

Lecture 9: Clustering

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

01 Clustering: Overview

02 K-Means Clustering

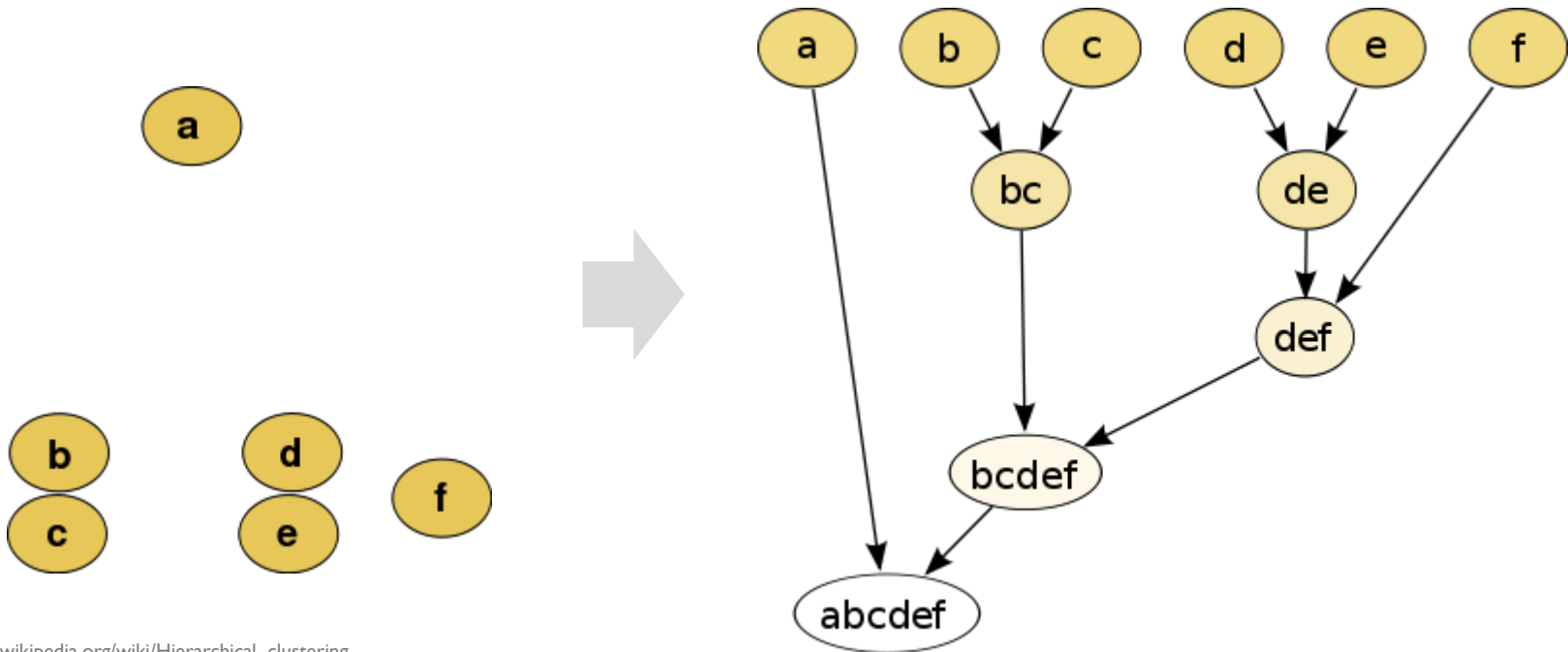
03 Hierarchical Clustering

04 Density-based Clustering: DBSCAN

04 R Exercise

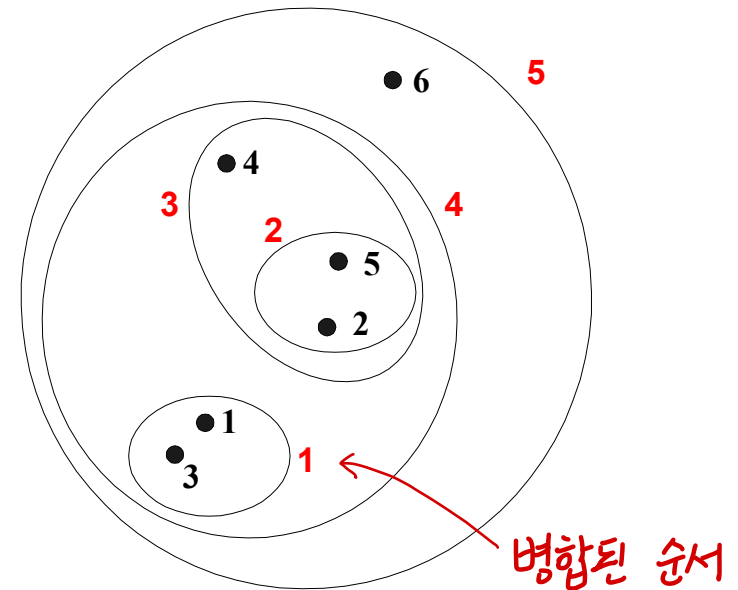
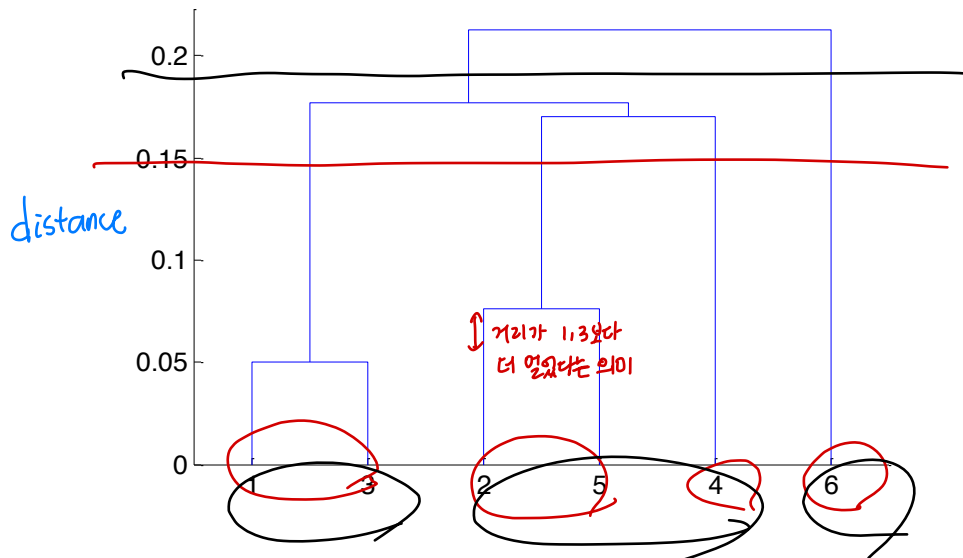
Hierarchical Clustering

