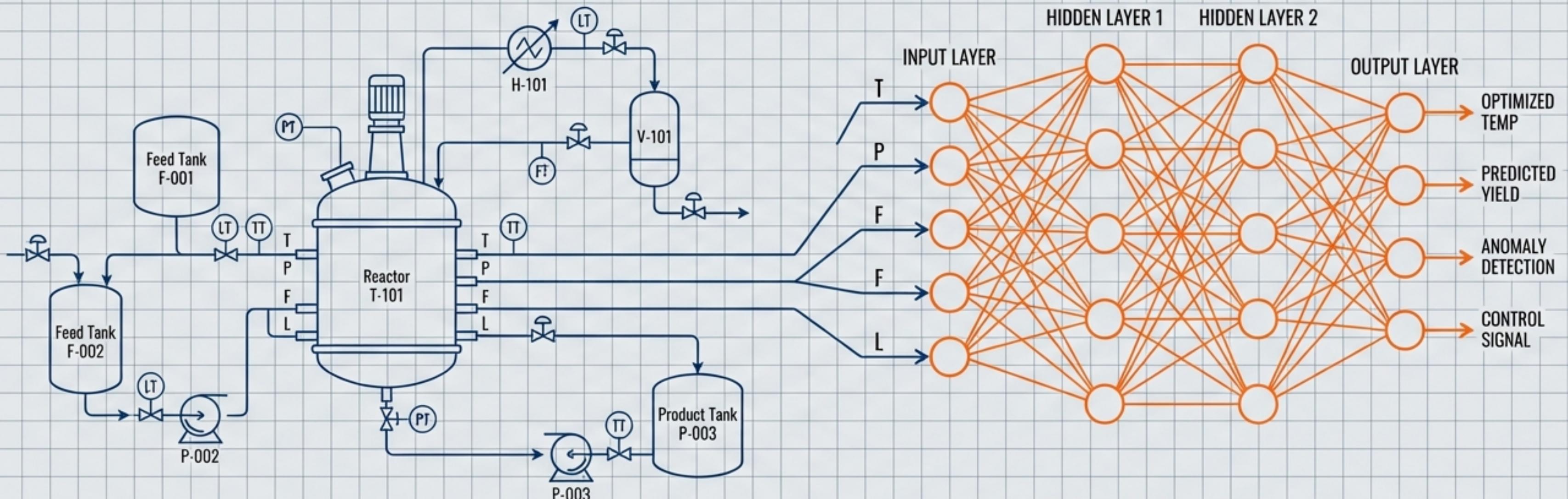


AI 在化工上的應用：高維數據視覺化

Unit 06 t-SNE 原理與實踐

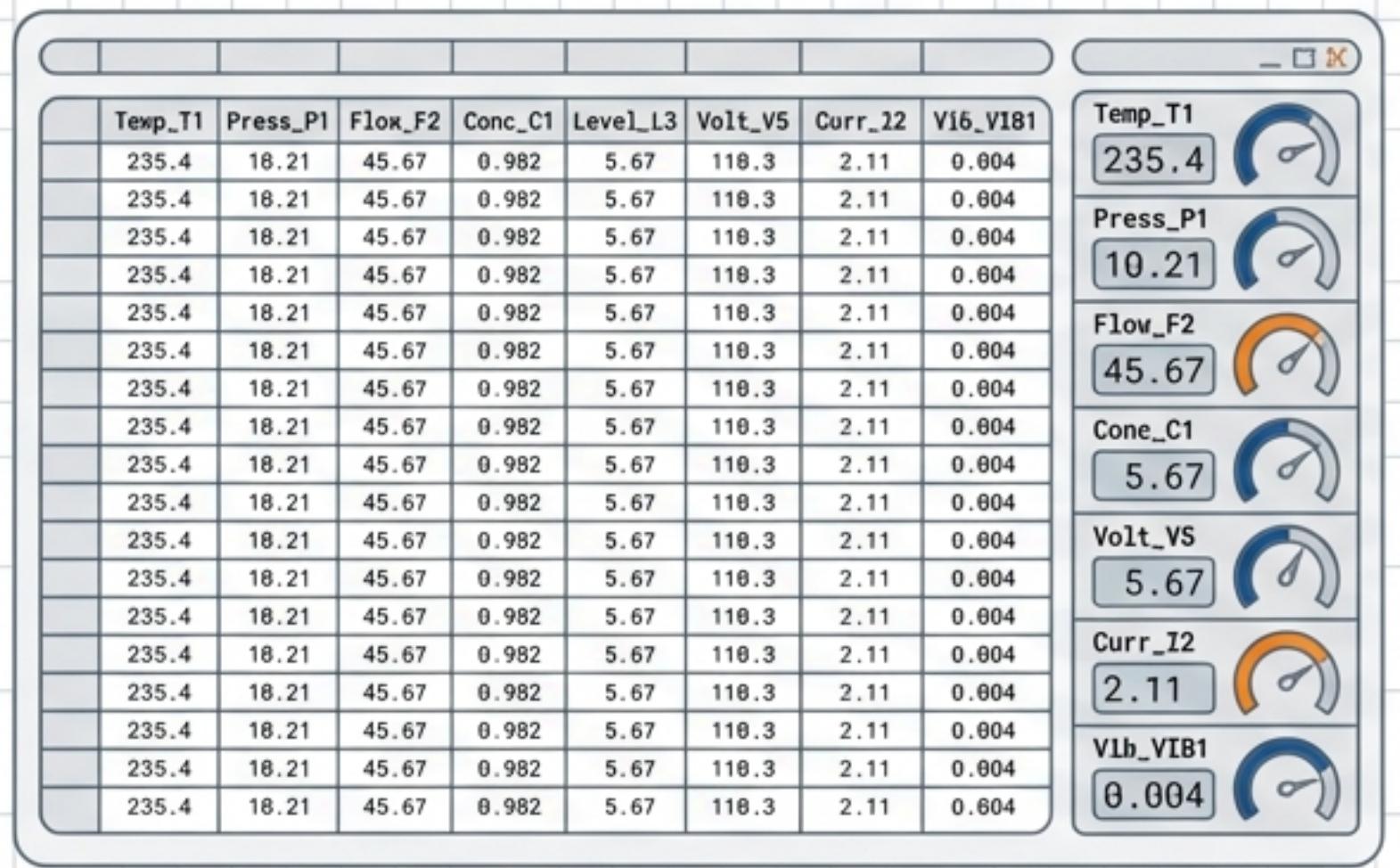


授課教師：莊曜禎 助理教授
學期：114學年度第2學期

課程目標：掌握非線性降維技術，
解鎖化工製程數據的隱藏模式

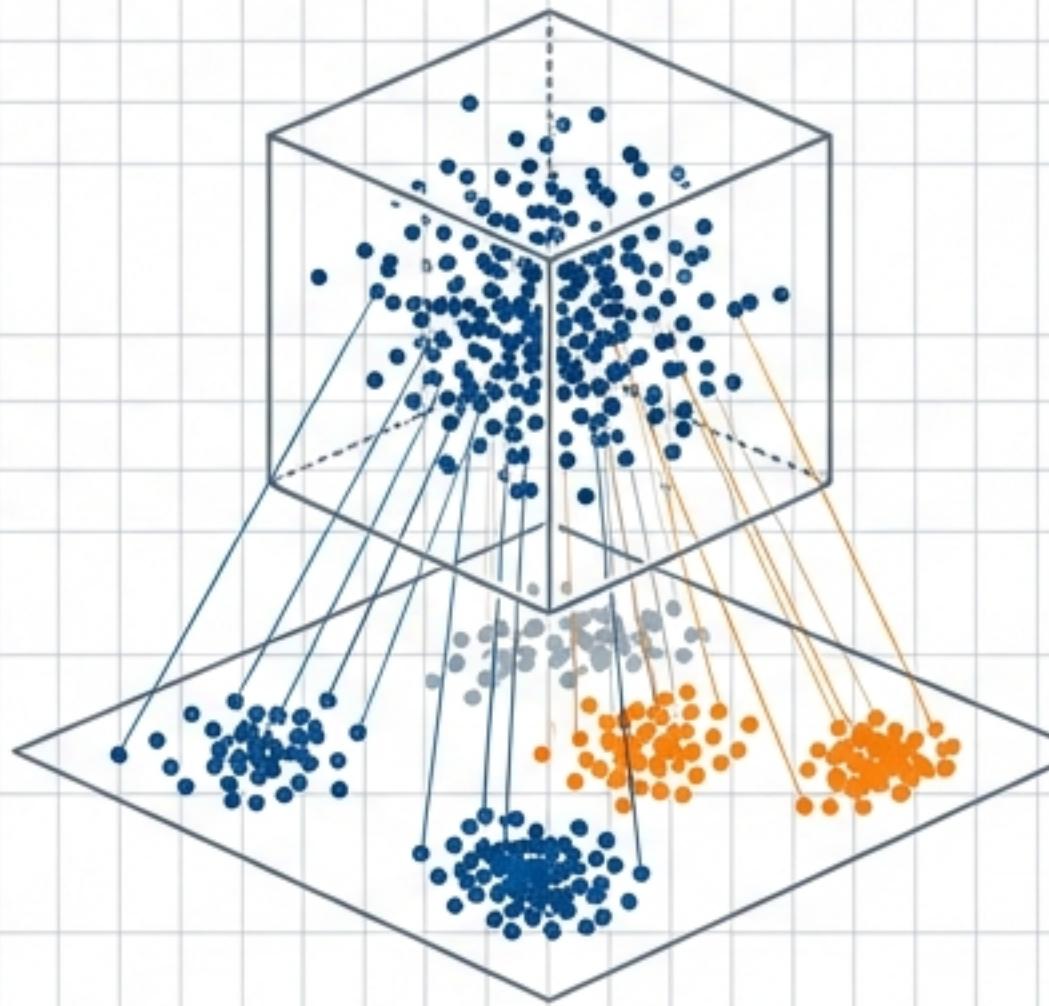
