**「台灣政府開放性資料之應用」議程**

時 間：108年1月12日(星期六) 上午8:40 ~ 17:30

地 點：中華經濟研究院 蔣碩傑國際會議廳

主辦單位：台灣經濟學會 中華經濟研究院 行政院主計總處

協辦單位：台大行為與資料科學研究中心 台大計量理論與應用中心

東華大學財務金融學系 中正大學經濟系

國泰金融控股股份有限公司

|  |  |  |
| --- | --- | --- |
| 時間 | 內容 | 主講人 |
| 8:40 -- 9:00 | 與會來賓報到 |  |
| 9:00 -- 9:20 | 開幕致詞  陳恭平 (台灣經濟學會理事長)  潘城武（主計資訊處處長） |  |
| 9:20 -- 10:10 | 主席: 吳中書 (台灣金融研訓院董事長) |  |
| 專題演講: Data Mining, Big Data to AI | 謝邦昌(臺北醫學大學 管理學院院長暨大數據研究中心主任) |
| 10:10 -- 10:40 | 使用R/Python整合開放數據源  作者: 何宗武，酆士昌 | 酆士昌(昊瀚資訊股份有限公司總經理) |
| 10:40 -- 11:00 | 休 息 時 間 | 發表人 |
| 11:00 -- 12:30 | 開放資料的應用實例與研究：不動產買賣實價登錄資料 |  |
| 主席: 林明仁 (國立台灣大學經濟學系系主任) |  |
| 1. 台灣房地產景氣循環週期之研究：開放資料之應用   作者: 林淑娟，洪敬瑞，李顯峰 | 洪敬瑞(南京資訊股份有限公司資深工程師) |
| 1. 大台北地區房價的大數據分析   作者: 林金龍，陳哲銘 | 林金龍 (國立東華大學財務金融學系教授) |
| (3) The Pricing of Liquefaction Risk: Empirical Evidence from Housing Market in Taiwan  作者: Tzu-Chang Forrest Cheng, Tai-Chi Wang, Jian-Da Zhu | 朱建達（國立台灣大學經濟學系助理教授） |
| 12:30 -- 14:00 | 午 餐 時 間 |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| 12:30 -- 14:00 | 午 餐 時 間 |  |
| 14:00 -- 15:30 | 開放資料的應用實例與研究：受雇人員/薪資與司法資料  主席: 葉俊顯 (中華經濟研究院副院長、中央研究院經濟所研究員) |  |
| 1. 以政府開放資料檢視受雇人員薪資與雇主經營盈餘之相關性   作者: 陳建良，管中閔，林樹明，李巧琳 | 陳建良(國立暨南國際大學經濟學系教授) |
| 1. 經濟成長、薪資停滯？初探台灣實質薪資與勞動生產力成長脫鉤之成因   作者: 林依伶，楊子霆 | 楊子霆(中央研究院經濟研究所助研究員) |
| 1. 大數據與人工智慧在法律經濟學研究的應用   作者: 林常青 | 林常青(成功大學經濟系副教授) |
| 15:30 -- 15:45 | 休 息 時 間 |  |
| 15:45 – 17:15 | 開放資料的應用實例與研究 |  |
| 主席: 林金龍(國立東華大學財務金融學系教授) |  |
| 1. 如何使用衛福資料中心資料進行健康研究   作者: 連賢明 | 連賢明（國立政治大學財政學系教授） |
| 1. Textual Data Analytics in Finance   作者: 蔡銘峰 | 蔡銘峰（國立政治大學資訊科學系副教授） |
| 1. 資料治理之發展趨勢與國際案例探討   作者: 許志義，賴姵瑀 | 許志義（亞洲大學經營管理學系講座教授、國立中興大學教授） |
| 17:15 -- 17:30 | 閉幕  主席: 林金龍(國立東華大學財務金融學系教授) |  |

**專題演講**

**謝邦昌 教授**

臺北醫學大學管理學院院長

臺北醫學大學大數據研究中心主任

臺北醫學大學生物科技高階管理

碩士在職專班教授

台灣人工智慧發展學會理事長

中華資料採礦協會榮譽理事長

中華市場研究協會理事長

世界中醫藥學會聯合會專業委員理事會副會長

北京對外經貿大學統計學院 講座教授

中國人民大學統計學院 兼職教授

北京大學新媒體研究院 兼職教授

廈門大學管理學院 講座教授

**Data Mining, Big Data to AI**

**謝邦昌 (**臺北醫學大學 管理學院院長暨大數據研究中心主任)

近年來隨著網路環境發達、雲端服務盛行及智慧型行動裝置之普及，造就了「儲存成本」與「資料取得成本」迅速的下降，伴隨而來的是資料快速且大量的累積，使得人們更容易地進行資料的搜集，也因此從資訊技術（Information Technology，IT）的階段邁入了資料技術（Data Technology，DT）的時代，是目前大數據快速竄升的背景。

