CS57300
PURDUE UNIVERSITY

APRIL 2, 2019

DATA MINING

ANNOUCENMENT

- Assignment 5 is out!
  - Clustering: K-means and agglomerative clustering
  - Due April 19 11:59pm

DESCRIPTIVE MODELING

DATA MINING COMPONENTS

- Task specification: **Description**
- Knowledge representation: **Partition-based, hierarchical, probabilistic model-based**
- Learning technique: **Scoring function + search**
- Evaluation and interpretation

DESCRIPTIVE MODELING: EVALUATION AND INTERPRETATION

DESCRIPTIVE MODEL EVALUATION

- Clustering evaluation
  - **Supervised**: Measures the extent to which clusters match external class label values, e.g., how likely a cluster contains only data instances of a particular class?
  - **Unsupervised**: Measures goodness of fit without class labels, e.g., how closely related instances within each cluster are and distinct instances across different clusters are?

DESCRIPTIVE MODEL EVALUATION

- Describe the current data precisely vs. Generalize to new data
- Example: in partition-based clustering, the model captures the data the best when k=n
- Strike a balance between between how well the model fits and the data and the simplicity of the model

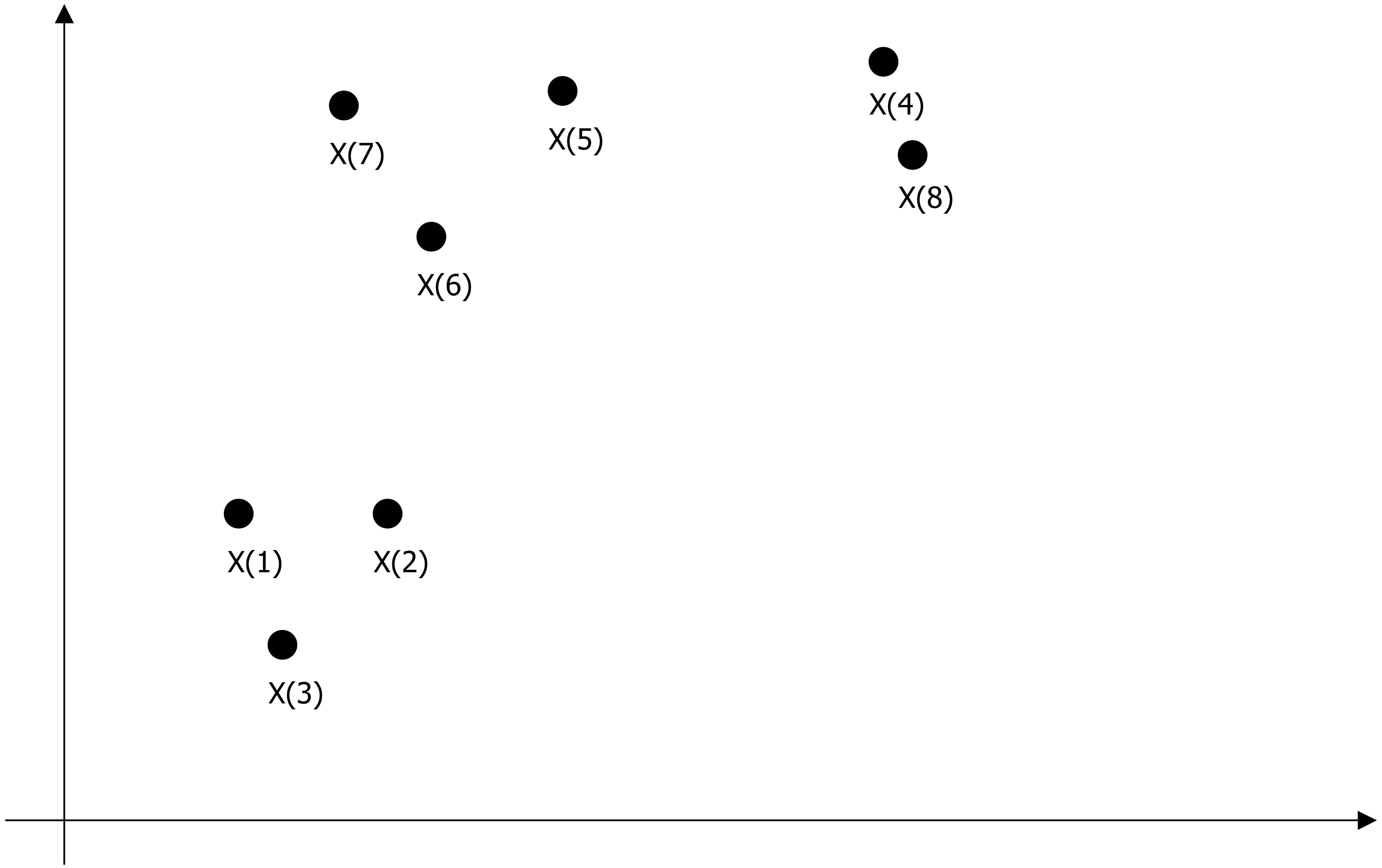
PARTITION-BASED CLUSTERING

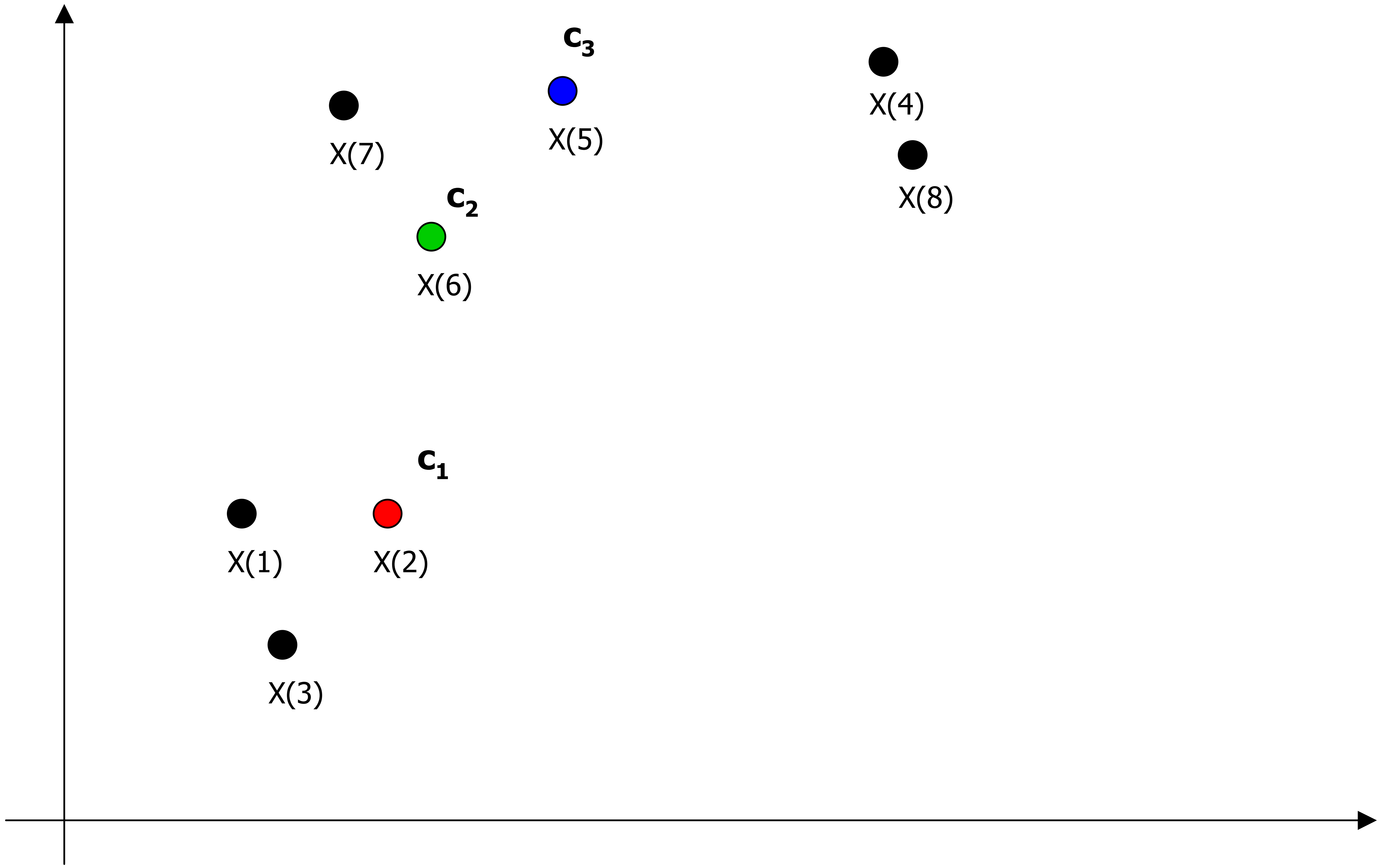
PARTITION-BASED

- Input: data D={**x**(1),**x**(2),...,**x**(n)}
- Output: k clusters $C=\{C_1,\ldots,C_k\}$ such that each **x**(i) is assigned to a unique C_j
- Evaluation: Score(C,D) is maximized/minimized
  - Combinatorial optimization: search among k^n allocations of n objects into k classes to maximize score function
  - Exhaustive search is intractable
  - Most approaches use iterative improvement algorithms