- Hierarchical clustering
 - ✓ Produces a set of nested clusters organized as a hierarchical tree
 - ✓ Can be visualized as a dendrogram
 - A tree like diagram that records the sequences of merges or splits



Hierarchical Clustering

- Hierarchical clustering
 - ✓ Produces a set of nested clusters organized as a hierarchical tree
 - ✓ Can be visualized as a dendrogram
 - A tree like diagram that records the sequences of merges or splits



Hierarchical Clustering

- Strengths of Hierarchical clustering

- ✓ Do not have to assume any particular number of clusters

- Any desired number of clusters can be obtained by **'cutting'** the dendrogram at the proper level

- ✓ May correspond to meaningful taxonomies

- Two main types of hierarchical clustering

- 이해하기 ✓ Agglomerative clustering 상향식

- Start with the points as individual clusters
 - At each step, merge the closest pair of clusters until only one cluster left

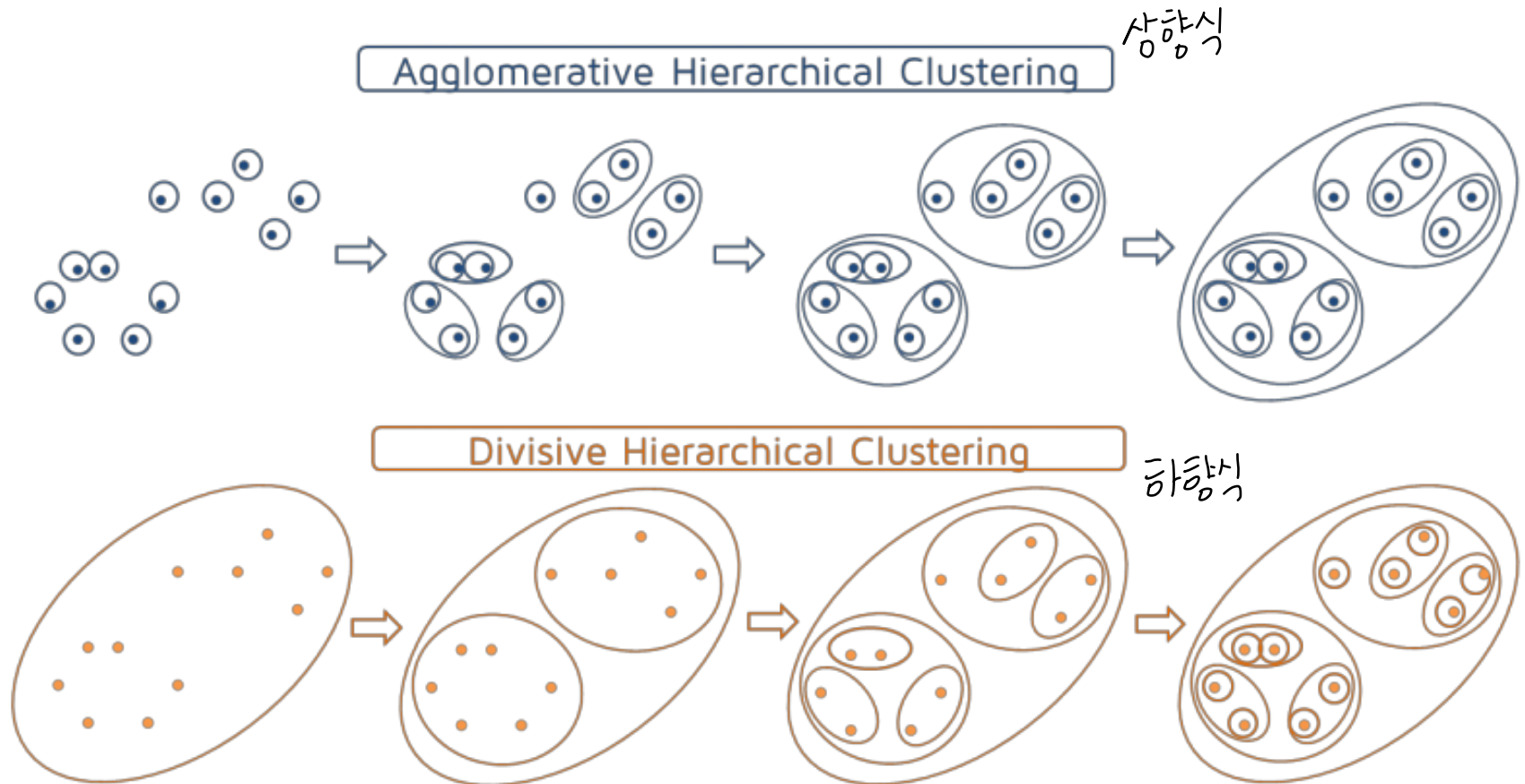
- ✓ Divisive clustering 하향식

- Start with one, all-inclusive cluster
 - At each step, split a cluster until each cluster contains a point

