集群/群集/分群/聚類
Clustering

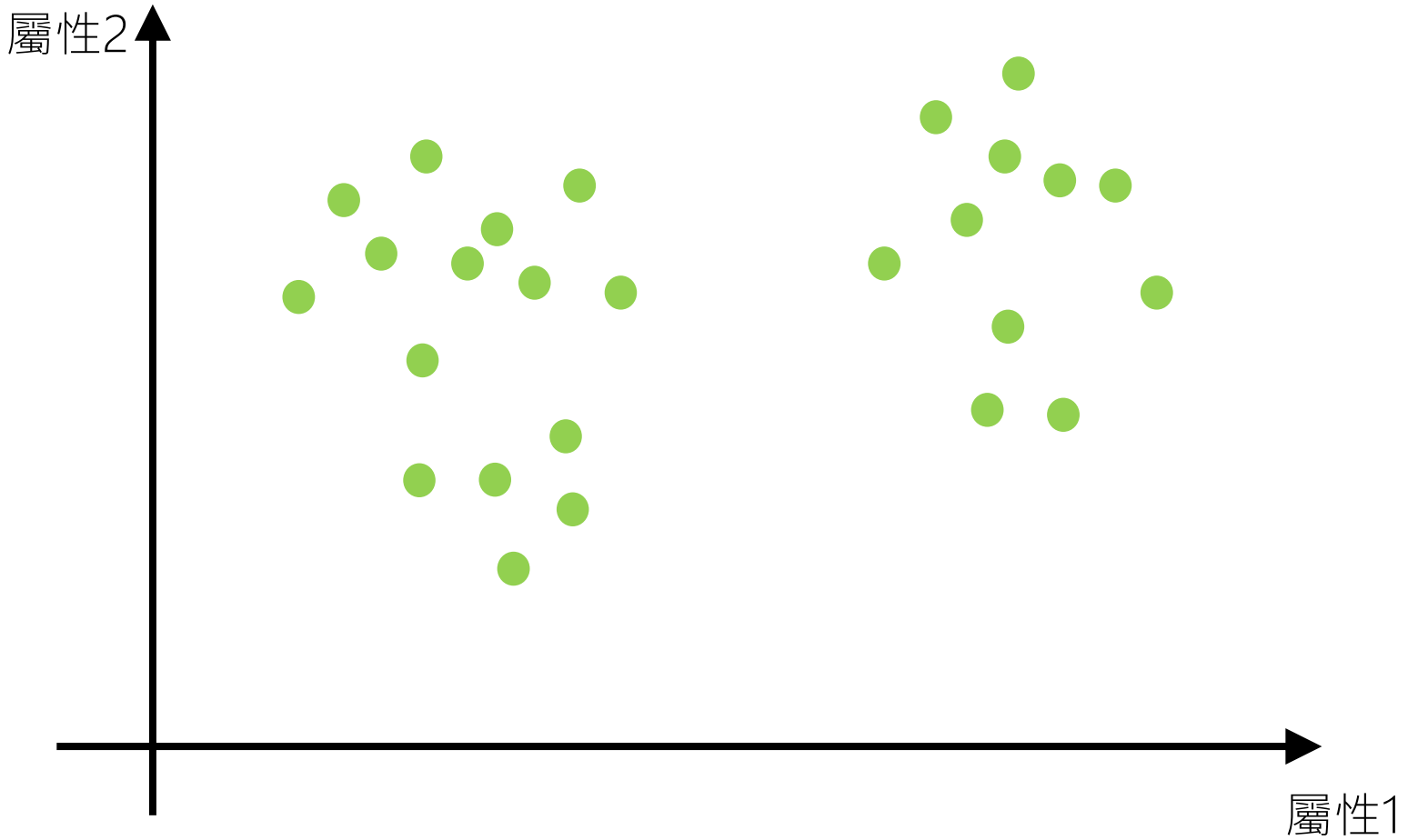
課程目標

- 瞭解集群的概念
- 瞭解衡量相似度（距離）的方式
- 學習簡單的集群演算法
  - 分割基礎(Partitional)
  - 階層基礎(Hierarchical)

