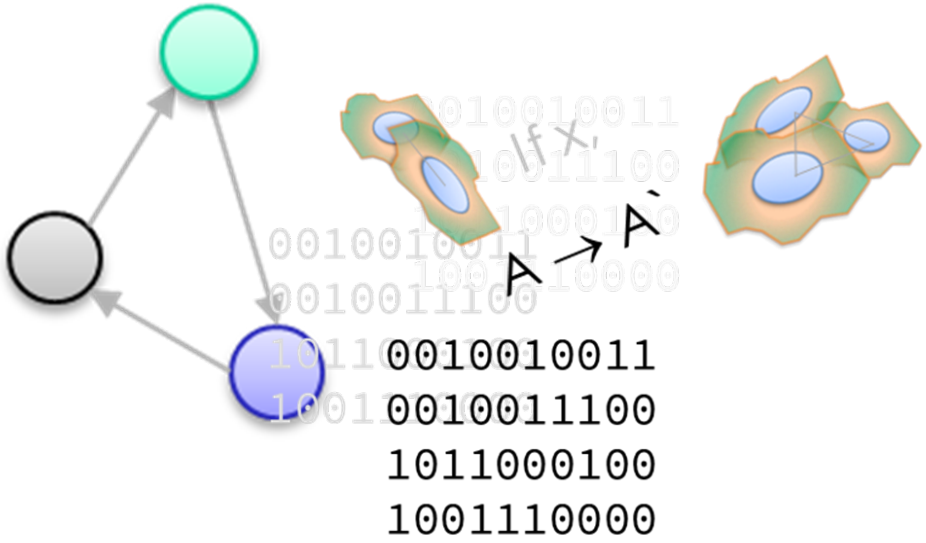
Computational Bioengineering
BME 4803 | BME 6313



Clustering Introduction

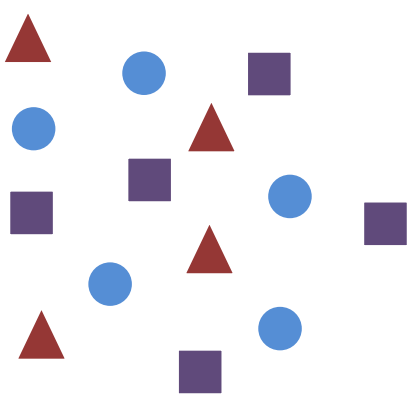
Slide Credits:

Dr. Chenyue Wendy Hu

Outline

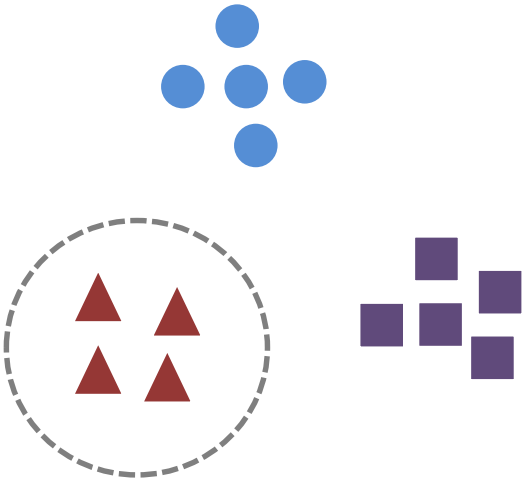
- What is cluster analysis?
  - concept, examples, application scenarios
- How does cluster analysis work?
  - Similarity (Euclidean distance)
  - K-means Clustering
  - Hierarchical Clustering
- How to pick the cluster number?
  - Elbow Method
  - Silhouette
- Best practices & review

