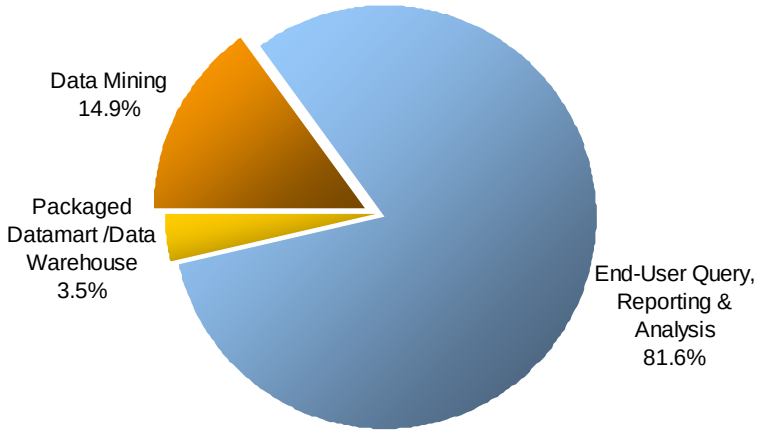
Ch 1資料採礦概念與應用

Data Mining的市場

- 資料採礦就是利用統計以及機械學習的演算法，啓發性地從大量資料中找尋隱藏具有商業價值的知識與規律，以作爲自動化商業策略之應用。
- MIT於2001/1「科技評論」中表示資料採礦將是未來改變世界的十大新興科技趨勢之一
- IDC研究總監Dan Vesset指出全球資料採礦市場市值已從2001年的4.55億美金成長至2002年的5.39億，而且預估在2006年將會成長至18.5億美金
- 2002/6 Smart Money預測資料採礦分析師將是未來十年中最熱門行業的第五名

Taiwan Business Intelligence Tools Revenue by Functional Market, 2003 (US$M)

Total Revenue = US$11.7M



Taiwan Business Intelligence Tools Revenue by Functional Market, 2003 H1 & H2 (US$M)

Functional	2003 H1	2003 H2	Growth Rate
Data Mining	0.8	1.0	22.7%
End-User Query, Reporting & Analysis	4.6	4.9	5.8%
Packaged Datamart /Data Warehouse	0.2	0.2	5.6%
Grand Total	5.6	6.1	8.2%

Source: IDC, Q1 2004

最終無法被簡化為數值問題者，終將無法被探究

法國實證學祖師 孔德

(Auguste Comte, 1798-1857)”

asiaMiner

Curve Up the Future

資料採礦名詞的來源

- 1960年代： Data Fishing
- 八十年代末期： Database Mining(Rakesh Agrawal)
- Data Mining一字是由Usama Fayyad於1991時，首次於他的博士論文中發表

