# Clustering Methods

PSY 504

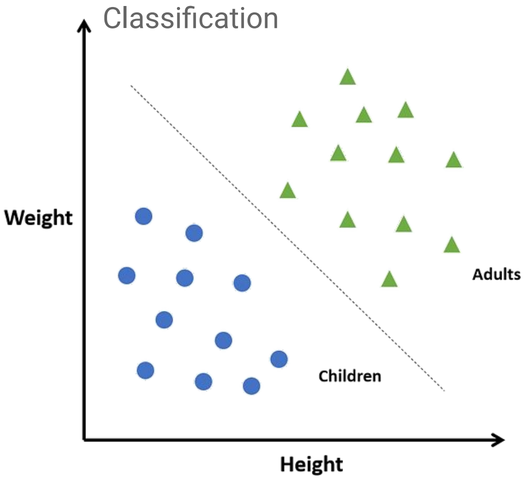
Jamie Chiu 2023-04-17

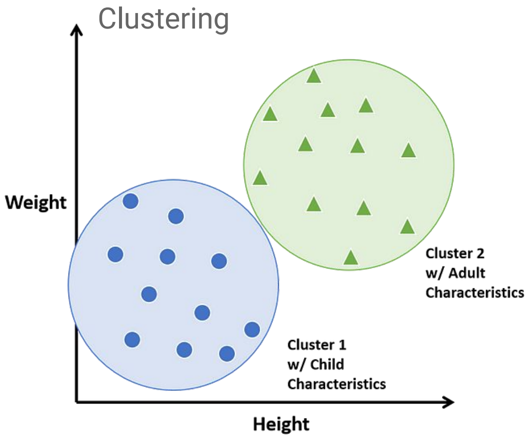
# What is Clustering?

Unsupervised problem where we are trying to discover structure in the dataset.

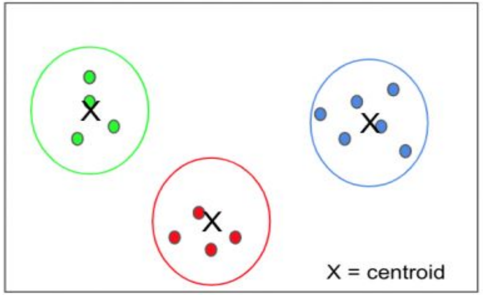
When we cluster the observations of a data set, we seek to:

1. partition them into distinct groups;
2. such that the observations within each group are quite similar to each other;
3. while observations in different  groups are quite different from each other.



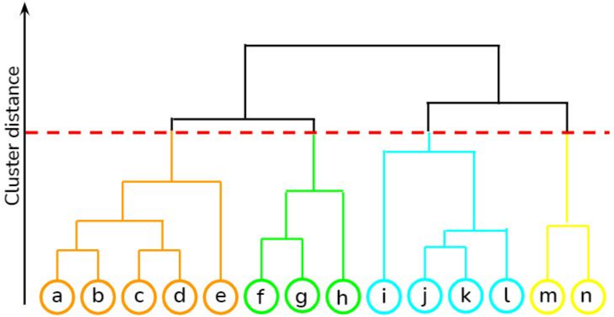


# K-Means Clustering



- Partition observations into K distinct, non-overlapping clusters (where K is pre-specified).
- Each observation can only belong in one cluster.
- Clusters are optimised for minimum within-cluster variation.

# Hierarchical Clustering



- Each observation begins as its own cluster, which fuses into larger and larger clusters as you move up the hierarchy.
- Therefore, K is not predetermined.
- Visual representation of clusters + subclusters.

# K-Means Clustering

Objective: Partition observations into K clusters such that within-cluster variation is minimised.

What is a good cluster?

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Where:

- C → cluster
- W(C) → within-cluster variation

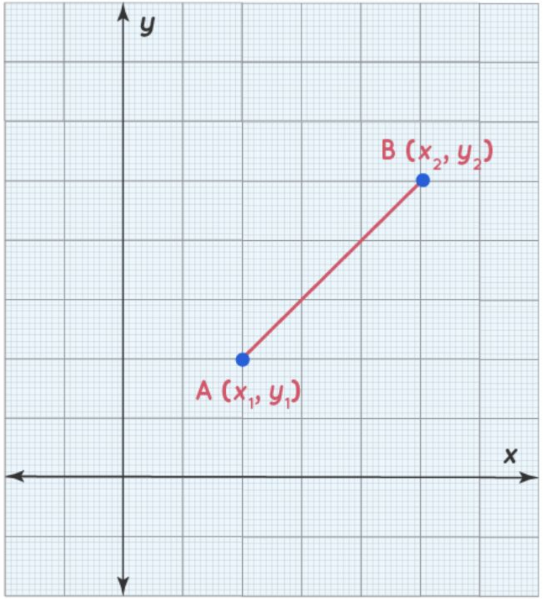
How do we define within-cluster variation?
Common method: *squared Euclidean distance*