Concept



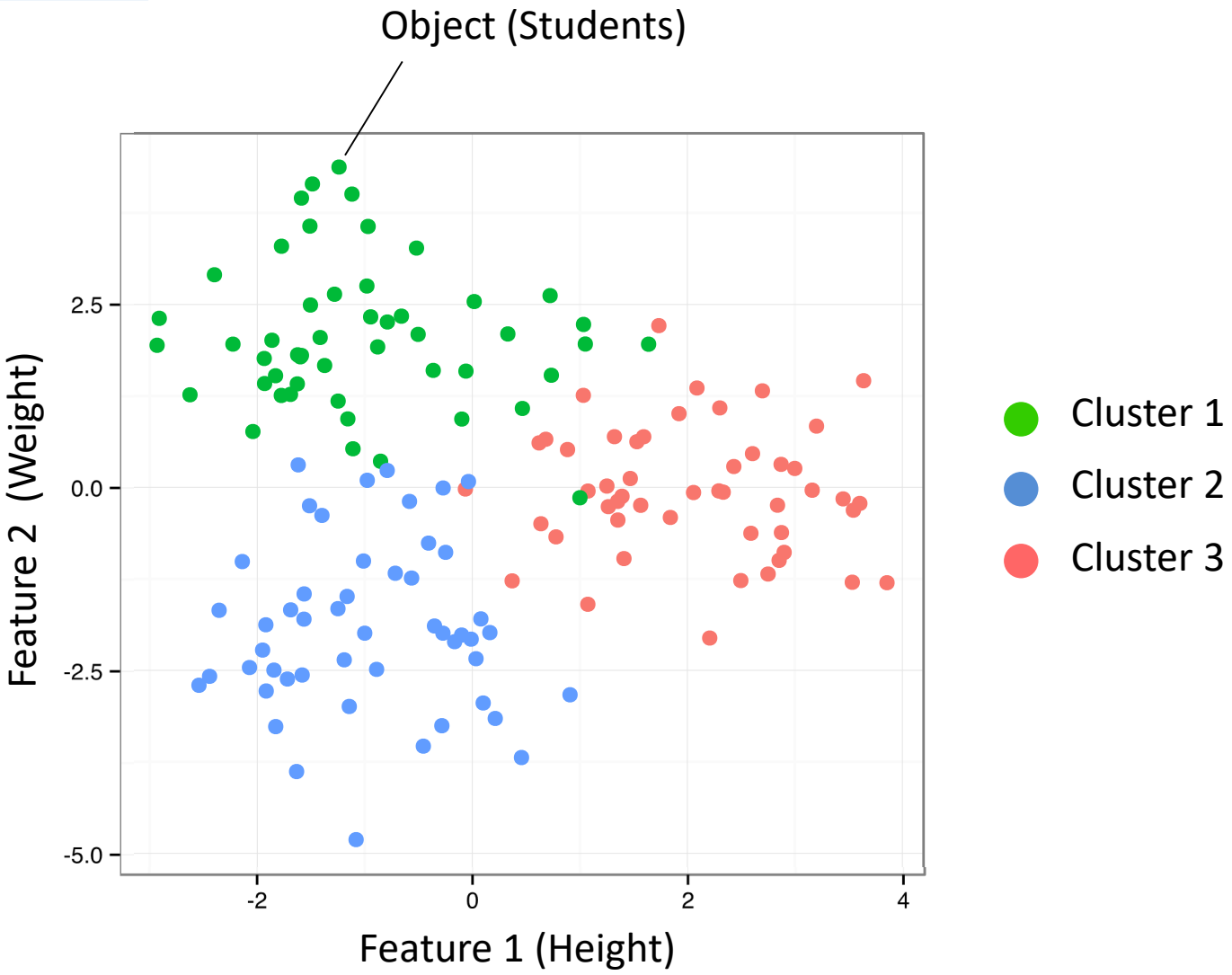
A set of objects
with known features

Group

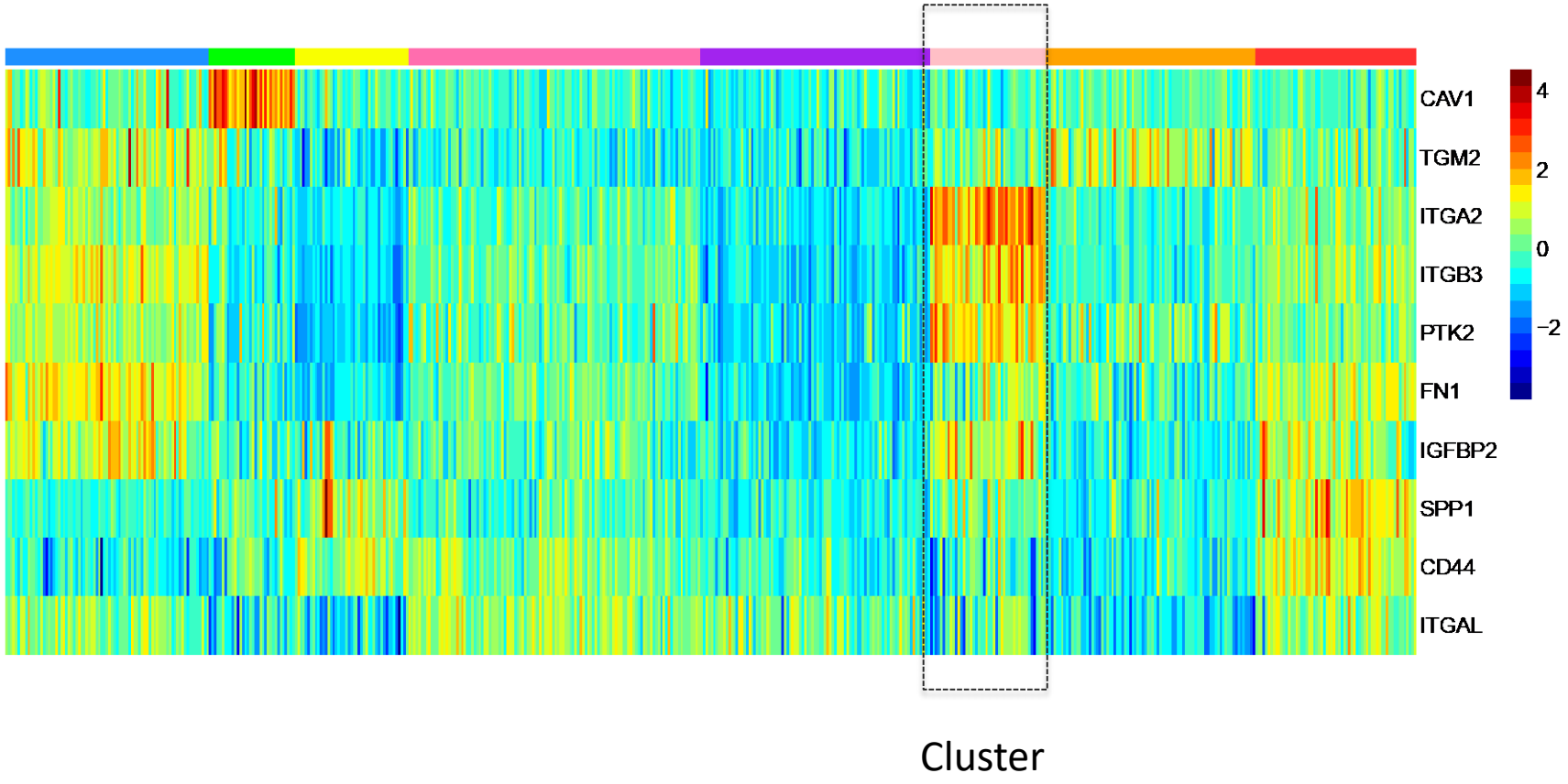


Cluster
Of similar objects

Example 1


Object (Students)
Feature 2 (Weight)
2.5
0.0
-2.5
-5.0
-2
0
2
4
Feature 1 (Height)
Cluster 1
Cluster 2
Cluster 3

Example 2


CAV1
TGM2
ITGA2
ITGB3
PTK2
FN1
IGFBP2
SPP1
CD44
ITGAL
4
2
0
−2
Cluster

Application

Application scenarios

- You have multiple features (> 1)
  - matrix or data table
- You believe the population is heterogeneous
- You don’t know the true grouping of objects

