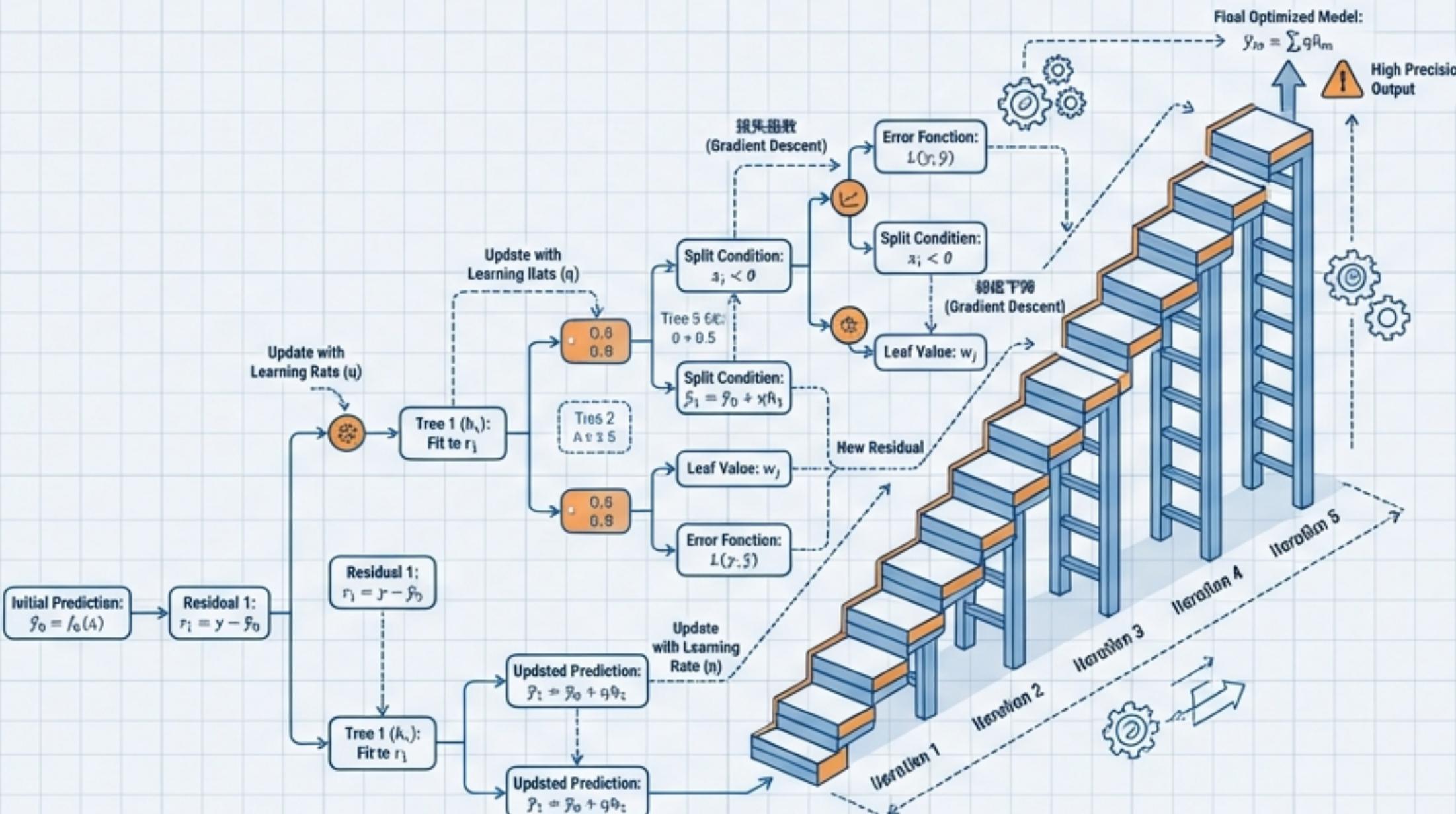


Unit 11: 梯度提升樹回歸 (Gradient Boosting Trees)

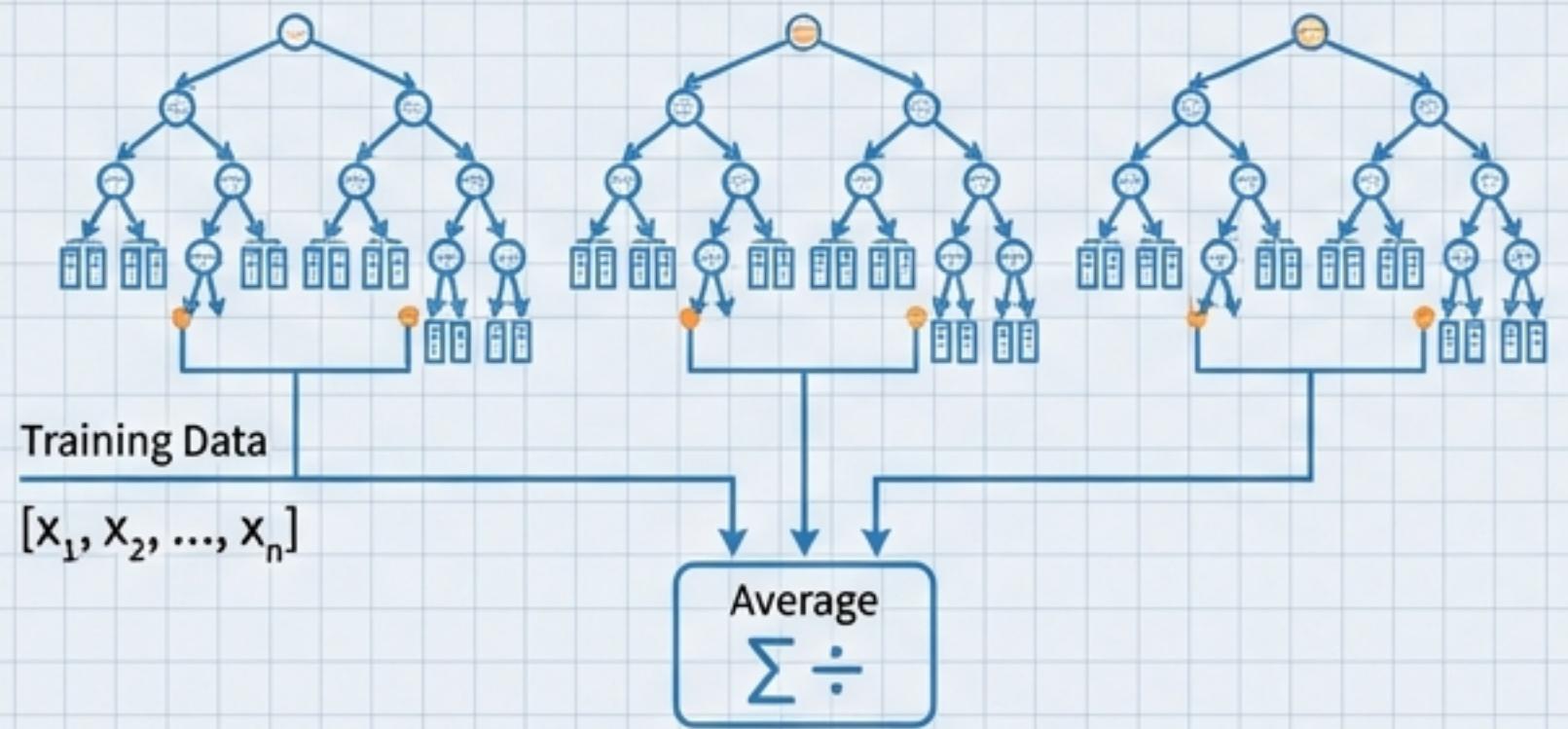
化工製程的精準優化與序列學習機制 (Precision Optimization & Sequential Learning)



