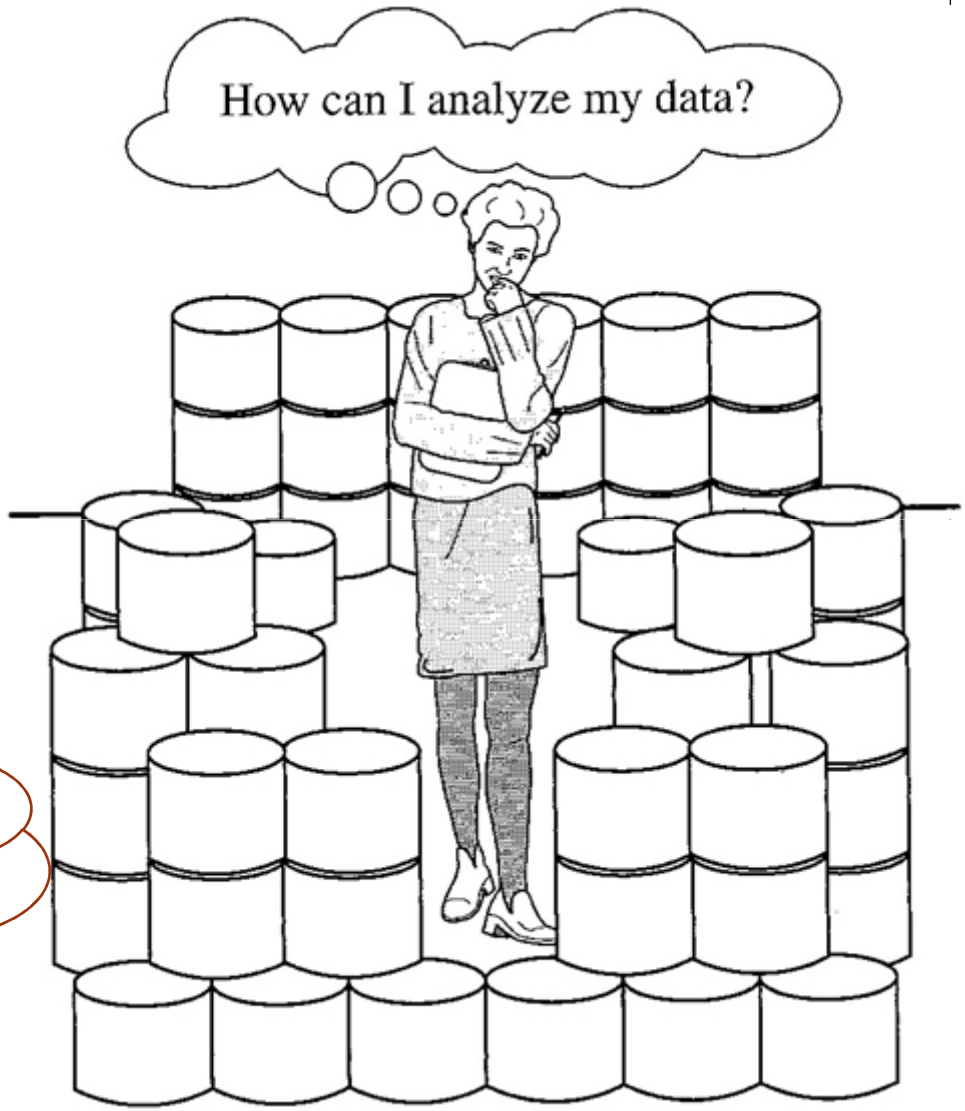
資料探勘
(Data Mining)

簡介
INTRODUCTION

動機

- 資訊通常「**隱藏**」在並非顯而易見的資料之中
- 分析師需花費數週才可發現有用的資訊
- 多數的資料並未經過分析



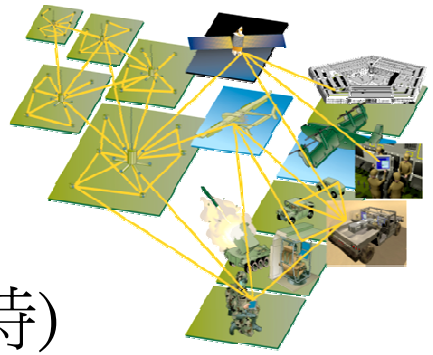
We are **drowning in data**, but **starving for knowledge**!

應用 (商業)

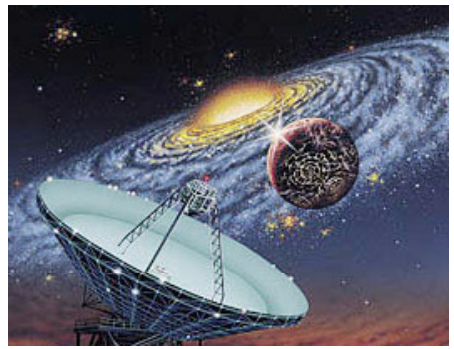
- 收集了大量的資料
  - 來自網站和電子商務交易
  - 來自商店的購物紀錄
  - 來自銀行和信用卡交易紀錄

