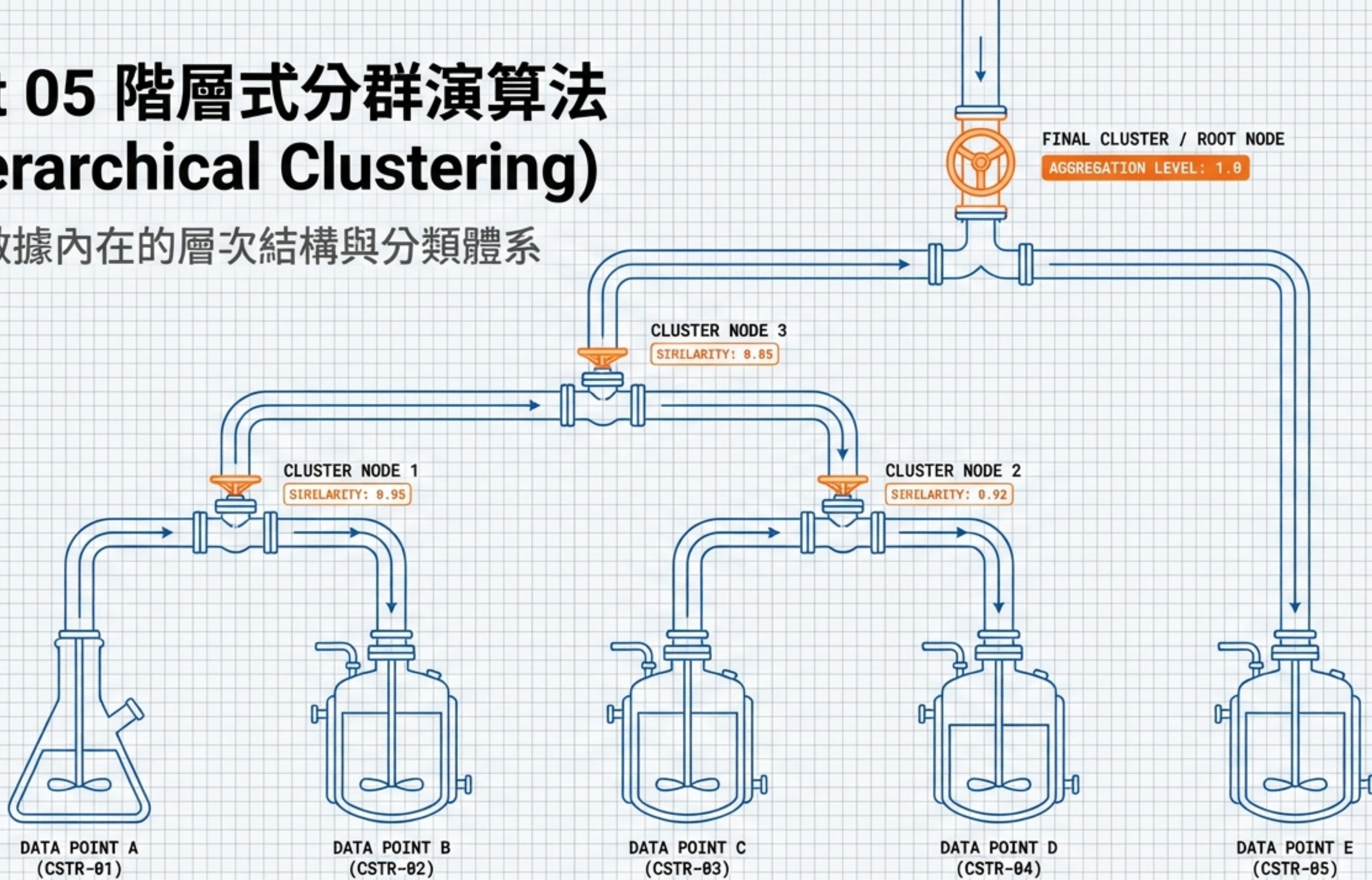


Unit 05 階層式分群演算法 (Hierarchical Clustering)

