



叢書總編：倪光南

善數者成

大數據改變中國

Digital: Big Data Changes China

涂子沛 鄭磊 等
/ 著

開明書店



「科技改變中國」叢書

編委會

叢書總主編

倪光南 中國工程院院士，中國科學院計算技術研究所研究員

叢書副總主編

寧濱 中國工程院院士，北京交通大學原校長

吳偉仁 中國工程院院士，國家國防科技工業局中國探月工程總設計師

徐宗本 中國科學院院士，西安交通大學原副校長

顧 翀 人民郵電出版社有限公司總經理

韓建民 杭州電子科技大學融媒體與主題出版研究院院長

編 委

武鎖寧 中國通信企業協會副會長，人民郵電報社原總編輯

陳 鍾 北京大學教授、博士生導師

馬殿富 北京航空航天大學教授、博士生導師

胡堅波 中國信息通信研究院總工程師

安 暉 中國電子信息產業發展研究院副總工程師

何寶宏 中國信息通信研究院雲計算與大數據研究所所長

陸 峰 中國電子信息產業發展研究院電子信息產業研究所副所長

推薦序

在我們的日常生活中，存在着各種形式的資料，比如文件、圖表、歌曲、演講視頻等，這些資料都是對社會經濟和生產生活片斷的記錄，這些記錄以數字化的形式存在、成為信息的載體時，就是數據。通俗地說，數據是數字化的資料，而大數據，就是大而複雜的資料集。

伴隨着過去半個多世紀信息技術的跨越式發展，上述數字化資料集開始以驚人的速度增長，數據排山倒海而來。如何處理這些數據，給科學家帶來了巨大的挑戰。但是我們也欣喜地發現，當數據積累到一定的量級，數據背後關於自然和社會的客觀規律也開始呈現出來，人類通過挖掘、分析，可以從龐大的資料集中判斷事物的特點、趨勢和相互關係，從而讓數據釋放出科學的偉力。預計在未來很長一段時間內，挖掘各領域數據的價值，從而實現由數據到信息再到知識和決策的轉換，將是一個基本的社會科學活動。大數據時代已然來臨。

雖然人類社會邁入這個新時代至今才不到十年，但世界各國都不約而同看到了大數據的價值。它既是重構社會經濟

的基本生產資料和促進生產力的利器，也是國家創新發展的核心驅動力，發展、普及大數據技術以及培養文化意識，十分迫切而且重要。

放眼世界，很多國家已經把經濟數字化作為實現創新發展的重要動能，一些先進國家還相繼出台了大數據發展規劃，把它上升到國家戰略的地位。就此而言，中國正處於全世界的第一梯隊，發展大數據具有獨特的優勢。一方面，這和我國數據資源豐富、市場規模巨大、互聯網普及程度高有關；另一方面，我國能夠集中力量辦大事，也保障了數據歸集、標準統一等大數據基礎性問題的解決。

《善數者成：大數據改變中國》是一本應時之作，以「深科普」的視角，關注當下大數據在中國催生的現象和變革，從社會管理到商業經濟，從交通醫療到環境生態，展現了大數據在各個領域前沿極具價值的應用場景，案例生動鮮活，筆觸溫暖生動，書中有不少思考和洞察，讓人耳目一新，受益良多。

本書兩位編者中，涂子沛先生是中國大數據領域的開拓者之一，也是極具影響力的大數據佈道者，著有一系列相關作品。另一位編者鄭磊教授一直堅持從事大數據領域的學術研究和決策諮詢，成果豐富。兩位編者深耕細作，本書的出版值得稱道。

誠如作者在書中所述，新的發展也帶來了新的問題，大數據時代出現的公共問題尤其值得我們關注，比如數據鴻溝、數據安全、數據主權、數據如何跨境流動，以及隱私保護等。只有解決了這些問題，才能更好地提煉和利用數據價值，從而有力推動經濟轉型和發展，提升國家治理現代化水平，在紛繁複雜的外部環境中打造新的國家競爭優勢。但要解決這些新問題，目前並沒有現成的方案，引用涂子沛先生《數文明》書中的一句話：通往美好社會的道路，永遠都在修建當中。這還需要學界、政界和業界不斷思考和努力。

總之，大數據改變中國的篇章才剛剛開始，讓我們拭目以待，迎接、建設這一新的時代。

中國科學院院士 徐宗本

前言

人類正處於一個前所未見的大數據時代。社交媒體、移動互聯網和物聯網的發展，讓人類經歷了空前的數據爆炸；而數據處理和分析技術的進步，更讓人類使用海量數據的能力得到了極大的提高。藉此，人類可以更好地發現知識、提升能力、創造價值，政治、經濟、學術等各大領域都出現了新的發展機遇。

大數據正在改變世界，也在改變中國。近年來，大數據產業發展日新月異，新興業態不斷湧現，大數據與實體經濟融合發展的水平穩步提升。我國政府數據共享開放的步伐也不斷加快，利用大數據提升行政管理、公共服務和社會治理水平初見成效。展望未來，我國在大數據領域的市場規模和數據資源優勢還將繼續發揮，關鍵技術研發有望繼續取得突破，大數據改變中國的進程才剛剛開始。

那麼，大數據正在如何改變中國？未來大數據還有望給我國帶來哪些變化？本書就將重點回答這些問題。在結構上，本書第一章首先介紹大數據的前世今生，介紹大數據時代從哪裏起步、有哪些特點。接着，本書第二章至第十章介

紹大數據在社會生活中的九個重要領域裏給我國帶來的改變，從政府公共服務與社會治理，到製造業、商業與金融業，再到與老百姓日常生活息息相關的交通、教育、醫療等領域，都能看到大數據給我國經濟社會的方方面面帶來的可喜變化。最後，本書展望大數據的未來，什麼將繼續改變、在改變的過程中還面臨哪些挑戰，以及什麼不應該被改變。

