

動手「毒」理科 - STEM抗毒教育計劃

數據科學



香港青年協會
the hongkong federation of youth groups

創意教育組
Creative Education Unit



保安局禁毒處
求助／諮詢

186 186

98 186 186

Facebook

Instagram

Urban legend...



How Tesco Increased Beer Sales by Placing them Next to Diapers!



Deepak Kanakaraju

Co-founder & CEO at LearnToday.com | Founder at PixelTrack®

13 articles

[+ Follow](#)

February 13, 2015



什麼是數據？

- 數據是一組通過觀測得到的特徵或資訊。 *OECD Glossary of Statistical Terms*. OECD. 2008. p. 119. [ISBN 978-92-64-025561](#).

數值數據

Numerical data

Pizza的重量：
0-???(克)

辣肉腸的數量：
0-???(塊)



分類數據

Categorical data

Pizza的成份：
麵包、芝士、辣肉腸、香料

味道：
鹹、辣、甜、酸、苦

食客對食物滿意：
非常同意、同意、不同意、
非常不同意

收集數據(data)就是數據科學嗎？

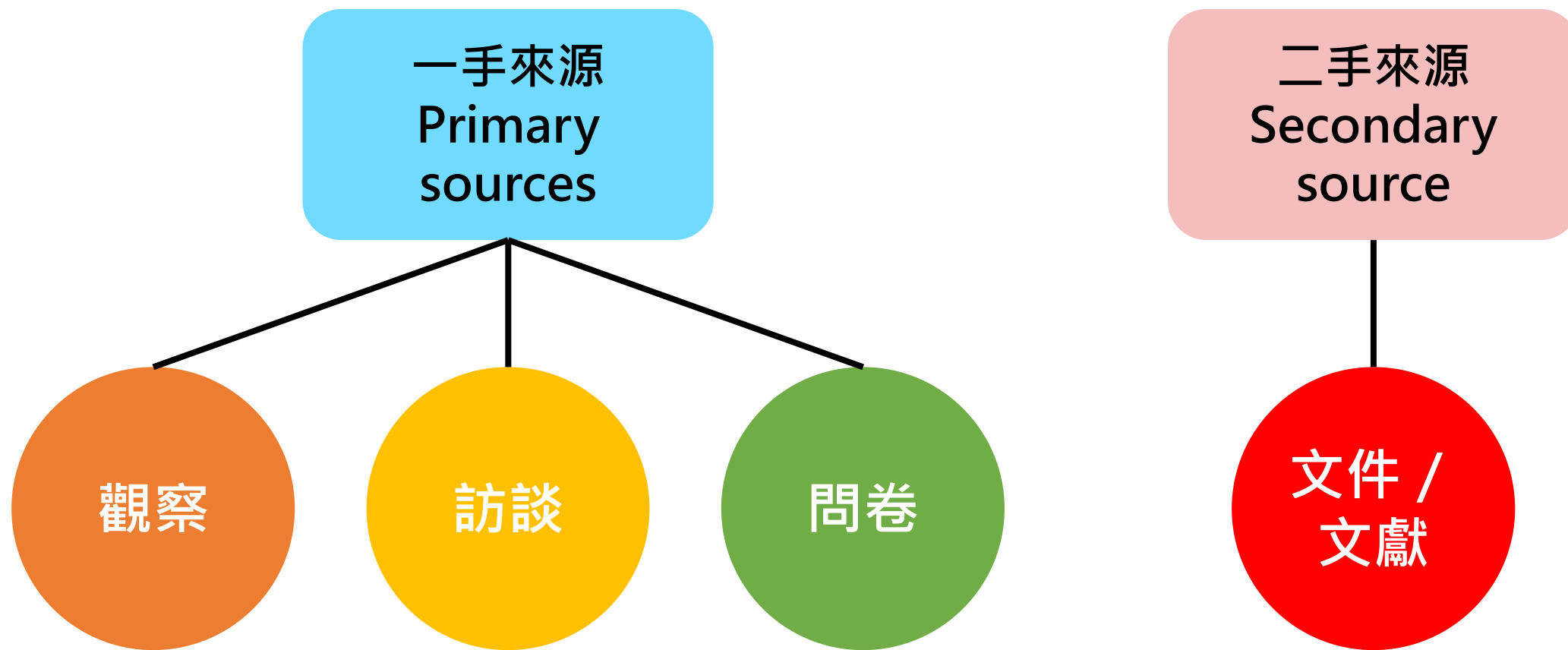
Data Science is...

the exploration and quantitative analysis of data to develop understanding, **extract knowledge and formulate actionable results**

