資料探勘:
概念與技術

─ 第一章─
─ 簡介 ─

第一至第七章內容

- 簡介
- 資料前處理
- 資料倉儲與即時線上分析技術: 簡介
- 進階資料方塊技術與資料一般化
- 探勘頻繁樣式, 關聯與相互關係
- 判別與預測
- 分群分析

第一章. 簡介

- 動機: 為什麼要資料探勘?
- 什麼是資料探勘?
- 資料探勘: 在何種資料?
- 資料探勘功能
- 所有樣式都是有趣嗎?
- 資料探勘系統分別
- 資料探勘工作基本項目
- 資料探勘系統與資料庫或資料倉儲系統的整合
- 資料探勘的主要議題

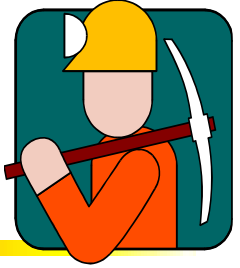
為什麼要資料探勘?

- 爆炸性成長的資料: 從 terabytes 到 petabytes
  - 資料收集與可用資料
    - 自動資料收集工具, 資料庫系統, 網際網路, 電腦化社群
  - 大量資料主要來源
    - 商業: 網際網路, 電子商務, 交易, 股票, ...
    - 科學: 遙控感應, 生物資訊, 科學模擬, ...
    - 社會與每個人: 新聞, 數位相機,
- 我們被資料所淹沒, 但是卻渴望知識!
- "需要是發明之母"—資料探勘—大量資料集自動分析

資料庫技術演進

- 1960s:
  - 資料收集, 產生資料庫, 資訊管理系統與網路化資料庫管理系統
- 1970s:
  - 關聯式資料模型, 關聯式資料庫管理系統實作
- 1980s:
  - 關聯式資料庫管理系統, 進階資料模型 (延伸關聯, 物件導向, 演繹等)
  - 應用導向資料庫管理系統(空間, 科學, 工程等)
- 1990s:
  - 資料探勘, 資料倉儲, 多媒體資料庫與網頁資料庫
- 2000s
  - 串流資料管理與探勘
  - 資料探勘與應用
  - 網際網路技術 (XML, 資料整合) 與全球資訊系統

什麼是資料探勘?



