WeST People and Knowledge Networks | Institute for Web Science & Technologies

UNIVERSITÄT KOBLENZ · LANDAU

Machine Learning and Data Mining WS21/22

"5 Clustering II"

Dr. Zeyd Boukhers

@ZBoukhers

Institute for Web Science and Technologies

University of Koblenz-Landau

November 24, 2021

From previous lecture

- Evaluating clustering results
  - Dunn index
  - Rand index
  - Silhouette coefficient
  - Mutual information
- K-Means
- Expectation Maximization

Agenda

- Other clustering techniques:
  - DBSCAN
  - Agglomerative Hierarchical Clustering
- Clustering for high-dimensional datasets
  - Dimensionality reduction
  - Superspace clustering

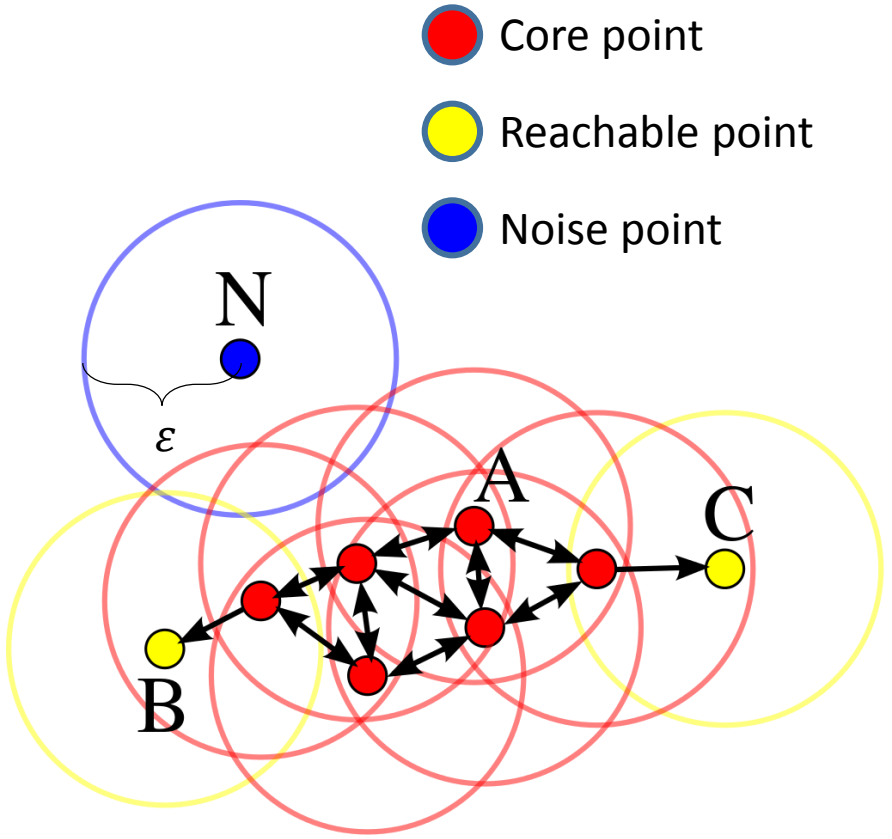
DBSCAN

DBSCAN: Density-based spatial clustering of applications with noise

- Parameters
  - q: Neighborhood size (e.g., 4)
  - ε: Neighborhood distance
- Properties
  - Density-based clustering
  - Non-parametric
    - Data does not require to fit a normal distribution.
  - $O(n^2)$
  - Deterministic on core and noise points (not at border)
  - DBSCAN* treats reachable points at border as noise, making it fully deterministic

DBSCAN

- Given a dataset $X = \{(\boldsymbol{x}_i)\}_{i=1}^{N}$, $V_\varepsilon(\boldsymbol{x}_i)$ is an hypersphere centred at $\boldsymbol{x}_i$ with a radius ε.
  - ε is a user-defined parameter
- Let $N_\varepsilon(\boldsymbol{x}_i)$ be the number of points of X lying in $V_\varepsilon(\boldsymbol{x}_i)$.
- Let q is a user-defined parameter, such that :
- $\boldsymbol{x}_i$ is $\begin{cases} \text{an interior (core)} & \text{if } N_\varepsilon(\boldsymbol{x}_i) \geq q \\ \text{noncore} & \text{Otherwise} \end{cases}$