The Concept of the Clustering

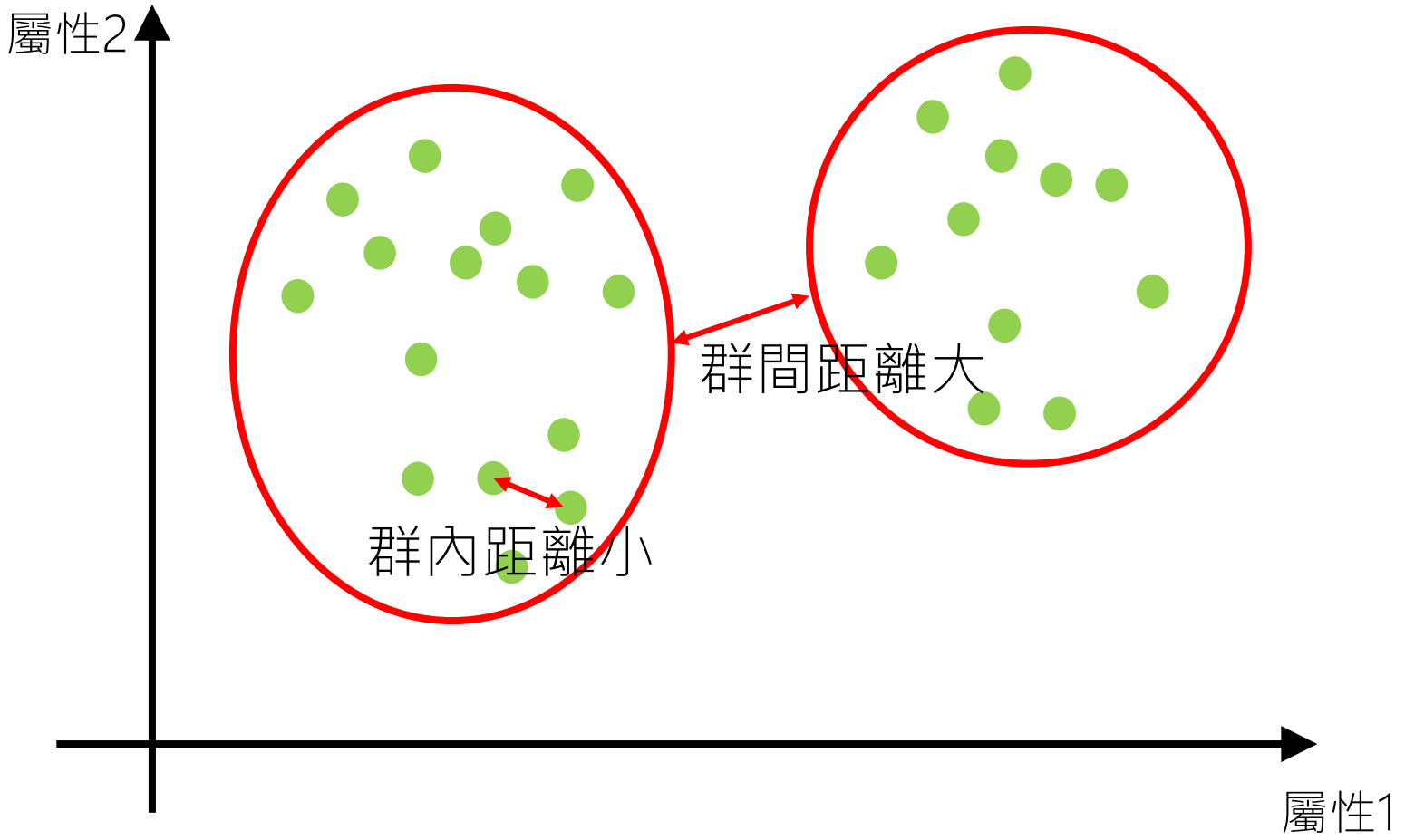
集群的概念

集群是什麼？

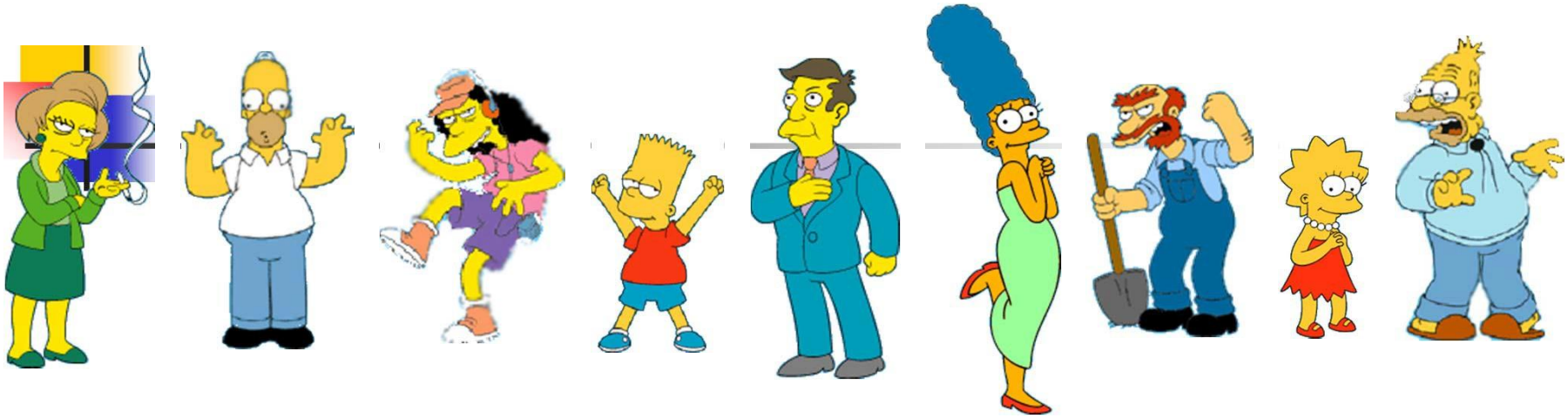


請問圖上的物件可以分成幾群？

集群是什麼？

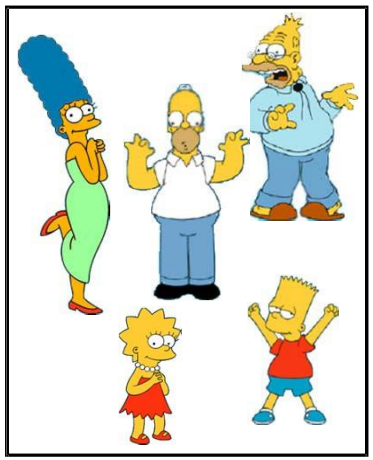


集群是什麼？



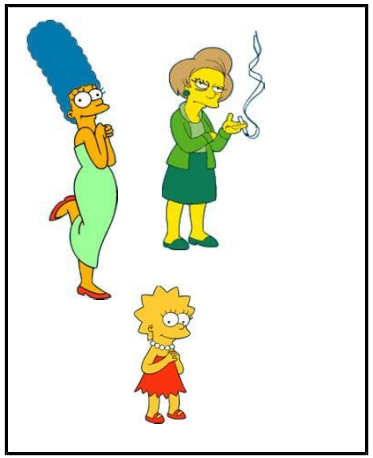
試著分群這些人物

集群是什麼？



Simpson's Family

School Employees



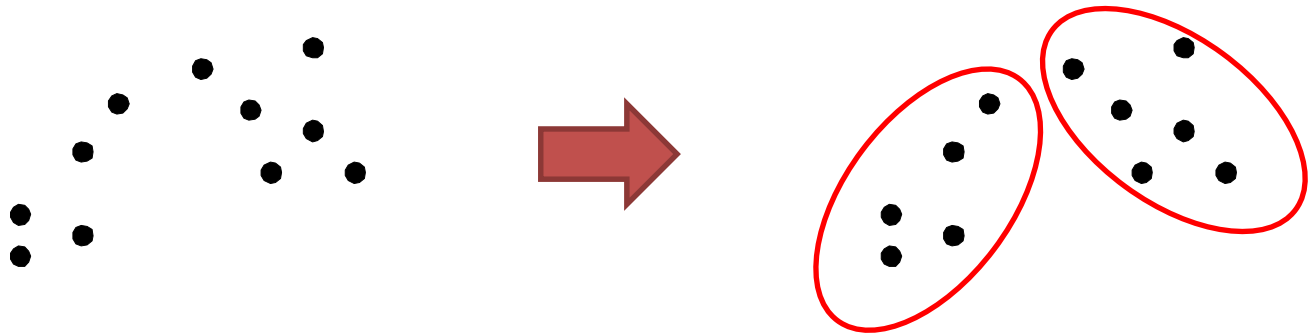
Females

Males

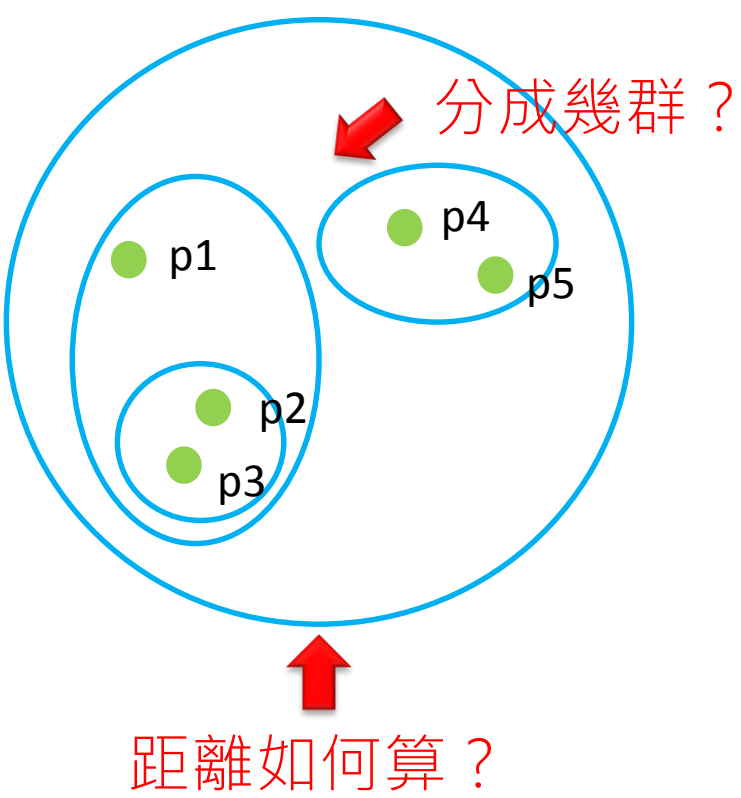
分群是主觀的

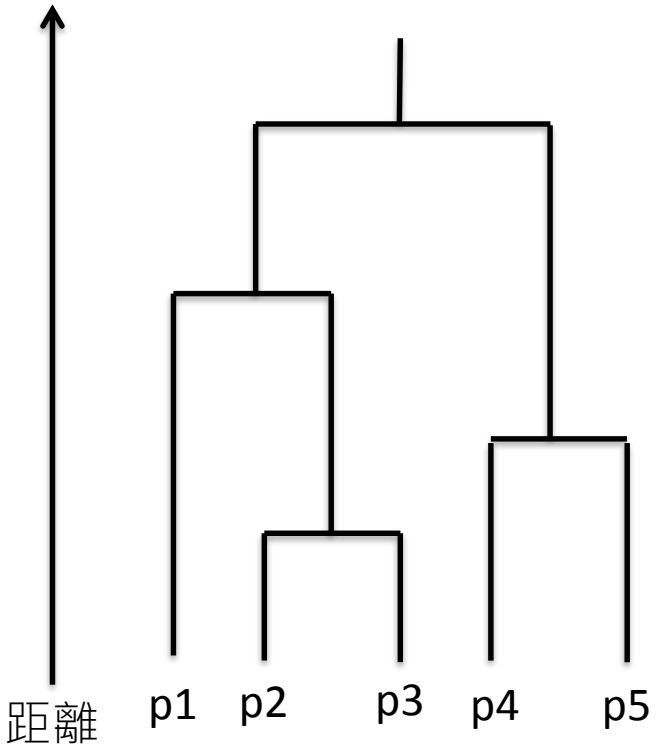
集群的定義

- 將未知類別的物件分成幾群，使得群內的物件相似性高，群間的物件相似性低

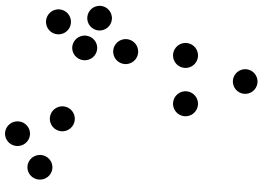


如何集群？

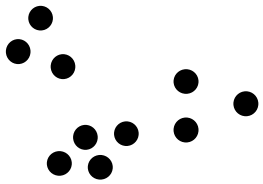


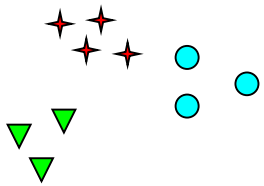


分成幾群？

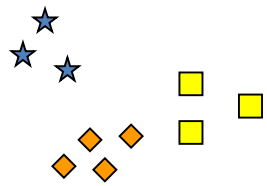


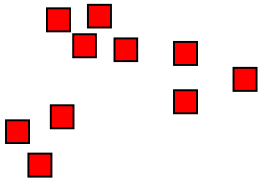
一群



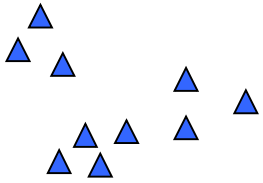


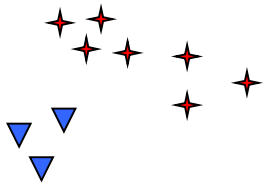
六群



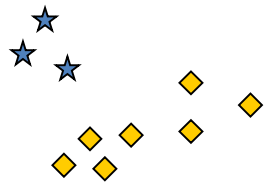


兩群





四群



自己決定！

The Measurement of the Similarity / Distance

相似度（距離）的衡量

相似度(Similarity)

- 定義 x 與 y 的相似程度為 $s(x, y)$
  - $s(x, y)$ 值越大表示 x 與 y 越相似
  - 對稱性

$$s(x, y) = s(y, x)$$

  - 標準化

$$0 \leq s(x, y) \leq 1$$

距離 (Distance)

