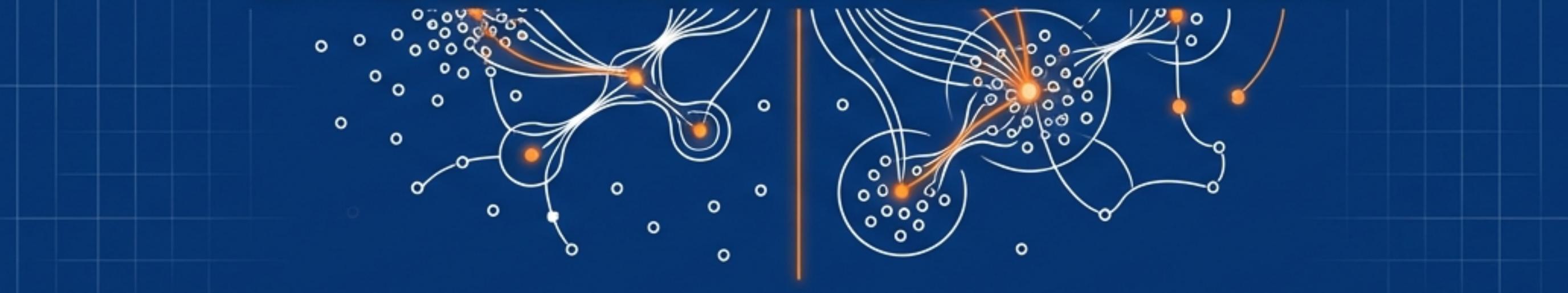


Unit 05: DBSCAN 分群演算法

基於密度的數據挖掘：從原理到化工應用

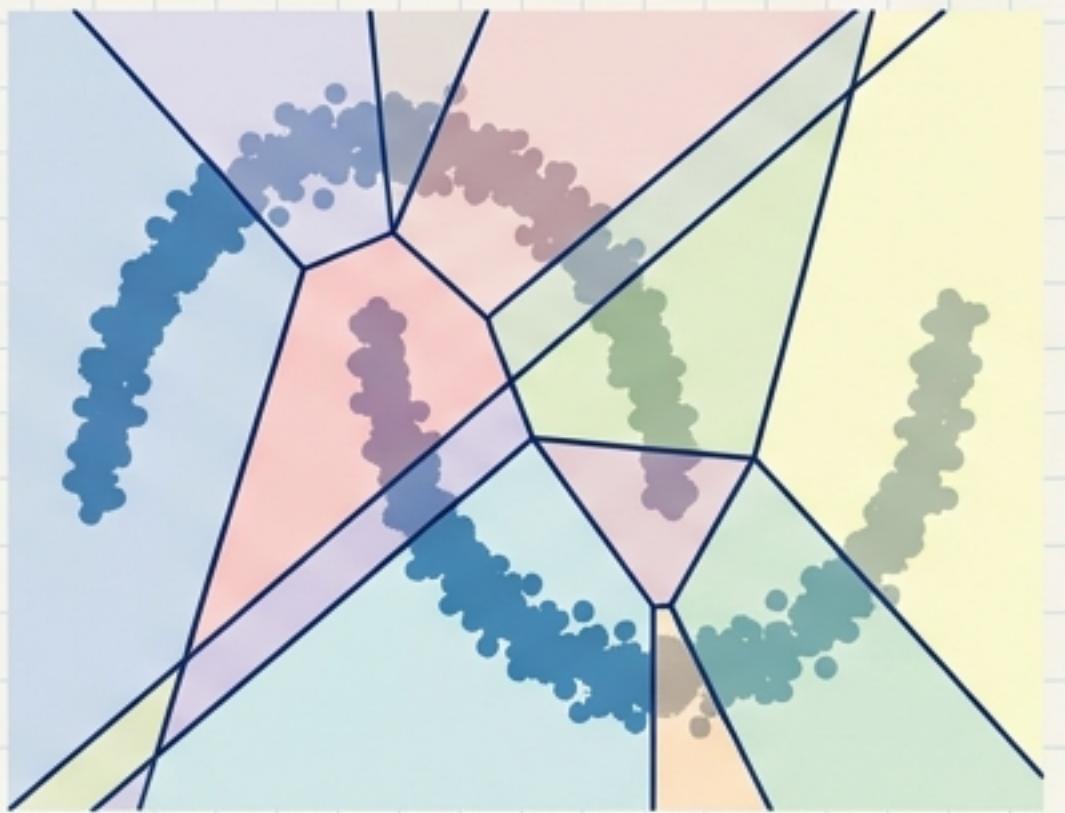


授課教師：莊曜楨 助理教授 | 課程：AI在化工上之應用 | 學期：114學年度第2學期

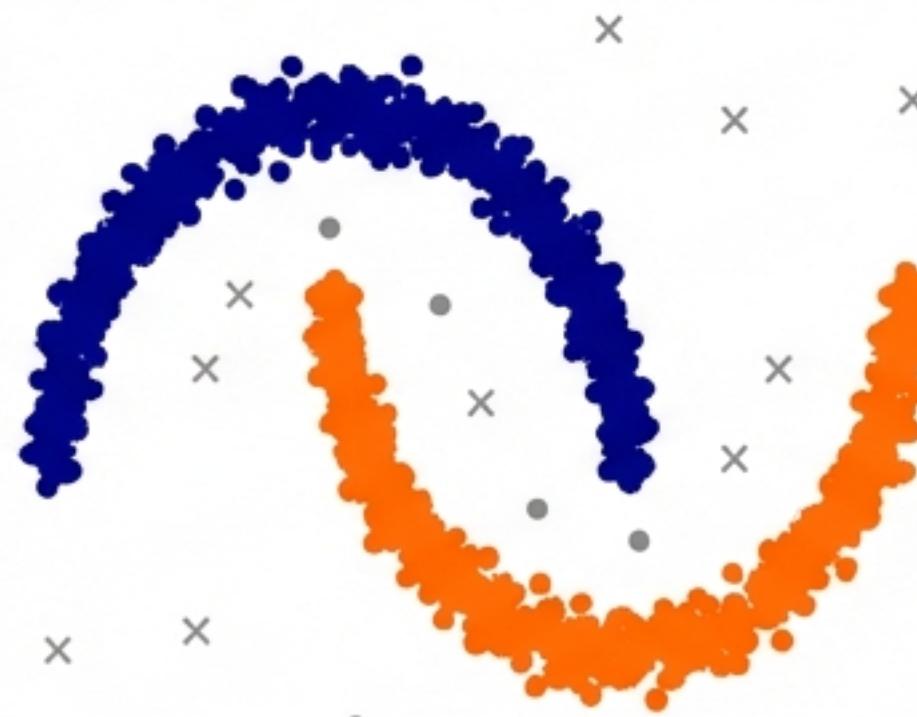
Department：逢甲大學 化工系 智慧程序系統工程實驗室

為什麼需要基於密度的分群？

K-Means：球形假設限制



