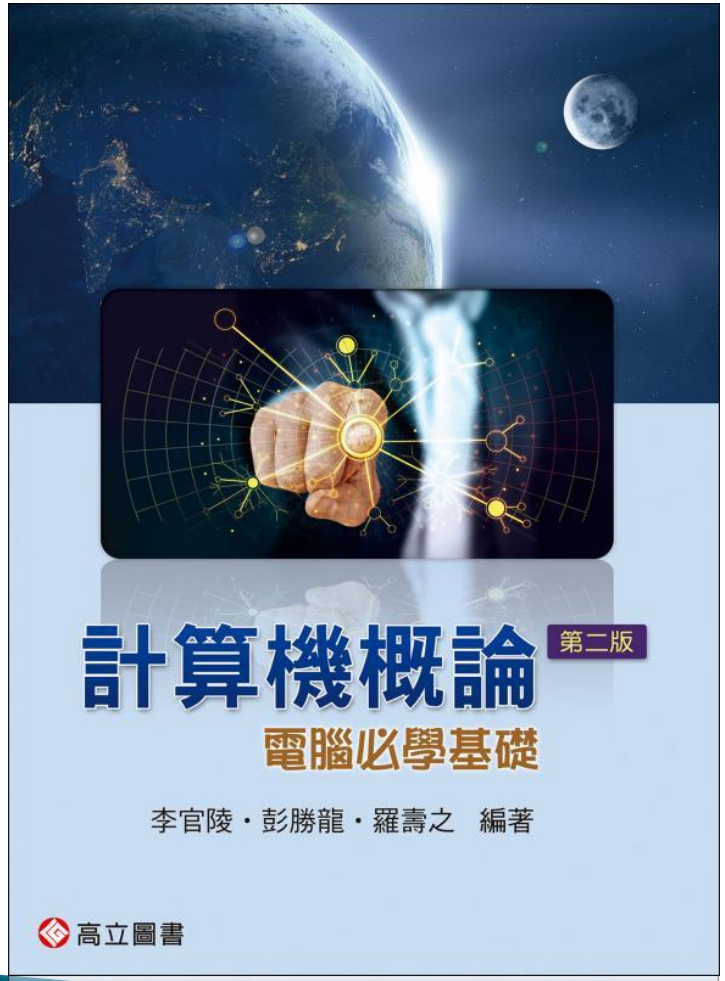


Ch12
巨量資料與雲端計算

李官陵 彭勝龍 羅壽之

- 隨著全球資料數位化的腳步、行動裝置的普及、雲端服務的進步以及社群網站的盛行，使得數位資料量以驚人的速度產生，這也讓全球進入了巨量資料 (big data) 的時代。
- 這些資料量的規模巨大到無法透過目前主流的軟體工具，在合理的時間內達到管理、搜尋、處理與分析等目的的資料，我們稱之為巨量資料。

圖 12.1　網際網路上的一分鐘（圖片來源：Intel 官網）

圖 12.2 Google 的搜尋推薦

圖 12.3 博客來上「姊姊的守護者」相關推薦

巨量資料

- 巨量資料具有三項主要特性：一為資料量 (amount of data) 非常大，二為資料增加的速度 (speed of data) 非常快，三為資料形式的多樣性 (range of data type)。
- 這些大量資料背後蘊藏著豐富的價值，若能適當地挖掘並好好的運用，可以帶給使用者更好的資訊服務，亦可為服務提供者帶來巨大的商機。

大量的資料

- 龐大的資料量是巨量資料最基本的特性。而在這裡所說的龐大的量，不是分散在數台電腦上就可以儲存下來的量，而是需要成百上千台的電腦才能存放之。當資料分散在數百台甚至數千台機器上時，如何有效率的管理與儲存這些資料就是一個重要的問題。

資料的快速增加

- 增加速度非常快速的資料大幅地增加了資料處理與分析的困難度。
- 如果我們處理資料的速度無法趕上資料產生的速度，則對於某些有即時需求的應用而言(如網路攻擊偵測、災害偵測等)，分析的結果就無法立即反應當下的狀況。

資料形式的多樣性

- 社群網路上被分享或傳送的文字、圖片、影像等等，又或者智慧型手機裡裝置的各式各樣感測器，可用來收集使用者的位置、移動速度、環境亮度等資料，都是具有多樣性的巨量資料來源，這些資料沒有固定的格式，因此被稱為非結構(unstructured) 性資料。非結構性資料是巨量資料的重要特性。

資料探勘

- 資料探勘即為在大量的資料中發現知識的過程。
- 因為資料探勘就是在大量的資料中挖掘出知識的過程，所以我們也會稱其為知識發現 (Knowledge Discovery in Database, KDD)。

