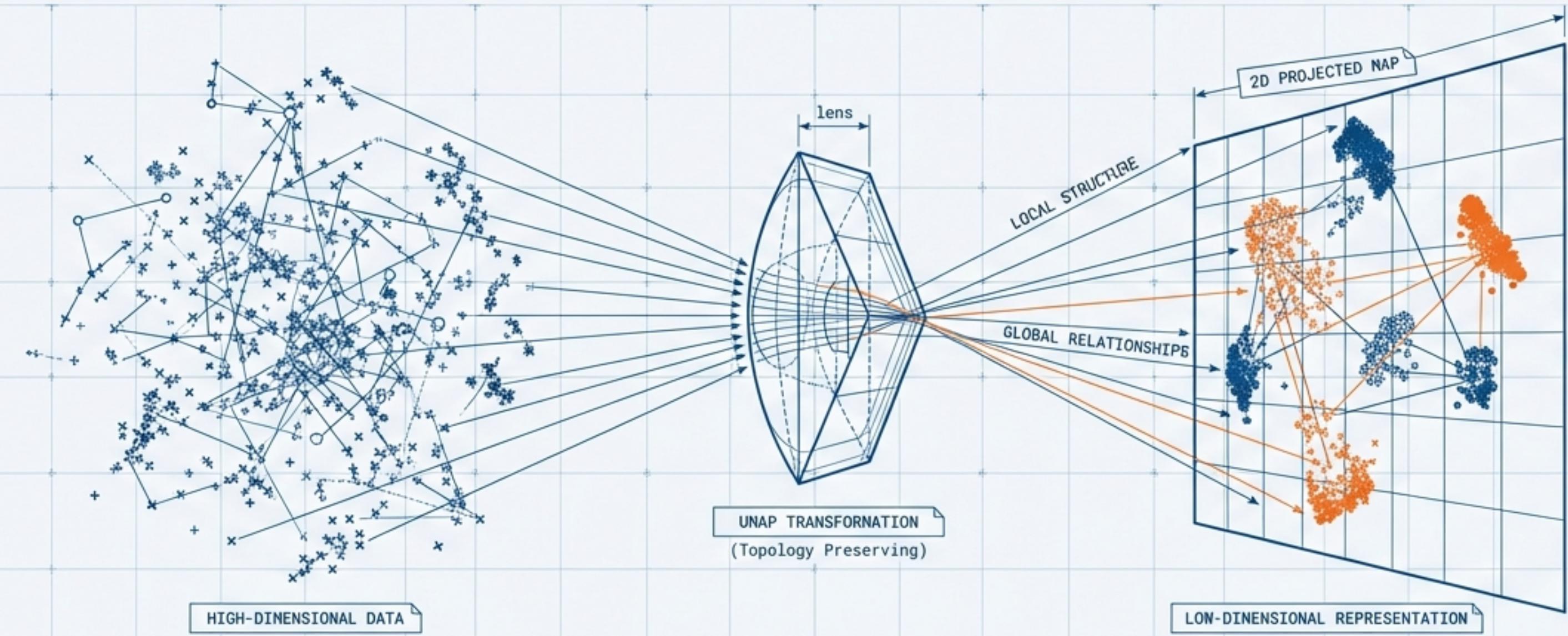


Unit 06: UMAP 大規模數據降維與視覺化

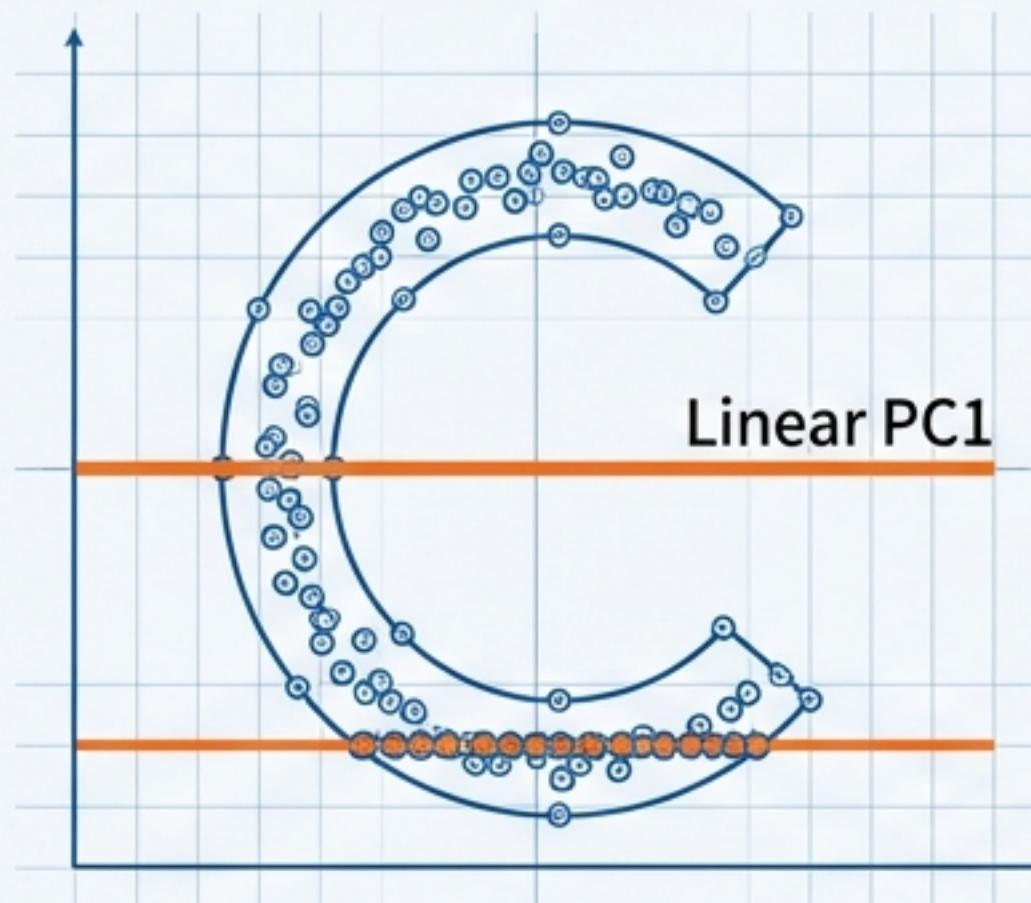
Uniform Manifold Approximation and Projection

