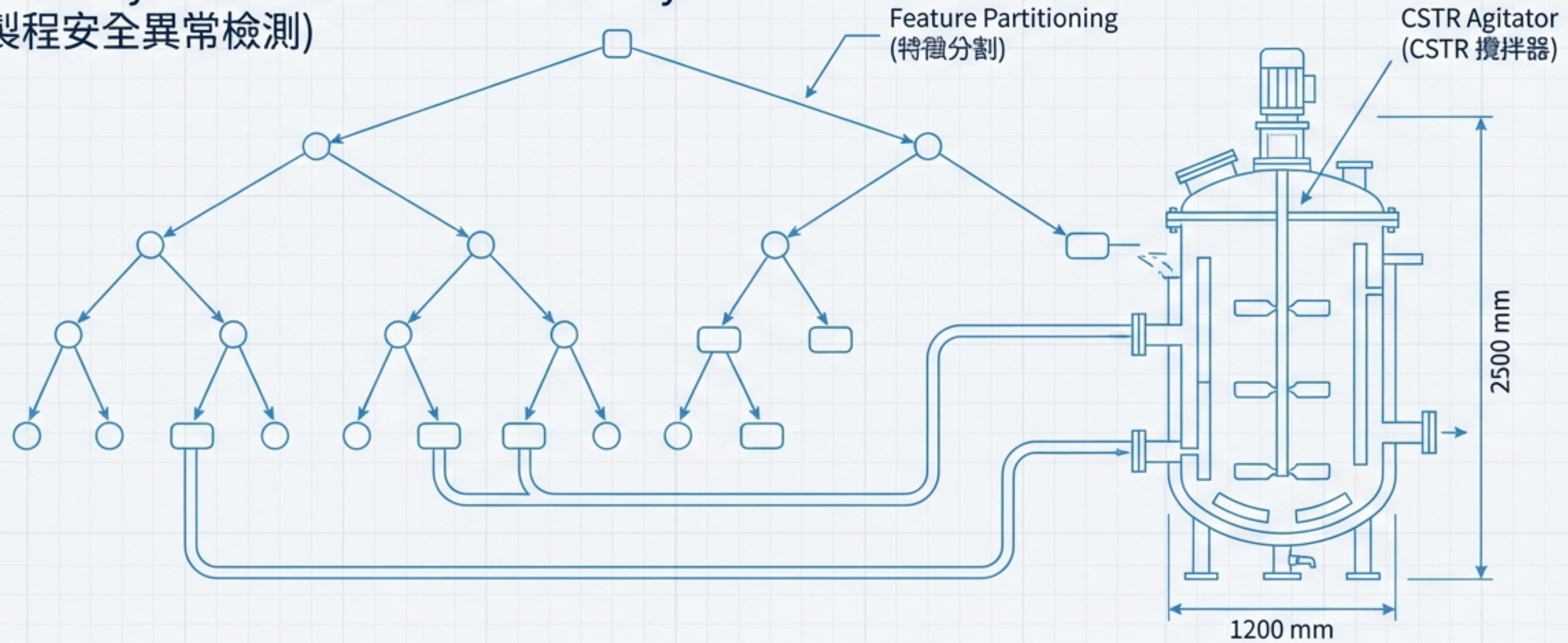


Unit 07

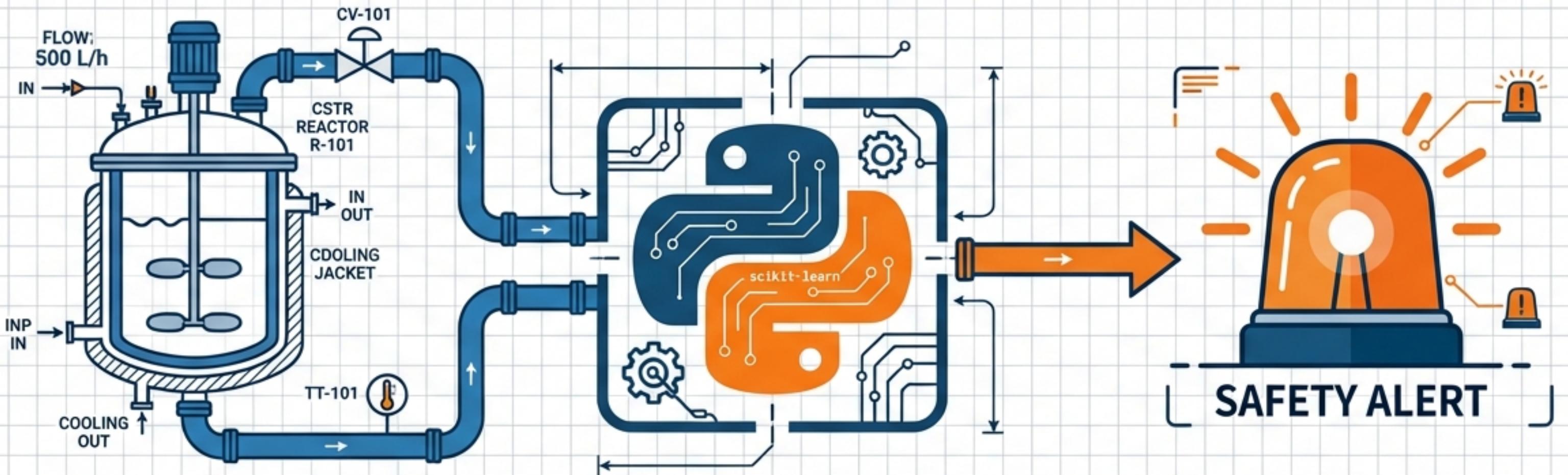
孤立森林 (Isolation Forest)

AI-Driven Anomaly Detection for Process Safety

(AI 驅動的製程安全異常檢測)



課程目標與應用藍圖 (Course Goals & Application Map)



原理理解 (Principles)

掌握孤立樹 (iTree) 的幾何
切分概念與路徑長度計算。

實作掌握 (Implementation)