作為一本「深科普」性質的讀物，本書的編寫主要有以下三個特點。

首先，力爭在理論與故事之間找到平衡。我們試圖通過實實在在的案例和真實的故事，為廣大讀者展示大數據給我國各行各業帶來的巨大變化，起到開闊視野、啟迪思考的作用。但案例和故事的背後，離不開國內外數據科學、計算機科學、信息管理、公共管理乃至哲學、歷史等各個學科的學術成果和理論框架。

其次，力爭在技術與人文之間找到平衡。儘管這是一本集中介紹科技成果的讀物，但我們認為科技的發展應該解放而非束縛人類。在展現技術力量的同時，我們時刻不忘人文的溫度，呼籲縮小數據鴻溝、保護數據隱私、反對「數據迷信」。大數據的發展應以人為中心，維護人的權利和尊嚴，促進人的全面發展，滿足人們對美好生活的嚮往，而不是走

向相反的方向。

最後，力爭在弘揚與反思之間找到平衡。儘管大數據在中國的發展高歌猛進，碩果累累，但我們必須時刻保持清醒的頭腦，絲毫不能沾沾自喜。現實與理想還有差距，成績和不足瑕瑜互見。我們用大量的篇幅介紹大數據應用的成功案例，但也反思存在的不足，更明言可能的挑戰。科技發展對社會進步的促進作用不是「短跑」，而是「馬拉松」，既要抓住機遇，也要應對挑戰，居安思危方能行穩致遠。

希望讀者朋友們能通過本書對大數據已經給中國帶來的巨大改變有一個直觀、深入的認識，並能對大數據即將給我們帶來的機遇和挑戰有全面和充分的準備，最後還能進一步對科技與人之間的關係應該如何改變和演化這一問題進行思考和討論。

本書的完成，首先要感謝編寫團隊為期半年的艱苦勞作，感謝「科技改變中國」叢書總主編倪光南院士對本書的悉心指導。

本書編寫人員分處廣州、上海、湘潭、南寧等地，地域分散，集結困難，書稿撰寫階段，寫作組每周定期召開微信電話會，交流心得，碰撞觀點，常在周末和假期的深夜還在加班加點、打磨文字。全書數易其稿，凝結了團隊全體成員

的辛勤汗水。

本書共十一章，第一章、第六章由涂子沛執筆，第二章由博士生王翔（復旦大學）執筆，第三章由朱曉婷（復旦大學）執筆，第四章由溫祖卿（復旦大學）執筆，第五章由涂斯婧（廣西中醫藥大學）執筆，第七章由葉俊傑博士（數文明科技）執筆，第八章由朱曉婷、溫祖卿執筆，第九章由杜為兮、李楠（數文明科技）執筆，第十章由張炳劍、石大義（數文明科技）執筆，第十一章由王翔、鄭磊執筆，全書由涂子沛、鄭磊統稿。

感謝數文明科技公司，以及復旦大學數字與移動治理實驗室的同學和業界朋友對本書創作的支持。李楠協助修改、整理書稿，把控項目進度，鄧志新對個別章節提出了寶貴的修改意見。珠海伊斯佳王德友董事長為編寫過程中的調研走訪提供了大力支持。還要特別感謝人民郵電出版社王威和賀瑞君等編輯對書稿提出的建設性意見，他們為本書的面世做了非常細緻的工作。

在中華人民共和國成立 70 周年之際，能以此書獻禮，我們既感榮幸，又感重任在肩。我們深知，本書只是對我國大數據發展的一個階段性小結。限於知識和能力，本書講述的故事和展開的討論難免掛一漏萬，還請各位讀者方家不吝指正。

目 錄

第一章 大數據的前世今生

- 1.1 正解大數據：世上本沒有數 / 2
- 1.2 存儲革命：摩爾定律推動的進化 / 7
- 1.3 社交媒體：每個人都是數據的生產者和協作者 / 17
- 1.4 數據挖掘如何「點數成金」 / 19

第二章 數字治理：用大數據提升政府管理與公共服務水平

- 2.1 從「告別奇葩證明」到「告別證明」 / 28
- 2.2 「12345」數據讓城市更美好 / 32
- 2.3 大數據辨識真假「鬼城」 / 34
- 2.4 「數據鐵籠」讓權力不再「任性」 / 37
- 2.5 「Gov Store」：開放數據，建立生態 / 40
- 2.6 數據跑不到的地方，用溫情來彌補 / 45

第三章 變革時空：數據再造出行與物流

- 3.1 城市「數腦」：改善交通擁堵的新方案 / 52
- 3.2 智慧物流：實現更貼心的最後一公里 / 59
- 3.3 數據開路：來一場說走就走的旅行 / 65
- 3.4 數據止痛：改變時間與空間的交錯 / 70
- 3.5 數據監管：立法規範進行時 / 73

第四章 教育升「溫」：用數據精準滴灌

- 4.1 教學科研：被大數據換上新顏 / 79
- 4.2 教育管理：因大數據而行穩致遠 / 85

- 4.3 教育與大數據：緣何走到一起 / 90
- 4.4 路在何方：人的全面發展與數據的底線 / 92

第五章 顛覆醫療：大數據助力健康中國

- 5.1 「智慧養老」：讓關懷永不缺席 / 97
- 5.2 「數」有所為：生命的「精算符」 / 106
- 5.3 「互聯網+醫療」：醫患和諧的「公約數」 / 114
- 5.4 醫療「聯姻」區塊鏈：念好隱私的「緊箍咒」 / 121

第六章 無僥倖天下：一個更安全的中國社會

- 6.1 要是此案在中國，早破了 / 126
- 6.2 城市「視網膜」如何看見 / 132
- 6.3 邊緣計算：驅動計算之網 / 137
- 6.4 軌跡追蹤：賦能公共安全 / 142
- 6.5 硬盤和眼藥水為什麼同時脫銷 / 149
- 6.6 無僥倖天下：大數據重建社會的安全和秩序 / 156

