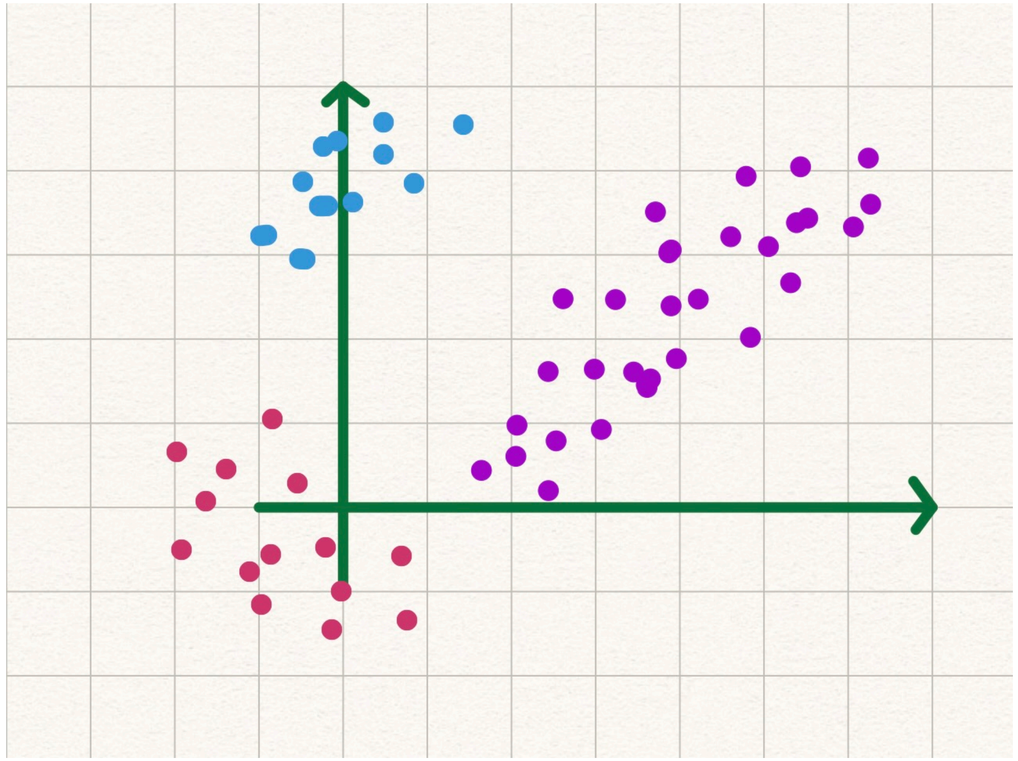
Applied Machine Learning

Clustering

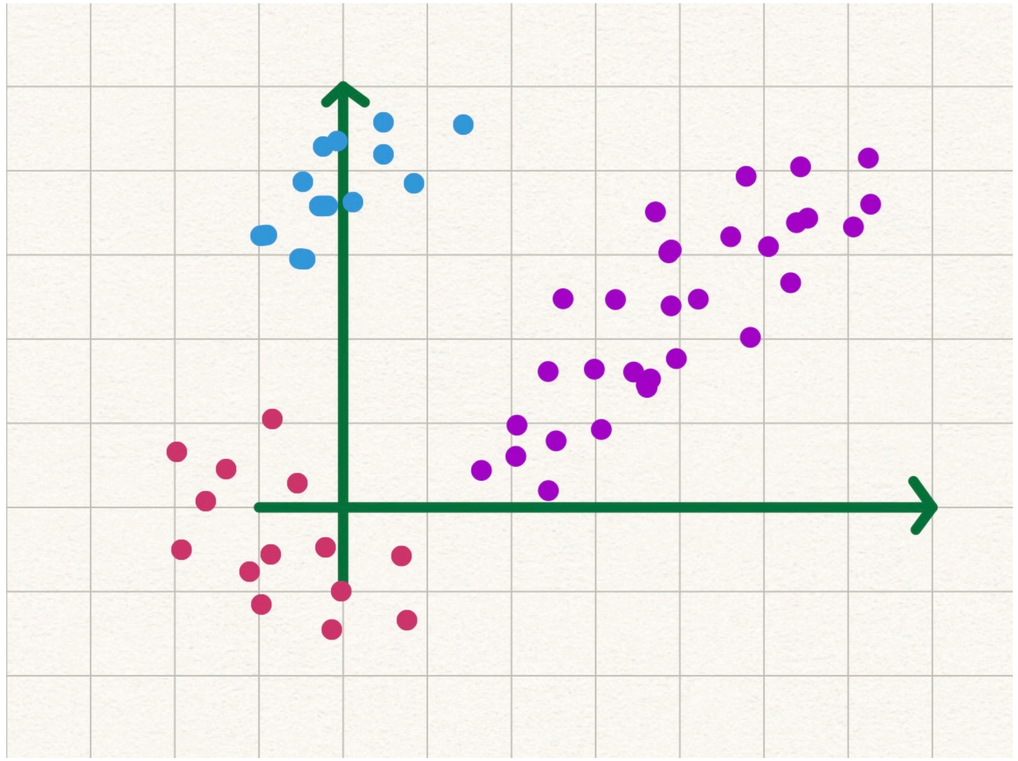
Clustering

- Overview
- Agglomerative Clustering
- Divisive Clustering

The Curse of Dimensionality

