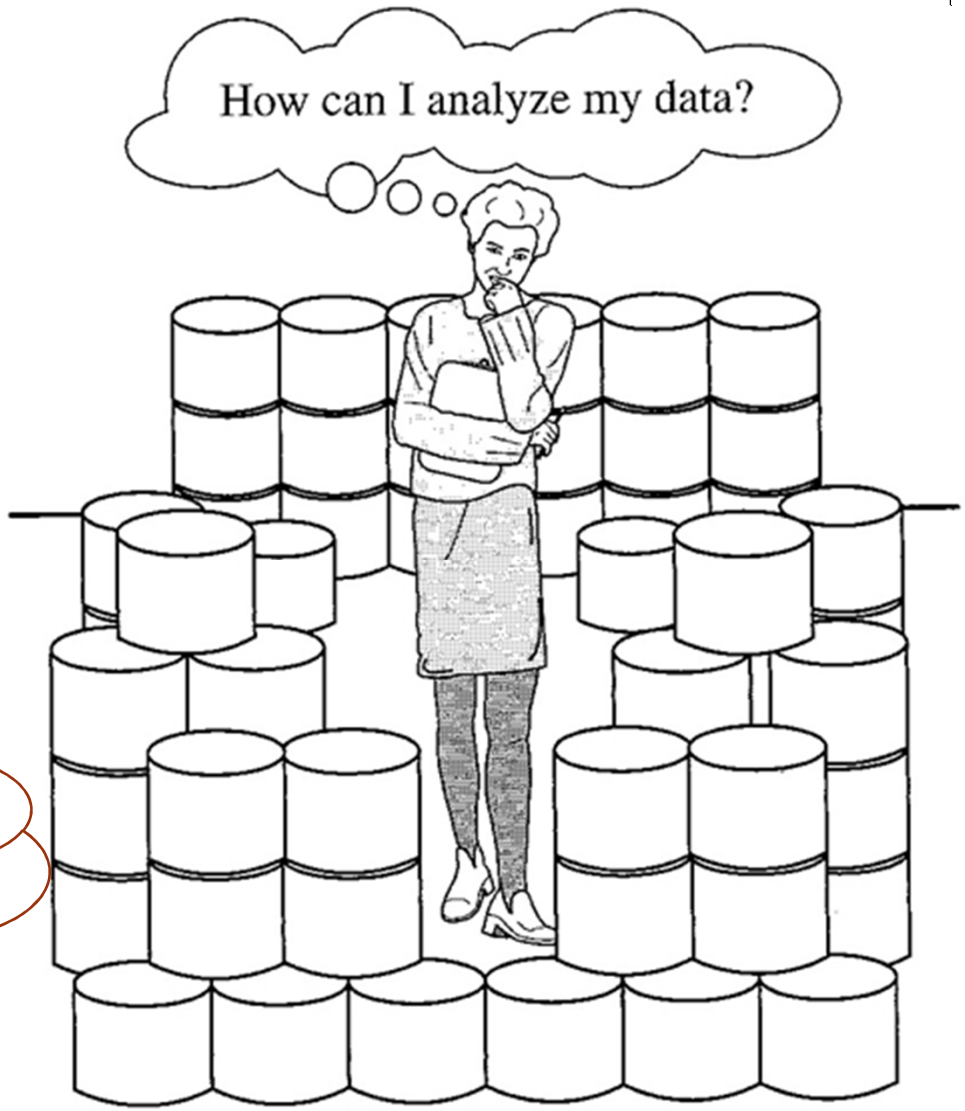
動機

- 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、客製化的服務作為競爭優勢（如顧客關係管理）

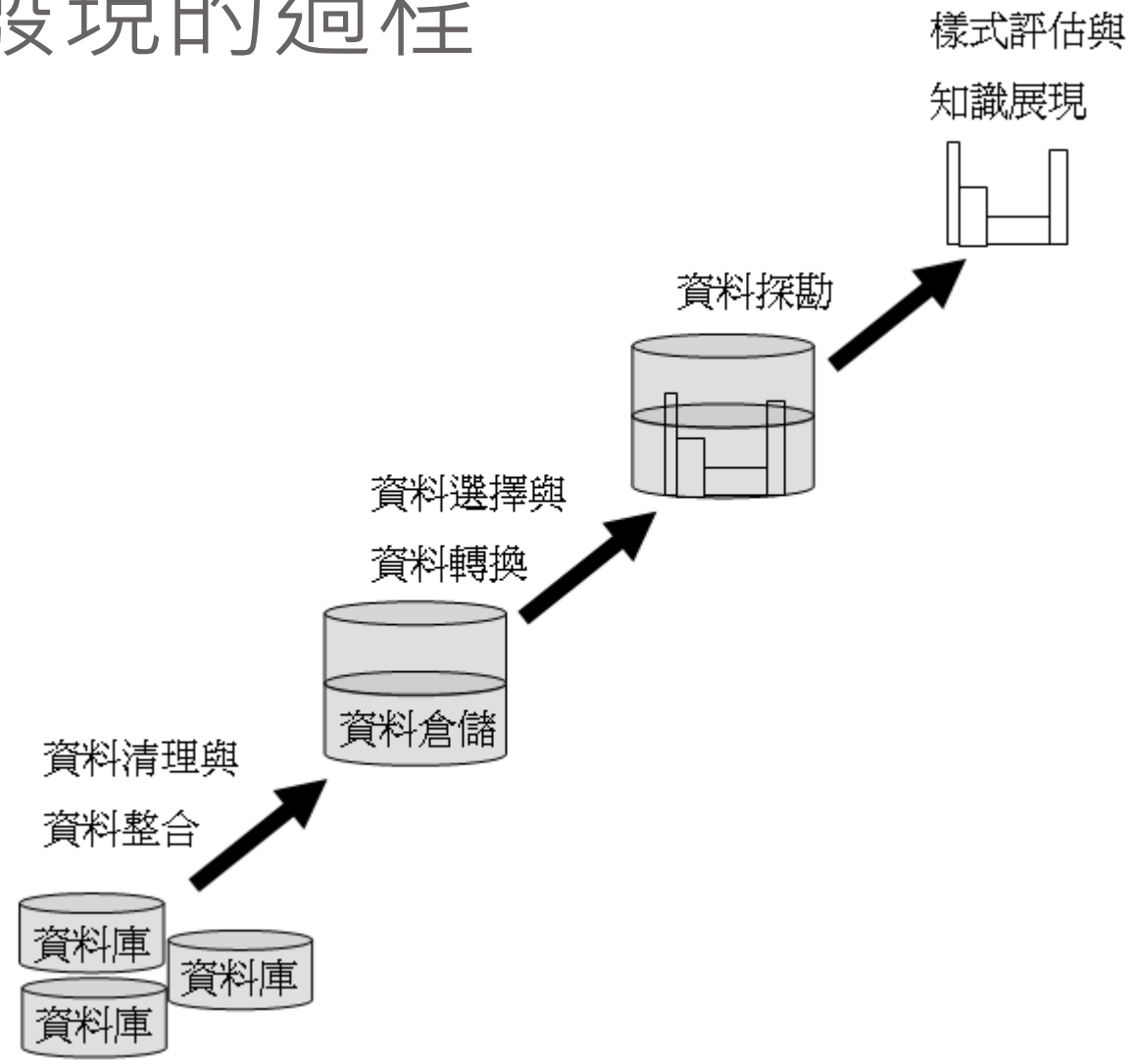
資料探勘（Data Mining）

- **資料探勘**（Data Mining）係經由自動或半自動的方式，探勘及分析大量資料，以建立有效的**樣式**（Pattern）、**規則**（Rule）或**模型**（Model）。資料探勘並不是無中生有，也不是一種魔術。
- 過去多年以來，企業大多是利用統計學，以人工的方式對資料庫作探勘，尋找可能有意義的統計樣式。其中有些是以「**由上而下**」的方式進行，稱之為「**假設檢定**」。另一種「**由下而上**」的方式，稱之為「**知識發現**」（Knowledge Discovery in Database，KDD）。
- 知識發現並不需要事先的假設，只需直接讓資料說話。此外，知識發現可分為兩大類—監督式及非監督式。

知識發現是發現知識的一連串過程

- 步驟1：**資料清理**（Data cleaning）：去除不一致或錯誤的資料。
- 步驟2：**資料整合**（Data Integration）：將各個來源的資料予以整合。
- 步驟3：**資料選擇**（Data Selection）：從資料庫中擷取相關資料並做分析。
- 步驟4：**資料轉換**（Data Transformation）：將資料轉換至適當的格式，以進行彙總。