使用 scikit-learn 建構並調校
模型參數。

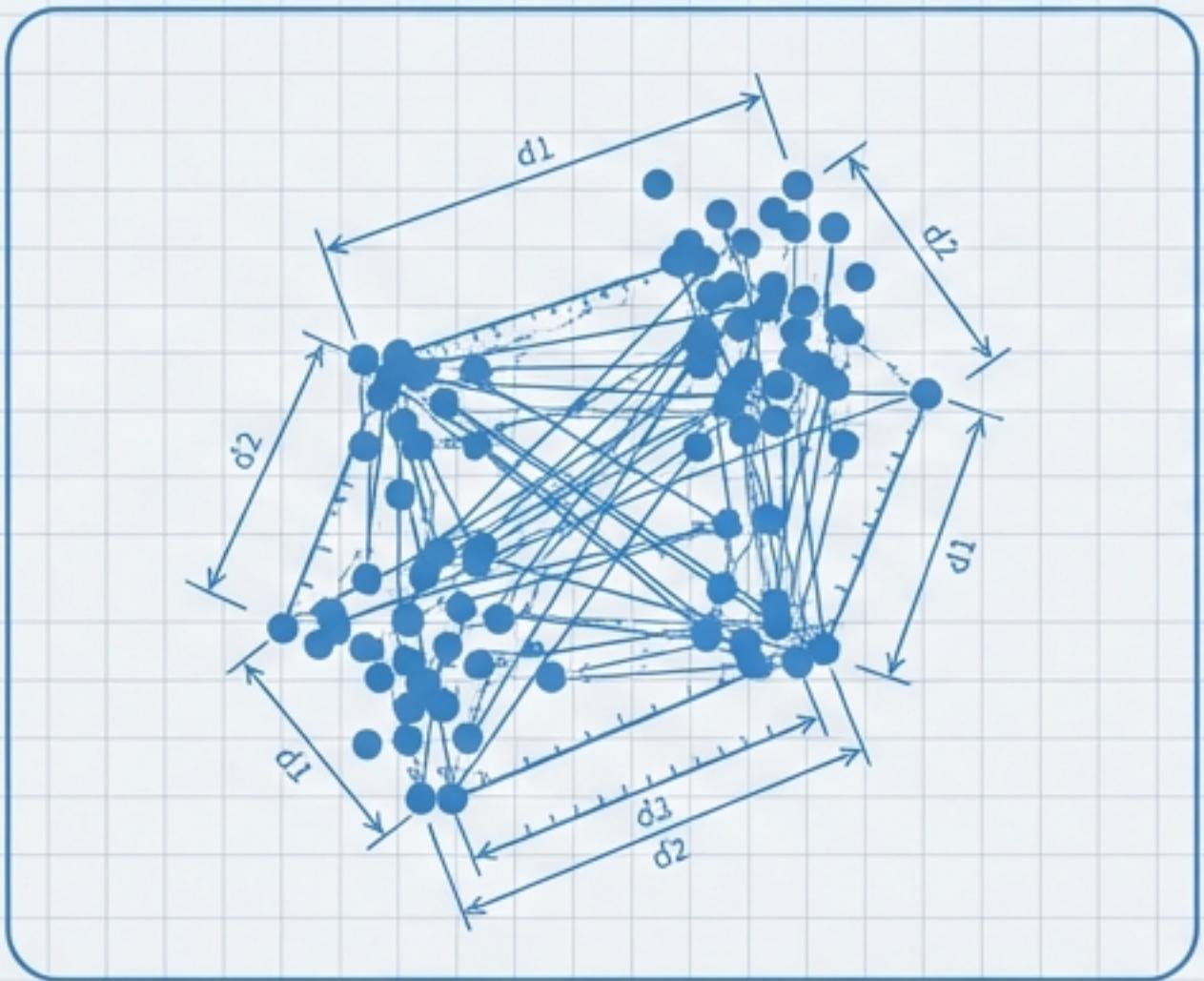
案例應用 (Case Study)

應用於 CSTR 反應器監控，
檢測溫度失控與攪拌故障。

為了在複雜化工廠中檢測稀有 (Rare) 且危險的事件，我們需要一種不依賴常態分佈假設的高效演算法。

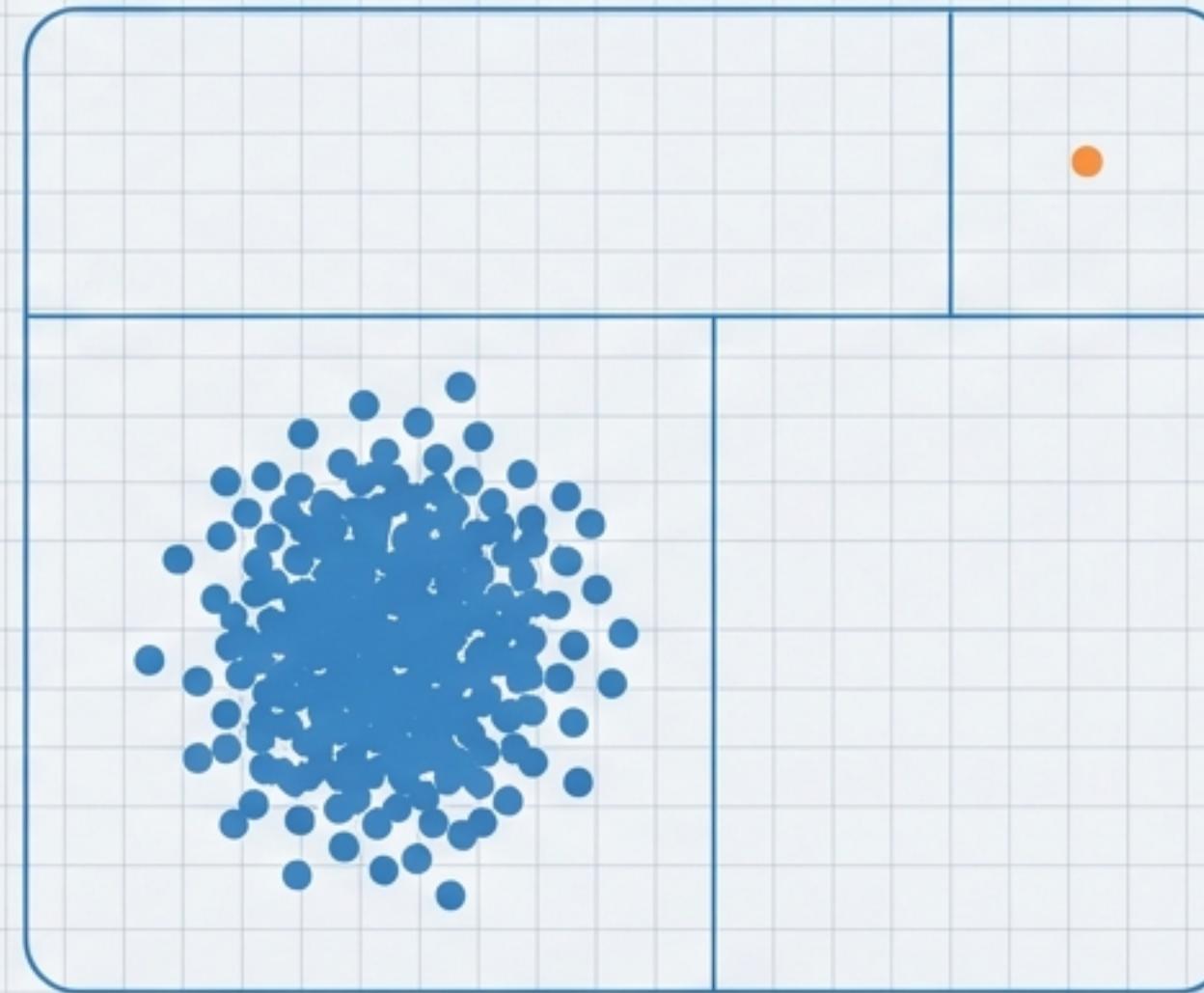
核心直覺：距離 vs. 孤立 (The Core Intuition: Isolation vs. Distance)

基於距離/密度 (Distance/Density)



計算昂貴 (Expensive)

