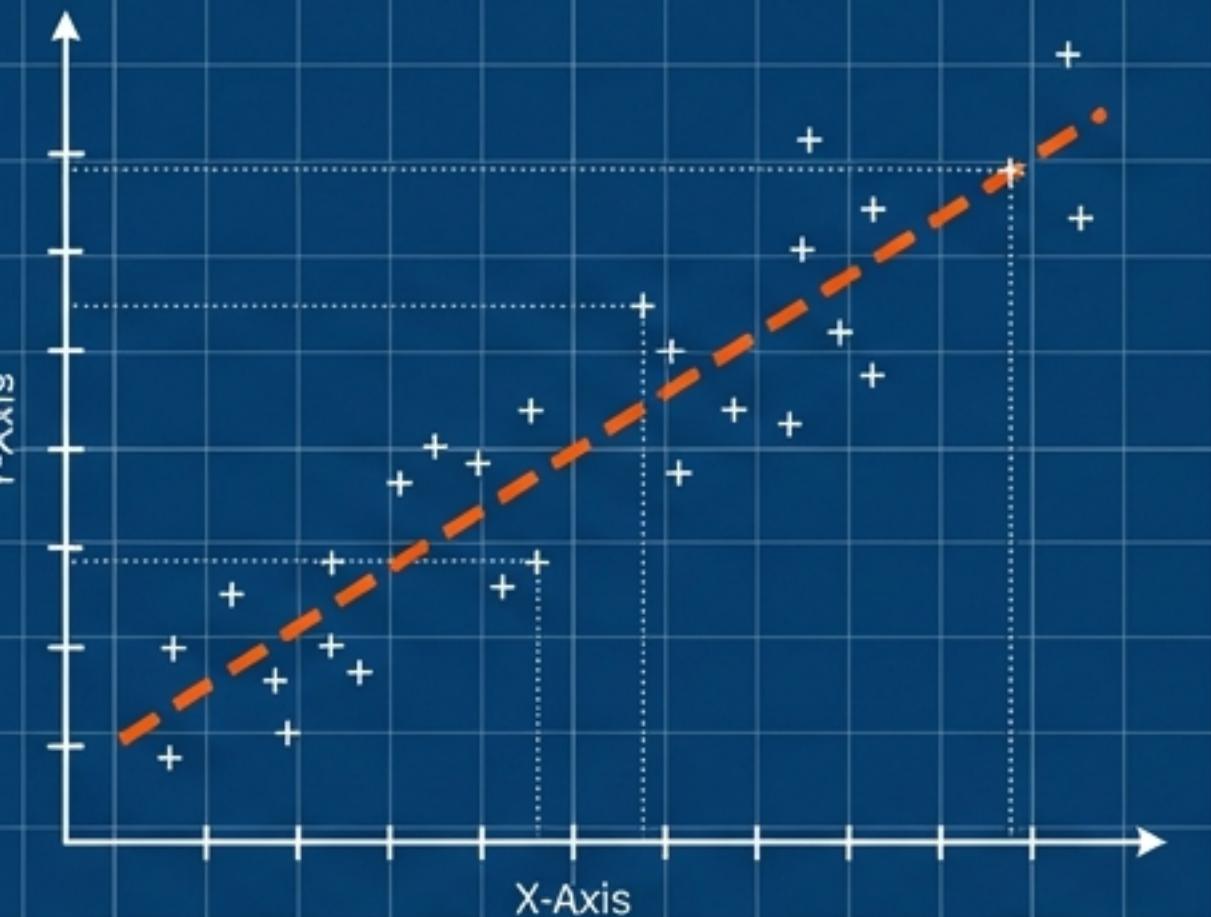


# Unit 10 線性模型回歸總覽

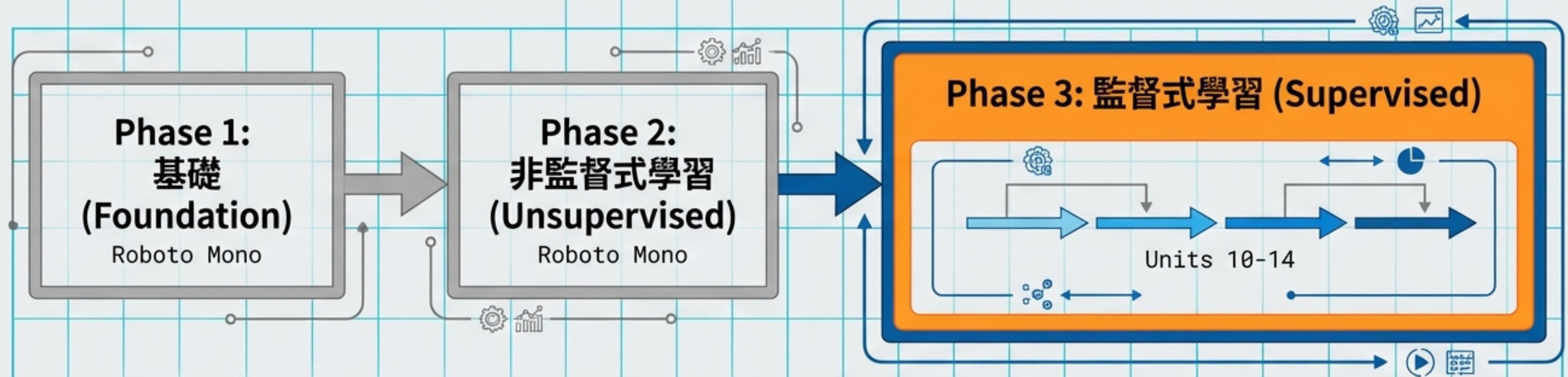
Linear Models Regression Overview



從基礎理論到化工實務應用

授課教師：莊曜禎 助理教授  
課程：AI在化工上之應用

智慧程序系統工程實驗室  
2026-01-28

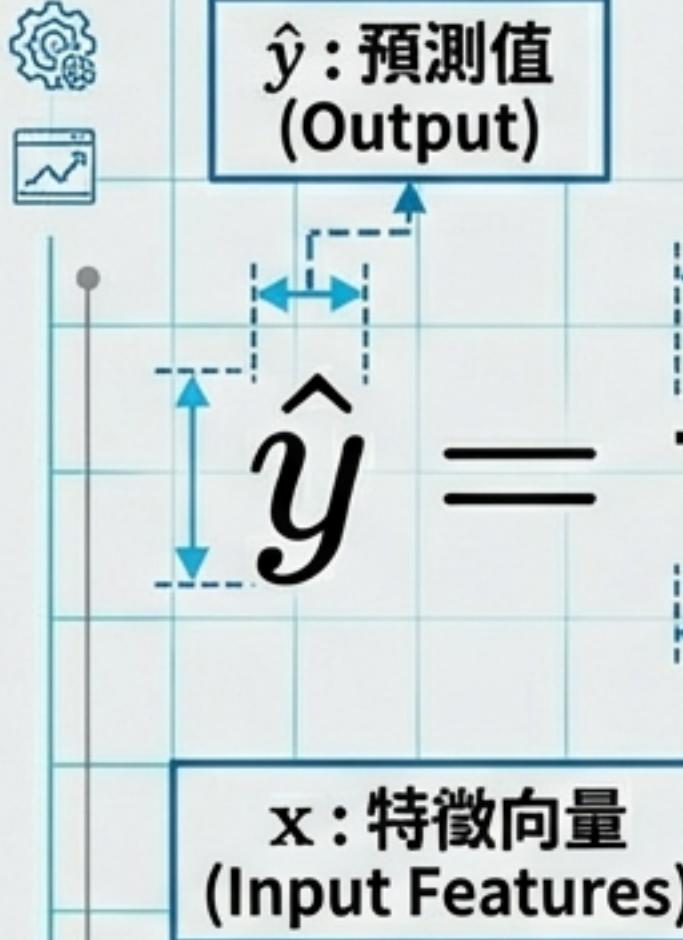


# 本單元課程目標 (Course Goals)

- 目標：理解線性模型原理、掌握 Scikit-learn 工具、應用於化工數據。
- 關鍵模組 (Key Modules) :
- 關鍵模組 (Key Modules) :
  1. 基礎理論 (Theory)
  2. sklearn 模型庫 (The Toolbox)
  3. 數據前處理 (Preprocessing)
  4. 模型評估 (Evaluation)
  5. 化工實例 (Case Studies)

