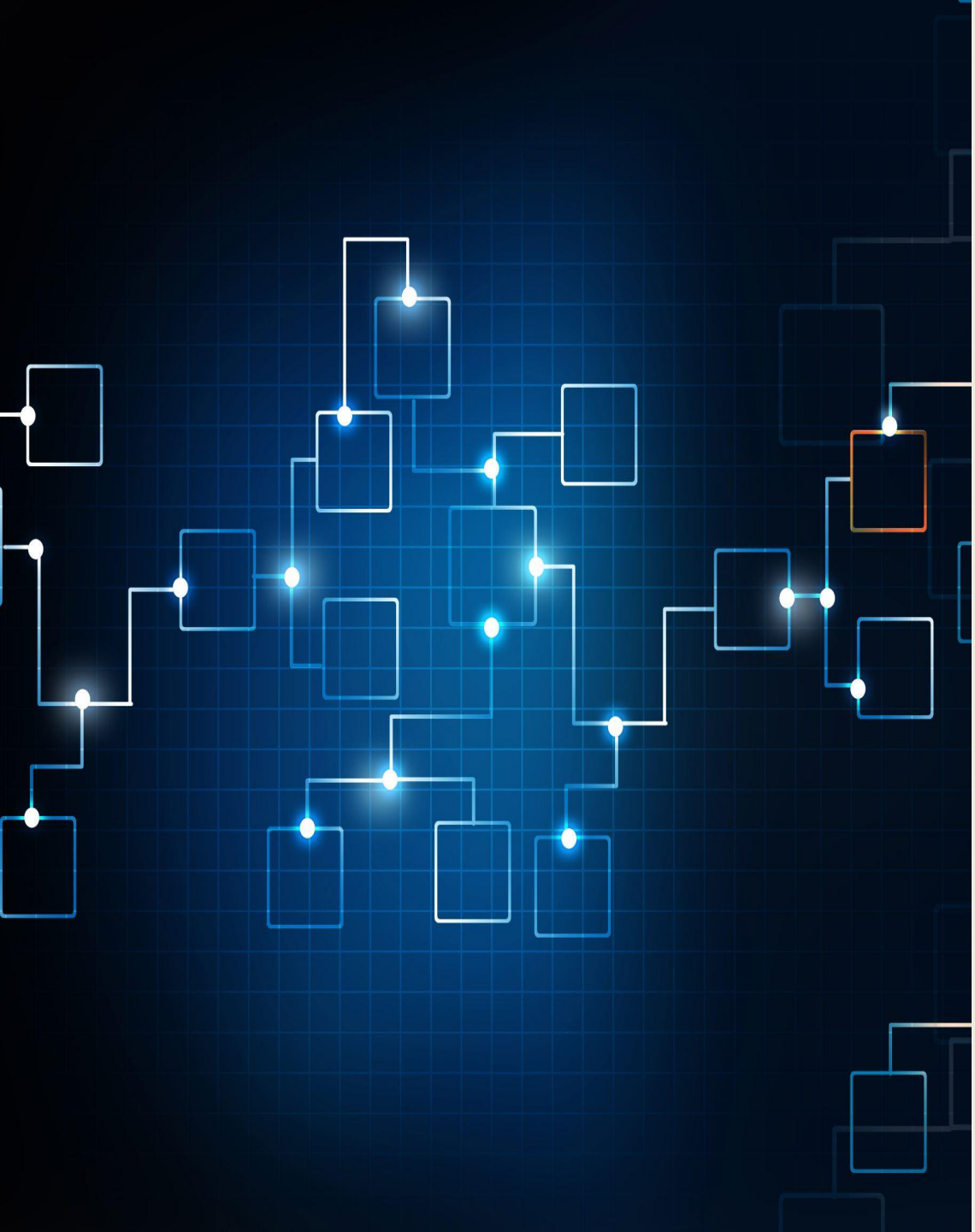


HIT391

MACHINE LEARNING: ADVANCEMENTS AND APPLICATIONS

- Lecturer: Dr. Yan Zhang
- Email: yan.zhang@cdu.edu.au
- Office: Level 6 Desk 5 Danana Camupus

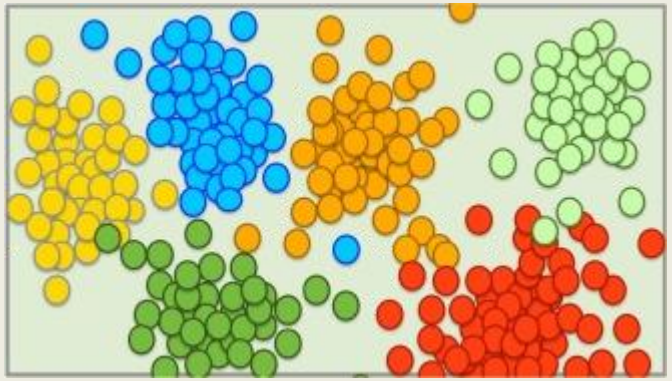
CHARLES DARWIN UNIVERSITY
AUSTRALIA

Week 8:
Unsupervised Learning – Clustering

- **Learning Outcomes**
  - K-Means Clustering
  - Hierarchical Clustering

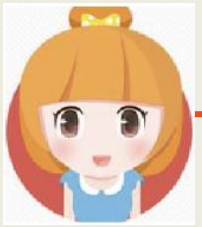
Unsupervised Learning

- **Supervised vs Unsupervised learning**
  - **Supervised Learning (Regression, Classification)**
    - **Given both** the n-dimensional features $X_1, X_2, \cdots, X_n$ and the label Y for each data sample.
    - **The goal** of supervised learning is to predict Y using $X_1, X_2, \cdots, X_n$
  - **Unsupervised Learning (known as Knowledge Discovery)**
    - **Given only** the n-dimensional features $X_1, X_2, \cdots, X_n$.
    - **The goal** of unsupervised learning is to discover hidden patterns and groupings in the data.