基於孤立 (Isolation) - 孤立森林

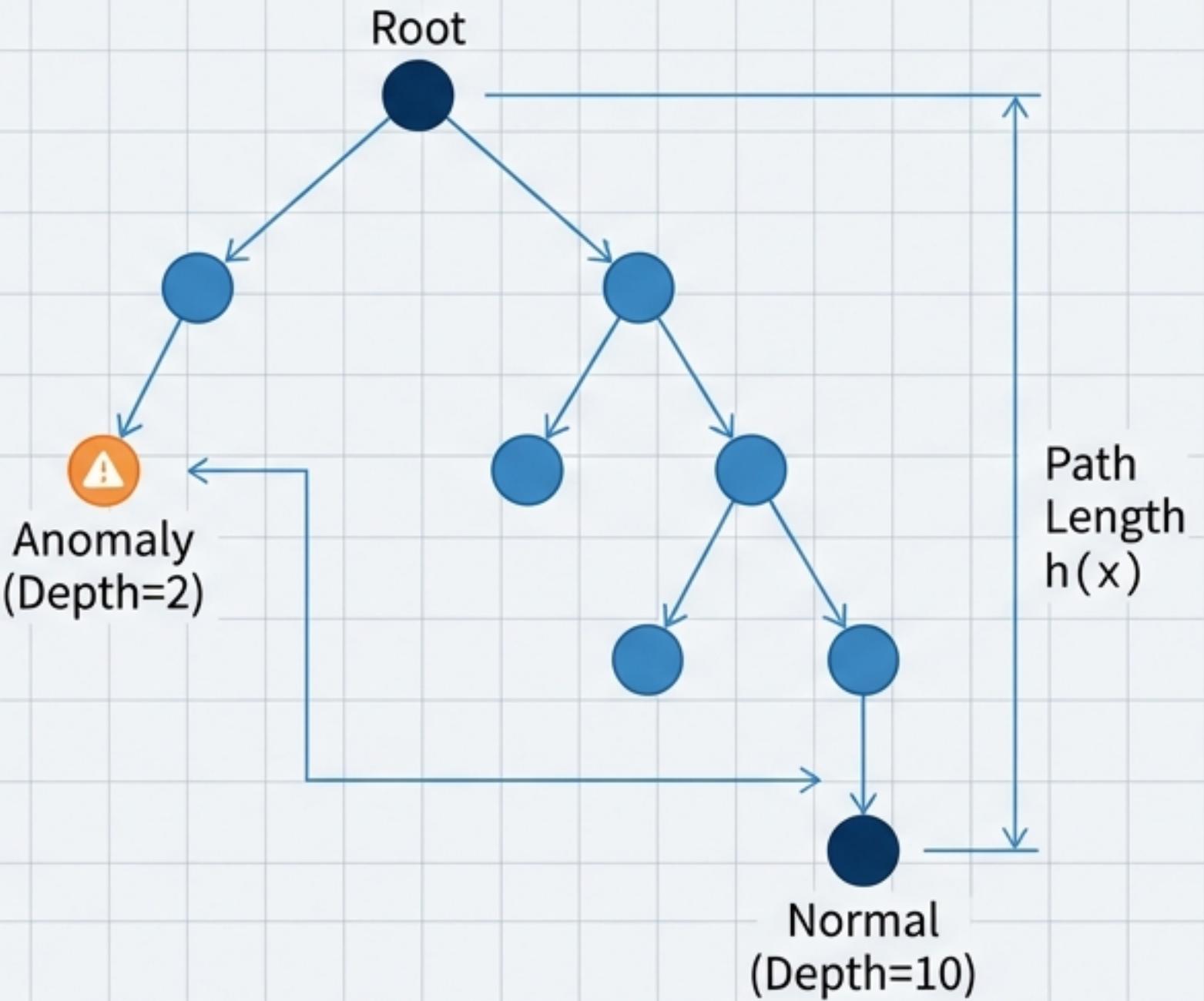


切分快速 (Fast)

異常的特質：異常點通常是「稀少 (Rare)」且「顯著不同 (Different)」。Short Path Length = Anomaly (路徑越短 = 越可能是異常)。

運作機制：孤立樹的建構 (Under the Hood: The Isolation Tree)

- 1. 隨機選擇特徵 (Random Feature) :**
從所有變數中隨機選取一個特徵 q 。
- 2. 隨機選擇分割點 (Random Split) :**
在該特徵的最大值與最小值之間，隨機選取分割點 p 。
- 3. 遞迴分割 (Recursive Partition) :**
如果數據點 $x < p$ 往左走，否則往右走。
- 4. 停止條件 (Stop) :**
節點只剩一個樣本，或達到樹的最大深度。



評分系統：從路徑長度到異常分數 (The Scoring System)

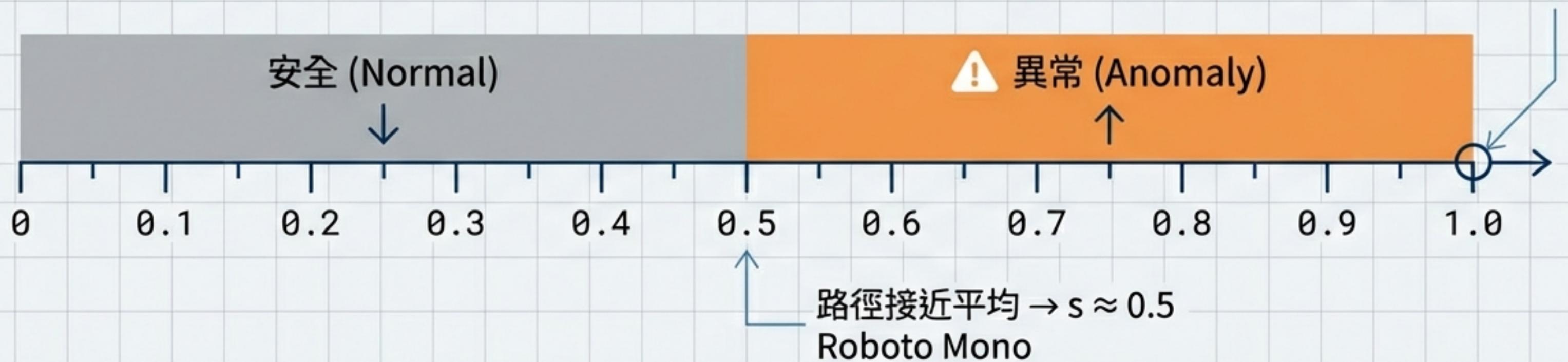
$$s(x, n) = 2^{-E[h(x)]/c(n)}$$

$h(x)$: 路徑長度 (Path Length) (Noto Sans TC)

$E[h(x)]$: 所有樹的平均路徑長度 (Roboto Mono)

$c(n)$: 歸一化常數 (類似平均路徑長度的期望值) (Noto

路徑極短 $\rightarrow s \approx 1$
Roboto Mono



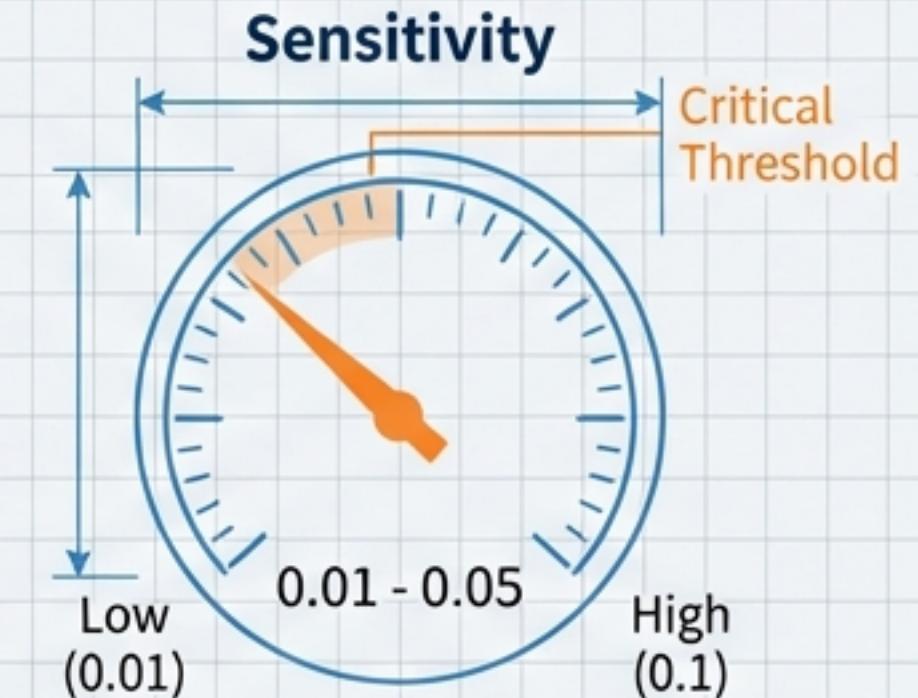
建構藍圖：Scikit-learn 實作 (Implementation)

```
from sklearn.ensemble import IsolationForest  
  
# 初始化模型  
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          # 確保可重現性  
)  
  
# 訓練與預測  
iso_forest.fit(X_train)  
y_pred = iso_forest.predict(X_test)
```

