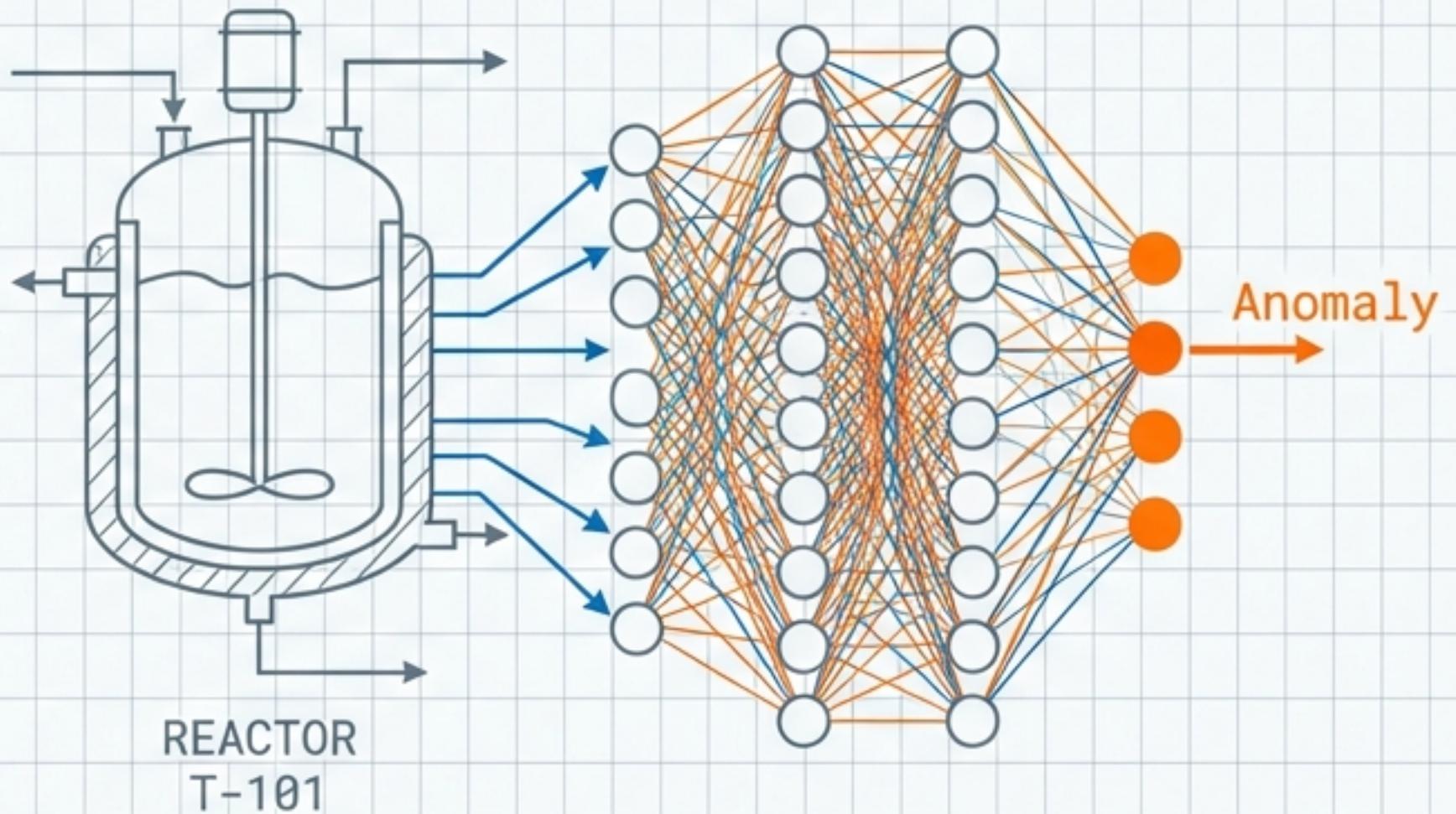


AI 在化工上的應用： 從原理到實踐

Unit 07 異常檢測總覽 (Anomaly Detection Overview)



授課教師：莊曜禎 助理教授

學期：114學年度第2學期

單元目標：建立異常檢測觀念，掌握 sklearn 主流演算法與評估策略

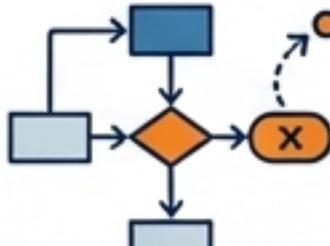
異常檢測概論：定義與術語 (Introduction to Anomaly Detection)

 **Core Definition:** 異常檢測 (Anomaly Detection): 識別與大多數數據顯著不同的樣本的技術。

 **Why it matters:** 早期發現製程故障、產品瑕疵或操作錯誤。

異常 (Anomaly)

