» 21 世紀的資源

資料與資訊的差異是？

在日常生活當中，經常會遇到需要「預測」、「評估」的情況（圖 1-1）。此時，除了仰賴經驗、直覺外，也可搜尋舊有資料，或者利用問卷調查、網際網路等蒐集所需的內容。**大量蒐集正確的資料，可提升預測、評估的準確率**，資料就好比「21 世紀的資源」。

這邊需要注意的是，**資料**和**資訊**的差異。一般而言，「資料＝未經處理的狀態」、「資訊＝經過處理的狀態」，這樣解釋的話，資料是尚未經過處理的內容。就直觀而言，資料是「羅列的數字」、「按固定形式蒐集的數據」。若說資料是電腦容易處理的內容，則資訊是「人類容易閱讀的訊息」、「促使對方採取下一步行動的訊息」（圖 1-2）。

將資料轉成資訊

例如，聽聞「現在氣溫 18 度」會產生什麼感想呢？這個「18 度」可視為資料，夏天會覺得「涼爽」；冬天會覺得「溫暖」。即使是同一個資料，資訊也會因周遭情況而變。

觀看隔天的天氣預報後，便利商店的店長可能會決定「增加冰淇淋的商品數量」、「減少關東煮的食物品項」；家長可能據此決定孩童隔天上學的服裝。

此時，需要的是氣溫「資料」、氣象預報人員所傳達的「資訊」，正確的資料會影響傳達內容的可信度。

圖 1-1 需要預測、評估的情況

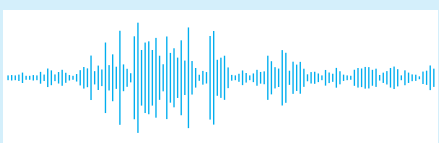
圖 1-2 資料與資訊的差異

資料

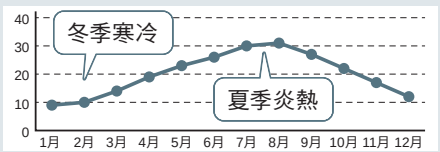
月	1月	2月	3月	4月	5月	6月
氣溫(℃)	9	10	14	19	23	26

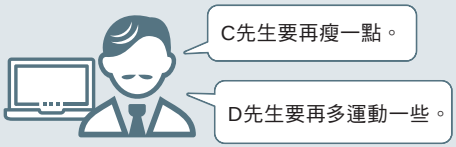
月	7月	8月	9月	10月	11月	12月
氣溫(℃)	30	31	27	22	17	12

名稱	A	B	C	D	E	F
身高(cm)	165	167	170	173	178	182
體重(kg)	55	61	80	72	73	78



體重





Point

- 大量蒐集正確的資料，可提升預測、評估的準確率
- 將電腦擅長處理的資料轉為人類容易閱讀的資訊，可促使他人採取行動

» 資料增加的原因

由資訊化社會邁入資訊社會

一般而言，煤炭帶來的輕工業機械化，稱為第一次工業革命；石油帶來的重工業機械化，稱為第二次工業革命；電腦帶來的機械自動化，稱為第三次工業革命。導入電腦後，資訊的重要性逐漸提高，自 1970 年左右邁入所謂的**資訊化社會**。

這股趨勢並未改變，如今已有第四次工業革命、工業 4.0 的說法，意指 AI（人工智慧）、**物聯網**（IoT：Internet of Things）帶來的高度自動化（圖 1-3）。然後，最近迎來**資訊社會**，不再是人類將資料轉為資訊的「資訊化」，而是**資訊的相關技術已經存在，人類可自由地使用資訊的社會**。