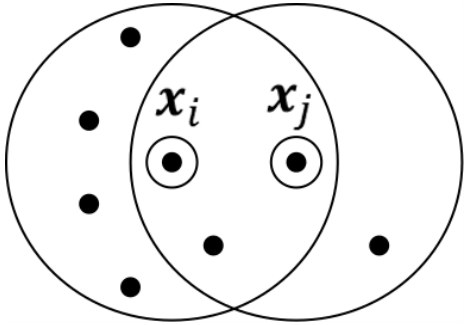


https://en.wikipedia.org/wiki/DBSCAN#/media/File:DBSCAN-Illustration.svg

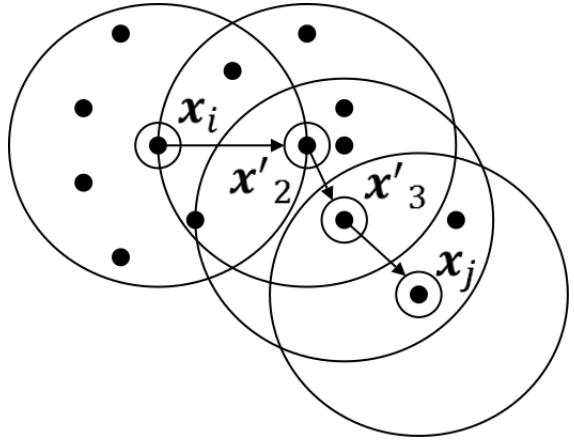
Definitions I

- $\forall\, \boldsymbol{x}_i, \boldsymbol{x}_j \in X, \quad i \neq j$

- $\boldsymbol{x}_j$ is directly density reachable from $\boldsymbol{x}_i$ if:
  - $\boldsymbol{x}_j \in V_\varepsilon(\boldsymbol{x}_i)$ and,
  - $N_\varepsilon(\boldsymbol{x}_i) \geq q$



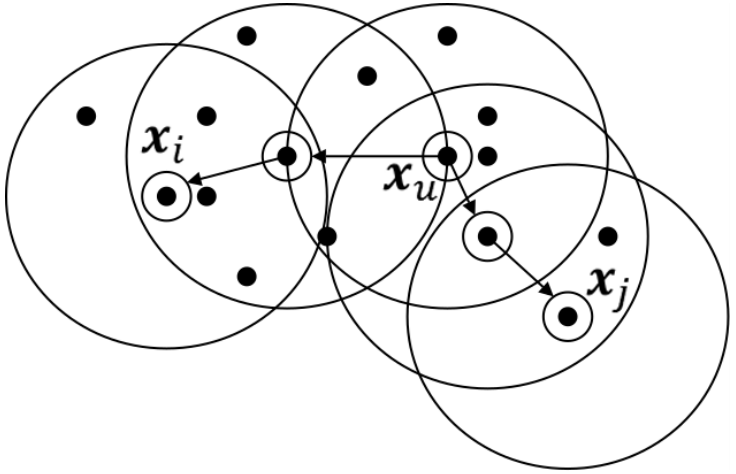
- $\boldsymbol{x}_j$ is density reachable from $\boldsymbol{x}_i$ if:
  - $\exists\, X' = \boldsymbol{x'}_1, \boldsymbol{x'}_2, \cdots \boldsymbol{x'}_p \in X$ and,
  - $\boldsymbol{x'}_1 = \boldsymbol{x}_i$ and,
  - $\boldsymbol{x'}_p = \boldsymbol{x}_j$ and,
  - $\forall\, \boldsymbol{x'}_u \in X'$ $\boldsymbol{x'}_{u+1}$ is directly density reachable from $\boldsymbol{x'}_u$



- Let "$\boldsymbol{x}_j$ is density reachable from $\boldsymbol{x}_i$" be $\boldsymbol{x}_j \leftarrow \boldsymbol{x}_i$

Definitions II

- $\forall\, \boldsymbol{x}_i, \boldsymbol{x}_j \in\ X, \quad i \neq j$



- $\boldsymbol{x}_j$ is density connected to $\boldsymbol{x}_i$ if $\exists \boldsymbol{x}_u \in X$ such that:
  - $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_u$ and,
  - $\boldsymbol{x}_j \leftarrow \boldsymbol{x}_u$
  - Let "$\boldsymbol{x}_j$ is density connected to $\boldsymbol{x}_i$" be $\boldsymbol{x}_j \leftrightarrow \boldsymbol{x}_i$