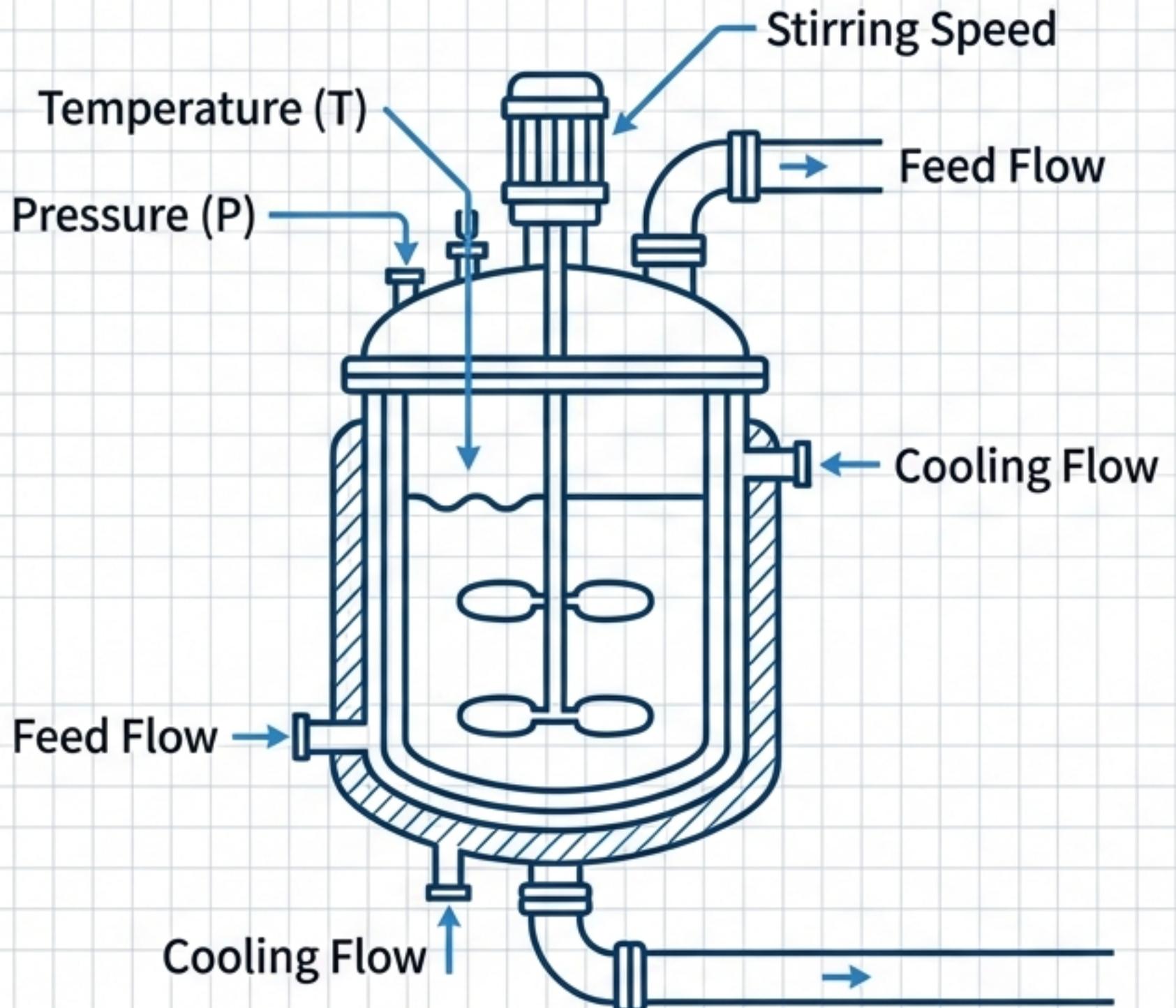
Critical Threshold Knob

關鍵參數調校：設定機器的靈敏度 (Critical Parameters)

Parameter	Description	Recommendation
contamination (異常比例) Roboto Mono/ Noto Sans TC Bold	最關鍵的參數。決定了 決策閾值 (Threshold)。 Noto Sans TC Regular	若未知，建議從 0.01 - 0.05 開始保守測試。 Noto Sans TC Regular
n_estimators (樹的數量) Roboto Mono/ Noto Sans TC	決定模型的穩定性。	100 (預設) 通常足夠。 大數據集可增至 300。 Noto Sans TC Regular
max_samples (取樣大小) Roboto Mono/ Noto Sans TC	控制每棵樹看到的數 據量。	維持 'auto' (256) 可 保證良好的隨機性。 Noto Sans TC Regular