揭示數據內在的層次結構與分類體系

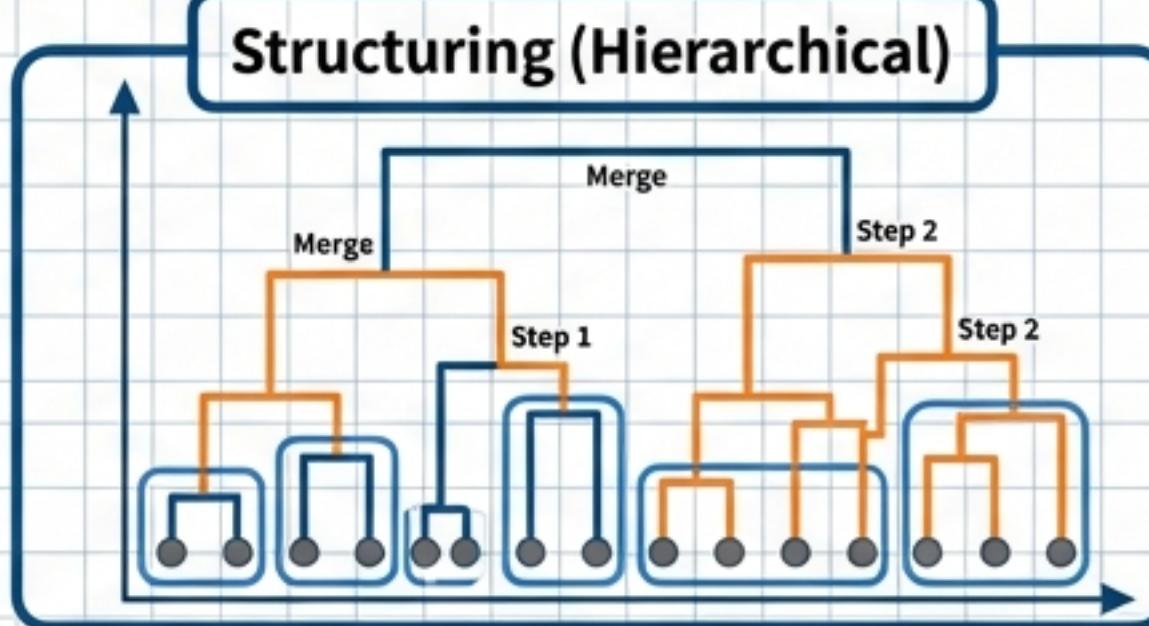
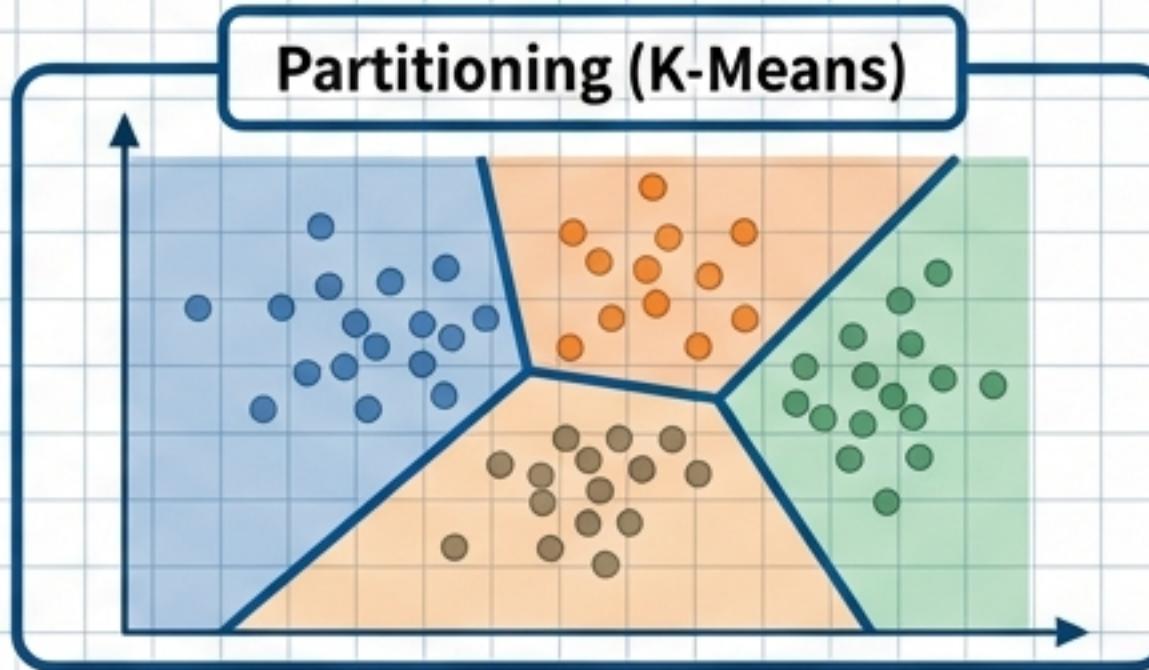


什麼是階層式分群？(Definition & Core Concept)

定義：一種建立數據點之間層次結構 (Hierarchy) 的分群方法。

關鍵差異 (Key Differentiator)：與 K-Means 不同，
不需要預先指定群集數量 (K)。

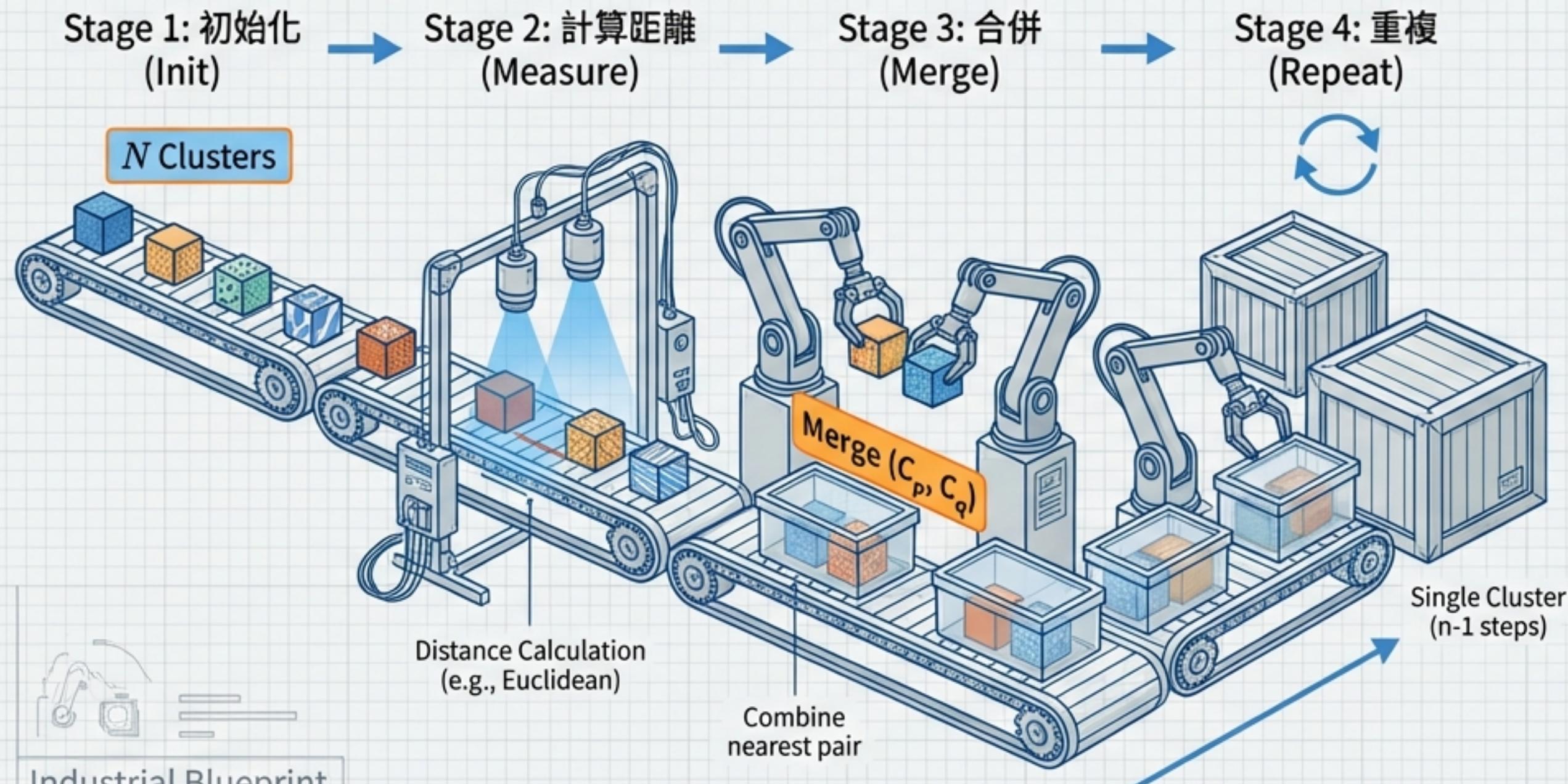
核心機制：透過逐步合併 (Agglomerative) 或分裂 (Divisive) 的方式，建立樹狀結構。



