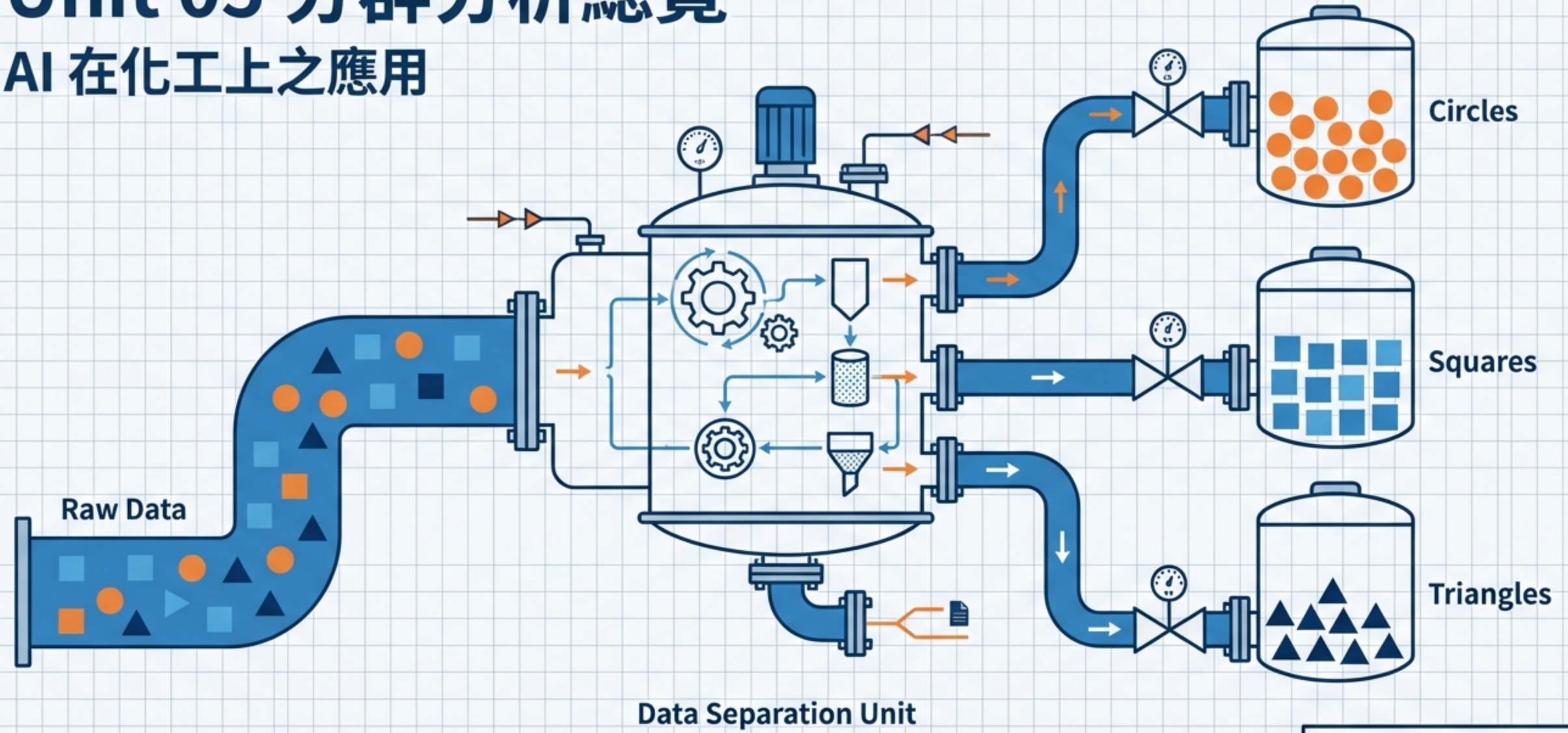
Unit 05 分群分析總覽

AI 在化工上之應用



製作：逢甲大學 化工系
智慧程序系統工程實驗室

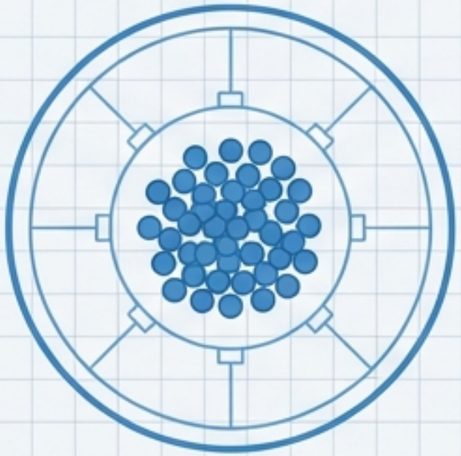
更新日期：2026-

單元操作定義：什麼是分群分析？

非監督式學習 (Unsupervised Learning) — 根據數據特徵自動發現內在結構，無需預先標記。

Concept Definitions

High Cohesion (高內聚性)



