巨量資料在經濟學研究的應用及挑戰

簡錦漢、劉玉晢*

一、巨量資料的定義

人工智慧的發展，奠基在巨量資料的運算和分析上。巨量資料不但為學術研究帶來突破，其研究結果也有深刻的政策意涵，可增進社會福祉。例如運用高達百萬筆的細部交通資料，探討計程車顏色是否對肇事率有影響（Ho et al, 2017）；或是結合網路評論和市政資料，探討餐廳衛生稽查的人力應如何配置方為最適（Glaeser et al., 2016）。此外還有許多醫療、企業管理、財務、公共行政上的應用。

長久以來，經濟學實證研究常使用巨量資料。「巨量資料」必須符合「數量大（Volume）、產生速度快（Velocity）、資料具多樣性（Variety）」這三個定義。資料量大，可以讓抽樣誤差降低；資料產生速度快，讓研究者獲取資料更為便捷；資料具多樣性，則可讓研究者獲取更多不同層面的變數，用來驗證何種假設為真，更利於假設檢定和因果推論。由於巨量資料包含的資訊量很龐大，經濟學者又稱這種變數個數高於樣本數的資料為高維度資料（high-dimensional data）。

巨量資料的種類很多，資料來源可能來自於公私部門各層面，例如稅務資料、出生登記資料、醫療保險資料、教育成績、消費紀錄、信用卡交易資料、互聯網資料等等。行銷和管理學常用的商品掃描資料（scanner data）、政府執行公務所累積的公務資料（administrative data）也屬於巨量資料。

巨量資料為社會科學提供更多資料來源，但也為固有的計量方法帶來挑戰。不論是個體或總體經濟學，皆已廣泛使用巨量資料作為研究資料來源。資料可及性也在不知不覺間主導了研究議程的進展與推論，決定了重大議題的走向。本文介紹巨量資料的種類與應用趨勢，並且簡介常用的巨量資料分析方

* 簡錦漢，中央研究院經濟研究所所長；劉玉晢，科技部人文社會科學研究中心博士後研究員。

法，最後介紹資料推論時應注意的事項。

二、巨量資料的應用

經濟學上運用甚廣的巨量資料，首推公務資料。到了 2010 年，在 AER、QJE、JPE 等頂尖經濟學期刊的論文，資料來源為公務資料者已超過六成，凌駕傳統的抽樣調查資料（Chetty, 2012）。就社會層面而言，公務資料對公共政策評估更至關重要。公務資料包含社會各群體或階層，避免了抽樣誤差，而且以個體為基礎，紀錄了個體的反應和行為；因此，若想探討社會中各群體間的差異或不均等效果，這些公務資料貢獻極大（Einav & Levin, 2014）。如果公務資料能達到跨部門的資料連結，更有助於政策因果關係的確認、以及處置效果的估計。

臺灣最有名也最常被使用的公務資料為健保資料庫等，由於資料完整、涵蓋人口廣，已累積相當多研究。例如 Chou, Grossman & Liu（2014）利用臺灣出生登記資料，以臺灣全民健保的施行時點作為自然實驗（natural experiment），研究全民健保對於社會不同部門的嬰兒死亡率的影響。Chen & Cheng（2015）則利用全民健保資料庫，研究論質計酬（pay for performance）支付誘因對於肥胖病人照護連續性（Continuity of Care, COC）的影響。這些研究強調嚴謹的因果推論，而非單純的顯示資料相關性，對於理解民眾醫療選擇、提出政策建議，皆甚有幫助。連賢明（2008, 2011）有系統地介紹健保資料庫的學術使用方式，並以家庭收支調查串聯全民健保資料庫，以補健保資料庫個人社經資料之不足，能夠更準確衡量疾病或制度對個人的影響，顯示出跨資料庫連結的重要性。

臺灣其他使用公務資料的重大研究，例如朱敬一等（2015）使用財政部財稅資料，探討臺灣的所得分配、跨代流動及不均等狀況，研究已被納入 Atkinson 與 Piketty 所創 World Top Income Database 之中。其資料包括國民的勞動所得和報稅檔案，也包括財產總歸戶之土地、房屋等資料，除了北歐國家，這是國際上唯一紀錄了家戶財產的公務資料，對於政策制定和學術研究的意義重大。近年來，全球和臺灣都把貧富差距和所得不均列為重要議題；然而，完整的政策推論需要以嚴謹的學術研究為基礎，若要進行實證研究，詳實的公務資料是不可或缺的。

