# Intro to Clustering

Machine Learning

# Clustering

- Clustering is an Unsupervised Learning Technique
- A Cluster: collection of objects that are **similar**
- Objective is to group similar data points into a group
    - Segmenting customers into similar groups
    - Automatically organizing similar files/emails into folders
- Simplifies data by reducing many data points into a few clusters

# Distance

- Do define “similarity” you need a measure of distance
- Examples of common distance measures
  - Manhattan Distance
  - Eucledian Distance
  - Chebyshev Distance

# Types of Clustering

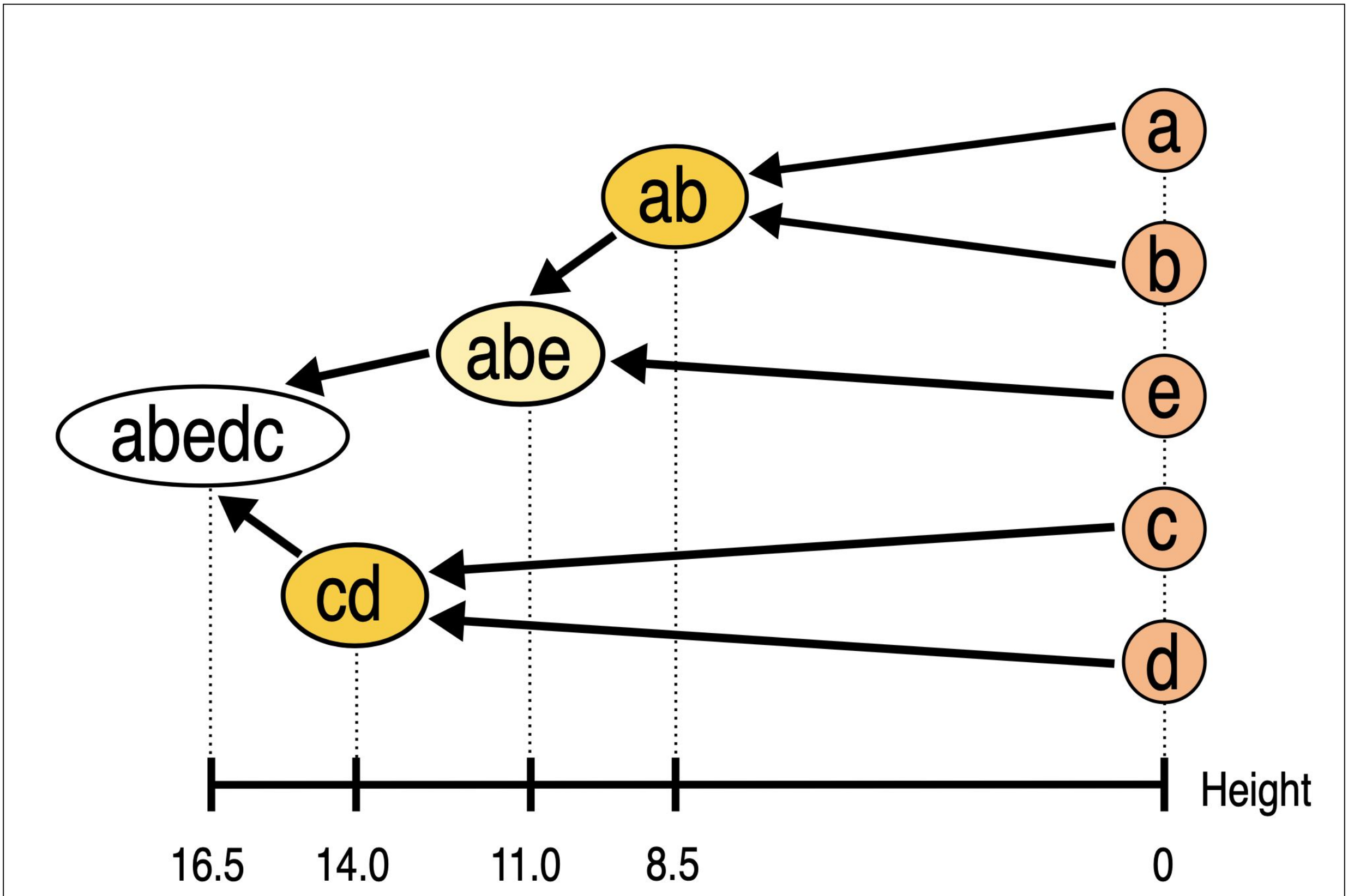
1. <u>Connectivity based clustering</u> (Hierarchical clustering): based on the idea that related objects are closer to each other. Can we then create a hierarchy of clusters/groups.

    - Useful when you want flexibility in how many clusters you ultimately want. For example, imagine grouping items on an online marketplace like Etsy or Amazon.

    - In terms of outputs from the algorithm, in addition to cluster assignments you also build a nice tree **(dendrogram)** that tells you about the hierarchies between the clusters. You can then pick the number of clusters you want from this tree.

    - In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis.

    - Algorithms can be agglomerative (start with 1 object and aggregate them into clusters) or divisive (start with complete data and divide into partitions).


a
ab
b
abe
e
abedc
c
cd
d
Height
16.5
14.0
11.0
8.5
0

# Types of Clustering

2. Centroid based clustering (Eg. K- Means clustering): The objective is to find K clusters/groups. The way these groups are defined is by creating a centroid for each group. The centroids are like the heart of the cluster, they “capture” the points closest to them and add them to the cluster.

    - Large K produces smaller groups and a small K produces larger groups
    - K-Means uses Eucledian distances and is the most popular
    - Other variants like K-medians and K-mediods use other distance measures