Clustering Analysis of Sample Data

Roberto Pagliarini

Practical information

- Course materials:
  - https://github.com/RobertoPagliarini/Introduction-to-Bioinformatics/tree/main
- My email: roberto.pagliarini@uniud.it
- Theorical lessons + R exercises
  - Slides + scripts

What is Bioinformatics?

- Bioinformatics is an interdisciplinary field that uses computational tools and techniques to analyze and interpret large and complex biological data, such as DNA, RNA, and protein sequences.
- It combines biology with computer science, mathematics, and statistics to manage, store, analyze, and disseminate biological information, helping researchers understand biological systems at a molecular level.
- Key applications include genome sequencing analysis, predicting protein structures, identifying disease-associated genes, and designing new drugs, making it crucial for modern biology and medicine.

What is Clustering?

Unsupervised learning: no labels provided

Goal: group objects that are similar to each other

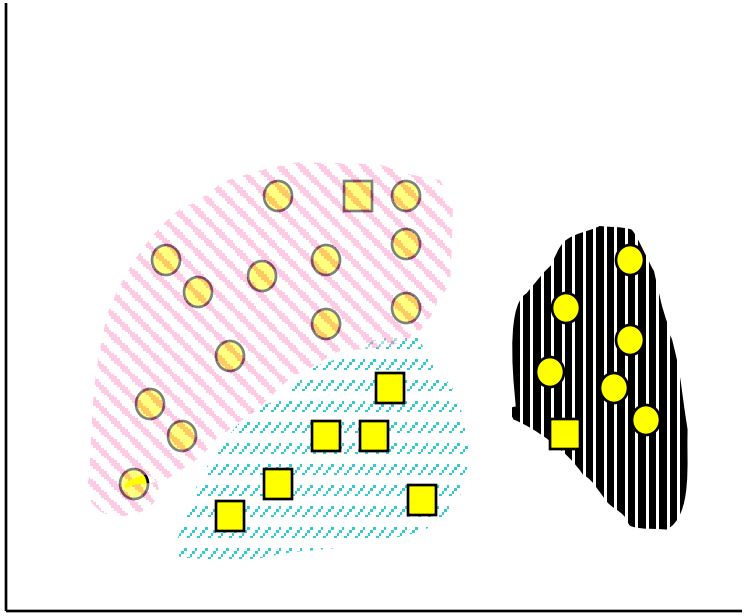
Widely used in:

- Genomics (gene expression, allele-specific expression)

- Market segmentation

- Image recognition

Clustering: a very simple example



- Find “natural” groupings in unlabeled data

Why is Clustering Important?

Reduces complexity of large datasets

Reveals hidden structures

Helps generate hypotheses

Provides insight for downstream analyses (classification, prediction)

Similarity - distance

- Distance d(x,y)
  - Measures the “dissimilarity between objects”
  - Similarity s(x,y)
  - S(x,y) ≈1/d(x,y)

- Properties

$$d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

$$d(\mathbf{x}_i, \mathbf{x}_i) = 0$$

$$d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$$

$$d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_k, \mathbf{x}_j)$$

Example in Bioinformatics

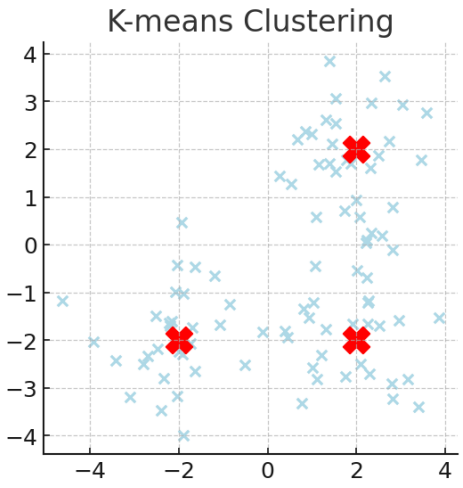
- Clustering gene expression across samples
- Identifying co-regulated genes
- Grouping patients based on molecular profiles
- Supports precision medicine