Unsupervised Learning *vs.* Supervised Learning

- **Unsupervised learning** is generally **more subjective** than supervised learning because there is **no explicit goal**, such as predicting **a label**.
  - **Unlabeled data** $X_1, X_2, \cdots, X_n$ is often easier to obtain — for example, directly from lab instruments or computer systems — compared to labeled data $\{X_1, X_2, \cdots, X_n, Y\}$, which usually requires human effort to classify or annotate.



Task: Rate books

Label: ?

Features of books

Feature: $X_1, X_2, \cdots, X_n$

Unsupervised Learning – Clustering

- Clustering refers to a very broad set of techniques for finding **subgroups**, or **clusters**, in a dataset.
  - We seek a **partition/classify** of the data into distinct groups so that the observations within each group are **quite similar** to each other,
  - It makes this concrete, we must **define** what it means for two or more observations to be **similar** or different.
- We will discuss several common clustering methods, including:
  - **K-means**
  - **Hierarchical Clustering**

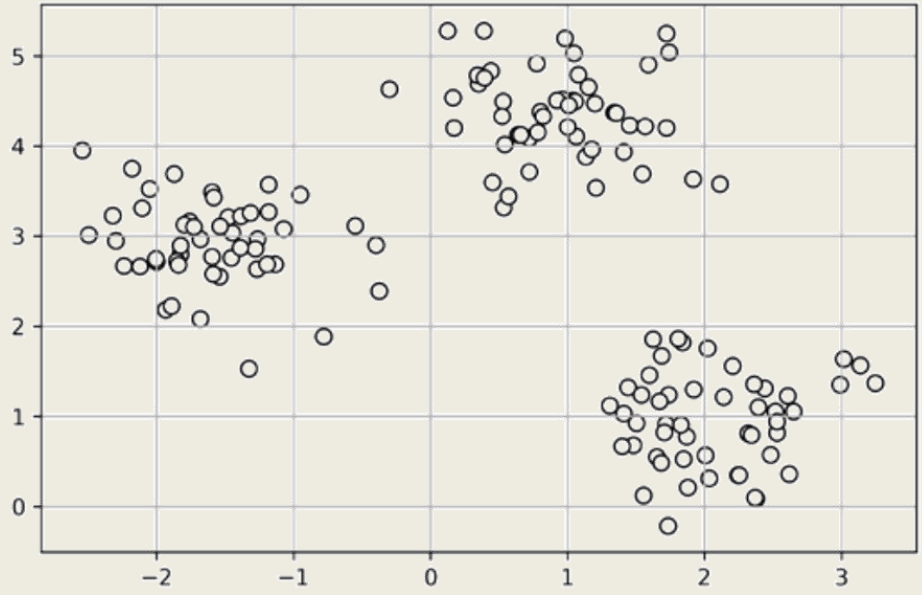
K–means *vs.* Hierarchical Clustering

- K-means Clustering
  - Partition data into a **pre-specified number** of clusters (**k**).
  - Must choose **k** before running the algorithm.
- Hierarchical Clustering
  - No need to pre-specify the number of clusters.
  - Produces a tree-like diagram, showing possible clusterings from **1** to **n** observations(data samples).
  - Allows **flexibility** to decide the number of clusters afterward.

K–Means Clustering

- It is used when we have **unlabelled data** which is data without defined categories or groups
  - Our goal is to **group the data smples based on their feature similarities**, which can be achieved using the k-means algorithm, as summarized by the following four steps:

- Scatter diagram of **n** data smples in 2-D space.



➢ How to partition data using K-Means?

Steps of K–Means

1. **Initialize: Randomly** select **k samples as initial centroids** (cluster centers).
2. **Assign:** each sample to the **nearest centroid**, $\mu_j, j \in \{1, \cdots, k\}$ (Similarity)
3. **Update**: **Move each centroid** to the mean of its assigned samples.
4. **Repeat:** steps 2 and 3.

   Stop when:

   a) cluster assignments no longer change or

   b) a user- defined tolerance (e.g., ε) is met or

   c) maximum iterations (e.g.,1000) are reached.

Similarity Definition

- Similarity can be viewed as the opposite of distance.
- In clustering with continuous features, a common choice is the **squared Euclidean distance**:

$$d(x,y)^2 = \sum_{i=1}^{p} (X_{1i} - X_{2i})^2 = ||X_1 - X_2||_2^2$$

  - where p is the number of features (in the p-dimensional space).
  - X_{1i}, X_{2i} are the values of the two points in the *i*-th feature.

Objective Function

- K-Means is an iterative approach for minimizing the **within-cluster sum of squared errors (SSE)**, which is sometimes also called **cluster inertia**:

$$SSE = \sum_{i=1}^{n} \sum_{i=1}^{k} w^{(i,j)} ||X_i - \mathbf{u}_j||_2^2$$

- where $\mathbf{u}_j$ is the centroid point for clusster j. $w^{(i,j)}$ = 1 if the sample X_i is in cluster j, or 0 otherwise.

$$w^{(i,j)} = \begin{cases} 1, & \text{if } x^{(i)} \in j \\ 0, & \text{otherwise} \end{cases}$$

Note: K–Means may get stuck in a local minimum — it doesn't guarantee the global best clustering unless initialized smartly (hint for K–Means++ if you want to extend).

– A greedy algorithm
– No 'lookahead'
– Random initialization

Algorithm of K–Means (in Practice)

1. **Initialize: Randomly** select **k samples as initial centroids** (cluster centers).
2. **Assign:** each sample to the **nearest centroid**, $\mu_j, j \in \{1, \cdots, k\}$
3. **Update: Move each centroid** to the mean of its assigned samples.
4. **Repeat:** steps 2 and 3.

