大數據的影響及反思

隨著資訊通訊設備及科技技術的日新月異，相關科學與知識不斷突破更新，21 世紀資訊快速成長，造就了現今資訊大爆炸的時代，資訊量的增加，讓人們在生活中、工作中更加便捷，但是管理這些資訊是相當困難的，如何管理這些巨量資訊變成一個重要的課題，這些數位資訊可能是這個時代的顯微鏡或望遠鏡，可能產生很多重要的，甚至是具有革命性的見解，其衍生的價值無與倫比。但是面對巨量又雜亂的資訊，其中甚至大部分的資訊都是意義不大或者無用的，我們卻又面臨另外一個課題：人的時間和記憶力是有限的，要如何正確又精準的篩選出想要的資訊呢?近年來興起的新興技術，甚至可說是最夯的議題「大數據」，什麼是大數據?本書的作者並沒有對大數據做出精準定義，因為究竟多大才是大，事實上使用大數據往往會縮減本身所用的數據，如果直接使用一大堆數據，想從龐大又雜亂的數據中發現問題的癥結點，就好像亂槍打鳥、大海中找出一根針一樣困難，為了從大數據中取得洞察力，使用正確的資訊，提出正確的問題才是重點。數據分析並不是一項新的技術，1990 年代的資料倉儲就是類似的應用，都是處理和儲存大量的資料，不同的是數據分析是處理鬆散的資訊，資料倉儲則是處理結構化的資訊，然而面對大量鬆散且多元的資訊，使用大數據分析顯然是一個具有效率及效果的，透過分析

正確的資訊，得到正確有用的數據，用來輔助判斷出更好的決策與應用。

本書作者歸納大數據擁有的四種力量：提供新類型的數據、提供誠實的數據、允許我們把焦點放在人口中的小子集、允許我們進行許多因果關係的實驗。第一種力量-提供新類型的數據，透過如何善加利用大數據，並詳細解釋大數據為何如此強大，數據在我們的生活扮演重要的角色，並且數據發揮的作用會越來越大，有效的數據並沒有我們想像的那麼複雜，也許每個人都是數據分析師，例如小時候用哭來吸引媽媽的注意、經常抱怨的人朋友就不想跟你出去，這些都是數據科學，也就是經驗，當無法透過數據分析時，我們透過累積經驗來判斷並決解問題，但是有時候在沒有嚴謹分析的指導下，我們的經驗或直覺可能錯得離譜，因為經驗僅僅累積於自身或周遭發生的事件，樣本數太小且不客觀，如美國職棒大聯盟傳統球探透過主觀的眼光挑選球員，挑選出來的球員往後卻不一定可以打出好的成績，而奧克蘭運動家隊是個市場小、規模不大、球隊薪資在聯盟裡敬陪末座的球隊，無法花大錢簽約明星球員，當時的球隊經理比利·比恩(Billy Beane)為了挽救球隊頹勢的戰績，在各種不利的條件情勢下，採用當時前所未見的創新方法-數據分析，透過數據分析挑選具有潛力而且還未被發現的球員，因為這種數據派的科學分析，讓運動家隊經常可以獨具

慧眼找到被低估的球員，進而幫運動家隊締造亮眼的成績。這個名為「Moneyball」的科學數據化的管理方法，改變了棒球界的選秀和交易策略，甚至被廣泛應用在其他領域上。另一個透過數據分析成功的例子-賽馬，在各種不利的條件之下，馬主人如何挑選一匹可以為自己賺錢的馬呢?傳統的挑選模式是分析馬的血統，挑一匹血統純正優良的馬，雖然馬的血統確實重要，但這只是一匹馬成為明星賽馬的一小部分原因，根據比賽紀錄，血統優良的馬匹獲得冠軍的比率大約只佔四分之一，數據告訴我們僅僅靠血統預測是否能成為優秀的賽馬，還存在很大的改進空間，一家名為 EQB 的公司負責人傑夫·賽德(Jeff Seder）透過評量分析賽馬的各種屬性，並找出哪些屬性跟賽馬場上賽馬的表現有關，最後找出馬的內臟大小會影響其表現，尤其是左心室的大小，透過大數據分析研究，傑夫・賽德（Jeff Seder）挑選一匹名叫 85 號的賽馬，它的左心室的大小是 99.61%，加上其他內臟大小的資料，傑夫・賽德（Jeff Seder）預測它是一匹 10 萬里選一，甚至百萬裡選一的賽馬。18 個月之後，這匹馬成為了 30 年來第一匹得到三連冠的賽馬。第二種力量-提供誠實的數據，數據分析就是完美的有效指標嗎？其實不然，作者提到 1950 年蒐集調查有關丹佛居民投票率的數據，看看調查結果與官方數據是否相符，結果當時居民受訪時的回答與官方公布的投票率差距甚遠，即使這項調查是以匿名

