天下文化
Science Culture

大數據

「數位革命」之後，「資料革命」登場：
巨量資料掀起生活、工作和思考方式的全面革新

BIG DATA

A Revolution That Will Transform How We Live, Work, and Think

by Viktor Mayer-Schönberger and Kenneth Cukier

麥爾荀伯格、庫基耶_著
林俊宏_譯

科學文化 156

大數據

Big Data

A Revolution That Will Transform How We Live, Work, and Think

國家圖書館出版品預行編目 (CIP) 資料

大數據 / 麥爾荀伯格 (Viktor Mayer-Schönberger), 庫基耶 (Kennech Cukier) 著 ; 林俊宏譯 . -- 第一版 .
-- 臺北市 : 天下遠見 , 2013.05
面 ; 公分 . -- (科學文化 ; 156)
譯自 : Big data : a revolution that will transform how we live, work, and think
ISBN 978-986-320-191-5(精裝)

1. 資訊社會 2. 網路社會 3. 社會變遷

541.415 102008422

原著 — 麥爾荀伯格、庫基耶
譯者 — 林俊宏
策劃群 — 林和 (總策劃)、牟中原、李國偉、周成功
科學叢書總監暨責任編輯 — 林榮崧
封面設計 — 張議文
內頁設計 — 江儀玲

出版者 — 天下遠見出版股份有限公司
創辦人 — 高希均、王力行
遠見・天下文化・事業群 董事長 — 高希均
事業群發行人／ CEO — 王力行
出版事業部總編輯 — 許耀雲
版權部經理 — 張紫蘭
法律顧問 — 理律法律事務所陳長文律師
著作權顧問 — 魏啟翔律師
地址 — 台北市 104 松江路 93 巷 1 號 2 樓

讀者服務專線 —（02）2662-0012 | 傳真 —（02）2662-0007；2662-0009
電子信箱 — cwpc@cwgv.com.tw
直接郵撥帳號 — 1326703-6 號 天下遠見出版股份有限公司

排版廠 — 極翔電腦排版有限公司
製版廠 — 東豪印刷事業有限公司
印刷廠 — 崇寶彩藝印刷股份有限公司
裝訂廠 — 源太裝訂實業有限公司
登記證 — 局版台業字第 2517 號
總經銷 — 大和書報圖書股份有限公司 電話／（02）8990-2588
出版日期 — 2013 年 5 月 30 日第一版第 1 次印行

ISBN 978-986-320-191-5（精裝）（英文版 ISBN 978-0-544-00269-2）

定價 — 360 元
書號 — CS156
天下文化書坊 — www.bookzone.com.tw

大數據 BIG DATA

A Revolution That Will Transform How We Live, Work, and Think

大數據

「數位革命」之後，「資料革命」登場：
巨量資料掀起生活、工作和思考方式的全面革新

BIG DATA

A Revolution That Will Transform How We Live, Work, and Think

by Viktor Mayer-Schönberger and Kenneth Cukier

林俊宏　譯

每十年，總是有極少數的書，能改變你看待一切的方式。
《大數據》正是這樣的書。
整個社會已經開始估算，巨量資料將帶來的變化。
這本書是一個非常重要的起點。

——萊斯格（Lawrence Lessig），哈佛法學院網路智慧財產權教授

《大數據》開闢了新境界，
告訴我們巨量資料如何從根本上，轉變我們對世界的基本理解……
這本書清楚說明了，企業如何釋放潛藏的價值，決策者如何因應新局，
以及每個人的認知模式需要如何改變。

