Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences
Western Switzerland

MSE | MASTER OF SCIENCE IN ENGINEERING

Machine Learning

T-MachLe

8. Unsupervised systems - Clustering

Jean Hennebert
Andres Perez Uribe

Plan

1. Recaps from class 1 about unsupervised learning and clustering
2. Clustering algorithms
3. k-Means algorithm

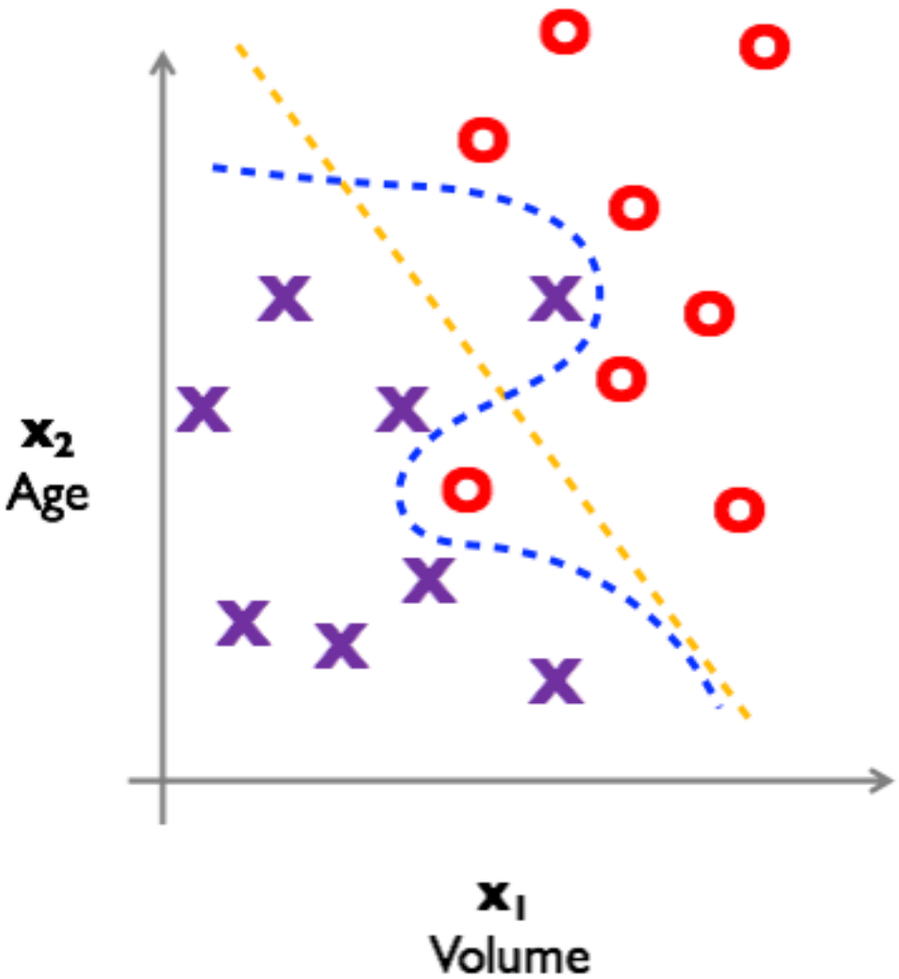
Practical Work 8

Unsupervised learning

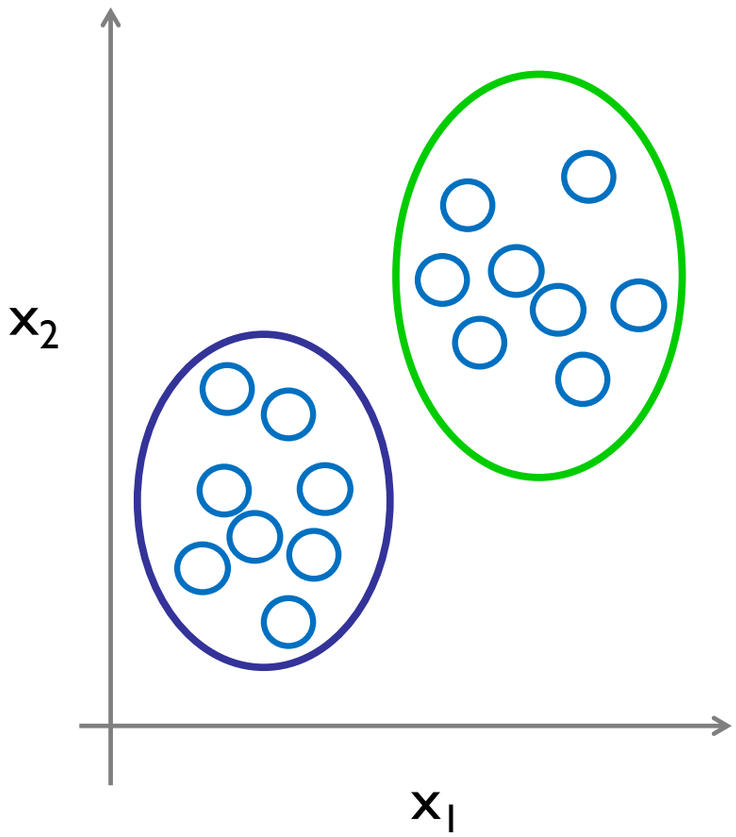
Clustering

RECAPS week 1

Supervised learning



Unsupervised learning

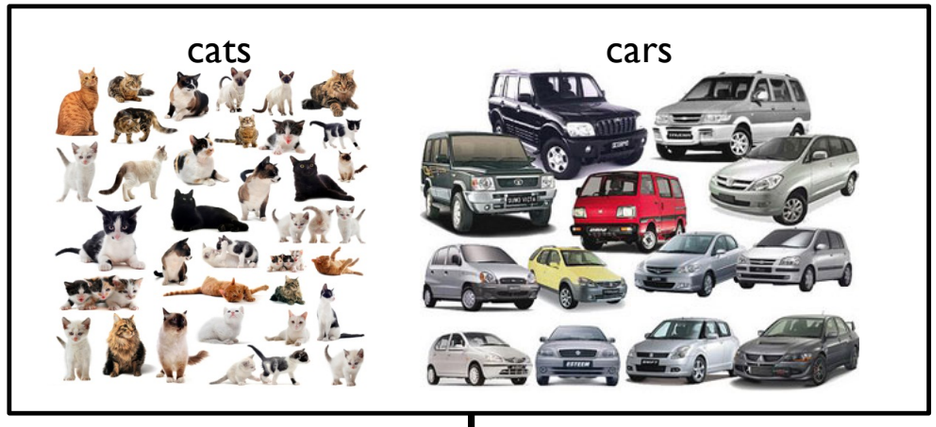


Unsupervised learning

RECAPS week 1

With **unsupervised learning**, the goal is to discover interesting structures from inputs **x** given a set of data called the training set.

Supervised



Unsupervised

Unsupervised learning - why?

RECAPS week 1

- Useful when we do not have the class labels for the training samples
  - Too much data to label
  - Too lazy to label
  - Or simply we don’t know the possible labels

Unsupervised learning - when?

RECAPS week 1

1. **Clustering**
   - Find and group data points with natural similarities
   - Often used to explore the potential grouping in a dataset
2. **Association**
   - Identify common co-occurrences among a list of possible events
   - Matrix completion: frequently used for market basket analysis, image in-painting, collaborative filtering

Unsupervised learning - when? RECAPS week 1

1. **Feature extraction**
   - Create new features from association or clustering of the attributes
   - Sometimes used to reduce the number of attributes to improve supervised techniques = dimensionality reduction
2. **Compression/Quantization**
   - Compress data into representative form
   - Commonly used to reduce bandwidth, e.g. lossy image / voice compression

Unsupervised learning - algos

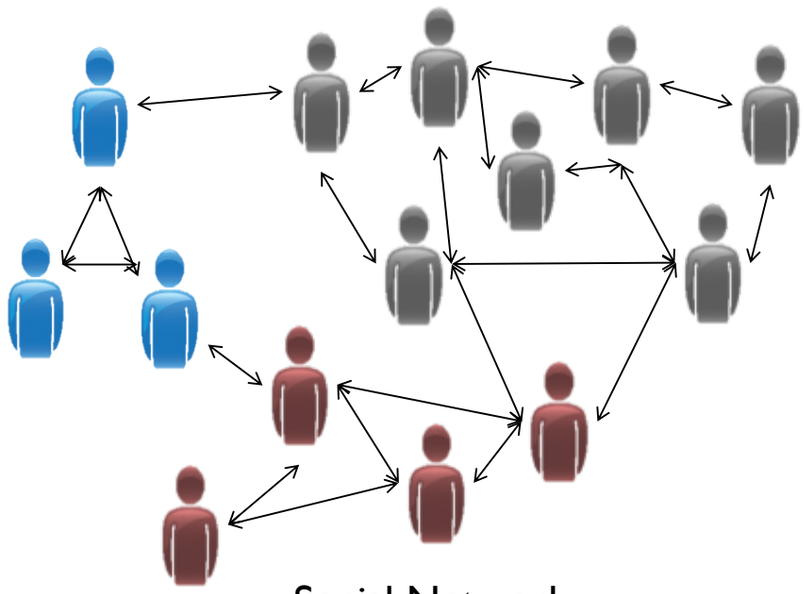
RECAPS week 1

- Clustering
  - k-Means clustering (partition-based)
  - Hierarchical clustering (hierarchical based)
  - Expectation-maximization algorithm (probabilistic model-based)
  - Kohonen’s Neural Network (Self Organizing Maps)
- Principal Component Analysis (PCA)
- Competitive Learning
- Auto-encoders
- Etc.

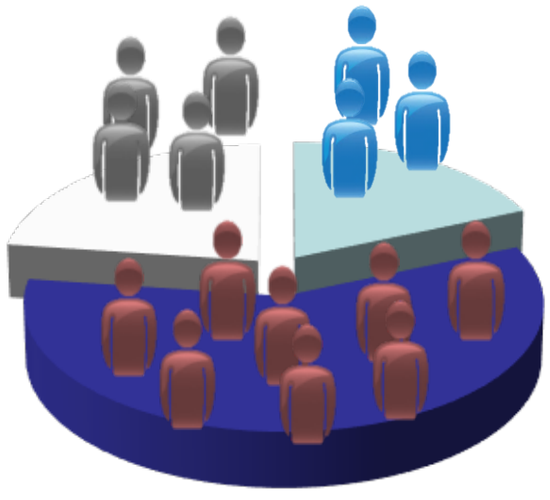
Unsupervised learning - apps

RECAPS week 1

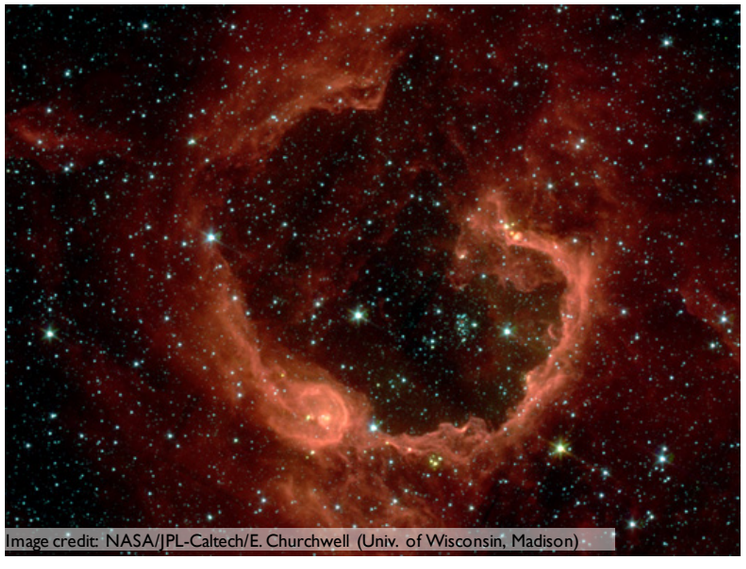
Organize computing clusters



Social Network



Market segmentation

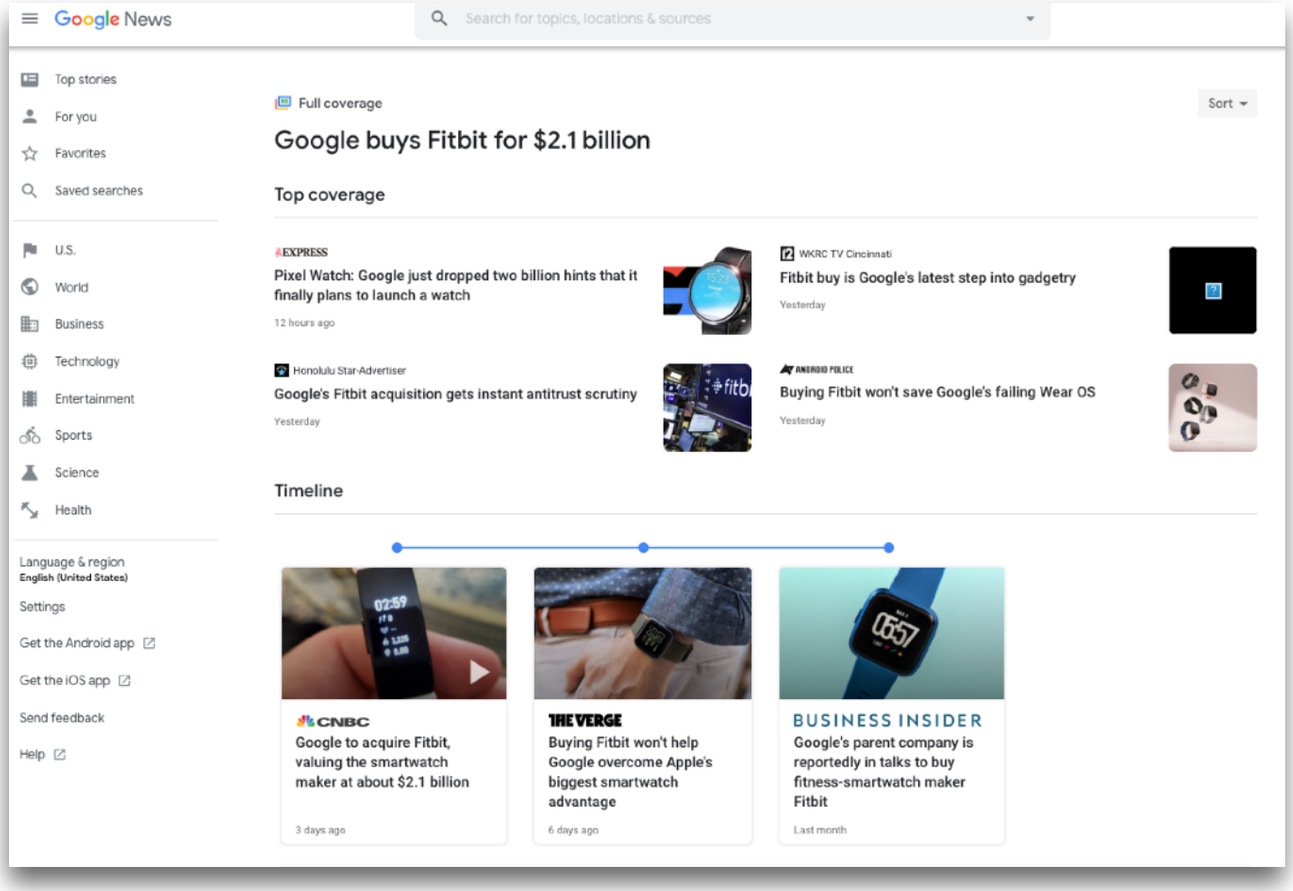


Astronomical data analysis

Example of text clustering

Activity

- Analyse Google news. How can they “group” similar news articles?
- Is it unsupervised learning? Under which conditions?
- How would you build such a system?
- How can we evaluate such a system?



