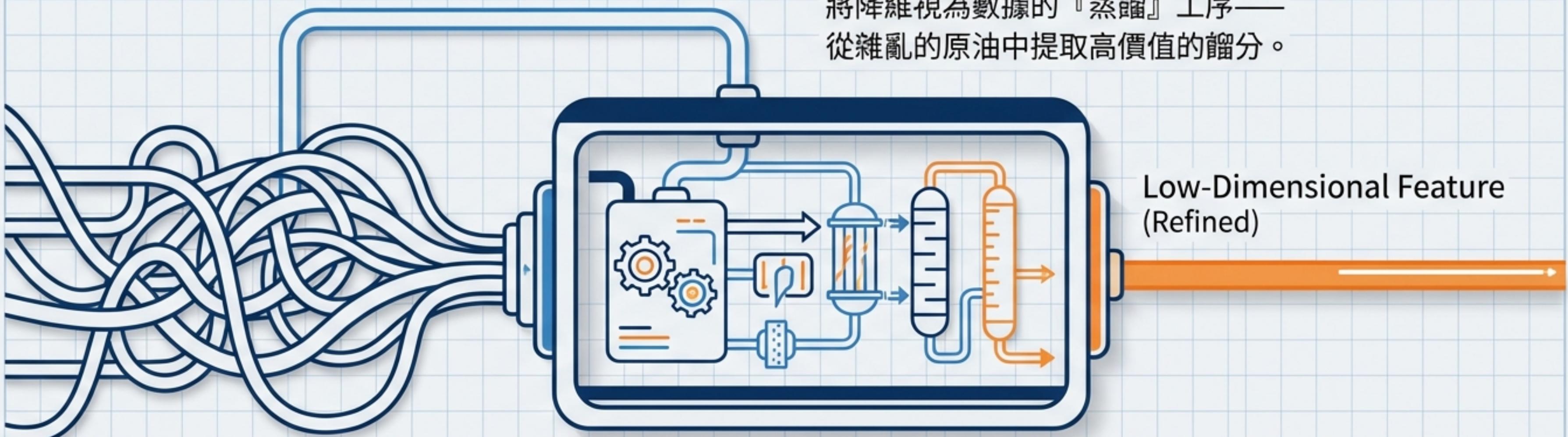


Unit 06 降維技術 (Dimensionality Reduction)

化工大數據的特徵萃取與視覺化 Noto Sans TC Regular



High-Dimensional Data (Raw)
Noto Sans TC Regular

DR-Unit-06

將降維視為數據的『蒸餾』工序——
從雜亂的原油中提取高價值的餾分。

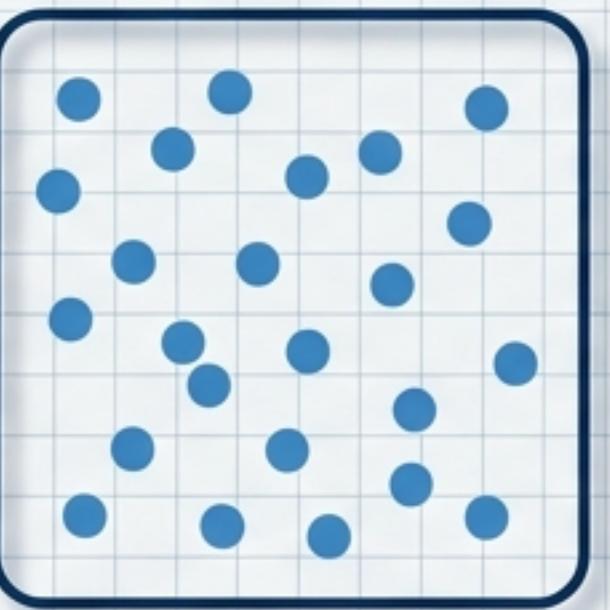
Low-Dimensional Feature
(Refined)

為什麼需要降維？維度詛咒的挑戰

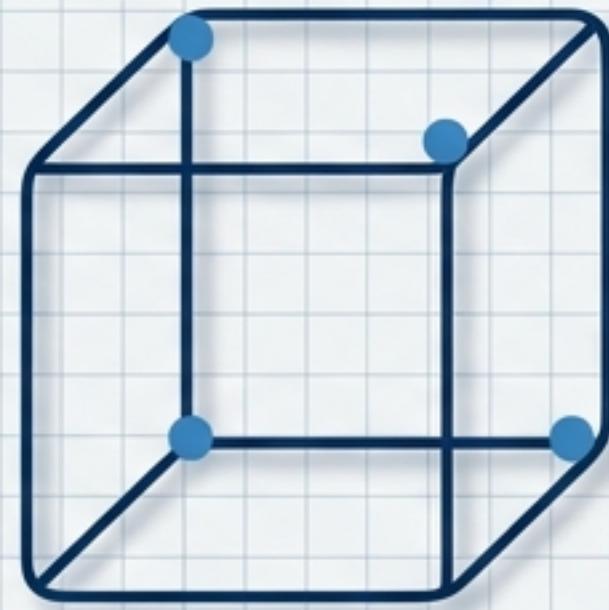
1D



2D



3D (Sparse)



$$l = p^{1/d}$$

當維度 d 很大時，
邊長比例 l 接近 1
(需覆蓋整個空間)



