Soft Computing
Clustering - I

Dr. Chun-Hao Chen

Outline

1、Types of Machine Learning (ML)

2、Clustering Algorithms and their Problems

3、Factors should be considered in Clustering

4、Implementation in Python I & II

Machine Learning

- ✓Teach machine to learn to recognize things
- ✓Divide into three categories

ML

- ✓Supervised learning
  - ●e.g., Classification (Decision Tree)
- ✓Unsupervised learning
  - ●e.g., Clustering (k-means)
- ✓Semi-supervised learning
  - ●e.g.,Generative Adversarial Network

Utilizing a set of attributes
with or **without** the labels
to build a model (classifier)
to identify label of new instances
are supervised and unsupervised learning

What is Label?

✓For Tennis, the label is “to play(Yes)” or “not to play(No)”

- e.g.

Consider only attributes for learning
➔ Unsupervised Learning (Clustering)

Value Value

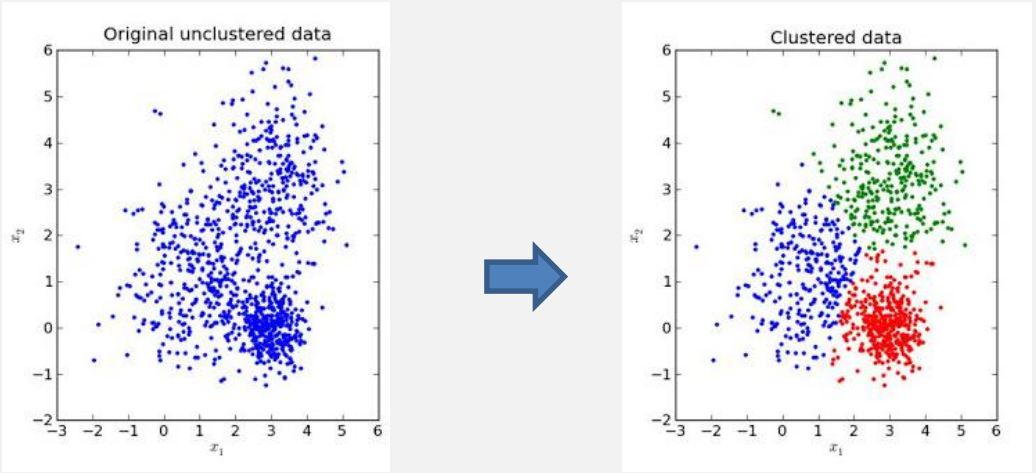
If Outlook is Sunny and Wind is Weak, then playTennis is Yes

Attribute Name Label

Consider attributes and label together for learning
➔ Supervised Learning

Clustering Concept

- ✓Divide instances into many groups by given attributes
  - ●e.g.
    - ✓Use *x* and *y* axis values to obtain three clusters
    - ✓Instances in the same cluster ➔ Similar instances



source: https://mropengate.blogspot.com/2015/06/ai-ch16-5-k-introduction-to-clustering.html

Case Study: BMW Dealership Dataset

✓9 attributes in the dataset

- CustomerID numeric, Dealership numeric, Showroom numeric, ComputerSearch numeric, M5 numeric, 3Series numeric, Z4 numeric, Financing numeric, Purchase numeric

✓ 20 instances selected from the dataset

@data
1,1,0,0,0,0,0,0,0
2,1,1,1,0,0,0,1,0
3,1,0,0,0,0,0,0,0
4,1,1,1,1,0,0,1,1
5,1,0,1,1,1,0,1,1
6,1,1,1,0,1,0,0,0
7,1,0,1,0,0,0,1,1
8,1,0,1,0,1,0,0,0
9,1,1,1,0,1,0,1,0
10,1,0,1,1,1,1,1,1

@data
11,1,0,1,1,1,1,1,0
12,1,0,1,1,0,1,0,0
13,1,0,1,1,0,0,1,1
14,1,1,1,0,0,1,1,0
15,1,0,1,1,1,1,0,0
16,1,1,1,1,1,0,1,1
17,1,0,1,0,0,0,1,1
18,1,0,1,0,0,0,1,0
19,1,1,0,1,1,0,0,0
20,1,0,0,1,0,0,0,0

Try to find customers with similar behavior from the 20 instances

Case Study: BMW Dealership Dataset (Cont.)

- ✓Use *k*-means algorithm for finding groups
- ✓When *k* = 5, via weak we can get following result

Attribute	Full Data (100)	Cluster# 0 (26)	1 (27)	2 (5)	3 (14)	4 (28)
Dealership	0.6	0.9615	0.6667	1	0.8571	0
Showroom	0.72	0.6923	0.6667	0	0.5714	1
ComputerSearch	0.43	0.6538	0	1	0.8571	0.3214
M5	0.53	0.4615	0.963	1	0.7143	0
3Series	0.55	0.3846	0.4444	0.8	0.0714	1
Z4	0.45	0.5385	0	0.8	0.5714	0.6786
Financing	0.61	0.4615	0.6296	0.8	1	0.5
Purchase	0.39	0	0.5185	0.4	1	0.3214

Clustered Instances

0	26	(26%)
1	27	(27%)
2	5	(5%)
3	14	(14%)
4	28	(28%)

Discussions

Attribute	Full Data (100)	Cluster# 0 (26)	1 (27)	2 (5)	3 (14)	4 (28)
Dealership	0.6	0.9615	0.6667	1	0.8571	0
Showroom	0.72	0.6923	0.6667	0	0.5714	1
ComputerSearch	0.43	0.6538	0	1	0.8571	0.3214
M5	0.53	0.4615	0.963	1	0.7143	0
3Series	0.55	0.3846	0.4444	0.8	0.0714	1
Z4	0.45	0.5385	0	0.8	0.5714	0.6786
Financing	0.61	0.4615	0.6296	0.8	1	0.5
Purchase	0.39	0	0.5185	0.4	1	0.3214

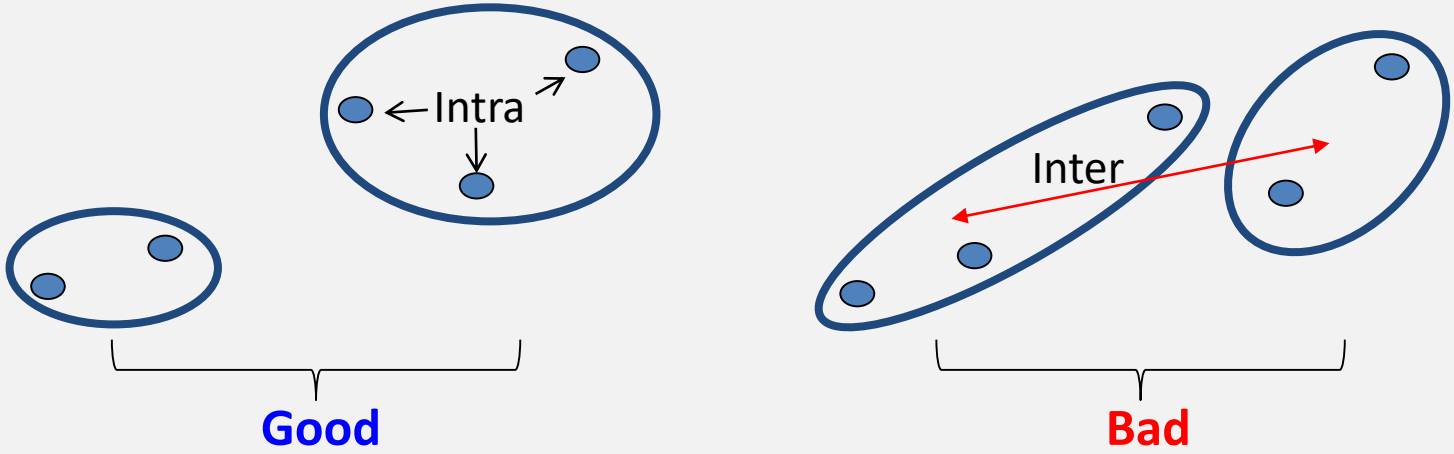
Clustered Instances

0	26	(26%)
1	27	(27%)
2	5	(5%)
3	14	(14%)
4	28	(28%)

What are the behaviors that you can find from the results?

The Well-Known Clustering Algorithm

- ✓The *k*-means clustering algorithm
  - ●Proposed by MacQueen, 1967
  - ●Unsupervised learning algorithm
- ✓Goal ➔ Minimize intra-cluster distance, and Maximize inter-cluster distance
  - ●e.g.



