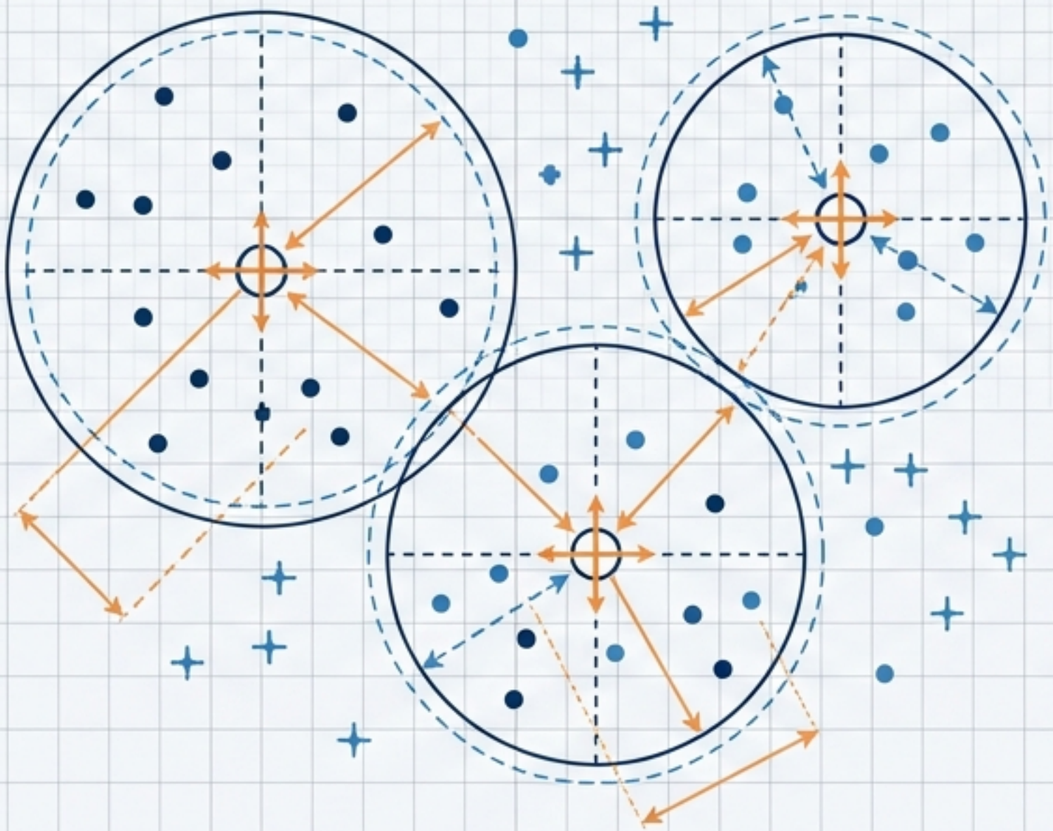
# Unit 05: K-Means 分群演算法

## 化工製程數據的非監督式學習



| | |
|---|---|
| SPEC: | ALG-KM-05 |
| CONTEXT: | Unit05_K_Means.md |
| SYSTEM: | Python scikit-learn |
| DATE: | 2026-01-28 |

# 演算法規格與應用場景

Definition

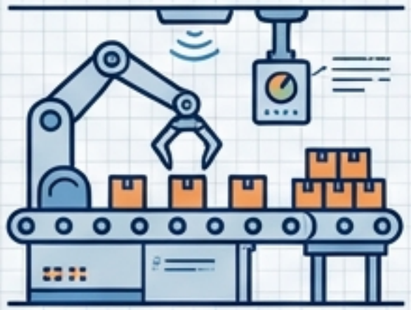
- **核心功能:** 將 n 個數據點劃分為 K 個群集 (Clusters)，使群內變異數最小化。
- **運作邏輯:** 基於距離 (Distance-based) 的非監督式學習。