第七章 數據造夢：為金融業挖出一座「金礦」

- 7.1 點石成金：餘額寶背後的大數據故事 / 166
- 7.2 技術升維：大數據風控破殼而出 / 172
- 7.3 火眼金睛：大數據金融監管走上舞台 / 177
- 7.4 數據信託：一個全新的大數據金融產品 / 185
- 7.5 浪潮席捲：一個無可限量的市場 / 187

第八章 撬動商業：新「規模經濟」，數最懂你

- 8.1 精準營銷：從廣而告之到瞄準目標 / 194
- 8.2 數據「智導」：再造影視創作模式 / 198

- 8.3 「數造」個性：以社群文化帶動新營銷 / 201
- 8.4 大數據何以撬動商業變革 / 204
- 8.5 數據創造價值：前提是規則和邊界 / 207

第九章 數據革新：正被重構的製造業版圖

- 9.1 個性化定製：規模化和差異化的結合 / 213
- 9.2 大數據「問診」：為製造開一劑良方 / 226
- 9.3 工業互聯網：不僅僅是機器換人 / 233
- 9.4 未來已至：覺醒的「智」造業大國 / 241

第十章 數治生態：行進中的美麗中國

- 10.1 關注點滴：知水方能節水 / 247
- 10.2 灌溉有「數」：別讓農作物「喝多了」 / 249
- 10.3 環保雲平台：打破「部門割據」強化監管 / 255
- 10.4 共建生態大數據，喚醒公眾參與熱情 / 262

第十一章 大數據的未來：數據主義還是人文回歸

- 11.1 大數據還將改變什麼 / 269
- 11.2 實現改變還面臨哪些挑戰 / 272
- 11.3 什麼不應該被改變 / 276
- 11.4 改變是為了什麼 / 279

參考文獻 / 281

第一章

大數據的前世今生

在互聯網經濟時代，數據是新的生產要素，是基礎性資源和戰略性資源，也是重要生產力。

——習近平總書記在中共中央政治局第二次
集體學習時做出的科學判斷^[1]

1.1 正解大數據：世上本沒有數^{1 [2]}

傳統意義上的「數據」，是指「有根據的數字」。數字之所以產生，是因為人類在實踐中發現，僅僅用語言、文字和圖形來描述這個世界是不精確的，也是遠遠不夠的。例如，有人問「姚明有多高」，如果回答說「很高」「非常高」「最高」，別人聽了，只能得到一個抽象的印象，因為每個人對「很」有不同的理解，「非常」和「最」也是相對的；但如果回答說「2.26米」，就一清二楚。除了描述世界，數據還是我們改造世界的重要工具。人類的一切生產、交換活動，可以說都是以數據為基礎展開的，例如度量衡、貨幣的背後都是數據，它們的發明或出現，都極大地推動了人類文明的進步。

如圖 1.1 所示，數據的來源分為測量、記錄和計算。數據最早來源於測量，所謂「有根據的數字」，是指數據是對客觀世界測量結果的記錄，而不是隨意產生的。測量是從古至今科學研究最主要的手段。可以說，沒有測量，就沒有科學；也可以說，一切科學的本質都是測量。就此而言，數據之於科學的

重要性，就像語言之於文學、音符之於音樂、形色之於美術一樣，離開數據，就沒有科學可言。

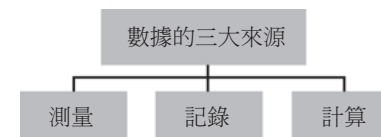


圖 1.1 數據的來源

除了測量和顯而易見的記錄，新數據還可以由老數據經計算衍生而來。測量和計算都是人為的，也就是說，世上本沒有數，一切數據都是人為的產物。我們說的「原始數據」，並不是「原始森林」這個意義上的「原始」。原始森林是指天然就存在的森林，而原始數據僅僅是指第一手、沒有經過人為修改的數據。

如圖 1.2 所示，傳統意義上的數據，和信息、知識也是完全不同的概念：數據是信息的載體，信息是有背景的數據，而知識是經過人類的歸納和整理，最終呈現規律的信息。

20 世紀 60 年代，軟件科學取得了巨大進步，數據庫被發明。此後，數字、文本、圖片都不加區分地保存在計算機的數據庫中，以「比特」為單位進行存儲，「數據」二字的內涵開始擴大。「數據」不僅指代那些作為「量」而存在的數據——也就是「量數」，還逐漸成為「數字、文本、圖片、音頻、視頻」等的統稱，即「信息」的代名詞，由於這些信息作為一種證據、根據而存在，因此可以稱為「據數」。

1 本章部分內容編選自本書編著者之一涂子沛 2014 年在中信出版社出版的《數據之巔：大數據革命，歷史、現實與未來》一書。

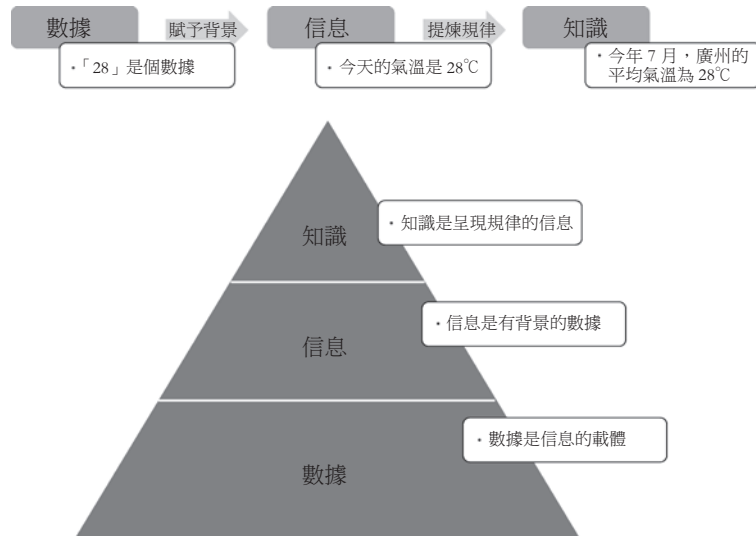


圖 1.2 數據、信息、知識的區別和聯繫

在此基礎上，關於大數據的定義，筆者主張用下面這樣一個式子來較為簡潔、精確地表示。

大數據 = 傳統的量數 + 現代的據數

(量數源於測量，如氣溫 28°C；據數源於記錄，如一張照片)

