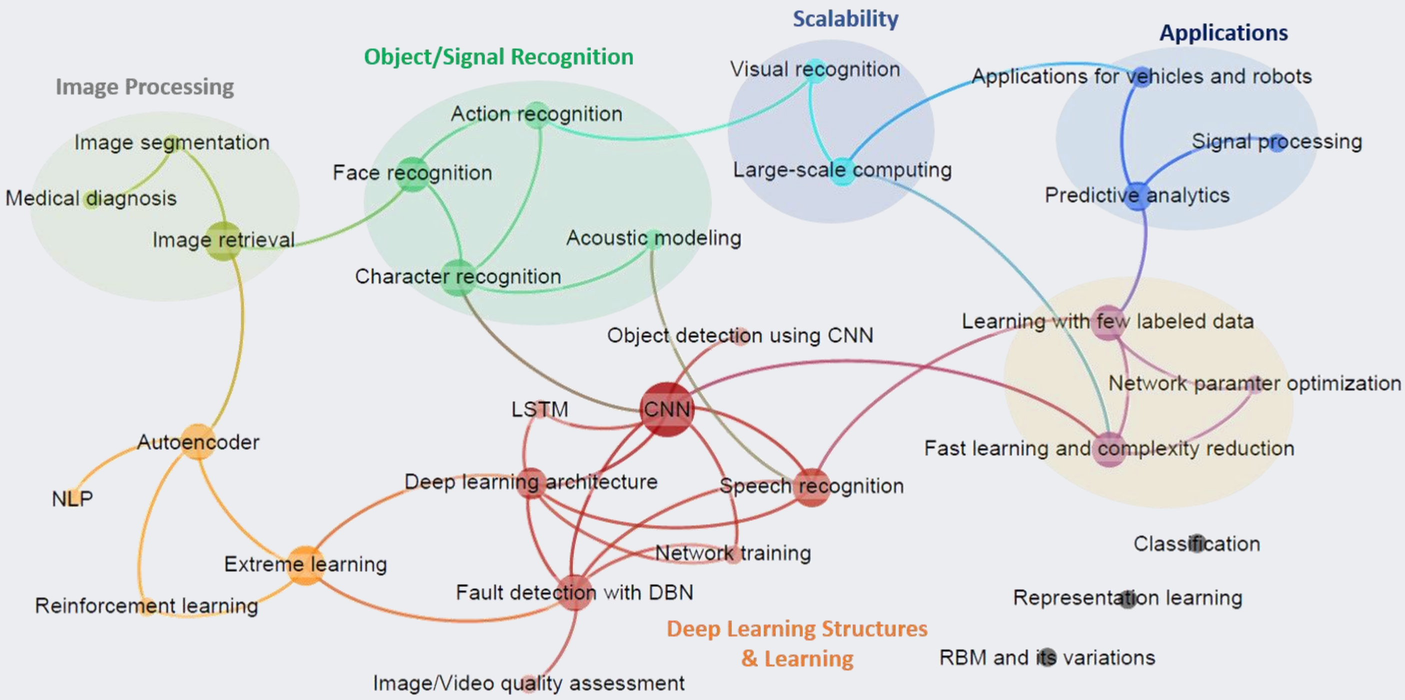


Lecture 4: Clustering

Pilsung Kang

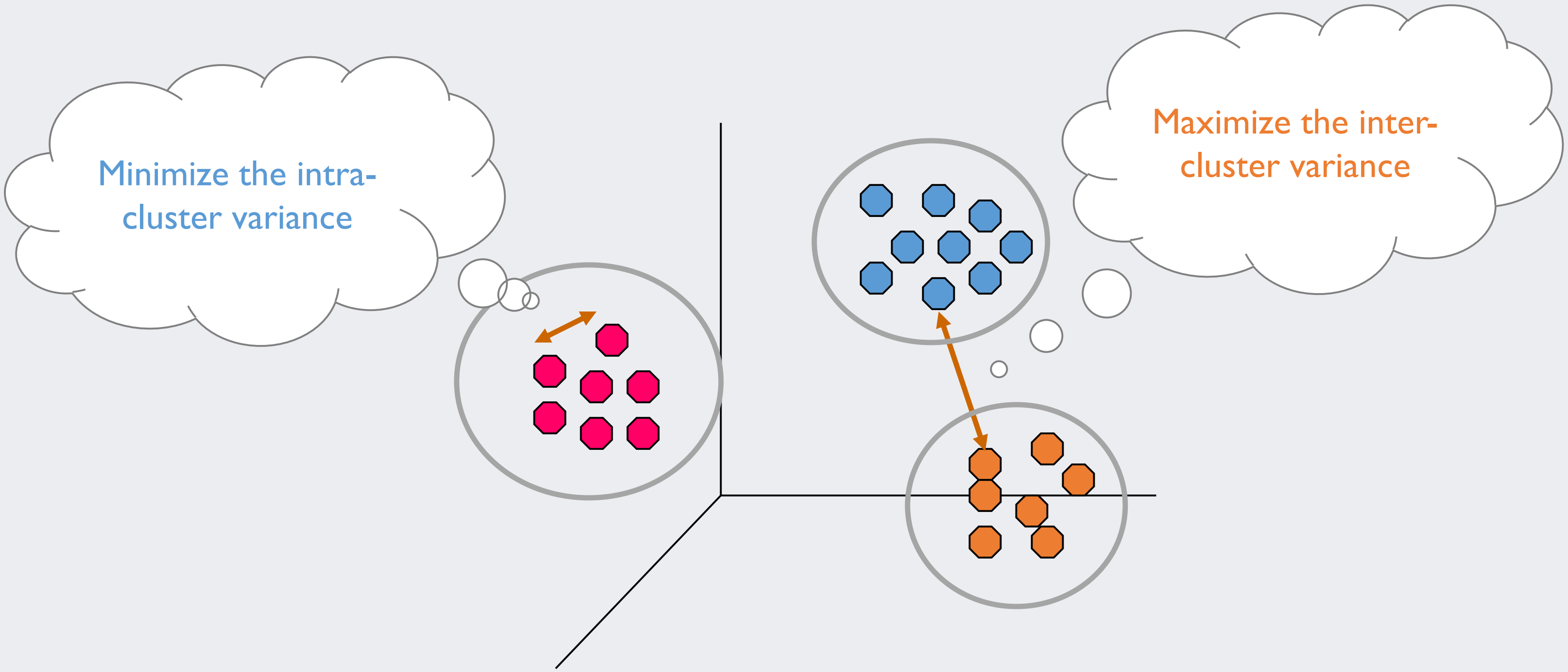
School of Industrial Management Engineering

Korea University

AGENDA

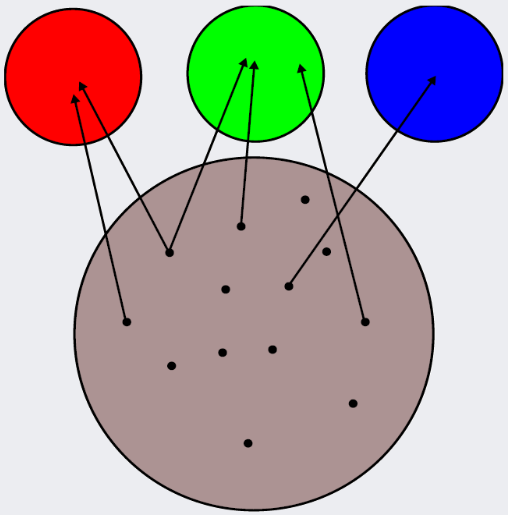
Clustering: Overview

- 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

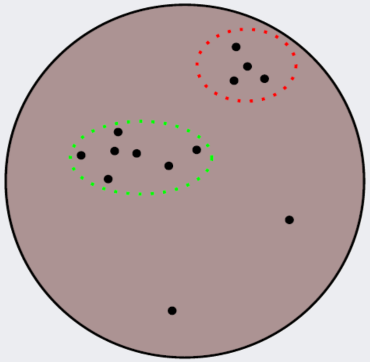


Clustering: Overview

- 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



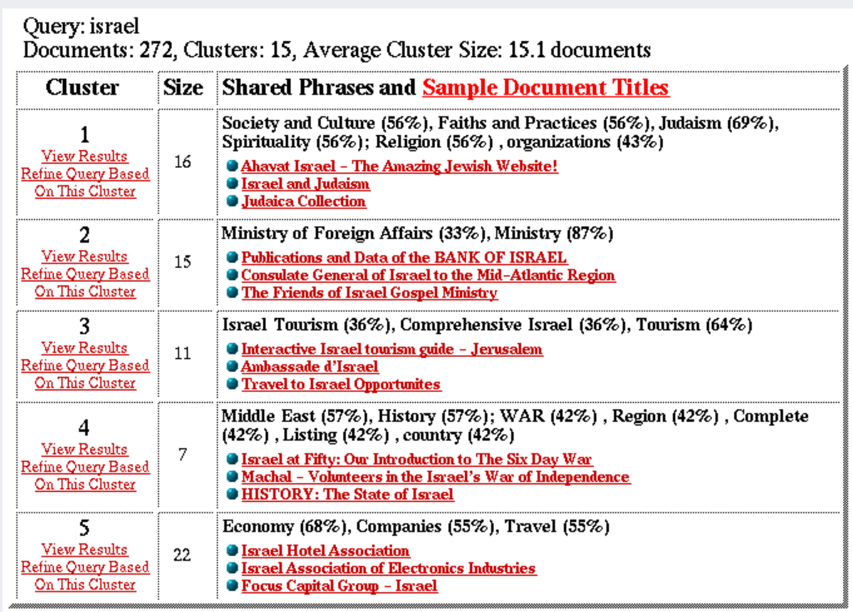
(a) Classification



(b) Clustering

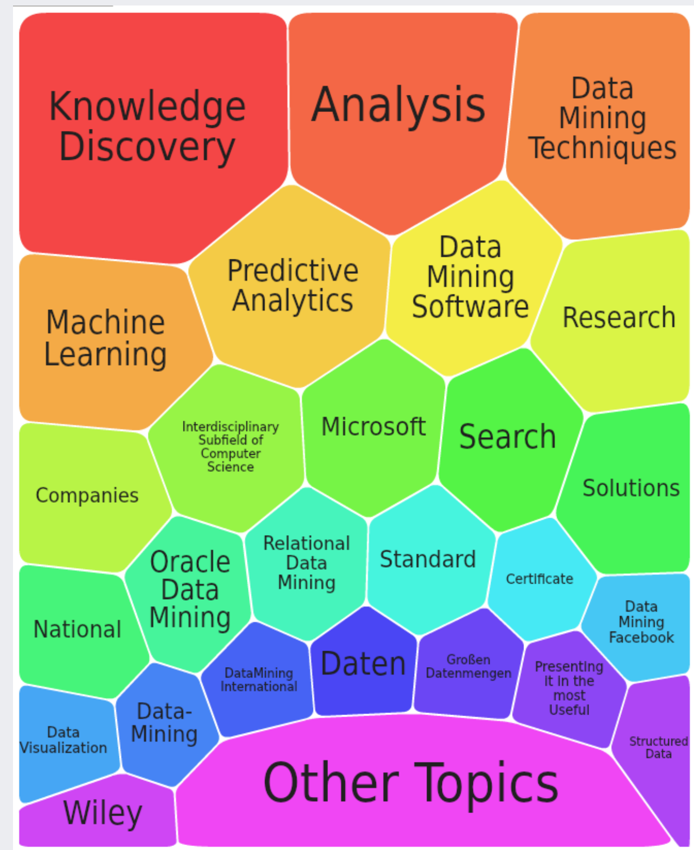
Clustering: Overview

- Where are clustering used?
  - ✓ "Understanding"
    - Related documents for browsing
    - Genes and proteins for similar functionalities
    - Stocks with similar price fluctuation



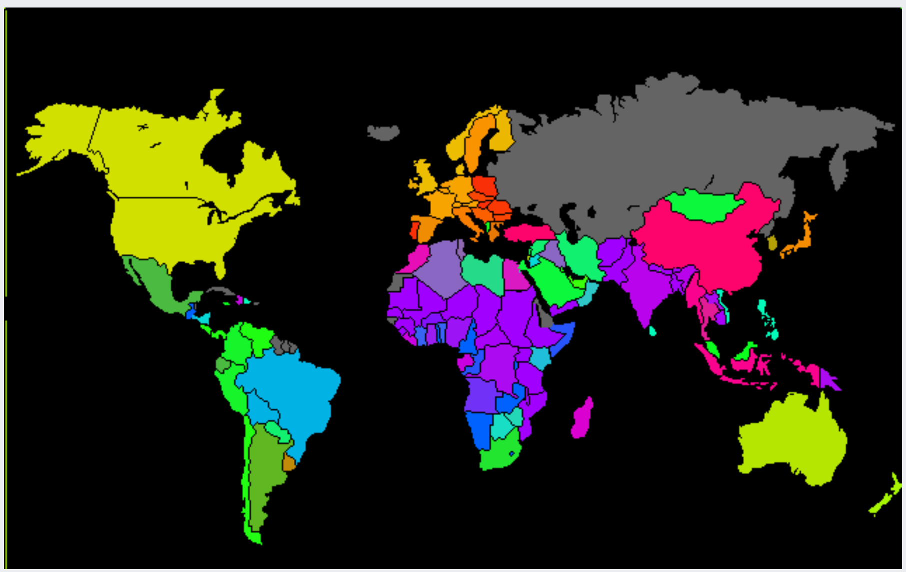
Query: israel
Documents: 272, Clusters: 15, Average Cluster Size: 15.1 documents

