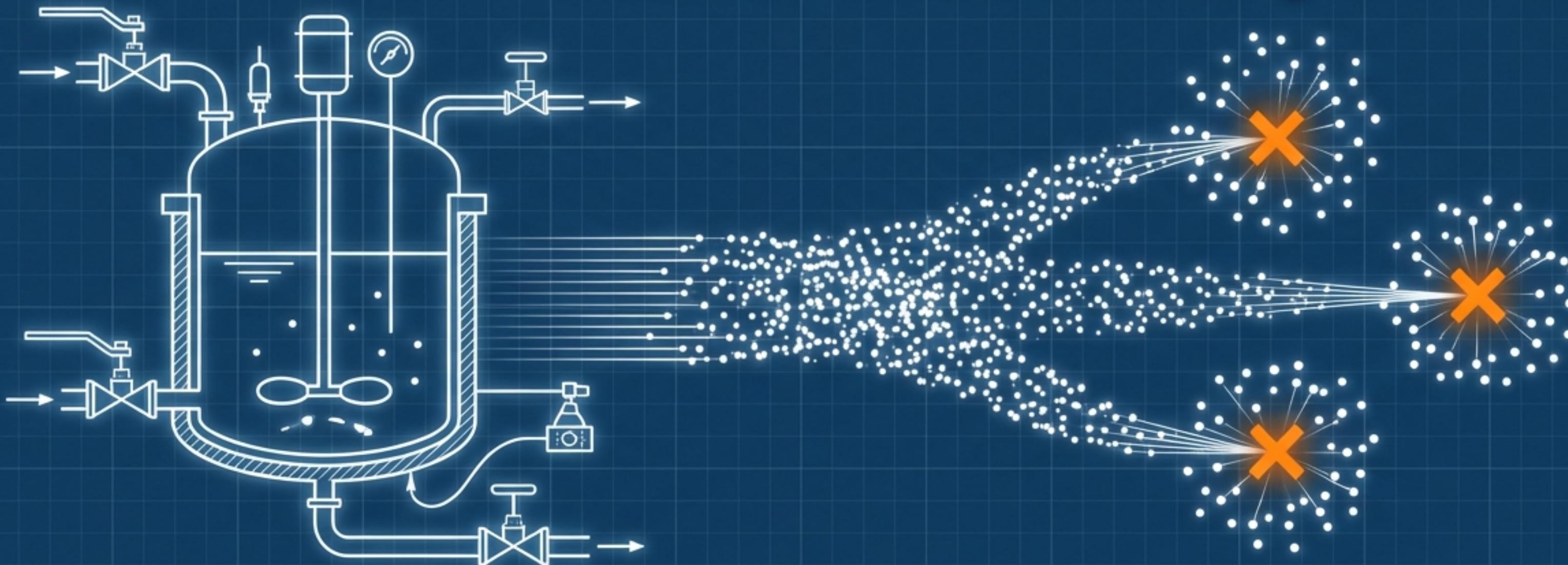


Unit 05

# K-Means 分群演算法

## AI 在化工上的應用：非監督式學習實務

從混沌數據中發現秩序 (Uncovering Order in Chaos)



# 典範轉移：從預測 (Prediction) 到發現 (Discovery)

## 監督式學習 (Supervised Learning)



**數據 (Data) :**

Input X + Output y (有標籤數據)

**目標 (Goal) :**

預測 (Prediction)

**比喻 (Analogy) :**

像是跟著老師學習標準答案。

## 非監督式學習 (Unsupervised Learning)



**數據 (Data) :**

Input X only (無標籤數據)

**目標 (Goal) :**

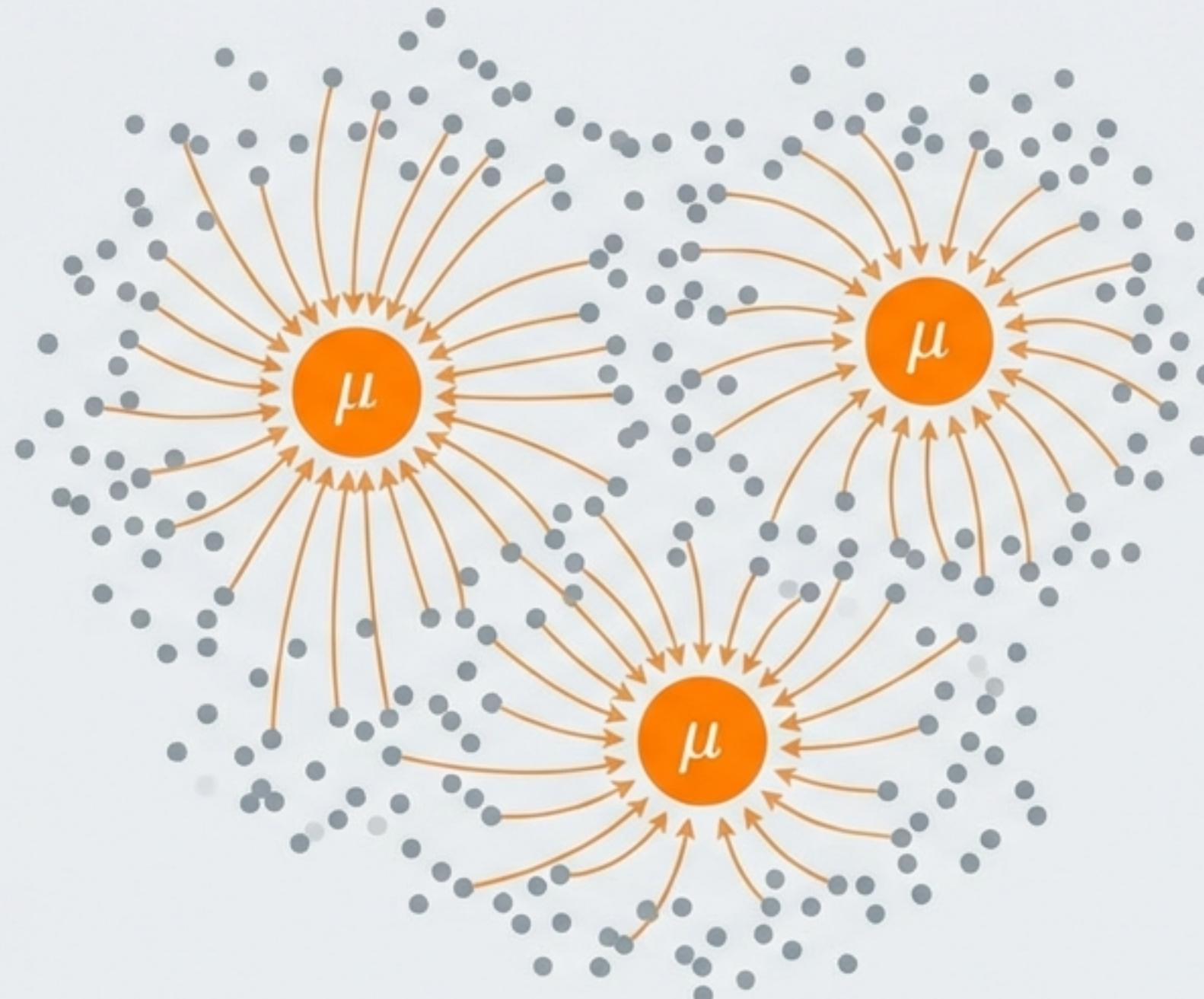
發現模式 (Discovery)

