朝向資料科學家之路

進入職場1.5 y之心得

蔡岳霖 2018/05/14

Contact: tan800630@gmail.com

自我介紹

蔡岳霖

學歷

- 成功大學心理學系(2009-2013)
- 成功大學心理學研究所(2013-2015)

現職

- 亞洲大學大數據研究中心　研究助理 (2017/01-至今)

踏入 資料分析 / 資料科學 的原因

- 特質

熱愛學習/嘗試

與數字/電腦相處融洽

- 技能

心理學研究法

類神經網路

統計學

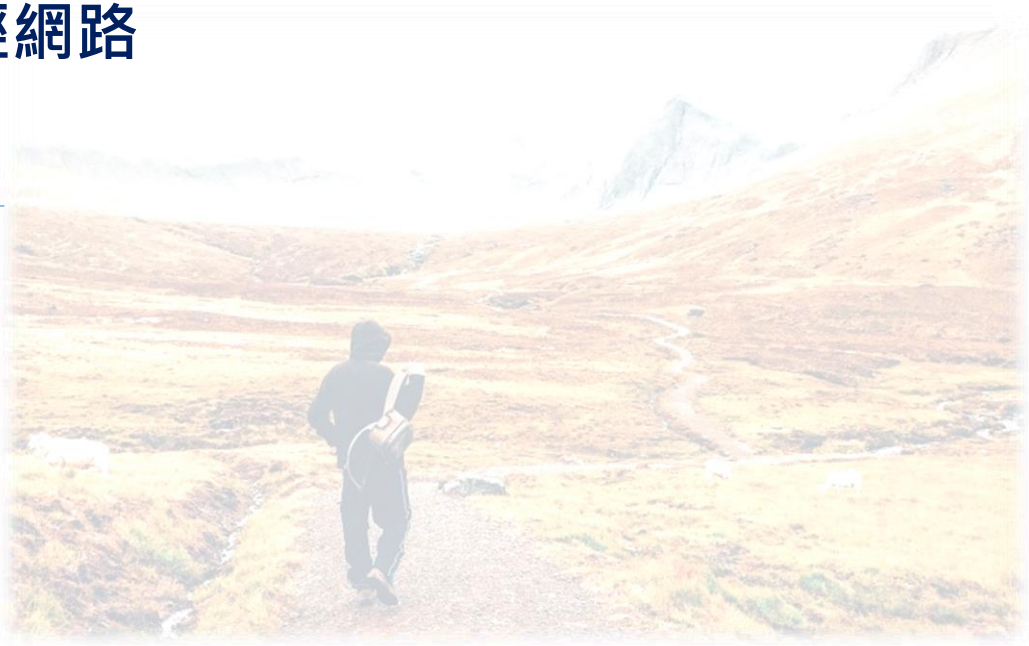
R語言

- 趨勢

資料科學

大數據

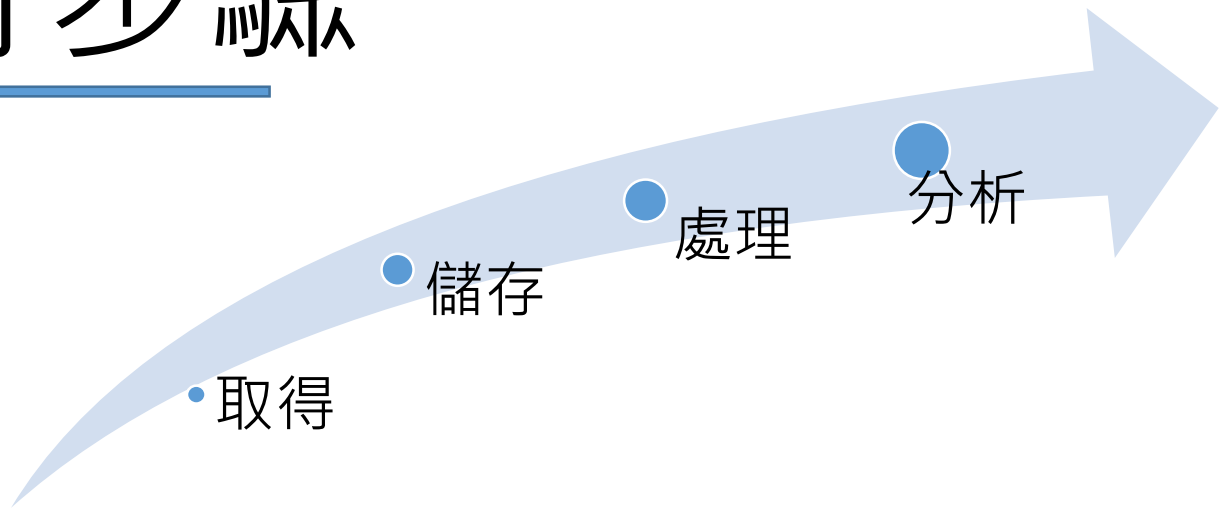
機器學習



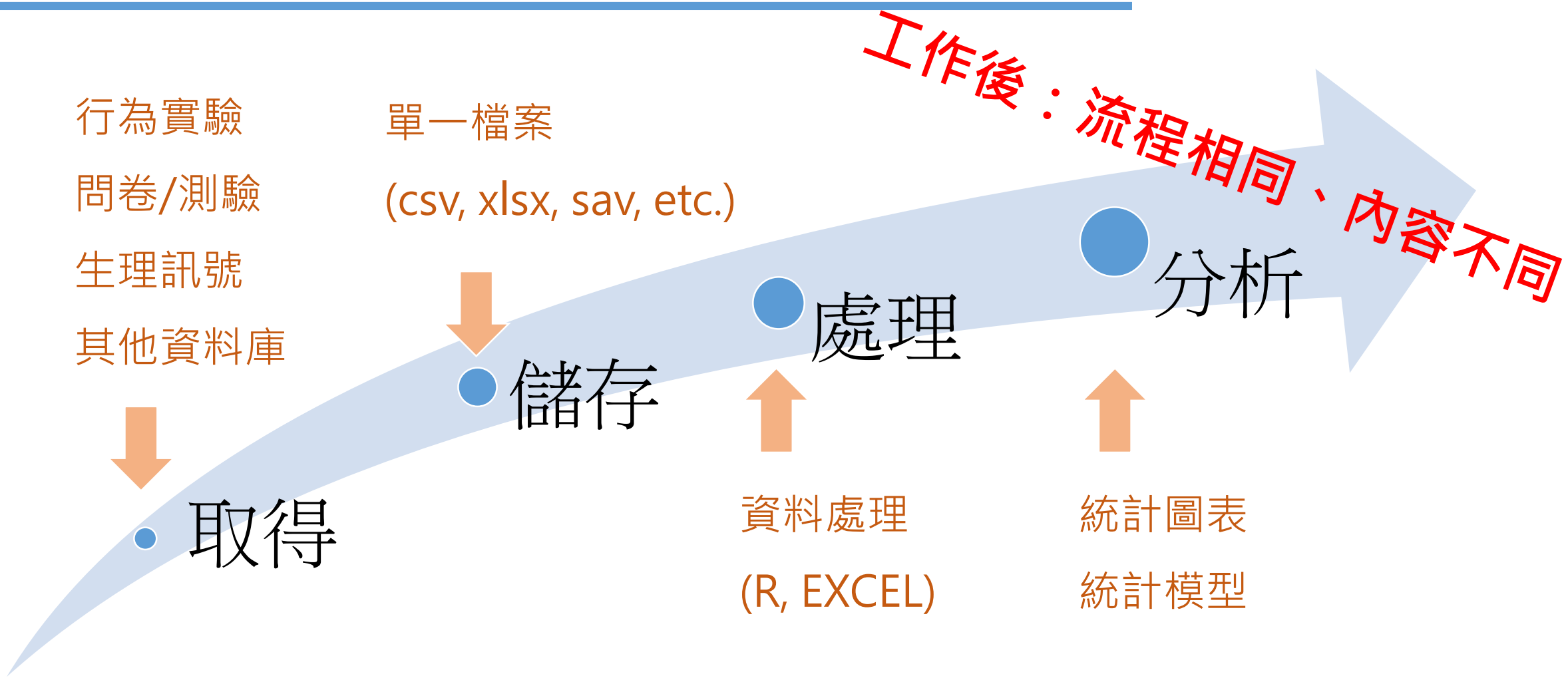
注意

以下內容高度包含主觀經驗與個人意見

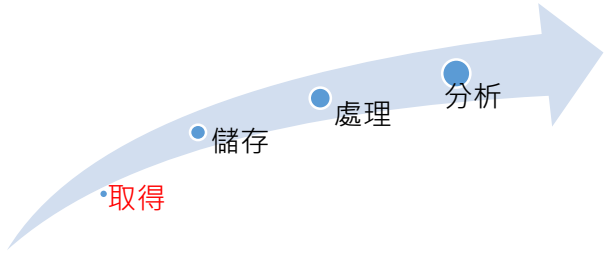
量化研究的執行步驟



量化研究的執行步驟-學生時期



資料取得



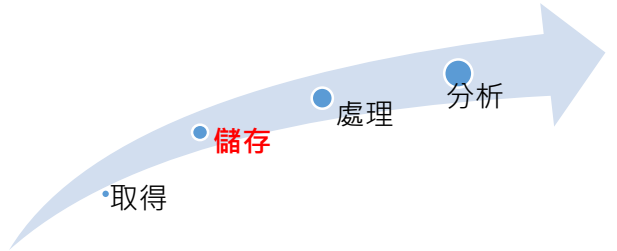
- 資料來源：
  - 行為實驗指標
  - 問卷 / 測驗
  - 廠商/合作單位資料庫
  - 社群媒體資訊

相關技能

SQL查詢(DQL)

網路爬蟲

資料儲存



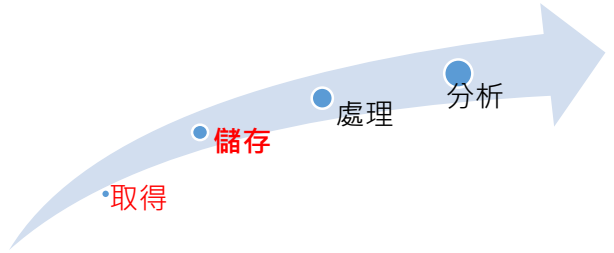
- 存放方式：
  - 單一/多檔案
  - SQL資料庫
  - NoSQL資料庫
- 資料格式：
  - 結構化資料(ex. csv format)
  - 非結構化資料(ex. jpeg format)
  - 半結構化資料(ex. xml format)

相關技能

SQL資料庫定義(DML, DDL)

XML, JSON格式解析

Case：人力銀行爬蟲



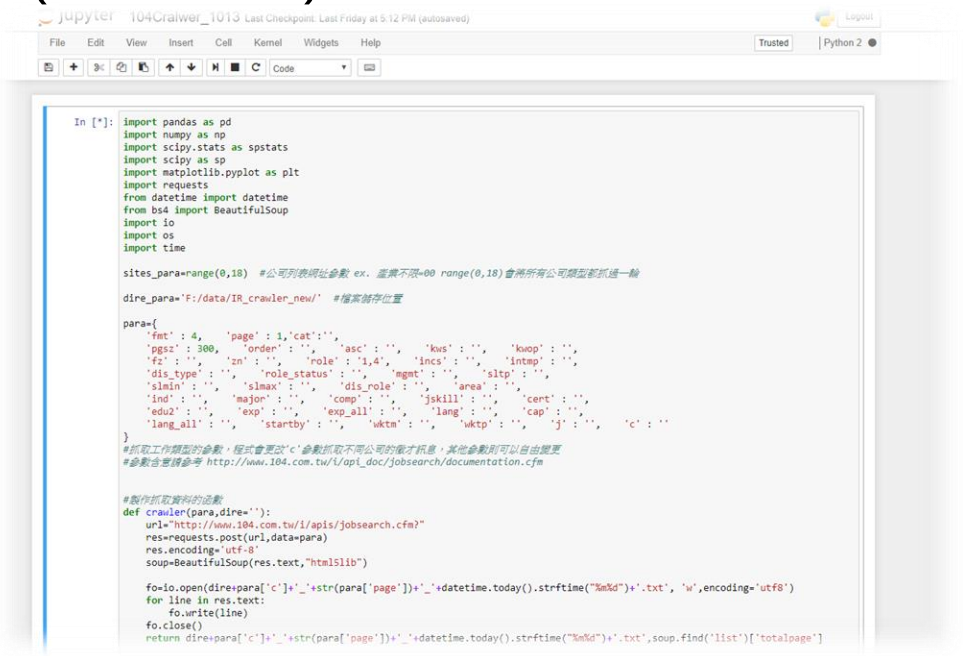
目的：了解就業市場人力需求狀況與趨勢

程式語言：Python 2.7

目標網站：104人力銀行(API服務)、1111人力銀行(網頁爬蟲)

頻率：每日擷取一次(約20萬筆職缺資訊/日)

儲存格式：xml檔案



```
In [*]: import pandas as pd
        import numpy as np
        import scipy.stats as spstats
        import scipy as sp
        import matplotlib.pyplot as plt
        import requests
        from datetime import datetime
        from bs4 import BeautifulSoup
        import io
        import os
        import time

        sites_para=range(0,18)  #公司列表網址參數 ex. 產業不限=00 range(0,18)會將所有公司類型都抓過一輪

        dire_para='F:/data/IR_crawler_new/'  #檔案儲存位置

        para={
            'fmt' : 4,    'page' : 1,'cat':'',
            'pgsz' : 300,    'order' : '',    'asc' : '',    'kws' : '',    'kwop' : '',
            'fz' : '',    'zn' : '',    'role' : '1,4',    'incs' : '',    'intmp' : '',
            'dis_type' : '',    'role_status' : '',    'mgmt' : '',    'sltp' : '',
            'slmin' : '',    'slmax' : '',    'dis_role' : '',    'area' : '',
