Clustering

Thanh-Nghi Do

Can Tho University
dtnghi@cit.ctu.edu.vn

Can Tho
Dec. 2019

Content

- Introduction
- Hierarchical clustering
- *k*-means

Content

- **Introduction**
- Hierarchical clustering
- *k*-means

Top 10 Data Mining Algorithms (Kdnuggets)

Here are the algorithms:

- 1. C4.5
- 2. k-means
- 3. Support vector machines
- 4. Apriori
- 5. EM
- 6. PageRank
- 7. AdaBoost
- 8. kNN
- 9. Naive Bayes
- 10. CART

Introduction

- Unsupervised learning (clustering)
  - Finding "natural" grouping of examples given un-labeled dataset
  - Clustering: dividing examples into a number of groups such that examples in the same groups are more similar to other examples in the same group and dissimilar to examples in other groups
  - Approaches: hierarchical methods, partitioning methods, density-based methods
  - Most popular: *k*-means, Dendrogram, SOM, EM

Introduction

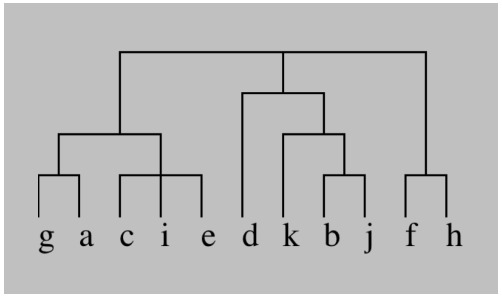
- Unsupervised learning (clustering)
  - Maximize the intra-cluster similarity and minimize the inter-cluster similarity
  - Similarity measure: distance
  - Distance: data types (binary, nominal, numerical)

Content

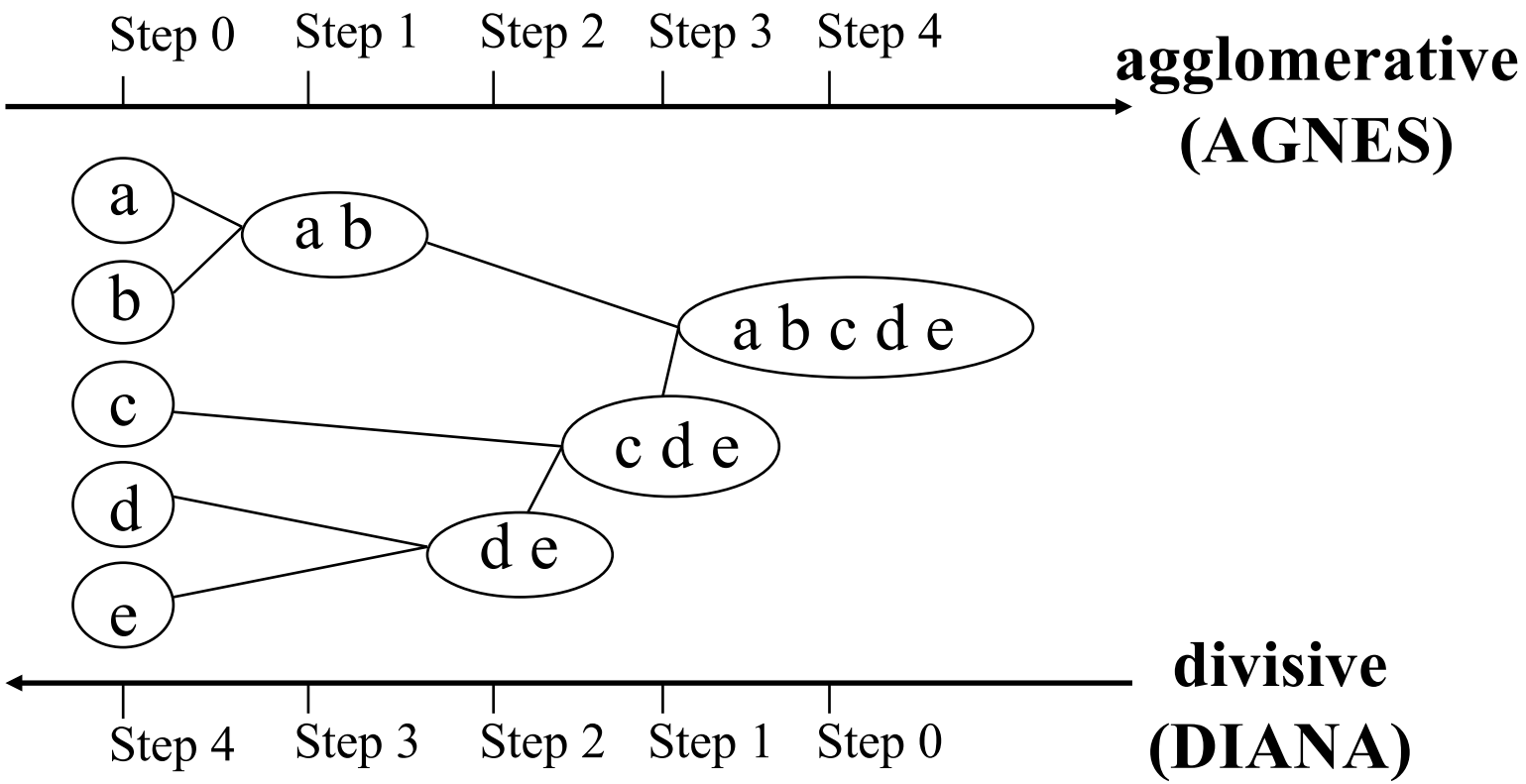
- Introduction
- **Hierarchical clustering**
- *k*-means

Hierarchical clustering

- Bottom up
  - Start with single-example clusters
  - At each step, join the two closest clusters
  - Design decision: distance between clusters (e.g. two closest examples in clusters, distance between means)
- Top down
  - Start with one universal cluster
  - Find two clusters
  - Proceed recursively on each subset



Hierarchical clustering



Hierarchical clustering

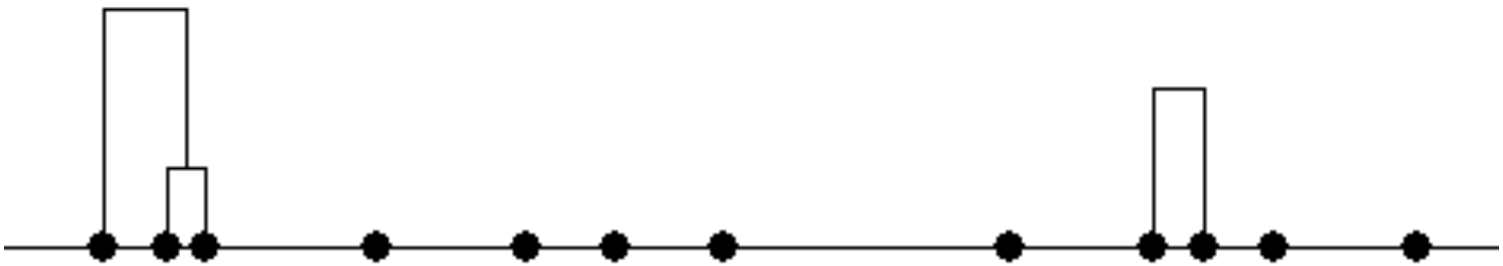
- Distance between two clusters
  - Single linkage: shortest distance between two examples in each cluster
  - Complete linkage: longest distance between two examples in each cluster
  - Average linkage: average distance between each example in one cluster to every example in the other cluster

Hierarchical clustering (single linkage)

