café選址樂

第三組

組長：周旻柔

組員：張家寧、史晏臺、許倫彬、 陳冠儒、龔宣銘

大綱

- 團隊介紹
- 專題簡介
- 資料蒐集
- 資料清理
- 模型建置
- 成果視覺化
- 未來展望

團隊介紹

周旻柔
組長

專案管理
環境建置

張家寧
組員

環境建置
成果視覺化

史晏臺
組員

資料蒐集
模型建置

許倫彬
組員

資料清理
資料蒐集

陳冠儒
組員

模型建置
資料清理

龔宣銘
組員

成果視覺化
資料蒐集

主講人:周旻柔

2 專題簡介

1. 研究動機

2. 資料需求

專題簡介：研究動機

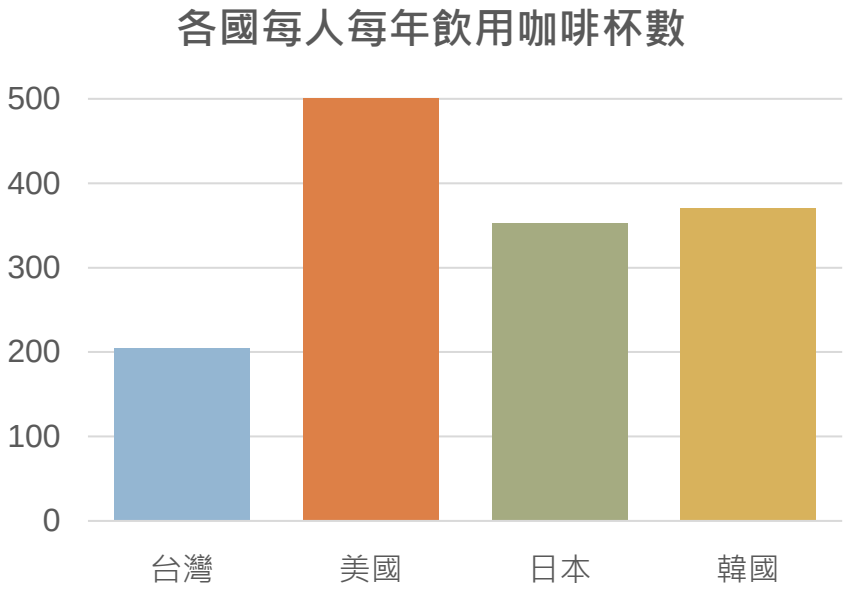
周旻柔

- 再忙也要跟你喝杯咖啡~
- 整個城市，就是我的咖啡館

專題簡介：研究動機

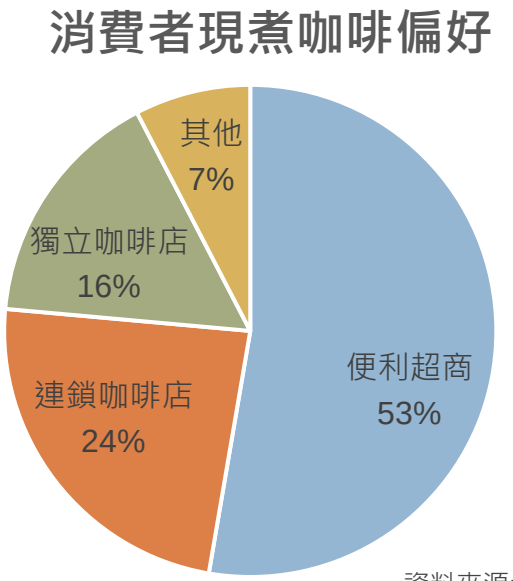
周旻柔

- 2020年台灣人一年就能喝掉**28億杯**咖啡，平均**每人200杯**咖啡，市場規模上看新台幣**800億元**。



資料來源: 國際咖啡組織（ICO）

- 消費者現煮咖啡偏好
  - 便利超商佔**5成**
  - 連鎖咖啡與獨立咖啡店佔**4成**。

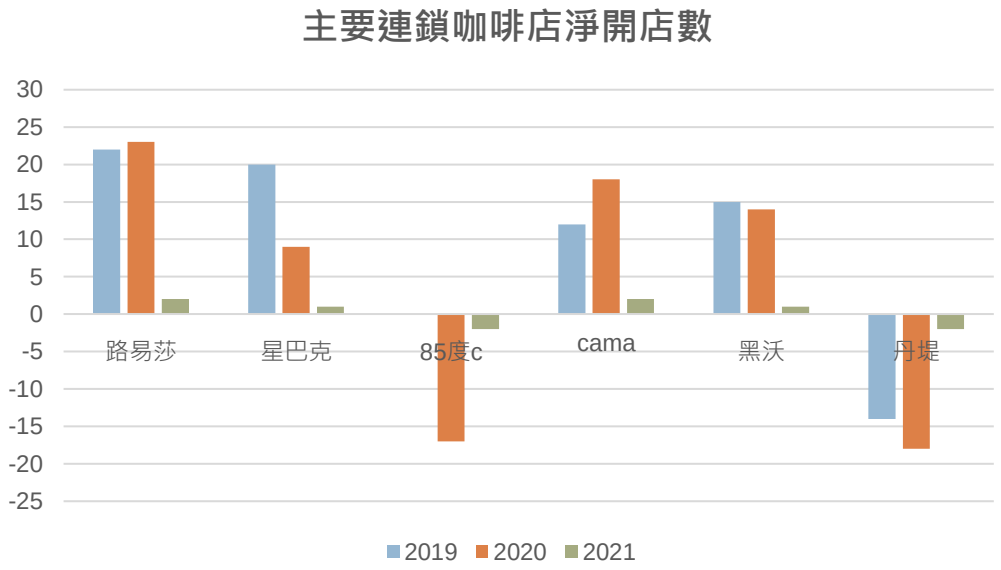


資料來源:《食力》2019年問卷調查

專題簡介：研究動機

周旻柔

- 台灣咖啡市場早已進入卡位戰。
- 全台有供應咖啡的店家超過**2萬**間。
  - 四大便利超商
  - 連鎖和獨立咖啡店
  - 超市和速食業者
  - 早餐店
- 台灣連鎖咖啡店2020年以前積極開店。



資料來源:《數據實驗室-Data Lab》

專題簡介：研究動機

周旻柔

- 決定開店地點後，營業額的 70% 就已決定
  - 開店超過1年的機率：10%
  - 開店超過 5 年的機率：1%
- 我們可以提供哪些服務?
  - 蒐集既有咖啡店的商圈資訊
  - 推薦可能有高營收的咖啡店開店位置

專題簡介：資料需求

周旻柔

- 商圈分析
  - 市場規模：商圈人口密度與成長率
  - 品質：商圈家戶所得
  - 商業氣息： 其他業別店家數
  - 競爭程度：同業店數
- 營收替代指標
  - Google商家評論數

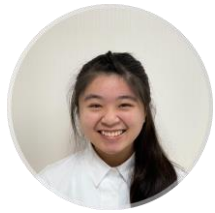
主講人:張家寧

3

環境建置：Hadoop叢集

1. 叢集架設的工具
2. 叢集建置的環境
3. 叢集內部結構、運作流程
4. Spark叢集
5. Hadoop MapReduce vs. Spark

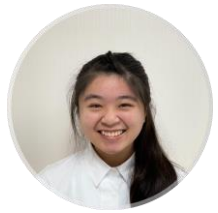
Hadoop叢集架設



張家寧

- jupyter + python™ + APACHE Spark™
- hadoop YARN
- hadoop HDFS
- Linux + VMware Workstation 16 + ubuntu-20.04.3