            'ind' : '',    'major' : '',    'comp' : '',    'jskill' : '',    'cert' : '',
            'edu2' : '',    'exp' : '',    'exp_all' : '',    'lang' : '',    'cap' : '',
            'lang_all' : '',    'startby' : '',    'wktm' : '',    'wktp' : '',    'j' : '',    'c' : ''
        }
        #抓取工作類型的參數，程式會更改'c'參數抓取不同公司的個才抓息，其他參數則可以自由變更
        #參數含意請參考 http://www.104.com.tw/i/api_doc/jobsearch/documentation.cfm


        #製作抓取資料的函數
        def crawler(para,dire=''):
            url="http://www.104.com.tw/i/apis/jobsearch.cfm?"
            res=requests.post(url,data=para)
            res.encoding='utf-8'
            soup=BeautifulSoup(res.text,"html5lib")

            fo=io.open(dire+para['c']+'_'+str(para['page'])+'_'+datetime.today().strftime("%m%d")+'.txt', 'w',encoding='utf8')
            for line in res.text:
                fo.write(line)
            fo.close()
            return dire+para['c']+'_'+str(para['page'])+'_'+datetime.today().strftime("%m%d")+'.txt',soup.find('list')['totalpage']
```

Case：人力銀行爬蟲

Date:1011 Job_count: 210837
----------example----------
MAJOR_CAT_DESCRIPT:工業工程相關@類目@類目
JOB:IE Engineer (NXP Kaohsiung)
J:4a62442b5c38415837343c652e30381b8423c446d
6038422323232673c2d2827563j97
JOBCAT_DESCRIPT:工業工程師／生產線規劃@類目@類目

C:373941243b353d6659313b1c1c1c1c5e143393930
98j55
NAME:NXP Semiconductors Taiwan Ltd._台灣恩智浦半導體股份有限公司
PROFILE:恩智浦半導體NXP Semiconductors N.V. (NASDAQ: NXPI) 在全球…
PRODUCT:恩智浦半導體為IC設計與製造垂直整合的半導體公司…
WELFARE:1. 年終獎金、激勵獎金、員工認股計畫、佳節獎金禮券及多項員工福利補助。 2. …

Data-content

Operate Date: 2017-10-12 08:10:05.597000
##################################

ConnectionError occurred when retrieving list of company :
https://www.104.com.tw/cust/list/index/?page=965&indcat=1000000000&order=1&mode=l&jobsource=checkc

2017-10-12 08:34:09.804000 :29237 companies collected.

