Unsupervised Learning: Part I

An Introduction to Cluster Analysis

Efthymios Costa

June 9, 2025

> The Milky Way is nothing else but a mass of innumerable stars planted together in **clusters**.
>
> *Galileo Galilei*

Types of Learning Revisited

	Supervised Learning	**Unsupervised Learning**	**Reinforcement Learning**
Input Data	Features + labels (e.g., X, y)	Only features (X)	Environment feedback (states, actions, rewards)
Objective	Predict output	Identify patterns	Optimise decision making
Output	Predicted label or value	Cluster assignment, reduced dimensions, etc.	Policy/Strategy for decision making
Feedback	Direct error signal	No explicit feedback; pattern discovery	Delayed reward signal from environment

Unsupervised Learning

- No explicit labels in the data, so cannot train a predictive model. 😨
- What else might be of interest, besides predictions? 🤔
- How about discovering patterns/structures in the data? 🧐

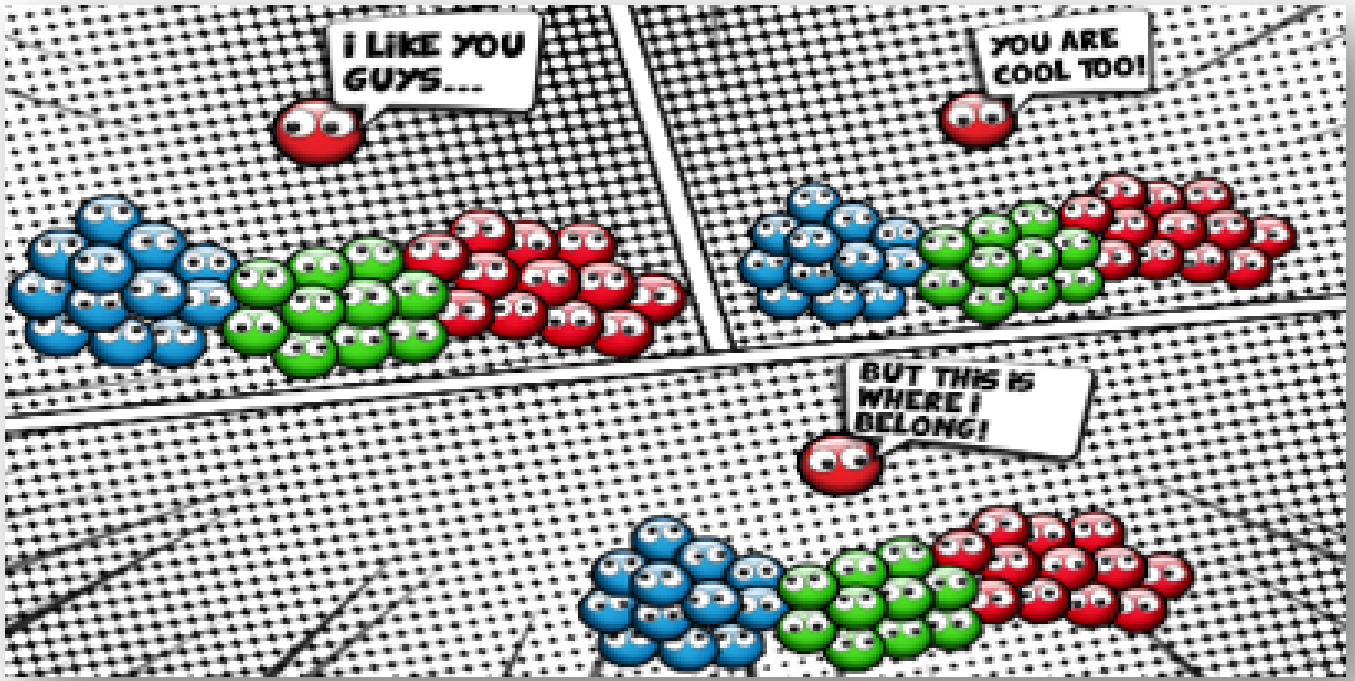
Example: Customer segmentation

- A retailer launches an advertisement campaign.
- They want to identify groups of people with common interests.
- This is a problem of **cluster analysis**.

