何時不能用資料驅動？

第一段指出許多資料在分析時的認知偏見。本文接下來要探討的是資料不能作的事，指的是在各種方法下，哪一種衡量方式與解釋本身能夠轉化資料。我們並不是要寫一篇「謊言、該死的假象與統計學」的論文。我們都知道資料可以有目的此用來混淆視聽，在此，我們的重點是在它是如何不小心造成混淆。特別是：

- 我們用在資料上的工具不正確。
- 我們用已知的偏見處理資料。

雖然我們的範例將重點放在生物學與財務資料上，但它們或多或少是可以衍生的。「資料」指的是根據經驗、觀察或實驗累積而得的任何一組未經加工的事實。

1. 資料多不見得一定比較好

統計學是一種表示法與近似值的科學。我們擷取或觀察的系統越多，就越能忠實呈現。統計學開端文字總會強調，當你增加樣本大小，你便能大膽減少信賴區間。換句話說，資料越多對於控制誤差範圍越有幫助（見圖 13-4）。

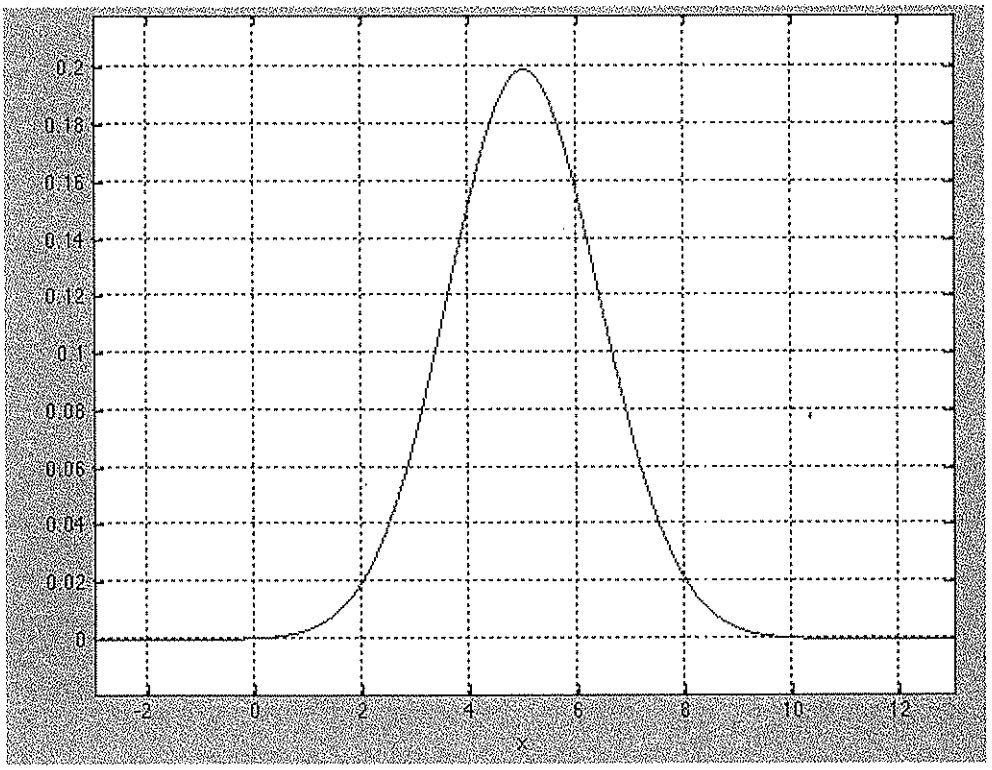

圖 13-4. 常態分佈（See Color Plate 49）

教科書上有個不錯的真理，薄紗外的世界，有些假定是必須驗證的。首先，你的資料分佈狀況如何？它必然是常態性的嗎？在眾多財務層面的例子上，分佈狀況是避開常態性的。生物醫學資料（例如特徵的表達）更常出現高斯分佈（Gaussian），但進化是沒有必要一定與中央極限定理（Central Limit Theorem）一致的。

