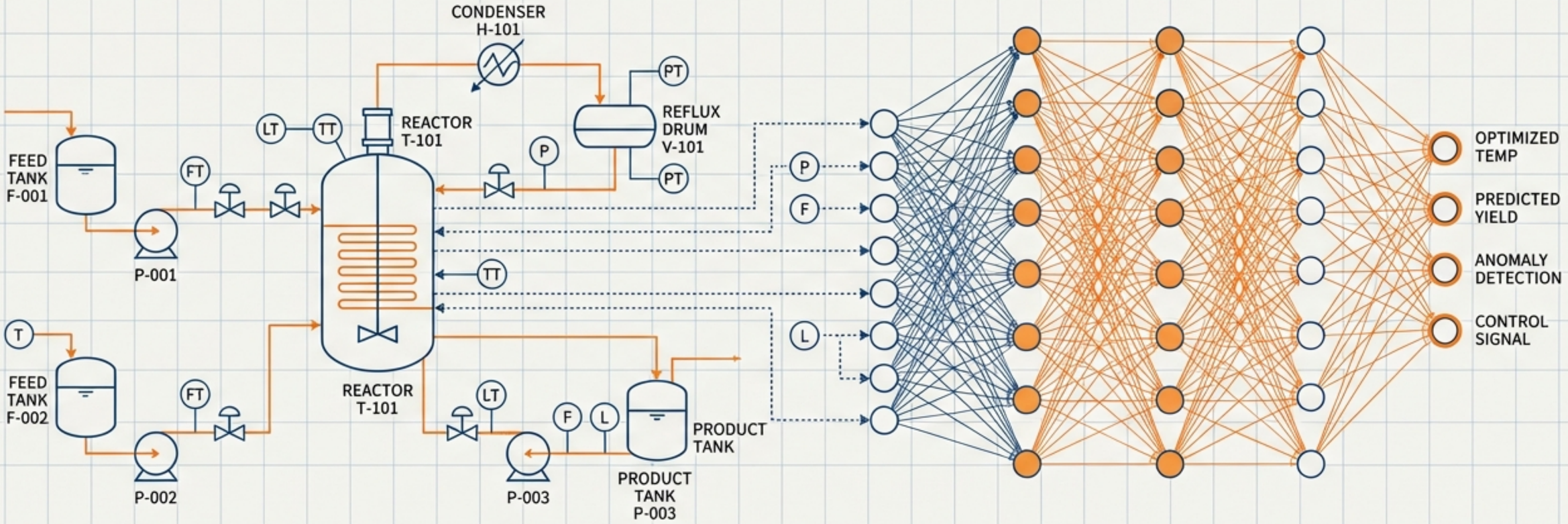
Unit 11 非線性模型回歸總覽 (Non-Linear Models Regression Overview)

Unit 11

AI 在化工上之應用 (AI Applications in Chemical Engineering)



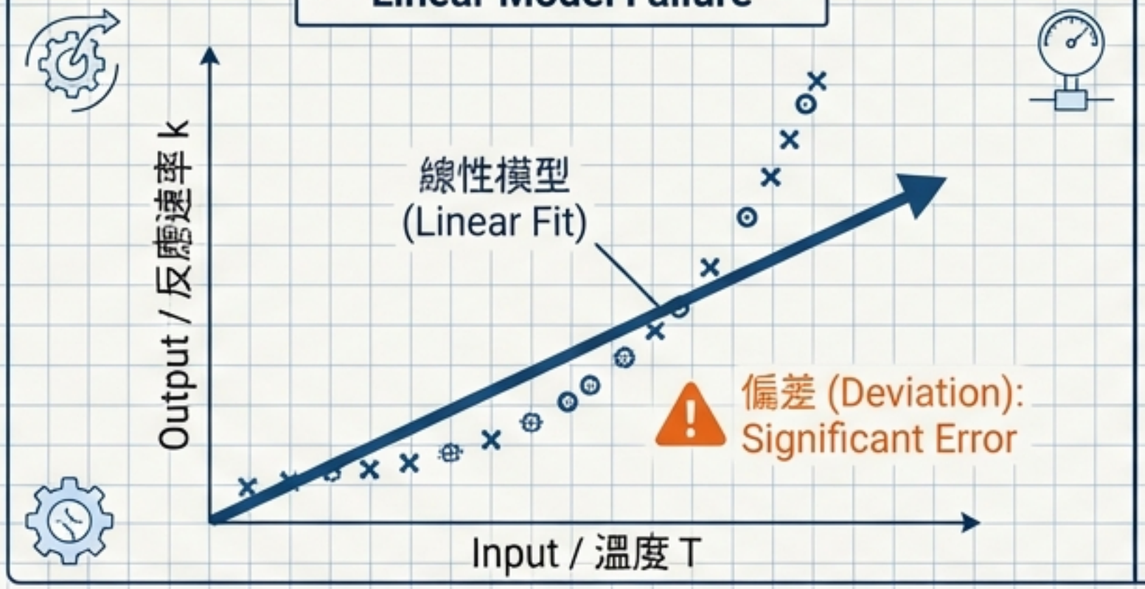
課程目標 (Course Goal)：建立人工智能與機器學習基礎，解決化工領域實際問題