Applications

### 1. 反應器操作識別

高/中/低產率區間判定

ZON2 ZONE



### 2. 批次品質分類

高品質 vs 次級品特徵

高品質 次級品

**數據處理中心**

### 3. 設備維護分群

健康/預警/故障狀態

健康 → 10.8 → 預警 → 10.0 → 故障 → 10.2 →

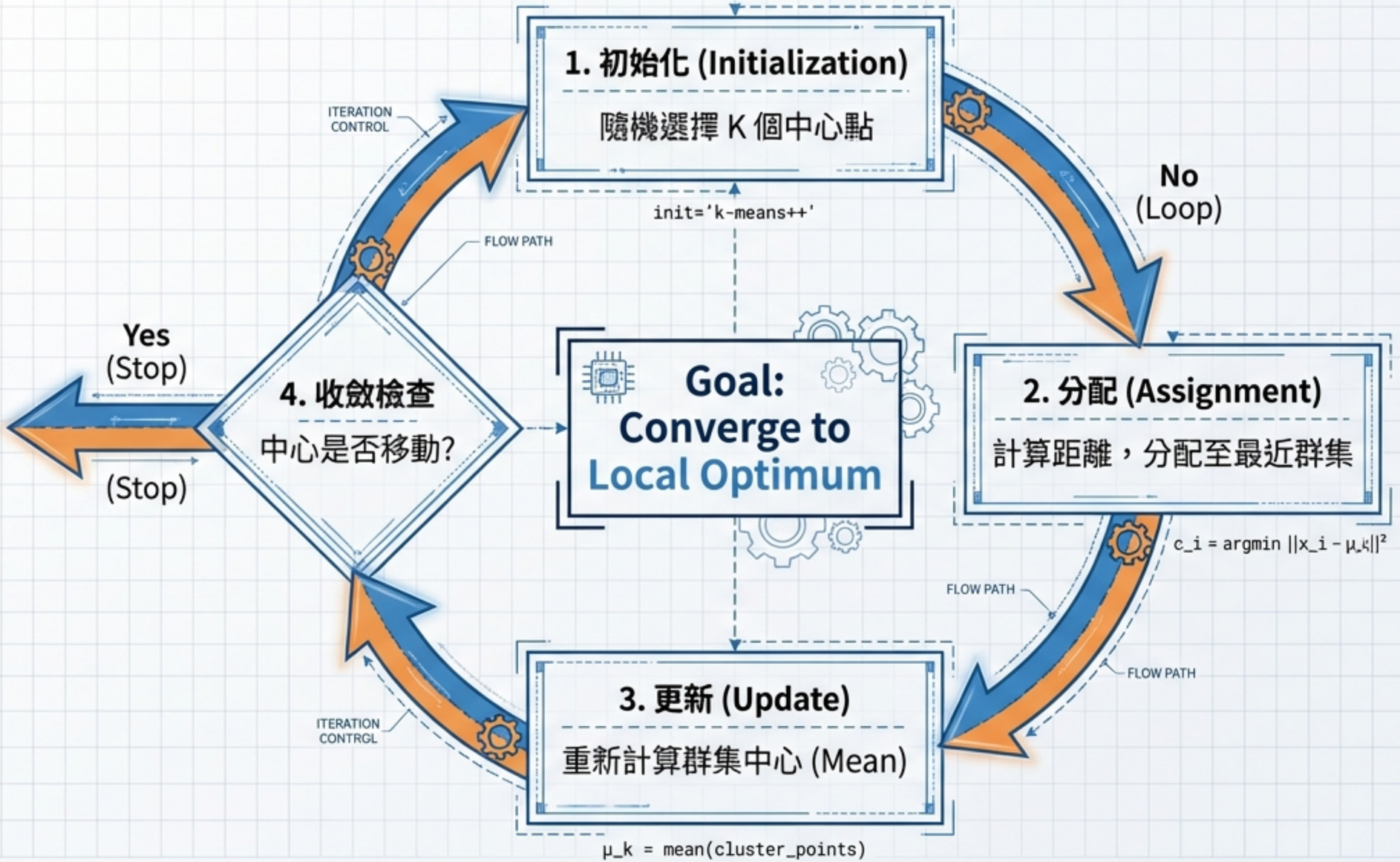
### 4. 供應商評級

原料品質自動分級

| A 級 (優) | B 級 (良) | C 級 (待加強) |
|---|---|---|
| ☑ 10% | ☐ 70% | ☐ 10.0% |

Applications

# 運作機制：迭代優化迴圈



# 數學核心：目標函數 (Inertia)

K: Number of Clusters

J: Inertia

$C_k$: Cluster k

$x_i$: Data Point

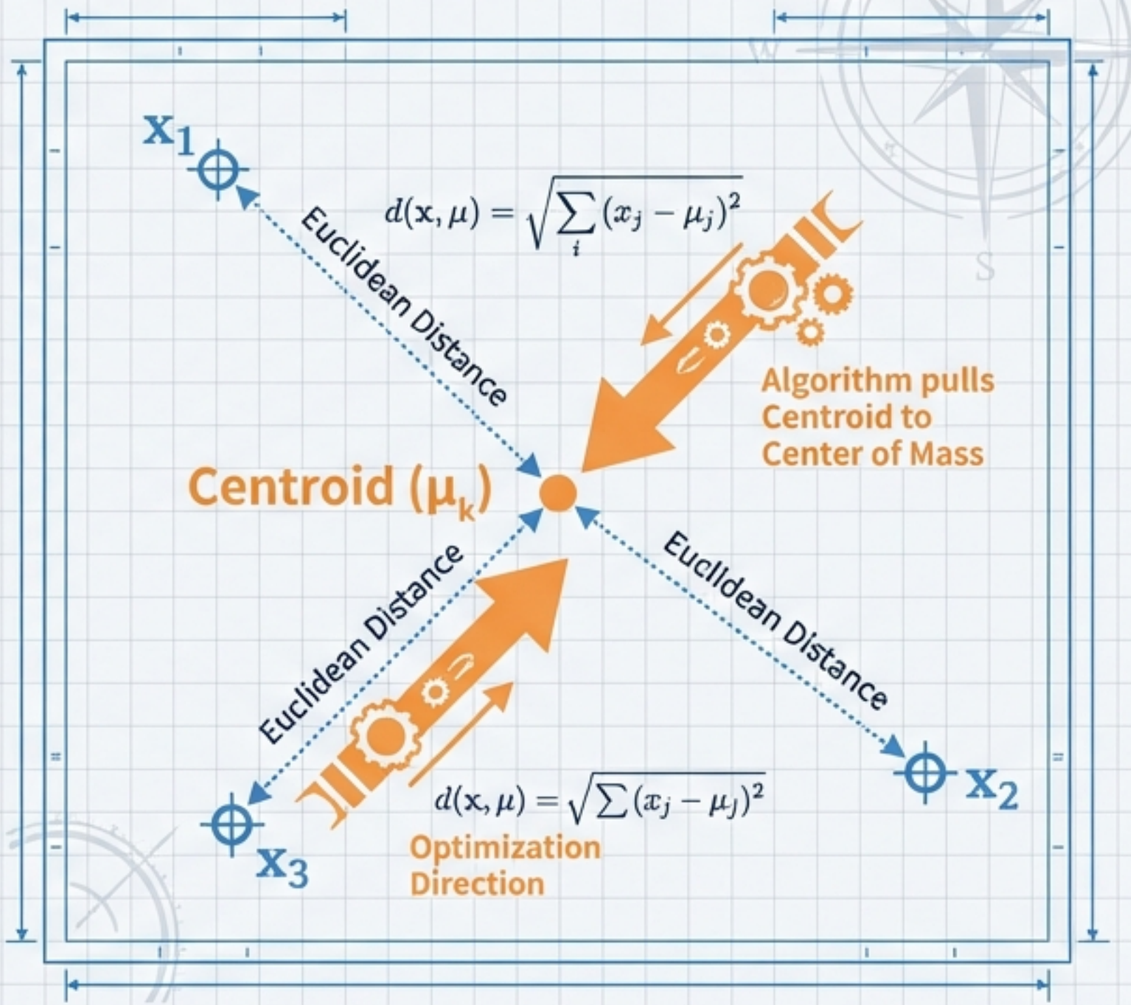
$\mu_k$: Centroid of Cluster k

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$

