21 世紀最性感工作——「資料科學家」的八種技能

資料來源:https://www.inside.com.tw/2015/03/27/8-skills-you-need-to-be-a-data-scientist

- 隨著大數據滲透進各行各業，負責淘洗數據、從中精煉價值的資料科學家無疑是這幾年最炙手可熱的職位，《哈佛商業評論》將之譽為「21 世紀最性感工作」[1]，因為優異的資料科學家就像獨角獸一樣珍貴難尋，而且可不是只有科技公司在搶人，傳統金融界、零售商、廣告、教育，幾乎所有產業都需要資料科學家從大量數據中萃取精華。根據去年七月 Indeed.com 的調查，**美國資料科學家每年均薪 12.3 萬美金**[2]，比起整體均薪多出 113%——當然，還是比每年平均可以領 74 萬美金的 CEO 還少，但也夠讓 99.99% 的上班族望塵莫及。

- 頂尖的資料科學家最好統計、數學、程式能力最好都要掌握，而且要能從中洞察意義，並且擁有非凡的直覺，用數據資料發聲，幫助公司制定重大決策。但是，其實就算同樣都是尋找「資料科學家」，Google 跟沃爾瑪超市要的人才，可能非常不一樣。別因你好像缺了哪個專長而打退堂鼓，如果仔細閱讀每家公司張貼的職缺敘述，你會發現說不定現有的技能就能進入資料科學的殿堂。Airbnb 資料科學家 Dave Holtz 把市場上所需的資料科學家概括成以下四類[3]：

菜鳥資料科學家說穿了就是資料分析師

- 有些公司需要的資料科學家，說白話就是資料分析師（data analyst），而資料分析師就是菜鳥資料科學家。你的工作包括從 MySQL 萃取數據或是一名 Excel 專家，也許要能繪製基礎的數據視覺圖表、分析 A/B 測試的結果或者管理公司的 Google Analytics 帳號。這種公司對抱負遠大的資料科學家來說，是很不錯的練功場所，當你變成老手了，也能開始嘗試新事物，擴充技能組合。