物聯網與感測器帶來的便利社會

前面提到的**物聯網**是「物體的網際網路」，邁入不只電腦、智慧手機，而是電視、空調、冰箱等各種家電皆可聯網的便利時代（圖 1-4）。

從外出地返家前先開啟空調，一到家就可享受舒適的溫度。在超市購物的時候，透過智慧手機確認冰箱中的庫存，預防漏買重要食材。

如果物聯網設備搭配**感測器**會變得更加方便。例如，房間變暗時自動關閉窗簾；有人走動時自動開啟電燈；溫度下降時自動啟動暖氣。

如上所述，在資訊社會中，除了供人類判斷情況外，資訊也扮演著連動設備的重要角色。

圖 1-3 工業革命的變遷

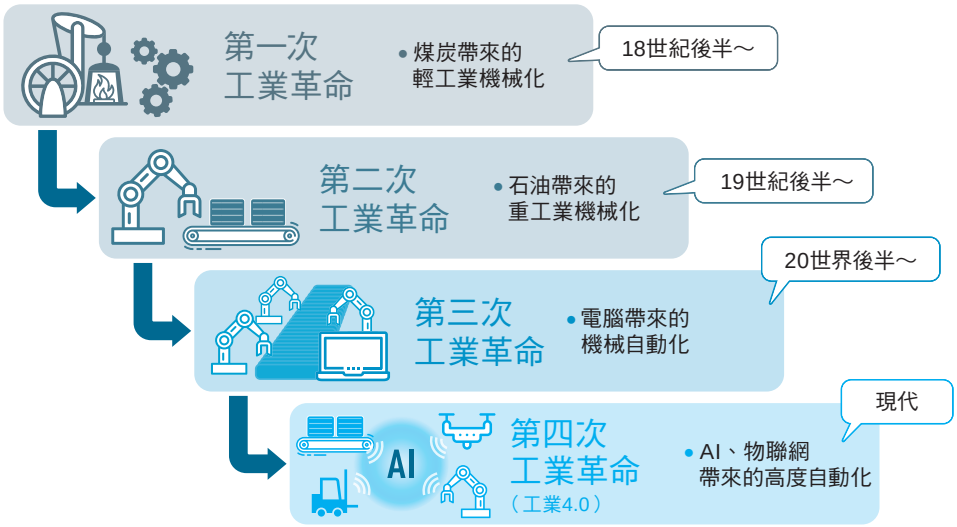
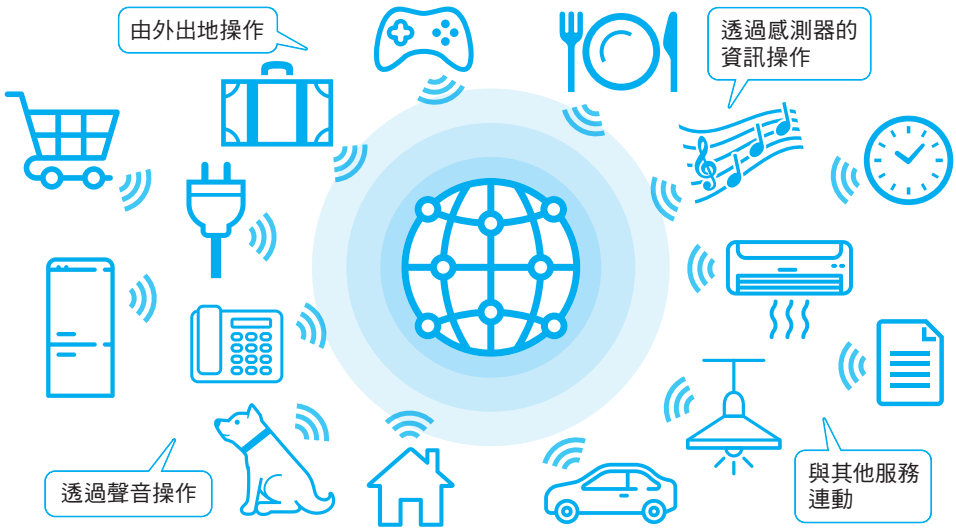


圖 1-4 物聯網可實踐的事情



Point

- AI、物聯網帶來的高度自動化，稱為第四次工業革命
- 各種設備可藉物聯網連接網際網路，讓我們的生活更加便利

» 結合各種知識進行分析

資料科學需要哪些知識？

在分析資料的時候，不可僅單純知道分析手法。即使了解數學上的分析手法，若缺少程式設計的知識，則無法編寫處理資料的程式。

再者，即使嫻熟程式設計，若不曉得資料背後的意義、欠缺商務方面的知識，對於如何編寫程式也會沒有頭緒。

如上所述，數學、統計等科學領域的知識；程式設計、伺服器架設等工程領域的知識；經濟、經營等商務領域的知識等，**資料科學**需要**結合各種知識進行資料分析**（圖 1-5）。

從資料獲得未注意到的見解

分析資料的時候，我們想要的是「新的見解」，期望從資料發現人類未注意到的觀點。此過程可比喻成從地底挖掘礦物（mining），稱為**資料探勘**（圖 1-6）。

資料探勘著名的例子有「買尿布的人也常買啤酒」，購買紙尿布的人父往往會順手添購啤酒。姑且不論真偽，這個傾向是相當有趣的發現。

如上所述，資料探勘是藉結合 AI 等技術分析龐大的資料，推導資料傾向並找出最佳組合的作業。由於需要進階的分析，一般是在大學等研究機構、企業的研發部門等進行，且著重在**我們人類如何運用得到的見解**。

圖 1-5 資料科學的相關領域

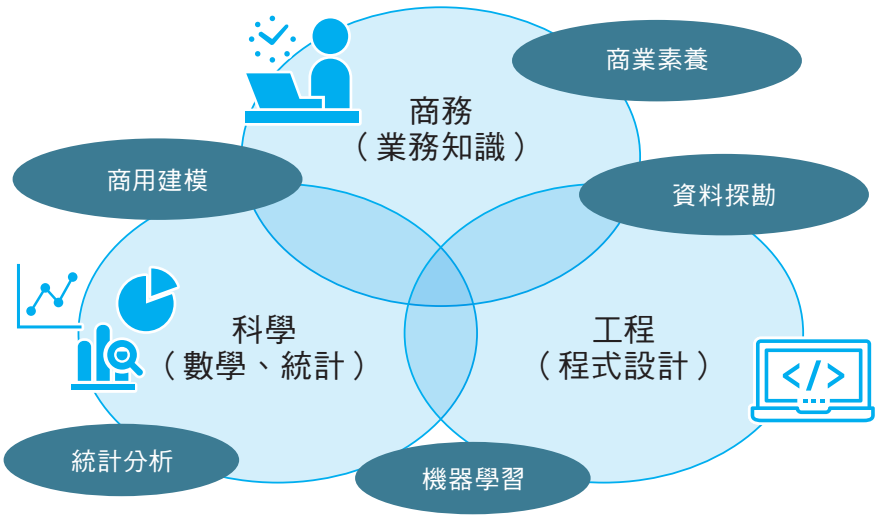
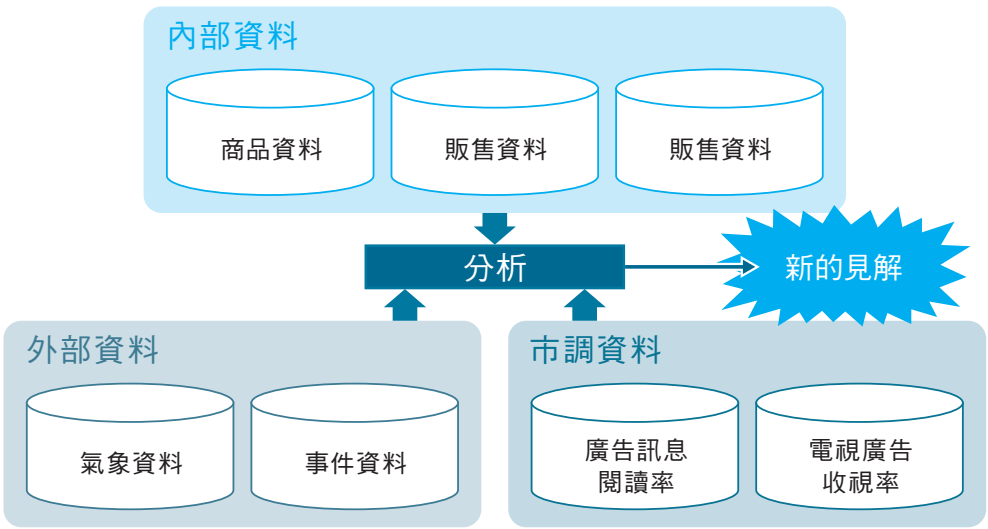