適用對象：化學工程學系學生
製作單位：逢甲大學 化工系 智慧程序系統工程實驗室

集成學習的進化：從並行平均到序列修正

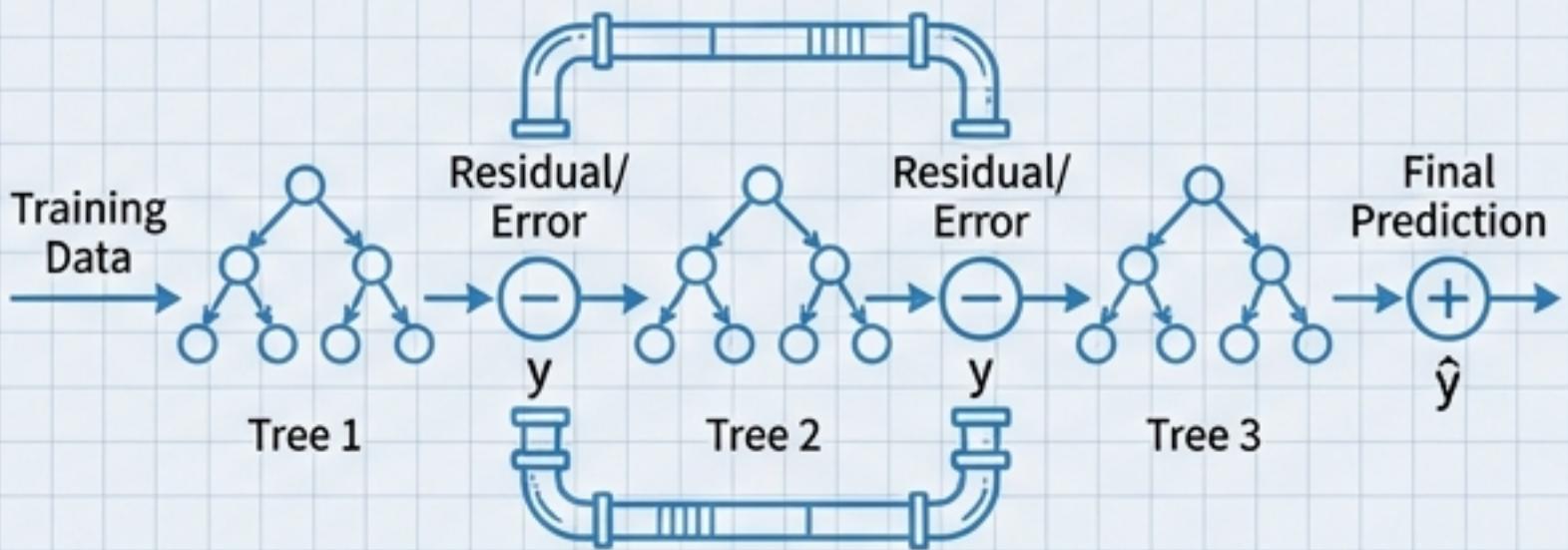
並行訓練 (Parallel) - 隨機森林



民主機制：多數決或平均

- 獨立訓練
- 深度樹 (Deep Trees)
- 旨在降低變異 (Variance)

序列訓練 (Sequential) - 梯度提升樹

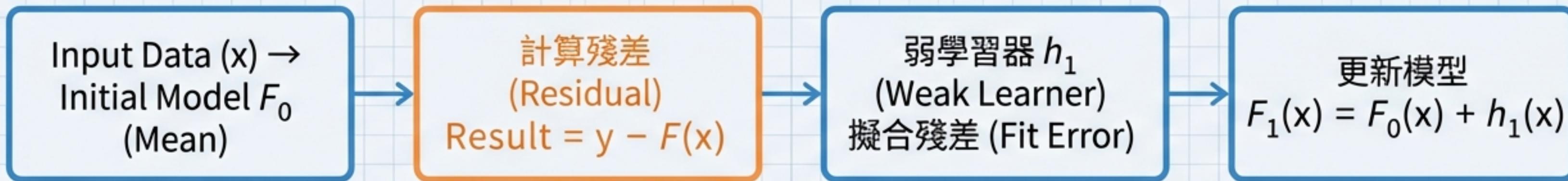


修正機制：專注於改進錯誤

- 依賴訓練
- 淺層樹 (Shallow Trees)
- 旨在降低偏差 (Bias)

