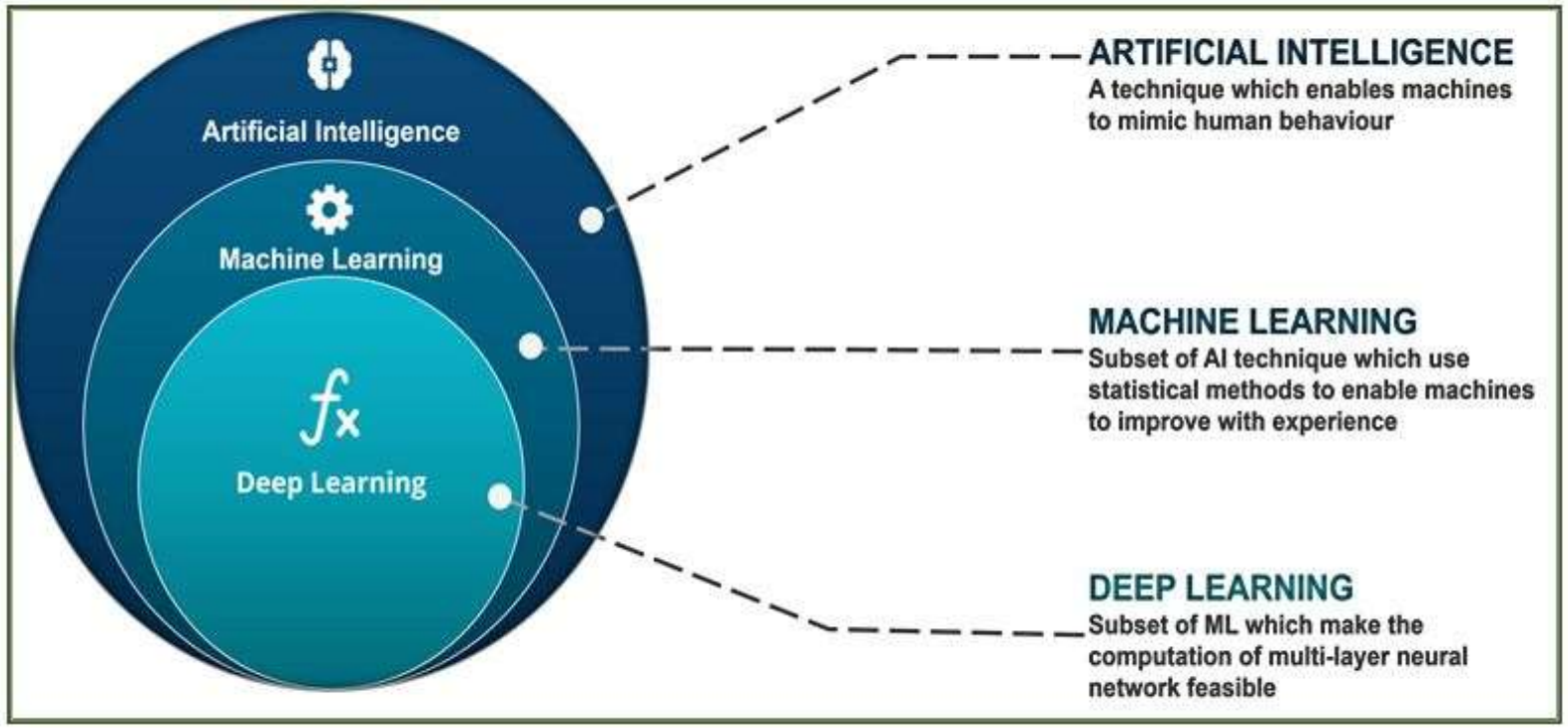
AI – ML – DL

- What is AI, ML and DL?



Machine Learning

- Supervised Learning
- Unsupervised Learning
- Semi supervised Learning
- Reinforcement Learning

ML

- Supervised Learning
  - Technique to achieve AI through algorithm trained with data
  - Function that maps and input to an output based on a series of examples
  - Examples: input-output pairs

Regression

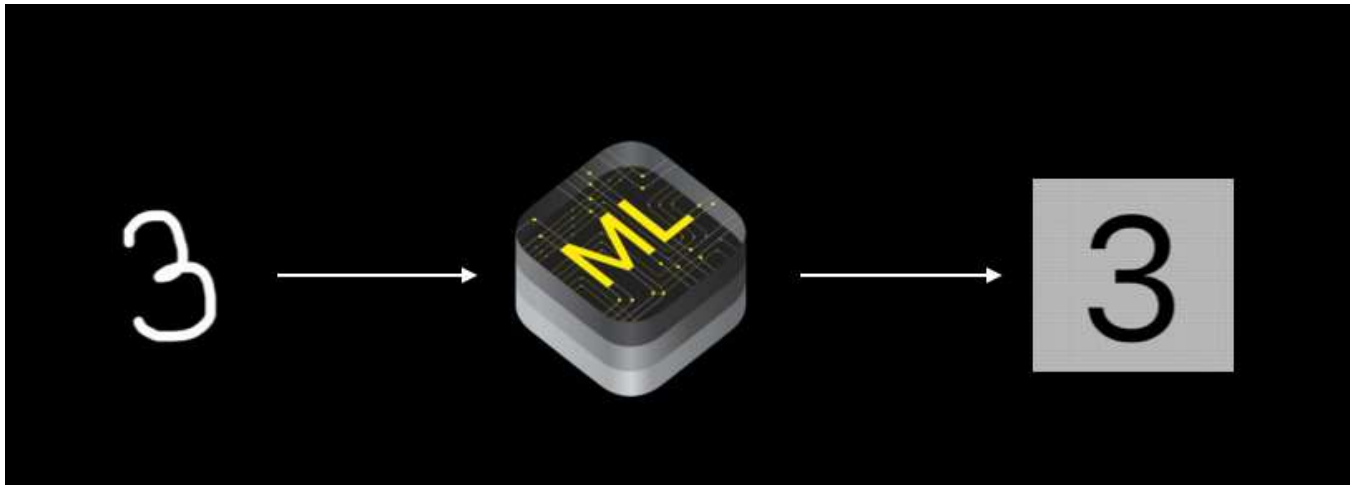
- To find/predict an output based on independent predictors
- The output is continuous
- Methods:
  - Linear regression; Decision tree; Random forest; NNETs; …

Classification

- To assign a class to an output (discriminate, separate)
- The output is discrete
- Methods:
  - Logistic regression; Linear discriminant; Support Vector Machines; Naïve Bayes; NNETs; DL based models; …

Examples

- Recognition of handwriting digits



ML

- Unsupervised Learning
  - Find patterns from input data without references to the outcomes