同一群集內的數據點相似度高。

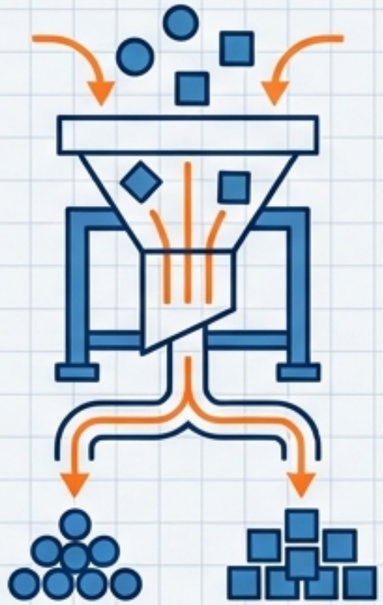
Low Coupling (低耦合性)

不同群集之間的數據點明顯分離。

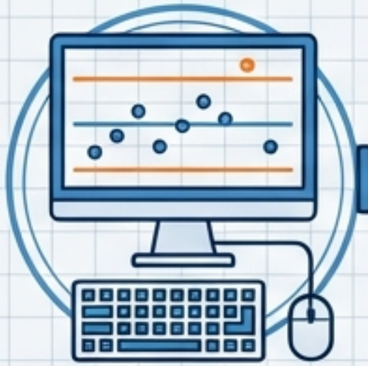
Supervised vs. Unsupervised

Supervised

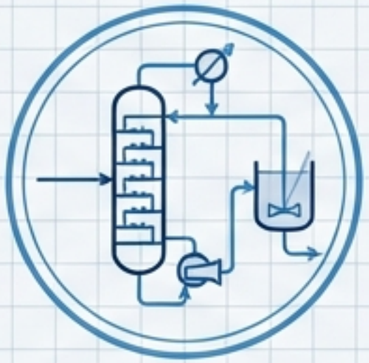
Unsupervised



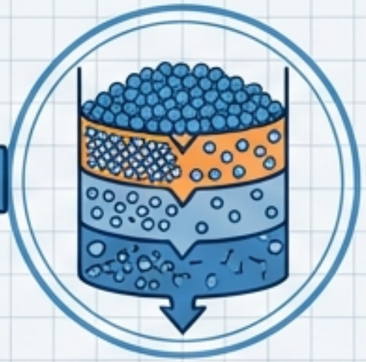
化工場域應用圖譜



製程監控
(Process Monitoring)



製程操作模式識別
(Operation Modes)



產品品質分級
(Product Grading)

溶劑與配方篩選
(Screening)

Clustering Engine
Noto Sans TC Bold

異常檢測
(Anomaly Detection)

數據物理學：相似度與距離

在高維空間中，距離越近代表化學性質越相似。

歐幾里得距離 (Euclidean)

$$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

適用於連續變數

曼哈頓距離 (Manhattan)

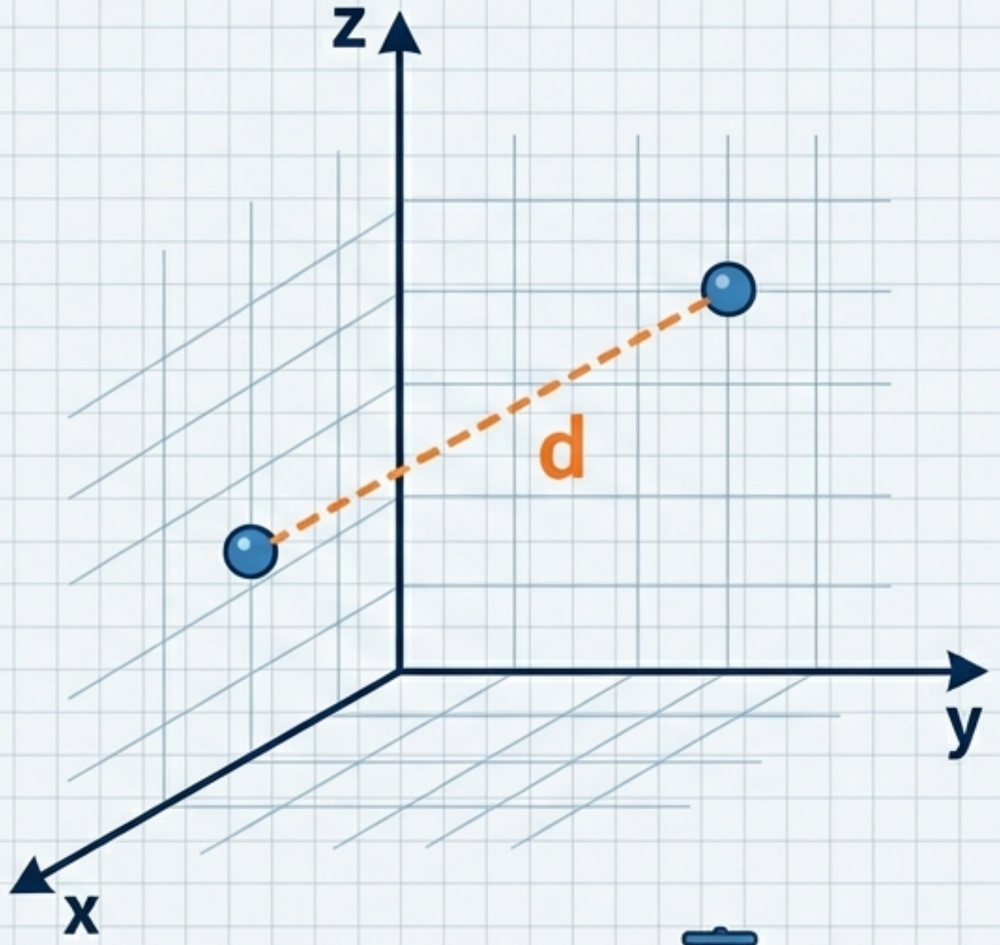
$$d(x, y) = \sum |x_i - y_i|$$

適用於高維度或異常值

餘弦相似度 (Cosine)

$$\text{similarity} = \frac{x \cdot y}{\|x\| \|y\|}$$

適用於方向向量



設備 A：K-平均演算法 (K-Means Clustering)