核心差異：RF 是「三個臭皮匠，勝過一個諸葛亮」；GBT 是「專家診斷，對症下藥」。

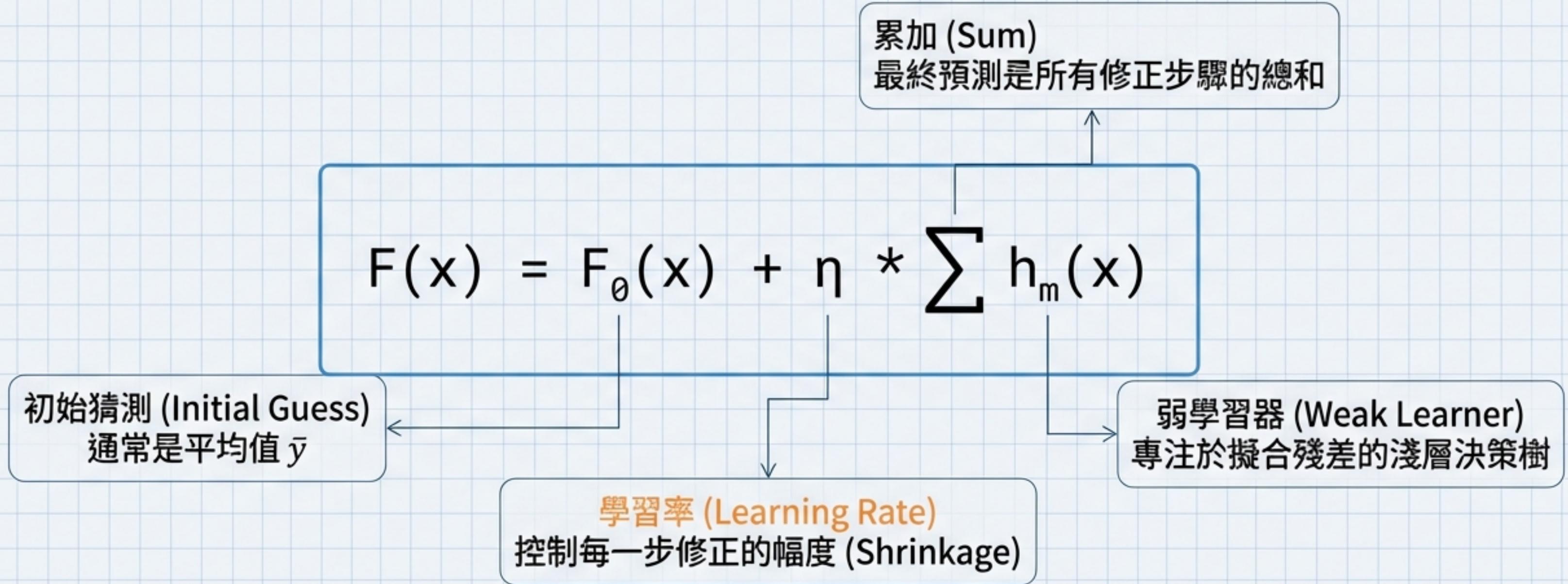
核心原理：我們不預測目標，我們預測「殘差」



梯度 (Gradient) \approx 殘差 (Residual)：在均方誤差 (MSE) 損失函數下，負梯度正好等於殘差 ($y - F(x)$)。

直觀理解：這就像打高爾夫球。第一桿打過去 (F_0)，離洞口還有距離（殘差）；第二桿 (h_1) 不對瞄準洞口，而是瞄準「剩下的距離」進行修正。

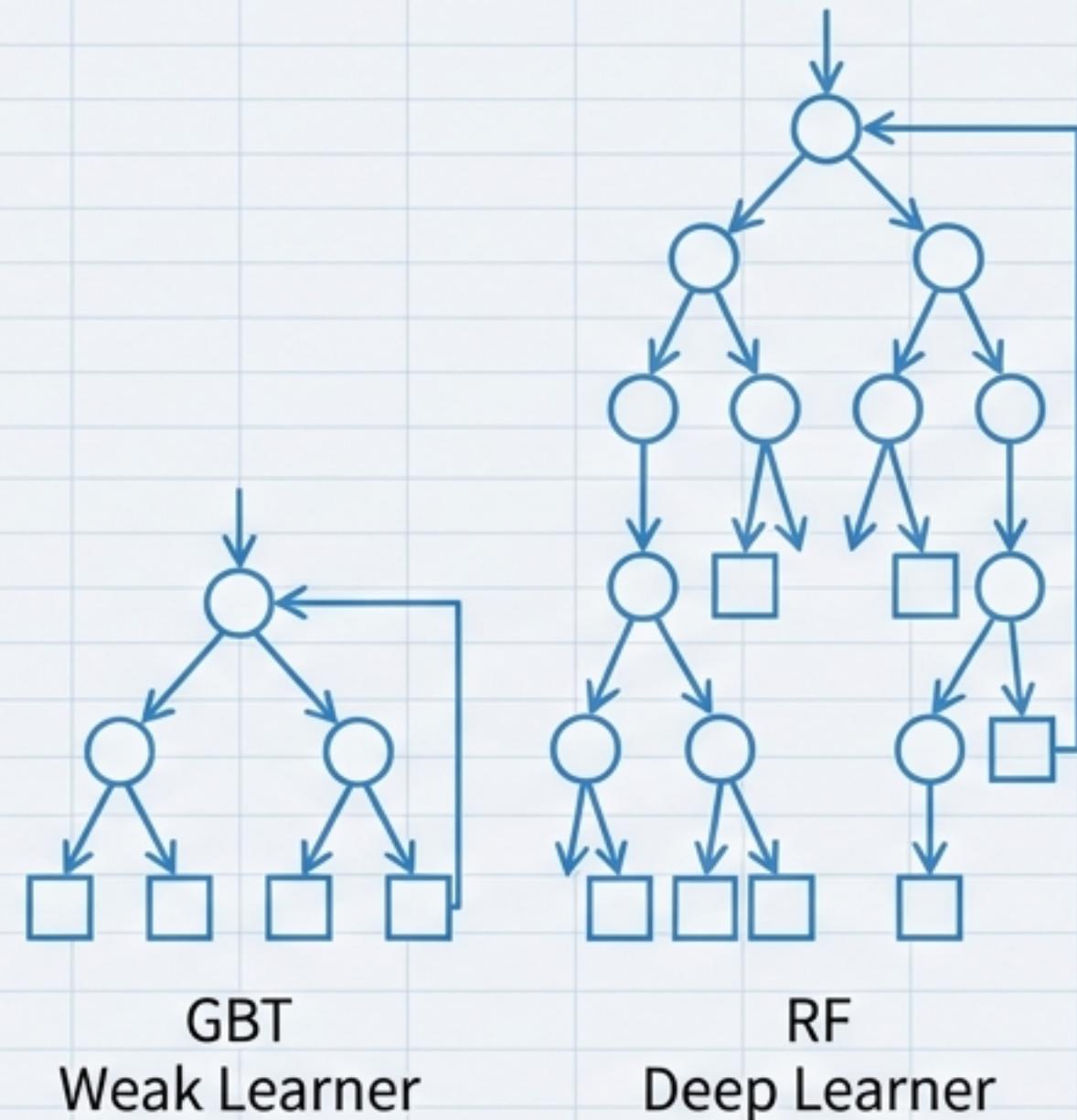
數學架構：加法模型 (Additive Model)