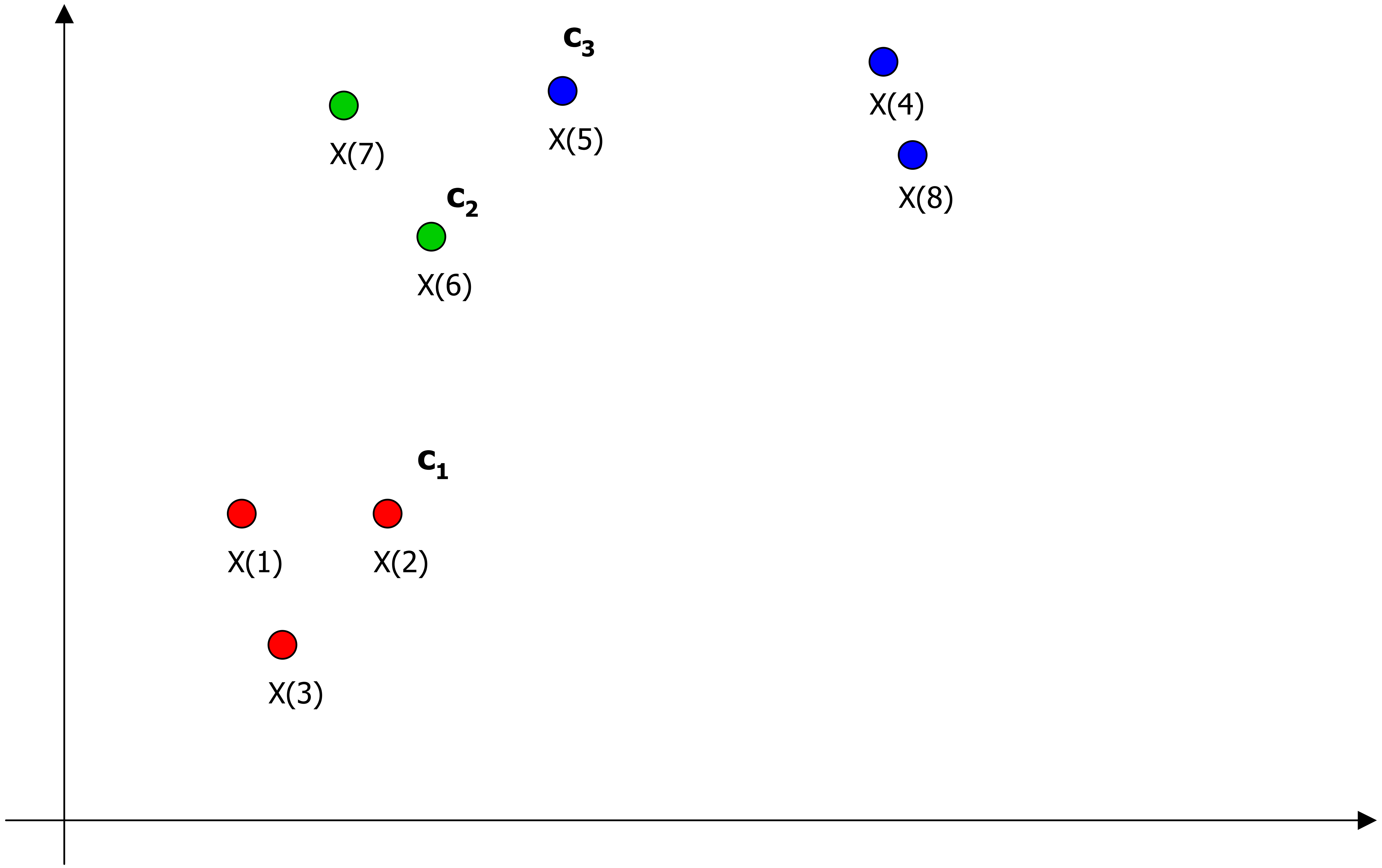
EXAMPLE: K-MEANS

- Algorithm idea:
  - Start with k randomly chosen centroids
  - Repeat until no changes in assignments
    - Assign instances to closest centroid
    - Recompute cluster centroids

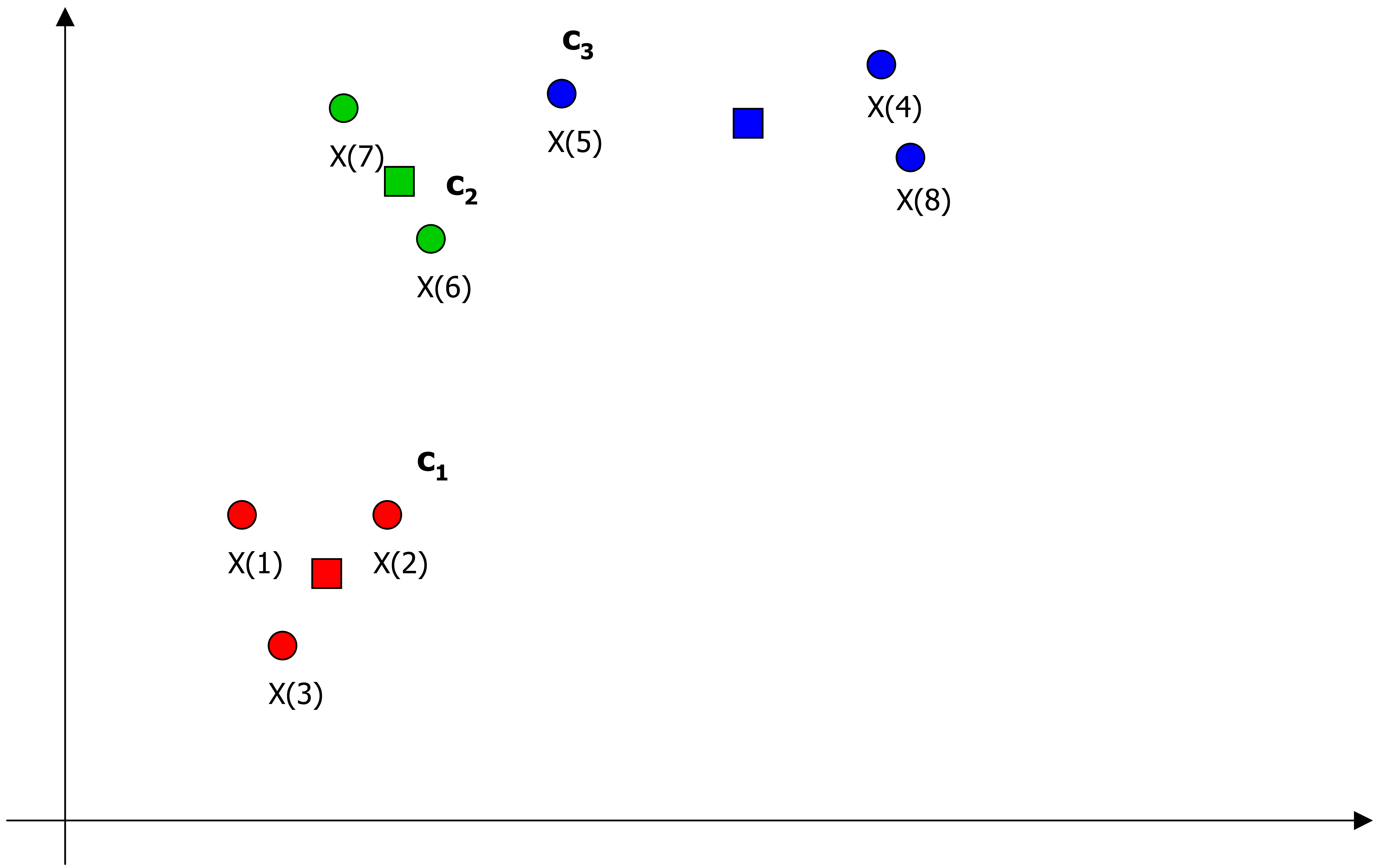

X(7)
X(5)
X(4)
X(8)
X(6)
X(1)
X(2)
X(3)

