數據怎麼說話？

郭位

我們常說要憑數據說話，但是不是都了解數字背後的含義呢？就以英文的 12 pm 為例，到底是指「正午 12 點」，還是「午夜 12 點」？再以一個人的成就與學業成績來說，常有人讀書時成績好，卻事業無成，因而怨天尤人。成績好的人沒有什麼成就，或是成績差的人成就很大，一點不出奇，因為一個人的成功，有千百個因素，成績好只是其中一個小小的數據點而已。

數據科學跨越多個學科，涉及統計、程式應用及專業知識，從大量複雜的數據中抽取有價值部分進行分析和深度研究。大數據適用於各行各業，如銀行與金融、醫護、資訊、運輸物流、零售及市場推廣、以至政府決策。

數據分析帶來的利益不勝枚舉，如何善用大、小數據十分重要。譬如台灣農業就因此得利而進化，德國巧用大數據而贏得 2014 年世界杯，川普則用大數據擬定競選策略而當選美國第 45 任總統。雲端運算因為大數據而存在；無人駕駛的汽車，也是靠感應器收集數據，分析路障狀況，幫助駕駛者順利上路。

電算功能的突飛猛進為大數據時代的來臨創造了條件。但是最重要的是具備知識與常識，否則再多的分析也沒有什麼價值。

大數據分析可以發掘許多有價值的信息，但是需要專家協助，才能夠了解數據背後代表的含意。在生活中，我們常聽到一些只憑個人經驗或主觀判斷做出誤導的結論。譬如許多對高等教育不甚了解的人，振振有詞認為小班教學效果比大班教學好。但據我幾年前做的一項調查發現，除了個別例子，在絕大部份情況下，大班與小班教學並沒有什麼顯著的區別。

在這個講究數據的時代，誤用統計數字，以訛傳訛的例子俯拾皆是，以下僅舉一例。四十年前在美國讀書，老美非常驚訝於台灣留學生成績特優，以為華人都是智慧出眾的天才。近幾年，統計結果顯示印度裔美國人是美國收入最高的族裔，所以有印度人富甲天下的說法。這類現象都是基於取自偏差的（biased）樣本，在美國讀書或移民美國工作的不是常態華人、印度人，他們成就突出，不足為奇，因此不可一概而論，誤以為華人、印度人全都出類拔萃。

不懂事的話，切不可亂引數據、亂說話。二十多年前，有人在可樂罐中發現針尖。這家公司的老總接受採訪時自信地說，「發生這種問題的機會只

有百萬分之一。」英文裏用 not in a million 表示發生的機率很低，但在這種情況下適用嗎？試想假使一天賣 500 萬瓶可樂，如果有百萬分之一的機會，豈不是說有五瓶可樂罐裏藏有針尖，這未免太可怕了吧！果不其然，這位老總隨後出來為自己的失言道歉、辭職。

最難的其實是只有小數據或者沒有任何數據。依一般經驗，數據不足往往變成言人人殊，莫衷一是。然而，完全沒有數據和經驗，有時候又必須做出決定。我當年在貝爾實驗室工作時，上司要求對全新的產品，在沒有數據的情況下，為可靠度做預測。愛因斯坦就在沒有數據和經驗的基礎上創立了相對論。我們的生活也有類似的例子，如初戀和第一次婚姻，也沒有數據和經驗下貿然投入。因此明白，雖然沒有數據，卻可以借鑑類似成品或他人經驗，舉一反三，做些結論。

僅僅靠著表面文章，容易做出誤導的結論。有時候，以幽默的方式談些數字或文字，可博君一笑。前兩年我在美航的飛機上，空姐對一般旅客並沒有任何表示，唯獨對一位體重極端超重、舉步維艱的旅客說：「You look good！」原來 good 可以有如此不同的解讀方式，這是有條件的。

有些人把數字當作形容詞使用，誇張吹牛，若是無傷大雅，倒也不值得計較。但最討厭的是，有些人隨便或者故意濫用數據，甚至惡意扭曲編造數據，絕不可取。

數字代表的不同意義，應該留待專業研究，才知所以。如果不懂該行業，或不了解數字在該行業的行情，最好不要插手。如果以亂講話作為當代社會的一個指標，那麼我們離國際化先進標準依然有相當的距離。

註：6 月 1 日於 108 年國立交通大學開業典禮演講。