從「黑盒」到「拓撲藍圖」——掌握高維化工數據的形狀



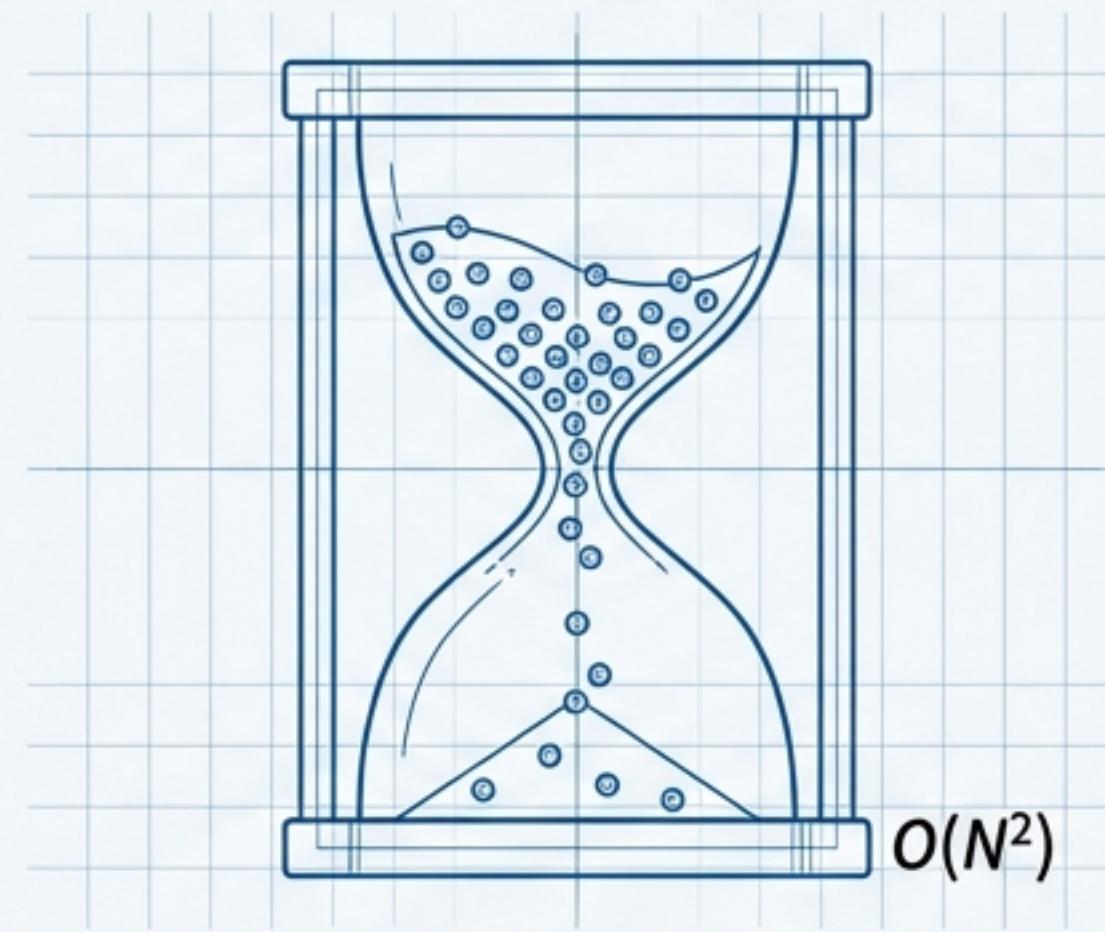
為什麼我們需要 UMAP？化工數據的維度挑戰

PCA：線性限制



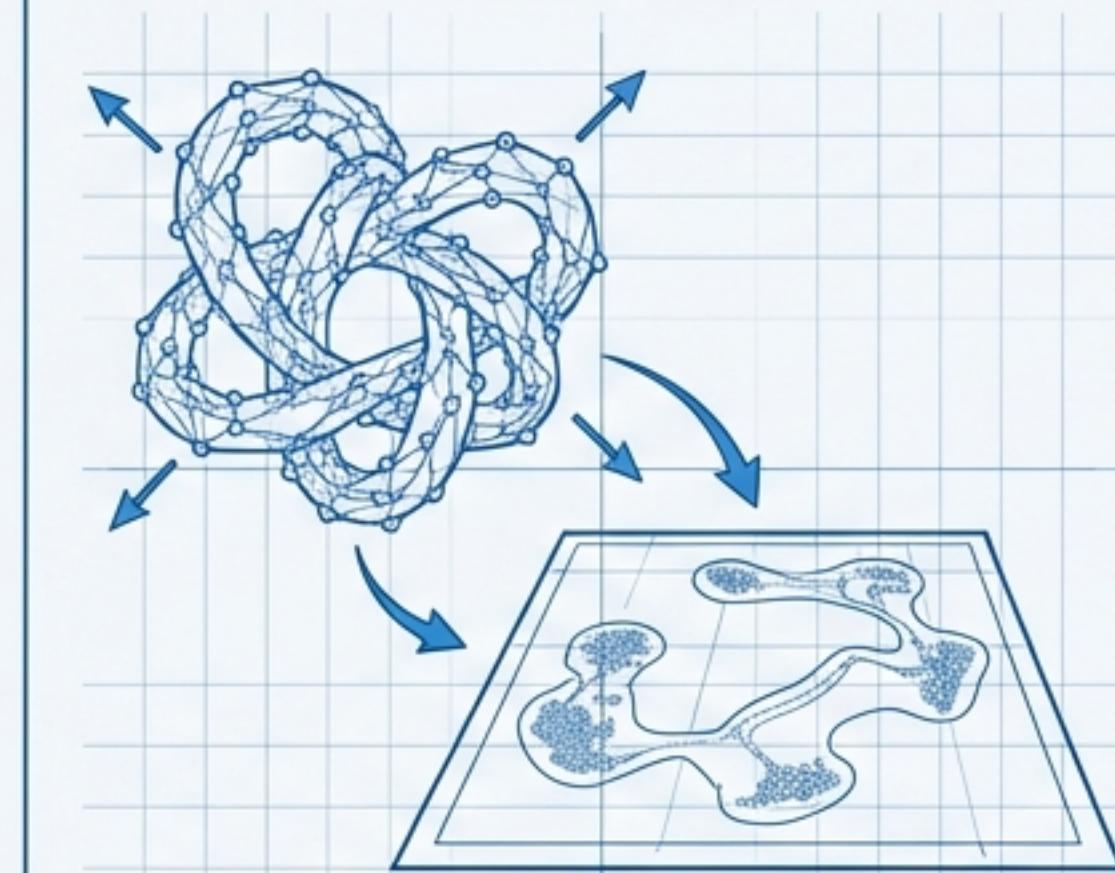
無法捕捉非線性動態

t-SNE：規模限制



計算過慢 ($O(N^2)$)

UMAP：最佳解



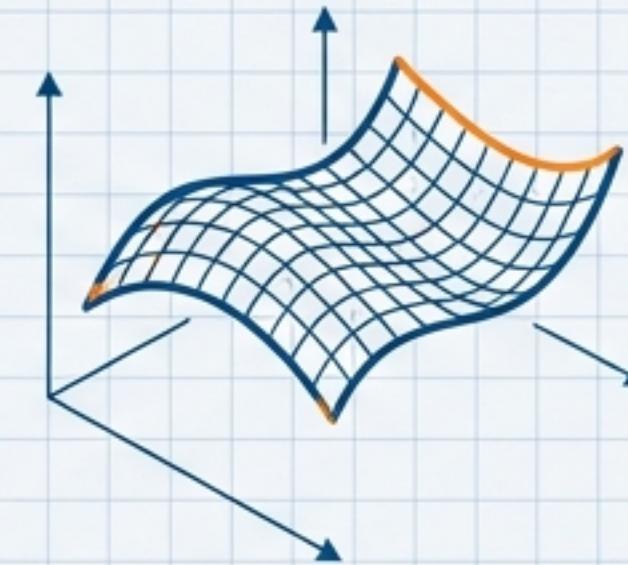
快速且保留全域結構

現實挑戰 (The Challenge)

- 化工廠產生海量數據 (Sensors × Time × Batches)。我們需要看見數據的「形狀」。
- PCA 的不足：線性降維，忽略了反應器內的複雜非線性關係。
- t-SNE 的不足：雖然視覺化效果好，但計算複雜度過高，且難以保留全域結構。
- **UMAP 的解答**：結合流形學習與拓撲數據分析，速度快、保留全域結構，且支援新數據投影。

演算法檔案：什麼是 UMAP？

UMAP (Uniform Manifold Approximation and Projection)	
發布年份 (Year)	2018 (McInnes, Healy, Melville)
核心理論 (Theory)	黎曼幾何 (Riemannian Geometry) & 代數拓撲 (Algebraic Topology)
運算複雜度 (Complexity)	$O(N \log N)$ (比 t-SNE 快 10-100 倍)
輸入 (Input)	高維特徵向量 \mathbf{X} (如：溫度，壓力，光譜)
輸出 (Output)	低維嵌入座標 \mathbf{Y} (2D 或 3D 座標)



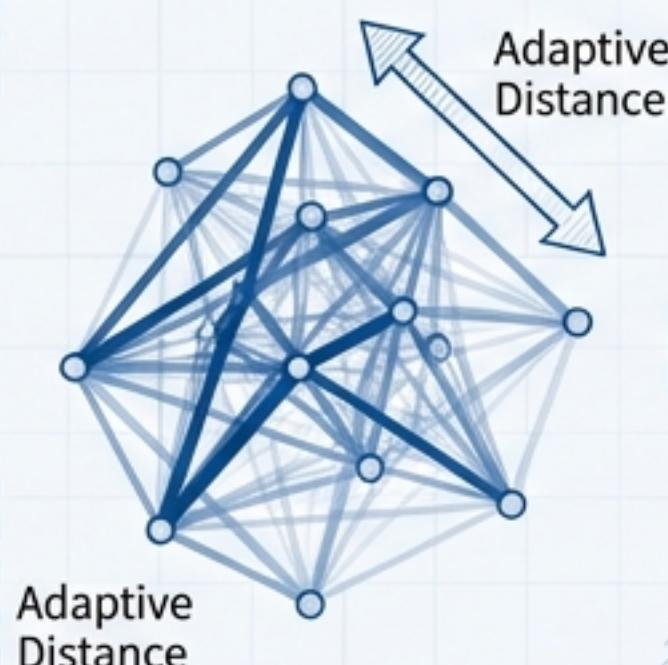
核心理念 (Core Philosophy)