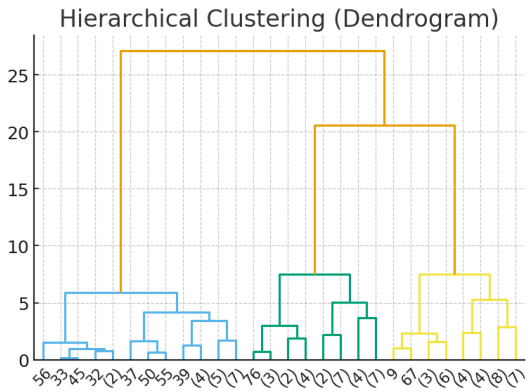
Key Challenges

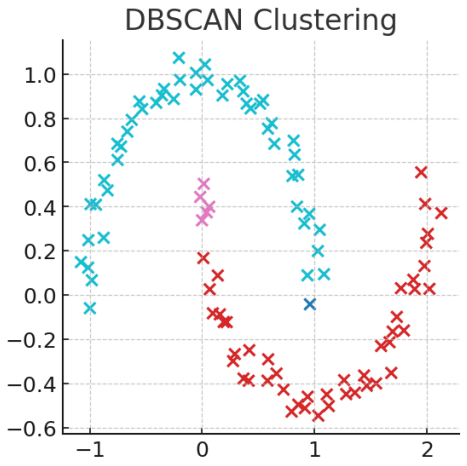
- Choosing the right number of clusters
- What variables matter most?
- Scalability: large datasets (e.g. RNA-seq)
- Interpretability: biological meaning of clusters

Common Methods

- **K-means: partitions data into k groups**
- Hierarchical clustering: tree-like structure
- Density-based (DBSCAN): detects clusters of arbitrary shape
- Model-based: assumes statistical distributions







Evaluation Metrics

Internal metrics: Silhouette score, Dunn index

External metrics: Adjusted Rand index (if labels exist)

Biological validation: overlap with known pathways/genes

Summary

- Clustering = grouping data without labels
- Essential tool for exploring high-dimensional data
- Useful in biology to find gene groups or patient subtypes
- Challenges: choosing methods, validating results

K-means Clustering Algorithm

What is K-Means Clustering

It is an unsupervised **machine learning algorithm** used for partitioning a dataset into a pre-defined number of clusters.

The goal is to group similar data points together and discover underlying patterns or structures within the data.

The aim is to minimize the distance between the points within a cluster.

K-means is a centroid-based algorithm or a distance-based algorithm

In K-Means, each cluster is associated with a centroid.

K-means Algorithm

- Algorithm *K-Means*(*D, k*)
  - $m \leftarrow D.size$ // number of instance
  - FOR $i \leftarrow 1$ TO k DO
    - $\mu_i \leftarrow$ random // select a random point
  - WHILE (termination condition)
    - FOR $j \leftarrow 1$ TO m DO // computing cluster membership
    - $h \leftarrow \text{argmin}_{1 \le i \le k} \text{dist}(x_j, \mu_i)$

