
大數據分析與人文社會科學跨領域研究應用

劉吉軒^{*}

摘要

二十一世紀以降，人類社會開始進入一個數位連接與數位沉浸的世代，物質實體世界與數位虛擬世界可以相互介接併聯，資料不僅是重要的無形資產，更是競爭力的來源，人腦與機器也更深入的互動共創。本文介紹大數據基本概念與主要分析技術，回溯資料計算與人文社會科學跨領域發展的趨勢脈絡，並以中華民國官職資料庫與政治研究、二二八事件台灣本地新聞史料彙編與歷史文本修辭語境研究、經濟部商業司商工登記資料與企業交叉持股行為研究等三個研究應用，呈現跨領域研究在資料內容性質、分析方法、探索發現等層面的多元樣貌。基於這些基礎，本文論述大數據與資料計算為人文社會研究所帶來的重要意涵，包括資料的原生性、資料的可操作性、資料的可連結性、資料的可呈現性、資料的可檢驗性及資料的高價值性，最後，本文總結呼籲，大數據時代需要人文社會學者與資料計算學者共同攜手，探討新研究範式中資料、議題、實驗設計驗證、理論模型建構的合理組成與循環修正，才能迎接豐碩的研究成果，進而開創新的知識體系。

關鍵詞：大數據分析、數位人文、計算社會學、跨領域研究

^{*}劉吉軒為國立政治大學創新與創造力研究中心主任及理學院資訊科學系教授。聯絡方式為：
liuj@nccu.edu.tw

壹、緣起

二十一世紀以降，電腦、平板、手機、網路的重度使用成為現代社會的普遍現象，人們的工作與生活進入生產資料與消費資料的快速循環，數位世界的疆域不斷擴大，成為各種社會行為的重要虛擬空間。在可預見的未來，數位世界將與實體世界共同演化，資料的涵蓋率與滲透率也將持續提升。資料為建構數位世界的基本元素，資料不僅是重要的無形資產，更是競爭力的來源。因此，如何掌握資料與善用資料成為具有高度共識的關鍵議題。

近幾年，大數據（big data）成為一個熱門而受到關注的名詞，許多大數據的創新應用，引起社會各界的興趣與重視。例如，Google 曾經推出 Google Flu Trends，以大量網路檢索紀錄資料的分析，監測或預測某一人口地區是否出現流行性感冒擴散的現象，預期能比政府部門透過各地醫療單位回報整合的資料，更快速而且及時的掌握公共衛生狀態（Ginsberg et al., 2009）。Facebook 於 2007 年至 2012 年之間推出 myPersonality 應用，依據使用者填寫的問卷及其在臉書帳戶上表現出的喜好行為（likes），再透過數百萬使用者資料的交叉分析，提供個人性格特質評量的服務（Kosinski et al., 2013）。大量資料的無所不在宛如無形的數位礦產，等待適當工具的開採與挖掘，創意想法的提煉與運用，未來想像充滿了豐富的可能。

大數據是一個抽象概念，象徵著人類社會發展已經進入到一個資料無所不在、各種行為全面紀錄的時代。各專業領域及生活層面的龐大資料不斷生產累積，再加上功能持續提升的資訊設備及日益精進的資料計算分析技術，大數據讓我們開始能以全面、宏觀、長期等角度去看待事物，從中解讀出規律或型態，而能預測未來、改善決策品質。大數據儼然成為創新的重要來源，不僅帶來重大的經濟價值，也將改變人們的生活、工作與思考方式（Mayer-Schönberger & Cukier, 2013）。大數據被視為企業組織改造提升、創新發展、開創利益的重要途徑（Schmarzo, 2013），部分國家政府也啟動大數據應用計畫，目標領域包括醫療照護、自然災害防護、國土安全防衛等（Kim et al., 2014）。大數據成為一個新時代的競逐場域，從個人、

組織到國家，都將是未來技術發展的利害關係人，都必須重視、掌握、運用大數據的能力（McKinsey, 2011）。

大數據的興起同時在四個面向上展現強烈的訊號。第一個面向為權威學術期刊的系列專題刊載，如 *Nature* 期刊從 2008 年 9 月的特別專題開始到 2015 年底，一共刊載 25 篇標題含有”big data”的文章，*Science* 期刊也於 2011 年 2 月以”dealing with data”為主題（*Science*, 2011），刊載一系列探討文章，其後至 2016 年 2 月止，共出現 15 篇標題含有”big data”的學術研究成果。第二個面向為指標性媒體的報導，例如英國經濟學人於 2010 年 2 月的專題中，報導資料的大幅成長與無所不在，並評論此現象對經濟與社會的影響（*The Economics*, 2010）。美國紐約時報於 2012 年 2 月的一篇報導中，宣告大數據時代的來臨（*Lohr*, 2012）。第三個面向為政府部門相繼推動的政策；例如，聯合國於 2009 年發起 *Global Plus* 行動，倡議以全球性的大數據研究增進人類的福祉；美國政府於 2012 年宣布大數據研究與發展的大型投資計劃。第四個面向為產業界的行動，包括 *IBM*、*Google*、*Facebook* 等企業對大數據能力的大量投資，專業研究機構 *Gartner* 也將大數據分析列為最重要的新興技術之一。大數據現象與其潛力得到前所未有的全面性關注。

如同自然的演化，大數據現象也是一個漸進的過程，隨著條件的成熟而發苞萌芽。在大數據之前，以資料觀點出發的相關學術研究早已在不同知識體系中，形成一個交互成長的脈絡。資訊科學（*computer science and information science*）領域的發展有很大一部分持續以 *data*, *information*, *knowledge* 為主要概念元素（*Aamodt & Nygård*, 1995；*Zins*, 2007），並嘗試建立彼此之間的系統性連結方式，從機器學習（*Michalski et al.*, 2013）、資料探勘（*Torra*, 2013）、知識管理（*Kebede*, 2010）到與其他專業領域結合的資訊應用，如 *medical informatics*（*Hersh*, 2002）、*health informatics*（*Dalrymple*, 2011）、*environmental informatics*（*Huang & Chang*, 2003）、*social informatics*（*Kling*, 2007）等。這一路發展、擴散的主軸都是在資料的掌握、資訊處理的駕馭、資料資訊價值的提升與利用。大數據讓這一主軸的演進更加彰顯，影響社會層面更廣，在某些工作與生活層面，甚至帶

來革命性的衝擊 (Kitchin, 2014)。

本文介紹大數據基本概念與主要分析技術，回溯資料計算與人文社會科學跨領域發展的趨勢脈絡，並以中華民國官職資料庫與政治研究、二二八事件台灣本地新聞史料彙編與歷史文本修辭語境研究、經濟部商業司商工登記資料與企業交叉持股行為研究等三個研究應用，呈現跨領域研究在資料內容性質、分析方法、探索發現層面的多元樣貌。基於這些基礎，本文論述大數據與資料計算為人文社會研究所帶來的重要意涵，包括資料的原生性、資料的可操作性、資料的可連結性、資料的可呈現性、資料的可檢驗性及資料的高價值性，最後，總結跨領域研究的展望。

貳、大數據與分析技術

一、大數據概念

大數據的定義通常以五個 V 為特徵：Volume、Variety、Velocity、Veracity、Value，前面三個 V 為原先被提出的基本描述，後來新增的兩個 V 則反映更多的實務面向 (Gandomi & Haider, 2015)。第一個 V 的 Volume 代表著大數據的最基本內涵為龐大的資料量，從 terabytes (2^{40} bytes or 2^{10} GB) 以上到 petabytes (2^{50} bytes) 的資料量，在不同應用領域都可被認定為大數據的範圍。第二個 V 的 Variety 指出各領域中的多元資料類型，從原來的結構性資料到更寬廣的各種非結構性資料，包括文字、圖片、影像、聲音等；與更廣泛的資訊內容，如使用者行為的紀錄、連結關係及各種機器設備狀態的偵測等。第三個 V 的 Velocity 則指生產資料、處理資料到應用資料的速度都在快速的提升。各種資訊設備大量進入到社會的各個層面，讓資料的傳輸與介接更為流暢，也讓各種分析與決策應用從資料的蒐集到採取行動反應，以更快的速度進行迴圈。第四個 V 的 Veracity 代表實務資料中隱含的不可靠性，包括各種不確定、不準確甚至是錯誤的資料，各種分析技術工具必須能適當的管理及利用這些大量但無法保證百分之百正確的資料。最後，第五個 V 的 Value 則是強調潛在的價值，大數據的資料通常是低價值密度，單一少量資料通常無直接明顯用處，但若同一類型的資料能被大量蒐集，甚至與其他相關資料整合，再經適當的分析利用，就可以產生高

度的價值。

大數據讓人類社會開始進入到一個數位連接與數位沉浸的世代，物質實體世界與數位虛擬世界可以相互介接、併聯，人腦與機器也更深入的互動共創。在大數據時代，大量資料的來源大致可分為四類：（1）透過網路進行的各種社會活動，如通訊、社交、言論、娛樂、資訊搜尋、市場交易等；（2）政府或專業機構各種管理面的紀錄資料，如產業發展與生產、所得收入、健保醫療等；（3）機器設備或儀器的感應偵測數據，如工廠運作、城市活動、生態氣候等；（4）學術活動的數位化，包括不斷生產的學術論文、各種原生數位的研究素材及持續回溯數位轉換紙本內容材料等。

大數據分析的主要技術基礎包括資訊學科的資料探勘、文字探勘、社群網絡分析、資訊視覺化等及傳統的統計學工具，各項技術之間也存在彼此搭配整合的可能性與必要性。本文介紹與人文社會研究較直接相關的分析技術為主。

二、文字探勘技術

人文社會科學領域的研究議題通常以人為主體，研究素材也通常牽涉大量文字資料的分析判讀與應用，因此，文字探勘技術實為大數據分析在人文社會科學研究上，非常關鍵的工具與利器。文字探勘與資料探勘同樣是嘗試從大量資料中，挖掘出未知的有用資訊，但更聚焦在人類語言的文字資料上（Feldman & Sanger, 2007）。因此，文字探勘必須結合自然語言處理、計算語言學及機器學習的專業技術，在技術項目上又可以更進一步展開：

- 資訊擷取（information extraction）：依據事先設定或經學習而得的文字表達型態，從大量文字中，自動辨識出人事時地物等語意元素字詞及彼此的對應關係（Doan et al., 2006）。資訊擷取技術能取得大量文件中的關鍵資訊元素，例如人名及其隸屬機關單位名、事件及發生的地點、藥物名稱及副作用等，能將大量非結構文件，快速萃取重要資訊元素，轉化成結構化的資料庫，是知識管理的重要工具，也是其他

文字探勘技術的重要基礎。

- 議題偵測與追蹤 (topic detection and tracking)：從一系列相同議題的文件內容中，透過詞頻與特徵分析，自動建立能代表該議題的關鍵詞組或語言表達模型 (Allan, 2012)。議題偵測與追蹤是一種以事件為基礎的資訊組織技術，用於未來大量新增文件中是否含有特定議題的自動偵測與篩選，而可以讓個人或組織持續追蹤蒐集該議題的最新資訊、被討論的熱度及發展趨勢。
- 意見挖掘與情感分析 (opinion mining and sentiment analysis)：針對特定議題或對象，從大量文件中，自動判別作者對該議題或對象所表達的意見類型或情感面向 (Pang & Lee, 2008)，可以快速的蒐集眾人對特定人物、事件、產品、地點等的支持或反對、喜好或厭惡的整體傾向程度，也可進一步掌握相關評價的描述與看法。

