【解牛集】— 刊於《信報》，2019 年 8 月 5 日

大數據時代個人隱私形同透明

黃昊

香港科大商學院會計學系副教授

今天，大數據的商業應用愈來愈廣泛，筆者曾在本欄撰文，指出大數據分析有局限性，分析結果往往只能顯示出相關性而非因果關係。另一方面，大數據還有一個更值得注意的地方，就是關於個人的隱私（Privacy）問題。

大數據應用在日常生活中如今無孔不入，但一般人顯然沒有覺察到，自身不想外人知道的隱私，其實有很多已暴露在別人眼底，被人「一覽無遺」，甚至拿來應用到自己不知道的地方。可怕吧！何以如此？

要言之，大數據可以讓人知道很多沒有「說出來」或「明示出來」的個人信息，這也是商家十分喜歡應用大數據的原因之一。譬如，行銷商可以利用大數據有效瞄準潛在目標顧客、或者有效掌握不同顧客對不同產品的需要；透過顧客在網上選購行為的「記錄」、性別、年齡、家庭狀況等個人資訊，商家可從中制訂生產什麼？銷售什麼、向誰銷售等有效的商業營利效策。

Cookie 內裡乾坤

很顯然，消費者在網上進行活動或購物，並沒有預期到或同意網站營運商「肆意」取用個人的資料，即使跟網站營運商有「取用個人資料」的「同意協議」，但相信大部分人都不完全清楚冗長的「同意協議」內容、自己胡裡胡塗給出了什麼樣的同意內容。

舉一個例子，當瀏覽人登入一個網站，這個網站為了辨別用戶身份而儲存在用戶終端（Client Side）、亦即存放在瀏覽者電腦上的 cookie，一般人都會同意，甚至不假思索便接受了這個 cookie。對此，大家都不會進一步去想，此舉有什麼後果？

其實，放置了 cookie 的網站營運商可以透過這個 cookie，知道瀏覽者繼後會去登錄什麼網站，瀏覽了什麼信息……，這些信息，可以讓營運商對該名瀏覽者整合出一個完整的「個人資料檔」（Personal Profile）。譬如，這個人接著經常登入那些專賣兒童服飾的網站，於是，就可以推斷其人有養育子女；或者經常登錄一些專科醫療網站，據此也就推斷此人健康有什麼毛病。可以看到，營運商掌握其人

的個人隱私資料，比任何人都多，瀏覽人發夢也沒有想到，自己如此多的個人隱私信息，幾乎在別人眼中完全「透明化」，赤裸裸一覽無遺。

「基因檢測」並非玩意

再舉一個例子。目前，不少人都喜做「基因檢測・Genetic Test」（需要寄口水樣本），此舉往往帶點「貪玩」心理去進行檢測，看看自己有沒有肥胖的體質、能不能喝酒、自己是南方人的體質抑或屬北方人體質等等。檢測前，填好了問卷，簽署了同意聲明。但同意些什麼，往往沒有深究。然而，進行基因測檢的公司並不會將測檢視為一項「玩意」，反之，基因測檢公司往往把這些收集得來的基因資料信息，將之分析，並跟製藥公司合作，去探索或發現有什麼基因的人，很有可能出現某種疾病。攜手去開發一些新藥物。

最近，美國著名基因公司 23andMe 便和製藥商 Glaxo 合作，去開發一種醫治腦部退化的柏金遜病症（Parkinson's disease，簡稱 PD）藥物。透過分析了 800 萬人的基因，研究有什麼基因的人更容易患上這柏金遜症。當掌握了相關信息並據此研發醫治的新藥物，按照美國食物及藥物管理局（FDA）的法律規定，推出新藥前，必須經過臨床測試和審批。研發商如何找出臨床試驗對象？顯然可以從 23andMe 的基因資料庫中，找到對口的測試者。由於有這種基因的人數量不多，1000 人中充其量只有 1 人。然而，從 23andMe 信息庫中，便找到 7000 人，不僅「輕而易舉」，且大大節約了製藥商可觀的搜尋成本。

看深一層，進行檢測的人會想到，自己的資料會成就這個商業行為結果嗎？究竟這些個人資料還作什麼用途？若資料交給了保險公司，保險公司清楚知道投保人的基因組合信息，日後可能引發某種嚴重疾病，此人的保費必然會提高，這個結果，進行基因檢測的人會想到嗎？

「匿名人」身份無所遁形

事實上，人們在網站上對自身隱私信息給出同意的聲明，但同意給出什麼內容或範疇，往往無法知道，也難以掌握，這樣一來，個人的隱私問題，在大數據時代下，便顯得紕漏百出，毫無保障。