Initialization
(隨機中心)

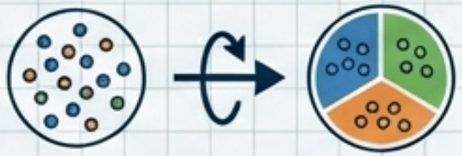
↓

Assignment
(分配點到最近中心)

↓

Update
(更新中心位置)

↺ **Repeat**
(直到收斂)



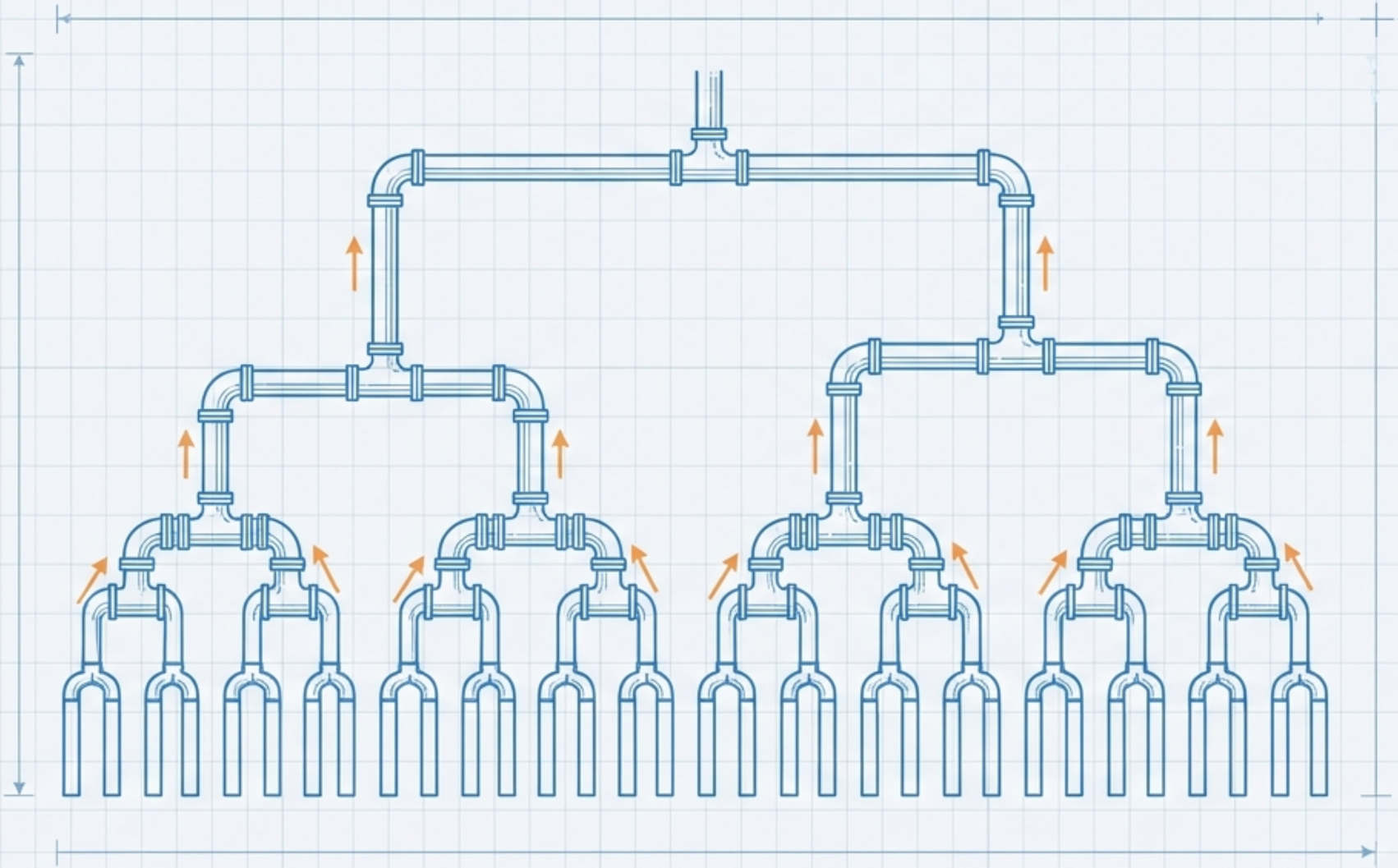
Phase Separation Analogy

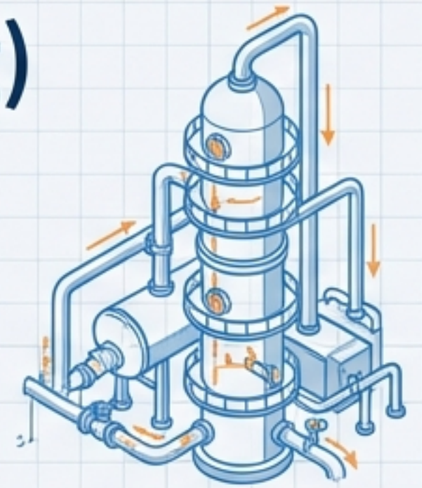
Analogy:
離心機 (Centrifuge)

