評 *Bit By Bit: Social Research in the Digital Age*

洪晨碩
美國麻薩諸塞大學（安姆斯特分校）社會學系

Bit by Bit: Social Research in the Digital Age. By Matthew J. Salganik. Princeton, NJ: Princeton University Press, 2018. xix+423 pages.

近年來，許多社會科學研究者注意到大數據（big data）具有探索人類行爲與社會趨勢的潛力。網際網路與社群媒體的發展，像是Facebook、Twitter，使民眾表達意見的管道更加即時且多元。另一方面，物聯網（Internet of Things）等資訊科技的普及，使得蒐集海量行爲資料也越來越容易。如何善用數位時代下大數據帶來的研究機會，提出更多關於社會運作的洞見，成爲許多社會科學家的研究目標。在 *Bit By Bit: Social Research in the Digital Age* 這本書（以下簡稱 *Bit By Bit*）中，社會學家 Matthew J. Salganik 以豐富的案例和討論，提供研究者結合數位科技、大數據與社會科學研究的指引。

Bit By Bit 共分七章。除了第一章導論與第七章結論之外，第二到第五章節分別對應四種常見的研究方法：行爲觀察、訪調、實驗、以及大眾協作。綜觀全書內容，*Bit By Bit* 將重點聚焦在兩個主題上。第一個是混合現成（readymades）與客製（custommades），指的是研究者在數位時代擁有更多的機會混合從企業或政府取得的海量數據，以及依其研究旨趣蒐集而來的特定數據。第二個主題是研究倫理

DOI: 10.6676/TS.201806_(35).06

（ethics），指的是因應大數據取得更加快速方便，研究者需要在研究設計階段更謹慎的考量數據蒐集與分析過程中對受試者可能帶來的風險。面對越來越多大數據研究出現倫理爭議，Salganik 另以第六章詳細討論數位時代對研究倫理帶來哪些衝擊，並提出建議。

Bit By Bit 的編排較接近一般教科書的呈現方式。透過書中大量的案例引用與延伸閱讀，可以看出作者 Salganik 希望本書能夠激發社會科學家對大數據的興趣。每一章節除了整理數位研究的特色外，也提供讀者未來執行研究時的設計建議。由於大數據在社會學仍屬新興議題，本書評將把重點放在引介該書內容。本文將先梳理 Salganik 對大數據的定義，接著討論數位時代下的社會研究以及研究倫理議題，最後總結本書貢獻以及未來可以繼續發展的研究方向。

一、大數據的「大」有哪些意涵？

不同於一般對大數據的定義著重在所謂的「4V」——數量（volume）、時效性（velocity）、種類（variety）、真實性（veracity）[1]——*Bit By Bit* 從社會科學研究出發，整理出十種大數據的特性。首先是數據量。在數位時代，更易取得的龐大數據量是大數據最吸引社會科學研究者的特色之一。Salganik 強調此特色主要在三個面向上有助於社會研究：從數據中找出罕見事件、揭露現象的異質性以及挖掘微小差異。舉例來說，經濟學家 Raj Chetty 與其同事（2014）利用 4000 萬筆稅收紀錄估算世代間流動機率在美國各州的分布。結果發現居住在隔離與收入不平等程度較低，教育資源、社會資本與家庭穩定性相對較高的城鎮，人口向上流動的機率較高。透過大量數據做爲基礎，該研究不只證實世代流動具有地區差異，還細緻

1 關於大數據的「4V」定義，最早可見 Gartner 研究員 Doug Laney 於 2012 年提出來的文章：“Deja VVVu: Others Claiming Gartner’s Construct for Big Data”（https://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/，取用日期：2018 年 7 月 8 日）。更多的中文介紹可見 http://www.statedu.ntu.edu.tw/bigdata/index.asp（取用日期：2018 年 7 月 8 日）。