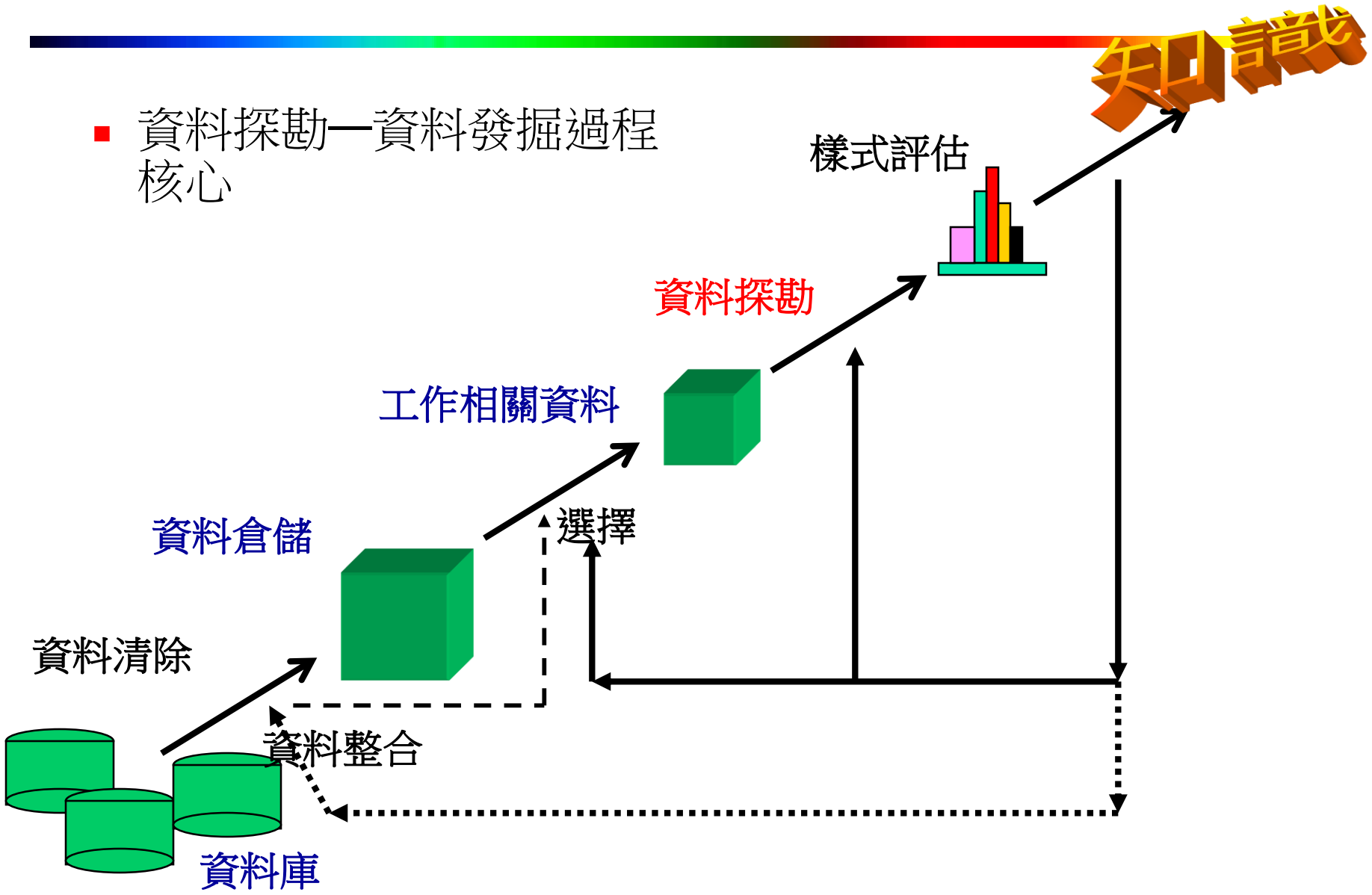
- 資料探勘 (從資料發掘知識)
  - 從龐大資料中擷取有趣 (不明顯, 隱含, 先前未知與有潛在用途) 樣式或知識
  - 資料探勘: 誤稱?
- 另外名稱
  - 在資料庫發掘(探勘)知識 (KDD), 知識擷取, 資料/樣式分析, 資料考古學, 資料疏濬, 資訊收穫, 商業智慧等.
- 注意: 所有都是 "資料探勘"?
  - 簡單搜尋與查詢處理
  - (演繹) 專家系統

為什麼資料探勘?—潛在應用

- 資料分析與決策支援
  - 市場分析與管理
    - 目標
    - 客戶關係管理 (CRM), 購物籃分析, 交互銷售, 市場區隔
  - 風險分析與管理
    - 預測, 客戶保持, 品管, 競爭分析
  - 詐騙發掘與發掘異常樣式 (離異值)
- 其他應用
  - 文字探勘 (新聞群組, 電子郵件, 文件) 與網際網路探勘
  - 串流資料探勘
  - 生物資訊與生物資料分析