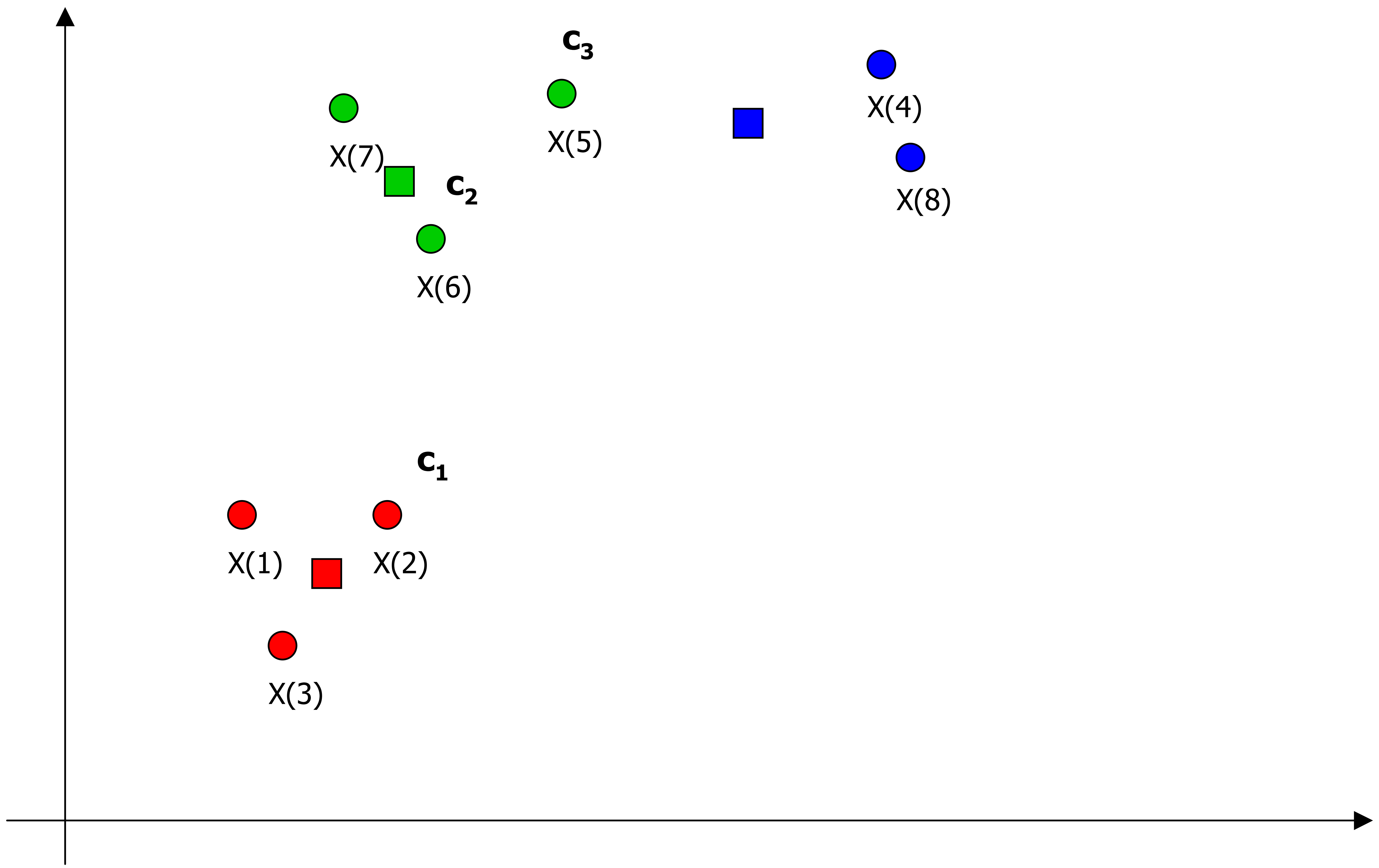


c_3
X(7)
X(5)
X(4)
X(8)
c_2
X(6)
c_1
X(1)
X(2)
X(3)