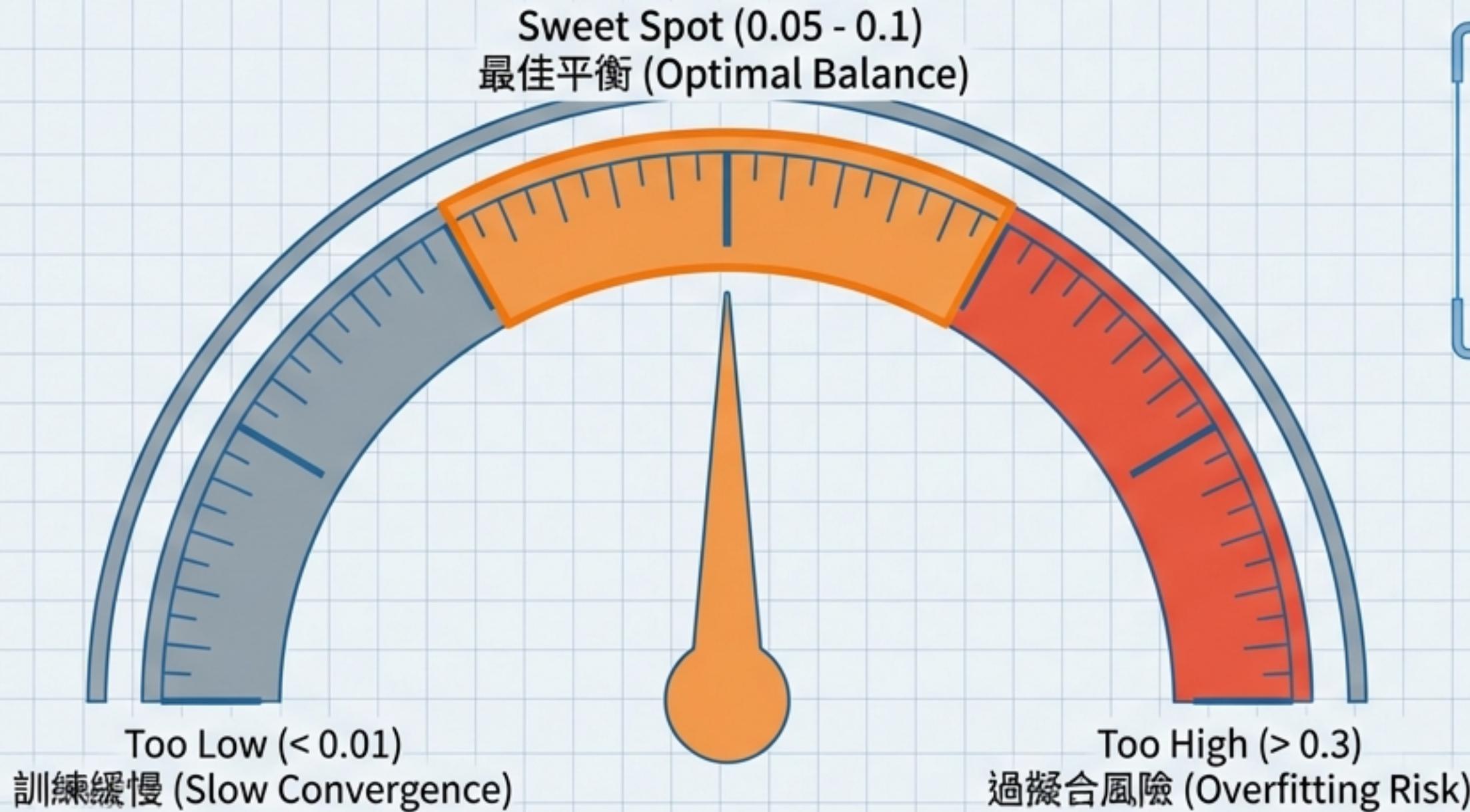
Takeaway: 透過積累許多微小的改進 (弱學習器)，最終構建出一個強大的預測模型。

控制面板：關鍵超參數 (Hyperparameters)

參數 (Parameter)	梯度提升樹 (GBT)	隨機森林 (RF)	作用 (Function)
n_estimators	50-200 (適中)	100-500 (多)	樹的數量 (總迭代次數)
max_depth	3-6 (淺層樹)	深 (無限制)	複雜度控制。 GBT 使用「弱學習器」，不需深樹
learning_rate	0.01-0.1 (關鍵)	N/A	收斂速度與穩定性 。GBT 特有參數



速度與精度的權衡：學習率 (Learning Rate)



實驗數據 (Experimental Data):

- 最佳設定 : learning_rate = 0.10
- 配合樹數量 : 100 棵
- 測試 R² : 0.769

權衡公式：較小的 η + 較多的樹 = 更好的泛化性能 (但計算成本增加)。

安全機制：防止模型過熱 (Regularization)



淺樹限制 (Shallow Depth)

設定 `max_depth=3~5`。限制單棵樹的複雜度，防止記憶噪聲 (Noise)。



隨機子採樣 (Stochastic Subsampling)

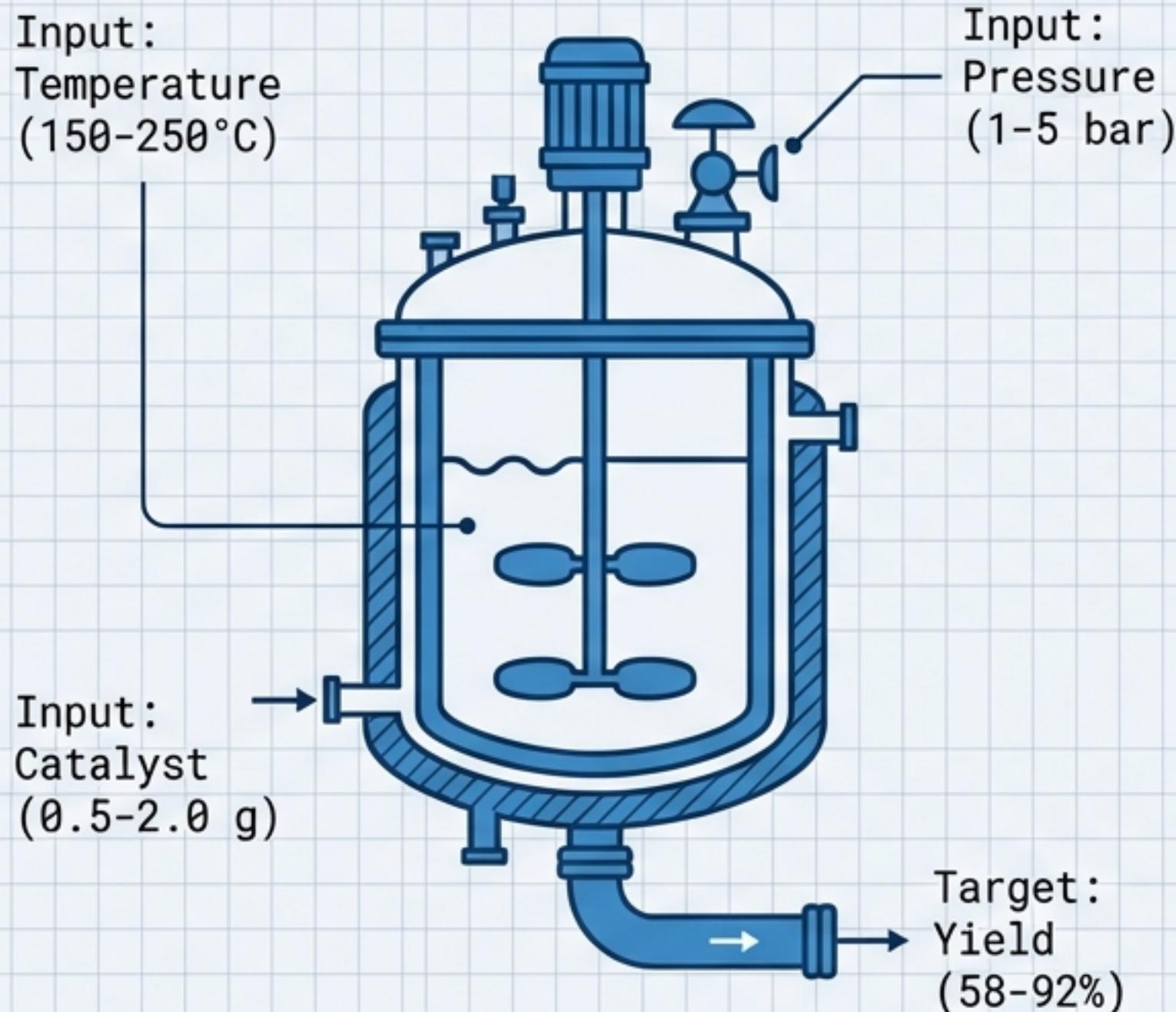
設定 `subsample=0.8`。每次迭代只用 80% 數據。引入隨機性，提升魯棒性。