高維數據的挑戰與 t-SNE 解決方案 (The Challenge of High-Dimensional Data & t-SNE Solution)

高維數據的挑戰 (The Challenge)



化工製程通常涉及數十甚至數百個感測器變數。傳統的 2D 散佈圖無法同時呈現這些變數之間的複雜交互作用。

t-SNE：看見不可見 (Seeing the Invisible)

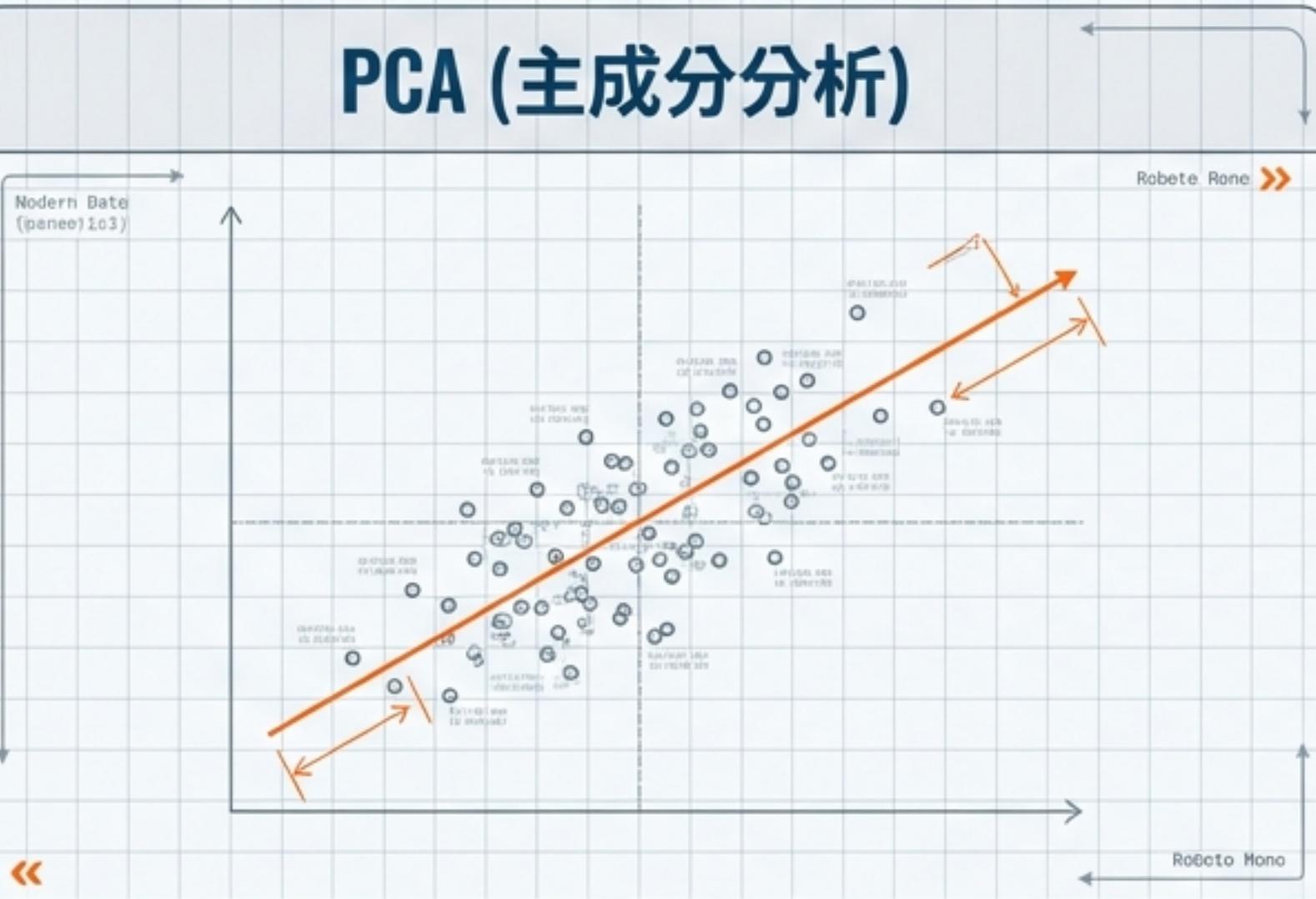


t-分布隨機鄰域嵌入 (t-SNE)