An Example

- ✓Assume
  - ●5 basketball players: P1, P2, P3, P4, P5
  - ●Two attributes: Speed and Weight

	P1	P2	P3	P4	P5
Speed	10sec(100m)	9sec(100m)	11sec(100m)	20sec(100m)	13sec(100m)
Weight	65kg	60kg	80kg	99kg	70kg

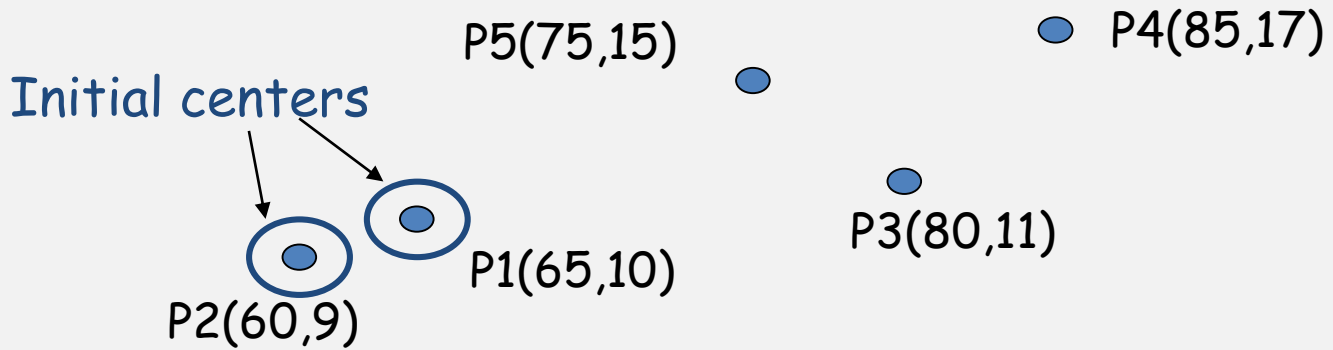
- ✓Illustrate how *k*-means clustering algorithm work
  - ●Number of cluster $k = 2$

Step 1

✓Initialize *k* centers

- Randomly select
- *k* = 2 ➔ Choose P1 and P2

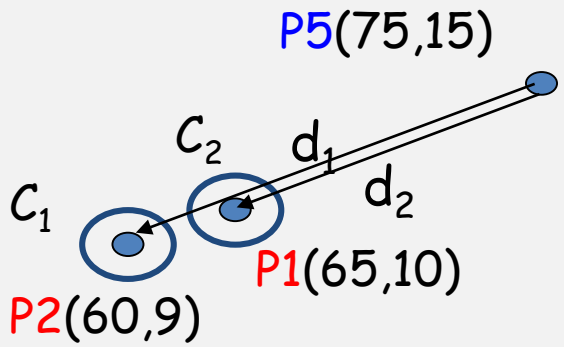
	P1	P2	P3	P4	P5
Speed	10sec(100m)	9sec(100m)	11sec(100m)	17sec(100m)	15sec(100m)
Weight	65kg	60kg	80kg	85kg	75kg



Step 2

✓Assign each object to the group

- According to the Euclidean distance
- e.g. P5



Center$_1$
P2(60,9)
P5(75,15)

Center$_2$
P1(65,10)
P5(75,15)

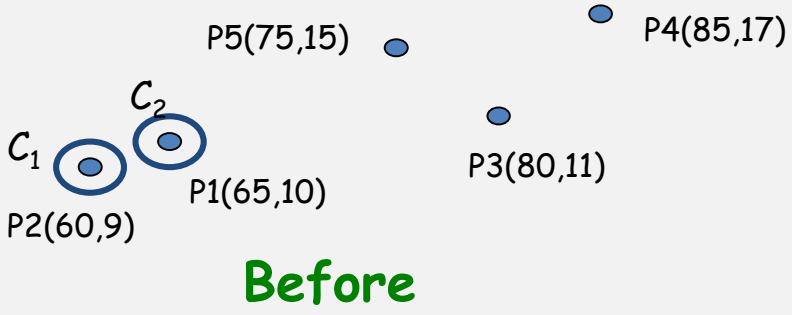
$[(60-75)^2+(9-15)^2]^{1/2} = 261^{1/2}$

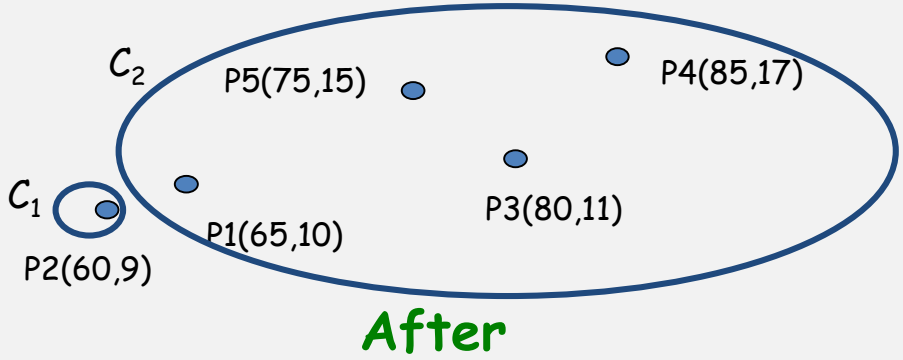
$[(65-75)^2+(10-15)^2]^{1/2} = 125^{1/2}$

Assign P5 to C_2

After Step 2

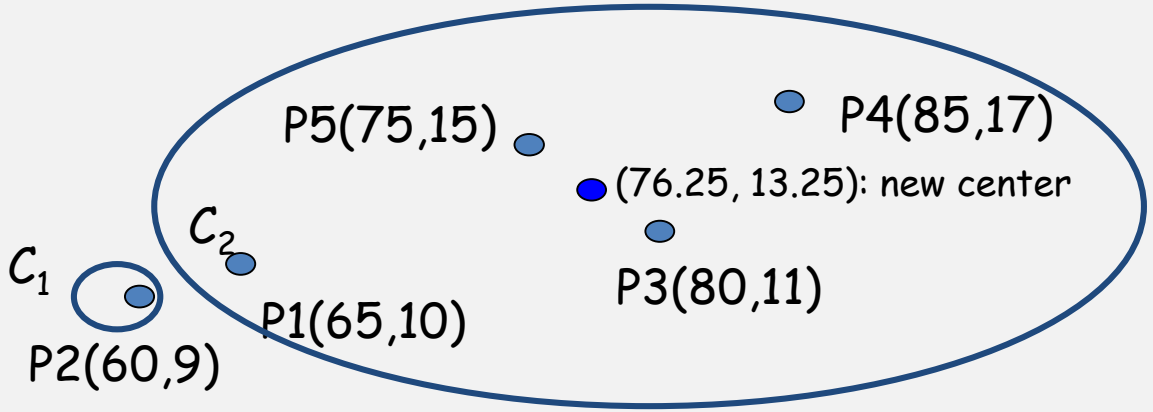
✓Assign other objects to the groups





Step 3

✓Recalculate the new k centers



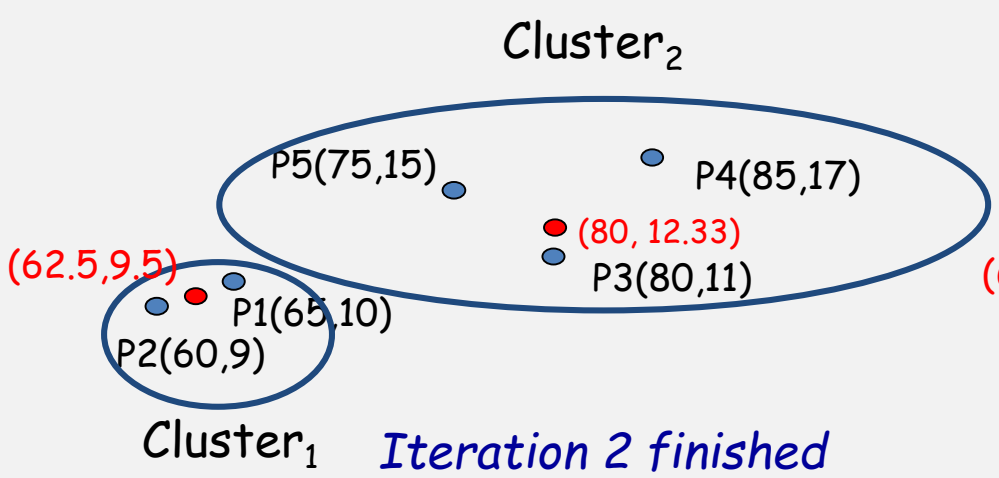
$Center_1$ ➡ $Center_1$
P2(60,9) P2(60,9)

Iteration 1 finished

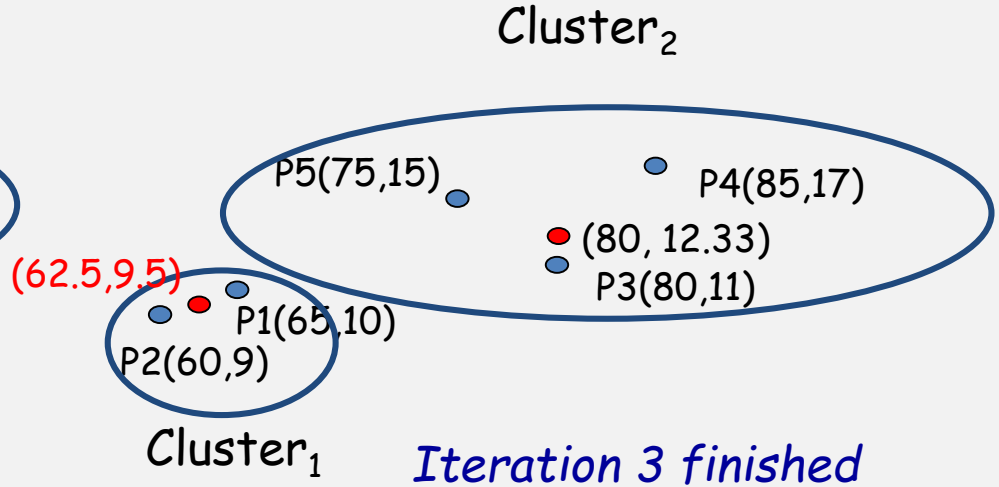
$Center_2$ ➡ New $Center_2$
P1(65,10) ((65+75+80+85)/4, (10+15+11+17)/4)
= (76.25, 13.25)

