大數據與演算法的虛像與實相

摘要

本文檢視國內外大數據與社會（科）學的研究文獻，歸納整理其共同點、特質與問題。本文的論述重點有三：一是大數據、小數據、社會大數據的定義與概念內涵，二是大數據在人文與社會科學的知識論與方法論問題，三是演算法的迷思。西方學者對這三個研究軸線均有頗多批判性的探討與反思，但台灣的人文、社會科學，乃至於法律學界，在這三個層面的研究仍有討論與改善的空間。唯有透過批判的思考與對話，公共政策的研擬與執行才能有紮實的知識論與方法論基礎。

關鍵字

大數據、演算法、知識論、方法論

Abstract

The paper reviews the journal papers and book chapters published in Taiwan, trying to depict their features and problems. There are three points in the paper. First, the definitions and conceptual implications of big data, small data, and big social data are explored. Second, the epistemological and methodological problems are considered. Finally, the myths of algorithms are debunked by demonstrating the built-in errors accompanied with and limitations of algorithms. The studies of big data in Taiwan and by Taiwanese scholars might as well focus on the three issue areas. Only then can we establish solid epistemological, methodological and theoretical foundations for the making of public policies.

Keyword

big data; algorithm; epistemology; methodology

壹、導言

大數據、巨量資料、海量資料（big data）是近年來最時尚的名詞或概念。不論是財經、資訊科技的雜誌，還是工科、商管、社會科學、人文、法律類的學術專書與期刊論文，大數據或 big data、資料科學（data science）等相關概念、主題的研究數量急遽增加。從事數據分析的業者與軟體系統、解決方案也迅速湧入市場。然而，財經商管雜誌與暢銷書，以及實務應用導向的學界，對大數據這個概念的定義與內涵，卻常有望文生義或因人而異的模糊理解等問題（Halavais, 2015: 584）。國內對大數據的定義也常流於浮面，鮮少深入探討大數據概念的定義、內涵與相關的理論、知識論、方法論議題。

大數據的研究浪潮與誇詞（hype）常讓人以為大數據無所不能，卻忽略大數據牽涉到的知識論、方法論與批判的理論問題。西方學者致力於釐清大數據的知識論、方法論與批判的論著不少，但從筆者在台灣蒐集與閱讀到的期刊論文與專書論文來看，認真探討大數據知識論、方法論與批判的論證者，非常少見。這種趕熱潮的現象，與之前的文創產業、文化節慶論文大量出籠的狀況，如出一轍，似乎反映出台灣學術界某種特殊的結構與生態。

有鑑於此，本文蒐集國內以大數據為主題或關鍵字詞的學術期刊論文與論文集，刻劃其共同點與問題，主張相關的研究能夠跳脫浮光掠影式的概念定義，認真思考大數據與演算法的知識論、方法論問題，提出批判的理論觀點與研究，希望能推動各個研究領域大數據研究的深度，包括公共政策的研究在內。

本文結構安排如下。第一節是研究動機與研究目的。第二節回顧國內期刊文獻以大數據為主題或關鍵字的期刊論文與專書論文，綜合整理其共同點與問題。第三節探討大數據、小數據（small data）與社會大數據、社群大數據（big social

data）的概念內涵與牽涉到的方法問題。第四節澄清有關演算法的迷思，特別是偏誤與局限。第五節結論，綜合本文論點，提出研究展望。

貳、國內文獻回顧

目前國內的大數據研究（期刊論文與專書論文）大致可分為概論、法律案例探討、社會科學的案例研究、實務應用等。

基本概念與議程的探討包括「向運算轉」的研究典範轉移、「計算社會學」（computational sociology）的介紹（江彥生、陳昇瑋，2016）、大數據與智慧決策（劉玉山，2016）等。

法律的論文以個資、隱私、去識別化等議題與案例探討為主，包括大數據時代對隱私權的影響與可能對策（許炳華，2016）；醫療照護與隱私保護（劉宜君，2016）；大數據與個資保護之法院判例分析（葉志良，2016）；大數據的個資規範與法律保護議題（劉靜怡，2017；葉志良，2017）；介紹歐盟與各國（美國、英國、日本、韓國、中華民國）個人資料隱私與去識別化的法規、資訊技術與保障原則（劉宜君，2017）；大數據產業的資料隱私問題與對策（丘昌泰、劉宜君，2017）；國際上因應大數據與物聯網對個資自主權利的法制修改途徑（劉定基，2017）；台灣個資法架構之下巨量資料的適法性法律風險（彭金隆等，2017）；日本的大數據與個資權利探討（范姜真媺，2017）；隱私權保護的法理探討（黃章令，2018）。

社會科學的案例研究以文本分析為主，在政治學界有 2012 年總統大選社交媒體的繁體與簡體字文本分析歸納出來的三種社群（鄭宇君、陳百齡，2012）；OpView 應用分析九二共識在台灣網路輿論中的趨勢、正負情緒比、關鍵字（杜聖聰等，2016）；統獨的網路聲量分析（劉嘉薇，2017）；習近平時代的科技威