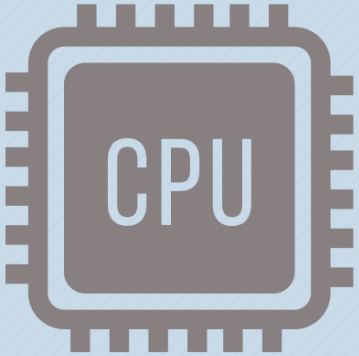
Hadoop叢集環境



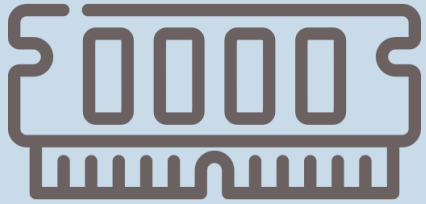
張家寧

實體主機 6 台

虛擬主機 3 台

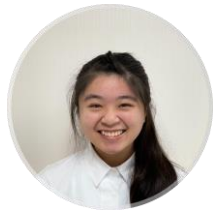


每台虛擬主機
皆配置4顆CPU

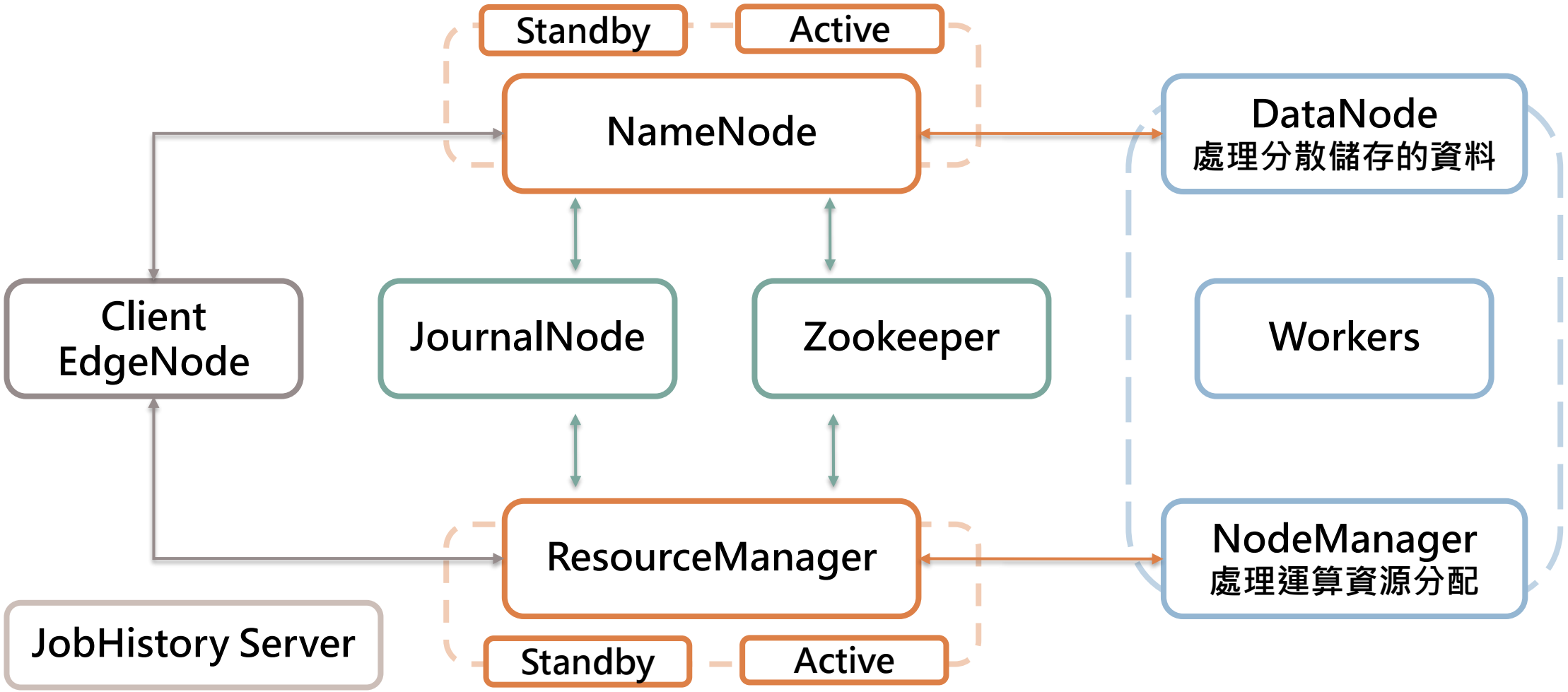


每台虛擬主機
皆配置 16G RAM

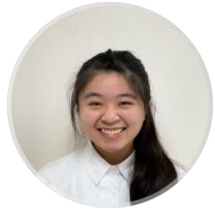
Hadoop叢集：節點介紹



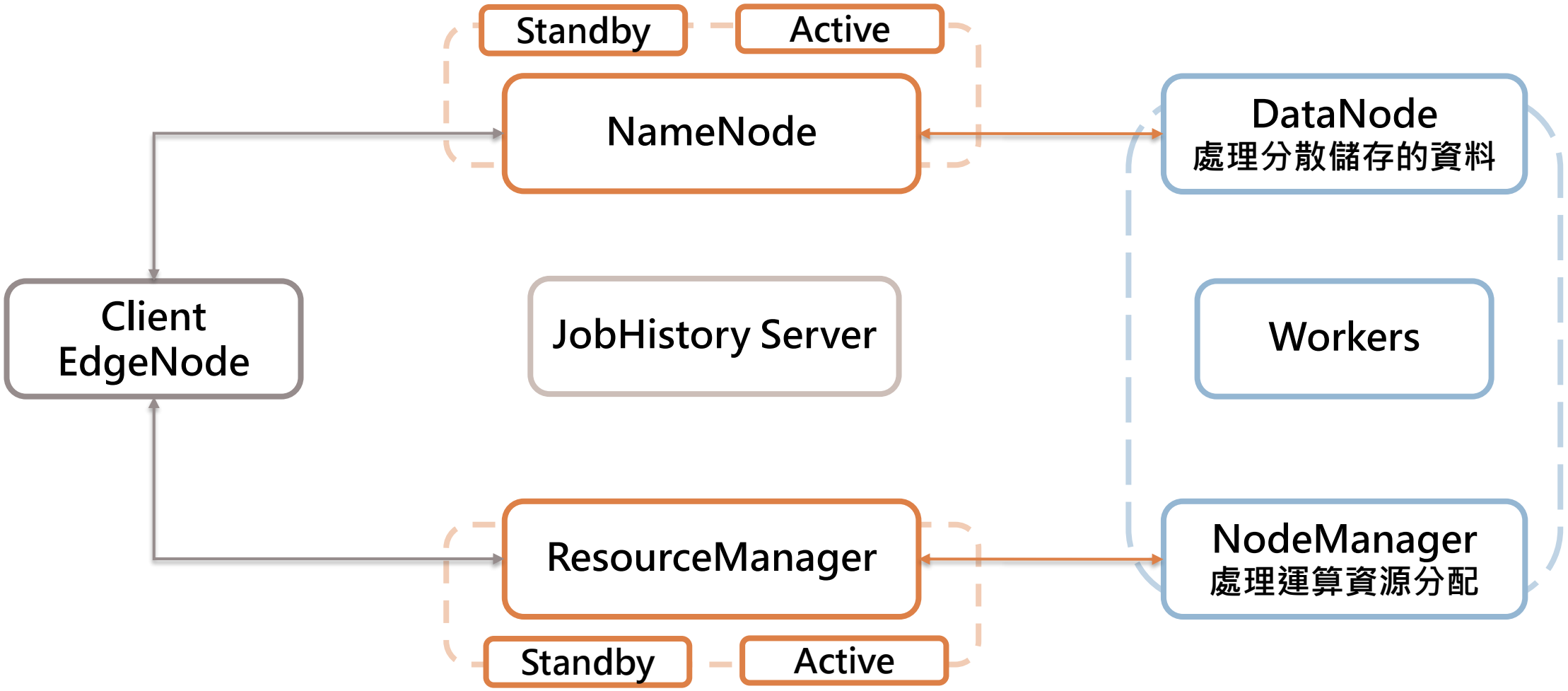
張家寧



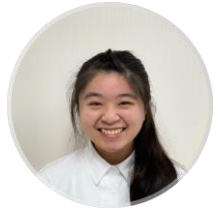
Hadoop叢集：節點介紹



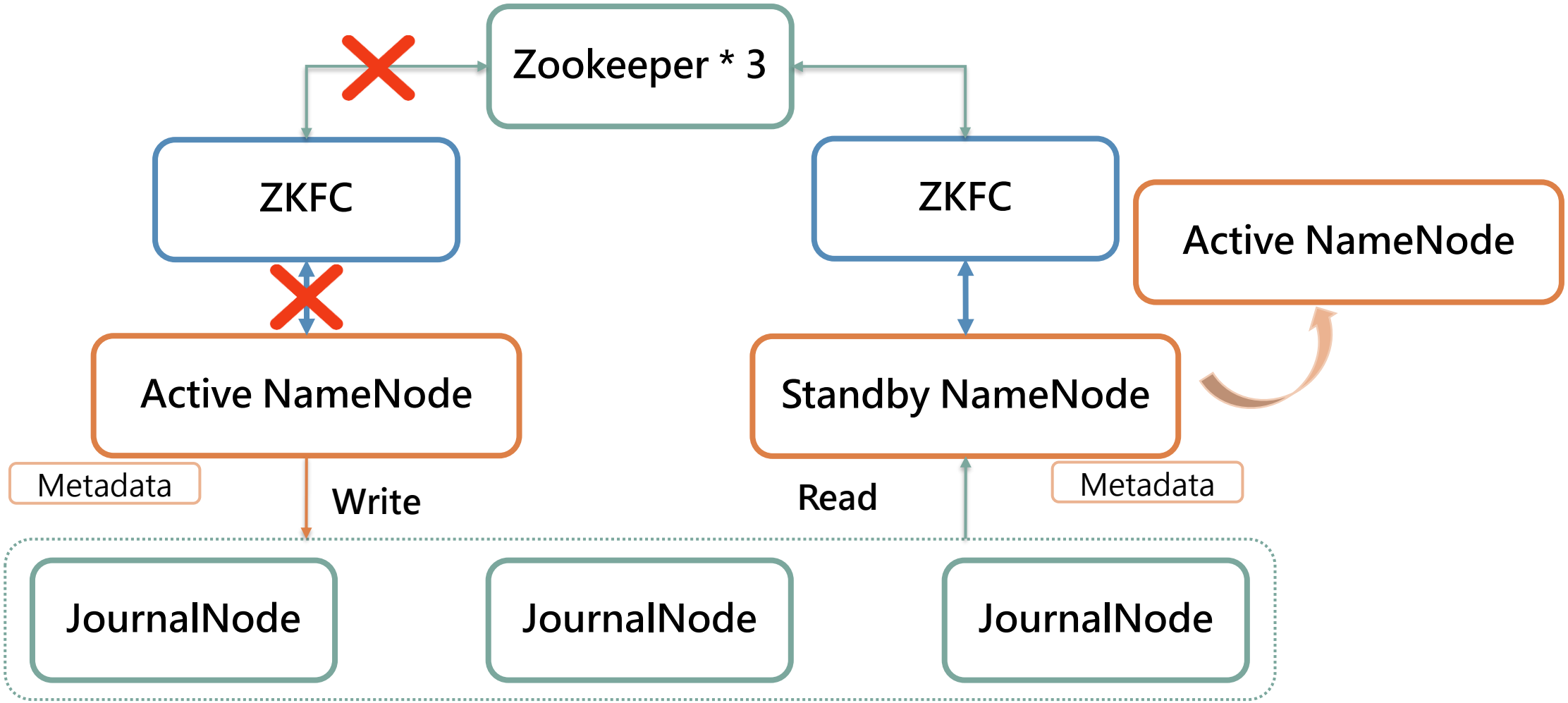
張家寧



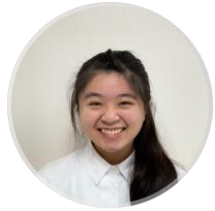
Hadoop叢集：ZKFC



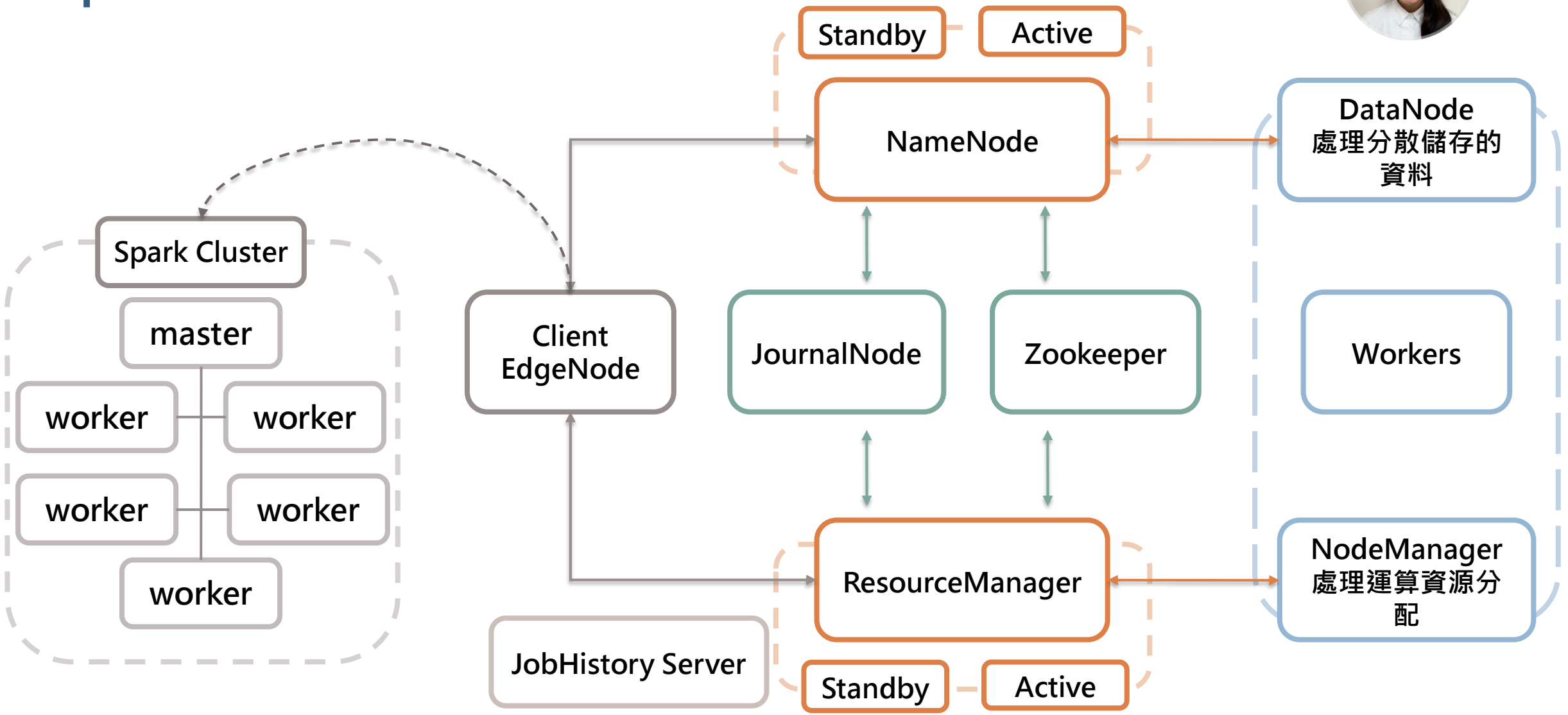
張家寧



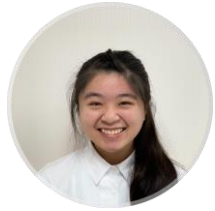
Spark叢集



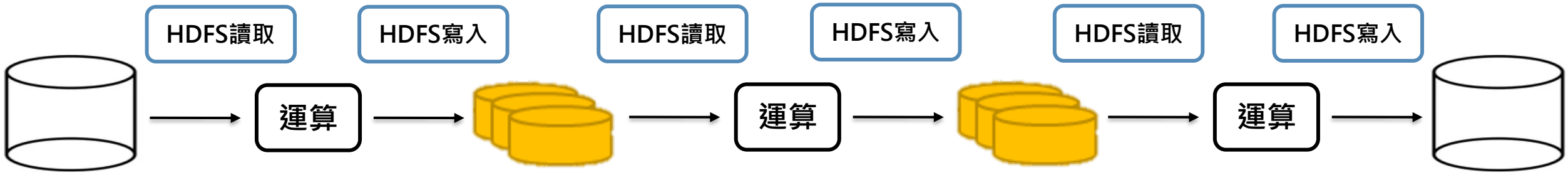
張家寧

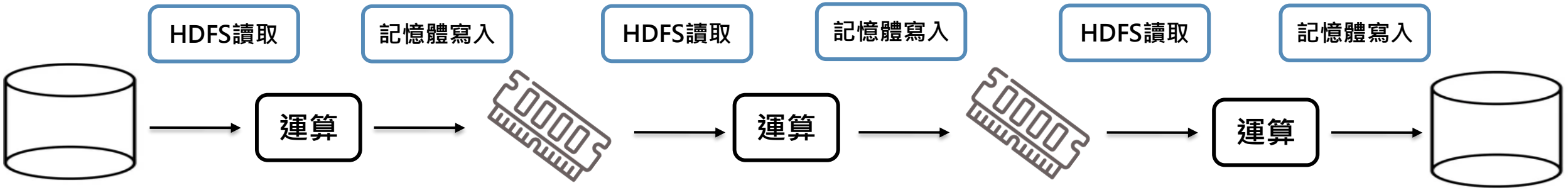

Standby
Active
NameNode
DataNode
處理分散儲存的資料
Spark Cluster
master
worker
worker
worker
worker
worker
Client
EdgeNode
JournalNode
Zookeeper
Workers
ResourceManager
NodeManager
處理運算資源分配
JobHistory Server
Standby
Active