Step 4

✓Repeat Steps 2 and 3



Iteration 2 finished



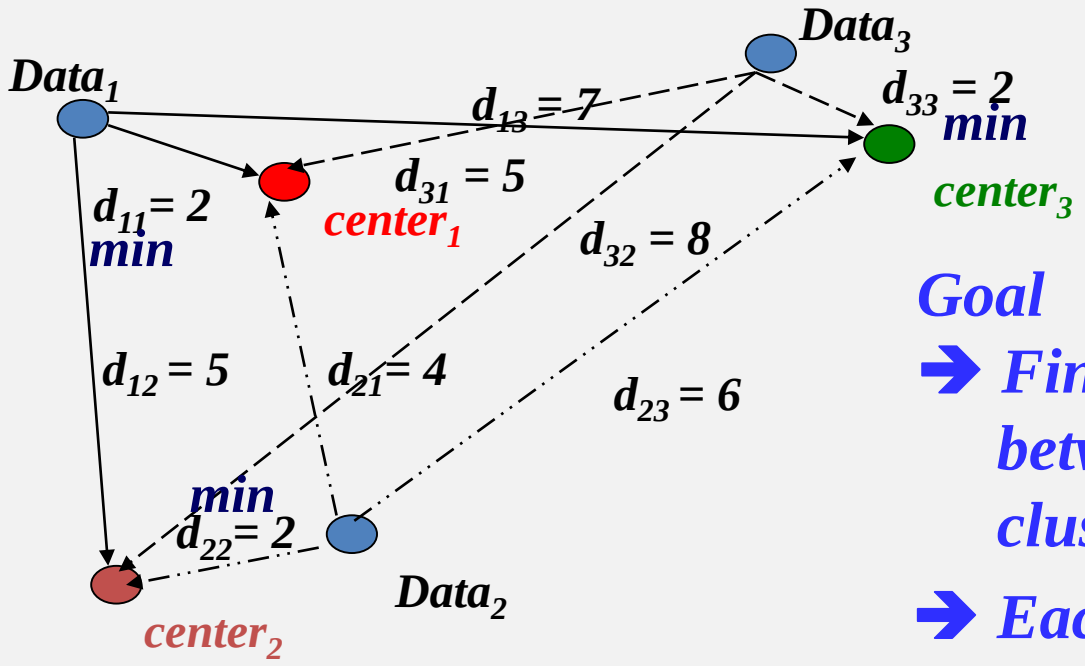
Iteration 3 finished

No change ➔ Stop

- Cluster center
- Instance

Concepts of *k*-Means

✓Number of clusters = 3



Goal

➔ ***Find minimum distance between data points and cluster centers***

➔ ***Each data point belong to only one group***

The *k*-means Algorithm

- ✓Parameter: A number of group k
- ✓Algorithm: 4 Steps
  - ●Step 1: Initialize *k* centers
  - ●Step 2: Assign each object to the group
    - ✓Find the closest center
  - ●Step 3: Recalculate the new *k* centers
  - ●Step 4: Repeat Steps 2 and 3
    - ✓Until the centers no longer move

Discussions

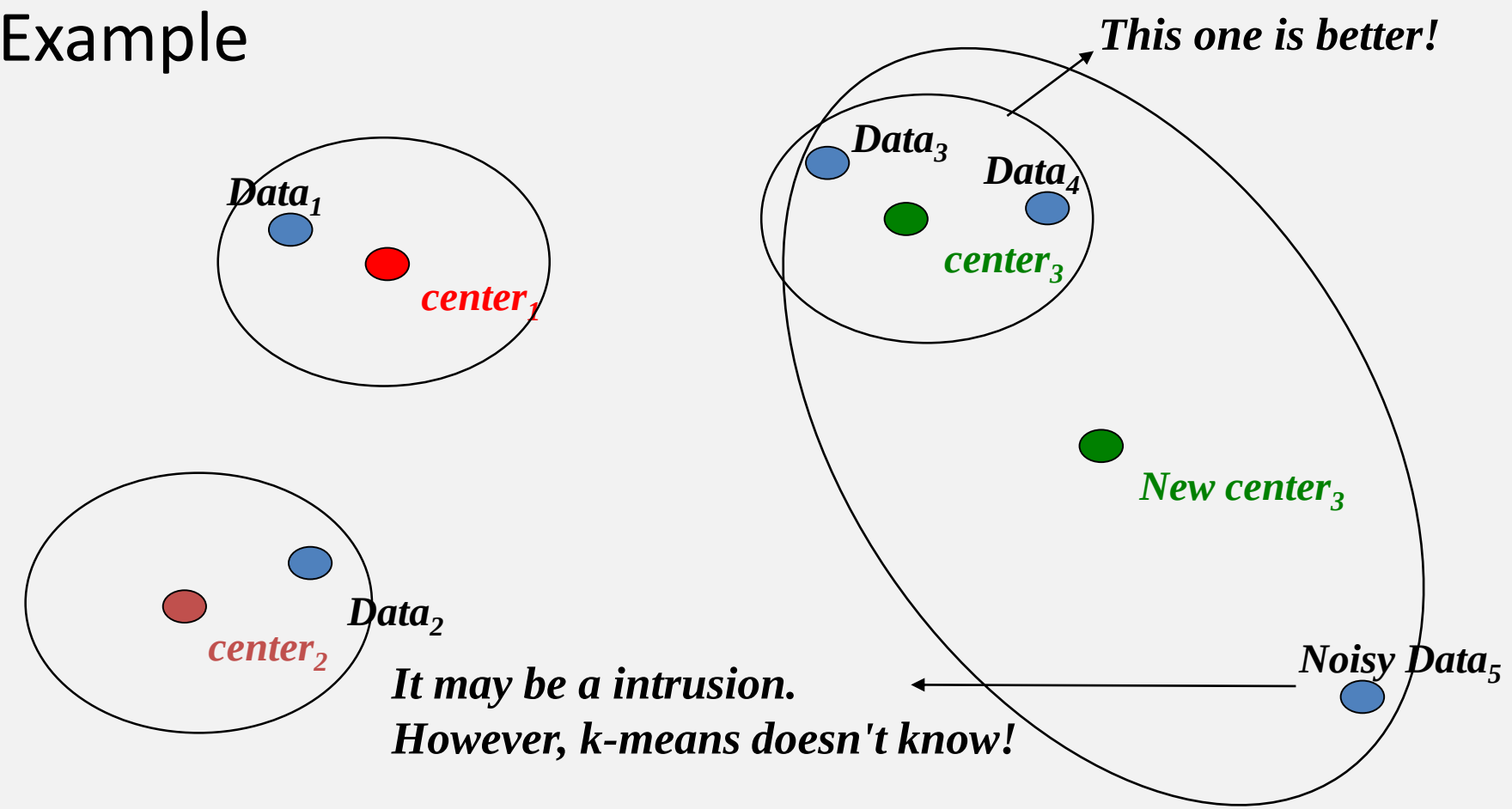
What are the problems of the *k*-means clustering algorithm that you can find?

Problems of *k*-means

- ✓Limitations
  - ●Can not process outliners (noisy data points)
  - ●Cluster convergence
  - ●Number of clusters dependency
  - ●Cluster degeneracy

