- **余創豪** chonghoyu@gmail.com

毋庸置疑，現在我們已經進入了機械學習與大數據的時代，其實，兩者原本是各自發展的科技趨勢，機械學習是基於人工智能的電腦系統，而顧名思義，大數據是基於大量資料的分析方法，但現在這兩種趨勢已經合二為一，因為一部機器需要龐大數據，才可以不斷學習、不斷更新、不斷完善自己，從而得到更加精準的結果。

然而，任何嶄新的事物都會惹來批評與懷疑，針對機械學習的批評聲音主要是擔心有朝一日人工智能會超越了人類智能，結果威脅到人類。而針對大數據的批評，主要是關於收集大數據會侵犯了人們的私隱，而且大數據演算法會出現誤判，加深了偏見、歧視、壓迫，從而令社會更加不平等。可是，如果細心觀察，這兩種批評是有點自相矛盾的，因為前者顧慮到機械學習越來越厲害，越來越精準，而後者卻認為大數據演算法並不準確。在這篇短文中，筆者將會討論第二點。

有關第二類批評的書籍已經為數不少，例如奧尼爾（ Cathy O'Neil）的〈大數據的傲慢與偏見〉（ Weapons of Math Destruction）、尤班克斯（ Virginia Eubanks）的〈自動化不平等：高科技工具如何分析、監管和懲罰窮人〉（ Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor）。這類作者舉出了不少例子，例如許多美國法院採用 COMPAS 演算法，去推測犯人在將來重操故業的可能性，從而決定犯人能否獲得假釋，但研究發現該演算法歧視黑人罪犯。此外，印第安納州在三年內拒絕了一百萬份醫療保健、食品券、現金福利的申請，因為電腦系統將任何錯誤輸入解釋為「拒絕合作」。

不過，若果細心查看，這些電腦系統和演算法的失敗並不是由於機械學習或者大數據， 剛剛相反，這是由於出問題的系統過分倚賴單一或者少數的因素，這類錯誤是古典統計學的問題：一竹篙打一船人。

我沒有可能在這有限的篇幅完整地介紹古典統計學的歷史，我只能夠舉出一個例子：奎特萊（ Adolphe Quetelet），奎特萊是橫跨十八、十九世紀的比利時天文學家、統計學家、社會學家，他曾經撰寫了一部經典巨著，名為〈社會物理學的文章〉（ Essays on Social Physics），在研究天文學的時候，奎特萊發現不同的科學家觀察同一天文現象時會得到不同的數據，他將這種天文物理學的現象應用到社會，他指出：在社會中不同的人在同一行為上會有不同的表現，於是乎形成一個高低不平的分布線，這就好像中國俗語所描述的「 十隻手指有長短」。當他由不同的數值計算出一個平均值，便可以發展出「 平均人」（ Average man） 這個概念，奎特萊認為「 平均人」 可以代表整個群體，例如法國人的平均身高是五呎八吋，這就是典型的法國人，套用現在的術語，這就是「刻板形象」（ Stereotype）。 我相信讀者已經可以猜測到這種思維的負面結果：往往人們只是憑着單一因素和某個群體的單一數字去判斷個人，結果造成了一竹篙打一船人。

以下筆者會舉出一個自己經歷的例子：筆者曾經在亞里桑拿州居住，有一次我計劃作越野旅行，所以需要一部四輪驅動的運動休閒車，我在網上向租車公司辦好了預訂手續，然後前往取車。可是，租車公司的櫃檯員工卻拒絕向我提供服務，他說：「本公司不會租賃車輛予持有亞里桑拿州駕駛執照的顧客。」 我追問原因，他解釋：「 過去很多持有亞里桑拿州駕駛執照的人將我們的運動休閒車駛去墨西哥，從此一去不回頭。」 我跟他爭論了很久，結果仍然不得要領。這就是全然倚賴單一因素和單一數字而發生的問題，試想像，若果大多數將車輛偷到墨西哥的竊賊是墨西哥人，那公司宣布拒絕租車予所有墨西哥人，這就是「 種族定性」（ Racial profiling），就是一竹篙打一船人！雖然亞里桑拿州居民並不是一個種族，但這種措施和「 種族定性」 沒有分別。其實，若果這租車公司安裝了機械學習和大數據系統，便可以基於多種因素去計算顧客是否屬於劉備般的「高危人士」，是否有可能出現「劉備借荊州，一借莫回頭」的情況。

在某個層面上，機械學習、大數據跟古典統計學背道而馳，古典統計學忽略了個體差異， 其結論重視對人口總體的概括（generalization to the population）；相反，機械學習和

大數據在檢視了整體數據之後，會精準地推測個人的行為，從而提供度身訂做的服務，例如個人化的治療（Personalized medicine ）、 智能輔導系統（Intelligence tutoring system ），還有，網飛的推薦系統（Netflix’s recommender ） 可以根據你的個人喜好而建議你應該看什麼電影。

我相信很多人都不會同意筆者的結論：我認為，要消除電腦系統和數據所造成的偏見、歧視、壓迫、誤判， 就是要加速發展尊重個體差異的機械學習與大數據，而不是仍然停留在「 平均人」、一竹篙打一船人的古典統計學。我希望下一次自己到租車公司出示駕駛執照和信用卡的時候，櫃檯服務員會馬上說：「本公司的演算法認為你是優質顧客，我會向你提供特價優惠，而且燃油費、保險費全免，還會向你贈送紅燒牛肉麵。」

2023 年 5 月 27 日

〈機械學習與大數據加深了歧視和壓迫嗎？〉

原載於澳洲《同路人》雜誌

更多資訊