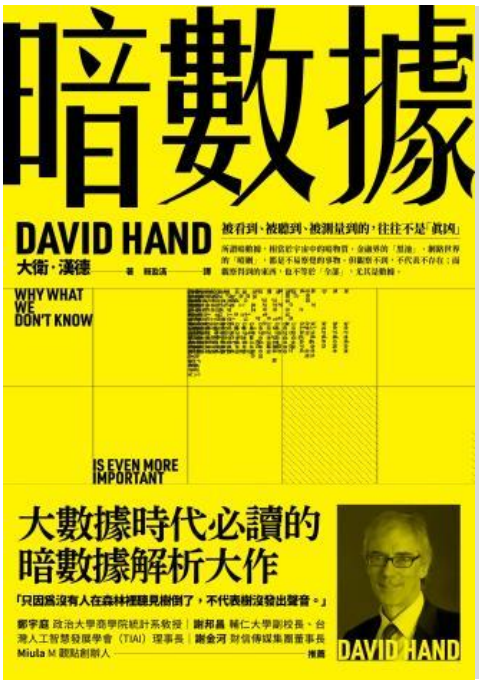

112 年 6 月份　推薦書目

暗數據：被看到、被聽到、被測量到的，往往不是「真凶」

作者：大衛・漢德
譯者：賴盈滿
出版社：大塊文化
出版日期：2021 年 5 月 27 日
語言：繁體中文
ISBN：9789865549961

作者簡介

大衛・漢德 David Hand

英國統計學權威，目前是倫敦帝國理工學院數學系榮譽教授暨資深研究調查員。曾擔任倫敦帝國理工學院及公開大學統計學教授，為合格統計學家及多家學術機構成員，多次獲獎，學術成就非凡。除了常上媒體接受訪問，也曾協助警方調查科學詐欺案件。

曾出版七本著作，其中包含《統計極簡介》（Statistics: A Very Short Introduction）、《資訊世代：資訊如何掌管世界》（Information Generation: How Data Rule Our World）、《不大可能法則》（The Improbability Principle）。

譯者簡介

賴盈滿

倫敦政經學院科學哲學碩士，現專事翻譯。

譯有《分心不上癮》、《跳舞的骷髏》和《成功的反思》等書。

以上資料取自博客來網路書店

內容簡介

身處大數據時代，不難以為我們擁有做出好決定的一切數據。但我們擁有的數據其實從未完整，甚至只取得冰山一角。就如同宇宙大部分由暗物質組成，雖然存在卻不被看見，資訊世界也充滿了暗數據，為我們所無視，非常危險。在這本《暗數據》中，數據專家大衛・漢德帶領我們踏上一趟啟發人心的精采旅程，走進我們看不見的數據的世界。

本書探討許多對於暗數據視而不見的情況，討論這些情況如何讓我們做出錯誤、危險，甚至災難性的結論與行動。作者檢視了現實生活中的例子，從挑戰者號太空梭爆炸到

複雜的金融詐騙，並分享一套務實的暗數據分類法，說明這些暗數據是如何產生，以便我們學會辨別與掌控暗數據。作者不僅教導我們要對未知事物造成的問題提高警覺，也闡述如何利用暗數據，從中得益，讓我們得到更深入的理解，做出更好的決定。

序言

大衛・漢德

這本書與坊間大多數講數據的書不一樣。不論是介紹大數據、開放數據、資料科學的科普書，或講解如何分析數據的專業統計書，講的都是已有的數據，亦即存在你的電腦資料夾、桌上檔案匣或你寫在筆記本裡的資料。然而，這本書講的是你沒有的數據。或許是你當初求之而不可得的數據，也可能是你希望擁有、以為擁有但其實並未擁有的資料。我在書中將舉出許多例子來論證一件事：你遺漏的數據不僅跟你擁有的數據同等重要，甚至更關鍵。看不見的數據可能會誤導你，甚至帶來災難。我將逐一點明這些災難，並闡述它們如何發生、為何發生。但我也會說明災難如何避免，需要留意什麼以防憾事發生。最後我將指出（或許有些出人意料），一旦我們明白暗數據如何產生及釀成大禍，就能藉此一百八十度翻轉

數據分析的老方法，藉由（聰明地）隱藏數據，得到更深入的理解，做出更好的決策與行動。

數據（data）一詞在英文裡到底該視作單數或複數，一直未有定論。過去通常看作複數，但隨著語言演化，現在許多人都視之為單數。我在書裡大多作複數處理，只有寫來特別怪時才視作單數。既然常言道「情人眼裡出西施」，我的認知很有可能與你不同。