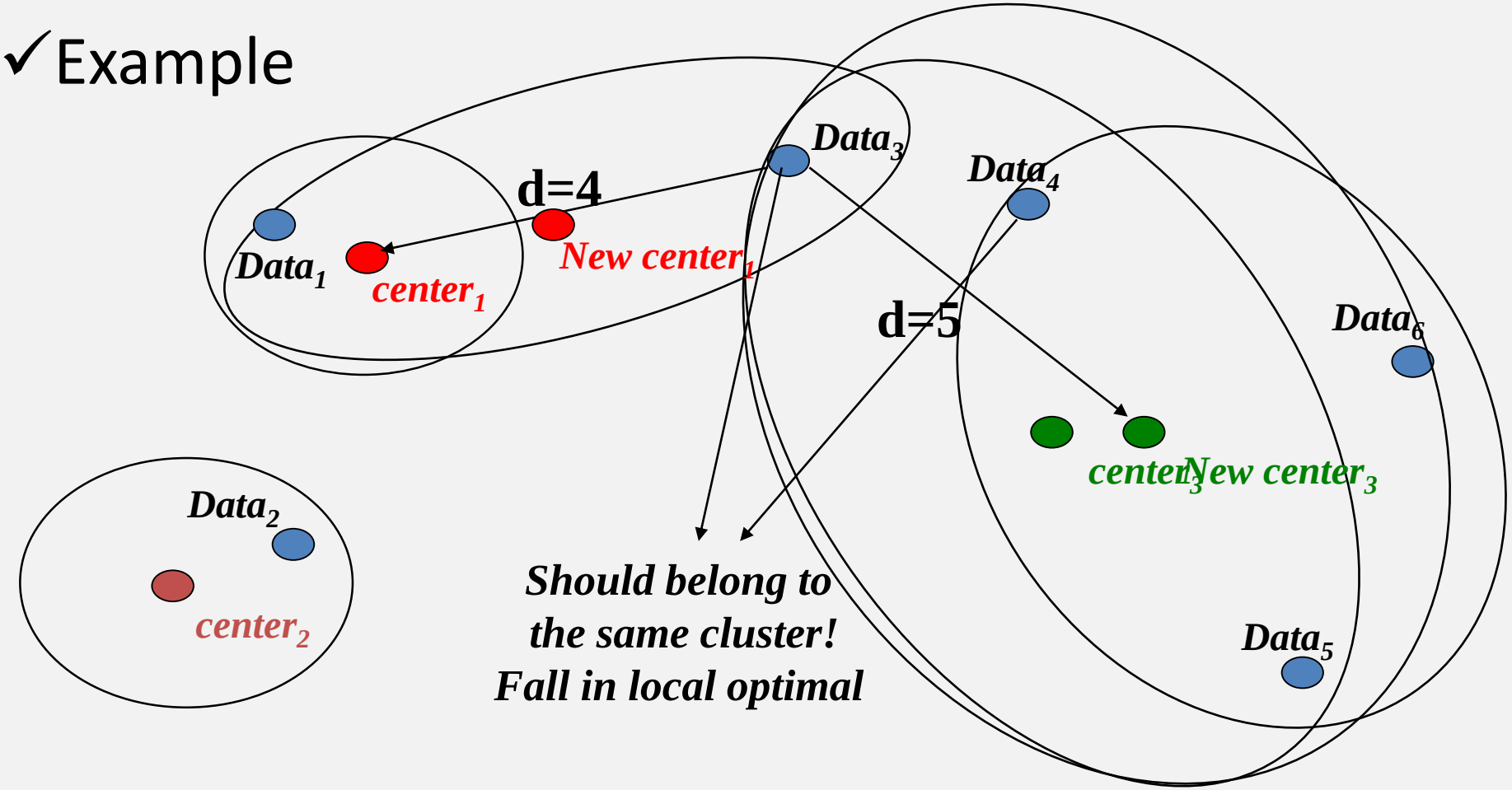
1. Noisy Data Points or Outliners

✓Example



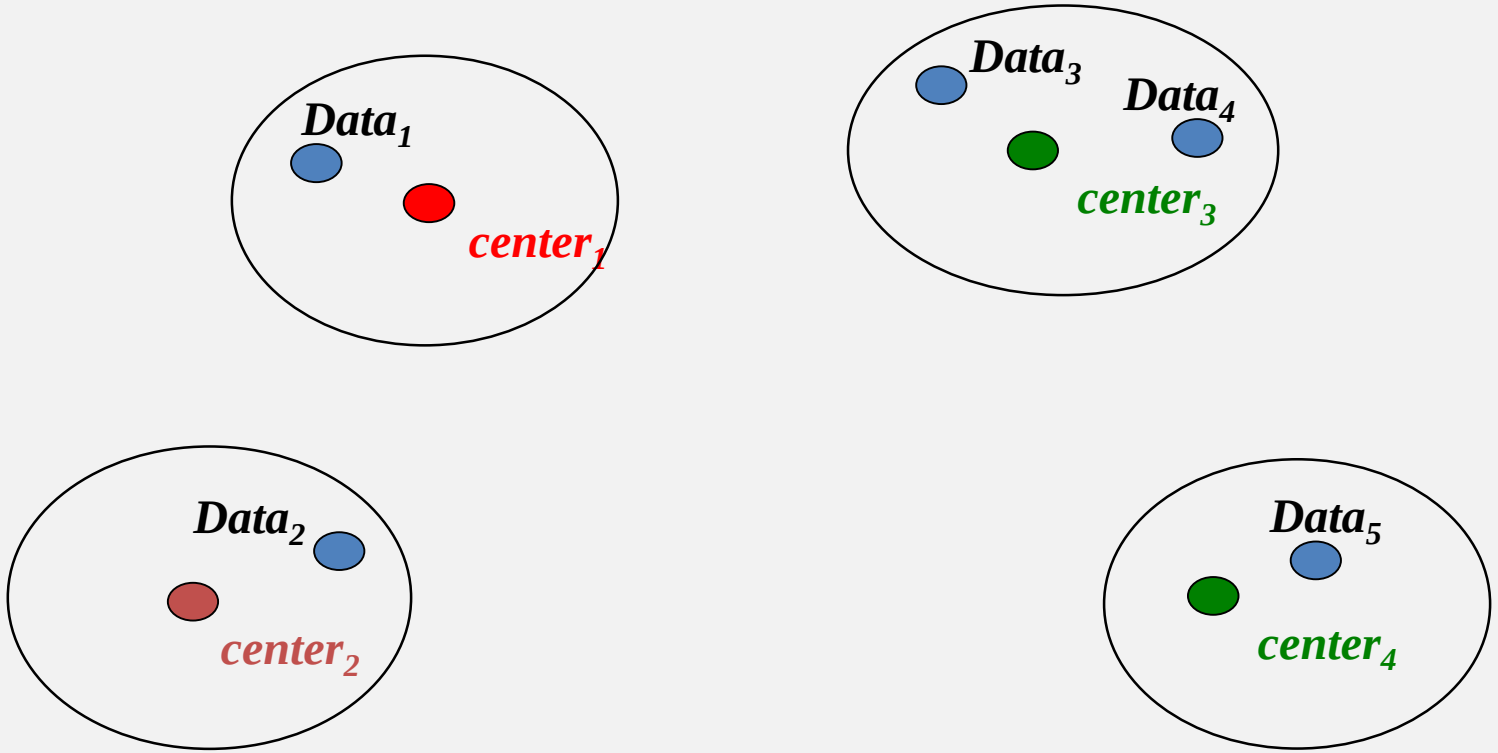
2. Cluster Convergence

✓Example



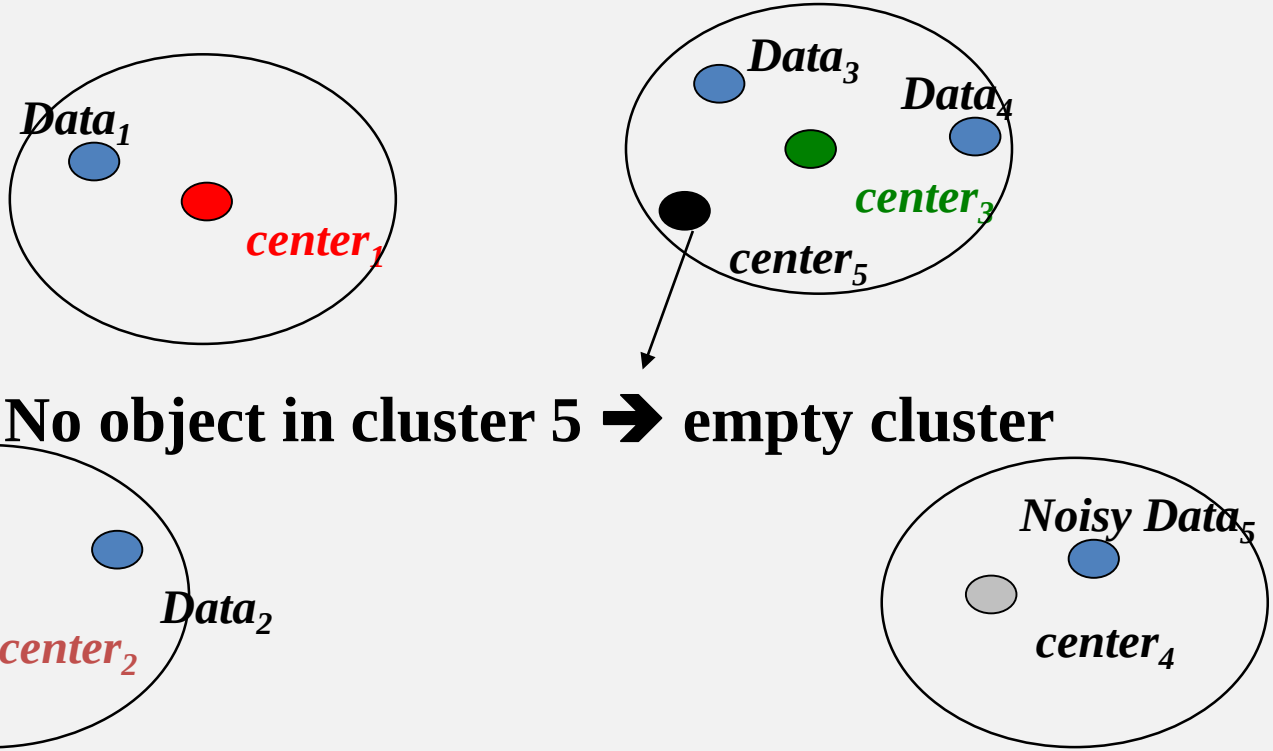
3. Number of Clusters Dependency

✓Example ➔ k = 4 is the best



4. Cluster Degeneracy

✓Example ➔ if $k = 5$



Discussions

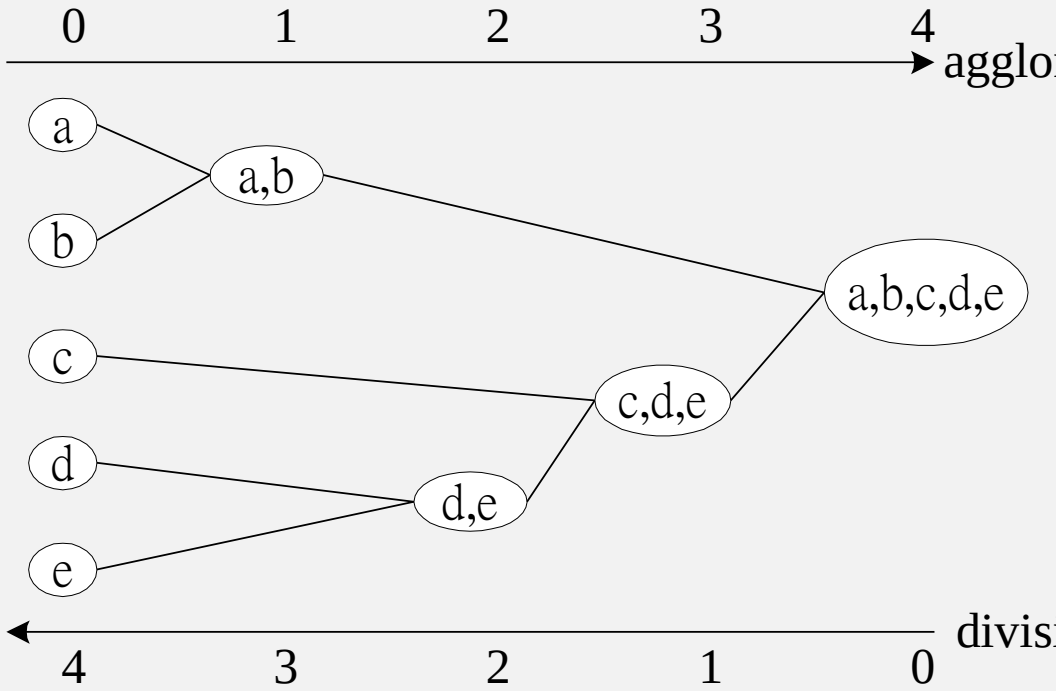
- ✓4 limitations of *k*-means
  - Can not process outliners (noisy data points)
  - Cluster convergence
  - Number of clusters dependency
  - Cluster degeneracy

How to solve the four mentioned problems?

