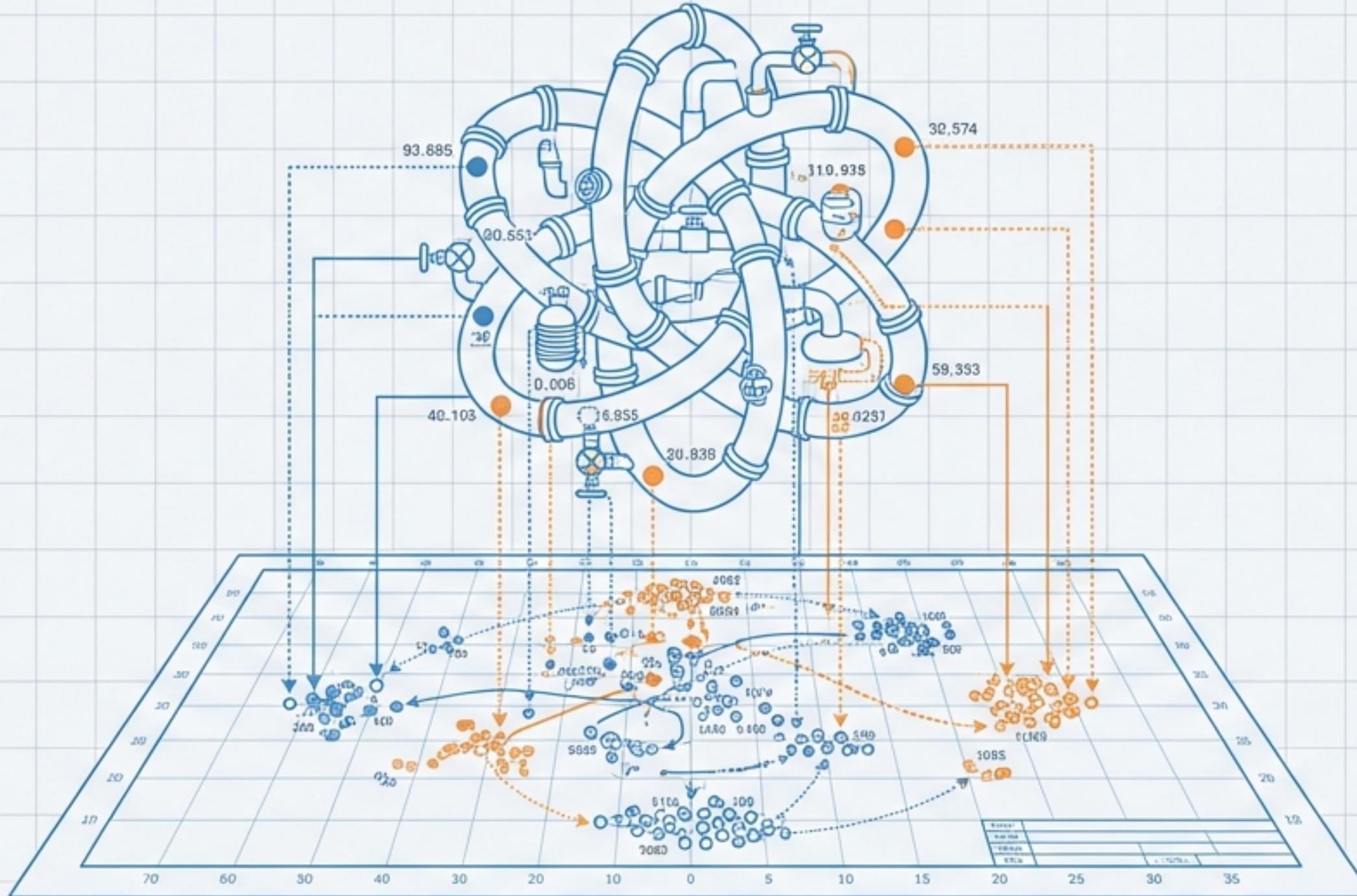


Unit 06: t-分布隨機鄰域嵌入 (t-SNE)

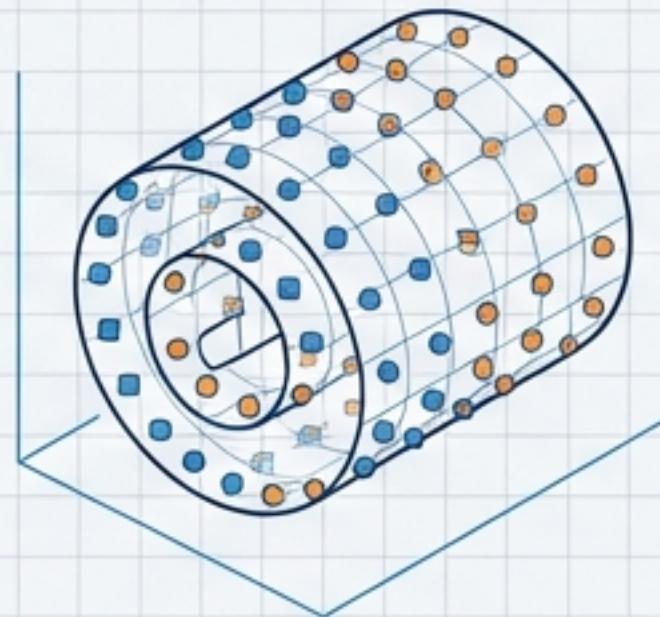
高維化工數據的非線性視覺化技術



高維數據的視覺化挑戰：看見非線性結構

- 化工製程涉及數十個變數（溫度 T、壓力 P、流率 F、濃度 C）。
- **維度詛咒 (The Curse of Dimensionality)**：
高維空間十分稀疏，距離度量變得反直覺。
- **線性限制：**
線性降維（如 PCA）無法有效展開彎曲、螺旋等複雜流形結構（Manifolds）。

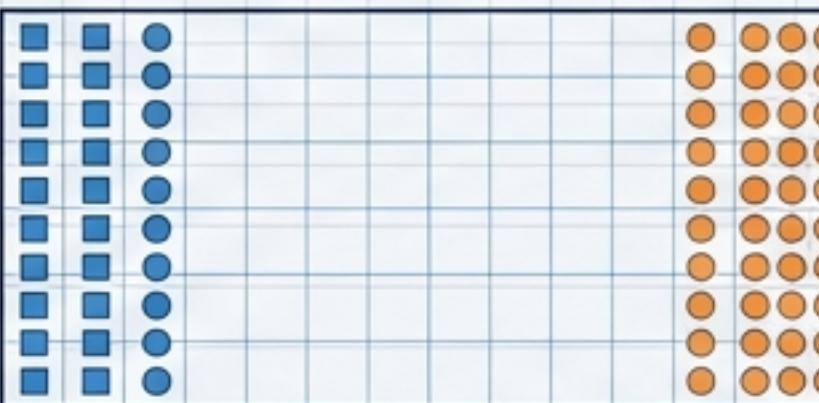
The Problem



3D Manifold
(Entangled)

