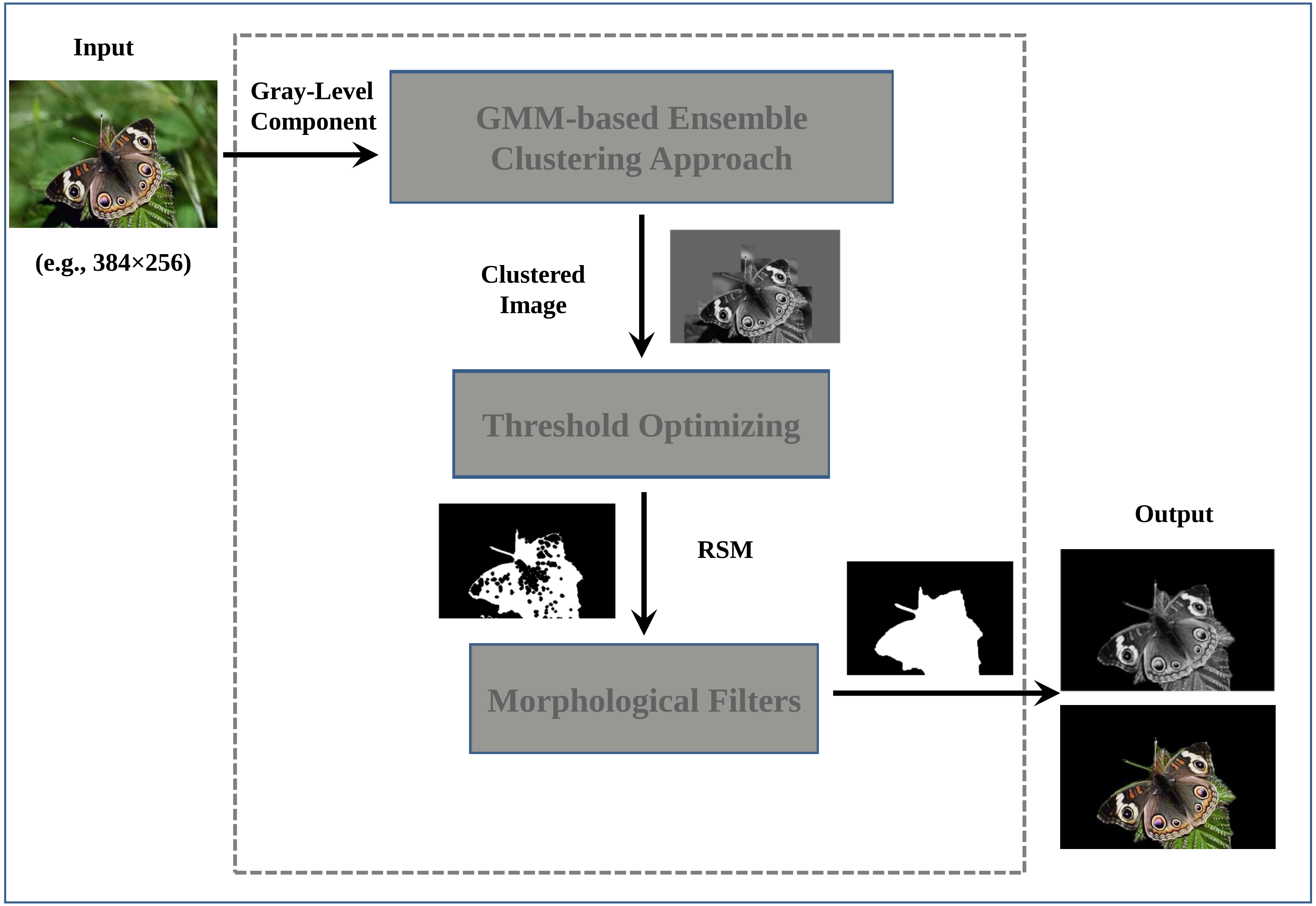
Clustering image pixels

Region-of-interest extraction from images

G. Rafiee, S.S. Dlay, W.L. Woo
Region-of-interest extraction in low depth of field images using ensemble clustering and difference of Gaussian approaches
Pattern Recognition., 46 (10) (2013), pp. 2685-2699

Clustering image pixels

Region-of-interest extraction from images



G. Rafiee, S.S. Dlay, W.L. Woo
Region-of-interest extraction in low depth of field images using ensemble clustering and difference of Gaussian approaches
Pattern Recognition., 46 (10) (2013), pp. 2685-2699

Clustering image pixels

Region-of-interest extraction from images

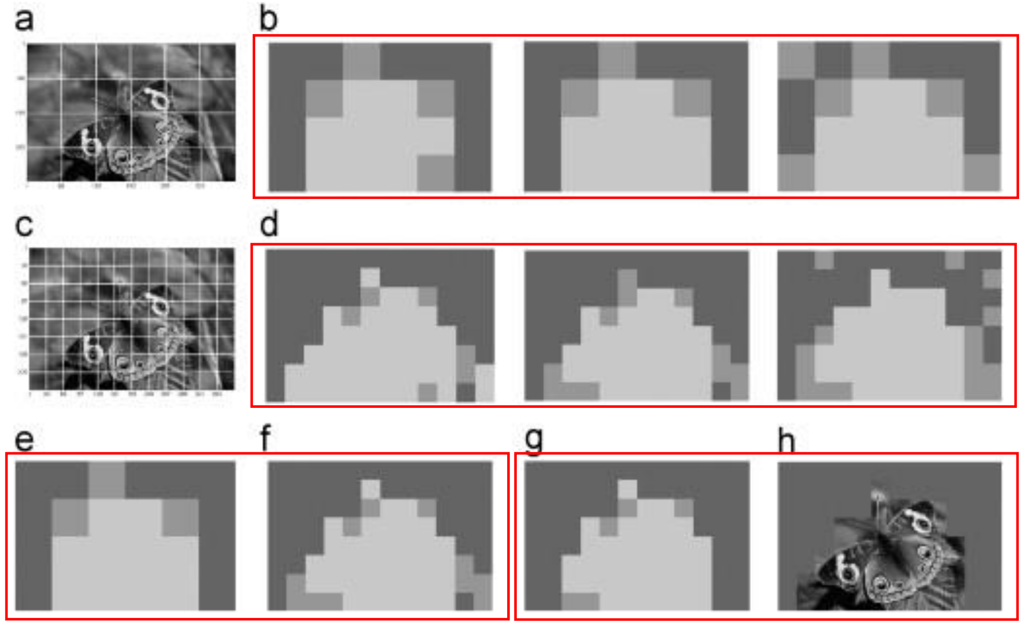


Illustration of different partitions and the fusion decision process. (a) and (c) Grayscale images with uniform partitioning at two consecutive levels, i.e., 64*64 and32*32. (b) and (d) Different partitions corresponding to different local optima at the first and second level, respectively. (e) and (f) Partitions after aggregating process in each level. (g) Final partition after combining (e) and (f). (h) Clustered image.

