行政院所屬各機關因公出國人員出國報告書

(出國類別：研討會)

探討大數據於央行政策之運用講習與研討會

服務機關：中央銀行

姓名職稱：繆維正 副研究員

出國地點：印尼

出國期間：107.7.22-107.7.27

報告日期：107 年 10 月

摘要

隨著網路科技及新的統計分析技術的發展，各種大數據已成為經濟政策制訂者做決策與分析的重要工具。央行所使用的大數據，除了傳統的高頻的經濟、金融指標之外，尚包括網路商店的物價資料，可用來編製與預測物價與通膨率；網路搜尋資料（如 Google Trends)，可用來預測各種總體經濟變數，如 GDP 與失業率等；文字資料，例如新聞、企業與銀行之財務報表、社群媒體、各種調查資料及其他央行的政策會議記錄，研究者可藉此瞭解一般民眾對市場的看法及新聞媒體關注的主題；金融市場的網路分析則可探討銀行間資金流動的關係，可由此分析各銀行之流動性風險，並確認金融體系中最重要的銀行。

大數據的發展需要跨領域的團隊合作，需要投入人力、時間、並建立適當的資訊基礎建設。本行可針對行內需要及其他國家已著手的研究主題開始進行研究，並且多關注國際間的研究成果，考慮與國際及國內之學界進行合作，俾利央行政策之制訂。

目錄

1 前言

職奉派於本（107）年 7 月 22 日至 7 月 27 日參加由印尼央行與國際清算銀行（BIS）Irving Fischer Committee 於印尼峇里島舉辦之「探討大數據於央行政策之運用講習與研討會（Workshop and Seminar on Big Data for Central Bank Policies)」研習課程。課程內容包括大數據之概念、機器學習演算法、文字探勘、總體經濟數據之即時預測及金融市場之網路分析。研討會中各國中央銀行共有 7 篇大數據相關的研究論文進行報告與討論，總體經濟方面包括即時預測、即時經濟指標建立、以文字探勘建立即時情緒指標，金融市場方面包括信用風險分析、主權債券與信用違約交換（CDS）市場等等。本次參加研習學員包括各國中央銀行代表共 68 人，講師包括 BIS 統計局局長 Bruno Tissot、IMF 統計局副局長 Gabriel Quiros-Romero、英格蘭銀行數據分析部門經理 Paul Robinson、德國央行統計局副局長 Robert Kirchner、法國央行統計局局長 Renaud Lacroix、ECB 資深顧問 Per Nymand-Andersen、日本央行研究與統計局副局長 Naruki Mori、新加坡貨幣管理局（MAS）首席資料科學家 David Roi Hardoon、荷蘭央行資深政策顧問 Iman van Lelyveld，及前芬蘭央行與紐約 Fed 研究人員暨 Financial Network Analysis 之共同創辦人兼 CEO Kimmo Soramaki。學界則包括麻省理工學院（MIT）管理學院教授 Roberto Rigobon、牛津大學（University of Oxford）經濟系教授 Stephen Hansen、美國加州 Santa Clara University 教授 Sanjiv Das。本文架構如下：第二節敘述各國使用大數據的現況及專家的看法，第三節與第四節說明目前央行大數據運用之其中兩項重要主題—大數據在物價預測與編製之運用及文字探勘，第五節為結語與建議。

2 各國使用大數據的現況及專家的看法

2.1 各國使用大數據的現況

1. 機器學習演算法：機器學習演算法係新一代的數據處理方法，在部分情境下，處理數據的效能優於傳統統計方法，可運用於以下方面：(1) 偵測錯誤之申報統計資料：荷蘭央行運用機器學習演算法偵測錯誤或異常之申報統計資料。傳統上統計資料的正確性需以人工判斷，機器學習演算法可偵測離群值（outlier），並探討離群值為數據錯誤或狀況異常。(2) 區分風險程度高的金融機構，強化金融監理：希臘央行使用深度學習與極值梯度（Extreme Gradient Boosting）演算法預測金融機構違約之可能性，其預測之準確性優於傳統的邏輯式迴歸。(3) 主權債券市場及金融監理：香港金融管理局（HKMA）以機器學習演算法發現主權債券（sovereign bond）市場可由經濟與金融指標，運用機器學習演算法進行預測。由此可評估主權債券之風險程度。(4) 編製物價資料：日本央行在運用大數據編製物價指數的過程中運用機器學習演算法。由於編製高頻的物價指數資料需要找出本月與上個月同時存在的物品，BoJ 使用機器學習演算法針對本月與上月的物品清單進行配對。(5) 即時預測：紐西蘭央行運用 550 個經濟指標進行即時預測，發現機器學習演算法之預測準確度優於傳統方法。英格蘭銀行亦有類似研究結果。

2. 網路模型：(1) 金融機構壓力測試：荷蘭央行利用網路模型，利用 CDS 市場交易資料，找出金融機構在全體金融市場的網路關係，做壓力測試，瞭解各金融機構的風險程度。(2) 洗錢之偵測：馬來西亞央行建立資金流出、流入之資料，建立匯款之網路模型，偵測可能為異常之匯款，偵測洗錢犯罪行為。

3. 文字探勘：印尼央行利用新聞之歷史資料庫進行文字探勘，找出經濟擴張與緊縮的關鍵字，由文字探勘的結果估計媒體對央行政策的預期，結

果顯示該預期結果與 Bloomberg 調查之相關係數高達 0.73。

4. Google Trends（以及 Google Correlate）：(1) 泰國央行使用 Google Trends 及 Google Correlate 找出與經濟指標相關的搜尋關鍵字，並找出與經濟指標相關的主成分（principal component），發現家戶所得、情緒指標、民間消費、失業率，與主成分之同期相關程度介於 0.7 與 0.83 之間。MIT 教授 Roberto Rigobon 指出，由於官方資料發布時間較落後，即使是同時指標，仍對政策制定者有參考價值。(2) ECB 運用 Google Trends 預測汽車銷售，發現加入 Google Trends 的預測模型之預測結果優於傳統基準模型。(3) 西班牙央行使用較高頻的資料，包括文字探勘得到的經濟政策不確定性指數（Economic Policy Uncertainty, EPU）、信用卡消費資料、ATM 提款資料、Google Trends 預測民間消費，發現適當的變數組合可勝過傳統預測模型。

5. 高頻、即時預測與即時資料之編製：馬來西亞央行利用網路上的房價資料編製房價指數，解決統計指標落後的問題。

2.2 專家對發展大數據分析的看法

1. 大數據的發展係跨領域、跨單位之合作：荷蘭央行統計局資深政策顧問暨阿姆斯特丹自由大學（Free University Amsterdam）金融系教授 Iman van Lelyveld 指出，由於大數據在央行可運用的範圍相當廣泛，包括貨幣政策、外匯、金融穩定、金融檢查等，該行發展大數據係行內跨單位之任務編組。MAS 首席資料科學家 David Roi Hardoon 與美國加州 Santa Clara University 財金系 Sanjiv Das 教授指出，由於要瞭解經濟、財金領域大數據問題需要該領域的專家，分析技術則需要數學、統計與資訊人才，因此大數據的發展需要結合經濟與數理人才一起合作。