The Goal

目標：將高維空間中「相似」的點，在低維地圖上聚在一起



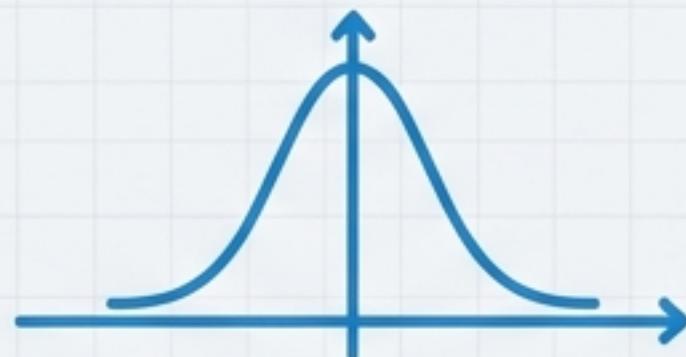
2D Unfolded
(Manifold Learning)

t-SNE 的核心理念：保持局部鄰域結構

定義：一種非線性降維技術，由 van der Maaten & Hinton (2008) 提出，專注於保持數據點之間的「局部鄰域」關係。

1. 高維空間

高斯分布 Gaussian



2. 低維空間

t-分布 t-Distribution

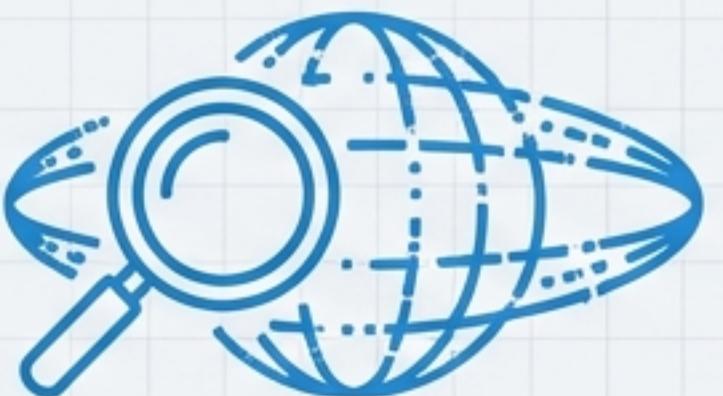


3. 優化目標

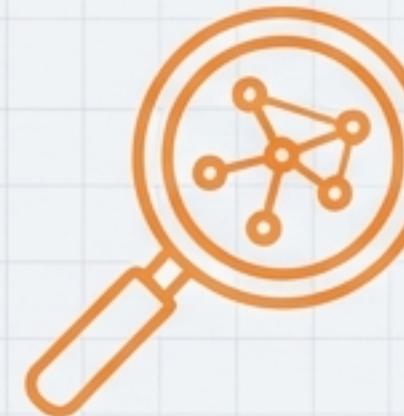
KL 散度



PCA：全局變異數最大化 (Global Variance)



t-SNE：局部相似性保持 (Local Neighborhood)



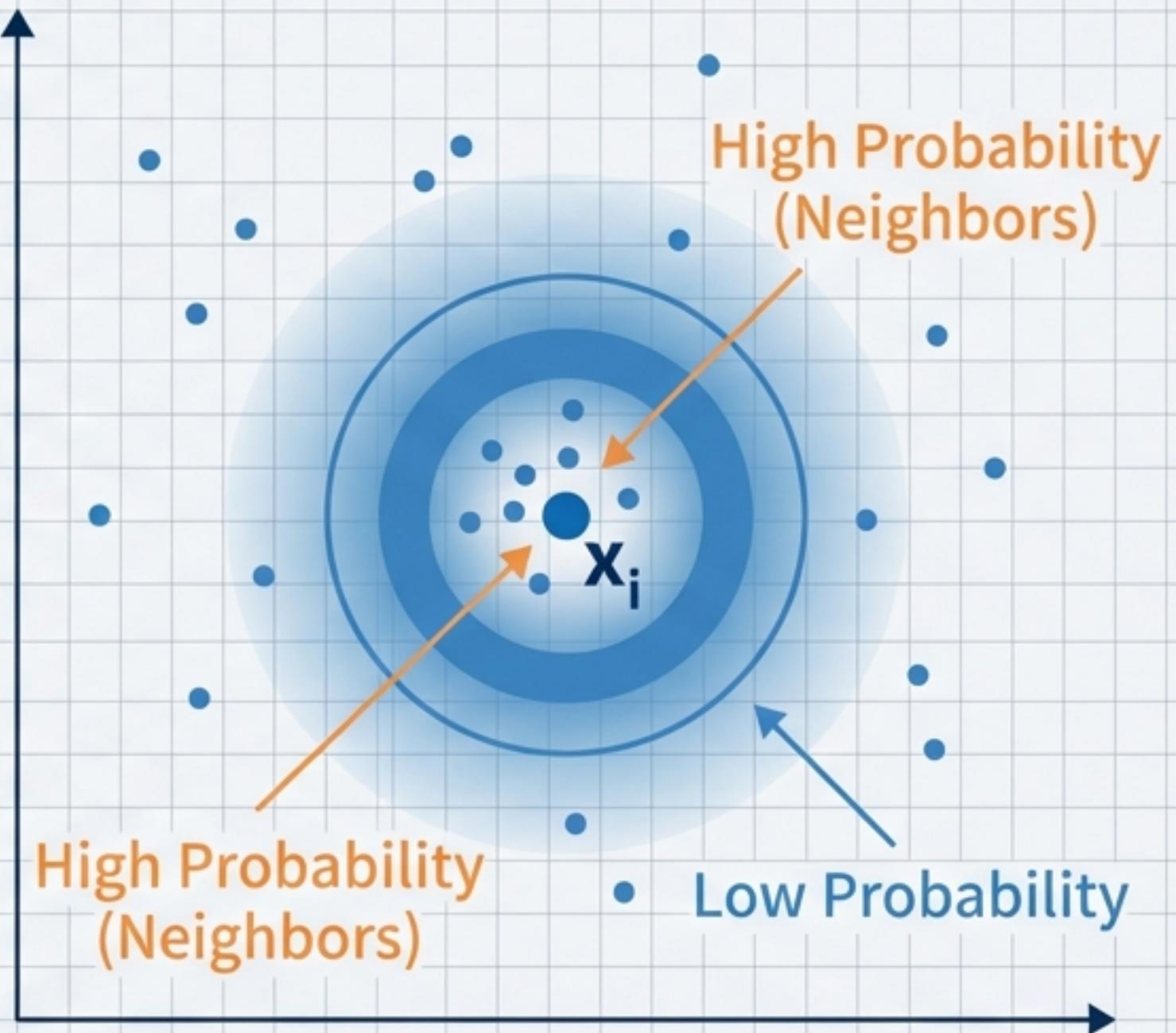
機制解析 (1)：高維空間的相似度度量

條件機率 (Conditional Probability)