# Noto Sans TC 數學原理 (First Principles)

Roboto Mono

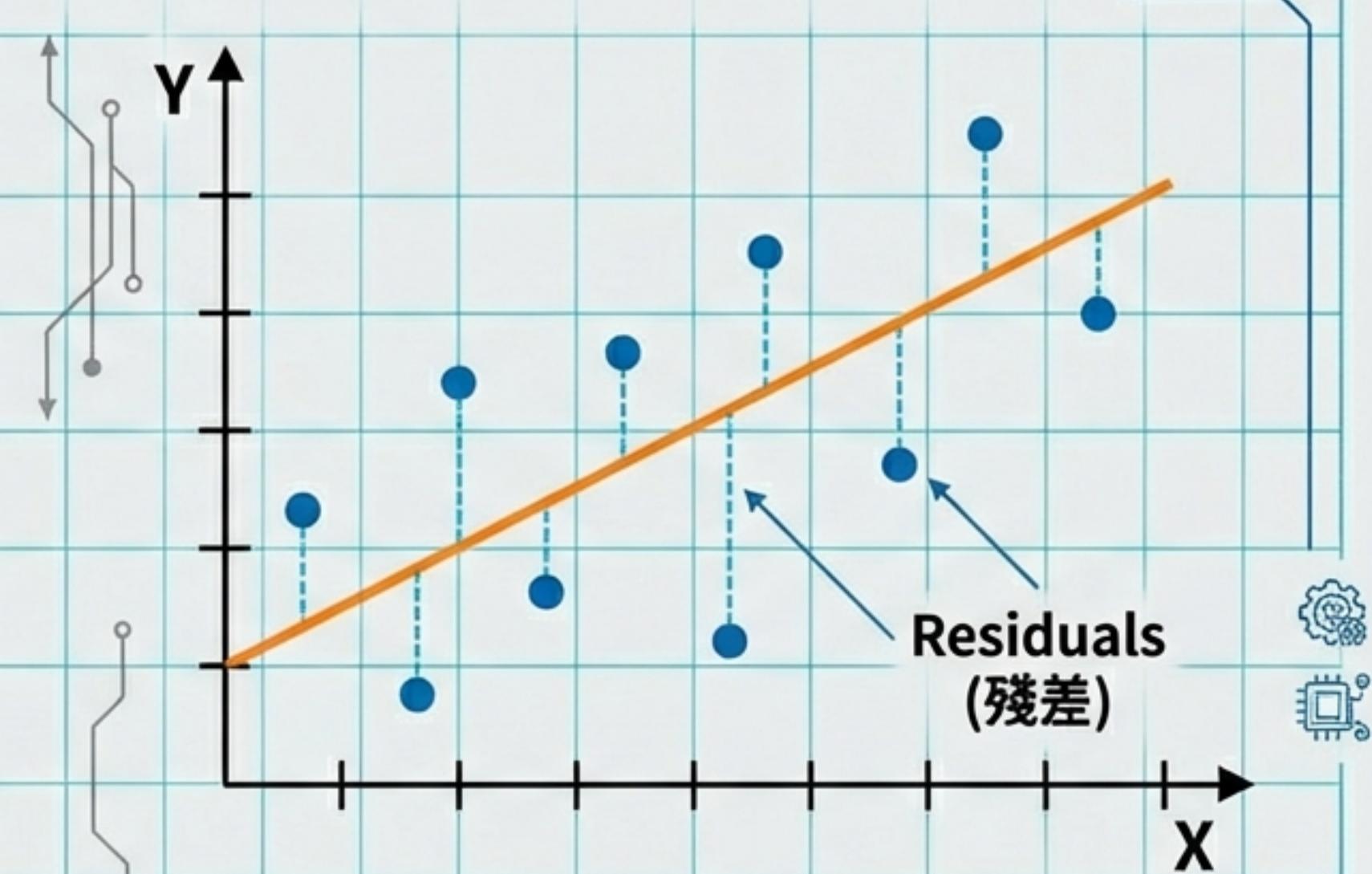


Noto Sans TC Roboto Mono

**Objective Function** Loss = Minimize MSE  $\rightarrow \min \frac{1}{m} \sum (y - \hat{y})^2$

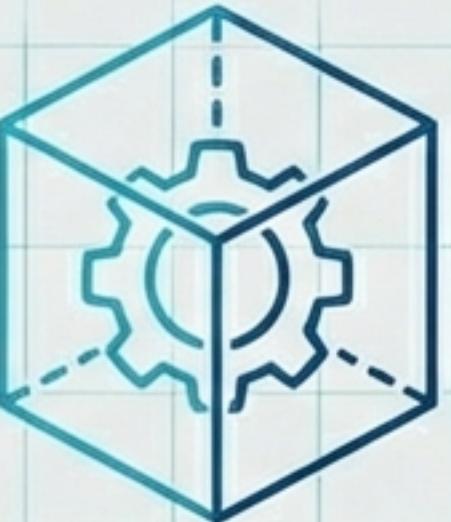
# Noto Sans TC 視覺化 (Visualization)

