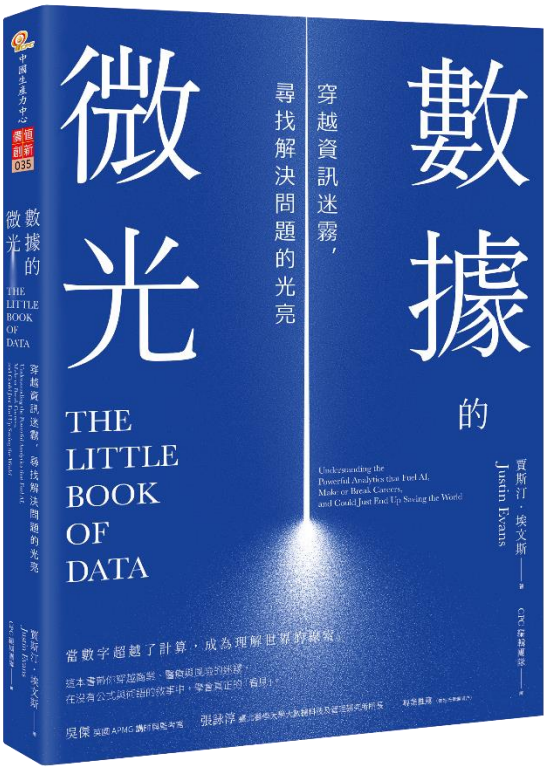

數據的微光

穿越資料迷霧，尋找解決問題的光亮

內文精選

第三章 資料科學

生死交關處的曙光

資料科學的躍進，似乎總發生在命懸一線的歷史瞬間。1665 年的倫敦，全城籠罩在恐懼之中。當時瘟疫已在阿姆斯特丹爆發，倫敦人深知自己的城市在劫難逃。這座城市顯然不堪一擊：人口過度稠密，貧民窟裡三、四十人擠在一棟房子裡。環境更是汙穢不堪，街道中央那溝壑般的排水道，竟同時充當下水道與垃圾場。人們隨手將一桶桶廢物往窗外潑灑；為了不把街上的糞便帶進家門，甚至在鞋底裝上小金屬台。

經歷過 1603 年、1625 年及 1636 年幾次大瘟疫的洗禮，他們深知染病後的慘狀：頭痛、發燒、嘔吐；腋下與腹股溝的淋巴結腫大如雞蛋，最終破裂溢膿。一旦染疫，那種劇痛足以讓人瘋狂嘶吼，且往往會奪走全家人的性命。在這種肅殺的氛圍下，人們瘋狂地翻閱每日印行的《死亡清單》(Bills of Mortality，一種專門記錄各教區死亡數據的晚報)，其焦慮情形，正如 2020 年的紐約居民，徹夜研讀《紐約時報》上的 COVID-19 七日平均疫情熱點圖。

在眾人驚惶失措之際，唯有一人保持了冷靜。約翰．格蘭特 (John Graunt) 是一名男裝裁縫，或許是因為職業使然，在與捲尺為伍的漫長歲月中，他對數字產生了一種近乎癡迷的執著。身處混亂之中，他暗自忖度：倫敦當局雖然花了大量時間盯著這些《死亡清單》，卻似乎沒能從中得出什麼實質的用處。

對格蘭特而言，這讓許多關鍵問題懸而未決：這場瘟疫到底奪走了多少倫敦人的性命？預計還會持續多久？那些負擔得起的人，是否該撤往人口稀疏、存活率較高的鄉間？而留下來的倫敦商人，又該如何規劃下一步？格蘭特深知，當時編纂《死亡清單》的方法相當粗糙且隨興。

他寫道：「每逢有人過世，鐘聲便會響起。」那喪鐘就像是一道集結令，召喚著一群被格蘭特稱為搜查員的人：她們是一群「宣誓就職的高齡婦人」。這些老婦會「前往陳屍之處」，透過目測屍體並「進行各項詢問」，藉此「判定死者是死於何種疾病或意外」。換句話說，這些「老婦」就是舊時代的法醫。你完全可以想像，要勝任這份工作需要多麼強悍的免疫系統。

那是一幅多麼奇特且堅毅的景象：每當鐘聲響起，這些老婦便步履蹣跚地朝教堂集結 (喪鐘總是為她們而鳴)。若用後來赫爾曼．何樂禮的術語來說，她們就是「列舉員」(詳第二章)。然而，即便她們的免疫力再強，也無法對「賄賂」免疫。房東們不希望自己的建築被劃上代表

