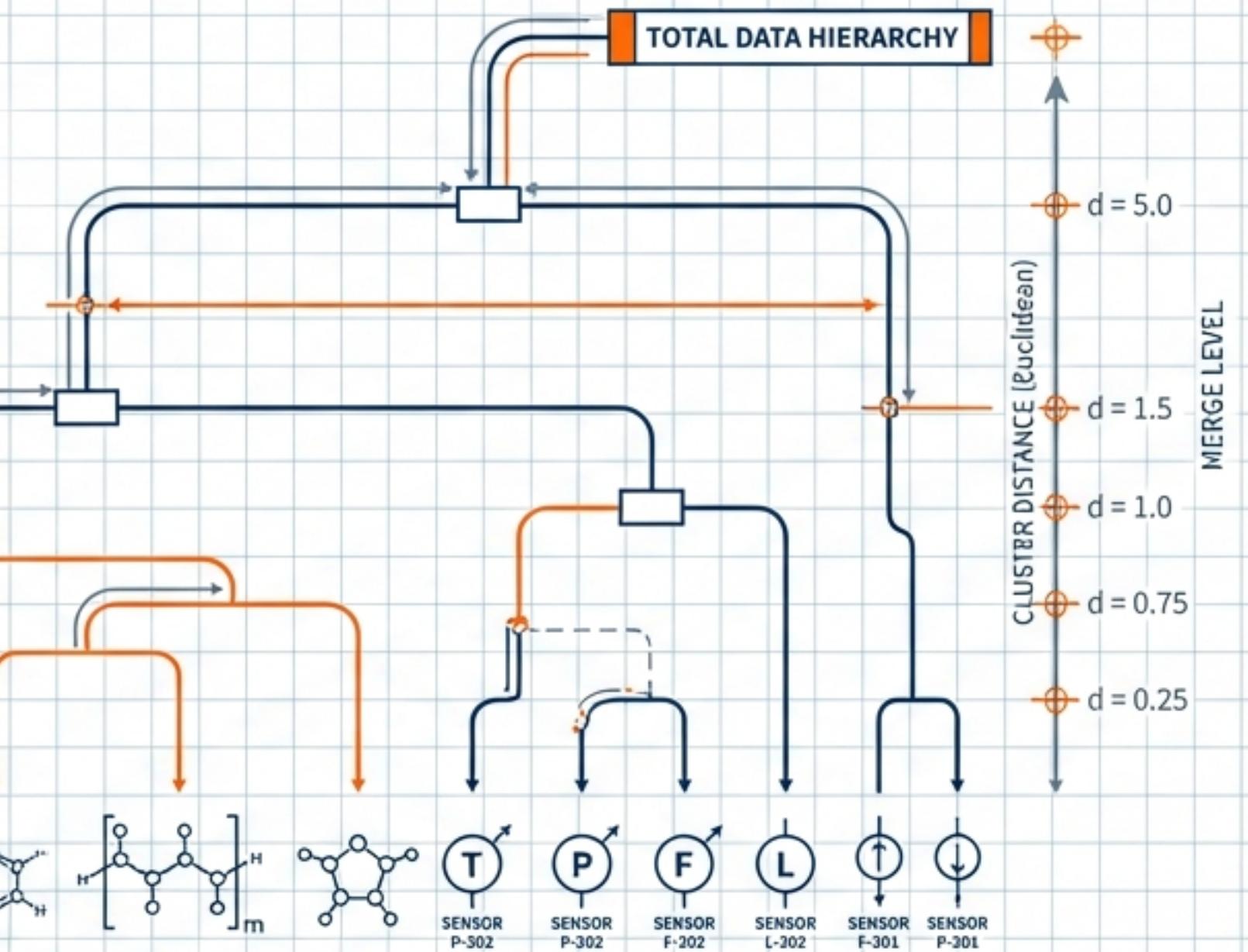
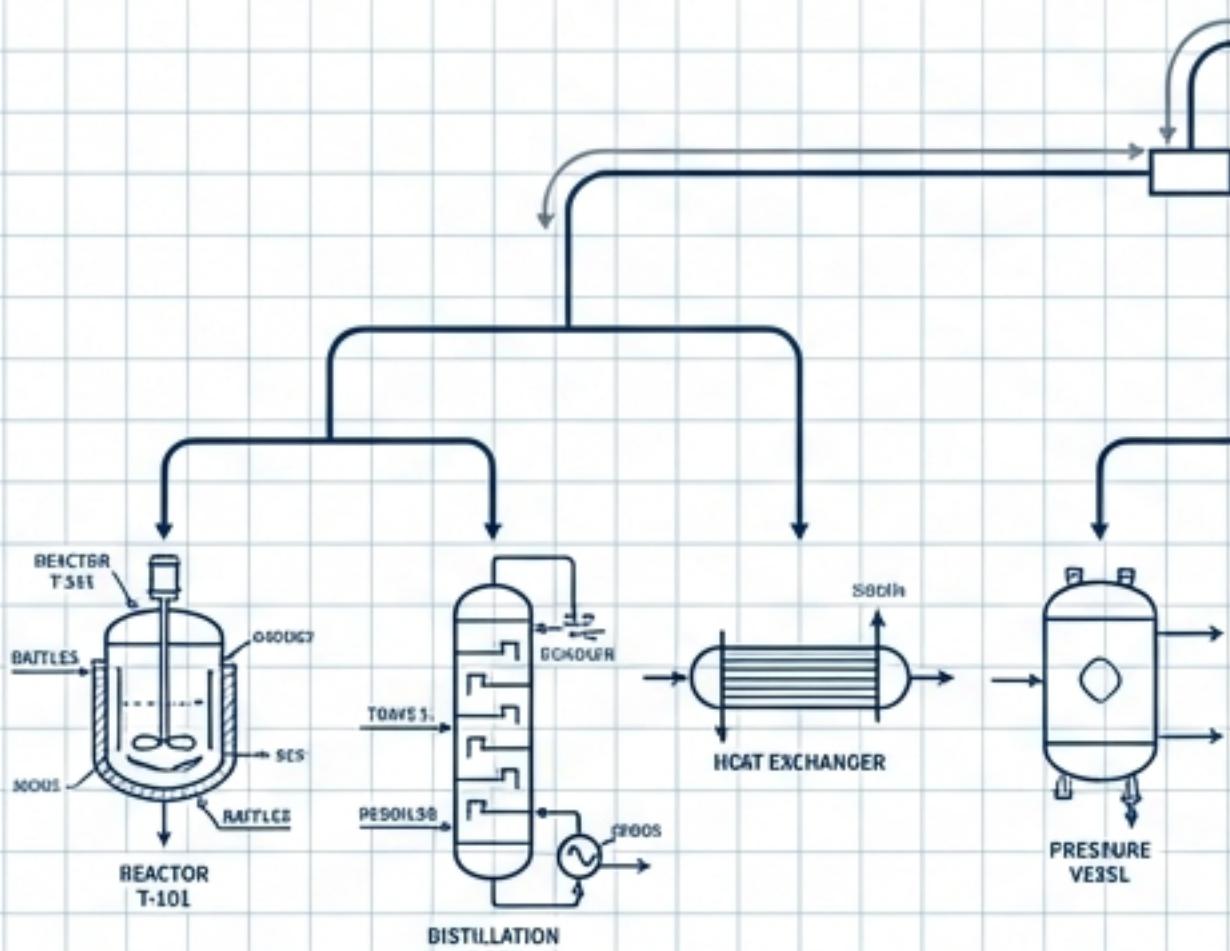


# AI 在化工上的應用：Unit 05

## 階層式分群演算法 (Hierarchical Clustering)

# 建立數據的層次結構與分類體系



授課教師：莊曜楨 助理教授 | 課程目標：理解數據層次結構，掌握樹狀圖 (Dendrogram) 解析與化工實務應用。

# 為什麼需要階層式結構？

## 傳統分群 (Traditional Clustering)

A scatter plot with 'Feature 1' on the horizontal axis and 'Feature 2' on the vertical axis. Three distinct groups of blue circular data points are visible. A large orange question mark is centered over the plot, with the text 'K=3?' written below it in orange.