早停機制 (Early Stopping)

監控驗證集誤差 (Validation Error)。當誤差不再下降 (`n_iter_no_change`) 時，自動停止訓練。效益：節省時間並防止過擬合。

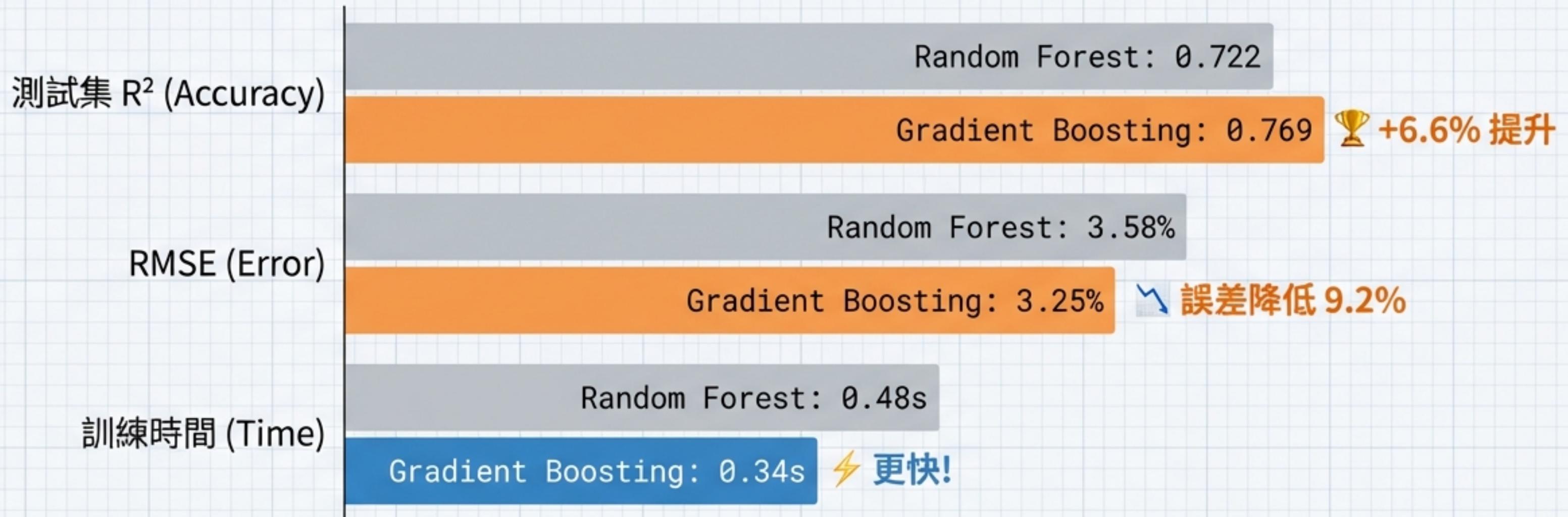
任務代號：CSTR 催化反應器產率優化



數據概況 (Data Profile):

- **挑戰：**非線性反應動力學，變數間存在複雜交互作用。
- **目標：**超越隨機森林的預測準確度。
- **變數：**3個輸入 (T , P , C)，1個輸出 (Yield)。

