嶄新展示

——立法會數據一目了然

撰文：胡辟礫、巢恬逸

引言：

端傳媒數據新聞作品《20 萬條投票紀錄帶你解碼香港立法會》搜集本屆立法會議員由 2012 年至 2015 年的立法會投票記錄，分析他們的投票取態和規律，並結合文字、圖表、視頻等多元的手法，將海量的信息以新穎、簡明、生動的方式呈現在公眾面前。這是一次頗為成功的數據新聞嘗試，對於今後的數據新聞發展具有重要的參考價值。

該報道獲 2016 年亞洲出版協會 (SOPA) 卓越資料圖像優異獎。

背景

「等埋發叔」、嫻姐「O 嘴」、田氏兄弟舌戰……2015 年 5 月香港立法會否決政改方案的投票場景仍歷歷在目。這些「甩轆」（掉鏈子、出意外狀況）場面以及其後各路政治人物的回應娛樂性太高，至今令人難忘。議事堂娛樂性雖高，監察議員表現卻是選民必定要做的事。但政治人物發言往往是騎牆之見，

隨風而動，選民很難了解議員們的真實想法。

正所謂：聽其言不如觀其行。議員們嘴上說了甚麼不重要，重要的是他們的行動——投票選擇能真實地反映他們或他們所代表的派別對於議案的想法。傳統的立法會議員派別劃分依賴對其政治背景及言論的考察，缺乏穩固的證據支撐。而且在各個派別內部，議員形象非常單一，泛民、建制、無黨派的簡單劃分抹殺了各個派別內部議員的政見分歧。在團結的表面下可能暗流湧動，各自為戰。為了檢驗民眾對於各政治派別及議員的既定印象，立法會投票紀錄作為各政治派別及議員對於特定議題的最終的意見表達，進入了我們的視線。

● 會議日期,投票結果及會議過程正式紀錄

日　期	會議議程	會議紀要/投票結果	會議過程正式紀錄(中文版)	會議過程即場紀錄本	網上廣播
2015 年 10 月 14 及 16 日	會議議程	會議紀要 投票結果 (PDF) (XML)	會議過程正式紀錄 (中 文 版) 14.10.2015 16.10.2015	會議過程即場紀錄本 14.10.2015 16.10.2015	
2015 年 10 月 22 日^^ (上午9時30分至11時)	會議議程	會議紀要	會議過程正式紀錄 (中 文 版)	會議過程即場紀錄本	
2015 年 10 月 28 及 29 日	會議議程	會議紀要 投票結果 (PDF) (XML)	會議過程正式紀錄 (中 文 版) 28.10.2015 29.10.2015	會議過程即場紀錄本 28.10.2015 29.10.2015	
2015 年 11 月 4 及 5 日	會議議程	會議紀要 投票結果 (PDF) (XML)	會議過程正式紀錄 (中 文 版) 4.11.2015 5.11.2015	會議過程即場紀錄本 4.11.2015 5.11.2015	
2015 年 11 月 11 及 12 日	會議議程	會議紀要 投票結果 (PDF) (XML)	會議過程正式紀錄 (中 文 版) 11.11.2015 12.11.2015	會議過程即場紀錄本 11.11.2015 12.11.2015	
2015 年 11 月 18 及 19 日	會議議程	會議紀要 投票結果 (PDF) (XML)	會議過程正式紀錄 (中 文 版) 18.11.2015 19.11.2015	會議過程即場紀錄本 18.11.2015 19.11.2015	
2015 年 11 月 25 及 26 日	會議議程	會議紀要 投票結果 (PDF) (XML)	會議過程正式紀錄 (中 文 版) 25.11.2015 26.11.2015	會議過程即場紀錄本 25.11.2015 26.11.2015	