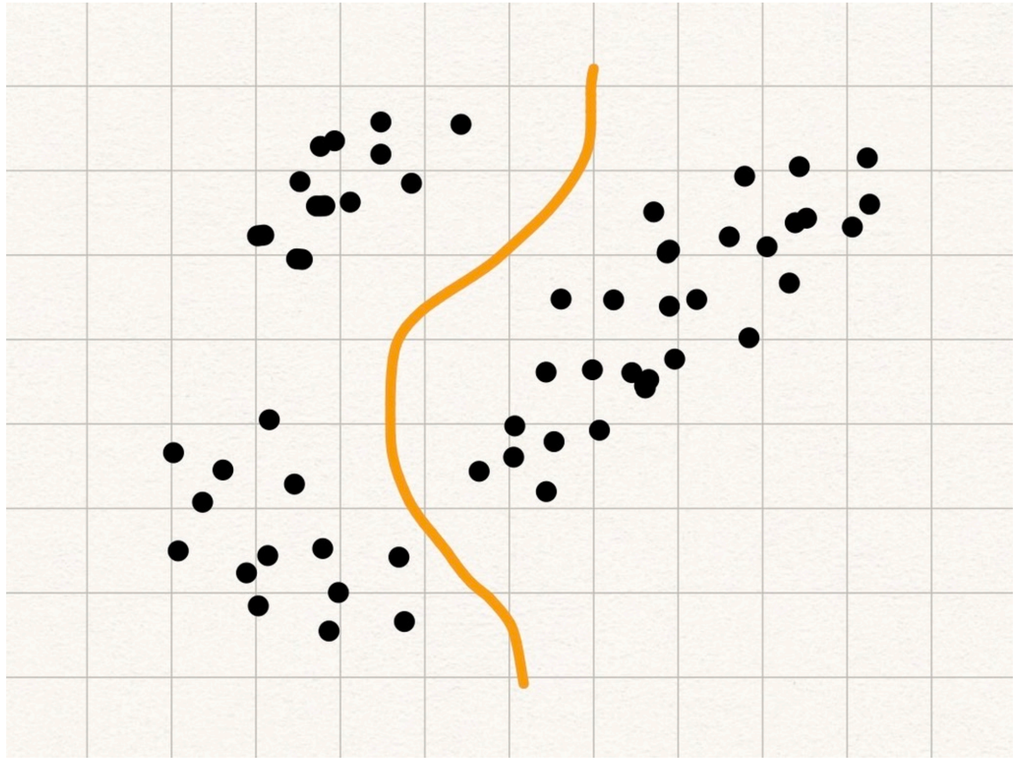
- Data in high-dimensions
  - as d increases, distances increase
  - mean and covariance
    - computational cost grows
    - need more data to get accurate estimations
- 

