

出國報告（出國類別：國際會議）

2018年美國聖地牙哥預測分析創新峰會 出國報告

服務機關：行政院主計總處主計資訊處

姓名職稱：洪理俊分析師

派赴國家：美國

出國期間：107 年 2 月 5 日至 107 年 2 月 11 日

報告日期：107 年 4 月

摘要

本總處為提升主計行政效能及品質，運用創新資訊科技，發展及推動共通性主計資訊系統，並透過系統集中維運環境彙集全國主計資料。因此如何運用數據預測分析技術及相關運用案例，發揮主計資料決策支援功能，促進政府預算資源妥適分配，將是本總處持續關注與發展的資訊服務。

創新商業媒體公司(Innovation Enterprise)舉辦之「預測分析創新峰會」係全球著名資訊創新服務應用研討會之一，今(2018)年研討會，共有22個主題及超過 100 名學界與業界專家分享預測分析應用技術與案例。參與本次研討會目的係希望藉由各界應用技術及案例之分享，能協助本總處找出主計資料運用方向，並作為日後創新主計資訊服務之參考。

目 次

壹、目的.....	3
貳、過程.....	4
參、重要內容摘要.....	8
肆、心得與建議.....	22

壹、 目的

隨著網際網路的蓬勃發展及資訊科技的快速演進，從 Google AlphaGo 到 Chatbot 聊天機器人、智慧理專、精準醫療、機器翻譯…，數位化資料的大量累積，促進人工智慧領域的機器學習 (machine learning) 技術發展迅速，機器已經能從大量資料中自我學習，識別數據內的規律，作出高度準確的預測。因此，各行各業如金融、零售、廣告、教育、物流、製造…，幾乎所有產業無不卯足全力，希望能從大量數據中找出模式，以預測風險或未來的營業狀況，早一步採取因應之道。

預測分析創新峰會為學界與行業領導者一個交流與分享平台，與會者以交付案例研究及相關專業知識等方式，分享其運用當今數據預測未來的獨特見解。對於各組織而言，僅利用最新預測分析技術與工具，絕無法成功預測未來趨勢，惟透過深入瞭解業務利基與困境，才可充分發揮新技術的優勢，為組織拓展業務及創造更大利潤與效益。參與本次會議，可於本總處導入預測分析相關技術時，提供有意義的實務案例，以協助本總處達成具前瞻性之資訊應用，並對於發展未來主計資訊服務能有所幫助。

貳、 過程

本次研討會於聖地牙哥舉行，共計有22個主題，由學界與業界專家分享預測分析應用技術與案例，其所參加之會議議程、場次及部分照片集錦如下所示：

研討會第1日議程（107 年 2 月 6 日）

時間	主題	演講者
08:00	_Breakfast	
08:50	Chairperson Overview	Seyed Sajjadi — Data Scientist at EA
09:00	Joint Presentation: Advanced Analytics at The Dow Chemical Company	Ameya Dhaygude — Advanced Analytics, Data Scientist at The Dow Chemical Company
09:30	Defining Strategies to Integrate Analytics with Business Intelligence Platforms	Ruben Quinonez, PhD — Associate Director, Big Data at AT&T
10:00	Embedding Predictive Analytics into business processes to drive operational value	Dallas Crawford — Advanced Analytics Executive at QueBIT
10:30	_Networking Coffee Break	
11:00	Understanding Customer Behavior with Predictive Analytics & a 360 Degree View of your Customers	Rich Fox — VP, Data Science & Analytics at Apex Parks
11:30	Avoiding Bias in Predictive Analytics	Federica Pelzel — Director of Data & Analytics Platforms at MasterCard
12:00	Lunch	
13:00	Deep Learning Fast Start	Alfred Essa — VP, Research & Data Science at McGraw-Hill Education

時間	主題	演講者
13:30	Intersections of the Marine Corps Planning Process and Data Science	Sean Forester — Director, Data Analytics at FSMAO West
14:00	Data Intelligence: Data-Driven Healthcare	Mike Thompson — Director, Enterprise Data Intelligence at Cedars-Sinai
14:30	The Challenges and Complexities of Retention at Sam' s Club	Eric Blabac — Director, Decision Science & Marketing Analytics at Sam's Club
15:00	Networking Coffee Break	
15:30	Predictive Analytics with Big Data in Health Care	Kenny Choy — Dir, Healthcare Data Science Systems at IEHP
16:00	Predictive Analytics: Developing Service Recommendation Systems	William Ford — Director, Data Science at Chegg
16:30	Navigating through Machine Learning Challenges	Ankur Uttam — Advanced Analytics Lead at PlayStation
17:00	Networking Drinks Reception	

研討會第2日議程 (107 年 2 月 7 日)

時間	主題	演講者
08:00	Registration & Breakfast	
08:50	Chairperson Overview	Rajat Agarwal — Senior Director, Investor Analytics at Lending Club
09:00	An EPiC adventure infusing analytics into Cottages.com	Jeremy TerBush — SVP, Analytics at Wyndham
09:30	Financial Time Series Forecasting Using Recurrent Neural Network	Jeffrey Yau — Chief Data Scientist at AllianceBernstein

時間	主題	演講者
10:00	Networking Coffee Break	
10:30	Analytics Powered by Machine Learning	Punnoose Isaac — Exec Dir, Analytics Development at Edmunds.com
11:00	Data Driven Cities	Brendan Bailey — Data Scientist at City of Los Angeles
11:30	Who Thought Binary Classification Would Be So Complicated!?	Seyed Sajjadi — Data Scientist at EA
12:00	Lunch	
13:00	P's of Data Science: Planning Collaborations to Create Products from Data	Ilkay Altintas — Chief Data Science Officer at San Diego Supercomputer Center
13:30	Mining Data to Optimize Design: Analytics for Targeted Affect and User Behavior	Elizabeth Owen, PhD — Director, Data Science at Age of Learning
14:00	Joint Presentation: How Can We Sell More Vehicles? Leveraging Predictive Analytics to drive customer retention and sell more cars.	Jeremy Coltin — Director, Predictive Analytics at Hyundai Capital America
14:30	Networking Coffee Break	
15:00	Applications of Predictive Analytics in Marketplace Lending	Rajat Agarwal — Senior Director, Investor Analytics at Lending Club
15:30	Forecast of Financial Measures from Opinions using Machine Learning	Yves Chauvin — Director, Investment Data Platform at AXA
16:00	End of Summit	

研討會會議照片集錦



參、重要內容摘要

本次研討共區分 22 個研討主題及100多位專家探討最新預測分析技術應用研討會，大部分參展主題及部分應用研討會議題與預測分析、機器學習及資料探勘相關，其重要內容摘要如下：

(一) 透過預測分析360度全方位視角了解客戶行為 (Understanding Customer Behavior with Predictive Analytics & a 360 Degree View of your Customers)

