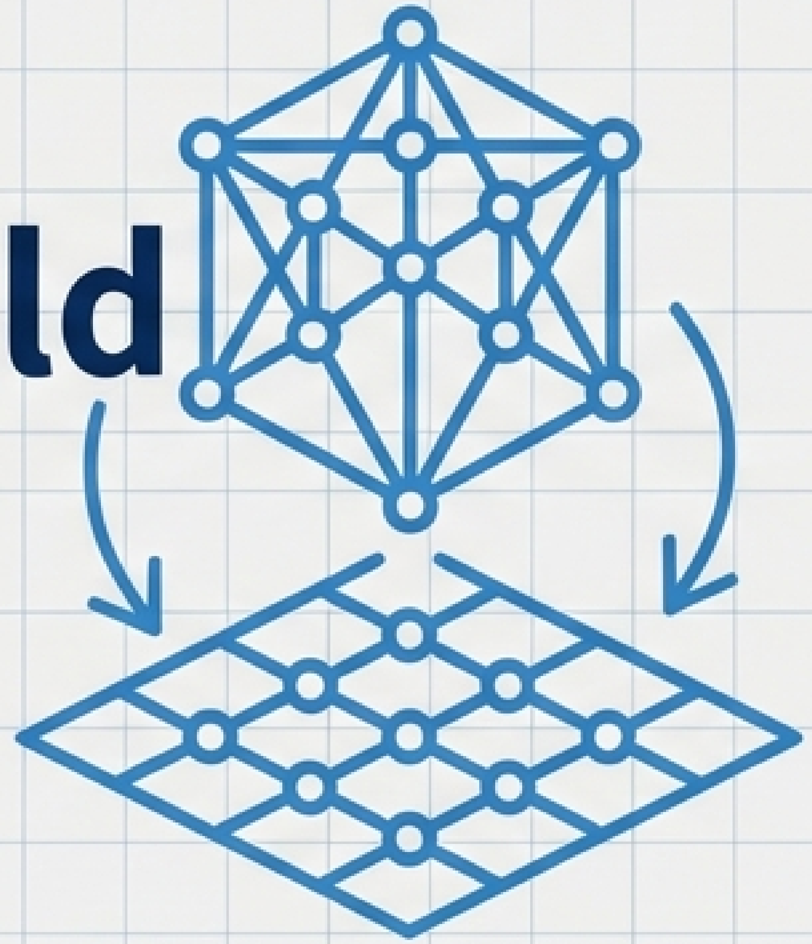
Unit 06: Uniform Manifold Approximation and Projection (UMAP)

大規模製程數據的非線性降維與視覺化



Project Metadata	
Domain	Chemical Engineering AI Applications (AI在化工上之應用)
Laboratory	逢甲大學 化工系 智慧程序系統工程實驗室
Instructor	莊曜禎 助理教授
Date	2026-01-28

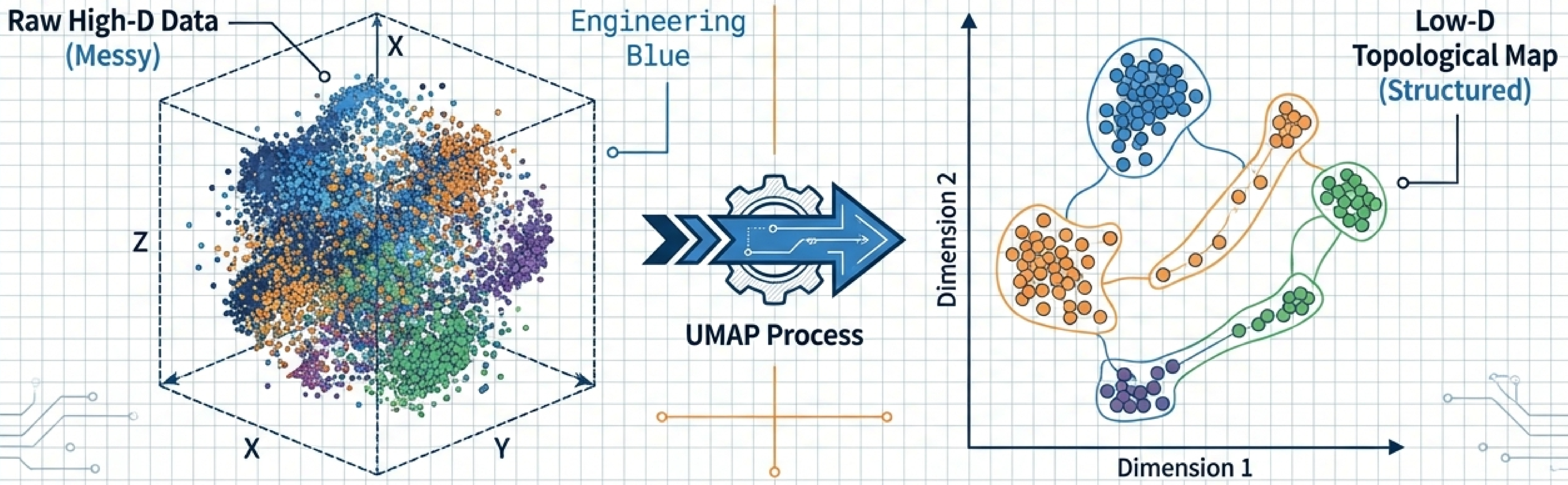
本單元學習目標 (Unit Objectives)

CHECKLIST

- [x] **核心原理 (Core Principles):**
  理解流形學習 (Manifold Learning) 與拓撲數據分析。
- [x] **效能對比 (Performance Benchmark):**
  掌握 UMAP 與 t-SNE、PCA 的差異與優勢。
- [x] **實作部署 (Implementation):**
  使用 Python `umap-learn` 套件建立模型。
- [ ] **參數調校 (Calibration):**
  調整 `n_neighbors` 與 `min_dist` 優化視覺化結果。
- [ ] **化工應用 (Applications):**
  反應器操作模式識別、批次軌跡分析與配方空間探索。

什麼是 UMAP？ (What is UMAP?)