Cluster Analysis: Definition

Definition

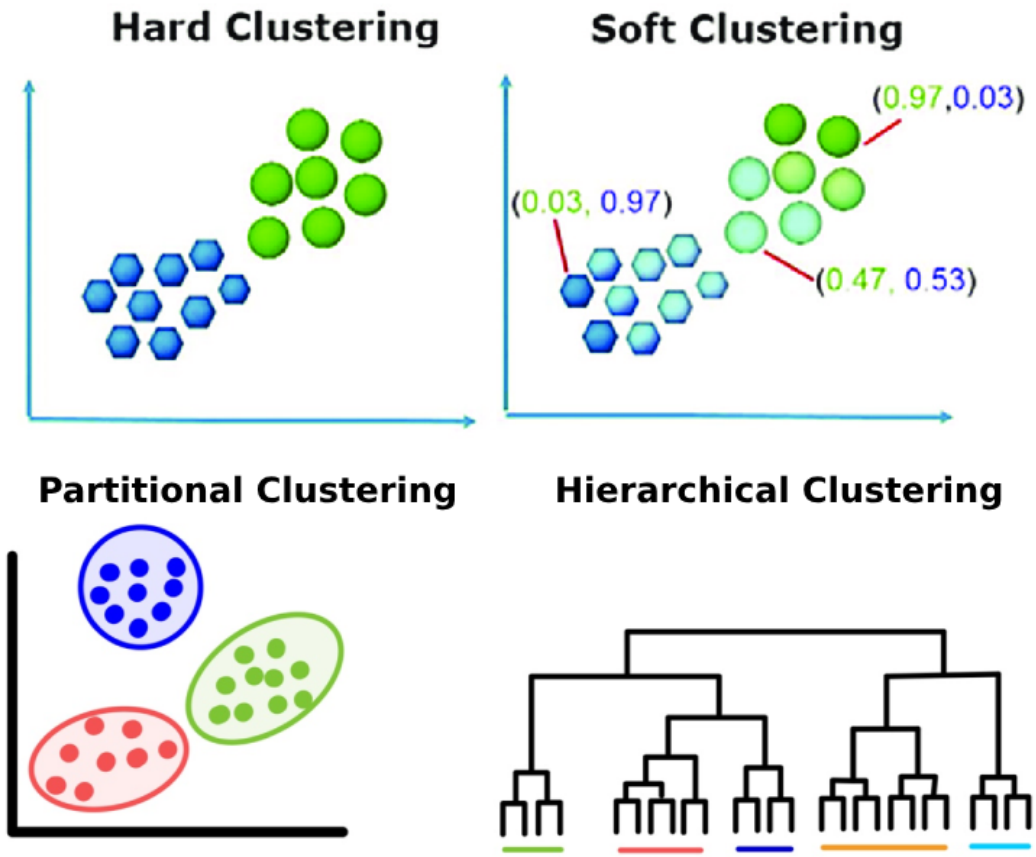
Cluster analysis (clustering) is the task of grouping a set of objects into groups (clusters) so that similar objects are assigned in the same cluster and less similar objects are assigned in distinct clusters.