   Stop when:

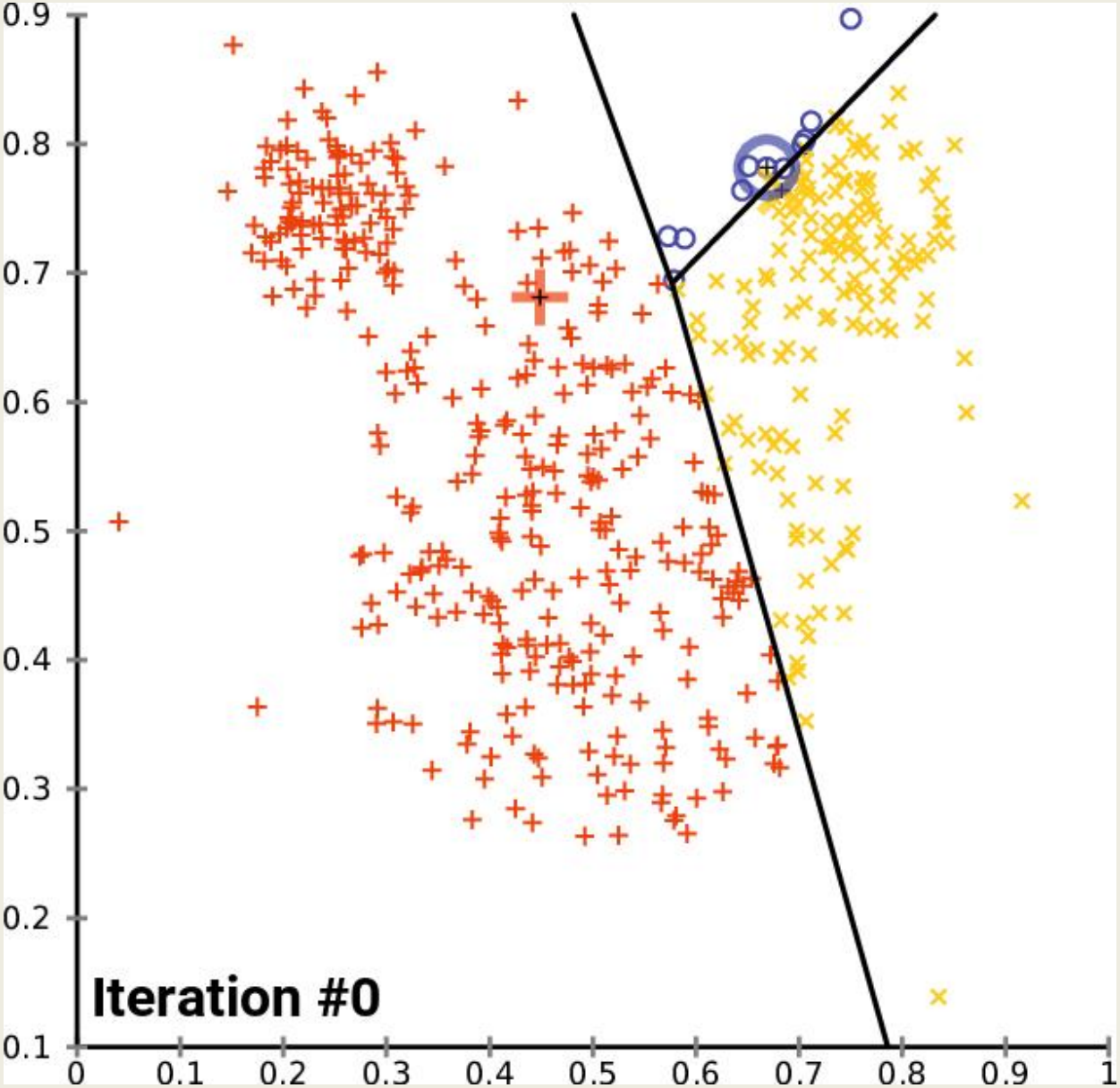
   a) cluster assignments no longer change or

   b) a user- defined tolerance (e.g., ε) is met or

   c) maximum iterations (e.g.,1000) are reached.

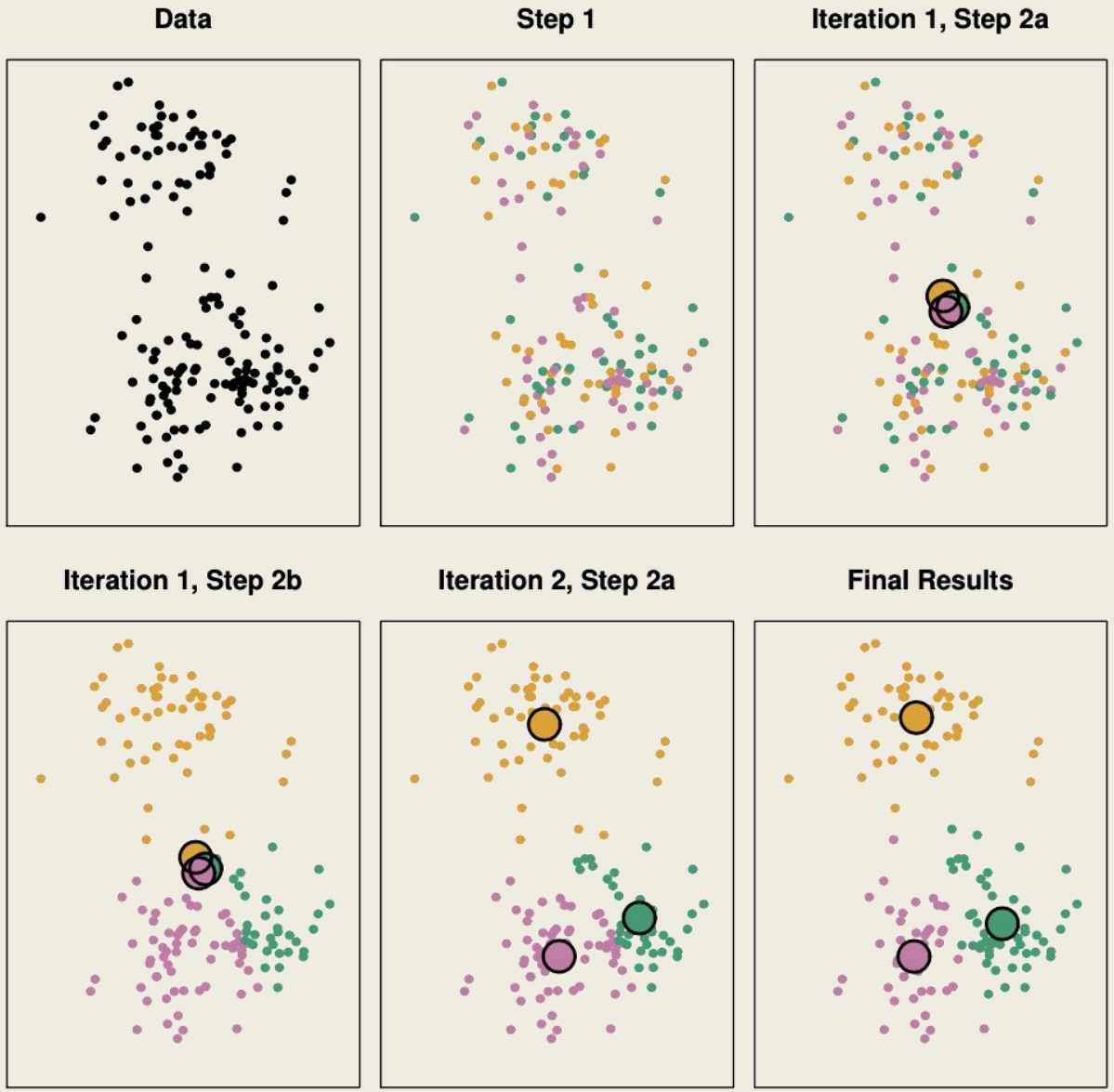
5. **We often carry out the k-means clustering algorithms n (e.g., n=10) times independently, with different random centroids to choose the final model as the one with the lowest SSE.**