**比喻 (Analogy) :**

像是在未知領域中繪製地圖。

在化工實務中，我們常擁有海量數據 (Sensor Data)，卻缺乏標籤 (Labels)。  
K-Means 幫助我們定義這些數據的『狀態』。

# 核心機制：尋找數據的重力中心



**群集中心 (Centroid)**

每個群集的代表位置 ( $\mu_k$ )。

**歐幾里得距離 (Euclidean Distance)**

衡量相似度的尺規 ( $d(x, \mu) = \|x - \mu\|^2$ )。

**慣性 (Inertia/WCSS)**

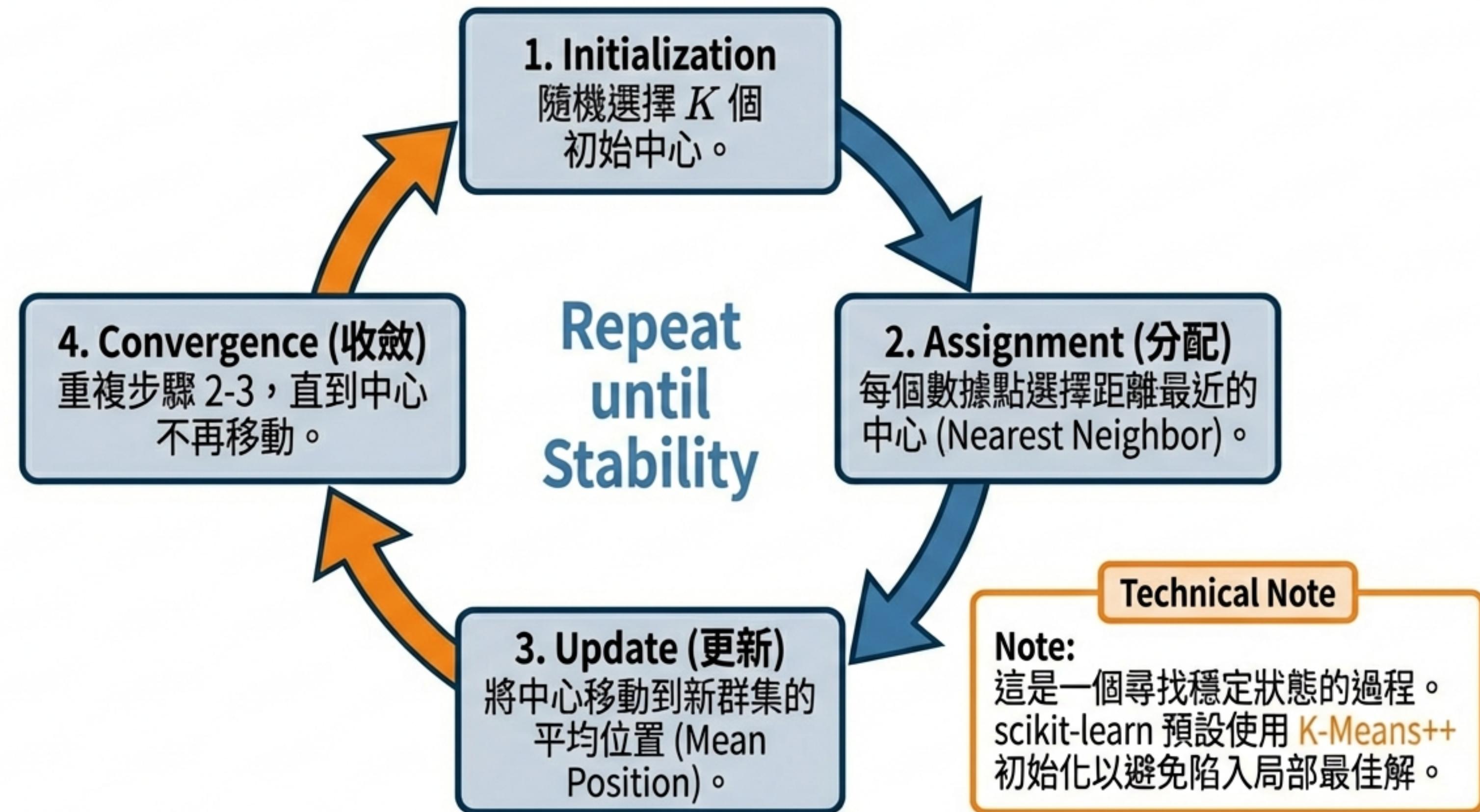
目標函數。我們希望群集內部越緊密越好  
(Minimize Variance)。