數據稀疏性
(Data Sparsity)：
距離度量失效



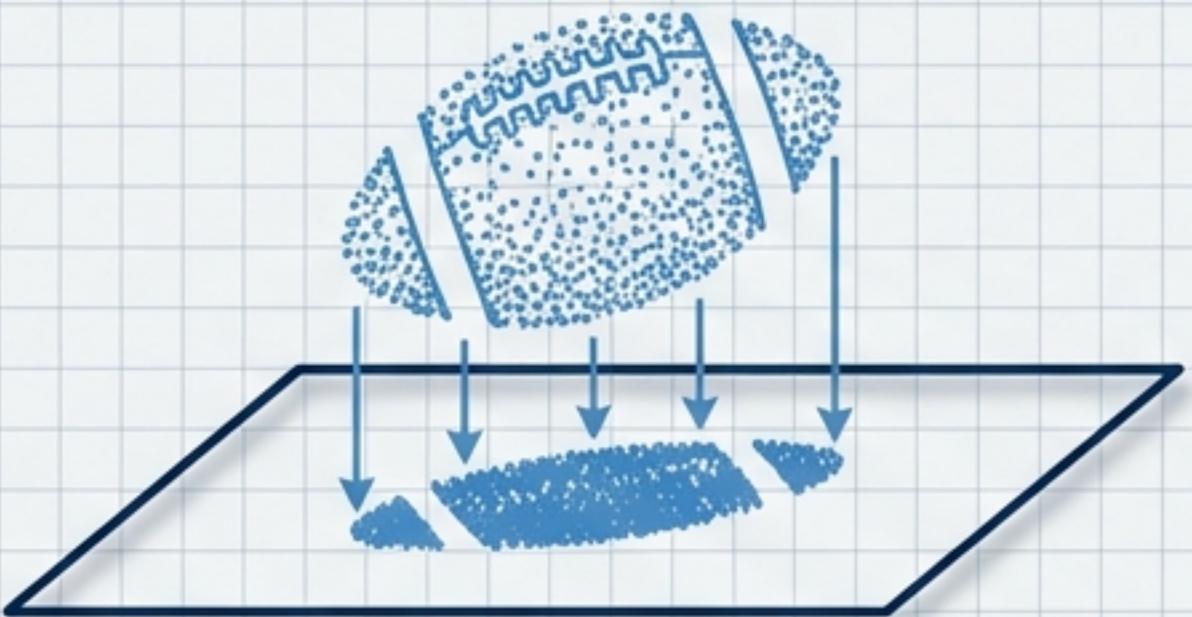
計算複雜度：
隨維度指數增長



模型過擬合
(Overfitting)：
雜訊被誤認為特徵

降維的數學本質：投影與流形

線性投影 (Linear Projection)

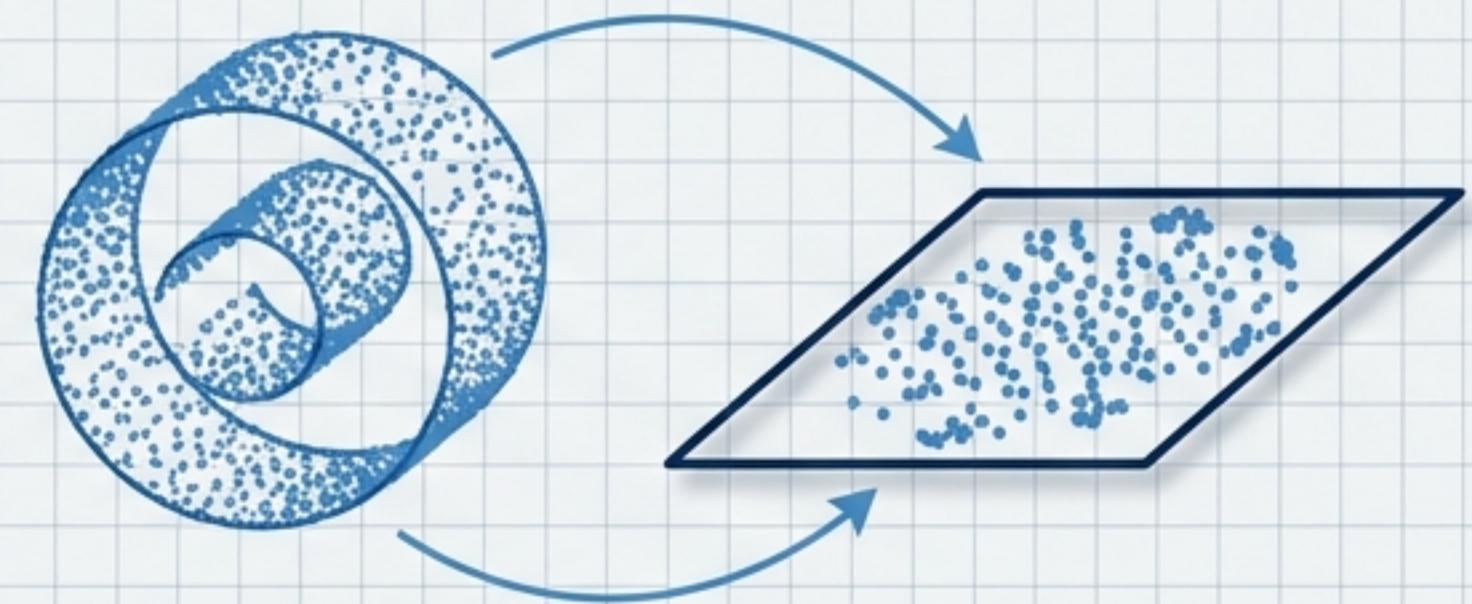


假設：變異可透過線性組合表達

Formula: $y = W^T x$

代表：PCA

流形學習 (Manifold Learning)



假設：數據位於非線性拓樸結構上

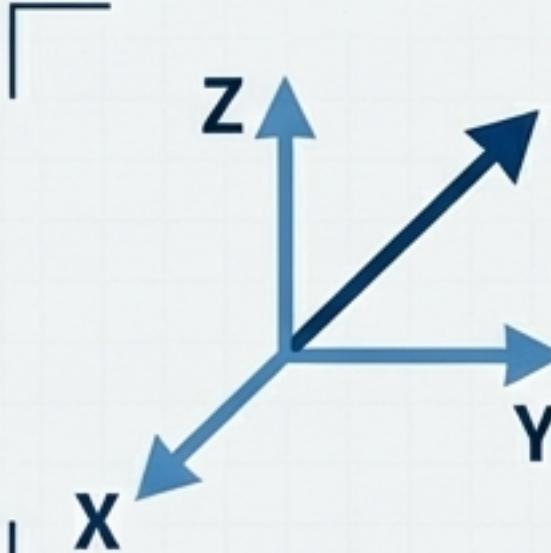
Formula: $y = f(x)$

代表：t-SNE, UMAP

$f: \mathbb{R}^D \rightarrow \mathbb{R}^d$, where $d \ll D$

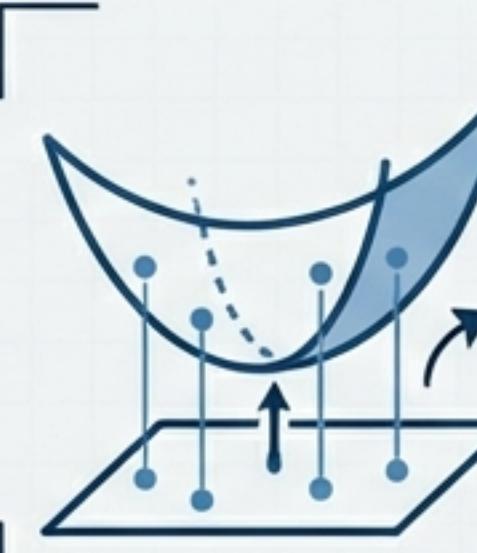
工程師的工具箱：主流降維演算法

PCA (主成分分析)



