11 個不為人知的資料管理秘密

可靠的資料管理策略可以為任何尋求利用資料價值的企業帶來回報。然而，資料驅動決策這條路仍然充滿挑戰和難題。

文／Peter Wayner　譯／雲翻譯

有些人稱資料為新石油。有些人則稱它為新黃金。哲學家和經濟學家可能會爭辯比喻是否恰當，但毫無疑問地，對任何希望達成資料驅動決策承諾的企業而言，資料組織分析都是一項重大工程。

為達到成效，關鍵在於可靠的資料管理策略：包括資料整治、資料操作、資料倉儲、資料工程、資料科學等，資料管理如果做得好，可以為各行各業帶來競爭優勢。

幸好，資料管理的許多方面已廣為人知，且數十年來的理論發展已十分健全。舉例而言，資料管理或許難以應用或理解，但多虧實驗科學家和數學家的共同努力，如今企業能夠廣泛採用資料演算框架，分析資料並得出結論。更重要的是，統計模型還可以繪製誤差線來界定分析範圍。

然而，儘管資料科學相關領域的研究成果帶來不少好處，有時資料仍然令人摸不著頭緒。企業經常遇到該領域的限制。有些悖論與收集大量資料並加以組織化的實務挑戰有關，有些是哲學面的問題。這些悖論在在考驗著企業的抽象推理能力。此外還有日益加深的隱私權顧慮，在進行大量資料收集之初就必須加以考量。

以下是資料管理不為人知的秘密，讓許多企業組織頭痛不已。

一、難以分析的非結構化資料

許多企業歸檔內儲存的資料毫無結構可言。我有個銀行業的朋友希望用人工智慧來搜尋電話客服人員的文字記錄，他認為記錄也許深藏洞見，有助於改善銀行的借貸服務品質。但這些記錄分別由好幾百個不同的人撰寫，他們接聽的電話內容不同，心中的想法也不同。除此之外，同仁們的寫作風格和能力也有所落差。有些人沒寫什麼東西，有些人則寫下太多資訊。文字本身大體而言不具結構，尤其這些文字堆積多年，還分別由成百上千名員工先後寫成，就算有結構，也變得微不足道。

二、結構化資料經常缺乏結構

優秀的科學家及資料庫管理員建立資料庫時，會明確分類欄位的結構。為了使資料更加結構化，有時他們甚至會限制特定欄位只能用整數或預設選項。即便如此，填寫表格的人還是有辦法寫得亂七八糟，破壞資料庫儲存的秩序：有些欄位空著沒填；或是認為問題不適用，就在欄位裡填了橫線或「N.A.」字樣。還有明明是同一個人，姓名卻每年、每天不一樣，甚至隔了一行寫法就換了。優秀的開發人員可以透過驗證發現部分問題。優秀的資料科學家也能透過資料清洗減少不確定性。但煩人的是，即使是最為結構化的表格也會有可疑條目 — 這些可疑條目可能會在分析中導入未知數，甚至造成錯誤。

三、資料庫綱要不是太嚴格就是太寬鬆

無論資料團隊如何努力闡述綱要限制，定義各種欄位數值的綱要成果不是太嚴格，就是太寬鬆。如果資料團隊限制過於嚴格，用戶便會抱怨無法在狹隘的可接受值列表中找到答案。如果綱要過於寬鬆，用戶便可能新增前後不一致的奇怪數值。想微調出合適的綱要幾乎不可能。

四、資料法規過於嚴格

隱私及資料保護的法規日益嚴峻。為符合 GDPR、HIPPA 等多種法規，資料收集可能非常困難，但任其閒置，曝露在駭客入侵的風險下更加危險。在許多情況下，企業寧可花更多的錢請律師，而不是請程式設計師或資料科學家。這些令人頭疼的問題就是某些公司會在資料處理完畢後立即銷毀的原因。