$x_i$: Data Point

$||...||^2$: Squared Euclidean Distance

**目標：最小化群集內距離平方和 (Minimize Inertia)**



# 技術評估：優缺點規格表

## 優點 (Advantages)

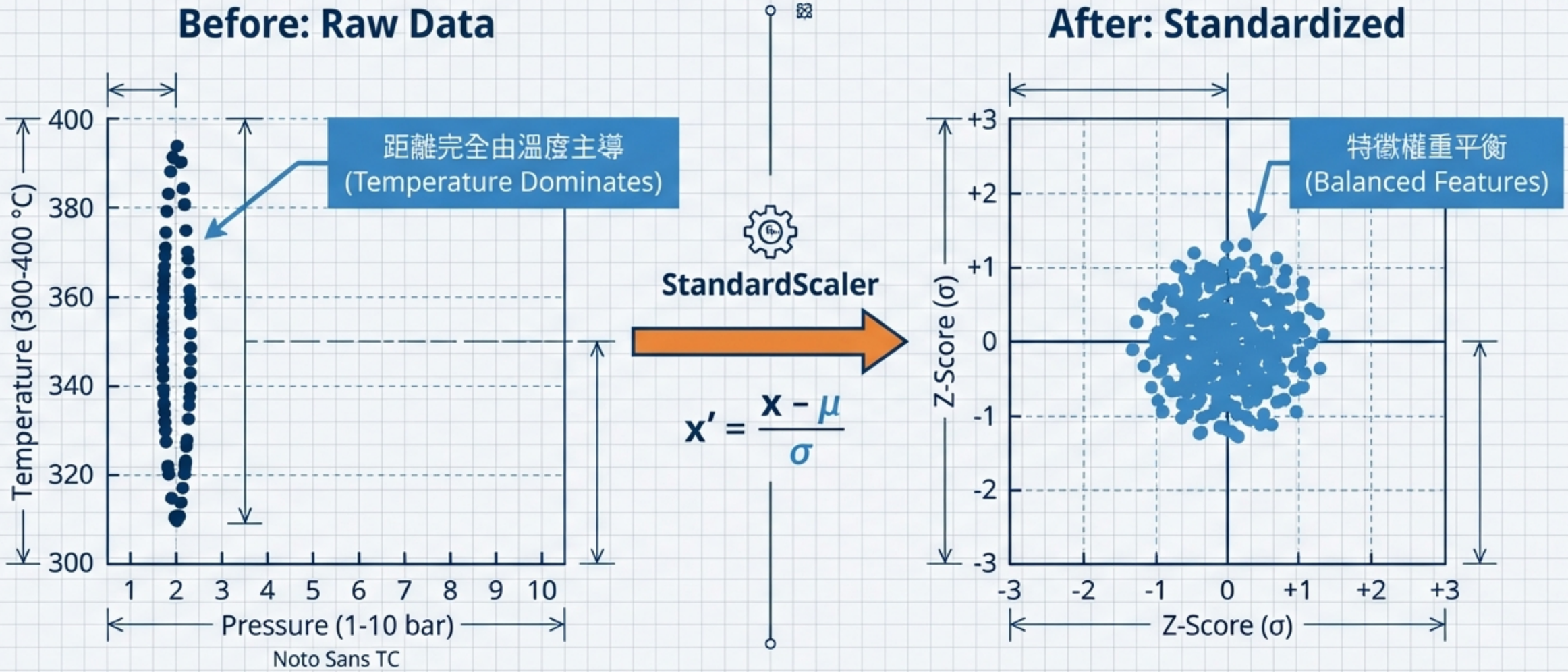
- 簡單直觀：容易解釋原理，實作門檻低。
- 高效率：適合大規模數據運算。
- 複雜度：$O(n \times K \times m \times I)$ (Linear Scalability)。
- 可擴展：易於平行化處理。

## 限制 (Limitations)

- 需指定 K 值：必須預先決定群集數量。
  → PRE-CONDITION
- 球形假設：僅適合凸形 (Convex) 且各向同性的群集。
  → GEOMETRY CONSTRAINT
- 對離群值敏感：異常值會顯著拉偏中心點。
  → NOISE SENSITIVE
- 尺度敏感：必須進行特徵標準化 (Scale Dependent)。
  → NORMALIZATION REQUIRED