就字面上來看，「大數據」這個詞容易使人誤解為「大量的資料」，而這僅是大數據的其中一項特性，目前對於大數據的特性普遍定義為4V，包含資料量（Volume）、資料傳輸速度（Velocity）、資料類型（Variety）及真實性（Veracity）。

「資料量」顧名思義就是巨大數據量，相較與以往人工耗費心力搜集資料的過程，現在由於技術的進步，各產業在其中的每項環節，皆可以透過物聯網的感測器來接收資料，也因此累積了巨量的資料，資料量很容易就能達到數百萬位元組（Tera Bytes，TB），甚至上看千百萬位元組（Peta Bytes，PB）或百萬百萬位元組（Exabytes，EB）的等級。

「資料傳輸速度」隨著感測器以及社群網站的蓬勃發展，產生龐大的資料量，公司與機構單位立即回應、反應這些資料的速度，成為他們首要的挑戰，許多資料要能即時得到結果才能發揮最大的價值，因此也將其視為「時效性」。

「資料類型」大致分為兩種，結構化資料以及非結構化資料，結構化資料是傳統分析的表格型資料，資料較整齊，便於分析；而非結構化資料包含文字、圖像、影視、音樂、電子郵件及網頁等，通常以不規則的形式存放，較不容易進行分析。

「真實性」是由於在資料搜集或是處理的過程當中，發生資料偏差、偽造及異常等情形，我們必須防止這些損害到資料系統的完整與正確性，因為這些情況會對決策產生相當大的影響。

大數據的時代下，資料種類更加多元化，大量的資料散佈於各個不同的單位，以結構化資料(如：數字型態)及非結構化資料(如：文字、聲音及影像)等形式存在，其資料存放方式差異很大，因此具備將資料整合與分析能力極為重要。在大數據之具體運用及整合上，包括可利用Google資料庫，檢索關鍵字句並結合疾病知識，運用數位模型判斷出流感傳播的途徑和趨勢，為公共衛生決策提供服務，以及運用交通大數據，于道路檢測、車輛出行、天氣、突發事件、重大活動等方面的資料來預測交通情況。另外，美國加州的洛杉磯警局，利用過往的犯罪資料庫並透過演算法，來產生可能有犯罪事件發生的地方的地圖，以預測及預防罪犯事件。

以「大數據應用於毒藥品防制議題」為例，在雲端資訊快速傳播的時代，罪犯不斷製造查緝中斷點，更需要藉由大數據分析的概念，分析出犯罪模式，對查緝人員提出預警。透過金流、物流、網路暗語等面向來分析及查緝毒品網路，針對可疑藥品原料來源管制清查，擴大追緝制毒工廠，以及分析供應來源或是犯罪熱點、運輸網路或模式等變項，以降低毒品的供給，防止毒品走私。此外，可分析比較不同地區間藥物濫用之差異及趨勢，找出警示因素和危險因數，進而提出相對應的防制策略。也可藉由分析藥品濫用者的消費模式、生活形態以及社會經濟資料，以探討更多樣化的預警因素。

透過大數據大量、快速及多變的特性，分析各種不同型態及類型的資料，將帶來更多有價值的洞見，藉由創新的處理方式規劃出精准的決策，可迅速掌握當前藥物濫用議題之情勢發展，以有效將低毒品犯罪及施用毒品人口，提升整體反毒成效。

Gartner預估到2021年之前，將有30%利潤會由AI來驅動，如果公司不投資AI，未來就少了30%掌握市場的機會。對於人工智慧應用與人才的培訓，Gartner研究總監Angela McIntyre也建議，可以先從大專院校中開始培養DNN與深度學習方面的技能，特別針對深度學習的資料科學家。但她也提到，現在不管是推動何種新方案，都需要跨部門的組成人才，除了資料科學家之外，需要真正使用到這個技術的業務團隊，還有行銷人員來定義採用的方法，任何專案在開始之前都需要有高階主管的投入與支持，這樣才可以把原先的計畫或項目大規模的採用。