在高維空間，我們不直接看歐氏距離，而是轉化為「機率」。如果 x_j 是 x_i 的鄰居，其機率密度就高。

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

σ_i (帶寬) 自動適應局部密度

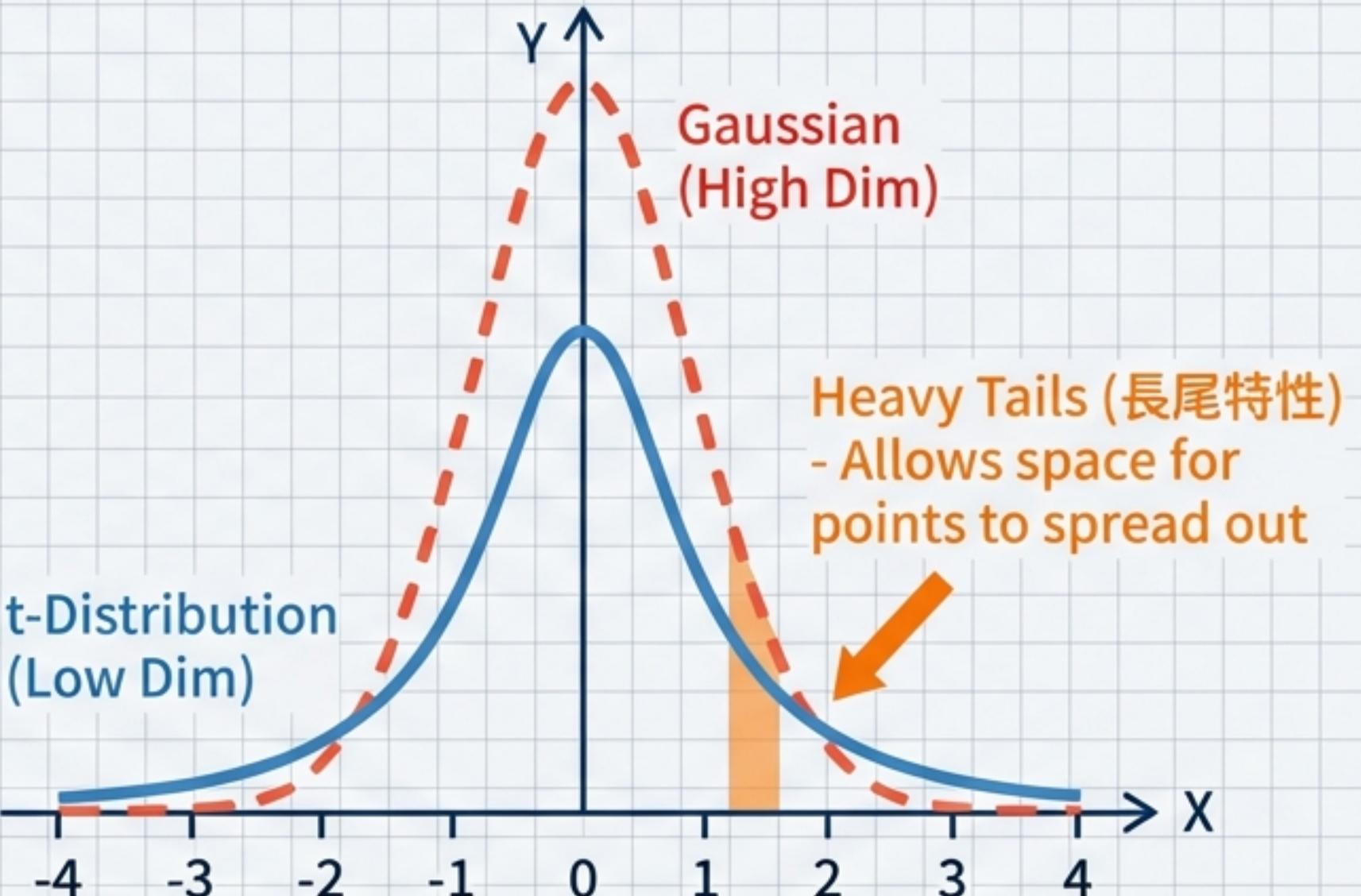


機制解析 (2)：低維映射與擁擠問題

擁擠問題 (Crowding Problem)

高維空間的體積遠大於低維空間。當我們把高維球體壓扁到 2D 時，中等距離的點會被迫「擠」在一起，破壞群集結構。

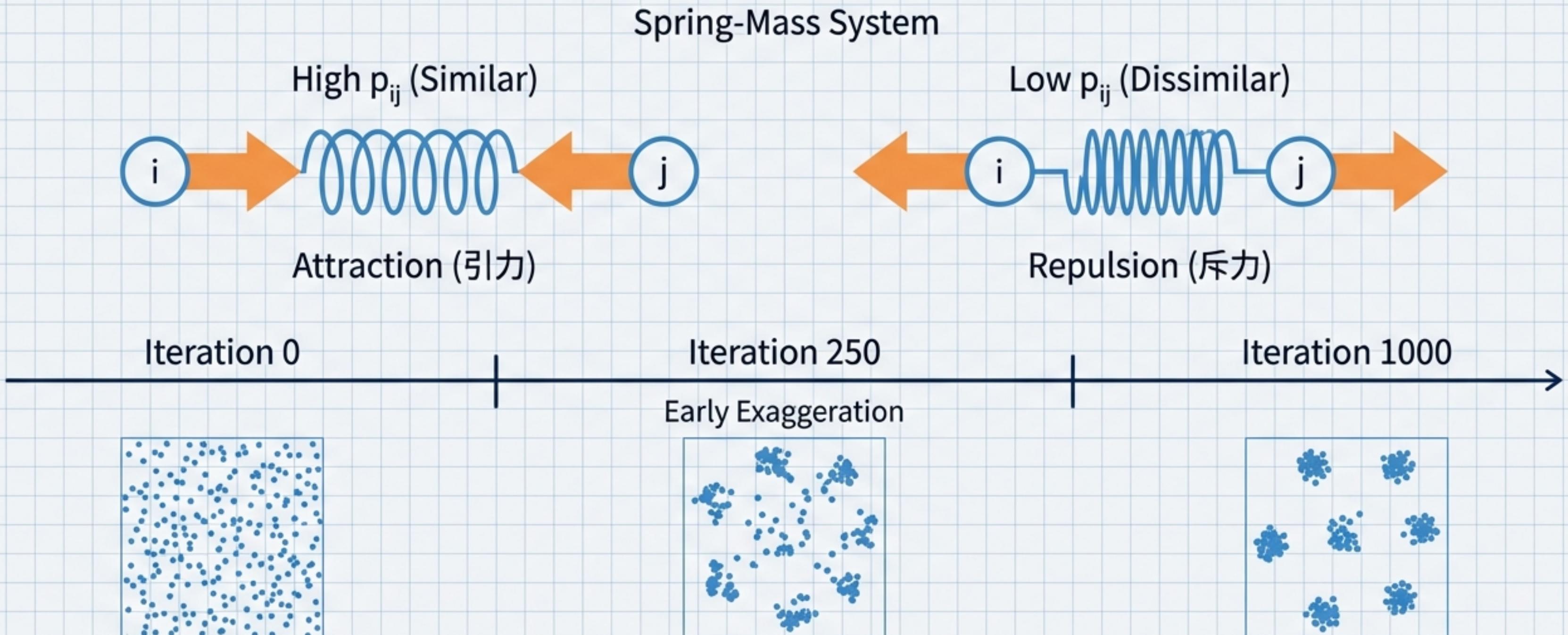
解決方案：在低維空間使用 Student t-分布 (1自由度)