Clustering

- Group similar data into Clusters of Blobs
  - Highlight underlying patterns
  - Pre-processing step
- Cluster parameters
  - Cluster representation
  - Assignment of data items to cluster
  - Number of clusters
  - Evaluation of clusters
- 

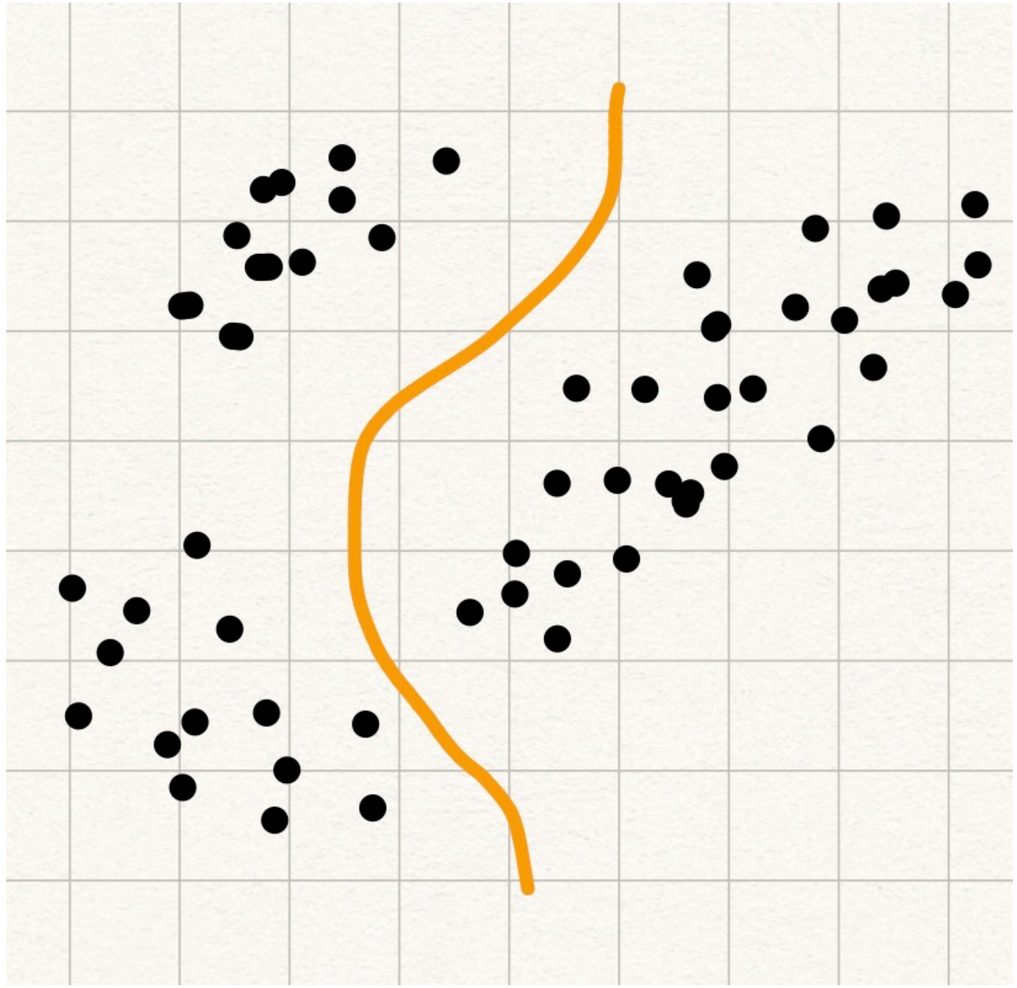
Types of Clustering Algorithms

- Divisive Clustering
  - start with a cluster for the whole dataset
  - repetitive split of clusters
- Agglomerative Clustering
  - start with a cluster per item
  - repetitive merging of clusters