疫區的「紅叉」，因此，只要給這些老婦一點小費，屋裡的屍體就可能被標上「肺癆」(即肺結核) 這種相對溫和的標籤。畢竟，那總比被標記為瘟疫好得多。

即便存在種種弊端，這套體制終究是在運作著。搜查員會將死亡人數與死因回報給教區秘書，教區秘書則在每週二彙整報告給大廳秘書。每逢週四，官方會發布統合後的週報，且每年進行一次年度統計。

身為一名引以為傲的倫敦人，格蘭特自覺能做得比官方更好。他蒐集了歷年的年度清單，統一了疾病定義，並清理了地理資料，釐清哪些教區應納入統計，哪些則不。接著，格蘭特將每年的死亡人數進行一致化與標準化處理，使其具備年度可比性。如此一來，趨勢便昭然若揭。

這或許顯得有些病態，但我發現格蘭特的記錄條目中，帶有一種宛如「蒙提．派森」(編註：英國超現實幽默表演團體)式荒誕喜劇的色彩。當然，清單中有些條目十分直觀，例如 1665 年的「謀殺與槍擊 9 人」或「投毒 1 人」(從現代美國人的觀點來看，這數據甚至顯得有些「古雅」得不可思議，畢竟美國現在每週有 800 人死於槍擊，不過這扯遠了)。然而，你還會看到死於俗稱國王惡病的「瘰癧」86 人、死於「心神喪失」 5 人、死於「絞痛與腸氣」134 人，甚至還有 23 人是單純被「嚇死」的。而在這些雜項之下，才是當年的重頭戲：瘟疫。那一年，瘟疫奪走 68,586 人的生命，死亡規模完全不在同一等級。讀到這裡，我笑不出來了，想必當時的倫敦當局也笑不出來。

格蘭特的工作處於資料解讀史的先河，這使他被譽為多項領域的開山鼻祖：第一位人口統計學家、第一位流行病學家、第一位進行時間序列分析的學者、第一位建立統計樣本的人，也是第一位關注嬰兒死亡率的人。

儘管只是名卑微的裁縫商，格蘭特仍出版了他的著作《對死亡清單的自然與政治觀察》（*Natural and Political Observations Made Upon the Bills of Mortality*）。當時一名官員塞繆爾．皮普斯（Samuel Pepys）購買了首版，這本書隨即引發轟動，英國皇家學會（Royal Society）更破格招募他為會員。自格蘭特之後，全歐洲的政界領袖開始統計生者與死者的人數，以便做出更明智的決策，將資訊發揮出更大的價值。

儘管如此，格蘭特仍在苦苦思索：資料究竟如何在某種程度上，成為解決現實問題的手段？他將自己的研究比作「愚笨學童」的信手塗鴉；他擔心自己的作品學術涵養不足，並試圖為自己辯護。「我之所以繼續研究，」他解釋道，「是為了從那些『虛幻的繁花』中，為世人呈獻些許『真實的果實』。」

在這裡，所謂的「虛幻繁花」，並非指那成千上萬死於瘟疫的軀體。正如一位歷史學家所寫：「整座城市聞起來就像個巨大的茅坑。」那景象與花香毫無關聯。不，那些「虛幻的繁花」指的是他所蒐集的數據表，是那些描繪死者的數據。

這些數據本身沒有重量，也缺乏實體。它們既不是救治病患的良藥，也不是撲殺帶疫跳蚤的殺蟲劑，更不是什麼刷洗消毒的衛生工具。它們僅僅是印著數字的表格。瞬息即逝，縹緲如煙。然而，「果實」卻是不同的——果實能提供營養，具有實用價值。

請注意，格蘭特並沒有選擇其他的隱喻，例如任何與「真相」或「純淨」有關的對比。他沒有說：「從汙穢的煤炭中為世人呈現鑽石」，也沒有說：「從混濁的深淵中為世人呈獻清泉」。他所希望的，是從虛幻的繁花中為世人呈獻果實，那是你可以實實在在咬上一口的、有感的東西！

我直到三十七歲，才接手人生第一份管理職。到了這個年紀，多數走管理職的人，少說也該有過在冰淇淋店帶班的經驗，但我完全沒有。我不確定原因，或許是我刻意規避，也可能是這類職位始終與我無緣。因此，管理他人對我而言是一場全新的挑戰；更何況這份新工作是個高階主管職位，隸屬於尼爾森公司（Nielsen Company）旗下的一家子公司，專門負責「統計建模」，這對我來說同樣是個陌生領域。