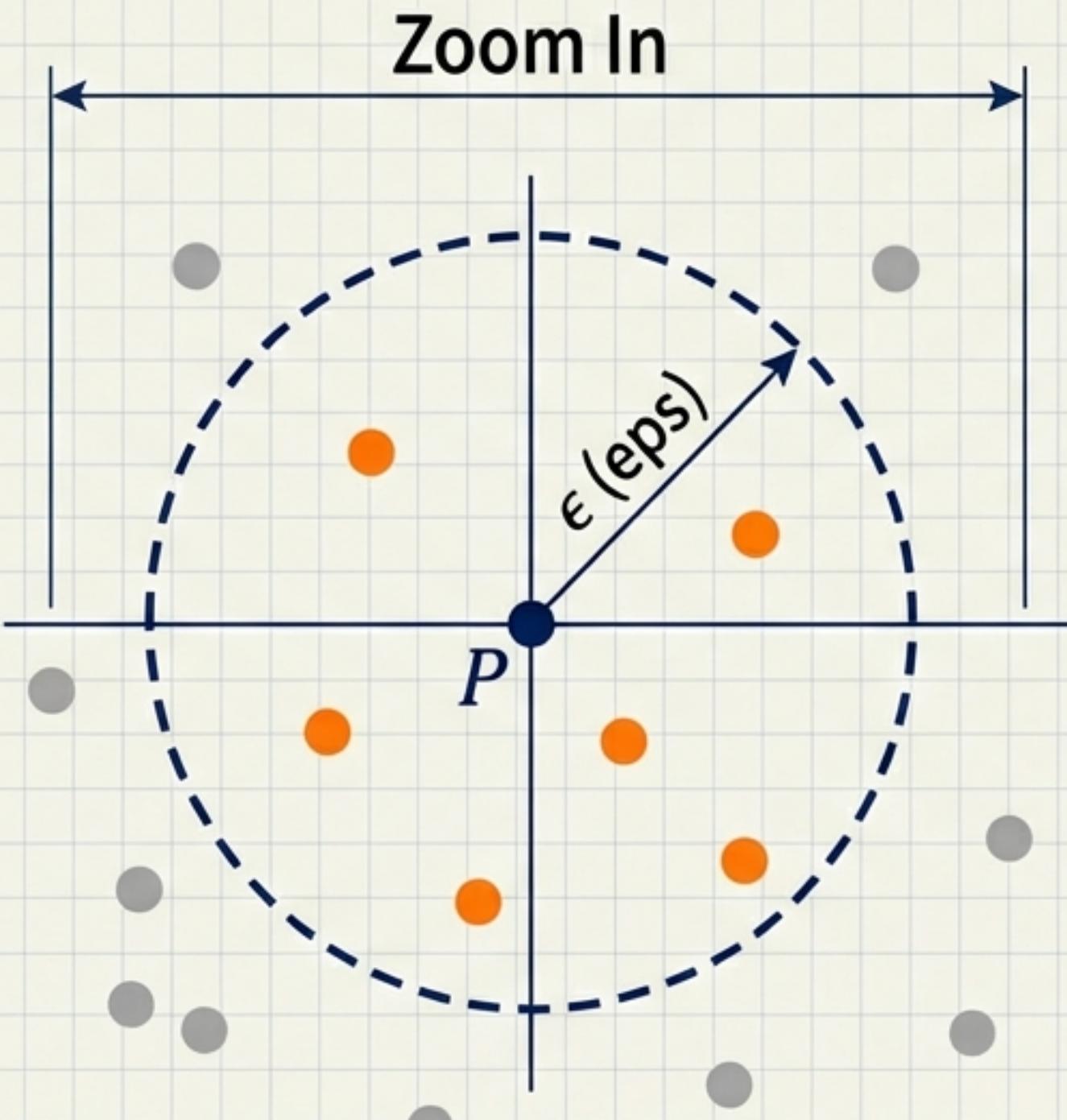
DBSCAN：任意形狀與噪音識別



核心思想：群集是「海洋中的島嶼」—被低密度區域（噪音）分隔的高密度區域。

- ✓ 無需指定 K 值：自動發現群集數量。
- ✓ 適應任意形狀：不受球形分布限制。
- ✓ 噪音容忍度：自動過濾異常值 (Outliers)。

演算法核心參數 (Core Parameters)



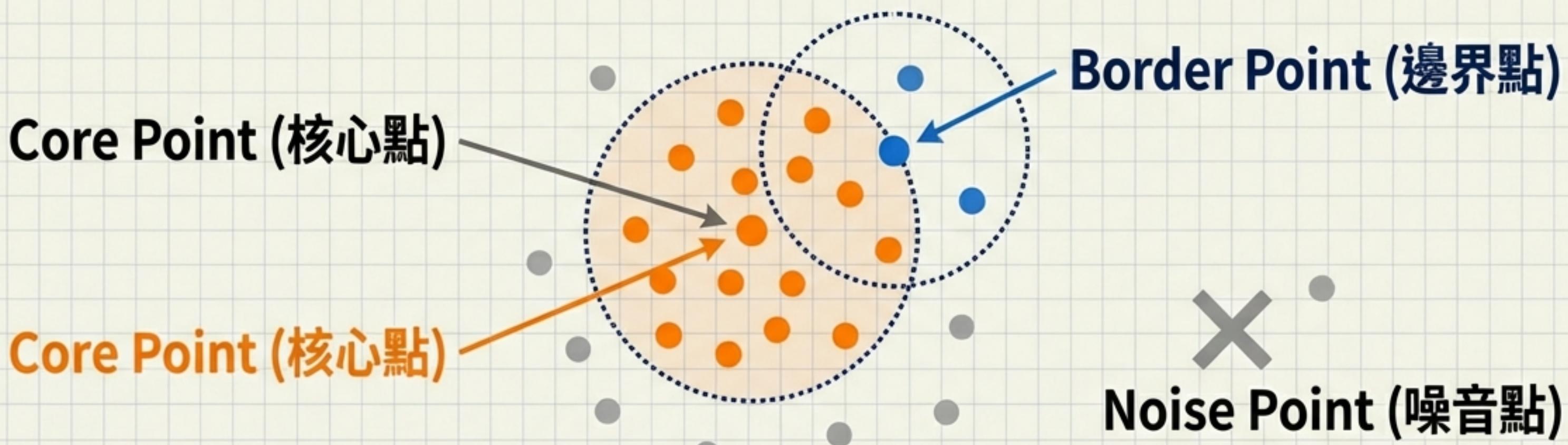
1. ϵ (eps): 鄰域半徑

- * 定義：以點 P 為圓心，半徑為 ϵ 的掃描範圍。
- * Analogy : "你的雷達探測距離有多遠？"
- * 數學定義： $N_\epsilon(p) = \{q \in D \mid \text{dist}(p, q) \leq \epsilon\}$

2. MinPts (min_samples) : 密度閾值

- * 定義：成為「核心點」所需的最小鄰居數量（包含自己）。
- * Analogy : "多少人聚集才算是一個群體？"
- * 化工建議值：通常設為 $2 \times \text{Dimensions}$ (維度)。

點的分類：核心、邊界與噪音



核心點 (Core Point)

