Hierarchical Clustering

Renato R. Maaliw III, *DIT*
College of Engineering
Southern Luzon State University
Lucban, Quezon, Philippines

Hierarchical Clustering (HC)

HC is a special way of grouping things where we start with everything in its own group, and then we slowly combine the most similar group until there's just one big group

Think of it like building a **family tree**:

- At the bottom, every person starts alone
- As you go up, people are connected based on who is most closely related to them
- At the top, all the people are connected in one big family

In HC:

- every data point starts in its own “family”
- then we merge the closest data points or clusters together step by step
- this process continues until everything is in one big “family”

Example: Grouping Friends

Imagine you have 10 friends, and you want to group them based on how much time they spend together. Here's how HC works:

Start with individuals: At first, each friend is in their own group

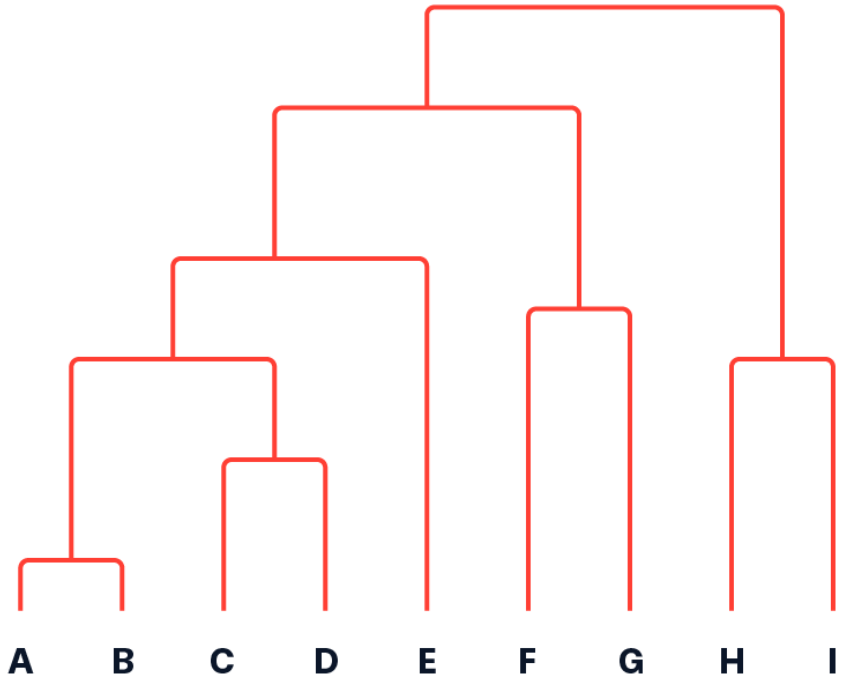
Find the closest friends: You check which two friends spend the most time together and group them.

Keep Grouping: Then you check which group of friends is closest to another and keep combining the groups.

One Big Group: At the end, you have one big group than shows how all the friends are connected to each other.

Dendrogram: The Tree of Clusters

Now, let's talk about the **dendrogram**, which is just a fancy word for the tree that HC creates.



You can "cut" this tree at any height to decide how many groups of friends you want.

If you cut it higher, you get fewer, larger groups. If you cut it lower, you get more, smaller groups.

Types of HC:

1. **Agglomerative (Bottom-Up Approach):** Start with each thing (or person) in its own group and merge until there's just one big group.

2. **Divisive (Bottom-Up Approach):** Start with one big group and split it into smaller groups. This is like starting with a whole group of friends and slowly dividing them into smaller cliques.

How do we know which cluster (friends) are closest?

1. Euclidean Distance: This is like drawing a straight line between two points on a map and measuring the distance. If two friends live close together, they're probably more likely to hang out.

How do we know which cluster (friends) are closest?

2. Manhattan Distance: Imagine walking only along streets on a grid, like in a city. This is another way of measuring how far things are.

Linkage Methods: How Do We Merge Groups?

When we start merging groups of friends, we need to decide how to measure the distance between groups, not just between individual people. Here are three ways we can do this:

1. Single Linkage

We look at the **two closest people** in each group and use that distance.

2. Complete Linkage:

We look at the **two farthest people** in each group and use that distance.

3. Average Linkage:

We find the **average distance** between all the people in the two groups.

* We usually use **Ward's Method**, which tries to make groups that are nice and balanced. *

4. Ward's Method:

Minimizes the total variance within clusters.:

* We usually use **Ward’s Method**, which tries to make groups that are nice and balanced. *

Thank you very much for listening.