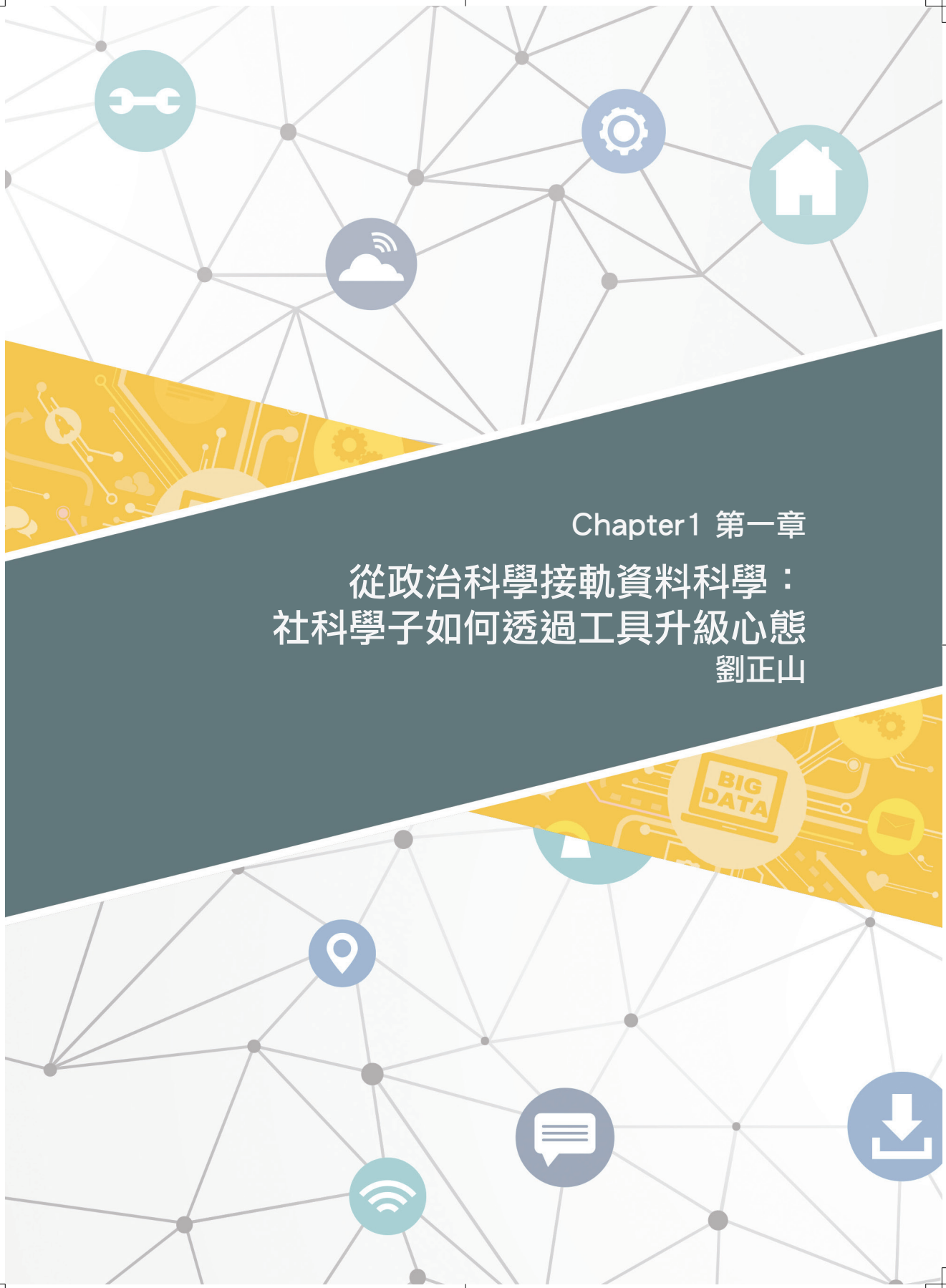

Chapter1 第一章

從政治科學接軌資料科學：社科學子如何透過工具升級心態

劉正山

各位尊敬的老師，還有各位長官各位同學大家早，我是劉正山，今天從中山大學今天專程回家。上次踏入復興崗已是20年前，在這邊受訓永遠不會忘記，所以今天回到家裡有很多的感觸，也還好我今天把全身的武藝都帶來了，也就是過往 20 年求學到這 10 年碰觸資料的很多心得，今天要用 1 小時左右的時間和大家分享。這裡是我歸零的地方，從大學畢業後接受預官訓練，這裡是我人生轉變的地方。歸零再出發，我看見了政治學界曾經的不足，我也看見了我們出國之後，歸零的好處。對每個人來說，歸零是一件非常辛苦的事情，學程式語言也是如此，不管做什麼事情，歸零是要放下一些我們習慣的東西，我們的習慣領域要解開，這是非常辛苦的事情，誰能幫助一個正在期待高飛的學子放下，那個地方就是他的寶地，所以，這裡就是我的寶地，所以非常非常興奮回到這裡來跟各位政戰同仁以及老師們分享在學界的一些觀察，在這裡又回到了預官的感覺了，向各位行一鞠躬，感謝各位為國家的付出。

從這裡我想要給大家的題目，是受莫大華老師之邀，就是我們如何從政治科學走向資料科學。大家可能注意到題目裡面並沒有大數據三個字，我會提一下為什麼。如何從我們的工具來轉變我們的心態，歸零這種心態。面對大數據大家只會想到我們又不是學程式語言的，我今天要碰觸這些數據、資料，應該覺得很辛苦。說實在的我們只有心態的問題，並沒有進入門檻的問題。2007 年開始有世界第一支手機，那時候大家在做什麼呢 ?2007 年才有耶！在座各位老師同學你有碰過手機對吧？可是 2007 年才有，連我現在才 3 歲的小朋友都可以自己滑了，可以自己滑手機這件事情表示這件事情對某些人沒有進入障礙。只有當心裡沒有障礙，就可以參與。我覺得我們現在最辛苦的地方在於，我們學會了一些功夫之後，我們要暫時放下學新的東西，但在學習新的東西開啟所謂大數據或是資料科學的時候，總是會徬徨，總是會覺得我們應該往哪裡去。那麼今天要跟大家說的就是如何從工具調整心態，這是一個路徑圖。我跟大家分享是一個知識地圖，或許大家可以從這邊看到我們政治學人在過去 10 年做方法論的時候看到什麼，這是對大數據一個簡單的定義。

我想大家都知道大數據是 5 個 V，就是量大數大，這個不需要特別強調，可是大家有沒有發現大數據這個事情最近愈來愈少人放在嘴邊，大家放在嘴邊的叫 AI 對吧？大家放在嘴上的是資料科學。所以在過去 5 年裡面，大數據的概念已經昇華，雖然我們可以講大數據，但是更多的人是講資料科學跟 AI，因為大數據已是 AI 的核心。至於真正的核心，到底是科學，還是講資料、數據、還是政治呢？政治科學也有碰到資料啊。我從這邊開始，先整理一下給大家聽，就是大數據這件事情在政治學過去幾年來大概在做哪些項目。很快的簡單講帶過之後，我們再來講有趣的個案。