-  **Definition:** 廣義的異常數據點 (General term).
-  **ChemE Example:** 任何偏離正常操作的數據。



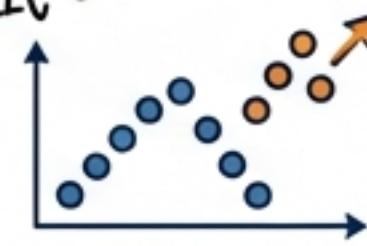
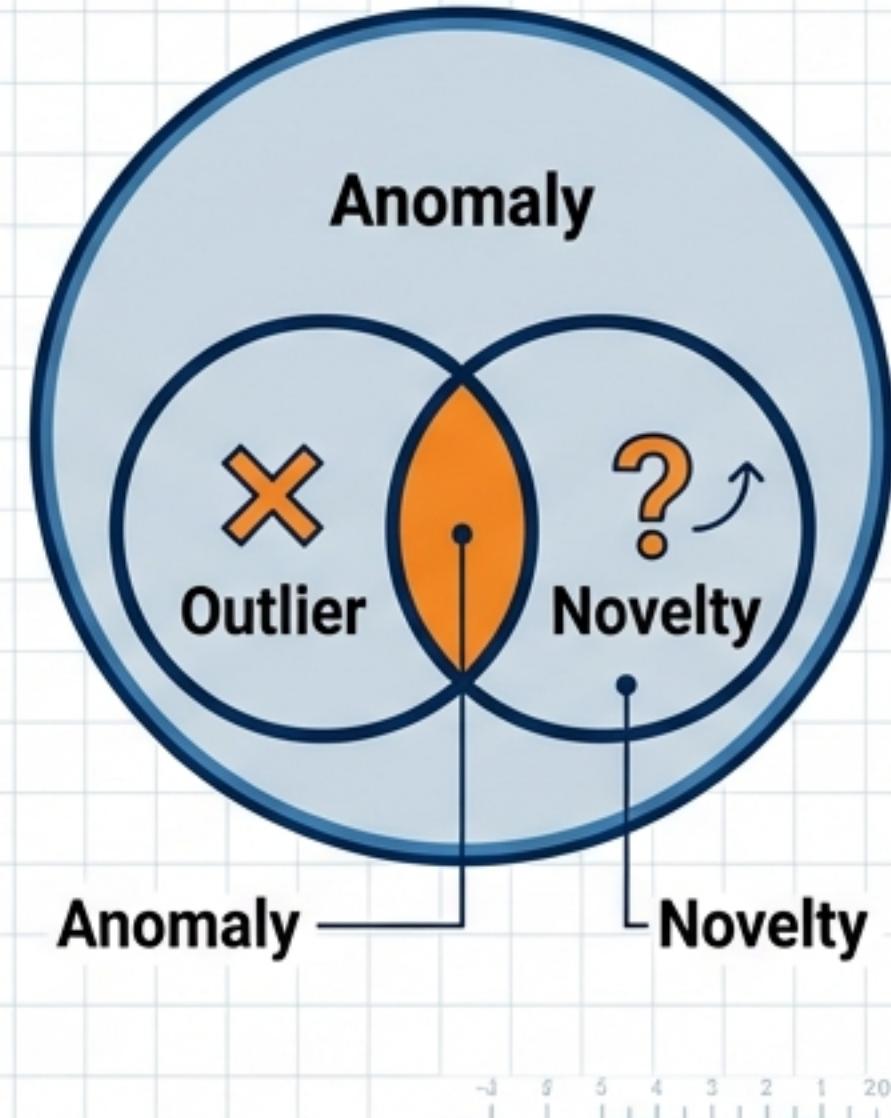
離群值 (Outlier)

-  **Context:** 訓練數據中的雜訊 (Noise).
-  **Goal:** 識別並移除 (Identify and remove).
-  **ChemE Example:** 感測器雜訊、操作失誤。



新奇點 (Novelty)

-  **Context:** 測試數據中的新模式 (New pattern).
-  **Goal:** 識別新事件 (Detect new events).
-  **ChemE Example:** 新產品試產、未見過的故障模式。

異常檢測在化工全生命週期的應用 (Applications Across the Chemical Lifecycle)

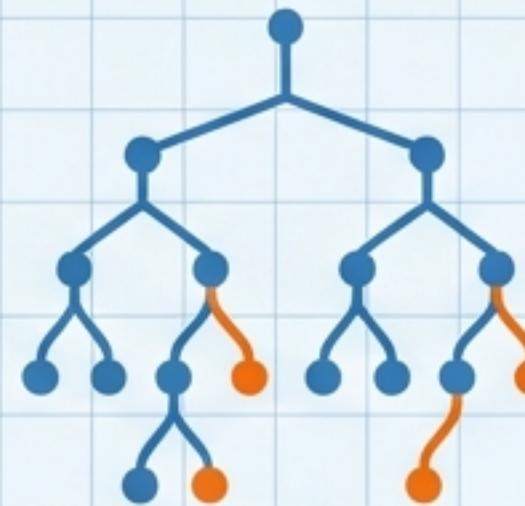


案例 (Case Study)

某石化廠使用 Isolation Forest 提前 15 分鐘預警反應器溫度異常，避免事故。

演算法地圖：sklearn 中的異常檢測工具箱 (Algorithm Map)

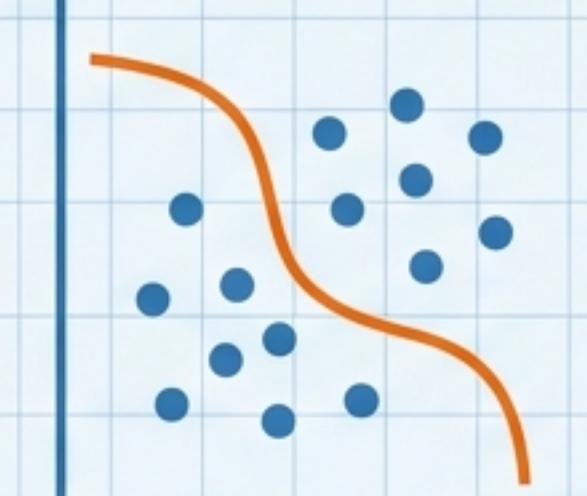
Isolation Forest (孤立森林)



基於樹
(Tree-based)

Superpower: 高效、適合高維大數據

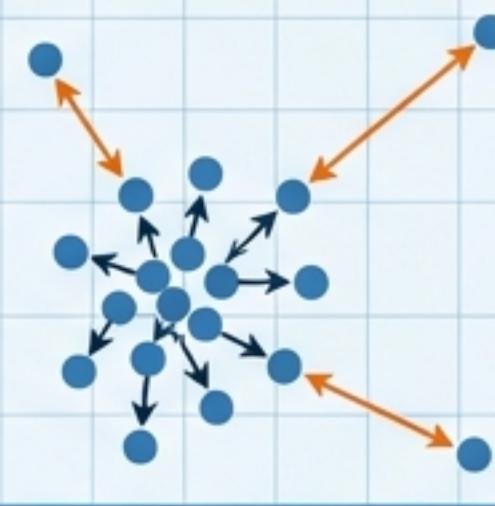
One-Class SVM (一類支持向量機)