2. 發展大數據需要編製新的統計資料：BIS 統計局局長 Bruno Tissot 指出，

傳統上中央銀行多為「資料使用者」，但金融危機後逐漸成為「資料編製者」。央行為了達到經營的目標，必須編製符合監控總體經濟與金融監理用途的新資料。編製資料的同時，必須與國內其他機構或國際組織合作。

3. BIS 與國際組織可提供各國央行的幫助：BIS 與其 Irving Fischer Committee 為全世界央行資訊交流的平台，任務之一係協助央行做統計資訊分享。此外，各國央行若有新的統計方法或經濟分析之研究結果，亦可投稿至 BIS 與 Irving Fischer Committee 以及其每兩年一次在 BIS 總部舉辦之研討會，經過內部之經濟學家與統計學家審核後可作為 Working Paper 刊登於其網頁。

3 高頻資料與大數據在物價編製與即時預測之運用

物價係所有經濟參與者，包括消費者、企業、金融市場投資者與政府共同關注的議題。包括我國央行在內，穩定物價為多數國家央行的重要職責之一。央行必須瞭解物價可能走勢，以制定適當的貨幣政策。由於央行採行影響物價的政策需要一段時間才能發揮作用，物價之即時資訊與預測更顯重要。

消費者物價指數反映一個典型消費者所面臨消費價格隨時間的改變。因此，該指數包含的項目為一個典型消費者日常生活必需消費的項目，包括食物、衣著、房租、醫療、教育、能源等等。粗略地說，物價指數即這些項目物價變化的加權平均。單一項目價格的改變所造成整體物價指數變化的大小，係與該項目在總指數中的權重有關。

物價指數的編製係各國國家統計局的職責所在，我國編製的機關為主計總處。包括我國在內，大多數國家每月發布一次物價指數。傳統上，各國編製物價指數係根據調查資料。近年來，由於電子商務的迅速發展，網路商店成為實體商店以外，消費者購物的重要選擇。因此，網路商店的價格已逐步成為各國國家統計局編製消費者物價指數之重要資料來源。此外，相較於傳統

實地訪查，由網路商店之價格編製物價指數可享有低成本、高效率的優點。

由於網路爬蟲（web scrapping）技術[1] 的發展，上網調查物價資料可由機器自動進行，並且以電腦計算該類的物價指數。網路商店無法包含物價指數部分服務價格資料，但其包含資料範圍已相當廣泛，對追蹤消費者物價指數的走勢相當有幫助。國際間以網路爬蟲搜集物價資料最有規模的研究計畫是麻省理工學院的「巨量價格數據計畫（Billion Price Project, BPP)」。該計畫搜集多國網路商店的資料，除了追蹤物價指數外，該資料亦用於總體經濟與貨幣政策的相關研究上，例如價格僵固性的研究，對貨幣政策的執行至為重要，惟價格僵固性的程度必須根據個體物價資料才能準確衡量。

前述的物價資料來源，包括實體訪查與網路資料的搜尋均專注於物價「編製」上。物價指數的「預測」方面，傳統上採用月頻率或季頻率的資料做預測，由於網路資訊等大數據的出現，這些較高頻率數據亦可用來預測通膨率。這種預測的方法亦可推廣至預測其他總體經濟變數，例如 GDP 與失業率等等。

3.1 以當期高頻資料進行通膨即時預測

本節敘述之物價即時預測，其想法與傳統計量模型預測類似，惟兩次傳統預測之間，會有新的經濟與金融指標數值出現，高頻預測係納入這些新的數值，不斷更新預測值。

3.1.1 傳統非即時預測與即時預測之差異

傳統通膨預測多假設預測當期之期初，所有當期前之經濟數據都可取得，並作為通膨預測的基礎[2]。這些解釋變數通常包括 GDP、物價、貨幣市場變數與匯率等等。預測模型之績效衡量亦根據此一假設，計算各期樣本外預測誤

[1]網路爬蟲或稱網路機器人（internet bots 或 web robots)，即搜尋引擎瀏覽大量網頁的技術。

[2]其他主要總體經濟變數之預測亦為如此。

差，即預測結果與實際公布數據的差異，取其平均值，平均誤差愈小，代表預測績效愈好[3]。

由於可用以預測通膨之經濟變數可能高達上百個，若能充分運用這些變數內涵之訊息，可能增進通膨預測之績效。傳統計量經濟學上處理大量變數的方法以主成分分析（principal component analysis）與因子分析（factor analysis）為主，惟這些方法通常亦假設預測期初知道當期前之所有經濟數據。實務上，新的經濟數據公布有時間落差，且在同一期中不同時間點發布，若能即時將這些新的資訊納入預測的訊息範圍，可能增進預測的績效。以台灣之通膨為例，若吾人欲在 2018 年 5 月 1 日發布當月之通膨預測值，現有之資料包括 3 月的通膨值與貨幣市場數據、4 月的市場利率與匯率，及上年第 4 季之 GDP 值。到了 5 月 7 日，4 月之消費者物價指數（CPI）發布，吾人可能因此修正 5 月之預測值。5 月 24 日發布貨幣市場數據，5 月 25 日發布第 1 季之 GDP 值，則可再修正當期或未來之預測值。此外，瞬息萬變的金融市場隨時都蘊含新的資訊，其中影響 CPI 最大的金融市場變數為油價（oil and gasoline price）及其他能源價格。吾人可在新資訊發布的同時，更新預測值。

3.1.2 美國 Fed 與 ECB 即時預測之經驗

以美國為例，Cleveland Fed 每上班日美東時間 10 點整公布當日四種物價指數之即時預測，包括 CPI、核心消費者物價指數（core CPI）、個人消費支出物價指數（PCE）、核心個人消費支出物價指數（core PCE）等四種物價指數之通膨率（參見Knotek and Zaman (2017) 與Knotek and Zaman (2016)）。該預測根據十個數列，其中八個數列為月資料，包括勞工統計局（BLS）編製之 CPI、核心 CPI、食物 CPI、家內食物 CPI、油價 CPI，及經濟分析局（BEA）編製之 PCE、核心 PCE、非飲食場所購買之食品飲料 PCE。另兩數列為零售能源價格，係能源資訊署（EIA）每週一發布之數據，及布蘭特原油價格，為日資料。歐洲央行（ECB）採用之資料大致相仿，主要差異在於加入經濟合作暨