Iteration Steps



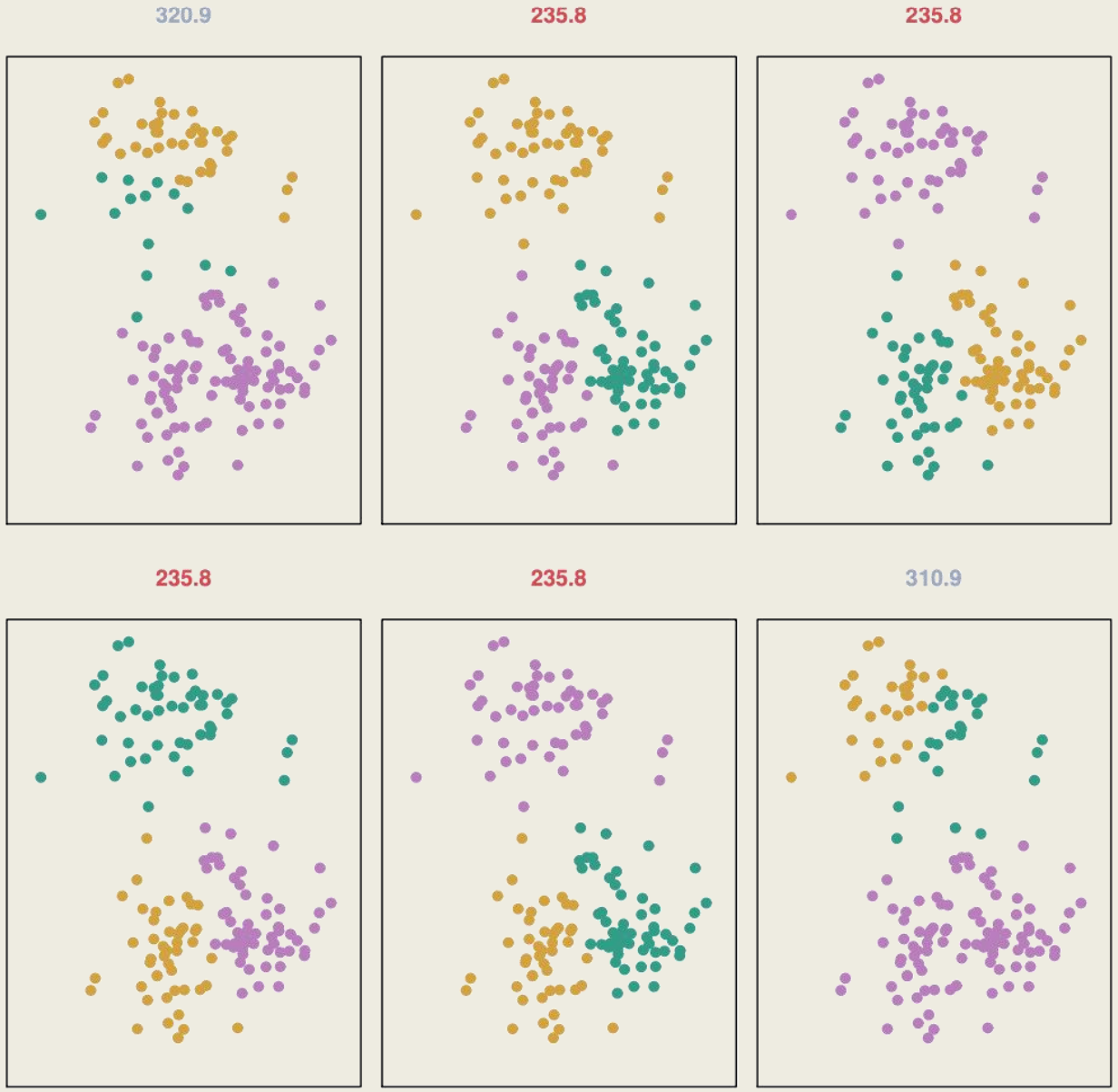
Example (k=3)

1. Step 1: Each observation is randomly assigned to a cluster;
2. Step 2a: **the cluster centroids** are computed by the **mean** of each points
3. Step 2b: Each observation is assigned to the **nearest centroid**
4. Step 2a**...**



Different Initiations

- Different random centroids will obtain varying clustering results.