文字探勘技術還包括文件分類與集群／分群 (classification and clustering)、文件歸納總結 (summarization) 等，相關技術可以互相搭配組合，應用範圍也非常廣，包括政治觀察與選舉預測 (Sobkowicz et al., 2012；Tumasjan et al., 2010；Vergeer, 2012)、熱門社會議題偵測與預測 (Li & Wu, 2010)、產品評價與市場預測 (Feldman, 2013；Tetlock, 2007)、緊急災害事件通報與應變協調 (Sakaki et al., 2013)、情報蒐集、市場行銷、顧客關係管理、產業競爭管理等 (Linoff & Berry, 2011；Netzer et al., 2012)。

面對人類語言文字表達的複雜性，文字探勘技術並無法提供百分之百正確的各式分析結果，但通常六成至七成以上的正確率已經可以對大量文本的分析產生很大的效益，尤其在宏觀性資訊的解讀與探討。人文學者可以在議題設定、資訊目標上與資訊技術學者充分討論，盡量降低文字探勘技術應用結果的錯誤率，同時在分析結果含有少部分錯誤的認知下，以專業領域知識為參考框架，審慎思考正確的解讀與結論。這種新研究方法的另一優勢在於建立探索實驗的可能性，每一個實驗週期可以大幅的縮短，議題結論與資料證據之間的關聯可以面向式的反覆檢視、修正，而得到更充分、更完整的研究結果。

三、社群網絡分析與探勘技術

在數位時代中，實體世界與資訊世界並存發展，人事物的運作與互動更加緊密連結。因此，大數據分析中不可或缺的技术為系統性探討群體互動機制與行為的社群網路分析與探勘（social network analysis and mining）（Carrington et al., 2005）。此技術發展的來源為社會學中的社會網路分析（Scott, 1988），針對由一群彼此具有某種關係的個體所組成的群體，以數學圖論（graph theory）中的網路連結模型，探討個體角色、群體結構、運作行為等局部或整體特徵，可以在個體層次或整體層次觀察，可以辨別群體中的核心、邊緣與中介，可以找出緊密連結的次團體，也可以觀測連結的強度、傳遞的路徑與方向等。社會學領域中的社會網路分析傳統上受限於資料取得與資料計算的條件不足，在研究範圍上通常以抽樣調查及由數十個個體組成的團體為限，著重於只需少量計算的分析結果，但在模型機制的建立與內涵意義的解讀上，有非常深厚的探討與發展（Kadushin, 2012）。

隨著數位時代的來臨，社會網路模型也吸引了資訊科學及物理學等不同領域的興趣，而發展出跨領域的社群網路分析與探勘技術，在研究範圍上可以涵蓋上萬個以上成員的群體，可以記錄成員之間的多重關係，可以操作複雜的大量運算，對模型、機制與分析方法也帶來新的視角，而產生新的研究框架與工具，如動態社群網路（dynamic social network）（Skrums & Pemantle, 2009）、社群網路演化（social network evolution）（Doreian & Stokman, 2013）、連結探勘與預測（link mining and link prediction）（Getoor & Diehl, 2005；Liben-Nowell & Kleinberg, 2007）等。另外，網路視覺化技術與工具讓抽象的網路模型與概念，能被具體的圖像化，而且能以互動方式操作檢視，以視覺資訊傳達分析結果，讓角色位置、結構差異、互動輪廓等概念能被清楚呈現，對於探索式研究與應用產生很大的幫助（Freeman, 2012；Brandes et al., 2001）。

社群網路分析與探勘的應用在許多領域上都可以發揮，包括產業生態（Pryke, 2004）、犯罪偵查（Chen et al., 2004）、國土安全（Ressler, 2006）、

政治權力精英集團(Liu, Ning, Chuang, 2013)、廣告行銷(Hill et al., 2006)。在各式社群媒體如 Facebook、Twitter、Instagram 等席捲全球社會下，大規模的社會互動與訊息流動等行為現象的探究，更是社群網路分析與探勘技術可以大力施展與實踐的場域(Ellison, 2007；Aral & Walker, 2012)。

參、資料計算跨領域發展脈絡 — 以數位人文為例

電腦的發明為人類帶來一個功能不斷演化的計算機器，過去數十年電腦工程與資訊科學領域對此計算機器軟硬體功能持續精進，而傳統的自然學科與人文社會學科也在此過程中摸索著以計算機器為新興研究工具，而逐漸形塑資料概念與計算技術向其他學科擴散的跨領域研究發展。此一漫長過程，如同河流的源頭涓涓細流，歷經數十年的延續，隨著各式計算與通訊裝置於 2000 年左右廣泛進入社會的各個層面，大量數位資料不斷堆積，計算技術與機器功能也達到前所未有的境界，資料計算與人文社會學科的跨領域研究乃漸蓬勃發展。例如，數位人文(digital humanities)即是一個近期受到重視的跨領域研究，國際上的學術興趣快速成長、研究社群遞增。熱烈的研究活動反映於 2005 年成立的數位人文組織聯盟(Alliance of Digital Humanities Organizations)、2006 年開始舉辦的 Digital Humanities 年度國際學術會議、2007 年創刊的 Digital Humanities Quarterly 國際學術期刊等。除了這些學術領域的正規奠基之外，許多地區更成立研究機構，正式投入組織資源，宣告並承擔長期推動數位人文的使命與任務。2007 年成立的 centerNet 為數位人文研究機構的國際網絡組織，其成員包括分布於 19 個國家的 100 個研究中心。這些發展趨勢展現了此新興領域的豐沛學術能量，同時也將促成更進一步的資訊與人文跨領域研究的典範轉移。

數位人文的起源其實可以追溯到更早的人文計算領域(Humanities Computing)。如果以國際學術會議、學術期刊、學會組織等三個面向來觀察一個領域的發展歷程，人文計算領域最早的學術會議為 1964 年的 Literary Data Processing，到 1970 年則開始 Computing in the Humanities 系列的研討會。學術期刊方面，Computers and the Humanities 於 1966 年創刊，一直持續出版至 2004 年。另外一個較晚的學術期刊為 1986 年創刊的 Literary and

Linguistic Computing，到目前仍持續出版，並成為此領域最重要、最權威的期刊之一。最後，在學會組織方面，1973 年成立了 Association of Literary and Linguistic Computing (ALLC)，1978 年又成立了 Association of Computers and the Humanities (ACH)。所以，從 1960 年到 1980 年左右大約是數位人文領域的前身，以人文計算為名的階段。數位人文 (Digital Humanities) 的名稱，於 1986 年開始出現在一個新成立的學會組織：Society for Digital Humanities，而在學術會議方面，ALLC 與 ACH 兩個學會組織於 1989 年開始，把原本獨立的年度研討會，整併為聯合會議；此一系列的年會後於 2006 年正式改名為 Digital Humanities，一直延續至今，成為近期最具代表性的跨領域學術社群。

根據 Digital Humanities Quarterly 的定義，數位人文涵蓋於資訊科技之中或透過資訊科技進行人文研究的實踐；及在人文與科技、媒體、計算方法等產生互動的過程中，對人文如何演化進行探索。而如同該權威期刊所觀察，數位人文是一個仍在發展當中、多樣化的新興領域，一些投入耕耘於此研究領域中的學者，也提出了各自的觀察與定義。例如，Davidson (2008) 提議數位人文是以科技與人文兩者共同搭配、整合，進行知識、教育、社群、政治、經濟等面向的實踐；Zorich (2008) 將數位人文界定為以數位技術與資源進行人文研究與教學的智力活動；Frischer (2011) 則認為數位人文是應用資訊科技協助保存、重建、傳送、解讀人類記錄等基本人文任務的完成。另外，Svensson (2010) 對數位人文領域的發展脈絡與現況提出觀察，剖析資訊科技與人文研究不同層面的互動方式與相互影響，並建議未來的可能方向。Presner (2010) 歸納指出第一波的數位人文浪潮是在 1990 年代晚期到 2000 年代初期，著重在大規模的數位化工作與基礎架構的建立。目前第二波的浪潮為數位人文 2.0 (digital humanities 2.0)，重心轉移至原生數位資源的產製、流通和互動所需的數位人文研究的環境與工具開發，開創新興研究場域、研究方法、出版模式，樹立跨領域研究的新典範。

數位人文代表著資訊科技與人文研究在環境條件成熟後，以更大的能量、更大的規模彼此再次交會，在名稱上雖以人文為代表，但在內容上也

涵蓋社會學科資料的計算與議題的探討，在性質上可視為資訊科技領域與人文社會研究追求共同目標的對等融合，因而開創真正的新興跨領域學門。在類似的跨領域脈絡下，社會科學與計算方法的結合也發展出計算社會科學（*computational social sciences*）、社會計算（*social computing*）、社會資訊學（*social informatics*）等新興分枝學門，不僅開拓了傳統的研究疆域，也帶來令人期待的新可能。例如，Conte et al. (2012) 觀察到日常社會行為的大量資料，提供了對複雜的社會系統進行量化的機會，透過計算社會科學的新研究範式，將展開一個可以對社會結構與功能在不同層次上解讀的時代，而更有機會發展出有助於社會議題良善決策的有用工具。Tinati et al. (2014) 代表新世代社會學者，嘗試突破傳統社會科學研究方法限制，提出以大數據結合適當的分析工具，以回應許多社會研究議題的規模、延展及動態的挑戰。另外，Facebook 也陸續提供社群互動與用戶行為大數據為學術研究使用，成為許多社會學門的新興研究工具與平台（Wilson et al., 2012）。

不論是數位人文或是計算社會科學，都是認知到資料計算跨領域結合的潛力而長期投入的努力。進入大數據的時代，資料的鋪天蓋地與精進計算技術結合的效益，再次以更強大的力量衝擊了人文社會學術體系。例如，Berry (2011) 認為在目前世界萬物的數位摺疊與包覆下，人文社會研究必須以新的視角去考慮數位表達與數位媒介的彈性，及對文化研究與認知所帶來的新方法與新機會，對傳統人文社會領域知識系統也將帶來根本的轉變。Kitchin (2014) 深入評論大數據將帶來的新知識與新範式，認為大數據結合新的分析方法，將挑戰人文與社會學科既有的知識體系，而以一種新型態的經驗主義，終結純理論模式，展開資料驅動而非知識驅動的科學研究，以完全不同的方式理解文化、歷史、經濟與社會。因此，我們可以確認資料計算與跨領域整合不僅是一個不可逆轉的長期趨勢，更將是未來人文社會創新發展不可或缺的核心學術能力。

肆、資料計算跨領域研究應用案例

大數據與人文社會科學跨領域發展脈絡論述，提供了通則性與概括性

切入觀點，具體研究應用案例則更有助於認知理解大數據與人文社會科學研究相互連結的各種可能性。以下三個研究應用案例，部分為已經發表過的研究成果，部分則為持續進行之研究議題，但皆以精簡內容扼要介紹，分為資料、計算分析方法與觀察解讀等三個面向，提供為人文社會科學與資料計算跨領域研究的參考。

一、中華民國政府官職資料庫與政治研究

中華民國政府官職資料庫（劉吉軒，2012）收錄中華民國體制以來七個時期總統府公報之人事異動命令，內容涵蓋民國元年迄今任職於政府機關之各職官任免概況，從北洋政府、軍政府、國民政府，直至中央政府遷臺、行憲後總統府等時期，因政府體制之沿革而留下不同公報歷史紀錄，期間所累積的文武官員任免資料相當豐富，是國內少數記錄中華民國政府官員任職動態的公開性資料庫，以超過百年的歷史紀錄，反映著國家機關在歷史發展各階段中，重要行政職位的人事權概況，更隱含著政府官職制度的內涵與行政組織架構的變遷。政府官員任免資料的來源是刊載於政府公報中公告發布之官員任免總統命令（如圖一、圖二），紙本原始文件由國家圖書館建置為文件掃描影像，收錄於政府公報資料庫，可供主題檢索與閱覽。