地找出可能造成地區差異的社會特徵。

大數據的第二個特性是數據流的持續性。一直以來，社會科學家都希望能夠擁有長時間的調查資料，用以追蹤行爲或態度的趨勢變化。這類縱貫型資料（longitudinal data）因爲所需經費不貲，往往一次調查與下一次調查之間需要間隔好幾年。相較之下，社群媒體或物聯網產生的資料是以秒不斷累積。更可貴的是資料會被長期保存下來，研究者不只可以進行即時運算，更可以追溯過去的資料。對追求描繪社會動態的研究者來說，大數據無疑是一個理想的資料來源。

然而大數據並非完美。*Bit By Bit* 最大的貢獻之一，就是指出大數據應用在社會研究上的挑戰。Salganik 強調，由於大部分的數位資料來自商業公司或政府，蒐集目的也不完全爲了研究，如此一來研究者一方面不一定能夠拿到完整數據，二方面數據也不一定具有代表性。舉例來說，微軟研究院研究員 Fernando Diaz 等人（2016）分析 Twitter 資料後指出，線上社群媒體資料應該被視爲一種「不完美的連續追蹤調查」（imperfect continuous panel survey）。在這類調查中，每次參與特定事件討論的群體經常處於變動的狀態，其人口屬性相對於母體（population）來說是有偏誤的（biased）。舉例來說，我們很容易可以想像選舉時期活躍的 Facebook 用戶組成跟世足賽時期的 Facebook 活躍用戶組成會有明顯差異。

另外，雖然許多大數據資料的蒐集過程較少干涉受試者的日常生活，但被動蒐集不等於毫無保留的紀錄。社會科學家能夠取得甚麼樣的大數據資料，很大程度受制於產出資料的技術或演算法。舉例來說，Johan Ugander 與他的同事（2011）蒐集 Facebook 用戶的好友數量後發現有相當高比例集中在 20 這個數字。深入研究之後才發現原來是因爲 Facebook 的演算法會一直鼓勵用戶加好友，直到數量達到 20 的時候才會停止鼓勵。如果 Johan 等人沒有注意到背後的演算法，很可能會做出 Facebook 最佳好友數是 20 的錯誤結論。

最後，*Bit By Bit* 提醒社會科學家使用大數據時，須注意數據是否包含研究者感興趣或符合研究目的的資訊。由於大部分的大數據來自

現成管道，不論是資料蒐集的方式還是資料品質都不一定會符合研究需要。Salganik 以「髒」（dirty）形容大數據的特色，除了呼應本書的主題之一——「混合現成與客製資料的必要性」之外，也提醒社會科學家不要輕易地相信大數據的資料內容，反而更應該謹慎考慮數據潛藏那些偏誤或假資訊。

二、數位時代下的社會研究

在 *Bit By Bit* 提出的四種研究方法中，Salganik 首先討論行爲觀察研究如何受益於數位科技與大數據。Salganik 以經濟學家 Henry Farber（2015）的計程車研究爲例，指出即使只是單純的描述統計，大數據也能提供過去難以得到的洞見。Farber 透過紐約市政府取得計程車乘車紀錄的資料，計算紐約市計程車司機的工時與日薪之間的關係。透過簡單的計算，Farber 發現整體來說計程車司機的載客行爲較符合新古典經濟學的預測，即日薪越高，司機的工時越久，不過此關係會隨著資歷累積後才開始明顯。而對於採取行爲經濟學預測結果，即日薪越高、工時越短的司機，則有比較高的機率會離開計程車這行業。在這個案例裡，乘車紀錄以數位形式即時儲存，使 Farber 可以分析完整的載客行爲資料。

除了利用大數據測試社會科學理論外，Salganik 也以 Google Flu Trends 爲例，指出大數據提供行爲觀察研究的另一種應用：近即時預測（nowcasting）。這個經典案例是由美國疾病管制局（Centers for Disease Control and Prevention, CDC）與 Google 研究員 Jeremy Ginsberg 帶領的團隊合作，結合 2003 年到 2007 年間 Google 使用者搜尋流感相關詞彙的文字數據，以及 CDC 定期從醫院蒐集的流感資料，準確預測 2007 到 2008 年的流感盛行率（Ginsberg et al. 2009）。不過，此案例也凸顯大數據的限制。雖然該計畫在初期的預測率相當高，但後期研究團隊卻發現 Google 搜尋演算法的設計和歷史事件，像是 2010 年 H1N1 流感疫情，會觸發使用者搜尋更多的流感詞彙，使得預測準確