K–means *vs.* Hierarchical Clustering

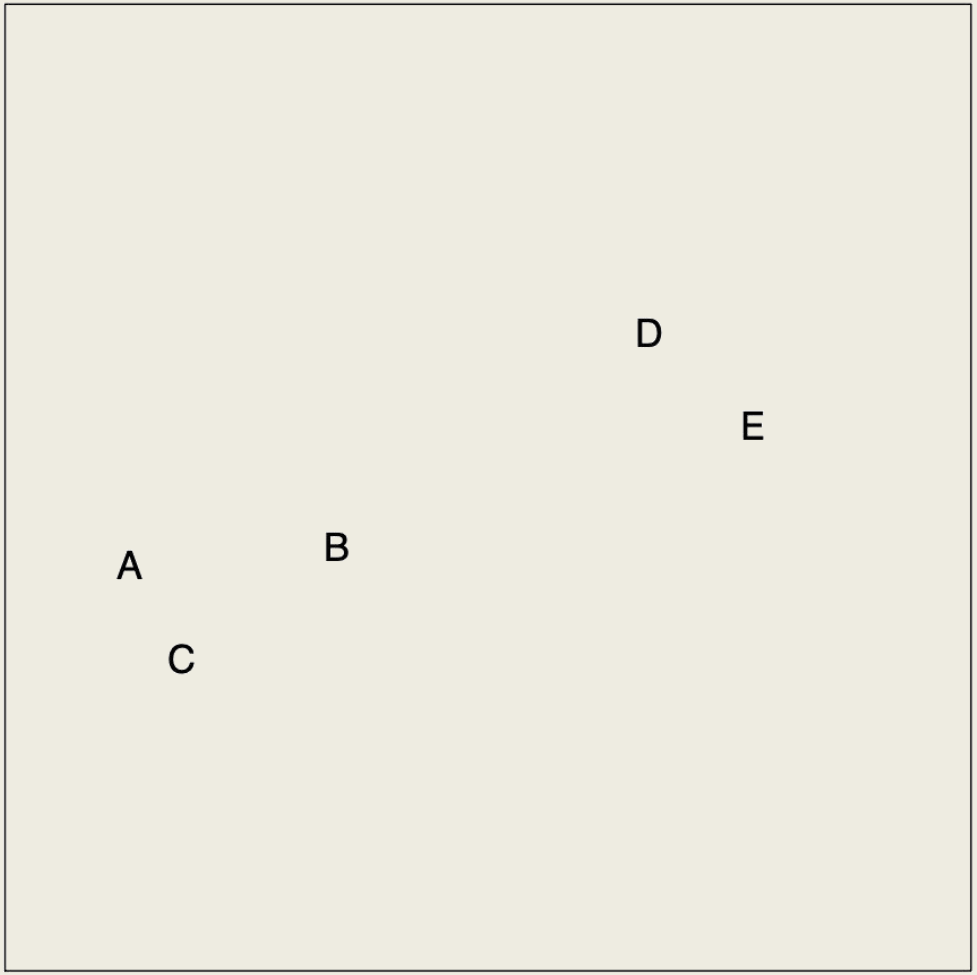
- K-means clustering
  - Requires pre-specifying the number of clusters (Fixed K).
  - Choosing the **right K can be challenging**.
- **Hierarchical clustering** is an alternative approach which
  - **does not require that we commit to a particular choice of K.**
  - Common approach: bottom-up clustering, which starts with **each point** as its own cluster and **merge** step-by-step.
  - Choose K later by cutting the dendrogram (Flexible K)

1. Start with **each point** in its own cluster.
2. Identify the **closest** two clusters and **merge** them.
3. Repeat 2.
4. Ends when all points are in a single cluster.

Hierarchical Clustering

- Builds a hierarchy in a “bottom-up” fashion

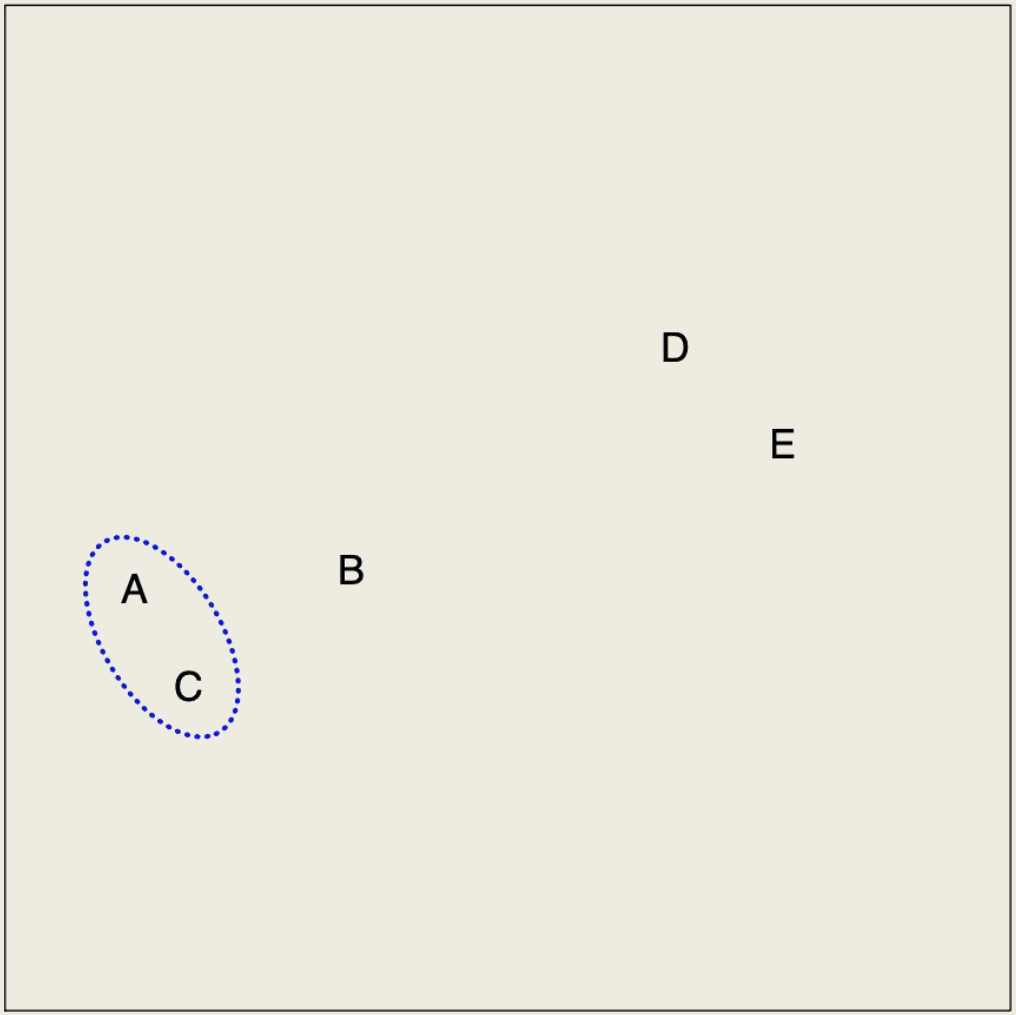
1. Start with **each point** in its own cluster.