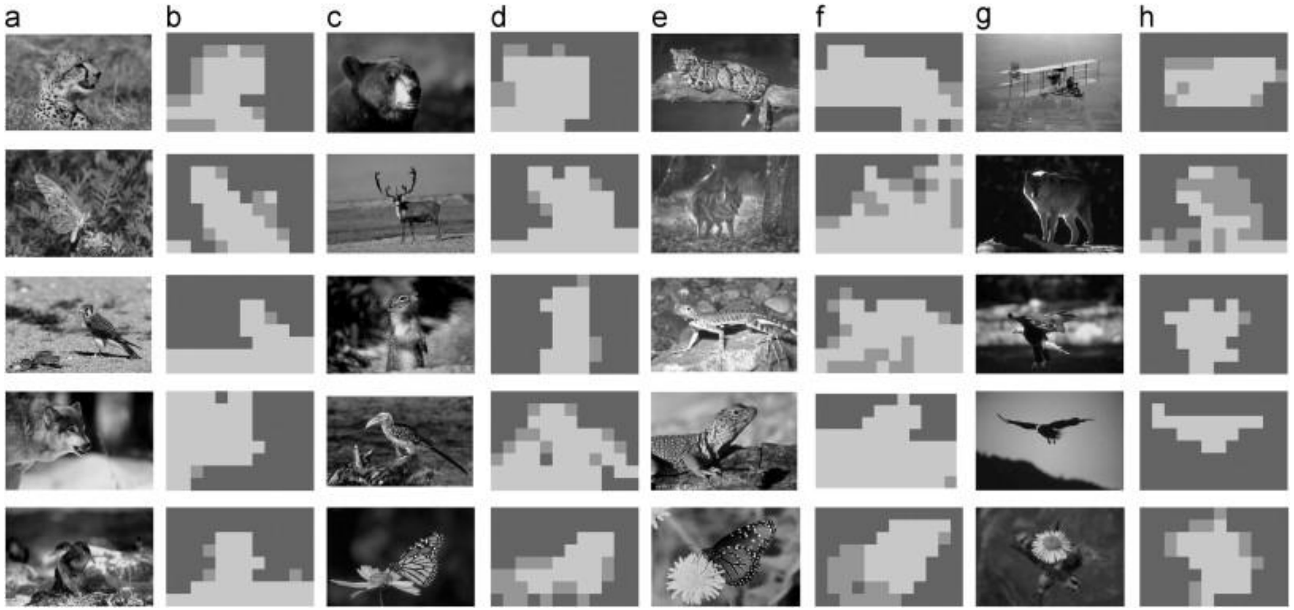
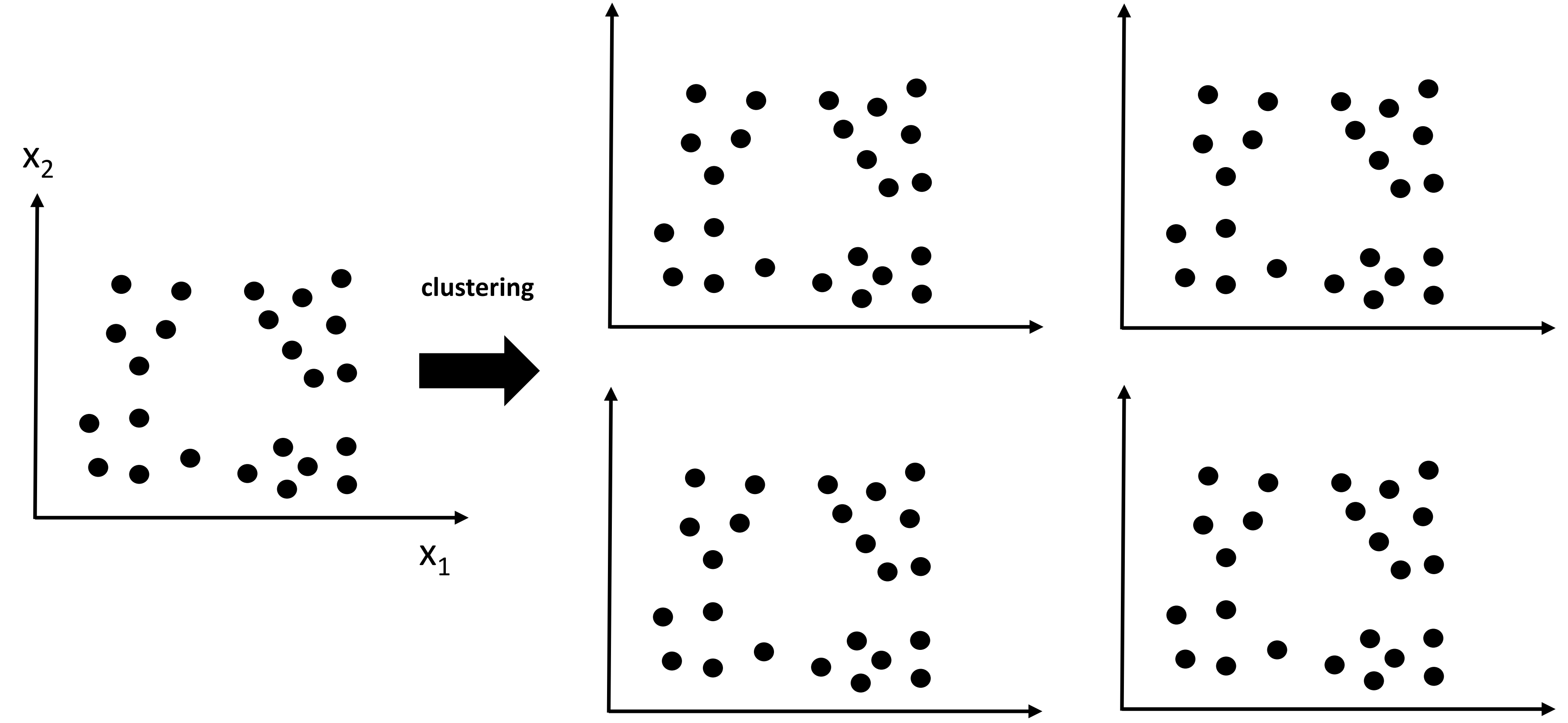


Illustration of final partitions for a number of images obtained from the algorithm withT_1=10. (a), (c), (e), and (g) Grayscale test images. (b), (d), (f), and (h) Final partitions after employing the combining process.

https://www.sciencedirect.com/science/article/pii/S0031320313001179#f0040

Number of clusters

Clustering concept



And more …

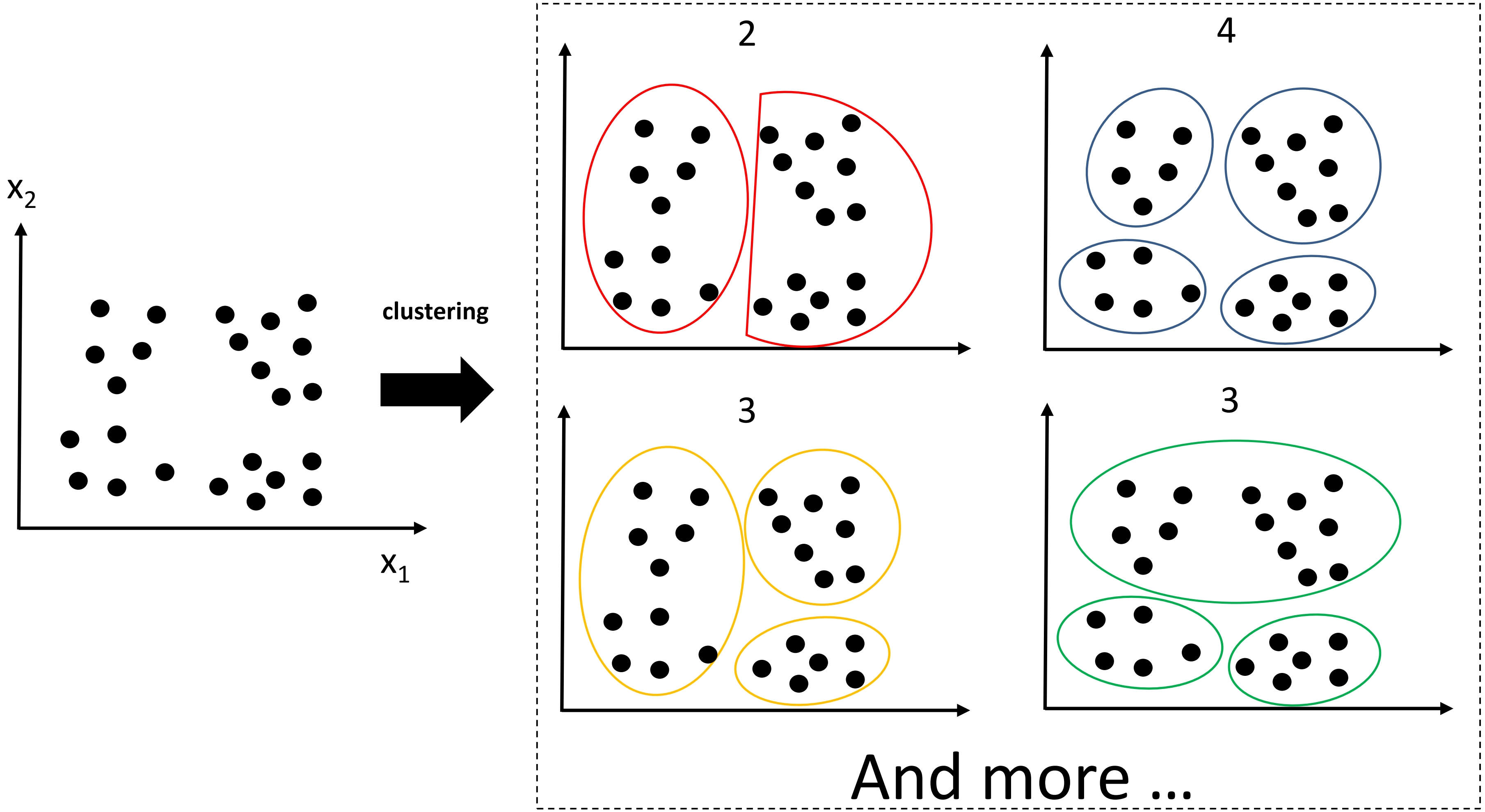
In case of applying an appropriate clustering method (with well-adjusted parameters/initialisation, bootstrapping and cross-validation techniques), we could have distinct groups (with possibly different number of clusters) but they might be meaningless!

