[Clustering]

[Rahmad Mahendra, M.Sc.]

Data Mining for Big Data

Pusat Ilmu Komputer Universitas Indonesia

16 – 20 Juli 2018

Agenda

- What is Clustering
- Major Clustering Methods
- Clustering Evaluation

What is Clustering?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Clustering
  - Grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

Examples of Clustering Applications

- World Wide Web
  - Cluster Weblog data to discover groups of similar access patterns
- Economic Science (especially market research)
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
pusilkom ui

Clustering Considerations

- What does it mean for objects to be similar?
- What algorithm and approach do we take?
- Do we need a hierarchical arrangement of clusters?
- How many clusters?
- Can we label or name the clusters?

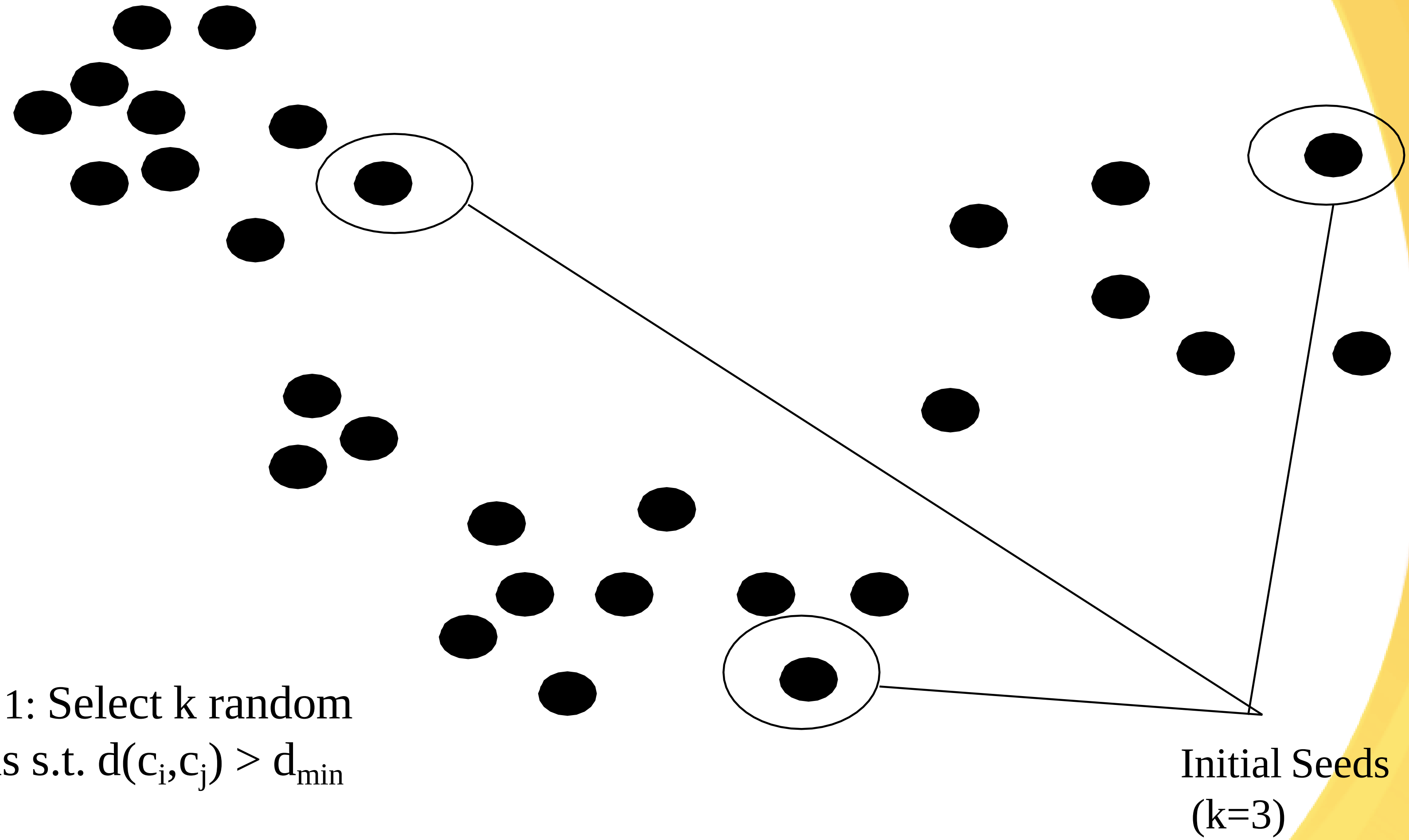
Major Clustering Approaches

- Partitioning algorithms: Construct various partitions and then evaluate them by some criterion
- Hierarchy algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Density-based: based on connectivity and density functions

Partitioning Algorithms

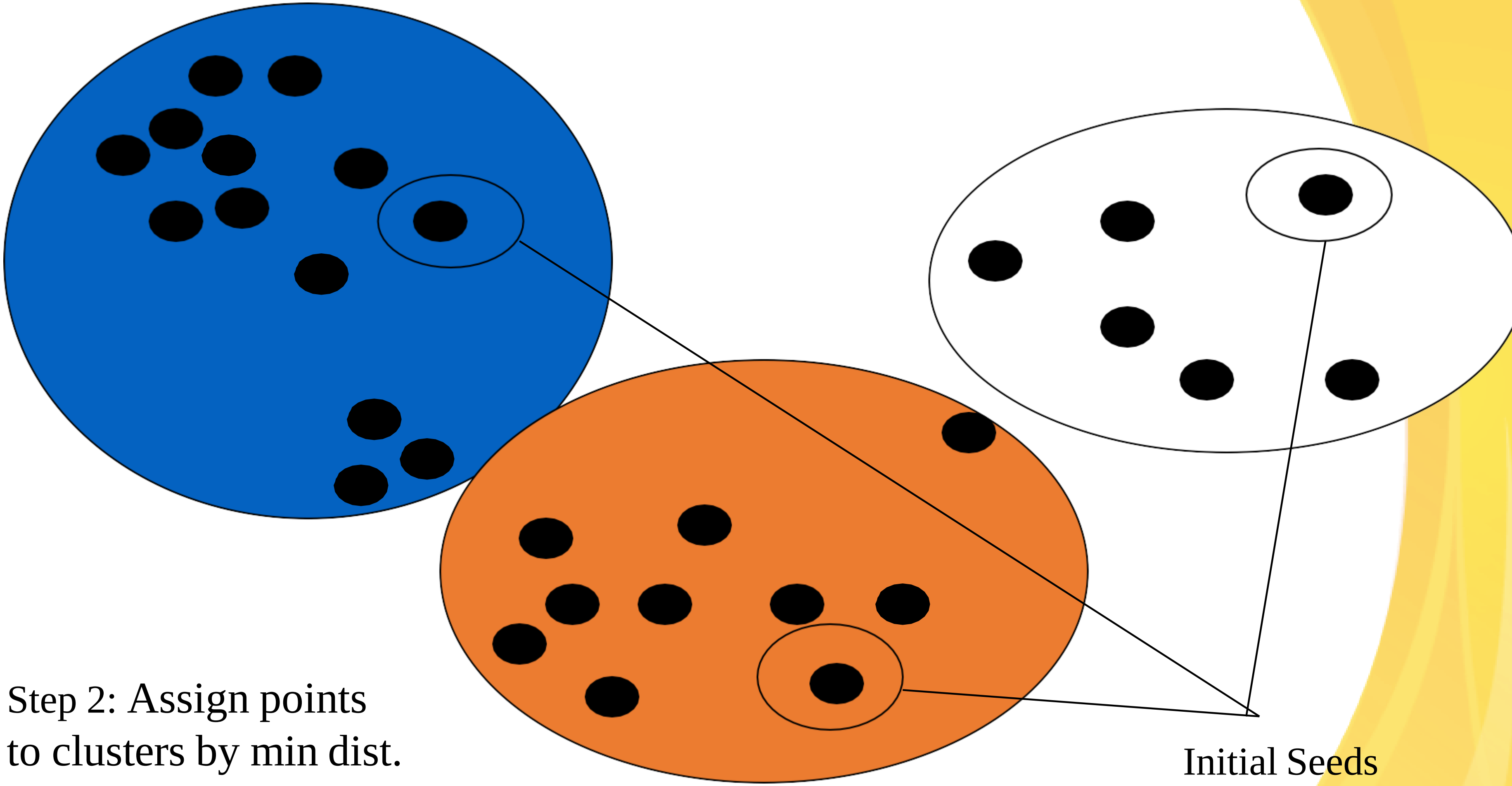
- Partitioning method: Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters
- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: *k-means* and *k-medoids* algorithms
  - *k-means* (MacQueen'67): Each cluster is represented by the center of the cluster
  - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

