

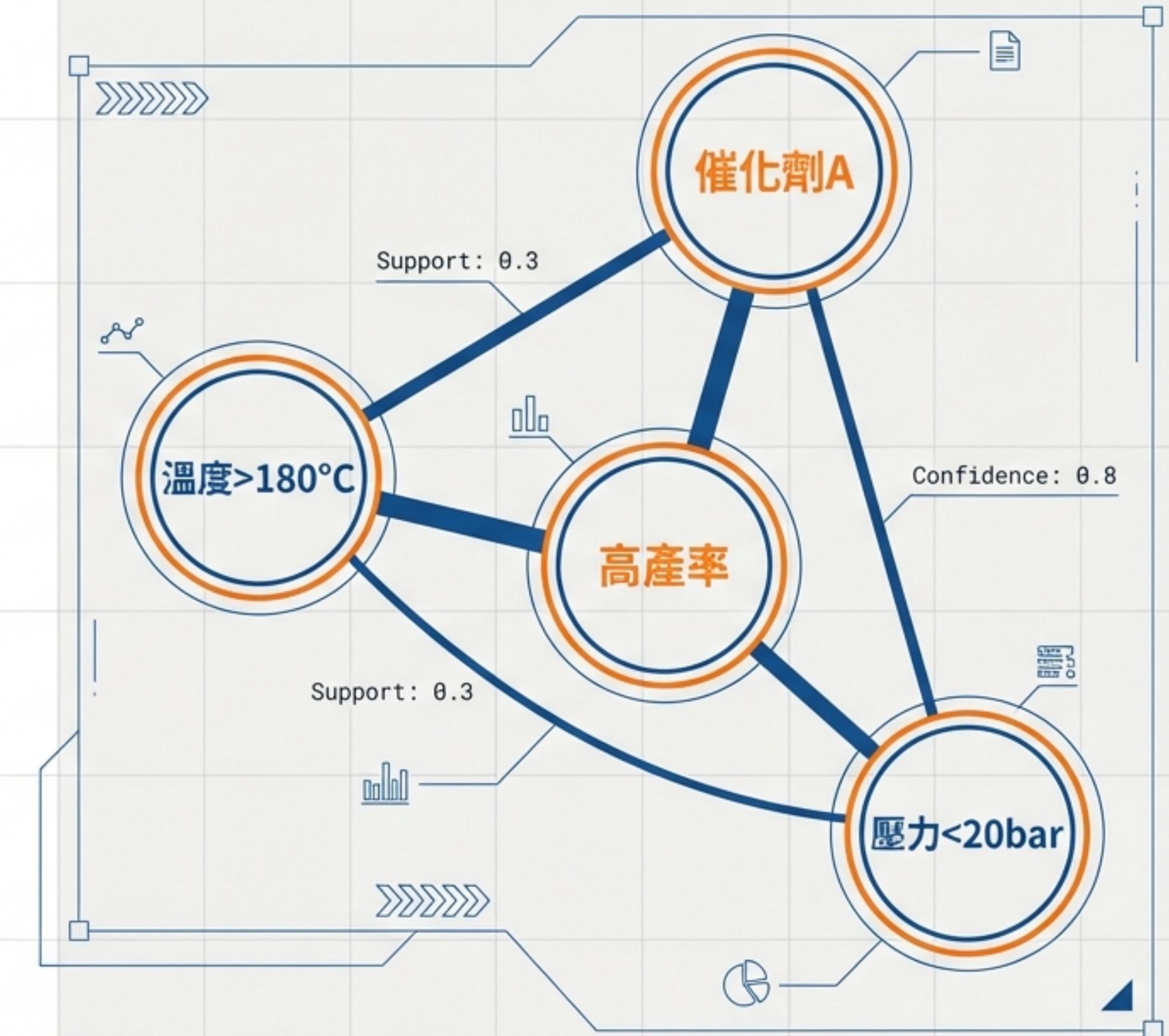
AI 在化工上的應用：從原理到實踐

Unit 08 關聯規則學習總覽

授課教師：莊曜禎 助理教授

學期：114學年度第2學期

實驗室：逢甲大學 化工系 智慧程序系統工程實驗室



化工工程師的演進：從經驗法則到數據驅動發現



傳統方法 (Traditional)

經驗法則與試錯 (Intuition & Trial-and-Error)

工程師基於有限經驗猜測配方組合。
(例如：憑感覺認為催化劑 X 適合溶劑 Y)



關聯規則學習 (ARL Method)

數據驅動發現 (Data-Driven Discovery)

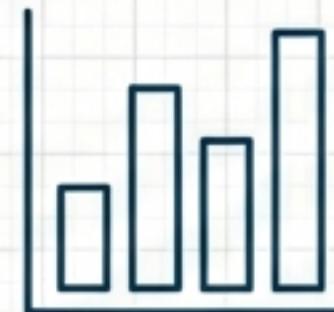
演算法自動挖掘歷史檔案中的隱藏模式。
 $\{Catalyst\ A, \ Temp > 180^{\circ}C\} \Rightarrow \{Yield > 95\%\}$

核心價值：不僅僅是驗證已知 (Checking what we know)，更是發現未知 (Discovering what we don't know)。

解密 ARL 核心：三大評估指標

基本定義：If X (前項), then Y (後項)

Support (支持度)



Key Word : 普遍性 (Frequency)

組合在數據中出現的頻率。

$$P(X \cup Y)$$

30%

Confidence (置信度)



Key Word : 可靠性 (Reliability)

當 X 出現時，Y 出現的機率。

$$P(Y|X)$$

60%

Lift (提升度)



Key Word : 關聯強度 (Correlation)