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$

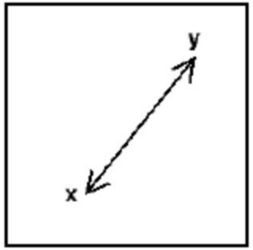
Where:

- |C| → observations in cluster

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$



$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



Euclidean

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

# K-Means Clustering

Objective: Partition observations into K clusters such that within-cluster variation is minimised.

Combining the two equations gives us the optimisation problem of K-means clustering:

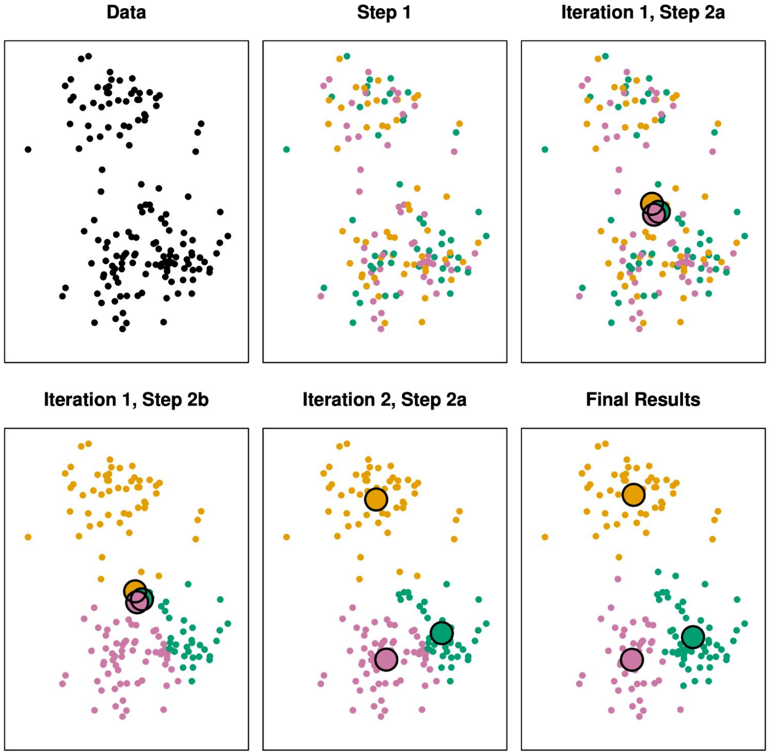
$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

# K-Means Clustering

Objective: Partition observations into K clusters such that within-cluster variation is minimised.

Pseudo-code:

```
1. Randomly assign a number, from 1 to $K$, to each of the observations.
   These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the $K$ clusters, compute the cluster *centroid*. The
       $k$th cluster centroid is the vector of the $p$ feature means for the
       observations in the $k$th cluster.
   (b) Assign each observation to the cluster whose centroid is closest
       (where *closest* is defined using Euclidean distance).
```



# K-Means Clustering

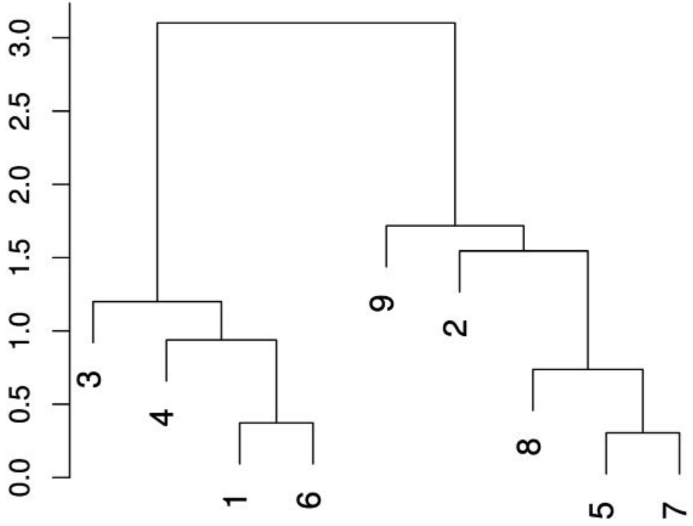
Objective: Partition observations into K clusters such that within-cluster variation is minimised.