Apex Parks Group是一家在美國各州擁有14座家庭娛樂中心的公司，為提高各地遊樂園的盈利能力，Apex 於2012年起開始利用預測分析技術提升產品獲利，透過遊樂園每月、每季、每年實際銷售資料，進行預測模式的比較，預測誤差是以近期、頻率、消費力 (Recency, Frequency, Monetary)三項指標進行評估，經由人工智慧、機器學習等技術找出資料規律性，建置客戶生命週期(Lifetime Value, LTV)營銷分段圖(如圖 1)，以預測模組分析工具評估新進(New)、經常(Repeat)、不經常(Infrequent)、流失(Lapsed)、無效經常 (Inactive repeat)等客戶群對提升LTV值(客源消費能力)的效用，協助找出影響獲利的因素，以因應投資人力、調整產品流程及規劃行銷策略，獲取最大利潤(如圖 2)。

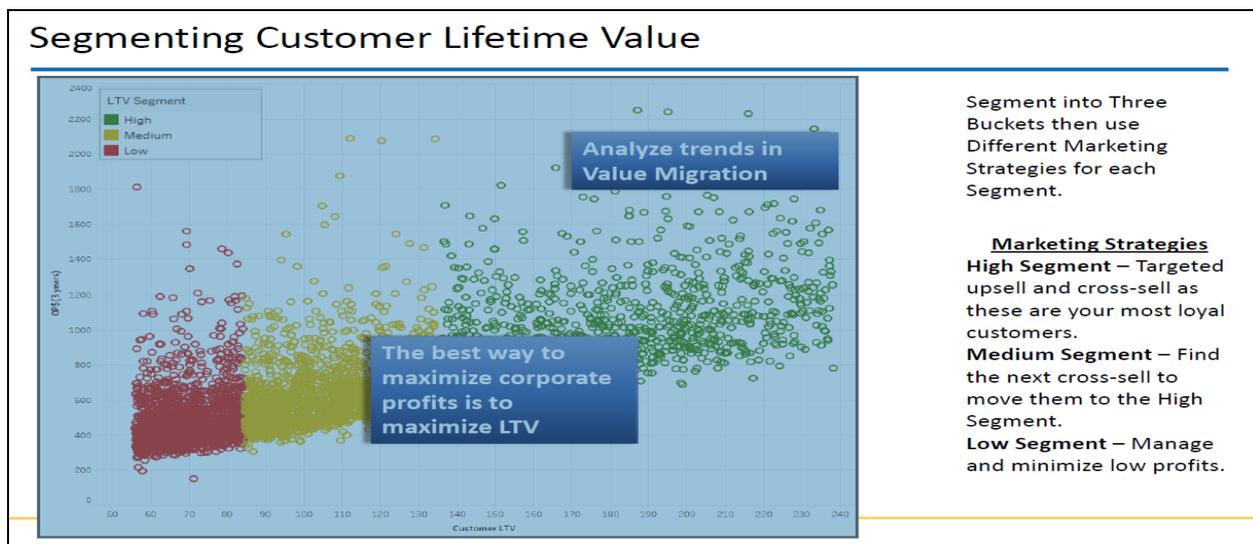


圖 1、客戶生命週期(Lifetime Value, LTV)營銷分段圖

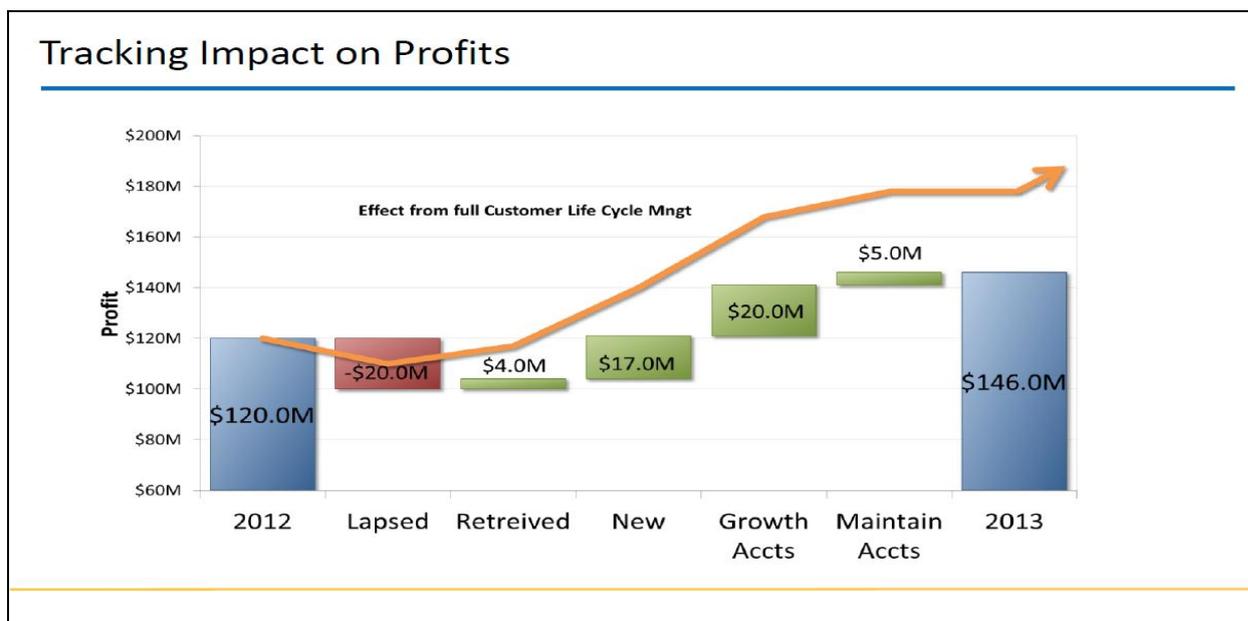


圖 2、產品流程及規劃行銷策略，獲取最大利潤

(二) 陶氏化學公司的先進分析技術 (Advanced Analytics at The Dow Chemical Company)

陶氏在全球的化工業務覆蓋面廣，包括56,000名員工，跨越34個國家的189個分支機構，為其客戶提供7,000多條產品線，這些產品線遍及175個國家和近十幾個不同的商業市場。陶氏公司一直被評為財富100強公司，其收入主要來自企業對企業銷售，且只有少數幾個市場的產品直接銷售給最終用戶。

本次主題討論重點為影響這樣一個範圍和規模的公司的機會和挑戰。主講人資料科學家Tim Licquia和Ameya Dhaygude利用資料科學來強調陶氏在各業務的複雜操作模型，藉以闡述陶氏目前的分析樣貌，這些論點也同樣獲得專業社群中的資料科學家認同。通過對「層級式時間序列的組合預測」案例(如圖3)研究得出結論，經證實對陶氏廣泛複雜的業務皆能節省成本，並且實現規模效益。