兩者關聯是真實的還是巧合？

規則

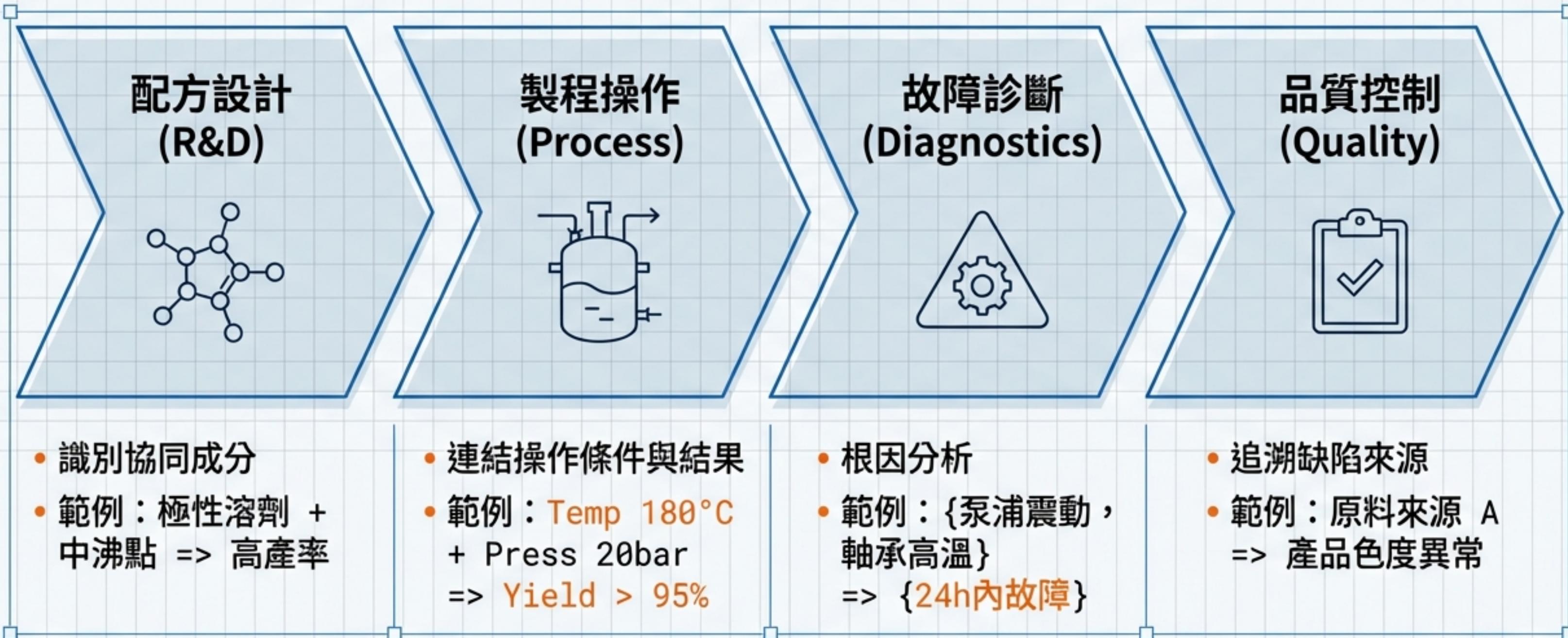
$\text{Lift} > 1$ (Positive), $\text{Lift} < 1$ (Negative)

1.5

化工範例 (Chemical Example)

規則 : {催化劑A} => {高產率} (使用催化劑A使高產率的可能性提升了 1.5 倍)

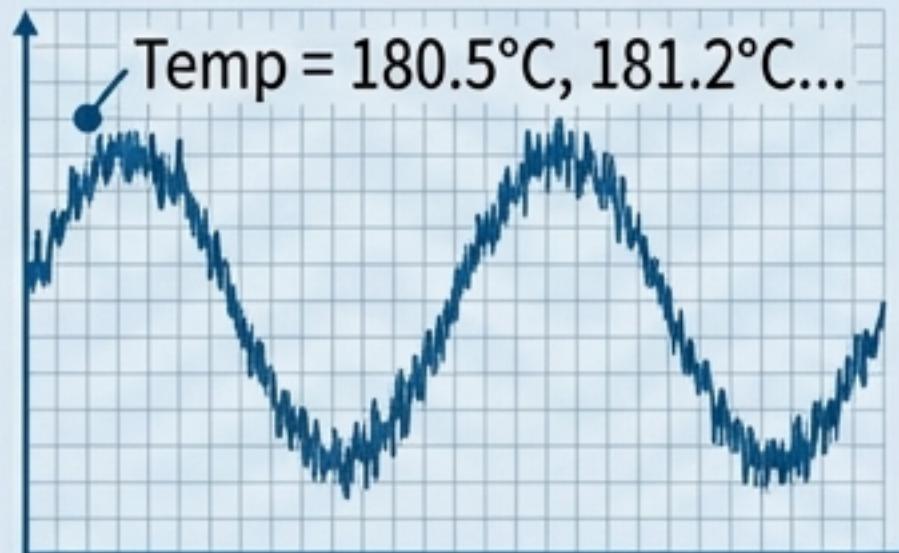
ARL 在化工全生命週期的價值



關鍵挑戰：連續變數的特徵工程

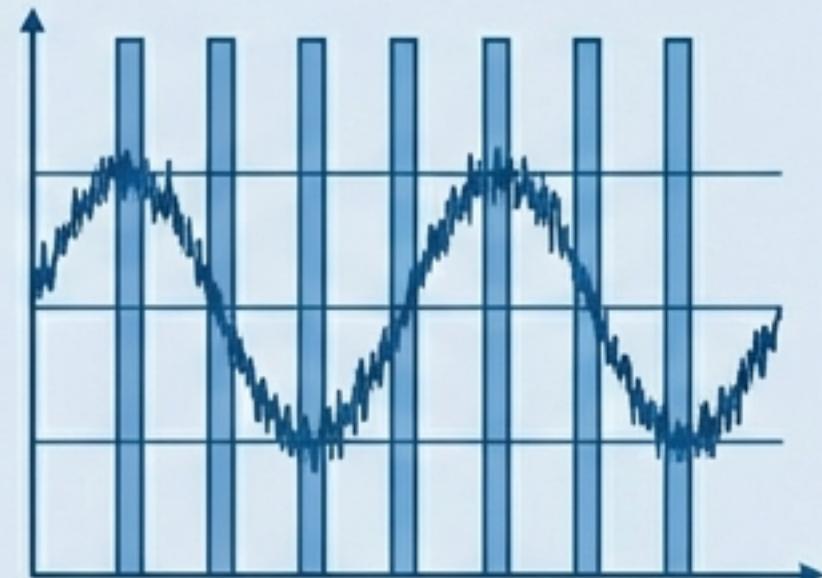
從連續信號到類別規則 (From Continuous to Categorical)

