

# Lecture 9: Clustering

Pilsung Kang

School of Industrial Management Engineering

Korea University

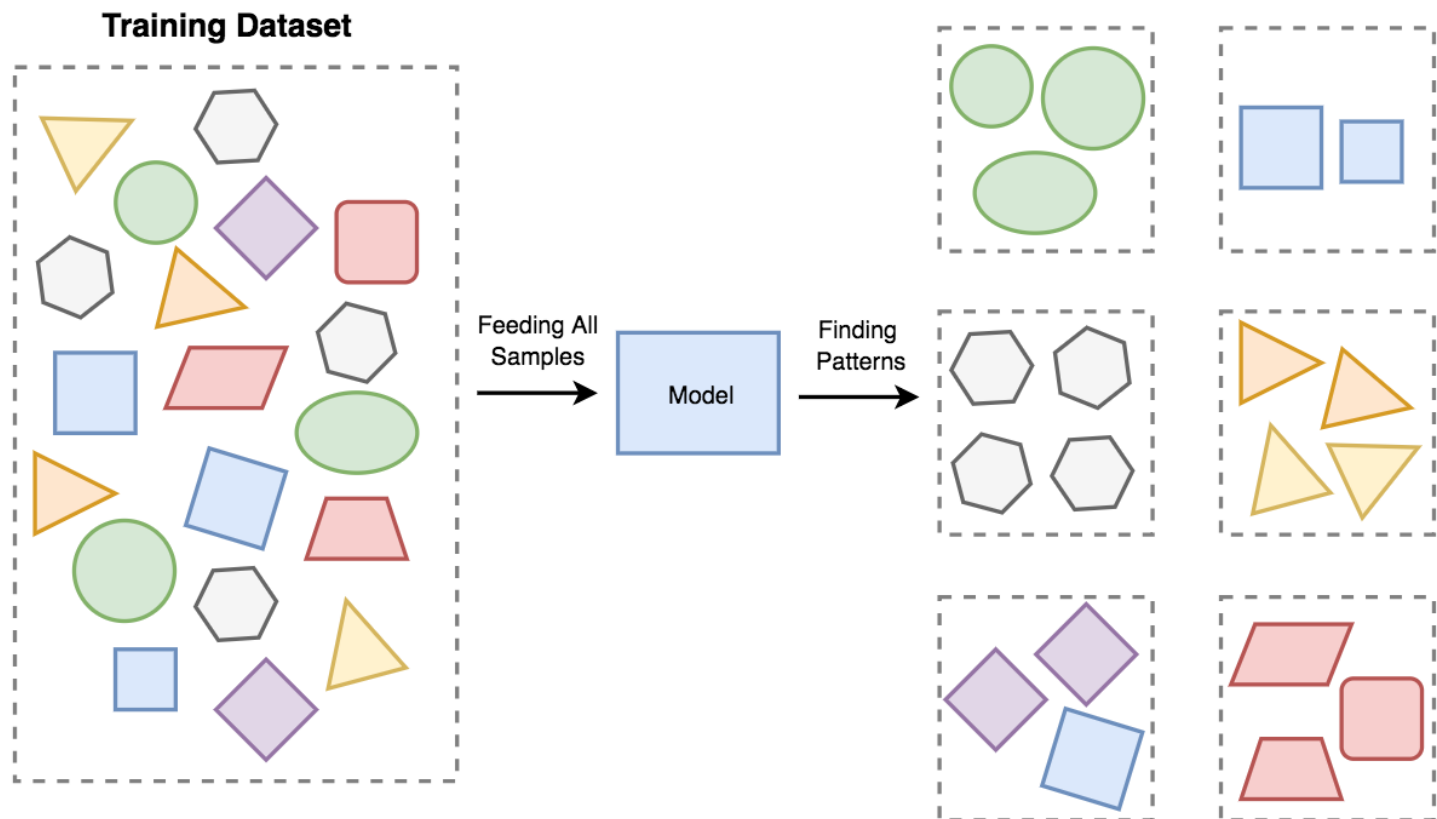
# AGENDA

- 01 Clustering: Overview
- 02 K-Means Clustering
- 03 Hierarchical Clustering
- 04 Density-based Clustering: DBSCAN

# Clustering: Overview

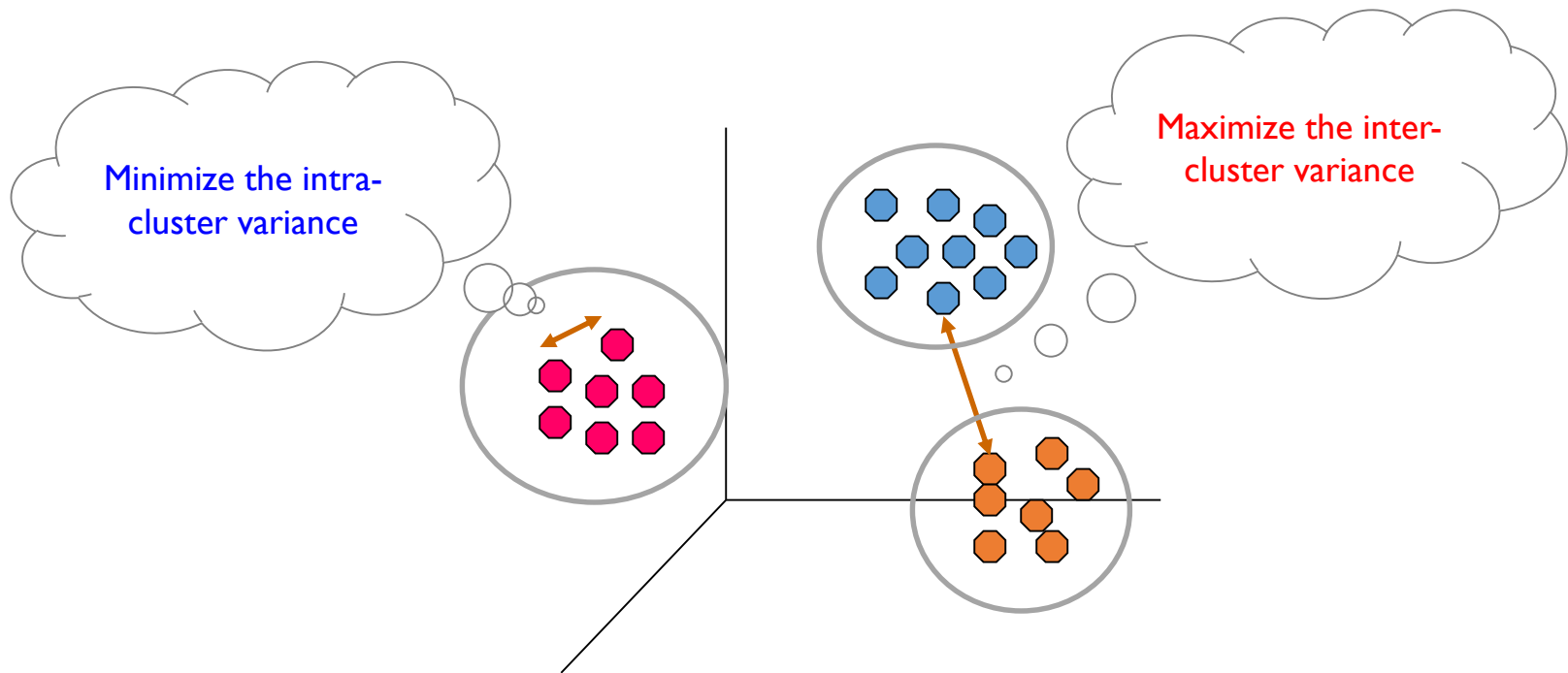
- Supervised vs. Unsupervised Learning

- ✓ Supervised: Find a function  $f$  that explains the relationship between the input  $X$  and the output  $Y$
- ✓ Unsupervised: Explore the features of the input  $X$