Several Models – Sample Hierarchy

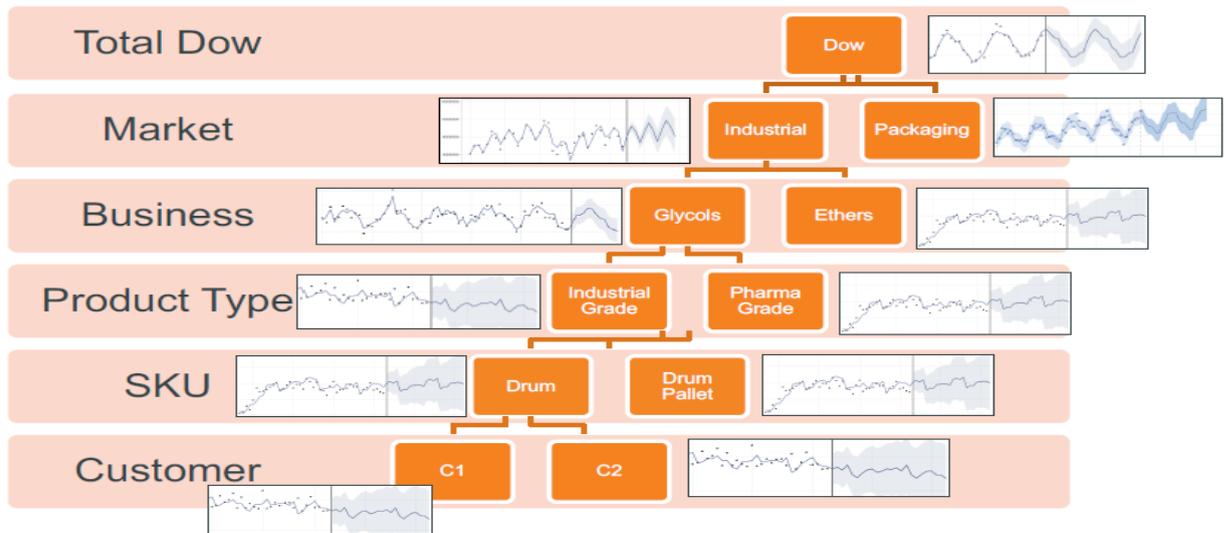


圖 3、層級式時間序列的組合預測

(三) 資料智慧：資料驅動的醫療保健(Data Intelligence: Data-Driven Healthcare)

Cedars-Sinai 是美國最大的非營利學術醫療中心之一，擁有4,500多名醫生、護士和其他醫療保健專業人員，平時除致力當地社區及外地病患提供醫療照護外，也提供醫療保健教育計畫以及廣泛的社區護理服務活動。

Cedars-Sinai於本次研討會說明巨量資料分析在過去與未來技術的差異性(如圖4)，並介紹如何利用巨量資料分析與深度學習科技，將先進的醫學研究成果，持續改善社區醫療照護品質，替民眾節省可觀的醫療成本開支。此外，他們還將預測分析技術運用在急診部就診量預測，透過收集醫療機構、藥品供應、病患體檢、醫療費用等資料，並結合醫學臨床、電腦科學、網路通訊、統計、財務等各專科領域專家知識，建立預測模型分析急診人數與季節、假日、天氣、地區各因子間的關係，得出醫療費用支出的預測值(如圖5)。

Healthcare's Past		Healthcare's Future
<ul style="list-style-type: none"> • Historical reporting - can only look at what has happened • Silo- knowledge in a handful of users 		<ul style="list-style-type: none"> • Predictive analytic models - based upon data patterns • Data Mining tools that can automate feature discovery
<ul style="list-style-type: none"> • Ability to view one patient at a time • Ability to view one patient visit at time 		<ul style="list-style-type: none"> • View all visits for a patient to see patterns and repeat visits • View populations of patients that share common characteristics • View patient data outside of hospital walls combined with internal data
<ul style="list-style-type: none"> • Limited range of data • Inability to accommodate data growth • Inability to store data that is not 'structured' to fit within current data architecture 		<ul style="list-style-type: none"> • Data growth of 300% per year - with a focus on integrated new data sources (weather, genomic, environmental, socio-economic, Fitbit)
<ul style="list-style-type: none"> • Users given a "dump" of all data - users sort through data on their own • Inability to interact with the data to explore hypothesis 		<ul style="list-style-type: none"> • Subject matter analytic apps built to answer specific questions • Focus on the appropriate metrics - reducing information overload • Machine Learning Techniques to derive most relevant data relationships
<ul style="list-style-type: none"> • Data/Analytic tools only used centrally • Data access limited • Numbers only - no visual analytics • No statistical capabilities 		<ul style="list-style-type: none"> • Data/Analytic tools used by end-users resulting in greater agility • Data access (with data governance) available to end-users • Focus on visual analytics and self-service

圖 4、巨量資料分析在過去與未來技術的差異性

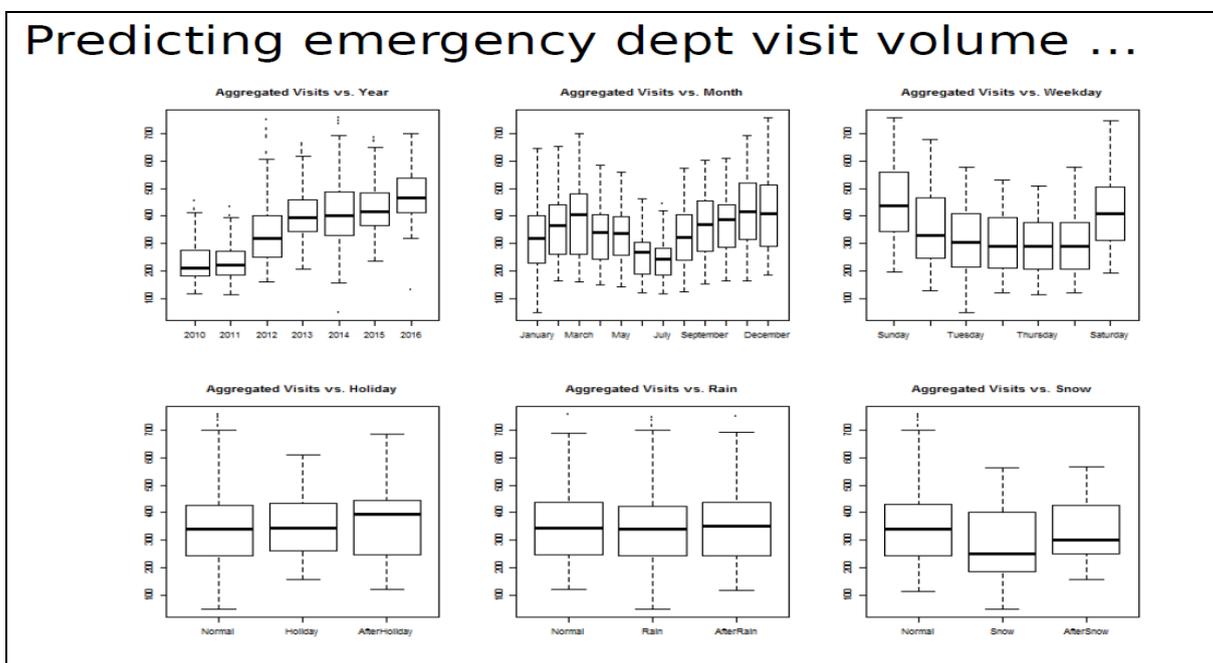


圖 5、醫療費用支出的預測值

(四) 制定結合分析與商業智慧平台的策略

分析和商業智慧平台的整合對於許多組織來說可能是一項艱鉅的任務。商業智慧平台

（如資料倉儲）通常被視為需要全面檢查的舊解決方案，而分析平台（例如資料湖泊 Data Lakes）被視為取而代之的代品。分析平台的資料湖泊架構有別於傳統的資料倉儲，傳統的資料倉儲的資料通常是品質較高且是被預先處理過的資料；而資料湖泊架構的設計是可擷取大量的、各種類型的資料，作為資料素材的儲存管理，以利分析應用，正因為資料湖泊架構範圍更廣、在資料分析上擁有更多彈性，很適合做為導入巨量資料分析應用的架構藍圖。

