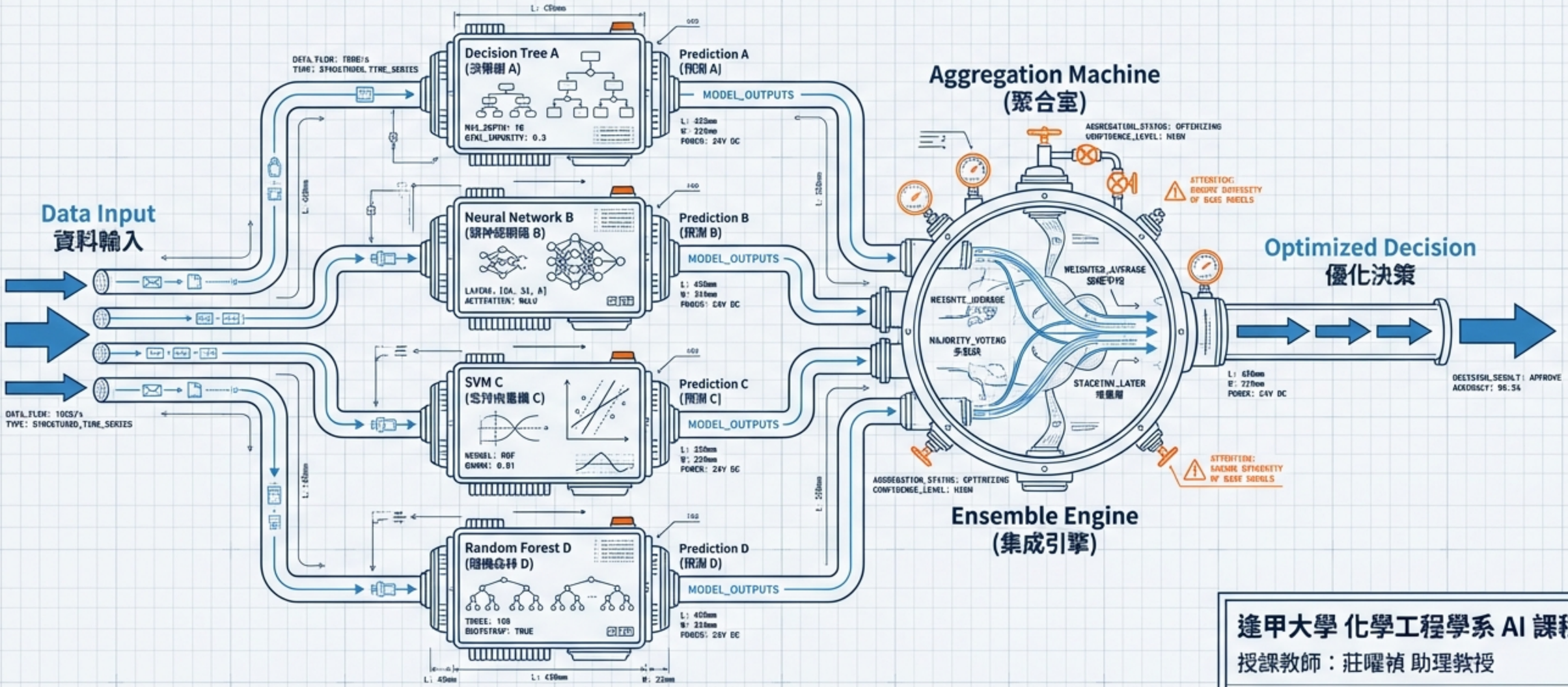
Unit 13 集成學習方法

建構高準確度與強健性的 AI 決策系統



逢甲大學 化學工程學系 AI 課程

授課教師：莊曜禎 助理教授

Designator: Industrial_Blueprint_v1

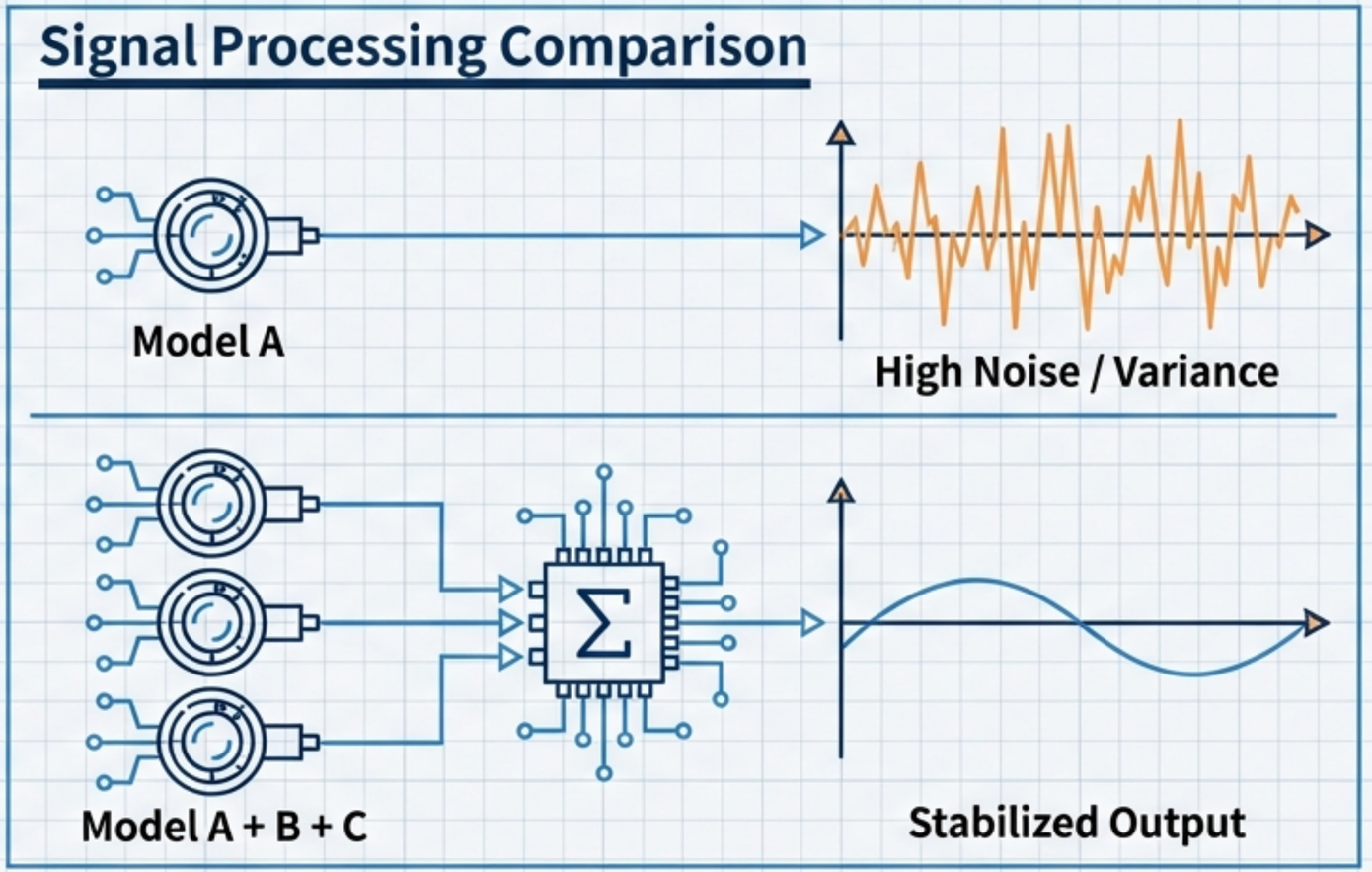
核心概念：為何我們需要「團隊」決策？

定義 (Definition)

透過組合多個基礎學習器 (Base Learner) 的預測結果，以獲得比單一模型更佳的性能。

核心哲學 (Philosophy)

「三個臭皮匠，勝過一個諸葛亮」



System Specs

準確性 (Accuracy)：
降低預測誤差

穩健性 (Robustness)：
抵抗異常值與噪音

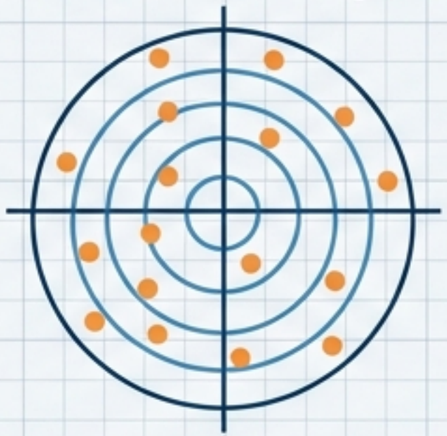
防過擬合 (Overfitting Reduction)：
減少變異