Solution I : Hierarchical Clustering

✓Hierarchical Clustering Techniques

- Distinction between agglomerative and divisive techniques



Button-up Strategy

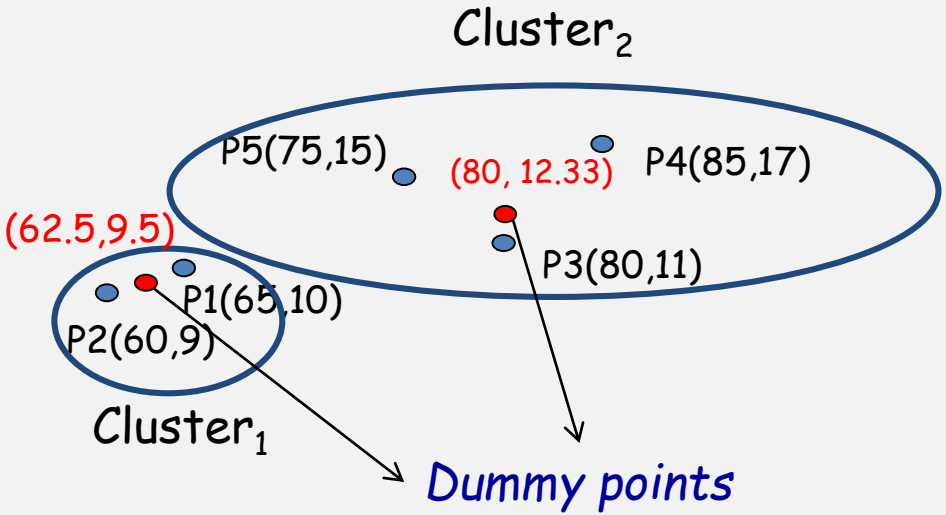
Top-down Strategy

Solution II : *k*-Medoids Clustering

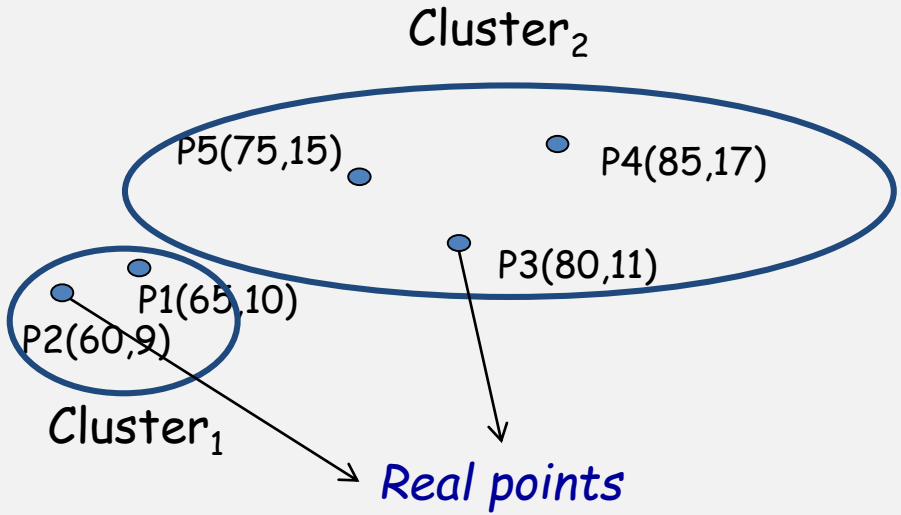
✓Difference between *k*-means and *k*-medoids

- How to form next center points

k-means clustering

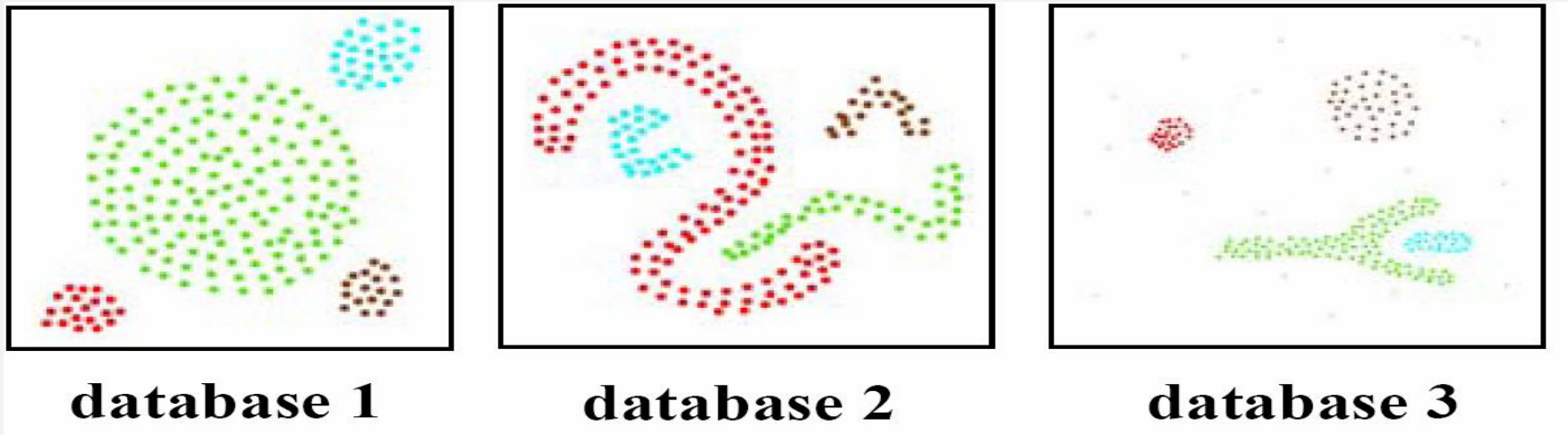


k-medoids clustering



Discussions: What if Dataset Like This..

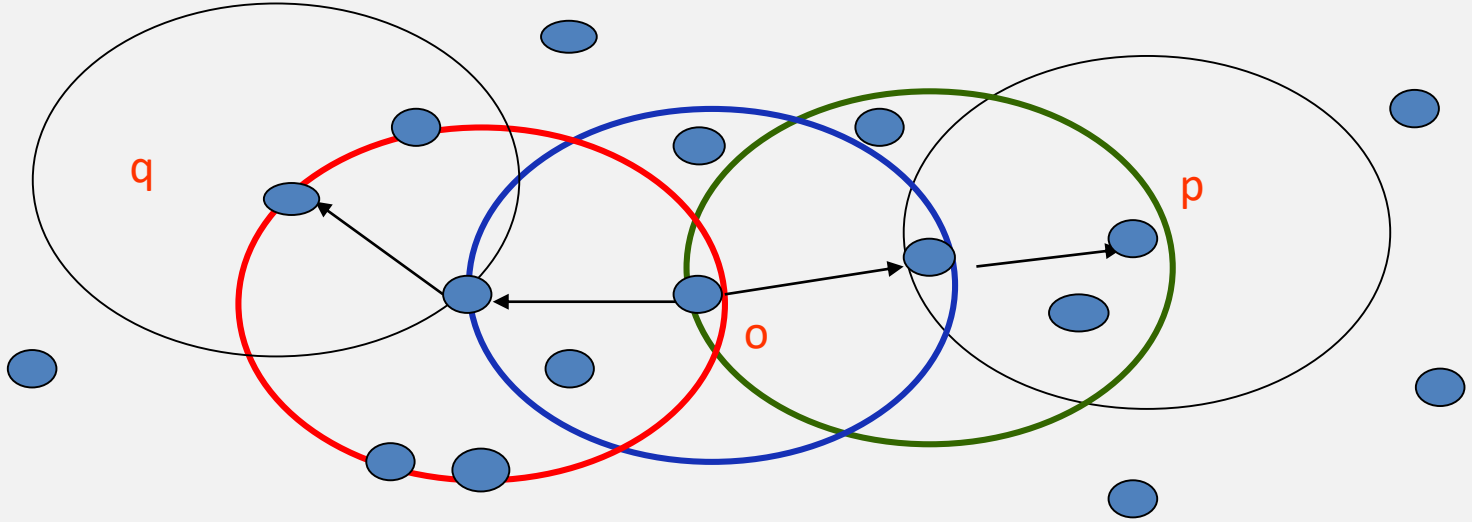
✓ Instances with the same color mean the same cluster



How to find correct clustering results?