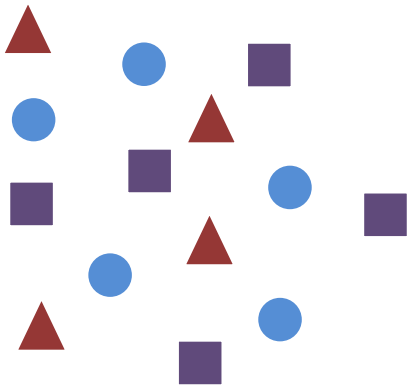
Application goals

✓ End-step: Define classes or categories

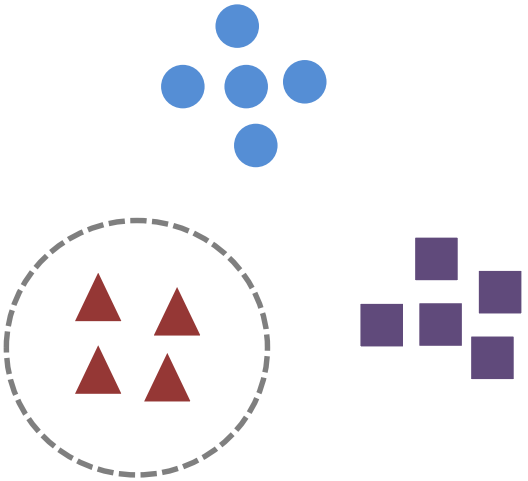
✓ Pre-step: Segment samples for subsequent analyses

Concept

Clustering Algorithm



Group



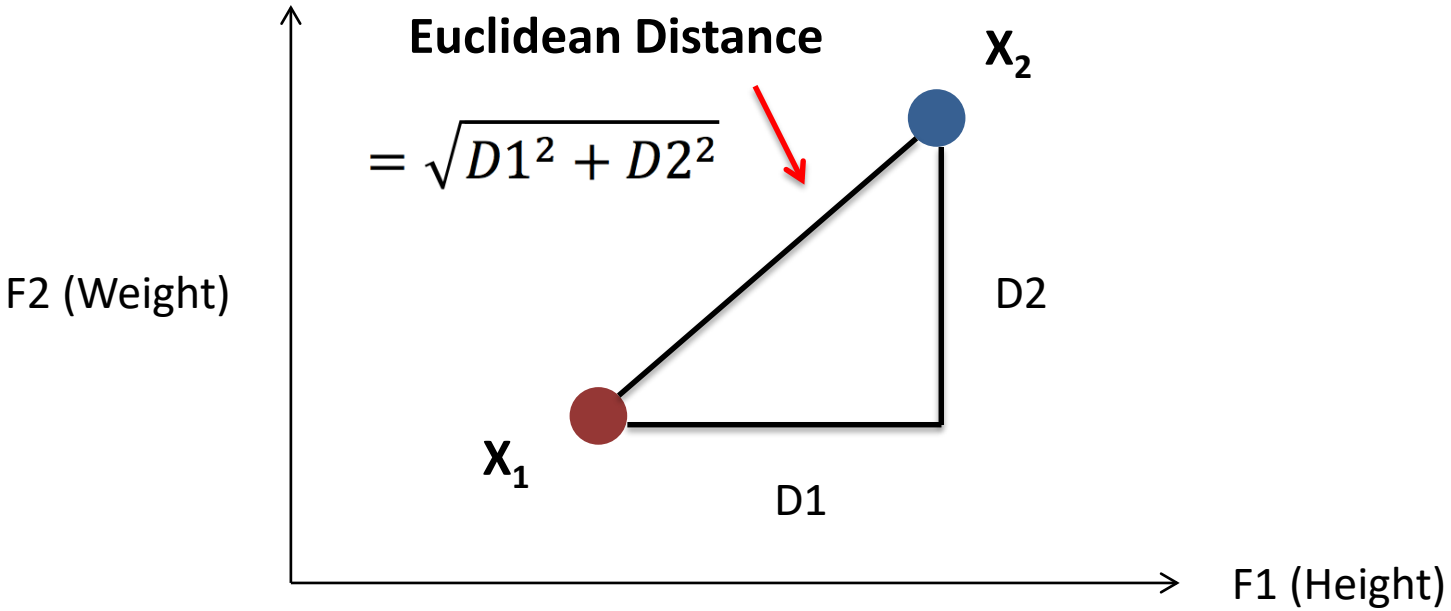
A set of objects with known features

Cluster Of similar objects

Defining similarity / distance

Similarity

Distance (Dissimilarity)