「數據自然具有層次，  
我們不僅要分群，  
更要了解它們的  
『演化關係』。」

需預先指定  $K$  值。  
假設群集為扁平結構，忽略  
了數據內部的演化關係。

## 階層式分群 (Hierarchical Clustering)

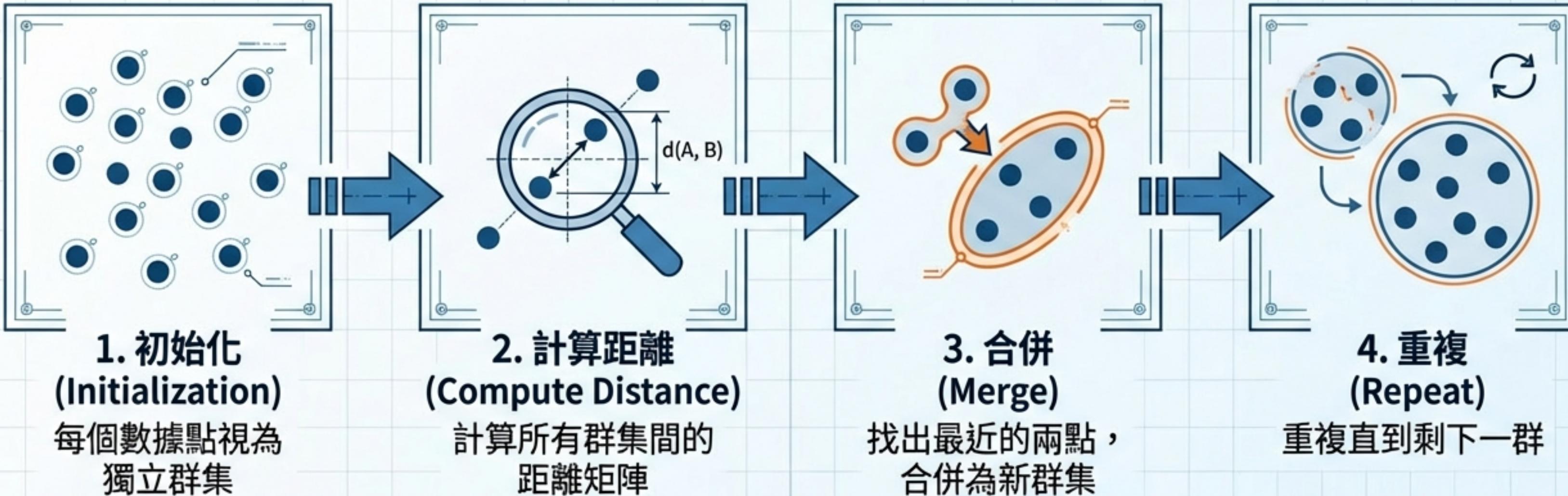
A dendrogram illustrating a nested hierarchical structure. The vertical axis is labeled 'Distance / Similarity'. The horizontal axis lists data points P1 through P30. The dendrogram shows the merging of clusters at various similarity levels, with orange lines indicating the formation of larger clusters from smaller ones. A legend on the right indicates that orange lines represent the 'Nested Hierarchical Structure'.

無需指定  $K$  值。  
建立樹狀圖 (Dendrogram)，揭  
示數據的自然層次與演化路徑。

NotebookLM

# 演算法核心：凝聚式策略 (Agglomerative Strategy)

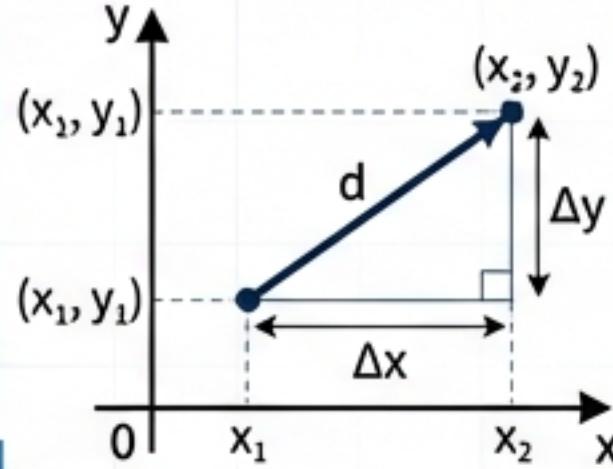
由下而上的建構過程 (Bottom-Up Approach)



迭代過程： $(C_p, C_q) = \operatorname{argmin} d(C_i, C_j)$ 。時間複雜度： $O(n^2 \log n)$ 。

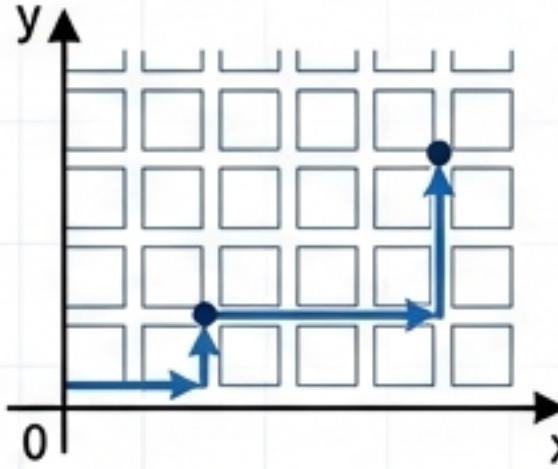
# 距離度量：定義「相似」的物理意義 (Distance Metrics)

## 歐幾里得距離 (Euclidean)



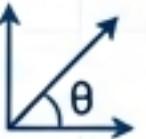
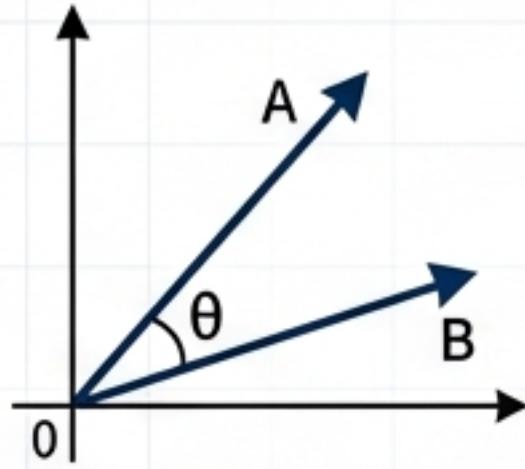
適用：物理意義相同、尺度一致的變數 (如：溫度/壓力)。  
公式： $\sqrt{\sum (x - y)^2}$