圖 1-6 資料探勘的例子



Point

- 資料科學需要廣泛的知識，如數學、統計學、程式設計、商業事務等
- 資料探勘是藉由分析大量資料，獲得人類未注意到的見解

» 找出資料價值的職業

資料分析的明星職業

以科學技術進行資料分析等作業的人，稱為**資料科學家**。資料科學家曾被喻為「21 世紀最性感的職業」，一時蔚為話題 ※1。

資料科學要結合科學、工程、商務等知識，來洞察大數據。然而，一個人難以掌握全數知識，且商務所需的知識亦因業務而異，故通常是集結不同領域的專家，建立團隊並以部門單位進行分析（圖 1-7）。

輔助資料科學家的職業

資料科學家善於分析龐大的資料，但若沒有資料也無用武之地。因此，輔助資料科學家的**資料工程師**誕生了。

該職種的主要業務是，整頓有利於資料科學家的分析環境，除了加工整合分析用的資料外，還包括架設伺服器等基礎設施、準備分析資料的雲端服務。不僅業務範圍廣泛，還需要豐富的 IT 知識（圖 1-8）。

涵蓋資料分析、諮詢顧問的職業

與資料科學家相似的職業還有**資料分析師**，如同其名，意指分析資料的人員，除了以資料探勘等手法進行分析，還負責諮詢顧問的相關業務。

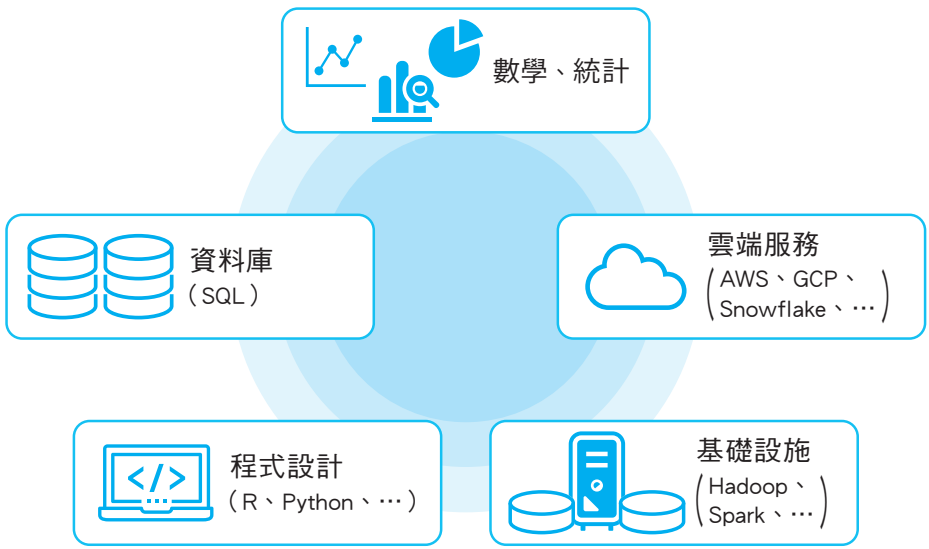
資料科學家是兼具資料工程師和資料分析師的人才，部分企業將其定位為兩者之上的進階職業。

※1 資料來源：Davenport, Thomas H., and D. J. Patil. "Data Scientist: The Sexiest Job of the 21st Century."Harvard Business Review 90, no. 10 (October 2012): 70 76.