權主義（王信賢，2018）等。社會（學）研究有宗教與多元性別網路輿情的情緒正負關係（楊嘵智，2015）；宗教與同運議題網路聲量與正負情緒分析（2013-2017 年）（楊嘵智，2018）。傳播研究的論文包括以大數據研究收視行為、大數據與傳統收視率調查的結合運用（賴祥蔚，2015）；英文傳播類核心學術期刊的大數據分析（作者一人或多人、主題、文章數量等）（江亦瑄、林翠娟，2015）；社交媒體使用者的虛實整合研究與社群資料分析（鄭宇君，2015）。

在公共行政、公共政策的研究，概念與一般討論包括我國大數據政策推動現況（鐘嘉德等，2015）、巨量資料分析與循證政府決策（陳敦源等，2015）；大數據與行政管理決策及思維的初論（莊文忠，2016）；資料驅動創新與跨域公共治理的議題討論（蕭乃沂、朱斌妤，2018）。主計方面有大數據挖掘主計資料金礦的討論（丘昌泰，2016）、主計資料的創新應用（鍾永淙，2017），但這些都是一般的討論，鮮少具體的理論架構或研究案例。

實務案例的討論包括交通運輸，如高速公路易甕塞路段分析（郭昌儒，2015）、台北捷運人潮移動的大數據分析（葉奕新，2017）、悠遊卡租賃 YouBike 與轉乘捷運行為（鍾智林、李舒媛，2018）等。經濟與傷商業類研究則有台灣數位音樂串流服務臉書粉絲專頁文本分析（余峰偉、王嵩音，2018）、數據經濟的問題（羅鈺珊，2018）。

犯罪防治與偵查方面有大數據發展趨勢與防罪防治（許華孚、吳吉裕，2015）、大數據運用於犯罪偵查（甘炎民等，2015）、大數據與家暴防治（韋愛梅，2015）、大數據探勘與監視器資料視覺化（江守寰，2016）等。

綜上所述，社會科學與法學界的大數據研究，可以歸納出幾個特徵。首先，多數論文不太注重大數據的定義問題與概念內涵，通常是用數量超大（volume）、

速度超快（velocity）、種類特別多樣（variety）等幾個 V 或特性就帶過去（葉志良，2016：3；劉宜君，2016b：236-238；丘昌泰、劉宜君，2017：32）。有的說大數據是典型資料庫與軟硬體工具無力處理的數據或資料數量（劉宜君，2016a：3；劉宜君，2017: 3-4；劉定基，2017: 267；劉嘉薇，2017：122）。少數比較深入的會引用國外的定義（帶入多樣、複雜 X 軸與位元組次方單位 Y 軸，以及企業資源系統、客戶關係管理系統、網路、大數據等維度），定義大數據等於交易資料、加上互動資料與觀察資料（許華孚、吳吉裕，2015）。

其次，前述論文多半不是很講究大數據的定義與內涵，它們所探討的議題，包括個資與隱私保護、資訊控制與政治（民主化、社會運動）等，在小數據、非大數據的發展之中也是重要且研究多年的議題。以監視與隱私為例，網際網路、虛擬社群、智慧型手機、手機與行動上網的研究所在多有，那麼大數據與個資、隱私、監視之間的關係是什麼？對這些議題與現象造成什麼質與量的重大改變？從前述的論文來看，多半只是假設大數據的量變會造成質變，卻鮮少深究這種量變造成質變的機制與過程為何。同樣的，家暴、犯罪防治等議題，即使在「大數據時代」來臨之前，各機關資料庫的連結互通與綜合運算處理，也已行之有年。大數據對家暴與犯罪偵防的影響，與非大數據有何不同，這些論文多半也未深究。正如 Burns and Thatcher（2015: 284）強調的，大數據已經，也將持續改變隱私在個人與社會層次的意涵。隱私的概念需要針對國家機關、企業與非營利組織存取（access）大數據、分析大數據與一連串相關行動的問題，重新概念化，包括法律論述、法律案例應用解釋、倫理規範等。相形之下，國內對這些理論、倫理議題的研究相對較少，多半還是概念探討與單點的案例應用分析。

國內這些論文探討大數據知識論、方法論與理論議題的著作，也不多見。少數例外有鄭宇君（2014）探討「向運算轉」的趨勢，提到鉅量資料的研究侷限與

倫理議題，算是國內最早聚焦於方法論議題的著作。賴祥蔚（2015）比較傳統收視率調查與大數據收視行為調查的差異，以及兩者的結合，在結論中順帶提到大數據研究收視行為的方法論局限。江彥生、陳昇瑋（2016: 189-193）簡介「計算社會學」的論文當中比較電腦模擬、實驗式與非實驗式大數據研究的方法論，算是開啟國內相關探討的論文。劉嘉薇（2017）在網路統獨聲量研究的論文中提到信度與效度、相關性與因果關係、代表性檢定、母體與樣本、「潛水者」與不表態者等議題，也強調網路聲量調查與實體（電話）問卷調查彼此相輔相成，各擅勝場，如前者可以觀察到較多年輕人的表態、網路匿名有助於研究對象勇於發表「真實的」意見。但該文是在研究主題之外附帶討論，並非專門針對知識論、方法論議題而立論。整體而言，這些作者觸及大數據的方法論局限，但並未進一步深究大數據的知識論問題。