許多文史主題之資料庫大都以收錄書刊文件之掃描影像為主，提供保存與取用之便利性，但其全文影像內容僅能由人工閱讀，大量細部主題資料仍難以取得使用，對於以計算技術為主之資料分析方法較難有發揮空間，而限制了許多人文社會學科的實證研究。

中華民國政府官職資料庫在初始設計上，即考量數位資料的應用價值，而以每一位官員個別的任免命令為資料單位，以後設資料摘錄任免命令中的重要資訊，如官員姓名、任職機關單位名稱、職位、職等、任免情形、日期等，將全文影像轉換為以精細的欄位資料為資訊組織單元，各欄位中的內容則是可由程式處理的文字資料，因而可以透過計算技術的協助，進行多面向多角度的切割整合，不論是特定條件的搜尋比對，或是更進階的資料探勘與發掘，都可以有效支援學者開創以資料為核心的實證分析研

究。

中華民國政府官職資料庫以橫跨百年的歷史縱深，收錄超過 90 萬筆官員異動資料，讓資料能脫離紙本的平面，進入多維度的數位空間中，而可依照使用者的需求自由切割組合，進而提供多角度、多面向的大數據觀察分析能力。本文摘錄過去的研究成果，以展示大量資料分析應用的部分意涵。

圖一：民國 37 年 5 月 21 日出刊之第 2 號總統府公報

總統府公報		
局五第府統總輯編	號貳第	元千一第份每售零
室報公局五第府統總行發	(半張一報公號本)	元萬一十第年售
廠工刷印局五第府統總刷印	類紙聞新類三第為認記登政郵華中經	元萬一十三年全價
加另外國及號掛內在費郵寄平內國		

國民政府令補登

三十七年五月十九日(補登)

國民政府現經改組，國民政府主席各地行轅，應隨同調職，所有東北行轅之職權與業務，著即歸併東北剿匪總司令部，北平行轅之職權與業務，著即歸併華北剿匪總司令部，武漢、西北、重慶、廣州各行轅，均著改為綏靖公署。此令。

國民政府令

三十七年五月十九日(補登)

查各省市縣參議會選舉補充規程第三條所選定之江蘇等省及南京等市參議員名單，前經明令公布在案，現以湖北省參議員袁繼德、漢口市參議員吳時俊均請辭職，自應分別另行遴選，茲將遴補名單公布之。此令。

湖北省
謝夢飛
漢口市
方顯庭

國民政府令

三十七年五月十九日(補登)

任命尤時初為內政部會計長。此令。

任命費繼洪為行政院新聞局秘書。此令。

任命主任民為北平市政府財政局局長。此令。

張敏之著以青島市政府參事試用。此令。

湯之鏡著以漢口市政府財政局局長試用。此令。

國民政府文官處呈，請任命龔公石、沈開選為科員。應照准。此令。

國民政府主計處呈，為歲計局科長尤時初另有任用，請免本職。應照准。此令。

國民政府主計處呈，請任命吳善揚為外交部會計處科長。應照准。此令。

國民政府主計處呈，請任命楊射斗補理財政部重慶貨物稅局會計主任職務。應照准。此令。

國民政府主計處呈，請任命沈源泉為國立編譯館會計主任。應照准。此令。

國民政府主計處呈，請任命景漢士為國立中央圖書館會計主任。應照准。此令。

國民政府主計處呈，請任命朱彥傑為國立政治大學會計主任。應照准。此令。

國民政府主計處呈，請任命董銘為國立四川大學會計主任。應照准。此令。

國民政府主計處呈，請任命林越為國立貴陽師範學院會計主任。應照准。此令。

國民政府主計處呈，請任命程翰丞為國立女子師範學院會計主任。應照准。此令。

國民政府主計處呈，請任命劉宗模為衛生部南京精神防治院會計主任。應照准。此令。

圖二：民國 67 年 4 月 4 日出刊之第 3338 號總統府公報

總統府公報 第三三八號

總統令

六十七年四月四日

總統令

台灣台北監獄典獄長張齊誠，台灣台中監獄典獄長趙益藩，台灣台南監獄典獄長王肇泰，台灣高雄監獄典獄長黃慶銓，另有任用，均應予免職。

派林能中為經濟部國際合作處秘書。

經濟部工業局技正吳容明另有任用；經濟部工業局技術處專員徐世光，呈請辭職。均應予免職。

交通部觀光局專門委員蔡益昌已准退休，應予免職。

交通部民用航空局飛航服務總台主任報務員熊剛已准退休，應予免職。

交通部台灣區高速公路工程局北區工程處繁工程司兼工務段副段長沈觀挺另有任用，應予免職。

任命沈觀挺為交通部台灣區高速公路工程局北區工程處副工程司兼工務段段長。

任命傅朝湘為台灣省政府農林廳林務局技正，潘世磐、路金鏐為台灣省物資局專門委員。

台北市大安區公所課長已馳，台北市雙園區公所課長黎國安，另有任用；台北市建成區公所主任秘書趙希江已准退休。均應予免職。

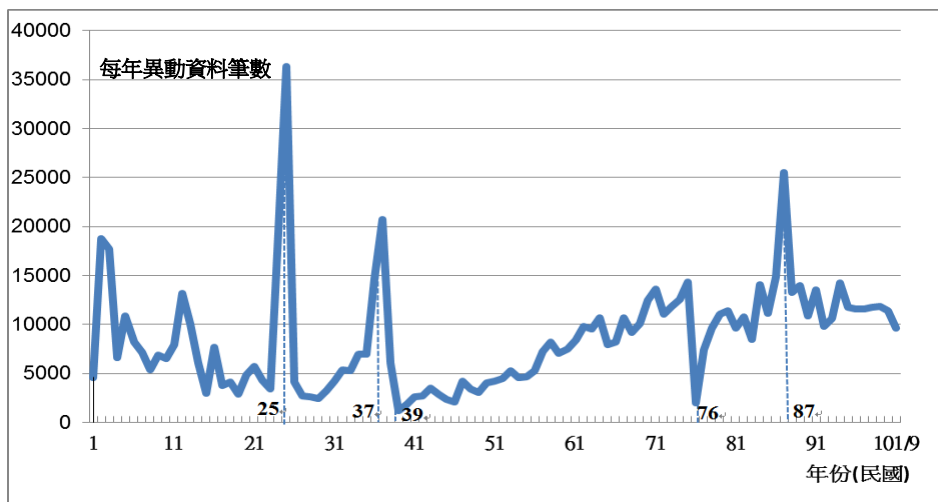
任命朱考權為台北市大安區公所視導。

總統 嚴家淦
行政院院長 蔣經國

1. 宏觀與微觀檢視分析

大量而長期涵蓋的資料提供了以數量角度，描繪特定事物在某個面向的全貌，包括變化趨勢、顯著局部等。透過中華民國政府官職資料庫的年度異動總次數統計，即可得到政府百年來的官員異動頻率變化綜覽，各年度總次數的折線圖（圖三）顯示出不同階段的變化特徵及明顯異常的高峰與低點，似乎代表著各時期的政治情勢發展差異及特定時間點發生之重大政治事件，可供微觀政治史研究之佐證或進一步驗證。本文試舉以下五個特殊年度為例，提供當時相關政治情勢背景：

圖三：民國元年至民國 101 年 9 月整體政府人事異動數量變化



- 民國 25 年的政府官員總異動次數超過 3 萬 5 千次，為歷史上最高峰，可能背景包括：國共內戰的情勢正值緊繃，還即將面臨日本的軍事入侵，國民政府與邊疆自治地區進行政治結盟，並大幅擴展軍事編制，政府組織歷經前所未有的規模擴張與人員更替。
- 民國 37 年政府官員總異動次數超過 2 萬次，為歷史上第三高峰，可能背景為：對日抗戰勝利後，國民大會於民國 36 年 12 月正式通過中華民國憲法，展開憲法制度的施行，因此，民國 37 年為完整的政府組織架構之始，產生大量的官員派任與異動。
- 民國 39 年的政府官員異動次數不到 2 千次，為歷史上之谷底，相關背景包括：國民政府於民國 38 年末正式播遷臺灣；民國 39 年以臺北為行政首都，縮減整體政府編制，同時因該年實施「動員戡亂時期臨時條款」與戒嚴令，而僅有少量的人事異動。
- 民國 76 年的政府官員異動次數亦為歷史上的另一低點，可能背景包括：當年 7 月，時任總統的蔣經國先生宣布解除戒嚴令，其後健康情形惡化，並於民國 77 年 1 月病逝。因此，民國 76 年的政府人事似乎有凍結情形。

- 民國 87 年的政府官員總異動次數超過 2 萬 5 千次，為歷史上第二高峰，可能背景為：中華民國政府於當年通過省虛級化（精省）方案，將原省政府所屬機關功能與人員整併於中央及地方政府，人事異動數量因此大幅攀升。

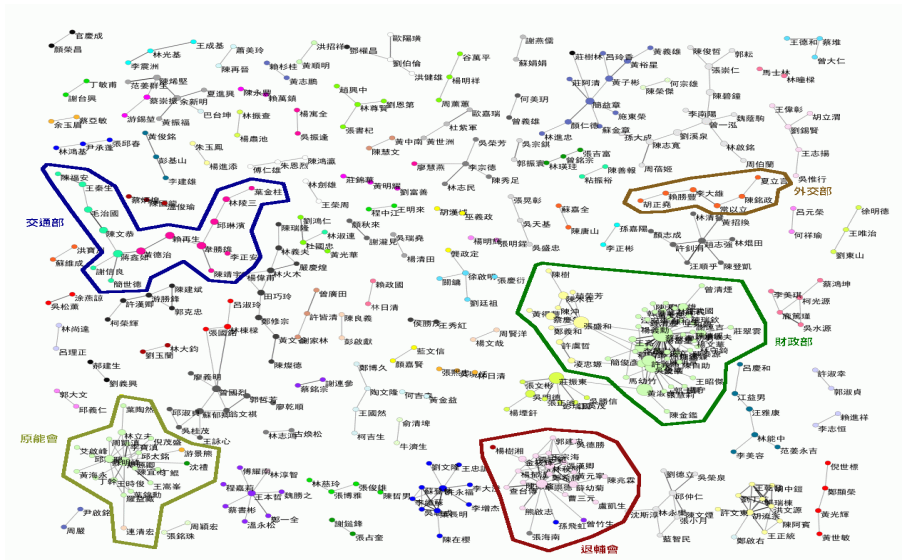
2. 資料視覺化

大量數位資料結合適當的分析模型及視覺化工具，可以將事物的複雜脈絡表示為圖形樣貌，透過人類的圖像認知能力，深入觀察系統中隱含的型態資訊。例如，以特定時間範圍及特定職等之內，高階官員的共同異動情形，透過關聯網路模型彙整相關資料，再以視覺化工具呈現，即可展示出部分政府官員職位接替的連動樣態（圖四），其中數個部會顯示出較為顯著的局部群聚，如左上方的交通部、左下方的原能會、右上方的外交部、右邊中間的財政部及中間下方的退輔會等。