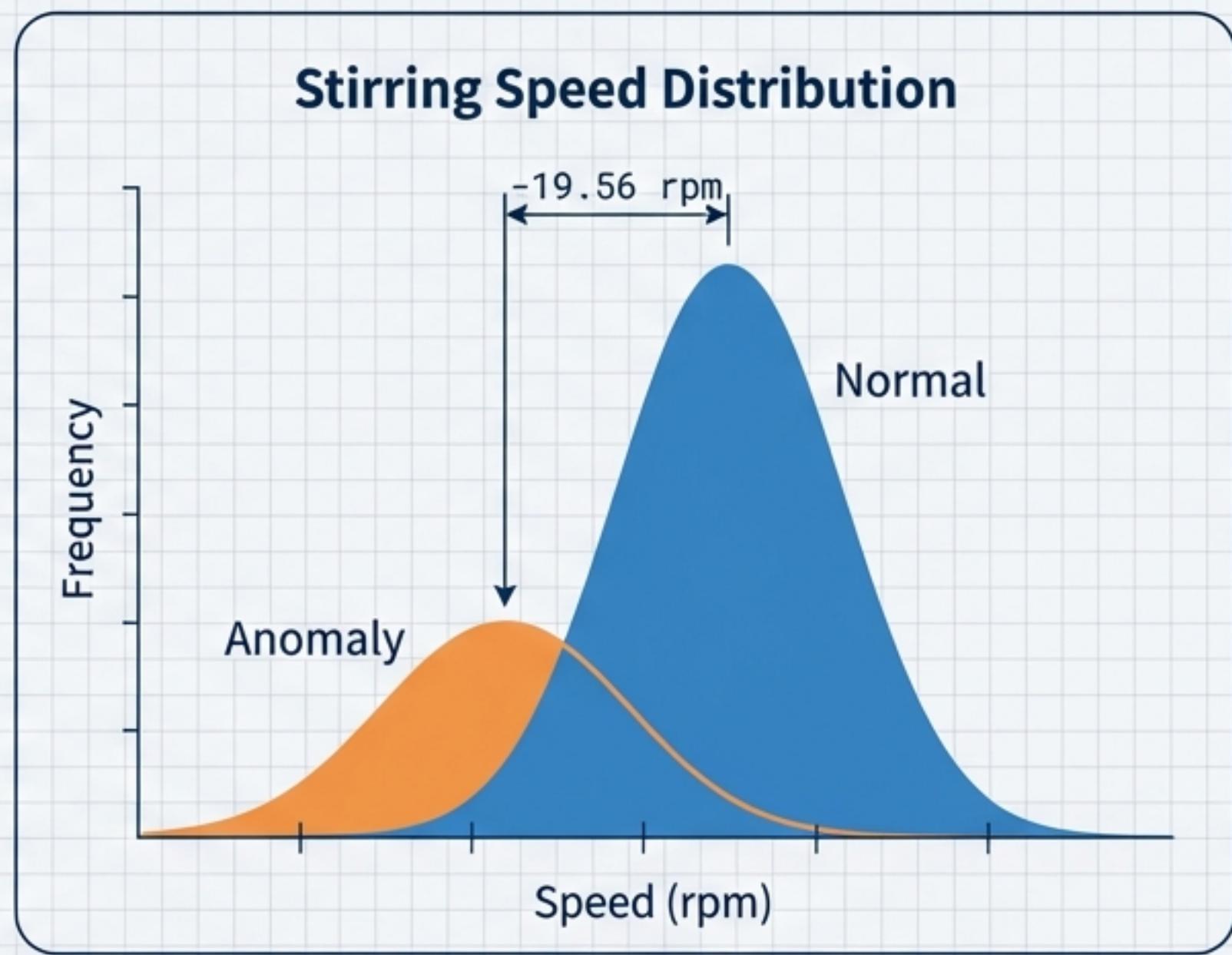
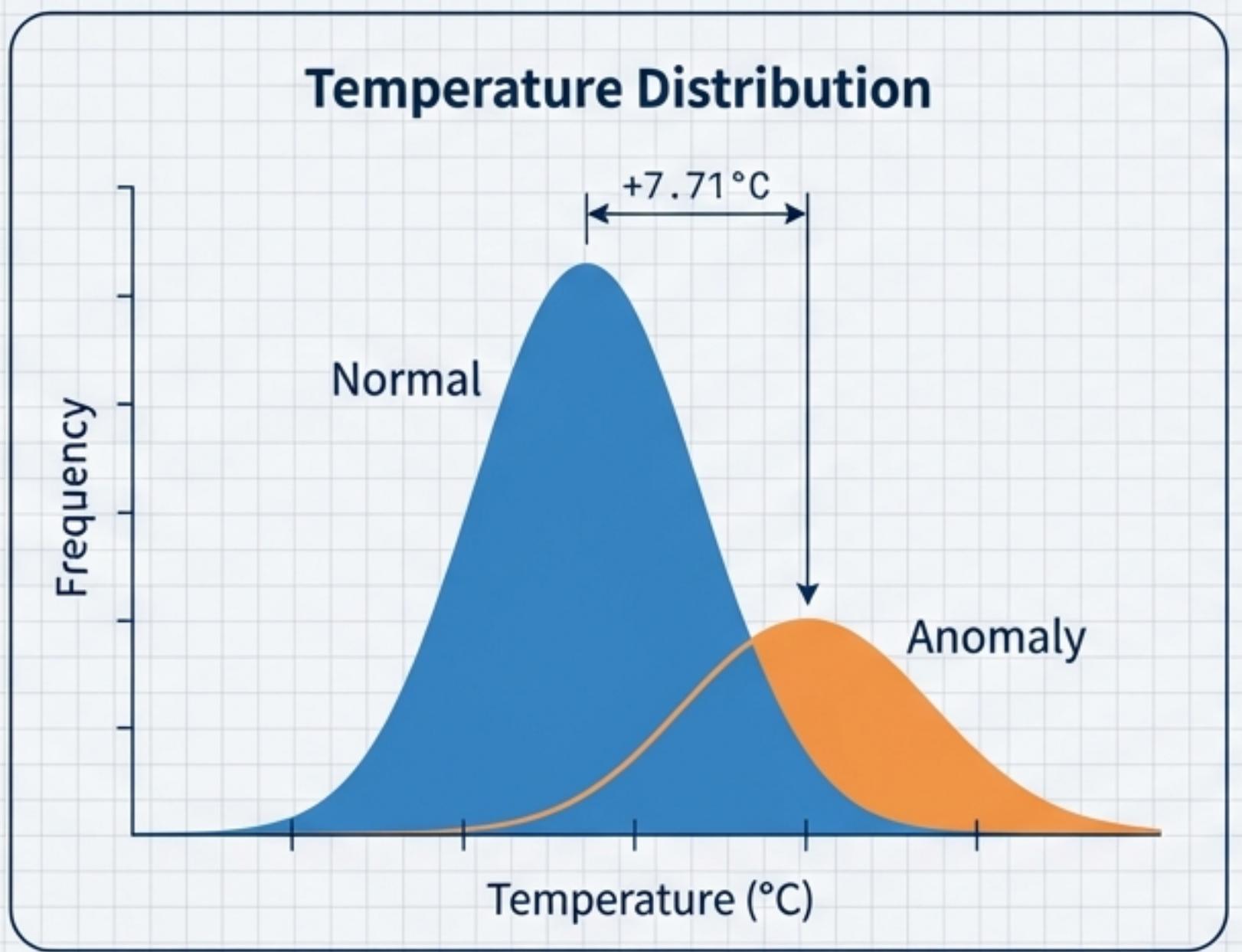
案例研究：CSTR 反應器監控 (Case Study: CSTR Monitoring)



The Challenge

- **多重共線性 (High Correlation)** : 變數之間高度相關。
- **稀有異常 (Rare Anomalies)** : 異常數據僅佔總體的 5% (0.05)。
- **目標** : 檢測反應失控 (Runaway) 或設備故障 (如攪拌器失效)。

數據透視：可視化「軟異常」(Visualizing ‘Soft’ Anomalies)



