NATIONAL RESEARCH UNIVERSITY

Predictive Modelling, 2020/2021

DAY 7

Sergey Shvydun,
shvydun@hse.ru

05.11.2020

SYLLABUS

1. **Introduction to predictive modelling**. Example of real life applications and projects (10/09/2020);
2. **Predictive modeling process**: basics of statistics, data splitting (17/09/2020).
3. **Reducing the dimension**: factor analysis, principal component analysis, etc (24/09/2020).
4. **Regression models**: linear/logistics regressions, probit and logit models, etc (1/10/2020).
5. **Time series analysis**: analysis of stationary time series, forecasting, etc (15/10/2020, 29/10/2020).
6. **Classification models**: k-nearest neighbors, SVM, neural networks, superposition models, etc.;
7. **Clustering and pattern analysis**: state-of-the-art clustering and patter analysis methods;
8. **Markov Chain Monte Carlo (MCMC) methods**: Markov processes, goals and applications of MCMC;
9. **Dynamic linear models**: Bayesian framework. State space models. State estimation and forecasting.

In total, 12 days.

LECTURE 7: PLAN OF THE DAY

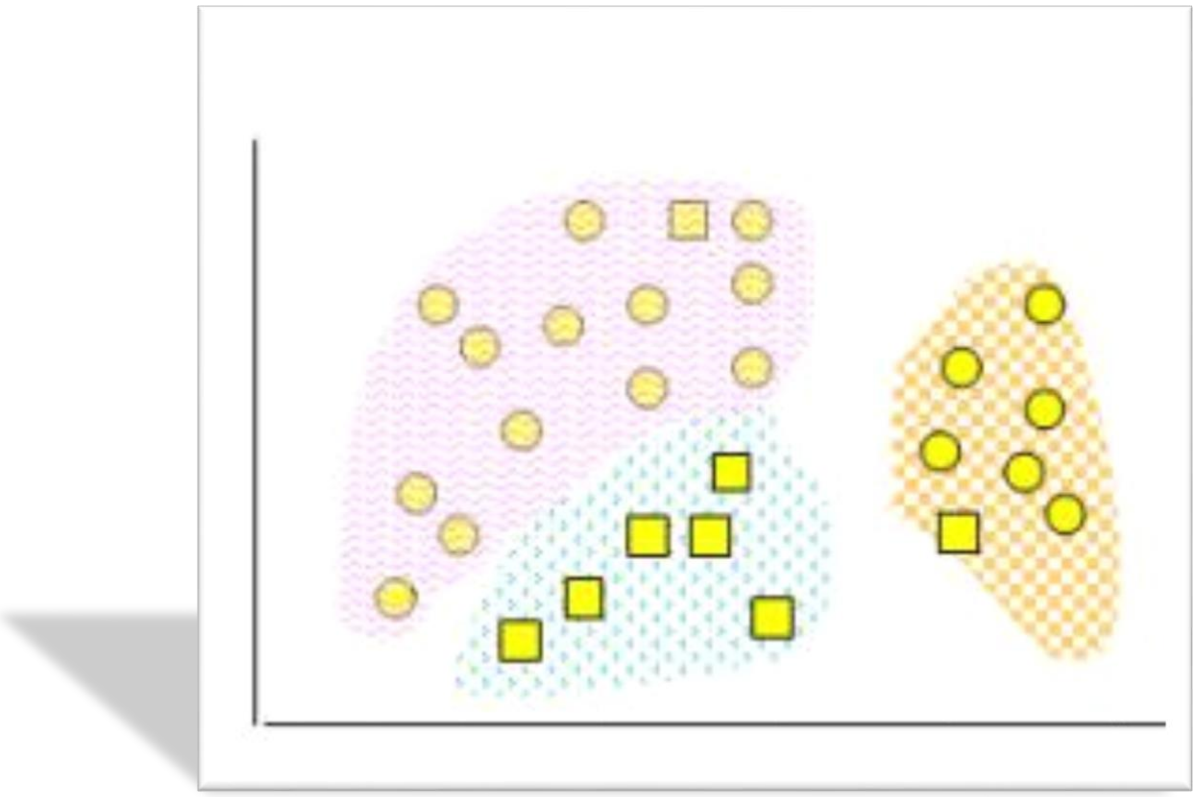
1. Clustering and pattern analysis: state-of-the-art clustering and patter analysis methods.

2. Exercises in Python.

CLUSTERING

Idea: the grouping of dissimilar unlabeled items into a certain number of similar subgroups or clusters. The attribute for clustering is unknown in advance.

- high intra-cluster similarity;
- low inter-cluster similarity.



- Clustering is **unsupervised** learning technique.

SUPERVISED VS UNSUPERVISED LEARNING

- **Supervised learning:** Given (X,Y) learn a function *f: X⟶Y, where*
  - *Categorical Y: classification;*
  - *Continuous Y: regression.*

- **Unsupervised learning**: Given only X can we infer the underlying structure of X?

WHY DO UNSUPERVISED LEARNING?

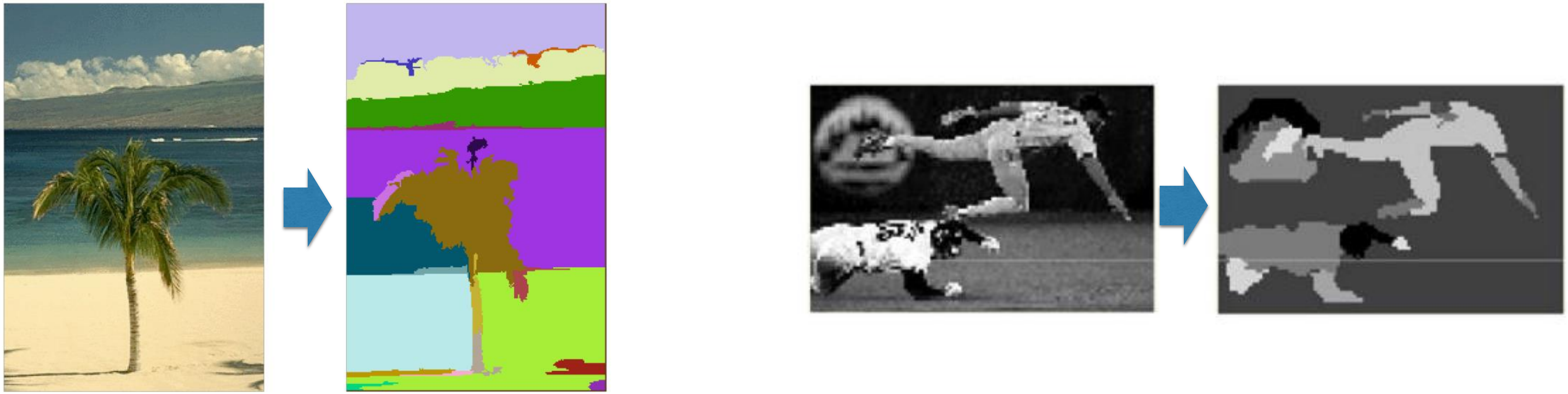
- Raw data cheap. Labeled data expensive.
- Save memory/computation.
- Reduce noise in high-dimensional data.
- Useful in exploratory data analysis.
- Often a pre-processing step for supervised learning.

CLUSTERING: EXAMPLES

Examples:

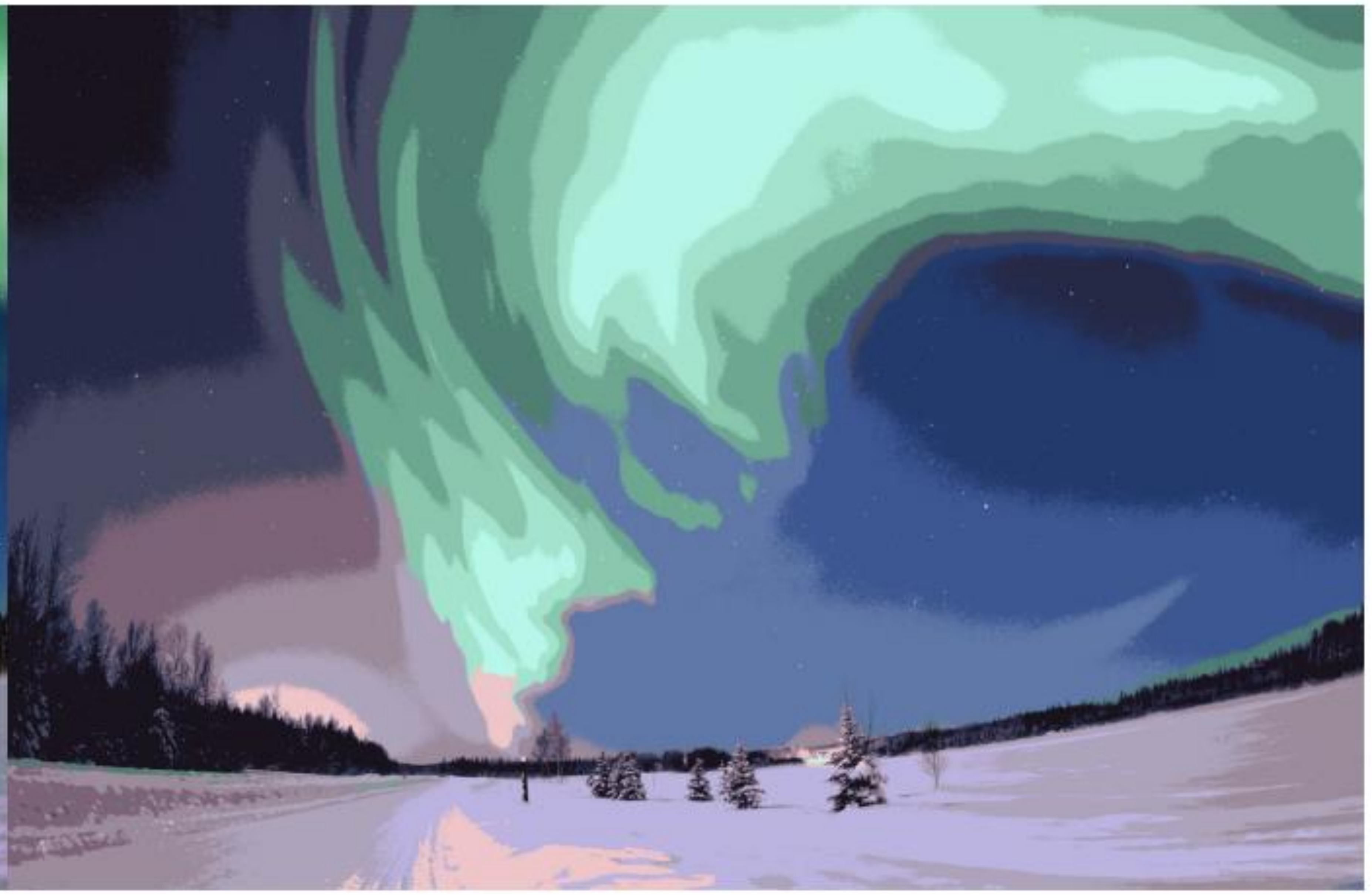
- A highlighted cluster of symptoms may indicate a disease.
- Various video films and music songs can be combined into one cluster of a certain subculture.
- Some services provided to customers of a telecommunications company can be bundled into a tariff (recommender systems).
- Segmentation of customers;
- Clustering outlets to develop marketing strategies.
- Separating normal data from outliers or anomalies.

CLUSTERING: EXAMPLES



- Image segmentation.

CLUSTERING: EXAMPLES

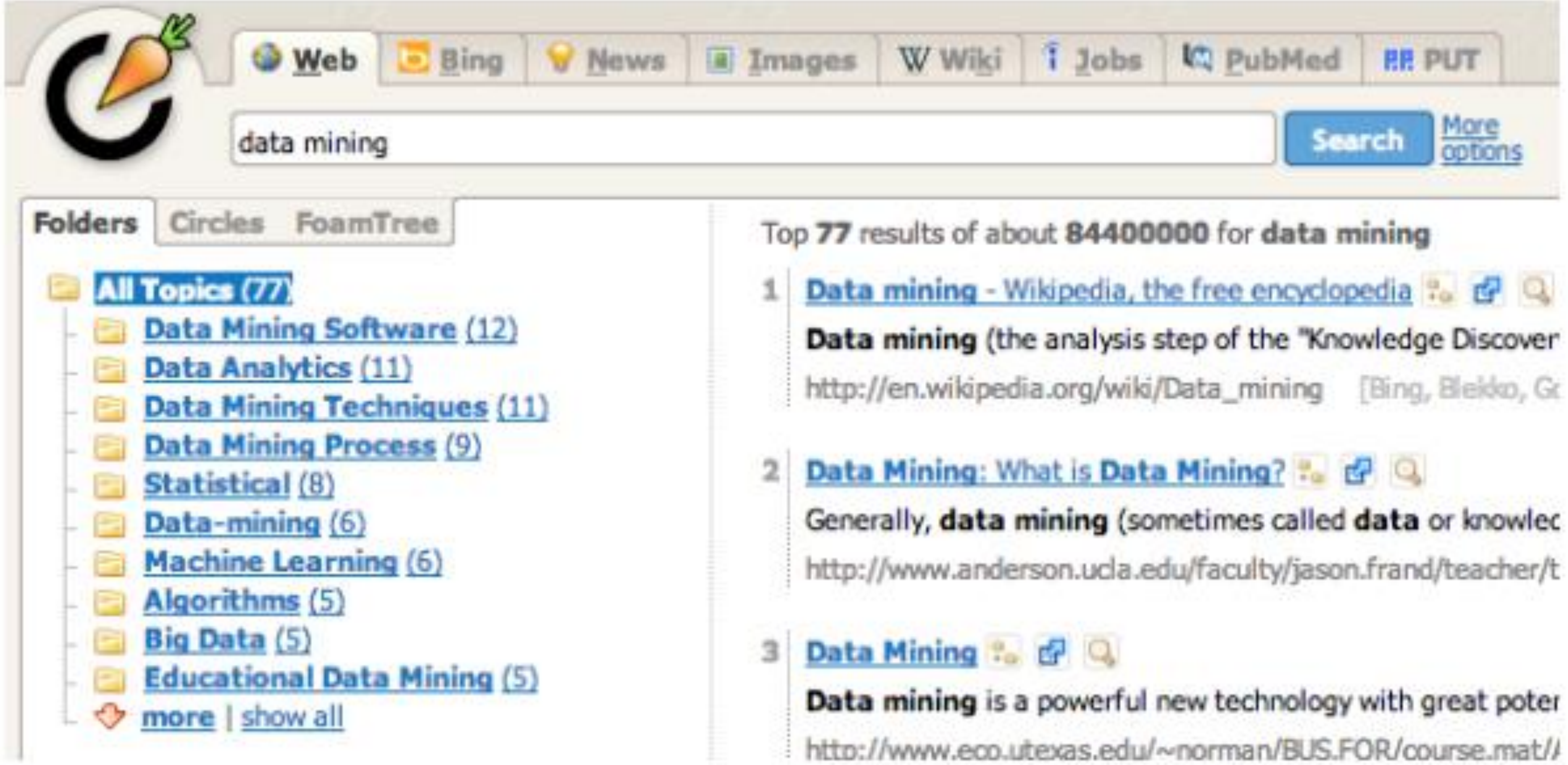


k-means (16 colors)

- Image compression.

