

Day 57

# 非監督式機器學習

## 階層分群算法



出題教練

周俊川

# 本日知識點目標

- 瞭解階層分群算法流程，及群數的定義
- 瞭解階層分群與 k-means 差異，及其優劣比較
- 階層分群的距離計算方式

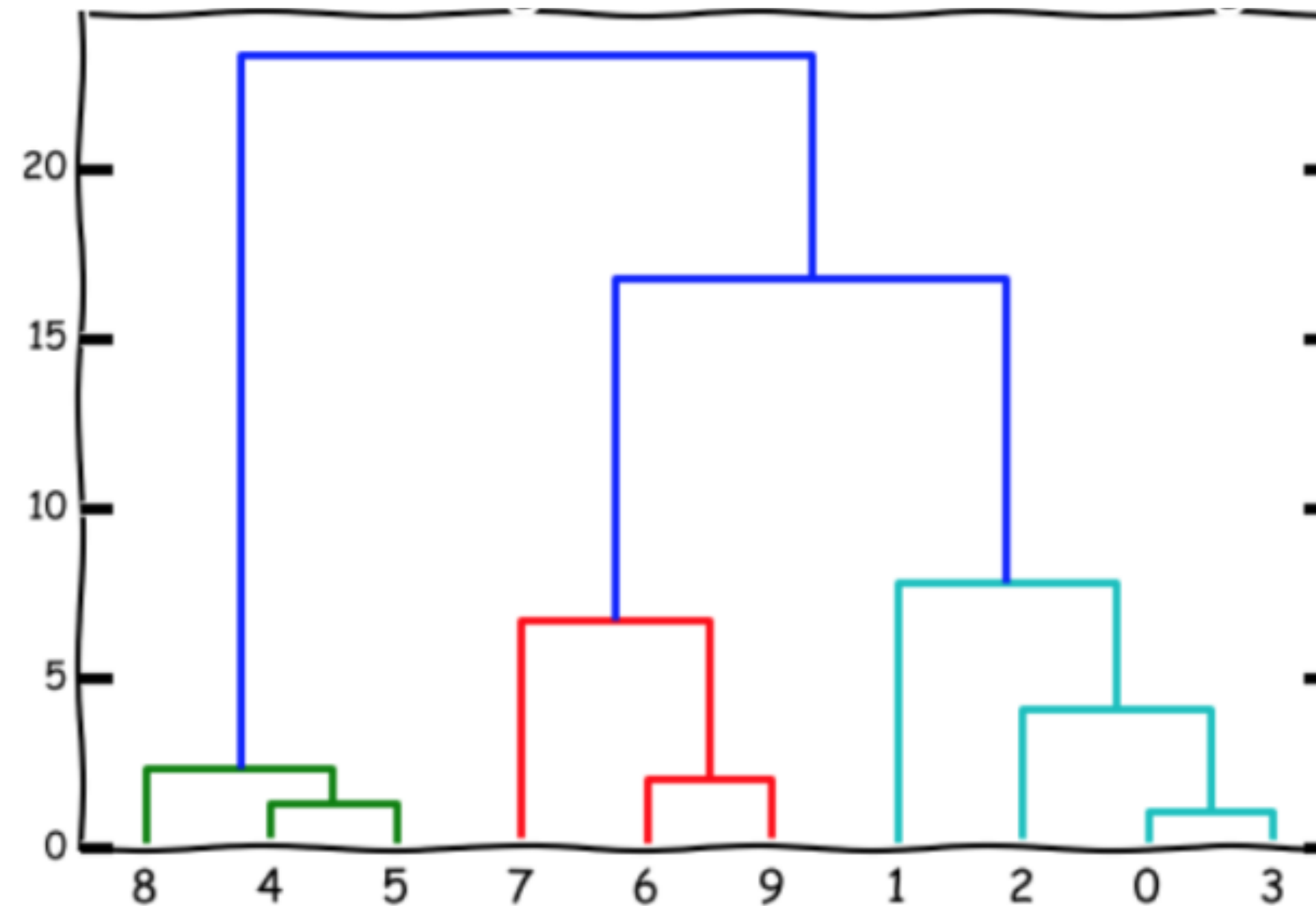


一種構建 cluster 的層次結構的算法。該算法從分配給自己 cluster 的所有資料點開始。然後，兩個距離最近的 cluster 合併為同一個 cluster。最後，當只剩下一個 cluster 時，該算法結束。

