「大數據」閱讀心得

沈翰祖

壹、 前言

從大學時期學習統計開始，一直被灌輸如何藉由取樣的技巧，推測母族群的樣貌，而且受限於電腦系統的運算速度、資料的儲存空間與資料的取得方式等，常無法取得大量數據並加以分析。然而近年由於資訊軟硬體發展迅速，上述的限制似乎漸漸不再成為資料分析的門檻，「大數據」這個新興學門也跟著蓬勃發展。

過去對「大數據」的認識是非常片段的，由於我從事的工作是生物技術研發，取樣也成為研究上的限制因子之一，此次看到天下文化出版的「大數據」一書，激起我藉由此書一窺其廟堂奧秘之心，期能對我的工作與世事的處理，有不同角度的認識。

貳、 心得分享

一、 認識「大數據」

大數據(Big Data)就是指巨量資料，亦即資料量一定要達到相當規模才能做的事，否則就無法實現，而且這些事將會改變現有市場、組織、公民與政府間的關係。

以網路巨擘谷歌 (Google)所掌握巨量搜尋資料為例，在 H1N1 躍上新聞頭條的幾星期前，旗下的幾位工程師在著名的《自然》科學期刊發表了一篇論文，解釋了谷哥能如何利用網路搜尋與流感相關的關鍵字，「預測」美國在冬天即將爆發流感，甚至還能精準定位到是哪些州。谷歌的系統真正做的，是要針對搜尋字眼的搜尋頻路，找出和流感傳播的時間、地區，有沒有統計上的相關性(correlation)。他們總共用上了高達四億五千萬種不同的數學模型，測試各種搜尋字眼，在與疾管局在 2007 年與 2008 年的實際流感病例加以比較。能一樣掌握流感疫情，而且不是數周後才完成預測，而是幾近即時同步掌握。谷歌並不需要去到處採樣、也不用到各醫院診所調查或要求他們通報，只是好好分析巨量資料。

利用巨量資料建立一個預測模型，並無須探究「為何如此」(why)，只須知道「正是如此」(what)，預測系統要有效，就必須擁有大量的數據資料。這些資料不再需要先整理成整齊的行列或資料表，因為各式各樣的資料來源，也充滿了各式各樣甚或是雜亂的資料內容，要如何從這些資料中尋找出脈絡，並解讀出其中的意義，就是「大數據時代」所要追尋的。

二、 樣本與母族群的關係改變

在大數據的時代，必須有三種思維要改變，第一，要具有針對特定主題分析龐大且全體資料的能力，而非退而求其次，分析較小的資料集；第二，願意接受真實的資料會是雜亂不清的事實，而不是一味追求精確；第三，要更看重相關性，而非不斷探索難以捉摸的因果關係。

我們現在所處的「數位宇宙」正在不斷擴大，從科學到醫療保健、從銀行到網路，涵蓋的行業各式各樣，持續的進行「資料畫」，也就是將所有類比形式的資料(像紙張、書籍、相片等)轉成可量化的數位資料格式，而且資料量也正在迅速增長，不僅超過機器所能處理的量，更遠超過我們的想像，例如，臉書在全球的使用者，每天約會按「讚」或留言超過 30 億次；谷歌每天要處理 24PB 的資料(1EB=10^3PB =10^6TB= 10^9GB= 10^{12}MB)；在 2007 年估算全世界儲存了 300EB 的資料，其中約只有 7%為類比形式的資料，可以想像嗎？在 7 年之前的 2000 年，全球的資料只有 25%是以數位方式儲存，其餘還都是類比方式。