Digital Unit Operation 06: The New Standard



極速運算 (High Speed)

O(N log N) 複雜度，比 t-SNE 快 10-100 倍。

全局結構 (Global Structure)

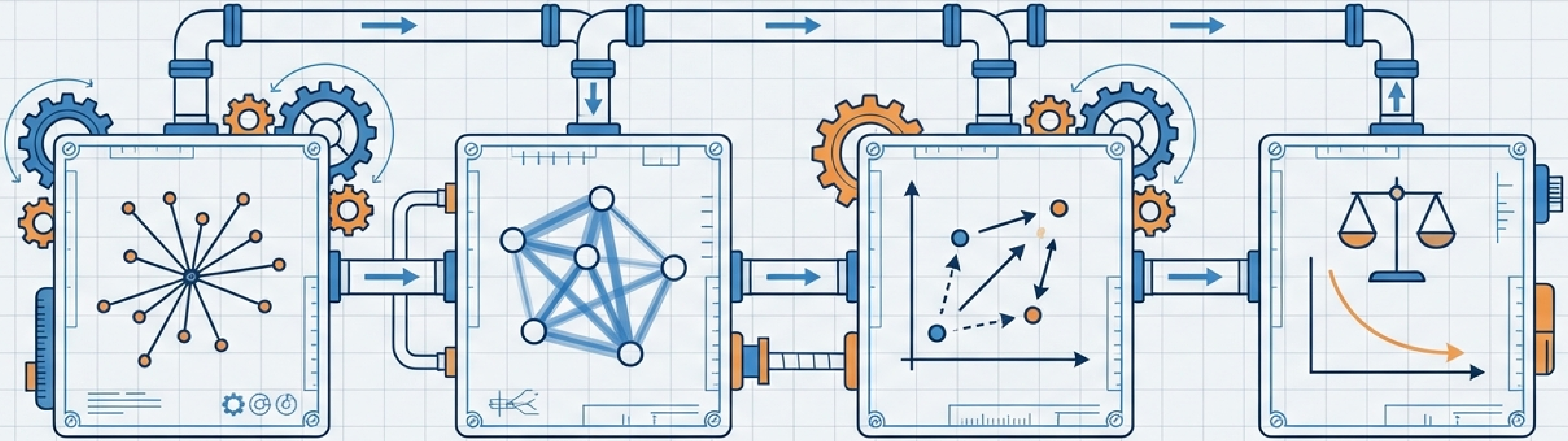
保留數據的整體拓撲結構與局部鄰域關係。

新數據投影 (New Data Projection)

支援 `transform()` 方法處理實時數據。

Industrial Blueprint v1

演算法核心機制 (Algorithm Engine)



1. 局部流形 (Local Manifold)

自適應距離:

$d_i = \max(0, \|\mathbf{x}_i - \mathbf{x}_j\| - \rho_i)$

2. 高維模糊拓撲 (High-D Fuzzy Topology)

隸屬度:

$w_{ij} = \exp(-d_i(\mathbf{x}_i, \mathbf{x}_j) / \sigma_i)$

3. 低維佈局優化 (Low-D Layout)

機率擬合:

$v_{ij} = (1 + a \|\mathbf{y}_i - \mathbf{y}_j\|^{2b})^{-1}$

4. 交叉熵最小化 (Cross-Entropy)