來清理我們亂糟糟的數據！

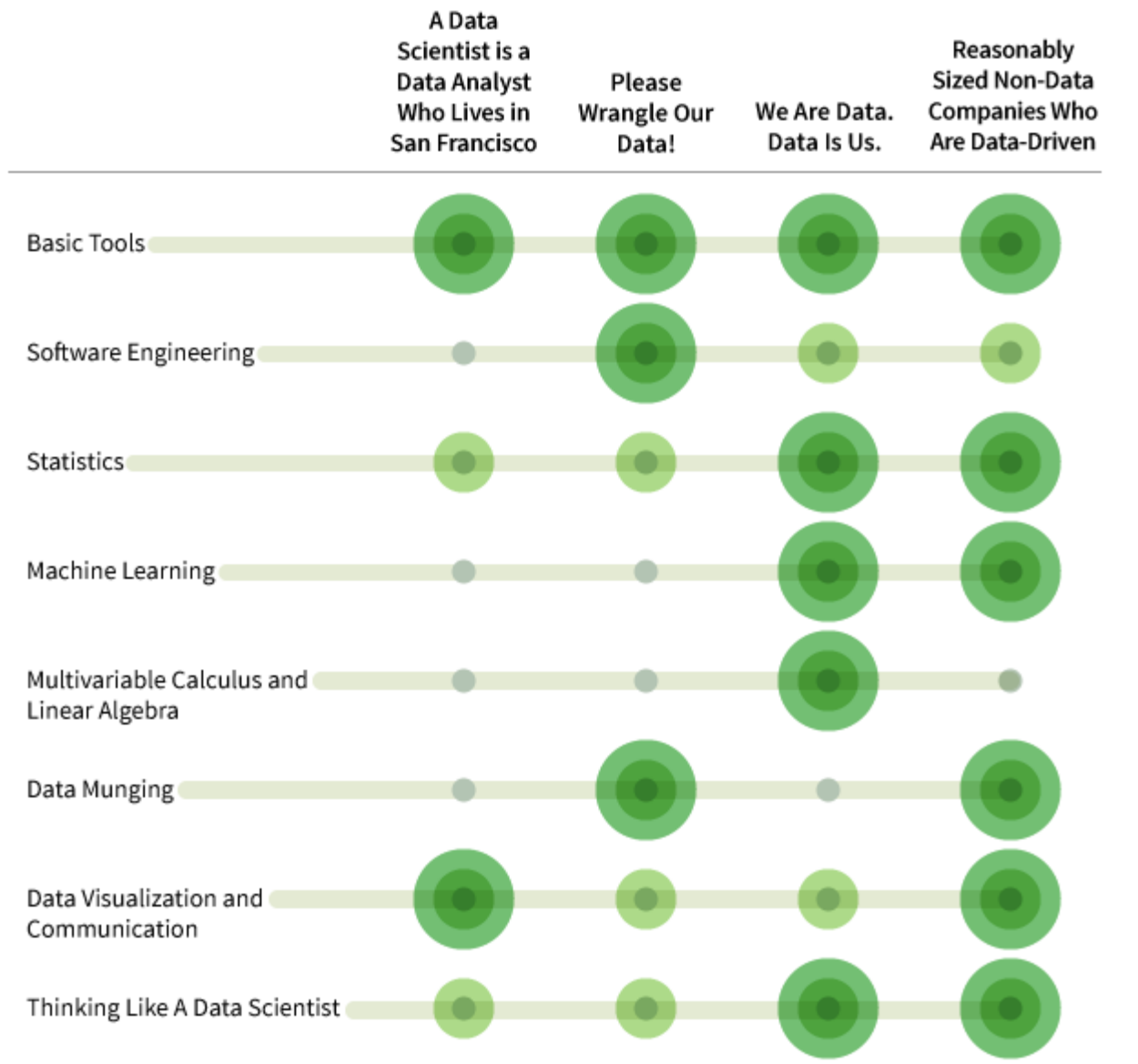
- 公司發展到了一定規模之後，累積一堆尚未理清的數據，而且持續大幅增加，因此他們會需要一個能夠建立資料基本設施（data infrastrucure）的人，以讓他們在這個基礎上繼續成長。由於你是第一個或第一批獲聘的資料相關人員，工作通常不會太難，不求統計學家或機器學習專家才能勝任。在這種公司裡面，帶有軟體工程背景的資料科學家就很吃香了，重點任務是提供數據到production code，關於數據的洞見與分析倒是其次。就像前面說的，你是這家公司的第一個數據探勘者，通常你不會獲得太多上層的支援，雖然反而更有機會大放異彩，不過因為比較缺乏真正的挑戰，也有可能面臨停滯不前的窘境。

我們就是資料，資料就是我們

- 也有很多公司，主要的產品就是數據（或數據分析平台）。如果你想進入這種公司，那你勢必要具備很高深的資料分析或機器學習功力。完美的人選應該是有正規的數學、統計、物理背景，而且有意繼續朝學術面鑽研。這些資料科學家的主要職責在於研發出色的資料產品，而非解答公司的營運問題。擁有大量消費者數據也以此作為主要營利來源的公司、或者提供基於數據的服務的公司，都歸屬此類。

產品並非數據、卻以數據驅動產品的公司

- 很多公司都屬這種類型。你可能會加入一組已經建立的資料科學家團隊，這家公司很重視數據，但稱不上一家數據公司。**你既要能夠進行資料分析、接觸 production code、也能將數據視覺化。**一般來說，這種公司要的人才要不是通才，就是他們團隊缺乏的某種特殊專才，比如資料視覺化或機器學習。想要通過這類公司的考驗，端看你對「大數據（比如 Hive 或 Pig）」工具的熟稔程度，以及過往處理雜亂無章數據的經驗。


A Data Scientist is a Data Analyst Who Lives in San Francisco
Please Wrangle Our Data!
We Are Data. Data Is Us.
Reasonably Sized Non-Data Companies Who Are Data-Driven
Basic Tools
Software Engineering
Statistics
Machine Learning
Multivariable Calculus and Linear Algebra
Data Munging
Data Visualization and Communication
Thinking Like A Data Scientist
Very important
Somewhat important
Not that important

