Unsupervised Learning: Concept & Applications

Unsupervised learning finds hidden structures in data without relying on ground-truth labels. Algorithms identify patterns — such as groupings, low-dimensional representations, or rare anomalies — using only the input features.

Key Applications

Clustering (Group Discovery)

Organise large image corpuses by visual similarity for efficient pre-labelling or exploration.

Anomaly Detection (Rare-Event Discovery)

Identify potential network intrusions from connection features, where outliers indicate security threats.

Clustering & K-Means

K-Means: objective, algorithm, and properties

The K-Means algorithm is a fundamental method in unsupervised learning for partitioning data.

Goal

Partition (n) samples into (k) clusters so each sample is assigned to the nearest cluster centroid.

Objective (WCSS)

$$J = \sum_{1-i}^{k} \sum_{x \in C_i} |x - \mu_i|^2$$

Minimize (J) (tight intra-cluster variance).

Algorithm (iterative coordinate descent)

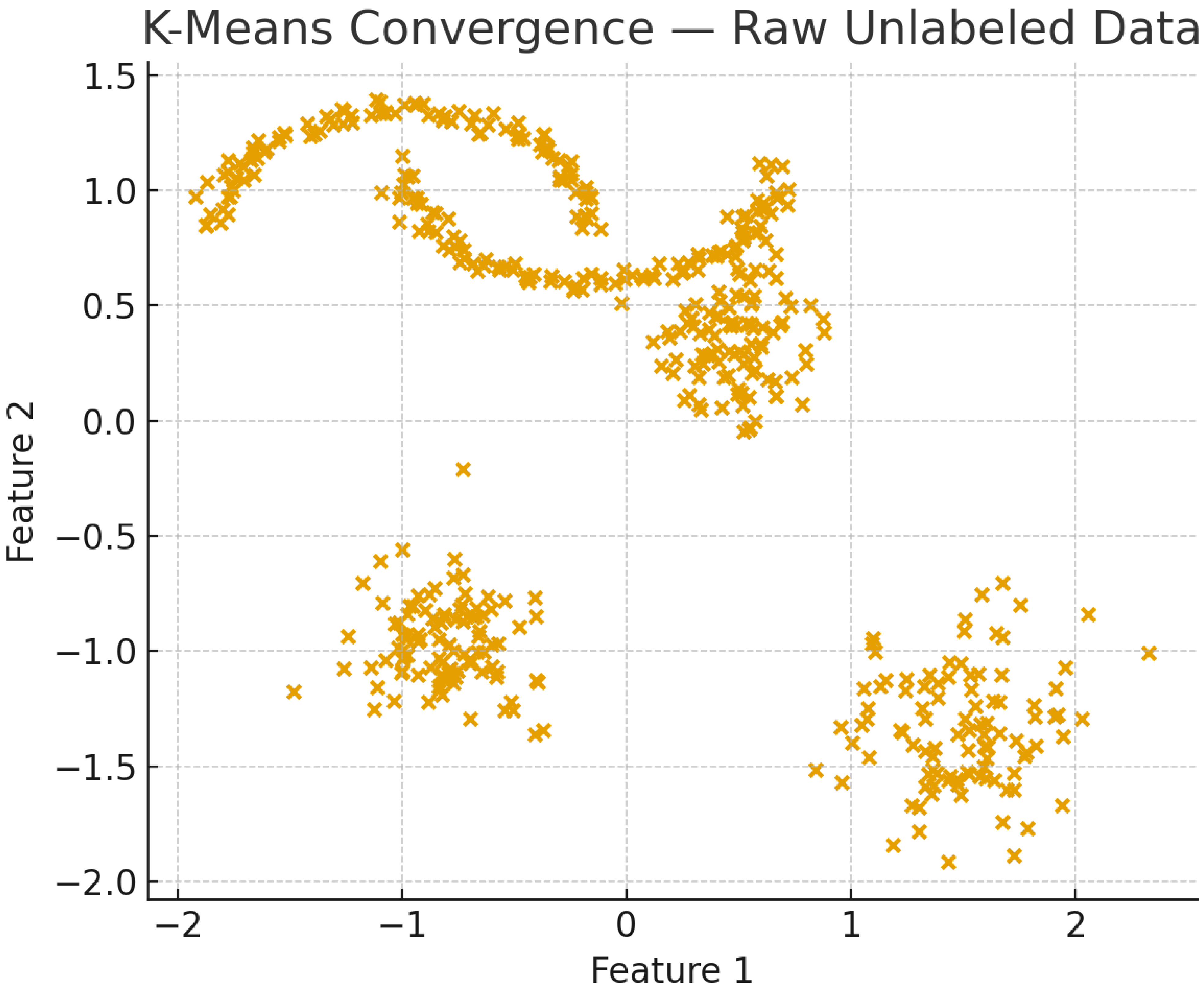
1. Initialization: choose (k) centroids (random, K-Means++).
2. Assignment: $c(\mathbf{x}) = \arg\min_j |\mathbf{x} - \boldsymbol{\mu}_j|$.
3. Update: $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$
4. Repeat until centroid movement ≈ 0.