      $\boldsymbol{C[h] \leftarrow x_j}$
    - FOR $i \leftarrow 1$ TO k DO

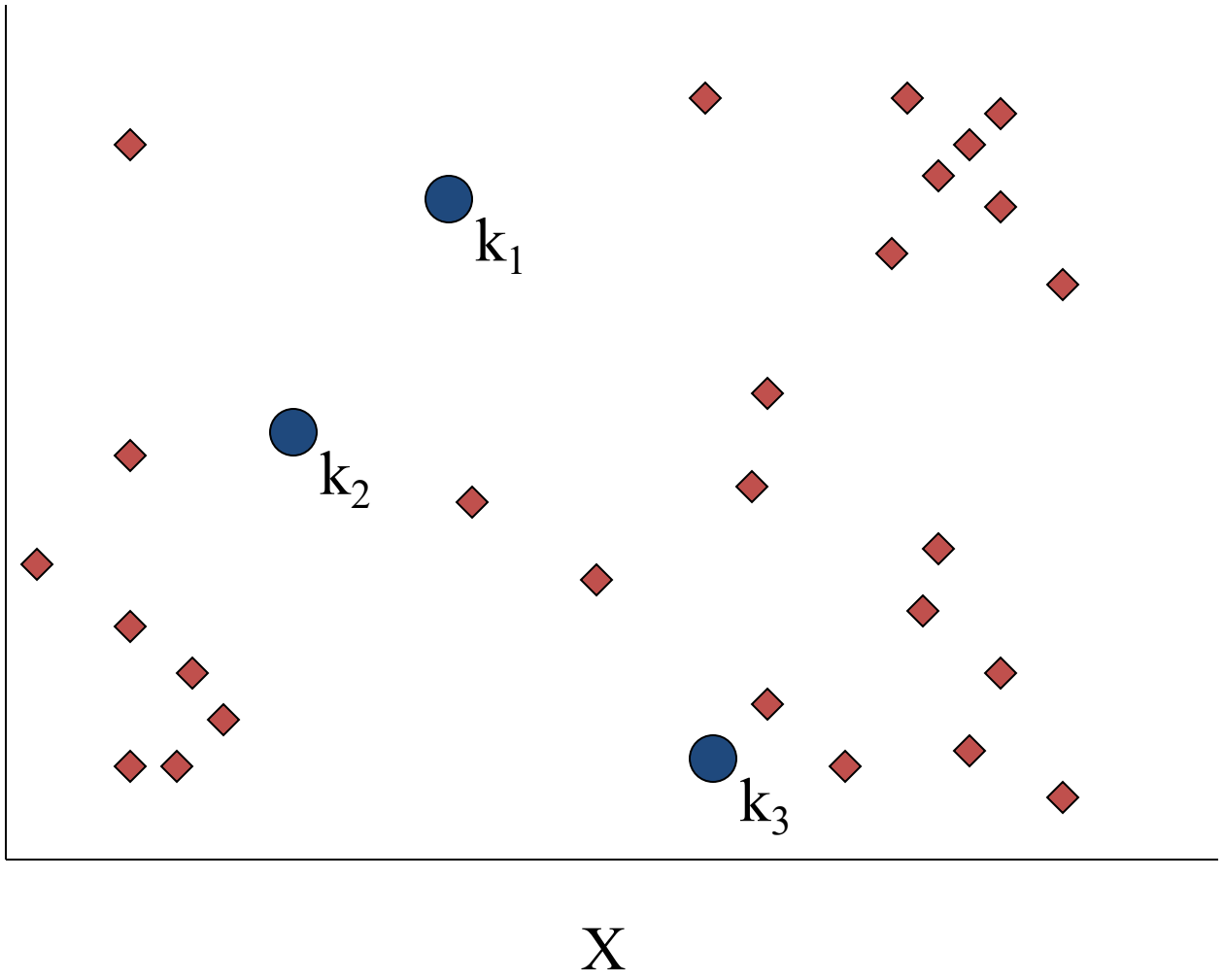
      – $\mu_i \leftarrow \frac{1}{n_i} \sum_{x_j \in C[i]} x_j$
  - RETURN *Make-Predictor* (*w*, *P*)

How K-Means Clustering Works?

1. **Initialization**: start by randomly selecting K points from the dataset. These points will act as the initial cluster centroids.
2. **Assignment**: for each data point in the dataset, calculate the distance between that point and each of the K centroids. Assign the data point to the cluster whose centroid is closest to it. This step effectively forms K clusters.
3. **Update centroids**: once all data points have been assigned to clusters, recalculate the centroids of the clusters by taking the mean of all data points assigned to each cluster.
4. **Repeat**: repeat steps 2 and 3 until convergence. Convergence occurs when the centroids no longer change significantly or when a specified number of iterations is reached.
5. **Final Result**: once convergence is achieved, the algorithm outputs the final cluster centroids and the assignment of each data point to a cluster.