非線性降維 (Non-linear Dimensionality Reduction) - 專門用於將高維數據映射到 2D 或 3D 空間，以進行視覺化探索。

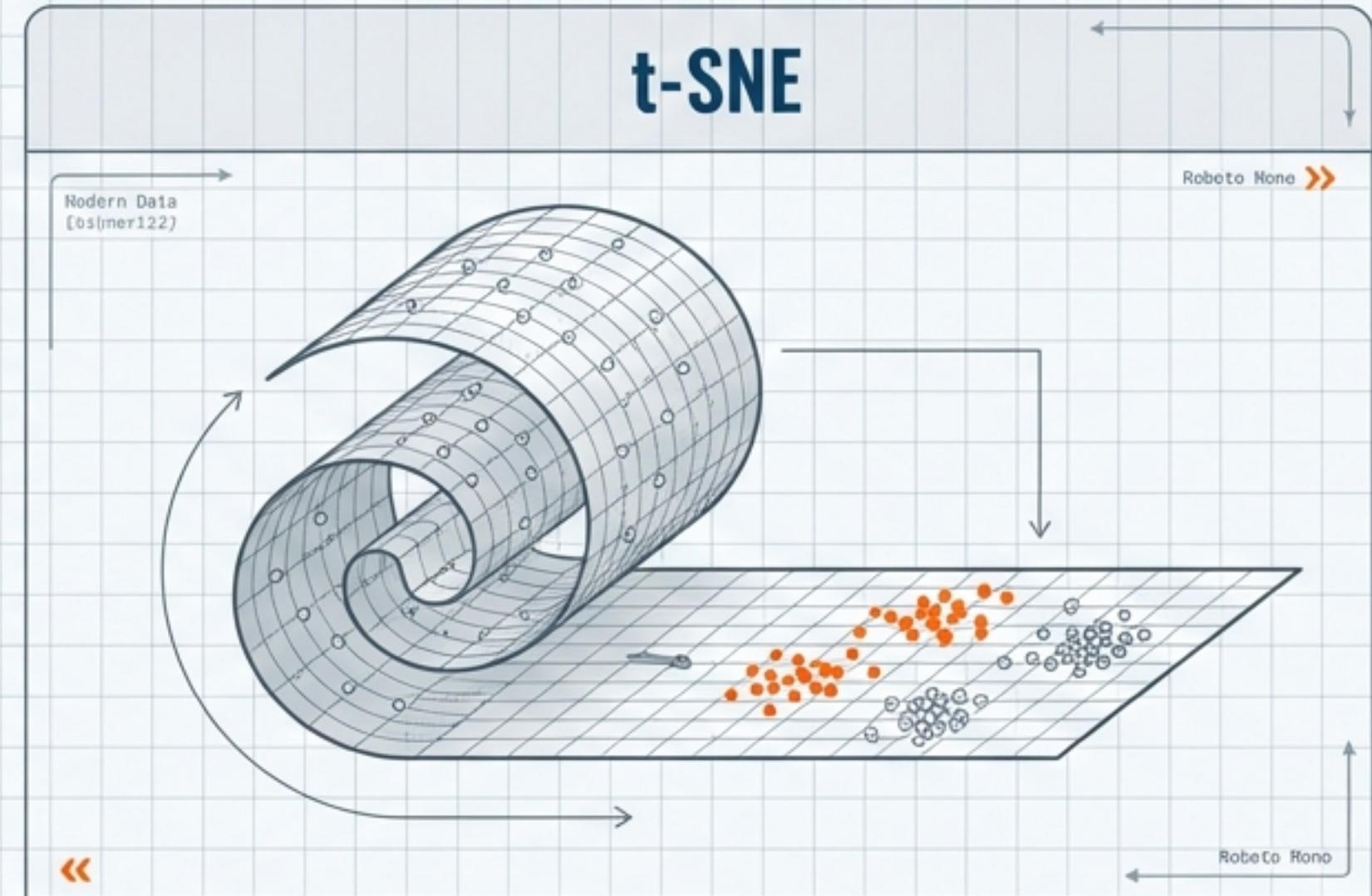
觀念解析：全局與局部的權衡

PCA (主成分分析)



- 保留全局變異數 (Global Variance)
- 像是從遠處看地圖，保留了整體的形狀，但細節可能模糊。
- 限制：僅能捕捉線性關係。

t-SNE

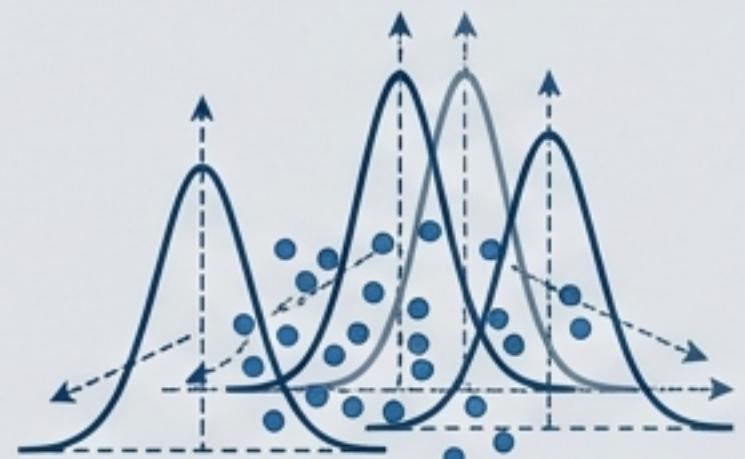


- 保持局部相似性 (Local Similarity)
- 像是關注每個人與其鄰居的關係，確保在低維空間中，鄰居依然是鄰居。
- 優勢：能揭示複雜的非線性結構與群集 (Clusters)。

t-SNE 演算法機制：從高維到低維的映射過程 (The t-SNE Algorithm Mechanism: From High-Dim to Low-Dim Mapping)

逐步解析 t-SNE 如何保留局部結構並解決擁擠問題。

1. 高維空間 (High Dim)



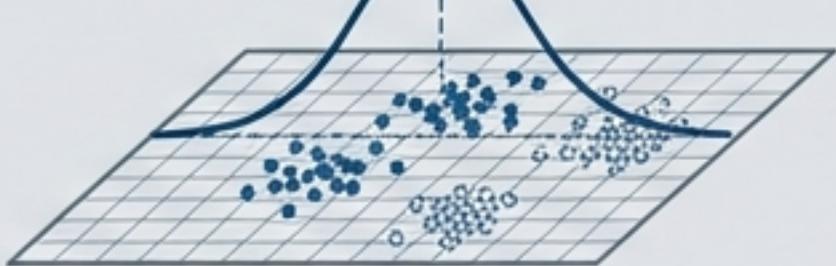
使用高斯分布 (Gaussian) 計算點與點之間的相似度條件機率 ($p_{j|i}$)。距離越近，機率越高。

$$\text{Formula: } p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

Roboto Nono

2. 低維空間 (Low Dim)

領域
知識轉換
(Domain Transformation)



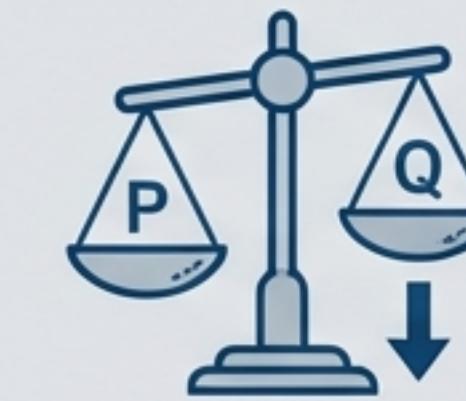
使用 t-分布 (Student t-distribution) 計算相似度 (q_{ij})。