這個子公司名為 Claritas，它之所以在數據圈聲名大噪，全歸功於一套名為 PRIZM 的消費者區隔方案。這套系統能根據全美各社區的人口普查特徵，將每個家庭歸類到特定的客群中。這些客群都有個俏皮的稱呼，例如「獵槍與皮卡車」或「藍血莊園」。每個區隔還會配上一張生動的人物縮圖，最後彙整成一張大海報，客戶總是對這張色彩繽紛、趣味十足，而且非常有故事感的海報愛不釋手。

然而，當時的我正處於焦慮的泥淖中。我一邊得摸索如何帶人、學習當個高階主管，另一邊還得強迫大腦塞進像「多變量分析」這種艱澀的統計概念。幸好，我身邊有戴夫．米勒（Dave Miller）。

戴夫在業界深耕三十年，是一位資深統計學家。他為人耐心、溫和且睿智，與我那種咖啡因攝取過量、既焦躁又憤世嫉俗的狀態形成鮮明對比，他簡直就是我這場職涯風暴中的避風港。與他共事的過程帶給我很大的啟發，我腦中很快萌生了一個念頭：PRIZM 這套方案過去一直被直效行銷人員壟斷，用來寄發百思買或好事達保險的精準廣告信，但我們或許能將它轉化為數位廣告的投放利器。（編註：直效行銷（Direct Marketing）是一種不透過中間商，直接將產

品或服務資訊傳達給特定目標客戶，並期望客戶能立即產生回應的行銷方式，強調互動性、可追蹤性及即時性，常運用郵件、電話、電子郵件、簡訊、電視購物或網路等媒介，目標是建立顧客關係與促成交易。)

在一場研討會上，我準備首次推銷這個構想。開會前一小時，我拉住戴夫不放，因為我還在努力理清腦袋裡那團雛形般的想法。我這人常這樣，不知道別人是否也如此：我會先產生一股強烈的直覺，然後才回過頭去證實或推翻它。

我一直擔心這套方法搬到數位環境後會失靈。戴夫的回應倒是很乾脆：既然它在直效行銷領域行得通，甚至剛幫我們拿下一座最佳模型獎，就說明它有其價值。當時評獎的關鍵在於「回覆率」。每一封實體廣告信都有專屬代碼，只要客戶撥打 800 免付費電話並報出代碼，行銷人員就能追蹤到這筆回應是來自哪一次活動；而我們的模型，正是回覆比例最高的一套。

我接著追問：「但我們該如何驗證它在數位廣告中同樣有效？」戴夫露出燦爛的笑容，給出了一個充滿智慧的回答：「其實，你只需要回答一個問題：這個方法，是否優於你目前『次好的替代方案』？」

戴夫為什麼露出那樣的笑容？他心裡清楚，我想要的其實是一個花俏的答案，或至少是一個在待會的客戶面前聽起來很專業的答案，最好是能把「多變量分析」這種字眼掛在嘴邊。然而，戴夫對自己的專業有足夠的自信，他知道何時該給出最簡潔的解釋。

「它是否優於你目前次好的替代方案？」換句話說：你用數據推導出的答案並不需要盡善盡美。它只需要比你昨天的答案更好就行了。只要做到了這一點，你現在擁有的答案就有價值，甚至可能價值連城。我從未忘記那段對話，只是當時我並未意識到，這個觀念其實早在幾十年

前就已經存在了。

約翰．圖基（John Tukey）是一位戰時統計學家。二戰期間，他致力於解決 B-29 轟炸機的技術挑戰、改良戰場測距儀，並參與破解納粹的密碼機 Enigma。到了冷戰時期，他任職於貝爾實驗室，協助研發 U-2 偵察機，讓美國得以偵測蘇聯可能的核武突襲。或許正是這種戰時思維，促使他最終「引爆」了統計學界。

1962 年，至今仍被視為統計學史上最具顛覆性的論文之一，題為〈數據分析的未來〉（Future of Data Analysis）。這篇論文名震遐邇，甚至擁有專屬的縮寫：FoDA。當時一家權威期刊刊載了這篇文章，與其並列的盡是些精確且充斥數學公式的研究。然而，約翰．圖基究竟是何許人也，竟敢炸毀自己投身的專業領域？