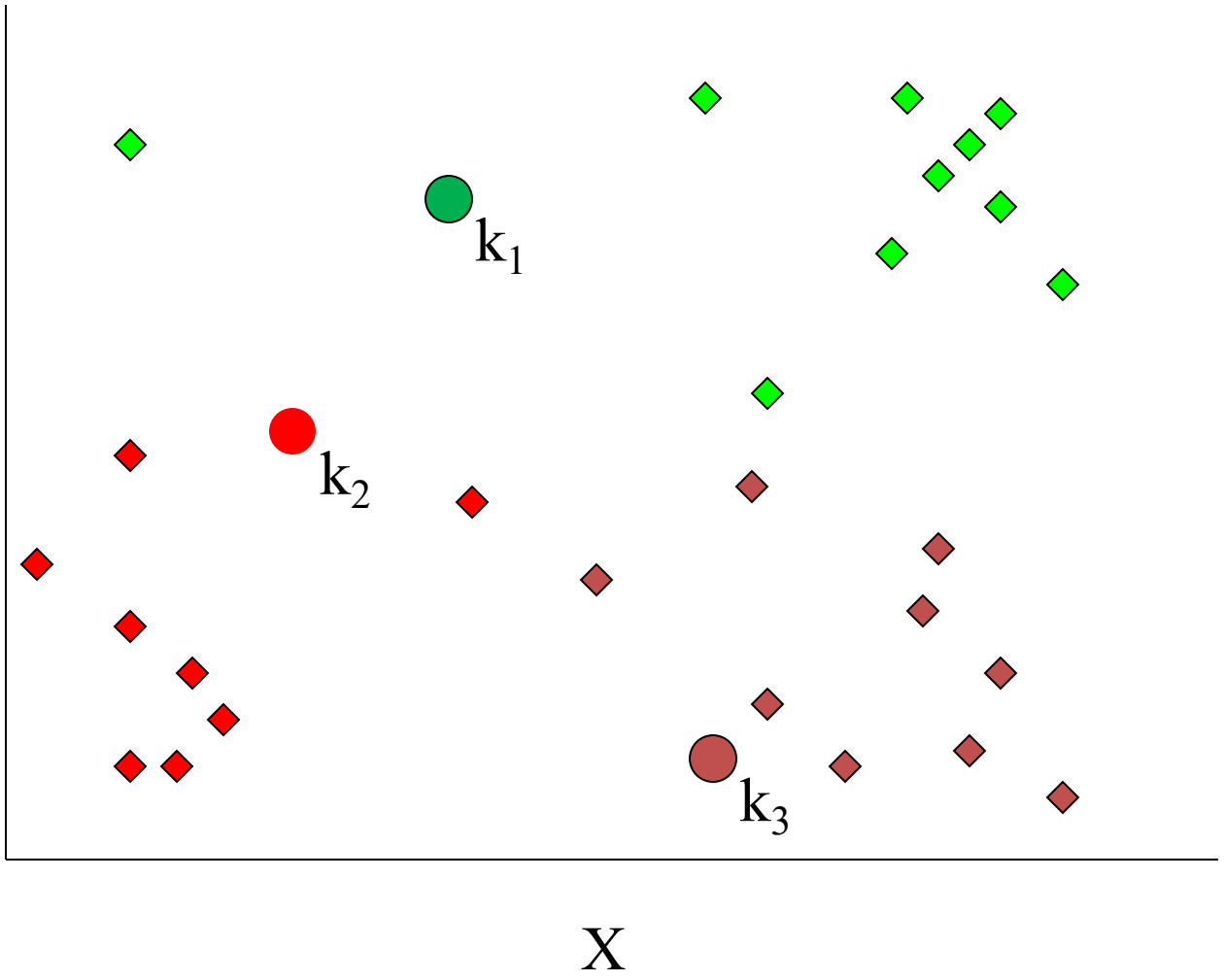
K-means example [1]

Selecting
3 random
centroids



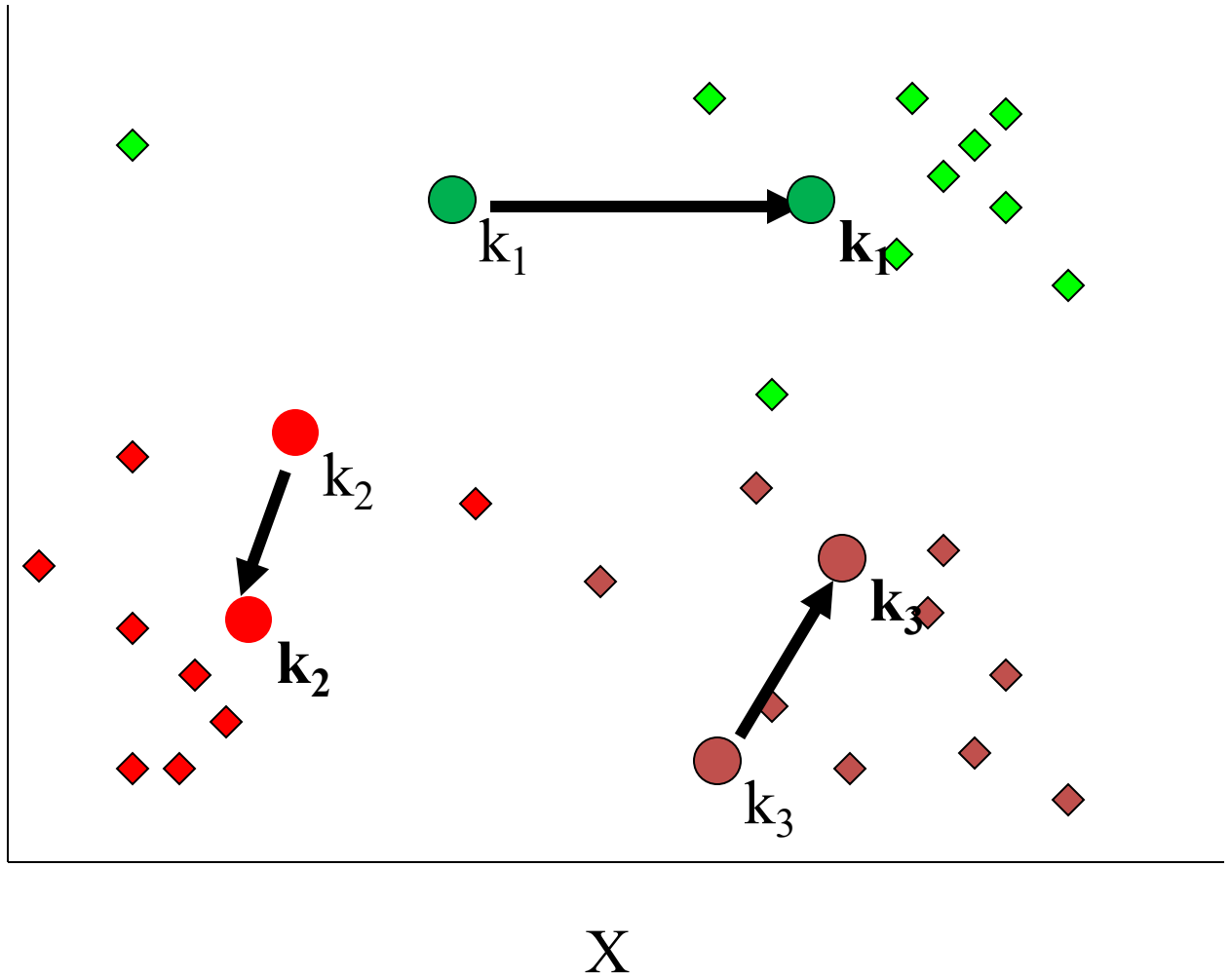
K-means example [2]