知識發掘過程

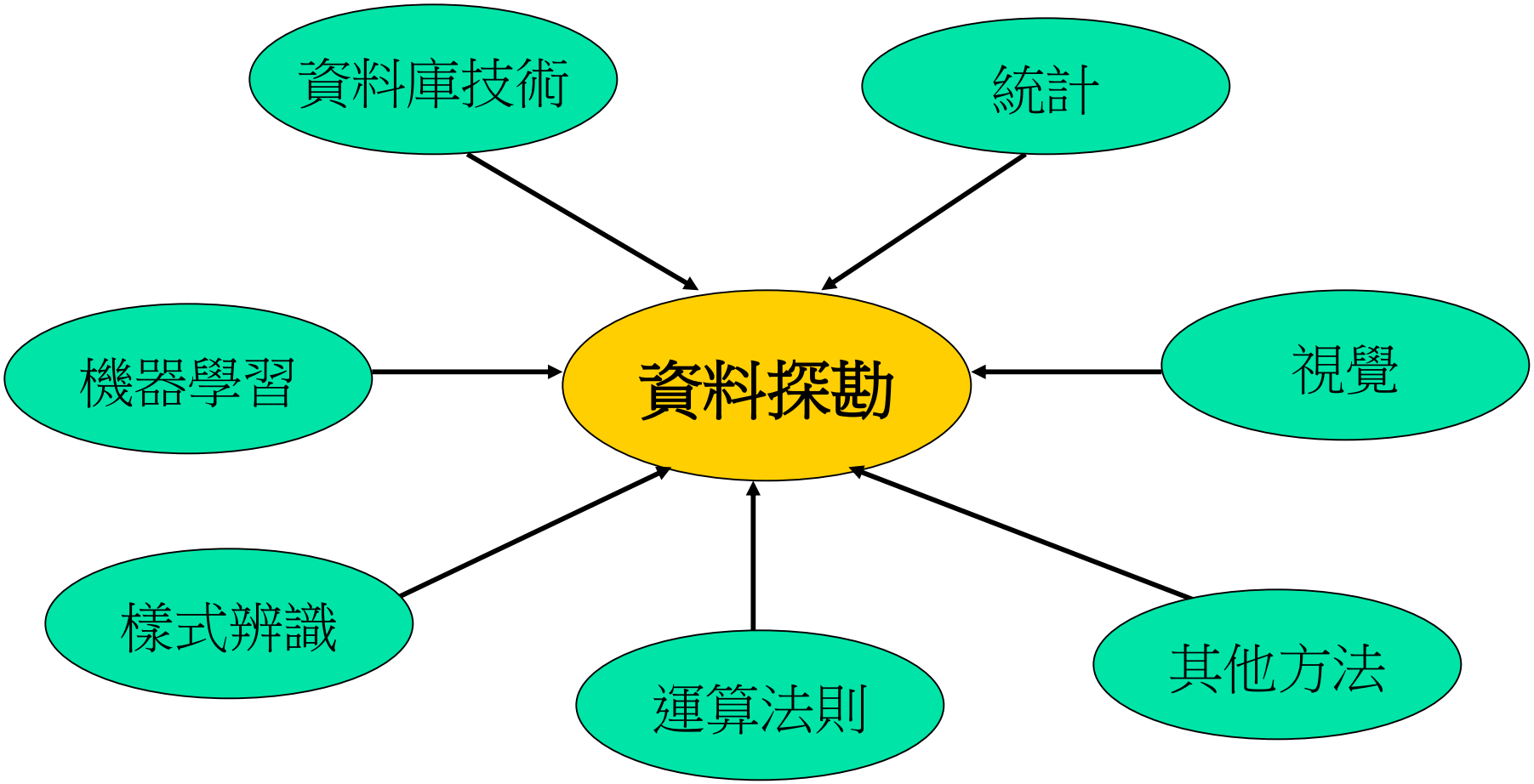
- 資料探勘—資料發掘過程核心



知識發掘過程: 關鍵步驟

- 學習應用範疇
  - 應用相關先前知識與目標
- 建立目標資料集: 資料選擇
- 資料清除 與前處理: (佔 60% 工作)
- 資料刪減與轉換
  - 尋找有用特性, 維度/變數刪減, 不變的表示
- 選擇資料探勘功能
  - 匯總, 判別, 迴歸, 關聯, 分群
- 選則探勘運算法則
- 資料探勘: 搜尋有趣樣式
- 樣式評估與知識呈現
  - 視覺化, 轉換, 移除多餘樣式等
- 使用探勘知識

資料探勘: 許多方法的匯合



多維度檢視的資料探勘

- 探勘資料
  - 關聯, 資料倉儲, 交易式, 串流, 物件導向/關聯, 主動式, 空間, 時間序列, 文字, 多媒體, 不同質, 遺贈, 全球資訊網路
- 探勘知識
  - 特徵化, 區別, 關聯, 判別, 分群, 趨勢/偏差, 離異值分析等
  - 多個/整合函數與多層次探勘
- 使用技巧
  - 資料庫導向, 資料倉儲 (OLAP), 機器學習, 統計, 視覺化等
- 套用應用
  - 零售, 電訊業, 銀行, 詐欺分析, 生物資料探勘, 股票市場分析, 文字探勘, 網際網路探勘等

資料探勘: 判別綱目

- 一般功能
  - 敘述性資料探勘
  - 預測性資料探勘
- 不同視野導致不同判別
  - 資料 視野: 探勘資料種類
  - 知識 視野: 探勘知識種類
  - 方法 視野: 使用方法種類
  - 應用 視野: 套用應用種類

資料探勘: 在何種資料?

