【解牛集】— 刊於《信報》，2019 年 5 月 28 日

大數據分析得不出因果性答案

黃昊

香港科大商學院會計學系副教授

在商業、經濟和其他領域使用數據分析，據此協助作出決策，做法並非新穎。過去，數據的搜集和處理能力受到一定制約，如難以搜集到數以萬亿計的數據信息，即使收集到，也缺乏有效的處理能力，遑論對巨量數據進行有效分析。今日，由於技術進步，傳統軟件難以在短時內處理的巨大數據量，如今可以輕易解決。在能夠便捷地發佈、搜集、分析數據的基礎上；加上全球大多數政府對高透明度的要求，使「大數據」（Big Data）成為目前個人和組織用來協助作出決策的「工具」。

利用大數據協助分析與決策，其實由來以久，譬如，在中國古代，發生了戰爭，兩軍對壘，出謀獻策的軍師可以通過觀察、統計對方軍隊駐紮地挖了有多少個煮飯的灶，去判斷敵軍的編隊人數。過去十年，尤其互聯網面世後，數據產生的速度愈來愈快，產生的量也日益龐大；加上數據儲存的成本比過往便宜得多，電腦運算的速度飛躍，並且處理和分析數據的算法——如分布式技术，機器學習（Machine Learning）也有很大進步，使大數據分析的應用愈來愈普及。

美總統大選結果出人意表

不過，大數據的分析亦有一定局限性，在一些地方也會「無能為力」，原因關乎數據的質量。究竟數據來自什麼地方？在採集過程中，有沒有出現誤差或缺漏，甚或帶有傾向性。以 2016 年美國總統大選為例，當時，據調查機構採集到巨量數據所顯示的分析結果，大多數人都相信，希拉莉可勝選。筆者記得，到

大選日晚上 10 時前，大家仍認定，希拉莉可以入主白宮，不過到凌晨二時左右，隨著點票陸續有結果，大家才「如夢初醒」，知道白宮新主人是特朗普而非希拉莉。

為何調眾多調查機構收集到那麼巨量數據，但分析的答案，竟然大錯特錯，令人費解，關鍵顯然在於所採集的數據本身帶有傾向性，譬如，有些受訪對象沒有把心目中的投票對象說出來，因而所以呈現出來的答案也變得失實。值得強調的是，數據採集過程中，所收集的數據並非隨機抽選出來。過去的調查，往往通過隨機抽樣方式，以電話「人對人」的方式進行。但如今很多都是透過互聯網收集，以非隨機的方式進行。一些帶有傾向性的人或組織樂於以這種方式參与和提供訊息，結果便可能形成帶有傾向性的民意。一旦數據帶有傾向性，分析結果難免與現實不相符。

對於帶有傾向性數據所導致的錯誤結果，有學者曾進行研究，發現由於數據傾向性的「偏見」（bias），即使採集到 230 萬條相關數據，還不如一個通過隨機抽樣，樣本數目只有 400 的調查來得有效。換言之，雖然數據量十分龐大，但非隨機性抽樣，使數據本身的「偏見」更容易在調查中呈現出來，得出了「誤導性」結果。

另有研究指出，通過大數據協助作出的決策，這個決策不定更好或更準確，反而只會增強對決策結果的信心。

賽馬評磅員實驗的啟示

1974 年，著名心理學家保羅・斯洛維奇（Paul Slovic）和 2002 年諾貝爾經濟學獎得主丹尼爾・卡尼曼（Daniel Kahneman）作了一項關於信息決策的「實驗」。二人召集了 8 位賽馬專業評磅員（Horse handicappers），告訴他們，想知道誰人能夠預測到賽事中那一匹馬跑勝出，表現最佳。「實驗」之目的，在於探究信息如何影響決策。

這 8 名賽馬評磅員專門評估馬匹的勝算，並初步定出賠率，在比賽中實在是關

鍵性人物。斯洛維奇教授告訴他們，對連續四輪包括 40 場賽事進行頭馬預測。在第一輪中，斯洛維奇教授給予評磅員每一匹馬 5 條信息，譬如，馬匹的騎師有多少年賽馬經驗等；第二輪，給了每名評磅員 10 條信息；第三輪 20 條；第四輪 40 條。實驗結果發現，評磅員預測的準確率，並沒有隨著信息量增加而上升，反而是增加對自己預測結果的信心。在這種情況下，評磅員相信，取得更多信息有助作出更好決策，故不斷強化自己對決策的信心。在賭博賽事中，若自己對預測結果有強烈的勝出信心，自然會下重注，後果當然會輸更多錢。研究結果不僅有趣，而且更顯示出，信息與決策的準確性沒有必然關係。

大數據應用範圍有局限

看深一層，大數據還有一個很值得一說的缺漏，就是數據本身只能夠告訴我們數據產生過程中的規律。亦即是說，只能告訴我們在數據本身範圍內的規律。這是什麼意思呢？假設，數據是人面識別的數據，若數據本身只含有白人的人面識別，一旦用來識別白人以外的人種，譬如黑人或黃種人，識別的準確性便大大降低。

幾年前，有學者研究，谷歌和亞馬遜應用的人面識別算法技術，識別黑人的準確性特別差勁。即使算法本不帶有傾向性，但由於識別算法的數據輸入，若只包含白人，在識別黑人時，算法的準確性便大打折扣。

目前，大數據很多應用到醫療保健上，譬如基因的數據，以位於美國加州山景城的 23andMe 基因技術公司為例，該公司逾百份之九十的基因數據都是採自白人，所以，這個數據算法可以對白人應用有效，卻不能拿到中國來直接應用。關鍵在於數據的「質」而非「量」。

相關性與因果性不容混淆

另一個同樣值得注意的大數據問題，是關乎「相關性」與「因果」問題。很多

大數據的分析結果其實都是相關性，而非因果關係。

舉一個例子，母乳喂哺可增加嬰兒的智商。醫學研究發現，母乳含有很多有益成份，但母乳喂乳是否增加嬰兒的智商，透過採集數據得出來的研究樣本發現，母乳喂哺可增加嬰兒的智商，這個研究結果只能說具有相關性，指出母乳喂哺與嬰兒的智商相關，而非因果關係，亦即母乳喂哺嬰兒，結果導致嬰兒智商增加。扼要言之，相關性是非必然，而因果性則具有必然性後果。我們利用大數據協助決策，所希望的，是可以得到因果關係的肯定性答案，可惜大數據简单分析得出的答案結果，只能是相關性。

再舉例來說，我們在媒體上看到一些醫療研究報告，有研究分析結果說，每天吃雞蛋好，但又有研究結果說，每天吃雞蛋不是那麼好。為什麼有這些區別？大部分醫療研究報告都是通過觀察到的數據得出的，而不是實驗。所以只能說明相關性，而非因果。如果平日吃雞蛋的人和不吃雞蛋的人有其他不同，比如其他飲食習慣，運動習慣，那麼得出的結果就不完全是吃雞蛋導致的。

事實上，大部分的數據只能夠告訴我們結果的相關性，而不能告訴我們答案的因果性。當然，有某些情況，我們取得答案的相關性便足夠了。譬如，在行營市場，利用大數據分析，結果發現，買Ａ貨品的人，很多會同時購買Ｂ貨品，於是營銷公司便可以在銷售Ａ貨品的網頁上，同時插入Ｂ貨品的銷售廣告，甚至決策考慮把Ａ貨品和Ｂ貨品打包銷售。利用這個大數據分析結果相關性所作的決策，基本沒有問題。

然而，譬如，當Ａ貨品的銷量下跌，應不應該及時把Ａ貨品的價格下調？很顯然，大數據分析的結果答案，要求具有因果性——指出銷量下跌，便需下調價格的結論。由於大數據只能告訴我們答案的相關性，若急不及待降價，決策便可能大錯。這一點，是我們應用大數據分析時，要明白到其分析的局限性。

個人隱私易受侵犯

最後，值得一提，相較幾年前，如今大數據產生，以及數據的採集，在我們日常生活中幾乎無孔而不入，使個人隱私很容易受到潛在侵犯。從商業角度說，若商家濫用取得的個人資料，可能引起顧客的反感。譬如，聽說有些 APP 裝置

了「竊聽」的程式碼，當取得下載者貨品喜好的信息，便會馬上送上相關貨品推廣優惠的信息，做法令人吃驚和反感。

究竟數據採集者搜集了我個人多少私生活數據？這些數據拿來作什麼用途？個人幾乎一無所知。譬如，我們下載一個應用程式，應用程式的研發或經營者會以合約形式，來說明所採集數據的用途，該公司和下載者的權利和義務，但有多少人會巨細無遺閱讀以至明白合約的內容？可以說，在今日「大數據時代」，個人的隱私保障是一個很嚴肅而且很重要的問題。囿於篇幅，這個問題筆者有機會再論。

〔本文由科大商學院傳訊部筆錄，黃昊教授口述及整理定稿〕