許輝煌VS.林志娟 揭巨量資料面紗

一流讀書人

文字／黃怡玲、王心妤整理
攝影／黃國恩、張令宜

資工系系主任許輝煌、統計系系主任林志娟　揭巨量資料面紗

在即將邁入2015年之際，大數據正方興未艾，其中的發展持續應用於各行各業，開始蔓延並改變你我的生活。由於全球網路應用在各行業當中，讓生活上的各種數位化資訊取得便利，所累積的海量數據資料開始被發現進而發展，逐步形成「大數據」概念，更突顯大數據分析的重要性。

本書從資訊本質中分析巨量資料所帶來的影響，提供我們在資訊化過程中的新見解。因此，本刊特邀資工系系主任許輝煌、統計系系主任林志娟，針對本書所提到的大數據趨勢觀點進行對談，不僅探討大數據趨勢發展，兩人以自身學術專長解構大數據的關鍵內容，以前瞻對該趨勢的深刻了解，並提及在此
洪流下所需具備的職場能力，以
掌握未來趨勢。

何謂大數據

大數據（Big data或Megadata）被稱為或稱巨量資料、海量資料，指的是各類的資料量規模已巨大到無法在合理時間內，以人工或軟體工具對其內容進行處理、管理、並整理出對人類有用的訊息，它包含「數據量龐大」、「數據量種類繁多」、「處理速度快」，以及「價值密度低」四個特徵。因此，如何從各類型數據中快速獲得有價值的資訊，需要運用到相關的軟硬體設備與分析技術，如MPP資料庫、統計分析、資料探勘、大型存儲系統等。

書名：大數據
作者：麥爾荀伯格、庫基耶
譯者：林俊宏
出版社：天下遠見出版公司
索書號：551.49／8836 102

記者：對於《大數據》書中所談的趨勢，兩位對談人是否都感受到了？從該書封面引

言的「這本書將會全面革新我們的思考方式」一句，兩位看法如何？

統計系系主任林志娟（以下簡稱林）：大數據」現已成流行語，曾在刊物看到描寫大數據現況，與大家分享：「每個人都在談論這件事情，並不是每個人都知道該怎麼做。他們認為自己之外的每個人都在進行這件事情，所以每個人都選擇自己正在做這件事情。」

我想這句話正好描述現況，因為大部分的人希望和Big Data有連結，在這樣趨勢下每個人都無法缺席，需要各式各樣的人才，如數據蒐集、整理分析等，大家都能參與其中。以統計系來看是以資料分析與整理為主，近幾年因這樣的趨勢，讓許多業者都找我們提供協助資料分析，例如健保局與統計系合作，將健保資料建立有效數據庫。我認為，大數據最重要的是「商業價值」，如何在有限的資料中找出它的價值，我想才是大數據的意義。

資工系系主任許輝煌（以下簡稱許）：我認同林主任所說大數據所帶來的「商業價值」。以資工系來說，我們的背景為計算機資訊科學，20年前已有資料探勘，涵蓋機器學習、人工智慧、圖書館學資訊分析等技術應用，但近些年發現，政府部門與企業都希望由資料探勘中去擷取有用的資訊，從中挖掘有用的資訊發揮價值。

Big Data的發展主因在於雲端科技技術的成熟，和從前相較可儲存大量數據並加以處理，當資料量太大，無法用資料探勘的方式建立模型來加以描述時，就會發展新技術進而應用；這也是書中所探討的，在擁有大量數據及雲端計算技術提升情況下，如何充分利用或解析資料。目前政府和企業開始感知並關注此趨勢的來臨，紛紛成立相關部門因應。之前校友向母校尋求合作，希望將每通客服電話內的資訊成立資料庫，讓電話內容不只是一通電話，而可以帶給企業更大的「商業價值」並為企業爭取最大的利潤。

記者：書中提到，巨量資料應用下可發揮預測功能，如預測疾病發展走向而預防等觀點，兩位的看法如何？

林：我不盡然同意書中的觀點，因為「預測」是有其關聯性，這些相關性會取決問題的設定，才能利用科學的統計方法對事物的未來發展進行定量推測，並計算概率置信區間，因此會強調「因果關係」。但在商業模式考量下，為因應快速商業變遷的環境，或許以相關性就能發展出致勝的商業模式，以購物網站推薦商品為例，它是根據消費者過往的消費紀錄、瀏覽商品歷程等數據，找出相關性來推薦可能合適的商品；但卻無法確切得知消費者購買商品的真正原因，或許是因為技術尚未成熟無法得知其中的因果關係。