- 資料庫導向資料集與應用
  - 關聯式資料庫, 資料倉儲, 交易式資料庫
- 進階資料集與進階應用
  - 資料串流與感應資料
  - 時間序列資料, 空間資料, 順序資料 (包含生物順序)
  - 結構資料, 圖形, 社會網路與多連結資料
  - 物件導性資料庫
  - 不同質與遺贈資料庫
  - 空間資料與空間時間資料
  - 多媒體資料庫
  - 文字資料庫
  - 全球資訊網路

資料探勘功能

- 多維度概念描述: 特徵與區別
  - 一般化, 匯總, 對照資料特性例如乾燥與潮溼區域
- 頻繁樣式, 關聯, 相互關係 與 意外
  - 尿布 → 啤酒 [0.5%, 75%] (相互關係 或 意外?)
- 判別與預測
  - 建立模型 (函數) 對未來預測描述並區別不同類別或概念
    - 例如, 根據(氣候)判別國家, 或根據(汽油公里數)來判別汽車
  - 預測一些未知或遺失數值

資料探勘功能 (2)

- 分群分析
  - 未知類別標籤: 群組資料產生新類別, 例如對房屋進行分群找出分布樣式
  - 不同類相似最大化 與 & 同類相似最小化
- 離異值分析
  - 離異值: 是指不同於一般資料行為的資料物件
  - 雜訊或例外? 用於詐騙檢測與極少事件分析
- 趨勢與演進分析
  - 趨勢與偏差: 例如迴歸分析
  - 順序樣式探勘: 例如數位相機 → 大 SD 記憶體
  - 週期性分析
  - 根據相似度分析
- 其他樣式-直接或統計分析

所有發掘樣式都是值得注意的嗎?

- 資料探勘會產生數以千計樣式: 並非所有的樣式都是值得注意的
  - 建議方法: 以人為中心, 以查詢為依據, 專注探勘
- 值得注意指標
  - 值得注意樣式為:人容易懂它, 在新或測試資料有一定層度的正確, 有潛在用處, 重要, 或驗證使用者要確認的某些假設
- 客觀與主觀的指標
  - 客觀: 根據統計與樣式結構, 例如 支持度, 信賴度等.
  - 主觀: 根據使用者對資料信心, 例如 意外, 重要, 可執行等

尋找所有或有趣樣式?

- 尋找所有有趣樣式: 完全
  - 資料探勘系統是否可以尋找所有有趣樣式? 是否需要尋找所有有趣樣式?
  - 啟發式(Heuristic) 或 徹底尋找
  - 關聯 或 判別 或 分群
- 僅搜尋有趣樣式: 一個最佳化問題
  - 資料探勘系統是否可以進搜尋有趣樣式?
  - 方法
    - 產生所有樣式並排除無趣樣式
    - 僅產生有趣樣式— 探勘查詢最佳化

為何使用資料探勘查詢語言?

- 自動或查詢導向?
  - 自動尋找資料庫所有樣式?—不實際。因為樣式會太多而且是無趣的。
- 資料探勘應為互動程序
  - 使用者主導探勘項目
- 使用者必須提供一組基本元素用於與資料探勘系統進行溝通
- 將這些基本元素包含於一個資料探勘查詢語言
  - 使用者互動更有彈性
  - 作為圖形化使用者介面設計的基礎
  - 資料探勘產業與實務的標準

資料探勘基本元素

- 探勘工作相關的資料
- 背景知識
- 樣式評估有趣度量
- 發覺樣式視覺化與呈現

基本元素 1: 工作相關資料

- 資料庫或資料倉儲名稱
- 資料庫資料表或資料倉儲資料方塊
- 資料選擇條件
- 相關屬性或維度
- 資料群組條件

基本元素 2: 探勘知識類型

- 特徵化
- 區別化
- 關聯
- 判別/預測
- 群組
- 離異值分析
- 其他資料探勘工作

基本元素 3: 背景知識

- 一項傳統背景知識: 概念階層
- 綱目階層
  - 例如. 街 < 城市 < 省或州 < 國家
- 集合群組階層
  - 例如. {20-39} = 青年, {40-59} = 中年
- 操作導向階層
  - 電子郵件地址: hagonzal@cs.uiuc.edu

    登入名稱 < 部門 < 大學 < 國家
- 規則階層
  - 低獲利 (X) <= 價格(X, P_1) 並且 成本 (X, P_2) and (P_1 - P_2) < $50