Hierarchical Clustering

- Strengths of Hierarchical clustering

✓ Agglomerative clustering vs. Divisive clustering

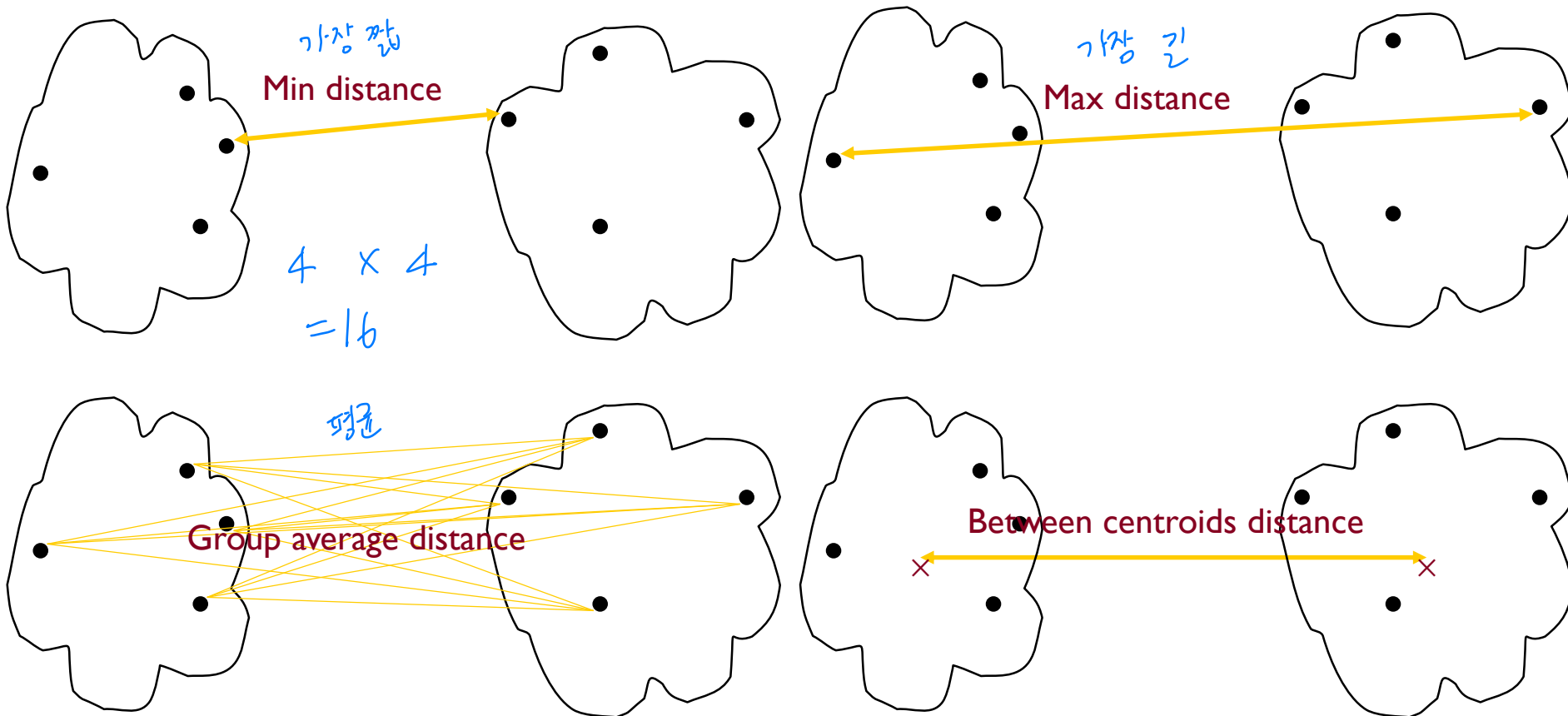


Hierarchical Clustering

- Agglomerative clustering algorithm

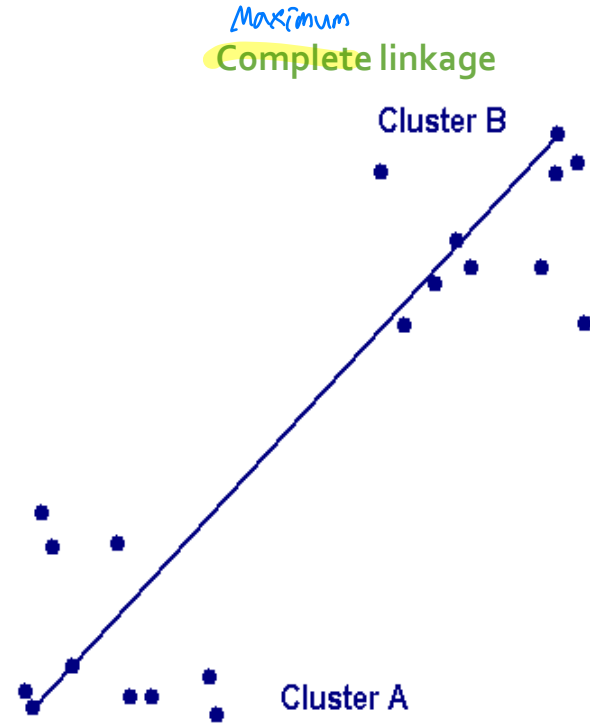
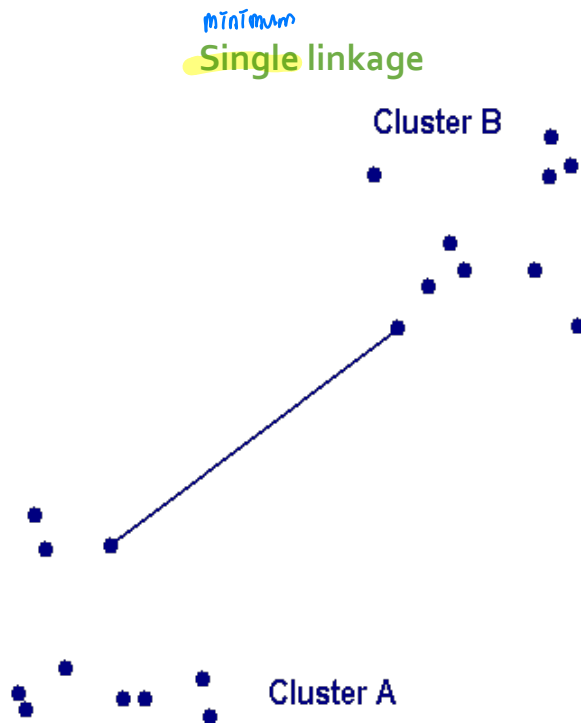
- ✓ Key operation: computation of the proximity of two clusters

- Min, max, group average, between centroid, etc.



Hierarchical Clustering

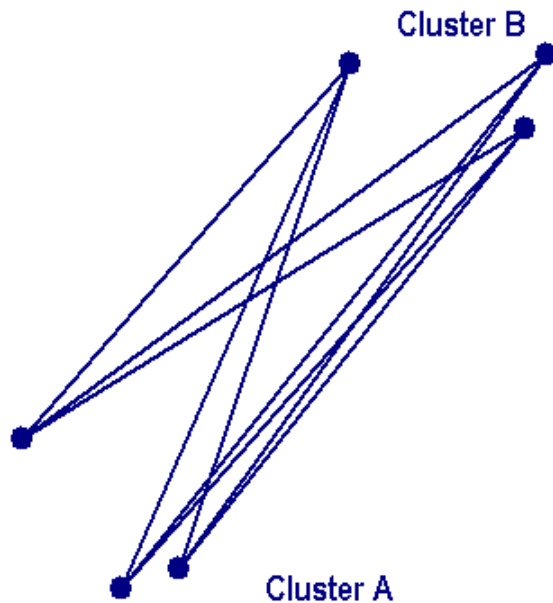