若資料並非常態，那麼資料再多也無法減少預期的誤差範圍。Karl Popper 說明了我們使用資料回答問題的不對稱，當沒有結果支援要確認的假設時，單一矛盾結果就能證明它不成立。更多的資料只是徒增少許確認性，單一案例便能推翻一世紀的信仰。

再者，偽陽性（false positive）的成本和偽陰性（false negative）相同嗎？就算你的資料是（或看起來是）正常的，你所關心的不同結果可能也不會對稱。例如未偵測出危及生命的疾病成本，可能比誤診還大。在這類案例中，改進診斷精確度的資料（透過分離出偽陰性的方式）會比將大量資料的偽陽性挑出來還要有用。

2. 資料較多不見得比較輕鬆

資料不一定要規模化。在資訊時代裡有句老格言，處理 10bit 的資料有多簡單，處理 10 TB 的也一樣容易，但一千萬個 widget 則比作 10 個要昂貴許多。

在某些情況下，清理及處理資料的成本都不小。特別是那些需要用人類肉眼確認的，像是讀 X 光片所帶來的意義，或者改寫問卷裡的資料代碼。以 Red Queen 的方式來說[註1]，更好的電腦與收集更多資料的能力，可以驅動（以及被驅動）新解析工具程式的開發及新方式的使用。

伴隨著更多資訊而來的還有認知成本。無論你是在超市裡選果醬還是在 401k 計劃裡選擇，研究顯示選項增加，我們花費的決定時間就越長，我們變得更有可能放棄不作任何選擇，而且對自己所做的任何選擇都不太滿意（Iyenger and Lepper 2000）。

[註1] Lewis Carroll 的 Red Queen，取自於愛麗絲夢遊仙境，表示「你得持續一直跑，才能維持在相同的位置。」這個想法被用來描述由於外部壓力的軍備競賽，而必須持續共同進化的系統。

最後，不自覺的代價便是，資料多了就會開始讓我們看不到其他可能性，尤其是當我們負責收集與整理時。不難想像，能看見更多資料，代表假設能夠得到更好的支持，偏見必然產生的結果與採樣議題先前已做過探討。

3. 資料本身不能作解釋

由人來做解釋。你可能聽說過，相互關係與因果關係造就了同床異夢。在統計學上顯然彼此有關的兩個變數，其因果關係正推、反推兩種方向皆可行，也可能毫無關聯。統計學家有個記錄相互關係濫用的習慣（更別說那一堆部落格了），像對現代世界中傳統價值觀衰敗現象發出不同意見的老太太。

新聞記者是這類統計學「tsks」的首選目標。舉例來說，在最近一篇 *Wall Street Journal*（Shellenbarger 2008）裡，提出由於婚前同居而產生較高離婚率的相互關係，沒有工作收入的夫妻可以避免住在一起，這樣能夠提升他們在婚後仍能長相廝守的機會。這份研究並未提供任何因果連結，不過是新聞記者針對「資料」提出她自己對夫妻們的建議。相互關係與因果關係間的取代性並不明確。

相互關係對於因果關係的替代性不需要那麼明確。在科學研究專案可接受的情況下，若發現該假定存在相互關係，也就暗示了其因果關係，就算一切是未知的。否則，何必徹底尋求研究專案問題的答案。大規模搜尋沒有因果關係的相互關係是一種投機的計算方式，並非科學。即便是在所謂大量資料下，科學仍持有相當程度的假說驅動程序。

經驗法則的研究限制不會讓我們徹底絕望，只是得小心往前推動我們的發現，不要老是想著它們之間的因果關係。只有人類會建立有關資料的故事，那是一種修訂一致性讓故事變得合理的能力。

4. 資料並非完美的單一解答

敘述性統計（Descriptive Statistics）可以隱藏細節。舉例來說，圖 13-5 的圖表，顯示四種看起來完全不同，但其實共享相同平均數與變異數分佈。敘述性統計（Descriptive Statistics）的兩個重要支點平均數與變異數，對分布而言能表達的極為有限（Anscombe 1973）。