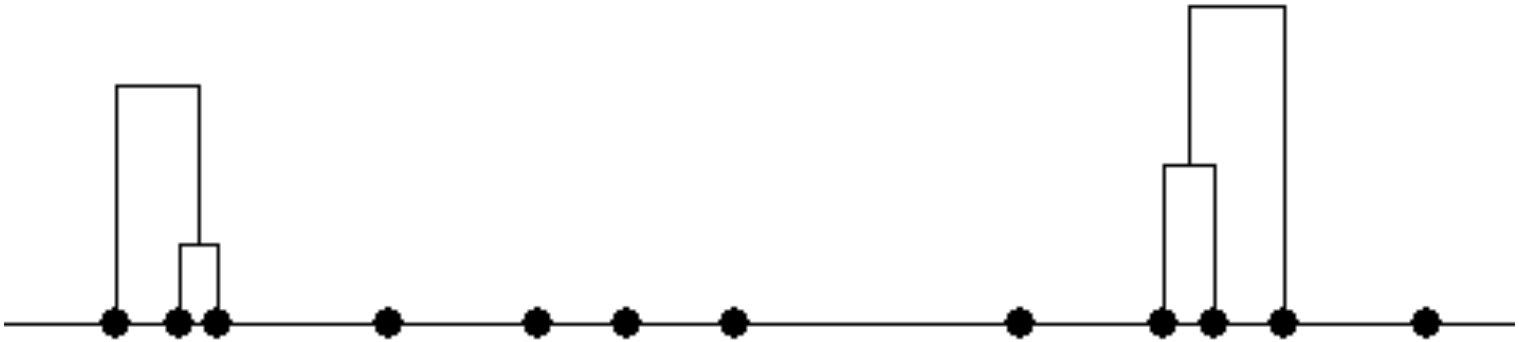
Hierarchical clustering (single linkage)

Hierarchical clustering (single linkage)

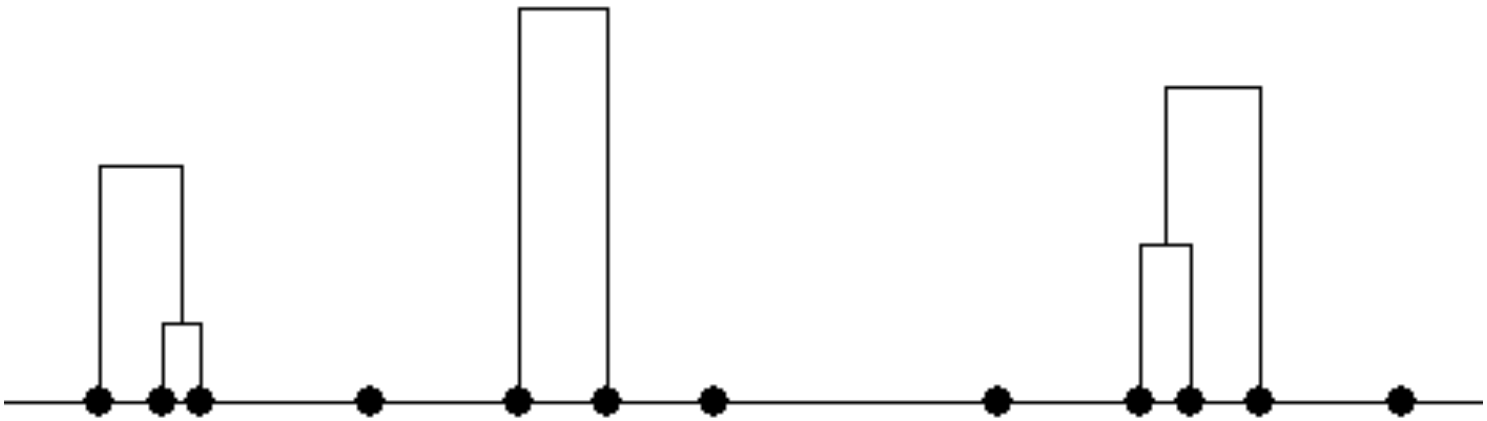
Hierarchical clustering (single linkage)



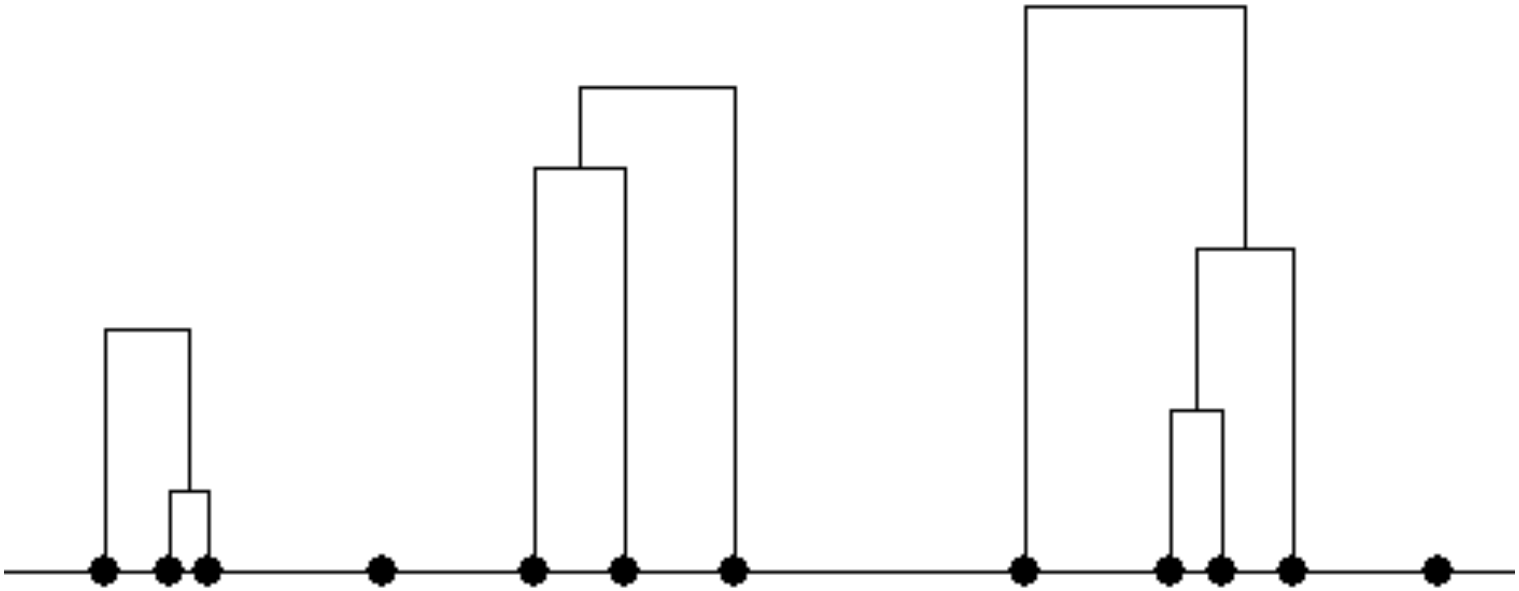
Hierarchical clustering (single linkage)