Roboto Mono



# 為什麼選擇線性模型？(Why Linear Models?)

Roboto Mono



## 可解釋性強 (Interpretability)

- 權重 ( $w$ ) 直接反映特徵對結果的影響。
- 例如：溫度每上升 1 度，產率增加多少。



## 計算效率高 (Efficiency)

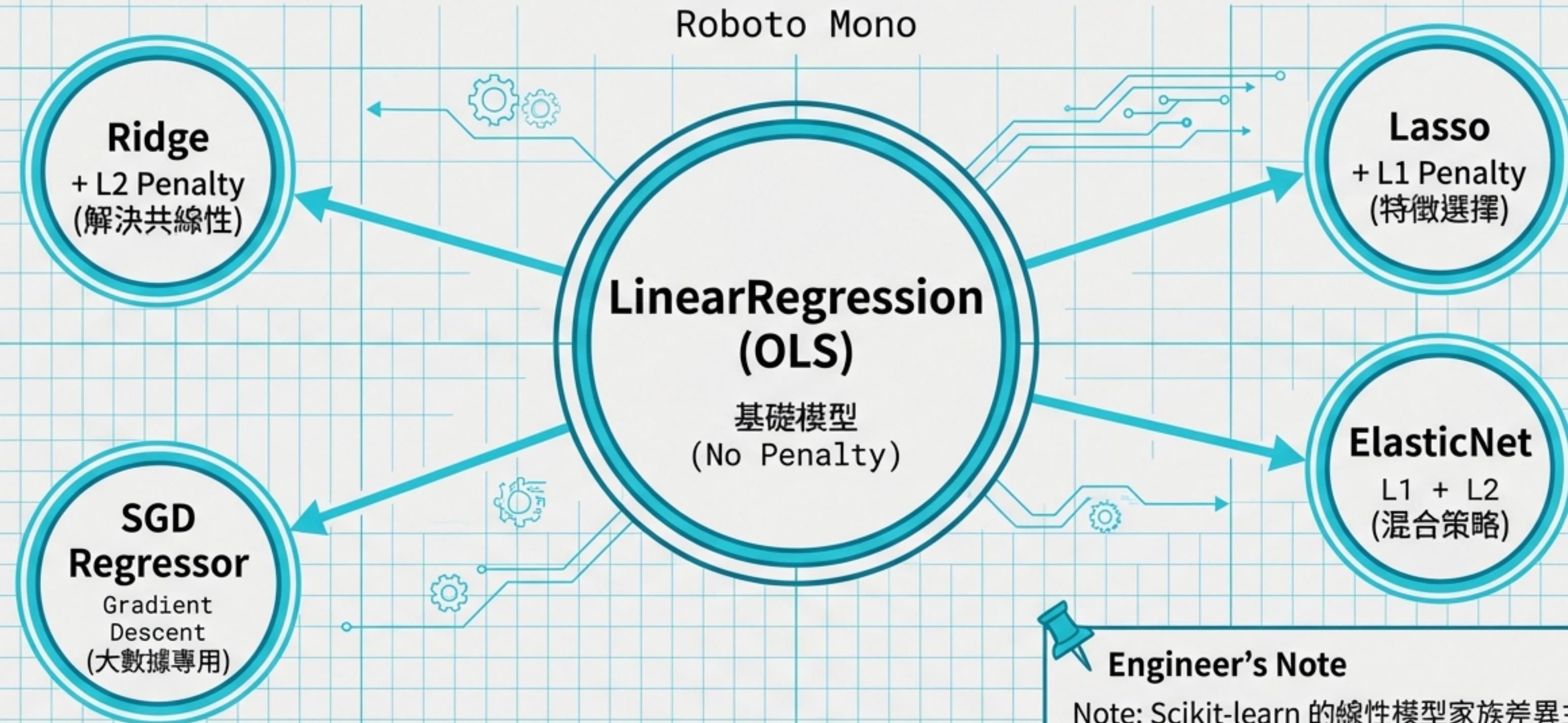
- 訓練速度快，適合大規模化工數據。
- 不易過擬合 (泛化能力好)。



## 基本假設 (Assumptions)

- 線性關係 (Linearity)
- 獨立性 (Independence)