- 步驟5：**資料探勘**（Data Mining）：應用某演算法將彙總的資料進行分析處理，以獲取資料的樣式。
- 步驟6：**樣式評估**（Patterns Evaluation）：定義有興趣了解的知識樣式。
- 步驟7：**知識展現**（Knowledge Presentation）：以視覺化方式展現探勘後的知識。

知識發現的過程



資料探勘為什麼是現在興起？

1. 大量資料的產生
2. 資料倉儲技術的興起
3. 電腦計算能力的提升
4. 競爭壓力強大
5. 所有的企業都將是服務業

資料探勘的兩種基本模式

1. **假設檢定**（Hypothesis Testing）：是一種由上而下的方式，檢視我們的想法是否成立
2. **知識發現**（Knowledge Discovery）：是一種由下而上的方式，從分析原始資料開始，找出我們所不知道的事實。

資料探勘的五大技術分類

1. 關聯法則（Association Rule）
2. 分類（Classification）
3. 預測（Prediction）
4. 叢集（Cluster）
5. 複雜型態探勘

推論出關聯法則

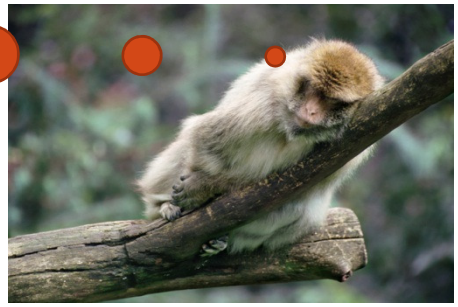
雜貨店中賣的四項產品:

1. 牛奶
2. 起司
3. 麵包
4. 蛋

可能的購買組合

1. 假如顧客買牛奶, 他們也會買麵包
2. 假如顧客買麵包, 他們也會買牛奶
3. 顧客若買牛奶與蛋, 也會買起司與麵包
4. 顧客若買牛奶, 起司和蛋, 他們也會買麵包

購買牛奶的事件會導致購買麵包的可能性有多大?



