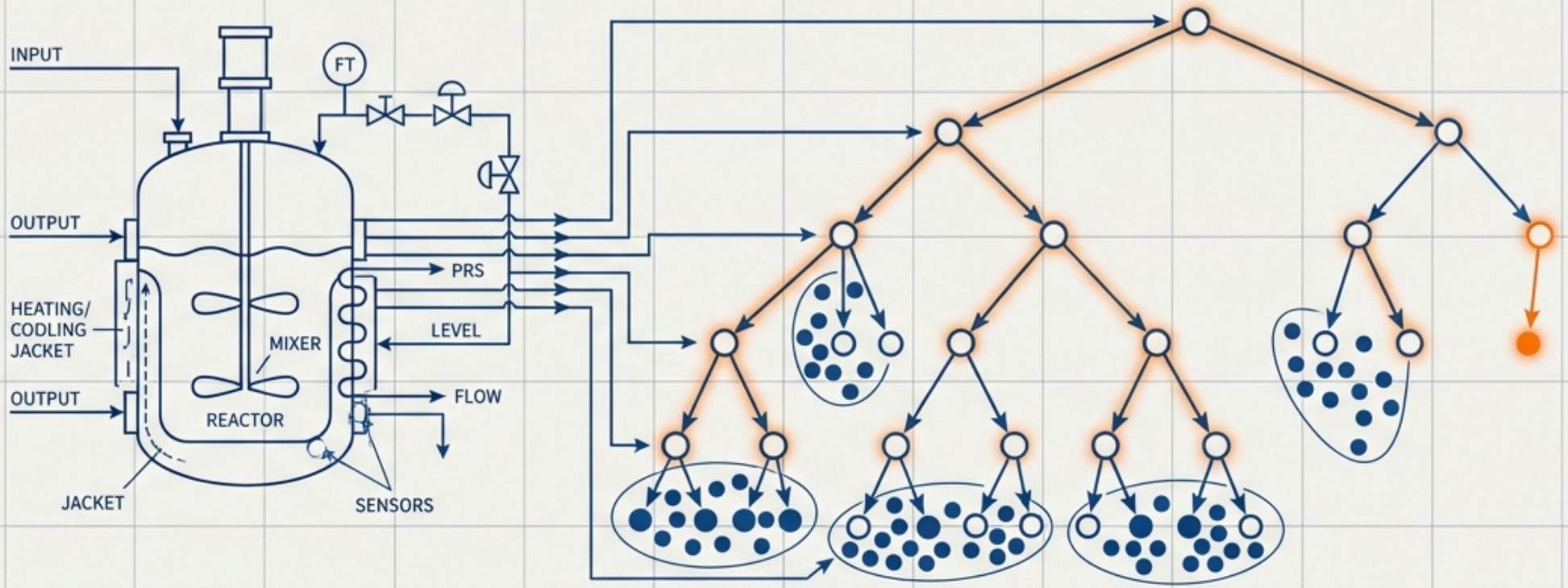
Noto Sans TC Roboto Mono

Unit 07 孤立森林 (Isolation Forest)

化工異常檢測實戰：從原理到 CSTR 反應器監控



授課教師：莊曜禎 助理教授

製作單位：逢甲大學 化工系
智慧程序系統工程實驗室

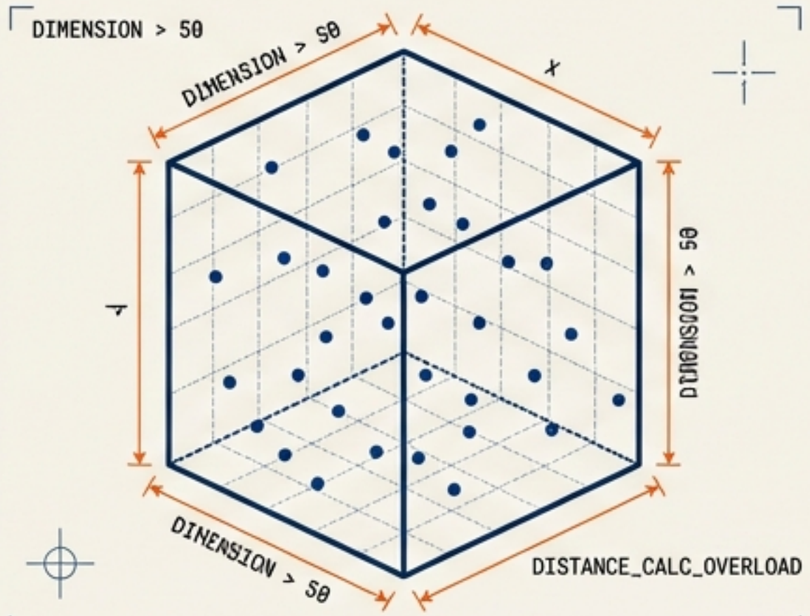
課程目標：掌握高維度製程數據的異常檢測技術，建立早期預警系統。

NotebookLM

為什麼選擇孤立森林？應對高維度數據的挑戰

傳統方法 (k-NN, LOF)

The Problem (高維度的詛咒)