在網站上留下或填報個人資料，網站營運商往往標榜瀏覽者為「匿名」，據此來讓大家「安心」，以為個人身份的隱私可以得到保護。但「匿名」的保障有效嗎？答案是：在大數據下，「匿名」是不管用的。

因為只要有一定的信息，便可以利用機器學習（Machine Learning）方法，將之整

合成為一個完整的「個人身份檔」。

日前，美國《紐約時報》（The New York Times）報導和引述了《Nature Communication》上一項科學研究，這項研究由英國倫敦帝國學院（Imperial College London）聯同比利時天主教魯汶大學（Université catholique de Louvain）的科學家，他們利用計算機算法，只需個人 15 個屬性（attributes）資料，便可以在「匿名」的資料庫中，識別出 99.98%美國人的身份。

15 條信息揭開廬山面目

這篇有趣的研究論文（Estimating the success of re-identifications on incomplete datasets using generative models）揭示，那些可以取得到的 15 個信息資料，包括 3 個基本屬性資料——個人的出生日期、性別和美國郵區編號（Zip code）；加上另外 12 個其他屬性資料，如種族、婚姻狀況、就讀學校、按揭記錄、職業等等，便可透過算法，將這些在「匿名」資料庫中的資料，還原成一個完整的個人身份。亦即是說，透過電腦計算法模型，毋須知道其姓名，可以識別出 99.98%美國人的身份，研究清楚展示出，「匿名」對個人身份的保障並不管用。「匿名人」的身份，在模型的分析和資料重組下無所循形。

以區塊鏈上的比特幣（Bitcoin）為例，箇中交易的信息雖然也是「匿名」，沒有明示交易人的身份，但理論上可以通過區塊鏈上的公開信息，譬如交易區域、交易時段、風險偏好選擇、甚至交易金額等等信息，還原出投資人是誰的身份。

可以說，在大數據時代，個人的隱私可以說毫無保障，被別人一覽無遺。譬如，你到某個落腳地點，該處設有 Wi-Fi，你又在該處用手機上網，在進入 Wi-Fi 時，須要同意一些進入條款。另一方面，由於每部手機都有一個獨特的網絡地址，根據你的手機網址，Wi-Fi 商便可以洞悉手持手機的人之前是否來過、來過的次數；也知道你手機登錄過什麼網站。換言之，透過 Wi-Fi 上顯示的手機網址，便可以知道你到了來。

隱私外洩到處是陷阱

即使你不使用該處的 Wi-Fi，若這個地方設有人臉識別（Face recognition）鏡頭，通過鏡頭，也知道你已來訪。如今一些公共地方都設有視像鏡頭，並公開說明有設置，不過，無論你同意與否，從 Wi-Fi 上得到關於你的信息，之後如何應用，對你來說，簡直是一個謎。

很顯然，匿名並未能好好保障個人身份不予外洩，因為手機的獨特網址，已「無

言地」把個人的身份顯露出來。順帶一說，無論在中外的示威衝突場面，我們看到一些示威者「幪面」，以為這樣做，其身份或廬山真面目便沒有人知道。理論上，對於違法行為的法律責任追究，假若政府能夠在監控法例准許下，取得在示威衝突場上，當時及當地出現過的手機網址，只要從電訊商處取得相關資料，基本上可追查到誰人——無論是否「幪面」，有份參與示威衝擊。若現場有人臉識別鏡頭，取得佐証，「幪面示威人」的身份更無所遁形。

正視個人隱私保護問題

日前，美國第五大信用卡發行商 Capital One Financial Corp 的客戶資料遭黑客入侵，約 1.06 億信用卡客戶的資訊被外洩。這名嫌疑女黑客是華盛頓州一名軟件工程師。據調查人員透露，她突破一個防範不嚴的防火牆，從而入侵了數據。值得注意的是，這些數據是存儲在亞馬遜公司(Amazon.com Inc., AMZN)的雲端上，而此駭客原是亞馬遜的職員，事件令人懷疑雲端儲存資料的安全性。

總括來說，大數據應用時代，個人隱私往往形同透明。月前，美國三藩市國會通過《反監控條例》，禁止當地政府及警察使用人臉辨識技術的城市。雖然人臉辨識技術在保安及監控工作上很有效。但美國一直有人權組織質疑，技術可能被濫用，三藩市作為美國的科技重鎮，成為首個禁用人臉辨識技術的城市，實在別具意義。

照目前情況看，作為網絡的消費者，個人資料不僅很難得到充份保障，而且自身也不自覺地成為商家的「產品」，如上文提及的基因檢測，個人的基因資料遭商家「產品化」，自身作為「產品」也沒有得到一分一毫的報酬，究竟在大數據時代，如何保障個人隱私免遭潛在的濫用和侵犯，問題必須正視。