# 樹狀圖

可定義 4, 5 是一群，或 8, 4, 5 是一群，看距離怎麼衡量 (y 軸要切在哪兒)

**Here's a dendrogram of our hierarchical clustering**



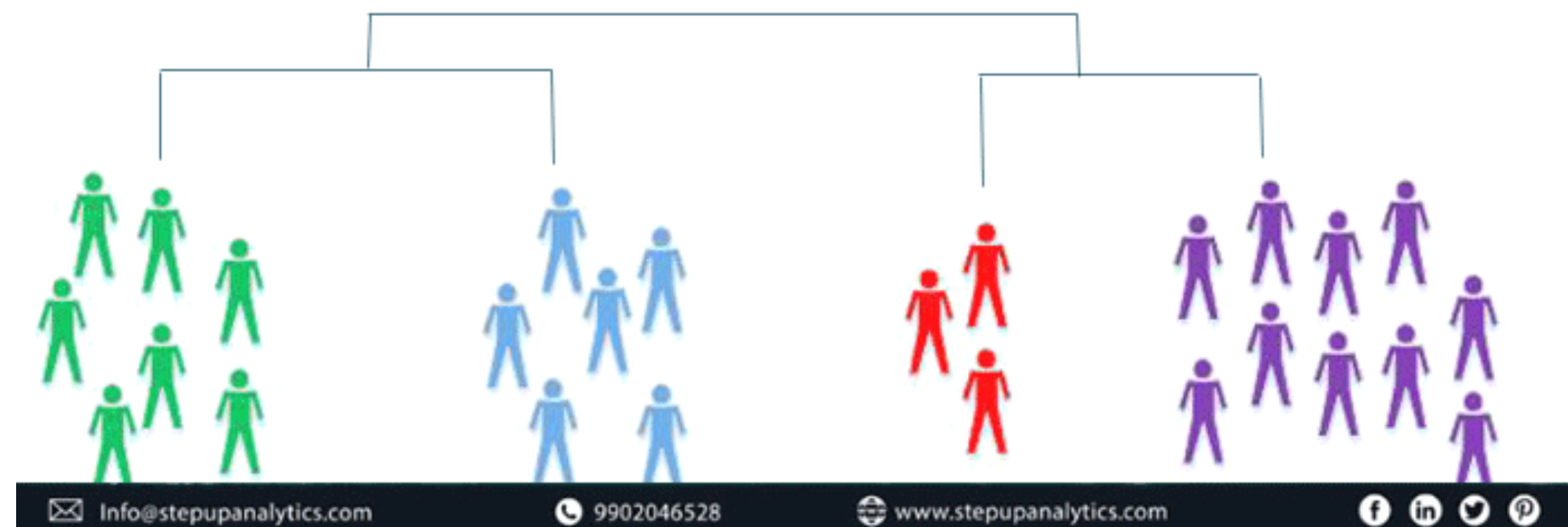
圖片來源：[Jocelyn Ong](#)

