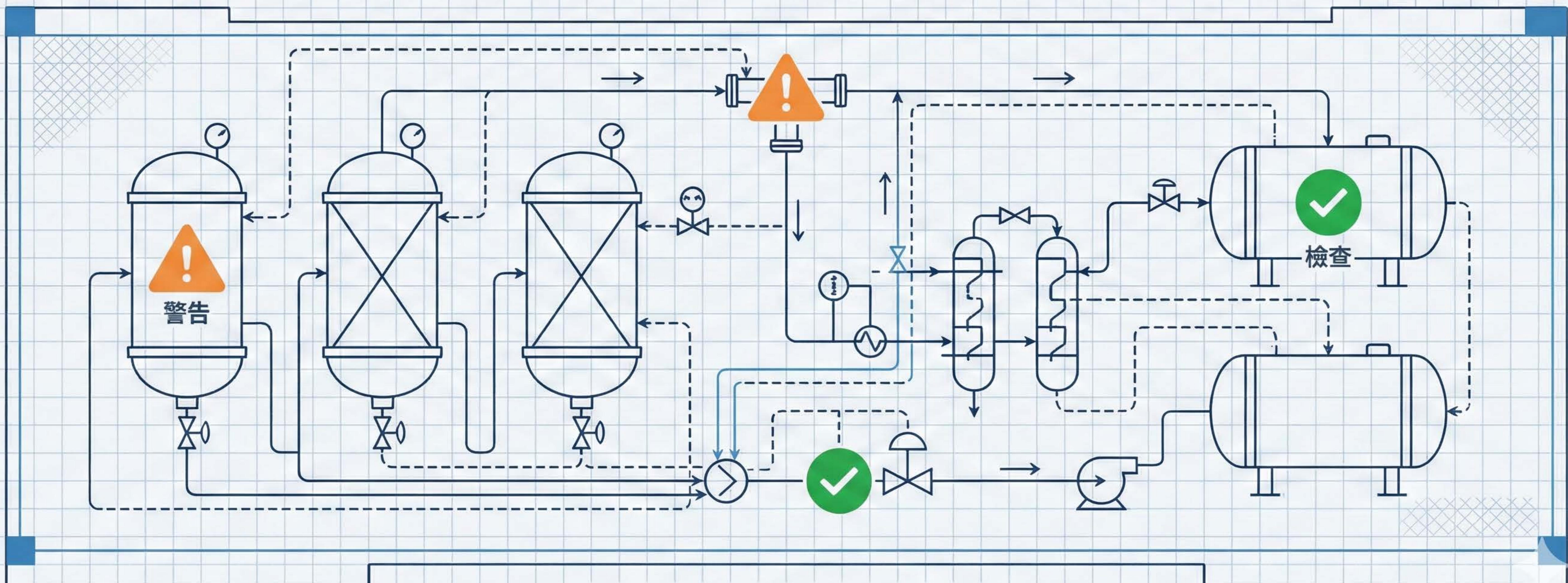
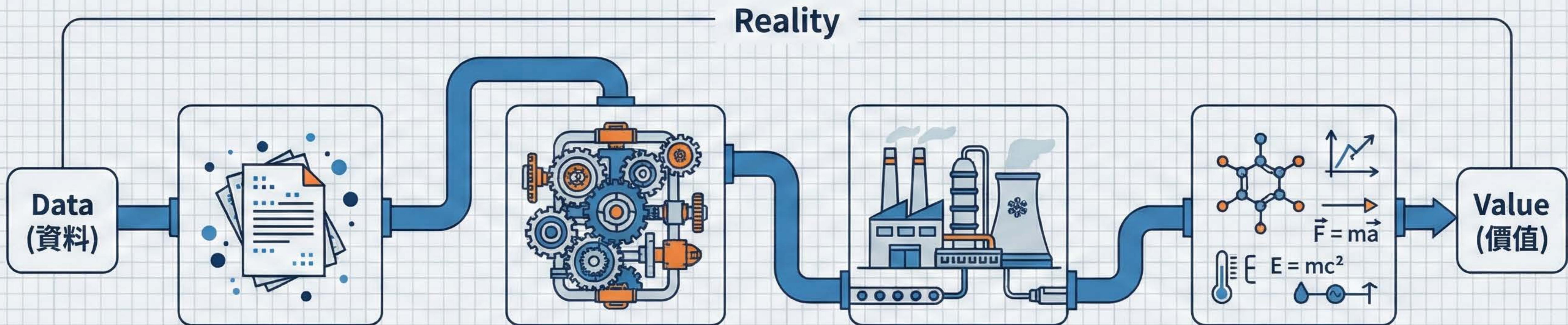
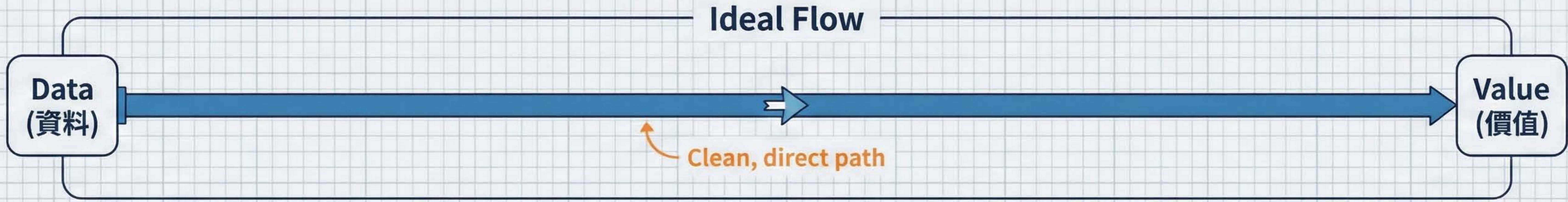


機器學習工作流程中的挑戰與解決方案

從理論到現場：AI 引擎的工程除錯指南



理想很豐滿，現實很骨感：ML 工程的四大戰場



1. 資料工程
(Data Engineering)
原料問題：雜訊、
缺漏、不平衡

2. 模型工程
(Model Engineering)
反應器調控：過擬合
黑箱、參數

3. 部署維運
(Deployment & Ops)
工廠放大：概念漂
移、算力

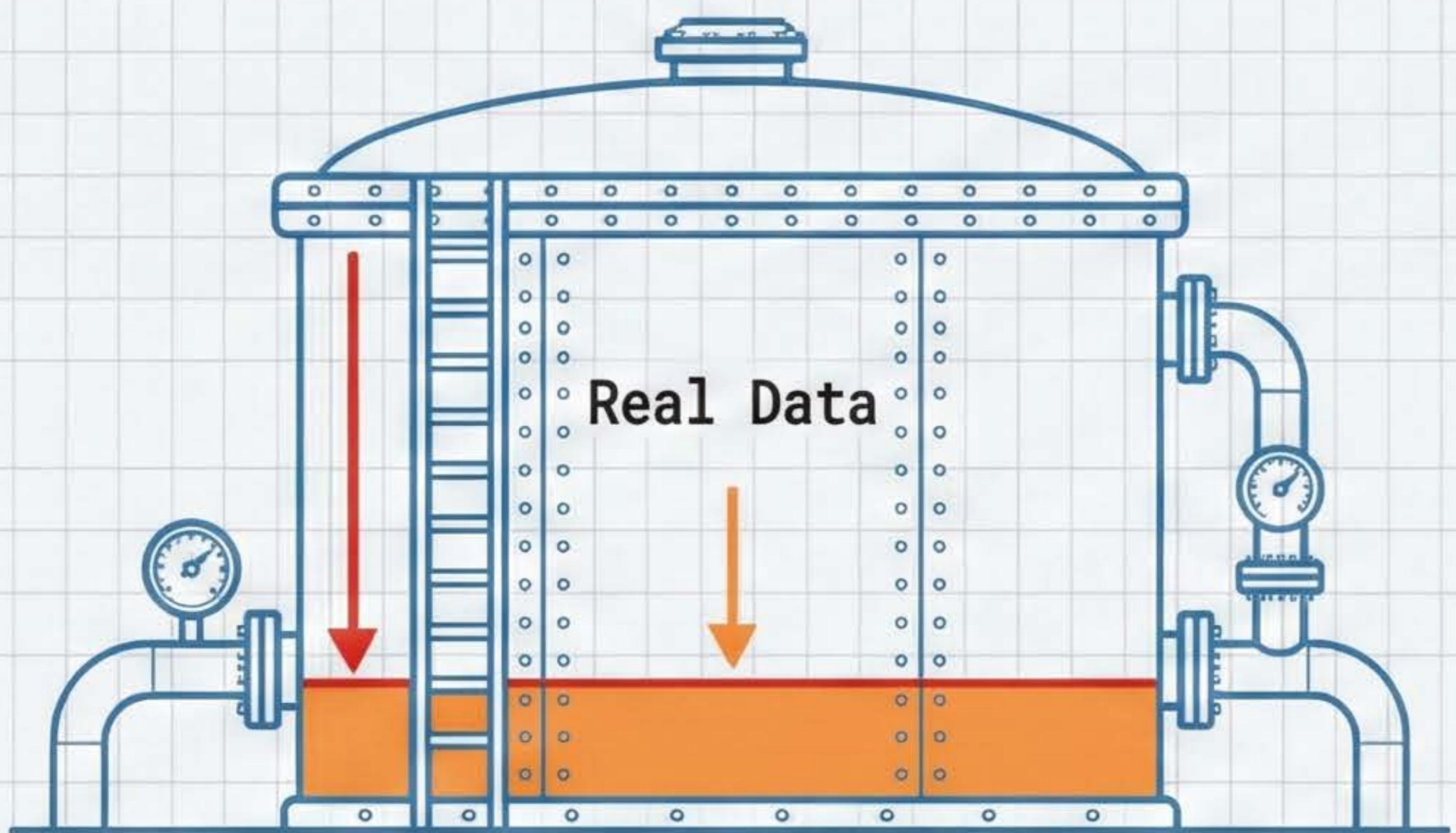
4. 化工特點
(Domain Specifics)
物理限制：守恆定律
批次變異