# Clustering: Overview

- What is clustering?
  - ✓ Find groups of objects such that the **objects in a group will be similar (or related) to one another** and **different from (or unrelated to) the objects in other groups**



# Clustering: Overview

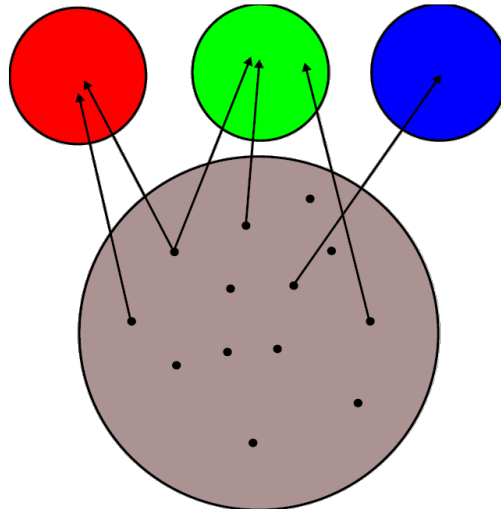
- Classification vs. Clustering

- ✓ Classification (supervised learning)

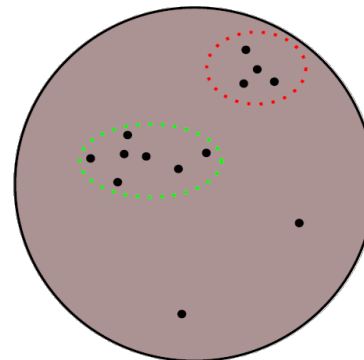
- The number of classes and the labels for all training instances are **known**
    - Goal is to find a function that links a set of input values to the target value

- ✓ Clustering (unsupervised learning)

- The number of clusters and memberships are **unknown**
    - Goal is to find an appropriate structure that can characterize the given dataset well



(a) Classification



(b) Clustering

# Clustering: Overview

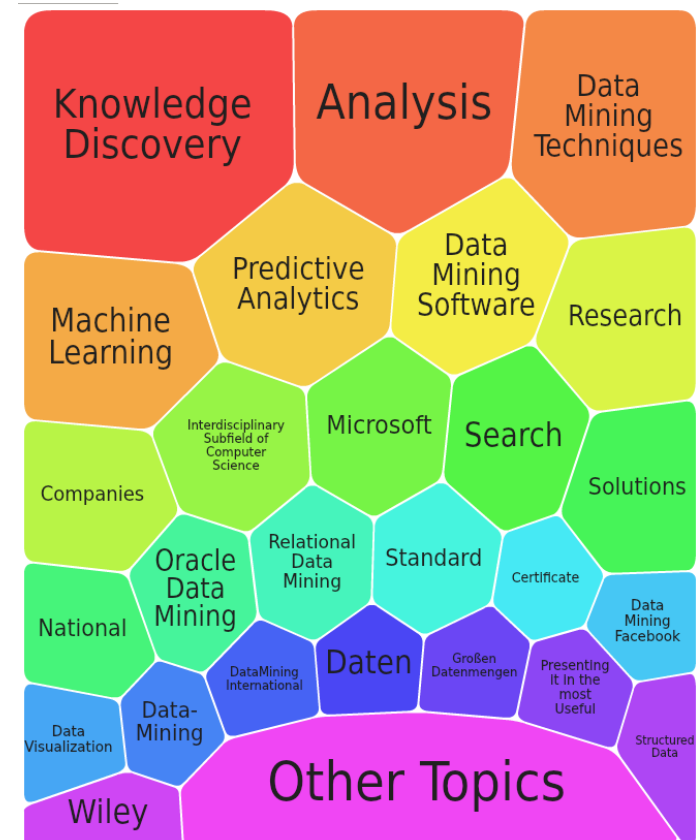
- Where are clustering used?

- ✓ “Understanding”

- Related documents for browsing
    - Genes and proteins for similar functionalities
    - Stocks with similar price fluctuation

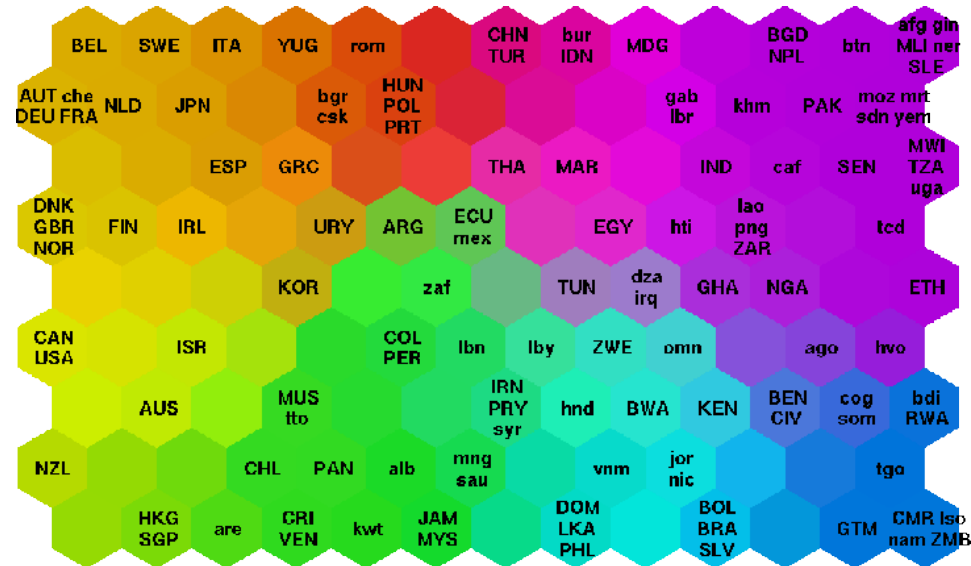
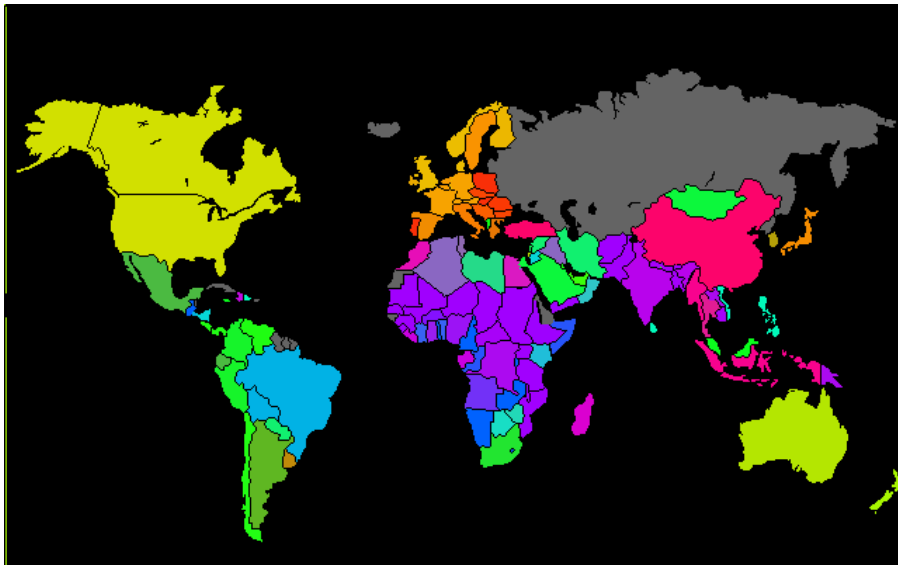
Query: israel  
Documents: 272, Clusters: 15, Average Cluster Size: 15.1 documents