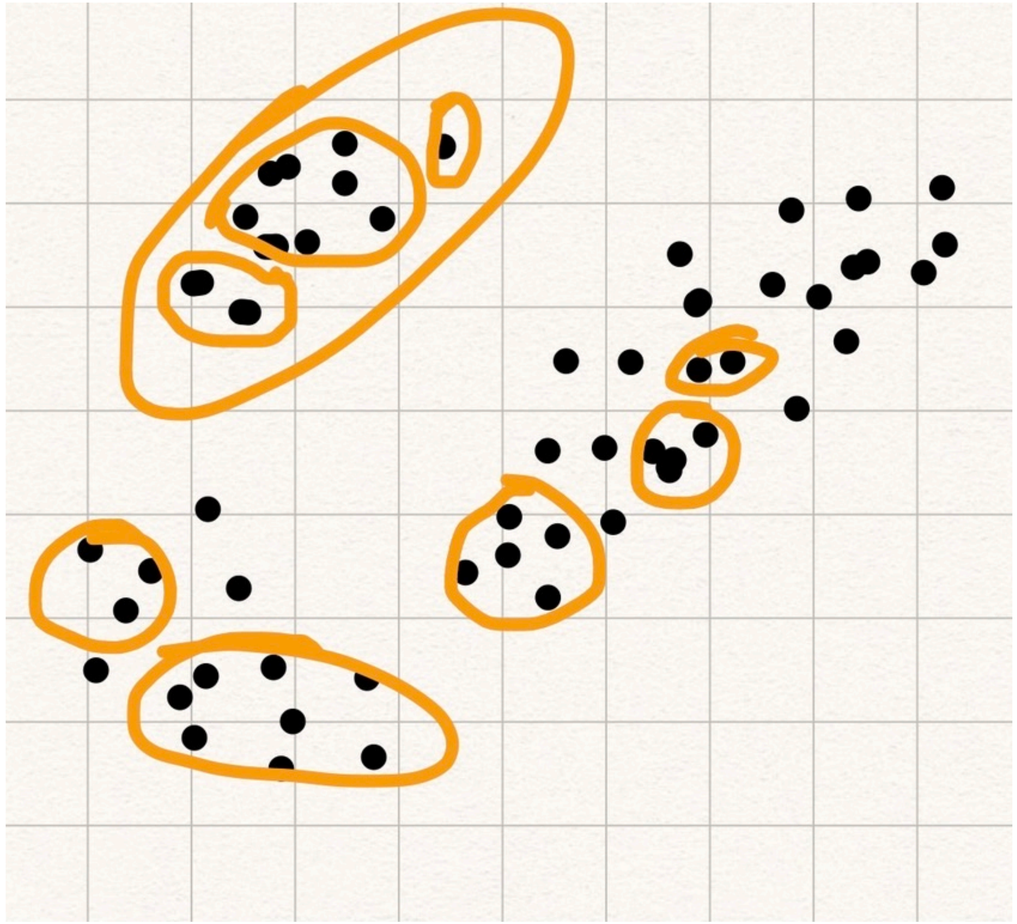
Divisive Clustering

- Splitting criteria
  - What cluster to split
  - Where to split the selected cluster
  - Application dependent



Agglomerative Clustering: Distances

- Computation of inter-cluster distance for clusters A and B
  - single-link clustering
    - $d(A, B) = \min_{a \in A, b \in B} d(a, b)$
  - complete-link clustering
    - $d(A, B) = \max_{a \in A, b \in B} d(a, b)$
  - group average clustering
    - $d(A, B) = \frac{\sum_{a \in A, b \in B} d(a, b)}{|A||B|}$