Clustering approaches

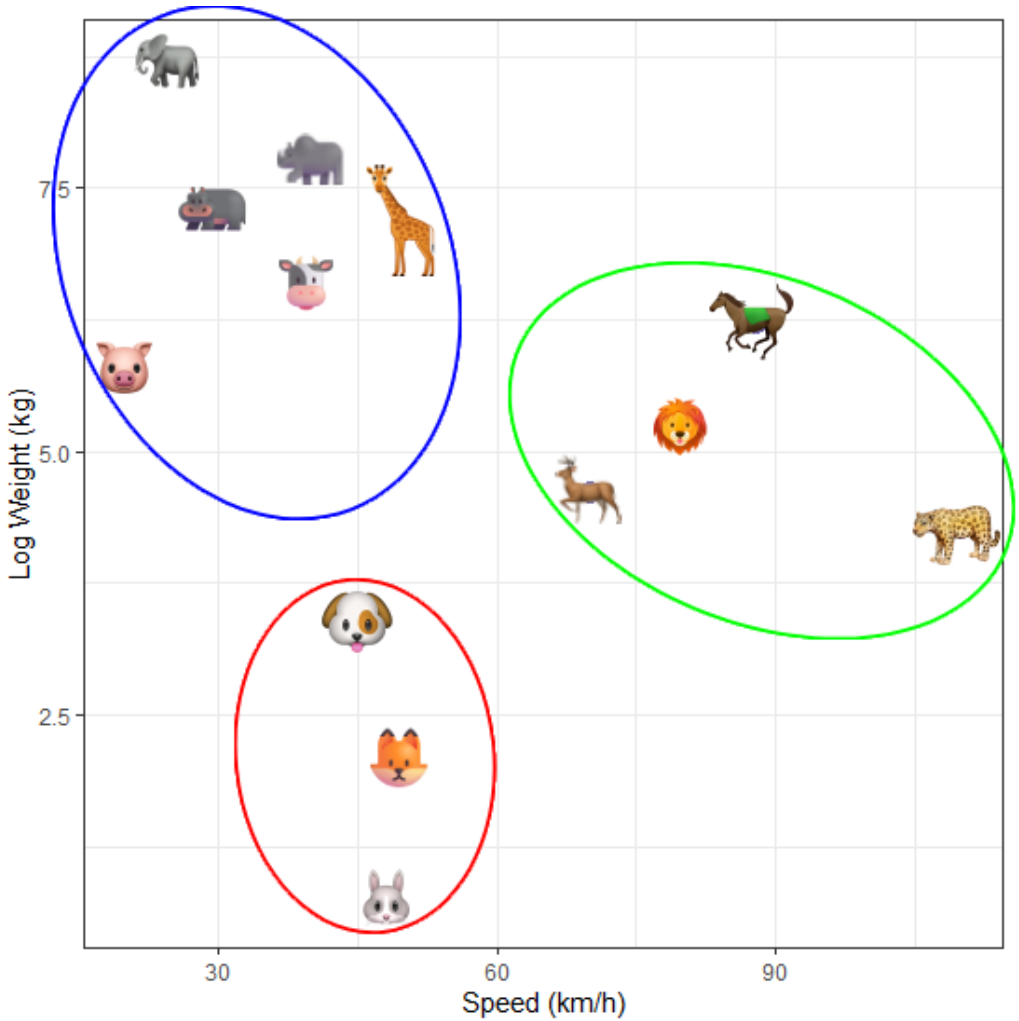
Clustering algorithms can be categorised in several ways:

- **By their output:** Hard/Soft clustering.
- **By their assumptions:** Distance-based/Model-based clustering.
- **By their structure:** Partitional/Hierarchical clustering.



Distance-based clustering

- **Main Idea:** Observations which are 'closer' are also 'more similar'.
- Can therefore assign observations in the same cluster if they are close and ensure distant observations are in different groups.



The K-Means Algorithm

- Let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ be our data, where $\mathbf{x}_i \in \mathbb{R}^p$.
- Assume we want to partition our data into K clusters.
- ***K*-Means Objective:** Find the best partition $\mathcal{C} = \{C_1, \ldots, C_K\}$ among the space of all partitions into K non-empty clusters $\mathcal{P}_K$ such that $C_i \cap C_j = \varnothing$ and:

$$\mathcal{C} = \underset{\mathcal{C}' \in \mathcal{P}_K}{\arg\min} \sum_{k=1}^{K} \sum_{i:\mathbf{x}_i \in C'_k} \|\mathbf{x}_i - \mathbf{m}_k\|^2, \quad \mathbf{m}_k = \frac{1}{|C'_k|} \sum_{i:\mathbf{x}_i \in C'_k} \mathbf{x}_i$$