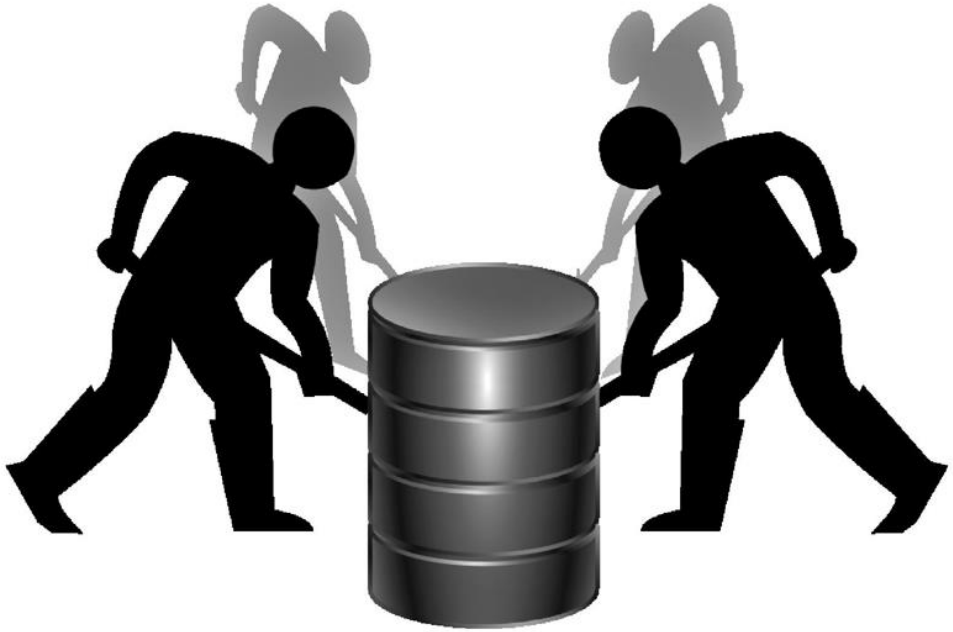


圖 12.4　資料探勘示意圖（圖片來源：維基共享資源）

資料探勘 (續)

- 知識發現的過程可分為下面所列之五大步驟：

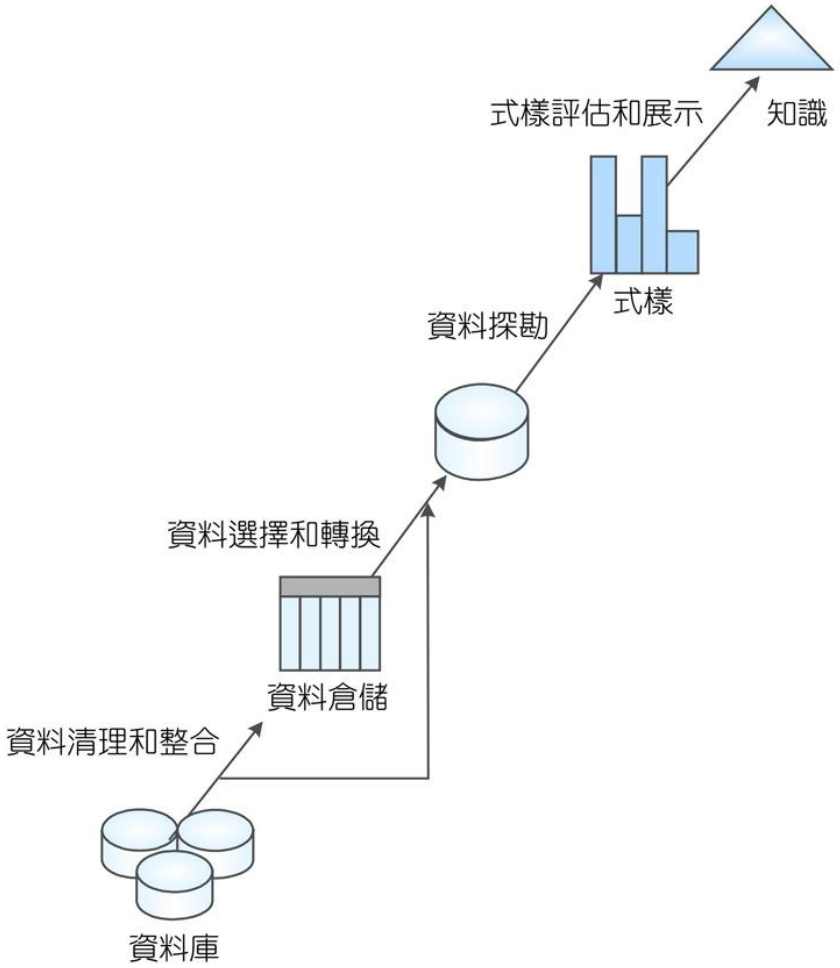


圖 12.5　知識發現流程圖

資料探勘 (續)