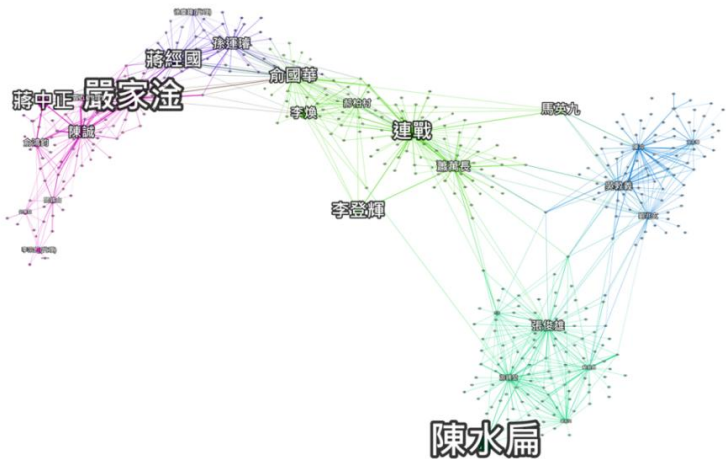
若將總統、行政院長及部會首長之間的上司與下屬關係資料，同樣建立為關聯網路模型，即可浮現出政府高階行政權力之演化脈絡（圖五），從早期國民黨政權的一脈相傳，直到李登輝、連戰、蕭萬長之後，出現政權轉移到民進黨的陳水扁、游錫堃、蘇貞昌、張俊雄等，之後再次的政權轉移回到國民黨的馬英九、劉兆玄、吳敦義、陳冲等。在兩次的政黨輪替之間，部分官員在各政權中連續出任部會首長以上之職務，如橫跨國民黨政權、民進黨政權、國民黨政權的彭淮南、張博雅（分別位於三角結構內部中的兩個明顯中心點）等，及橫跨兩個國民黨政權的馬英九等，而形成了一個三角形結構的權力演化網絡。

不論是官員職位連動樣態或是政府高階行政權力演化脈絡，都是以資料視覺化工具，展現主題事物的圖像樣貌，輔助研究議題的探索性認知。透過研究設計、資料範圍及實驗參數的調整等，可以很快地看到不同情境下的樣貌變化，進行比較性的觀察分析，而對事物的內涵得到更接近真實的掌握與理解。

圖四：官員職位連動樣態



圖五：政府高階行政權力演化脈絡。



3. 資料探勘與分析應用

文官職等升遷預測研究（劉吉軒、賴隆平，2011）利用政府官職資料庫中許多個別文官之職務變動與職等升遷序列，嘗試從一個文官過去的職務變動與職等升遷序列，預測其未來數年內可否升遷至更高職等。此研究將官員職等升遷之預測視為機器學習與資料探勘領域中的分類問題，並採用支撐向量機與決策樹為分類演算法，以過去已發生之官員職務變動與職等升遷序列為訓練資料，建立有效的分類判斷模型，而能對尚未發生的官員職等升遷做出預測。本研究選取資料庫中近二十年有足夠資歷之文官族群為實驗對象，實驗結果顯示，不論是支撐向量機或是決策樹皆可達到 77% 以上之預測正確率。此研究結果似乎代表：（1）中高階文官的職等年資與職務資歷，皆為職等晉升之重要關鍵因子；（2）我國中高階文官的陞遷具有相當程度的穩定性與合理性。

高階機關首長資格評選與預測研究（Liu & Ning, 2011）以機關單位歷任首長及其擔任首長之前的職務經歷資料，透過社會網路分析模型，根據官員資歷資料，建立官員與職位的相互連結關係，再以連結預測理論及節點相似度演算法，針對特定機關首長出缺時，對可能人選資格進行評量排序，並預測最可能獲得派任之官員名單序列。此研究之基本假設為，與歷任首長資歷之綜合相似度愈高之官員，因具備良好之晉升條件，愈可能獲得派任之機會。實驗發現愈高階的文官首長職位，愈能準確評估與預測人選，而各種高階職位的整體評估預測準確率是在可接受的範圍。此研究結果似乎也驗證了我國現代文官系統遵循穩定的功績晉升制度。

政治權力菁英團派觀察研究（Liu, Ning, Chuang, 2013）之背景假設為政治權力之行使通常以政府官職為依託，包括人事權與政策之主導。政府組織的運作架構包括垂直指揮與水平分工，因此，政治權力菁英通常也以具有某種凝聚力的團派方式共同發展。目前臺灣已經透過選舉進行政治權力的轉換與接替，高階政府官職的異動也通常伴隨著中央政府與地方政府首長的新任或階段性的人事權行使。本研究嘗試以社會網路模型與分析方法，先建構高階政府官員職位異動的連動關係，辨識具有顯著職位連動關

係的官員群組，再藉由一連串時間點之間的連動官員群組之變化分析，找出真正能持續存在的政治精英團派。實驗結果發現專業部會中接班梯隊的存在；另外，針對從臺北市市長到總統的政治領袖，本研究透過政府官職資料庫的官員異動資料，也可觀測到以該政治領袖為核心的政治精英團派。

二、二二八事件台灣本地新聞史料彙編與歷史文本修辭語境研究

二二八事件台灣本地新聞史料彙編，為財團法人二二八事件紀念基金會委託政治大學新聞系林元輝教授，彙編事件發生期間官方及民營報紙的相關報導，發行出版為公共使用之電子書。本研究將電子書內容轉換為全文字之文本，成為歷史文本語料研究之素材。文本研究通常著眼於語言文字使用行為的議題探討，相較於其他類型資料，數位文本是大量文字內容的非結構化資料，需要自然語言處理技術的協助，完成字詞單元的辨識及語句結構的拆解，才能更進一步進行文字探勘的應用研究。中文文本通常可以利用中研院中文斷詞系統（CKIP）的線上服務功能，建立文本語料的斷詞及詞類標記結果（Chang & Chen, 1995）。另外，中研院的漢語平衡語料庫（Sinica Corpus）蒐錄大量文章，兼顧主題領域的分布，提供文句的斷詞及詞性標記（Huang & Chen, 1992），最新的 4.0 版達 1100 萬目詞以上，可做為現代漢語語句的代表性大樣本。

歷史文本修辭語境研究（Liu, Ning, Lee, 2016）探討二二八事件台灣本地新聞史料中情態詞彙的使用，觀察當時政府在面臨治理危機時，透過報紙向社會大眾宣傳的修辭，並推論其所代表的立場及態度。研究理論立基於語言學中漢語情態語意框架（Li, 2004）：第一類為認知不確定型

（epistemic uncertainty），以客觀的可能性估量某事物是否會成真，包含的情態用詞有「能」、「能夠」、「會」、「可」、「可以」、「得」；第二類為認知預測型（epistemic probability）：預測或評論一個特定事件或狀態的必要性，表現的情態用詞包含「該」、「應該」、「要」、「得」；第三類的能力型（ability），表示參與者主觀預估的內在能力、功能、性質或品質，情態詞彙為「能」、「能夠」、「會」、「可」、「可以」、「得」；第四類的需求型（need），為事件

參與者主觀的內在需求，與希望、企圖、利益有關，而導致一個動作或事件的發生。情態詞彙包含「要」、「需要」、「須」、「必須」、「得」；第五、六類的允許與間接可能型（*permission and circumstantial possibility*），表達事件成真的可能性，基於道義的法則、規定、權威或非道義的客觀情況。情態詞彙包括「能」、「能夠」、「可」、「可以」、「得」；第七、八類的義務與間接必要型（*obligation and circumstantial necessity*），表示義務的必要性，來源可能是道德觀或社會習俗。情態詞彙包含「要」、「該」、「應該」、「應」、「當」、「應當」、「須」、「必須」、「得」。

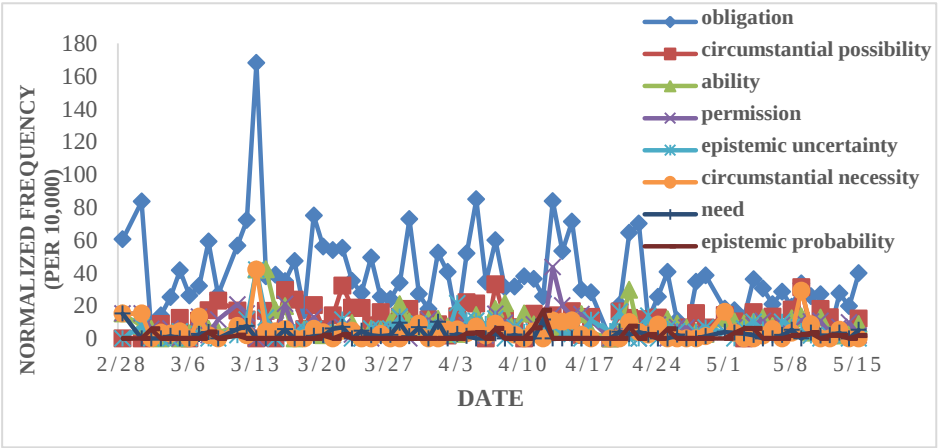
本研究以當時兩大官報之一的台灣新生報為觀察對象，並以中央研究院的漢語平衡語料庫 3.1 版，做為大樣本的參照基礎。台灣新生報文本語料共有 23 萬 7 千目詞，平衡語料庫 3.1 版則包含 573 萬目詞，多出了約 24 倍的詞量。兩個文本語料之間的比較，可以計算標的詞彙在個別文本中每百萬字出現次數的正規化頻率，得到相互校準的基礎。表一顯示在台灣新生報中，「應」、「應該」與「須」、「必須」的使用頻率遠大於平衡語料庫及其新聞子語料庫，可以觀察到官方言論的修辭行為表現出鮮明的態度與強烈的立場。

本研究進一步判讀台灣新生報文本中情態詞彙於情態框架中的分類，統計各類別出現的次數，以觀察不同態度及立場的分布。圖六、圖七分別顯示各情態類別於每日、每週詞彙正規化頻率之變化趨勢。結果顯示台灣新生報文本中的修辭行為有很大的比例集中在「義務」類型的情態使用，反映出官方言論中對廣大民眾的強烈社會責任要求與極力說服。以時間而言，義務情態使用快速於第二週達到高峰，接著在第六週、第七週再次成長，而後逐漸消減，似乎能呼應社會動亂爆發後官方試圖迅速恢復秩序的宣傳與遊說過程。整體而言，本研究觀察到台灣新生報修辭語境中，對於二二八事件發生後的社會秩序控制與恢復，展現出強烈態度及堅定立場。另外，本研究也顯示文本情態詞彙的量化分析，有助於從敘事語境中辨別說話者的立場，同時，提供了一種觀察語料情境之有效方法。

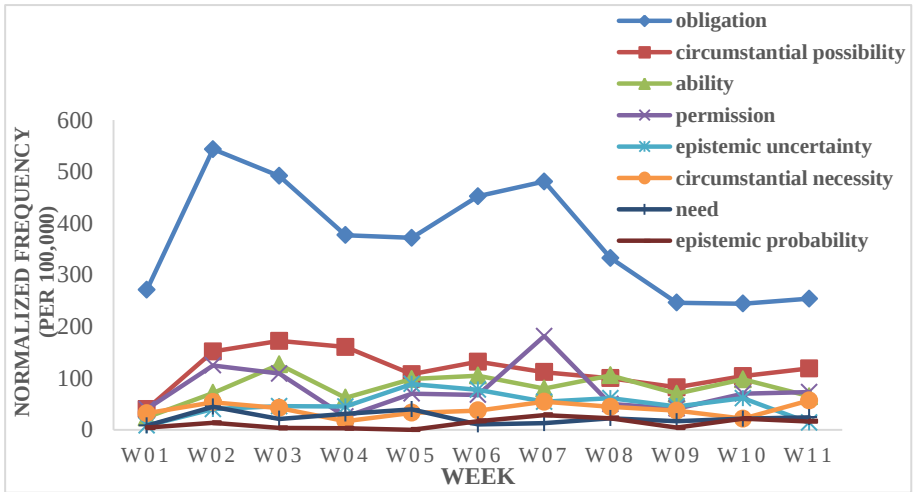
表一：情態語意使用量的跨文本比較

情態詞彙	台灣新生報		中研院語料庫		增加 比例	中研院語料庫 新聞類文章		增加 比例
	絕對	正規	絕對	正規		絕對	正規	
	頻率	頻率	頻率	頻率		頻率	頻率	
應, 應該	673	2839.7	6037	1052.1	269.9%	1750	1038.4	273.5%
可, 可以	622	2624.5	17864	3113.3	84.3%	4420	2622.6	100.1%
要	476	2008.4	15783	2750.7	73.0%	3135	1860.2	108.0%
能	403	1700.4	10867	1893.9	89.8%	3038	1802.6	94.3%
須, 必須	293	1236.3	3967	691.4	178.8%	1025	608.2	203.3%