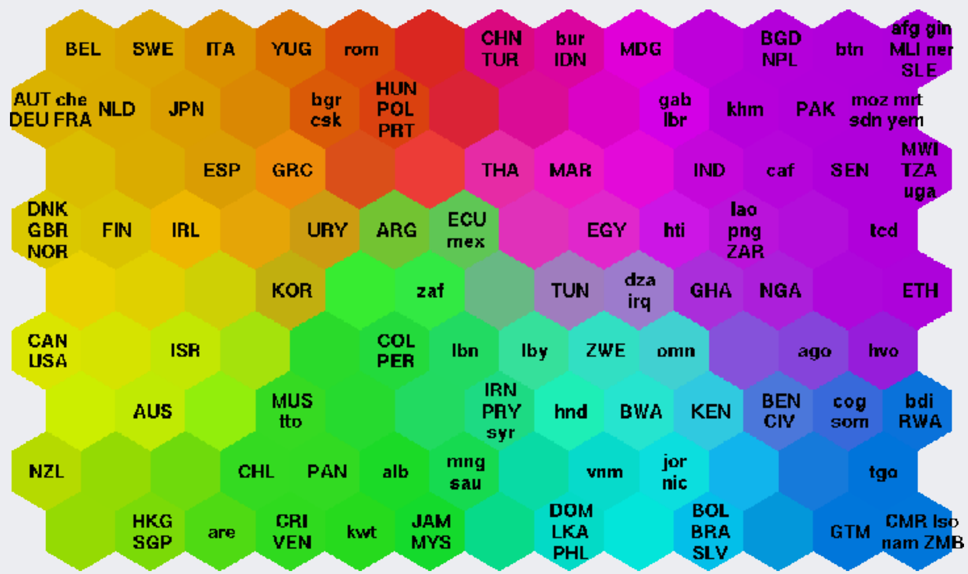
Cluster	Size	Shared Phrases and Sample Document Titles
1 View Results Refine Query Based On This Cluster	16	Society and Culture (56%), Faiths and Practices (56%), Judaism (69%), Spirituality (56%); Religion (56%) , organizations (43%) Ahavat Israel - The Amazing Jewish Website! Israel and Judaism Judaica Collection
2 View Results Refine Query Based On This Cluster	15	Ministry of Foreign Affairs (33%), Ministry (87%) Publications and Data of the BANK OF ISRAEL Consulate General of Israel to the Mid-Atlantic Region The Friends of Israel Gospel Ministry
3 View Results Refine Query Based On This Cluster	11	Israel Tourism (36%), Comprehensive Israel (36%), Tourism (64%) Interactive Israel tourism guide - Jerusalem Ambassade d'Israel Travel to Israel Opportunites
4 View Results Refine Query Based On This Cluster	7	Middle East (57%), History (57%); WAR (42%) , Region (42%) , Complete (42%) , Listing (42%) , country (42%) Israel at Fifty: Our Introduction to The Six Day War Machal - Volunteers in the Israel's War of Independence HISTORY: The State of Israel
5 View Results Refine Query Based On This Cluster	22	Economy (68%), Companies (55%), Travel (55%) Israel Hotel Association Israel Association of Electronics Industries Focus Capital Group - Israel



Clustering: Overview

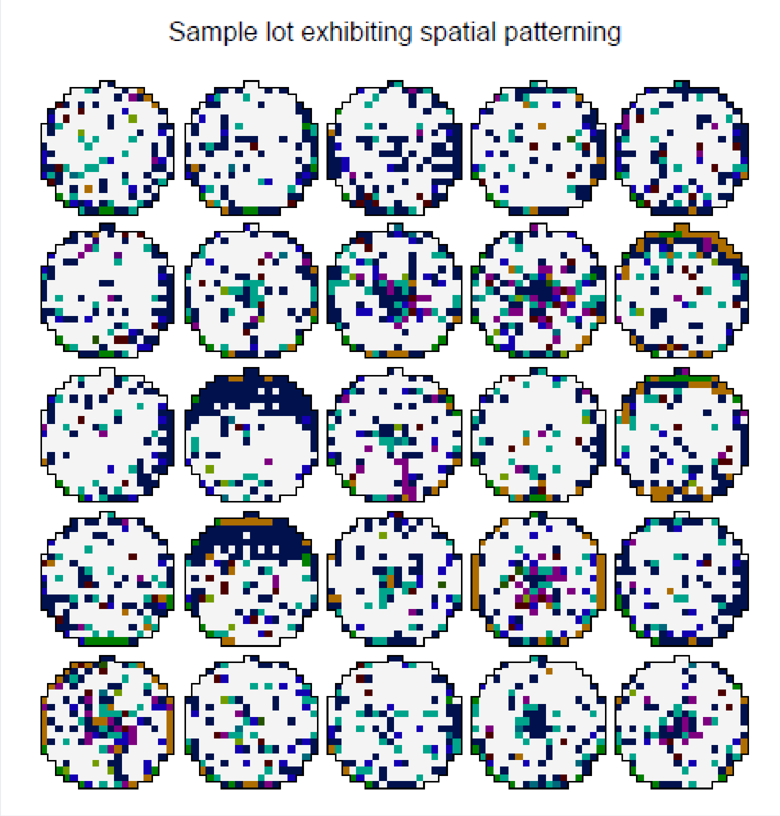
- Where are clustering used?
  - ✓ "Summarization"
    - Reduce the size of large data sets
  - ✓ Closely linked to "Visualization"

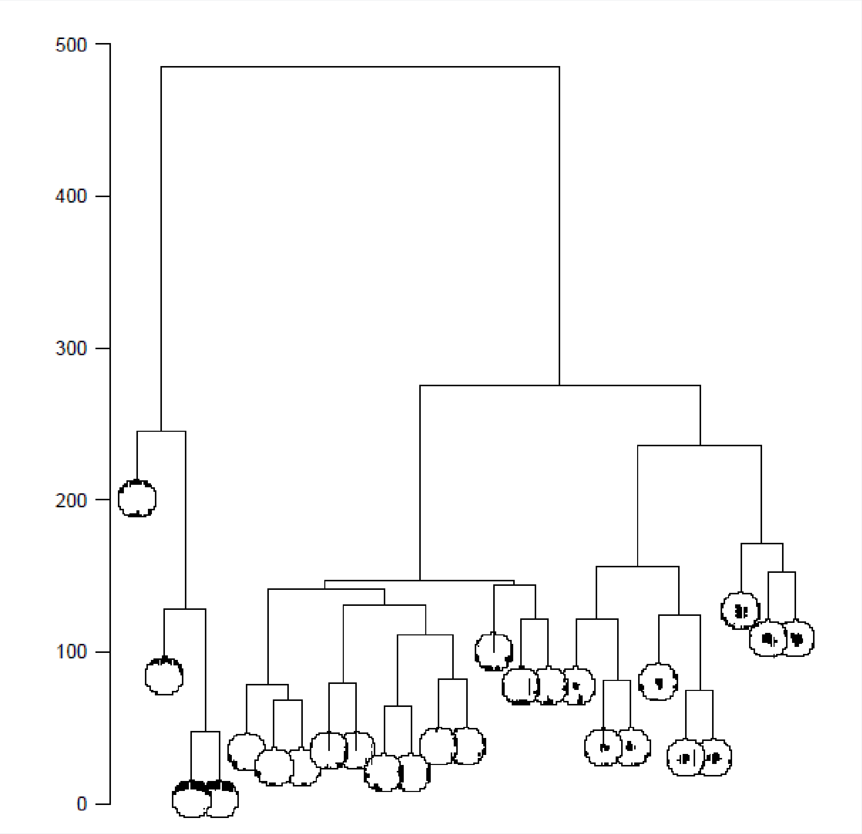




Clustering: Overview

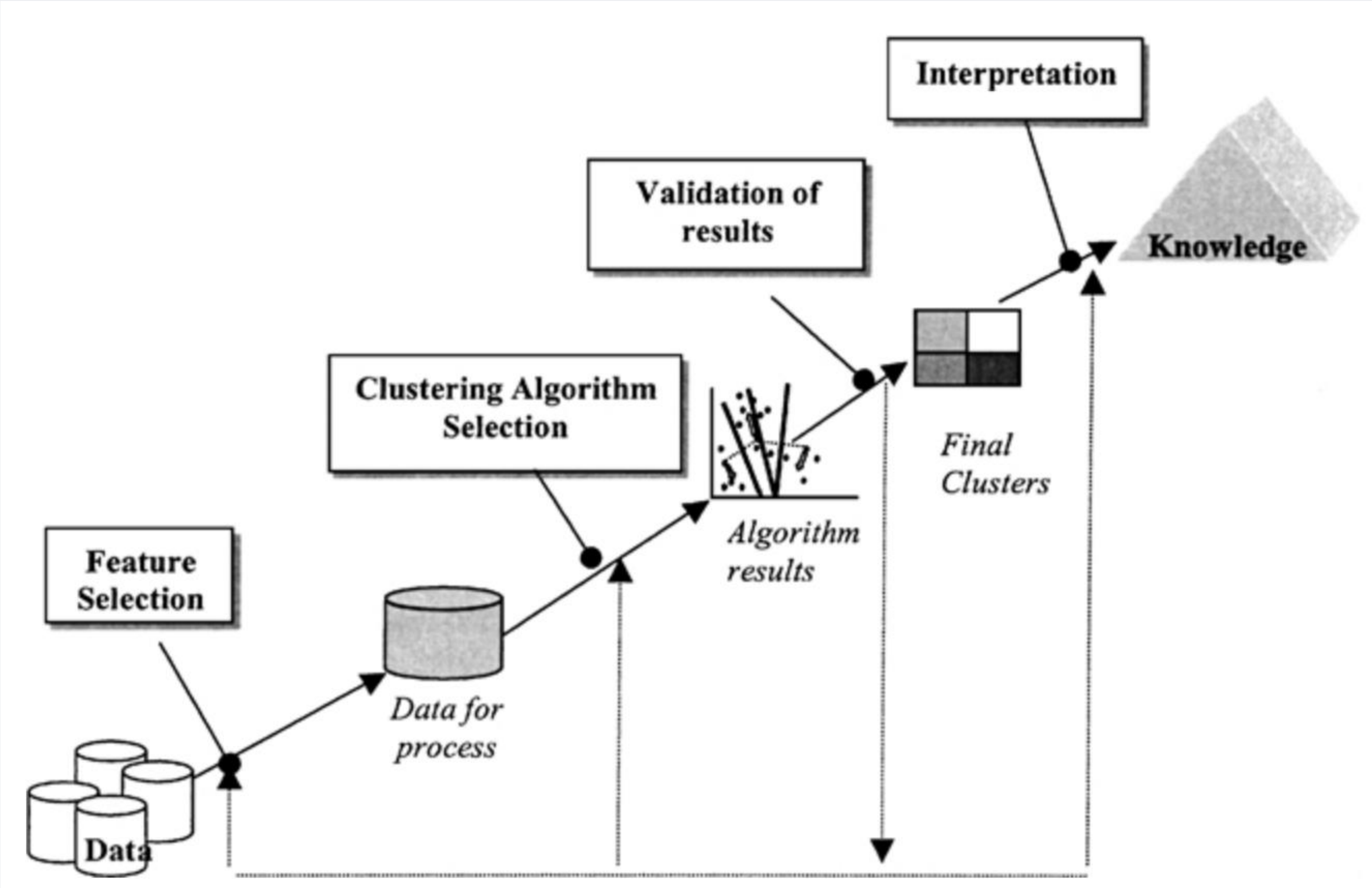
- Where are clustering used?
  - ✓ In-depth analysis





Clustering: Overview

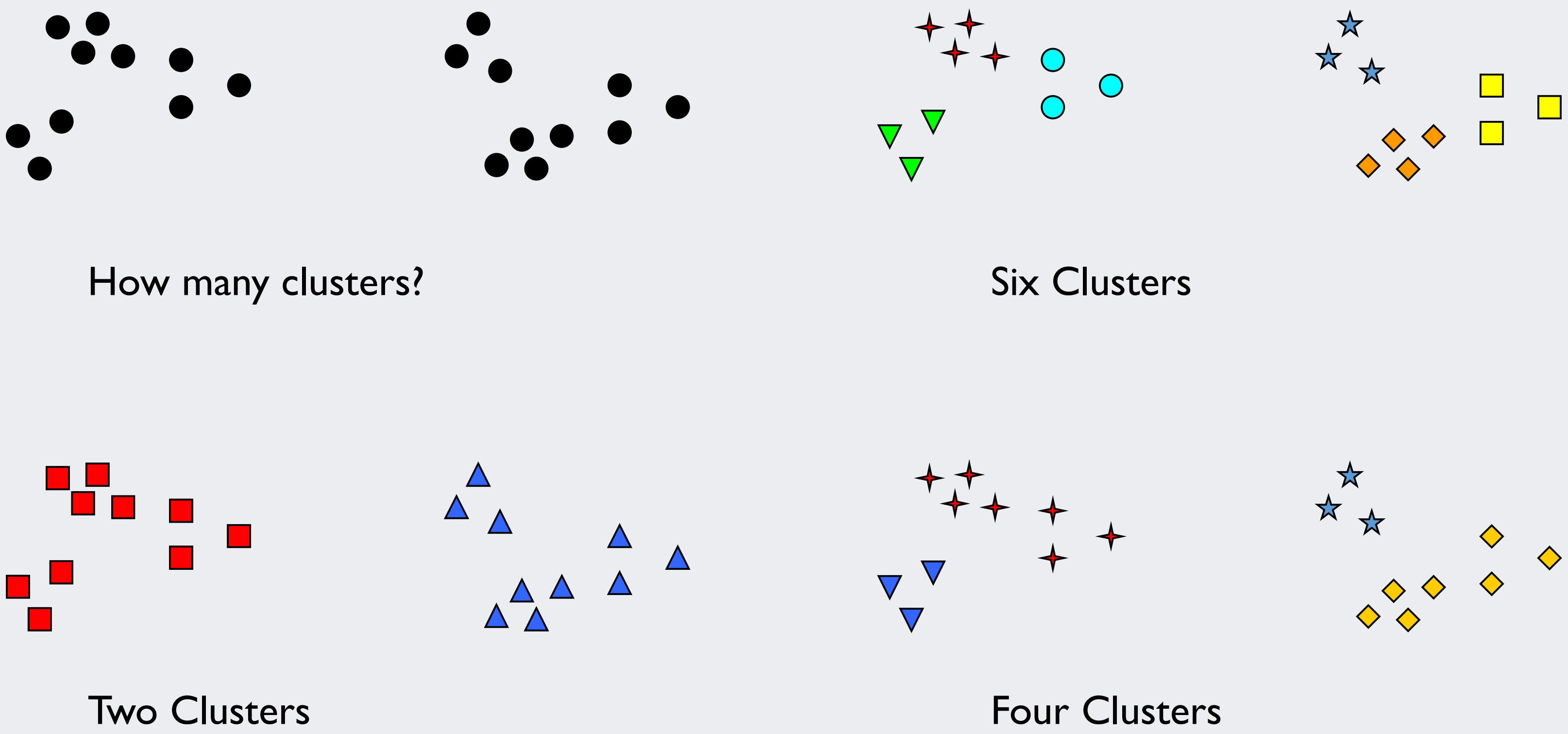
- Standard clustering procedure



http://web.itu.edu.tr/sgunduz/courses/verimaden/paper/validity_survey.pdf

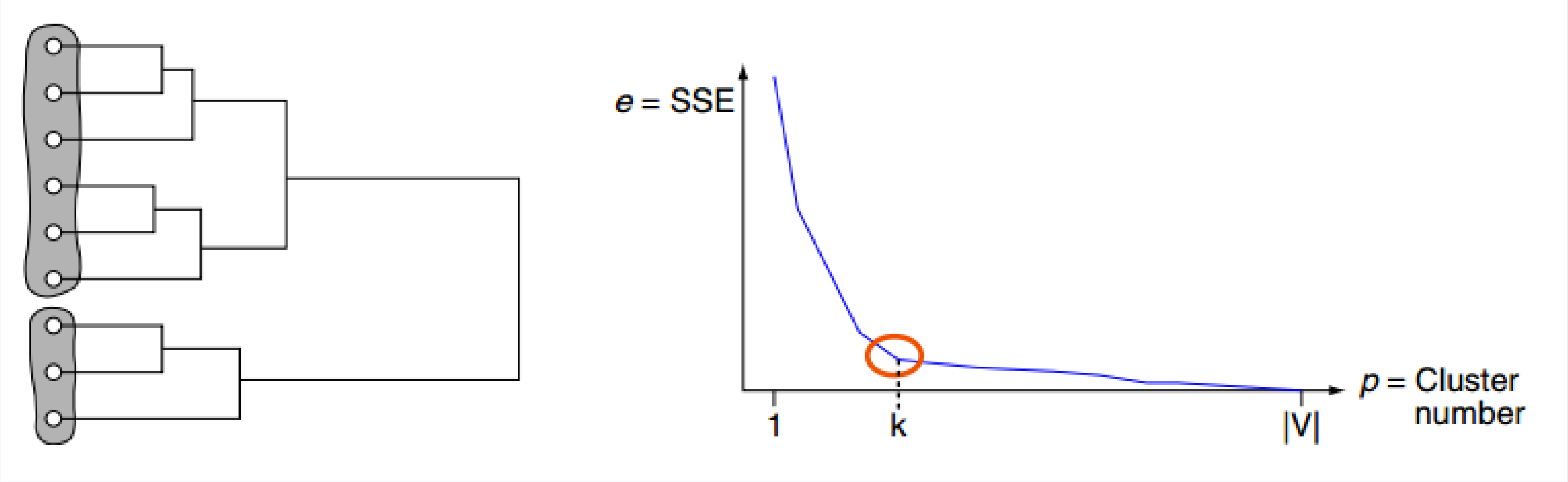
Clustering: Issues

- How many clusters are optimal?