1. 流形假設 (Manifold Hypothesis)：數據分布在在高維空間中的低維流形上。
2. 局部歐氏性 (Local Euclidean)：在局部鄰域內，流形近似為平坦的歐氏空間。
3. 拓撲保持 (Topology Preservation)：在低維空間中尋找最相似的拓撲結構。

核心機制：從高維到低維的拓撲映射

1. 構建模糊單純複形

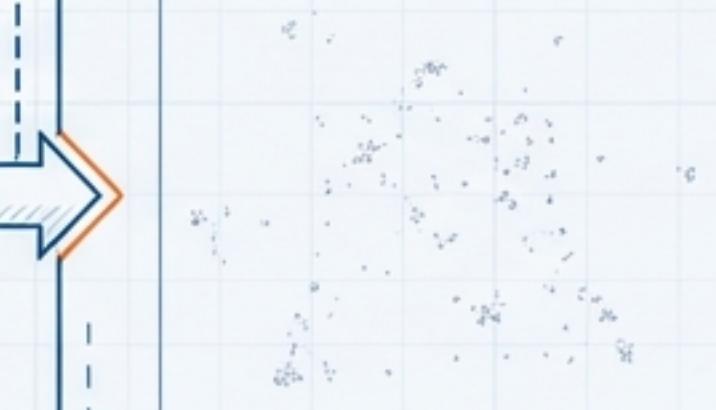
High-D Fuzzy Complex



2. 低維初始化

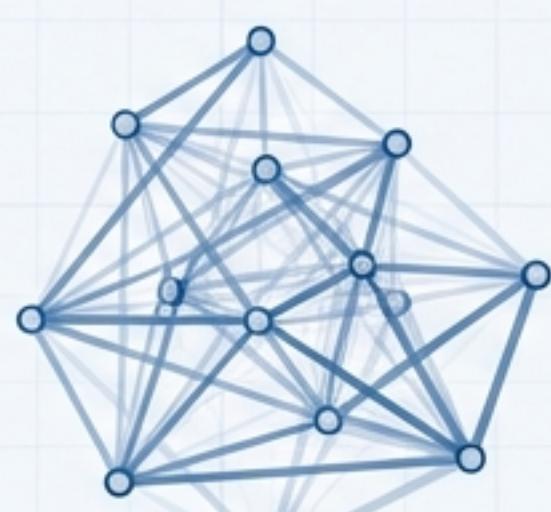
Low-D Initialization

Topological Representation



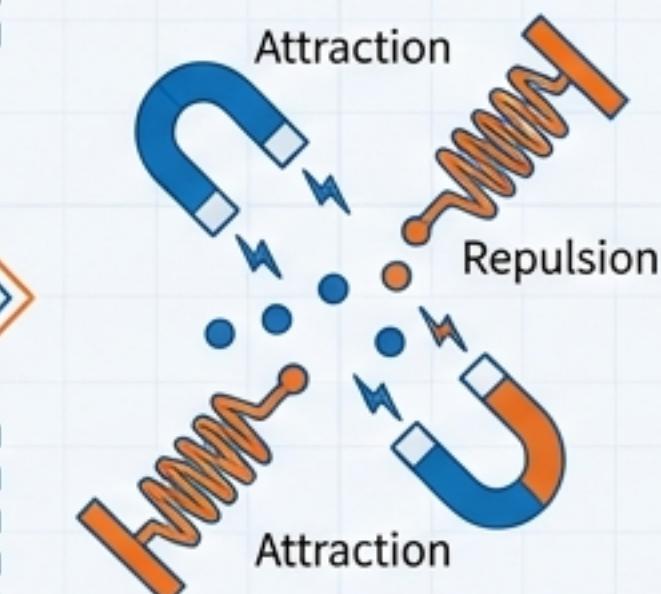
3. 構建低維拓撲

Low-D Fuzzy Topology



4. 交叉熵優化

Cross-Entropy Optimization



自適應距離 (Adaptive Distance) 處理數據密度差異。

Spectral Embedding
保留全域概貌。

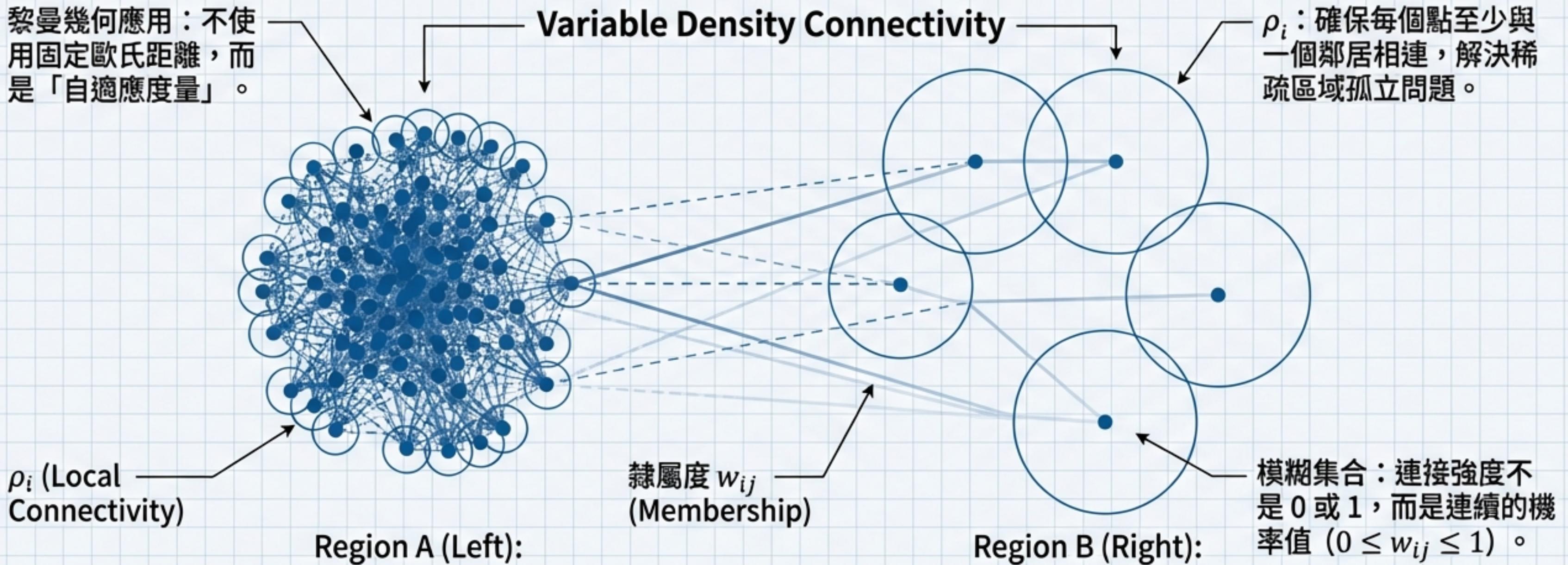
定義相似度 ν_{ij} (類似 t-Distribution)。

最小化高維與低維差異。