- * 條件： ϵ -鄰域內點數 $\geq \text{MinPts}$ 。
- * 意義：群集的內部骨架，密度最高處。

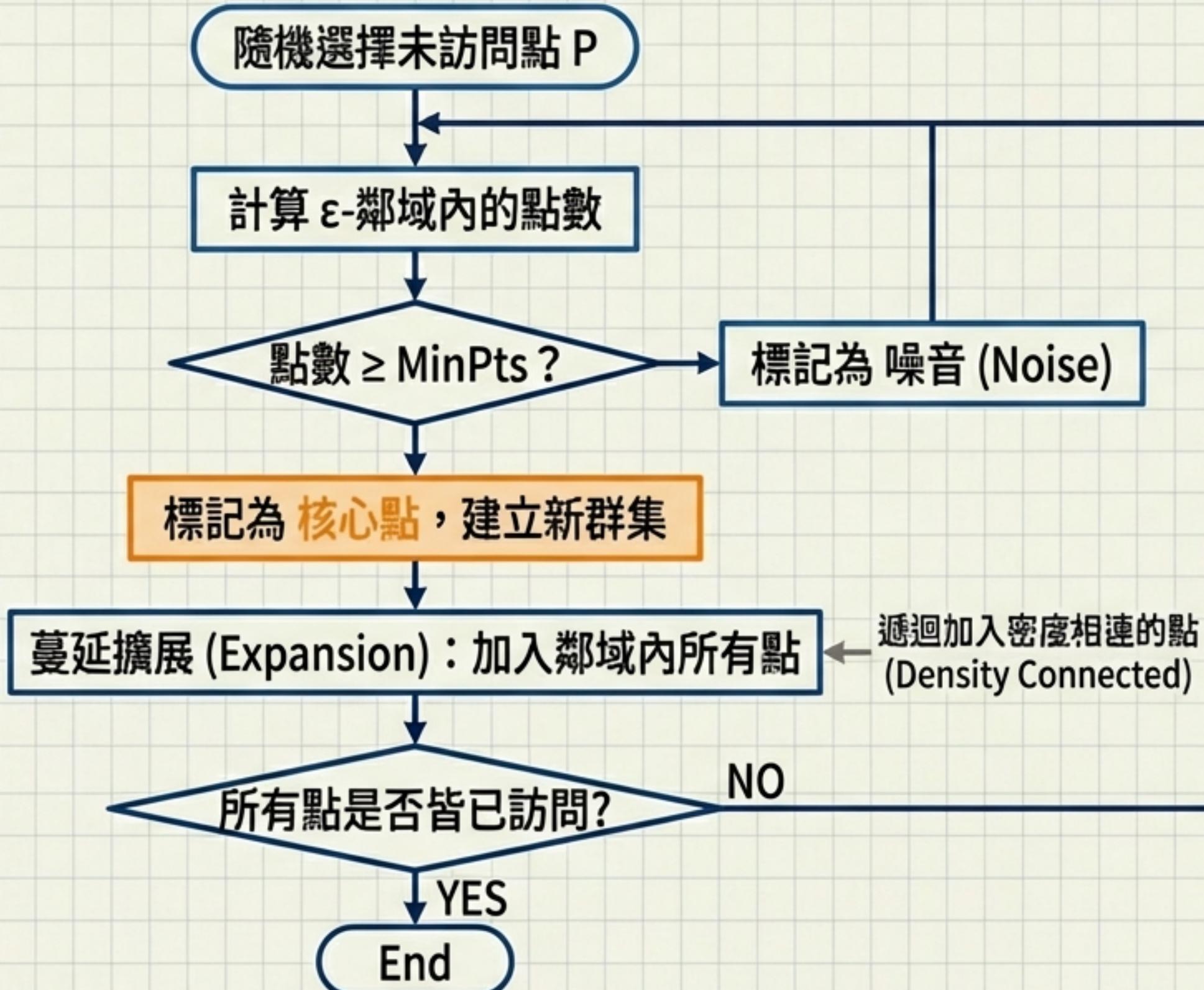
邊界點 (Border Point)