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

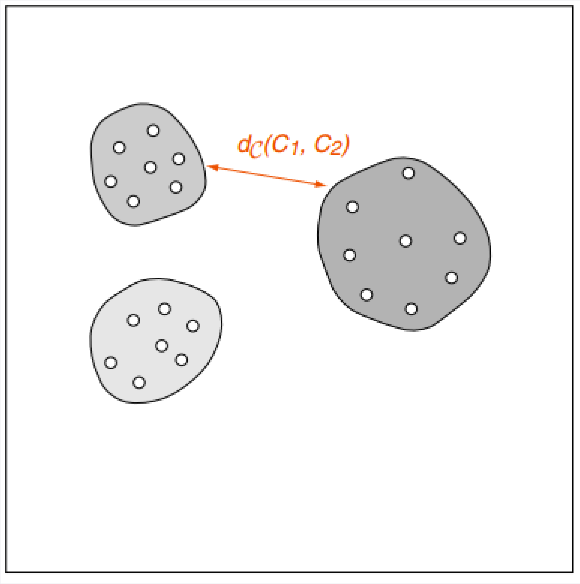
- Examples of clustering validity measures

External	Internal	Relative
Rand Statistic	Cophenetic Correlation Coefficient	Dunn family of indices
Jaccard Coefficient	Sum of Squared error (SSE)	Davies-Bouldin (DB) index
Folks and Mallows index	Cohesion and separation	Semi-partial R-squared
(Normalized) Hurbert Γ statistic		SD validity index
		Silhouette

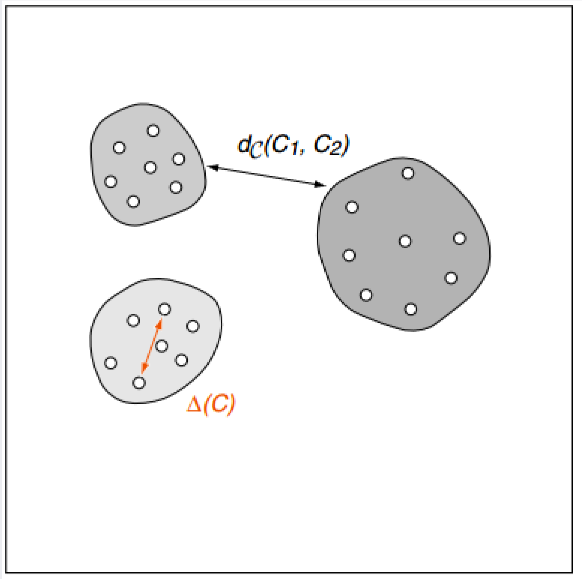
Clustering: Issues

- 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

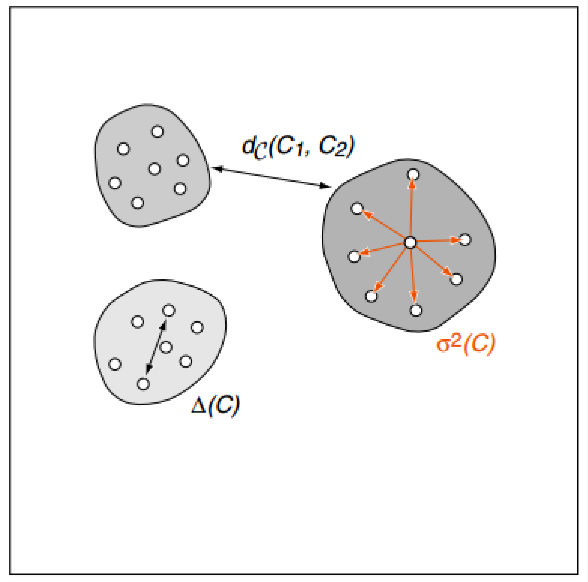
(1) Distance between two clusters



(2) Diameter of a cluster

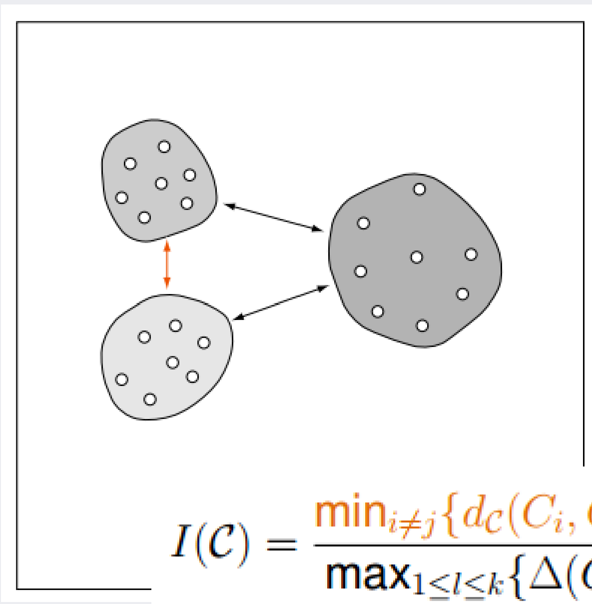


(3) Scatter within a cluster (SSE)



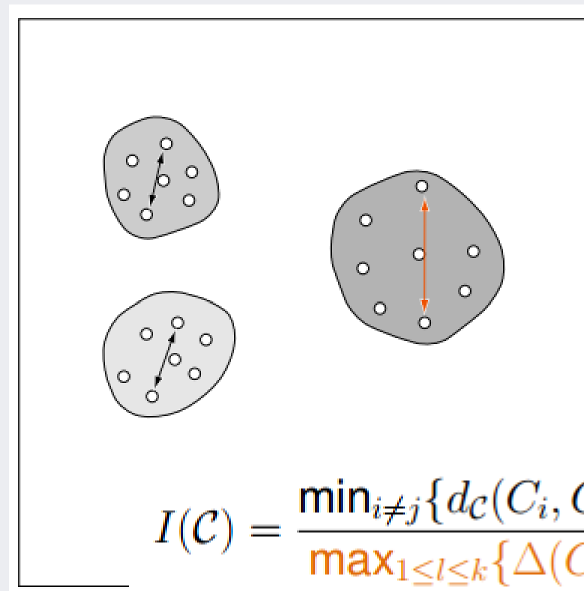
Clustering: Issues

- Clustering Validity Measure Example: Dunn Index
  - ✓ Dunn index is defined the ratio of (1) the minimum distance between two clusters to (2) the maximum diameter of the clusters



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

$$I(\mathcal{C}) \rightarrow \max$$



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

$$I(\mathcal{C}) \rightarrow \max$$

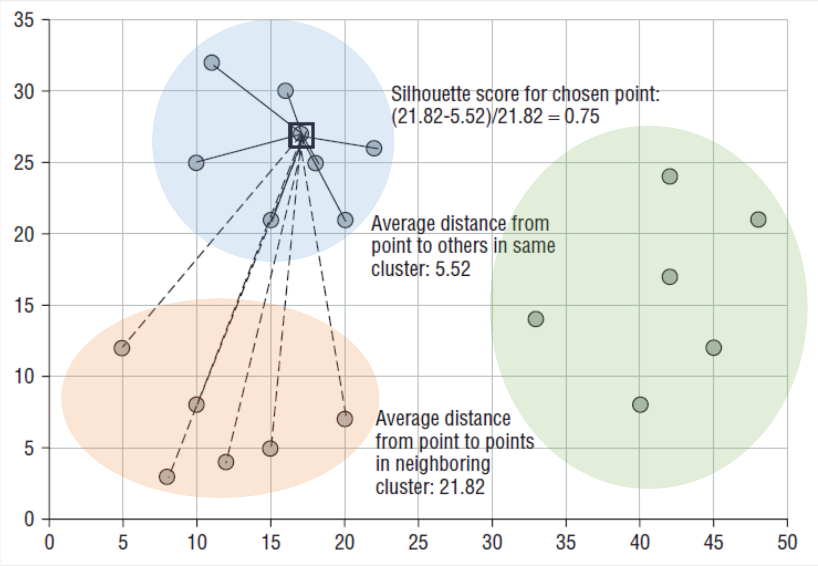
Clustering: Issues

- 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 is a cluster to which the instance i does not belong

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

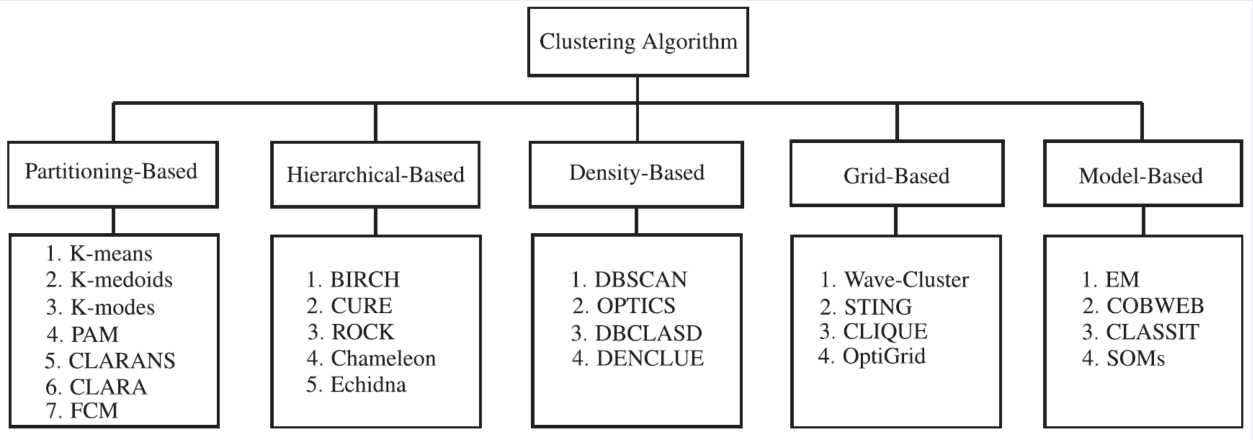
$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

$$-1 \leq s(i) \leq 1$$



Clustering: Types

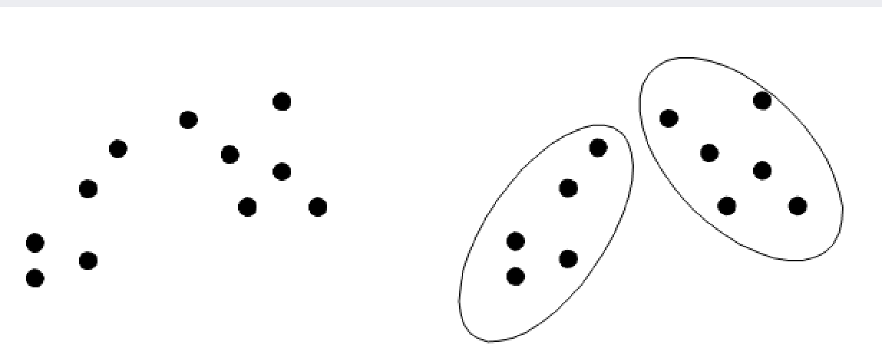
- Hard clustering vs. Soft clustering
  - ✓ Hard Clustering (Crisp Clustering)
    - Results in non-overlapping clusters
    - Each instance belongs to only one cluster



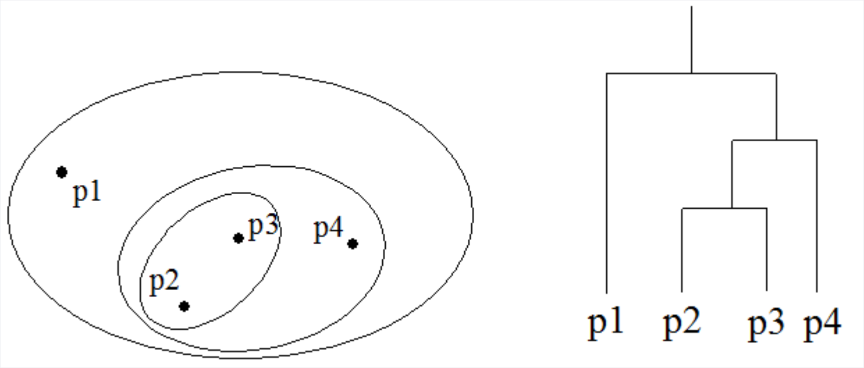
  - ✓ Soft Clustering (Fuzzy Clustering)
    - Possible to result in overlapping clusters
    - Each instance can belong to more than two clusters

Clustering: Algorithms

- <u>Partitional clustering</u>
  - ✓ Divide data into non-overlapping subsets such that each data object is in exactly one subset