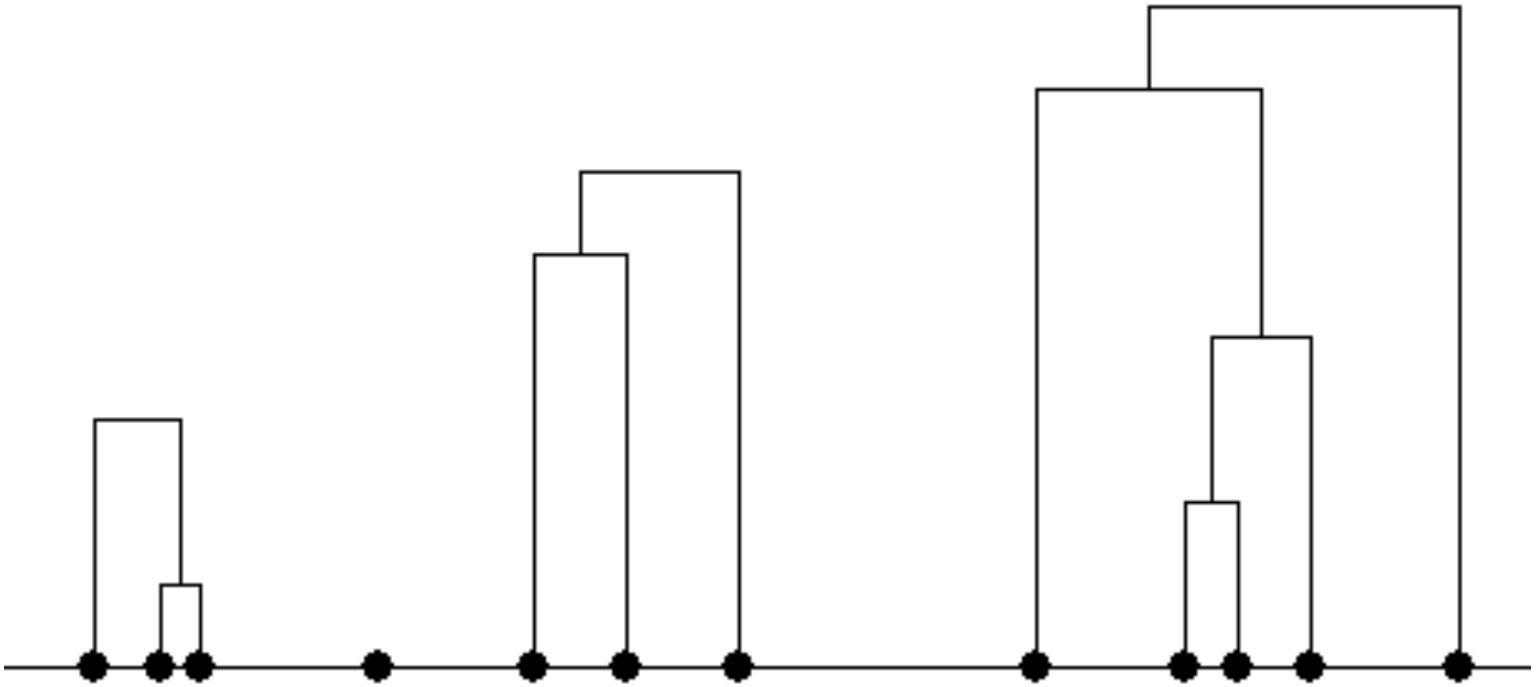
Hierarchical clustering (single linkage)



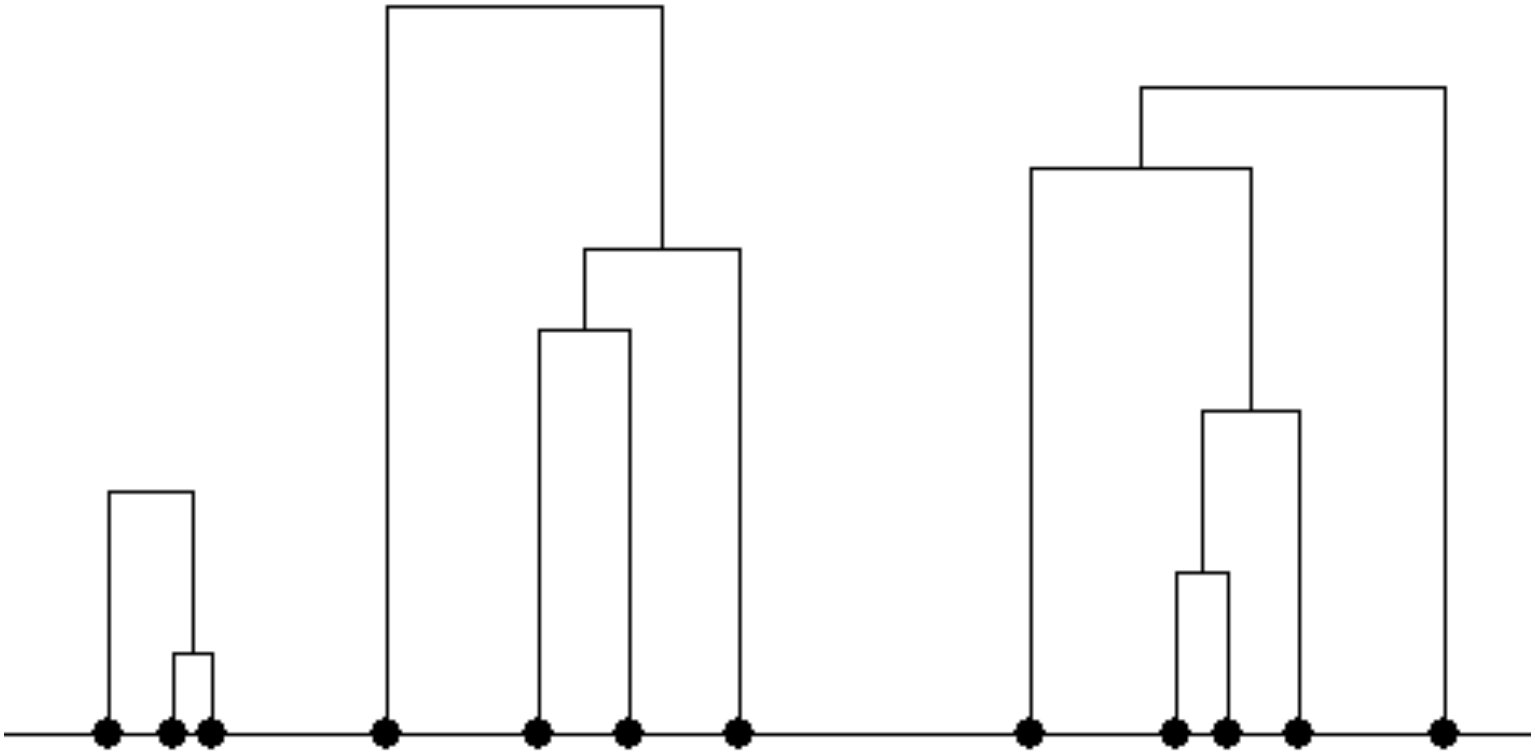
Hierarchical clustering (single linkage)