Specification Card

規格 (Specifications)
Pros (優點)： ✅ 簡單高效、適合大數據
Cons (缺點)： ⚠️ 需預先指定 K 值、假設群集為球形
Best Use Case (最佳應用)： 反應器多模式操作識別

設備 B：階層式分群 (Hierarchical Clustering)



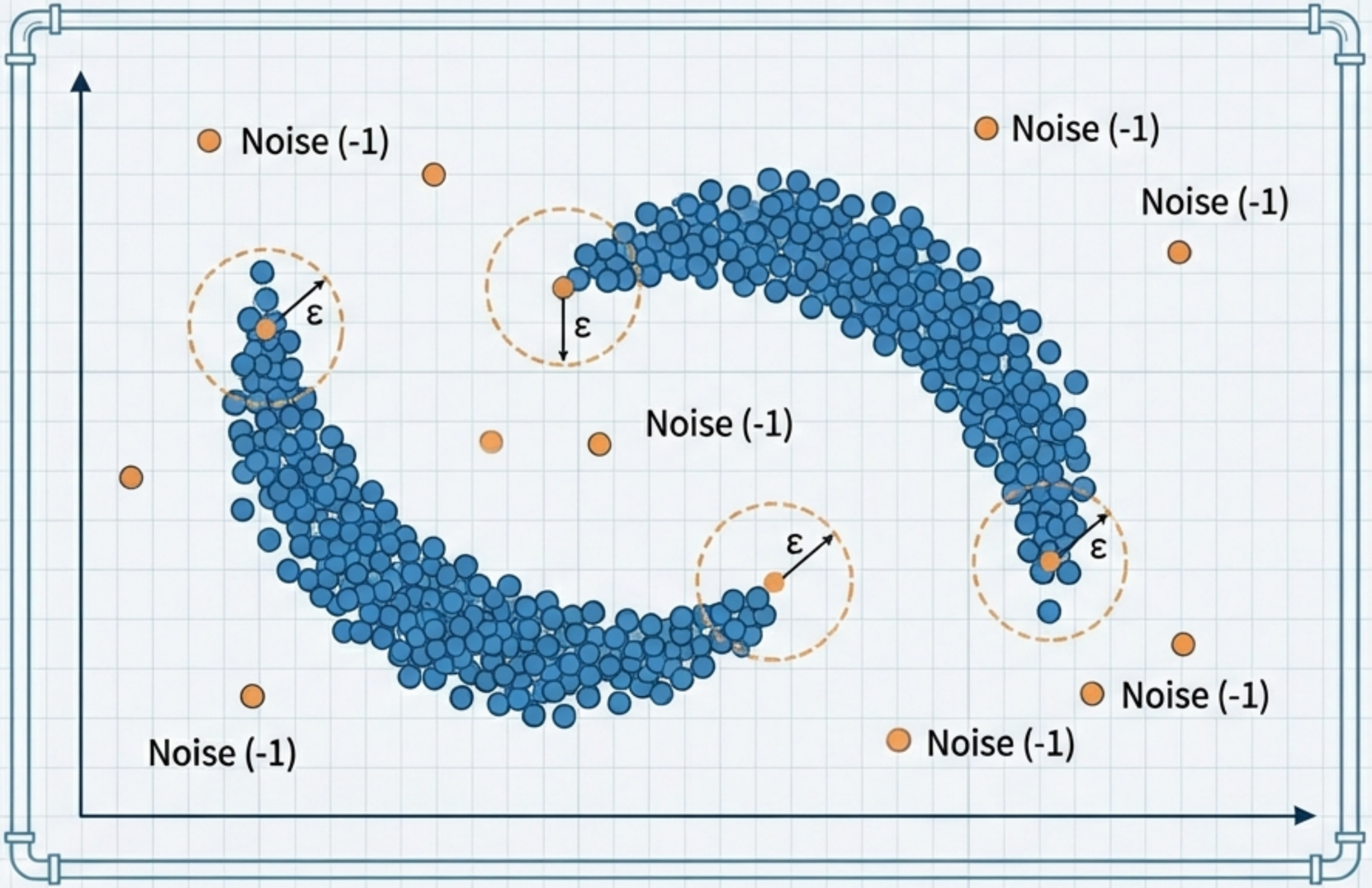


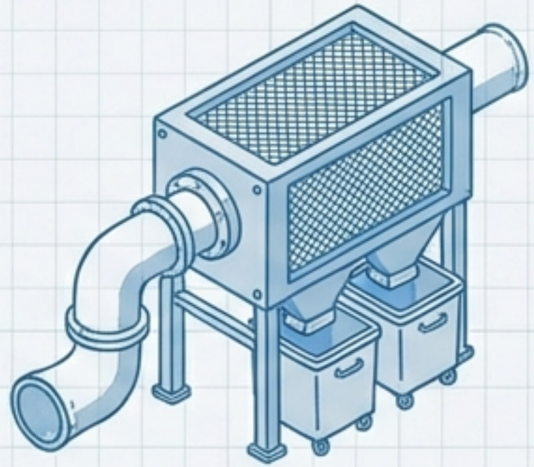
Analogy: 蒸餾塔 (Distillation Column)