- 常態分佈 (Normality)
- 無多重共線性 (No Multicollinearity)

# sklearn 線性模型家族 (The Model Family)



## Engineer's Note

Note: Scikit-learn 的線性模型家族差異主要在於「正則化技術」(Regularization)。

# 正則化技術：Ridge Regression (嶺回歸)

## 穩定控制系統 (Stabilization)

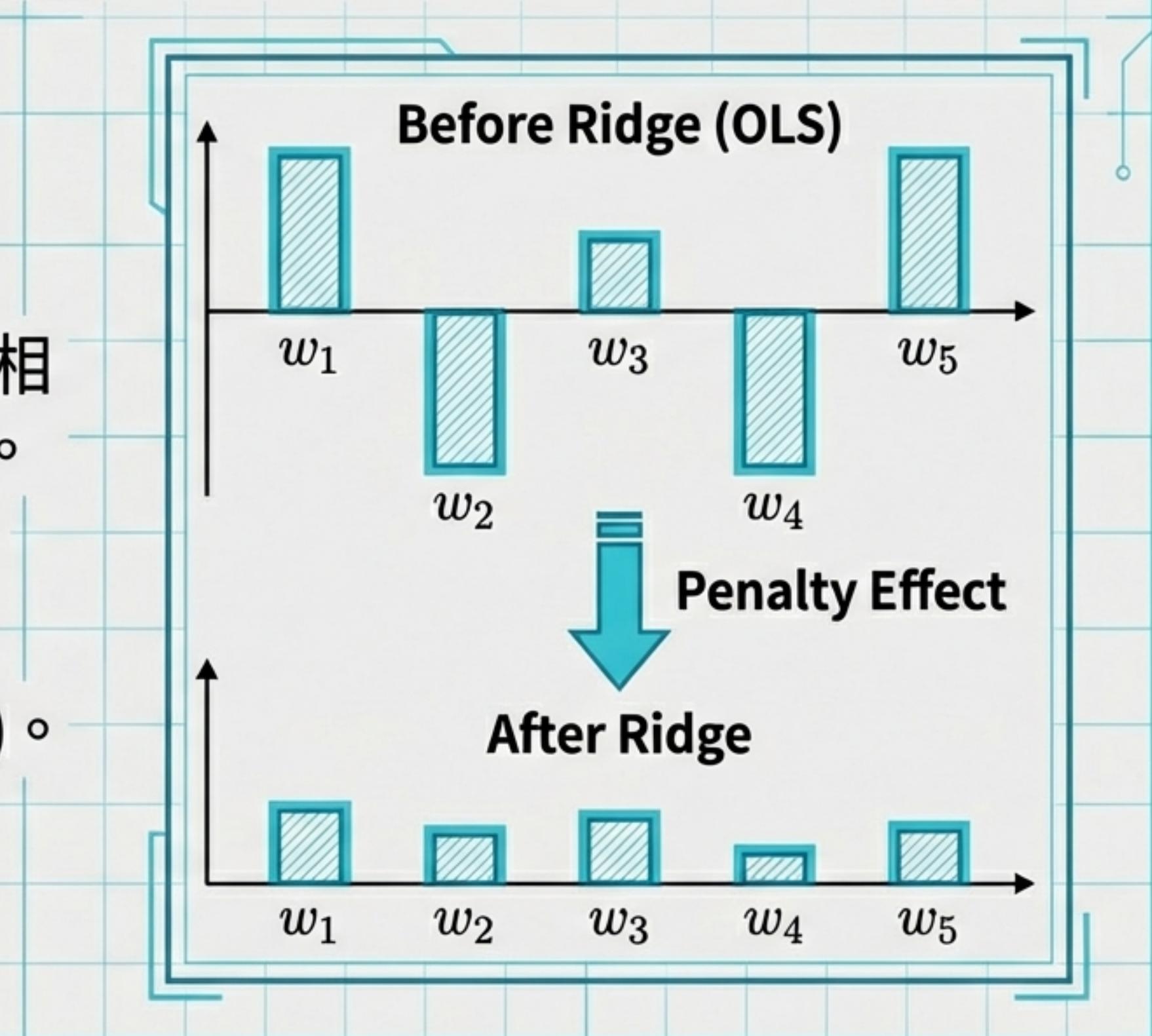
**問題 (Problem)**：多重共線性 (Multicollinearity)。