再者，我們可以看出來，前述論文多半是概念的淺論與實務案例的研究，很少有具體的政治、社會-文化理論作為研究的立論基礎或參考架構。沒有既定的或深刻的理論觀點，這麼多大數據的期刊論文與專書論文能把我們帶到哪裡去？我們對大數據研究是否能有深入的理解與反思？大數據研究的熱潮與宣稱的實務應用，與前大數據年代（pre-Big Data era）的理論與經驗性研究在質與量上有何重大差異？或者說大數據與非大數據研究彼此是否能夠、需要相互結合？這些問題，我們在國內既有的論著當中看不到討論的蹤跡或線索。

參、大數據、小數據、社會大數據

大數據的名詞或概念從何而來，國內外說法不一。國內的期刊論文有說最早由 IBM 提出，麥肯錫顧問公司（McKinsey & Company）提出大數據專題研究報告（甘炎民等，2015）。有的說麥肯錫最早提出（韋愛梅，2015）。有的說是麥

肯錫先指出數據的重要性，世界經濟論壇（World Economic Forum）提出名稱帶有大數據的研究報告,《紐約時報》專欄認為大數據時代已經來臨(劉宜君,2016a: 2-3；劉宜君，2016b：235）。有的則說是顧能諮詢公司（Gartner Analysis）最先提出 3V 定義（high-volume、high-velocity、high-variety）（許炳華，2016：133）。也有人說大數據之稱不脛而走，乃因牛津大學教授 Viktor Mayer-Schönberger 與《經濟學人》資深編輯 Kenneth Cukier 於 2013 年出版的書籍《大數據》（Big Data: A Revolution That Will Transform How We Live, Work, and Think）所致（丘昌泰、劉宜君，2017：31）。

大數據最早是誰提出來的問題，不一定那麼重要，重要的是大數據的定義與概念內涵。首先,「大」要多大才算大？是整個推特（Twitter）或臉書（Facebook）貼文的文本內容，還是某些社群的組合？天文學、物理學研究的大數據，數量遠非社群網站能比。所謂「大」，似乎沒有一定的標準。基本的統計原理與長期的社會（科）學統計發展歷史，已經告訴我們：樣本到達一定的數量（N=1068?），即足夠學者從事推論統計。儘管會有誤差，但只要研究討論注意到誤差造成的問題與局限，研究就有其價值。即使樣本大到 10,680，得到的統計資料 1,068 樣本數所得完全相反的機率，恐怕極小。那麼，超大數量資料的意義在哪裡（Halavais, 2015: 585）？如果「大」是指資料數量門檻達到拍位元組（Petabyte, 1024 TB），那麼這個標準的理論依據何在？[1] 隨著電子商務、Fintech、智慧製造或工業 4.0、

[1] 有的學者認為，從社會科學的內容分析（content analysis）來看，社會大數據的「大」要看編碼員判定分析單元需要花的時間。這有兩種判準，一是複雜的問題與大型文件的資料組超過 1 萬，可謂「大」。二是小型文件，編碼一個單位所需時間較少的話，超過 10 萬的資料組，對手工處理來說，可謂「大」（Guo *et al*., 2016: 333）。

物聯網、穿戴式裝置、雲端運算、資料壓縮與儲存技術的迅速發展，拍位元組作為大數據資料量門檻的說法，恐怕更讓人懷疑是否合宜。

Symons and Alvarado（2016: 3-5）指出，大數據的概念與定義要回到當初出現的脈絡：巨量的資料與相對有限的電腦儲存容量與運算能力。早在 1990 年代，美國航太總署研究中心（Ames Research Center）面對的是巨量資料視覺化對電腦能力的挑戰。超過 100 GB 以上的資料量，對當時的桌上型電腦、硬碟容量與外接硬碟、磁帶，都是無法處理的負擔。我們現在習以為常的雲端運算與 USB 工具、技術，在 1990 年代已有相關的概念與技術（分散式運算 distributed computing、網格運算 grid computing），但雲端技術與服務在當時尚未成熟。時至今日，無論是學者、工程專家，還是一般人的個人電腦、筆電，還是行動硬碟、USB 等儲存工具，容量與運算能力均遠超過 1990 年代的學者專家與電腦用戶。今日大數據數量的標準與當時已不可同日而語，現在所說的「大」，在十年或二十年後還會算「大」嗎？如果這個「大」是隨著時間、技術演進而調整與改變的，那麼大數據的「大」字就不是那麼容易定義或定著的形容詞。既然「大」是隨著時間與技術而變動的，資料地景或資料本體（data ontology）是動態的，那麼研究者可以取得所有母體的論斷就是不確實的，抽樣偏誤在大數據中一樣會發生。