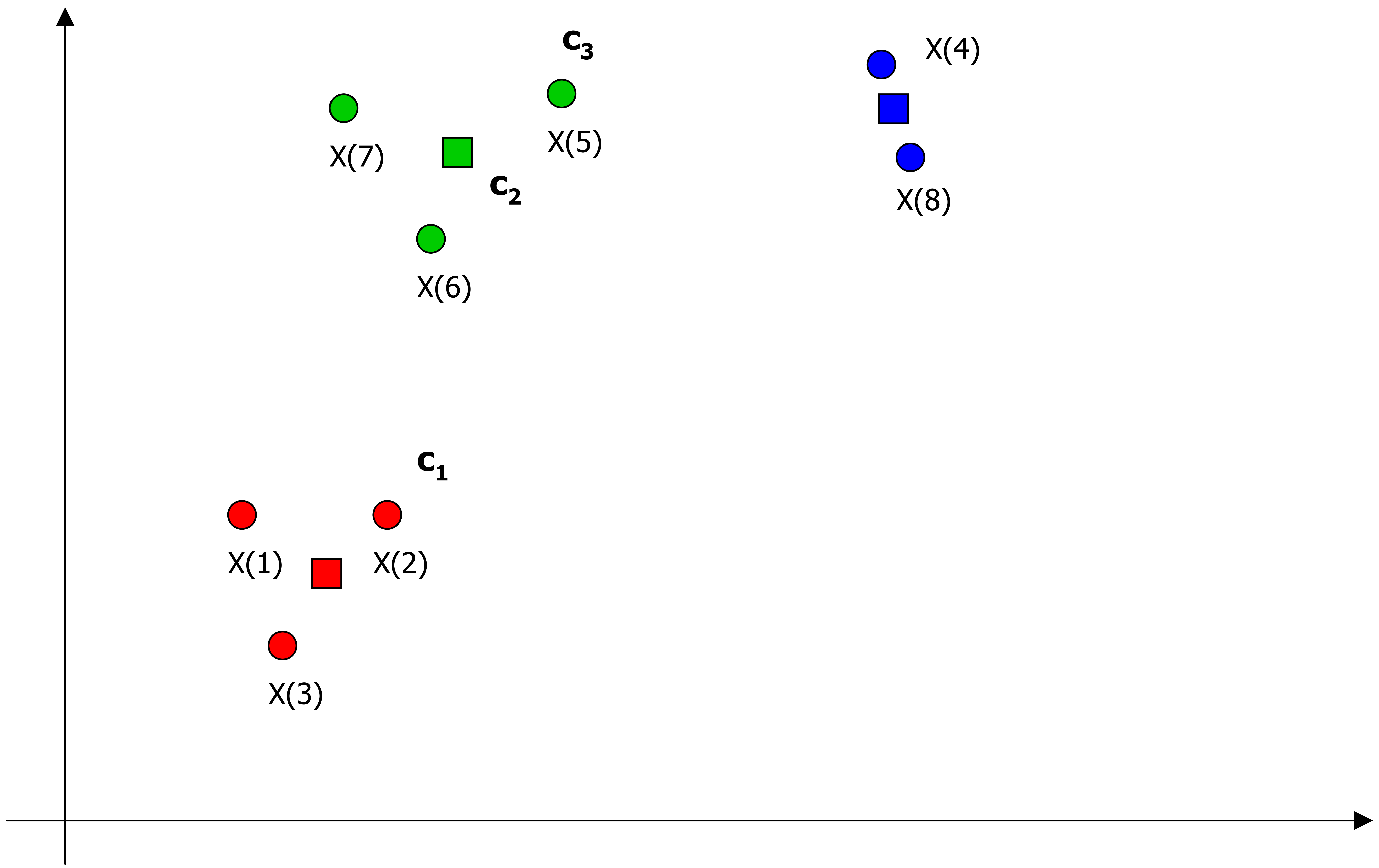


c3
X(7)
X(5)
X(4)
X(8)
c2
X(6)
c1
X(1)
X(2)
X(3)


c3
X(4)
X(5)
X(7)
c2
X(8)
X(6)
c1
X(1)
X(2)
X(3)



c_3
X(7)
X(5)
X(4)
c_2
X(8)
X(6)
c_1
X(1)
X(2)
X(3)


c3
X(4)
X(7)
X(5)
c2
X(8)
X(6)
c1
X(1)
X(2)
X(3)

Algorithm 2.1 The `k-means` algorithm

Input: Dataset D, number clusters k
Output: Set of cluster representatives C, cluster membership vector $\mathbf{m}$
    /* Initialize cluster representatives C */
    Randomly choose k data points from D
    Use these k points as initial set of cluster representatives C
    repeat
        /* Data Assignment */
        Reassign points in D to closest cluster mean
        Update $\mathbf{m}$ such that m_i is cluster ID of ith point in D
        /* Relocation of means */
        Update C such that c_j is mean of points in jth cluster
    until convergence

SCORING FUNCTION OF K-MEANS

- What scoring function is K-means trying to optimize for?

Score function: $$wc(C) = \sum_{k=1}^{K} wc(C_k) = \sum_{k=1}^{K} \sum_{x(i) \in C_k} d(x(i), r_k)^2$$

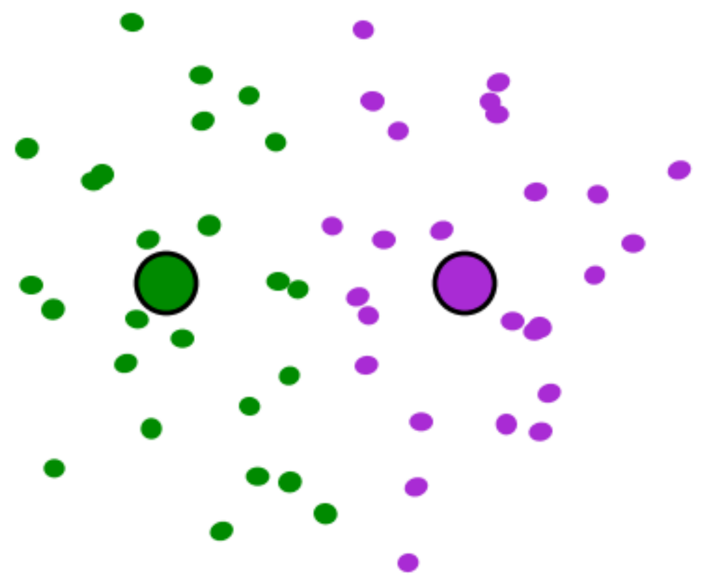
- An alternating optimization approach
  - Fix r_k, optimize for membership of C(x(i)): $min \sum_{i=1}^{N} (x(i) - r_{C(x(i))})^2$
  - Fix C(x(i)), optimize for r_k: $min_{r_k} \sum_{i=1}^{N} (x(i) - r_{C(x(i))})^2 = \sum_{k=1}^{K} \sum_{x \in C_k} (x - r_k)^2$
    - Take derivative with respect to r_k and set to 0 leads to $r_k = \frac{1}{|C_k|} \sum_{x \in C_k} x$