Clustering algorithms

Definition

Inputs - outputs

Evaluation criteria

Overview of 5 algorithms

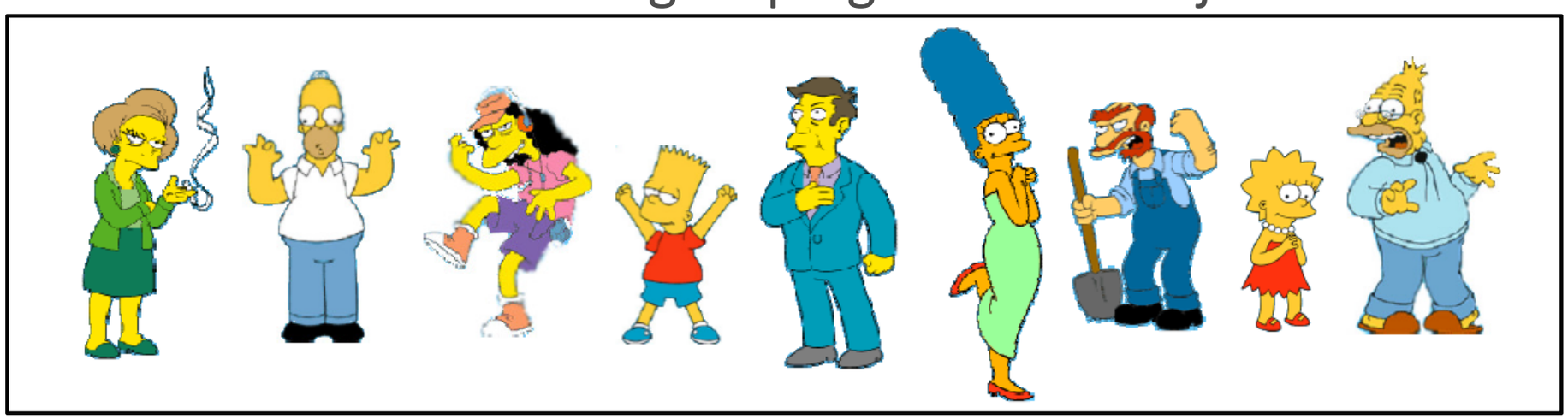
Definition

Clustering is finding "**cohesive**" clusters forming a partitioning of the input data. By cohesive, we mean to group the input samples into classes of similar objects with a high intra-class similarity and low inter-class similarity.

- We only have input variables in the training set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
  - We don't have "supervision" with targets $\{y_1, \ldots, y_N\}$
- In other words, we want to find "***natural***" groupings among objects
  - The groupings are the classes we want to "discover"

What do we mean by cohesive and natural?

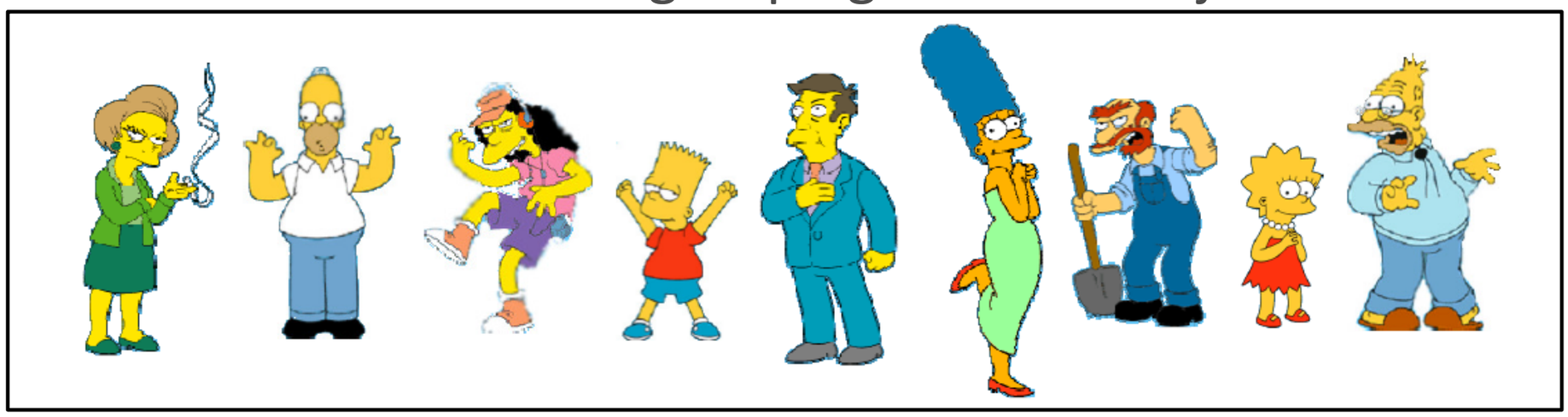
What is a natural grouping for these objects ?



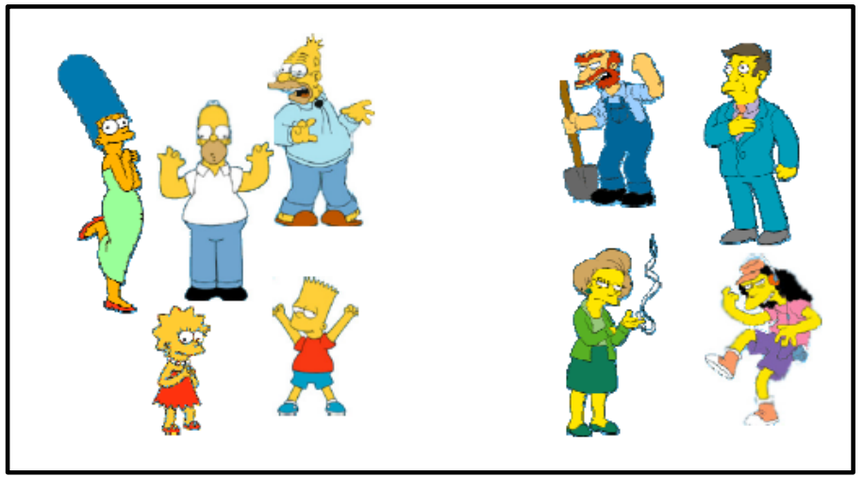
Source: http://www.cs.us.es/~fran/curso_unia/clustering.html

What do we mean by cohesive and natural?

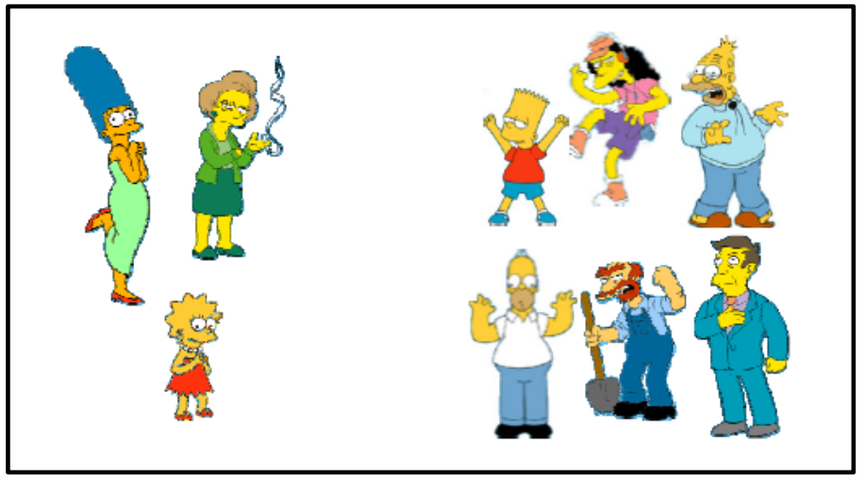
What is a natural grouping for these objects ?



Simpson family VS School employees ?



Females VS Males ?

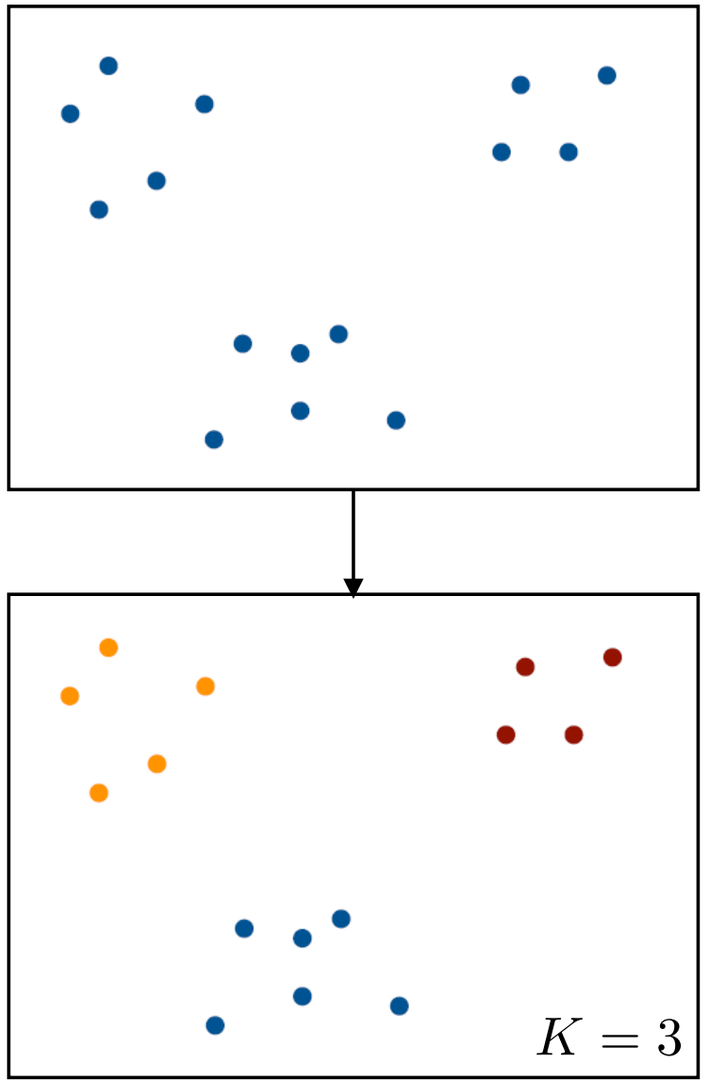


Source: http://www.cs.us.es/~fran/curso_unia/clustering.html

The user has to chose the relevant features that will actually define what is meant by “natural”.

Clustering input and output

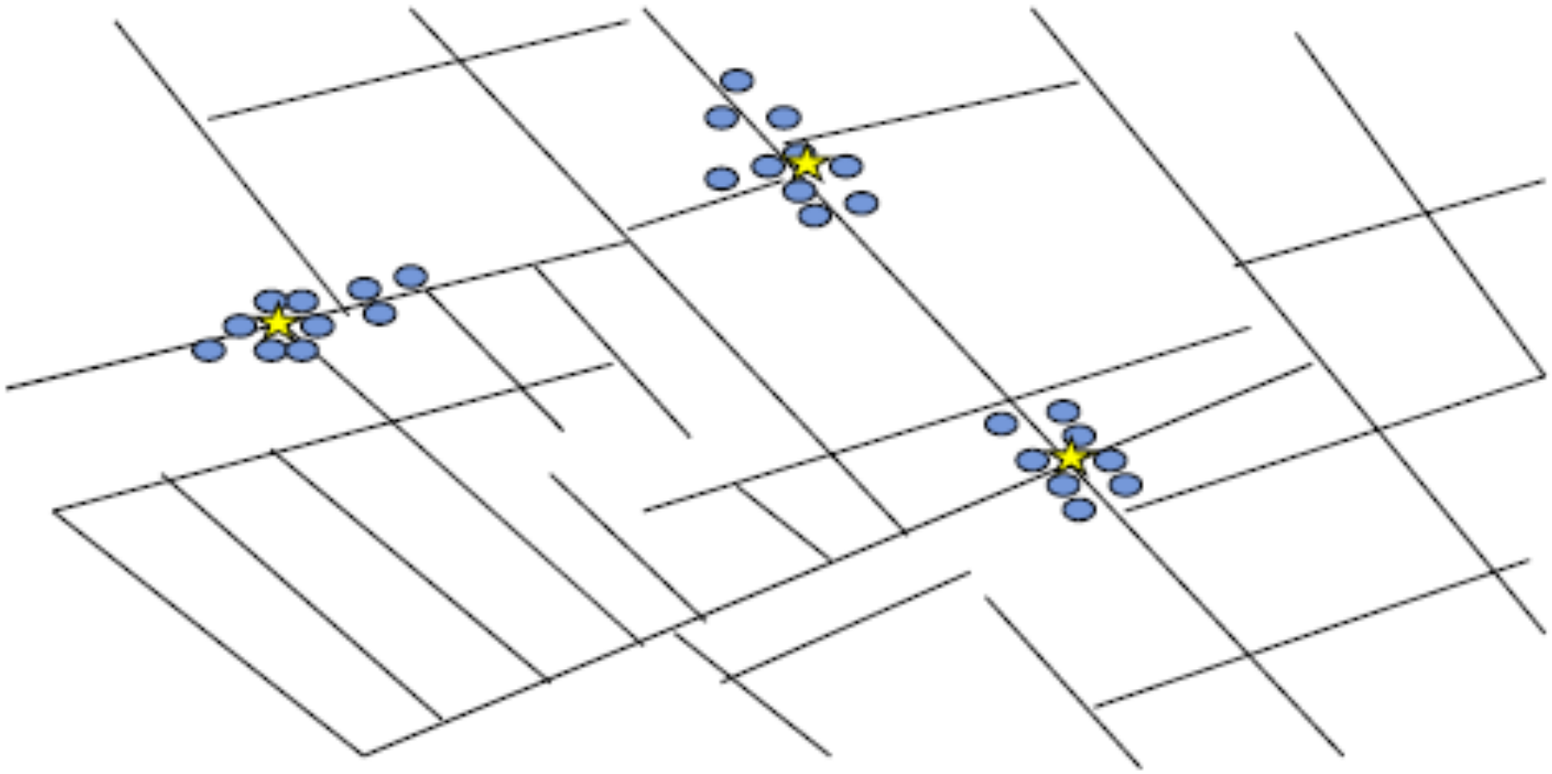
- Input = training samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
- Output = partitioning of the data into K regions of the input space



- Sometimes K is pre-defined as constraint of the problem
- In other cases, we will have to discover the best value for K

Example of clustering

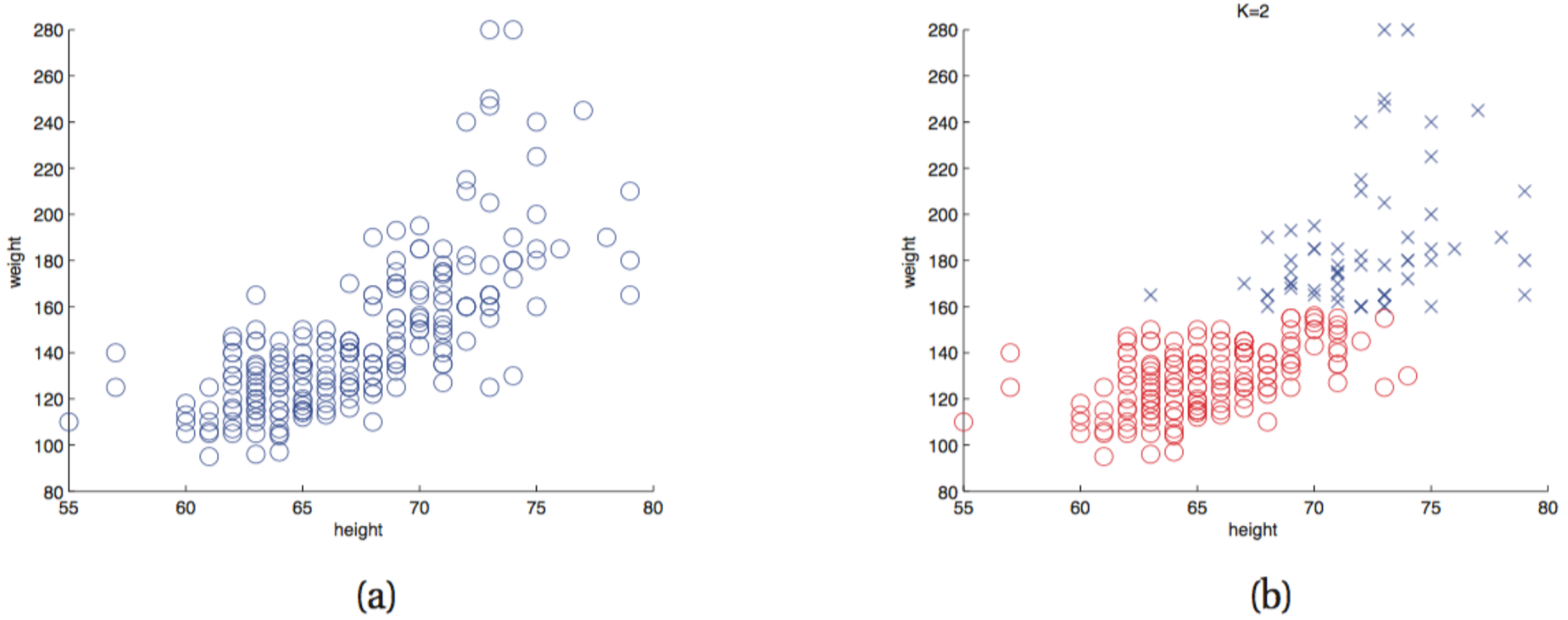
- The first known application of clustering is the cholera map of John Snow, a London physician who plotted the location of the deaths from cholera on a map (1850s)
- He found that deaths were clustered around specific wells



“ Snow was a sceptic of the then-dominant miasma theory that stated that diseases such as cholera and bubonic plague were caused by pollution or a noxious form of "bad air".