2017-10-12 08:34:10.195000 :Crawling 1/1 pages after retry
----------End of collection----------

----------End of the program----------2017-10-12 10:03:40.721000

#Company list#
574149723b3d456e3739416a3f453d208303030744863637119j55
5a7043273c6c3f2548423c1d1d1d1d5f2443a363189j97

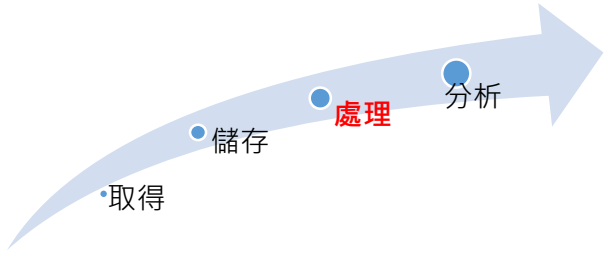
Crawler-log

```
<list Date="2017-10-17">
        <item  JOB="職務名稱"
               LOC="工作地點"
               SKILL="工作技能"> </item>

        <item …>
        </item>
        …
</list>
```

XML-format

資料處理



- 可重複執行的處理流程：
  - 簡潔易修改的程式碼
- 資料處理效率：
  - 資料處理套件使用
  - 決定資料處理演算方式

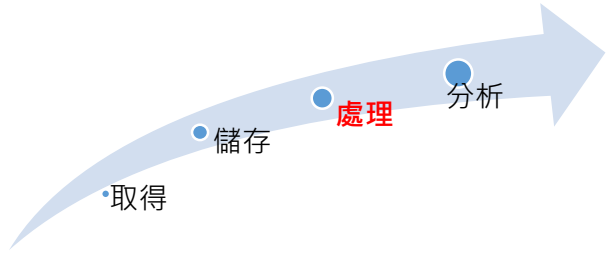
相關知識/技能

R套件使用

Soft skill (coding style)

演算法 / 資料結構

Case：資料處理效率



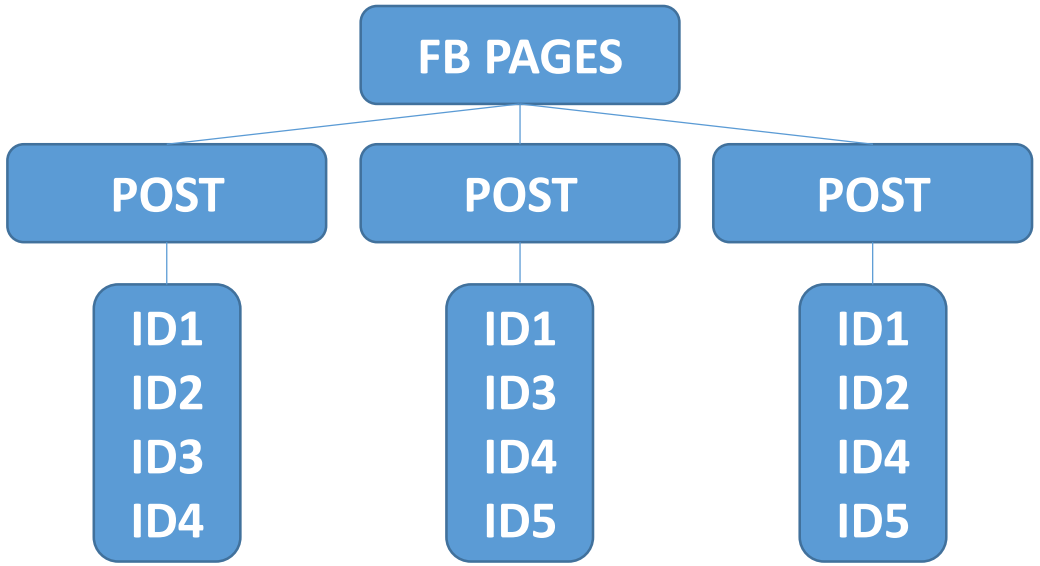
目的：統計某時間間隔下，每一位使用者(ID)對特定粉絲專業貼文的按讚次數、最早按讚日期、最後按讚日期

運算邏輯A：兩層迴圈(貼文 & 按讚ID)

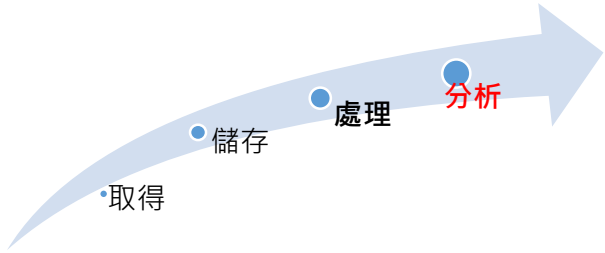
運算邏輯B：apply系列+data.table::rbindlist

dplyr (tidyverse系列)

data.table



資料分析(建立模型)

