Machine learning: Clustering

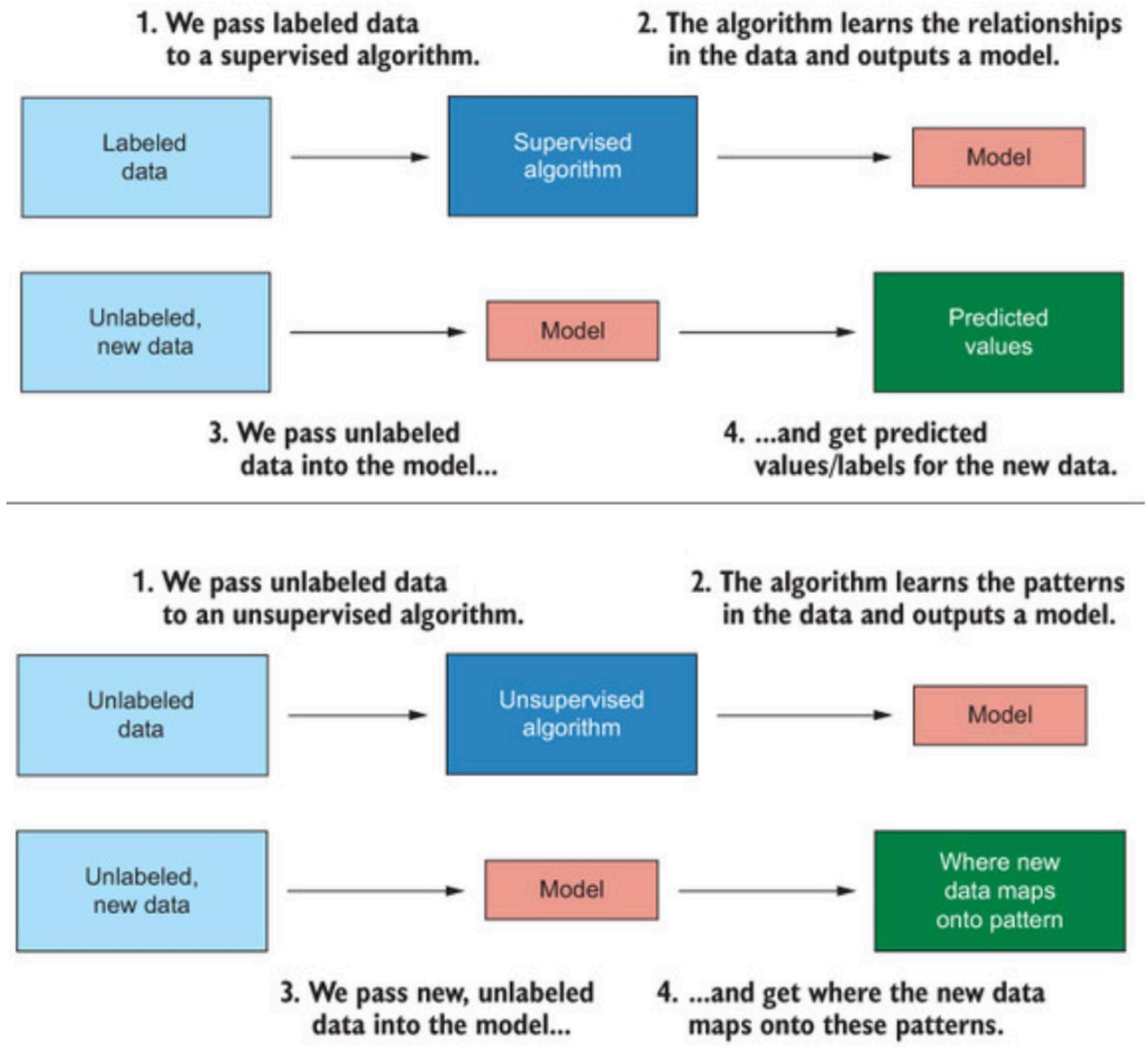
Alex Di Genova

09/07/2024

Machine learning algorithms

Classes

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Dimension Reduction
  - Clustering
- Semi-supervised

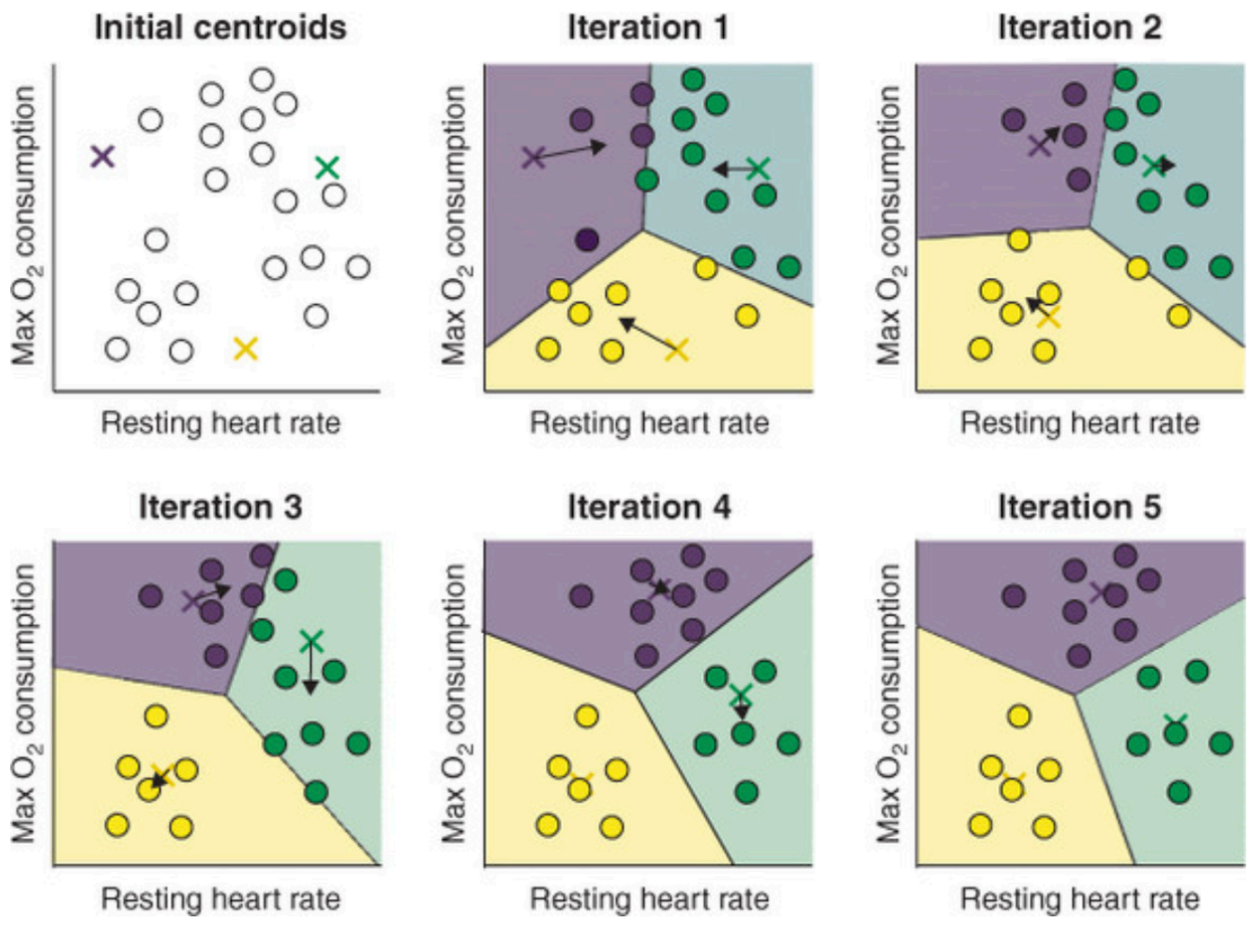


What is clustering?

- A cluster is a set of cases that are more similar to each other than they are to cases in other clusters.
- We use clustering when we don’t have any prior knowledge about class membership or whether there are distinct classes in the data
- Methods:
  - Clustering by finding centers with k-means
  - Hierarchical clustering
  - Clustering based on density: DBSCAN and OPTICS
  - Clustering based on distributions with mixture modeling

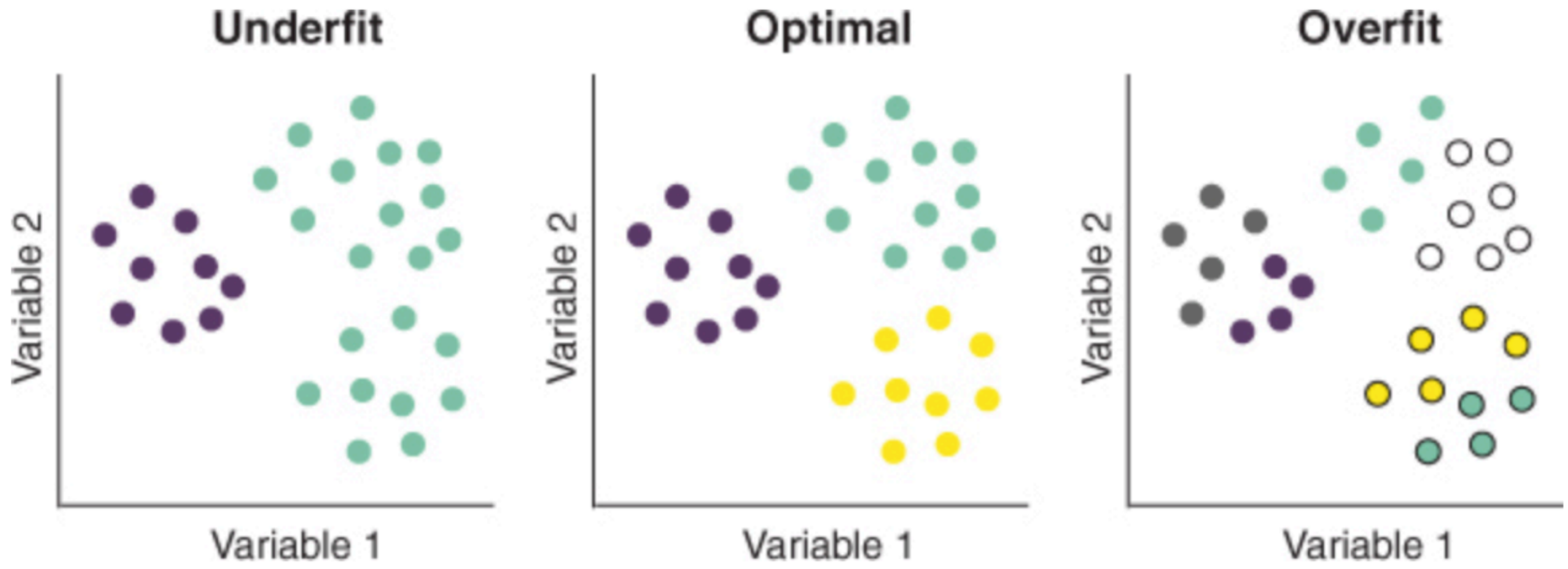
K-Means

- **K-means clustering** is an unsupervised machine learning algorithm used to partition a dataset into k distinct clusters.
    - **Initialization**: Choose the number of clusters k and randomly initialize k centroids (cluster centers).
    - **Assignment**: Assign each data point to the nearest centroid based on the Euclidean distance. This forms k clusters.
    - **Update**: Recalculate the centroids as the mean of all points assigned to each cluster.
    - **Repeat**: Repeat the assignment and update steps until the centroids no longer change significantly or a maximum number of iterations is reached.
    - **Convergence**: The algorithm converges when the centroids stabilize or the changes in their positions fall below a threshold.
- The result is a set of clusters with minimized intra-cluster variance (or squared differences).



K-Means

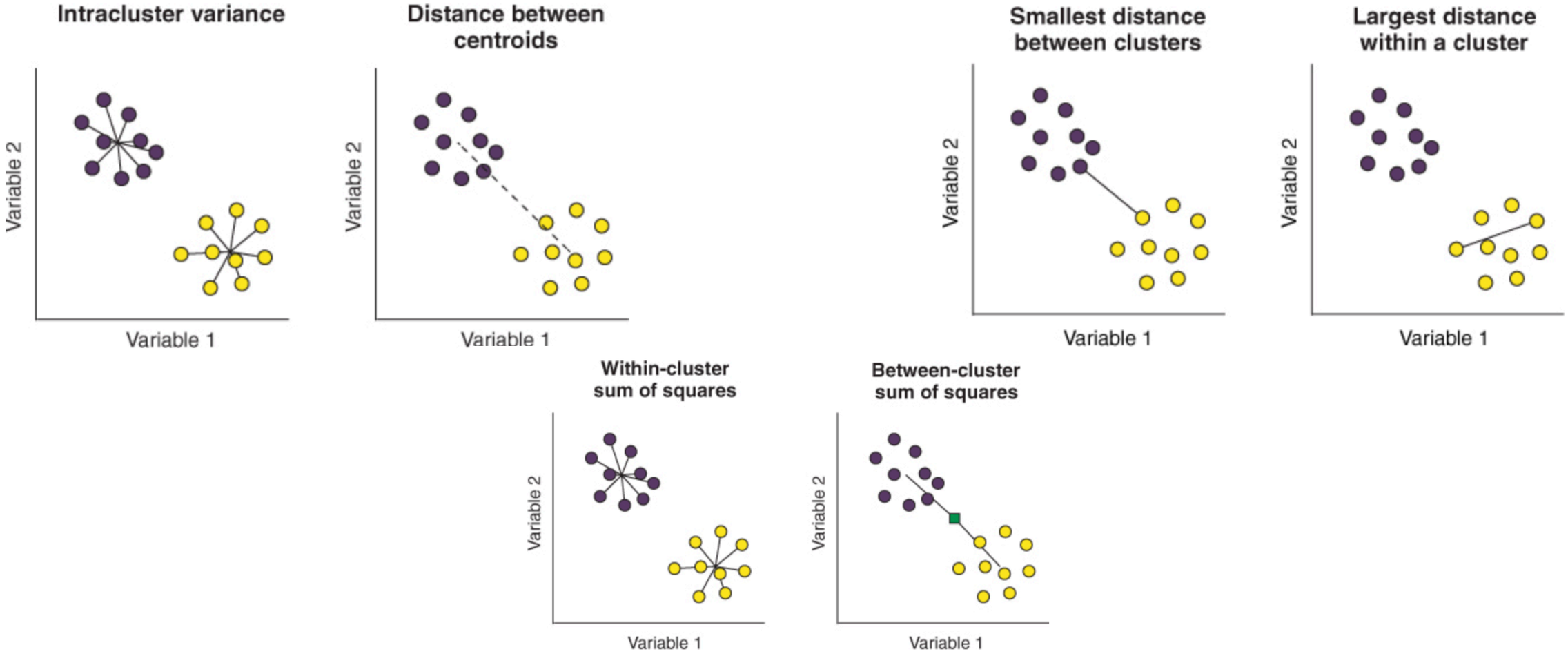
How to choose the number of clusters



- “good-quality” clusters typically mean that each cluster is as compact as possible, while the distances between clusters are as large as possible.
- A common approach is to compute multiple clustering models over a range of cluster numbers and compare the cluster metrics for each model to help choose the best-fitting one
    - Davies-Bouldin index
    - Dunn index
    - Pseudo F statistic

K-Means

How to choose the number of clusters



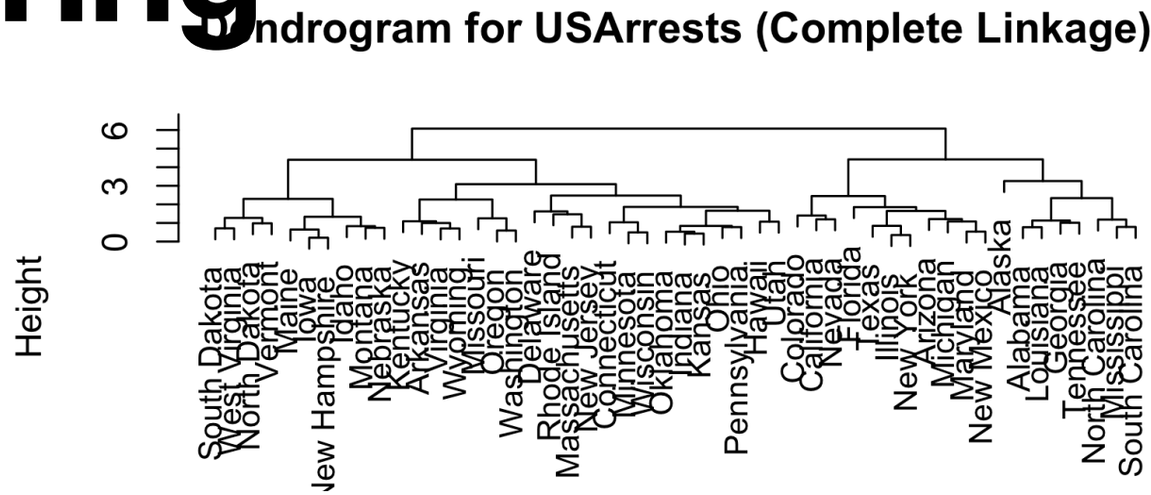
- Davies-Bouldin index: quantifies the average separability of each cluster from its nearest counterpart (The smaller the better).

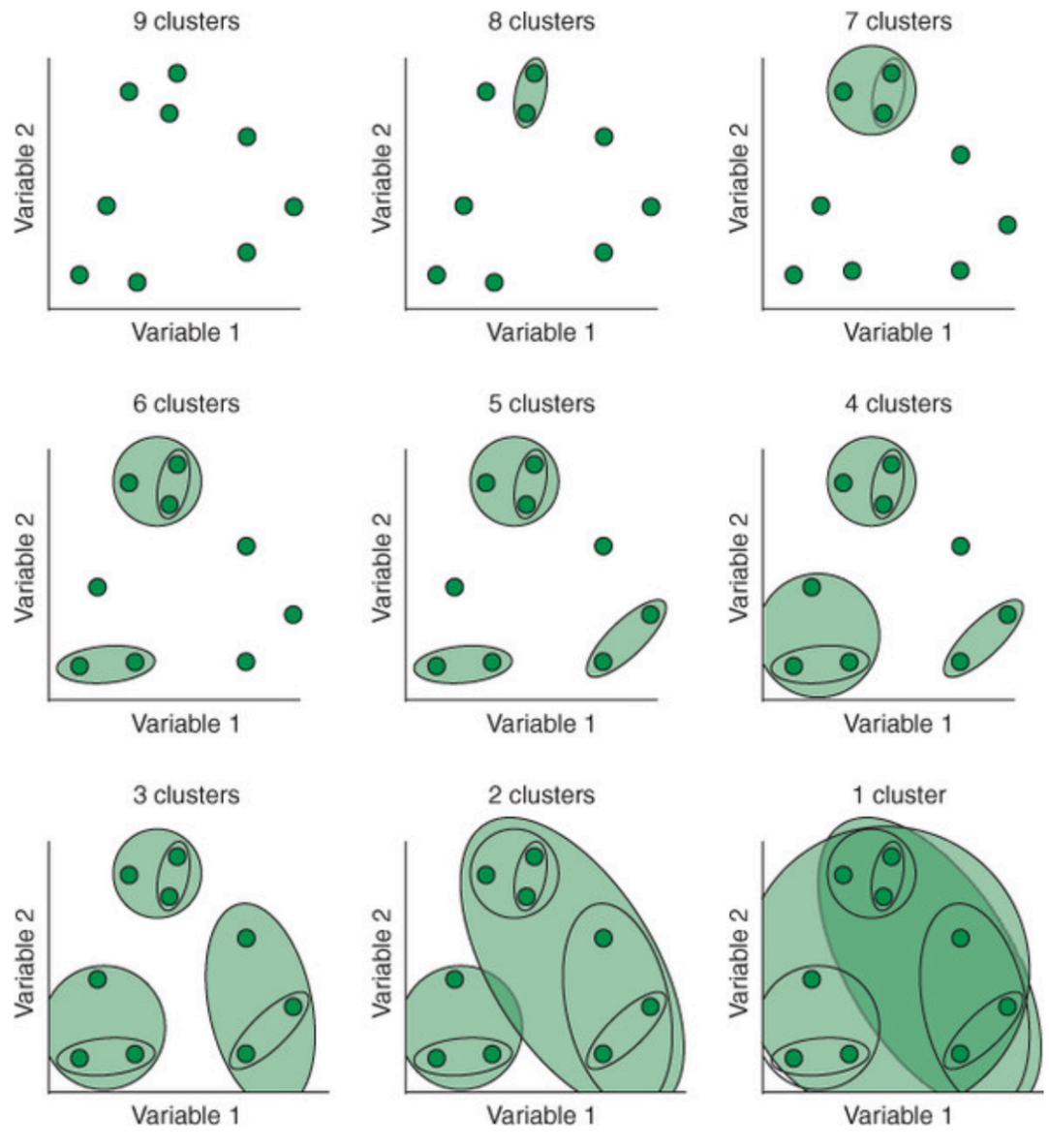
- Dunn index: quantifies the ratio between the smallest distance between points in different clusters, and the largest distance within any of the clusters (The higher the better).

- Pseudo F statistic: quantifies the ratio of between-cluster variance to within-cluster variance (The higher the better).

Hierarchical clustering

- A tree of clusters within clusters
- Agglomerative (Bottom-Up) Approach
  - Starts with each data point as a single cluster
  - Iteratively merges the closest pairs of clusters.
  - Continues until all points are merged into a single cluster or a specified number of clusters is reached.
- Divisive (Top-Down) Approach:
  - Starts with all data points in a single cluster.
  - Iteratively splits the most appropriate cluster into smaller clusters.
  - Continues until each point is its own cluster or a specified number of clusters is reached.





AHC

Steps

Steps in Agglomerative Hierarchical Clustering:

1. **Compute the Distance Matrix**:
    - Calculate the pairwise distances between all data points using a distance metric (e.g., Euclidean distance).
2. **Create Clusters**:
    - Each data point starts as its own cluster.
3. **Merge Clusters**:
    - Find the pair of clusters with the smallest distance between them and merge them.
    - Update the distance matrix to reflect the merge.
4. **Repeat**:
    - Repeat the merging process until all points are in one cluster.
5. **Form the Dendrogram**:
    - A dendrogram is a tree-like diagram that records the sequences of merges or splits. It helps visualize the clustering process and decide the number of clusters.

Book:

Machine learning with R

Questions?

Practice!!!