當輸入變數 (如溫度 T 與壓力 P) 高度相關時，模型權重會劇烈震盪且不穩定。

**解法 (Solution)**：L2 Regularization

$$\text{Loss} + \alpha \sum w^2$$

**效果**：壓縮權重但不歸零 (Shrinkage)。



# 特徵工程師：Lasso Regression 過濾雜訊訊號 (Signal Filtration)

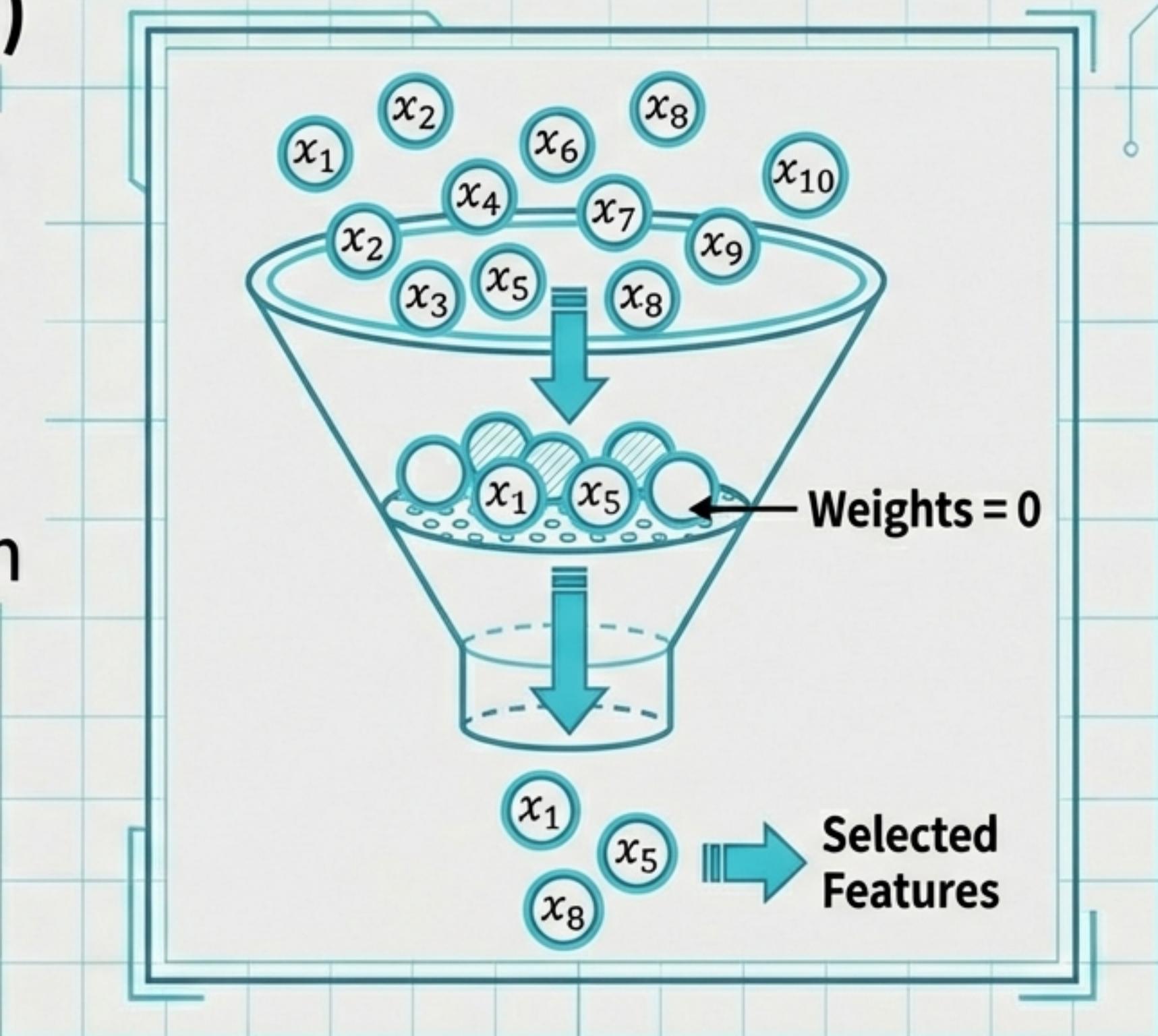
問題 (Problem)：高維度稀疏資料。

化工實例：在數千個分子描述符中，  
找出真正影響  $T_g$  的關鍵因子。

解法 (Solution)：L1 Regularization

$$\text{Loss} + \alpha \sum |w|$$

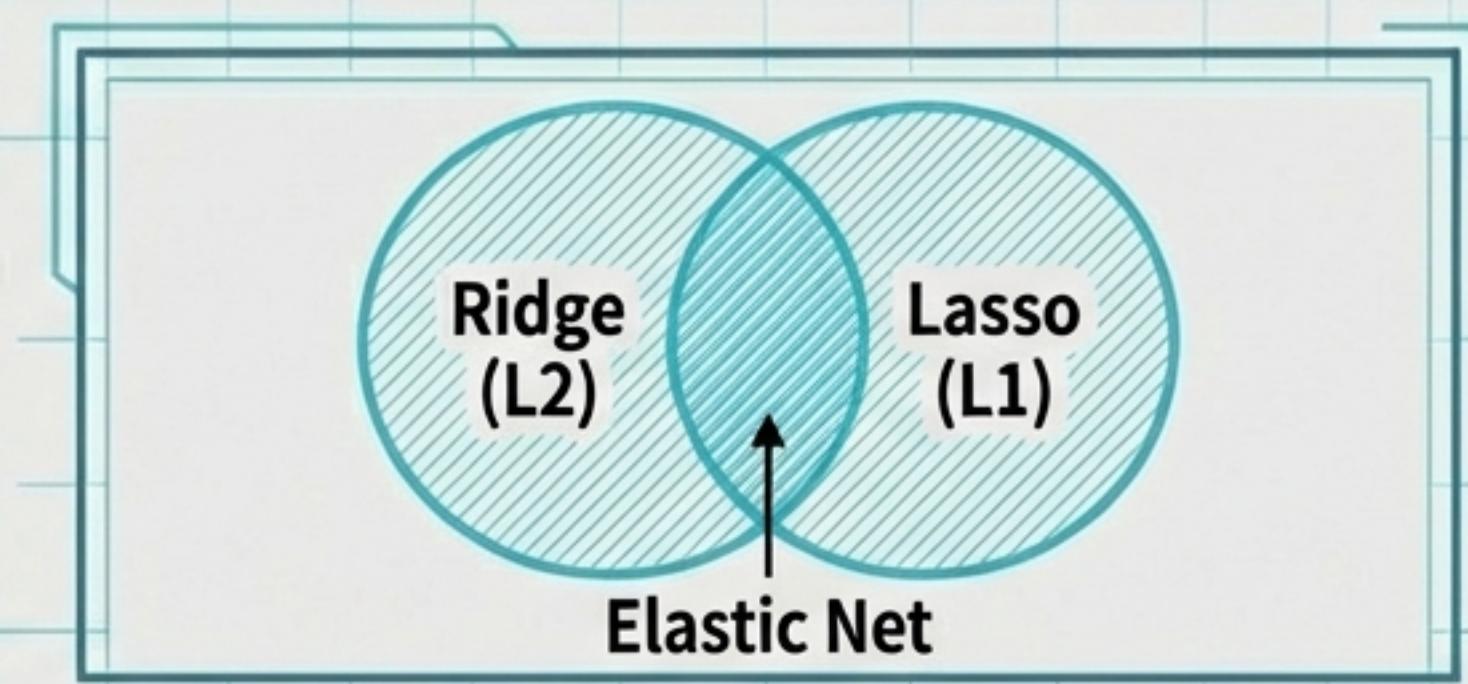
效果：強制不重要特徵權重歸零  
(Feature Selection)。



# 混合策略與大數據處理 Roboto Mono

## Elastic Net (彈性網路)

- 結合 L1 與 L2 的優點 ( $\rho$  參數控制比例)。
- 適用場景：特徵數量 > 樣本數，或特徵間存在群組相關性。



## SGD Regressor (隨機梯度下降)

- 使用 Gradient Descent 迭代求解，而非一次性矩陣運算。
- 適用場景：大規模資料集 (Large Scale Data) 或線上學習 (Online Learning)。

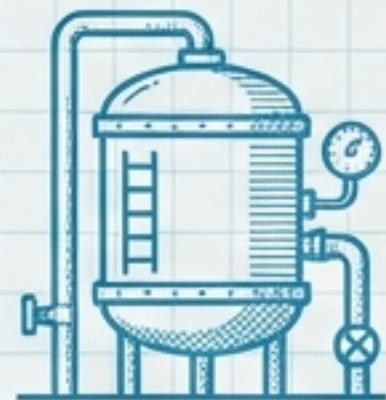