## 曼哈頓距離 (Manhattan)



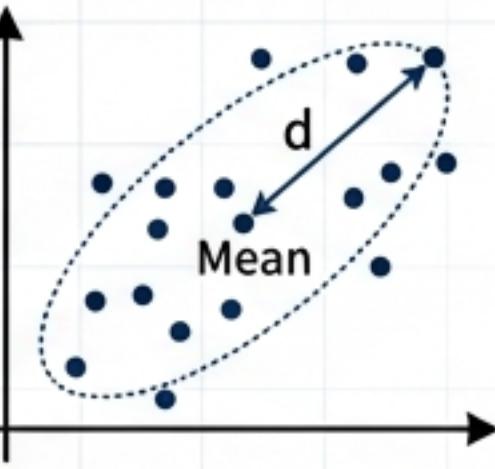
適用：對離群值較穩健，數據分布偏斜時。  
公式： $\sum |x - y|$

## 餘弦距離 (Cosine)



適用：關注「方向」與成分比例 (Composition)。  
公式： $1 - \cos(\theta)$

## 馬氏距離 (Mahalanobis)



適用：考慮變數間的相關性 (Correlation)。

工程筆記：不同單位的變數 (如  $^{\circ}\text{C}$  vs  $\text{kg/hr}$ ) 必須先進行標準化 (Standardization)。

# 連結方法：決定群集的形狀 (Linkage Methods)

## Method (方法)

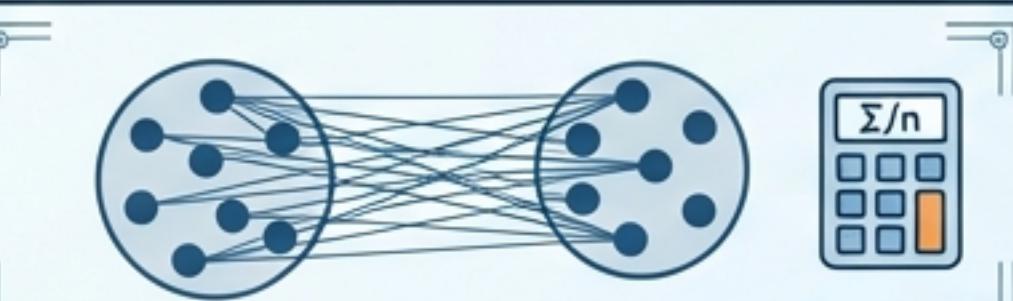
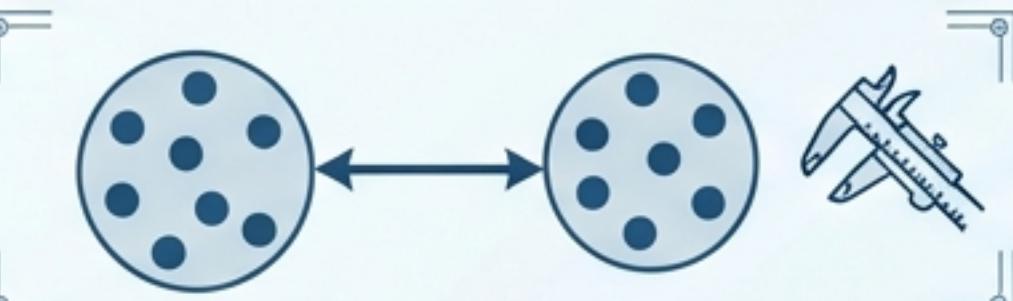
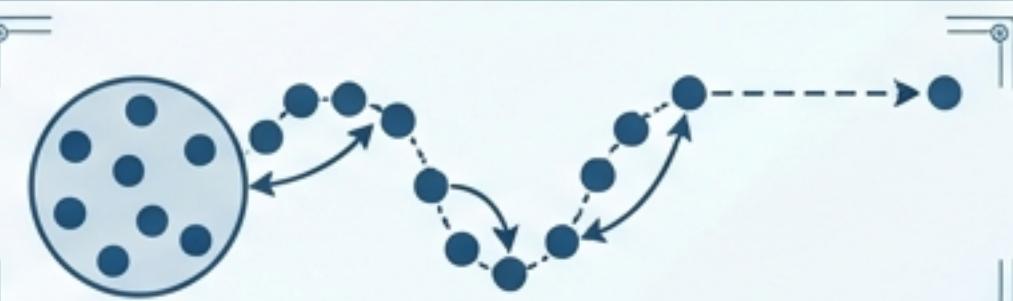
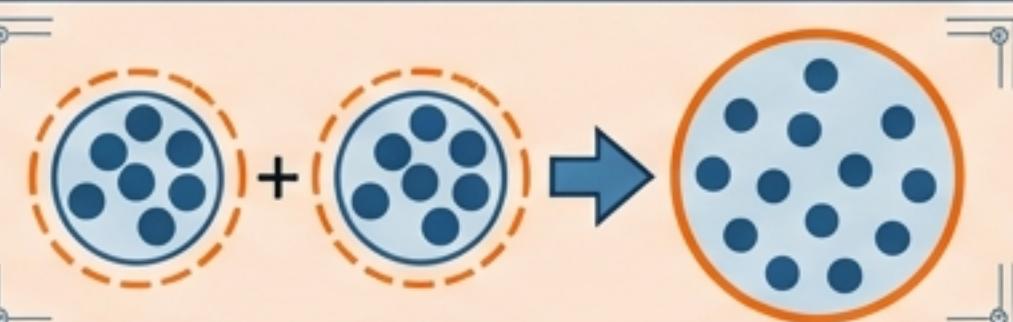
**Ward (華德法) [化工首選]**  
最小化群集內變異數  
(Minimize Intra-cluster Variance)