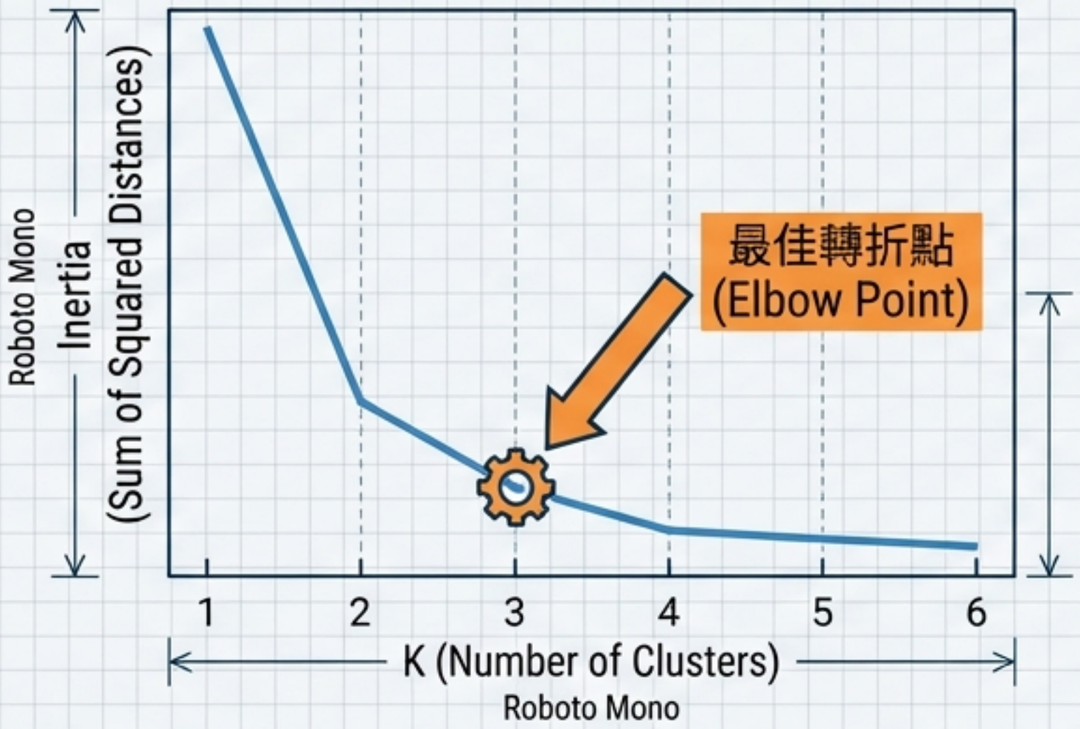
# 預處理協議：特徵標準化

**CRITICAL WARNING: Distance Calculation is Scale-Dependent**

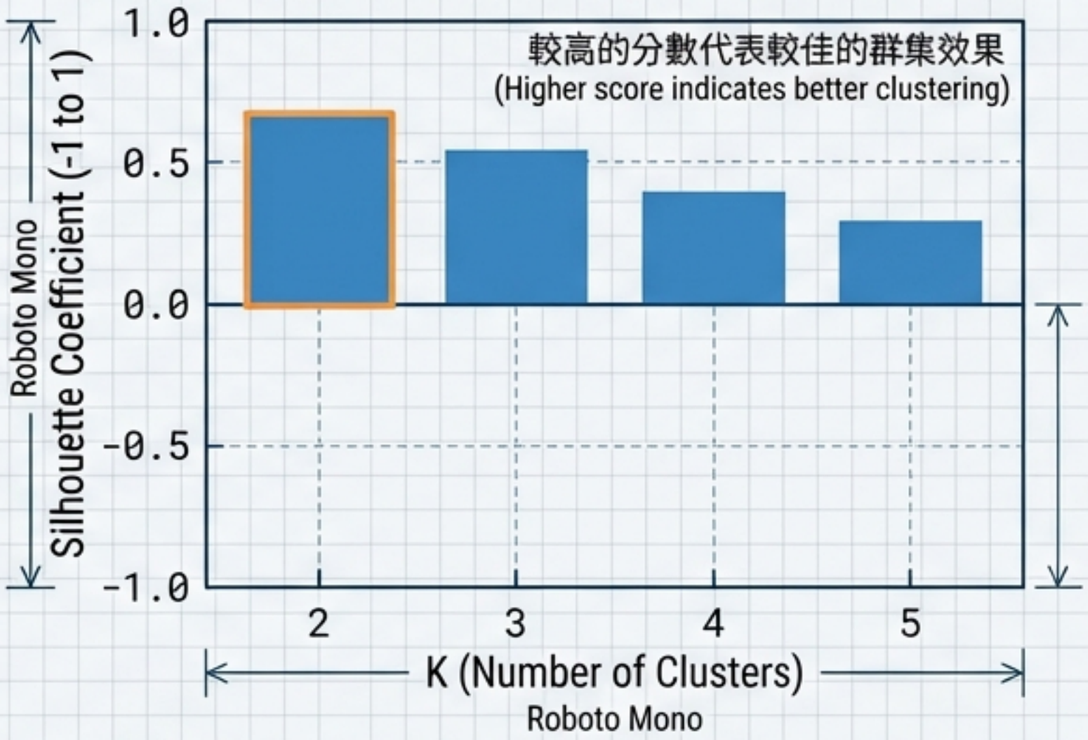


# 參數校正：決定 K 值的方法

## Method 1 : 手肘法 (Elbow Method)

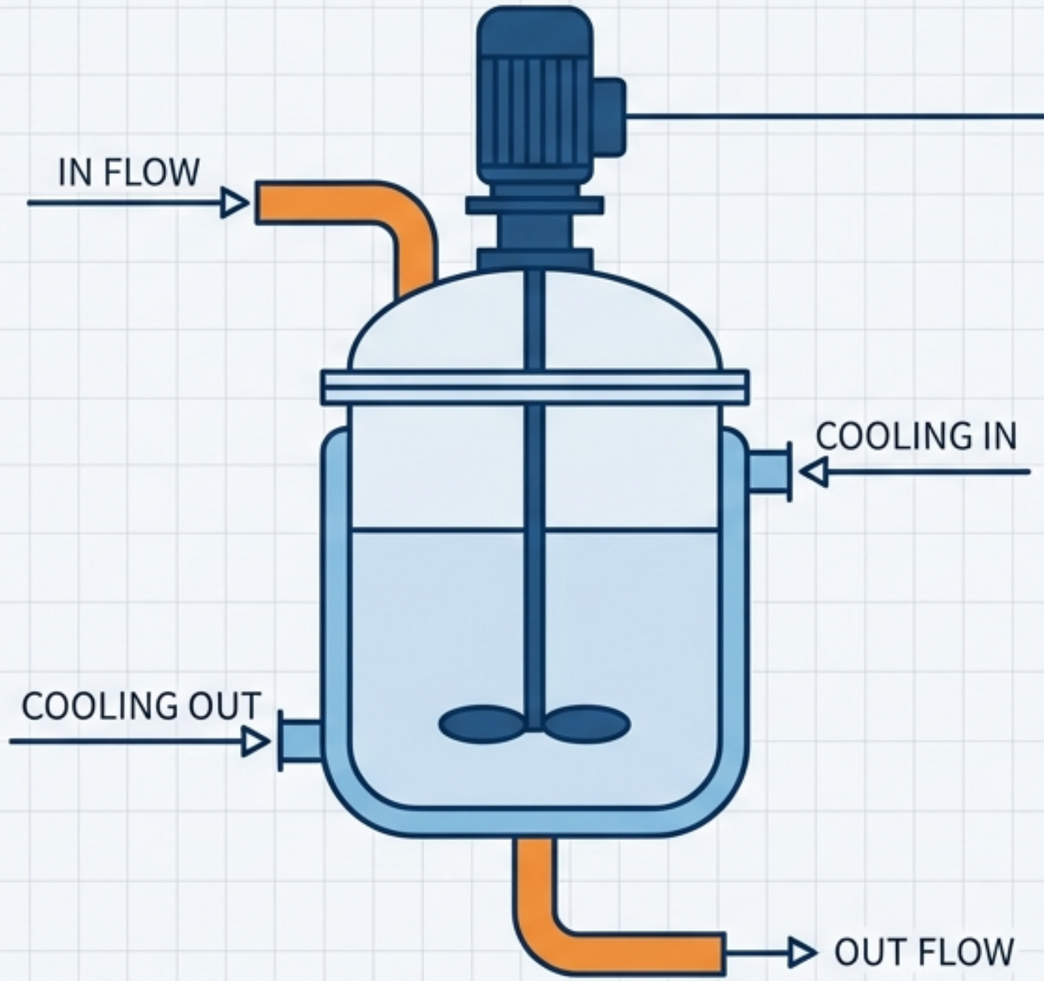



## Method 2 : 輪廓分析 (Silhouette Analysis)




⚙ **綜合考量：統計指標 + 物理意義**