Clustering

- Methods:
  - K-means; Hierarchical; Mean-shift; Density-based …

Dimensionality reduction

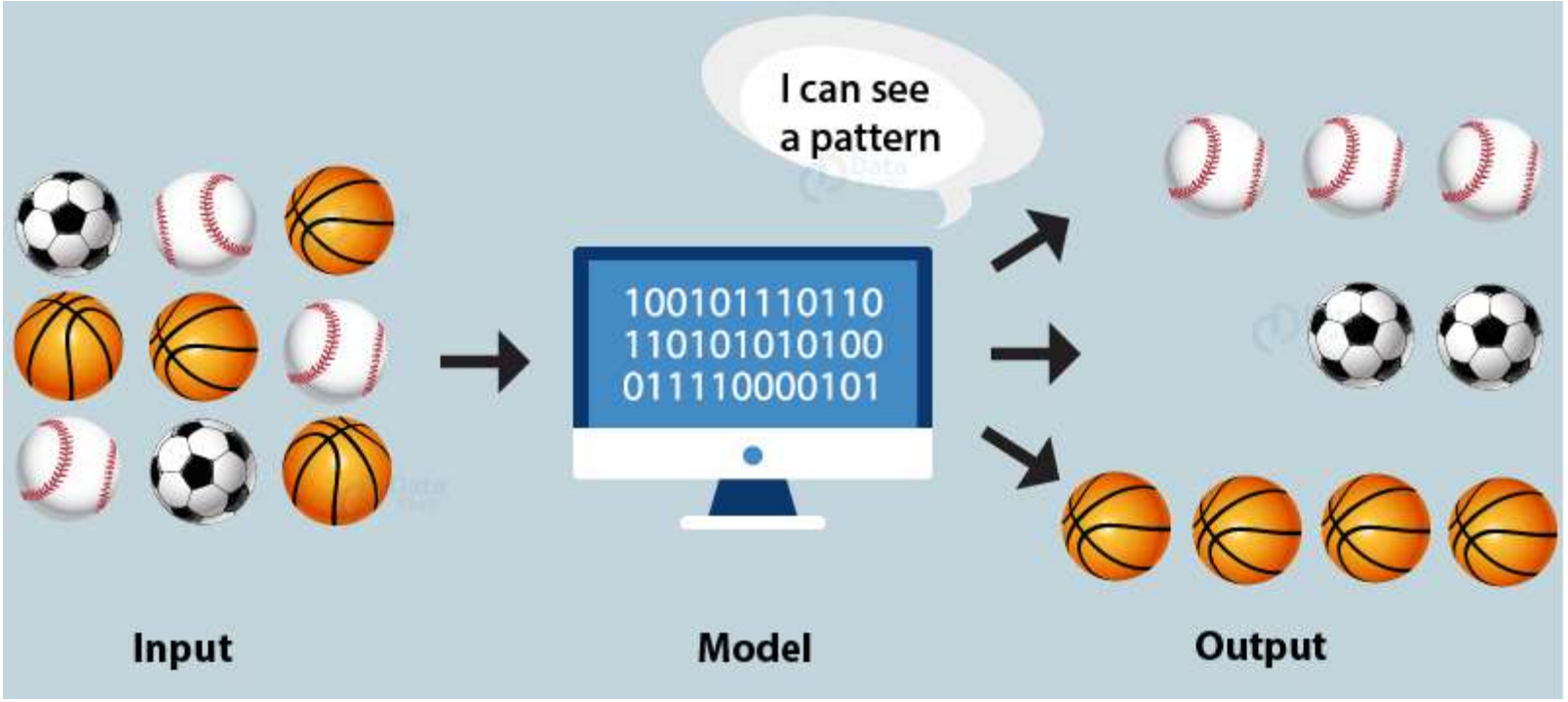
- Elimination or extraction of variables
- Methods:
  - PCA; …

Tecnológico de Monterrey

K-means clustering

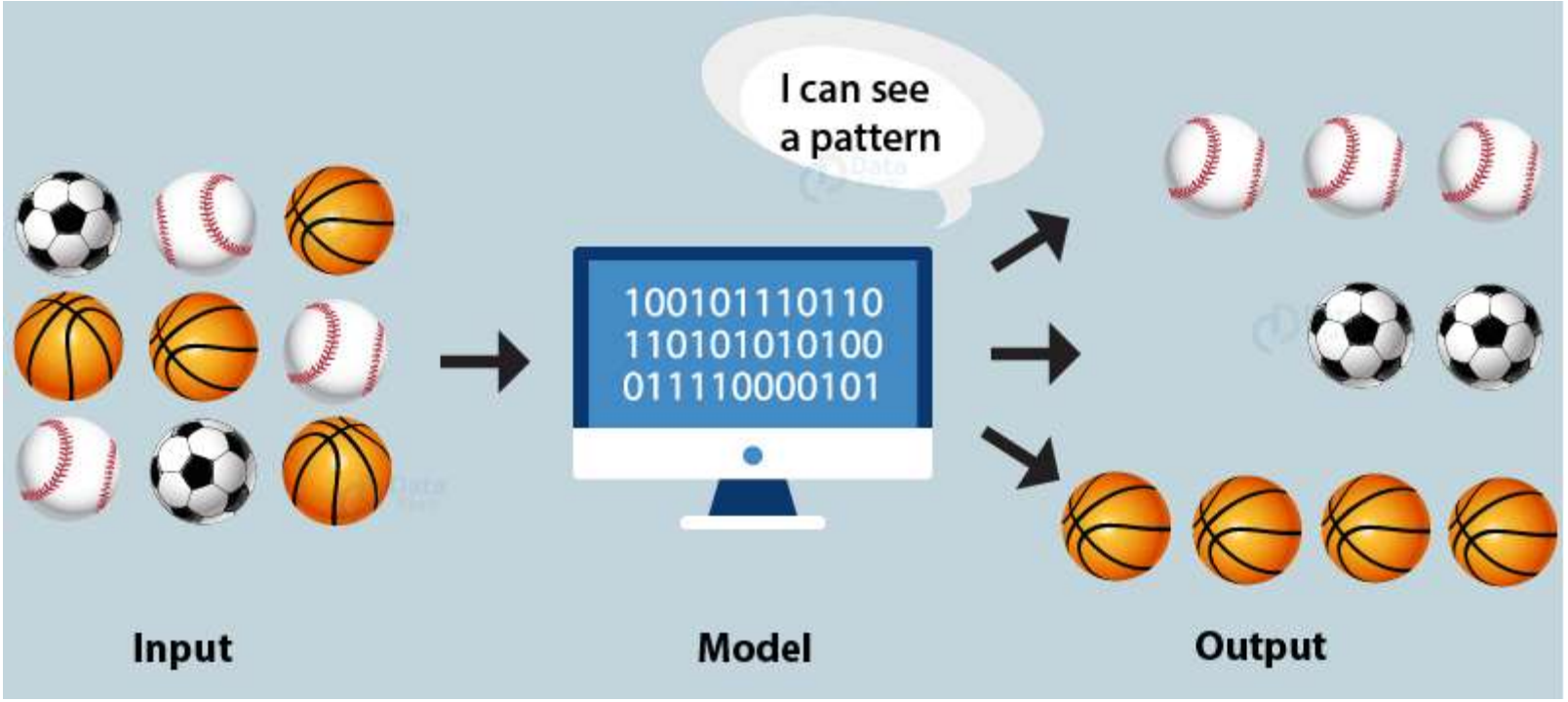
Javier M. Antelis
mauricio.antelis@tec.com
Escuela de Ingeniería y Ciencias
Tecnológico de Monterrey
Guadalajara, México

What is Clustering?