誠然，他曾是戰時統計學家，但他本質上仍是不折不扣的統計學者。他在新英格蘭長大，父親是高中的拉丁語老師，母親則在同一所學校代課。他畢業於布朗大學，對母校熱愛至極，總是繫著布朗大學的領帶。隨後他在普林斯頓大學獲得了教職與研究職位，並在那裡度過了整個職業生涯。約翰．圖基深愛他的妻子，兩人在土風舞會上相識。幾十年後妻子過世，他留下了唯有統計學家才說得出口的哀悼。他說：「一，遠遠小於二。」

在照片中，圖基穿著西裝外套、扣領襯衫並繫著領帶；他一頭灰髮剪得很短，身材有些厚實，看起來就像個認真、執著且成功的 20 世紀東岸白人精英。他的笑容中閃爍著一絲靈動，眼神則帶著某種疏離感，彷彿在與你交談的同時，腦袋裡正精算著某個天文數字。這樣的人，會是個革命家嗎？

「長期以來，」約翰．圖基在 FoDA 的開頭寫道，「我一直以為自己是一名統計學家。」等等，過去式？「以為」自己是統計學家？從文章一開頭，你就能察覺他所埋下的伏筆。「但當我注視著數理統計學的演進時，」他繼續寫道，「我有了理由去懷疑與動搖。」

是的，我們付錢給統計學家，就是為了讓他們懷疑、挑毛病、施加嚴密的審視。但約翰．圖基不只是個會審視的統計學家，他還是一個會「想像」的人。他在論文中構思了一幅全新的統計學藍圖：讓統計學脫離數學系中最枯燥的分支角色，獨立成為一門科學，也就是「資料科學」。圖基當時並未創造這個詞，那是後來的事，但他滿腔熱血地勾勒出這門學問的真諦：

- **數據分析必須追求實用性，而非安全性。**
- **數據分析必須甘於犯下適度的錯誤，以便在證據不足的情況下，也能指引出正確的方向。**
- **數據分析必須將數學論證視為判斷的基礎，而非證明的基礎。**

這是一次破舊立新的宣言，圖基認為，這門學科不該只是專業人士在瑣碎樣本間的小心求證，或是為了追求那絕對的「精確」而束手縛腳。與其追求虛假的無懈可擊，不如擁抱適度的不完美。我們應當學會在證據有限的情況下採取行動，將數學視為輔助判斷的利刃，而非供奉在殿堂裡的真理。畢竟，數據分析的靈魂在於實用而非證明，核心思想在於：做出有用的東西，並且與那些數學可能不如你、卻是你所研究領域的專家合作。

難道你沒聽到約翰·格蘭特正在喝采嗎？你不覺得格蘭特當年一定也曾像這樣，與那些「搜查員」們並肩而坐，仔細探詢她們的工作細節？他一定曾詢問她們：對於佝僂病與肺結核之間的辨別，她們究竟掌握了多少專業知識？圖基在研發新型測距儀時，難道沒去請教那些步兵與

軍官嗎？當有人想把你炸上天而你必須反擊時，我們該如何幫你更準確地判斷距離？

在戰火紛飛或瘟疫橫行的年代，我們沒有時間等待完美。數據分析，或者說資料科學，必須具備實用價值。這正是格蘭特的抱怨所在：所有閱讀《死亡清單》的人，都沒能真正「利用」它。而三百年後，致力於偵察機研究的圖基走上前說：「我完全明白你的意思，我們需要一門全新的科學來達成這件事。」

這門學問的準則為何？基本上，就是我的統計學家好友戴夫所說的：「你憑藉判斷力、利用不充足的證據所推導出的任何結果，是否優於你目前次好的替代方案？」如果答案是肯定的，那就足夠了。對於想躲避瘟疫的倫敦人、正發射肩扛式火箭的步兵，或是試圖不被納粹潛艇擊沉的艦長來說，「足夠好」就代表了一切。

我相信，圖基當年在論文中之所以展現出如此急切的口吻，還有另一個原因：他預見了數據革命的到來。四年後，圖基發表了另一篇論文，雖然名氣不及〈數據分析的未來〉，但他在此文中列舉了「當今影響數據分析的四大因素」。這四項因素中，有兩項與我們現在所說的「大數據」(Big Data) 息息相關。關於大數據，已有無數書籍與文章討論過，因此我簡單定義為：不再只是侷限於「樣本」，而是針對任何主題所蒐集的海量數據。