雖然量數比據數更接近「數」，但從歷史上看，據數的出現要早於量數。人類早期對自身活動的記錄，即「史」，就是早期的據數，也可以說，據數是歷史的影子。量數則是在記錄的實踐中慢慢產生的，其核心要義是精確。量數是否充沛，直

接決定了科學是否發達。從這個角度出發，數據的來源就不再只是對世界的測量，而是對世界的一種記錄。所以信息時代的數據又多了一個來源——記錄。

進入信息時代之後，數據成為信息的代名詞，兩者可以交替使用。一封郵件雖然包含很多條信息，但從技術的角度出發，可能還是「一個數據」。就此而言，現代意義上的數據的範疇，其實比信息還大，如圖 1.3 所示。

除了內涵的擴大，數據庫問世之後，還出現了另外一個重要現象，那就是數據的總量在不斷增加，而且增加的速度在不斷加快。

20 世紀 80 年代，美國就有人提出了「大數據」的概念。那個時候，其實還沒有進入數據大爆炸的時代，但有人預見到，隨着信息技術的進步，軟件的重要性將下降，數據的重要性將上升，因此提出「大數據」的概念。那時候的「大」，如「大人物」和「大轉折」之「大」，主要指價值上的重要性。到了 21 世紀，尤

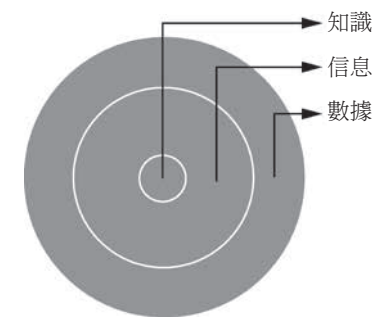


圖 1.3 現代數據的範疇

其是 2004 年社交媒體產生之後，數據開始呈爆炸式增長，國際數據公司（IDC）的數據顯示，2011—2018 年，全球的數據量增長了 18 倍，大數據的提法又重新進入大眾的視野並獲得了更大的關注。這個時候的「大」，含義也更加豐富了：一是指容量大，二是指價值大。

到底多大才算大呢？針對這一問題，十多年來爭議頗多。這首先涉及衡量數據大小的單位。2000 年的時候，一般認為「太字節（TB）」級別的數據就是大數據了，當時擁有「太」級別數據的企業並不多，但自此之後，互聯網企業開始崛起，這些企業擁有各種各樣的數據，其中大部分都是文本、圖片和視頻，其數據量之大，傳統企業根本無法望其項背。

延伸閱讀

理解幾個主要的存儲單位

一首標準音質的歌曲 ≈ 4 兆字節（MB）

一部標準畫質的電影 ≈ 1 吉字節（GB，1 吉字節 = 1 024 兆字節，相當於 250 首標準音質歌曲的大小）

一個普通圖書館的藏書 ≈ 1 太字節（TB，1 太字節 = 1 024 吉字節，相當於 1024 部標準畫質電影的大小）

其實不僅僅是互聯網行業，各行各業的數據都在爆炸，只是規模不同。如果僅僅把大數據的標準限定在互聯網企業，認為只有互聯網企業才擁有大數據，那就嚴重狹隘化了大數據的意義。畢竟容量只是表象，價值才是本質，而且大容量並不一定代表大價值。大數據的真正意義還在於大價值，價值主要通過數據的整合、分析和開放而獲得。從這個方面來看，大數據的真正意義是，人類擁有了前所未有的能力來使用海量的數據，在其中發現新知識、創造新價值，從而為社會帶來「大知識」「大科技」「大效益」和「大智能」等發展機遇。

以上論述是從概念上分析「數據」和「大數據」的區別，而掌握一個概念最好的方法，還是得從動態上了解其成因。大數據的形成，不僅是因為人類信息技術的進步，還是信息技術領域不同時期多個進步交互作用的結果，其中最重要的原因，當數摩爾定律的持續有效。

1.2 存儲革命：摩爾定律推動的進化

1965 年，英特爾公司的創始人之一戈登·摩爾（Gordon Moore）在考察了計算機硬件的發展規律之後，提出了著名的

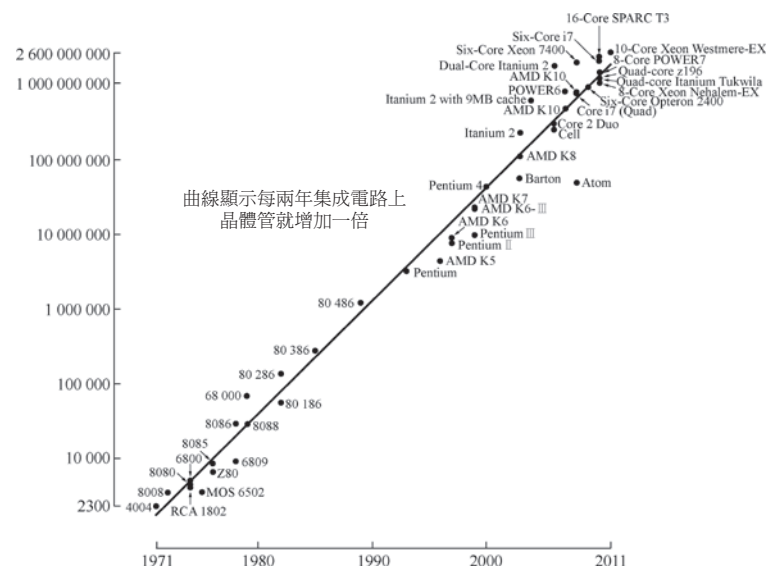
摩爾定律。該定律認為，同一面積芯片上可容納的晶體管數量，一到兩年將增加一倍。¹

要理解這種增加的意義，並不簡單。摩爾的本意是，由於單位面積芯片上晶體管的密度增加了，計算機硬件的處理速度、存儲能力，即其主要性能，一到兩年將提升一倍。本來性能提升了，價格也應該上升才對，但實際情況恰恰相反：半個多世紀以來，硬件的性能不斷提高，但價格卻持續下降。這背後的主要原因，竟然是因為晶體管越做越小，這種體積的縮小也使得其成本下降；再加上人類對晶體管的需求越來越大，大規模的生產也使得價格不斷下降。