- We can conclude that a cluster C is defined as non-empty subset of X such that:
  - If $\boldsymbol{x}_i \in C$ and $\boldsymbol{x}_j \leftarrow \boldsymbol{x}_i \implies \boldsymbol{x}_j \in C$
  - $\forall\, \boldsymbol{x}_i, \boldsymbol{x}_j \in C \implies \boldsymbol{x}_j \leftrightarrow \boldsymbol{x}_i$

Definitions III

- Let $\boldsymbol{C} = \{C_1, \cdots, C_k\} \subset X$ be the subset that contains all the points belonging to the k found clusters
- $\forall \boldsymbol{x}_i \in X \mid \boldsymbol{x}_i \notin \boldsymbol{C}$
  - $\boldsymbol{x}_i$ is known as noise.

Algorithm

- Set $X_{un} = X$
- Set $m = 0$
- While $X_{un} \neq \emptyset$ do
  - Arbitrary select $\boldsymbol{x}_i \in X_{un}$
  - If $\boldsymbol{x}_i$ is a noncore point
    - Mark $\boldsymbol{x}_i$ as a noise point
    - $X_{un} = X_{un} \backslash \{\boldsymbol{x}_i\}$
  - Else if $\boldsymbol{x}_i$ is a core point
    - m++
    - Determine all density-reachable points X' in X from $\boldsymbol{x}_i$
    - Assign $\boldsymbol{x}_i$ and X' to C_m
    - Assign border points that may have been marked as noise are assigned to C_m
    - $X' = \emptyset$
    - $X_{un} = X_{un} \backslash C_m$
  - End if
- End while

Remarks

- The results are greatly influenced by the choice of the parameters ε and q
- OPTIC (Ordering points to identify the clustering structure) is an extension of DBSCAN that can solve the problem of carefully choosing the parameters.
- It is not well suited for high-dimensional data.
- Not appropriate when the clusters exhibit significant differences in density.

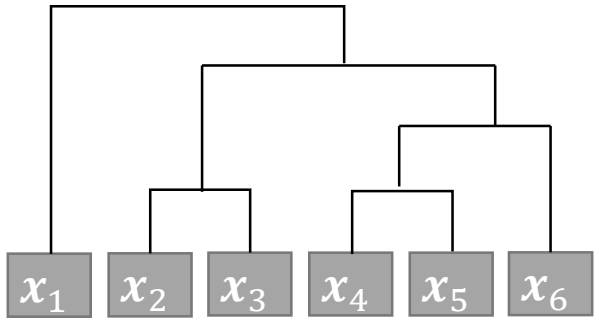
Bonus

- DBSCAN implementation and visualization:
  - If you joined later and you did not submit one of the assignments.
  - Deadline: January 5 at 23:59
  - What to consider:
    - Implement DBSCAN (with basic libraries) for two-dimensional data and visualize the steps of the algorithm.

Agglomerative Hierarchical Clustering

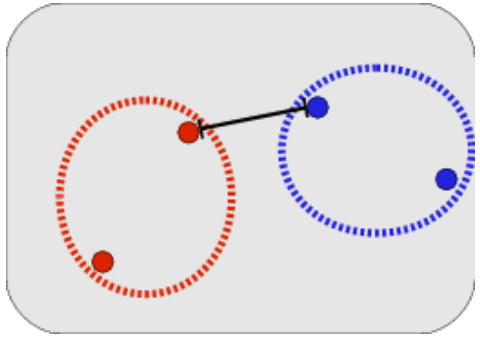
Agglomerative Hierarchical Clustering (AHC)

- Characteristics:
  - "Bottom-up approach"
  - Hierarchical clusters
  - Quadratic runtime (at least!)
  - Deterministic
  - Provides more structural information about clusters
- Parameters
  - Criterion for merging cluster
  - Distance Δ for clusters required