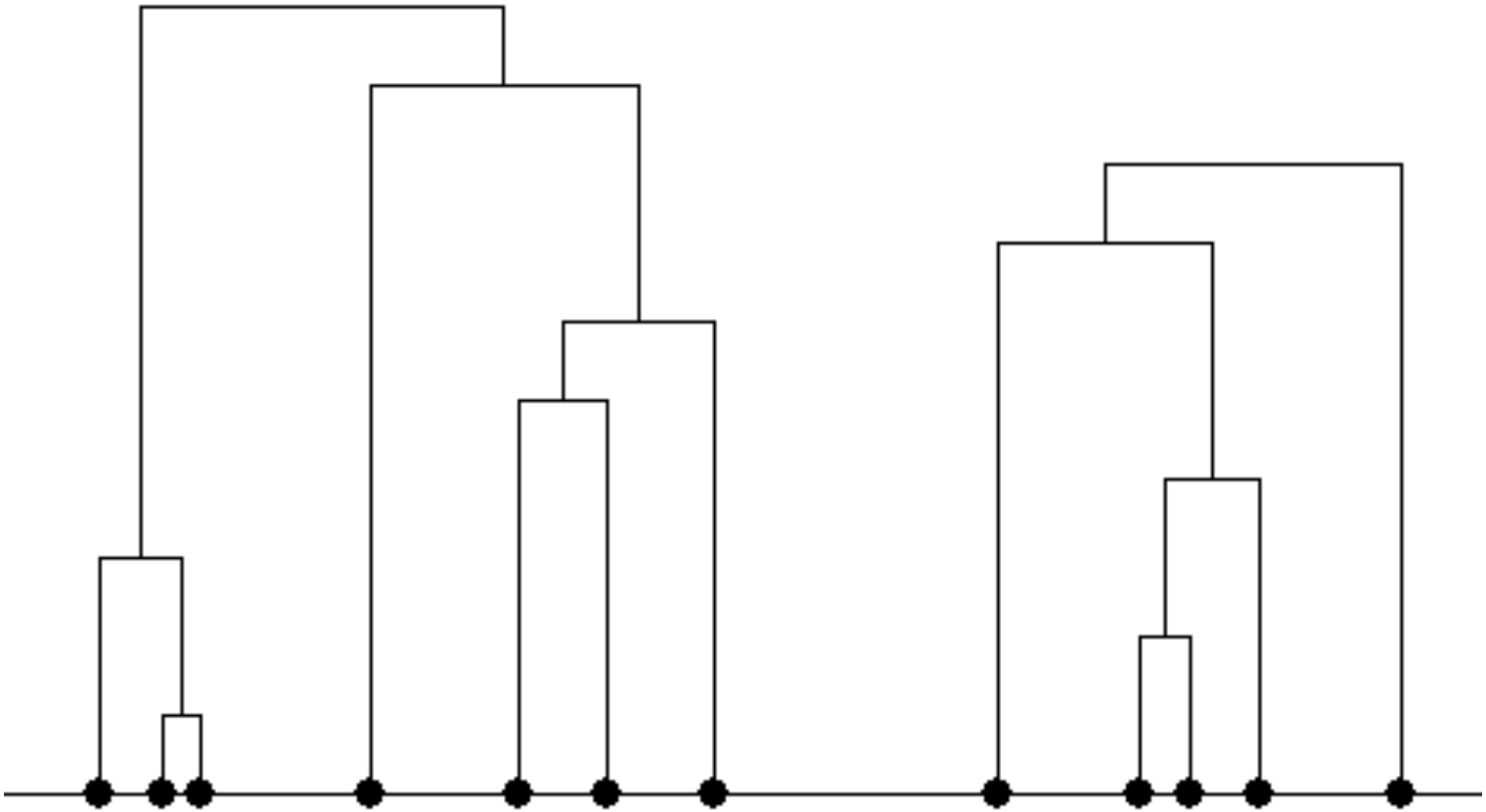
Hierarchical clustering (single linkage)



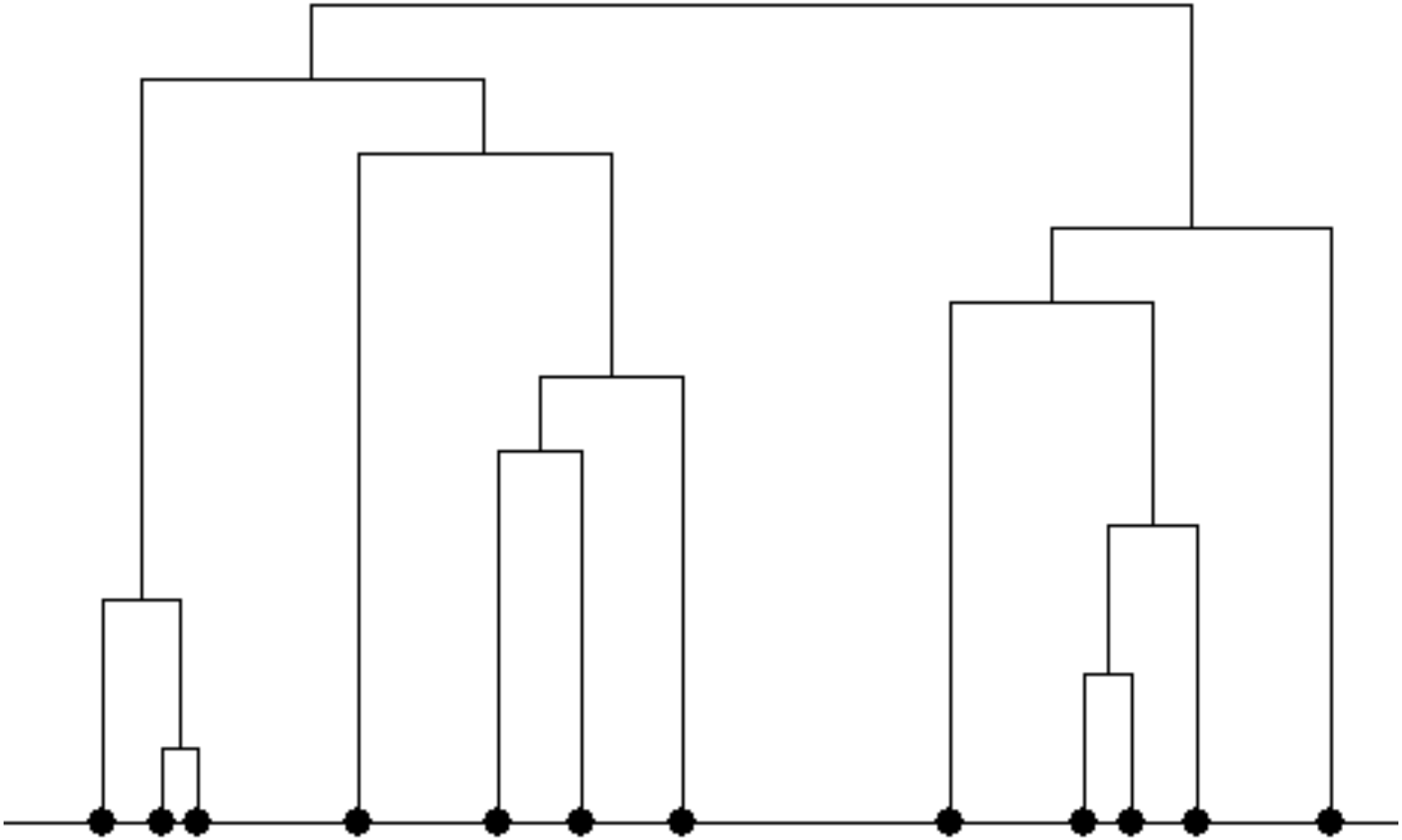
Hierarchical clustering (single linkage)