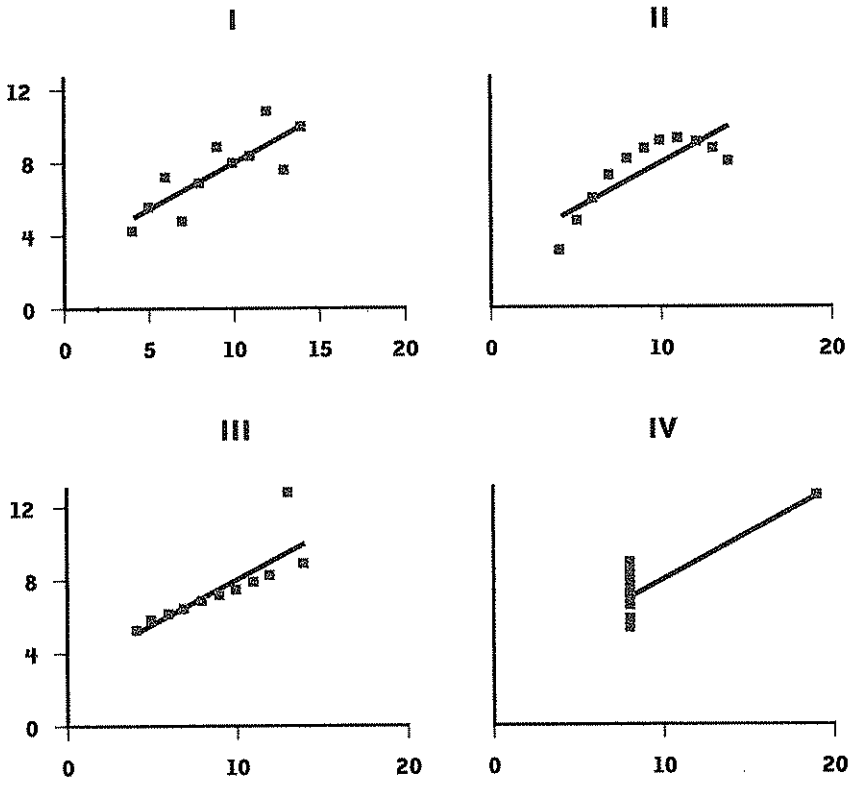

圖 13-5. Anscombe 的四組數據：每個資料集皆擁有相同的平均數與變異數。

使用資料做決策的時侯，我們傾向於認為分佈好像應該有一個很好的解釋。我們可能需要以二選一的決策基礎。美國應該宣戰？FDA 應核核准麻藥？預言誰將贏得選戰？或者在不明確的資料裡的摘要敘述，美國人有多富裕？地球的氣侯五年內會變得如何？無論變異數為何，決策才是最重要的。

人們根據結果思考，而非根據分佈情況。思考一下個人財務決策，我應該花多少錢投資股票、債券與現金？就算過去的財務表現或許能夠預測未來的回報率（若這個財務顧問是法律上認可的，他就不會這麼作）。也就是說，即使我們瞭解分佈的形態，我們仍有很多風險與報酬率的組合可選，而在這些分佈裡也帶有許多可能性結果。在既定風險層級下，一個人的退休生活可以是富裕的也可能貧乏，很難想像這些未來性同時存在（人們傾向於假設處於平均值，有時則覺得會是最好的狀況。此即所謂的「planning fallacy（規劃謬誤）」）。

決策科學家團隊建立了一套有趣的工具程式，幫助投資者瞭解結果分佈裡可能性的範圍（見圖 13-6）。參與者可以調整 100 個「可能性單位」形成分佈曲線。舉例來說，他們可能會將所有單位放在薪水的 75%，或者平均分配百分比層級的變數。接著，按下

go 檢視這些單位，會發現它們一個接著一個隨機消失。最後一個還在的便是「結果」（Goldstein et al. 2008）。因此，風險層級不但並非模稜兩可的分布曲線，而且是一組（在此為 100）機率相同的可能性。