Cluster	Size	Shared Phrases and <a href="#">Sample Document Titles</a>
1 <a href="#">View Results</a> <a href="#">Refine Query Based</a> <a href="#">On This Cluster</a>	16	Society and Culture (56%), Faiths and Practices (56%), Judaism (69%), Spirituality (56%); Religion (56%), organizations (43%) <ul style="list-style-type: none"> <li>● <a href="#">Ahavat Israel - The Amazing Jewish Website!</a></li> <li>● <a href="#">Israel and Judaism</a></li> <li>● <a href="#">Judaica Collection</a></li> </ul>
2 <a href="#">View Results</a> <a href="#">Refine Query Based</a> <a href="#">On This Cluster</a>	15	Ministry of Foreign Affairs (33%), Ministry (87%) <ul style="list-style-type: none"> <li>● <a href="#">Publications and Data of the BANK OF ISRAEL</a></li> <li>● <a href="#">Consulate General of Israel to the Mid-Atlantic Region</a></li> <li>● <a href="#">The Friends of Israel Gospel Ministry</a></li> </ul>
3 <a href="#">View Results</a> <a href="#">Refine Query Based</a> <a href="#">On This Cluster</a>	11	Israel Tourism (36%), Comprehensive Israel (36%), Tourism (64%) <ul style="list-style-type: none"> <li>● <a href="#">Interactive Israel tourism guide - Jerusalem</a></li> <li>● <a href="#">Ambassade d'Israel</a></li> <li>● <a href="#">Travel to Israel Opportunities</a></li> </ul>
4 <a href="#">View Results</a> <a href="#">Refine Query Based</a> <a href="#">On This Cluster</a>	7	Middle East (57%), History (57%); WAR (42%), Region (42%), Complete (42%), Listing (42%), country (42%) <ul style="list-style-type: none"> <li>● <a href="#">Israel at Fifty: Our Introduction to The Six Day War</a></li> <li>● <a href="#">Machal - Volunteers in the Israel's War of Independence</a></li> <li>● <a href="#">HISTORY: The State of Israel</a></li> </ul>
5 <a href="#">View Results</a> <a href="#">Refine Query Based</a> <a href="#">On This Cluster</a>	22	Economy (68%), Companies (55%), Travel (55%) <ul style="list-style-type: none"> <li>● <a href="#">Israel Hotel Association</a></li> <li>● <a href="#">Israel Association of Electronics Industries</a></li> <li>● <a href="#">Focus Capital Group - Israel</a></li> </ul>



# Clustering: Overview

- Where are clustering used?
  - ✓ “Summarization”
    - Reduce the size of large data sets
  - ✓ Closely linked to “Visualization”



# Clustering: Overview

- Where are clustering used?

- ✓ “Strategy Planning”

- Asset management based on stock clustering
    - Stocks are clustered based on their 6 month profit and volatility
    - Select stocks from “maximum performance” and “minimum volatility group” for portfolio management



<https://quantdare.com/hierarchical-clustering/>

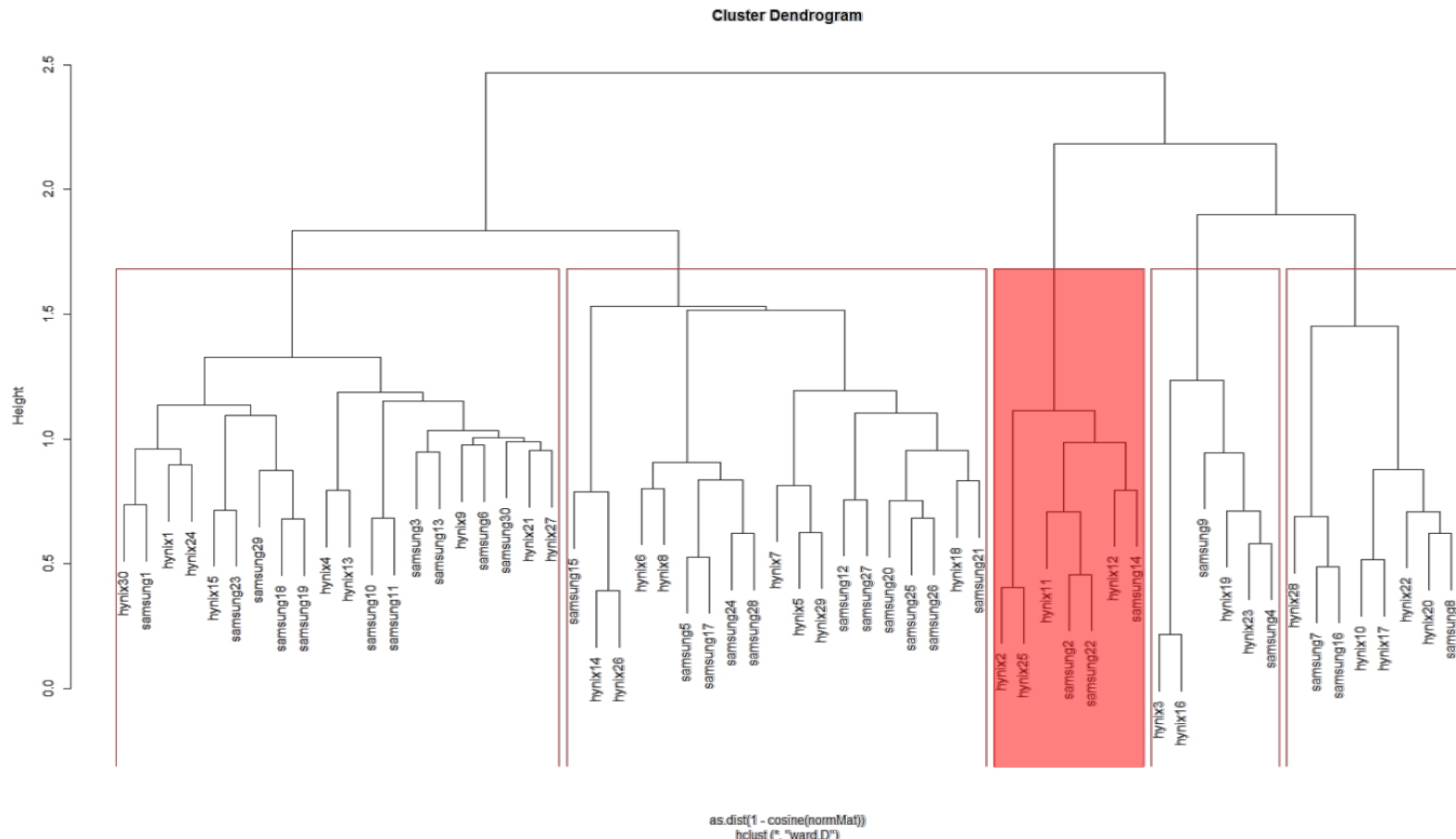


# Clustering: Overview

- Where are clustering used?

- ✓ “Strategy Planning”

- Patent analysis to understand pros and cons compared with the rival company



# Clustering: Overview

- Where are clustering used?