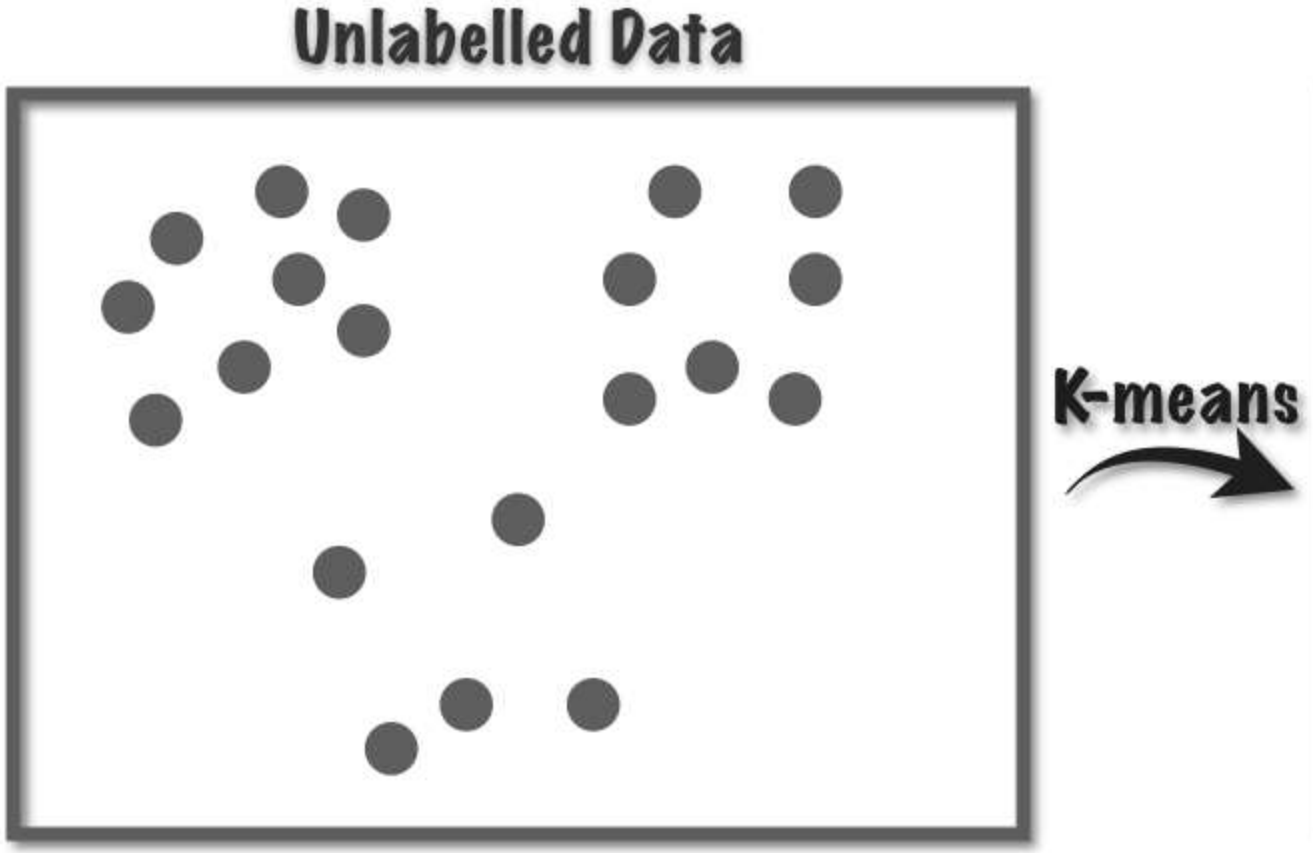
- Clustering is a set of "computational" techniques used to partition data objects into groups (or clusters)

What is Clustering?

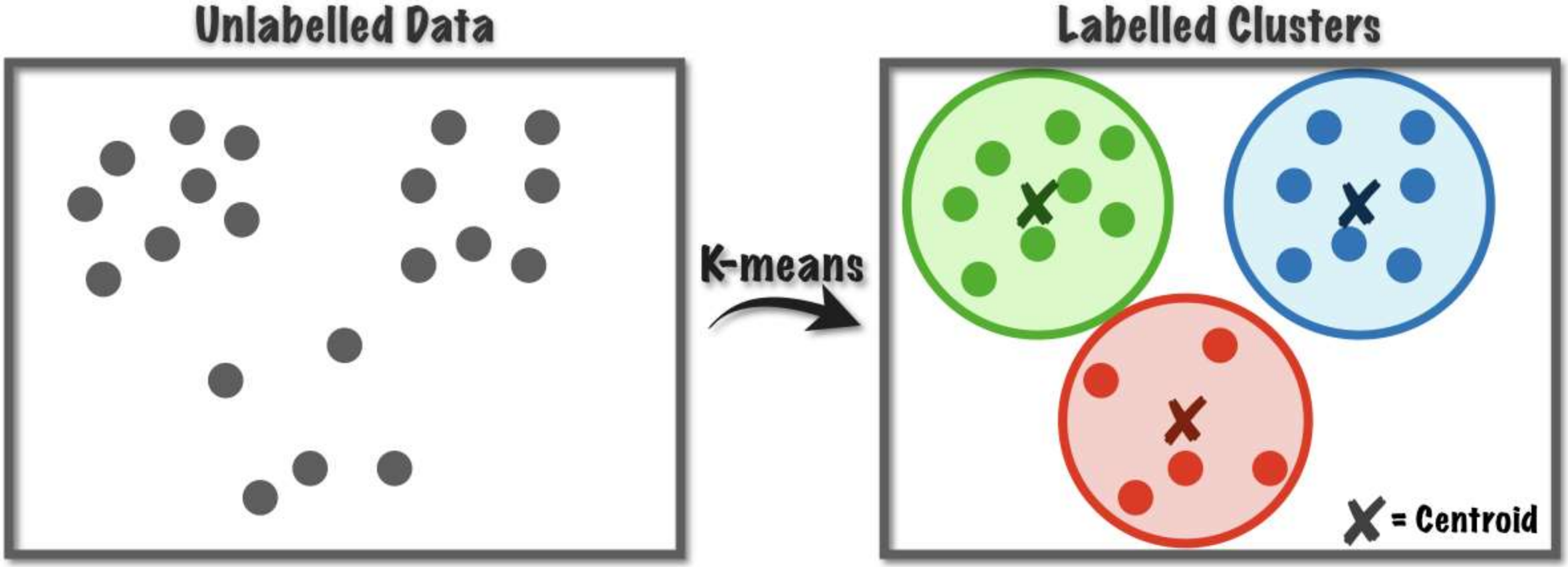


- Clusters are defined as groups of data objects that are more similar to other objects in their cluster than they are to data objects in other clusters

Clustering algorithms



Clustering algorithms



Types of clustering algorithms

- K-means Clustering: Using this algorithm, we classify a given data set through a certain number of predetermined clusters or “k” clusters.
- Hierarchical Clustering: follows two approaches Divisive and Agglomerative.
- Fuzzy C means Clustering: The working of the FCM Algorithm is almost similar to the k-means clustering algorithm, the major difference is that in FCM a data point can be put into more than one cluster.
- Density-Based Spatial Clustering: Useful in the application areas where we require non-linear cluster structures, purely based on density

What is K-means?

- K-means clustering method is an unsupervised machine learning technique used to identify clusters in data.

- K-means is an iterative algorithm that divides a group of data points into K subgroups/clusters based on the similarity of their mean distance from the centroid of that particular subgroup/formed.

- K is the pre-defined number of clusters to be formed by the algorithm

What is K-means?

