



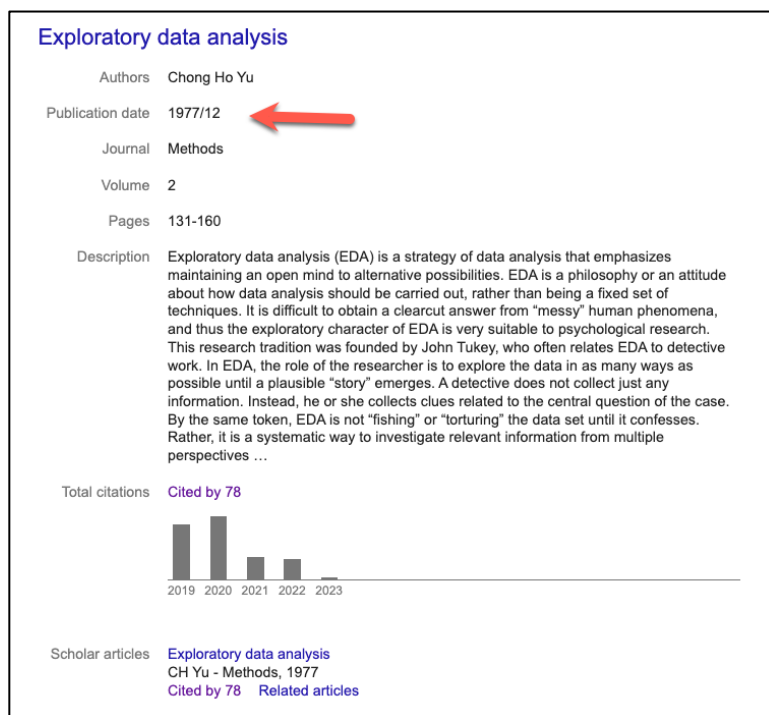
- 余創豪 chonghoyu@gmail.com

筆者要隆重地向各位宣佈：本人在數據科學的範疇裏面是首屈一指的學者，我在 1977 年發表的學術著作，直到今天仍然受到很多刊登在權威學術期刊的論文引用。熟悉我的朋友會馬上拆穿我的西洋鏡：「你今年幾多歲？你有可能在 46 年前發表學術著作嗎？」其實這是一個令人啼笑皆非的誤會，以下我會交代是什麼一回事。

搜索引擎龍頭公司谷歌不單止收集網站資料，而且亦搜羅學術著作，谷歌使用稱為「機器人」（robots）或「爬蟲」（crawlers）的自動化軟件來收集學術資料，並且將這些文章和書籍與原作者聯繫起來，更厲害的是，這些著作在什麼地方被徵引過，也進入了資料庫，谷歌根據作品數量和被引用的次數為作者評分，這個分數名為 H-index，H-index 是在 2005 年由哈書（J. E. Jirsch）教授在美國國家科學院提出來的，谷歌和很多其他的學術資料庫都採用了這評分標準。要建立學術成就的學者便需要變成「火麒麟」：周身引。

每當谷歌搜索引擎發現有些新作品的作者姓名跟數據庫的作者相似時，便會向該作者發出查詢：「這是你的文章嗎？」前一陣子谷歌問我是否撰寫某篇在 1977 年發表的論文？起初的我自然反應是：「沒有可能」，但細心看一看題目，我的確曾經寫過相同題目的論文，文章的題目是：〈探索性資料分析〉（exploratory data analysis）。我在 2017 年曾經在牛津大學的網站上用這個題目發表了一篇文章，在今年初又發表了更新版，但出版年份絕對不是 1977 年。

上面提過，谷歌將原本的文章和引用這篇文章的其他著作連結起來，直到目前為止，有七十八位學者引用我那一篇「1977 年的著作」，我翻閱注釋和書目，發現我的名字是對的，撮要也是對的，其實他們是參考了我在 2017 年出版的文章。1977 年，數據分析大師約翰·圖基（John Tukey）出版了一本同樣名稱的經典巨著，這本劃時代的奇書是現代數據科學的先鋒，我的文章平平無奇，根本無法與圖基的相提並



論。我相信其中一位作者將我和約翰圖基的著作混淆了，一位作者搞錯是不足為奇的，但為什麼其餘七十七人會犯同一錯誤呢？我相信聰明的讀者已經猜到答案，這是引上引、抄上抄的連鎖反應，其他作者都沒有打開牛津大學網站查證，更遑論細心閱讀我的文章，大有可能他們只是直接將引文條目搬到自己的文章。

2021 年，以帕夫洛維奇（Vedrana Pavlovic）為首的一個生物醫學研究團隊發表了一篇研究報告，審查過 1540 篇論文之後，他們發現大概 15% 的文章至少有一處引用的東西不正確，最常見的問題是作者引述的所謂科學發現根本並不存在（38.4%），其次是錯誤地解釋引用的東西（15.4%），五分之一的錯誤引用是源自「不準確的引用鏈」（chains of inaccurate citations），即是甲徵引時出錯，乙採用甲的徵引，丙把乙的東西搬至過紙，丁重蹈覆轍……，結果造成連環錯誤。

在其他文章中，我曾經質疑一些指出中國在科研上已經超越美國的報告，這些報告提供的證據是，中國學者出版論文的數量和論文被引用的次數都已經凌駕於美國，其實，論文引用的次數很多，並不一定表示該篇論文的質素高。很多時候，作者為求方便快捷，只是引上引、抄上抄，根本沒有翻開過原文！

美國統計學家泰力布（Nassim Nicholas Taleb）指出，人們很容易被運氣和隨機發生的東西愚弄，他曾經舉出這個看似荒謬、卻十分真實的例子：有位科學家在撰寫論文時需要為某個論據加入注腳和參考書目，他只是隨意選擇了三篇文章，學術期刊出版了他的論

文之後，那三篇論文被引上引、抄上抄，突然之間，那三篇論文的作者變成大紅人、天之驕子！

對泰力布的說法，我只同意一部份，的確，文章被引用的次數未必和文章的質素有關，但這現象並不全然是隨機的，當中有跡可尋。2019 年，瑞士學者埃爾根迪

（Mohammed Elgendi）分析過為什麼有些文章被引用的次數很高，有些則很低，他發現人氣極高的文章有幾個特徵，其中一個是題目很短，通常只有七至十三個字，這是很容易解釋的，通常人們在圖書館目錄找出了一大堆文章之後，會先看看題目，然後再決定是否看撮要及詳細內容，水蛇春長度的題目會令讀者望而卻步，簡單而容易明瞭的會受到青睞。

請讀者不要誤會，我絕對無意抹黑著作等身、作品廣泛地被引用的作者，我亦不是說 H-index 完全沒有價值，但我認為這一切只可以作為參考，而不是絕對標準。但願我們能夠擺脫「火麒麟」的迷思，不再追求周身引。

2023 年 5 月 16 日

原載於澳洲《同路人》雜誌

[更多資訊](#)