K-Means Clustering: Initial Data Points



Step 1: Select k random seeds s.t. $d(c_i,c_j) > d_{min}$

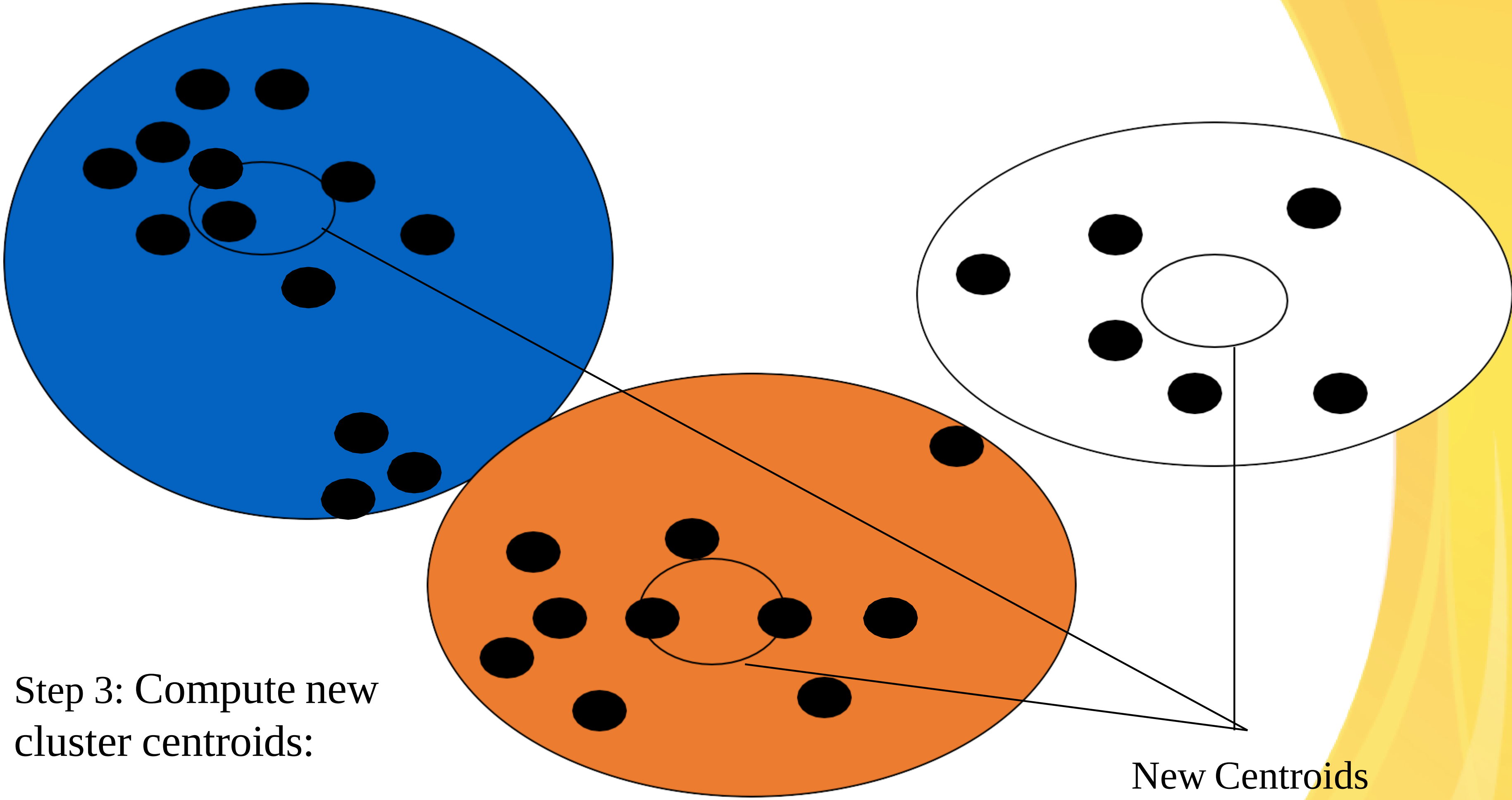
K-Means Clustering: First-Pass Clusters



Step 2: Assign points to clusters by min dist.

$$Cluster(\overrightarrow{p_i}) = \underset{1 \le j \le K}{\text{Argmin}}\, d(\overrightarrow{p_i}, \overrightarrow{c_j})$$