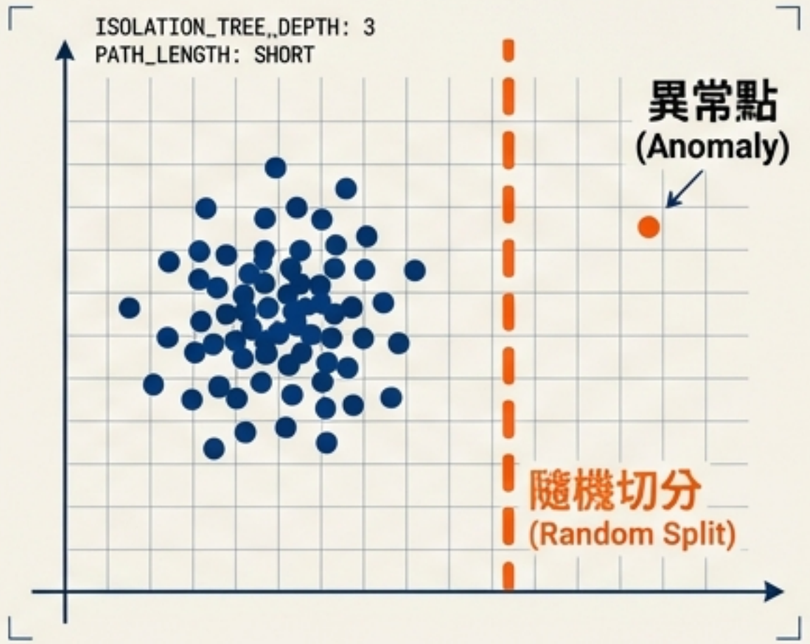


依賴計算點與點之間的「距離」。
當變數 > 50 (溫度、壓力、流量...)
計算成本極高且效率低落。

典範轉移
(Paradigm Shift)

孤立森林 (Isolation Forest)