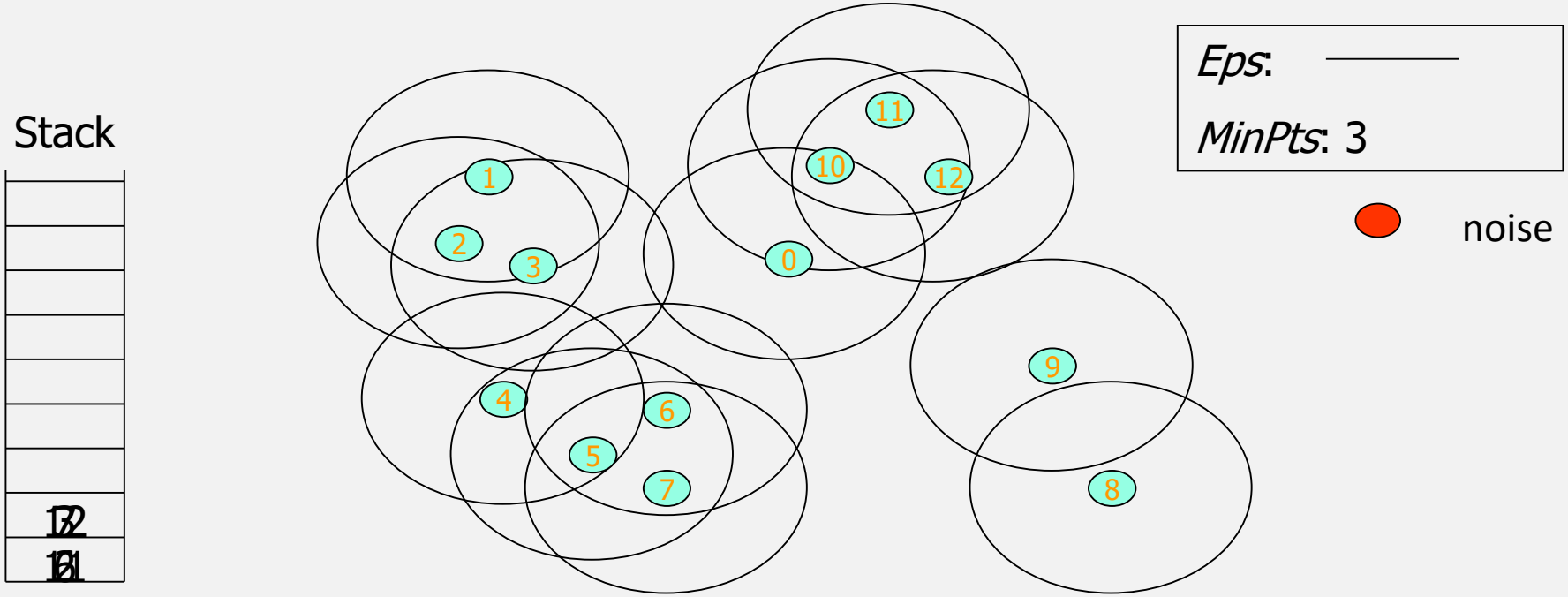
Solution I: DBSCAN Clustering

✓ Main Concept



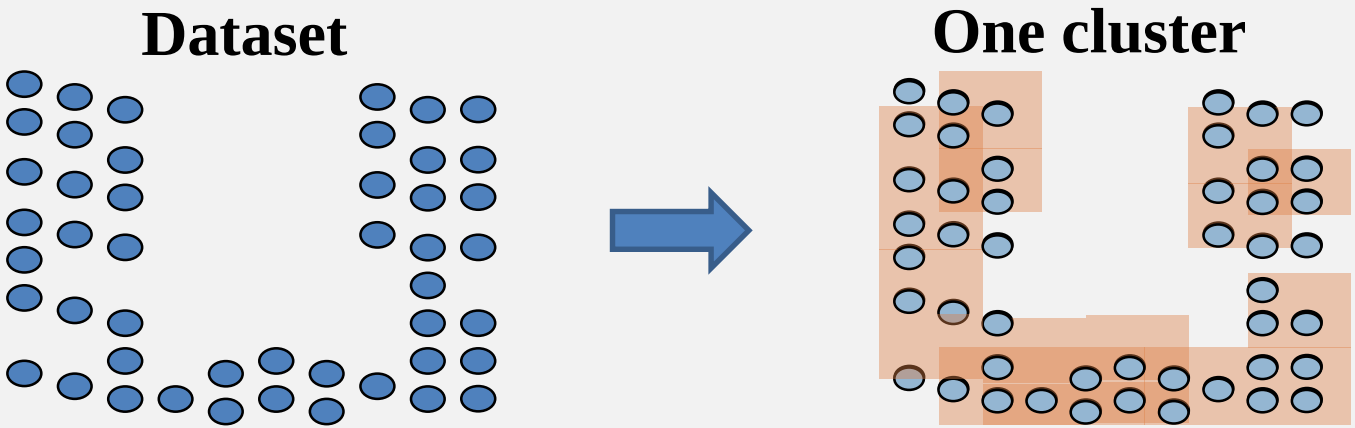
MinPts: 4

DBSCAN Demo



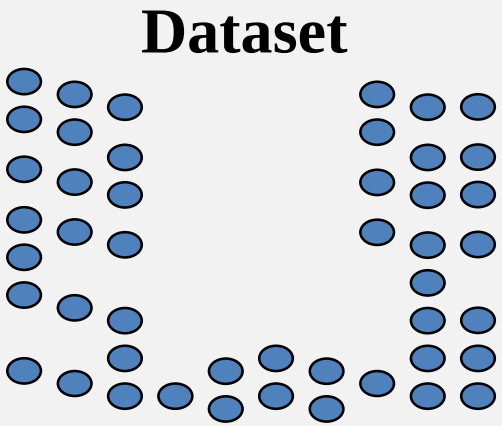
Solution II: Grid-based Clustering

✓ Concept



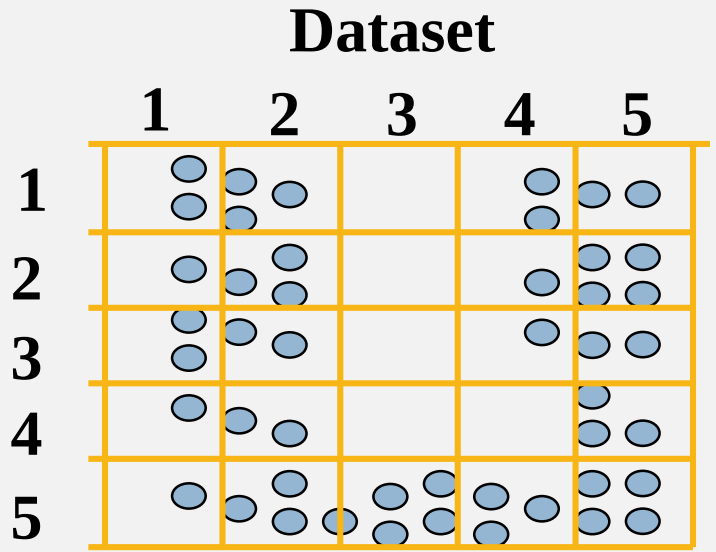
An Example (Grid-based Clustering)

- ✓ Two parameters
  - Size of the predefined grids
  - Threshold of the significant cells



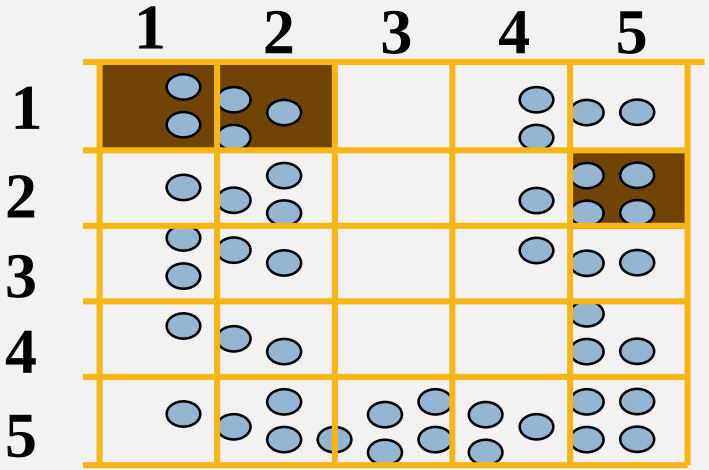
Size of The Predefined Grids

- ✓ Example
  - Predefined Grids ➔ 5*5



Threshold of The Significant Cells

- ✓ Example
  - ● Threshold = 3



Cell(1, 2) = 3 ➔ Significant Cell

Cell(2, 5) = 4 ➔ Significant Cell

Cell(1, 1) = 1 ➔ Not a Significant Cell

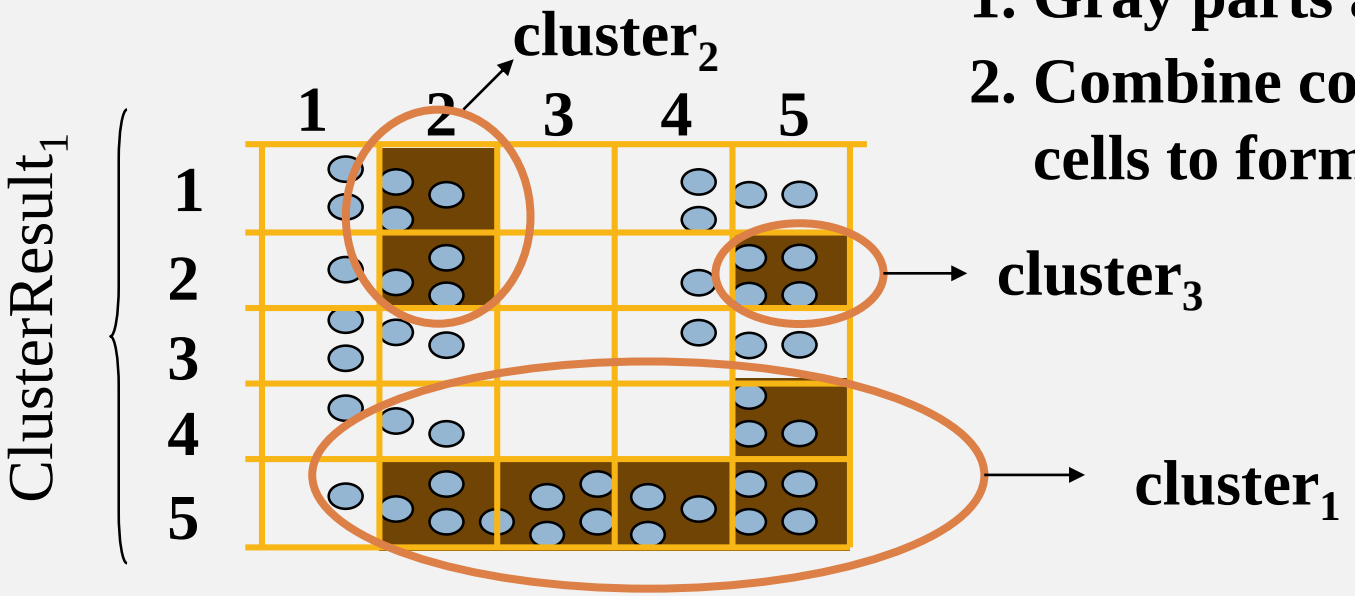
ADCC Approach

- ✓Adaptable Deflect and Conquer Clustering (ADCC)
  - ●1. Call SGDC (first time) ➔ Generate $ClusterResult_1$
  - ●2. Call SGDC (second time) ➔ Generate $ClusterResult_2$
  - ●3. Call DedflectAndConquer procedure
    - ✓Merge $ClusterResult_1$ & $ClusterResult_2$