基本上我們政治學者在做大數據的時候，就觸及到資料，都還沒有講到知識路徑圖這時候大家都跳進來用資料的模式做文本探勘、資料探勘、自動內容分析、電腦輔助文本分析、自動情緒分析、GIS 地理資訊系統、網絡的輿情分析，這幾個名詞念下去之後，大家可能都注意到這都是方法、都是工具。這些分析如果這些都是工具、都是方法，那大家必須要問的是它們求的是什麼？它們所謂的知識是什麼？在政治科學社群裡面，這些工具的運用可以分為幾個項目，第一個，已經有很多文獻在這裡，給大家看一下這些學術研究整體來講有 6 到 8 個面相是在做相關方面的研究，第一個叫做資訊工具應用概論，其實這主題已經快過時了，因為這主題的文獻就是在說大數據有多好，大數據可以幫我們產生什麼不同的問題意識等等。整體來講，我們的應用概論有很多是在講這個，但我們其實已經專注到不講概論了。最近三年講的其實是文字探勘可以帶給我們知道什麼，直接講專注的題目。

第二個主題我們會看到很多人在研究公共言論的趨勢，言論的趨勢這件事情從文字探勘來講是很好的題目，現在有很多研究都圍繞在公民到底在談什麼，例如我們現在有 10 個公投題目，那麼大家最常談什麼？公投題目大家有注意到有 10 個嗎？大家有特別注意到應該怎麼投嗎？輿論的變化與分析這是第二個應用大數據的題目與題材。其他的研究課題我們看一下：包含了政治立場的辨識和追蹤，大家聽過中間選民吧？大家知道中間選民會投誰嗎？你會說都已經是中間了怎麼知道他會投誰，他就是選人不選黨，可是從我自己做研究來看，不一定如此，電話打過去他會說自己是中間的，但是關起門來就開始罵了，有沒有這個情況？受訪所講的和真實情況不一定一樣，那你怎麼從文字感知到大家的偏好、政治立場的辨識和追蹤，這是有學問的，現在大數據正好可以切進來，因為你可能和電話調查公司說我中立，導致大約有 30-40% 的人都說他是中立的，所以民調公司永遠都算不準，因為有百分之 30 的人要被拿掉了所以無法分析。當你說一個候選人的支持度有 30% 的時候，記者也不知道這個數字代表什麼，因為有 40% 的人不表態。這種資料不完整的問題不是很嚴重嗎？也就是當媒體在說這個百分比的時候其實他們沒有完全的掌握真相。可是大家在粉絲頁留言就不一樣了，文字探勘就是讓我們回頭和朋友講話時候的文本被留下來，所以文字探勘有助於我們去瞭解政治立場，我們現在還有政治言論的管制策略以及公共政策形成的探討，這些題目也都是和輿情的走向有關。

我從高雄上來，目前高雄很常談的題目不是候選人而已，還包含了輕軌要不要蓋下去的問題，因為輕軌要經過的路太小條了，如果這條路輕軌蓋下去之後占用車道，到底是誰在歡喜呢？這樣的公共政策討論，其實用大數據或數據分析，其實是可以抓到一些內容的，但是我們的民調卻常常問「你支不支持在

大順路上蓋輕軌」，是贊成或是不贊成，民調有沒有辦法問更漂亮的題目？沒有！你支不支持死刑？電話只會這樣問，還能問出什麼名堂呢？所以會導致傳統的資料蒐集工具變得膚淺，不能問出好題目的結果，我們必須回歸到文本，也就是大家到底在日常生活中說些什麼，其實大家現在的拇指很厲害對吧？拇指比什麼指頭都厲害，很會打字，用手機就可以回傳一句話兩句話，所以這樣的文本對於大數據分析來講是重要的，不只是對台灣如此，微博上也是如此，所以未來的大數據分析，所有的 90 年後出生的這一代，他們擅長的不是在民調的回應，而是在文本的創造上，那這些文本創造之後，我們有沒有新一代的知識分子或者是研究者可以去抓取這些東西？這是我們一個大挑戰也是非常有趣的一個事情，所以我們學院全力在推這個方向是令人驚豔的。我們現在再往下看還有另外的其他主題。

其實這樣數下去，大家可以發現光是政治學領域大家就做不完啦，何況還有心理學、傳播學對不對？這些領域都是因為有資料的出現導致有新的議題我們可以做，像我們後來看到政治言論的語意分析、選舉預測的運用、國際關係研究，以前的國際關係研究是沒有很多 data 的，但是開始有 data 了。以往國際關係我們只能從理論上講，國際關係現在開始有資料可以分析了，在國際上面資料的取得變得容易。對於我們心戰或是政治作戰而言，雖然我不是在這個圈裡，但是我總覺得，資料的來源微博上面很多，就是我們言論的掌握上並不難，我的同事陳至潔教授已經完成了人民日報上面的分析，因為是公開的資料。他是我一個很好的同事，他做文本分析也有一些經驗，他發現了一個有趣的事情，中共北京在說人權的時候你覺得他是進步還是別有意圖？就我們的概念來講，講人權應該是種進步吧？原來北京也會講人權了？開始講究人權是不是一種走向民主化的前兆？答案不是！如果真的去爬文然後經過解讀之後會發現，人權一詞在那個文本有兩層意思，提升中共服務和效率就是人權，你能理解嗎？我們的人權指的是我們的個人自由要保障對嗎？可是在那塊土地上面，人權指的是國家或是政府把服務做好，提升你的效能感，這就是你的人權，這是從爬文，也就是從人民日報的分析之後，所看見分析出來的結果。意義對於大資料探勘是非常重要的一環，不是只有技術面而已，所以 why not big data? 我們為什麼要大數據卻也同時不一定要大數據？因為知識的論點很多，我們到底要求什麼知識？我們的知識之路到底以什麼作為目標？

我先跟大家列四個目標，大家可以想想看，這張圖不是很清楚，但是大家隱約可以看到四個大框框。我們很快的整理一下，這四個大框框是我們知識論上的四個境界，在做的學者彙整起來基本上都不出這四個框框，第一個叫做實證主義，第二個叫做實存主義，第三個叫做實用主義，第四個叫做詮釋主義。

我簡單講一下，或許大家就可以對號入座了，這四種我們叫四個大陸板塊，這四個大陸板塊上的人們，有完全不同的知識信仰，不同的知識信仰看大數據就有完全不同的使用方式，那麼你是不是應該先知道尤其當我們拿到 PHD，一定要先知道我們的知識立場或者我們所在的知識大陸是長什麼樣子，所以這四塊基本上涵蓋了目前所知道的百分之九十九的學者類型，第一個看一下實證主義者，也就是現在的大宗，實證主義者希望拿資料這件事情來做理論發展及因果推論。下雨會不會導致一個人不投票？這就是一個假設，所以天氣跟投票意願的關係，到底是有或沒有？很多人相信有，但是有沒有證據告訴我們不一定，或是真的有？這就是實證主義的發問方式，也就是求因跟果然後想要知道因和果是不是相關聯，你看看這是不是主流？我們大部分的訓練都在這裡。