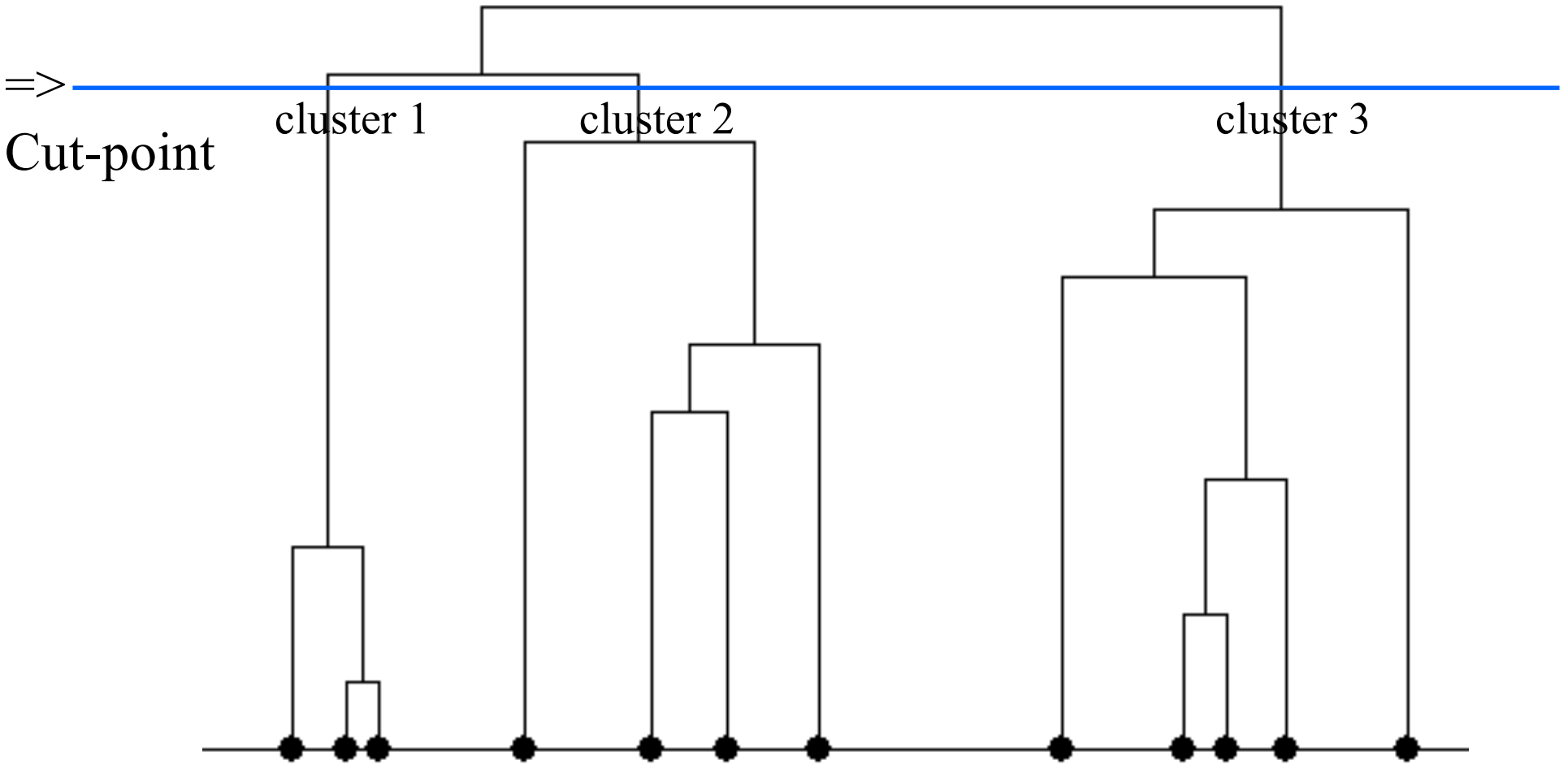
Hierarchical clustering (single linkage)



Hierarchical clustering (single linkage)



Hierarchical clustering (single linkage)



Hierarchical clustering (single linkage)

- Comments
  - Simple, comprehensive, non-parametric
  - Resultat: dendrogramme
  - Algorithmic complexity: expensive

Content

- Introduction
- Hierarchical clustering
- ***k*-means**

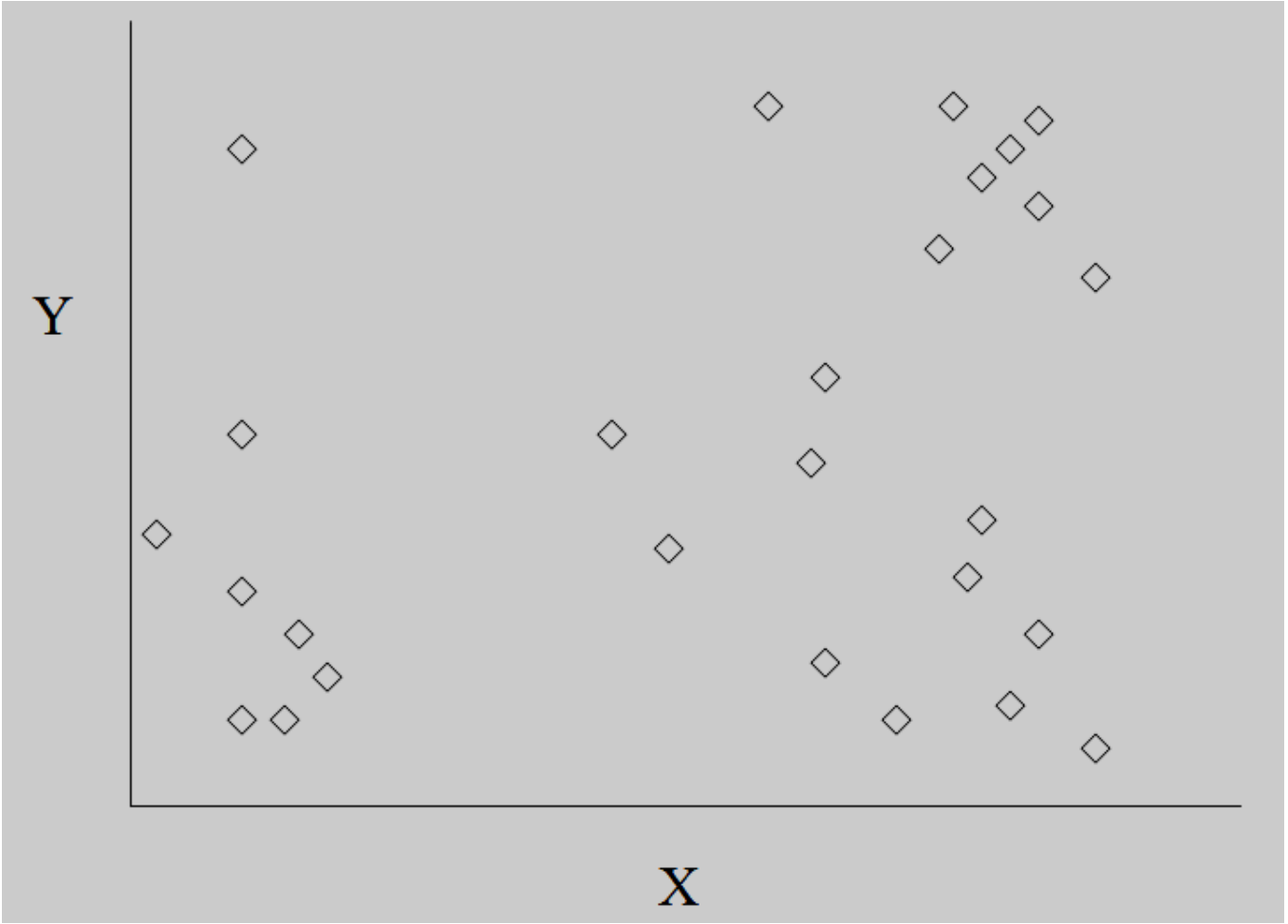
k-means

- Principle
  - Partition examples into *k* clusters
  - Cluster center: mean of examples in the cluster
  - Assigning the example to its nearest cluster center