Minimize

$C = \sum (w \log(w/v) + (1-w) \log((1-w)/(1-v)))$

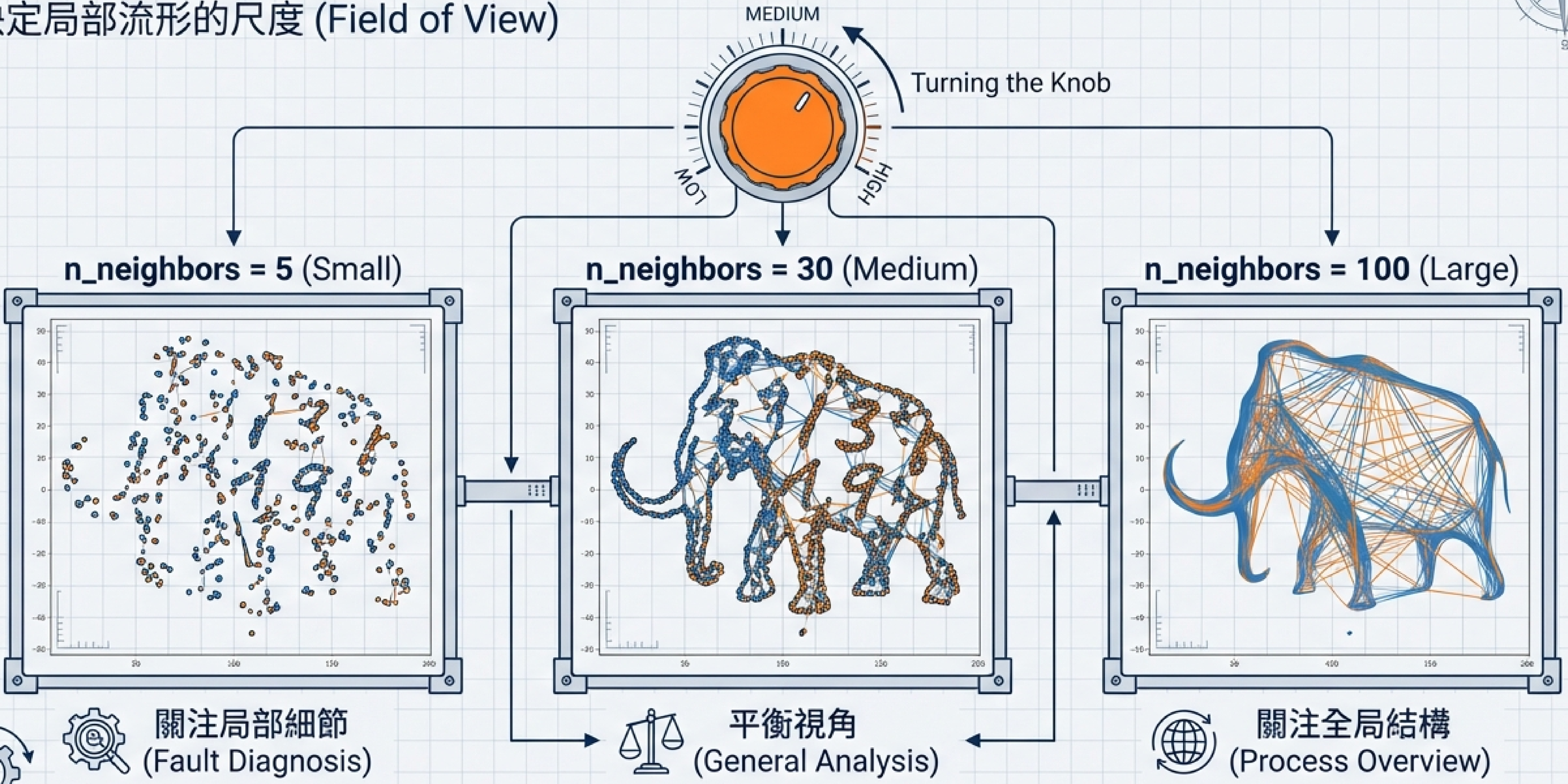
降維方法效能評比 (Performance Benchmark)

	PCA (Legacy)	t-SNE (Standard)	UMAP (Advanced)
運算速度 (Speed)	極快 (Fastest)	慢 (Slow)	快 (Fast)
線性/非線性 (Type)	線性 (Linear)	非線性 (Non-linear)	非線性 (Non-linear)
局部結構 (Local)	差 (Poor)	極佳 (Excellent)	佳 (Good)
全局結構 (Global)	佳 (Good)	失真 (Distorted)	佳 (Good)
實時投影 (Transform)	支援 (Yes)	不支援 (No)	支援 (Yes)

Key Takeaway: UMAP 是兼顧速度與結構的大規模化工數據首選。

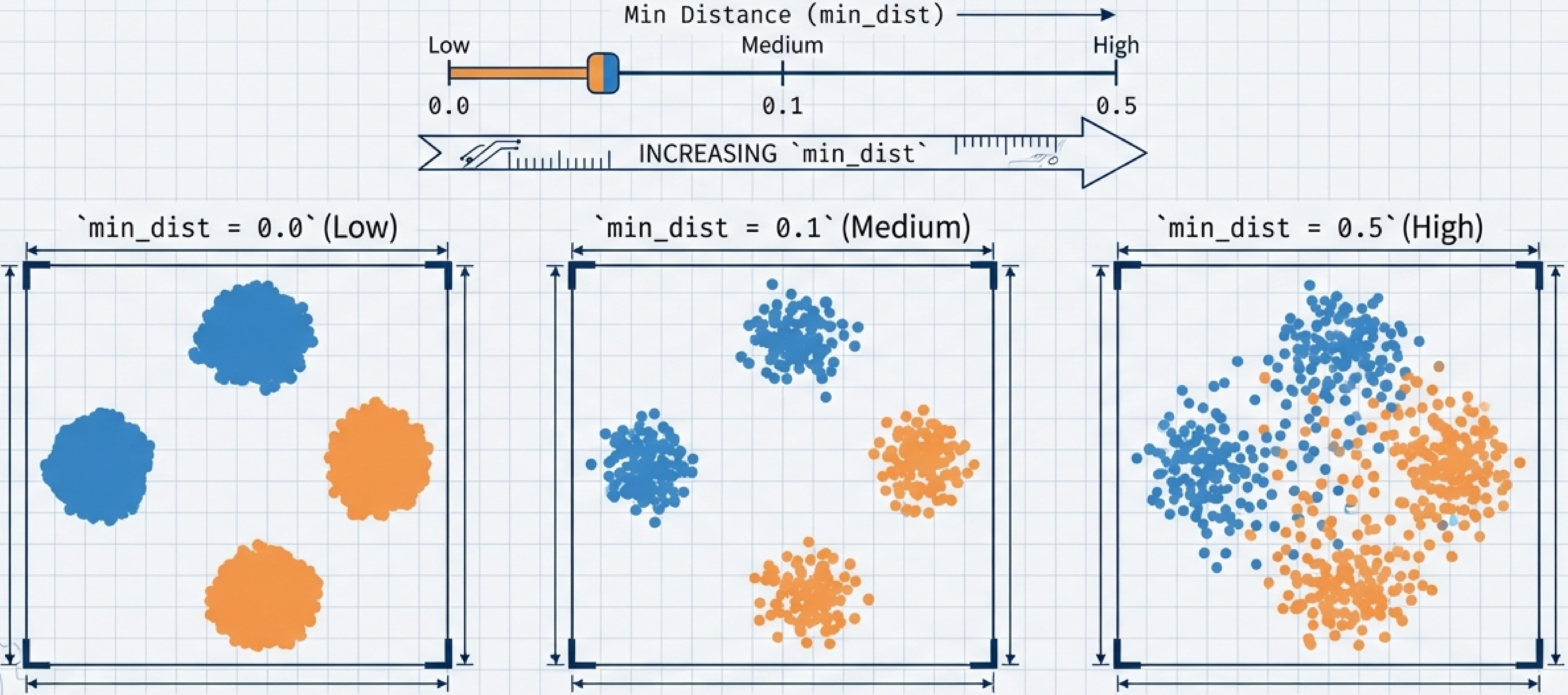
參數調校：鄰居數 (Calibration: `n_neighbors`)