——伊藤穰一（Joi Ito），麻省理工學院媒體實驗室主任

任何人如果想要保持領先地位，確定未來的商業發展趨勢，
都必須閱讀《大數據》。

——貝尼奧夫（Marc Benioff），salesforce.com 董事長兼執行長

《大數據》很樂觀而務實的看待巨量資料革命——
你只要伸頭看看周遭發生的大變化，就會明白這場革命已然開始了，
更大的變化即將衝擊而來。

——多克托羅（Cory Doctorow），boingboing.com

我們敢肯定的是，
《大數據》將是在討論這方面的未來時，一言九鼎的文本。

——富比士網站

有太多書籍在頌揚資訊社會的技術奇蹟，
但是唯有這本《大數據》對資訊的本質，進行了原創的分析。

——《柯克斯書評》（*Kirkus Reviews*）

這本書充滿了偉大的見解、駕馭資訊的新途徑，
並且對於未來趨勢，提供了很有說服力的願景，
這是任何使用巨量資料的人、或受到巨量資料影響的人，
都不可或缺的讀物。

——喬納斯（Jeff Jonas），IBM 首席科學家

這本出色耀眼的書，撥開了圍繞在巨量資料周邊的迷霧。
不論你從事的是商業、資訊科技、公共行政、教育、醫療，
或者你只是單純對未來趨勢感到好奇，都必須閱讀這本《大數據》。

——布朗（John Seely Brown），全錄帕羅奧圖研究中心主任

正如水是濕滑的，然而單個水分子卻不是；
巨量資料也能顯現個別資料無法揭露的訊息。
作者向我們展示了龐大、複雜、凌亂的資料，若是集合起來，
竟能用來預測購物行為、流感爆發……的一切模式，真是令人驚駭。

——薛奇（Clay Shirky），社會媒體理論家

作者讓「巨量資料」這個名詞的內涵變得非常清晰，
這個名詞的重要性已遠遠超過矽谷的其他流行語彙……
沒有哪一本書能夠提供這樣的可讀性和平衡報導，
告訴我們繼續迷戀數據和資料的諸多好處及缺點。

——《華爾街日報》

「巨量資料」是企業管理階層、技術官僚的流行語之一，
如果你想知道他們都在談論些什麼，那麼《大數據》正是為你而寫的。
這本書深入淺出、而且很有意思的切入這個大題目……

——《波士頓環球報》

大數據 BIG DATA

A Revolution That Will Transform How We Live, Work, and Think


NOW

第1章

現在
該讓巨量資料說話了

2009年又冒出了一種新的流感病毒，稱為H1N1。這種新菌株結合了禽流感和豬流感病毒，迅速蔓延。短短幾星期內，全球的公共衛生機構都憂心忡忡，擔心即將爆發流感大流行。有些人發出警訊，認為這次爆發可能與1918年的西班牙流感不相上下，當時感染人數達到五億人，最後奪走數千萬人的性命。雪上加霜的是，面對流感可能爆發，卻還沒有能派上用場的疫苗，公共衛生當局唯一能努力的，就是減緩流感蔓延的速度。為了達到這項目的，必須先知道當前流行感染的範圍及程度。

在美國，疾病管制局（CDC）要求醫生一碰到新流感病例，就必須立刻通報。即使如此，通報的速度仍總是慢了病毒一步，大約是慢上一到兩星期。畢竟，民眾覺得身體不舒服之後，通常還是會過個幾天才就醫，而層層通報回到疾管局也需要時間，更別提疾管局要每星期才整理一次通報來的資料。但是面對迅速蔓延的疫情，拖個兩星期簡直就像是拖了一個世紀，會在最關鍵的時刻，讓公共衛生當局完全無法掌握真實情況。

巨量資料初試啼聲

說巧不巧，就在H1N1躍上新聞頭條的幾星期前，網路巨擘谷歌（Google）旗下的幾位工程師，在著名的《自然》科學期刊發表了一篇重要的論文，當時並未引起一般人的注意，只在衛生當局和電腦科學圈裡引起討論。該篇論文解釋了谷歌能如何「預測」美國在冬天即將爆發流感，甚至還能精準定位到是哪些州。谷歌

的祕訣，就是看看民眾在網路上搜尋些什麼。由於谷歌每天會接收到超過三十億筆的搜尋，而且會把它們全部儲存起來，那就會有大量的資料得以運用。

谷歌先挑出美國人最常使用的前五千萬個搜尋字眼，再與美國疾病管制局在2003年到2008年之間的流感傳播資料，加以比對。谷歌的想法，是想靠著民眾在網路上搜尋什麼關鍵詞，找出那些感染了流感的人。雖然也曾有人就網路搜尋字眼做過類似的努力，但是從來沒人能像谷歌一樣掌握巨量資料（big data，直譯為大數據），並具備強大的處理能力和在統計上的專業技能。

