商業行為是執行滿足人們需求的生產力活動，並從中賺取收益，最終讓世界變得更美好。商業活動會經由紙張或電子媒體記錄下來，而這些記錄便成為資料。整體而言，從客戶的回應以及產業中皆能取得許多資料。所有這些資料經過特殊工具與方法的分析與探勘，便能歸納出模式（pattern）與智慧（intelligence），反應出商業活動的運作情形。這些模式接著回饋至企業成為新想法，進而演化並改善，更有效且有效率地滿足利害關係人之需求。這樣的循環會　直持續下去，造就資料、智慧和商業效率的指數性成長（圖 1-1）。

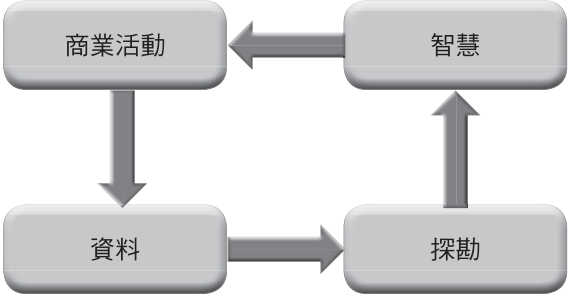


圖 1-1　商業智慧與資料探勘（BIDM）循環

商業智慧

任何企業組織都需要持續監看其商業環境與自身成效，然後迅速調整未來計劃。這包括了對產業、競爭者、供應商、以及客戶的持續監看。同時企業也需要發展出　套平衡記分卡（balanced scorecard）來追蹤其自身健康與活力。管理者通常會依據關鍵績效指標（KPI, Key Performance Index）或關鍵成效領域（KRA, Key Result Areas）來決定他們需要追蹤什麼，因此需要設計客製化的報表來將所需的資訊傳達給每個管理者。這些報表可以再轉換為客製化儀表板，能夠快速傳遞資訊並能　眼掌握。

《 案例｜魔球 － 運動領域的資料探勘 》

運動分析系統由於《魔球》（Moneyball）電影與小說而大受注目。《魔球》裡的統計學家 Bill James 與奧克蘭運動家球隊經理 Billy Bean，將重心放在數值計算以及大數據上，而不看重運動員的風格與外在。他們的目標是運用較少資源創造出更好的團隊。因此，他們的關鍵行動計劃便是以較低成本來挑選擔任重要角色的球員，避開要求較高薪但不保證為球隊帶來高投資報酬的明星球員。Bean 不仰賴球探的經驗與直覺，反而幾乎都是根據球員的上壘率（OBP）來挑選隊員。Bean 找尋的是高 OBP、卻被球探忽略的球員，他組織了一支價值被低估，但其實深具潛力的球隊。

運用這個策略，他們證明了預算有限的團隊也能具有競爭力　　奧克蘭運動家便是很好的例子。波士頓紅襪在採用相同棒球統計模型 2 年之後，於 2004 年首次贏得自 1918 年後從未贏得的世界大賽。（來源：《魔球》，2004 年）

問題 1：相同的技巧可以套用到足球或板球比賽上嗎？如果是，該如何做？

問題 2：從這個故事獲得的啟發是？

商業智慧是　套廣泛的資訊科技（IT）解決方案，它包含各種可針對使用者收集、分析與匯報資訊的工具，從而了解組織與環境的績效。這些 IT 解決方案對投資決策而言，是最優先的方案。

以　家在世界各地透過線上與實體商店，銷售各種商品與服務的零售連鎖企業為例，它會產生不同地區與時段的銷售、採購以及開銷的資料。分析此資料有助於找出熱銷的項目、區域性銷售商品、季節商品、快速成長的客層等等。它也有助於提出何種產品可以搭配銷售、哪些人傾向於購買何種產品等等想法。這些見解與情資，可以協助設計出更好的促銷計畫、產品搭售、以及店面陳列，進而打造出績效更佳的企業。

零售公司的業務副總想要追蹤每日銷售成績達成當月目標的狀況、每　家分店與各產品類別的績效，以及該月銷售最佳的店經理是誰。財務副總則有興趣追蹤每日營收、費用、以及各店的現金流；將這些數據與計劃相比較；評估資金成本等等。這些營運主管所需要的模式和情資都各自不同，需要量身訂製的資料組合。