率下降。

大數據也幫助研究者更容易執行類實驗研究。舉例來說，經濟學家 Liran Einav 等人想知道在真實交易情境下商品起始定價對於拍賣結果的影響。由於當中的因果關係受到很多因素的干擾，像是商品性質、賣家評價等等，Einav 使用配對法（matching）鎖定單一賣家同一產品的所有拍賣清單，再分析裡面的不同訂價與拍賣結果的關係。傳統上，此類研究方法很容易受限於可供配對的樣本數。不過 Einav 等人利用 ebay 此數位市場平台成功取得成千上萬的配對清單，使他們能夠更精準的估計當中的因果關係與誤差（Einav et al. 2015）。

在行爲觀察研究的討論中，Salganik 花了不少篇幅介紹現成大數據如何融入到社會研究，但是在訪調方法上，Salganik 將重點轉至抽樣方法論的修正以及現成與客製資料的結合。首先，訪調（survey）在社會科學界已經行之有年，其理論基礎建立在「機率抽樣」（probability sampling）方法論上。然而在數位時代裡，Salganik 認爲訪調應把重心放到「非機率抽樣」（non-probability sampling）上。他以 Xbox 用戶選舉民調爲例（Wang et al. 2015），指出透過統計模型，例如事後分層（post-stratification）與多階層模型（multilevel regression）的方式，即使是代表性不足的樣本一樣可以取得不偏的估計結果。比起回應率日漸低迷的傳統民調，Salganik 相信大數據可以成爲另一項用來蒐集民眾意見的利器。

另外，Salganik 認爲研究者也可以善用數位科技來客製化訪調的資料蒐集。像是利用攜帶智慧型手機紀錄出獄者生活處境（Sugie 2016）的「生態瞬間評估法」（ecological momentary assessment），或是同時結合開放式和封閉式問卷的「維基訪調」（wiki survey）（Salganik and Levy 2015）。過程中，Salganik 強調由於現今各種調查氾濫，如何吸引參加者的眼球成爲訪調能否成功的要素之一。藉由將訪調過程「遊戲化」（gamification），讓參與訪調的受試者享受接受調查的樂趣，也是 *Bit By Bit* 認爲數位時代下社會研究的特色。

訪調也是 *Bit By Bit* 主題中，混合現成與客製資料的一項重要方

法。由於大數據蘊含的資料不一定能夠揭露受試者的內在想法，訪調能夠補足此方面的不足，提供大數據分析時的脈絡資訊。除了受試者的內在想法，對於較敏感的人口屬性如收入，訪調與大數據結合還能夠進行大規模的預測。舉例來說，Joshua Blumenstock 與其同事蒐集 150 萬盧安達人民的行動電話通話紀錄，並且隨機抽樣一千名民眾邀請回答包含社經地位在內的問卷。透過結合訪調結果與通話紀錄，Joshua 透過機器學習模型從通話紀錄的特徵中自動預測用戶的財富狀況。最終該模型成功預測鄉鎮地區的財富分布（Blumenstock et al. 2015）。

雖然行爲觀察與訪調能夠一窺民眾想法與行爲趨勢，要探究當中的因果關係，實驗法是一項相當重要的方法。過去不管是實驗室還是田野實驗，大多會遭遇外推性或執行範圍的限制。*Bit By Bit* 認爲數位實驗至少帶來三項好處。首先是可以招募更多的受試者而不會大幅增加成本，第二是可以取得更完整的實驗前資訊，第三是可以提供長時期的實驗組，並記錄連續性的實驗資料。最有名的數位實驗工具是 Amazon 提供的線上群眾外包平台 Mechanical Turk，此平台可以讓實驗者付費招募受試者。不過，*Bit By Bit* 強調不只有數位平台可以進行實驗，研究者還可以自建環境、產品，或跟商業公司合作。當中，與企業合作可以協助研究者部署更大規模的實驗，但也可能受限於商業利益。例如曾有實驗透過與 Facebook 合作探討社會影響與投票行爲之間的關係。結果在三組介面設計中，其中一組被分配給 98%的受試者，因爲 Facebook 認爲所有用戶都應該使用這組介面（Bond et al. 2012）。