1. 資料清理及整合 (data cleaning and integration)：消除雜訊或不一致的資料，再將多種資料來源、型態、格式整合在一起，以供後續之探勘演算法使用。
2. 資料選擇與轉換 (data selection and transformation)：包含資料的淨化 (clean)、格式的轉換以及資料正規化 (normalization)，將資料轉換成適合探勘演算法的格式。
3. 資料探勘 (data mining)：根據應用以及目的，選擇適合的資料探勘演算法。將隱藏於上一步驟所完成的資料中，有用的資訊挖掘出來。這個過程對於資料探勘的應用成功與否有決定性的影響。

資料探勘 (續)

4. 分析及評估 (analysis and evaluation)：分析及評估所挖掘出來的知識是否真的有價值，以便將資料探勘的結果去蕪存菁，過濾掉沒有用的資訊，將有價值的知識提供給使用者。
5. 知識呈現 (knowledge presentation)：將複雜的資料探勘結果做一個淺顯易懂的呈現 (如圖形化介面)，使得這些有趣的知識可以容易的被了解使用 。

資料關聯分析

- 資料關聯 (data association) 主要的目的是找出項目與項目間的關聯性，
- 「麵包 => 牛奶 [10%, 70%]」為一條關聯規則，代表的意思為：10% 的顧客會買麵包和牛奶，而在買麵包的顧客中，有 70% 的比例會買牛奶，這就是一個典型的關聯規則。
- 在本例中，10% 稱為 {麵包，牛奶} 這項目集的支持度 (support)，代表的是包含這個項目集的交易個數佔全部交易總數的比例。而 70% 則代表著這關聯規則的信心度 (confidence)，也就是在已知買麵包的交易當中，會一起購買牛奶的比例。