基於邊界
(Boundary-based)

Superpower: 精確邊界、適合小樣本

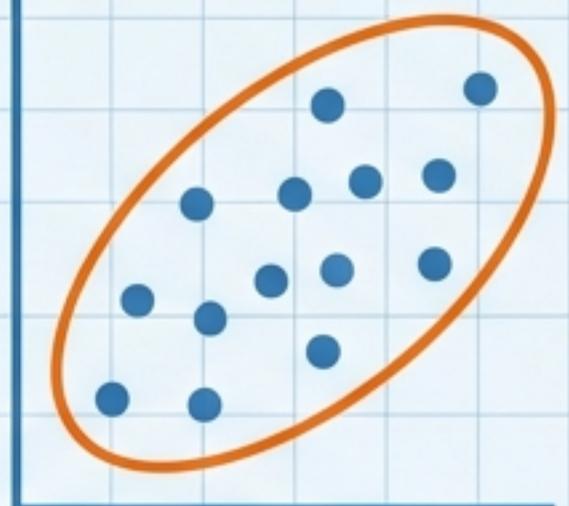
Local Outlier Factor (LOF)



基於密度
(Density-based)

Superpower: 處理密度不均、局部異常

Elliptic Envelope (橢圓包絡)



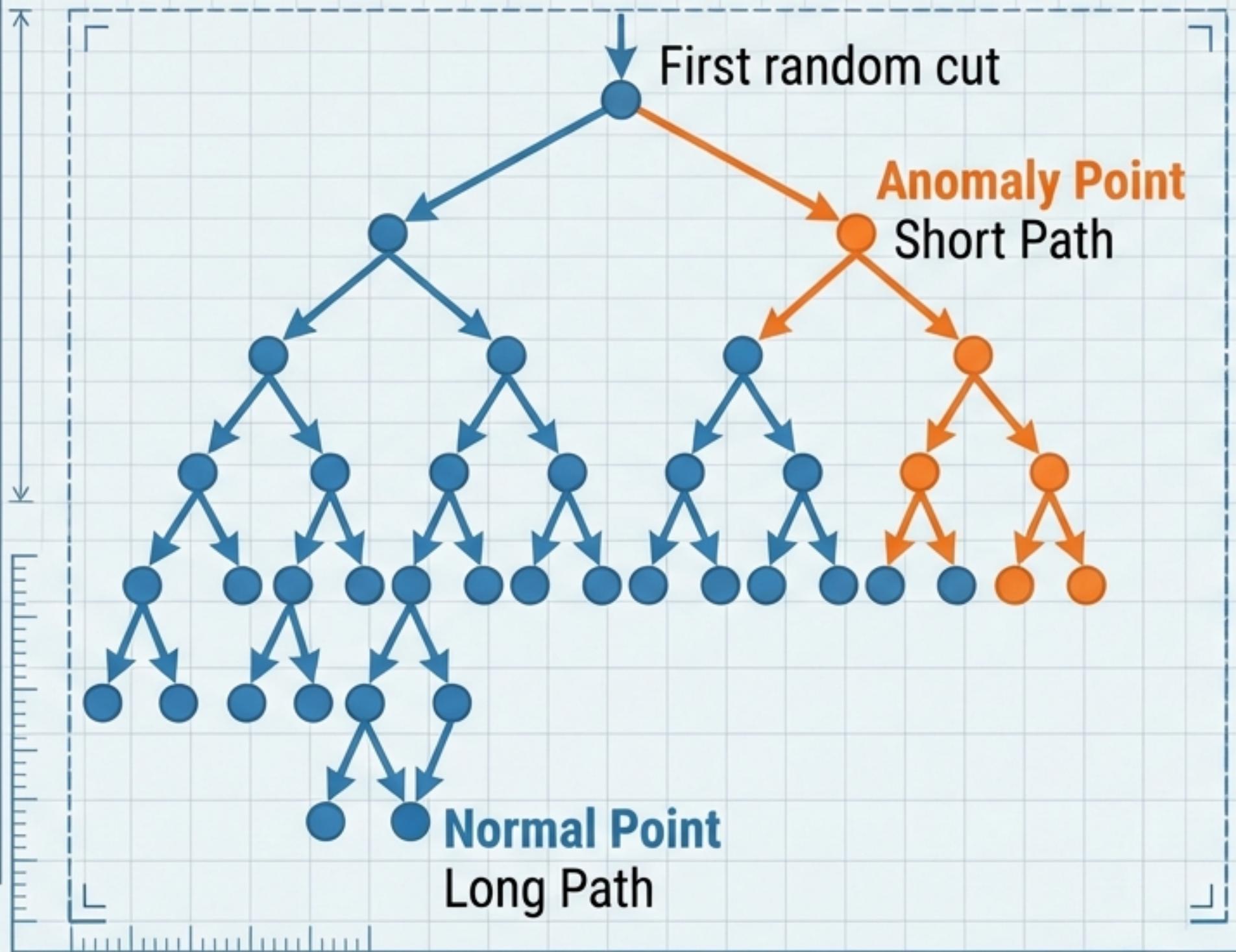
基於統計
(Statistics-based)

Superpower: 快速、高可解釋性 (Gaussian)

不同的化工數據特性（維度、分布、樣本數）決定了演算法的選擇。

深度解析 1：孤立森林 (Isolation Forest)

大規模製程數據的監控首選



- 核心思想：異常點“容易被孤立”
(Anomalies are easier to isolate)。
- Pros: $O(n \log n)$ 高效運算、不需假設分布、適合高維數據。