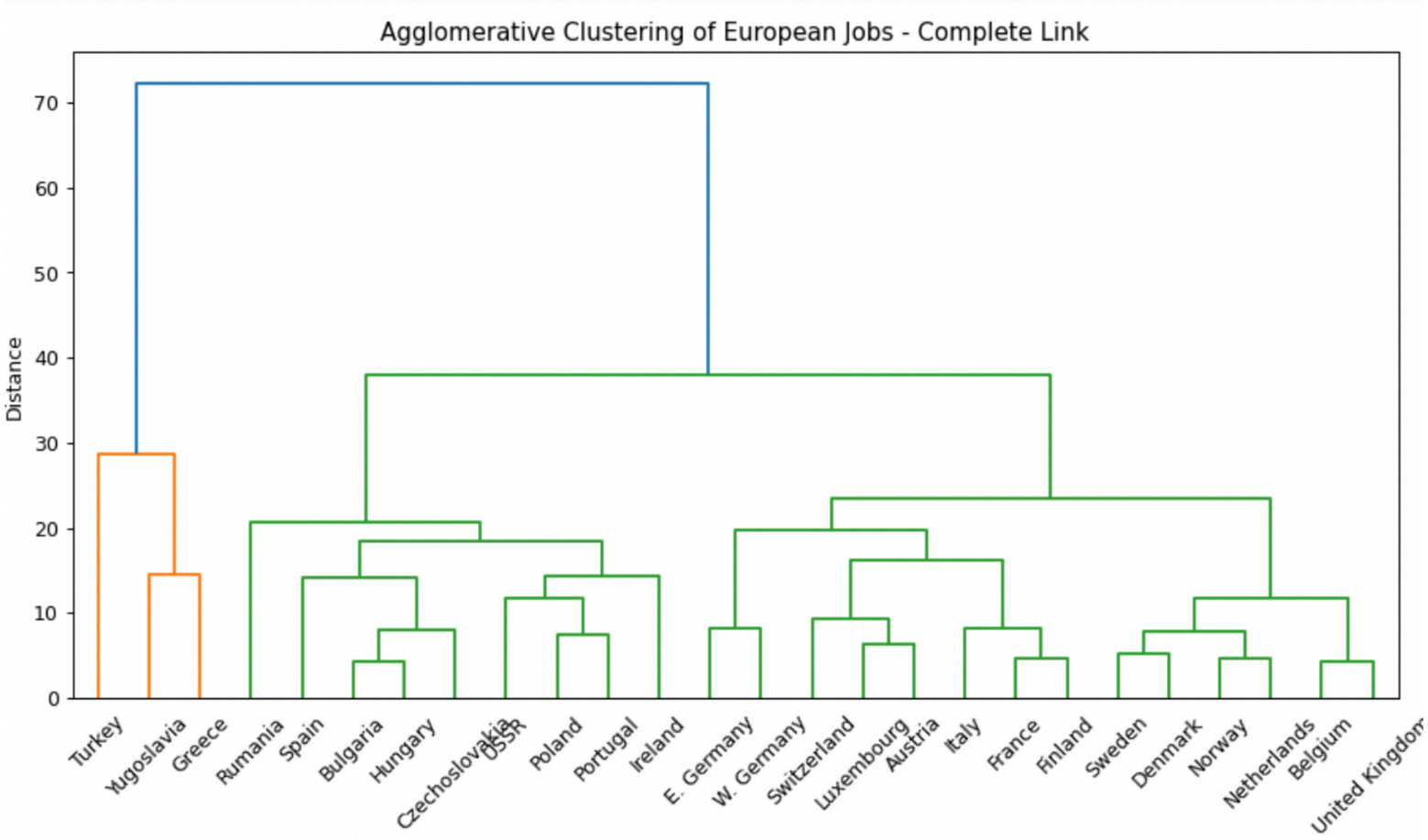


Agglomerative Clustering: Feature scale

- Scale to unit variance
  - original dataset X
  - new dataset Y:
    - $\mathbf{y}_i = \frac{\mathbf{x}_i - mean(X)}{\sigma_X^2}$
- De-correlation
  - original dataset X
  - $\Sigma = \text{Covmat}(X)$: compute eigenvectors U and eigenvalues Λ
  - new dataset Y:
    - $\mathbf{y}_i = (\Lambda^{\frac{1}{2}})^{-1} U^\top \left(\mathbf{x}_i - mean(X)\right)$

Dendogram

- hierarchy of clusters
- x: data items
- y: inter-cluster distance
- Useful to select level of clustering



Clustering

- Overview
- Agglomerative Clustering
- Divisive Clustering

Applied Machine Learning

Clustering