回顧這半個多世紀的歷史，硬件的發展基本符合摩爾定律，如圖 1.4 所示。以物理存儲器為例，其性能確實不斷上升，與此同時，價格不斷下降。1955 年，IBM 推出了第一款商用硬盤存儲器，1 兆字節容量的存儲器需要 6 000 多美元。此後，其價格不斷下降：1960 年下降到 3 600 美元；1993 年，下降到約 1 美元；2000 年降至約 1 美分；到 2010 年，每兆字

¹ 摩爾 1965 年提出該定律時，認為這個周期是一年；1975 年，他修訂為兩年。也有人認為這個周期是 18 個月。

節價格約為 0.005 美分。半個多世紀以來，存儲器的價格下降到原來的約一億分之一，這種變化巨大而且劇烈，令人瞠目結舌。事實上，縱觀人類全部的歷史，沒有其他任何一種產品，其價格的下降空間能夠如此巨大！



注：縱坐標為晶體管數量，橫坐標為年份。該曲線表明，1971—2011 年，大概每兩年相同面積的中央處理器集成電路上的晶體管就增加一倍。需要注意的是，縱坐標從 2 300 到 10 000 再到 100 000，其實不成比例。如果嚴格按比例作圖，這將是一條非常陡峭的曲線，頁面將無法容納（資料來源：維基百科）。

圖 1.4 1971—2011 年中央處理器上的晶體管數量和摩爾定律關係示意

延伸閱讀

晶體管的產量多過全世界的大米顆粒

晶體管由硅構成，相當於一個開關，通電的時候表示「1」，不通電的時候表示「0」，是電子產品最小的組織單元。一台筆記本電腦大概有 400 億個晶體管，一部智能手機約有 10 億個晶體管。晶體管行業（即半導體行業）堪稱人類歷史上最高產的行業。現在全球一年生產的晶體管比一年消耗的大米顆粒還要多：2002 年，人類生產的晶體管數量大概是小米的 40 倍，買一粒米的錢可以購買 100 個晶體管^[3]；2009 年，晶體管的產量上升到大米的 250 倍，一粒大米的價錢可以購買 10 萬個晶體管^[4]。

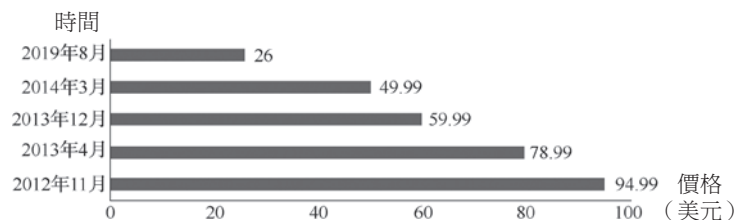
摩爾定律發展到今天，一根頭髮尖大小的地方，就能放上萬個晶體管。當然，晶體管不可能無限縮小，所以十幾年來，業界曾圍繞以下問題展開激烈爭論：摩爾定律所揭示的現象還會不會持續，即單位面積上的晶體管還能不能繼續增加甚至翻倍？如果能，又能持續多久？

2003 年，作為摩爾定律的發現者，戈登·摩爾也被問到這個問題。他認為：「創新無止境，下一個 10 年摩爾定律可能

還將有效。」

事實證明，摩爾是對的。2011 年，英特爾公司宣佈發明了 22 納米工藝的 3D（三維）晶體管，這使爭論暫時畫上了句號。此前最小的晶體管為 31 納米工藝，22 納米的晶體管小了大約 1/3。因為小，新的晶體管總是更便宜、更節能。2012 年，英特爾宣佈將投資 50 億美元在美國亞利桑那州建廠，在 2014 年投產 14 納米工藝的晶體管，這比 22 納米工藝的尺寸又縮小了 1/3。在 2019 年 1 月，英特爾又向外界展示了其首批 10 納米工藝的 Ice Lake 處理器，相當於在 1 平方毫米中塞下了 1 億個晶體管。該產品於 2019 年正式推出並供應市場。^[5]

英特爾公司的發明使大部分科學家相信，晶體管的微縮至少在十年內還是會持續，摩爾定律的生命周期尚未終結。未來，1 太字節硬盤容量的價格將相當於 1 杯咖啡的價格，其價格趨勢如圖 1.5 所示。美國的國會圖書館是全世界最大的圖書館，其印刷品館藏數據量約為 15 太字節，一所普通大學的圖書館，其館藏數據量可能只有 1~2 太字節。也就是說，在不久的將來，只需花上一杯咖啡的錢，就可以把一個圖書館的全部信息拷貝進一個小小的硬盤。信息保存的過程如此方便、成本如此低廉，歷史上從來沒有過。



注：筆者跟蹤了亞馬遜和京東網站上希捷硬盤在不同時段的報價，2012—2019年，1太字節硬盤容量價格下降顯著。

圖 1.5 1 太字節硬盤容量的價格變化

現在，摩爾定律已經成為描述一切呈指數級增長事物的代名詞，它給人類社會帶來的影響非常深遠。正是因為存儲器的價格在半個世紀之內經歷了空前的下降，人類才可能以非常低廉的成本保存海量的數據，這為大數據時代的到來鋪平了硬件道路。低價存儲器相當於物質基礎，沒有它，大數據無異於水中月、鏡中花。