雖然谷歌已經猜到，民眾的搜尋字眼可能與流感有關聯，像是「止咳退燒」，但有沒有因果關係並不是真正的重點，他們設計的系統也不是從這個角度出發。谷歌這套系統真正做的，是要針對搜尋字眼的搜尋頻率，找出和流感傳播的時間、地區，有沒有統計上的相關性（correlation）。他們總共用上了高達四億五千萬種不同的數學模型，測試各種搜尋字眼，再與疾管局在2007年與2008年的實際流感病例加以比較。他們可挖到寶了！這套軟體找出了一組共四十五個搜尋字眼，放進數學模型之後，預測結果會與官方公布的全美真實資料十分符合，有強烈的相關性。於是，他們就像疾管局一樣能夠掌握流感疫情，但可不是一、兩星期之後的事，而是幾近即時同步的掌握！

因此，在2009年發生H1N1危機的時候，比起政府手中的資料（以及無可避免的通報延遲），谷歌系統能提供更有用、更即時的資訊。公衛當局有了這種寶貴的資訊，控制疫情如虎添翼。

最驚人的是，谷歌的這套方法並不需要去採集檢體、也不用登門造訪各家醫院診所，而只是好好利用了巨量資料，也就是用全新的方式來使用資訊，以取得實用且價值非凡的見解、商機或服務。有了谷歌這套系統，下次爆發流感的時候，全球就有了更佳的工具能夠加以預測，並防止疫情蔓延。

巨量資料功能強大，可以讓許多領域改頭換面，公共衛生領域不過是其中之一，而商業領域也正在經歷這個過程。例如買飛機票就是一個很好的例子。

2003年，伊茲奧尼（Oren Etzioni）打算從西雅圖飛往洛杉磯參加弟弟的婚禮。早在幾個月前，他就已經上網買了機票，一心認為愈早預訂，票價就愈划算。但是在航程中，他出於好奇，問了坐在隔壁的乘客票價以及購票時間，結果那個人明明是最近才買的，票價卻是便宜得多。一氣之下，伊茲奧尼一個又一個的問下去，發現大部分人的票價都比他的更便宜。

對於大多數人來說，等到收回托盤、豎直椅背、準備下機的時候，這種覺得被敲竹槓的火氣，也差不多消了。但伊茲奧尼身為美國頂尖的資訊科學家，可沒這麼好打發。在他看來，整個世界就是由一連串關於巨量資料的問題構成的，而這正是他拿手的領域。追溯到1986年，伊茲奧尼可是哈佛大學第一位主修資訊科學的畢業生，之後進入華盛頓大學任教；而且早在巨量資料這個詞出現之前，他就已經開了數家處理巨量資料的公司。例如，他曾協助打造了最早期的網路搜尋引擎之一、於1994年推出的MetaCrawler，不久便由當時的網路巨擘InfoSpace公司買下。另

外，他也共同創立了史上第一個大型比價購物網站Netbot，後來出售給Excite公司。至於他的另一間公司ClearForest，則是處理如何從文件中取得語義資訊，後來由路透社收購。

客機著陸之後，伊茲奧尼已經下定決心，要讓人能夠知道自己在網上看到的票價，究竟是撿到便宜還是被人坑了。如果把飛機機位看成商品，同一航班的座位基本上也沒什麼不同，但票價卻是天差地別。這裡有許多因素，只有航空公司自己才曉得。

伊茲奧尼認為，這種系統並不需要真的去解出票價背後千絲萬縷的糾纏因素，只要能預測出未來票價是漲是跌就夠了。這其實不困難，只要先取得特定航線售出的所有票價資訊，再與出發前的天數做比較即可。