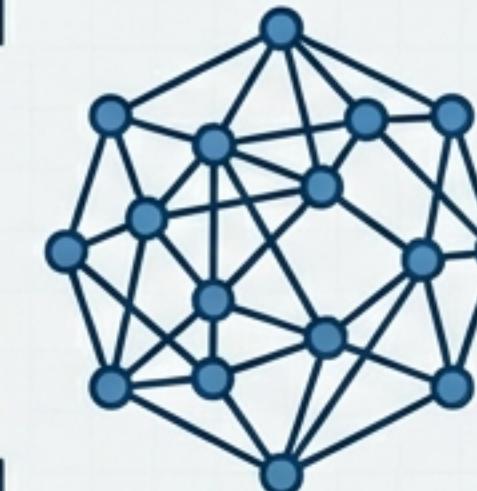
線性 (Linear)
用途：製程監控的主力
(The Workhorse)
關鍵：最大變異數方向

Kernel PCA

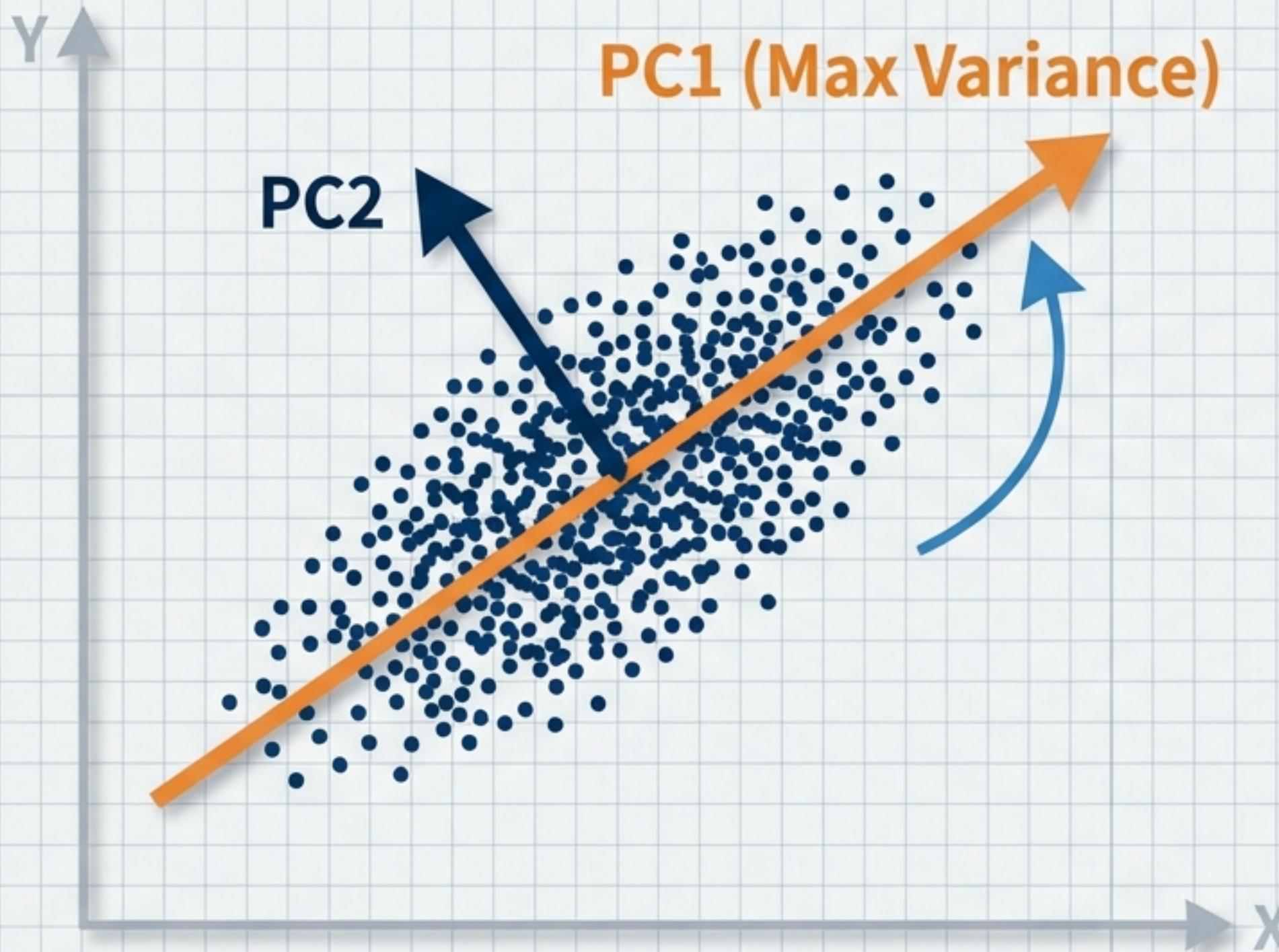


非線性 (Non-linear)
用途：處理非線性邊界的
反應器
關鍵：Kernel Trick (核技巧)