誤差的物理學：偏差-方差分解

$$\text{Error} = \text{Bias}^2 + \text{Variance} + \text{Noise}$$

High Bias
(Underfitting)

Solution:
Boosting
(e.g., XGBoost)



High Variance
(Overfitting)

Solution:
Bagging
(e.g., Random Forest)

Low Bias & Low Variance

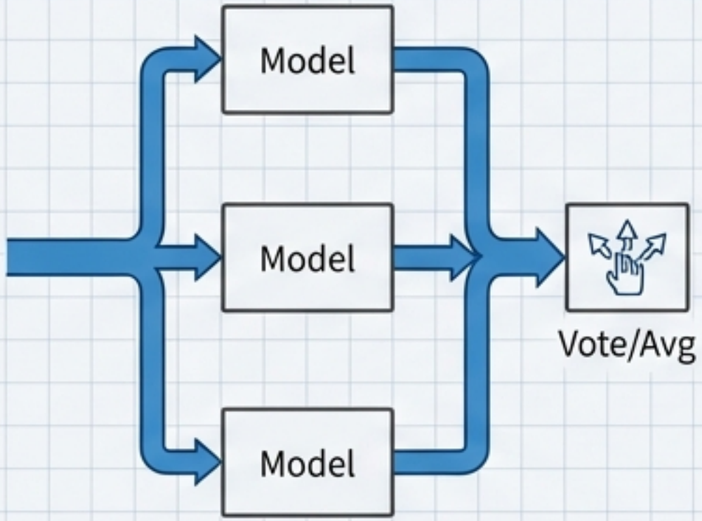
Goal:
Stacking

定義 (Definitions)

Bias (偏差)：擬合能力不足，模型過於簡單。
Variance (方差)：模型過於敏感，對訓練數據微小變化反應過度。

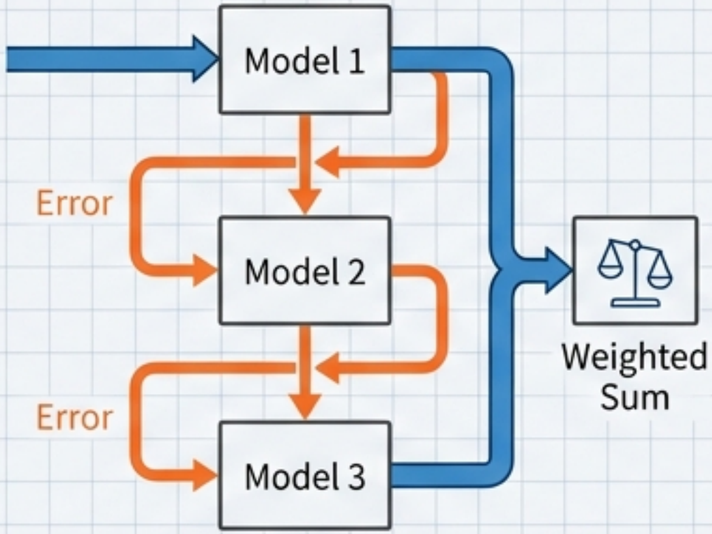
系統架構總覽：三種建構策略

Bagging (Parallel)