原始數據 (Raw Data)

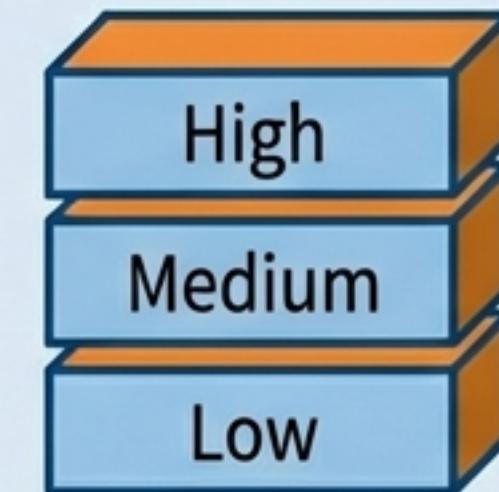


問題：ARL 需要類別型數據
(Items)

離散化 (Discretization / Binning)



轉換後特徵 (Transformed Features)

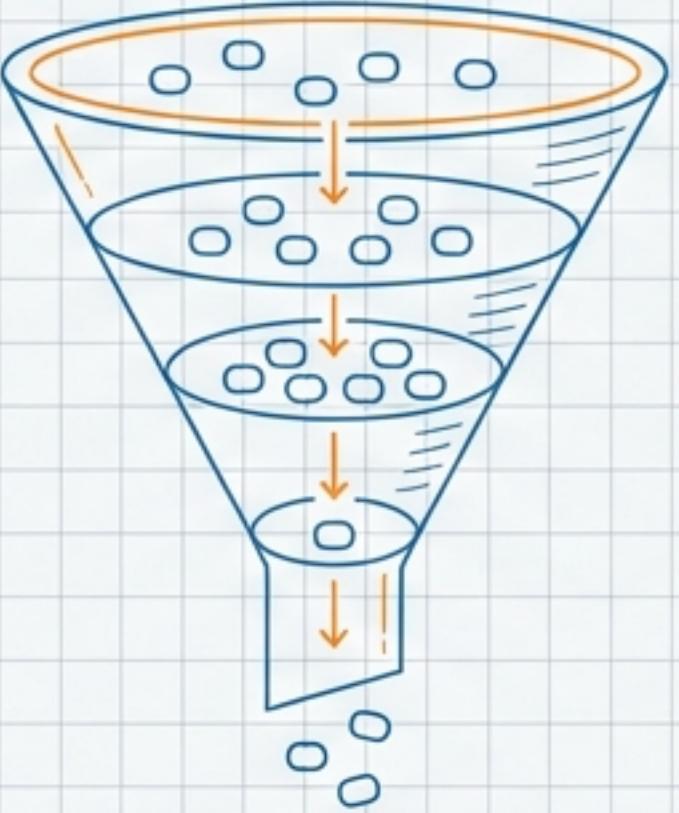


Temp_State =
{High, Medium, Low}

- 最佳實踐 (Best Practice)：使用領域知識 (Domain Knowledge) 設定閾值 (例如：沸點、凝固點、安全界限)。
- 警告：Garbage In, Garbage Out。

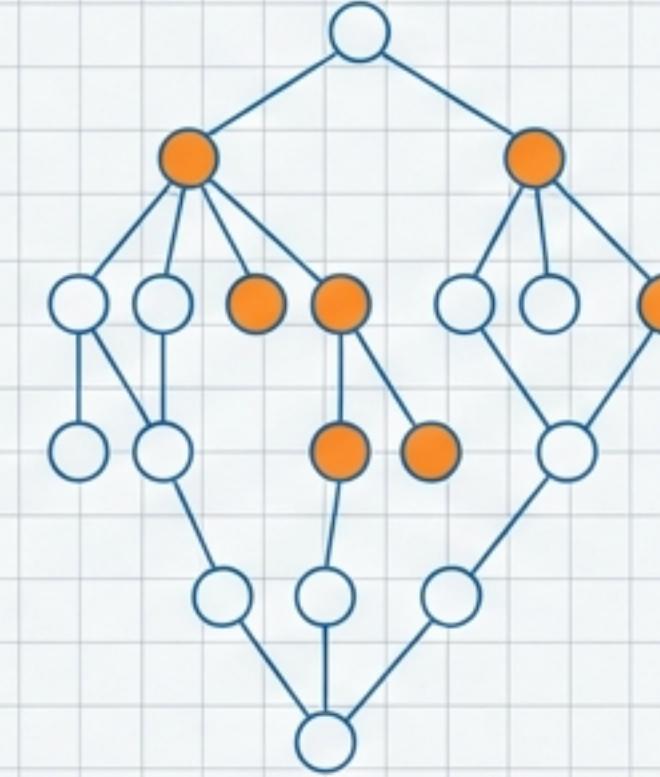
演算法比較：如何挖掘規則？

Apriori 演算法



指標	Apriori	FP-Growth
• Speed	Slow (慢)	Fast (快)
• Memory	Low (低)	High (高)
• Complexity	Simple (簡單)	Complex (複雜)

FP-Growth 演算法



- 機制：候選生成與剪枝 (Candidate Generation)
- 特點：易於解釋，但 I/O 成本高
- 適用：實驗室數據、教學、小規模數據 (<10k)