Snow later used a dot map to illustrate the cluster of cholera cases around the pump. He also used statistics to illustrate the connection between the quality of the water source and cholera cases. He showed that the Southwark and Vauxhall Waterworks Company was taking water from sewage-polluted sections of the Thames and delivering the water to homes, leading to an increased incidence of cholera. Snow's study was a major event in the history of public health and geography. It is regarded as the founding event of the science of epidemiology”

Example of clustering



(a) (b)

Source: K. Murphy, "Machine Learning – A Probabilistic Perspective", 2012

Figure 1.8 (a) The height and weight of some people. (b) A possible clustering using $K = 2$ clusters.

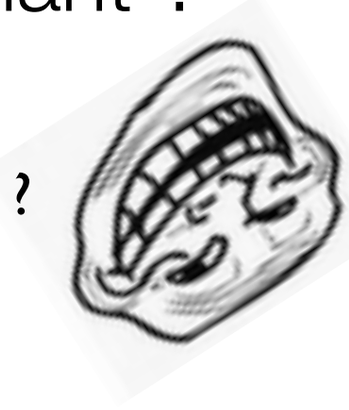
- A typical problem with unsupervised learning is to discover structure in the data, such as the presence of clusters. Once the clusters discovered, we may attempt to given them a signification.

What is needed for clustering?

1. **A distance measure** $d(\mathbf{x}_i, \mathbf{x}_j)$ defining a dissimilarity between two samples
   - Are two samples distant?
2. **A criterion to evaluate the quality of a partition** i.e. a loss function $J(partition)$
   - Are the clusters optimal?
   - Are the clusters meaningful ?
3. **An algorithm to perform the clustering**
   - Typically attempting to minimise the loss J

Distance measures

- A “good” distance measure is application dependant!
- The clusters should be invariant to transformations that are “natural” for the problem
- Rotation invariant ?


Same ?



Same ?

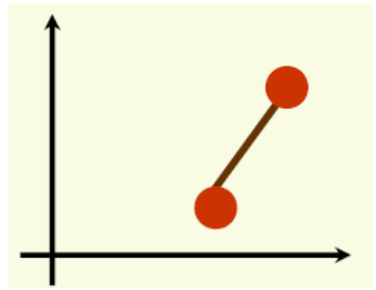

- Importance of features - colour for car brands clustering?


Similar ? Same ?

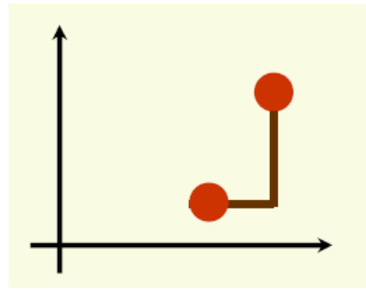
Distance measures

- Typical distance measures used:
  - Measure **dissimilarity**
    - Euclidean distance
    - Manhattan distance (aka city-block)
  - Some distance measures the **similarity**
    - Cosine similarity
      - Smaller angle —> similar
      - Invariant to scale
    - Correlation coefficient
      - Popular in image processing

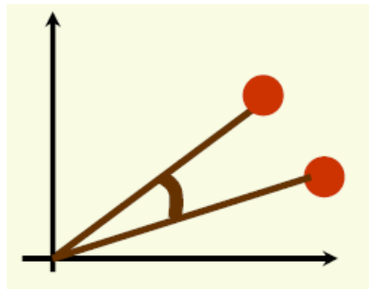
Euclidian



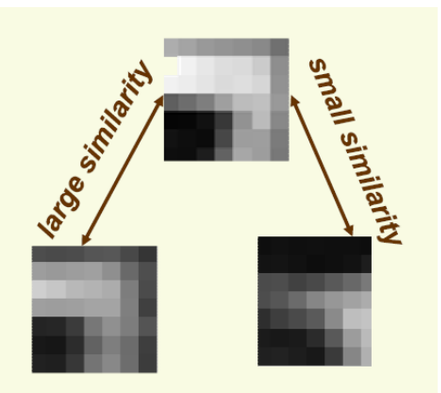
Manhattan



Cosine

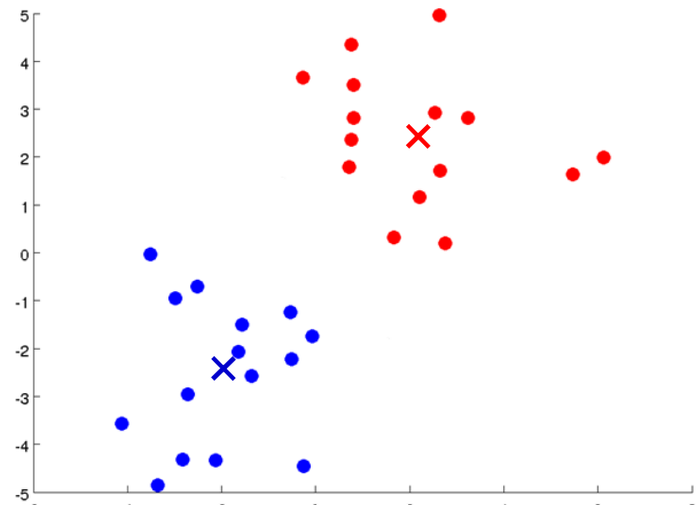


Correlation



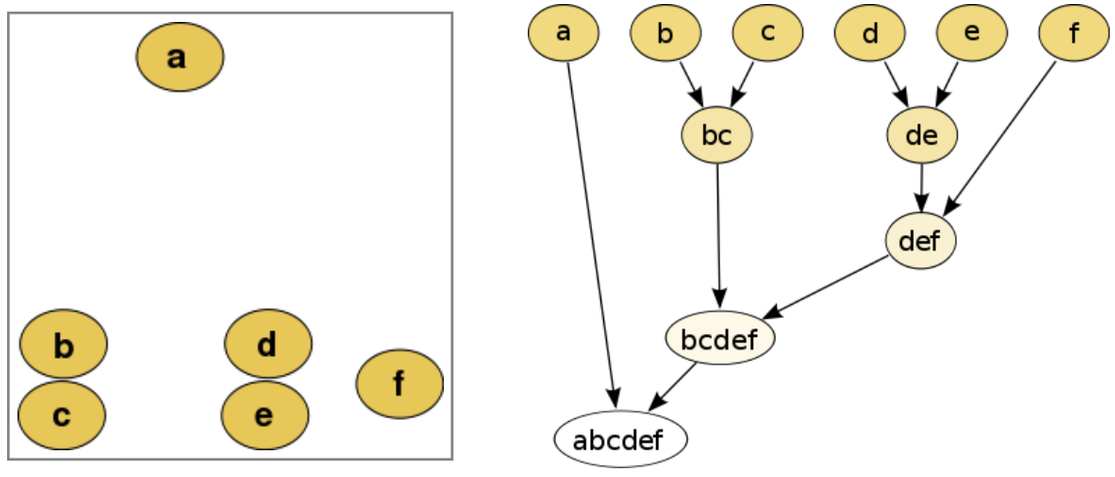
Criterion to evaluate partitions

- Distortion
  - How close are we to a “center of gravity” defining the partition?
  - Typically used by *k*-means



By A. Ng, Machine Learning Class, Stanford

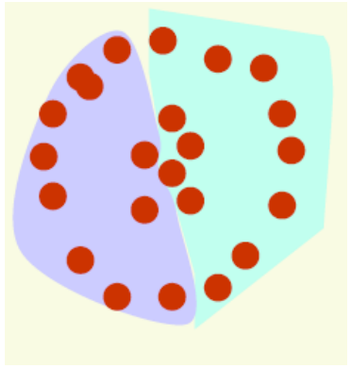
- Connectivity of points
  - How close are pairs of points?
  - Typically used in hierarchical clustering
  - Example of nearest neighbour clustering on the right



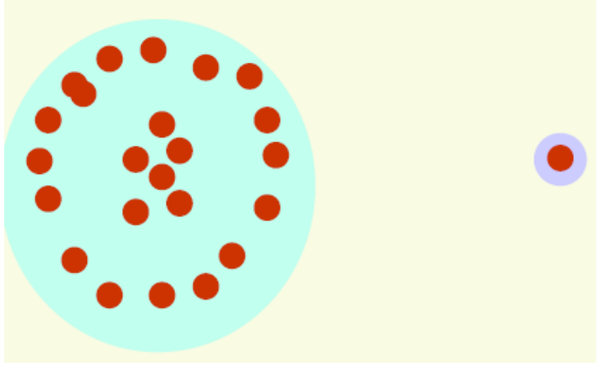
By [[File:Hierarchical_clustering_diagram.png#file]]: Stathis Sideris on 10/02/2005

Choice of the criterion

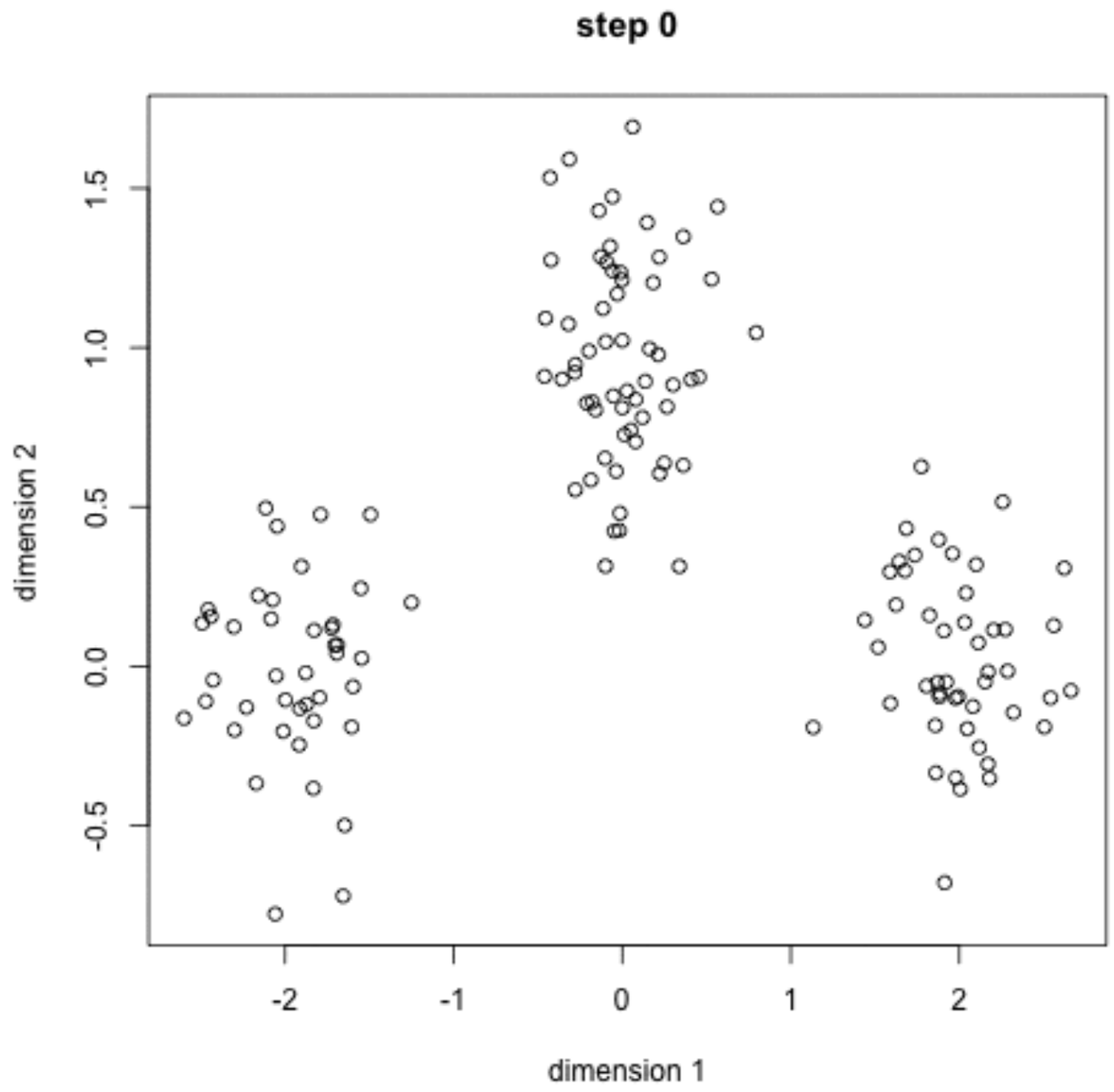
- Will depend to the nature of the problems and to the natural distributions of training points
- For example, with *K=2*
  - distortion will fail on