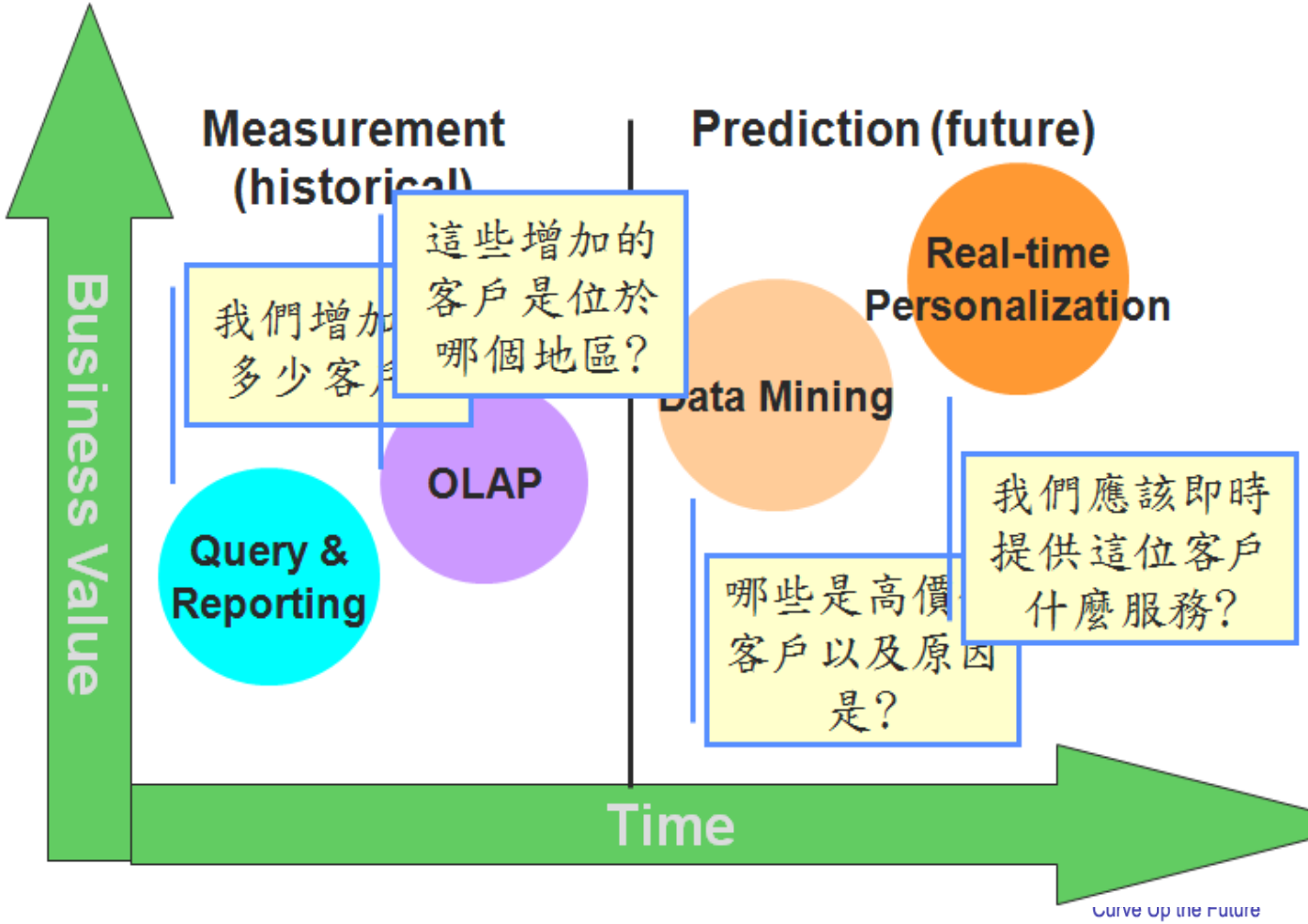
asiaMiner
Curve Up the Future

資料採礦一詞的定義

- **Extracting knowledge from large amount of data—Jiawei Han**
- **…..is the exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules—J. A. Michael Berry**
- 資料採礦是利用統計以及機械學習的演算法，啓發性地從大量資料中找尋隱藏具有商業價值的知識與規律，以作爲自動化商業策略之應用。

asiaMiner
Curve Up the Future

商業智慧分析工具


Business Value
Measurement (historical)
Prediction (future)
這些增加的客戶是位於哪個地區?
Real-time Personalization
Data Mining
OLAP
Query & Reporting
我們應該即時提供這位客戶什麼服務?
哪些是高價客戶以及原因是?
Time

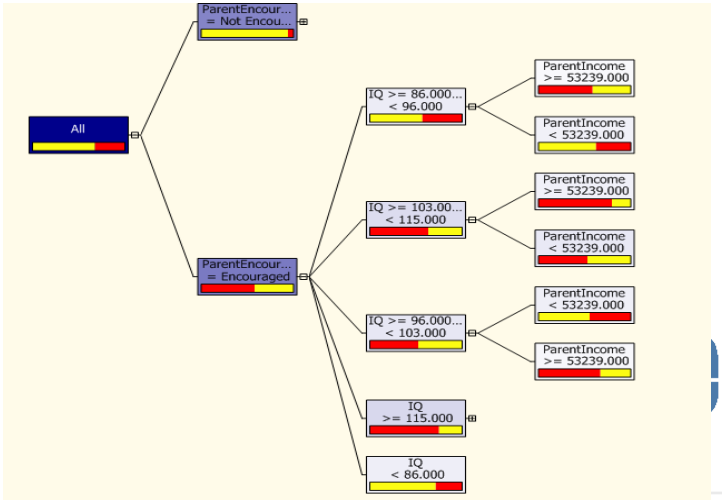
資料採礦的功能

- 分類(Classification)
- 推估(Estimation)
- 群集化(Cluster)
- 同質分組(Affinity Group)
- 序列(Sequential)
- 描述(Description)