QUEEN'S UNIVERSITY BELFAST

SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Number of clusters

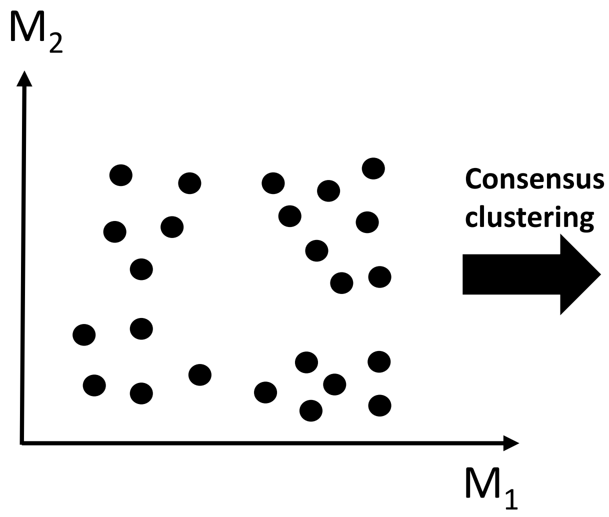
Identifying the number of clusters is a very challenging task

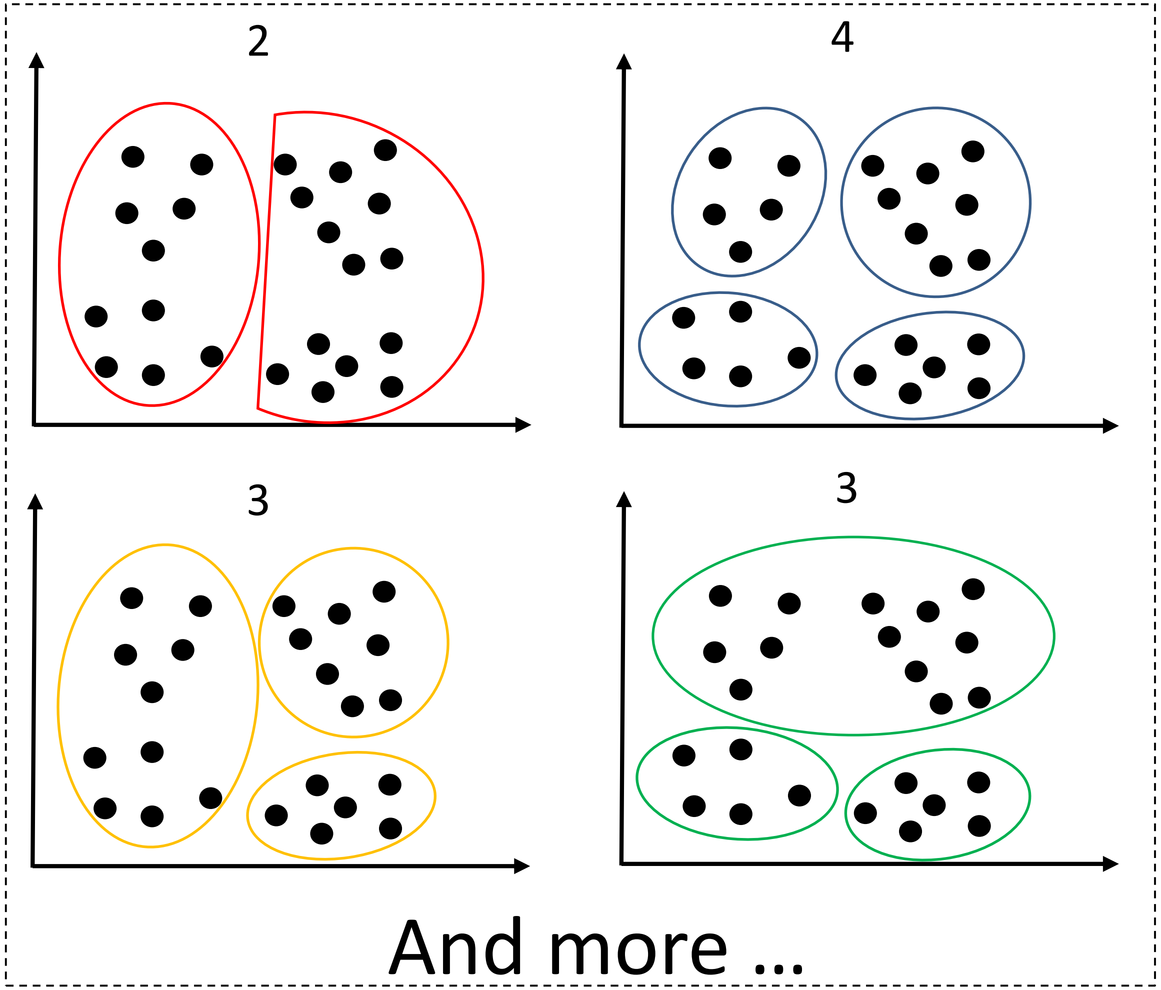


In case of applying an appropriate clustering method (with well-adjusted parameters/initialisation, bootstrapping and cross-validation techniques), we could have distinct groups (with possibly different number of clusters) but they might be meaningless!

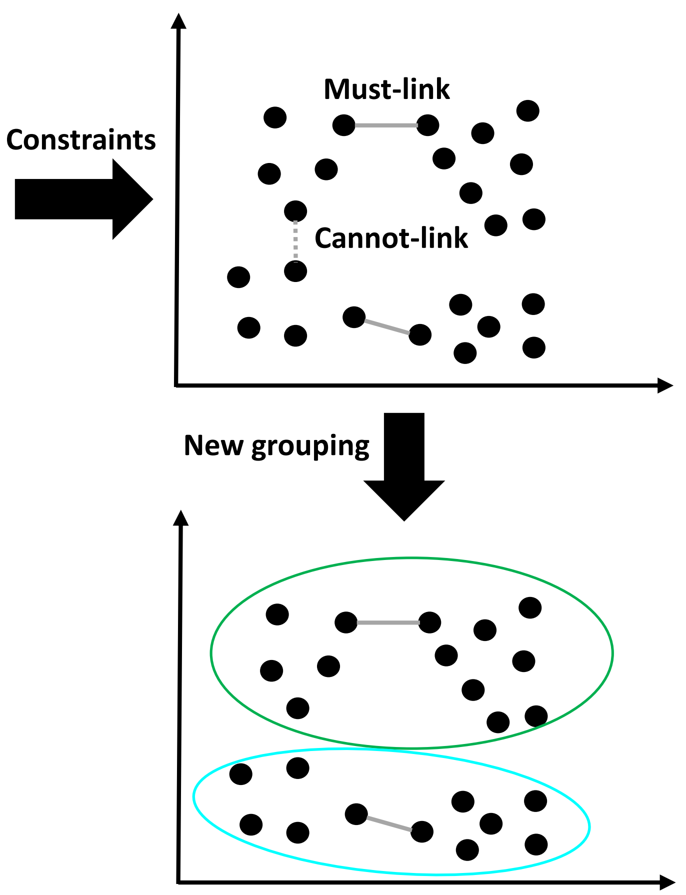
Prior knowledge – constrained clustering

Sometimes prior knowledge could help in finding the correct number of clusters





A must-link constraints is the knowledge/information which we are aware of some samples' connection



In case of applying an appropriate consensus clustering method (with well-adjusted parameters/initialisation, bootstrapping and cross-validation techniques), we could have distinct groups (with possibly different number of clusters) but they might be biologically meaningless!

