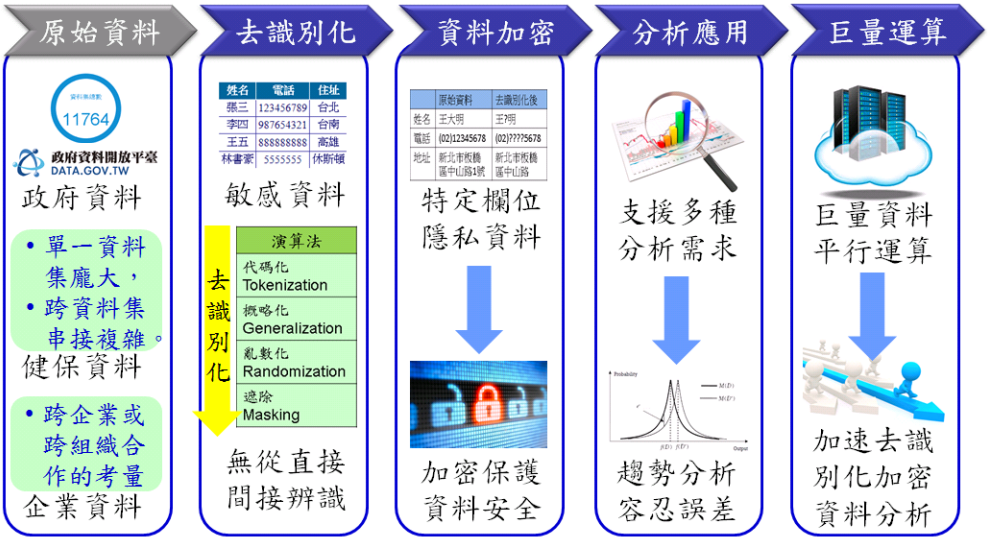
雲報專欄：大數據下資料隱私保護議題與技術應用

--中研究院特聘研究員 黃彥男/技術專家委員會委員

由於網際網路以及雲端儲存/運算平台的興起與普及，巨量資料的運用成為熱門的研究議題。巨量資料的來源多元，除了網路使用者的各式使用資料外，行動設備持有者的位置移動資訊、物聯網設備在不同定點所收集的資訊，其數量亦相當龐大。過去如此龐大的資訊難以儲存、分析，但在資訊科技的進步之下，透過分散式儲存與運算，巨量資料若能經過有效的聚合與分析，將可產生許多過去無法得知的研究結果。除了民間企業外，政府單位也加入巨量資料分析的風潮。民國 104 年 6 月財政部曾利用稅務資訊進行「大數據薪資分析」來分析國人的薪資實況，了解企業近 3 年薪資變化狀況，相對於過去的傳統統計方式，可看到公共政策的更多面向，融入大數據分析方法，作為政府在擬定公共政策時不同的決策參考資料。而世界各地政府或國際組織目前也積極推行 Open Data，將政府擁有的各式數據加以開放，供民間做各類分析，以求集眾人之智慧，發掘並解決社會面臨的問題。目前我國政府所提供許的開放資料項目數已達 11918 項，且已有許多應用程式運用之。國家衛生研究院全民健康保險研究資料庫也提供健保資訊供學術研究之用，自 1996 年起至今，使用該資料庫所進行的研究已累積超過 1800 篇。除此之外，許多民間機構也嘗試了解將自身擁有的資料對外開放，是否能帶來營運上的效益，如對天氣風險管理公司的研究，以及對人力仲介公司開放資料的研究，都在在顯示開放資料不僅對於政府有重要的意涵，亦可能成為民間企業嘗試獲取競爭優勢的有效手段之一。下圖表達資料隱私保護的處理流程，包含從原始資料出發，經過去識別化、資料加密、加密後資料分析，以及在實務應用上所需的整體系統架構和巨量運算加速技術都涵蓋於內。