Hadoop MapReduce vs. Spark



張家寧



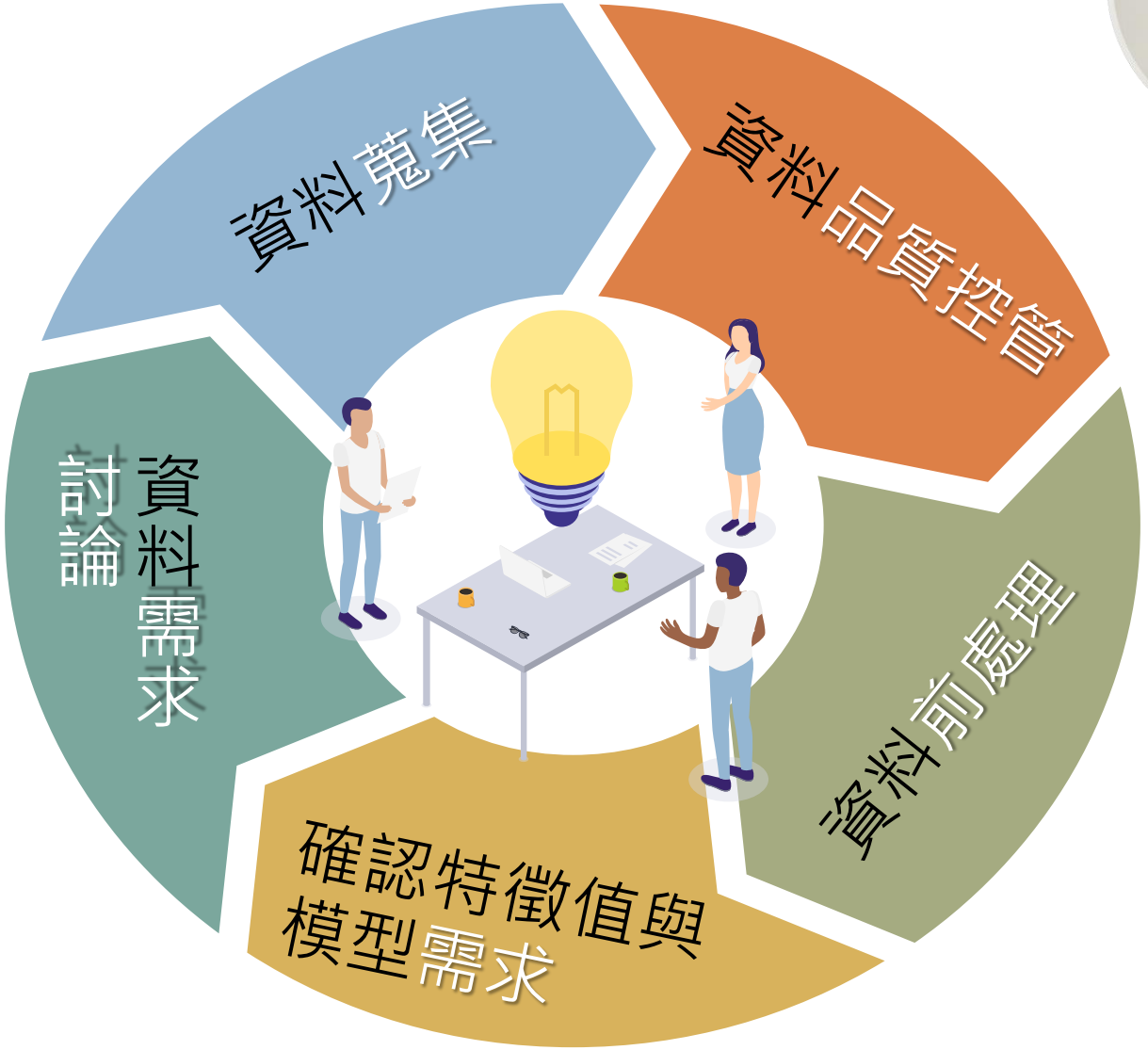


主講人:史晏臺

4

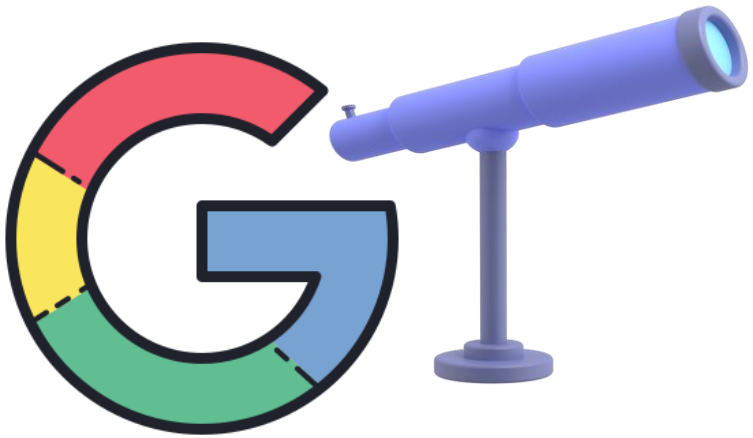
資料蒐集

1. 撰寫程式
2. 資料前處裡
3. 系統排程
4. Google Cloud 使用



使用工具及資料來源

史晏臺

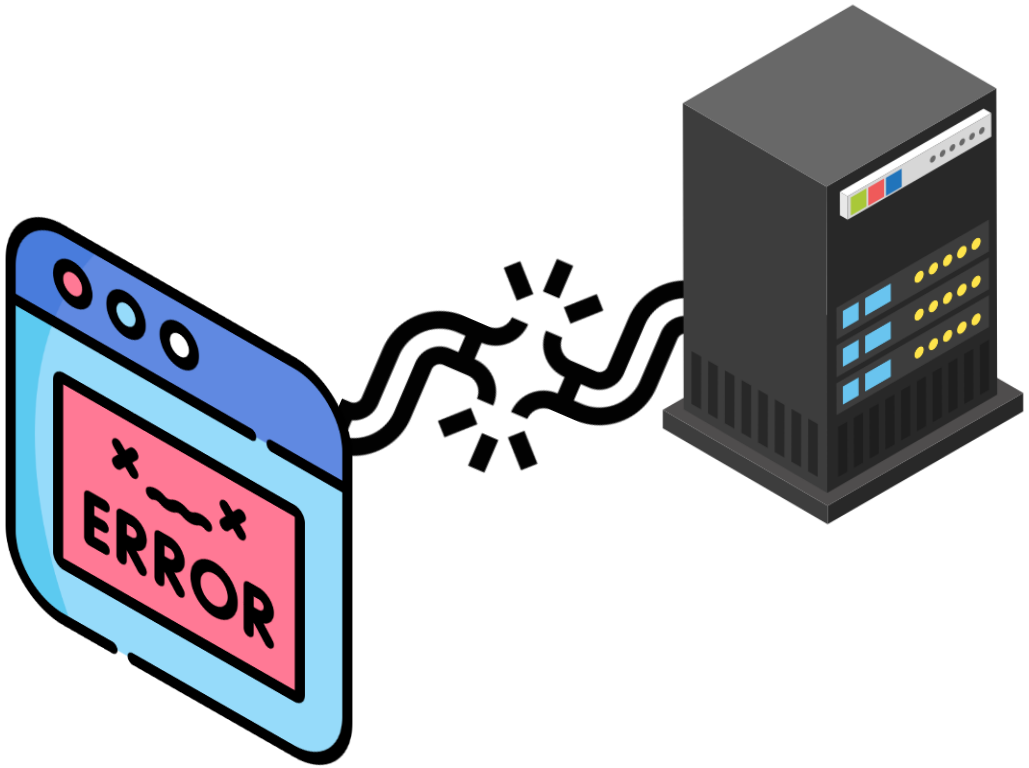


pandas

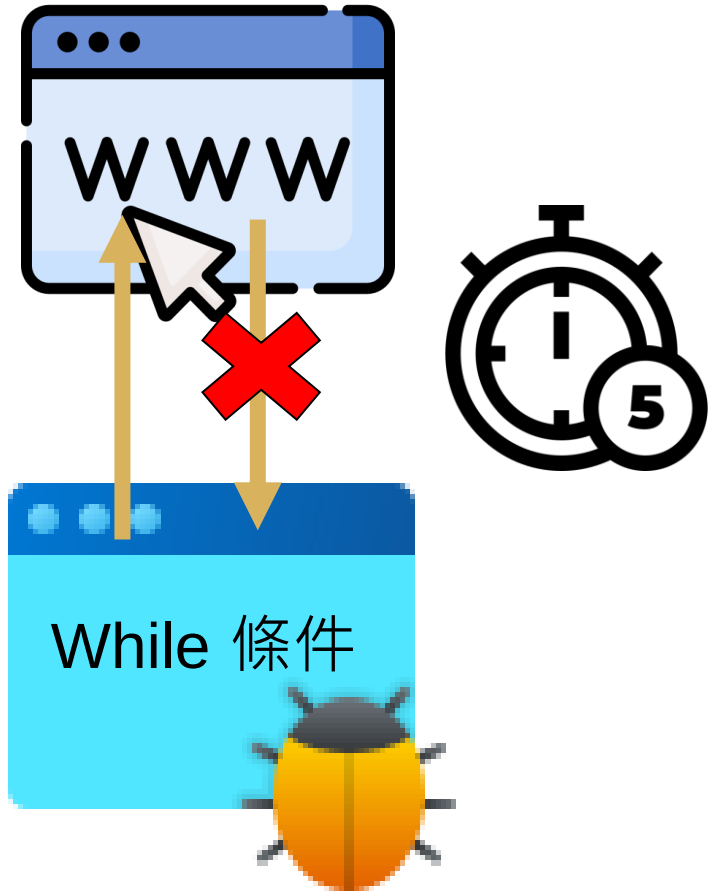
案例1. 氣象擷取問題

史晏臺

問題發現:
拒絕請求。



解決方案:
指定條件相符才能停止。



案例2. 評論數、商家特徵擷取問題

史晏臺

問題發現:
問題資料程式中止。

解決方案:
測試例外處理完成。

問題資料:

YUME Cafe

4.4 ★★★★½ 123 則評論

永久停業

正常資料:

917好事咖啡創意廚房-不限時咖啡廳|台北市大安區六張犁|早午餐下午茶|聚餐聚會讀書工作|課程講座會議場地租借|謝師宴尾牙迎新送舊|外帶美食79折

4.2 ★★★★☆ 587 則評論 · $$

咖啡廳

程式測試:

```
彼得好咖啡 ok!
 河田咖啡館 ok!
良日激動所（已遷址）找不到 例外訊息:list index out of range
雲上咖啡 ok!
Coffee Sweet ok!
```