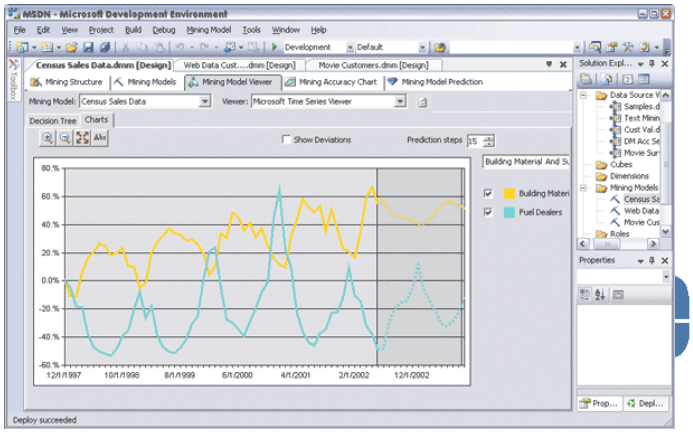
asiaMiner
Curve Up the Future

分類

- 預測類別變數(Discrete Variables)的過程我們稱之爲「分類」
- 在分類問題中，除了提供預測的分類結果之外，同時也會提供發生這個分類結果的可能機率
- 分類問題的應用包括：
  - →信用風險預測(是否會呆帳違約?)
  - →交叉銷售(客戶是否會購買?)
  - →顧客流失(顧客是否流失?)

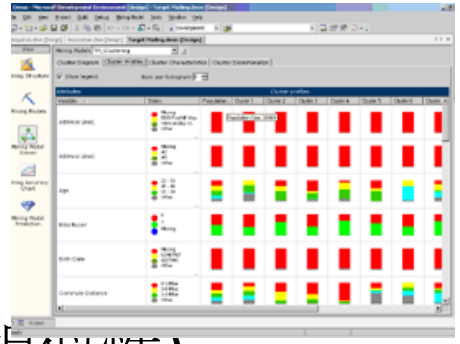


- 預測連續變數(Continuous Variables)的過程我們稱之爲「推估」
- 推估問題的重點在於如何透過已知的屬性來推估未知的連續數值的走向與趨勢
- 推估問題的應用包括：
  - →金融商品價格趨勢變化預測
  - →進貨、銷貨、存貨價量變化趨勢預測
  - →顧客貢獻度、顧客價值預測



群集化

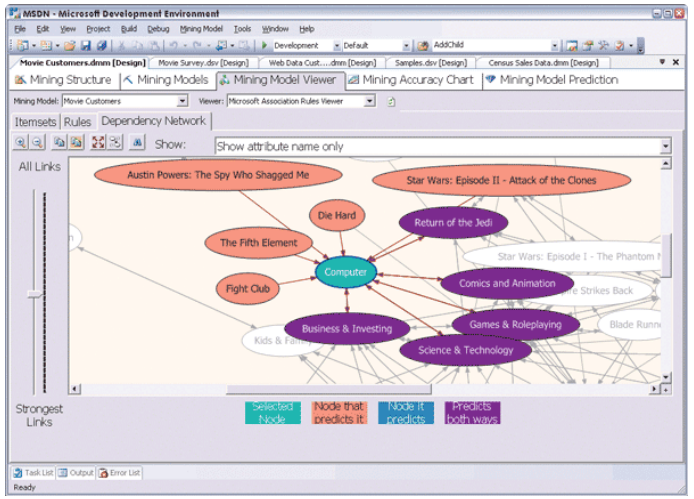
- 根據相似性，將相似的事物分群，這個步驟我們就稱之爲群集化。
- 群集化與分類有個最大的差異點
  - →分類是根據一個明確但尚未發生的分類事實
  - →群集化則沒有所謂的分類準則
  - →分類是對於**未知事實的預測**，而群集化則是找出**事物相似性的內部結構**
- 群集化的應用包括：
  - →顧客分群(根據顧客屬性相似性)
  - →文件分類(根據關鍵字相似性)
  - →晶圓製程瑕疵分布(根據瑕疵分布空間相似性)



asiaMiner
Curve Up the Future

同質分組

- 同質分組就是從歷史資料中，找出哪些物件/事件總是相伴發生
- 又稱之爲關聯規則(Association Rule)或者是購物籃分析(Basket Analysis)
  - →尿布與啤酒
  - →下架的藍乳酪
- 同質分組的應用包括：
  - →產品交叉銷售、自動化推薦
  - →網頁結構分析
  - →文件(病歷、專利文件)關鍵字關聯性分析



asiaMiner
Curve Up the Future

序列

- 同質分組中是找出哪些事物會相伴發生，但是透過序列，我們可以找出事物「先後」發生的順序
- 有時稱這樣的規則爲時序規則(sequential pattern)
- 序列的應用包括：
  - →產品提昇銷售(Up-selling，包括金融、零售、電信等產業)
  - →網頁瀏覽序列分析
  - →逾期繳款行爲模式分析