巨量資料可以為總體經濟學提供更具個體基礎的經濟衡量指標，與現行調查或官方指標作參照。例如 Vosen & Schmidt（2011）以 Google Trends 衡量私人

消費，計算各種消費類別被搜尋的次數，並且把這個指標跟密西根大學消費者信心指數和美國消費者信心指數兩者相較，指出新指標在樣本內及樣本外的預測能力皆優於傳統的消費者信心指數。Choi & Varian（2012）則使用 Google Trends 來衡量汽車銷售、失業救濟、消費者信心指數等常見的經濟指標，這些資料由個體的網路行為累積而得，有別於由公務單位或權威機構進行的經濟調查，為總體經濟提供另一種衡量方法。麻省理工的 “The Billion Prices Project” 則彙整數百家線上零售商每日報價資料，商品種類超過五百萬種，橫跨七十餘國；Cavalla & Rigobon（2016）使用該資料所計算出來的阿根廷通貨膨脹率，甚至比阿根廷政府的官方數據更貼近現實。以上也是使用互聯網資料進行總體經濟研究的實例。

互聯網資料包含交易行為，也有訊息傳遞行為或資訊蒐集行為，這些都是個體經濟實證研究的主題。因此，互聯網也是個體經濟學研究的資料來源，例如 Chen, Chou & Huang（2013）利用臺灣奇摩拍賣網站所蒐集到的資料進行拍賣行為的研究，研究「立即買」（Buy It Now）選項存在時，買賣雙方的最佳策略為何。Nardo, Petracco-Giudici & Naltsidis（2016）則對股價網站的網路評論進行文字探勘和情緒分析，探討投資人情緒對股價的影響。Heiberger（2015）則使用網路搜尋紀錄作為社會集體注意力的指標，探討特定公司被搜尋次數對其股價的影響。

電子商務公司也是巨量資料的來源。有別於總體經濟的加總後資料，電子商務資料的優點在於微觀尺度的數據，以個體為單位，可讓經濟學家挖掘出不同個體間的差異、進行更準確的因果推論。例如 Einav, Knoepfle & Sundaresan（2014）比較不同州但跟賣家距離相等的買家對於同一商品的點擊次數，比較不同稅率對於買家意願的影響。更有甚者，若企業願意配合，還可以進行經濟學的現場實驗（field experiment），測試不同條件下顧客的反應。網路企業由於進行實驗的成本較低，因此配合實驗的意願較傳統產業更高。例如 Ostrovsky & Schwarz（2011）與 Yahoo! 合作，探討不同的拍賣底價對於廣告拍賣的影響；或是 Blake 等（2015）與 eBay 合作，探討若關閉特定品項的 Google 搜尋，該商品交易將產生何種變化。

金融市場的逐筆交易（tick-by-tick）資料也是一種巨量資料。逐筆交易資料是在每個時間切分下最優買單和賣單的變動，對行為財務學和計量金融學研究貢獻良多。Barber et al.（2009）使用 1995-1999 年臺灣股票市場投資人下單與成交的逐筆交易資料，包括價量、時間、投資人身分資訊，可找出各種投資人的

交易行為及損益結果，探討交易造成的各群體福利變化。即使在今天，如此完整的資料依舊不易取得。這篇研究顯示出，完整的資訊對於學術研究十分有助益，也對社會整體福利和政策制定皆有貢獻。其他巨量資料來源，還包括衛星遙測影像資料、手機使用資料等等，這兩者都對都市經濟學和發展經濟學的研究甚有助益（Blumenstock et al., 2015; Donaldson et al., 2016）。

三、巨量資料的分析方法

巨量資料所含的資訊量和變數項眾多，且各變數間可能是非線性且非單向的關係，因此，在處理巨量資料時，常面臨資料異質性（heterogeneity）、過度配適（over-fitting）和模型選擇（model selection）的問題。所謂「異質性」是因為巨量資料中的變異性太大，誤差項不符合變異數相同的假說；「過度配適」和「模型選擇」的問題起因則是巨量資料所含的參數太多，即使以模型的預測能力為挑選標準，也難以決定哪一個模型是最適的；在這種狀況下，所獲得的模型一般化（generalized）的能力反而降低了。各種處理巨量資料的方法，不論是機器學習方法還是經濟學的計量方法，都在盡力克服這些問題。