**Single (最小連結)**

**Complete (最大連結)**

**Average (平均連結)**

## Visual Logic (視覺邏輯)



## Characteristics (特性)

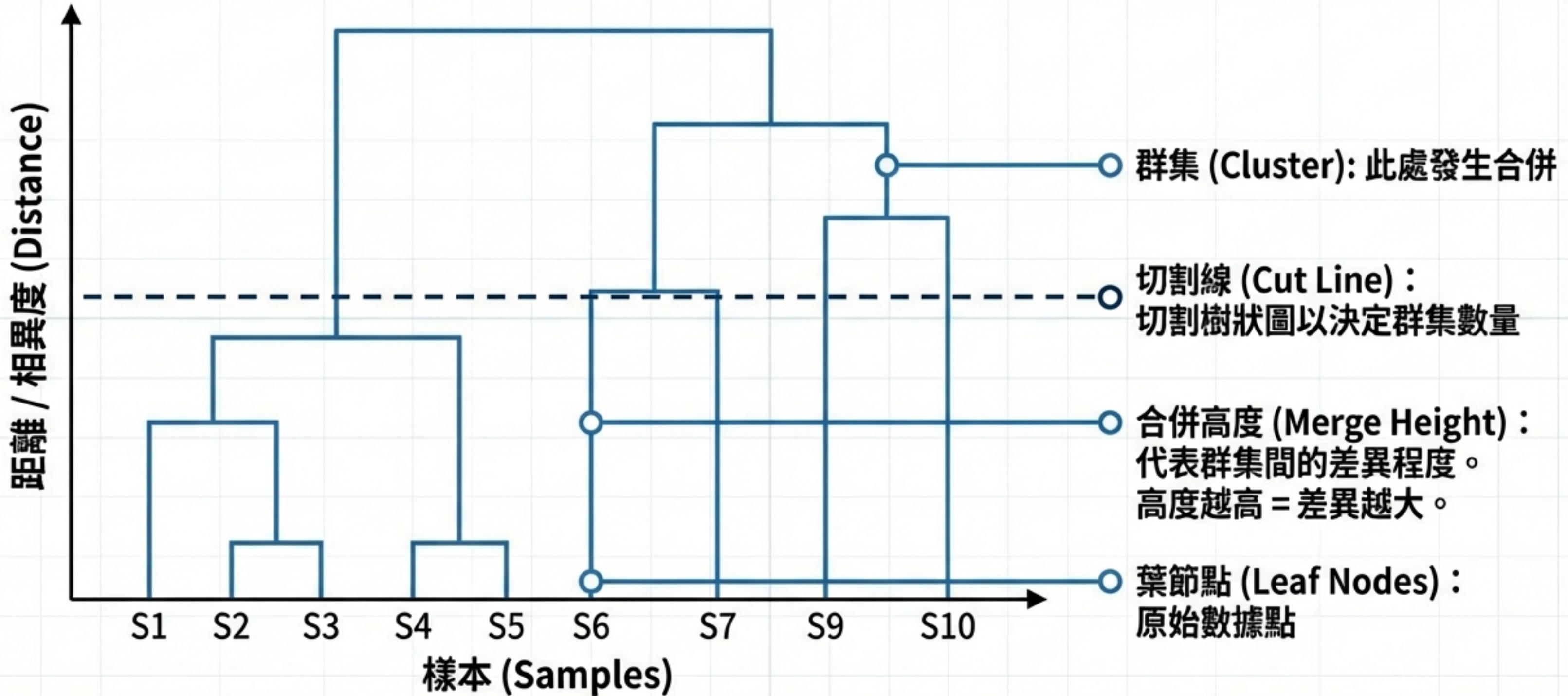
產生大小平衡、緊湊的球形群集

**最近點距離**  
易產生細長型群集與「鏈接效應」，受噪音影響大

**最遠點距離**  
產生緊湊球形，但受離群值影響

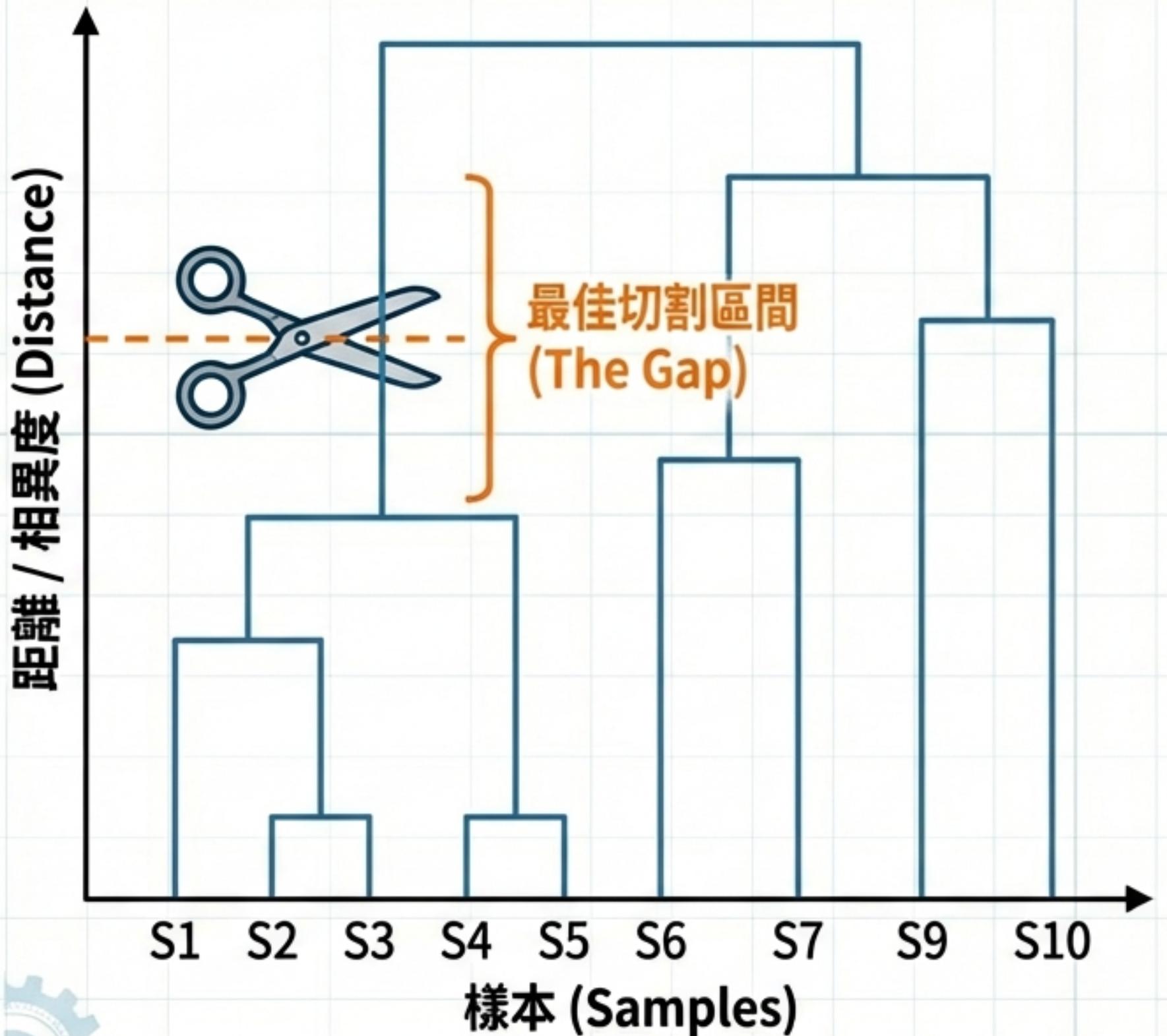
**所有點對平均距離**  
折衷方案，計算量較大

# 解讀樹狀圖 (Deciphering the Dendrogram)



樹狀圖不僅是圖表，它是數據結構的地圖。觀察高度的『跳躍 (Jump)』是決定分類數量的關鍵。

# 決策時刻：如何切割樹狀圖？(Cutting the Tree)



1. 觀察高度跳躍 (The Gap)
  - 尋找 Y 軸上垂直線最長的一段區間。這代表群集間有顯著差異，是自然的物理邊界。
2. 業務邏輯指定 (Business Logic)
  - 根據實際需求 (如：需要分 3 級產品) 直接指定  $K$  值。
3. 距離閾值 (Threshold)
  - 設定固定的相異度  $t$ ，低於此差異視為同類。
4. 輪廓係數 (Silhouette Analysis)
  - 計算不同  $K$  的分數，尋找統計最優解。