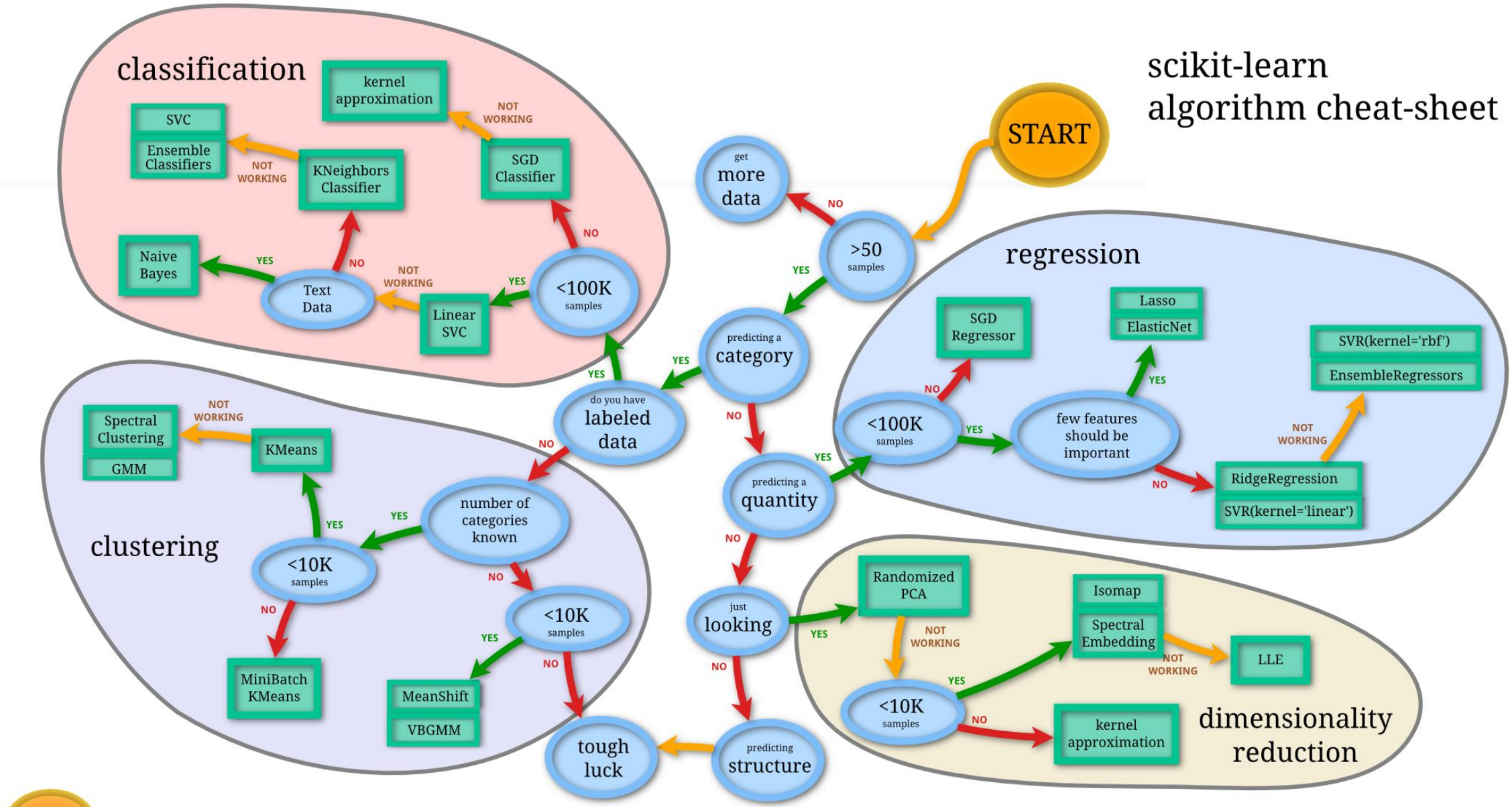


- 各類統計與機器學習模型
  - 回歸模型
  - 決策樹、關聯性法則
  - SVM、MLP
  - Glove、word2vec

- 依照問題與資料類型選擇模型

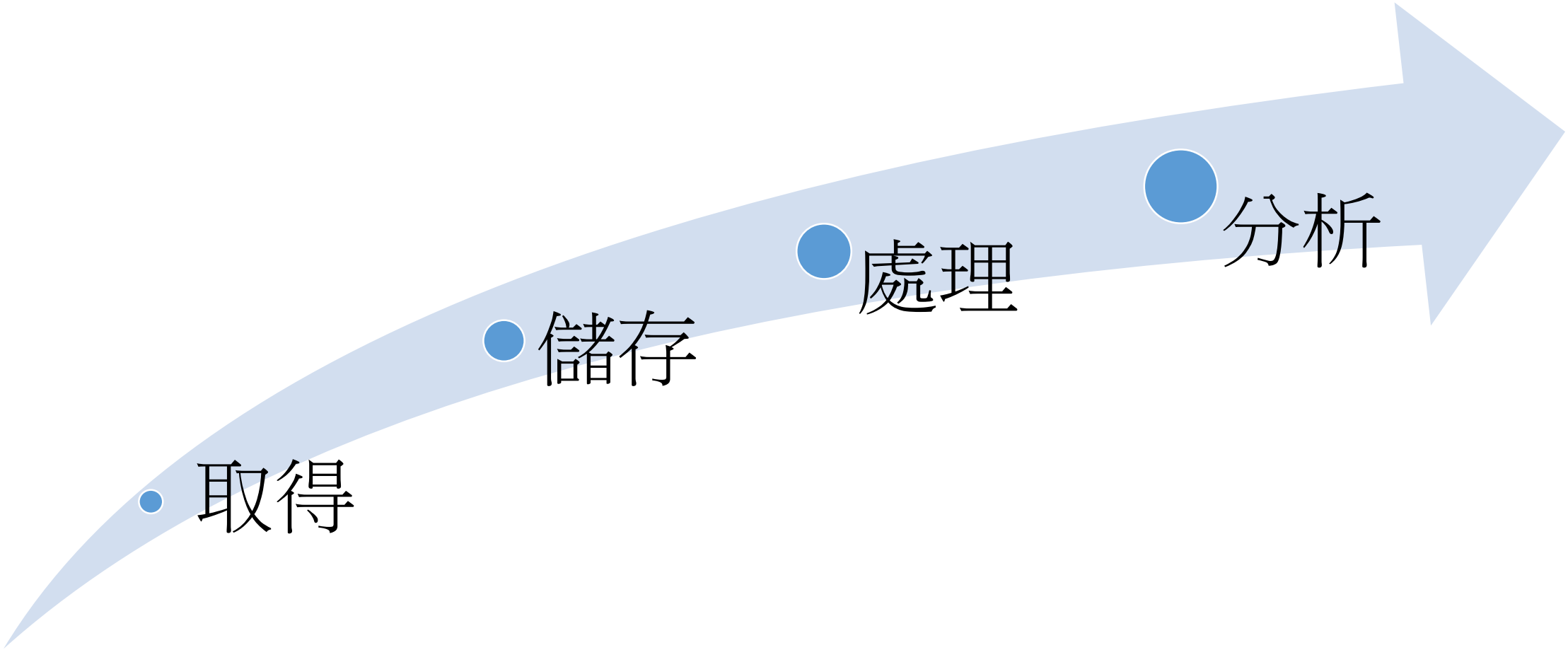
不同領域的人對模型也有不同偏好

scikit-learn algorithm cheat-sheet



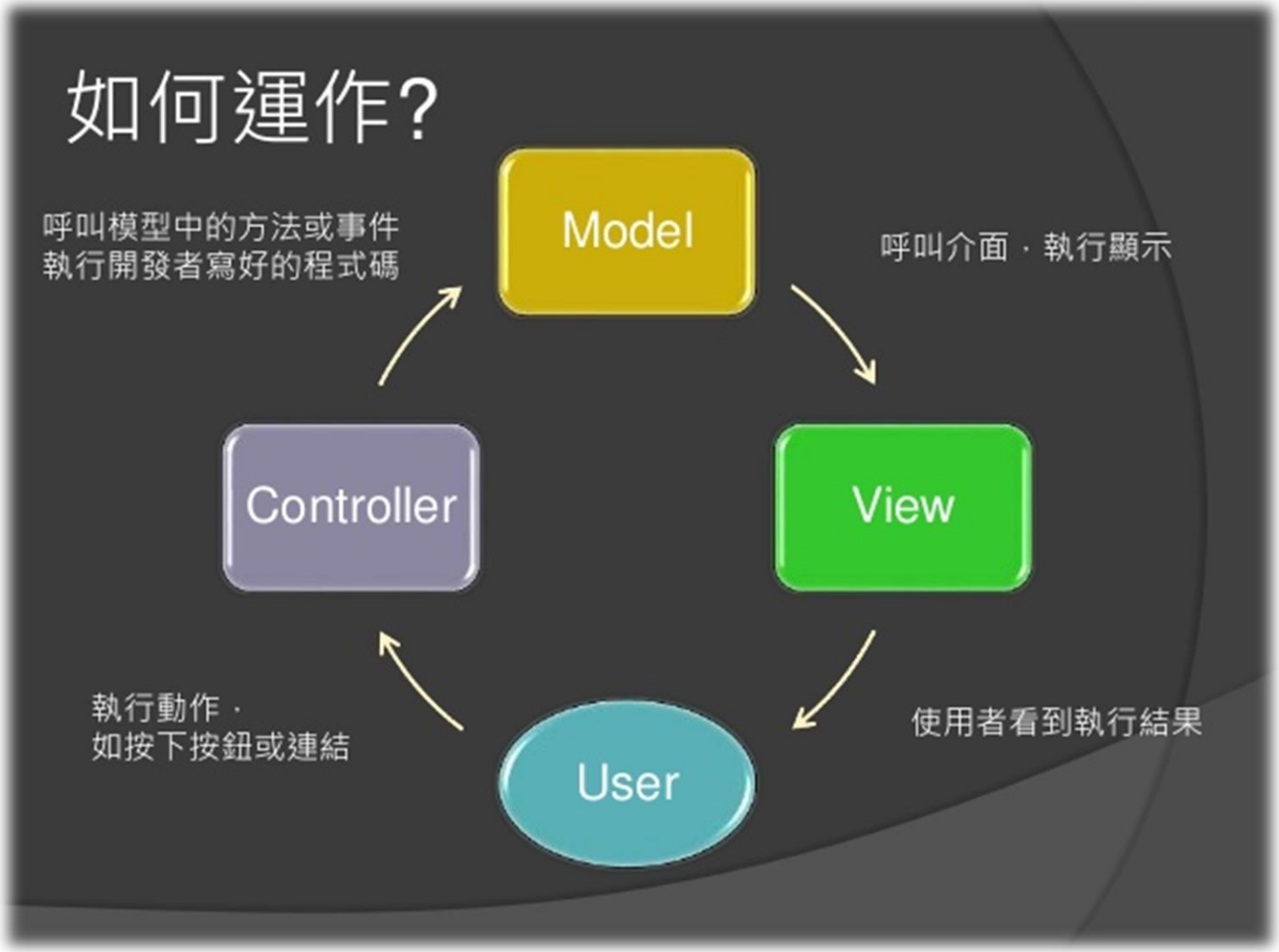
Back

I got a model, and then ?



