

A man in a red shirt and tan pants is longboarding down a winding road through a mountainous landscape. The sun is setting, casting a warm glow over the scene.

最後要確認每一頁右上有目前位置  
字體

# 科系、職務與待遇： 大數據時代下的社會學與勞動市場

- 游騰林
- 202011



# 游騰林 數據分析師

---



東華大學 社會學研究所



國泰世華 數據部



<https://github.com/TLYu0419>



提問方式介紹  
別懷疑趕快去下載 App ~

slido

Join at  
[Slido.com](https://www.slido.com)  
**#meet20**

# 今日綱要

迷路的時候可以看看右上角，找到目前在講什麼

介紹

服務旅程

SNA

總結

大標題由左往右出現，  
然後子標題向下依序出現

## 我的 社會學旅程

## 數據 分析

## 傑華老師的煩惱

## 總結

- |           |          |             |          |
|-----------|----------|-------------|----------|
| • 社會學教什麼? | • 分析流程介紹 | • 一份突如其來的工作 | • To 系上  |
| • 畢業之後呢   | • 反面應用案例 | • 資料蒐集      | • To 學生  |
| • 職涯規劃    |          | • 職缺分析      | • 一些學習資源 |

A man in a maroon hoodie and light-colored pants is longboarding down a winding asphalt road. He is leaning into the turn, looking back over his shoulder. The background consists of rolling green hills under a bright, slightly cloudy sky.

# 我的社會學旅程

- 社會學教什麼？
- 畢業之後呢？
- 職涯規劃

# 到底社會學在教什麼>”<

# 社會系課程規劃理念

強調基礎知識之掌握，  
並培養思辨與批判的能力

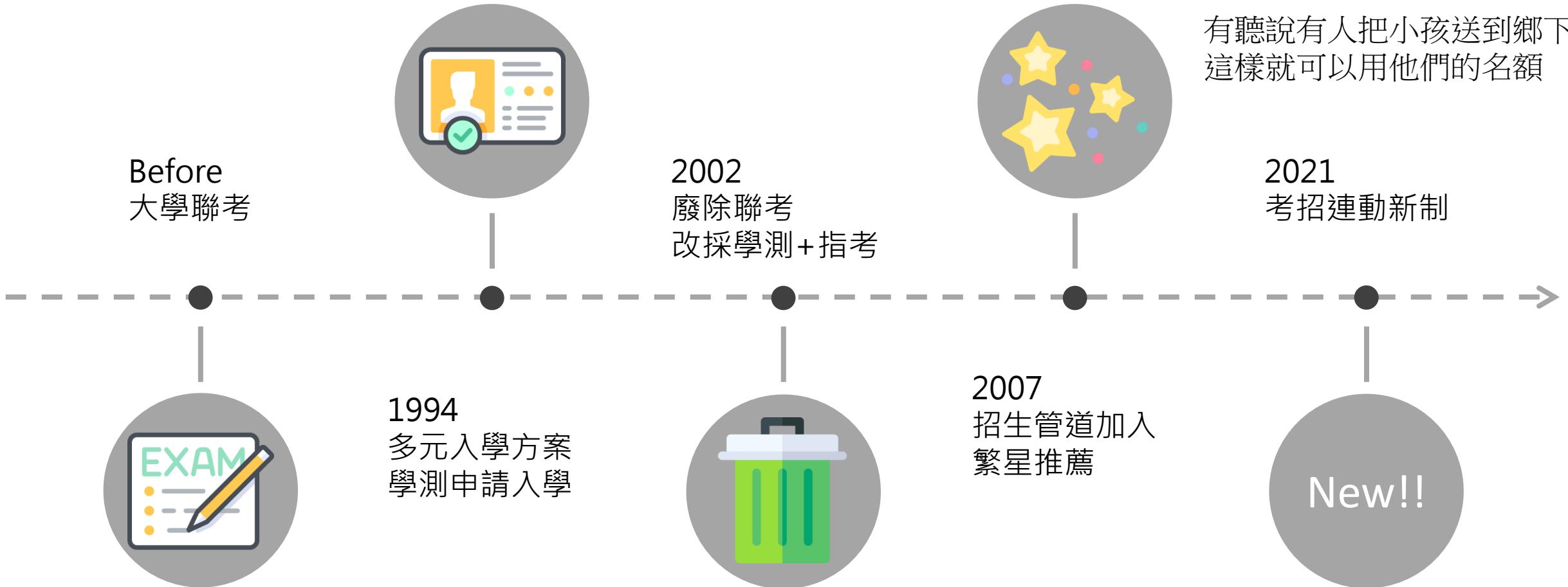


<https://spa.ndhu.edu.tw/p/412-1059-17584.php>

科系、職務與待遇：大數據時代下的社會學與勞動市場  
Mail: TLYu0419@gmail.com

# 為什麼要教育改革?

教改的目的是希望讓學生獲得的教育機會均等，但…有嗎？



- Ref: 1. 駱明慶. (2002). 誰是台大學生?-性別, 省籍與城鄉差異. *經濟論文叢刊*, 30(1), 113-147.  
2. 駱明慶. (2018). 誰是台大學生?(2001-2014)-多元入學的影響. *經濟論文叢刊*, 46(1), 47-95.

# 為什麼要年金改革? 改、不改以及怎麼改？

社會保險年金模式 (social insurance pension model)		多層年金模式 (multi-pillar pension model)	
政策目標		所得維持	濟貧
公共年金體系	制度資格水準	繳費式的社會保險 就業地位 高，根據繳費年限	均一給付的基礎年金 公民資格或資產調查 低於貧窮線
財務設計	隨收隨付制度	稅收或隨收隨付制度	稅收或隨收隨付制度
私人年金體系	低度發展	高度發展，且多樣化	國家規範的私人年金
代表國家	德國、法國、希臘、義大利和台灣 日本與韓國、美國	英國和紐西蘭等	瑞典、荷蘭、丹麥等北歐國家

- Ref: 1. [年金改革：笨蛋！問題在制度！](#)  
2. [臺灣年金制度改革之內涵與思考](#)

# 太陽花學運在吵什麼？ 吵什麼？

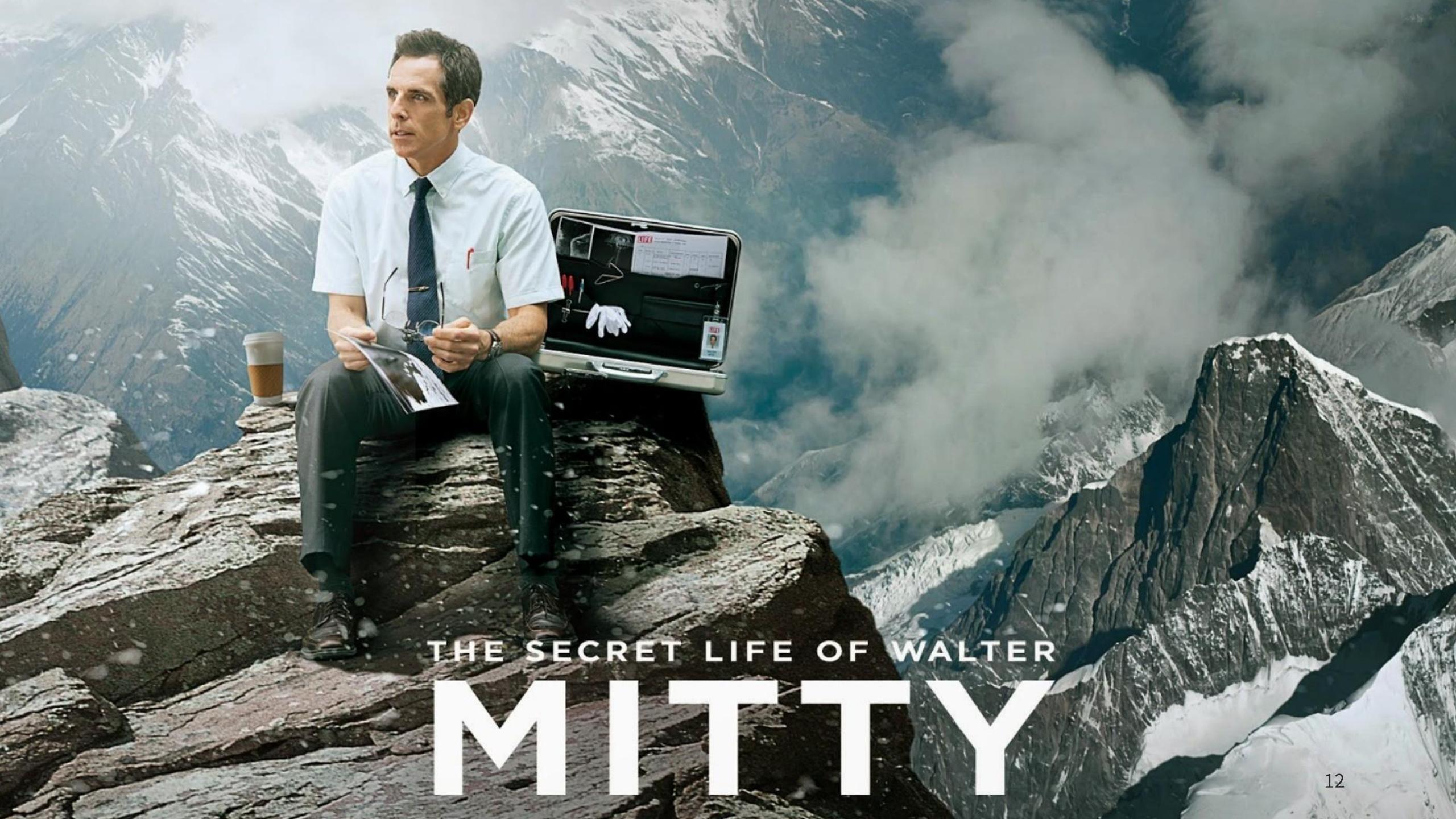


社會學不是反對的社會的經濟成長，  
而是希望討論清楚誰會從中受益，誰又會被犧牲？  
對於被犧牲的產業/群體有沒有對應的補償/輔導機制？



Ref: 1. [社會學在衝啥？在照亮黑箱政經結構](#)  
2. [〈島嶼天光〉 \(ISLAND'S SUNRISE\) 藝術公民計劃 太陽花運動歌曲](#)