決定局部流形的尺度 (Field of View)



參數調校：最小距離 (Calibration: `min_dist`)

控制低維空間中點的緊密程度 (Clump Tightness)



邊界銳利 (Clustering Focus)

標準設定 (Default)

分布均勻 (Topology Focus)

Python 實作流程 (Implementation Blueprint)

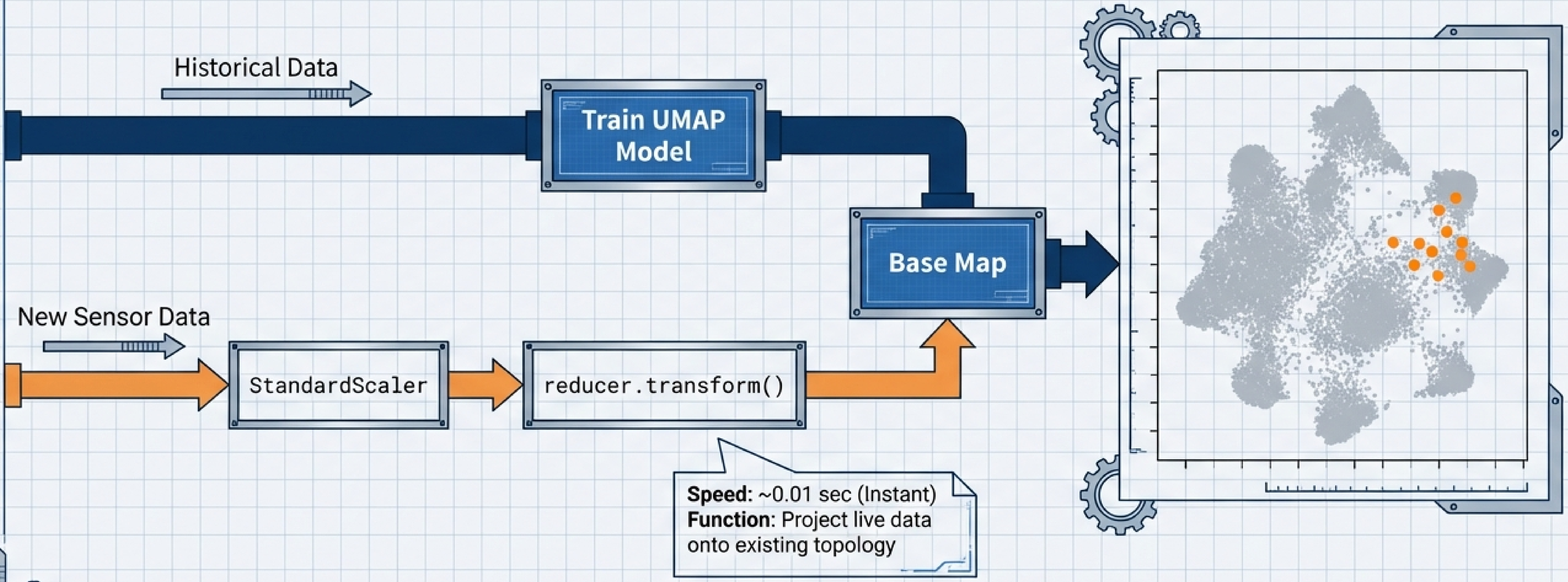
INPUT DATA

```python
import umap
from sklearn.preprocessing import StandardScaler

# 1. 前處理 (Critical Step!)
# 務必先進行標準化，因為距離計算對尺度敏感
scaled_data = StandardScaler().fit_transform(raw_data)

# 2. 建立模型 (Configuration)
reducer = umap.UMAP(
    n_neighbors=50,
    min_dist=0.1,
    n_components=2,
    random_state=42
)

# 3. 擬合與轉換 (Execution)
embedding = reducer.fit_transform(scaled_data)
```

Alert: Always Scale Data First!