[3]誤差平均通常以均方根誤差（RMSE）計算。

發展組織（OECD）編製的原物料價格指數。Fed 採用之模型大致分為四大部分：第一部分預測核心物價，第二部分預測食物物價，兩者皆以近期價格預測即可獲得良好結果。第三部分即時預測油價，係假設當日油價可描述近期油價之走勢，再做季節調整。Modugno (2013) 的模型架構則較複雜，包含因子模型與狀態空間模型，由於外生變數的資料頻率與物價並不一致，該文假設較低頻率的觀察資料有些樣本點為缺失資料（missing data），由計量方法補齊這些缺失資料後，再以該架構預測通膨。

一般而言，月（季）末之即時預測較月（季）初佳，係由於期末之即時預測根據的資訊較期初多。因此，若其他預測機構發布預測時間在當期內的時間點較晚，則可能有較優異之預測績效表現。平均而言，Fed 之即時預測準確度優於其他即時預測，然而並不是每一次的預測結果皆較佳。然而，未預期的衝擊無所不在，可能使實際通膨值遠離即時預測值。Fed 與 ECB 研究皆驗證了高頻資料在通膨即時預測的貢獻，ECB 之研究更顯示高頻資料不僅有助於提升即時預測的準確度，亦使一年期之通膨預測更精確。

3.1.3 對通膨預測最有貢獻的解釋變數

由於可能有助於通膨預測的資料相當多，若研究者能確認對預測最有幫助的變數，不但可瞭解通膨的動態，更可簡化模型之建構。確認邊際貢獻（marginal contribution）最大之變數，可根據經驗，亦可透過分析過去預測結果確認。由於核心 CPI 與核心 PCE 本身變動幅度較小，且其即時預測根據之參考資料亦較少，因此，兩者即時預測的變動並不頻繁。油價在 CPI 與 PCE 的即時預測上扮演重要的角色，因此其變動亦深刻影響兩者的預測值。油價本身難以預測，因此除了 CPI 與 PCE 發布時，通膨預測有較大的改變以外，其他時間點之預測改變多由於油價的改變。Giannone et al. (2008) 的研究則顯示，Federal Reserve Bank of Philadelphia Surveys 對預測績效的邊際貢獻相當大。

3.2 以網路商店之價格資訊編製與預測物價指數

傳統上，物價指數的編製係仰賴統計調查員實地訪查該指數內所有成分價格之變化，再加總得到物價指數。隨著網路科技的發展，物價指數內大多數的成分可由網路商店取得。由於網路商店在消費者的消費行為扮演重要的角色，其價格具備相當程度的代表性。這些價格可即時反映市場上商品與部分服務的價格。若吾人可搜集並整理網路上的價格資訊，加總則可獲得即時之物價指數，該指數比一般官方每個月公布一次的價格指數頻率更高。搜集該資料較傳統調查更有效率，一方面可部分替代編製物價指數需要的調查，另一方面，由於其資料的即時性，該方法亦可有效預測官方公布的 CPI，央行可根據這些資訊做即時決策。國際間以網路商店價格編製物價指數的研究計畫中，規模最大的是麻省理工學院（MIT）的巨量價格數據計畫（BPP）。

3.2.1 巨量價格數據計畫

MIT 的 BPP 由該校兩位教授 Alberto Cavallo 與 Roberto Rigobon 主持，主要目的係藉由搜集網路商店的價格，即時編製高頻率之物價指數（Cavallo and Rigobon (2016)）。傳統上，大多數國家物價指數之編製係根據每月調查事先選取的商品與服務項目，加總得到物價指數。然而，這一程序昂貴、複雜，且通常相當緩慢。新品上市、品質改變使編製過程更為繁瑣。此外，調查統計常面臨無回應（nonresponse）的問題。金融危機更凸顯即時資訊的重要。線上價格係一實用的選擇。物價資料散佈在成千上萬個網站中，網路爬蟲技術的發展使研究者能快速搜集網路上的價格資訊，準確、迅速且成本低廉。BPP 蒐集資料之店家為同時具有實體通路與線上通路之商店，例如 Walmart，而忽略僅以線上通路銷售之電商。Cavallo (2017) 後續的研究顯示線上商店的價格具有相當程度的代表性，並建議國家統計機構可以線上價格取代實體調查。該研究團隊將取得的資料運用於下列研究：

1. 提出阿根廷政府虛構通膨數據之證據

以線上價格編製物價指數的構想最先來自阿根廷發布不實的通膨值。到了 2007 年，政府掩蓋實際通膨數值的證據已經相當明顯。Cavallo (2013) 的研究顯示，根據阿根廷政府發布的通膨數據，2007 至 2011 年的平均通膨率為 8%，但線上價格的資訊卻顯示其通膨率超過 20%，與大多數經濟學家的估計、民眾的通膨感受大致接近。

2. 價格僵固性之研究

貨幣政策傳遞效果與價格調整速度有關，因此，藉由個體價格資料瞭解價格調整速度係貨幣政策實證研究之重要議題。由於線上價格資訊可即時取得，研究者能較以往更精確地計算物價調整的速度。Cavallo (2018) 之研究結果指出，物價之變動速度較以往文獻所認知的更緩慢，意即物價僵固程度較以往的認知更高。

3. 實質匯率與國際相對價格之研究

個體價格資料的出現使各國間相同商品之價格比較更容易，因此能藉此探討影響實質匯率的因素與事件。Cavallo et al. (2014) 研究顯示，商品的單一價格法則（law of one price）在貨幣同盟中成立，在貨幣同盟以外的地區則否，即使盯住名目匯率亦無法使名目相對價格相等。Cavallo et al. (2015) 以拉脫維亞加入歐元區前後之時期做研究，顯示該國加入歐元區後，網路商店價格與德國相同商品價格近乎相同的比率由 6% 增加至 89%，價差中位數由 7% 下降至零。Borraz et al. (2016) 以烏拉圭邊境之商品價格做研究，結果顯示相距 30 公里的城市，其物價價差較相距 10 公里的城市大；過去研究的結果多顯示兩者價格差異不大。造成該研究與過去研究不同的原因係過去的研究低估了運輸成本，而該研究亦凸顯出個體商品價格資訊的價值。

3.2.2 政府之物價統計

如前所述，網路商店之價格資訊已成為各國編製政府物價統計之重要資訊來源。以荷蘭國家統計局（Statistics Netherlands）為例，其物價部門已開始由網路蒐集部分物價資訊，例如機票價格、書籍、CD 與 DVD 等等。這些資料可由網路爬蟲的電腦程式蒐集，並儲存於近端資料庫。

由網路蒐集資料的方式有很多種，典型的方法係由腳本語言軟體（script language），例如 Perl 或 Python，再經由搜尋引擎軟體與專用的機器人工具進行資料蒐集。荷蘭國家統計局的資料來源包括四個航空公司網站、一個房市網站及一個無人加油站連鎖店網站。這些資料在夜間蒐集，以降低網站的負荷。

3.2.3 挑戰

由網路機器人蒐集網站資料的過程中，該網站會察覺到外來機器正在蒐集其資料，增加網站負荷，網站亦可設定防止特定網路位址之機器人蒐集其資料。因此，蒐集資料之禮貌（etiquette）甚為重要。一方面，資料蒐集者必須與網站管理者有良好溝通；另一方面，若蒐集之資料為日資料或頻率更低之資料，可在夜間，網站流量較低時蒐集，以降低網站之負荷。