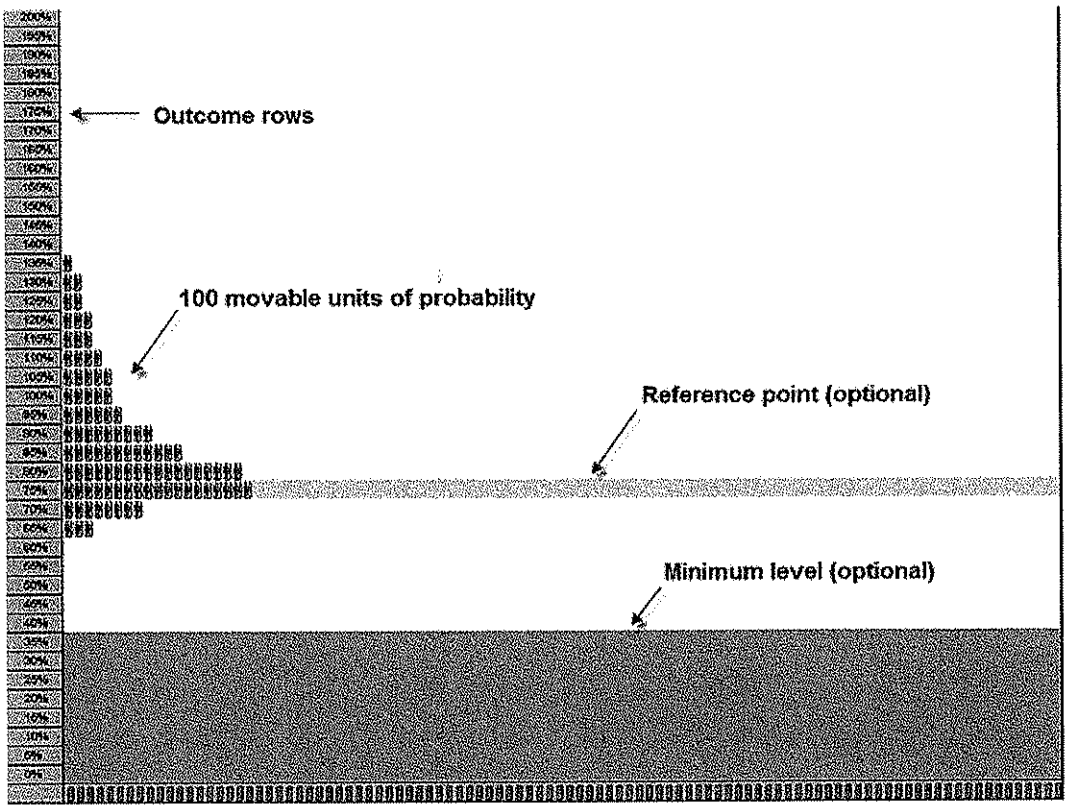

圖 13-6. 由 Goldstein 等人創作的工具程式，幫助人們瞭解一組結果的分佈情況。（See Color Plate 50）

生物學家 Stephen Jay Gould 進一步闡明與敘述性統計（Descriptive Statistics）同等的問題。「中間值並不等於要表達的訊息（The Median is not the Message）」是 Gould 對於癌症診斷後被警告他「有八個月可以活」之後做出的反動。癌症相關文獻裡揭露右斜分佈情況是建立在「整個大環境下的醫囑」，這也就是說活得很長的患者，是建立在過去醫療條件的假設下。

宣稱「八個月」就等於錯過一大片美好的人生風景。Gould 優雅地描繪出科學家的殘忍：

> 與時俱進的生物學家都知道，變數本身天生就有無法還原的特質。變數是殘忍的現實，而不是為了朝中央靠攏而產生的一組不完整測量。平均值與中間值都是抽象的。

5. 資料不是預言

建立模式（預言明天的天氣、2012 Super Bowl 的結果，還是 Fortune 500 的命運）是一門吸引人的藝術。確實，科學冒險的重要延伸 —解釋我們周遭的世界 —便是試著瞭解這個世界將會變成怎樣。

在特定領域裡，也就是實際世界裡可控制的 *-cosms*，為可能預言接近肯定的結果。會高度依循過去的事件所產生的未來結果有水熱了就會變氣體、在真空狀態下，墜落的物件會以每秒 9.8 公尺的速度加速、當生物的心臟靜止時，它就死了。認知上，Popper 的可證偽性概念沒什麼好爭論的，但它有可能導出合理性，讓接受前述三個假說的社交生活不言自明。