案例3. 人口、房屋特徵擷取問題

史晏臺

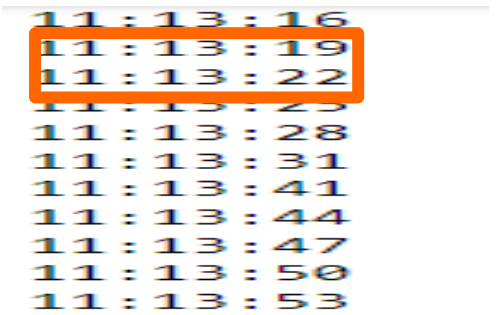
問題發現:
不易進入目標網頁。

解決方案:
撰寫仿人點擊。

效率測試:
Selenium影響作業時間。

目標網頁

Selenium: 1筆/3sec	主程式: 8筆/1sec



主程式使用套件

案例4. 分散工作設計腳本

史晏臺

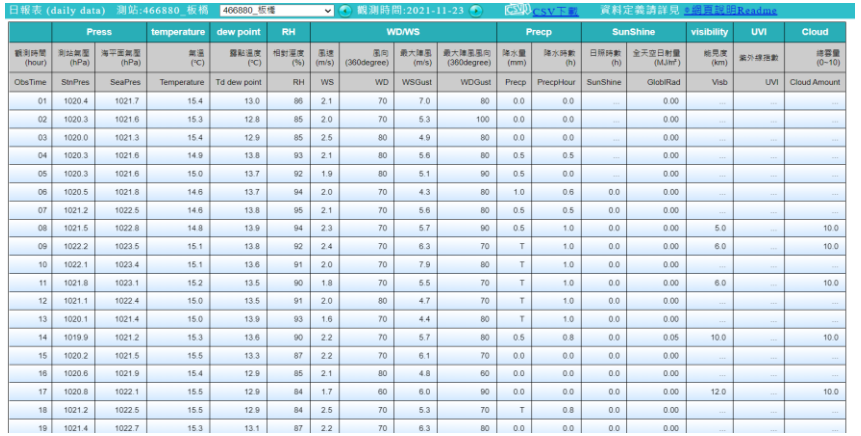
測試完可行程式，使用crontab掛載。

```
 GNU nano 4.8                                  /etc/crontab
# command to install the new version when you edit this file
# and files in /etc/cron.d. These files also have username fields,
# that none of the other crontabs do.

SHELL=/bin/sh
PATH=/usr/local/sbin:/usr/local/bin:/sbin:/bin:/usr/sbin:/usr/bin

# Example of job definition:
# .---------------- minute (0 - 59)
# |  .------------- hour (0 - 23)
# |  |  .---------- day of month (1 - 31)
# |  |  |  .------- month (1 - 12) OR jan,feb,mar,apr ...
# |  |  |  |  .---- day of week (0 - 6) (Sunday=0 or 7) OR sun,mon,tue,wed,thu
# |  |  |  |  |
00 6,12,18,24  *  *  * humboldt /home/humbolt/crawl.py
30 8,16,24  *  *  * humboldt /home/humbolt/crawl1.py
00 13,23  *  *  * humboldt /home/humbolt/crawl2.py
00 9,17  *  *  * humboldt /home/humbolt/clockin.sh
17 *    * * *   root    cd / && run-parts --report /etc/cron.hourly
25 6    * * *   root    test -x /usr/sbin/anacron || ( cd / && run-parts --rep
47 6    * * 7   root    test -x /usr/sbin/anacron || ( cd / && run-parts --rep
52 6    1 * *   root    test -x /usr/sbin/anacron || ( cd / && run-parts --rep
#
```

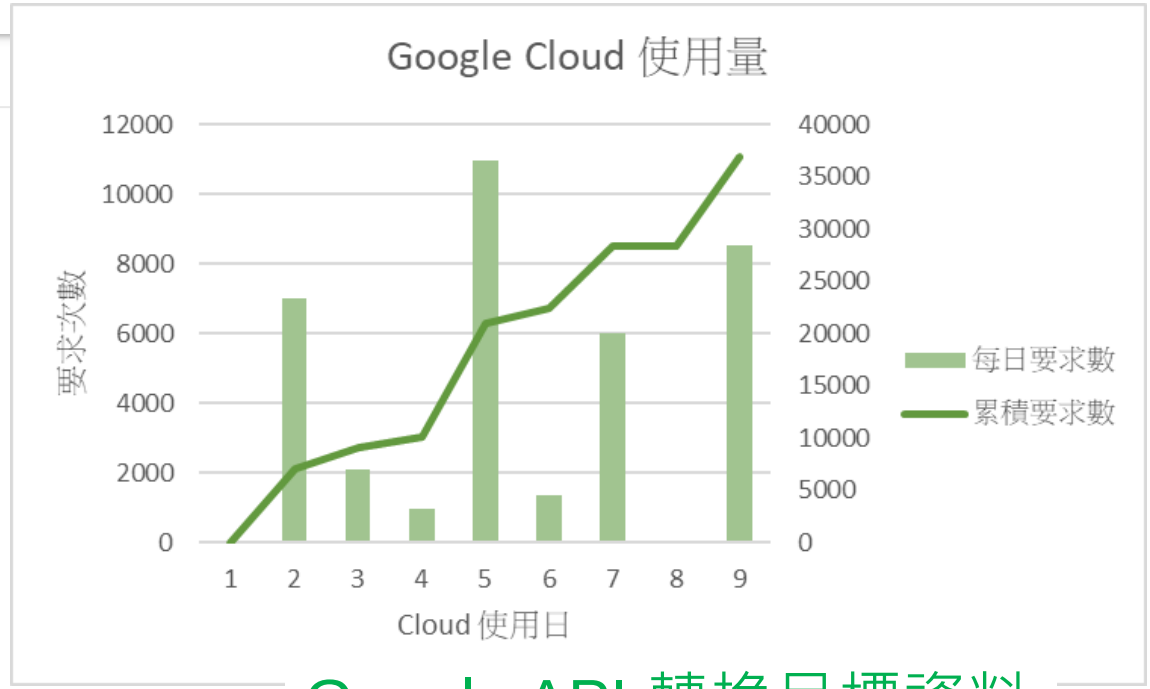
執行下載網頁取得特徵值。



Google Cloud

案例5. 使用Google Cloud

史晏臺



篩選 篩選條件

名稱	要求	錯誤百分比	延遲時間中位數 (毫秒)
Geocoding API	28,413	0	183
Places API	8,909	4	418
Compute Engine API	471	0	162

監控程式效率及是否傳輸正常。

Google API 轉換目標資料

轉換條件	轉換的資料
實際地址	經緯度
商店名稱	評論數
商店名稱	Google評級

主講人:許倫彬

5

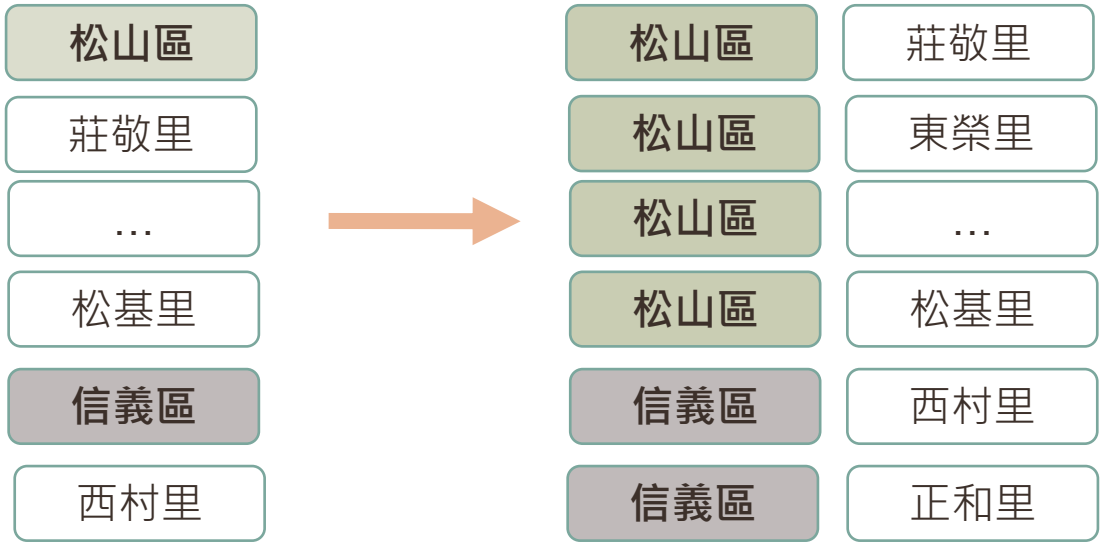
資料清理

1. 鄰里人口相關資訊
2. 地址資訊
3. 計算距離
4. 彙整資料表

資料清理：鄰里人口資訊

許倫彬

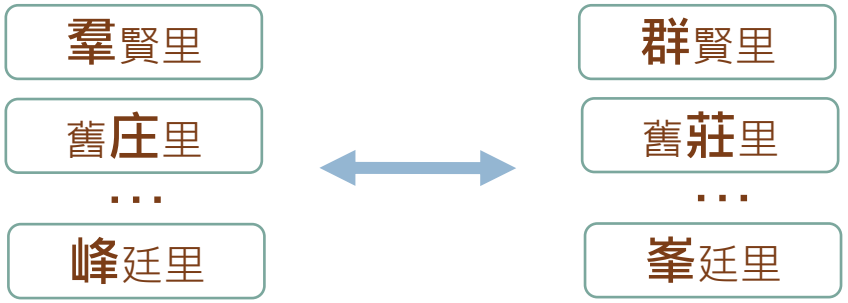
- 各里人口數
  - 區里資訊未在同一列



資料清理：鄰里人口資訊

許倫彬

- 計算里人口密度
  - QGIS：計算村里面積
  - 核對各區人口密度資料
- 合併年家戶所得資訊
  - 里名稱用字不同



資料清理：彙整地址資訊

許倫彬

整理各點地址

Google Geocoding API

各點區里資訊

經緯度轉換

GeoPy Nominatim

```
['cama cafe, 敦化南路二段, 義安里, 大安區, 六張犁,
臺北市, 10675, 臺灣']
```

正規化: \s[\u4e00-\u9fa5]{2}里,

資料清理：計算距離

許倫彬

GeoPandas

轉換資料型態
geometry

轉換
座標系統

計算距離

long	lat	geometry
121.549153	25.032824	POINT (121.549153 25.032824)

WGS84
TWD97

```
df.geometry.apply(
    lambda g:
        df2.distance(g))
```

資料清理：彙整資料表

許倫彬

各點對應的距離表

	cafe_id	point_id	point_type	distance
0	0	0	0	0.0
1	0	1	0	3270.0
2	0	2	0	2259.0
3	0	3	0	3049.0
4	0	4	0	3718.0
...	...	...	...	...

咖啡店： **1.5億** 筆
出租點： **2.9億** 筆

咖啡店/出租點的資訊表

area	brand	name	addr	filter	long	lat	density_2021	ave_pop_growth	2019_income
TPE	cama	台北敦南店	台北市敦化南路二段5號	大安區義安里	121.549153	25.032824	30537.582094	-1.038945	3844691.0
TPE	cama	民生松江店	台北市中山區民生東路二段133號	中山區中庄里	121.532388	25.058088	34199.348503	-1.879027	1857272.0
TPE	cama	台北長安店	台北市中山區長安東路二段94號	中山區朱園里	121.534597	25.048324	16690.778379	-0.500099	2769267.0
TPE	cama	台北長春店	台北市中山區長春路133-5號	中山區中吉里	121.531175	25.054950	30029.689819	-0.402234	2750941.0
TPE	cama	台北行天宮	台北市中山區民權東路二段97號	中山區新福里	121.532380	25.062712	37267.488548	-1.043902	1839588.0

資料清理：彙整資料表

許倫彬

300m距離下的彙整表

id	area	rating	comment	brand	name	addr	filter	long	lat	...	beverage	fastfood	supermarket	MRT
0	TPE	4.0	217.0	cama	台北敦南店	台北市敦化南路一段5號	大安區義安里	121.549153	25.032824	...	0.0	0.0	2.0	0.0
1	TPE	3.7	107.0	cama	民生松江店	台北市中山區民生東路二段133號	中山區中庄里	121.532388	25.058088	...	2.0	1.0	0.0	1.0
2	TPE	3.7	103.0	cama	台北長安店	台北市中山區長安東路二段94號	中山區朱圍里	121.534597	25.048324	...	2.0	0.0	2.0	0.0

主講人:陳冠儒

6 模型建置

1. 特徵值數據分析
2. 模型建置

我們想要模型做甚麼??

陳冠儒

city	filter	price	area	long	lat
TPE	大安區建倫里	280000	40	121.550889	25.040660
TPE	松山區吉祥里	200000	58	121.558054	25.049059
TPE	文山區萬隆里	45000	26	121.538167	25.001472
TPE	內湖區內湖里	450000	220	121.591466	25.082321
TPE	大同區光能里	188000	30	121.519051	25.053937
...	...	...	...	...	...

出租點資料

訓練好的模型

好結果與壞結果的標準是??