- 定義 x 與 y 的距離為 $d(x, y)$
  - $d(x, y)$ 值越大表示 x 與 y 越不相似
  - 非負性

$$d(x, y) \geq 0$$

  - 對稱性

$$d(x, y) = d(y, x)$$

  - 三角不等式

$$d(x, y) + d(y, z) \geq d(x, z)$$

相似度 vs. 距離

$s(x, y)$越大，$d(x, y)$越小

The Distance Between Points

點與點的距離

距離衡量

- Euclidean distance（歐式距離）

$$d_2(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

距離衡量

- City Block Distance

$$d_1(x, y) = \sum_{k=1}^{n} |x_k - y_k|$$

距離衡量

- Minkowski Distance

$$d_p(x,y) = \sqrt[p]{\sum_{k=1}^{n}(x_k - y_k)^p}$$

距離衡量

- Cosine-Correlation

$$s(x,y) = \frac{\sum_{k=1}^{n} x_k \cdot y_k}{\sqrt{\sum_{k=1}^{n} x_k^2 \cdot \sum_{k=1}^{n} y_k^2}}$$

範例

- $x = (1,0,1,0)$、$y = (2,1,-3,-1)$
  - Euclidean Distance

$$d_2(x,y) = \sqrt{1+1+16+1} = \sqrt{19}$$

  - City Block Distance

$$d_1(x,y) = |1+1+4+1| = 7$$

  - Minkowski Distance (p=3)

$$d_2(x,y) = \sqrt[3]{1+1+64+1} = \sqrt[3]{67}$$

  - Cosine-Correlation

$$s(x,y) = \frac{2+0-3+0}{\sqrt{(1+0+1+0)(4+1+9+1)}} = \frac{-1}{\sqrt{30}}$$

The Distance Between Clusters

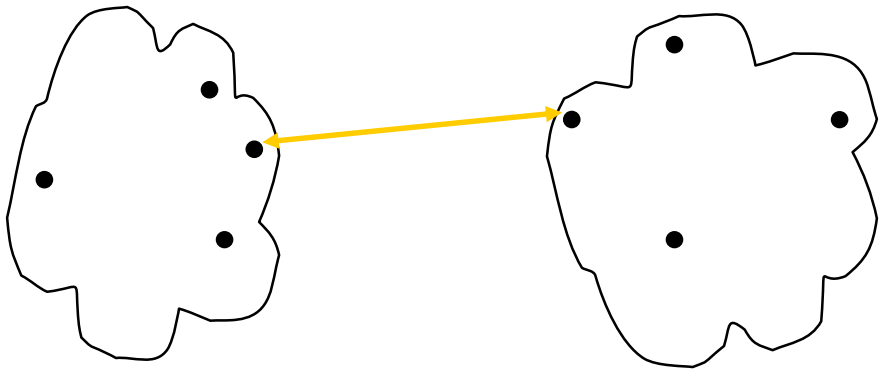
群與群的距離

距離衡量

- 最近距離(Single Link)

$$d(C_i, C_j) = \min d(p_i, p_j)$$

其中$p_i \in C_i$、$p_j \in C_j$

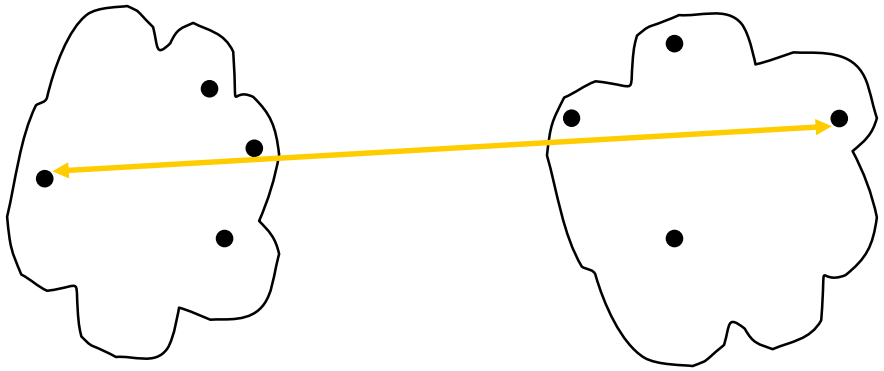


距離衡量

- 最遠距離(Complete Link)

$$d(C_i, C_j) = \max d(p_i, p_j)$$

其中$p_i \in C_i$、$p_j \in C_j$

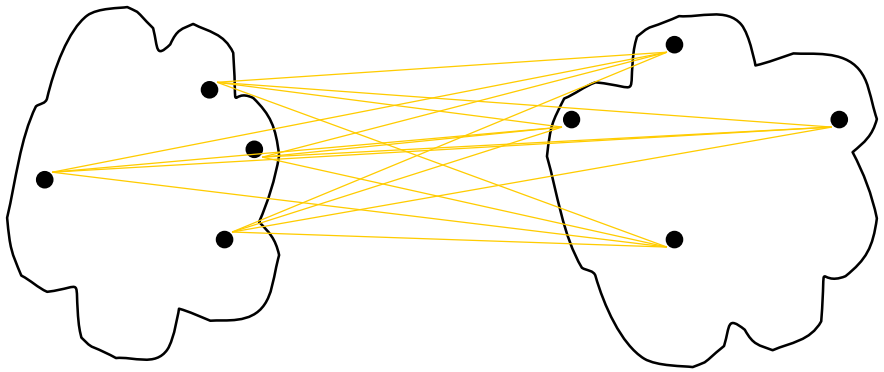


距離衡量

- 平均距離(Average Link)

$$d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{t=1}^{n_j} d(p_k, p_t)$$

其中$p_i \in C_i$、$p_j \in C_j$

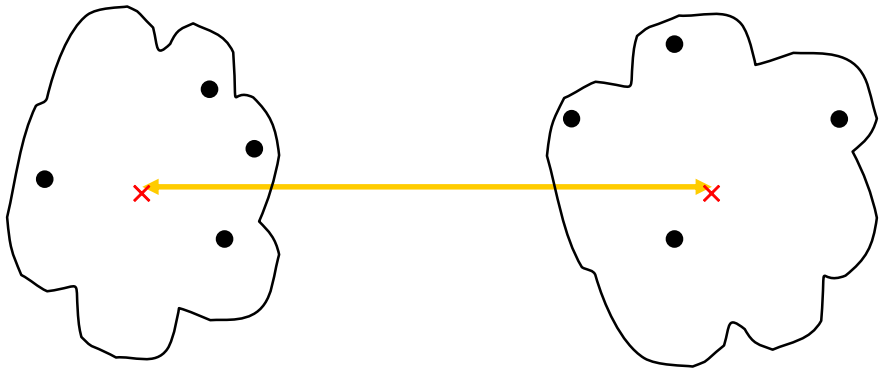


距離衡量

- 中心距離(Centriod)