平衡引力與斥力。

數學直覺：模糊拓撲與流形

黎曼幾何應用：不使用固定歐氏距離，而是「自適應度量」。

Variable Density Connectivity



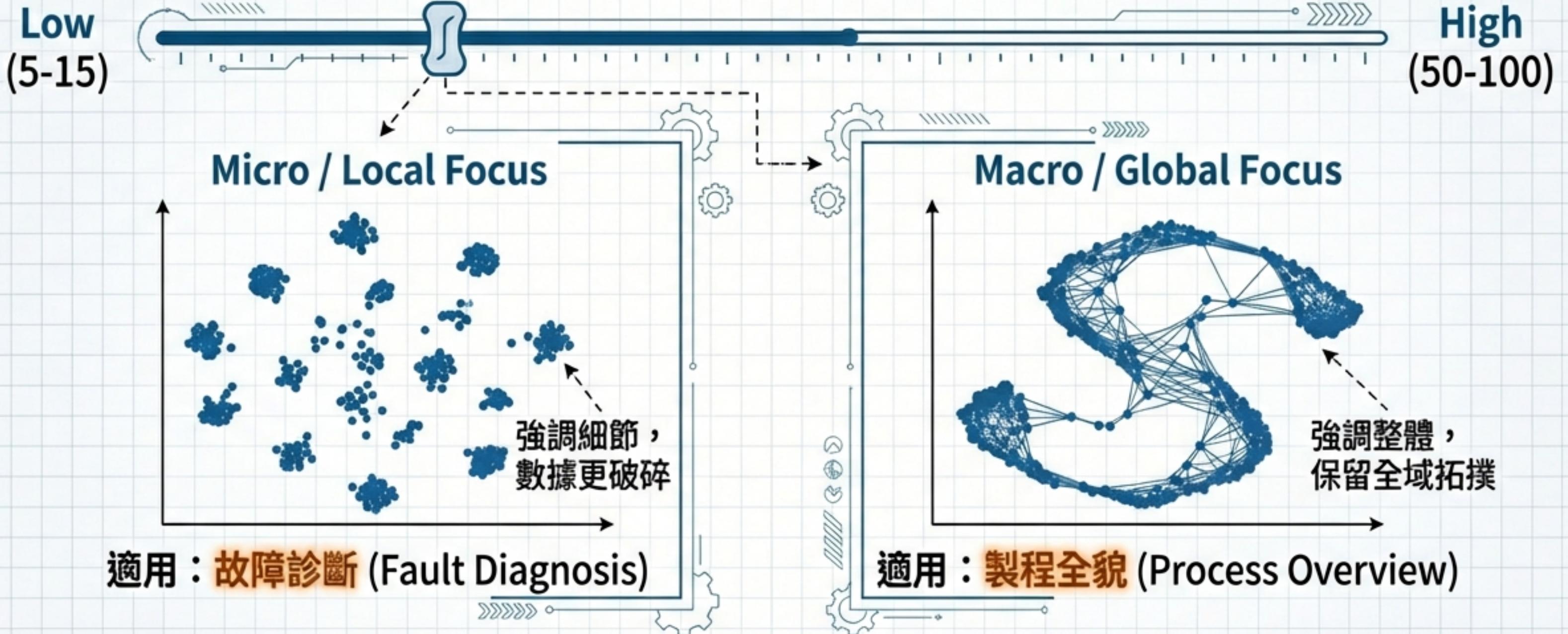
這就像是為數據建立了一個「彈性網格」，在稠密處收緊，在稀疏處放鬆。

演算法對決：UMAP vs. t-SNE vs. PCA

	PCA	t-SNE	UMAP
性質 (Nature)	線性 (Linear)	非線性 (Probabilistic)	非線性 (Topological)
速度 (Speed)	極快 (Fastest)	慢 (Slow)	快 (Fast)
全域結構 (Global Structure)	優 (Excellent)	差 (Poor)	良 (Good)
群集分離 (Separation)	差 (重疊嚴重)	極優 (Excellent)	優 (Great)
新數據投影 (New Data)	支援 (Yes)	不支援 (No)	支援 (<code>transform()</code>)

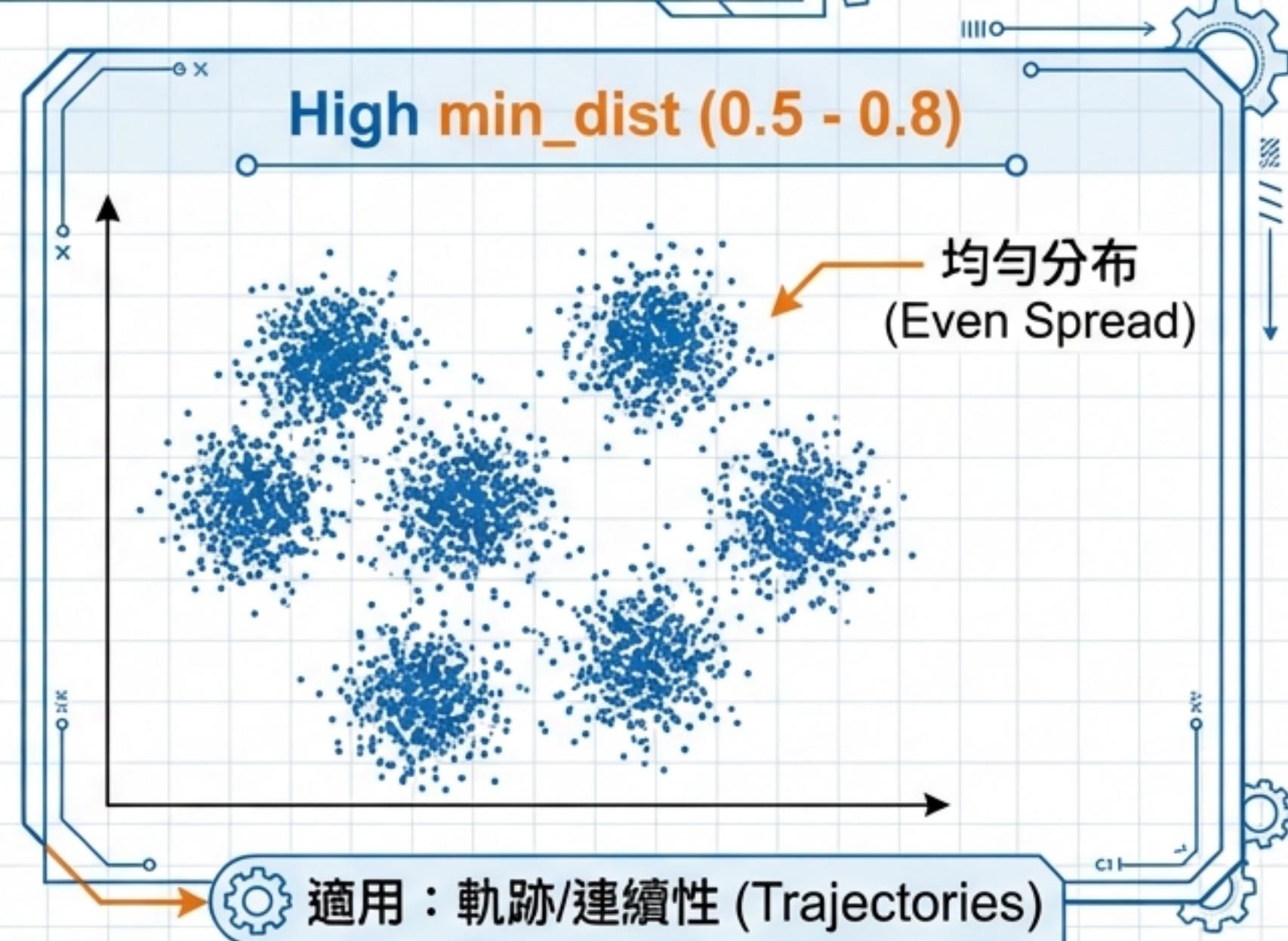
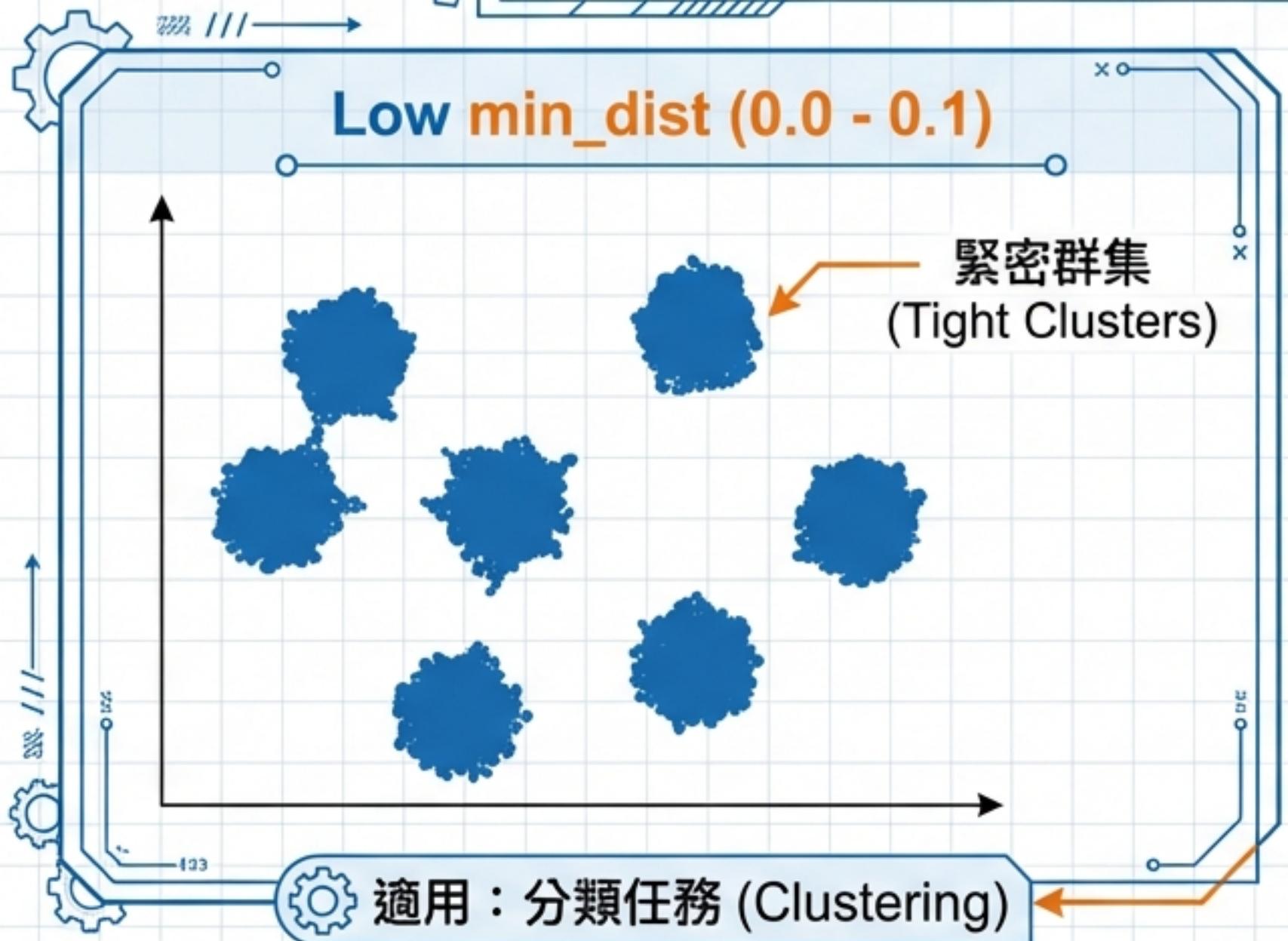
關鍵結論：UMAP 結合了 t-SNE 的視覺化品質與 PCA 的速度/擴展性。