挑戰 1：資料量不足 (Data Scarcity) — 餵飽你的 AI 引擎

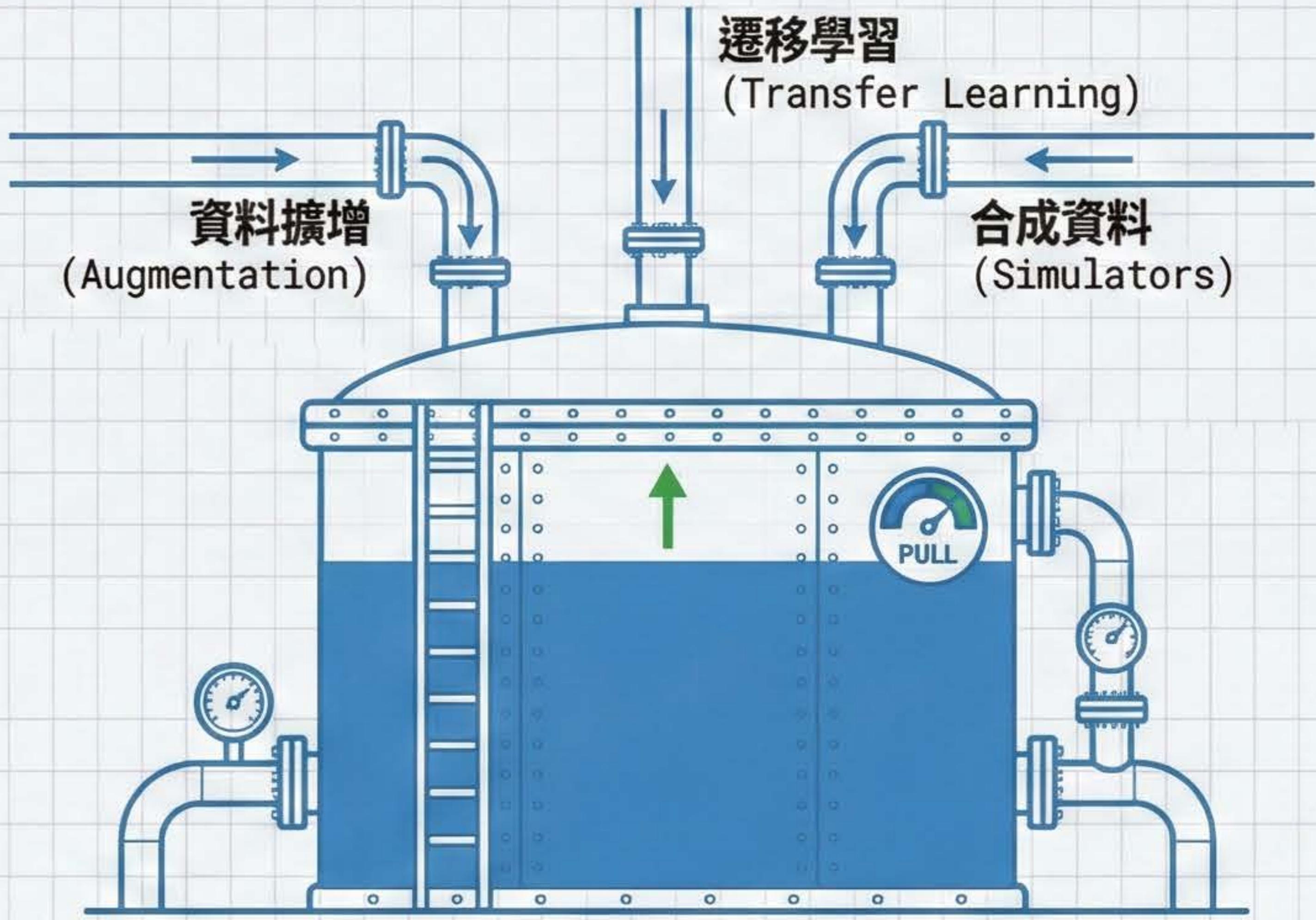
問題 (Problem)

症狀 (Symptom):

- 化工資料標記成本高昂且耗時 (High cost/time)。
- 機器學習需要足夠樣本才能捕捉有效模式。

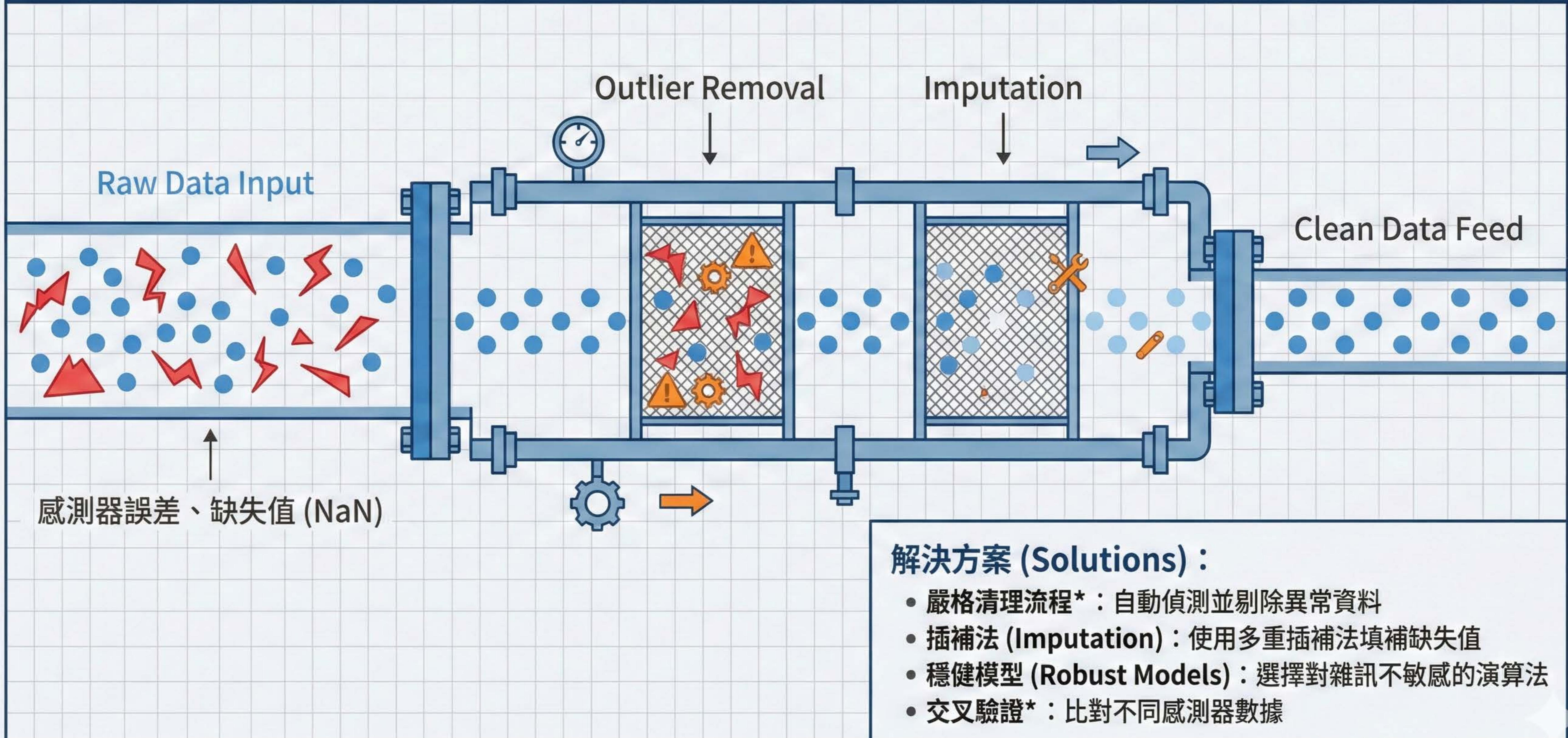


解決方案 (Solutions)