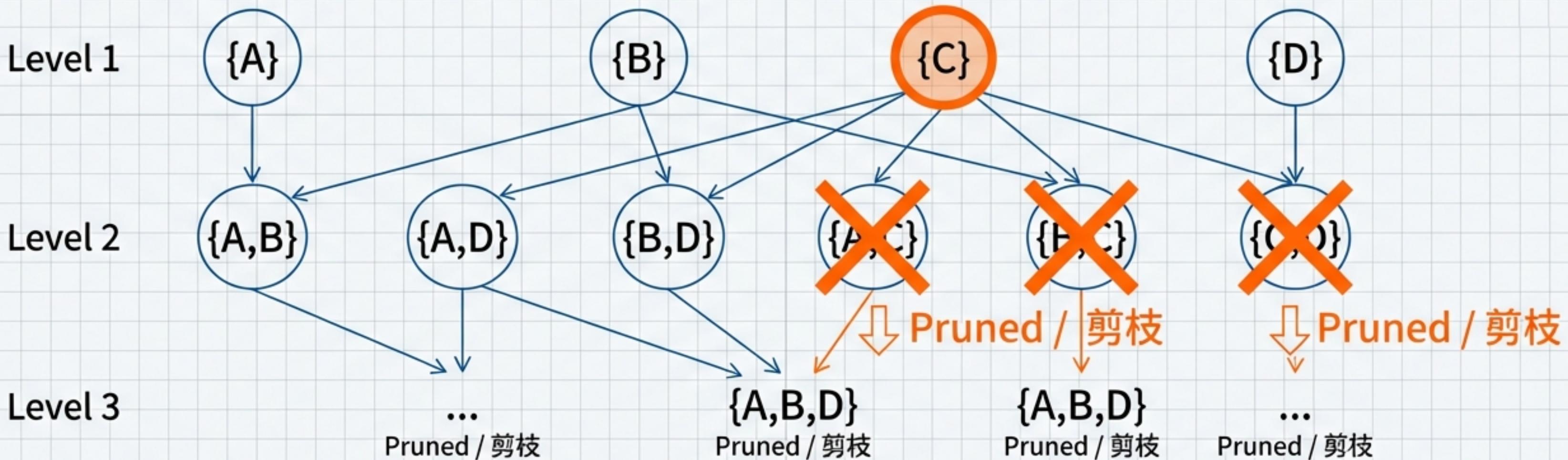
- 機制：模式增長 (Pattern Growth)
- 特點：無候選生成，建立 FP-Tree，速度快
- 適用：大數據、即時 SCADA 分析、生產環境

深入解析：Apriori 原理

向下封閉性 (Downward Closure Property)

核心邏輯：若一個項目集不頻繁，其超集也不頻繁。

$$\text{Support}(\{A\}) < \text{Min_Sup} \Rightarrow \text{Support}(\{A, B\}) < \text{Min_Sup}$$



適用場景：適合分析實驗室筆記本數據 (Small Data)。

深入解析：FP-Growth 演算法

頻繁模式樹 (Frequent Pattern Tree)

Raw Data (原始數據)