圖六：情態類型正規化次數每日變化



圖七：情態類型正規化次數每週變化



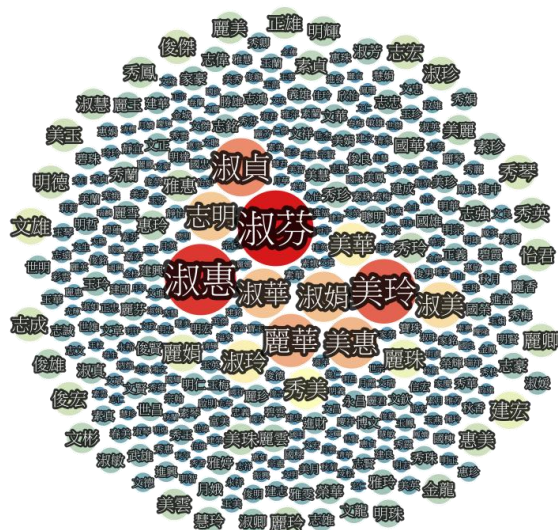
三、經濟部商業司商工登記資料與企業相互持股行為研究

在政府開放資料（government open data）政策的推動下，經濟部商業司定期公布台灣公司登記、商業登記、工廠登記等商工行政資料，包括公司統一編號、公司名稱、資本額、董事會成員、各成員出資額或持股比率等，近期公布之版本已達 300 萬筆以上之登記資料。目前已有部分廠商開發商工行政資料整合查詢的服務，另外，民間組織台灣零時政府（g0v）也利用視覺化技術，開發了互動式的台灣公司網路圖，可提供線上查詢探索，檢視特定公司與其他公司的依存關係（陳佩君，2015）。

本研究以學術角度，嘗試利用商工行政資料於各種議題之探索。首先，300 多萬筆之資料必須依研究目的先進行清理及格式轉換，我們先選取其中的公司登記資料約 147 萬筆，再依據使用登記狀況為核准設立、重組、核准報備、核准認許等四類選取，得到約 62 萬筆公司登記資料，若再依照實收資本額登載大於 5000 萬以上之準上櫃公司條件，則剩約 3 萬 2 千家公司。這些商工行政的實際資料，隱藏了許多可能有意義的資訊。例如，62 萬筆公司登記的董事會成員名字，提供了一個從事特定社會活動族群名字廣泛特徵的觀察機會。圖八以文字雲概念利用 Gephi 視覺化工具軟體，呈

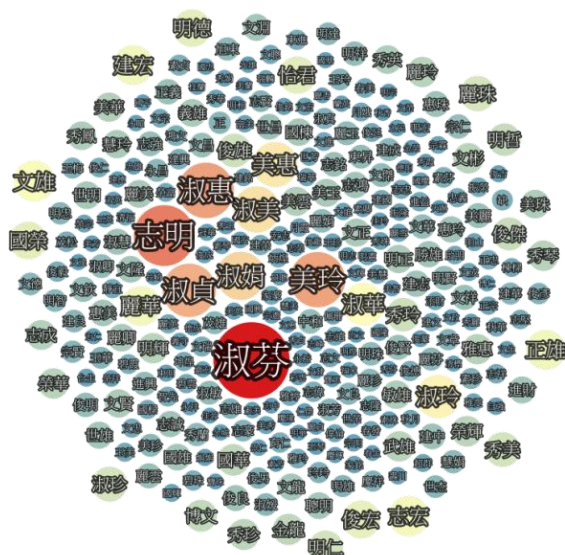
現一般公司高階經營族群去除姓氏之後的名字分布，字型愈大則頻率愈高。全部登記姓名共約 121 萬 6 千餘個，頻率最高的前三名分別為淑芬約 2000 餘次、淑惠約 1900 餘次、美玲約 1700 百餘次。

圖八：一般公司高階經營族群名字分布

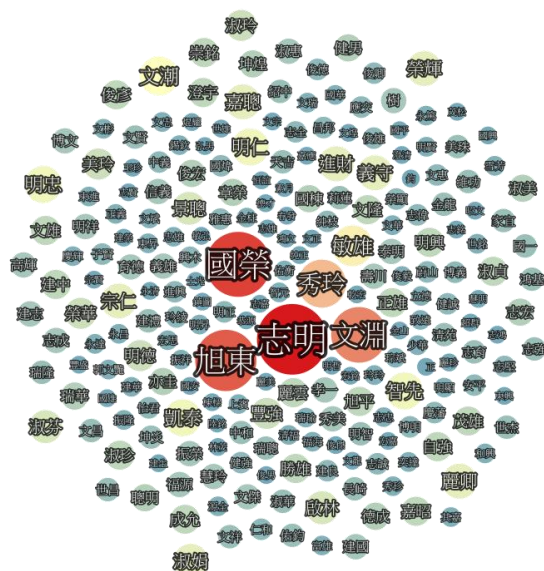


圖九為準上櫃公司高階經營族群的名字群像，全部登記姓名共約 13 萬 7 千餘個，頻率最高的前三名分別為淑芬約 220 餘次、志明約 170 餘次、淑惠約 150 餘次。圖十為登記實收資本額 10 億以上公司（約 2100 家）高階經營族群的名字群像，全部登記姓名共約 1 萬 5800 餘個，頻率最高的前三名分別為志明 35 次、國榮 32 次、旭東 30 次。從名字群像中，我們可以觀察到顯著分布之常用名字及性別優勢，也可推測此特定族群的年齡區間。另外，隨著公司規模的擴大，高階經營族群的常用名字也開始改變性別優勢。這是大量資料透過資料篩選條件設定，並搭配視覺化技術，可以快速浮現顯著資訊及相關差異的簡單範例。

圖九：準上櫃公司高階經營族群名字分布

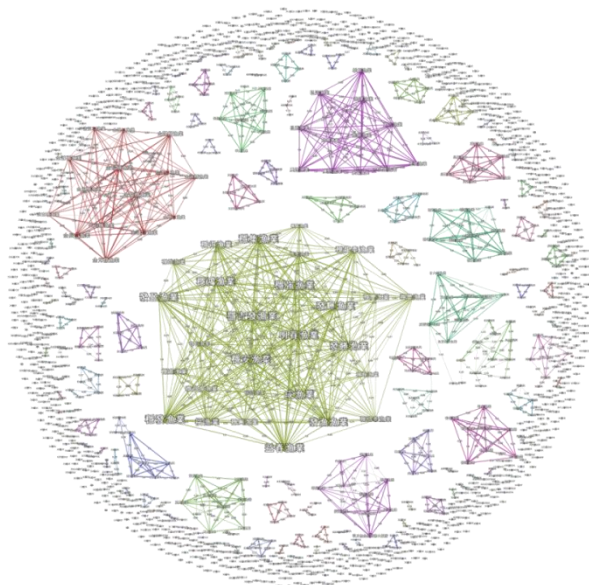


圖十：實收資本額十億以上公司高階經營族群名字分布

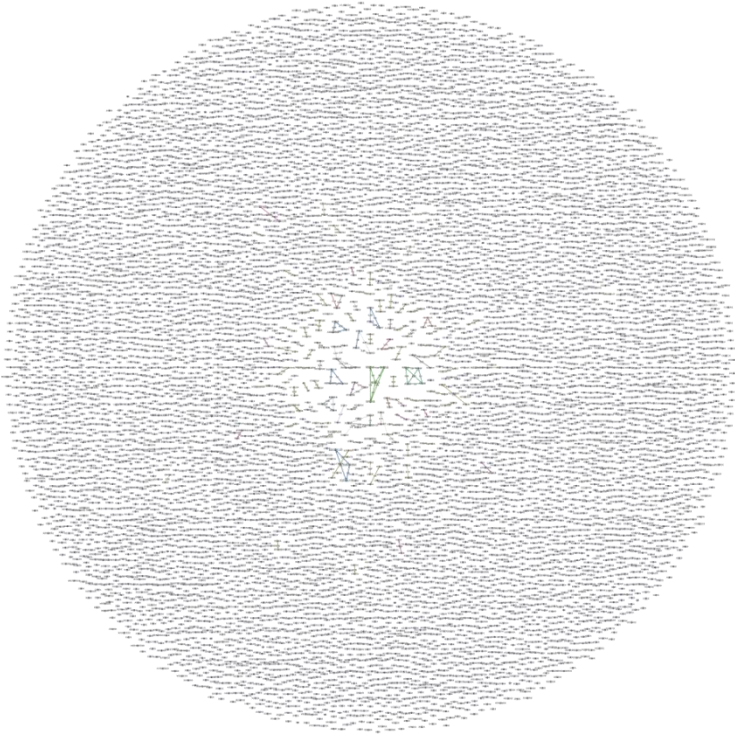


企業交叉持股研究（王連成等，2015）探討台灣企業之間的交叉持股行為，依據約 62 萬筆公司登記資料，計算出兩家公司之間的共同董事會成員及相互持股比率，再以社群網絡分析與探勘技術，建立公司之間的資本關係網絡。公司法規定，一家公司持有另一家公司超過百分之五十以上的股份或資本額，即具有從屬控制關係，稱為關係企業。本研究先以公司名稱後半段之特徵字串及其登記營業項目資料進行產業分類，嘗試觀察特定產業中關係企業群體及非關係企業群體的分布與結構，初步抽樣九個產業：生醫、建設、科技、食品、漁產、工業、旅遊、交通、投資，發現其中最大的三個產業為科技業（5 萬 8 千多家）、工業（5 萬 5 千多家）及金融投資業（3 萬 9 千多家）；最小的三個產業為漁業（1 千多家）、旅遊業（4 千多家）及食品業（6 千多家）。各產業中關係企業群體所占比例最高者為漁業（19.2%）、最低者為食品業（1.7%）。圖十一顯示台灣漁業的資本關係網絡，獨立的點代表非關係企業的群體，相互連結的點則為關係企業群體，其中似乎存在三大集團。圖十二則為台灣食品業的資本關係網絡，絕大部分為獨立企業，具有資本關係的企業非常稀疏。