第二個叫實存主義，就是一個中間選民，到最後如何變成有政黨傾向？怎麼變的？一個選區為什麼會翻盤？怎麼變的？「怎麼變的？」是實存主義者問問題的方法。大數據在這一塊不太使得上力，因為一件事情來龍去脈怎麼變成這個樣子其實是很有趣的問題，但在學界裡很不好做。大家可以想到大數據可以用在哪裡？在剛剛那個比較好用對吧？就是我可以找到因、找到果，兩個變數、兩個欄位之間，就可以找到關係了，大數據可以在這裡用，也是大數據目前使用在實證主義最火紅的時候。實存主義會講說這台車是怎麼發動的，我們的經濟如果上的來是怎麼上來的，這個變化太複雜了，大數據不見得能幫上忙。我自己在做的實存主義相關研究，是在回答一個問題：當所有人都不看電視新聞的時候，當所有人只看手機追劇的時候，這個社會是會更多元還是更極化？這很難回答，這個因果關係是連不上來的，怎麼變？所以我實存主義的研究事實上在做模擬的方式來理解，這個就更複雜了，因為複雜到大數據可能無法理解全貌，因為大數據的欄位，很多欄位，不一定能夠告訴我它怎麼變的。第二個叫實存主義，是我覺得很有趣，但一般大數據還使不上力的地方，這是第二塊大陸，科學實存主義。

第三塊大陸在最左手邊叫詮釋主義，他強調的是意義，等一下就有例子給大家看意義是什麼。幸福是什麼概念？幸福是不是收入所能代表的？意義這件事情跟大數據有什麼關係，只要我們追求意義看到大數據就會拼命問這個資料背後的意義是什麼，就像我剛剛舉的例子，人權的意義是不一樣的，大數據可以幫助我們理解人權不同的意義，所以陳至潔教授就是從左手邊，他是個國關的詮釋主義者，他曾經作匪情研究，他可以分析，可以知道很多意義，當他加上大數據之後那就不得了了，因為他可以看到數據背後的意思，所以如果各位不是數據出身，而你本身對於意義的解讀是很有概念的，那你應該比任何人都歡迎大數據，因為它讓你的詮釋有了新的基準、新的工具，當你能說出一些資

料背後的意義的時候，很多的大數據也就是工科或是電腦科學家，是沒辦法跟我們相比的，所以以詮釋主義者來講，有什麼好抗拒的？學習的過程本來就是有點痛，但是當你能夠挖角到自己的資料的時候，你又有說故事的能力，那你就無敵了，這就是目前看到大數據一個非常無窮的潛力，而這個球是落在人文社會科學這邊。

第三個叫實用主義，實用主義其實就是以解決問題為知識方向的領域，這個其實在政治學門一直被埋沒，今天還有沒有人講經世致用？我想應該很少，但是拿大數據做經世致用卻火紅，最近火紅的剛好就是拿大數據來解決社會問題，大家不知道有沒有注意到這個趨勢，不管是警政大數據，用在犯罪研究，或是用在交通的閘道控管，你只要是有助於解決民生問題，從不塞車到減少犯罪到垃圾桶要設多少到所謂的智慧城市，每一個都是屬於實用主義的範圍，而政治學界將這個埋沒了二十年，至少就我所知，在我們的學界裡面不太強調經世致用而因為大數據的出現，逼著我們政治學者開始得注意這個思維方式，那麼現在大家想想看這四個區塊的人如果坐在一起，大家可能不太能夠說出誰最有貢獻，因此大數據已經融入了一個新的概念叫 Data Science。

資料科學若只是大數據加上統計學，那麼我們所謂的政治學、心理學、社會學、傳播學是否在資料科學中就找不到定位了？目前一般說的資料科學是屬於工科的再進化，而不是屬於人文的再進化。但是未來的資料科學對我們人文的人來講有一個非常好的契機，就是我們也應該擁有資料科學內涵的話語權，因為我們會資料分析而且我們也會詮釋，我們不只會描述資料還會說故事，所以資料科學對我們來講是一個很好的機會。

政治學裡實證主義目前是大宗，大家可以看到很多文章。哈佛大學有位知名政治學教授叫Gary King，他所做的研究值得一談，Gary King老師不懂中文，但他 APSR 出了兩篇關於中國，他在 2013 年出了一篇文章，就是我們如何用文本探勘的方式去挖掘北京的意圖，他的研究發現，其實他怎麼做的呢？他在所有的微博裡面做假帳號，雖然今天已經不合法了，違反倫理，但是在 2013 年那時候還沒有被限制，所以他有許多的機器人散布在微博裡面，他可以觀察到，所有在微博上面的機器人抓到的文字，指出北京可能在使用五毛黨在做注意力轉移的工程，這是 2017 年剛發表的文章，大家去 Google Gary King 會找到這兩篇 APSR 的文章，他本來是實證主義者，也就是他在政治學界是一個旗手，強調因果關係，可是因為他的技術能力很強，所以他跳進了實用主義的這個區域，他用資料科學的概念告訴大家：今天其實沒有理論也沒有關係，我派我的機器人下去看微博上到底發生什麼事情。他是第一個鼓吹我們政治學者要走向理論，卻又是第一個要我們放下理論的，這對政治學界來講是一個巨大的衝擊，

他的衝擊導致了現在大部分在美國求學的博士，都走向資料科學這條路。因為對實用主義者來講，理論可以先放下，因為事實或是從資料中可以發展出的樣貌，發現這些樣貌當作問題意識才是最重要的。這對我們政治學界起了翻天覆地的變化，當我們不必特別強調因果關係的時候，我們其實可以用大數據看到我們從未想過的問題。Gary King 說的這些問題意識，剛好成為實用主義在政治科學復興的基礎。換句話講，有些實證主義者不太認識他了：他一個實證主義者不談實證主義、不談理論，卻談中共是怎麼樣用五毛黨來分散群眾注意力。他又開啟了一個新的可能。當然等一下會講到大數據的一些限制。

他的這篇文章我拿給我們中山大學的大陸留學生看，結果一看就說錯誤百出，誰其實是傾向五毛黨，或者誰啟動了五毛黨這些人都說錯了，真正瞭解文本的人，也就是大陸來的人，看著不認識中文的美國學者所寫出來的政情分析會搖頭。可是我們學者會說這是 APSR 耶！好歹人家登上去了國際期刊，可是不懂脈絡的人寫了一個和自己脈絡無關的文章，這是什麼意思呢？它（大數據）有無窮的可能，但是我們到底在幹什麼？真正瞭解他們（中共）的不是我們嗎？真正瞭解他們用五毛黨在做什麼的不是我們嗎？結果是由哈佛大學教授，用完全不懂中文的概念，放了機器人進去之後，把文爬出來，然後用 google translator 翻譯一下，他就好像懂中文了，就可以寫頂尖的政治學期刊文章了，有時候看到這裡，就想把書丟在地上。早知道我們做的比你好，我們如果做的話一定比你好，因為我們知道所有的內幕，我們知道中共在想什麼，如果想什麼結合大數據，我們說出來的故事當然是 APSR 等級的，APSR 是美國政治學期刊的縮寫，American Political Science Review，是我們學界最重要的期刊，那麼我來看看他到底為我們說了什麼。

他手上拿的這張照片，背後寫了一個白板，他的白板上面寫了，我是大數據的科學家。「我想做什麼」我把他唸給大家聽。大數據可以用來評估政策，可以用來瞭解大家的 PO 文到底說了什麼，甚至可以瞭解死亡的原因，大家現在都希望長壽，我們可以用大數據，用公共衛生資料庫的方式，來知道我們要怎麼樣長壽，我們可以用大數據來瞭解怎麼樣劃分選區是最公平的，大家可以懂他的意思嗎？我們現在劃選區的方式其實是非常土法的，到底人口結構如何？重劃選區也可以 用數據來做，這都是他在白板上面所寫的這些發想，還可以 reverse engineer China’s censorship program，就是 2013 年那篇文章，他的發現是在網路上說民主不一定會被抓，網軍在抓什麼人，他用逆向工程的方式去做，大家能不能猜的出來，在網路上講什麼話會被抓起來？當然現在有很多隱晦的字眼對不對？但是在 2013 年的時候會被抓起來的是任何有關會帶動群眾集會的字眼，這個人會被抓起來，不一定是他講了民主這兩個字，這個叫做逆