推論出關聯法則(Cont.)

假設總共有 10,000筆顧客交易涉及牛奶的購買, 其中同時買牛奶又買麵包的交易有 5,000筆, 則**信賴度**為 5000 / 10000 = 50%

假設總共有 20,000筆顧客交易涉及麵包的購買, 其中同時買麵包又買牛奶的交易有 5,000筆, 則**信賴度**為 5000 / 20000 = 25%

信賴度和支持度

一個規則的信賴度不會提供另外一些重要的資訊，如一個關聯法則中所找到的屬性值，它們在所有交易中所佔的比率。所以對一條規則而言，此統計學的方法就是所謂的**支持度**（support）。簡單地說支持度就是在特定的關聯法則中包含所有項目目錄的資料庫中之最小的交易百分率。

推論出關聯法則(Cont.)

- **支持度**的定義為決策變數在資料庫中所出現的比例，表現形式為Sup (X)，也就是在整個資料庫L中出現的比例，支持度越高，越值得重視。支持度代表事件的發生機率。Sup(X→Y)代表同時發生X和Y兩個交易事項的機率，支持度介於0%和100%之間。

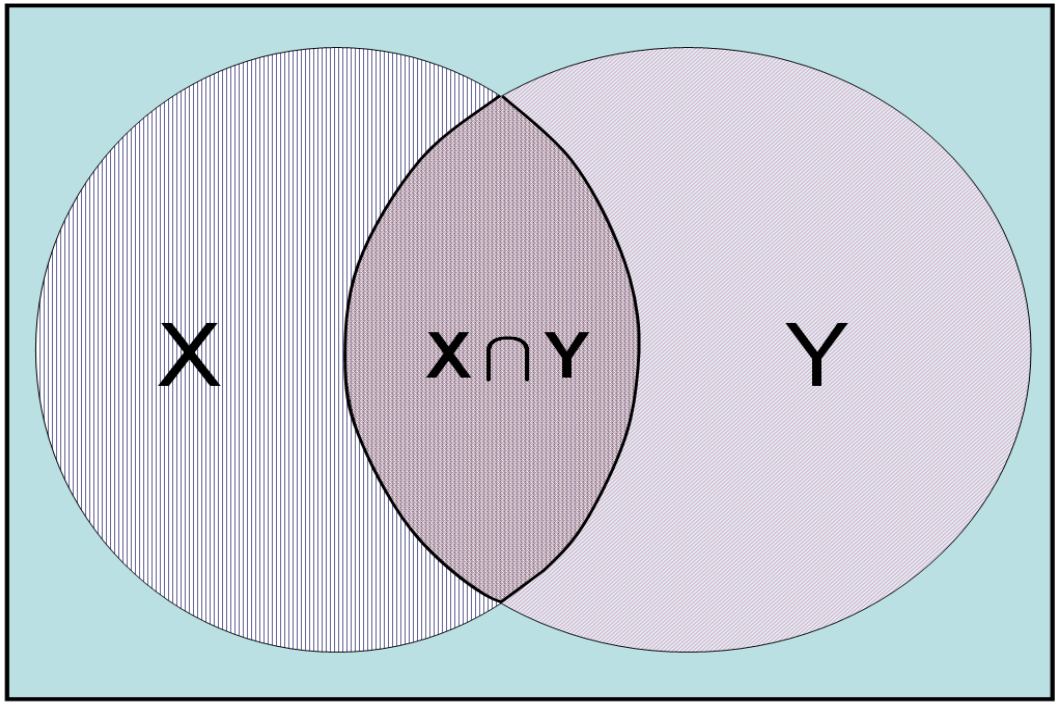
$$Sup(X) = \frac{\text{項目集合X在資料庫中出現的總次 數}}{\text{資料庫中的總交易筆數}}$$

推論出關聯法則(Cont.)

- 可靠度的定義此關聯性法則可信的程度，也就是某決策變數X已確知或成立時，另一決策變數Y發生或成立的機率，與統計中的條件機率相同，表現形式為Conf (X→Y)。Conf(X→Y)代表發生X的交易事項下，發生Y交易事項的機率，可靠度介於0%和100%之間。

$$Conf(X \rightarrow Y) = \frac{Sup(X \cap Y)}{Sup(X)}$$

推論出關聯法則(Cont.)



- 一般而言，關聯性法則的**支持度**及**可靠度**皆必須分別大於使用者訂定的最低限制，才能據此判定其為有意義的關聯性法則 。