Note: 適合探索未知結構的數據，提供從微觀到巨觀的完整視角。

演算法原理：凝聚式策略 (Agglomerative Approach)

由下而上的組裝流程 (Bottom-Up Assembly)



技術說明 (Technical Notes)

分裂式 (Divisive)：由上而下 (Top-Down) 的策略，計算成本較高 $(O(2^n))$ 。

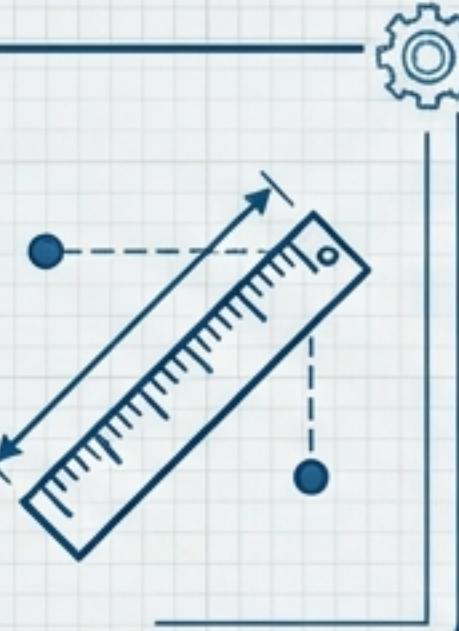
時間複雜度 (Time Complexity)：一般凝聚式 實作為 $O(n^2 \log n)$ 。

距離度量 (Distance Metrics) : 定義相似的標準

1. 歐幾里得距離 (Euclidean)

$$\sqrt{\sum(x_i - x_j)^2}$$

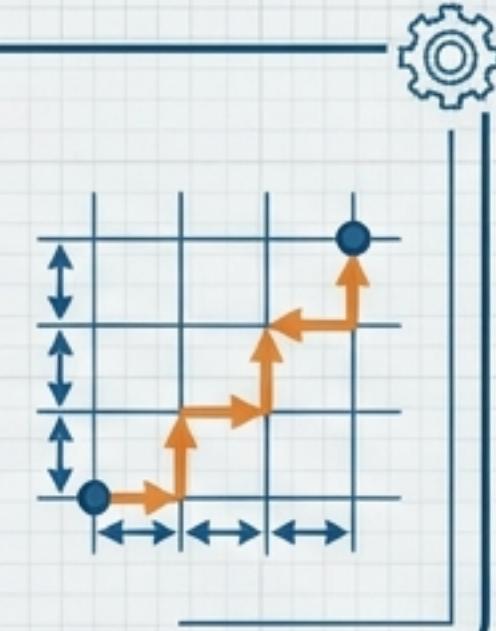
適用：操作參數（溫度、壓力），物理意義相同時。



2. 曼哈頓距離 (Manhattan)

$$\sum |x_i - x_j|$$

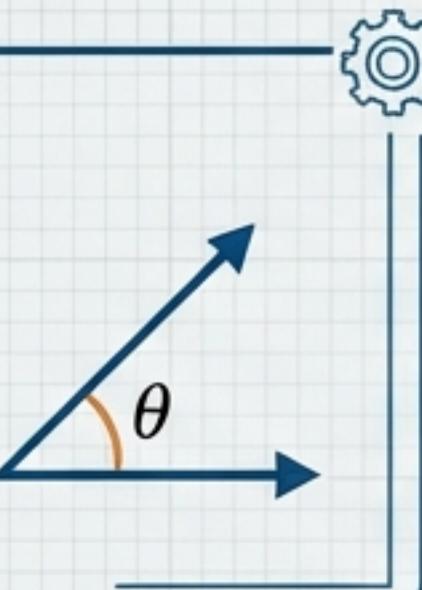
適用：穩健性需求，對離群值較不敏感。



3. 餘弦距離 (Cosine)

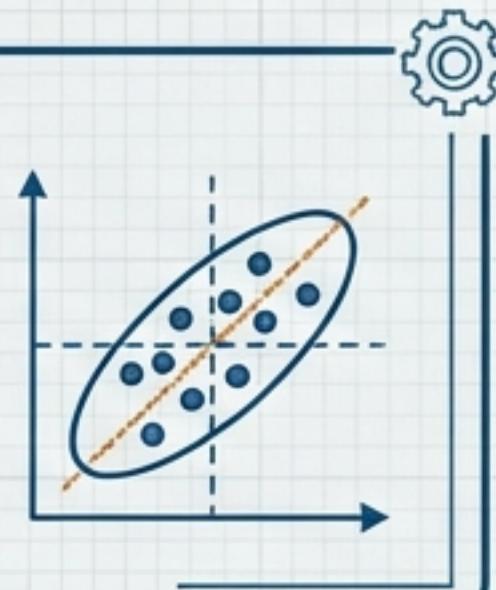
$$1 - \cos(\theta)$$

適用：配方比例，關注成分比例而非總量。



4. 馬氏距離 (Mahalanobis)