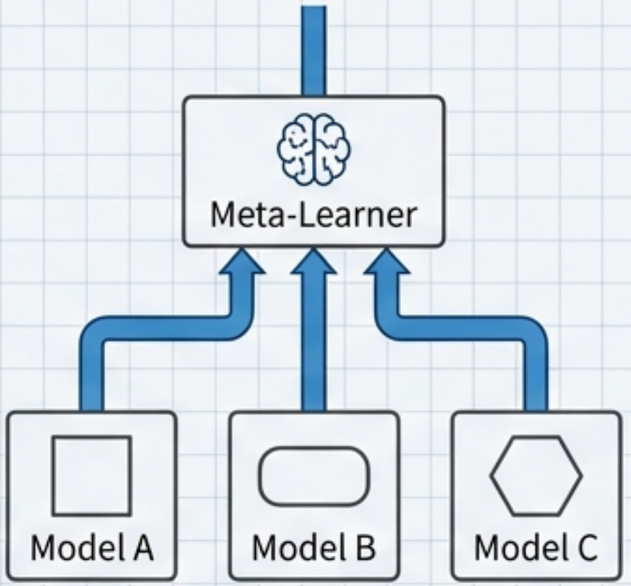
- 並行訓練 (Parallel)
- 隨機性強
- 適用高方差模型 (如決策樹)

Boosting (Sequential)



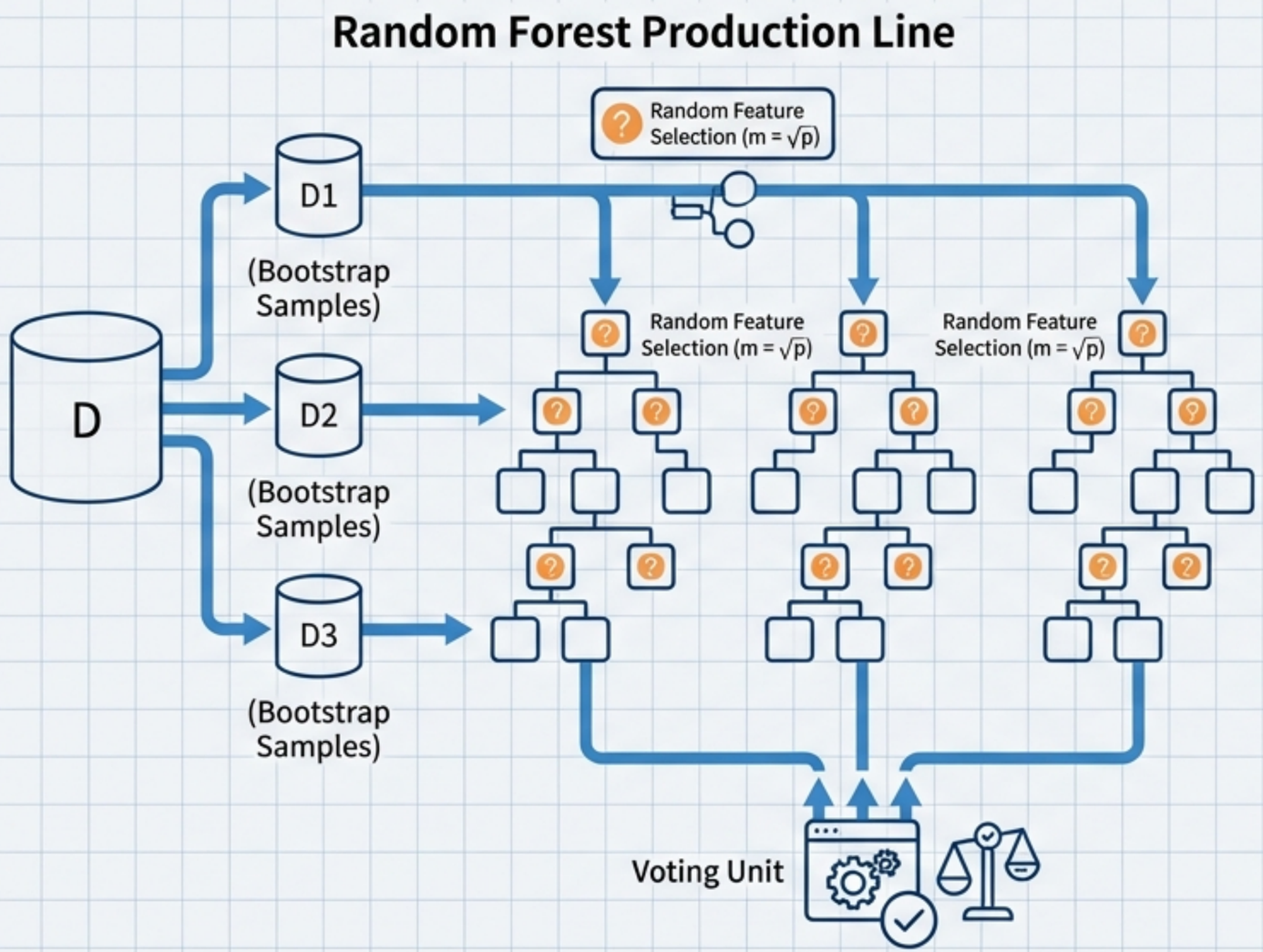
- 序列訓練 (Sequential)
- 修正錯誤
- 適用追求高準確度

Stacking (Hierarchical)