切割高度決定了分類的粗細度：  
切得越高 → 分類越粗 (大類)；  
切得越低 → 分類越細 (子類)。

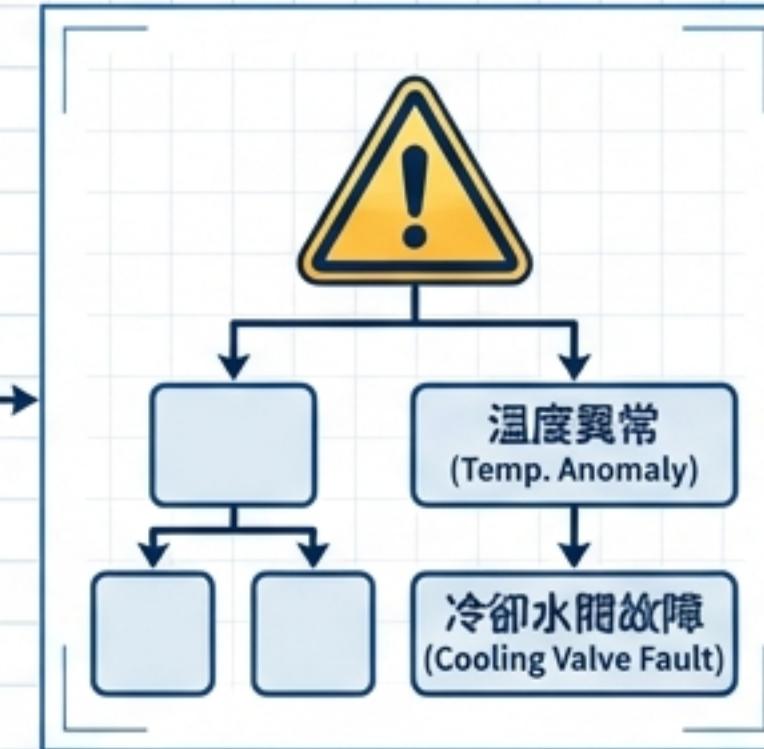


# 化工領域的實際應用場景 (ChemE Applications)



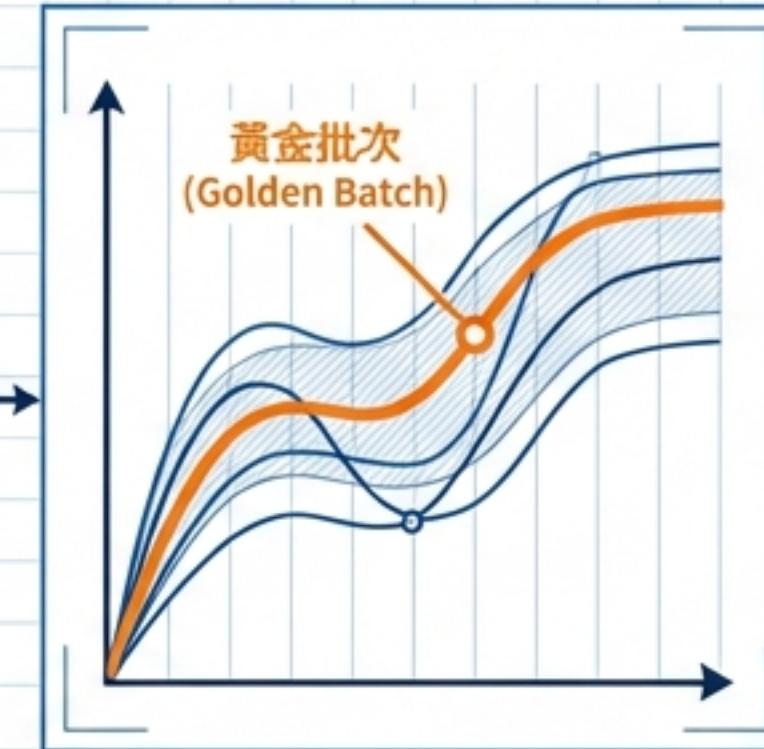
## 1. 產品分類體系

根據配方性質建立階層  
**(大類/中類/小類)**。幫助新產品開發與庫存管理。



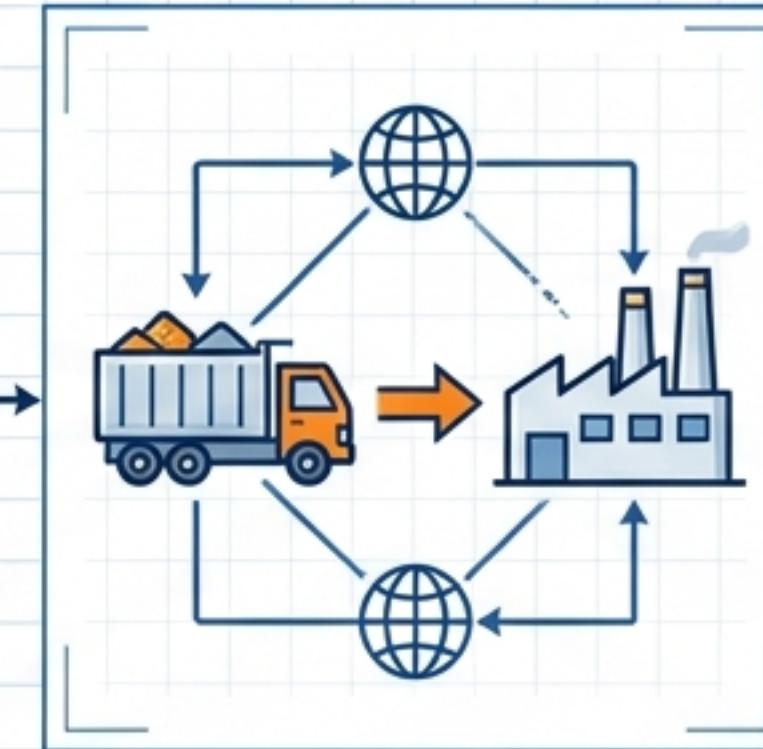
## 2. 故障診斷 (Fault Diagnosis)