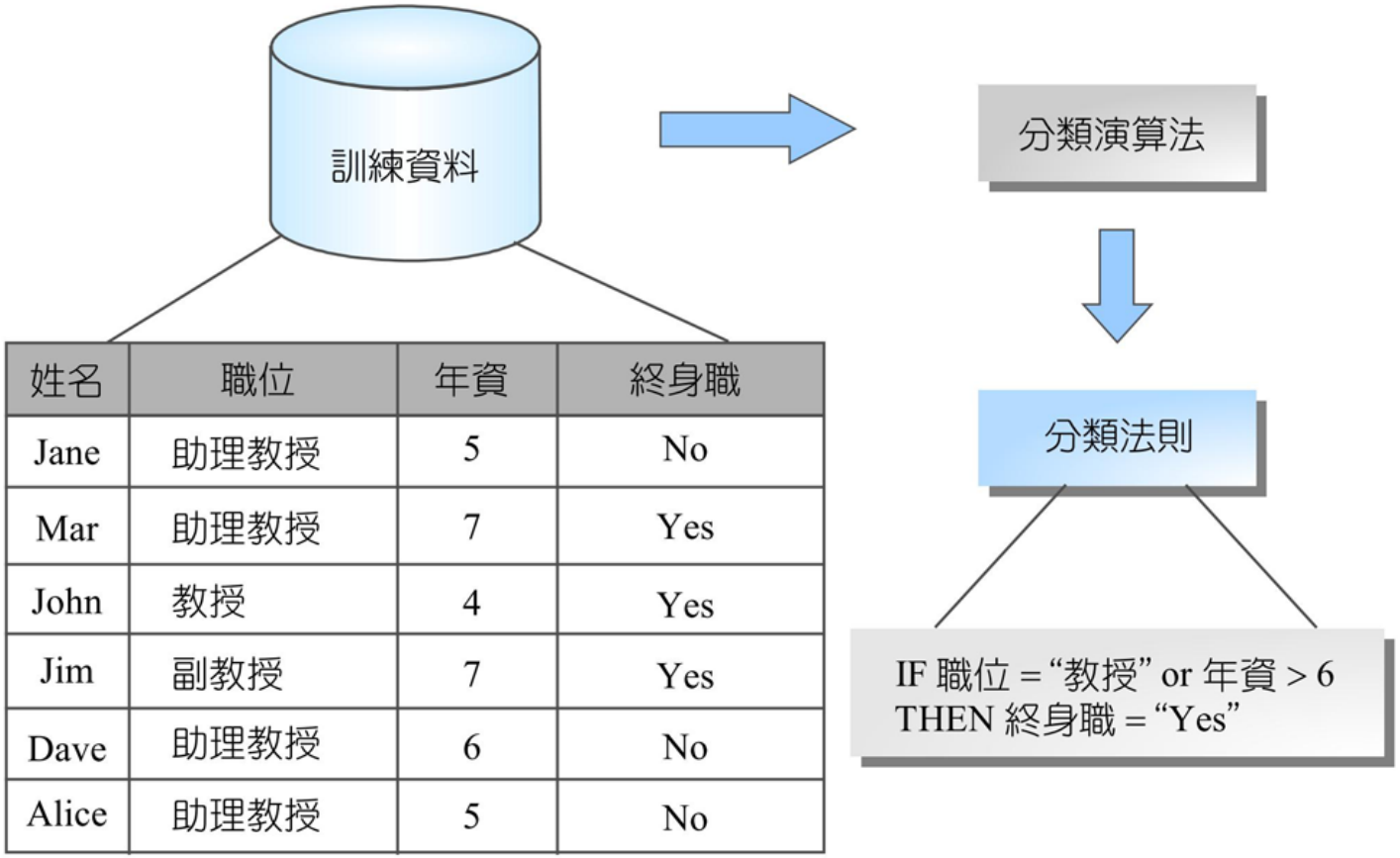
資料分類

- 資料分類 (data classification) 是根據已知資料的相關屬性 (attributes) 以及該資料所屬之類別 (class label)，建立資料的分類模型 (classification model)，然後透過該模型來預測新進資料的類別歸屬。
- 分類基本上可以分為三個階段：第一階段為建立分類模型，也就是利用現有的資料將資料的分類規則找出來。第二階段為評估分類模型的準確度，在一開始建立分類模型時會將現有的資料分成二組：一組為訓練樣本 (training dataset)，另一組則為測試樣本 (testing dataset)。一開始我們在建立模型時只有利用訓練樣本，而在第二階段就利用測試樣本來評估分類模型的準確性。第三階段為使用模型，在評估完分類模型的效能 (如準確度) 後，若是這模型的效能是可被接受的，就可以開始進入使用階段。

資料分類 (續)

表 12.1 分類資料

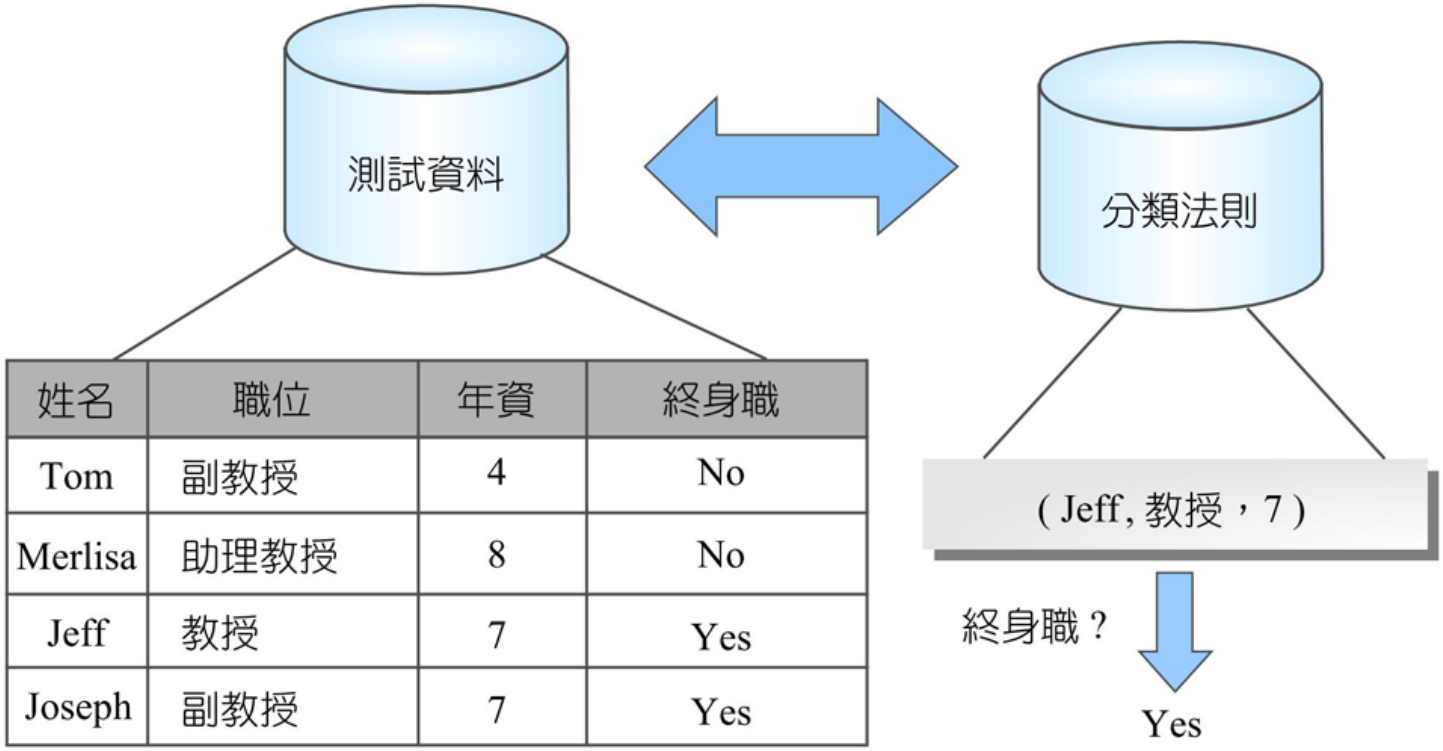
資料分類 (續)