延伸閱讀

摩爾定律促使硬件成為大眾消費品

摩爾定律使得硬件價格大幅下降，最終使曾經昂貴的硬件成為大眾消費品，原來高端的產品，如激光打印機、

服務器、智能手機，已經逐漸從科研機構、大型企業進入普通家庭。由於這些設備的普及，美國的一些公司甚至出現了一種新趨勢：鼓勵員工自己帶設備來上班（Bring Your Own Device, BYOD），公司只提供網絡和辦公場地，成為「輕」公司。

除了便宜、功能強大，摩爾定律也導致各種計算設備變得越來越小。這個現象在 1988 年被美國科學家馬克·韋澤（Mark Weiser）概括為「普適計算」。普適計算理論認為，計算機發明以後，將經歷三個主要階段：第一階段是主機型階段，指的是很多人共享一台大型機，一台機器就佔據半個房間；第二階段是個人計算機階段，計算機變小，人手一機，韋澤當時就處於這個時代，這似乎已經是很理想的狀態，但韋澤天才般地預見到，人手一機不是時代的終結；在第三個階段，計算機將變得很小，小得將從人們的視線中消失，人們可以在日常環境中廣泛部署各種各樣微小的計算設備，在任何時間、地點都能獲取並處理數據，計算設備最終將和環境融為一體，這個階段被稱為普適計算階段。

今天，普適計算第三階段的浪潮正向我們奔湧而來，小小的智能手機，其功能已經毫不遜色於一台計算機，各種傳感

器正越做越小，RFID（射頻識別）標籤方興未艾，可穿戴式設備又向我們走來。

RFID 標籤已經在零售、醫療、城市管理、動物飼養等領域得到了廣泛應用。近兩年，上海、烏鎮等地陸續展開智能垃圾桶應用，在垃圾桶內安裝 RFID 傳感器，實時感知垃圾投放數量及存放量，垃圾桶還可以自動「通知」環衛工人哪處垃圾已滿需要清理，大大提升了城市管理工作效率。RFID 也在改寫航空業。2019 年，中國東方航空針對行李托運部署了 RFID 技術，為「行李」這位「不會說話的旅客」解鎖「表達」技能。旅客通過微信小程序，即可查詢到托運行李的運輸狀態，精準鎖定位置，實時掌握動態，猶如為行李安裝了 GPS 定位系統。^[6]

「可穿戴元年」可以追溯到谷歌眼鏡面世的 2012 年，隨後，各種智能可穿戴設備層出不窮。可穿戴式設備是指可以穿戴在身上、不影響個人活動的微型電子設備，這些設備可以記錄佩戴者的物理位置、熱量消耗、體溫、心跳、睡眠模式、步數以及健身目標等數據。2015 年亞洲杯上，中國國家足球隊身穿黑色「比基尼」訓練的新聞圖片曾一度躋身熱搜榜。其實，這件看似性感的訓練背心，就是一款名為「GPSports」的可穿戴設備，能夠對運動員的跑動距離、路線、速度、加速度

以及心率變化等參數進行採集和監測。通過對數據的對比和進一步分析，教練人員可以制訂訓練計劃，安排比賽陣容，做出臨場指揮的關鍵決策。^[7]

法國的運動器材製造商 Babolat 還把傳感器安裝在了網球拍的手柄上，它可以記錄球員擊球時的狀態參數，例如正反拍、擊球點、擊球的力量、球速、球的旋轉方向等。這些數據以幾乎實時的速度傳到現場的智能手機和平板電腦上，運動員和教練可以隨時查看。

2014 年在澳網奪冠的中國網球「一姐」李娜，用的就是這個品牌的球拍。為了配合這種球拍的使用，2013 年，國際網球聯合會（International Tennis Federation, ITF）已經修改了章程，從 2014 年 1 月起，允許運動員在國際比賽中使用帶有傳感器的球拍，以記錄、分析自己的數據。在未來的比賽中，如果運動員同意，這些數據甚至可以實時出現在比賽場地的大屏幕上，供觀眾分析參考。

除了足球、網球領域，傳感器也在快速進入棒球、橄欖球等領域。美國的一些研究機構認為，美國運動產業的營收近年內會有大幅增長，主要原因就是，基於傳感器的數據收集和分析技術將改變整個產業的生態。

除了運動，可穿戴式設備還有很多其他應用，甚至連一

片小小的紙尿褲也開始了自己的智慧升級。2015年，一個名為「貝肯熊」的國產品牌研發出一款新型智能紙尿褲，通過在紙尿褲中植入一個輕巧的濕度傳感智能硬件，連接藍牙，使之與看護者的手機綁定，一旦寶寶尿了，靈敏的智能硬件就會用鈴聲或震動的方式通知看護者。

此外，作為可穿戴式設備最經典的產品而風靡一時的谷歌眼鏡，其同類產品也在娛樂之外得到了更廣泛的應用：2018年2月，鄭州鐵路警方在全國鐵路系統中率先使用了人臉比對警務眼鏡，新聞報道說這款眼鏡可以通過人臉識別，篩查出旅客中的不法分子，有效淨化列車的治安環境。^[8]

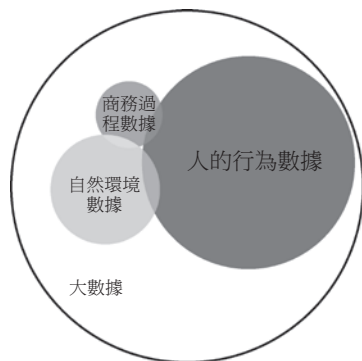
普適計算的本質，是在人類生活的物理環境中廣泛部署微小的計算設備，實現無處不在的數據自動採集，這意味着人類收集數據能力的增強。在此之前，電子化的數據主要由各種信息系統產生，這些信息系統記錄的主要是商業過程中產生的數據。而傳感器的出現及其技術的成熟，使人類開始有能力大規模記錄物理世界的狀態，這種進步推動了大數據時代的到來。