Generalized Euclidean Distance
Between X_i and X_j
In n dimensional space

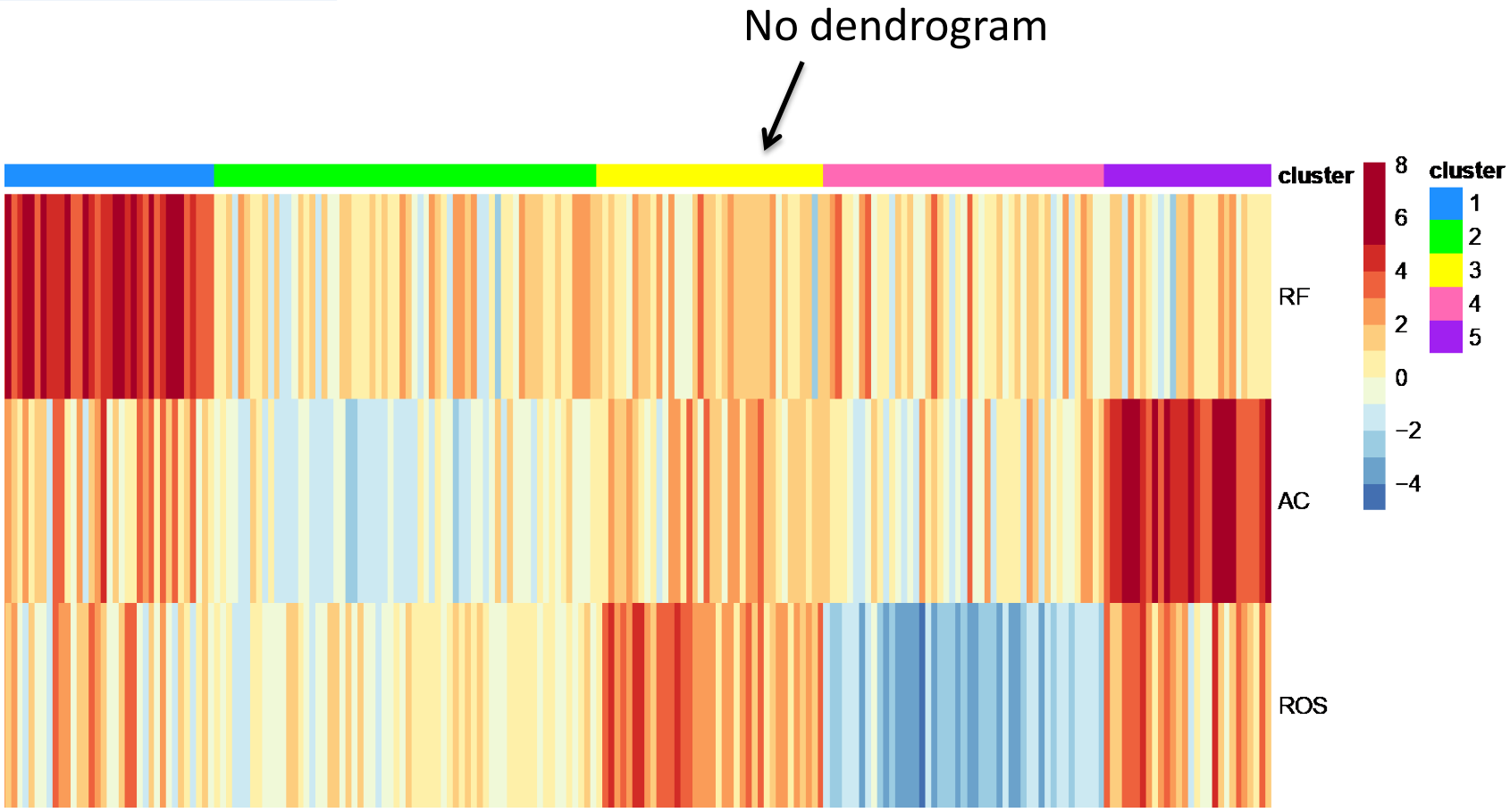
$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(D_{ik} - D_{jk}\right)^2}$$

Algorithms

Common Clustering Algorithms

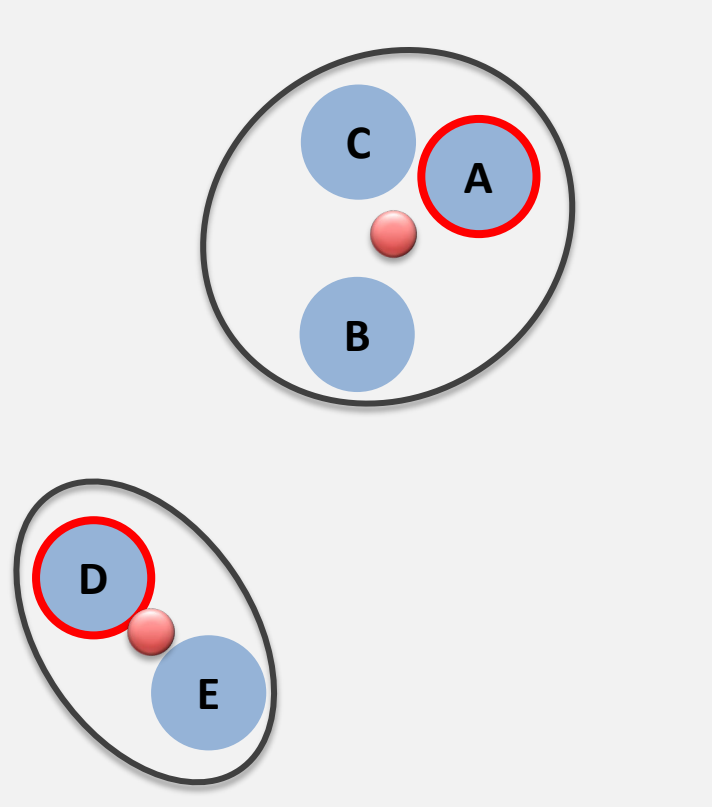
- **Hierarchical Clustering**
- **K-means**
- Affinity Propagation
- Spectral Clustering
- DBSCAN

K-means



K-means

- Start with K initial random cluster centers
- Repeat the following steps:
  - Assign objects to their closest cluster centers
  - Recompute cluster centers
- End when the solution is optimized (nothing to do)