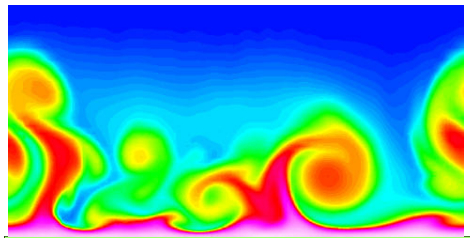
- 電腦設備的功能越來越強大，且價錢越來越便宜
- 競爭壓力越來越高
  - 以提供更好、客製化的服務作爲競爭優勢（如顧客關係管理）

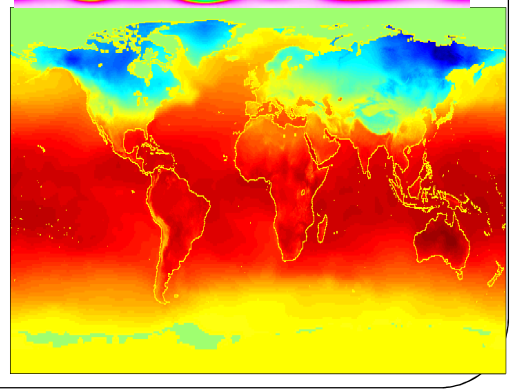
應用 (科學)



- 資料收集和儲存技術大幅提升 (GB/小時)
  - 利用衛星收集資料
  - 太空望遠鏡收集氣候資料
  - 微陣列技術產生基因的描述性資料
- 傳統技術無法分析這些原始資料
- 資料探勘可以協助科學家
  - 分類資料
  - 形成假設檢定







資料探勘 (Data Mining)

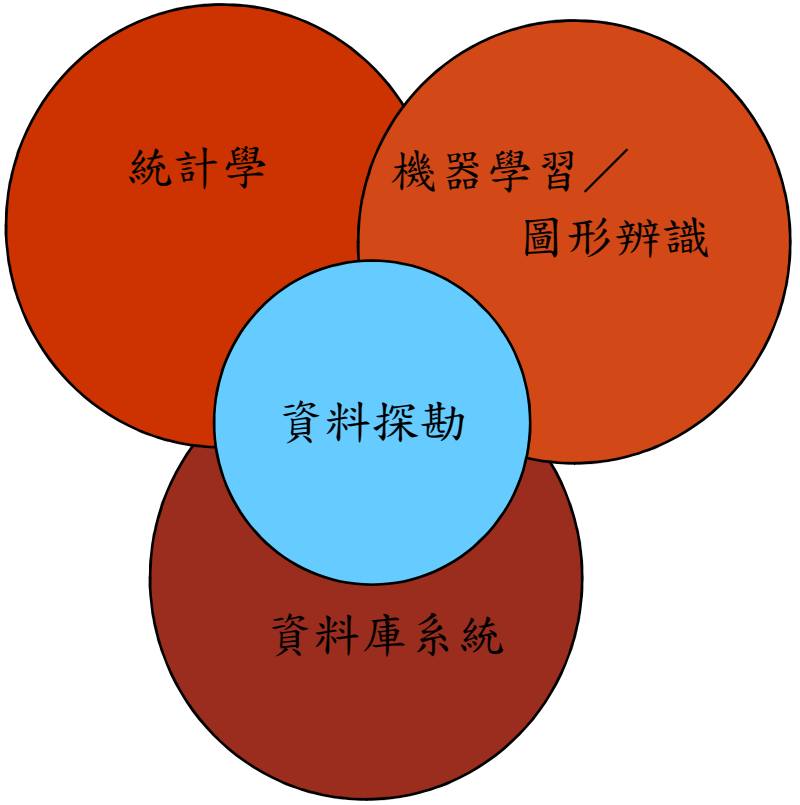
- **定義** (Definition)
  - 一種從整個資料庫裡的資料，利用一種或多種電腦技術以自動**分析**或去**擷取知識**的過程。
- **目的**
  - 在資料中發現**趨勢**與**樣式**這些知識是前所未知的
- **方法**
  - 歸納法學習
  - 概念明確、具體可知的例子，構造出通用的概念定義
- **例如：**
  - 唱片公司在專爲老年人設計的雜誌中，打饒舌音樂的廣告是否有意義？

知識發現 (Knowledge Discovery in Databases, KDD)