長尾效應 (Heavy Tail) - 解決擁擠問題，允許中等距離的點分得更開。

$$\text{Formula: } q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_k - y_i\|^2)^{-1}}$$

Roboto Nono

3. 優化目標 (Optimization)



最小化 KL 散度
(Kullback-Leibler Divergence)

強迫低維分布 Q 盡可能模仿高維分布 P。

$$\text{KL}(P \parallel Q) = \sum_{ij} p_{ij} * \log(p_{ij}/q_{ij})$$

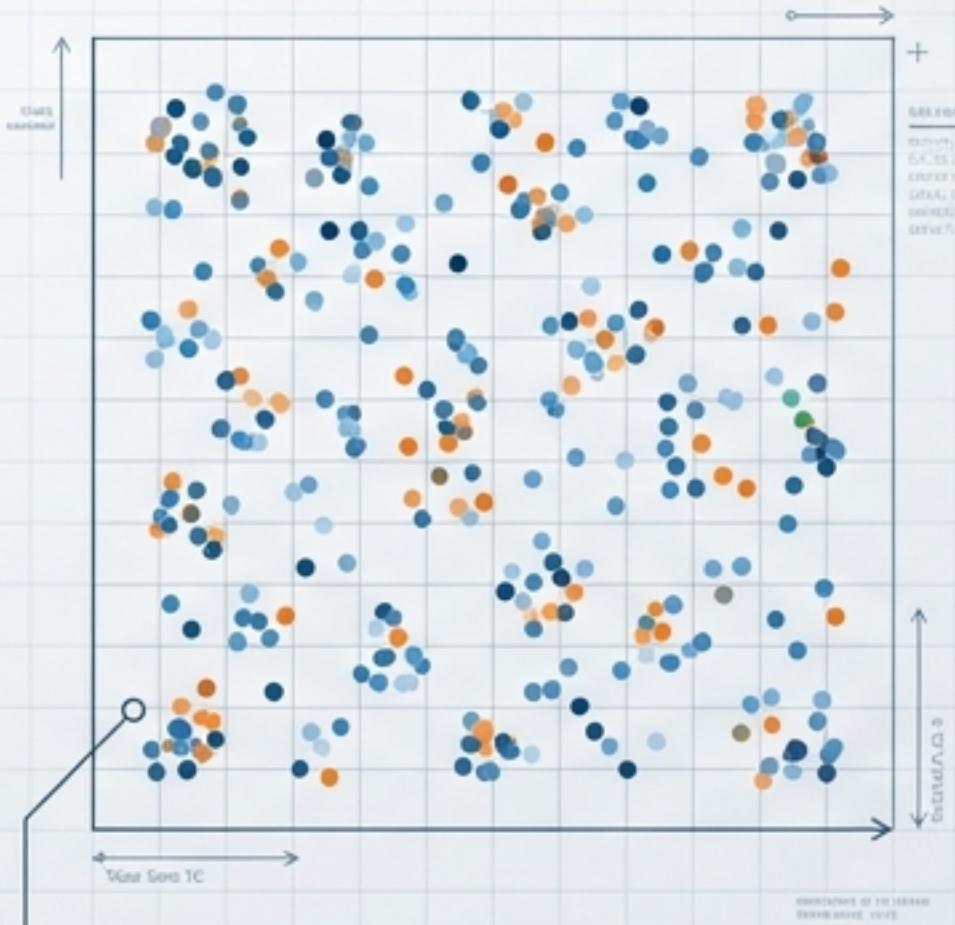
Roboto Nono

t-SNE 透過這種機制，在低維空間中有效地展示了數據的局部結構和群集 (clusters)。

核心超參數：困惑度 (Perplexity)

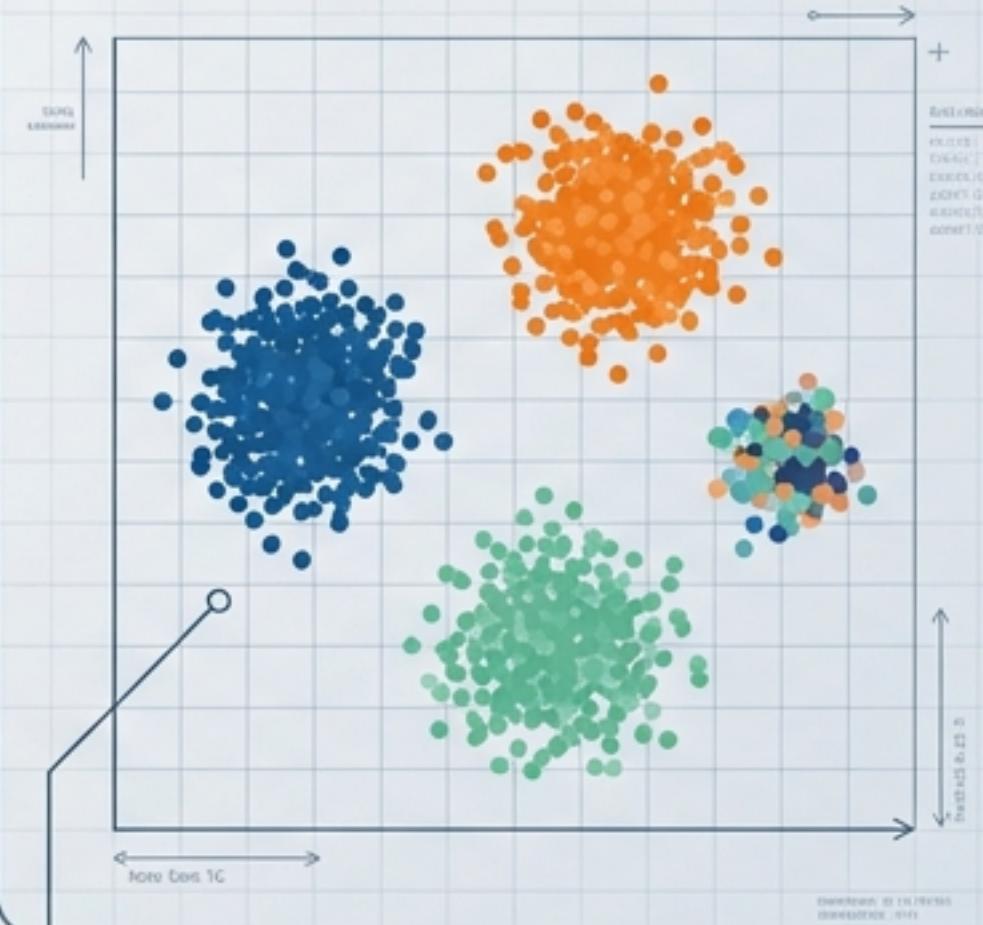
困惑度代表每個數據點的『有效鄰居數量』。通常建議值：5 - 50。

Perplexity = 5 (過低)