陳冠儒

營業收入

評論數

特徵值的選擇

陳冠儒

以咖啡店為中心方圓300公尺內的店家、交通站、學校、銀行數目

陳冠儒

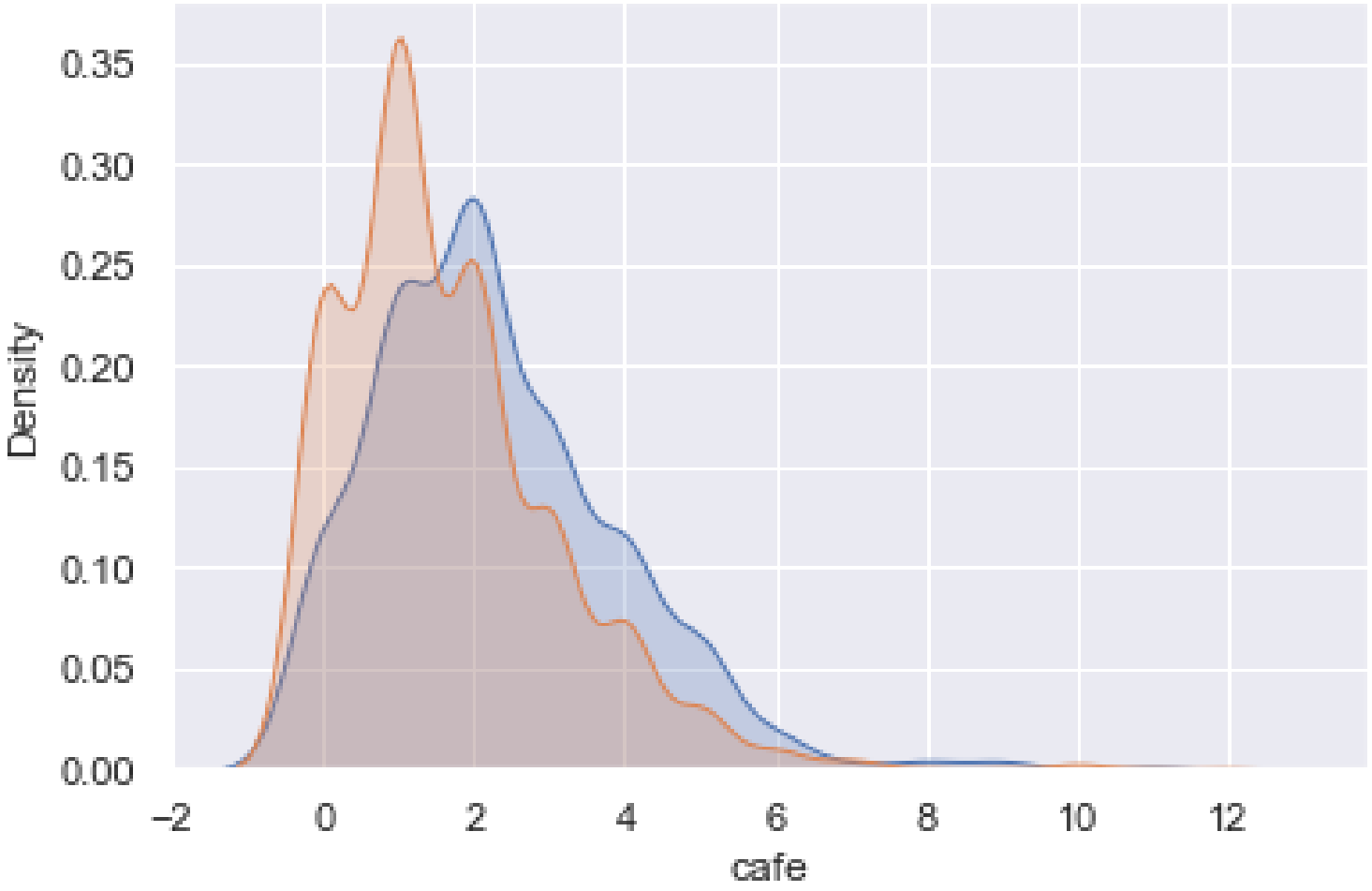
資料表

以平均值分成高評論數與低評論數

	density_2021	ave_pop_growth	2019_income	cafe	small_cafe	breakfast	beverage	fastfood	MRT	parking_space	CVS	bank	comment	comRank	result	
0	30537.582094	-1.038945	3844691	3	9	1	0	0	0	1	5	19	217	low_comment	**0**	0.0
1	16690.778379	-0.500099	2769267	3	7	1	2	0	0	1	7	13	103	low_comment	**1**	0.0
2	30029.689819	-0.402234	2750941	2	6	2	2	2	0	1	8	24	172	low_comment	**2**	0.0
3	37267.488548	-1.043902	1839588	3	9	1	0	1	0	0	10	5	205	low_comment	**3**	0.0
4	21085.129310	0.000600	3990206	3	8	0	0	1	0	0	8	14	234	low_comment	**4**	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	**...**	...
1813	31773.631713	-1.430861	3068587	0	1	0	0	0	0	0	0	0	960	high_comment	**1813**	1.0
1814	44539.769277	-0.616275	1559838	4	3	2	0	0	0	2	4	2	663	high_comment	**1814**	1.0
1815	33570.083333	-1.308009	2988501	2	10	0	2	0	1	0	8	12	432	high_comment	**1815**	1.0
1816	16248.877608	-0.256942	4560710	2	2	1	1	2	0	4	2	5	151	low_comment	**1816**	0.0
1817	56261.390813	-0.794574	1832720	2	6	4	3	2	1	1	6	6	191	low_comment	**1817**	0.0