k-means

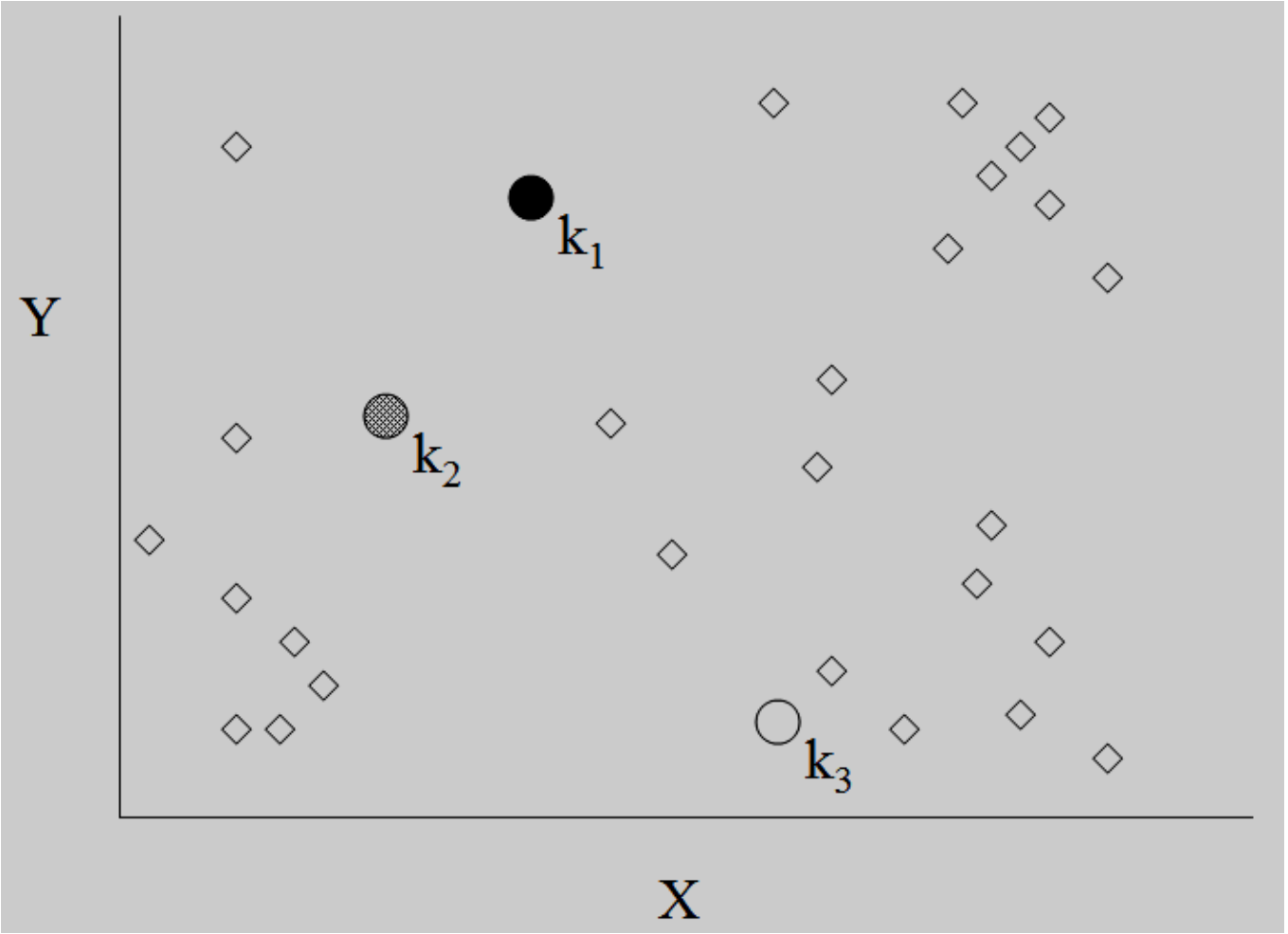
- Training algorithm
  1. Pick a number *k* of cluster centers (at random)
  2. Assign every example to its nearest cluster center (e.g. using Euclidean distance)
  3. Move each cluster center to the mean of its assigned examples
  4. Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