# 案例研究：反應器操作模式識別



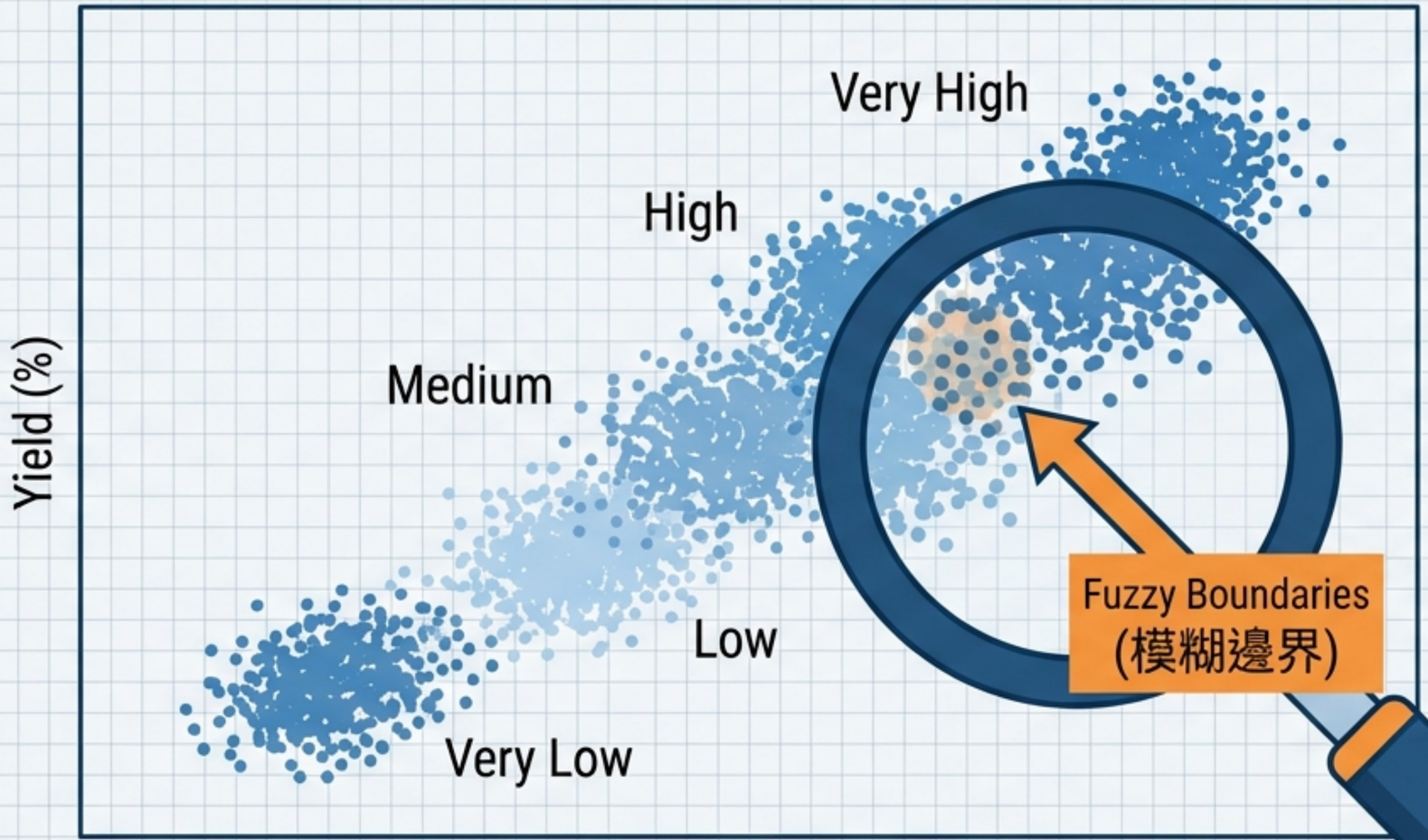

**Data Input Block**

**FEATURE VARIABLES**

1. Temperature (°C) [T]
2. Feed Flow ($m^3/h$) [F_in]
3. Cooling Flow ($m^3/h$) [F_cool]
4. Stirring Speed (rpm) [RPM]
5. Product Conc (mol/L) [C_p]
6. Yield (%) [Y]

Mission: 從 600 筆歷史數據中，識別穩定操作模式 (Operating Modes)

