《數據、謊言與真相：Google 資料分析師用大數據揭露人們的真面目》閱讀心得

在選擇閱讀此書前，在了解內容是以何種主題研究所撰寫而成前，吸引我閱讀的其中一項原因是他的原文標題-《Everybody Lies》。美國知名電視劇《怪醫豪斯》(House, M.D.）中的主角豪斯醫生，最令人記憶深刻的名言便是” Everybody Lies”他認為所有患者不管何種原因入院，對醫生的問診絕對不會據實以告，病患總會掩蓋事實、隱瞞真相，但不起眼的生活或飲食習慣都會影響豪斯醫生對疾病的判斷。而這也是本書作者-賽斯．史蒂芬斯—大衛德維茲，透過數據分析來告訴世人，所見所聞並不代表真相，因為只要是人都會說謊。

究竟我們為什麼需要數據？數據對現代社會是否重要？作者在擔任 Google 數據學家期間，藉由探索許多爭議性議題、透過各種網路實驗，分析人類經由網路、紙本、訪談等其他來源數據，蒐集而成令大眾意想不到的真相，就像第一段提到的豪斯醫生名言:Everybody Lies。人們對醫生說謊、對初次見面的陌生人說謊、填寫問卷時會說謊，這些都是間接或直接影響最終結果的變數，但人類對網路、對 Google 搜尋欄可不這麼想，大多數人會透過 Google 搜尋自己難以啟齒或不願承認的事情，而這些發自人類內心真實的解答正是作者想要

的數據，讓人一窺人類內心世界。作者於前言以一個最讓人匪夷所思的事件作為開場白，關於美國總統-唐納·川普的當選，他無疑是位成功的商人、企業家，任誰都想不到這位言論極具虛假、充滿爭議性、且帶有種族色彩，更直白說法所謂”政治不正確”的總統候選人，能在沒有擔任過任何軍職或公職的情況下於選戰中脫穎而出，成為第四十五任美國總統。事實真是如此嗎？大數據可不這麼認為，作者透過數據分析告訴我們，川普可能贏得大選的蛛絲馬跡。一開始會懷疑，我們真的能依賴大數據嗎？而作者在書中也清楚表明，他相信Google搜尋是有史以來針對人類心靈所能蒐集到的最重要數據集，大選前的各種民調、專家觀察都表示川普沒有機會勝出，而顯然有太多人對民調撒謊，這些不表態或聲稱還沒決定的選民也許內心早有適任人選，那些表明會出門投票的選民或許最後並沒有出門投票，以上變數都讓支持度、預估投票率出現偏差，但透過 Google 搜尋趨勢就會發現極大的相關資訊，作者從許多層面研究選舉相關主題，不管是搜尋喜歡或討厭的某位候選人頻率、對特定或少數族群的敵意、種族主義的搜尋率甚至是黑鬼總統這樣仇恨意味濃厚的用詞等，都是作者用來預測川普可能會勝選的線索，而這些都是民調機構無法發現的資訊。

對美國總統大選研究只是簡單的開場，Google 數據集也不會是藉由網路理解人類行為的唯一工具，因為還有更多的資料庫與搜尋引

擎甚至網路最大的色情網站之一，都是提供作者深入了解人類內心的完整數據。

那麼，究竟要蒐集多少數據、獲得多大的量，才會是一個好數據？鍵入的關鍵字質量是否影響最後研究結果？作者在本書進入關鍵分析前先替讀者打了預防針：我不會對大數據做出一個精準定義。因為大數據本身就是個籠統的概念，我們可用的資訊在數量與質量上每天都是海量、爆炸性的產生，Google 與其他社群媒體每天都產生更多新資訊、新數據，本書重點以過去舊有市場、研究為基礎，尋找他們缺乏的龐大數據集，人類每天平均產生數十、數百萬兆位元組的數據，該如何善用現今新穎的方式處理並分類，才是此書要討論的核心。