帶有不明確特性的那些領域，像是人類或身體的行為，模組化為有助於解釋這些模式的重要工具。然而，當我們熱心的想讓資料說話時，很有可能會讓模組超適（overfit）。

思考一下，使用 Doppler 光速尋找系外行星的問題（我不想假裝我懂得比表面上那些還多，基本上，閃亮的恆星會讓行星很難看到，所以天文學家認為 Dopper 位移的組合，只會出現在行星繞恆星而行的時侯）。模組的滿意度是很難測試的，但僅 15 個觀察，就可能讓資料符合圖 13-7 這個性感的正弦曲線（Ge et al. 2004）！

當我們超適模組時，便失去了預言的權力。而且，若我們願意接受任何最相符於現存資料的模組，無論其複雜度或敏感度如何，我們就犯了好幾個錯誤。首先，我們忘了因果關係，對資料造成損害，過度調整的模組，無法解釋任何事。

再者，我們忘了資料（或資料收集）有其界限，但這個世界一直在變動。想一下從現在開始試著預言世界氣侯 200 年的問題。歷經一段長久的時間，有幾個高辯識度的關鍵證據，也就是來自化石記錄與冰核（ice cores）的全球溫度資料。氣候學家也可以從日誌與年輪推論出本地溫度與降雨量，但精確度卻大有不同。18 世紀的 storm glasses 和20 世紀帶有 GPS 的氣象氣球是不一樣的。而且，誰知道導致 21 世紀氣侯事件的同樣一套相互作用是否在 20 世紀也有相同結果。

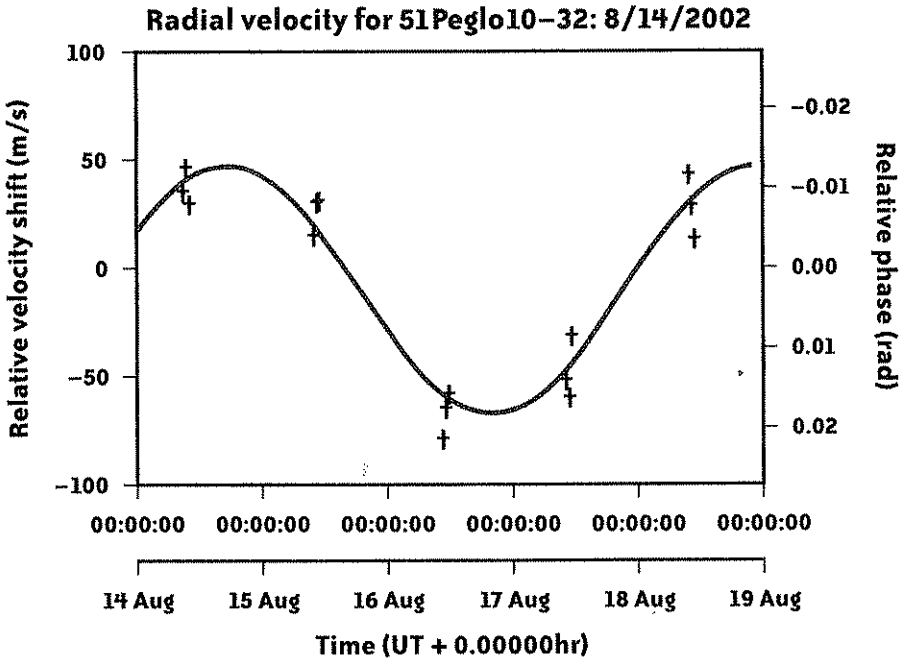

圖 13-7. 太陽系外行星的識別模組

類似的情況，1914 年的 Ford Motor Company 和 1975 年和今日的 Ford 也不會一樣，但仍有許多財務模式假定市場景氣最後的循環動態亦能夠解釋其未來效能（而模組也因考量的相關時期不同，而作出完全不同的假設）。值此之故，風險分析模組可能只有些微（可接受範圍）偏離大多數時期，不過當「預期之外的事」（例如房市崩潰時期）發生時就可能整個瓦解。