關鍵發現: 單一變數難以區分 (Overlap)，必須依賴多變數的聯合分析。

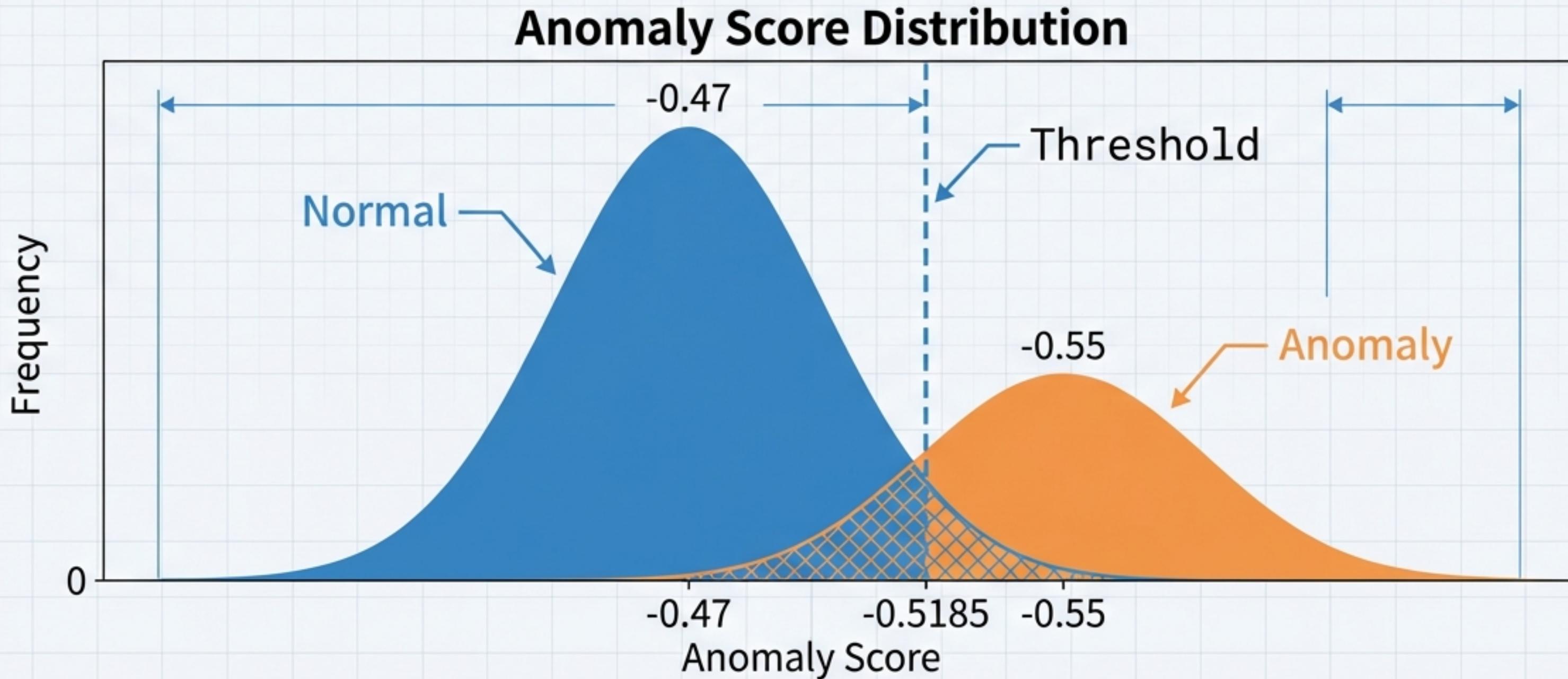
訓練結果與效能測試 (Training & Detection Results)