 **t-SNE**
非線性 / 視覺化
用途：顯微鏡 -
專注局部細節
關鍵：局部群集分離

 **UMAP**
非線性 / 混合
用途：現代化高效分離塔
**關鍵：平衡全局與局部結構，
速度快**

主成分分析 (PCA)：製程監控的核心



核心原理

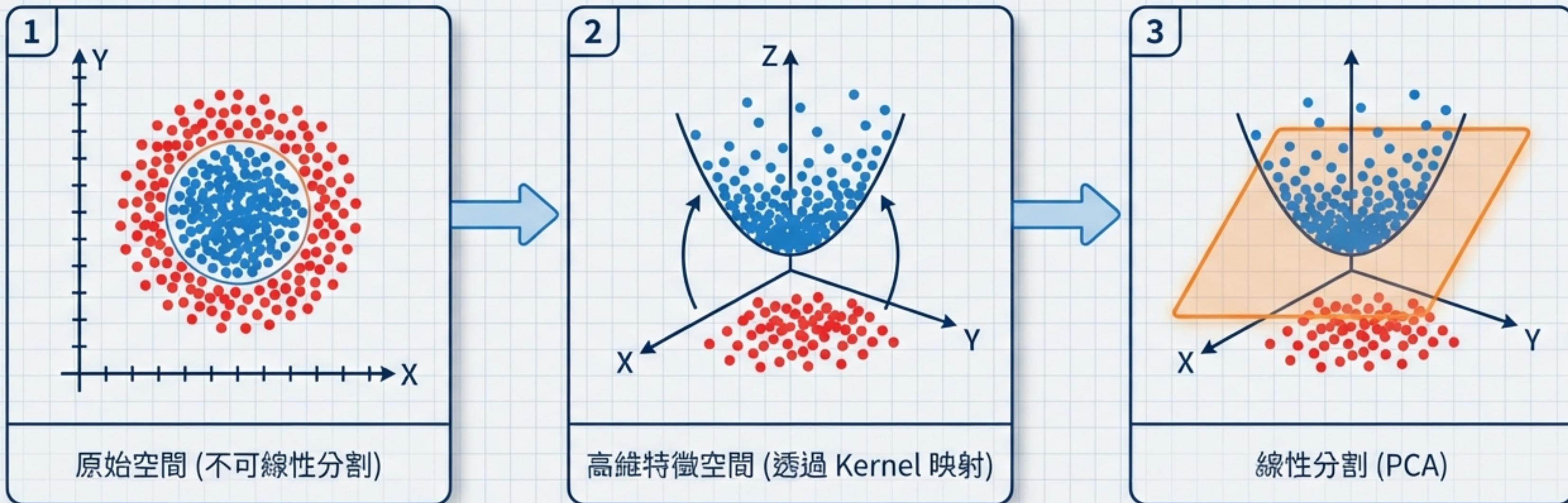
尋找數據變異數最大的方向，建立新的正交座標系。

$$w_1 = \operatorname{argmax} \operatorname{Var}(Xw)$$

化工應用

1. 建立 T^2 與 SPE 控制圖 (異常檢測)
2. Loading 分析：回推溫度、壓力對變異的貢獻度。

核主成分分析 (Kernel PCA)：駕馭非線性

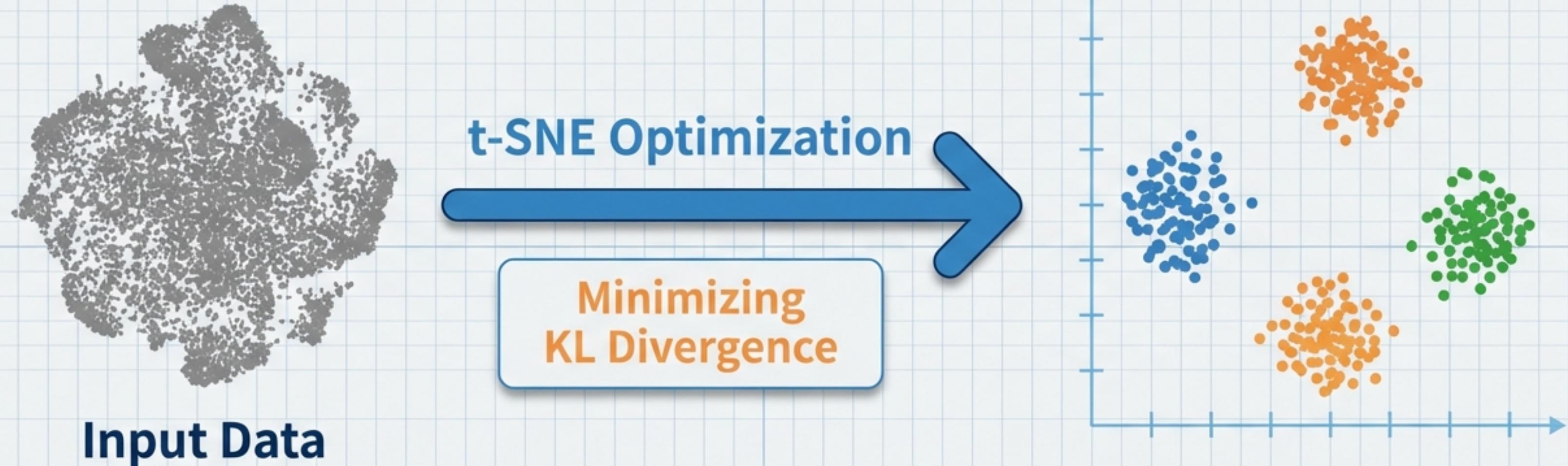