我認為，醫藥或生命科學研究中，不僅是會以相關性研究為主，也會了解導致疾病蔓延的因果關係等內容，所以才能提出預防措施防治疾病，舉例來說，天氣熱讓飲料和

小兒麻痺的指數呈正相關，這是只有關聯性，但真正導致小兒麻痺人數上升是因天氣炎熱讓買飲料人數上升，也讓病毒成長，所以就會有科學家以這樣的數據、資料分析、研究方法等了解其相關性、因果關係等，開發出適合的藥品。所以「追根究柢」應該是要看想要解決甚麼問題後，再來尋求所應用的方式。

許：傳統的資訊學習上，有些技術應用不太會注意相關性或因果關係，但會關注資料中的「INPUT」與「OUTPUT」的關聯性，有些應用面如同林主任所提因果關係並不是太重要，也是這本書想強調的重點，就是大家太注重前因後果，會阻礙進步，因為目前有許多應用也是順其自然的發展，並沒有太大的原因。

書中有提到，應用Google的搜尋工程在2009 年中預測出流感的爆發，而今年最火紅的議題就是「伊波拉病毒」，日前在加拿大有間公司就是以大數據分析，對該病毒做擴散的分析，很多國家或世界衛生組織都去請該公司幫忙，我想結果就算不是百分之百正確，依然是有很高的可信度。從此例看出，他們可以從數據中分析，病毒可能會擴散地點、散播途徑、擴散時間等等，這樣就有機會事前防堵。所以，在大數據分析中，除了有大量資訊外，一定還要有各領域的專家共同加入，才會讓研究有意義。大數據一定要有過往的資料及相關性才能預測未來發展。

記者：知名女星安潔莉娜因基因檢測後而施行手術，這是大數據應用嗎？

許：我認為這個例子不能完全說是大數據的應用，比較像是基因學下的產物。她會選擇切除乳房，是從家族病史中研究顯示有高達87% 的機率會罹患乳癌。這與大數據應用並沒有太大的關聯，比較是個人選擇部分。

林：我也贊同許主任的說法，「要使用大數據就要先有全部完整的資料」，若提到在生物學上的應用，我認為可以用在蒐集基因資料，因為基因數量太龐大了，用人工選取方式需要花費相當大的資源，如果我們能利用大數據作為輔助，相信可節省大量的時間。

記者：書中第三章提到：「從精確走向可能性」，請問這方面來說有容許錯誤的空間嗎？

許：其實我對於這部分還是抱持著疑問，但是書中第三章提到要「擁抱不精確」，以前我們需要把錯誤挑出來，但現在我們需要接受它，就算擁有錯誤還是可以從中得到有用的資訊。我認為，書中所想要闡釋的是，在如此龐大的資料量下，若挑出錯誤必須要花費相當大的精神。以實用面來看，如果我們能「擁抱不精確」是不是得出一個「還可以用」的結論就夠了呢？這對我而言仍是思考中的議題。

林：我也同意許主任所說，做學問最重要的是務實，在學校我們告訴學生的是歷經多位學者專家驗證後的理論，一出社會後卻是要馬上上手的技能面。所以我也在思考「我們真的需要Big Data嗎？」因為，蒐集資料和分析資料就可應用各種統計方法以達

到精確，對統計系來說，在統計過程看到的「差異」並不是錯誤或瑕疵，反而會促成研究議題的產生，所以不應該將「錯誤」視為不對或忽視，而是要正視它進而解決它。

記者：民調中心的調查與和網站上的新聞投票意見調查，兩者相較之下，如何相信是可信度的？

林：我覺得這要考慮的最大因素是資料蒐集的來源，因為是利用網路，我們無法得知受訪者是否是誠實。假設受訪者皆誠實，我們就可以利用大數據去研究兩者關係，也許是因為候選人的政見，而使女性受訪者有較高的意願選擇該位候選人，如果可以找出政見與選民意願的關係，候選人就能藉此修正政見內容進而獲選，我想即是大數據的應用。

許：對我來說，奇摩上的投票可信度並不高，因為他們在取樣上比較輕率，不能蒐集到各年齡層的資訊，僅能做為參考。但我相信，如果有學者要用學術性的方式討論投票率，結果是可以相當精確的。

記者：書中特闢一章討論資訊倫理和隱私權，兩位的看法如何？

林：我想，書中一語「Data可以說是未來的石油，誰掌握Data就是下一個石油大亨」就是這樣的概念，舉Google為例，常常我們瀏覽網頁時，廣告訊息出現的類型因人而異，Google 根據使用者的瀏覽歷程所提供的廣告訊息。我覺得，這如同書中最後提到的，我們只能依靠個人素養和道德，雖然我們還是會使用，但是人性就變得至關重要。