- * 條件：鄰域點數 $< \text{MinPts}$ ，但在核心點鄰域內。
- * 意義：群集的邊緣。

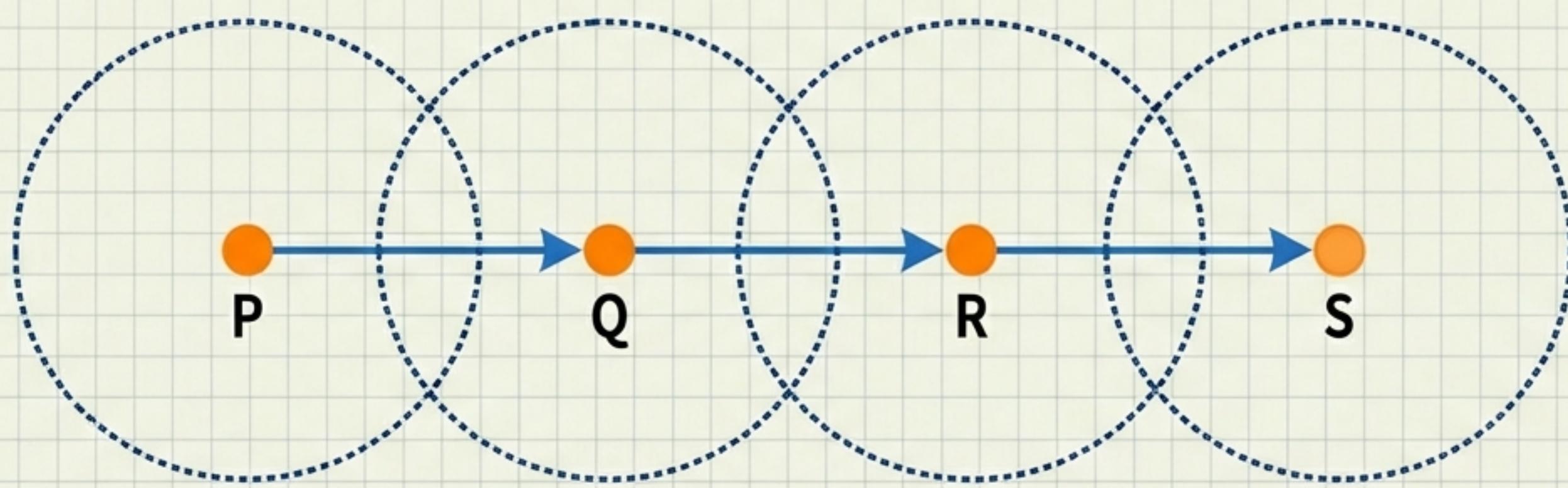
噪音點 (Noise Point)

- * 條件：既非核心點，亦非邊界點。
- * 意義：異常值 (Outliers) 或測量誤差。
- * 化工意義：設備故障或異常操作。

DBSCAN 演算法工作流程



密度連接概念 (Density Connectivity)



1. 直接密度可達
(Directly Density-Reachable)

$P \rightarrow Q$ (一步可達)

2. 密度可達
(Density-Reachable)

$P \rightarrow \dots \rightarrow R$
(通過連鎖反應傳遞)

3. 密度連接
(Density-Connected)

定義了同一個群集 (Cluster)
的最大範圍。

Key Insight : 只要密度不中斷，群集就會一直延伸，這就是DBSCAN能發現任意形狀的原因。

參數選擇：K-距離圖法 (K-Distance Graph)



步驟 (Methodology)：

1. 設定 $k = \text{MinPts}$ (通常為維度 $\times 2$)。
2. 計算每個點到其第 k 個最近鄰居的距離。
3. 將距離由小到大排序並繪圖。
4. 尋找「肘部」(Elbow Point)：曲線斜率急劇變化的位置 = 最佳 ϵ 。

Engineering Insight :

- * ϵ 過小 → 數據過度分割，噪音極多。
- * ϵ 過大 → 不同群集被合併，失去分辨力。

化工領域應用場景 (ChemE Applications)



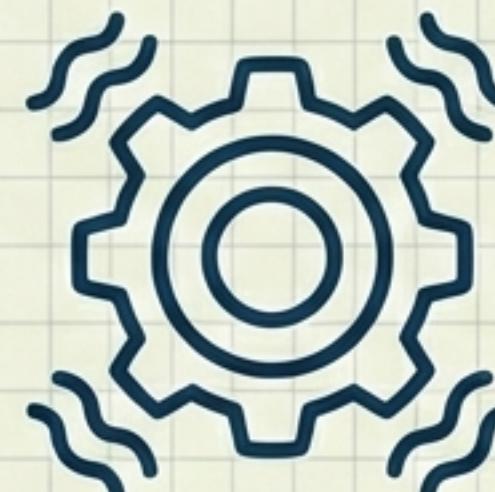
**異常操作模式
(Anomaly Detection)**
場景：反應器故障前的異常軌跡。
優勢：自動標記不屬於任何穩態的點為「噪音」。