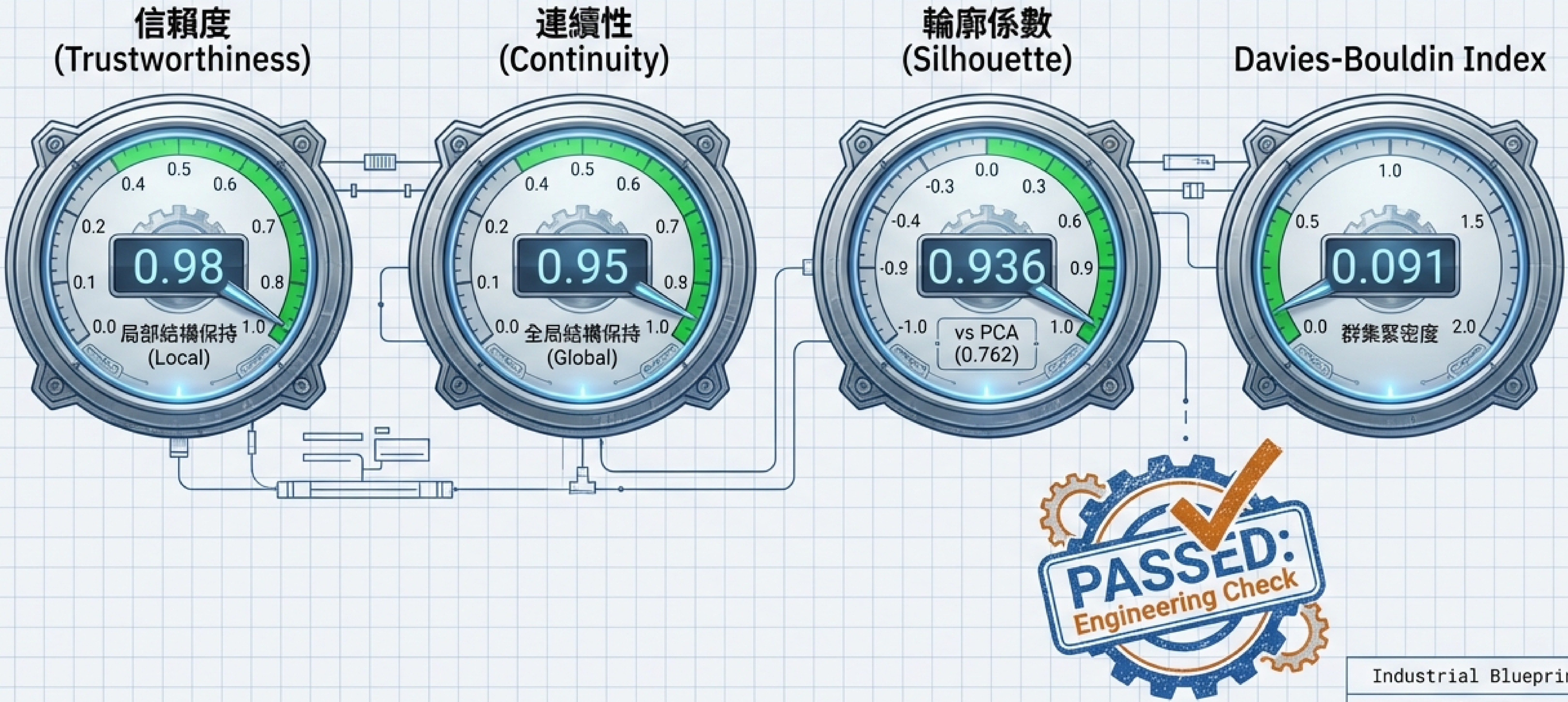
EMBEDDING OUTPUT

實時製程監控 (Real-Time Process Monitoring)



視覺化品質評估 (Quality Control Metrics)