另一方面，人文學與社會科學家所做的大數據分析，其實應該稱作社會大數據或社群大數據（big social data），資料與數據通常是來自社群媒體、社群網站、網路搜尋，或是政府開放資料（open data）等資料。比起天文學、氣象學、生醫與基因研究等學科與領域的資料，社會大數據的規模與數量，不一定那麼大，更沒那麼全面。平台業者不是慈善團體或人道組織，他們釋出資料，必然有商業考量，重要的是營收與利潤。研究者或許可以無償獲得一點資料做分析，但他們不可能獲得全部資料，也就是研究資料沒那麼「大」。如果要拿到多一點資料，勢

必要找到研究補助經費，但這筆經費通常不是小數目。要拿到這麼大的經費，若不是要研究非常傑出，可以獲得審查者的青睞，就得有點「辦法」或手腕。換言之，我們面對大數據的熱潮，應用大數據分析，還要注意到新數位落差（new digital divide）或大數據落差（Big Data Divide）的問題，也就是學者與學者之間，大學與大學之間研究大數據的存取資源落差，運算技能的問題（哪些學科比較有能力操作應用程式介面 application programming interface, API？）。研究選題的空間也有限，不能冒犯到平台業者，因為資料是他們的「財產」，平台業者若認為研究題目影響到他們的聲譽，不太可能無償或有償提供給學者做研究（boyd and Crawford, 2012: 673-674）。

大數據的崛起與流行，並不表示小數據（small data）就此完全沒有價值（boyd and Crawford, 2012: 670）。小數據也可以與大數據結合，彼此截長補短，依據不同的尺度（scales），研究學者關心的現象與議題（Kitchin and Lauriault, 2015: 464-465）。大數據研究不是要取代小數據研究，或者說目前還不能取代小數據；大數據是另一種觀看、理解世界的方式。小數據分析仍有其獨特的價值，很多小數據探討的問題是大數據這種匯集的資料無法處理的，如人類學民族誌、社會學符號互動論與俗民方法論、扎根理論經常研究的微觀情境與場域（安寧病房、特定團體的社會空間）。小數據常含有大數據沒有的資訊，可以與大數據在具體研究案例上相互結合（Lazer *et al*., 2014），如學者結合大數據與線上問卷調查，分析臉書用戶收發政治新聞的過程與模式（Wells and Thorson, 2017）。

大數據也不能取代理論，理論包括概念、命題、預設等，有助於引導研究方向，聚焦研究議題，也可以用資料驗證、否證或修正、重建。理論與大數據之間是演繹法、歸納法的辯證過程。純粹歸納即可得到相關、模型與知識的推論或看法，恐怕站不住腳（Burns and Thatcher, 2015: 446-447）。沒有理論的聚焦與引

導，我們可能只會得到許多零碎的經驗性研究，看不到整體的政治、社會與經濟圖像、機制與結構。國內外社會（科）學與人文學研究都流行傅科式的研究（Foucauldian research），經常參考他的著作文本，論述監視（surveillance）、規訓（disciplinary）、知識與權力、生物政治（biopolitics）與生物權力（biopower）、身體的控制、知識與權力對主體性（subjectivities）的構成等。國外多有學者援引傅科的概念與理論，對大數據與演算法提出批判的觀點與分析，但國內在這些主題領域的貢獻卻仍付諸闕如。研究大數據與個資的論文多為法律應用與公共政策的研究，但似乎就此止步，未再進一步往理論化的方向走。理論與政策、理論與實務並非完全對立，一個完整的理論，可以作為我們研議、擬定、執行政策的參考。沒有長足的、深入的理論化工作，政策制定與分析可能會趨向單點的、零碎的案例分析與比較。

大數據、小數據、社會大數據的概念定義，不只是操作型定義的問題，更牽涉到知識論與方法論的問題。大數據分析往往預設一種素樸的實證論（Positivism）或邏輯實證論（Logic Positivism）。問題是在科學哲學與社會（科）學的歷史上，批判實證論與邏輯實證論的學者所在多有，包括批判理論陣營的阿多諾（Theodor Adorno）等人與維也納學圈（Vienne Circle）、邏輯實證論之間曾展開長期的辯論。與維也納學圈相熟的波普（Karl Popper），也質疑歸納法，提出可否證性（falsifiability）與否證（falsification）的概念。波普的立論與（邏輯）實證論驗證原則（verification）的立論頗有區別（Rosenberg, 2000: 120-122），我們的大數據研究卻仍停留在（邏輯）實證論的知識論與方法論界域之內，缺乏批判的觀點與反思。

哈伯瑪斯（Jürgen Habermas）這位第二代法蘭克福學派的學者，也提出三種知識構成的興趣（knowledge-constitutive interests）與邏輯-方法論規則（logic-

methodological rules）。經驗-分析的科學途徑（empirical-analytic sciences）包含技術的認知興趣（technical cognitive interest），以命題的假設-演繹與實驗法，建構可驗證的理論、法則或定律，以及可預測的知識，自然科學與社會行動的系統化科學（systematical science of social action，包括經濟學、社會學、政治學），均屬此類。歷史-詮釋的科學（historical-hermeneutic sciences）基於實際的認知興趣（practical cognitive interest），主要方法是文本詮釋（interpretation），促成互為主體所達成的共識與意義的理解。批判取向的科學（critically oriented sciences）基於解放的認知興趣（emancipatory cognitive interest），批判的社會科學與意識型態的批判，旨在超越未經反思的法則或定律知識，在解放的認知興趣之下採用自我反思（self-reflection），將主體從依賴假定某些實在性的權力（hypostatized power）之下解放出來（Habermas, 1972: 307-311）。