核技巧 (Kernel Trick)：無需顯式計算高維映射，直接計算相似度。

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

Common Kernels: RBF, Polynomial, Sigmoid

t-SNE：極致的群集視覺化



Pros



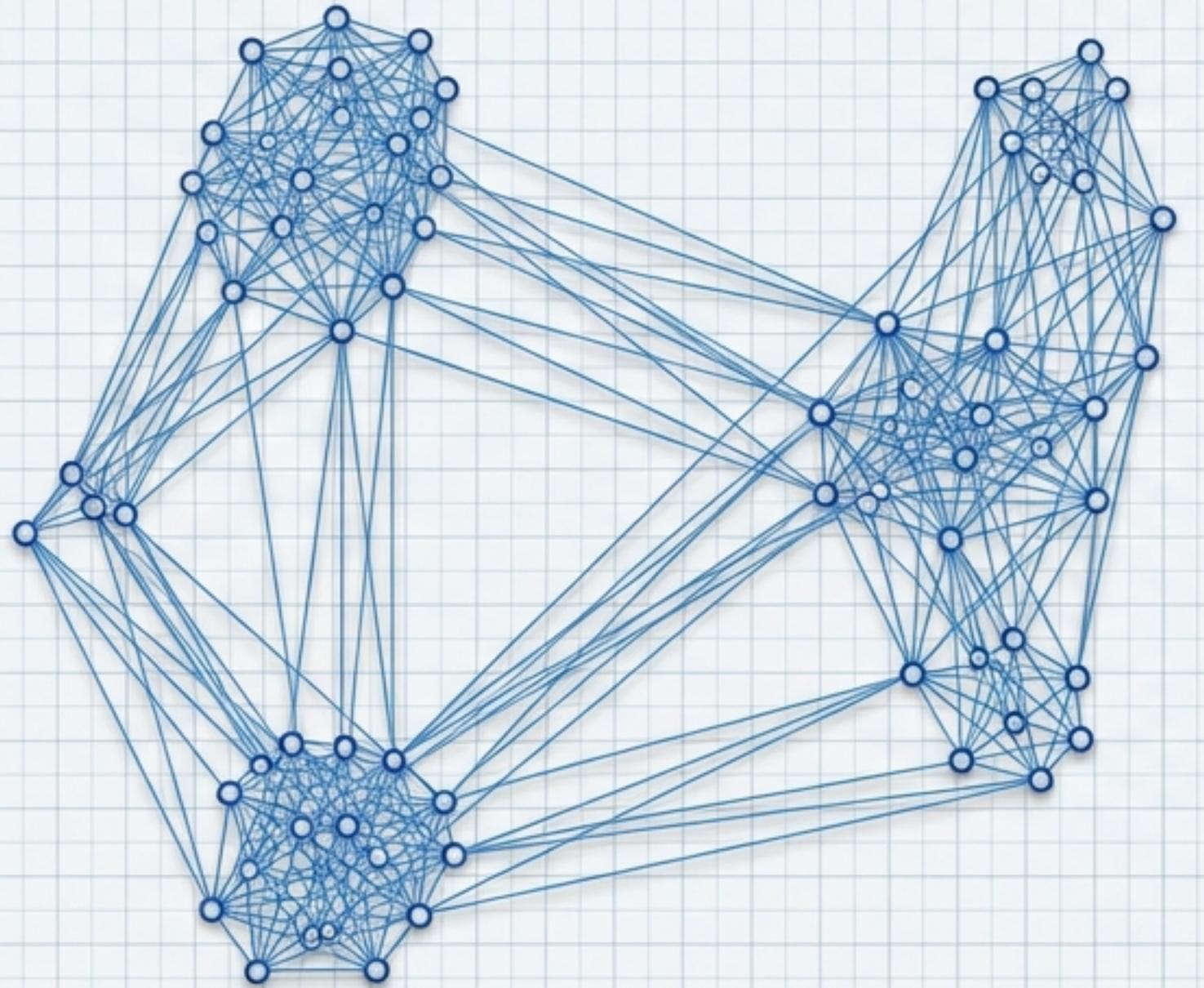
局部結構保持：能清晰分開不同的操作模式或產品等級。

Cons



計算成本高： $O(n^2)$ ，不適合大規模數據。
隨機性：結果不可重現 (需固定 Seed)。

UMAP：速度與結構的平衡

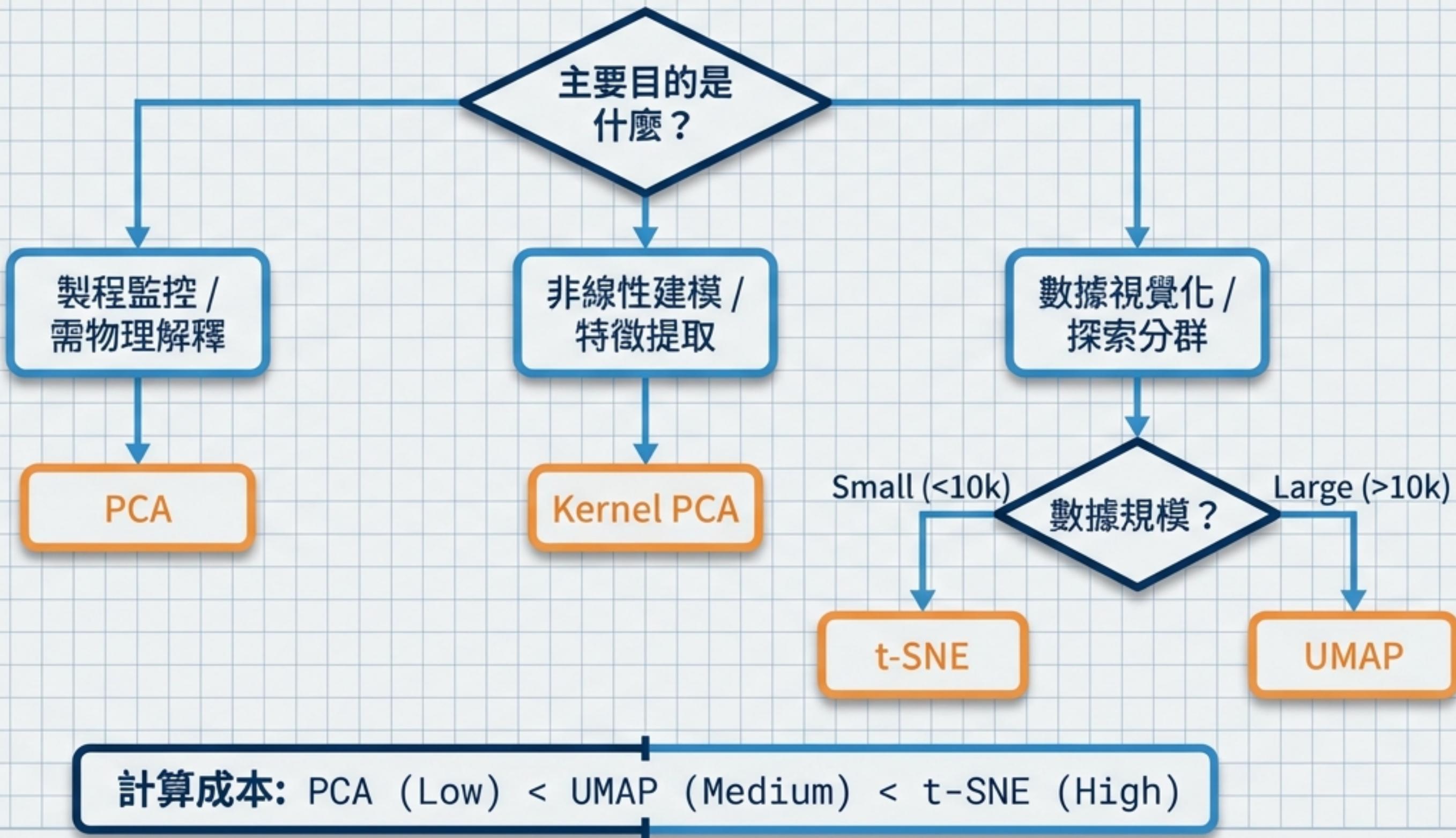


Simplicial Complex

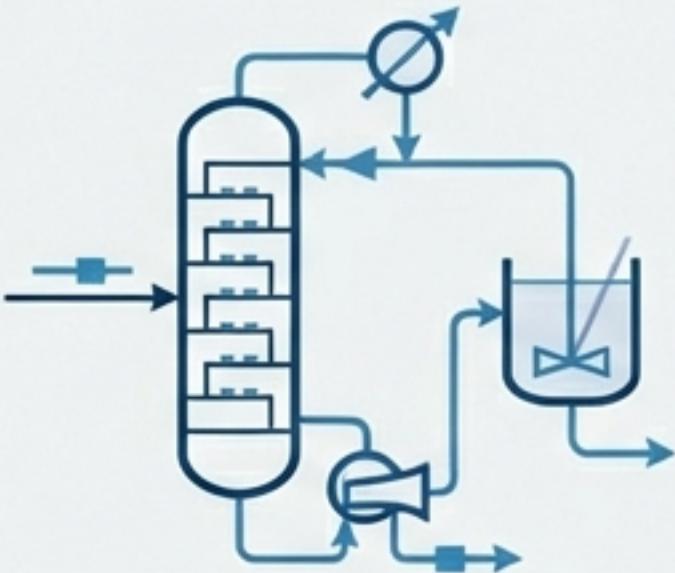
為何選擇 UMAP?