- Agglomerative clustering algorithm
 - ✓ Single linkage: minimum distance between two data points in different clusters
 - ✓ Complete linkage: maximum distance between two data points in different clusters



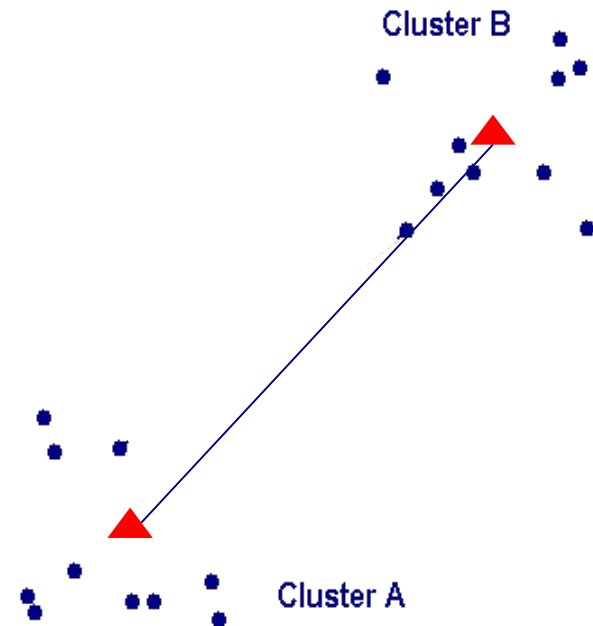
Hierarchical Clustering

- Agglomerative clustering algorithm
 - ✓ Average linkage: mean distance between two data points in different clusters
 - ✓ Centroid linkage: distance between centroids in different clusters