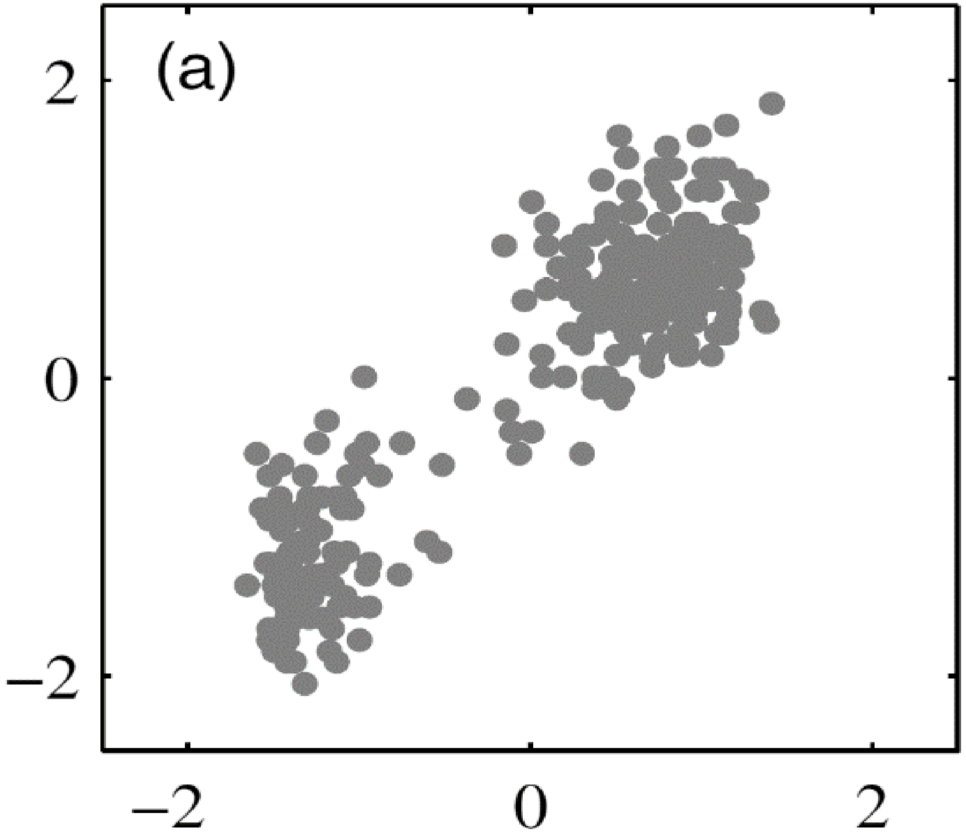
Must-link and cannot-link constraints are indicated by solid line and dashed line, respectively (e.g., 3 constraints).

K-means clustering

- k-means is one of the simplest unsupervised learning algorithms
- It classifies a given data set through a certain number of clusters (let's say k clusters)

K-means clustering

Let’s cluster the following data points using k-means algorithm

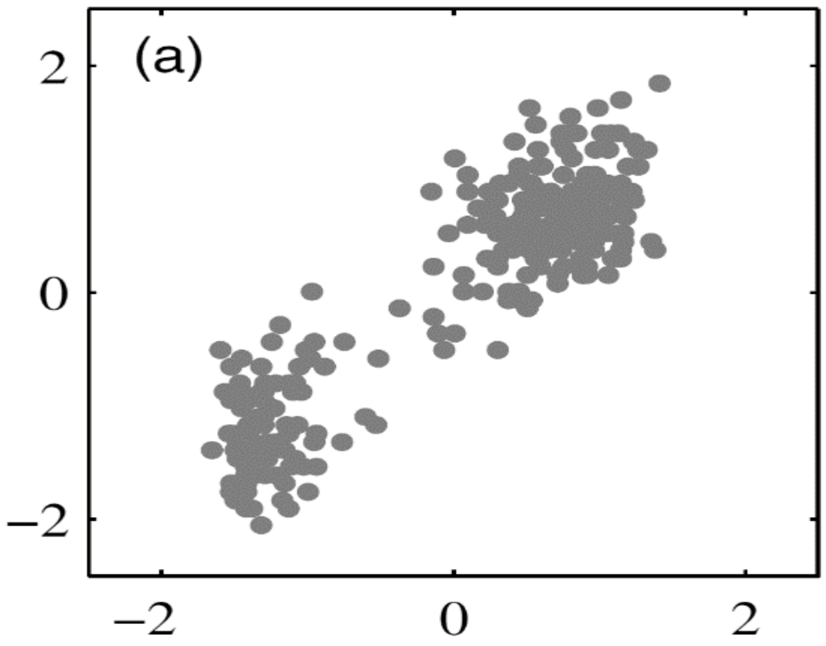


Images originated from Pattern Recognition and Machine Learning by Christopher M. Bishop

K-means clustering

Basic algorithm:

- Step 1: select k (number of clusters)
- Step 2: randomly select k initial cluster centers (or cluster centroids)



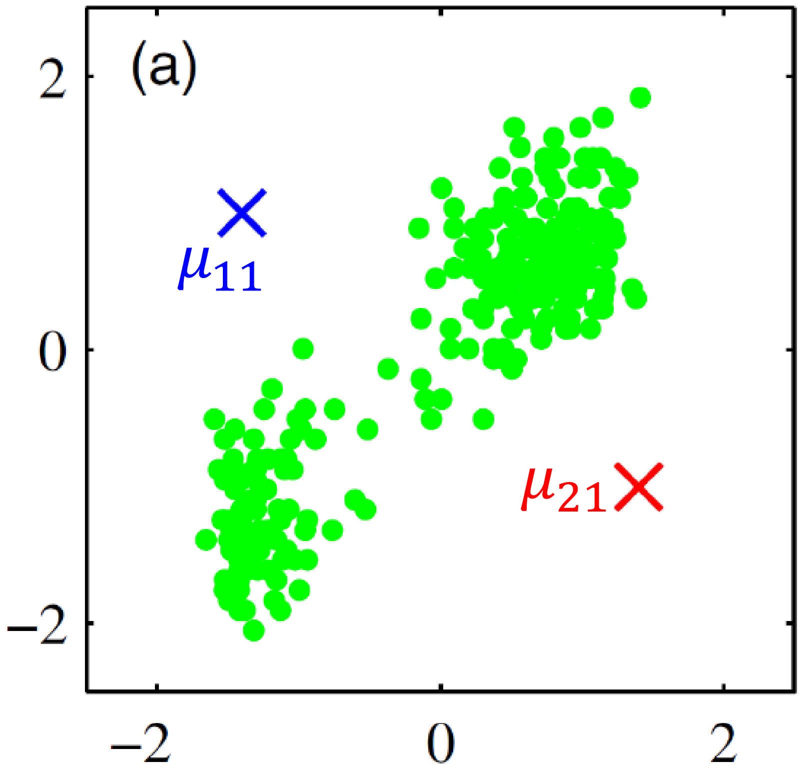
Images originated from Pattern Recognition and Machine Learning by Christopher M. Bishop

K-means clustering

Basic algorithm:

- Step 1: select k (number of clusters)
- Step 2: randomly select k initial cluster centers (or cluster centroids)

k=2

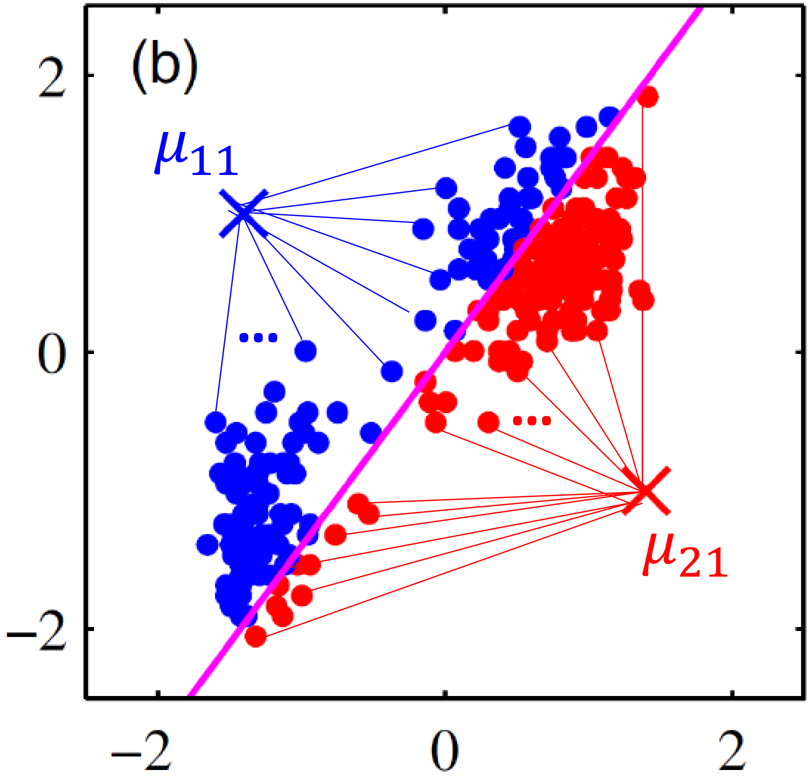


Images originated from Pattern Recognition and Machine Learning by Christopher M. Bishop

K-means clustering

Basic algorithm:

- Step 1: select k (number of clusters)
- Step 2: randomly select k initial cluster centers
- Step 3: calculate distance from each data point to each cluster center
  - What type of distance should we use? E.g., Euclidean distance
- Step 4: assign each data point to the closest cluster center (centroid)

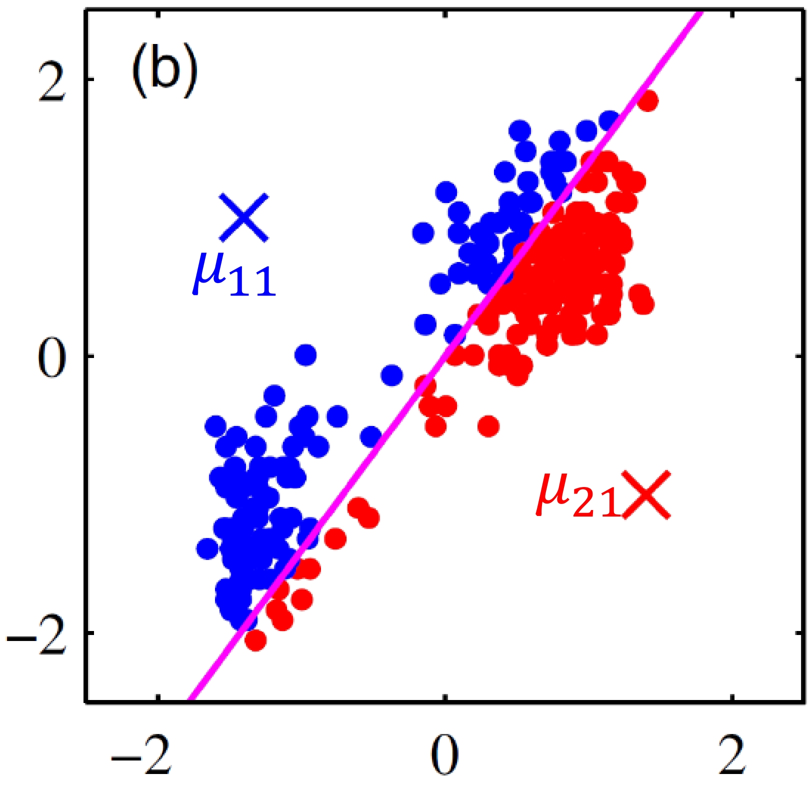


Distances partially illustrated

K-means clustering

Basic algorithm:

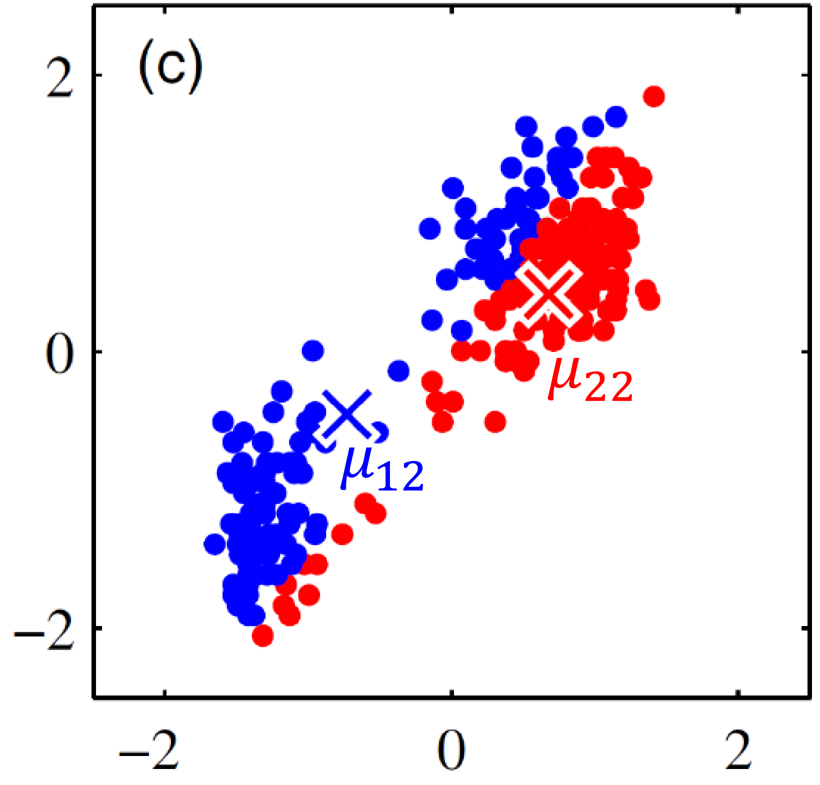
- Step 1: select k (number of clusters)
- Step 2: randomly select k initial cluster centers
- Step 3: calculate distance from each data point to each cluster center
  - What type of distance should we use? E.g., Euclidean distance
- Step 4: assign each data point to the closest cluster center (centroid)



K-means clustering

Basic algorithm:

- Step 1: select k (number of clusters)
- Step 2: randomly select k initial cluster centers
- Step 3: calculate distance from each data point to each cluster center
  - What type of distance should we use? E.g., Euclidean distance
- Step 4: assign each data point to the closest cluster center (centroid)
- Step 5: calculate new centroids as the mean of the data points that belong to the centroid of the previous step

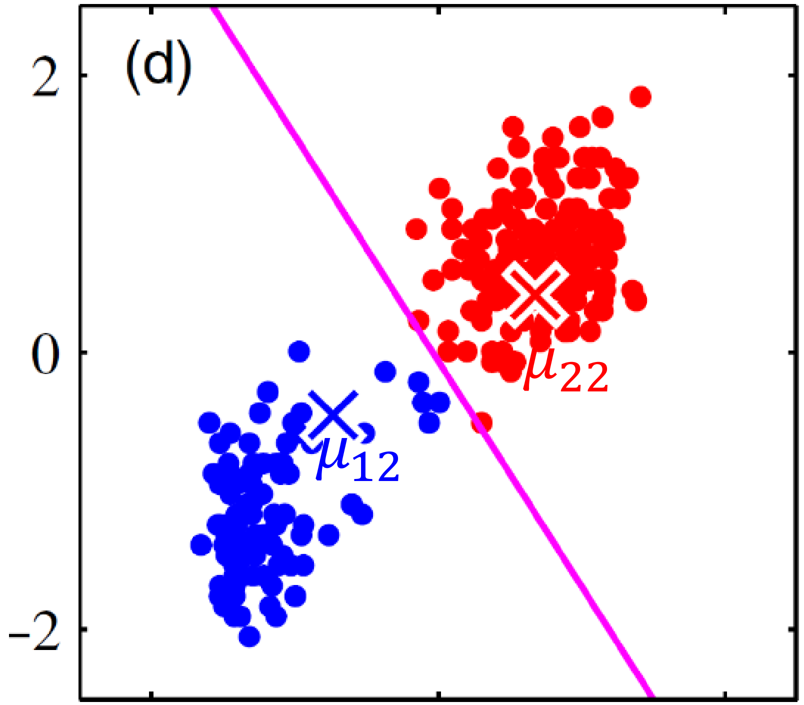


Distances partially illustrated

K-means clustering

Basic algorithm:

- Step 1: select k (number of clusters)
- Step 2: randomly select k initial cluster centers
- Step 3: calculate distance from each data point to each cluster center
    - What type of distance should we use? E.g., Euclidean distance
- Step 4: assign each data point to the closest cluster center (centroid)
- Step 5: calculate new centroids as the mean of the data points that belong to the centroid of the previous step

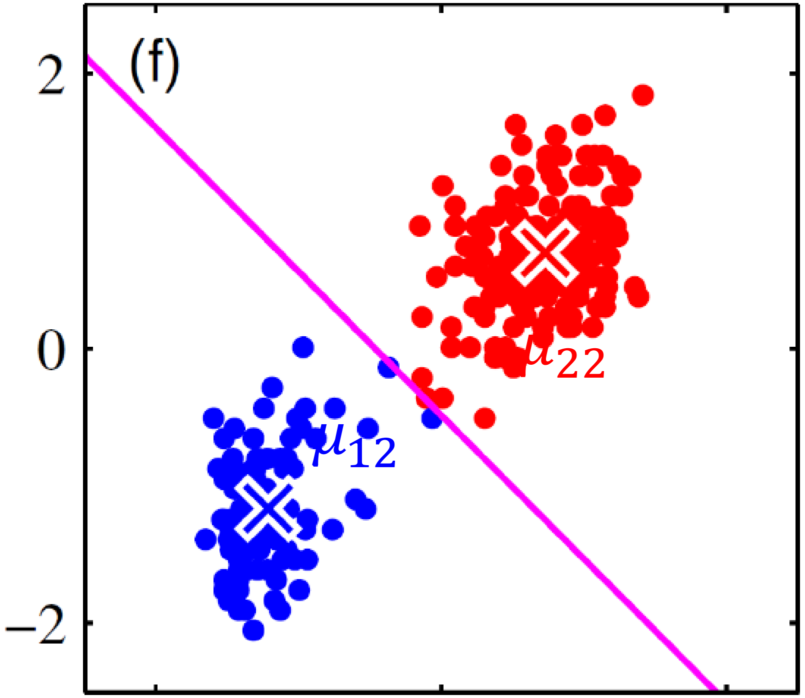


Distances partially illustrated

K-means clustering

Basic algorithm:

- Step 1: select k (number of clusters)
- Step 2: randomly select k initial cluster centers
- Step 3: calculate distance from each data point to each cluster center
  - What type of distance should we use? E.g., Euclidean distance
- Step 4: assign each data point to the closest cluster center (centroid)
- Step 5: calculate new centroids as the mean of the data points that belong to the centroid of the previous step
- Repeat Step 3-5 until a final stop condition

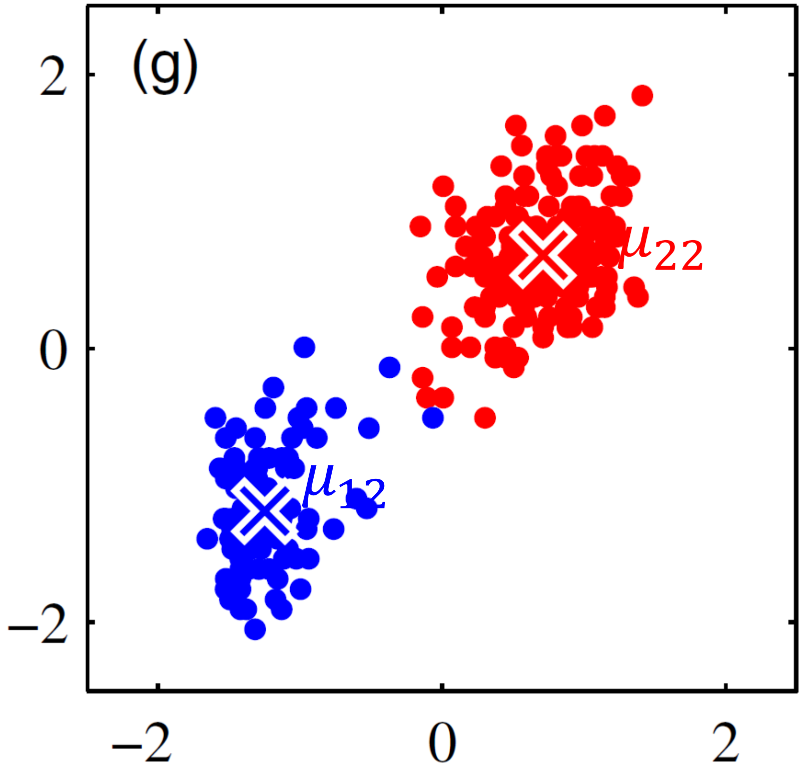


Distances partially illustrated

K-means clustering

Basic algorithm:

- Step 1: select k (number of clusters)
- Step 2: randomly select k initial cluster centers
- Step 3: calculate distance from each data point to each cluster center
  - What type of distance should we use? E.g., Euclidean distance
- Step 4: assign each data point to the closest cluster center (centroid)
- Step 5: calculate new centroids as the mean of the data points that belong to the centroid of the previous step
- Repeat Step 3-5 until a final stop condition

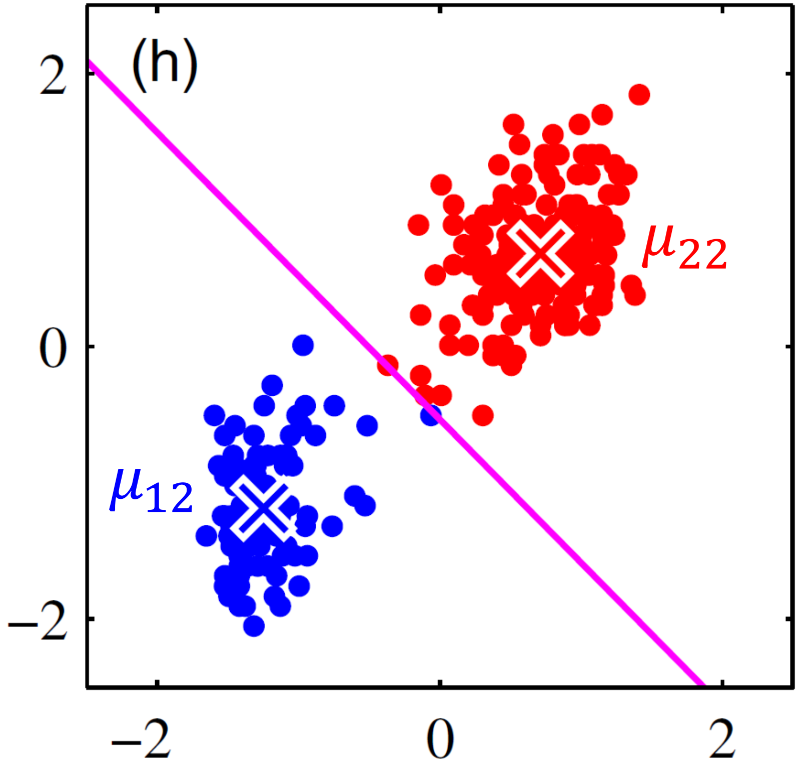


Distances partially illustrated

K-means clustering

Basic algorithm:

- Step 1: select k (number of clusters)
- Step 2: randomly select k initial cluster centers
- Step 3: calculate distance from each data point to each cluster center
  - What type of distance should we use? E.g., Euclidean distance
- Step 4: assign each data point to the closest cluster center (centroid)
- Step 5: calculate new centroids as the mean of the data points that belong to the centroid of the previous step
- Repeat Step 3-5 until a final stop condition