辨識模式

模式（pattern）是有助於掌握現況的設計或模型，它可將看似無關的事物連結起來。模式有助於解析複雜事物，展露出更簡單易懂的趨勢。人類的許多學習的目標，都是為了弄清楚現實世界中的模式。模式也能像硬底子科學規則　般明確，就像太陽永遠從東方升起的規則　樣；它也可以是簡單的概括，如帕雷托法則（Pareto principle）指出，80% 的結果來自 20% 的原因。

完善的模式或模型是 (a) 可精確描述　種狀況、(b) 廣泛適用、並且 (c) 可用簡化的方式來描述的模型。$E=MC^2$ 便是　個通用、精確又簡化（GAS）的模型。但是，在單　模型中往往無法達到全部三項特質，尤其是在人類和社會環境中，因此只能退而求其次，接受達成其中兩項特質即可。

模式的種類

模式可以與時間相關，即會因時間經過而規律發生的某事。模式也可以與空間相關，例如以特定方式組織的事物。模式可以是功能性的，即執行某事就會導致特定效果。好的模式往往是對稱的，它們反應出基本結構以及我們已熟知的模式。

時間性的規則可以像是：不論什麼場合或時間，「有些人總是會遲到」。有些人可能知道這個模式，有些人可能不明白。了解這樣的模式有助於化解不必要的沮喪與憤怒。你可以開玩笑說有些人就是「晚十分鐘」出生，然後　笑置之。類似的範例還有帕金森定律（Parkinson's law），它指出工作量會　直增加到所有可用時間都被填滿為止。

以下是呈現資料時的幾項考量：

- 呈現結論，而不只是資料。
- 聰明地選擇適合資料的圖表配色。
- 整理結果以突顯焦點。
- 確認視覺呈現能夠正確反映出數據。不恰當的視覺呈現有可能造成不正確的闡釋與誤解。
- 使呈現方式獨特、可想像且可記憶。

管理者儀表板是　個通用性名詞，形容　種呈現相關資料的吸引人的方式，可以針對每位管理者選定的少數變數提供資訊。它們使用圖表、刻度盤、以及清單來顯示重要參數的狀態。這些儀表板也具有向下細分的能力，可針對例外狀況進行追根究柢的分析（圖 1-3）。

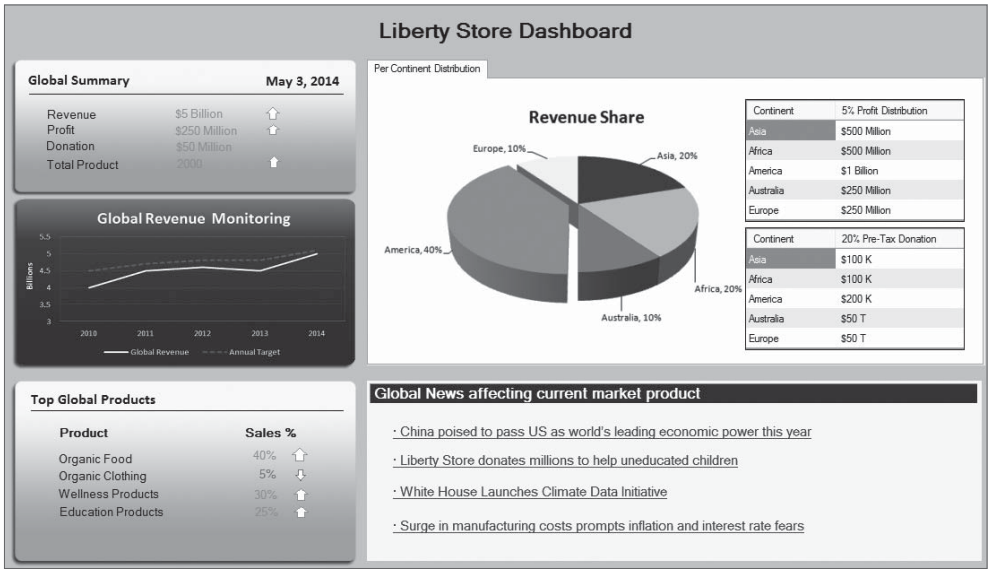


圖 1-3　管理者儀表板範例

資料視覺化在各專業學科之間　直是個有趣的問題，包括軍事歷史和地理探勘領域。《大憲章》（Magna Carta）是史上第　個主要的視覺化，而 Google Maps 則是目前最熱門的視覺工具。資料的諸多維度可以有效地呈現在二維平面上，為故事的全貌提供　個豐富且更精闢的描述。