由此觀之，新聞媒體中報導的大數據分析給人的意象（或無所不見、無所不知的誇詞），國內期刊論文中大數據的概念探討（或歡呼）與實務案例分析，可說是偏向經驗-分析的科學途徑與技術興趣。但這種分析無法幫助我們探索其他兩種知識興趣、知識類型與相關方法論，甚至限縮或化約我們對人文與社會世界的知識與理解。大數據分析預設的知識是實證論與經驗主義所建構的知識，忽略知識有很多種面向。知識的來源不只是網路、感測器、刷卡資料、物聯網、穿戴式裝置。在特定制度、機構、情境中的日常生活與生活世界（life-world）中的訊息、資料與知識，多非現有大數據分析的管道與技術能夠處理。大數據分析到的不一定是個整全的個體資料，而是數據構成的資料主體（data subjects），這種資料主體性（data subjectivity）不完全能還原為肉體的個人與個體性（Matzner, 2017: 44）。

再者，大數據的「數據」預設的是數值資料與量化，無法或不容易處理非量化資料或材料、文本的意義與意涵。我們在國內既有研究上看到的網絡資料或網絡關係圖、情緒分析（正負情緒比）、網路聲量分析（文字雲、座標軸等），其實是非常粗略的側寫（profiling），分析的廣度是有，但深度頗有商榷餘地。在社會情境之中，情緒是很複雜的一種社會表達與社會互動，很難用正負情緒比明確區分。網路聲量可以指涉一個人或一件事在網路上的熱度，都不一定等於集體意志、情緒的表達，與街頭巷尾、人際關係網絡內部的議論也不完全相同。文本分析的意義解讀有很多層面與複雜精微之處。同樣的文本，不同人的詮釋，得到的意義理解可能大異其趣。目前的演算法、人工智能（artificial intelligence, AI），抑或是監督與非監督式的機器學習（supervised and unsupervised machine learning），似乎還沒有辦法處理文字資料中的多層次意義，包括隱喻、反諷、反話、幽默、戲謔、賞析等。

肆、演算法的迷思

從大數據的蒐集、儲存、處理、運算到分析，演算法始終扮演核心與關鍵的角色。我們也可讀到很多政治與政策、社會、行銷與廣告分析都針對網路聲量、語意分析、情緒分析、網絡連結分析。問題是何謂演算法？演算法真的是百分之百、絕對正確的嗎？演算法有哪些局限？

一般而言，演算法是結構化界定的程序與步驟，處理輸入的（input）資料與指令，產生程式設計師、研究者想要的產出（output），作為進一步分析的資料依據。演算法可以自動化處理鉅量與即時資料，辨識人眼無法找出的模式，資料最佳化，找出資料之間的關聯與推薦，預測趨勢與模擬，視覺化呈現資料，極小化人工處理可能造成的誤差與自我校正，大幅減少資料處理的成本（相對於人工

而言）。然而，從人文與社會科學的角度來看，演算法不只是一種軟體與程式工具，更已穿透、中介、型塑我們日常生活的各個層面，構成演算法治理（algorithmic governance）或演算機器（algorithm machines）控制的社會。演算法將社會的規訓與控制自動化，也大幅提升資本積累的效率（Kitchin, 2017: 14-15）。我們應該要追問的，可能不是演算法的定義，而是演算法與日常生活的結合，穿透我們的日常生活，有什麼樣的政治、法律、經濟、社會與文化意涵？演算法看似中壢、客觀的特性，內部有什麼與生俱來的問題？許多學者指出，系統性的偏誤來源，就是個大問題。

演算法的設計鑲嵌在整個運算的技術架構之中，但偏誤（bias）的可能性自系統設計開始就鑲嵌在其中。演算法的偏誤來源，包括個人（態度）與社會（制度）；技術的侷限與考量（電腦工具、去脈絡化的演算法、隨機產生的數字、人的構念的形式化）；突現的（emergent，如使用者與系統設計之間的錯搭 mismatch）。推特的趨勢分析（Twitter Trend）將某些議題優先排在頂端，背後是根據設計時的某些價值。也就是說，推特將某些模式與認知排序自然化（naturalize），但這種結果並不自然，而是演算法設計之初的考量與安排必然的結果（Willson, 2016: 9-10）。也就是說，演算法的偏誤問題不只是除錯（debug）的技術問題，更是個系統問題。