**Blueprint Specification**

Objective Function:  $J = \sum_{i=1}^n \|x_i - \mu_{c_i}\|^2 \rightarrow \min$

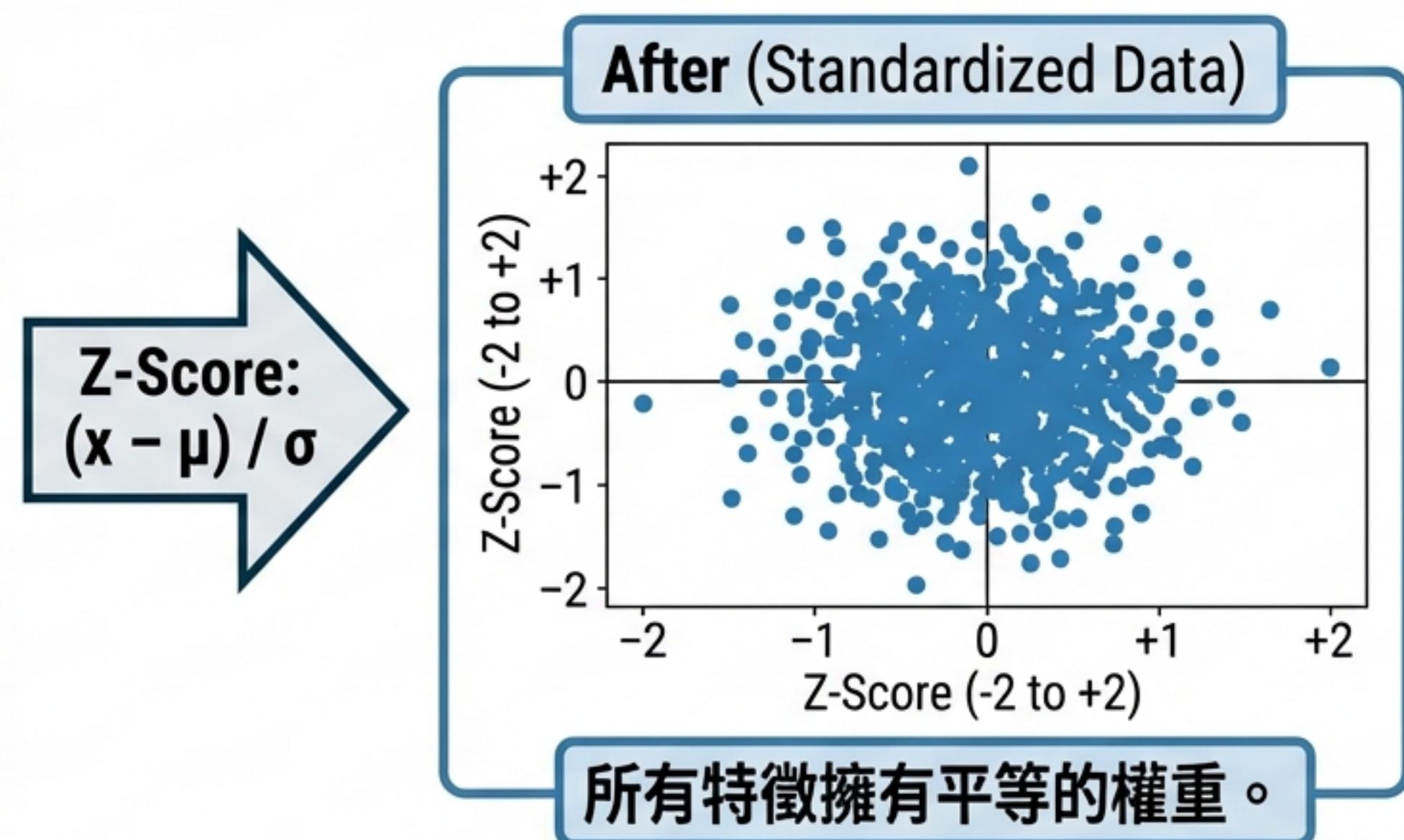
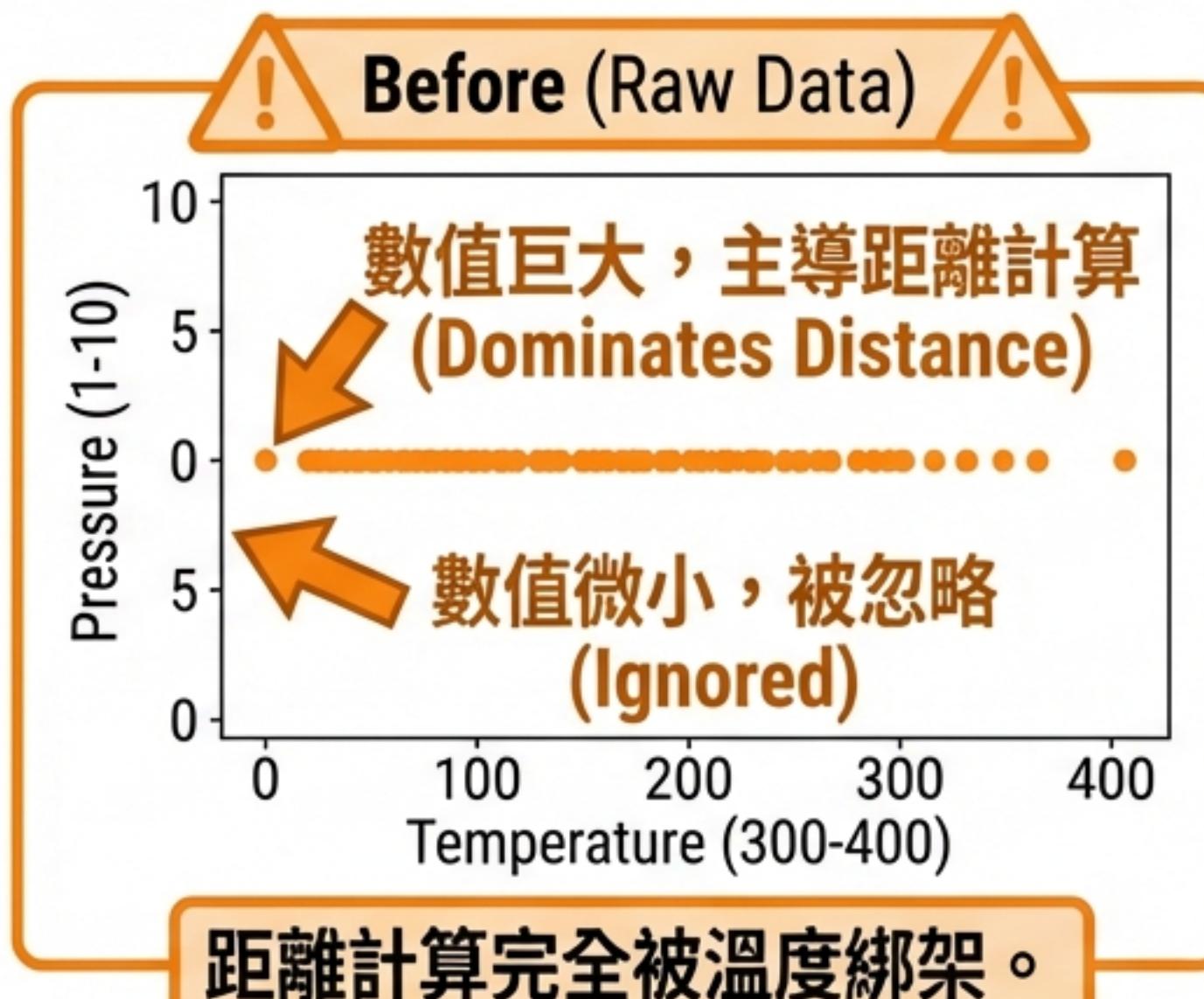
演算法的目標很簡單：讓群集內的差異最小化，群集間的差異最大化。