圖 1-7 企業中的資料科學家

配置方式	優點	缺點
集結少數的天才型人才 （一個人精通所有領域）	●可有效率地分析 ●成本較低	●難以尋得人才 ●一個人的負擔較大
橫跨部門的配置 （利用主要業務的空檔進行分析作業）	●可活用商務知識 ●成本較低	●難以撥出時間給非主要業務的分析 ●難以獲得顯著的成果
按部門集結人才的配置 （集結各種背景的人才進行分析）	●容易集結人才 ●取得重大成果時的效果顯著	●未取得成果時會不斷墊高成本 ●可能逐漸偏離實際業務

圖 1-8 資料科學所需的知識



Point

- 一個資料科學家難以掌握廣泛的知識，故通常採取團隊進行分析
- 與資料科學家相似的職業，還有資料工程師、資料分析師

» 資料需要加工處理

電腦容易處理的資料

電腦在處理資料時，**得先讓程式知道儲存的檔案布置（資料的排列和架構）等**。

例如，處理 CSV 通訊錄的程式，會將目標檔案存成 CSV 格式。第一列輸入名稱、第二列輸入郵遞區號、第三列輸入住址……等，得先決定各列的儲存內容才有辦法處理（圖 1-9）。

這種事前決定檔案資料結構，方便電腦處理的資料，稱為**結構化資料**。除了通訊錄等表格形式外，還有 XML、JSON 等各種檔案格式（format）。

結構化資料具有容易搜尋、重新排序等特色。在通訊錄中，可搜尋名稱含有特定字句的人物，也可按郵遞區號重新排序。

人類經常使用的資料

另一方面，備忘錄、日記等單純排列文句的資料，稱為**非結構化資料**（圖 1-10）。即使日記中提到某人的名稱，電腦也難以判斷該詞為人名。

人類能夠理解文章意思做出判斷，但電腦無法理解裡頭想要傳達的內容，搜尋時可找出一致的關鍵字，卻不易判斷名稱是否含有特定字詞。

除了文章之外，圖像、影片、聲音也有類似的問題。近年伴隨 AI 的問世，人臉辨識技術也跟著發展起來，不過目前的辨識準確率仍不甚理想。

圖 1-9 結構化資料的例子

CSV檔案

```
名稱,郵遞區號,住址,電話號碼,郵件地址
山中太郎, 105-0011, 東京都港區芝公園, 03-1111-1111, t_yamada@example.com
鈴木花子, 112-0004, 東京都文京區後樂, 03-2222-2222, h_suzuki@example.co.jp
佐藤三郎, 160-0014, 東京都新宿區內藤町, 03-3333-3333, s_sato@example.org
```

以試算表軟體開啟

名稱	郵遞區號	住址	電話號碼	郵件地址
山田太郎	105-0011	東京都港區芝公園	03-1111-1111	t_yamada@example.com
鈴木花子	112-0004	東京都文京區後樂	03-2222-2222	h_suzuki@example.co.jp
佐藤三郎	160-0014	東京都新宿區內藤町	03-3333-3333	s_sato@example.org

圖 1-10 非結構化資料的例子

日記、部落格等

8月15日
晴天

僅有聲音、影片、圖像檔案，無法進行搜尋

今天和○○

一起去逛了XX。

整天都是好天氣，非常愉快。

希望之後還有機會再去一次。

無法知道文章哪邊提到人名、哪邊提到地點

Point

- 事前決定檔案架構、方便電腦處理的資料，稱為結構化資料
- 電腦難以從日記等非結構化資料找出人名、地點

» 巨量資料是座寶山

何謂 3V ？

資料科學備受注目的理由，包括累積的資料量已多到人類難以處理。隨著網際網路的興盛，愈來愈多人上網發送訊息，再加上物聯網技術中的感測器，物件也開始會發送訊息（圖 1-11）。

這般巨量的資料稱為**大數據**，普通的電腦難以有效處理。大數據具有「Volumn（巨量性）」、「Velocity（即時性）」、「Variety（多樣性）」等特性，三者合稱為 **3V**。

Volumn 如同其名意指巨大的數量；Velocity 意指不做批次處理，而是即時處理頻繁更新的資料；Variety 意指不僅結構化資料，也能夠處理非結構化資料（圖 1-12）。

透過分析這樣的大數據，期望獲得過往未發現的見解。

4V 或者 5V

「4V」是「3V」加上「Veracity（真實性）」；「5V」是「4V」再加上「Value（價值性）」（圖 1-13）。Veracity 意指徒有巨量的資料沒有意義，要蒐集有用的、高可信度的資料；Value 意指**光擁有資料沒有意義，必須藉資料分析等解決社會議題，或者孕育新的價值**。

如今，另外再加上「Virtue（道德性）」，也開始講究處理資料時的道德倫理。

圖 1-11 資料驟增的理由

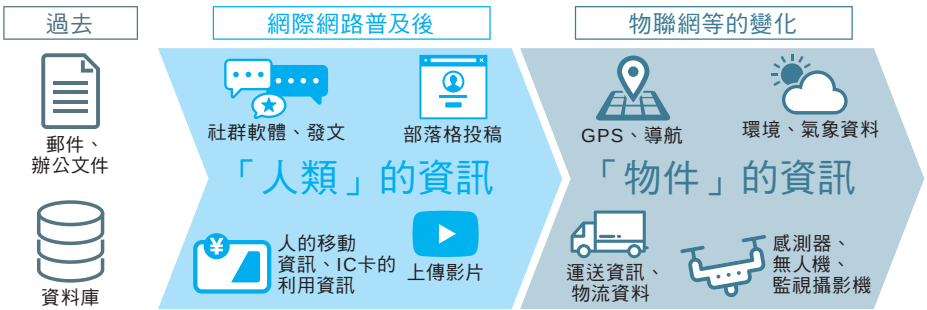
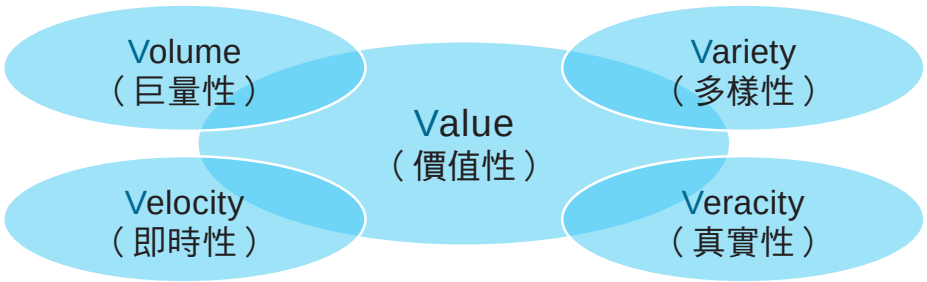