的方式進行，但大多數人還是誇大自己的投票行為，拉到最近 2016 年的美國總統大選，當時民調預估川普支持率低於 2 個百分點，但是最終川普勝選了，為什麼匿名調查時仍會產生錯誤的資訊?因為「誘因」人們沒有誘因跟調查說真話。除非願意跟調查說真話，否則我們沒有其他的資訊來源。但是現在情況改觀了，Google 搜尋，讓人們承認自己在其他地方不願意承認的事情，人們在心裡好奇想要獲得解答的問題會上網搜尋，任何疑難雜症，只要連上網路輸入關鍵字 Google 搜尋即時提供解答，方便又快速，Google 還有一項巨大的優勢，就是讓人們願意說出真話，因為獨自一人、沒有人進行調查，作者針對 Google 搜尋大數據分析揭露許多人們難以啟齒的問題，揭露人性的真面目，例如美國同志問題、種族問題、有關性方面的問題等等，以種族問題為例，美國是一個多元種族組成的國家，但是翻開美國歷史，黑人飽受偏見與歧視，即便美國白人很少會承認自己是種族主義者，但是因為潛在性偏見，所以影響他們對待黑人的方式，美國人搜尋「黑鬼」的次數，竟然跟搜尋「偏頭痛」一樣多，這種明確的種族主義的確對美國黑人產生重大的影響，我們對偏見的刻板印象不只出現在種族歧視，也出現在其他地方，以景氣衰退期間虐童情況為例，許多專家擔心景氣衰退會大量產生虐童案件，後來官方公布數據，虐童案件卻有減少的趨勢，但是當景氣不好，人們陷入壓力與沮喪中，

虐童案件真的減少了嗎?作者透過Google數據進行分析,發現景氣衰退這段期間,幼童搜尋的關鍵字令人揪心,而且這樣的搜尋次數隨著失業率提升而激增,所以我們不能盲目的相信數據,因為我們看到的結果有可能受數據蒐集的方法有缺失而造成假象。第三種力量-允許我們把焦點放在人口中的小子集,作者分析美國職棒大聯盟-紐約大都會隊的球迷,數據研究發現在1962年及1978年出生的這群人中,大都會隊非常受歡迎,原來在那段期間1969年及1986年大都會隊贏得2次世界大賽冠軍,先前提到的那群球迷年紀大約是7、8歲,分析其他球隊的球迷也是類似情形,所以預估男孩是哪一隊球迷的一項重要指標就是,檢視男孩在7、8歲時是哪一隊獲得世界大賽冠軍;政治偏好的起源也是一個研究重點,研究人員發現政治觀點形成的方式其實跟運動團隊偏好形成的方式相同,都有一個關鍵時期,大數據允許我們有意義的放大檢視數據的細部獲取新見解,除了年齡以外,還可以放大其他不同的面向。第四種力量-允許我們進行許多因果關係的實驗,作者分析美國出名人士來自哪裡,當他研究出這些數據時發現,大學城和大城市造就名人的機率實在驚人,為什麼會這樣呢?可能是基因的關係、可能因為很早就接觸到創新;還有另一個變數是個人出生郡的移民比例,一個地區外國出生人口比例越高,日後那裏的孩童有名的比例就越高。研究發現顯示暴力電影可能煽動暴力行為,