# 演算法流程：迭代優化循環 (Iterative Optimization)



# 關鍵前處理：單位與尺度的陷阱

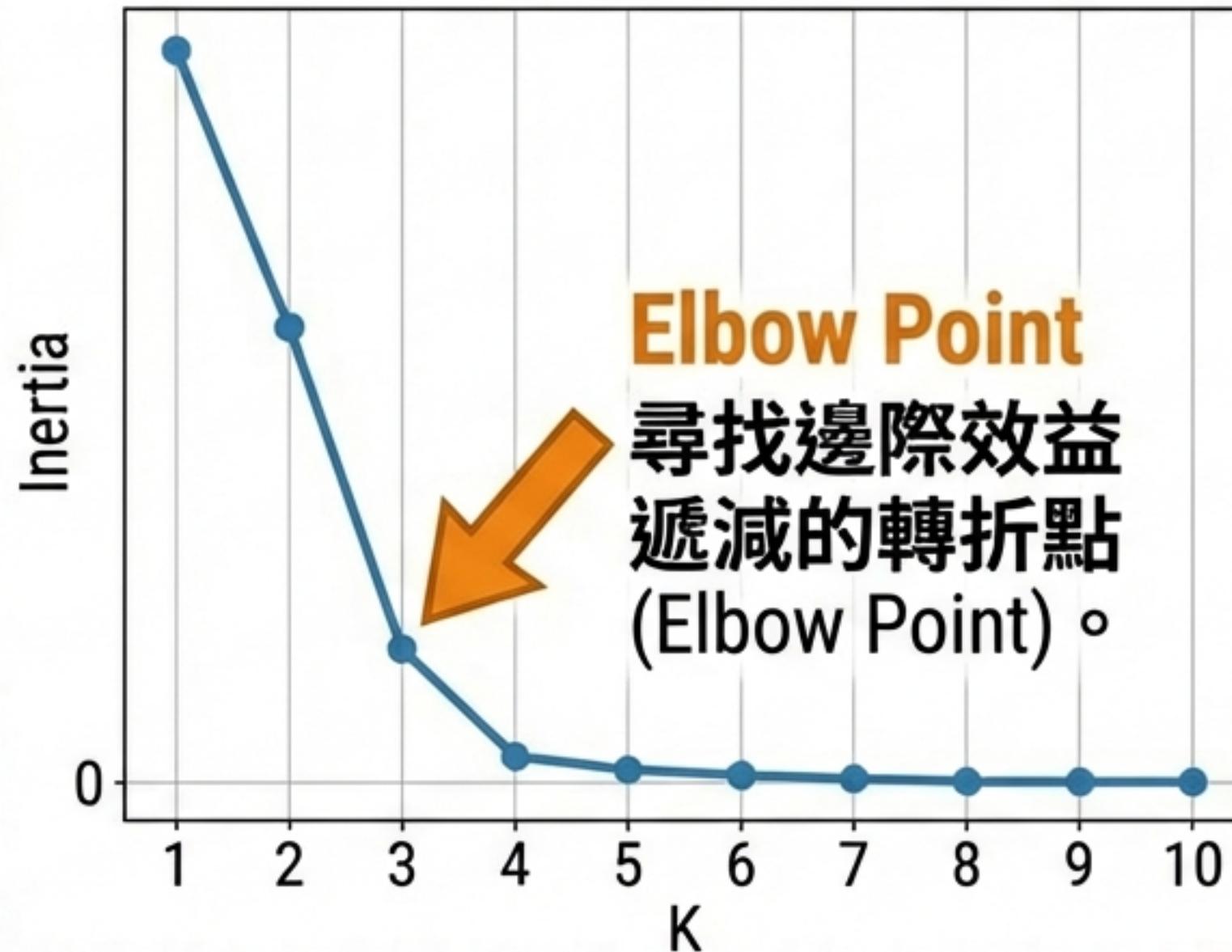
原則：K-Means 計算的是距離，因此特徵尺度 (Scale) 決定一切。



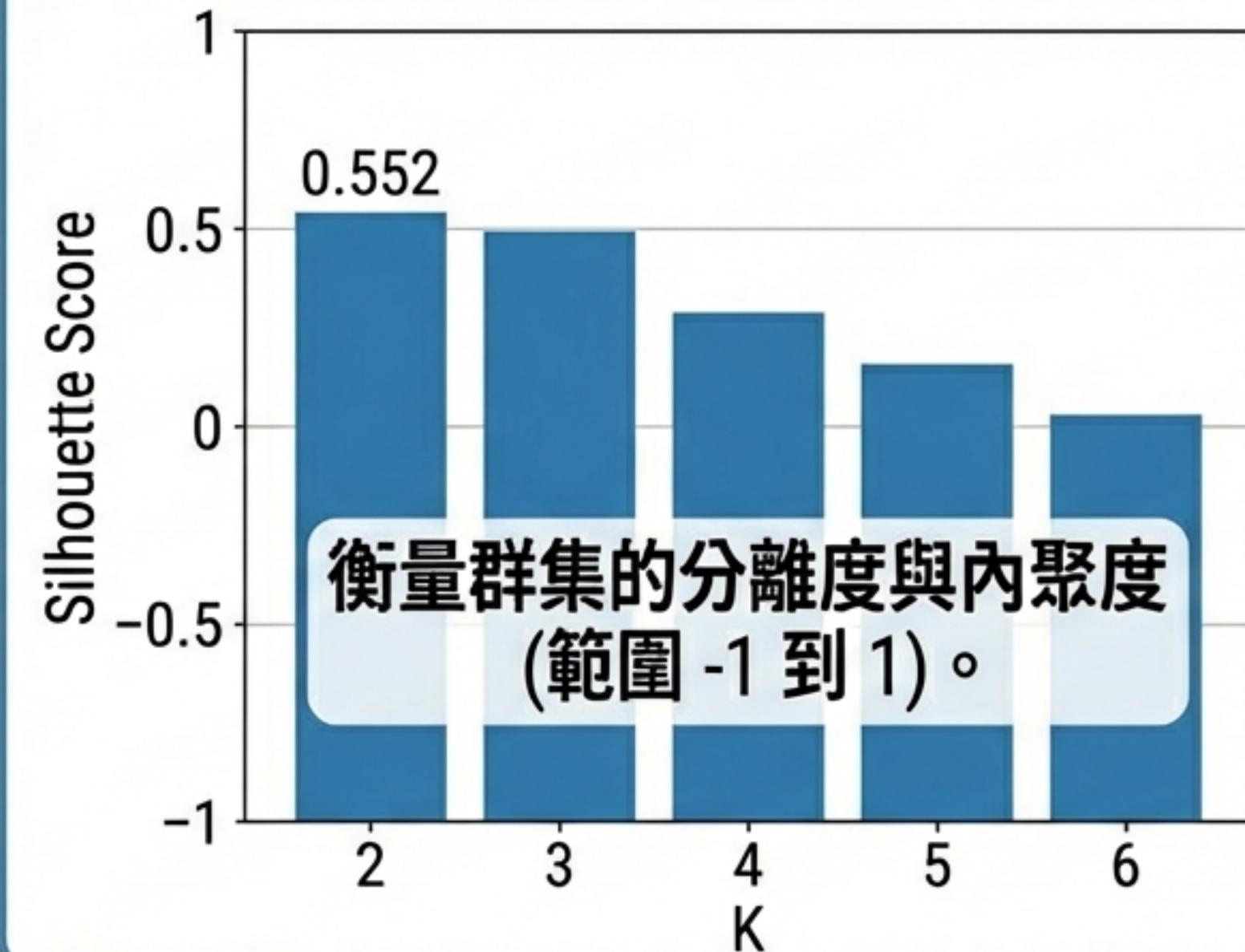
```
scaler = StandardScaler()  
X_scaled = scaler.fit_transform(X)
```