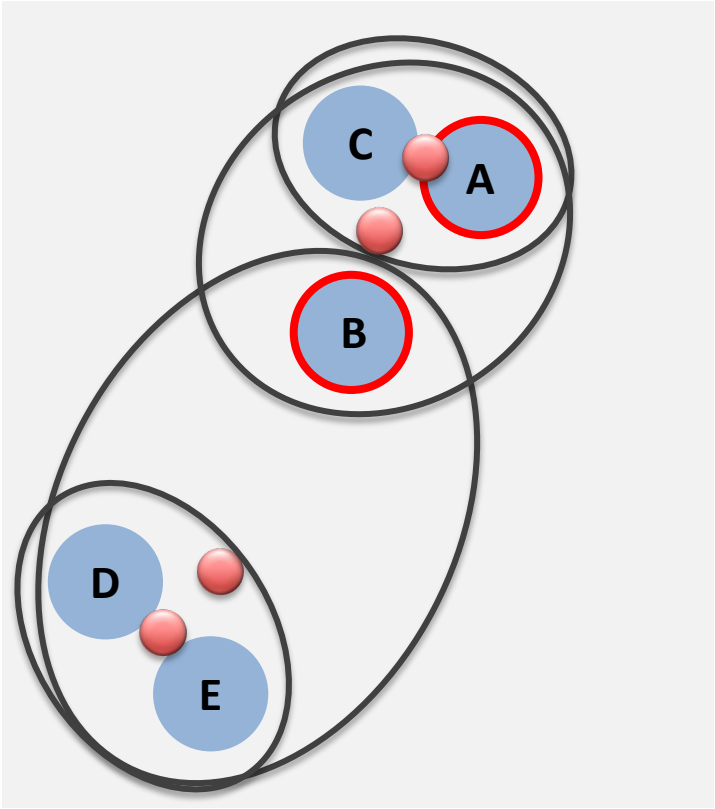
Initialization: A, D

Cluster: {A, B, C}, {D, E}

Cluster: {A, B, C}, {D, E}

K-means

- Start with K initial random cluster centers
- Repeat the following steps:
  - Assign objects to their closest cluster centers
  - Recompute cluster centers
- End when the solution is optimized (nothing to do)



Initialization: A, D

Cluster: {A, B, C}, {D, E}

Cluster: {A, B, C}, {D, E}

Initialization: A, B

Cluster: {A, C}, {B, D, E}

Cluster: {A, B, C}, {D, E}

Cluster: {A, B, C}, {D, E}

K-means

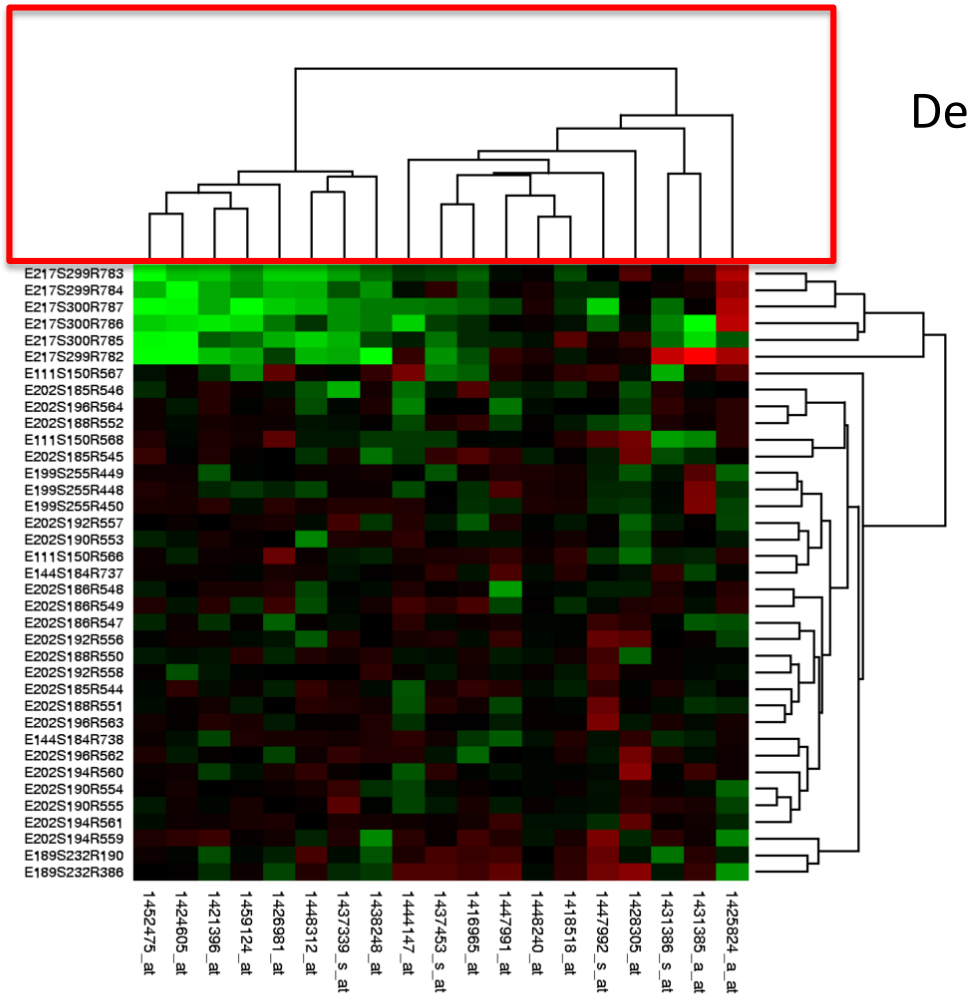
Summary

- Need to provide K (number of clusters)
- Iterative assignment and center update
- Optimize within-cluster distances
- Converge to local optimum

- Not necessarily the global optimum
- Different random initiations -> different solutions
- Make sure to run the algorithm multiple times (>=100)