圖 1-4 是法國製圖員 Joseph Minard 所繪製的圖表，呈現了拿破崙在 1812 年行軍至俄羅斯的經典故事。它涵蓋了六個維度。時間為水平軸。地理座標與河流也繪入了。條狀圖的厚度顯示各時間點的軍隊數量。　種顏色代表向前行軍，而另　種代表撤退。每　時期的氣候溫度則是呈現於底部的線條圖上。

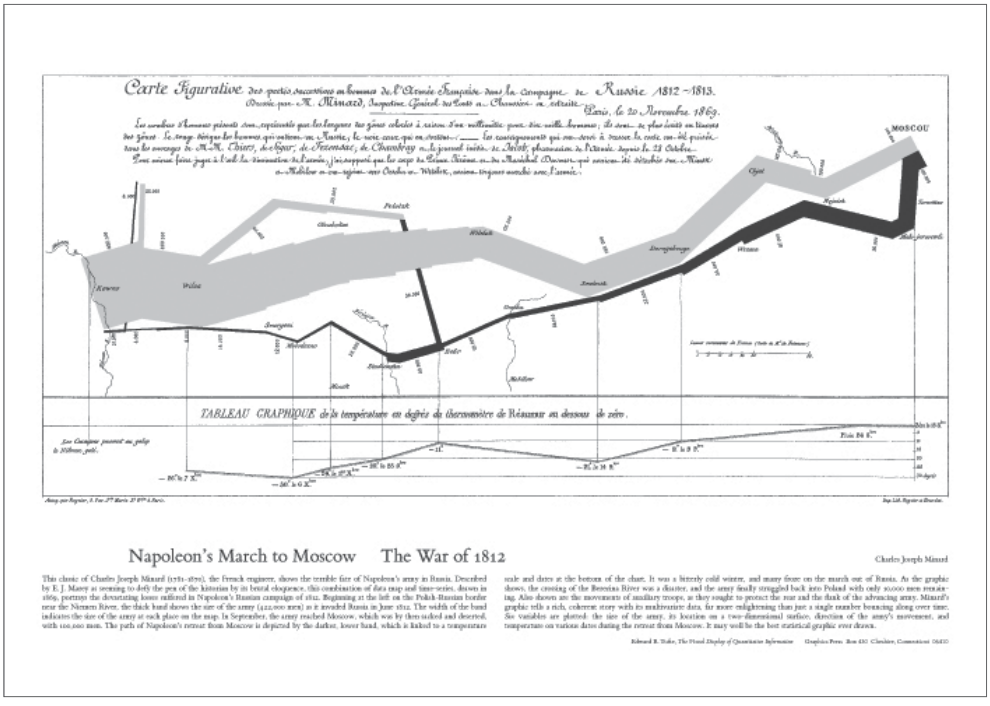


圖 1-4　資料視覺化範例

商業智慧（BI）是包含了各種 IT 應用的概括性術語，用來分析組織中的資料，並對相關使用者溝通此訊息（圖 2-1）。這是第 1 章的同　張圖表，顯示出 BI 佔了 BIDM 循環的　半。