CLUSTERING: EXAMPLES

- Clustering of search results.



data mining

All Topics (77)
- Data Mining Software (12)
- Data Analytics (11)
- Data Mining Techniques (11)
- Data Mining Process (9)
- Statistical (8)
- Data-mining (6)
- Machine Learning (6)
- Algorithms (5)
- Big Data (5)
- Educational Data Mining (5)
- more | show all

Top **77** results of about **84400000** for **data mining**

1 Data mining - Wikipedia, the free encyclopedia

Data mining (the analysis step of the "Knowledge Discover

http://en.wikipedia.org/wiki/Data_mining [Bing, Blekko, Go

2 Data Mining: What is Data Mining?

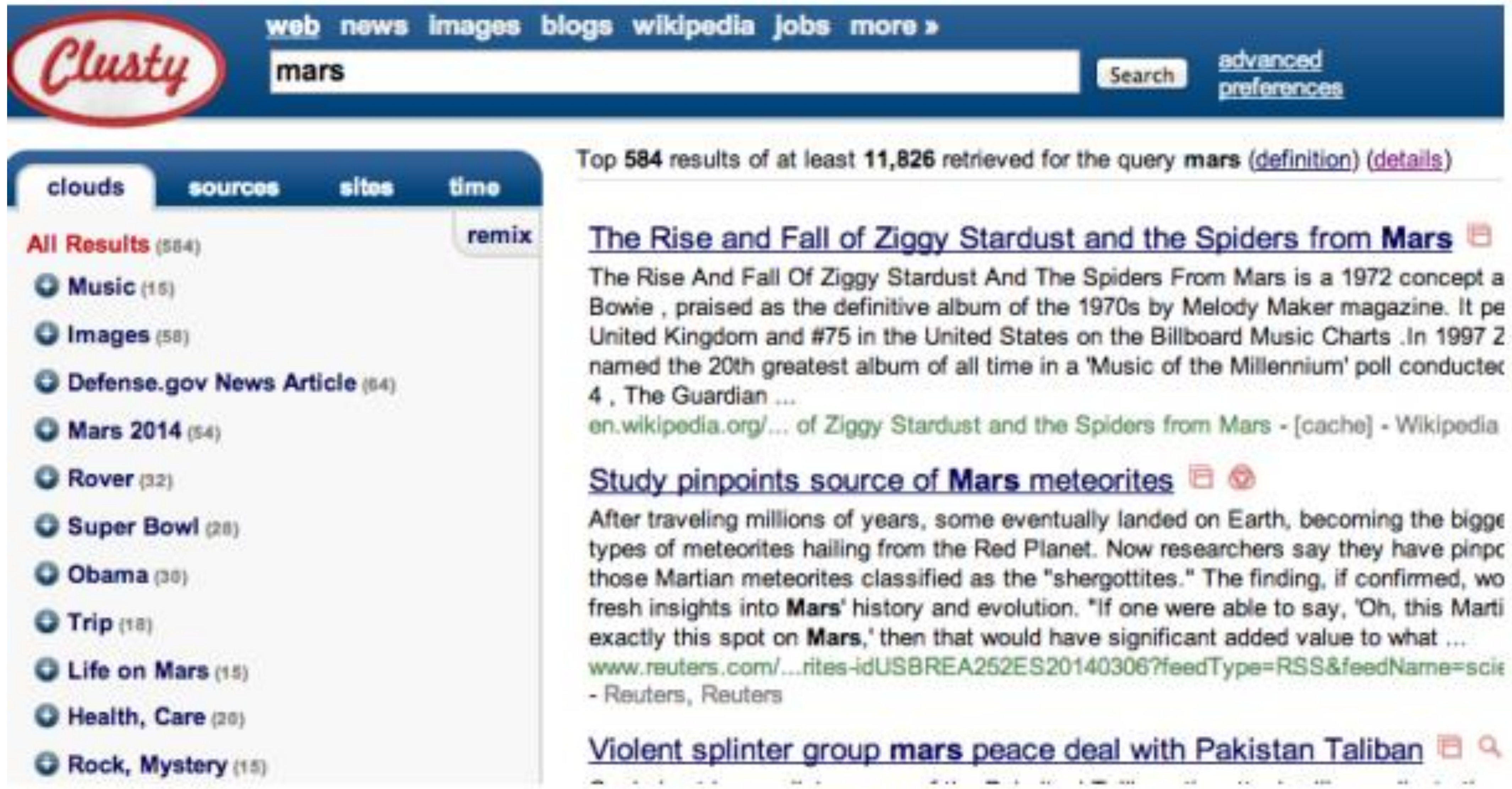
Generally, **data mining** (sometimes called **data** or knowlec

http://www.anderson.ucla.edu/faculty/jason.frand/teacher/t

3 Data Mining

Data mining is a powerful new technology with great poter

http://www.eco.utexas.edu/~norman/BUS.FOR/course.mat/A



mars

All Results (584)
- Music (15)
- Images (58)
- Defense.gov News Article (64)
- Mars 2014 (54)
- Rover (32)
- Super Bowl (28)
- Obama (30)
- Trip (18)
- Life on Mars (15)
- Health, Care (20)
- Rock, Mystery (15)

Top **584** results of at least **11,826** retrieved for the query **mars** (definition) (details)

The Rise and Fall of Ziggy Stardust and the Spiders from **Mars**

The Rise And Fall Of Ziggy Stardust And The Spiders From Mars is a 1972 concept a Bowie , praised as the definitive album of the 1970s by Melody Maker magazine. It pe United Kingdom and #75 in the United States on the Billboard Music Charts .In 1997 Z named the 20th greatest album of all time in a 'Music of the Millennium' poll conductec 4 , The Guardian ...

en.wikipedia.org/... of Ziggy Stardust and the Spiders from Mars - [cache] - Wikipedia

Study pinpoints source of **Mars** meteorites

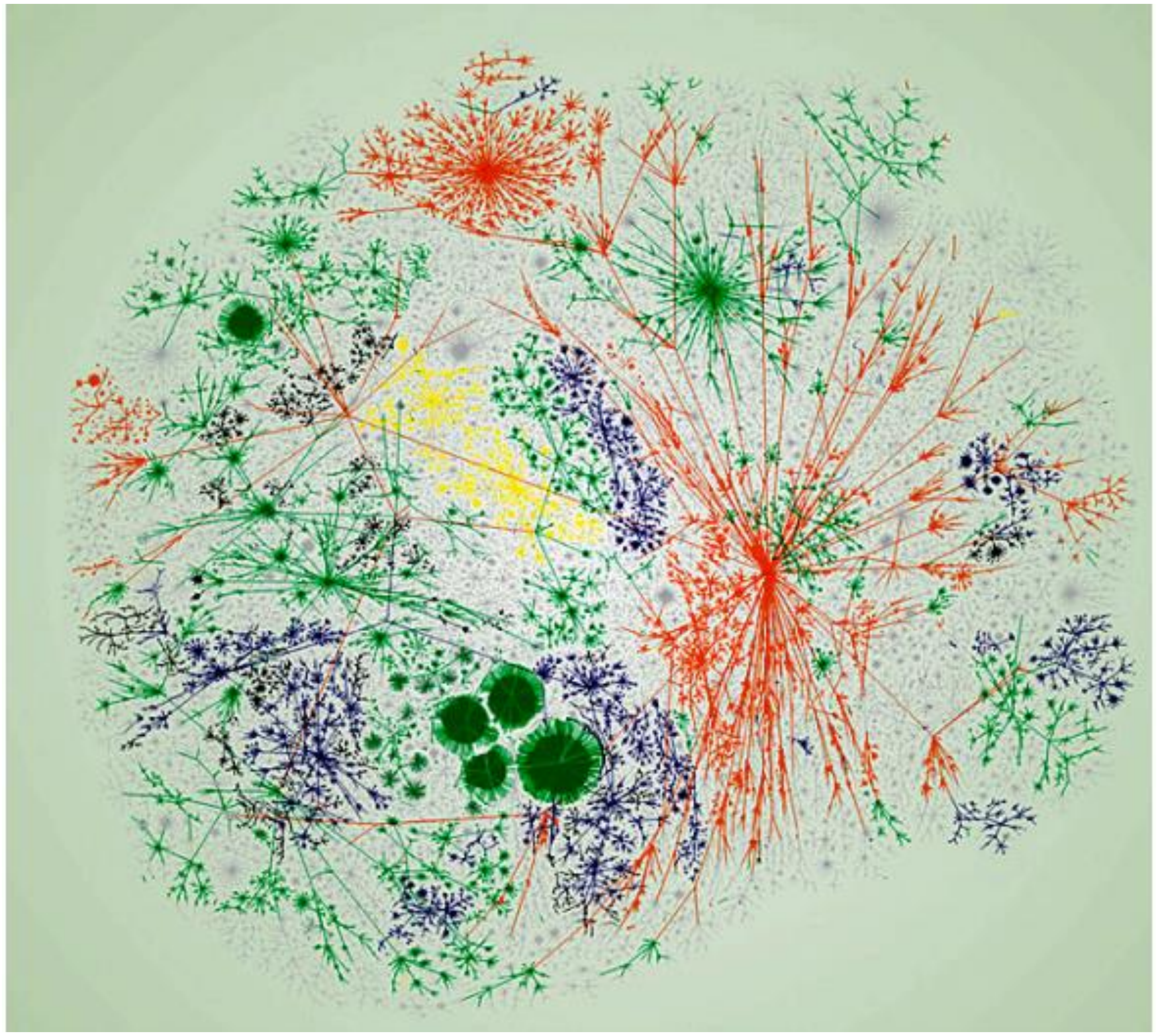
After traveling millions of years, some eventually landed on Earth, becoming the bigge types of meteorites hailing from the Red Planet. Now researchers say they have pinpc those Martian meteorites classified as the "shergottites." The finding, if confirmed, wo fresh insights into **Mars**' history and evolution. "If one were able to say, 'Oh, this Marti exactly this spot on **Mars**,' then that would have significant added value to what ...

www.reuters.com/...rites-idUSBREA252ES20140306?feedType=RSS&feedName=scie

- Reuters, Reuters

Violent splinter group **mars** peace deal with Pakistan Taliban

CLUSTERING: EXAMPLES



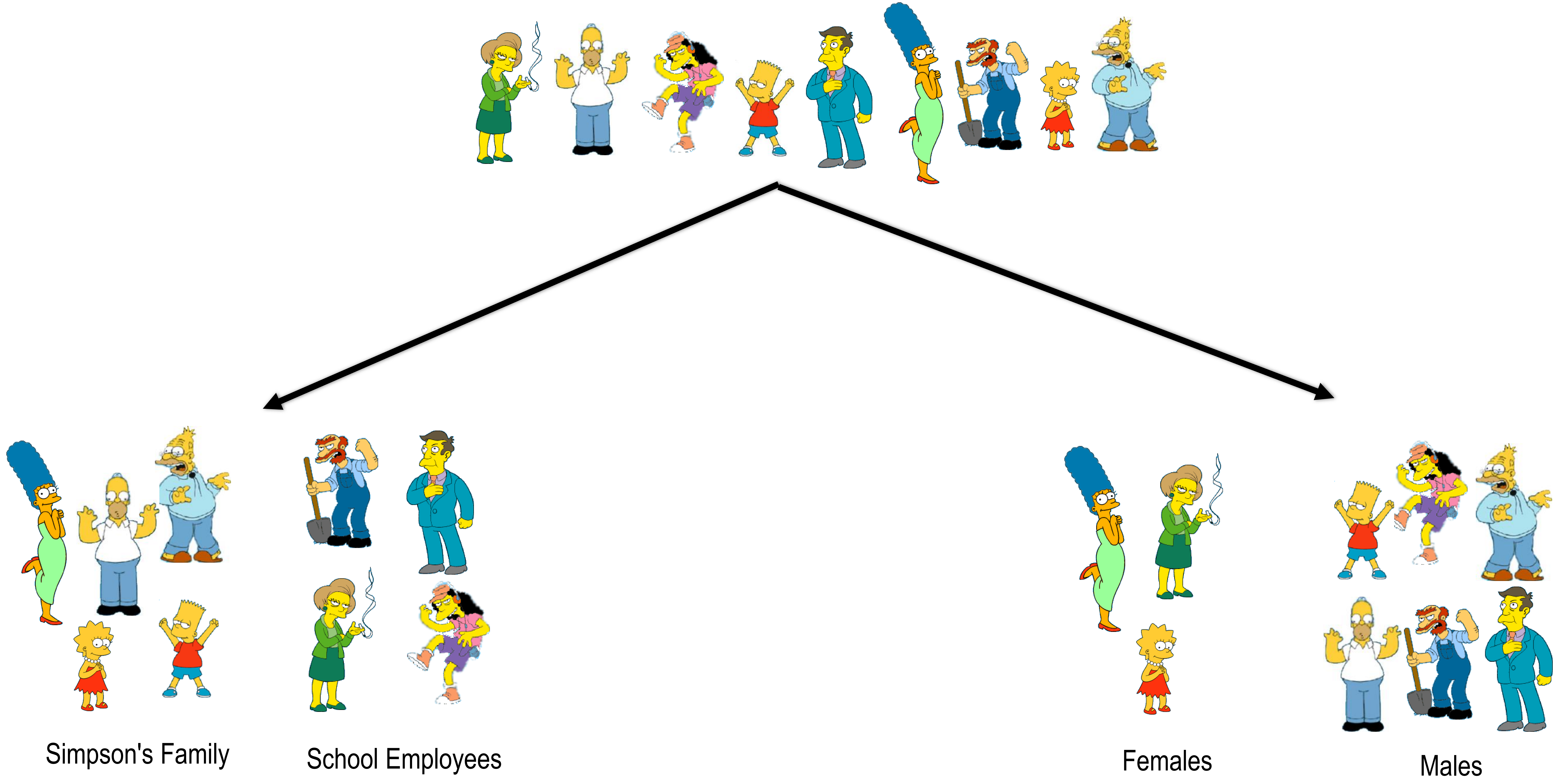
- Clustering graphs

CLUSTERING

Cluster analysis methods allow solving the following tasks:

- Classification of objects, taking into account signs that reflect the essence, nature of objects. The solution of such a problem, as a rule, leads to a deepening of knowledge about the set of classified objects;
- Verification of the put forward assumptions about the presence of some structure in the studied set of objects, i.e. search for an existing structure;
- Construction of new classifications for poorly studied phenomena, when it is necessary to establish the presence of connections within a set and try to bring structure into it.

CLUSTERING IS SUBJECTIVE



Simpson's Family

School Employees

Females

Males

COMPONENTS OF CLUSTER ANALYSIS

Components:

1. A dissimilarity function between samples;
2. A loss function to evaluate clusters;
3. Algorithm that optimizes this loss function.

DISSIMILARITY FUNCTION

- How to define a dissimilarity function $D(x_i, x_j)$?

- Choice of dissimilarity function is application dependent.

- Need to consider the type of features (categorical, ordinal or quantitative).

DISSIMILARITY FUNCTION

Suppose each data point x_i has features $x_{ij}, j = 1, \dots, p$.

One choice of dissimilarity function is the **Euclidean distance**:

$$D(x_i, x_k) = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{kj}\right)^2}.$$

- Resulting clusters **are not invariant to scaling**. Thus, if the features have different scales, standardize the data.

STANDARTIZATION VS NORMALIZATION

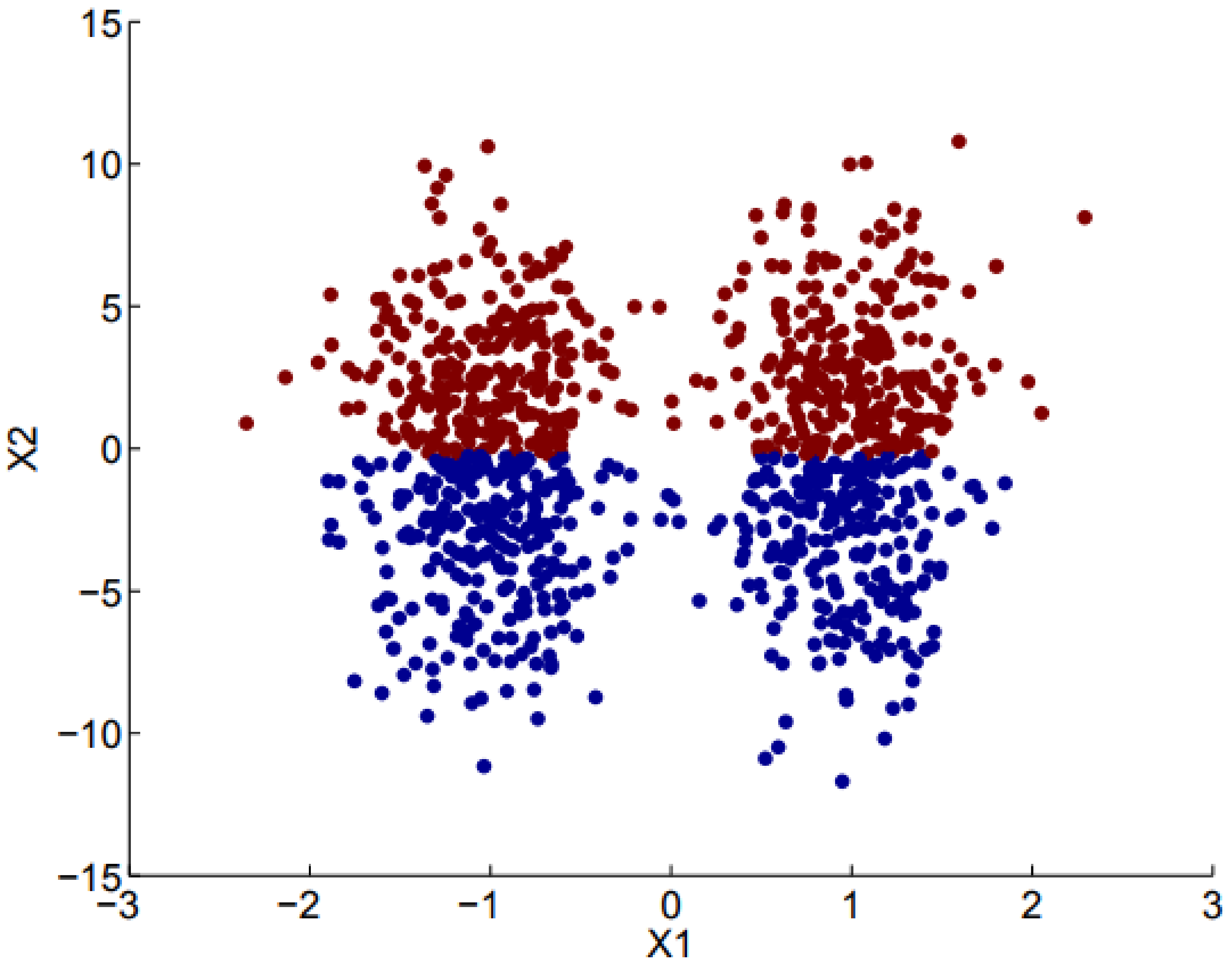
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