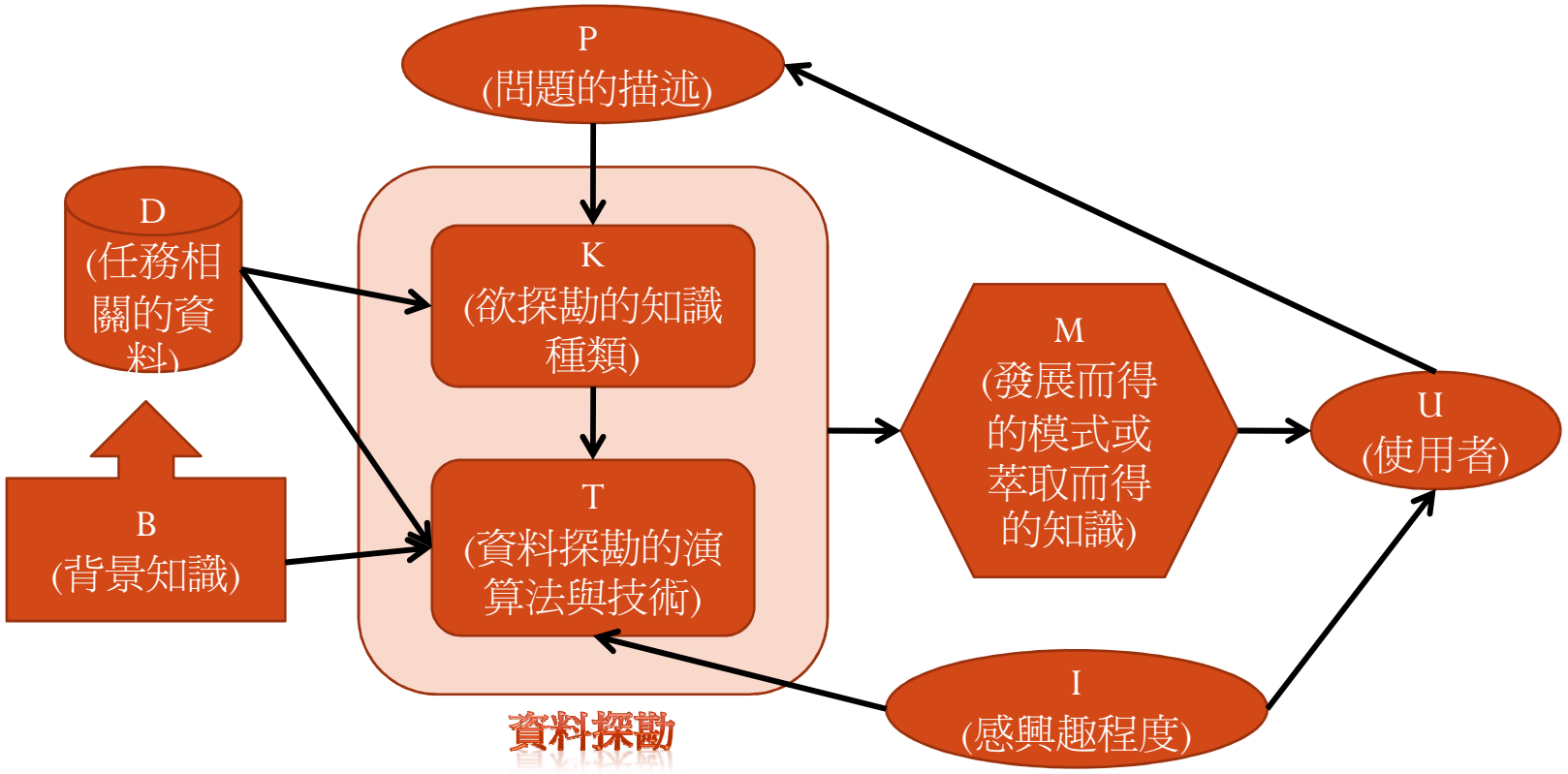
- 一種常被用來與資料探勘互換使用的術語
- 一種運用科學方法來做資料探勘的應用
- 典型的知識發現處理模型還包括
  - 資料擷取
  - 資料準備
  - 資料探勘後需採取的決策支援