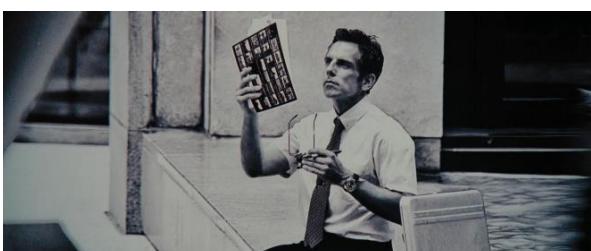
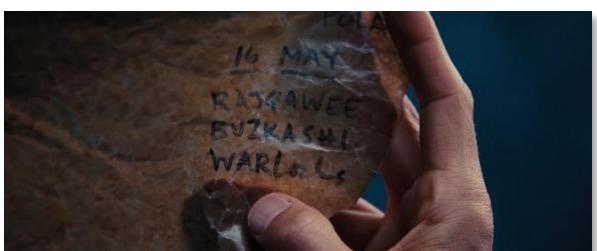
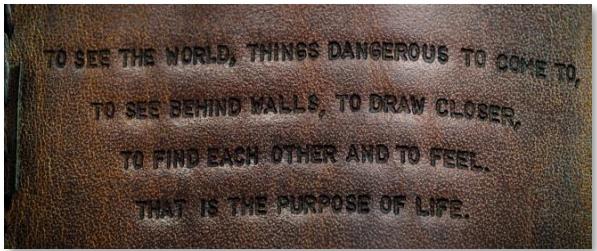
# 畢業之後呢？



THE SECRET LIFE OF WALTER  
**MITTY**

難道對於職涯沒有幫助嗎qq  
難得的周末我們先來看個影片輕鬆一下XD





# 那麼職涯規劃呢？

# 關於職涯規劃

看看畢業學長姊們去了哪裡



National Dong Hwa University Department of Sociology

國立東華大學社會學系

[首頁](#) [最新公告](#) [系所簡介](#) [課程規劃](#) [我們的老師](#) [我們的學生](#) [招生資訊](#) [法規及表單下載](#) [新生Q&A](#) [校友風雲榜](#)

[首頁](#) / [招生相關資訊](#) / [未來大學生參考資料](#) / [社會學系過來人經驗分享](#) / [線上同學會](#)

95級-警察-孫英哲 96級-農創-林坤泰 96級-傳播-鍾效京 96級-出版傳播-鍾宛君 96級-社工-張智凱 96級-都計建築-廖一上

96級-設計師-呂孟達 97級-教師-王韋憲 97級-資訊科技-林如珊 98級-非營利組織-顏若卉 98級-教師-黃中一 100級-社工-林思婷

101級-教師-張元鳳 101級-創業-陳貞穎 104級-教師-詹于頃 104級-公職-盧亭榕 104級-社工-鍾政榮 104級-鐵路交通-謝居璋

104級-農漁會-吳博軒 105級-創業-游承輝 105級-公職-賴文婷 105級-人資-林吉偉 105級-公職-曾亞媛 105級-人資-謝佩鈞

105級-斜槓-李英吉 105級-金融保險-王姿雅 106級-警察-古汶玉 107級-人資-蔡忻穎 107級-非營利組織-施芃年 108級-升學-林宜萱

108級-金融-蔡易玲 108級-教師-王慧仙 108級-公職-陳韋誌 108級-升學-蔡佳錚 109級-升學-許冠澤 109級-升學-張利琦

109級-升學-洪紹予 109級-僑生-李文豪 109級大學部經驗分享-黃聖文

Ref: [社會學系過來人經驗分享](#)

科系、職務與待遇：大數據時代下的社會學與勞動市場

Mail: TLYu0419@gmail.com

# 關於職涯規劃

看看畢業學長姊們去了哪裡



National Dong Hwa University Department of Sociology

國立東華大學社會學系

首頁 最新公告 系所簡介 課程規劃 我們的老師 我們的學生 招生資訊 法規及表單下載 新生Q&A 校友風雲榜

首頁 / 招生相關資訊 / 未來大學生參考資料 / 社

95級-警察-孫英哲	96級-農創-林坤泰	96級-
96級-設計師-呂孟達	97級-教師-王韋憲	97級-
101級-教師-張元鳳	101級-創業-陳貞穎	104級-
104級-農漁會-吳博軒	105級-創業-游承輝	105級-
105級-斜槓-李英吉	105級-金融保險-王姿雅	106級-
108級-金融-蔡易玲	108級-教師-王慧仙	108級-
109級-升學-洪紹予	109級-僑生-李文豪	109級-

林如珊 社會發展系97級 國立東華大學社會學研究所碩士  
現職：垣創科技/喬昱科技 專案經理

社會學是一門很廣泛的學科，就因為它的廣泛，所以一般人對於它的印象會覺得廣而不精，但是，就因為此學門含括的範圍極廣，法政經社均有所涉獵，讓我在修讀課程的期間，才更能去摸索探討出自己有興趣的方向去深究研讀它。

研究所期間，有感於自己家鄉-宜蘭-因為雪山隧道的開通，對於交通有了巨大的改變，交通的便利性帶來生活圈的變化，人潮車潮的湧入讓宜蘭許多地方有了明顯的成長，故在研究所就讀時就訂定了論文主題目標。也讓我朝著「都市社會學」這學門開始精進，運用「都市社會學」的核心理論加上「GIS地理資訊系統」為工具，搭配政府的工商業普查統計數據，讓我作出了宜蘭地區在雪隧開通前後五年的都市產業分布的變化。

在論文發表過關之後與辦理離校手續的這段期間，就開始了我的求職面試之旅，每週從花蓮北上台北面試，最後順利面試上台灣第一大GIS原廠資訊公司，也慶幸在研究所期間的學習，讓我帶論文順利錄取與所學相關的工作，更開啟了我從文組背景的學歷意外進入到資訊科技產業領域，也一直發展至今，雖然後來工作的轉職未繼續在GIS產業，但是也讓我意外的深耕了資訊科技產業，至今還是非常感謝系所的多元課綱，讓我更清楚的找到了自己有興趣的發展，並勇於走出自己的路！

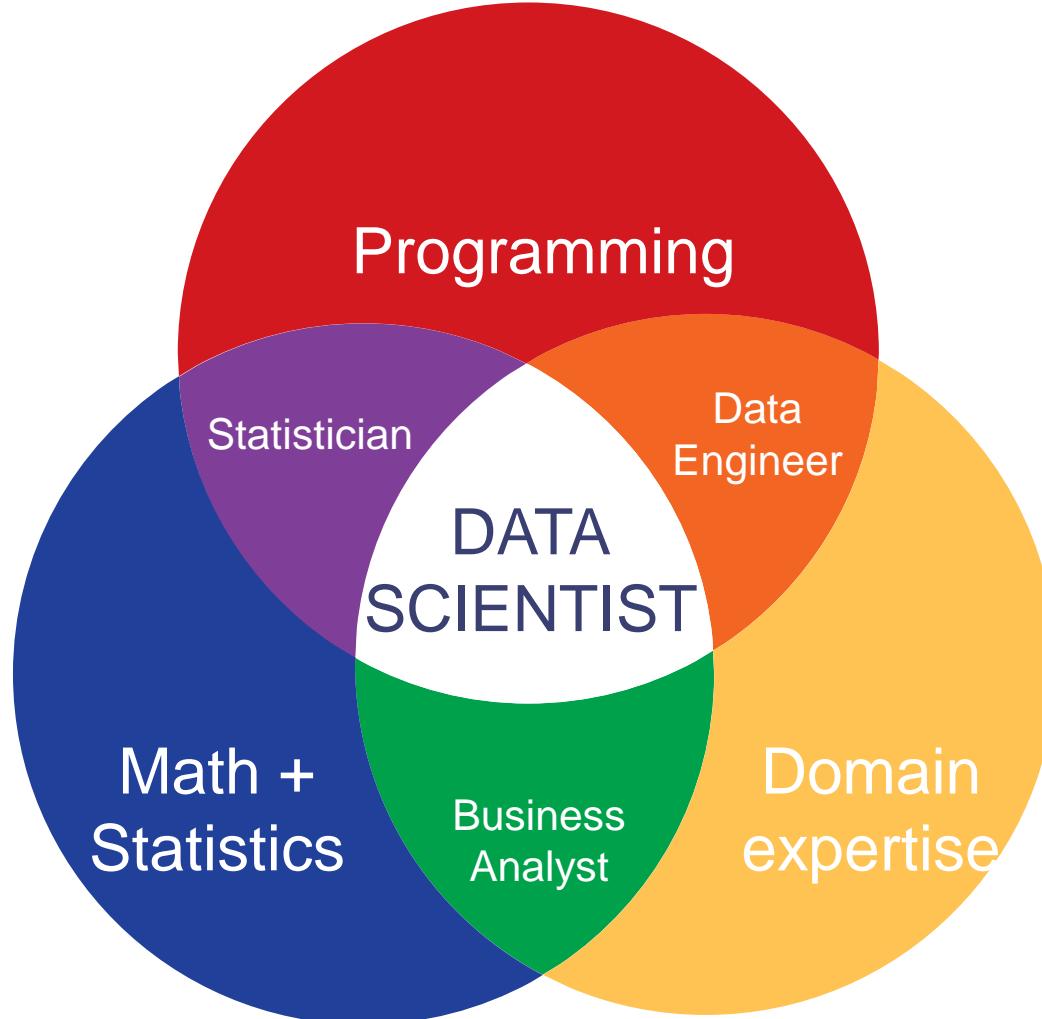
Ref: 社會學系過來人經驗分享

科系、職務與待遇：大數據時代下的社會學與勞動市場

Mail: TLYu0419@gmail.com

# 關於職涯規劃

## 資料科學領域技能與職務需求



### Math + Statistics

- Statistical and Mathematical knowledge (Linear Algebra, Calculus, Metrics and Probability).
- Learn some popular tools like R, Python, and/or Tableau.

### Domain expertise

- Good enterprise expertise helps you build better ML model tools and processes and infer the best business decisions.

### Programming

- Database: Hadoop, SQL and NoSQL
- Cloud Server Architecture: AWS
- toolkits: sklearn, Numpy, Pandas, Keras, etc.
- Algorithms: KNN, LR, NB, RF, GBM, etc.
- ...

Ref: 1. [Data Scientist、Data Analyst、Data Engineer 的区别是什么？](#)  
2. [真 · 資料團隊與分工](#)

# 關於職涯規劃

提供幾個找到自己興趣的幾個方法



- 性向測驗
- 多回想一些平常在做什麼事情的時候會忙到忘記時間？
- 參加業界的實習
- 來上學長今天的課程(笑)

# 我們來看看 QA 吧!

A man in a maroon hoodie and light-colored pants is longboarding down a winding asphalt road. He is leaning into the turn, with his front foot on the board and his back foot pushing off. The road is surrounded by lush green hills and mountains under a bright, slightly cloudy sky.