Slower than Hierarchical, but better clustering quality

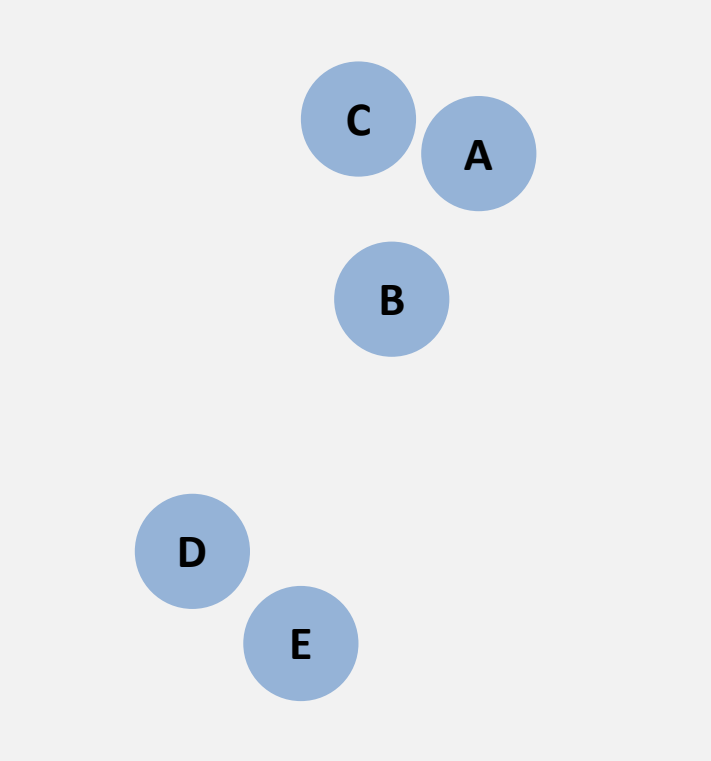
Hierarchical

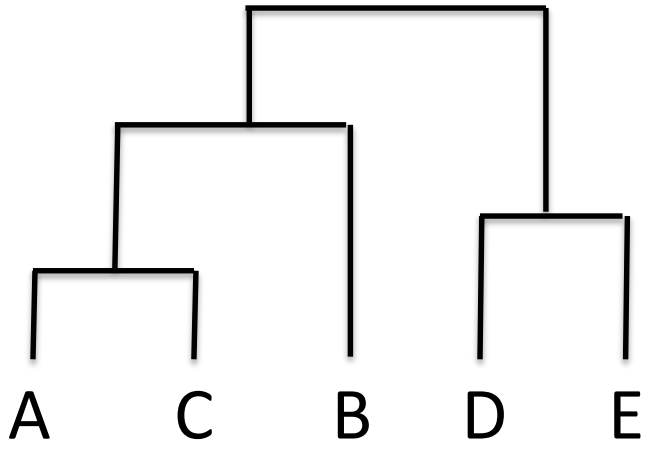


Dendrogram

Hierarchical

- Start with each object as a cluster
- Repeatedly connect the closest two clusters
- End when there is only one cluster left