  - connectivity will fail on

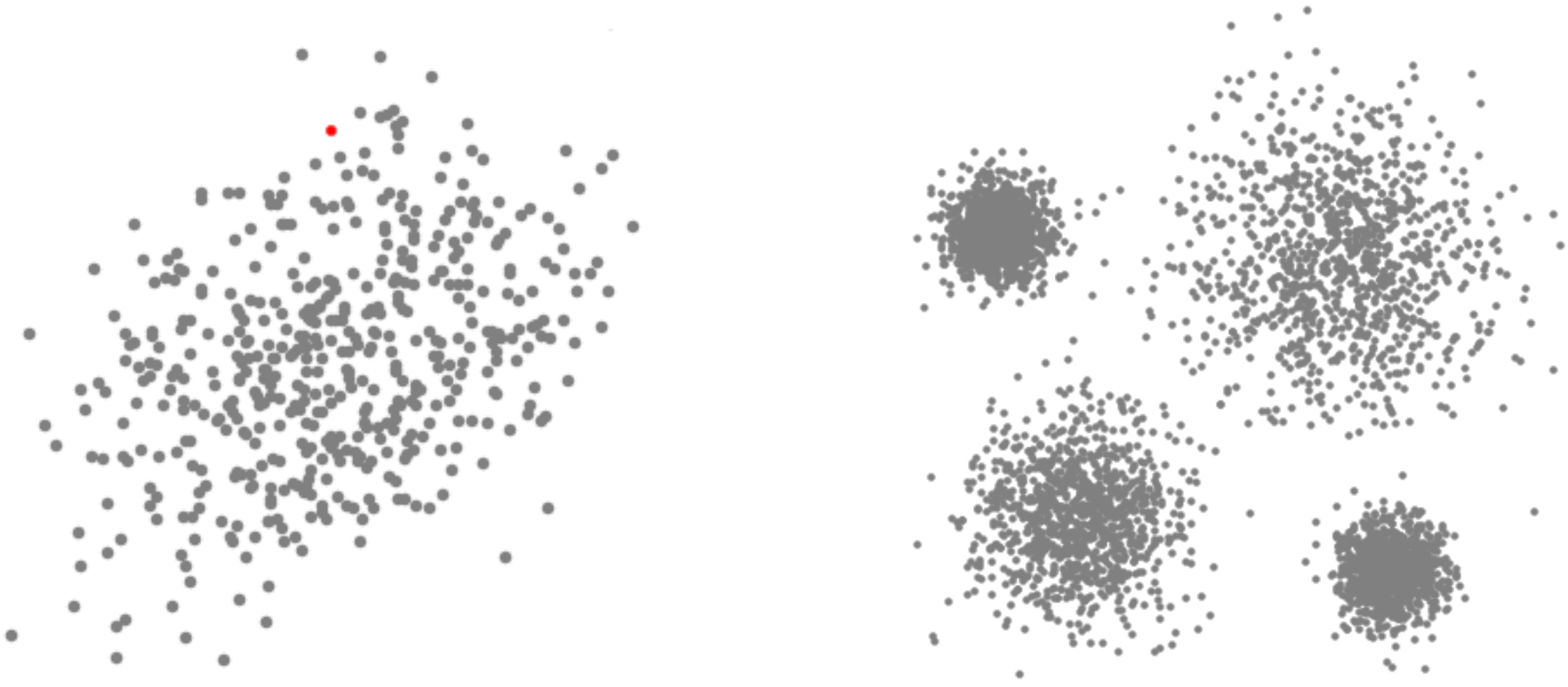


Overview of algorithms - K-Means



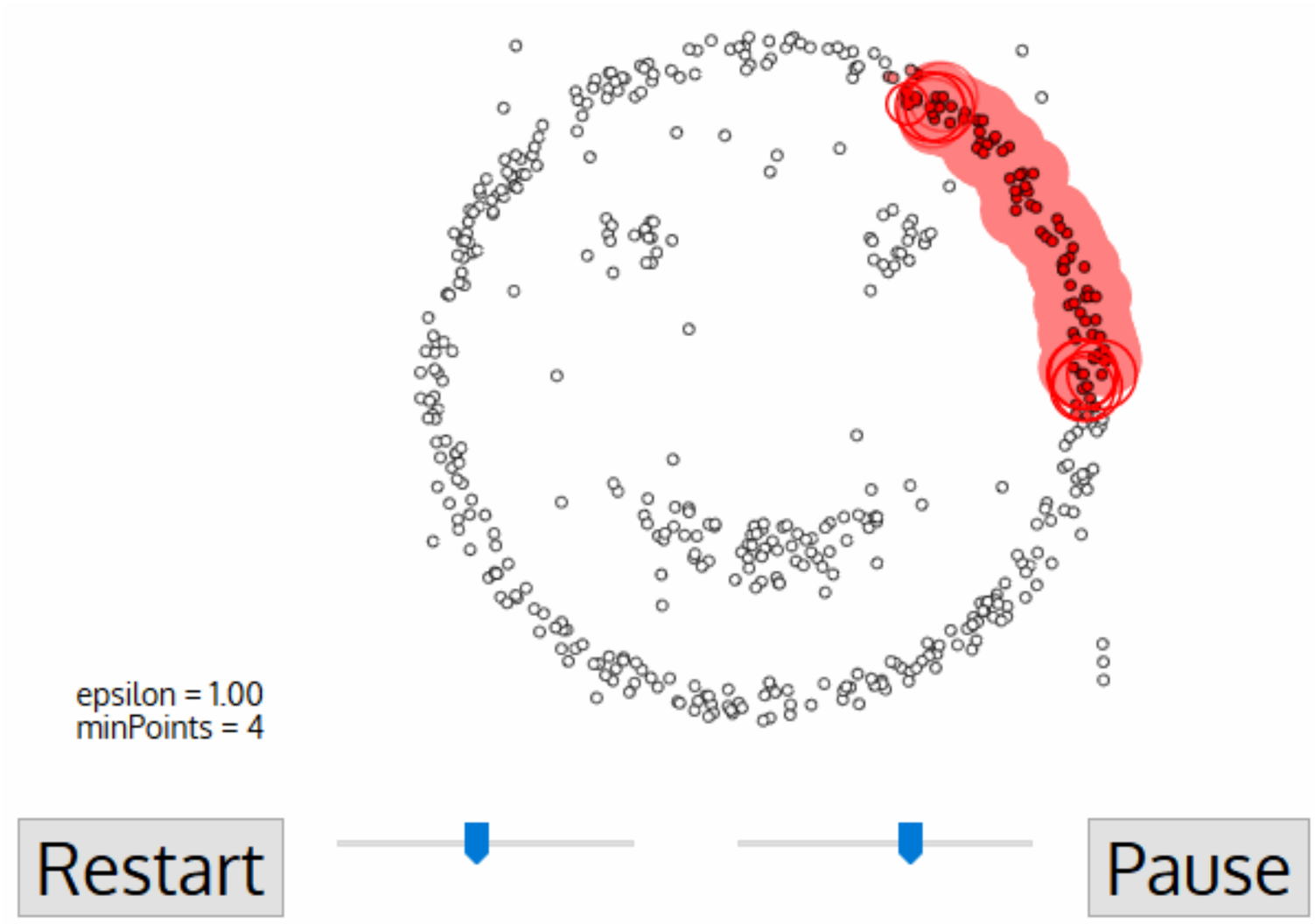
- See next Section for more info

Overview of algorithms - Mean-Shift Clustering



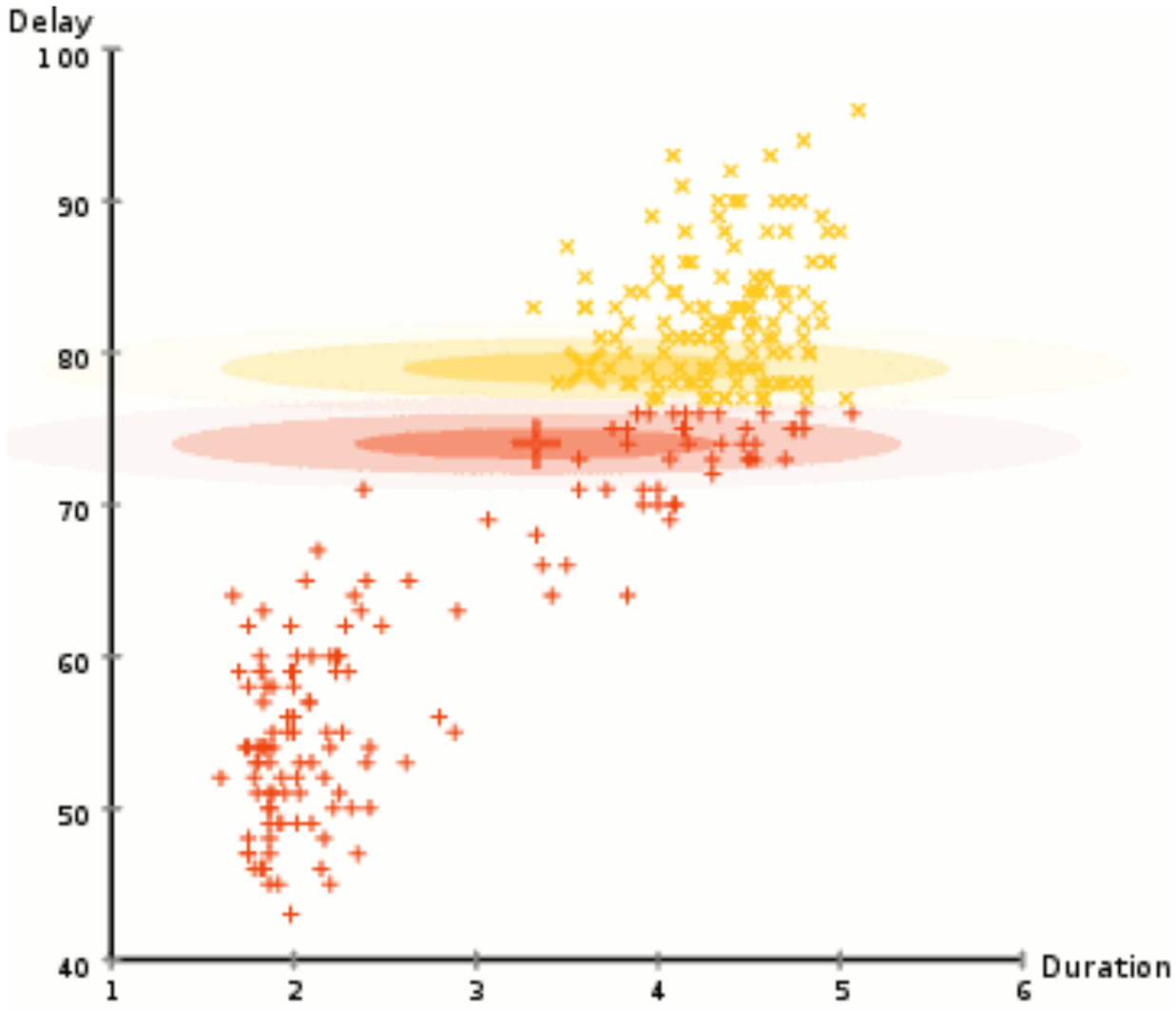
- Sliding window based algorithm: find the dense areas of data points
- Update the candidate center points to the mean of the points of the window

Overview of algorithms - DBSCAN



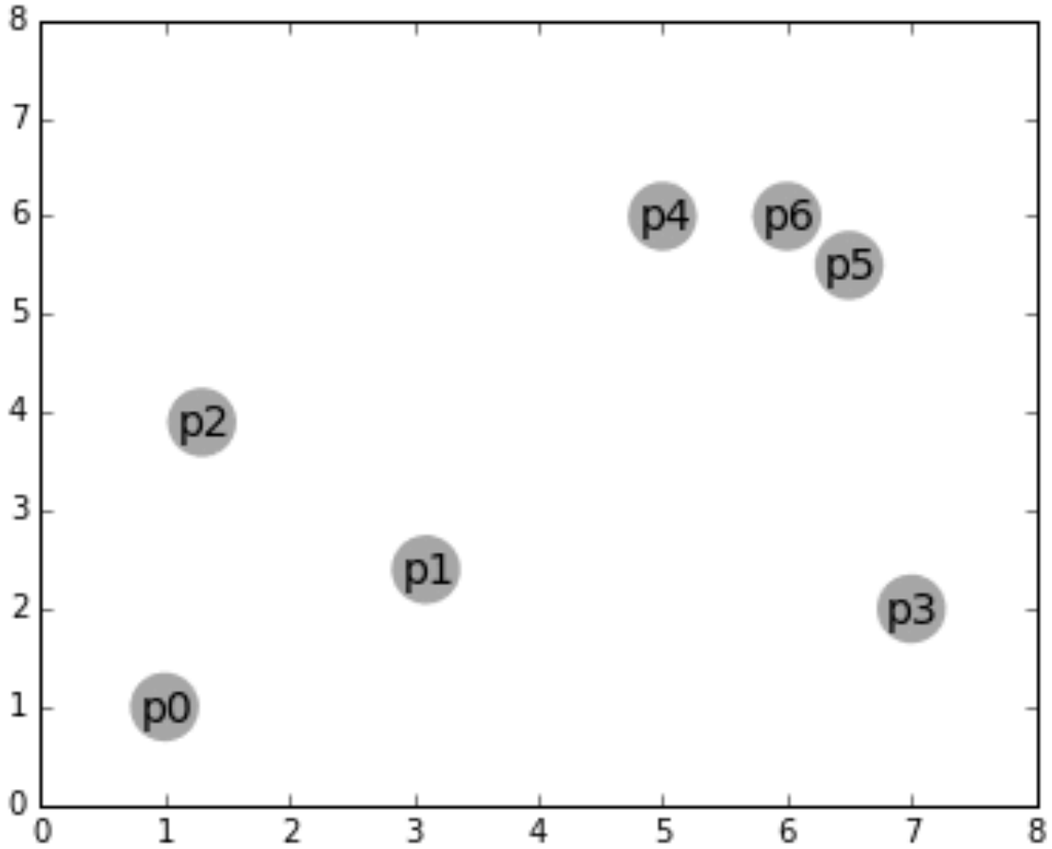
- DBSCAN = Density-Based Spatial Clustering of Applications with Noise
- Points are added to a cluster within a neighbourhood of the previously added points to the cluster.

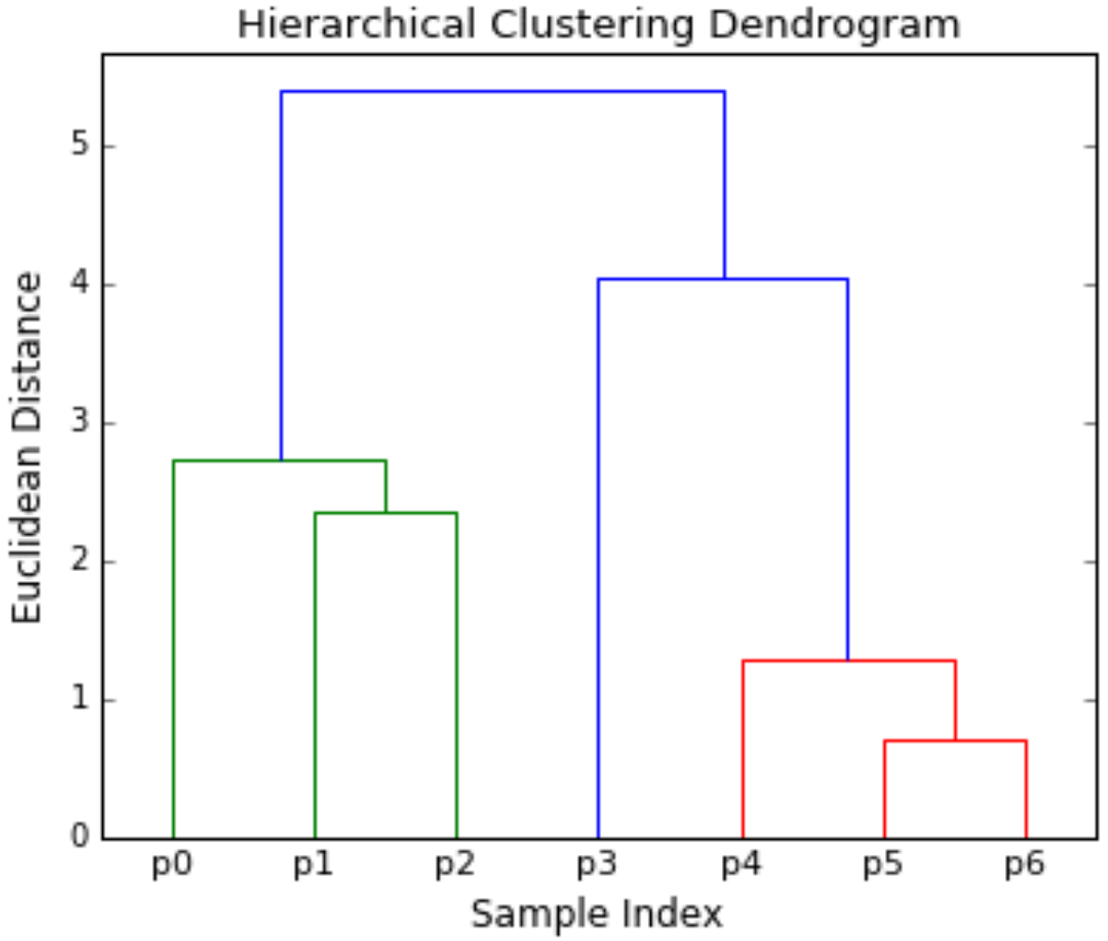
Overview of algorithms - Gaussian Mixture Models