Average linkage



Centroid linkage



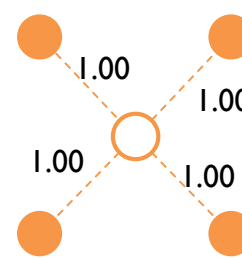
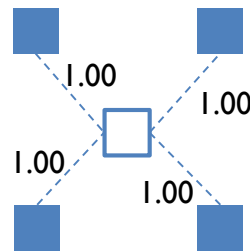
Hierarchical Clustering

- Agglomerative clustering algorithm

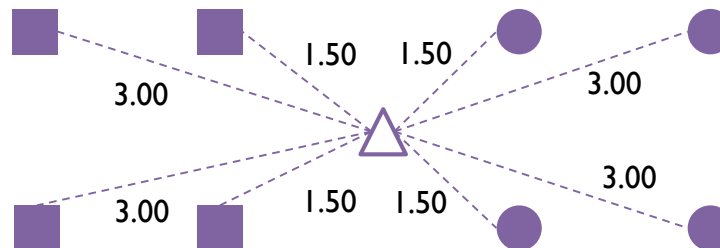
병합된 점과 후의 분산

✓ Ward method: Compare the sum of squared error (SSE) before and after the merge

- SSE before merge: $1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 = 8$



- SSE after merge: $4 \times 1.5^2 + 4 \times 3^2 = 45$



얼마나 커졌는가?

병합 후에도 작은것이
먼저 병합

- Ward distance: $45 - 8 = 37$

군집내 평방중의 제곱합

Hierarchical Clustering

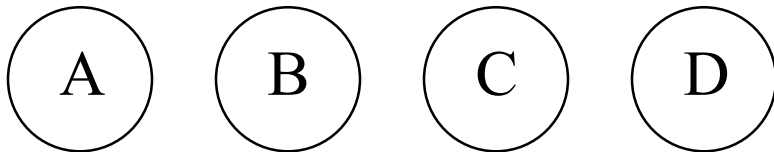
- Agglomerative Clustering Procedure

- ✓ Step 1: Assume that each data point is an individual cluster, compute the cluster distance
- ✓ Step 2: Repeat the following procedure
 - Step 2-1: Merge the two closest clusters
 - Step 2-2: Update the cluster distance matrix → cluster 가 1개만 남을 때까지
- ✓ When all data points are merged as a single cluster, stop