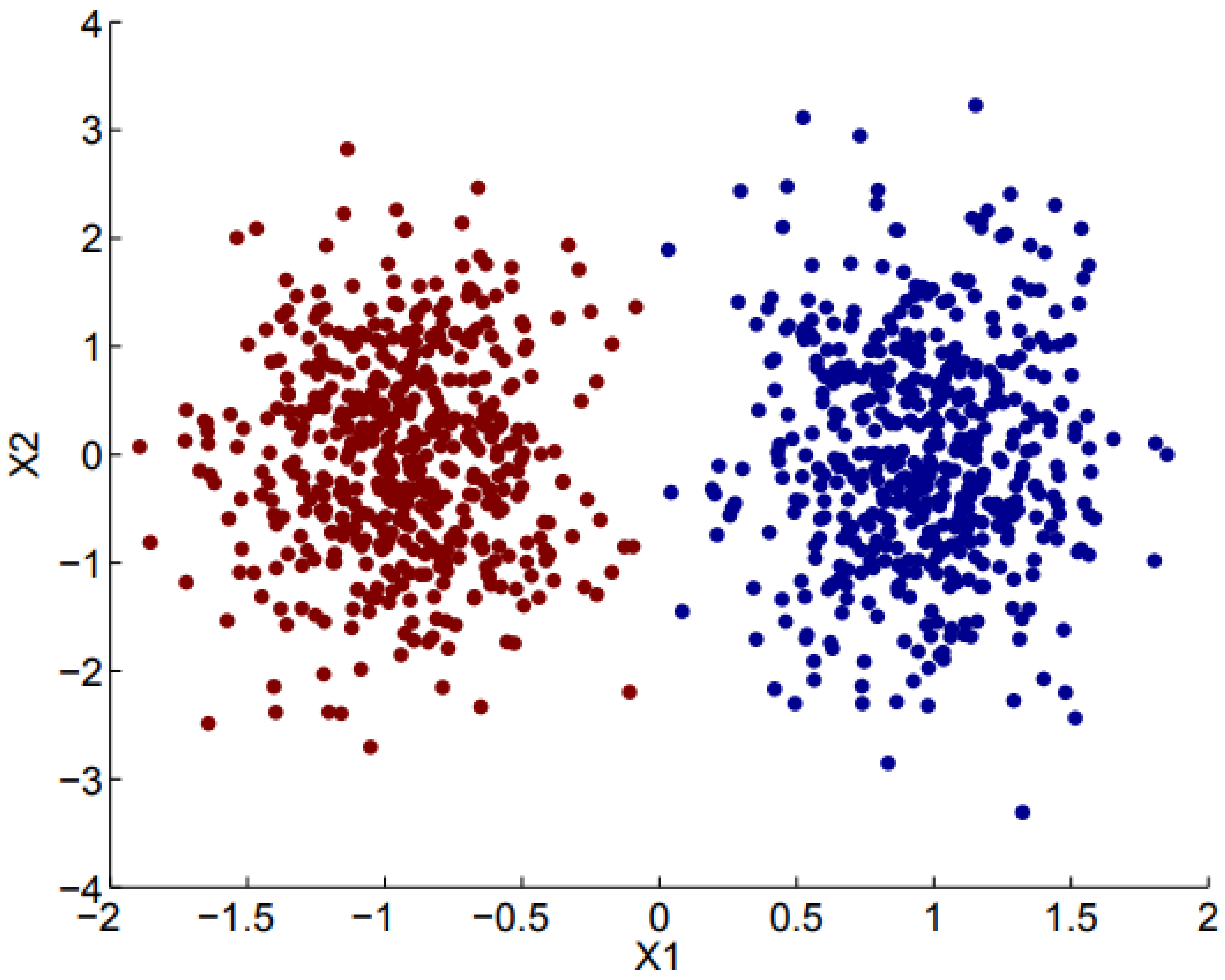
Standardization is another scaling technique where the values are centered around the mean $\bar{X}$ with a unit standard deviation σ_x. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X' = \frac{X - \bar{X}}{\sigma_x}$$

STANDARTIZATION

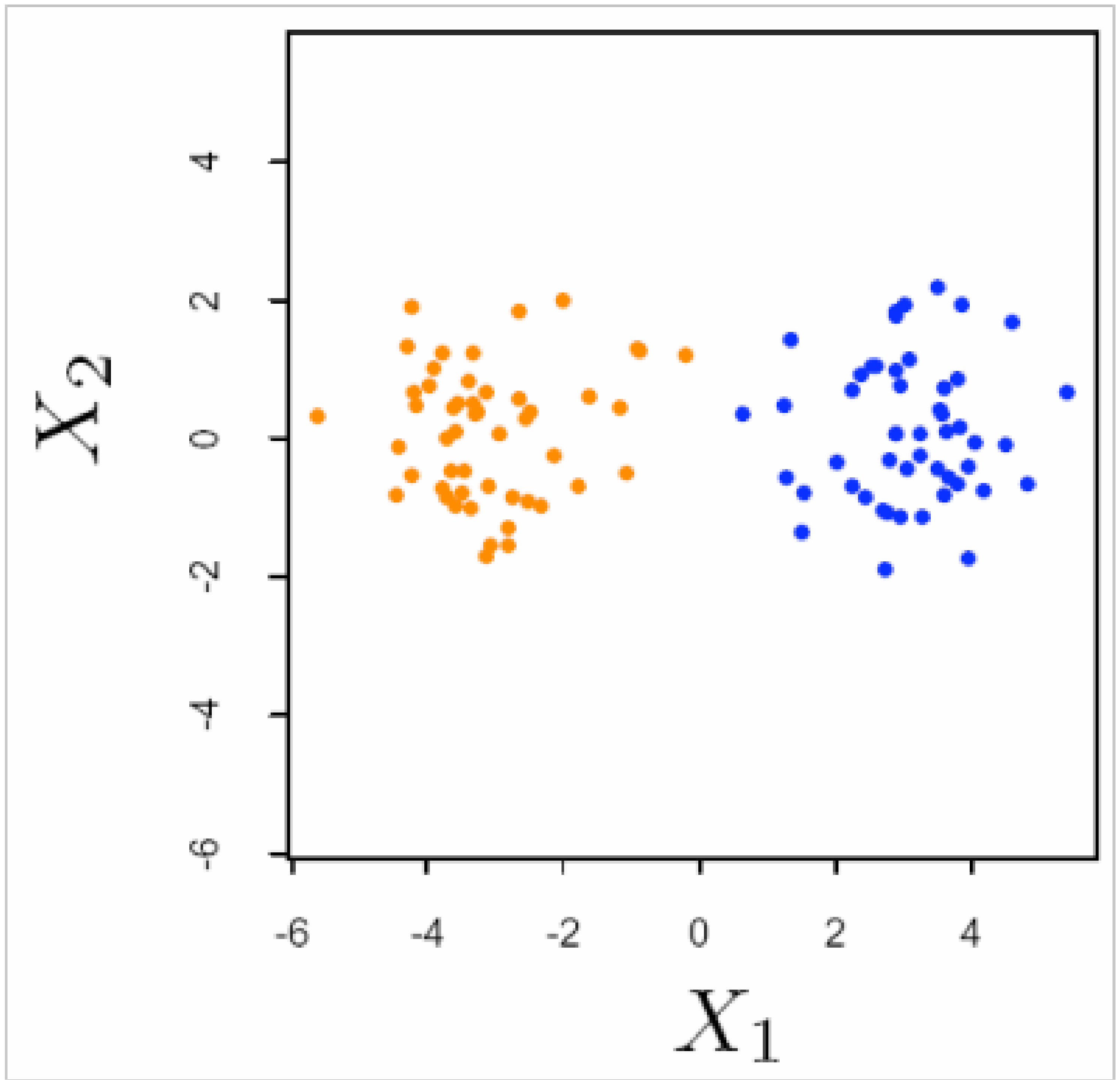


Without standardization

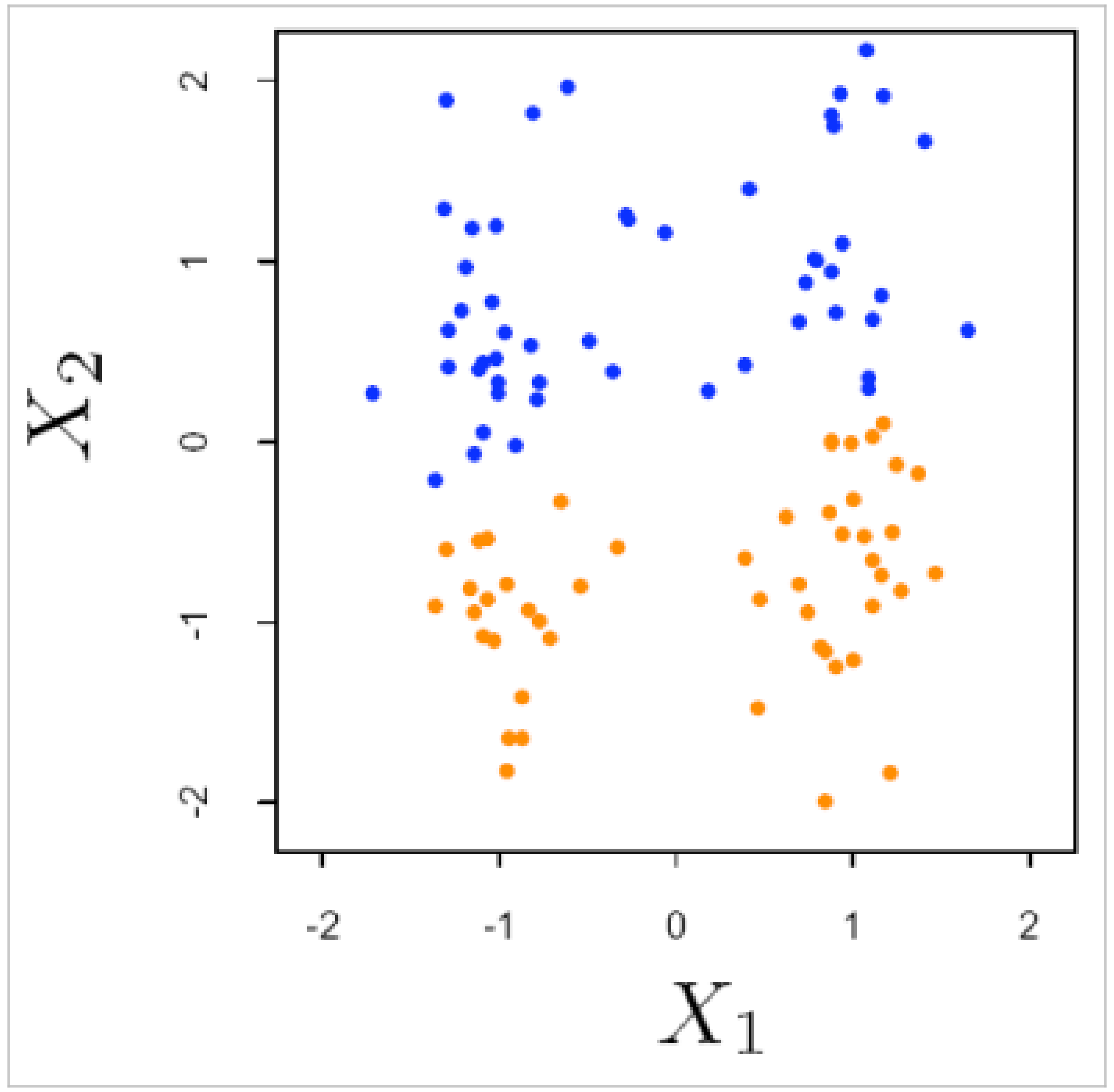


With standardization

STANDARTIZATION



Without standardization



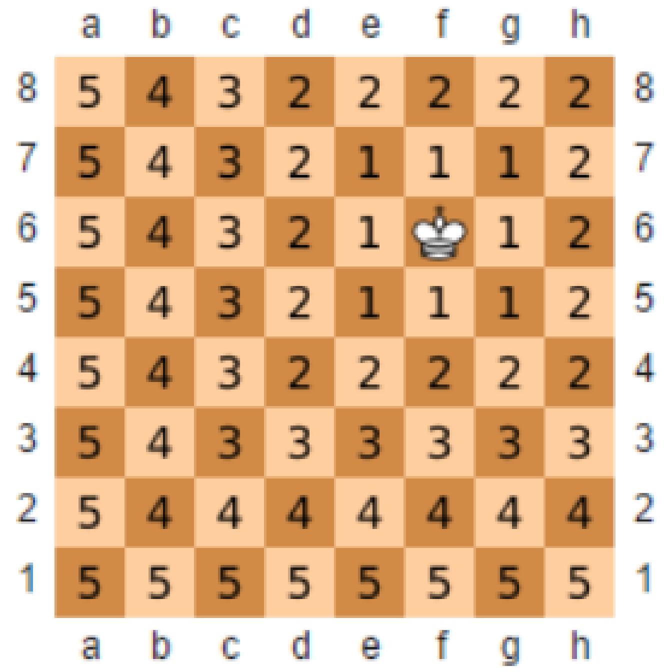
With standardization

- Standardization is not always helpful.

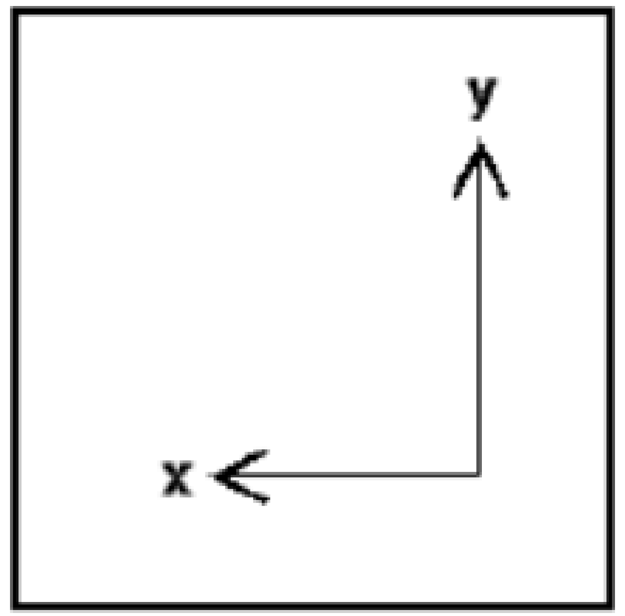
OTHER DISSIMILARITY FUNCTIONS

Chebyshev distance: $D(x_i, x_k) = \max_j(|x_{ij} - x_{kj}|)$.

Manhattan distance: $D(x_i, x_k) = \sum_{j=1}^{p}(|x_{ij} - x_{kj}|)$



	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1	♔	1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	



Manhattan

Correlation coefficient: $D(x_i, x_k) = \frac{\sum(x_i - \mu_{x_i})(x_k - \mu_{x_k})}{\sigma_{x_i}\sigma_{x_k}}$

Distance based on Pearson correlation: $D(x_i, x_k) = \frac{1 - corr(x_i, x_k)}{2}$

etc.