另一方面，網頁的特別設計會增加爬蟲程式設計的難度。例如有些網站係動態網頁，這種網頁設計會增加爬蟲程式的複雜度，但可在設計程式時解決。最嚴重的問題在於網頁改版：每當網頁改版，以荷蘭國家統計局的經驗而言，需要大約 8 到 40 小時重新撰寫程式碼，時間長短取決於網頁修改幅度及網頁本身複雜度。若資料蒐集者未察覺網站設計已經改版，可能蒐集到錯誤資料，或完全無資料。為了處理這個問題，需要發展「穩健」的腳本程式（robust scripts），檢測下載的資料是否正確。

3.3 以網路搜尋大數據預測通膨

搜尋引擎巨擘 Google 自 2008 年 8 月起，公布其搜尋引擎使用者在一定期間內搜尋一字詞之搜尋量，這一紀錄稱為 Google Trends，由此可看出網路搜尋者對某一特定字詞搜尋頻率隨時間之變化[4]。由於 Google Trends 之關鍵字搜尋量係代表所有網路使用者關注某一事件之程度，該數據可能反映當時經濟狀況，亦可能為經濟狀態的領先指標。由於經濟數據的公布皆落後實際發生一段時間，無論該數據反映當時或未來的經濟狀態，皆可用於預測。

Google 移除網路使用者之成長趨勢，並將關鍵字搜尋量標準化，其值範圍設定於 0 至 100 之間。Google Trends 網站提供之關鍵字資料以時間數列呈現，最早可回溯至 2004 年 1 月 1 日。此外，Google Trends 網站亦針對地理區域進行設定，可獲得單一國家或較小範圍（如州、縣、市）的關鍵字搜尋資料。

Ginsberg et al. (2009) 係早期利用 Google Trends 做預測最具代表性的例子之一，該文根據網路搜尋流行性感冒之相關用字，用來做流行性感冒之預測，稱為 Google Flu Trends。掌握感冒之流行趨勢，可使衛生當局能及早針對疫情做反應，防止疫情擴大。

經濟預測方面，Ettredge et al. (2005) 係最早以網路搜尋關鍵字資料做經濟預測之文獻。該文採用的資料為 Rivergold Associates 每週公布搜尋頻率最高之 500 個關鍵字。該文發現，關鍵字搜尋資料有助於預測失業人數。最早使用 Google Trends 做預測之文獻為Choi and Varian (2009) 與Choi and Varian (2012)。兩位作者以 Google Trends 預測汽車銷售、失業補助、消費者信心等，發現加入 Google 搜尋變數可提升預測績效[5]。Schmidt and Vosen (2012) 發現破車換現金計畫（“cash for clunkers” program）的關鍵字搜尋頻率可提升法國、德國、義大利、美國民間消費之預測績效。Guzmán (2011) 以 Google 搜尋關

[4]例如在 Google Trends 內輸入「crisis」一詞，可看出 2008 年 10 月金融危機爆發時，該詞在全世界 Google 搜尋引擎使用者的搜尋量達最高點，顯示該時間點全球對 crisis 一詞的關注程度高於其他期間。

[5]Hal Varian 為 UC Berkeley 經濟系教授，Google 首席經濟學家。

鍵字估計通膨預期，並發現該文建構之通膨預期數列可預測通膨率。以色列央行研究人員Suhoy (2009) 發現 Google 搜尋資料有助於民間消費之估計，並根據數值模擬的結果，在 2007 年 12 月可預測 2008 年 4 月金融危機造成的經濟衰退。Simeon and Torsten (2011) 發現 Google Trends 預測民間消費的能力勝過 Michigan Survey。由於 Google Trends 的關鍵字甚多，研究者的挑戰之一在於選取對未來走勢有預測能力的關鍵字。

早期 Google Trends 預測之研究多針對已開發國家，晚近由於開發中國家的資料增加，開發中國家相關文獻如雨後春筍般出現。Carrière-Swallow and Labbé (2013) 建構智利的汽車銷售指數，發現該指數預測能力超越傳統上的 IMACEX index。Seabold and Coppola (2015) 之研究針對拉丁美洲各國的物價，選取與物價有關的關鍵字，建立預測模型，發現若一國 Google Trends 之資料品質夠佳，則該資料對預測物價較有幫助，例如哥斯大黎加與薩爾瓦多等國。

4 文字探勘在經濟上的應用

文字探勘（text mining）或自然語言處理（natural language processing，簡稱 NLP）係挖掘文字資訊的量化方法。文字探勘與閱讀相似，兩者目的都是為了萃取文字資訊，但電腦文字探勘有下列兩項優勢：第一，電腦閱讀文字的速度遠較人類快；第二，人類閱讀文章可能受限於情緒或個人成見，以致於忽略文章重要內容，電腦的閱讀則相對冷靜，可補足此類不足。

對央行而言，有助於政策制定的文字資料包括新聞、企業與銀行之財務報表、社群媒體、各種調查資料及央行的政策會議記錄。藉由文字探勘技術，吾人可分析這些文字資料裡的訊息。例如，新聞資訊與社群媒體的文字資料可使決策者了解市場參與者對市場情緒的看法，以及新聞媒體關注的主題，吾人可由這些資訊編製市場情緒指標。媒體報導與一般民眾對市場的看法，隱含對未來經濟狀況的預期，實證結果顯示，這些文字資訊可預測股價與金

融市場的走勢。

相較於政治學或行銷，文字探勘較少應用於總體經濟學與經濟政策分析。一般經濟學家所使用的文字探勘工具通常僅限於 Google 搜尋，其他工具較少運用，可能係因相較於其他社會科學，經濟學領域擁有較多數字資訊。然而，由於文字資訊無所不在，若能有效地將重要的文字資訊加入研究範圍內，有助於央行政策之制定。

4.1 文字探勘之前置處理及基本統計量

4.1.1 前置處理

文字探勘最的基本方法需要計算帶有訊息的字詞在文章中出現的頻率。以英文為例，文字探勘前必須將文件經過以下的前置處理，以萃取出文件的重要訊息：

(1) 斷詞（tokenization）：即將文字字串分解成較小的個體，例如字詞（words）、數字與標點符號，通常需要處理連字號（hyphen）或縮寫的句號，細節詳見Manning et al. (2009)。

(2) 除去停止詞（stopword）：即意義比較不重要的詞，例如冠詞 the、a 及介系詞 to、for 等等。

(3) 除去字尾（suffixes），將同一字詞的詞類變化歸為一類：例如將 prefer、prefers 與 preference 歸為同一字詞。

(4) 多單字構成的詞（multi-word phrase）：例如「central bank」一詞，處理時必須將兩個字看成一個複合字詞。

以上的步驟需要仰賴預先建立好的詞庫，包括哪些是停止詞、哪些是同一字的詞類變化、哪些是複合字詞。

4.1.2 文件-字詞頻率矩陣

文件-字詞矩陣（document-term matrix）係一矩陣，描述文件內各詞出現的頻率。例如，假設文件 1 由下列句子組成：

I love mathematical analysis and time series analysis.