T1: {A, B, E}

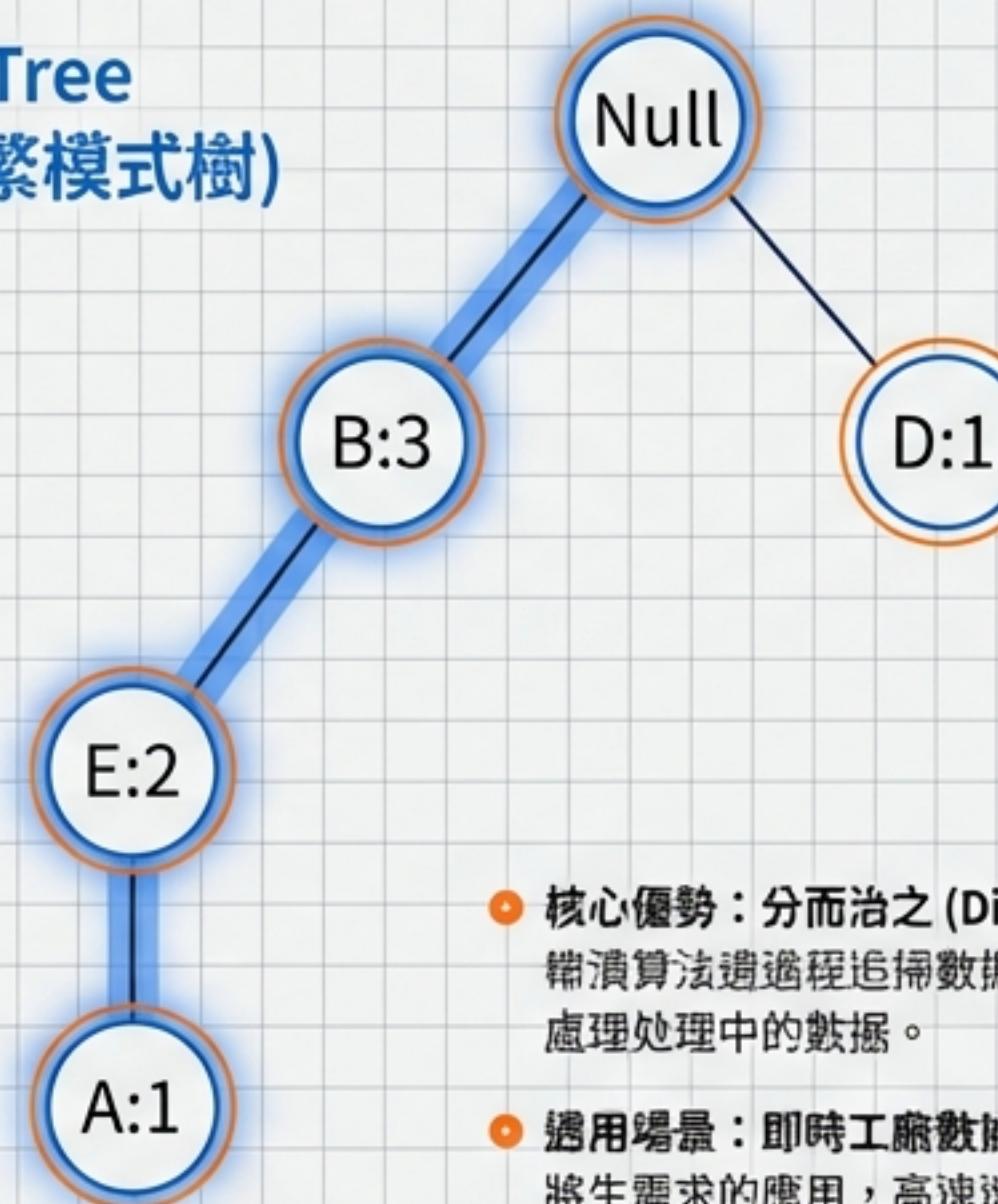
T2: {B, D}

T3: {B, E}

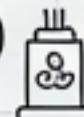


Compress & Map
(只掃描兩次)

FP-Tree (頻繁模式樹)



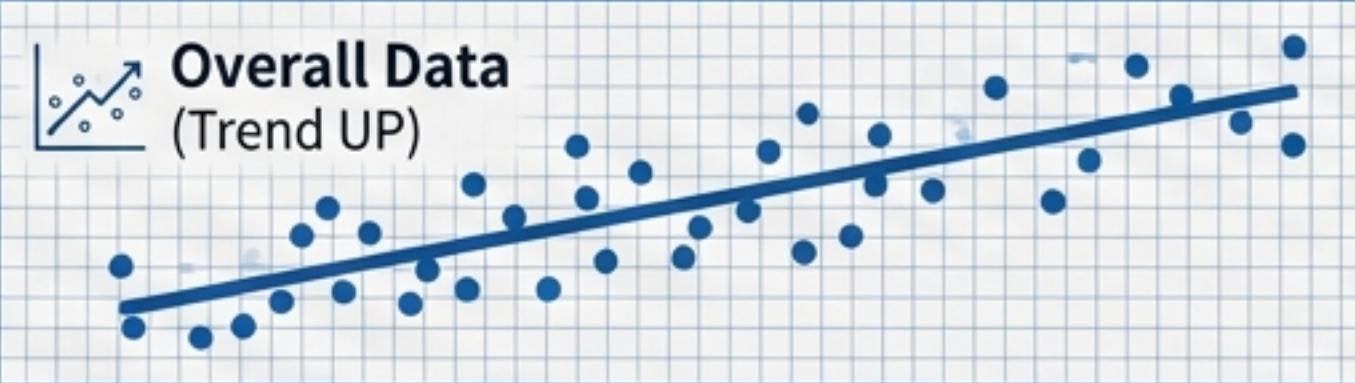
- 核心優勢：分而治之 (Divide and Conquer)
掃描算法在過程追掃數據的過程中，跳過處理中的數據。
- 適用場景：即時工廠數據 (SCADA Logs)
將生需求的應用，高速率數據流。



評估陷阱：高置信度不等於因果性

辛普森悖論 (Simpson's Paradox)

辛普森悖論視覺化 (The Paradox Visualized)



趨勢在分組後可能逆轉！



進階評估指標 (Advanced Metrics)

➤ Conviction (確信度)

測量意外性 (Unexpectedness)。



➤ Leverage (槓桿值)

測量與獨立性的偏差。



➤ Zhang's Metric

對稱性關聯指標。



警告：Correlation ≠ Causation。必須在不同操作條件下驗證規則。



過濾雜訊：尋找黃金規則

Unit 01

Raw Rules (1000+)

Roboto Mono

去除冗餘 (Redundancy)

Occam's Razor: 移除複雜且無增益的規則。

統計顯著性 (Statistical Sig)

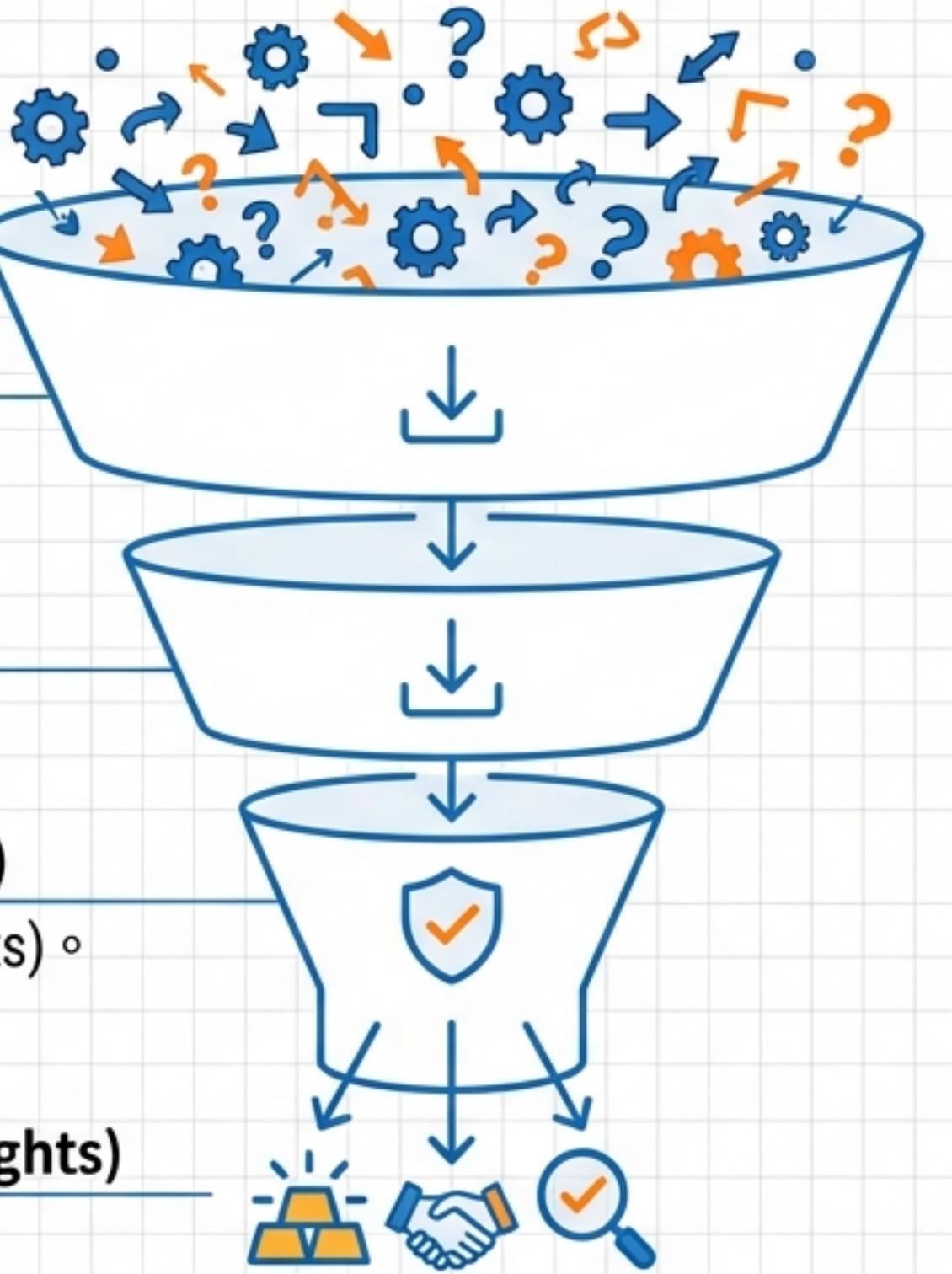
Chi-square Test ($p\text{-value} < 0.05$)