如果平均票價呈現下跌趨勢，買票這件事當然就可慢慢來。如果平均價格呈現上漲趨勢，系統則會建議馬上以目前顯示的價格購票。換句話說，當初伊茲奧尼是在一萬公尺高空中，逐一詢問其他乘客的票價，而現在這個系統就是個加強版。雖然說這絕對還是一個資訊工程的龐大問題，但與過去相同，這對他而言仍然能夠迎刃而解。於是，他動工了。

伊茲奧尼花了四十一天，從某個旅遊網站取得超過一萬兩千筆票價資料，做為樣本，並建立一個預測模型，讓模擬的乘客都省下了大筆鈔票。這個模型並不懂「為何如此」（why），只知道「正是如此」（what）。換言之，模型完全不知道各種影響票價的因素，像是未售出的機位數、淡旺季、或是星期幾的機票較便宜之類；模型所做的預測，都是基於手中確實的資訊，也就是從其他

航班所蒐集到的相關資料。

伊茲奧尼思思念念的，就是「要買還是不買」的問題——像極了莎翁名劇《哈姆雷特》的經典獨白：「生存還是毀滅，這是一個值得考慮的問題。」正因如此，伊茲奧尼把這個研究計畫命名為「哈姆雷特」。

原本的小小研究計畫，後來發展成投入大量資金的創業計畫「Farecast」，藉著預測機票票價可能上漲或是下跌，Farecast就能讓消費者知道是否該立刻點選「購買」鍵。在過去，消費者從來不可能得知這些資訊。Farecast堅持一切應該透明，所以甚至還會對自己的預測加上可信度評分，提供給使用者參考。

預測系統要有效，就必須擁有大量的數據資料。為了提升效能，伊茲奧尼從航空業的一個航班預訂資料庫下手。資料庫存有全年美國商業航空公司各航班、各座位的資料，能做為系統預測的基礎。現在，Farecast手中大約有近兩千億筆票價紀錄，用以做出預測。如此一來，消費者就能省下大筆的金錢。

伊茲奧尼有一頭黃褐色的頭髮，露齒微笑、一臉天真，看起來實在不像是會讓航空業損失數百萬美元潛在收入的人。但事實上，他的目標還不止於此。

到了2008年，伊茲奧尼打算將這套辦法再應用到其他商品，像是飯店客房、音樂會門票、二手車，只要是產品差異性小、價格變化大、而且有大量數據資料的商品，都能適用。但他還沒來得及讓計畫成真，微軟就已經找上門，用大約一億一千萬美元買下Farecast，結合到Bing搜尋引擎之中。到了2012年，該系統平均

有75%的預測準確率，讓每位旅客省下五十美元。

Farecast正是一個巨量資料公司的縮影，也是世界未來的走向。如果是五年或十年前，伊茲奧尼絕不可能建立起這種公司。他說：「這本來是不可能的任務，」所需要的計算能力和儲存容量都還太過昂貴。然而，讓計畫成真的原因當中，雖然科技進展是關鍵因素，但還有一個更微小、卻也更重要的因素，就是關於該如何使用資料的思維，已經有所改變。

過去認為資料是靜態、靜止的，一旦完成原本蒐集的目的（例如飛機已降落、或在谷歌網頁完成了一次搜尋），便不再有用處。但是現在，資料是新的商業生產原料、重要的經濟資源投入，可以創造出新形式的經濟價值。如果心態和思維正確，就能巧妙重複運用資料，不斷帶來創新和不同的服務。只要夠謙卑、有意願、也有工具可傾聽，資料就能讓種種祕密躍然眼前。

讓巨量資料說話

不論是每個人口袋裡的手機、背著到處走的電腦、又或是辦公室所使用的伺服器系統，都是資訊社會明顯而豐碩的果實。但相較之下，「資訊」本身就不那麼引人注意。自從電腦在半世紀前進入主流社會以來，累積的資料已經到了一定程度，開始帶來全新而特殊的改變。現在，世界上不僅是資訊量前所未見，資訊成長的速度更是一日千里。規模的改變已經開始導致狀態的改變；換句話說，就是從量變引發了質變。譬如天文學、基因組學之類