- Similar to k-Means but considers a Gaussian probability density function instead of a distance metric

Overview of algorithms - Agglomerative Hierarchical Clustering





- Bottom-up approach: merge clusters hierarchically from most granular (all points are individual clusters) to less granular (all points are 1 cluster).
- Works well when there is a pre-supposed notion of hierarchy in the data

Clustering with k-means

Definition

Algorithm

Examples

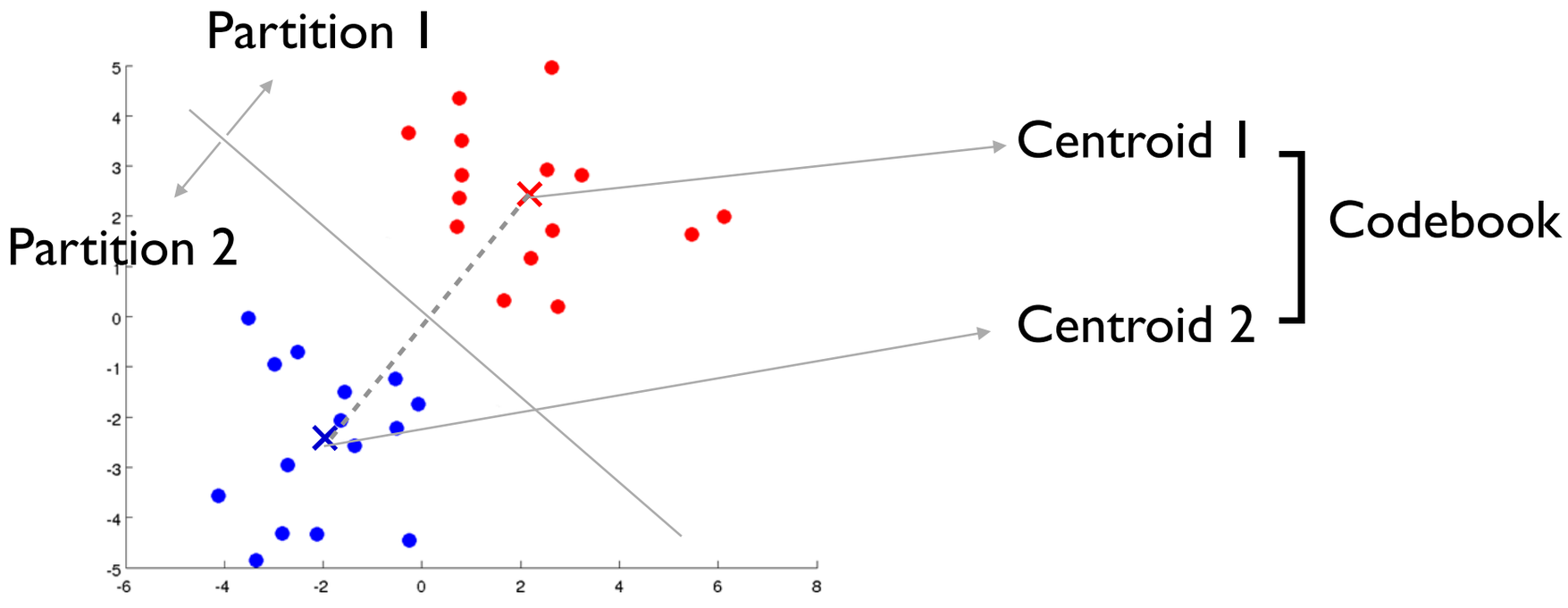
k-means algorithm

Terminology

A **centroid** is the center of a cluster

A **codebook** is the ensemble of all centroids

A **partition** is the ensemble of samples attributed to a centroid (to a cluster)



k-means algorithm

Algorithm

a) Chose a value for K, the number of clusters

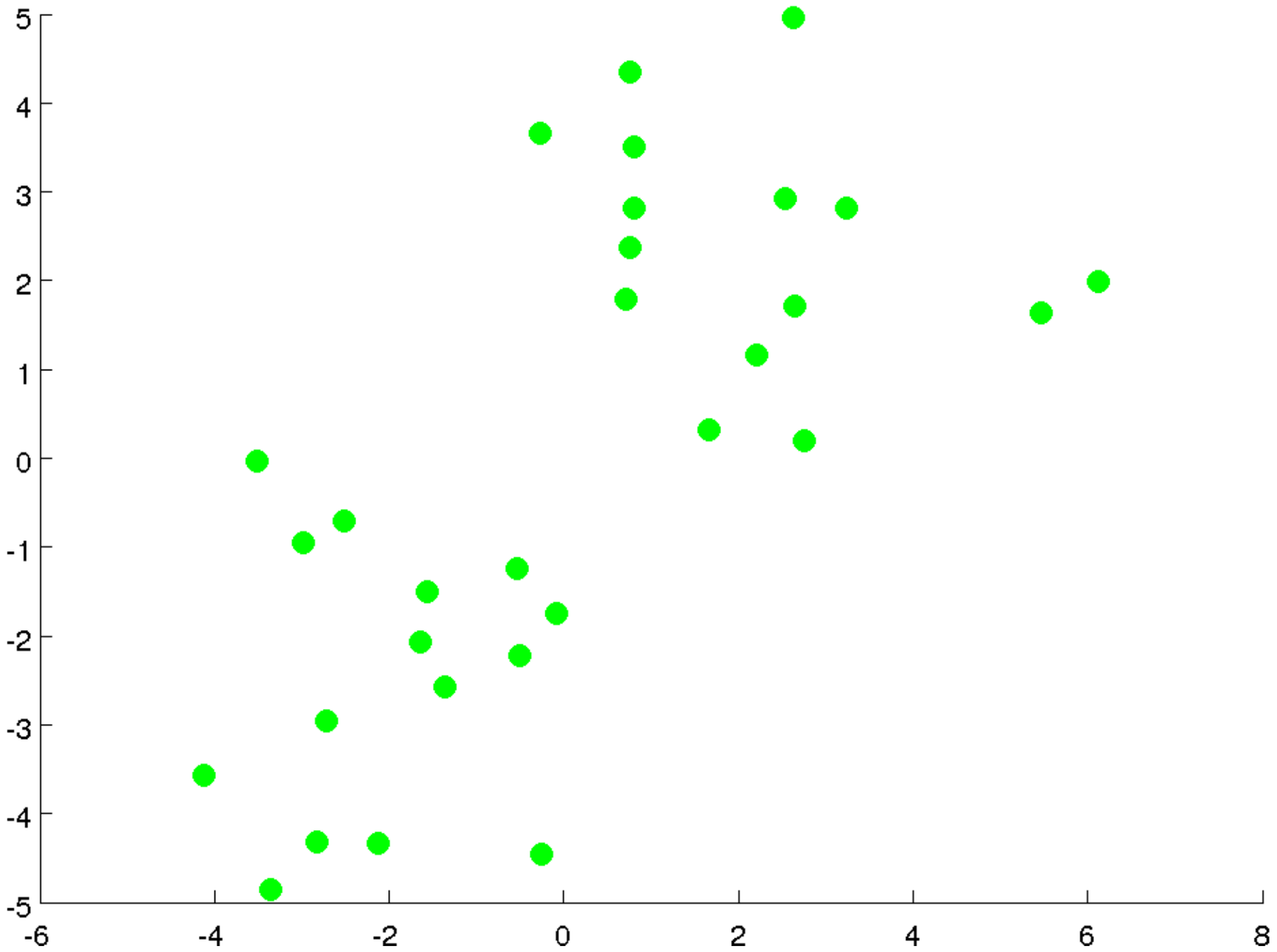
b) Initialize the centroids $\mu_1, \mu_2, \ldots, \mu_K$ randomly

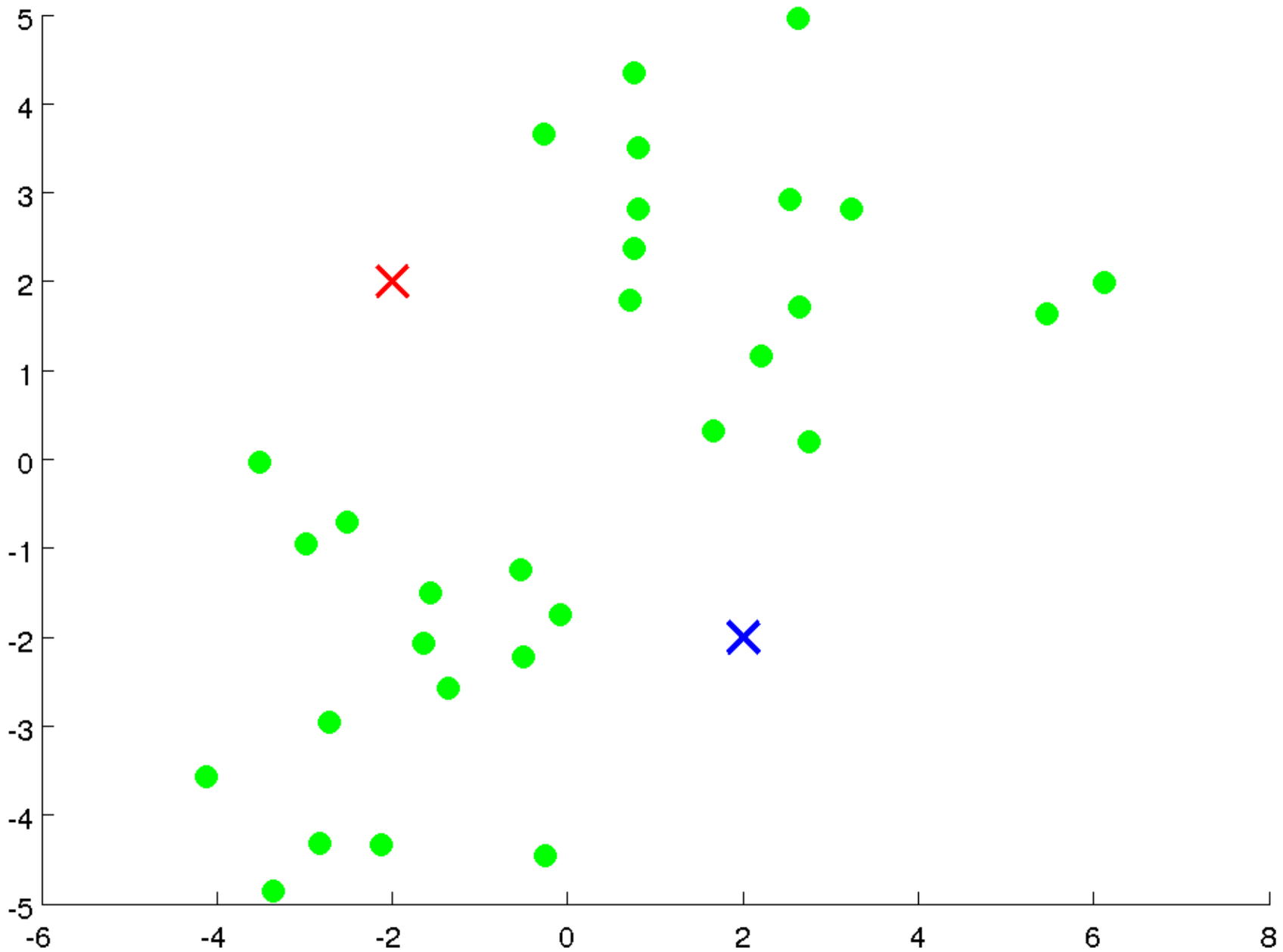
c) Repeat until convergence :

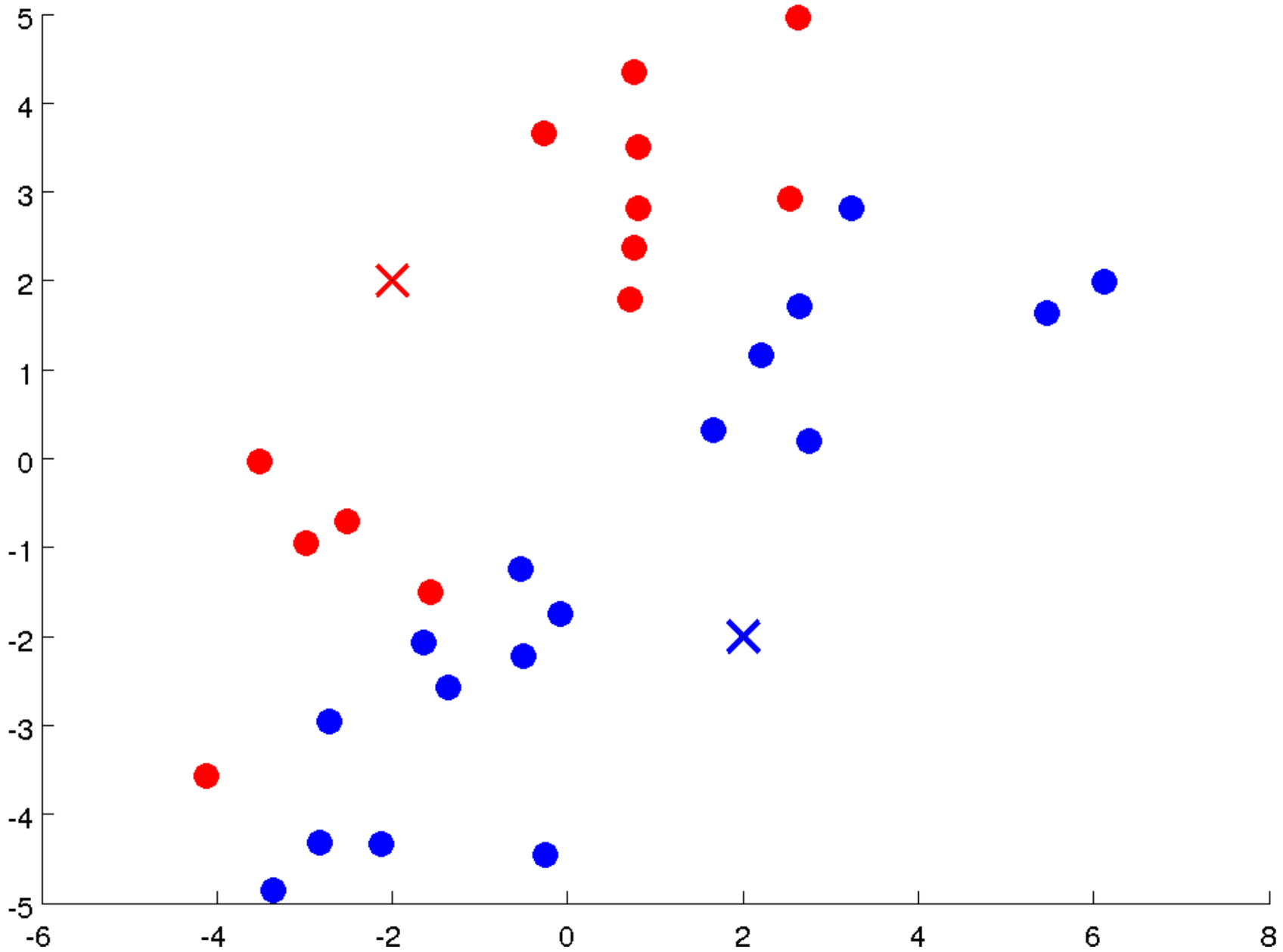
  i) For each training sample x_n, set c_n as the index of the centroid closest to x_n

