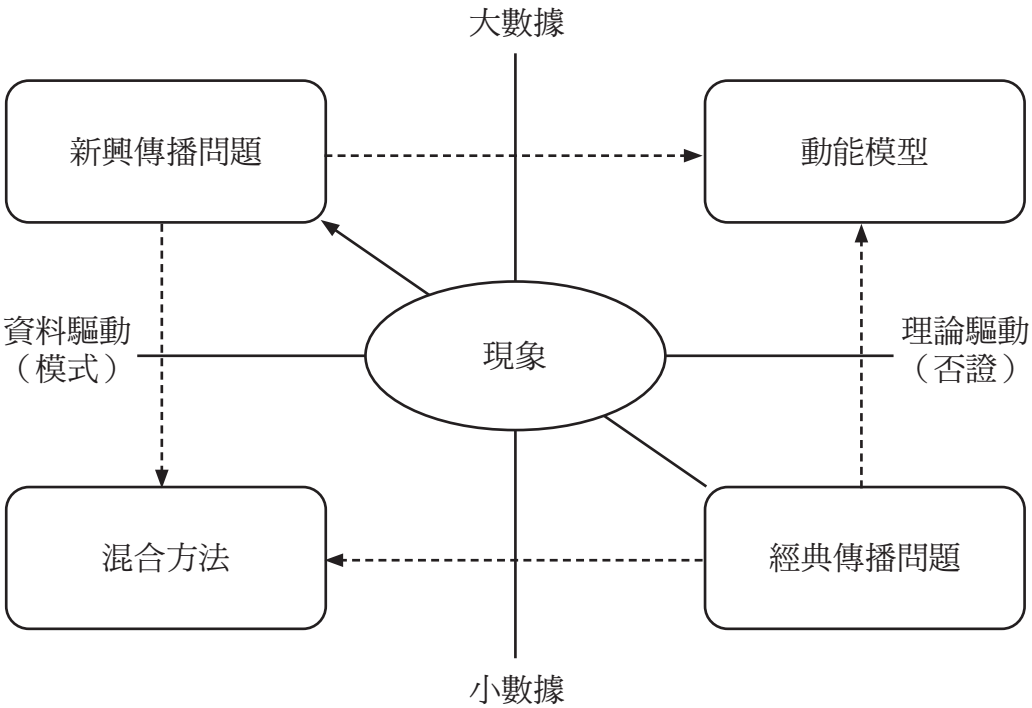
大數據與傳播研究：運算傳播研究的可能發展方向

陶振超*

人類活動之數位資料的出現，因其記錄人數眾多（可達數萬、甚至以上）、觀測面向廣（如到過哪裡的定位資訊、瀏覽哪些網頁、個人網絡與人際互動等各式各樣變項）、保留過程（也就是提供時間軸，在連續、多個時間點重複測量，而非僅在單一或少數時間點測量一次）而被稱為大數據，遠非其他研究方法所能企及，被認為具備回答傳播研究重要問題的潛力而吸引許多研究者投入，促成新興研究領域「運算傳播研究」的出現，並改變了傳播研究的版圖。我們將從以下兩個面向，檢視這樣的改變與未來可能的發展方向（參見圖一）。



圖一：傳播研究版圖的可能發展

* 國立陽明交通大學傳播與科技學系教授

第一個面向，考慮研究發現是「理論驅動」或「資料驅動」。理論驅動指研究著重變項間的**因果**關係，變項間的模型來自理論，資料作為否證（falsification）模型之用，研究目的在**解釋**。因為重點在變項間的因果關係，分析工具偏向使用統計方法，適合觀測面向（自變項）少的簡單模型。模型選擇的標準，在模型既能符合理論、又與資料適配。資料驅動指研究著重變項間的**相關**關係，資料作為偵測模式（pattern）之用，資料中的一部分用於導出變項間的模型，剩餘的資料則用於驗證模型，研究目的在**探索或預測**。因為重點在變項間的相關關係，分析工具偏向使用演算法，較適合觀測面向（自變項）多的複雜模型。模型選擇的標準，在對新資料進行預測有最佳的準確性。

第二個面向，考慮採用的研究資料是「小數據」或「大數據」。兩者除了在記錄人數、觀測面向、過程長短上有巨大差異外，更重要的是研究設計與資料蒐集是一起進行還是分開進行。小數據研究的**研究設計與資料蒐集結合在一起進行**。在蒐集資料前，須先建立研究假設與研究問題，並決定蒐集哪些變項及這些變項的操作化程序；資料蒐集時，僅測量事先決定的變項；資料蒐集後，因變項的操作化程序已經確定，所得多為結構化資料，易於轉為量化指標，以目前普遍被接受的統計方法進行假設驗證。大數據研究的**研究設計與資料蒐集則是分開進行**。資料蒐集通常由數位平臺進行，記錄人類在這些數位平臺上的各式各樣活動，但在資料蒐集前，並沒有建立任何研究假設與研究問題，也沒有決定蒐集哪些變項及這些變項的操作化程序，使得這些資料大多未結構化；研究設計則通常由研究者進行，面對已經蒐集的未結構化資料，提出研究假設與研究問題、決定變項及其操作化程序，使用一般社會科學研究者陌生、偏向資訊科學的程式語言整理資料與轉為量化指標，最後以演算法進行預測。