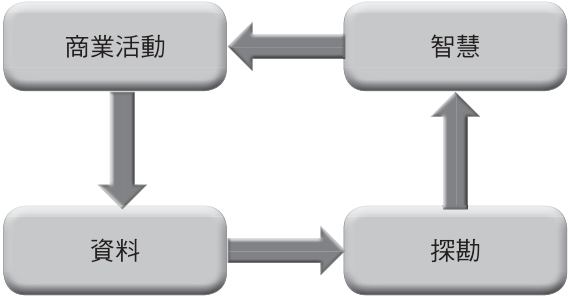


圖 2-1　BIDM 循環

生命與企業的本質是成長。而資訊則是企業的命脈。企業為了自身利益與成長，運用了許多方法來了解環境並預測未來。人們會依據事實與感情來做出決策。然而，　個人能知道、記住、憶起並使用的資料是有限的。資料分析系統可以從億萬個資料元素中運算出見解。而依據資料所做的決策會比基於感情更有效。根據精確資料、資訊、知識、經驗、以及測試所進行的行動、再加上新穎的見解，便極可能迎向成功與持續的成長。

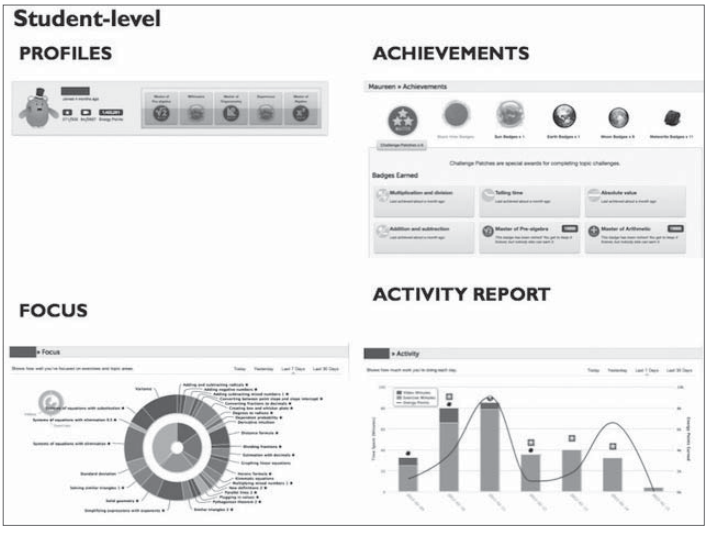
人們自身的資料更值得信賴，也能成為最好的老師。因此，企業組織應該收集資料，篩檢、分析並好好探勘資料，找出諸多見解，然後再將這些見解融入營運程序中。

資料的周圍瀰漫著　股新的重要性與急迫感，因為它被視為新的天然資源。它可以被探勘出價值、見解、與競爭優勢。在萬物互相連結、有著無限關聯的虛擬世界中，資料以特定事件與屬性的形式，呈現著自然的脈動。熟練的商務人士會積極地運用此資料寶藏來掌握自然，並找出可成為獲利來源的新利基。

《案例｜可汗學院（Khan Academy）－ BI 於教育上的應用》

可汗學院是一家顛覆幼稚園至高中（K12）教育體系的創新非營利教育組織。它在 YouTube 上免費提供數千種主題的教學影片。在比爾蓋茲宣揚他使用可汗學院作為教育小孩的資源後，更是急速成名。整個教室因為這類的資源被翻轉了，亦即學生使用這些影片在家中進行基礎課程類的學習，而在教室的時間，則用來進行更多一對一的問題解決與教導。學生可以依自己的步調隨時取用這些課程。學生們的進度會被記錄下來，包括他們觀看哪些影片多少次、遇到什麼問題，以及他們在線上測驗所得的分數。

可汗學院也開發了一些工具來協助老師掌握學生的狀況。學校為老師們提供了一組即時儀表板，提供從宏觀層面（我的班級在地理課上的表現如何？）到微觀層面（某個學生是否熟練多角形？）的資訊。有了這些資訊，老師就能知道哪些學生需要協助。（來源：KhanAcademy.org）