建立故障樹。從「溫度異常」細分到「冷卻水閥故障」。



## 3. 批次製程分析

識別**黃金批次 (Golden Batch)** 與**異常批次**，分析批次間相似度。



## 4. 供應鏈管理

根據原料性質對供應商進行分層管理，尋找替代料源。

# 案例研究：塗料配方自動分類 (Case Study)

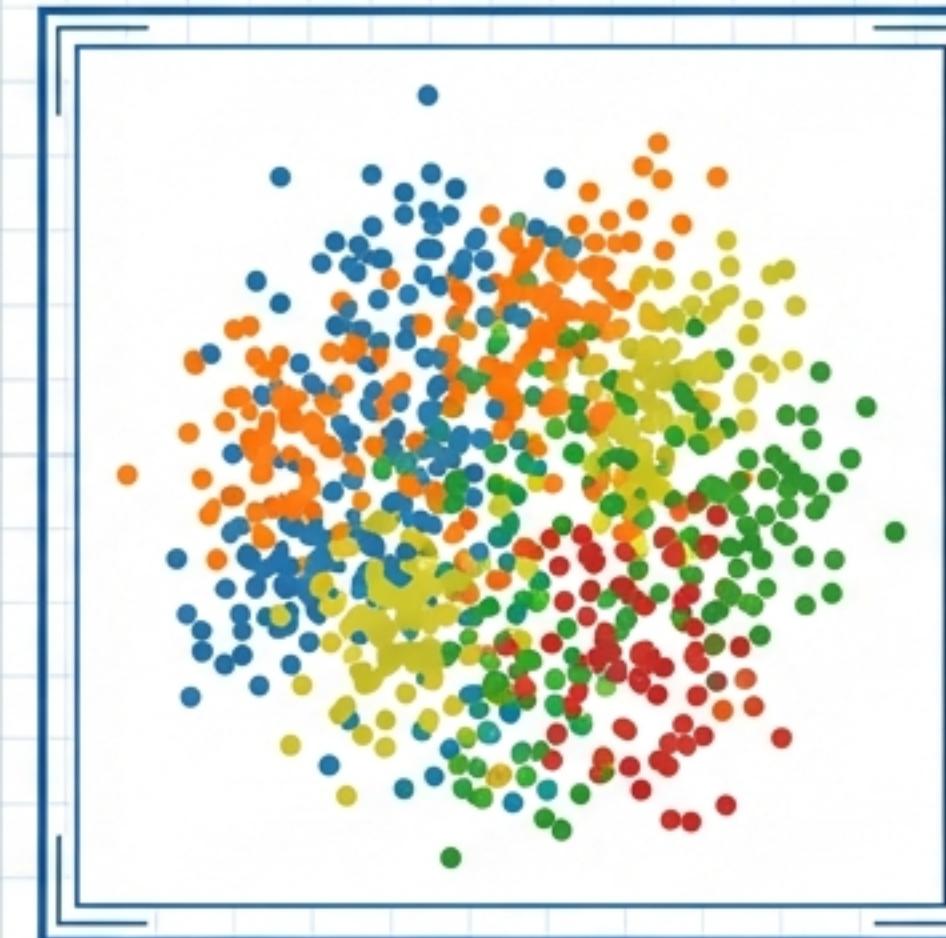
## Data Input Specification

**目標 (Goal) :** 利用階層式分群，將 500 筆混亂的配方數據自動建立分類體系。

**數據集 (Dataset) :**  
500 Samples (Formulations)

**特徵維度 (8 Features) :**  
\* 成分: Resin, Solvent, Pigment, Additive  
\* 性質: Solid Content, Viscosity, Drying Time  
\* 經濟: Cost (\$/kg)

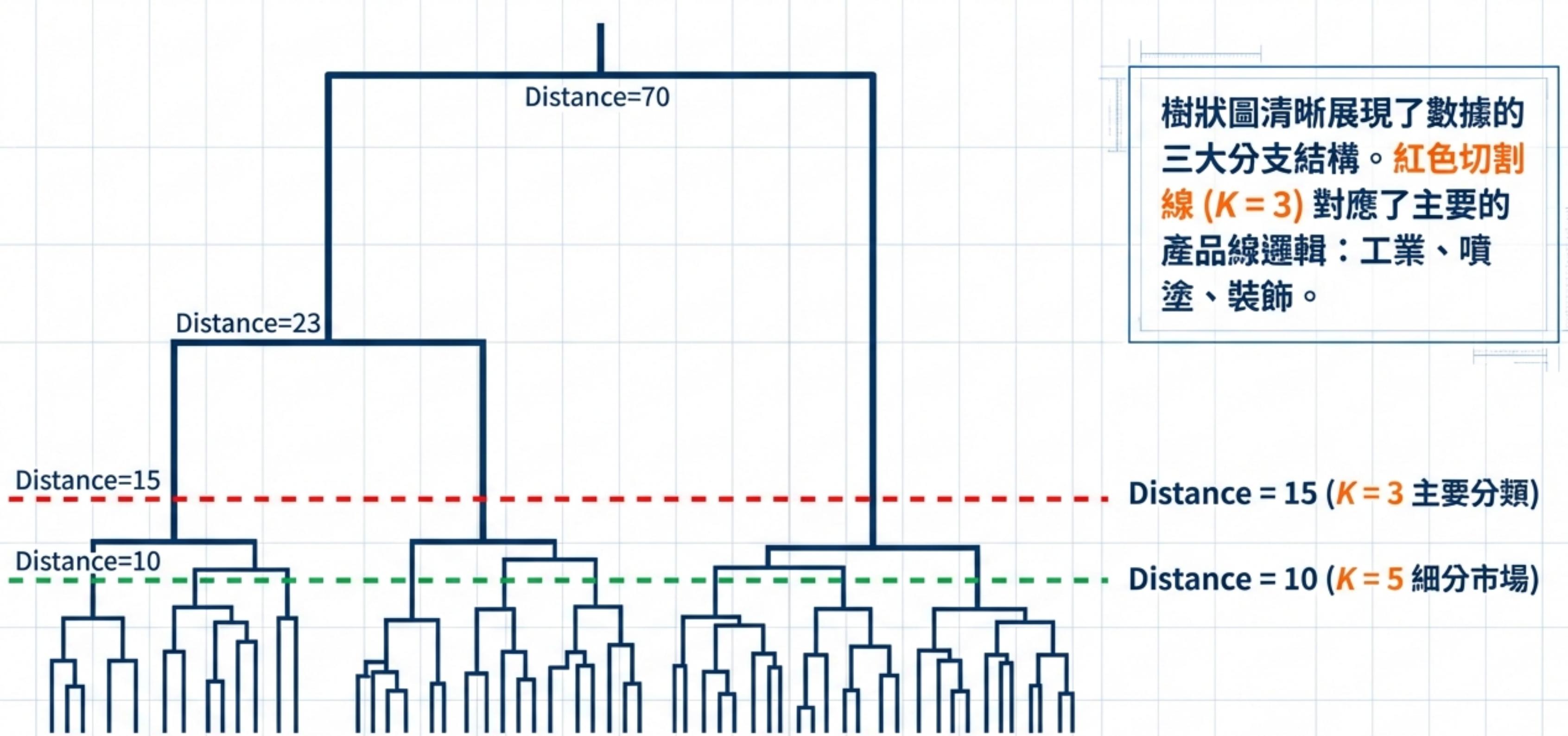
## 原始數據 (Raw Data)