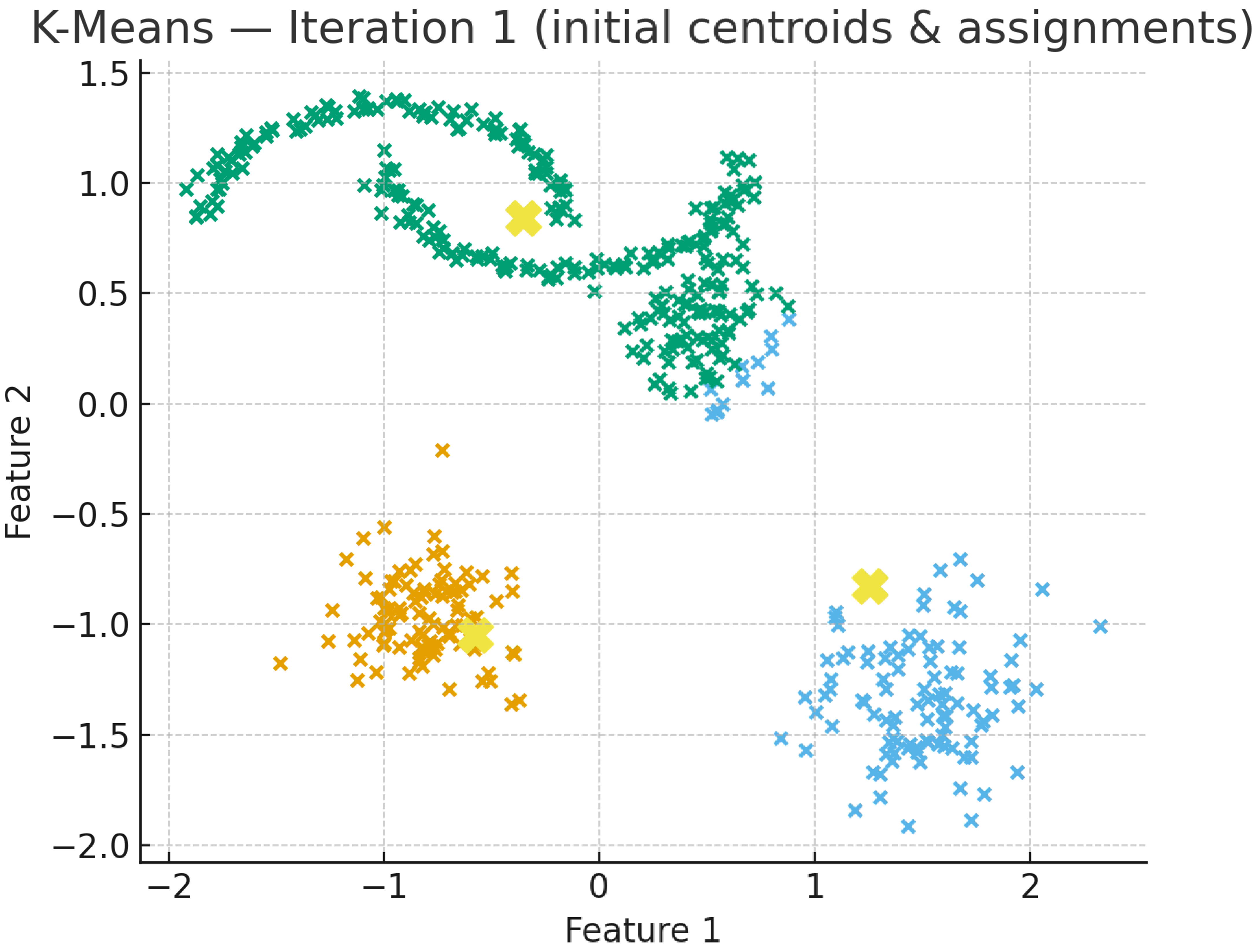
Complexity

O(nkd,i) for (n) samples, (k) clusters, (d) dims, (i) iterations.