$$d(C_i, C_j) = d(m_i, m_j)$$

其中m_i是C_i中心、m_j是C_j中心



Type of Clustering Algorithm

集群演算法介紹

集群演算法

- 原型基礎 (Prototype-based)
  - K-Mean
- 階層基礎 (Hierarchy-based)
  - AHC

Clustering Algorithm–k-Mean

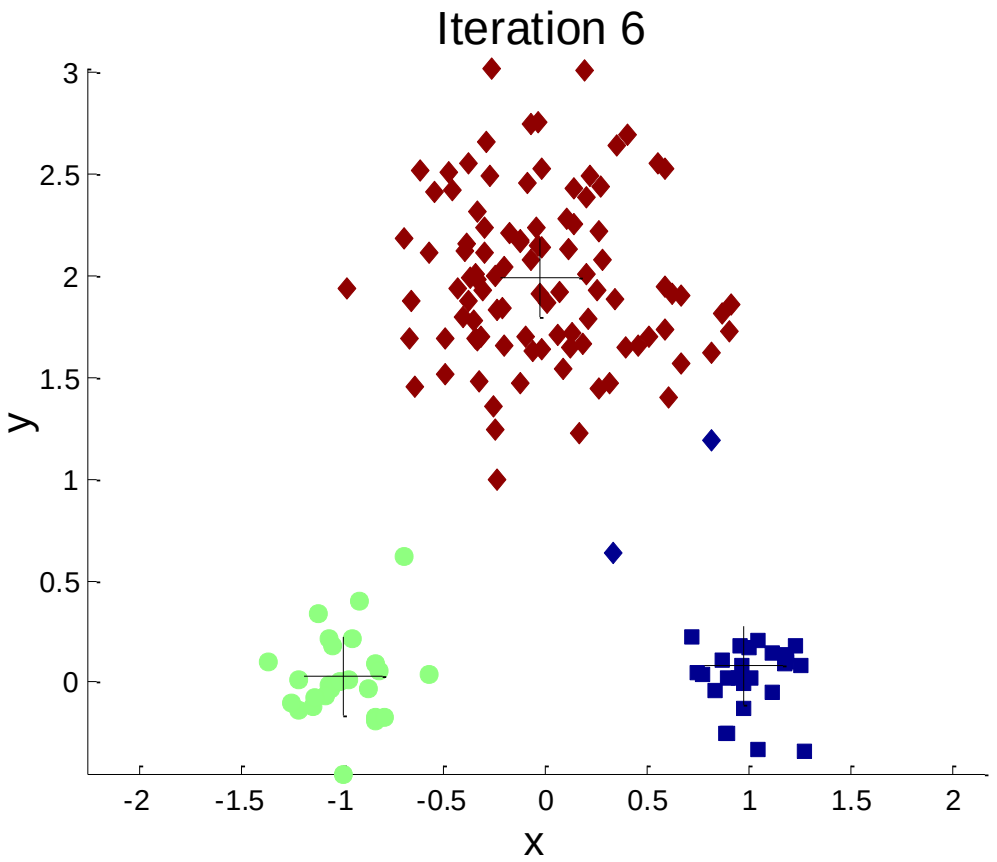
集群演算法 - K-MEAN

K-Mean演算法

- 分割式的群集演算法
- 每一個群集有一個代表中心
- 每一個點被分到最近的代表中心的群集
- 非常簡單

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

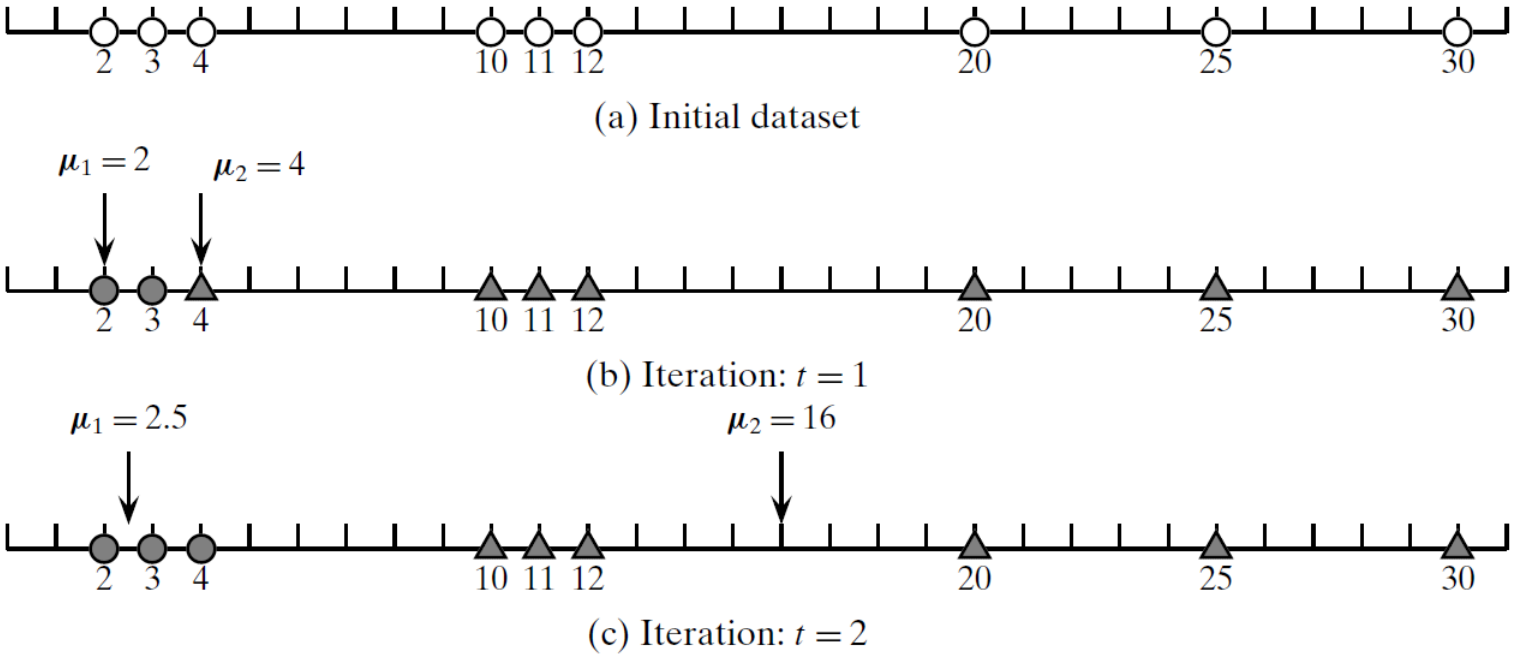
K-Mean演算法



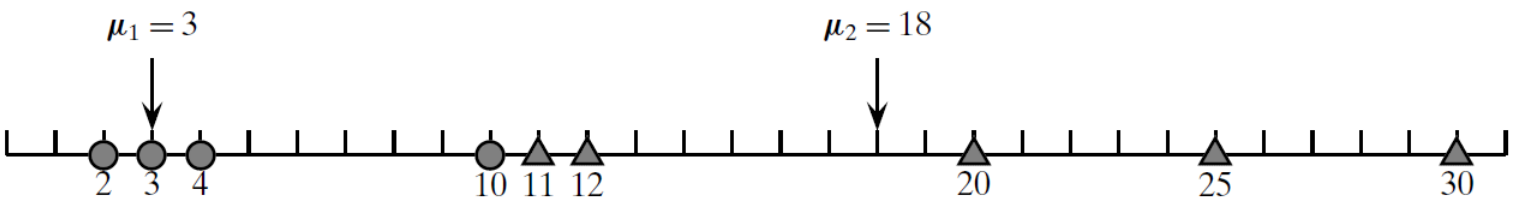
練習一下

- 將2、3、4、10、11、12、20、25、30分成2群
- 初始中心為2、4

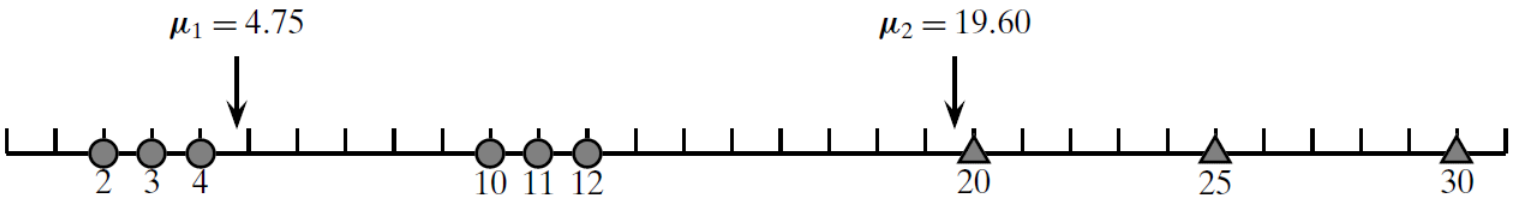
練習一下


2 3 4 10 11 12 20 25 30
(a) Initial dataset
$\mu_1 = 2$
$\mu_2 = 4$
2 3 4 10 11 12 20 25 30
(b) Iteration: $t = 1$
$\mu_1 = 2.5$
$\mu_2 = 16$
2 3 4 10 11 12 20 25 30
(c) Iteration: $t = 2$