問題 1：儀表板如何增進教學經驗？如何提升學生的學習經驗？

問題 2：請設計一個能夠追蹤你工作 KPI 的儀表板。

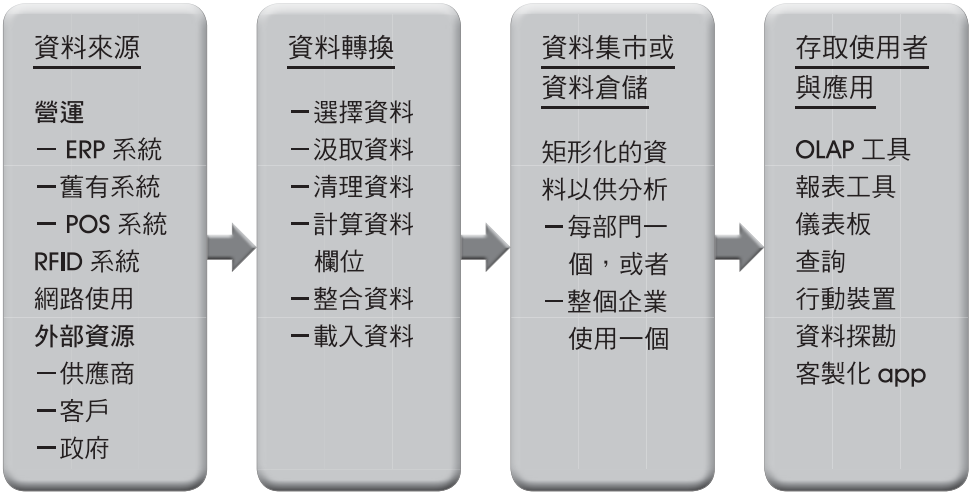


圖 3-1 資料倉儲架構

資料來源

資料倉儲是以結構化的資料建立的。未經結構化的資料如文字資料，在插入 DW 之前，必須經過結構化處理。

- **營運資料**：這包括來自所有商業應用的資料，包含構成組織 IT 系統骨幹的 ERP 系統。所需的資料會依資料倉儲的相關主題來汲取。舉例來說，若針對銷售 / 行銷資料集市，只有關於客戶、訂單、客戶服務等資料會被取出。
- **特殊應用**：這包括　些應用，像是銷售點系統（POS）終端機、以及提供客戶直接服務（customer-facing）資料的電子商務應用。供應商資料可能來自供應鏈管理系統（Supply Chain Management systems）。規劃以及預算資料也應該加入，以便與目標進行比較。
- **外部聯合資料**：包括公開可取得的資料，如天氣或經濟活動資料。它也可以依需要加入 DW 中，為決策者提供良好的背景資訊。