基本元素 4: 評估有趣樣式的度量

- 簡單性

  例如. (關聯) 規則長度, (決策) 數大小

- 確定性

  例如. 信賴度, P(A|B) = #(A and B)/ #(B), 判別信賴部或正確率, 確定因素, 規則強度, 規則品質, 區別權重等

- 使用性

  潛在使用, 例如. 支持度 (關聯), 雜訊界限值 (描述)

- 重要性

  先前未知, 驚訝 (用於移除多於規則, 例如., 伊利諾 相對於 香檳 規則隱含支持度比例)

基本元素 5: 發掘樣式呈現

- 不同背景/使用需要不同型態表示
  - 例如. 規則, 資料表, 交互表, 圓形/長條圖等
- 概念階層也是重要的
  - 當發掘知識以高抽象階層表現是比較容易懂得
  - 互動的向上/向下鑽探, 樞紐分析, 切片與切塊 提供不同資料面像
- 不同類型知識需要不同表示: 關聯, 判別, 分群等

DMQL—資料探勘查詢語言

- 動機
  - 允許使用者與資料探勘系統能有彈性的互動
  - 藉提供像SQL的標準化語言
    - 希望能達到SQL在關聯資料庫的相似效果
    - 系統發展與演進的基礎
    - 加速資訊交換, 技術轉移, 商業化與答, 大量接受度
- 設計
  - DMQL用先前敘述基本素進行設計

DMQL查詢範例

範例 1.11 探勘判別規則

假設你身為全電公司經理，你想要根據顧客購買樣式對顧客進行判別。你感到有興趣的顧客條件是薪資不低於$40,000，購買總金額超過$1,000，並且所購買項目單價要不低於$100。你特別對顧客年齡、收入、購買項目類別、購買地點、項目製造地有興趣，並且要把結果顯示成規則的形式，這個資料探勘查詢以DMQL[3]顯示如下：

(1) use database 全電公司資料庫
(2) use hierarchy 位置階層 for T.分店, 年齡階層 for C.年齡
(3) mine classification as 有希望客户
(4) in relevance to C.年齡, C.收入, I.類型, I.製造地, T.分店
(5) from 客户 C, 項目 I, 交易 T
(6) where I.項目編號 = T.項目編號 and C.客户編號 = T.客户編號 and
C.收入 ≥ 40,000 and I.價格 ≥ 100
(7) group by T.客户編號
(8) having sum (I.價格) ≥ 1000
(9) display as 規則