控制面板：鄰居數 ($n_{neighbors}$) 的影響



$n_{neighbors}$ ：決定了構建拓撲結構時參考的鄰居數量。如同顯微鏡的變焦旋鈕。

控制面板: 最小距離 (min_dist) 的影響

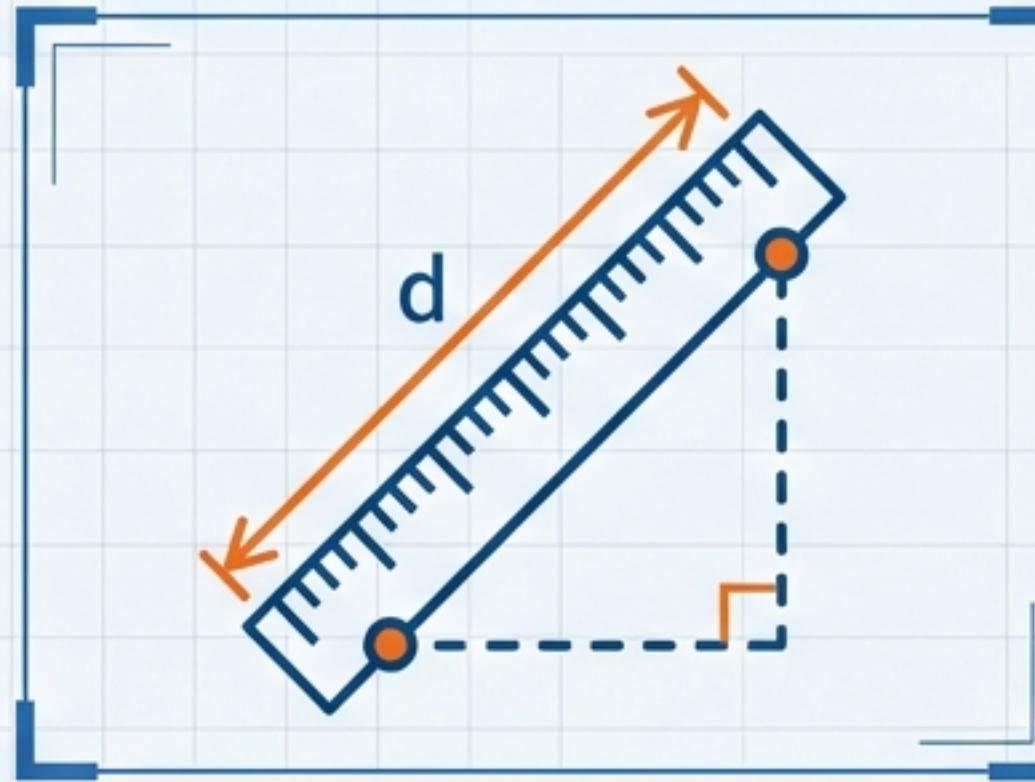


關鍵結論

- **min_dist:** 控制低維空間中點之間的最小允許距離 (Clumpiness)。
- **ChemE Insight:** 觀察批次反應演化 (Batch Evolution) 時，避免設得太低，以免軌跡斷裂。

特徵工程與距離度量 (Feature Engineering & Metric Selection)

原則：**Garbage In, Garbage Out.** 幾何結構取決於「距離」定義。

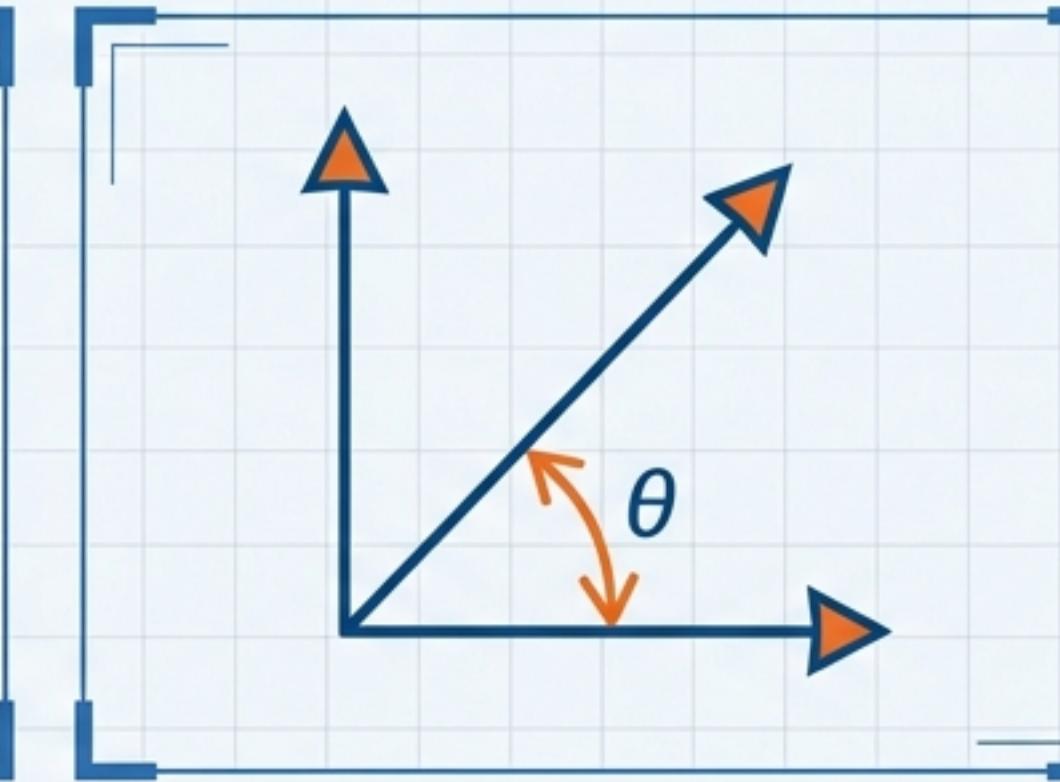


Euclidean (歐氏距離)