設計服務架構

- 系統平台服務
  - 自動化分析模組
  - MVC架構
- API服務
  - 資料
  - 統計圖表
  - 預測模型



Case：系統平台

共通職能	職能類別
共通職能	人際互動
共通職能	持續學習
共通職能	問題解決
共通職能	創新
共通職能	溝通表達
共通職能	團隊合作

專業職能	職能類別
專業職能	人力資源規劃
專業職能	人員管理
專業職能	工作分析
專業職能	面談技巧
專業職能	員工訓練發展
專業職能	員工關係管理
專業職能	專案管理

平台設立與功能簡介

本平台的設立是由謝邦昌教授所帶領的研究團隊並在李艦老師的幫忙下逐漸成形。
不同於以往的數值資料分析，潛藏在文字之間的資料量越來越龐大，而其中透露出來的趨勢以及信息亦讓人無法忽略。因此我們利用 文字探勘技術 結合 爬蟲技術 ，去建立一個 輿情語意分析平台 讓使用者能夠利用平台，獲取所需的資訊。

本平台有三個主要功能，功能如下:

Crawl

可以供使用者輸入疾病的關鍵字詞，並且利用本平台將相關文章自動爬取下來，以利於接下來的文字探勘分析。

Text mining

可以供使用者自行輸入txt檔文本，並進行文字探勘分析：詞雲圖，集群分析，脈絡分析及關聯分析。

Facebook

對臉書粉絲團進行爬文，並可分別對管理員貼文及粉絲留言進行分析

CDMS官網
台北醫學大學管理學院
台北醫學大學大數據研究中心
Fanpage

API服務

http://10.95.3.41:8080/IR_API/q_jobTrend.jsp?major_cat=資訊工程相關

[{"時間周期":" '2018-01', '2018-02', '2018-03', '2018-04', '2018-05’”},
{"職缺":"Internet程式設計師","職缺數量":["1246.2","1249.43","1246.57","1261.57","1268.83""]},
{"職缺":"MIS程式設計師","職缺數量":["992.43","974.86","995.4","995.43","1008"]},
{"職缺":"系統維護 / 操作人員","職缺數量":["899","882.57","914.5","893.5","894.57"]},
{"職缺":"軟體設計工程師","職缺數量":["4306","4179.43","4336","4247.75","4316.71"]},
{"職缺":"韌體設計工程師","職缺數量":["1135","1094.71","1125","1120.5","1130.71"]}]

http://10.95.3.41:8080/IR_API/predict_model?gender=1&college=CD&scc=1&itest=82&etest=1&drank=0.21

[{”work_prob”: 0.52},{”study_prob”:0.28}]