製作單位 (Producer)：逢甲大學 化工系 智慧程序系統工程實驗室

現實世界的物理限制：為何我們需要非線性模型？

The Physics of Non-Linearity

Linear Model Failure



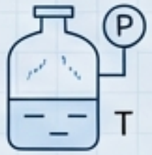
Non-Linear Model Success

Reaction Kinetics (Arrhenius)

$$k = Ae^{-\frac{E_a}{RT}}$$

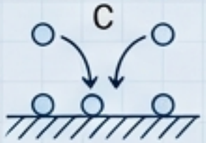
Phase Equilibrium (Antoine)

$$\log_{10} P = A - \frac{B}{C+T}$$



Adsorption (Langmuir)

$$q = \frac{q_{max} K C}{1 + KC}$$



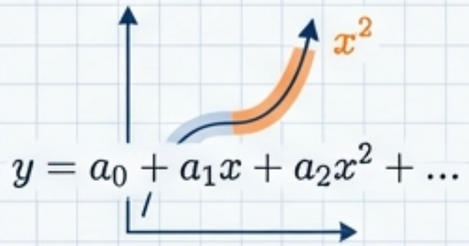
核心觀點：化工現象本質上是非線性的 (The chemical world is rarely linear).
Linear models fail when physics dictates complexity like saturation and interactions.

非線性模型工具箱地圖

The Non-Linear Model Landscape

Non-Linear Models

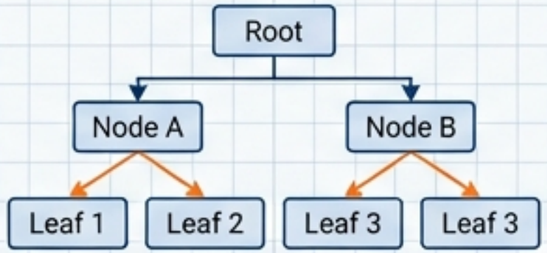
1. 參數擴展型 (Parameter Expansion)



Mechanism: Extending feature space (增加特徵高次項)

Representative:
- 多項式回歸 (Polynomial Regression)

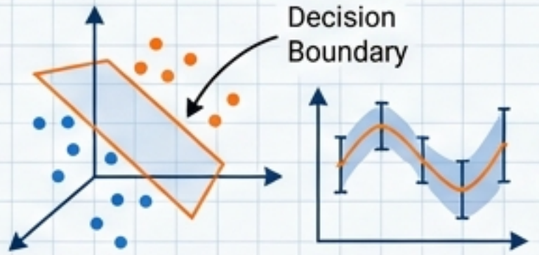
2. 樹狀模型 (Tree-based Models)



Mechanism: Step-wise logic rules (分層決策)

Representatives:
- 決策樹 (Decision Tree)
- 梯度提升樹 (Gradient Boosting / XGBoost)

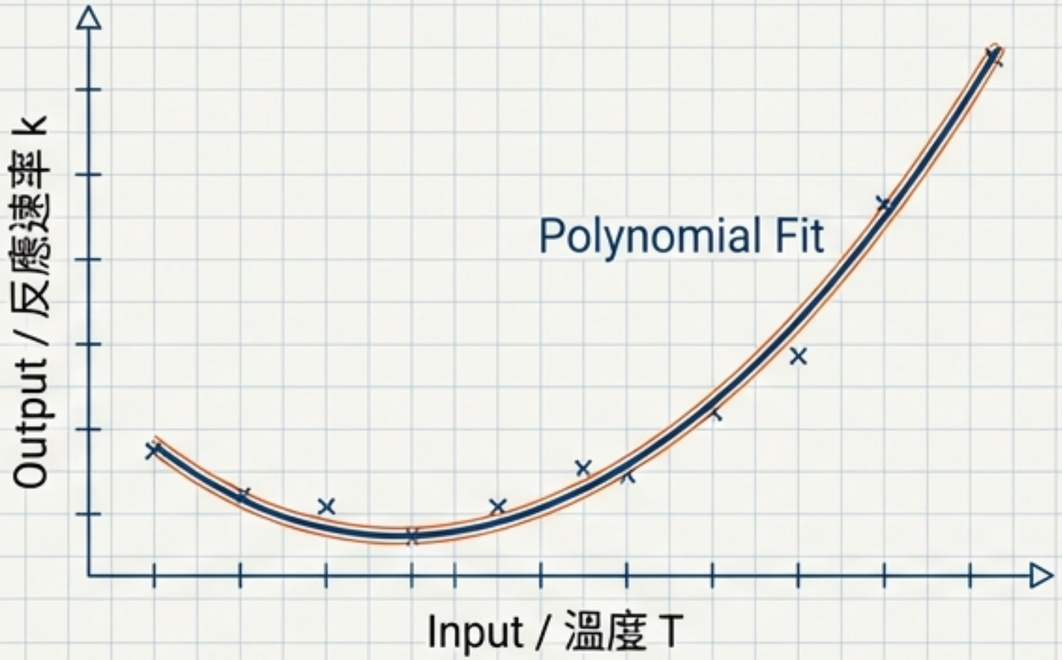
3. 核方法與機率模型 (Kernel & Probabilistic)



Mechanism: Mapping to high dimensions or Bayesian inference

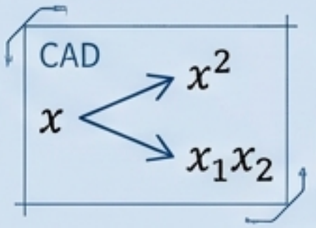
Representatives:
- 支持向量機 (SVM)
- 高斯過程回歸 (GPR)