資料探勘的起源

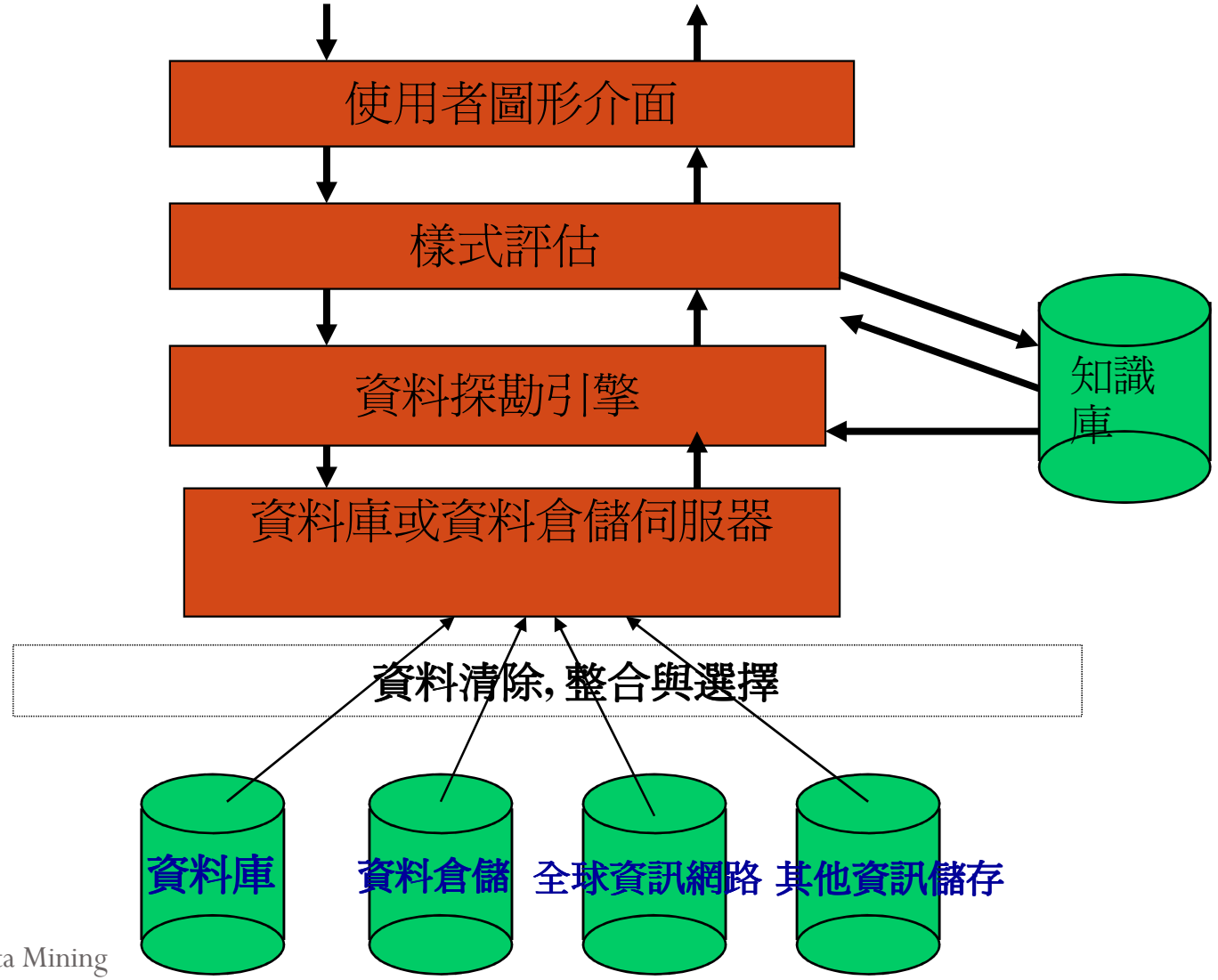
- 採用來自**機器學習**、**圖形辨識**、**統計學**和**資料庫系統**等領域的想法
- **傳統的技術**可能不適用於處理
  - **大量**的資料
  - **高維度**資料
  - **異質**和**分散性**的資料