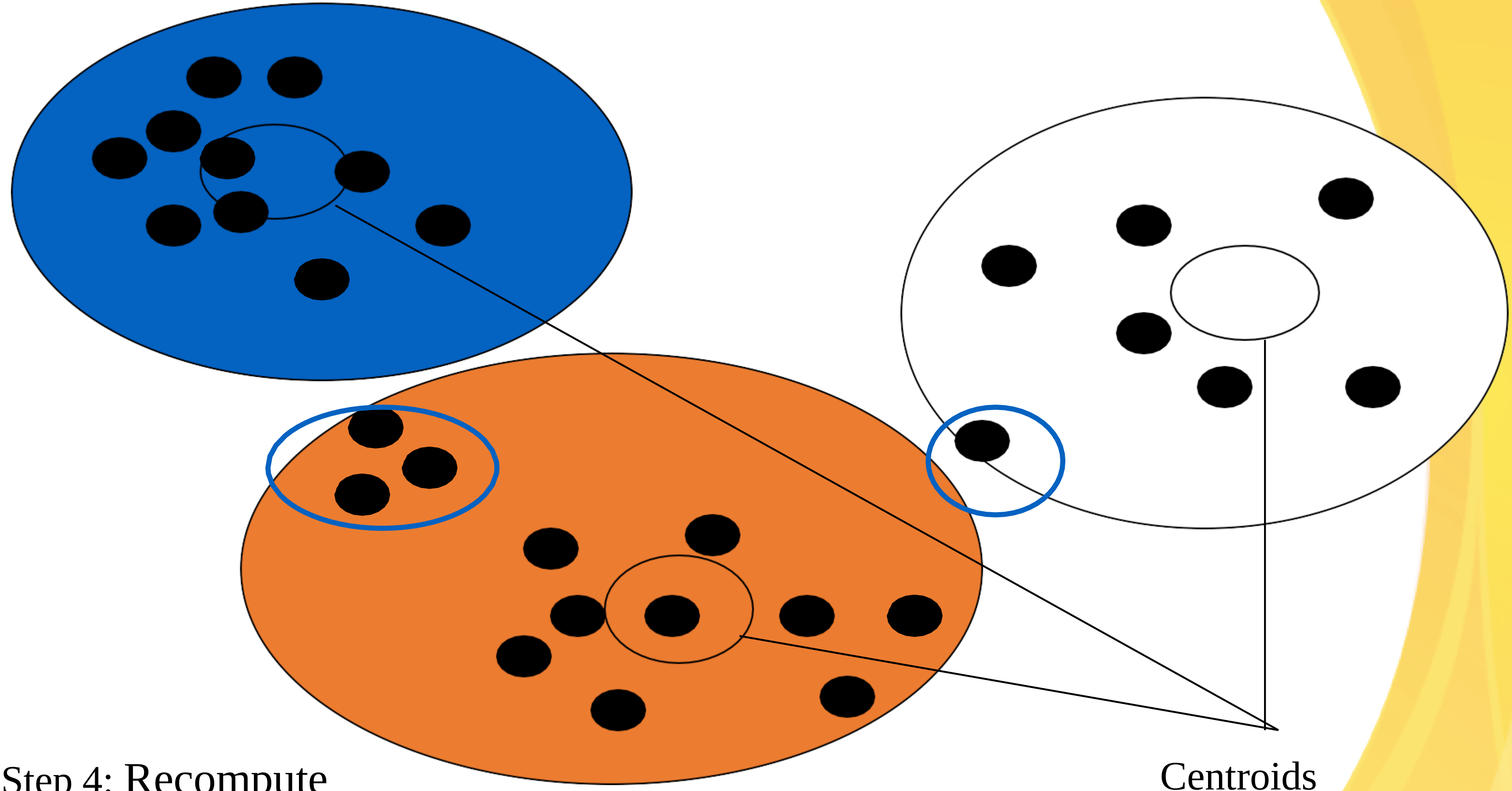
K-Means Clustering: Seeds → Centroids



Step 3: Compute new cluster centroids:

$$\vec{c}_j = \frac{1}{n} \sum_{Cluster(\vec{p}_i)=j} \vec{p}_i$$

K-Means Clustering: Second Pass Clusters

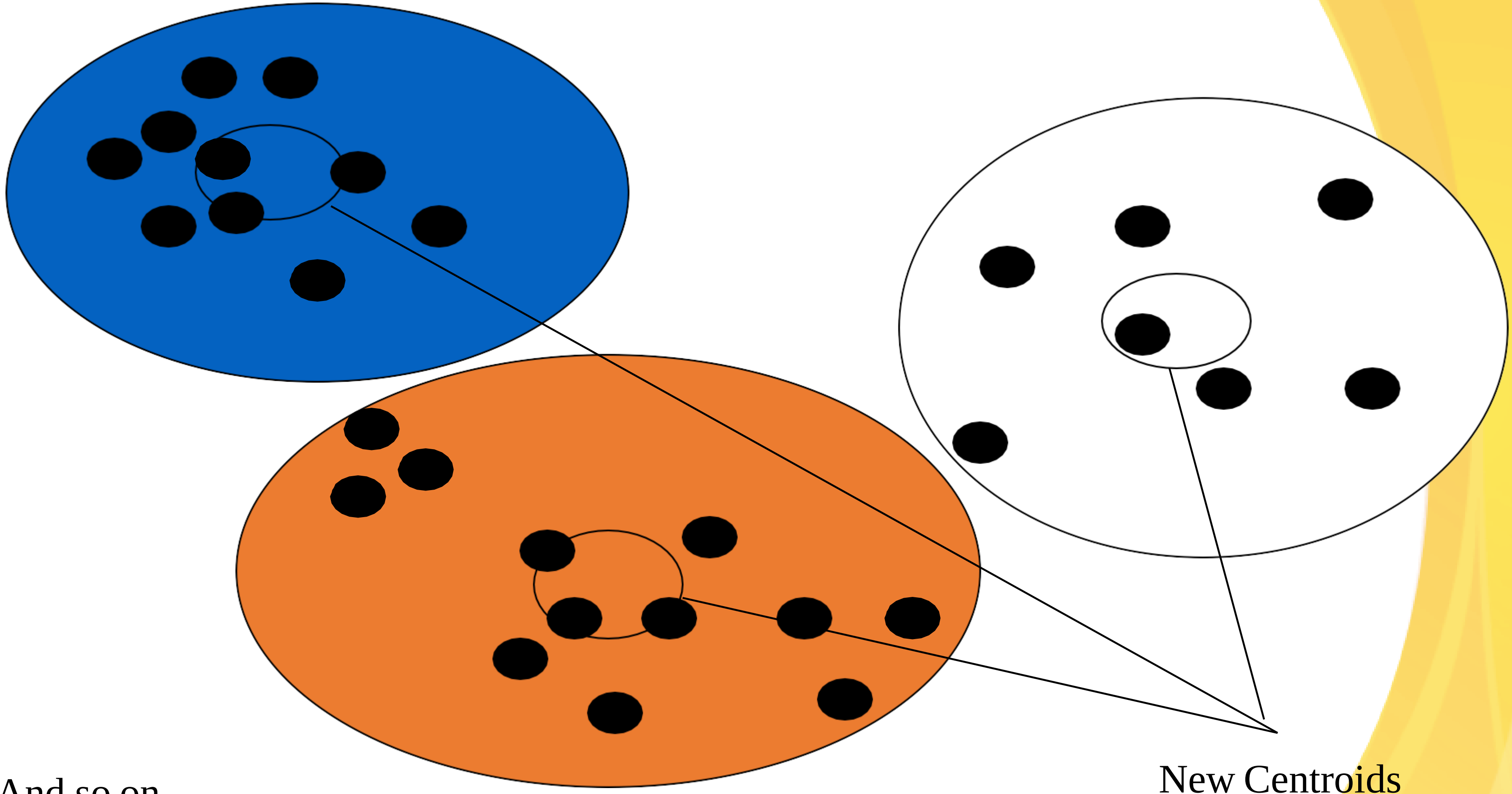


Step 4: Recompute

Centroids