姓名	職位	年資	終身職
Jane	助理教授	5	No
Mar	助理教授	7	Yes
John	教授	4	Yes
Jim	副教授	7	Yes
Dave	助理教授	6	No
Alice	助理教授	5	No

圖 12.6　資料分類模型建置

資料分類 (續)



姓名	職位	年資	終身職
Tom	副教授	4	No
Merlisa	助理教授	8	No
Jeff	教授	7	Yes
Joseph	副教授	7	Yes

圖 12.7　模型準確度測試

資料分群

- 資料分群 (data clustering) 也稱之為群集分析。
- 群集分析的主要目的是分析資料彼此間的相似度，並根據資料間的相似度將資料分成數個群集 (cluster)，使得同一群集內的資料彼此間有較高的相似度 (high intra cluster similarity)，而不同群集的資料彼此間的相似度就較低 (low inter cluster similarity)。一般而言，群集分析有兩個用途，一為用於了解資料的分布特性，另一則為當作其他工具的前製處理工作，用以減少資料的量，增加效能。

資料分群 (續)

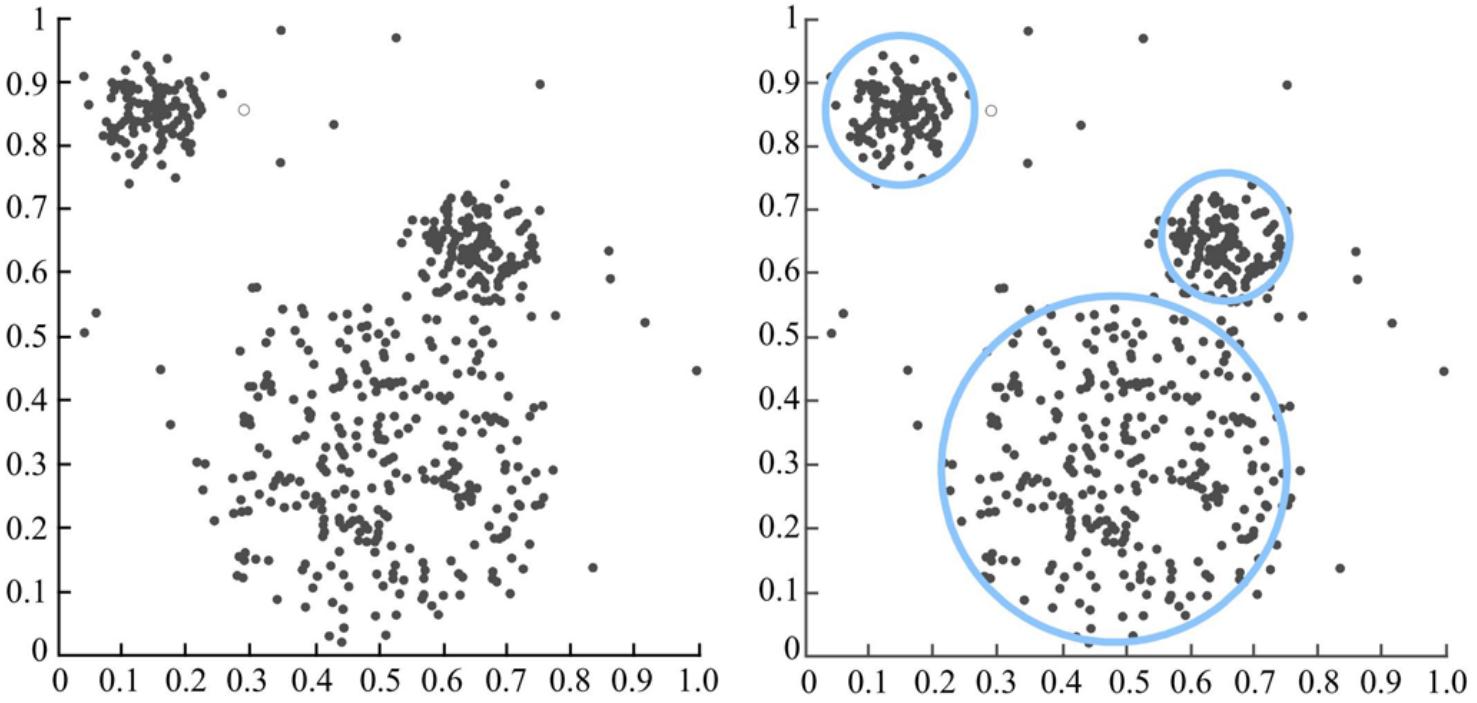


圖 12.8　資料群集示意圖

雲端計算

- 網際網路的興起帶動資訊相關產業的盛行，網路的相關設備，如網路介面卡、網路硬碟與伺服器主機等，需求大增。全球資訊網崛起，更帶動網路線上模式的商機。
- 此時，提供網路服務的大型公司，如谷歌、雅虎與亞馬遜等也陸續出現，這些公司為了服務全世界廣大的用戶，到處建置網路伺服器、購置大量硬碟。
- 當這些建置的設備達到某個規模時，所能提供的計算能力與儲存空間已遠遠超過本身的需求。於是，開始出現租用伺服器與硬碟空間、甚至販售資訊的新商業模式，這種新的網路服務觀念，造就熱門的雲端計算 (cloud computing)。

雲端計算 (續)