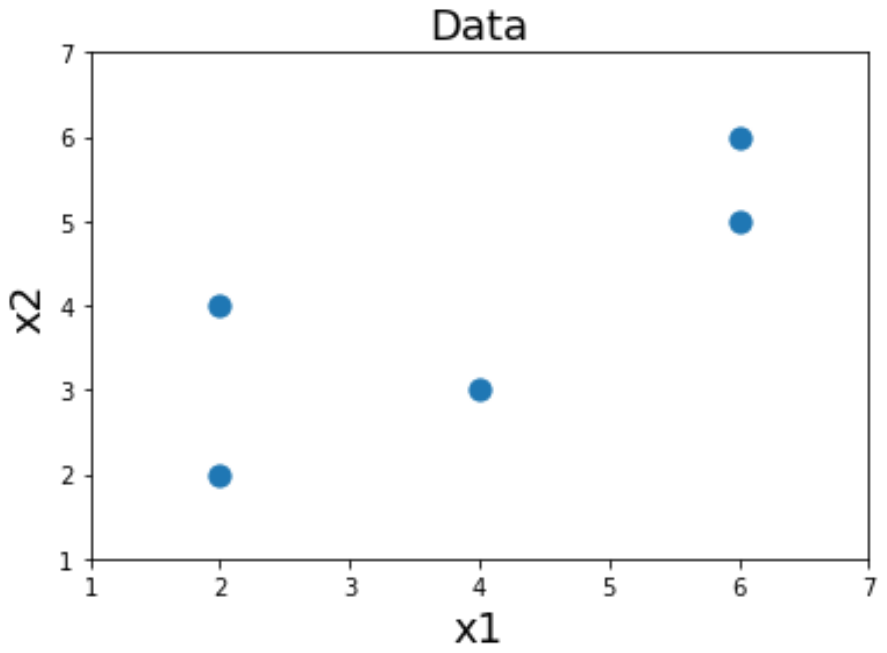
Algorithm 1 k-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

Intuitive explanation on K-means clustering

- We have a set of data points that we want to cluster
- We do not know the number of clusters in advance, that is, the data is not labelled