Noto Sans TC | Roboto Mono



Industrial Blueprint v1

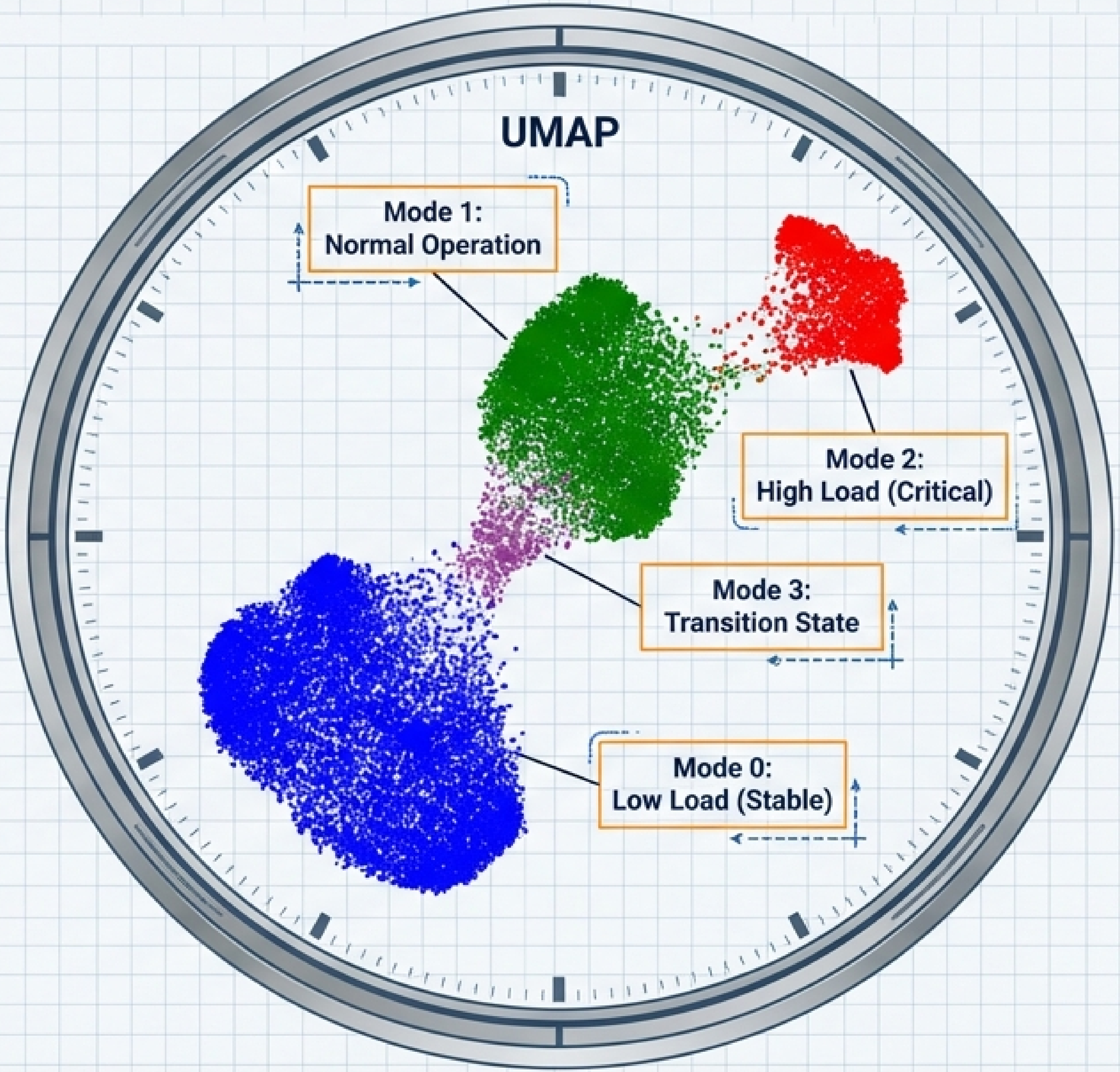
NotebookLM | Note

案例一：反應器操作模式識別 (Reactor Operation Modes)

Data Source:
Simulated Reactor (模擬反應器)

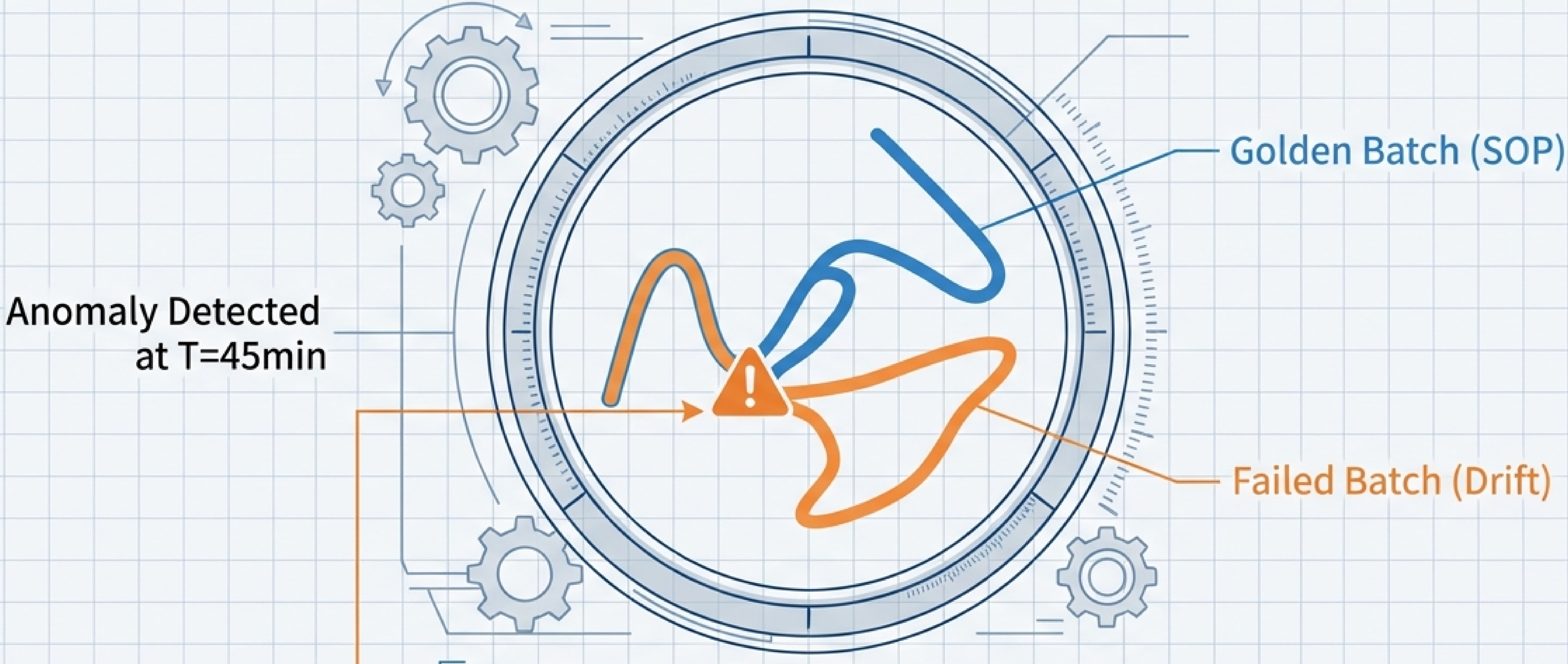
Variables:
50+ (Temp, Pressure, Flow)