基本工具（Basic Tools）

- 無論哪一類公司，統計程式語言如 R 或 Python，以及資料庫查詢工具像 SQL 大概都是資料科學家必備的常識。

基礎統計學（Basic Statistics）

- 對統計起碼要有基本認識，才稱得上及格的資料科學家，一名擁有許多面試經驗的人資說，很多他曾面試的人連 p-value 的定義都講得不清不楚。你應該熟悉統計測試、分佈、最大似然法則（maximum likelihood estimators）等等。機器學習也很重要，但更關鍵的能力，是你能否判斷不同狀況該用什麼不同的技術。統計學適用於所有類型的公司，但對那些主要產品並非數據、卻大幅依賴數據的公司來說尤為必備能力，老闆需要的是你能不能利用數據幫助他們進行決策，以及設計、評估實驗與結果。

機器學習（Machine Learning）

- 假如你是在握有大量資料的大型企業，或是產品本身就是以數據為賣點的公司工作，機器學習就是你用來吃飯的傢伙。雖然 KNN 演算法（k-nearest neighbors）、隨機森林（random forest）、集成學習（ensemble methods）這類機器學習的流行術語好像不懂不行，不過因為事實上很多技術都可以用 R、Python 程式庫解決，所以即使你不是演算法的世界頂尖專家，並不代表就毫無希望。比較重要的是，能夠縱觀全局，每種狀況出現都能找出最契合的技術。

多變量微積分、線性代數（Multivariable Calculus and Linear Algebra）

- 就算你即將面試的公司並未要求機器學習或統計學知識，基礎多變量微積分與線性代數問題十之八九都是逃避不了的必考題，因為資料科學就是由這些技術型塑而成。儘管很多事情可以交給 sklearn 或 R 自動執行，但是未來如果公司想要建立自有的方案，這些基本知識就變得很重要了。如果你置身於「數據就是產品」，或者預測績效僅因小小進步或演算法優化就能帶來驚人效益的公司裡面，微積分、線性代數等數學概念都需了解通透。

清理數據（Data Munging）

- Data Munging 是最容易令人不耐的過程，你面對的是亂七八糟的數據。這些數據包含消失的數值、不一致的字串格式（比如「New York」與「new york」與「ny」）、數據格式（「2015-03-26」、「03/26/2015」，「unix time」、「timestamps」等等），必須勞心費神梳理這些龐雜的數據。雖然這工作吃力不討好，但只要是資料科學家，大概都避免不了，而如果你是某家小公司的先遣資料科學家，或是在一家產品非與數據相關，但是數據卻扮演重要角色的公司裡工作，清理數據的任務格外重要。

資料視覺化與溝通（Data Visualization & Communication）

- 把枯燥繁瑣的數據轉成圖像，以及向外界溝通的技能愈來愈重要，尤其是在年輕的公司制定由數據驅動的決策，或者協助其他組織進行數據決策的公司。「溝通」二字的真諦在於，面對技術人或一般人，你都能準確的傳達研究發現，並能讓他們輕易理解。至於視覺化，如果可以熟悉 ggplot、d3.js 等軟體的運用，會有很大的助益，當然工具只是表象，能否參透資料視覺化的原則，才是最需費心的地方。

軟體工程（Software Engineering）

- 如果你是公司資料科學團隊的草創元老，擁有強悍的軟體工程背景十分重要，你會負責處理很多資料登錄（data logging），也有可能需要參與開發以數據為本的產品。

像個數據科學家般思考（Thinking Like A Data Scientist）

- 所謂資料科學家，就是你解決問題的方法奠基於數據資料。在面試過程中，主考官可能會出一些比較艱澀的問題，比如公司想要執行的某個測試，或者計劃開發的數據產品。判斷事情的輕重緩急、作為資料科學家如何與工程師和產品經理互動、知道該用什麼方式解決問題，都是你該培養的能力。