**複雜製程狀態
(Operating Modes)**
場景：聚合反應器中不同產品等級的操作區域。
優勢：準確識別非球形的操作窗口。



**批次軌跡分析
(Batch Process)**
場景：發酵製程的生長曲線分類。
優勢：區分正常批次與失敗批次。



**故障診斷
(Fault Diagnosis)**
場景：壓縮機振動數據分析。
優勢：識別罕見的故障類型。

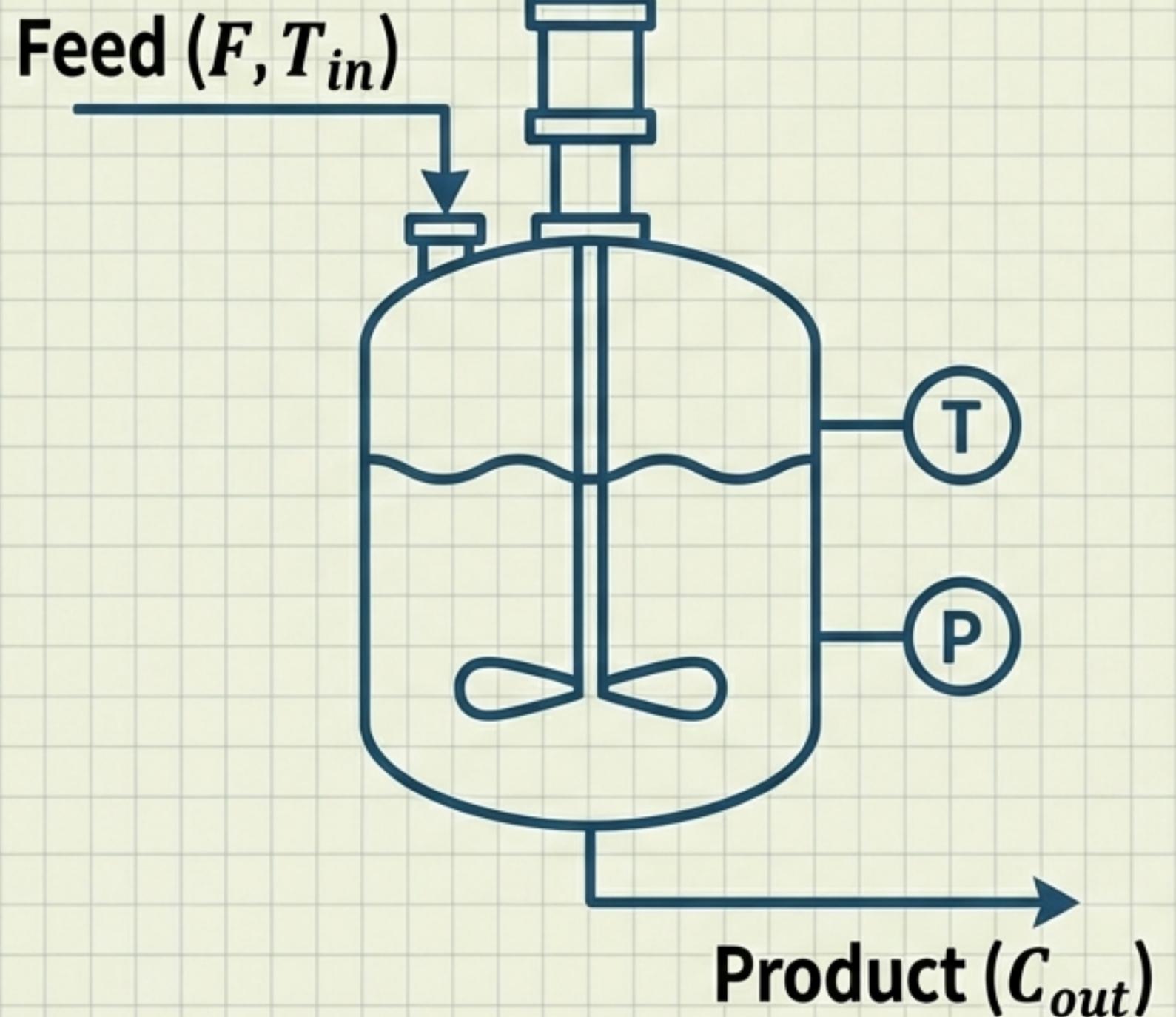
實作指南：scikit-learn 與 數據預處理

```
●●●  
from sklearn.cluster import DBSCAN  
from sklearn.preprocessing import StandardScaler  
  
# 1. 數據標準化 (CRITICAL STEP!)  
# 將數據縮放至均值=0，標準差=1  
X_scaled = StandardScaler().fit_transform(X_raw)  
  
# 2. 模型訓練  
# eps: 鄰域半徑，min_samples: 最小鄰居數  
db = DBSCAN(eps=0.5, min_samples=5)  
db.fit(X_scaled)  
  
# 3. 取得標籤  
# 標籤為 -1 的點即為噪音 (Noise)  
labels = db.labels_
```

⚠ 必須進行標準化 (Standardization is Must)

- 原因：DBSCAN 基於歐幾里得距離 (L_2 Distance)。
- 問題：若變數單位不同 (例如：溫度 $T = 300K$ vs 濃度 $C = 0.1M$)，數值大的變數(溫度)會完全主導距離計算。
- 解決：使用 StandardScaler 將所有特徵縮放至同一尺度 $((x - \mu)/\sigma)$ 。