ALGORITHM DETAILS

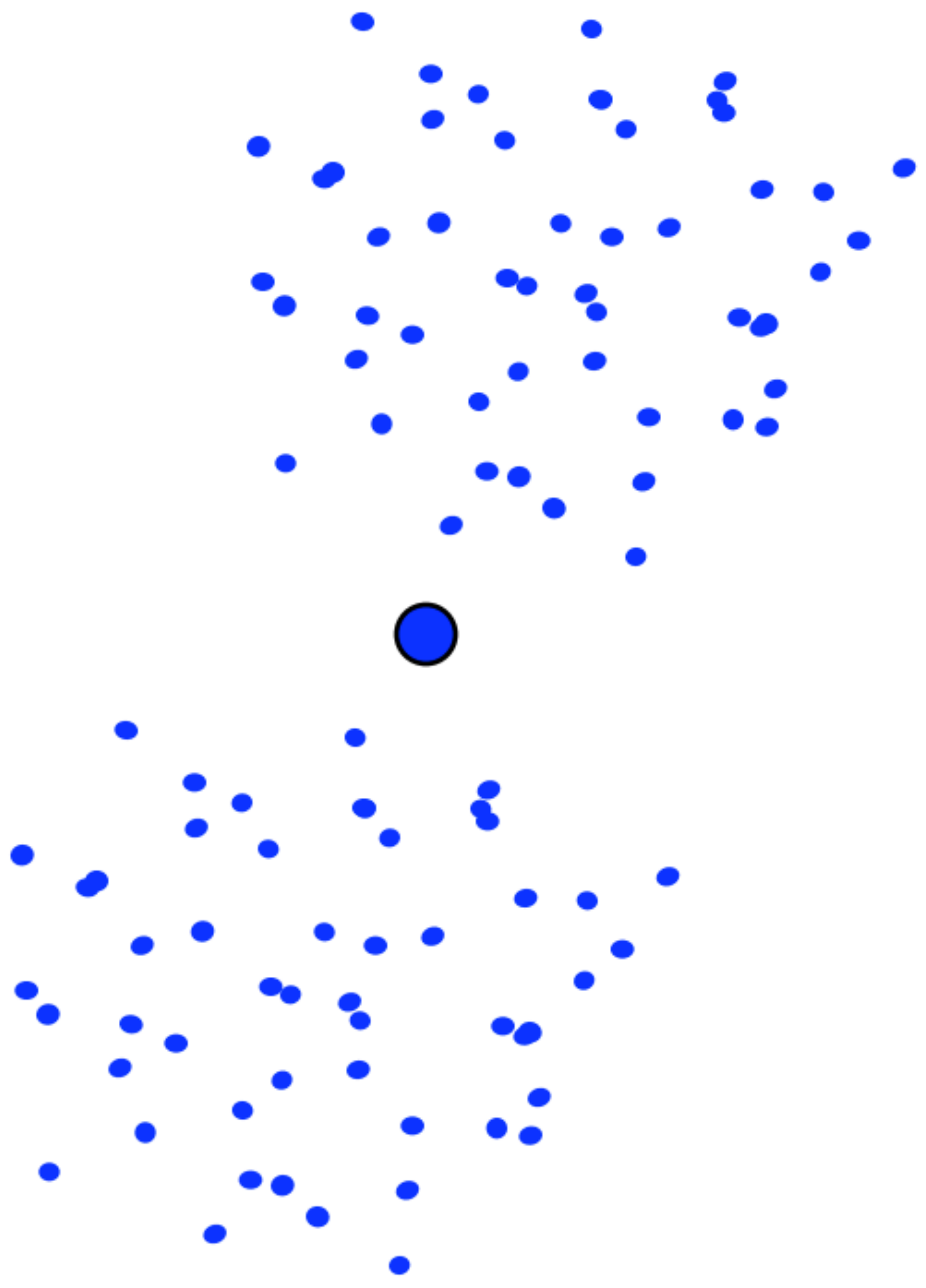
- Does it terminate?
  - Yes, the objective function decreases on each iteration. It usually converges quickly.
- Does it converge to an optimal solution?
  - No, the algorithm terminates at a local optima which depends on the starting seeds.

K-MEANS IS SENSITIVE TO INITIAL SEEDS

A local optimum:



Would be better to have
one cluster here



… and two clusters here

K-MEANS

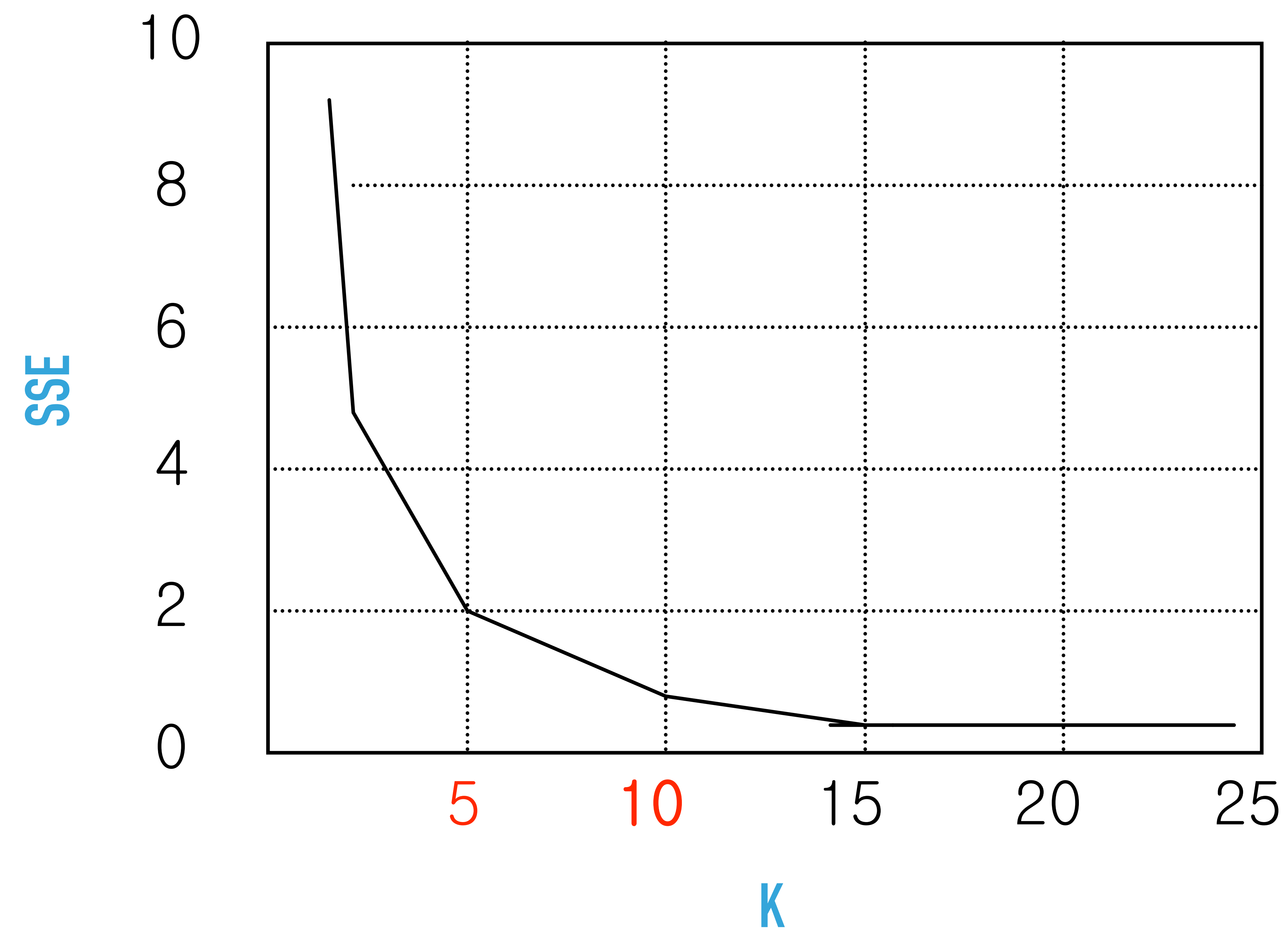
- Strengths:
  - Relatively efficient (time complexity is O(K·N·i), where i is the number of iterations)
  - Finds spherical clusters
- Weaknesses:
  - Terminates at local optimum (sensitive to initial seeds)
  - Applicable only when mean is defined
  - Need to specify K
  - Susceptible to outliers/noise