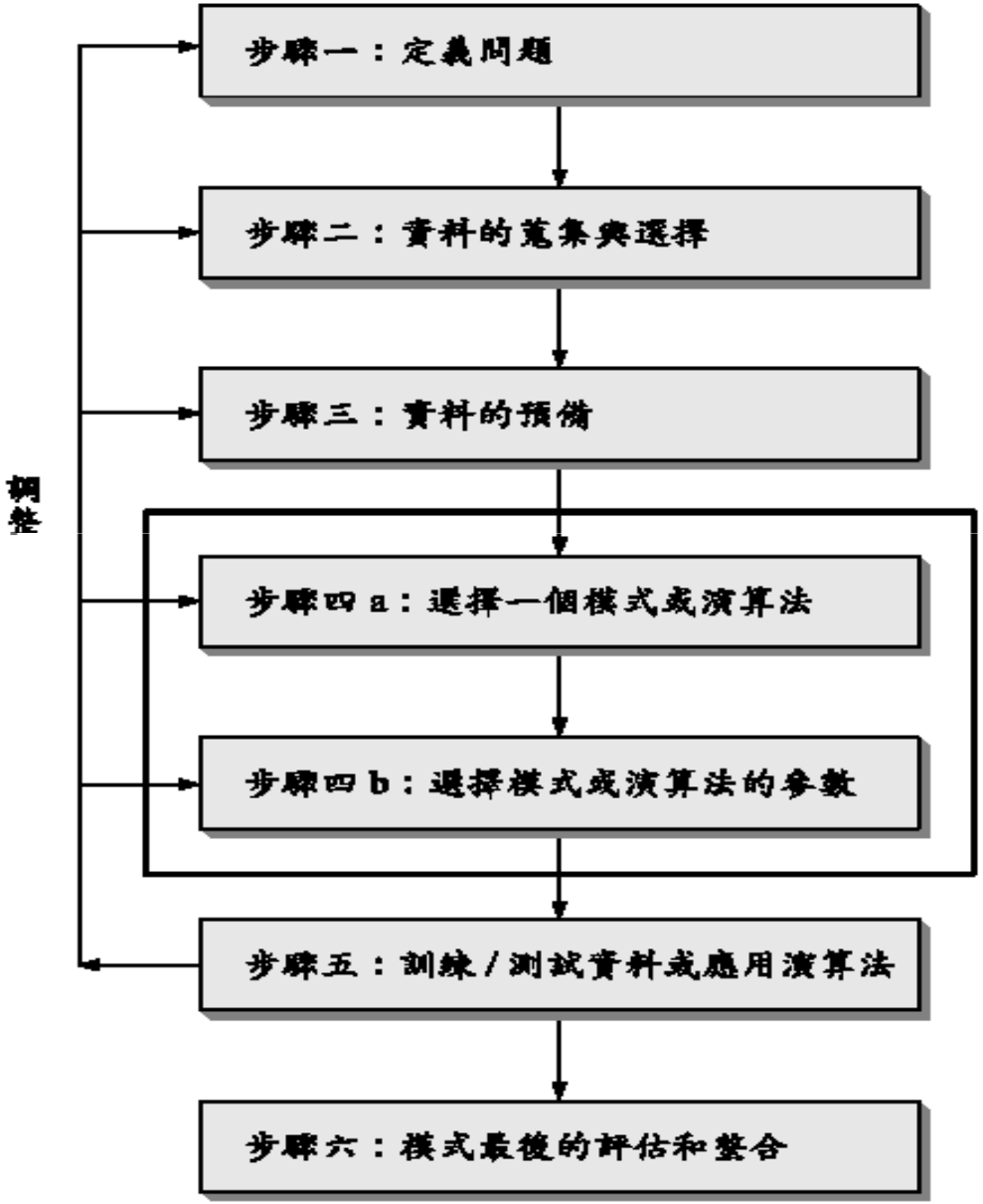
資料探勘的基本元件


P
(問題的描述)
D
(任務相關的資料)
K
(欲探勘的知識種類)
M
(發展而得的模式或萃取而得的知識)
U
(使用者)
B
(背景知識)
T
(資料探勘的演算法與技術)
I
(感興趣程度)
資料探勘

資料探勘的概念性架構


使用者圖形介面
樣式評估
資料探勘引擎
資料庫或資料倉儲伺服器
知識庫
資料清除, 整合與選擇
資料庫
資料倉儲
全球資訊網路
其他資訊儲存

資料探勘的流程


步驟一：定義問題
步驟二：資料的蒐集與選擇
步驟三：資料的預備
調整
步驟四 a：選擇一個模式或演算法
步驟四 b：選擇模式或演算法的參數
步驟五：訓練/測試資料或應用演算法
步驟六：模式最後的評估和整合

資料探勘的流程 (Cont.)

- 步驟一:**定義問題**
  - 定義資料探勘的標的
  - 定義目標
  - 問題的分析
- 步驟二:**資料的蒐集與選擇**
  - 有任何需要連結到本身或外部的資料庫嗎？如果有的話，該如何進行
  - 這些將被探勘的資料在經過探勘之後，是否會被改變？可否再次的被利用
  - 有什麼內部或外部的資訊有助於此次的分析
  - 這些資料與商業目標間有什麼關係
  - 資料庫中的資料表間需要什麼聯合
  - 在這些資料中是否具有可用的統計資訊

資料探勘的流程 (Cont.)

- 步驟三：**資料的預備**
  - 問題
    - 資料有哪些限制？哪個資料探勘的階段需要用到這些資料
    - 哪些資料的轉換在分析中是必要的
    - 這些資料的處理以及改變，是否可以被使用者接受
    - 這些資料是否有偏差？是否需要利用對數或平方轉換的方法來使資料能夠一致化
    - 需要對資料進行正規化嗎
    - 是否需要將資料轉換爲其他格式，例如：將是 / 否轉換爲1 / 0
  - 資料集合
    - 訓練資料集
    - 測試資料集
    - 評估資料集

資料探勘的流程 (Cont.)

- 步驟四:**選擇資料探勘的方法**
  - 選擇一個模式或演算法
    - 資料探勘想要的格式
    - 要用哪些技術
  - 選擇模式或演算法的參數
- 步驟五：**訓練 / 測試資料或應用演算法**
- 步驟六：**模式最後的評估和整合**
  - 此模式的錯誤率，是否可以接受？是否可以改進？
  - 是否有其他資料可以有助於改進模式的效率？
  - 輸出的結果是否需採用SQL的語法？
  - 是否可以整合獲得的知識到決策支援系統中，可以的話，該如何進行？

主要技術

- **預測模式**
  - 建立一個將**目標變數**視爲**解釋變數**的函數之模式
  - 預測模式有兩種：
    - **分類**模式：應用在目標變數爲**離散型**的資料上
    - **回歸**模式：應用在目標變數爲**連續型**的資料上
  - 範例：花型的預測
- **關聯規則分析**
  - 用來發現資料中**特徵屬性間**具有**高度關聯**的一種樣式
  - 範例：購物籃分析

主要技術 (Cont.)

- **分群分析**
  - 發現一群具有**相似特質**的**觀察值**，而這群具有相似特質的觀察值具有一些和**其他觀察值**不一樣的特性
  - 範例：文件分群
- **異常偵測**
  - 從一群資料中找出一些具有**顯著差異**的**觀察值**出來
  - 範例：信用卡詐騙的偵測