但是暴力電影的影響究竟有多大,研究團隊數據分析在暴力電影上映的周末,現實生活中的犯罪率卻明顯下降,這項研究顛覆了我們的認知,第一,想想看誰可能看暴力電影。是年輕人,尤其是逞凶鬥狠的年輕男性,暴力電影上映,要狠好鬥的年輕男性會進電影院看這部片,但如果週末上映的是愛情文藝片或者搞笑喜劇,上述的年輕男性可能不會想去看,所以就會去上酒吧、俱樂部或去打撞球,而這些地方的犯罪發生率就會提高,原來暴力電影讓潛在的暴力者留在電影院而遠離街道,但是當暴力電影散場之後為什麼犯罪率沒有提升呢?研究團隊最後了解到犯罪的主因是酒精,因為美國電影院不賣酒,所以散場之後犯罪率沒有提升。隨機對照實驗是證明因果關係的黃金標準,而在數位世界的實驗更具有龐大的優勢,在數位世界裡隨機對照實驗既省錢又省時,只要透過滑鼠的拖移動作和點擊次數,透過設計程式自動分析,就可以找到因果關係的方法,在大數據的時代,整個世界就是一間實驗室。

2019 年 12 月底新型冠狀病毒在中國武漢市發現首例開始,除了對中國影響甚鉅以外,短短 3 個月已經造成全球大流行,各國政府無不繃緊神經紛紛採取應變措施,緊靠中國的臺灣,在原先不被看好的情況下,但是目前卻能有效控制疫情,除了政府記取 2003 年對抗 SARS 的經驗,各項防疫措施超前部屬以外,善用大數據分析也是政府對抗

新冠病毒的手段之一，疫情爆發初期便面臨口罩缺貨問題，政府與民間合力開發口罩地圖，利用大數據分析藉由取得健保藥局的位置及各藥局的即時存量資料，解決口罩荒有效遏止口罩亂象；確診人數高達 697 人的鑽石公主號郵輪曾有旅客於基隆港下船觀光，政府亦靠相關部會及電信業者的資料連結和大數據分析熱點，回溯出下船旅客們的旅遊足跡，做出警示路線與行徑時間，提醒民眾是否曾經接觸過郵輪旅客；利用健保署建置的全國健保資料庫的資料，結合移民署掌管的出入境紀錄，藉由大數據分析來掌握特定對象(如確診、曾接觸確診民眾)的健康狀況、回溯過去 14 天航班紀錄，讓醫療人員在診療過程中即時掌握特定對象出入境及旅行史。以上都是政府透過大數據結合相關的智慧科技應用提出對抗新型冠狀病毒的因應措施成功案例，「美國醫學會期刊」網站更在 3 月 3 日刊登文章，以「臺灣因應 2019 年冠狀病毒疾病：大數據分析、新科技與積極主動採檢」為標題，詳細介紹臺灣的防疫措施，在疫情仍然持續加溫的情勢下，如何善用大數據結合其他面向，（如人工智慧、醫療、生物、基因科技等），在病毒分析、疫情追蹤、病患篩檢等方面提供各種協助，提升防疫效率及防疫效果是接下來急需面對的一個重要課題。

隨著大數據時代來臨，大數據在我們的生活中扮演著重要的角色，大數據在各行各業中廣泛的存在，如企業分析使用者在搜尋引擎中搜

尋的關鍵字，進而推播相關的商品；企業透過蒐集用戶的購買記錄，分析消費習慣，再適時推出用戶喜好商品的優惠或發送折價卷等促銷吸引消費；各種運動型穿戴式產品現世，可以偵測使用者的身體狀況和健康管理，這些資料是企業在從事產品研發、設計和行銷等活動時的研究分析基礎，然而大數據雖然帶來巨大的好處與便利，但是大數據的管理卻存在挑戰，在大數據時代的浪潮中，每個人都是主動或是被動的數據創造者，但無論主動或被動，在這些便利、智慧的分析下，大数據使用的道德及到個人資料的保護面，亦是個重要的課題。例如新冠肺炎感染者被社會大眾知道後，感染者本身或是其家人否會被貼上標籤感覺被社會遺棄，造成身心的陰影，進而影響未來的人生及規劃，就如同衛生福利部陳時中部長所說的，知道這些數據及資訊的真相後是否會對社會造成的恐慌，是我們必須去慎重思考的。