The Solution

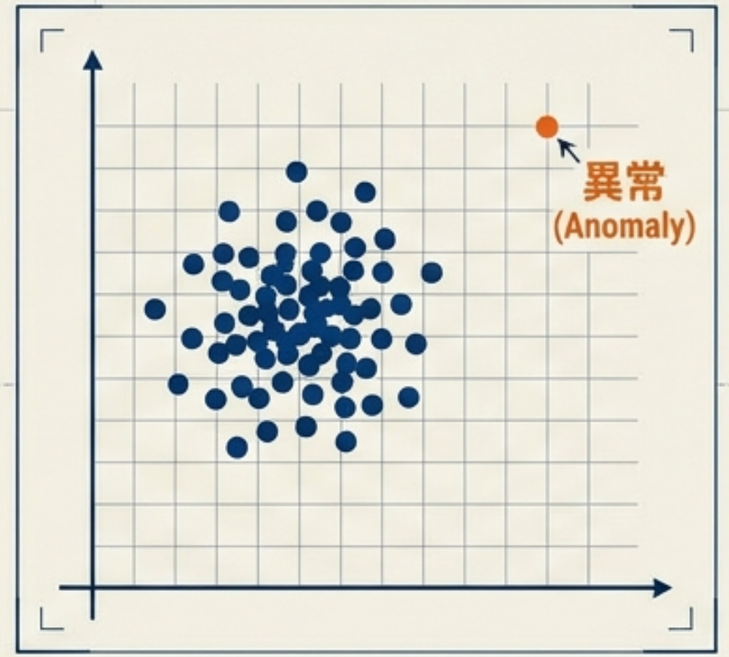


利用「稀少且不同」的特性。
異常數據點更容易被隨機切分「孤立」。

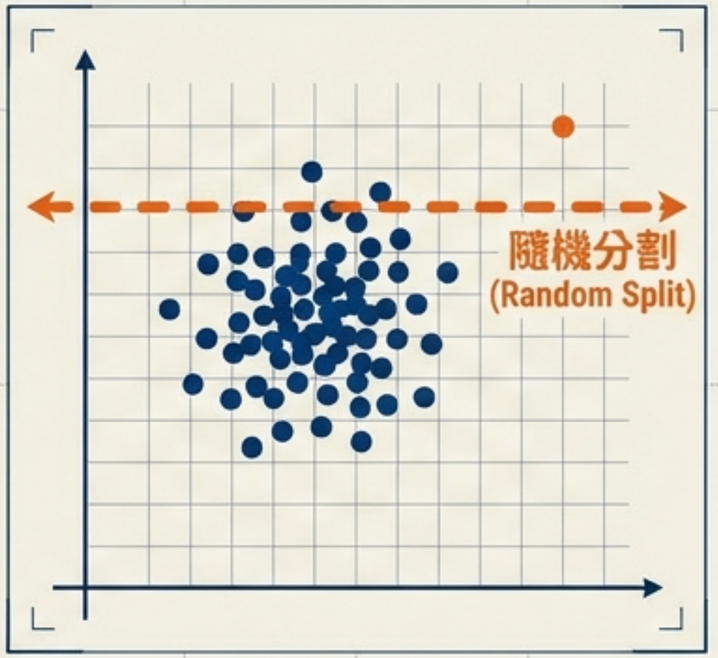
NotebookLM

核心直覺：隨機切割與路徑長度

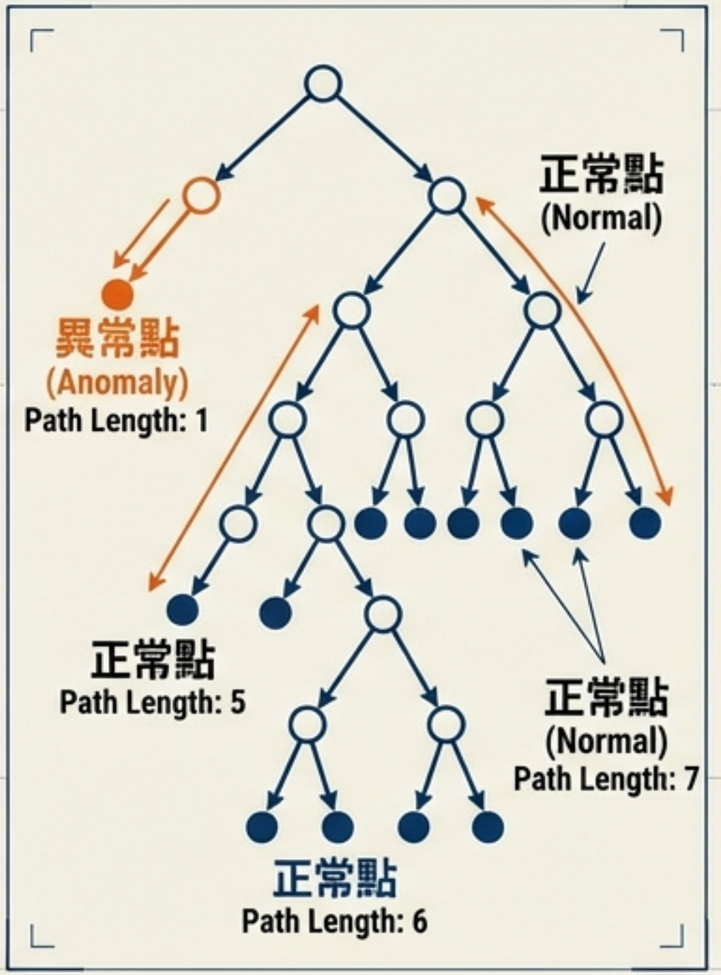
Step 1



Step 2



Result



Key Metric

路徑長度（Path Length）：

- **短路徑**（Short Path）→ **異常點**（容易被切分）
- **長路徑**（Long Path）→ **正常點**（深埋在群體中）

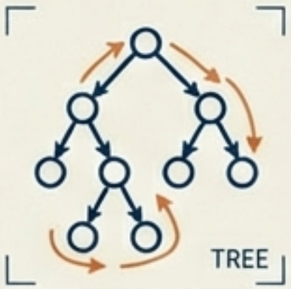
NotebookLM

演算法機制：孤立樹 (Isolation Tree) 的建構

建構流程 (Process)

1. 隨機選擇特徵 (Random Feature)
從所有製程變數 (q) 中隨機選一個。

2. 隨機選擇分割點 (Random Split)
在 Max 和 Min 之間隨機選一個值 (p)。



3. 遞迴分割 (Recursive Partitioning)
重複直到節點剩一個樣本或達最大深度。

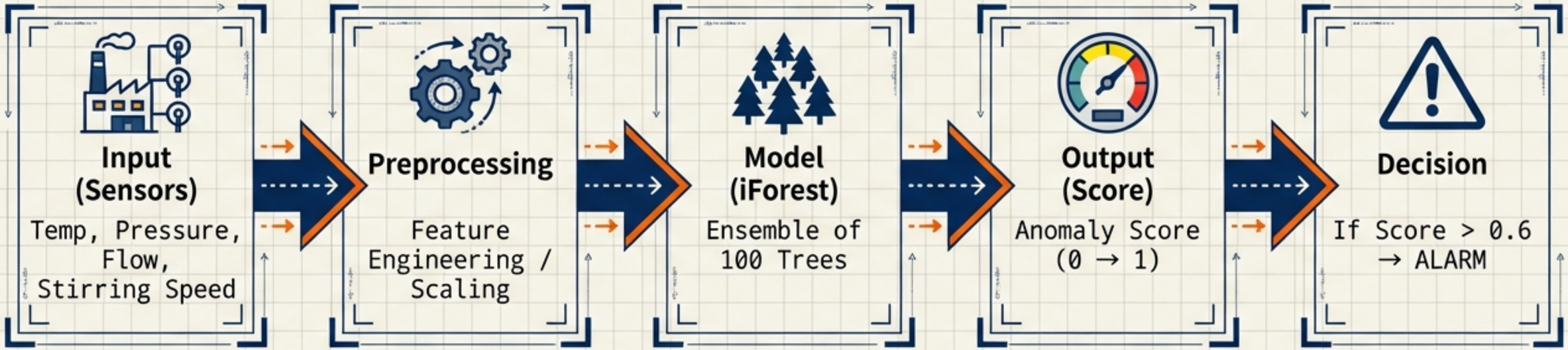
異常分數公式 (The Math)

$$s(x,n) = 2^{-\frac{E[h(x)]}{c(n)}}$$

- $h(x)$：路徑長度 (Path Length)
- $c(n)$：平均路徑長度的期望值 (Normalization)
- 判定規則：若 $s \approx 1$ 則為高度異常 (High Anomaly Score)

NotebookLM

工程實作：從感測器數據到異常評分



關鍵優勢 (Key Advantage)