# 模型選擇指南 (Decision Matrix) Robo Mono

模型 (Model)	正則化 (Penalty)	特徵選擇 (Selection)	適用場景 (Best Use Case)
Linear Regression	None	No	簡單線性問題 (Simple Linear)
Ridge	L2	No	多重共線性 (Collinearity) - Default
Lasso	L1	Yes	稀疏資料/特徵篩選 (Sparse)
Elastic Net	L1 + L2	Yes	複雜特徵關係 (Complex)
SGD Regressor	Any	Varies	超大數據量 (Massive Data)

Safety Orange  
Recommended Starting Point

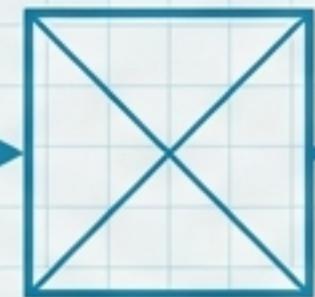
# 數據前處理：關鍵的一步 (Feed Preparation)



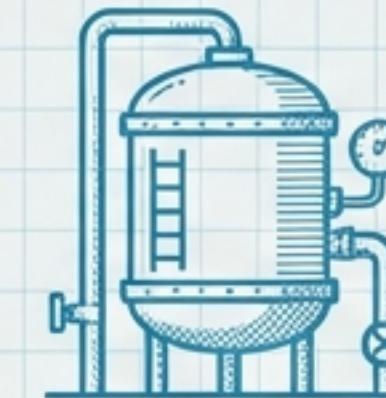
Raw Data

Different Scales  
(K, Bar, kg/hr)

$$\text{Z-score Formula: } z = \frac{x - \mu}{\sigma}$$



StandardScaler



Clean Data

Unified Scale  
(Mean=0, Std=1)

## ■ 獨熱編碼 (One-Hot Encoding)

- 處理類別變數 (e.g., Reactor Type A/B)
  - ['A'] → [1, 0]
  - ['B'] → [0, 1]