我對暗數據的理解是一點一滴從工作中累積的。一路上多虧許多人的挑戰，我不僅慢慢明白這些都是暗數據的問題，這些人也和我一起研究各種因應之道。這些問題涵蓋了醫療研究、製藥業、政府、社會政策、金融業和製造業等，沒有一個領域能倖免於暗數據的威脅。

我尤其感謝撥冗閱讀本書初稿的人：克里斯多佛洛斯・安納諾斯托普羅斯（Christoforos Anagnostopoulos）、尼爾・錢農（Niall Channon）、奈爾・亞當斯（Neil Adams）和出版社找的三位匿名讀者，讓我免於犯下太多難堪的錯誤。感謝我的經紀人彼得・塔拉克（Peter Tallack）大力支持我，協助我找到合適的出版社，並慷慨給予建議，提點我這本書該把重點與方向擺在哪裡。謝謝普林斯頓出版社的責編英格莉德・訥理奇（Ingrid Gnerlich），她是不可多得的聰明嚮導，協助我將初稿修剪成形。最後，我要特別感謝我的妻子雪莉・錢農（Shelley Channon）教授，謝謝她細心評論了好幾版的初稿，大大提升了本書的品質。

以上資料取自博客來網路書店

內容試閱

第 1 章　暗數據：我們看不見的事物形塑了我們的世界

數據鬼魂

讓我先從一個笑話講起。

前幾天我在路上遇見一位老人，他走在馬路中央，每隔五十步左右就在路上撒一小堆粉末。我問他在做什麼，他說：「我在撒大象粉。大象受不了這種粉末，所以都不會靠近。」

我說：「但這裡沒有大象。」

他回答：「沒錯！你瞧這粉末多有效！」

笑話講完了，來點正經的。

全球每年有將近十萬人死於麻疹，每五百名感染者就有一人死於併發症，其餘則是終生耳聾或大腦受損。幸好該傳染病在美國極為罕見，一九九九年只有九十九起通報病例。然而，二〇一九年一月華盛頓州麻疹爆發，導致該州宣布進入緊急狀態，其餘各州的通報病例也顯著增加。美國以外的國家也有類似情形。二〇一九年二月中旬，烏克蘭的麻疹爆發病例已經超過二萬一千例。歐洲二〇一七年有二萬五八六三例麻疹，二〇一八年卻暴增高達八萬二千多例。羅馬尼亞從二〇一六年元旦至二〇一七年三月底，則有四千多起麻疹通報病例，造成十八人死亡。

麻疹是可怕的惡疾，由於感染之後要過幾週才會有明顯症狀，很容易悄悄蔓延而不被察覺，根本還不曉得它在傳播，就已經被感染了。

然而，麻疹是可以預防的，只要接種疫苗就能免於被傳染的風險。而美國施行的全國免疫計畫也確實非常成功，應該說太成功了，使得施行這類計畫的國家的大多數家長，一輩子都沒見過或經歷過這種可預防疾病的可怕。

因此，當政府建議家長帶孩子去打疫苗，好預防這種他們從來沒見過或聽過親朋好友左鄰右舍得過、疾病預防管制中心還曾宣布絕跡的疾病，家長自然會對這樣的建議半信半疑。

為了不存在的東西挨一針？感覺就跟撒大象粉一樣。

只是麻疹和大象不同，威脅並未消失，始終千真萬確。只不過家長遺漏了做決定所需的資訊與數據，所以才看不到風險。

凡是遺漏的資訊與數據，我一概以「暗數據」（dark data）稱之。暗數據隱而不顯，單憑這點就可能導致誤解、錯誤結論及壞決定。簡單說，就是無知會讓人出錯。

暗數據一詞發想自物理學的暗物質（dark matter）。宇宙有二七％由這種神祕物質構成。由於它不跟光和電磁輻射作用，肉眼不可見，進而使得天文學家長年不知其存在。直到觀察星系旋轉，發現距離星系中心較遠的星體移動速度並不比距離較近的星體慢，違反我們對重力的理解，天文學家才察覺不對。於是，有人假設星系的總質量比望遠鏡觀察到的星體和其他物體的質量總和還大，這樣就能解釋星系旋轉的反常現象。由於我們看不見那多出來的質量，所以稱之為暗物質，而且這種物質可能分量（我差點就說「質量」）驚人：據估計，我們所在的銀河系擁有的暗物質是一般物質的十倍左右。

暗數據與暗物質很類似——我們見不到那些數據；那些數據沒有紀錄，卻會大大影響我們的推論、決定與行動。本書稍後將會舉例說明，除非我們察覺四周潛藏著未知的事物，否則後果可能不堪設想，甚至致命。