領域知識 (Domain Knowledge)

物理限制過濾 (Physics Constraints)。

Golden Rules (Actionable Insights)

Roboto Mono



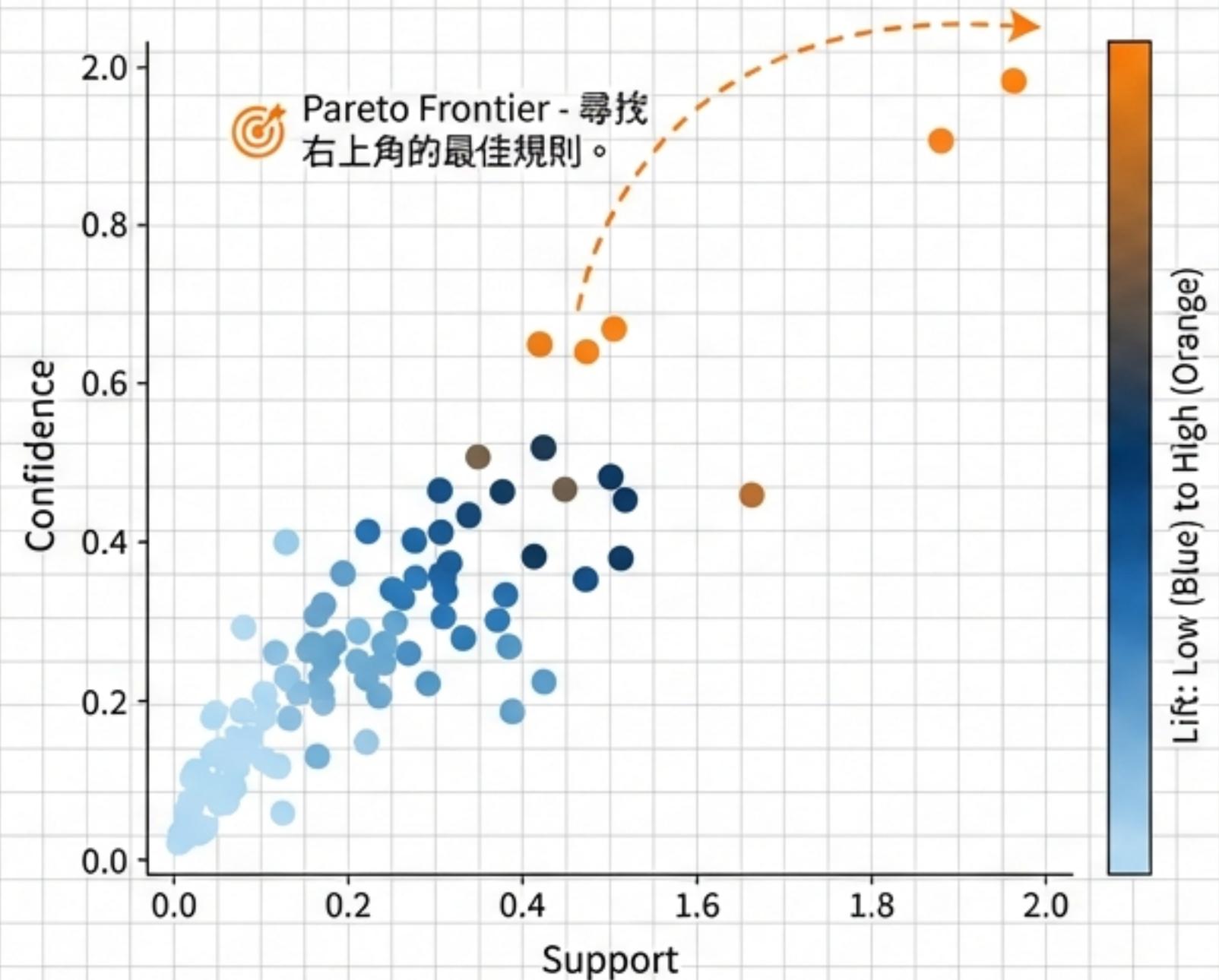
不平衡比率 (Imbalance Ratio)