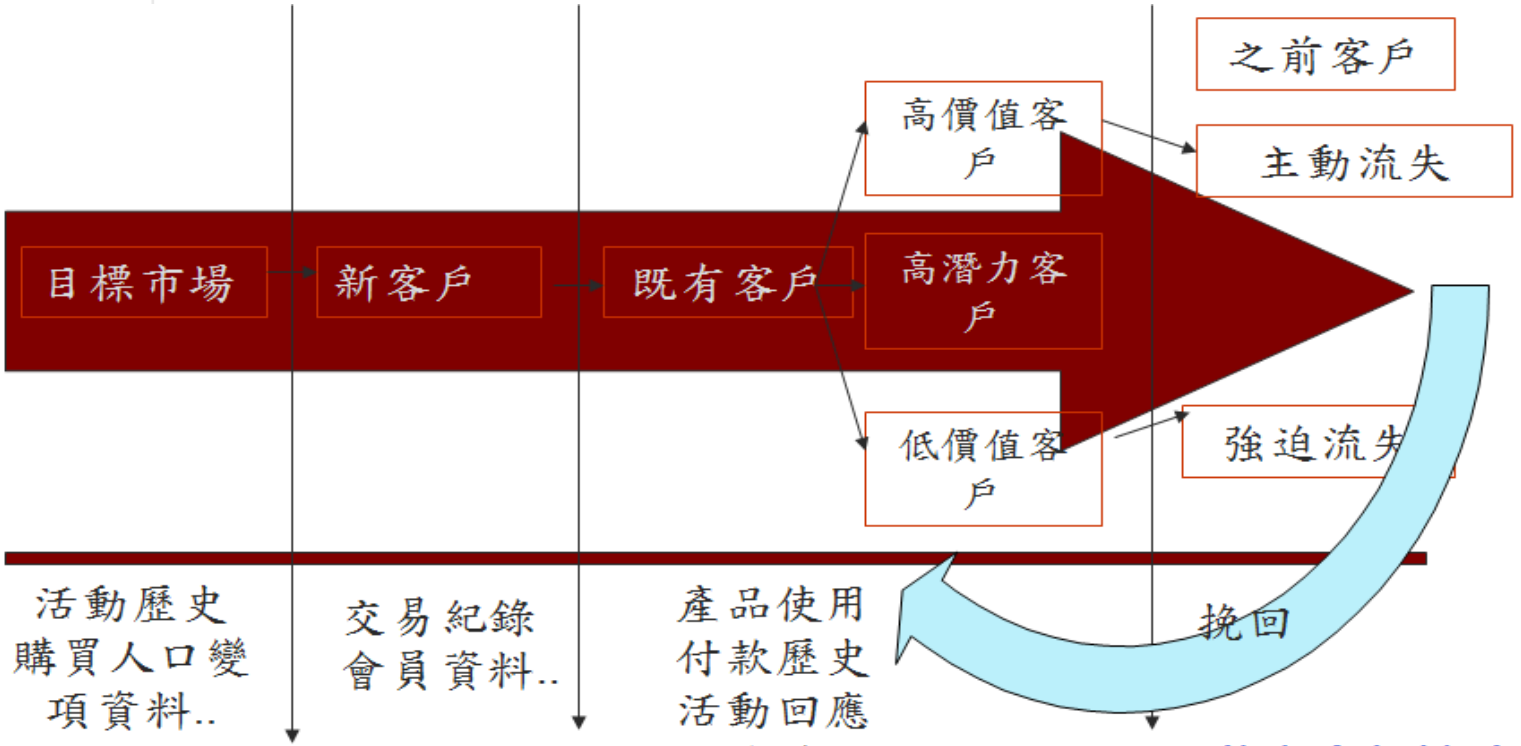
asiaMiner
Curve Up the Future

描述

- 資料採礦是透過演算法來找出潛在的規則，但是人類的觀察力其實是不可被取代的演算法
- 透過資料視覺化的呈現以及對於資料敏銳的觀察力，但是卻同樣能發掘出潛在的規則