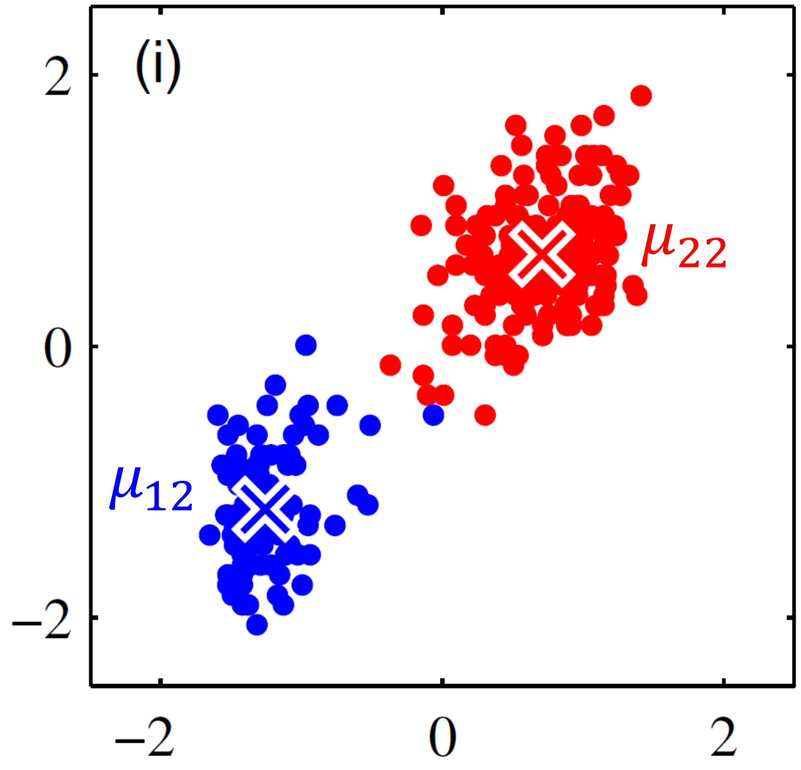


Distances partially illustrated

K-means clustering

Basic algorithm:

- Step 1: select k (number of clusters)
- Step 2: randomly select k initial cluster centers
- Step 3: calculate distance from each data point to each cluster center
  - What type of distance should we use? E.g., Euclidean distance
- Step 4: assign each data point to the closest cluster center (centroid)
- Step 5: calculate new centroids as the mean of the data points that belong to the centroid of the previous step
- Repeat Step 3-5 until a final stop condition

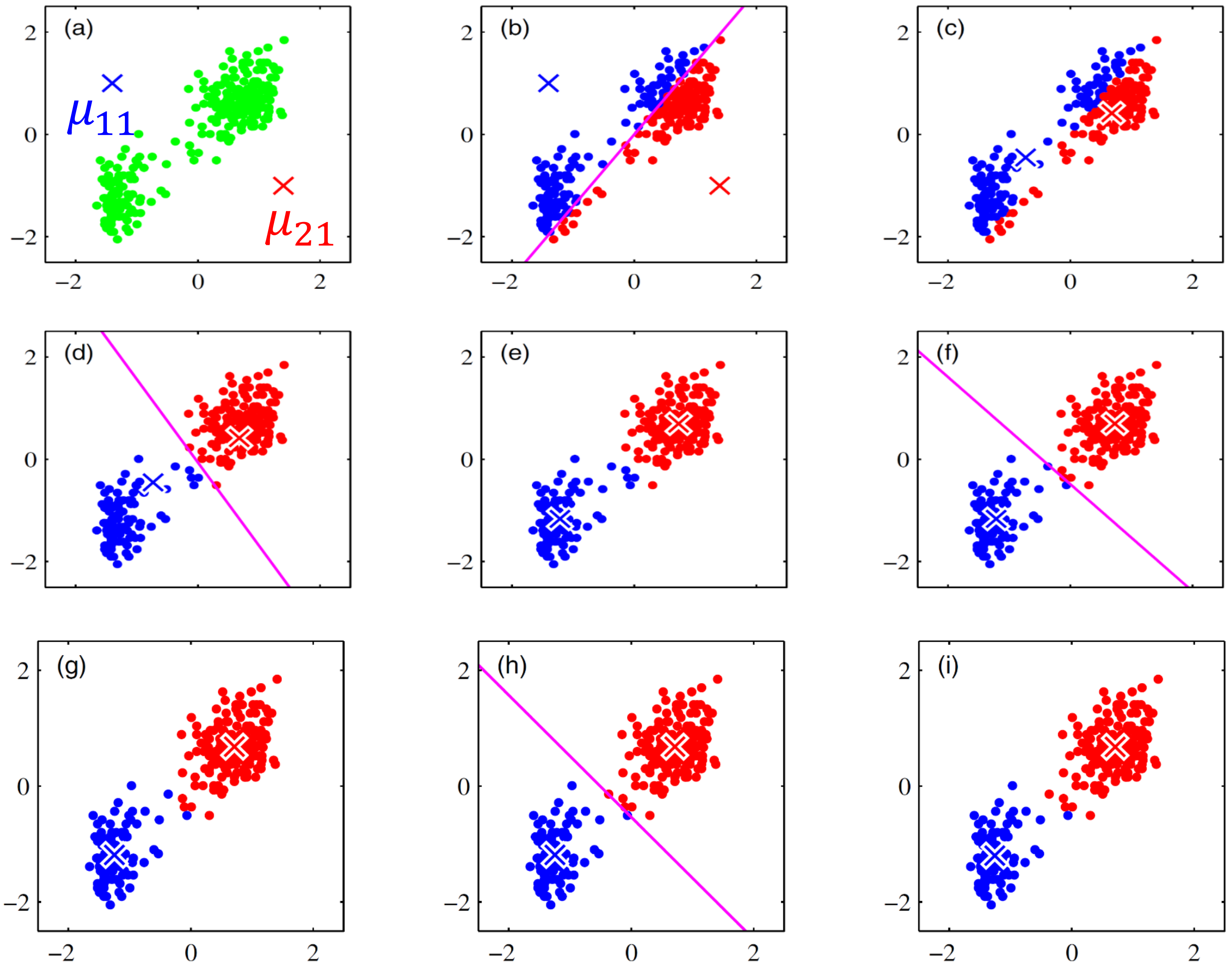


Distances partially illustrated

K-means clustering

Initial cluster centres: μ_1 μ_2

Illustration of *k*-means algorithm
(a) Green points denote the data set in a two-dimensional Euclidean space



Images originated from Pattern Recognition and Machine Learning by Christopher M. Bishop

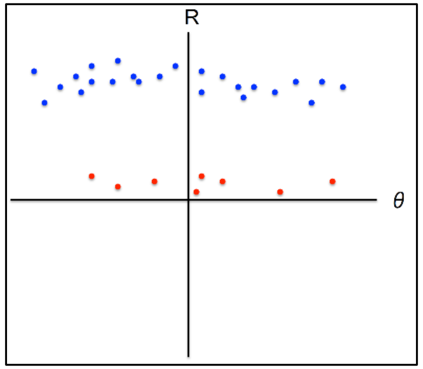
K-means clustering

Basic algorithm:

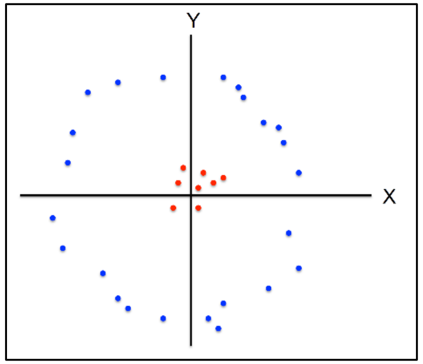
- Step 1: select k (number of clusters)
- Step 2: randomly select k initial cluster centers
- Step 3: calculate distance from each data point to each cluster center
  - What type of distance should we use? E.g., Euclidean distance
- Step 4: assign each data point to the closest cluster center (centroid)
- Step 5: calculate new centroids as the mean of the data points that belong to the centroid of the previous step
- Repeat Step 3-5 until a final stop condition (if no data point was reassigned then stop).