資料湖泊資料可包含各式的MPP 資料庫、In-Memory 資料庫及 HDFS 分散式儲存資料庫，各種來源、各類型的資料在資料湖泊經由不同方式的儲存、處理、淨化、管理後，有彈性的產出各種分析資料。不同的是，資料湖泊可同時包含「未被清理的資料」（Unclean Data），保持其最原始的形式，故分析者可取得最原始模式的資料，以減少資源上處理數據的必要，讓來自各機關的資料來源更易於結合。

資料湖泊主要有四點特性：以低成本保存巨量資料（Size and Low Cost）、維持資料高度真實性（Fidelity）、資料易取得（Ease of Accessibility）及資料分析富彈性（Flexible），但新一代資料庫技術，並不會去破壞現有大企業習於使用的 OLTP 及OLAP 等結構化資料處理與分析的市場。本次議題中，與會者學習將舊版商業智慧與分析平台集成，而不是取代舊策略的策略；研究顯示資料湖泊和資料倉庫是互補的(如圖 6)，而不是競爭的解決方案。

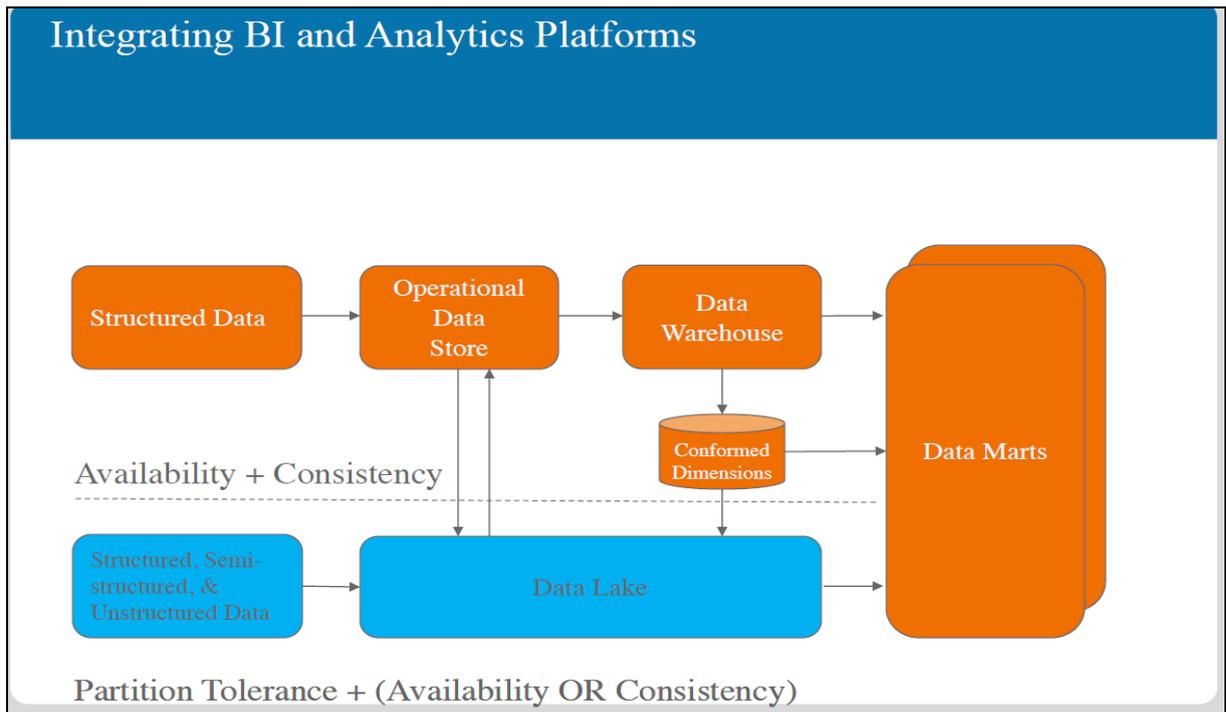


圖 6、整合 BI 和分析平台

(五) 預測分析融入業務流程以提升運營價值(Embedding Predictive Analytics Into Business Processes to Drive Operational Value)

為提升供應商獲利能力，QueBIT 協助客戶將供應鏈時事因子之動態價格預測分析技術導入商業流程，提升營運價值。傳統的價格預測分析多以歷史價格與經濟模型，分析過去的價格趨勢或推導出合適的機率模型，並預測未來的價格變化區間，但實際應用時，卻發現**時事**往往才是影響價格變動的主因，也牽動理論模型的預測成效。QueBIT「結合供應鏈時事因子之動態價格預測分析」巨量資料分析模型，探討供應鏈之上下游原物料／產品，或相關／競爭物料之事件發生，對標的物價格的影響，並建立事件對標的物價格影響的量化模型，透過事件所產生的供需變化，預測標的物未來的價格波動。以辦公家具業物料為樣本試驗場域，透過QueBIT「需求輸入營運調整」(圖 7)數據分析技術，有效提升家具物流產品(Bookcases、Office Suppliers)營運分析預測準確度(圖8)，為企業組織提供可信賴的決策輔助。