- <u>Hierarchical clustering</u>
  - ✓ A set of nested clusters organized as a hierarchical tree



AGENDA

01	Clustering: Overview
02	K-Means Clustering
03	Hierarchical Clustering
04	R Exercise

K-Means Clustering

- K-Means Clustering (KMC)
  - ✓ Partitional clustering approach
    - Each cluster is associated with a centroid
    - Each point is assigned to the cluster with the closest centroid
    - Number of cluster, K, must be specified

$$\mathbf{X} = C_1 \cup C_2 \ldots \cup C_K, \quad C_i \cap C_j = \phi, \quad i \neq j$$

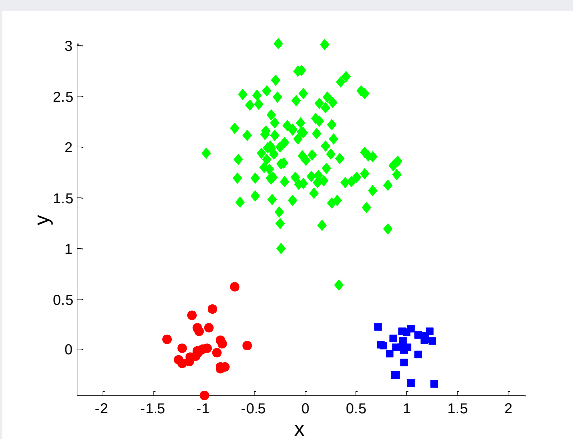
$$\arg\min_{\mathbf{C}} \quad \sum_{i=1}^{K} \sum_{\mathbf{x}_j \in C_i} ||\mathbf{x}_j - \mathbf{c}_i||^2$$

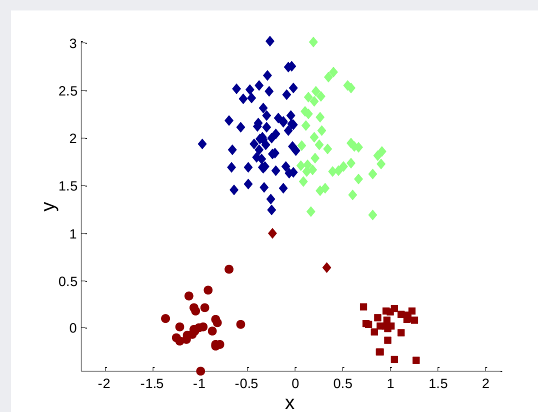
K-Means Clustering

- K-Means Clustering Procedure

Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change

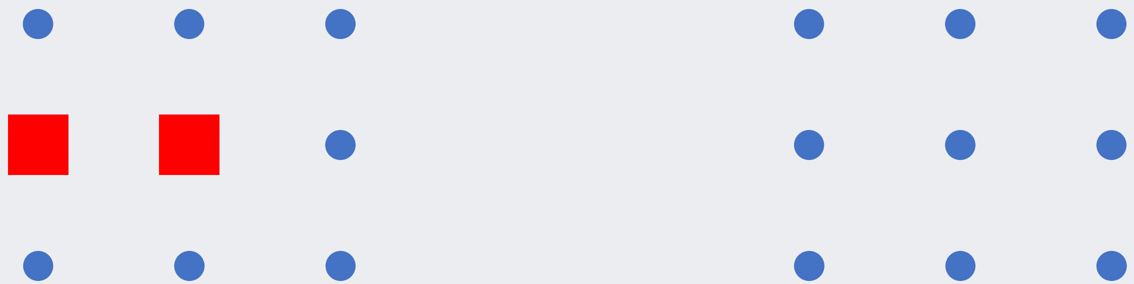
✓ Initial centroids are often chosen randomly: clustering results vary according to the initial centroid selection



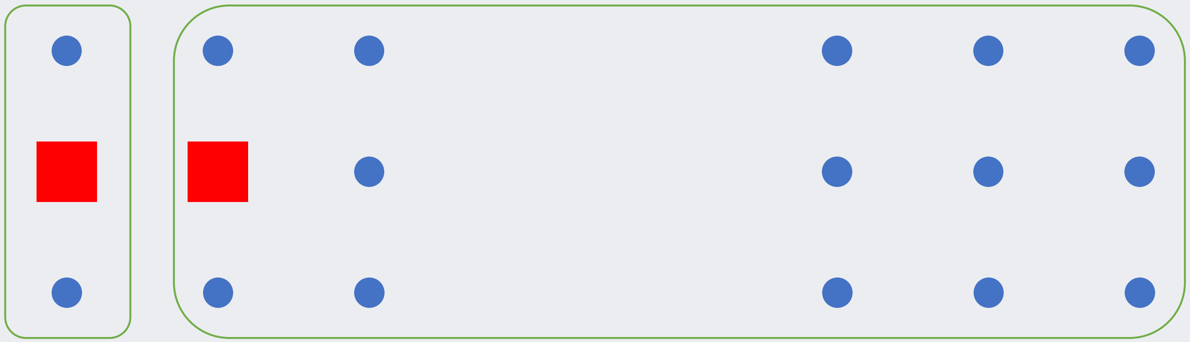


K-Means Clustering

- Example
  - ✓ Step 1: Initializing K centroids

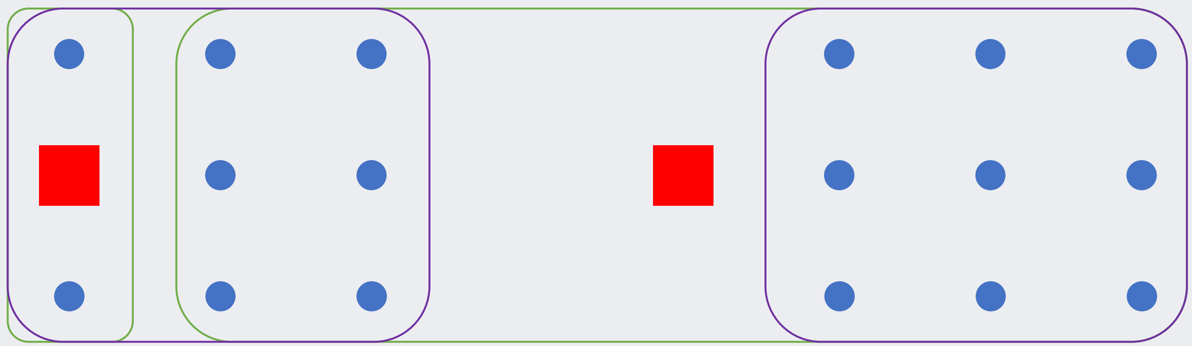


  - ✓ Step 2-1 (1st): Assign each instance to the closest center
  - ✓ Step 2-2 (1st): Re-compute the centroids based on the assigned instances

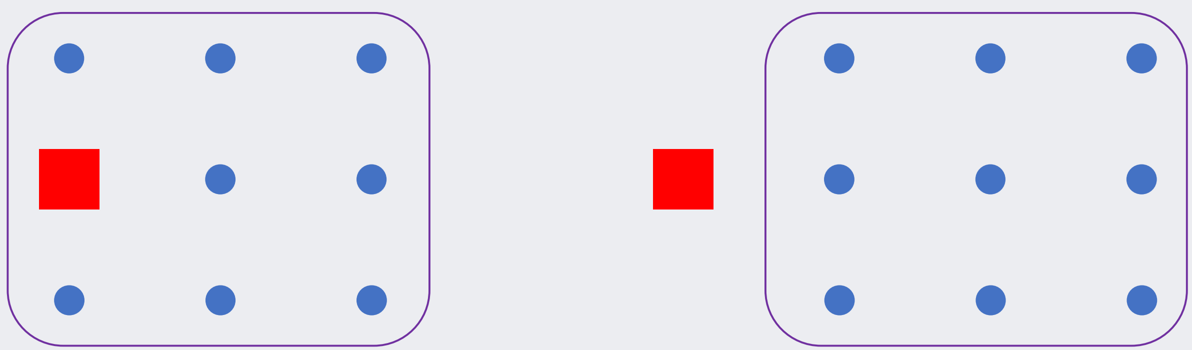


K-Means Clustering

- Example
  - ✓ Step 2-1 (2nd): Assign each instance to the closest center



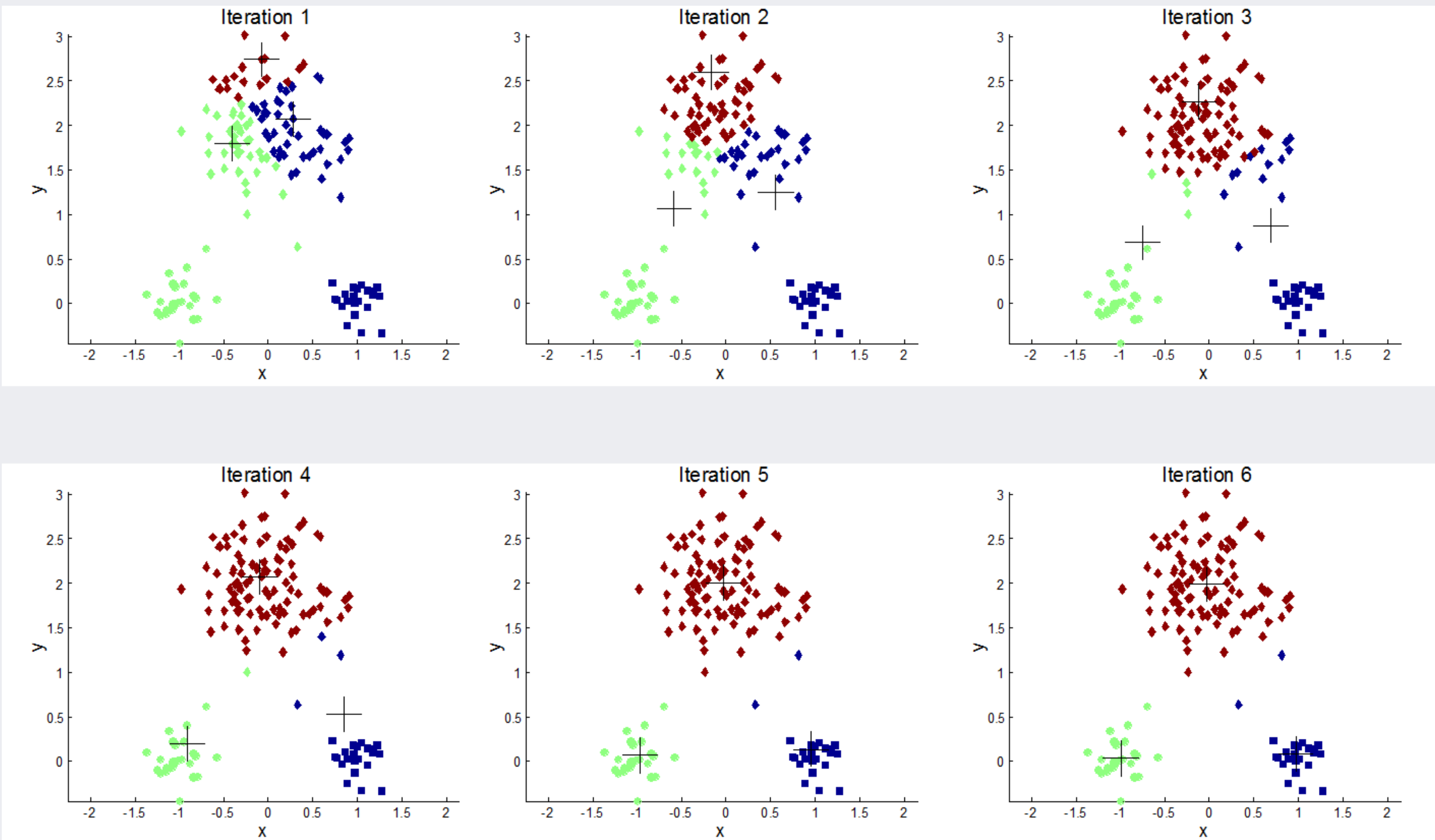
  - ✓ Step 2-2 (2nd): Re-compute the centroids based on the assigned instances



  - ✓ Stop the algorithm because there is no change for centroids and membership assignment