多項式回歸 (Polynomial Regression)：線性模型的延伸



核心機制 (Mechanism)

- 本質上仍是線性模型 (Linear in parameters)
- 透過 `PolynomialFeatures` 創造交互項 (Interaction) 與高次項



Formula Transformation:

From: $\hat{y} = w_0 + w_1 x$ (Linear)

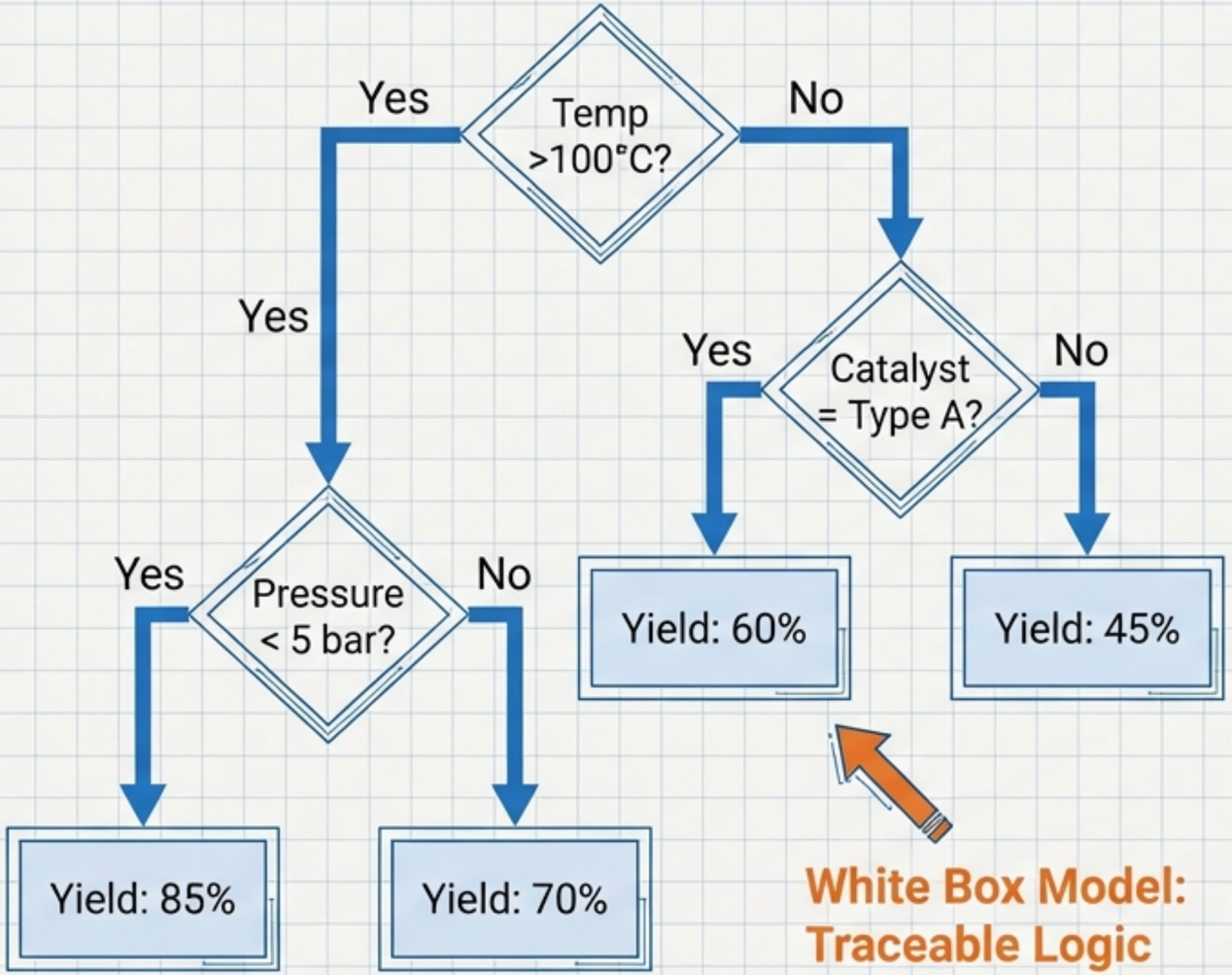
To: $\hat{y} = w_0 + w_1 x + w_2 x^2 + w_3 x_1 x_2$ (Polynomial)

優缺點 (Pros & Cons)

Pros:	✓ 簡單 (Simple) ✓ 可解釋 (Interpretable) ✓ 適合平滑曲線 (Smooth curves)
Cons:	⚠ 高次項易過擬合 (Overfitting) ⚠ 外推能力差 (Poor extrapolation)

ChemE Use Case: Sensor calibration (感測器校正) - Physics suggests smooth quadratic relationships.