$$Cluster(\overrightarrow{p_i}) = \underset{1 \le j \le K}{\text{Argmin}}\, d(\overrightarrow{p_i}, \overrightarrow{c_j})$$

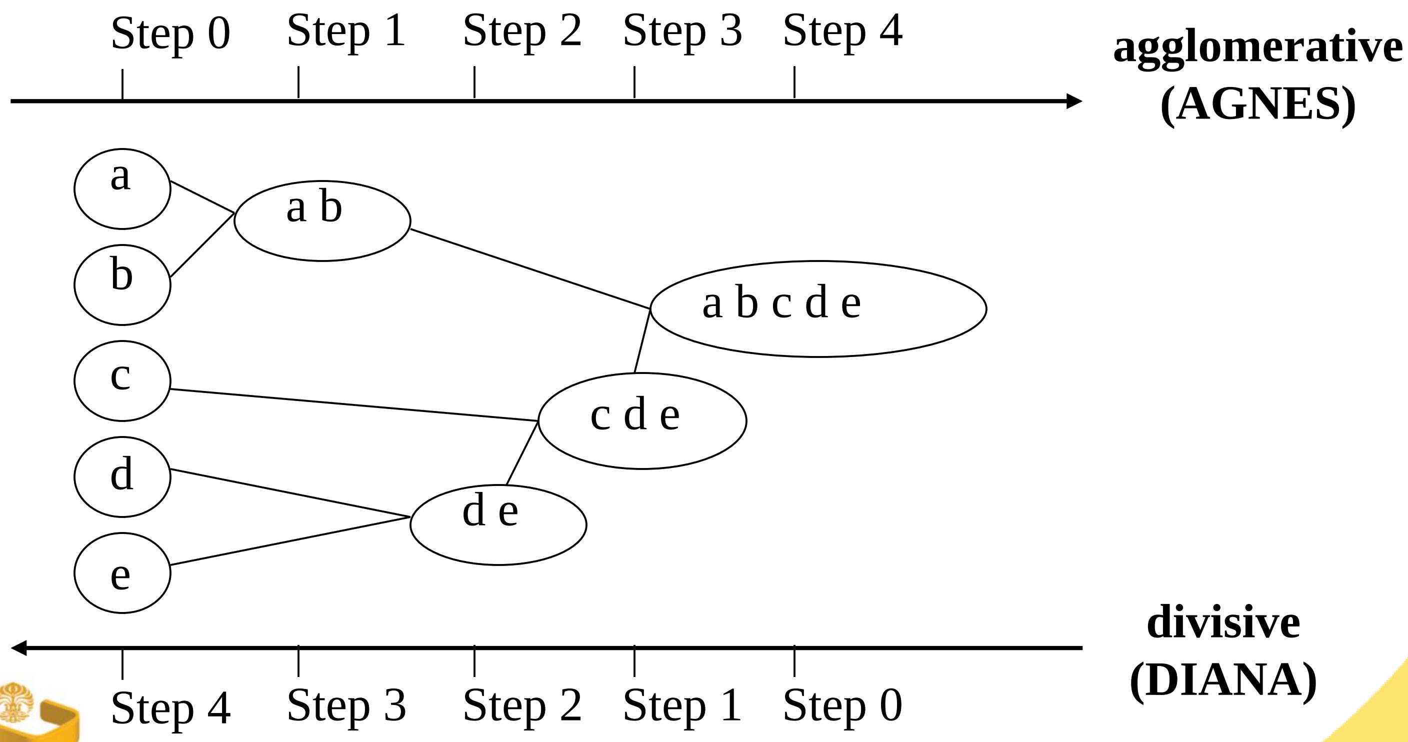
K-Means Clustering: Iterate Until Stability