休息~

- 10分鐘後回來>” <

A man in a red shirt and white pants is running on a road through a green, hilly landscape under a bright sky.

# 數據應用案例

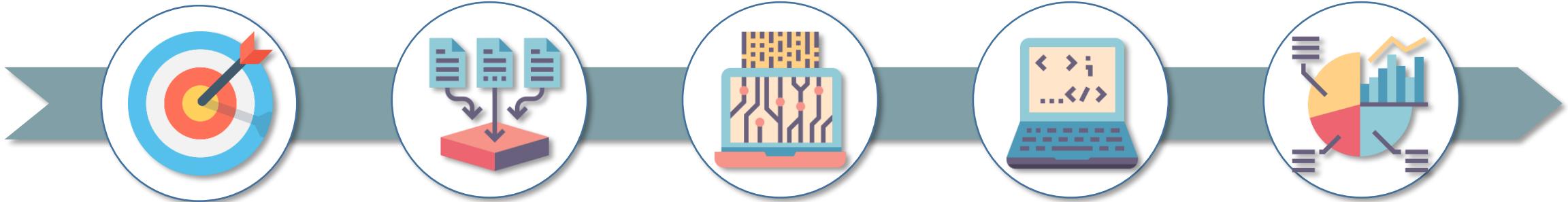
- 關於分析流程
- 數據分析的應用案例
- 大數據時代下的社會學有什麼挑戰

**阿不就分析，還有流程？  
要記得很清楚，因為面試工作時會問！**

# 記得量化課程曾做過的事嗎？

# 分析流程介紹

找工作面試的時候會問!  
要清楚掌握，不能只說分析



## 確認目標

- 問題重要嗎？
- 問題定義

## 資料蒐集

- 資料在哪裡?
- 資料品質
- 要怎麼取得資料?

## 預處理

- 遺漏值
- 離群值
- 標準化
- 特徵工程

## 模型建置 專案佈署

- 演算法選擇
  - 選定評估指標
  - 調參
  - 不平衡資料
- 撰寫分析報告
  - 專案上線

# 分析流程\_確認目標 把問題說清楚，講明白！

## 為什麼這個問題重要？

- 好玩  
預測生存遊戲 [PUBG](#) 誰可以活得久
- 企業核心問題  
如何對用戶更精準的投放廣告 [ADPC](#)
- 公眾利益 / 影響政策方向  
[停車方針](#)、[計程車載客優化](#)
- 對世界很有貢獻  
[肺炎偵測](#)

## 要回答什麼問題？

- 問題要能夠被轉化為具體的結果
- 根據目標變數型態會需要使用不同的衡量指標
- [PUBG](#)：生存時間、遊戲排名
- [ADPC](#)：廣告點擊率、商品成交率
- [肺炎偵測](#)：影像辨識的準確度
- ...

# 分析流程\_確認目標 把問題說清楚，講明白！

知乎 | 首发于 王喆的机器学习笔记

2019-05-05

宁肯  
一个清华大学的，人生价值在于研究让人更多地点击广告，这不是很可悲的一件事吗？  
1

12 条回复

王喆 (作者) 回复 宁肯  
2019-05-05  
清华大学一年只要出十个为国为民的大家就够了，其他人做一些养家糊口的事情挺好  
49

行者 回复 宁肯  
2019-05-05  
对于一个普通人来说，好好工作就是对国家和社会的最大贡献；而且和答主相比，我觉得大多数人都是造粪机器（比如我）

5

**深度学习中不得不学的Graph Embedding方法**

王喆 ✨  
数据挖掘等 3 个话题下的优秀回答者

1,033 人赞同了该文章

这里是「王喆的机器学习笔记」的第十四篇文章，之前已经有无数同学让我介绍一下 Graph Embedding，我想主要有两个原因：

- 一是因为 Graph Embedding 是推荐系统、计算广告领域最近非常流行的做法，是从等一路发展而来的 Embedding 技术的最新延伸；
- 二是因为已经有很多大厂将 Graph Embedding 应用于实践后取得了非常不错的线上效果。

Ref: [深度学习中不得不学的Graph Embedding方法](#)

科系、職務與待遇：大數據時代下的社會學與勞動市場

Mail: TLYu0419@gmail.com

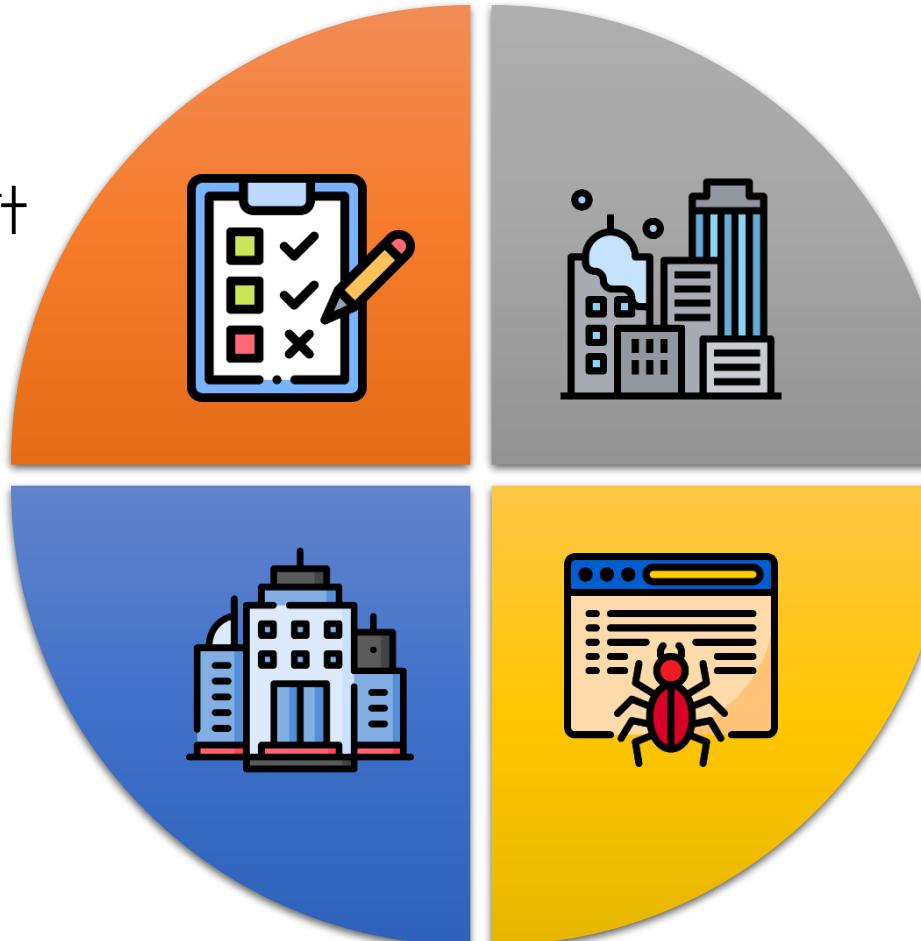
# 分析流程\_資料蒐集

## 資料在哪裡？要怎麼取得資料？

### 調查問卷

依專案需求，由領域專家設計與發放問卷的方式蒐集資料

- [台灣社會變遷基本調查](#)



### 政府公開資料

政府部門提供之公開資料

- [經濟地理資訊系統](#)
- [捷運各站進出量統計](#)

### 公司資料庫

依公司個別業務累積的客戶行為、消費、位置……等等資料

### 網路爬蟲

透過網路爬蟲程式擷取資料

- [GooglePlay](#)
- [松果購物](#)

# 分析流程\_資料蒐集

## 資料在哪裡？要怎麼取得資料？

The image shows two screenshots illustrating data collection methods:

- Google Play Store Screenshot:** Shows the main interface of the Google Play Store. A blue box highlights the search bar at the top. The sidebar on the left is expanded, showing categories like 娛樂 (Entertainment), 應用程式 (Applications), 電影 (Movies), 圖書 (Books), and 裝置 (Devices). A blue box highlights the "APP 名稱" (App Name) field in the search bar.
- Facebook App Page Screenshot:** Shows the app page for the Facebook app on the Google Play Store. A blue box highlights the "APP 名稱" (App Name) field in the search bar. The page displays the app icon, name, developer, rating (4.2 stars from 104,679,238 reviews), and a large green "安裝" (Install) button. A blue box highlights the "整體滿意度" (Overall Satisfaction) section, which includes a star rating chart and review counts. Another blue box highlights the "評論" (Reviews) section, showing a sample review by "jolin shen" dated November 4, 2020, with 30 likes.

**整體滿意度**

4.2  
★★★★★  
共 104,679,238 則評分

**客戶名稱、時間、個人滿意度、評論內容、被讚數**

jolin shen ★★★★★ 2020年11月4日  
從9月中以後，廣告幾乎是沒有效果，我問其他同樣管理粉絲專頁的人，大家都是同樣反應 請問為何繳了費用，為何廣告效益差那麼多？台灣沒有疫情，沒有封城，大家都正常上班上課 是FB系統在轉換或更新嗎？如果是如此，是否可以歸還廣告費用呢？

30

Ref: <https://play.google.com/store/apps>

# 資料蒐集 我想知道電商的價格與客戶的評論

卖家中心 | 下载 | 追蹤我們 [Facebook](#) [Instagram](#) [Line](#)

通知總覽 [帮助中心](#) [註冊](#) | [登入](#)

蝦皮購物

看更多11.11免運商品。

11.11

11.11 最強購物節

全面11折<sup>up</sup> 天天0免運 \$11<sup>up</sup>標豪宅

萬人瘋傳攻略搶先看

免運！蝦皮直送 超市\$299免運 領最強商城券 蝦皮電子

條件篩選

分類

- 電腦周邊配件 (5萬+)
- 電腦零組件 (2萬+)
- 鍵盤滑鼠 (2萬+)
- 居家生活 (2萬+)

更多 ▾

出貨地點

- 台灣
- 海外

運送方式

- 全家
- 7-11
- 黑貓宅急便
- 萊爾富

更多 ▾

品牌

筆記型電腦

通知總覽 [帮助中心](#) [註冊](#) | [登入](#)

