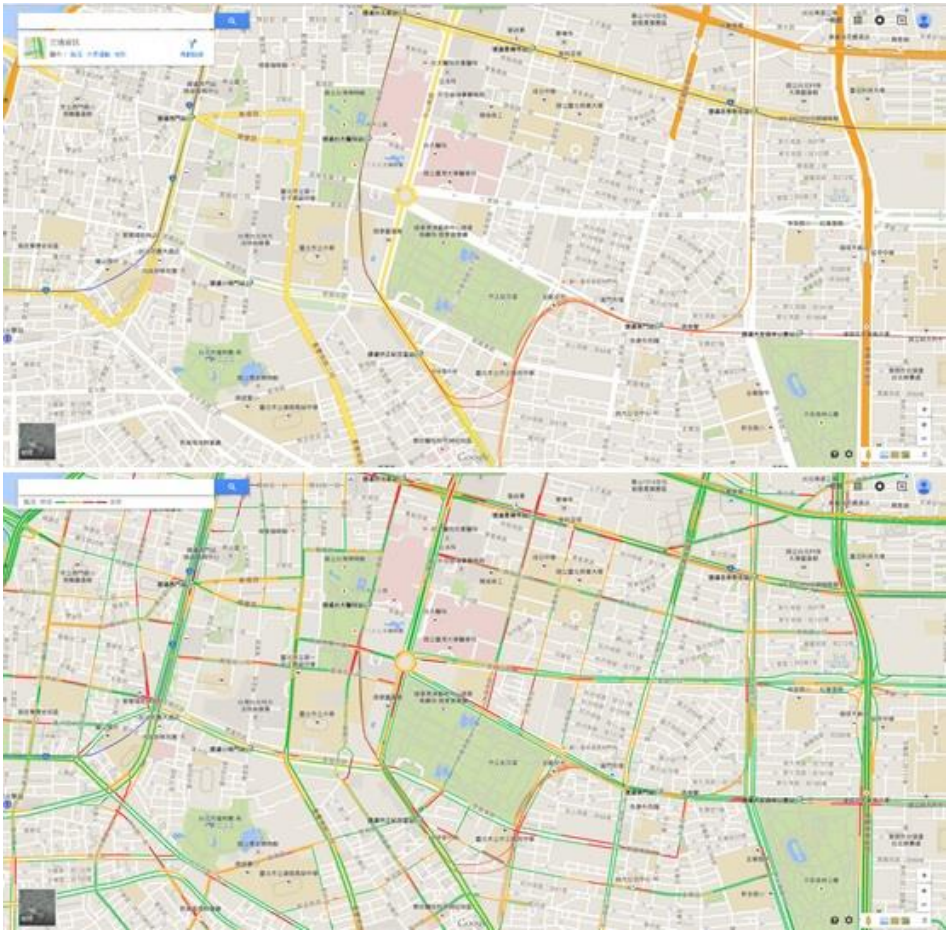
「大」數據一定要大？

在2014年12月行政院院長毛治國甫上任時，就提出「開放資料」、「大數據」及「群眾外包」等新聞媒體時稱「毛式三箭」的政策，希望利用網路與科技協助政府創造有感施政，其中「大數據 (big data)」這個概念其實更早已盛行一段時間，然而如果詢問台灣民眾什麼是「大數據」時，得到的答案常常不外乎就是直接從字面上的解釋，也就是和龐「大」的數據資料有關等模稜兩可、模糊籠統的回答。其實大數據並不一定與「大」有關，並且龐「大」或者是非常「大」量的數據資料也不一定就能夠稱之為大數據。

其實大數據簡單來講，就是當我們所能蒐集到的有效樣本數等於或趨近於母體的所有的樣本數時，我們也就更能精準、精確的去分析，從而得知母體的真實情況。不過等於或趨近於母體的樣本數量並不一定就是非常大量的，因為這必須視母體的數量或者大小來決定。勤崴國際科技黃晟中副總經理舉例，如果說今天真正的母體數量是五百萬，而我們蒐集到的樣本數為一百萬，那麼這確實可以稱之為大數據，但是如果今天我們所要分析的母體數量只有五百，而蒐集到的樣本數量只有一百的時候，將它稱做為大數據就好像就有點太過了。

黃副總認為，其實我們現在所謂的「大數據」就是以前我們所學的統計分析方法之延伸，差別在於過去因科技較沒有現在發達，所以當我們要去採集樣本數時，能用的方式既不多也可能需耗費大量的人力物力等資源，使得樣本無法快速並簡單的取得，導致採集到的有效樣本數可能只是母體總數的一部分甚至是一小部分，而如果我們以這部分的樣本數來解釋、描繪母體時，可能就會出現誤差甚至誤差會大到不足以解釋母體的情況發生。但是隨著科技的進步，人們有更多並且更輕鬆的方法來蒐集資訊的時候，我們所可以利用的資訊也就更能完整描繪出母體的樣貌出來。

那如果這樣說，不就還是代表我們因為能夠取得大量的資料而使得我們能更清楚、完整的描繪出母體的樣子嗎？其實更正確的說，黃副總表示，應該是我們能夠以更精準、更簡單、更直接的方法，取得我們所需要的樣本來進行分析，而非先蒐集一大堆數據資訊後才篩選出我們所要的資訊。每次蒐集的數據結果並不一定是大量的，這是由母體數量的多寡來決定的，但相較於以前，我們現在所使用蒐集樣本的方式卻有了更多樣化的選擇並且可以同時進行，而這些方法讓我們能夠快速的大量累積樣本，造成許多人誤以為所謂的大數據就是不分青紅皂白的大量蒐集樣本，然後才從這當中去挖掘、過濾出我們所想要的資訊。換個方式說，大量的樣本數就有點像是伴隨大數據這個概念所產生的結果，而真正的大數據其實是因為透過現今的高科技，讓我們得以擁有更多元、更直接、更簡單、更快速的方法來獲取我們所需要的樣本，而這些樣本比起以前所採集到的樣本，不只是數量更多，還可以讓我們更精準的來描述、形繪出母體的真實狀況。



google map不只提供導航，還提供了即時路況以及各式道路、交通資訊

（圖：google map）

勤崴不只與google合作，創造出年輕人甚至是大多數民眾最愛用的google map，還囊括臺灣市場七成以上的地圖資訊供應商角色，大數據自然功不可沒。黃副總提到，因為智慧型手機的盛行，所以現在有很多的道路交通分析都可以倚靠用路人的手機來進行樣本的採集並馬上進行分析。副總舉例，之前他們就曾利用此方法分析過台北車站忠孝西路公車專用道拆除前以及拆除後的車流速度比對。副總認為臺灣有許多的道路是可以使用此種方式來分析追出原因，並進而解決問題，不管是道路設計、流量分析還是事故原因等等，都能夠利用大數據的方法來改善。一項政策的好壞數字會告訴我們一切，因此若我們能好好利用大數據，那麼它將不只會為我們發掘並改善許多我們不知道的問題，還可以讓我們有足夠的證據與民眾溝通並說服他們，使得政策的實施也能夠更加容易、更加暢通。

口述/勤崴國際科技副總經理 黃晟中

採訪撰文/中華民國運輸學會 李明展