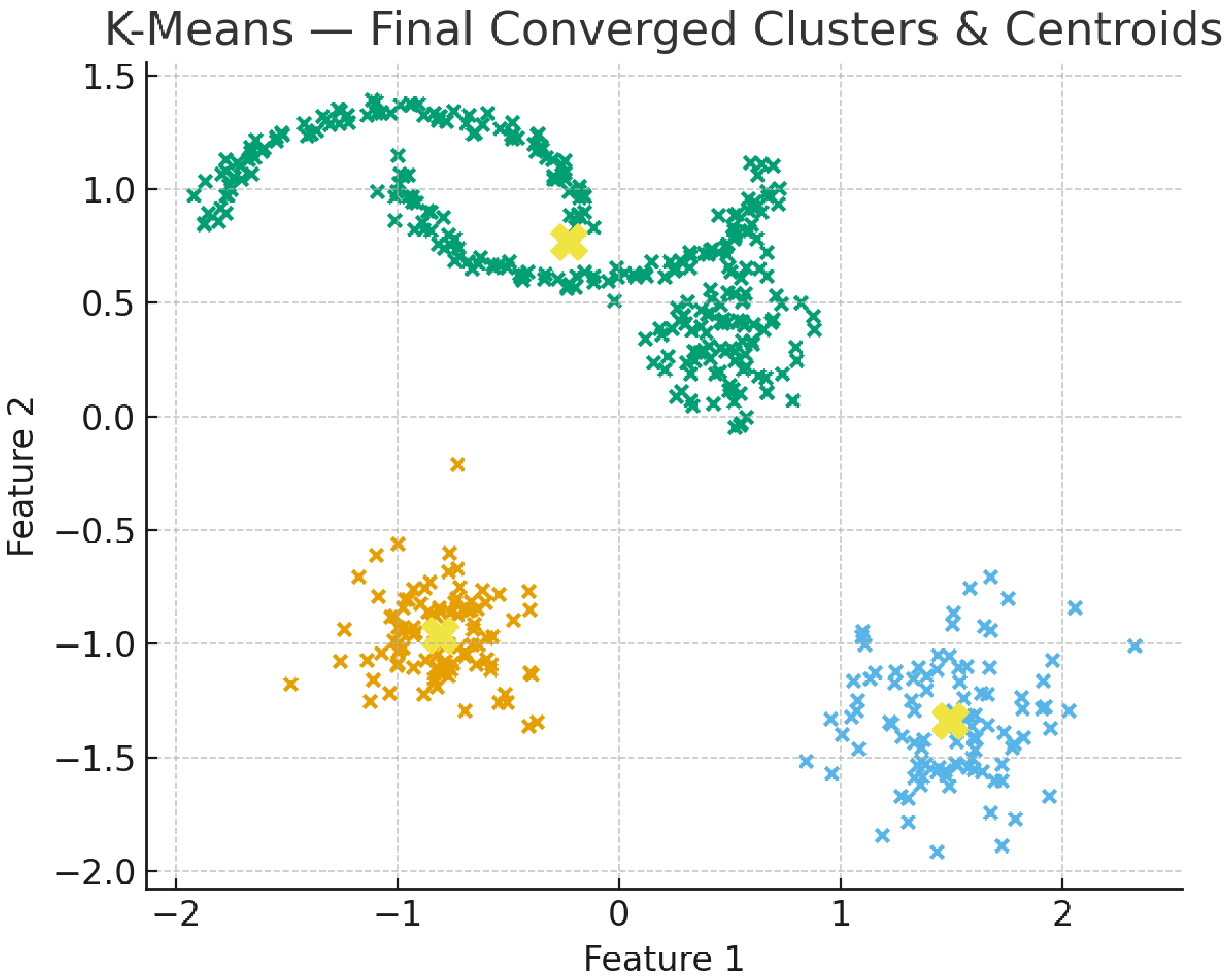
Strengths / Limitations

fast and simple; assumes spherical clusters, sensitive to initialization and chosen (k).


K-Means Convergence — Raw Unlabeled Data
Feature 2
Feature 1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2



K-Means — Iteration 1 (initial centroids & assignments)
Feature 2
Feature 1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2


K-Means — Final Converged Clusters & Centroids
Feature 2
Feature 1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2

Hierarchical Clustering & DBSCAN

Hierarchical clustering (dendrograms) and density-based clustering (DBSCAN)

1

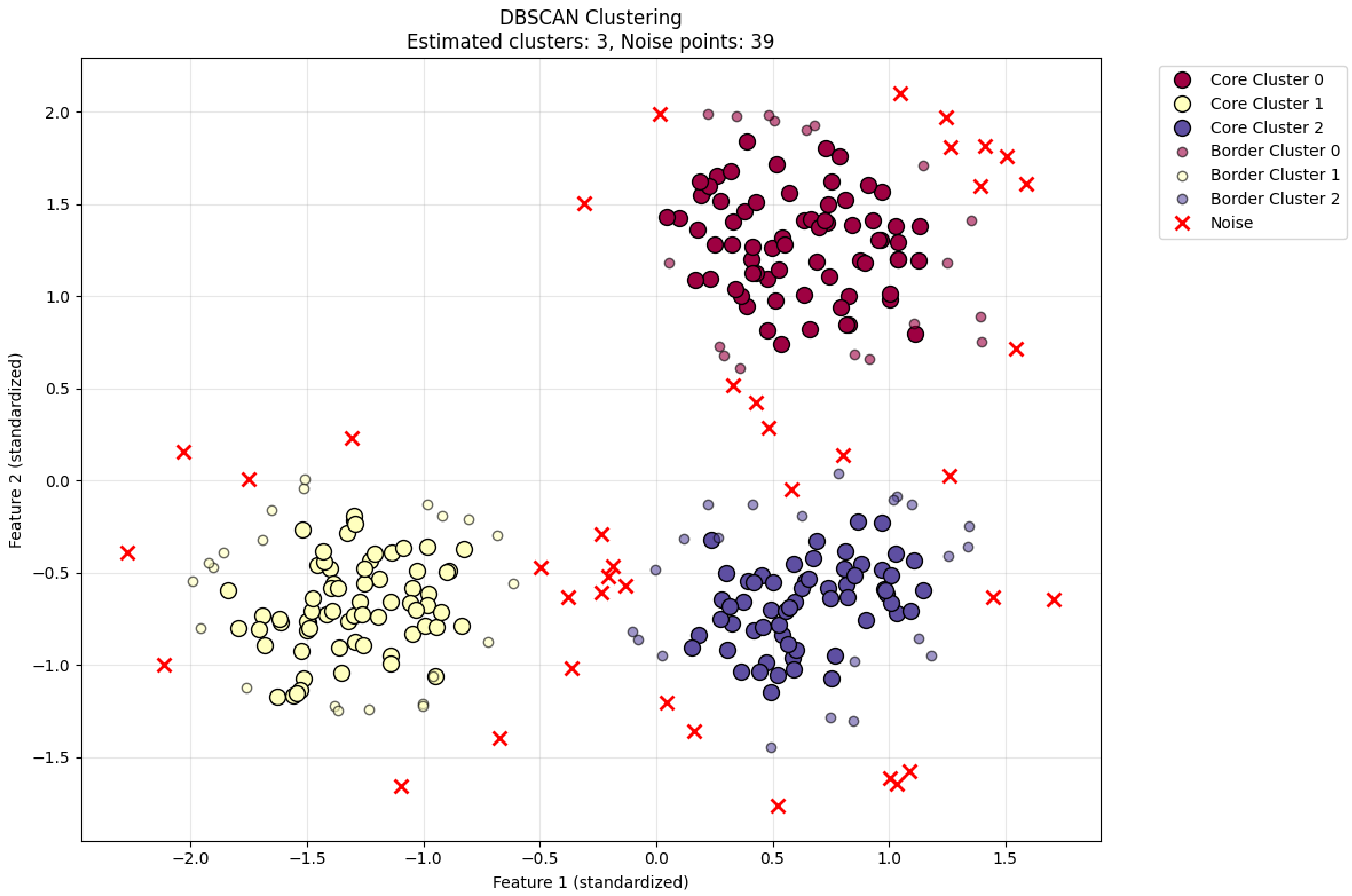
Hierarchical (tree-based)