- 一般公司若要提供某種網路服務，如線上訂房與網購等，均採用主從架構的模式。
- 對小公司來說，建置與維護都是不小的成本。如果有一家大型的網路公司提供完善的網路環境與管理，客戶可以依據需求租用不同等級的伺服器、不同容量的硬碟空間與不同頻寬速度的網路，就可以大幅降低投入的成本，也不必擔心資料遺失或系統故障的問題。這種模式就是雲端計算希望提供的多元網路服務架構。
- 對個人而言，雲端計算也可以提供線上應用程式的執行。例如編輯與排版文件的服務，個人端的電腦不用安裝昂貴的軟體，只要有網路連線的功能，就可以連至雲端上的伺服器，啟動租用的軟體以編輯個人的文件。資料的儲存也可以放置雲端上租用的硬碟空間。

雲端計算 (續)

- 雲端計算的服務類型有三種

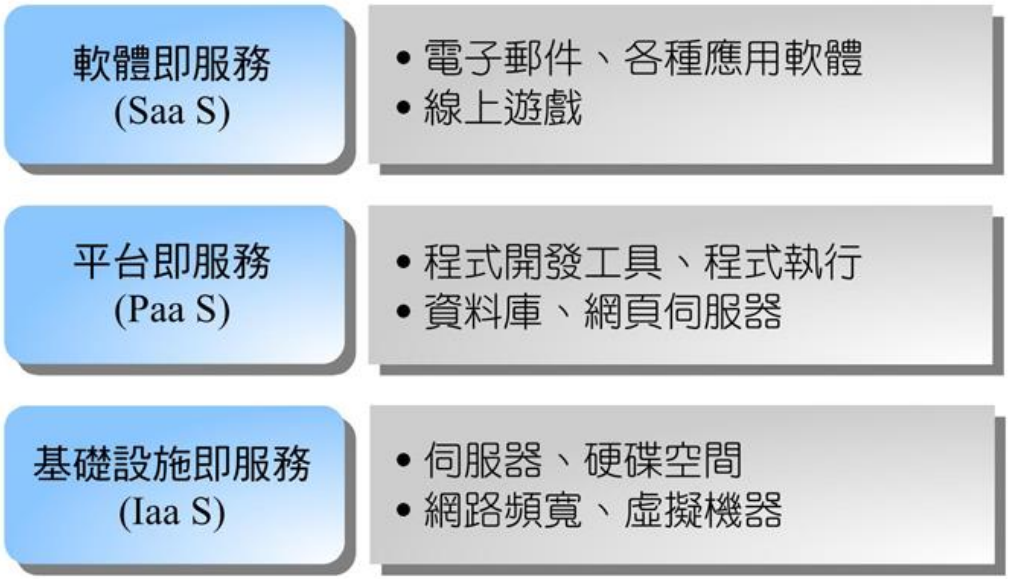


圖 12.9 雲端計算服務類型

雲端計算 (續)

1. 軟體即服務 (Software as a Service, SaaS)

   這是一種租用應用軟體程式的商業模式。服務供應商將各種應用軟體集中安裝於雲端伺服器，客戶端只要利用網路連接進來，就可以執行這些軟體，處理自己的資料。

2. 平台即服務 (Platform as a Service, PaaS)

   這是一種租用計算平台的商業模式，服務供應商在雲端上建置好電腦作業系統的執行環境、提供應用程式開發的各種工具、甚至提供資料庫供大量資料的管理，以及現成的網頁伺服器等。客戶端只要開發並上傳自己的應用程式，後續的程式執行與資料的儲存管理，全部交給雲端。

雲端計算 (續)

3. 基礎設施即服務 (Infrastructure as a Service, IaaS)

這是一種租用雲端設備的商業模式，可租用的設備需求包括主機個數、記憶體容量、硬碟空間與網路頻寬等。客戶自行管理租用的電腦資源，如安裝作業系統與應用軟體等。

雲端計算 (續)