$$q_{ij} \propto (1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}$$

優化引擎：KL 散度與梯度下降

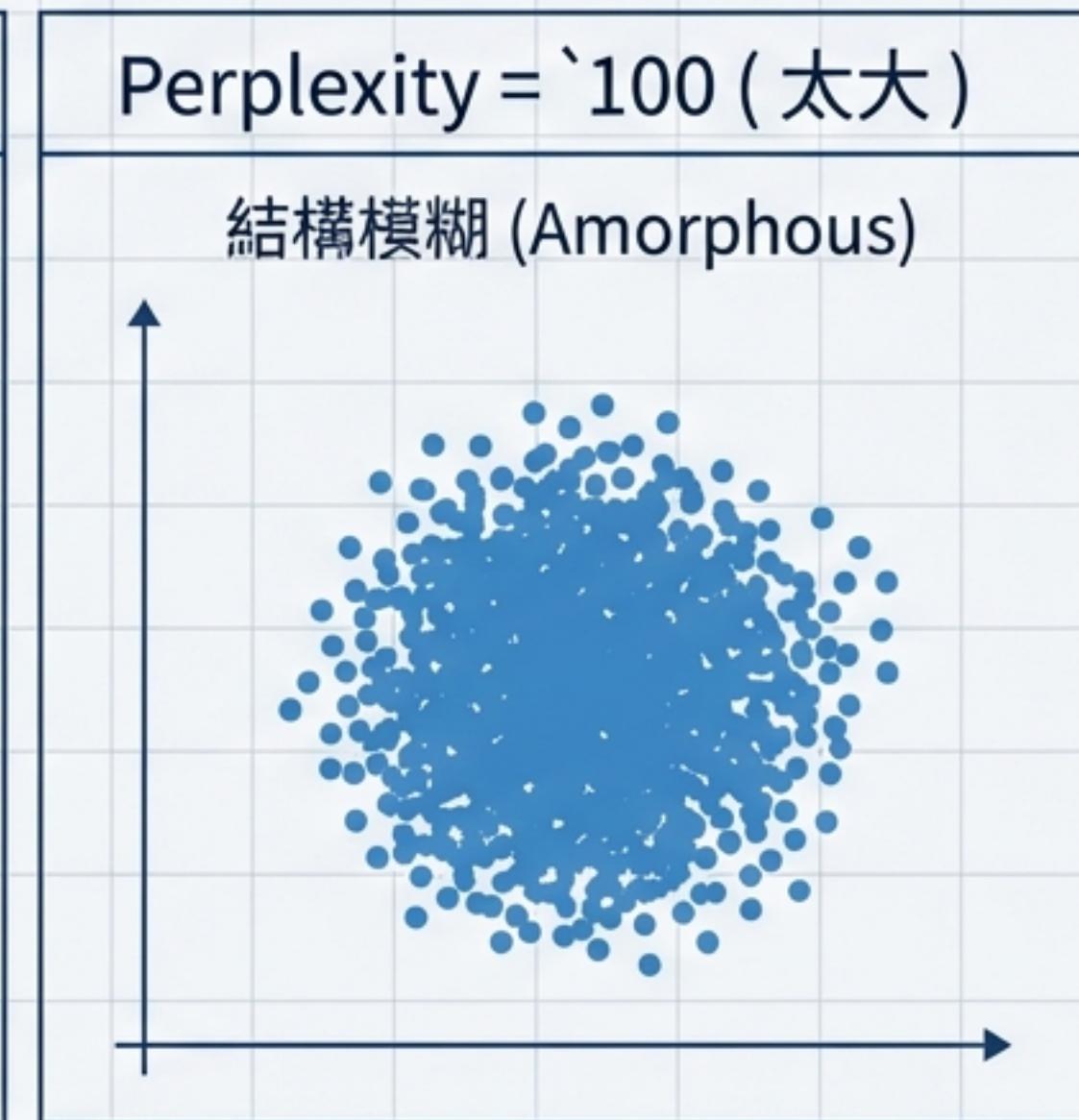
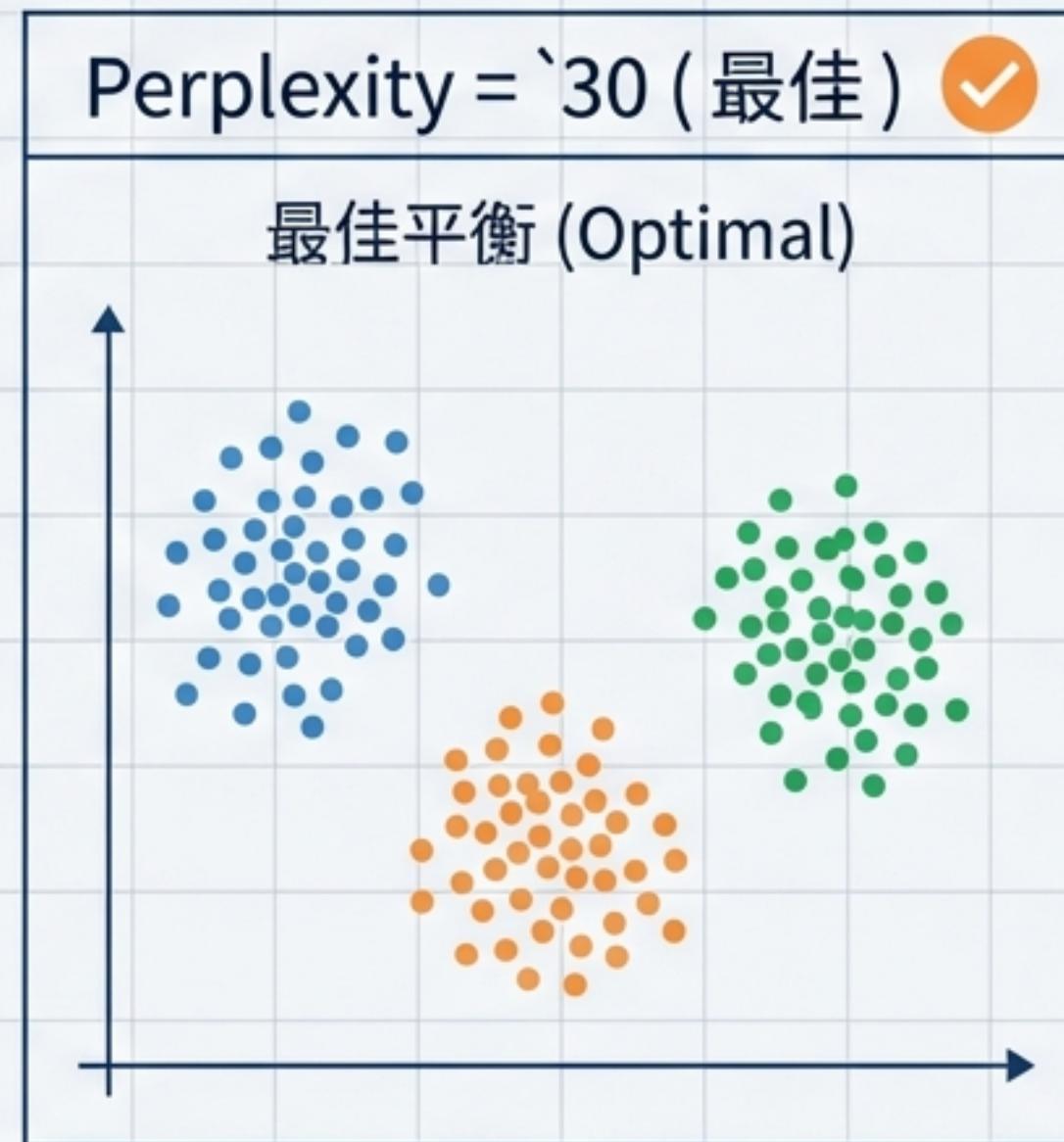
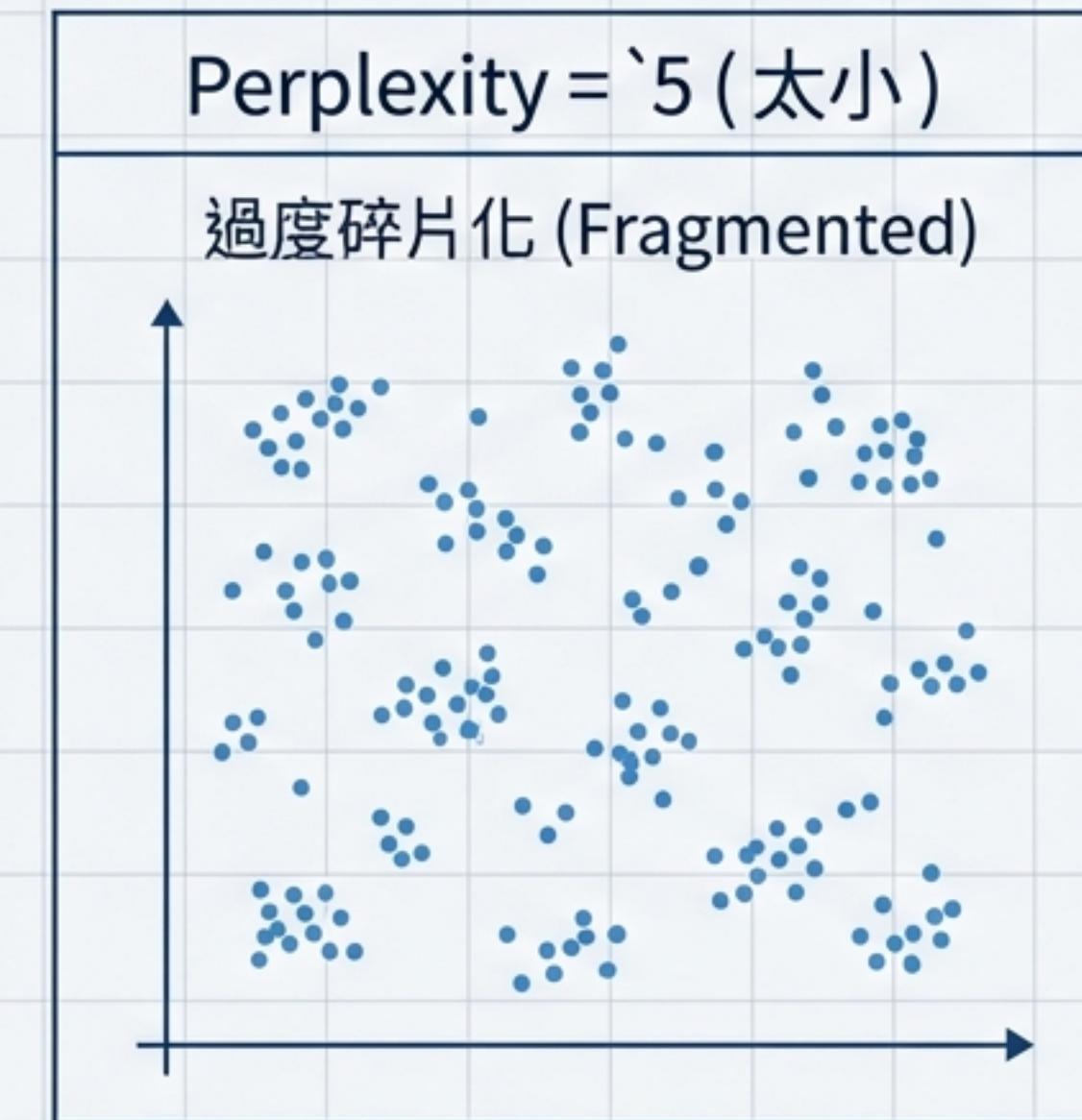
目標函數：最小化高維分布 P 與低維分布 Q 之間的差異。 $KL(P || Q) = \sum p_{ij} \log(p_{ij}/q_{ij})$