規格 (Specifications)

- **Key Concepts:**
  - 策略: 凝聚式 (Agglomerative)
  - 無需指定 K 值
- **Pros (優點):**
  ✅ 揭示階層結構、視覺化強
- **Cons (缺點):**
  ⚠️ 計算量大 $O(n^2)$、不適合大數據
- **Best Use Case (最佳應用):**
  溶劑分類體系建立

設備 C：基於密度的分群 (DBSCAN)



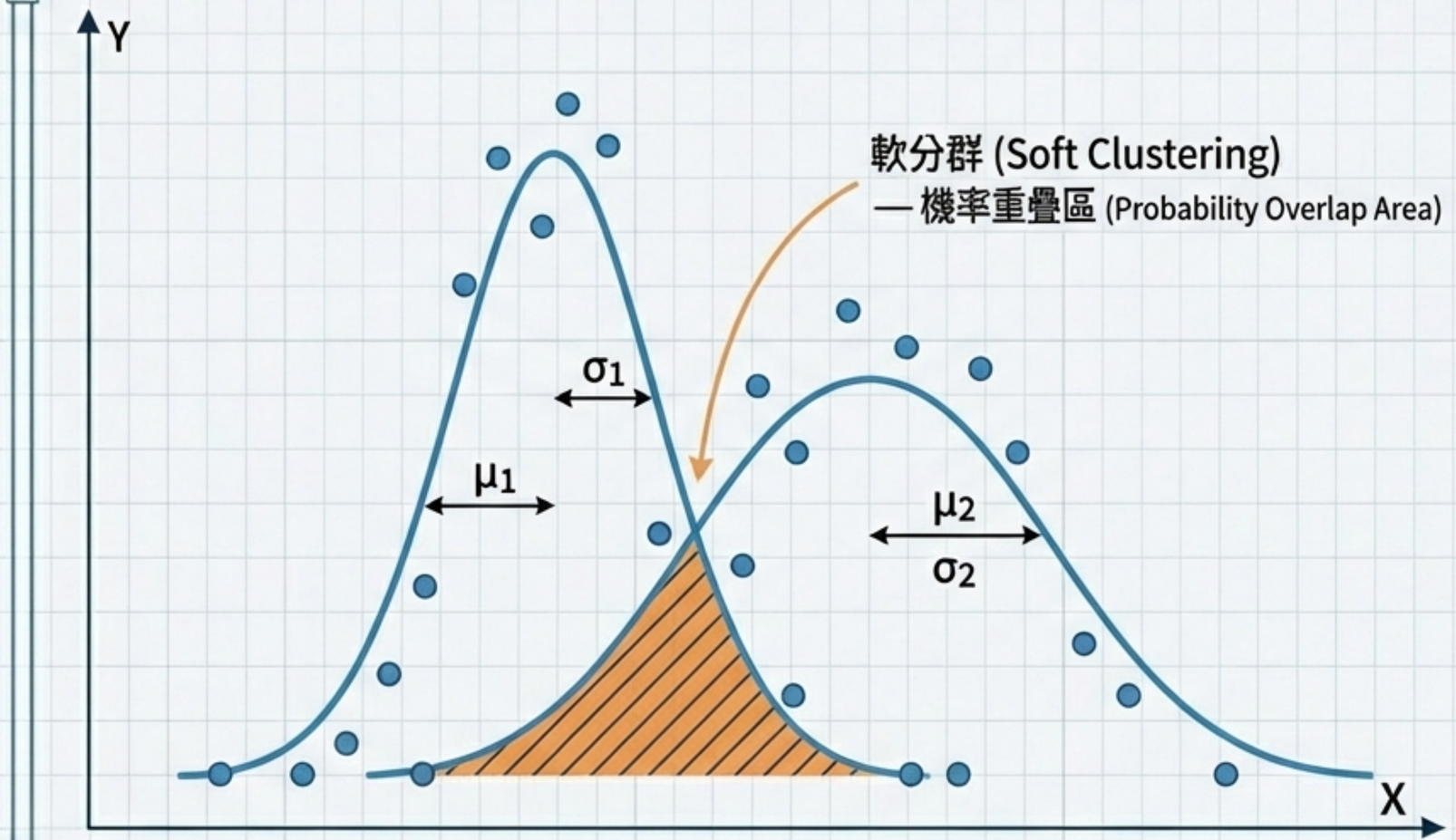


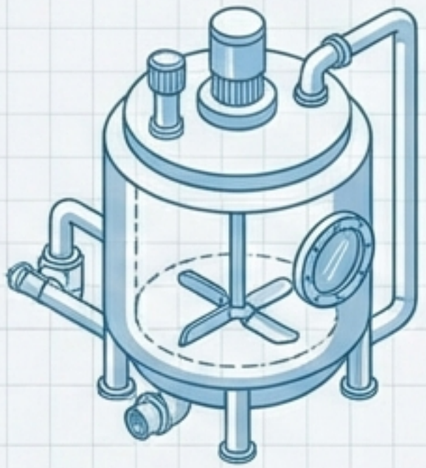
Analogy: 過濾器 (The Filter)