# 決策難題：如何選擇最佳的 K 值？

## 手肘法 (Elbow Method)

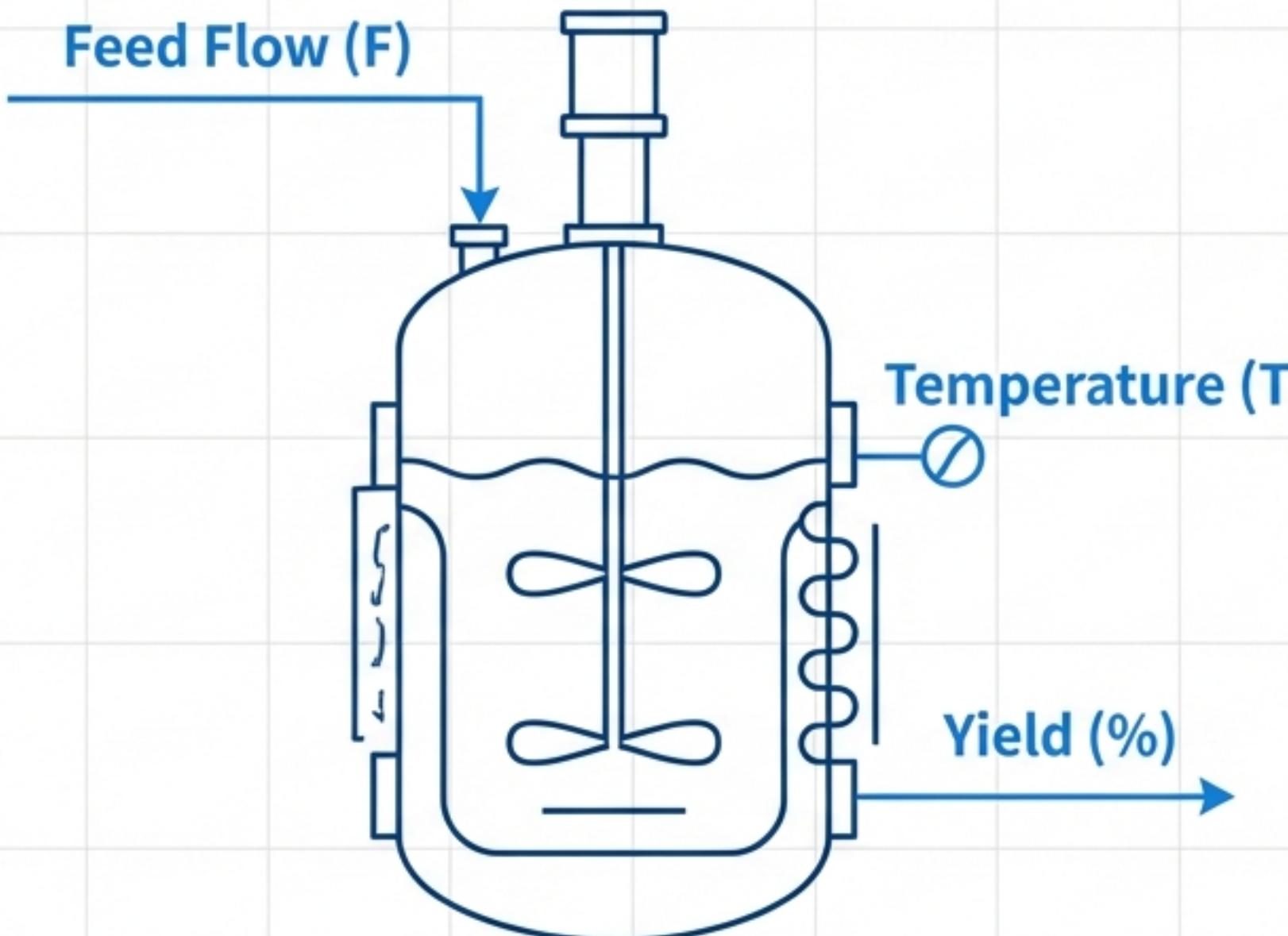


## 輪廓分析 (Silhouette Analysis)



統計指標提供了客觀建議，但通常沒有唯一的標準答案。

# 化工案例：CSTR 反應器操作模式識別



- Scenario & Data

- 反應器運行多年，操作員憑經驗知道有「不同模式」，但無法量化。

- Goal

- 自動將操作數據分類，建立操作基準 (Baseline)。

- Data

- 600 個歷史樣本，6 個特徵 (Features)

| Sample | Temp (T) | Flow (F) | Yield (%) | ... |
|--------|----------|----------|-----------|-----|
| 001    | 375.2    | 45.1     | 82.1      | ... |
| 002    | 334.8    | 24.5     | 56.3      | ... |
| ...    | ...      | ...      | ...       | ... |

數據包含隱藏的 5 種真實模式，且邊界模糊 (重疊 15-20%)。

# 決策困境：統計最優 vs. 工程現實

The Statistician  
(統計觀點)



**K = 2**

Silhouette Score: 0.552  
(Highest)

“數學上最清晰的分類是  
 $K=2$  (高/低)。”

The Domain Expert  
(工程觀點)



**K = 5**

Theoretical Modes: Very\_High,  
High, Medium, Med\_Low, Low

“物理告訴我們有 5 種模式，  
但數據重疊嚴重。”



當數學建議  $K=2$ ，但物理告訴我們有 5 種模式時，工程師該如何選擇？

# 工程思維：尋找平衡點 (Trade-off)

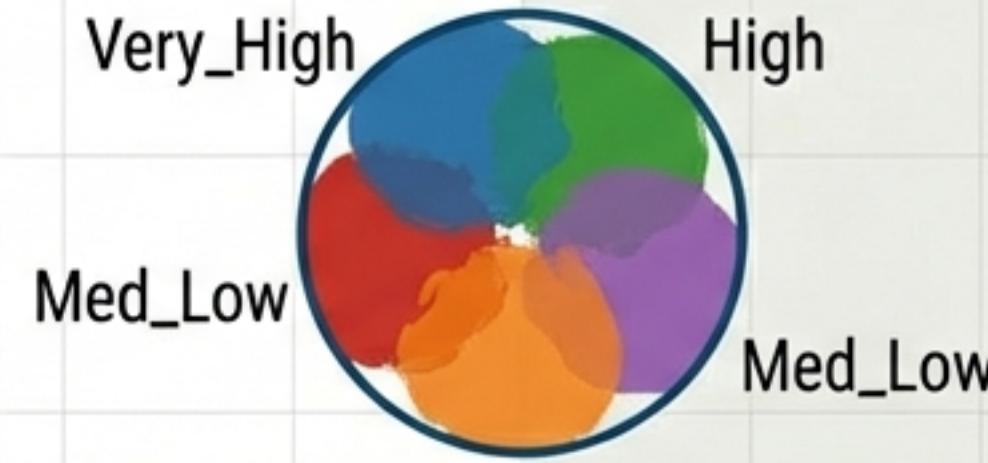
K=2  
(Math Winner)