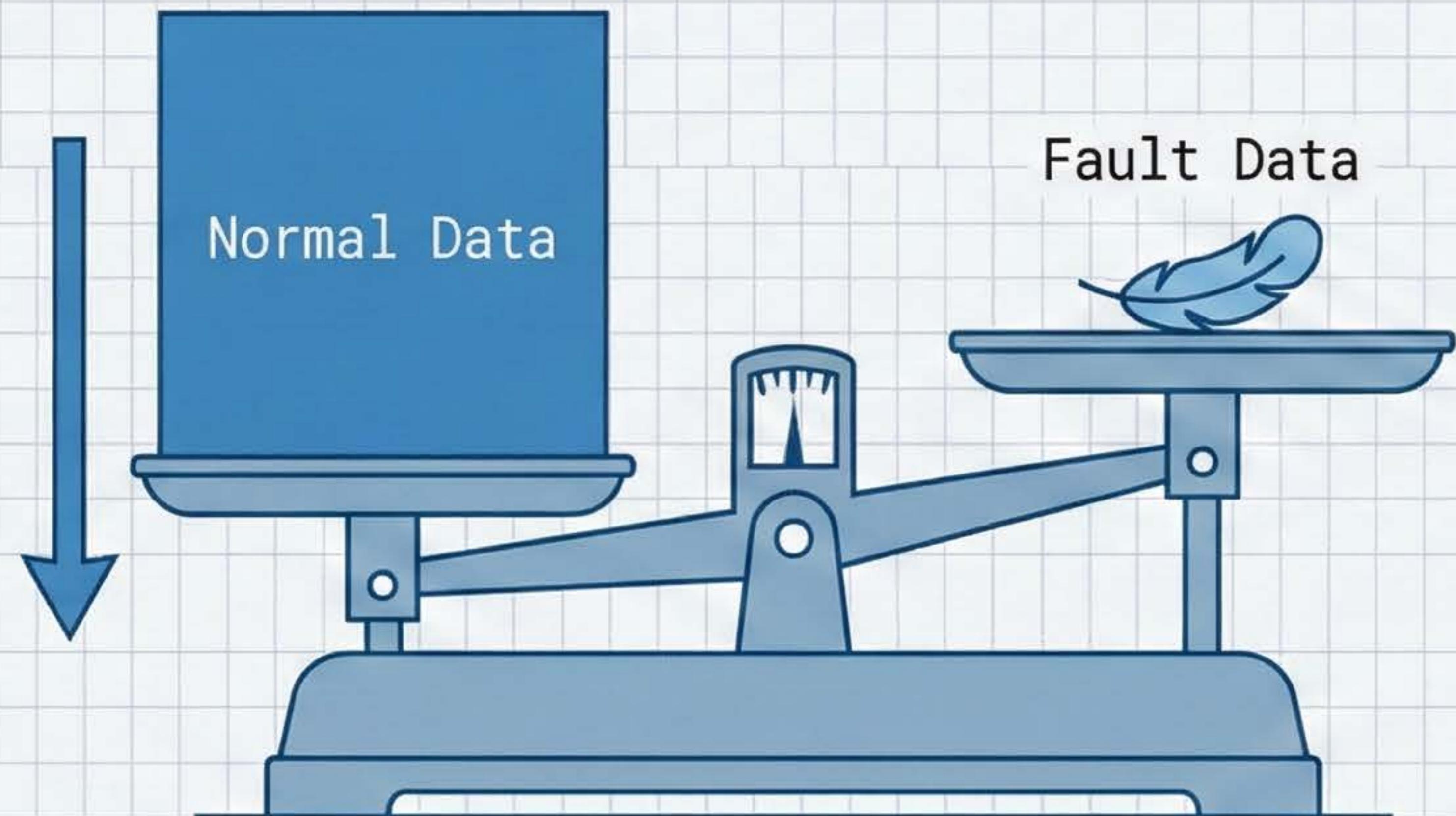
- 資料擴增 (Data Augmentation)：透過轉換生成更多訓練樣本
- 遷移學習 (Transfer Learning)：利用相似製程的預訓練模型
- 合成資料 (Synthetic Data)：使用製程模擬器生成虛擬訓練資料
- 主動學習 (Active Learning)：智慧選擇最有資訊量的樣本進行標記

挑戰 2：資料品質 (Data Quality) – 垃圾進，垃圾出



挑戰 3 & 4：不平衡與維度詛咒

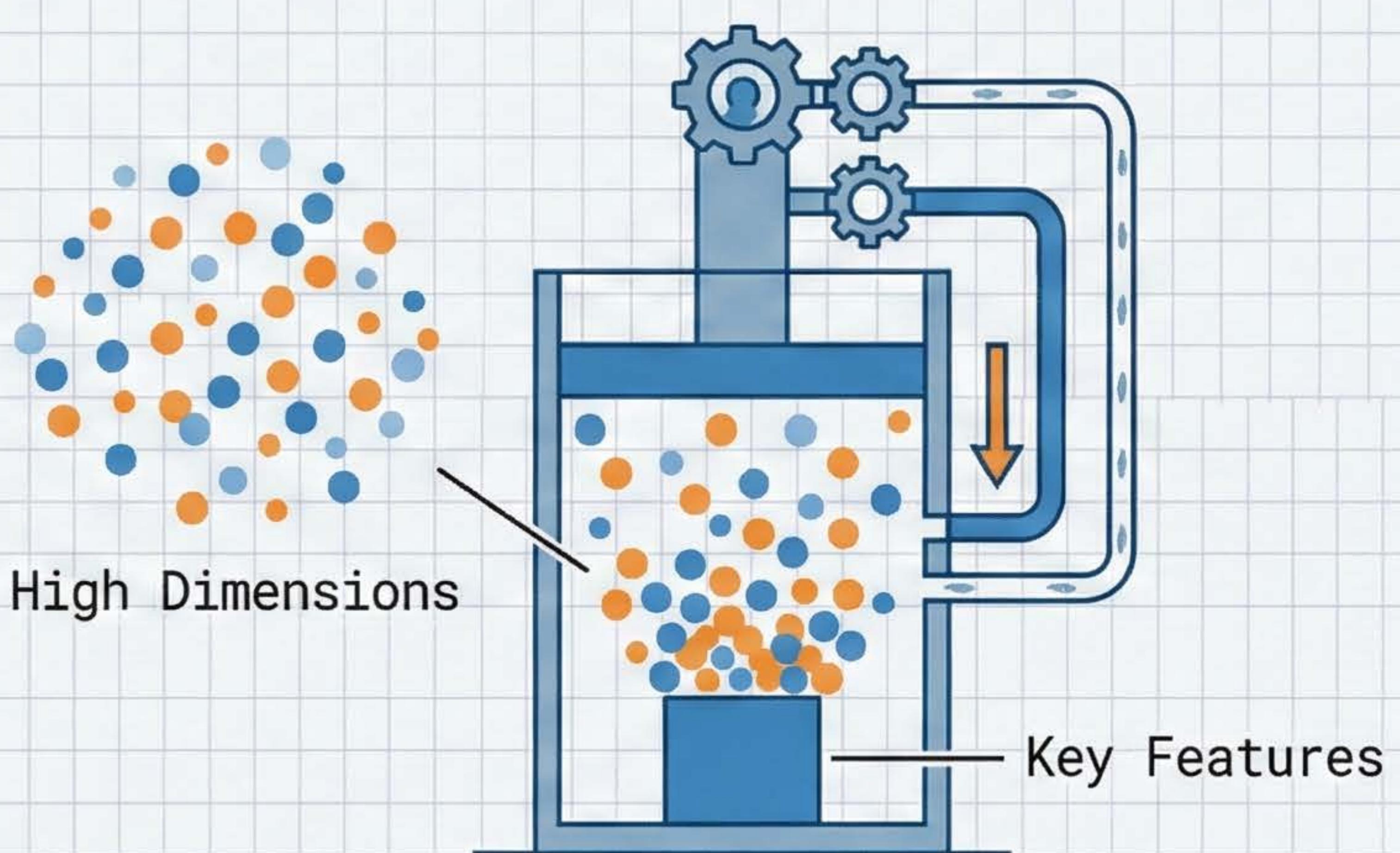
不平衡 (Imbalance)



正常操作數據 >> 異常數據，模型忽略少數類別。

解法：**重新採樣 (Resampling) (SMOTE, 過/欠採樣) & 調整權重**

維度詛咒 (Dimensionality)