規格 (Specifications)

- **Key Parameters:**
  - `eps (ε)`: 鄰域半徑
  - `min_samples`: 最小樣本數
- **Pros (優點):**
  ✅ 能發現任意形狀、自動識別噪音
- **Cons (缺點):**
  ⚠️ 對參數敏感
- **Best Use Case (最佳應用):**
  製程異常檢測

設備 D：高斯混合模型 (GMM)





Analogy: 混和器 (The Blender)

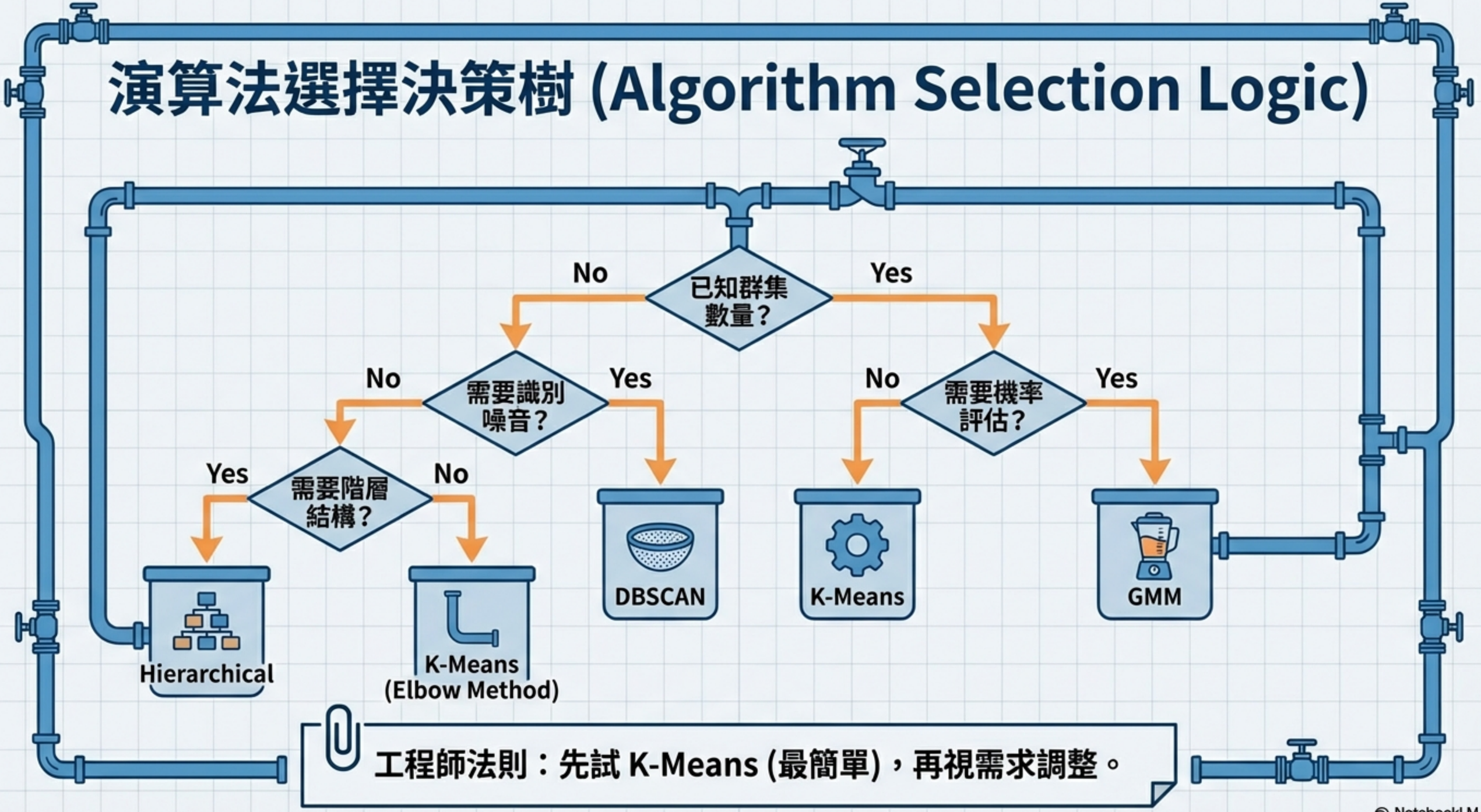
規格 (Specifications)

- **Key Feature:**
  - 數據是由 K 個高斯分布混合而成
  - — 提供屬於某群的「機率」
- **Pros (優點):**
  ✅ 提供機率估計、適應橢圓形群集
- **Cons (缺點):**
  ⚠️ 計算複雜、易陷入局部最佳
- **Best Use Case (最佳應用):**
  產品品質分布建模