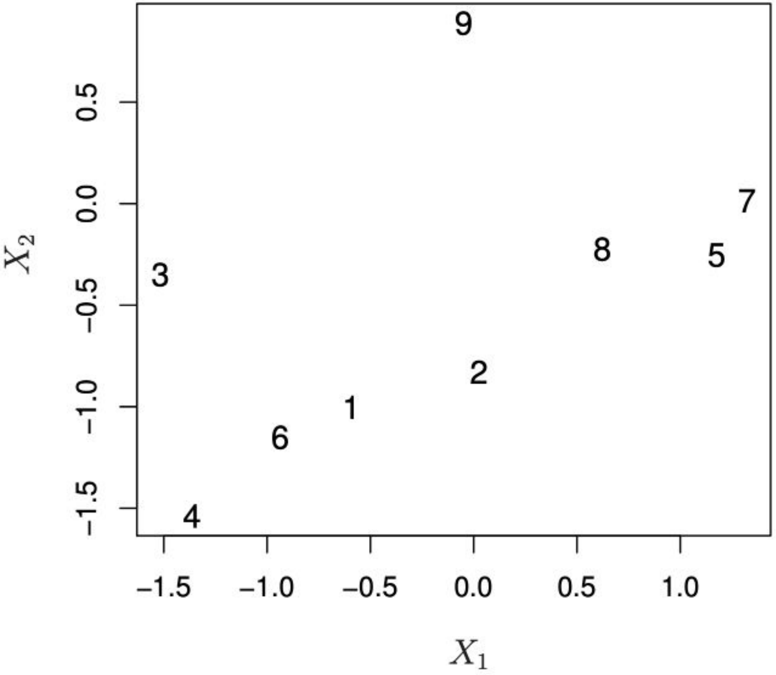
Considerations:

- When to use? If there is a specific number of clusters in the dataset, but the group membership is unknown, then K-means can be useful.
- K-means is good at capturing the structure of the data if the clusters have a spherical-like shape.
- Which can also be a disadvantage: If your clusters are shaped differently, K-means can do a poor job clustering.
- K-means begin with random initialization, so clustering results can be different depending on the run. Depending on your dataset, clusters may not be reproducible on another independent dataset.
- K-means can also be unstructured and difficult to interpret.
- Each observation has to be in one cluster only, thus, cannot support nested relationships.

# Hierarchical Clustering

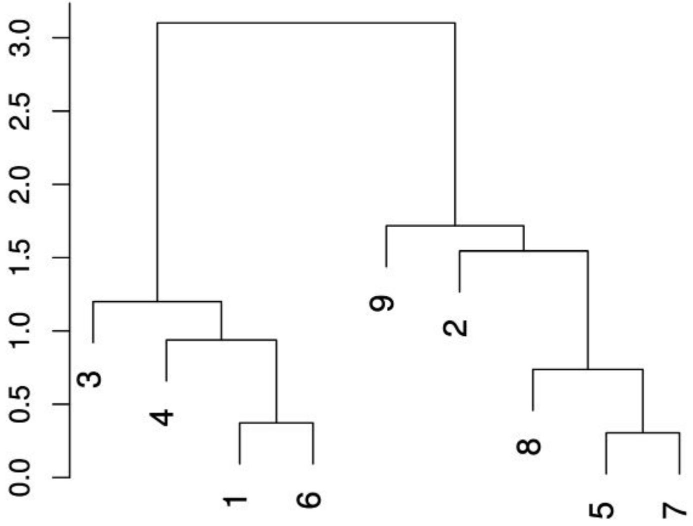
Objective: Organise observations into nested levels of similarity and dissimilarity.





# Hierarchical Clustering

Objective: Organise observations into nested levels of similarity and dissimilarity.