We assign each point to the closest cluster



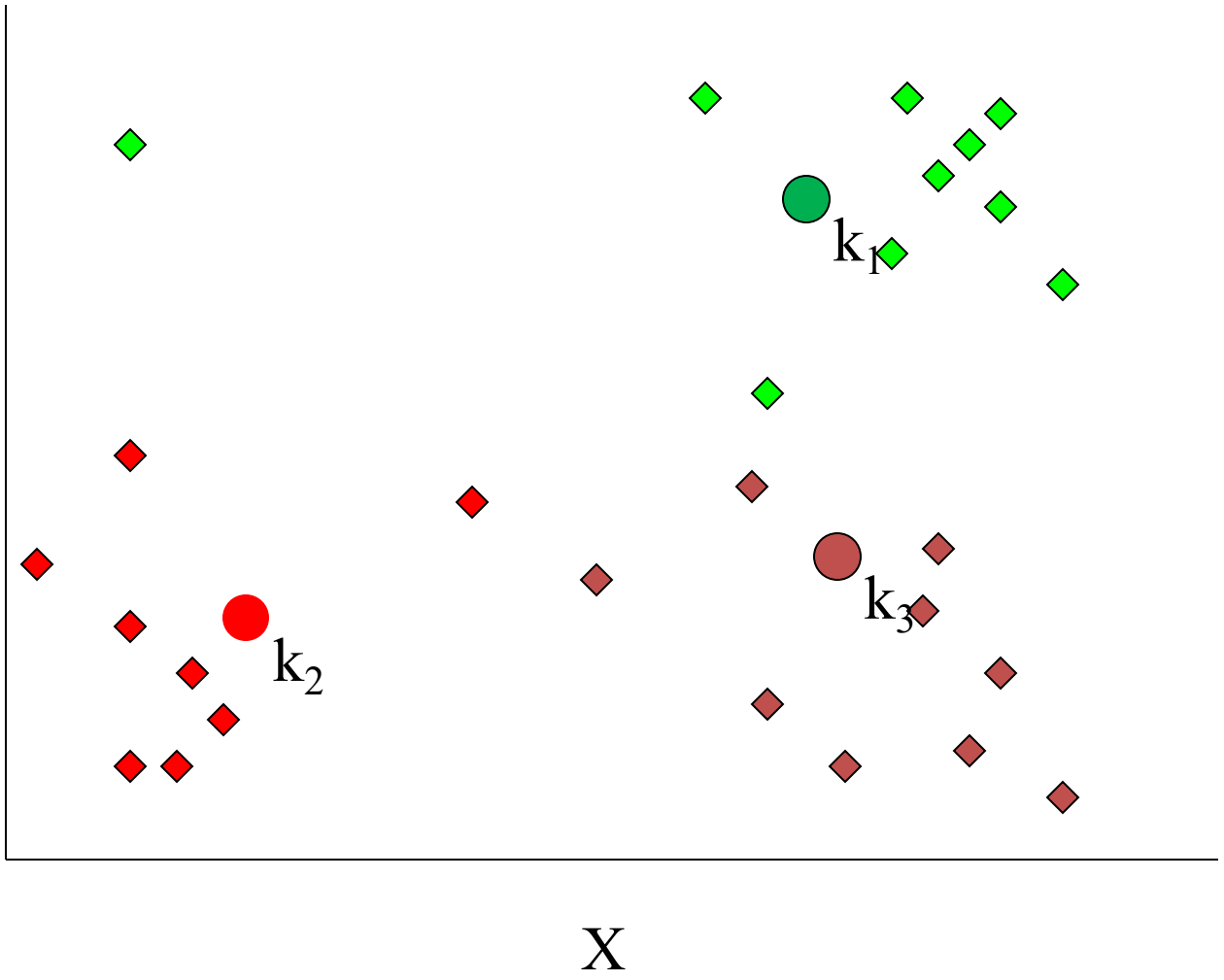
K-means example [3]