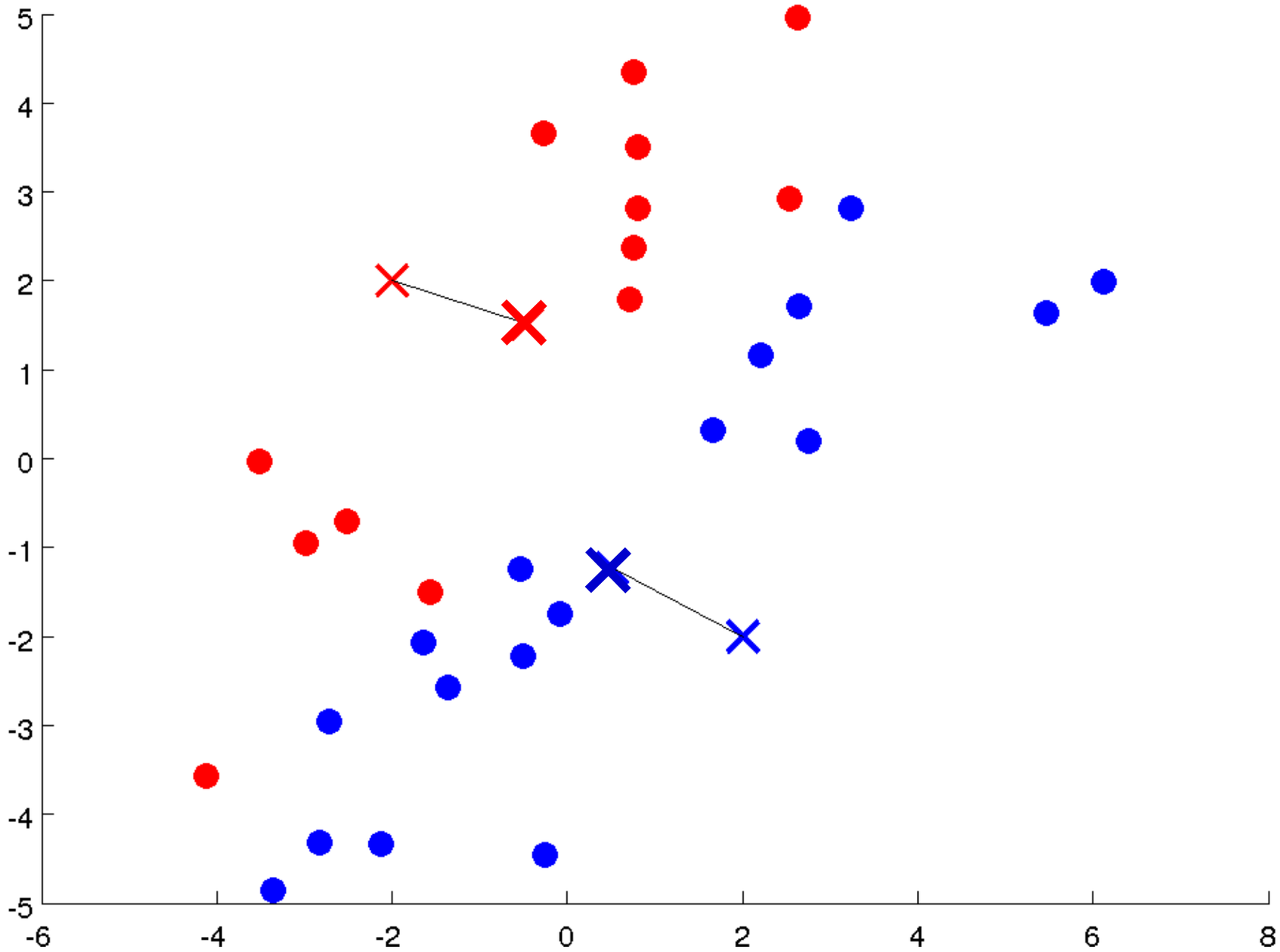
  ii) For each centroid, compute its new value μ_k as the average of training sample associated to cluster k, i.e. using the set of samples $\{\mathbf{x}_n; c_n = k\}$

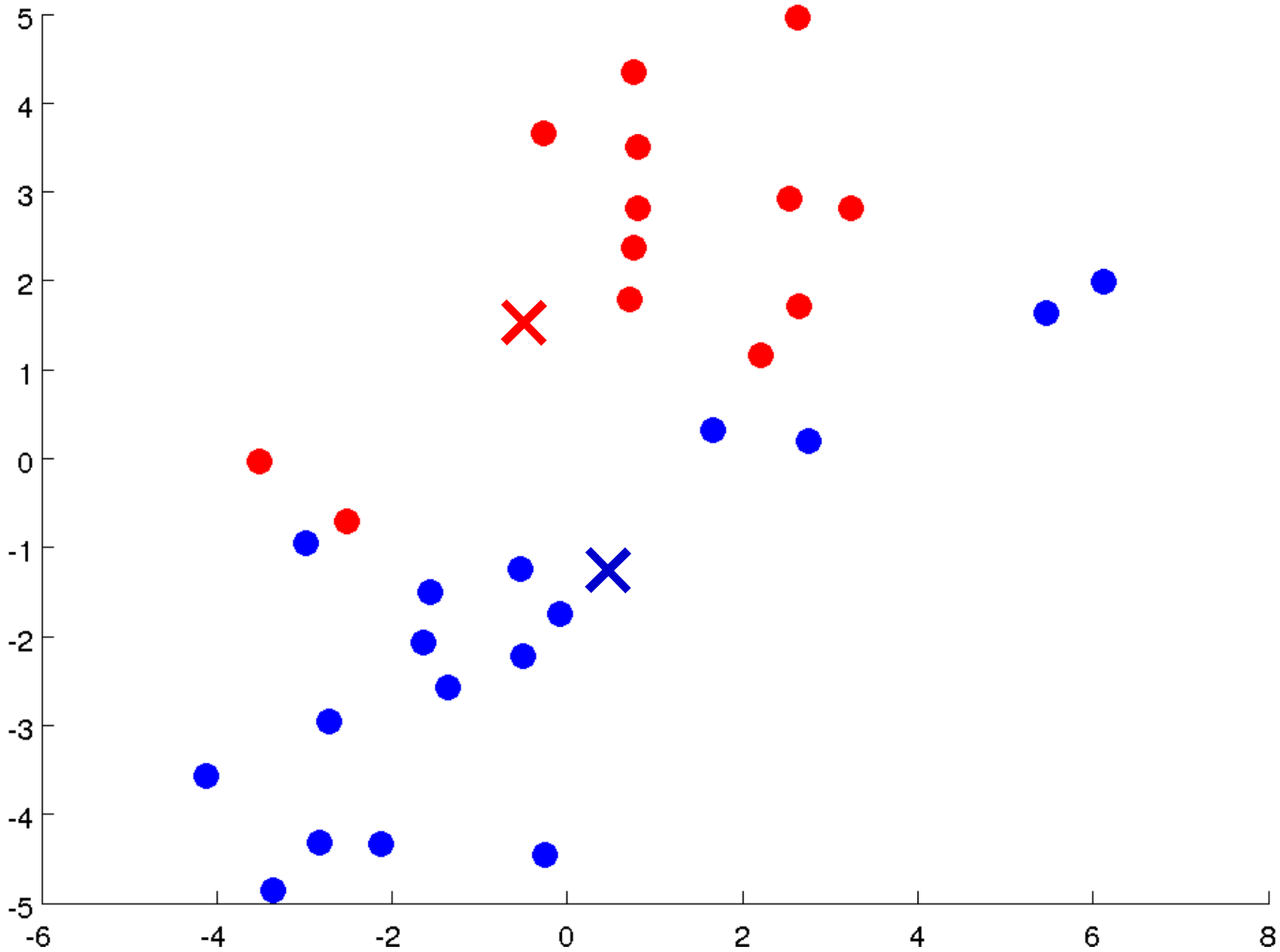
- This algorithm is mathematically guaranteed to converge
- i) is called **accumulation phase** = partition computation
- ii) is called **adaptation phase** = codebook update