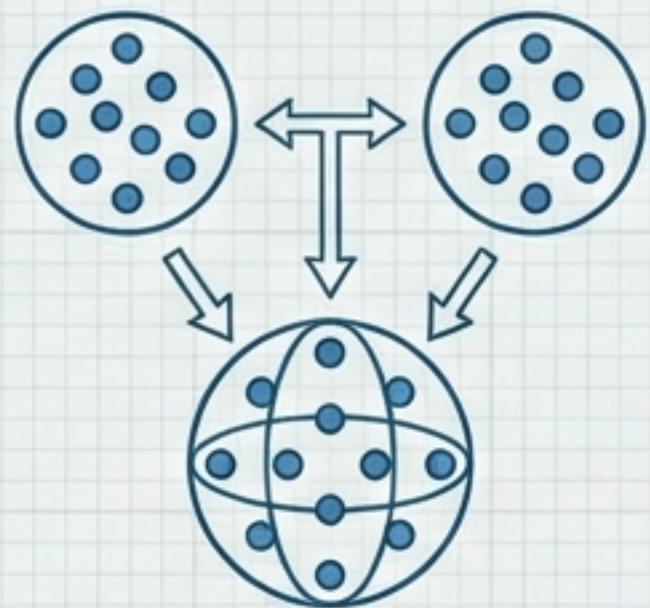
適用：相關性變數，考慮數據分布結構。



⚠ 注意：在化工應用中，單位不同的變數（如溫度 vs 濃度）必須先進行標準化 (Standardization)。

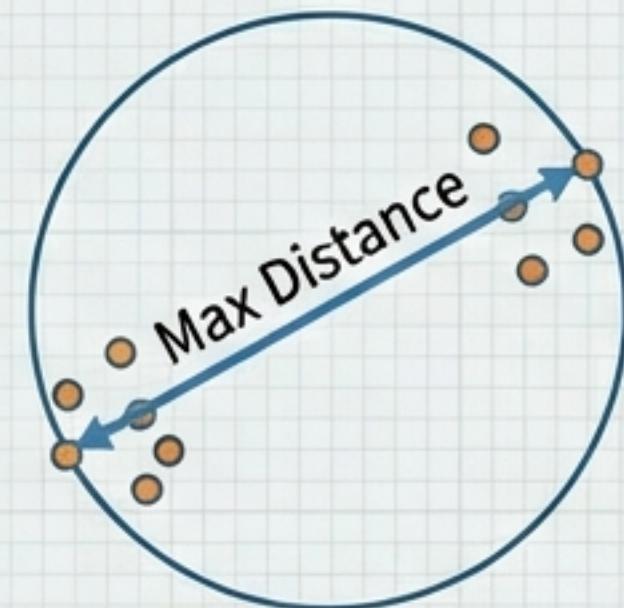
連結方法 (Linkage Methods)：決定群集的形狀

Ward (華德法)



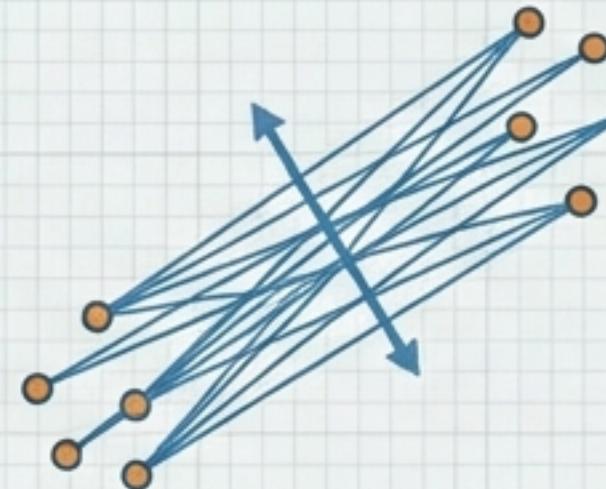
最小化變異數。產生球形、大小相近群集。

Complete (最大連結)



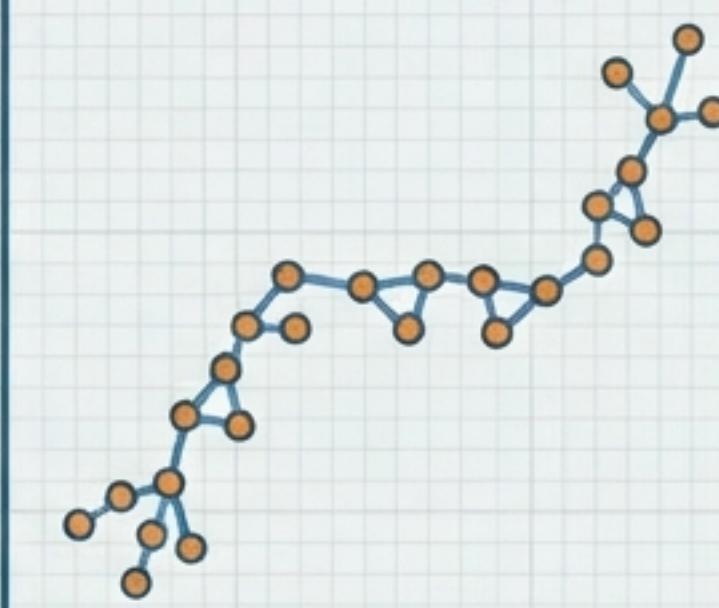
最遠點距離。產生緊湊球形，對離群值穩健。

Average (平均連結)



所有點對距離平均。
平衡選擇。

Single (最小連結)