	x1	x2
0	2	2
1	4	3
2	2	4
3	6	6
4	6	5



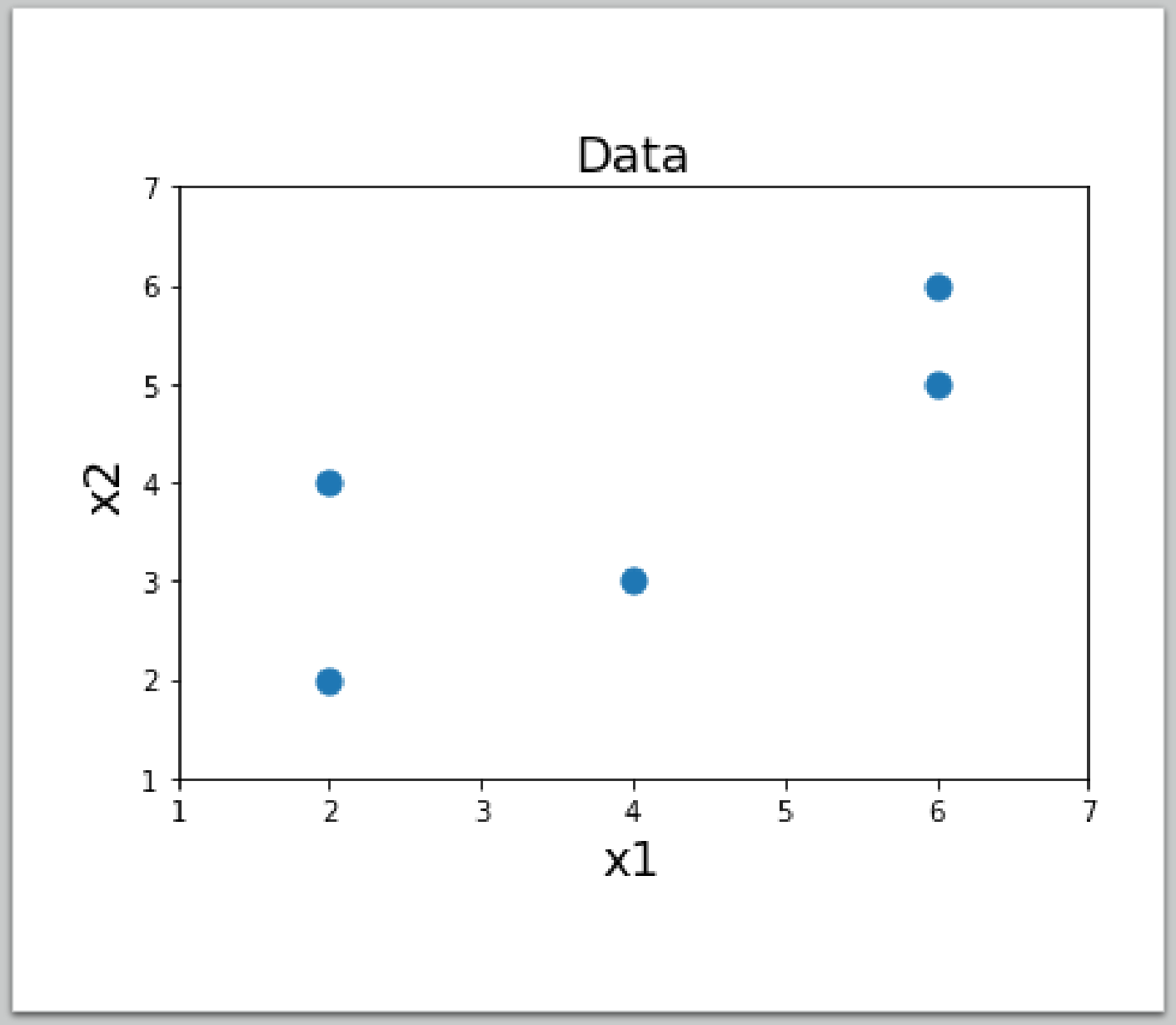
Graphical illustration

- Data

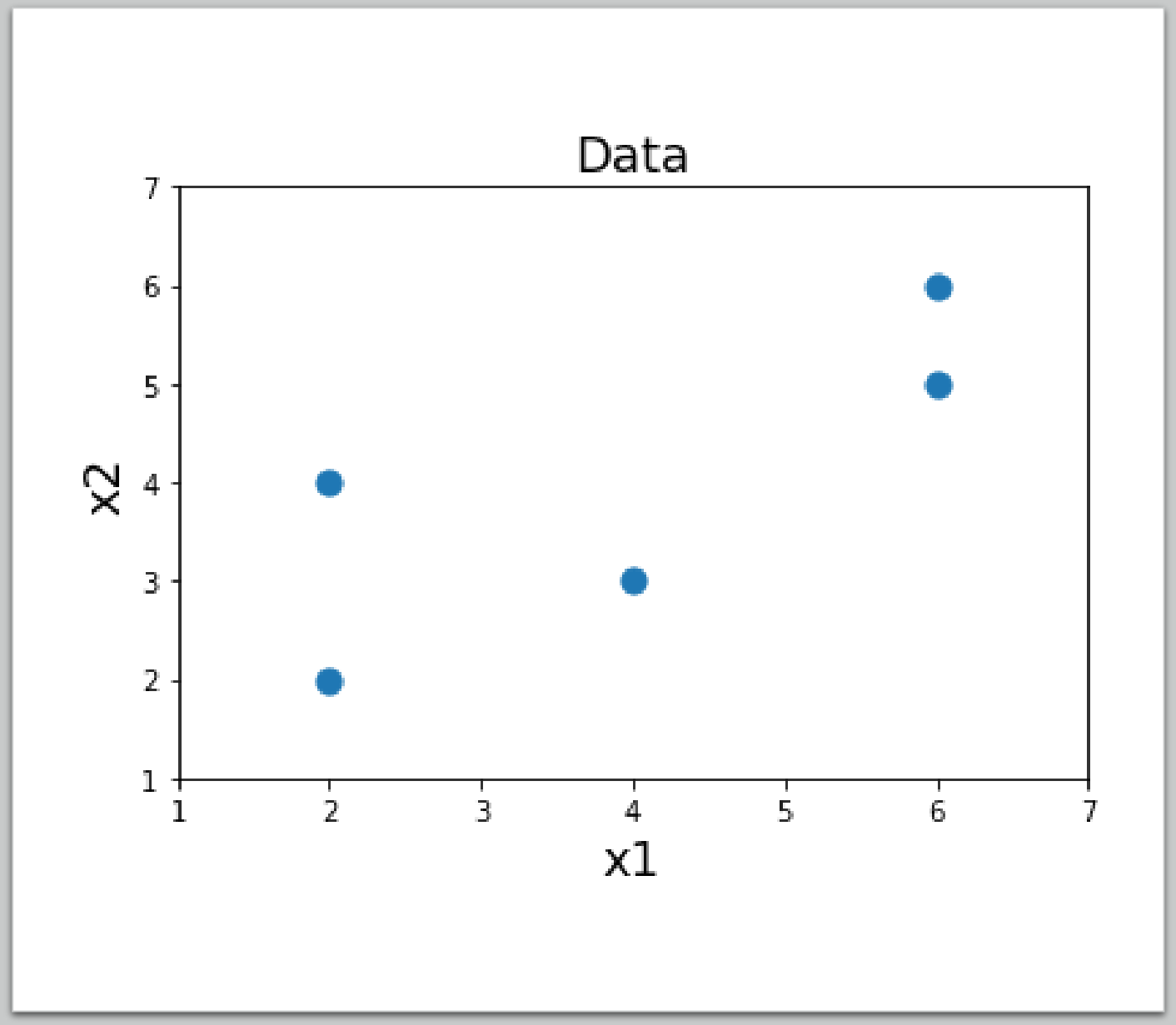
	x1	x2
0	2	2
1	4	3
2	2	4
3	6	6
4	6	5

Graphical illustration

- Data



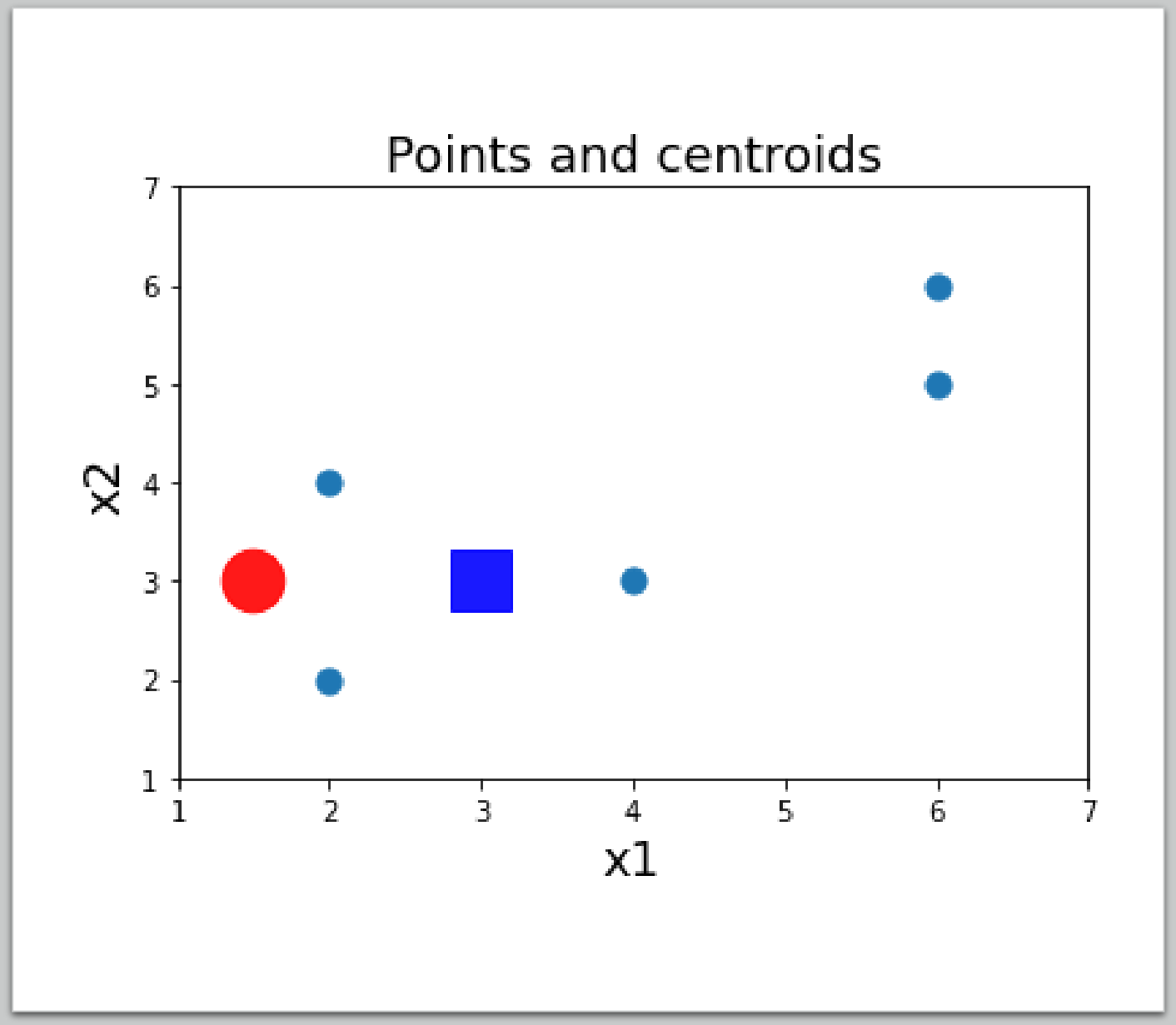
Graphical illustration