本書嘗試探討暗數據如何出現，以及為何出現。書中將檢視各種暗數據；瞭解這些數據的成因；說明哪些步驟可以避免暗數據出現，防範未然；介紹察覺自己被暗數據蒙蔽時該如何處置；最後指出只要夠聰明，有時還能利用暗數據，從中得益。雖然聽來奇怪又矛盾，但我們確實能夠利用無知和暗數據，思考做出更好的決定與行動。說得更具體一點，就是讓我們生活得更健康、賺更多錢，並明智運用未知來降低風險。這不代表我們應該對別人隱瞞資訊（雖然本書之後幾章會提到，刻意隱瞞的數據是常見的一種暗數據），實際作法比這複雜許多，而且所有人都會受益。

暗數據有各式各樣的形態，成因也五花八門，因此本書建立了一套分類法，以「DDTx」表示「X型暗數據」，並將暗數據分成十五種類型。然而，這套分類並不完全。暗數據的成因太多，可能永遠無法完全分類，而且某個暗數據實例可能同時展現不只一種暗數據的影響。不同型的暗數據可以聯手，甚至產生不幸的加乘效應。儘管如此，覺察這些暗數據類型，檢視暗數據生成的實例，還是能讓你在問題浮現時立即發現，免於受害。我在本章結尾列出了所有暗數據類型（DD-Tx），按相似度粗略排列，並且將在第十章詳加說明。書中有些例子，我會明白指出這是某一型暗數據，但我刻意避免每個例子都標明，以免妨礙閱讀。

正式開始之前，讓我再舉一個例子。

在醫學領域，創傷是一種重傷害，可能留下嚴重的長期後患，或可導致過早死亡與殘障，是「壽命減損」的最重大事由之一，也是四十歲以下人口最常見的死因。創傷審計與研究網路（TARN）擁有歐洲最大的醫學創傷資料庫，蒐集的創傷紀錄來自全歐兩百多所醫院，除了英格蘭和威爾斯九三%以上的醫院，還包括愛爾蘭、荷蘭和瑞士的各級醫院。不論研究創傷病例的預後或治療的有效性，這個網路顯然都是非常豐富的寶藏。

英國萊徹斯特大學的艾夫吉尼・莫克斯（Evgeny Mirkes）博士的研究團隊，檢視了創傷審計與研究網路的部分數據。他們研究十六萬五五五九個創傷病例，發現其中有一萬九二八九個病例結果不明。在創傷研究中，所謂「結果」是指病患受創三十天以後是否存活。因此，一一%的創傷病人三十天後是否存活，我們不得而知。這是很常見的一型暗數據——DD-T1：我們知道漏掉的數據。我們知道這些病人一定有結果，只是不曉得結果是什麼。

你可能會想，這有什麼問題？只要分析我們知道結果的那十四萬六二七〇位創傷病人，從中得出理解與預後就好。畢竟十四萬六二七〇是個大數字，至少醫學上如此，所以我們當然可以很有把握地說，從這些數據得出的結論是正確的。

可是，真的是這樣嗎？說不定少掉的那一萬九二八九人的數據，跟其餘病人很不一樣。畢竟他們顯然有一個不同點，就是結果不明，因此設想他們可能還有其他方面和其餘病人不同，也就不無道理。相較於納入全體創傷病人，只分析結果已知的十四萬六二七〇位病人可能會造成誤導，據此採取的作為也可能出錯，可能導致錯誤的預後、不正確的處方、不當的治療方案，對病人造成不幸甚至致命的後果。

讓我們暫時撇開現實，舉個極端的例子吧。假設結果已知的那十四萬六二七〇位病人，未受治療都存活下來並康復了，而結果不明的那一萬九二八九名病人都在就診後的兩天內死亡。這時要是忽略結果不明的病例，我們就會信誓旦旦地下結論說，不用擔心，所有創傷病人都會康復，面對新的創傷病人也都覺得他們自己會好，因而不進行任何治療，結果卻驚慌又困惑地發現怎麼會有一一%以上的病人性命垂危。

在往下說之前，我想先請讀者放心，我舉的極端例子是最嚴重的狀況，我們大可相信現實不會這麼糟，而且莫克斯博士和他同事是研究遺漏數據的專家。他們很清楚箇中危險，也一直努力研發統計方法來處理這類問題，本書稍後會介紹這些方法。但這個例子給我們的教訓是，事情可能不是外表看上去那樣。事實上，如果你要我用一句話總結這本書，我可能會用這句話。擁有大量數據是好事，也就是所謂的「大數據」，然而不是量多就好。要瞭解真實情況，我們不知道和不擁有的數據，可能比我們擁有的數據還重要。不論如何，我們之後就會明白暗數據的問題不只發生在大數據，小數據也躲不過。暗數據的問題無所不在。