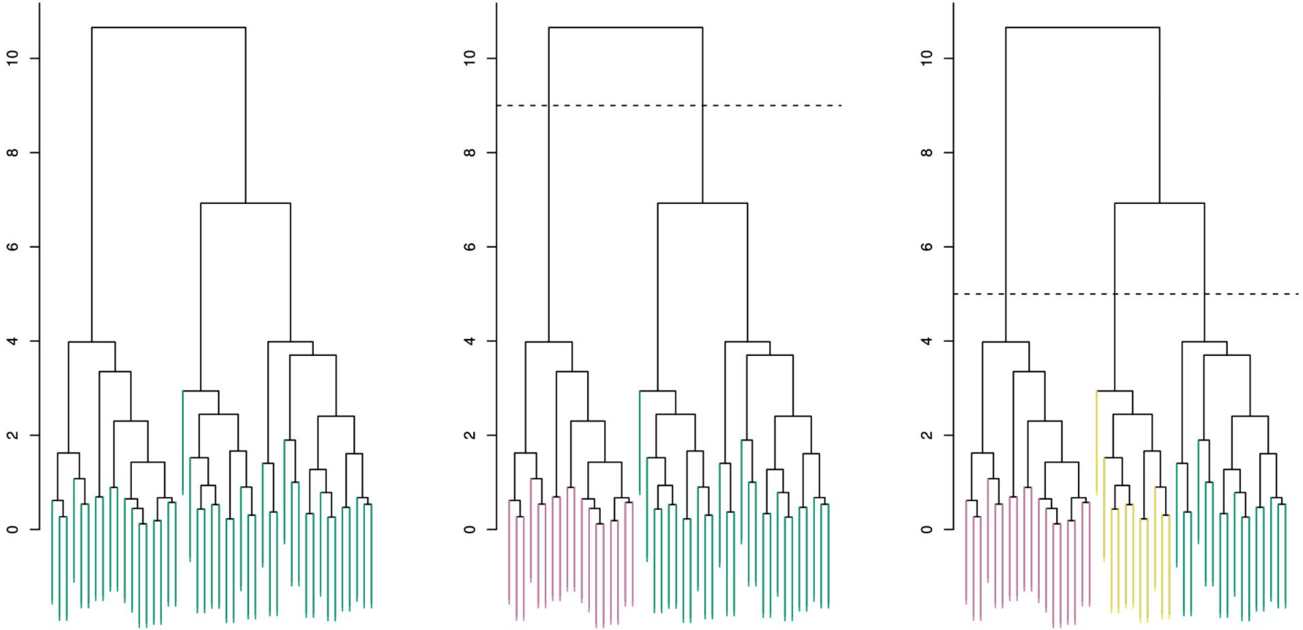


Interpreting Dendrograms:

- Bottom-most level (leaves) represent the individual observations.
- Where the branches join represent formation of a cluster / sub-cluster.
- Similarity is depicted by the Y-axis (NOT the X-axis). The higher up the joint, the less similarity there is.

# Hierarchical Clustering

Objective: Organise observations into nested levels of similarity and dissimilarity.



# Hierarchical Clustering

Objective: Organise observations into nested levels of similarity and dissimilarity.

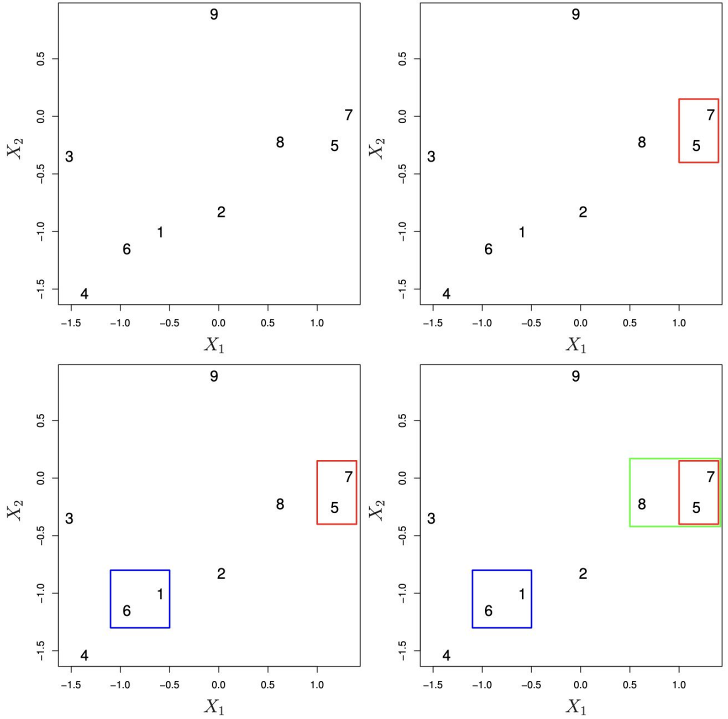
## Pseudo-code:

```
1. Begin with $n$ observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

    (a) Examine all pairwise inter-cluster dissimilarities among the $i$ clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

    (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.
```



# Hierarchical Clustering

Objective: Organise observations into nested levels of similarity and dissimilarity.