深度演算法、計算力、資料量以及應用場景共同推動AI迎來爆發期

自從1956年美國電腦協會組織的達特莫斯（Dartmouth）學會上提出“人工智慧”一詞，這個概念隨著眾多理論、技術和應用的出現而被不斷豐富。經過半個多世紀的發展，人工智慧已經成為時下最為熱門的話題之一，“AI”與“AI+”亦成為一級市場的最火熱的創業/投資領域。

在經歷了誕生、黃金時代、遭遇障礙、繁榮、低潮等數個階段之後，人工智慧迎來了爆發期。總結本次AI爆發的驅動力主要是包括目前主流應用的基於多層網路神經的深度演算法和以及包括晶片、超級電腦、雲計算等在內的計算力等。此外，被記錄下的海量資料和越發豐富的應用場景也同樣推動了AI爆發期的到來。

使用R/Python整合開放數據源

何宗武(臺師大管理學院全營所教授)

**酆士昌**先生(昊瀚資訊股份有限公司總經理)

本專案嘗試使用開源軟體開發串接開放數據的程式，除了台灣政府<https://data.gov.tw/> 的數據，全球開放數據也在計畫之中。目前<https://data.gov.tw/>的格式問題較為嚴重，使用度不可靠也不理想；要標準化並不容易；財經數據則較為可靠。簡報分四點報告：

1. Open Data的資料範疇

2. 全球開放數據建構現況：近3萬筆清單的清理。

3. 台灣政府<https://data.gov.tw/> 開放數據的處理現況：格式問題

4. R語言的爬蟲套件

5. 範例程式與輸入輸出的標準化

6. 未來的發展規劃: R套件和Python

台灣房地產景氣循環週期之研究：開放資料之應用

林淑娟（財政部臺北國稅局）

**洪敬瑞**（南京資訊有限公司）

李顯峰（國立台灣大學經濟學系）

本文嘗試應用頻譜分析(Spectral Analysis)以及運用R程式語言，探討台灣2002年至2016年間房地產景氣循環週期，研究方法迥異於現有探討台灣房地產景氣循環的文獻，補充迄今分析台灣房地產景氣變化的方法。將所有與房地產相關政府開放的經濟變數，由時域(time domain)轉換成頻域(frequency domain)，並在頻域中分析出組合所有頻率的頻譜，尋出各變數中影響性較大的頻率後，再篩選符合之變數，將這些影響性較大的各頻率進行變數間的頻譜連動性(Spectral Coherence)分析，並以此分析結果來度量兩時間序列的連鎖效果(ripple effect)大小，最後加以整理並估計得房地產景氣循環的主要週期(Major Cycle)。

實證研究結果發現，台灣2002年至2016年間之房地產景氣循環週期約1~2年，顯示台灣房地產的波動較為快速，與以往文獻所分析出的週期較不相同。另外以往所認定與房地產景氣循環較具有相關性的樣本變數，如房價所得比、房貸負擔率、租金指數等，其頻率與其他變數較不相同外，頻譜連動性亦不高 ，不必然為房地產景氣循環之構成要素。

關鍵字：房地產景氣循環、房地產綜合指標、頻譜分析、傅立葉變換(Fourier Transform)、頻譜連動性、開放資料應用

**大台北地區房價的大數據分析**

**林金龍** (國立東華大學財務金融學系)

陳哲銘 (國立東華大學財務金融學系)

本文應用「遞迴分區迴歸樹模型」（Model-based recursive partitioning、MOB）來分析影響房價的關鍵因素，並建構房價預測模型。MOB係迴歸與決策樹的混合模型，對於已知是線性或非線性的影響變數以線性或非線性迴歸模型處理之；對於影響非常複雜的變數則納入決策樹中，作為區分變數。本文使用的資料庫是不動產交易實價查詢，這是一個登錄正確性極高且交易的細節皆很完整的資料庫。該資料庫除了房價及房屋基本資訊外，額外的加入「與最近捷運站的距離」、「與最近公園的距離」、「公園的面積」，「移轉樓層」，「使用目地」及「經緯度」等變數。這些變數都對房價有很大的影響且相當完整，經由經濟理論可以得知其中某些變數適合以母數迴歸模型處理，如房屋的面積等；有些變數如經緯度則宜以納入決策樹中作為分區變數。混合模型能夠兼顧估計效率，具有非常高的彈性，且適宜處理連續與間斷型值性變數，實證結果發現MOB對於房價有很好的預測能力。

**The Pricing of Liquefaction Risk: Empirical Evidence from Housing Market in Taiwan**

Tzu-Chang Forrest Cheng (Department of Economics, National Central University)

Tai-Chi Wang (National Center for High-performance Computing)