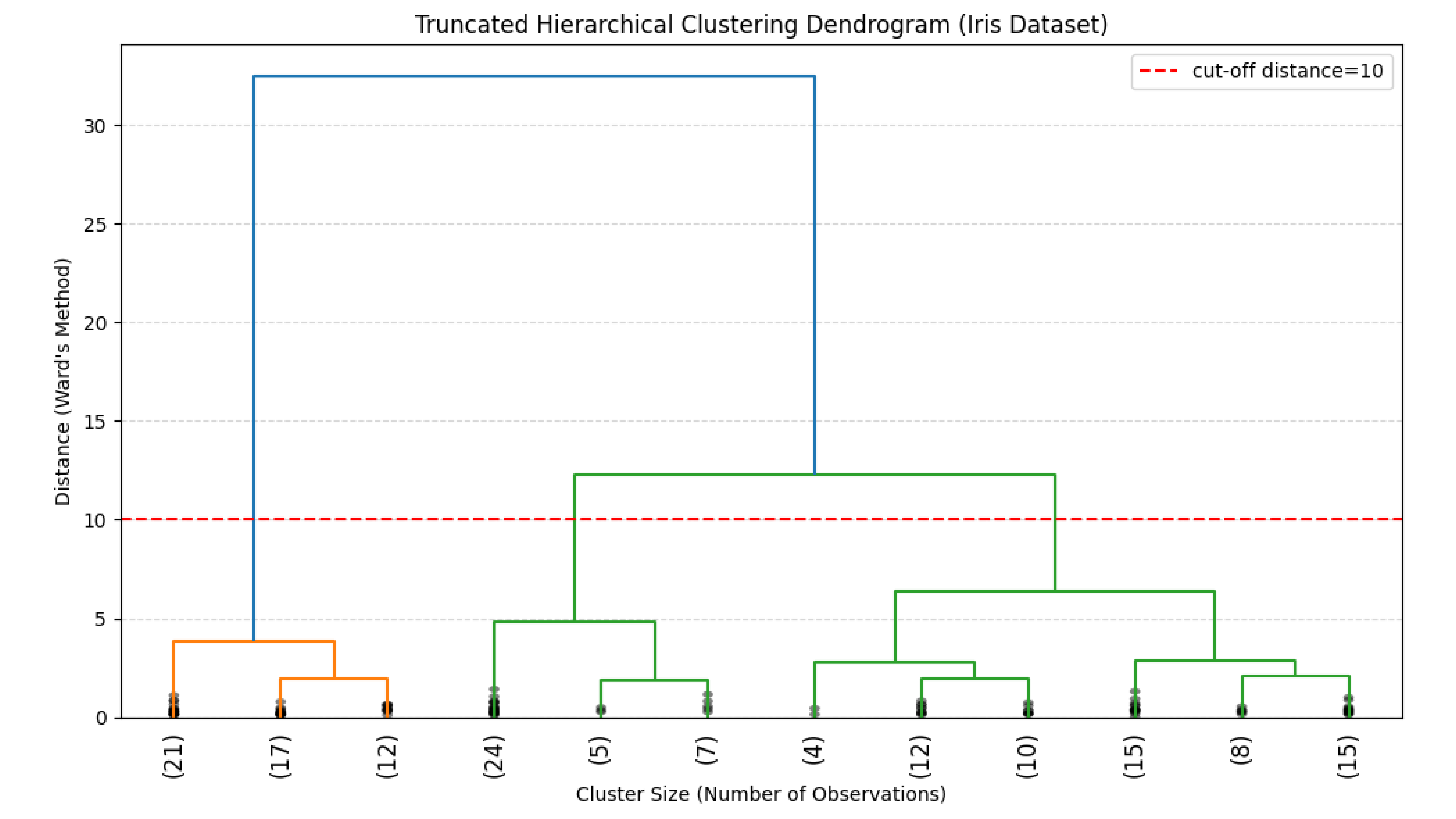
- **Concept:** produce a nested cluster tree (dendrogram); no fixed (k) required.
- **Agglomerative (bottom-up):** start with singleton clusters and merge by linkage:
  - **Linkages:** Ward (minimize increase in WCSS), complete, average, single.
- **Divisive (top-down):** recursively split the dataset.
- **Output:** dendrogram; choose cut height to extract clusters.

2

DBSCAN (density-based)

- **Parameters:** radius ε and $min\_samples$.
- **Point types:**
  - **Core:** $|\mathcal{N}_\varepsilon(\mathbf{x})| \geq \text{min\_samples}$.
  - **Border:** $\varepsilon - neighborhood$ of a core but not a core.
  - **Noise:** neither core nor border.
- **Mechanics:** expand clusters from core points by density connectivity — finds arbitrary-shaped clusters and flags outliers.
- **Pros / Cons:** robust to noise; no (k) required; sensitive to ε and varying density.


DBSCAN Clustering
Estimated clusters: 3, Noise points: 39
Core Cluster 0
Core Cluster 1
Core Cluster 2
Border Cluster 0
Border Cluster 1
Border Cluster 2
Noise
Feature 1 (standardized)
Feature 2 (standardized)
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Truncated Hierarchical Clustering Dendrogram (Iris Dataset)
cut-off distance=10
Distance (Ward's Method)
0
5
10
15
20
25
30
(21)
(17)
(12)
(24)
(5)
(7)
(4)
(12)
(10)
(15)
(8)
(15)
Cluster Size (Number of Observations)

Internal Clustering Evaluation Metrics

To assess the quality of clustering results objectively, internal metrics evaluate the structure inherent in the data without requiring external ground-truth labels.

1

Silhouette Coefficient

Measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation), based on pairwise distances.

2

Davies-Bouldin Index

Evaluates the average "similarity" between each cluster and its most similar one, considering the dispersion within clusters and separation between their centroids.