商品名稱、價格、滿意度、銷售量

商品名稱	價格	滿意度	銷售量
15.6吋 筆電 win10 超速 戰勝 asus acer imac 打遊...	\$3,499 - \$8,999	★★★★★	已售出 147 新北市永和區
MITPC*acer宏碁 S3 391 UltraBook 超輕薄筆電 ...	\$5,500	★★★★★	已售出 473 台中市北屯區
14吋 全新筆電 win10 超速 快感 戰勝 asus 筆電 ipa...	\$2,999 - \$8,999	★★★★★	已售出 381 新北市永和區
15.6吋 win10 打lol 英雄聯盟 戰勝 asus 筆電 筆記...	\$3,499 - \$7,999	★★★★★	已售出 127 新北市永和區
ASUS 15.6吋 i5 4G/240G SSD K55A Win10二手手...	\$4,600 - \$5,100	★★★★★	新北市永和區
ASUS X205TA 珍珠白 WIN10	\$4,500	★★★★★	已售出 19 新北市中和區
ASUS 華碩 E410MA 【... 滿額贈】	\$10,900 - \$9,999	★★★★★	已售出 2 台北市中正區
清倉【ASUS】華碩 S410UA (i5-...)	\$13,500	★★★★★	已售出 1 彰化縣彰化市
14吋 全新筆電 win10 超速 快感 全新 中文 筆電 筆...	\$2,999 - \$7,999	★★★★★	已售出 186 新北市永和區
3C客製筆電 i7 i5 intel AMD獨顯 筆電 筆記型...	\$4,500 - \$13,000	★★★★★	已售出 141 桃園市八德區

Ref: <https://shopee.tw/>

# 資料蒐集 我想知道候選人的動態與網路聲量

The screenshot shows a Facebook profile for Tsai Ing-wen. On the left, there are two posts: one from 9 hours ago about movie theaters and another from yesterday morning about medical workers. The main post in the center is highlighted with a blue border and features a large image of Tsai speaking at a podium with microphones from various media outlets. Overlaid on the image is a block of text in yellow and white. Below the image is a call to action: "通關密語 1203 記得收看我的直播". To the right of the main post is a blue-bordered box containing the post's details (916 Comments, 267 Shares) and a list of three replies from Tsai Ing-wen, each with a yellow border. The replies discuss public opinion and political accountability.

因為疫情的影響，我們的生活有了一些改變，但常進行，一樣能走進電影院、舉辦頒獎典禮。前陣子，我參與了金馬影展 TGHFF 的社群活動，台灣的防疫日常。金馬獎，是每年台灣的電影盛事，相信喜愛電影在還在院線，而且有入圍今年金馬獎的台灣電影... 查看更多

林雨蒼和其他 6.1 萬人

蔡英文 Tsai Ing-wen 昨天上午 3:50  
台灣能擁有許多有專業、有熱忱的醫療工作者，今天，我在總統府接見了 #醫療奉獻獎 的得獎人

追蹤中

不論是民心還是民調，民主國家的人民，有自己自由的意志。民調好，更要全力衝刺；民調不好，就深刻檢討。

通關密語 1203 記得收看我的直播

蔡英文 @iing

蔡英文 Tsai Ing-wen 9 小時  
不論是民心還是民調，民主國家的人民，有自己自由的意志，不是政治人物可以下令要人民做什麼事情，政治人物只能拜託人民，為人民做事情。  
從大選開始到現在，有很多民調，不能說民調都是做出來的。民調好，我們更要全力衝刺，民調不好，我們就深刻檢討。

#我的通關密語1203  
#鎖定直播

17K 916 Comments 267 Shares

Like Comment Share

Most Relevant

Author 蔡英文 Tsai Ing-wen https://facebook.com/events/473640546595775/?fref=cl Like Reply 2h 141

Author 蔡英文 Tsai Ing-wen https://www.facebook.com/events/473640546595775/?fref=cl Like Reply 2h 101

呂玉產 之前一位中國人在問，中國已經是全球的第二經濟體了，為何台灣人不願意接受一國兩治，我的回答是，中華民國已經是主權獨立的國家，而且永遠不會變。北

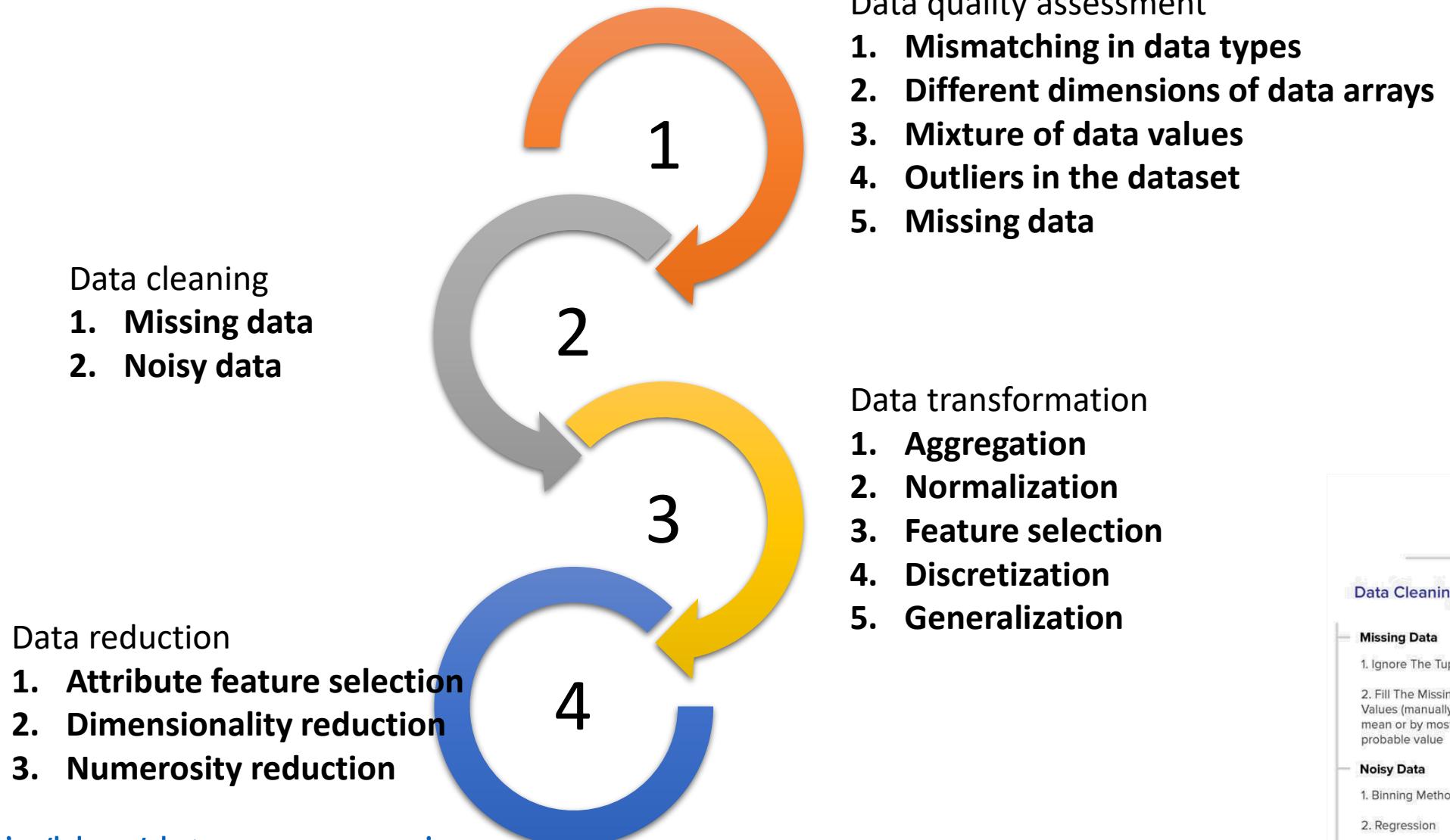
Write a comment...

貼文人、  
貼文時間、  
貼文內容、  
按讚數、  
留言、分享數

留言人、  
留言時間、  
留言內容、  
按讚數、  
回覆內容

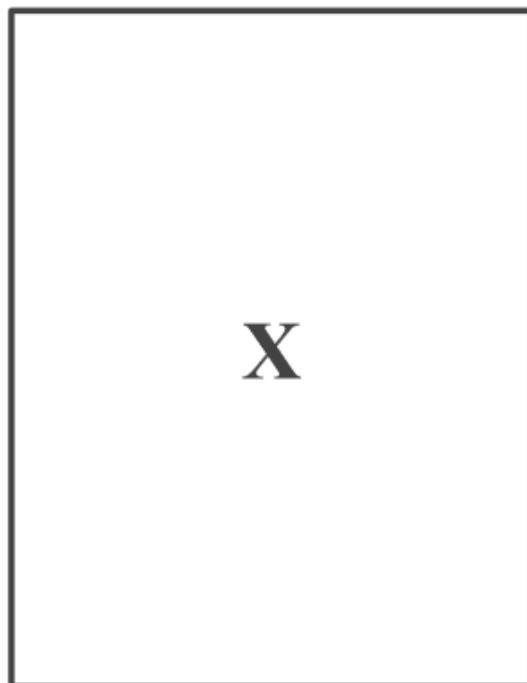
Ref: <https://www.facebook.com/tsaiingwen>