圖十一：台灣漁業資本關係網絡

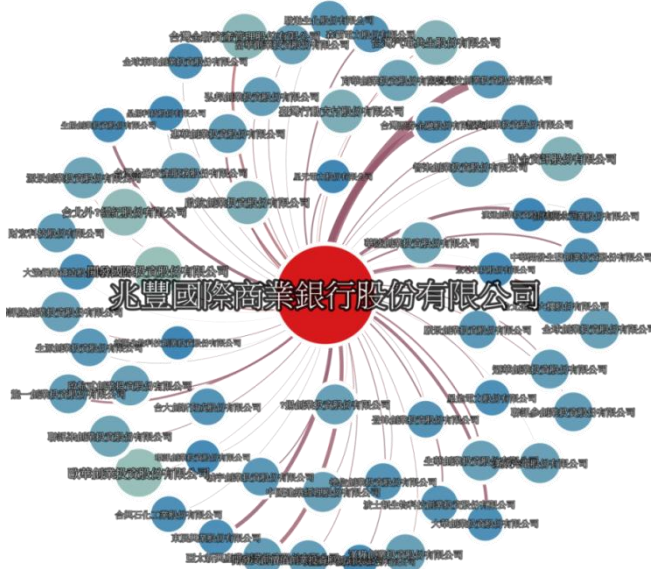


圖十二：台灣食品業資本關係網絡

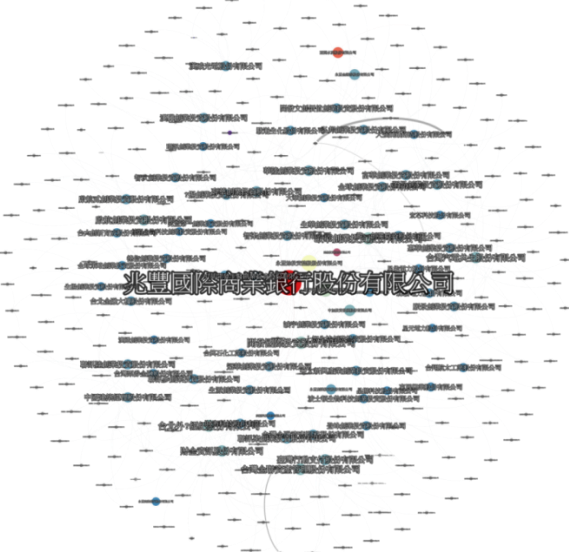


企業集團與地理脈絡研究（陳俊宏、劉吉軒，2015）探討準上櫃公司的關係企業群體規模及地理脈絡現象，依據約 3 萬 2 千家準上櫃公司登記資料，透過董事會成員之法人代表及持股比率，建立準上櫃公司之間的資本關係網絡。研究發現，可能的持股集團中心前三名為兆豐商銀、裕隆汽車、統一企業，其中兆豐商銀直接投資其他 60 家準上櫃公司，如圖十三所示。若將投資與被投資關係一併考量，以兆豐商銀為中心、關聯度為 2 層的資本關係網絡共計有 422 家公司，如圖十四所示；若繼續展開至 n 個連結拓撲，整個資本關係網絡共計 7,838 家公司，佔準上櫃公司比例達 24%，如圖十五所示。

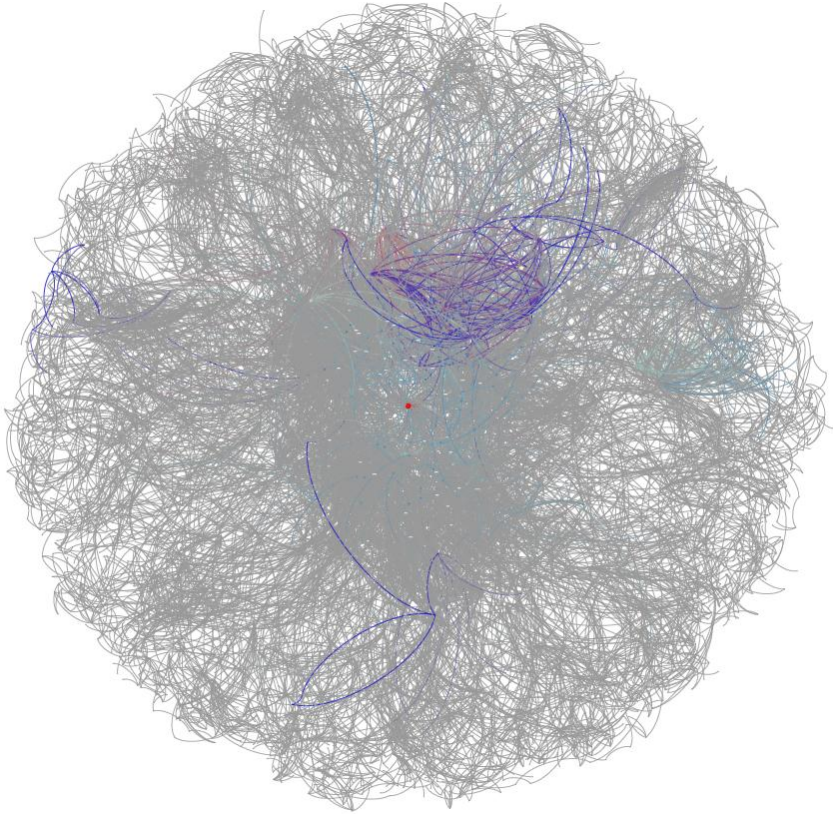
圖十三：兆豐商銀直接投資之準上櫃公司網絡



圖十四：兆豐商銀資本關係 2 層連結網絡



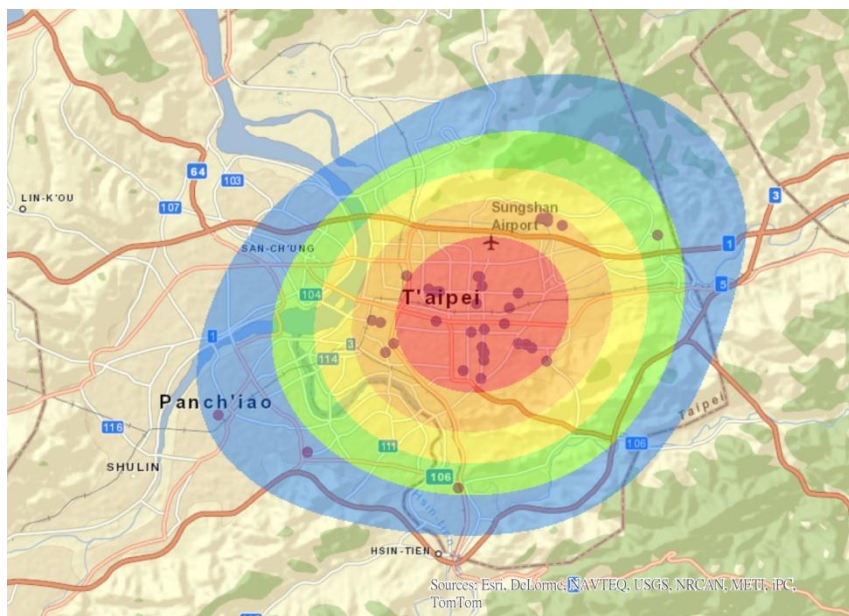
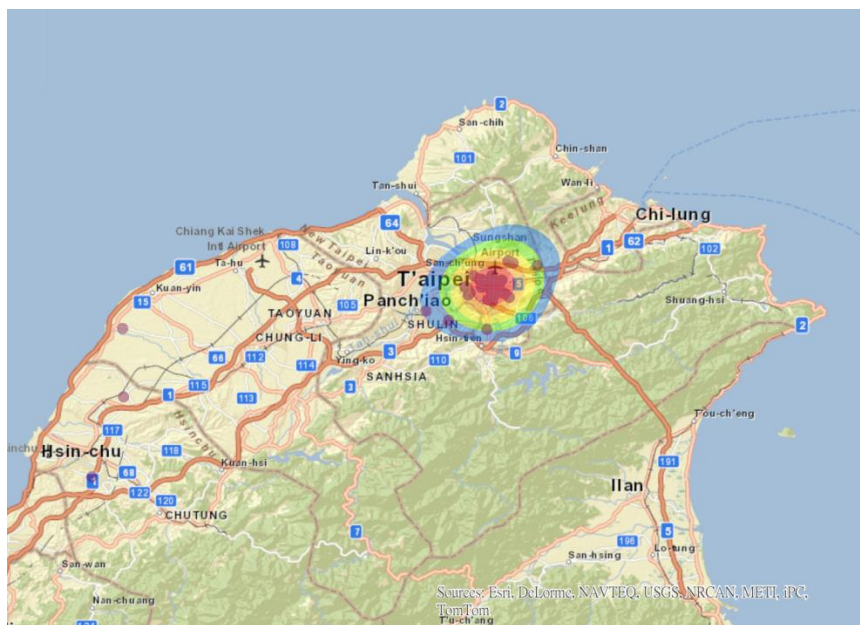
圖十五：兆豐商銀資本關係 n 層連結網絡



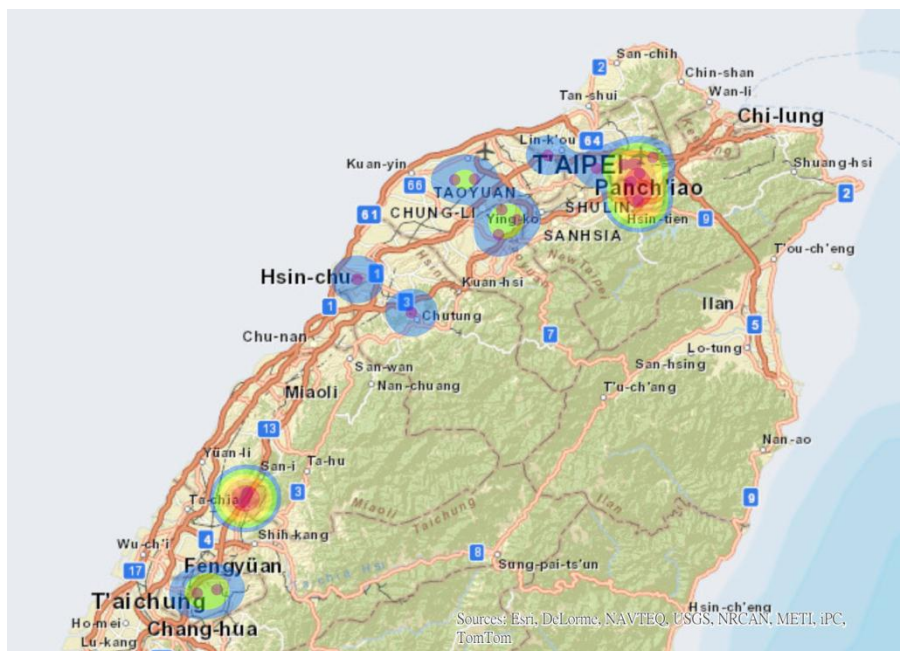
本研究進一步探討投資對象與地理空間聚落的關聯，以前三大持股集團中心為觀察焦點，發現各中心的直接投資對象有明顯的地理群聚現象：兆豐商銀的直接投資聚落為大台北地區（圖十六），裕隆汽車的直接投資聚落分布於中北部（圖十七），統一企業的直接投資聚落則集中於台南地區（圖十八），似乎顯示企業進行投資時，會部分考慮地理位置因素。

以上三個研究主題分析應用案例，初步展示了人文社會議題與資料計算框架跨領域結合的可能性與多樣性，更專精的研究議題與更深入的研究發現有待相關領域學者的嚴謹剖析與提煉。