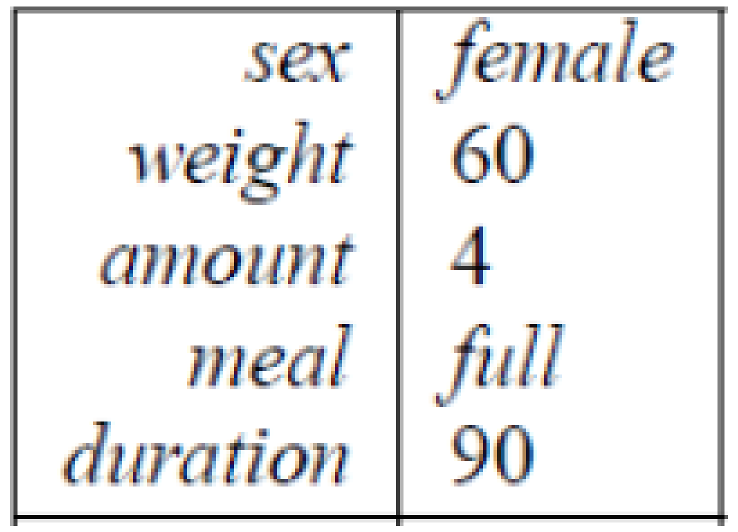
DISSIMILARITY FUNCTION: PROPERTIES

- **Symmetry:** $D(x_i, x_k)$=$D(x_k, x_i)$;
- **Constancy of self-similarity:** $D(x_i, x_i) = 0$;
- **Positivity (Separation):** $D(x_i, x_k) = 0 \; iif \; i = k$;
- **Triangular Inequality:** $D(x_i, x_k) \leq D(x_i, x_l) + D(x_l, x_k)$

DISSIMILARITY FUNCTIONS

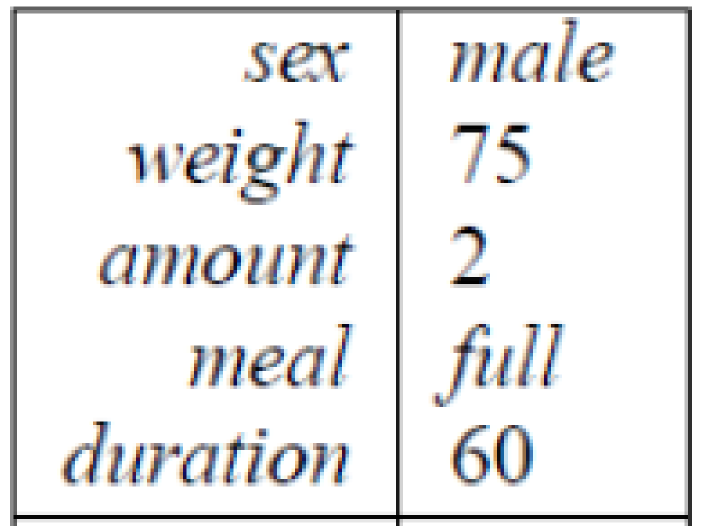
What is the distance between these objects?

x_1



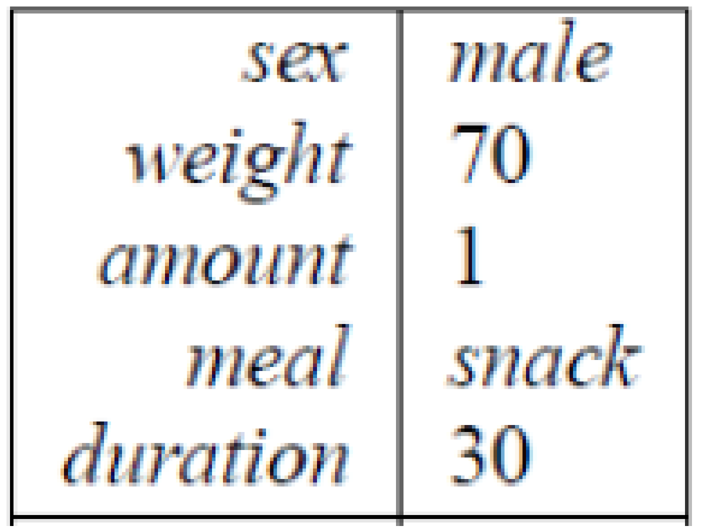
sex	*female*
weight	60
amount	4
meal	*full*
duration	90

x_2



sex	*male*
weight	75
amount	2
meal	*full*
duration	60

q



sex	*male*
weight	70
amount	1
meal	*snack*
duration	30

$$D(x_{ij}, x_{kj}) = \begin{cases} \dfrac{|x_{ij} - x_{kj}|}{\max\limits_i x_{ij} - \min\limits_i x_{ij}} \; if\ j\ is\ numeric, \\ 0, if\ j\ is\ symbolic\ and\ x_{ij} = x_{kj} \\ 1, otherwise \end{cases}$$

CLUSTERING ALGORITHMS

1. K-means;
2. Mean-shift clustering;
3. DBSCAN;
4. EM using GMM;
5. Hierarchical clustering;
6. Spectral clustering.

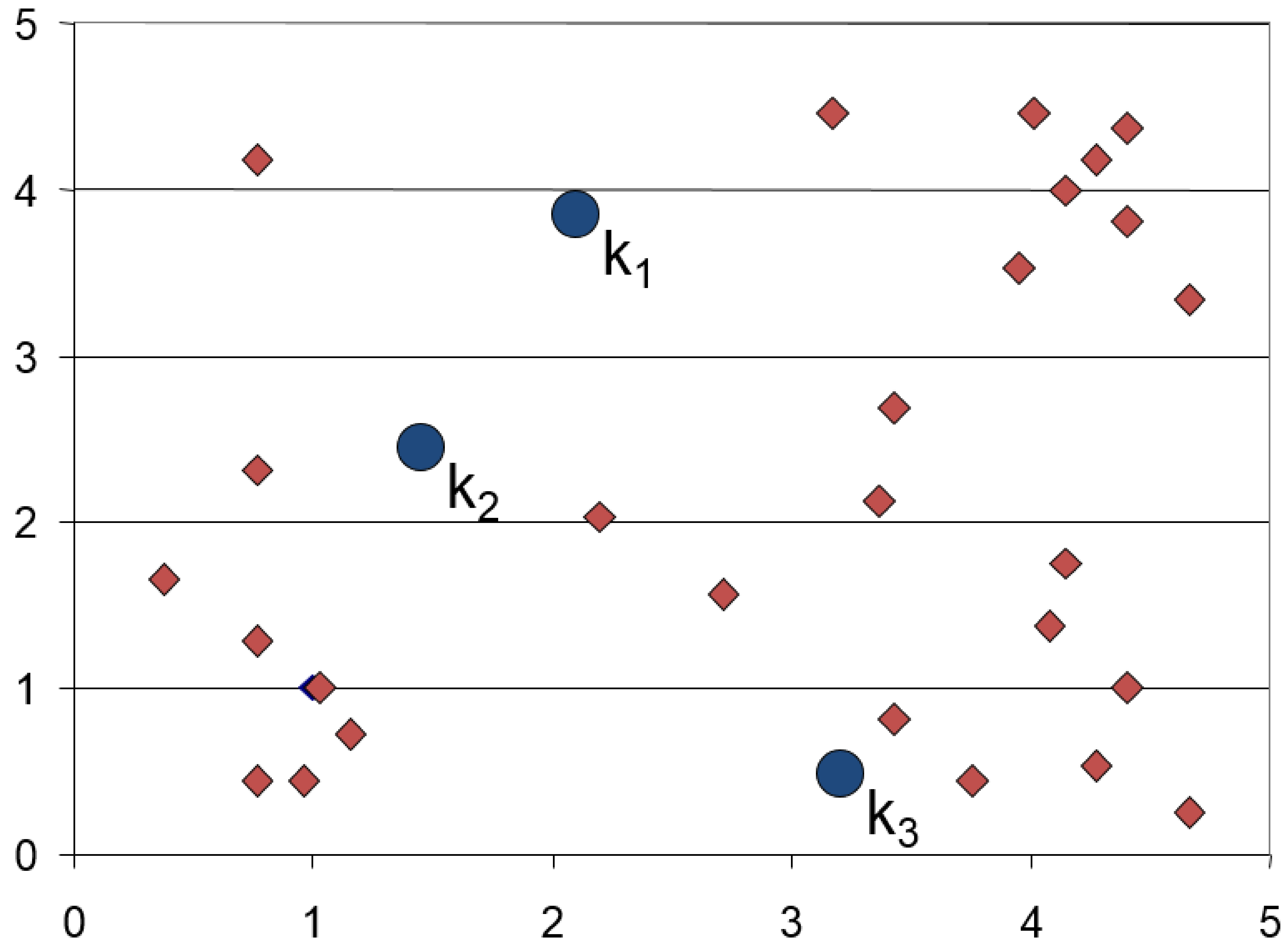
1. K-MEANS

- One of the most common clustering algorithms.
- Provides exactly k clusters.

Main steps:

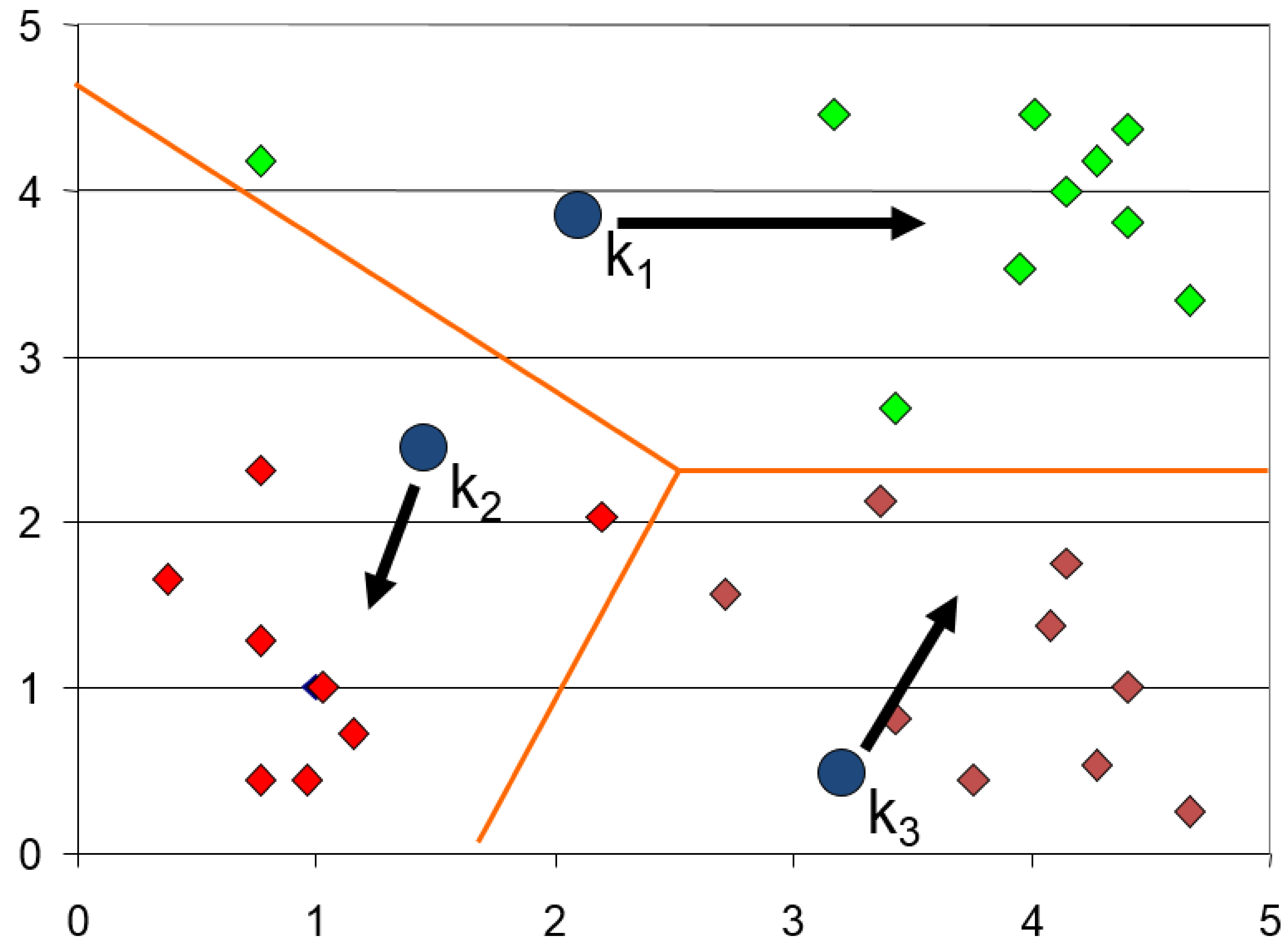
1. Initialize *k* center points (group centers).
2. Classify each data point by calculating the distance between the particular points and each group center. The point belongs to the group whose center is the nearest to it.
3. Recalculate the group center (mean of all the vectors in the particular.
4. Repeat the procedure until you ensure that the group centers do not vary much between iterations.