前言至此，會覺得作者拋出的看法有那麼點難以理解，身在數位時代的我們，每天在世界各地產生的新聞量透過網路快速傳遞，每個人都能輕鬆取得最新資訊或轉發消息，若不依賴這些數據集帶來的便利性又能如何憑直覺斷定一份調查需要多少樣本數，才能排除存在這之中的異常數值？而這也是 Google 帶來最有趣地方，我們靠著搜尋引擎解決生活中疑難雜症，我們靠它解決食衣住行等方面的需求，更進階點，有許多研究人員、企業主、甚至政府單位都仰賴他或其他知識庫帶來的便利性，每一次的搜尋足跡都會是另一組人的研究數據，透過鍵盤我們會自動吐露不願被人知道的私密問題，以各種方式表達

自己的心聲或渴望，甚至不必承擔後果，這些發自人類內心最真實的回饋正是作者追求的，史蒂芬斯—大衛德維茲用輕鬆詼諧的方式，透過多組大數據比對，解開我們過去對某些議題的誤解，重新認識這個世界的真相。

但是怎樣的數據才算是大數據呢？作者已經將話講明，他不會將大數據做出定義，但他透過研究與文字逐步論證大數據的力量，首先是「提供新類型的數據」，美國的金融業者靠著每月失業率來觀察股市，每個金融公司都想搶在第一時間獲得由勞工統計局藉著電話調查取得的資訊，因為只要眨眼間，這項統計就能影響廣大市場，然而這種計算方式不僅費時費工，也消耗龐大資金，金融業是分秒必爭的行業，如何運用大數據替失業率抓出粗略的結果，便能使這家公司取得市場最大先機。那麼 Google 是怎麼站穩搜尋引擎的龍頭？作者先用網路搜尋關鍵字做介紹，如何靠單字與網站內容之關聯性運算出最合適連結？Google 創辦人針對不同主題的所有意見做彙整，蒐集成千上萬新聞、部落客、網民等的意見，做出與這個單字最有關連性的搜尋結果，比起計算詞彙出現次數更有價值，而這些連結也是過去許多搜尋引擎沒有考慮到的數據，Google 不是單憑比其他搜尋引擎蒐集更多數據，而成為搜尋引擎的龍頭，而是藉由找到更好的數據類型才

能脫穎而出。接著作者用較輕鬆的比喻來解釋 Google 運作方式：如何成為賽馬明星？

看似無俚頭且不相關的標題，卻讓人能快速理解 Google 蒐集、分類數據的關鍵理由，誰能知道一位非傳統路線訓練出的養馬人，靠的不是觀察馬匹血統，而是評量賽馬各種屬性與他在場上的表現有關，這位養馬人的策略便是以數據科學為基礎，多年反覆尋找成為明星賽馬的關鍵因素，就像 Google 蒐集更好的數據集一樣，我們必須保持開放的態度並懂得變通，考慮使用非傳統數據來源，這些新類型數據將帶來更大回報。

大數據的第二種論證「提供誠實的數據」，考量到每個人都會說謊，作者運用人類對 Google 搜尋都據實以告的特性調查同志比例、仇恨言論、性隱私等問題，許多人在做各類問卷調查被問及令自己困窘的提問時，總會選擇違背事實的答案，即便這些問卷是匿名，人類還是想維持自己社會上形象，也因此讓問卷調查出現一定比例上的誤差值，好比前言講到的投票率就是很好的對照，官方統計與數據調查呈現強烈反差，這也成為民調結果無法預測最終是川普勝出的原因之一。

人類針對敏感話題總會選擇避而不答或說謊逃避，因為沒有一個「誘因」促使人們與調查說實話，但面對搜尋引擎卻能讓人輕易承認

自己不願承認的事實，沒有任何誘因讓人接受調查時說出自己遭受病痛所苦，但確實有誘因以搜尋引擎了解自己出現的症狀及可行的治療辦法。作者在章節裡針對不同議題做出更多分析，他利用另一個大數據來源「臉書」，來計算同志人口的分佈比例，透過人們對自己臉書資訊的編輯，將公開感興趣性別的用戶按照區域編碼分類，就能看出贊成與不贊成同志的人口差異，越支持同志的州，居住的同志人口相對占多數，反之亦然，但這其中有的變數是，居住在不贊成同志州內的同志，便不會透露自己真正性向，或像是高中生，很少能選擇自己要居住在哪裡，也成為影響這項數據的因素，正因如此我們才不能僅依靠一項大數據就斷定同志人口的數量或分布地區。還有像是針對不同種族的偏見，大部分人們不會直白表達對特定族群的看法，卻會在 Google 搜尋將穆斯林與恐怖分子畫上等號，當這種刻板印象深植人心時，透過數據變化都能觀察到這些煽動仇恨言論是如何爆發的，潛藏在人心的歧視、對不同族群潛在性的偏見，我們口頭上否認，但卻明確的影響 Google 數據變化。

接著，大數據允許我們進行許多「因果關係的實驗」，大數據讓隨機對照實驗找到真正因果關係的方法，作者於本章節介紹由 Google 工程師進行的「A/B 測試」，他們僅需藉著網路資源在線上將用戶隨機分組，就能透過回訪次數統整出他們想要的答案，看上去沒什麼創

新的手法，但對比在現實世界做測試所要耗費的大量資源要省時省力的多，且獲得的數據同樣令人信服，而且他還能避開人類說謊的可能性，只要使用者在線上，幾乎隨時隨地都能進行，整個網路世界就是大型實驗室。作者接著解析美國前總統歐巴馬競選團隊，怎麼在競選活動中善用 A/B 測試，讓更多人在造訪候選人網頁的同時，能採取更實際的行動。這些測試內容包含更改首頁照片、按鈕文字的變化，團隊測試多組搭配並回收用戶點擊按鈕頻率，找出最適合首頁招呼訪客的組合，這種測試不僅省錢又輕鬆完成，也因為我們對人性的了解還不足，透過測試也能填補差異，從錯誤中不斷修正與學習。

史蒂芬斯—大衛德維茲透過這些年收集而來的數據，用精闢且幽默的方式解釋大數據的能與不能，解釋他令人著迷的地方，但在文章最後仍用些許反例提醒過分迷戀大數據，會讓我們忽略更重要的考量，當人們越陷越深就越會遭數字誘惑。透過臉書我們可以知道什麼文章被按讚、被分享，但研究臉書數據的科學家卻得不到使用者真正的使用體驗，他不會知道兩位使用者之間是否有增加現實世界的連結，因此要使大數據發揮最大效用，都會使用小數據來填補漏洞，且人類數千年來發展出的傳統方式也無法被大數據去除掉，兩者必須是相輔相成。他也提到數據可能會面對的道德問題，假設我們在預防任何犯罪狀況發生前，就依靠數據分析哪些地區的仇恨言論搜尋比例增高，更

甚至縮小範圍到可能犯罪的個人，但這不僅完全侵犯到個人隱私，且數據只能證明一個可怕關鍵字被搜尋，不代表這些搜尋就會發生可怕行徑。

閱讀過這些章節，我也自行上Google搜尋趨勢瀏覽臺灣的紀錄，第一時間肯定會覺得，新冠狀病毒、武漢肺炎、coronavirus，會是臺灣搜尋熱度最高的關鍵字，當然答案是肯定的。儘管這三個月來，他一直是搜尋熱度前幾名，但隨臺灣掌控疫情的速度，搜尋熱度逐週遞減，從一月中至二月份的搜尋高峰，到三月份逐漸下降，近期數據與其他國家相比搜尋率反而降低許多，但這也僅代表單一關鍵字所能看出的數據，且這並不代表臺灣人對新冠狀病毒已經不在乎。

最後的結論，作者將整本書做簡潔有力的統整，社會科學正在成為一門真正的科學，而這門貨真價實的新科學將有助於改善我們的生活，在書內他討論這些創新數據集，並抨擊各領域專家學者，忽視數位時代造成的數據爆炸，他們眼睜睜看著網路垂手可獲取的海量資源，仍舊埋頭於傳統的測試方法，在這片廣大領域只有具前瞻性思想的人們願意接受如此創新的思維。他不像生硬的、需要不斷反覆閱讀的艱深書籍，透過作者專業描述替社會科學開啟新的道路，搭配輕鬆易懂的實證舉例，看到最後一頁我想每個人都能會心一笑。