資料科學的五大迷思

2015/08/06 讚 223 分享

這一切大概開始於我搬到紐約後的 2012 年。當時，「資料科學」一詞還沒人用，大數據是大數據、機器學習是機器學習（其實廣義來講也算統計學的資料探勘），而人工智慧也只稱作人工智慧，這些都是截然不同的專業領域。大數據，專注於建置可儲存和處理大量資料的系統平台；機器學習專注於設計用資料趨勢來分類與預測的模型；而人工智慧，廣義而論是在設計和模擬人類般的知識管理和決策能力。這些領域之間或許相輔相成，但並不被視為同一專業。

說穿了，所有的科學都是資料科學。世界上有甚麼科學不用資料的嗎？

不料幾年後，資料科學一詞充斥整個業界。根據 維基百科 ，資料科學是「從大量的結構性與非結構性資料中萃取知識，實為資料探勘的延伸，另稱知識發現與資料探勘（Knowledge discovery and Data Mining）」。而「知識發現」與「資料探勘」等用詞其實在學術界已經使用數十年，並不是甚麼新領域。

同時我們透過美國知名的實務訓練學校 General Assembly 或是知名線上教學網站 Coursera 來了解資料科學的範疇，不難看出這些資料科學課程的課綱摻雜了大數據和機器學習。但說到機器學習也只是點到為止，這些課程只討論基本的統計迴歸分析（Regressions）、決策樹（Decision Trees）以及樸素貝斯分類法（Naive Bayes classifier），這些只能算是機器學習的紮馬步。

大數據和機器學習，無可置疑地，都是相當有價值且應用性極高的專業技能，但「資料科學」這模糊的名詞使得很多產業迷思難以破除。在思考資料科學時，我們要了解資料科學並非科技媒體鼓吹的萬靈丹，更要了解大數據、資料探勘與人工智慧之間在目標、用途和訓練需求上不同，才能夠利其器以善其事。

今天，讓我們來談談資料科學的一些產業迷思。

- **一、資料科學是門新學問**

資料科學應該算是所謂的噱頭詞（Buzzword）。一門新學問、新領域應該要創造新的知識、新實作方法或是新應用面。然而資料科學與「雲端運算」、「物聯網」和「使用者經驗」一樣，並沒有改變相關學術與產業的知識、方法或應用面，而只是把特定的產品、技能和族群納入一模糊的概念之中。

噱頭詞的用途在於炒作特定的產品或技術，這本身並非甚麼壞事。但一旦大家把噱頭詞當作專業領域討論，我們很容易把自己限制於特定做事的方法，而在沒有足夠的專業知識的情況下，這些

方法常常被誤用。打個比方，迷信雲端使我們相信把東西丟到伺服器上就一定比單機運算有效；迷信物聯網，使我們創造太多不必要的內嵌式系統裝置，而低估了一般電腦與平板的運算能力；迷信使用者經驗，常常使人懶惰，在不下功夫研習設計、心理學和產業知識前就草草把樣品丟給使用者測試，而浪費很多不必要的時間。

而迷信資料科學會有甚麼後果呢？其後果就是讓我們不知道甚麼時候該停止使用大數據和機器學習。大數據和機器學習是不同的專業領域，他們不需要被綁在一起，他們也不適用於許多軟體解決方案。我們對資料科學的態度不應該是想用就用。

- **二、資料科學會計算出最好的結果**

過去一兩年創業界有個現象，那就是新創公司很喜歡說：「自己的產品能用機器學習去解決問題。」

不，機器學習並不是一個解決方案，他只是一種解決問題的方法。而當新創公司沒有大量的資料時，機器學習是個非常糟糕的選擇。

我相信你一定有用過蘋果的 Siri、Google Now 或是微軟手機和 Xbox 的 Cortana 吧？這些平台的語音辨識系統至少吃下了數以 TB 計的英語語音資料，而即便如此，語音辨識仍然錯誤百出。

我們應該以此為鑑，在考慮使用機器學習的時候就要想到機器學習產生的模型就算有大量資料常常都還有很大進步的空間。

對此，我們應該謹記兩點：第一點較哲學，那就是機器學習的用意在於模擬人類的分類和預測能力，如果你的團隊本身就沒有相關產業的工作經驗，機器學習絕對不會幫你找到答案。第二點則是應用面問題，那就是機器學習不管在甚麼情況下都不會取代產品設計。機器學習不是魔術，你自己都不知道使用者的問題如何解決的時，資料科學是不會跑出結果的。

- **三、資料科學會幫我改善產品**

上段我們討論到資料科學並非魔術，而跟多數科學與藝術一樣，資料科學的「眼界」受限於你使用的資料和方法。正如你今天帶著單色眼鏡到處跑，你絕對不會看到「藍色」；資料科學也一樣，能找到甚麼取決於你設計的系統帶上甚麼樣的眼鏡。

在資料科學（或應該說是機器學習）中，這種限制可稱為 **監督（supervision）**。簡單地說，監督式學習（supervised learning）就是尋找特定形式的答案。比如說在使用迴歸分析，你在找能解釋資料趨勢的方程式，而在使用樹狀決策時，你希望能夠產生一樹狀圖來形容決策的許多可能路線。其他的限制可能是 **變數選擇（variable selection）**，也有可能是單純的 **取樣偏差（sample collection bias）**。

打個簡單的比方，若你今天去蒐集許多人的手機型號和身高的資料，想利用這些資料去找線性模型來用手機型號來預測身高，這種作法不論在方法、取樣還是在選擇變數都有很大的問題，你可能窮忙老半天也找不到像樣的模型。