VARIATIONS

- Selection of initial centroids
  - Select first seed randomly and then pick successive points that are farthest away
  - Run with multiple random selections, pick result with best score
  - Use hierarchical clustering to identify likely clusters and pick seeds from distinct groups
- When mean is undefined
  - K-medioids: use one of the data points as cluster center
  - K-modes: uses categorical distance measure and frequency-based update method

HOW TO SELECT K?

- Plot objective function (i.e., within cluster SSE) as a function of K, and look for "elbow" in plot



K-MEANS SUMMARY

- Knowledge representation
  - K clusters are defined by canonical members (e.g., centroids)
- Model space the algorithm searches over?
  - All possible partitions of the examples into k groups
- Scoring function?
  - Minimize within-cluster Euclidean distance
- Search procedure?
  - Iterative refinement correspond to greedy hill-climbing

HIERARCHICAL CLUSTERING

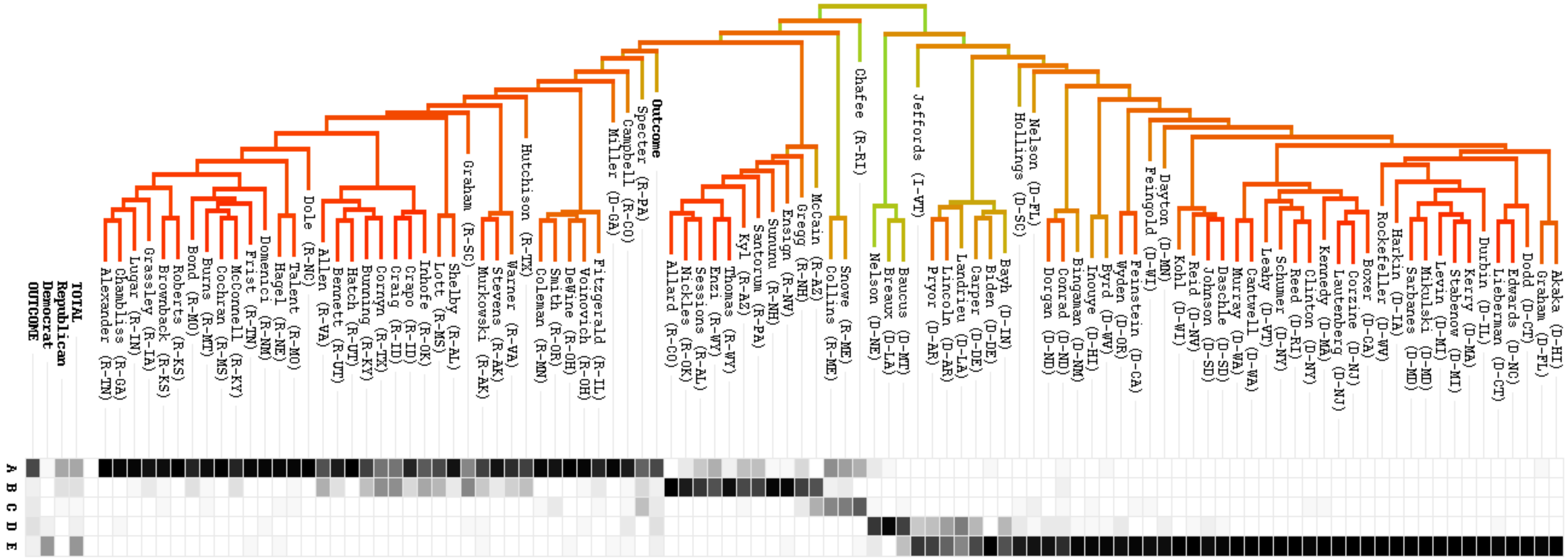
HIERARCHICAL METHODS

- Construct a hierarchy of nested clusters rather than picking K beforehand
- Approaches:
  - Agglomerative: merge clusters successively
  - Divisive: divided clusters successively
- Dendrogram depicts sequences of merges or splits and height indicates distance

AGGLOMERATIVE

- For $i = 1$ to n:
  - Let $C_i = \{x(i)\}$
- While $|C|>1$:
  - Let C_i and C_j be the pair of clusters with min $D(C_i,C_j)$
  - $C_i=C_i \cup C_j$
  - Remove C_j