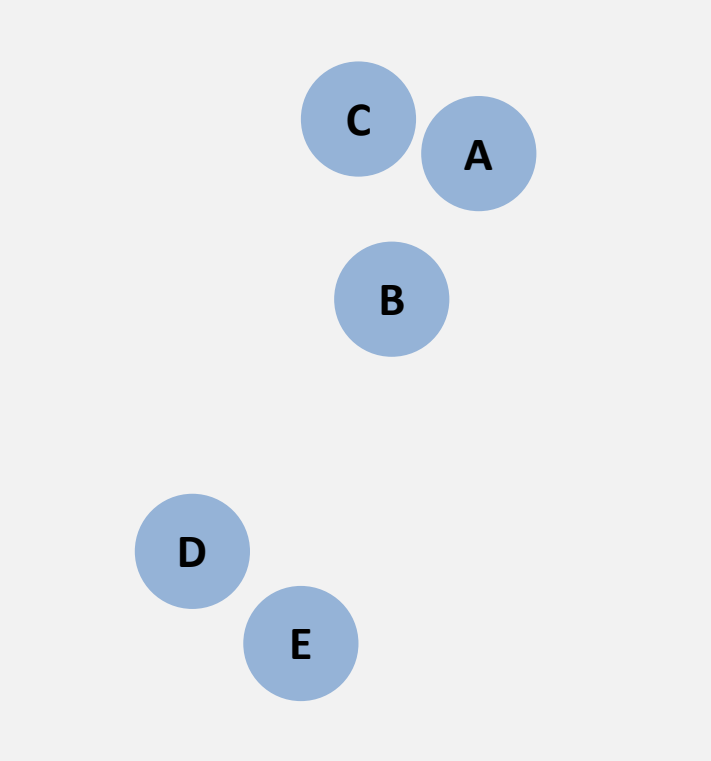


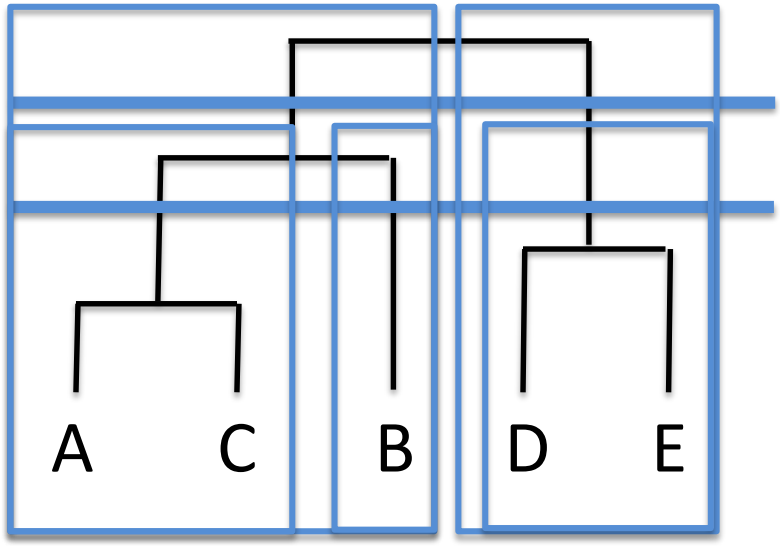


Minimum **Linkage**: the smallest distance between elements of each cluster

Hierarchical

Cut tree to obtain a clustering solution





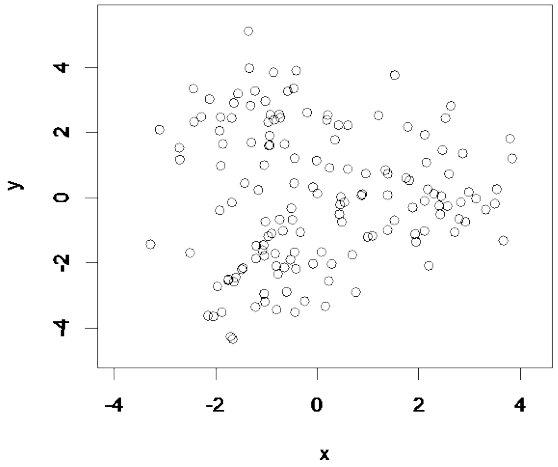
Hierarchical

Summary

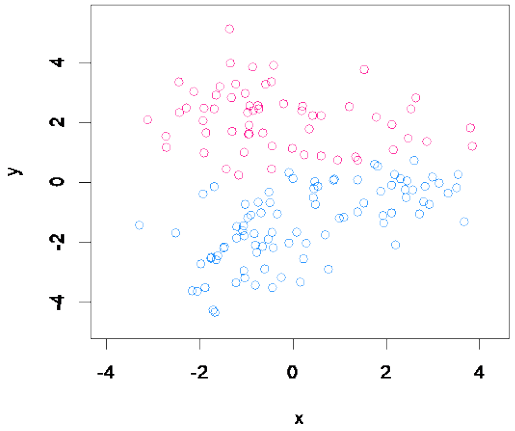
- Dendrogram (tree structure)
- Bottom-up or agglomerative approach
- Linkage Method
- Cut-tree

- Linkage Method Choice: Maximum (Complete), Minimum (Single), Average, Ward
- Pro: Good visualization, Fast Computing
- Con: Biased towards outliers