asiaMiner
Curve Up the Future

資料採礦的應用



From : Data Mining Theory and Applications

Curve Up the Future

直效行銷

- 透過資料採礦，可以根據過往的行銷成果，比對購買商品客戶以及婉拒銷售客戶的屬性差異，來找出預測電話行銷結果的規則
- 將客戶根據電話行銷回應機率由高到低排序，來作爲分配電話行銷名單、執行直效行銷策略規劃之基礎
  - →預測回應率
  - →話術配對

asiaMiner
Curve Up the Future

交叉銷售

- 提高「市場佔有率」vs.「荷包佔有率(wallet share)」
- 交叉銷售的策略可以根據企業營運而使用不同的資料採礦技術。
  - →**金融業**：商品種類較少，而且獲利較高，因此可以針對個別產品建置回應模型
  - →**零售業**：產品種類太多，而且可能產品生命週期較短而且個別產品的獲利佔比較低，可以使用同質分組或者是序列的技術來找出產品之間的關聯模式

asiaMiner
Curve Up the Future

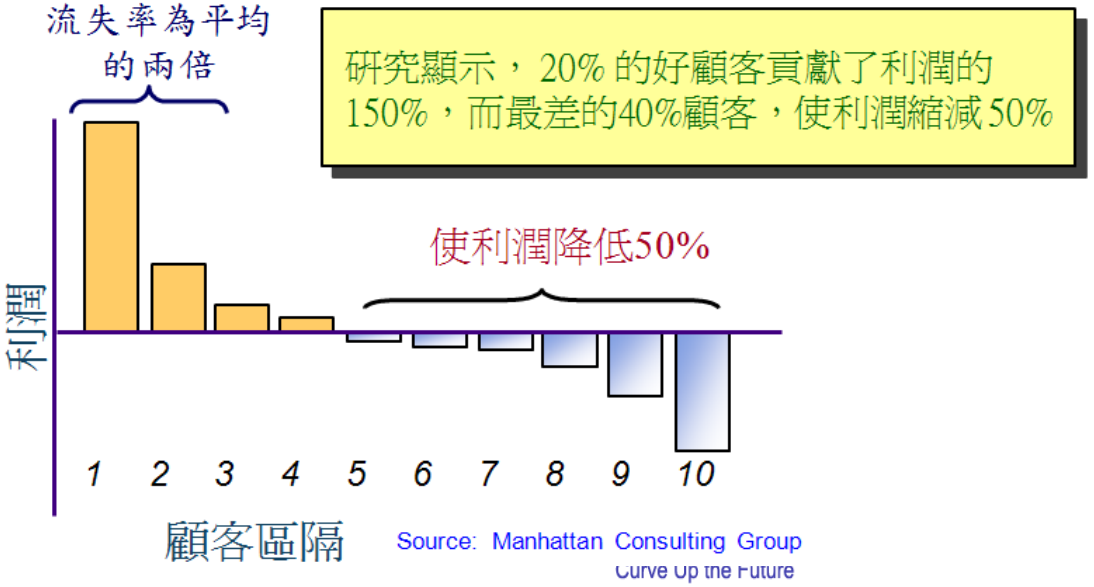
信用風險管理

- 國際清算銀行（BIS）下的巴塞爾銀行監理委員會（The Basel Committee on Banking Supervision)爲促進跨國銀行經營之健全性於1988 年提出以規範信用風險爲主的「巴塞爾資本公約」
- 若是銀行希望能夠有效的降低資本成本，就必須透過大量歷史資料的分析，建立風險預測模型，以內部評等法(IRB)來更精準的評估風險資產的違約風險以計算客戶違約機率(PD, Probability of default)、曝險額(EAD, Exposure at default)以及違約損失率(LGD, Loss given default)。

asiaMiner
Curve Up the Future

流失分析

- 根據「一對一的未來」作者Peppers與Rogers指出，**若能將客戶流失率減少5%，利潤將會有100%的成長**
- 利用資料採礦技術預估客戶流失的可能性



其他應用

- **保險業理賠濫用**
- **航空業預測No Show**
- 信用卡僞卡偵測
- **刑案偵辦**
- 顧客終身價值(Lifetime value)預估
- 製造業製程良率分析
- 網際網路資料採礦(Web Mining)

asiaMiner
Curve Up the Future

Q&A

asiaMiner
Curve Up the Future