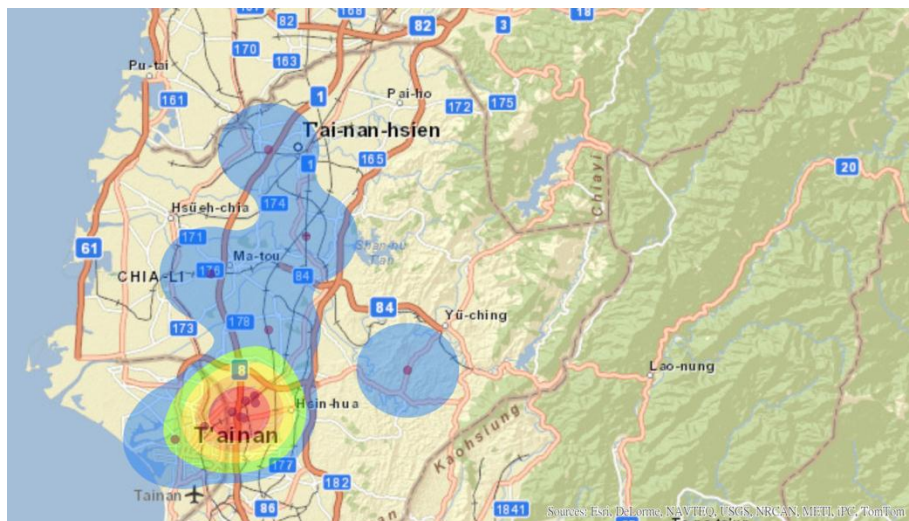
圖十六：兆豐商銀直接投資地理聚落為大台北地區



圖十七：裕隆汽車直接投資地理聚落為台灣中北部



圖十八：統一企業直接投資地理聚落為台南地區



伍、資料計算與大數據意涵

人文社會研究以人為主體，針對人的思想、情感、創作、行為、活動、體制等層面進行探討與思辯，研究資料通常為人類個體或群體的各項表現紀錄，在資料利用上，大致以文獻素材的消化反芻及抽樣調查或局部觀測數據的統計模型分析為主，研究限制條件包括資料的覆蓋率、資料處理的成本及人腦對資料接收與解讀的認知能力等。而在現今數位世界中的資料計算與大數據框架下，許多現象與行為成為可以被廣泛紀錄觀察，語言文字也被轉化為可供計算分析的資料，個人的生活樣貌及人群之間的互動更是豐富的、包覆式的展現並保存於各項社交工具平台，人文社會的知識體系正面臨鋪天蓋地的數位浪潮衝擊，挑戰了傳統的資料認知與研究方法，但也帶來巨大的、等待開發的潛能，並將啟動新世代的研究典範轉移。

本文基於作者個人的有限研究經驗，並且以資訊技術為主要的角度，嘗試歸納出大數據與資料計算為人文社會研究所帶來的幾個意涵：

1. 資料的原生性（nativeness）

許多社會層面以數位化流程運作後，各種人類活動留下豐富而大量的數位足跡。對人文社會研究而言，資料已是原生、多樣而充沛，政府部門、專業機構、企業、社群平台等不斷累積龐大紀錄資料，許多文件書籍等紙本資料，也持續被回溯數位化建置。因此，人類歷史與現代樣貌某種程度的被投射到一個數位空間，是研究議題不可忽視的巨大存在，也是充滿無限可能的機會之地。

2. 資料的可操作性（operability）

數位資料除了在保存、傳輸、取用等面向上打破了傳統的時空限制之外，其數位本質更是讓各種計算技術有非常寬廣的操作空間。資料可以被切割成顆粒單元，可以在各種維度上重新組合、以各種角度檢視，並可供各種計算方法嘗試達成各種分析目的。以文本資料而言，人類的語言文字轉化為數位資料後，可以被拆解為一個字元，甚或一個音節的資料單位，而從最小單位，到字詞、詞組、句子、段落、篇章、特定主題文稿集或無

所不包的語料庫，都可以被設定為操作對象。各類表單紀錄的結構化資料則是提供了便捷的資料操作框架，讓資料使用更容易擴大。認識並理解各種數位資料操作概念與技術，有助於人文社會研究數位取徑的開發。

3. 資料的可連結性（linkability）

隨著人類資料的大幅成長，各種資料之間也將產生在主題、對象或概念上的連結，而能提供更完整的意義與樣貌。資料之間可能存在顯而易見的關聯，也可能隱藏著微妙的關聯，無論是主觀的認定，或是客觀的檢驗，都可以透過資料的計算分析過程，進行主導性的建立或是探索性的辨識，而取得資訊的連結與整合，協助研究人員建構出更完整的脈絡資訊與知識體系。

4. 資料的可呈現性（representability）

資料的數位性質在各式計算方法的操作下，可以產生高度彈性的資料觀察能力，在各種分析工具的參數設定下，資料的維度與面向可以被輕易的選擇、調控與投射，而可以多視角的呈現資料內容的豐富樣貌。當資料量遠超過一般人有限的資料消化與認知能力時，資料的彈性呈現能力，可以提供宏觀的全貌描繪，快速取得統整性資訊或浮現出顯著資訊，也可以聚焦到特定對象或區塊，進行細微的爬梳檢視。另外，資料視覺化技術更以多元圖像呈現資訊意涵，展現出更有力的資訊傳遞效果。

5. 資料的可檢驗性（inspectability）

資料的數位型態讓資料便於公布流通，可以讓不同研究團隊之間就相同的議題、相同的方法，對各自的實作分析結果，相互或前後驗證；也可以用不同的方法，比較結果的差異，測試方法的優劣；或是探討不同的議題，開發不同的發現；或是累積整合為更大的研究成果。資料就如同實驗室中的材料，可以反覆操作測試，實驗過程與結果都可以被複製檢驗，而提升研究成果的客觀性與正確性。

6. 資料的高價值性 (value)

在大數據時代中，資料的累積速度遠遠超過資料被利用的程度。不斷生產堆積的資料大幅度的捕捉了人類行為與社會運作的真實樣貌，提供了偵測、探究、理解個人與群體的全新途徑。大數據就如同巨大的礦源，蘊含著豐富的資訊與知識，等待鑽探與挖掘。因此，數位資料礦源不論是對學術研究、政府公共治理、知識經濟產業，都是具有高度價值而應積極掌握利用的重要資產。

另外，大數據與人文社會研究也存在一個必須介接的認知與理解。大數據是以資料規模的龐大為最基本的定義，但何謂龐大，從資訊技術角度而言，是存在一個量化門檻，其訂定條件是以造成現有機器設備及軟硬體運作功能上的技術障礙為考量，如作業系統效能、記憶體容量、資料庫存取、資料計算時間等。Gandomi & Haider (2015) 認為一般是以超過 1 tera-bytes (TB)，也就是 1024 giga-bytes (GB) 的資料量為大數據的門檻。若以一般相機或手機拍攝大約為 2 mega-bytes (MB) 的照片為基準，則必須累計超過 524,288 張以上的照片資料量，才進入大數據的技術門檻。Beaver et al. (2010) 的研究指出 Facebook 儲存的照片資料規模已經達到 20 peta-bytes (PB) (20 x 1024 TB)，而且每星期使用者上傳的照片資料量約為 60 TB。因此，所謂大數據資料規模的技術門檻，並非絕對客觀不變，也可能隨著資訊技術的進步與資料的快速成長而提升。

本文認為人文社會科學與資料計算框架的跨領域研究，未必需要硬性套用所謂大數據資料規模的技術門檻。人文社會研究資料通常是以文字為主，大部分的資料規模並不會達到前述機器設備的技術門檻。以本文所介紹的應用研究資料為例，中華民國政府官職資料庫約 90 萬筆異動資料的資料量約為 150 MB，中研院平衡語料庫 4.0 版含有 1100 萬以上詞目文本原始文字檔案 (XML 格式) 約為 200 MB，轉為資料庫格式之後則約為 475 MB，經濟部商業司商工登記資料含有 300 萬餘筆的資料量約為 1.8 GB，皆未達機器技術的大數據門檻。許多人文社會研究議題相關的本地資料集，在規模上幾乎不可能達到技術性上的龐大，但其資料涵蓋面已遠超過傳統人文

社會領域的人工蒐集採樣方法範疇，其資料量也遠超過人腦的認知能力與記憶容量，或可視之為人腦概念大數據，或是相對於過去人文社會資料的大數據，對人文社會研究展開資料計算框架的探索，仍是一個必須先建立的起點，同時，也具有實質的研究價值與潛在的研究貢獻。大數據的概念對人文社會研究更重要的意義在於新研究取徑的啟發，為原已啟動的資料計算跨領域研究帶來更大的能量，也喚起社會更廣泛的想像與重視。

陸、研究限制與討論

本文介紹大數據基本概念與主要分析技術，回溯學術發展趨勢，並搭配本地三個資料計算與人文社會科學跨領域應用研究案例，進而總結研究意涵，並嘗試為人文社會的資料計算研究在大數據情境下，找到合適的定位。相關內容以作者個人之領域知識及研究成果經驗為主，希望為許多人文社會學者及資訊學者在彼此陌生、缺乏互動的隔閡下，建立一條可以相互接近的參考路徑。因此，本文的性質並非相關議題學術領域的全面盤點調查研究，在大數據分析技術的說明上，也有相當的侷限性，僅能涵蓋作者相對熟悉的項目。在應用研究的列舉上，則是聚焦於作者實際操作的部分成果，以政治、新聞與歷史及商管應用的案例，試圖展現各種資料計算在解讀與發現的多元性，並提供資料意涵的對應參照。這些應用案例在性質上大都屬於探索研究，主要呈現資料面與計算方法的可能性，而在人文社會研究議題的設定與尋求解答上，仍然有許多進一步開展的空間。

本文在資訊分析技術或資料計算意涵的論述上，可能引發部分人文社會學者的不同解讀。事實上，資訊學者與人文社會學者在人文研究內涵與資訊技術能耐上，因為學術養成體系的隔閡與學術知識專精的差異，的確存在相互的認知落差。絕大部分的資料分析技術，並無法提供百分之百的效能，但若其效能可以達到某一個基本門檻，則其技術可行性已經開始建立。資訊學者通常會論述此技術可以發揮的目標空間，一邊展開應用實驗，一邊持續改善技術，循環精進。而各種人文議題與資料領域也存在許多獨特困難與障礙，必須個別因應克服。因此，本文所描述的分析技術與資料計算意涵，並不代表其可輕易的直接套用到各種跨領域研究議題上，也不

能視之為問題的有限解答，而是提供了一些可能性與發展性的參照。另外，本文也未對人文社會學者與資訊學者的跨領域研究分工互動有所設定，不論是階段性的清楚切割，或是全程的交互共創，在不同條件與情境下，皆是可行的模式，也存在各種可能。

學海浩瀚，漫無疆界。不論是大數據分析，或是人文社會的資料計算跨領域研究，都是體系分枝龐雜，並且不斷演化的有機體。單一學者、一篇論文大致只能描繪一小部分的面貌，更多更完整的圖像可能有賴主題手冊（subject handbook）的編撰，廣邀各領域專精學者共襄盛舉，才能呈現兼具廣度與深度的研究框架與階段性成果盤點。本文希望能鼓勵人文社會資料計算跨領域研究的興趣，促進討論與互動，開啟更多的跨領域合作嘗試，進而發展出更多相互連結的多元路徑。

柒、跨領域研究展望

在二十一世紀的學術發展脈絡下，學科之間的互動合作與交融協作，已經是一個必須面對與重視的趨勢。跨領域帶來各種創新的機會，當然也存在許多門檻與障礙。各學術領域之間在學術目標、研究方法、專業語言、甚至價值體系等方面，都存在不可忽視的差異，必須在雙方真誠合作，以同理心溝通磨合，相互學習、理解對方的知識技能，才能真正打破彼此的隔閡，進入到平等共創的穩健模式。如同國內重量級學者的敏銳觀察，人文社會科學與資料計算的跨領域結合，代表著一個新研究典範開創的機會，但必須以相關學術社群的共同投入與正確認知為基礎。例如，王汎森（2014）指出，數位和人文的結合，關鍵還是在於建立一種夥伴式的合作關係，研究對象的整體脈絡及內在的意義結構必須由人來解讀，而數位工具可以對資料進行梳理與探勘，幫助釐清研究素材細節、獲取客觀證據。人與數位工具兩者之間的往反與對話，才是最具生產力的工作模式。項潔（2014）認為，數位科技與人文的結合，並非只有將人文引向科學化、實證化一個面相，數位人文並不意味著要和傳統人文研究重疊，如何系統性深化與擴大人文議題的研究，亦是應當要思考的方向。