關鍵參數控制 (1)：困惑度 (Perplexity)

定義：困惑度 \approx 每個數據點的「有效近鄰數量」。這是最重要的調優參數。

建議範圍：5 - 50 (預設值: 30)

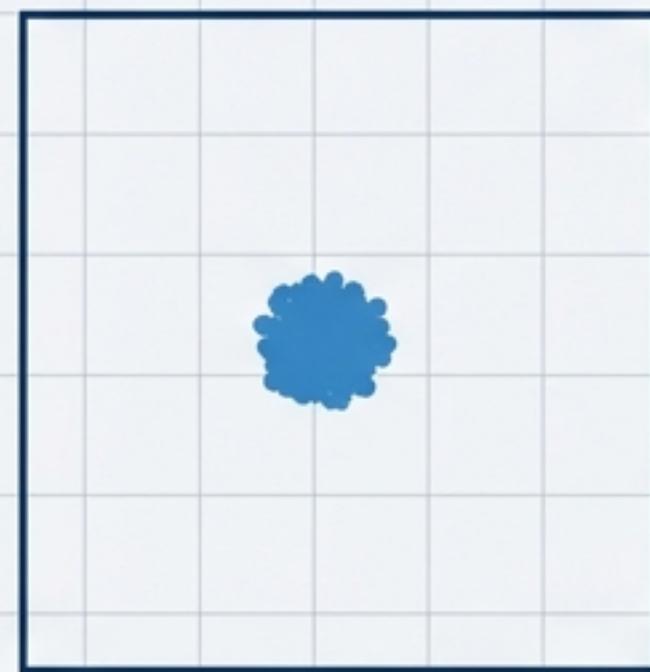


Trade-off: 平衡局部細節 (Local) 與全局結構 (Global)。

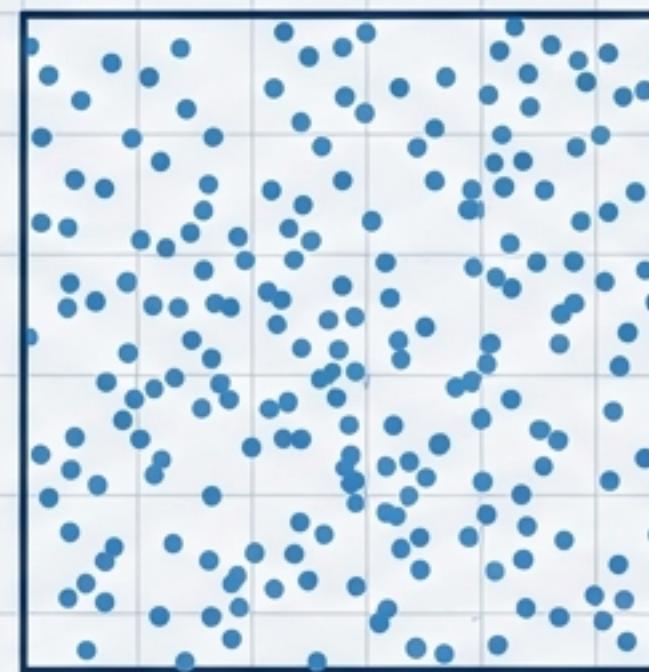
關鍵參數控制 (2)：學習率與迭代次數

學習率 (Learning Rate)

- 控制梯度下降的步長。
- Range: 100 – 1000 (Default: 'auto').



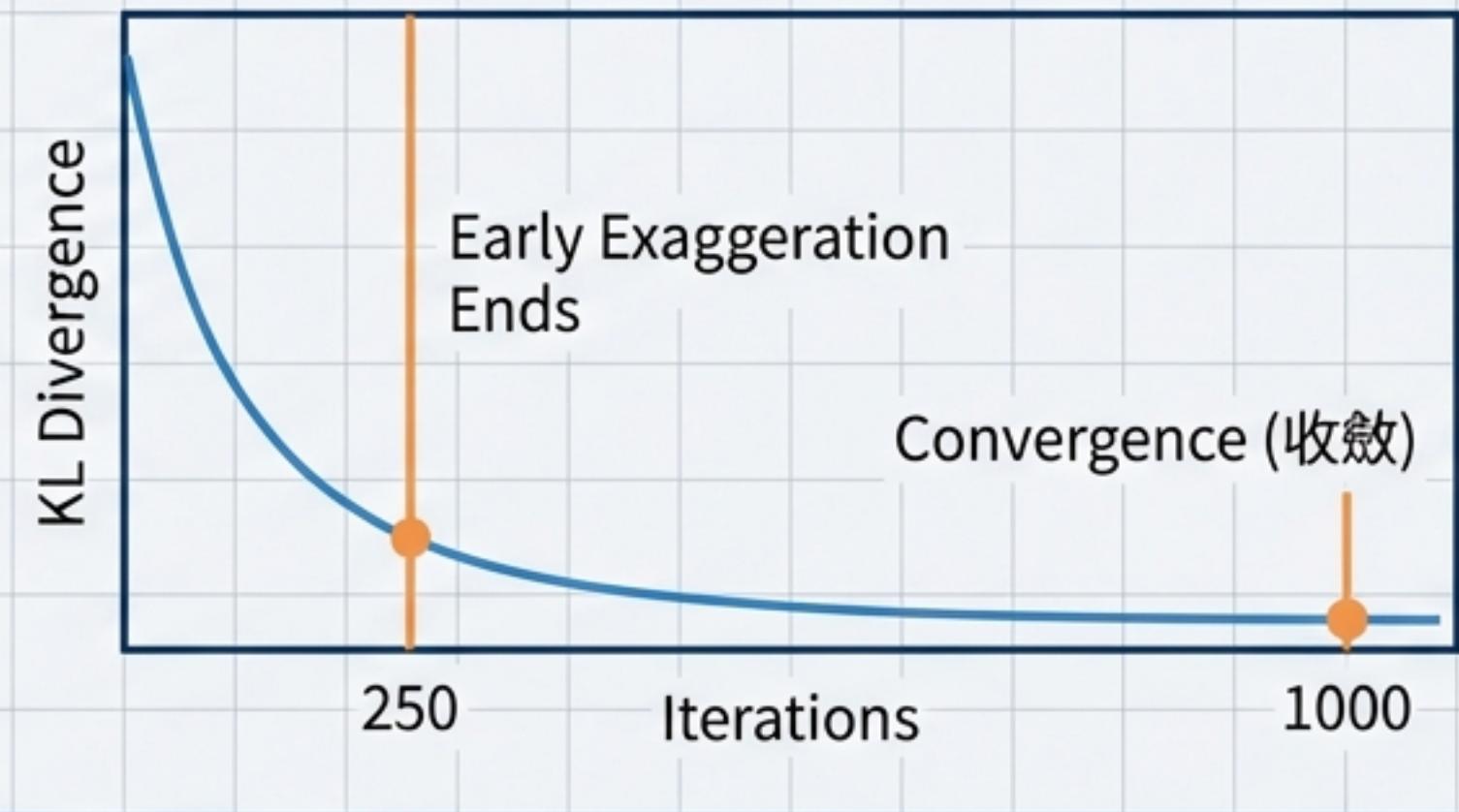
Too Low (太小)



Too High (太大)

迭代次數 (Iterations)