- ✓ “Strategy Planning”

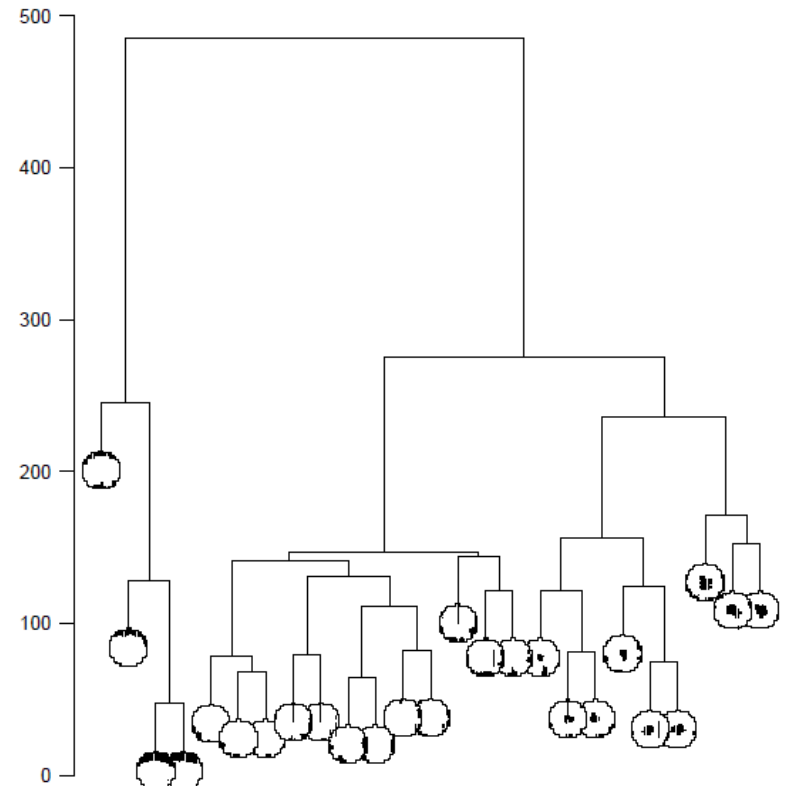
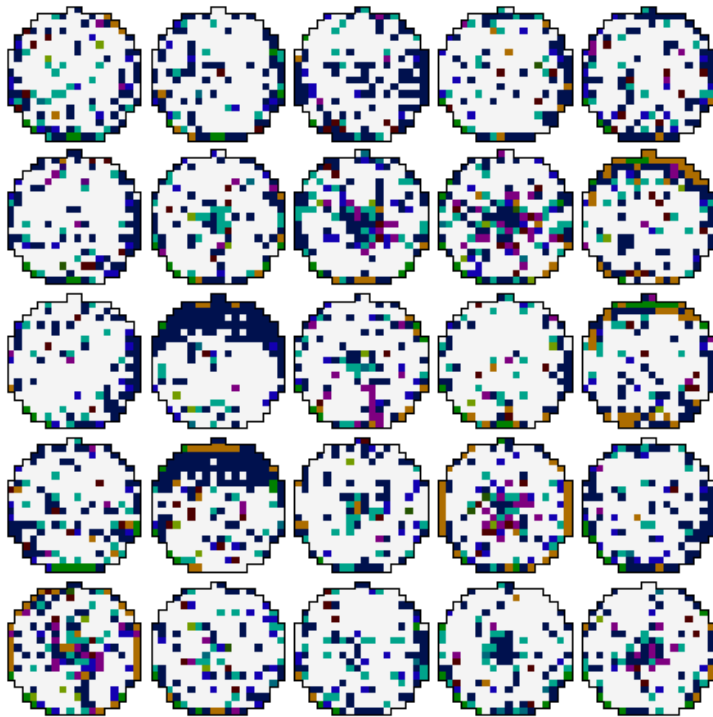
- Patent analysis to understand pros and cons compared with the rival company

1	회사	일련번호	특허명	조록
2	SK하이닉스	2	멀티 레글레이터 회로 및 이를 구비한 집적회로	본 기술에 따른 레글레이터 회로는, 입력전압을 일정한 전압 레벨로 레글레이팅하여 출력하도록 구성된 레글레이터 및 복수개의 전압 생성 코드 들에 의해 결정되는 내부 저항값들에 따라 상기 레글레이터의 출력 전압을 분배한 분배전압들을 각각 출력하도록 구성된 복수개의 전압 분배회로를 포함한다.
3	SK하이닉스	11	내부 전압 생성 회로 및 그의 동작 방법	펄핑 동작을 통해 내부 전압을 생성하는 내부 전압 생성 회로에 관한 것으로, 다수의 펄핑부를 포함하며, 목표 전압 레벨에 대응하는 최종 펄핑 전압을 생성하기 위한 펄핑 전압 생성부, 및 상기 목표 전압 레벨에 대응하여 상기 다수의 펄핑부의 활성화 개수를 제어하기 위한 활성화 제어부를 구비하는 내부 전압 생성 회로가 제공된다.
4	SK하이닉스	12	자기 메모리 장치를 위한 라이트 드라이버 회로 및 자기 메모리 장치	비트라인과 소스라인 간에 접속되며, 비트라인 방향으로 인접하는 한 쌍의 자기 메모리 셀이 소스라인을 공유하는 복수의 자기 메모리 셀로 이루어진 메모리 셀 어레이를 포함하는 자기 메모리 장치를 위한 라이트 드라이버 회로로서, 정의 기록전압 공급단자와 부의 기록전압 공급단자 간에 접속되어, 라이트 인에이블 신호 및 데이터 신호에 따라 정의 기록전압 또는 부의 기록전압에 의한 전류를 비트라인에 선택적으로 공급하는 스위칭부를 포함하는 자기 메모리 장치를 제공한다.
5	SK하이닉스	25	전압 레글레이터 및 전압 레글레이팅 방법	전압 레글레이터는 출력전압을 전압 출력단으로 출력하는 전압 출력부와, 제1 제어코드의 제어에 따라 분배 저항값을 조절하는 제1 저항분배 스테이지와, 제1 저항분배 스테이지에서 결정된 분배 저항값을 제2 제어코드의 제어에 따라 조절하는 제2 저항분배 스테이지를 포함하며, 전압 출력단을 통해서 출력되는 출력전압의 전압레벨은 제1 및 제2 저항분배 스테이지를 통해서 결정된 상기 분배 저항값과, 기준저항의 저항값 비율에 따라 조절되는 것을 특징으로 한다.
6	삼성전자	2	전압 공급 장치 및 그것을 포함한 불휘발성 메모리 장치	본 발명에 따른 전압 공급 장치는 전원 전압을 승압하고, 상기 승압된 전압을 출력 라인으로 제공하기 위한 전하 펌프 및 상기 출력 라인의 전압 레벨을 목표 전압 레벨로 유지하기 위한 전압 제어 회로를 포함한다. 본 발명에 따른 상기 전압 제어 회로는 웰 상에 형성된 제 1 영역 및 제 2 영역을 포함하고, 상기 제 1 영역 및 제 2 영역 사이의 리치 스루(reach through)를 이용하여 상기 출력 라인의 전압 레벨을 제어하기 위한 리치 스루 소자를 포함한다.
7	삼성전자	14	파워 공급 회로 및 이를 구비하는 상 변화 메모리 장치	파워 공급 회로 및 이를 구비하는 상 변화 메모리 장치가 개시된다. 본 발명의 제 1 실시예에 따른 반도체 메모리 장치는 파워 공급 회로, 스위치들 및 선택기들을 구비한다. 파워 공급 회로는 상기 블록들의 메모리 셀들에 사용되는 제 1 전압 및 제 2 전압을 생성한다. 스위치들은 상기 파워 공급 회로와 상기 제 1 전압이 전달되는 제 1 라인 및 상기 제 2 전압이 전달되는 제 2 라인으로 연결되고, 제어 신호에 응답하여 상기 제 1 전압 및 제 2 전압 중 하나를 대응되는 블록으로 인가한다. 선택기들은 블록 선택 신호 및 디스차이지 성공 신호에 응답하여, 상기 제어 신호를 생성한다. 본 발명에 따른 파워 공급 회로 및 이를 구비하는 상 변화 메모리 장치는 셀 블록마다 별도의 파워 스위치를 구비함으로써 파워 공급 회로의 동작 시간 및 동작 전류를 감소시킬 수 있다. 또한, 기입 전압을 디스차이지한 후 다른 레벨의 전압을 공급함으로써, 상 변화 메모리 장치의 오작동이 방지될 수 있다.
8	삼성전자	22	전압 안정화 장치 및 그것을 포함하는 반도체 장치 및 전압 생성 방법	본 발명은 전압 안정화 장치 및 그것을 이용하는 반도체 장치에 관한 것이다. 본 발명의 기술적 사상의 실시예에 따른 전압 안정화 장치는 제 1 전압을 생성하는 제 1 레글레이터 및 상기 제 1 전압보다 낮은 제 2 전압을 생성하는 제 2 레글레이터를 포함하고, 상기 제 2 레글레이터는 상기 제 1 전압의 레벨과 미리 정해진 기준 전압의 레벨의 비교 결과에 기초하여 상기 제 1 전압 또는 상기 제 1 전압보다 높은 제 2 전압을 선택적으로 이용하여 상기 제 2 전압을 생성한다. 본 발명의 기술적 사상의 실시예에 따르면 제 1의 전압 > 제 2의 전압의 관계를 유지하면서, 동시에 제 2의 전압을 고속으로 전위 변환시킬 수 있다.