決策樹 (Decision Tree)：模擬工程師的邏輯思維



核心邏輯 (Core Logic)

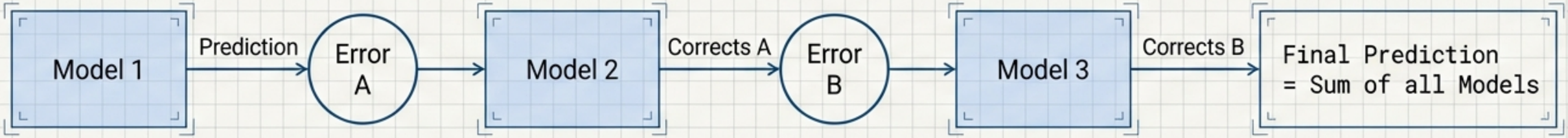
- Splitting Strategy: Minimizing MSE (均方誤差)
- Structure: Recursive partitioning (遞迴分割)

優缺點 (Pros & Cons)

✔ Pros: 高度可解釋 (White Box)，無需特徵縮放，自動處理交互作用

⚠ Cons: 不穩定 (Unstable)，容易過擬合，預測為階梯函數

梯度提升樹 (Gradient Boosting)：工業界的準確度冠軍



機制與變體 (Mechanism & Variants)

Boosting: Learning from residuals (從殘差中學習)

Optimization: Gradient Descent in function space

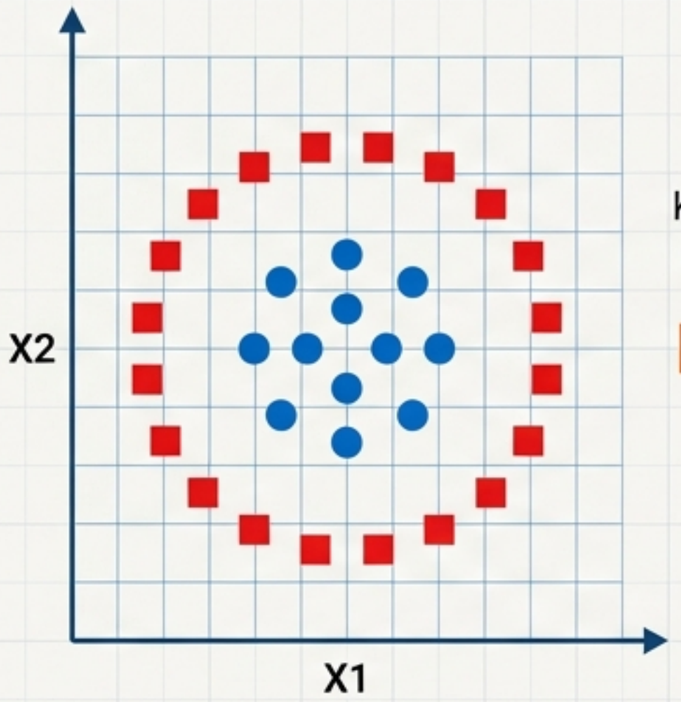
Tools: XGBoost, LightGBM, CatBoost

優缺點 (Pros & Cons)

Pros: 極高準確度 (Kaggle winner)，穩健 (Robust)

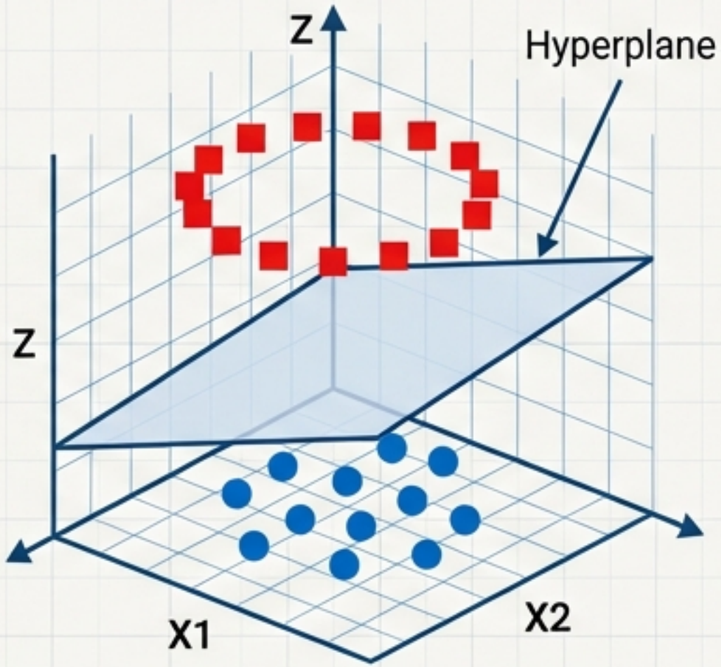
Cons: 黑盒 (Black Box)，訓練慢，參數眾多

支持向量機 (SVM)：高維空間的邊界尋找



Non-separable in 2D

Kernel Mapping Function $\phi(x)$



Separable in 3D

關鍵概念 (Core Concepts)