案例研究：CSTR 反應器異常檢測



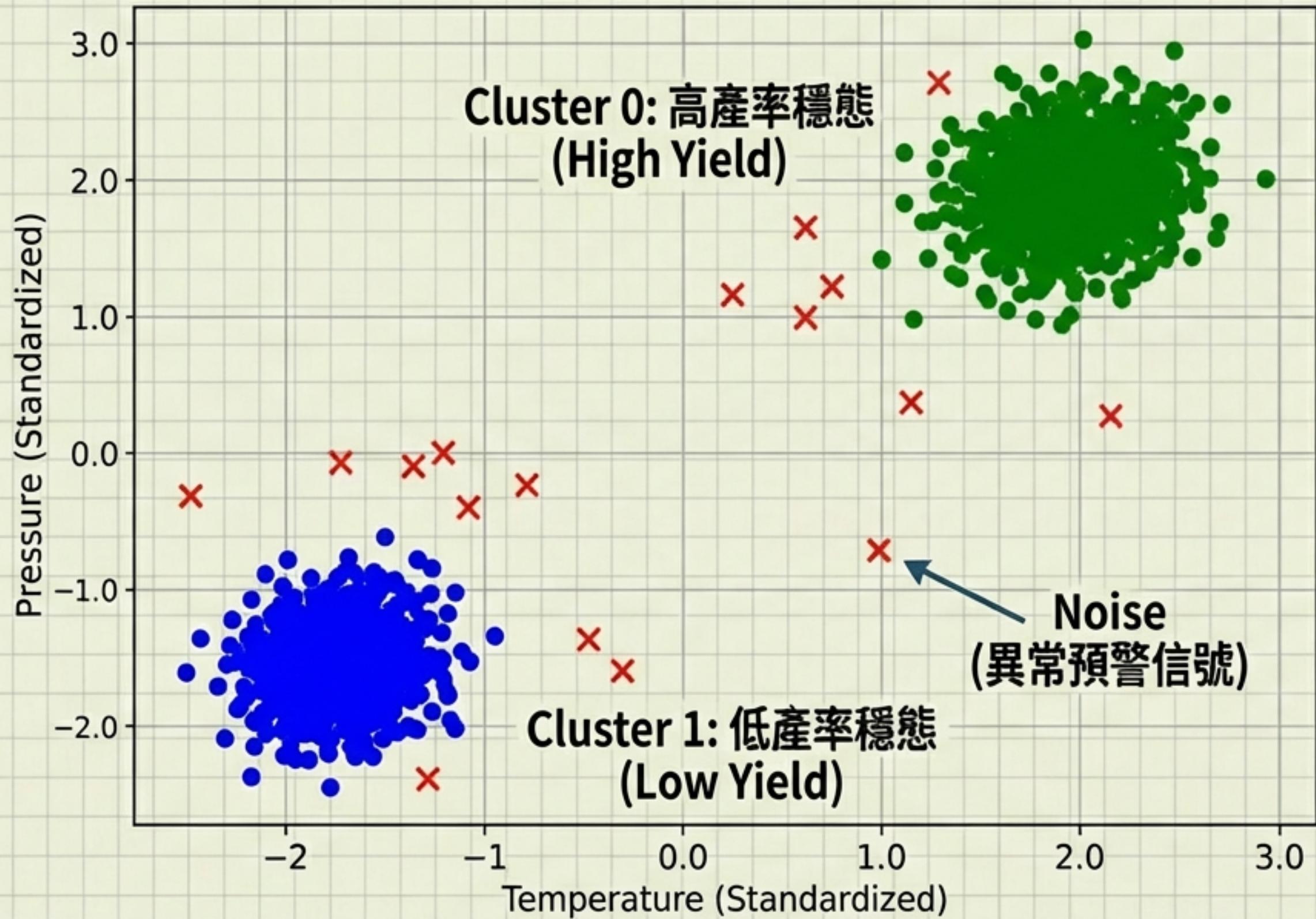
問題定義 (Problem Definition):

- 設備：連續攪拌槽式反應器 (CSTR)。
- 監測變數：反應溫度 (T)，壓力 (P)，進料流速 (F)，攪拌速度 (RPM)。
- 目標：自動識別三種狀態，無需人工標籤：
 1. 操作模式 A (高產率/High Yield)
 2. 操作模式 B (低產率/Low Yield)
 3. 異常操作 (Anomalies) – 安全隱患

數據特徵 (Data Characteristics):