2013年底，香港立法會開始發布每次會議投票的XML格式數據——這是一種機器可讀、結構化的數據格式。（端傳媒圖表）

數據收集

做數據新聞，巧妙構思離不開原始數據支撐，能否搜集到扎實的數據，總是最關鍵的一步。

2013 年底，香港立法會開始發布每次會議投票的 XML 格式數據——這是一種機器可讀、結構化的數據格式。我們編寫 Python 腳本抓取這些 XML 文件，抓取過程不到十分鐘時間就完成了。這份數據包括今屆立法會審議的二千九百二十一項議案的詳細信息，如每項議案的動議時間、名字、動議人和七十位議員的投票紀錄，甚至精確到議員按下表決器的時間。

換作以前，公開的投票記錄都存放在 PDF 檔案中，搜集數據全靠人工抄寫，集齊一屆立法會的投票信息，且精確到這麼細的力度，需要至少一個人工作四個月。

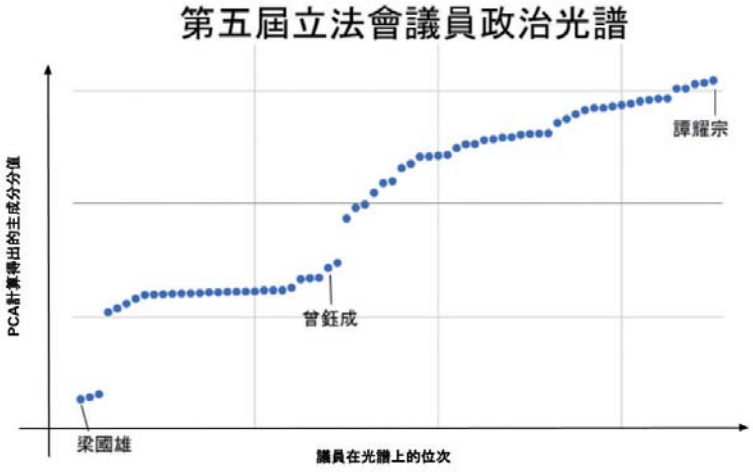
數據分析

這個龐大的數據庫可以回答很多問題，誰的提案最多？誰的提案通過率高？誰出席的次數多？……不過我們最關心的，是投票紀錄顯示出的政治派別以及每位議員的偏好：誰更激進？誰更保守？哪些人自成小圈子相互支持？帶着這些問題出發，我們開始了漫漫的數據摸索之路。

我們蒐集到的數據庫中，總共有二千九百二十一個議案，每位議員對每個議案可能出現五種不同結果：贊成、反對、棄權、出席或缺席。對於部分議案，議員們分歧不大，每張票的指標意義就相對較小——一致的票無法區分議員的投票傾向。相反，對於某些議案，議員們分歧嚴重，此時議員們投出的票對於區分議員的投票傾向具有重要價值。

我們用一種叫主成分分析（Principal Component Analysis，簡稱 PCA）的機器學習方法，將這二十萬票的價值「綜合」起來，計算出了一個虛擬的「超級議案」（即「主成分」），使得議員們在這個「超級議案」上的分歧最大。這意味着，議員們對這個虛擬議案最為態度鮮明，這可以成為我們區分七十個人立場的依據。

根據不同議員對這個「超級議案」的反應，我們為每位議員計算出一個分值，記為 PCA 計算得出的主成分分值。兩位議員得到的分值愈接近，則說明他們的投票傾向也愈接近。所以，按照這個分支由小到大排序，就形成了一條數據驅動的「政治光譜」（如下圖所示）。這條由純數據分析得到的「政治光譜」從投票的角度印證了「建制派」與「泛民主派」的分野。這條光譜的兩端，分別是梁國雄和譚耀宗，這說明兩人的投票風格差異最大，和其他議員相比，他們的投票風格更加激進。在這條「光譜」中間，是從不投票的曾鈺成。愈靠近曾鈺成的人，他的投票風格就愈溫和。愈遠離曾鈺成的人，他的投票風格就愈激進。