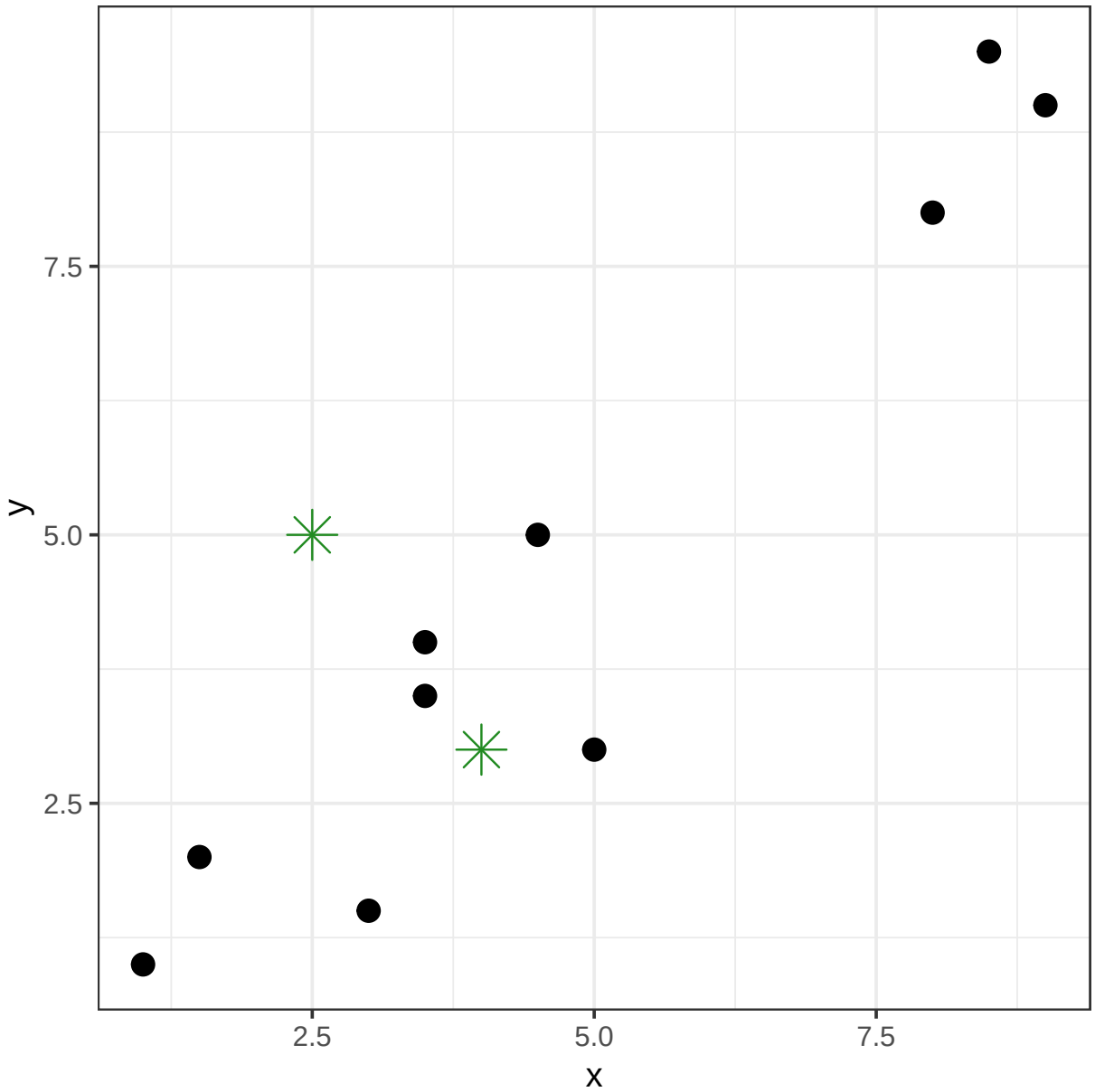
- $\mathbf{m}_k$ is the mean vector of values in cluster k - the kth **centroid**.
- $\|\cdot\|^2$ is the Euclidean distance (measure of dissimilarity).
- $|C_k|$ is the size of cluster k (i.e. how many observations it includes).