適用：一般製程傳感器數據（溫度，壓力，流量）。



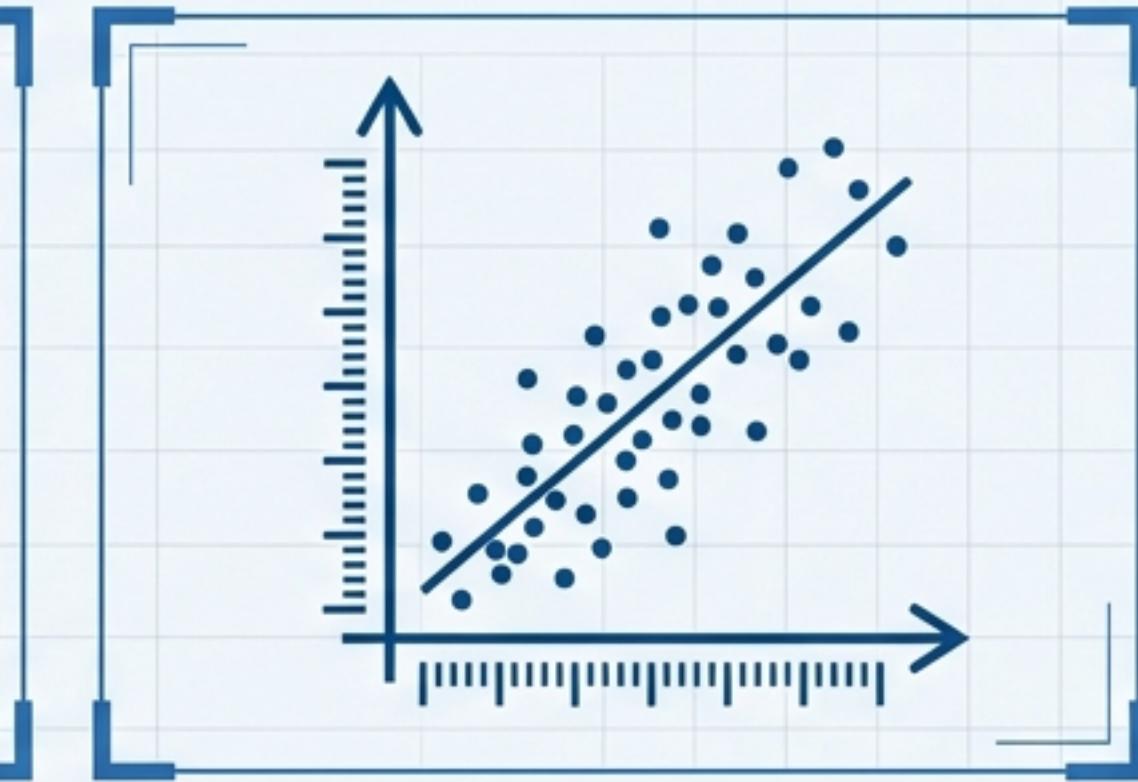
注意：必須先進行標準化 (Z-score)，避免大數值變數主導。



Cosine (餘弦距離)

適用：高維稀疏數據、光譜數據 (NIR/Raman)、文本。

原理：關注向量的方向（形狀）而非大小。



Correlation (相關係數)

適用：基因表現數據或特定生物化工應用。

關注變數變化的趨勢相似性。

評估指標：我們如何知道它是對的？

DASHBOARD



信賴度 (Trustworthiness)

局部結構保持

高維的近鄰在低維是否還是近鄰？



連續性 (Continuity)

全域結構完整

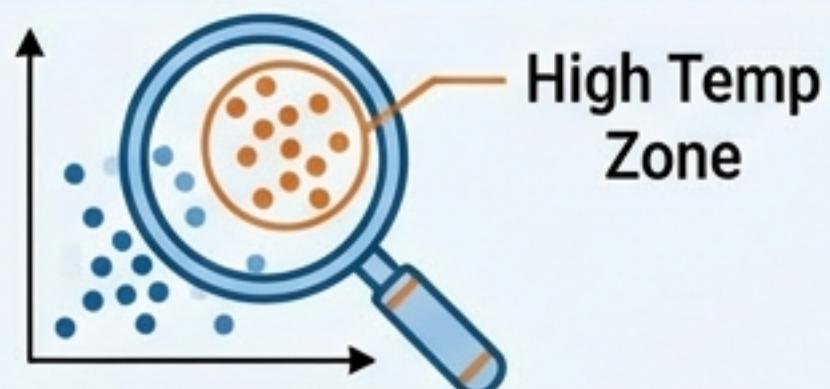
有沒有過度撕裂數據？



輪廓係數 (Silhouette Score)

群集分離度

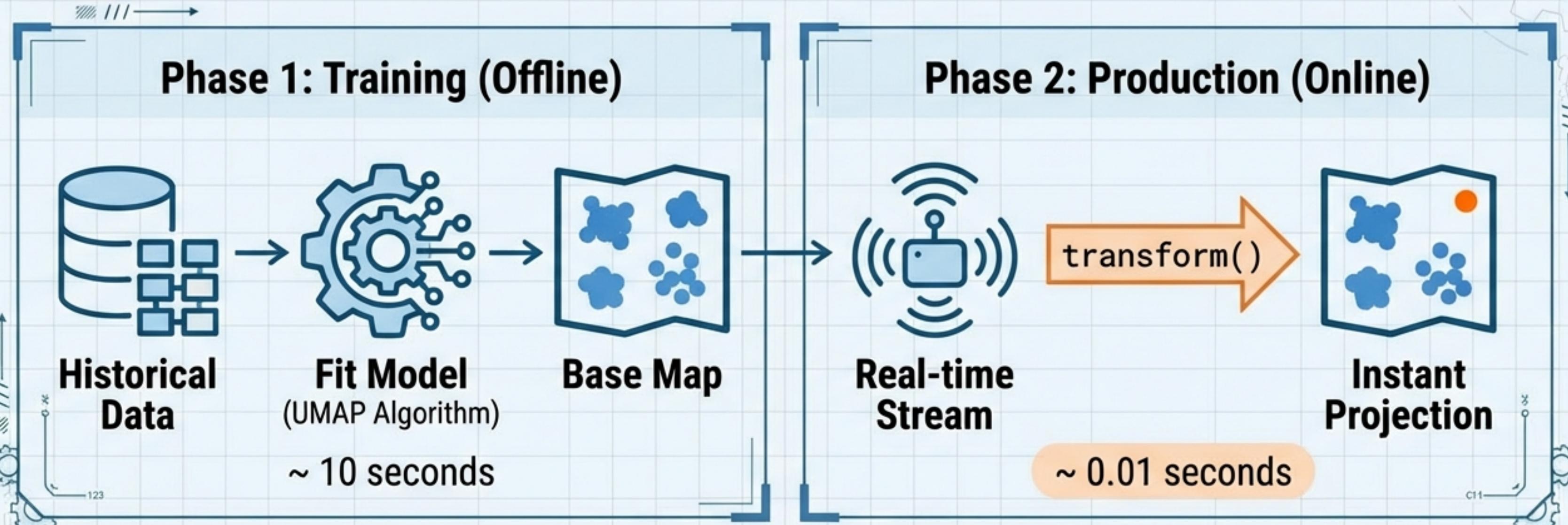
群集邊界是否清晰？



DOMAIN CHECK (領域驗證)