- t-SNE 需要足夠時間「解開」糾結。
- Guidance: 至少 1000 次。



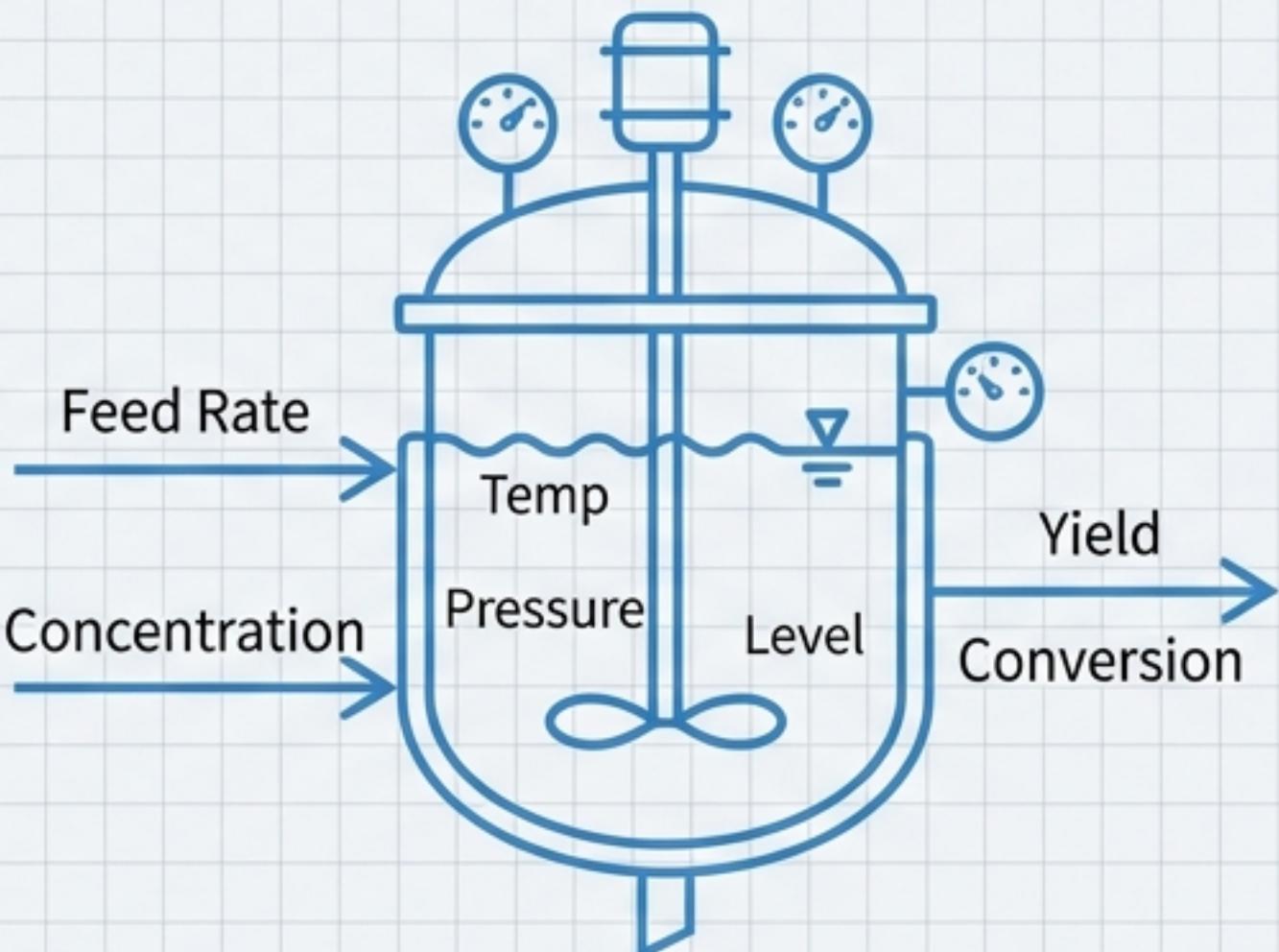
Initialization: 推薦使用 'pca' 初始化而非 'random'，結果更穩定。

實作案例：批次反應器多操作模式模擬

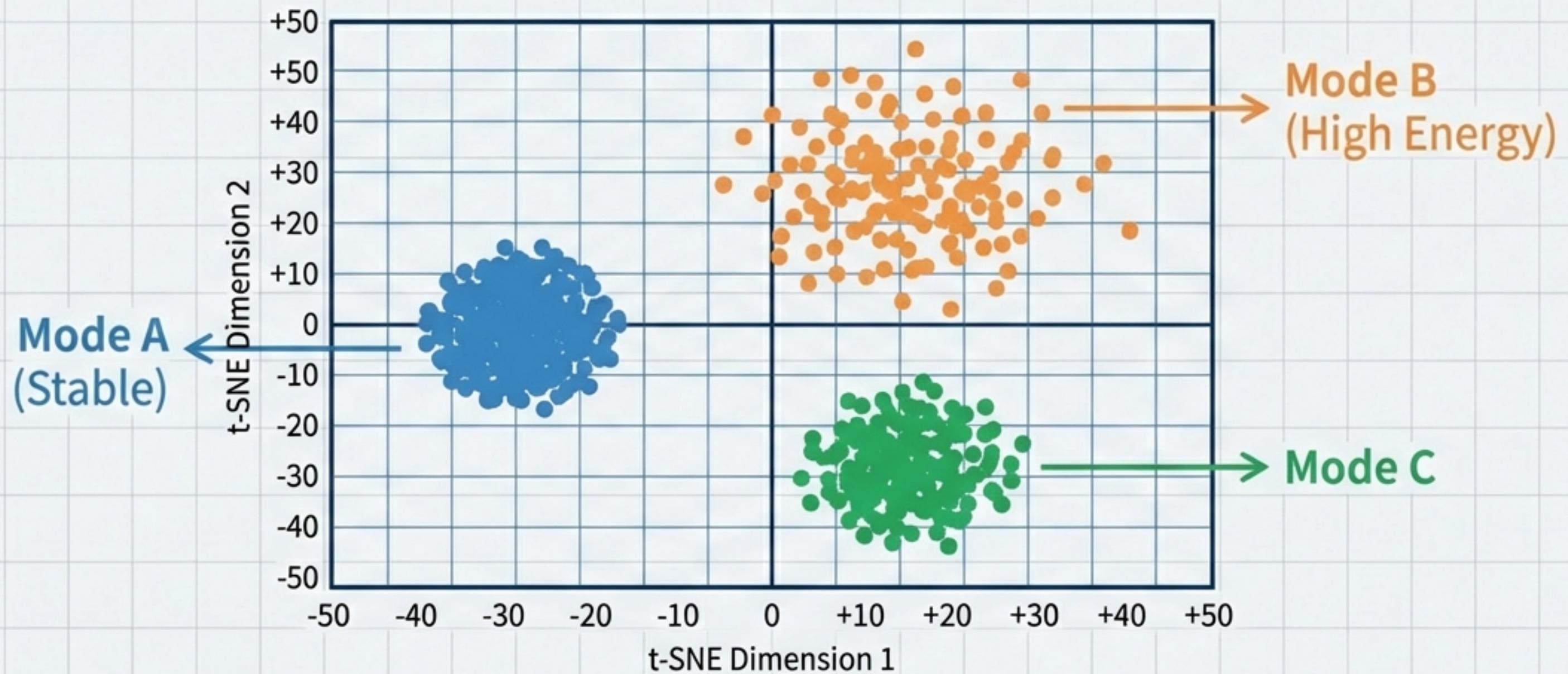
場景：一個批次反應器 (CSTR) 在 3 種不同模式下運行。

數據規格： $N=600$ 樣本, $D=10$ 製程變數。

- **Mode A (Blue)**: 低溫低壓 (Low T/P), 高穩定性。
- **Mode B (Orange)**: 高溫高壓 (High T/P), 高產率但高變異。
- **Mode C (Green)**: 中溫中壓 (Mid T/P), 平衡模式。