# 資料預處理



Ref: <https://serokell.io/blog/data-preprocessing>

## *Supervised Learning*

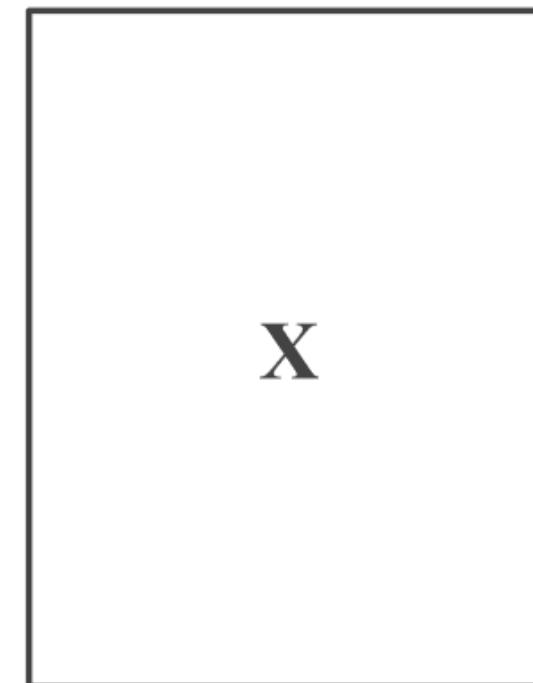


Input variables or  
features



Response  
variable

## *Unsupervised Learning*



Input variables or  
features



Unobserved  
variable

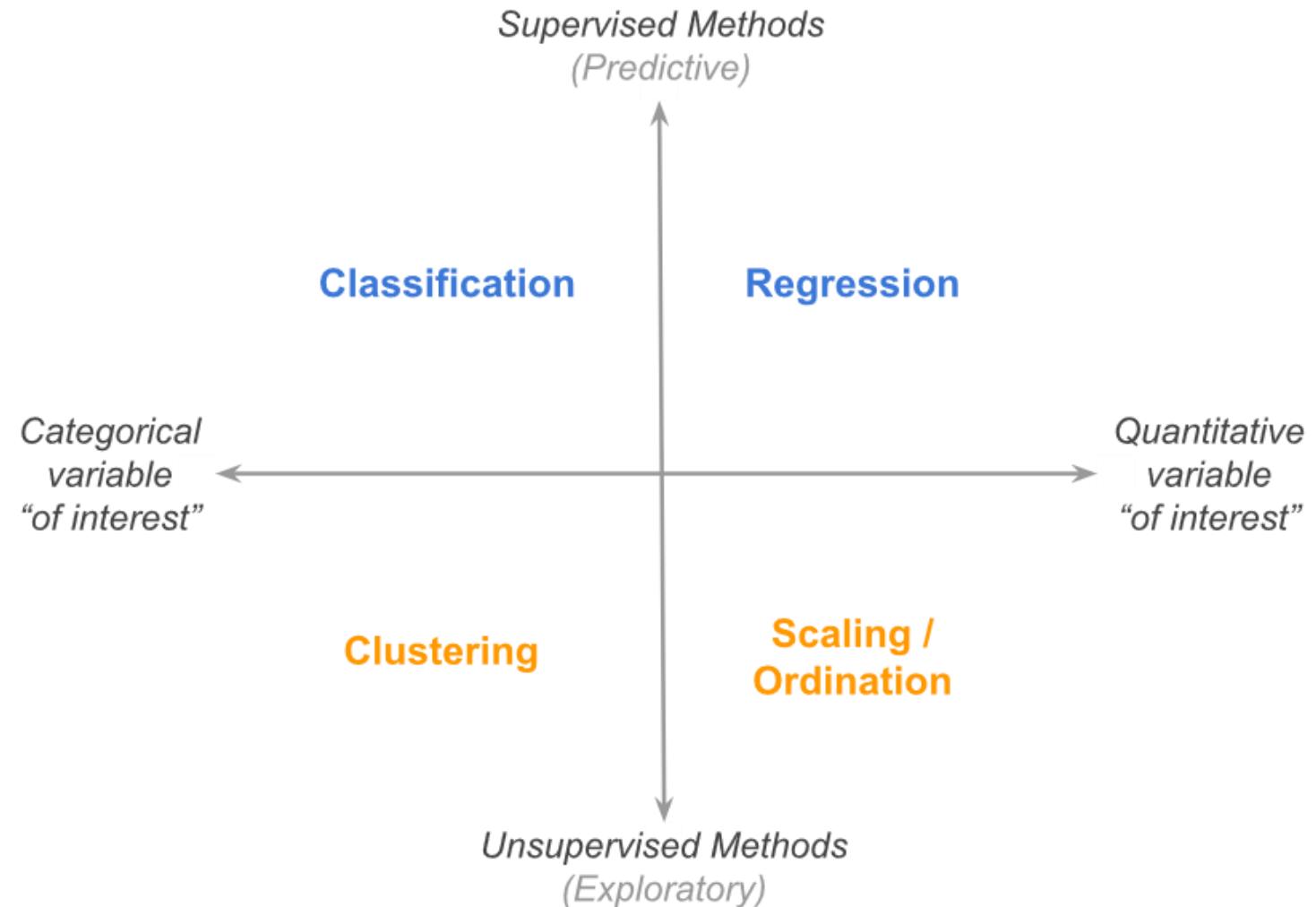
Ref: <https://allmodelsarewrong.github.io>

# 資料預處理

## 什麼是資料尺度

類別	數學特徵	範例
類別	=, ≠	居住縣市，科系名稱...
順序	=, ≠, >, <	滿意度，消費頻率...
等距	=, ≠, >, <, +, -	溫度，年份...
等比	=, ≠, >, <, +, -, *, /	營業額，來客數...

# 模型建置



# 分類模型案例

- (預測)分類：想知道100萬名客戶中哪些人會購買商品
- 為什麼要建模預測？單維度的標籤往往很難精準的預測
- 學期成績 過/不過 (老師的角度)

# 回歸模型案例

- (預測)回歸：想知道100萬名客戶中每位客戶各自會買多少錢
- 想知道客服中心接下來5天各自有多少通電話進線
- 為什麼要這樣預測？
- 學期成績會拿到多少分 (老師的角度)

# 降維模型案例

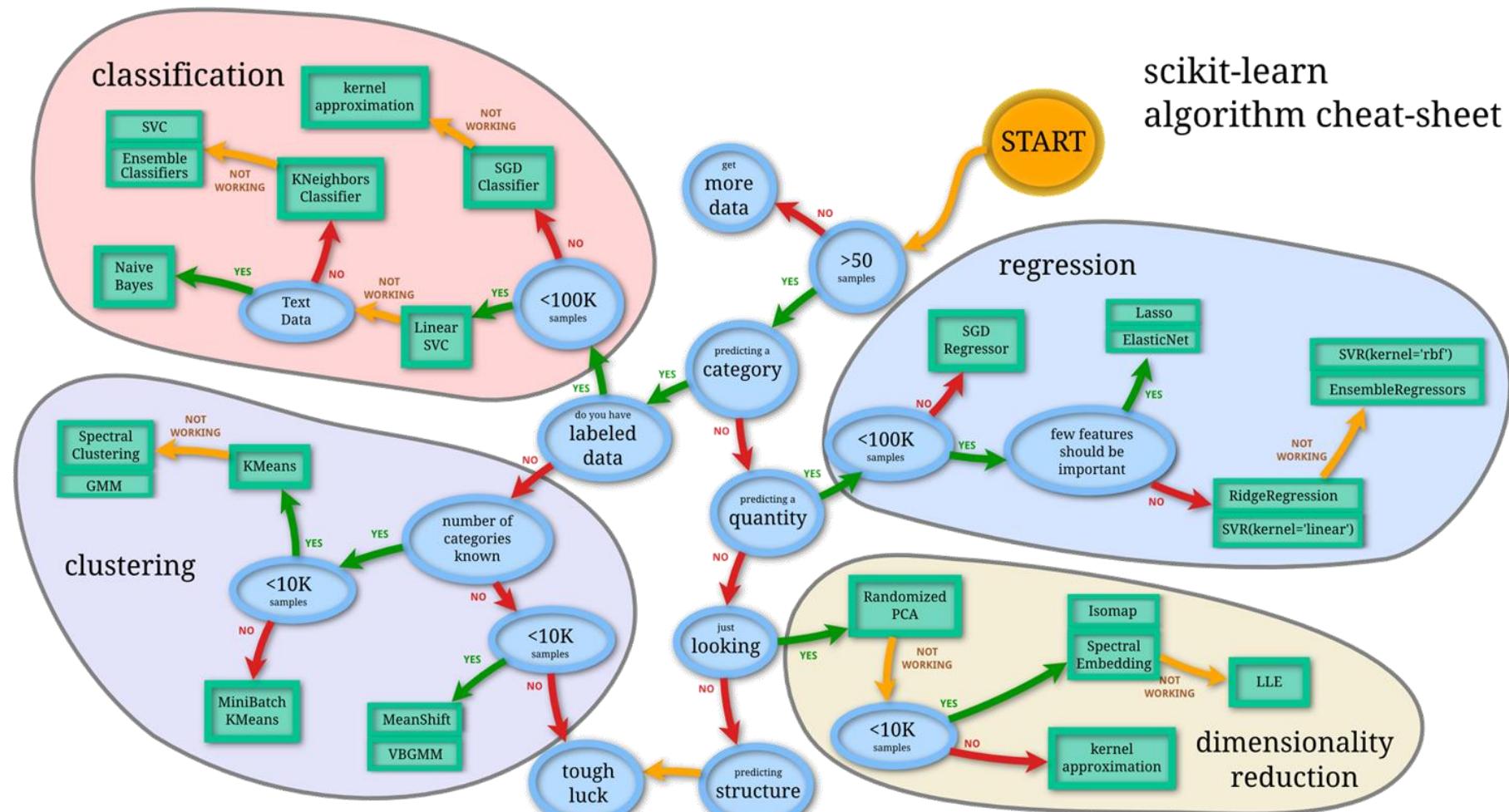
- (敘述)降維：想要萃取潛在結構???
- 通常用來做視覺化
- 系上的課程有很多，要精簡成績的構面(歷史、地理、公民)變成社會
- 物理化學變成自然

# 分群模型案例

- (敘述)分群：想知道客戶長什麼樣子???
- 班上學生的成績可以怎麼樣分群
- 我想要更直接一點，直接跟我說可以分成幾群，各群長什麼樣子

- - 回歸問題 (預測值為實數)
  - - RMSE, Root Mean Square Error
  - - Mean Absolute Error
  - - R-Square
- - 分類問題 (預測值為類別)
  - - Accuracy
  - - F1-score
  - - AUC, Area Under Curve