化工應用場景

蒸餾塔監控 (50+ sensors, T/P/F/L)。數據量大，變數多，需要即時反應。



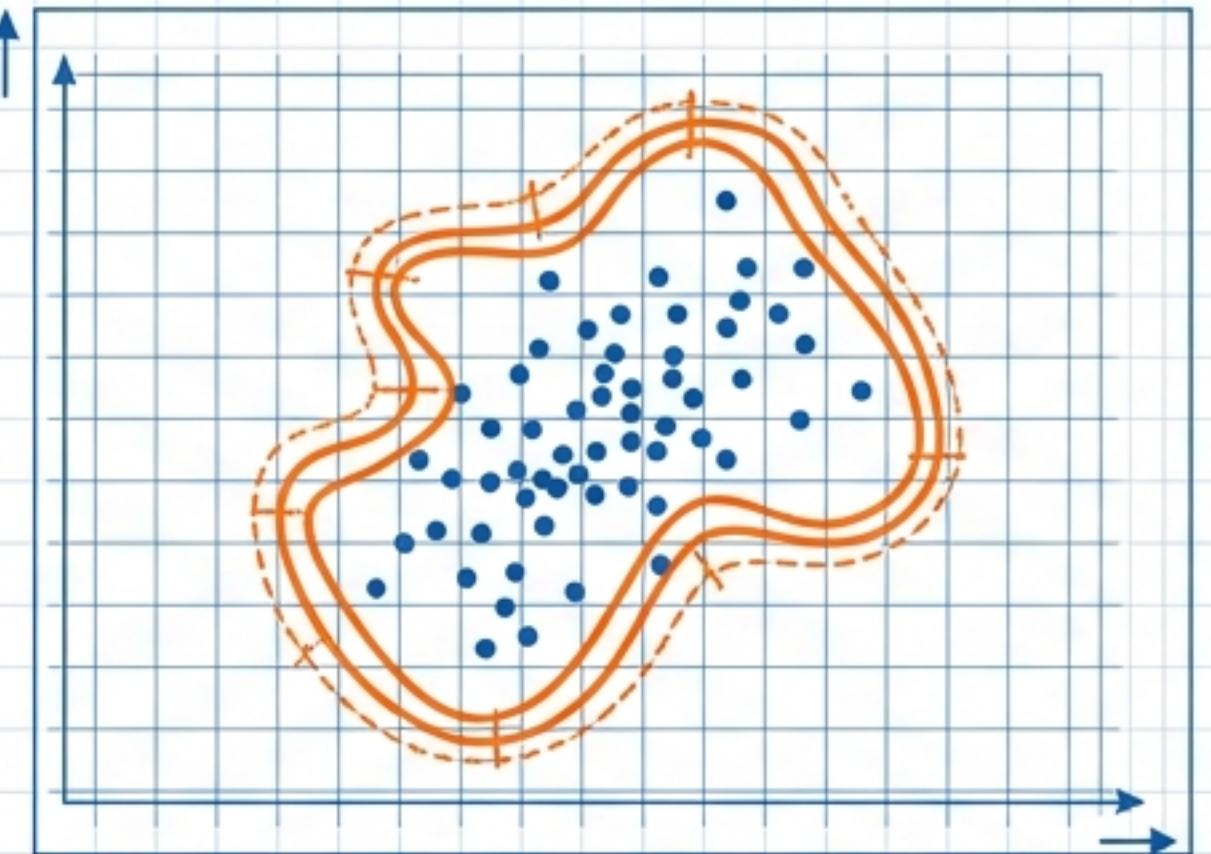
`sklearn.ensemble.IsolationForest`



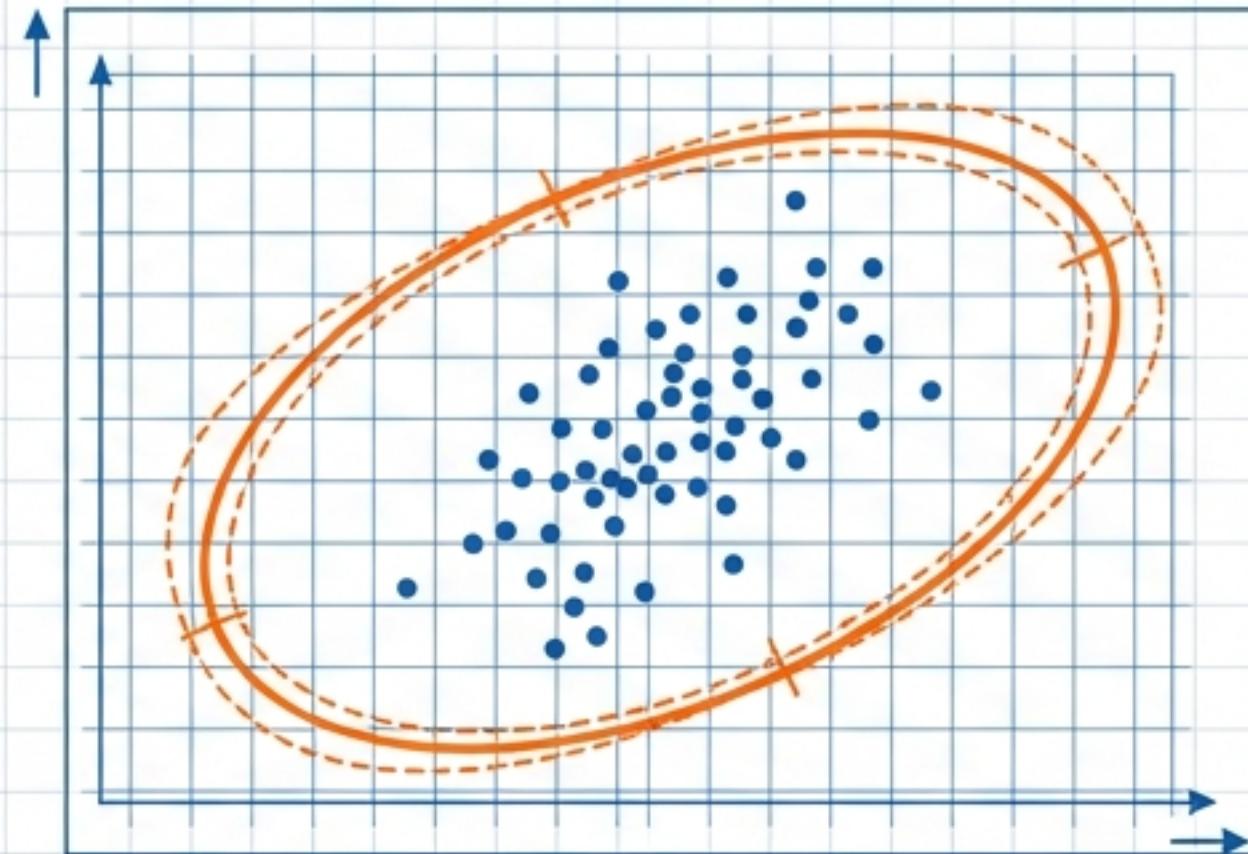
深度解析 2：邊界與統計方法 (Boundary & Statistical Methods)