許：我也贊成林主任所說，舉Facebook來說，其實有簡易的隱私權設定，但大家常常忽略，所以最終回到個人上的使用。隱私權的權益不能全權交由網站為我們把關，自己在使用科技產品上也需要多留心，呼籲大家多了解網站隱私權應用。

記者：對兩位而言，認為《大數據》中最有挑戰性、突破原有觀念議題為何？為甚麼？

林：對於相關性這個部分有些疑慮，但也不完全推崇因果關係。我認為比較難突破的部分是，大數據發展快速憂心學校的教育是否能跟上趨勢，希望教導學生「質」的因應，如何讓學生養成大數據的思維及找答案的能力的培養勢在必行。

「擁抱不精確」中，反思「我們真的需要Big Data嗎？」而當中提到的「容許誤差」是在統計當中「相信自己的能力無法做到完美」，因無法掌握而允許容錯。

許：我覺得有趣的議題是第二章「樣本等於母體」、第三章「擁抱不精確」，這些都是在挑戰傳統的統計方法，而「樣本等於母體」未來是有機會實現的，是因為資訊發展及雲端儲存的技術進步。在「擁抱不精確」探討的是資料量夠多夠大時，其實不需要太精確也是能從中得到有價值的資訊。

記者：那在大數據時代的趨勢下，兩位老師在教學方面是否有些改變？期待帶給學生甚麼樣的視野？

林：臺灣在以往沒有數據分析師證照，於今年開始引進，誰能在大數據這個新的趨勢當中找到商機與關鍵，將帶動領域的潮流。因此大數據的出現是危機也是轉機，希望讓學生在不同階段及領域的整合，培養跨領域的溝通能力，達成共同的目標，畢竟科技的運用最終還是要由人類來主導，希望多培養資訊的倫理。

許：大數據時代中，演算法也可以帶動數據分析的部分，尤其雲端技術。對資工系而言，數據資料的處理及運用是最基本的，許多相關的研究正大量投入，預期有許多新技術正持續開發，目前也開設雲端課程。在此希望學生勿忘「謙卑與人性」，這是學科技的人永遠不能忘記的。

大數據時代 職場必備

許輝煌說：

曾有人提出成為資料工程師必備28項能力，這當中與電腦相關如資料庫、資料結構、程式語言等課程在本系皆有開課，相信學生是可因應這股趨勢的。無論任何科系要學好系上開辦的專業科目都是有難度的，而且也不必每個人都成為資料工程師，所以同學們在面對大數據時代所要做好的準備，就是要有「溝通的能力」，尤其是對各行業和專業科目，都要有基本認識和學業專精，才能發揮所學專長。

林志娟說：

現在談「跨領域」人才，尤其大數據分析更需要資訊科學、統計分析以及相關專業領域知識的整合。統計系現有開辦「數據分析師」專業證照研習課程，也有統計分析相關課程來輔導學生與大數據時代接軌，更重要的是需具備「溝通」能力，以表達自己的專業能力和整合客戶的需求。因此我認為擁有「溝通」與「整合」的能力，是未來必備的能力。

《大數據》目錄簡介

1. 現在 該讓巨量資料說話了
2. 更多資料 「樣本=母體」的時代來臨
3. 雜亂 擁抱不精確，宏觀新世界
4. 相關性 不再拘泥於因果關係
5. 資料化 當一切成為資料，用途無窮無盡
6. 價值 不在乎擁有，只在乎充分運用
7. 蘊涵 資料價值鏈的三個環節
8. 風險 巨量資料也有黑暗面
9. 管控 打破巨量資料的黑盒子