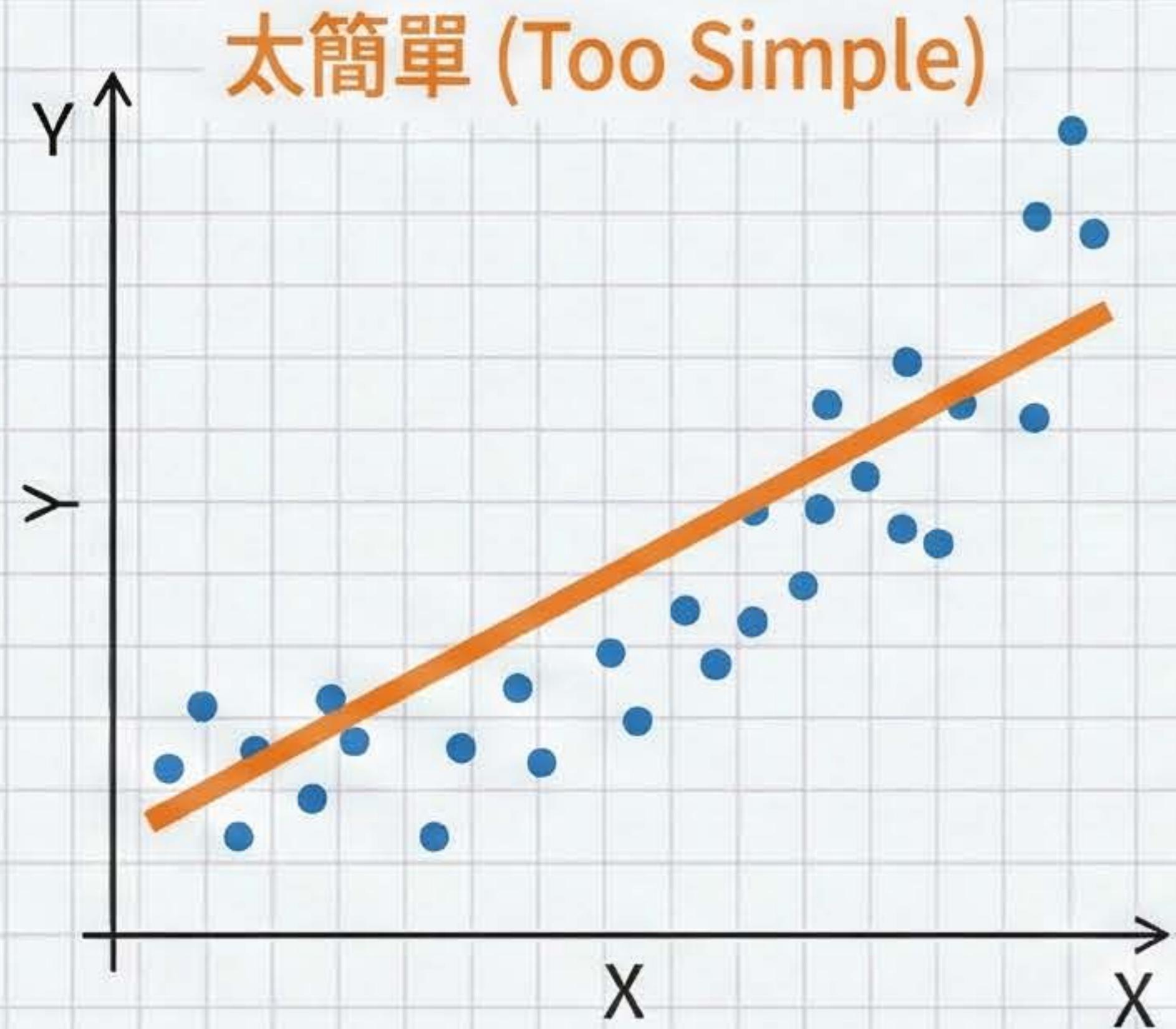


變數過多導致過擬合與計算量大增。

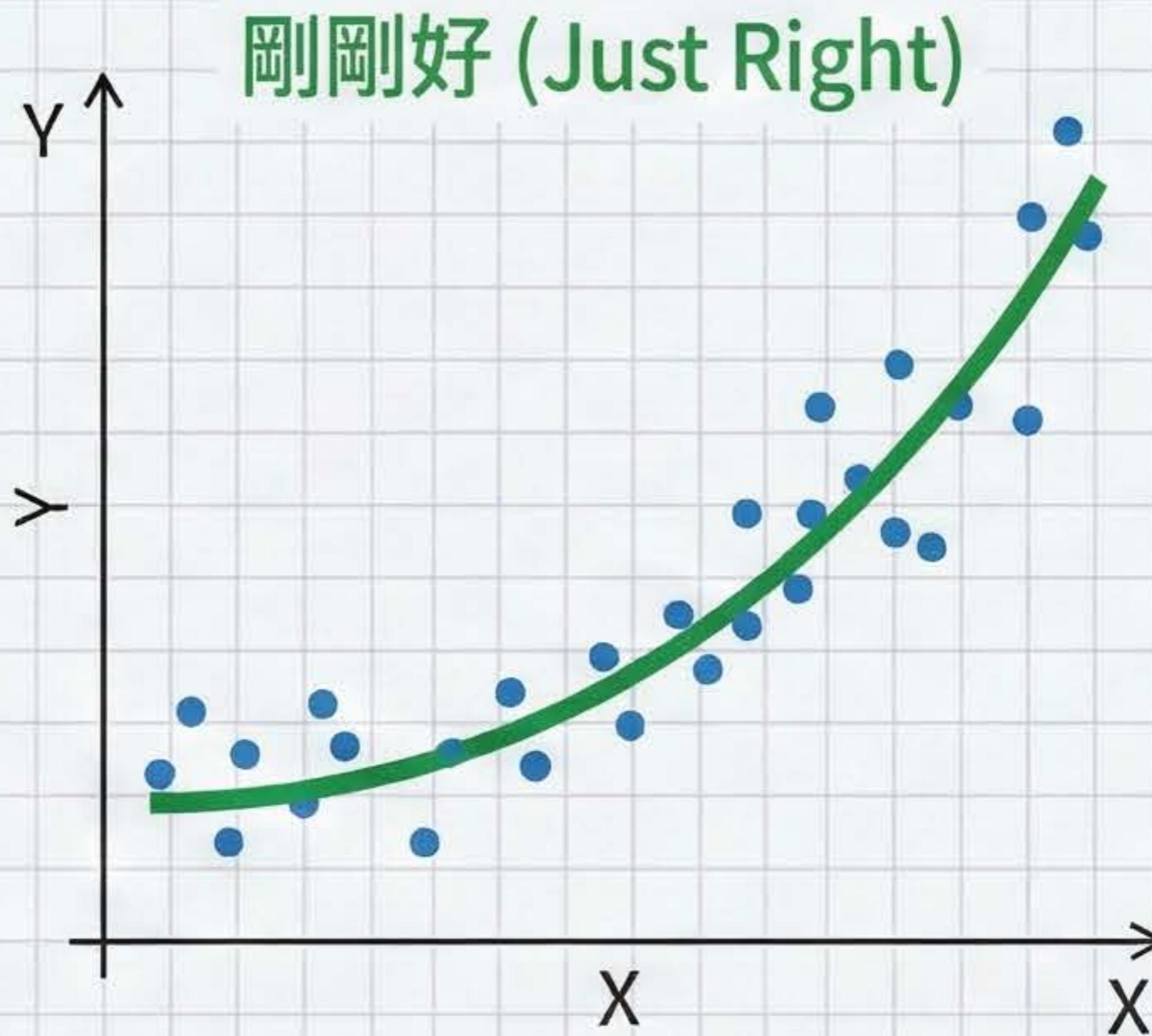
解法：**特徵選擇 & 降維技術 (PCA, t-SNE)**

挑戰 5：穩定性 (Stability) — 過擬合 vs. 欠擬合

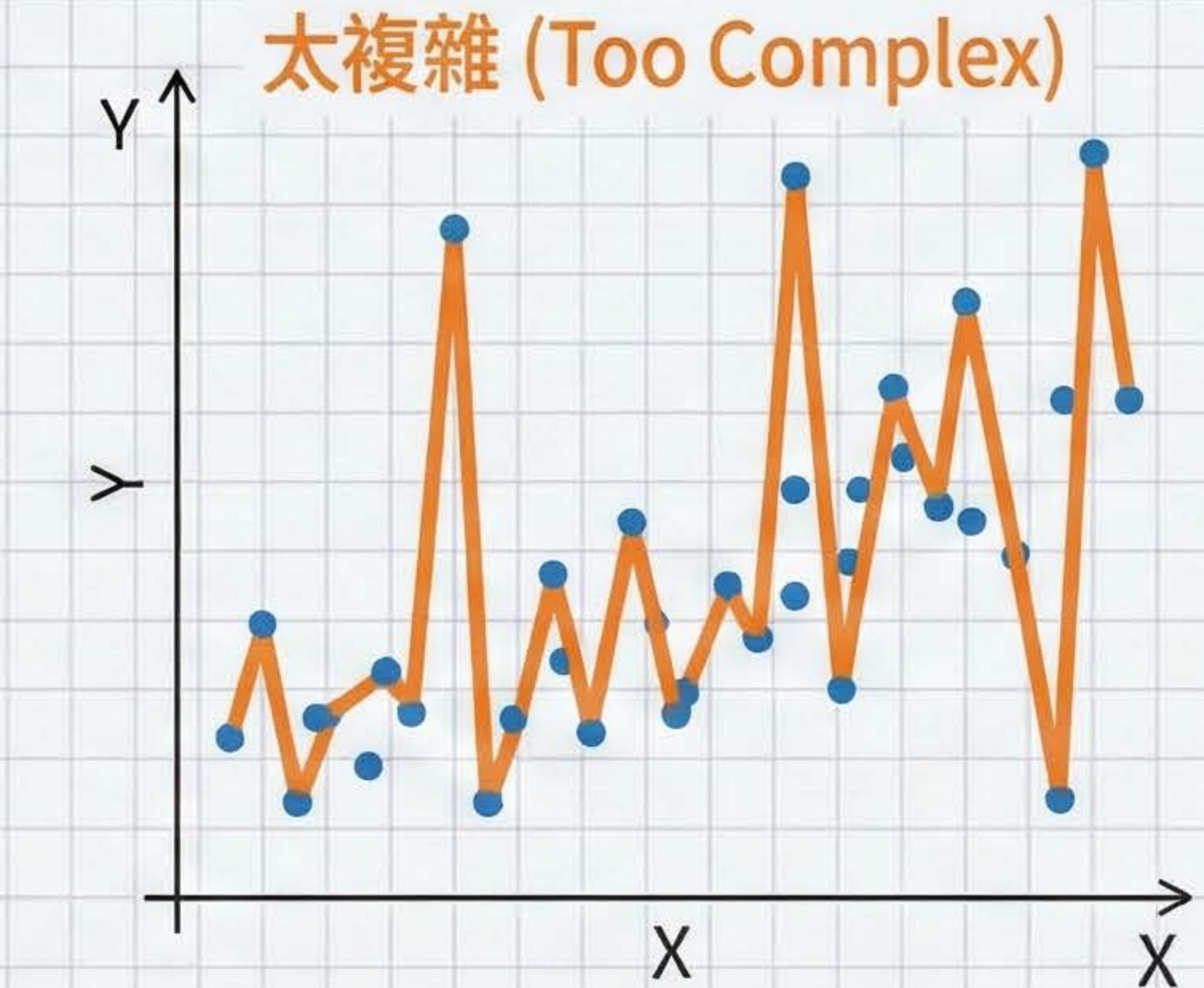
欠擬合 (Underfitting)



適配 (Robust)



過擬合 (Overfitting)