資料庫之知識發掘的相關技術

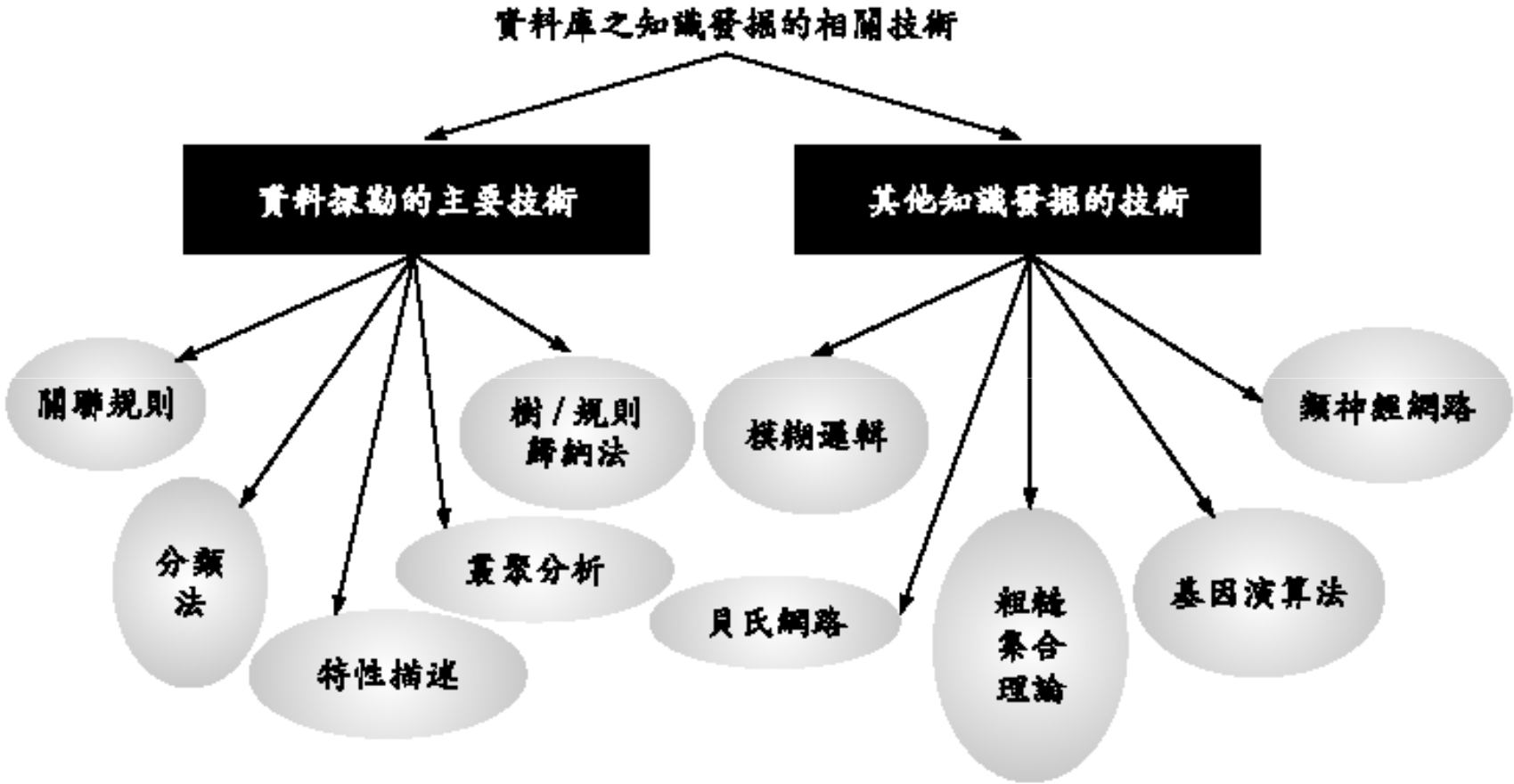


圖 1.14 資料庫之知識發掘的相關技術

資料探勘的資料來源

- 數位圖書館
- 影像檔案庫
- 醫學資料庫
- 財務與投資
- 生產與產品
- 商業與行銷
- 電信網路
- 科學領域
