Slide 1:

Good afternoon, everyone!
My name is Mayur Mali, and I'm here with my classmate Arya Memane. Today, we're going to discuss an interesting topic in Machine Learning – **Hierarchical Clustering**. And we'll walk you through the key concepts, the process, and its importance.

Slide 2:

So, what is Hierarchical Clustering?
In simple terms, it's a clustering algorithm that groups data points based on how similar or close they are to each other. Unlike other algorithms that create just one set of clusters, hierarchical clustering builds a "tree" of clusters, where each branch represents clusters within clusters.

The process is step-by-step. We can either start with small clusters and keep merging them to form bigger ones, or we can start with one large cluster and divide it into smaller ones. This step-by-step structure is why it's called "hierarchical" – the clusters get organized at different levels, kind of like layers.

A big advantage is that we can visualize hierarchical clustering with something called a **dendrogram** – which we'll talk more about later. Unlike other methods like K-means, we don't need to know how many clusters we want in advance, so it's really useful when exploring data for the first time.

Slide 3:

So, why choose Hierarchical Clustering?
Firstly, **it's flexible**. We don't have to decide on the number of clusters beforehand, which is an advantage over K-means. It also gives us a great visual representation in the form of a dendrogram. This visual "tree" helps us see the relationships between data points and understand how clusters form at different levels.

Plus, it's easy to interpret. Hierarchical clustering shows us how closely data points are related at different levels of similarity, which is useful when we want a clearer picture of the data.
And it's versatile too – it works with both numbers and categories, making it a good choice for exploring smaller or medium-sized datasets. However, for large datasets, it can be computationally expensive.

Slide 4:

Now, let's talk about the types of hierarchical clustering.
There are two main approaches:

1. **Agglomerative Clustering:** This is a "bottom-up" approach where each data point starts as its own separate cluster. The algorithm then merges the closest clusters one by one until we're left with one big cluster. We'll focus more on this today since it's the more common approach in machine learning.
2. **Divisive Clustering:** This is the opposite, a "top-down" approach. All data points start in one big cluster, which is then divided into smaller clusters until each point stands alone.

Slide 5:

Let's dive deeper into Agglomerative Clustering.
In this approach, each data point starts as its own little cluster. Then, the algorithm finds the two clusters that are closest to each other and merges them into a single cluster. It keeps doing this until we're left with just one big cluster.

If you've worked with Merge Sort before, you might see a similarity. In Merge Sort, we split the list into small parts and then merge them back together in sorted order. Here, we start with individual data points and keep merging the closest ones, building up to larger clusters

---

Slide 6:

Let's break down the steps for Agglomerative Clustering:

1. Each data point begins as its own cluster. So, with N data points, we start with N clusters.
2. Next, we calculate the distance between every pair of clusters. We often use Euclidean distance for this.
3. We find the two clusters that are closest to each other and merge them into a larger cluster.
4. We update the distance matrix to include the new cluster's distances from the remaining clusters.
5. This continues until we're left with one large cluster.

6. To make sense of it, we use a dendrogram, which is like a tree diagram showing each step of clustering. It helps us see which clusters merged at each stage.

Slide 7:

let's talk about dendrograms what they are and why they're useful.
A dendrogram is a tree-like diagram that shows the entire clustering process. You can think of it as a map of how clusters formed over time.

- At the very bottom, each point or "leaf" represents an individual data point.
- The branches show how these points group together, forming larger and larger clusters.
- The height of each branch tells us the distance, or how different clusters were from each other when they were merged.

A helpful feature of the dendrogram is that we can decide the number of clusters by "cutting" it at a certain level. For example, cutting at a higher level might give us two or three clusters, while cutting lower gives us more clusters.

Slide 8:

Finally, let's talk about how we measure the distance between two clusters.
There are a few common ways to do this, called "Linkage" methods:

1. **Single Linkage:** Uses the minimum distance between any two points in the clusters.
2. **Complete Linkage:** Measures the longest distance between points in the clusters.
3. **Centroid Method:** Measures the distance between the centroids (geometric centers) of the two clusters.
4. **Average Linkage:** Calculates the average distance between all points in one cluster and all points in the other.

Now Aarya will explain the working of the Agglomerative clustering with the help of a case study.