# 如何選擇合適的演算法

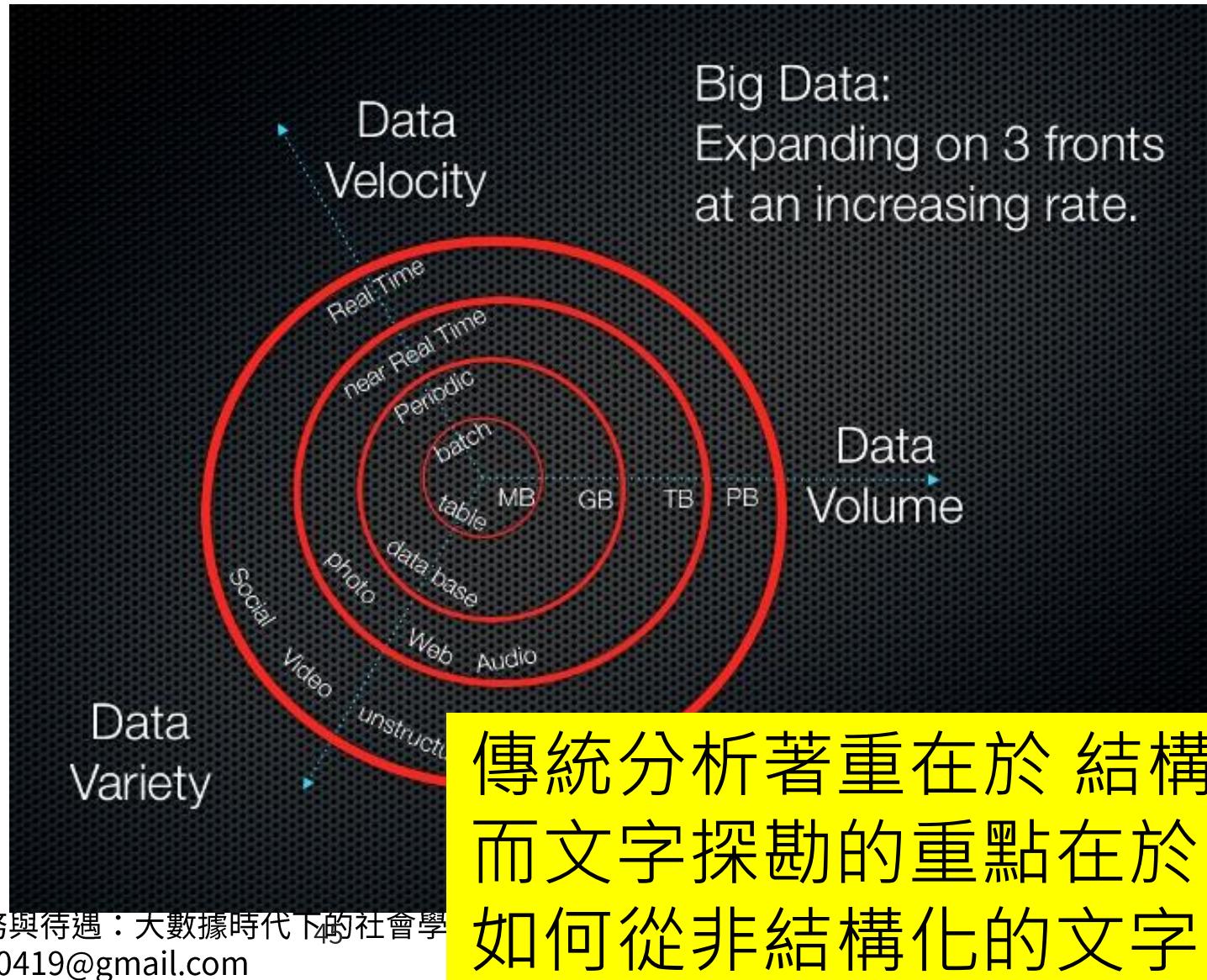


- 檢視模型實施後是否能有良好的預測、解釋效果。並且也可以在這個階段找出進一步優化模型的客能
- 設定評估指標
- 如果是分析案，會需要用到視覺化的技巧，協助長官/讀者理解
- 如果是模型案，會需要設計追蹤的機制定期檢視成效

學長碎碎唸：- 要有獨立完成分析專案的能力，但你不能只會分析

# 大數據給社會學什麼挑戰？

- 大數據的3個 V



✓ 資料量

✓ 形態

✓ 即時性

- ref :

社會學視角下的大數據方法論及其困境

邱泽奇：大数据给社会学带来了什么挑战？

# 反面案例

- [https://www.informationsecurity.com.tw/article/article\\_detail.aspx?tv=&aid=8311&pages=1](https://www.informationsecurity.com.tw/article/article_detail.aspx?tv=&aid=8311&pages=1)

## 案例2

- Youtube推薦機制
- 狠準，但也造成了沉迷
- (教大家一招，如果男朋友沉迷在這上面可以偷偷幫她點不要推薦我這個頻道XD)

## 案例3

- 了解規則之後被玩弄規則
- 網軍、買讚

- 應用的反例：假新聞

- 買讚(還真的有人會買XD)

- 要強調不是說技術很可怕，所以我們要排斥他
- 反而我們也得要學會這些技術，雖然技術能力可能還是比不上資訊背景的人，但至少要具備基本的工程、分析能力!

# 到底大數據是好還是壞> “<

# 總結為什麼變成反面

- 用的方式不對
- 沒有完美個模型，永遠都有誤差，但被錯誤預測的人是無辜的
- 預測會有準確度的問題(什麼是型一錯誤，型二錯誤)

# 什麼是型一型二錯誤

- 什麼叫做預測錯誤
  - 對於回歸模型叫做殘差
  - 分類模型有型1型2錯誤
- 
- 颱風天，預測會不會停課
  - 預測成功，

# 休息~

- 10分鐘後回來>” <
- 有問題可以在Slido上發問&幫別人點讚



# 我們來看看 QA 吧!

A man in a maroon hoodie and light-colored pants is longboarding down a winding asphalt road. He is leaning into the turn, looking back over his shoulder. The road is surrounded by lush green hills and mountains under a bright, slightly cloudy sky.

# 傑華老師的煩惱

- aaaa

就在上週  
校長把傑華老師找進辦公室…

# 資料蒐集 我想知道勞動市場的職缺資訊

人力銀行 My104 家教 外包 找才 教育 更多 好企業哪裡找？快看企業品牌大賞

python 地區 職務類別 搜尋

更新日期 上班時段 薪資待遇 經歷要求 公司相關 更多條件 排除條件 清空條件 檢視或訂閱條件

全部(3961) 全職(3716) 兼職(159) 高階(4) 其他 第 1 / 150 頁 符合度排序 強力放送

坐15年冷板凳！12強中美大職先發投手吳昇峰的故事

資料工程師 (Data Engineer)  
和泰汽車股份有限公司  
台北市中山區 | 2年以上 | 大學  
1. 負責資料分析流程執行與優化，參與企業內部數據應用、顧客分析等實際專案執行。  
2. 負責維護ETL系統與資料處理、運算等作業，能協助開發/維護資料分析介面。 \*錄取後需至TOYOTA經銷商據點見學二個月  
月薪 34,000~69,000元 上市上櫃 員工550人  
儲存 電郵

AI架構師 / Algorithm Engineer / Data Engineer / Data Scientist  
台泥企業團\_臺泥資訊股份有限公司  
台北市中山區 | 經歷不拘 | 碩士  
<AI大數據工程師> 1.進行跨部門溝通，了解需求、分析議題並規劃Data Collection Schema 2.執行Data Exploration, Pre-stage visualization, Data Reduction, Feature Reduction與Machine Learning. 3.模組化與優化Data Collection /Cleaning Stage and 待遇面議 員工90人  
儲存 電郵

11/21 Python Software Engineer  
hyve solutions\_海峰電腦股份有限公司 | 電腦系統整合服務業  
桃園市龜山區 | 2年以上 | 大學  
Hyve Solutions is looking for a talented junior/mid-level Python software engineer to help write Python programs. You are  
待遇面議 外商公司 員工1003人  
儲存 電郵

11/28 Python Developer  
訊真科技股份有限公司 | 電腦軟體服務業  
台北市大安區 | 2年以上 | 大學  
科系、職務與待遇，大數據時代下的社會學與勞動市場  
Mail: TLYu0419@gmail.com

## • 人力銀行

## • 資料內容

- ✓ 公司名稱
- ✓ 工作地區
- ✓ 所在地
- ✓ 職務內容
- ✓ 學歷要求
- ✓ 薪資
- ✓ ...

# 我不想做資料科學的工作 學這個對我會有幫助嗎？

- 會阿，你可以找到高薪的工作  
(有人覺得這個不重要嗎？然後說不一定啦，因為每天3萬其實還滿有吸引力的XD)
- 而且找到學會這個還可以幫你找到高薪的工作
- 跟學弟妹互動，你覺得這個重要嗎?
  - 找到說重要的學弟妹，
  - 然後我反問妳社會系的，  
你覺得賺錢跟對社會有幫助哪個重要？
  - 然後問為什麼？
  - 有些說重要

接到傑華老師任務的你們，開始回想量化研究、社會學期末報告的分析流程...(請學生分享) ,

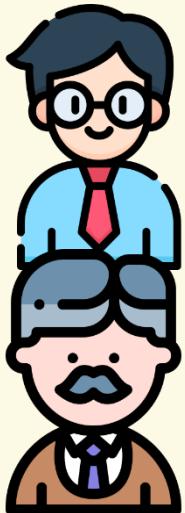


想要「討論」一下關於社會系學生的就業狀況(實際上是想要請傑華老師寫報告)，討論後希望傑華老師能幫忙了解以下幾個問題，並約傑華老師 1個月後繳交書面資料

市場上哪些產業/工作會想要找社會學系的學生？這些工作的工作內容是什麼？

不同科系職缺的就業狀況有哪些差異？社會學的優勢是什麼？

這幾年大數據很熱門，也有很多公司在做數位轉型，社會學系的學生在這個方面有哪些挑戰與機會？



...。(怎麼又把要給教育部的報告丟給我做，而且也不早點  
跟我說，只給我 1個月怎麼來的及...  
)



有問題嗎？

好的，沒問題

這邊用校長和老師的人頭來呈現對話

校長：傑華阿，你們社會系一直是以量化研究分析聞名的

傑華：這是系上師生一起努力的成果！(糟糕，該不會是有)