Computing the new centroids

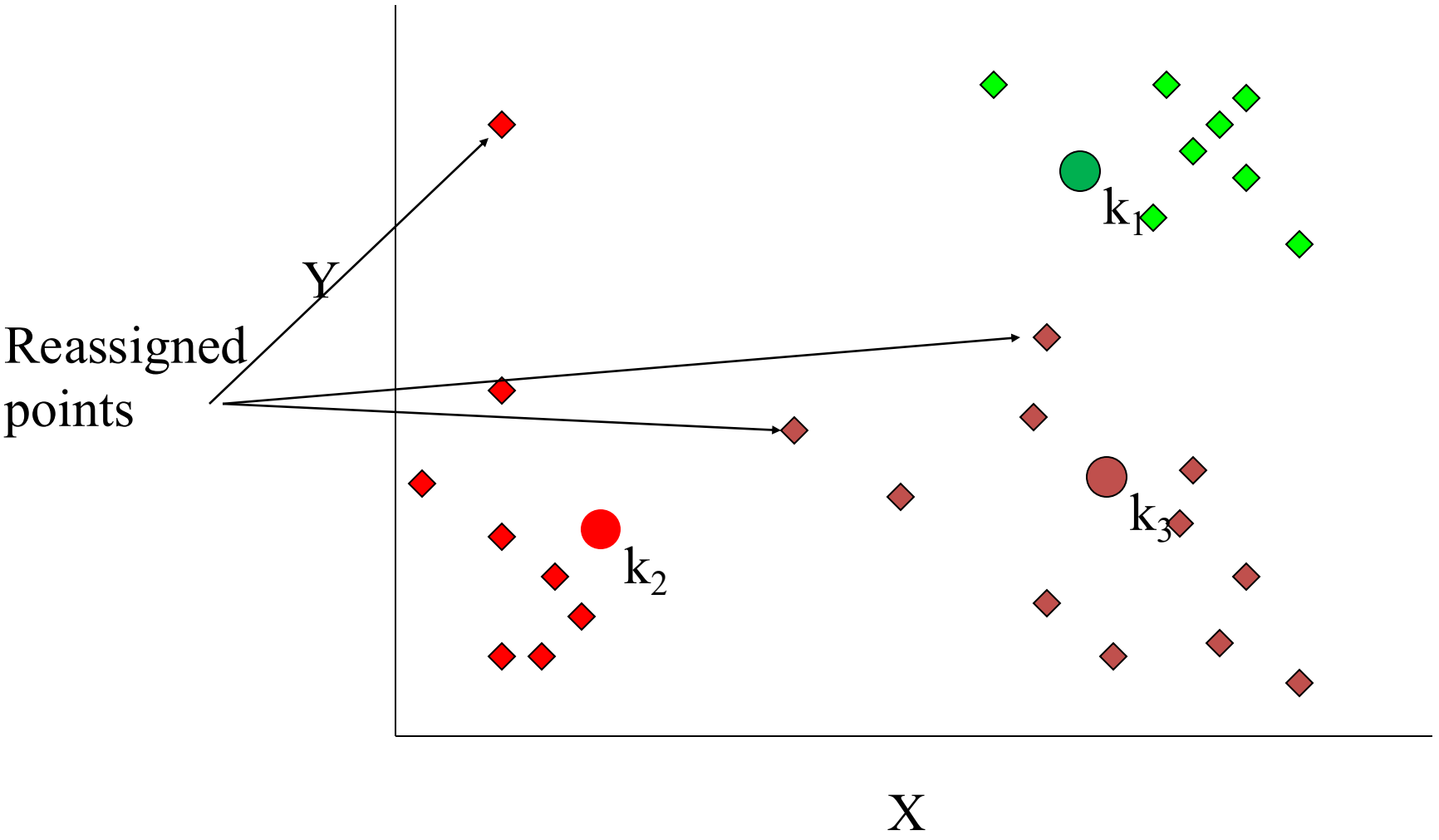


K-means example [4]