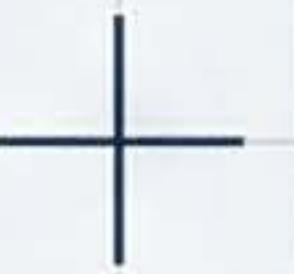
解法：增加特徵、提高複雜度

目標：學習規律而非記憶

解法：正則化、早停、更多數據

****關鍵工具**：**驗證集 (Validation Set) 與交叉驗證 (Cross-Validation) 是檢測此問題的標準工具。

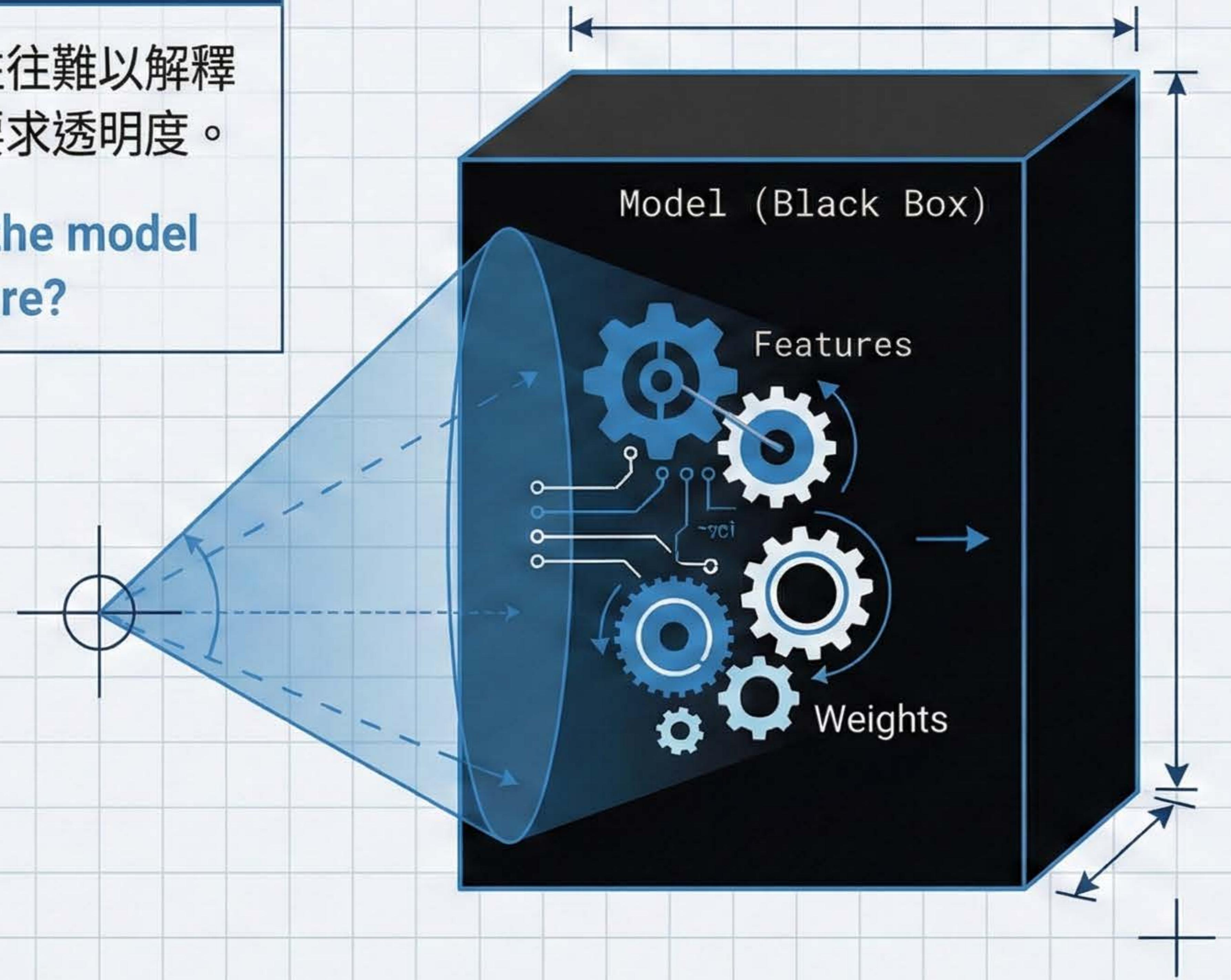
挑戰 6：可解釋性 (Interpretability) — 打開黑盒子



挑戰：

深度學習模型往往難以解釋，但化工安全要求透明度。

Q: Why did the model predict failure?