Optimization Review & Adjustments

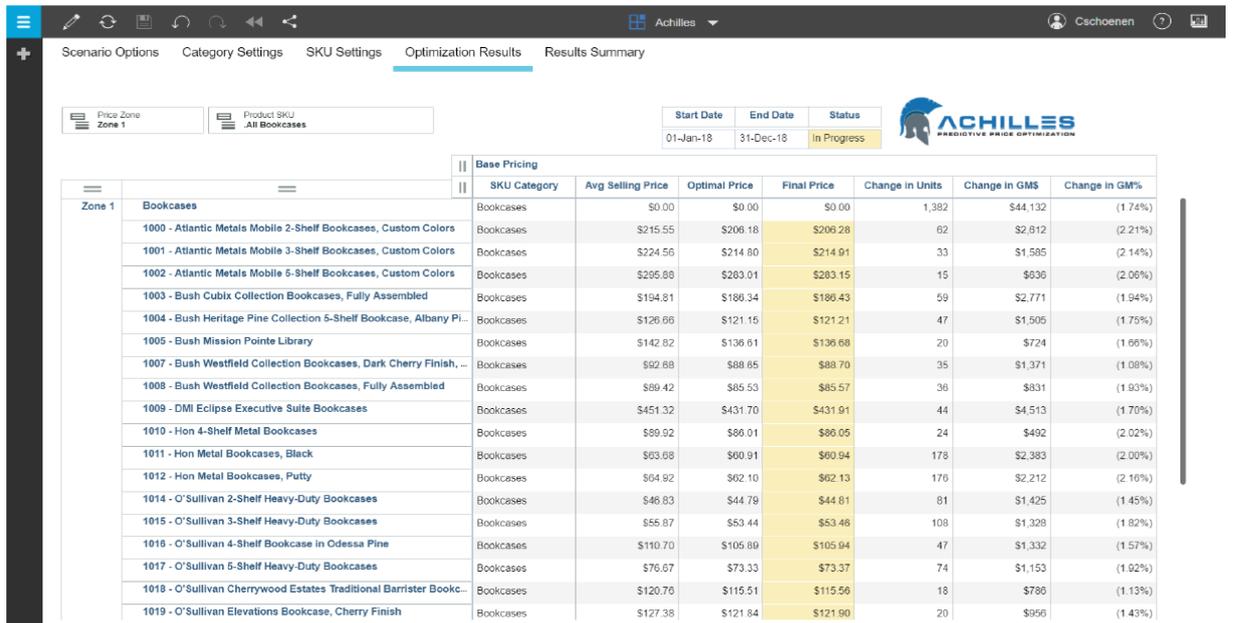


圖 7 需求輸入營銷調整

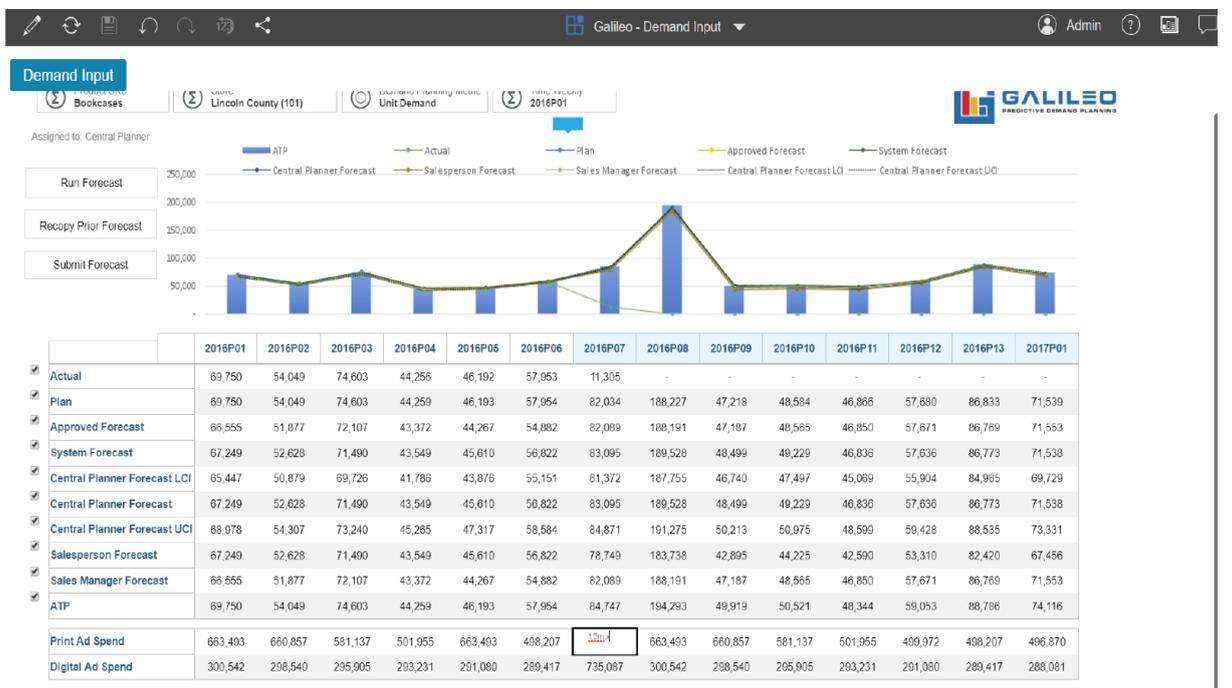


圖 8 營運預測分析報表

(六) EPiC 融入性分析在 Cottages.com 的應用(An EPiC adventure infusing analytics into Cottages.com)

數據可以說是 Cottages.com 的衣食父母，用以確保對顧客做個人化的推薦、價格的優化、鎖定目標客群等。對 Cottages.com 來說，這些數據就像神奇的魔法藥水，讓他們與顧客之間的關係更加緊密，像是「價格清單」、「雇員推薦」、「租賃歷史」等。利用 EpiC 全球分析能力(圖9、圖10)來幫助優化我們的價格並推動整個賽季的變化這改善了別墅出租業主和客戶的入住率和出租別墅淨收入。透過機器學習技能和創建預測模型的工具也將幫助組織更好地管理財產招聘和保留客戶(例如「租賃同樣別墅的人，也看了什麼度假別墅」、「客戶曾經瀏覽過的度假別墅」)，並使分銷更具成本效益。