無需標籤數據 (Unsupervised Learning)。直接使用歷史數據進行訓練。

NotebookLM
NotebookLM

Python 實作：Scikit-learn `IsolationForest´

CODE EDITOR - ISOLATION_FOREST_PYTHON.PY

```python
from sklearn.ensemble import IsolationForest

# 1. 建立模型
iso_forest = IsolationForest(
    n_estimators=100,      # 樹的數量
    contamination=0.05,    # 預期異常比例
    max_samples='auto',    # 子樣本大小
    random_state=42
)

# 2. 訓練模型 (傳入正常數據)
iso_forest.fit(X_train)

# 3. 預測與評分
y_pred = iso_forest.predict(X_test)
scores = iso_forest.decision_function(X_test)
```

核心 API 解析

`fit(X)`

非監督學習，不需要標籤 y。模型自動學習數據的結構。

`predict(X)`

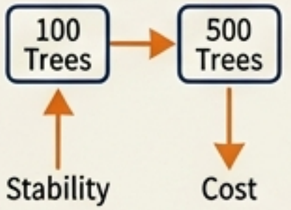
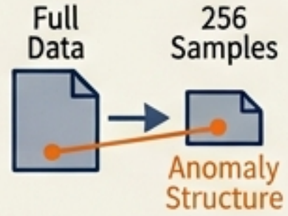
回傳 {-1, 1}。
-1 代表異常，1 代表正常。

`decision_function(X)`

回傳原始異常分數。這對於微調閾值 (Threshold) 至關重要。

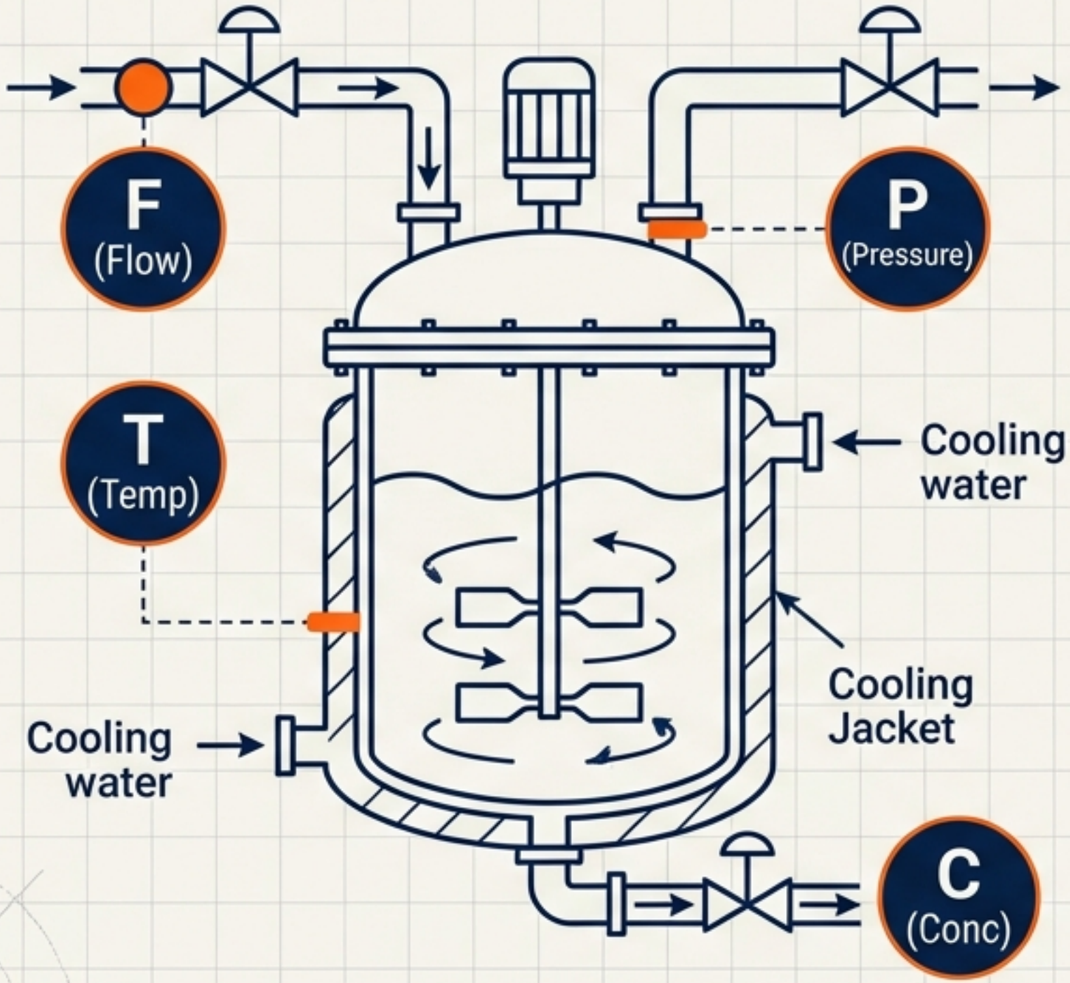
NotebookLM

調校引擎：關鍵超參數 (Hyperparameters)