練習一下



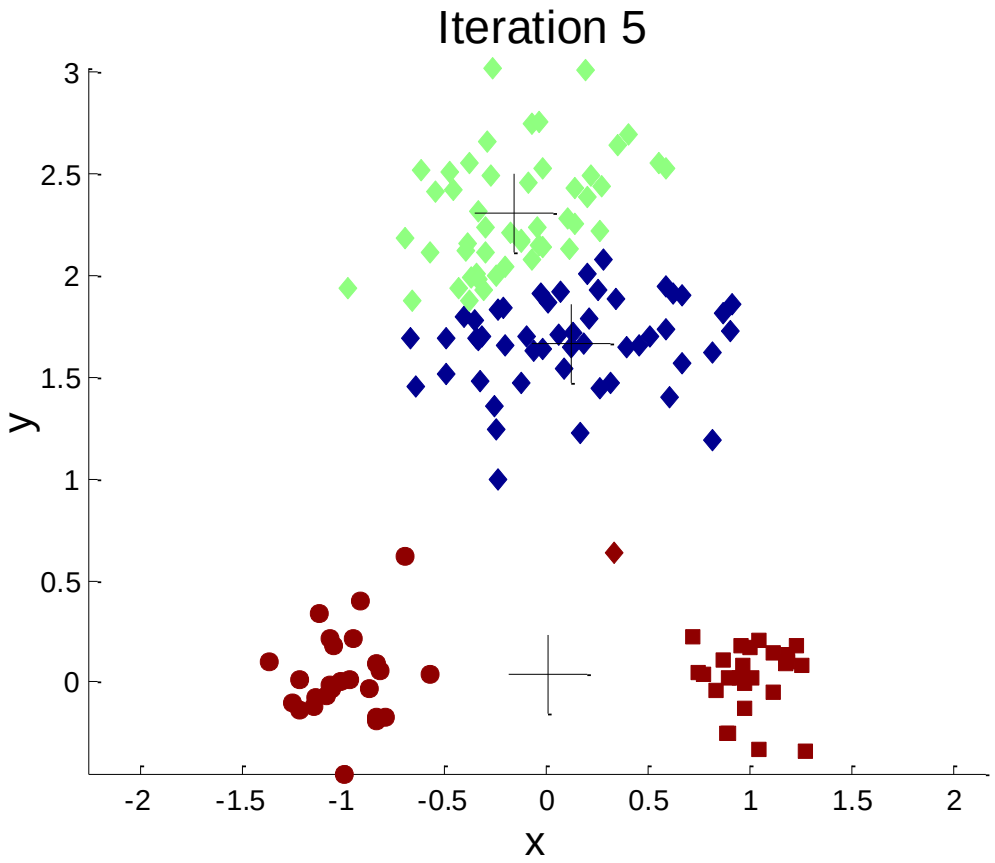
(d) Iteration: $t = 3$



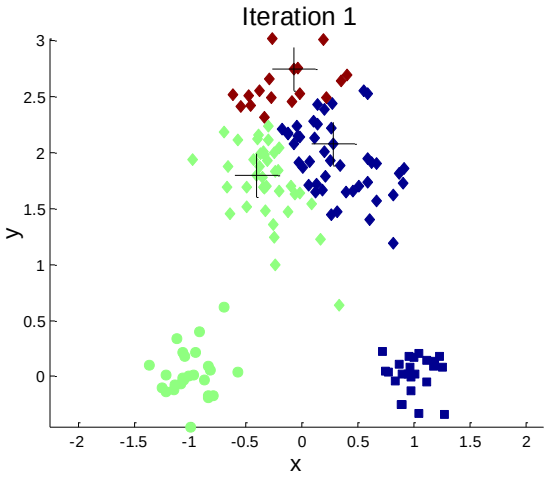
(e) Iteration: $t = 4$

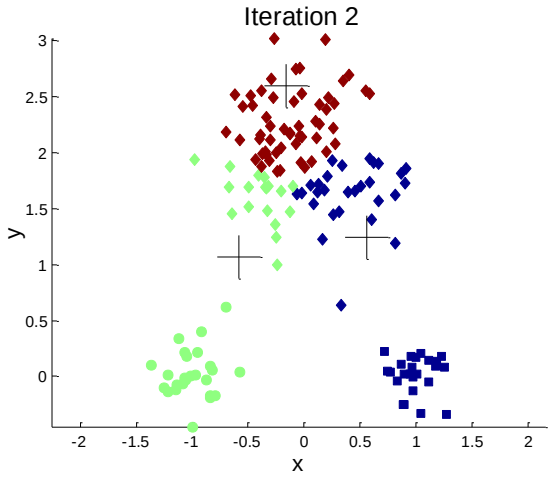
(f) Iteration: $t = 5$ (converged)