# K-means vs. 階層分群



**K-mean 要預先定義群數  
(n of clusters)**

## Comparison of Kmeans and Hierarchical Clustering



**階層分群可根據定義距離來分群  
(bottom-up)，也可以決定羣數做  
分羣 (top-down)**

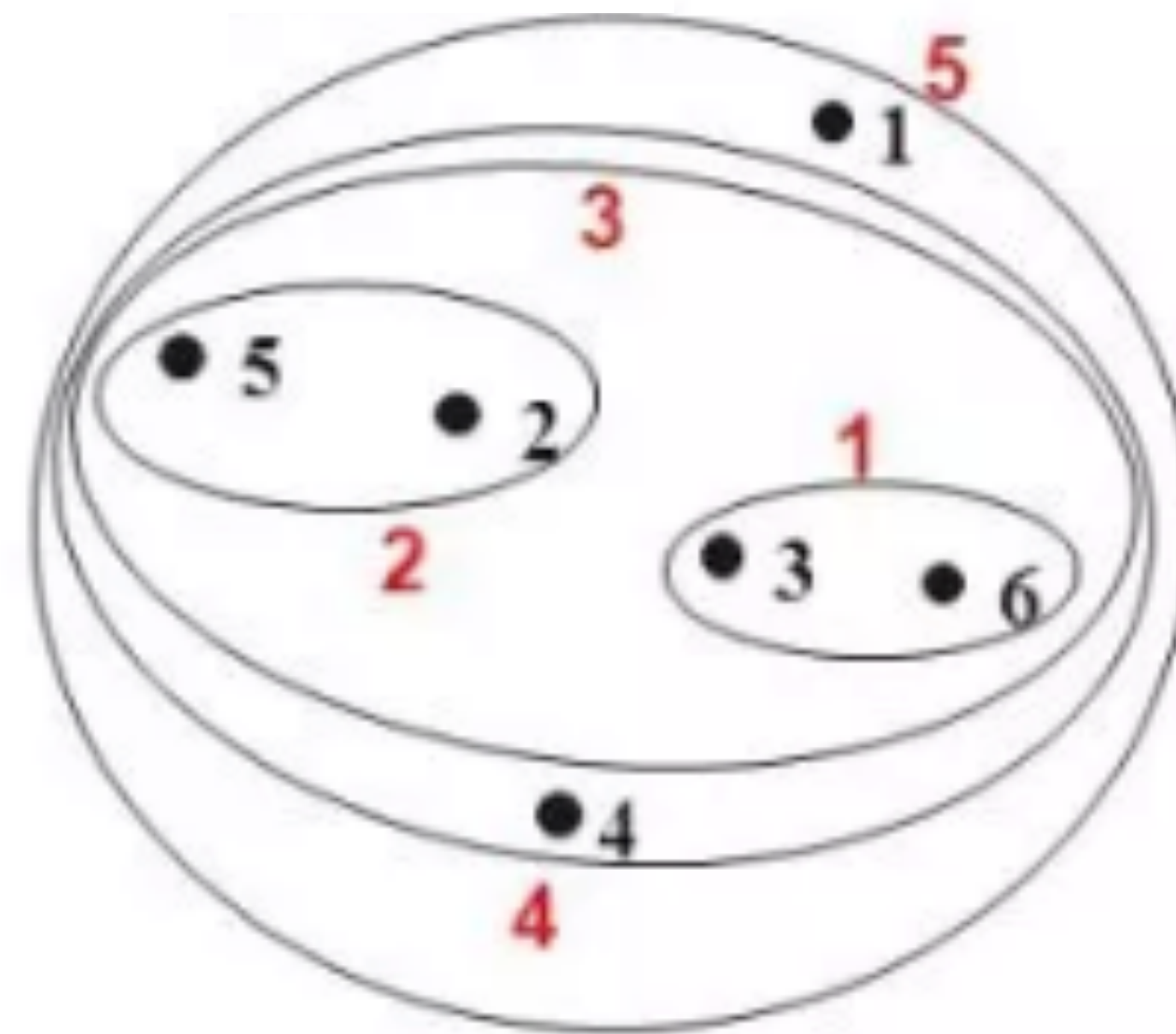
## 不指定分群的數量

- 每筆資料為一個 cluster
- 計算每兩群之間的距離
- 將最近的兩群合併成一群
- 重覆步驟 2、3，直到所有資料合併成同一 cluster

