「數據、謊言與真相」閱讀心得

一、 閱讀緣起

這本書取名為數據謊言真相，作者會取這種書名，即是很明白指出人們會對朋友說謊、會對調查說謊、會對自己說謊。為的是讓外人看自己是完美的假象。因為人都會希望自己體面、正直、善良，即使是在匿名問卷調查中，我們仍可能在「希望自己看起來剛好」的心態作祟下，「不由自主」填出偏離自己本性的答案。這邊是真的「不由自主」，並不是故意說謊。像是作者舉出一些比較敏感的議題上，人們會被「政治正確」驅使，無意識的選擇自己認為比較「恰當」的答案。但可能潛意識裡，自己並不是這樣想。像是難民、伊斯蘭教、同性戀...等議題。還有一些則是比較「難以啟齒」的議題，像是亂倫、性虐待、親密關係，面對這些議題，人們往往會因為害羞或是無法坦然面對，對匿名調查選擇說謊。那我們要怎麼樣才能知道「大家到底怎麼想」的真實答案呢？作者的答案是 Google 搜尋引擎。本書講述大數據的力量如何推翻我們的直覺，以及我們如何透過 Google 搜尋引擎，透露出我們的「喜好」、「不安」以及「政治立場」。書中諸多的實例都令人印象深刻，他告訴我們數據不只是揭發已然發生的狀況，更能夠作為一種預測工具。

二、 作者介紹

賽斯・史蒂芬斯—大衛德維茲 Seth Stephens-Davidowitz。《紐約時報》撰稿人暨華頓商學院客座講師，曾為 Google 數據科學家。史丹佛大學哲學系畢，哈佛大學優等生榮譽學會（Phi Beta Kappa）成員暨經濟學博士，目前定居紐約市。史蒂芬斯—大衛德維茲的研究使用新的大數據來源，揭露人們潛藏的行為和態度，並已刊登在《公共經濟學期刊》(Journal of Public Economics）等聲望卓著的出版物。

三、 本書介紹

本書作者深入研究 Google、推特(Twitter)、臉書(Facebook)、警察局紀錄、電影票收據、維基百科、色情網站、棒球球員個人成績表，和你想像不到的數位與傳統資料來源之後，發現這些資料來源有個共通的特點：他們提供的是大數據，亦即我們每個人每秒鐘無意識的反射，而非根據民意調查而來的一小部分民眾的意見樣本。

四、 心得分享

（一）大數據的魔力

大數據是近期社會上討論度極高的話題，不論公部門或私部門都不斷在強調，最早提出「大數據時代已經到來」的機構是全球知名諮詢公司麥肯錫。2011 年，麥肯錫在題為《海量數據，創新、競爭和提高生成率的下一個新領域》的研究報告中指出，數據已經滲透到每

一個行業和業務職能領域，逐漸成為重要的生產因素；而人們對於海量數據的運用將預示著新一波生產率增長和消費者盈餘浪潮的到來。大數據是一個不斷演變的概念，目前大部份的機構將大數據的特性歸類為「3V」要大，包括資料量（Volume）、資料類型（Variety）與資料傳輸速度（Velocity）。到目前為止，大數據的 3V 到底要多大或多即時，並沒有明確的共識或定義。近年來大數據的定義又從最早的 3V 變成了 4V ——第四個 V 代表 Veracity ，意指資料真實性 Veracity ，討論的問題包括：資料收集的時候是不是有資料造假、即使是真實資料，是否能夠準確的紀錄、資料中有沒有異常值、有異常值的話該怎麼處理… 等等。本書即是針對此點在探討數據真實性，作者認為我們可以透過躲在網路搜尋的真實發言，獲得最正確的數據。

（二）環境影響了我們的思想

作者從他自身與弟弟對於棒球的喜好有極大差別，明明長相相近基因相似，但一個極度喜歡棒球一個很厭惡棒球，因而提出很有趣的觀點，究竟我們興趣養成與甚麼有關，利用臉書的資料按讚人數去做假定，歸納分析後得出，男孩8歲時如果該球隊獲得世界大賽的冠軍，男孩極有可能成為該球隊的球迷並且終其一生支持該支球隊。同樣的假設也可以放在政治上，人們在 18 歲時該政黨在社會上的評價也會