Hierarchical Clustering

- Example

Initial Data Items



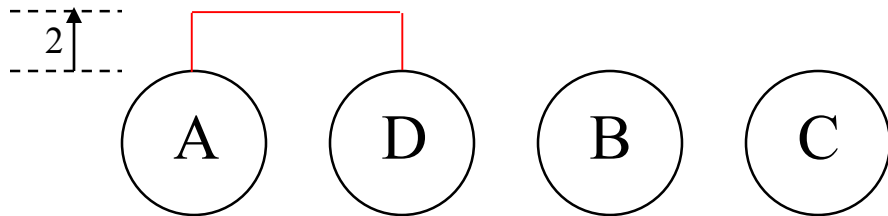
Distance Matrix

Dist	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

Hierarchical Clustering

- Example

Current Clusters



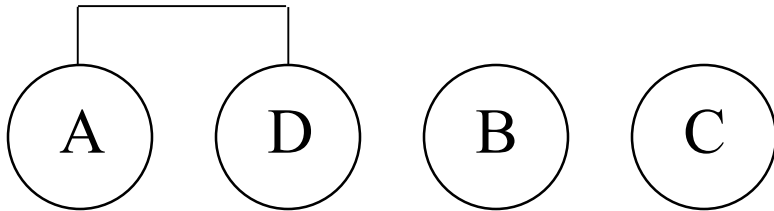
Distance Matrix

Dist	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

Hierarchical Clustering

- Example

Current Clusters



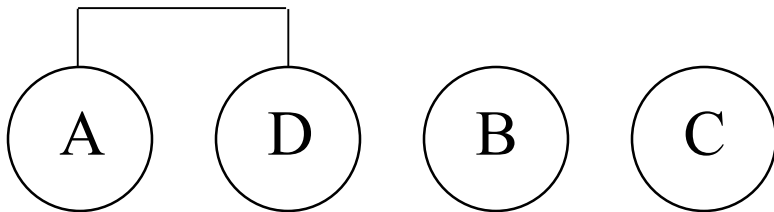
Distance Matrix

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Hierarchical Clustering

- Example

Current Clusters



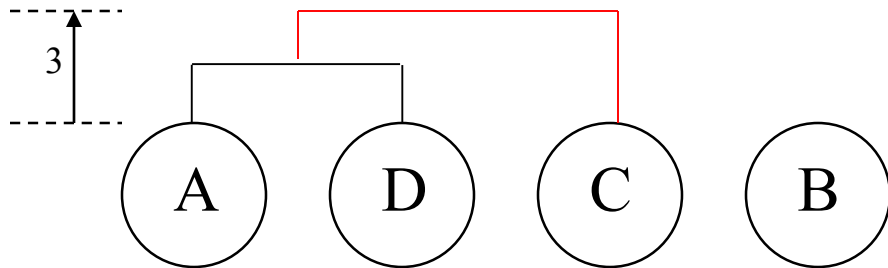
Distance Matrix

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Hierarchical Clustering