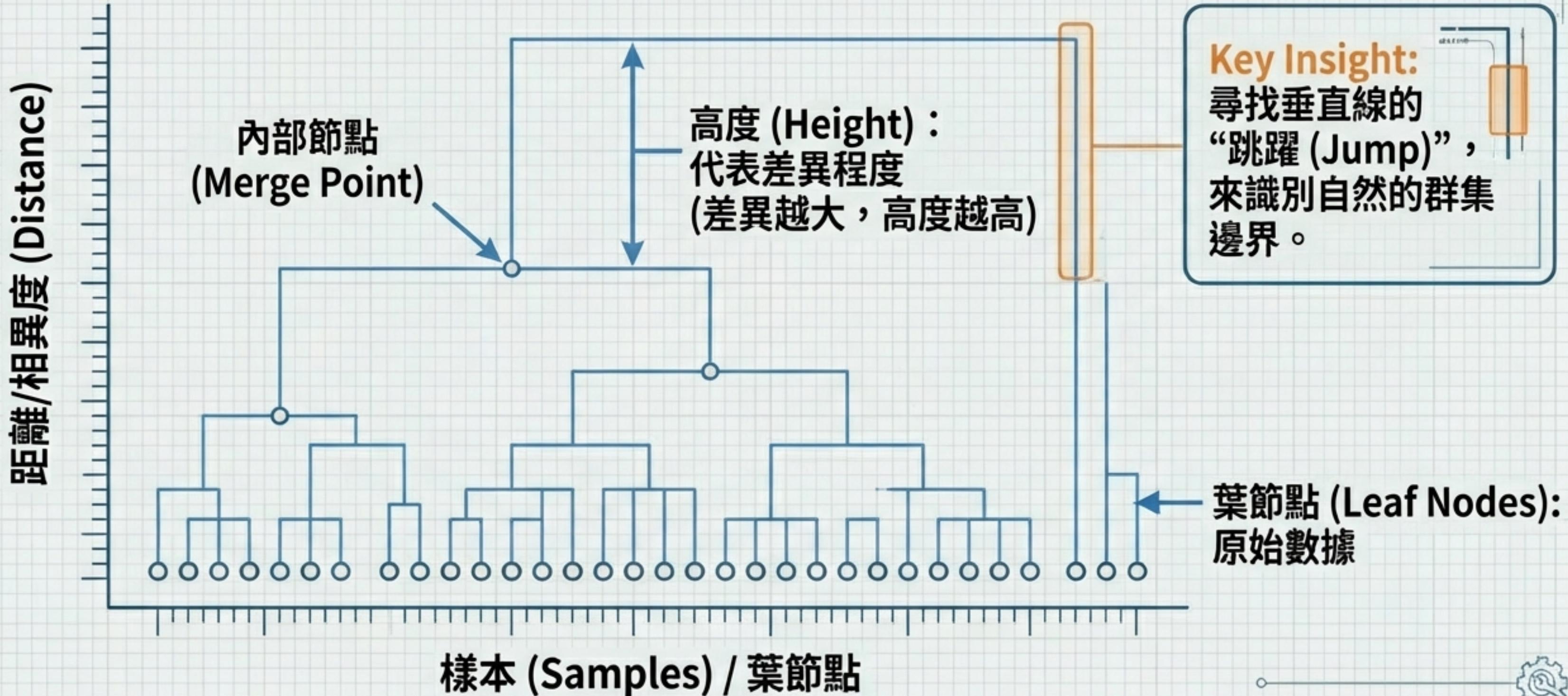


最近點距離。產生細長型 (Chain Effect)。

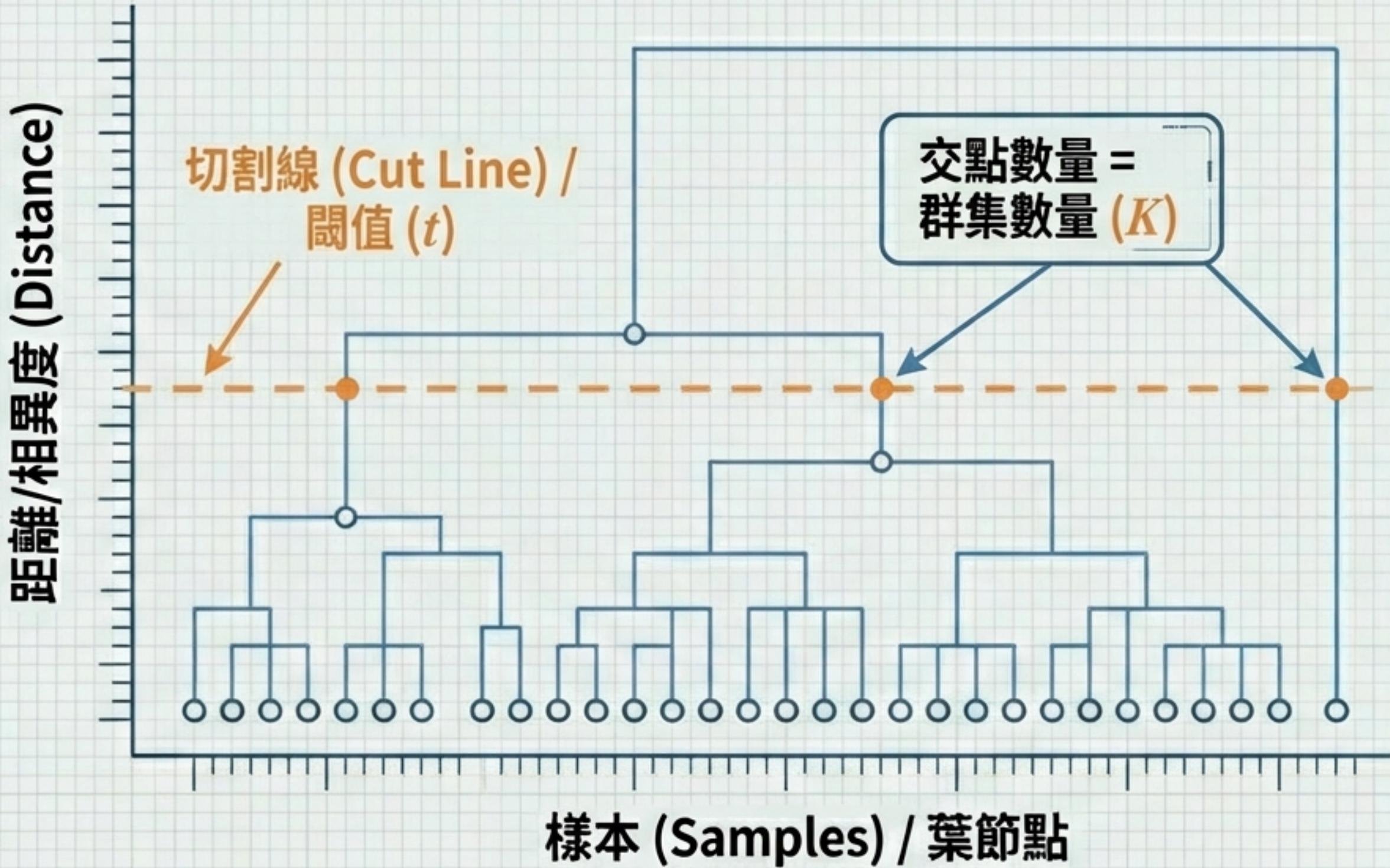
⚠ 風險



樹狀圖 (Dendrogram) 解讀指南



決定群集數量：切割樹狀圖



三種切割方法：

- 觀察高度跳躍 (最直觀)：在垂直線最長且不交叉的區間切割。
- 指定數量 (K)：依業務需求 (如：大中小三類) 反推高度。
- 距離閾值：設定最大容許差異值。