關於巨量資料的分析方法，不同領域的著眼點各有不同。在資管領域，巨量資料的分析方法傳統上以資料採礦（data mining）和機器學習（machine learning）為主，目的在於預測，比較著重「探索性」（Exploratory Analysis）的數據分析；但經濟學及其他社會科學常用的計量方法，其目的在於進行因果推論（casual inference）、或是估計處置效果（treatment effects），比較著重「驗證性」（Confirmatory Analysis）的數據分析。「探索性分析」沒有預設立場，重點在整理出資料結構，找出某些變項的相關性，呈現出資料大致趨勢；而經濟學著重的「驗證性分析」是在控制部分變因下，確認關鍵變數之間的因果關係，甚至驗證既有的理論或假設。這兩種方法並不衝突，但既然目的迥異，這兩種方法所找出來的變數關係有可能截然不同。如果沒有探究個體在各種狀況下面臨的真正誘因與反應，僅從歷史數據中推斷事件的關係，那麼，即使是監督式機器學習得出的預測模型，仍有可能僅測得虛假的表面關係，這是資料分析時應注意的。

以上兩種方法並非完全對立，近年來，經濟學家致力於結合機器學習和因果推論兩者，以確認因果關係，並對處置效果（treatment effects）進行估計，例如 LASSO 法結合傳統迴歸模型和 GMM 確認因果關係（Cheng & Liao, 2015），是用迴歸樹（regression tree）來處理異質化問題，改進因果關係模型（Athey &

Imbens, 2016），增進對於巨量資料的分析能力。

四、巨量資料應用的趨勢與限制

使用巨量資料進行學術研究、甚至輔助公共決策，是未來不變的方向。巨量資料為總體經濟提供更具個體基礎的衡量方式，也幫助個體經濟學更精確地掌握個體的動態行為。目前應用的趨勢，是強調對於個體行為的預測能力；這並不意味消除資料匿名性或侵犯個人隱私，而是使用資料推估或對照的技巧，推算出各代表性群體在不同社經條件下的行為模式。巨量資料在公共行政上更有相當廣泛的運用，例如交通規劃、消防或食物安全檢查等（Ho et al., 2017; Athey et al., 2017），目的在於極大化施政效益並極小化行政成本。

儘管巨量資料應用廣泛，但在分析資料時，仍應注意推論過程的合理性。Athey（2017）提到，「預測」不等於「因果推論」，從「資料分析」跨越到「決策」時，必須審慎為之。舉例而言，機器學習可以幫助企業找出最易流失的顧客；但這不表示企業應該將大部分資源投注在這些顧客身上，因為這批顧客可能對企業的任何措施都不敏感，資源投注在他們身上，未必能得到最有效率的回報。資料分析的結果是找出目標顧客，然而企業真正關心的因果關係是「投資在客戶管理上是否有相應的報酬」，這是可能存在的落差。

此外，巨量資料例如互聯網資料和企業資料等，其代表性也常受質疑。因為這些資料雖然量大，但不一定涵蓋所有消費者或所有商品，也可能不是隨機抽樣。例如線上零售商的商品很有可能集中在某些種類，分布跟母體有所差異，這些差異有可能導致推論偏誤。又例如對社群網站和網路論壇進行文字探勘，活躍使用者的意見可能就被放大。再者，社群網路、網路論壇、搜尋引擎有其自定的演算法，例如搜尋引擎必須進行最適化（search-engine optimization），研究者蒐集到的網路資料已是演算法淘選後的結果，可能有演算法偏誤（algorithm bias）。

此外，跨部會的公務資料聯繫不易，個別觀察值的社經背景資料有可能不足，例如健保資料庫雖然包括就醫和健康紀錄，但缺乏個人的收入、教育程度、家中排行、子女數目、婚姻狀況等社會經濟變數；又例如教育資料庫缺乏個別學生的家庭背景、家庭收入、排行、父母教育程度等資料。這些缺失都使得個體的個別狀況難以被指認、重要因素難以被控制，對因果推論造成困難。若能在保護隱私權的前提下，進行跨部會的公務資料庫串聯，例如以村里加總