3

Calinski-Harabasz (CH) Index

Calculates the ratio of between-clusters dispersion to within-cluster dispersion across all clusters, analogous to an F-statistic in ANOVA.

Internal clustering evaluation

Internal metrics (no ground truth)

1

Silhouette coefficient

Per-sample:

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)},$$

where *a(i)* = mean intra-cluster distance, *b(i)* = mean nearest-cluster distance.

Range: (-1) (misplaced) to (+1) (well separated). Use the mean *s(i)* as a global score.

2

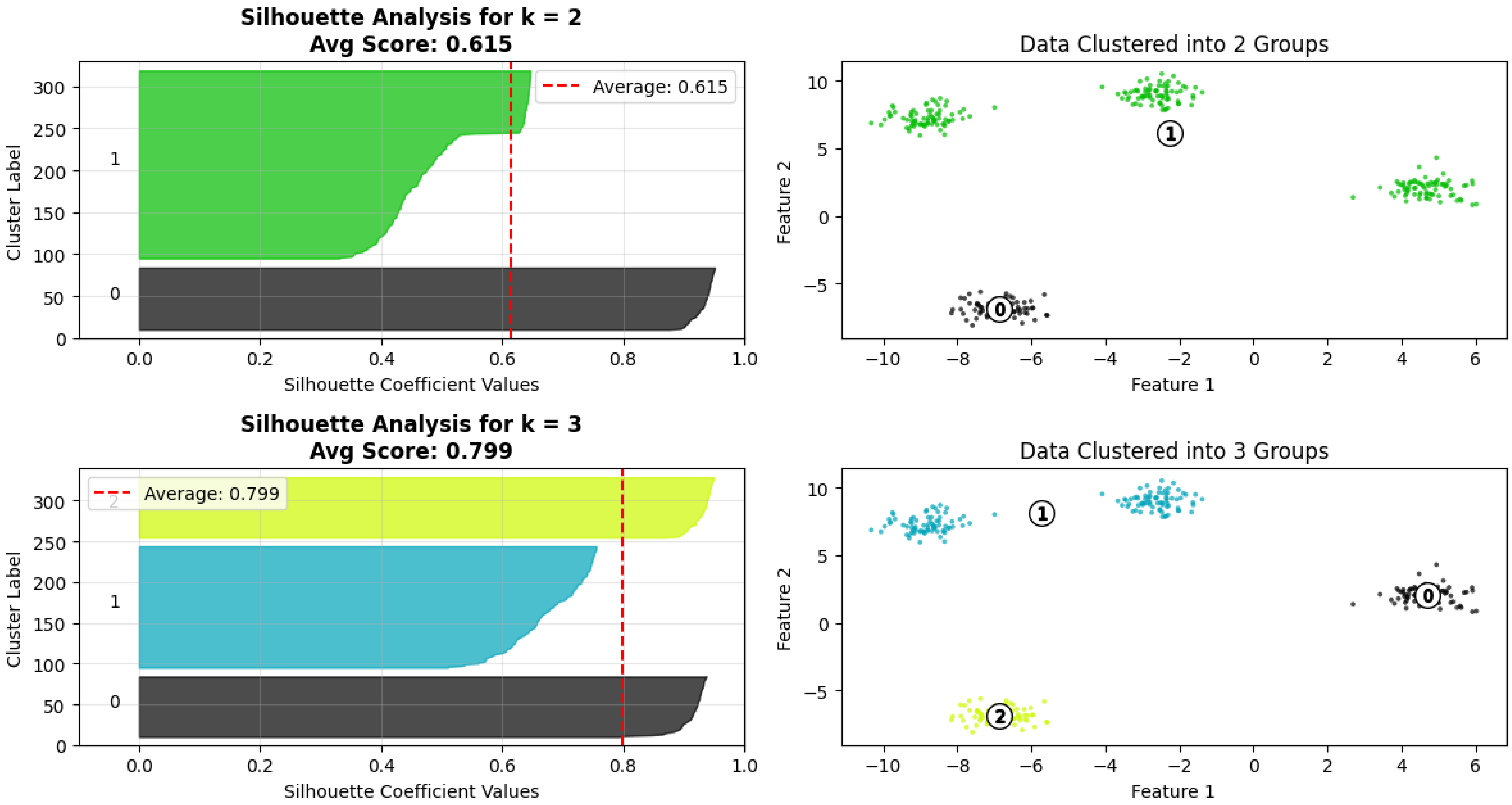
Davies–Bouldin index (DB)

For (k) clusters:

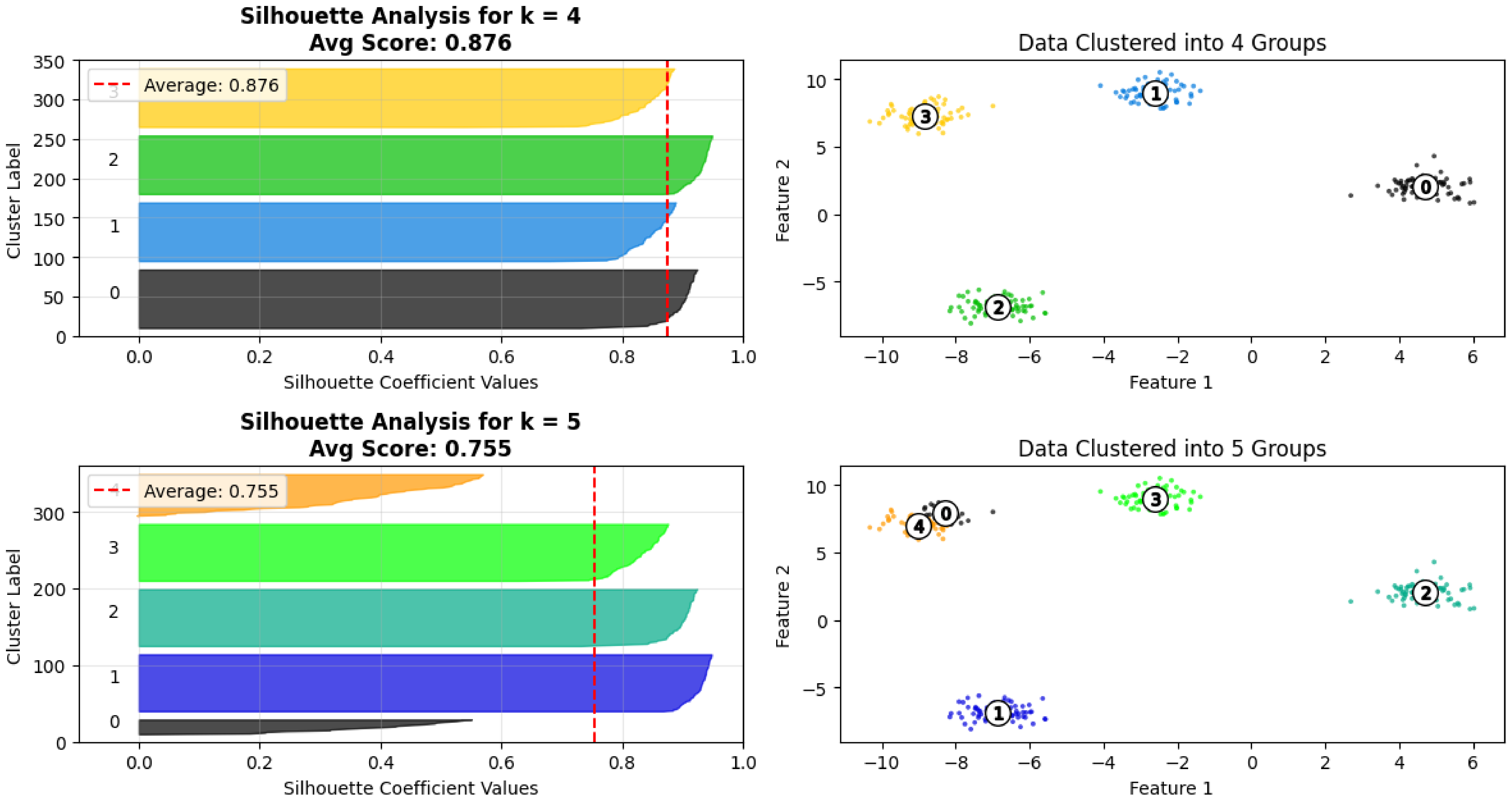
$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j=i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)},$$

where $\sigma_i = average\ distance$ of cluster (i) points to centroid $c_i, d(\cdot,\cdot) = centroiddistance.$

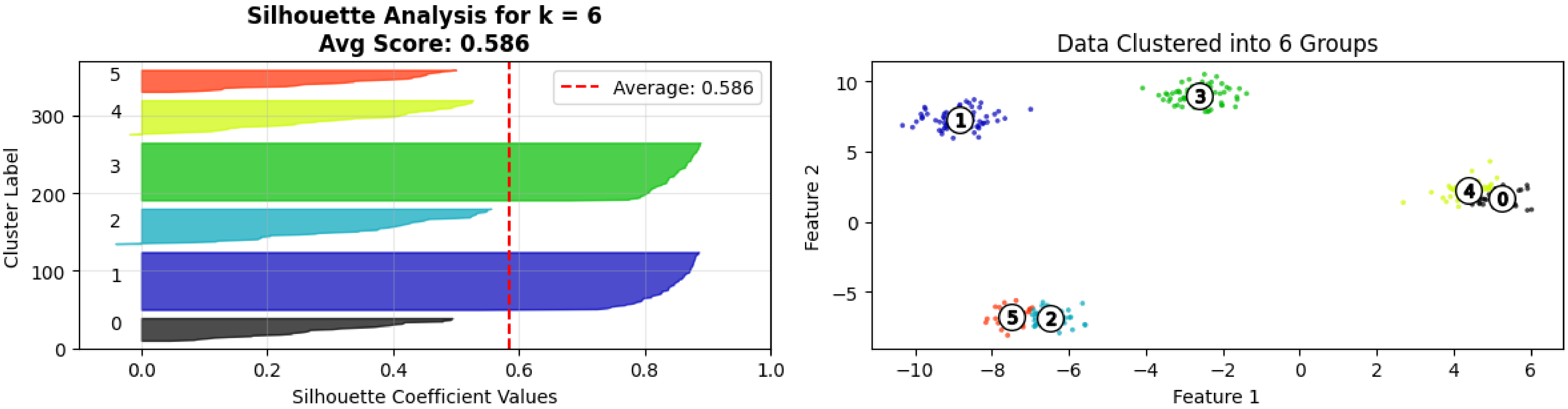
Interpretation: lower DB = better (compact & well separated).