Digital Readout

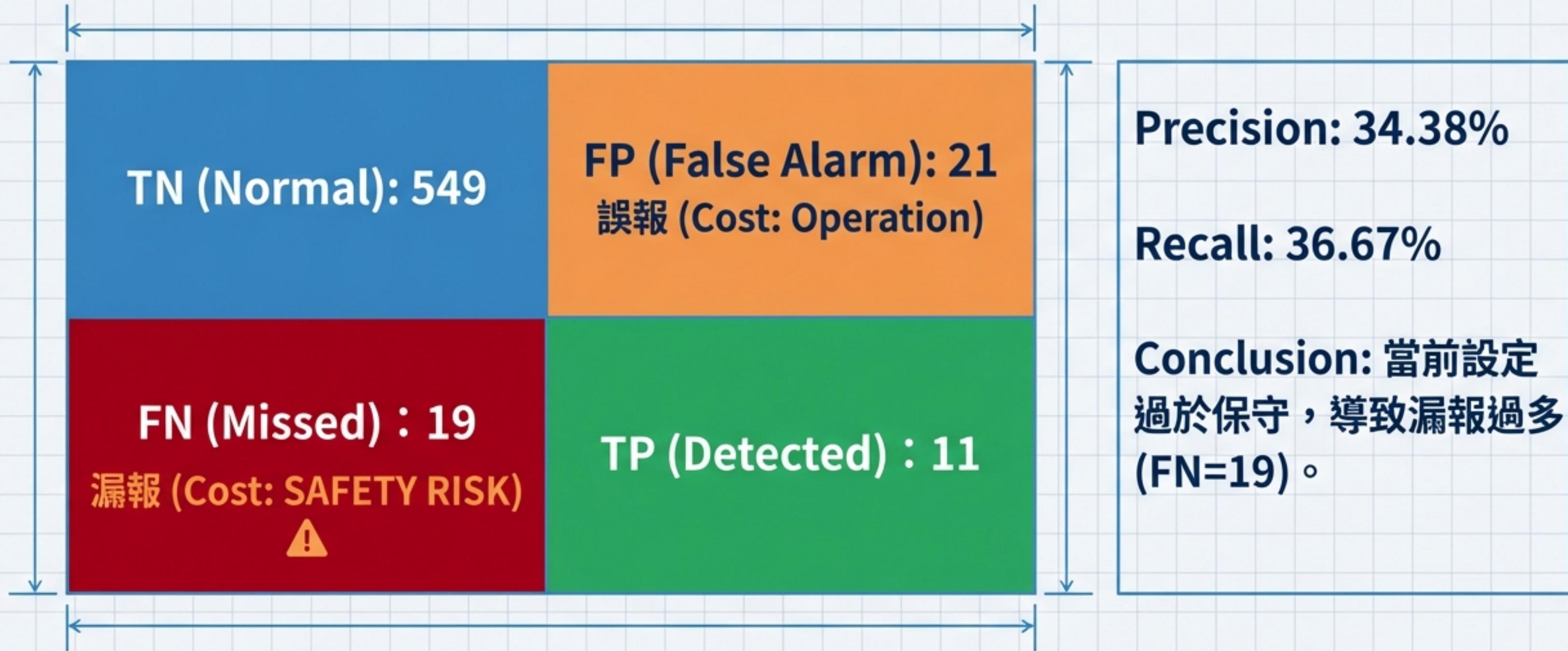
Training Time: 0.2060s

Digital Readout

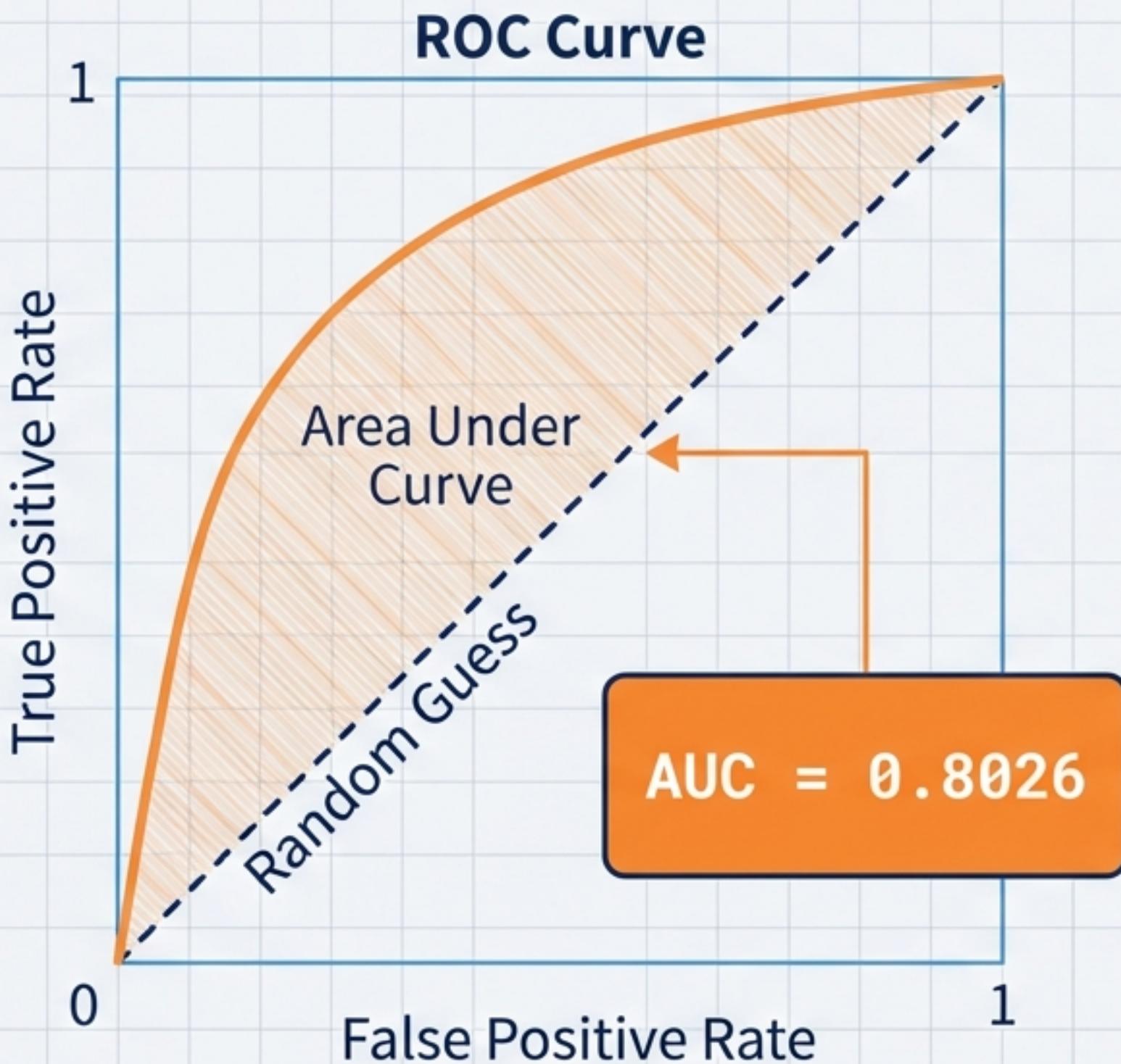
Threshold Offset: -0.5185



評估一：混淆矩陣 (The Confusion Matrix)

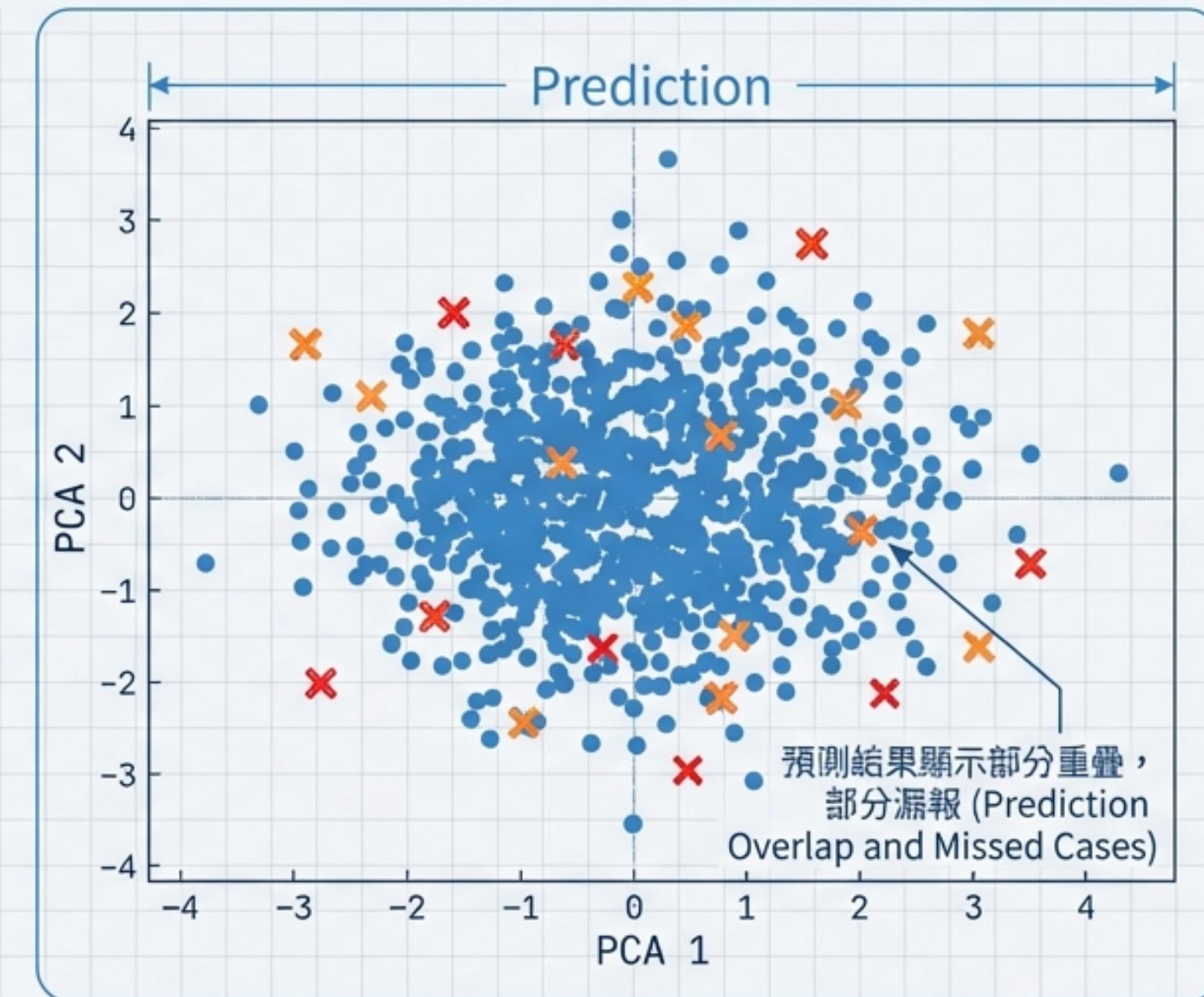
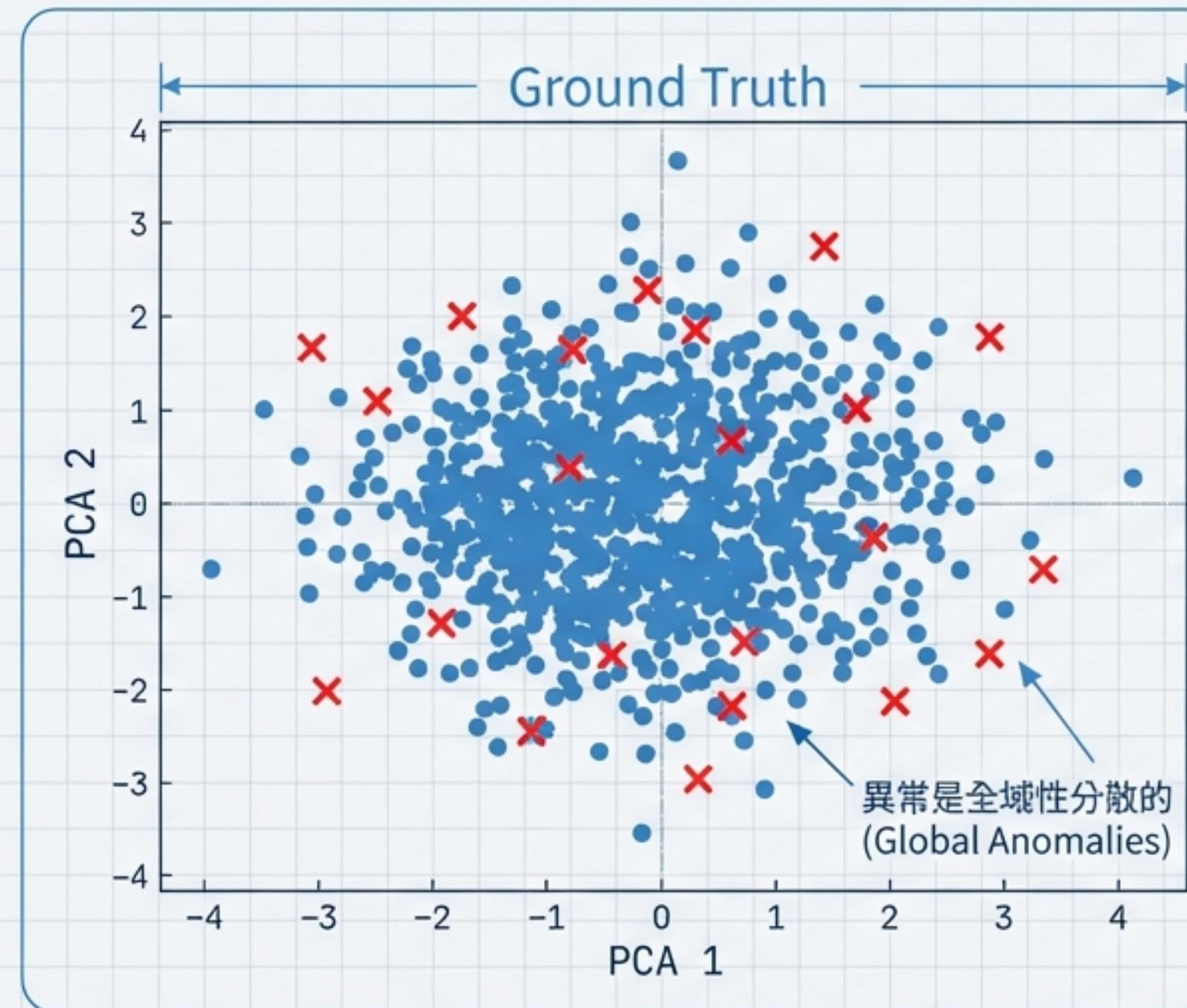