文件 2 由下列句子組成：

I love economic analysis.

文件 1 中出現的字包括 I、love、mathematical、analysis、and、time、series；文件 2 包括 I、love、economic、analysis。其中 analysis 在文件 1 中出現 2 次，其餘字在文件中出現一次。吾人可將文件與文字出現的頻率寫成矩陣形式：

$$\begin{array}{c} \\ \text{文件 1} \\ \text{文件 2} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \text{I} & \text{love} & \text{mathematical} & \text{analysis} & \text{and} & \text{time} & \text{series} & \text{economic} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 2 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \quad (1)$$

矩陣的第 i,j 元素代表第 i 個文件中，第 j 個字出現的次數。例如，該矩陣第 $1,4$ 元素值為 2，代表第 1 文件中，第 4 個字，即 analysis 字出現 2 次。有些文獻將該矩陣寫成轉置的形式：

$$\begin{array}{c} \\ \text{I} \\ \text{love} \\ \text{mathematical} \\ \text{analysis} \\ \text{and} \\ \text{time} \\ \text{series} \\ \text{economic} \end{array} \begin{array}{c} \begin{array}{cc} \text{文件 1} & \text{文件 2} \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 2 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \quad (2)$$

4.1.3 字詞在文件中出現的頻率（tf-idf）

假設有一文件集包含 1000 份探討總體經濟狀況的文件。由於每一份總體經濟相關文件多少都可能觸及探討景氣循環的議題，某一文件出現 cycle（意為景氣循環）一詞的頻率高於 bank（銀行）一詞，未必代表該文件特別強調景氣循環的主題，其重視程度高於 bank 一詞。然而，若該文件出現 cycle 一詞相對 bank 一詞的頻率較另外 999 分文件高，則可代表該文件特別強調景氣循環，其強度勝過 bank 一詞。因此，為了衡量某一字詞在該文件中的重要程度，必須將該字詞在該文件出現的頻率，以該詞在所有文件出現的頻率比較，以標準化，這種做法稱為 tf-idf（term frequency–inverse document frequency），一個常用的標準化方法為：

$$tfidf(t) = f_{t,d} \cdot log\frac{N}{n_t} \tag{3}$$

其中 $f_{t,d}$ 代表 t 詞在 d 文件中出現的頻率，n_t 代表 t 詞在文件集中所有文件出現次數的總和，N 為文件集中所有字詞的總數。假設 $tfidf(t)$ 很高，不僅代表 $f_{t,d}$ 很高，而是代表 $f_{t,d}$ 的值，經過該字在文件集中所有文件出現的頻率標準化之後很高。

4.2 資訊檢索

資訊檢索（Information Retrieval, IR）係由文字資料萃取情緒（通常區分為正面或負面情緒）的方法。隨著資訊時代的來臨，文字資訊的數量較過去增加，電腦的發展使分析文字資料較以往更容易。若吾人能藉由分析文字資料訊息萃取市場的情緒，及當下媒體最關注的主題，將有助於經濟預測與政策制定。

事實上，資訊檢索的概念已經存在已久。William Peter Hamilton 於 1902 年至 1929 年間擔任華爾街日報主編，寫過關於 255 篇預測股市的評論。為了探討 Hamilton 的評論能否預測股價走勢，Cowles (1933) 將其評論文章之情緒

區分為「牛市」、「熊市」、「懷疑」三種，並且根據這些情緒指標編製交易策略。研究者聘請 5 位讀者，這些讀者讀完每一篇文章後標記其情緒感受，由三種情緒中選擇其情緒感受，最後將五位讀者的情緒以多數決，作為該篇評論的情緒。研究者若發現文章情緒為「懷疑」，則不做交易；若情緒為「牛市」，則買入；若情緒為「熊市」，則賣出。研究結果顯示，Hamilton 若運用此交易策略買賣道瓊工業指數，1903 年至 1929 年的收益為 12%，惟該指數在該期間之收益為 15.5%，顯示 Hamilton 之預測遜於消極之交易策略。

上述的文字分析實例係由讀者主觀判斷其情緒。現代的文字探勘研究中，由文字資訊分析市場情緒，最基本的方法係建立一個辭典，該辭典必須包含正負向用字以及類別用字。舉例而言，若要分析媒體的經濟新聞，吾人可能關心的議題包括「談論的主題是股市、匯市、物價或就業市場」，以及「媒體對該市場的報導是正面或負面」。若將市場與正負方向結合，吾人可將媒體的訊息區分為「物價上升」、「就業市場改善」等等。

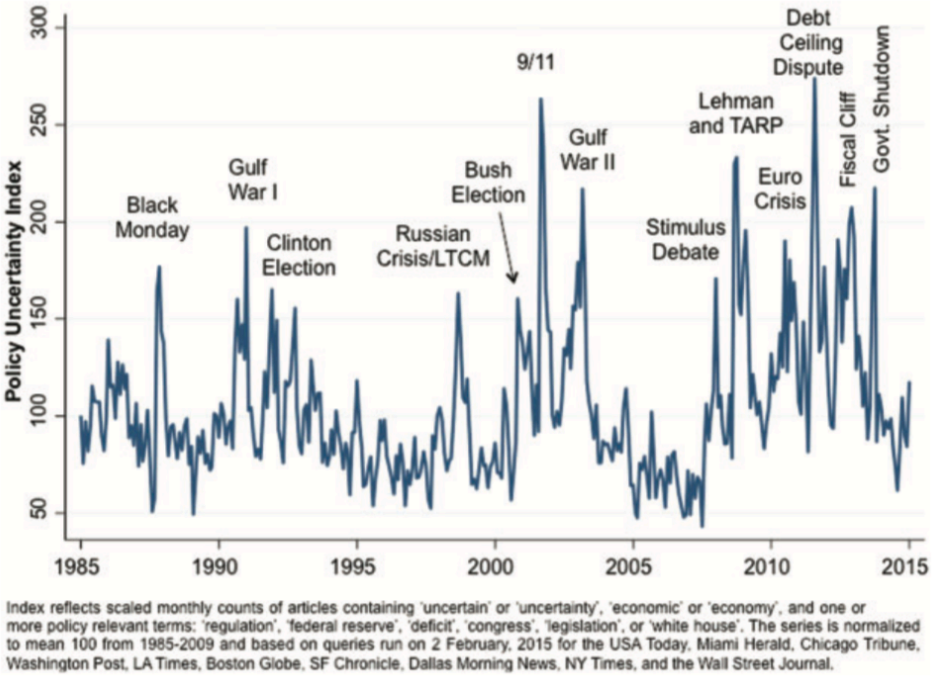
運用資訊檢索分析文字的第一步係建立一個辭典，就英文而言，最常用的辭典之一為哈佛心理學辭典（Harvard Psychology Dictionary）。然而，由於英文有許多同義字，各領域用字不盡相同，發展各領域適用的辭典作為文字探勘的依據甚為重要。舉例而言，「liability」一字在大多數領域意為累贅、負擔、風險，為負面字，在金融領域為資產負債表的「負債」，未必代表負面意義。因此，有些經濟或金融領域學者針對其研究的文件，自行建立辭典，以探討文字內的正向與負向情緒，例如Loughran and McDonald (2011) 建立之辭典，廣為後續經濟與金融領域之研究採用。該文發現哈佛心理學辭典中，大約有四分之三的負面用字在金融上並無負面意義，例如 tax、cost、capital、board、liability、foreign 與 vice，有些一般使用上的負面字在某些產業有其他意義，例如 mine、cancer、crude (oil)、tire 或 capital 等等。該文運用自行開發的辭典，發現新聞中存在之金融負面用字與現金增資（seasoned equity offering）有高度相關。