- 雲端服務面對的是全球廣大的客戶，每個人的需求不盡相同。
- 雲端服務的供應商如何同時滿足這些不同的需求呢？答案是使用一種稱為虛擬化 (virtualization) 的技術，在一台伺服器主機上利用軟體的方式模擬成多台的電腦。這種虛擬機器 (Virtual Machine, VM) 可透過軟體設定的方式，成為不同規格的電腦，滿足用戶不同的需求。
- 布建雲端計算服務的方式可依據使用的限定性，區分私有雲 (private cloud) 與公眾雲 (public cloud)。私有雲的限定性較高，如公司或校園內部使用；公眾雲則開放一般用戶使用，但部分加值的服務需要付費。

穿戴式計算

- 智慧型手機可視為整合行動手機 (mobile phone) 與PDA 功能的設備，具有作業系統的管理平台，可以安裝各種應用軟體。現在更多的智慧型攜帶裝置，如智慧手錶與智慧手環亦相繼出現，這些方便攜帶又提供資料處理能力的設備，統稱穿戴式計算 (wearable computing)。
- 智慧型手機的平台系統主要有三大陣營：iPhone、Android 以及Windows Phone。谷歌發行的 Android 作業系統，目前擁有最多的廠商加入，市佔率也是最高。 執行於智慧型手機的應用軟體稱為 APP，一般是由作業系統廠商提供平台，讓程式開發人員於其上開發與銷售 APP。

穿戴式計算 (續)

- 智慧型手機具備多項軟硬體功能 (如圖 12.10)，通訊部分可以使用 3G/4G 等行動通訊網路，網際網路的連線也可以使用 Wi-Fi 無線網路。耳機與喇叭等個人周邊設備則可以使用藍芽通訊與手機配對。而近距離通訊 (Near-Field Communication, NFC) 技術則讓手機可以讀取晶片卡。
- 基本的相機功能讓手機可以拍照與攝影，透過相機的鏡頭在觀看實物時，甚至可以套用擴充實境 (Augmented Reality, AR) 的技術。手機內建的衛星定位系統 (Global Positioning System, GPS) 接收器，讓用戶隨時掌握目前的位置，並可結合提供位置服務的 APP，查詢附近的街道圖，以及導引至目的地的路徑規劃等。
- 智慧型手機的另一個特色是內建多種感測器 (sensors)，如加速度感測器、方向感測器、陀螺儀感測器、光線感測器及接近感測器等。

穿戴式計算 (續)

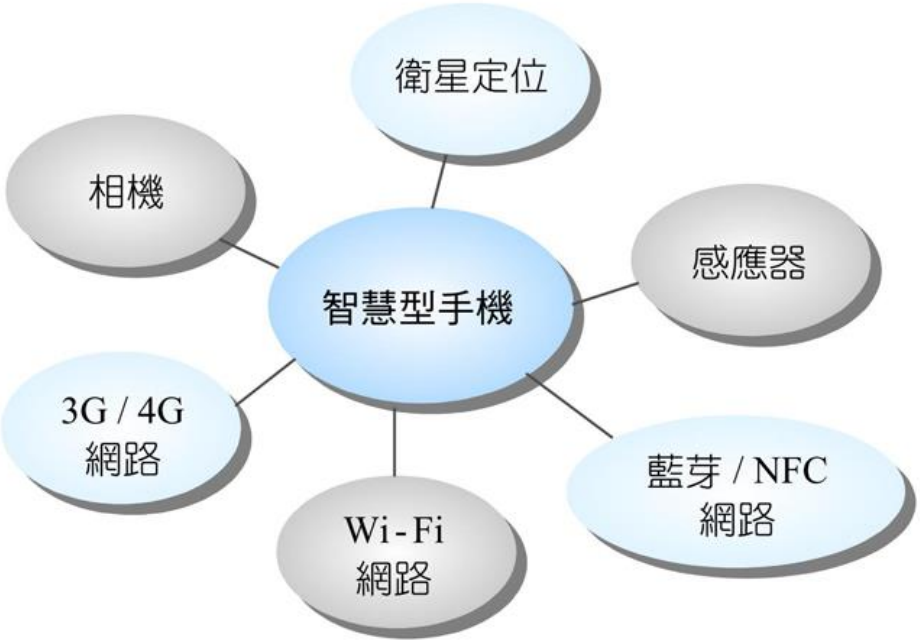


圖 12.10　具備多功能的智慧型手機

穿戴式計算 (續)