Hierarchical Clustering

- Builds a hierarchy in a “bottom-up” fashion

2. Identify the **closest** two clusters and **merge** them.

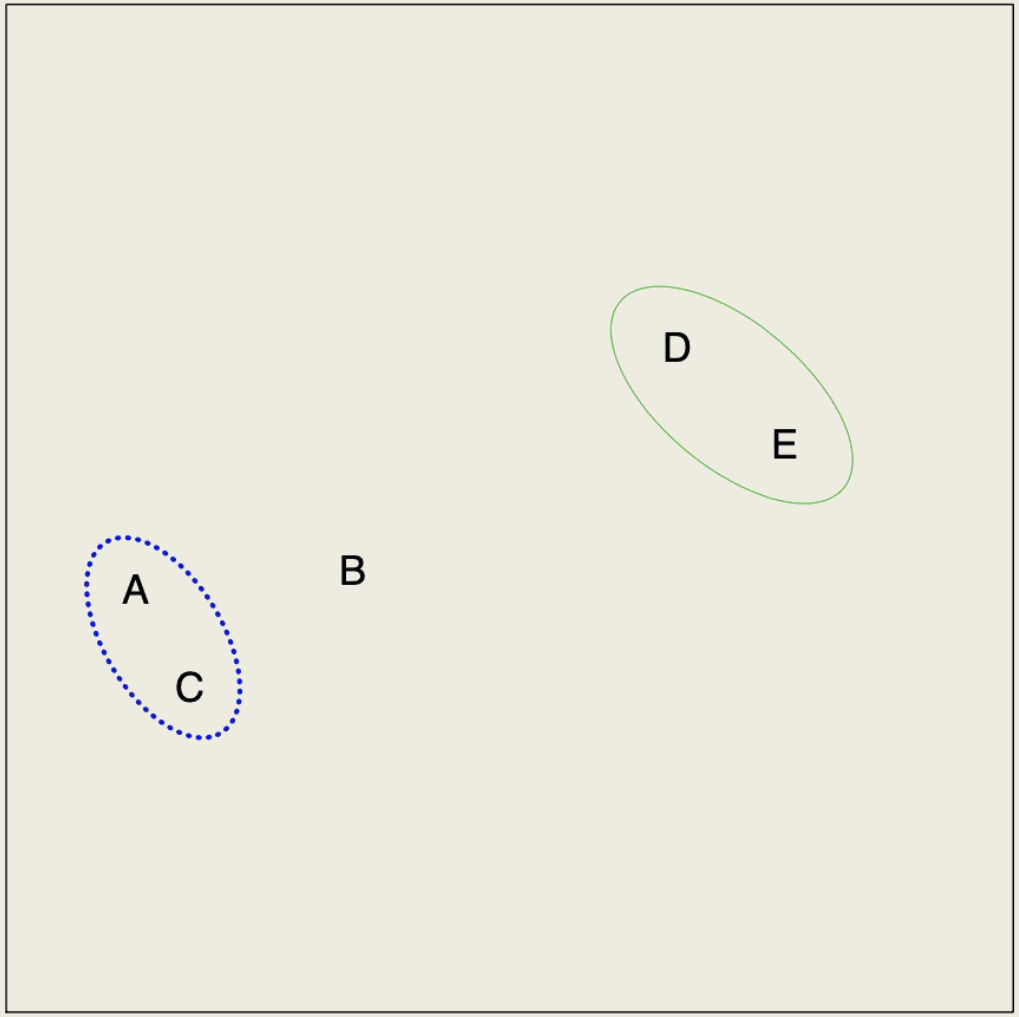


Hierarchical Clustering

- Builds a hierarchy in a “bottom-up” fashion

3. Repeat 2

Identify the **closest** two clusters and **merge** them.