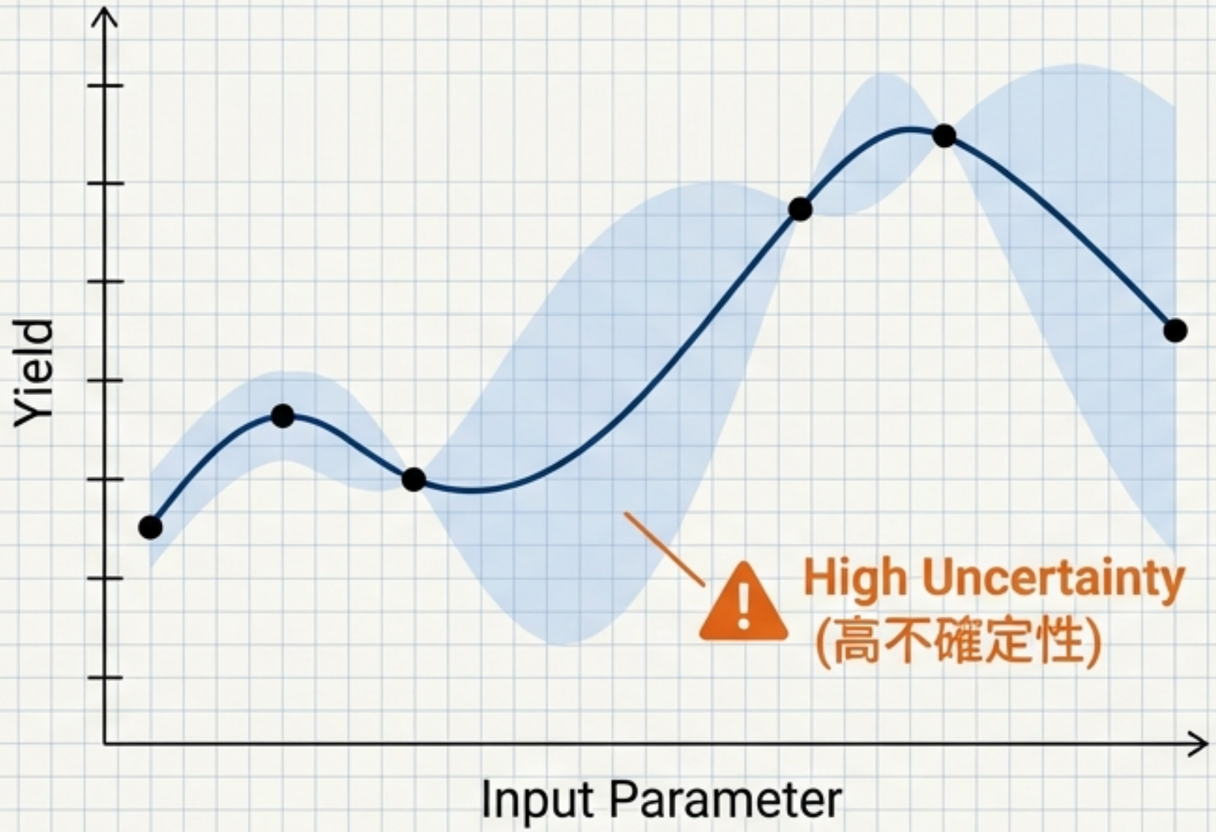
Kernel Function (核函數): $K(x_i, x_j)$ (e.g., RBF)

Support Vectors: Only points near boundary matter

重要前提 (Requirement): Feature Scaling is mandatory (特徵縮放是必須的). SVM relies on distance calculations.

ChemE Use Case: High-dimensional QSAR/QSPR (Molecule property prediction).

高斯過程回歸 (GPR)：不僅預測值，更預測風險



不確定性量化 (Uncertainty Quantification)	優缺點 (Pros & Cons)
• Prediction = Mean (μ) + Variance (σ^2) • Mechanism: Bayesian Inference (貝氏推論)	✅ Pros: 適合小數據 (Small Data), 支援貝氏優化 ⚠️ Cons: 計算成本極高 ($O(N^3)$)

Strategic Use: Essential for Safety and Experimental Design (Bayesian Optimization).

模型選擇指南：工程師的決策矩陣 (Model Selection Matrix)

模型 (Model)	準確度 (Accuracy)	可解釋性 (Interpretability)	速度 (Speed)	不確定性 (Uncertainty)	資料量 (Data Size)
Polynomial	★★	★★★★	★★★★★	✕	Medium
Decision Tree	★★★	★★★★★	★★★★★	✕	Small/Med.
XGBoost (GBM)	★★★★★	★★	★★	✕	Large
SVM	★★★★★	★	★★	✕	Small/Med.
GPR	★★★★★	★	★	★★★★★	Very Small

關鍵前處理：特徵工程與縮放 (Preprocessing & Feature Engineering)

原始數據
(Raw Data)

- Sensor_T1
- Flow_Rate

工程特徵
(Engineered Features)

- Reynolds Number (Re)
- Reaction Rate (k)
  (= A * exp(-E/RT))

更好的模型
(Better Model)

- Higher Accuracy + Interpretability

Scaling (縮放)

- **Critical for:** SVM, GPR (Distance-based)
- Not needed for: Trees, XGBoost
- Method: Standardization (**z-score**)