- Example

Current Clusters



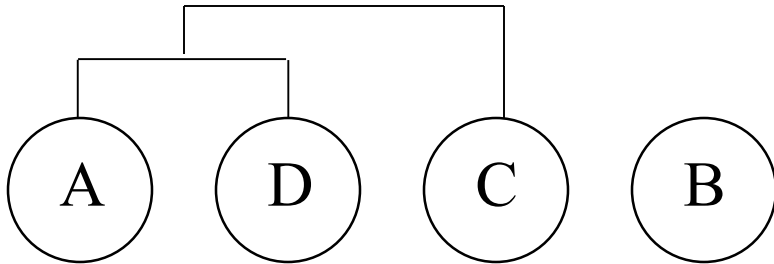
Distance Matrix

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Hierarchical Clustering

- Example

Current Clusters



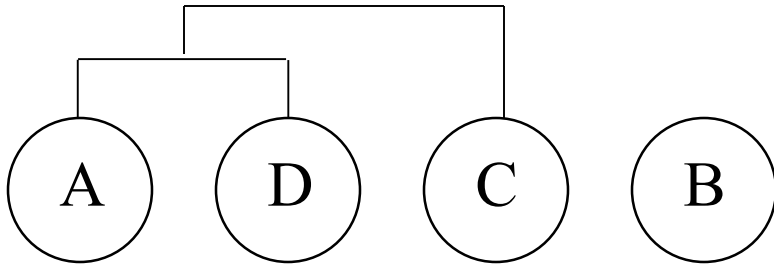
Distance Matrix

Dist	AD C	B		
AD C		10		
B				

Hierarchical Clustering

- Example

Current Clusters



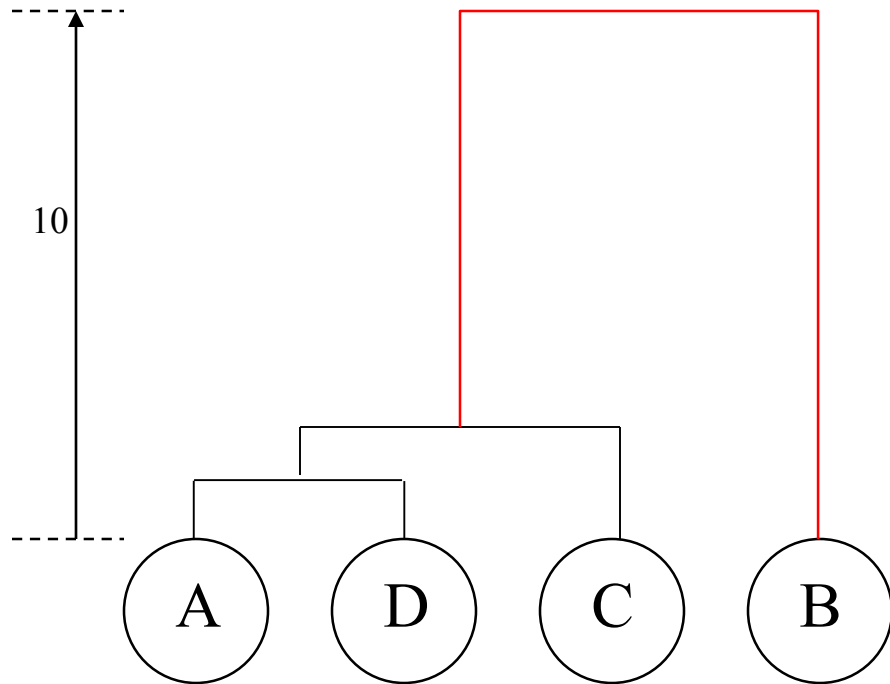
Distance Matrix

Dist	AD C	B		
AD C		10		
B				

Hierarchical Clustering

- Example

Current Clusters



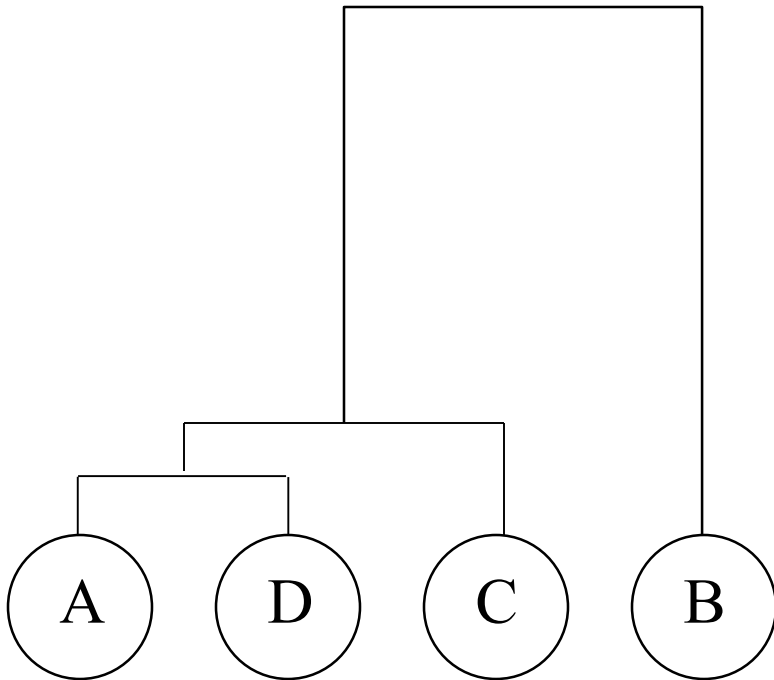
Distance Matrix

Dist	AD C	B		
AD C		10		
B				

Hierarchical Clustering

- Example

Final Result

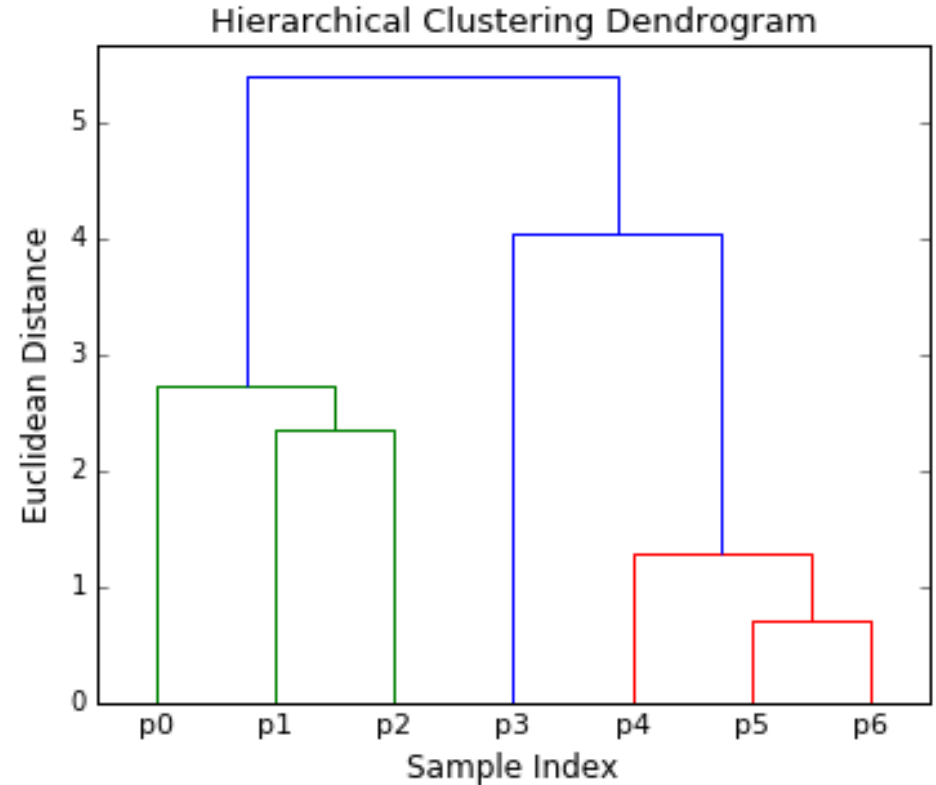
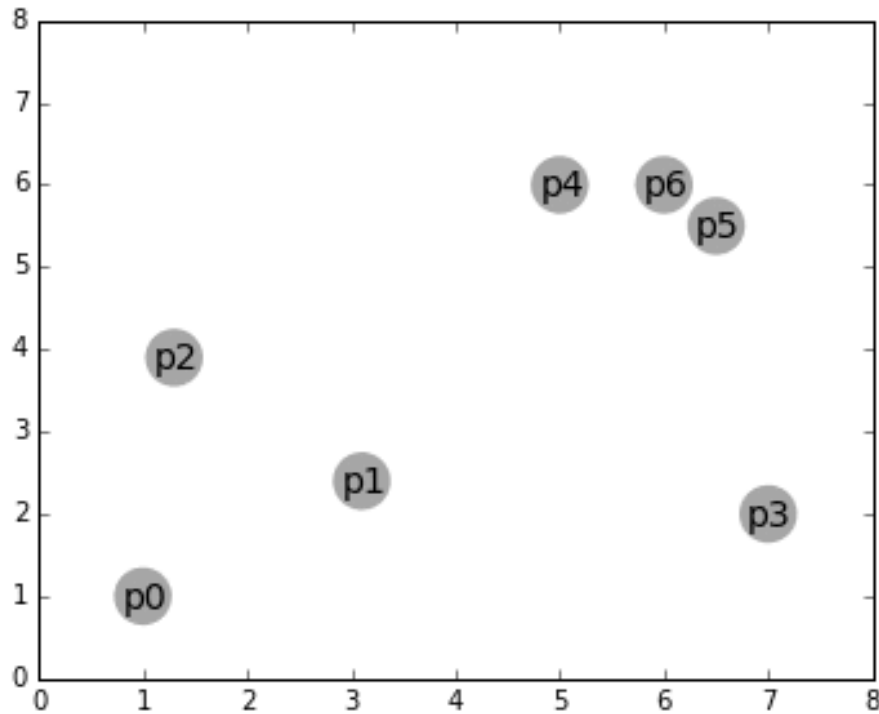