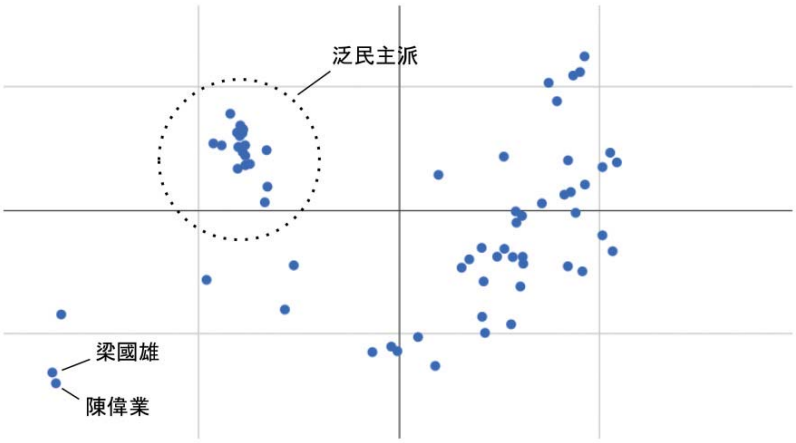
這條由純數據分析得到的「政治光譜」從投票的角度印證了「建制派」與「泛民主派」的分野。（端傳媒圖表）

數據可視化

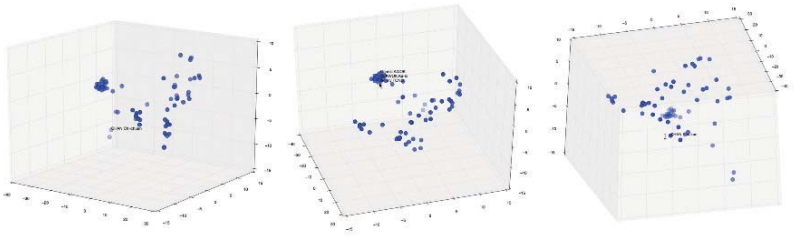
製作這條光譜固然是項目中最複雜的一步，但它只是數據分析和可視化的開始。

我們迅速繪製了二維散點圖和三維關係網，以呈現派別內部的親疏關係。又找來兩三個同事給建議，大家認為，兩張圖表都過於「geek」，一般讀者很難立刻讀懂，甚至有可能被嚇到。

立法會議員投票傾向之二維散點圖

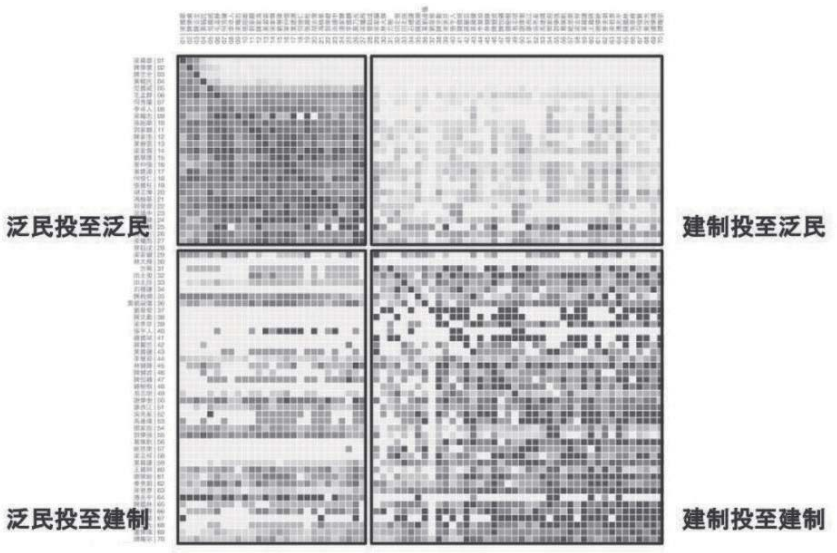


立法會議員投票傾向之三維關係網



我們又想到，熱度圖是表現每兩個個體關係的好辦法，不如用它來看看議員的相互支持程度。於是我們做了一張 70X70 的熱度圖，計算出每位投票人對每位提案人的支持程度，用格