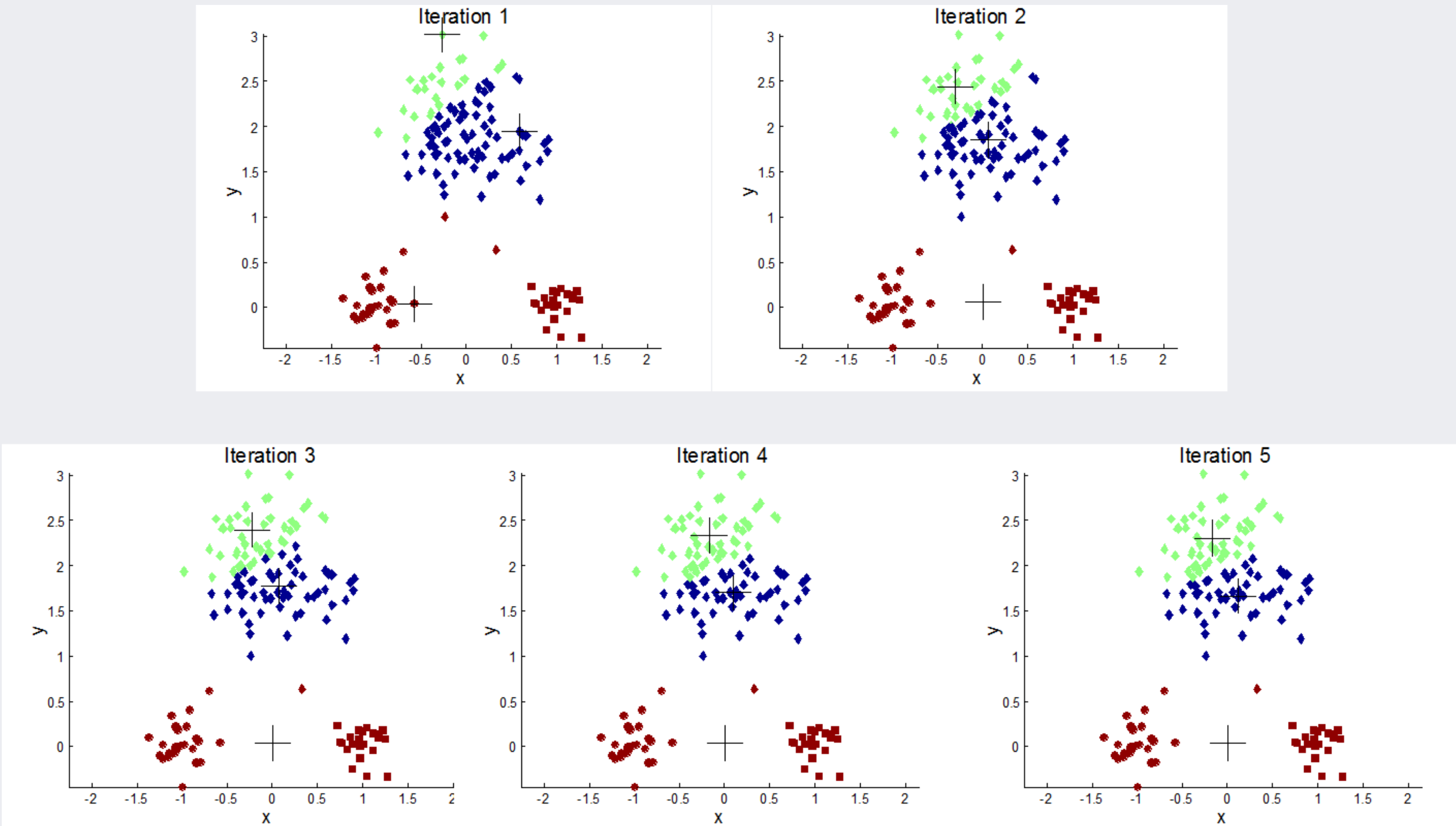
K-Means Clustering

- Effect of initial centroids
  - ✓ Desirable centroid selection



K-Means Clustering

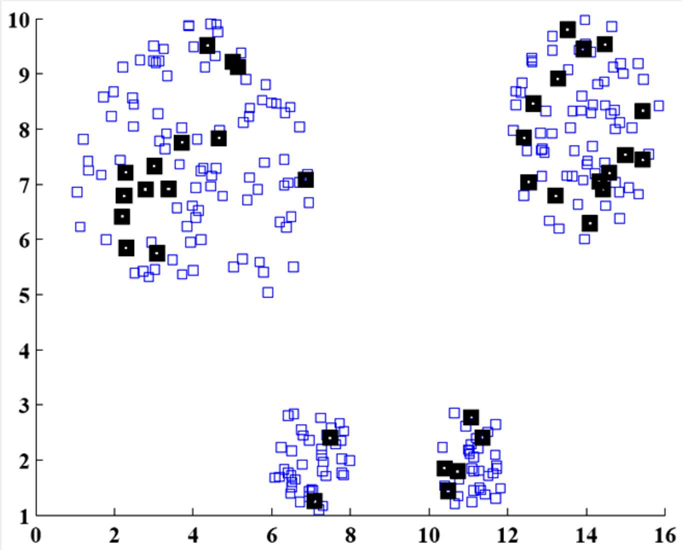
- Effects of initial centroids
  - ✓ Undesirable centroid selection

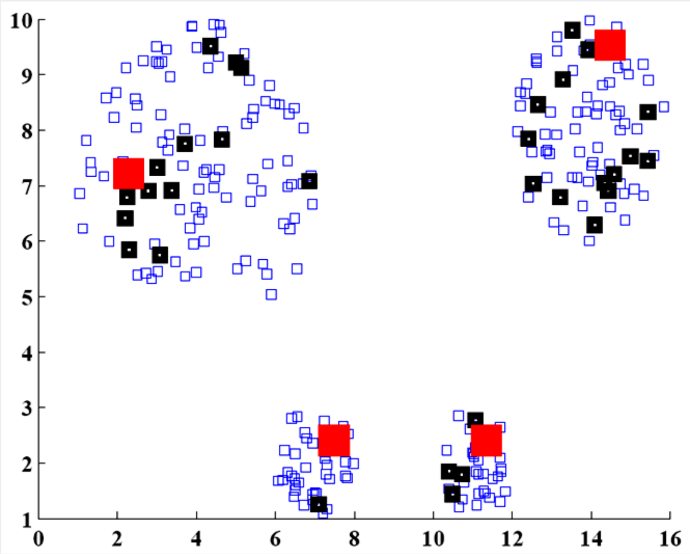


K-Means Clustering

- Some remedies for initial centroid selection
  - ✓ Multiple runs
  - ✓ Sample and use hierarchical clustering to determine initial centroids
  - ✓ Preprocessing & Postprocessing

$$\mathcal{L}(\mathbf{x}_s|\mathbf{S},\mathbf{C}) = d_G(\mathbf{x}_s,\mathbf{S}) \times \frac{1}{1+\exp\big(-d_R(\mathbf{x}_s,\mathbf{S})\big)}$$

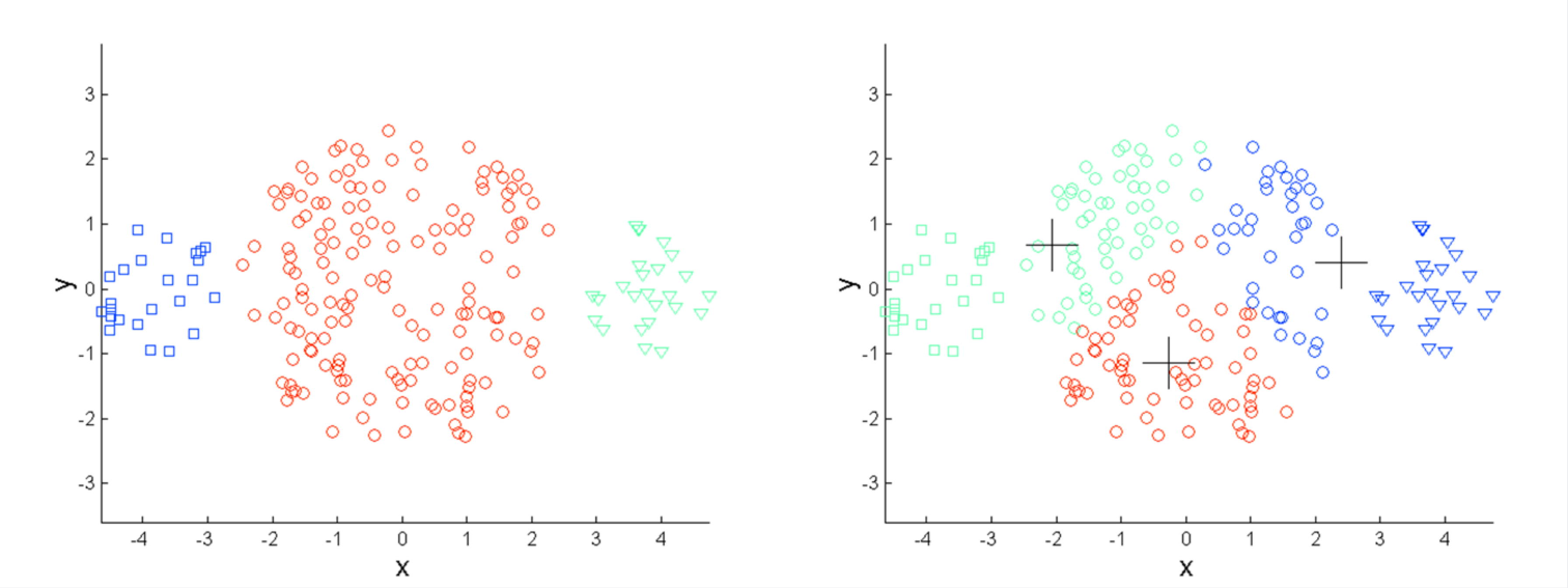




Pilsung Kang and Sungzoon Cho. (2009). K-Means clustering seeds initialization based on centrality, sparsity, and isotropy. *The 13th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2009)*, Burgos, Spain. E. Corchado and H. Yin (Eds.), ***Lecture Notes in Computer Science LNCS 5788***, 109-117.

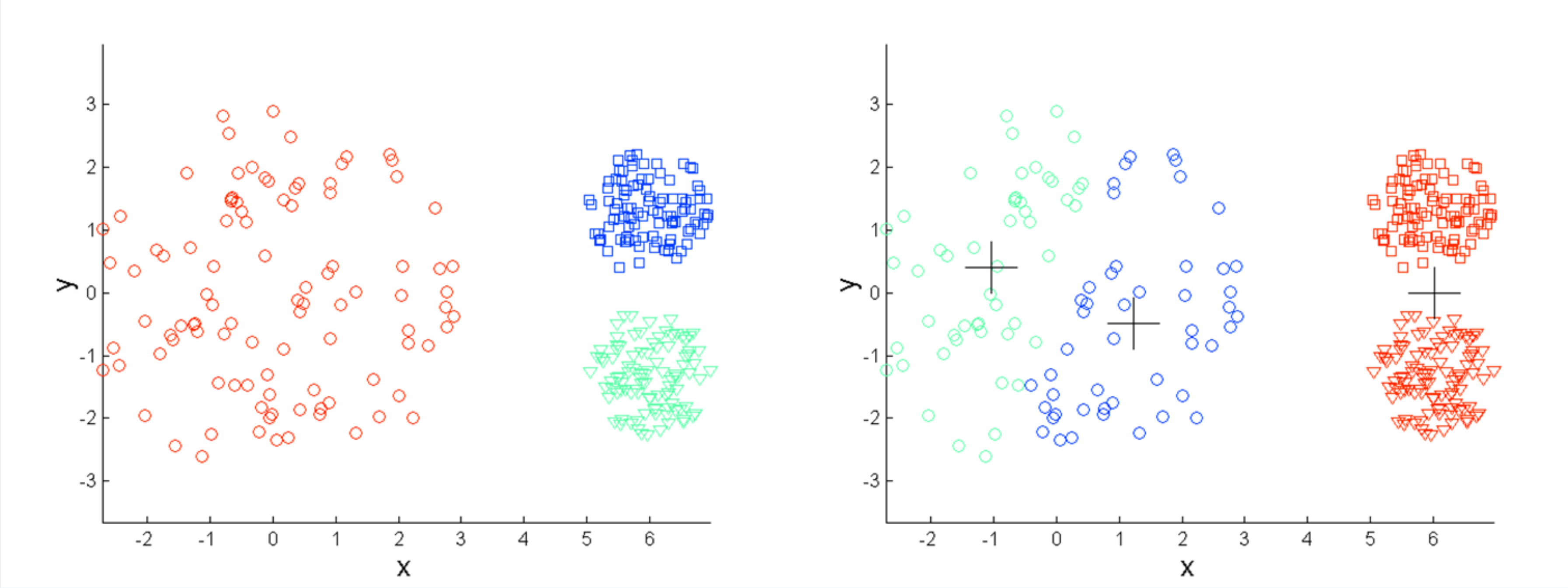
K-Means Clustering

- Limitations of K-Means Clustering
  - ✓ Cannot cope with different sizes



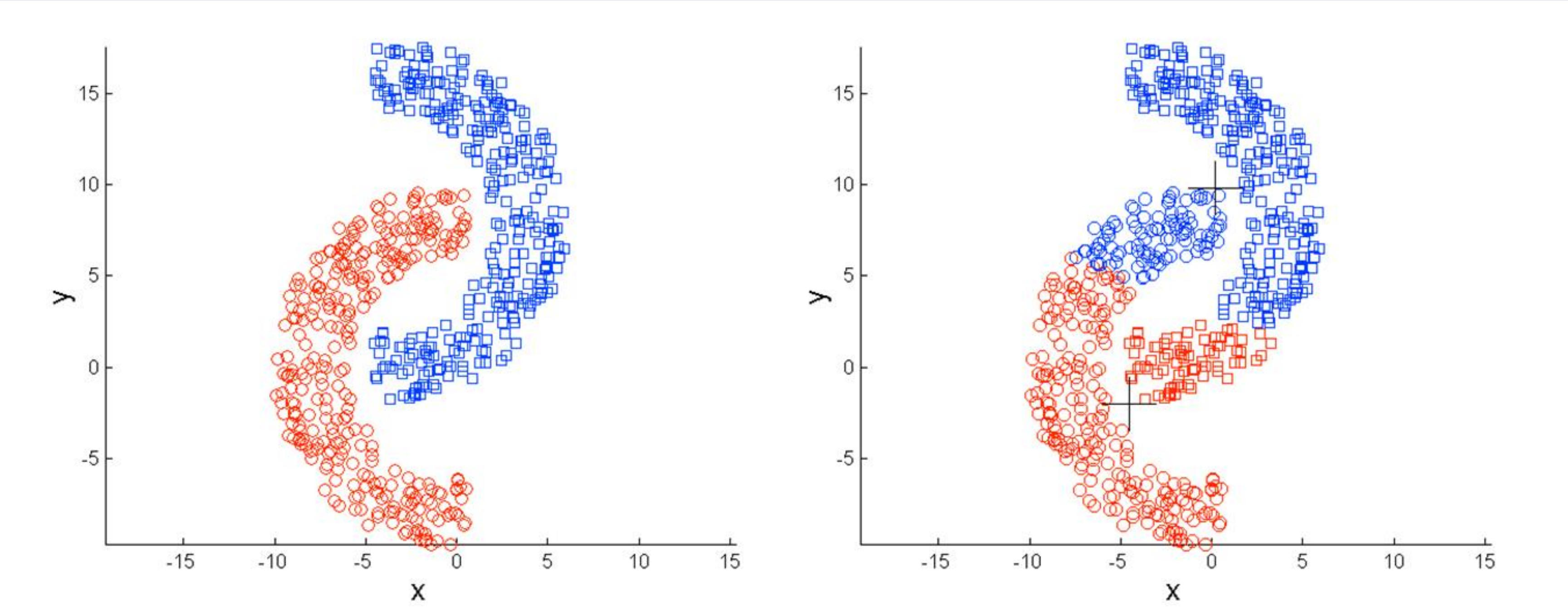
K-Means Clustering

- Limitations of K-Means Clustering
  - ✓ Cannot cope with different densities



K-Means Clustering

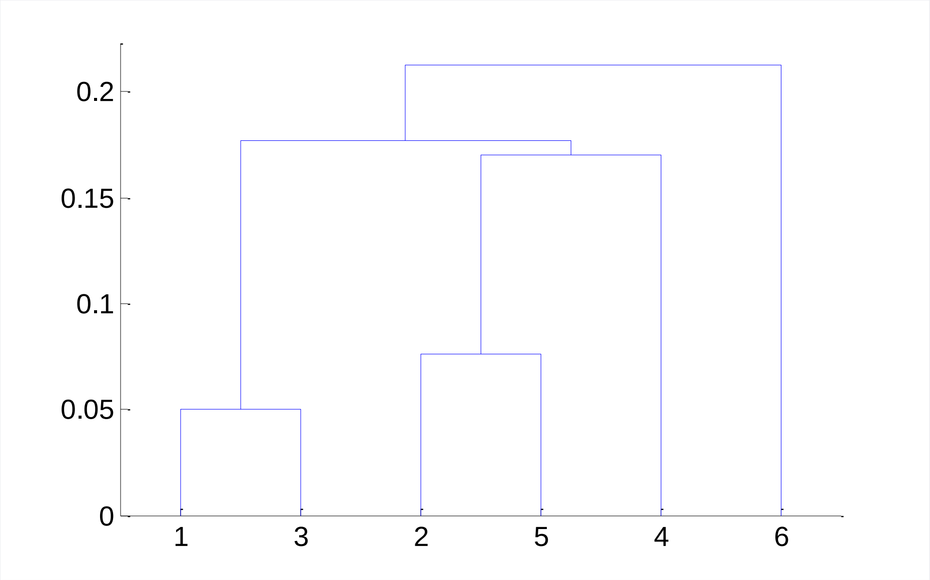
- Limitations of K-Means Clustering
  - ✓ Cannot cope with non-globular shapes