太粗糙 (Too coarse)

**Verdict:**  
雖然分數最高，但無法提供足夠的操作細節。

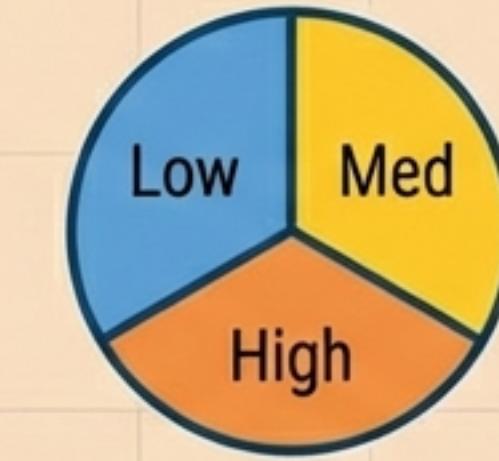
K=5  
(Physics Winner)



邊界太模糊 (Too noisy)

**Verdict:**  
易導致誤判，實務上難以精確操作。

K=3  
(Engineering Balance)



最佳平衡 (Optimal Balance)

1. 手肘點清晰 (Clear Elbow)
2. Silhouette 0.437 (Good)
3. 實務上「高、中、低」三檔最易於管理。

數據科學不僅是計算，更是權衡 (Trade-off)。我們選擇 K=3 來兼顧統計品質與實務可操作性。

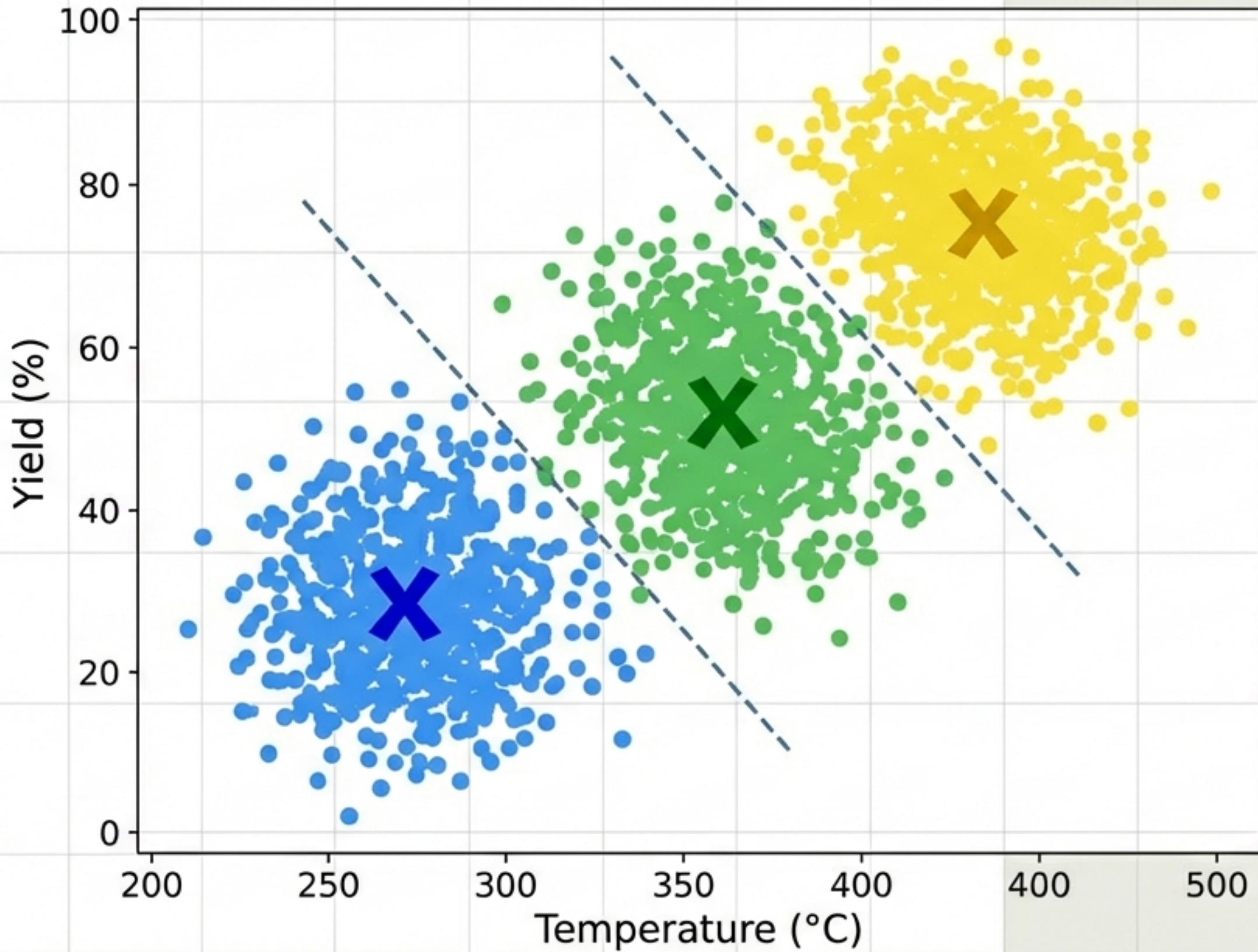
# 結果解讀：群集中心的物理意義

## 三個操作模式的物理特徵分析



Insight: 演算法自動重新發現了化學反應動力學 (Kinetics) 的規則。

# 視覺化驗證：操作區域劃分

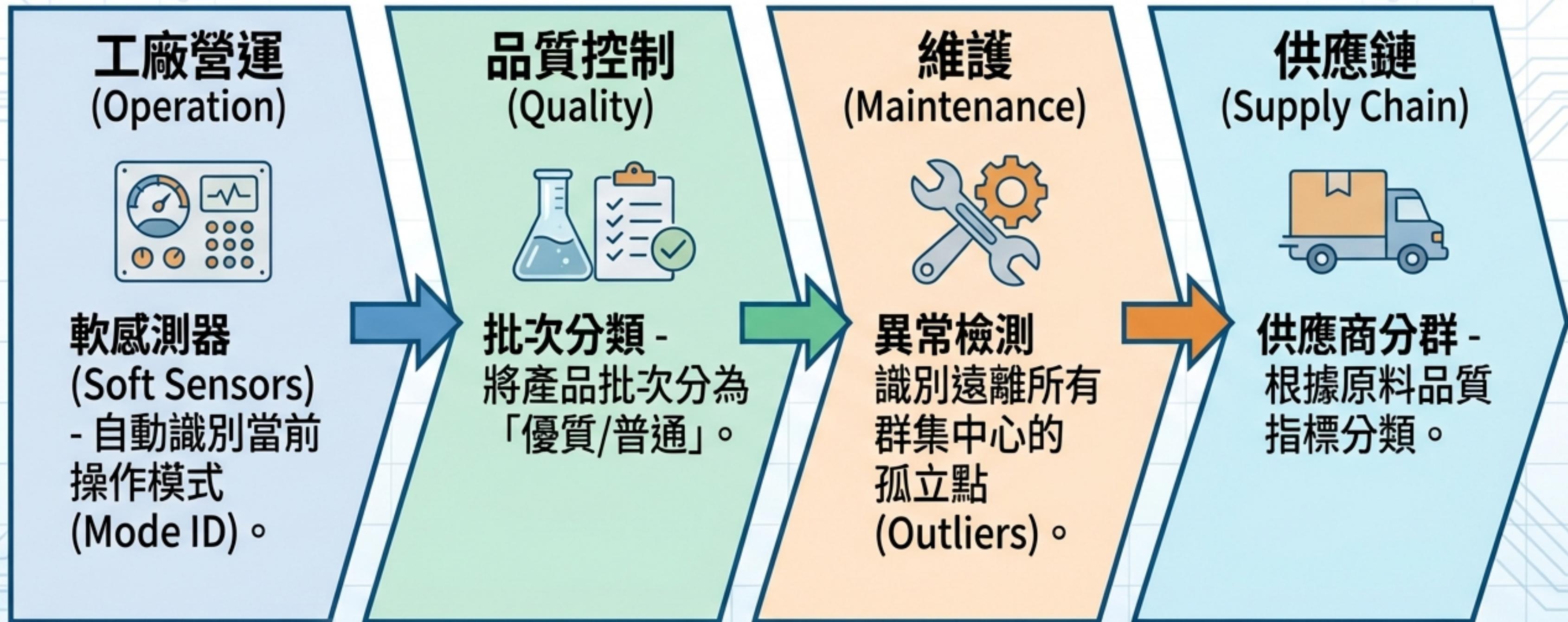


## 區域分析

- Cluster 1 (Yellow)  
合併了真實的 High + Very\_High 模式。
- Cluster 0 (Blue)  
合併了 Low + Medium\_Low 模式。