如何發現結構？

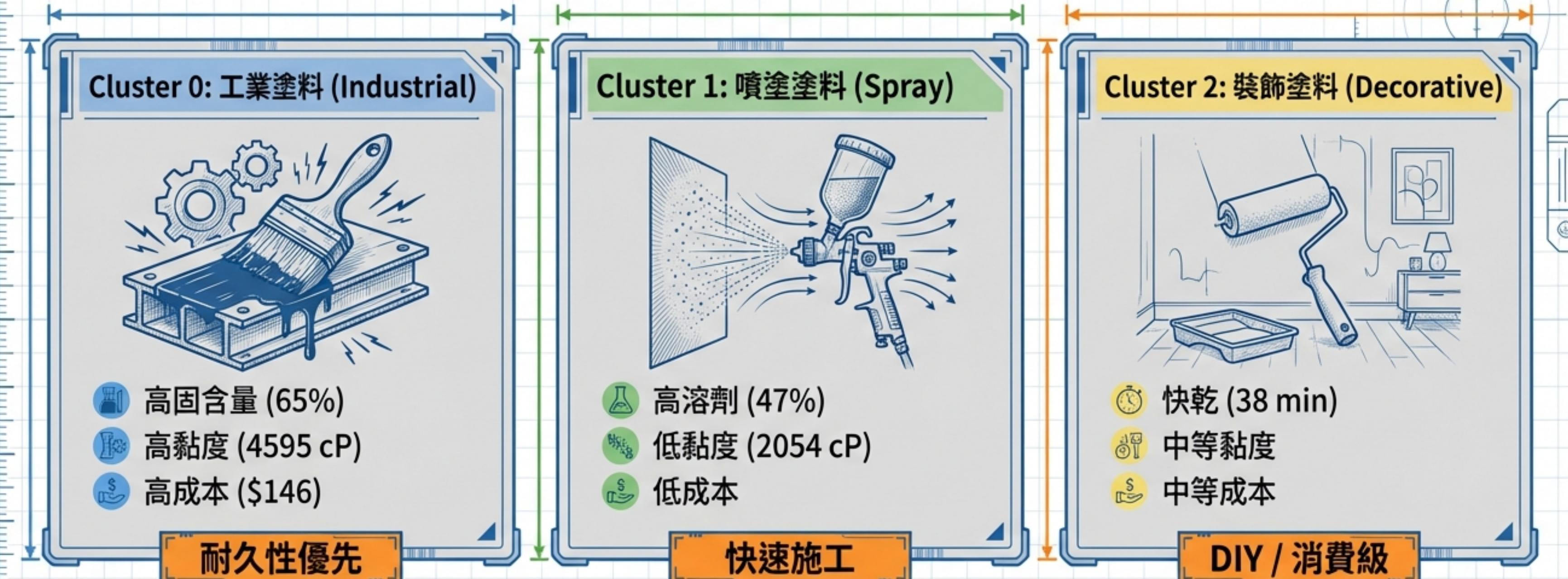
數據存在重疊 (Overlap) 與多維度交互作用，人工分類極為困難。

# 從數據到結構：配方樹狀圖 (From Data to Structure)



樹狀圖清晰展現了數據的三大分支結構。紅色切割線 ( $K=3$ ) 對應了主要的產品線邏輯：工業、噴塗、裝飾。

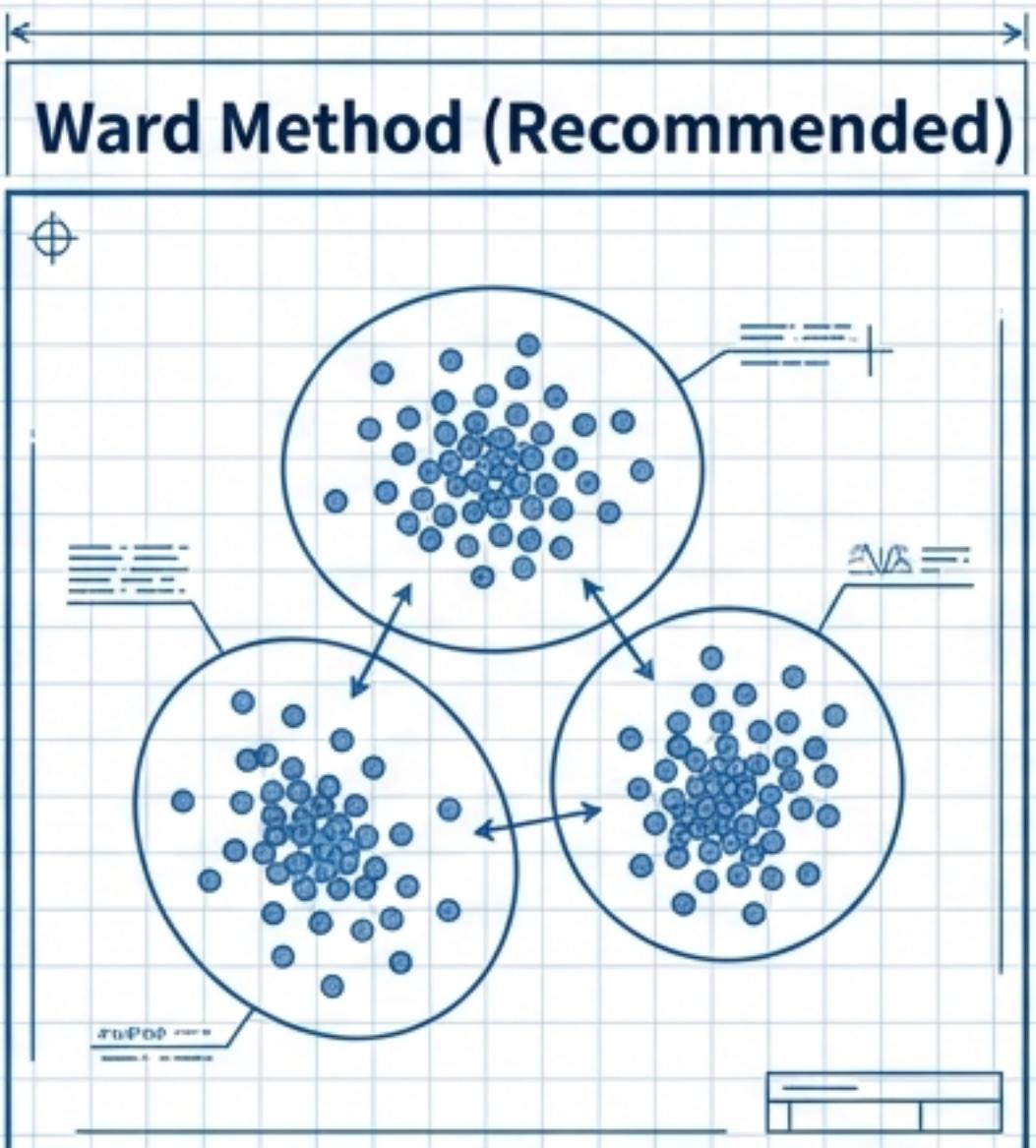
# 解碼群集：AI 發現了什麼？(Decoding the Clusters)