校長：你們去年是不是還有請一位畢業的校友回來分享大

傑華：對的，當時是請他回來分享網路爬蟲和文字探勘的

可以透過Python快速的爬取像是電商、社群、AP

校長：太好了，我這邊剛好有一份工作想要請你幫忙，

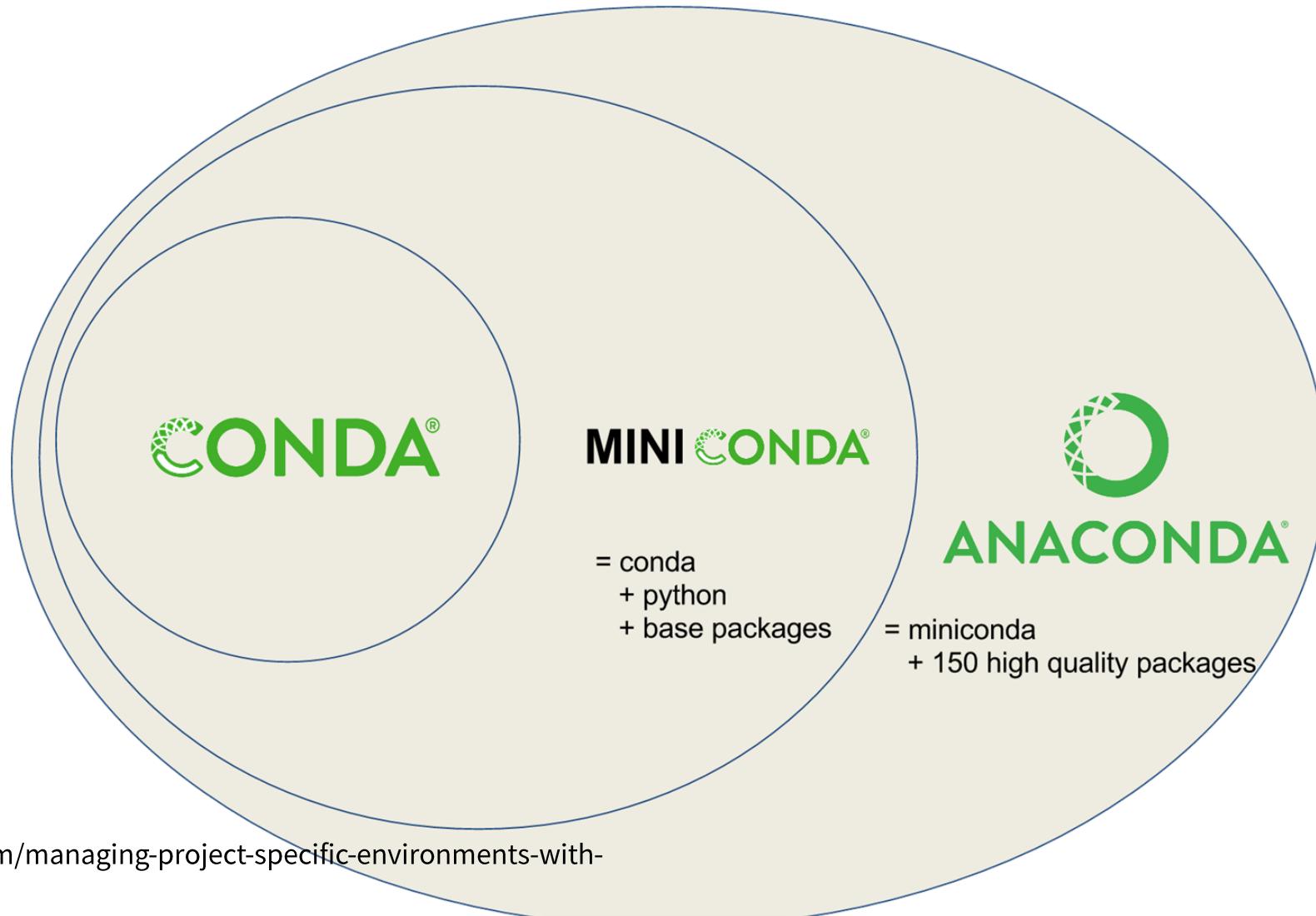
我這邊有接到教育部的報告，想請你幫我抓人力

傑華：...

在煩惱中傑華老師靈光一閃想到了今天來參加課程的你們，在課程中的你們學會了透過 Python 爬取資料與資料分析

# 來實做一下網路爬蟲

# 建立分析環境

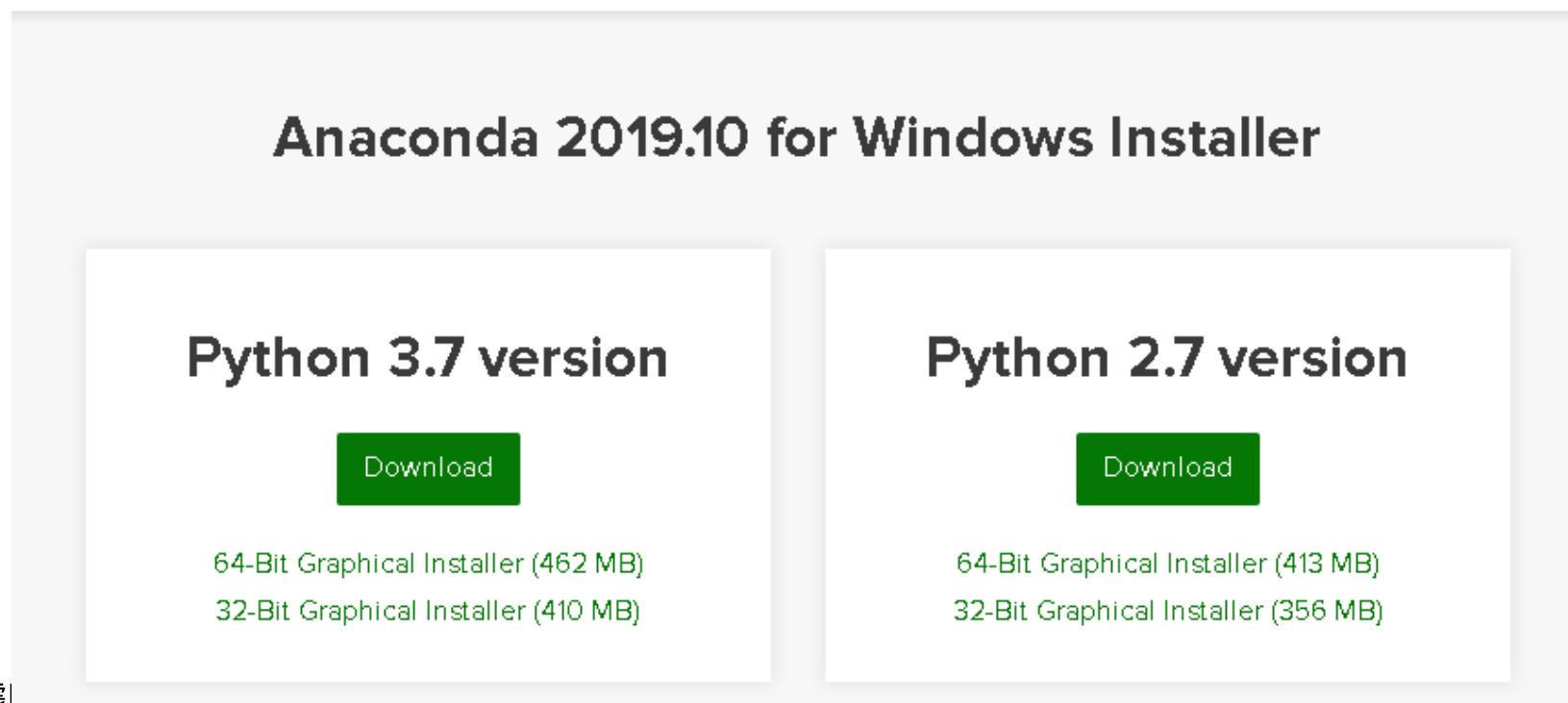


<https://towardsdatascience.com/managing-project-specific-environments-with-conda-b8b50aa8be0e>

# 環境設定

## Python 安裝

- <https://www.anaconda.com/distribution/>



# 啟動jupyter

# 人力銀行網路爬蟲實作

帶他們看一下網站結構  
練習爬外層的資料  
反爬蟲機制介紹  
加速

- 學長去年有來講怎麼爬蟲，但有點忘記該怎麼做了，我們來回想一下爬蟲要怎麼做...(時間回到1年前...)
- 可以講headers要怎麼設定，如果沒有放的話會怎樣
- 104有限制資料筆數的問題，所以要組合地區和職務來細分職缺
  - 地區代碼的表格是哪裡找到的
  - 職務代碼的表格怎麼找
  - json怎麼轉成DataFrame
- 怎麼寫發現規則 & 寫爬蟲程式
- [HumanResource\\_104\\_v2.ipynb](#)
- 學長碎碎唸：如果想要成為優秀的分析師，還是需要具備基礎的資料工程能力

# 安裝套件

- Conda install xxxx

費盡千辛萬苦，總算把職缺都  
抓下來了！  
我們可以看到抓下來的結果長  
什麼樣子…

# 我們來看看 QA 吧!

A man in a maroon hoodie and light-colored pants is longboarding down a winding asphalt road. He is leaning into the turn, looking back over his shoulder. The road is surrounded by lush green hills and mountains under a bright, slightly cloudy sky.

# 傑華的煩惱II



傑華：報告校長，上次你交代的任務已經完成囉，我們花費了千辛萬苦，總算把職缺都透過大數據網路爬蟲的技術抓下來了(努力邀功一下)

校長：非常好!!既然你都抓好資料了，那我想要再多麻煩你一下，請你幫我作一下分析

傑華：OK

# 再回顧一下分析流程

# 資料清理

- 方法比較

	統計分析	大數據
目的	驗證假設	發現知識
方法	檢定因果關係	找出相關性
資料量	要多少樣本才有足夠的代表性	大量的資料與更大量的資料
成本	高	低
效率	低	高
類型	批次資料	即時資料

- 方法比較

	統計分析
目的	驗證假設
方法	檢定因果關係
資料量	要多少樣本才有足夠的代表性
成本	高
效率	低
類型	批次資料



# • 資料結構

The screenshot shows the IBM SPSS Statistics Data Editor interface. A table is displayed with columns: id, zip, stratum2, qtype, year, year\_m, b\_dtime, b\_dtm, and b\_. The 'year' column is highlighted in yellow. The data consists of 735 rows, with the first few rows shown below:

	id	zip	stratum2	qtype	year	year_m	b_dtime	b_dtm	b_
1	100102	台北市中正...	都會核心	卷一	2007	96	7221030	7	
2	100106	台北市中正...	都會核心	卷一	2007	96	7161933	7	
3	100111	台北市中正...	都會核心	卷一	2007	96	8012100	8	
4	100112	台北市中正...	都會核心	卷一	2007	96	8011905	8	
5	100116	台北市中正...	都會核心	卷一	2007	96	8011617	8	
6	100120	台北市中正...	都會核心	卷一	2007	96	7161817	7	
7	100121	台北市中正...	都會核心	卷一	2007	96	7211627	7	
8	100125	台北市中正...	都會核心	卷一	2007	96	7242046	7	
9	100127	台北市中正...	都會核心	卷一	2007	96	7242135	7	
10	100131	台北市中正...	都會核心	卷一	2007	96	7222045	7	
11	100133	台北市中正...	都會核心	卷一	2007	96	7281027	7	
12	100136	台北市中正...	都會核心	卷一	2007	96	7242215	7	
13	100145	台北市中正...	都會核心	卷一	2007	96	7222145	7	
14	100151	台北市中正...	都會核心	卷一	2007	96	7182119	7	
15	100152	台北市中正...	都會核心	卷一	2007	96	7281135	7	
16	100156	台北市中正...	都會核心	卷一	2007	96	7221135	7	
17	100161	台北市中正...	都會核心	卷一	2007	96	8041725	8	
18	100162	台北市中正...	都會核心	卷一	2007	96	7170945	7	
19	100167	台北市中正...	都會核心	卷一	2007	96	8011807	8	
20	100302	台北市中正...	都會核心	卷一	2007	96	7231700	7	
21	100310	台北市中正...	都會核心	卷一	2007	96	7231001	7	
22	100312	台北市中正...	都會核心	卷一	2007	96	7310915	7	
23	100316	台北市中正...	都會核心	卷一	2007	96	7201945	7	
24	100317	台北市中正...	都會核心	卷一	2007	96	7261750	7	
25	100319	台北市中正...	都會核心	卷一	2007	96	7190915	7	