Reassign
the points to the
clusters

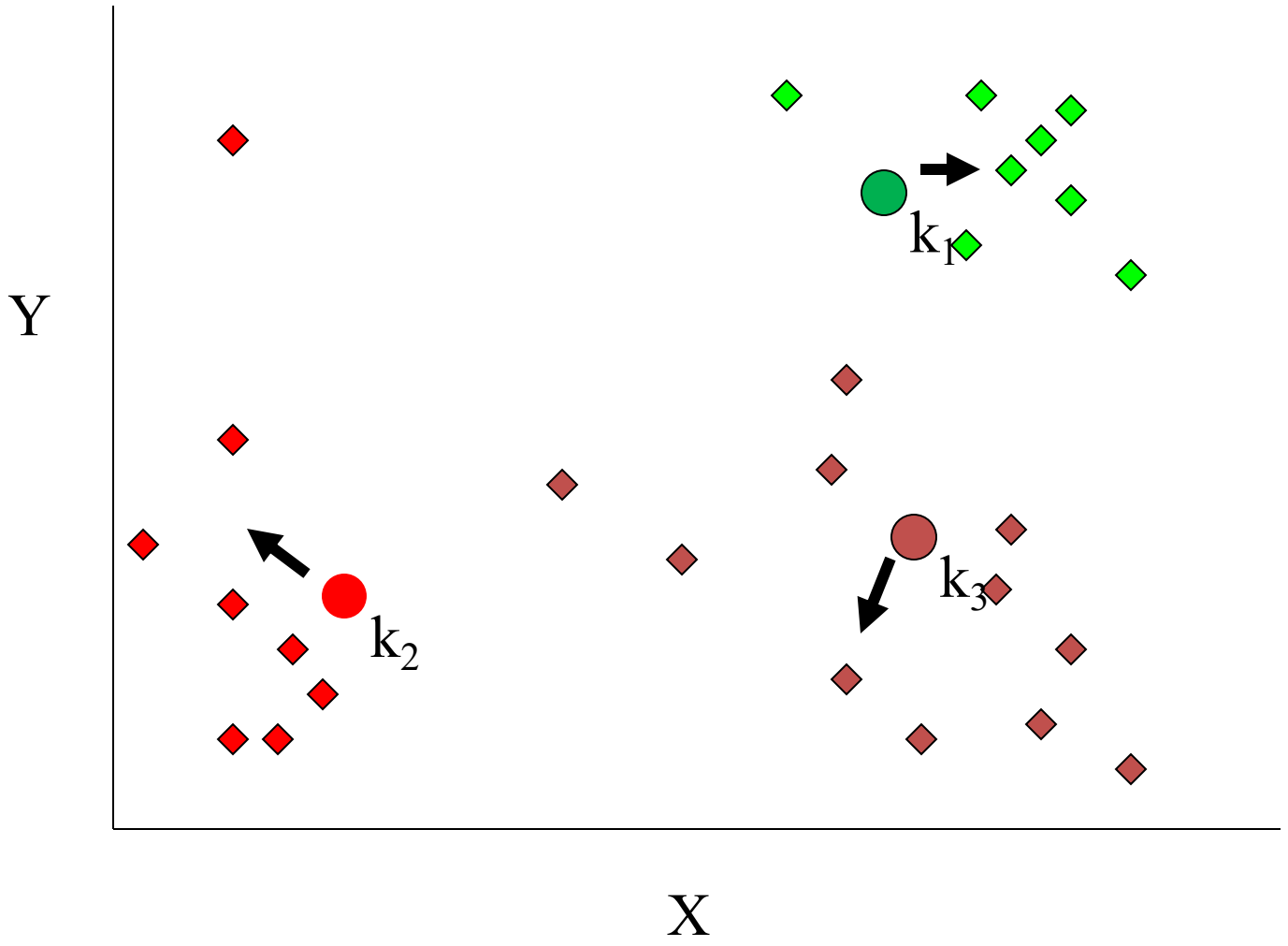


K-means example [5]

