

Data, Visualization and EDA

陳君厚

中央研究院統計科學研究所

Abstract

大數據 (Big Data) 與資料科學 (Data Science) 已是大眾耳熟能詳的詞彙，而經由 Drew Conway 的 Data Science Venn Diagram¹ 在網路上廣為流傳，大家也理所當然的認為數學與統計知識 (Math & Statistics Knowledge) 在處理大數據與進行深層分析 (Deep Analytics) 時乃必備之工具。然而大學四年的統計系與多年的統研所課程，老師教的、自己學的夠用了嗎？我回顧了下過去自己在大學四年 (1980~1984) 所修的相關課程，如果老師都認真地教，我也有效率地學習，怎麼會不夠用？再加上這 30 年後新的學習環境：包含統計諮詢、統計計算、機器學習等新的課程，與網路學習。。。。

缺了什麼嗎？我個人覺得是 **對資料的感覺 (sense about data)**

John W. Tukey 在探索式資料分析 (Exploratory Data Analysis, EDA) 一書中開宗明義提到“*It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.*” 學習可以做什麼，有助於在資料分析的過程中達到事半功倍的效果。EDA 的作用在於從「看」資料獲得資料所傳達的訊息，所注重的是簡單的算術與容易建構的圖、表。透過 EDA 對於圖表中所顯露的型樣 (pattern) 做一初步的認知與描述，再進一步以人類的心智 (human mind) 對所接受的訊息做全面的分析與判斷，以探索潛藏於資料中的訊息。強調的是探索式的分析而非嚴謹的模式確認。

本演講前半段將由講者帶大家實際操作一個講者已使用 25 年的互動式與視覺化的探索式資料分析軟體：Data Desk，提供大家一個容易對資料有感覺的 EDA 環境。若還有時間後半段將介紹相關研究團隊多年來發展之矩陣視覺化 (Matrix Visualization) 環境：Generalized Association Plots (GAP 廣義相關圖)。本演講將經由各種實例介紹 Data Desk 7 與 GAP 工作環境。

- 有興趣參加之同學請至依指示於課前裝好 Data Desk 7 軟體 (Special Edition, Expires July 1st, 2015) 與 demo data sets :
- 欲使用 GAP (continuous and binary version) 同學請至以下網頁下載 Window version:

<http://gap.stat.sinica.edu.tw/Software/GAP/index.htm>

¹ Drew Conway 的 Data Science Venn Diagram:

(<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>)