處理罕見事件 (如工安事故) 時，需動態調整 Support 閾值。

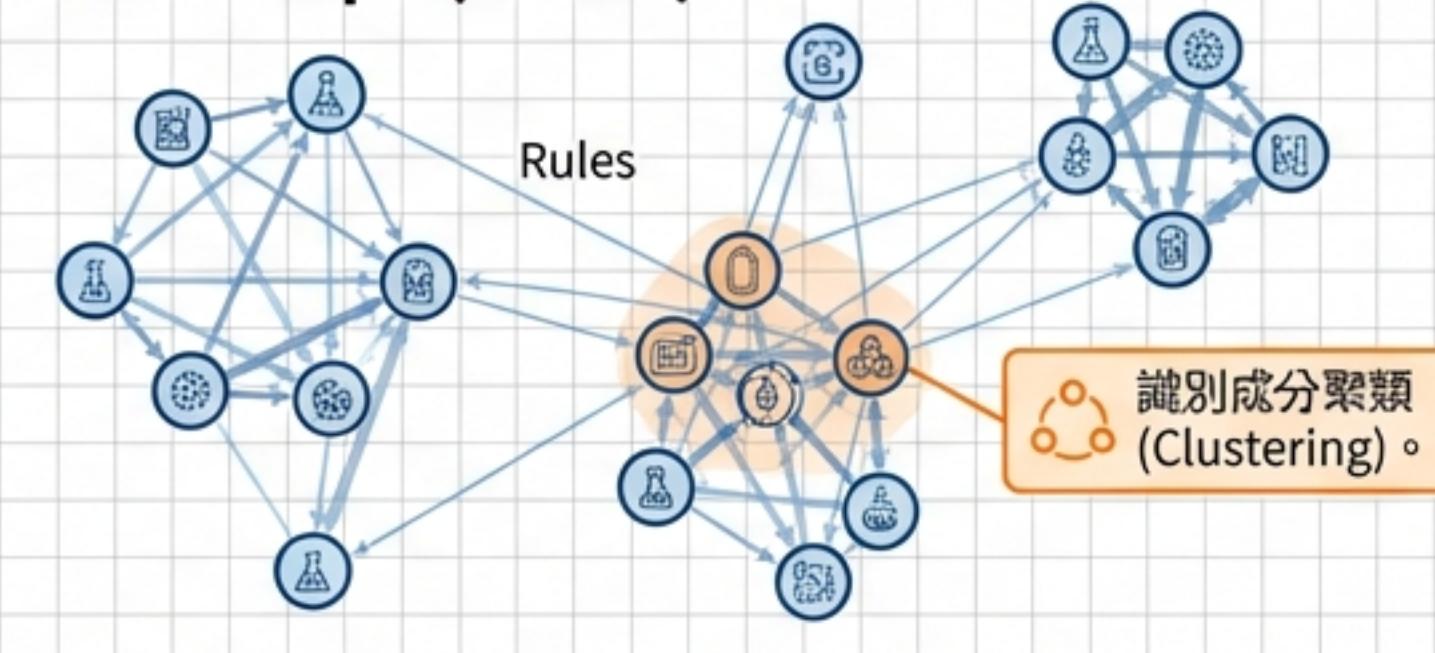
視覺化呈現：將數據轉化為洞察

如何向管理層展示結果

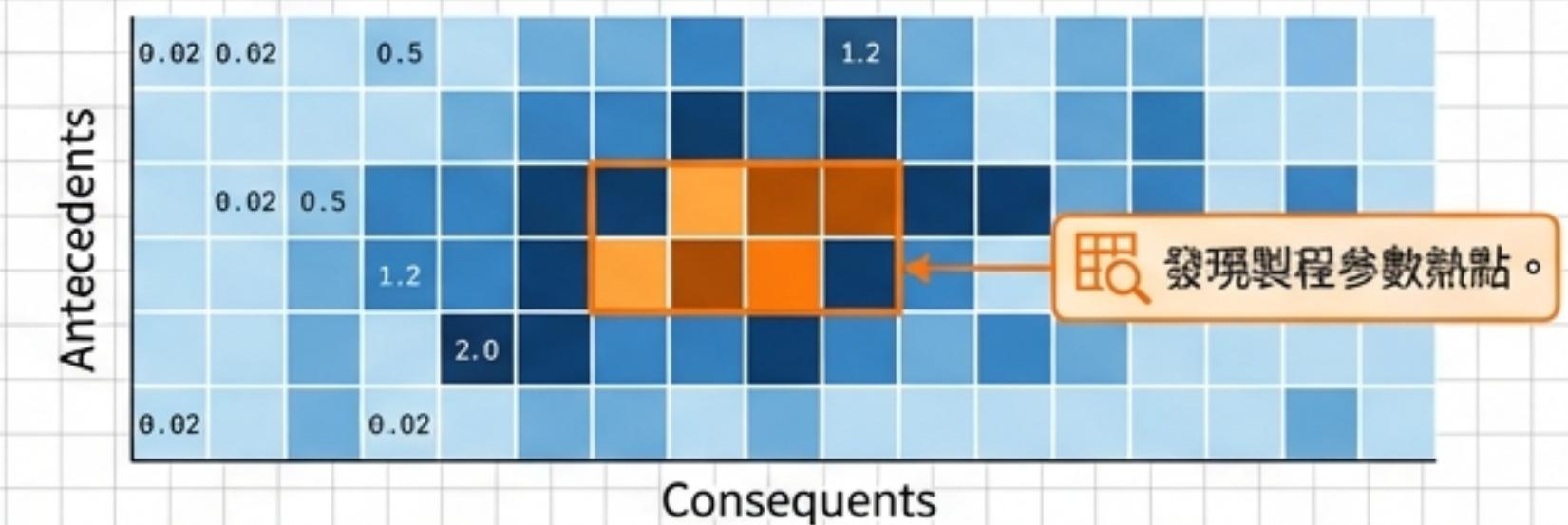
Scatter Plot (散佈圖)



Network Graph (網絡圖)