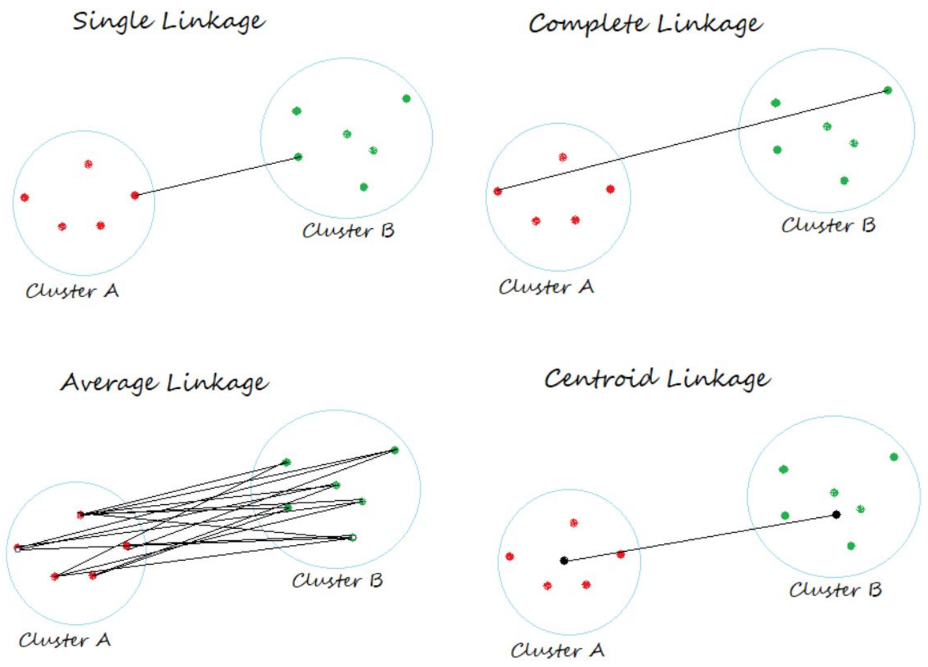
Different methods used to determine dissimilarity:

- Euclidean distance
- Correlation-based distance

And for how distance between clusters are aggregated:

- “Linkage”: complete (max), single (min), average, centroid.

Choose these parameters based on your data and question.



# Hierarchical or K-Means Clustering?

Example:

Suppose our dataset corresponds to a group of men and women, evenly split among Americans, Japanese, and French.

The best division into two groups might split these people by gender, and the best division into three groups might split them by nationality.

In this case, the true clusters are not nested – and would not be well-represented by hierarchical clustering.

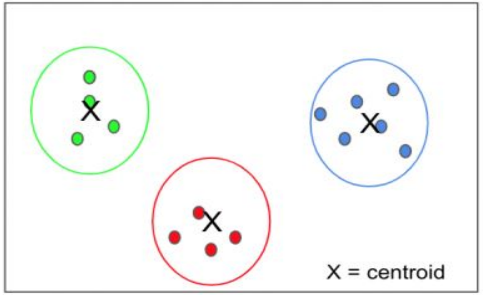
# Clustering

Considerations:

- Clustering will always result in clusters; whether it makes sense or not.
- Even though these are data-driven approaches (unsupervised), you should still think about your dataset and question.
- In practice – try several methods of clustering and look for the one with the most useful or interpretable solution.
- Clustering is often used in the form of descriptive modelling rather than predictive.
- K-means and hierarchical clustering both assign each observation to a cluster – however, that might not always be appropriate, and there are mixture models that better accommodate outliers.

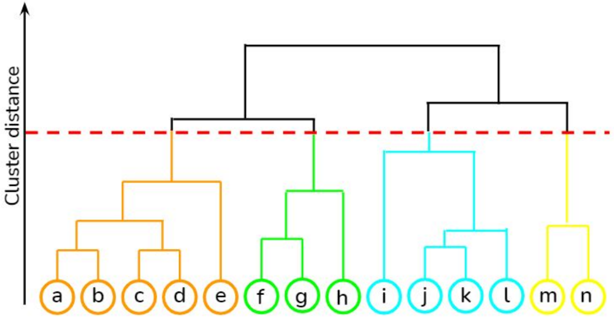
**Questions?**

# K-Means Clustering



- Partition observations into K distinct, non-overlapping clusters (where K is pre-specified).
- Each observation can only belong in one cluster.
- Clusters are optimised for minimum within-cluster variation.

# Hierarchical Clustering



- Each observation begins as its own cluster, which fuses into larger and larger clusters as you move up the hierarchy.
- Therefore, K is not predetermined.
- Visual representation of clusters + subclusters.