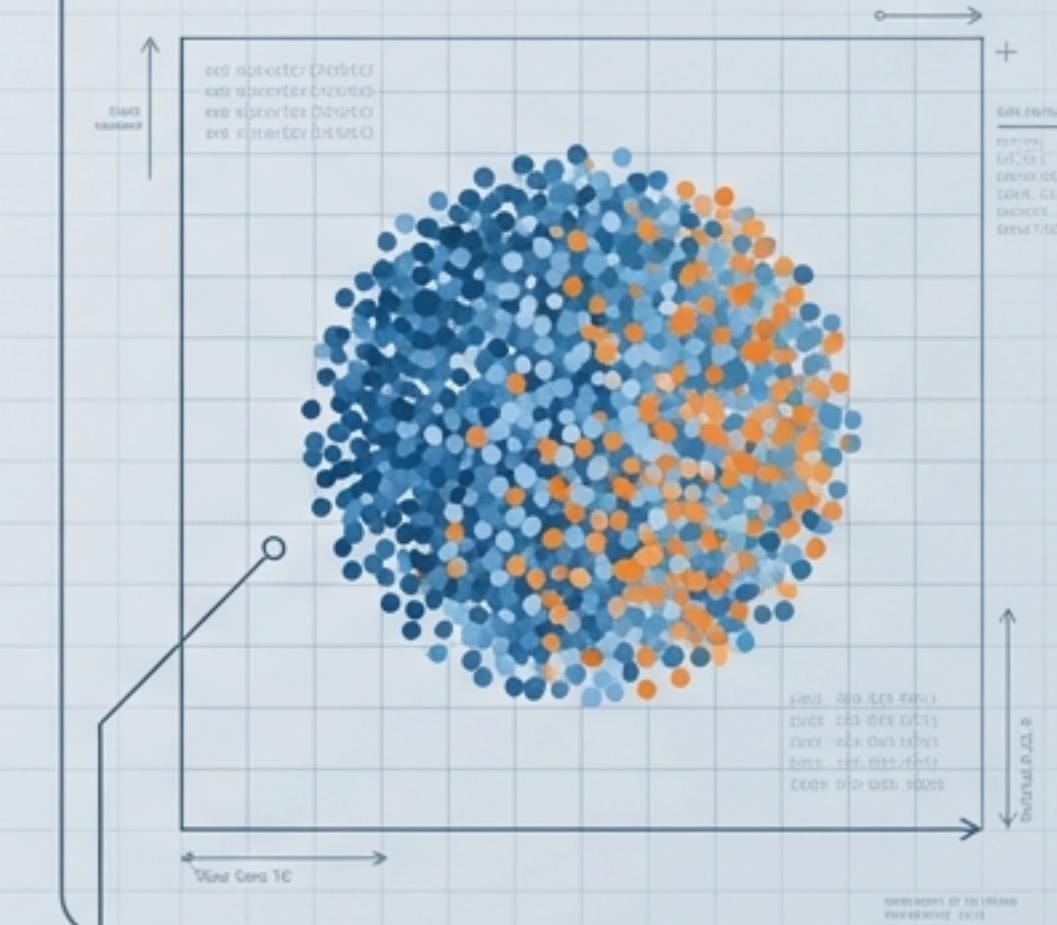
- 過度關注局部細節，導致群集破碎 (Noise)。

Perplexity = 30 (最佳)



- 平衡局部與全局結構，群集清晰可見。

Perplexity = 100 (過高)



- 過度平滑，喪失局部細節，僅保留全局形狀。

標準工作流程：從數據到洞察

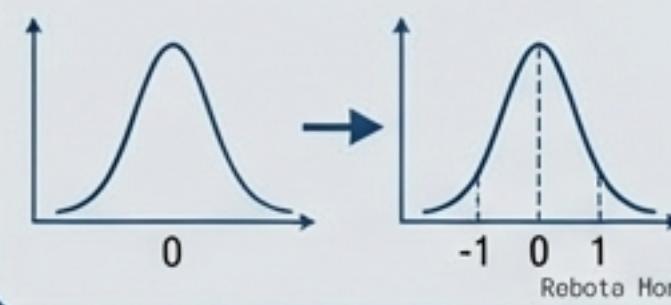
t-SNE 標準工作流程：從數據到洞察

1. 前處理 (Pre-processing)



標準化 (Standardization) 是必須的。t-SNE 基於距離計算，若變數尺度不同（如 壓力 vs 濃度），結果將無效。

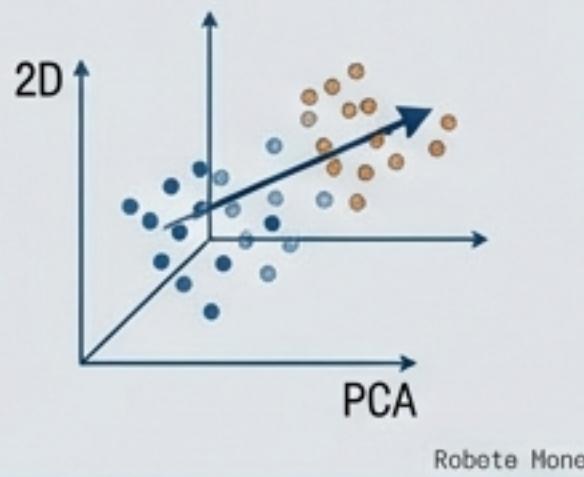
Formula: Z-score: $(x - \mu) / \sigma$



2. 初始化 (Initialization)



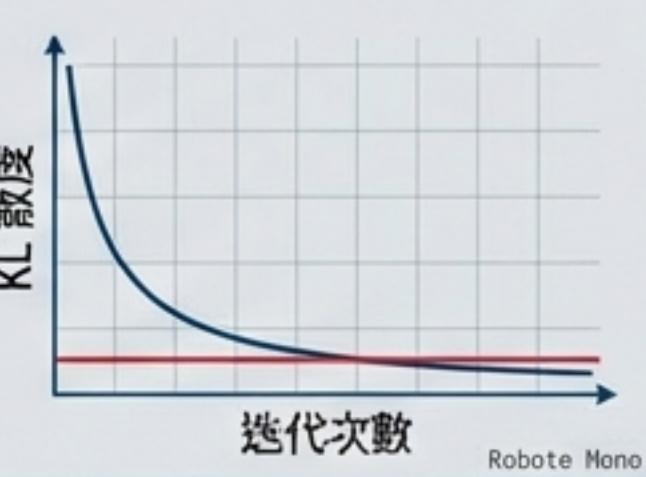
建議使用 PCA 初始化而非隨機初始化。提升結果的穩定性與可重複性，加速收斂。



3. 優化 (Optimization)



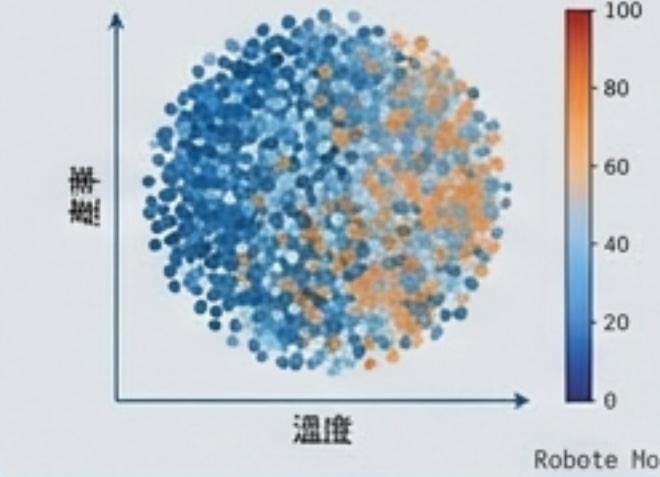
調整 Perplexity 與 Learning Rate。監控 KL 散度曲線，確保模型收斂 (至少 1000 次迭代)。



