巨量資料與統計分析 Fall 2024

Assignment #1 10/18/2024 Due

1. 不當使用大數據建立的模型，可能威脅個人生計、甚至造成重大損害，2008年金融大海嘯即是值得我們警惕的教訓，這類型大數據模型大多具有三項特質：「不透明」、「大規模應用」、「造成傷害」。請各組舉出實例說明不當地套用大數據模型，會衍生出那些不公平現象及社會問題。（註：參考《大數據的傲慢與偏見》或《Weapons of Math Destruction》。）
2. 大數據時代為各產業帶來新氣象，對資料分析的需求高於先前資料採礦(Data Mining)時代，其中需要統計專業的協助，但也對統計發展造成潛在威脅。請各組參考書籍、期刊及各國統計學會網站，整理相關研究論文後提出建議，說明現代統計學家（或資料科學家）需要哪些訓練，以及如何區隔統計、資訊科學扮演的角色。（註：請詳細註明參考文獻及其出處。）
3. 資料品質往往比資料量更為重要，Google Flu Trends 就是一個典型範例。
   (a) 請至 Google 公司蒐集資訊，建議可提高流感盛行預測正確性的方法。
   (b) 合適的資料蒐集方法及樣本很重要，我們可從傳統抽樣理論中激盪出新想法。請從「未來事件交易所」、「韓流怎麼造出來的」或網路討論區挑選一個範例，說明如何挑選合適樣本，並討論這種想法的其他可能應用。
4. 資料偵錯、敘述性統計量（集中趨勢、散佈趨勢）、變數的關聯性分析，大致可視為資料探索性分析（EDA）的三個重要單元，透過圖形及表格的視覺化可提供我們重要資訊。請分析血液基金會的捐血問卷，並分別執行偵錯、敘述性統計量、關聯性分析，依序回答以下問題。
   (a) 各組說明資料偵錯的原則，並舉例如何處理有問題的資料。
   (b) 以「一、基本資料」為例，比較臺北與其他捐血中心。（註：加總組員的學號後兩碼，除以五後之餘數分別選取新竹、高雄、花蓮、台中、台南為臺北的對照組。）
5. （問題：抽樣理論在大數據的角色？）延續上一題(b)的分析，隨機抽取 1%、5%、10%資料，再與全部資料的分析結果比較，討論抽樣是否可用於大數據分析。