向工程，倒數第二個，當然他後面還會講到，對國際現勢會發生什麼事情，英國脫歐之後會發生什麼事情，其實用大數據做預測，也是有可能的，這是他的雄心壯志。

以實證主義者來講，當然還有一個流程圖，這張圖就是有左右兩邊，左邊是傳統的，有一個理論過來我就拿來用大數據驗證，可是這本 2016 年最新的教科書告訴大家，拿大數據做理論驗證是不夠的，我們要開發理論，實證主義者拿大數據來開發理論這件事情，我們苦口婆心講了很多年，可是就是沒有人承認，終於在 2016 年有人用教科書的方式把它寫出來了，我在這裡再把它講清楚一點，今天在座如果做學術研究，一定會有一個理論要我們去套用對嗎？某某人寫了什麼東西，然後我們引用他，然後我們來驗證他講的對不對，這叫驗證，也就是我們左手邊的流程圖，現在大數據出現之後我們要強調的是右邊這塊，也就是他們倆個併行，右邊這塊叫做發展理論、發現理論，這又是驚天動地的變化。因為這是實用主義的觀念帶入實證主義的區域，實證主義不完全為了驗證，Gary King 跟我們講的，就是我們可以來發現理論，五毛黨真正的作用，如果不是為了抓人、抓這些反動的那麼五毛黨到底在幹什麼？如果你能夠抓到五毛黨這樣一個文字的大軍，他的一些現象，你有機會發展出你自己的理論，大家可以感受到這個潛力嗎？就是我們可以沒有理論，但如果你是做台灣研究，你可以針對台灣現象，發展出屬於你的理論，在台灣有個很有趣的理論，是由民間發展出來的，我不知道大家有沒有聽過，政治學者很少去做這方面的研究，叫做鐘擺理論，大家知道這個理論嗎？就是今天選這個黨，但下一次為了制衡他又盪過去選另一個黨，有沒有聽過這個說法？有沒有由民間發動上來的？就是我們今天西瓜偎大邊，叫西瓜理論或西瓜效應，這一次誰比較有可能贏我們就靠過去，可是呢我們又觀察到一個理論叫鐘擺，就是選民會從這邊又回到另一邊，所以請問一下，西瓜跟鐘擺誰是對的？咦？有趣啦！因為你也不知道，可是每一個人都在用，淺顯易懂，所以當民眾出現一個叫西瓜理論的時候，好有學問對吧？這個就是由現象或民情觀察出來產生理論，那麼西瓜理論要不要被驗證？要。今年這個選區，到底是屬於偏向大的這邊的西瓜理論或是西瓜效應呢，還是想換人做做看的鐘擺？大家從這個比喻可以看到這左右兩邊的互惠的必要性，這個在過去 10 年的教科書沒有出現過，所有在過去 10 年的政治學教科書都強調否辨，就是誰講的理論、他是大師，所以那個大師的理論要經過大家用驗證的，我們所知道大數據的應用也都限制在這裡，就是我們把大數據拿來驗證理論，可是很少人會去想，大數據是用來發現理論的。這個給所有人文社會科學帶來了不起的機會，大家可以看到滿地都是黃金嗎？滿地都是等待被發掘出來的理論。不是一個知名學者還是可以發展理論的。只要你有數據，

你就可以從數據裡面說話。等一下給大家幾個例子。

這裡有一張照片，大家看一下這是 Google，大家看到 Google 會想到什麼？他擁有大數據對吧？他的數據涵蓋了全世界多少人口，我覺得至少有一半吧？除了窮鄉僻壤，否則人人都有一個 G-mail，擁有 G-mail 的這家公司，擁有我們多少資料？大概沒辦法想像對吧？他可以說我是普查，光是台灣他就可以做研究了，可是大家要想到，Google 即使是擁有大數據，這張圖告訴你，他還是派出了小車，在 2016 年，在美國 50 個大城市，去做廂型車專訪，他如果有大數據幹嘛做這種事？他把人邀請上車去問他的 Google+ 好不好用，他在私底下做這件事情的時候有誰知道？我們說我們有大數據，可是某種程度，擁有大數據的人是不限於大數據在做開發理論的工作，就是說我想要從哪裡看，我要從人性看，我要從人的角度看事情，就是說大數據也有，厚資料也有，也就是說真實的感受的資料也有，可是哪個要哪個不要？今天踏入大數據的我們，如果我們的矛頭轉對了，具有政戰的基本素養，又有大數據的素養，你覺得呢？我覺得這是一個非常大的潛力。不管是傳統的民調或是訪談，都有和大數據互補的機會，學術界和國際上這種書愈來愈多了，這三本書基本上都可以找到中譯版，也就是意義探勘的工程，大家現在在數據探勘、文字探勘，文字探勘背後還有意義發覺的必要性，所以我列出了幾本書，我中文書名唸給大家聽，一個叫做《無知的力量》、一個叫做《小數據獵人》small data、一個叫 sense making《演算法下的行銷優勢》，這三本上上個月剛翻譯出來，我覺得我們要做大數的同行、或後起之秀們，其實應該要看一下這種書，也就是這是我們的底韻，我們的人性如何拿來掌握大數據，這裡面有很多人性的東西在討論，sense making，我會花一點時間講這個。

大家想看看，除了 CIO 之外，chief information officer，除了大公司裡面有這種人以外，以後，或許 10 年內，會有一個 chief sense-making officer，專門把資料還原出他的原始意義的工程師，我覺得今天回到這裡沒有什麼比我們做心戰、做人性、瞭解人的需求還更適合這個職位的工作了。我在中山大學政治學研究所，我們強調的就是人性這件事情，我們看見了大數據的發展，我們也在玩大數據，我們所上有 8 個老師，4 個在用 R，我自己教 R 教了 10 年，我們並不是碰不碰數據，我們更在乎的是資料過來以後怎麼解讀。如果我們方向對了，大概我們說出來的東西，是沒有辦法被複製的，所以知識界的立場怎麼調和，我在這邊給大家一個概念，就是讓資料科學稍微有人性一點，就是你會帶著一個清楚的人性關懷來做資料的蒐集和解讀。這看起來有點空泛對吧？但是基本上我用英文來講，data science for extracting facts and discovery meanings，今天我們大數據的發展太強調只發掘事實，可是資料放在眼前可能

不知道怎麼解讀它，我覺得我們的專業有一部分是還要發展出它的意義，所以這裡已經無關大或小數據了，也就是 data science 這件事情，我們已經進化到是用資料在做科學探索，大數據本來就在裡面，那麼小數據去哪裡？各位熟悉國關的資料檔或是我們自己建的資料表，這些小數據還是有它的價值，大小的概念，在資料科學裡已經消融了很多，這是目前 2018 年的趨勢。