# Clustering: Overview

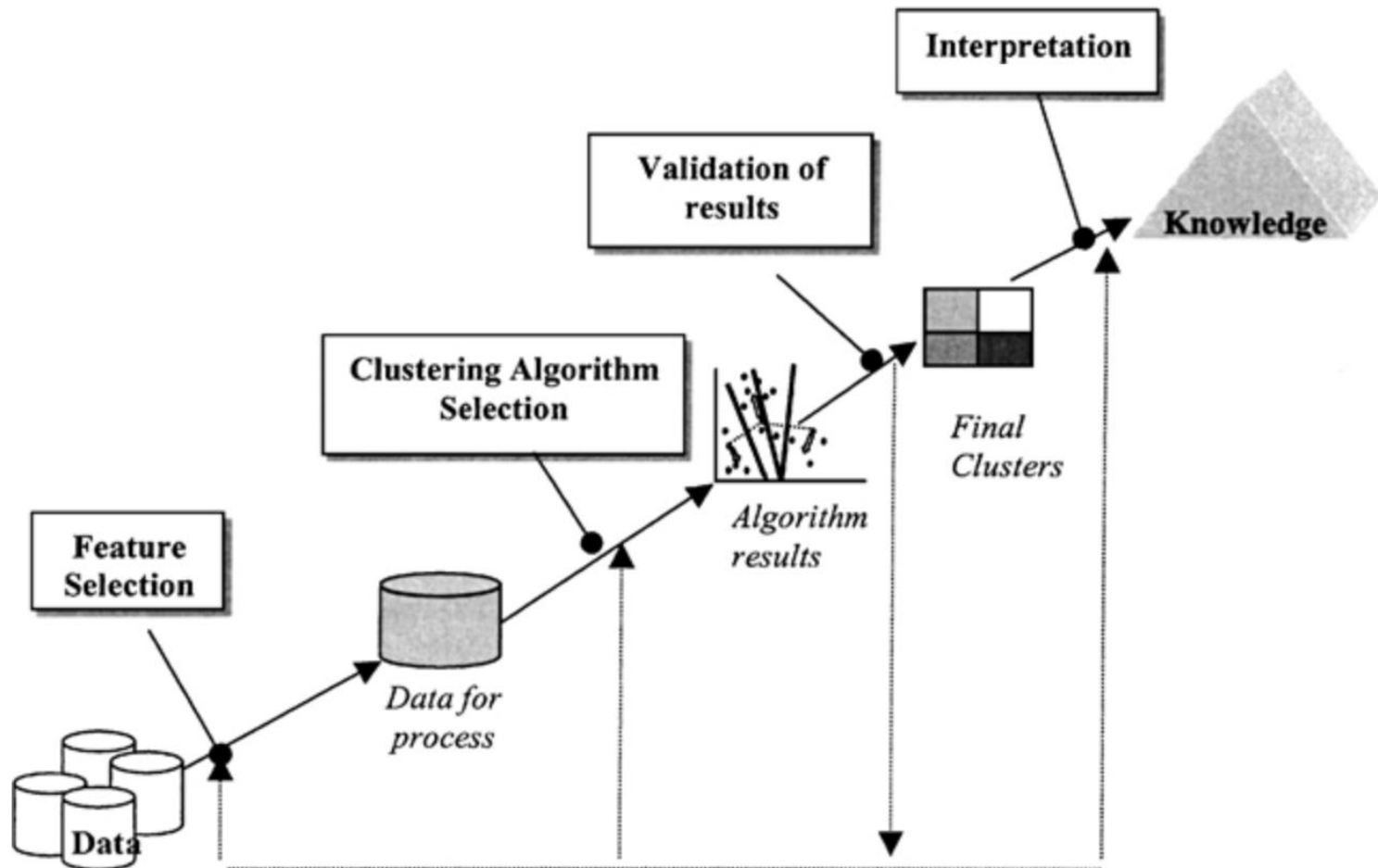
- Where are clustering used?
  - ✓ In-depth analysis

Sample lot exhibiting spatial patterning



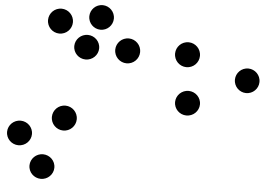
# Clustering: Overview

- Standard clustering procedure

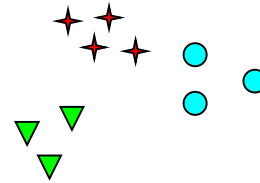
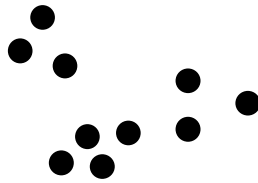


# Clustering: Issues

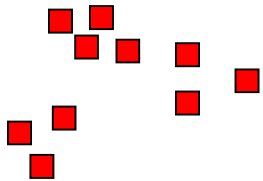
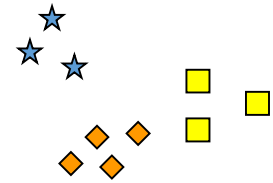
- How many clusters are optimal?



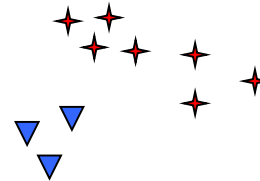
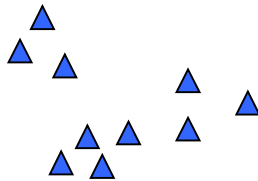
How many clusters?



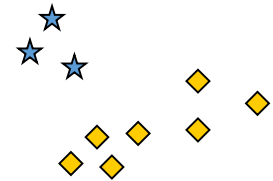
Six Clusters



Two Clusters

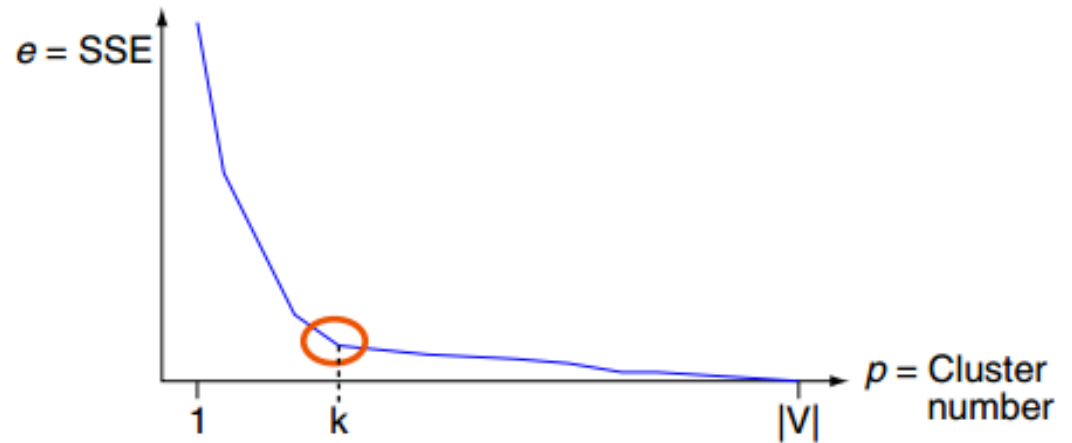
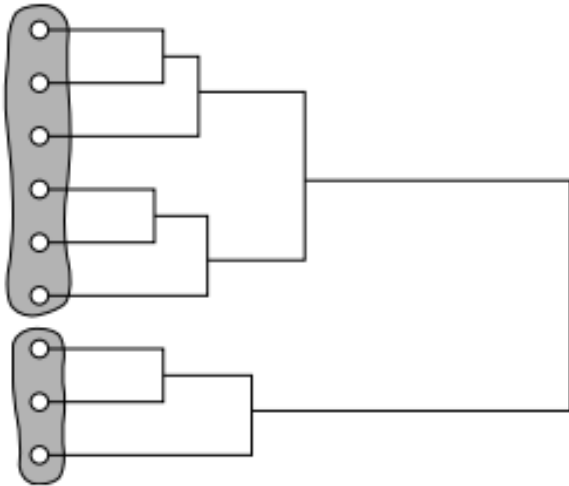