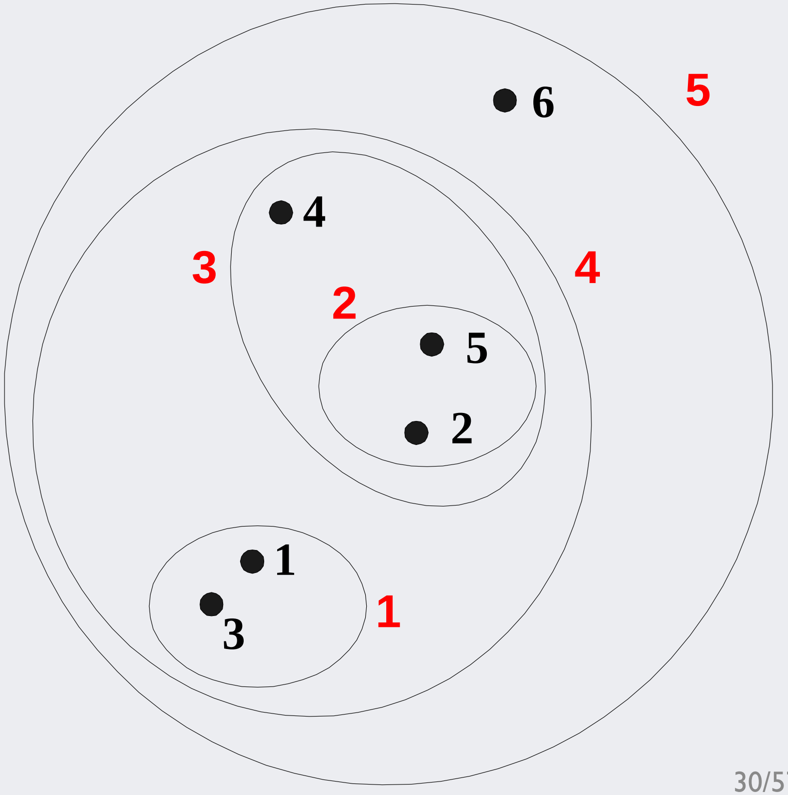


AGENDA

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



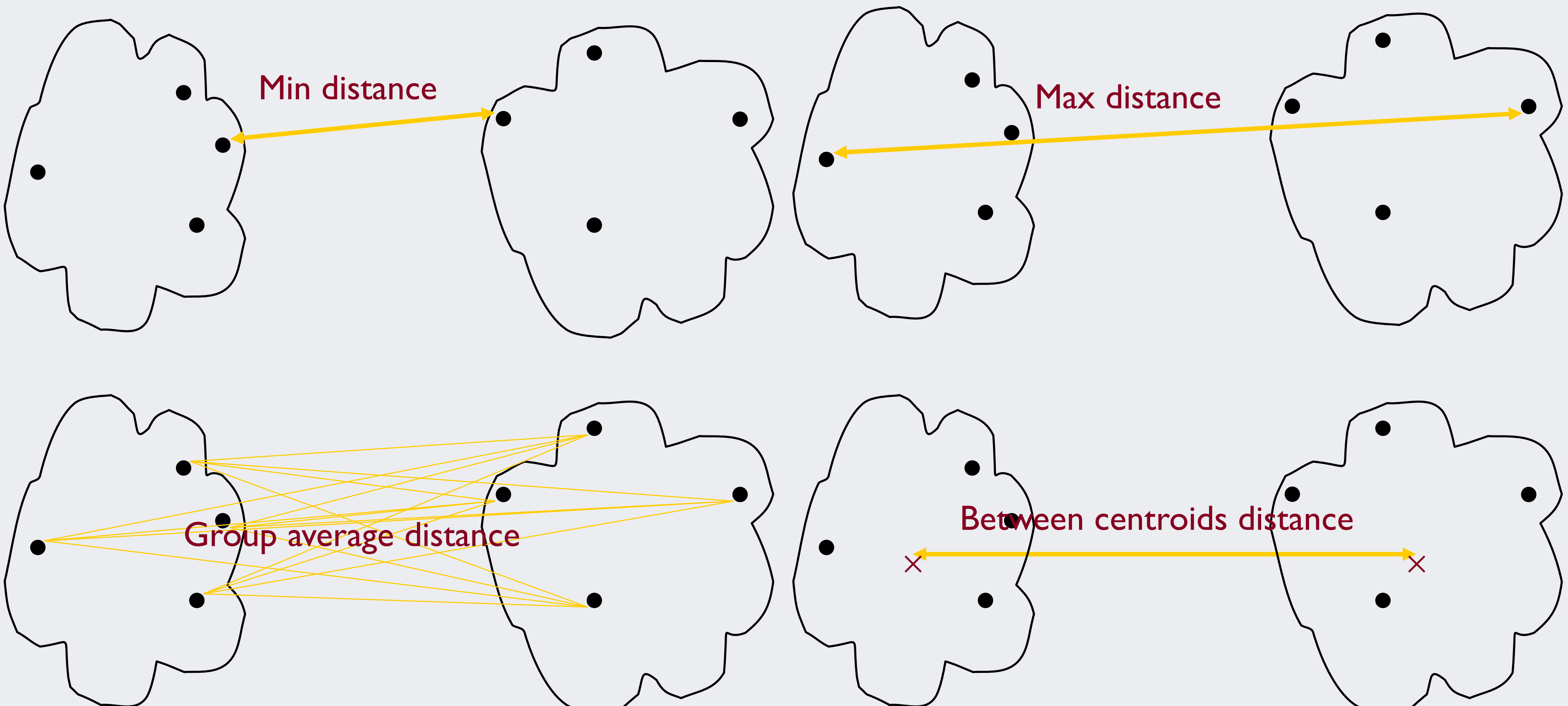


Hierarchical Clustering

- 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
  - ✓ Agglomerative clustering
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster left
  - ✓ Divisive clustering
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point

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
  - ✓ Complete linkage: maximum distance
  - ✓ Average linkage: mean distance
  - ✓ Centroid linkage: distance between centroids

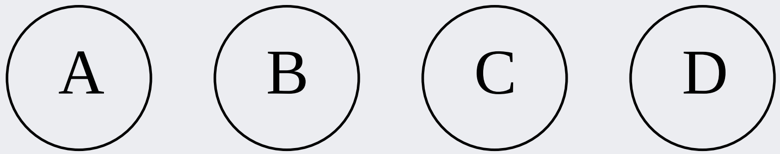
Hierarchical Clustering

- Agglomerative clustering algorithm
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. **Repeat**
     1. Merge the two closest clusters
     2. Update the proximity matrix
  4. **Until** only a single cluster remains