我今天沒有那麼多時間來講方法論的倡議，我這邊其實有一個 data assist meaning mining 或是 data assist meaning netting，這個是由幾個學者幫我一起構想出來的概念，英文縮寫我不能在這裡唸，這是他們創造出來的，但這是一個蠻有趣的縮寫，原本是叫 DAMM，可是他們故意要改成 netting，所以大家可以去看一下，我們其實是在想要知道資料輔助的意義探勘跟文字探勘一模一樣重要。

我們來看一下資料，今天的講題叫做「如何從工具來提升我們的心態」，我就來簡單講幾個例子給大家聽，一般來說探索式資料分析是蠻復古的動作，其實 1970 年代就有了，探索式資料分析基本上就是在做統計分析，就是這個地區有多少人支持 A、多少人支持 B，然後畫個 bar chart 就出來了，今天講到探索式資料分析好像沒有人想要做，因為太淺了，應該是要來做個迴歸吧，但是現在真的在復古，探索式資料分析可以很創新，因為你如果結合了意義探勘的動作，就會有效果，這有幾本 R 的書，就是我現在要舉的例子的基礎，各位有用過 R 的請舉手，至少有聽過吧？謝謝，現在有個新的東西叫 Python，所以大家在學習技術的時候總會覺得心臟壓力很大，但是工具可以處理大數據也可以處理小數據，現在就跟大家講你有 R 這個工具以後，你不用怕大數據也不用怕小數據，你有小數據可以做出跟大數據一樣的事情，如果你有大數據的話就更不一樣、更不得了。

一般我們講到民調，無非就是 yes/no，你贊不贊成死刑這種，但如果我可以告訴你比你單問一題更多的結果會是怎麼樣呢？這個研究方法，基本上就是告訴大家，原來我們所做的 factor analysis 可以用在民調的小數據上面，我把節奏放慢下來跟大家講一下，如果今天我們要做一個簡單的調查，訪問你三題，都是 yes/no 的問題，請問你喝不喝下午茶？喝哪一種茶？紅茶、綠茶還是抹茶？然後你喝茶加不加糖？你看這三題是不是太平凡，高中生如果第一次做民調讓他練習可能就是這樣，可能就到餐廳裡訪問一下，這東西有什麼意義？有什麼用？大家會覺得三題的分析沒什麼用，注意看，如果用這個方法，這個叫 multiple correspondence analysis，它其實就是原本的因素分析用在民調上，我在中山大學就是在做這個事情，回到這個例子，如果我告訴你，我可以找到一個樣貌，是所有喜歡喝下午茶的人都會選綠茶，而且選綠茶還會加糖，你可

以聞到市場的潛力嗎？三題沒什麼了不起，如果我把這三題的共同成分組合告訴你就變得完全不一樣，這是小數據，才三題，大數據有幾百個欄位，這些欄位裡面你可以找到的不只是變數之間的關係了，你可以找到選項之間的關係，這是什麼樣的世界？這是過去 3 年左右，我不小心挖寶挖到的，這東西是由法國學者發出，其實早就有了，但因為 R 的關係，它有了套件，所以我們開始使用。

我們突然發現，就算是民調的資料，傳統的小數據也有它的潛力，這本教科書早就在了，可是為什麼我們現在才知道？大家可以想想看剛剛舉的喝茶的例子，用在任何的問卷調查都適用，所以我等一下馬上就舉一個我自己做的例子讓大家看，我這邊有 3 筆 data，一種是傳統的面訪調查，很貴的那種，大概 200 萬，另外有電話調查樣本、網路調查樣本，網路調查樣本就是大數據，從網路上爬來的東西沒有代表性，我今天不管是網路上爬的東西或者是在台北車站訪問人，這些人不能代表全部的選民，他們沒有代表性，但是內容可能很豐富，從很有代表性的資料到沒有代表性的資料，我來做一些事情，我用剛剛的方法找了 30 題的問卷題，來問他們的認同，從政黨認同、你是不是炎黃子孫、我們的疆域包不包含中國大陸，你看這些題目不是很傳統的問法嗎？都是認同題，問了 30 題沒辦法全部唸出來，其中有幾題我覺得很有趣的，就是當說道什麼歷史事件是你覺得重要的，要留給下一代知道，有的人說推翻滿清建立民國、有人說 8 年對日抗戰勝利、有人說 228 事件、有人說台灣民主化，這些事情，也是 30 題裡面的 4 題，30 題總理出來以後，出現兩個維次，都是小數據出現的，上面的黑點，是 1900 多個受訪者撒上去的樣子，經過意義推導以後，X 軸的右邊代表「我是中國人」，就是以中國人做一個民族的概念，向左代表「我是台灣人」，也就是 X 軸的最左邊說的就是我是台灣人 only，最右邊是我就是中國人沒有其他什麼可以加上去了，那麼在中間的呢？大家可以想一下，這是 X 軸的部分，那 Y 軸在這 30 題裡面我們可以挖掘出意義，愈靠上面的人愈不認同中華民國的歷史正當性，而愈下面的人愈在乎這件事情，其實這個題目我覺得今天來這邊，回到政戰來講格外的會有共鳴才對。我覺得這邊的聽眾是在乎的。

我們把受訪者撒在 X 跟 Y 的二維圖像上面，我們可以看到在這 5 個世代，我們把他們依照政治世代切割，就是他們在 18 歲的時候經歷過什麼重大歷史事件，那麼第一世代在 1931 年前出生的他在哪裡？如果是 1、2、3、4 象限他們會在哪裡？會在第 4 象限，你可以看到這 1900 多個人的分布大概就是我愛中華民國，我是中國人，接下來我們繼續往下看。第二世代，當然以人口比例來講會比較多，他們會在哪裡？看到了嗎？大家注意到，這還只是小數據，就已經可以做到這樣子了，我可以把 4 個世代分類放上來，基本上我可不可以說他比

較偏右？就是在政治立場上認為我是中國人是比較多的。第三世代基本上是散開來，1954 到 1968 年出生的這個世代，已經分散在四面八方，大家可以看看，是上多還下多？下多對吧，但是我們現在腦袋都可以開始說故事了。第四世代，1969 年到 1978 年出生的在哪裡？各為用手指一下，我們看一下我們指的一不一樣，好像是這樣，用眼睛看而已，如果是同一張維次圖，這 1900 個人撒在同一張圖上面，我們會看出同在一個時空下的人們不一樣的認同觀。再來看第五世代，也差不多，那麼我們把全部都打出來的話，1、2、3、4、5，你如果看的清楚的話，紫紅色的地方就是全部的人裡面，他所屬的年齡層就是紫紅色標出來的，那麼這是剛剛我們看到的每一個圖像，所有的藍點就是全部 1900 多個受訪者，有代表性的樣本所打在圖上的樣子，我常常跟同學開玩笑，大家看了這張圖之後，你知道這張圖是什麼時候的嗎 ?2013 年，2014 年出現太陽花學運、2016 年總統大選，如果我在 2013 年就可以看到這個圖像，我會說什麼？大多數的藍點都聚集在哪裡？上還是下？下，所以我就跟同學開玩笑，如果我更早以前就知道這個研究方法，那麼我一定會知道，多數的人或是說選票，一定會在下面這裡，那麼最好的行銷語言會是「維持現狀」，中華民國現狀，因為這張圖下半部剛好是中華民國正當性、最多人的地方，當然這叫事後諸葛對吧，已經被講掉了，如果我們現在是用 2013 年的資料來看，我們會說故事的話，你會看到剛剛那 4 張的分鏡圖，給大家看到，就是有一些在右邊、一些在左邊，但是很多人在下面，這告訴我們什麼故事？中國人的中華民國跟台灣人的中華民國，這是我們看到剛剛那張圖能說出來的故事，這個我剛剛做出來的時候，如果我要打擊台灣的話那太容易了，我可以讓台灣完整分裂，但如果用用在對的人手上我可以讓台灣稍微融合一點，這要看我們怎麼用，小小一張圖卻具有巨大的潛力，我們叫 insight，直接可以指穿到底選民在想什麼，那麼現在大家看一下這兩個圖，另外一個笑話我跟同學們講的就是，當這兩個世代，分別是祖跟孫坐在一張餐桌上面的時候會發生什麼事情？孫子跟爺爺坐在那邊談選舉，結論應該就是翻桌了，話不投機半句多，這是兩種不同的思維方式，但生活在同一個時空裡，這是給大家的第一個例子。大家想想看這是一個小數據，我們用小數據產生出來，但是我們是用 R，是用資料科學的工具所做出來的，其實語法給你，你自己執行一遍不用半小時，學 R 的人沒有那麼痛苦，因為 R 的語法檔是可以拿來複製的，需要的話待會可以分享給你看，但重點在於，你拿到這些圖片之後能不能說故事呢？