你或許會覺得我舉的例子很蠢，是很蠢沒錯，但是這跟新創公司使用資料科學的方式其實相差不遠。請記得你的資料科學成效取決於你團隊的產業經驗。若你的團隊對產業的趨勢和動向沒有概念，那你們將很難找到合適的資料、合理的機器學習方法，那其實用資料科學反而是在浪費你們的時間。

所以說，資料科學並不能取代專業知識。

- **四、使用大數據的解決方案比較好**

話不能這麼說。

在商業上，一解決方案的好壞與否取決於其解決問題的全面性和效率。而說到商業問題，大數據只是在建置軟體平台的方法上做改進，並不會直接影響商業問題的解決方式。

大數據恰如其名，是一系列能將大量資料分散儲存並平行處理的技術，因傳統的單伺服器資料儲存和處理技術已無法滿足高容量和高流量的大型網路公司哥吉拉級資料量。現在的大型網站已經從過去總共幾 TB 的資料成長至每日至少數 TB 的資料。唯有將資料量和處理時間分配給數個伺服器節點（甚至上百、上千個伺服器），才能夠滿足資料處理需求。

因此，就算非常乏味的資料處理工作（如計算平均數）若資料量大到 PB（約一千 TB）等級那都可能需要用到大數據。相對地，一複雜的線性代數運算若資料量不到 TB，不用大數據也能輕鬆解決。
所以說，並沒有用大數據建置的解決方案就是好方案的說法。

事實上，大數據技術如 Hadoop 因為要管理多伺服器節點並將資料從記憶體移動至資料庫，常常會需要上至幾分鐘的時間才能開始運算。除非你的資料量已經大到可以容忍這些大數據平台的啟動延遲，不然使用大數據還可能會比一般的資料處理方案更慢，而且更昂貴。

- **五、三個月內速成資料科學家**

如前所述，資料科學的概念相當模糊，實際上其涉獵的領域是多個截然不同的專業領域，各個領域都需要長時間的訓練和產業知識方能培育。

不管是大數據還是機器學習，在語音辨識、自然語言、DNA 排序、因果網路、航線最佳化等不同產業的應用上都有不同的資料儲存需求、運算時間配置需求以及平行化演算法設計需求。

過去，資料庫管理、資訊科學、資訊工程、統計學等都需針對特定產業的需求進行專業化，資料科學亦然。今天，一位只懂得 K-means、 Logistic 迴歸分析法的資料科學家，並不比一位統計學新生在產業界中更具有競爭力。

一位好統計學家、資料庫管理員、資訊科學家所需的訓練和經驗，過去三個月培育不出的，不會因為噱頭詞改了突然生出一堆資料科學家。

- **小結**

本文的用意不在於批評資料科學的實用性，資料科學涉獵的專業在今天的科技業非常重要。我本身是資訊科學出身，我的工作幾乎每天都會碰到機器學習和資料儲存問題。但是，在許多情況

下，一個設計得當的演算法的實用性和效率遠超過用資料訓練出來的模型。

我了解很多人認為叫機器用資料去搞出解決方案聽起來 **很酷** ，但是這種作法其實是脫褲子放屁，因為若問題本身就沒有未知數，那機率早就變成純邏輯了。換句話說，如果你已經知道如何解決問題，那為什麼還要強迫用資料科學去猜呢？

最後，我們應該謹記：資料科學的目標其實就是在猜，也就是在模擬決策模型。而資料科學的效益絕對取決於團隊的產業專業與經驗，才能為資料科學決定合宜的目標、變數以及資料採集方法。

對於資料科學的主題，如果你有什麼看法或問題，歡迎在下面留言交流！

Photo Credit: Wikipedia

（本文轉載自合作夥伴《alphacamp》；未經授權，不得轉載）

點關鍵字看更多相關文章：

Big data　資料科學家

0則回應

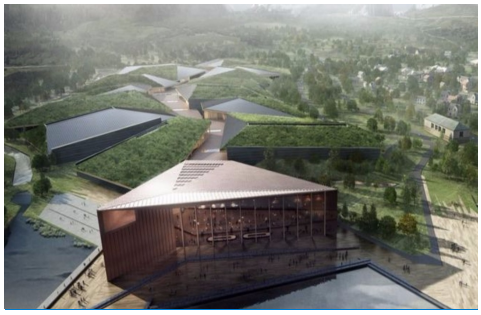
排序依據 最舊

新增回應……

Facebook 留言外掛程式

Google 統計結果大出爐：全世界都在問「怎麼打領帶」，猜猜看台灣人最愛問什麼？

大數據讓人類活得更久：預測糖尿病與病毒突變模式、保護新生兒它通通做得到！



全球最大數據中心落腳北極圈！100 % 再生能發電，成本全歐最便宜

「大數據」滾出黃金：美國教授用 Twitter 推文來預測股票動向，精確度高達 86%！

8.4 億用戶大數據首次開放，微信指數助掌握火熱議題

SHARE 91

【年薪 330 萬台幣好誘人】美國企業卯起來培育資料科學人才，那台灣呢？

SHARE 112

資料科學家、資料工程師和軟體工程師差在哪？這張圖讓你秒懂

SHARE 756

淺談 21 世紀初的科學：這世紀的科學因科技大放異彩，但短期內不會有重大突破

SHARE 59

【台灣最美資料科學家】專訪林郁珊： 美國資工學生不只在意分數，更會思考如何學以致用

SHARE 5.7 K

營業額增加 25%！呼叫台灣團隊 SkyREC 幫你「算出」商品怎麼擺最賣

SHARE 39