影響人們支持政黨的傾向。

作者認為得出的假設若要得到驗證，必須有一筆很大的數據，如果數據不夠大或者是不夠精準，那麼獲得的結果就不盡然是正確。幸好我們目前的社會，網路世界的無遠弗屆及虛擬空間讓我們能得到越來越多的數據，電腦也能精準去計算，有時候數據科學家也可以變身成為人類學家。作者甚至利用研究維基百科的數據獲得你能成為名人的機率，發現大城市及大學城能夠造就名人的機率十分驚人，但名人不一定等於晉升中產階級，反而父母來自新移民的新生兒更有機會成為名人，這個論點真的十分新奇及特別。

（三）整個城市都是實驗室

相關性及因果分析是我們做數據分析必須被解決的一個大問題，甚麼是相關性呢?例如適當飲酒有益身體健康這是相關性，但要如何判斷是否適當飲酒就會改善人們的身體狀況，這就是因果關係的證明，但我們要如何建立因果關係呢?目前人類最常使用的方法是隨機對照實驗，目前這在社會科學各個領域中都是被普遍適用。然而 20 年前的 2000 年 2 月 17 日，GOOGLE 利用這個因果關係的隨機對照實驗改變了網路世界的創舉，將用戶分成兩組，一組看到 20 個連結另一組看到 10 個連結，比較兩者滿意度，看似沒創新，但這卻大大啟發了GOOGLE，因為他們發現了網路世界的優勢，線上隨機實驗比起實體隨

機實驗所需資源及少，在數位世界裡，隨機對照省錢又省時，作者告訴我們這就是大數據擁有的第四種力量:大數據讓隨機對照實驗變得更容易進行，在大數據時代，整個世界就是一間實驗室。

在 20 年後的現在，大數據讓我們更便利使用隨機實驗，因此也有相關的新興行業因運而生，進行 A/B 測試就是這家公司的主要業務，在 2012 年美國總統大選，雙方候選人都有聘請該家公司協助進行隨機實驗測試，也令人十分好奇在大選競爭都很激烈的台灣，是否有類似的作法?

（四）大數據與人類的真實面

還記得 2016 年美國總統大選，川普勝選讓全世界都傻眼的那刻嗎?沒有一個投票專家認定川普可以勝選，因為很多選民認為川普冒犯了他們，但作者卻說網路上有很多線索顯示，川普可能會贏。作者運用了這讓人跌破眼鏡的大選結果引人入勝去了解為什麼這一次的大選所有資料都遇預測不成真呢?

美國前任總統歐巴馬的勝選是美國的一個跨世紀的指標，因為他是美國第一位非裔美籍總統候選人，距離美國廢除總族隔離政策不到 50 年，即使歐巴馬只有一半非裔血統，但畢竟他的身分就並非純正白人血統，所有人都認為美國切實達成所謂種族大拼盤的平等國家。然而作者在此時發現了 GOOGLE 趨勢，發現美國人在問卷調查上或許

會說謊，但在 GOOGLE 搜尋上絕對不會說謊，他驚訝發現人們一而再再而三利用 GOOGLE 搜尋有關貶低黑人的話語以及帶有強烈種族主義的情緒性字眼，原來這才是我們所認知的世界背後的真相，這也是傳統在數據蒐集上絕對得不出的結果。

作者甚至利用這些搜尋結果獲得很多與我們原本認知都不一樣的地方，例如我們原本以為恐怖攻擊會引發大眾普遍焦慮與空荒，但作者檢是歐美地區發生重大恐怖攻擊後的幾天，與焦慮有關的搜尋是否增加，發現居然沒有，原來這個世界跟我們想像的不太一樣啊。作者也提出別讓直覺蒙蔽我們獲得正確的數據分析，他認為如果我們只是仰賴我們所聽到的事情或只仰賴個人經驗時，我們對於世界的看法往往會有問題，雖然良好數據科學的方法論是直觀的但結果卻往往違反直覺。

（五）政策執行度落實檢視

公家機關擔負國家時常推行許多政策，但到底實際上落實與否我們時常無從得知，但如果透過 GOOGLE 搜尋其實我們間接了解到民眾是否真的有收到相關資訊，例如監理所這幾年所推行的汽燃費約定扣繳，我利用 GOOGLE 搜尋關鍵字，發現以關鍵字「汽燃費 約定扣款」所作搜尋討論度並不高，反而是「汽燃費 信用卡」是出現的熱門選項，這部分其實我們可以分別作許多討論，一部分是監理所花許多人

力及經費所宣導的汽燃費約定扣款政策並大部份人民並不知道，是否是我們宣導政策與民眾接觸的管道有所落差，或者是必須思考透過另一個方式宣導。第二是汽燃費約定扣款政策並不被民眾所接受認同，這個政策是否值得繼續推行下去，這是值得檢討的地方。第三是我們由主動化成被動，由民眾的需求做為政策推行的出發點，也就是說我們可以藉此分析民眾透過信用卡進行汽燃費繳納的比例是否有逐年增加趨勢，而採取其餘繳納的宣導方式。

（六）網軍真的存在嗎

網軍議題一直是最近社會彼此爭倫的話題，網路上有許多論壇，台灣常見的有 PTT、Mobile，另一方面網路的普及也造就許多社交媒體，例如臉書、Instagram、微博，這些都是網路上一般民眾可以發表言論的地方，也可以看到別人發表的言論。那麼這時就有可能出現所謂帶風向的情形發生，也就是說除了你可以發表言論，但同時也會看到別人發表的言論，就像是早期媒體一樣，觀眾的想法有可能會被媒體影響，然而現在更簡單了，每個人都有可能成為影響其他人的腳色，這也是全新的網路時代造就所謂網軍存在的契機。

最近武漢肺炎消息滿天飛，而中國資訊戰更是趁機開打，從肺炎之初的口罩政策到近期的確診人數爭議，我們都可以很清楚看到有中國網軍滲入的情形，他們想要透過假訊息影響台灣的輿論，這些假訊

息的內容多是利用一些模組套版製造不實訊息，收集台灣圖片素材進行變造，以社群平台為集散地廣發假訊息。更研究兩岸的語法差異跟輿論鋪陳散佈策略，提升假消息辨認難度，這些假消息也將會成為大樹據統計上的黑數。

（七）網路數據缺點

除了上述假消息會造成網路數據的不正確性外，利用網路數據做為統計樣本亦有可能有其他方面的缺失，其中之一就是不可忽略數位落差所造成的影響，所謂數位落差是指因資訊通訊科技取用與否，在個人或群體之間所產生的斷裂缺口，早期調查的項目著重於家戶是否使用電話、電腦與網路等設備，而最為關注的則是學校因經費多寡形成擁有電腦與網路的設備差距議題。早期多半從「擁有」與「沒有」電腦、網路使用等接近資訊設備的機會與運用能力之有無，來區別數位落差存在的現象，然時代變遷，資訊設備與觀念條件隨之更新，數位落差衍生新的定義與更多的討論面向。國內外研究發現，數位落差的現象在性別、年齡、教育程度、都市化程度、種族、職業、收入、區域、國別之間等因素皆存在著程度不一的狀況；衡量指標從早期的電腦及網路擁有率、普及率，到中期的上網率及現今的應用能力，每種衡量指標與研究結果都有其針對性，探討此類議題時應視問題的情境與切入的視角，才能掌握問題核心。

五、 結語

本書帶給我們很不一樣的大數據觀點，藉由網路上搜尋軌跡得到比問卷更加正確的數據，這些都是值得公部門服務的我們可以效法並且提出相關政策建議的地方。