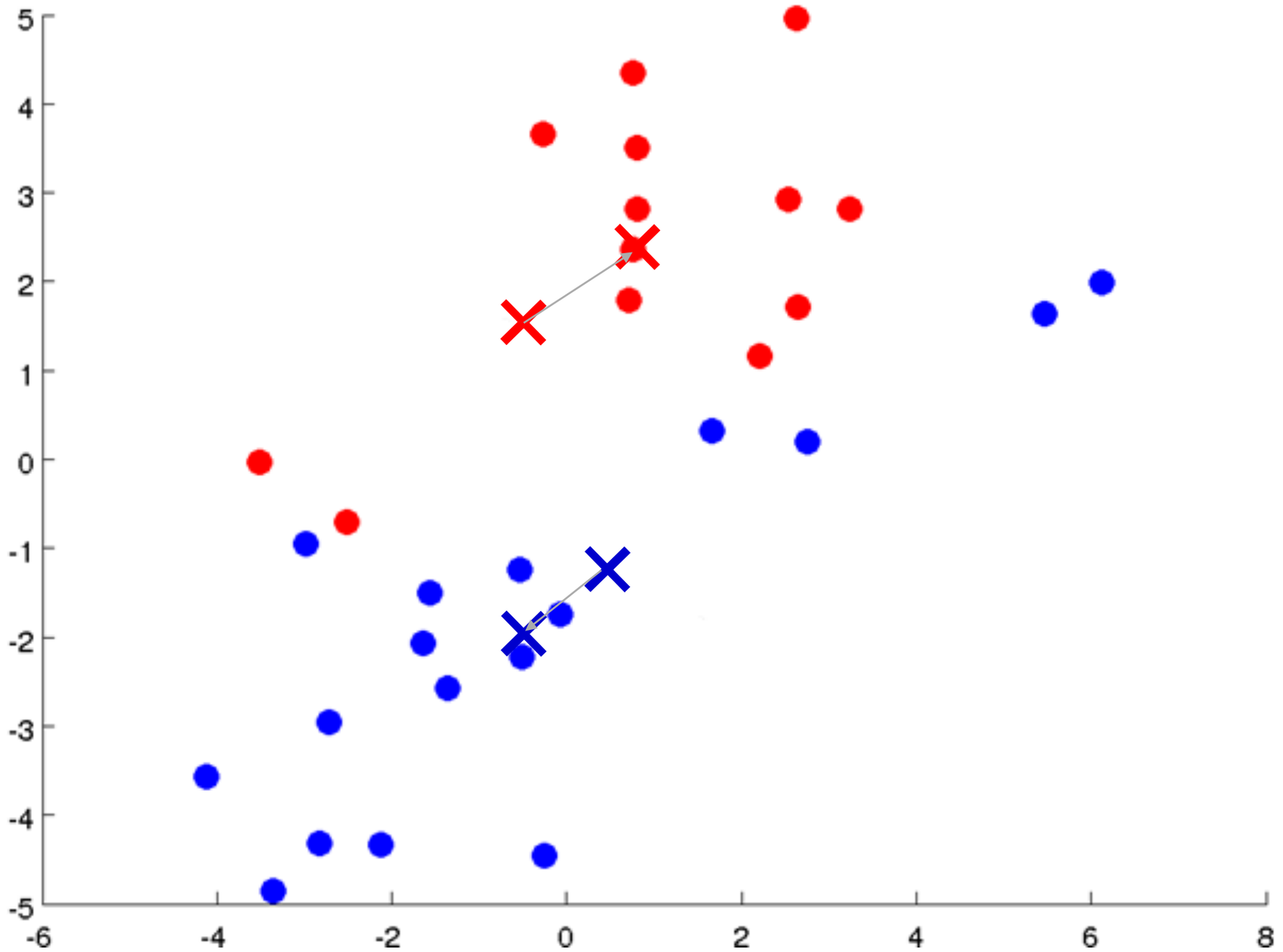

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

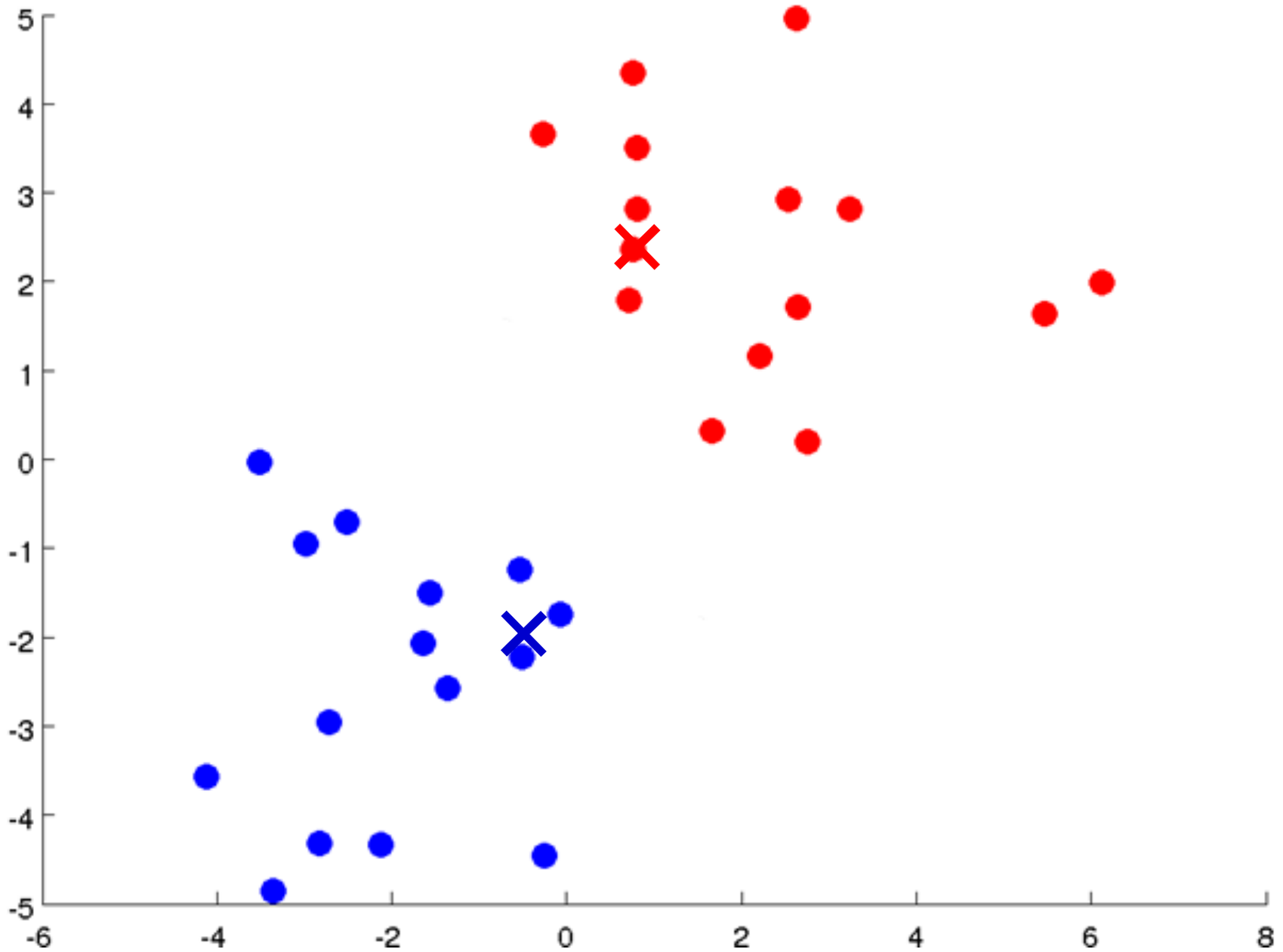

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

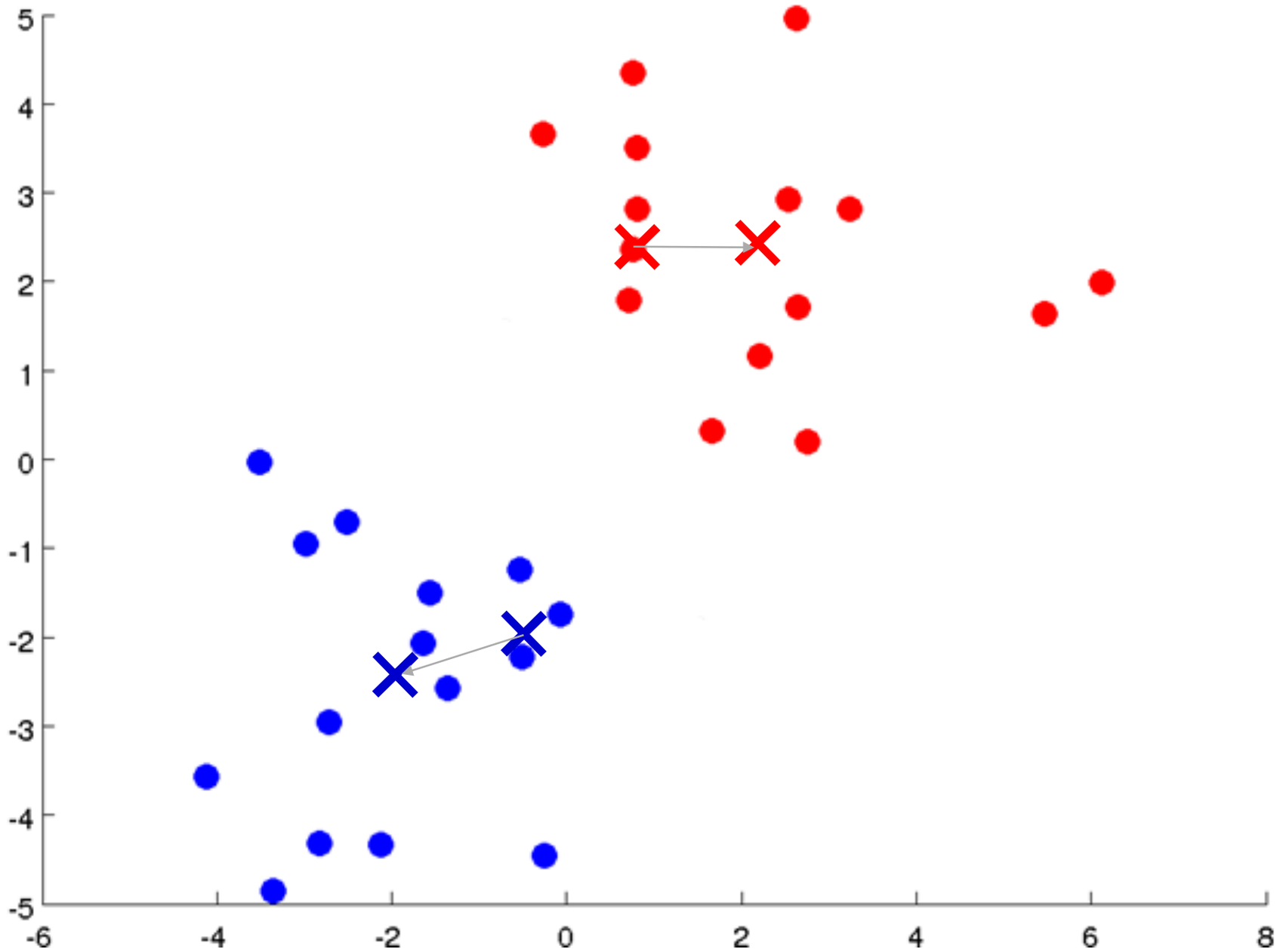

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

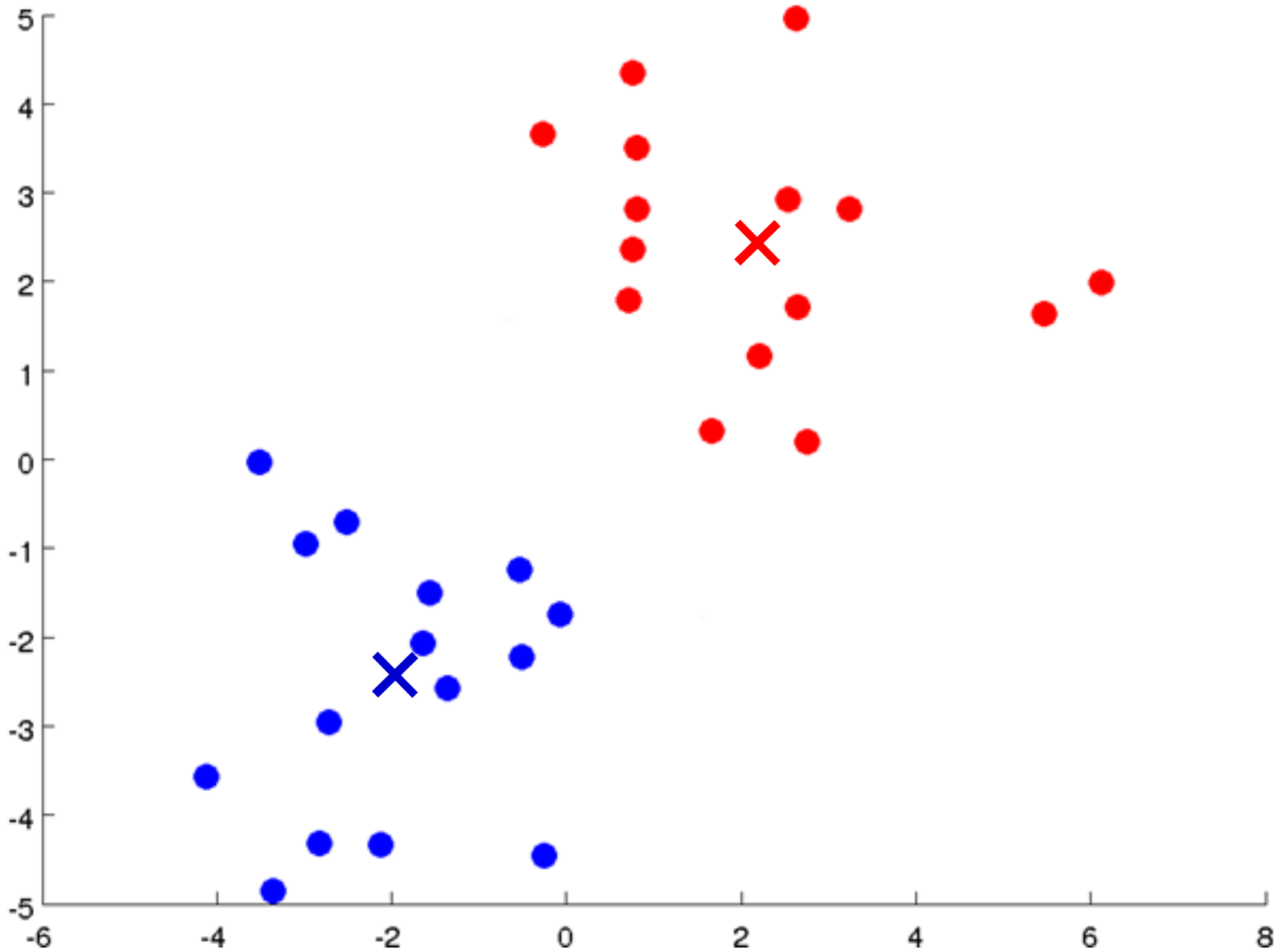

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8



From this point on, more iterations will not change the positions of the centroids

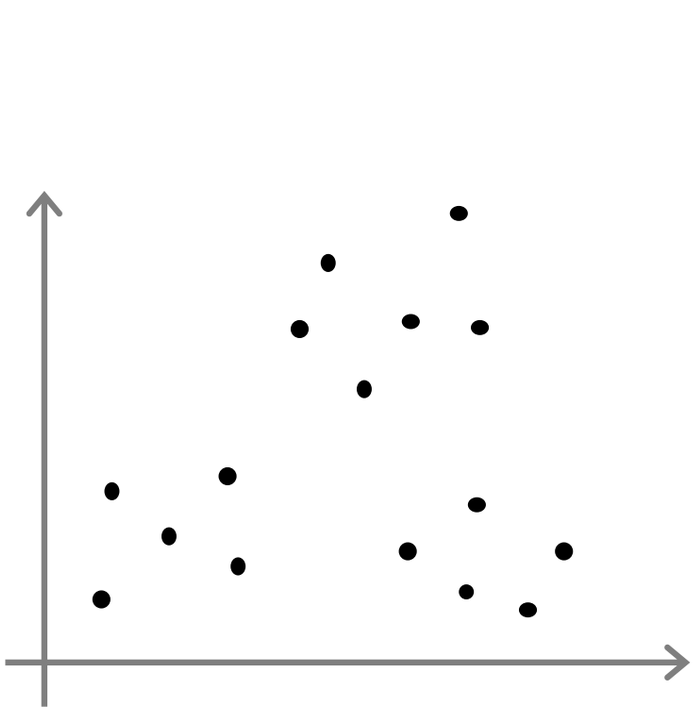
k-means optimisation objective

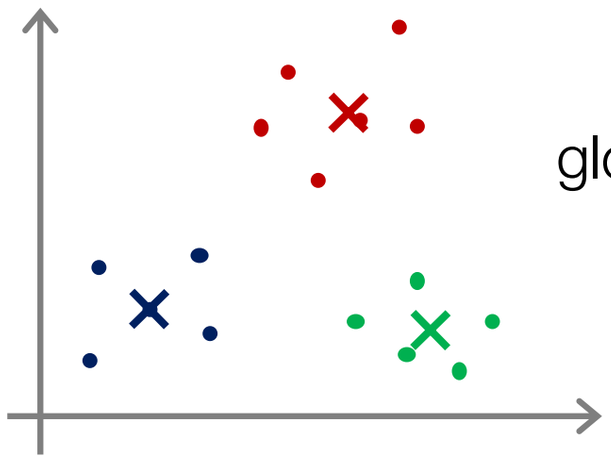
- The *k*-means algorithm is mathematically guaranteed to converge to a minimum of the distortion.
- It will actually minimise iteratively the distortion defined with

$$J(c, \mu) = \sum_{n=1}^{N} d(x_n, \mu_{c_n})^2$$

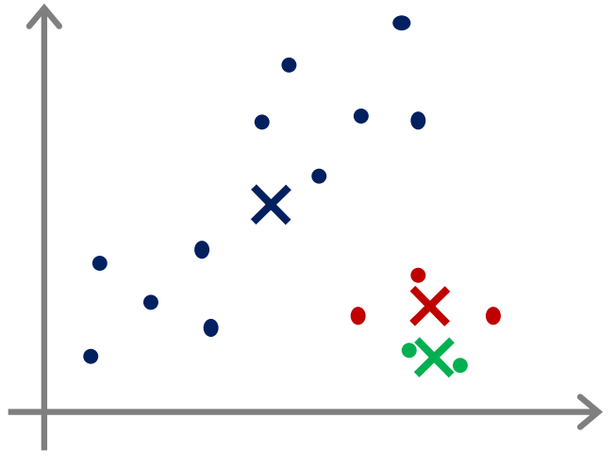
- The *k*-means may fall into local minima (as *J* is non-convex).
  - An option to avoid local minima is to run *k*-means many times using different initial values for the cluster centroids.
  - Then, out of all the different clustering, keep the one that gives the lowest distorsion *J*.