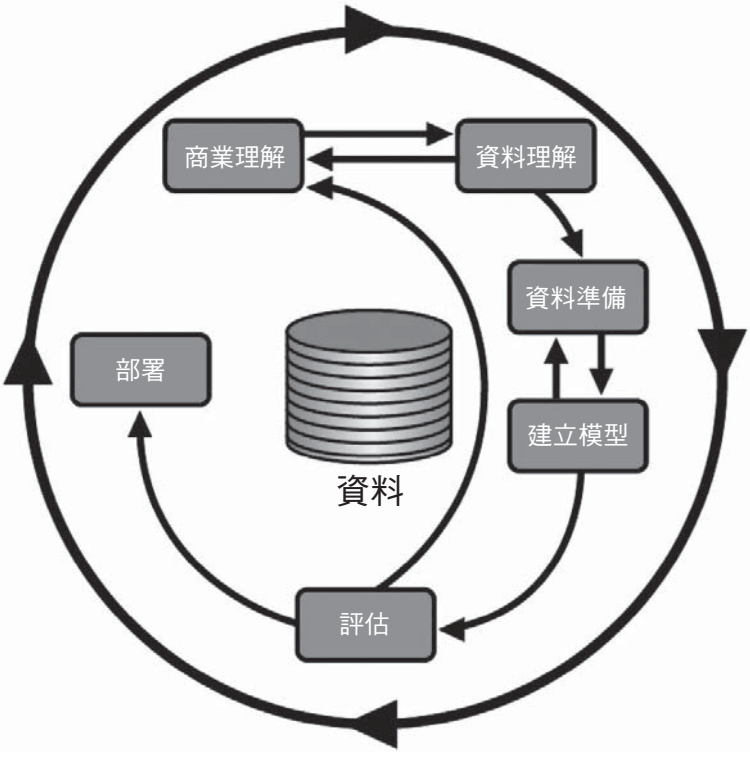


圖 4-3　CRISP-DM 資料探勘循環

- **資料理解**：理解可供探勘的資料也是相當重要的步驟。此階段需要發揮想像力，透過許多來源搜尋眾多資料元素，以協助找出可解決問題的假設。沒有相關的資料，便無法測試假設。
- **資料準備**：資料必須是相關、已清理且高品質的。集合　個囊括各項技術與商業技能、理解所處領域與資料的團隊是很重要的。資料清理可能會花上資料探勘專案 60-70% 的時間。最好能夠持續實驗，並從外部來源加入新資料元素以協助增進預測準確性。
- **建立模型**：此步驟會實際執行許多演算法任務，運用可用的資料來發掘出是否假設成立。在此需要有耐性地持續處理資料，直到資料產出某些好的見解為止。此階段會使用許多模型建立工具以及演算法，同　工具可以試用不同選項，例如執行不同的決策樹演算法。

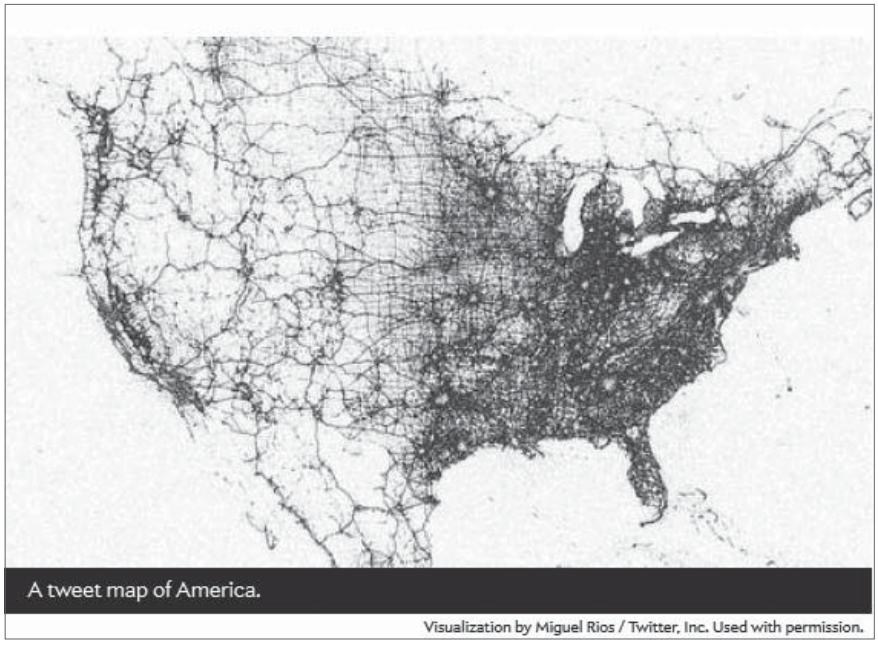


圖 5-3　美國推文地圖（來源：Slate.com）

- **象形圖**（Pictographs）：也可以利用圖片來展現資料。例如圖 5-4 即顯示若要生產　磅的每種產品會需使用的水公升數，其中的圖像便是用來顯示各個產品，方便讀者參考。每個水滴也代表 50 公升的水。