區分出與研究主題相關的關鍵字後，下一步便是找出各文件中某一主題或

情緒的強度，這些主題或強度的變數可作為後續研究的解釋變數。例如，若吾人欲研究社群媒體上正面或負面的情緒對隔日的股市是否有正面或負面的影響，或是這些情緒對隔日的股市是否有預測能力，吾人可藉由迴歸分析等統計方法，以股價作為被解釋變數。

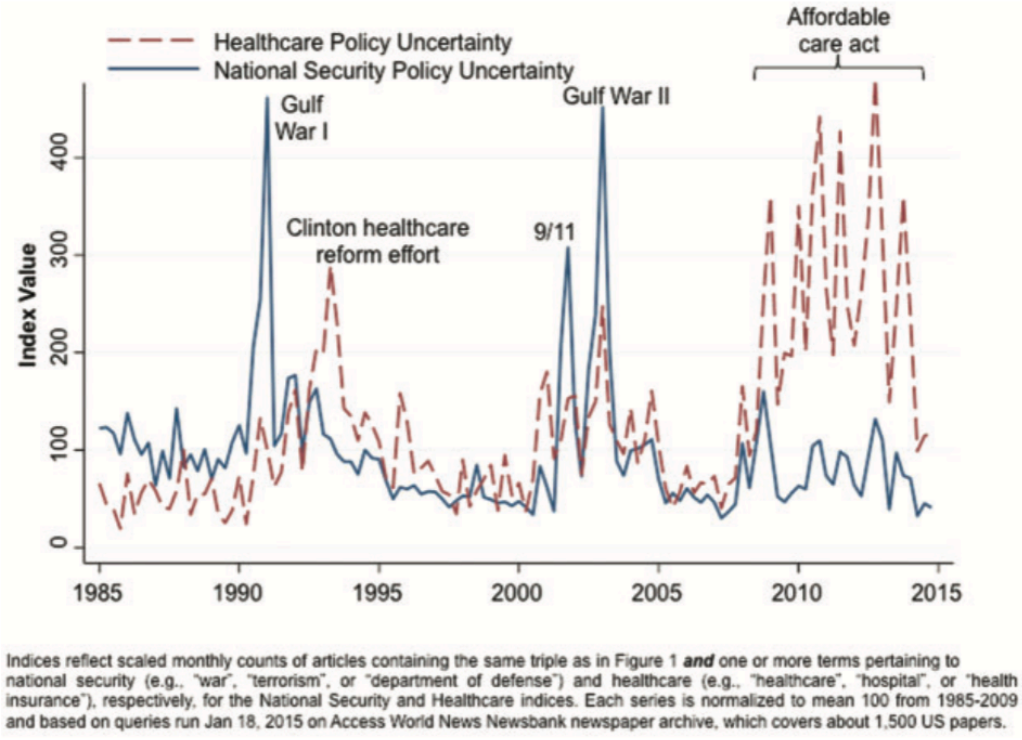
Baker et al. (2016) 堪稱文字探勘在經濟學領域最成功的應用之一。該研究分析 1985 年至 2012 年，10 家主要報紙的新聞報導，建立與政策不確定性有關的詞庫，並由關鍵字分析建構各時間點的「經濟政策不確定性指數」(Economic Policy Uncertainty index, EPU index)，如圖1。由圖可知，該 EPU 在特殊政治或經濟事件發生時達高峰，與理論預期大致相符。此外，該文另將 EPU 各成分依照政策種類分為財政（分為稅務與政府支出兩類）、貨幣、醫療、國防、監理、主權債務與金融危機、應享權益方案（entitlement program）、貿易政策等類，並編製各政策類別的不確定性指數。醫療與國防政策的不確定性指數如圖2。

圖 1: 美國政策不確定性指數（EPU Index）



資料來源：轉引自Baker et al. (2016)，p.1600。

圖 2: 美國國防與醫療政策不確定性指數
（National Security and Health Care EPU Index）



Indices reflect scaled monthly counts of articles containing the same triple as in Figure 1 ***and*** one or more terms pertaining to national security (e.g., "war", "terrorism", or "department of defense") and healthcare (e.g., "healthcare", "hospital", or "health insurance"), respectively, for the National Security and Healthcare indices. Each series is normalized to mean 100 from 1985-2009 and based on queries run Jan 18, 2015 on Access World News Newsbank newspaper archive, which covers about 1,500 US papers.

資料來源：轉引自Baker et al. (2016)，p.1601。

金融領域方面，Tetlock (2007) 係首先發現媒體報導可預測股市活動的文獻。該文分析華爾街日報專欄「Abreast of the Market」1984 年至 1999 年的報導，藉由這些文字資訊建立每日的悲觀因子（pessimism factor），並發現該悲觀因子的變化可預測美國股價報酬與交易量。採用華爾街日報的理由，係該報之發行量與流通量在同業中最大，且在投資人中的聲譽極佳。

該文建立的悲觀因子，係 General Inquirer（GI）根據哈佛心理學辭典的 77 種情緒，運用主成分分析法，取其第一主成分。由於該主成分與悲觀用字的相關程度相當高，作者稱此變數為悲觀因子。

理論上，悲觀因子有可能反映市場上的悲觀情緒，亦有可能預測股價走勢或交易量。Tetlock (2007) 發現，市場上的悲觀情緒會對股市造成壓力，預示股價下挫以及市場波動度上升。若悲觀情緒極高，則市場上的交易量會暫時

大增，股價下挫，但一段時間後股價會由過度反應的情緒中回升。

後續的研究中，Loughran and McDonald (2011) 認為傳統研究所採用的哈佛心理學辭典可能不適用於金融的情境。舉例而言，英文字中的 liability、cost 與 foreign 在一般情境中為負面用字，但在金融或財務報表內則非如此。大約有四分之三的日常負面用字並不具負面意義。因此，作者編製了一個專屬於金融領域專用字辭典，包括負面用字（Fin-Neg）、正面用字（Fin-Pos）、不確定用字（Fin-Unc）、訴訟用字（Fin-Lit; litigious）等。