特徵值數據分析：
300公尺內有多少連鎖咖啡廳

陳冠儒



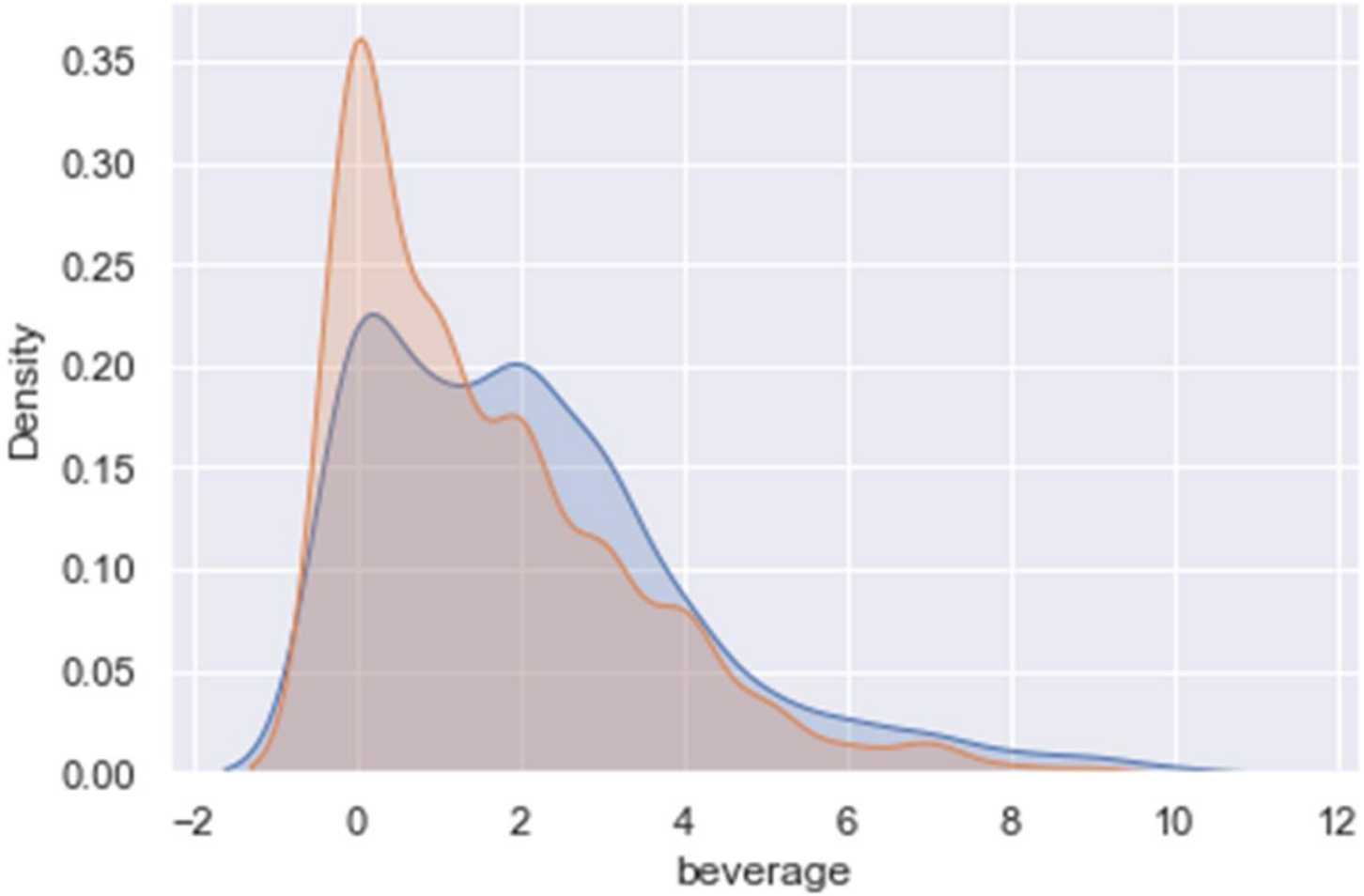
高評論數

低評論數

觀察結果：
高評論數的咖啡店
附近通常都有一間
以上的連鎖咖啡廳

特徵值數據分析：
300公尺內有多少手搖飲料店

陳冠儒



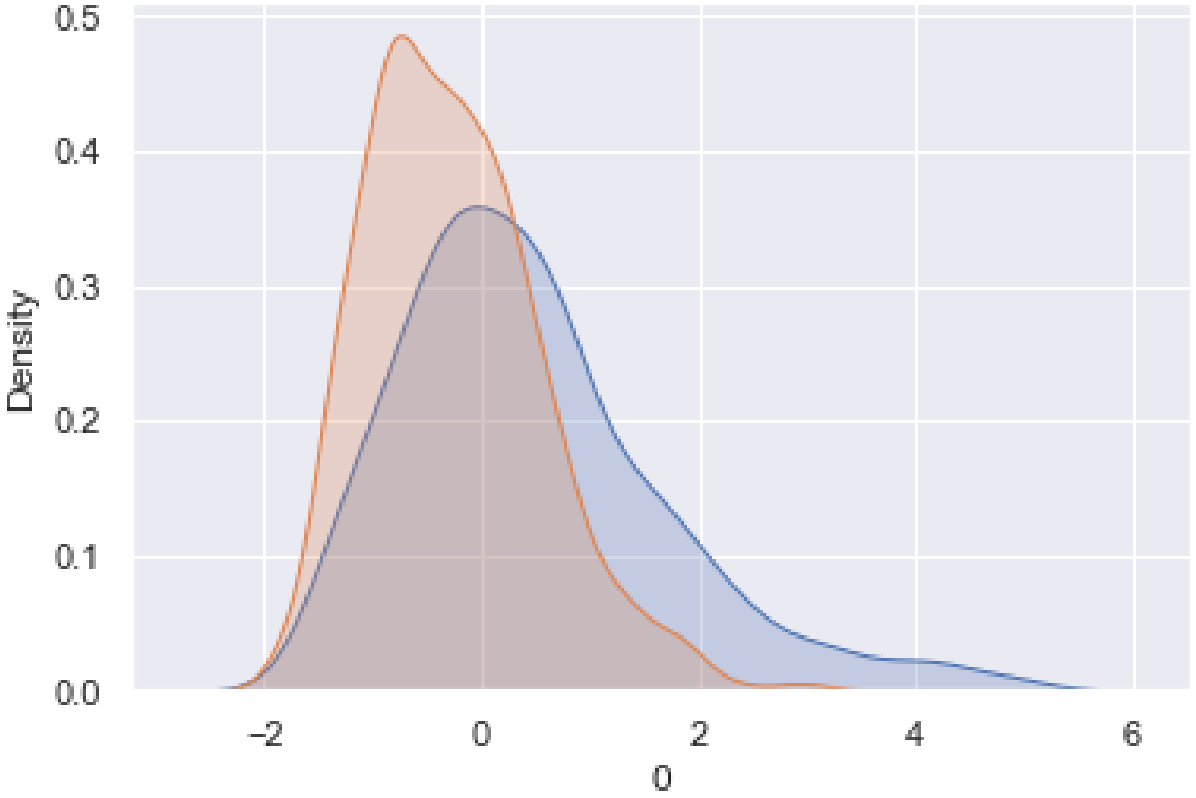
高評論數

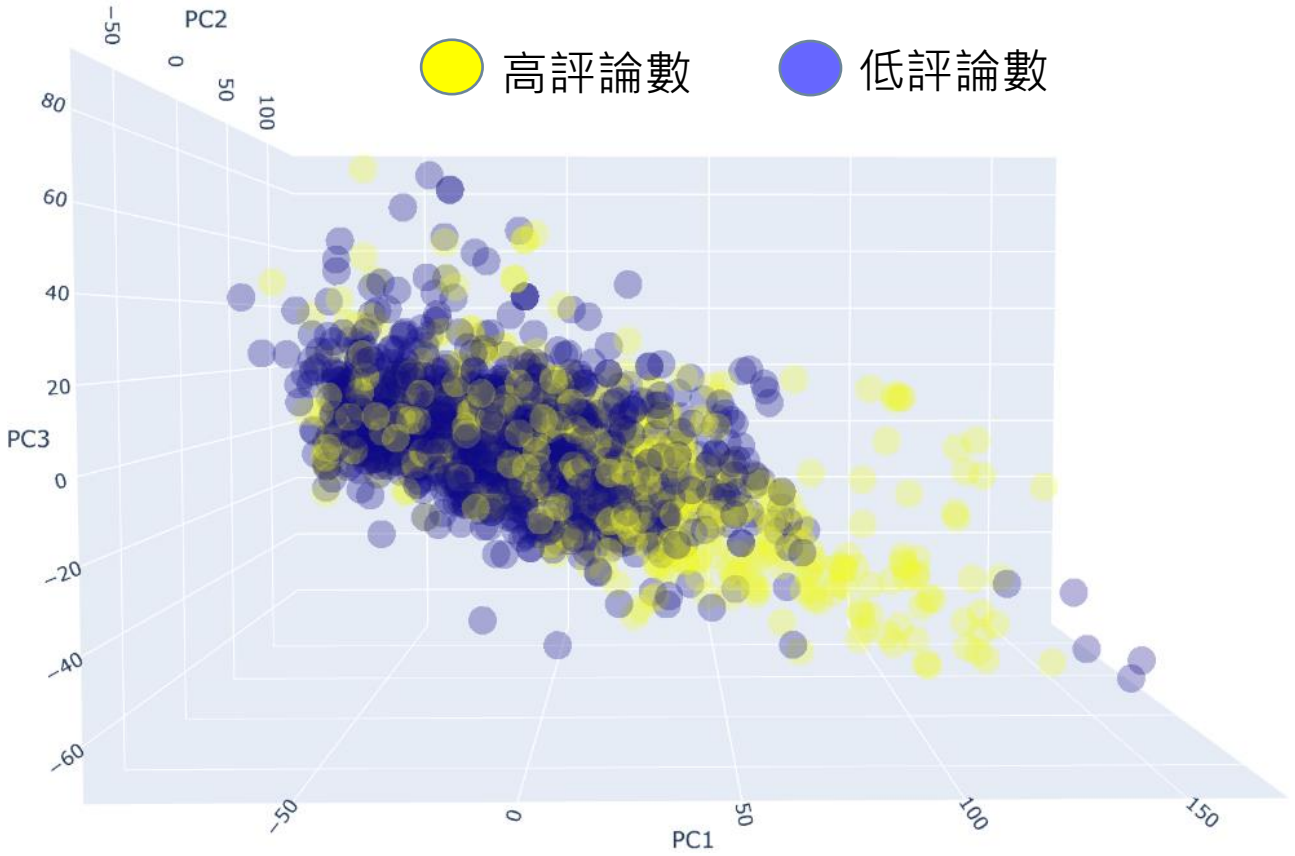
低評論數

觀察結果：
低評論數的咖啡廳
附近有很高的比例
沒有手搖飲料店

陳冠儒

觀察維度轉換後的效果 LDA vs. NCA



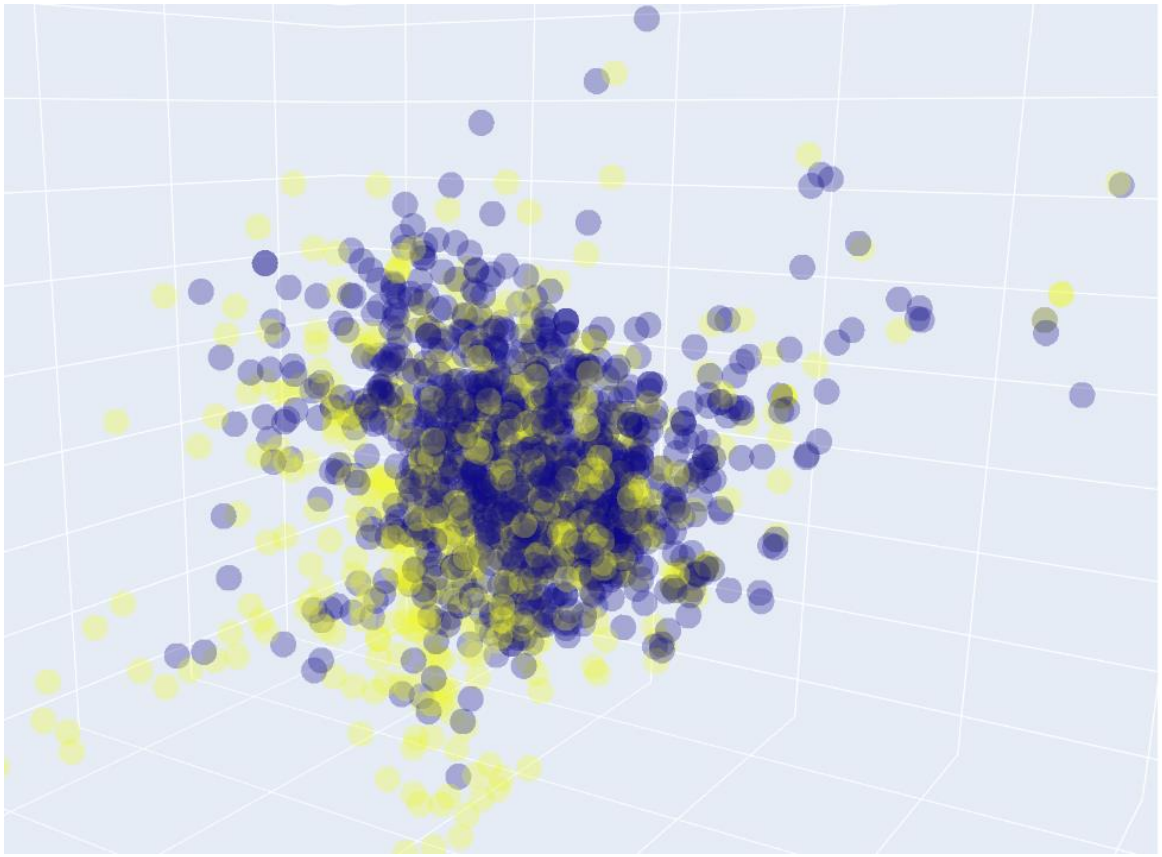


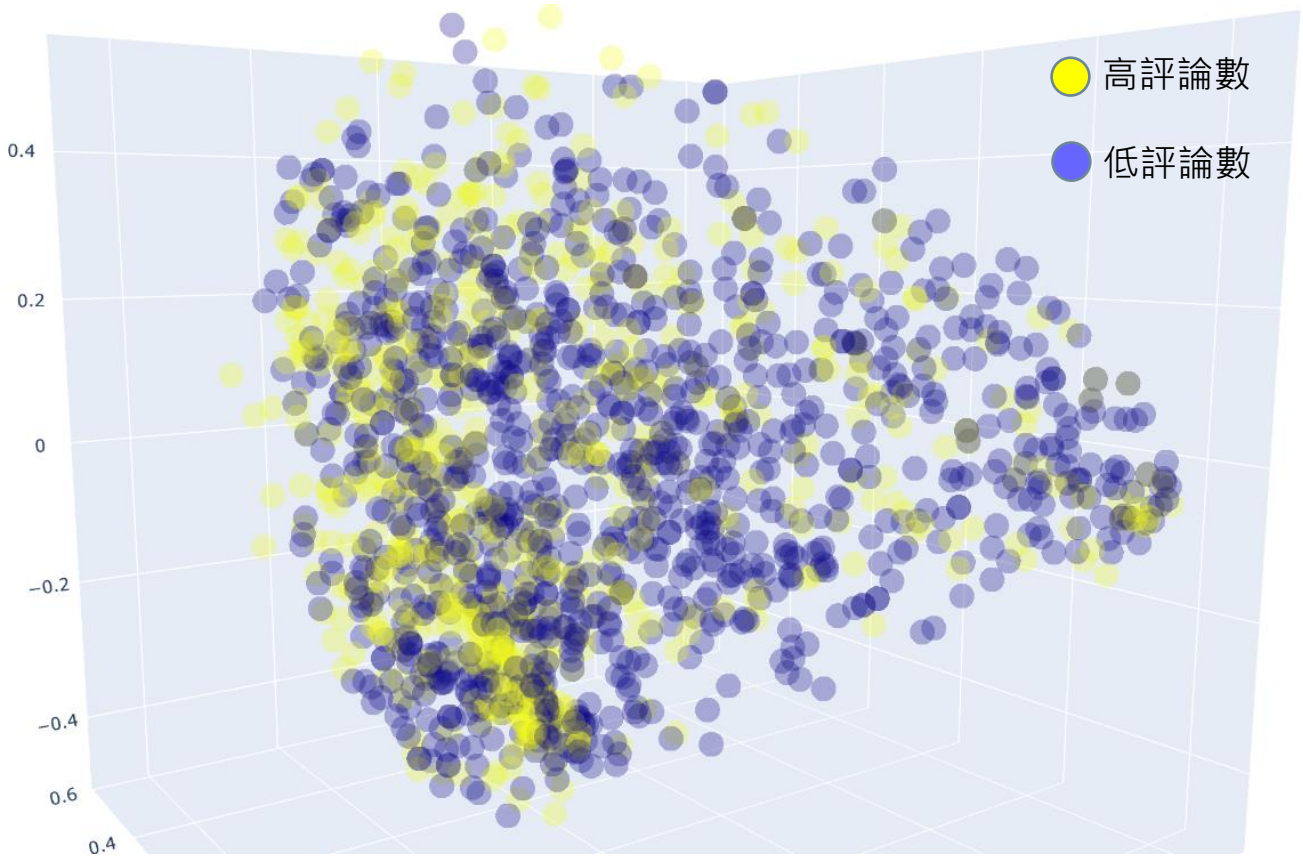
LDA(降到一維)

NCA(保留25%特徵值)

陳冠儒

觀察維度轉換後的效果 PCA vs. KPCA





PCA(保留48%特徵值)

KPCA(保留44%特徵值)