- 多層訓練 (Multi-layer)
- 異質整合
- 適用複雜任務

Bagging 與隨機森林：並行處理與多樣性



核心機制 (Core Mechanism)：Bootstrap Aggregating

- **Bootstrap**：有放回抽樣 (Resampling)。每個子集約含 63.2% 原始樣本。
- **Aggregating**：回歸取平均 (Average)；分類取投票 (Vote)。

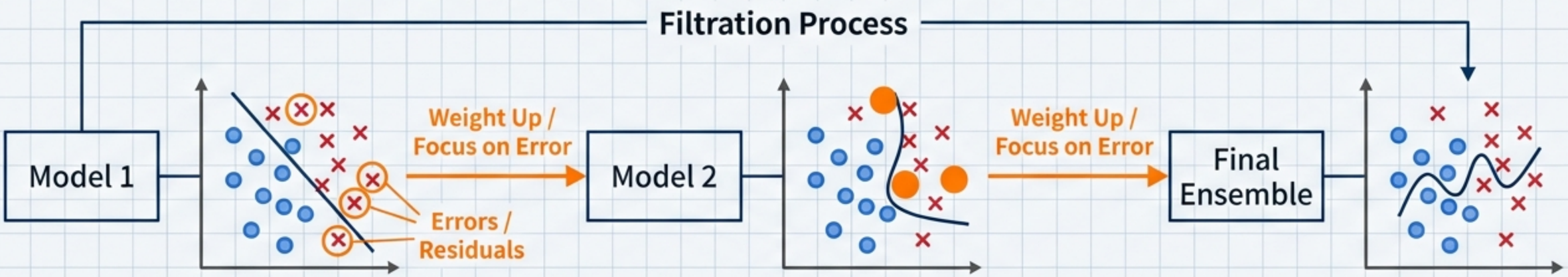
隨機森林參數 (RF Specs)

- **Data Randomness**: Bootstrap sampling.
- **Feature Randomness**: 分裂時隨機選取特徵。

品質檢測 (Validation)

- **Out-of-Bag (OOB) Error**: 使用未被抽樣的數據進行自我驗證，無需額外驗證集。

Boosting 概念：迭代式強化學習



核心邏輯 (Core Logic)

1. **Sequential Training**: 後一個模型依賴前一個模型的結果.
2. **Focus on Errors**: 專注於被錯誤分類或**殘差較大**的樣本.
3. Weighted Combination: $F(x) = \sum \alpha_t h_t(x)$

AdaBoost (1995): 調整樣本權重 → **Gradient Boosting (2001):** 擬合負梯度 (Pseudo-residual) → **XGBoost (2016):** 系統優化

極致梯度提升：XGBoost (Extreme Gradient Boosting)

Blueprint Spec Sheet

Technical Upgrades

二階泰勒展開 (Taylor Expansion):
利用一階與二階梯度加速收斂.

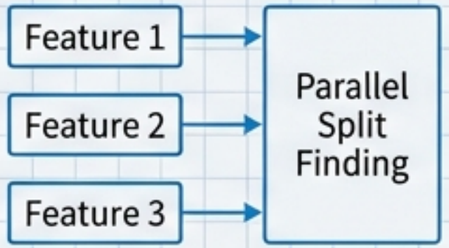
缺失值處理 (Missing Values):
自動學習缺失值的最優分裂方向.

$$Obj = L(\theta) + \Omega(\theta)$$

Loss Function:
準確度 (Accuracy)

Regularization:
複雜度懲罰 (Simplicity)
-> 防止過擬合

系統優化 (System Optimization):
特徵並行化計算 (Parallelization)
與 樹剪枝 (Pruning).



進階 Boosting 框架：LightGBM 與 CatBoost

Product Card

LightGBM
(Efficiency Specialist)

Histogram-based
連續特徵離散化 (Memory friendly)

Leaf-wise Growth
每次分裂增益最大的葉子
(High Speed)

Best For:
Big Data (>100k samples)

Product Card

CatBoost
(Category Specialist)