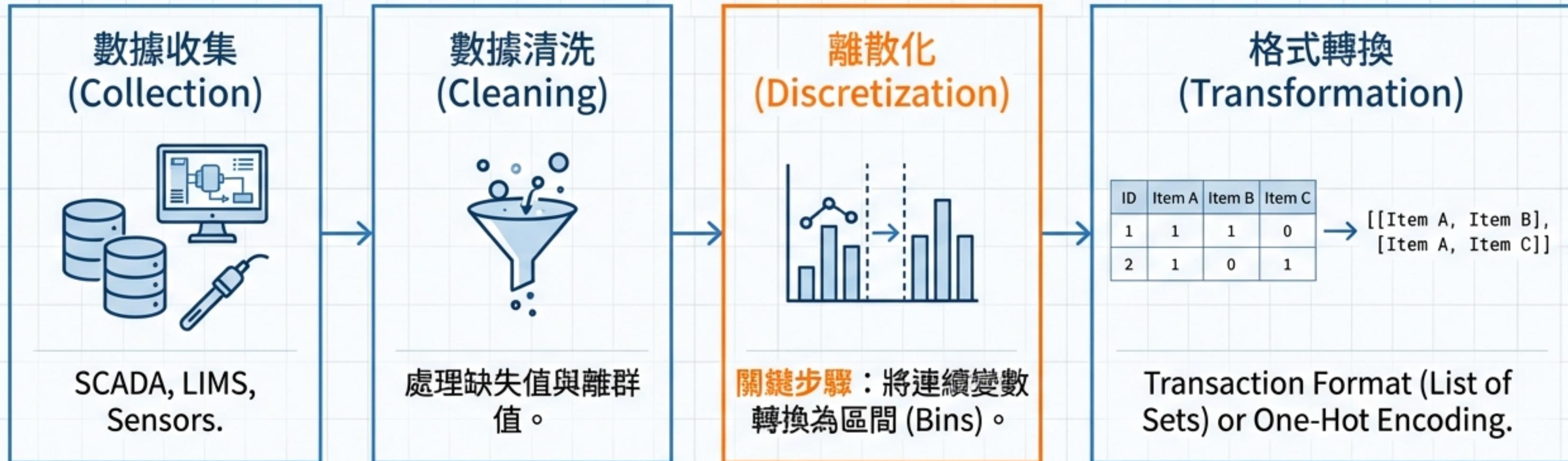


Matrix View (熱力圖)



實作流程 Part 1：數據準備

Data to Transaction Format



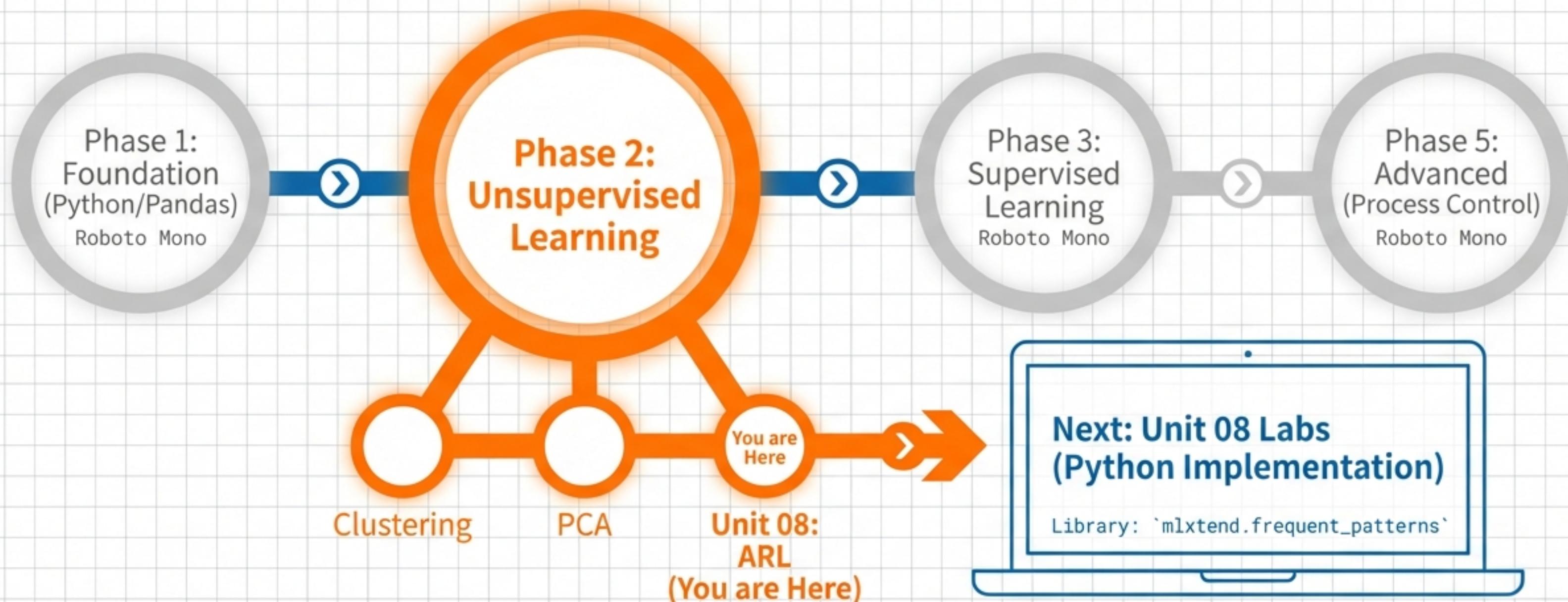
原則：Garbage In, Garbage Out. 分箱策略決定規則品質。

實作流程 Part 2：決策與衝突處理

Unit 01



課程地圖：學習路徑



結語：您是未來的定義者

數據發現關聯 (Correlation)；工程師判定因果 (Causality)



下一步 (Next Step)：前往 Unit 08 Labs

開始挖掘您的“Digital Gold”。

```
from mlxtend.frequent_patterns import apriori
```