# 數據現況：重疊的真實世界




**Total Samples:** 600

Roboto Mono

**True Modes:** 5

Noto Sans TC

**Challenge:**

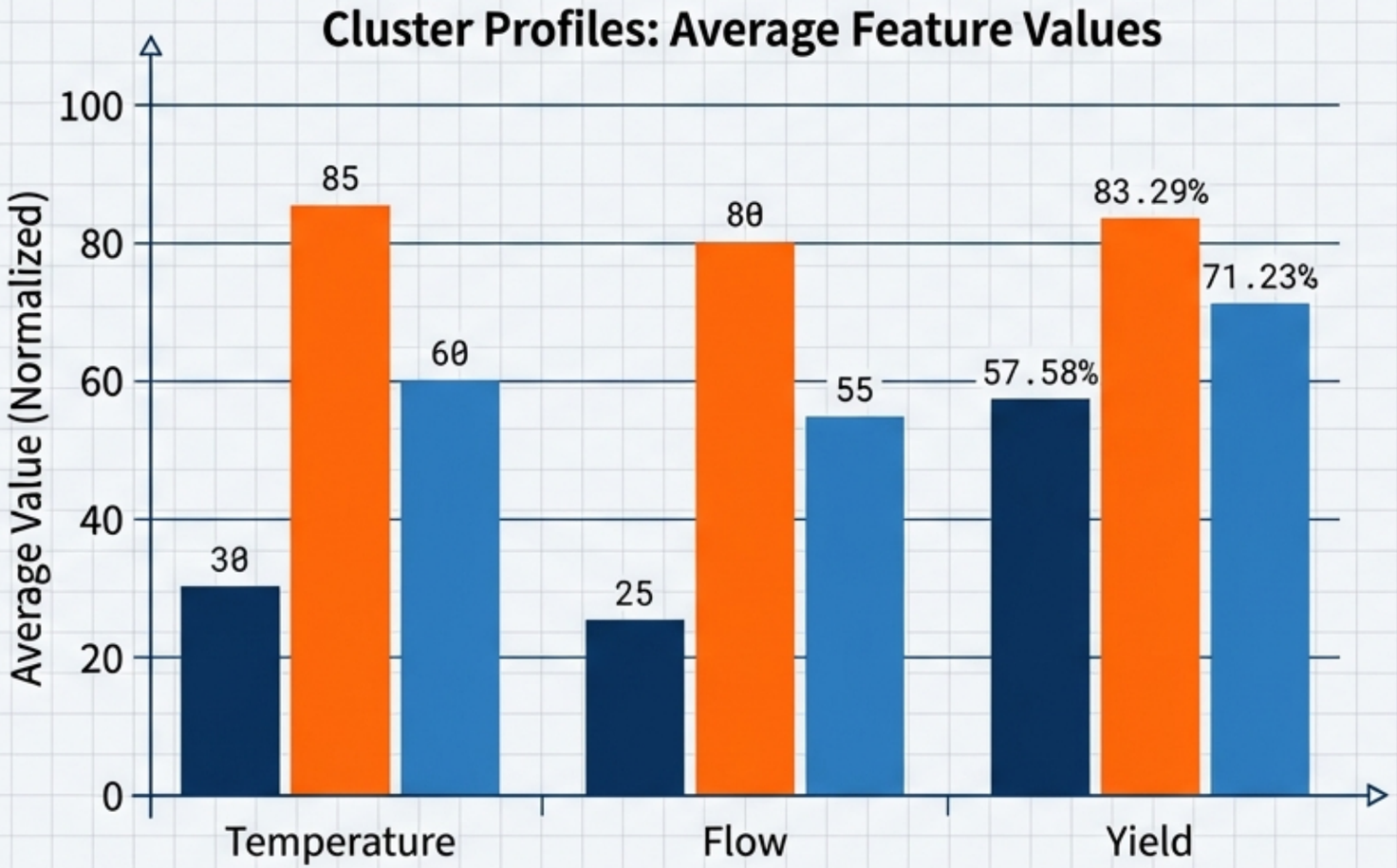
模式之間存在顯著重疊，非離散類別。

# 決策矩陣：統計 vs. 實務

| | K=2 | K=3 (Selected) | K=5 |
|---|---|---|---|
| Statistical Score | Excellent (0.552) | Good (0.437) | Acceptable (0.346) |
| Engineering Meaning | 高/低<br>(太粗糙) | 高/中/低<br>(最佳平衡) | 過度細分<br>(邊界不清) |
| Visual |  |  |  |

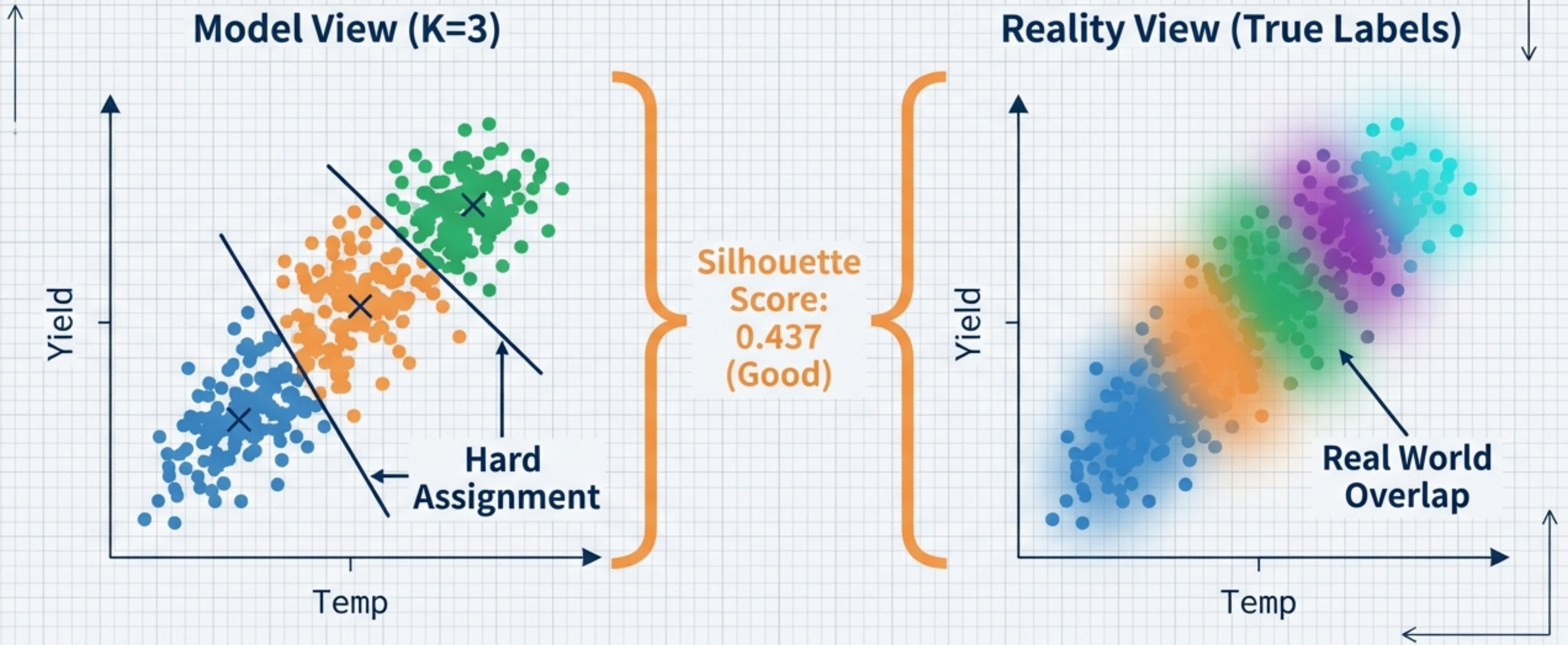
Decision: Select K=3. Trade-off between statistical purity and operational utility.