**Jian-Da Zhu** (Department of Economics, National Taiwan University)

November 2018

This research applies the difference-in-differences method to investigate the effect of the announcement of the soil liquefaction risk map on housing prices in the Taipei metropolitan area. The result shows that housing prices in the high potential risk area significantly dropped 4.32% in the first three months following the announcement, compared to those in the no risk area, and there was no effect in the moderate and low potential risk areas. In addition, this information effect was temporary, and housing prices returned back quickly three months later. Furthermore, this effect only showed up for those less earthquake-resistant buildings, such as apartments without elevators, older houses, and buildings built before 1999, and there was no effect for those safe buildings. To further study the information effect near the boundary, the result shows that housing prices close to the center of high potential risk area dropped more sharply than those near the high-moderate boundary although the risk information was the same based on the map. In addition, the effect has diminished in different speeds. For those apartments near the boundary, the effect was small, and it rebounded back quickly. However, the effect continued until the end of 2016 for those apartments located in the center of high risk area, and the effect even continued existing for those less quake-resistant buildings.

以政府開放資料檢視

受雇人員薪資與雇主經營盈餘之相關性

**陳建良**（國立暨南國際大學經濟學系）

管中閔（國立臺灣大學財務金融學系）

林樹明(Bucknell University (USA), Department of Economics)

李巧琳（國立暨南國際大學國際企業學系）

台灣勞動市場平均實質薪資長期停滯，原因眾說紛紜；一個可能的理由是雇主不願將利潤分享員工，是為“慣老闆假說”。本研究採用財政資訊中心的報稅資料，連結勞動部和內政部的受雇人員特性資料，在標準的工資函數設定之下，討論雇主盈餘和受雇人員薪資兩者之間的彈性。結果發現，在其他條件控制之下，雇主保留盈餘對於受雇人員薪資的彈性是1.5%；以一個30人規模的公司而言，雇主的盈餘有一半左右分給受雇人員。進一步區分雇主的當期盈餘與前期盈餘，前者的邊際效果是後者的兩倍。顯然，台灣的雇主願意分享盈餘給員工；政府開放資料並不支持“慣老闆假說”。

經濟成長、薪資停滯？

初探台灣實質薪資與勞動生產力成長脫鉤之成因

林依伶（中央銀行經濟研究處）

**楊子霆**（中央研究院經濟研究所）

過去15 年來，台灣的勞動生產力（每單位勞動的實質產出)持續成長，但實質薪資卻出現停滯。過去文獻多半將此一現象歸咎於我國勞動報酬份額持續下滑，亦即這是經濟成長果實「分配不均」所致。本文利用經濟模型與統計資料，將實質薪資與勞動生產力的成長差異拆解為三項因素的變動趨勢:(1) 勞動報酬份額; (2) 薪資佔勞動報酬之比率; (3) 產出對消費之相對價格(GDP 平減指數對消費者物價指數之比值)。與過去看法不同，實證結果顯示在2002-2014 年間，絕大部分(87%) 的每工時實質薪資與勞動生產力的成長差異，是來自於我國產出價格相對消費價格的下滑，而非勞動報酬份額的變化。進一步分析產出對消費之相對價格下跌的原因，並與南韓資料比較，我們發現台灣的實質薪資與勞動生產力之成長脫鉤現象可能與台灣貿易條件惡化有關。

關鍵詞：薪資、勞動生產力、貿易條件

JEL分類代號：J24, J31, F16**大數據與人工智慧在法律經濟學研究的應用**

**林常青** (成大經濟系)

專業人士「作決定」會受到什麼因素的影響，一直是經濟學家、心理學家與其他社會科學家想探究的問題。又因法官為高教育、訓練有素 、且常進行專業判決的職業，他們的決定－特別是外在事件發生時－常成為法律與經濟學實證研究的對象。然而，實證最大的困難在於判決的相關記錄不一定被詳細記載。臺灣雖然有詳盡的裁判書查詢系統而有發展的潛力，但需要大量文字探勘的技術。近期，司法院於「政府公開資訊平台」開放了司法院及所屬各級法院之「終結案件資料」。這個資料把主要的案件特徵作相當有系統的整理，能作為法律經濟學實證的第一手資料。本場次將介紹 (1) 「終結案件資料」的內容; (2) 該資料庫的優缺點; (3) 司法裁判書分析的潛在困難; (4) 文字大數據與 AI 能帶來的輔助 (5) 以及講者曾嚐試過幾個法律經濟學實證上的議題。最後，將以Chang, Chen, Lin (2018) 法院「錨定效應」的實證，作為示範的例子。