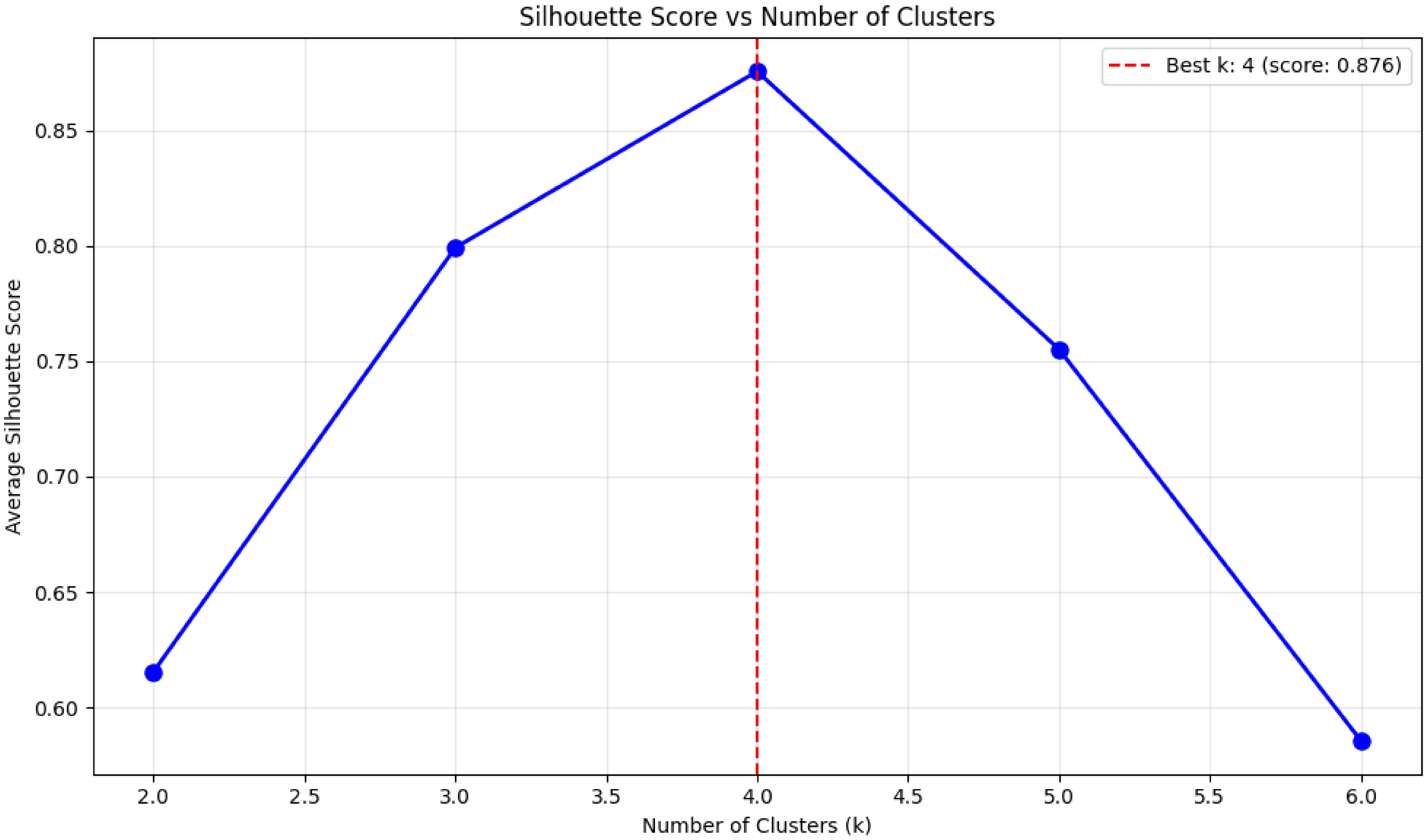
Silhouette Analysis for k = 2
Avg Score: 0.615
Average: 0.615
Cluster Label
Silhouette Coefficient Values
Data Clustered into 2 Groups
Feature 1
Feature 2
Silhouette Analysis for k = 3
Avg Score: 0.799
Average: 0.799
Cluster Label
Silhouette Coefficient Values
Data Clustered into 3 Groups
Feature 1
Feature 2



Silhouette Analysis for k = 4
Avg Score: 0.876
Average: 0.876
Cluster Label
Silhouette Coefficient Values
Data Clustered into 4 Groups
Feature 1
Feature 2
Silhouette Analysis for k = 5
Avg Score: 0.755
Average: 0.755
Cluster Label
Silhouette Coefficient Values
Data Clustered into 5 Groups
Feature 1
Feature 2



Silhouette Analysis for k = 6
Avg Score: 0.586
Average: 0.586
Cluster Label
Silhouette Coefficient Values
0
1
2
3
4
5
0
100
200
300
0.0
0.2
0.4
0.6
0.8
1.0
Data Clustered into 6 Groups
Feature 1
Feature 2
−10
−8
−6
−4
−2
0
2
4
6
10
5
0
−5


Silhouette Score vs Number of Clusters
Best k: 4 (score: 0.876)
Average Silhouette Score
0.85
0.80
0.75
0.70
0.65
0.60
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Number of Clusters (k)

Additional metrics & practical comparison

Calinski–Harabasz and metric comparison notes

Calinski–Harabasz (CH)

Definition:

$$CH = \frac{\mathrm{trace}(B_k)/(k-1)}{\mathrm{trace}(W_k)/(n-k)}$$

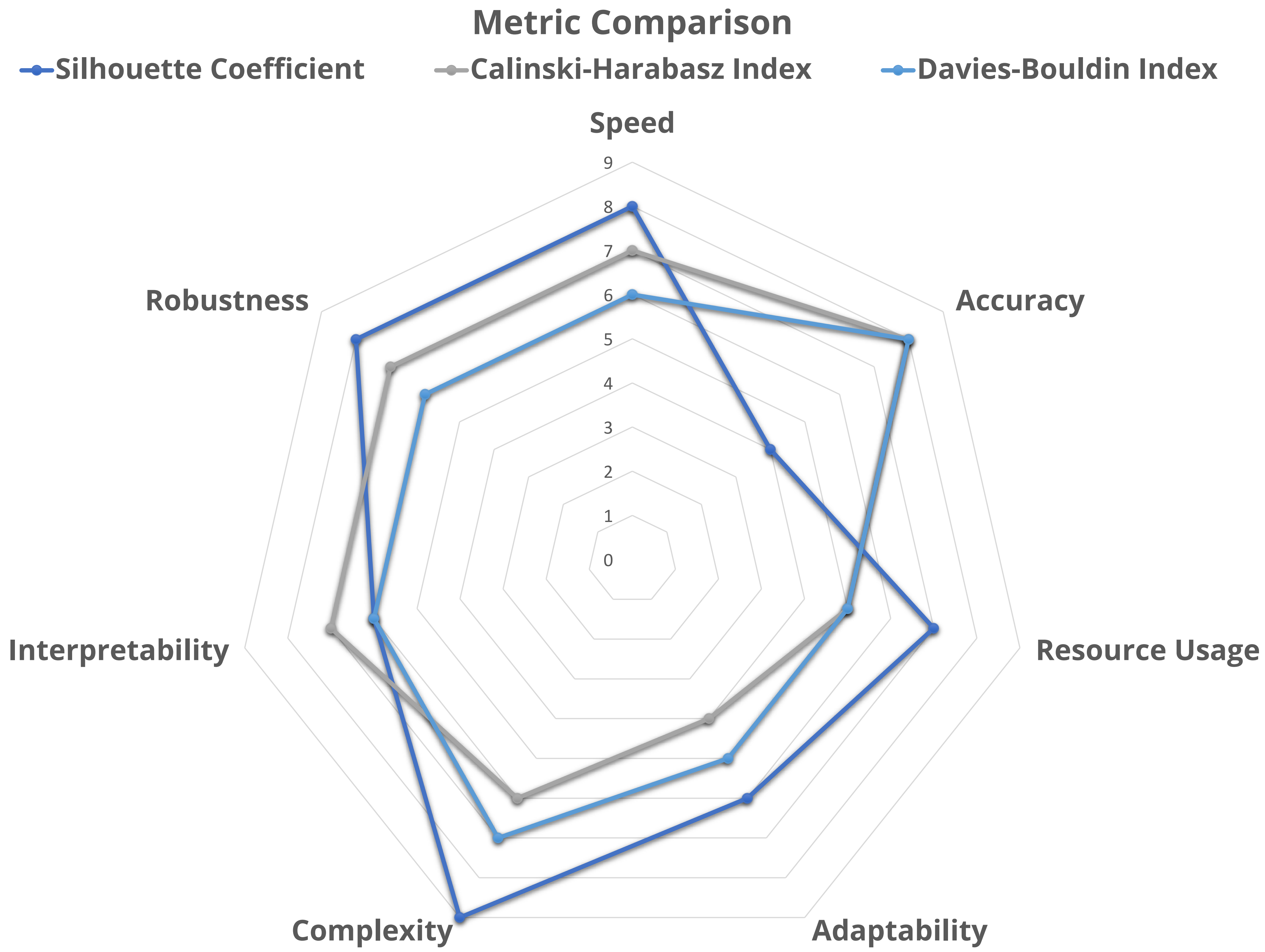
where B_k = between-cluster dispersion matrix, W_k = within-cluster dispersion matrix.

Interpretation: larger CH = better separated, compact clusters.

Metric goals (practical checklist)

- **Maximize:** Silhouette, CH (higher → better cohesion/separation).
- **Minimize:** Davies–Bouldin (lower → better).

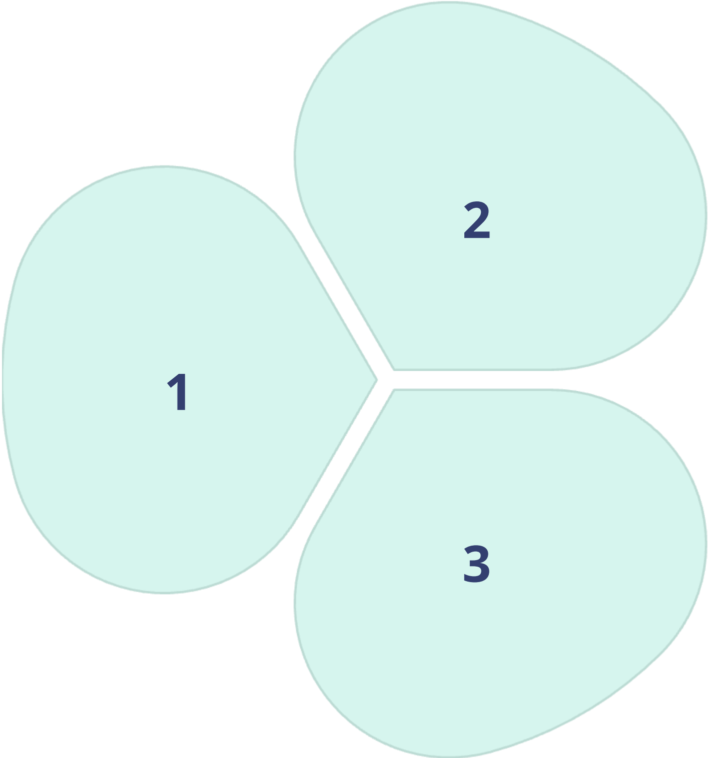
Caveat: metrics have different numeric scales—compare relative rankings or normalize before aggregation.


Metric Comparison
Silhouette Coefficient
Calinski-Harabasz Index
Davies-Bouldin Index
Speed
Accuracy
Resource Usage
Adaptability
Complexity
Interpretability
Robustness
0
1
2
3
4
5
6
7
8
9

Applications & concise conclusion

Key applications and takeaways

Applications



Anomaly detection

treat DBSCAN noise or low-density / far-from-centroid points as anomalies.

Recommendation systems

user segmentation (cluster users by behavior) and item clustering (recommend within same cluster).

Feature engineering & EDA

clustering to derive categorical features or reduce label sparsity.

Conclusion

Clustering converts unlabelled data into structure—choose algorithm and metric based on shape, density, interpretability, and downstream needs.