1. K-MEANS



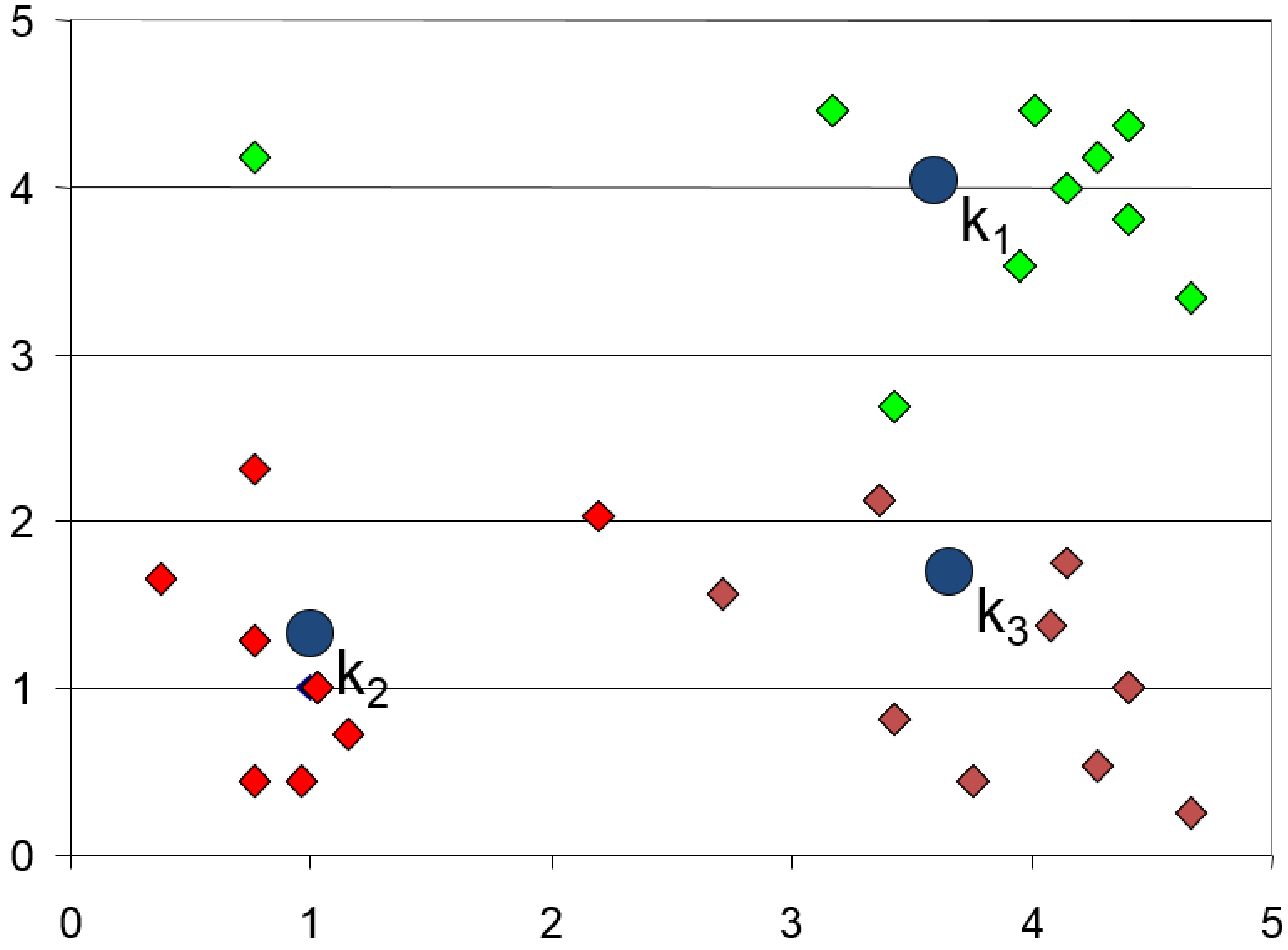
Step 1

1. K-MEANS



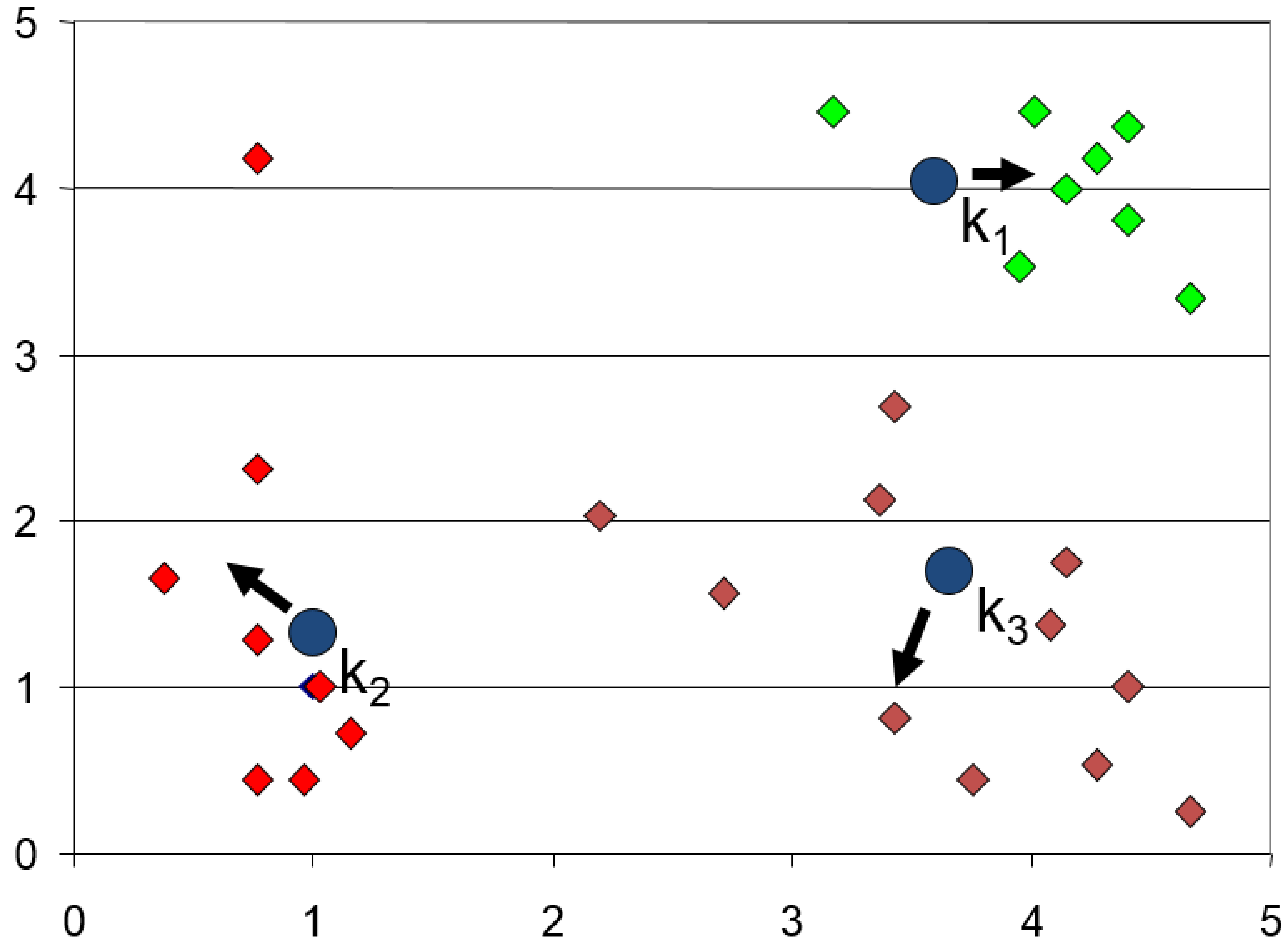
Step 2

1. K-MEANS



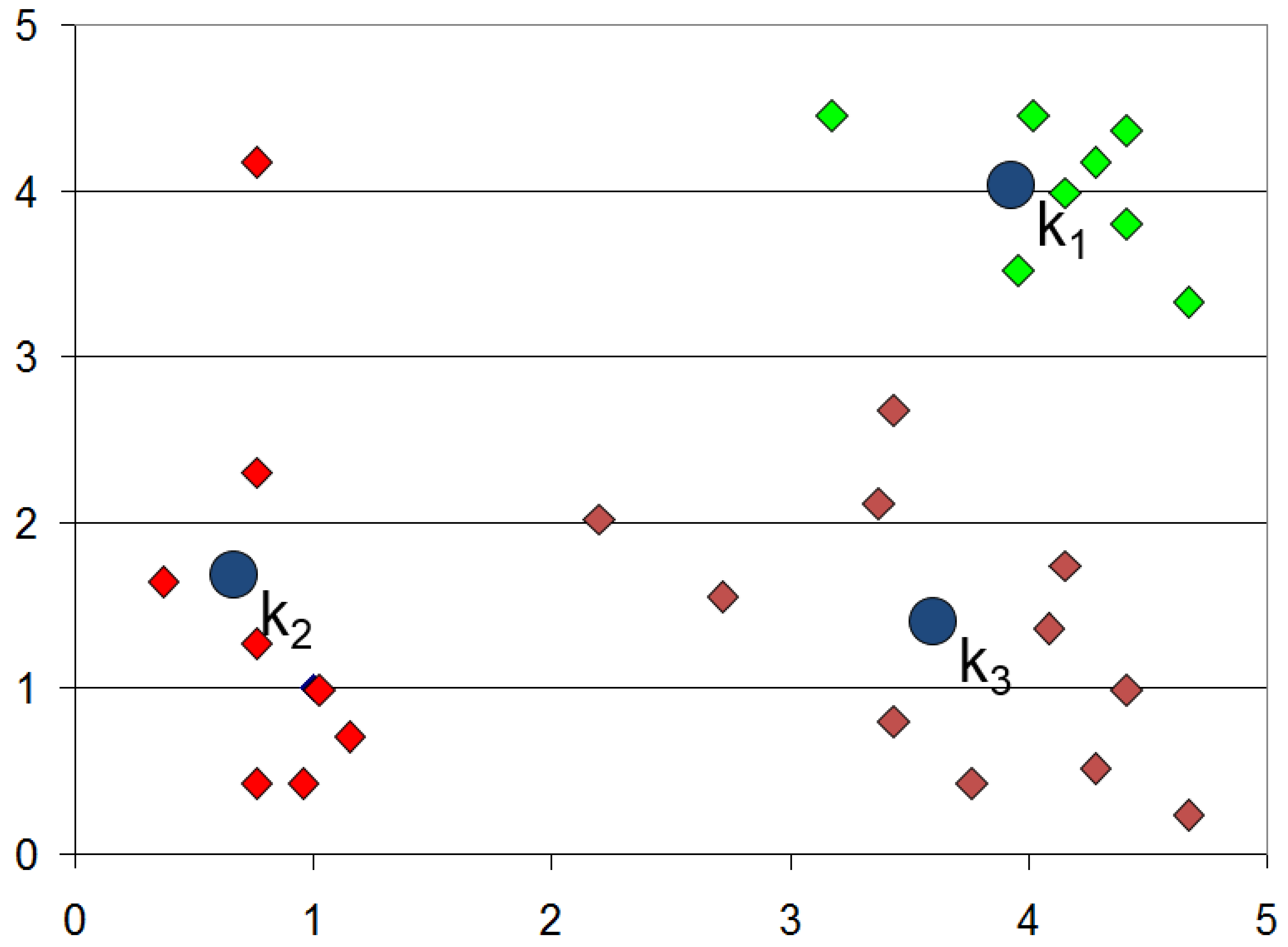
Step 3

1. K-MEANS



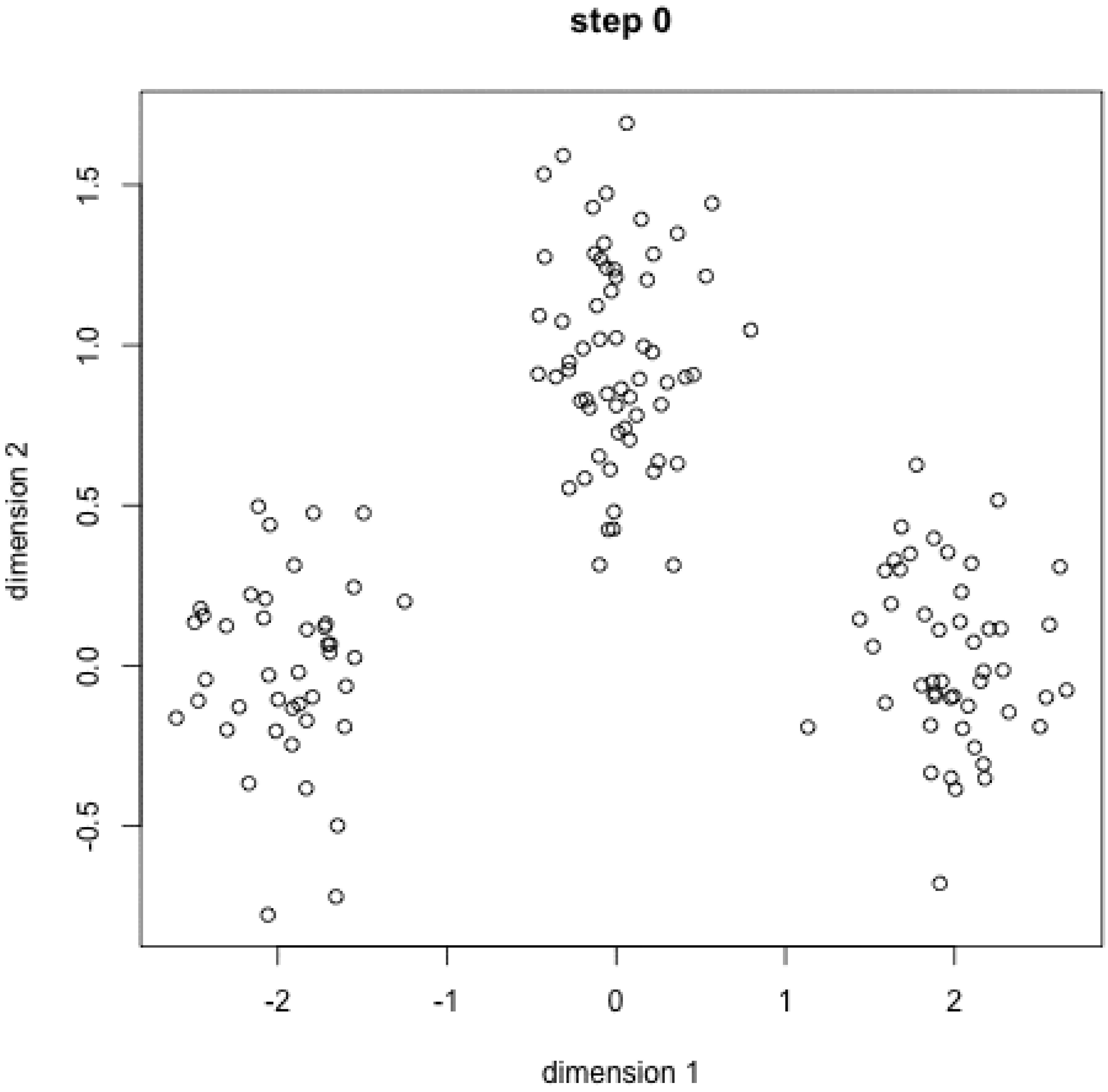
Step 4

1. K-MEANS



Step 5

1. K-MEANS: VISUALIZATION



1. K-MEANS

Pros:

- Pretty fast: a linear complexity O(n).

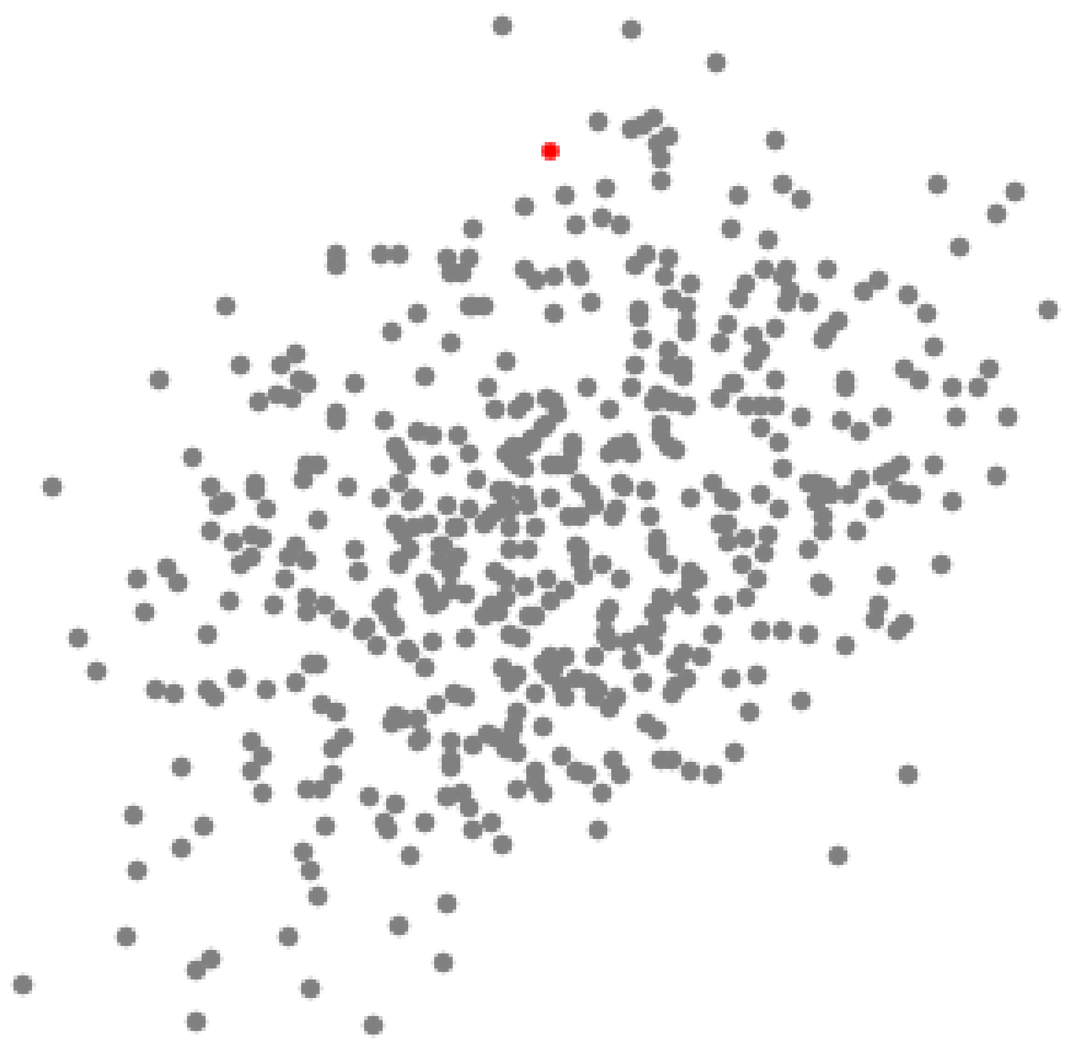
Cons:

- Requires to select how many groups/classes there are.
- Since the initial group centers are random, it may yield different clustering results on different runs of the algorithm.
- Hard assignments of data points (unlike GMM).
- Sensitivity to outliers.

Counterparts of k-means: k-median (less sensitive to outliers), k-medoids (non-Euclidean distance loss function).

2. MEAN-SHIFT CLUSTERING

- Mean-shift is a sliding window type algorithm;
- A centroid-based algorithm meaning that the goal is to locate the center points of each group/class, which works by updating candidates for center points to be the mean of the points within the sliding-window.



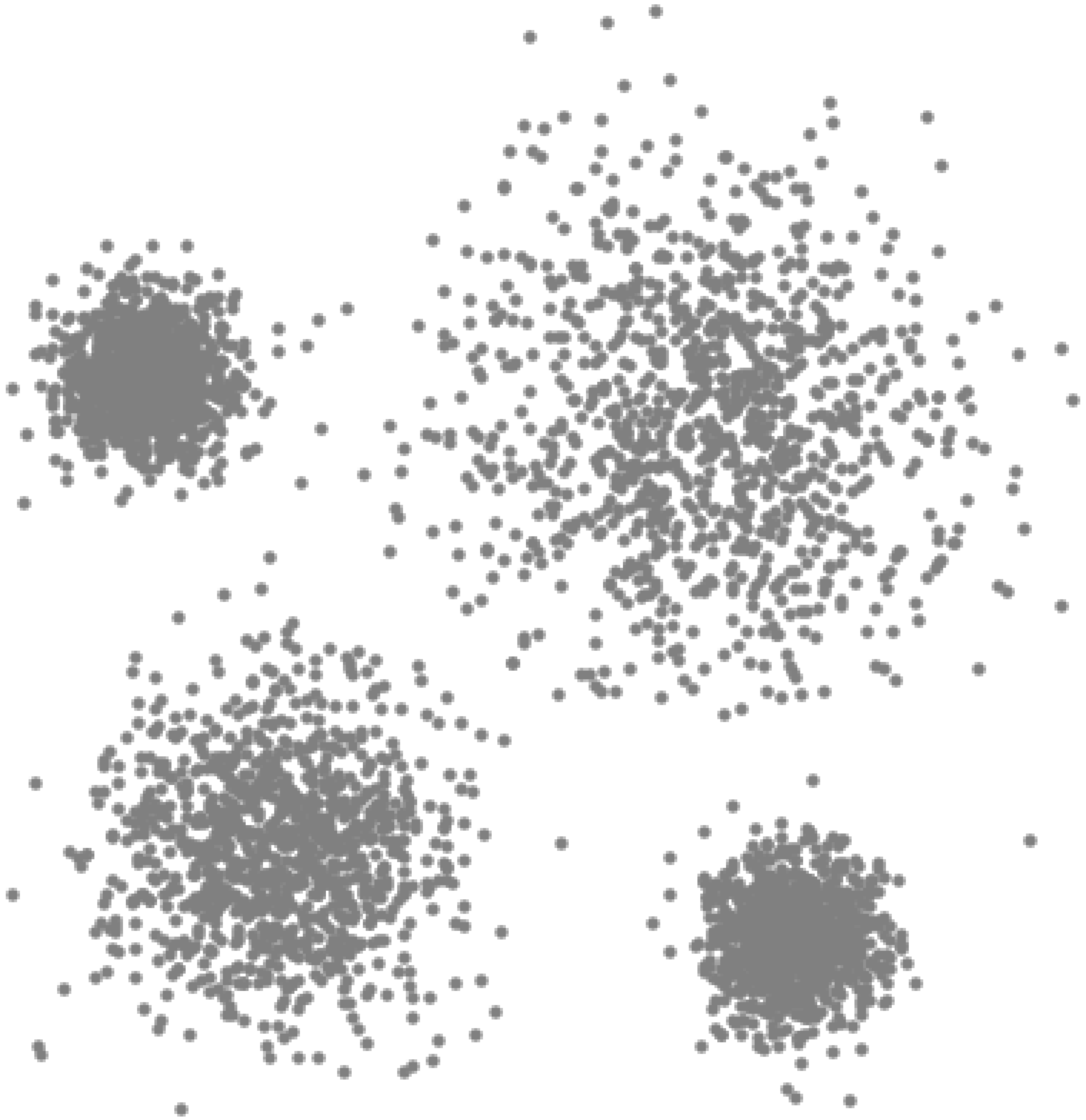
Candidates for center points
(single sliding window)

2. MEAN-SHIFT CLUSTERING

Main steps (two-dimensional space):

1. Begin with a circular sliding window having its center at a randomly selected point *C* with radius *r* as the kernel.
2. At every iteration, the window shifts towards the denser regions by changing the center point to the mean of the points within the window.
3. Continue shifting the window according to the mean until you reach the point where you accommodate the maximum number of points within it.
4. Repeat this process with multiple sliding windows until you come to a situation wherein all the points will lie within a window. In the case of overlapping of windows, the window having a higher number of points will prevail.