大數據的出現，將傳播研究從圖一右下角以理論驅動、小數據為主的經典傳播問題研究，擴展到左上角資料驅動、大數據為主的新興傳播問題，尤其數位平臺的興起，累積了大量的文字資料可供分析（Shah, Cappella, & Neuman, 2015）。這些研究針對人們在真實世界的實際行為進行分析，不是從自我報告而來的測量（van Atteveldt & Peng, 2018），一方面回應了社會當前的需要（如假訊息、假新聞、深度偽造，或社交／群媒體在政治上的影響、網路口碑在行銷上的力量），另一方面符合長期鼓勵與推動的跨領域合作（如傳播與資訊科學），迅速成為熱門的研究取徑。文字資料一向是傳播領域研究者關注的焦點之一，也有內容分析等方法將文字資料轉換為量化指標。伴隨大數據出現的新興分析工具（如情感分析、機器學習等），對傳播研究者來說是如虎添翼，可以自動內容分析對數位平臺上的大量文本進行意義層面的辨識與分類。尤其資料驅動、大

數據取徑強調預測性模型，與理論驅動、小數據強調解釋型模型不同（Yarkoni & Westfall, 2017; Shmueli, 2010; 不同看法，見 Fricke, 2015）。人類活動之數位資料因觀測面向廣，其中可能存在許多現有理論沒有包含的概念，使得以解釋性模型為主的現有理論不易用來檢驗這些數位資料。然而，預測性模型在此狀況卻可發揮長處，將數量龐大及新興概念納入一起檢驗，可能揭露現有理論未考慮、潛在的因果關係供未來進一步研究。這樣的作法提升研究發現的實用性，使得學術研究符合社會需求，彰顯學術研究的價值。比較可惜的是，這方面的研究目前以敘述性模型較多，預測性模型有待更多研究者投入。

然而，近來研究者開始反思大數據引發的研究，是否真的回答了傳播研究的重要問題、促成傳播研究進步。越來越多不同社會科學領域的研究者了解到大數據、資料驅動型態的研究，不能取代小數據、理論驅動型態的研究，兩者應該建立互補的關係。首先，大數據的資料大多在研究開始前已被蒐集，此時規劃的變項操作化程序可能涉及新興方法（如機器學習），使得經由這些操作程序所得的僅是「近似」欲研究的理論構念（Adjerid & Kelley, 2018）。舉例來說，以字詞的屬性及頻率對文本分類以代表特定理論構念（如文本的立場是支持或反對某個議題），後者是否反應前者是測量問題，測量誤差依然存在。雖然研究者對相較於小數據，大數據的測量誤差變大還是變小有不同看法，但針對大數據之操作化程序的信度與效度其實比小數據還需要重視與檢視。其次，小數據存在的統計問題，在大數據中仍存在，甚至被放大。舉例來說，大數據的記錄人數雖然眾多，樣本的代表性並非較佳；有時因為來自特定網站，其樣本代表性更偏差（Hargittai, 2015; van Atteveldt & Peng, 2018）。記錄人數眾多（即樣本數大）容易達到高統計顯著性（Gandomi & Haider, 2015）及多重假設假定（Clark & Golder, 2015），提升偽相關的可能性，p 值甚至被建議減少使用。記錄人數眾多再加上觀測面向廣，資料與模型容易形成過度適配，使得模型不適用於樣本以外的其他資料，導致概推性降低（Adjerid & Kelley, 2018; Domingos, 2012; Yarkoni & Westfall, 2017）。省略變項（未納入模型中考慮但影響依變項的因子）在大數據中比在小數據中還嚴重，導致估計值偏誤（Clark & Golder, 2015）。畢竟資料的量與資料的質是兩件不同的事情。接著，大數據的近用仍存在問題（boyd & Crawford, 2012）。大數據的使用可能集中在少數研究者或企業獨占，大多數研究者仍接觸不到，如 Bakshy, Messing, & Adamic（2015）分析一千萬美國臉書使用者的同溫層現象。如何促進開放資料與共享資料是未來的重要議題。更進一步，若要在數位平臺上利用眾多的使用者進行隨機分配的實地實驗，檢驗幾個變項在較長過程中的因果關係，如 Bond et al.（2012）比較臉書上社交訊

息的政治動員效果；Muchnik, Aral, & Taylor（2013）在新聞網站進行的社會影響研究；Kramer, Guillory, & Hancock（2014）極具爭議在臉書上進行的情緒傳染實驗，僅有非常少數的研究者有機會進行，在臺灣幾乎沒有。最後，大數據相關的研究倫理仍在發展中。從參與者知情同意（如知情同意的研究參與者，常常無意間揭露未提供知情同意的朋友與其貼文，這樣蒐集到的資料是否符合知情同意的要求）、隱私保護與去識別化（如數位足跡可能揭露原本應該匿名保護的研究參與者身分），以及研究結果是否可能被其他組織不當使用等，具體守則尚未訂定（Adjerid & Kelley, 2018; boyd & Crawford, 2012; van Atteveldt & Peng, 2018）。這些皆為採用大數據的研究者不能忽視的重要議題。