One-Class SVM



Elliptic Envelope



Concept：尋找包覆正常數據的最小超球面。

Pros：非線性、邊界精確。

Cons：計算成本高，不適合大數據。



化學應用場景：高價值藥品批次監控
(小樣本、高精度)。

Concept：假設高斯分布，建立橢圓邊界 (Mahalanobis Distance)。

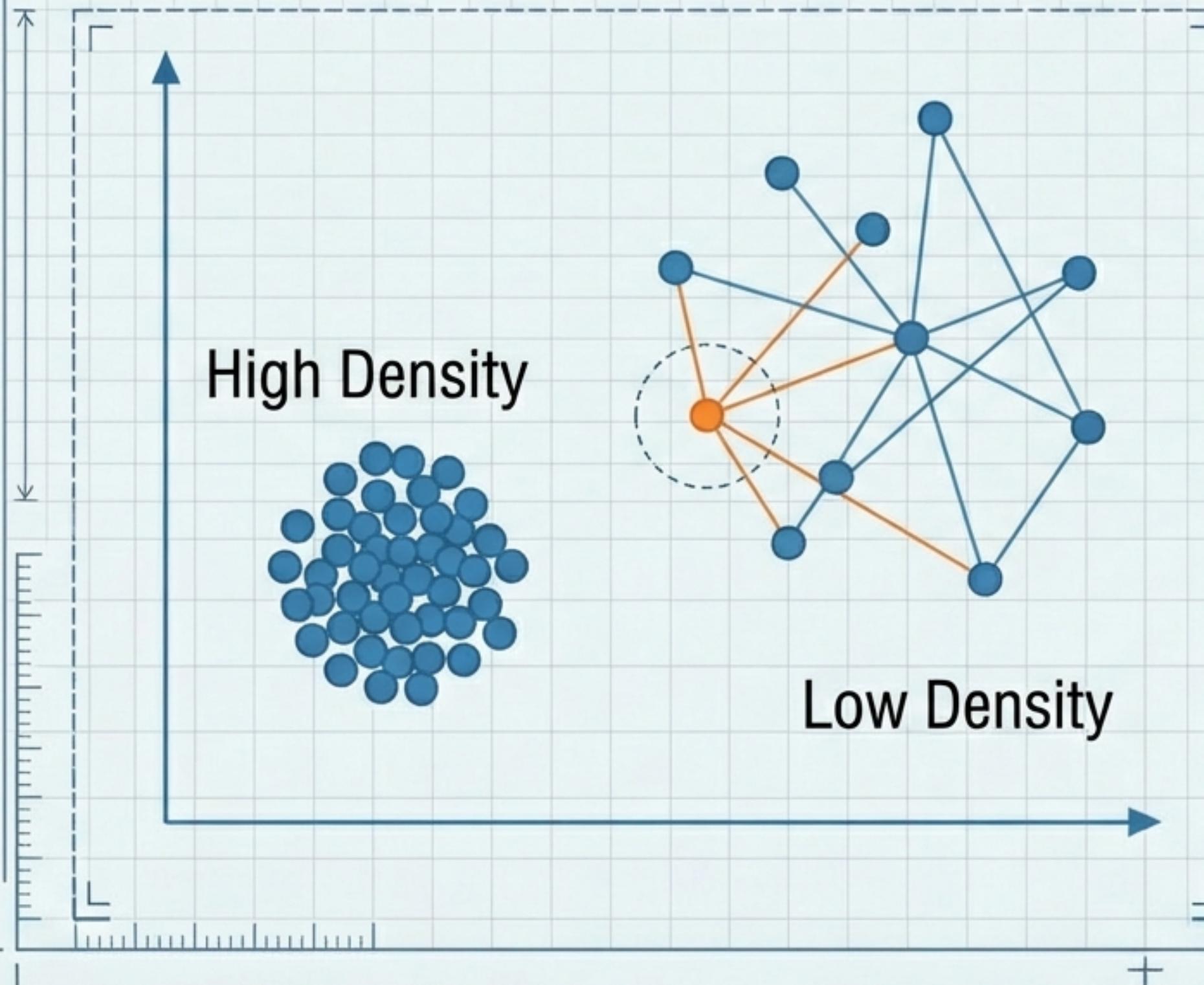
Pros：極快、物理意義明確。



化學應用場景：穩態操作下的產品品質管制
(Steady-state QC)。

深度解析 3：區域性離群因子 (Local Outlier Factor, LOF)

解決多模式與密度不均問題



- 核心思想：並非看「絕對距離」，而是看「相對密度」。
- Rule: $\text{LOF} >> 1$ implies Anomaly.