Related R package:plumber

資料科學

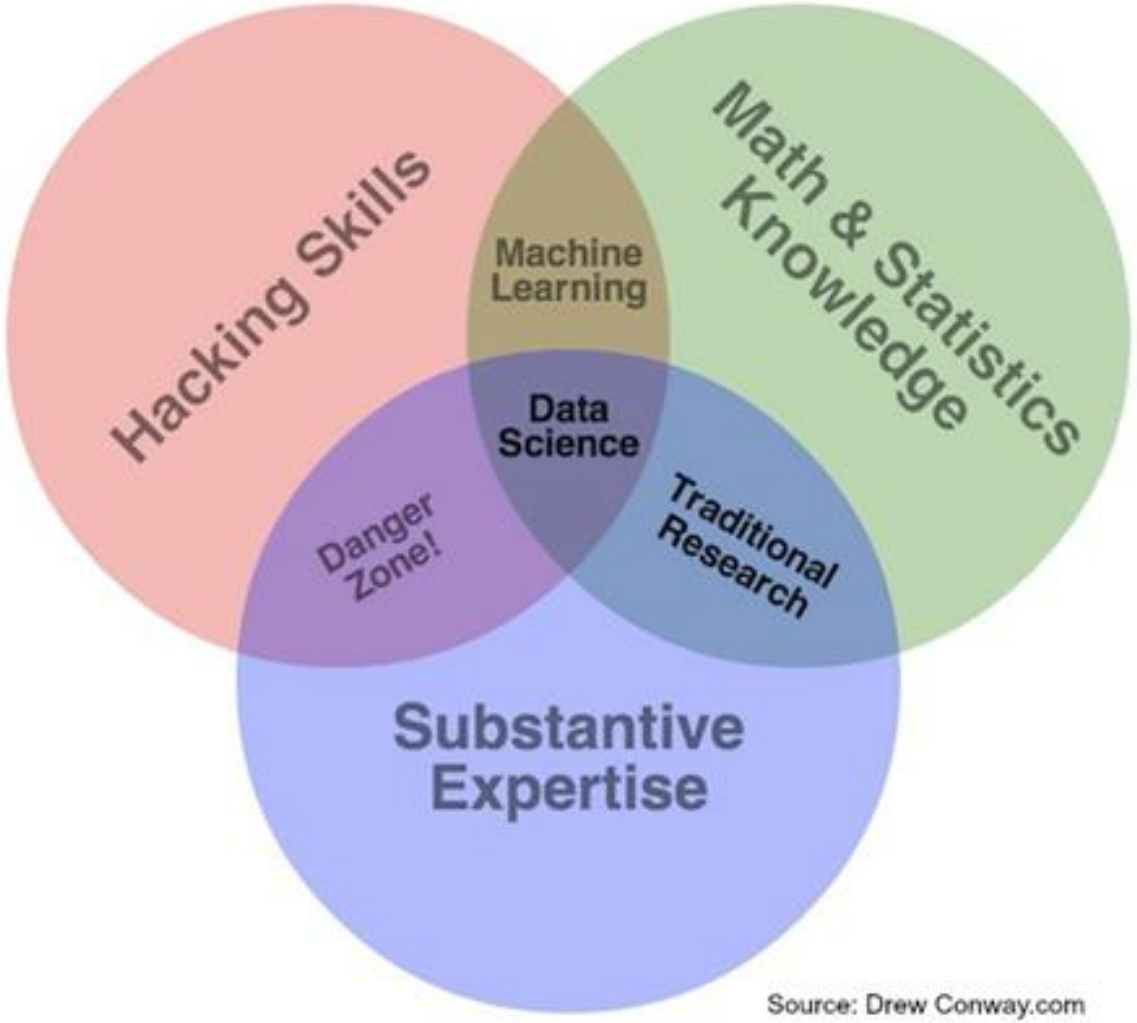
跨領域的學門

工業工程　　資訊工程

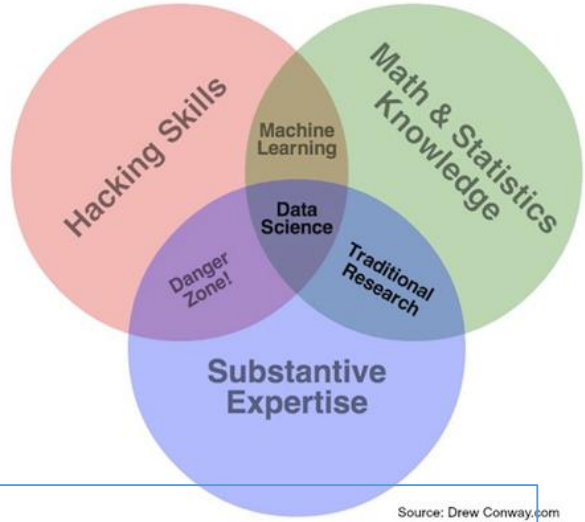
心理學　　財務金融

資料科學-跨領域的學門

- 不同領域的交集
  - 知識 / 技能
  - 對問題的切入點
  - 對模型的選擇偏好



資料科學-跨領域的學門



統計學的六大領域

目的在於找出實際狀況的**社會調查法**

目的在於找出原因的**流行病學及生物統計學**

目的在於測量抽象概念的**心理統計學**

目的在於機械式分類的**資料採礦**

目的在於處理自然語言的**文字勘察**

著重在推論的**計量經濟學**

西內啟— 統計學，最強的商業武器

資料科學-跨領域的學門

ID	性別	身分別	院別	級別	整體滿意
1	0	3	4	2	5
2	1	3	4	4	5
3	0	3	4	1	4
4	1	3	2	1	5
5	1	3	1	1	5

資料量	大	小
解釋變項(x)意義	模糊	明確
(較)常用模型	Machine learning	Statistical model

Case：背景不同，想法不同

學業成績

社團參與

打工紀錄

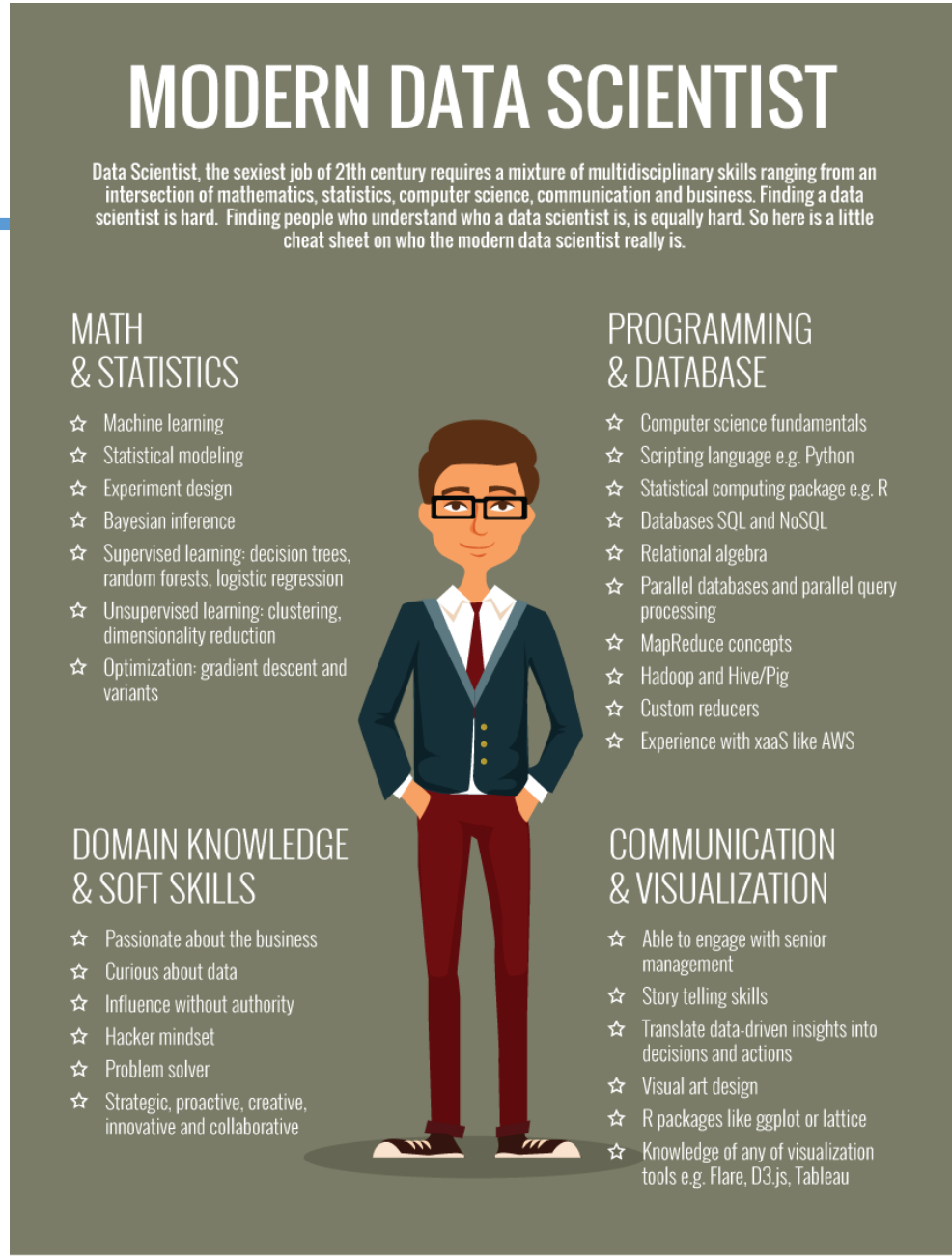
哪些因素影響學生
畢業後走向

如何選擇模型?

模型能夠預測多準

資料科學-跨領域的學門

- 目前仍以電腦科學/資訊工程領域為主流
  - 產業結構
  - 硬實力(門檻較高)
  - 應用能力強

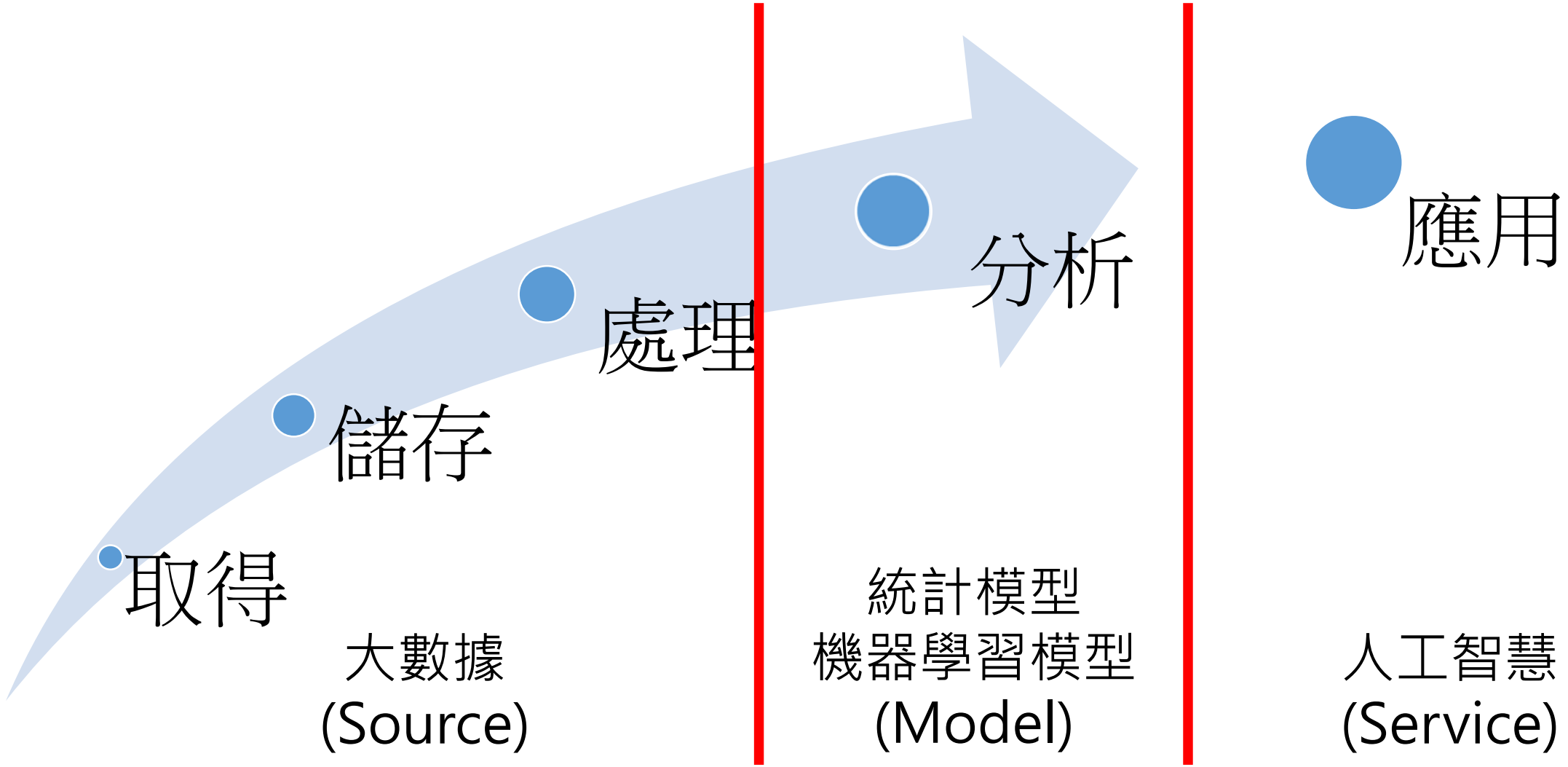


社會科學訓練下的強項為何?