我舉的TARN資料庫的例子可能很誇大，但很有警惕作用。那一萬九二八九位病人的結果沒有紀錄，可能恰恰因為他們都在三十天內過世了。畢竟如果結果是入院三十天後才測

量，過世者顯然沒辦法回答問題。除非我們意識到這個可能，否則永遠不會記錄到過世的病人。

這件事乍聽之下有點蠢，其實還滿常發生的。例如我們根據之前接受某項治療的病患的結果建立了一個模型，用來判斷新進病人的預後，決定他們是否要接受該項治療。但要是之前設定的時間對某些病患來說太短了，來不及出現結果呢？對於那些病患，我們其實並不曉得最終結果。如此一來，只建立在結果確定的病患上的模型便有可能造成誤導。

民調也有類似的狀況，「未回應」往往會造成問題。研究者通常會有一份名單，上頭是他們希望回答問題的人，但通常不是所有人都會作答。要是作答和不作答的人在某些方面有所不同，研究者就得擔心統計數據能否切實反映母群體的狀況。畢竟如果某家雜誌進行訂戶調查，只問訂戶一個問題：你有回覆本刊的調查嗎？我們也不能因為回覆調查的人答「有」的比例百分之百，從而推論所有訂戶都有回覆。

前面這些例子都是第一型暗數據。即使不是所有 TARN 病人的量測值都有記錄下來，我們確信他們都有數據。我們也知道所有接受民調的人心中都有答案，只是有些人沒有作答。我們通常知道數值一定在，只是不曉得是多少。

接下來是另一型暗數據（DD-T2：我們不知道漏掉的數據）的例子。

許多城市都有路面坑洞的問題。冬天水會滲進路面縫隙，然後結凍，將裂縫撐大，接著又被車子的輪胎不停碾過，形成惡性循環，最後弄出足以損壞輪胎或車軸的大洞來。美國波士頓市決定運用現代科技來解決這個問題。市府推出一款手機應用程式，使用手機裡頭的加速度感測器偵測車輛經過坑洞時的震動，再用 GPS 將坑洞位置傳回市府單位。

這招真是太帥了！這下高速公路養護工程大隊肯定知道上哪兒填補坑洞了。

這又是一個運用現代數據分析技術，輕鬆漂亮解決實務問題的好例子——只不過有車又有手機的人通常集中在收入較高的地區。因此，收入較低地區的路面坑洞可能不會被偵測到，坑洞位置也不會送出，某些區段的路面坑洞可能永遠不會補好。結果，這個方法非但沒有徹底解決問題，反而可能加劇了社會不平等。這個例子跟 TARN 的例子不同。TARN 的例子是我們知道數據有遺漏，這個例子我們則是不知道數據存在。

以下是這型暗數據的另一個案例。二〇一二年十月底，又名「超級珊蒂」的珊蒂颶風襲擊美國東岸，不僅造成美國史上第二慘重的颶風災情，也是自有紀錄以來最猛烈的大西洋颶風，財物損失估計高達七百五十億美元，共有八個國家兩百多人死亡。美國有二十四州受到影響，包括佛羅里達、緬因、密西根和威斯康辛，金融市場也因為停電而關閉。這場颶風還間接造成九個多月後生育率突然飆升。

除此之外，現代媒體也在這場颶風中大獲全勝。珊蒂颶風所到之處，推特也颳起一場訊息風暴，分享即時現況。推特的功用就是在第一時間告訴你哪裡發生了什麼事，還有發生在誰身上。這是個讓人即時掌握事情動態的社群媒體平台，而珊蒂颶風來襲期間正是如此。二〇一二年十月廿七日至十一月一日，推特上出現了兩千萬則颶風的相關貼文。於是我們可能會想，這些貼文應該可以讓我們持續掌握颶風的發展，找出哪些地區受創最重，哪裡需要緊急救援吧？

然而，事後分析顯示，珊蒂颶風相關推文最多來自曼哈頓，只有少數推文來自洛克威海灘或康尼島等地。這表示洛克威海灘和康尼島受創較不嚴重嗎？的確，曼哈頓區的地鐵和街道都淹水了，但很難說是受創最重的地區，就算只論紐約亦然。想也知道，實情是推文較少的地區之所以如此，不是因為受到颶風衝擊較小，純粹是因為那裡的推特用戶較少，比較少人有手機可以貼文。