本文呼應認為人文社會與資料計算跨領域的開創與拓展，需要各相關

領域的偕同並進、互動探索。以資訊科學領域而言，各種數位技術與工具的研究主要是以計算模型與方法上的創新發展為主要訴求，同時也希望對問題領域的應用能開創價值與貢獻。而對人文社會科學領域來說，數位技術與工具可能提供了另一種形式的觀察分析方法，但是資料內涵的議題仍然是其學科知識上所關切或可延伸的範圍，而在大量資料透過數位技術與工具的協助下，能以過去人力所未逮的觀察角度，取得更明確的具體客觀證據，而得到更充分、更完整或是全新的研究結果與發現。以資訊科學領域為出發的跨領域，必須掌握人文社會研究的問題特性與資料計算方法應用的適當性，而取得應用成果的創新；以人文社會研究議題為主體的跨領域，則必須認識理解各種數位工具的功能概念與限制，而藉由全新的人文資料與數位實驗方法，得到研究取徑的開創及領域知識的突破。兩者之間必須深度互動與密切融合，透過實際的跨領域參與及實踐，發展出能結合人文社會探索研究與數位工具的實驗能力，並以回饋－再回饋的反覆修正、持續擴展深化，才能建立重要的跨領域研究典範。

大數據的時代需要人文社會學者與資料計算學者共同攜手，探討新研究範式中資料、議題、實驗設計驗證、理論模型建構的合理組成與循環修正，才能迎接豐碩的研究成果，進而開創新的知識體系。本文呼籲國內學術社群可以共同思考，在可供數位實驗之人文社會資料集、開放性數位工具及數位實驗平台等三個層面的資源，如何共建共享，讓更多有興趣的學者專家，更容易投入於相關研究議題的探討，相互結合並累積研究成果，啟動本質性、結構性、系統性的人文社會研究創新，進而協助產業界創新應用，成為提升國家競爭力的重要基石。

參考文獻

- 王汎森 (2014)。〈數位人文學之可能性及限制：一個歷史學者的觀察〉，項潔 (編)《數位人文研究與技藝》，頁 26-35。臺北：臺大出版中心。
- 王連成、張博城、劉吉軒、甯格致 (2015)。〈應用社會網路分析於企業交叉持股探討〉，《2015 人工智慧技術與應用研討會論文集》，頁 160-165。台南：國立台南大學。
- 陳佩君 (2015)。〈經濟部商業司開放資料現況與未來〉，《政府機關資訊通報》，329：1-10。
- 陳俊宏、劉吉軒 (2015)。〈公司交叉持股與地理脈絡之關聯結構研究〉，《2015 人工智慧技術與應用研討會論文集》，頁 148-153。台南：國立台南大學。
- 項潔、陳麗華 (2014)。〈數位人文：學科對話與融合的新領域〉，項潔 (編)《數位人文研究與技藝》，頁 9-23。臺北：臺大出版中心。
- 劉吉軒 (2012)。〈中華民國政府官職資料庫發展與應用〉，《圖書與資訊學刊》，4 (2)：1-32。
- 劉吉軒、賴隆平 (2011)。〈臺灣文官職等升遷預測〉，項潔 (編)《從保存到創造：開啟數位人文研究》，頁 113-130。臺北：臺大出版中心。
- Aamodt, A., & Nygård, M. (1995). Different roles and mutual dependencies of data, information, and knowledge: An AI perspective on their integration. *Data & Knowledge Engineering*, 16(3), 191-222.
- Allan, J. (Ed.). (2012). *Topic detection and tracking: event-based information organization. The Information Retrieval Series*, vol. 12. New York: Springer Science & Business Media.
- Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, 337(6092), 337-341.
- Beaver, D., Kumar, S., Li, H. C., Sobel, J., & Vajgel, P. (2010). Finding a Needle in Haystack: Facebook's Photo Storage. In *proceedings of the 10th USENIX symposium on operating systems design and implementation*, vol. 10 (pp. 1-8).

- Berry, D. M. (2011). The computational turn: Thinking about the digital humanities. *Culture Machine*, 12, 2.
- Brandes, U., Raab, J., & Wagner, D. (2001). Exploratory network visualization: Simultaneous display of actor status and connections. *Journal of Social Structure*, 2(4).
- Carrington, P. J., Scott, J., & Wasserman, S. (2005). *Models and methods in social network analysis*. New York: Cambridge University Press.
- Chang, L. P., & Chen, K. J. (1995.10). The CKIP part-of-speech tagging system for modern Chinese texts. *Proceedings of the international conference on computer processing of oriented languages* (pp. 172-175). Hawaii, U.S.A.
- Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: A general framework and some examples. *Computer*, 37(4), 50-56.
- Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J. & Nowak, A. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1), 325-346.
- Dalrymple, P. W. (2011). Data, information, knowledge: The emerging field of health informatics. *Bulletin of the American Society for Information Science and Technology*, 37(5), 41-44.
- Davidson, C. N. (2008). Humanities 2.0: Promise, perils, predictions. *Publications of the Modern Language Association of America*, 123(3), 707-717.
- Doan, A., Ramakrishnan, R., & Vaithyanathan, S. (2006). Managing information extraction: State of the art and research directions. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* (pp. 799-800). ACM.
- Doreian, P., & Stokman, F. (Eds.). (2013). *Evolution of social networks*. London: Routledge.
- The Economics. (2010). Data, data everywhere. February 2010.

- Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210-230.
- Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82-89.
- Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. New York: Cambridge University Press.
- Freeman, L. C. (2012). Methods of Social Network Visualization. In Robert A. Meyer (Ed.), *Computational Complexity* (pp. 2981-2998). New York: Springer.
- Frischer, B. (2011). Art and science in the age of digital reproduction: From mimetic representation to interactive virtual reality. *Virtual Archaeology Review*, 2(4), 19-32.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144.
- Getoor, L., & Diehl, C. P. (2005). Link mining: A survey. *ACM SIGKDD Explorations Newsletter*, 7(2), 3-12.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data, *Nature*, 457(7232), 1012-1014.
- Hersh, W. R. (2002). Medical informatics: Improving health care through information. *Journal of the American Medical Association*, 288(16), 1955-1958.
- Hill, S., Provost, F., & Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2), 256-276.
- Huang, G. H., & Chang, N. B. (2003). The perspectives of environmental informatics and systems analysis. *Journal of Environmental Informatics*, 1(1), 1-7.

- Huang, C. R., & Chen, K. J. (1992). A Chinese corpus for linguistics research. In *Proceedings of the 1992 International Conference on Computational Linguistics* (pp.1214-1217). Nantes, France.
- Kadushin, C. (2012). *Understanding social networks: Theories, concepts, and findings*. Oxford: Oxford University Press.
- Kebede, G. (2010). Knowledge management: An information science perspective. *International Journal of Information Management*, 30(5), 416-424.
- Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-data applications in the government sector. *Communications of the ACM*, 57(3), 78-85.
- Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.
- Kling, R. (2007). What is social informatics and why does it matter? *The Information Society*, 23(4), 205-220.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802-5805.
- Li, R. Z. (2004). *Modality in English and Chinese: A typological perspective*. Unpublished doctoral dissertation, University of Antwerp.
- Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48(2), 354-368.
- Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7), 1019-1031.
- Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: For marketing, sales, and customer relationship management*. Indiana: John Wiley & Sons.
- Liu, J. S., & Ning, K. C. (2011). Applying Link Prediction to Ranking Candidates for High-Level Government Post, In *Proceedings of the 2011*

international conference on advances in social networks analysis and mining (ASONAM) (pp. 145-152), Kaohsiung, Taiwan.

Liu, J. S., Ning, K. C., Chuang, W. C. (2013). Discovering and Characterizing Political Elite Cliques with Evolutionary Community Detection. *Social Network Analysis and Mining*, 3(3), 761-783.

Liu, J. S., Ning, K. C., & Lee, C. Y. (2016.07). Evaluating modal use in news corpus for constructing rhetorical context of historical event. In Proceedings of the 2016 Digital Humanities conference abstracts (pp. 262-266). Krakow, Poland.

Lohr, S. (2012.02.12). The age of big data. *New York Times*, 11.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved Feb. 12, 2016, from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Berlin: Springer Science & Business Media.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521-543.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.

Presner, T. (2010). Digital Humanities 2.0: A report on knowledge. Connexions. Retrieved Mar. 27, 2016, from <http://cnx.org/content/m34246/1.6>

Pryke, S. D. (2004). Analysing construction project coalitions: Exploring the application of social network analysis. *Construction Management and Economics*, 22(8), 787-797.

- Ressler, S. (2006). Social network analysis as an approach to combat terrorism: Past, present, and future research. *Homeland Security Affairs*, 2(2), 1-10.
- Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4), 919-931.
- Science. (2011). Special issue: Dealing with data. *Science*, 331(6018), February 2011.
- Schmarzo, B. (2013). *Big data: Understanding how data powers big business*. Indiana: John Wiley & Sons.
- Scott, J. (1988). Social network analysis. *Sociology*, 22(1), 109-127.
- Skyrms, B., & Pemantle, R. (2009). A dynamic model of social network formation. In Gross T. & Sayama H. (Eds.), *Adaptive Networks* (pp. 231-251). Berlin: Springer.
- Sobkowicz, P., Kaschesky, M., & Bouchard, G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, 29(4), 470-479.
- Svensson, P. (2010). The landscape of digital humanities. *Digital Humanities*, 4(1), sn: 1938-4122.
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139-1168.
- Tinati, R., Halford, S., Carr, L., & Pope, C. (2014). Big data: Methodological challenges and approaches for sociological analysis. *Sociology*, 48, 663-681.
- Torra, V. (Ed.) (2013). *Information fusion in data mining, Studies in fuzziness and soft computing series*, vol. 123. Berlin: Springer.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4), 402-418.
- Vergeer, M. (2012). Politics, elections and online campaigning: Past, present and a peek into the future. *New Media & Society*, 15(1), 128-148.

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, 7(3), 203-220.

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, 58(4), 479-493.

Zorich, D.M. (2008). A survey of digital humanities centers in the United States. Washington, DC: *Council on Library and Information Resources*. Retrieved Feb. 18, 2016, from <http://www.clir.org/pubs/reports/pub143/pub143.pdf>.

Big Data Analytics and Interdisciplinary Research Applications in Humanities and Social Sciences

Liu Jyi-Shane

Abstract

In the twenty-first century, human society enters an era of digital connection and digital immersion. Physical world are interfaced and linked with virtual world. Data become competitive resources. Human and machines collaborate in creation. This paper introduces fundamental concepts of big data, data analytic techniques, and the retrospective context of interdisciplinary development between data computing and humanities/social sciences. Three research applications of data oriented interdisciplinary study are presented to showcase the aspects of data content, analytic method, and exploratory discovery. The interdisciplinary research examples include: the R.O.C. government post database and political power, Taiwanese local news archive for the 228 event and rhetoric context of historical event, business registration data from the ministry of economics and corporate interlocking.

This paper further suggests essential implications of big data and data computing for humanities and social sciences research, such as data nativeness, data operability, data linkability, data representability, data inspectability, and data value. Finally, this paper calls for a joint research venture by data scientists, humanists, and social scientists to explore and refine appropriate construction of data, issue, experimental design, validation, and theoretic model in the new research paradigm. This exciting endeavor will lead to promising and fruitful research results, and eventually prosperity of new knowledge discovery.

Keywords: big data analytics, digital humanities, computational social sciences,
interdisciplinary research