五、資料清洗的成本高昂

許多資料科學家坦承，90% 的工作在於收集資料、將其以一致的方式呈現，並處理層出不窮的漏洞或錯誤。資料持有人總是會說，「資料都在 CSV 檔案裡，隨時可以使用。」但他們沒有提到如何處理空白欄位或錯誤描述。在資料科學專案中，進行資料清洗備用相當沒效率，往往要花費多達十倍的時間，還不如利用 R 語言或 Python 重新啟始排程、實際執行統計分析。

六、用戶對於企業使用資料的疑慮日益加深

終端用戶和客戶對於企業資料管理使用的疑慮日益加深，運用人工智慧演算法往往只會加劇恐懼，許多人認為一舉一動遭到資料監控而感到非常不安。 這些擔憂助長了法規監管，也損傷了一般企業，甚至是立意良善資料科學家的公關印象。不僅如此，人們還故意假造數值或錯誤答案

干擾資料收集。以致資料收集的工作泰半都在處理惡意合作對象和客戶。

七、整合外部資料雖然有些益處，卻也帶來災難

一般而言，企業會持有其收集的資料，並交由 IT 部門和資料科學家掌控，但有越來越多積極進取的公司學會了整合內部資訊、第三方資料，以及茫茫網路大海之中的客製化訊息。有些工具公開標榜它會鉅細靡遺地收集每一個客戶的購買資料，以建立客製化檔案。這種說法跟情報單位有什麼兩樣？追蹤快餐購買紀錄和信用評分，弄得像是要追擊恐怖份子。難怪人們會擔心和恐慌。

八、監管機構針對資料使用嚴格執法

沒有人知道，精巧的資料分析何時會越界，一旦問題發生，監管機構就會介入。加拿大最近有個例子，政府發現有的甜甜圈店會追蹤那些也去競爭對手店裡光顧的客人。最近一份新聞稿宣布：「調查發現快餐店 Tim Hortons 與一家美國第三方定位服務供應商的合約用語過於模糊和寬容，以至於允許該公司為自身營利目的出售『去識別化』的定位資料。」公司為什麼這麼做？為了賣更多甜甜圈嗎？監管機構如今已越來越關注個資議題。

九、資料方案性價比低

我們期望出色的演算法可以讓一切變得更加有效率、更有利可圖。有時這樣的演算法實際上是可行的，但價格也可能太高。例如，消費者 — 甚至公司 — 越來越質疑精心設計的目標市場行銷資料管理方案是否真有價值。有些人指出，我們經常買完某樣東西，卻還一直看到相同商品的廣告，因為廣告追蹤器還沒發現我們已經不是目標客群了。其他精妙的資料方案也有同樣的命運。有時候，嚴格的資料分析會找出表現最差的工廠，但這並不重要，因為公司已經簽了為期 30 年的工廠租約。公司需要做好準備，就算資料科學再怎麼神機妙算，最終得出的答案也不一定符合預期。

十、所謂的資料決策不過是人為的主觀判斷。

數字可以提供足夠的精準度，但人類如何解釋資料才是最重要的。資料分析和人工智慧再強大，多數演算法還是需要人為決策數值的門檻高低。有時科學家希望 p 值低於 0.05。有時候，警察會針對超速 20% 的汽車開罰單。這些門檻通常只是隨機值。對於所有資料應用科學和數學而言，許多「資料驅動」流程中的灰色地帶比我們發現的要多，儘管公司可能傾注資源投入資料管理實踐，但是到頭來，決策依舊是取決於直覺。

十一、資料儲存成本爆炸性增長

硬碟轉速確實越來越快，每 TB 成本也在降價，但是程式設計師收集資料的速度還是比降價速度快。物聯網（IoT）裝置不斷上傳資料，用戶也希望可以永遠瀏覽這些豐富的儲存庫。與此同時，審計人員和監管機構也不斷要求提供更多的資料，以備將來審計之需。

如果真的有人會檢視資料內容也就算了，但是一天的時間不過才 24 小時，人們實際上會重複存取的資料越來越少，擴充儲存空間的成本卻依舊扶搖直上，造成另一項隱憂。