演算法出錯最著名的例子之一，當為谷歌的流感趨勢（Google Flu Trends, GFT）。這個預測的演算法剛剛出台之時，號稱可以即時追蹤實體世界中的流感趨勢，準確地預測到流感趨勢，後來卻屢次出錯。學者指出，這個演算法出現高報與低報誤差，有兩大原因：一是大數據論述的傲慢（hubris），二是演算法的動態（algorithm dynamics）。大數據的傲慢是指誤以為大數據可以取代傳統的研究方法，特別是統計方法基本的量測、構念效度與信度、資料間依賴（dependencies

among data）的問題。後來的研究顯示，疾病管制中心的滯後模型（lagged models）資料，預報的準確度也比此一演算法要高。大數據預測不能取代傳統的統計分析，但兩者可以結合運用，並不斷調整該演算法，得到更準確或更完整的研究發現與知識。在演算法動態方面，工程師的修改測試與網民的搜尋行為，都會影響到搜尋引擎的演算法，後者的搜尋行為又受到前者的影響，導致搜尋到的資料量膨脹。另一種資料量膨脹的原因，來自政治人物與業者刻意的操作（Lazer et al., 2014），如我們常說的網軍或「婉君」操作，塑造網路聲量熱議的印象，或是散播謠言、假新聞，以達到他們在實體世界的目的。

另一個演算法出錯的例子是亞馬遜（Amazon.com）。2009 年 4 月，作家 Mark R. Probst 發現，有些同志羅曼史書籍在亞馬遜的銷售排名書單中消失無蹤，亞馬遜給他回答的是公司的政策是過濾掉帶有「成人」等關鍵字詞的書單。他在自己的部落格中揭露、批評該公司的做法，傳統媒體隨即跟進報導與評論。針對數萬本書籍因此消失在該公司網站的書單之上，亞馬遜發言人的回應說是工程師不小心造成的演算法失誤（Striphas,2015: 395-396）。從常理推斷，亞馬遜應不致於刻意用一個關鍵字詞篩選掉所謂兒童不宜的內容，但這個例子顯示，在大數據與演算法日益穿透我們日常生活的世界，一個思慮不周的指令或設定，可能會造成巨大的困擾。亞馬遜這個例子，只是令特定的群體不快，一時影響商譽。但下錯的指令或設定可能影響到集體安全，造成傷亡，也是值得思考與防範的議題。

不少學者提醒，我們常把網路或社群媒體平台上的資料當作百分之百正確的資料來分析，卻常忽略業者常基於商業機密或智慧財產權等理由或考量，並不公開演算法的內容。推特等社群媒體業者能夠讓人存取的資料數量也常有相當的差異或出入。有的人可以掌握到資料串流的 firehose，有的人只能拿到 garden hose（公共推文的 10%）。還有只能拿到一杯（spritzer，約公共推文的 1%）。API 抽

取隨機樣本的方法也有差異（每小時資料流量的前數千則，或是網絡圖中特別的區塊）。得以分析的推文並未包括受保護的帳號，也不一定是所有公共推文的代表性樣本（boyd and Crawford, 2012: 668-670）。如果資料或數據來源是演算法產生出來的，學者或分析者卻完全倚賴演算法產生的資料，這等於是仰賴黑盒子的資料而作業（Striaphas, 2015: 406-407; Beer, 2017: 2-3; Kitchin, 2017: 15, 20），因此才有學者以「黑盒子社會」（Black Box Society）稱呼當代日益被演算法控制的社會（Pasquale, 2016）。

用演算法分析資料或數據，即使沒有偏誤，也要面對資料分析本身的局限。例如，學者以電腦輔助的內容分析法，找出歐巴馬（Barack Obama）與羅姆尼（Mitt Romney）競選美國總統期間期間的推特文章，發現潛在的狄利克雷分布（Latent Dirichlet Allocation, LDA）這種非監督的機器學習演算法（unsupervised machine learning algorithms），比字典為基礎的文本分析（dictionary-based textual analysis）更能辨識主題（topic），前者的效能與效度均高於後者。但 LDA 也不是沒有缺點，它與字典為基礎的分析法都無法處理諷刺、反諷等語意。進階的自然語言處理（natural language processing）或許有助於解決這個問題，監督的機器學習，也可能是一種解決方法。無論是哪一種演算法，都要面對社群媒體文字內容特殊的挑戰：拼錯字、文法錯誤、表情符號、截斷字詞等，這些都需要研究者預先處理與清理，才能做下一步的演算分析（Guo *et al.*, 2016）。

伍、結論

本文綜合整理國內以大數據為主題或關鍵字的期刊論文與專書論文，勾勒其特質與問題。接著探討大數據、小數據、社會大數據的定義與概念內涵。大數據可說是一種流動的概念，「大」的概念或門檻，往往隨著科技發展而調整與改變。