- **理論基礎**：基於 Riemannian geometry 與拓撲數據分析。
- **速度優勢**：複雜度接近 $O(n)$ ，比 t-SNE 快得多。
- **結構保持**：更能保留數據的全局結構 (Global Structure)，適合跨廠區比較。
- **支持新數據**：具備 'transform' 方法，可應用於即時監控 (t-SNE 做不到)。

決策指南：如何選擇演算法？

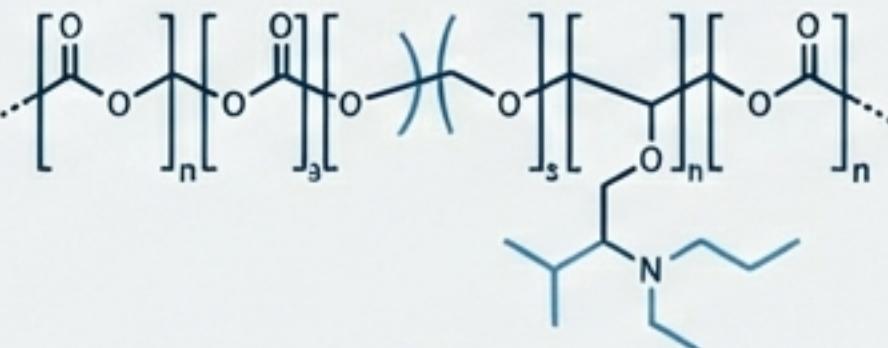


化工實戰案例



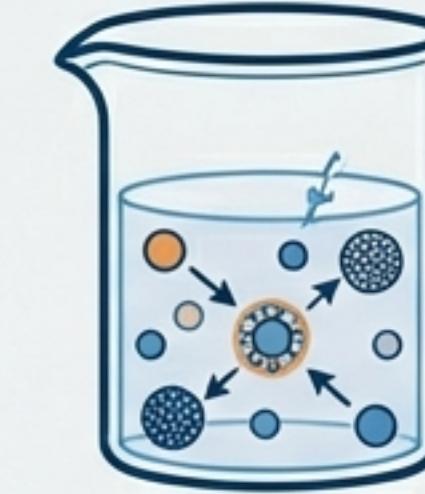
CSTR 反應器監控

痛點：**50+ 感測器**，難以監控。
對策：PCA ($T^2 + SPE$ Charts)。
效益：**降低 40% 誤報率**，
提前 15 分鐘檢測異常。



聚合物品質預測

痛點：參數多重共線性。
對策：PCA top 10 components -> 迴歸。
效益： **R^2 從 0.72 提升至 0.88**，訓練快 70%。

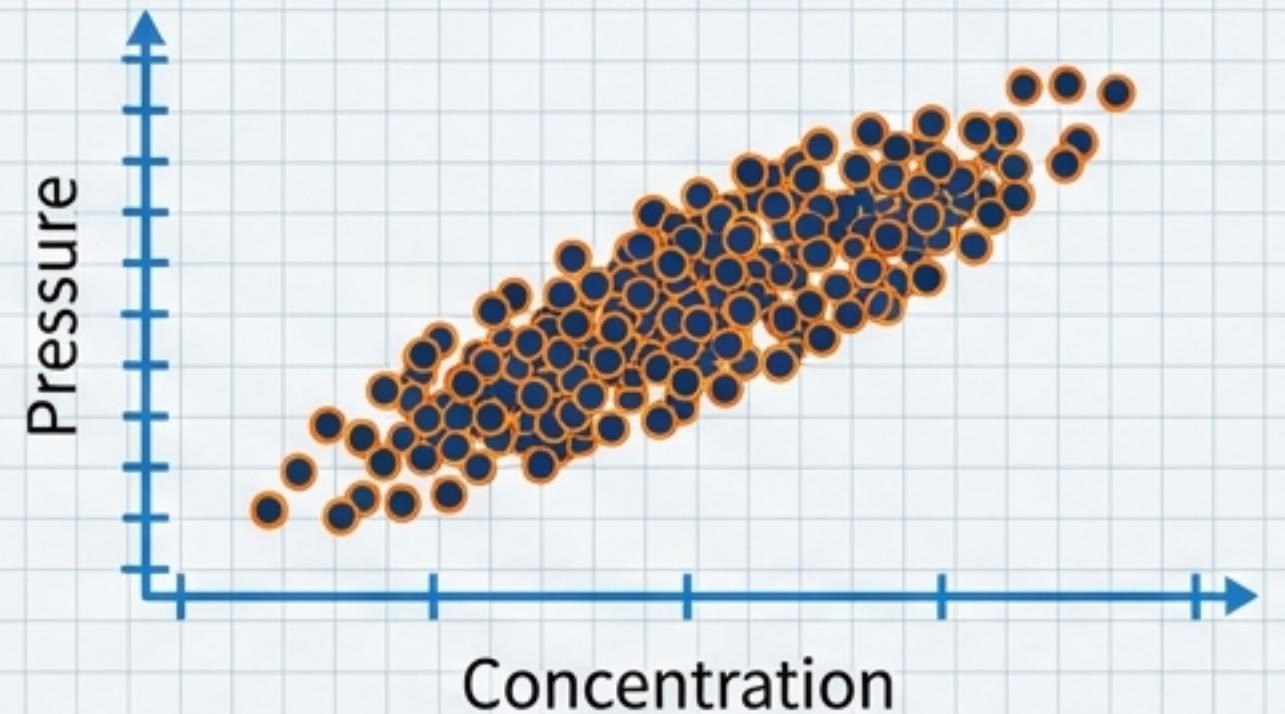


觸媒篩選

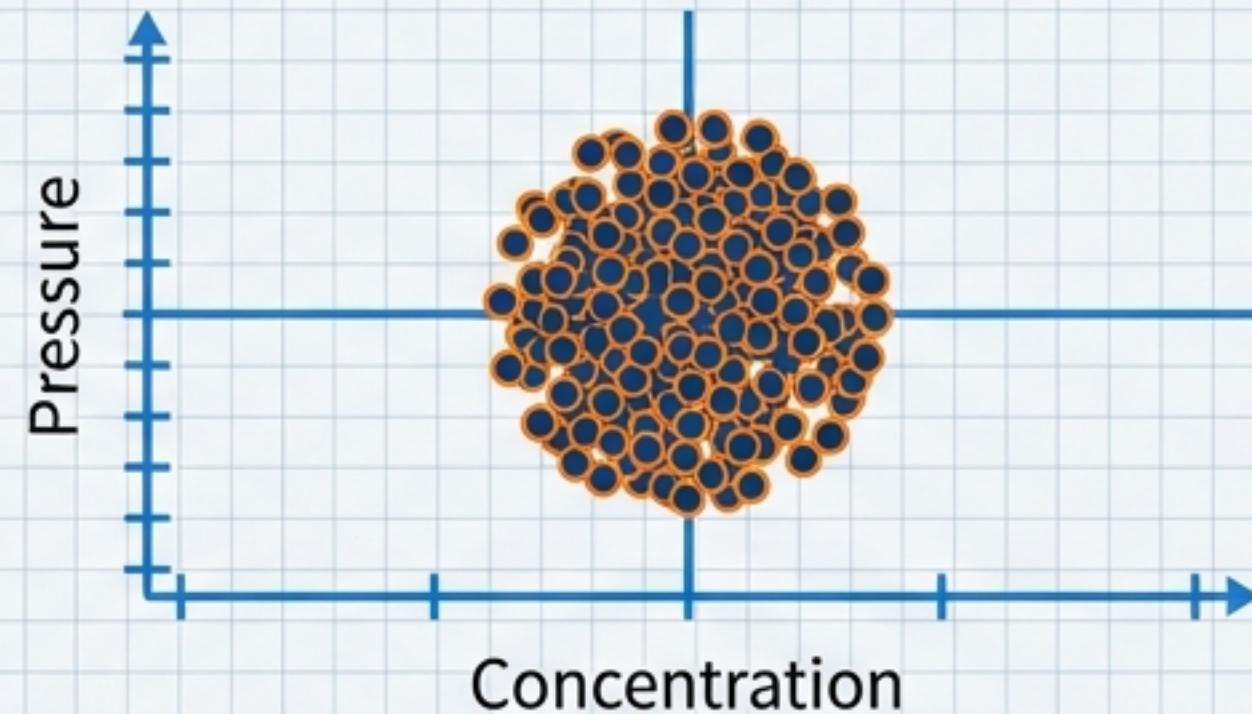
痛點：**100 種配方**，非線性關聯。
對策：UMAP 視覺化 + Kernel PCA。
效益：**縮短篩選時間 50%**。