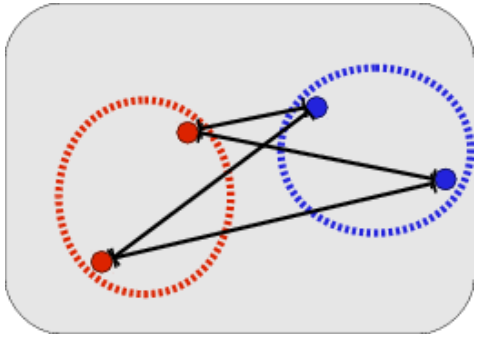


AHC: Algorithm

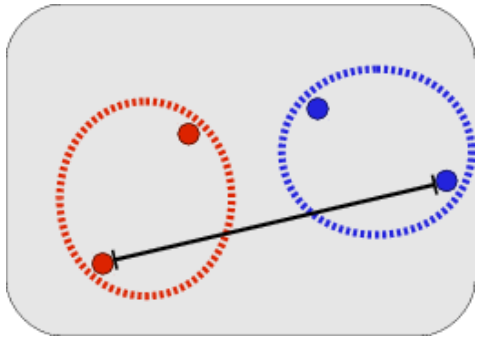
- Initialize N clusters, one per instance
- While $|\Omega| > 1$
  - Merge clusters ω_i and ω_j with minimal $\Delta(\omega_i, \omega_j)$