6 維數據，包含測量誤差 (噪音)。

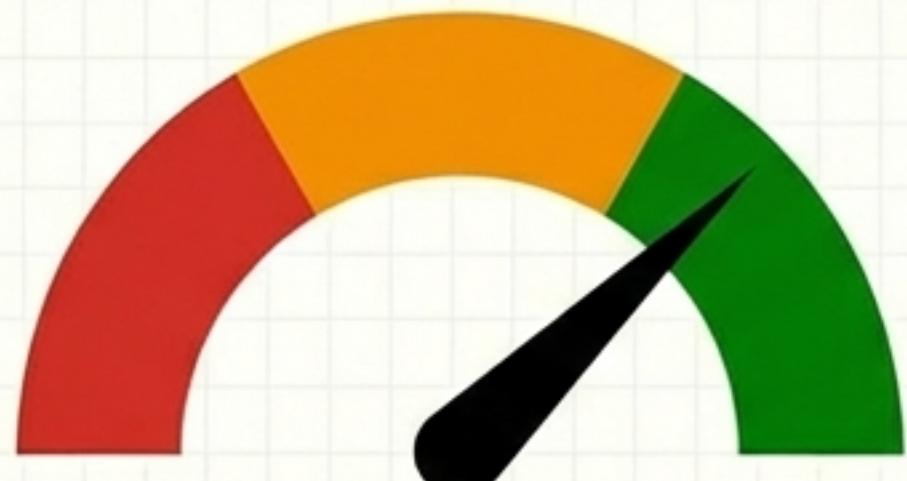
實驗結果：操作模式與異常識別



- 參數設定： $\text{eps} = 0.9$ 、 $\text{min_samples} = 10$ (經 K-Dist 優化)。
- 發現：DBSCAN 成功將異常點分離為 'Label -1'。
- 應用：當新數據點落入「紅色區域」時，立即觸發警報。

模型評估指標 (Evaluation Metrics)

輪廓係數
(Silhouette Score)

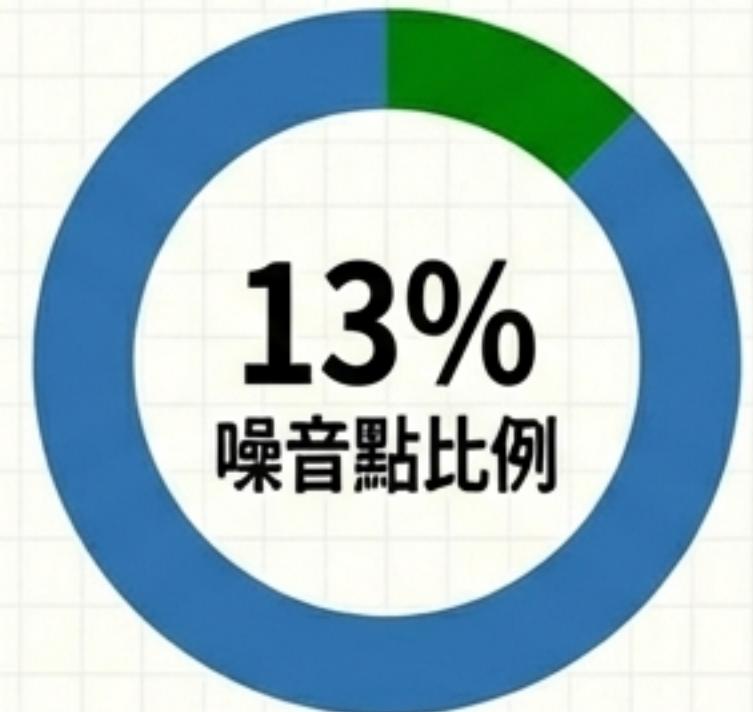


0.70

輪廓係數

> 0.5 表示群集分離良好。

噪音點比例
(Noise Ratio)



Goldilocks Zone (合理範圍)：

- < 1%: 參數太鬆，漏報異常。
- > 30%: 參數太嚴，誤報過多。

Davies-Bouldin
Index

0.43

Davies-Bouldin Index

數值越低越好。

在化工應用中，噪音往往代表最重要的信息（故障或異常）。

算法對決：DBSCAN vs. K-Means

特性 (Feature)	DBSCAN	K-Means
群集形狀	任意形狀 (適合複雜製程)	僅限球形 (凸集)
噪音處理	自動識別 (-1)	強制分配 (受異常值影響大)
參數需求	ε , MinPts (難調參)	K (需預知群集數)
最佳場景	異常檢測、狀態監控	數據壓縮、一般分類

現實世界的挑戰與對策 (Challenges & Solutions)



密度不均 (Varying Densities)

- ⚠ 挑戰 (Challenge) :
- 若群集 A 密度極高，群集 B 密度低，單一 ϵ 難以同時捕捉。

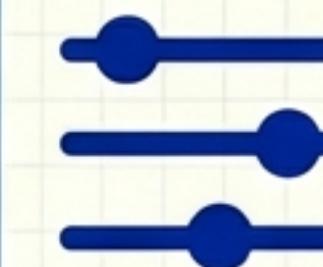
- ✓ 對策 (Solution) :
- 使用 **HDBSCAN** (分層式) 或多次運行不同參數。



維度詛咒 (Curse of Dimensionality)

- ⚠ 挑戰 (Challenge) :
- 高維空間 ($D > 10$) 中，距離度量變得無意義。

- ✓ 對策 (Solution) :
- 先使用 **PCA** 降維，或進行特徵工程 (Feature Engineering)。



參數敏感性 (Parameter Sensitivity)

- ⚠ 挑戰 (Challenge) :
- ϵ 微小的變化可能導致結果劇變。

- ✓ 對策 (Solution) :
- 善用 **K-距離圖** 輔助選擇，不要憑感覺猜測。

在化工應用中，應對這些挑戰是實現可靠模型的關鍵。

總結與下一步 (Summary & Next Steps)



Data has shape; let the density define it.
(數據具有形狀，讓密度來定義它。)