資料前處理：品質的關鍵 (GIGO)

Raw Data (變異數不均)



Scaled Data (標準化)



StandardScaler (Z-score)

$$\frac{x - \mu}{\sigma}$$

PCA 標準配備。

MinmaxScaler

Scale to [0, 1]

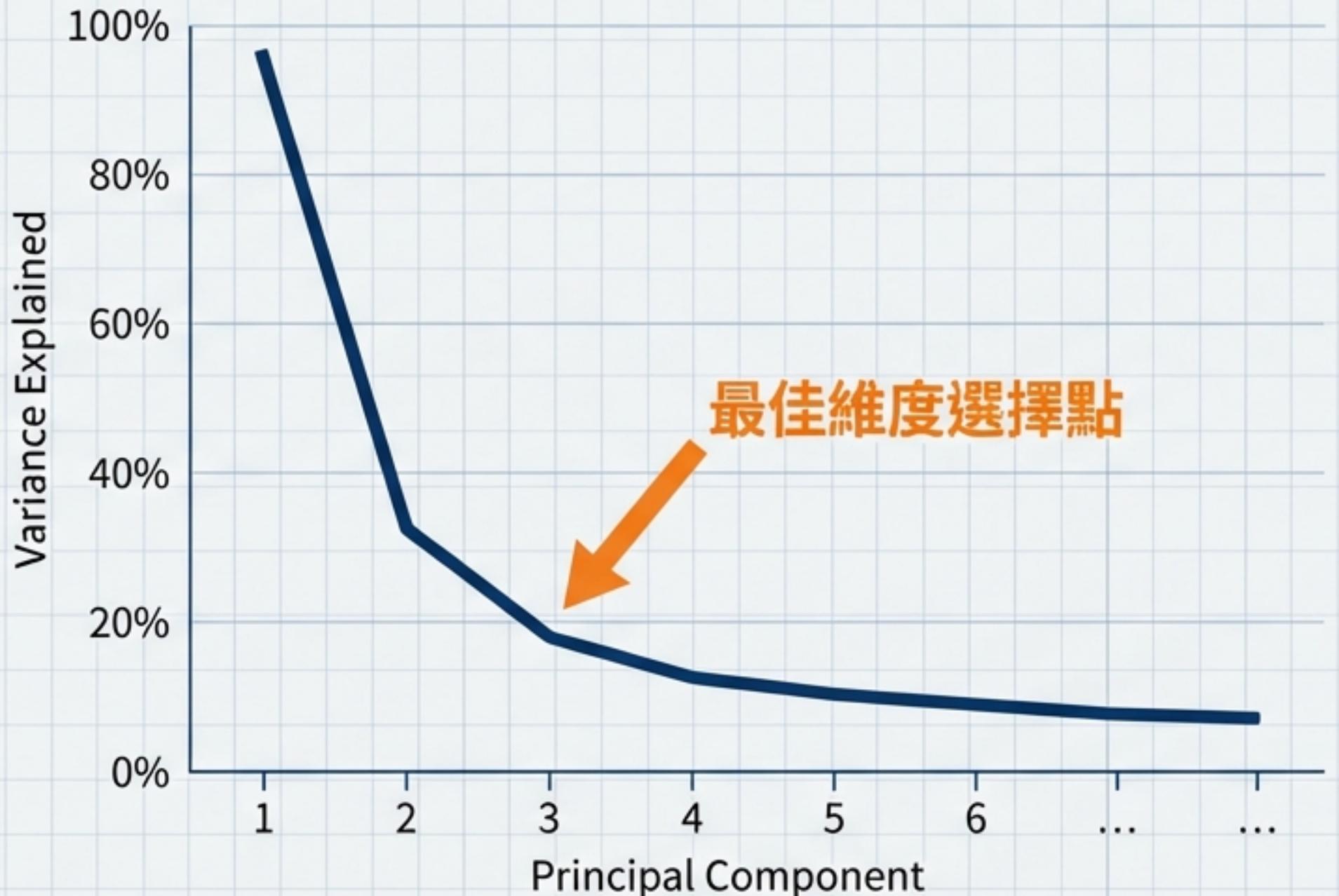
t-SNE/UMAP 常用。

RobustScaler

使用中位數與 IQR，抗異常
值干擾。

模型評估：QC 指標

Scree Plot



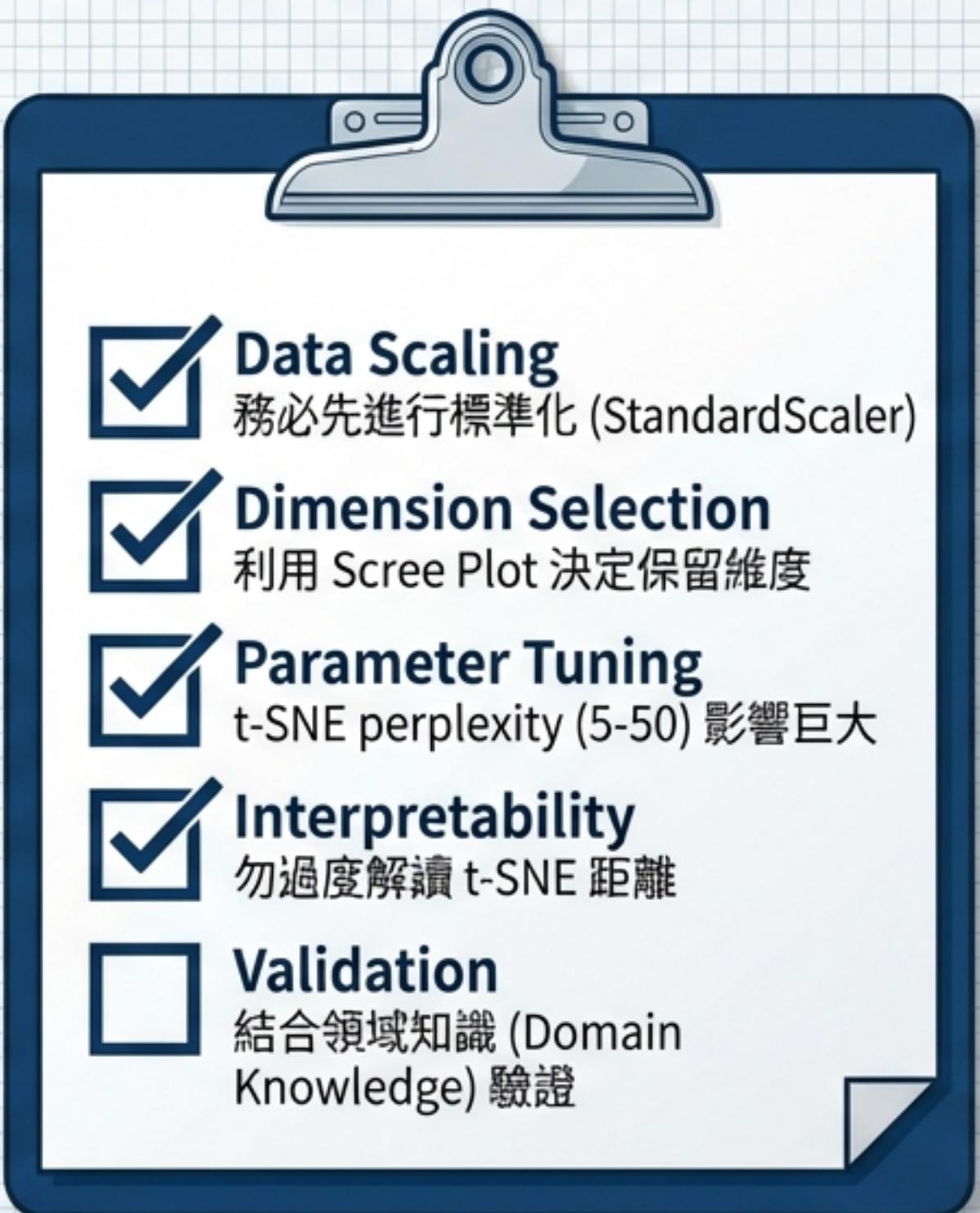
Explained Variance Ratio



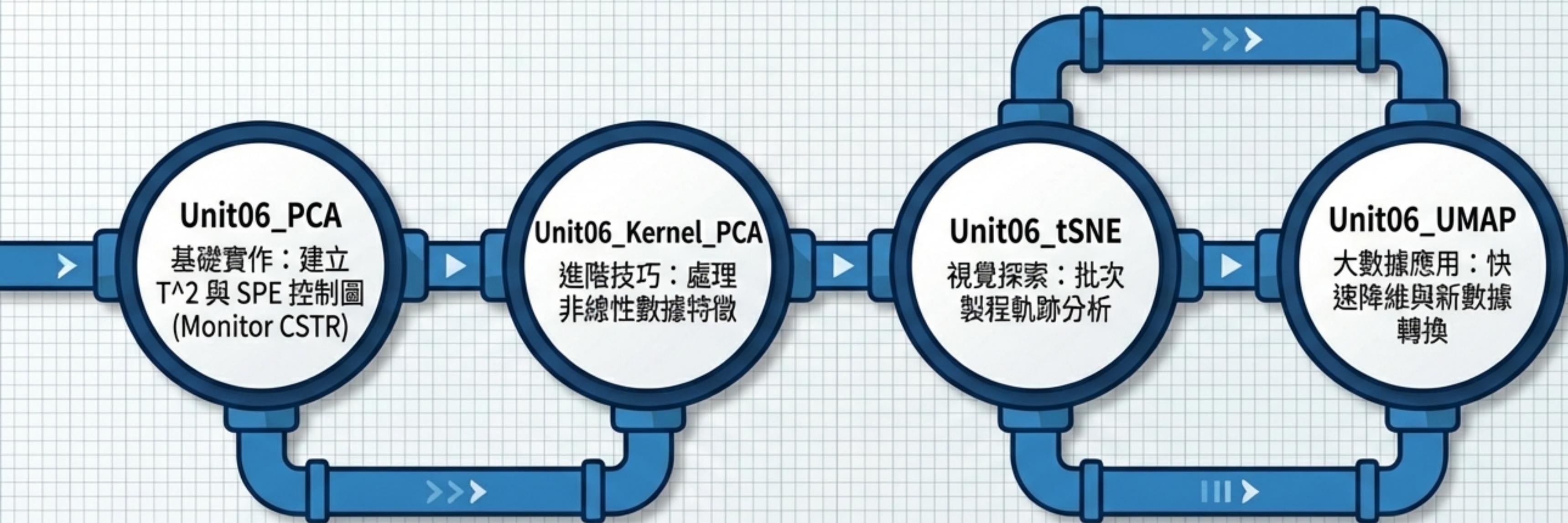