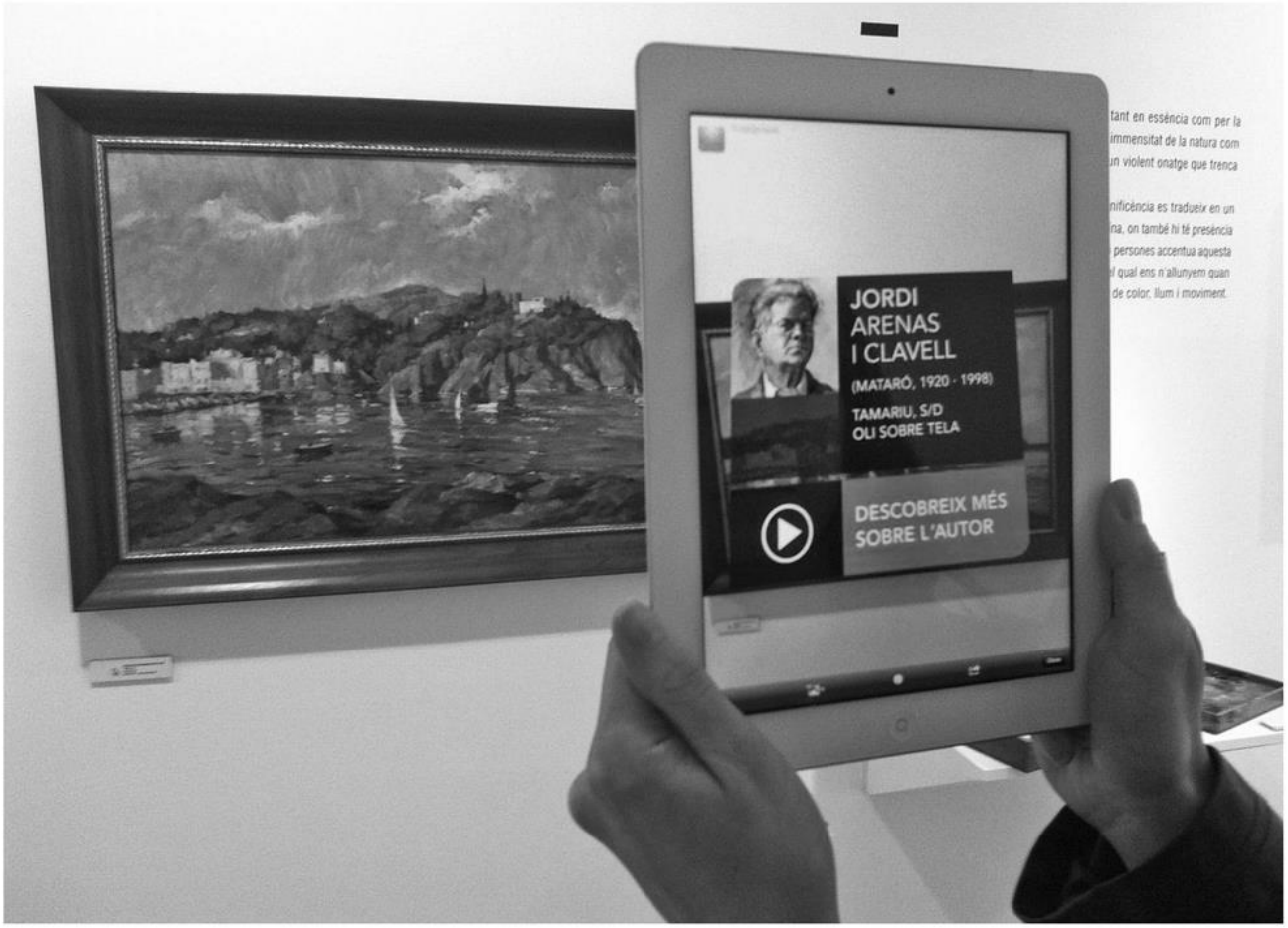


圖 12.11　AR 應用（圖片來源：維基共享資源）

穿戴式計算(續)

- 除了以上與手機使用者動作相關的感測器之外，亦可作為身體健康狀況測量的感測器。

圖 12.12 兼具運動與健康監測的智慧手環與智慧手錶（圖片來源：林士凱）

穿戴式計算 (續)

- 感測器的種類與應用相當多，可以測量個人健康狀況外，也可以測量環境如溫度與溼度的變化。如果感測器也兼具通訊的能力，能相互傳遞訊息，甚至連至網際網路，就可以在居家或戶外環境中大量布置，形成一個智慧空間 (smart space)。
- 當感測器越做越小、價格越來越便宜時，所有的物品都可以嵌入感測器，讓物品間可以交換訊息。在網際網路上查看這些物品，形成一種物聯網 (Internet of Things, IoT)。

穿戴式計算(續)

- 想像未來生活的一種情境，家裡的冰箱自動得知牛奶快過期了，主動發送手機訊息給主人，並提示是否自動進行網路購買的處理；家裡的床墊感測主人已經起床，聯繫廚房的咖啡機自動沖泡一杯咖啡等。科技的進步，讓未來的生活似乎更加便利與舒適。