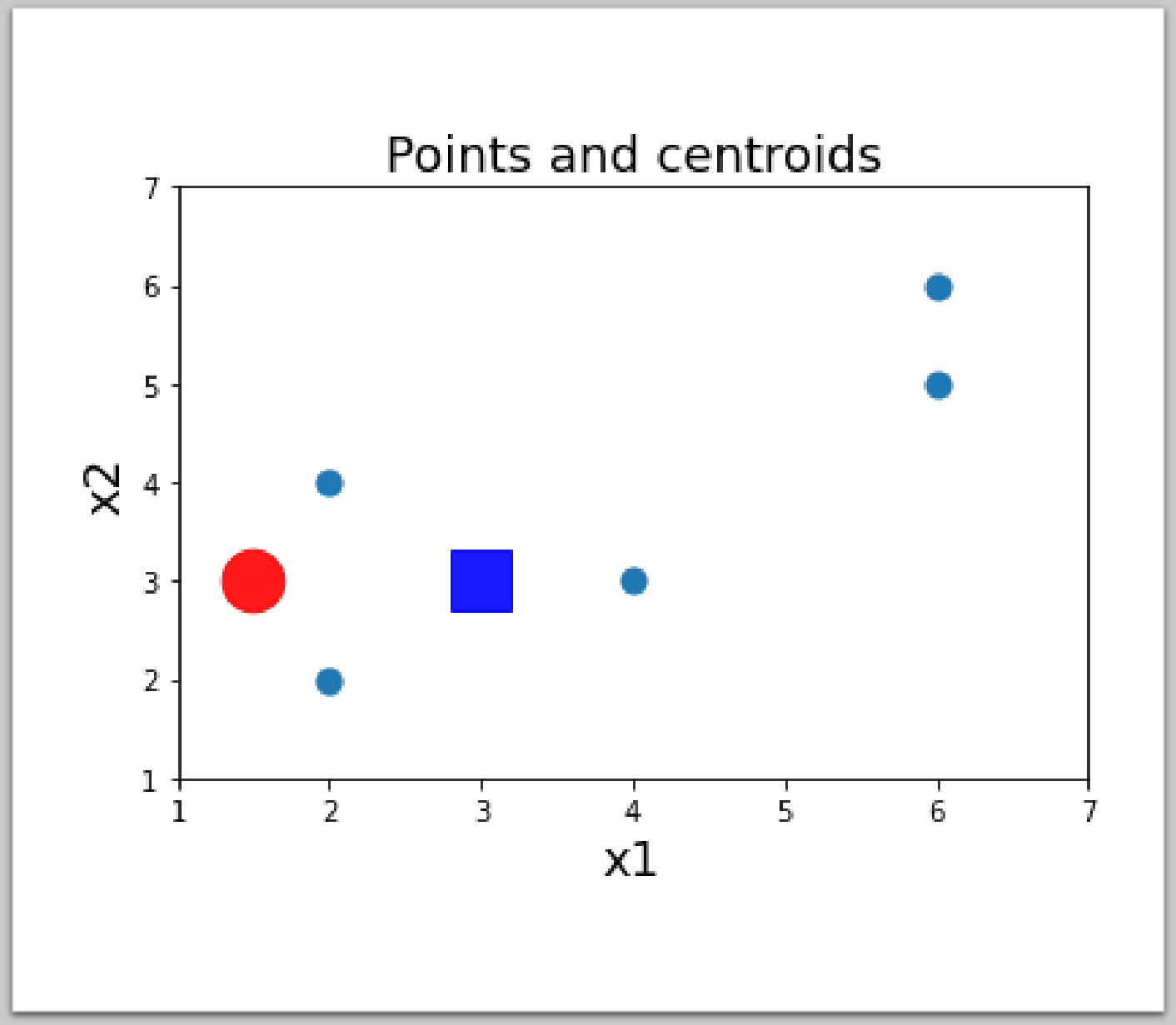
- 1: Specify the number of clusters. K=?

Graphical illustration

- 2: Initialize the centroids of the clusters (randomly)

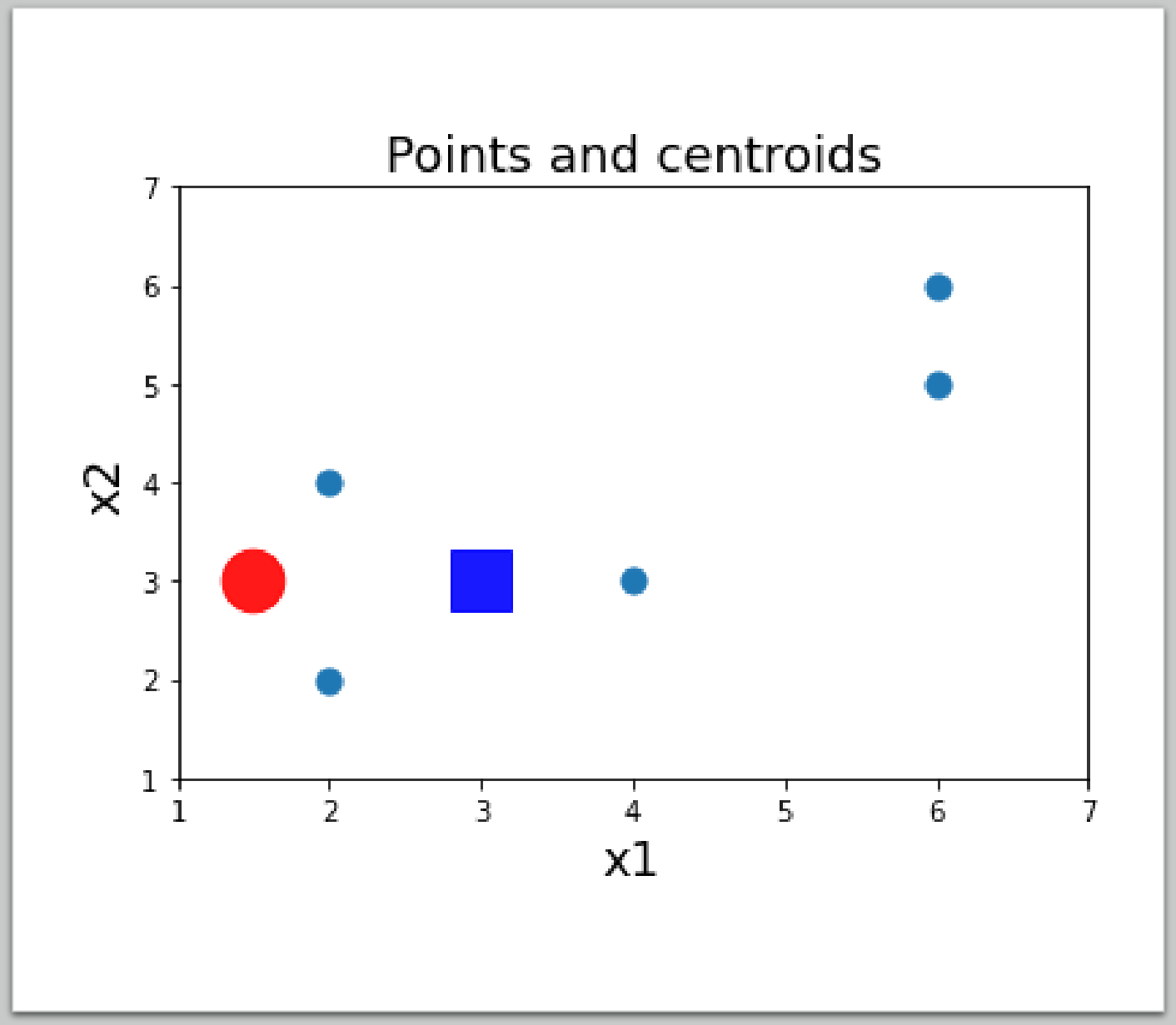


Graphical illustration



- 3: Repeat the following (with the current centroids)