圖 1-12 3V 的相關技術

3V	條件與技術
Volume	需要儲存巨量的資料 例）利用雲端服務、利用高擴展性的儲存服務
Velocity	需要存放並即時處理頻繁更新的資料 例）準備高速網路設備、就近保存與處理、利用快取記憶體
Variety	需要存放並分析多樣的資料 例）利用 NoSQL、利用構詞分析、聲音辨識等技術

圖 1-13 5V



Point

- 大數據具有數量龐大、需要即時處理、種類多樣等特性
- 除了資料的巨量性之外，最近也講求真實性、價值性等其他要素

» 人類和電腦容易處理的資料不同

人類容易閱讀的資料與電腦容易處理的資料

在發表簡報的時候，會整理成圖表數據等方便人類閱讀的格式。Excel 等試算表軟體能夠簡單轉為表格，但使用程式處理時可就不同了。

例如，對人類來說，圖 1-14 是經過整理、容易閱讀的圖表數據，但對程式來說卻是難以處理的內容。即便是相同意義的資料，程式比較容易處理圖 1-15 的格式。

如圖 1-14 的資料稱為**雜亂資料**（messy data）；如圖 1-15 的資料稱為**整潔資料**（tidy data）。根據 Hadley Wickham 論文的 ※2，整潔資料具備下述三個條件，亦即**直列表示變數項目；橫行表示觀測資料**：

1. 每列僅有一個變數
2. 每行僅有一個觀測對象
3. 每個表格僅有一種觀測單位

整潔資料的優點

使用整潔資料統計人數時，只需要相加人數欄位的數值即可。然後，想要知道特定部門的總人數、男女人數，或者哪個部門人數最多的時候，可用試算表軟體篩選直列單位，簡單調查出來（圖 1-16）。

整潔資料不僅容易新增、刪除、更新資料，也可簡單地重新排序顯示。

※2 Wickham, H. . (2014). "Tidy Data". Journal of Statistical Software, 59(10), 1 23.
（https://www.jstatsoft.org/article/view/v059i10）

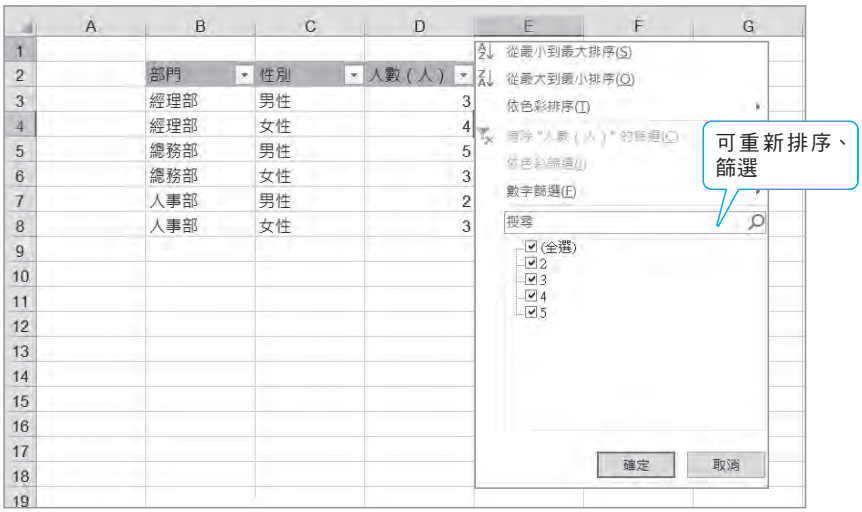
圖 1-14 雜亂資料的例子

	經理部	總務部	人事部
男性	3人	5人	2人
女性	4人	3人	3人

圖 1-15 整潔資料的例子

部門	性別	人數（人）
經理部	男性	3
經理部	女性	4
總務部	男性	5
總務部	女性	3
人事部	男性	2
人事部	女性	3

圖 1-16 整潔資料的優點



Point

- 即便同樣是圖表數據，電腦比較容易處理整潔資料
- 整潔資料不僅方便新增、刪除內容，也容易重新排序、篩選來分析

» 描述資料本身的資料

企業統一管理資料

企業在建立資料庫的時候，共同所需的資料稱為**主檔資料**（master data）。例如，若未登錄顧客的姓名、住址等資訊，則無法寄送顧客購買的商品；若未登錄商品的資訊，則無法記錄販售資料。

如上所述，**對企業來說，當作基礎的主檔資料非常重要**。一般來說，主檔資料會簡稱為「主檔」，建立「顧客主檔」、「商品主檔」等表格。然後，將這些主檔資料連動其他表格，來實踐各種應用程式（圖 1-17）。

相反地，同份資料存於多個位置、需要按部門更換 ID 等，若主檔資料尚未整理，得先統整各項資料才行。

描述資料本身的資料

為了有效率地管理資料，**得掌握資料有哪些項目、何種儲存格式等**。這類內容會因資料而異，需要描述資料本身的資料——**元資料**（metadata），按照資料管理項目、格式（圖 1-18）。