- 特定領域的知識(Domain knowledge)
  - 產品/服務仍然針對人做設計
- 對資料品質的敏感度
  - 誤差/隨機性

優勢：資料取得設計

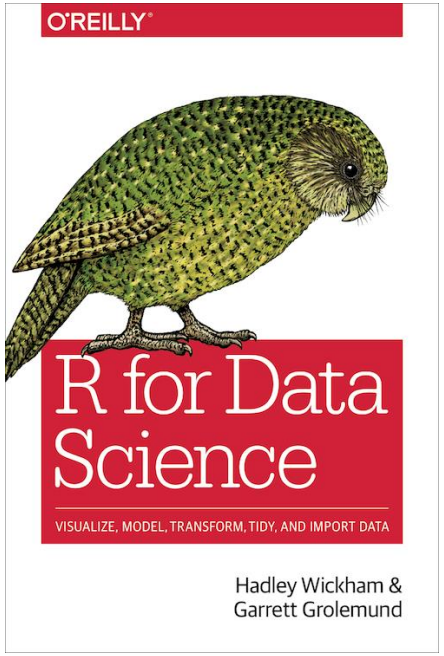
大數據、AI、資料科學、機器學習？



在資料科學的浪潮之中

- 持續學習
  - 廣泛涉略 v.s. 專精領域？

- 幸運的是，學習資源永不缺乏
  - **Coursera** Machine learning (Andrew Ng)
  - **PTT** R_language版
  - **Facebook** 台灣資料科學同好交流區
  - **資料科學年會系列活動**



http://r4ds.had.co.nz/

感謝聆聽

歡迎任何問題、討論

工作至今接觸過的資料源

- 工廠工單資料
- 公車票卡紀錄資料
- 學生學籍資料、學習歷程資料
- 人力銀行職缺資料

工單資料範例

车间生产派工单

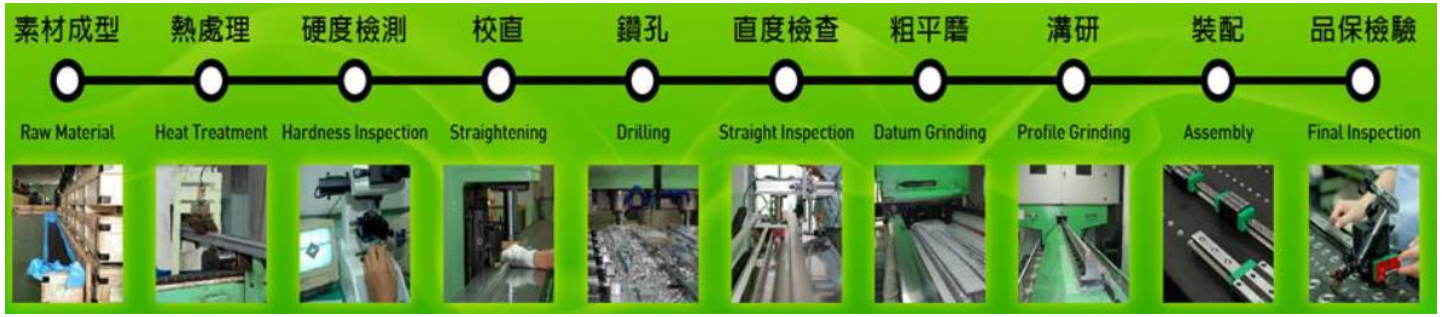
生产定单:	20091204000002	派工单号:	KA1234567890-0	油漆颜色:		定单标识:	
产品型号:	CB252-IV	工 程 号:	0001	工 作 号:		批　　次:	
组件代号:		项目代号:	2KA.021.034.5.7	项目名称:	断路器总装	图号版次:	
登记日期:	2009-12-22	上 工 序:		下 工 序:		计划数量:	20.00

简图

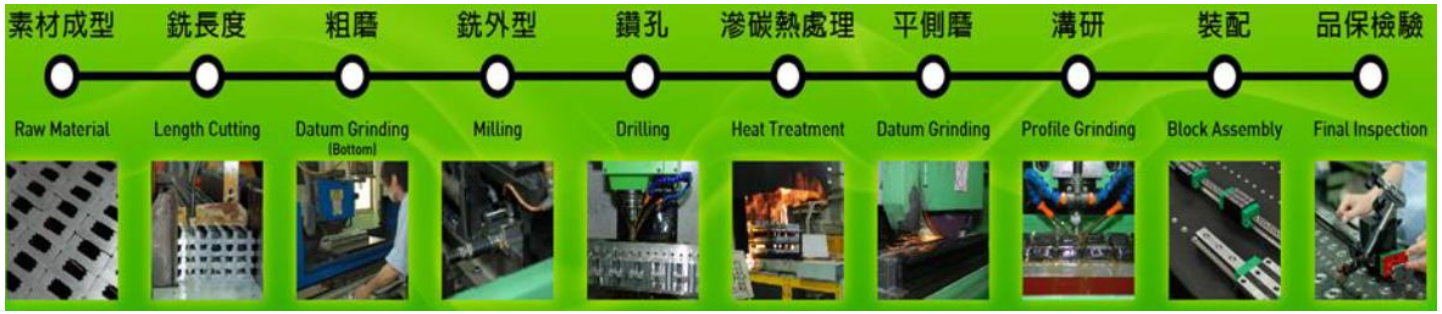
工序号	工序代码	工序名称	加工车间	工作中心	加工设备	准备工时	加工工时	单件数量	合格数量	回用数量
2	BL	备料	ZZ	ZX003				20		
3	GRSX	固溶时效	ZZ	ZX003	SB03			20		
5	DM	打磨	ZZ	ZX003				20		
10	QZ	钳装	16	JJZX	SB01			20		
12	ZH	装焊	32	ZX002	SB02			20		
13	HJ	焊接	32	ZX002				20		
15	QL2	清理	Z2	ZX001				20		
9	T	镗	16	JJZX	SB03			20		
14	XZ	校正	32	ZX002	SB02			20		
11	QYSY	气压试验	16	JJZX	SB03			20		
16	PX	配线	Z2	ZX001	SB02			20		
17	ZX	组线	Z2	JJZX	SB03			20		

工單資料-生產製程

滑軌



滑塊



組合件

排程問題：如何決定訂單製作順序/機台分配?