評估二：ROC 曲線與 AUC (Performance Metrics)



- 模型效能評級：良好 (Good)。
- AUC 顯示模型確實學到了異常異常模式 (遠優於隨機猜測 0.5)。
- 證明孤立森林能有效處理變數間的非線性關係。

結果可視化：PCA 投影 (Visualizing the Result)



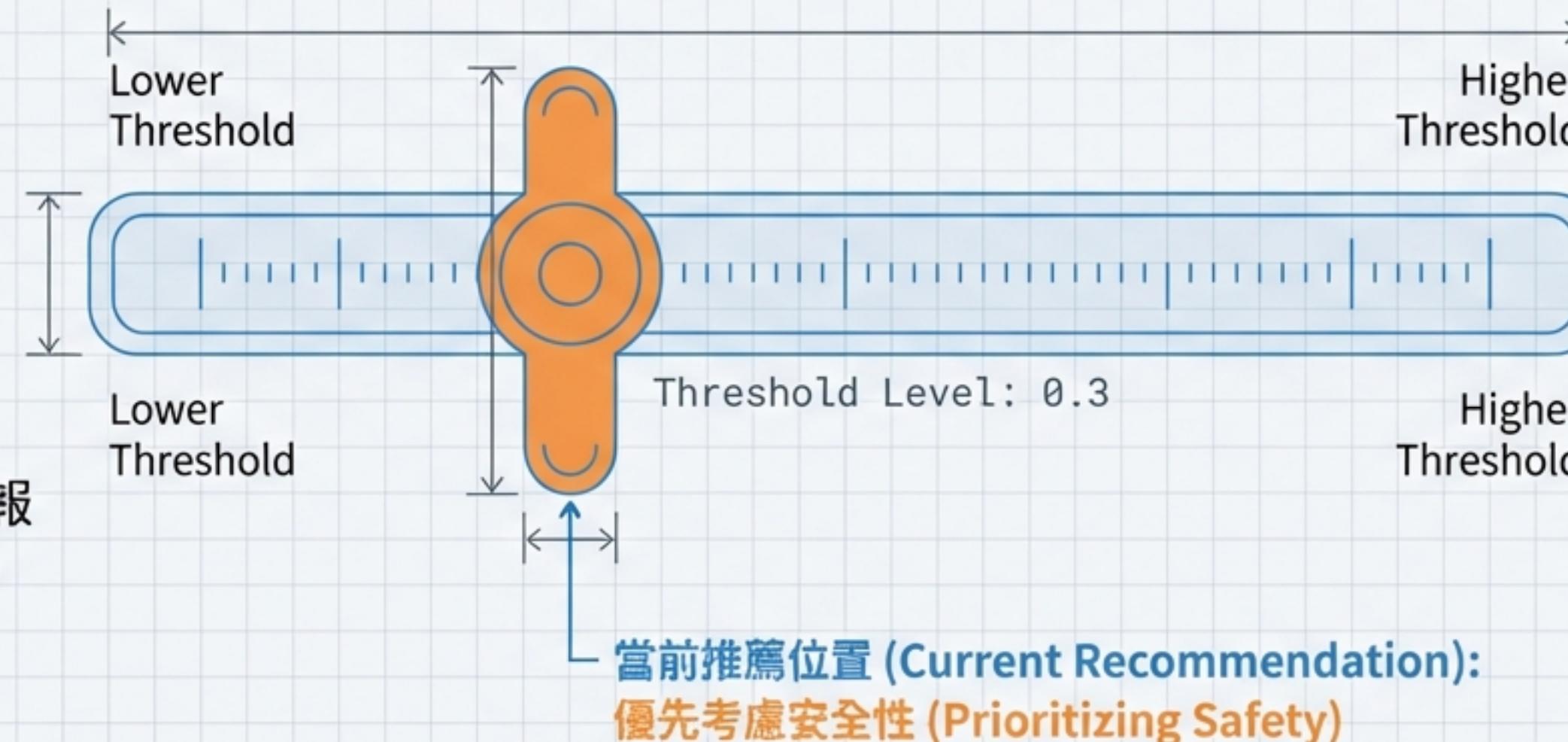
孤立森林優勢: 正是因為異常點分散且孤立，隨機切分策略才能生效。

工程權衡：閾值調整策略 (Engineering Trade-offs)



**High Recall
(Safety)**

製程安全:
寧可誤報，不可漏報
(降低 Threshold)



**High Precision
(Efficiency)**

維護規劃:
避免浪費人力
(提高 Threshold)

總結與工具選擇 (Summary & Selection Guide)

Checklist: When to use iForest?

- **Big Data:** > 10,000 samples
- **High Dimensions:** > 50 vars
- **Global Anomalies:** Scattered, not clustered
- **Label-free:** Unsupervised

Comparison

vs. LOF

LOF 適合局部異常，但計算慢。

vs. One-Class SVM

SVM 適合小數據，但對高維吃力。

Next Step: Unit 08 (LOF) - Local Outlier Factor