圖基撰寫這些論文的 1960 年代，並非大數據時代。然而，他當時已洞察到「眾多領域正迎向規模日益龐大之數據集的挑戰」，並預見「量化研究將在各行各業中受到前所未有的重視」。他進一步指出：「當務之急，是去理解各領域數據分析之間的共通性，無論是在核子物理、細胞核生理學、抗病毒藥物研發，還是民意調查之中。」他預見了一個多種學科，或許是所有學科，都將產生大數據的世界。這將對更快速、更精密的分析能力產生巨大需求。他也預言，數

據處理「不僅需要達到某種程度的即時性」，且其表現必須「足以媲美現有的專家判斷」。而這一切，帶領他做出了另一項重大預測：「電腦的發展速度將會大幅躍進。」

面對生物學、醫學、核子物理領域中噴湧而出的數據，人類心智已顯得力不從心，更遑論變幻莫測的民意。在那數據的汪洋中，體量無窮無盡，複雜度更是深不見底。我們需要協助，才能處理這一切。於是，為了與這份「無窮性」搏鬥，資料科學被迫演化成一個「仿生」領域，它是人類與機器的聯手協作，更是一場屬於「半機人」(Cyborgs)的競技。

而圖基的這番願景，還需要另一位 20 世紀計算領域的遠見卓識者相助，那個人就是——艾倫．圖靈（Alan Turing）。

約翰·圖基與艾倫·圖靈極可能曾有交集，而普林斯頓大學正是兩人學術軌跡的交會之處。在二戰期間，普林斯頓是全球尖端數學的核心陣地；當時圖基在那裡任職，而圖靈也曾前去訪問。兩人不僅年齡相仿（圖基僅小了三歲），更同樣投身於破解 Enigma 密碼機的工作。儘管背景如此相似，我卻很難想像他們能成為朋友，畢竟圖靈的性格實在太過孤僻怪異。若將兩人並列觀察，圖基與圖靈呈現出一組耐人尋味的鮮明對比：

- **圖基：身材魁梧、謙遜、接受在家教育，是不折不扣的美國派。**
- **圖靈：骨瘦如柴、不修邊幅，出身自謝伯恩公學與劍橋大學國王學院。**
- **圖基：勇猛果敢，如同一支勢如破竹的軍隊。**
- **圖靈：思維跳躍，如同一道橫跨星系的雷射。**

- **圖基：活出了完整的一生，八十八歲時在好友簇擁下與世長辭。**

- **圖靈：四十一歲時自殺身亡。當時距離圖基預言「電腦發展」將成為資料科學的主要動力，還有十年之遙，而圖靈本人基本上就是電腦的發明者。**

「電腦」(Computer) 的舊定義，其實更接近我們今天所說的「計算機」(Calculator) : 那是一個接收特定問題（例如：147 乘以 29 等於多少？)、隨即回傳答案（4,263）的裝置。在當時，有些團隊專門負責執行運算任務，而這些人就被稱為「計算員」(Computers)。說來諷刺，這位電腦之父在當時卻是個不稱職的計算員，打起字來總是亂七八糟。

二十四歲那年，圖靈將眾人帶往了更接近現代電腦概念的領域。他當時正鑽研德國數學家大衛．希爾伯特 (David Hilbert，1862–1943，被譽為數學界的無冕之王及最後一位全才數學家)，在幾年前提出的難題「判定性問題」(Entscheidungsproblem)。這個命題的核心在於：對於任何數學命題，你必須能夠證明它是「真」還是「偽」。

圖靈的恩師後來回憶，他之所以會投入這個問題，很可能始於旁聽了自己的一場講座。「我想，一切的開端，是他來聽了我的課。當時我也許提到，希爾伯特的判定性問題本質上只是某種機械式的計算；我甚至可能隨口說過：這件事交給機器來做就行。圖靈把這個想法聽進心裡，並試著一路追問下去。」

（本章未完，精彩內容請參閱本書）

目錄

作者介紹

賈斯汀．埃文斯（Justin Evans）

在數據與科技交會的領域深耕二十年，不僅是創造數億美元營收的商務專家，更是賦予枯燥數據靈魂的轉譯者。足跡橫跨三星、康卡斯特與尼爾森等全球頂尖企業，協助領導者在演算時代找回主導權。兼具人文與商管背景，擅長將艱澀的 AI 概念化繁為簡，其專欄《The DataStory》廣受業界關注。

他的文字魅力同樣展現在文學創作中，他的小說《A Good and Happy Child》曾獲《華盛頓郵報》年度選書殊榮，並吸引派拉蒙影業改編影視。曾獲選為美國最權威的學術榮譽學會 Phi Beta Kappa 會員，並獲紐約大學史登商學院 MBA 學位。目前在紐約市工作與生活。