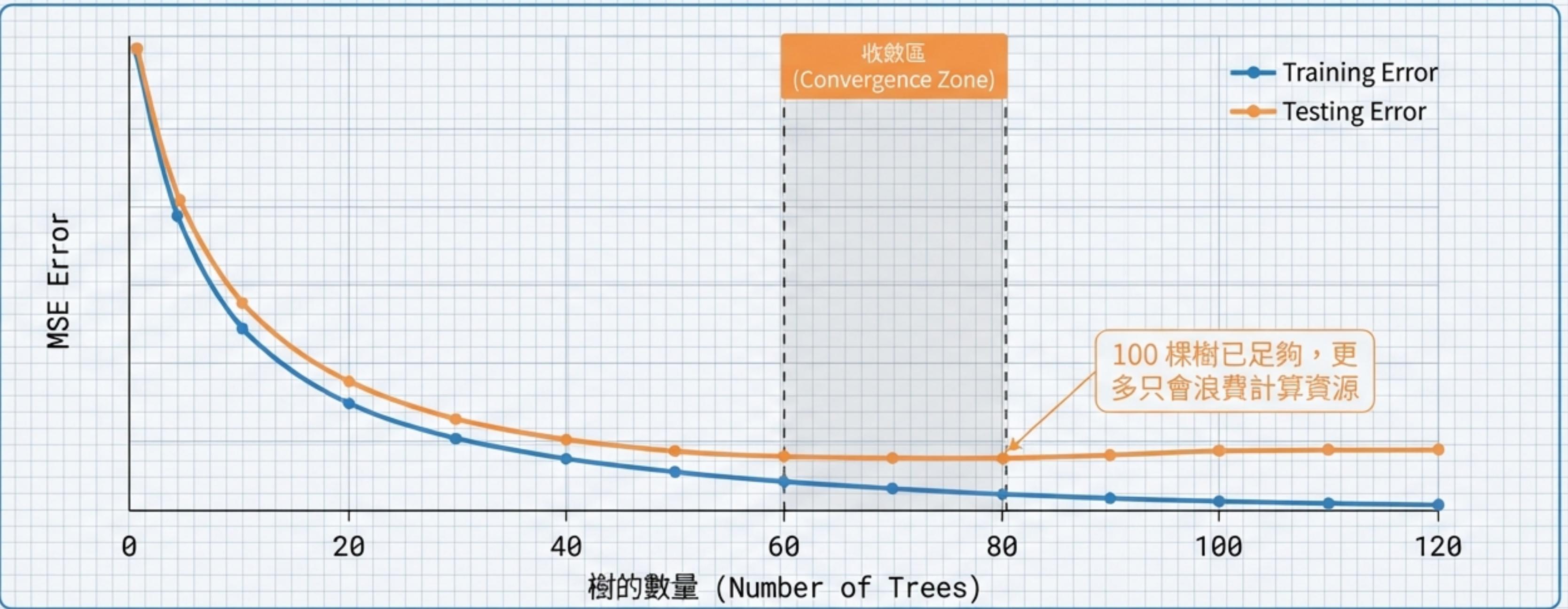
性能對決：GBT vs Random Forest



Insight: 為什麼 GBT 更快？

因為它使用淺樹 (Depth 5 vs RF Depth 12) 且樹的數量較少 (100 vs 300)，模型結構更精簡。

學習曲線分析：收斂與效率



Takeaway:

透過早停機制 (Early Stopping)，我們可以在最佳點自動停止，無需猜測樹的數量。

關鍵變數識別：特徵重要性 (Feature Importance)

Temperature (溫度)

Roboto Mono Noto Sans TC



41.4% (RF: 38.2%) - **主導因素**

Pressure (壓力)

Noto Sans TC



35.7% (RF: 33.5%)

Catalyst (催化劑)

Noto Sans TC



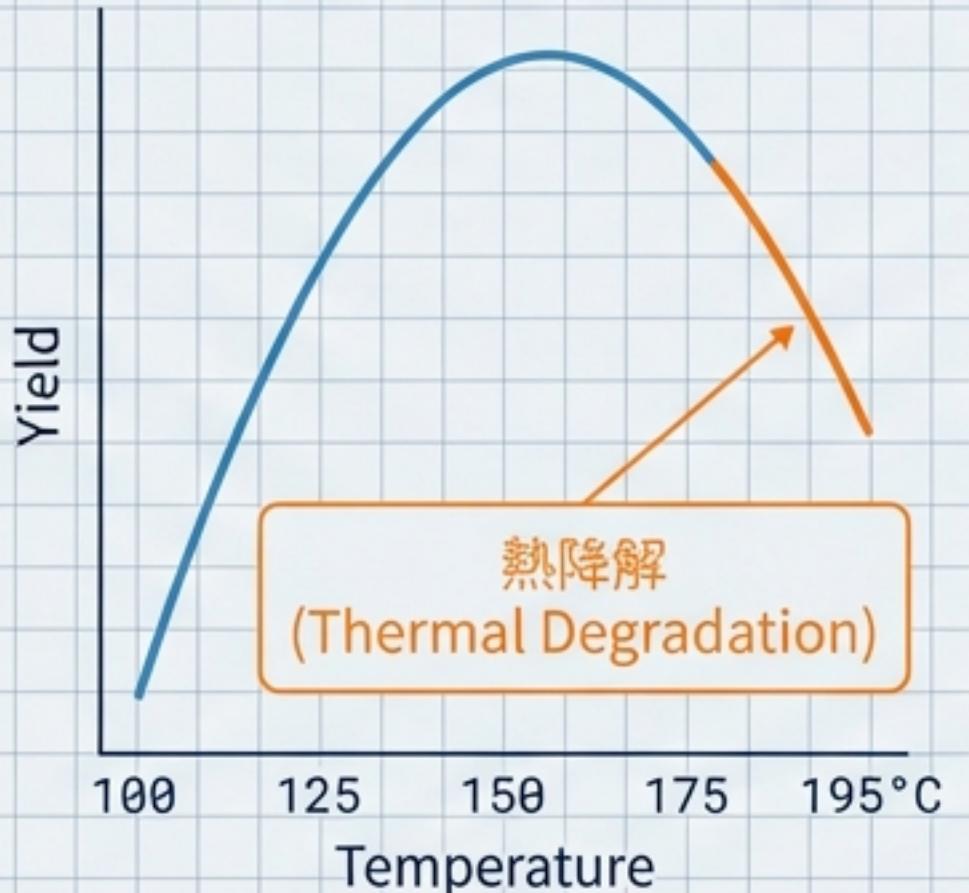
22.9% (RF: 28.3%)

物理意義解析：

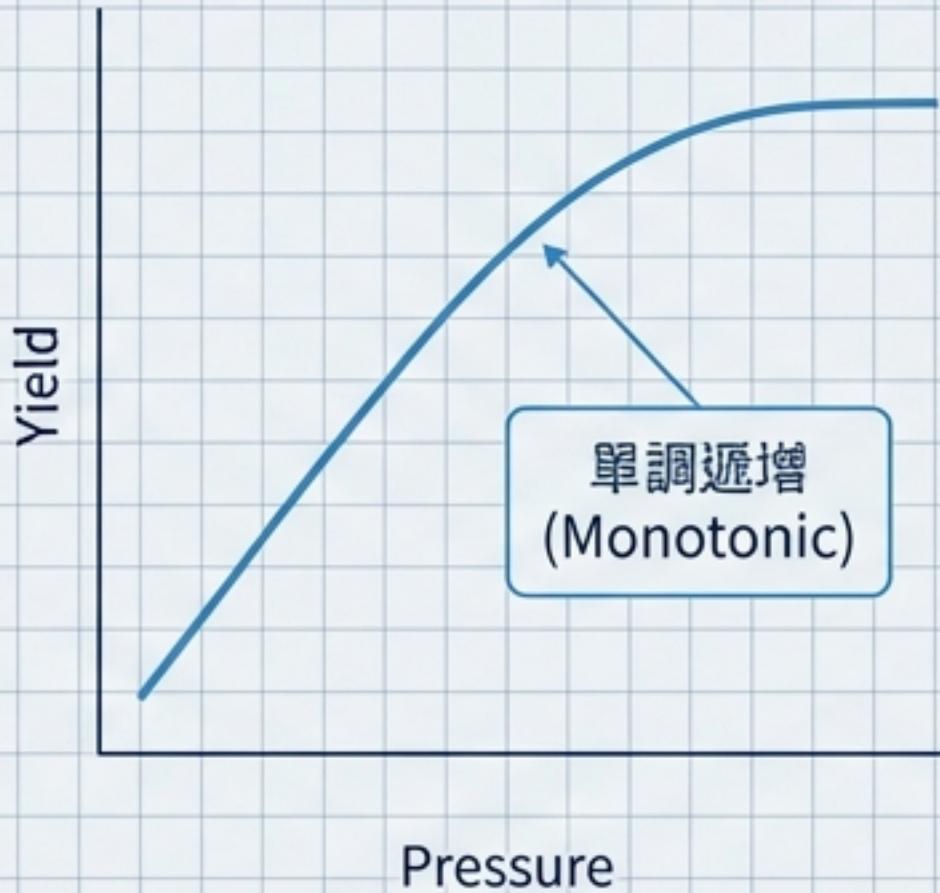
- GBT 權重分佈比 RF 更集中。
- 反應動力學中，溫度 (Arrhenius equation) 和壓力通常是指數級級影響，GBT 成功捕捉到了這種主導效應。