實務案例：塗料產品配方階層分類

Problem Statement:

化工廠擁有 500 筆產品配方，需建立多層次分類體系以優化庫存與生產。

Spec Sheet

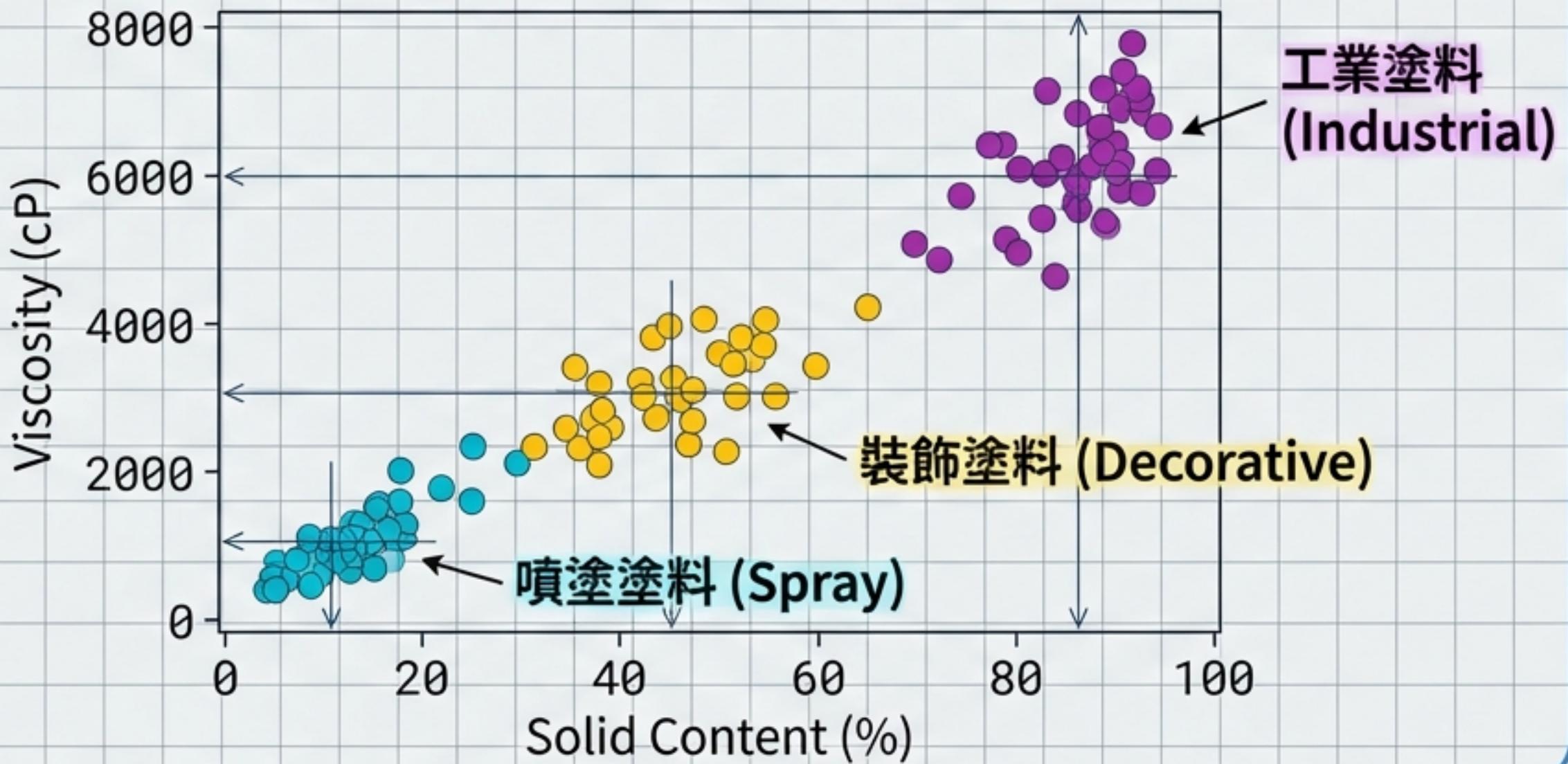
 Sample ID	 Resin (%)	 Solvent (%)	 Pigment (%)	 Viscosity (cP)	 Cost (\$)
001	37.7	38.1	16.4	3296	125
002
...