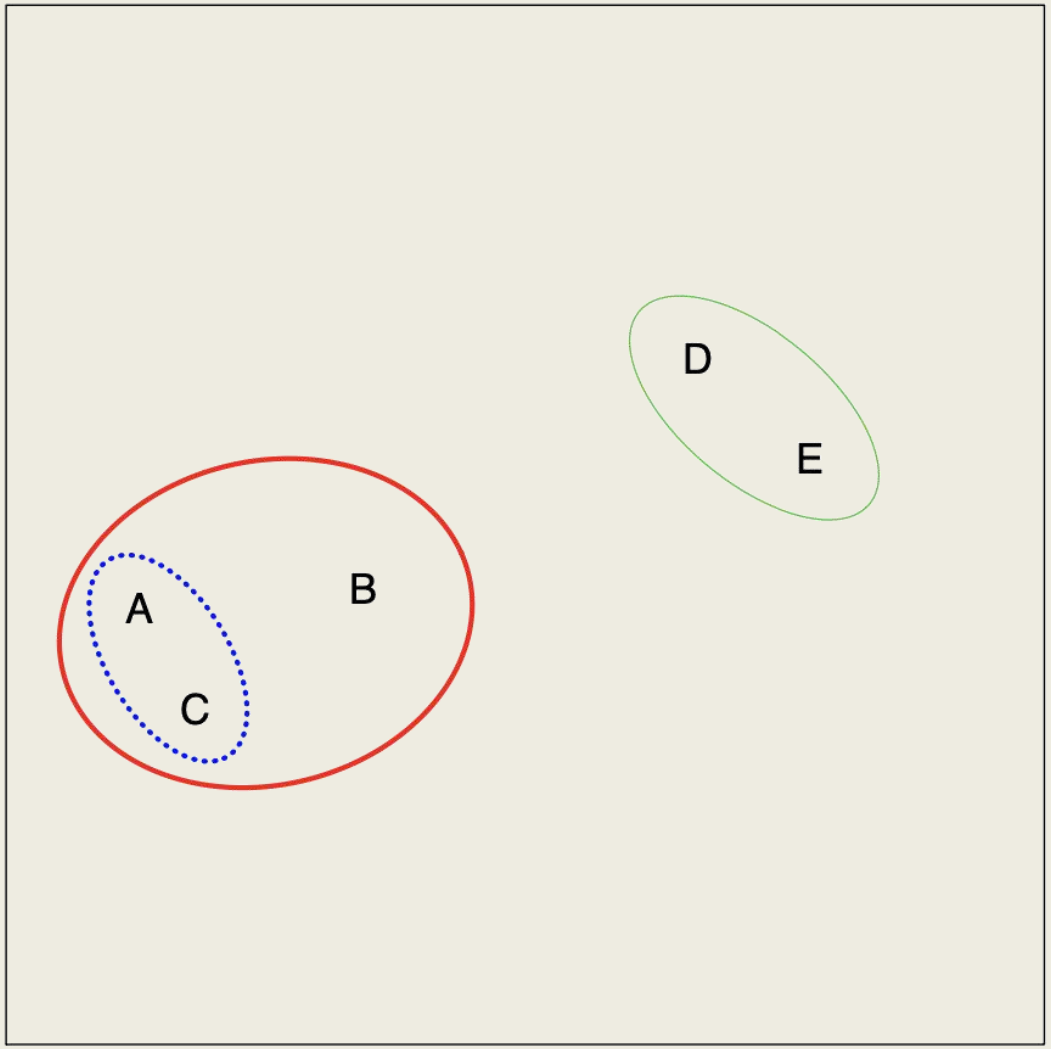


Hierarchical Clustering

- Builds a hierarchy in a “bottom-up” fashion

3. Repeat 2

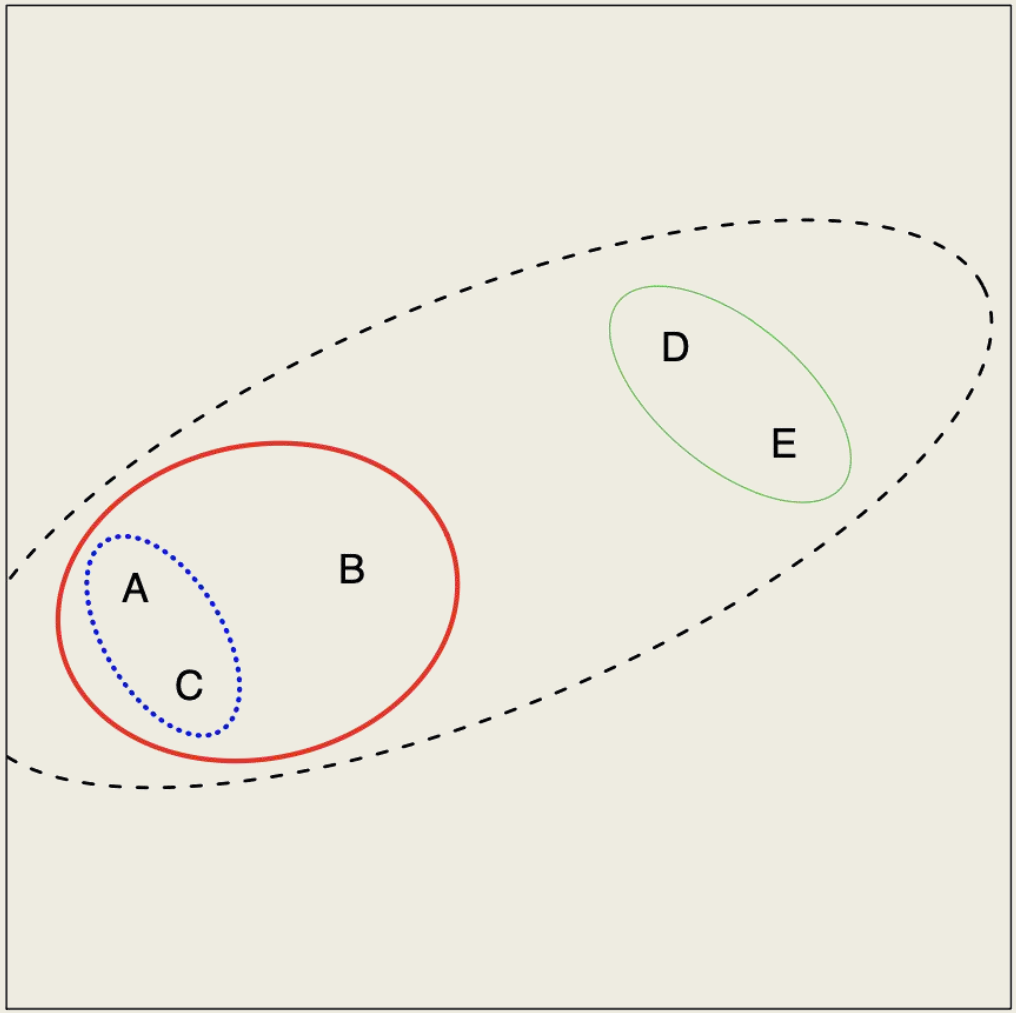
Identify the **closest** two clusters and **merge** them.