打開黑盒子：部分依賴圖 (Partial Dependence)

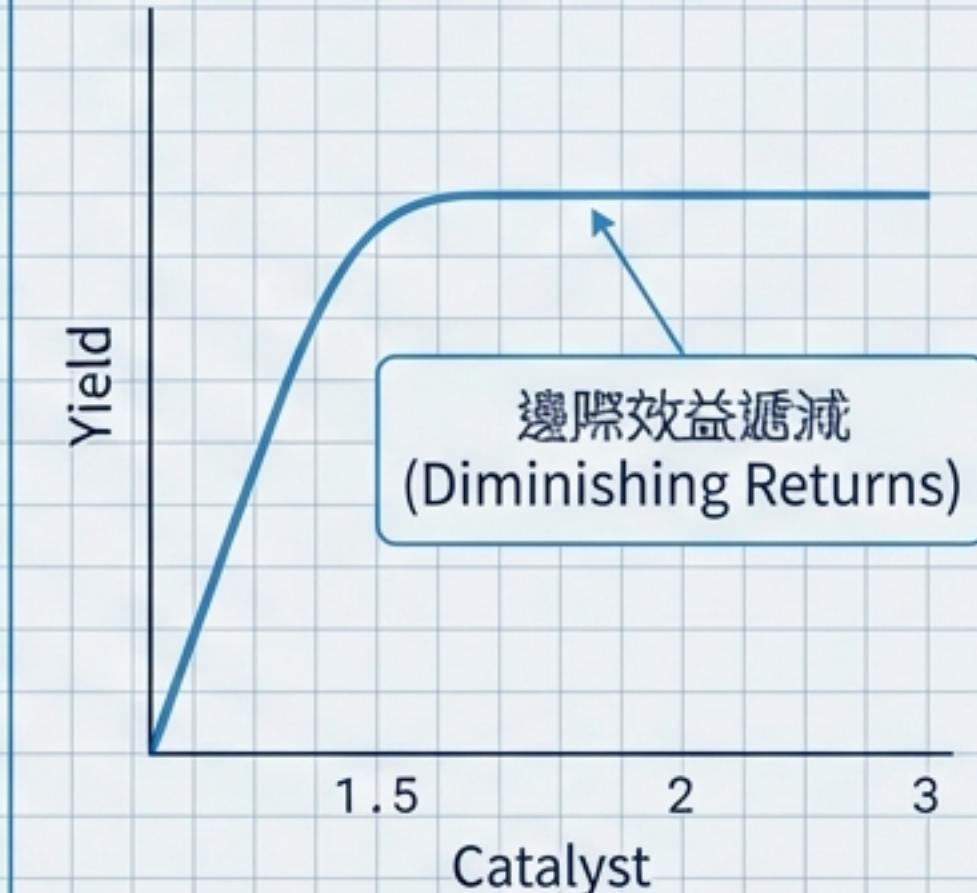
Temperature Effect



Pressure Effect



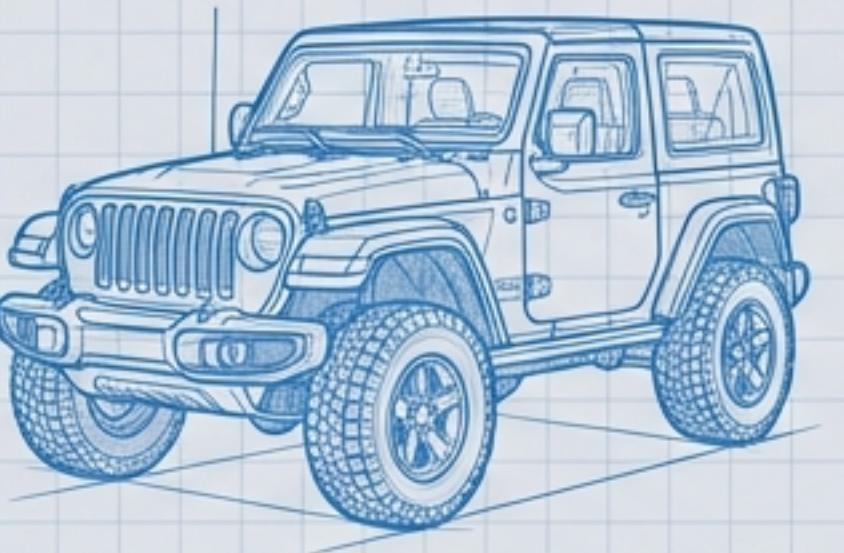
Catalyst Effect



結論：模型不僅預測準確，而且學到了正確的物理化學現象 (Physical Consistency)。

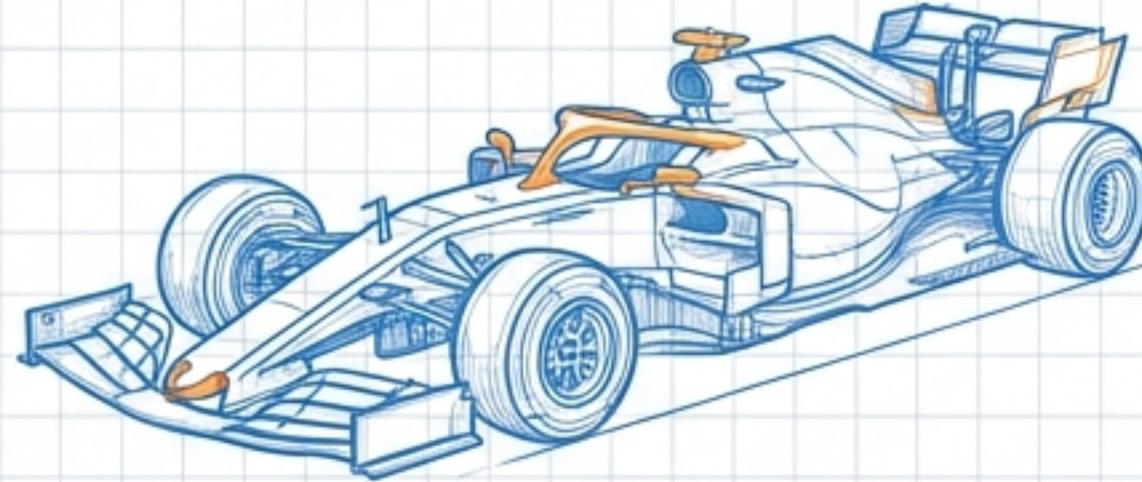
決策指南：何時選擇梯度提升樹？

選擇 Random Forest



- 數據充滿噪聲或異常值 (Outliers)
- 需要快速原型開發
- 擔心過度擬合，不想花時間調參
- 適合：初步探索、含噪數據

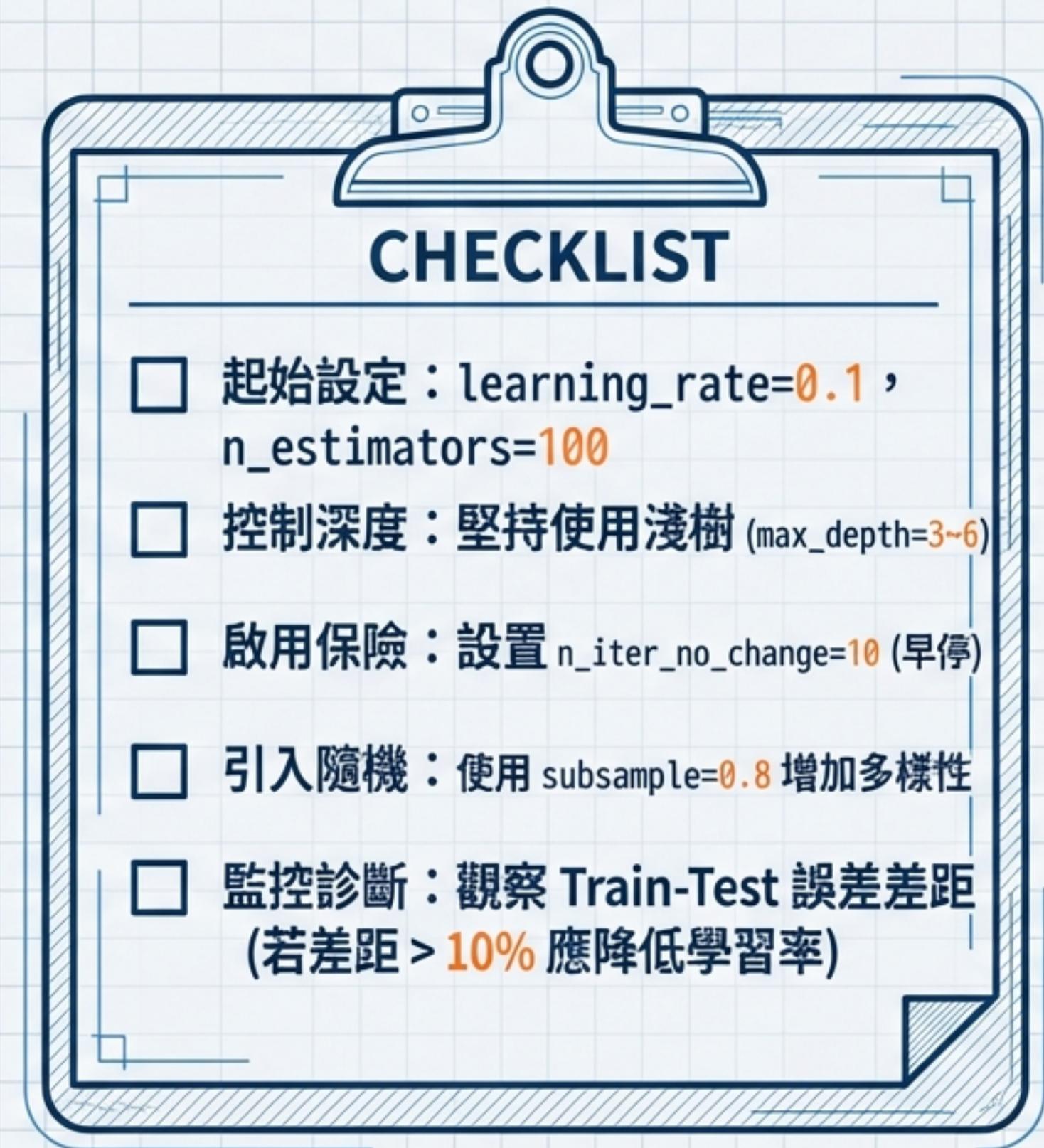
選擇 Gradient Boosting