Eureka! 演算法在沒有標籤的情況下，成功根據物理性質重新發現了產品的市場定位。

# 連結方法的影響：為何選擇 Ward？

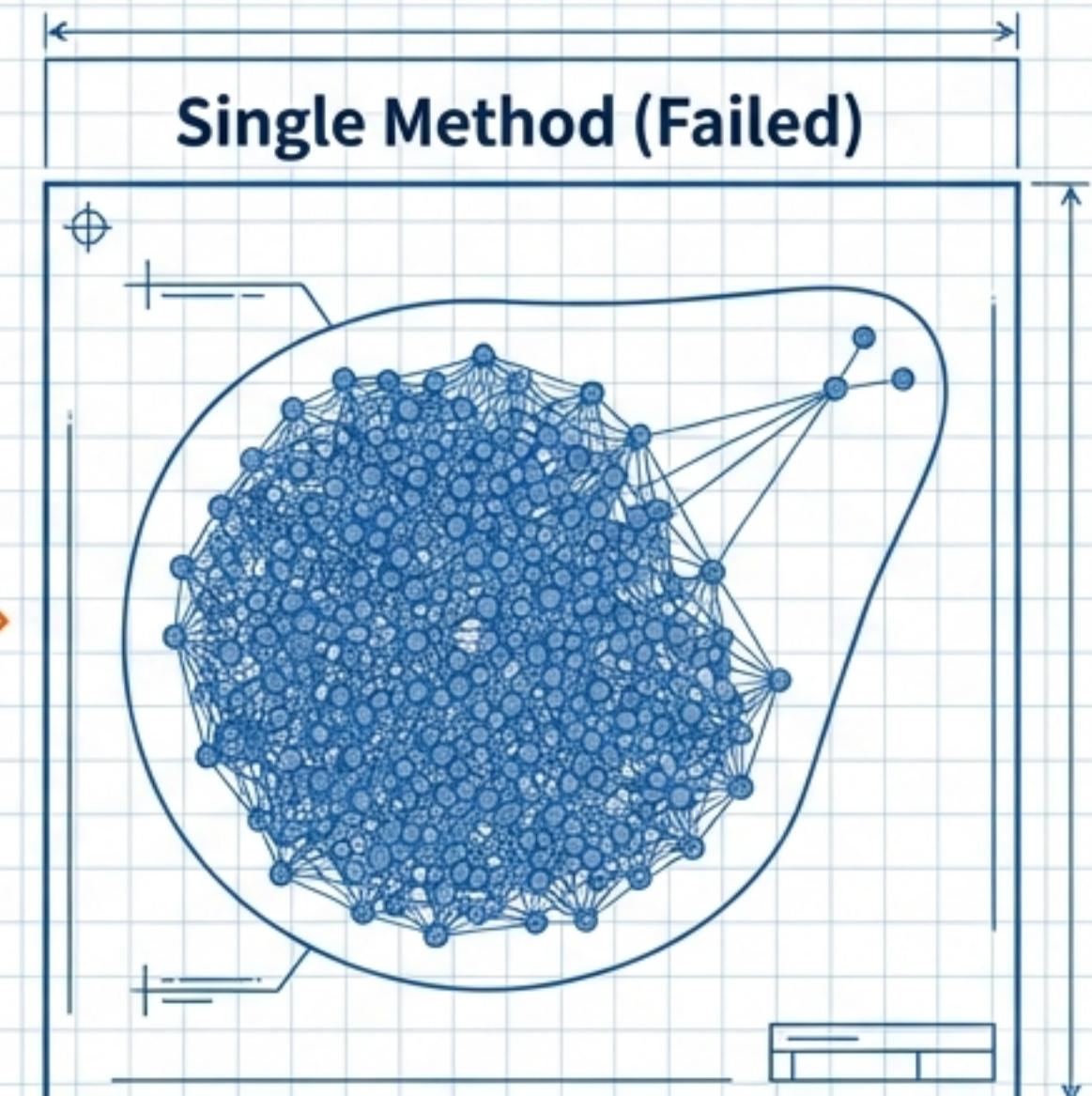
避免鏈接效應 (Avoid Chain Effect)



+

**關鍵教訓：不要只看數學指標 (Metrics)。必須檢查群集的大小平衡性 (Balance) 與物理意義。**

A central text box with an orange border contains the following text: "關鍵教訓：不要只看數學指標 (Metrics)。必須檢查群集的大小平衡性 (Balance) 與物理意義。" (Key Lesson: Do not only look at mathematical metrics. Must check cluster size balance and physical meaning.) The text is surrounded by orange arrows pointing towards it, and there are orange icons of gears, a lightbulb, and a balance scale at the top and bottom.



群集大小平衡 (201 / 160 / 139)。  
結構清晰，符合業務需求。

極度不平衡 (498 / 1 / 1)。發生「鏈接效應」，將絕大多數樣本視為同一群。

# 決策制定：統計最佳解 vs. 業務需求



Level 1  
(Business Logic)

所有配方  
(All Formulations)

Level 2  
(Sub-categories)

業務最優 ( $K = 3\$$ )

工業塗料

噴塗塗料

裝飾塗料

高階工業 標準工業

汽車噴塗 家具噴塗

室內牆面 戶外防護

統計細分 ( $K = 6\$$ )



「AI 提供地圖，工程師決定目的地。」



## 關鍵洞察 (Key Insights)

- 統計建議：Silhouette score 建議  $K = 2$  (數學最優)。
- 業務現實： $K = 2$  無法區分噴塗與裝飾市場。
- 解決方案：採用多層次分類體系。



# 工具選擇指南：Hierarchical vs. K-Means

## Hierarchical Clustering



### PROS

無需指定  $K$  | 提供層次結構 | 結果可重現



### CONS

計算量大 ( $O(n^2)$ ) | 僅適合中小數據 ( $N < 10k$ )



### BEST FOR

探索性分析、建立分類體系

## K-Means



### PROS

計算速度快 | 適合大數據



### CONS

需預設  $K$  | 結果受隨機初始影響

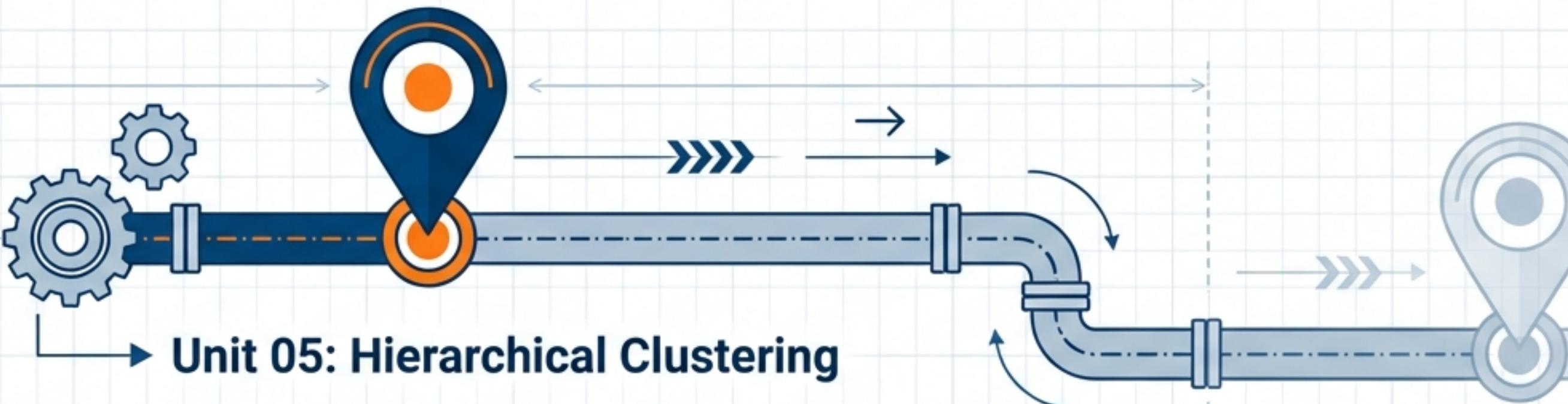


### BEST FOR

大規模分群、已知類別數量

最佳實踐：先用 Hierarchical 探索結構與  $K$  值，再用 K-Means 處理大規模數據。

# 總結與下一步 (Summary & Roadmap)



## Unit 05: Hierarchical Clustering

1. **結構化**: 樹狀圖是解讀數據層次的鑰匙。
2. **參數化**: Ward Linkage + Euclidean 是化工黃金組合。
3. **決策化**: 結合統計指標與領域知識決定  $K$ 。

## Unit 06: DBSCAN

學習處理不規則形狀與含噪音數據

從混亂到有序：階層式分群是理解複雜化工數據結構的第一步。