Hierarchical Clustering

- Example

Initial Data Items



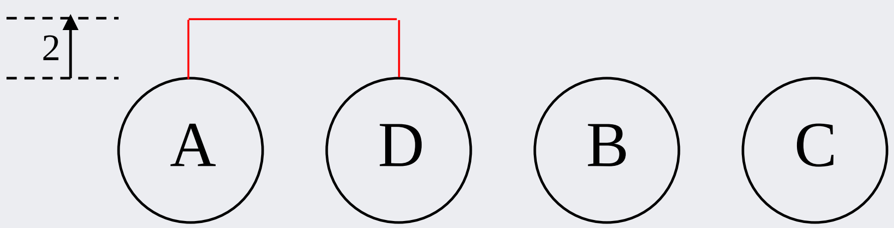
Distance Matrix

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

Hierarchical Clustering

- Example

Current Clusters



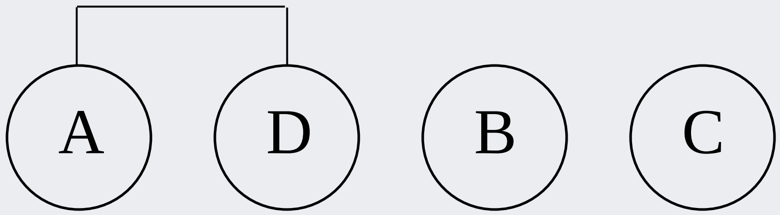
Distance Matrix

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

Hierarchical Clustering

- Example

Current Clusters



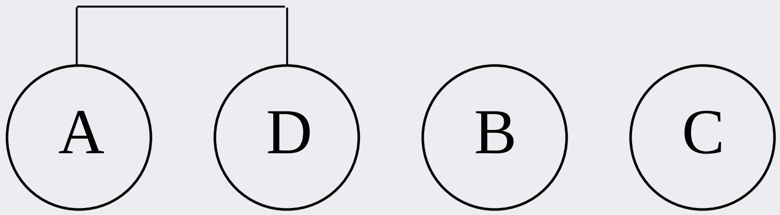
Distance Matrix

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Hierarchical Clustering

- Example

Current Clusters



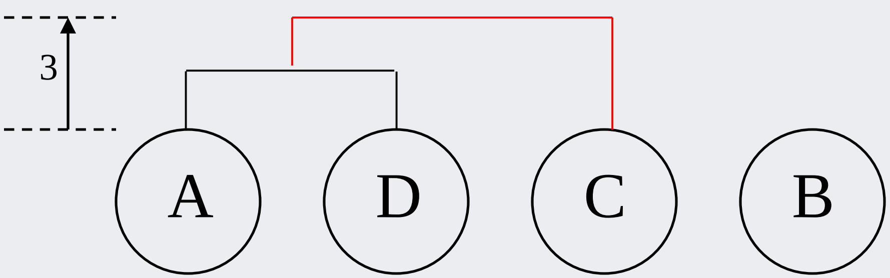
Distance Matrix

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Hierarchical Clustering

- Example

Current Clusters



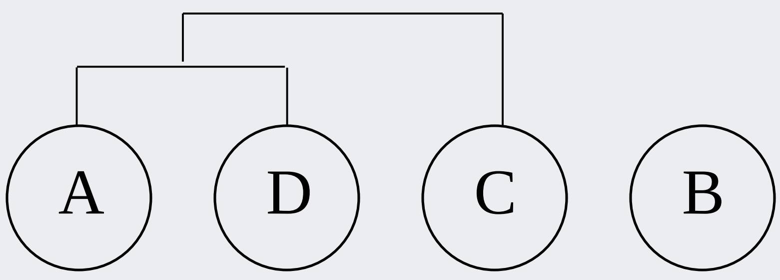
Distance Matrix

Dist	AD	B	C	
AD		20	3	
B			10	
C				

Hierarchical Clustering

- Example

Current Clusters



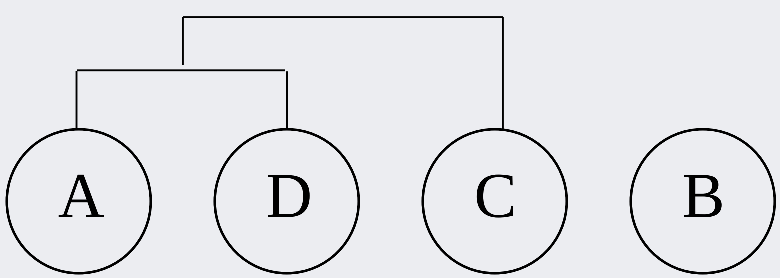
Distance Matrix

Dist	AD C	B		
AD C		10		
B				

Hierarchical Clustering

- Example

Current Clusters



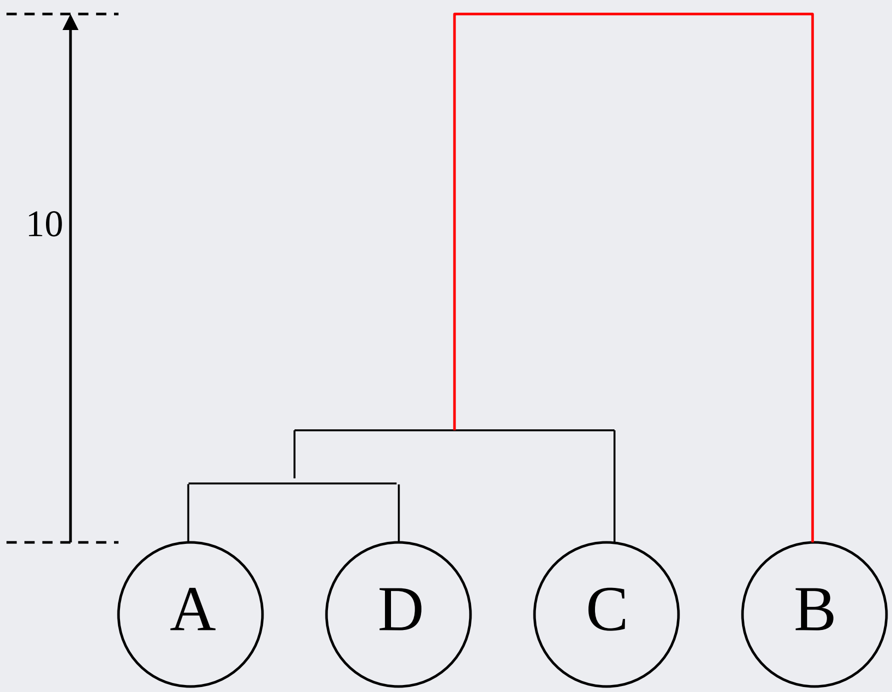
Distance Matrix

Dist	AD C	B		
AD C		10		
B				

Hierarchical Clustering

- Example

Current Clusters



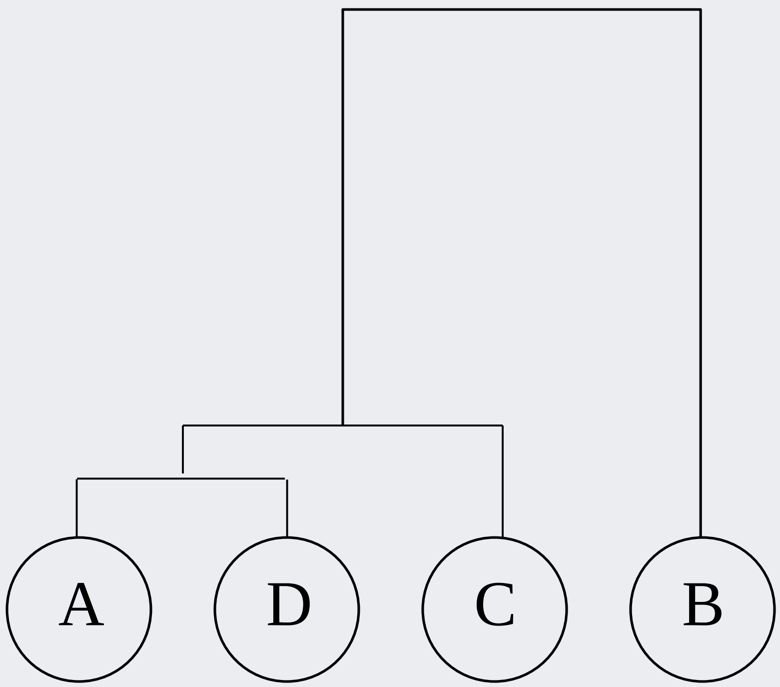
Distance Matrix

Dist	AD C	B		
AD C		10		
B				

Hierarchical Clustering

- Example

Final Result

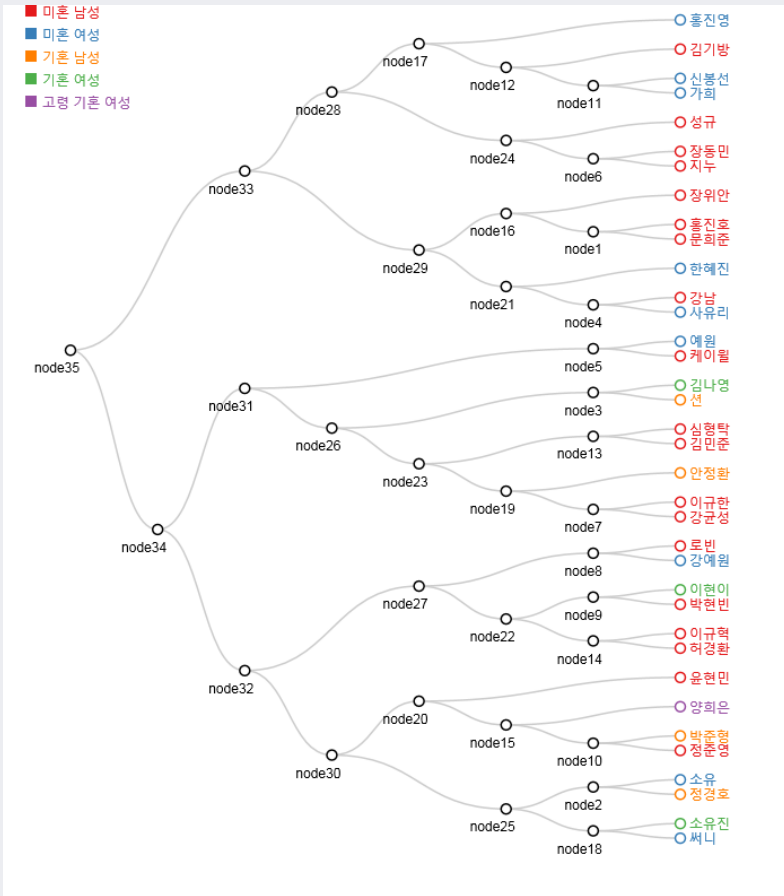


Distance Matrix

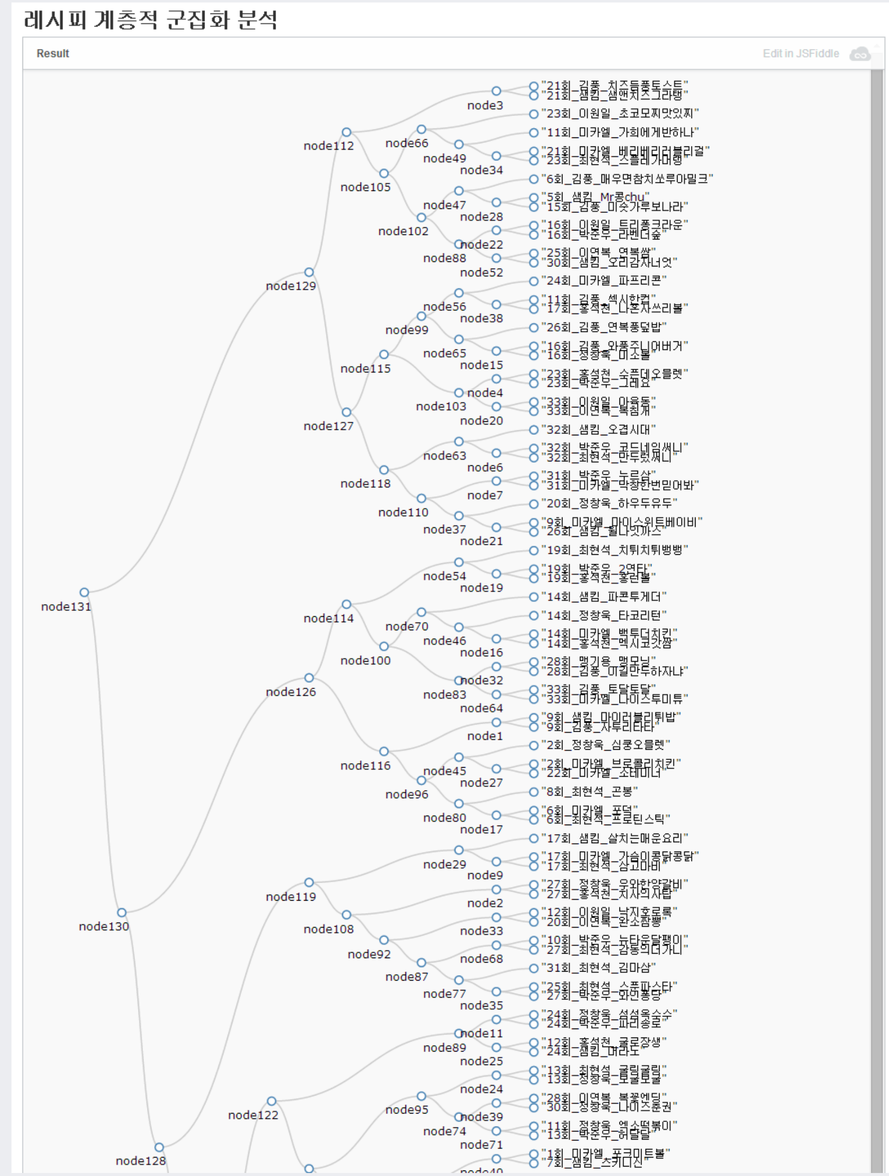
Dist	AD CB			
AD CB				

Hierarchical Clustering

- 냉장고를 부탁해!



냉장고 재료를 이용한 게스트 군집화



AGENDA

R Exercise: K-Means Clustering

- R packages providing K-Means Clustering
  - ✓ stats, kml, kml3d, RSKC, skmeans, sparcl, etc.
- Use the "iris" dataset

```r
# Package for cluster validity
install.packages("clValid")
library(clValid)

# Load the Iris dataset
data(iris)

# Part 1: K-Means Clustering ----------------------------------------------
# Remove the class label
newiris <- iris
newiris$Species <- NULL
rownames(newiris) <- paste("I", 1:150, sep = "_")

# Perform K-Means Clustering with K=3
kc <- kmeans(newiris,3)

str(kc)
kc$centers
kc$size
kc$cluster
```

R Exercise: K-Means Clustering

- Clustering results
  - ✓ Centroids, the number of instances in each cluster, cluster memberships, etc.

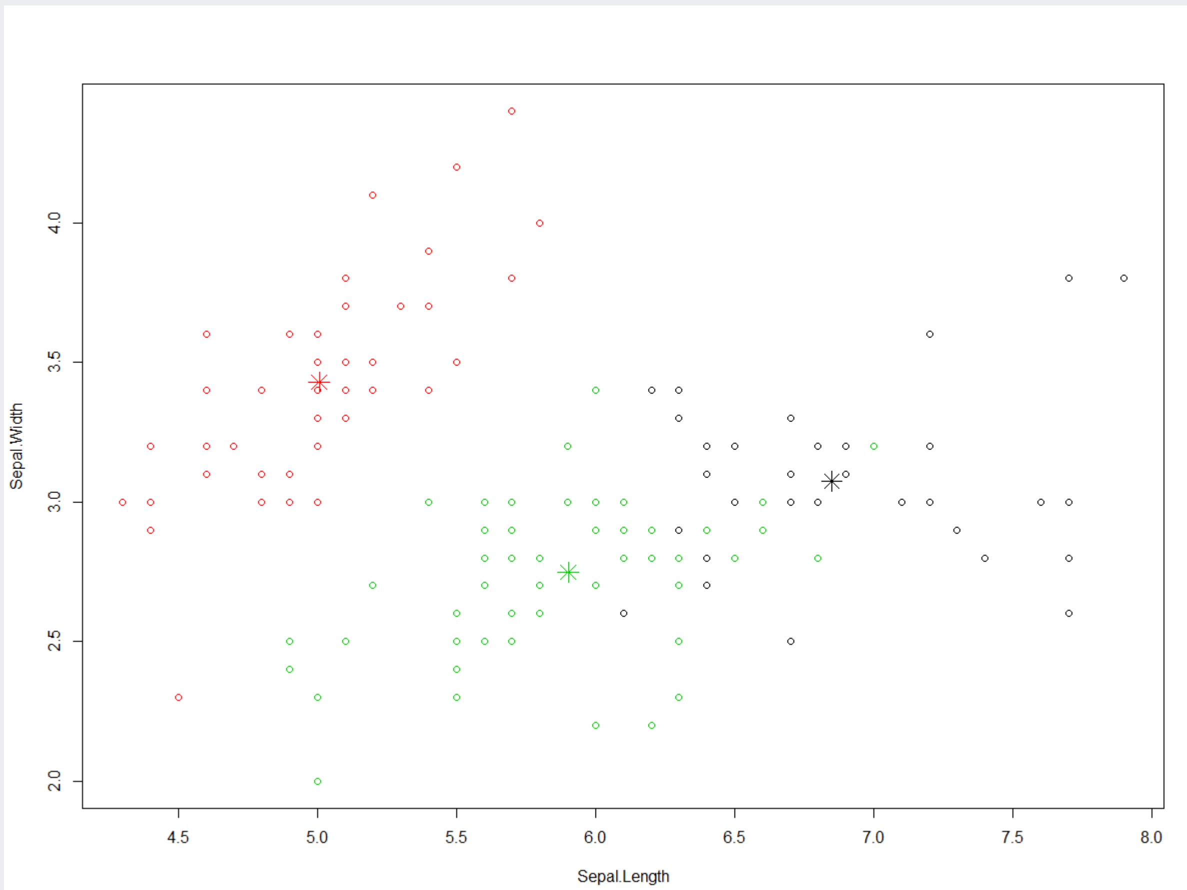
```
> str(kc)
List of 9
 $ cluster     : int [1:150] 2 2 2 2 2 2 2 2 2 2 ...
 $ centers     : num [1:3, 1:4] 6.85 5.01 5.9 3.07 3.43 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3] "1" "2" "3"
  .. ..$ : chr [1:4] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width"
 $ totss       : num 681
 $ withinss    : num [1:3] 23.9 15.2 39.8
 $ tot.withinss: num 78.9
 $ betweenss   : num 603
 $ size        : int [1:3] 38 50 62
 $ iter        : int 2
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
> kc$centers
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.850000    3.073684     5.742105    2.071053
2     5.006000    3.428000     1.462000    0.246000
3     5.901613    2.748387     4.393548    1.433871
> kc$size
[1] 38 50 62
> kc$cluster
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 1 3 3 3 3 3 3 3 3
 [62] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 1 1 1 3 1 1 1 1 1 1 3 3 1 1 1 1 3 1 3
[123] 1 3 1 1 3 3 1 1 1 1 1 3 1 1 1 1 3 1 1 1 3 1 1 1 3 1 1 3
> # Compare the assigned clusters and the Species
> table(iris$Species, kc$cluster)

              1  2  3
  setosa      0 50  0
  versicolor  2  0 48
  virginica  36  0 14
```

R Exercise: K-Means Clustering

- Clustering result visualization

```r
plot(newiris[,c("Sepal.Length", "Sepal.Width")], col = kc$cluster)
points(kc$centers[,c("Sepal.Length", "Sepal.Width")], col = 1:3, pch = 8, cex=2)
```