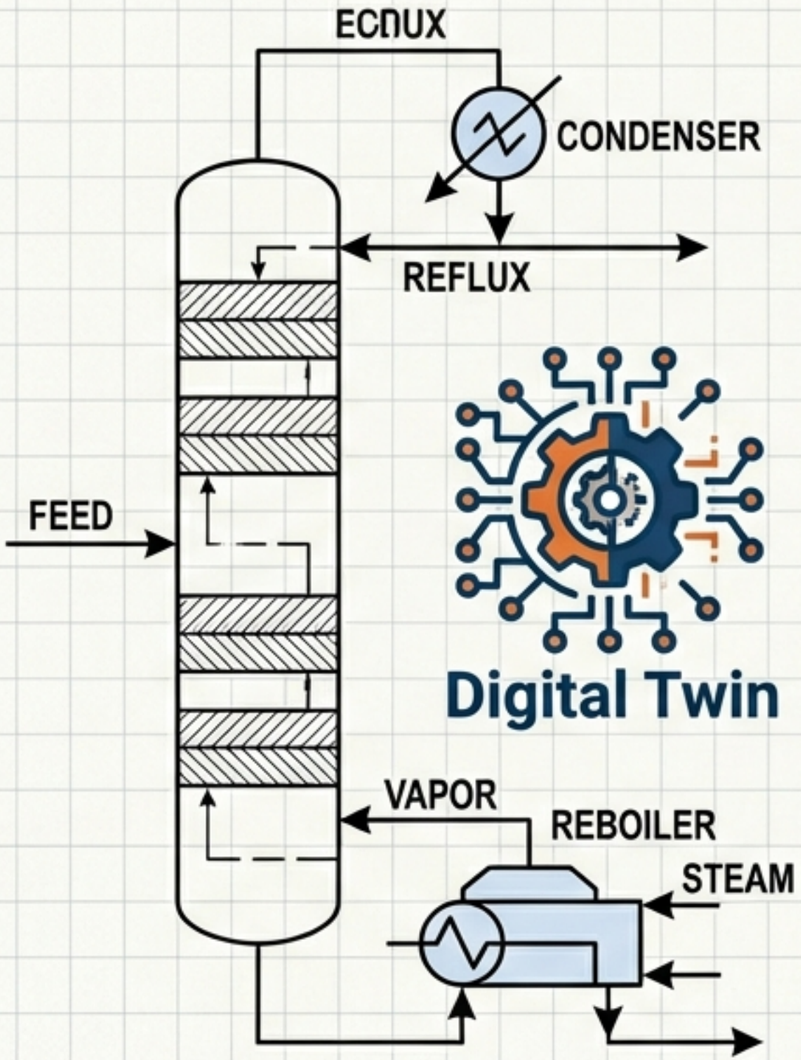
Encoding (編碼)

- One-Hot Encoding:
- For categorical data (e.g., Reactor ID)

Feature Engineering

- **Domain Knowledge** is key.
- Input physics-based parameters over raw sensor data.

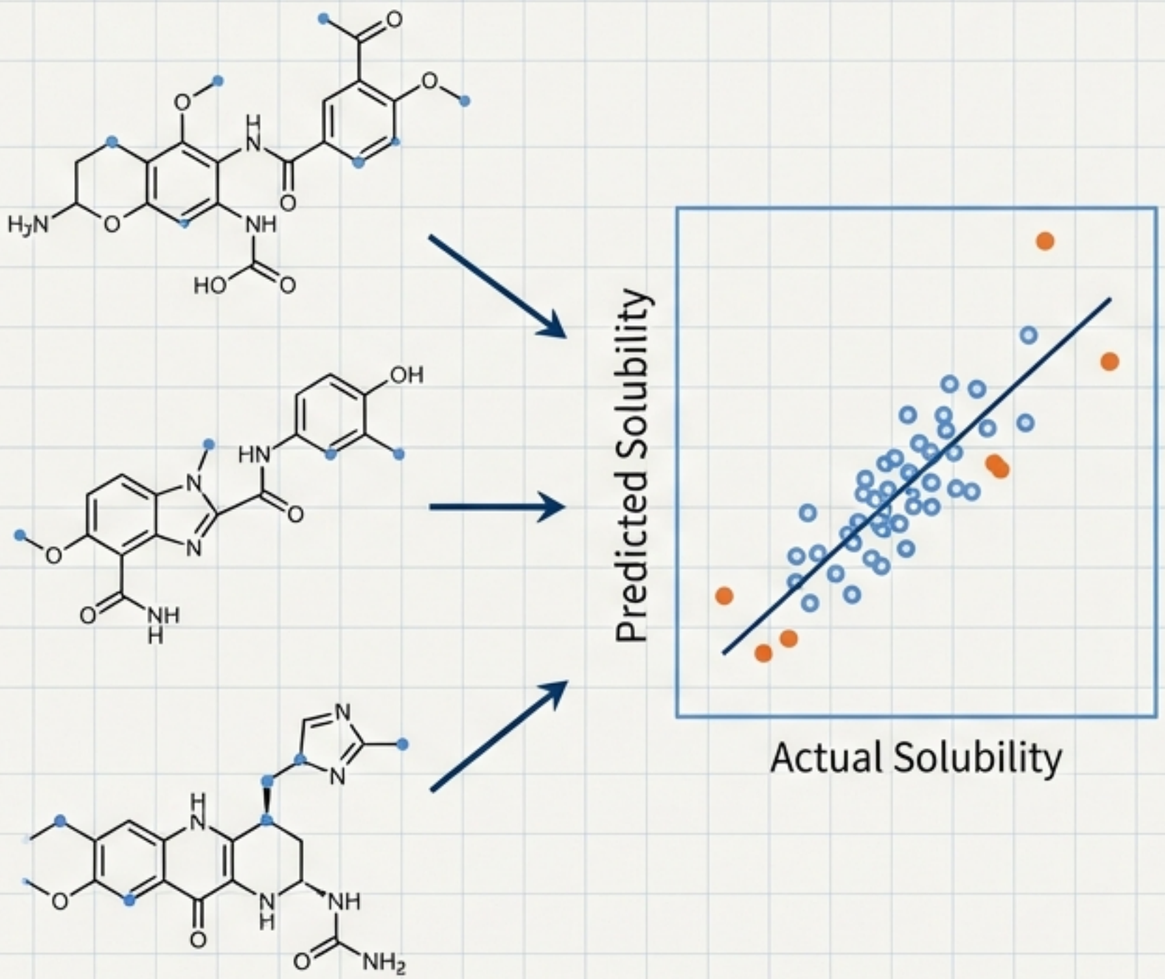
實戰案例 1：蒸餾塔操作建模 (Distillation Column Modeling)



Project Card (Noto Sans TC)
Problem: Predicting top product purity (塔頂產品純度) for control.
Data: Flow rates, Reflux ratio, Temperatures (Time-series).
Challenge: Dynamic process with time delays (時間滯後).
Solution: XGBoost with lag features (t-1, t-2).