Hierarchical Clustering

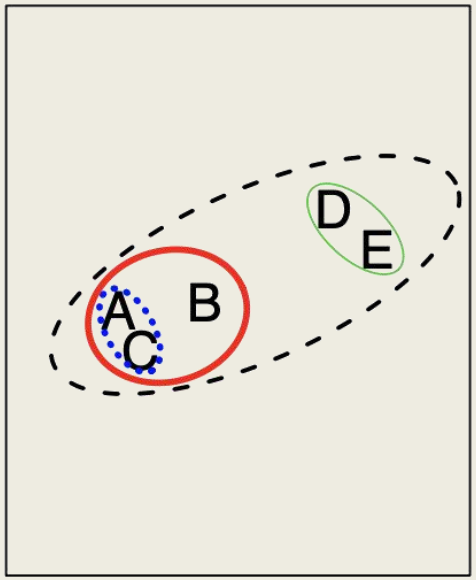
- Builds a hierarchy in a “bottom-up” fashion

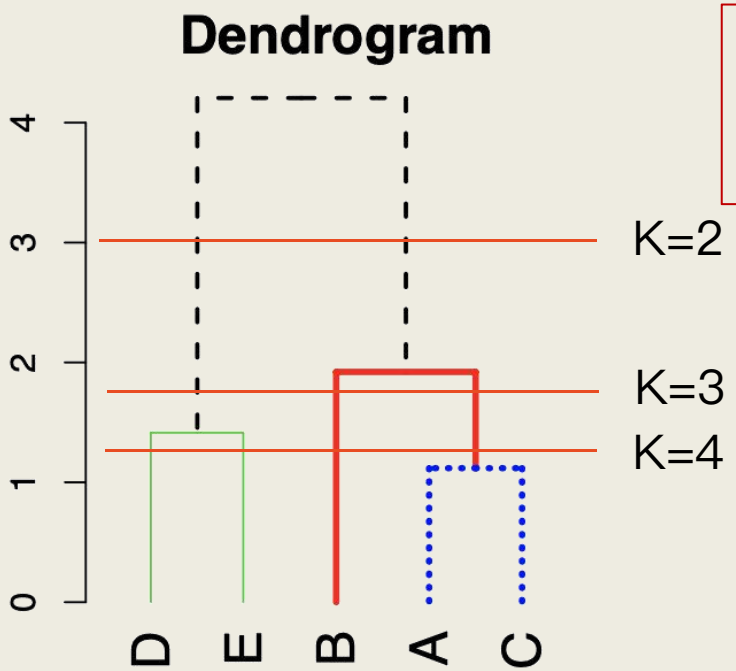
4. Ends when **all points** are in a **single cluste**r.



Hierarchical Clustering Algorithm

- The approach in words:
  - Start with each point in its own cluster.
  - Identify the **closest** two clusters and merge them.
  - Repeat.
  - Ends when all points are in a single cluster.

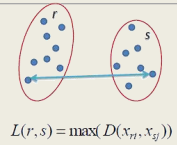
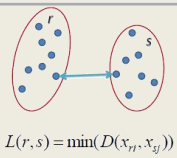
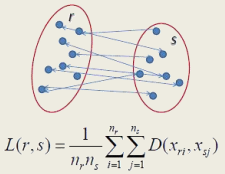




Choose K by cutting the dendrogram (Flexible K)

Cluster Linkage

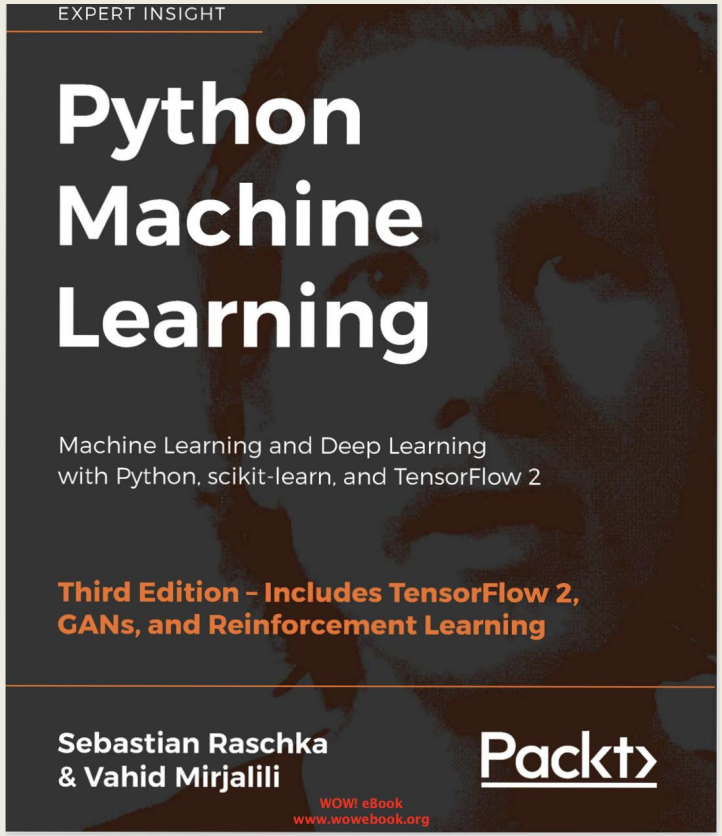
In order to decide which clusters should be combined or where a cluster should be split, a measure of dissimilarity between clusters of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate distance d, such as the Euclidean distance, and a cluster linkage which specifies the dissimilarity of clusters.

Maximum or complete-linkage clustering	$\max_{a \in A,\, b \in B} d(a,b)$	
Minimum or single-linkage clustering	$\min_{a \in A,\, b \in B} d(a,b)$	
Unweighted average linkage clustering (or UPGMA)	$\frac{1}{\lvert A\rvert \cdot \lvert B\rvert} \sum_{a \in A} \sum_{b \in B} d(a,b).$	
Weighted average linkage clustering (or WPGMA)	$d(i \cup j, k) = \frac{d(i,k) + d(j,k)}{2}.$	
Centroid linkage clustering, or UPGMC	$\lVert \mu_A - \mu_B \rVert^2$ where μ_A and μ_B are the centroids of *A* resp. *B*.	

Summary

- Unsupervised learning is important for understanding grouping structure of a set of unlabeled data, and can be a useful pre–processor for supervised learning – obtain label data.
- K-Means Clustering
- Hierarchical Clustering

References



- https://en.wikipedia.org/wiki/K-means_clustering
- https://en.wikipedia.org/wiki/Hierarchical_clustering
- https://www.statlearning.com/resources-python
- Sebastian Raschka, etl., 'Python machine learning', Third Edition