Four Clusters



# Clustering: Issues

- How many clusters are optimal?
  - ✓ Use a clustering validity measure to evaluate the clustering result
  - ✓ Find the elbow point






# Clustering: Issues

- How to evaluate the clustering result?
  - ✓ There is no globally accepted validity measure
  - ✓ Because clustering is an unsupervised learning task, we do not know the exact answer
- Three categories for clustering validity measures
  - ✓ External: Compare the clustering structure with the known answer (**unrealistic**)
  - ✓ Internal: Focusing on the **compactness** of clusters
  - ✓ Relative: Focusing on both the **compactness** of clusters and **separation** between clusters

# Clustering: Issues

- Examples of clustering validity measures

External	Internal	Relative
		
<input type="checkbox"/> Rand Statistic	<input type="checkbox"/> Cophenetic Correlation Coefficient	<input type="checkbox"/> Dunn family of indices
<input type="checkbox"/> Jaccard Coefficient	<input type="checkbox"/> Sum of Squared error (SSE)	<input type="checkbox"/> Davies-Bouldin (DB) index
<input type="checkbox"/> Folks and Mallows index	<input type="checkbox"/> Cohesion and separation	<input type="checkbox"/> Semi-partial R-squared
<input type="checkbox"/> (Normalized) Hurbert $\Gamma$ statistic		<input type="checkbox"/> SD validity index
		<input type="checkbox"/> Silhouette



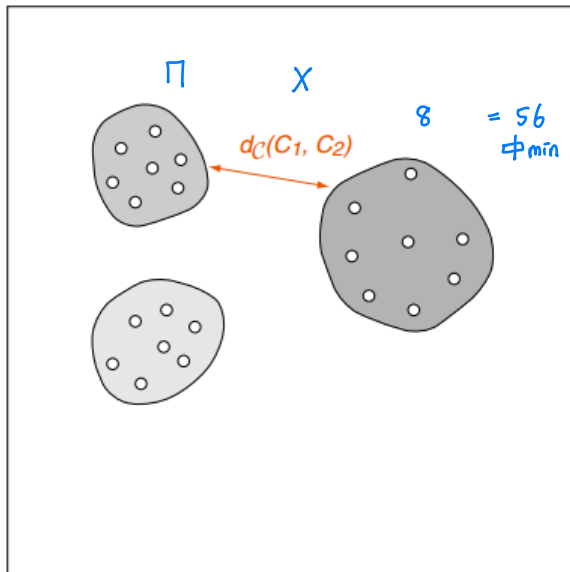
# Clustering: Issues

- Clustering Validity Measure Example: Dunn Index

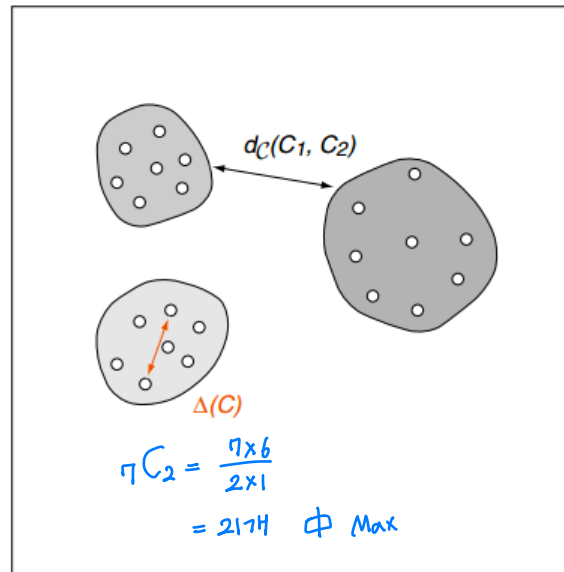
✓ If the clustering is well performed,

- The value of (1) will be large and the values of (2) and (3) will be small

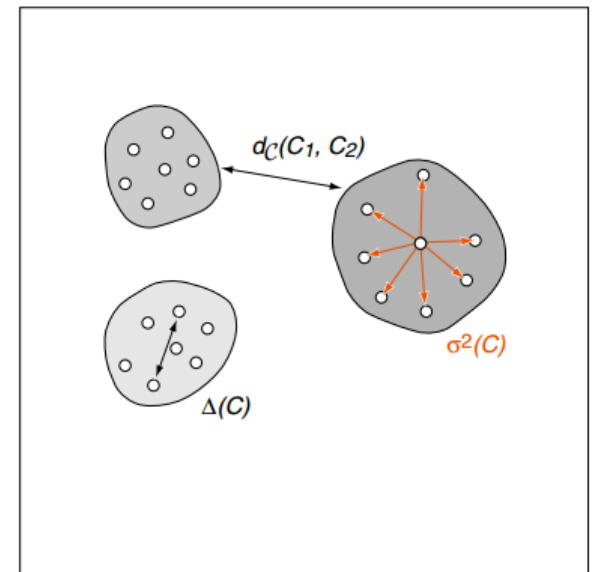
(1) Distance between two clusters



(2) Diameter of a cluster



(3) Scatter within a cluster (SSE)



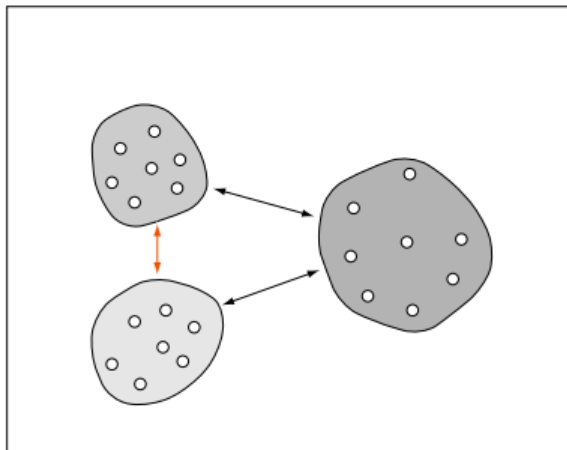
# Clustering: Issues

- Clustering Validity Measure Example: Dunn Index

멀면 멀수록 좋음

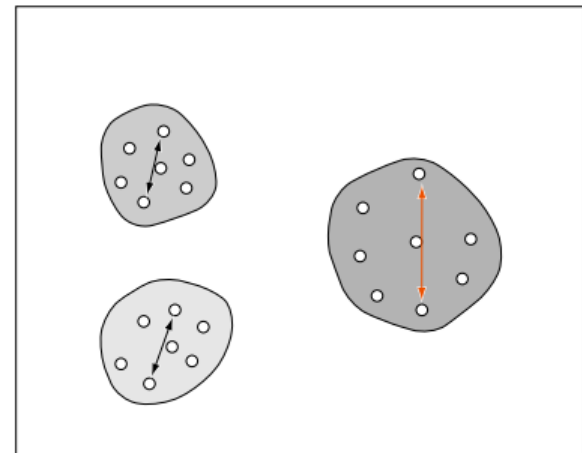
- ✓ Dunn index is defined the ratio of (1) the minimum distance between two clusters to (2) the maximum diameter of the clusters