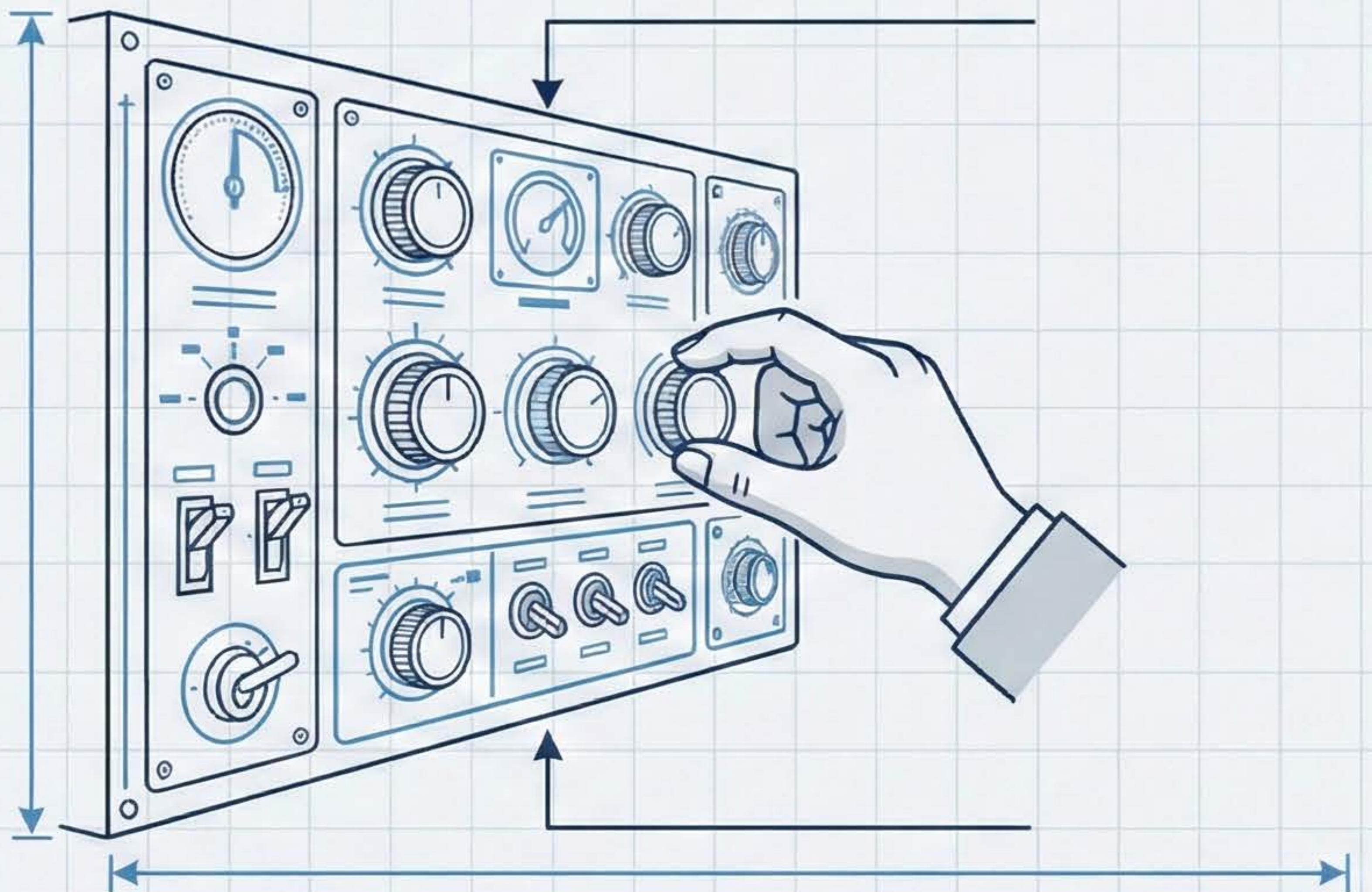


解決方案：

- **可解釋模型**：高風險場景優先使用線性回歸或決策樹。
- **事後解釋工具 (Post-hoc)**：
 - **SHAP**：量化每個特徵對預測的貢獻
 - **LIME**：局部解釋決策邊界
- **領域知識驗證**：確認邏輯符合物理化學直覺

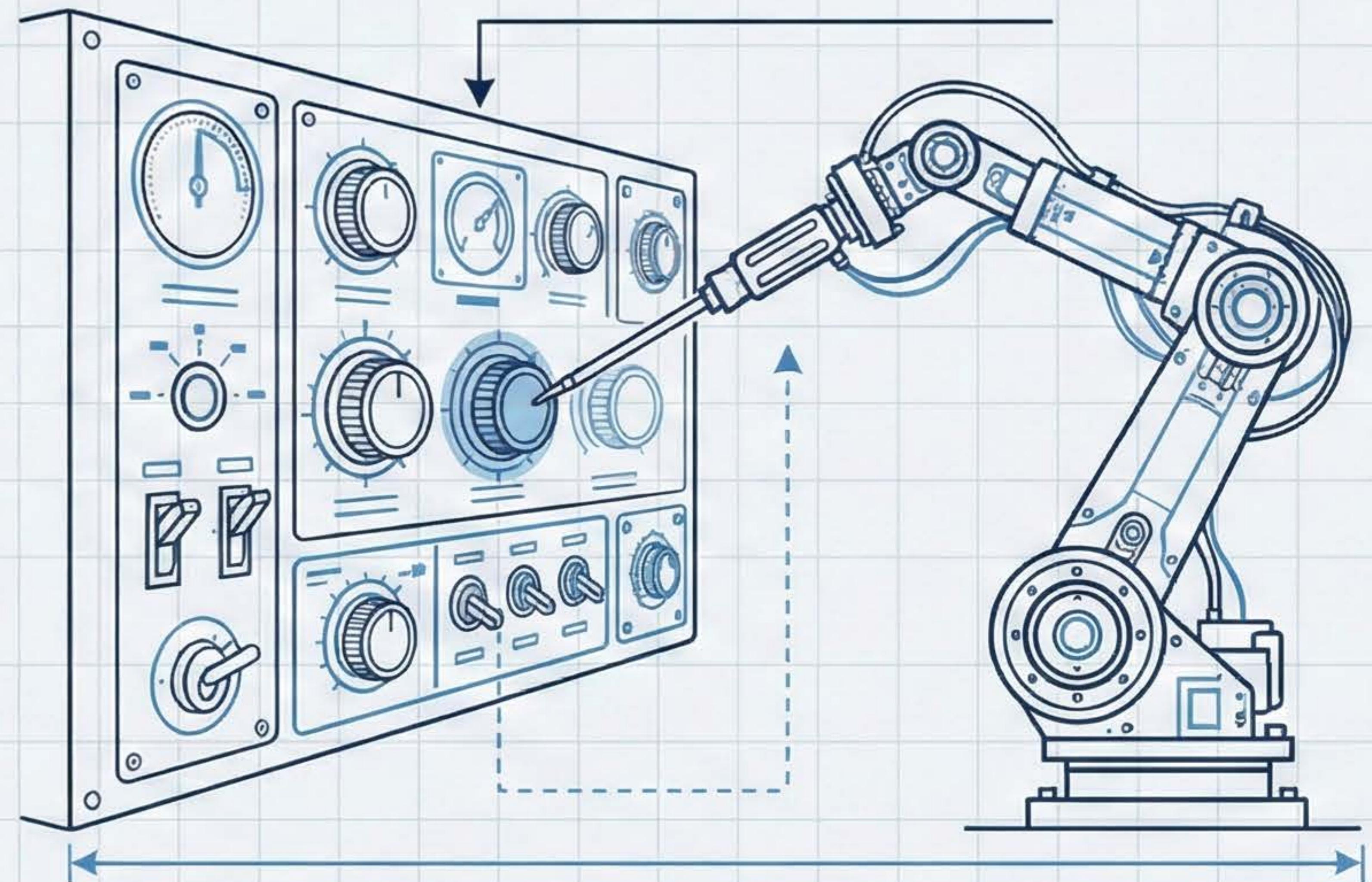
挑戰 7：超參數調整 (Hyperparameter Tuning)

網格搜尋 (Grid Search)



暴力窮舉，效率低，耗時長。

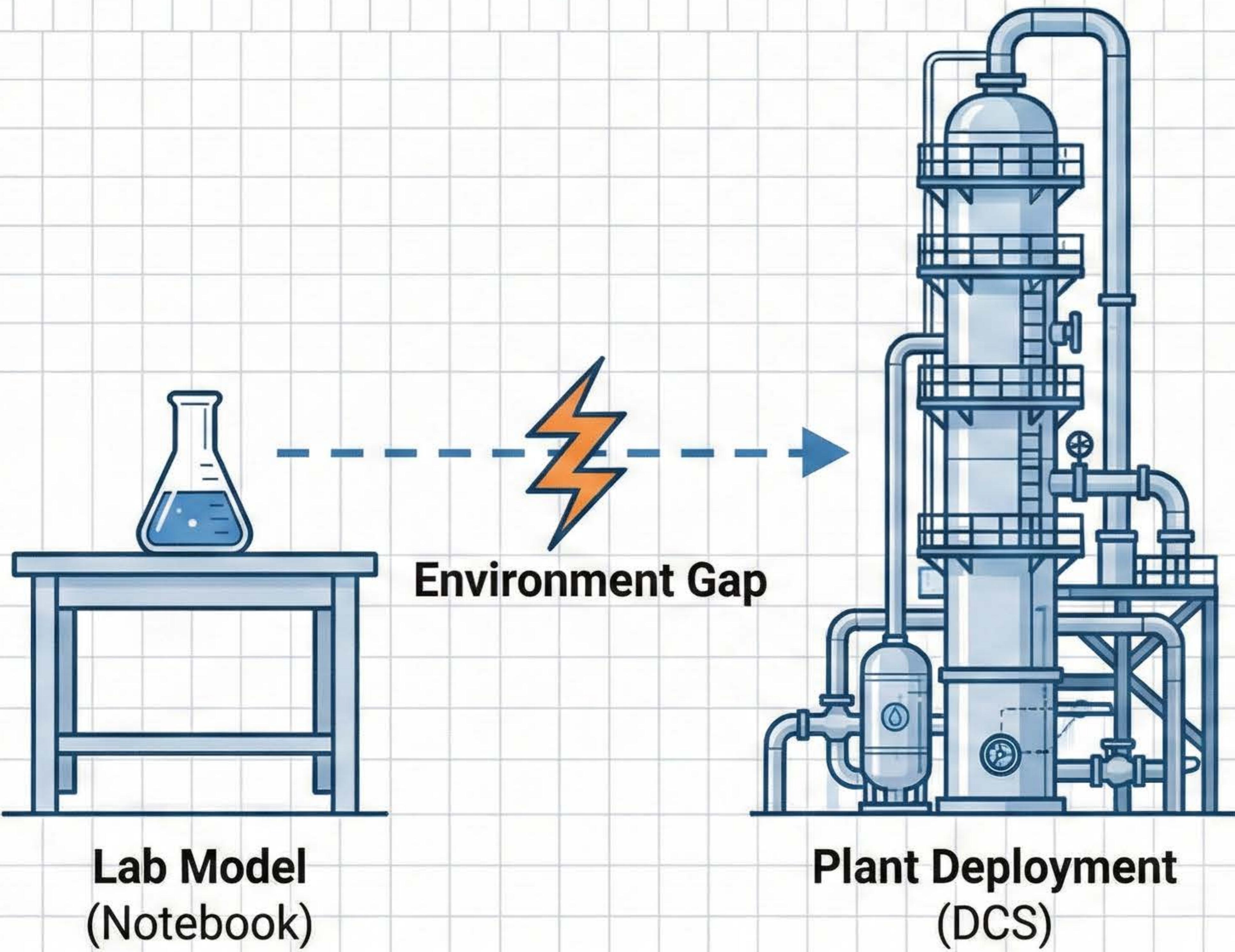
貝氏優化 (Bayesian Optimization)