子的深淺表示，顏色愈深代表愈支持，顏色愈淺代表愈不支持。再將議員按照計算出的投票傾向光譜排序。

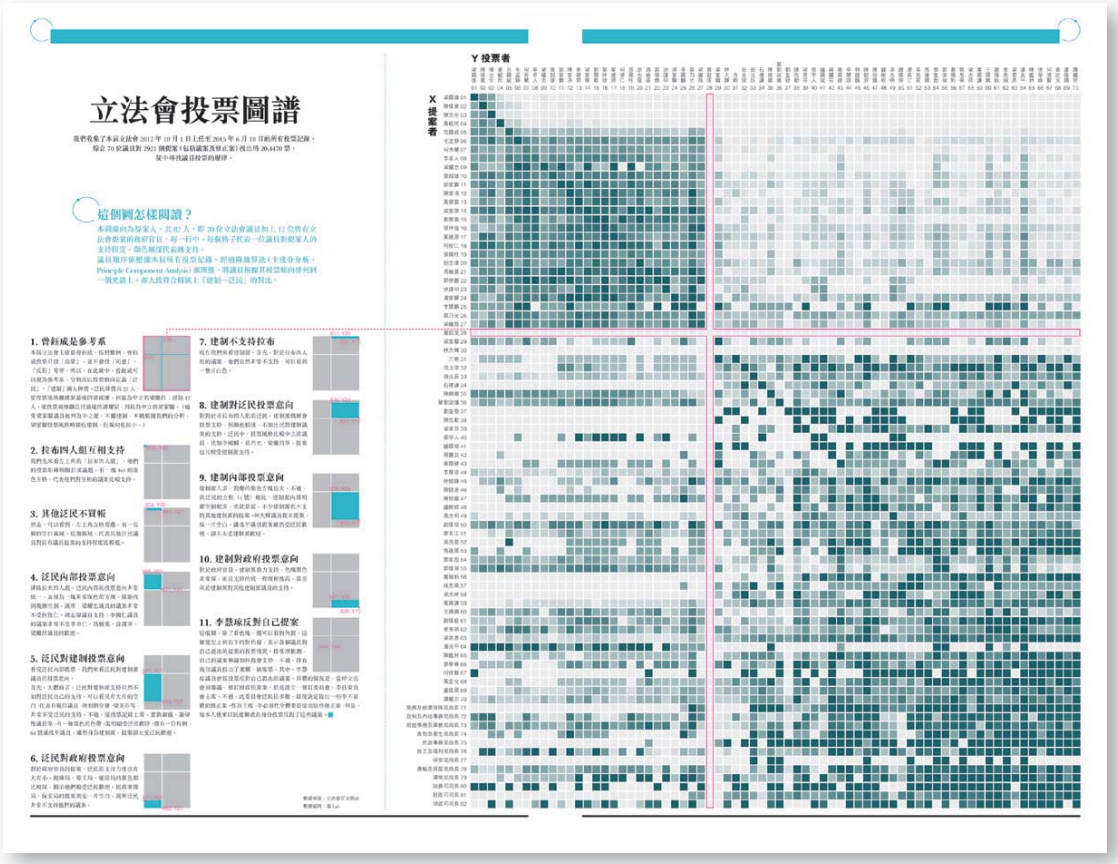


分區熱度圖，圖中四塊區域顏色從深到淺依次是：泛民投至泛民、建制投至建制、泛民投至建制、建制投至泛民。(端傳媒圖表)

這張密密麻麻的熱度圖，包含着大量「待開採」的信息。我們先分析整張圖能否呈現規律，很明顯，圖中四塊區域顏色從深到淺依次是：泛民投至泛民、建制投至建制、泛民投至建制、建制投至泛民。這說明，除了「拉布四人組」（左上角四人，梁國雄、陳偉業、陳志全和黃毓民），泛民內部的投票意向非常統一，與之相比，建制內部分歧更大一些；跨陣營的支持關係上，泛民對建制議案的支持程度大過建制對泛民議案的支持程度。

找出規律後，再轉換策略，仔細檢驗每一個格子去尋找有沒有異常情況——這可能隱藏着新聞點。接下來的幾天，我們對其進行地毯式「掃描」，果然得到很多有趣的發現，很多都寫進了最後的報道和動畫短片中。