****最重要的指標**：地圖上的群集是否對應實際的物理意義
(如：反應器高溫區、產品不純區) ?

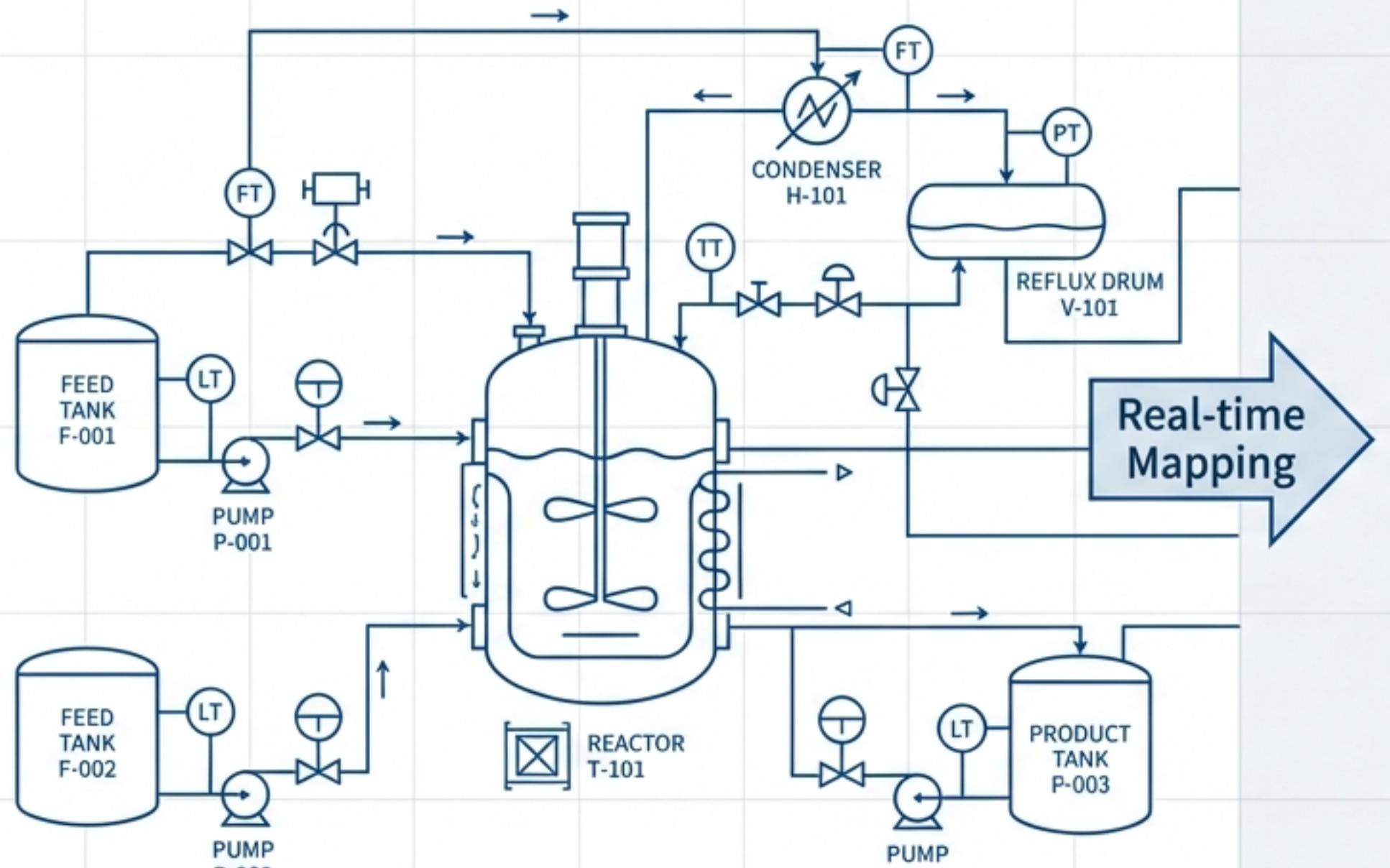
殺手級功能：新數據投影 (Projecting New Data)



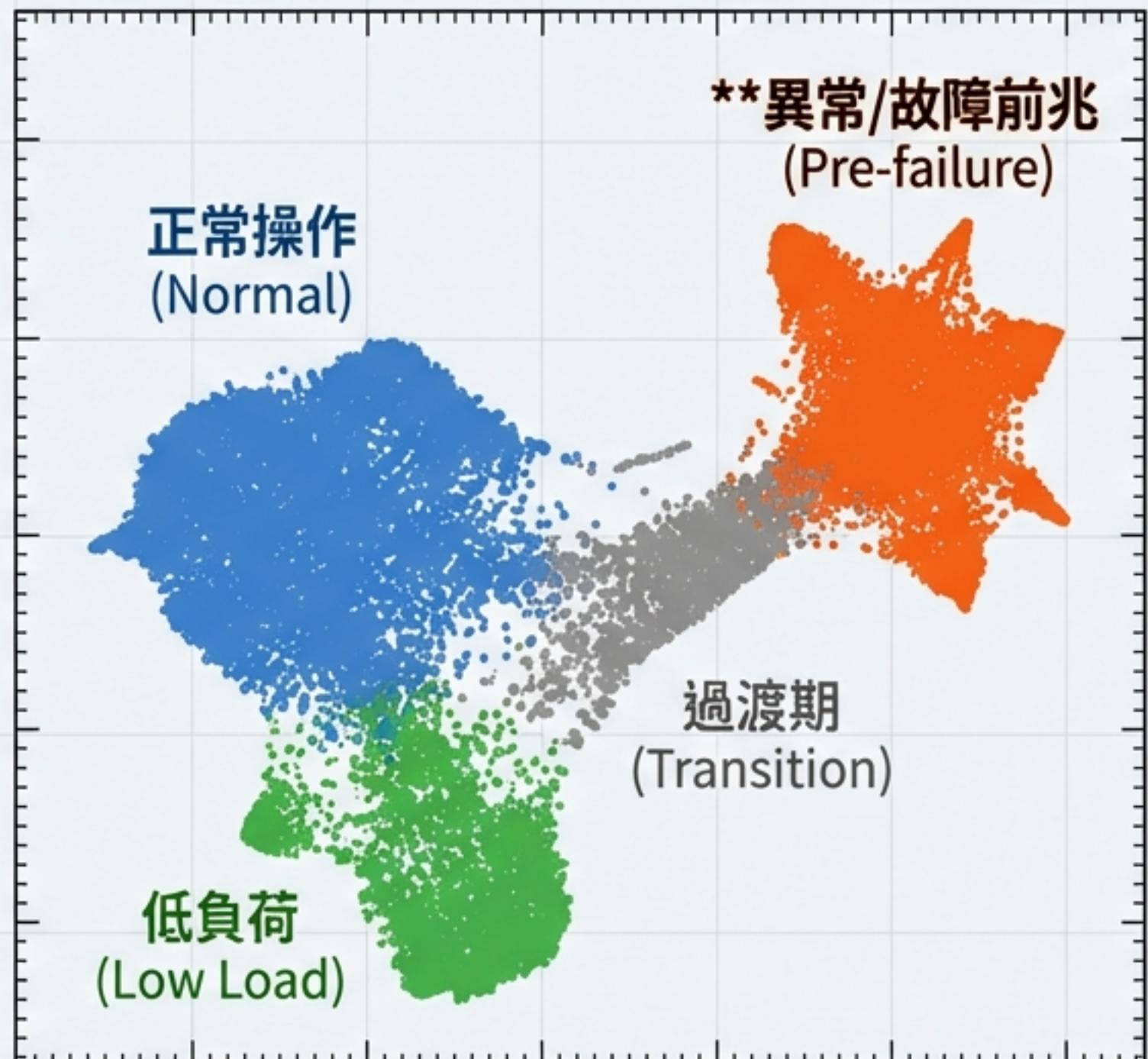
t-SNE 無法做到此點（需重新計算）。

UMAP 學習了參數化的映射關係，支援**即時異常檢測**：若新數據點落在正常群集之外，立即觸發警報。

化工應用一：製程狀態監控 (Process Monitoring)