2. MEAN-SHIFT CLUSTERING



2. MEAN-SHIFT CLUSTERING

Pros:

- Unlike the K-means clustering algorithm, you need not select the number of clusters.
- The cluster centers converging towards the point of maximum density is a desirable aspect as it fits well in the data-driven sense.
- Less vulnerable to outfits.

Cons:

- The selection of the window size or the radius t is a non-trivial issue.

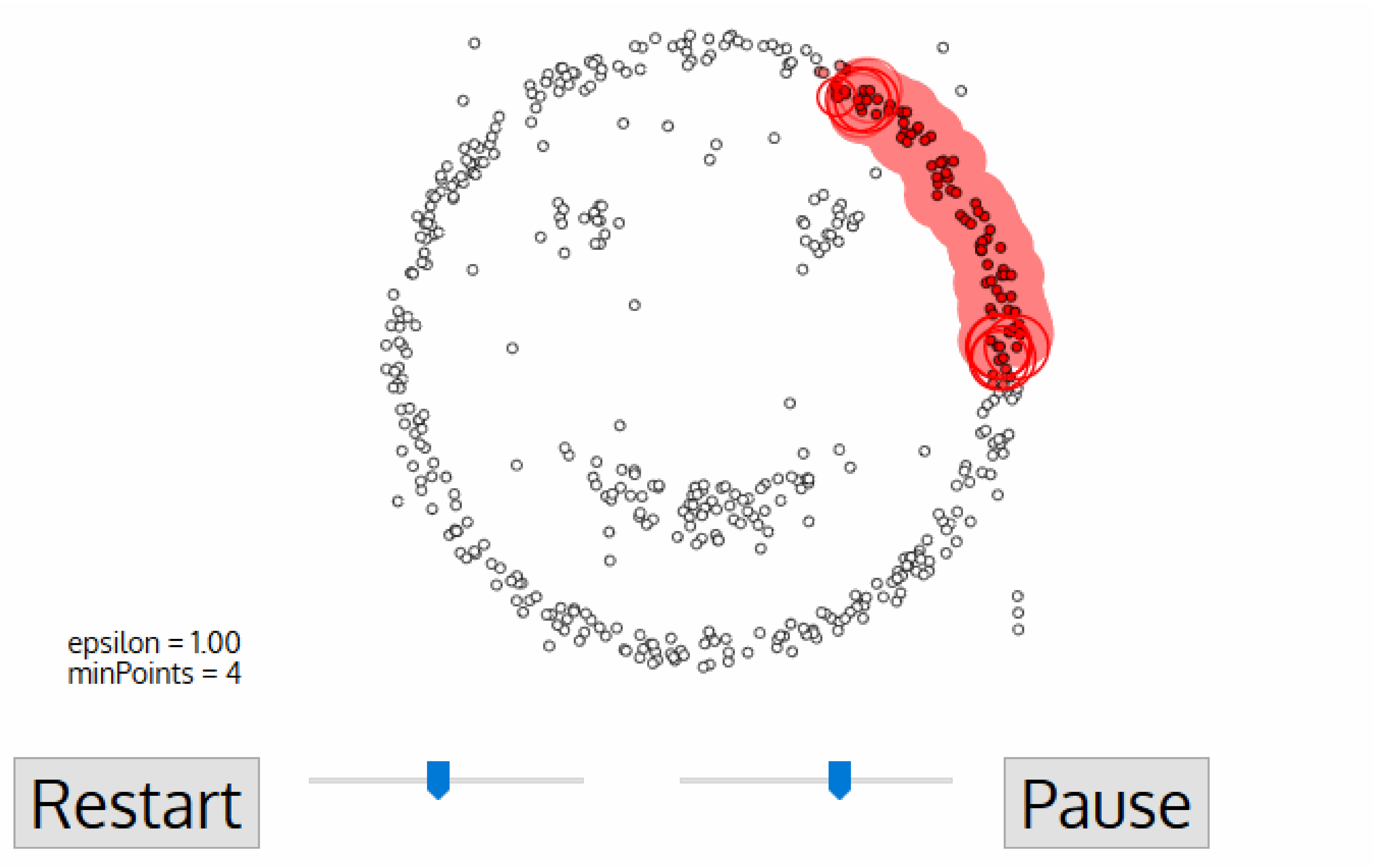
3. DBSCAN

- DBSCAN – Density-Based Spatial Clustering of Applications with Noise

Main steps:

1. Start with a random unvisited data point. All points within a distance ε are the *neighborhood points*.
2. If # of *neighborhood points* > *minPoints,* the current data point becomes the first point in the cluster. Otherwise, the point gets labeled as 'Noise.'
3. All points within the distance ε become part of the same cluster.
4. Continue with the process until you visit and label each point within the ε neighborhood of the cluster.
5. Repeat the procedure for all the new points added to the cluster group.
6. On completion of the process, start again with a new unvisited point thereby leading to the discovery of more clusters or noise. At the end of the process, you ensure that you mark each point as either cluster or noise.

3. DBSCAN



epsilon = 1.00
minPoints = 4

3. DBSCAN

Pros:

- It does not require a pre-set number of clusters.
- It identifies outliers as noise, unlike the Mean-Shift method
- It finds arbitrarily shaped and sized clusters quite well.

Cons:

- It is not very effective when you have clusters of varying densities. There is a variance in the setting of the distance threshold ε and the minimum points for identifying the neighborhood when there is a change in the density levels.
- If you have high dimensional data, the determining of the distance threshold ε becomes a challenging task.

4. EM USING GMM

- EM using GMM – Expectation-Maximization (EM) Clustering using Gaussian Mixture Models (GMM).

Assumptions:

- Data points are Gaussian distributed (normal distribution).
- Two parameters to describe the shape of each cluster: the mean μ_c and the standard deviation σ_c. As we have a standard deviation in both X and Y directions, a cluster can take any elliptical shape. Every single cluster has a Gaussian distribution.

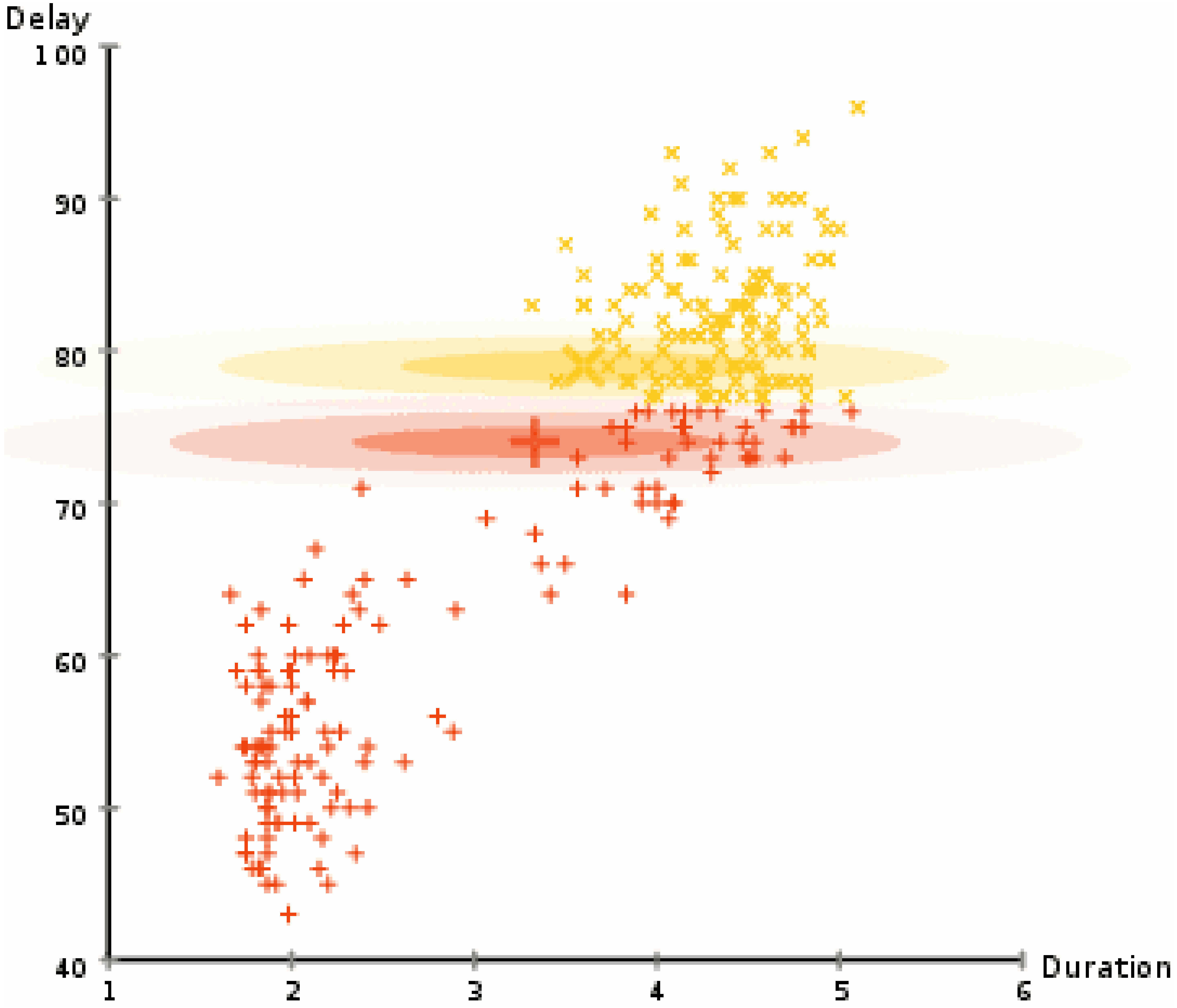
Idea: to find out the parameters of the Gaussian for each cluster

4. EM USING GMM

Main steps:

1. Define the total number of clusters and randomly initialize the Gaussian distribution parameters for each one of them.
2. With this background, calculate the probability of each data point belonging to a particular cluster. The closer the point is to the Gaussian's center, the better are the chances of it belonging to the cluster.
3. Determine a new set of parameters for the Gaussian distributions to maximize the probabilities of data points within the clusters. A weighted sum of data point positions is used to compute these probabilities. The likelihood of the data point belonging to the particular cluster is the weight factor.
4. Repeat the steps 2 and 3 until convergence where there is not much variation.

4. EM USING GMM



4. EM USING GMM

Pros:

- Higher level of flexibility as compared to the K-means clustering.
- Multiple clusters per data point.

Cons:

- Often terminates at a local optimum.
- Requires to select how many groups/classes there are.
- Not suitable to discover clusters with non-convex shapes.

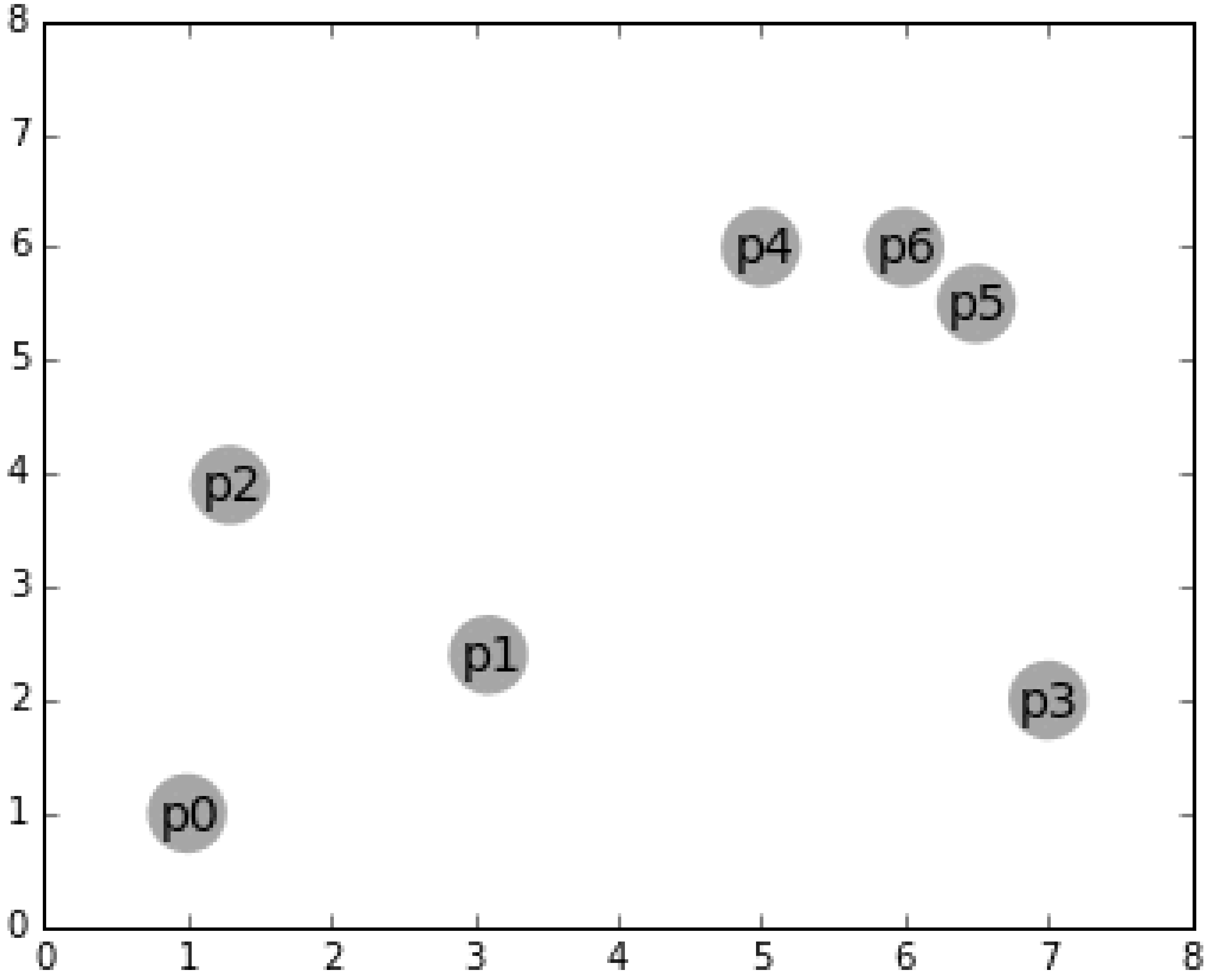
5. AGGLOMERATIVE HIERARCHICAL CLUSTERING

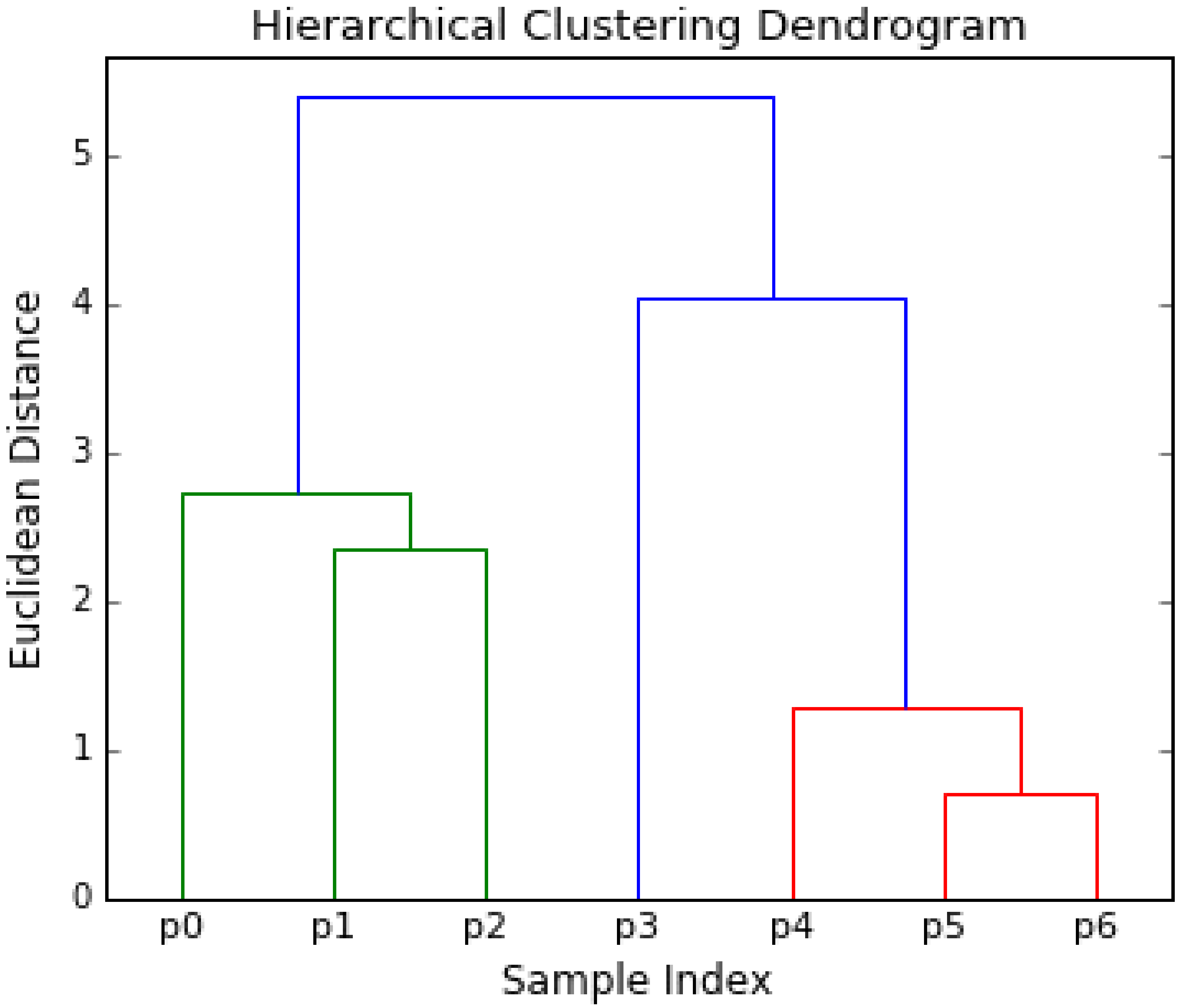
Idea: agglomerative - merge pairs of clusters until you have a single group containing all data points, divisible – sequentially find the best split in 2 parts.