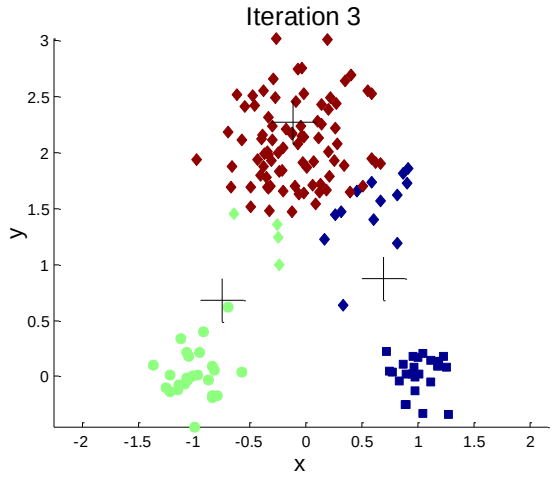
初始中心的選擇重要

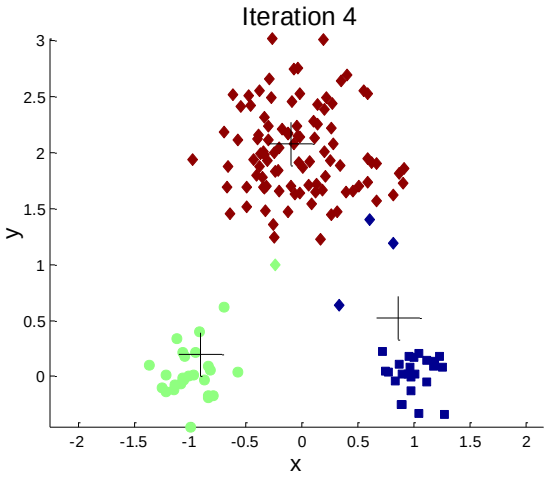


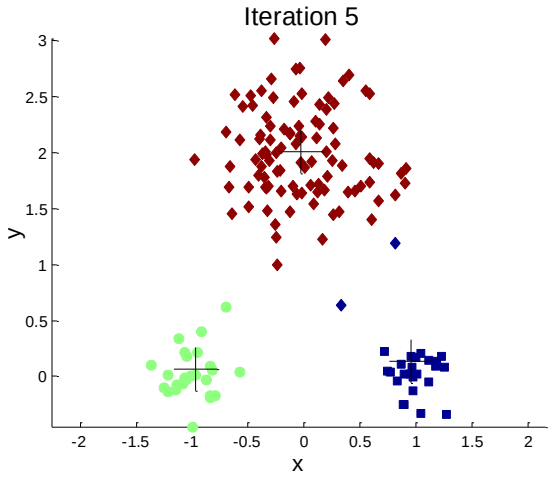
初始中心的選擇很重要

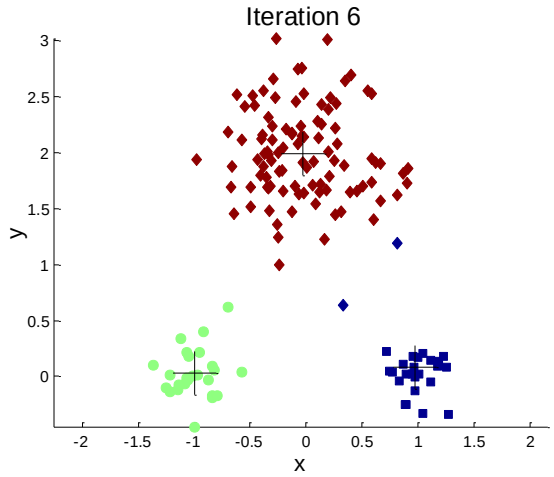

Iteration 1
x
y


Iteration 2
x
y

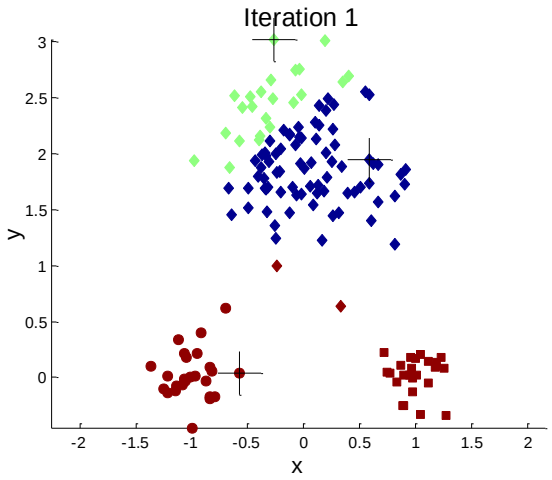

Iteration 3
x
y

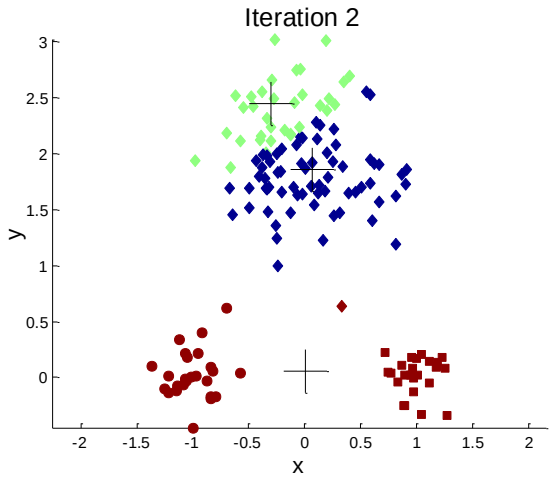

Iteration 4
x
y

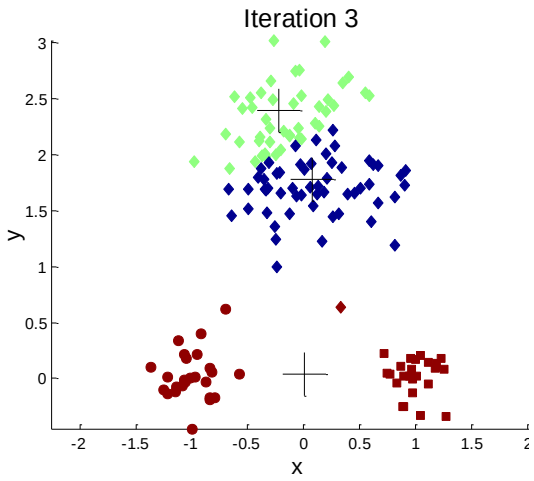

Iteration 5
x
y

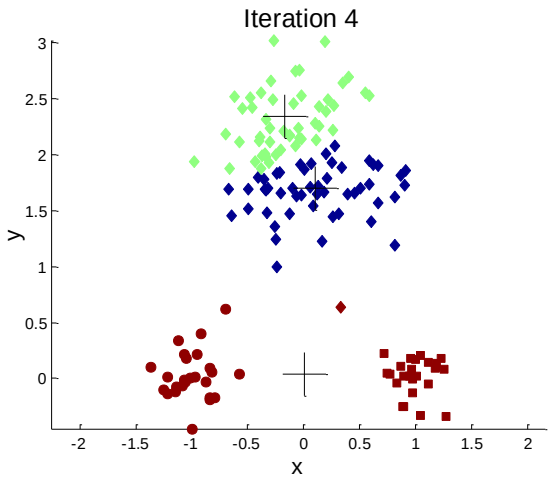

Iteration 6
x
y

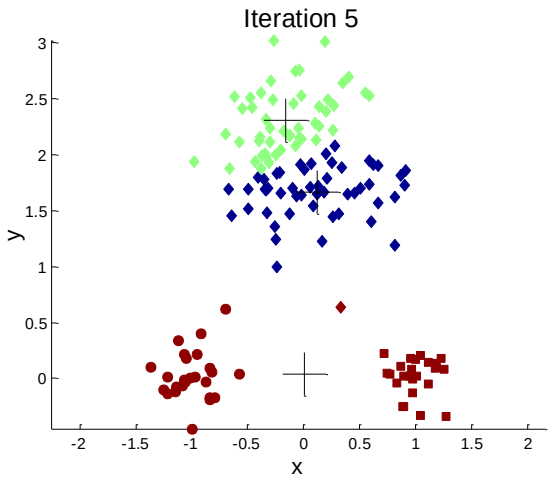
初始中心的選擇很重要


Iteration 1
x
y


Iteration 2
x
y


Iteration 3
x
y


Iteration 4
x
y

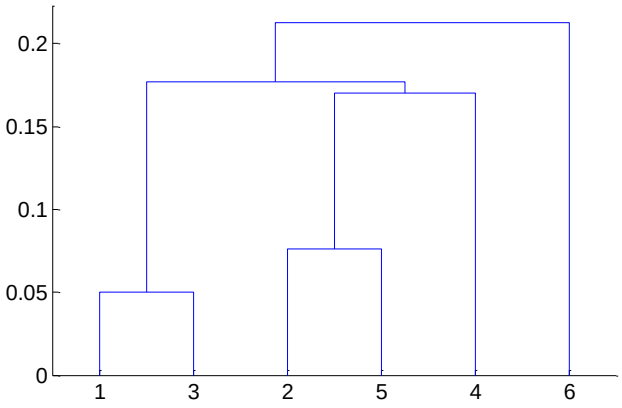

Iteration 5
x
y

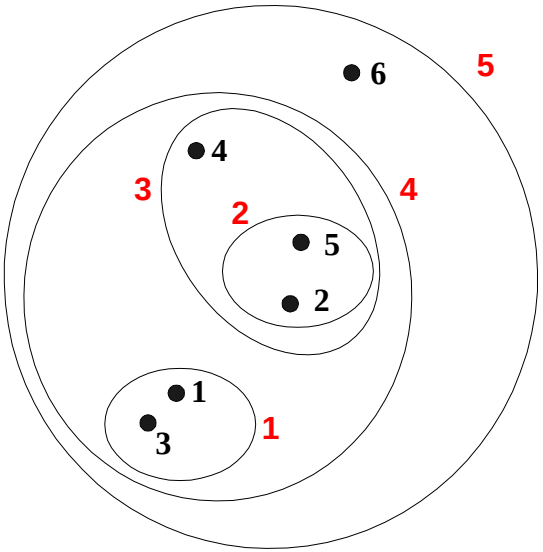
Clustering Algorithm–Agglomerative Hierarchical Clustering

集群演算法 - 階層分群

階層分群演算法

- 將相近的群或點連結起來成一個階層樹
- 可用樹狀圖(dendrogram)表示



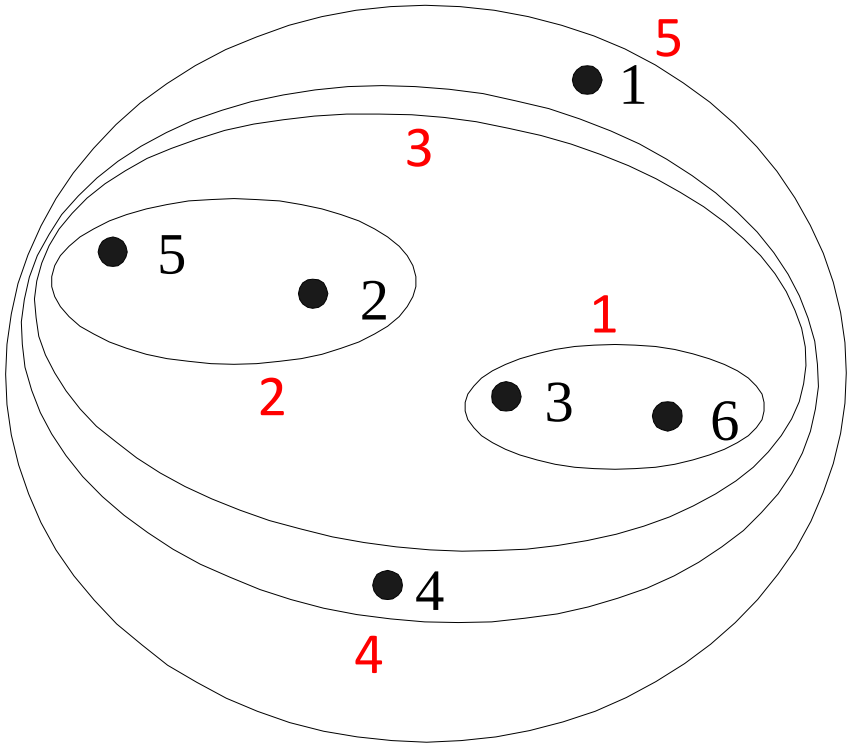


階層分群演算法

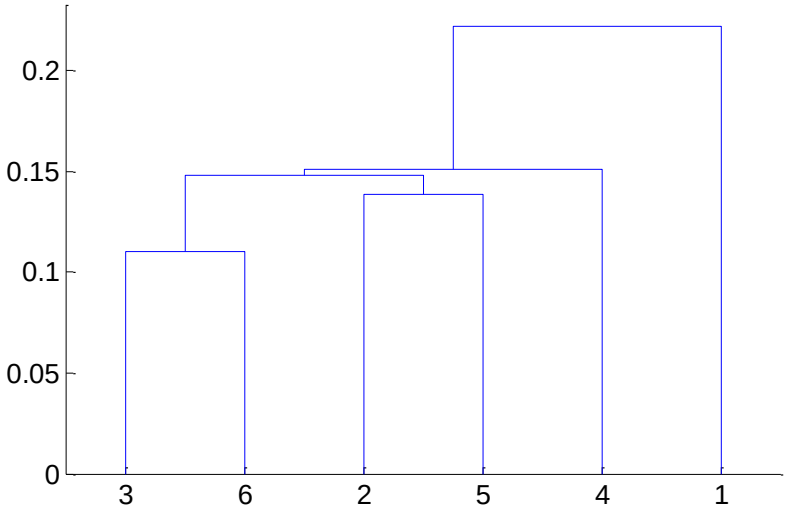
- 演算法

  Compute the proximity matrix
  Let each data point be a cluster
  Repeat
    Merge the two closest clusters
    Update the proximity matrix
  Until only a single cluster remains