작으면 작을수록 좋음



$$I(C) = \frac{\min_{i \neq j} \{d_c(C_i, C_j)\}}{\max_{1 \leq l \leq k} \{\Delta(C_l)\}},$$

$I(C) \rightarrow \max$



$$I(C) = \frac{\min_{i \neq j} \{d_c(C_i, C_j)\}}{\max_{1 \leq l \leq k} \{\Delta(C_l)\}},$$

$I(C) \rightarrow \max$

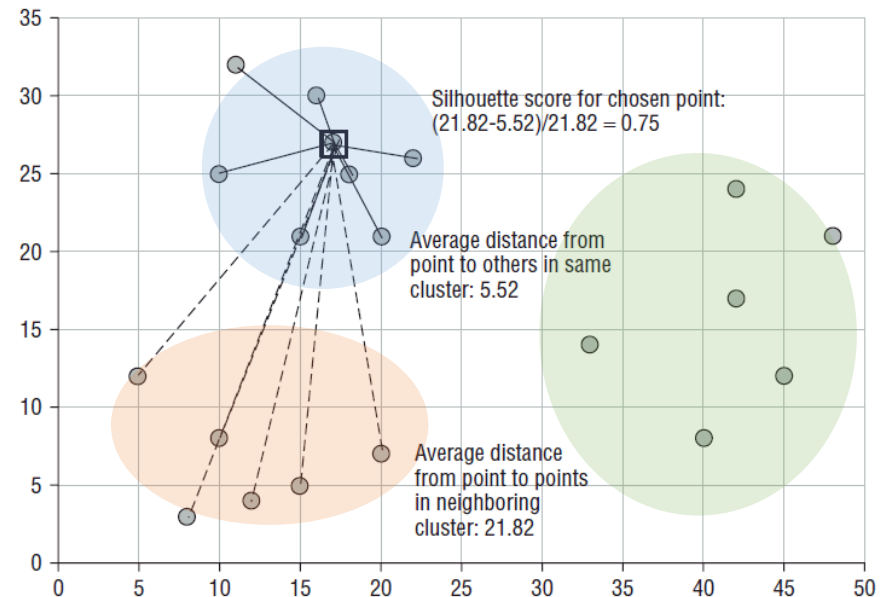
# Clustering: Issues

- Clustering Validity Measure Example: Silhouette

- ✓  $a(i)$ : the average distance between an instance  $i$  and the other instances in the same cluster
- ✓  $b(i)$ : the minimum of the average distances between an instance  $i$  and the instances in a cluster to which the instance  $i$  does not belong

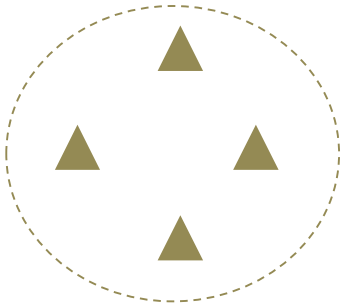
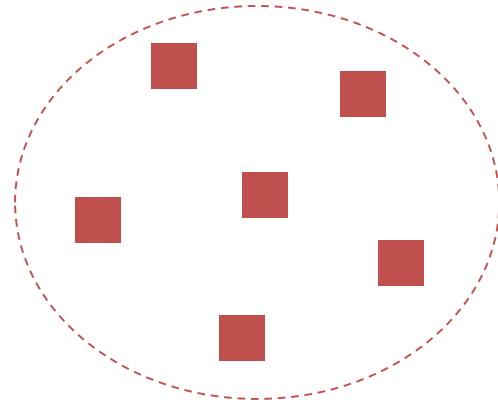
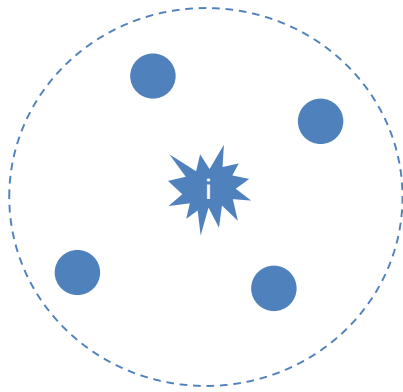
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$



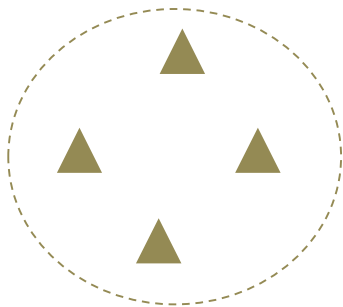
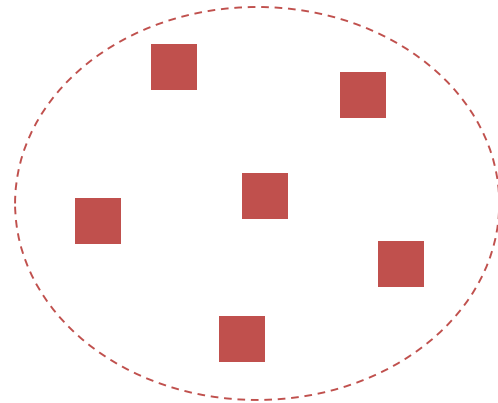
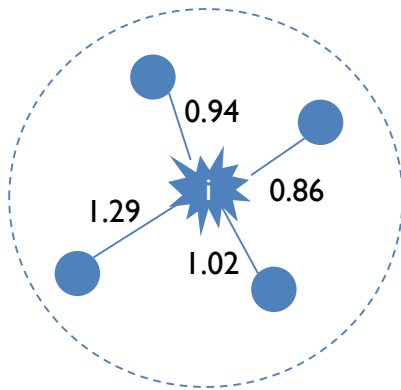
# Clustering: Issues

- Clustering Validity Measure Example: Silhouette



# Clustering: Issues

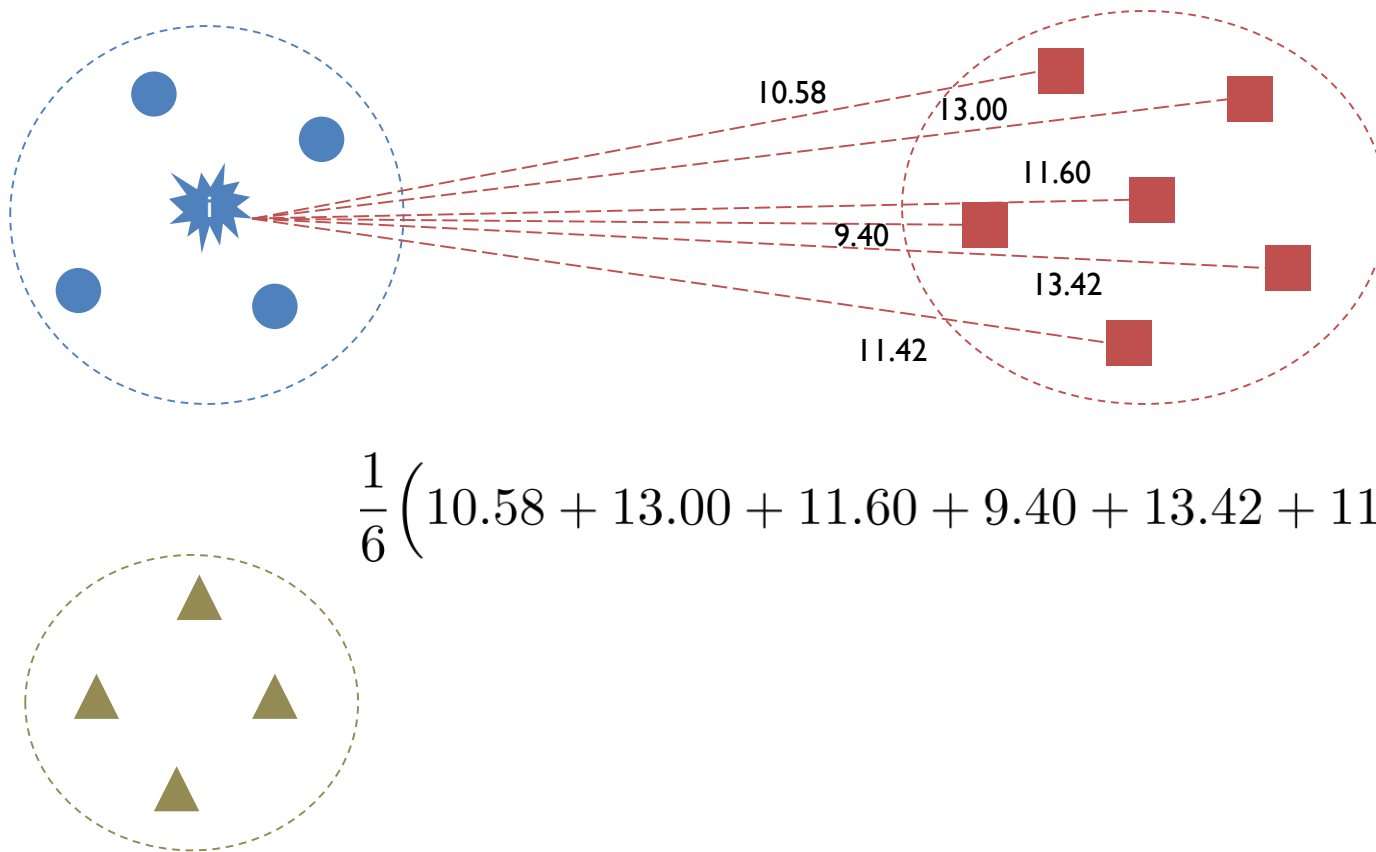
- Clustering Validity Measure Example: Silhouette