到目前為止，大數據還不能取代小數據，大數據分析發掘與建構的知識，還不一定能取代小數據研究累積的知識，因為兩者的情境脈絡與研究方法大不相同，儘管大數據與小數據的研究方法可以相互結合運用。在人文與社會科學研究之中，許多號稱大數據的研究，其實稱為社會大數據，或是資料採礦（data mining），似乎比較妥當。社會大數據實際分析的數據量，好像也沒有大到天文數字的地步，這是因為他們能夠從網路社群媒體平台業者取得的資料數量往往是有限的。所謂有限，是指業者願意釋出多少資料，還有業者蒐集到的資料是否真的能涵蓋所有社群媒體的文本資料。如果分析的資料只是業者手中資料的一小部分，就趕流行稱為大數據研究，恐怕有點過頭。

相對於國外學術界的研究論著，台灣對大數據研究的知識論、方法論與理論問題，至今仍處於低度發展狀態。既有的大數據分析隱約接受實證論、邏輯實證論與經驗主義的知識論、方法論預設，缺乏批判的思考與檢視，忽略不同尺度情境的資訊與知識建構。深刻的人文與社會-文化理論觀點更是付諸闕如，結果就是我們看到為數眾多的案例與應用研究，但無法給我們整體的理論觀點與社會-文化圖像，以及大數據引發的倫理問題。

展望未來的研究議程，人文與社會科學，乃至於公共政策的研究，除了政策、法律的案例與應用研究之外，似乎可以思考知識論、方法論與理論的研究軸線。就知識論而言，大數據建構的知識有哪些特質與問題、局限？哈伯瑪斯分殊的三種知識興趣、研究途徑與技術，如何看待大數據與演算法對知識的認知與建構知識的模式？方法論方面，大數據與演算法的研究操作程序如何處理、減少系統性偏誤的問題？在政策分析與人文、社會科學的具體與經驗性研究上，大數據與小數據如何或是否可能結合，以兼顧不同尺度與社會情境的觀點與資料？理論方面，哪些公共行政與社會-文化的理論可以用於大數據的研究？例如，傅科式的理論

觀點如何看待、解析開放資料（open data）、開放政府（open document）的政策論述？

這裡提出的未來研究建議只是舉例，不能窮盡未來可以發展的研究議程，但總算是一個起步，希望能夠拋磚引玉，促成國內大數據研究在公共政策與人文學、社會科學研究的深化與廣化。