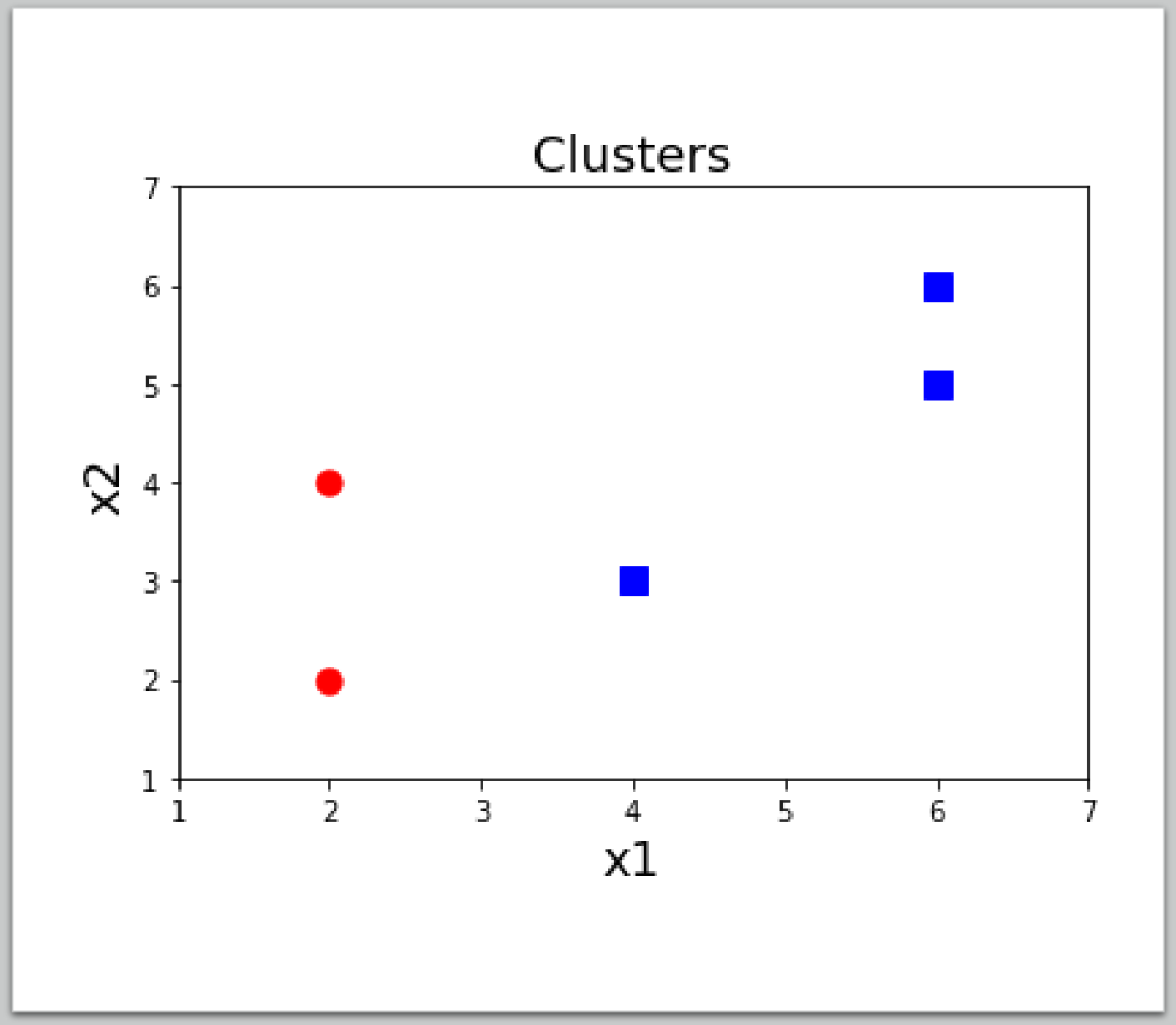
Graphical illustration



- 4: Assign each point to its closest centroid

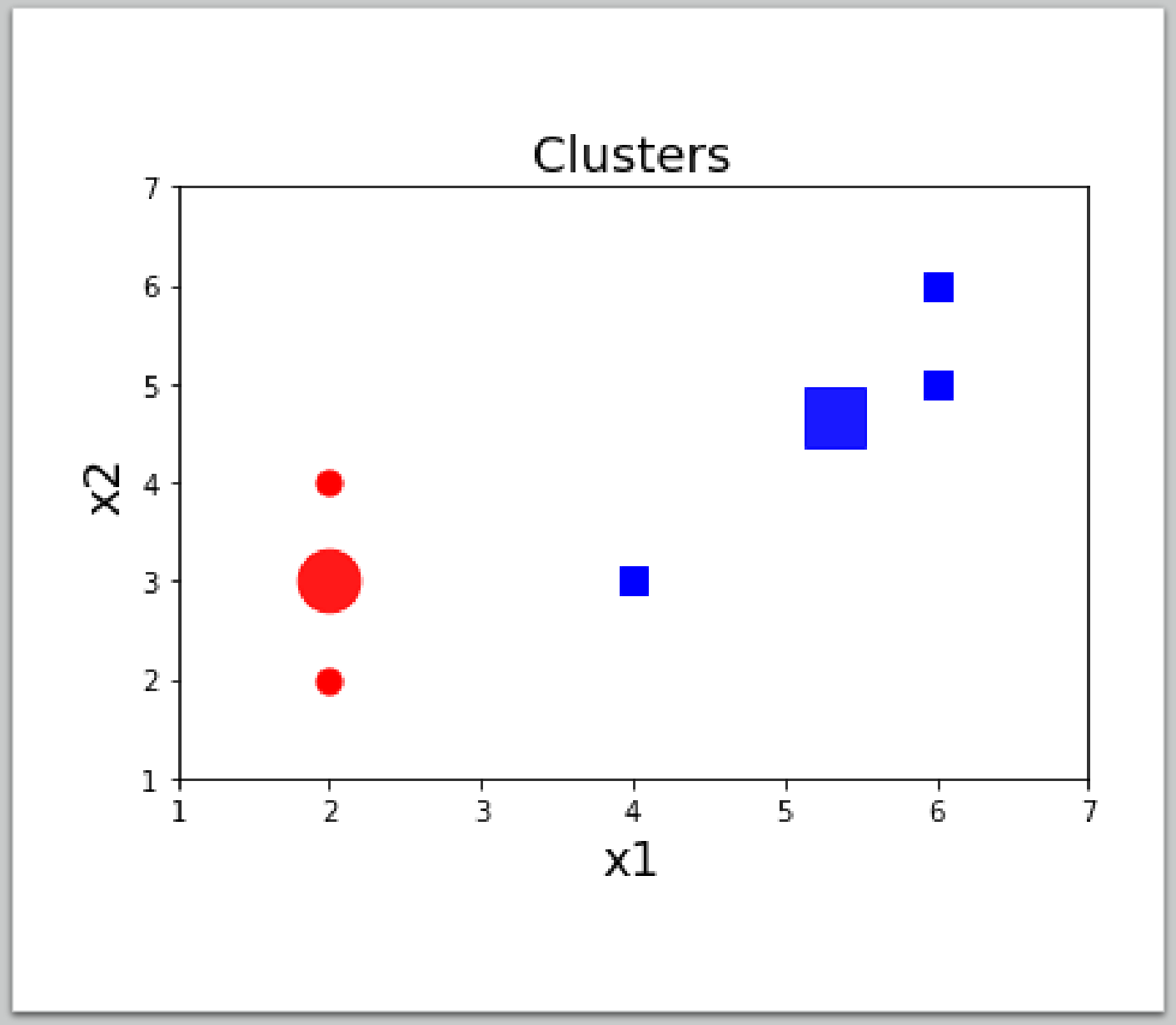
Graphical illustration

- 4: Assign each point to its closest centroid



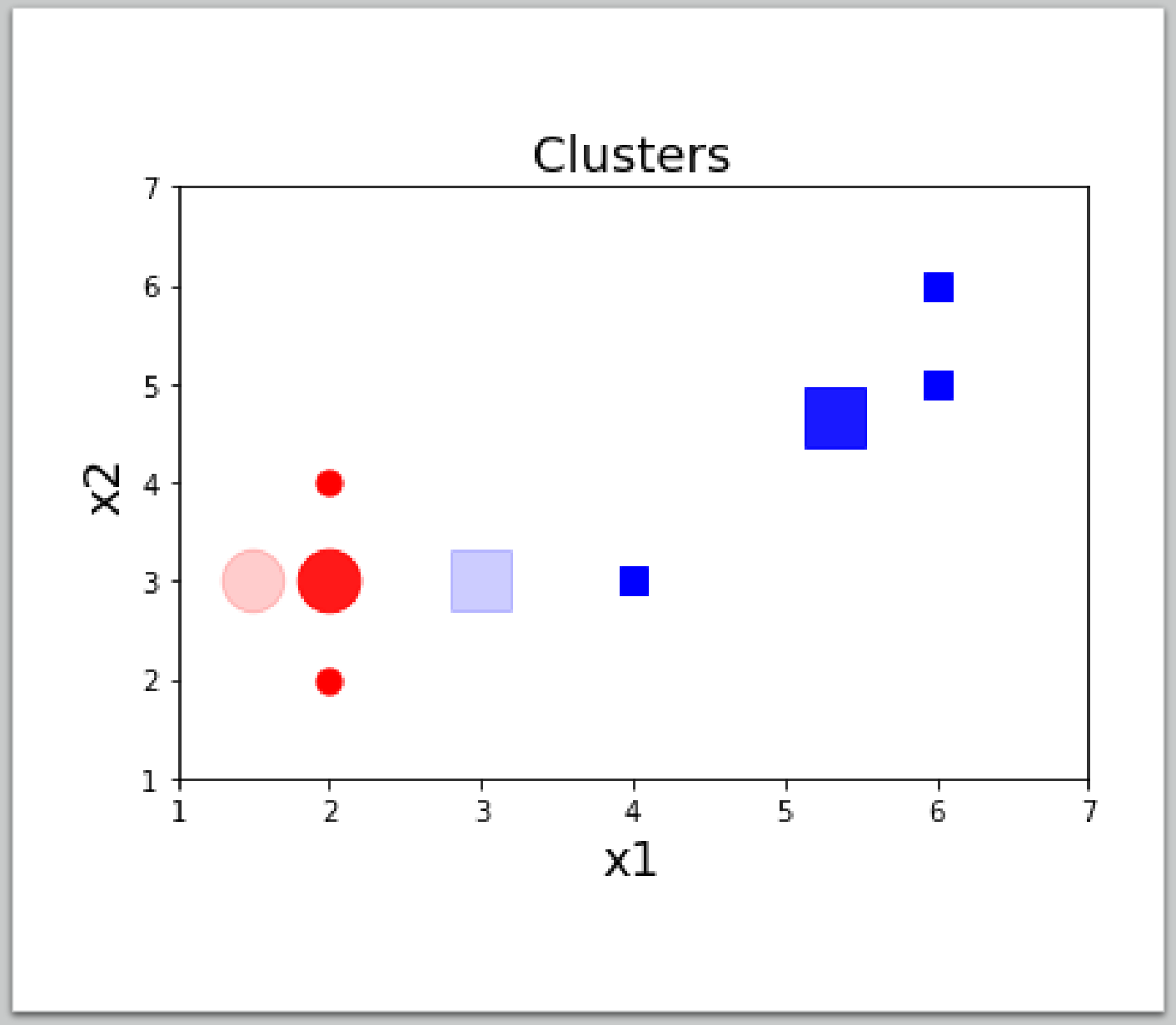
Graphical illustration