$$a(i) = \frac{1}{4} (0.94 + 0.86 + 1.02 + 1.29) = 1.03$$

# Clustering: Issues

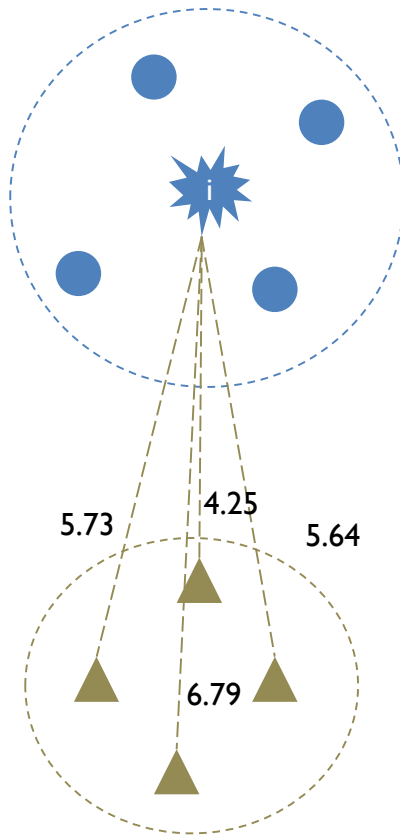
- Clustering Validity Measure Example: Silhouette



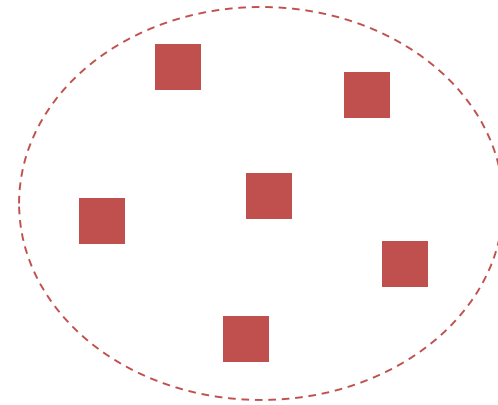
$$\frac{1}{6} \left( 10.58 + 13.00 + 11.60 + 9.40 + 13.42 + 11.42 \right) = 11.57$$

# Clustering: Issues

- Clustering Validity Measure Example: Silhouette



$$\frac{1}{4} (5.73 + 6.79 + 4.25 + 5.64) = 5.60$$



$$b(i) = \min(11.57, 5.60) = 5.60$$

$$s(i) = \frac{5.60 - 1.03}{\max(1.03, 5.60)}$$

$$= \frac{4.57}{5.60} = 0.82 \quad \text{크면 클수록 좋음}$$

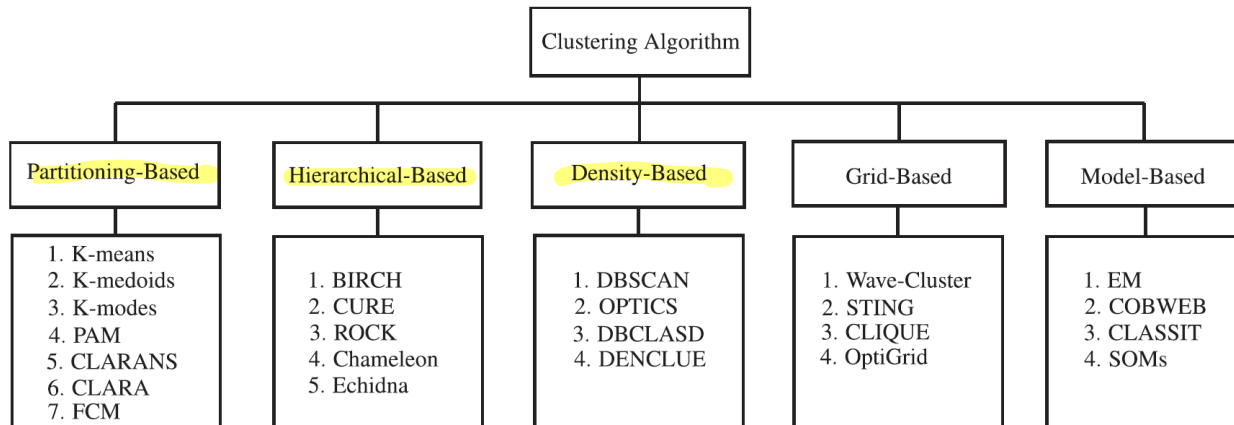
# Clustering: Types

하나의 객체가 두개  
이상의 군집에 관계  
하는 것 허용 X  
= 허용 O

- Hard clustering vs. Soft clustering

- ✓ Hard Clustering (Crisp Clustering)

- Results in non-overlapping clusters
- Each instance belongs to only one cluster



- ✓ Soft Clustering (Fuzzy Clustering)

- Possible to result in overlapping clusters
- Each instance can belong to more than two clusters



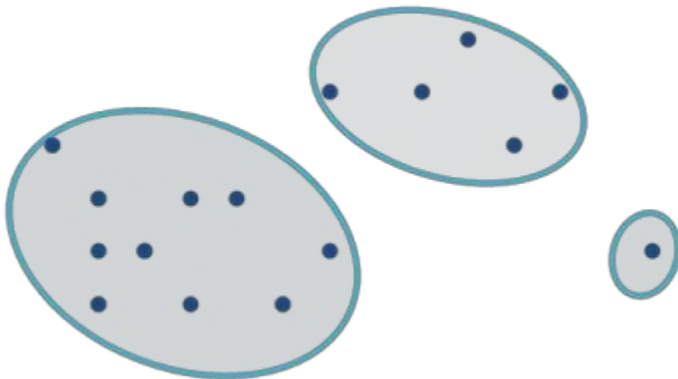
# Clustering: Algorithms

한번에 구분 done

- Partitional clustering

- ✓ Divide data into non-overlapping subsets such that each data object is in exactly one subset

Partitional Clustering



무엇을 무엇을 처리

- Hierarchical clustering

- ✓ A set of nested clusters organized as a hierarchical tree

Hierarchical Clustering