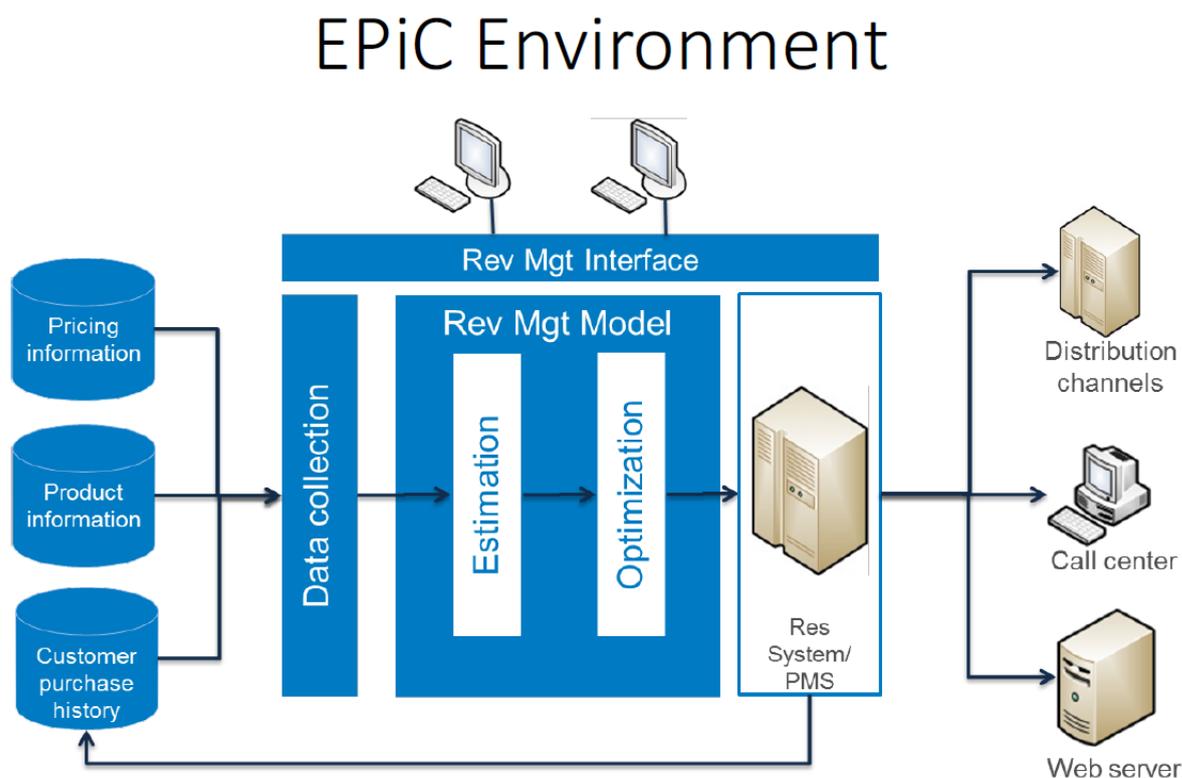


圖 9、EpiC 系統架構

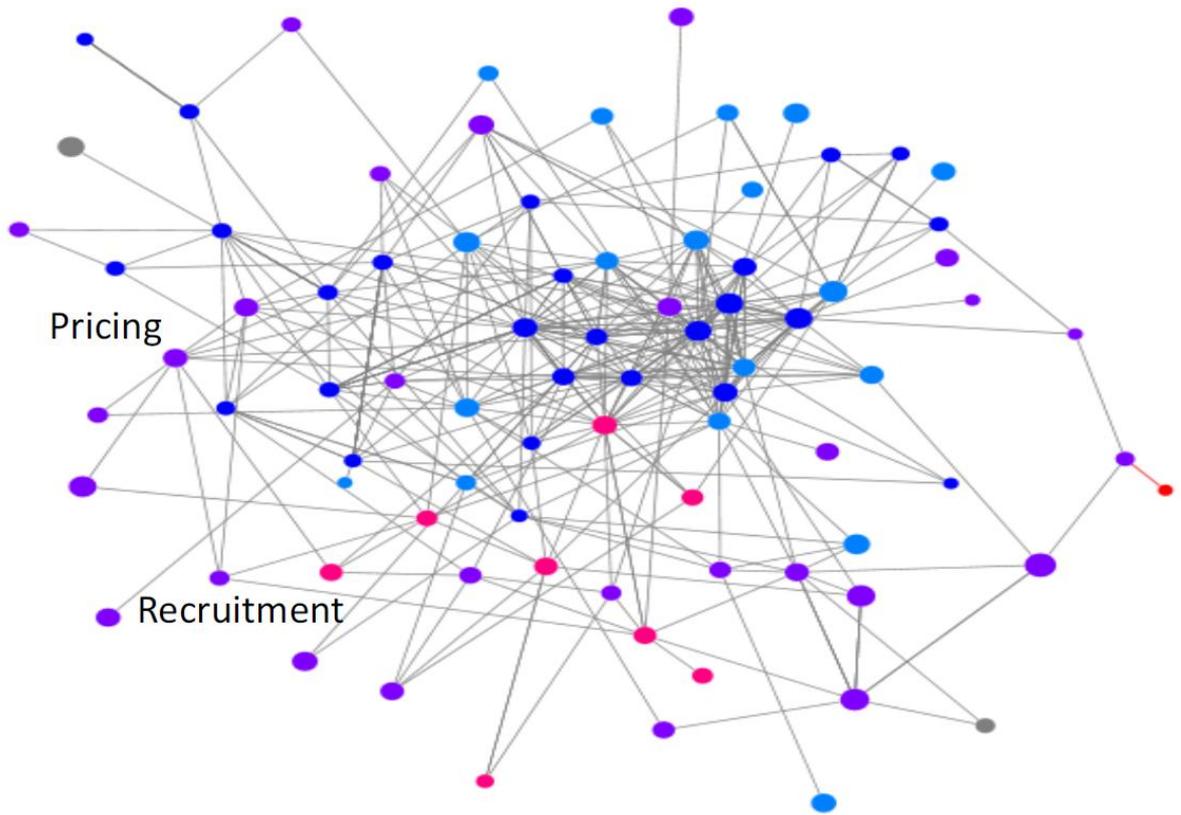


圖 10、價格、雇員的 3D 立體分析圖

(七) 數據驅動城市 (Data Driven Cities)

加州洛杉磯市長授權的創新團隊如何利用資料解決警方招聘和無家可歸問題。主講人在洛杉磯辦公室負責該市市政數據化推動，其主導的創新服務團隊透過密集的資料分析，為警局公平錄用組提供數據分析和預測，使之即時調整錄用程序，以保持警隊適當性別和種族比例。

講者並強調全國各地和全球各地城市都建立了一個名為「創新團隊」的跨職能內部諮詢團隊，以解決當地的重大問題。包括建立執行級資料儀表板、可視化地圖，並建立預測模型，幫助利益相關方就未來作出明智的決策。另外，講者也強調市政服務工作或任務可釋放給業界去發揮，才能發揮服務最大的效用；業界在服務創新的能量上遠大於政府機關，這些服務工作包括自動化流程，以便市政工作人員可以專注於高價值任務，而非重複性任務。歸納幾點作法如下：

- 政府部門應跨部門結合產官學界雲端運算、演算法、深度學習、大數據分析、人工智慧應用等核心技術及研究領域專家學者，共同成立數據分析預測創新研究中心，主要目的在於鼓勵學業界投入數據分析預測創新技術及應用研究，培育數據分析人才並打造眾多創新服務工作團隊。
- 由政府公部門主導建置開放資料預測分析中心，引進前述創新服務工作團隊，以協助各界民間或網路社群，從公部門開放資料平台取得資料，進一步對資料進行訓練或預測，再將預測分析結果提供給其他民間業者或一般民眾加值運用，並回饋公部門的資料開放平台，達到擴大資料應用範圍及提升服務效率雙重價值。
- 為解決外界具有創意的人才因不會使用資訊工具及技術，致埋沒創新應用服務貢獻機會，應積極推動數據分析創新應用職能教育，營造人才培訓和交流環境，讓資料分析技術可以在民間各階層生根，透過合作把有創意的服務實作出來。
- 透過個案導入引進外界創新服務工作團隊人才知識及經驗，培養公部門業務專家及外部預測分析專家合作之默契，並於內部效仿學習數據分析預測相關知識及技術，逐步累積公務人員的資料預測分析知識、經驗及能力，當預測分析能力獲得提升後，讓資料分析技術真正內化成為政府管理決策的一環。

(八) 機器學習在預測分析中的應用((Analytics Powered by Machine Learning))