# 結果分析：群集特徵定義 (Cluster Profiles)

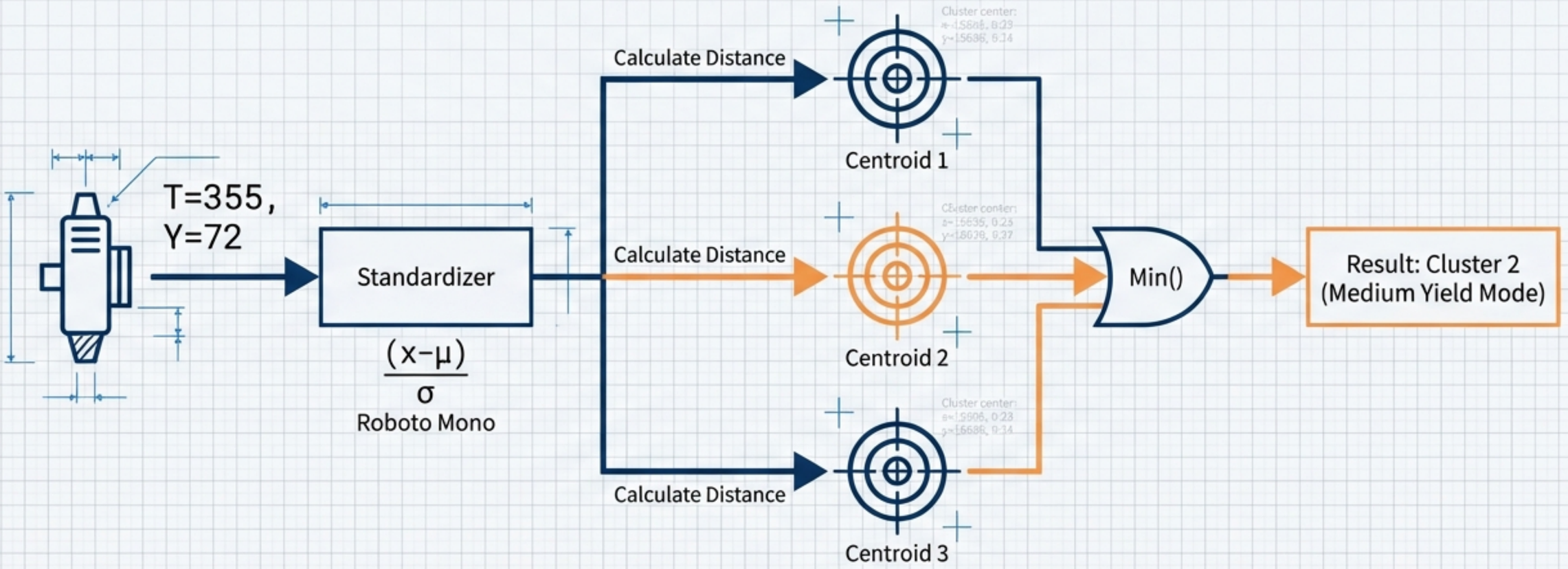



- Cluster 1 (High Yield):
  高產率區 (High Temp / High Flow)
  -> Yield 83.29%
- Cluster 2 (Medium Yield):
  中產率區
  (Medium Temp / Medium Flow)
  -> Yield 71.23%
- Cluster 0 (Low Yield):
  低產率區
  (Low Temp / Low Flow)
  -> Yield 57.58%