1.3 社交媒體：每個人都是數據的生產者和協作者

人類數據的真正爆炸發生在社交媒體時代。

從2004年起，以臉譜網（Facebook）、推特（Twitter）為代表的社交媒體相繼問世，拉開了一個互聯網的嶄新時代——Web 2.0。在此之前，互聯網的主要作用是信息的傳播和分享，其最主要的組織形式是網站，但網站是靜態的。進入Web 2.0時代之後，互聯網開始成為人們實時互動、交流協同的載體。

除了把交流和協同的功能推到了一個登峰造極的高度，社交媒體的另外一層重要意義就是，給全世界無數的網民提供了平台，使其隨時隨地都可以記錄自己的行為、想法，這種記錄其實就是貢獻數據。前面我們談到過，所有的數據都是人為產生的，所有的數據都是對世界的測量、記錄和計算。從1946年人類發明第一台計算機並進入信息時代算起，到社交媒體產生之前，主要是信息系統、傳感器在產生和收集數據，但由於社交媒體的橫空出世，人類自己也開始在互聯網上生產數據，他們發微博和微信，記錄各自的活動和行為，這部分數據也因此被稱為「行為數據」，如圖1.6所示。



注：各類數據間存在交互、影響。商務數據中自然會包含和產生人的行為數據與自然環境數據，人的行為數據與自然環境數據也相互包含、交叉並影響。過去，是我們選擇什麼東西需要記錄，才對它進行記錄；在大數據時代，是選擇什麼東西不需要記錄，才取消對它的記錄。隨着記錄範圍不斷擴大，可以肯定，人類的數據總量還將滾雪球式地增大。

圖 1.6 各種數據的大小和種類

由於社交媒體的出現，全世界的網民都開始成為數據的生產者，每個網民都猶如一個信息系統、一個傳感器，不斷地製造數據。這引發了人類歷史上迄今為止最龐大的數據爆炸。

除了數據總量驟然增加，社交媒體還使人類的數據世界更為複雜。在大家發的微博中，你的帶圖片，他的帶視頻，大小、結構完全不一樣。因為沒有嚴整的結構，在社交媒體上產生的數據也被稱為非結構化數據。這部分數據的處理遠比處

理結構嚴整的數據困難。2019年3月15日，新浪微博發佈的《2018 微博用戶發展報告》顯示，截至2018年第四季度，新浪微博日均文字發佈量為1.3億條，日均圖片發佈量1.2億幅，日均視頻/直播發佈量150萬次以上。而在過去50年，《紐約時報》產生的信息量總共也不過30億個單詞。

在這種前所未有的數據生產速度下，目前全世界的數據大約75%都是非結構化數據。今天回頭看，社交媒體的出現，給了大數據一錘定音的力量。基於以上分析，我們也可以這樣認為：

大數據 = 結構化數據 + 非結構化數據

但我們前面談到，大數據之大，不僅在於其大容量，更在於其大價值。價值在於使用，如同埋在地底下的石油，遠古即已有之，人類進入石油時代，是因為掌握了開採、冶煉石油的技術；現在進入大數據時代，最根本的原因，也是人類使用數據的能力取得了重大突破和進步。

1.4 數據挖掘如何「點數成金」

數據使用能力的突破集中表現在數據挖掘上。

數據挖掘是指通過特定的算法對大量的數據進行自動分析，從而揭示數據當中隱藏的規律和趨勢，即在大量的數據

中發現新知識，為決策者提供參考。數據挖掘的進步，根本原因是人類能夠不斷設計出更強大的模式識別算法¹，這其實是軟件的進步。其中最重要的里程碑，是1989年美國計算機協會（Association for Computing Machinery, ACM）下屬的知識發現和數據挖掘小組（Special Interest Group on Knowledge Discovery and Data Mining, SIGKDD）舉辦了第一屆數據挖掘學術年會，出版了專門期刊，此後數據挖掘發展得如火如荼。

正是通過數據挖掘，近幾十年來，各大公司譜寫了不少「點數成金」的傳奇故事。例如阿里巴巴憑藉長期以來積累的用戶資金流水記錄，涉足金融領域，在幾分鐘之內就能判斷用戶的信用資質，決定是否為其發放貸款；沃爾瑪通過捆綁「啤酒和尿布」提高門店商品銷量；奈飛公司（Netflix）利用客戶的網上點擊記錄，預測其喜歡觀看的內容，實現精準營銷等。

近年來，數據挖掘的應用還在不斷推陳出新，有望到達一個新高度。例如，曾與我們「相看兩不厭」數千年的菜市場，正在走向發展的拐點。2019年初，在阿里巴巴本地生活鮮夥伴大會上，「餓了麼」提出要「改變菜市場」，建立全新的生鮮開放平台，把菜市場搬到線上，讓傳統菜市場告別

1 算法是運用數學和統計學的方法和技巧，解決某一類問題的特定步驟。

數千年單兵作戰、看天賣菜的模式，並讓平台協作賣菜成為主流。

怎麼實現協作呢？關鍵利器就是數據挖掘。傳統菜市場最大的痛點就是信息不對稱，進貨的商戶找不准市場真實需求而導致商品積存或出現質量問題。而「餓了麼」背靠阿里巴巴的海量數據資源，可以為商戶提供最精準的用戶畫像，從而指導其進貨行為。從此，菜市場的進貨行為不再隨機，決策過程被外包給了算法，由算法來決定賣什麼，這種數字化營銷讓商家與平台共振，可以激發出極大的商業價值。這種模式已經被市場所驗證：「叮咚買菜」在入駐「餓了麼」之後，2018年全年平台單量增長20倍，月交易額超千萬元。^[9]

還有一則關於數據挖掘的小故事。2012年6月歐洲盃足球賽期間，我國出現了多篇「男人一看球，女人就網購」的相關報道^[10]。報道稱，根據淘寶網的銷售數據，歐洲盃開賽以來，女性網購的成交量明顯上升，而且「網購的高峰期延時兩個小時，變成了23點到24點」。此外，在「凌晨1點45分第一場球結束到凌晨2點45分第二場球開始前」，出現了一個新的網購高峰，這個新的高峰和賽前的同時段相比，成交量「增長超過260%」。