K-Means 成功識別出三個主要的操作區域，儘管真實模式存在重疊。

# AI 在化工全生命週期的價值 (Clustering Edition)

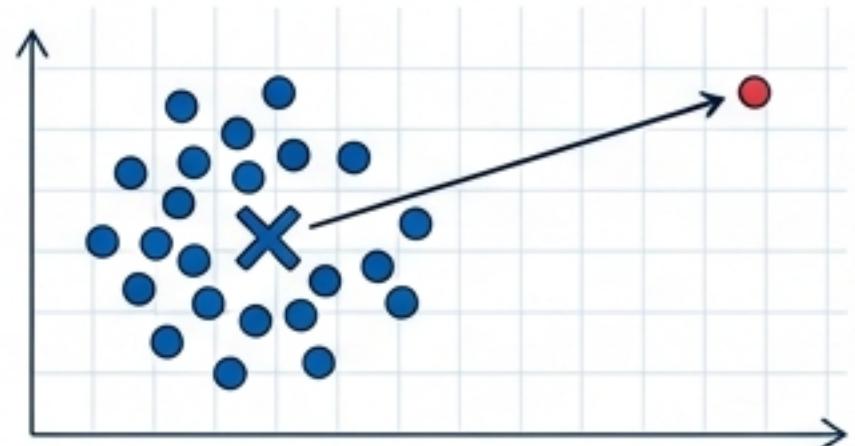


# 現實世界的挑戰 (Challenges in the Real World)

K-Means 在真實數據中的限制與解方。



## 1. 離群值敏感 (Outlier Sensitivity)

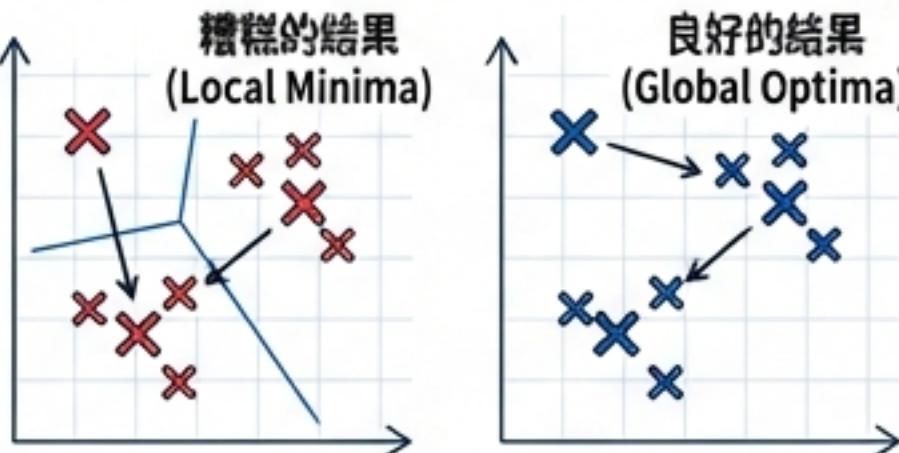


一個極端異常值就會拉偏整個群集中心。

→ **解決方案 (Solution):**  
Fix: 移除 Outliers 或使用 K-Medoids。

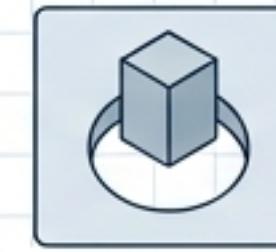


## 2. 初始值風險 (Initialization Risk)

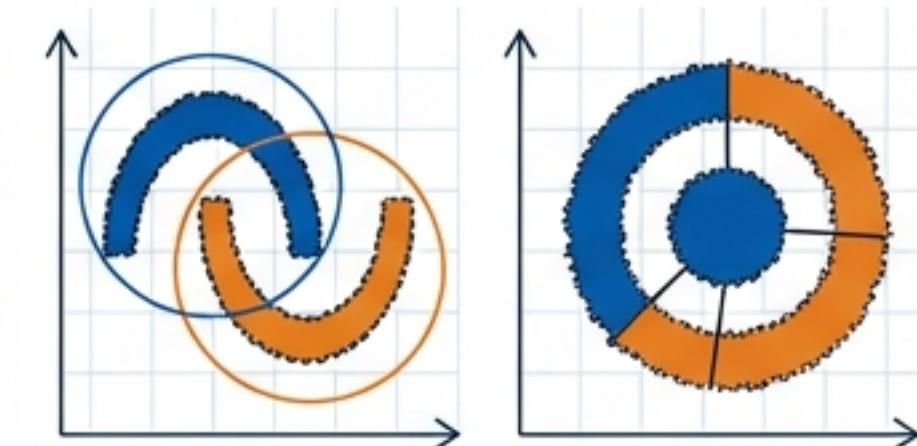


糟糕的起點導致糟糕的結果 (Local Minima)。

→ **解決方案 (Solution):**  
Fix: 始終使用 `n\_init=10` 和 `K-Means++`。