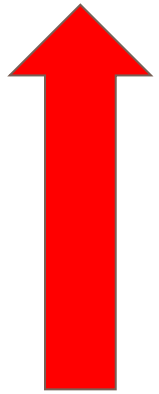
選擇模型評估指標

陳冠儒

考量因素：開店成本極高，且需要時間才能證明營收。

希望當模型預測是高評論的結果時，要有較高的準確度

$$\text{Precision} = \frac{\text{真●高評論數}}{\text{預測結果是高評論數時}}$$



模型選擇：
XGBoost vs. RandomForest

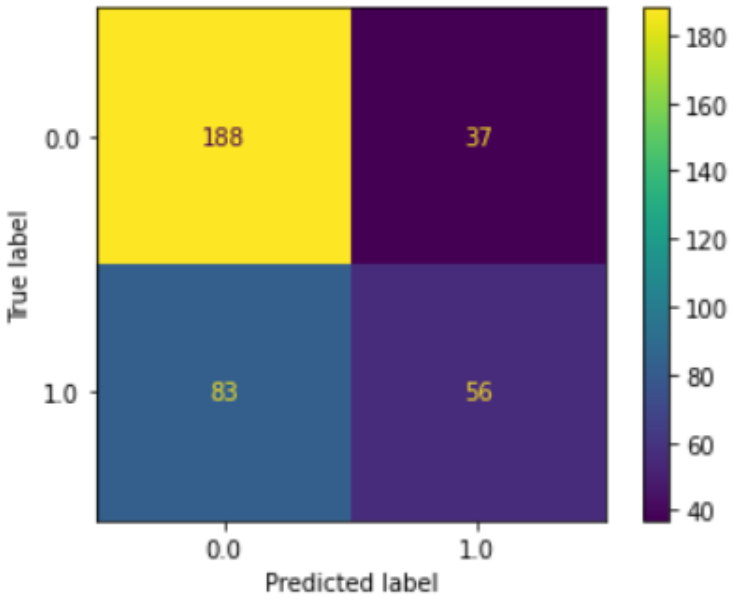

陳冠儒

XGBoost： Boosting 則是透過序列的方式生成樹，後面一顆與前面一顆相關。

RandomForest： Bagging透過隨機抽樣。

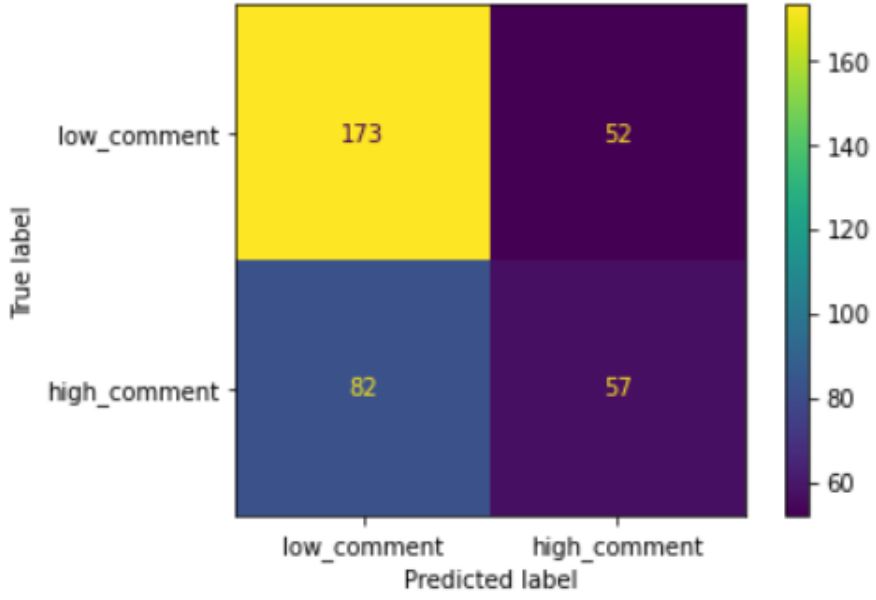
陳冠儒

XGBoost



	precision	recall	f1-score	support
0.0	0.69	0.84	0.76	225
1.0	0.60	0.40	0.48	139
accuracy			0.67	364
macro avg	0.65	0.62	0.62	364
weighted avg	0.66	0.67	0.65	364

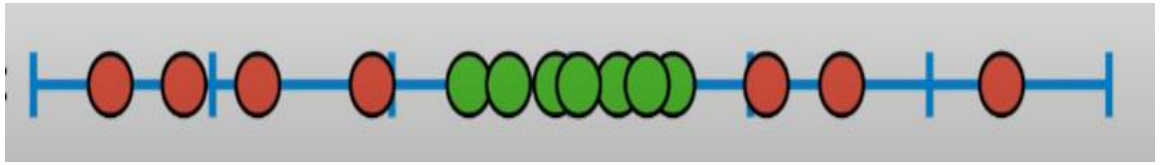
RandomForest

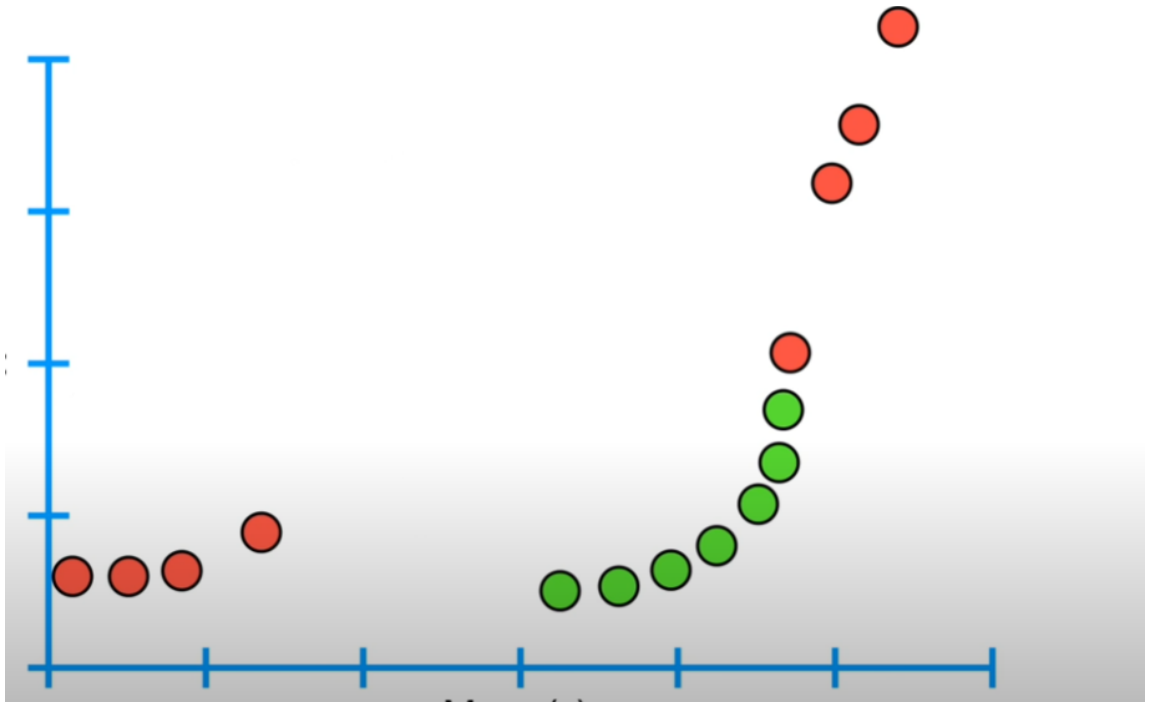


pre:	precision	recall	f1-score	support
0.0	0.68	0.77	0.72	225
1.0	0.52	0.41	0.46	139
accuracy			0.63	364
macro avg	0.60	0.59	0.59	364
weighted avg	0.62	0.63	0.62	364

模型選擇：SVM (Kernel: poly)

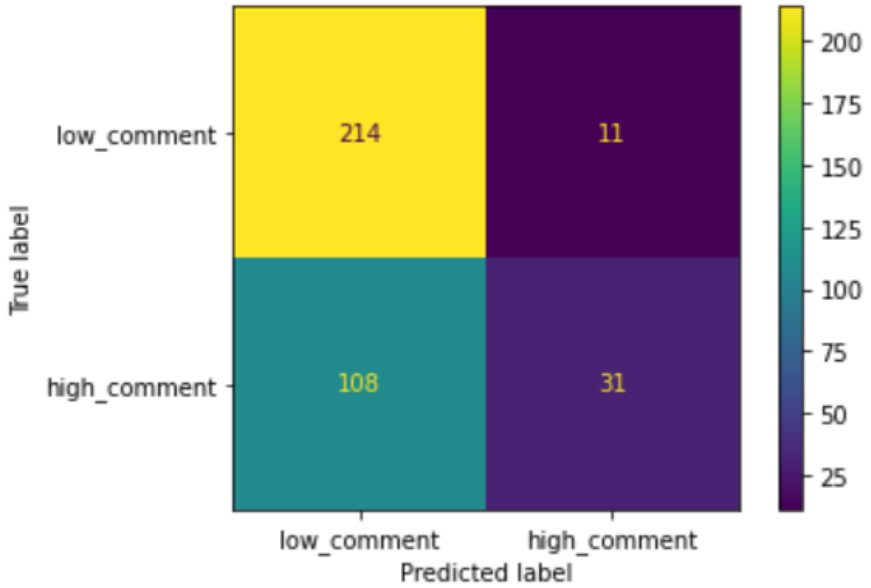
陳冠儒





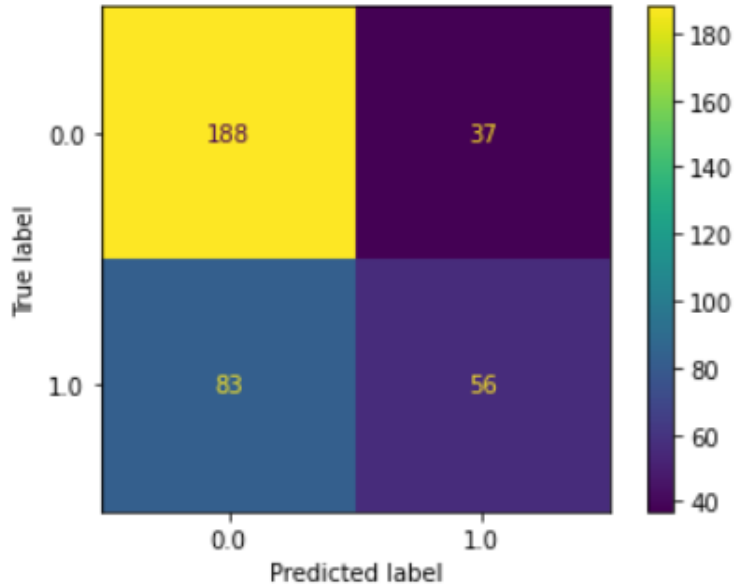
陳冠儒

SVM



	precision	recall	f1-score	support
0.0	0.66	0.95	0.78	225
1.0	0.74	0.22	0.34	139
accuracy			0.67	364
macro avg	0.70	0.59	0.56	364
weighted avg	0.69	0.67	0.61	364