And so on.

Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition

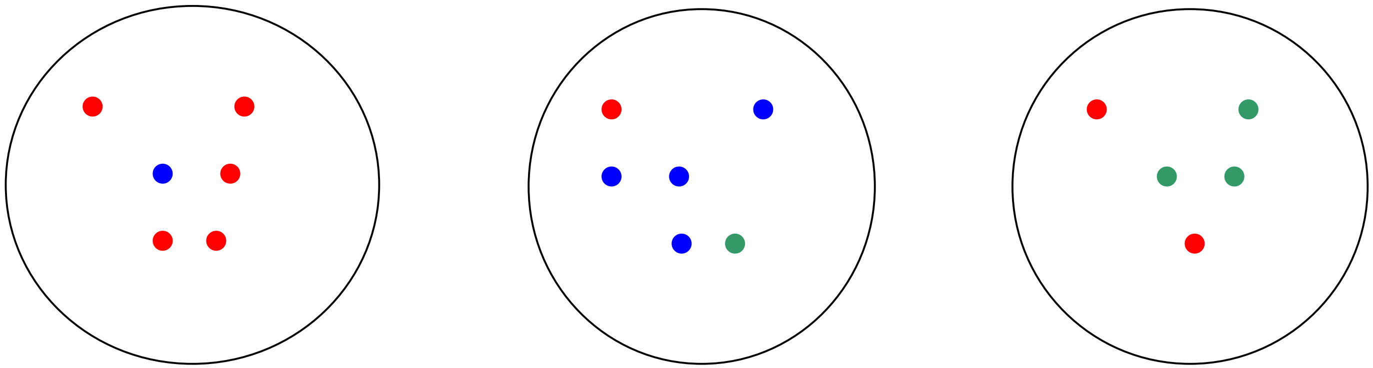


Evaluating Clusters

- Internal Criteria
  - the intra-class (that is, intra-cluster) similarity is high
  - the inter-class similarity is low

- External Criteria
  - Quality measured by its ability to discover some or all of the hidden patterns or latent classes in gold standard data.
  - One of simple measure is **Purity**

Purity metric evaluation



Cluster I   Cluster II   Cluster III

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

Cluster I: Purity = 1/6 * (max(5, 1, 0)) = 5/6

Cluster II: Purity = 1/6 * (max(1, 4, 1)) = 4/6

Cluster III: Purity = 1/5 * (max(2, 0, 3)) = 3/5

Overall: Purity = 1/17 (5+4+3) = 0.71

Summary

- Cluster analysis groups objects based on their similarity and has wide applications
- Measure of similarity can be computed for various types of data
- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
pusilkom ui

Thank You