設備規格比較表 (Equipment Selection Guide)

	K-Means	Hierarchical	DBSCAN	GMM
形狀限制 (Geometry)	球形 (Spherical)	任意 (Any)	任意 (Any)	橢圓形 (Elliptical)
噪音處理 (Noise)	差 (Sensitive)	差 (Sensitive)	優 (Robust)	中 (Moderate)
運算速度 (Speed)	快 (Fast)	慢 (Slow)	中 (Medium)	慢 (Slow)
參數需求 (Params)	K 值	None	eps, min_samples	K 值

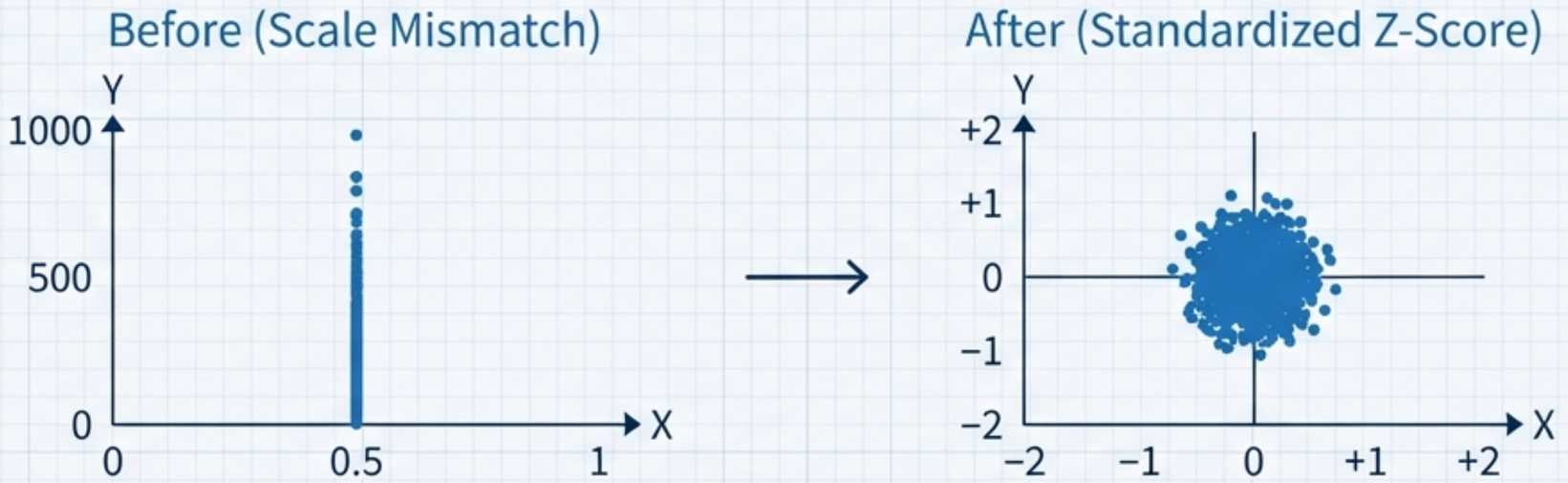
演算法選擇決策樹 (Algorithm Selection Logic)



工程師法則：先試 K-Means (最簡單)，再視需求調整。

進料前處理：數據標準化 (Feed Preparation)