最近距離



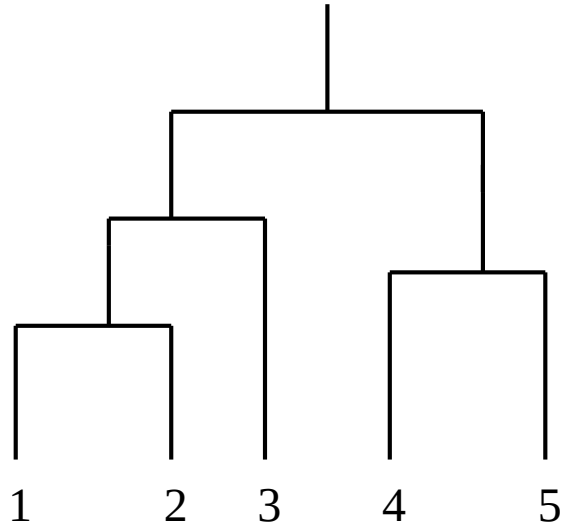
Nested Clusters



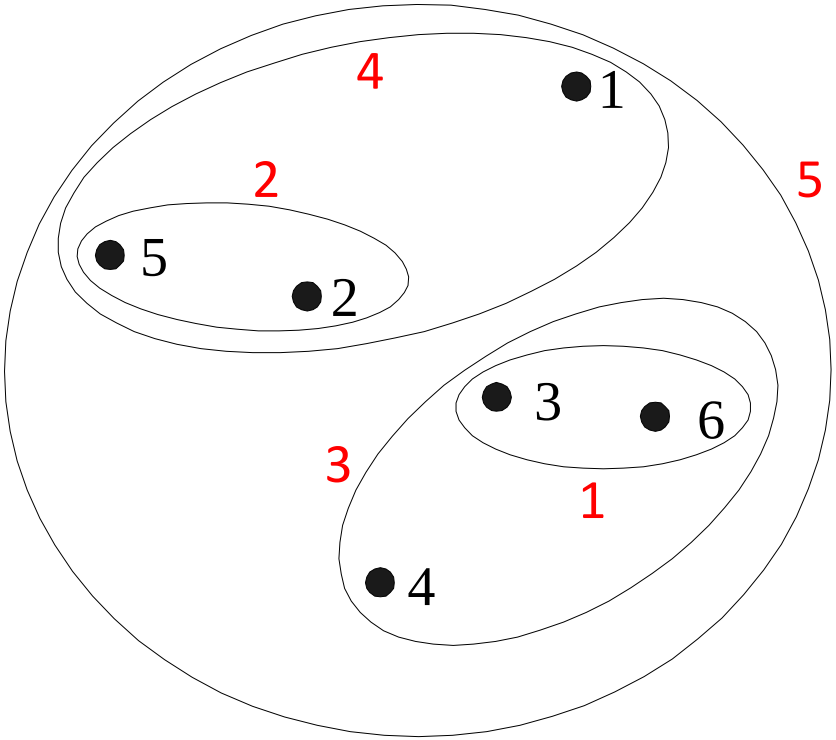
Dendrogram

最近距離

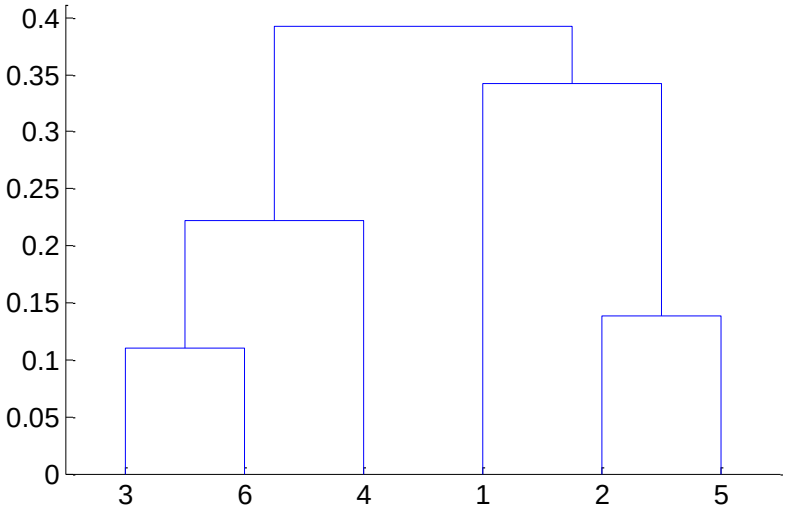
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



最遠距離



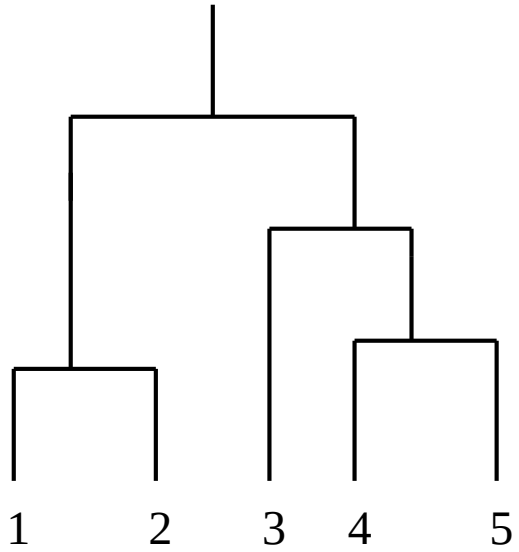
Nested Clusters



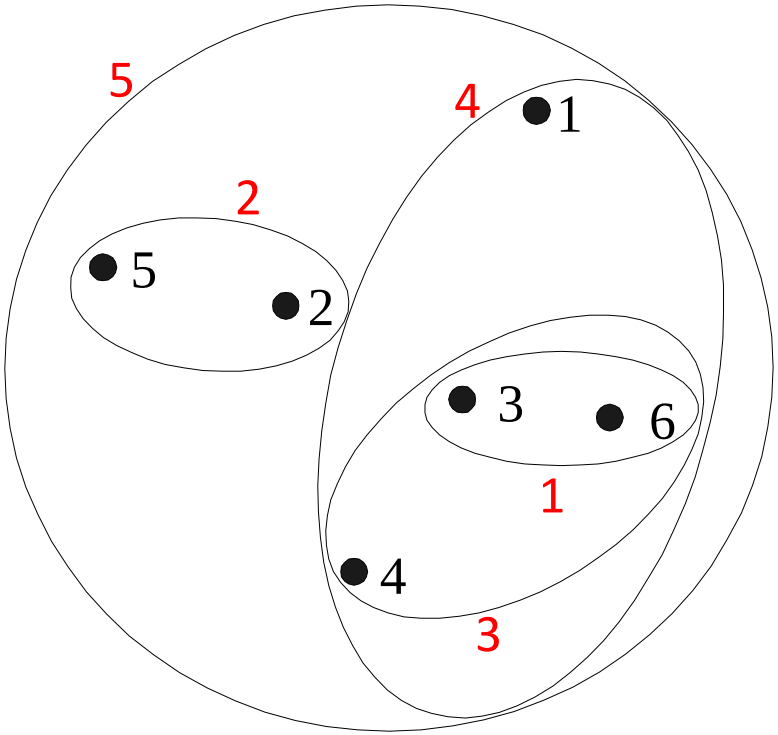
Dendrogram

最遠距離

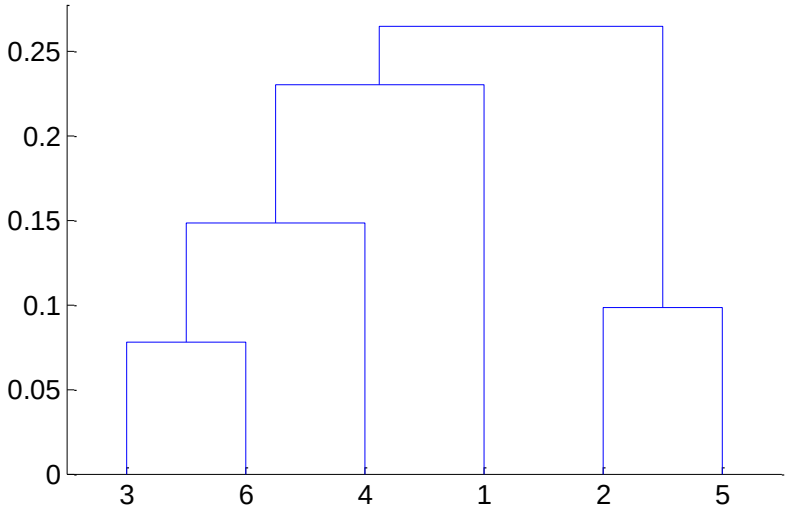
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