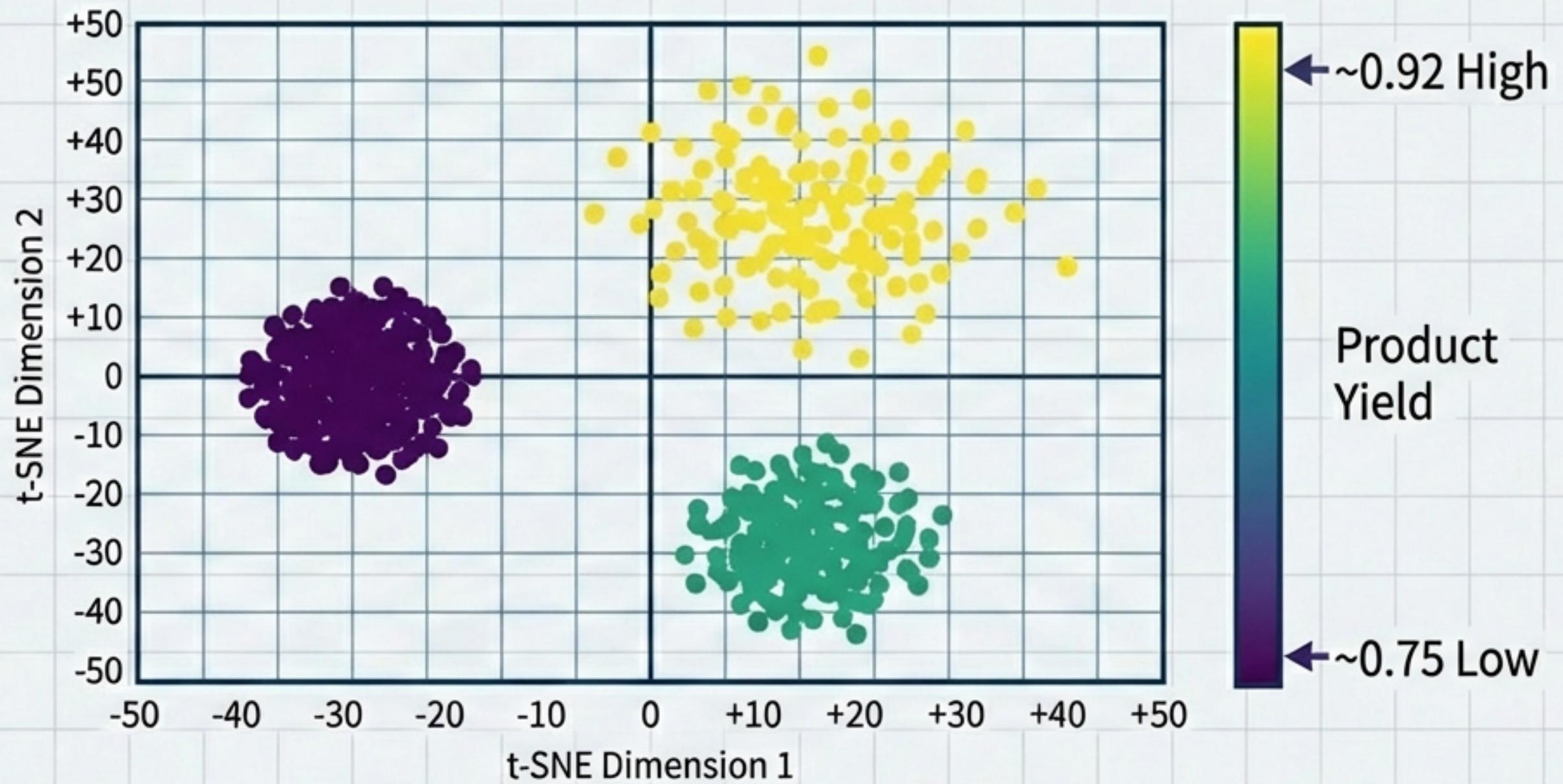


視覺化結果：操作模式的清晰分離



洞察：t-SNE 成功將 10 維數據降至 2 維。Mode B 的分散形狀反映了高溫高壓操作下的不穩定性 (High Variance)。

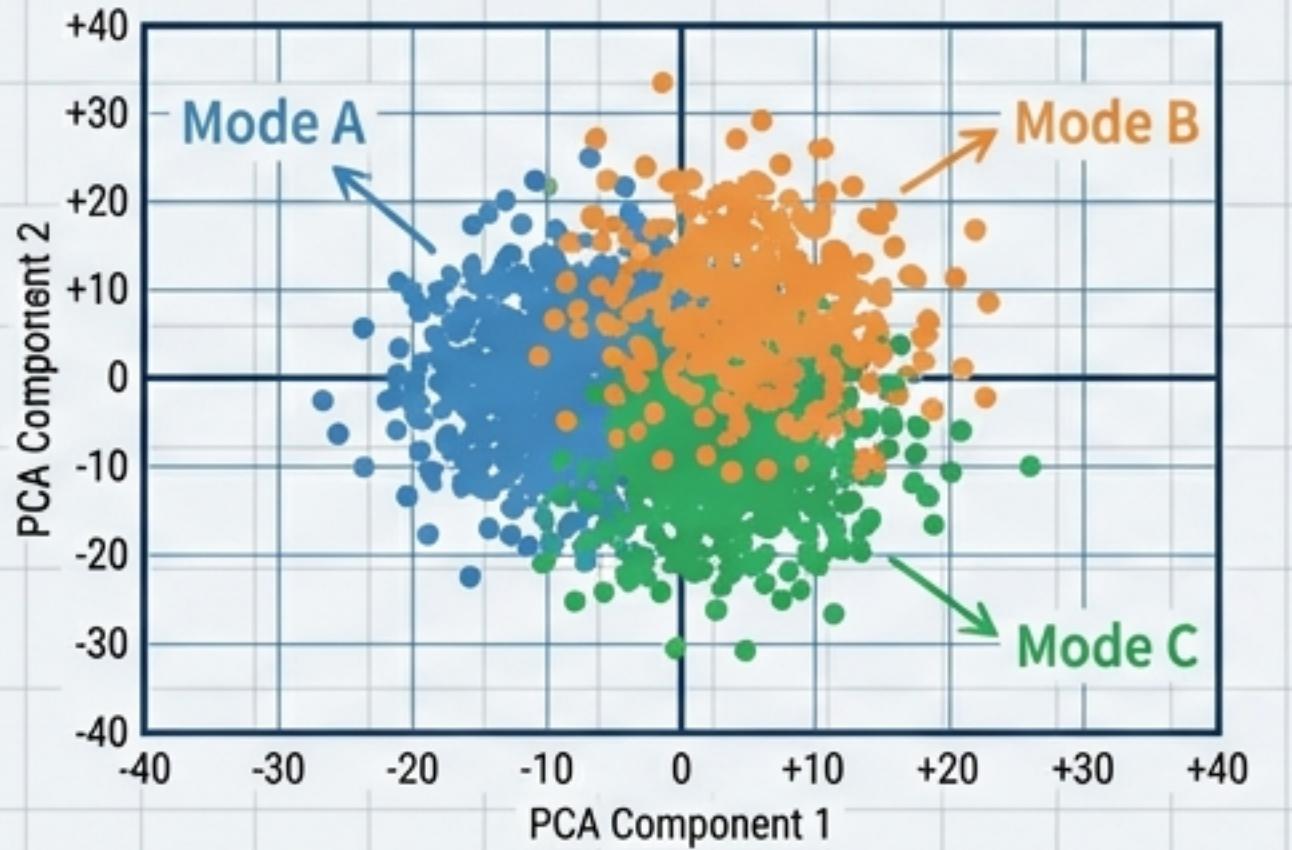
特徵映射：產物產率 (Yield) 分析



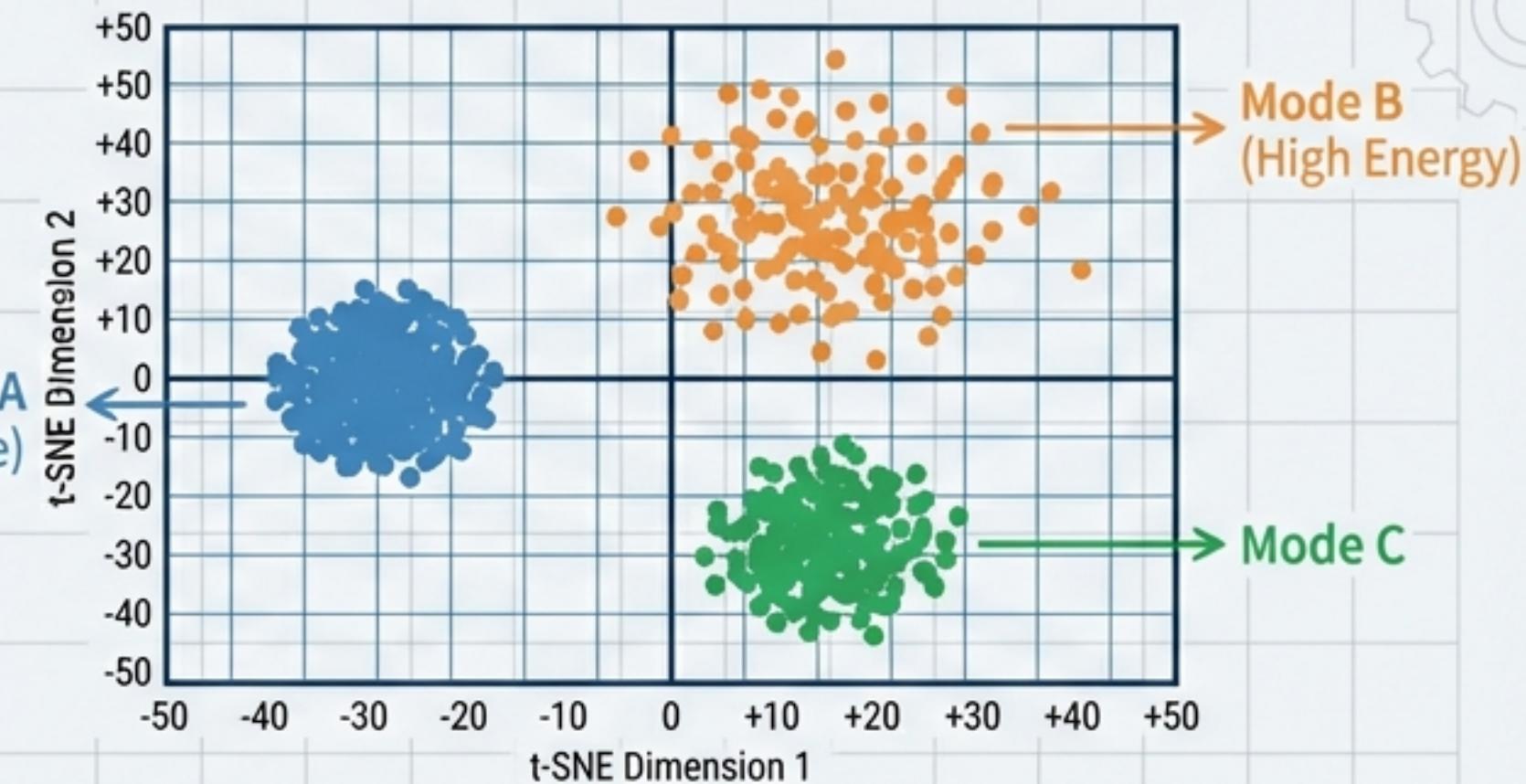
物理意義：顏色變化平滑連續。t-SNE 揭示了「高風險高回報」的權衡—Mode B 雖然變異大（不穩定），但產率最高。

效能對決：t-SNE vs. PCA

PCA Result



t-SNE Result



關鍵指標

Roboto Mono
Noto Sans TC

Silhouette Score: 0.695
Separation: Blurred (模糊)

關鍵指標

Roboto Mono
Noto Sans TC

Silhouette Score: **0.802 (+15.4%)**
Calinski-Harabasz: **7807 (+151.8%)**
Separation: Sharp (清晰)

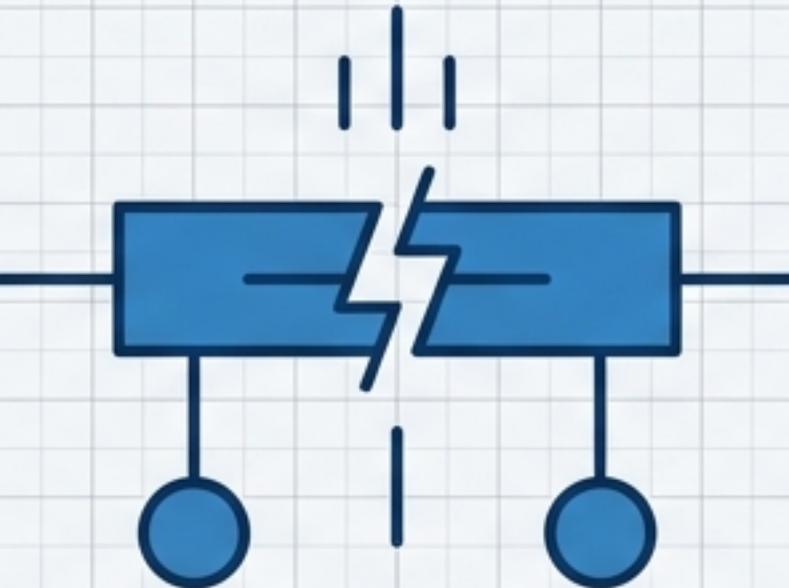
結論：對於非線性群集結構，t-SNE 提供顯著更優異的視覺分離度。