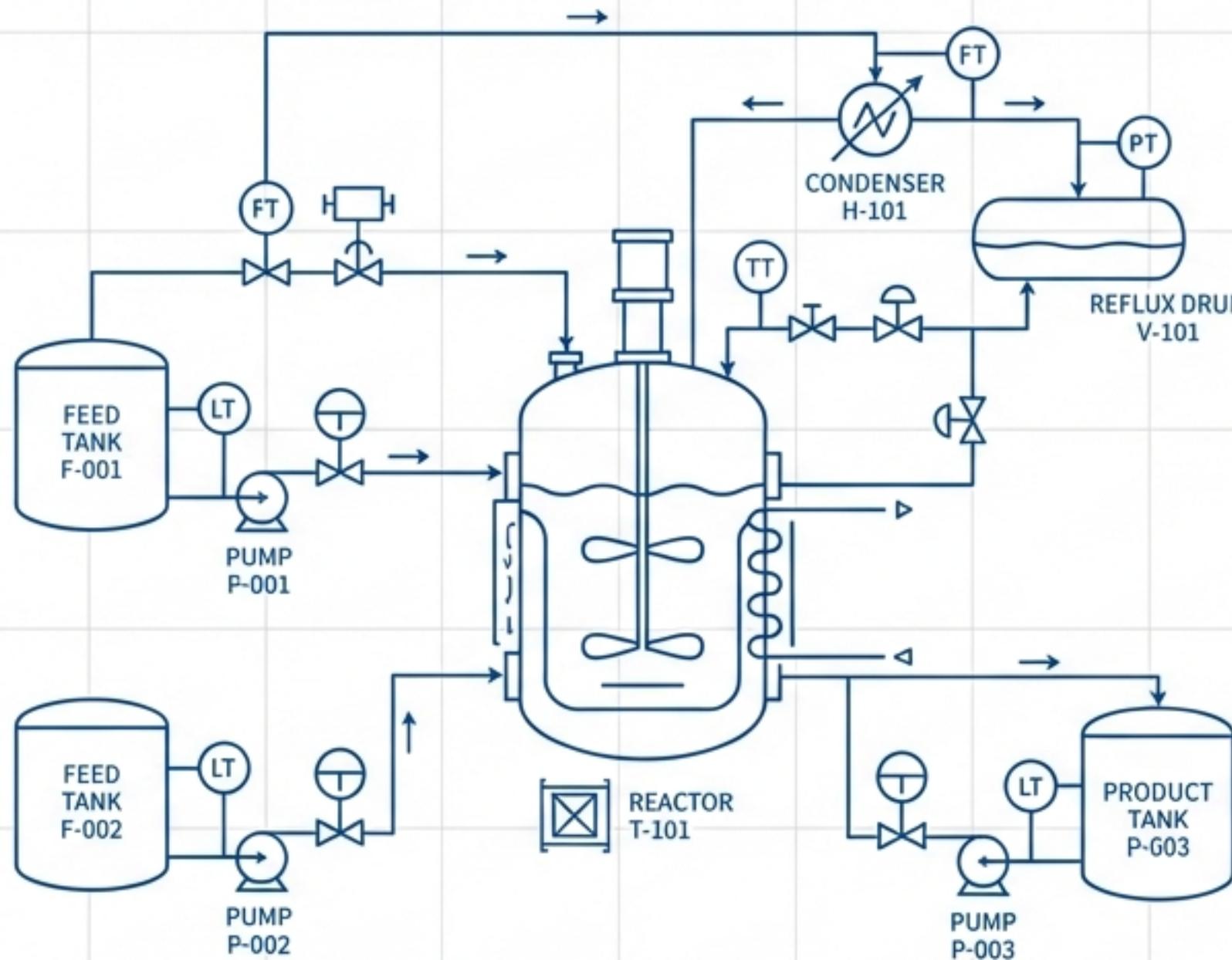
4. 視覺化 (Visualization)



將 2D 座標作圖，並根據物理性質（如產率、溫度）進行著色分析。



實戰案例：批次反應器操作模式識別 (Batch Reactor Case Study)



Data Specification

輸入數據 (Input Data):

- 10 個製程變數 (溫度，壓力，流率，濃度等)
- 樣本數：600 個批次

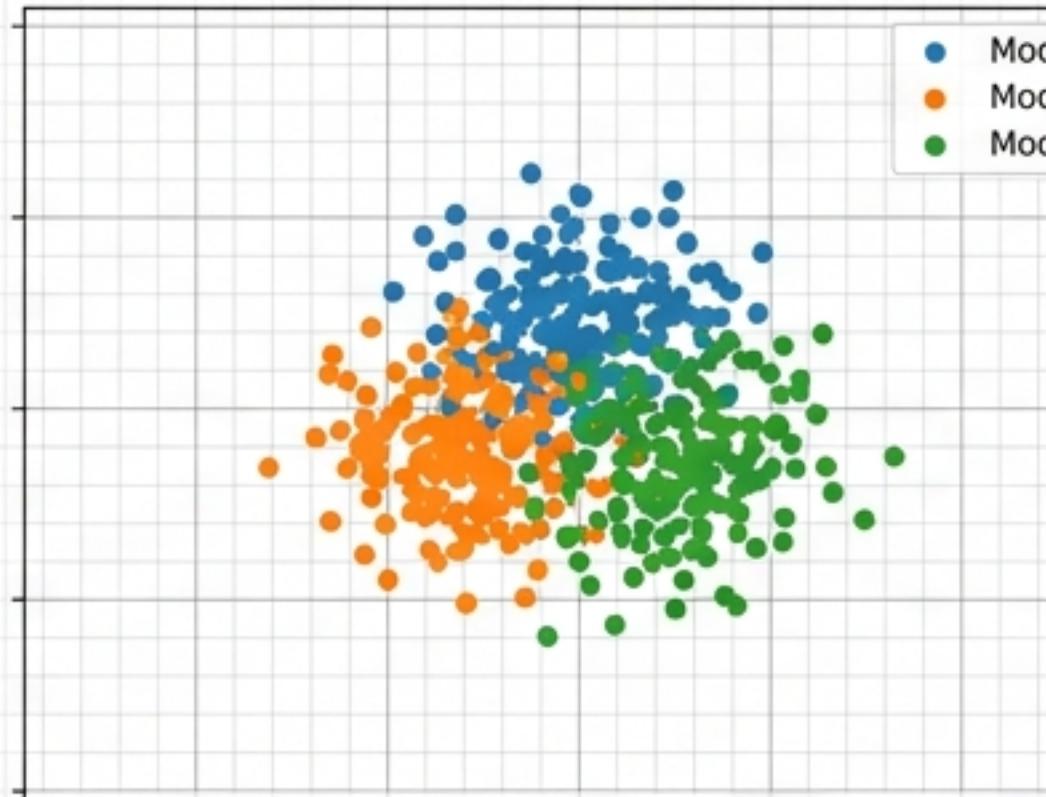
Hidden Modes

- Mode A (Blue Dot) : 低溫低壓 (Low T, Low P) – 操作穩定，產率較低。
- Mode B (Orange Dot) : 高溫高壓 (High T, High P) – 高產率，但高能耗且不穩定。
- Mode C (Green Dot) : 中間過渡態 (Transitional) – 平衡性能。