參考文獻

中文部分

王信賢（2018）。〈科技威權主義：習近平「新時代」中國大陸國家社會關係〉。《展望與探索》，16(5): 111-127。

丘昌泰（2017）。〈以大數據挖掘主計資料金礦〉。《主計月刊》，721: 246-52。

丘昌泰、劉宜君（2017）。〈大數據產業的資料隱私問題與對策〉。《產業與管理論壇》，19(1): 28+30-51。

甘炎民、郭士豪、黃冠豪、李承龍（2015）。〈大數據資料系統分析運用在偵查實務之研究〉。《警察通識叢刊》，5: 140-159。

江亦瑄、林翠娟（2015）。〈採用大數據探討媒體使用之學術期刊文獻分析〉，頁355-368，收入彭芸主編，《大數據、新媒體、使用者論文集》。新北市：風雲論壇。

江彥生、陳昇瑋（2016）。〈簡介「計算社會學」：一個結合電腦與數位科技的新興社會學研究〉。《臺灣社會學》，32: 171-201。

余峰偉、王嵩音（2018）。〈臺灣數位音樂串流服務 Facebook 粉絲專頁溝通策略分析：鉅量資料分析取徑〉。《電子商務研究》，16(1): 59-102。

杜聖聰、楊喨智、巫家宇（2016）。〈九二共識的臺灣網路輿論發展：以 OpView 為調查工具〉。《人文社經論叢》，3: 65-89。

范姜真媺（2017）。〈大數據時代下個人資料範圍之再檢討：以日本為借鏡〉。《東吳法律學報》，29(2): 1-38。

韋愛梅（2015）。〈大數據時代的家庭暴力防治〉。《警專論壇》，17: 88-97。

張嘉玲（2016）。〈案以群分：大數據解析回饋型群眾募資成功關鍵要素〉。《臺灣經濟研究月刊》，39(2): 27-37。

許炳華（2016）。〈大數據時代下隱私權之保護：可能之影響暨對策〉。《興大法學》，20: 129-191。

許華孚、吳吉裕（2015）。〈大數據發展趨勢以及在犯罪防治領域之應用〉。《刑事政策與犯罪研究論文集》，18: 341-375。

陳敦源、蕭乃沂、廖洲棚（2015）。〈邁向循證政府決策的關鍵變革：公部門巨量資料分析的理論與實務〉。《國土及公共治理》，3(3): 33-44。

彭金隆、陳俞沛、孫群（2017）。〈巨量資料應用在台灣個資法架構下的法律風險〉。《臺大管理論叢》，27(2S): 93-118。

黃章令（2018）。〈重塑大數據時代下的隱私權法理：以隱私權概念為主要內容〉。《月旦民商法雜誌》，62: 131-162。

楊嘵智（2015）。〈宗教與多元性別議題之網路輿情觀察：觀測時程：2013 年 9 月至 2013 年 12 月〉。《臺灣宗教研究》，14(2): 53-63。

楊嘵智（2018）。〈宗教與同運議題之網路輿情觀察〉。《玄奘佛學研究》，29 : 59-81

葉志良（2016）。〈大數據應用下個人資料定義的檢討：以我國法院判決為例〉。《資訊社會研究》，31: 1-36。

葉志良（2017）。〈大數據應用下個人資料的法律保護〉。《人文與社會科學簡訊》，19(1): 31-36。

葉奕新（2017）。〈臺北捷運系統之人潮移動分析〉。《中國統計學報》，55(2): 69-95。

劉玉山（2016）。〈智慧決策與大數據資訊應用：觀念性架構與研究議題建構〉。《東亞論壇季刊》，494: 55-68。

劉定基（2017）。〈大數據與物聯網時代的個人資料自主權〉。《憲政時代》，

42(3): 265-308。

劉宜君（2016a）。〈大數據時代對於醫療照護影響與醫療隱私保護之研究〉。《前瞻科技與管理》，6(1): 1-25。

劉宜君（2016b）。〈大數據與政策創意之研究〉，頁 233-259，收入彭芸主編，《數位匯流時代：創新、創意、創世紀》。新北市：風雲論壇。

劉宜君（2017）。〈大數據時代的個人資料隱私與去識別化之探討〉。《前瞻科技與管理》，7(2): 1-34。

劉嘉薇（2017）。〈網路統獨的聲量研究：大數據的分析〉。《政治科學論叢》，71: 113-165。

劉靜怡（2017）。〈巨量資料相關應用的規範省思〉。《人文與社會科學簡訊》，19(1): 20-26。

鄭宇君（2014）。〈向運算轉：新媒體研究與資訊技術結合的契機與挑戰〉。《傳播研究與實踐》，4(1): 67-83。

鄭宇君（2015）。〈鉅量資料時代下的使用者研究〉，頁 301-312，收入彭芸主編，《大數據、新媒體、使用者論文集》。新北市：風雲論壇。

鄭宇君、陳百齡（2014）。〈探索 2012 臺灣總統大選之社交媒體浮現社群：鉅量資料分析取徑〉。《新聞學研究》，120: 121-165。

蕭乃沂、朱斌妤（2018）。〈資料驅動創新的跨域公共治理〉。《國土及公共治理》，6(4): 74-85。

賴祥蔚（2015）。〈大數據趨勢下的收視行為研究〉，頁 339-354，收入彭芸主編，《大數據、新媒體、使用者論文集》。新北市：風雲論壇。

鍾永淙（2017）。〈主計資料結合商業智慧，創新應用模式〉。《主計季刊》，58(1): 41-49。

羅鈺珊（2018）。〈數據經濟下共融成長的挑戰：大數據的兩面刃〉。《經濟前瞻》，178: 87-93。

鐘嘉德、柴惠珍、高崎鈞、曹元良（2015）。〈我國大數據政策推動現況〉。《國土及公共治理》，3(4): 77-84。

英文部分

Beer, D. (2017). "The Social Power of Algorithms." *Information, Communication & Society* 20(1): 1-13.

Bilić, P. (2016). "Search Algorithms, Hidden Labour and Information Control." *Big Data & Society* 3(1): 1-9.

boyd, D. and Crawford, K. (2012). "Critical Questions for Big Data." *Information, Communication & Society* 15(5): 662-679.

Burns, R. and Thatcher, J. (2015). "Guest editorials: What's So Big about Big Data? Finding the Spaces and Perils of Big Data." *GeoJournal* 80(4): 445-448.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., and Ishwar, P. (2016). "Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling." *Journalism & Mass Communication Quarterly* 93(2): 332-359.

Habermas, J. (1972). *Knowledge and Human Interests*. Translated by Jeremy J. Shapiro. London: Heineman.

Halavais, A. (2015). "Bigger Sociological Imaginations: Framing Big Social Data Theory and Methods." *Information, Communication & Society* 18(5): 583-594.

Kitchin, R. (2017). "Thinking Critically about and Researching Algorithms."

Information Communication & Society 20(1): 14-29.

Kitchin, R. and Lauriault, T. P. (2015). "Small Data in the Era of Big Data." *GeoJournal* 80(4): 463-475.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343: 1203-1205.

Matzner, T. (2017). "Opening Black Boxes Is Not Enough: Data-Based Surveillance in *Discipline and Punish* and Today." *Foucault Studies* 23: 27-45.

Pasquale, F. (2016). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Rosenberg, A. (2000). *Philosophy of Science: A Contemporary Introduction*. 2nd Edition. New York and London: Routledge.

Striphas, T. (2015). "Algorithmic Culture." *European Journal of Cultural Studies* 18(4-5): 395-412.

Symons, J. and Alvarado, R. (2016). "Can We Trust Big Data? Applying Philosophy of Science to Software." *Big Data & Society* : 3(2), 1-17.

Wells, C. and Thorson, K. (2017). "Combining Big Data and Survey Techniques to Model Effects of Political Content Flows in Facebook." *Social Science Computer Review* 35(1): 33-52.

Willson, M. (2016). "Algorithms (and the) Everyday." *Information, Communication & Society* 20(1): 1-14.