工程評估：優缺點與限制

優勢 (Advantages)



- ✓ 最佳群集視覺化
(Best-in-class Visualization)
- ✓ 處理非線性幾何結構
(Manifolds)
- ✓ 無監督學習 (無需標籤)

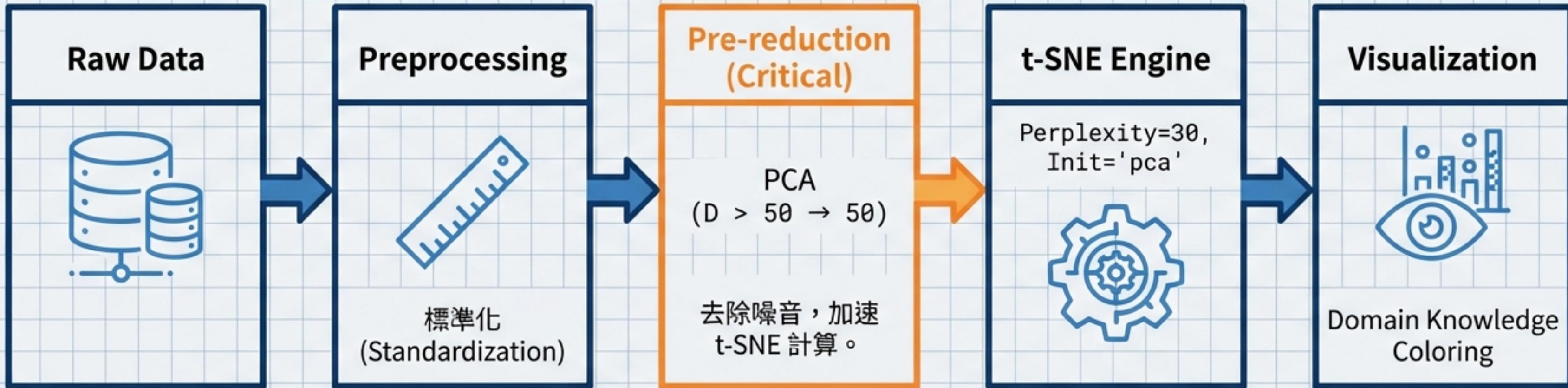


限制 (Limitations)



- ⚠ 計算成本高： $O(N^2)$ ，不適合
 $N > 10,000$
- ⚠ 隨機性 (Stochastic)：需固定
Random State
- ⚠ 不可投影 (Non-parametric)：
無法處理新進數據
- ⚠ 軸無意義：距離僅具局部參考性

最佳實踐工作流程 (Best Practices)



Always fix `random_state` for reproducibility.

總結：化工數據分析的策略選擇

PCA (The Monitor)

- 用途：製程監控、故障偵測、即時系統。
- 特性：全域結構、可解釋性高、計算快。
- 關鍵：可投影新數據 (Online Projection)。

t-SNE (The Explorer)

- 用途：探索性分析、模式識別、歷史審查。
- 特性：局部結構、視覺效果極佳、非線性。
- 關鍵：揭示隱藏的群集結構。

Next Unit 07: UMAP - Faster, Global-structure preserving.