K-means clustering

- Strengths
  - Simple & fast and can be applied to high-dimensional large data
  - Finds cluster centres that minimize conditional variance (good representation of data)
  - Easy to implement
- Weaknesses
  - Need to choose k
  - Sensitive to outliers
  - Prone to local minima and no guarantee of optimal solution (local optima)
    - Repeat with different starting values
  - Difficult to guess the correct “k”



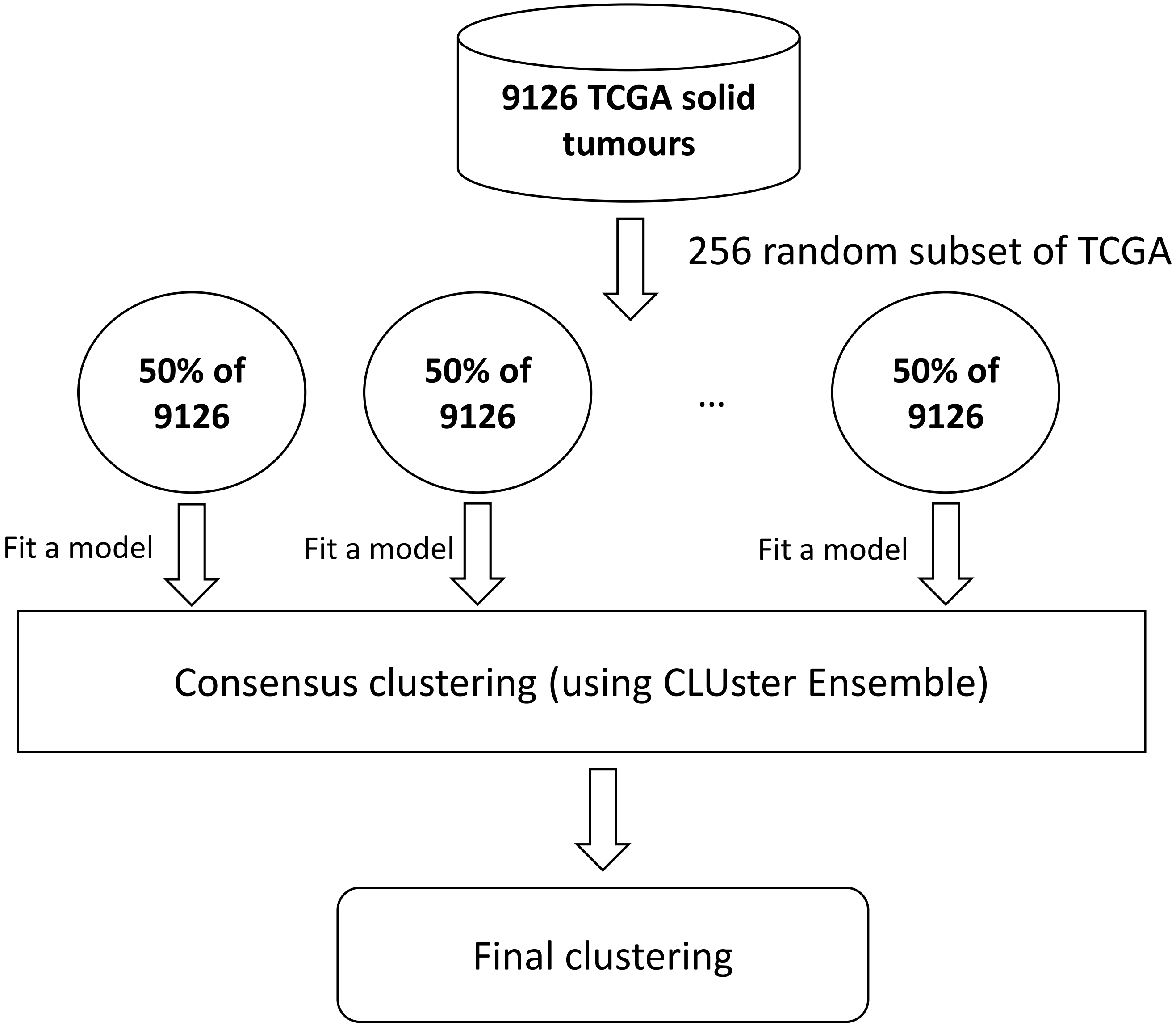
Changing features & distance function



K-means algorithm is not able to properly cluster this data points

Original approach (the immunity paper)

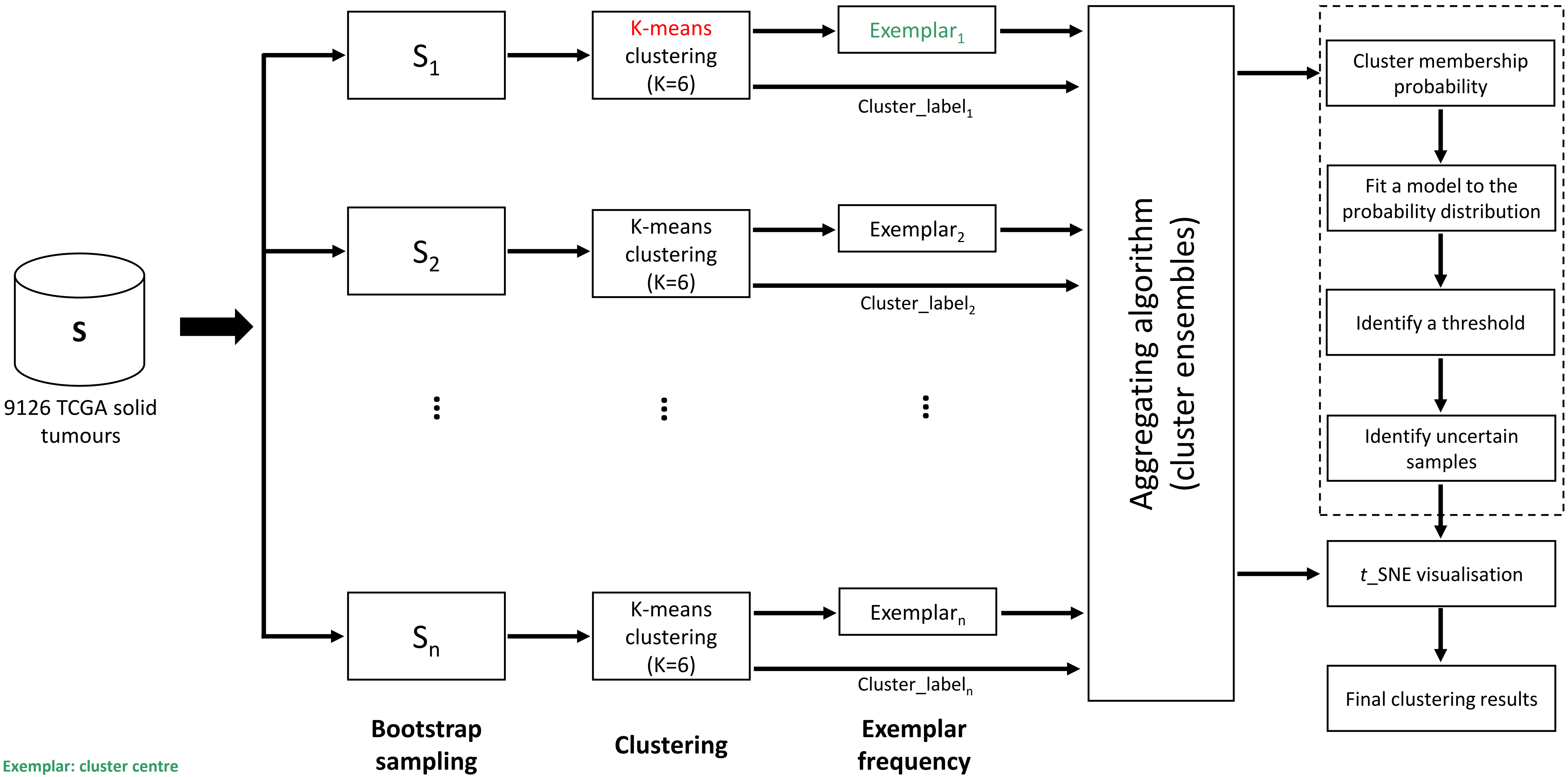
A comprehensive approach for unsupervised clustering of 9126 solid tumours **(440 genes)**



QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Consensus approach

A comprehensive Ensemble approach for unsupervised clustering of 9126 solid tumours (440 genes) with **the objective of identifying uncertain samples in clustering**



Exemplar: cluster centre

Dr Reza Rafiee: Bootstrap aggregating (Bagging) approach

Any Questions?