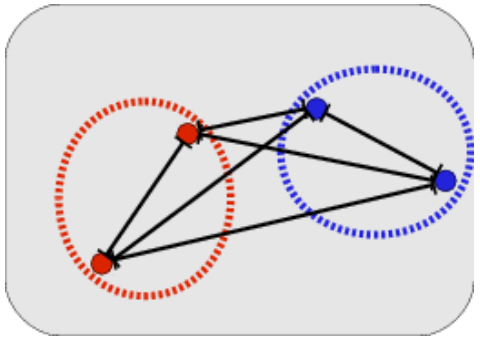
Single Link



Average Link

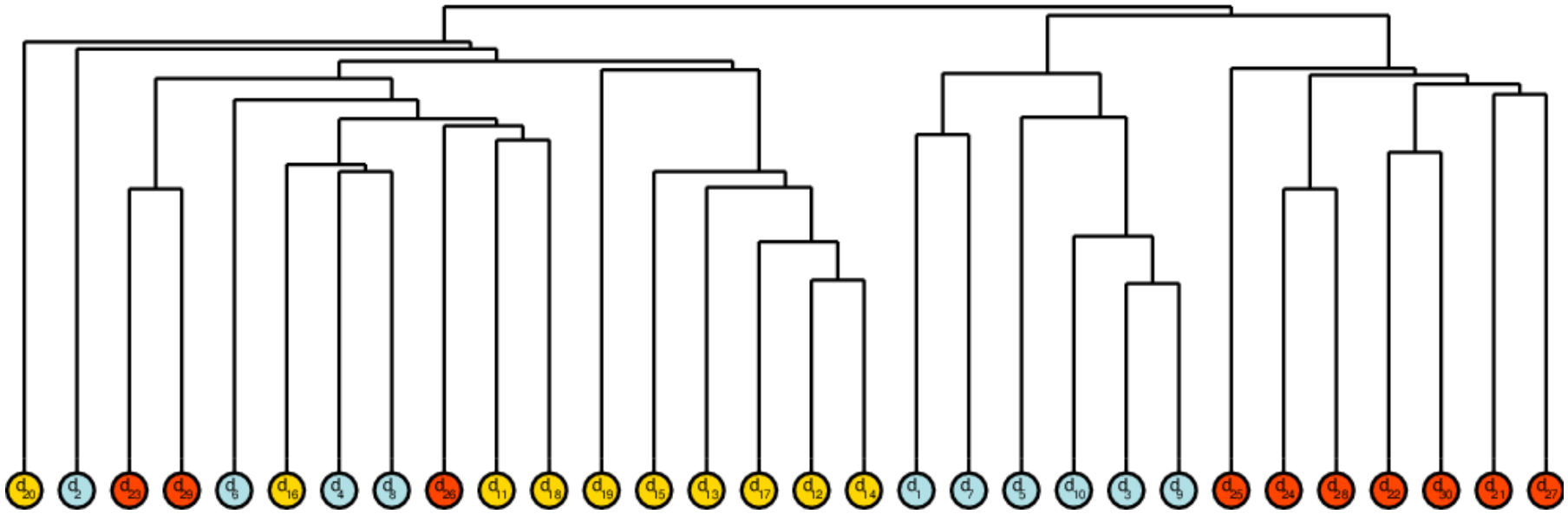


Complete Link



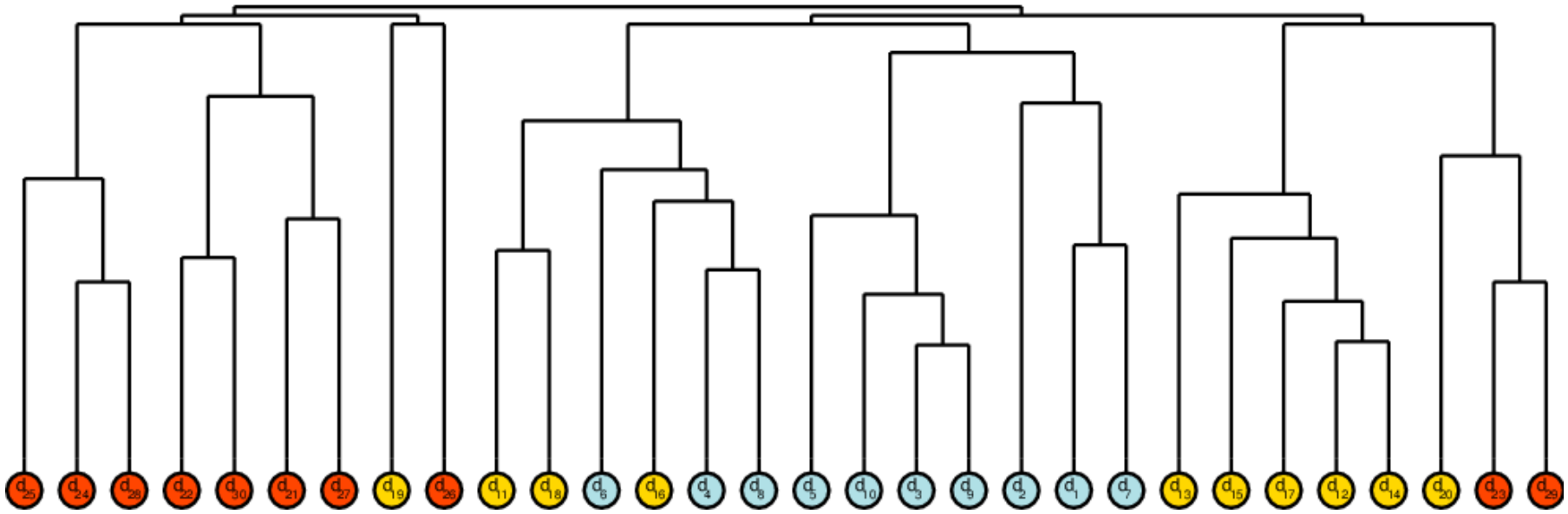
Ward Method

AHC: Single Link



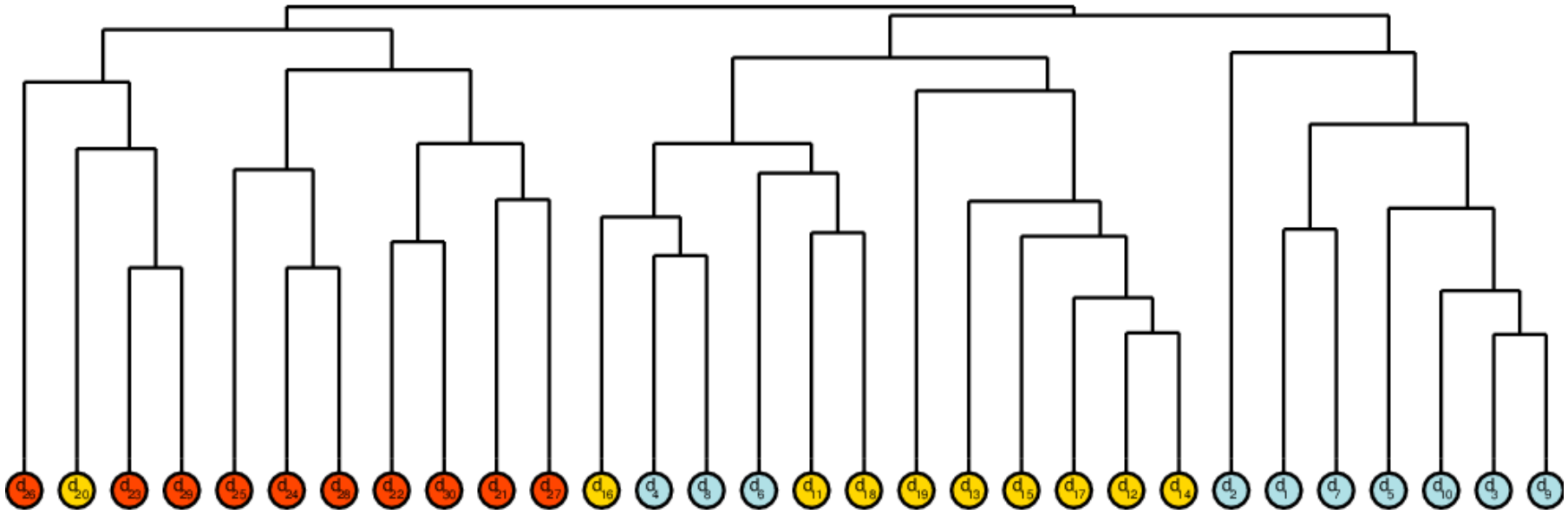
$$\Delta(\omega_i, \omega_j) = \min_{a \in \omega_i, b \in \omega_j} d(a, b)$$