中文方面，由於語法結構較英文複雜，資源相對匱乏。雖然中國大陸與台灣都各自有許多相關技術發表於研究領域，但免費的語料庫與程式的發展卻相對落後。中央研究院古倫維研究員的研究團隊開發了中文情感語意分析套件（CSentiPackage, Chinese Sentiment Package），開放供研究目的的使用者自由免費下載。其中資料集的部分有中文構詞資料集（Chinese Morphological Dataset）及中文意見樹庫（Chinese Opinion Treebank）；語意字典則包括了台灣大學意見詞詞典（NTUSD）以及增廣意見詞詞典（ANTUSD）；語意分析工具則有用於中文意見分析的語意挖掘計分工具（CopeOpi）及一個深度社群立場分析模型（UTCNN）。關於該系統介紹，參見Chen and Ku (2018)。

4.3 主題模型與相似性分析

前述資訊檢索法係先定義一組關鍵字，由關鍵字在各文件中出現的頻率，找出文件中的主題。本節探討的問題係兩個文件主題的類似程度，及在一群文件中找出共同主題及各自文件中這些主題的強度。這些方法必須仰賴文件中出現頻率較高的字作為關鍵字，但不需預先定義關鍵字。研究者可以各主題強度作為解釋變數，探討文件中各主題對經濟與金融情勢的影響。

4.3.1 兩文件相似程度：餘弦相似性

文件經過前置處理後，潛在語意分析的第一步係計算文件-字詞頻率矩陣，每一列代表一個語詞，每一行代表一個文件（詳見4.1.2 小節矩陣 2）。由於吾人

假設每一個文件的訊息都可由其中每一字詞出現的頻率來代表，文件-字詞頻率矩陣已經涵蓋這些文件集中所有訊息。矩陣中的每一行代表這個文件的特性，若兩個文件的內容相當類似，則我們會預期兩文件各字詞出現的比例大致接近，用向量的語言而言，兩條向量接近平行。衡量兩條向量是否接近的方法係計算其餘弦值（cosine），這種衡量方法稱為「餘弦相似性」：

$$cos(\theta) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} \tag{4}$$

其值愈高，則代表兩文件愈相似[6]。Hoberg and Phillips (2010) 取用美國 Securities and Exchange Commission 下之公司財務報表資料，計算各公司之間之餘弦相似性，相似度高的公司為類似的產業。他們發現此種分析方法計算所獲得的相似性即為產業的相似性，亦為產品間的替代性程度。

4.3.2 奇異值分解與潛在語義分析：以 FOMC 相關文獻為例

潛在語意分析（Latent Semantic Analysis, LSA）係一演算法，由眾多文件中找出文件群中最重要的共同主題（這些主題通常不止一個），並且找出個別文件中各主題的成分。

以美國聯邦公開市場委員會（FOMC）之議事錄摘要（minutes）為例，文件的數量大約有一兩百個。這些文件中，可能有若干文件是類似的，其中之主題包括經濟成長、物價、景氣循環、就業市場、貨幣政策等等，惟每一次議事錄摘要中各主題的重要性不盡相同。Boukus and Rosenberg (2006) 以潛在語意分析法研究 1987 年至 2005 年，152 份 FOMC 的議事錄摘要所探討的主題。某些議事錄摘要多針對消費，論及其他主題的成分可能較少；某些議事錄摘要多強調景氣循環的成分較高。

文件經過4.1.1小節的前置處理後，潛在語意分析的第一步係計算文件-字詞頻率矩陣，每一列代表一個語詞，每一行代表一個文件（參閱4.1.2 小節

[6]由於兩向量內的數值皆為正，該餘弦值必定不為負；最大值為 1，最小值為 0。

矩陣 2)。FOMC 文件中，常出現的字包括 *spend*、*consum*、*econom*、*busi*、*inflat*、*dollar*、*restraint* 等等[7]。一個典型的矩陣如矩陣5：

$$\begin{array}{c} \\ \text{spend} \\ \text{consum} \\ \text{econom} \\ \text{busi} \\ \cdots \\ \text{dollar} \\ \text{inflat} \\ \text{restraint} \end{array} \begin{array}{c} \begin{array}{cc} \text{文件 1} & \text{文件 2} \end{array} \\ \begin{pmatrix} 4 & 7 \\ 9 & 10 \\ 12 & 10 \\ 5 & 7 \\ \cdots & \cdots \\ 3 & 1 \\ 8 & 7 \\ 4 & 3 \end{pmatrix} \end{array} \tag{5}$$

這 152 份文件中，不重複的詞共出現 2402 次。因此，矩陣5共有 2402 列，152 行。如何萃取各文件中最重要的主題？由於有 2402 個不重複的字，主題最多有 2402 個。直覺上，最重要的主題似乎即是這 2402 個字中，出現頻率最高的詞。假設頻率最高的詞是 growth，且在這些文件中，growth 一詞出現時，spend 一詞亦同時出現，則可將 growth 與 spend 歸類為同一主題。另一方面，如果把每個文件中各字詞的組合看成一個主題，則主題最多有 152 個。然而，經濟上關心的主題應該遠小於 152 個，若文件 1、2、3 的主題大致相似，即各字詞出現的頻率完全相同，則可將該三個文件的主題濃縮成一個主題。如前所述，矩陣中的每一行代表這個文件的特性，若兩個文件關心的主題相當類似，則我們會預期兩文件各字詞出現的比例大致接近，意即兩條向量接近平行，或是代表一個文件的向量大致接近另一文件向量的倍數。假設 152 份文件中，可找到少數重要的主題，使每一次的會議記錄大致都可寫成這些主題的倍數及其加總（即這些主題的線性組合），吾人便可分析文章中最重要的主題，以及這些主題在各文件中被強調的程度。

[7]由於除去字尾，這些英文單字未必是完整的字，例如 *consum* 可能代表名詞 consumption 亦可能代表動詞 consume。

找出這些文件主題的方法係以奇異值分解法（singular value decomposition）分析字詞文件矩陣，萃取一群文件中最重要的數個主題。這種計算方式與主成分分析法類似。各主題包含數個關鍵字，吾人再由關鍵字的解讀，將這群關鍵詞判定為特定的經濟主題。該文發現，前十大重要主題可解釋變異數中的一半，前一百大主題可解釋變異數中的 94%。前五大主題如下：

表 1: 主題

	主題	關鍵字
主題 1	消費者信心、經濟擴張	spend, consum, domest, expans, pressur, and growth
主題 2	總體經濟情勢、景氣循環	econom, weak, busi, ease, declin, price, growth, and increas
主題 3	通膨壓力、緊縮貨幣政策	inflat, rise, demand, labor, tight, and price
主題 4	外匯市場	dollar, market, inflat, condit, and pressur
主題 5	貨幣政策情勢	polici, slightli, restraint, and consider