Native Categorical Handling
無需 One-Hot Encoding

Ordered Boosting
解決預測偏移

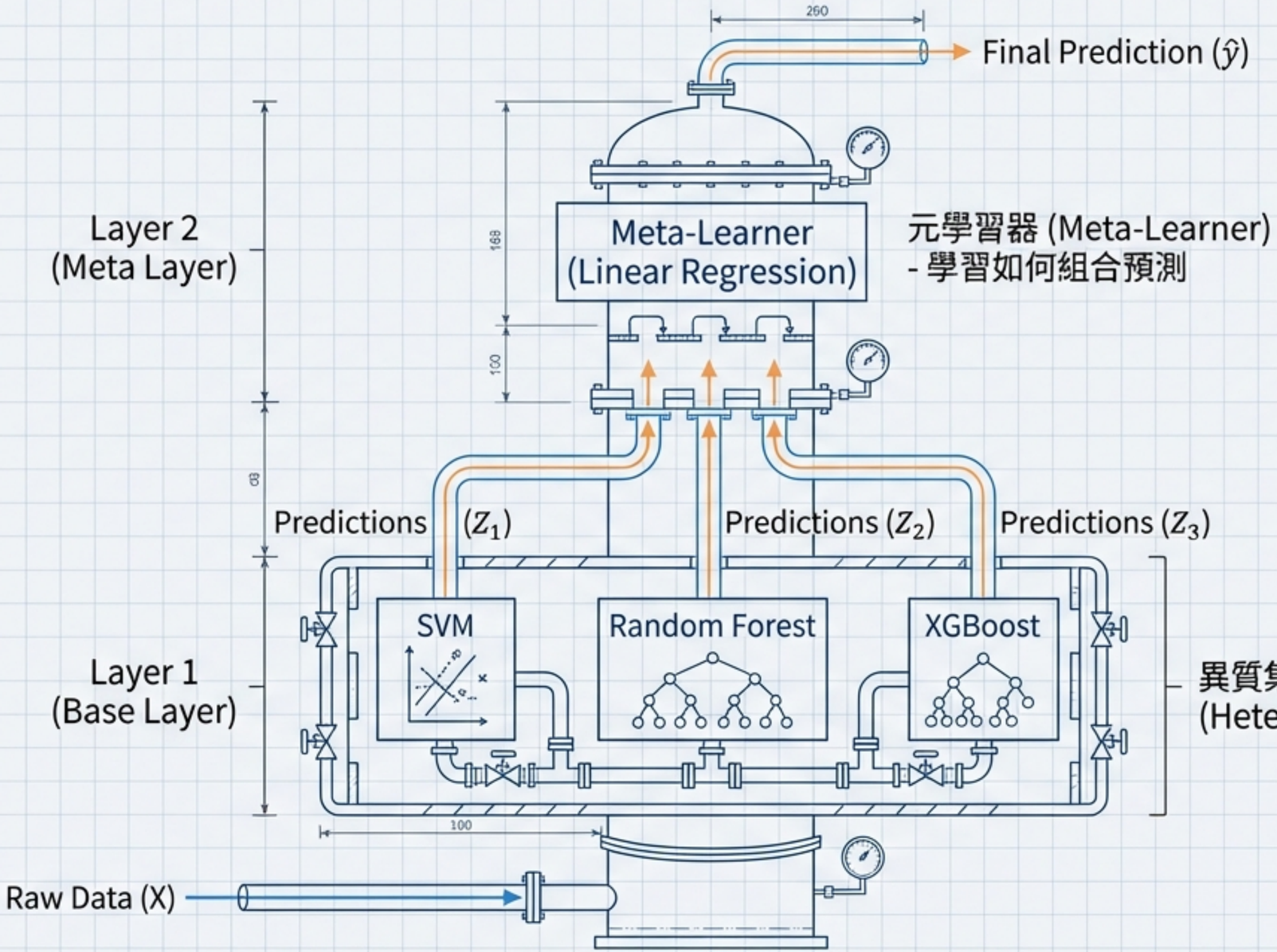
Symmetric Trees
對稱樹結構 (Fast Prediction)

Best For:
Data with many categories

框架對比 (Framework Comparison): Speed vs. Accuracy vs. Features

Framework	Speed	Accuracy	Categorical Handling
XGBoost		✓	✓
LightGBM		✓	✓
CatBoost		✓	🏷

Stacking：多層模型堆疊 (Model Stacking)



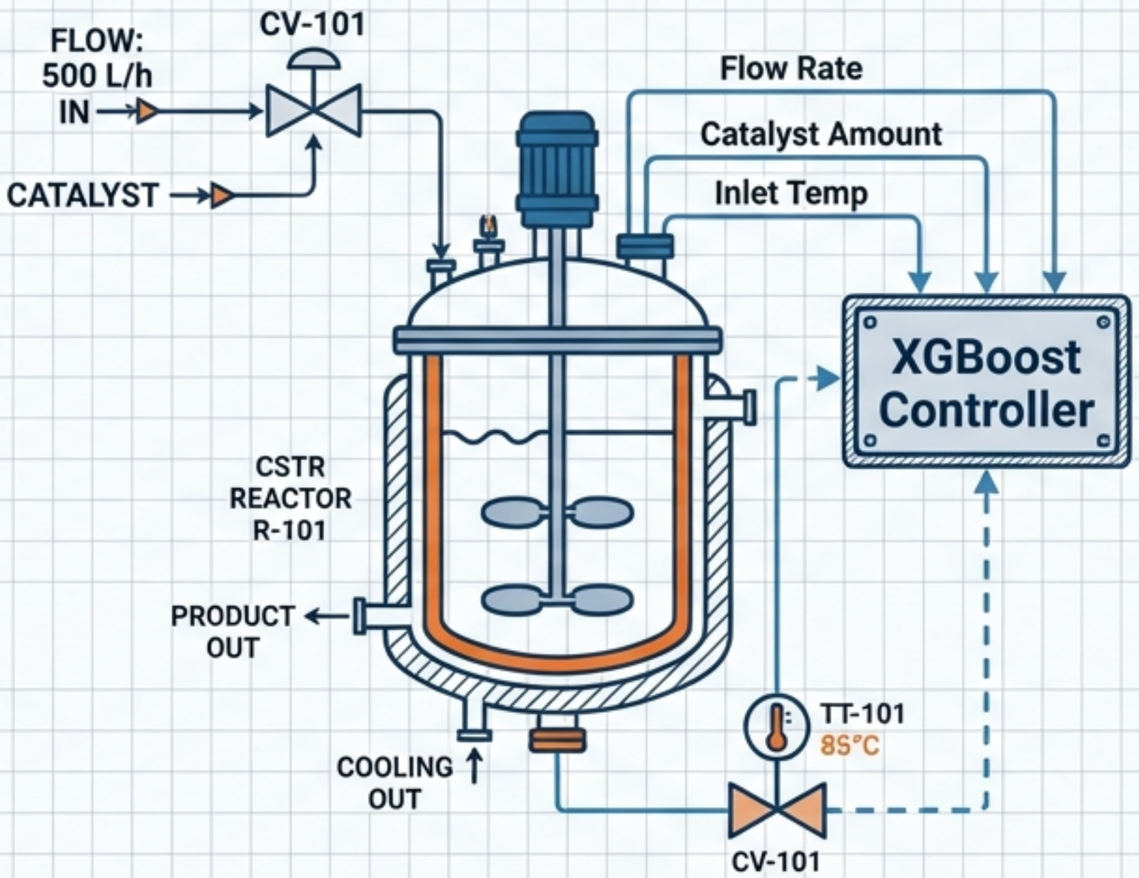
****Workflow Insight****

使用 K-Fold Cross-Validation 生成訓練 Meta-Learner 的預測值 (Out-of-Fold predictions)，防止數據洩漏。