Call SGDC (First Time)

✓Example

- The Predefined Grids ➔ 5*5
- Threshold = 3

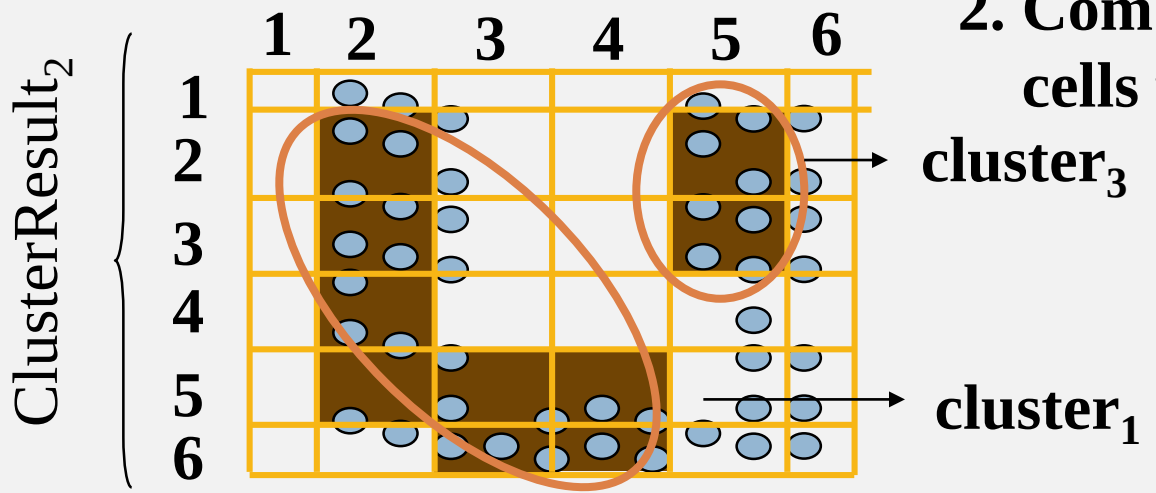


1. Gray parts are significant cell
2. Combine connected significant cells to form cluster

Call SGDC (Second Time)

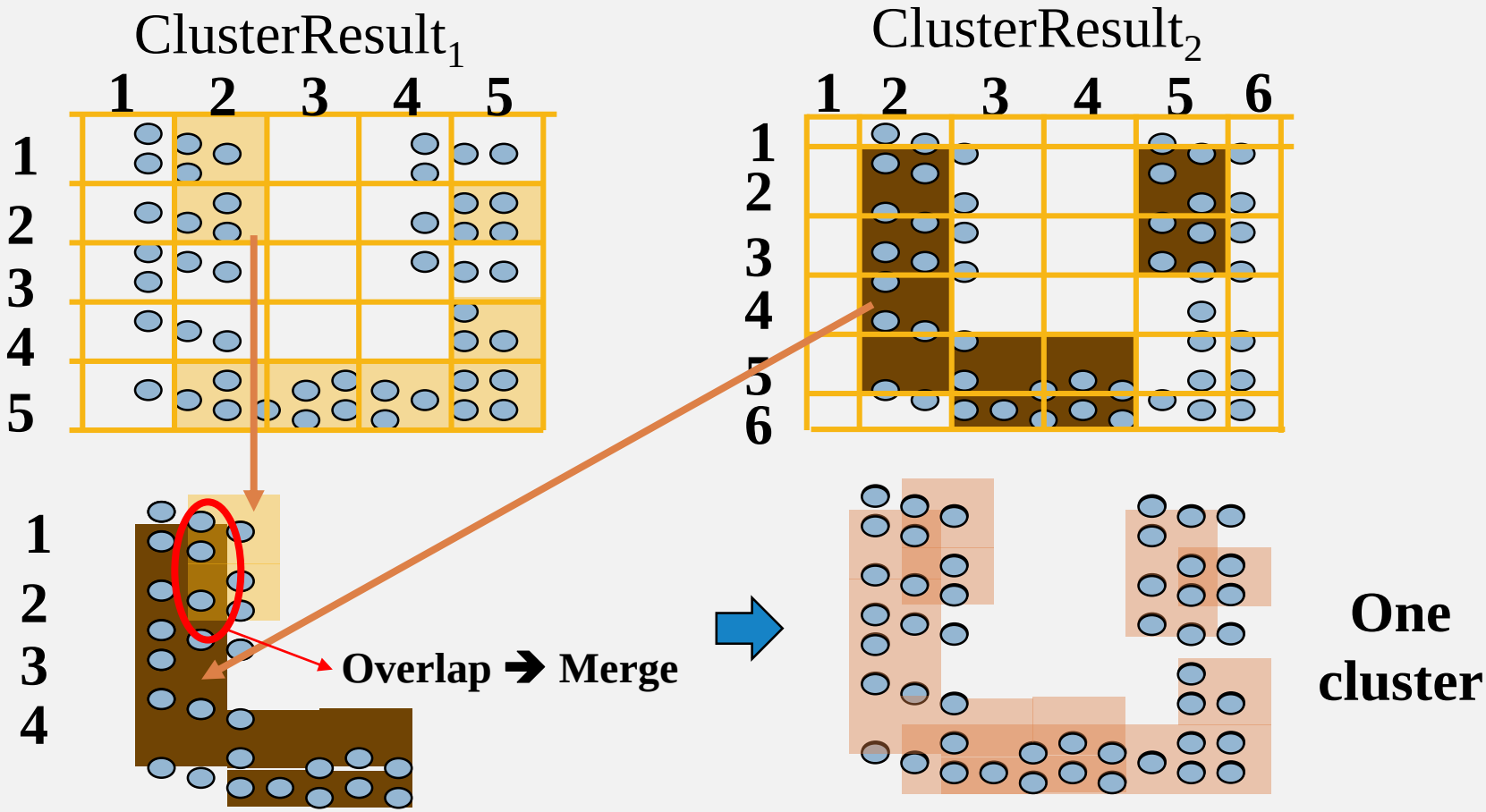
✓Example

- ●Shift half cell to form new grids ➔ 6*6
- ●Threshold = 3



1. **Gray parts are significant cell**
2. **Combine connected significant cells to form cluster**

Merging Clustering Results



Types of Clustering Methods

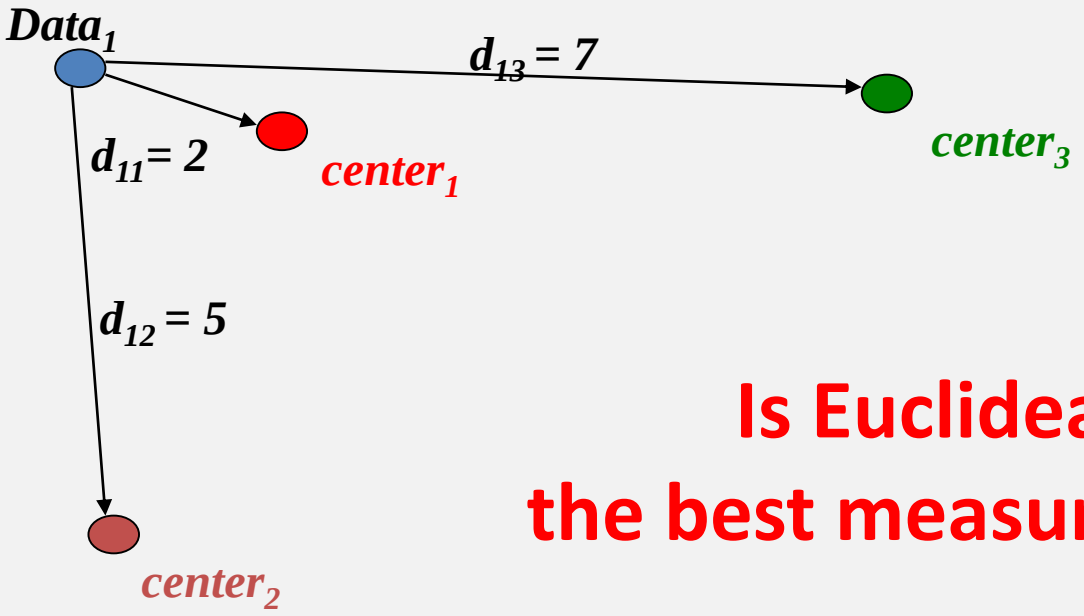
- ✓ Partitioning
  - K-Means, K-Medoids, PAM, CLARA, CLARANS, CAST, …
- ✓ Hierarchical
  - HAC, BIRCH, CURE, ROCK, CHAMELEON, …
- ✓ Density-based
  - DBSCAN, OPTICS, CLIQUE, WaveCluster, …
- ✓ Grid-based
  - ADCC, STING, CLIQUE, WaveCluster, …

Discussions

Could you find any possible ways to get a better clustering result?