關鍵參數 (CRITICAL PARAMETER)		工程影響 (ENGINEERING IMPACT)	
`contamination` (最關鍵)	預期數據中的**異常比例** (Expected Anomaly Rate)。		決定了**閾值** (Threshold) 的切分位置。若要求高安全性（寧可誤報），設**高一點** (e.g., 0.1)；若要減少干擾，設**低一點** (e.g., 0.01)。
`n_estimators`	森林中樹的數量。		通常 100 棵足夠。更多樹 = 結果更穩定，但計算成本增加。 
`max_samples`	建構每棵樹時的抽樣樣本數。		不需要使用全量數據。256 個樣本通常足以捕捉異常結構 (Sampling Efficiency)。 

NotebookLM

實戰案例：CSTR 反應器異常監控

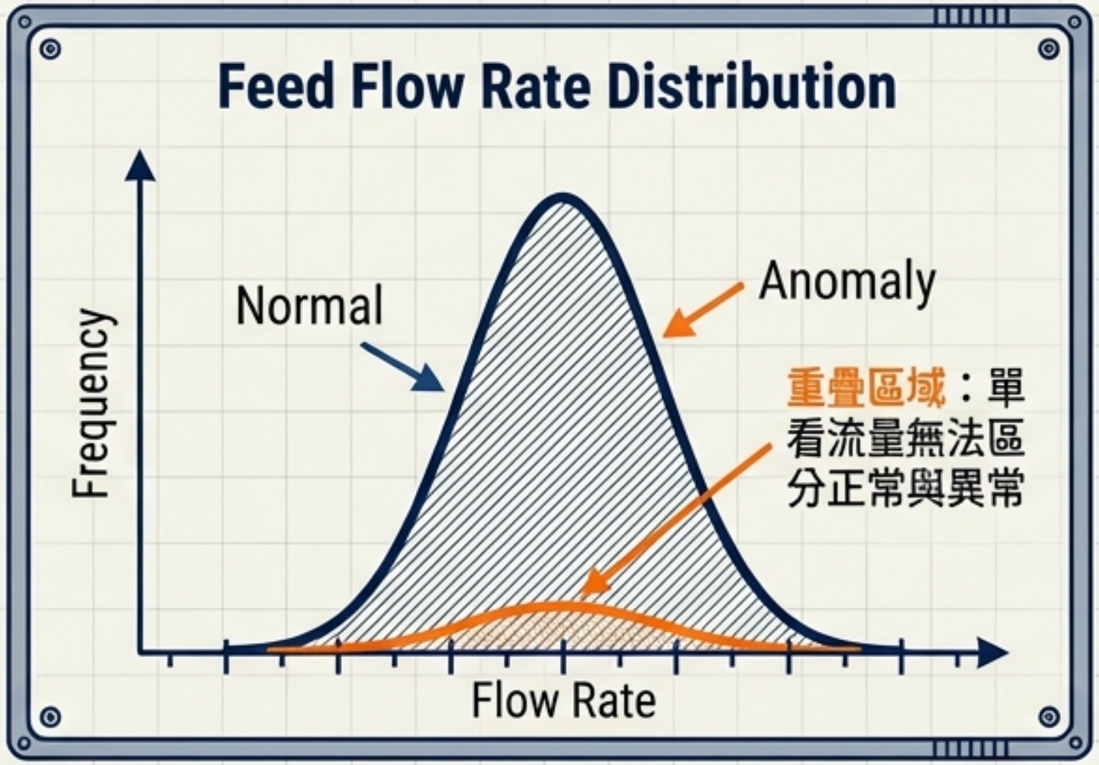


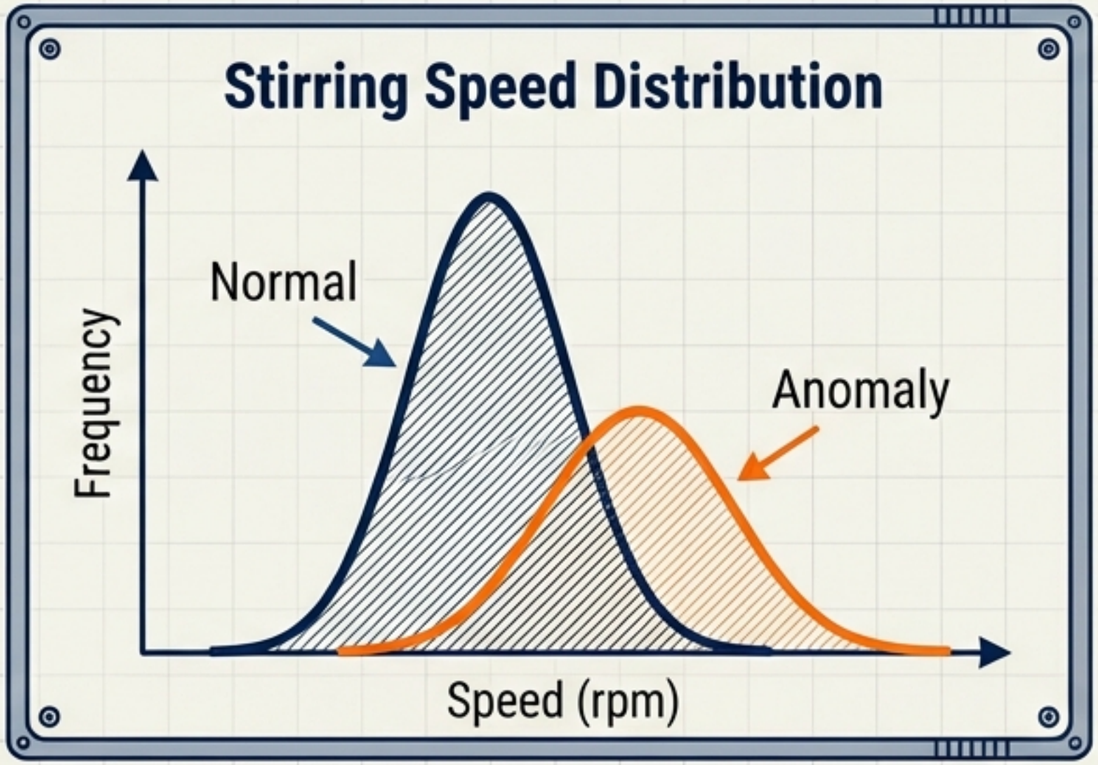
System Specification 'Roboto Mono'