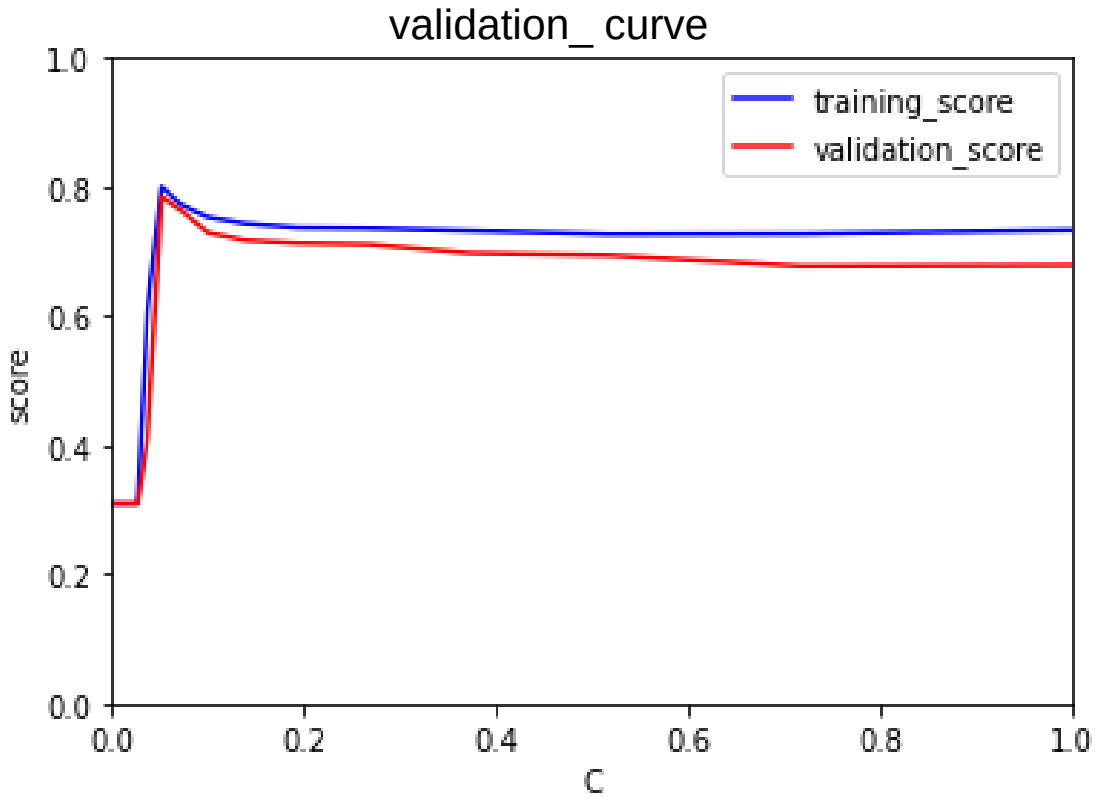
XGBoost

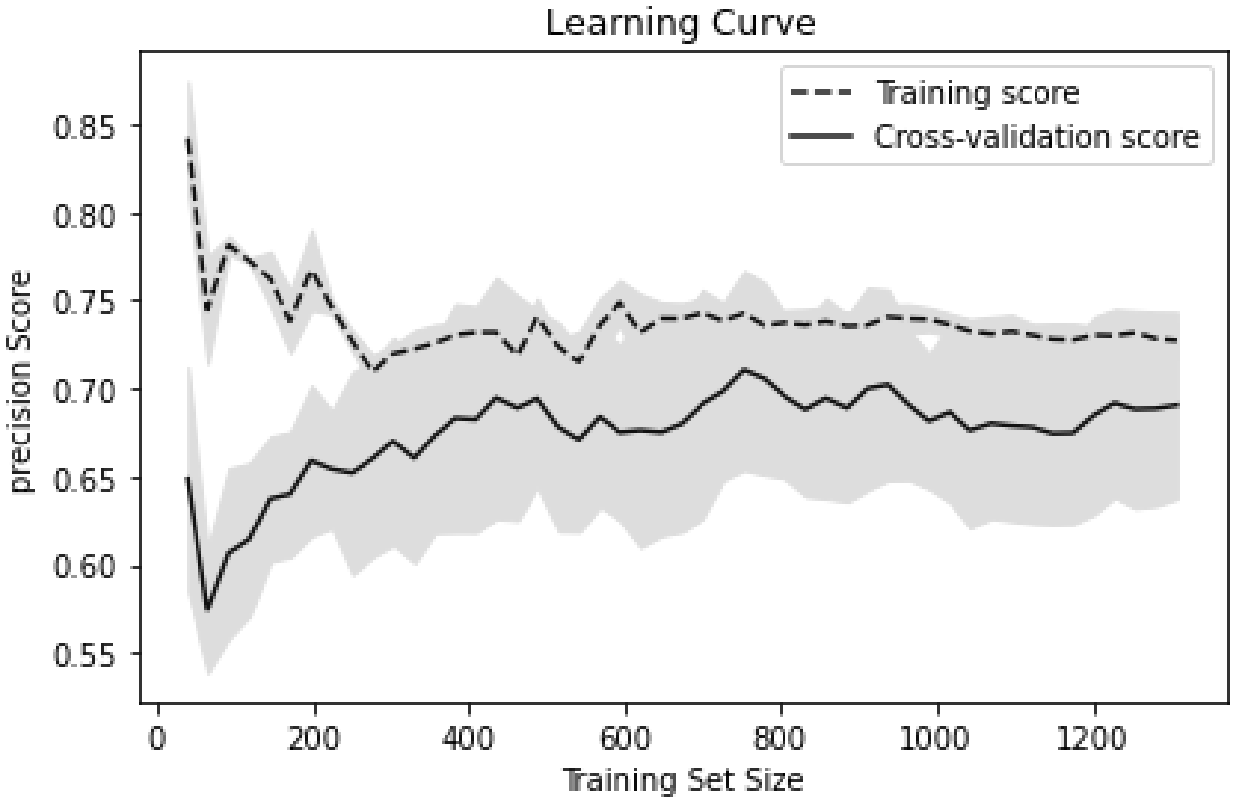


	precision	recall	f1-score	support
0.0	0.69	0.84	0.76	225
1.0	0.60	0.40	0.48	139
accuracy			0.67	364
macro avg	0.65	0.62	0.62	364
weighted avg	0.66	0.67	0.65	364

模型評估

陳冠儒

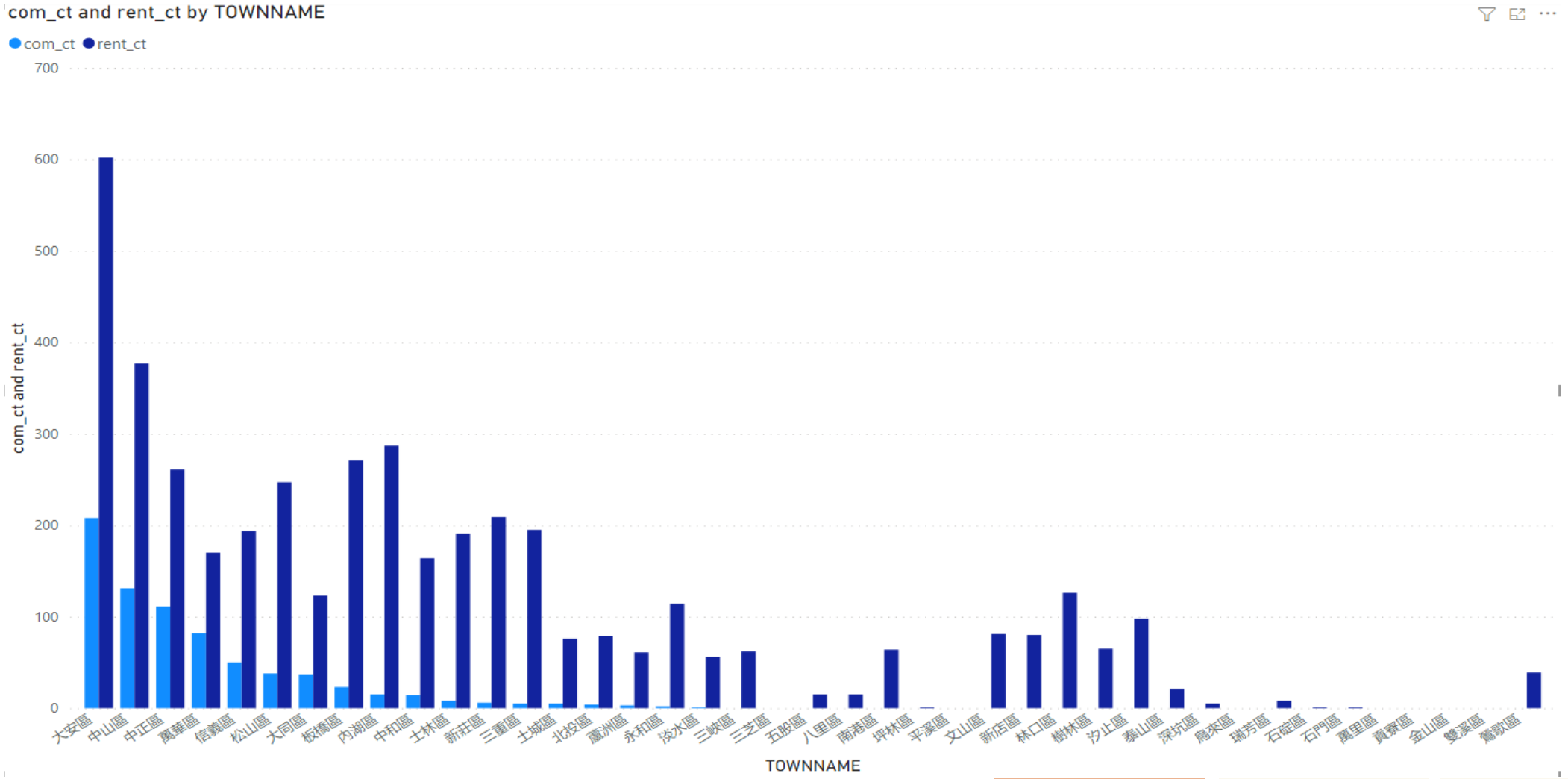




模型優化方向：更多的資料與更多的特徵值來增加高評論數的Precision

出租點：模型預測結果

陳冠儒



主講人:龔宣銘

7

成果視覺化

1. 使用工具
2. 網頁呈現

成果視覺化：使用工具

龔宣銘

前端

地圖應用

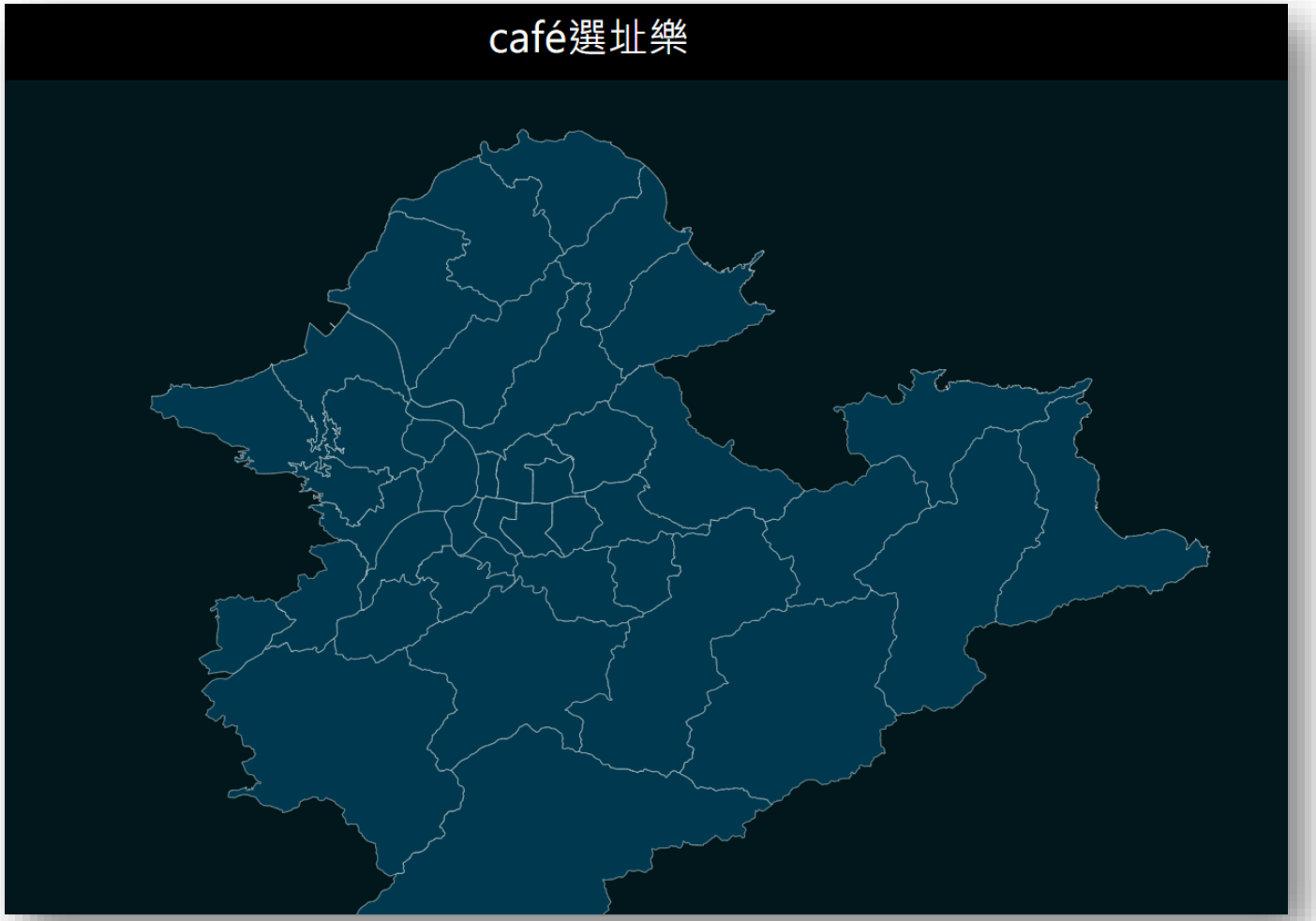
雲端佈署

後端

資料庫

成果視覺化：café選址樂網頁

龔宣銘



主講人:周旻柔

8 未來展望

未來展望

周旻柔

- 優化方向
  - 地區擴大: 由雙北擴大至全台
  - 增加特徵: 找到其他攸關的特徵，讓模型準確度更好
- 未來應用
  - 一般客戶：運用至其他業別的開店推薦，如超市、便利超商等
  - 客製化服務：結合公司內部資料，作為展店營收預測工具

café選址樂

快速找到最佳的開店地點