其他指標

- 重建誤差 (Reconstruction Error): 越小代表資訊保留越多
- 輪廓係數 (Silhouette Score): 評估群集分離度 (-1 to 1)
- Trustworthiness: 評估鄰近關係保持 (>0.9 為佳)

實務建議與操作清單



學習路徑與程式演練



下一步：從 `Unit06_PCA.ipynb` 開始您的實作練習。

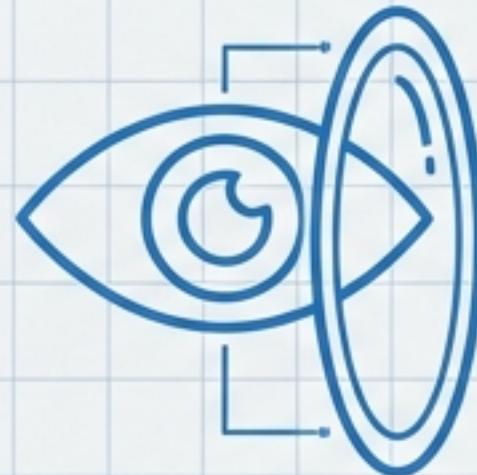
結語：數位化工廠的「壓縮機」

Simplify (簡化)



將數百個感測器讀值壓縮為關鍵指標。

Visualize (看見)



看見高維數據中的隱藏結構。

Optimize (優化)



提升下游模型效率與準確度。

降維不僅是數學運算，更是將『數據』轉化為『情報』的關鍵單元操作。