結構化資料

The screenshot shows a JSON file named 'records.json'. It contains three entries, each representing a record with a track ID, reporting date, longitude, and latitude.

```

{
    "trackid": "AA-1234",
    "reported_dt": "12/31/2019 23:59",
    "longitude": "-111.12500000",
    "latitude": "33.37500000"
},
{
    "trackid": "BB-7890",
    "reported_dt": "12/31/2019 23:59",
    "longitude": "-113.67500000",
    "latitude": "35.87500000"
},
{
    "trackid": "CC-4545",
    "reported_dt": "12/31/2019 23:59",
    "longitude": "-115.57500000",
    "latitude": "37.67500000"
}

```

半結構化資料

The screenshot shows the DevTools of a browser (specifically Google Chrome) displaying the source code of a webpage. The page content is heavily obfuscated with long strings of characters, indicating unstructured or semi-structured data.

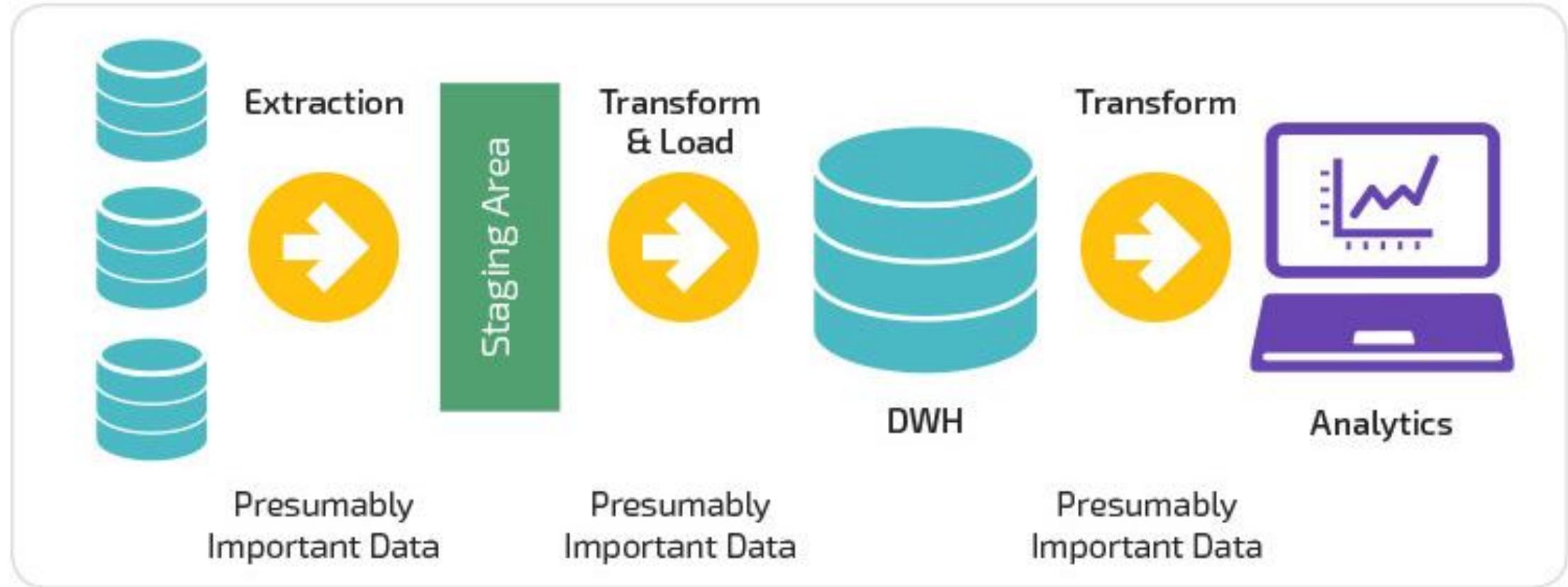
```

<body>
  <div id="root">
    <div class="LoadingBar"></div>
    <div></div>
    <main role="main" class="App-main">
      <div>
        <div class="SearchTabs SearchTab-bottomShadow" data-za-detail-view-path-module="SearchSwitchTabBar" data-za-extra-module="{"card":{"content":null}}"></div>
        <div class="SearchContainer">
          <div class="SearchMain" id="SearchMain">
            <div class="ListShortcut">
              <div class="List">
                <div class="data-za-detail-view-path-module="SearchResultList" data-za-extra-module="{"card":{"content":null}}">
                  <div class="Card SearchResult-Card" data-za-detail-view-path-index="0" data-za-extra-module="{"card":{"content": {"type": "Answer", "token": "91300972"}}, "attached_info_bytes": "Op0BCgtwbGfjZWhvbGRlchIgzt4ND1jMmQ0Y2zNU1NDzMHzUxYnJ1njc4YWZkNGEYBCIJHTE: HjUyKgk1Nz2zDcxTUyCDI0TgyNzgxOgkyOTEzNzY5HzdkF+eiseioiC1pkFn1bjm5og5beu! ABYAWABmAFAoAEBAqAEAsAEBuAEAOHc717CbdgBEuABBIAAbgCAA=="}}></div>
                  <div class="Card SearchResult-Card" data-za-detail-view-path-index="1" data-za-extra-module="{"card":{"content": {"type": "Answer", "token": "479307195"}}, "attached_info_bytes": "Op0BCgtwbGfjZWhvbGRlchIgzt4ND1jMmQ0Y2zNU1NDzMHzUxYnJ1njc4YWZkNGEYBCIJHTE: HjUyKgk1Nz2zDcxTUyCDI0TgyNzgxOgkyOTEzNzY5HzdkF+eiseioiC1pkFn1bjm5og5beu! ABYAWABmAFAoAEBAqAEAsAEBuAEAOHgloYXibdgB0JABIAAbgCAA=="}}></div>
                  <div class="Card SearchResult-Card" data-za-detail-view-path-index="2" data-za-extra-module="{"card":{"content": {"type": "Answer", "token": "576473399"}}, "attached_info_bytes": "Op0BCgtwbGfjZWhvbGRlchIgzt4ND1jMmQ0Y2zNU1NDzMHzUxYnJ1njc4YWZkNGEYBCIJHTE: HTE0Kgk1Nz2zDtz0TkycDNwNOTczNzxQgkzhDk1Q1NzZKF+eiseioiC1pkFn1bjm5og5beu! ABYAWABmAFAoAEBAqAEAsAEBuAEAOHgloYXibdgB0JABIAAbgCAA=="}}></div>
                  <div class="Card SearchResult-Card" data-za-detail-view-path-index="3" data-za-extra-module="{"card":{"content": {"type": "Answer", "token": "10A1660133"}}, "attached_info_bytes": "Op0BCgtwbGfjZWhvbGRlchIgzt4ND1jMmQ0Y2zNU1NDzMHzUxYnJ1njc4YWZkNGEYBCIJHTE: HTE0Kgk1Nz2zDtz0TkycDNwNOTczNzxQgkzhDk1Q1NzZKF+eiseioiC1pkFn1bjm5og5beu! ABYAWABmAFAoAEBAqAEAsAEBuAEAOHgloYXibdgB0JABIAAbgCAA=="}}></div>
                </div>
              </div>
            </div>
          </div>
        </div>
      </div>
    </main>
  </div>
</body>

```

非結構化資料

- ETL



網頁資料大多是非結構化的資料  
在分析前需要先進行資料處理才能進行分析

- 分析方式比較
- SPSS點選式，不用寫語法(適合一次性的分析，如學術研究)
- SPSS再處理資料很不方便
- Python/R 撰寫語法(企業需要定期執行程式)
  
- 社會學的議題很容易太大，要改變
- 對於學生的職涯規劃要有更積極的幫助

- 不是每個老師都喜歡計量研究，做報告前請先調查好老師的口味
- 學一下做簡報的技巧與設計的美感、配色
- 不要小瞧自己，但也不要委屈自己

- 發現問題，不要再慢慢研究了，要快快研究，採取行動，如果改變沒有達到成效，繼續想辦法改善！
- 鼓勵大家從小地方開始著手
- 透過怎麼樣的方式蒐集資料、做了什麼分析、帶來哪些改變
- 我覺得再小都是很有價值的事情！

- 所以社會學背景如果要走資料分析的領域要怎麼走?
- 建議往商業分析師的方向發展，技術能力也許沒這麼強，但一定要加強業務知識
- 如果有一些不服輸的精神，可以跟我一樣往資料科學家發展> “<[https://blog.v123582.tw/2020/10/18/%E7%9C%9F%E3%83%BB%E8%B3%87%E6%96%99%E5%9C%98%E9%9A%8A%E8%88%87%E5%88%86%E5%B7%A5/?utm\\_source=Facebook\\_PicSee&fbclid=IwAR22NWGUjrv0MTzBOfsIGiddawZylnTWSFBRc\\_xhfsiVtlUFK7G\\_TyD8bJs](https://blog.v123582.tw/2020/10/18/%E7%9C%9F%E3%83%BB%E8%B3%87%E6%96%99%E5%9C%98%E9%9A%8A%E8%88%87%E5%88%86%E5%B7%A5/?utm_source=Facebook_PicSee&fbclid=IwAR22NWGUjrv0MTzBOfsIGiddawZylnTWSFBRc_xhfsiVtlUFK7G_TyD8bJs)

# 學習資源

- Facebook社團
- Udemy