在圖像、聲音、影片等檔案前面，預留放置元資料的空間，與其他資料存成一個檔案。

另外，DBMS（資料庫管理系統）的資料字典（Data Dictionary），是資料庫管理、保存元資料的場所。

圖 1-17 主檔資料的架構

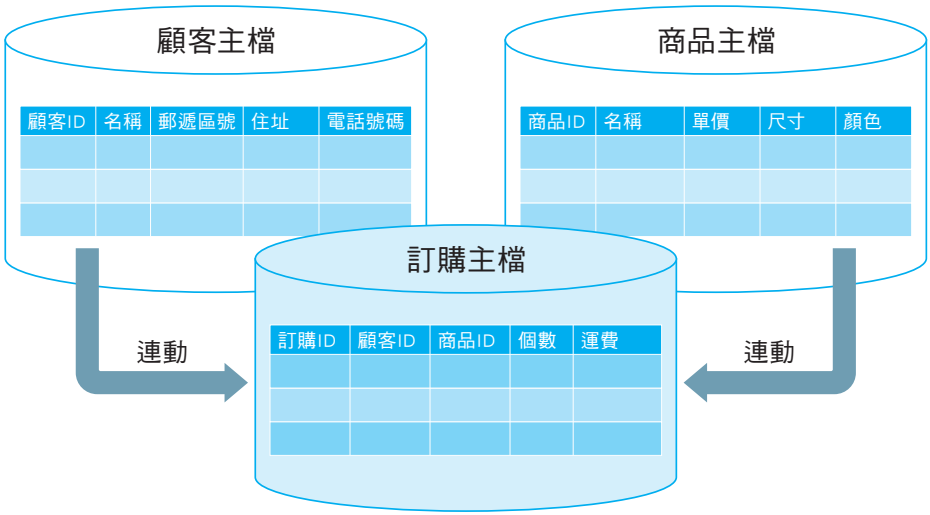
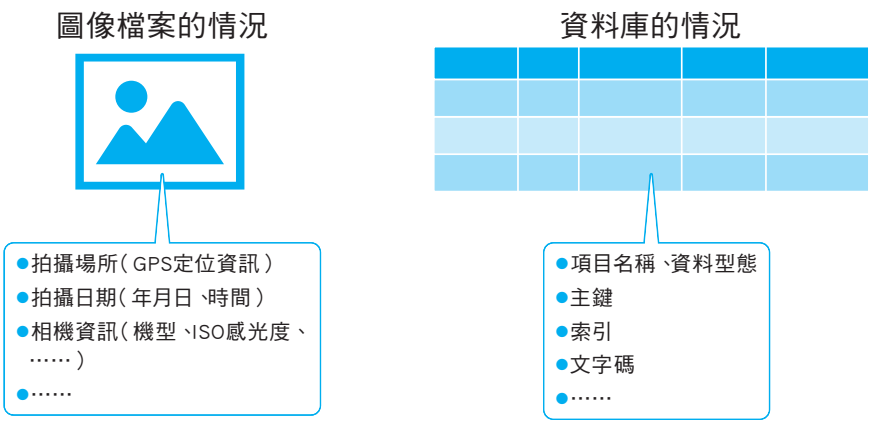


圖 1-18 元資料



Point

- 主檔資料，是當作企業資料庫基礎的重要資料
- 元資料是「描述資料本身的資料」，用來詮釋資料的內容、有效率地管理資料

» 將資料集結起來

建立分析資料的基礎建設

零散分布於各處的資料，難以結合起來分析。有鑑於此，需要建立累積資料再視情況隨時提取的系統群組——**資料基礎建設**（資料分析平台）（圖 1-19）。

除了存放資料的資料庫外，通常還要有整合管理的設備機制，如進行處理的伺服器、視覺化分析結果的程式。若選擇使用雲端服務，**即使身邊沒有高階電腦，分析人員也可利用高速的分析環境**。