k-means example of local optima

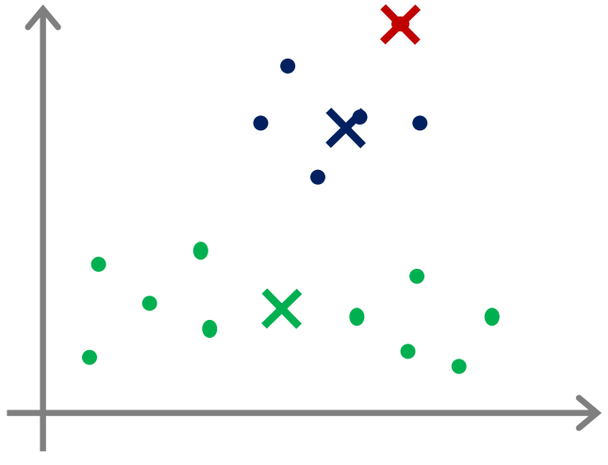




global optimum



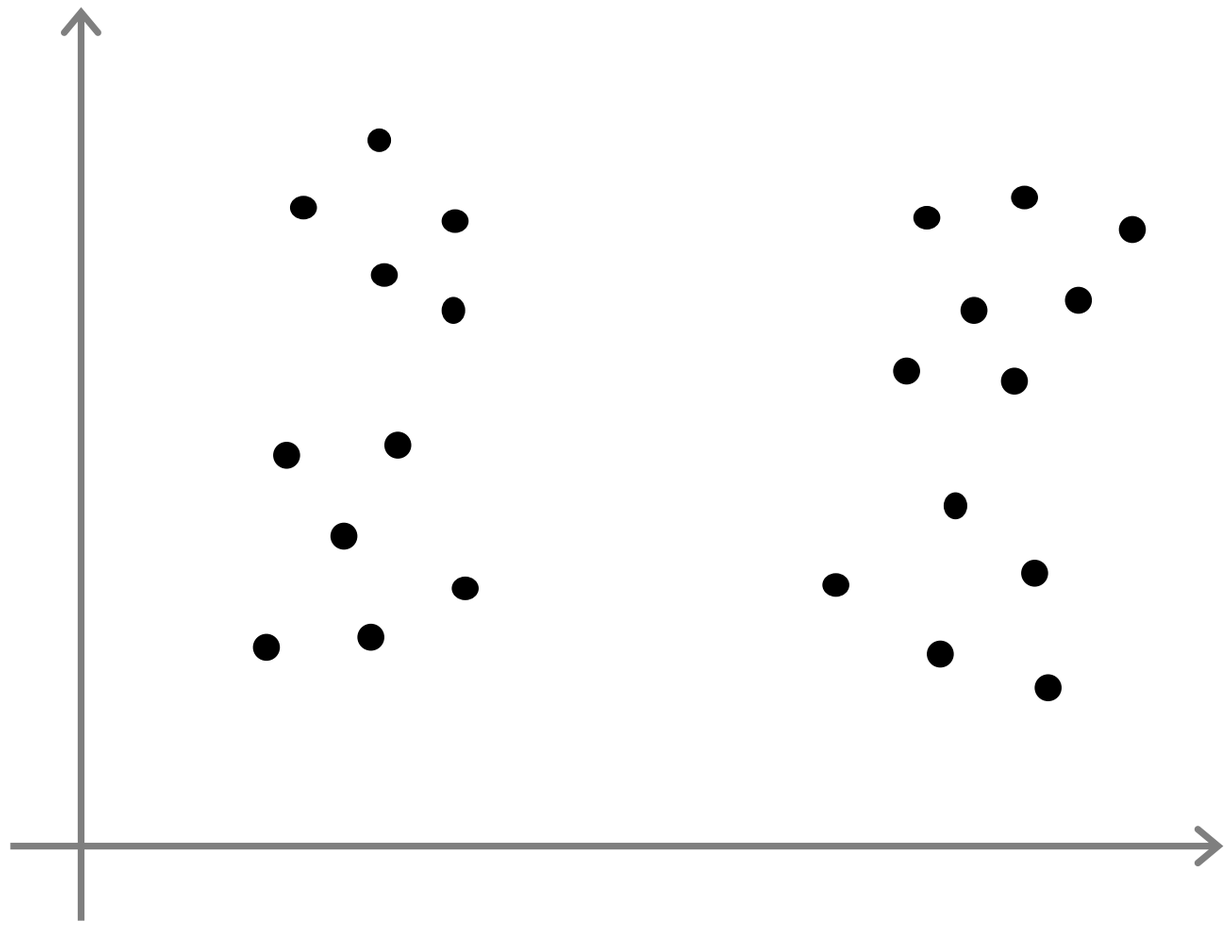
local optimum



local optimum

Choosing *K*

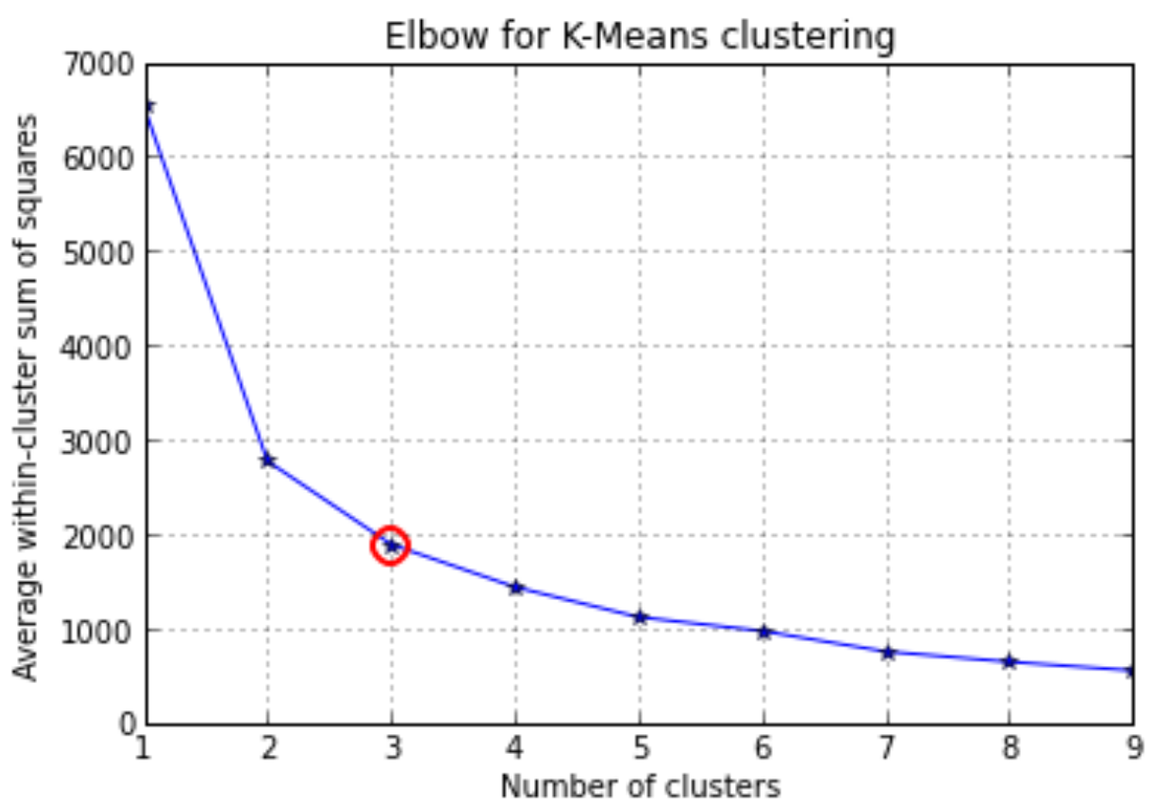
- What is the right value of *K* here?



Choosing K - when clusters are “clearly” separated, we can use the elbow method

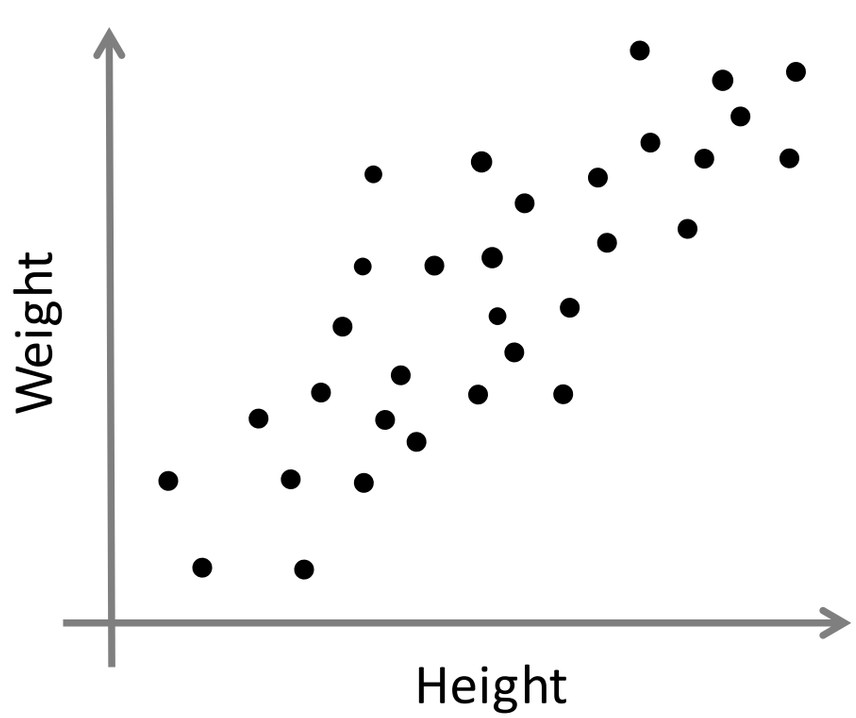
- Increase *K* and observe the average distortion in each cluster
  - Large decreases indicate we are not yet to the optimal value for *K*
  - Small decreases indicate that having more centroids does not help

J

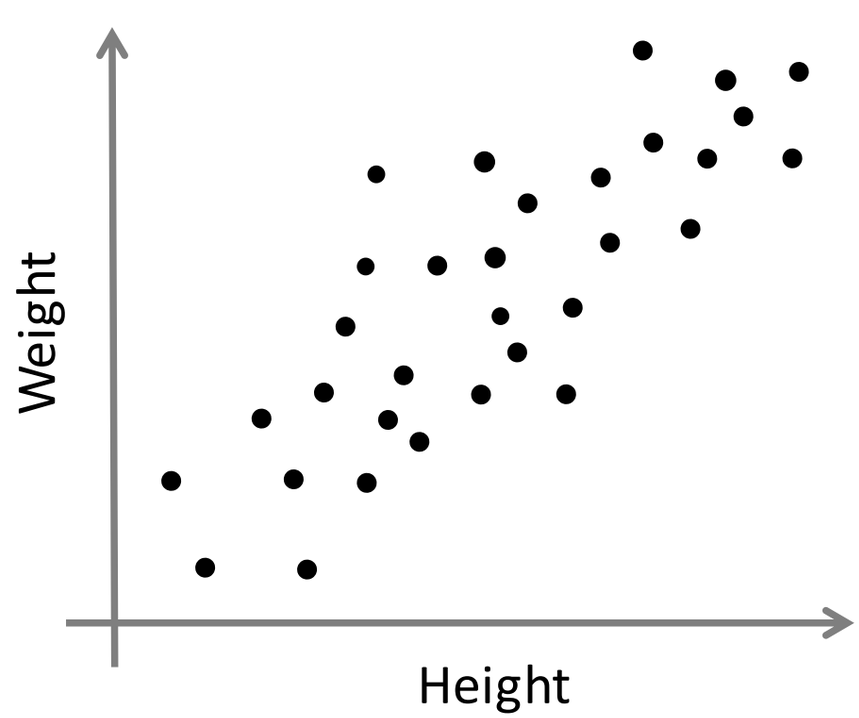


Choosing *K* -a priori- , i.e. due to the application purpose

K=3 S,M,L

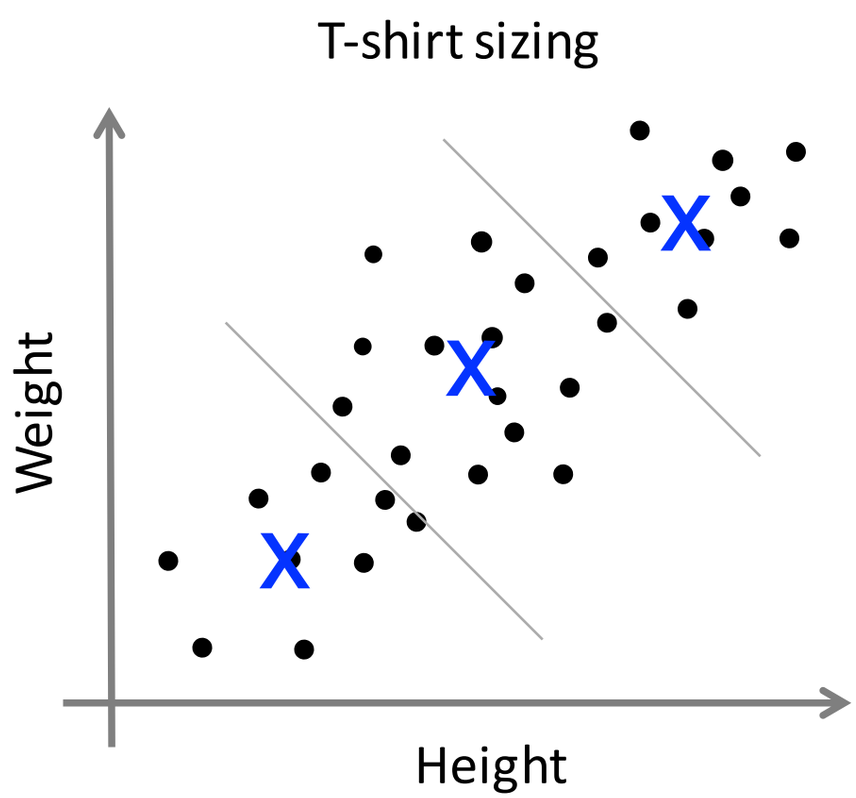


K=5 XS,S,M,L,XL

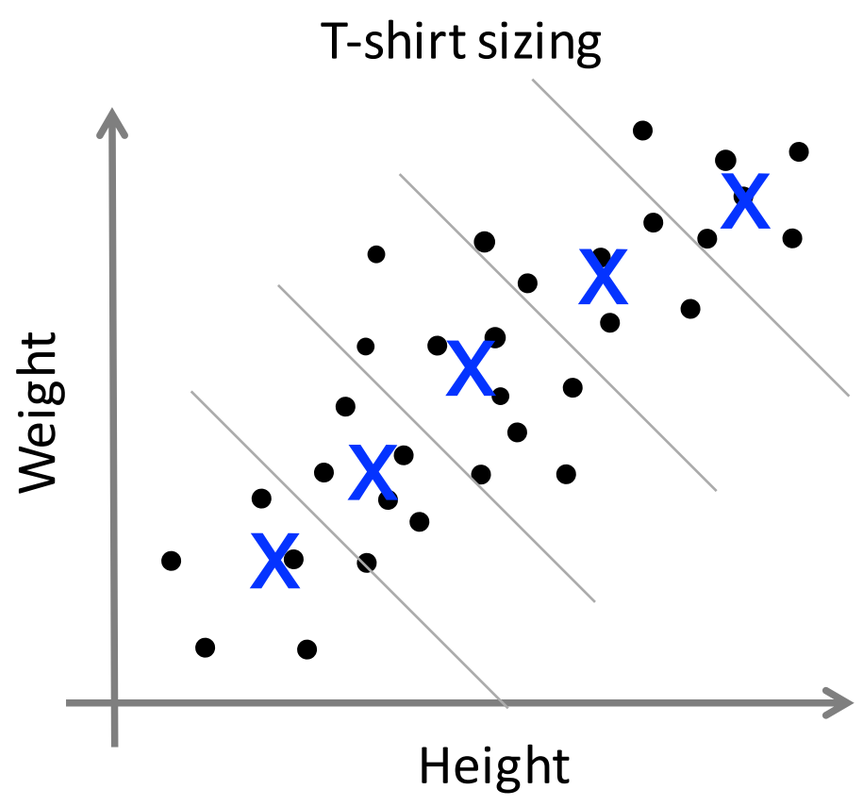


Choosing *K*- a priori set due to the application purpose

K=3 S,M,L



K=5 XS,S,M,L,XL



Use of *k*-means

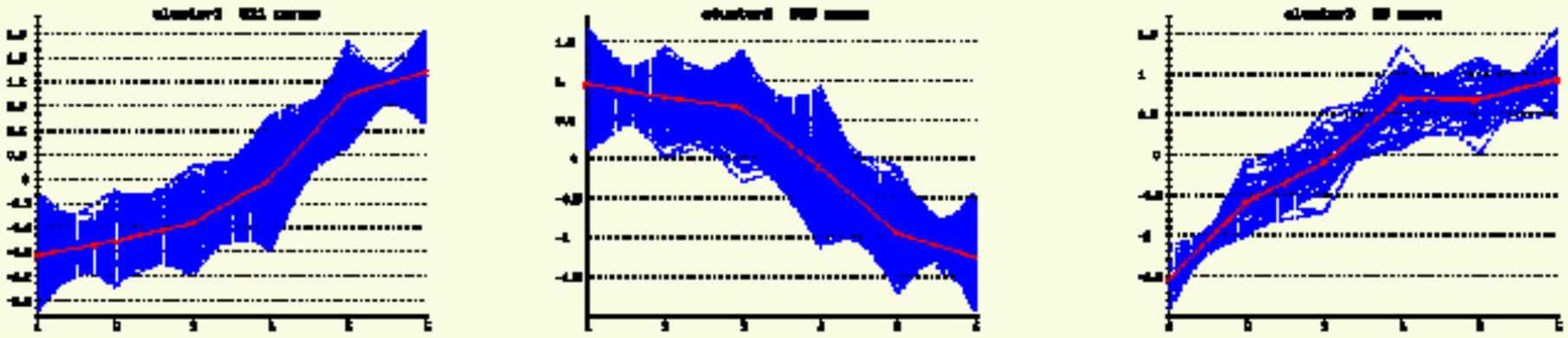
- Marketing: finding groups of customers with similar behaviour
- Production: finding how many t-shirt to produce according to sizes
- Biology: grouping of unknown plants and/or animals
- Libraries: book ordering
- Web: documents classification (e.g. news grouping), users grouping

- Other use:
  - Compression - to encode the input space in a smaller representations (the centroid indices)
    - using a codebook with 256 centroids, any sample data can be encoded with 8 bits
    - This is also called Vector Quantization
  - Initialisation for other algorithms such as GMMs

Example of vector quantisation - reducing the color space

Example of clustering in bio-informatics

U18675 4CL -0.151 -0.207 0.126 0.359 0.208 0.091 -0.083 -0.209
M84697 a-TUB 0.188 0.030 0.111 0.094 -0.009 -0.173 -0.119 -0.136
M95595 ACC2 0.000 0.041 0.000 0.000 0.000 0.000 0.000 0.000
X66719 ACO1 0.058 0.155 0.082 0.284 0.240 0.065 -0.159 -0.010
U41998 ACT 0.096 -0.019 0.070 0.137 0.089 0.038 0.096 -0.070
AF057044 ACX1 0.268 0.403 0.679 0.785 0.565 0.260 0.203 0.252
AF057043 ACX2 0.415 0.000 -0.053 0.114 0.296 0.242 0.090 0.230
U40856 AIG1 0.096 -0.106 -0.027 -0.026 -0.005 -0.052 0.054 0.006
U40857 AIG2 0.311 0.140 0.257 0.261 0.158 0.056 -0.049 0.058
AF123253 AIM1 -0.040 0.002 -0.202 -0.040 0.077 0.081 0.088 0.224
X92510 AOS 0.473 0.560 0.914 0.625 0.375 0.387 0.019 0.141



From:De Smet F., Mathys J., Marchal K., Thijs G., De Moor B. & Moreau Y. 2002.
Adaptive Quality-based clustering of gene expression profiles, Bioinformatics, **18**(6), 735-746.

Conclusions

- Unsupervised learning = discover a good internal representation of the input from which we can make some sense
- Clustering is a good and widely used strategy to find inherent structure in data
- Clustering needs:
  - A distance measure
  - A criterion to evaluate the partition
  - An algorithm to perform the clustering
- *k*-means is such an algorithm
  - Typically an Euclidian distance
  - Criterion = distortion = proximity of points to centroids
  - Algorithm is guaranteed to converge