rthk.hk 香港電台網站
傳媒透視
MEDIA DIGEST
http://rthk.hk/mediadigest

# 「大數據」分析局限 乃傳統統計學問題

「大數據」已成為當今炙手可熱的資訊科技，商務、醫療、社交、教育、政務等領域紛紛廣泛採用「大數據」技術去深入分析相關的網絡資訊，提升應用系統的智能及效率。舉例而言，「智慧城市」是《2017 年施政報告》的重點發展之一，香港資訊科技總監楊德斌指出，多個國際先進城市如紐約和巴塞隆納的「智慧城市」規劃都不約而同地圍繞六大範疇而設，包括「智慧交通」、「智慧生活」、「智慧市民」、「智慧政府」、「智慧環境」及「智慧經濟」。他進一步解釋這些範疇如何利用「大數據」科技去改善市民的生活，刺激經濟；由此觀之，「大數據」可說是「智慧城市」計劃的「主菜」。


## 「大數據」分析之潛在問題

「大數據」的廣泛應用源於美國。自從美國總統奧巴馬 2012 年 3 月推出 2 億美元的「大數據研究及發展計劃」後，世界各大小經濟體陸續仿效，大力投資相關領域。全球資訊科技企業亦不敢怠慢，積極推出適合的大數據資訊科技方案及產品，更大灑金錢推廣大數據分析的優點及其所能帶來的商機。據觀察，近期不少從事金融、醫療、社會工作、工商業、政務等範疇主管都已被潛移默化，鼓吹「大數據」的功能及效益。然而，「大數據」真的是萬能的嗎？本文引用不同的國際專家報告，反映「大數據」分析之潛在問題。

首篇報告題為《谷歌流感的比喻:大數據分析的陷阱》("The Parable of Google Flu : Traps of Big Data Analytics")，描述了谷歌公司曾利用「大數據」分析推算 2011/2012 年度美國流感的趨勢，但結果卻強差人意，估計的流感個案數目遠超過實際數目。而谷歌利用的數據是來自用戶使用的關鍵詞(如「禽流感」)次數及分布作推算分析。專家認為構成嚴重誤差的主要原因是谷歌盲目地廣泛收集關鍵詞，以為越多越好，卻沒有了解用戶查詢時的出發點，結果收集得的數據大部分來自非流感病患者，因此在數據採集階段已嚴重犯錯，自然推算失準。若數據分析全力集中在流感病患者，結果便會截然不同。

第二位專家是美國加州大學柏克萊分校的國際知名學者米高佐敦(Michael

Jordon)教授，他最近接受美國 IEEE 學會雜誌訪問，在題為"Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts" 一文中指出，「大數據」在現今商業市場被過分炒作，它最後可能只是一場空歡喜，教授更預測「大數據」的「冬天」即將來臨。他認為「大數據」用戶作出假設的速度將會超越大數據的統計範疇，在這情況下數據分析結果難免會出現錯誤，造成大量噪音，影響推算的可靠性。

從另一角度看，「大數據」用戶往往忽略數據的「動力」(dynamics)。例如在變幻無常的商務環境中，用戶的需求不停在變，那麼昨天的「大數據」分析結果能有效地應用於今天的商務環境嗎？能夠滿足用戶今天的需求嗎？若然不能，我們需要重新進行分析，但昨天採集商務數據的方法能滿足用戶今天的新需求嗎？歸根究底，什麼時候開始分析及什麼時候停止既是統計學應用的老問題，亦是「大數據」分析必須嚴肅面對的問題，但在千變萬化的應用及數據環境下，要應對這個問題更是難上加難。因此佐敦教授進一步指出「大數據」分析服務提供者有責任清楚說明分析推算法的質量標準及其誤差度，做好用戶的「期望管理」(Expectation Management)。

## 「大數據」的十大局限

「前車可鑑」，因此用戶在使用「大數據」技術時不容掉以輕心，必須緊慎考慮它在操作上的「盲點」(局限性)。歸納而言，這些「盲點」大致是由於以下網絡數據的不健康特性而產生：

- 噪音性：網上數據泛濫，資訊內容五花八門，格式也參差不一。要從中過濾與應用需求無關的數據，既複雜亦耗時。

- 真實性：由於網絡資訊自由，即使在找出相關數據之後，內容的真假亦難以分別。例如去年在美國總統大選期間，在網絡媒體上謠言滿天飛，虛假新聞層出不窮，滲透全美每一角落；“教宗贊助特朗普”、“希拉里向伊斯蘭國（IS）販賣軍火”等假新聞在《臉書》上的分享及點評率遠比傳統紙媒為高。然而，「垃圾入，垃圾出」(Garbage In Garbage Out)，基於偽造資訊的「大數據」分析，難免會適得其反。

- 代表性：真實的數據並不一定具代表性。若然系統錯誤地使用了缺乏代表性的資料作分析的話，結果便會弄巧反拙。

- 完整性：利用非完整的數據進行分析，結果以偏概全，不盡不實，容易引致誤判。

- 時效性：某類數據在事件發生當刻可能大派用場，但當事件或時限過後，其影響力未必復再。若然過量的舊數據被用作分析，結果未能反映現況。再者，適時

的數據往往因為比舊數據少而很容易被忽略。

- 解釋性：在「大數據」的分析過程中，基於輸入的數據，算法便會產生及輸出分析結果。在分析過程中，數據輸入如何產生輸出的理據及兩者的因果關係並不清晰，如黑箱作業。

- 預測性：世事變幻莫測，以前從未發生過的意外絕不罕見，但卻難以預料（分析出來）。因此，有專家認為「大數據」分析是規範的（prescriptive）而不具預測性（predictive）的功能。

- 誤導性：使用假資訊或錯誤分析算法均會影響结果的可靠性。「盡信書則不如無書」，未經核實及驗證的分析結果可能會造成嚴重的反效果。

- 合法性：數據內容、採集方法及其使用過程極有可能涉及個人私隱、商業機密及公眾權益等資訊。因此，資訊的安全性和合法性對「大數據」應用十分之關鍵，可是不少企業只顧賺錢，而罔顧這些因素。

- 價值性：「大數據」不是免費的，企業切忌盲目跟風。數據本身、分析軟件等均所費不菲，因此成本效益的衡量是企業採用「大數據」的另一關鍵考慮點。

**推廣「大數據」缺乏人才**

有數學專家指「大數據」實際上只是統計數據系統的升級版而已，上述的「盲點」問題，在統計學上早已解決過，而方案亦適用於今天「大數據」應用。筆者不完全同意這觀點。從統計技術而言，這觀點確實無可口非。首先，應用這技術在「大數據」範疇時，用家必須徹底關注上述統計學的基本問題，避免出錯。再者，在應用層面，「大數據」比純統計複雜，是一門跨領域、應用主導的工程課題。要有效地使用「大數據」技術很需要領域知識及經驗的適當配合，絕對不能單靠統計分析理論。

所以，推廣「大數據」的樽頸在於缺乏人才，在現今職場中熟悉統計學或個別應用學問的大有人在，但要聘請精通兩者而具多元知識的人才絕不容易。正因如此，香港各大學近年相繼推出數據科學(Data Science)課程，培養高質素的「大數據」精英。

**黃錦輝**
香港中文大學工程學院副院長(外務) 及 創新科技中心主任

3.2017