# 模型評估指標 (Evaluation Metrics) Roboto Mono

MSE / RMSE



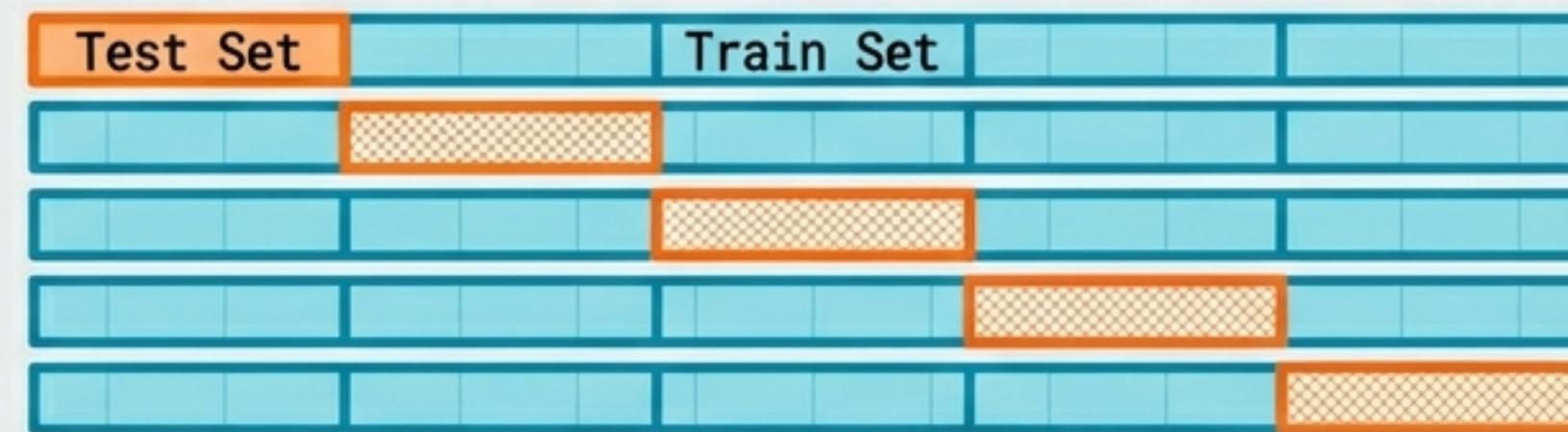
MAE



$R^2$  Score

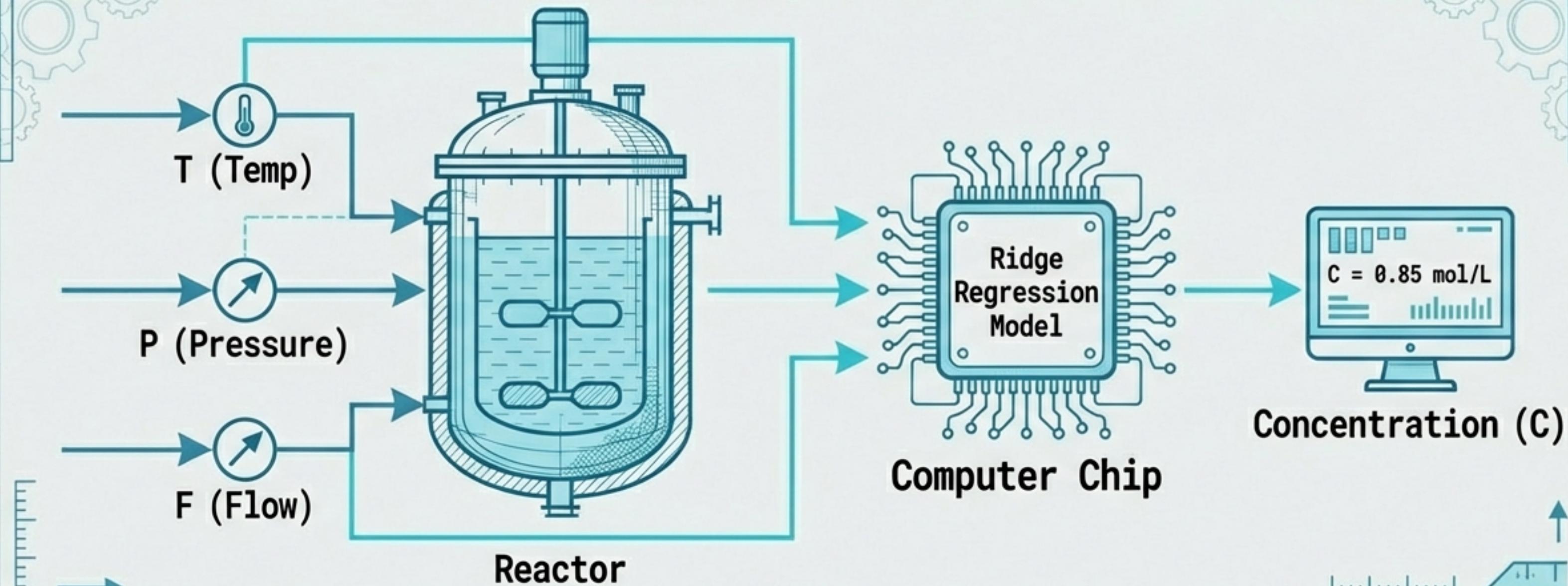


## 交叉驗證 (Cross-Validation)



K-Fold: 確保模型穩健性，避免過擬合 (Robustness check)