以單一畫面顯示資料的狀況

逐一確認各項分析結果過於耗費時間，通常會以單一畫面統整顯示圖表、試算表等。如此一來，既不需要個別確認資料，也不用一一比較多張圖表。

這種複合畫面的**儀表板**，可按觀測者整理所需的資訊，對經營人員顯示營業額、股價等；對前線人員顯示當前的系統運轉情況、當天的作業目標等（圖 1-20）。

自動加工資料

若選擇手動從各種資訊源頭蒐集，加工並累積至資料基礎建設，不但費時也耗費精力。大型系統每天不斷增加的資料，需要自動執行這些作業來分析。

這種工作機制稱為**資料管道**（data pipeline），採取批次處理等方式，每晚加工並累積一天份的資料。

圖 1-19 累積資料的基礎建設

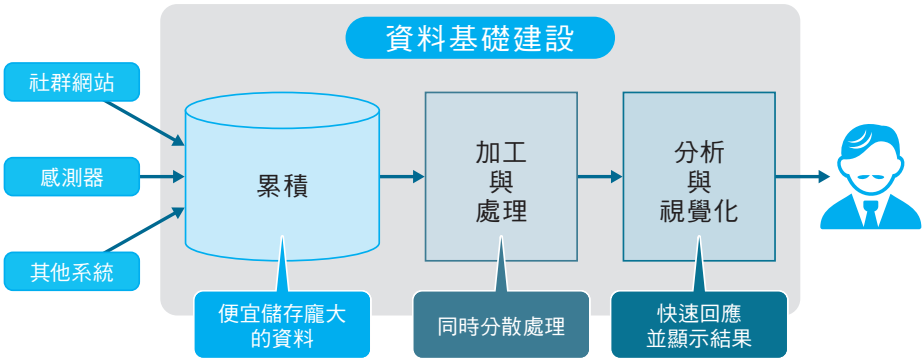
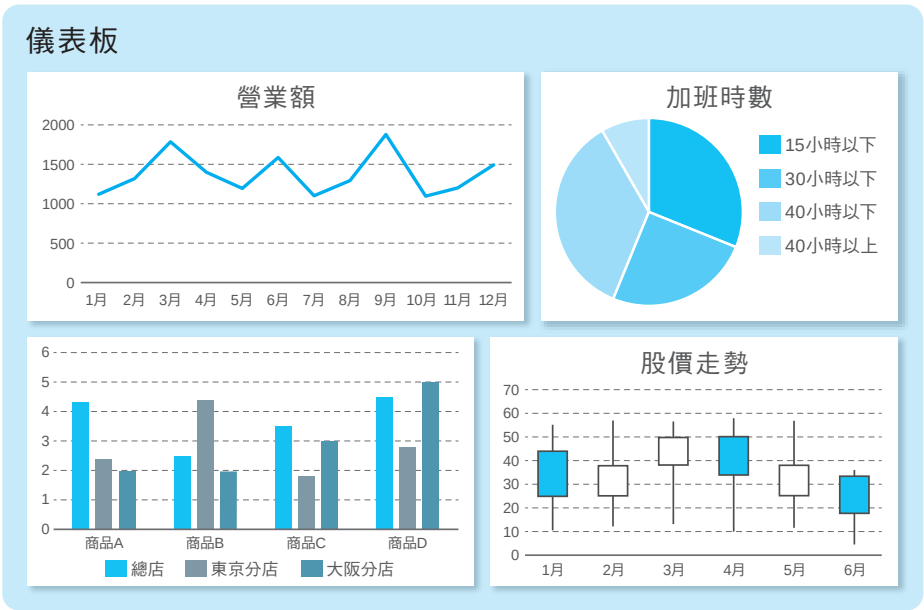


圖 1-20 儀表板的示意圖



Point

- 資料基礎建設是可隨時取出資料的系統，包含資料庫、伺服器、分析與視覺化的程式等
- 儀表板是可統整顯示圖表、試算表等的單一畫面

» 檢討高效率的處理程序

了解演算法與資料結構

演算法是解決問題的步驟、運算方式。給定某個問題的時候，縱使相同輸入輸出同樣答案，仍有好幾種推導過程（圖 1-21）。不過，只要知道一種方式，任誰皆可得到相同的結果。

在程式設計中，演算法是指電腦解決問題的步驟、程式的實作內容。當有多個相同輸入得到同樣結果的步驟時，不同的原始碼寫法、處理順序，需要的執行時間、記憶體大小也會有所不同；想辦法改進寫法、處理順序，有可能會縮短處理時間。

程式資料的存放方式也會影響演算法。例如，假設記憶體上儲存多個資料，選擇在連續空間存放資料再按位址提取，還是附加下個資料位置來依序存取資料，程式的處理方式會有所不同。程式處理時資料的存放方式，稱為**資料結構**（圖 1-22）。

演算法的處理時間

實作某演算法的程式，其執行時間跟輸入的資料量密切相關。例如，10 件資料僅需要一瞬間，1 萬件的資料處理起來肯定會耗費許多時間。

此時，需要討論輸入件數與處理時間的變化。當資料量變成 10 倍、100 倍時，處理時間是否變為 100 倍、1 萬倍，藉此判斷演算法的好壞。在分析資料的時候，**若未事前預測處理時間的話，處理起來可能耗費龐大的時間**。

圖 1-21 多種得到相同答案的方法

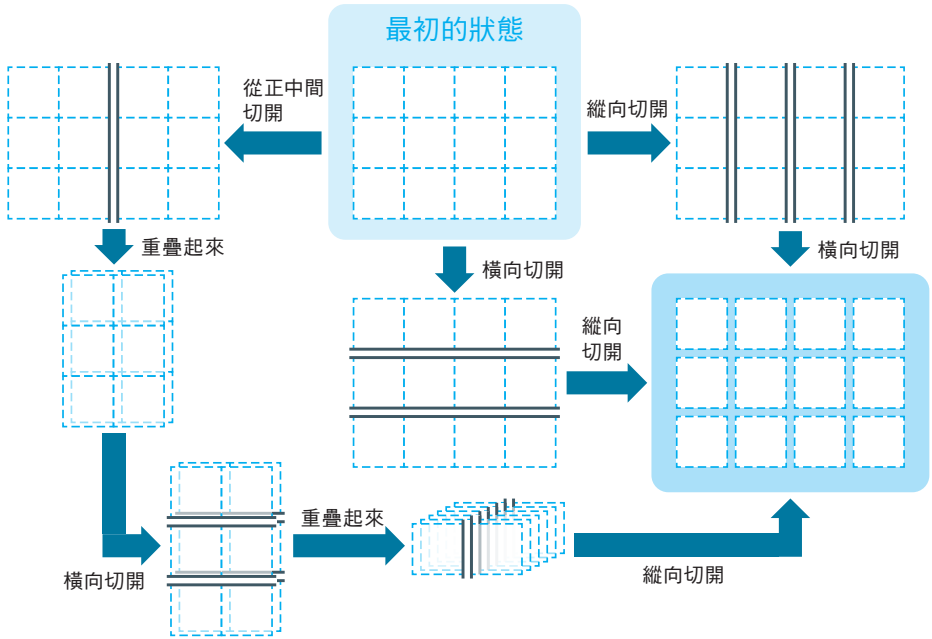
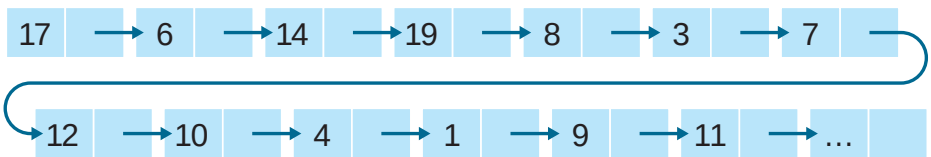