除了觀察、訪調與實驗法，*Bit By Bit* 還提出大眾協作（mass collaboration）作爲結合大數據與數位科技的社會研究方法。大眾協作強調參與者主動加入資料蒐集與分析的過程。在數位時代下，大眾協作的規模可以大到幾十萬人同時在線，使答案的產生更快更精準。其中最常見的協作模式是「人力協同運算」（human computation）。這類方法經常用於需要大量人力才能完成、但不須經過專業訓練的任務，

像是分類圖片、文章等等。然而，如果該任務牽涉較多主觀判斷，則運算結果容易因爲參與者的背景產生偏誤。第二種方法則是目前資料科學領域的主流——「公開徵選」（open call）。這種方法並非由研究者來解答謎題，而是設計一套評估流程，公開徵求群眾提出最佳答案。第三種則是讓群眾直接蒐集資料，這類研究方法在生態學等領域相當常見，像是讓愛鳥者上傳自己的觀察資料到資料庫。

由於大眾協作在社會研究的應用不多，與其他章節比起來，*Bit By Bit* 對大眾協作的討論大多引用天文學、生態學或是都市計畫的案例。Salganik 認爲大眾協作最大的潛力在於結合群眾智慧，具有「民主化」研究過程的潛力。換言之，一般研究者即使沒有高額經費聘任助理或訪員，也能利用大眾協作執行需要大量人力物力的研究。這類研究設計最大的風險是無人自願參加，也因此 Salganik 提醒研究者設計大眾協作研究時，需要多加思考如何提升參與者動機、讓他們能夠專注在提供高品質的資料、適時提供驚喜的同時也注意是否違反研究倫理。

強調倫理在設計研究時扮演的角色，可說是 *Bit By Bit* 的另一項重要特色。在 Salganik 討論的眾多案例中，其中以 Facebook 情緒擴散實驗受到大眾極大的關注（Kramer et al. 2014；Verma 2014）。在這項實驗中，Facebook 研究員刻意刪除 68 萬人的部分動態內容，以驗證當用戶動態牆上的正向或負向情緒用字減少時，用戶的情緒是否會跟著產生變化。這項實驗最大的爭議在於沒有取得受試者的知情同意，也沒有評估對用戶是否會帶來負面影響。Salganik 指出大數據研究最大的倫理隱憂是許多研究經常是在受試者不知情的情況下進行。Salganik 在書中討論幾項設計研究時可以權衡的原則，像是尊重受試者的自主性、平衡利益與風險及兩者的分配，以及考慮法律和公眾利益。

三、社會研究的未來——混雜、趣味、倫理

隨著大數據逐漸成爲重要的資料來源，*Bit By Bit* 提供社會研究者

精彩又務實的研究設計建議。建立在原有的社會研究方法上，Salganik 不只提供結合大數據與數位科技的建議，對於每種研究方法也有清楚易懂的介紹，相當適合作爲研究方法的上課教材。美中不足的地方是，全書只有在訪調一章深入到大數據對方法論的修正，其他方法像是行爲觀察、實驗法以及大眾協作，大部分仍停留在實作層次。對於爲何選擇四種而不是更多種社會研究方法，Salganik 也沒有提出太多解釋。最可惜的地方是，Salganik 雖然提供很多精彩的研究案例和設計建議，但對於未來社會研究可以發展的方向卻只有點到爲止。以下，我將延續 Salganik 的討論，提出幾點目前仍待發展的方向。