Samples:
100,000+



案例二：批次製程軌跡分析 (Batch Trajectories)

從靜態點到動態線 (Dynamic Time-Series Evolution)

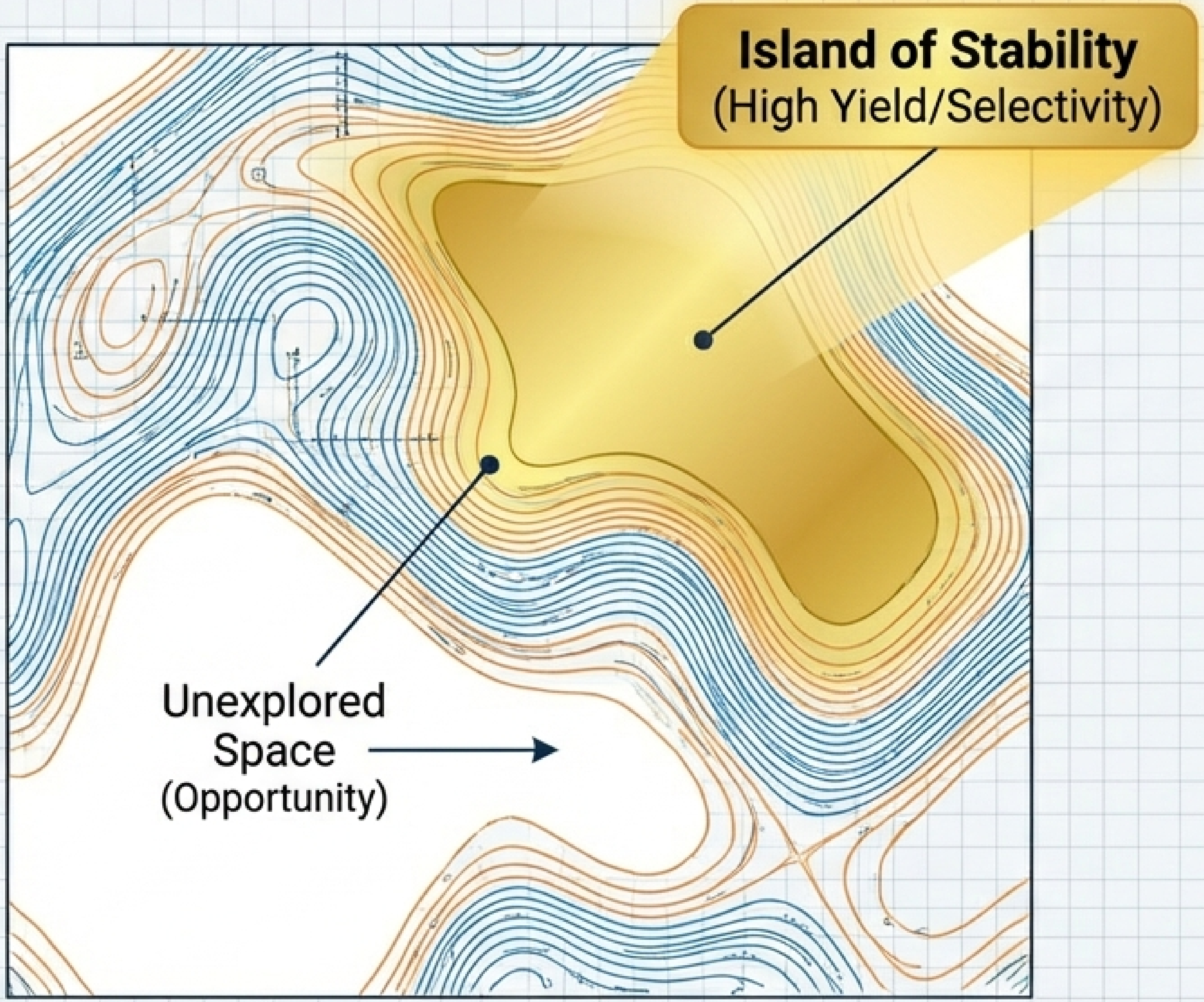


Early Warning System (早期預警):
Detecting deviation before the batch is complete.