(1) 資料來源：Boukus and Rosenberg (2006)。

(2) 各文件之關鍵字係來自分析 FOMC 議事錄摘要的關鍵字出現之頻率，主題係由關鍵字的組成解讀。

確認出最重要的主題後，吾人亦可計算各主題成分隨時間的變化。為了確認這些主題是否有正確的經濟意涵，該文檢驗這些主題隨時間的改變與經濟變數、金融市場的關係。失業率與主題 4、5 的成分改變有顯著的負向關係；經濟成長與主題 2、3 有正向、顯著的關係。整體而言，主題 2、3 與未來的經濟與金情勢關係最強。主題 2 與景氣循環有關，當主題 2 的正向成分可預期短期利率上升、能源價格上升、經濟成長。主題 3 與物價相關，與長期利率、陡峭的殖利率曲線、高經濟成長同期相關。主題 4 與較低的長期利率、正的 S&P 500 報酬、低失業率同期相關。主題 5 與貨幣政策情勢相關。

在預測結果方面，本文顯示 FOMC 有新的訊息內涵：FOMC 公布議事錄摘要當天，公債市場之波動度增加，且這些主題可以解釋未來長期利率的變化。2 年期公債報酬率受主題 1、2 衝擊的影響，係數為負，亦受主題 3 衝擊的影響，係數為正；10 年期公債報酬率，顯示市場對中長期利率之預期受 FOMC 探討之特定主題驅動。

5 結語與建議

大數據與央行業務有關的範圍相當廣泛，包括經濟預測、情緒分析、金融市場、金融檢查、統計資料之校正等等，多數央行皆已建立研究團隊進行研究，並有研究成果。建議本行可針對行內需要及其他國家已著手的研究主題開始，投入人力進行研究，並且多關注國際間的研究成果，參與國際研討會。行內人力不足或面臨較缺乏的專業知識時，可考慮與行外學者專家合作。

大數據對央行政策制訂上可帶來很大的幫助，惟發展大數據需要有足夠資源，包括人力與資訊基礎建設。以 MIT 的巨量價格數據計畫為例，以網路大數據建構即時物價指數，需要搜尋世界多國的網路商店，該團隊聘請校內瞭解各種語言的學生撰寫網路爬蟲程式碼。網路商店之網頁改版時，程式碼必須重新撰寫。牛津大學教授 Stephen Hansen 指出，文字探勘的研究，在分析文字前需要經過前置處理，將檔案轉成存文字格式，套裝軟體未必能達到需求，有些部分需要人工處理。其他研究計畫亦需要足夠人力撰寫程式，並擬定研究方向。資料的蒐集方面，傳統上央行為資料使用者，發展大數據有可能需要蒐集新的資料，並建立資訊交流的管道，俾利資料之研究與分析。

參考文獻

Baker, S. R., Bloom, N., Davis, S. J. (2016), "Measuring economic policy uncertainty," *The Quarterly Journal of Economics*, 131, 1593–1636.

Borraz, F., Cavallo, A., Rigobon, R., Zipitria, L. (2016), "Distance and political boundaries: Estimating border effects under inequality constraints," *International Journal of Finance & Economics*, 21, 3–35.

Boukus, E., Rosenberg, J. V. (2006), "The information content of FOMC minutes."

Carrière-Swallow, Y., Labbé, F. (2013), "Nowcasting with Google trends in an emerging market," *Journal of Forecasting*, 32, 289–298.

Cavallo, A. (2018), "Scraped data and sticky prices," *Review of Economics and Statistics*, 100, 105–119.

Cavallo, A. (2013), "Online and official price indexes: Measuring Argentina's inflation," *Journal of Monetary Economics*, 60, 152 – 165.

Cavallo, A. (2017), "Are online and offline prices similar? Evidence from large multi-channel retailers," *American Economic Review*, 107, 283–303.

Cavallo, A., Neiman, B., Rigobon, R. (2014), "Currency unions, product introductions, and the real exchange rate," *The Quarterly Journal of Economics*, 129, 529–595.

Cavallo, A., Neiman, B., Rigobon, R. (2015), "The price impact of joining a currency union: Evidence from Latvia," *IMF Economic Review*, 63, 281–297.

Cavallo, A., Rigobon, R. (2016), "The Billion Prices Project: Using online prices for measurement and research," *Journal of Economic Perspectives*, 30, 151–78.

Chen, W.-F., Ku, L.-W. (2018), "Introduction to CSenti package," *Journal of Library & Information Science*.

Choi, H., Varian, H. (2009), "Predicting initial claims for unemployment benefits," Technical Report, Google.

Choi, H., Varian, H. (2012), "Predicting the present with Google trends," *Economic Record*, 88, 2–9.

Cowles, A. (1933), "Can stock market forecasters forecast?" *Econometrica*, 1, 309–324.

Ettredge, M., Gerdes, J., Karuga, G. (2005), "Using web-based search data to predict macroeconomic statistics," *Communications of the ACM*, 48, 87–92.

Giannone, D., Reichlin, L., Small, D. (2008), "Nowcasting: The real-time informational content of macroeconomic data," *Journal of Monetary Economics*, 55, 665 – 676.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L. (2009), "Detecting influenza epidemics using search engine query data," *Nature*, 457, p. 1012.

Guzmán, G. (2011), "Internet search behavior as an economic forecasting tool: The case of inflation expectations," *Journal of Economic and Social Measurement*, 36, 119–167.

Hoberg, G., Phillips, G. (2010), "Product market synergies and competition in mergers and acquisitions: A text-based analysis," *The Review of Financial Studies*, 23, 3773–3811.

Knotek, E. S., Zaman, S. (2016), "Inflation nowcasting: Frequently asked questions."

Knotek, E. S., Zaman, S. (2017), "Nowcasting U.S. headline and core inflation," *Journal of Money, Credit and Banking*, 49, 931–968.

Loughran, T., McDonald, B. (2011), "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *The Journal of Finance*, 66, 35–65.

Manning, C. D., Raghavan, P., Schütze, H. (2009), Introduction to Information Retrieval, Cambridge University Press.

Modugno, M. (2013), "Now-casting inflation using high frequency data," *International Journal of Forecasting*, 29, 664 – 675.

Schmidt, T., Vosen, S. (2012), "Using internet data to account for special events in economic forecasting," *Ruhr Economic Paper*.

Seabold, S., Coppola, A. (2015), "Nowcasting prices using Google trends an application to Central America," Technical Report, World Bank Group.

Simeon, V., Torsten, S. (2011), "Forecasting private consumption: survey-based indicators vs. Google trends," *Journal of Forecasting*, 30, 565–578.

Suhoy, T. (2009), "Query indices and a 2008 downturn: Israeli data," Technical Report, Research Department, Bank of Israel.

Tetlock, P. C. (2007), "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, 62, 1139–1168.