- **Accuracy**: $R^2 = 0.97$
- **Impact**: Prediction 15 mins ahead (提前 15 分鐘預測)
- **Reduced** purity fluctuation by **40%**

實戰案例 2：藥物溶解度預測 (Drug Solubility Prediction)



Project Card

Problem
Screening drug candidates for water solubility.

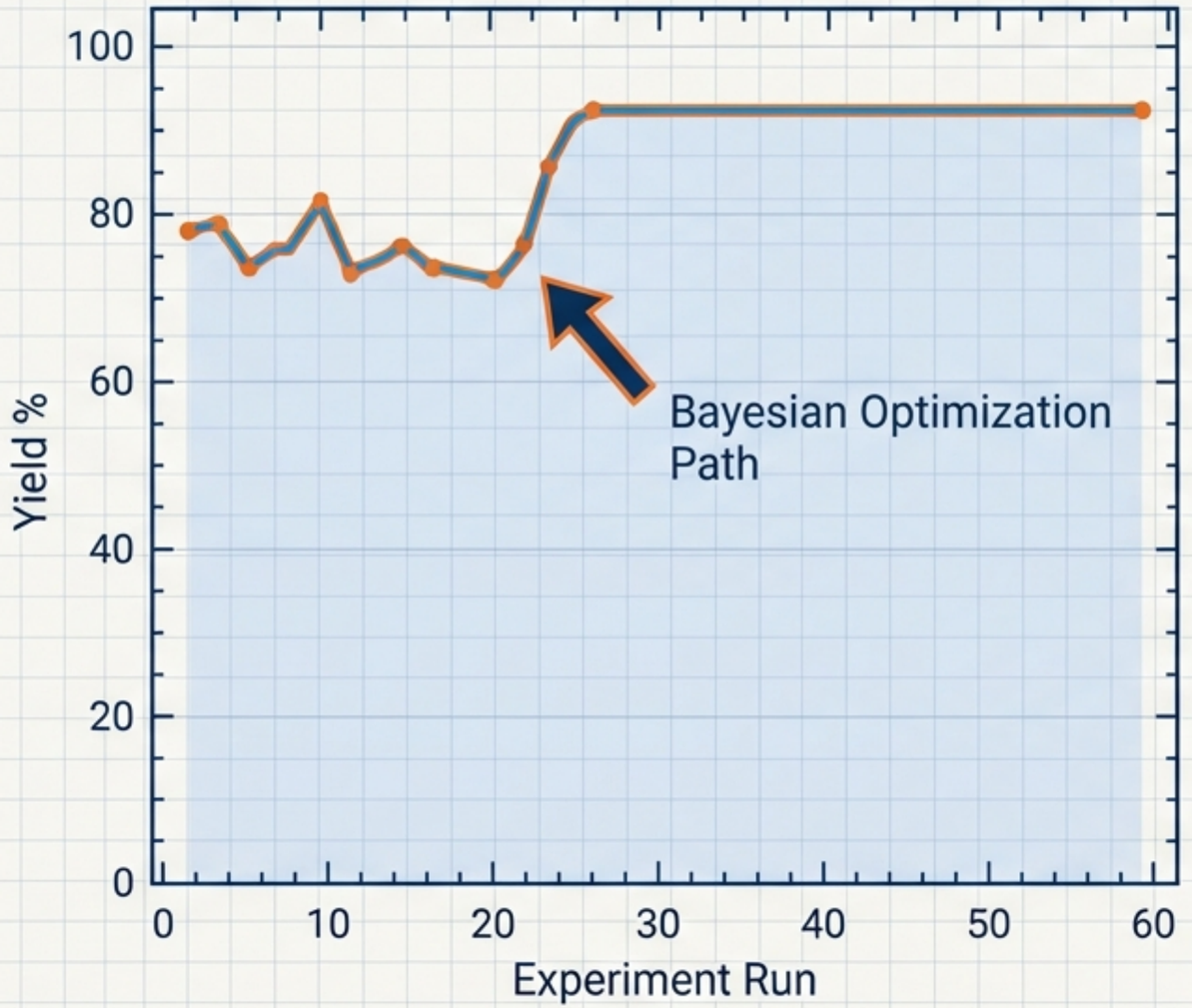
Data
Molecular descriptors (Weight, LogP, H-bonds).
High dimension (>200 features).

Challenge
Complex structure-property relationship.

Solution
SVM (RBF Kernel) + Lasso Feature Selection.

Results
Accuracy: $R^2 = 0.89$
Key Insight: LogP and TPSA identified as critical factors.
Impact: Reduced candidates by 80% (Virtual Screening)

實戰案例 3：反應器貝氏優化 (Reactor Bayesian Optimization)



Project Card

Problem:
Optimizing temperature profile for batch reactor.