以上兩種研究取徑彼此間不應該是對立、甚至緊張的關係，應該是互補的，可以搭配在一起進行。研究同時兼顧理論層面的解釋性與資料層面的實用性更能提升研究的貢獻。有兩個可能的發展方向。第一個發展方向，是根據大數據進行之資料驅動研究所得到的研究發現，提出新的假設，並以小數據檢驗**資料型式**的解釋力，檢視過去未考慮的變項間因果關係，並進一步發展成新理論（見圖一左下角）。這樣的發展方向稱為「**混合方法**」取徑，重點在發展新的因果關係。此發展方向的挑戰，在大數據進行之資料驅動研究所提出的新假設，與現有理論是進行融合還是挑戰。無論小數據進行之理論驅動研究的結果與大數據進行之資料驅動研究所提出的新假設一致或不一致，研究者在理論與資料層面的討論與後續研究規劃，是最重要的步驟。以質化研究對個案進行深入討論，對此發展方向也有很大的助益。第二個發展方向，是根據小數據進行之理論驅動研究所得到的研究發現，以真實世界、大量、包含時間軸之動態資料檢驗**理論模型**的預測力，檢視已獲支持之變項間因果關係放在真實世界是否能運作，確認理論的實用價值（見圖一右上角）。這樣的發展方向稱為「**動態模型**」取徑，重點在檢視既存因果關係的實用性，也被認為可能解決近來社會科學領域發生的研究重複驗證危機（Sakaluk, 2016）。此發展方向面對的挑戰，在小數據進行之理論驅動研究中的各個變項，也就是傳播研究中的理論構念，研究者要提出適當的操作化程序，並檢視測量的信效度問題，將數位平臺所記錄的非結構化資料轉為變項。另外，針對較長過程、涉及時間軸的資料，一方面在蒐集上有一定的困難度；另一方面在統計分析傳播領域的研究者較陌生，皆延緩了傳播領域採納動態模型的時間。

綜上所述，以大數據進行之資料驅動研究的出現，不但提供傳播領域的研究者新興工具將大量文本以自動化方式轉為量化指標，並進一步引用預測性模型，使傳播研究結果具更高的實用性，能具體回應社會需求。但要能善用此一

新興發展方向，傳播領域的研究者不僅在運算知識與能力上需要提供，並將其融入研究所與大學部的教育中；更重要的是，大數據研究在測量指標的信效度上可能比小數據研究更重要，輕忽這項議題會降低預測性模型的準確性與概推性，如何評量有待更多學者投入與釐清；接著，在一篇學術論文中結合大數據與小數據研究，無論是從大數據研究中提出新的假設、在小數據研究中驗證與提出理論解釋，或從小數據研究中建立因果關係、在大數據研究中以動態模型進行預測，皆能促使傳播領域在知識上的進展。最後，在大數據資料的共享與對研究倫理與隱私保護的重視是促使運算傳播研究學術社群成長的關鍵，需要傳播領域的研究者立刻採取行動。

參考文獻

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist,* 73(7), 899-917. doi: 10.1037/amp0000190.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science,* 348(6239), 1130-1132. doi: 10.1126/science.aaa1160.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature,* 489(7415), 295-298. doi: 10.1038/nature11421.

boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication & Society,* 15(5), 662-679. doi: 10.1080/1369118x.2012.678878.

Clark, W. R., & Golder, M. (2015). Big data, causal inference, and formal theory: Contradictory trends in political science? *PS: Political Science & Politics,* 48(1), 65-70. doi: 10.1017/s1049096514001759.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM,* 55(10), 78-87. doi: 10.1145/2347736.2347755.

Fricke, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology,* 66(4), 651-661. doi: 10.1002/asi.23212.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management,* 35(2), 137-144. doi: 10.1016/j.ijinfomgt.2014.10.007.

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *ANNALS of the American Academy of Political and Social Science,* 659(1), 63-76. doi: 10.1177/0002716215570866.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America,* 111(24), 8788-8790. doi: 10.1073/pnas.1320040111.

Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science,* 341(6146), 647-651. doi: 10.1126/science.1240466.

Sakaluk, J. K. (2016). Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology,* 66, 47-54. doi: 10.1016/j.jesp.2015.09.013.

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *ANNALS of the American Academy of Political and Social Science,* 659(1), 6-13. doi: 10.1177/0002716215572084.

Shmueli, G. (2010). To explain or to predict? *Statistical Science,* 25(3), 289-310. doi: 10.1214/10-sts330.

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures,* 12(2-3), 81-92. doi: 10.1080/19312458.2018.1458084.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100-1122. doi: 10.1177/1745691617693393.