Garbage In, Garbage Out.
未經縮放的數據 = 錯誤的結果

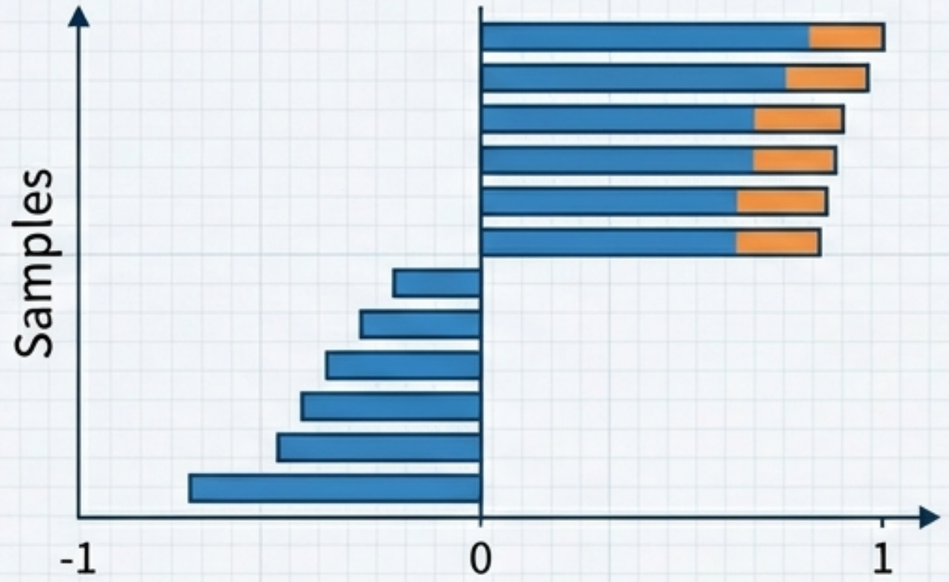


Standardization (Z-Score): $z = (x - \mu) / \sigma$

分群依賴「距離」，不同單位的變數會導致計算偏差。

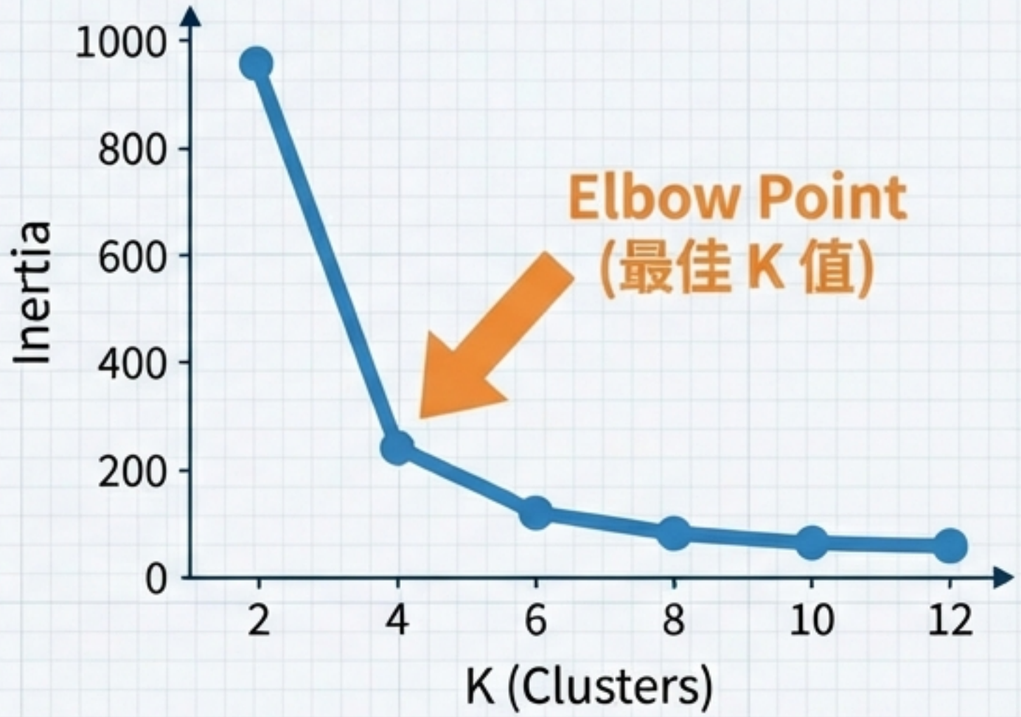
內部品質管制 (Internal Quality Control)

Metric 1 : 輪廓係數 (Silhouette Score)



Range: [-1, 1]
Goal: 接近 1 (高內聚、低耦合)

Metric 2 : 手肘法 (Elbow Method)

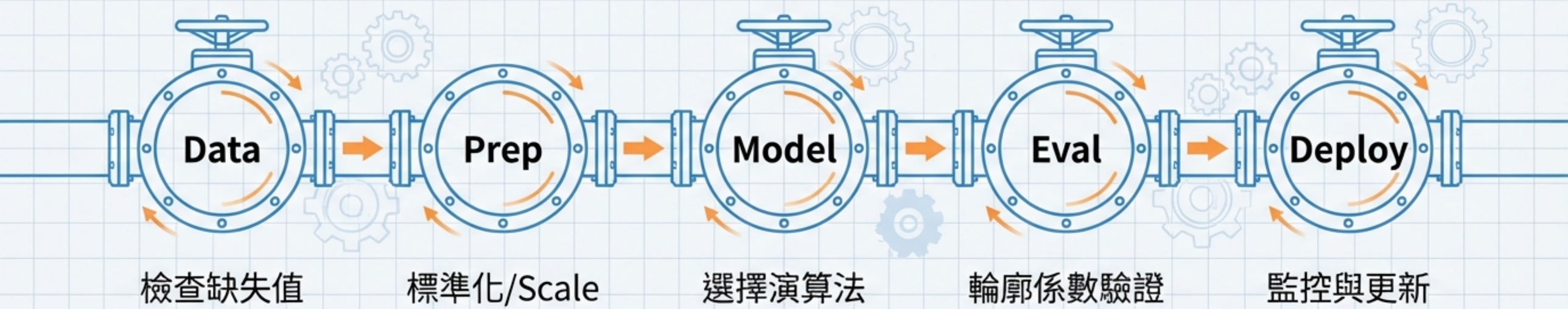



```
sklearn.metrics.silhouette_score(X, labels)
```

工程驗收：物理意義確認 (Engineering QC)

「數學上的最佳解，不一定是工程上的最佳解。」

標準作業程序 (SOP: Standard Operating Procedure)



Best Practice:
始終保留原始數據備份，並記錄所有預處理參數 (Save Scalers!)

值班總結 (Shift Summary)

Key Takeaways

- **工具箱:** K-Means 是**通用板手**，DBSCAN 是**精密過濾器**。
- **前處理:** **標準化** (Standardization) 是**不可妥協**的步驟。
- **驗證:** **統計指標** (Silhouette) + **工程知識** (Domain Knowledge) = **成功分群**。

Next Step: 前往 Unit 06 降維方法 (Dimensionality Reduction)