關鍵詞: 終結案件資料, 決策制定, 量刑決定, 法官行為, 錨定效應, 文字探勘

JEL: D91, K41, K42

**如何使用衛福資料中心資料進行健康研究**

**連賢明**（國立政治大學財政學系）

在進行衛生學術研究以及討論公共衛生政策中，最常使用數據資料是衛福部的「衛生福利資料科學中心」。這資料中心涵蓋健保署的全民健保資料庫；醫事處的醫療機構和評鑑資料；疾管局法定傳染病資料；國健署癌症登記、癌症篩檢、健康行為資料；統計處的死因檔、疾病主題檔、各項衛生調查；社家署的弱勢兒童和身心障礙資料；保護司的家暴檔，還有警政署的交通事故檔，以及原民會的原住民分份檔案，是目前台灣具備健康數據最完整的資料中心。

演講在第一部分中針對衛生福利資料科學中心的申請規定、資料收費、以及操作要點進行說明，使聽者能瞭解如何使用這資料中心。第二部分則以作者的研究實例來具體說明如何使用健康資料進行研究。作者主要是針對台灣社會保險的殘障給付，針對切除子宮所產生的金錢誘因引發的道德風險討論。作者使用歷年健保資料，找出請領年紀限制（45歲）前後5年（40~49歲）的子宮切除者，在區分婦女的投保身分後得到幾點發現。第一、勞保婦女子宮切除比率，自42歲開始明顯增加，在45歲前子宮切除比率達到高峰，約為42歲的3倍左右；但過了45歲後這手術比率迅速下降；第二、相同趨勢也出現在可請領生育失能給付的公、農保女性中，惟農保婦女的波動遠比公保婦女來的小。第三、對無法領取失能給付的女性投保人（非勞、公、農保），這些婦女子宮切除比率並沒有隨年齡而明顯波動。第四、若以50歲為基準計算，公、勞保婦女在40~49歲已切除子宮比率，相較於沒有職業保險婦女高出20％以上。這研究成果對勞委會和銓敘保討論修正勞保和公保殘障給付做出實際政策建議。

**Textual Data Analytics in Finance**

**蔡銘峰**（國立政治大學資訊科學系）

The growing amount of public financial data makes it more and more important to learn how to discover valuable information for financial decision-making. This talk presents our recent studies on exploring and mining soft information in finance, which usually refers to text, including opinions, ideas, and market commentary. This talk will cover the study of financial risk among companies via financial reports, the work of discovering new finance keywords via word embedding techniques, and the demonstrations on our developed web-based information systems, FIN10K and FRIDAYS, to showcase their abilities for textual analytics in finance.

資料治理之發展趨勢與國際案例探討

**許志義** （亞洲大學經營管理學系、國立中興大學）

賴姵瑀（國立中興大學）

資料治理(data governance)是當前全球關注之公共政策課題。尤其在歐盟於2018 年5 月25 日全面實施「一般資料保護規範(GDPR)」之後，更引起產官學各界共同探討如何在資料私有財與公共財之間，得到最適的平衡點。

有鑑於政府是各種巨量資料的持有者，也是資料治理政策的規劃者與執行者，對於如何將物聯網與大數據時代的各種不同屬性之資料，加以分門別類妥適規範，並透過公私夥伴關係(public private partnership, PPP)予以有效應用，發揮有關數據產業(data industry)附加價值，產生創新經濟的正向動能，謀求整體社會最大的福祉，實為當務之急，更屬責無旁貸。

綜觀歐美重要先進國家及世界銀行、世界經濟論壇等國際組織，已倡議並推動「開放政府資料」多年，以此作為政府服務創新與提升國家競爭力之重要策略。善用民間資源與創意創新，利用雲端運算與邊界運算(edge computing)，整合運用政府公共財開放資料與民間私有財資料，發展各項跨機關跨部門之創新服務與便民服務，並建構以巨量資料為導向的新經濟生態系，已蔚為國際潮流。上述有關資料治理之課題、案例及新趨勢，乃本文之研究動機與探討主題。

本文首先探討資料治理之重要性與經濟意涵，接著藉由探討先進國家開放資料與治理政策之實際個案，印證當代公共政策與私人抉擇，均積極朝向以巨量資料為決策基礎之典範移轉大趨勢。最後，本文歸納出結論與建議。

關鍵字：資料治理、開放資料、大數據、智慧揭露、一般資料保護規範、歐盟S plan、Uber 運動、透明加州、GDPR