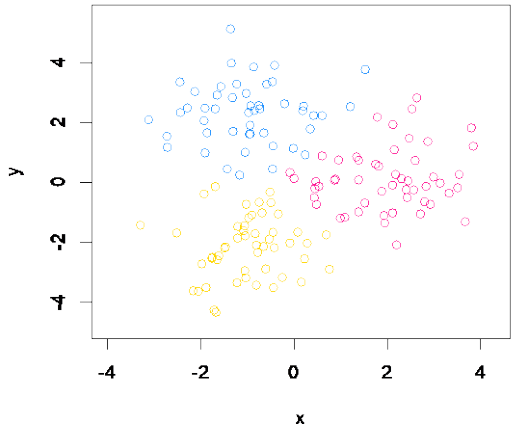
Cluster Number



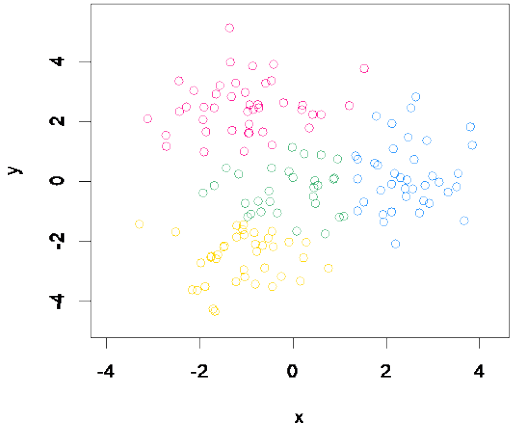
K=2



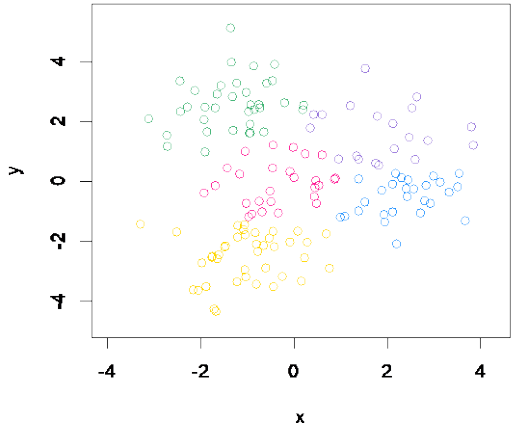
K=3



K=4



K=5



Cluster Number

Computational Methods

- Elbow Method
- Silhouette
- Gap Statistic
- Progeny Clustering*

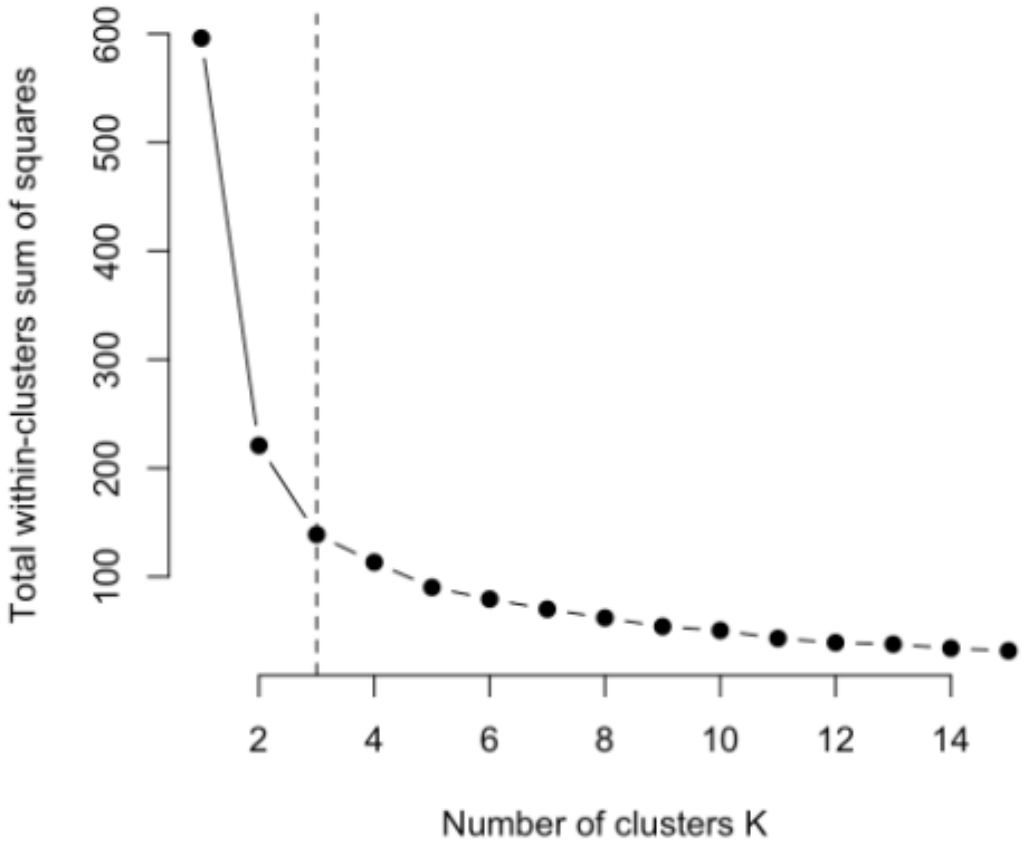
Procedure

1. Cluster data repeated with varying cluster numbers (K=2, …, 10)
2. Pick the number with best clustering quality

*Progeny Clustering (Hu et al.): https://www.nature.com/articles/srep12894

Cluster Number

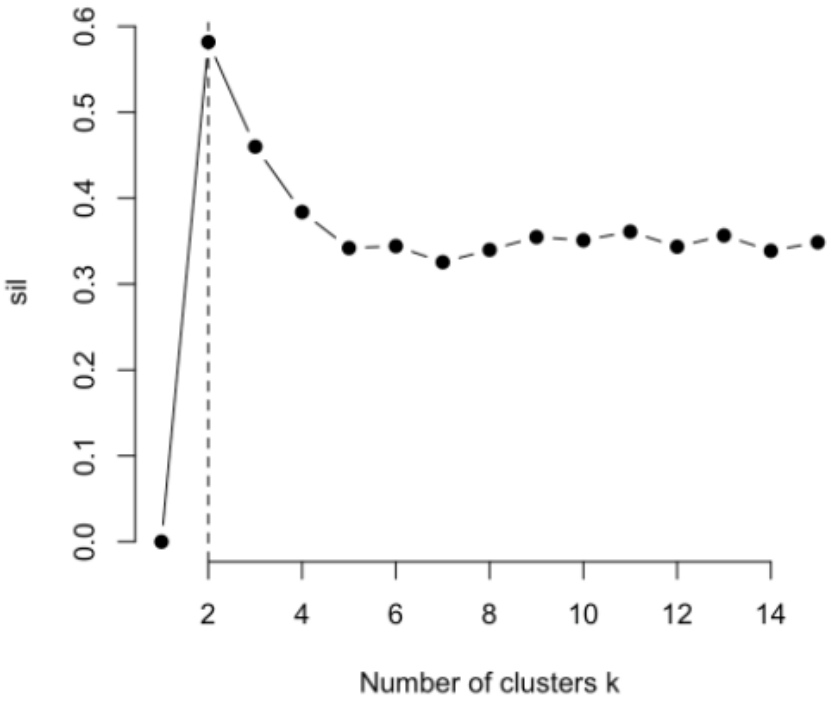
Elbow Method

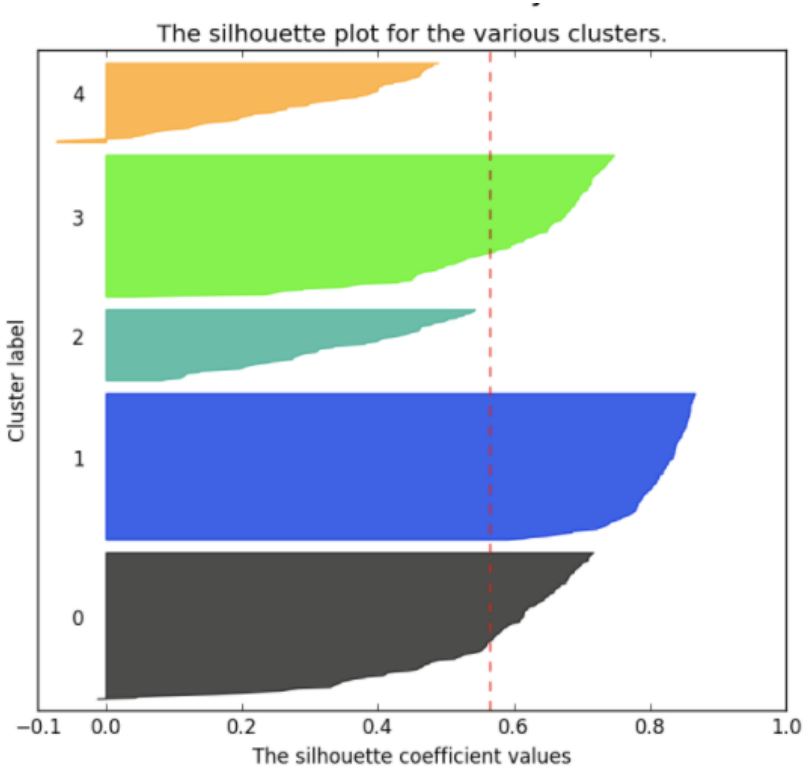


Cluster Number

Silhouette

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad -1 \leq s(i) \leq 1$$





Best Practice

- Data Normalization
  - Scaling -> Equal Variance
- Clustering Algorithm Choice
  - K-means with multiple runs
- Cluster Number Choice
  - Silhouette or your expert knowledge
- Visualization
  - Heatmaps with dendrograms or annotation bars
  - Scatter Plot with Principal Components (PCA)

Outline

- Concept
- Application
- Best Practices
- Circular Heatmap
- Tutorial

✓ what is a heatmap?

✓ when do I need a heatmap?

✓ how should I choose the color?

✓ what is a circular heatmap?

✓ I can draw heatmaps!