最佳化問題

公車搭乘紀錄分析

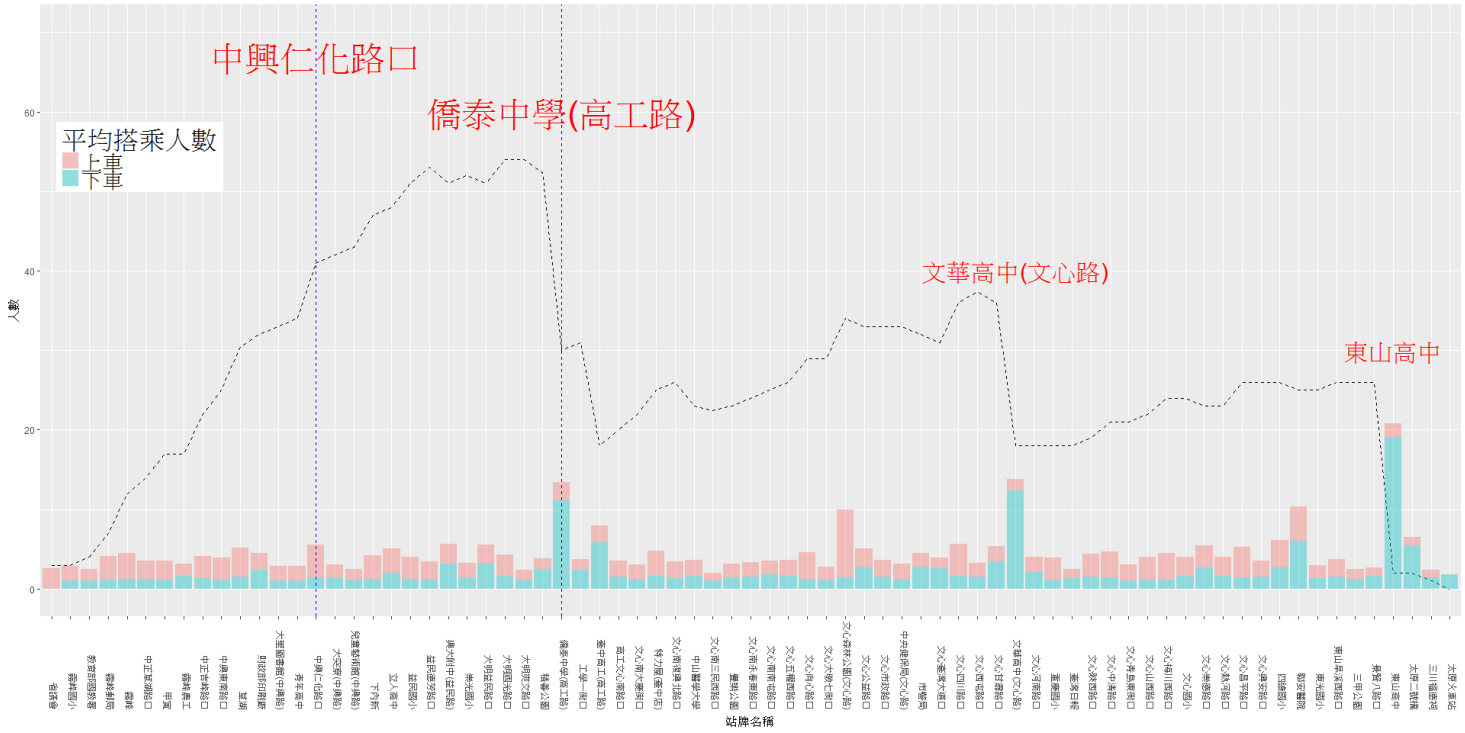
	Id	OperateDate	RouteNo	RouteStart	RouteEnd	Plate	CardCompany	TicketType	CardNo	BoardTime
▶	b-8163941a1f17	2015-05-04	93	高鐵臺中站	銅安厝	062-U8	ECC	全票	2591017487	2015-05-04 17...
	00000089-c50...	2015-12-28	53	0	0	849-U5	ECC	全票	262F9B95	2015-12-28 16...
	0000009a-81e...	2015-02-01	105	四張犁國小	龍山國小	558-FQ	ECC	全票	2689306493	2015-02-01 11...

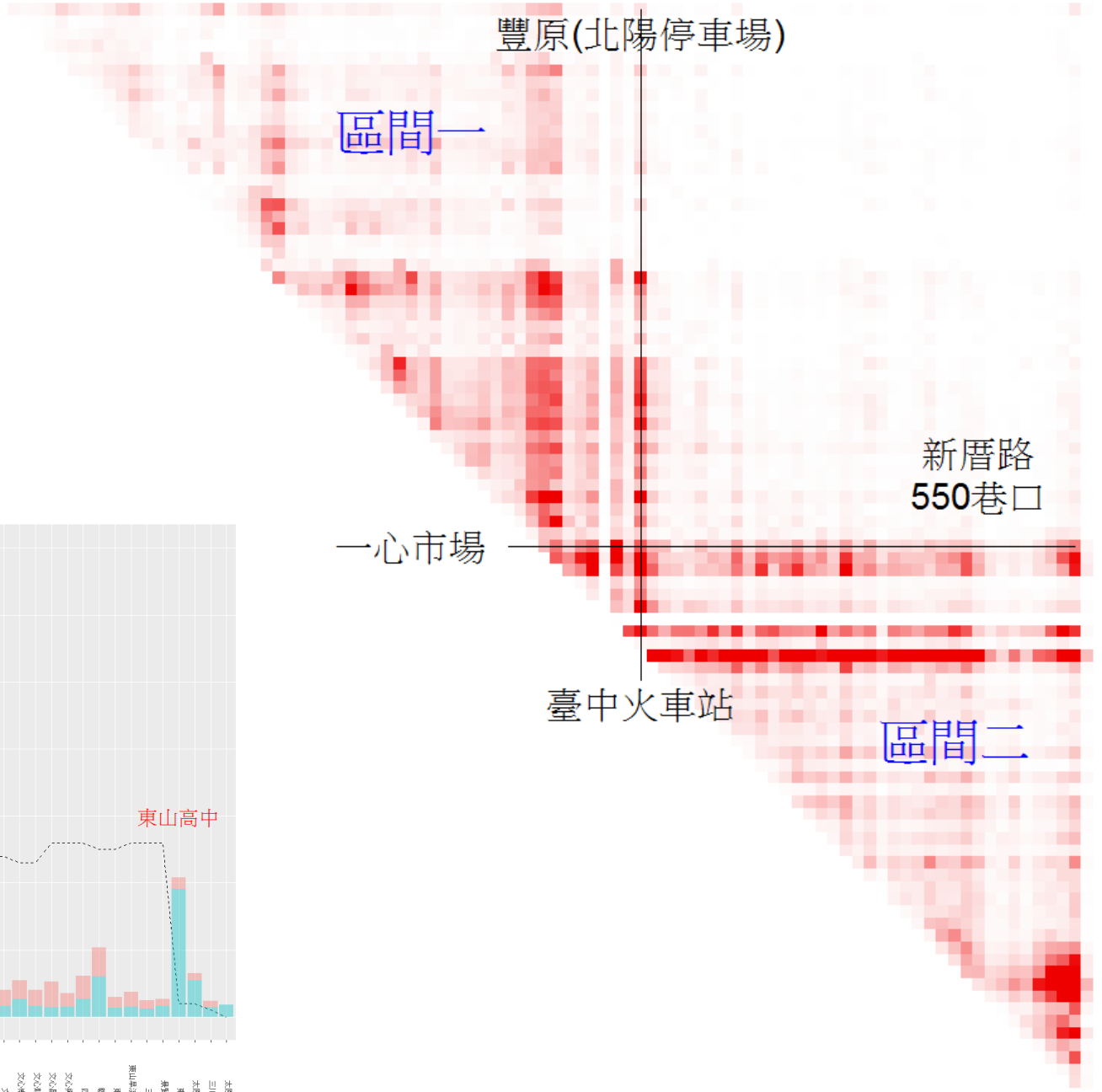
BoardCostM...	BoardTransfe...	BoardDiffere...	BoardStopSe...	BoardStop	BoardCharge...	BoardTransfe...	AlightTime	AlightCostMo...	AlightTransfer...
20	0	21.596	61	清水高中	清水高中	1	2015-05-04 17...	21	0
20	0	21.596	47	工學一街口	臺中高工(高工路)	1	2015-12-28 17...	2	0
20	0	21.596	51	九張犁	九張犁	1	2015-02-01 12...	0	0

AlightDifferen...	AlightStopSer...	AlightStop	AlightCharge...	AlightTransfer...	Direction	SubTotal	FileName	CreatTimeUtc	AbsoluteBoar...
0	89	日南	日南	1	1	42.596000671...	D:\abc\Taichu...	2016-02-12 03...	4129
0	71	霧峰郵局	霧峰	1	1	8.8000001907...	D:\abc\Taichu...	2016-06-30 05...	4258
0	69	臺中火車站	臺中火車站九...	1	2	21.596000671...	D:\abc\Taichu...	2016-02-16 02...	994

AbsoluteAlig...	AbsoluteBoar...	AbsoluteAlig...	IsArrange
4125	False	False	True
3978	False	False	True
1824	True	True	True

公車搭乘紀錄分析


中興仁化路口
僑泰中學(高工路)
文華高中(文心路)
東山高中
平均搭乘人數
上車
下車
人數
站牌名稱


豐原(北陽停車場)
區間一
新厝路
550巷口
一心市場
臺中火車站
區間二

人力銀行職缺分析-文字探勘