The K-Means Algorithm - Pseudocode

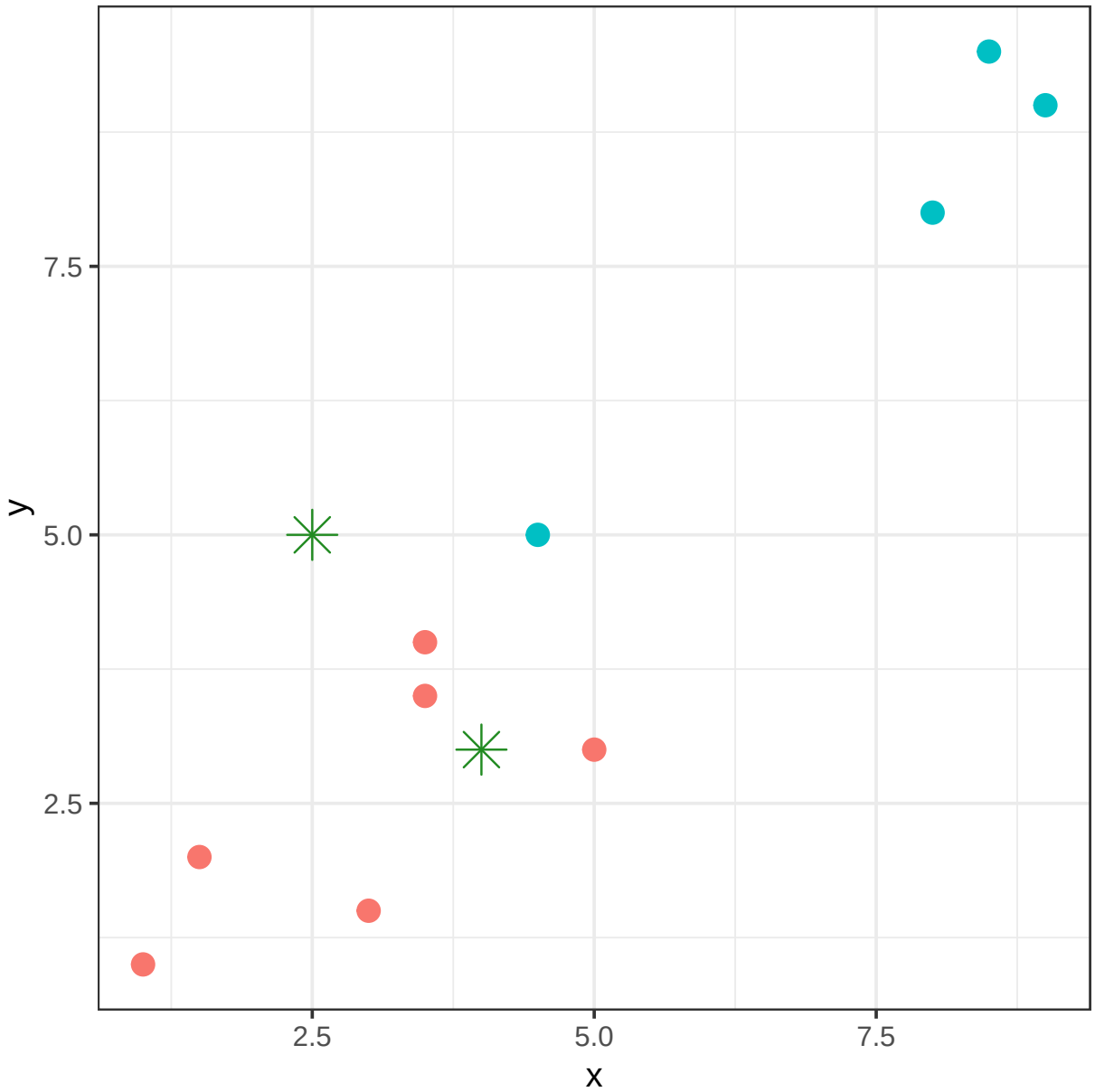
Input: Data matrix $\mathbf{X}$, number of clusters K, number of iterations $t^{\max}$.

1. Select K random points in $\mathbb{R}^p$ $\mathbf{m}_1^0, \ldots, \mathbf{m}_K^0$.
2. Compute pairwise distances between observations and centroids $\mathbf{m}_1^0, \ldots, \mathbf{m}_K^0$.
3. Assign each observation to the cluster with the closest centroid.
4. Update cluster centroids by re-computing the mean vectors.
5. Repeat Steps 3 - 4 until cluster assignments remain unchanged or max number of iterations $t^{\max}$ is reached.
6. Repeat all the above steps several times and choose the solution for which the K-Means objective is minimised.