Distance Matrix

Dist	AD CB			
AD CB				

Hierarchical Clustering

- HC example



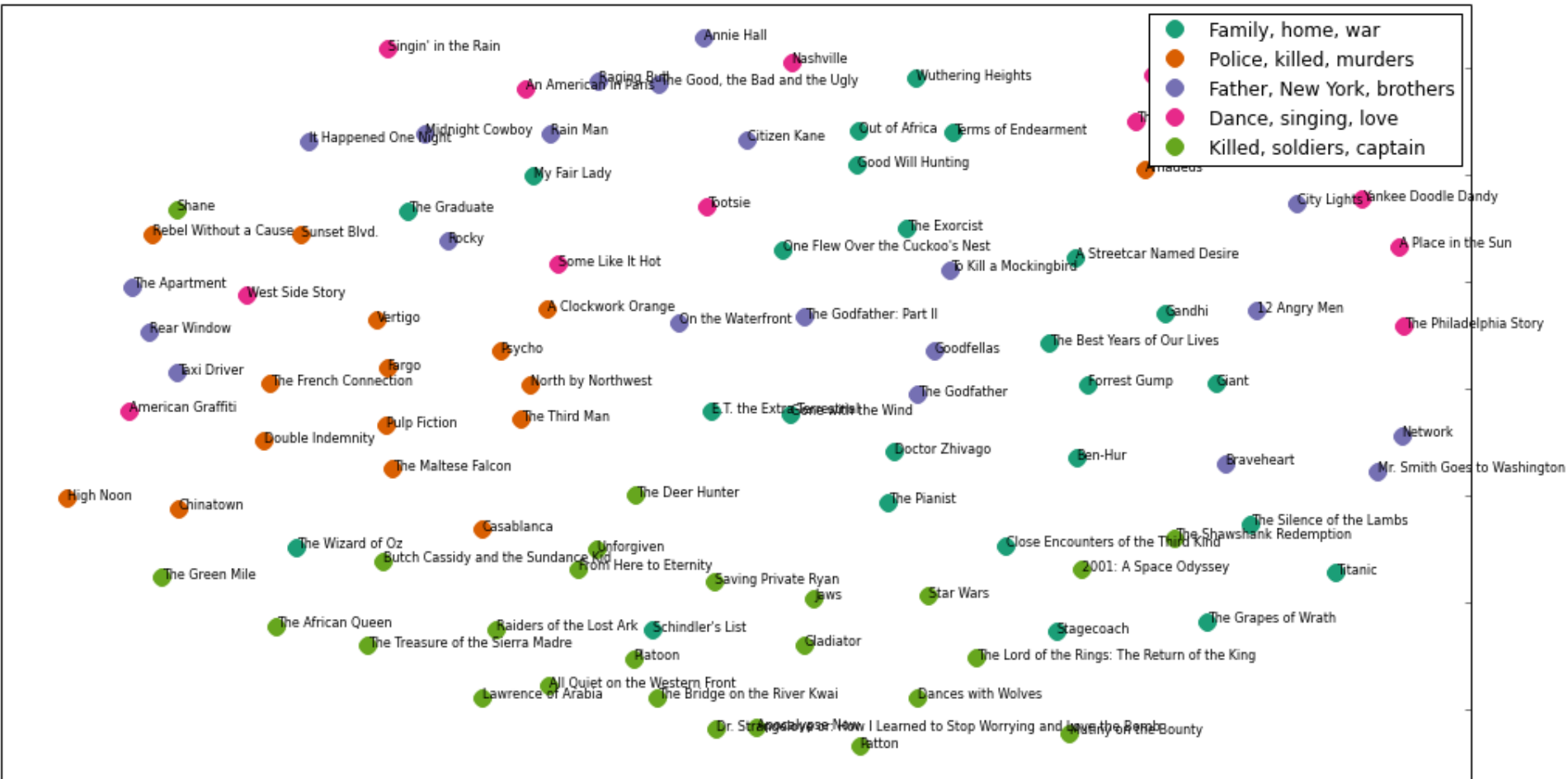
<https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>

Hierarchical Clustering

- Clustering top 100 film synopses (<http://brandonrose.org/clustering>)
 - ✓ Tokenizing and stemming each synopsis
 - ✓ Transforming the corpus into vector space using [tf-idf](#)
 - ✓ Calculating cosine distance between each document as a measure of similarity
 - ✓ Clustering the documents using the [k-means algorithm](#)
 - ✓ Using [multidimensional scaling](#) to reduce dimensionality within the corpus
 - ✓ Plotting the clustering output using [matplotlib](#) and [mpld3](#)
 - ✓ Conducting a hierarchical clustering on the corpus using [Ward clustering](#)
 - ✓ Plotting a Ward dendrogram
 - ✓ Topic modeling using [Latent Dirichlet Allocation \(LDA\)](#)

Hierarchical Clustering

- MDS result



Hierarchical Clustering

- Hierarchical clustering