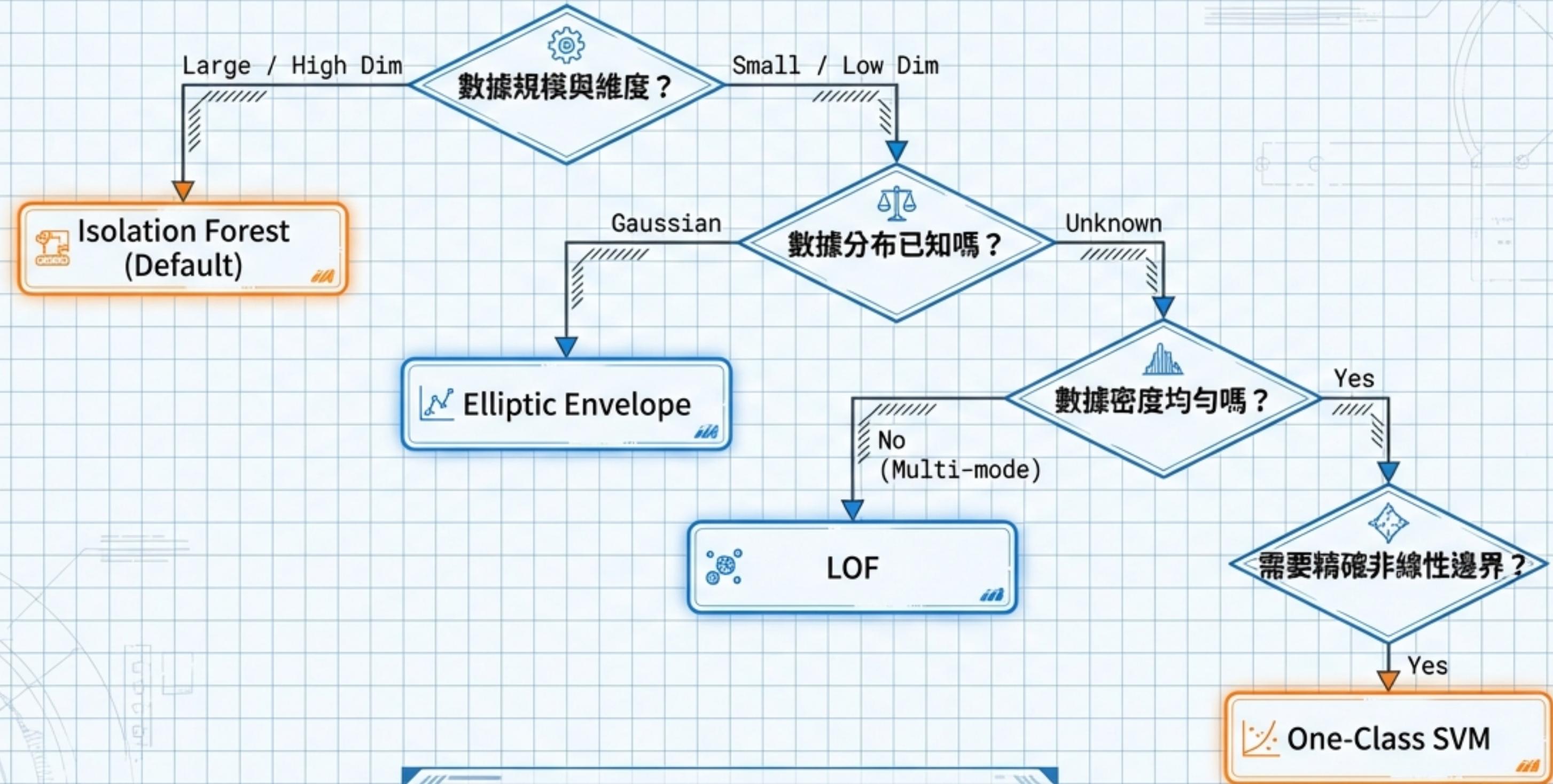
化工應用場景

多模式操作 (Multi-mode Operation)
- 如開車、停車、穩態。不同階段密度不同，單一閾值無法適用。



sklearn.ensemble.IsolationForest

決策指南：如何選擇合適的演算法？(Decision Guide)