智慧搜尋，利用過往結果指導方向 (如 Optuna)。

問題：參數組合空間龐大，人工調整易陷入**局部最佳解**。

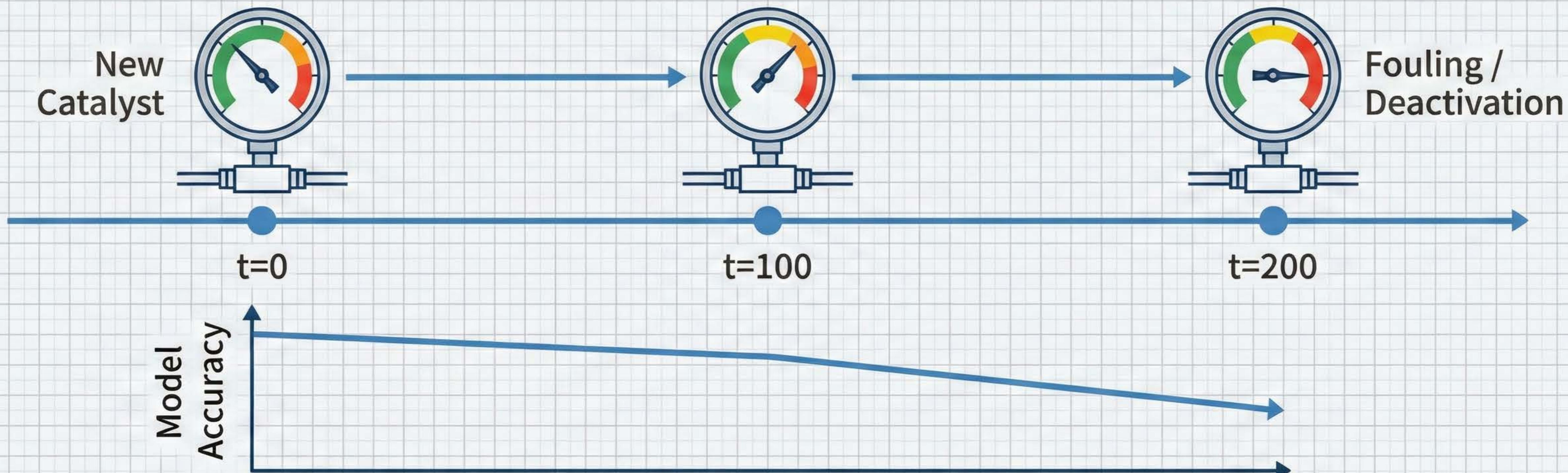
挑戰 8：實驗室到工廠的落差 — 放大效應



如何跨越鴻溝？

- 影子模式 (Shadow Mode)：
模型平行運行，只記錄不控制，驗證真實表現。
- A/B 測試：逐步導入流量，
比較新舊模型差異。
- 真實數據訓練：避免僅依賴
清洗完美的實驗數據。

挑戰 9：概念漂移 (Concept Drift) — 製程的老化

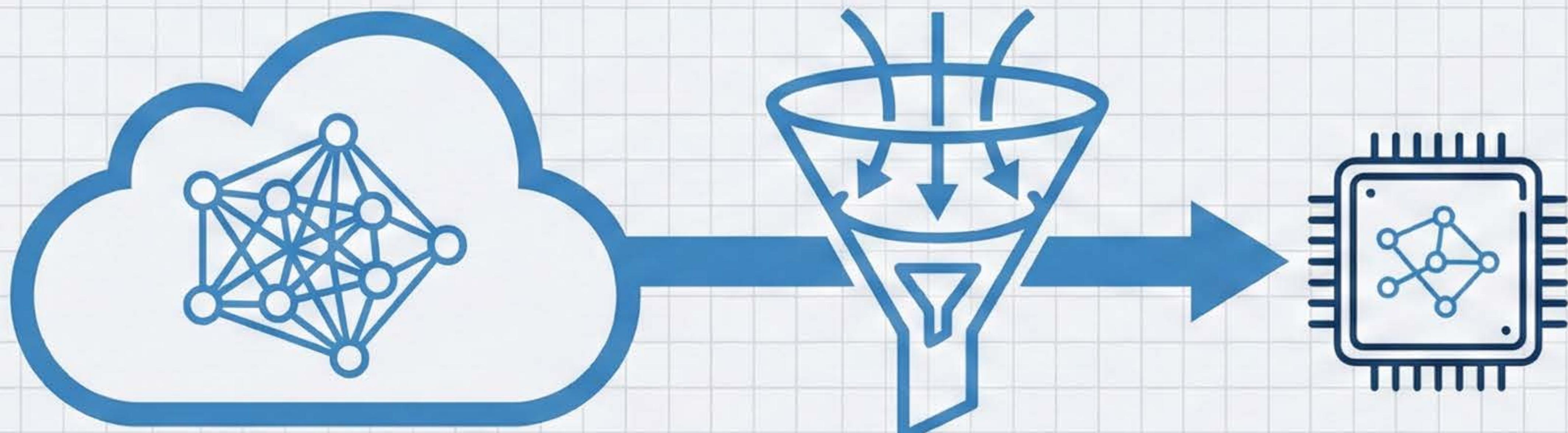


催化劑失活、設備結垢導致物理關係改變，
模型效能衰退。

如何應對 (Solutions)

- 定期重新訓練 (Retraining)：排程更新模型
- 線上學習 (Online Learning)：動態微調參數
- 漂移偵測：監控數據分佈 (Distribution Monitor)

挑戰 10：計算資源 — 邊緣運算 (Edge Computing)



Server
(Training)

Model Distillation
/ Compression

Edge Device
(Inference)

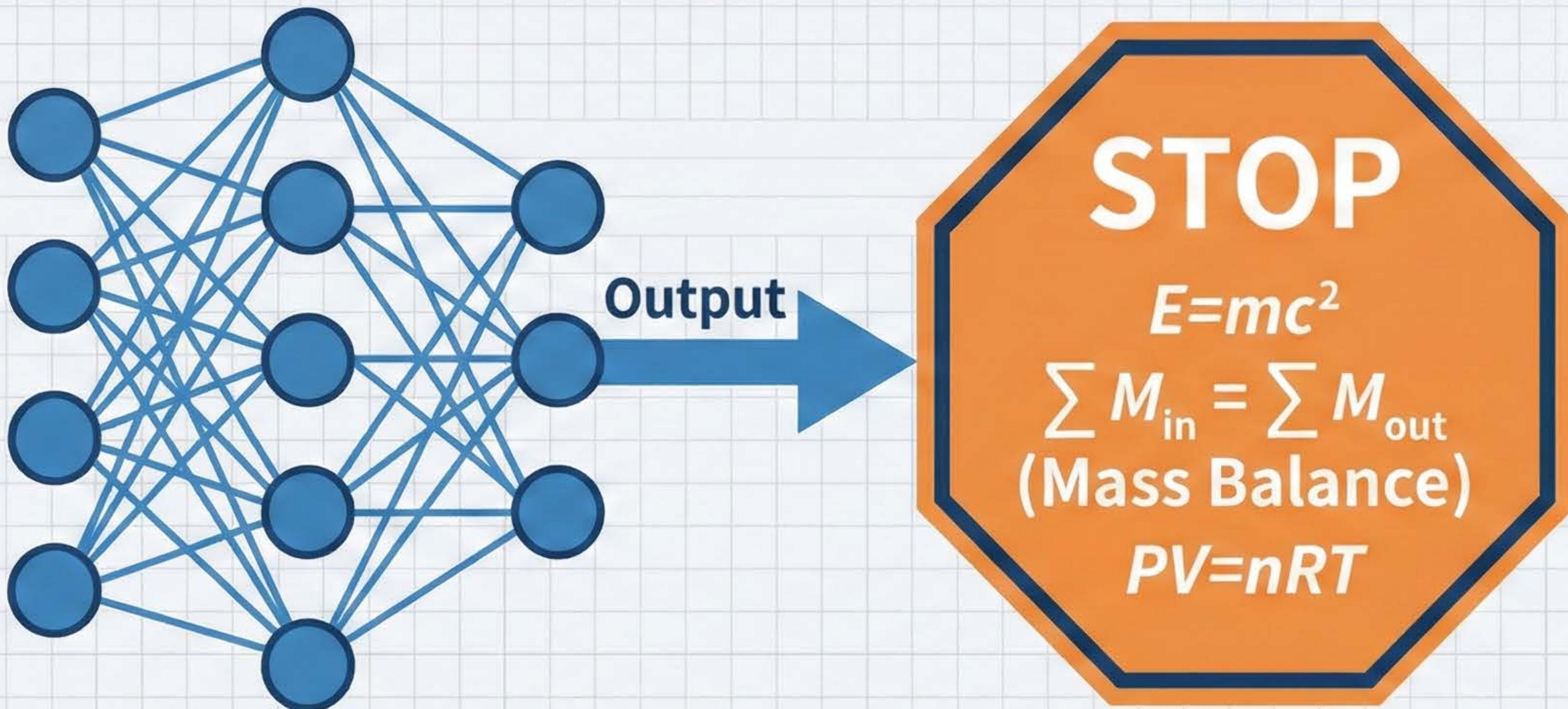
問題 (Problem)