- 全球資訊網
- 生物鑑定

資料探勘的資料類型

- 關聯式資料庫

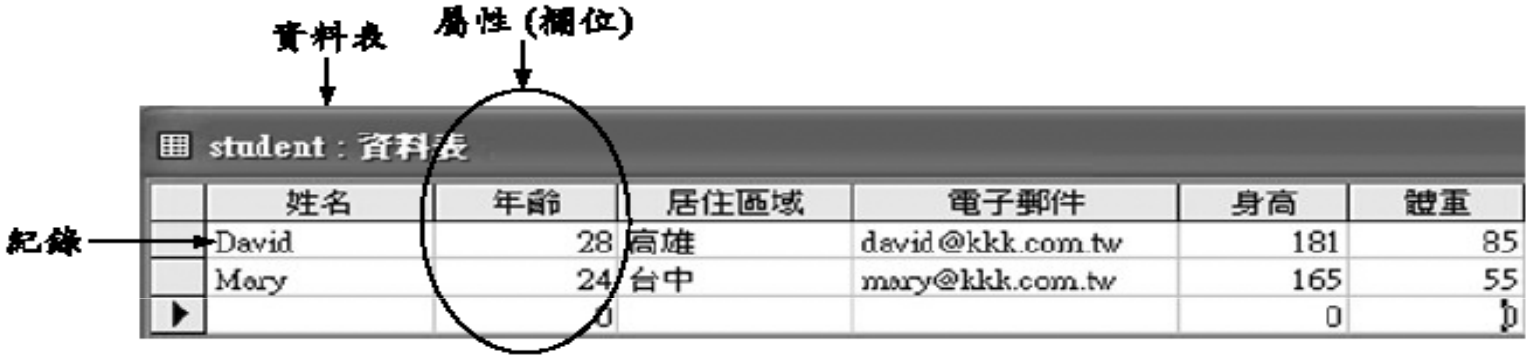
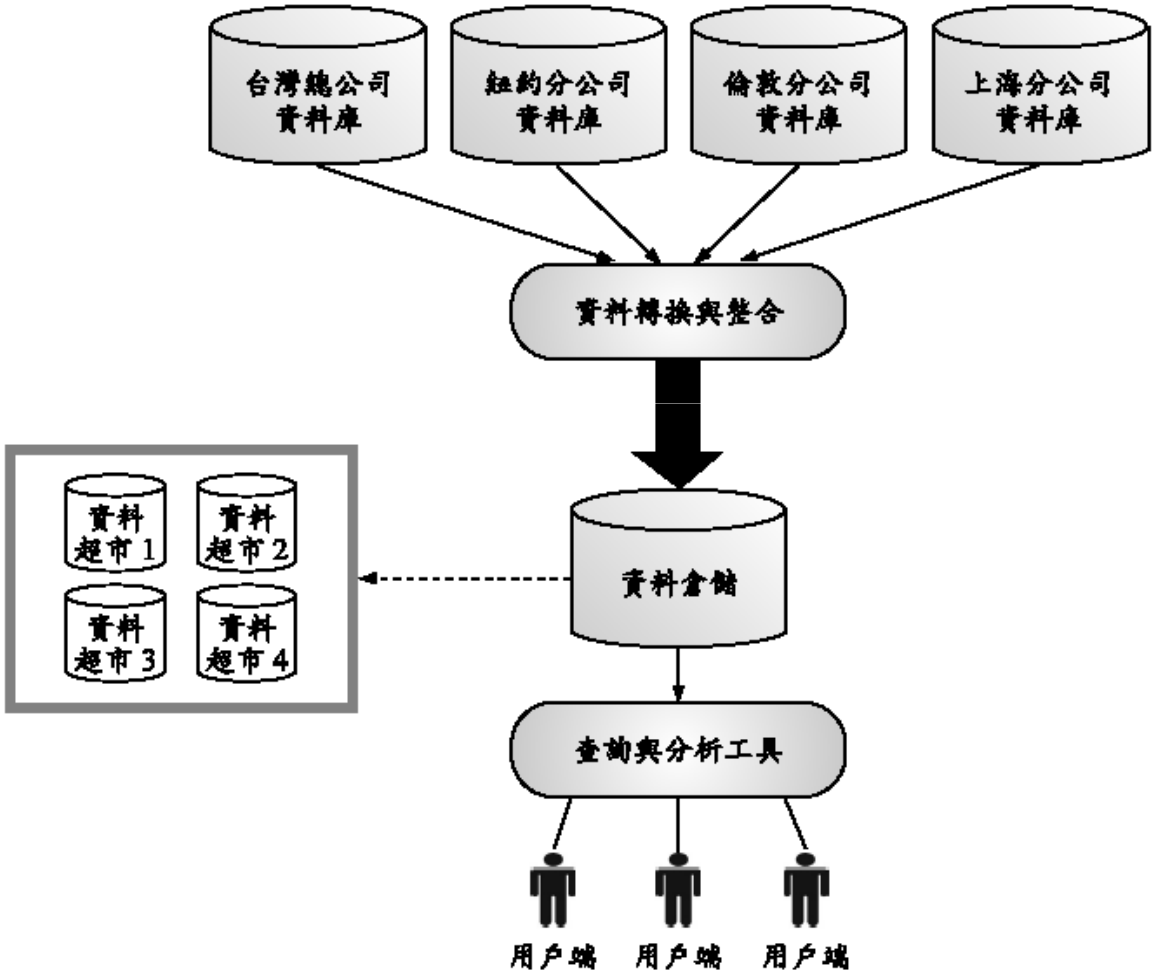


圖 1.9 關聯式資料庫基本概念示意圖

資料探勘的資料類型 (Cont.)