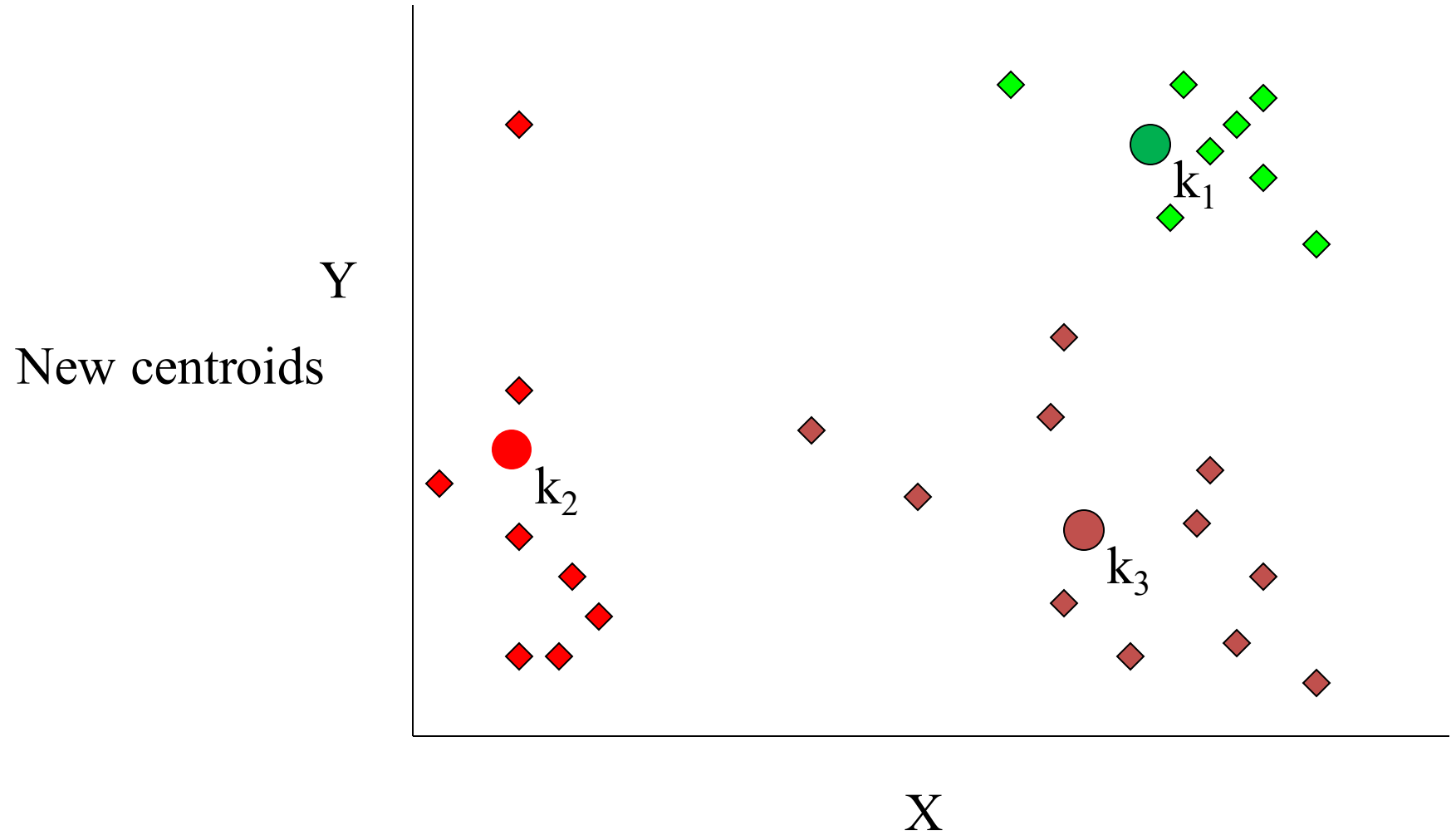


K-means example [6]

Computing new centroids



K-means example [7]



Questions

Will different initialization lead to different results?

- Yes
- No
- Sometimes

Will the algorithm always stop after some iterations?

- Yes
- No (We have to set a maximum number of iterations)
- Sometime

K-means clustering in R

We will use the pasilla data from Bioconductor.

The Pasilla Dataset

This dataset is available from the Pasilla Bioconductor library and is derived from the work from Brooks et al. (Conservation of an RNA regulatory map between Drosophila and mammals. Genome Research, 2010).

Alternative splicing is generally controlled by proteins that bind directly to regulatory sequence elements and either activate or repress splicing of adjacent splice sites in a target pre-mRNA. Here, the authors have combined RNAi and mRNA-seq to identify exons that are regulated by Pasilla (PS), the Drosophila melanogaster ortholog of the mammalian RNA-binding proteins NOVA1 and NOVA2.