無論是巨量資料的收集或是政府推行 Open Data 的開放，都對個人的隱私及資料安全性形成嚴重的挑戰。巨量資料本身可能包含個人的隱私資訊，若不能有效的加以編碼與屏蔽，則政府只能傾向不提供具潛在爭議性的政府資料，但此類資料往往具再利用價值，例如政府財政預算及交易資料、公司登記、土地登記等相關資料。實際上，世界各國政府提供公眾有用之資料集佔不到全世界政府資料的 10%，呈現各國 Open Data 政策實行還有很大進步空間。而企業在進行跨組織合作時也將受限於個人資料保護法，而無法有效的交換資料，共同進行分析。以學術研究為例，過去學者向政府取得研究資料時以書面約定不外洩資料為主，但新版個資法 16 條已規範，公務機關如果要提供相關資料供研究所須，必須要使資料無從識別特定當事人後方得釋出。但個資法中所謂的「無從識別當事人」，僅僅透過部分資料欄位的遮蔽恐無法達成。例如在開放查詢的資料庫中，如果能用一系列的欄位資訊查詢到一位特定的使用者，此時再追加的查詢屬性若也繼續有回復一筆查詢結果，則可斷定擁有這些同一人永遠所有的屬性。是故目前民間意見認為即使部分欄位有遮蔽或進行亂碼化(scramble)仍不足以達到完全的資料隱私，實為有據。但是從技術上而言，任何的去除資料識別性(de-identification)的嘗試，均不免必須對資料做出更動，而影響到後續運算的準確性，意即資料的去識別化程度與資料後續運算的準確程度有一種取捨(trade-off)的關係。

目前對於資料去識別化的研究，學術界提出不同的方法來達到去識別化，像是 k-anonymization、t-closeness、l-diversity 之類的概念都陸續有被提出，但是現階段對去識別化技術來說，最嚴格也最被一般學術界所接受的標準則是差分隱私

(differential privacy)，因為其有嚴格隱私保護保證並且其相對應的實作方式也都相對於其他去識別化技術來得簡單容易。差分隱私保護機制的概念是為了保證任一個體在資料集中或者不再資料集中時，對最終的查詢結果幾近零差別，也就是說，若有兩個幾乎完全相同的資料集（兩者的區別在於一筆資料不相同），分別對這兩個資料集進行任意查詢動作，同一查詢在兩個資料集上產生同一結果的機率接近１。

除了去識別化資料隱私保護機制外，針對某些特別敏感的資料，資料擁有者可能會對資料做加密之後才釋放給發佈出來給資料使用者。而資料使用者在獲取資料之後，對獲取的加密資料進行運算。在得到加密後的運算結果，會將加密後的運算結果丟回給資料擁有者幫忙解密，資料擁有者解密後給予資料使用者分析結果。在這樣架構下有多個可能的研究方向，但是最重要的就是如何既能加密資料，又能讓資料使用者進行分析。一個可能的作法就是目前最新型態的 fully homomorphic encryption。這種的 fully homomorphic encryption 可以保證資料使用者可以對加密資料做算數的四則運算。Fully homomorphic encryption 雖然有這樣強大的計算與安全上的保證，但是其運算速度非常慢。雖然自從 2009 年發現以來，迄今實作上已經加速了 1000000 倍的速度，但是仍遠遠偏離可實際應用的地步。

依目前現有的研究，對於上述的個別問題均已經有一定的累積，但是同時兼具去識別化、加密、高分析精確度以及具串流特性的研究則尚屬少數，且暫無突破性的結果。我們認為面對未來巨量資料來源的多元化，網路資料、行動設備資訊及物聯網收集資訊等龐大的巨量資訊，都需要經過可靠的去識別化、有效的聚合與高精確度地分析，才能獲得更多過去無法得知的研究結果，而「去識別化技術與針對加密保護後資料進行分析的機制」為研究主要之創新點。