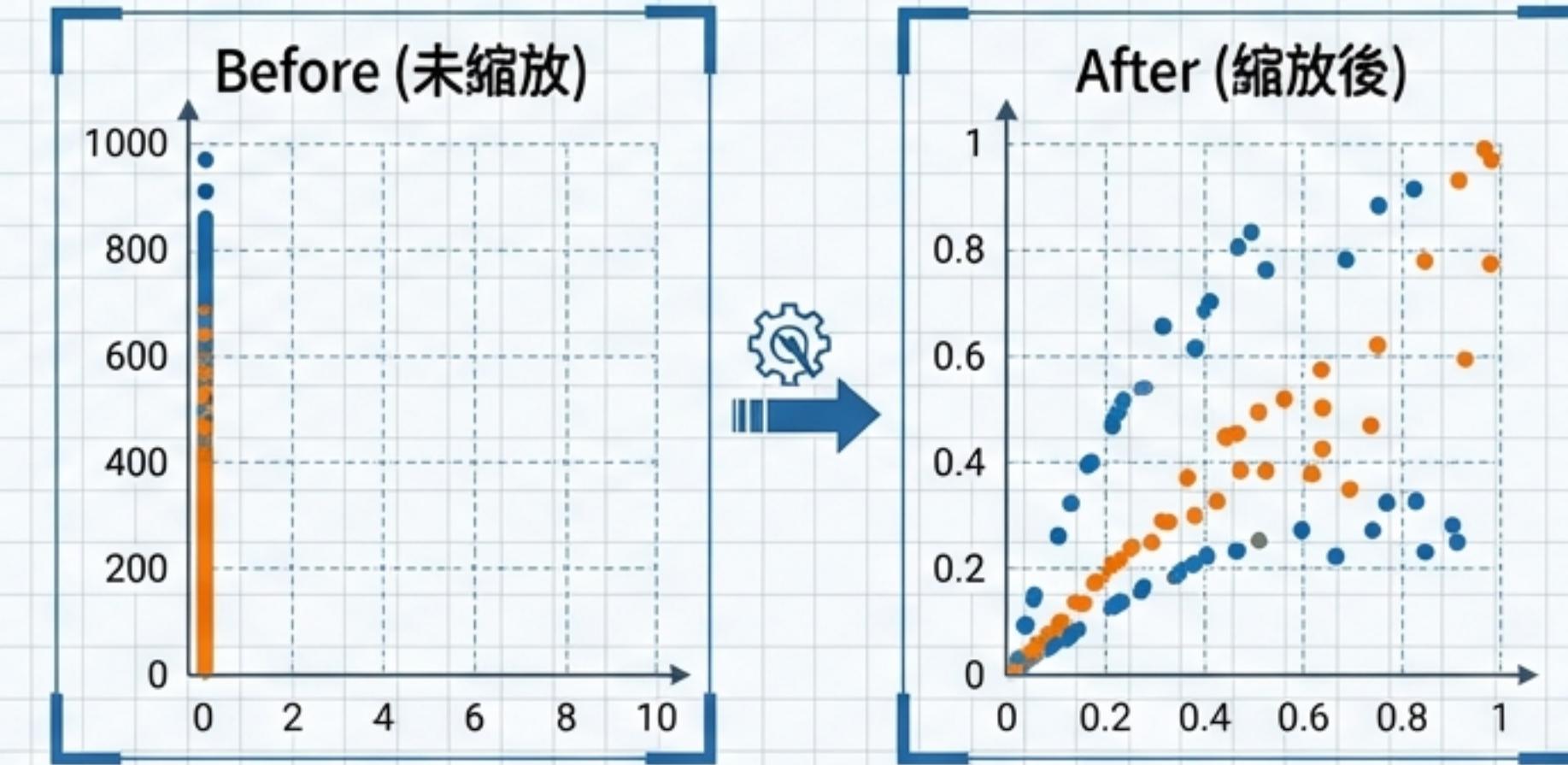


Big Data/IoT -> Isolation Forest
Quality Control -> Elliptic Envelope / SVM
Complex Process -> LOF

工作流程 Part 1：關鍵前處理 (Critical Preprocessing)

⚠ Garbage In, Garbage Out : 為什麼異常檢測對尺度極度敏感？

- Concept: 距離計算受數值大小影響 (溫度 300 vs 壓力 10)。
- Techniques:
 1. **StandardScaler (z-score)**: 預設選項 (SVM, LOF, Envelope)。
 2. **MinMaxScaler**: 保留邊界。
 3. **RobustScaler**: 當訓練集包含離群值時使用 (Important!)。



⚠ 資料洩漏 (Data Leakage): Fit on Train, Transform on Test. 嚴禁在訓練時看到測試集的分布。

模型評估：當擁有歷史標籤時 (With Historical Labels)

(Model Evaluation: Supervised with Confusion Matrix)

	Predicted Normal	Predicted Anomaly
Actual Normal	真陰性 (True Negative) 	誤報 (False Alarm) - Efficiency Loss
Actual Anomaly	漏報 (Missed Alarm) - Safety Incident Risk 	真陽性 (True Positive)

Wasted Resources,
Operator Fatigue

High Cost,
Safety Concern

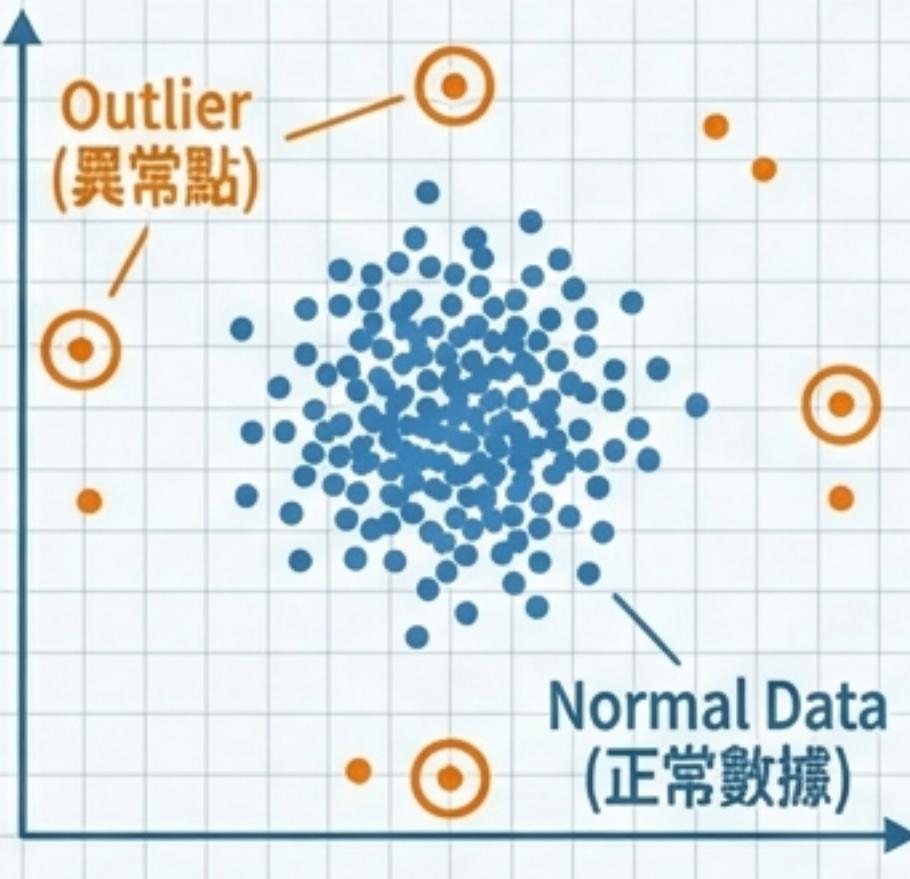
評估指標定義 (Metrics Definitions)

- **Recall (召回率)**：寧可殺錯，不可放過 (Crucial for Safety)。
- **Precision (精確率)**：避免“狼來了”效應 (Crucial for Efficiency)。
- **F1-Score**: 平衡點。
- **ROC-AUC**: 整體效能指標。

ChemE Context: 在製程安全中，我們通常優化 Recall；在維護建議中，我們優化 Precision。

模型評估：當沒有標準答案時 (Without Labels)