- 5: Compute the new centroid (mean) of each cluster

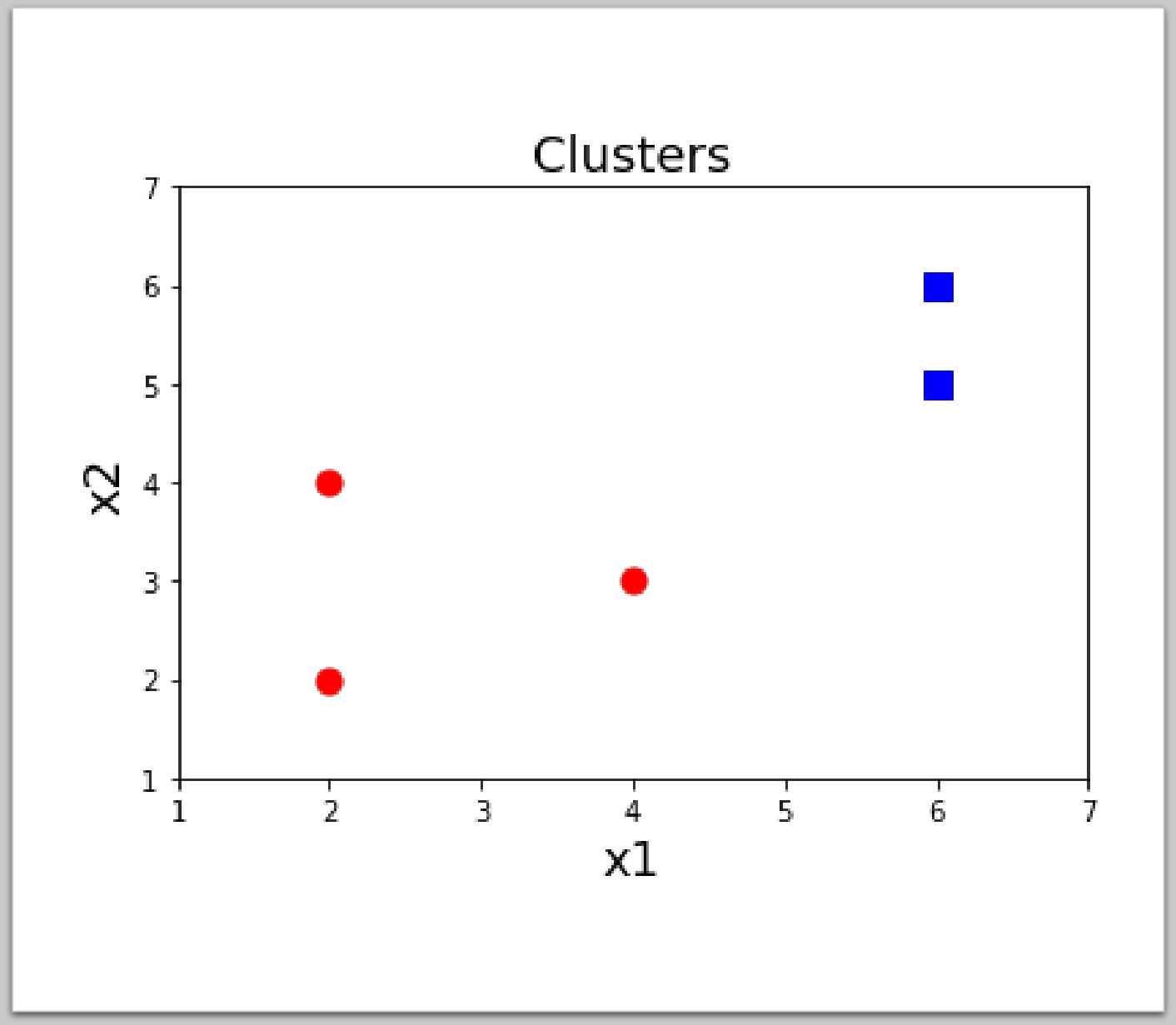


Graphical illustration

- 6: Until the centroids do not change



Graphical illustration



- Final solution

How to choose the number of clusters K

- Elbow method gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids.