讓鮮魚呈現自己的美味

漫談數據、科學、計算

作者：**謝復興**　譯者：**吳宏達**

謝復興現任教於美國加州大學戴維斯分校，他的研究興趣為數據力學與複雜系統之整合模式推論。

在與幾位年輕數理統計學家的閒聊聚會裡，有人問了一個似乎很難理解的問題：為什麼數據科學（Data Science，或譯資料科學）的重要性**現在**才冒出來呢？對許多專業統計學家來說，這個問題始終被忽視，因而從未被好好提問過，但這是個錯誤，而且有嚴重的後果。

為了回答這個問題，我們要先思考一個「子問題」：數據科學的專業學科目前設在哪裡？現在數據科學的大部分課程都設在社會科學院或商學院，跟統計系幾乎八竿子打不著甚至形同陌路。統計學這樣一個專門處理與分析數據的學門，怎麼落到如此處境，許多人都感到不解，不管是圈內人還是圈外人。從科學史的角度來說，這問題的答案既有趣又重要，對統計學的整體未來更是如此。

前述子問題的首要答案就是：「錢，一切都是錢」。當數學、統計、物理、化學以外的所有非核心科學開始盲目或雄心勃勃的蒐集數據，而變得越來越量化，這些領域的科學家就有了金錢上的誘因，讓他們把所有的研究工作及成果，連同學生所有修課學分都掌握在自己手上。他們無意與統計學家分享自己的研究，儘管統計學家認為自己才是數據分析的「合法專家」。顯然在當下此刻，許多科學家並不認為統計學家是唯一的專家。

下一輪的討論變成了指責，譬如：「這些人不曉得怎麼正確處理數據」、「缺乏紮實的訓練，甚至不知道怎麼正確使用統計方法」、「他們只是試遍所有的套裝軟體，只求做出自己想要的結果」、「他們一點也不清楚怎麼進行合宜的統計推論，遑論發展出自己的方法」。

在這幾位數理統計學家看來，如果這些非核心科學領域學者竟然也有許多疑問，似乎非常沒道理。但事實上，底下他們所提的幾個問題，相當大聲的呼應了 50 多年前，杜奇（John Tukey）1962 年發表在《數理統計年報》（*Annals of Mathematical Statistics*）的文章〈數據分析的未來〉（The future of data analysis）：

1. 在美國甚至全世界的所有統計系，有多少統計教授在處理實際的數據時肯扮演「黑手」的角色？
2. 有多少統計學家除了統計期刊之外，也在真正的科學期刊發表過論文？
3. 有多少統計學家會與自己研究主題的科學家經常保持聯繫？
4. 有多少統計學家真的解決過統計學以外任何一個實際的科學問題？

對於這些提問，這群數理統計學家所提供的訊息和答案相當令人憂慮。他們不但自許為數理統計學界的明日之星，還認為自己對這些數據分析的污名問題自動免疫。如果有人這樣問：數理統計學家是否理所當然有資格做「數據分析」？面對這種權利問題，他們的反應直白得令人意外：「當然有」。這很容易讓我們想起杜奇。他原本認為自己是統計學家，但在看到許多數理統計學家的以及他們對科學的態度以後，他改變了心意，而且在 1962 年的那篇文章的第一段，就開宗明義表明自己對數據分析的興趣。

在這個網際網路與大數據的時代，回頭研究一下杜奇為何這麼想，會是相當耐人尋味的事，這類反思也有助於說明數據科學的迅速竄紅。我們先用一

句話交代一下數據分析（data analysis）從 1962 年到現在的發展：杜奇文章裡提出的所有問題幾乎都仍然沒有解決。這句話驚人的短，但對我們理解現狀應該頗有幫助。

一方面，數理統計學家強調自己是在做數學；就這點，許多數學家是不同意的。而另一方面，正如前面提到的，統計學家一直自認是在做數據分析，儘管更精確的描述應該是：統計學家拿數據來適配（fit）他們所提出的模型，並且教導學生這些模型具有相當的「普適性」（universality）。相較之下，杜奇則明白指出，統計學家要嘛只在數學方面，要不然就只在數據分析上，找到自己研究工作的正當性（validity）。這個正當性不可能在數學與數據分析兩端同時發生。事實上，統計學在過去半個世紀的發展是極其偏頗的。數理統計學家很自豪地宣告「統計在科學上屬於數學」。但很不幸的，對杜奇來說，數學**並非**科學，主因是數學中人造的公理系統。另一方面，數據分析這個面向卻完全受到忽視，說得更確切些，應該是受到壓制。

被忽視了超過半世紀之後，很少有統計學家敢向外接觸實際科學問題，而到最後，數據分析在各個科學領域都活躍起來，現在還被大力推銷，改稱為「數據科學」。由於數據在所有的科學領域數量越來越龐大，因此迫切需要數據分析的技術。這個早該出現在統計圈外的覺醒，在數理統計學家眼裡，其實不該像是突然或意外間冒出來似的。

儘管如此，現階段與其說「數據科學」是一門真正的科學，還不如說它是個行銷用語，因為這個領域裡尚未建立起廣為人知的原理，而現在也正是時候，就「數據」與「科學」之間的關係給出一個簡

科學家錄下每隻被取樣的鳥在求偶期的歌聲。（楊雅棠 攝）

要而精確的敘述。

類別模式匹配

數據的價值主要在本身涵藏的資訊與知識內容。就某一個科學主題而言，知識內容所指的就是一組反應特徵（response feature）與另一組共變特徵（covariate feature）之間的連結性。科學知識最基本的形式之一，可以表現為「**類別模式匹配**」（categorical pattern-matching）的概念：共變端的某個模式類別，唯一包含反應端的單一模式類別中的部分成員。我們把這個關係稱為一個知識軌位（knowledge locus），因為有一部分的反應模式類別可以單獨用共變模式類別來解釋。透過反應及共變兩端模式所對應到的兩種表現，我們可以從字面上解讀出這個解釋。如果一個反應類別可由共變端兩個以上的模式類別來解釋，那麼反應模式類別所具有的**異質性**（heterogeneity）就算是確立了。

舉例來說，假設某單一反應特徵包括了鳥類學家確立的幾個鳥種類別，而在共變端，科學家錄下每隻被取樣的鳥在求偶期的歌聲。有了這些數據，科學家就能運用**類別模式匹配**技術，建立起哪種鳥唱哪些求偶曲目的相關知識。僅具有單一結構架構的邏輯斯迴歸（logistic regression），居然沒辦法納

入多重反應類別（即超過兩個類別），實在令人驚訝。

若能把求偶歌的模式類別當作是多重反應特徵，科學家就可以採用同樣的類別模式匹配技術，將求偶歌的模式類別關聯到從某隻鳥翅骨架測量到的幾個特徵的模式類別，同時也能關聯到幾個鳥喙特徵的模式類別。同樣令人驚訝的是，目前在統計學裡仍缺乏一個結構性的架構，可以納入多重反應特徵。實際上，杜奇在 1962 年文章裡已經提出這些問題，但 55 年已經過去了，這些著名的問題仍然沒有解決。杜奇甚至在文章裡已經暗示，分類學（taxonomy）及分類（classification）的技術可能會有幫助，一如前面提到的類別模式匹配。為何年輕統計學家完全忽視或不知道杜奇的提示呢？

類別模式匹配之所以在過去半世紀沒有發展起來，有以下幾個理由：

1. 本質上這是一個系統概念。
2. 它是資料驅動（data driven）導向的。
3. 沒有單一數學函數可含納這個概念。
4. 它無法遵循概似原理（likelihood principle）或中央極限定理（central limit theorem）。

由於在同一系統中，所有的研究對象和特徵都透過可能很曲折複雜的相依結構彼此關聯。另外，除非研究的系統是獨立且相同分布（independent and identically distributed，i.i.d）的，不然使用概似原理與中央極限定理在大部分系統裡並不妥當。因而在邏輯上，類別模式匹配的概念對數理統計學家而言並不是理想的研究題材。

此外，類別模式匹配的資料驅動取向使它幾乎無法被單一數學架構所含納。再者，一個系統可能在反應端與共變端包含許多機制。儘管這些未知機制導致的相依結構，正是等著從數據之中揭露的真正知識，但在為此建立數學模型及進行分析時，這些未知的機制不可避免將導致更多的阻礙與困難。但是諷刺的，正因為所有現實科學問題都處於複雜系統中，為了避開這些麻煩，數理統計學家理所當然的選擇在 i.i.d. 的假定下做研究，而不想去發展「探索式數據分析」（exploratory data analysis，EDA）所需要的方法。事實上，《探索式數據分析》正是杜奇一本著作的書名。為了偶爾掉一下書袋，杜奇的著述經常被統計學家所引用，但他的數據分析哲學卻完全被忽略。

數理統計學家與科學家的差異

曾經有位博士生向我請假取消例行討論，因為她要去史丹佛大學參加一個為數據科學家舉辦的就業博覽會。我問她是哪個系贊助這個活動，她告訴我是應用物理系，於是我請她幫我打聽一下為什麼是應用物理系。她回來之後，把她從史丹佛應用物理系的朋友那兒聽來的答案告訴我：我們研究應用物理的一直都在做數據分析，所以舉辦這樣的博覽會是很自然的事。

物理學家研究物理系統的時間，比統計學家早了幾百年，他們既研究系統靜態的特徵模式，也研究動態的模式如混沌（chaos）或紊流（turbulence）。統計力學這個物理領域，肇始於波茲曼（Ludwig Boltzmann）和吉布斯（J. W. Gibbs）的研究工作，即使是現在也仍然非常活躍。他們發展出來的模型

與分布，跟統計學家比起來相對較少也較簡單，其中主要的模型是易辛模型（Ising model），還有很常用的波茲曼分布（Boltzmann distribution）。事實上，統計學並不使用波茲曼分布，因為它不具解析形式。此外，波茲曼分布的系綜（ensemble）概念也跟統計學家基於 σ 體（sigma-field）建構的分布概念截然不同。

物理學家與統計學家在數據分析方面的真正差異，在於他們尋求的目標不同。物理學家會運用他們認為能攫取關鍵機制的簡單模型，看模型是否能顯現出從數據中觀察到的模式。也就是說，他們關注的焦點放在宏觀尺度的模式，而不是細微尺度的細節。相反的，統計學家會設法調校一堆已知的模型以找出一個「最佳」模型，來適配所觀察到的數據，而含在數據中的主要模式資訊則不是目標，所以通常會迷失在最小均方誤差之中。這兩者間的明顯差異，在領域科學上是否重要呢？

我個人的觀點是：很重要。一個系統當然是由確定性（deterministic，又譯決定性）的結構及隨機的隨機性所組成。確定性的結構形成能量最小的巨觀態，而所有的微觀態除了必須符合巨觀態，還會受制於系統特定的隨機性。這就是為什麼在以系統為導向的研究裡，擬仿（mimicking）是必要而基本的目標，因為我們必須將確定性結構和隨機性精確的擷取出來，才能進行一致的擬仿。這種擬仿的概念與統計上的拔靴自助抽樣法（bootstrapping）非常不一樣。拔靴法被譽為最重要的統計發展之一，但很可惜它主要還是仰賴獨立性或平穩性，而這在科學家感興趣的系統裡通常是不存在的。

因此，數據分析的重要工作，就是要忠實地從數據中抽繹出一致的確定性結構與隨機性。這項任務的重要性，可從數據矩陣很容易看出來。數據矩陣是一種呈現數據的基本形式。我們可以利用行與列的排列，在不做任何假設的情況下，純粹由計算來揭露出潛藏的區塊模式（block pattern）。這些潛藏的區塊模式就是確定性結構，至於區塊內的均勻性（uniformity）則是隨機性。確定性與隨機性這兩個擷取出來的成分合在一起，就可以擬仿數據矩陣。有了反應矩陣與共變矩陣，科學家就能利用這樣的計算進行類別模式匹配，而無須外加任何不真實的模型假設和建模結構。把探索式方法取得的結果組織起來，這些基於模式匹配的知識軌位可以幫助科學家發現各自領域中的原理。然而這種無監督式機器學習（unsupervised machine learning）的知識發現架構，雖然是跟隨杜奇在數據分析上的腳步，卻不太被統計學界所接受。

在物理學家重拾網絡（network）以研究大型複雜系統之後，現今在統計社群裡相當熱門的是高斯網絡建模（Gaussian Network modeling）的課題。在這個主題上，數理統計學家想要展現他們在網絡系統研究也能做出一些工作。但再一次，從我個人的觀點，這個課題同樣充分顯現出這群數理統計學家與現實脫節的思維及自毀心態，因為這項課題完全無視數據之資訊內容，更對它本應協助的真正科學領域完全幫不上忙。

網絡建模的架構是將數據矩陣轉換成變異數 / 共變異數矩陣（variance-covariance matrix），在此基礎上再假設類網絡的相依結構為已知。然後，當牽涉到極高維參數之估計時，為了使估計所需之計算可行，就引入稀疏（sparsity）假設。從兩個方

面來看，這種做法其實傷害都很大。

從資訊面來看，轉換成變異數 / 共變異數矩陣的程序，將導致大量資訊損失，因為除了相關性二元關係以外的所有資訊內容全都消失了。值得注意的是，只有在目標母體（target population）持續保持高維高斯分布，而且涉及個體的所有向量變數均為 i.i.d 時，這樣的變異數 / 共變異數矩陣才是有效的。在這類系統研究中，不允許有異質性存在。此外，基於稀疏假設的估計讓此建模太不實際，難以當成一回事。面對真實性與有效性的質疑時，這群數理統計學家就認真裝作是在做數學。在科學上，這不是誠實的做法。

另外，從數據分析的未來面來看，這樣的高斯網絡建模不但很可能無法實質解決任何實際科學問題，最重要的是，這種做法鮮明的展示了數理統計學家對數據與科學普遍的無知。這樣的無知當然會使這一行的學者對投入科學研究踟躕不前，數理統計社群可能因此更加孤立。與此同時，竄起的「數據與科學」（如果不稱為數據科學的話）將會取得勝利，協助那些在意自己手上的數據、致力於建構系統知識的科學家。

數據與科學間的循環鏈

由於數據主要是從科學系統產生或蒐集來的，從科學連結到數據的關係顯然是由領域科學家賦予的。另一方向，從數據連結到科學的關係，就必須由執行數據分析的計算科學家來促成。這兩個連結其實形成了一個迴圈。科學藉由科學家的實驗操作，將潛藏的資訊嵌入數據之中；而數據分析則是藉由計算科學家為了回饋科學的不斷探索，揭露了知識軌位的組織結構。這是數據與科學間良性的循環互動。在這樣的良性互動下，數據分析能做出貢獻，科學也得以有所進展。然而當這個循環鏈被打斷時，數據分析與科學雙方都會停滯下來。

在這個網際網路與大數據時代，很不幸的，數據與科學的這種關係，已經遭受到雙重威脅，一是數理統計學家的虛矯做作，另一則是看似普遍適用的機器學習演算法與套裝軟體。過去經常誤以為數理統計學家有能力處理數據分析的領域科學家，現在開始尋找現成可用的機器學習演算法。結果發現這些演算法根本不是特定設計的，甚至不適用於他們的數據型態。舉例來說，許多行為科學家為了使用網絡分析軟體來分析自己研究的社會行為數據，而把他們的數據轉換成網絡型態，得到的計算結果卻是以各種中心度（centrality）、群聚係數（clustering coefficient）或中介度（betweenness）來概括呈現的。這些進行網絡分析的人並沒有意識到，將原始數據轉換成基於兩端連結性的網絡（dyadic connectivity-based network）已經導致資訊損失，例如所有三元及多樣的相依結構都會消失。除了損失重要的資訊內容外，他們多半也很難用合適的行為意義來解釋計算的結果。從這些網絡特徵獲得的見解通常都模糊不清。這當然不是科學研究的常道。因此顯而易見的，一旦忽視數據與科學的循環回饋關係，就會拖延科學的進展。

雖然數理統計學家的成功方法幾乎與實際數據無關，而機器學習套裝軟體的設計卻主要都是商業市場導向，但是整體一致的資料驅動計算仍然存在發展的餘地，完善數據與科學的循環鏈。我認為實務數據分析的關鍵，在於願意花工夫做計算，**讀取數**

據並把數據視覺化、探究並呈現數據的資訊模式，以及願意去發現並組織由計算得出的知識軌位。

我所謂的「讀取數據並把數據視覺化」，意思就是將所有一維特徵變數，計算出可能帶有間隙的直方圖，以便同時顯出每一個特徵變數的個別特點及共同特性。每有一對這樣的直方圖，就可以導出基於定向條件熵（directed conditional entropy）的關聯性。相較於不切實際的線性關聯性方法，非線性的關聯性將促成任何特徵變數集合體都能重組成「協同 vs. 對抗」（synergistic-vs-antagonistic）的群聚組合。科學家花了那麼多時間、金錢及努力之後，如果只是把數據亂餵給統計或機器學習套裝軟體，卻未好好審視，即便不算罪過，也是很浪費的。

我所說的「探究並呈現數據的資訊模式」，則是指計算每一個數據矩陣的資訊區塊模式。這些數據矩陣是由共同媒介節點空間（media node-space，比如個體）及一個協同特徵叢聚所定義出來的。由於區塊描繪出特徵叢聚跟媒介節點叢聚之間的特定互動或耦合關係，且所有區塊都具有區塊內部的均勻性，因此模式的整體組合其實已顯現出基於模式的幾何結構，正適合當作數據矩陣的資訊內容。我們應該讓這種探索式分析完全取代、而不只是替代盲目的建模。要喜歡由資料驅動的真正模式資訊（pattern-information）必須有誠心。

而所謂的「發現並組織由計算得出的知識軌位」，意思是要從反應矩陣到共變矩陣來做類別模式匹配。將知識軌位好好組織起來，就能讓領域科學家從自己的數據集了解他們所感興趣的系統。目標系統相依結構的複雜性，最好是擺在一個簡單的平台上，讓它自己說話。至少在數據分析上，解開人為的糾結總是個好主意。

總而言之，為數據與科學間的循環鏈做計算時，每一步最好都要保持簡單、基本而不華麗，就如臺灣人煮魚湯的方法：只要在清水裡放一點點鹽巴、薑、蔥和新鮮的魚，讓鮮魚「說出」自己的美味。我們設想，在簡單的基礎上帶出豐富的味道，這個道理對所有的正宗美食都是真確而必要的，不管是法國菜、義大利菜、日式料理還是墨西哥美食。對於數據、科學與計算，同樣的原則也依然適用。∞

譯者簡介

吳宏達畢業於臺大流行病學研究所，現任教於中興大學應數系與統計所。

延伸閱讀

▶ Tukey, John “The future of data analysis”, *Annals of Mathematical Statistics* 33（1962）No.1, IMS。

▶ Tukey, John *Exploratory Data Analysis*（1977）, Pearson。

▶ Donoho, David "50 years of Data Science"（2015）。2017 年發表於 *Journal of Computational and Graphical Statistics* 26（2017）Issue 4。