## 3. 球形假設 (Spherical Assumption)



K-Means 假設群集是圓的。如果數據是長條形或環形，效果會很差。

→ **解決方案 (Solution):**  
Fix: 考慮使用 DBSCAN (下一單元)。

# 最佳實踐檢查清單 (Best Practices Checklist)

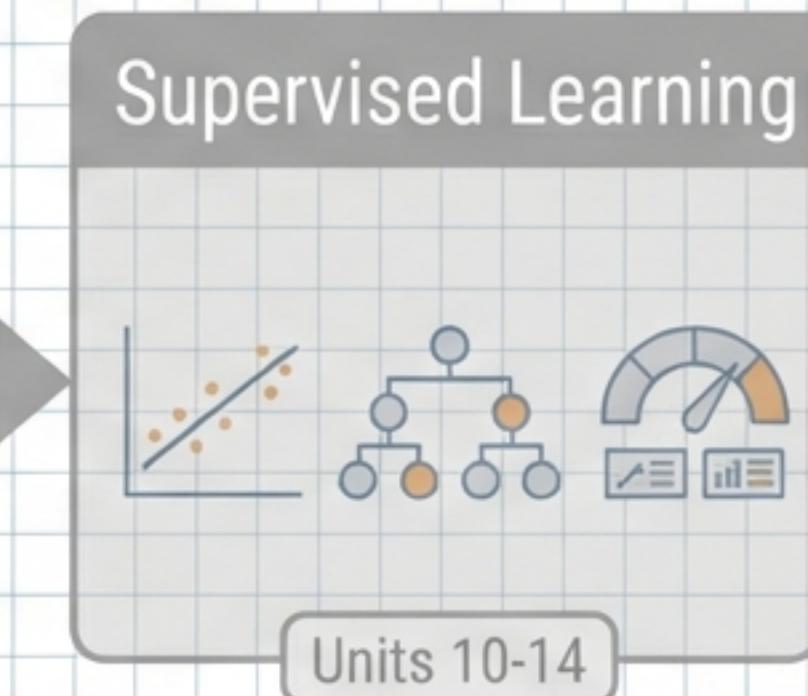
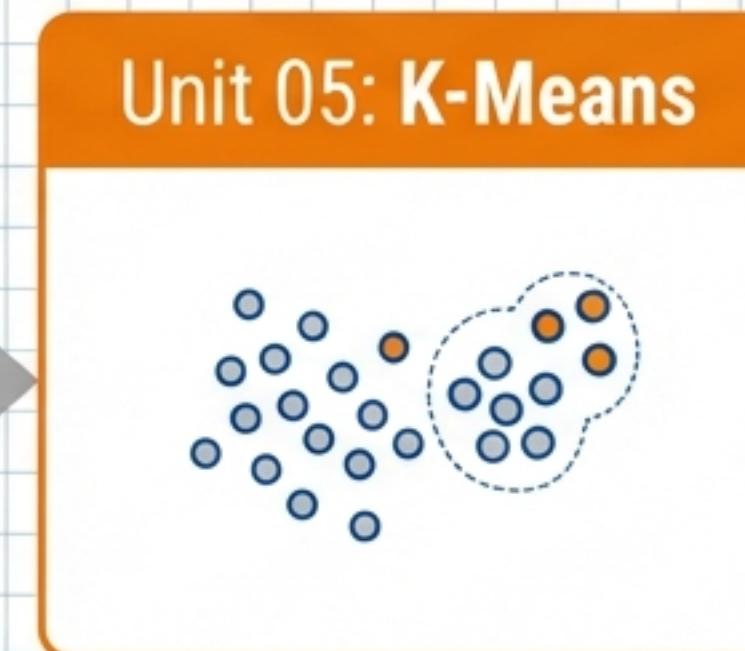
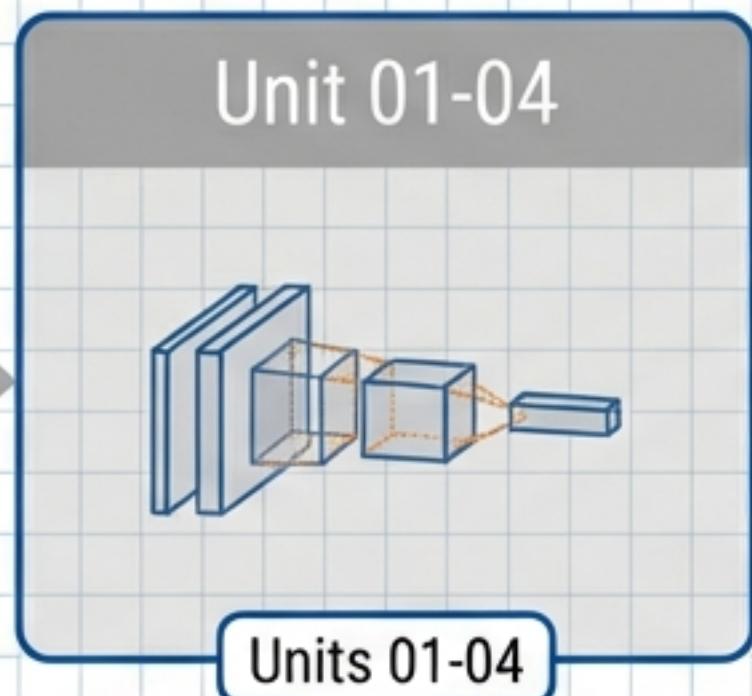


- 資料前處理：檢查缺失值，並務必進行 標準化 (Standardization)。
- 離群值處理：使用 Box Plot 或 Z-score 移除異常點。
- K 值選擇：結合手肘法 (Elbow) 與輪廓分析 (Silhouette)。
- 領域驗證：關鍵步驟—確認群集中心具有物理意義 (Physical Meaning)。
- 穩定性測試：多次運行模型，確保結果一致。

“演算法提供建議，工程師做最終決定。”

# 下一步：超越球形分群

K-Means 是最經典、快速的分群工具，適合處理結構簡單、分界清晰的化工數據。但如果數據形狀不規則？或是充滿噪音？



處理複雜結構與雜訊  
(Complex Shapes & Noise)