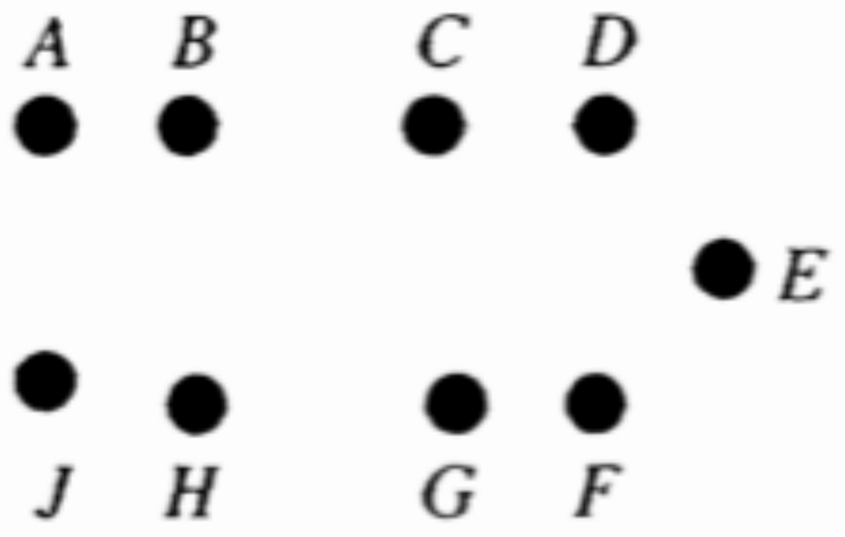
HIERARCHICAL CLUSTERING



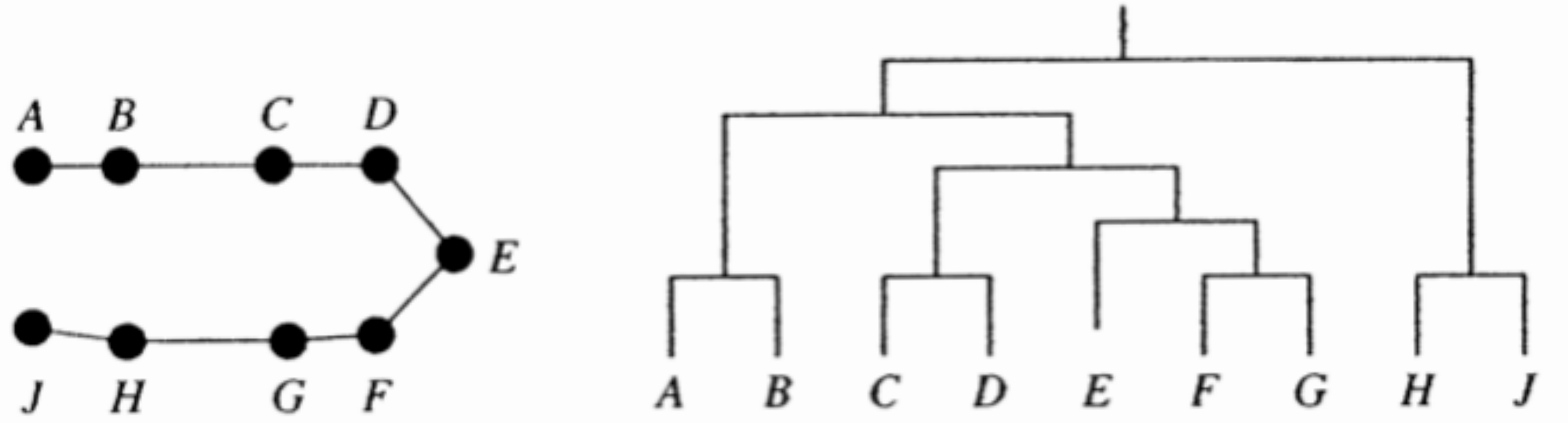
Clustering represented with dendrogram

DISTANCE MEASURES BETWEEN CLUSTERS

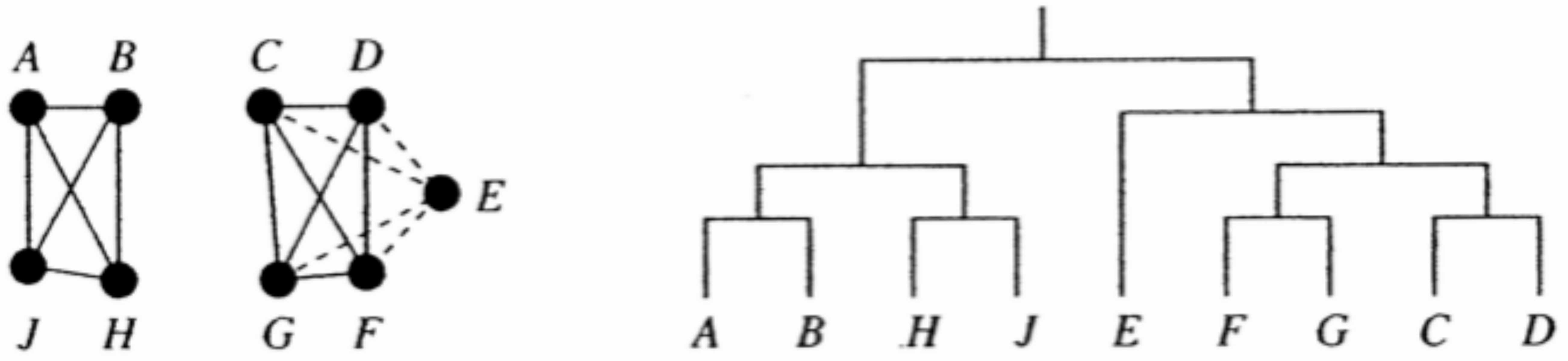
- Single-link/nearest neighbor:
  - $D(C_i,C_j) = \mathbf{min}\{\, d(x,y) \mid x \in C_i,\ y \in C_j \,\}$ ⇒ can produce long thin clusters
- Complete-link/furthest neighbor:
  - $D(C_i,C_j) = \mathbf{max}\{\, d(x,y) \mid x \in C_i,\ y \in C_j \,\}$ ⇒ is sensitive to outliers
- Average link:
  - $D(C_i,C_j) = \mathbf{avg}\{\, d(x,y) \mid x \in C_i,\ y \in C_j \,\}$ ⇒ compromise between the two



(a) Data set



(b) Clustering using single linkage



(c) Clustering using complete linkage

HIERARCHICAL CLUSTERING SUMMARY

- Knowledge representation
  - Dendrogram represents a hierarchy of clusterings
- Model space the algorithm searches over?
  - All possible dendrograms (i.e., hierarchies of partitions from 1 to N)
- Score function?
  - Locally minimize across-cluster distance (e.g., single link)
- Search procedure?
  - Local greedy search

DIVISIVE

- While $|C| < n$:
  - For each C_i with more than 2 objects:
    - Apply partition-based clustering method to split C_i into two clusters C_j and C_k
    - $C = C - \{C_i\} \cup \{C_j, C_k\}$
- Example: spectral clustering