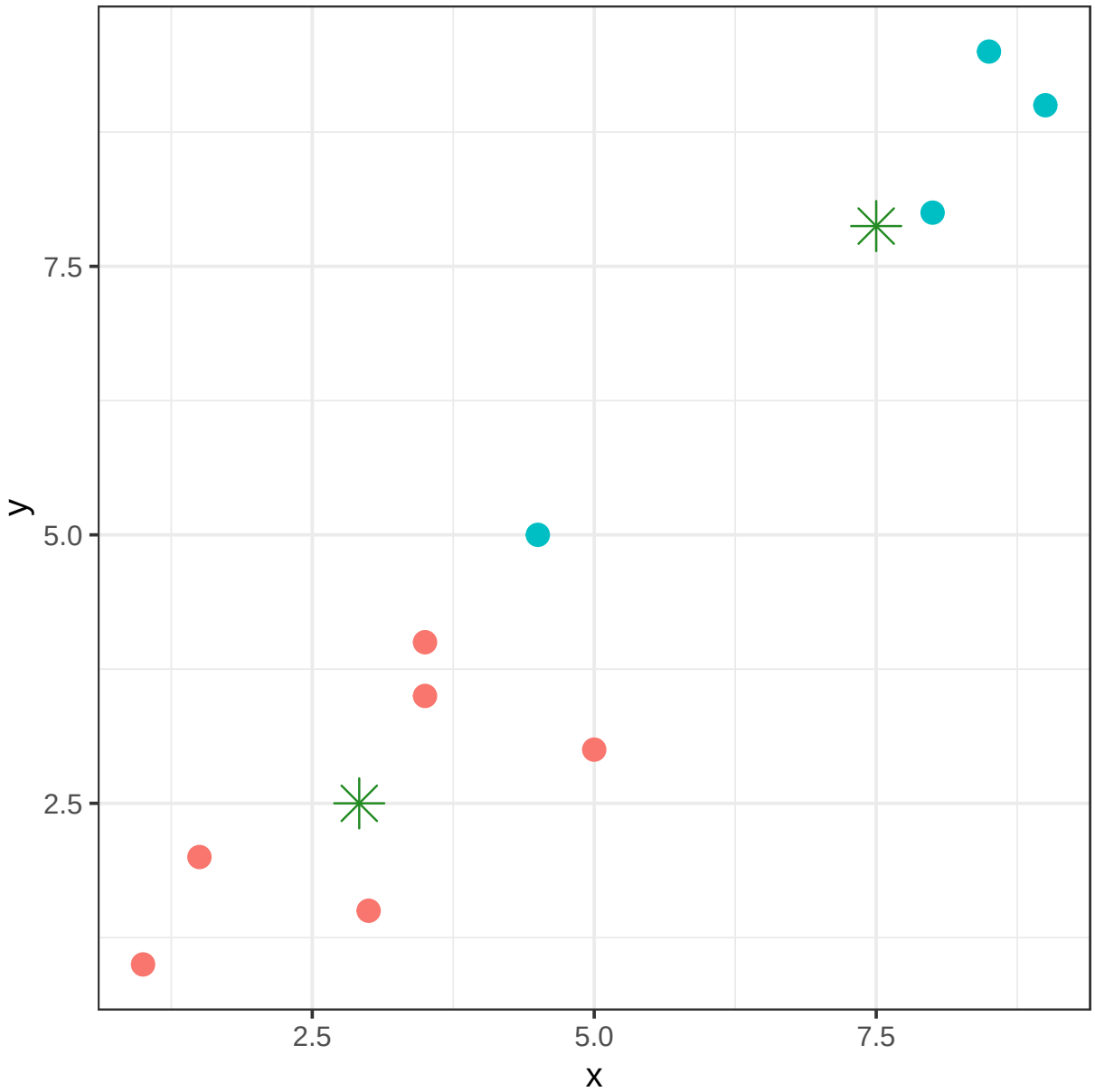
K-Means in practice - Step 0



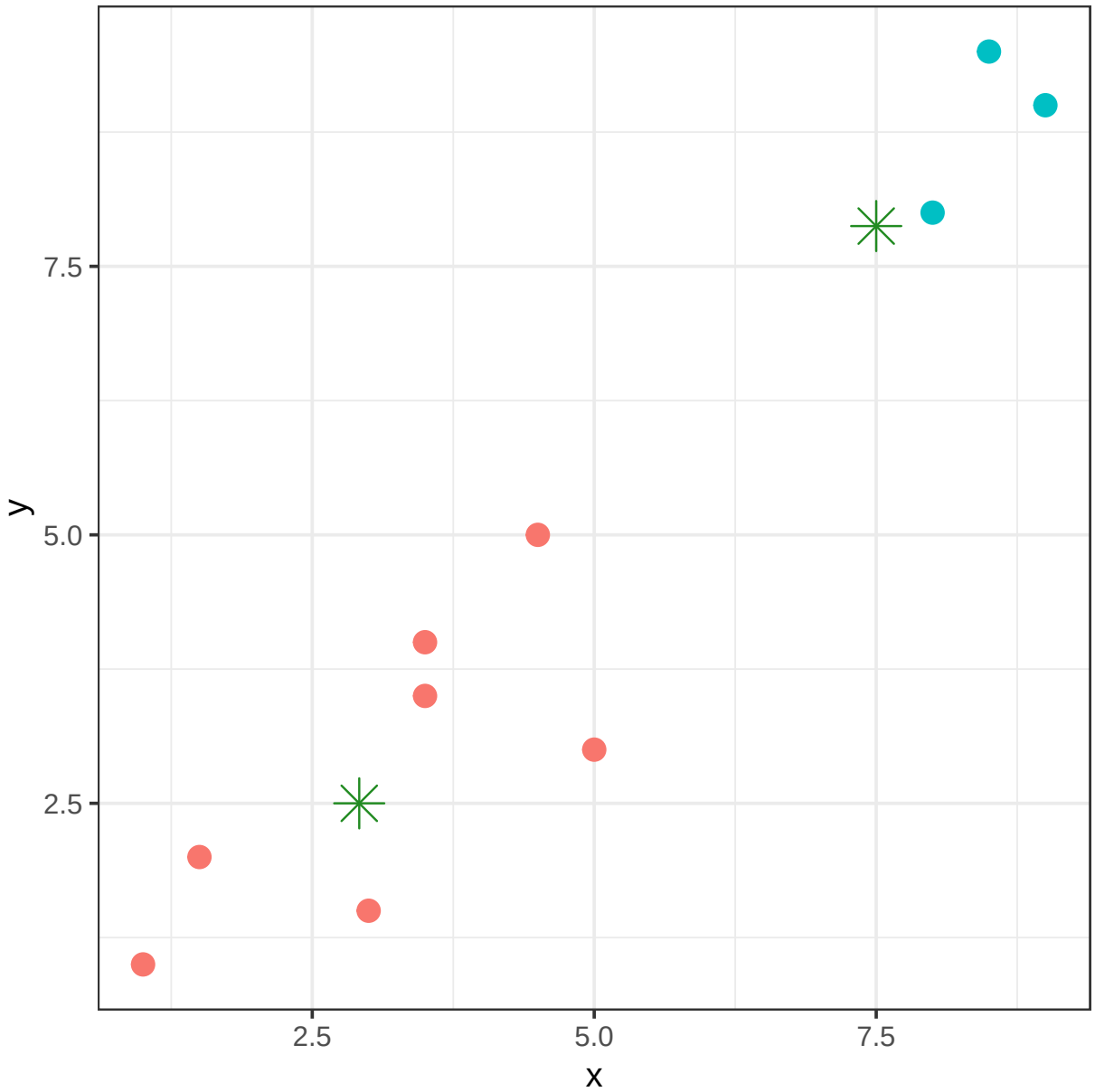
K-Means in practice - Step 1



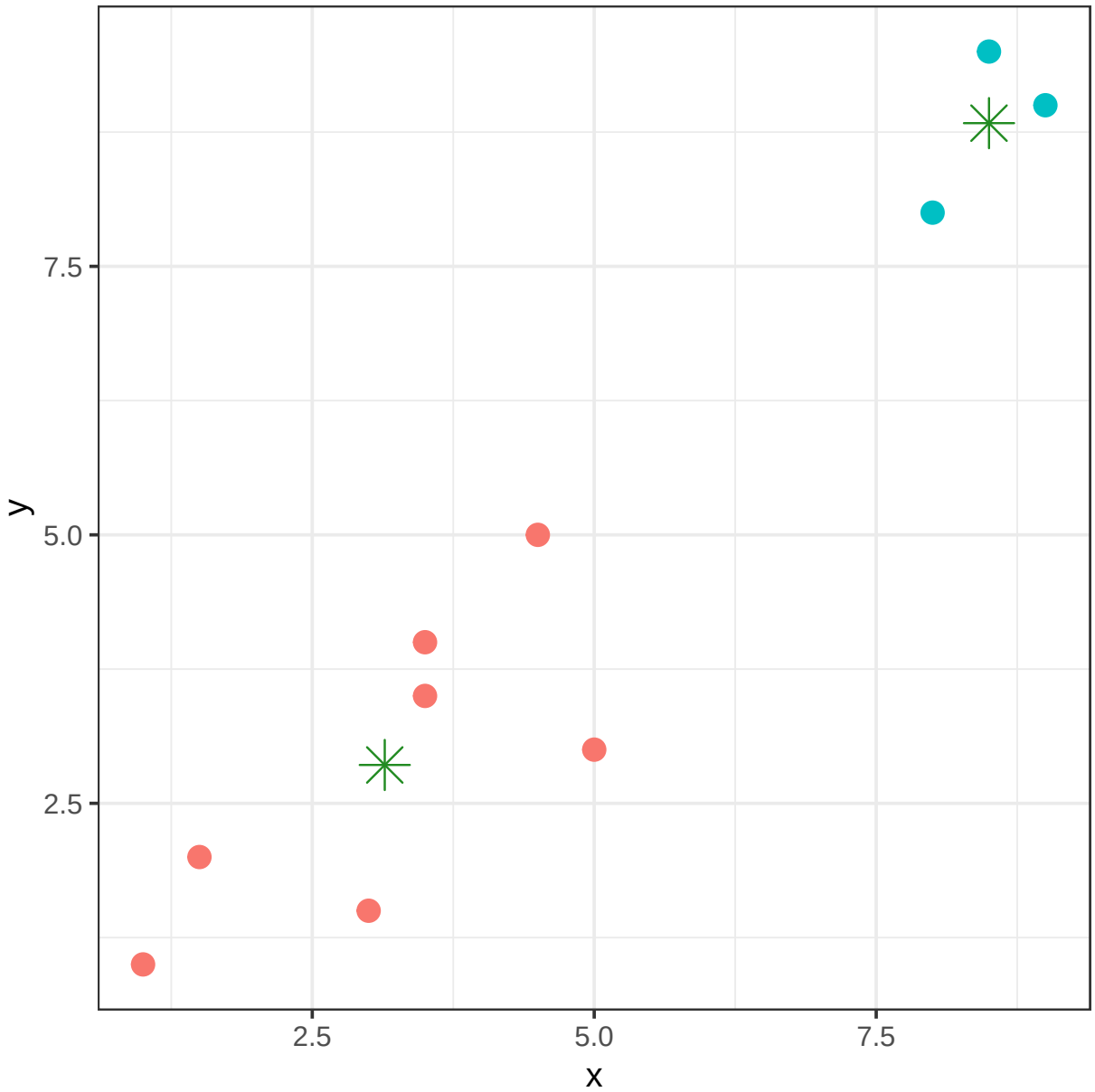
K-Means in practice - Step 2



K-Means in practice - Step 3



K-Means in practice - Step 4



Summary

- **Unsupervised learning:** No labels, no target output available for training.
- **Cluster analysis:** The task of finding groups (clusters) in data.
- **Clustering objective:** Similar observations are assigned in the same cluster, dissimilar items are placed in distinct clusters.
- ***K*-Means Algorithm:** Basic clustering algorithm that uses the Euclidean distance to define dissimilarities.

Useful Resources

- The Elements of Statistical Learning (Friedman, Tibshirani, and Hastie, 2001) - §14.3.
- `R` function `kmeans` provides a good implementation of K-Means.
- Code to reproduce plots in earlier slides can be found in: