〈轉載自清流雙月刊 106 年 7 月號〉

工作人都該懂的大數據實務運用

法務部調查局資通安全處 雷喻翔

隨著通訊網路的爆炸性發展以及物聯網應用的普及性大數據的應用一時之間成為顯學，產官學界也趨之若鶩，希冀能夠運用大數據的威力，提升整體業務效能。何謂大數據？其實也不是甚麼嶄新的技術，數據分析早已行之有年，從最基本的統計分析到經由人工智慧演進的資料探勘機器學習，相關領域都有專家學者競相投入研究。之所以近幾年大數據成為朗朗上口的話題是因為電腦處理速度的提升，使得原本需要耗費大量時間資源才得以獲得的分析結果，如今在短時間之內即可知道隱藏在茫茫資料海中細如針線的重要資訊。

2011 年，時任紐約市長的彭博為了改善頻傳的火警，除了政策面的訂定外，同時也決定從資料面著手，重新檢視消防局及建築部所有相關的建築資料，大數據一躍成為熄滅紐約市烈焰光火的利器。經由數據挖掘與分析，紐約

消防隊發現高危險群的房屋數量從原本的 13%竄升至 70%，有了更詳盡、更周全的目標之後，消防人員得以進一步針對需要檢驗的建築物進行分類，篩選出屬於高風險群而需要特別關注的建築物。透過大數據的著力，紐約市的火災數量確實獲得明顯的改進。

本篇文章不在數學統計等多加著墨，主要是探討大數據實際的應用，不需要龐大的資料庫亦無需多麼艱澀的數學模型，僅利用政府公開資料以及人人皆可使用的開放原始碼應用程式便可進行大數據的分析，進而激發工作上的新思維，資料俯拾即是，實作勝於一切。

資料蒐集及所用之工具

首先蒐集包含從臺北市政府開放資料及網路公開資料等 6 種不同的資料，其中包含臺北市 1)竊盜統計、2) 里資訊、3) 捷運站地址、4) 公車站地址、5) 百貨公司賣場地址及 6) UBIKE 站地址，各資料表來源如下表一所列，其中臺北市竊盜統計資料表又可分為房屋竊盜、汽車竊盜及腳踏車竊盜案件。

表一、本文蒐集之資料表統整

資料	資料來源	資料筆數
臺北市竊盜統計	臺北市政府公開資料 http://data.taipei/	(房屋/汽車/腳踏車) 1,012 / 177 / 607
臺北市里資訊	臺北市政府民政局 http://ca.gov.taipei	456
捷運站列表	臺北捷運公司 http://www.metro.taipei	108
公車站列表	臺北市公車資訊 http://5284.taipei.gov.tw/	5,026
大型商場資訊	OneLife 生活網 http://onelife.tw	46
UBIKE 站列表	UBIKE 網站 http://taipei.youbike.com.tw	274
實價登錄	信義房屋	402

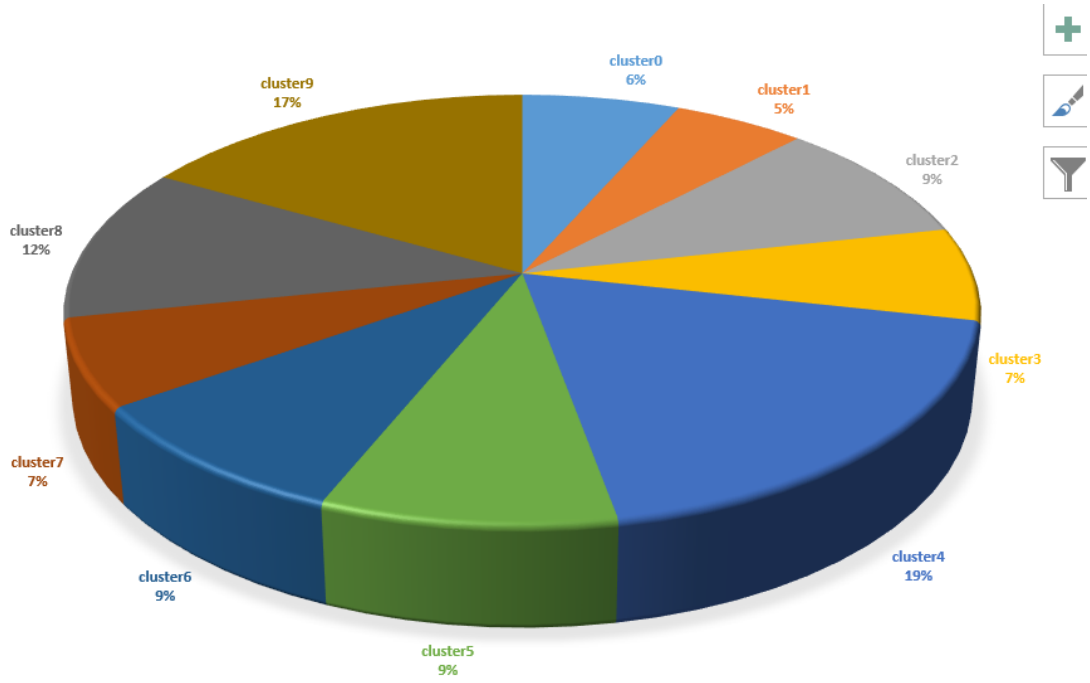
資訊		

由於每個資料表擁有共通的屬性：地址，所以我們利用地址資訊將各資料表整合起來。考量臺北市的行政區域由大到小以區、里、鄰劃分，如下圖一所示，共有 12 區、456 里、9, 594 鄰，若是以一個區做為行政區域之劃分，其所涵蓋的區域面積過廣，對於在一個區內之東部或西部發生的犯罪案件並無太高的相關性，且行政區之面積大小也有很大的差異；而若是以鄰做為行政區域之劃分，其所涵蓋的面積又過於狹小，以鄰為歸類之所屬案件資料不足，各鄰之間的屬性（特徵）也較難有差異性。因此，我們取位於中間的里做為本專題之行政區域劃分，並以里為基準結合所有的資料表。在處理臺北市里資訊時，由於各個里的範圍不一，要找到里的實際中心較為複雜，因此我們採用里辦公室地址來代表每個里的中心座標。

圖一、臺北市行政區劃分。

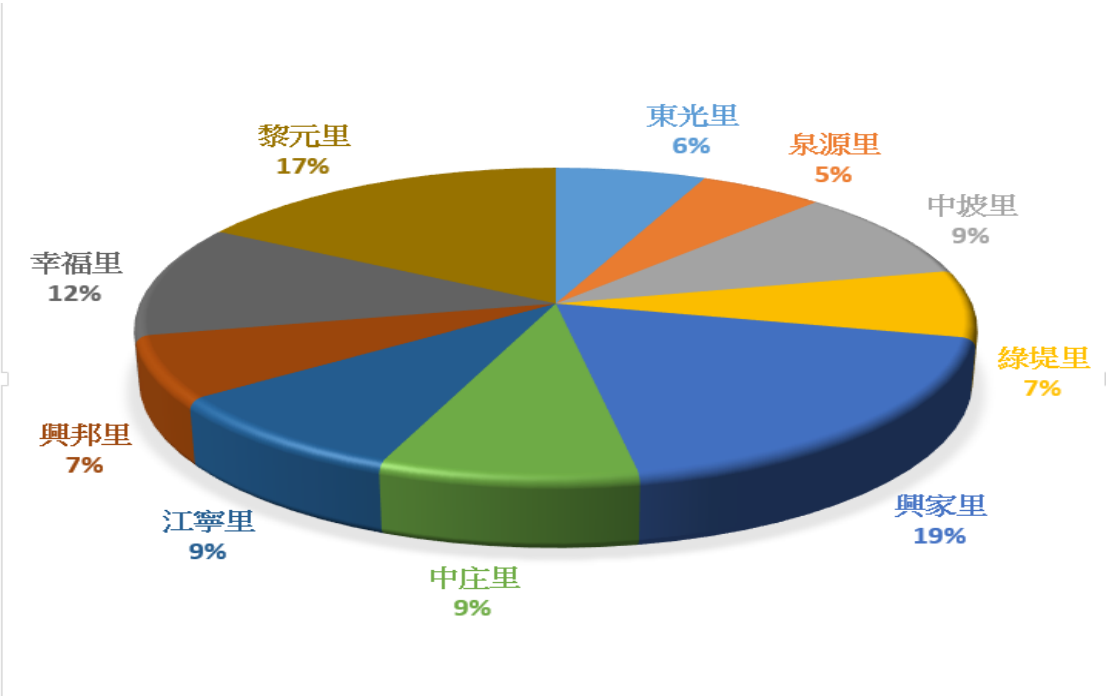
資料歸納分析

首先統計各個里所發生的竊盜案件次數、公車站總數、百貨公司總數以及里距離最近之捷運站名稱及其距離等資訊，接著先將全臺北市 456 個里分成 10 個群，進一步分析每個群的屬性。本文採用大數據領域中廣為使用的開放原始碼應用程式 WEKA 為分群工具，其結果如圖二所示。



圖二、各群所佔之里數量比例

從上圖可以看出各群中里數量的分布從最高 19%到最低 5%，沒有特別集中於某一群，顯見臺北市的住宅環境並無明顯趨於一種型態。為了凸顯每個群的屬性，我們進一步在每一群中找出地理位置上距離中心點最近的里做為代表，其結果如圖三所示:



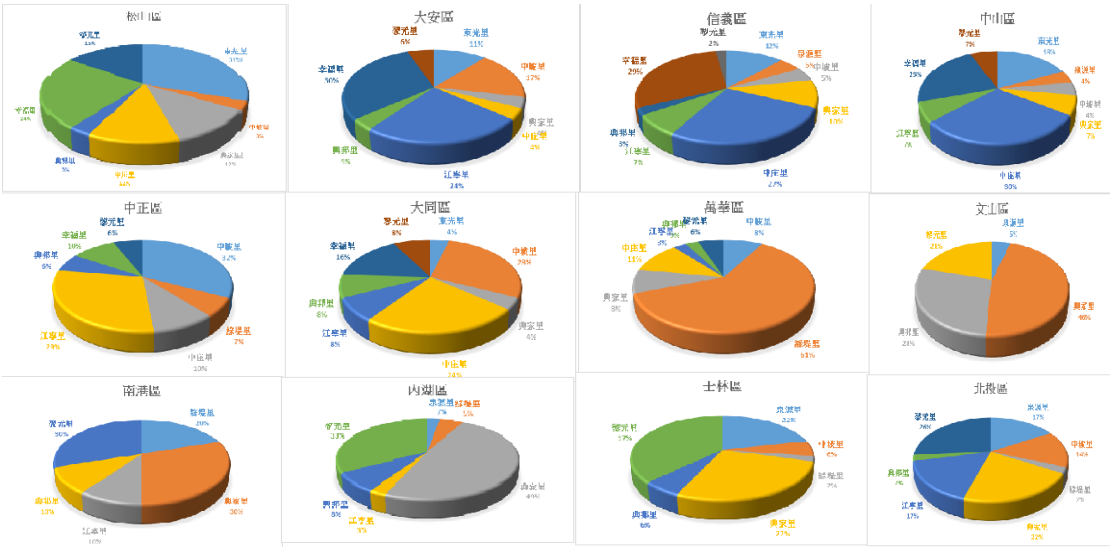
圖三、以離各群之中心最近距離之里做為代表

有了以上的轉換之後可以更明顯的看出群內的屬性關係，舉例而言，幸福里位於中正區善導寺捷運站旁邊，該位置不僅交通方便，離各大百貨也很近，同時房價亦居高不下，由此可知幸福里所代表的群其屬性較偏上述特徵。

接著再針對一些比較特殊的群來做分析，首先是以泉源里為代表的群組 1。檢視該群的特性，我們可以發現群組 1 距離捷運站最近，但是一公里內的捷運數反而不多，同時可能也因為人口數不多、非鬧區(MRT、公車站牌數、UBIKE 站數不多)，所以竊盜次數最少；群組 0 及群組 4 的

特性則是人口數較多、戶數較多；群組 8 則是房價較高，屬鬧區的型態，以大安區、信義區及中山區為主。

另外也比較臺北市 12 個行政區區內的差異度，從每個區內所屬不同群的比例可以了解該區內的差異性。從分析結果來看，萬華區及內湖區內的屬性較為集中，兩者皆有接近 50%的里被分至同一群中；其餘的區皆無明顯的集中屬性，其中中山區更是被分類成 8 群，顯見其區內的分散性。這項分析提供另一種對於臺北市地區的傳統思維，一般而言，提到信義區、大安區時總讓人聯想到較為熱鬧、交通便利的區域，然而透過分析資料，得以更進一步地篩選出真正符合期待的區域及地點。總結分析結果如圖四：



圖四、各群之間總結圖示

本文之數據分析亦可做為後續研析的前導，文中打破臺北市 12 個行政區的分界，重新將區下轄的里重新分類，歸納出隱藏其中的關聯性。擁有同樣屬性的里，其人口組成及習性也有較高的機率偏向同一群，此種做法不侷限於臺北市，可套用至全臺各縣市地方。後續若能再整合更多的資料類型，例如貪瀆案件發生的地點、毒販常出沒的區域、經濟犯罪的熱點或人潮聚集軟目標，如此一來或可提供執法機關一種新的犯罪追查手法。

總結

處在被智慧網路包圍的社會，生活週遭無時無刻在產出數據。當然我們可以無視這些資訊，一如往常地執行日常工作，但是若能搭上這一波大數據熱潮，從中提升自己的工作能量，何樂而不為？

然而大數據並非百利而無一害，由於經手龐大的資料量，隨之而來的是資料隱私權的問題。在極為注重個人隱私的現今，享受大數據所帶來的便利之餘，不得不注意個人資料的保護。現就公開資料和私有資料分別討論：公開

資料既為公開，相關牽涉到個資的部分已去識別化，因此當無洩漏隱私權之疑慮，關鍵部分為政府私有資料。政府有關部門就各自職掌範圍蒐集民眾個資有其法律上之規範稅捐部門就其稅捐稽查業務、衛生部門就健保業務等都早已行之有年。然而大數據的應用可貴之處在於大量異質性資料的整合，如同前述所提紐約市於火警預防的研析就觸及不同部門間的資料，於此必然會產生跨業務資料使用權限的問題。對此，筆者認為若有適切的稽核規範，明確限定何人、何時、何地可以接觸整合後的資料，且有專職人員定期檢查使用情形，如此在個資運用上則無侵犯或洩漏個資之虞。

資訊技術是未來工作的趨勢，不用將它視為高不可攀的技術，畢竟我們不是鑽研統計數學的專家學者，所要做得是掌握概念、操作應用。藉由簡單的資料蒐集及分析，原本單純的結果便產生重要的附加價值，這正是大數據應用的精神所在。看完本文後，也一起來動手吧！