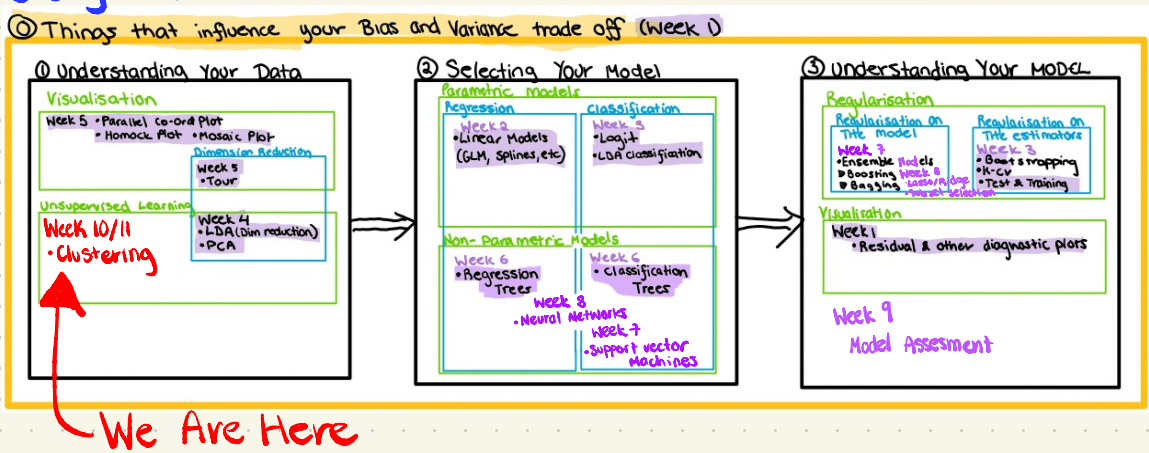
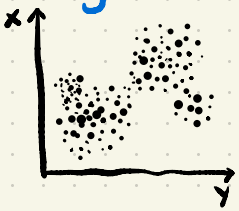
Week 10 : Hierarchical Clustering

① Big Picture



We Are Here

② What Do We Need



- How Many Clusters?
- Where are the lines drawn?

↳ Need a formal way to group observations together

↳ Must Define
- Method (Hierarchical)
- Linkage
- Distance.

③ Hierachical Clustering

Idea: Every observation is its own cluster and we combine them 1 by 1.

Algorithm 12.3 *Hierarchical Clustering*

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.

Algorithm

← Start with own clusters

← Fuse two most similar clusters

← Repeat until we have no more clusters.

Textbook Pg 526

When the clusters contain multiple observations, how do we define its distance to other clusters?