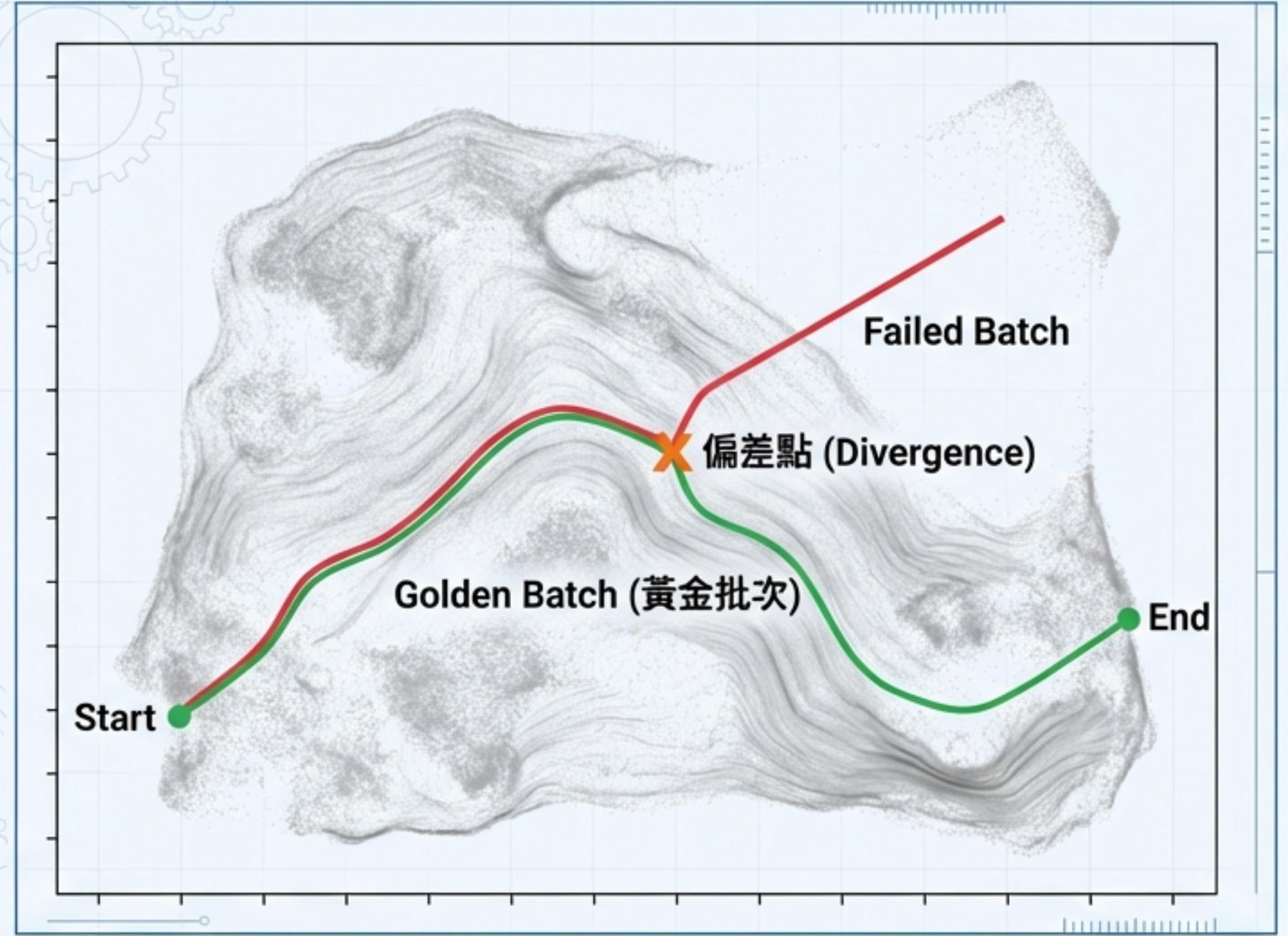


高維輸入: 50+ Sensors



價值：讓操作員一眼看出目前的反應器處於什麼「狀態」，而非盯著 50 個跳動的儀表。

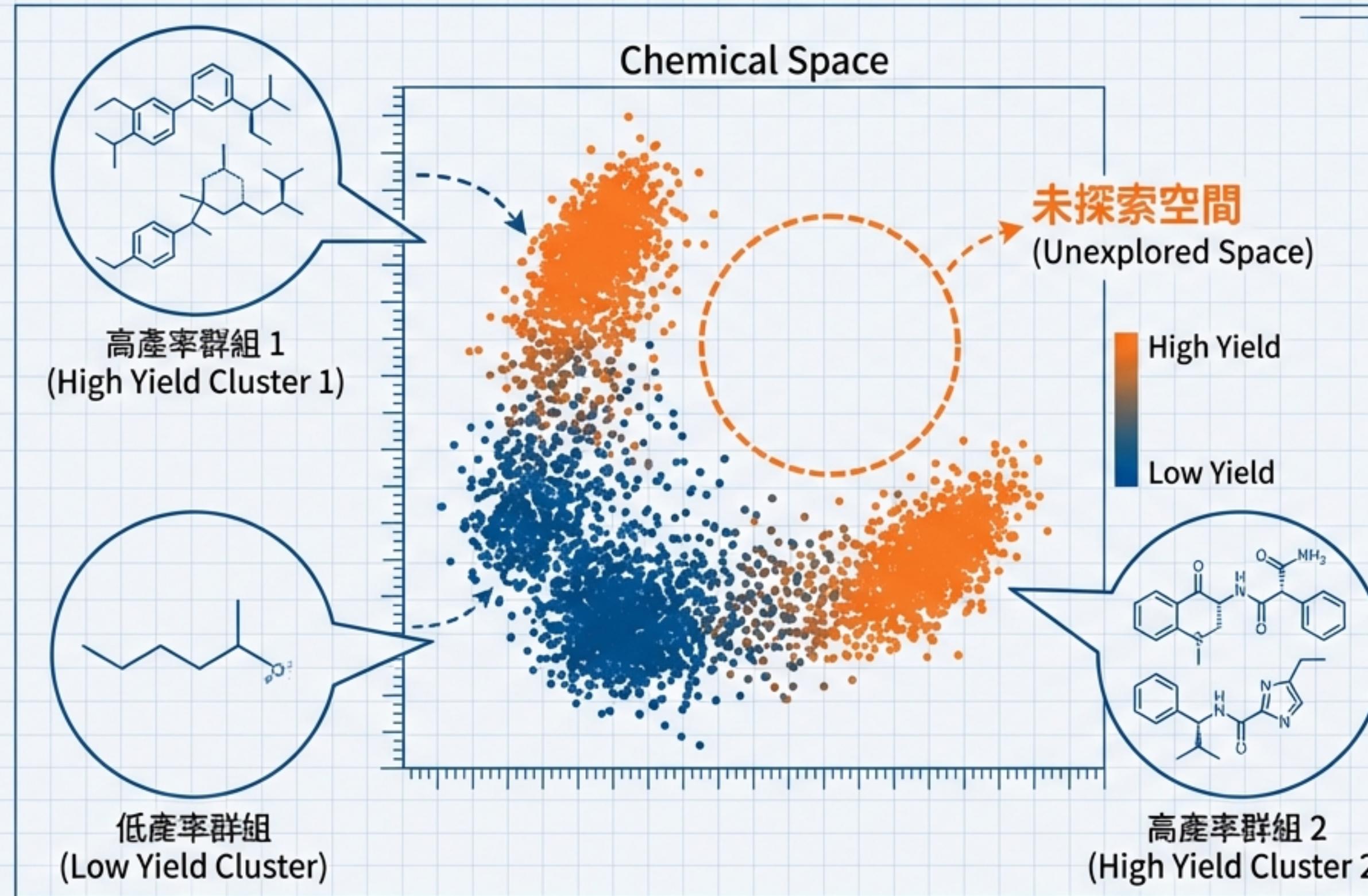
化工應用二：批次軌跡分析 (Batch Trajectory Analysis)



主要分析與價值 (Key Analysis & Value)

- **動態演化**：利用 UMAP 比較數百個批次的時間序列。
- **參數建議**：使用較高的``min_dist`` (0.3-0.5) 和``n_neighbors`` (30-50) 以保持軌跡連續性。
- **價值**：早期發現批次偏離，及時進行修正控制。

化工應用三：化學空間與配方探索



- 發現：找到高產率分子的「群島」(Islands of Performance)。
- 距離度量：使用 Tanimoto 或 Cosine 距離處理分子指紋。
- 應用：R&D 加速，指導下一步實驗設計 (DoE)。

Noto Sans TC 總結：定義未來的工程師

- UMAP 是現代標準:** 比 PCA 精細，比 t-SNE 強大。
- 工程視角:** 它不僅是演算法，更是透視高維數據的「透鏡」。
- 實戰關鍵:** 掌握 `n_neighbors` (縮放) 與 `min_dist` (緊密) 的調整。

下一步 (Next Step)



前往 Unit06_UMAP.ipynb 進行實作。嘗試將你的製程數據放入 UMAP，看看它長什麼樣子。

“AI 不會取代化工工程師；但懂得使用 AI 看見數據形狀的工程師，將取代看不見的人。”