Main steps:

1. Consider each data point as an individual cluster.
2. Select a distance metric to measure the distance between the two groups.
3. At each iteration, merge two clusters with the smallest average linkage into one.
4. Repeat the above step until we have one large cluster containing all the data points.
5. Based on the dendrogram choose an appropriate number of clusters.

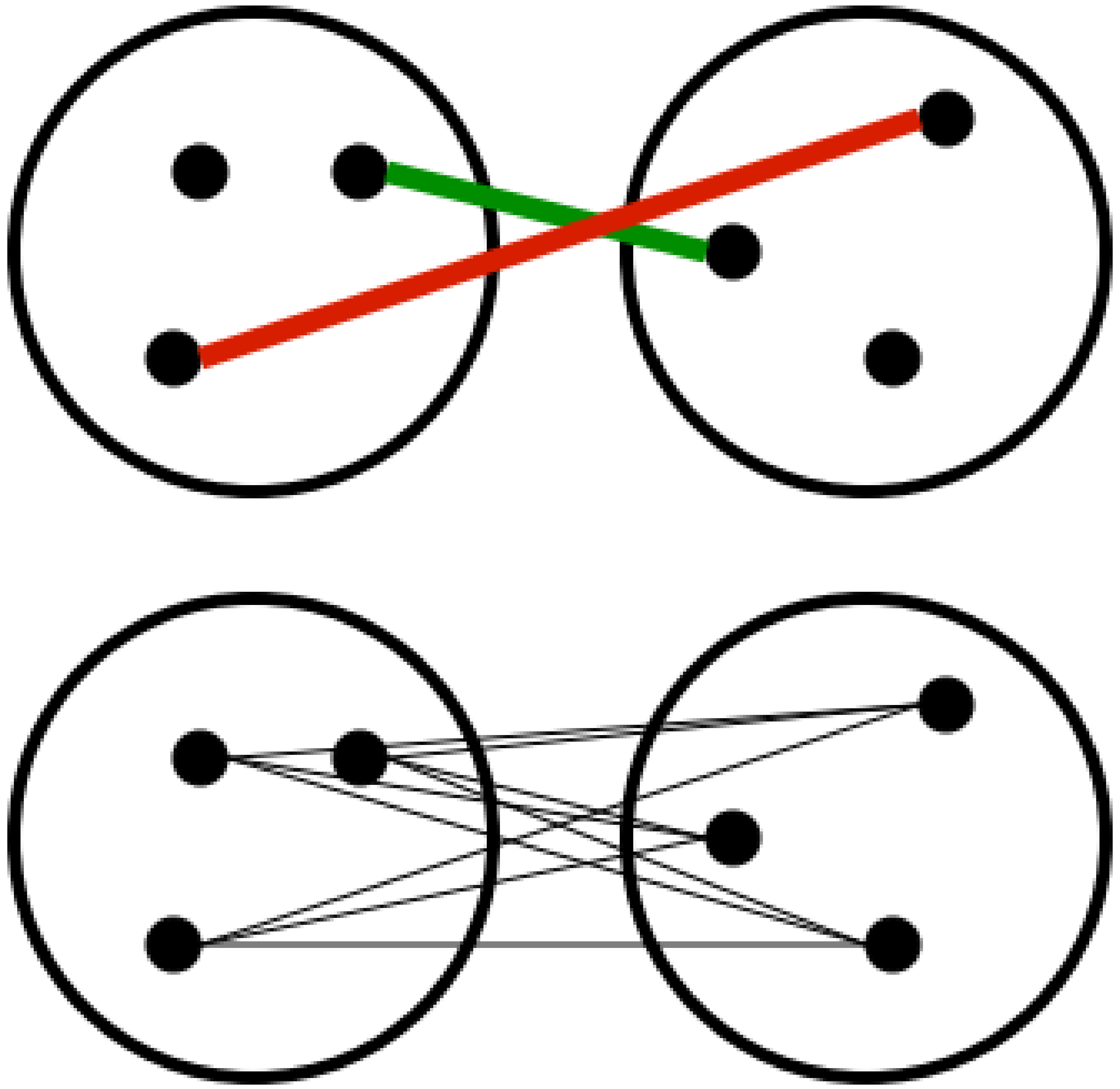
5. AGGLOMERATIVE HIERARCHICAL CLUSTERING





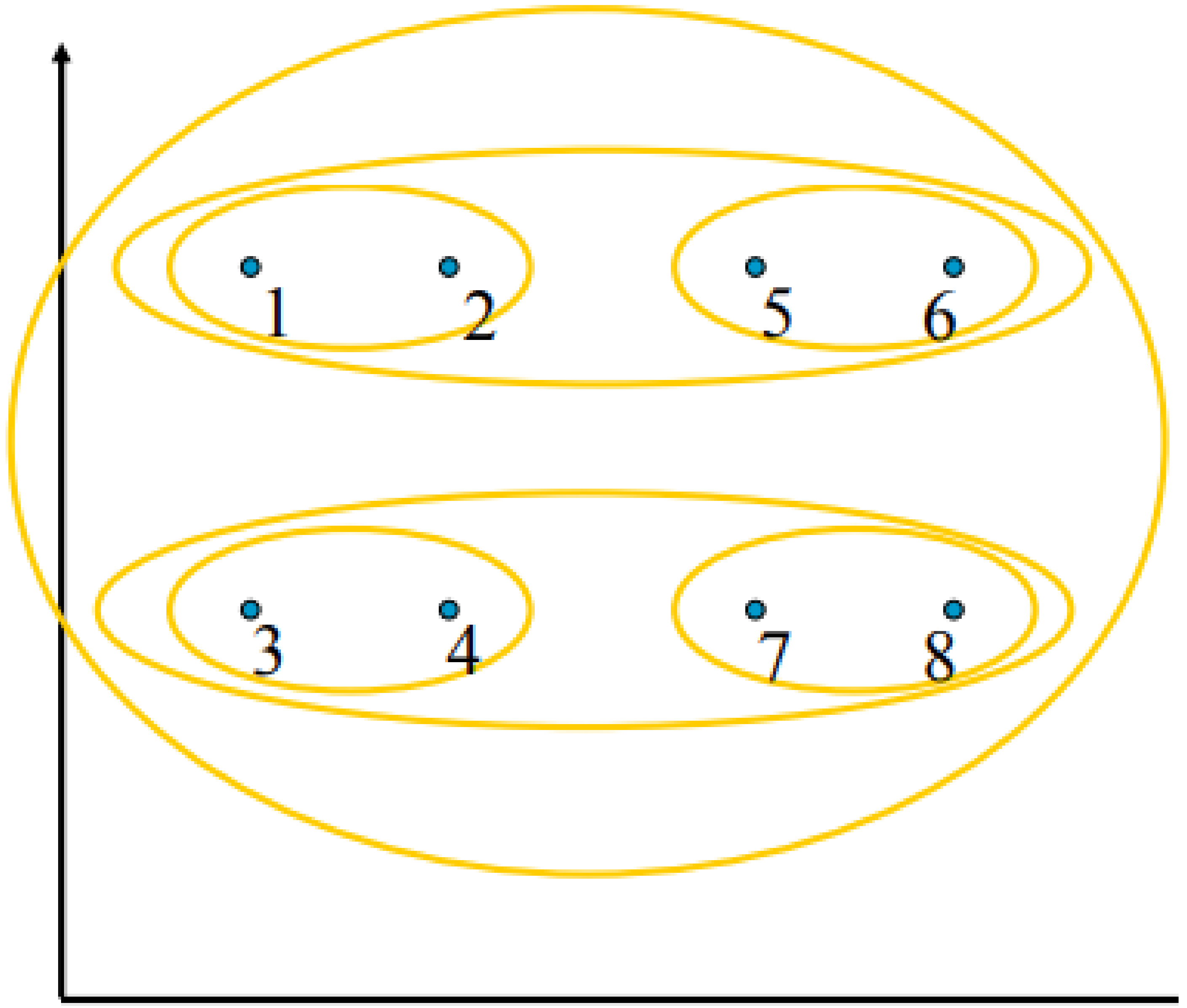
DISTANCE BETWEEN CLUSTERS

1. Average linkage;
2. Nearest neighbor;
3. Furthest neighbor;
4. Centroid clustering;
5. Median clustering;
6. Ward's method;
7. etc.

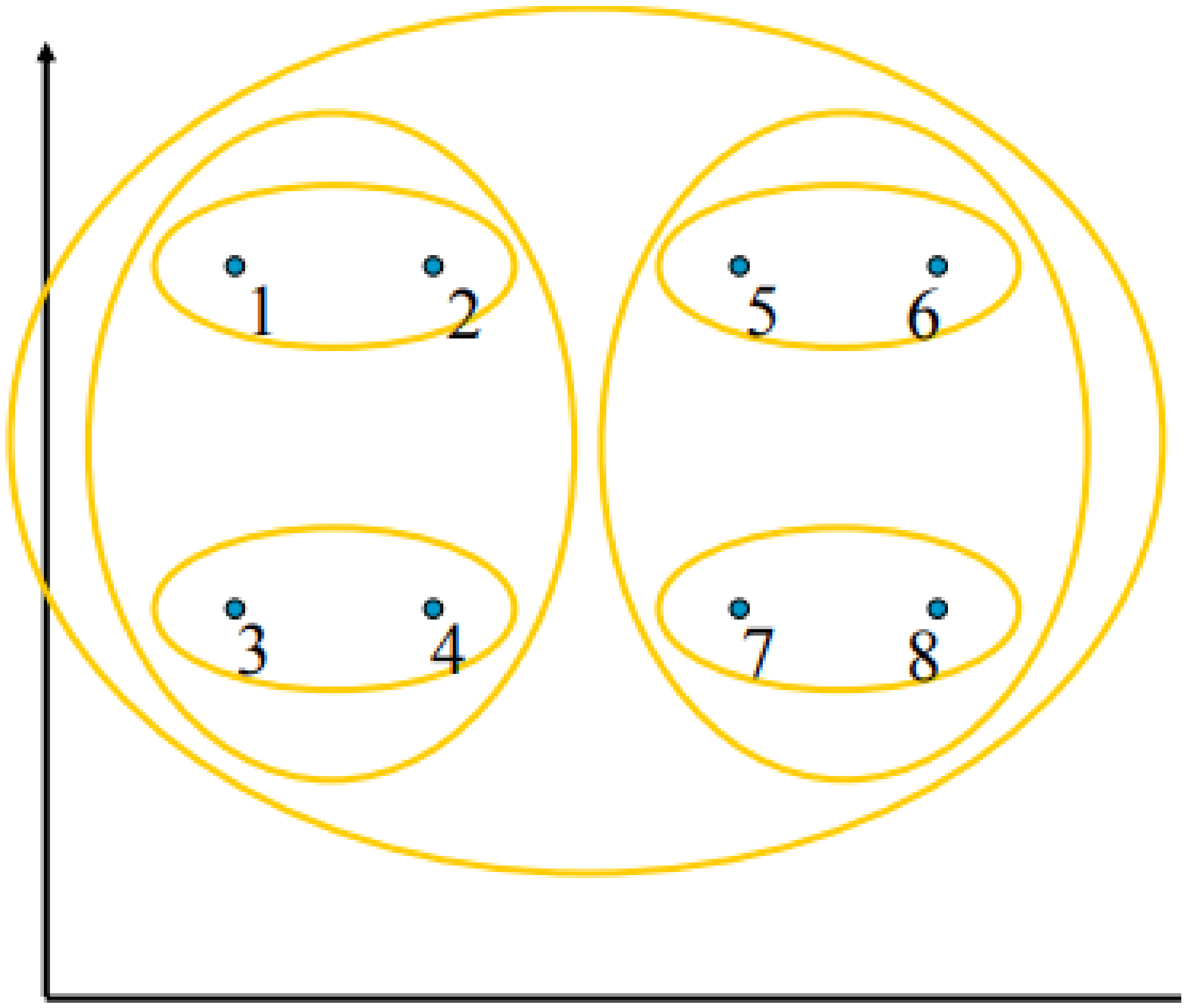


Different choices create different clustering behaviors.

DISTANCE BETWEEN CLUSTERS



Nearest neighbor



Furthest neighbor

5. AGGLOMERATIVE HIERARCHICAL CLUSTERING

Pros:

- No need to specify the number of clusters. You have the option of choosing the best-looking clusters.
- This algorithm is not sensitive to the choice of distance metric.

Cons:

- Complexity (naively, $O(n^3)$);

5. AGGLOMERATIVE HIERARCHICAL CLUSTERING

Pros:

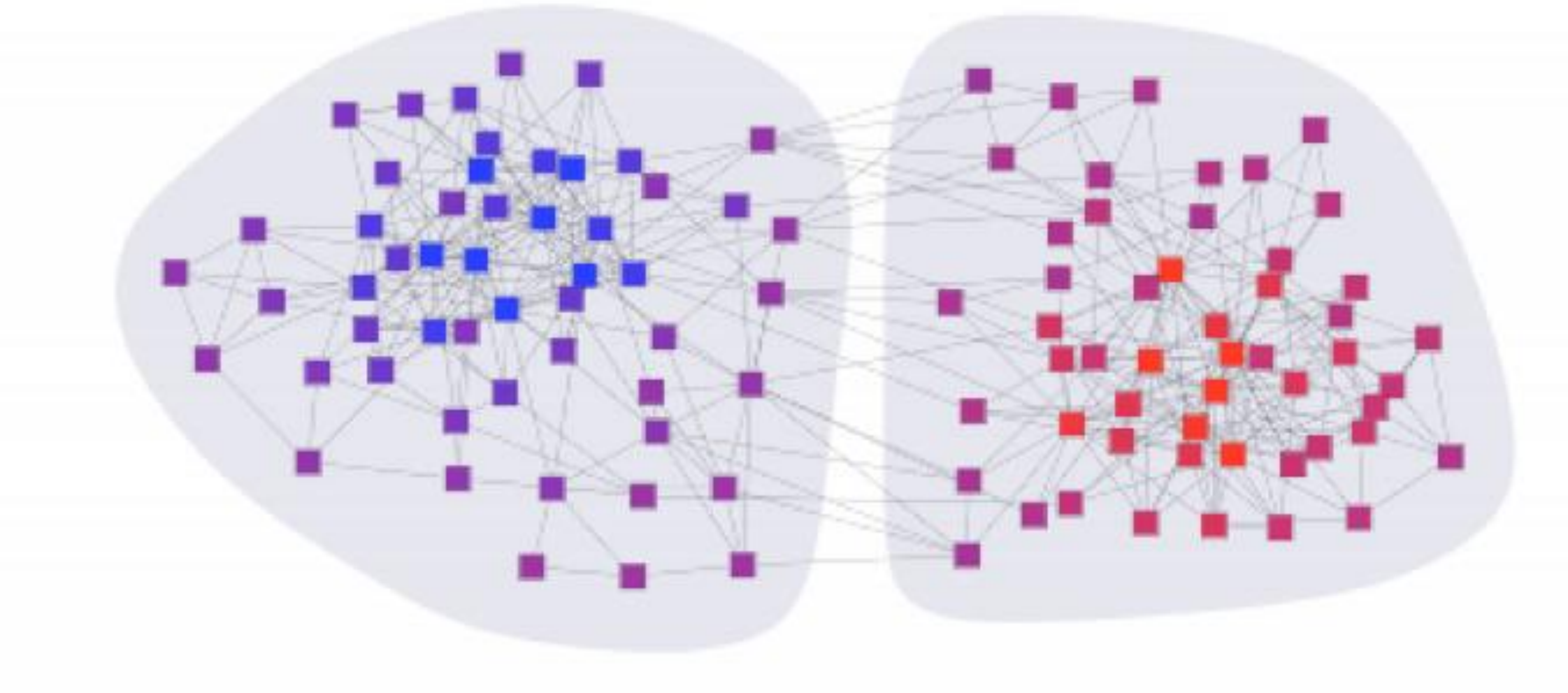
- No need to specify the number of clusters. You have the option of choosing the best-looking clusters.
- This algorithm is not sensitive to the choice of distance metric.