化工應用 (1)：程序控制與優化

反應器溫度控制 (Reactor Control)

- **問題：**非線性動力學，傳統 PID 難以精確控制。
- **解法：XGBoost** 預測出口溫度，實時調整冷卻水流量。



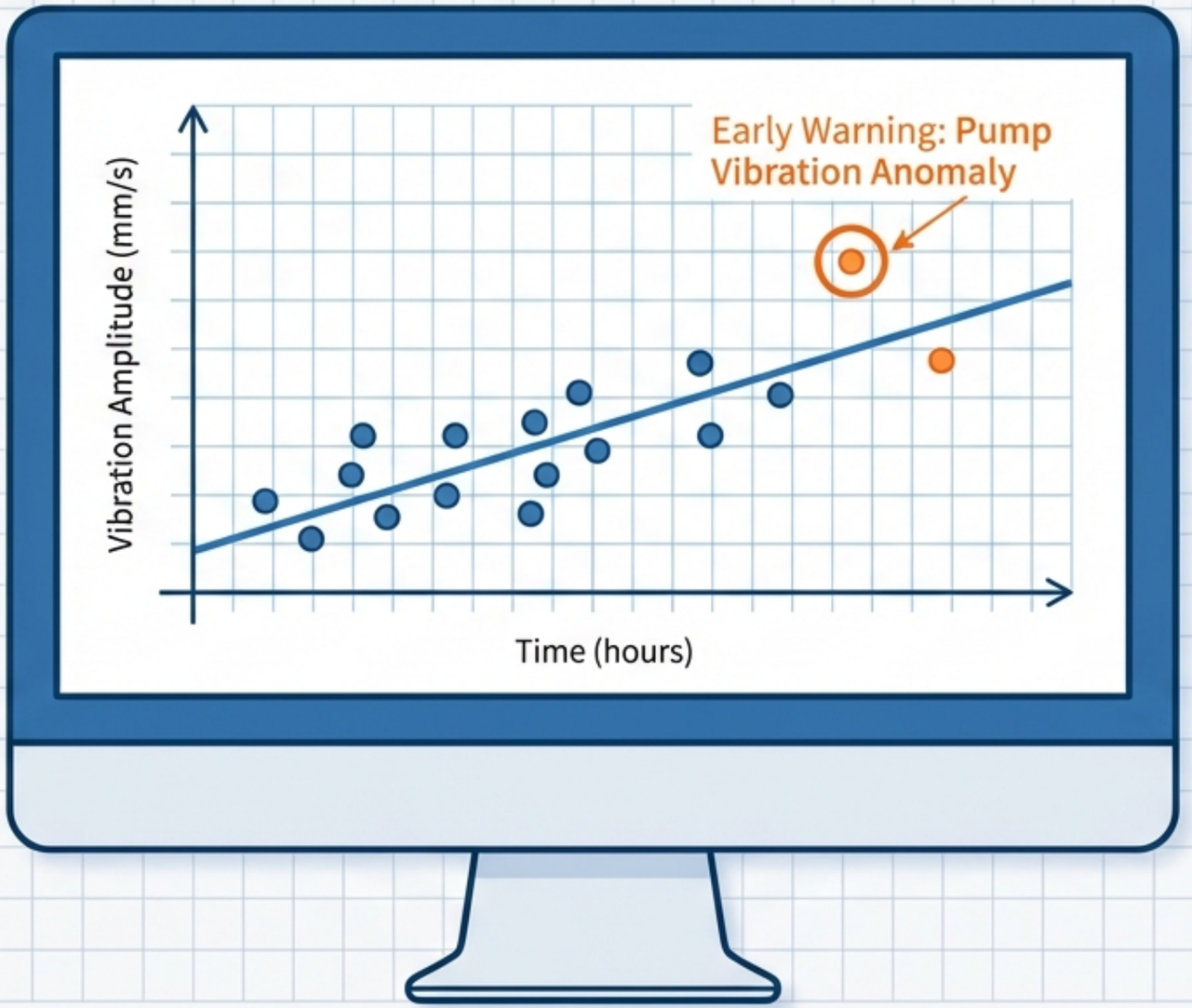
蒸餾塔優化 (Distillation Optimization)

- **解法：Random Forest** 分析關鍵變數 (Feature Importance)，優化產品質量。

化工應用(2)：故障診斷與預測性維護

設備健康監控 (Equipment Health)

- **目標:** 泵浦振動 (Vibration) 或 換熱器結垢 (Fouling)。
- **方法:** **Random Forest Classifier** 識別故障類型 (e.g., 軸承磨損 vs. 對中不良)。

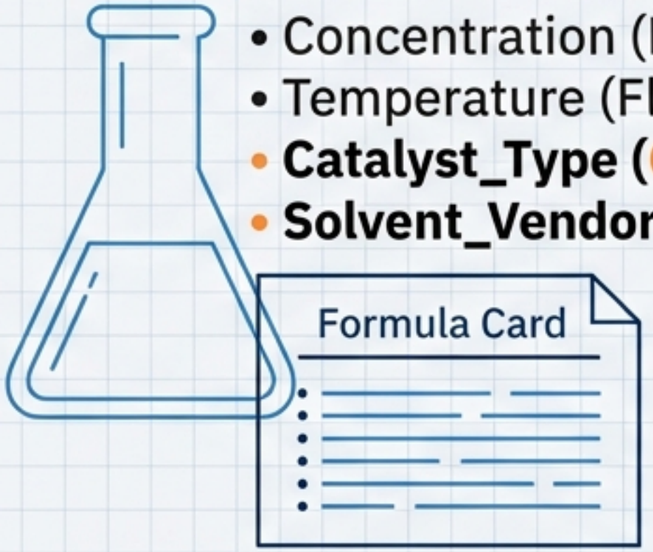


異常檢測 (Anomaly Detection)