挑戰：能否在不預先知道標籤的情況下，透過數據自動發現這三種操作模式？

降維效果比較：PCA vs t-SNE (Dimensionality Reduction Comparison: PCA vs t-SNE)

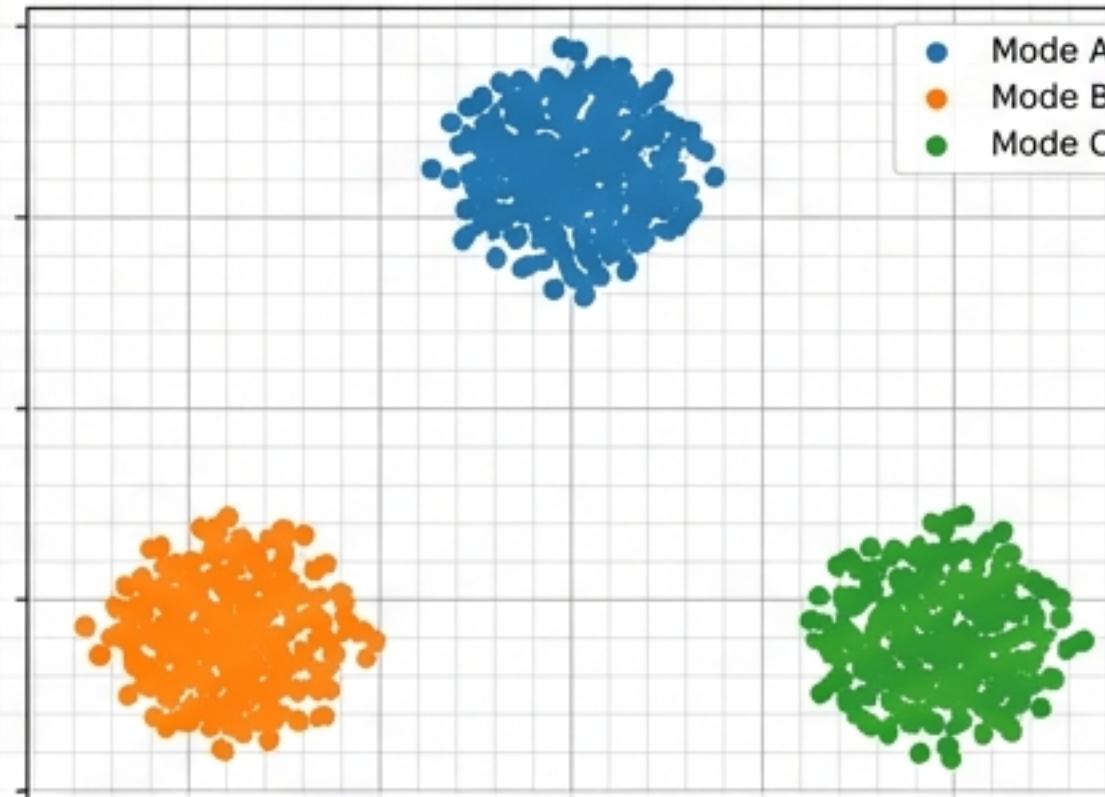
PCA 結果 (2D)



Silhouette Score: 0.695

分析：僅能捕捉線性變異，群集邊界模糊，難以區分 Mode C 與其他模式。

t-SNE 結果 (2D)



Silhouette Score: 0.802 (+15.4%)

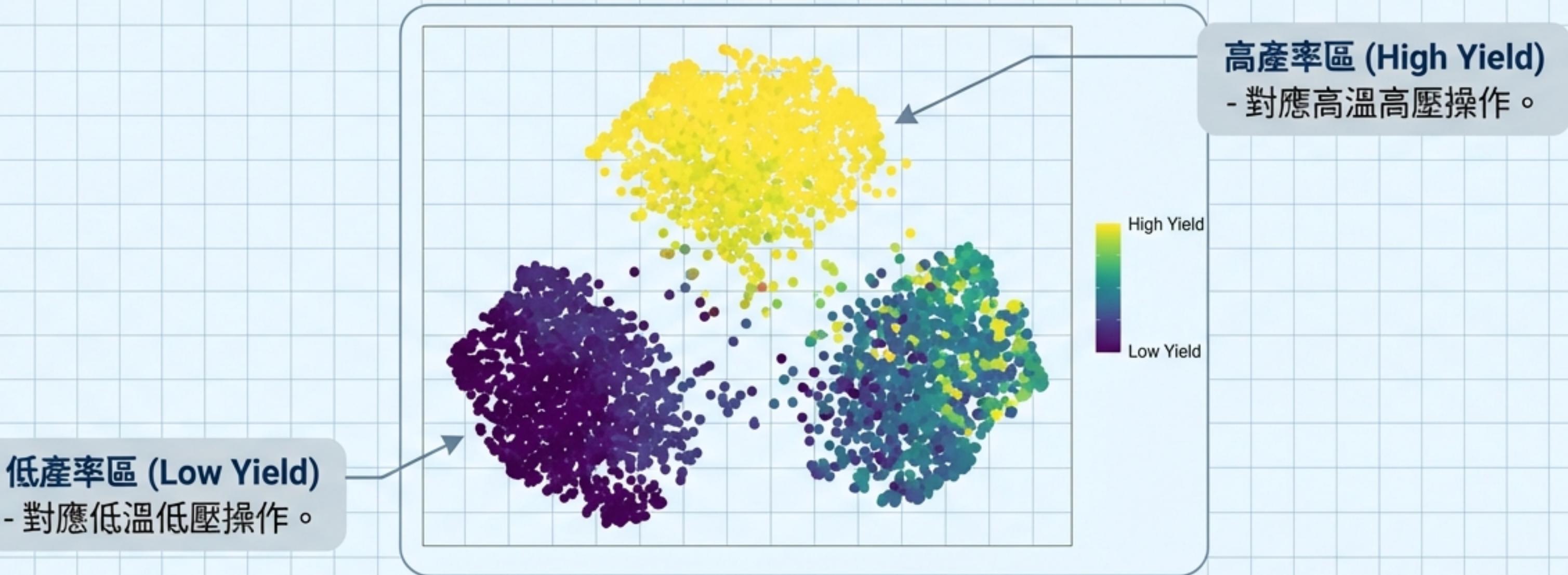
分析：成功捕捉非線性結構，三種操作模式完全分離 (Clean Separation)。

對於複雜的化工數據，t-SNE 提供了比 PCA 更清晰的視覺區分度。

10 mm

Reboot: Mono

解讀地圖：產率與操作模式的關聯



t-SNE 不僅能分群，還能視覺化連續變數（如產率、能耗）在操作空間中的分布。
顏色呈現平滑過渡，證實 t-SNE 成功保留了數據的局部物理特性。

現實世界的限制 (Reality Check)



優勢 (The Good)