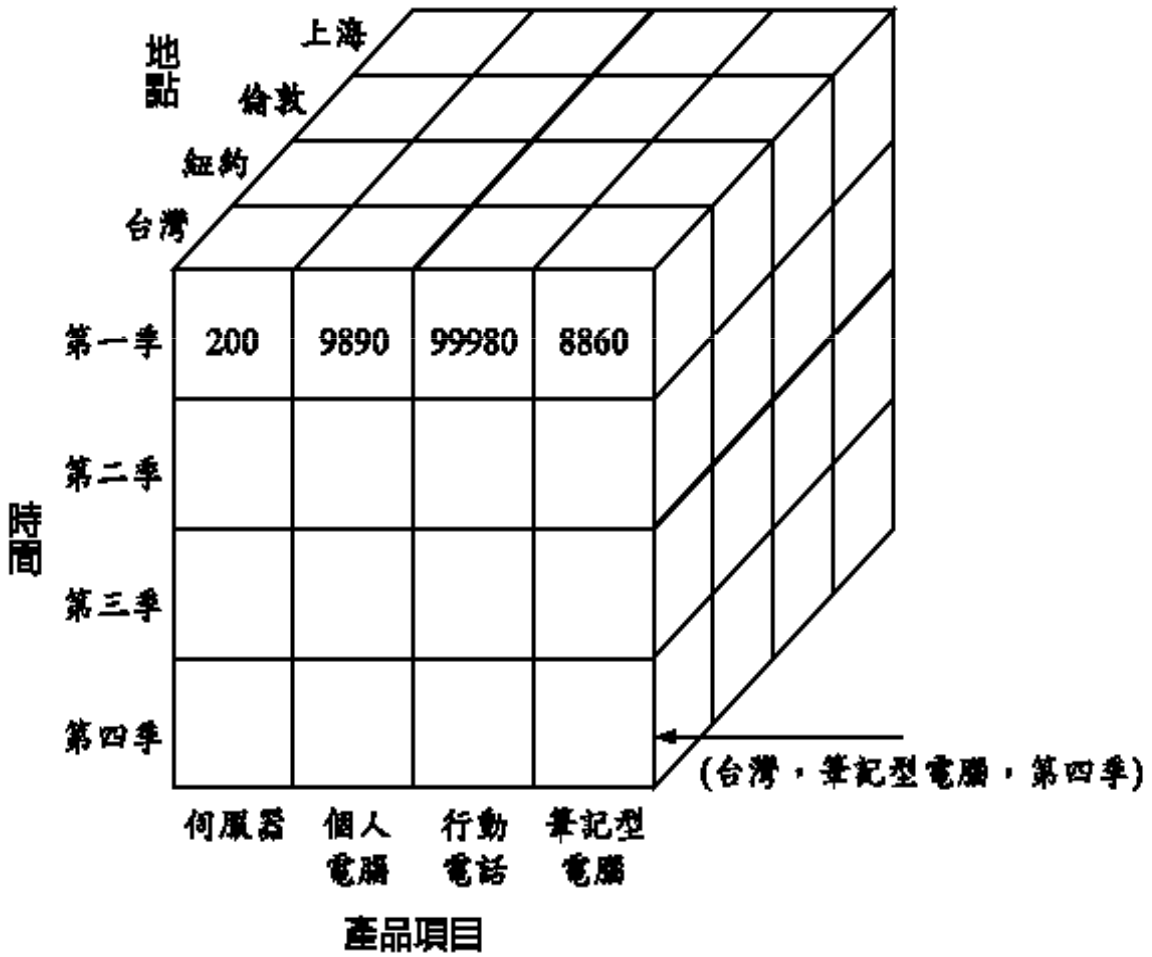
- 資料倉儲



1.10 資料倉儲處理過程及架構示意圖

資料探勘的資料類型 (Cont.)

- 資料倉儲
  - 資料方塊



1.11 多維度資料方塊

資料探勘的資料類型 (Cont.)

- 交易資料庫

1.12
交易資料庫範例

銷售：資料表

交易編號	交易時間	產品項目編號
001	2004/4/4	1,3,5
002	2004/4/4	10,12,13,14
003	2004/4/5	7,8,9,10

資料探勘的資料類型 (Cont.)

- 其他進階的資料庫系統和應用
  - 物件導向資料庫
  - 物件－關聯資料庫
  - 空間資料庫
  - 時間序列資料庫 (Time-Series Database)
  - 文字 (Text) 資料庫與多媒體 (Multimedia) 資料庫
  - 全球資訊網

資料探勘面臨的挑戰

- 擴展性
- 高維度的問題
- 異質性及複雜性的資料
- 資料品質
- 資料擁有者與分散性
- 非傳統式的分析

資料探勘的未來發展

- 應用領域的開發
- 資料探勘技術與其他技術的整合
- 有擴展性的(scalable)資料探勘方法
- 網際網路上的資料探勘(Web mining)
- 複雜資料型式(complex data types)的探擷
- 視覺化的資料探勘(visual data mining)

什麼是電腦可以學習的?

- 學習可分爲四種等級
  - **事實** (Fact): 事實就真相的簡單敘述
  - **概念** (Concept): 由具有特定特性的一群**物件**、**特徵**或**事件**所成的集合
  - **程序** (Procedure): 是爲達某一特定目的所進行的**連續步驟**
  - **原則** (Principle): 最高層次的學習。原則是以一些真相爲基礎，所形成的**通律**或**定律**
- **概念**是資料探勘的結果

概念觀

三個條件均成立

若　年收入 >= 30,000
且　目前職務的年資 >= 5
且　擁有自用住宅 = 是

則　優良信用風險 = 是

- 概念的結構:
  - 樹狀、規則、網路狀及數學方程式
- **標準概念觀** (Classical view)
  - 具有確定定義屬性的概念
- **可能性概念觀** (Probabilistic View)
  - 不要求概念表達但要有明確的屬性
- **範例概念觀** (Exemplar view)
  - 給定的例子與一或多個已知的概念範例夠相似

•持續按時繳納貸款金額的人，平均收入是 30,000
•大部份擁有優良信用風險的人，在同一家公司工作至少5年
•大多數擁有優良信用風險的人，擁有自己的房子

範例一:
年收入 = 32,000
在同一家公司服務年資 = 6
持有房屋

範例二:
年收入 = 52,000
在同一家公司服務年資 = 16
目前租屋

範例三:
年收入 = 28,000
在同一家公司服務年資 = 12
持有房屋