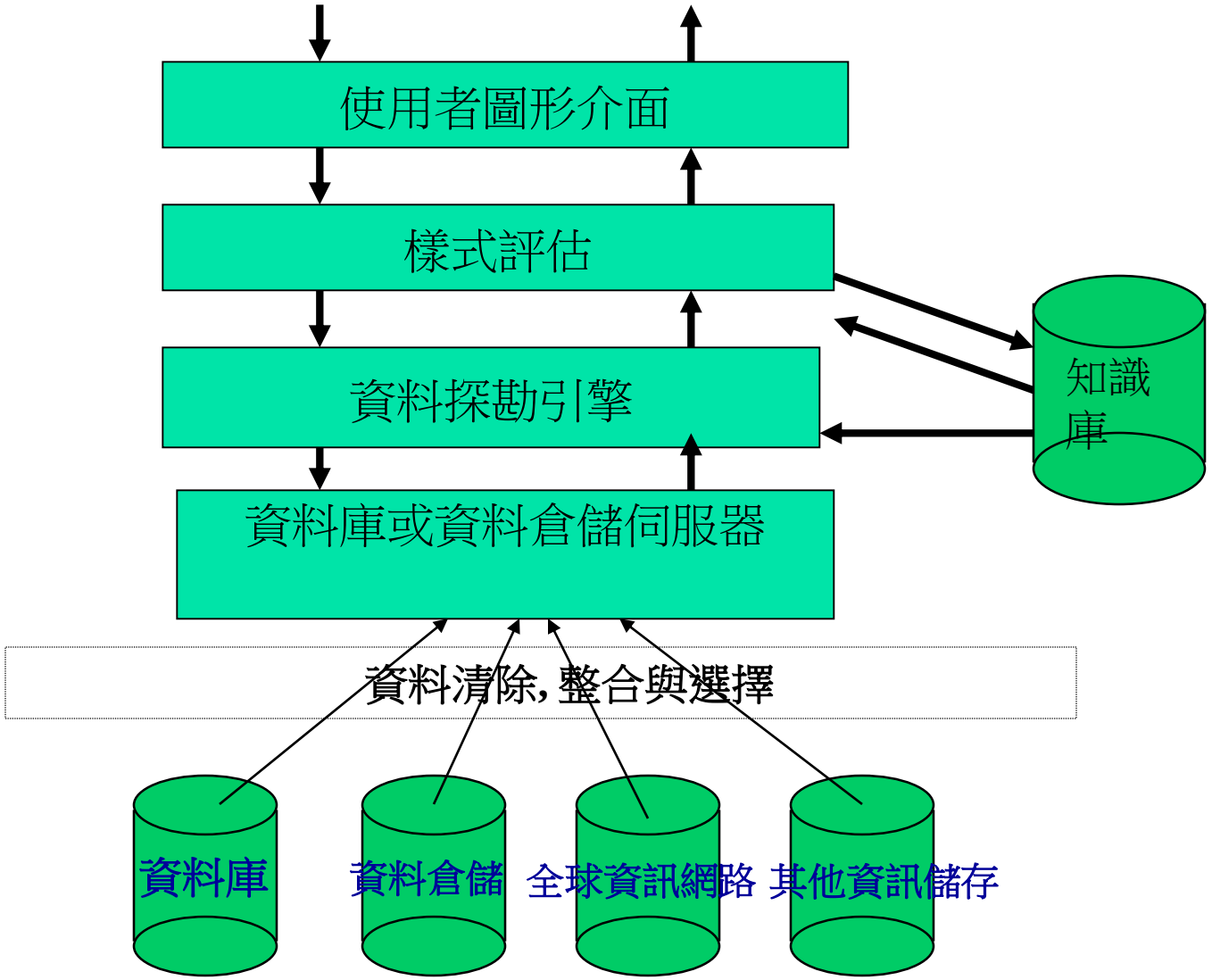
資料探勘與資料倉儲整合

- **資料探勘系統, 資料庫管理系統, 資料倉儲結合**
  - 沒有結合, 鬆散結合, 半緊密結合, 緊密結合
- **即時分析探勘資料**
  - 整合探勘與即時分析處理技術
- **互動式探勘多層次知識**
  - 透過鑽探/捲動, 樞紐分析, 切片/切塊來探勘不同抽象階層的知識與樣式
- **整合許多探勘功能**
  - 特徵化判別, 先分群再關連

資料探勘與資料庫/資料倉儲結合

- 沒有結合—扁平檔案處理, 不建議
- 鬆散結合
  - 從資料庫/資料倉儲擷取資料
- 半緊密結合— 加強資料探勘效能
  - 某些有效率的資料探勘基本元素是由資料庫系統或資料倉儲系統提供，這些基本元素包含排序、索引、聚合、直方圖分析、多向連結、某些重要統計度量的事先計算
- 緊密結合— 一個一致資訊處理環境
  - 資料探勘系統與資料庫系統或資料倉儲系統進行完美整合，資料探勘子系統被視為一個資訊系統功能成分，資料探勘查詢與功能根據探勘查詢分析、資料結構、索引方案與資料庫系統或資料倉儲系統的查詢方法進行最佳化。

架構: 傳統資料探勘系統


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

資料探勘主要議題

- 探勘方法
  - 在資料庫探勘不同類型知識
  - 資料探勘方法的效率與可量度性
  - 樣式評估：有趣問題
  - 包含背景知識
  - 處理雜訊與不完全資料
  - 平行化、分散式與遞增式探勘方法
- 使用者互動
  - 資料探勘查詢語言與特別資料探勘
  - 資料探勘結果呈現與顯示
  - 在不同抽象層次進行互動知識探勘

總結

- 資料探勘: 從大量資料中發掘有趣樣式
- 由於龐大需求與應用 ， 資料庫技術自然演進
- 從資料進行知識發掘包含資料清除、資料整合、資料選擇與轉換、資料探勘、樣式評估與知識呈現
- 可在不同資訊儲存模式執行探勘
- 資料探勘功能: 特徵化, 區別化, 關聯, 判別, 分群, 離異與趨勢分析等
- 資料探勘系統與架構
- 資料探勘主要議題