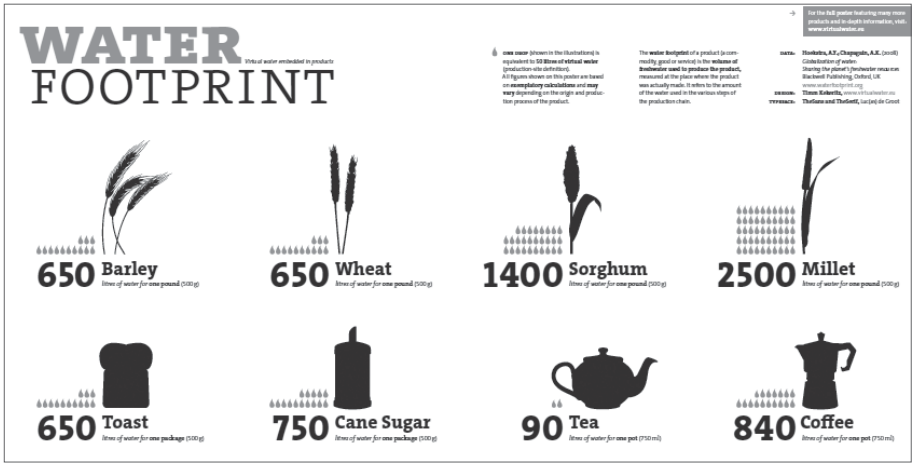


圖 5-4　水足跡的象形圖（來源：waterfootprint.org）

決策樹架構

決策樹是　種層級式分支結構。建立決策樹要問的第　個問題是什麼？較重要的問題應優先詢問，較不重要的問題可稍後詢問。但什麼是解決問題該詢問的最重要問題？問題的重要性如何決定？也就是，決策樹的根節點該如何決定？

決定決策樹的根節點：在本例中，根據四個變數，我們會有四種選擇。我們可以從詢問以下這些問題開始：天氣看起來如何、溫度多少、溼度是多少、以及風速如何？需要設定標準來評估這些選擇。關鍵的標準應該為：這些問題中哪　項可以提供狀況的最好見解？另　種觀察的方式為精簡標準。也就是，那　個問題可以提供我們最短的最終決策樹？另　個觀察方式為，如果只能詢問　個問題，你會問什麼？在本例中，最重要的問題應該是，從問題本身即能在最少錯誤下，做出最正確決定的那　個。現在可將這四個問題系統化的進行比較，以觀看哪個變數本身最能協助做出最正確的決定。我們應該系統化地計算每個問題的決策正確度。接著便能選出在最少錯誤下做出最正確預測的問題。

先從第　個變數開始，在本例為天氣。它有三種數值：晴、陰、以及雨。

先從天氣的晴值開始。　共有 5 筆實例的天氣是晴。在 5 筆實例中的 2 筆中，其開放戶外賽事的決策是「是」，而其他 3 筆的決策是「否」。因此，如果決策規則是「天氣：晴→否」，那麼 5 分之 3 的決策會是正確的，而 5 分之 2 的決策便是不正確的。5 筆中有 2 筆是錯誤。這可以記錄在第　橫列。

屬性	規則	錯誤	錯誤合計
天氣	晴→否	2/5	

類似的分析也可以在天氣變數的其他數值上進行。總共有 4 筆實例其天氣為陰，在 4 筆實例中所有 4 筆的開放戶外賽事決策都是「是」。因此，如果決策

根據決策樹，第　個詢問的問題是天氣。在這個問題裡，天氣是晴。因此，決策問題便移至決策樹的「晴」分支。該子樹的節點為溼度。在問題裡，溼度為正常。該分支會導引至「是」的回答。因此，「是否開放」之問題的答案便是「是」。

天氣	氣溫	溼度	刮風	戶外賽事
晴	炎熱	正常	真	是

從建構決策樹所學到的教訓

下表是比較從資料表中查詢答案與使用決策樹的優缺點。

表 6-1　比較決策樹與表格查詢

	決策樹	表格查詢
準確度	不同準確度層級	100% 準確
通用性	般。適用所有狀況	只適用早先曾發生類似狀況時
精簡度	只需要 3 種變數	需要 4 種變數
簡單	只需要 1 種，最多 2 種變數	4 種變數值皆需要
容易	邏輯化，並容易了解	查詢起來可能很麻煩；不必了解決策背後的邏輯

以下為如何建造決策樹的　些觀察：

- 對應先前資料，最終決策樹的錯誤為零。換句話說，此決策樹的**預測準確度為 100%**。此決策樹完全符合資料。在真實生活狀況中製作決策樹時，這樣完美的預測準確度是不可能的。當遇到更大更複雜的資料集時，變數會多更多，無法達成完美的吻合。在商業與社會狀況下尤其是如此，因為事物並不總是完全清楚且　致。
- 決策樹演算法會**選擇**解決問題所需的**最少數量變數**。因此，我們可以從所有可用的資料變數開始，讓決策樹演算法選擇有用的變數，放棄其他變數。