Noto Sans TC

案例三：高維配方與優化探索 (Formulation & Optimization)

- **Design Space Exploration:**
  視覺化數千種配方
- **Optimization:**
  鎖定最佳操作窗口
- **New Discovery:**
  識別空白區域進行開發



操作限制與最佳實踐 (Safety Protocols)

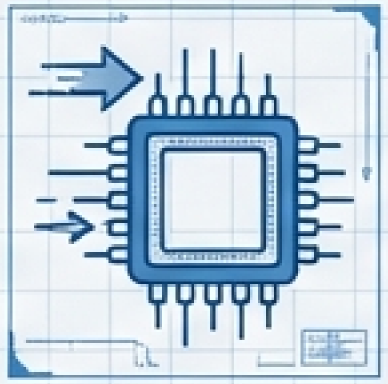
Limitations (限制)

- **Non-linear Distances:** 距離不代表絕對物理長度。
- **Randomness:** 結果受初始值影響，請固定 `random_state`。

Best Practices (最佳實踐)

- **Data Size:** 適合 > 10,000 筆數據 (小數據請用 t-SNE)。
- **Preprocessing:** 必須使用 StandardScaller。
- **Automation:** 結合 DBSCAN 進行自動分群。

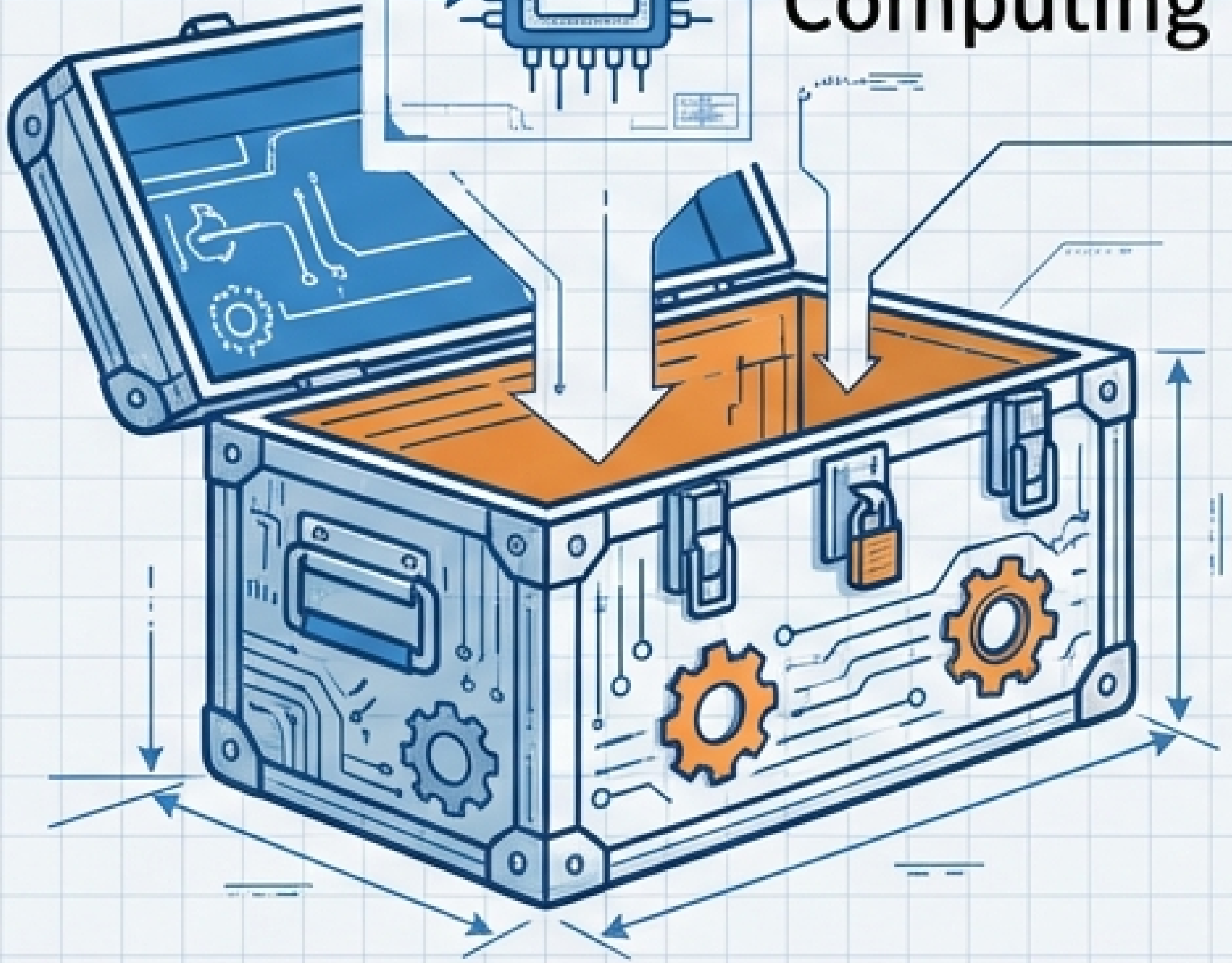
總結：您的數位工具箱已升級



Fast Computing

Global Topology

Real-time Monitor



UMAP = Fast + Global + Nonlinear

Next Step: 啟動 `Unit06_UMAP.ipynb` 開始處理您的反應器數據。

The Digital Toolkit is ready. (數位工具箱已備妥)