好的科學家對劣質模式的危險會有所警覺，但很難不被太適切以致於不真實的情況所誘惑。就拿 Moody（Dwyer）2005 年討論某單位使用超適模組的經驗報告為例（已作了修正，但—其實是事後諸葛—是以不夠大量的方式）：

> 當新的模組化方法論的產出大幅增加時，是該有某種程度的懷疑。它通常是搭配資料收集機制的結果，而非真實底層行為的關係。此外，這些問題在之後都很清楚，但當我們在處理一般預先模組化資料整理過程中卻沒發現…
>
> 最好的情況下，超適的動作會導致模組產生不必要的複雜情況且對使用者失去公信力。最糟則可能造成投資組合的風險評估產生系統性錯誤。

除了預言之外還有很多不得不建置模組的理由，包括像是探索場景與闡明假說，想瞭解最詳盡的內容，可以參考 Joshua Epstein 在 2008 年的隨筆「Why Model？」

6. 機率無法直覺理解

這是統計學家確定的另一個推動口號，也是個很好的藉口。統計學家不厭其煩地發明新玩意證明表面上看來是常識的答案，其機率可能不正確，且條件與連結的機率亦非直覺。當數學家與醫生們被這些玩意愚弄時他們就特別開心。

在某個美國城市裡，一百萬裡大約有 1000 個（或說 0.1%）居民 HIV 呈陽性。

診斷 HIV 的新測試有 1% 的失敗率，一百次檢驗裡會有一個，會出現有 HIV 的人遭不正確的診斷 HIV 呈陰性的個案，而有 1% 的次數會不正確地診斷出某個有 HIV 的人其 HIV 呈陰性。

假設某個人作了測試，被診斷出為 HIV 呈陽性。他有多少的機會感染此病毒呢？

很多人都會回答他有 99% 的機會感染病毒，因為測試有 1% 的失敗率。事實上，由於染病的人口比率太小，任何一個人得到這個病的機會 —— 即便是診斷後 —— 很低：只有 9.9 %。（在 999,000 位 HIV 呈陰性的居民中，有 9,990 會被告知他們感染病毒，而 HIV 呈陽性的居民中有 990 個人是真的呈陽性。在確認陽性的診斷下，真正 HIV 呈陽性的機會是 990/9,990，或 9%）。

醫生 —無論如何，未經證實 —全軍覆沒。

在很多情況下，先驗值不會消失。當我們用資料回答問題的時侯，並不知道哪些證據除外，還有如何權衡應該含括的部份。Daniel Kahneman —孜孜不倦的概念命名者 —稱此為基數謬誤（Base-rate fallacy）。

7. 機率全都無法直覺理解

不是只有機率理論難以掌控，個別機率亦稍縱即逝。在缺乏因果關係的解釋將事件與一組結果串連起來的情況下，個別的部份就仰賴過去的觀察才能推估其機率。而觀察則通常是以一種帶有偏見的方式收集（特別是當它們是透過經驗產生，但多半是經由實驗進行收集的時侯），而且相當難以文件化、一致化、衡量、保存與查詢。

8. 真實世界不會產生隨機變數

一開始，地球上沒有 *form* 和 *void*。然後 *Fisher*【譯註：統計學家】說，「讓我們有 *z* 分數（*z-score*）和變異數分析（*ANOVA*；*Analysis of Variance*）吧」接著 *z* 分數就存在了。之後 *Fisher* 覺得回歸不錯，就把它從沒有統計意義的地方抽取出來讓它具統計意義。

統計學家的創新似乎對人們的影響很大，我們得時時記得這些並非自然法則。

想像一個平行宇宙，其實際存在統計學上的重要臨界點已設為（隨意，在我們的宇宙下）p=0.01 或 p=0.06，而非當下的 p=0.05。想想被核准與禁止的藥物 —在環境變數與健康影響間錯置正確性的例子 —你可以省下大把汽車保險的錢。