早期要面對如此巨量的資料，發展出了統計學，希望能從少量的資料中，取得最豐富的結果，於是從各種規範、流程，甚至獎勵措施中，盡量減少資訊量。理想的統計是取得測量完整母族群的資料，但如果其規模太大，便不大可能完成，導致大多數時候只能利用抽樣，蒐集全部資料的一小部分，而且又只篩選這些資料中的一小部分，以便於分析，這是不得不的自我設限，此時「樣本＜母族群」；統計學家為了提高抽樣的準確度，最好的方法不是增加樣品數，而是進行隨機抽樣，如此能降低蒐集資料的成本，又能以推斷方式準確的描繪母族群。就本質而言，隨機抽樣是將巨量資料縮減為較易處理的資料管理問題，但由於無法蒐集完整的資料加以分析，而且最後推估結果的精確度又取決於抽樣時是否達到真正的隨機，要真正的隨機更是難上加難。

現今隨著各種資料蒐集裝置與方法的快速發展，例如各種感應器、手機、網路等，都能有形或無形的進行資料蒐集，而電腦處理資料的速度也一日千里，因此有許多領域已經逐步放棄抽樣，而蒐集完整的資料，也就是「樣本=母族群」。不進行抽樣而利用全面資料也有其限制的條件，包括需要有足夠的資料處理與儲存能力、要有頂尖的分析工具、需要有簡易且可負擔的資料蒐集方式。谷歌的流感預測就是使用了最完整的資料，並未使用隨機抽樣，因此可簡單的說，是否為巨量資料的判斷標準，就在於不使用隨機抽樣。

巨量資料的「巨量」不是絕對的、而是相對的，指的是要蒐集完整的資料。完整的資料或全部的資料也並不一定就是龐大的資料，而是「樣本=母族群」。

三、　巨量資料常伴隨著雜亂

在以前小量資料的世界裡，為了要減少資料錯誤、確保資料品質，需要不斷改善工具，讓量測結果更為精確，以減少誤差；但在巨量資料的世界裡，容忍各種不精確(也就是雜亂)、放寬容許的誤差值，常能得到更多的資料，在此處資料愈多常比品質愈好更為重要；重點是讓資料從「精確」走向「可能性」。

雜亂包括三種，第一，資料點愈多，發生錯誤的可能性就愈高，例如谷歌的流感預測是分析所有網路關鍵字搜尋資料，其中不乏是有人對流感好奇或誤按的情況；第二，為了結合不同的資料源頭與類型，資料彼此不一定相容，導致發生雜亂；第三，資料格式不一，需要先整理過才能使用。要處理雜亂的巨量資料，需要更高速的電腦處理能力，所幸電腦處理能力這幾年依循「摩爾定律」(Moor's Law)發展，就是單一晶片上的晶體總數與處理效能，約 18 個月會增加一倍。

四、　不再凡事探究因果關係

巨量資料的時代，挑戰著我們的生活方式以及邏輯觀念，其中最重要的是拋下對因果關係的執著，轉而注重相關性，也就是 A 現象與 B 現象有相關性，不表示 A、B 現象是互為因果，許多事情都不用知道「為何如此」，只要知道「正是如此」就可以了，推翻了過去數世紀的思維。

相關性的核心概念，就在於將兩個資料值之間的統計關係加以量化，二者間的相關性強，代表著如果其中一個值有所變化，另一個值就極有可能更著改變。在谷歌流感預測就看到相關性，就是在某個特定區域，有愈多人搜尋某些詞彙，就發現有愈多人得到流感。有相關性並不代表兩者間有確定性，只是有可能性，因此，相關性並不能真的預知未來，只能說有一定的可能。

在巨量資料出現前，相關分析就早已十分重要，當時的統計學家挑選相關性指標時，通常以抽象的理論作為參考，形成假說，再依假說蒐集資料並進行相關分析，以確認指標是否適當。然而，挑選指標、建立假說時，都有可能會參雜自己的偏見。但在現今有了如此多的資料與強大運算能力的電腦，就不用再精心挑出少數幾個指標，再一一驗證。在谷歌流感預測中，電腦就嘗試了近 5 億個數學模式後的成果。

人類在解釋和理解世界時，總是有一種尋求因果關係的直覺作為，就算是根本沒有甚麼原因，我們還是會假設出原因，A 現象與 B 現象有相關性，是可以用數學方式加以證明，而 A、B 現象是互為因果，並無法以簡單的數學方式進行計算，必須藉由原因分析、假說建立、實驗證明等方式，再由控制組與對照組的結果進行比對，以確認是否有因果關係。而相關性分析也可以協助找出真正有因果關係的資料，讓實驗設計更為精準，並降低因果分析之成本。

五、　掌握資料就掌握價值

巨量資料的核心價值在於預測，不在於教電腦如何像人一樣思考，而是計算大量資料後推算機率。一個有效率的預測系統，最重要的基礎就是要掌握大量資料，其次是必須能隨著時間自動改進，從新增的資料中，判斷出最佳的預測模式。例如，亞馬遜(Amazon)已經能推薦讀者最想看的書、谷歌排序出最相關的網站、臉書找出我們可能認是哪些人，目前陸續應用到診斷疾病、建議療法、預防犯罪等領域。

在數位的進展之下，已經能更有效的管理資訊，將大量的類比資料轉換成電腦可讀取與分析的格式，於是也降低儲存與處理的成本，但由於分析資料的人還無法完全擺脫類比資料處理的思維，總覺得資料只有某種特定的用途才有其價值，或是努力找出資料的因果關係，於是在處理資料時就延續著這種偏見，我們必須秉除這種舊思維，就能像尋寶一樣，在巨量資料中找出潛在的寶藏。

所以資料的價值已不再是只能應用在某種目的上，而資料也成為一種可交易的重要商品，誰掌握資料，誰就掌握了價值。一般物質性的東西一旦開始使用，其價值通常就會逐漸降低；但資料卻不然，可以一次又一次的選擇相同或不同的

目的來重複利用，價值不會因而減少，也就是經濟學所謂的「非競爭性」(non-rivalrous)商品，某個人的使用並不會妨礙其他人的使用。資料能發揮的真正價值，會遠大於原始的使用價值，有如漂浮在海上的冰山，第一眼看到的只有一小部分，有很大一塊都藏在海面之下，資料的使用者若能找出隱藏的價值，就能獲取巨大的利益。

資料的價值，就是選擇要如何利用的各種項目所產生的價值之總和，稱之為資料的「選項價值」(option value)，要釋放出資料的選項價值，有三種重要方式，包括要重複使用資料、合併資料集、讓資料買一送一。其中「重複使用資料」以關鍵字搜尋為例，在搜尋的當下似乎已完成使用者的目的，其留下的資訊似乎只是一大堆關鍵字與網址等內容，乍看之下一文不值，其實不然，例如可以從搜尋流量來了解消費者的喜好、從關鍵字作為即時經濟指標的預測等，均可重複使用。至於「合併資料集」，就是說到想要釋放潛在的資料價值，就必須與其他資料結合，甚至和截然不同的資料結合而達到創新，例如結合某地區民眾的醫療紀錄、商業行為、人口統計等資料，來預測某些疾病的發生或某些商品的潛在商機等。「讓資料買一送一」就是在設計要如何取得資料時，就規劃資料的多種用途，例如谷歌的街景車在街道繞一圈除了取得地圖街道實景影像與 GPS 資料外，其實其資料也規劃作為該公司研發自動駕駛汽車之用。

大多數的資料會隨著時間而失去部分價值與效用，而且舊資料常會導致新資料的價值變低，因此需要不斷的檢視資料，並剔除已經失去價值者。

六、　巨量資料的使用要省思

巨量資料靠著演算法，我們試著預設各種事情的可能性，像是預測是否會罹患心臟病（於是保更多險）、貸款是否會變成呆帳（因而拒絕放款）、是否會有人犯罪（或許就能先發制人）。這會引發一項倫理問題：自由意志和資料獨裁究竟孰輕孰重？如果使用巨量資料，從統計學的角度得知應做的裁定，但卻與個人意願相違背，那該如何抉擇？

在不久的將來，資料能夠提出許多預測結果，但我們卻不一定都能解釋背後的因果關係。如果醫生使用巨量資料來做診斷，就像是要問別人別問原因、直接相信某個黑盒子就對了，這會有什麼影響？現在的司法系統看待嫌疑犯，判斷標準是「相當理由」，但未來可能會改成「犯罪機率」，這對人的自由和尊嚴，又有什麼影響？是大家要一同留意並省思的課題。

七、　應用與發想

我們在規劃並執行試驗研究時，由於受到人力、物力的限制，不得不進行隨機抽樣，但因著「大數據」的概念，從新思考曾經做過以及目前進行的研究，是否能有機會達到「樣本=母族群」，即或不然，也盡量使樣本接近母族群。

例如在進行植物(作物)品種判別時，除了看它的外觀外，尚可利用其 DNA 的特異性進行判別，若由種子階段萃取其 DNA 進行分析，可免除需要先播種後再由生長之植株進行外觀辨識，以縮短檢測時效，但如此之成本非常高昂，故免不了要進行隨機抽樣。因此，是否有可能利用自動化儀器、透種子外觀的特性，

建立作物品種的判別模式？在此又有一衍生的問題，種子在儀器中可能有不同的角度、位置、方向等，是否又造成了某些困難？其實不然，若以大數據的概念來看這些問題，分析 10 粒、100 粒種子可能產生非常大的變數，但若設計一種儀器它是以類似輸送帶的方式，將每粒種子分別進行資料的紀錄，分析 1 萬粒、10 萬粒.....甚或更多，除了記錄其外觀特性，同時也記錄種子在儀器中的角度、位置、方向、重量等資料，運用數學運算找出之間的相關性，進而建立該品種的判別模式。如此可應用於辨識大量種子樣品中具有哪些我們曾建立判別模式之品種，甚至可加以自動分類，減少人工篩檢的支出。

八、 結論

巨量資料使分析資訊的方式產生三大改變：第一，能夠取得、分析的資料量大幅增加。過去因為資訊不足而發展出來的抽樣方法，漸被分析完整母族群所取代。第二，我們面對極大量的資料時，就不會堅持要求一切都要做到精準。減少了抽樣誤差，所以能接受較多的量測誤差。第三，放下長久以來對因果關係的堅持。簡言之就是有時候不用追根究柢、找出真正的原因，只要能做出更好的選擇、得到改變，就已足夠。

雖然人類科技發展的速度一日千里，讓我們能夠測試的更快、探索的更多，但能夠蒐集和處理的，永遠只是世界上一小部分的資訊，更不可能得到真正完美的資訊，所以資料分析與應用也就必然會出錯，所以巨量資料永遠只是工具，需要重視的仍然是人的創造力、直覺、知識的爆發力等，使用這項工具時，我們必須懷有更謙卑的心，畢竟人類的聰明才智才是世界進步的泉源。

參、 閱讀書籍資料 (資料來源：天下文化)

一、 書名：大數據 (BIG DATA)

二、 出版商： 天下文化

三、 作者：

麥爾荀伯格 (Viktor Mayer-Schönberger)：

牛津大學網路研究所教授，並擔任微軟、世界經濟論壇等大公司和組織的顧問，是大數據（巨量資料）領域公認的權威，寫過八本書以及上百篇專論。

庫基耶 (Kenneth Cukier)：

《經濟學人》雜誌資料編輯，巨量資料思潮評論員，經常於《紐約時報》、《金融時報》、以及《外交事務》期刊發表財經文章。

四、 譯者：林俊宏

師大譯研所畢業。現為自由譯者，並就讀於師大譯研所博士班。譯作《建築為何重要》曾獲 2013 開卷翻譯書獎。另譯有《英語的祕密家譜》、《大數據》、《漫畫費曼》等等。