- **方法:** **LightGBM** 處理大規模歷史數據 (Big Data)。
- **應用:** 實時監控製程參數，於故障發生前發出警報。

化工應用 (3)：配方設計與物性預測

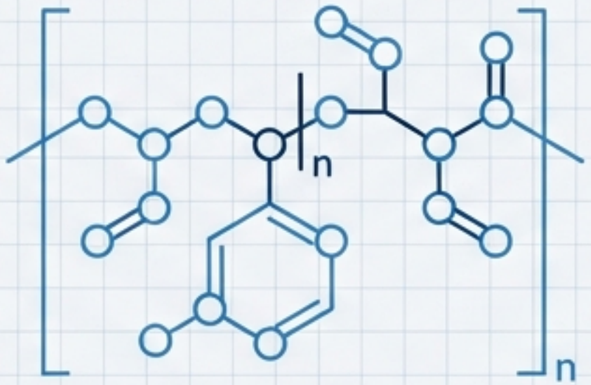
Mixed Input Data



- Concentration (Float)
- Temperature (Float)
- **Catalyst_Type (Category)**
- **Solvent_Vendor (Category)**

CatBoost Model

Property Prediction



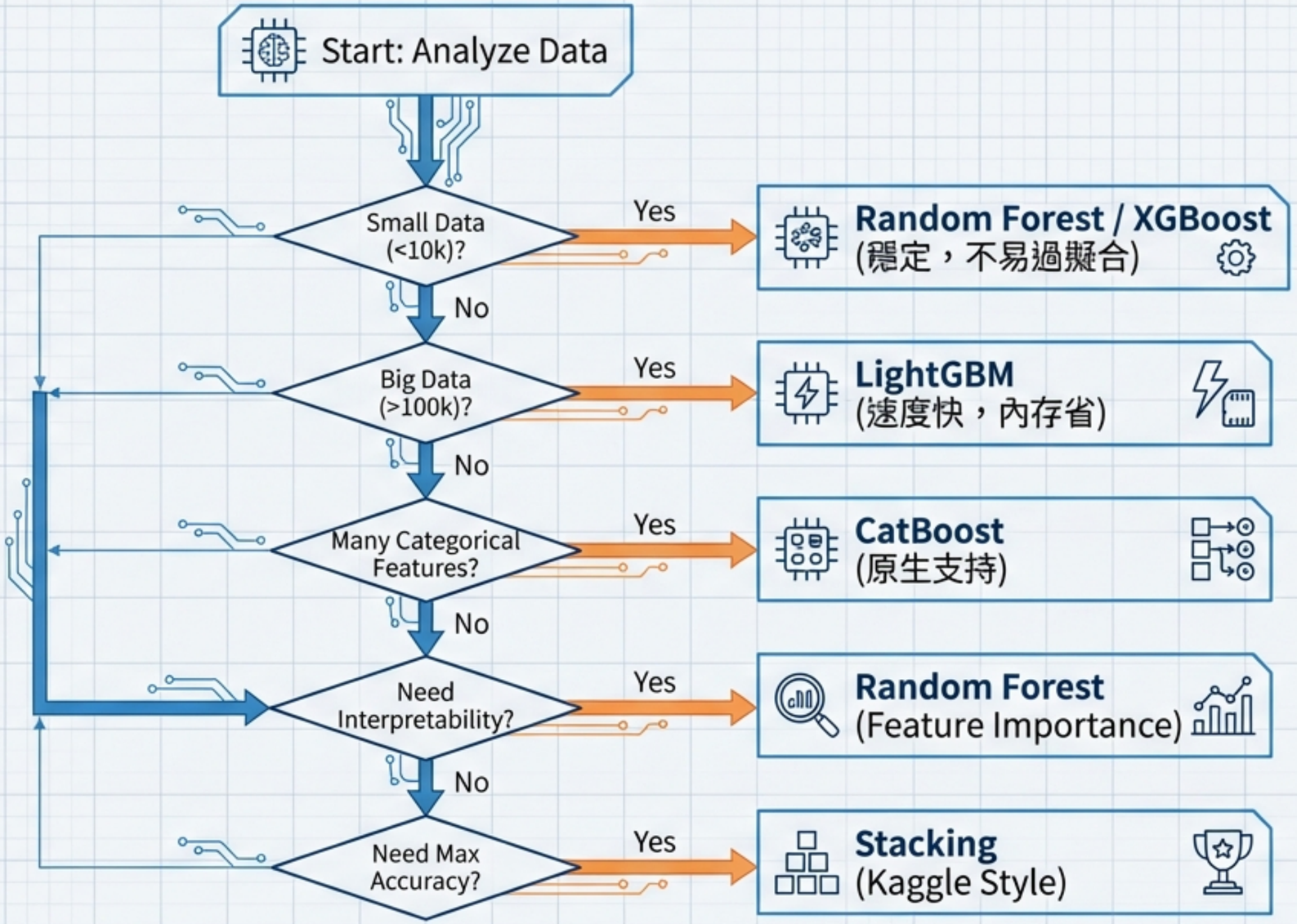
- Viscosity: 120 cP
- Yield: 98.5%

挑戰: 混合數據類型 - 連續型 + 類別型.