# 階層分群距離計算方式：single-link

群聚與群聚間的距離可以定義為不同群聚中最接近兩點間的距離。

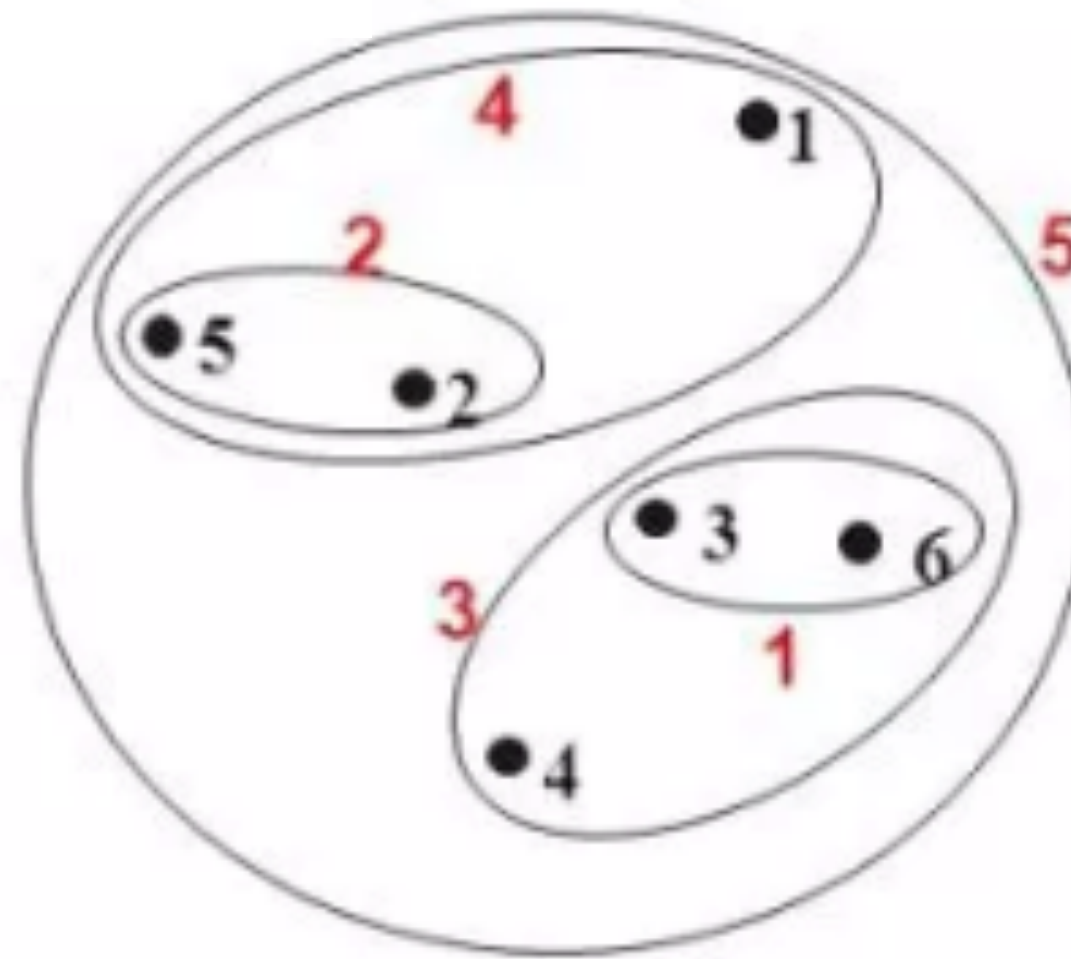
## Single-link



# 階層分群距離計算方式：complete-link

群聚間的距離定義為不同群聚中最遠兩點間的距離，這樣可以保證這兩個集合合併後，任何一對的距離不會大於  $d$ 。

## Complete-link

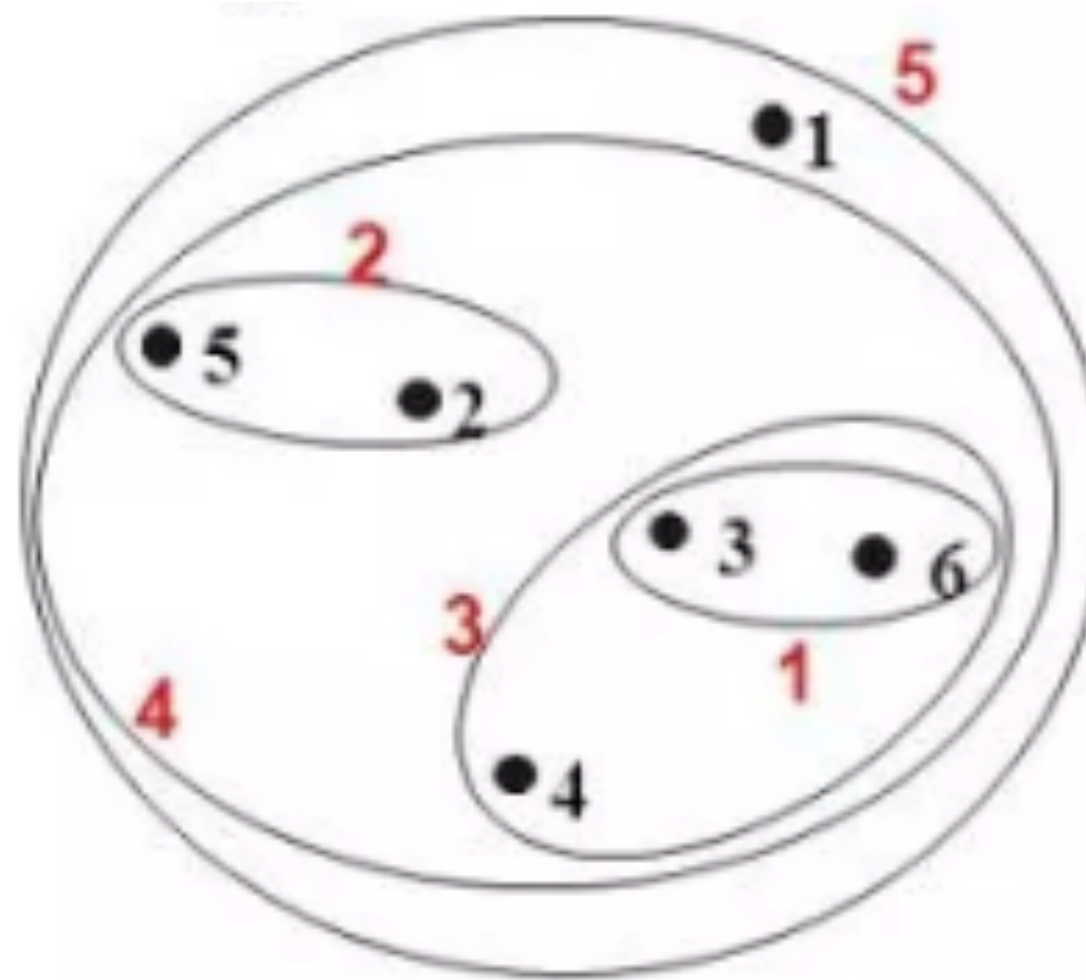




# 階層分群距離計算方式：average-link

群聚間的距離定義為不同群聚間各點與各點間距離總和的平均。

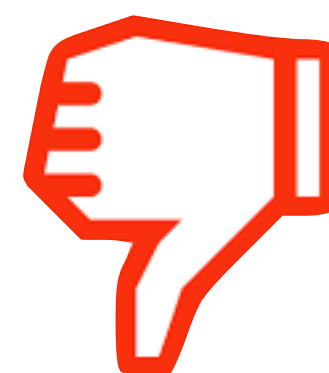
## Average-link





## 優點

1. 概念簡單，易於呈現
2. 不需指定群數



## 缺點

只適用於少量資料，大量資料  
會很難處理

- 階層式分群在無需定義群數的情況下做資料的分群，而後可以用不同的距離定義方式決定資料群組。
- 分群距離計算方式有 single-link, complete-link, average-link。
- 概念簡單且容易呈現，但不適合用在大資料。

# 解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業  
開始解題