10.未來 巨量資料只是工具，勿忘謙卑與人性

對談人響應書中話語

許輝煌呼應書中P104：

第五章「資料化」：當一切資料化，用途無窮無盡。我認為同學必須了解大數據的可能性為何，就能從分析的技術，得到用途。

林志娟呼應書中P172：

第六章「價值」：不在乎擁有，只在乎充分運用。我想鼓勵同學，要找出Data的價值，從不同的角度去觀看。

淡江未來大數據應用

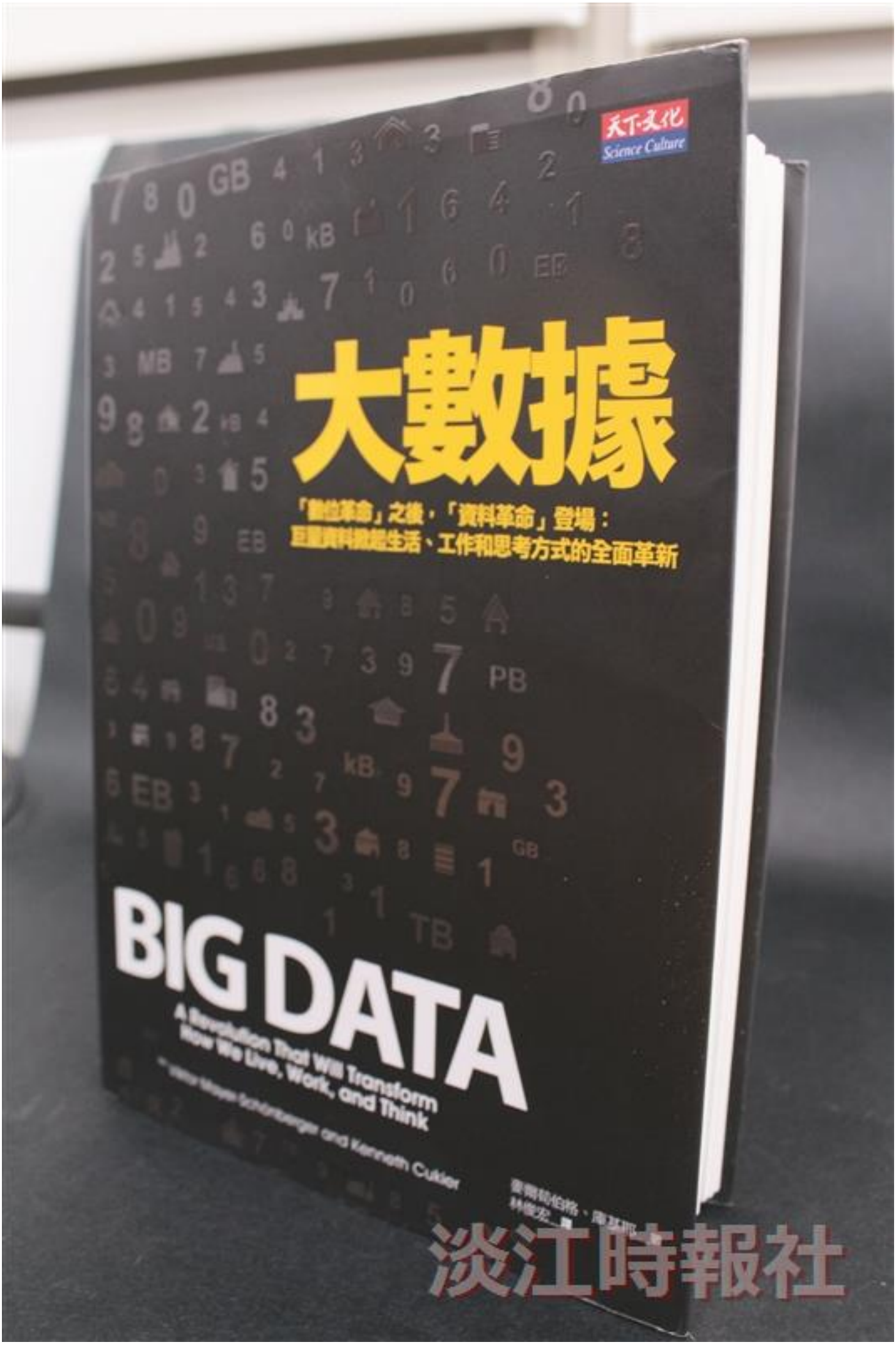
許輝煌建議：可運用於「改善教學」，這可透過教學評鑑的資料，亦或加入全班平均及歷年該科被當人數的數據、學生課堂表現等分析，將會更具參考價值。

還有「招生部分」，可更進一步分析學校各科系特色，以及學生背景，進而加強重點學校的宣傳。

林志娟建議：本校已有24萬校友可充分運用他們的課堂成績、在學表現等分析其資料，以加強與校友互動，前提是不能違法個資法。還可將校外租賃調查資料數位化，以即時分析並得知學生在外租屋的品質。另外，可持續累積校內各類型資料量後並加以分析運用。


VIVA BIG DATA
錢江時報社


淡江時報社


天下文化
Science Culture
大數據
「數位革命」之後，「資料革命」登場：
巨量資料徹底改變生活、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
淡江時報社