Cons:

- Complexity (naively, $O(n^3)$);

6. SPECTRAL CLUSTERING

- Represent data points as vertices *V* of a graph *G*.
- Each pair of vertices is connected by an edge.
- Edges have weights W . Large weights mean that adjacent vertices are similar.
- The graph construction depends on the application.

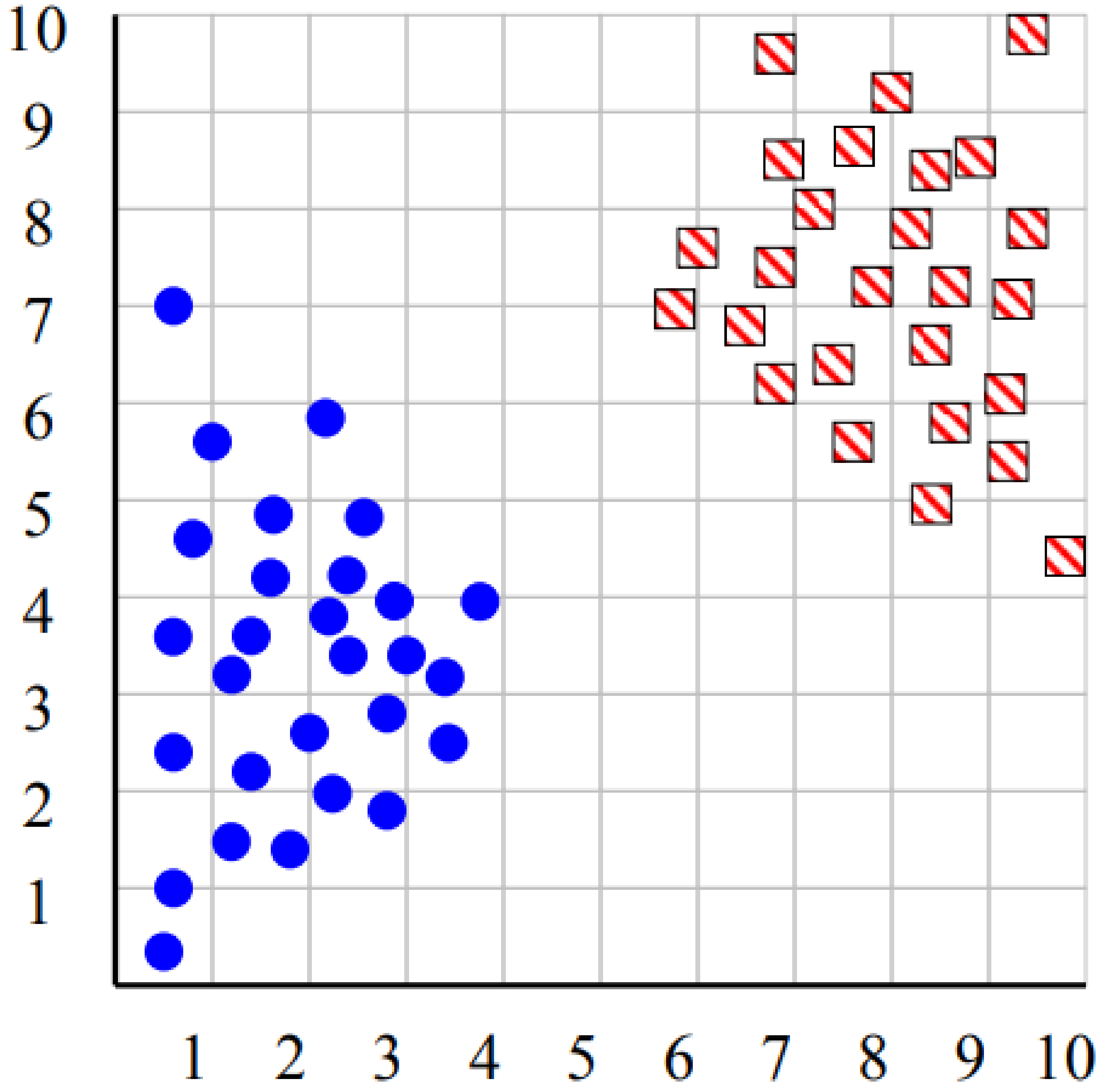


HOW CAN WE TELL THE RIGHT NUMBER OF CLUSTERS?

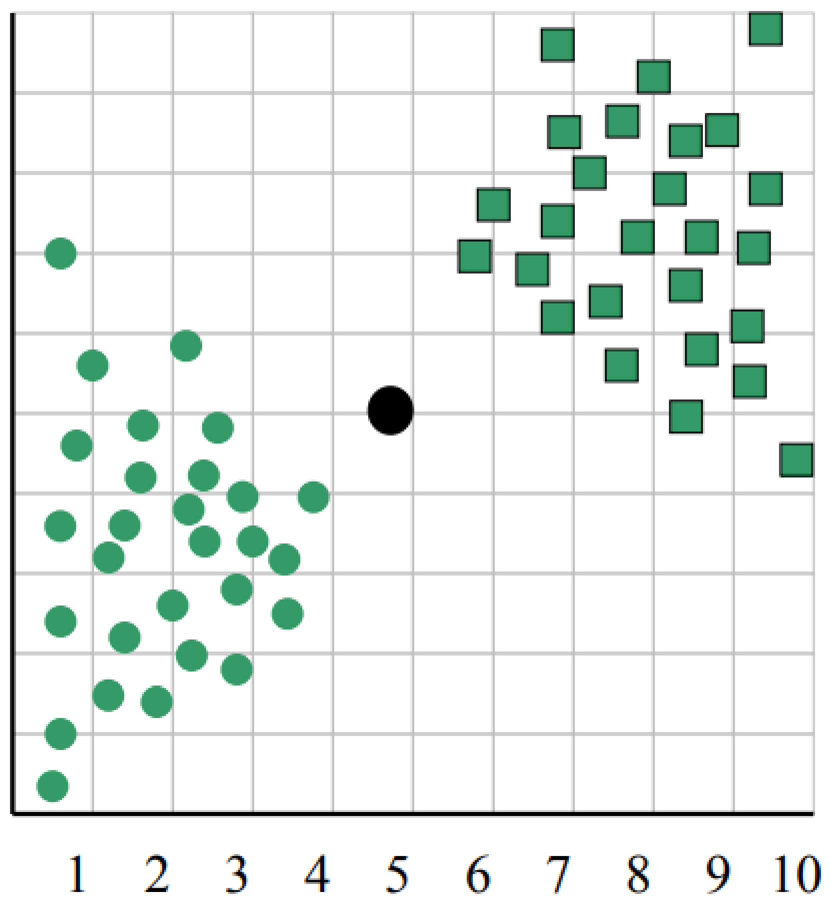
NUMBER OF CLUSTERS

In general, this is a **unsolved problem**. However there are many approximate methods.

Example:

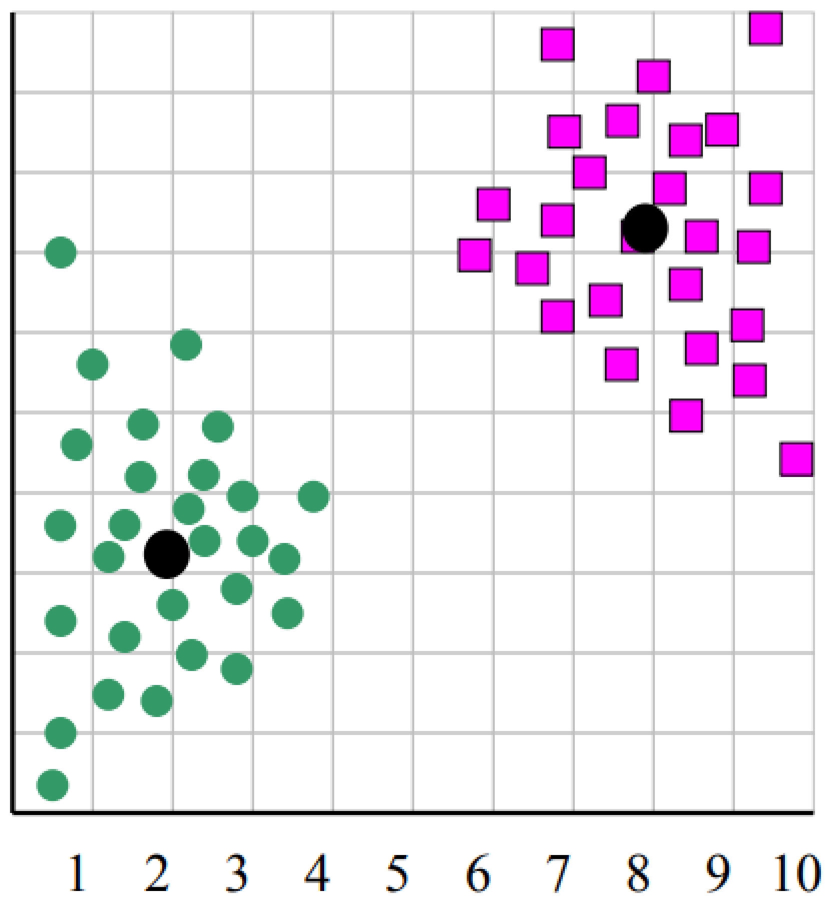


NUMBER OF CLUSTERS



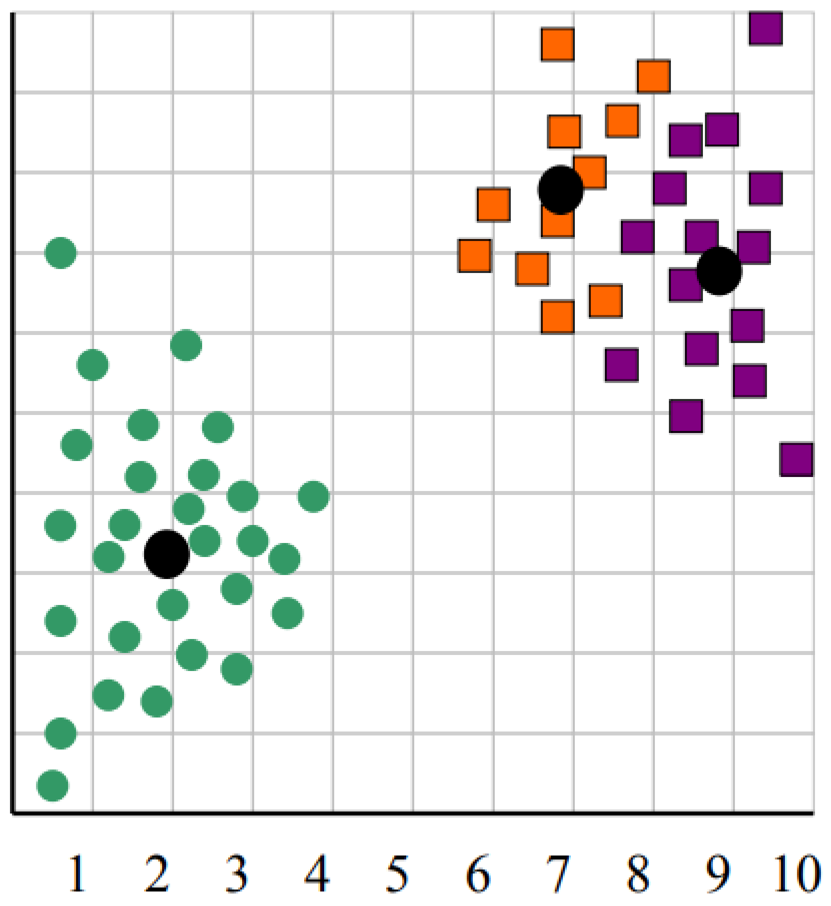
$k = 1$

$f = 873$



$k = 2$

$f = 173$

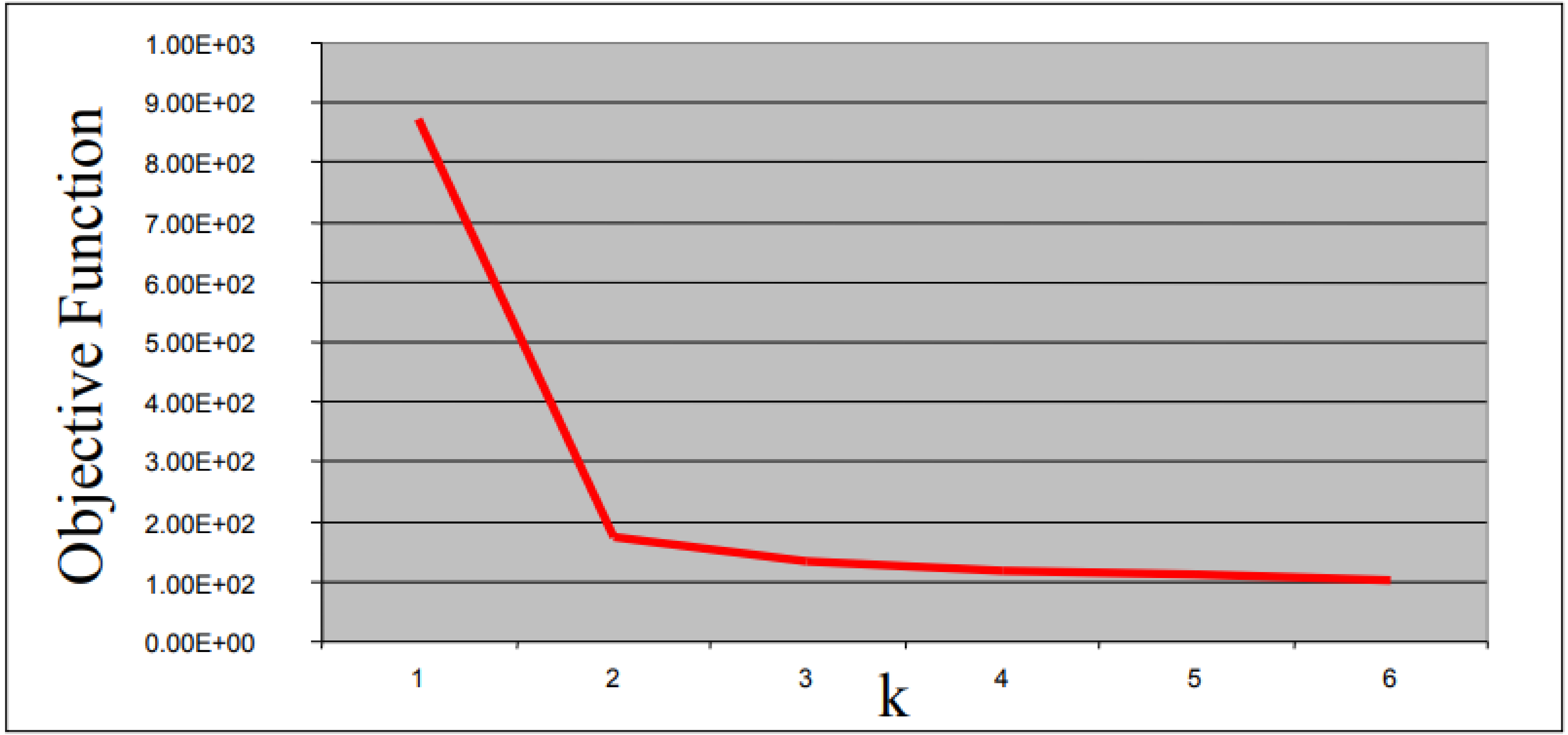


$k = 3$

$f = 133$

NUMBER OF CLUSTERS

Example:



The abrupt change at k = 2, is highly suggestive of two clusters in the data. This technique for determining the number of clusters is known as “knee finding” or “elbow finding”.

CONCLUSION

- Clustering is useful;
- There are different types of clustering algorithms based on different assumptions;
- There are some unsolved issues: number of clusters, initialization, etc.

HIGHER · SCHOOL · OF · ECONOMICS

NATIONAL RESEARCH
UNIVERSITY

Thank you!