Hierarchical clustering:

- Two main types:
    - Agglomerative:
        - Start with every point in its own cluster
        - At each step, merge the two closest clusters
        - Stop when every point is in the same cluster
    - Divisive:
        - Start with every point in the same cluster
        - At each step, split until every point is in its own cluster
- Agglomerative Clustering Algorithm:
    - 1.Let each point in the dataset be in its own cluster
    - 2.Compute the distance between all pairs of clusters
    - 3.Merge the two closest clusters
    - Repeat 3 & 4 until all points are in the same cluster
- Hierarchical clustering:
    - At every step, we record which clusters were merged in order to produce a dendrogram:
    - We can “cut” the dendrogram at any threshold to produce any number of clusters
        - Finding the threshold with which to cut the dendrogram requires exploration and tuning. But in general hierarchical clustering is used to expose a hierarchy in the data (ex: finding/defining species via DNA similarity).
        - To capture the difference between clusterings you can use a cost function, or methods that we will discuss later when we look at clustering aggregation.
- Distance functions:
    - Let’s first define:
    - Distance between points: $\mathbf{d(p_1, p_2)}$
    - Distance between clusters: $\mathbf{D(C_1, C_2)}$
- single - link distance:
    - Is the **minimum** of all pairwise distances between a point from one cluster and a point from the other cluster.
        - $D_{SL}(C_1, C_2) = \min\{d(p_1, p_2) \mid p_1 \in C_1, p_2 \in C_2\}$
        - Can handle clusters of different sizes
    - But… Sensitive to noise points; Tends to create elongated clusters
- Complete-link distance:
    - Is the **maximum** of all pairwise distances between a point from one cluster and a point from the other cluster.
        - $D_{CL}(C_1, C_2) = \max\{d(p_1, p_2) \mid p_1 \in C_1, p_2 \in C_2\}$
    - Less susceptible to noise; Creates more balanced (equal diameter) clusters
    - But… Tends to split up large clusters; All clusters tend to have the same diameter
- Average-Link Distance:

    - the **average** of all pairwise distances between a point from one cluster and a point from the other cluster.
        - $$D_{AL}(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{p_1 \in C_1, p_2 \in C_2} d(p_1, p_2)$$
        - Less susceptible to noise and outliers.
        - But… Tends to be biased toward globular clusters
- Centroid Distance:
    - The distance between the centroids of clusters.
    - $$D_C(C_1, C_2) = d(\mu_1, \mu_2)$$
- Ward's Distance:
    - difference between the spread / variance of points in the merged cluster and the unmerged clusters
- $$D_{WD}(C_1, C_2) = \sum_{p \in C_{12}} d(p, \mu_{12}) - \sum_{p_1 \in C_1} d(p_1, \mu_1) - \sum_{p_2 \in C_2} d(p_2, \mu_2)$$