我們再來看另外一筆資料，這是一個語法的小東西，我把很多題抓出來，你有看到很多列出來嗎？字小一點看不清楚沒關係，但要告訴大家的是說，一個資料檔裡面很大有很多欄位，但我可以取我要的欄位出來，形成一個小的資

料檔，用它來做探索式分析，其實大數據在分析的時候終究也是小數據，大數據不可能全部一起跑，大數據在跑的時候一定要切割出小數據來分析，所以不是大而以，做成小數據終究是最後的 ABC、ㄅㄆㄇ。我們選出幾個題目，這邊不一一的唸，我們來看下這是我個人研究中比較喜歡的，中間選民的這一塊，這個是有代表性的資料，我們把這個檔案左邊紫色區域就是全部的受訪者撒在這個區域上面，右邊代表支持政府、相信政府，左邊則反之，下面代表隊民主有熱情、民主是最好的制度，而上面則反之，圖上右邊紅色和黑色的點，不曉得大家看不看得清楚，黑色的點就是自稱沒有政黨傾向的人，分布都在 Y 軸以上，黑點座落在上面、上半部，還記得下午茶的比喻嗎？我們如果讓這些沒有政黨傾向的人，放在二維的概念圖上面，他們會是在什麼樣的狀況？對民主比較沒感覺，不一定是反民主，但是會比較沒感覺，政治不關我的事、投票也不想投，以這個圖為基礎，我們來看看藍綠又怎麼想，這是同一張圖，但是是以政黨來做切點，可以看到，這讓我非常訝異，應該不會是這樣才對，我們做資料科學的人，看到這張圖，應該是會流冷汗的，這是什麼意思？應該看起來藍綠都喜歡民主才對啊？好歹也輪替幾次了，大家對於民主都有基本的概念，怎麼會出現這個樣子。左手邊是國民黨支持者，右手邊是民進黨支持者，自稱的，這是 2016 年的資料，那你會說，那時候的執政者是誰，敗選不爽，當然就不喜歡民主，可不可以這樣講？有可能對吧？我不會講國民黨的支持者就是怎麼樣，但是如果你回到當時的脈絡，我選輸當然就不爽阿，選贏的當然對民主充滿了熱情，是不是這樣呢？所以我給這張圖的意思是說，如果你是我，你是會直接相信這張圖，還是會繼續去找資料？如果你是新一代的資料科學家，你應該是會繼續找資料的，這個東西很嚇人對吧？如果這個到國際會議上去發表，台灣人對民主的熱情是依政黨來分類，那這張圖出去就會嚇到人了，所以如果你是研究者，你應該會往下問，如果回到 2008 年的資料，會不會是這個樣子，但我自己做了，答案是沒那麼嚴重，但是你從 2008 年看到 2016 年，怎麼會變嚴重了呢？這裡面是不是有理論？這就是開發理論的機會。

所以剛剛介紹了這個工具，還有一個例子，大家可能會想到，實證主義者在剛剛圖上這四個大陸區，知識論上有不同的領域，你到底屬於哪一種人，你是喜歡意義追求的詮釋主義者，還是你是喜歡直接找問題解答、政策解答的實用主義者、還是你是喜歡整個脈絡、整個過程的科學實存主義者、還是說其實你是喜歡因果關係的，有因有果，我要找到原因的實證主義者，在這個地圖上面，大家各自有自己的城堡，但是新一代的資料科學家，卻是嫁接在四個島嶼、四個板塊上的那塊玻璃，在這片透明玻璃上的這群人，是可以遊走在不同的知識國度的。這是我們資料科學家未來的位置。我們資料科學家其實並沒有設限

自己應該在實證主義者，或是限縮自己一定是詮釋主義，資料對我們來講應該是一種求真的方式，不管資料是小的、大的、音樂的、漫畫的還是日記，這些在資料科學裡面通通叫做 data，這些資料怎麼發掘出意義，或者是你融合不同型態的資料，可以做出一些很厚的東西，厚實、有人文內涵的東西。

大家可以想想看剛剛那四個世代分裂的對嗎？那張圖，如果加上不同世代的專訪會有什麼樣的效果？民調大概告訴我們，四個象限的人各自會說什麼話，但是你如果找出這些受訪者能夠真的談出一些話，你的資料如果是大數據的話，它的厚度就會往下走，這時候是大還是厚比較好？沒有，它們已經融在一起了，既厚且大，可是既厚不一定要大，所以厚的概念是我今天帶過來，希望我們要一起努力的，那麼在最後，我們當然還可以來幾個比較刺激的問題，用這個觀點來看，我們是做民調也好、大數據也好，大數據也可以去爬文做分類，但是我想要呈現的是大數據能做到的，小數據也能做到，而且會做的更好，大家想想看我們文字爬文是不是要經過很多苦工去整理？文字不管多少，到最後在做情緒分析的時候還是會變成類別對吧？那麼傳統的工具，如果可以單刀直問一個人的偏好，那不是更好嗎？大數據是間接取得一個人的偏好，可是如果民調好好的問，任何的研究問題研究設計好好的問，你甚至可以取得一個含金量更高的資料檔。

我再來舉一個例子。目前大家講統獨這個議題，你看看民調會怎麼問？電話拿起來就說，關於台灣海峽兩岸現況，有幾種說法，請問你比較支持哪一種，一是立刻統一、維持現狀後再統一、永遠維持現狀、維持現狀以後走向獨立，我現在問完了，是不是常聽到這種問法？然後最後就一個光譜對不對？這個問題最後就由媒體、國際去炒作，炒作到後來沒有人真正清楚台灣人到底在想什麼，有沒有人真的問過什麼叫做統一？那麼獨立又在說什麼？你看，這麼簡單、基礎的問題，可能要用大數據爬，對吧？大數據可以去爬，因為你可以爬出大家對統一的論述、對於獨立的論述，但我們可以用民調直接問，現在看到的，我們先看統一這個，我們可以很簡單的問一句話，這是我在執行的微笑小熊調查小棧的平臺，是一個以精準的發問，作為民意探勘的一個實驗，會員有 8000 多個人，活躍會員有 1000 多個在填寫問卷，他們所表達出來的立場，這是大概一年多前做的，我們問「說到統一，你最先想到什麼？」，這個用電話大概沒辦法問，因為電話還沒問完大概就掛了，因為選項太多，可是如果用網路調查就有可能，所以我們可以很清楚的列出所有有關於統一的想像，包括知名超商都可以放進去，受訪者很自然的狀態下回答出來的東西沒有代表性，但是在某種意義上面，我們可以說故事，想想看如果這個東西是大數據的話，我們可以爬出的內涵不知道有多少，這只是其中一題而已。