其實，同樣的狀況我們可以推到極端。假設有個地方被珊蒂颶風徹底摧毀，那個地方就不會有推文出現，結果可能讓人以為那裡一切無恙。這可真是暗數據，黑暗得很。

和第一型暗數據一樣，第二型暗數據（我們不知道有所遺漏的數據）也是無所不在，只要想想沒被查到的詐騙案或查無凶殺案的被害者訪查報告，就會明白我的意思。

對於前兩型暗數據，你可能覺得似曾相識。前美國國防部長朗斯菲德（DonaldRumsfeld）在那場名震全球的記者會上，曾經一語道破箇中奧妙：「這世上有已知的未知，也就是有些事我們知道自己不知道。但這世上還有未知的未知，也就是有些事我們不知道自己不知道。」朗斯菲德因為這句晦澀的發言而遭到大量媒體奚落，但那些批評並不公道。朗斯菲德說的不僅有道理，而且完全正確。

不過，這兩型暗數據只是開胃菜而已。下一節我們會再介紹幾型暗數據。這些和之後談到的暗數據，就是本書的全部內容。你將會明白，暗數據類型千變萬化。除非我們察覺到數據可能不完全，觀察到東西不代表觀察到全部，測量可能不準確，測量到的可能不是我們想測量的東西，否則很可能對事實狀況產生偏頗的認知。只因為沒有人在森林裡聽見樹倒了，不代表樹沒發出聲音。

你以為你有的數據就是全部？

顧客推著裝滿商品的推車來到超市結帳櫃台，掃描器逐一掃過商品條碼，收銀機一邊發出電子嗶聲，一邊加總金額，最後顧客拿到帳單，然後付帳——只不過這不是最後的結果：顧客購買的各樣商品及價格都會送到數據庫儲存起來。之後，統計學家和資料科學家會鑽研這些數據，包括顧客買了哪些商品、哪件和哪件商品一起購買，以及購買這些商品是哪類顧客，從中掌握顧客的行為樣態。這樣做肯定沒有數據遺漏了吧？超市必須掌握交易數據，才知道要收顧客多少錢，除非遇到停電、收銀機故障或有人詐騙。

感覺上，收銀機蒐集到的數據顯然就是我們能蒐集到的所有數據了。它蒐集到的不是部分交易或部分商品的資料，而是超市裡所有顧客購買的所有商品、進行的所有交易的紀錄，就像有些人說的，資料＝全部（data=all）。

然而，真是這樣嗎？畢竟這些數據描述的是上週或上個月的事，雖然有用，但要管好一家超市，我們真正想知道的或許是明天、下週或下個月會發生什麼事：哪些商品不快點補貨就可能讓顧客買不到？哪些牌子更受顧客青睞？我們想知道的是還沒被測量到的數據。第七型暗數據（DD-T7：隨時間而異）就是在講時間讓數據變得隱晦的特性。

其實，撇開這點麻煩不談，我們可能想知道，要是換成其他商品、換個陳列方式或開店時間，顧客會有什麼反應？這些叫作反事實疑問，因為它們和事實相反，討論事實上沒發生的事要是發生了會如何。反事實是第六型暗數據DD-T6：可能會如何。

想也知道，不是只有超市經理會在意反事實問題。我們都服過藥。你信任開藥給你的醫師，同時認為那些藥經過檢驗，能夠有效緩解症狀。但要是你發現那些藥其實未經檢驗，藥廠並未蒐集那些藥是否有效的數據，甚至吃了其實會讓症狀更嚴重，你會有什麼感覺？或者那些藥確實經過檢驗，也證實有效，但沒有跟「什麼都不做」比較，看是吃藥還是自然痊癒比較快好，你會怎麼想？又或者那些藥並沒有和其他藥物比較過，看它是否更有效，你又會作何感想呢？在大象粉的例子裡，只要一拿「什麼也不做」來比，就會發現當你什麼也不做，驅離大象的效果跟撒粉一樣好，進而察覺根本沒有大象需要趕跑。

回到「資料＝全部」這個概念。覺得我們可以擁有「全部」數據，這個想法許多時候顯然是無稽之談。就拿你的體重來說吧。你的體重很好量，只要站到體重計上就好。但只要量第二次，就算和第一次時間相隔很短，你也可能得到稍微不同的結果，尤其量到盎司或公克的話。所有物理測量都可能不精確，因為可能有量測誤差或環境細微變動造成的隨

機紊變（DD-T10：量測誤差與不確定）。為了克服這個問題，研究人員測量某個現象（例如光速或電子的帶電量）的值都會重複測量數次，然後取平均值。他們可能會記錄十次或一百次的測量值，但顯然不可能記錄「全部」的次數。這種情況下沒有「全部」可言。

以上資料取自博客來網路書店