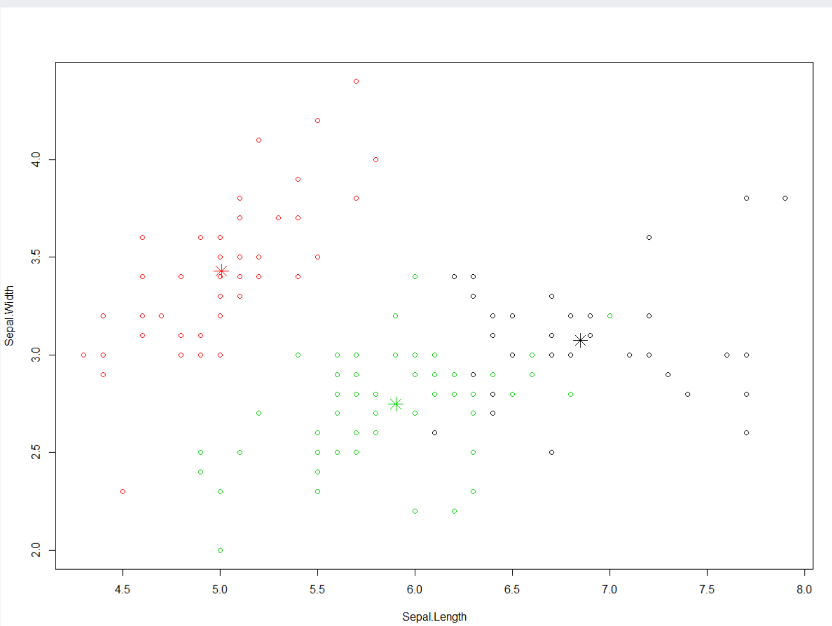
R Exercise: K-Means Clustering

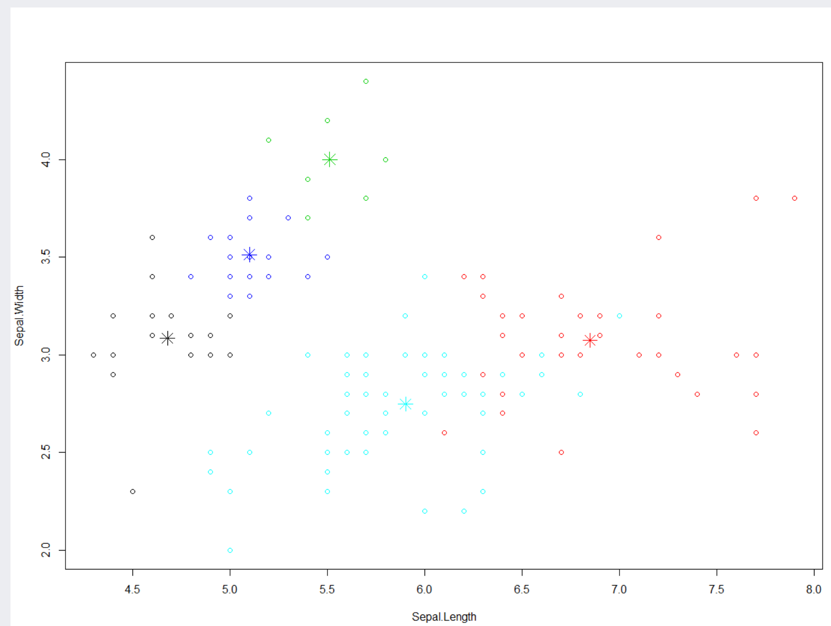
- Re-run KMC with k=5

```r
# Perform K-Means Clustering with K=5
kc <- kmeans(newiris,5)

# Compare the assigned clusters and the Species
table(iris$Species, kc$cluster)

plot(newiris[,c("Sepal.Length", "Sepal.Width")], col = kc$cluster)
points(kc$centers[,c("Sepal.Length", "Sepal.Width")], col = 1:5, pch = 8, cex=2)
```





R Exercise: K-Means Clustering

- Comparing clustering validity measures

```r
# Evaluating the cluster validity measures
newiris.clValid <- clValid(newiris, 2:10, clMethods = "kmeans", validation = c("internal", "stability"))
summary(newiris.clValid)
```

```
> summary(newiris.clValid)

Clustering Methods:
 kmeans

Cluster sizes:
 2 3 4 5 6 7 8 9 10

Validation Measures:
                           2       3       4       5       6       7       8       9      10

kmeans APN            0.0130  0.0630  0.1572  0.2394  0.1680  0.1954  0.2212  0.2198  0.2619
       AD             1.2223  0.9390  0.8722  0.8149  0.7309  0.6946  0.6804  0.6489  0.6306
       ADM            0.0562  0.1131  0.2803  0.3316  0.2293  0.2340  0.2523  0.2245  0.2593
       FOM            0.4990  0.3935  0.3590  0.3534  0.3354  0.3144  0.3131  0.3050  0.3009
       Connectivity   6.1536 10.0917 17.5194 27.9373 36.4873 33.9595 38.9556 49.9901 58.0988
       Dunn           0.0765  0.0988  0.1365  0.0823  0.0853  0.0872  0.0872  0.0617  0.0684
       Silhouette     0.6810  0.5528  0.4981  0.4887  0.3648  0.3609  0.3556  0.3360  0.3391

Optimal Scores:

             Score  Method Clusters
APN          0.0130 kmeans 2
AD           0.6306 kmeans 10
ADM          0.0562 kmeans 2
FOM          0.3009 kmeans 10
Connectivity 6.1536 kmeans 2
Dunn         0.1365 kmeans 4
Silhouette   0.6810 kmeans 2
```

R Exercise: Hierarchical Clustering

- Clustering bank customers: Personal Loan dataset

Data Description:

ID	Customer ID
Age	Customer's Age in completed years
Experience	#years of professional experience
Income	Annual income of the customer ($000)
ZIPCode	Home Address ZIP code.
Family	Family size (dependents) of the customer
CCAvg	Avg. Spending on Credit Cards per month ($000)
Education	Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage	Value of house mortgage if any. ($000)
Personal Loan	Did this customer accept the personal loan offered in the last campaign?
Securities Account	Does the customer have a Securities account with the bank?
CD Account	Does the customer have a Certificate of Deposit (CD) account with the bank?
Online	Does the customer use internet banking facilities?
CreditCard	Does the customer use a credit card issued by UniversalBank?

R Exercise: Hierarchical Clustering

- Clustering bank customers: Personal Loan dataset
  - ✓ Use Pearson correlation coefficient to compute the similarity between customers
  - ✓ Use complete linkage to compute the distance between clusters

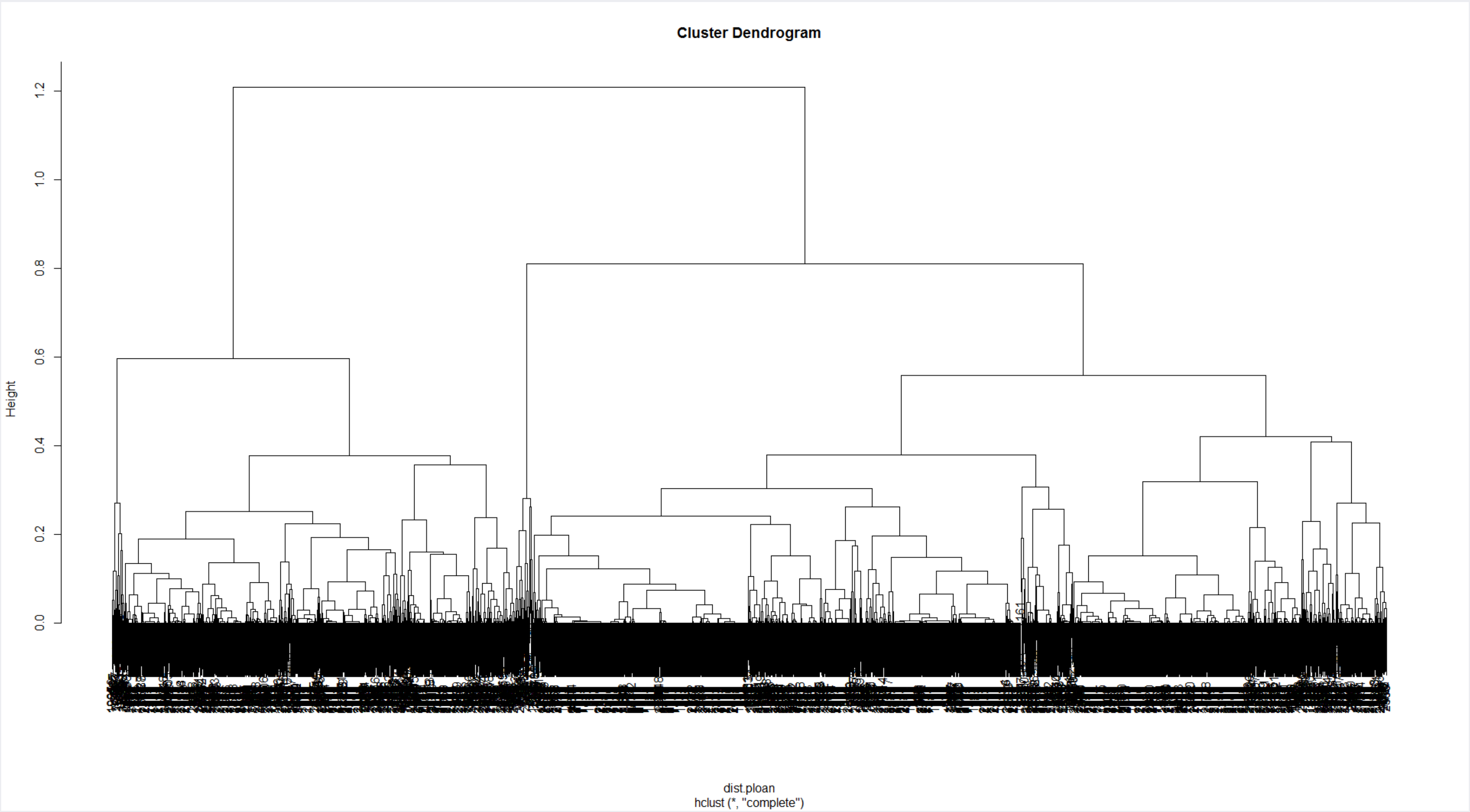
```r
# Part 2: Hierarchical Clustering ----------------------------------------
ploan <- read.csv("Personal Loan.csv")
ploan.x <- ploan[,-c(1,5,10)]

# Compute the similarity using the spearman coefficient
cor.Mat <- cor(t(ploan.x), method = "spearman")
dist.ploan <- as.dist(1-cor.Mat)

# Perform hierarchical clustering
hr <- hclust(dist.ploan, method = "complete", members=NULL)
```

R Exercise: Hierarchical Clustering

- Dendrogram

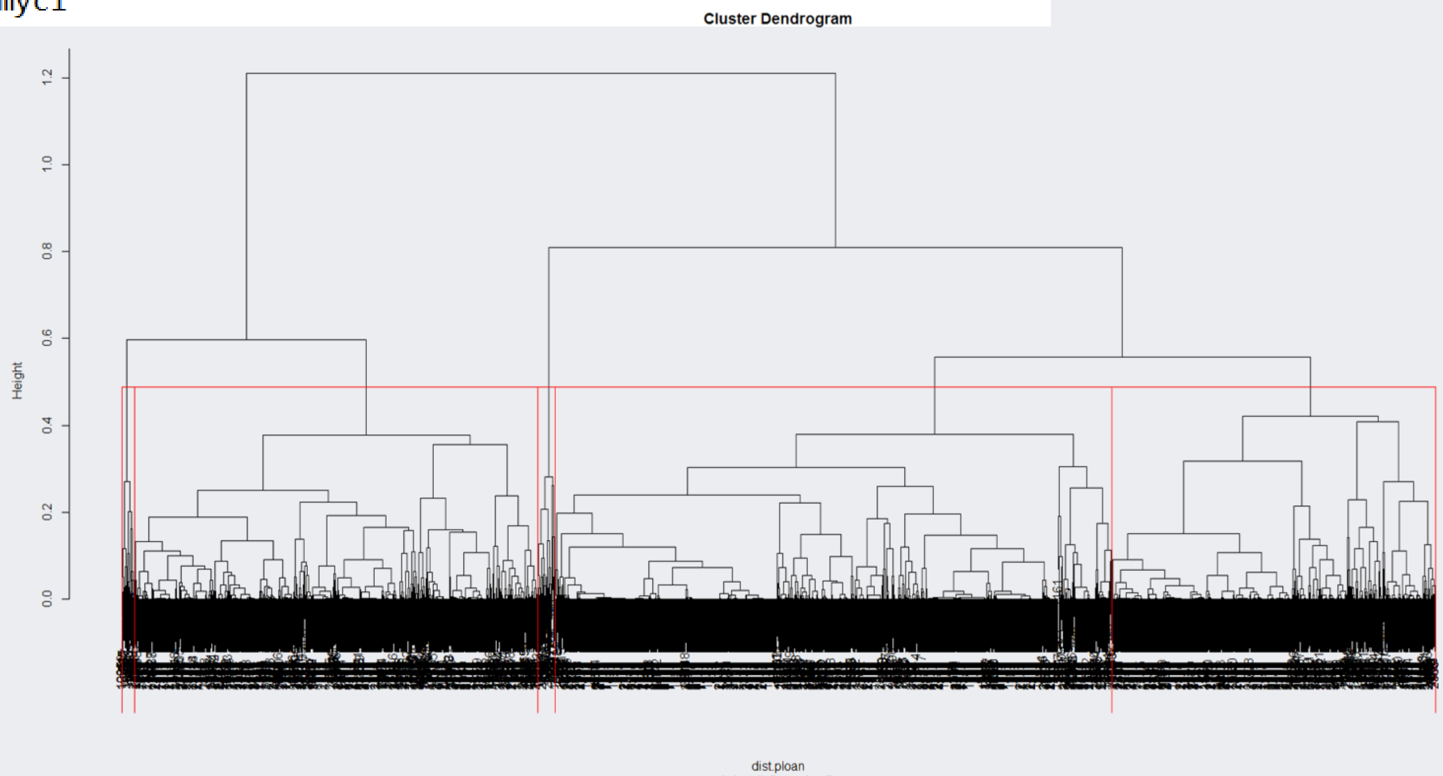


R Exercise: Hierarchical Clustering

- Perform clustering with k=5

```
# plot the results
plot(hr)
plot(hr, hang = -1)
plot(as.dendrogram(hr), edgePar=list(col=3, lwd=4), horiz=T)

# Find the clusters
mycl <- cutree(hr, k=5)
mycl
```



R Exercise: Hierarchical Clustering

- Compare the clusters

```r
# Compare each cluster
segment.ploan <- cbind(ploan.x, ploanYN = ploan[,10], clusterID = as.factor(mycl))
segment.summary <- data.frame()

for (i in 1:(dim(segment.ploan)[2]-1)){
  segment.summary = rbind(segment.summary,
                          tapply(segment.ploan[,i], segment.ploan$clusterID, mean))
}

colnames(segment.summary) <- paste("cluster", c(1:5))
rownames(segment.summary) <- c(colnames(ploan.x), "LoanRatio")
segment.summary
```

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
Age	45.14919736	45.86962190	46.91558442	24.96969697	25.8400
Experience	19.96411709	20.67796610	21.62012987	-0.60606061	0.0800
Income	87.71671388	73.39504563	52.86688312	71.30303030	80.5200
Family	2.33899906	2.42503259	2.44805195	2.90909091	3.1600
CCAvg	2.69259679	1.86147327	0.78404221	1.82878788	2.2272
Education	1.71671388	1.86440678	2.09740260	2.18181818	2.0800
Mortgage	0.00000000	180.92568449	0.00000000	0.00000000	188.0400
Securities.Account	0.10103872	0.11473272	0.11688312	0.12121212	0.1200
CD.Account	0.06421152	0.07953064	0.04220779	0.03030303	0.0000
Online	0.46553352	0.59061278	0.83441558	0.57575758	0.6000
CreditCard	0.28706327	0.27770535	0.31168831	0.33333333	0.2400
LoanRatio	0.13408876	0.11734029	0.03246753	0.09090909	0.0400

R Exercise: Hierarchical Clustering

- Compare the clusters

```
# Radar chart
segment.summary <- t(segment.summary)
stars(segment.summary, locations = c(0, 0),
        radius = TRUE, key.loc = c(0, 0), col.lines = 2:6,
        main = "Customer Segmentation", lty = 1, lwd = 2)
```