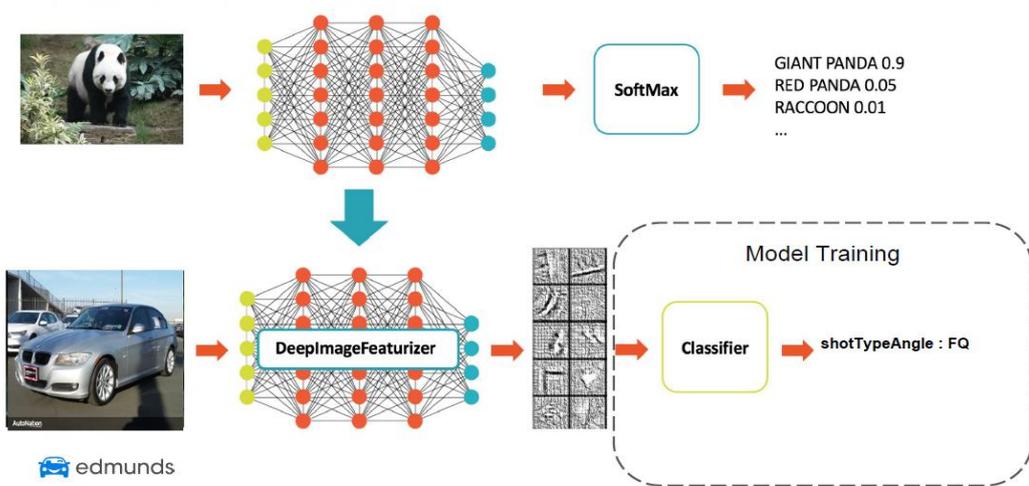
自網際網路和各式行動裝置普及之後，每天都有超過一百萬TB的數位資料產生，其中有一大部分是數位影像資料。大量的數位影像資料如果經過適當的自動化處理、抽取出其中的資訊，就能成為貼心的服務、發揮出數位資訊驚人的妙用。從基本的手寫文字辨識、物件識別、人臉辨識，到自動化圖像描述(Image Captioning)、無人駕駛車(Self-Driving Car)，還有最新的馬賽克還原技術，都是機器學習和影像辨識整合後的應用。

- Edmonds公司以機器學習在影像辨識上的實例(圖 11)，透過人工智慧(Artificial

Intelligence)和機器學習(Machine Learning)來處理影像辨識，幫助客戶徹底改變購車體驗。

- 這是一個時代的轉變，它定義了什麼樣的分析對”現在”(非”將來”)來說是可靠的。隨著像 SageMaker和 CloudML這樣的工具的出現，機器學習技術已經不再掌握在少數幾位資料科學家的手中，只要能夠取得資料的分析師，如今皆可使用機器學習去訓練模型，並且建構生產系統。雖然業務問題可能相同，但解決方案已不同。為了保持相關性並建立市場上一流的分析產品，我們必須擁抱這種思維方式。
- 隨著2018年到來，機器學習展現的力量已滲透到我們所做的一切。透過主題模型預測價格變化、主題模型建置改善網站導航，根據預測的領先收盤價動態重新定價經銷商，在金融股票投資領域中也能創造高價值客戶群以獲得更高的投資回報等，這些都是透過機器學習提供的創新服務應用。

Image classification: Re-using what you've already learned



36

圖 11、影像辨識

(九) 預測分析：開發服務推薦系統 (Predictive Analytics: Developing Service

Recommendation Systems)

在過去的十年中，Chegg從一家低成本課本出租的零售公司發展成為ed-tech的主要品牌，現在的幾條業務線除提供租賃服務和靜態內容外，尚提供許多其他服務。圍繞P2P教育的預測分析專業系統是ed-tech必須開發，以便在企業重整時維持產品差異化，Chegg也順勢將服務推薦系統視為內容推薦系統的下一代系統發展。

Chegg於2001年開始從事分散式運算模組建置，創新團隊在過去兩年於產品、業務和工程等領域的交集點，專注先進的教學者關鍵績效指標(KPI)預測分析技術(圖12)，提供結合教學專業(經驗)、課程取得、教學效果等評估方法改善業務成果，以支持組織快速增長。

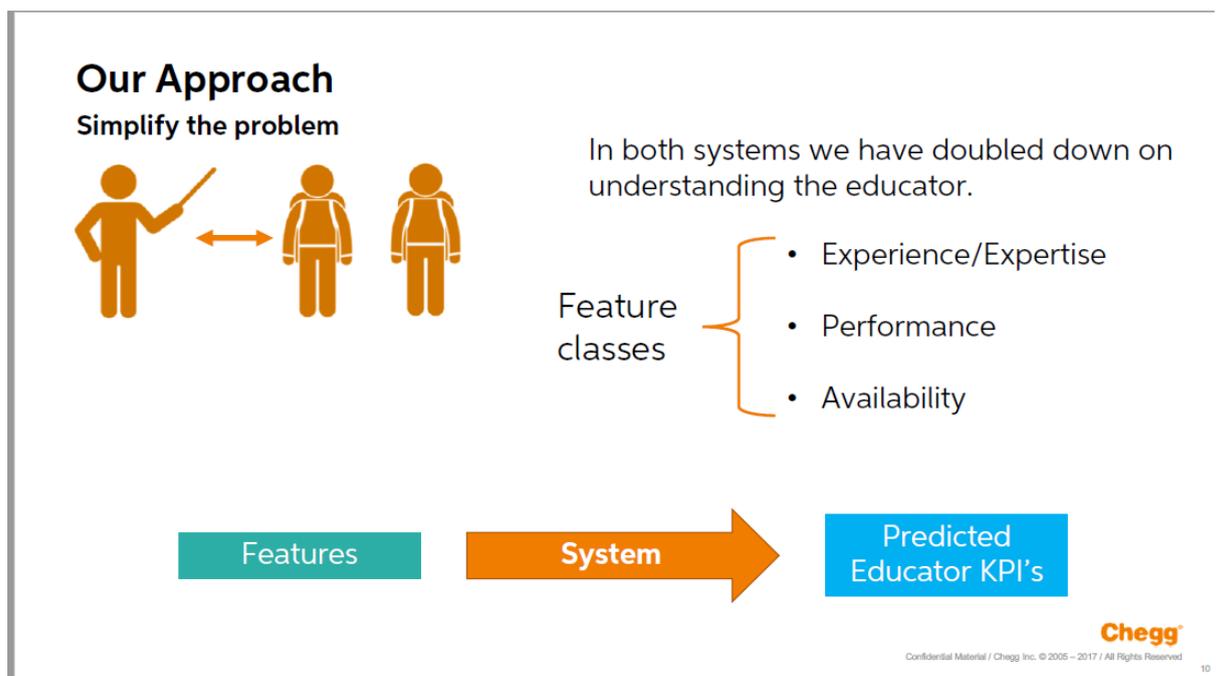


圖 12、關鍵績效指標(KPI)預測分析技術

(十) 資料探勘優化設計：針對預設目標影響和用戶行為進行分析 (Mining Data to Optimize Design: Analytics for Targeted Affect and User Behavior)

設計理想的使用者行為的數字體驗對於資料探勘成功至關重要，特別是對於具有數字產品或面向客戶端門戶之企業。當期望行為是一個難以衡量的結構（如情感和時刻參與）和基於成長的行為（例如隨著時間學習）時(圖 13)，這變得特別具有挑戰性。因此多模式資