Goal:

建立 大類 (Level 1) → 中類 (Level 2) → 小類 (Level 3) 的階層管理系統。



數據探索：配方特徵分佈

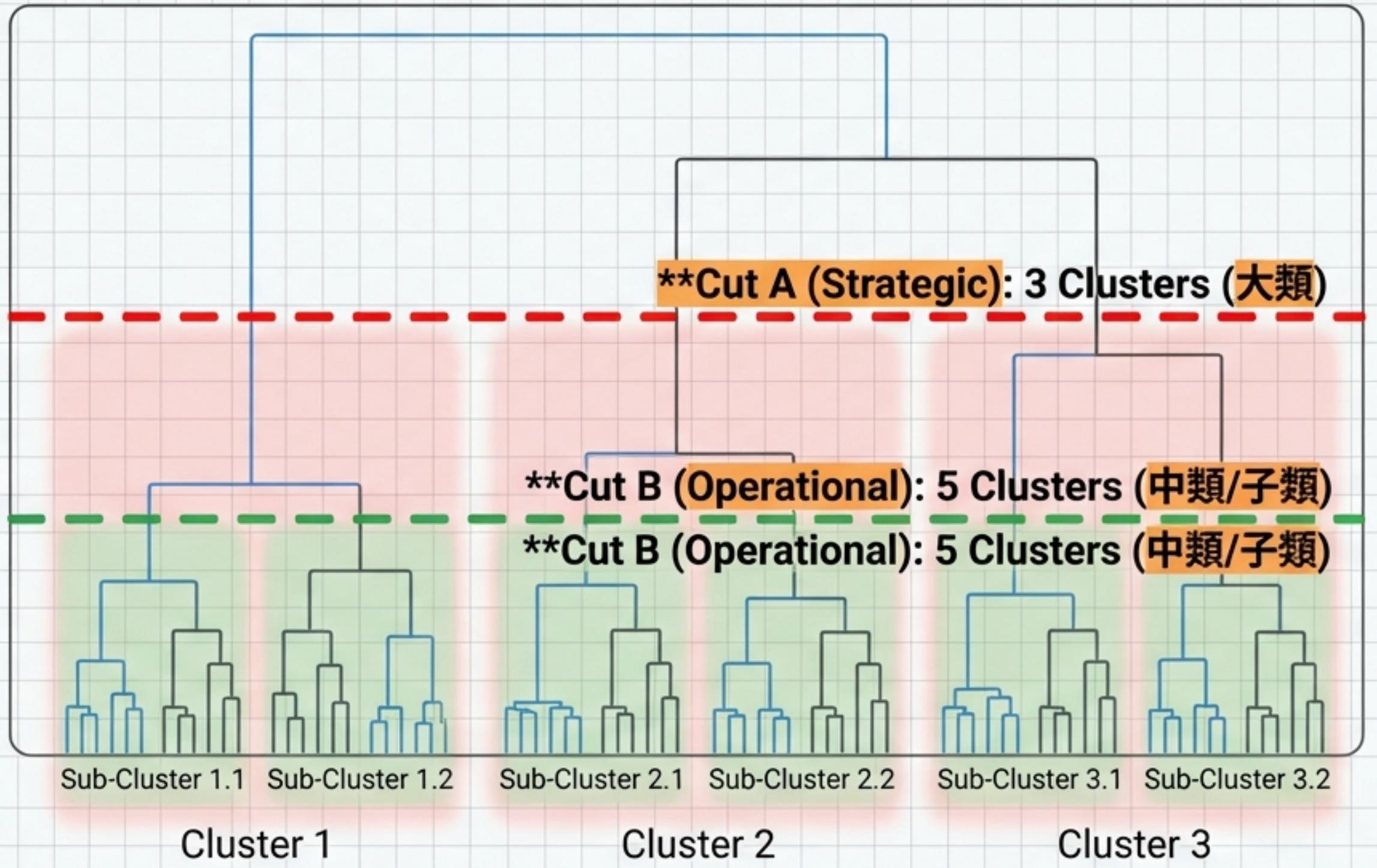


觀察：工業塗料與其他類別差異顯著；噴塗與裝飾塗料在部分特徵上有重疊。

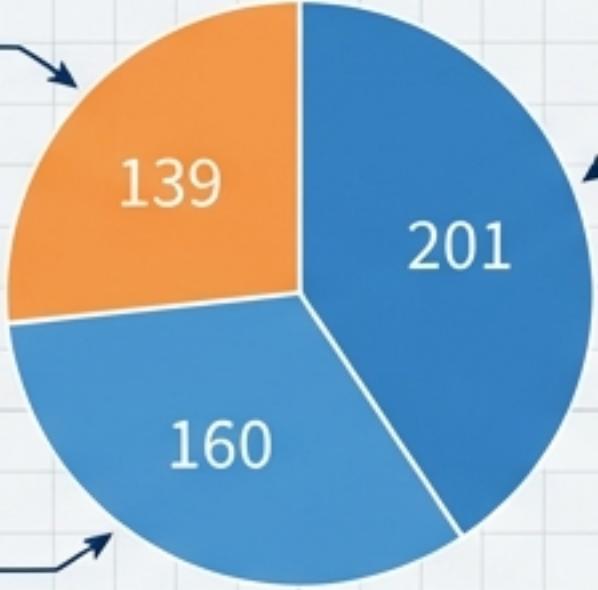
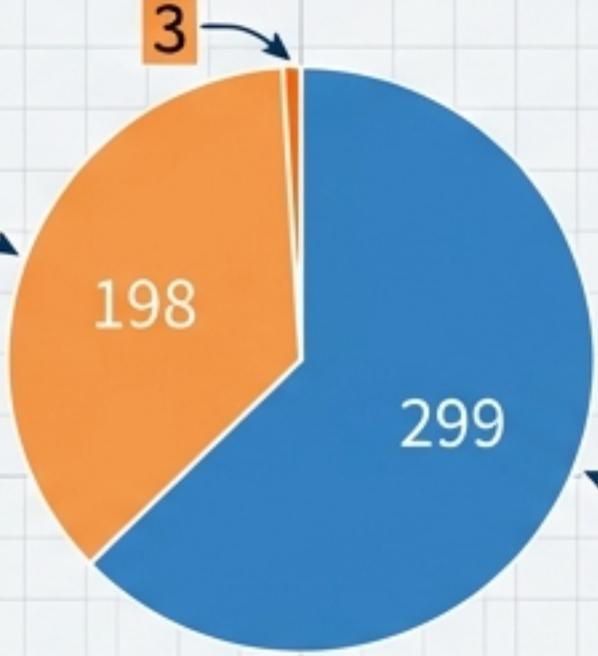
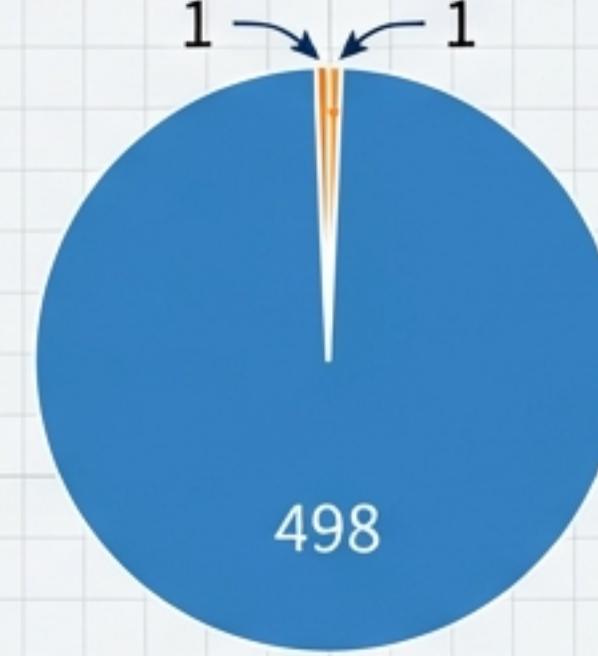
案例分析：配方樹狀圖解析 (Ward Linkage)