圖 1-22 資料結構的例子

陣列（在連續空間存放資料）

位址	0	1	2	3	4	5	6	7	8	9	10	11	12	…
值	17	6	14	19	8	3	7	12	10	4	1	9	11	…

鏈接串列（附加下個資料位置）



Point

- 演算法是解決問題的步驟、運算方式
- 除了演算法外，製作程式時也要檢討資料存放和資料結構

» 套用推論規則

由資料生成模型

假設在人工智慧的研究與分析巨量資料後，得到某項不錯的結果。然而，這到底僅是對給定的資料得到不錯的結果。

該分析是否也適用不同領域的資料，實際情況往往不盡理想。不過，**透過簡化內容直擊核心，可能創造通用的推論規則**。

這種由資料掌握推論核心的概念，稱為**模型**。例如，登山時會體驗到氣溫隨標高增加而下降。如圖 1-23 所示，量測後可知是往右下分布的直線。雖然該直線公式無法用於別的資料，但還有許多同樣可用直線描述的情況，這類直線關係統稱為「線性模型」。建立模型後，能夠說明資料背後的關聯性。

建立模型並反覆修正

藉由建立模型、套用觀測的資料來解釋現象的過程，稱為**建模**（modeling）。對於無法直接看出結論的資料，可透過圖表等視覺化手法，對未知的資料預測結果，來獲得有益的資訊。

此時，相同的資料會因分析人員建立、利用的模型，產生不同的解釋、使用方式。分析資料時沒有絕對正確的模型，重要的是**分析人員選擇適當的模型**。

實際上，世界各地的研究人員已開發諸多模型，分析人員常是依據面臨的課題，選擇適當的模型再做修正、微調（圖 1-24）。

圖 1-23 氣溫隨標高增加而下降的資料與關係圖

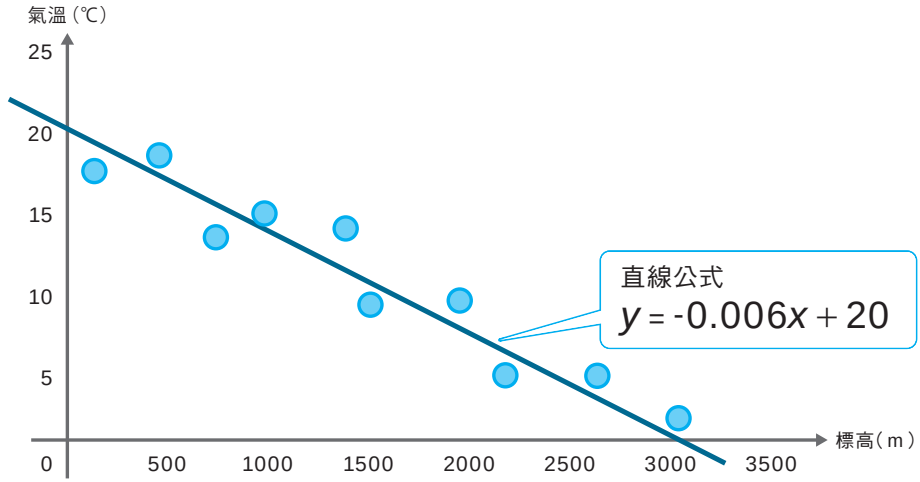
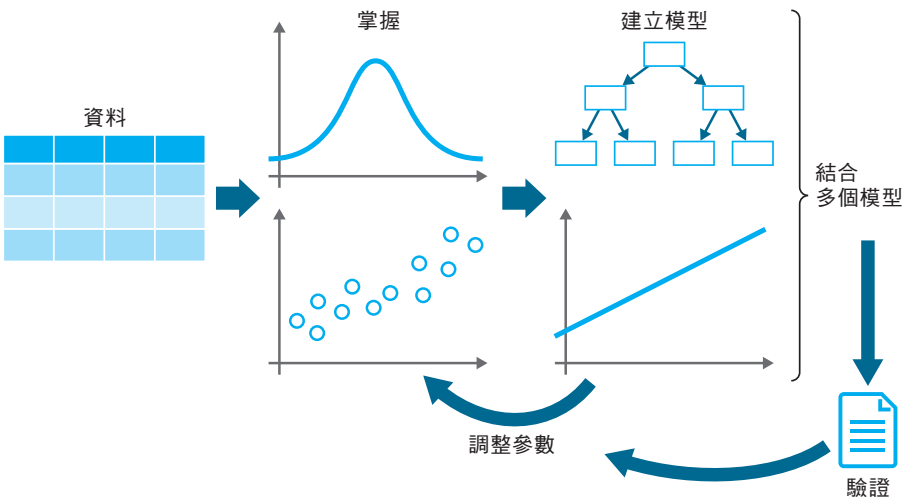


圖 1-24 建模程序



Point

- 模型可由資料掌握推論的核心，並說明資料的關聯性
- 建模是建立模型並套用資料來解釋現象