- 視覺化效果極佳 (Best-in-class Visualization)
- 能處理高度非線性數據
- 不需要標籤 (Unsupervised)



限制 (The Bad)

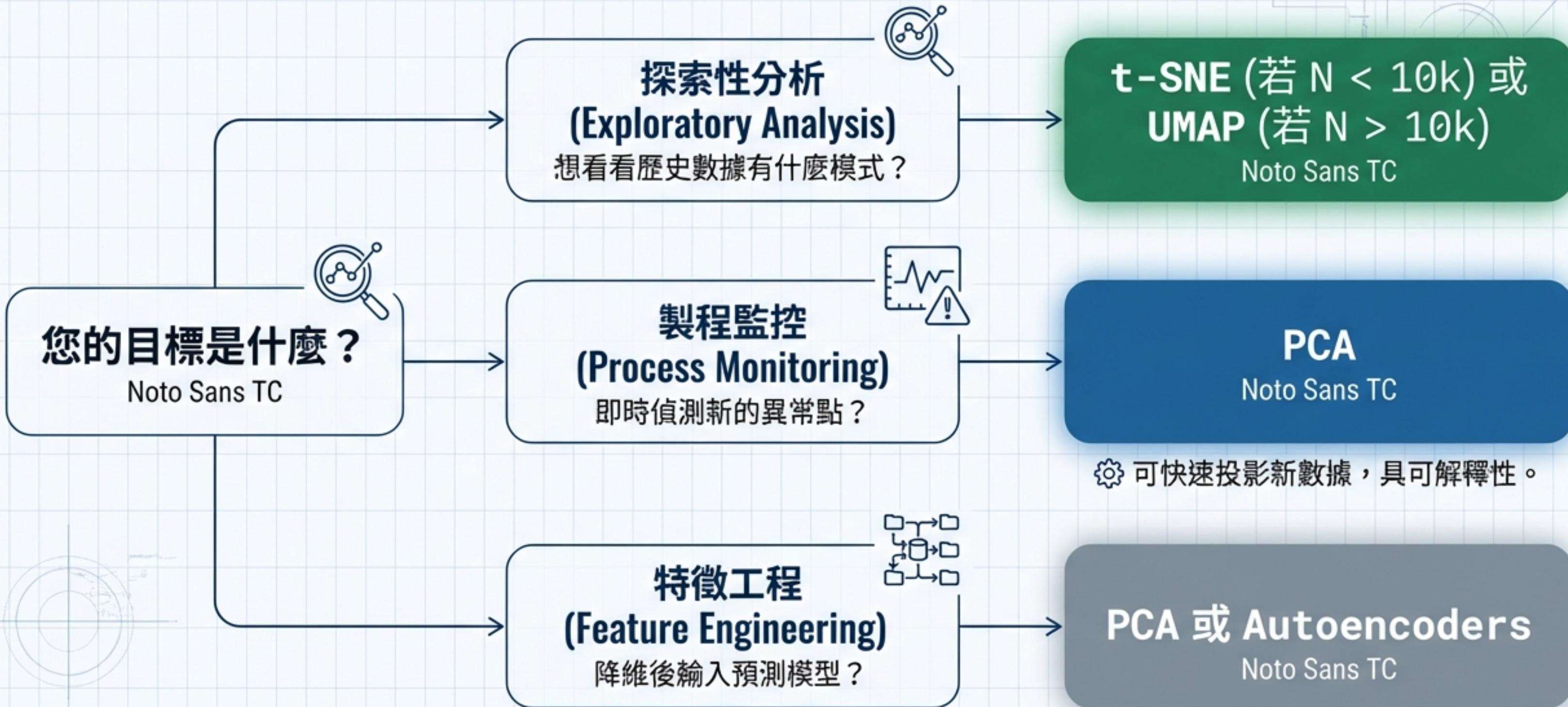
- **計算成本高**：複雜度 $O(N^2)$ ，數據量 $> 10,000$ 時極慢。
- **無法投影新數據**：新數據進來必須重新跑整個演算法。不適合即時線上監控。



注意事項 (Note)

- **距離無絕對意義**：群集之間的距離不代表實際相似度，僅局部距離有意義。

如何選擇正確的工具？(Choosing the Right Tool)



化工師的最佳實踐 (Best Practices)



Pre-processing is Key: 不同單位的感測器數據（溫度 K vs 壓力 bar）必須先進行 **Z-Score 標準化**。



Handling Big Data : 若變數 > 50 ，先用 PCA 降至 30-50 維，再做 t-SNE (PCA 初始化)。這能去除噪音並加速計算。



Stability : t-SNE 具有隨機性。務必設定 **Random Seed (隨機種子)** 以確保結果可重複，這對於工程報告至關重要。



Interpretation : 不要過度解讀群集之間的距離。專注於群集的形成與內部的密度。

進階前沿：UMAP

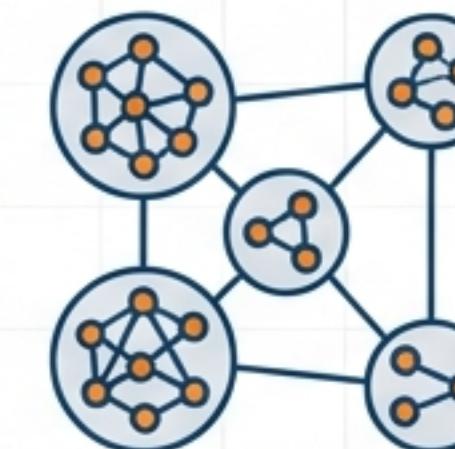
UMAP (Uniform Manifold Approximation and Projection)

- 適用情境：當數據量超過 10,000 筆，或需要保留更多全局結構時。



Speed

速度比 t-SNE 快得多
 $(O(N \log N))$ 。



Structure

更好地保留了群集之間的相對位置（全局結構）。

t-SNE 是黃金標準，但 UMAP 是大數據時代的強大後繼者。

結語：賦予數據視覺的力量

t-SNE 讓我們能夠『看見』反應器的內部狀態，將枯燥的感測器數值轉化為直觀的操作地圖。

影響力 (Impact)



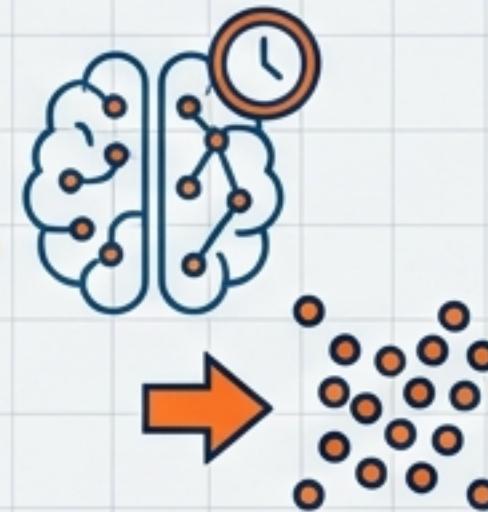
- 加速異常診斷
(Faster Diagnosis)



- 優化操作策略
(Operational Strategy)

下一步 (Next Step) :

前往 Unit 07 : UMAP - 更快、更強大的降維工具。



Roboto Mono

“AI 不會取代化工工程師，但懂得使用 AI 工具（如 t-SNE）的工程師將無可替代。”