料採掘（從結構發現到預測以及超越）可以發現新興的行為模式，為優化的用戶體驗提供反覆的數據驅動設計。本次討論提出了這些方法的本體論，在數字教育背景下的大數據，結合數據挖掘、機器學習和統計等技術應用，進行用戶行為分析推斷方法。這些方法在系統不同的開發階段中，對關鍵互動和使用者行為進行反覆的資料探勘優化設計提供重要見解。

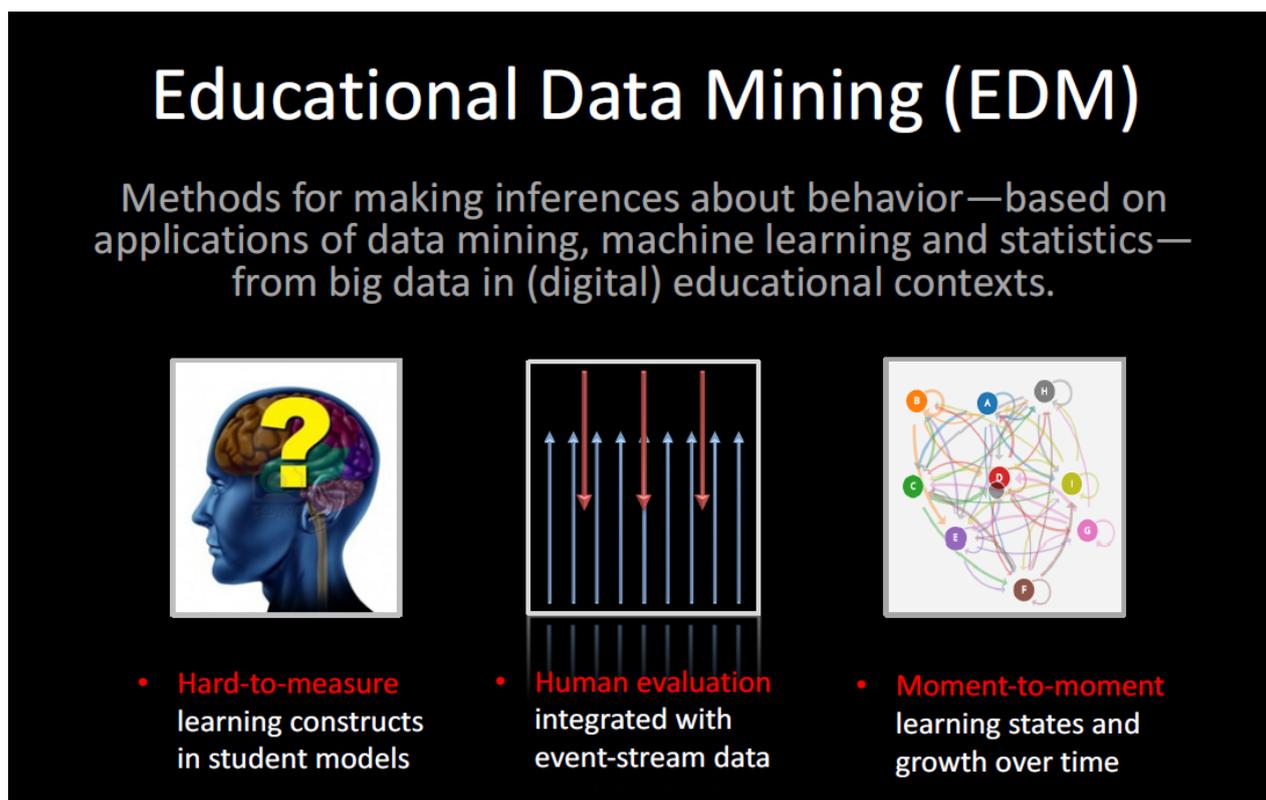


圖 13、EDM 行為推測分析

(十一) 深度學習快速入門 (Deep Learning Fast Start)

深度學習目前已經成為人工智慧和機器學習最有前途的領域之一。本次會議為深度學習和人工神經網路提供“快速入門”介紹，從對人工神經網路的概念性理解開始，對主要技術概念進行介紹，並透過實際程式碼建立預測模型來得出結論。本次會議目的是揭開深度學習的神秘面紗，並為那些剛接觸深度學習的人提供一個快速入門的介紹。

- 深度學習功能強大，但想要改進深度學習的效果常比傳統的機器學習模型更難。深度學習的類神經網路元件可以調整的設定和參數很多，想要在眾多參數選項中

找到一個最好的排列組合，是一項極艱鉅的任務。因此，通常會套用一些現成的模式作為新模式的基礎。

- 深度學習網路就像黑盒子，使用者不理解模型中各網路層的內涵，就無法做出有效的調整。大家應該根據各自數據的不同，選擇相對應的分析預測手段，而深度學習神經網路不該被視為唯一的工具。
- 人工智慧、機器學習、深度學習這三者間緊密聯繫，基本上可稱作彼此的子集。也可以說是深度學習驅動機器學習，最後實現了人工智慧。

肆、心得與建議

在參與多場主題演講後，除對國外的數據預測分析專案的成功經驗印象深刻外，深感資訊科技跨領域創新應用也值得台灣借鏡，對於想深入了解資料探勘、機器學習、預測分析等技術之資訊人員相當有助益，從本次研討會內容所啟發之心得如下：

綜觀本次研討會中各企業及政府部門在數據預測分析上的經驗，不難發現幾乎所有企業普遍認為資料分析人才養成確實不易。以洛杉磯創新團隊以數據驅動城市願景為例，該團隊於組織內部積極培育具數據分析能力的資料分析人才，並結合來自不同領域的專業知識，進行密集交錯的資料分析，讓市政多元化工作流程自動化，以便市府員工專注於高價值的任務而非重複性的任務。本總處刻正導入視覺化分析工具及辦理實務訓練課程，以因應管理分析報表需要，俾利掌控各機關業務執行狀況，適時提供決策參考。未來可利用現有數據分析工具及訓練技術資源，引進外部經驗與支援，並於內部效仿學習數據分析預測相關知識及技術，作為厚植數據分析專業職能之基礎。

本總處掌理全國之歲計、會計及統計業務，累積大量數據資料，如何因應業務發展，有效運用數據分析技術，挖掘主計資料價值，係本總處未來主計資訊應用發展的重要課題。綜合本次研討會獲取外界不同組織企業對於數據分析的成功經驗及數據分析技術發展趨勢等議題，提出以下一些建議以供參酌。

- (一) 強化數據預測分析知能，以因應主計業務發展需求：本總處可透過參加數據分析技術研討會、教育訓練等方式，持續瞭解數據分析技術及經驗，並依據本總處歲計、會計及統計之業務發展需求，將相關技術運用於現行資訊服務。
- (二) 運用數據分析技術與工具，以視覺化方式呈現歲計會計資料：依據歲計會計業務需要，本總處已建置歲計會計資訊系統支援本總處業務單位及各主計機構辦理預決算報表編製及例行性會計帳務處理工作，未來將依各層級主計機構所提管理需求，運用數據分析技術，增加圖表等視覺化方式，呈現更易於閱讀及理解的財務管理資訊。