AHC: Complete Link



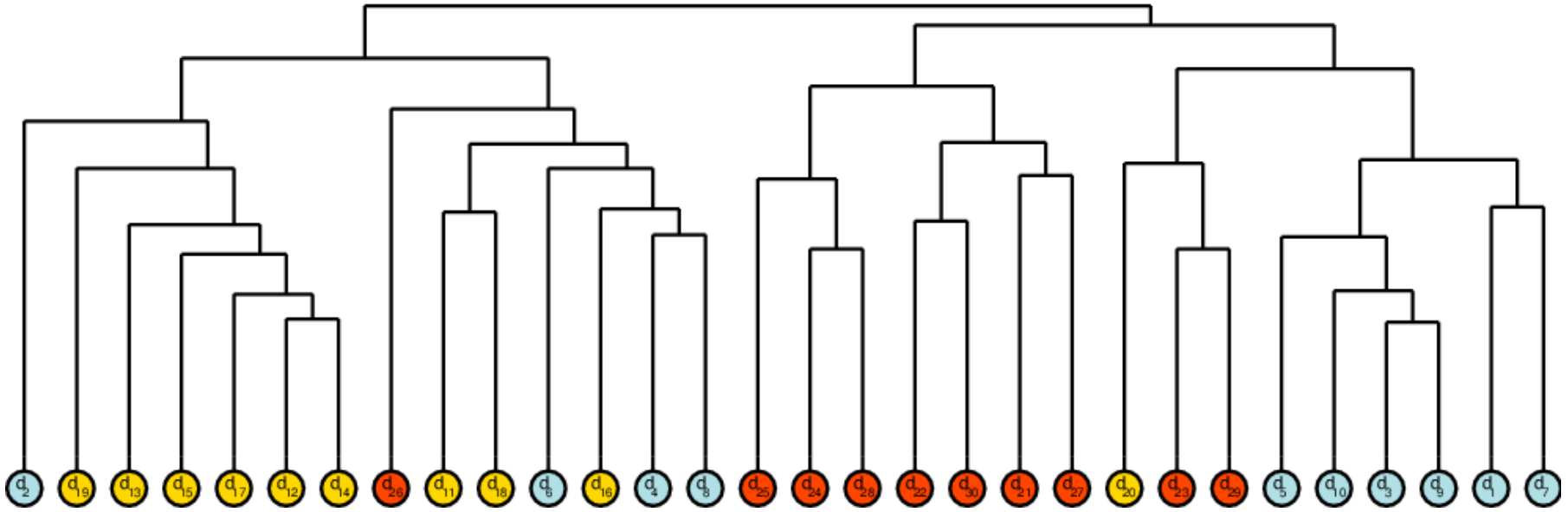
$$\Delta(\omega_i, \omega_j) = \max_{a \in \omega_i, b \in \omega_j} d(a, b)$$

AHC: Average Link



$$\Delta(\omega_i, \omega_j) = \frac{1}{|\omega_i| \cdot |\omega_j|} \sum_{a \in \omega_i, b \in \omega_j} d(a, b)$$

AHC: Ward Method



Recursive definition for N objects:

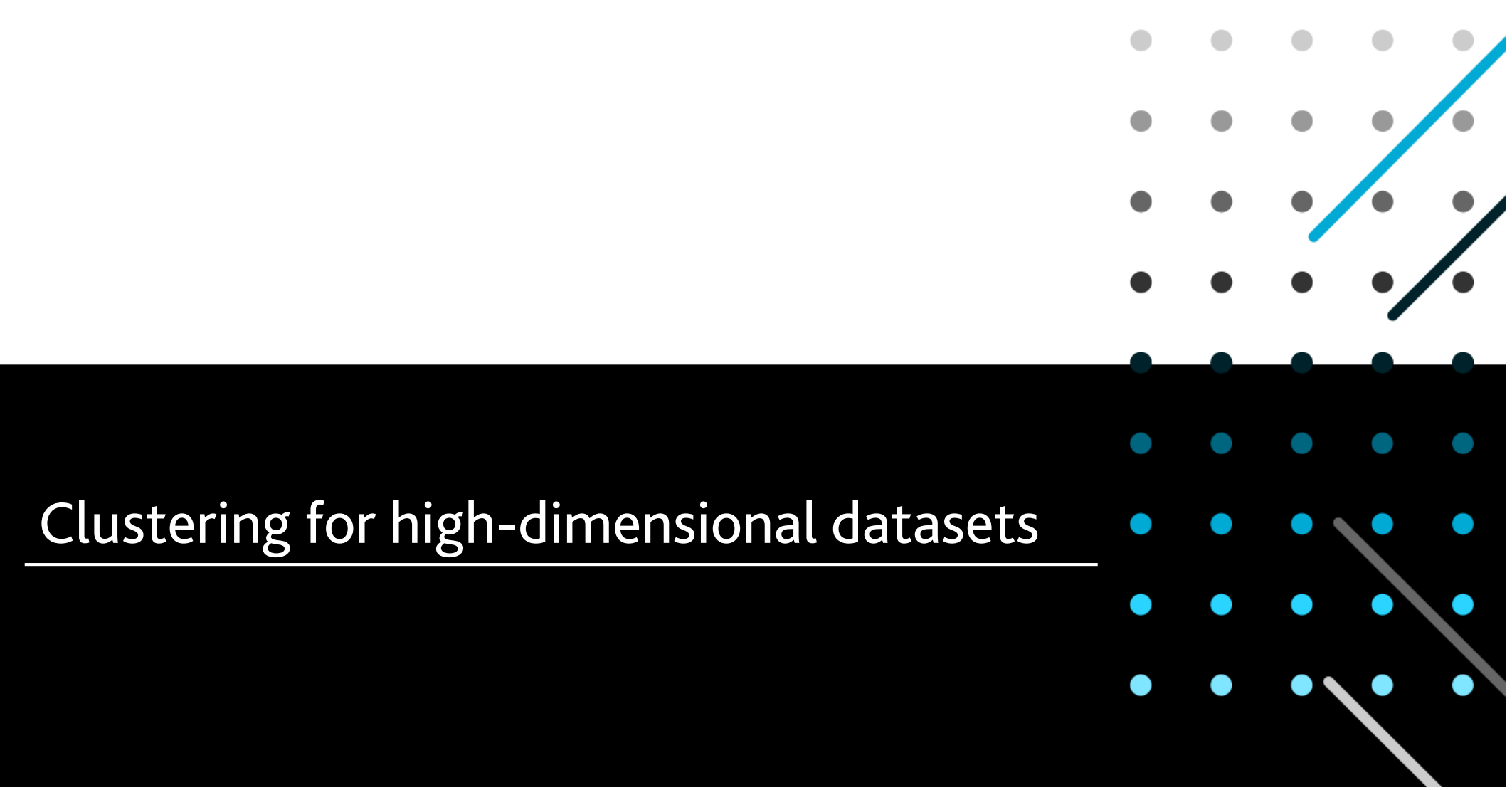
1. Start: $\Delta(\{\boldsymbol{a}\},\{\boldsymbol{b}\}) = (\boldsymbol{a}-\boldsymbol{b})^2$

2a. Given distance for all pairs $\Delta(\omega_k, \omega_l)$, merge 2 closest clusters ω_i, ω_j with sizes n_i, n_j

2b. Recompute cluster distances to new cluster $\omega_i \cup \omega_j$, minimizes the total within-cluster variance

$$\Delta(\omega_i \cup \omega_j, \omega_k) = \frac{n_i + n_k}{N}\Delta(\omega_i, \omega_k) + \frac{n_j + n_k}{N}\Delta(\omega_j, \omega_k) - \frac{n_k}{N}\Delta(\omega_i, \omega_j)$$

Clustering for high-dimensional datasets



Clustering for high-dimensional datasets

- Most of the approaches discussed so far consider all dimensions of the features space simultaneously.