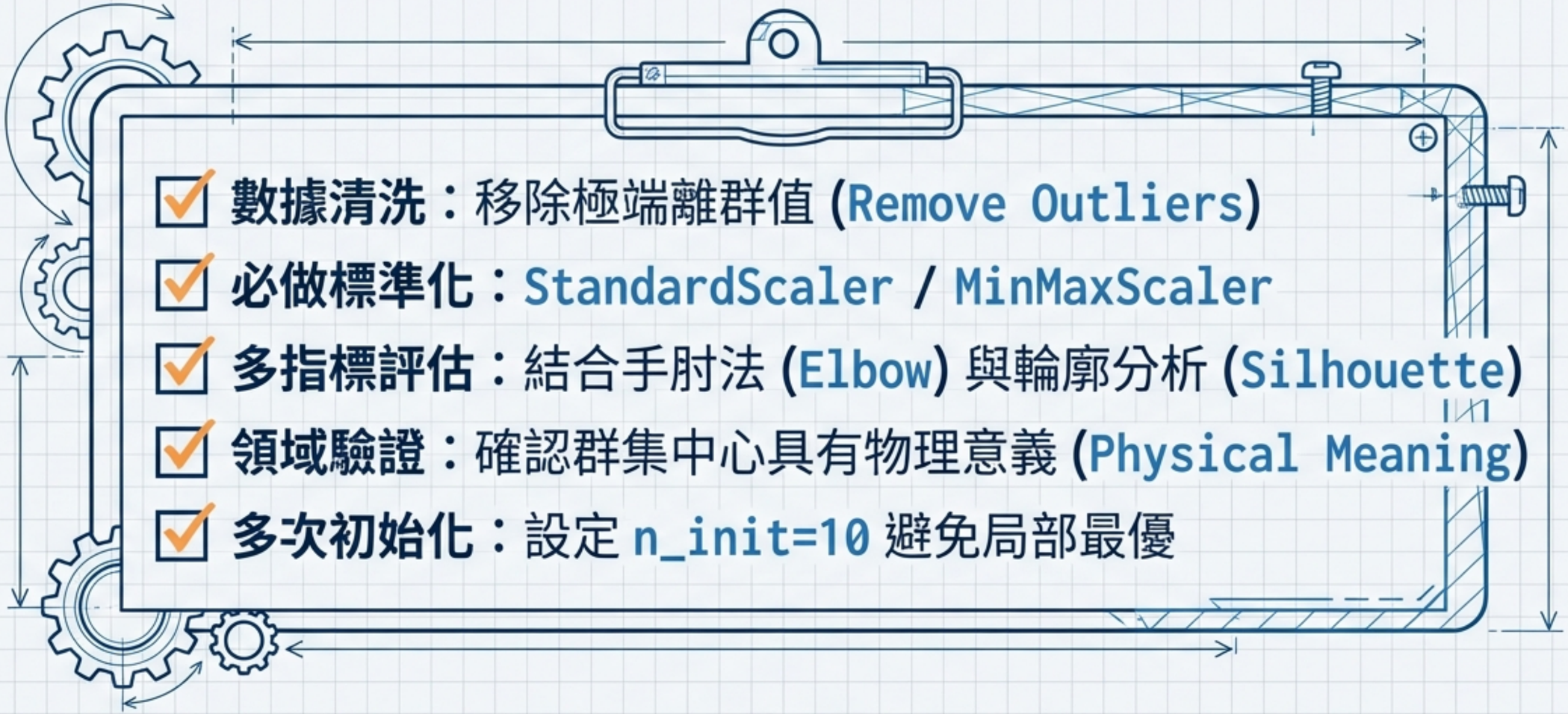
# 視覺驗證：模型預測 vs. 真實標籤



# 實際部署：新數據預測流程



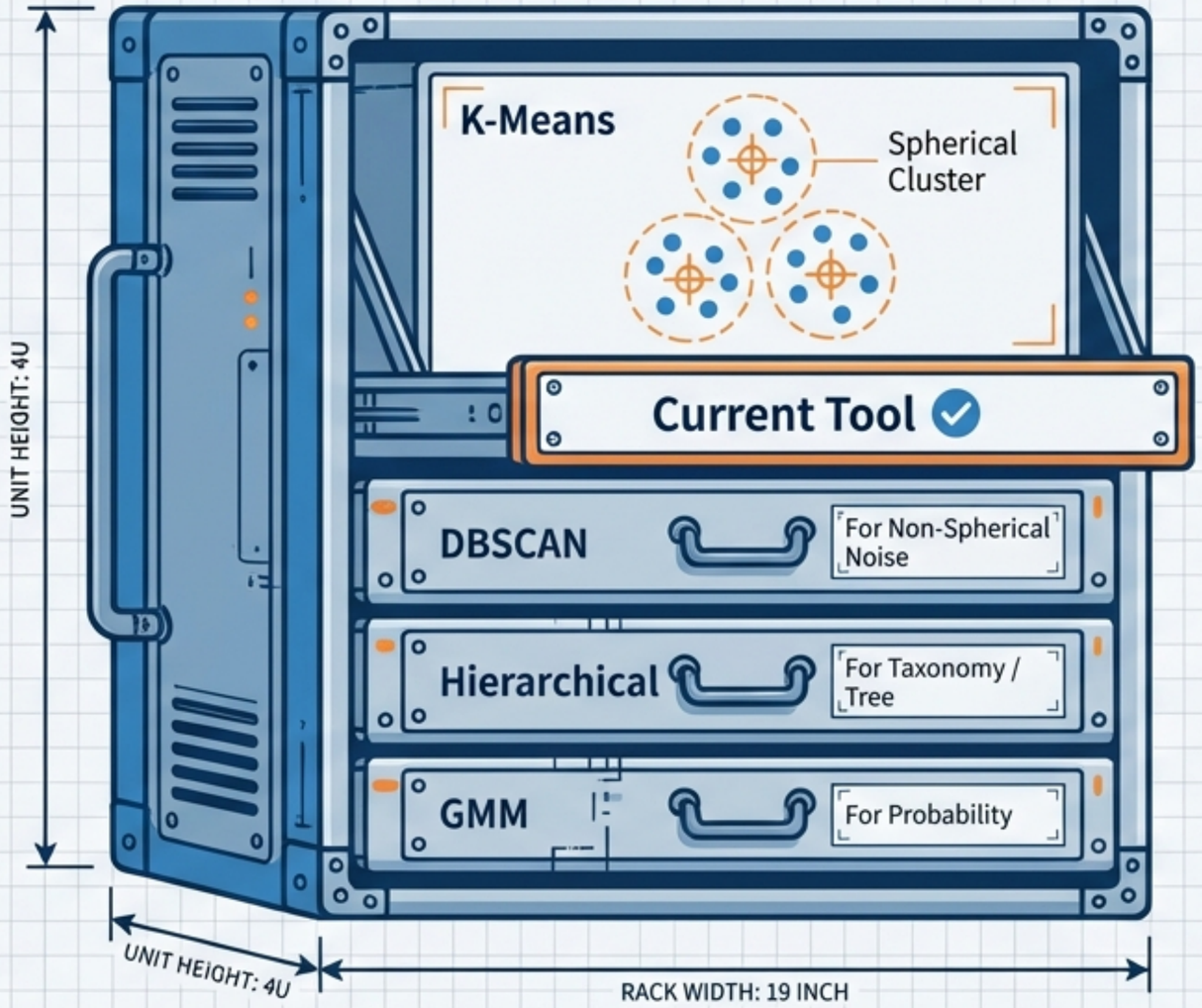
# 標準作業程序 (SOP: Best Practices)



- ☑ **數據清洗**：移除極端離群值 (Remove Outliers)
- ☑ **必做標準化**：StandardScaler / MinMaxScaler
- ☑ **多指標評估**：結合手肘法 (Elbow) 與輪廓分析 (Silhouette)
- ☑ **領域驗證**：確認群集中心具有物理意義 (Physical Meaning)
- ☑ **多次初始化**：設定 n_init=10 避免局部最優

*統計指標提供建議，但工程師必須做最終決策。*

# 工具箱總結與下一步




## Next Unit: Hierarchical Clustering (階層式分群)

進一步探索層次結構

Swiss Style meets Industrial Blueprint