- **場景**：連續攪拌槽反應器，需防止**熱失控**。
- **監控變數 (6 Dimensions)**：
  1. Reaction Temperature (T)
  2. Reactor Pressure (P)
  3. Feed Flow Rate
  4. Cooling Water Flow
  5. Stirring Speed (rpm)
  6. Product Concentration (C_A)
- **挑戰**：**軟異常** (Soft Anomalies) —— 單一變數看起來正常，但變數間的關係已破壞。

NotebookLM

數據探索：為什麼單變量分析不夠？

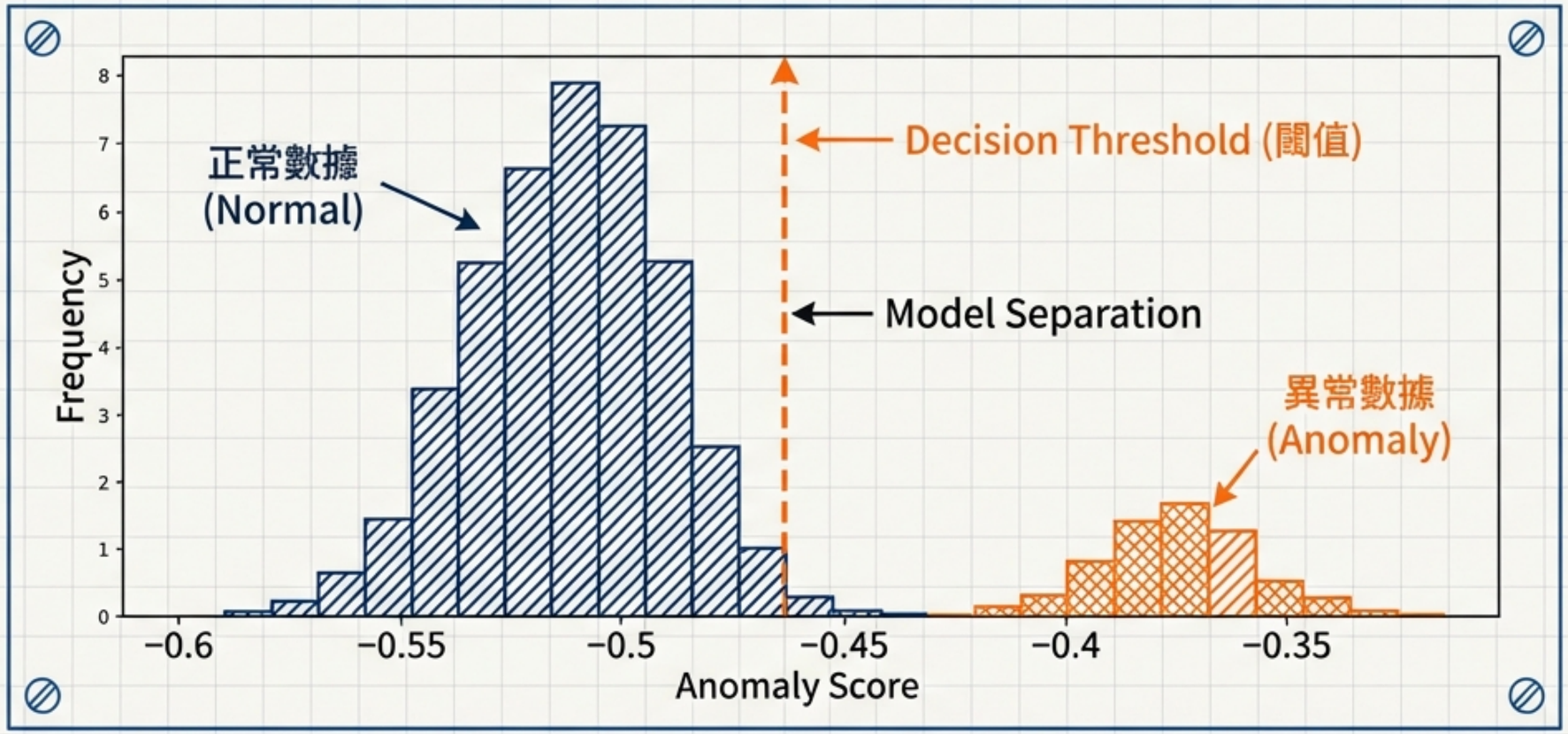




結論 (Conclusion)：單變量分析 (Univariate) 容易失效。我們需要捕捉 6 個變數之間的交互作用 (Interactions)，這正是孤立森林擅長的領域。

NotebookLM

訓練結果：異常分數分布 (Anomaly Score Distribution)