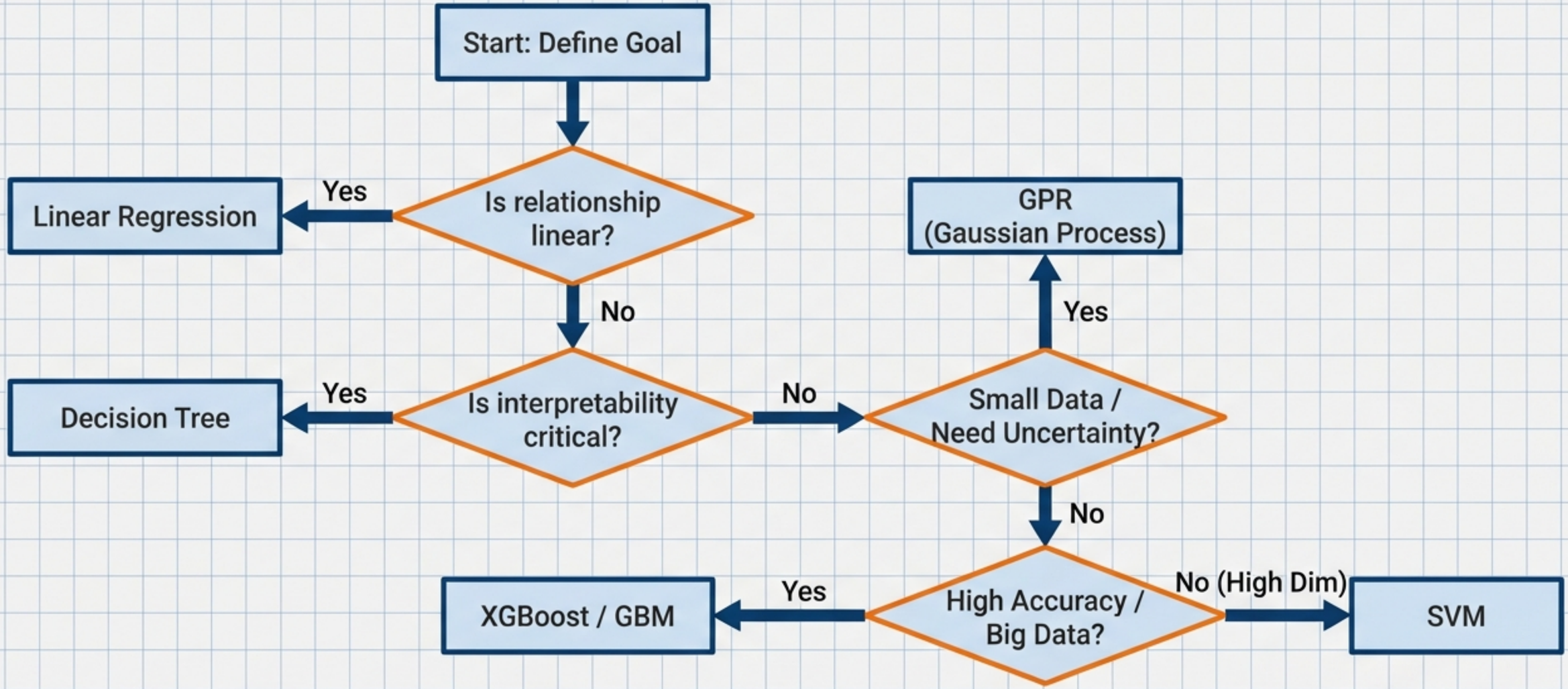
Constraint:
Experiments are very expensive (High cost).
Small data only.

Solution:
Gaussian Process Regression (GPR) + Expected Improvement.

Results

- **Efficiency:** Found optimum in 50 runs (vs. random search).
- **Yield:** Increased from 78% -> 92%.
- **Key Value:** Active Learning minimized waste.

決策流程：如何選擇正確的模型？ (Decision Framework)



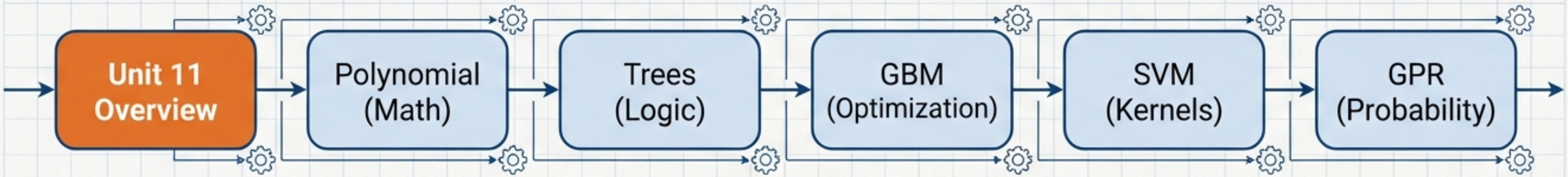
No Free Lunch Theorem: Data understanding dictates model choice (資料特性決定模型選擇).

結語與學習路徑 (Summary & Next Steps)

✓ **Gradient Boosting** is the industrial workhorse (工業首選).

✓ **GPR** is vital for experimental design (實驗設計).

✓ **Scaling** makes or breaks distance-based models (SVM/GPR).



Mastering these non-linear tools bridges the gap between theory and real-world chemical complexity.