Strategy 1: 視覺化 (Visualization)



Strategy 2: 領域專家驗證 (Expert Validation)



Strategy 3: 歷史回溯 (Back-testing)



使用 t-SNE 或 PCA 降維，檢查異常點是否位於邊緣或稀疏區。

Human-in-the-loop: 將 Top-20 異常點交給製程工程師確認。

對照過去一年的維修紀錄。模型是否在故障發生前發出訊號？

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

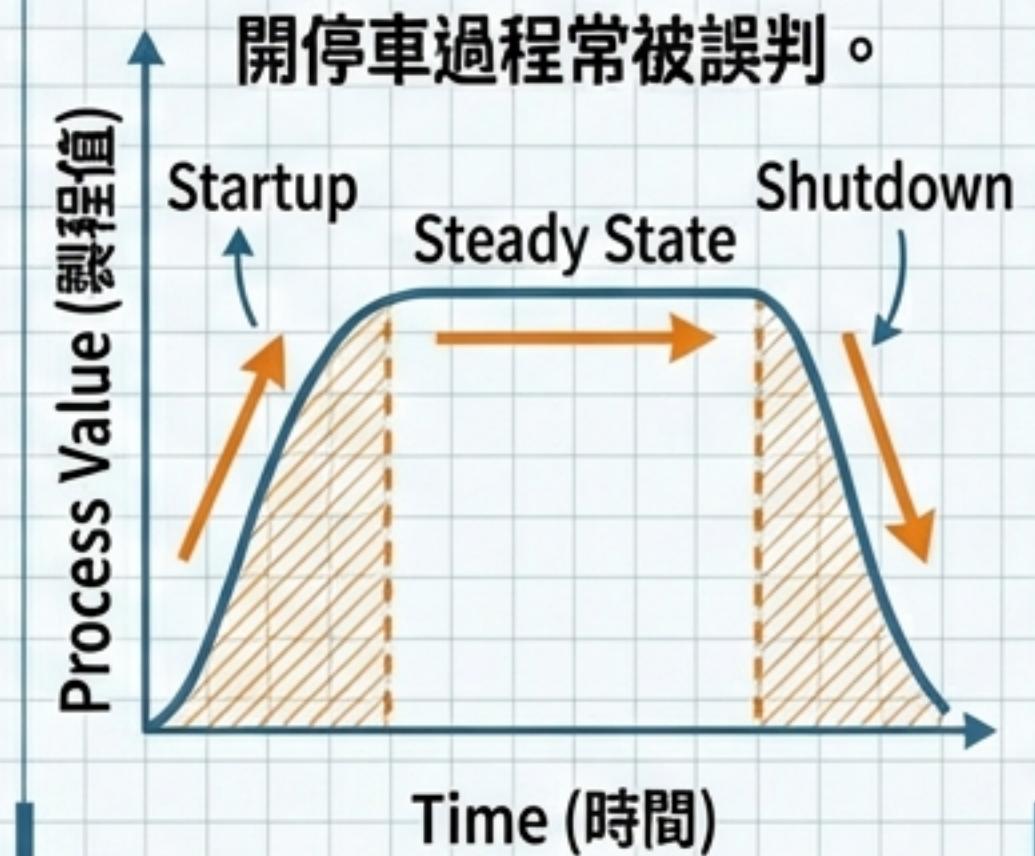
成本不對稱性 (Cost Asymmetry)



告警疲勞 (Alarm Fatigue)



非穩態操作 (Non-Steady State)



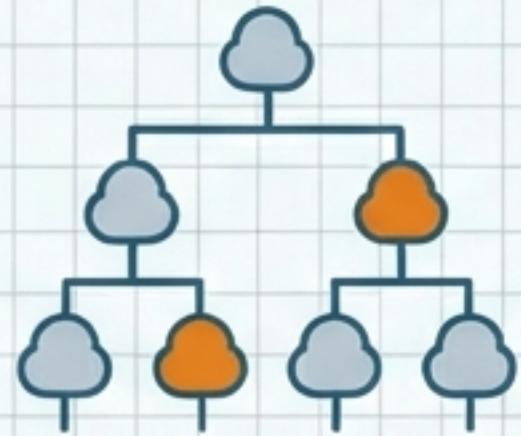
Solution: 調整閾值，接受較高誤報率以換取安全。

Solution: 告警抑制 (Alarm Suppression)
—連續 N 點異常才報警。

Solution: 結合狀態變數
或分段建模。

最佳實踐與總結 (Best Practices & Summary)

1. 從簡單開始 (Start Simple)



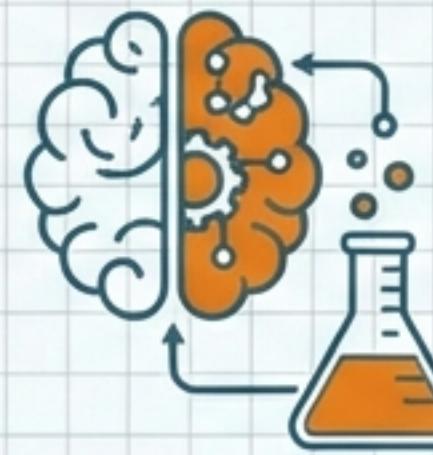
先用 Isolation Forest 或 Envelope 建立基準。

2. 重視前處理 (Data Prep)



標準化 (Scaling) 是成敗關鍵。

3. 結合領域知識 (Expertise)



讓工程師參與評估，解釋異常原因。

4. 動態更新 (Iterate)



製程會老化，模型需定期重訓練。

異常檢測不是為了取代操作員，而是為了給他們 "第三隻眼" (A third eye)。

結語：從發現問題到解決問題

異常檢測是 AI 在化工最直接的應用。
它連結了 數據 (Data) 與 安全 (Safety)。

前往 Unit 08: 實作與比較 (Implementation)

Preview: 我們將使用 Python 實作 Isolation Forest
與 One-Class SVM，並調整參數觀察結果。