第一個選擇，我唸一下選擇，看到最多的那個 bar 叫做中國併吞台灣；第二個叫中華人民共和國統一全中國；第三中華民國統一全中國；中華人民共和國消滅中華民國；中國人的新時代；中國夢實現；台灣人最好的選擇；在一國兩制下台灣高度自治；獨立戰爭；某知名超商企業。說到統一這兩個字，我們問的其實是在當下對於一個概念的直覺，這是詮釋主義者的作法對吧？詮釋主義者追的是它的意義、它的概念，可是當你知道原來大家在說統一的時候是有主流的，雖然有主流，但是是很分散，所以如果還是只用極統、極獨來問，到底是誰偏了呢？那當然，比較緊張的題目就是說到獨立你想到什麼，其實以我們政治學者來講，我們終究要碰觸這塊東西，而這一塊東西出來的結果卻令人非常的驚訝，我拿這個結果去問實際的同事，他們都不敢相信結果竟然是這樣，就以這 900 多個受訪者，沒有什麼代表性、大多是年輕人，他們所填出來的，我從上面唸下來，選項唸下來大家看一下，我們在設計問卷我們可以讓類別一直打開，我們只要他勾心中的答案就可以了，這種方法是網路調查很強大的一種地方，第一個，台灣是個新國家；中華民國是主權國家；領土上與中國大陸不相隸屬；有自己的軍隊、司法、行政系統和國家符號，例如國旗、國徽；台灣人自己的國家；兩岸必然有戰爭；台灣要切斷和中華民國的關係；台灣要切斷和中華人民共和國的關係；完全切斷與中國有關的東西；中華人民共和國理解下，台灣如國家般運作；自主和尊嚴；最後一個，是最多人選的，在剛剛所唸的各種關於獨立的定義裡面，分離主義，不管是分離哪個地方，但是這個問題下去之後，所謂的超級敏感題，都讓你看到讓你選了，卻不是最多人選的，最多人選的答案是「得到世界各國官方承認」，管他承認什麼。就是說當你這邊只能單選的時候，你最先想到什麼？而他所選的答案，是不是我們以為的那個定義呢？這裡面大家是不是可以看到滿地的黃金呢？就以概念來講就有很多可以發揮、很多說故事的空間，在蒐集文獻上，在文字探勘和爬樹上，其實是有很多彈性的。

這是給大家的另外一個例子，就是說當我們從意義出發的時候，我們得到的大數據或小數據，背後還有很多我們能夠做人性、做心戰或是做政治學、民意研究的人需要持續探索的。結束之前，給大家一張圖，我很喜歡的一個。現在 AI 很盛行，AI 的對面是什麼 ?human， human intelligence，就是說是由我們來賦予資料意義的，當然 AI 會把我們的詮釋拿去用，但是不管是 AI 還是大數據，在這本書的封面上呈現出來的就是互補或是競爭的關係，我們做為資料的解讀者、開發者，解讀絕對是我們唯一可以勝過資訊工程師的強項，今天來這邊其實想給大家的訊號是，如果我們擁抱資料的話，那大家不得了，因為我們本來就懂人，懂人的專業現在已經要從政治學爬回來了，我在中山大學在努

力這件事情，我回來這裡為什麼那麼興奮，是因為這裡本來就是在挖掘人性、瞭解人性的地方，我覺得沒有地方比這裡更適合做人性和資料結合的堡壘。我們在這裡看見了，這是我興奮的原因。我們再說幾個結語，就是我們要把大數據的船開出去和世界競爭，真的對兩岸民情有話語權的基地，我們會有什麼地方需要把事情做對。尤其是對各為新手、即將進入這個領域的人來說，大數據呈現的可能是真相也可能是假象。這其實是老生常談，例如剛剛所說的，國民黨支持者是不是真的不喜歡民主？這個問題要留待我們去解答對吧？但 2016 年第一筆資料發出來之後，怎麼會有這個現象？這是需要被解讀的，或許不是真的，但如果有人拿這個資料去招搖撞騙的話怎麼辦？如果拿去公開場合說了，會不會有人挑戰他？

這是另外一本書，就是剛剛我說的那三本很有名的書的其中一本，叮嚀我們的，我們一直無法抗拒對確定感的追求，因為我們的眼睛一直在找 pattern，就是我們一直想要確定的東西，所以人有預測的需求，你有沒有辦法讓自己處在一個沒有答案的狀態很久？你能不能夠抗拒這種壓力？我看到一筆資料、兩筆資料我都還不敢講它是真實啊。我覺得用這種戒慎恐懼的態度來做大數據，就有人的味道了。太想要確定感的話，最後會做的像氣象預報。我常常跟家人開玩笑，氣象報告不準，你們都不罵氣象局，但是民調不準你們就罵我，因為你們都想要預測，但卻對於預測這件事情有不同的容忍度，該下雨沒下雨、該放假沒放假，然後大家就罵翻了，這是我們對於確定感的需求。對大數據來講我們必須抗拒這件事情，所有的 pattern 找到了，我們必須確認再確認，這個不容易，因為一個不小心，我們就會摧毀一群人的既有信念或相信、對既有信念產生轉變。從攻擊的角度來講可以，就是譬如說我們想要改變什麼輿情，我們把握了數據，你就以為是真相，這是一般人最喜歡、最容易切進去的，當我們視覺化一件事情以後，你就覺得真實是這樣，所以它是兩面刃。

擁有大數據的人，就是擁有尚方寶劍的人，你會被我剛才的圖所說服，那代表別人更無法抗拒，如果你剛剛發現這些圖有一點說服力的話，那麼一般人就沒有抗拒能力，那我們的角色就和魔術師差不多，只是說我有 PHD 了，那這個事情，是很多書上再三叮嚀我們要小心的，經驗本身不能被拿來做為詮釋的材料，所以我們有資料科學，但是科學家，或是我們學者所做的詮釋又如何可信，這就有點像我們既要學 R、又要學大數據的技術、又要通人性、要能夠把故事說圓，這個是另外一個學問，這樣沉重的負擔是兩肩扛，我覺得是只有在對的地方、對的人才能做對的事情。目前有很多資料科學家或很多單位，都只偏向一擔，就是我把技術做對了，然後把 pattern 告訴你，就可以安身立命，可是我總覺得我們做人文社會科學，總還有一擔要肩負，就是詮釋、說故事要