階層優勢

- High Level (K=3): 對應主要產品線 (工業、噴塗、裝飾)。
- Low Level (K=5+): 識別產品線內的特殊規格。



方法比較：為何選擇 Ward 連結法？

Ward 連結法 (Winner)	Average 連結法 (The Trap)	Single 連結法 (The Failure)
 <p>Sizes: 201 / 160 / 139</p>	 <p>Sizes: 299 / 198 / 3 Silhouette Score 最高 (0.38)，但產生孤立小群集。</p>	 <p>Sizes: 498 / 1 / 1 嚴重的鏈接效應 (Chain Effect)。</p>
<input checked="" type="checkbox"/> 實用 (Balanced)	⚠ 高分但無用	✗ 失敗

工程教訓：不能只看數學指標，必須確認群集大小的平衡性。

應用價值：從數據到決策



Cluster 0: 工業塗料

Traits: 高黏度 (4595 cP)，高成本 (\$146)

- Action: 專線生產，B2B 高價策略。



Cluster 1: 噴塗塗料

Traits: 高溶劑 (47%)，低黏度

- Action: 監控 VOC，經濟型定價。



Cluster 1: 噴塗塗料

Traits: 高溶劑 (47%)，低黏度

- Action: 監控 VOC，經濟型定價。



Cluster 2: 裝飾塗料

Traits: 快乾 (38 min)，多彩

- Action: B2C 零售，強調便利性。

營運優化

庫存共用原料

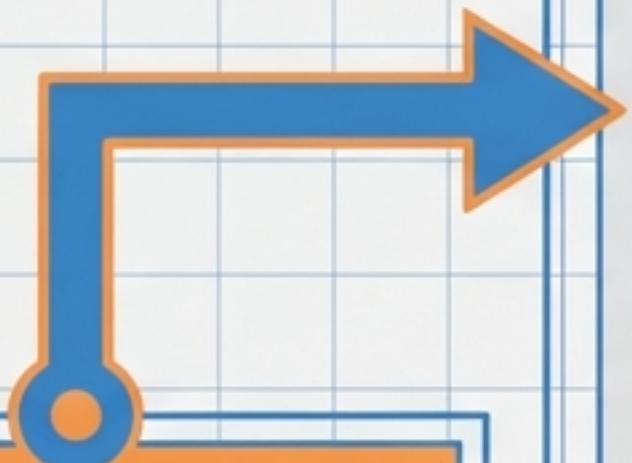
按黏度順序排程 (低->中->高)

差異化品管

實作指南：Scikit-Learn 程式碼範例

```
from sklearn.preprocessing import StandardScaler  
from scipy.cluster.hierarchy import dendrogram, linkage  
from sklearn.cluster import AgglomerativeClustering  
  
# 1. Standardize (Crucial for chemical data)  
# 確保黏度(cP)與成分(%)具有相同權重  
X_scaled = StandardScaler().fit_transform(df)  
  
# 2. Compute Linkage Matrix & Plot  
Z = linkage(X_scaled, method='ward')  
dendrogram(Z)  
  
# 3. Fit Model with chosen K  
model = AgglomerativeClustering(n_clusters=3)  
labels = model.fit_predict(X_scaled)
```

****關鍵步驟****：化工數據單位差異大，未標準化會導致錯誤。



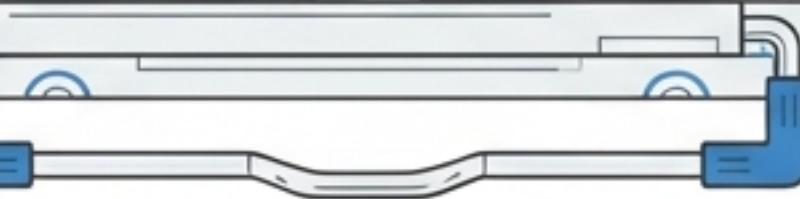
綜合評估：階層式分群 vs. K-Means

階層式 (Hierarchical)	K-Means
<ul style="list-style-type: none">• K 值：無需預先指定• 結構：提供層次結構 (Tree)• 計算：慢 ($O(n^2)$)• 適用：中小型數據 ($N < 10k$)	<ul style="list-style-type: none">• K 值：需預先指定• 結構：平坦結構 (Flat)• 計算：快 ($O(n)$)• 適用：大規模數據

品質指標 (Quality Metrics)

- Silhouette Score (-1 to 1)：凝聚度與分離度。
- Calinski-Harabasz：越高越好。

總結與最佳實踐 (Best Practices)



CHECKLIST

- 數據標準化**：確保不同單位的特徵具有相等權重。
- 首選 Ward 連結法**：除非有特殊形狀需求，否則從 Ward 開始。
- 善用樹狀圖**：尋找高度跳躍點 (Gaps) 來決定 K 。
- 驗證平衡性**：避免使用導致單一巨大群集的連結法 (如 Single)。
- 混合策略**：對於超大數據，先用 K-Means 分群，再對群中心進行階層分析。

"階層式分群是理解數據結構的藍圖 (Blueprint)，而不僅僅是分類工具。"