在我們非 Fisher 式的世界裡，沒有這樣獨立的隨機變數。事實上有許多東西都高度相關。好的實驗會在盡可能的範圍下，控制其相依性，但從屬性就很難定位了。一如我們最近所學到的，假定分散的事件（例如屋主拖欠其抵押貸款）獨立的，而在沒那麼迫切需要的情況下，以此假定（例如可買賣的金融商品分期）蓋大房子，就可能犯下大錯。

預言市場以及群體決策程序的運作出奇地好，在某些情況下，比一群專家的評估還好。然而，當資訊湧入與相依性加入系統時，它們就會瓦解（Bikhchandani et al. 1998）。

9. 資料並非獨一無二

在真實世界裡作決策時，資料會有很多種型式。少有資料是整理好並打包成做好標籤的試算表或矩陣的檔案，反而我們經常需要透過主觀與定量的資訊作出結論。

以決定是否要借款給某個線上的人為例（為了獲利，此乃貸款市場建立的原因之一）。同事與我，在顯示了各種模型（混用模型、神經網絡、決策樹、回歸）的點對點平台 Prosper.com 以 350,000 個資料集作貸款融資與償還的施行分析，預知誰可取得貸款，誰又將即時償還，準確率大約只有 75%。數量龐大的資料 —針對網路各個成員，含括超過 100 項個人財務健檢指標 —可以餵算式進去處理，但對於哪個申請人將得到好的待遇，哪一個又得不到借貸仍是無法確定的。

我們的模型以試著量化主觀功能達成部份改良。當個人決定是否借錢給網路會員時，貸方（與銀行不同）可以考量一些「柔性」因子，借方的用途敘述、隨附的影像、拼法、語法與其他個人資料的資訊。為了將這些功能併入模組裡，我使用人工方式（從 Amazon

的 Mechanical Turk）為 Prosper.com 會員的影像進行編碼，先從內容開始 —是否有描繪個人、家庭與交通工具等等的影像 —然後接著是「可信賴性」分數，指的是有關「你願意借錢給這個人嗎？」問題的答案。

不過這個模式仍有不足之處，社交因子會以非預期的方式影響借貸動態。這和我們的假定是相矛盾的，借款決策並非單獨決定。甚至，在投票時會有些許群體行為的跡象，貸款者會跟著貸款者的腳步，而當更多投票聚積在借貸上時，就會讓每單位時間的投票加速。

就算有這些以及其他的社交因子可以用來考量，但仍有許多貸款者作出次優的決定。Prosper 理論上是趨近於完美的資訊市場，幾乎所有人皆可存取網站 API 並重作我們的分析。有許多貸款者仍持續接受低報酬的高風險投資，統計在一定預期收益下，做出極差賭注的數字很令人驚訝。就算有完備的資訊（與提供主觀資料的適切代理人），也並非一定是直接透過資料作決策，反過來，資料也只能解釋部份的人為決策。

10. 觀察者不能不密切注意資料

最後，就算在可能提供強而有力因果關係解釋的地方，由 Fisher 以及（若我們的學生真的想要的話）Bayes 最聰慧的學生以公正的方式收集資料並謹慎進行模組化，然後解釋其模組的變異與驗證（且對其結果仍保持懷疑），仍然還是有相當多認知上的偏見籠罩著我們的思考。在真實世界裡，我們盡可能處理這些假性機率。

一如統計學家在他們部落格裡叨叨絮絮提及的行為經濟學家在他們的領域下也有惡名昭彰的紀錄史。敘述謬誤（narrative fallacy）、確認偏見（confirmation bias）、選擇的矛盾（paradox of choice）、風險不對稱（asymmetry of risk-taking）、基數謬誤（base rate fallacy）與雙曲貼現（Hyperbolic discounting）先前都已提過。心理學家編列其他部份，範圍從定錨（在做決策時，過度依賴單一最近資料點）到 Lake Wobegon 效應（佔一半以上人口數的個人確信自己優於水準之上的現象）都有。

當這些效應有完整的說明文件後，我們就可以開發工具與直覺理解的知識協助取得資料的表面價值（我工作的一部份便是著重在開發財務決策的工具程式）。

在某些意義層面，解決方案其實很簡單，若你不瞭解資料的限度，它其實幫不上什麼忙。