- 追求**極致準確度** (Kaggle/High Value)
- 數據結構**清晰** (Structured Data)
- **有時間**進行超參數微調
- 適合：**製程優化**、精準預測

本案例結論：在反應器優化中，數據結構清晰且追求高產率，GBT 完勝。

工程師教戰守則：GBT 建模檢查清單



總結：序列學習帶來的精準躍升



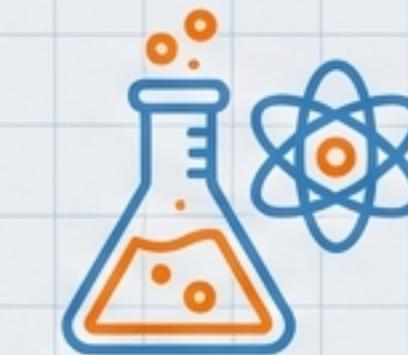
精準度 (Accuracy)

測試 R^2 提升 **6.6%** (**0.769 vs 0.722**)。誤差顯著降低。



效率 (Efficiency)

訓練速度更快 (**0.34s**)，模型體積更小 (淺層樹優勢)。



物理性 (Physics)

準確捕捉溫度倒 U 型與壓力非線性效應，符合化學原理。

Closing Thought: 隨機森林提供了很好的基準，但當你需要榨出最後 5% 的性能時，梯度提升樹是你的最佳武器。

Next Unit: 支持向量機 (SVM) →