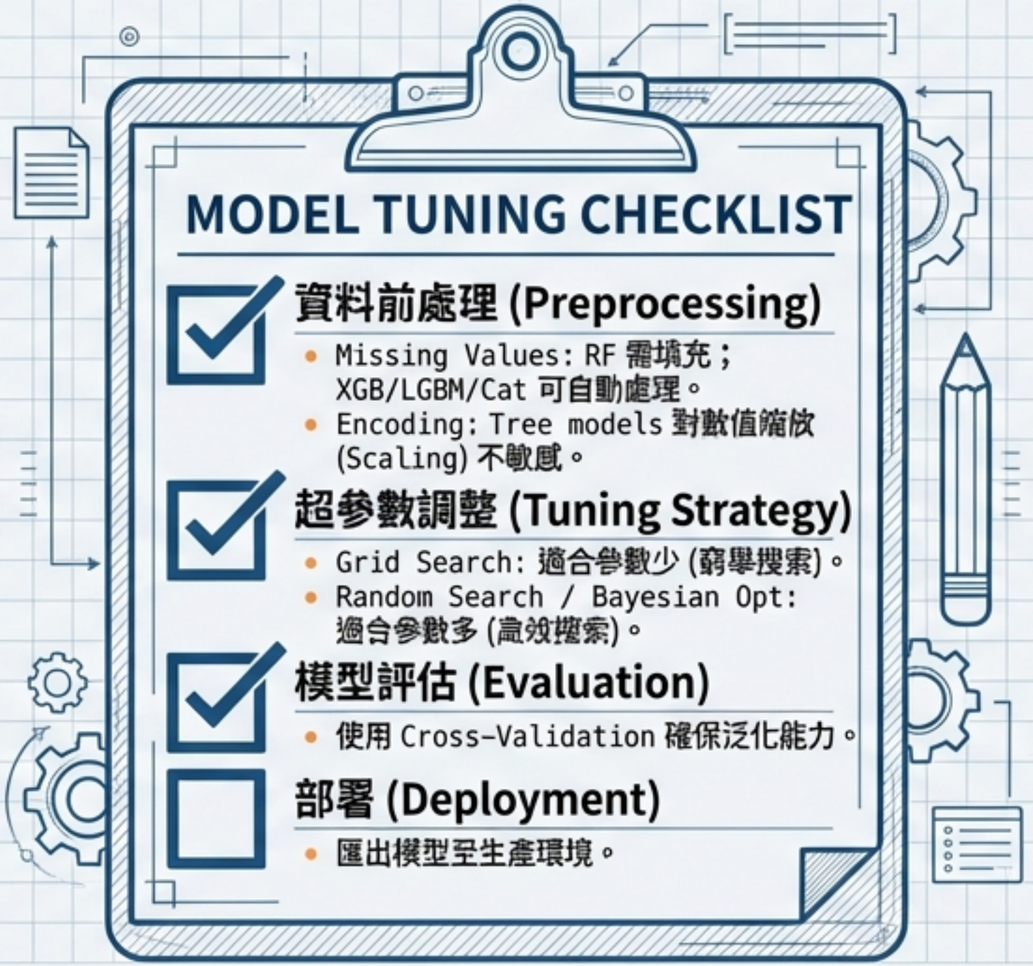
解決方案: CatBoost

1. 原生處理類別特徵 (Native Categorical Support).
2. 無需繁瑣的 One-Hot Encoding.
3. 應用於聚合物分子量分佈預測、塗料性能預測.

模型選擇指南 (Selection Guide)

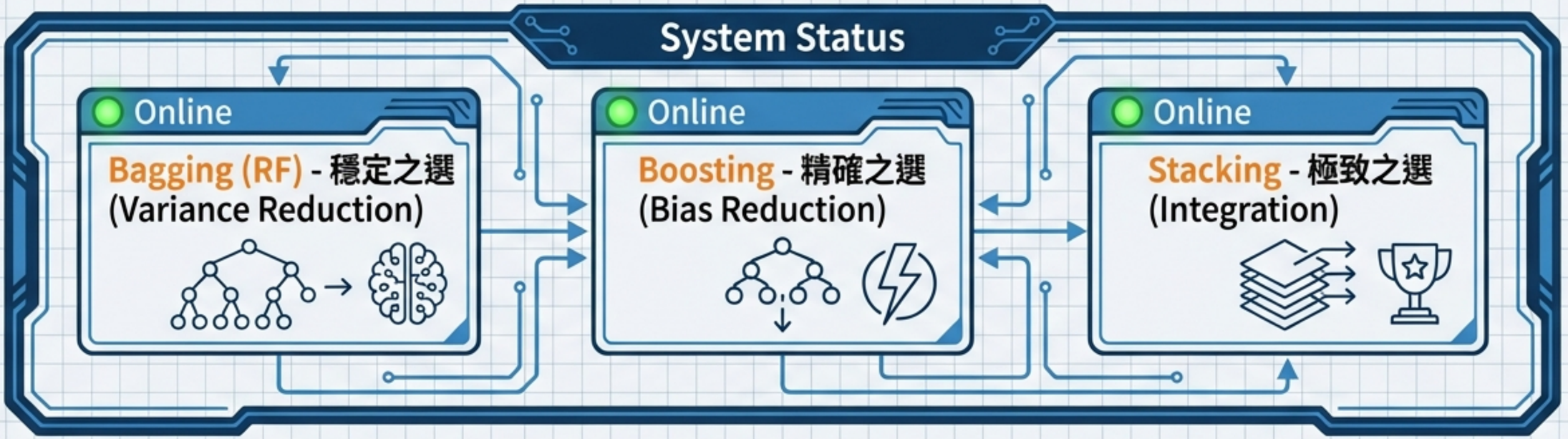


實務建構清單 (Implementation Checklist)



"A machine is only as good as its calibration."

單元總結與下一步 (Summary & Roadmap)



Actionable Next Steps

> 在化工數據集上比較 XGBoost, LightGBM, CatBoost。
> 練習 Unit 13 Jupyter Notebooks.
> System Status: Unit 13 Complete.

Unit 13 結束 | 前往 Unit 14