這個現象背後的邏輯不難理解。球賽期間，男性沉迷於

球賽，冷落了妻子（女朋友）和孩子。女性，特別是已婚女性會覺得沮喪、惱火、失落。每天晚上球賽開始的時候，在個體層面，每位女性都有很多選擇，她可以做家務、跟閨蜜聊天、和母親通電話或上網購物等，其行為具有不確定性，她究竟會做什麼，難以預測。但是，當我們把幾個電子商務平台的交易數據一匯總、一分析，就會發現，女性群體的行為有規律可循。隨着球賽的開始，女性在網上購物的成交量就開始增加，其中的高檔物品也較平時明顯增多，也就是說，平時捨不得買的東西，這時候終於出手了。在大數據時代之前，「男人一看球，女人就網購」永遠是一個猜測，無法得到證實。但在大數據時代，這很容易就能證實，甚至連成交的商品有什麼特點，都可以進行分析。等到下一年球賽再開始的時候，商家的廣告就可以更有的放矢，不僅可以把廣告對象瞄得更准，推廣的商品也會更有針對性，猜測上升為知識，知識將創造利潤。

除了上述商業應用，用數據挖掘來解決社會問題，也正變得越來越普遍。2013年7月，有報道稱，華東師範大學的一位女生收到校方的短信：「同學你好，發現你上個月餐飲消費較少，不知是否有經濟困難？」^[11]這條溫暖的短信也要歸功於數據挖掘：校方通過挖掘校園飯卡的消費數據，發現其每頓的餐費都偏低，於是發出了關心的詢問。但隨後發現這是一

個美麗的錯誤——該女生其實是在減肥。可以想到，誤會之所以發生，還是因為數據不夠「大」，大數據的特點除了「量大」，還有「多源」，如果除了飯卡，還有其他來源的數據作為輔助，判斷就可能更加準確。

雖然數據挖掘仍如日中天，但在一定程度上，數據挖掘已經不是大數據的前沿和熱點，取而代之的是機器學習。當下興起的機器學習憑藉的也是計算機算法，但和數據挖掘相比，其算法並不是固定的，而是帶有自調適參數的，也就是說，它能夠隨着計算、挖掘次數的增多，不斷自動調整自己算法的參數，使挖掘和預測的結果更為準確，即通過給機器提供大量的數據，讓機器可以像人一樣通過學習逐步自我改善提高，這也是該技術被命名為「機器學習」的原因。

除了數據挖掘和機器學習，數據的分析、使用技術已經非常成熟，並且形成了一個體系。數據倉庫、聯機分析處理（OLAP）、數據可視化、內存分析都是該體系的重要組成部分，在人類數據技術的進步過程中，都扮演過重要的角色¹。

回顧半個多世紀人類信息社會的歷史，正是因為晶體

1 關於人類數據分析技術的演進，有興趣的讀者請參閱本書編著者之一涂子沛所著《大數據》一書第四章「商務智能的前世今生」中的闡述。^[12]

管越做越小、成本越來越低，才形成了大數據現象的物理基礎。這相當於鑄器，人類有能力製造巨鼎盛載海量的數據。1989年興起的數據挖掘，則相當於把原油煉成石油的技術，是讓大數據產生「大價值」的關鍵，沒有技術，原油再多，我們也只能「望油興歎」。2004年出現的社交媒體，則把全世界每個人都轉變成了潛在的數據生成器，向摩爾定律鑄成的巨鼎貢獻數據，這是「大容量」形成的主要原因，如圖 1.7 所示。

分析了大數據的靜態概念和動態成因，我們更清楚地理解了大數據的特點，現在可以從圖 1.8 所示的以下幾個角度來理解、定義大數據。正如前文討論的，當前人類的數據約 75% 都是非結構化數據，大記錄的表現形式主要就是非結構化數

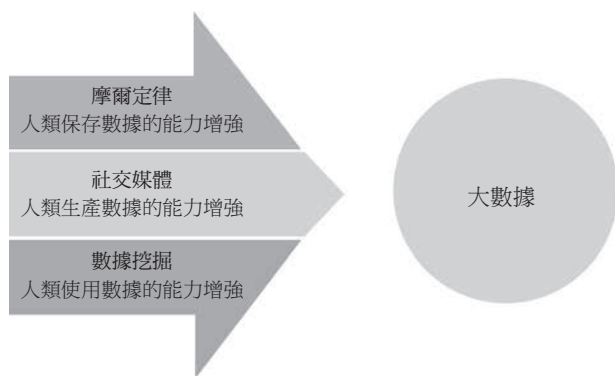


圖 1.7 大數據的三大成因

據，而大記錄、非結構化數據要體現出價值，當前主要的處理方法，還是把它們轉化為有嚴整結構的數據，即傳統的小數據。因此筆者認為，大數據的價值維度主要體現在傳統的小數據和結構化數據之上，而大數據的容量維度主要體現在現代的大記錄和非結構化數據兩個方面。

大數據浪潮興起之後，全世界的科學家都在預測和展望——這股由信息技術掀起的新浪潮將對人類社會產生何種影響，將帶領中國和世界走向何方？在下面幾章中，我們選幾個側面來嘗試剖析。

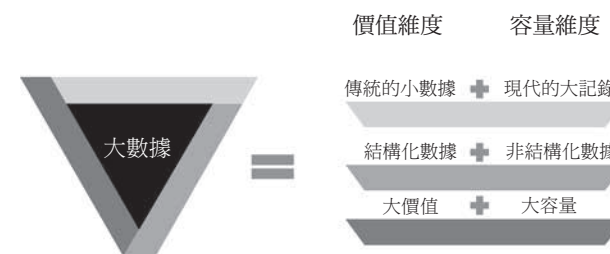


圖 1.8 大數據的概念和維度