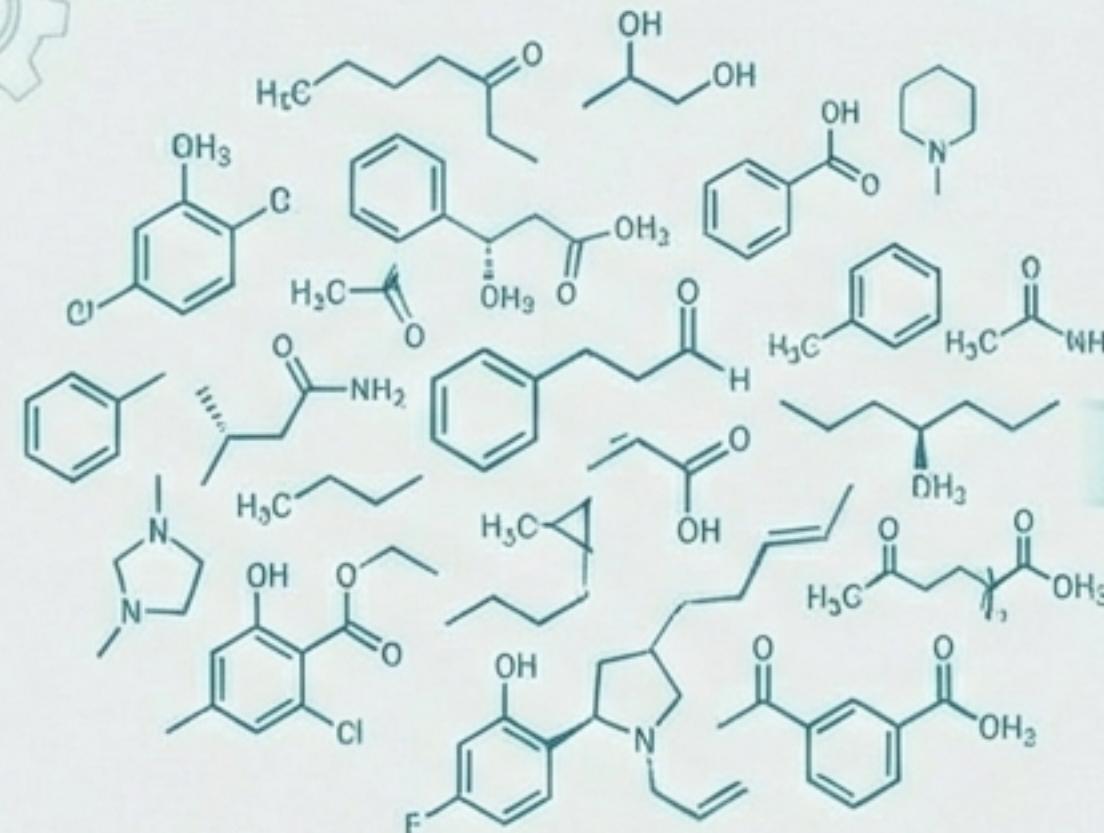
# 應用案例 I：軟感測器 (Soft Sensors) Roboto Mono



- 情境：催化裂解反應產率預測
- 挑戰：變數間高度相關 (High Correlation)

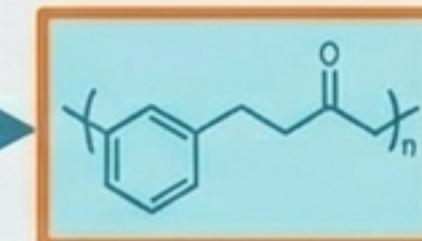
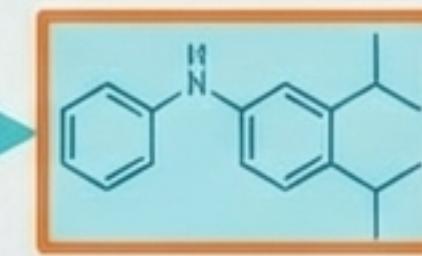
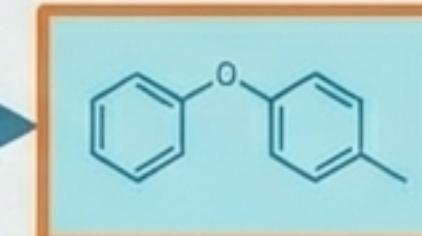
- 模型：Ridge Regression
- 結果： $R^2 = 0.92$  (即時製程優化)

# 應用案例 II：特徵篩選 (Feature Selection) Roboto Mono



Lasso Regression

Lasso Regression



3 Prominent Chemical Structures

- 情境：聚合物性質預測 (Polymer QSPR)
- 挑戰：找出決定玻璃轉化溫度 ( $T_g$ ) 的關鍵因子

- 解決方案：Lasso 強制將不重要權重壓縮為 0
- 結果：篩選出 15 個關鍵結構因子 ( $R^2 = 0.88$ )

# 現實世界的挑戰與限制 (Real World Challenges)

## 非線性關係 (Non-Linearity)

Arrhenius eq  
(Rate = A \* exp(-Ea/RT))

線性模型可能失效  
(Linear Model Fails)

→ 化學反應速率 (Arrhenius eq) 通常是非線性的。線性模型可能失效。

## 外推限制 (Extrapolation)

訓練範圍  
(Training Range)

反應失控溫度  
(Runaway Temp)

→ 模型無法預測訓練範圍以外的行為 (例如：反應失控溫度)。

## 異常值 (Outliers)

感測器故障數據  
(Sensor Fault Data)

OLS 結果  
(OLS Results)

→ 感測器故障數據會嚴重影響 OLS 結果。

# 總結與學習路徑 (Summary & Roadmap) Roboto Mono

- 線性模型是 ML 的「第一原理」：簡單、可解釋。
- Ridge (L2) 處理共線性，Lasso (L1) 處理特徵選擇。
- 數據前處理 (Scaling) 對線性模型至關重要。

## Next Step: Python 實作

深入各個子單元  
(Unit 10\_Ridge, Unit 10\_Lasso)

「AI 不會取代化工工程師，  
但懂得使用 AI 的化工工程師將取代不懂的人。」