資料代替個人資料，避免個人隱私外洩，就能對學術研究和政策建議甚有助益。這是未來學者、公務機關、民眾需要取得共識之處。

資料開放有益於政策討論和學術研究，有助於揭發社會分配的不均，如能避免侵犯隱私權，資料開放應是全民努力的目標。另一個隱憂則是：資料的開放程度，也可能反過來決定了何者會被研究、何者會被討論。如果某議題資料欠缺、無法進行實證研究，也就無法形成嚴謹的政策討論。然而，資料開放的腳步並不是均等的，不同資料會遇到不同的障礙，可能是技術障礙、也可能是行政上的障礙；所以，資料可及性也在不知不覺間決定了目前學術研究或政策討論的議程，這是學者和整個社會在進行資料分析時必須意識到的。

參考文獻

Athey, S. (2017). Beyond prediction: Using big data for policy problems. Science, 355(6324), 483-485.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353-7360.

Barber, B. M., Lee, Y.-T., Liu, Y.-J., & Odean, T. (2009). Just How Much Do Individual Investors Lose by Trading? The Review of Financial Studies, 22(2), 609-632.

Blake, T., Nosko, C., & Tadelis, S. (2015). Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment. Econometrica, 83(1), 155-174.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076.

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. Journal of Economic Perspectives, 30(2), 151-178.

Chetty, R. (2012). Time Trends in the Use of Administrative Data for Empirical Research, slides presented at NBER Summer Institute, July 2012.

Chen, C.-C., & Cheng, S.-H. (2016). Does pay-for-performance benefit patients with multiple chronic conditions? Evidence from a universal coverage health care system. Health Policy and Planning, 31(1), 83-90.

Chen, J.-R., K.-P. Chen, C.-F Chou, and C.-I. Huang (2013). A Dynamic Model of Auctions with Buy-Out Options: Theory and Evidence, Journal of Industrial Economics, 61, 393-429.

Cheng, X., & Liao, Z. (2015). Select the valid and relevant moments: An information-based LASSO for GMM with many moments. Journal of Econometrics, 186(2), 443-464.

Choi, H. & Varian, H. (2012). Predicting the present with Google trends. Econ. Rec. 88, 2-9.

Chou, S.-Y., Grossman, M., & Liu, J.-T. (2014). The impact of National Health Insurance on birth outcomes: A natural experiment in Taiwan. Journal of Development Economics, 111, 75-91.

Chu, C., Chou, T., Hu, S.-S., Kan, K., Cheng, P.-C., Lin, M.-J., Liao, P.-J., Yu, R.-R. (2015). Top Incomes in Taiwan, 1977-2013. WID Working Paper Series No. 6/2015.

Donaldson, D., & Storeygard, A. (2016). The View from Above: Applications of Satellite Data in Economics. Journal of Economic Perspectives, 30(4), 171-198.

Einav, L., & Levin, J. (2014). Economics in the age of big data. Science, 346(6210), 1243089-1243089.

Einav, L., Kuchler. T., Levin, J., Sundaresan, N.(2014). “Learning from seller experiments in online markets,” NBER working paper no. 17385.

Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. (2016). Predictive cities crowdsourcing city government: Using tournaments to improve inspection accuracy. The American Economic Review, 106(5), 114-118.

Heiberger, R.H. (2015). Collective Attention and Stock Prices: Evidence from Google Trends Data on Standard and Poor's 100, PLoS ONE, 10(8): e0135311.

Ho, T. H., Chong, J. K., & Xia, X. (2017). Yellow taxis have fewer accidents than blue taxis because yellow is more visible than blue. Proceedings of the National Academy of Sciences, 114(12), 3074-3078.

Nardo, M., Petracco-Giudici, M., & Naltsidis, M. (2016). Walking down Wall Street with a Tablet: A Survey of Stock Market Predictions Using the Web. Journal of Economic Surveys, 30(2), 356-369.

Ostrovsky, M., & Schwarz, M. (2011, June). Reserve prices in internet advertising auctions: a field experiment. In Proceedings of the 12th ACM conference on Electronic commerce (pp. 59-60). ACM.

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google Trends. Journal of Forecasting, 30(6), 565-578.