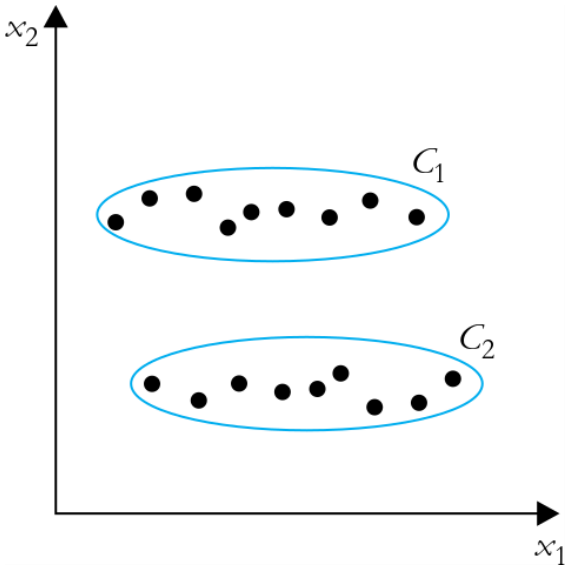
BUT

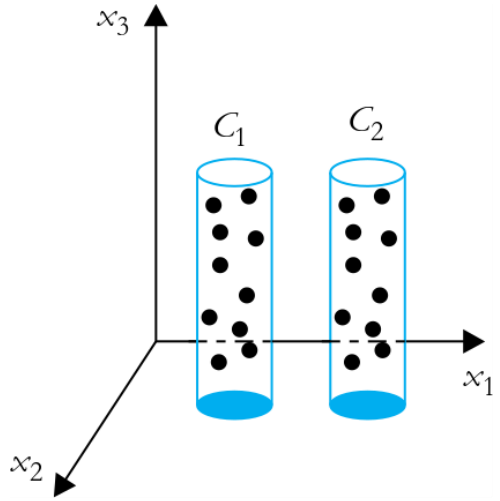
- In some applications such as bioinformatics and web mining, the dimensionality of the datasets can be as high as few thousands.

Problematic and not practical

Problems

- Fixed number of data points and an increasing feature-dimensionality make the data points spread out in the space.
  - They become equidistant → the meaning of similarity and dissimilarity is lost.
- In very high dimensional spaces, often only a small fraction of the features contributes to the formation of the clusters.
- Solutions:
  - Dimensionality reduction
  - Superspace clustering

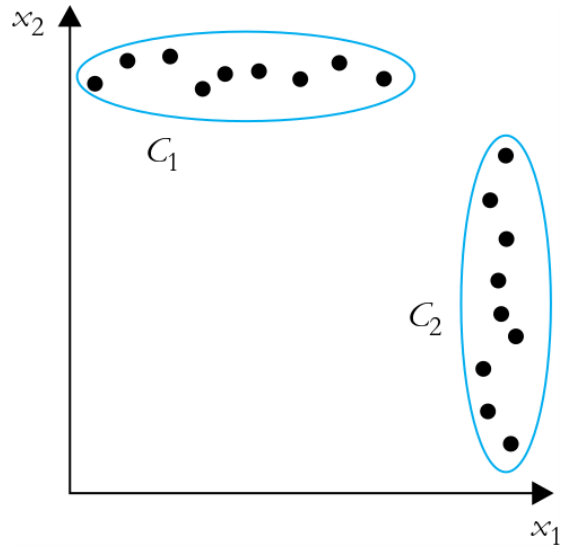




Dimensionality reduction

- PCA: Principal Component Analysis
- SVD: Singular Value Decomposition
- Autoencoder
- Etc.

But it is not practical in all cases.



Superspace clustering

- This technique reveals the clusters and their superspaces (supposed to be different).
- There exist two algorithm categories:
  - Grid-based
    1. Identify the superspaces that are likely to contain clusters.
    2. Determine the clusters lying in each superspace.
    3. Obtain the result.
  - Point-based
    - The clusters as well as the superspaces are simultaneously determined.

Summary

Summary

- Clustering
  - DBSCAN
  - AHC
  - Clustering for high-dimensional datasets

Thank you!

Zeyd Boukhers

E-mail: Boukhers@uni-koblenz.de
Phone: +49 (0) 261 287-2765
Web: Zeyd.Boukhers.com

University of Koblenz-Landau
Universitätsstr. 1
56070 Koblenz