模型成功將異常點推向分數較低（或較高，視實作定義）的極端區域，形成可切分的邊界。

NotebookLM

效能評估：混淆矩陣與代價 (Confusion Matrix)

	Predicted Normal (預測正常)	Predicted Anomaly (預測異常)
Actual Normal (實際正常)	True Negative (TN): 549 (Correct Normal)	False Positive (FP): 21 誤報 (操作員困擾)
Actual Anomaly (實際異常)	False Negative (FN): 19 漏報 (安全風險)	True Positive (TP): 11 (Detected Anomaly)

Performance Metrics

Roboto Mono

Recall (召回率): 36.67%
(偏低 – 需改進)

Precision (精確率): 34.38%

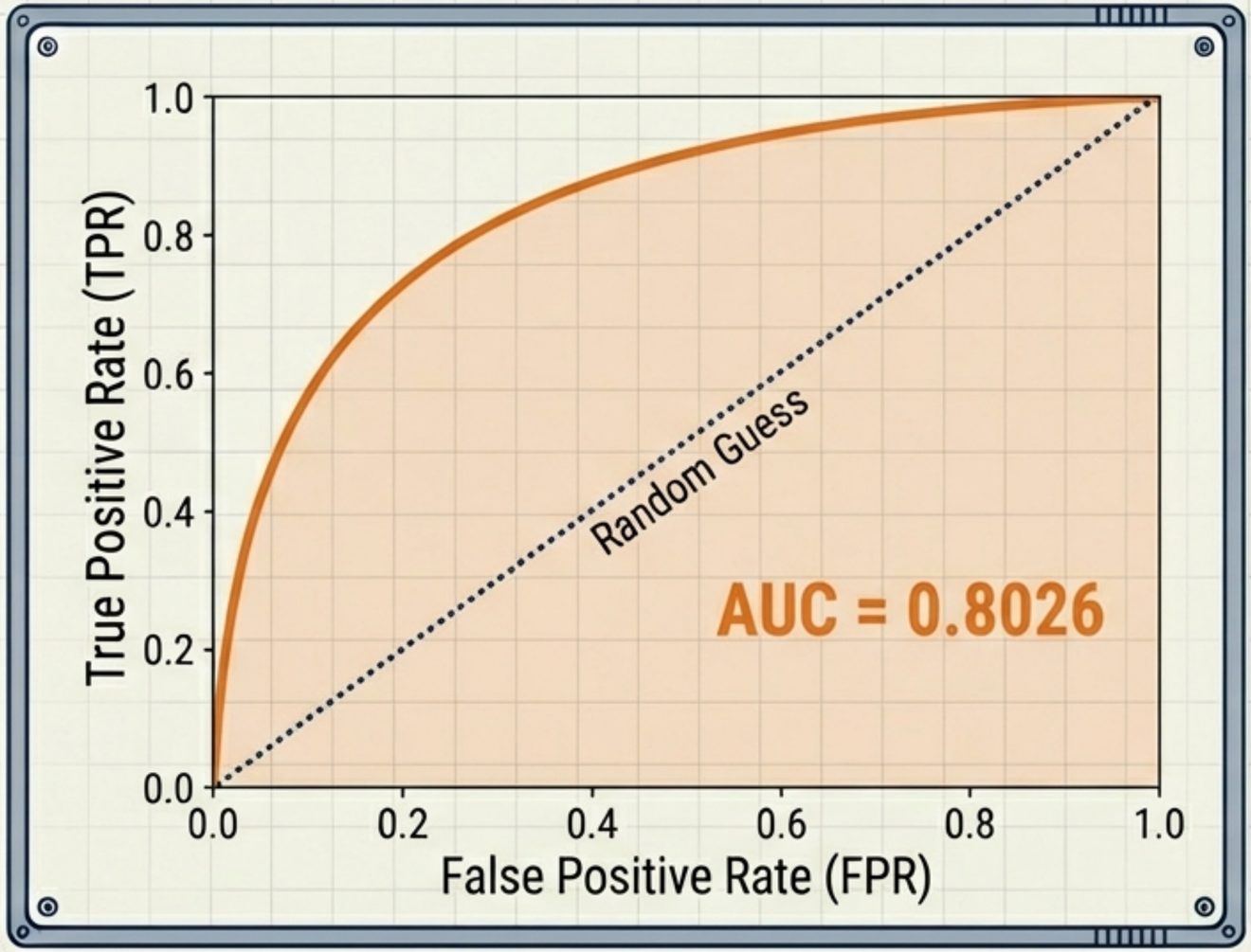
Noto Sans TC

Engineering Insight:

目前 `contamination=0.05` 設定過於保守。為了提升安全性，我們應降低閾值以捕捉更多異常 (減少 FN)。

NotebookLM

進階評估：ROC 曲線與 AUC



AUC (Area Under Curve) 解讀:

Roboto Mono

- 0.80 - 0.90: 良好模型 (Good)
- > 0.50: 優於隨機猜測

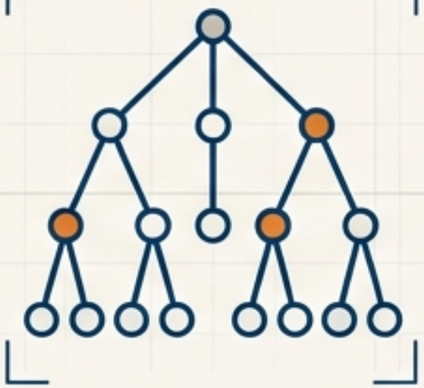
Noto Sans TC

Takeaway: 雖然 F1-score 受閾值影響看起來較低，但 AUC 證明模型本身對異常的排序能力是強健的 (Robust)。

NotebookLM

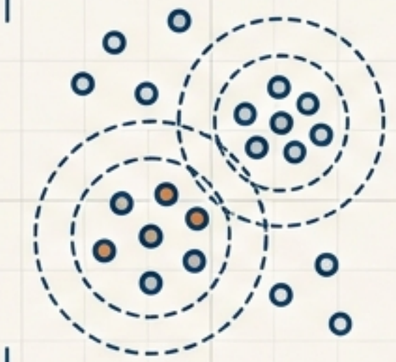
方法比較：如何選擇正確的工具？

Isolation Forest



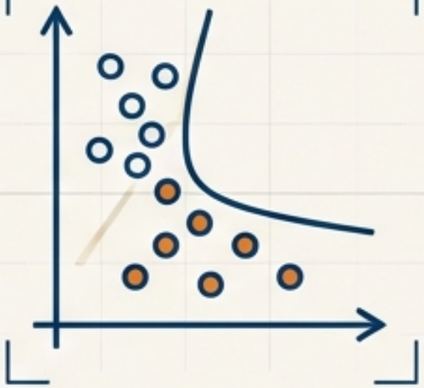
- Pros: 速度快 ($O(n \log n)$)，適合高維度，全局異常
- Cons: 對群聚型異常較弱
- Best For: 大數據、即時監控

Local Outlier Factor (LOF)



- Pros: 擅長檢測局部密度異常
- Cons: 計算慢 ($O(n^2)$)，受維度災難影響
- Best For: 小數據、複雜分布

One-Class SVM

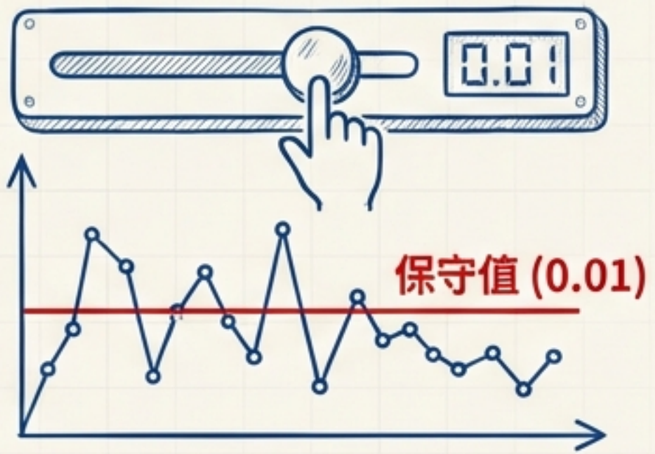


- Pros: 邊界定義精確
- Cons: 訓練極慢，參數敏感
- Best For: 小樣本、非高斯分布

NotebookLM

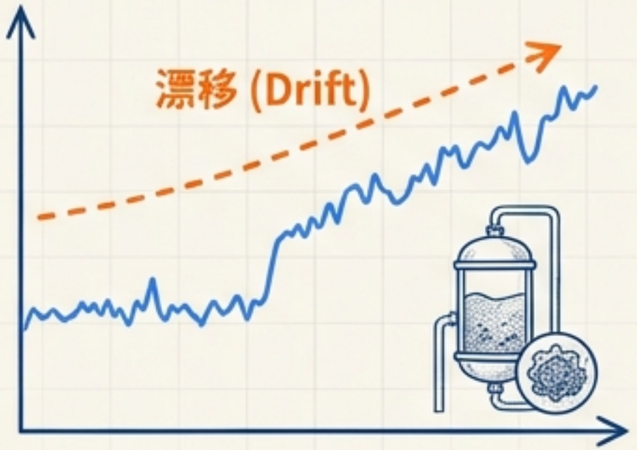
實務挑戰：化工廠不是 Kaggle 競賽
(Practical Challenges: Chemical Plants Are Not Kaggle Competitions)

閾值設定難題
(The Threshold Dilemma)



現實中沒有標籤。
Solution: 從保守值 (0.01) 開始，結合**領域專家** (**Expert**) 審查。

概念漂移
(Concept Drift)



觸媒老化導致正常定義改變。
Solution: 定期**重訓練** (**Retraining**) 或使用滑動視窗。

告警疲勞
(Alert Fatigue)

誤報過多會被忽略。
Solution: **告警抑制** (**Alert Suppression**) — 連續 **N** 點異常才發報。

NotebookLM

結語與下一步 (Key Takeaways & Next Steps)

Summary Box

- 高效 (Efficient): 訓練只需 0.2 秒，適合即時系統。
- 強大 (Powerful): 無需標籤即可處理高維度交互作用。
- 靈活 (Flexible): 可根據安全需求調整靈敏度。

前往 Unit07_Isolation_Forest.ipynb

任務 (Task):
親手執行程式碼，嘗試調整 `contamination` 參數，觀察 Precision/Recall 的變化。

AI 不會取代工程師，但懂得使用 AI 的工程師將取代不懂的人。

NotebookLM