k-means



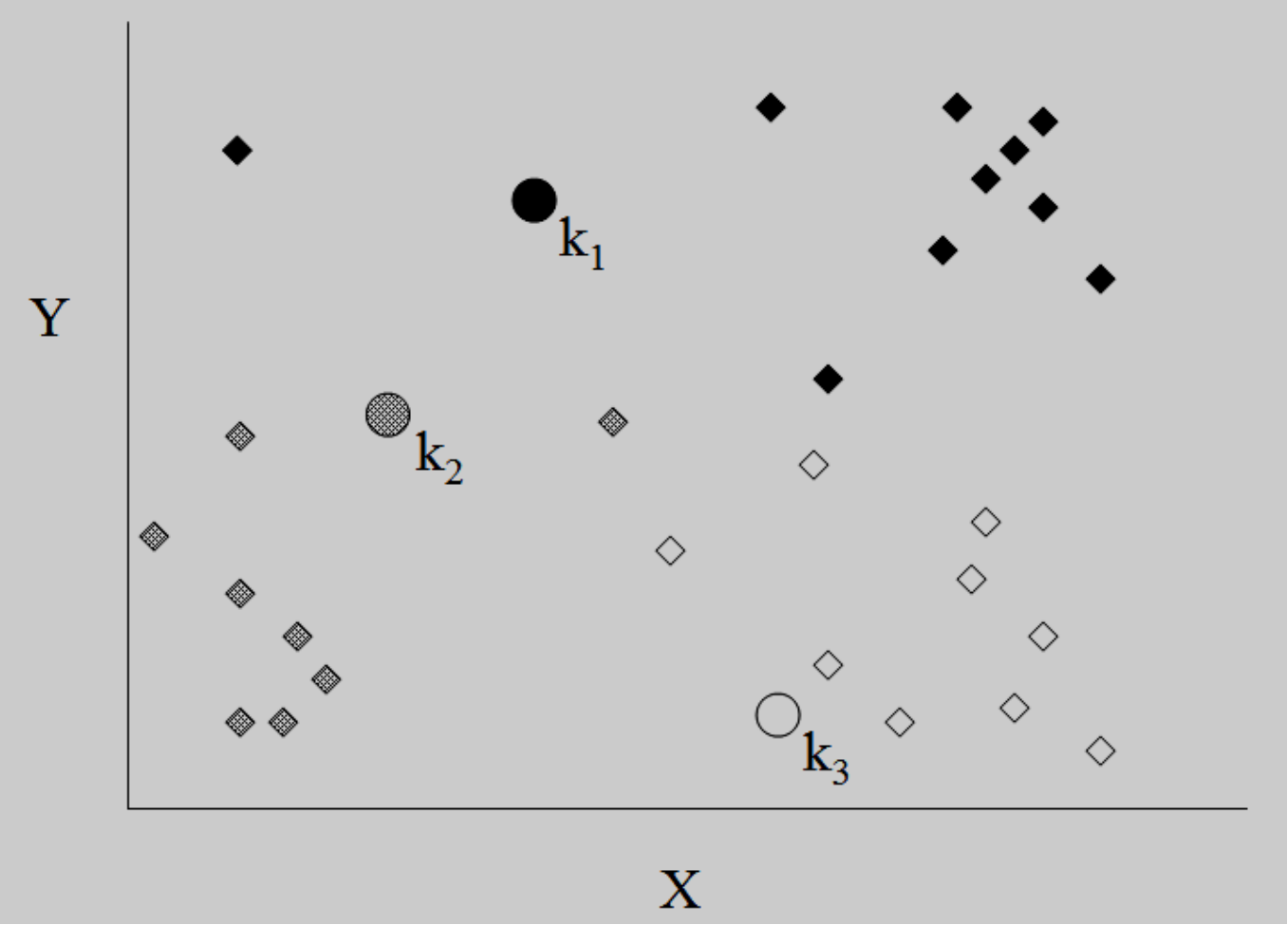
k-means

3 random cluster centers



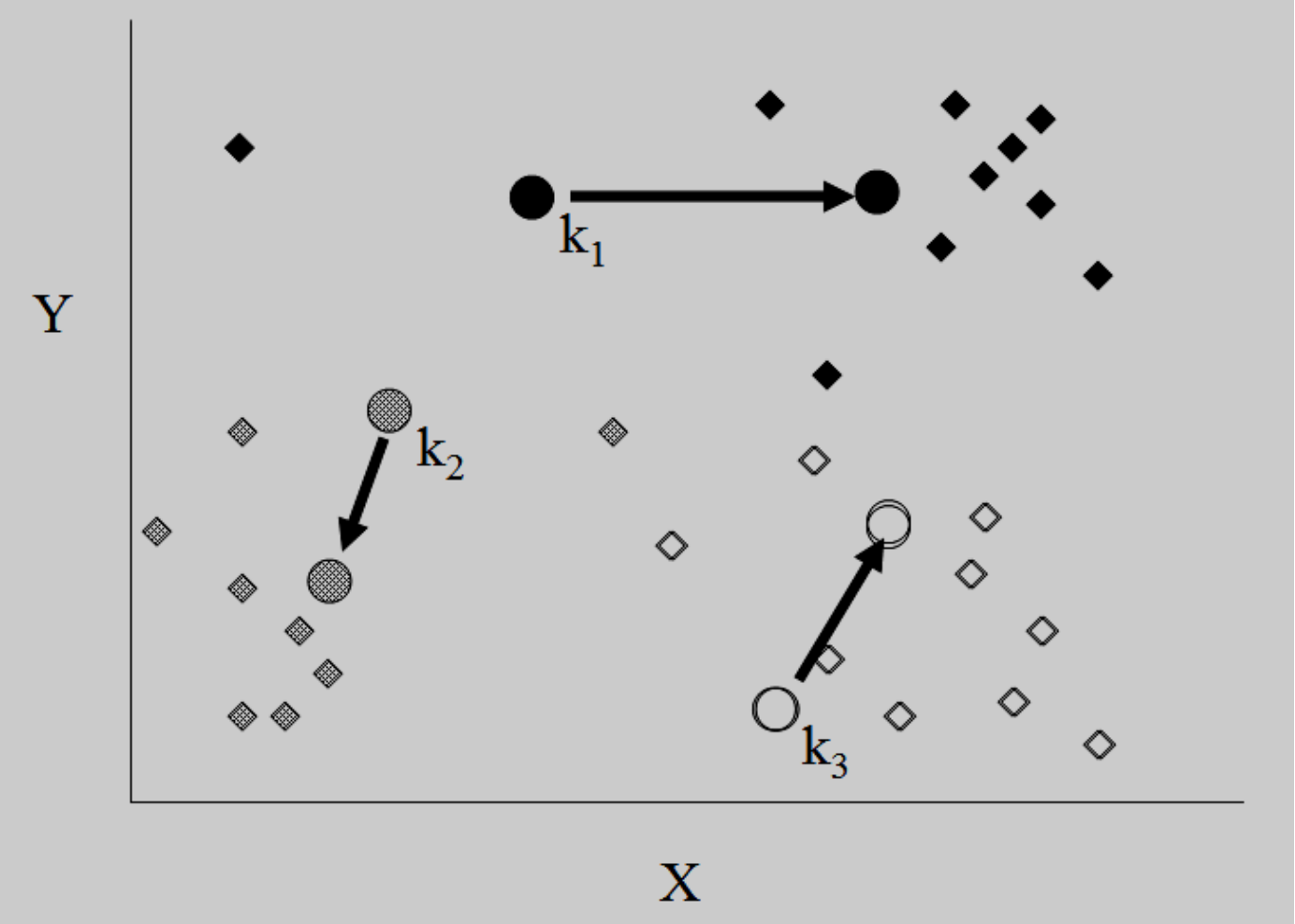
k-means

assign every example to its nearest cluster center



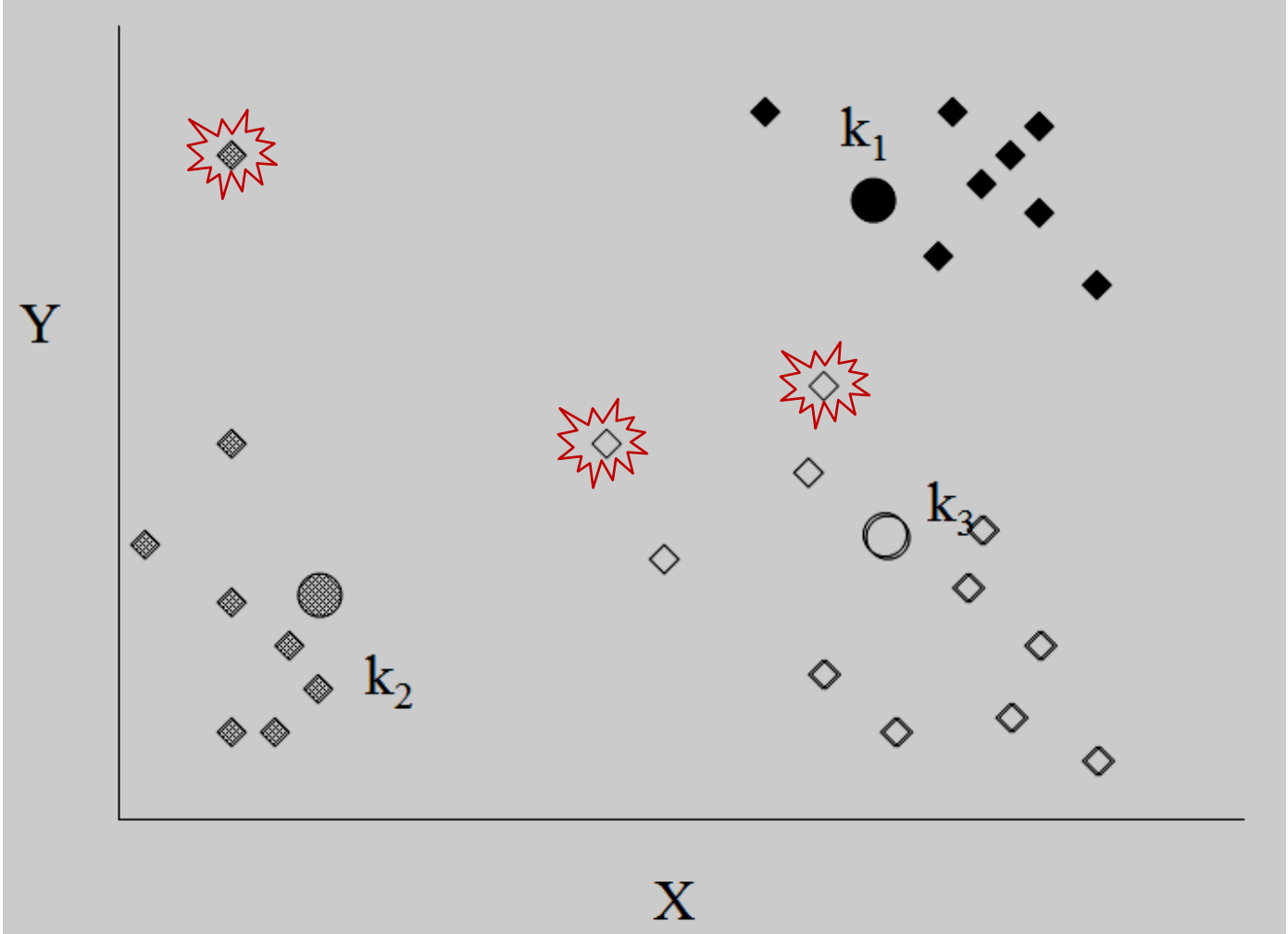
k-means

update 3
cluster centers



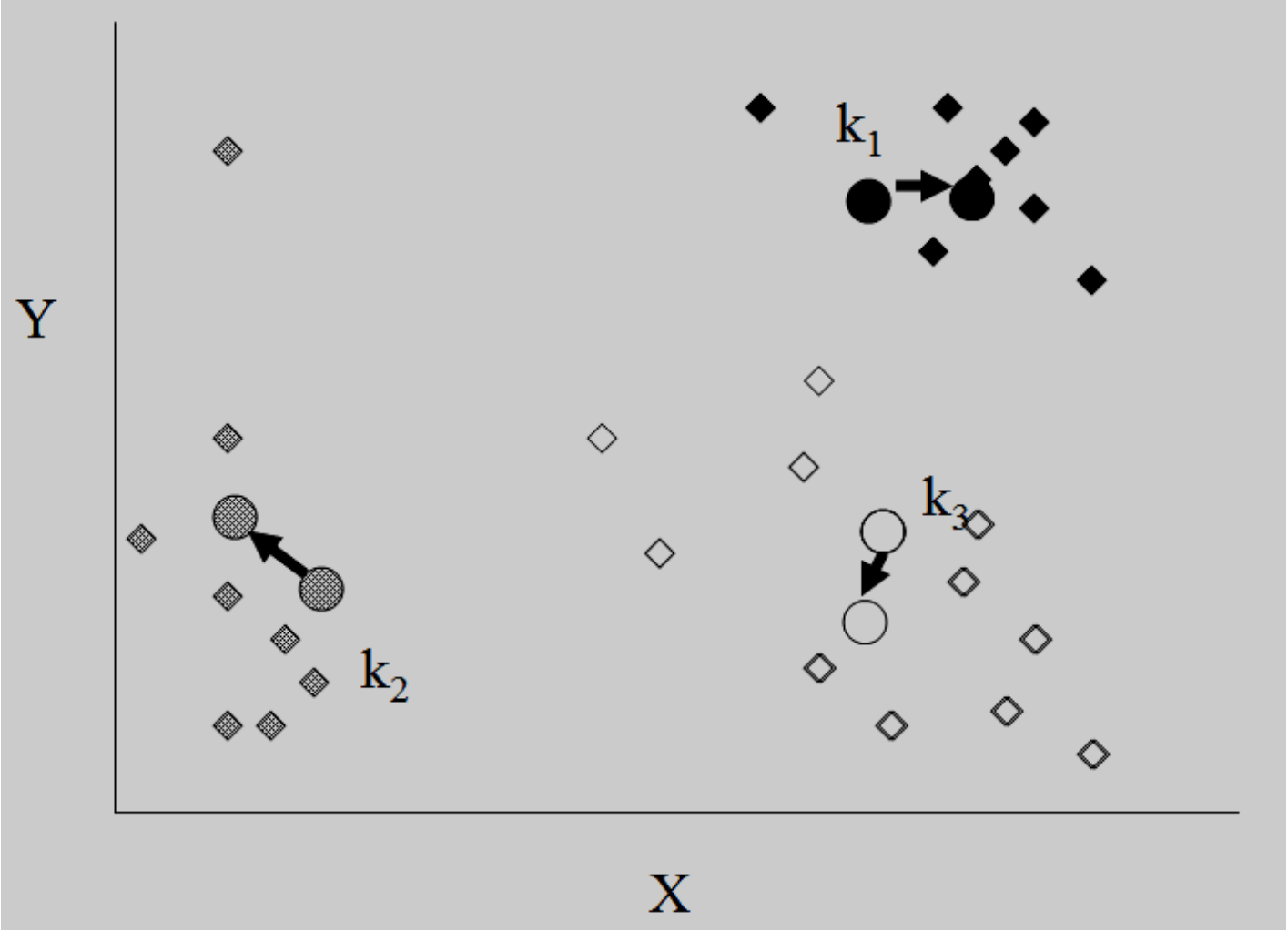
k-means

assign every example to its nearest cluster center

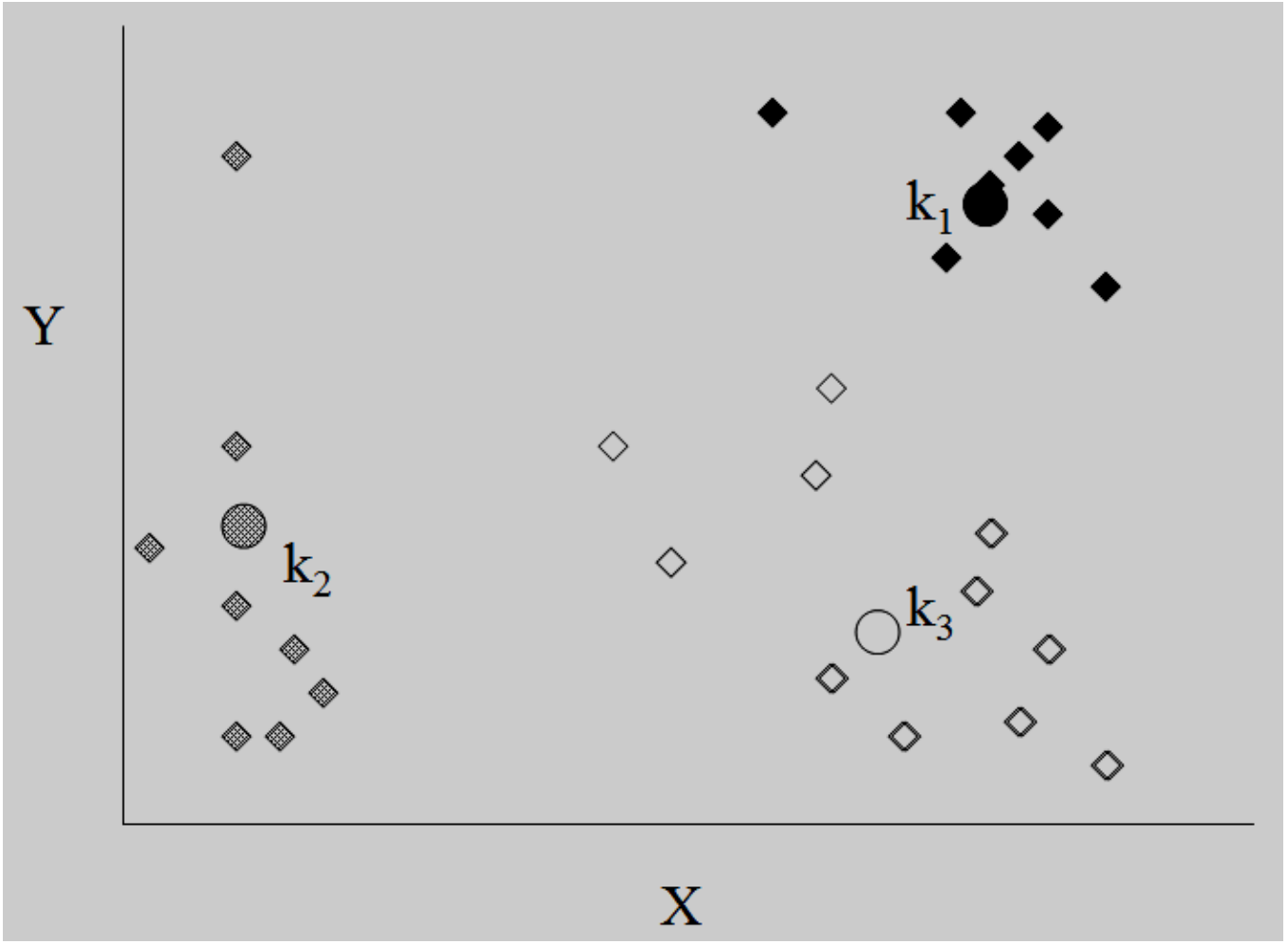


k-means

update 3
cluster centers



k-means



k-means

- Comments
  - Simple, comprehensive
  - Parameter *k*
  - Result can vary significantly depending on initial choice of seeds (number and position)
  - Too sensitive to outliers
  - Shapes of clusters (with radius equal to the distance between the centroid and the furthest data point): spherical

k-means: variations

- *k*-medoids
  - Instead of mean, use medians of each cluster
  - Median advantage: not affected by extreme values
- Fuzzy *k*-means
  - Each example can belong to more than one cluster
- Model-based clustering: EM
  - Assumes that the data were generated by a model and tries to recover the original model from the data
  - Instead of assigning examples to clusters, the EM clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions
  - Maximize the overall probability or likelihood of the data, given the (final) clusters

Merci !