接着，我們在辦公室裏做起「用戶測試」，邀請更多同事來讀圖。這一次，許多同事反饋說，熱度圖能提供非常詳盡的



立法會熱度圖及導讀。（端傳媒圖表）

信息，尤其是聽完我們的介紹，他們能夠很快找出規律和異常情況。不過，編輯部同事平日就浸泡在政治新聞、社會議題之中，而我們的讀者，也許沒有這麼充足的領域知識，他們能夠耐心地讀完甚至摸索這張圖嗎？並且，熱度圖是一種理解「門檻」較高的圖表類型，處理不好就會讓讀者頭暈眼花，有甚麼方法能既準確又淺顯地將這些有趣的分析結果呈現給讀者呢？

回顧用戶測試過程，我們發現，每次有新用戶來看圖時，我們都會介紹許多背景：橫軸代表甚麼、縱軸代表甚麼、圖是怎樣製作的……重複幾次後，「拍條短片」的想法擊中了我們。短視頻既可以吸引眼球，又能起到「教育」作用，迅速教會讀者怎樣看圖、怎樣摸索。最後，我們耗時四個星期製作了一條三分鐘的動畫短片，作為報道的開篇。看過短片並有興趣繼續探索的讀者，可以閱讀我們的調查報道，甚至在交互頁面上自

行探索，發現更多故事。

在短片末尾，我們拋出了兩個問題：「為甚麼李慧琼會給自己投反對票？」、「為甚麼建制派議員潘兆平獲得來自泛民的支持比來自同儕的更多？」，這成為後續文字記者追蹤的重點，她通過聯繫議員本人和翻閱立法會議事記錄，探詢原委。

最終我們的報道以一條動畫開頭，配合文字報道，為讀者解密二十萬個立法會投票記錄。大數據分析搭配生動圖像，這篇報道在網絡上迅速傳播，兩週內即獲得超過二萬次瀏覽。報道亦收到來自立法會議員和本地技術社群的肯定。

評價

這是一次頗為成功的、在新聞中運用數據挖掘和數據可視化的探索，靜態圖表、交互頁面和視頻動畫相互配合。獨到的視角讓讀者能夠從一個全新的角度審視立法會議員和香港的政治生態，撥開政治的迷霧重重，還原一個最真實的香港立法會。

新聞報道有很多方式，數據新聞才初露鋒芒。身處於這個碎片閱讀的「看圖」時代，簡單的信息圖表佔據了讀者大量的注意力，而基於數據挖掘、可視化的報道模式，鮮有所見。

立法會的開放數據很豐富，可以做的研究與新聞作品遠不止於本篇分享的「投票熱力圖」。比如，若能夠把提案和修正案分開分析、或把不同議題（政治改革、經濟、民生、人權、環境等……）下的提案分開分析，一定會帶來更多發現。又或者，可以製作互動遊戲，讓讀者對議案投票，再對他們的座標與其他議員的座標做可視化展現，這對未來的投票也帶來幫助。除去投票數據，立法會也有公開的會議記錄、會議視頻，若通過自然語言處理，也能分析出更多的洞見。

提示：

- 與翻查資料、親臨現場和採訪關鍵人物一樣，挖掘數據亦是獲取新聞線索的方式。數據作為消息來源的一種，過去被新聞人忽視了，如今又被捧得過高。在我們看來，各類消息來源同等重要，組合起來一齊使用，效果更佳。
- 在這信息爆炸的年代，數據很多、也很少。多數情況下，我們接觸到的是結構不規整、格式不友好的數據，需要有人把它們整理好、開放給大眾使用。政府有這個責任，卻沒做好，如今許多媒體和組織接過來做。用開放和分享的態度來做這件事，事半功倍，慢慢地，愈來愈多的人會加入到開放數據的浪潮中，這是我們公開所有數據類新聞項目原始資料與分析方法的初衷。

其他：

如果你想探索這張 70X70 的交互熱度圖，獲取本文的開放數據與開放源碼資料庫，請掃描以下 QR code 前往項目網站：

作者簡介：

胡辟礫，前端傳媒首席技術官，入行一年，專責數據分析。

巢恬逸，前端傳媒技術部協調專員，入行一年，專責調研。