工廠控制系統 (DCS/PLC) 運算能力有限且要求即時性。

解決方案 (Solutions)

- 模型蒸餾 (Distillation)：
大模型教導小模型
- 邊緣運算：
將輕量化模型部署於現場
- 雲端-邊緣協作：
雲端訓練，邊緣推論

挑戰 11：物理限制 (Physical Constraints) — 遵守自然法則



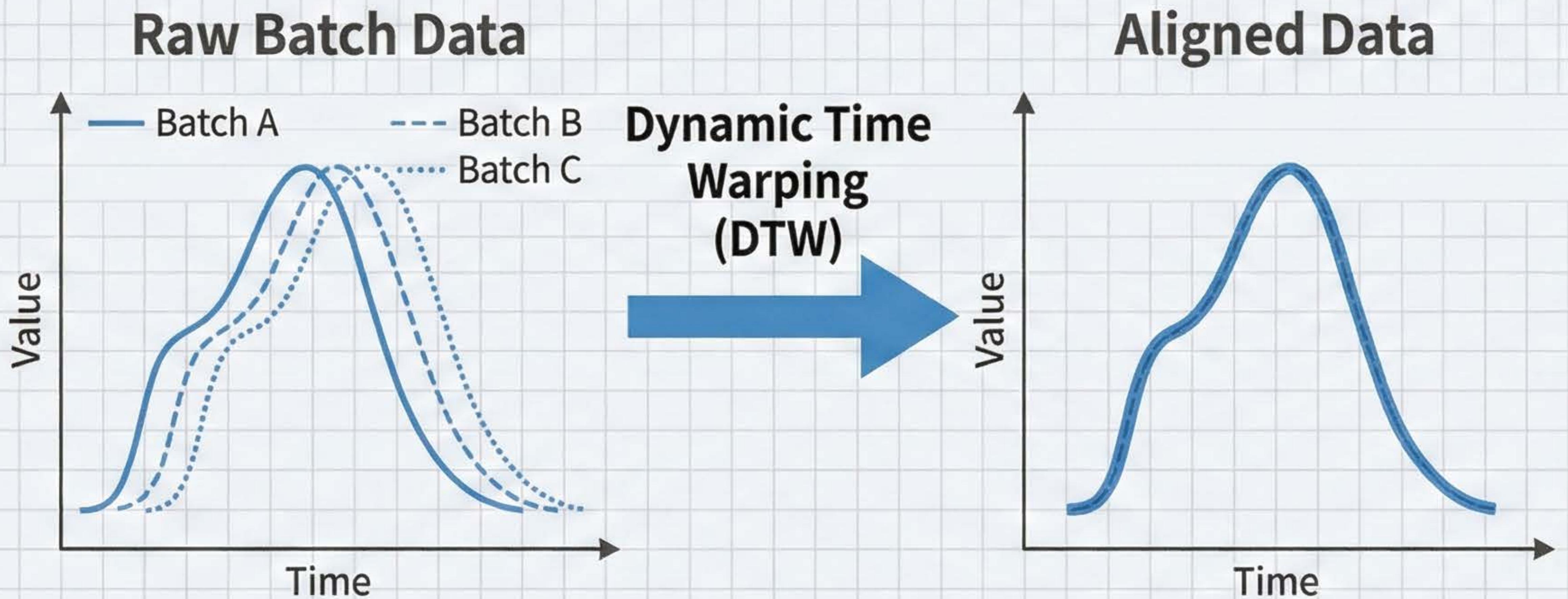
問題 (Problem)

純數據驅動模型可能違反質量守恆或熱力學定律。

解决方案 (Solutions)

- 1. 物理約束嵌入 (PINNs):**
將物理方程式加入 Loss Function。
- 2. 混合建模 (Hybrid Modeling):**
第一原理模型 + ML 殘差修正。
- 3. 安全層 (Safety Layer):**
強制執行輸出上下限。

挑戰 12：批次處理差異 — 對齊時間軸



問題 (Problem)

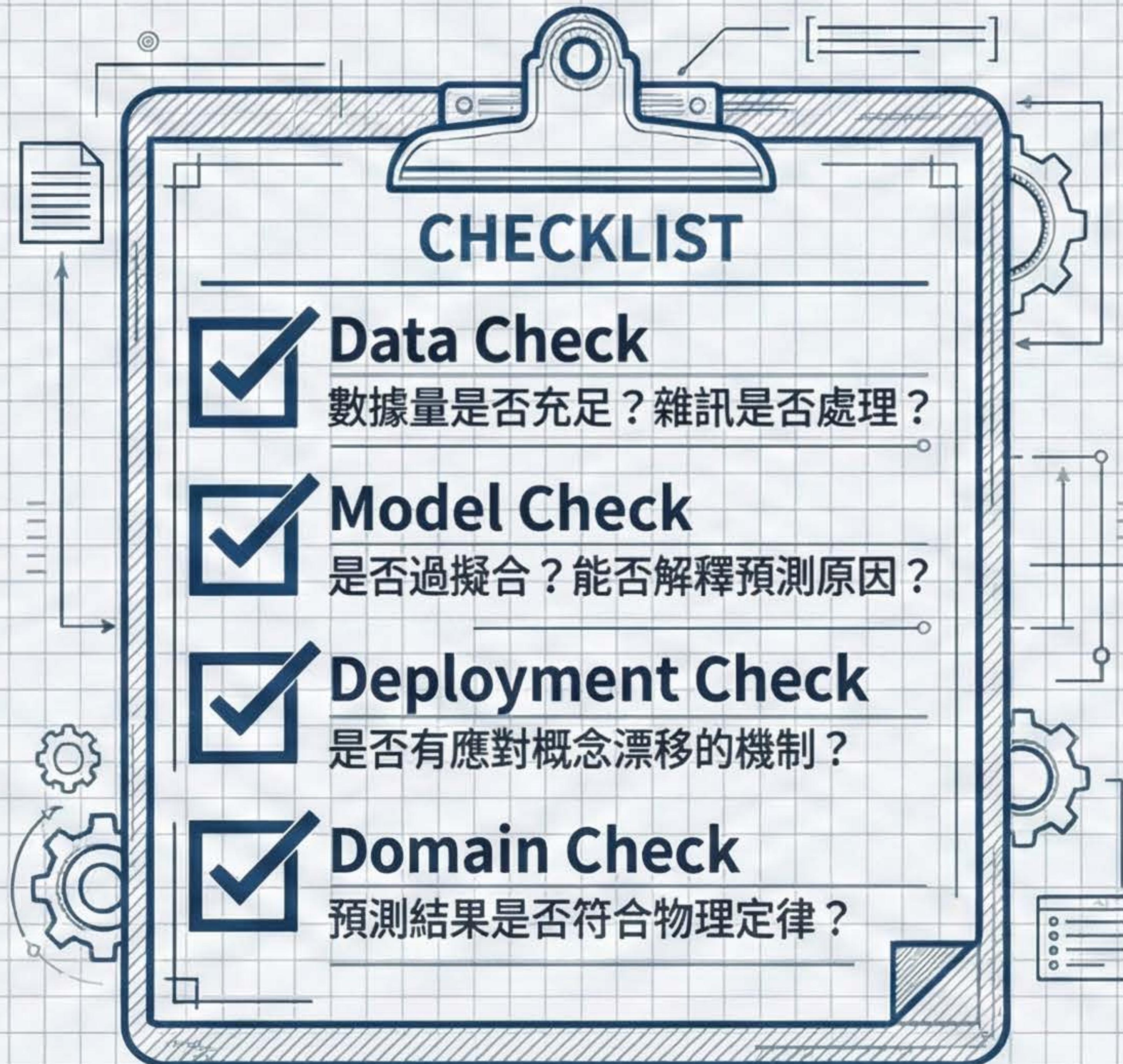
批次長度不同、起始點不一，難以比較。

解決方案 (Solutions)

1. 動態時間校正 (DTW)、多任務學習 (Multi-task Learning)、批次對批次 (Batch-to-Batch) 建模。



總結：ML 工程師的起飛前檢查表



「機器學習的效能上限由資料品質決定。」