首先是貫穿全書的主題之一：混合現成與客製數據。在 *Bit By Bit* 中，Salganik 對該主題最明確的討論是結合大量行爲資料與少數受試者的訪調（Blumenstock et al. 2015）。前者來自現成數據，後者則是針對研究者感興趣的問題設計問卷。然而 *Bit By Bit* 較少討論到：同樣的組合方式是否適用在組合其他方法？像是把現成的行爲觀察數據加上研究者設計的實驗，或是利用大眾協作產出的客製數據結合訪調。更進一步說，大數據是否提供研究者更好的機會混合不同研究方法，是本書較少討論的部分。另一方面，混合現成與客製數據的分析經常需要應用資料科學技術，但對於這類技術如何影響社會研究，本書著墨也不深。事實上，資料科學技術強調預測能力，但許多社會研究更看重解釋能力。*Bit By Bit* 雖然兩者都有提到，但當結合兩種目標不同的技術可能會遭遇甚麼挑戰方面，仍有可以探討的空間。

接著是 Salganik 在結論章提到的「以人爲中心的資料蒐集」。在討論訪調、實驗與大眾協作時，有一個概念重複被提及：讓研究過程變得更有趣。這種以使用者爲中心的設計思考，在過往社會研究中的確較少提及，或經常只從金錢物質酬賞的角度切入。然而，哪些研究議題可以加入娛樂性？如何遊戲化資料蒐集過程而不至於產生研究誤差？雖然 Salganik 認爲趣味化會是未來社會研究的趨勢之一，但如何把這種思維融入研究設計中並沒有交代得很清楚，未來值得繼續發展。

最後，在研究倫理主題上，我認爲*Bit By Bit*對於大數據研究中的跨組織合作問題，仍有細緻發展的空間。由於大數據的發展過程與國家治理還有商業發展密不可分，當研究者透過政府或企業合作取得資料時，如何權衡當中的權力不對等是一個相當重要的問題。舉例來說，資料分析結果若會被政府用來排除弱勢族群，研究者倒底應不應該執行此研究？當研究者的角色介於公司雇員與學術工作者時，我們又該依循甚麼樣的研究倫理，相信都會是未來大數據研究者需要面對的課題。無論如何，*Bit By Bit* 已經爲社會研究者提供一套實用的指引，相信在不久將來，我們會有更多有創意的大數據社會研究，提供我們關於人與社會更豐富的洞見。

參考文獻

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. “Predicting Poverty and Wealth from Mobile Phone Metadata.” *Science* 350(6264): 1073-1076.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. “A 61-Million-Person Experiment in Social Influence and Political Mobilization.” *Nature* 489: 295-298.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. “Where Is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States.” *Quarterly Journal of Economics* 129(4): 1553-1623.

Diaz, Fernando, Michael Gamon, Jake M. Hofman, Emre Kiciman, and David Rothschild. 2016. “Online and Social Media Data as an Imperfect Continuous Panel Survey.” *PLoS ONE* 11(1): e0145406.

Einav, Liran, Theresa Kuchler, Jonathan Levin, and Neel Sundaresan. 2015. “Assessing Sale Strategies in Online Markets Using Matched Listings.” *American Economic Journal: Microeconomics* 7(2): 215-47.

Farber, Henry S. 2015. “Why You Can’t Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers.” *The Quarterly Journal of Economics* 130(4): 1975-2026.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. “Detecting Influenza Epidemics Using Search Engine Query Data.” *Nature* 457: 1012-1015.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. “Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks.” *Proceedings of the National Academy of Sciences* 111(24): 8788-8790.

Salganik, Matthew J. and Karen E. C. Levy. 2015. “Wiki Surveys: Open and Quantifiable Social Data Collection.” *PLoS ONE* 10(5): e0123483.

Sugie, Naomi F. 2016. “Utilizing Smartphones to Study Disadvantaged and Hard-to-Reach Groups.” *Sociological Methods & Research*: 0049124115626176.

Ugander, Johan, Brian Karrer, Lars Backstrom, and Cameron Marlow. 2011. “The Anatomy of the Facebook Social Graph.” arXiv:1111.4503 [Physics], November. http://arxiv.org/abs/1111.4503.

Verma, Inder M. 2014. “Editorial Expression of Concern: Experimental Evidence of Massivescale Emotional Contagion through Social Networks.” *Proceedings of the National Academy of Sciences* 111(29): 10779.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. “Forecasting Elections with Non-Representative Polls.” *International Journal of Forecasting* 31 (3): 980-991.