Appropriate Similarity Measurement?

✓ The *k*-means clustering algorithm



Is Euclidean Distance the best measurement function?

How You Think About the Similarity?



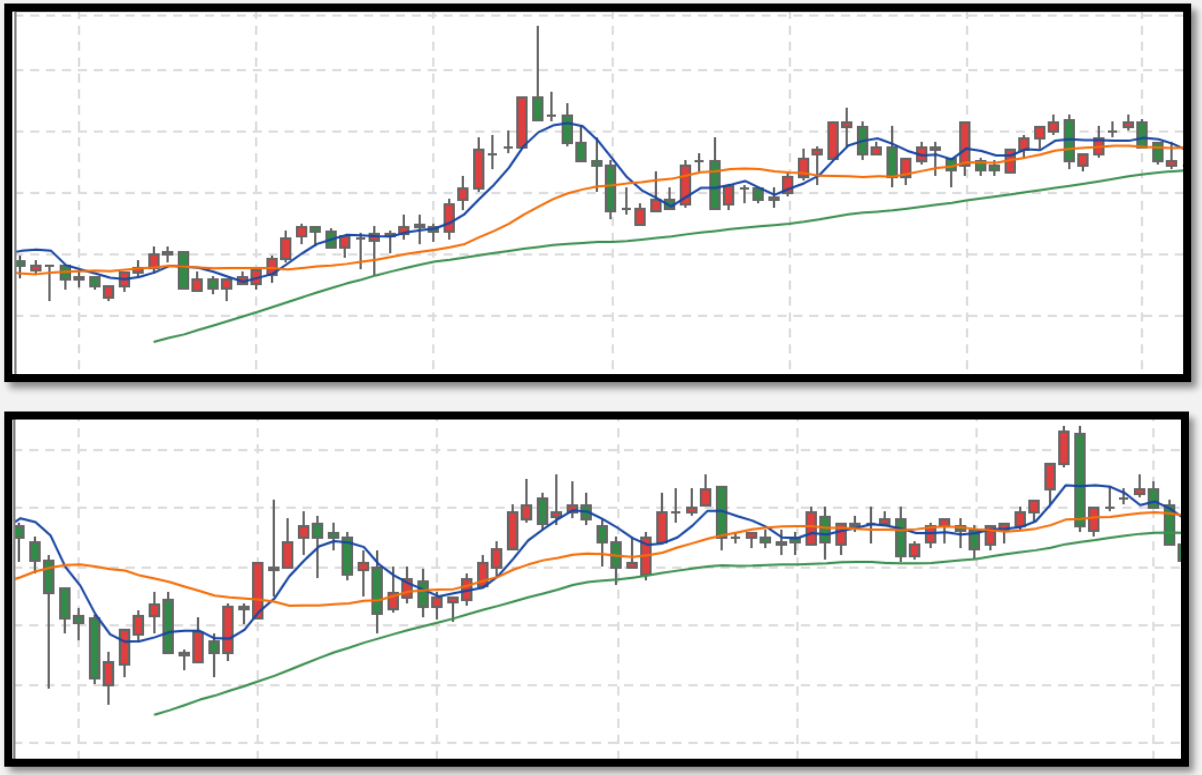

Source: https://www.twgreatdaily.com/cat74/node1999089

People know they are different photos.

However, computer may identify them as similar photos.

Why?

How You Think About the Similarity (Cont.)?



How you think about the similarity of the two stock price series?

Similarity Measurements

- ✓Types of similarity measurements
  - Distance measurements
  - Correlation coefficients
  - Association coefficients
  - Probabilistic similarity coefficients

Correlation Coefficients

- ✓Pearson correlation coefficient (1892)
  - Most popular correlation coefficient
- ✓Correlation between
  - $X=\{X_1, X_2, \ldots, X_n\}$ and $Y=\{Y_1, Y_2, \ldots, Y_n\}$:

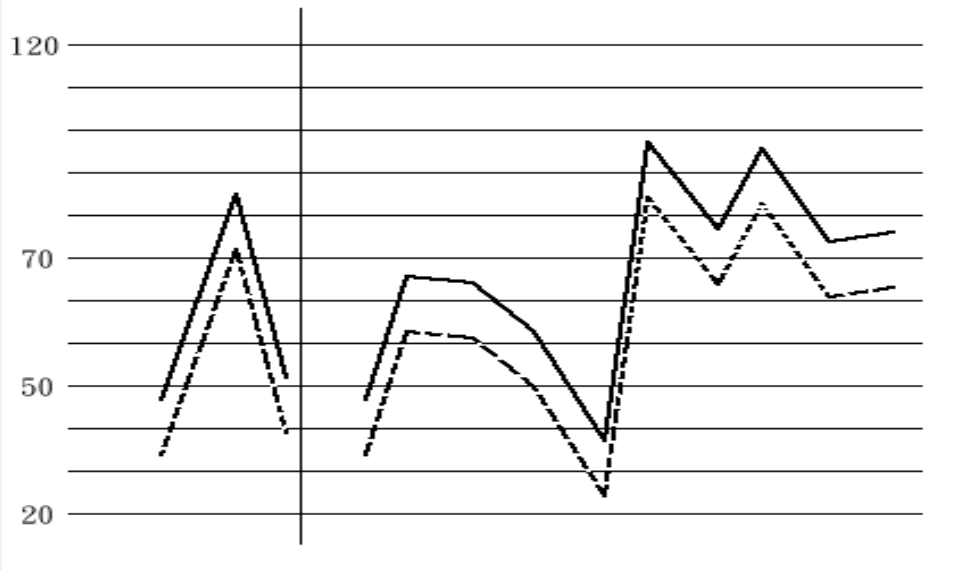
where $$r = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{X_k - \overline{X}}{\sigma_X}\right)\left(\frac{Y_k - \overline{Y}}{\sigma_Y}\right)$$

$$\sigma_G = \sqrt{\sum_{k=1}^{n}\frac{\left(G_k - \overline{G}\right)^2}{n}}$$

Correlation Coefficients (Cont.)

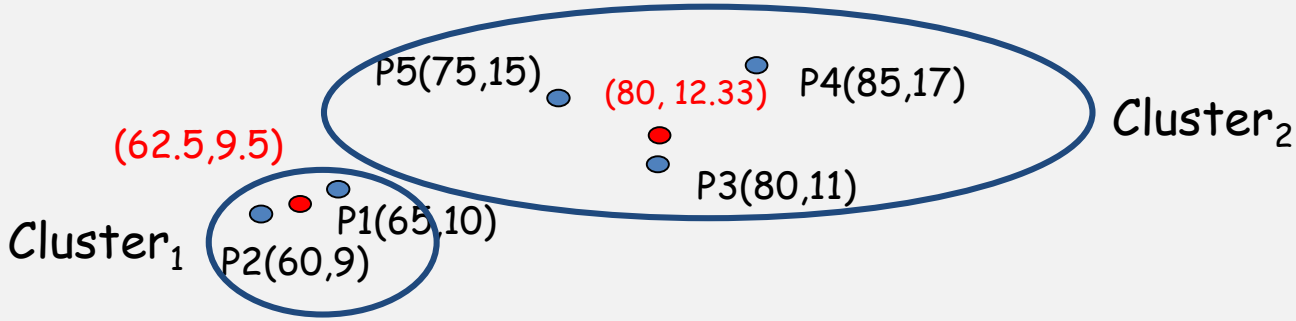
- ✓ Capture similarity of the “**shapes**” of two expression profiles
- ✓ Ignore differences between their magnitudes

r = 1.0



Discussion 1

k-means clustering

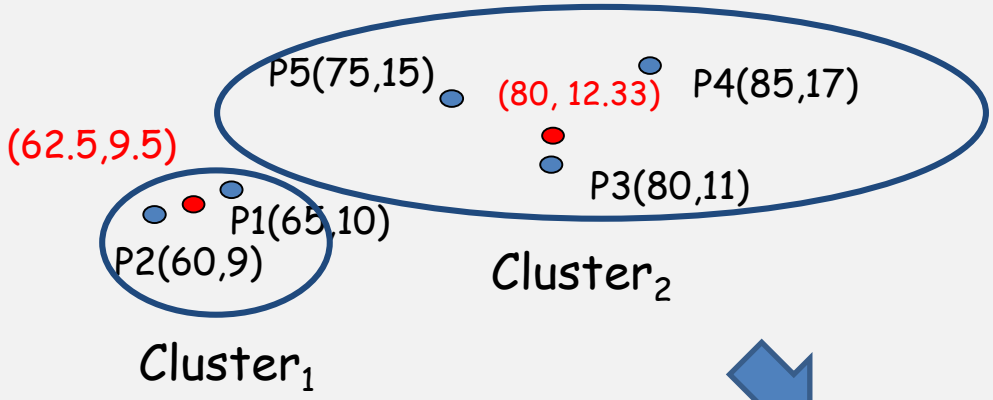


In *k*-means clustering, each data point should belong to a group.
How do you think about that?

Discussion 2

✓Continue previous example

- Number of clusters $k = 2$



Do we have any approach to get clusters without the *k* value?

Implementation In Python

- ✓ I – Using Height & Weight for clustering
- ✓ II – Using make_blobs to generate dataset for clustering