對的地方。

所以我們的結語有幾點。大數據分析已成為資料科學的一環，所以我們寧可說我們是 data scientist。Data Scientist 包含了我們對於資料的尊重，卻又不被它駕馭的概念，所以資料科學家是一個新的名詞。政治科學必須迎向資料分析的時代，政治科學在某種程度已經不被人相信，因為我們做的東西不太接地氣，我們有很多的理論，我們在驗證理論，但是我們不一定有資料和故事來充實我們的理論，我們可能覺得因果關係確認就好了，可是背後還有很多人性的東西沒有被挖掘。放開理論，政治學給了我們什麼？其實不太多，我們政治學有沒有談人性？現在有一個東西叫道德心理學非常火紅，道德心理學這件事情，才是我覺得政治學應該做的，不是心理學而已，道德心理學這件事情，我舉個例子給大家聽，你知道好人為什麼會分裂嗎？美國有所謂的民主、共和兩黨，現在分裂的不得了，一邊擁槍、一邊反槍，一邊支持女權、一邊反墮胎，分裂的很嚴重，可是有學者研究出，依據道德理學，人有七種道德光譜。如果你這樣講，事情就可以講得圓。民主黨在乎的人生價值是自由，保守的這邊在乎的是聖潔，這兩種價值給你這樣看，你會不會覺得兩邊都是好人？道德心理學某種程度上是走進人的價值觀做研究，而當看這樣的書籍的時候，我總覺得我們新一代的政治學與大數據結合之後，有太多閃亮的題目可以做，但在告訴我們之前，所有的學者只是在追尋分裂的原因、或是分裂的多嚴重，卻沒有人去解釋這兩群人的道德觀、道德信仰是完全不同的。

大家可以回到剛剛那五個世代，你有五個世代，各自有各自的國家想像，那麼接下來你是要拿刀把他們切割的更清楚呢？還是讓他們融合在一起？這不是很好的題目嗎？事實在這裡，我們實用主義者，怎麼樣讓它變成有效的心戰力量。但如果這個東西給敵方要破壞台灣社會的人拿到就麻煩了。就是說，有時候我自己在做研究的時候，看到一個現象是我們自己懂的，我們才有辦法詮釋，網路的世代，對方的手伸進我們的民情研究也是有的，但是詮釋這件工程只有我們自己能夠掌握，同樣的道理。我們在網路世代我們去微博做機器人也有可能，但我們又瞭解他們的民情到什麼程度？要挖到什麼程度？知道他們想什麼、害怕什麼？當然也要瞭解我們想什麼、害怕什麼。我覺得這就是對的地方，也就是這裡、對的人、在用對的工具所以可產生的綜效，這是我今天給大家最後的期許。

大、小數據關於技術這件事情其實已經平民化了，R 我教了 10 年了，以前不敢推廣它，可是 R 現在真的變容易了，一個工具或套件，密集的訓練坦白講不需要超過三個禮拜，那以各為現在肩負了很多任務來講，還是有可能學得起來。工具不是要你成為程式設計師，這個東西留給資料工程師就可以，這個東

西就交給資料的分析師，也就是我們，我們分析資料所需要的總能量，是可以算出來、是可以掌握的，不要把學英文這件事情當做要學英文就要用英文唱歌和寫詩，然後來告訴老師其實你不太能學英文，大家可以瞭解這個比喻嗎？誰要你去寫詩？

學資料科學本來就是要拿 R 去做與當下資料有關的事情，你掌握了好的套件、掌握了好的語法，所以以我過去 10 年的教學經驗我可以告訴各位，進入這個資料分析的門檻，已經不是要不要進去的問題，而是你為什麼不進去的問題，所以從資料中發現有意義的樣貌，去發掘，就是我今天有一筆資料給你，像微笑小熊我們一直在做這樣的調查，資料都上傳、分析出來、現在年底選戰在 11 月，我們好多政治調查在進行，這些資料全部都公開出來，你有技術和方法就可以去分析這群民眾究竟在想什麼，是不是晚上不睡覺在用手機的人比較容易投給什麼人？而他在投給他的時候，其實是因為他喜歡電競？來自不同題目，所能構成不同的圖像，可能民調還是得幫上忙，所以民調不能只問一種問題，例如 : 請問你支不支持成立電競中心？你支不支持用台灣名號參加東奧？這種 yes/no 的問題，在今天和大家分享資料科學的角度來看，不覺得很無聊嗎？單薄。就是我們的問題用單問的這種方式，在資料科學眼中已經單薄了。我們要有一種複合式的問題，而這複合式的問題剛好就是大數據帶給我們的精神，大數據的精神就是我不怕有多重欄位、我不怕有幾萬筆的資料，我的欄位多，表示我分析的對象還有產生出來的故事就大，像健保，去識別化的健保有什麼效果？現在很多公衛 paper 都密集的產生出來，因為我們的健保是世界第一，我們去識別化的資料釋出之後也是世界第一，因為我連我的血糖曾經有得到什麼問題什麼狀況，跟我一樣有這個問題的人還會有什麼狀況，這件事情，在大數據的時代就有可能。可是大家想想看，把這個概念用在民意、輿情、文字探勘的結合上面，那潛力實在是太大了，所以政治科學，或是社會科學的學者，剛剛講的四大科學的領域，不管你對號入座你覺得你是什麼人，不管你喜歡理論、因果關係、你喜歡一個機制或過程、還是你喜歡一個解答或解方、還是你喜歡意義，這四種人其實在資料科學裡都可以找到自己的地方。

畫面上的這八個字就是今天的結尾。學 R 不難，但是我們有人剛剛大家看到這些技術，怎麼樣在學習技術的時候，提升自己的心態，這個心廣又要虛，廣就是我不怕來問什麼問題，什麼欄位都可以，你有創意你就來個欄位，來個欄位就通通畫近來，但是你要虛，因為你不是對的，坦白講人在知識面前永遠是瞎子摸象，我摸到的象耳是具體的，可是沒有人說這就是象貌，所以 teamwork 在未來是極度重要的事情，那剛剛孫老師和我講的，我們這邊的大數據中心也就是一個綜效的概念，這是非常令人興奮的事情，就是綜合大家的專

長，避免大家進入瞎子摸象的局面，這是一個對的方向，也是讓人非常欽佩的方向。手要勤要黑，就是 get your hands dirty，就是現在手要下去摸資料、碰資料，這就是黑手，坦白講資料科學就是黑手的工作，你以為資料清理很華麗嗎？資料清理遇到更多文字資料那手更髒，就是你清理資料之後的手是非常的髒，那你要願意讓手髒，就是你下廚房就不怕手油，這個是鼓勵大家一起做，我自己也在做，所以大家一起努力，互相勉勵，最後是一些參考資料，這是我們所上的幾位老師，所有對意義解讀、開發有興趣的，我們所上有這些老師們都在走這個概念，不管是做國際關係還是比較政治，其實都是在想這件事情，所以也非常期待未來大家有問題互相交流合作，今天再次感謝我們政戰學院，各位長官、各位老師、各位同學，來一起分享我過去 10 年的反思之路，自己做了然後跟大家分享，這裡是我歸零的起點，我能夠說出這些，是因為我歸過零。我在台大政治系是唸政治思想，我從這裡歸零之後出去碰的，就是實證主義的訓練，這是多痛的過程，我唸思想的人要面對做假設檢證，我沒有辦法說服我自己，還要從它拿下學位，這個是痛苦的過程。但當你碰過這樣歸零重學的經驗，你能夠和多種人對話，我現在非常享受這個過程，我既可以和資料科學家對話，我也可以回到亞里斯多德的境界去談人性，你不覺的這兩種事情很美好嗎？有人性的訓練、又有資料的訓練，你說台灣有哪裡能這樣做？我想只靠中山政研所來打造這樣人才力量太小。所以非常期待、非常讚許，我可以說我的母校嗎？這個讓我歸零的家鄉，在做這件偉大的事情，並一起合作努力。恭喜各位，再次謝謝大家，謝謝。