平均距離



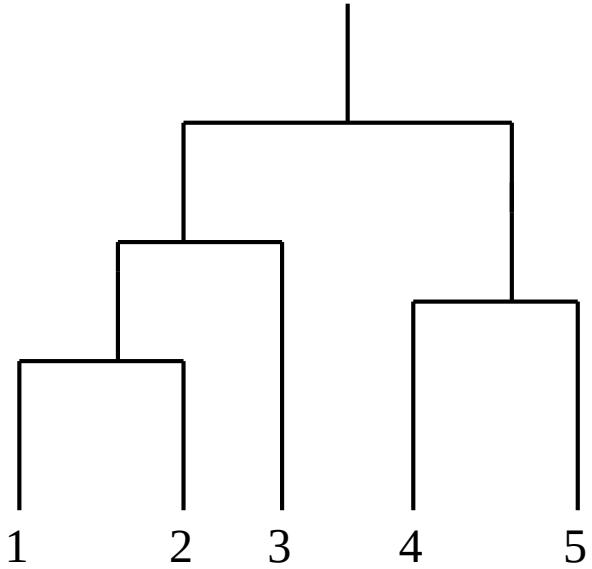
Nested Clusters



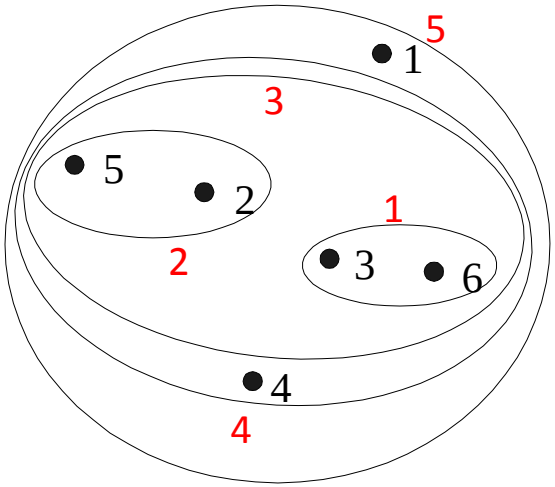
Dendrogram

平均距離

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

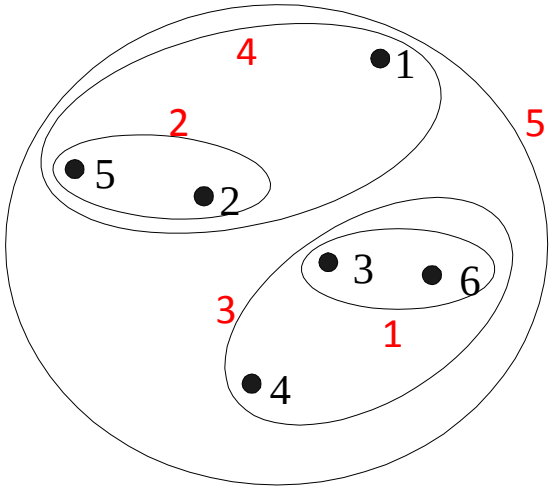


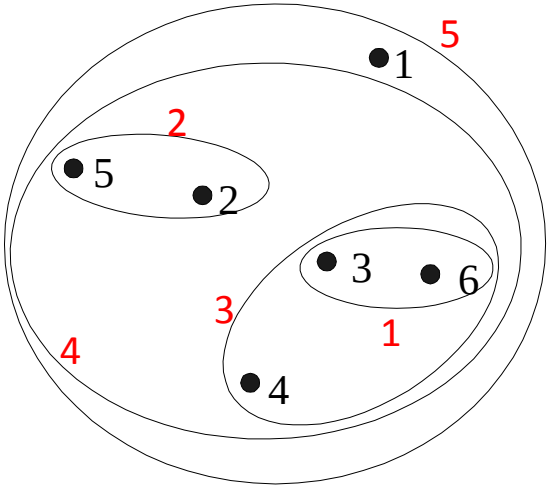
使用不同距離計算的結果



MIN

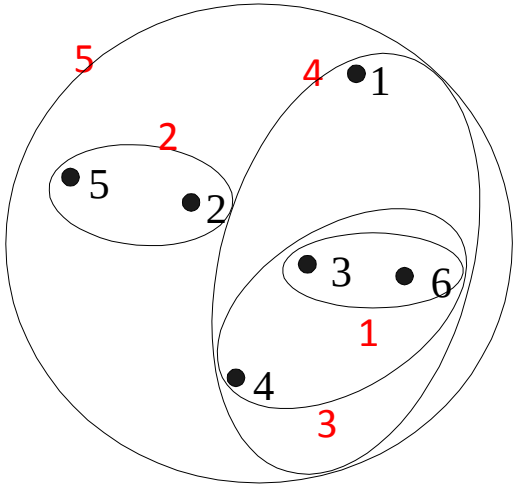
MAX





Group Average

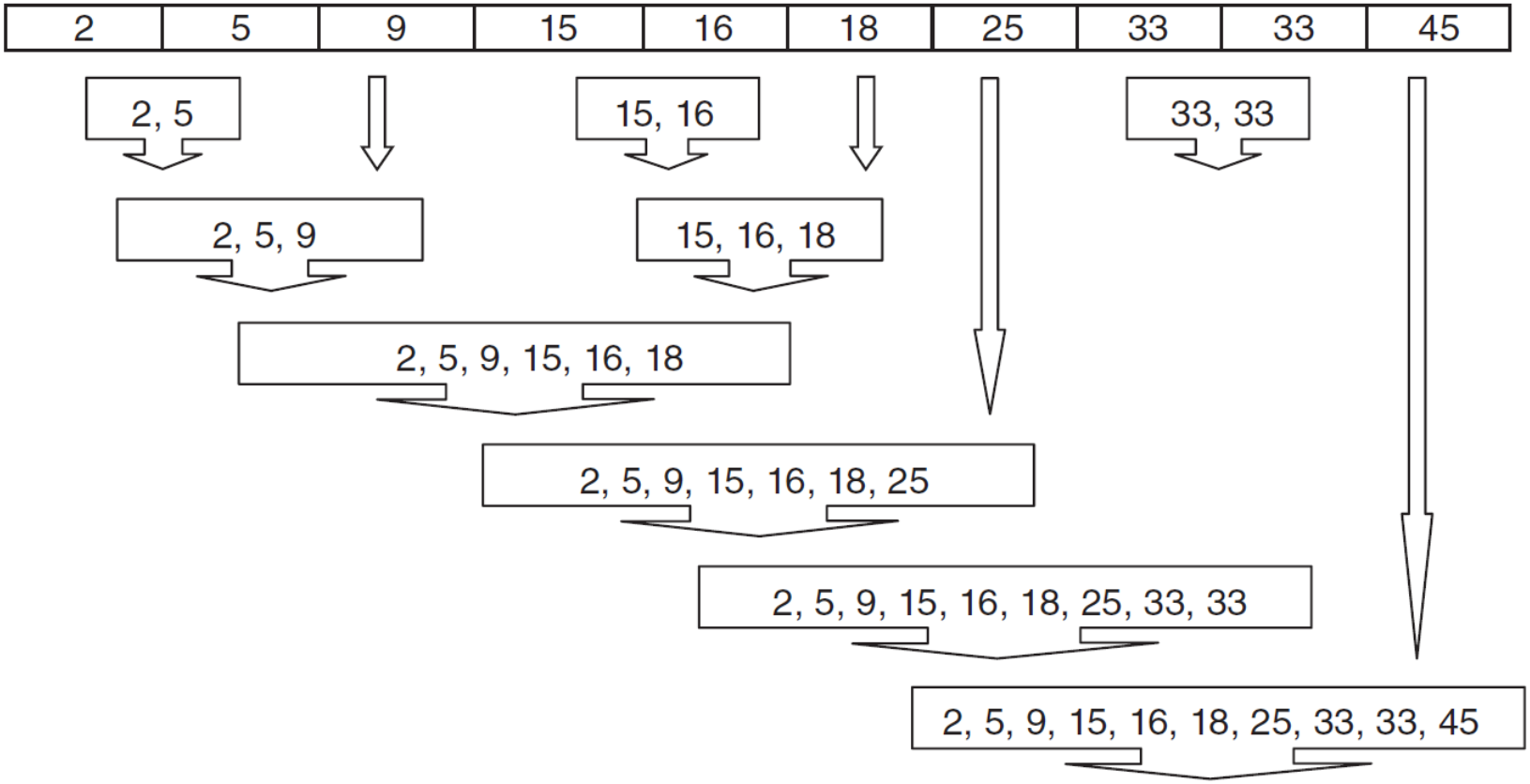
Ward's Method



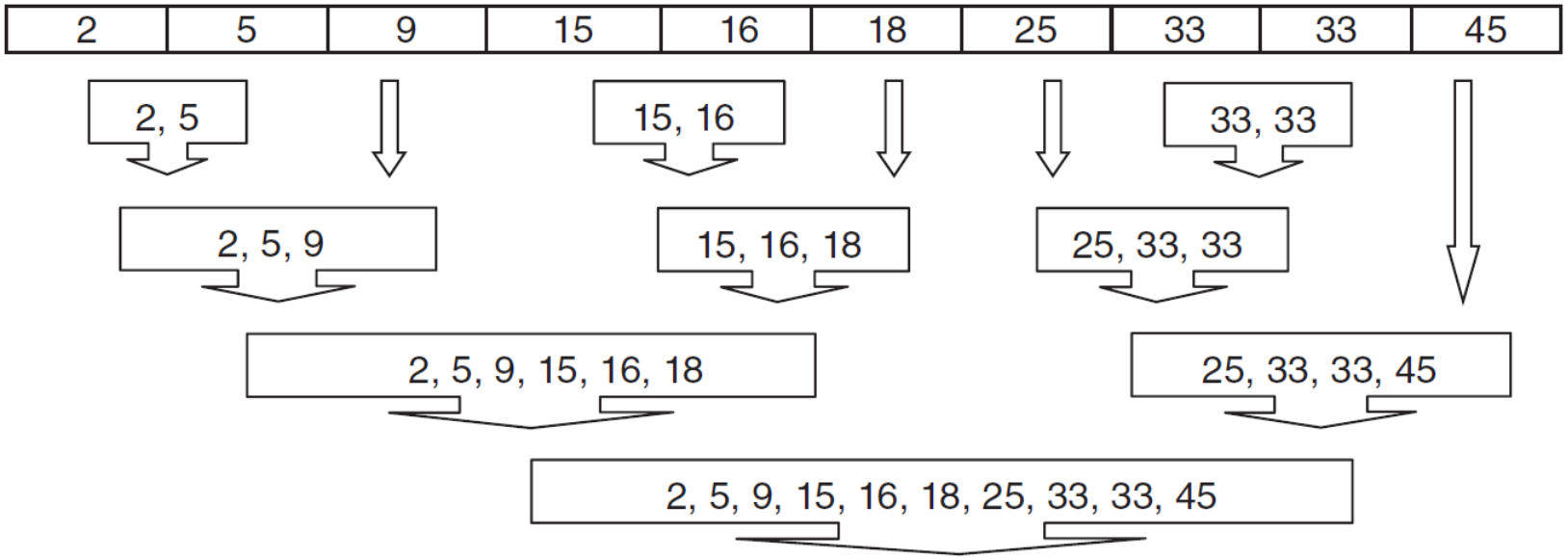
練習一下

- 將2、5、9、15、16、18、25、33、33、45利用最近距離與最遠距離分群

練習一下 ~ 最近距離


2
5
9
15
16
18
25
33
33
45
2, 5
15, 16
33, 33
2, 5, 9
15, 16, 18
2, 5, 9, 15, 16, 18
2, 5, 9, 15, 16, 18, 25
2, 5, 9, 15, 16, 18, 25, 33, 33
2, 5, 9, 15, 16, 18, 25, 33, 33, 45

練習一下 ~ 最遠距離


2
5
9
15
16
18
25
33
33
45
2, 5
15, 16
33, 33
2, 5, 9
15, 16, 18
25, 33, 33
2, 5, 9, 15, 16, 18
25, 33, 33, 45
2, 5, 9, 15, 16, 18, 25, 33, 33, 45

本單元課程結束

感謝您們的參與