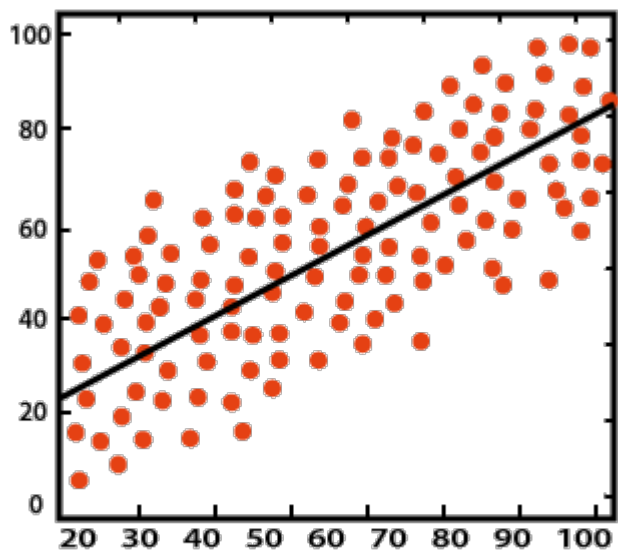
數據科學是...

藉著探索和分析數據，**挖掘出有用的資訊，並作出可行動的決策**

收集數據

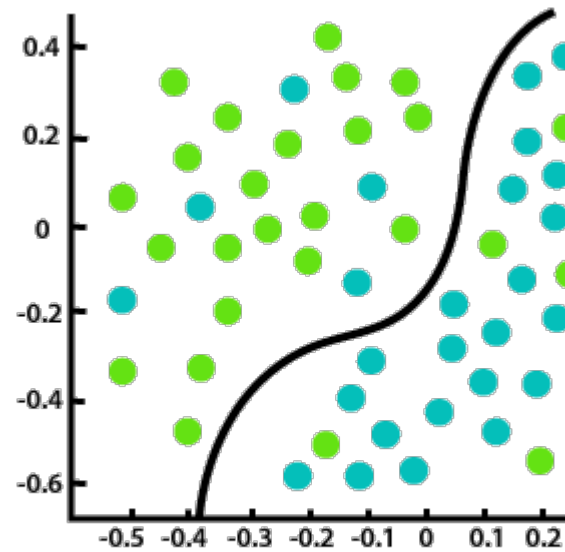


數據科學的兩大運用



Regression

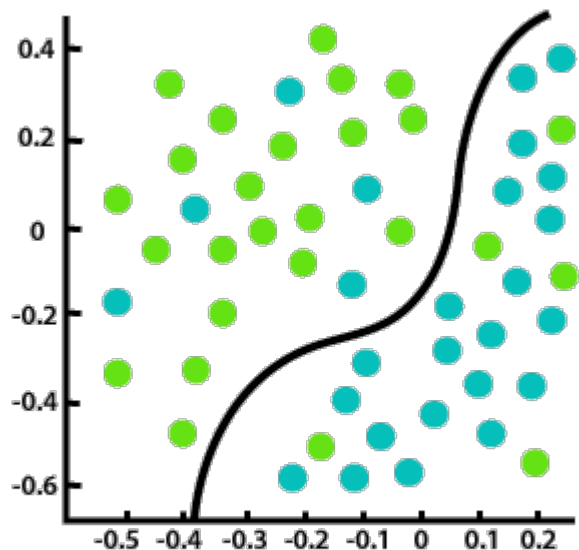
回歸



Classification

分類

數據科學的兩大運用



Classification

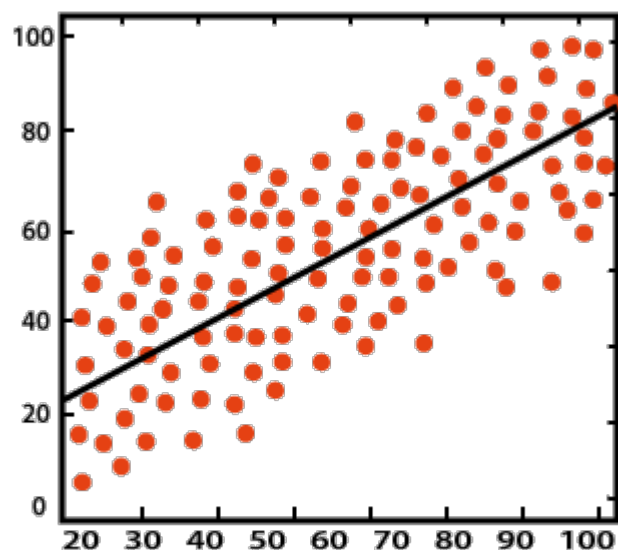
解決分類問題

- 將數據以不同參數(parameter)分類
- 類別的數量是預設的
- **答案是一個類別**
- 例如: 根據顧客的背景, 估計他對某產品是否感興趣

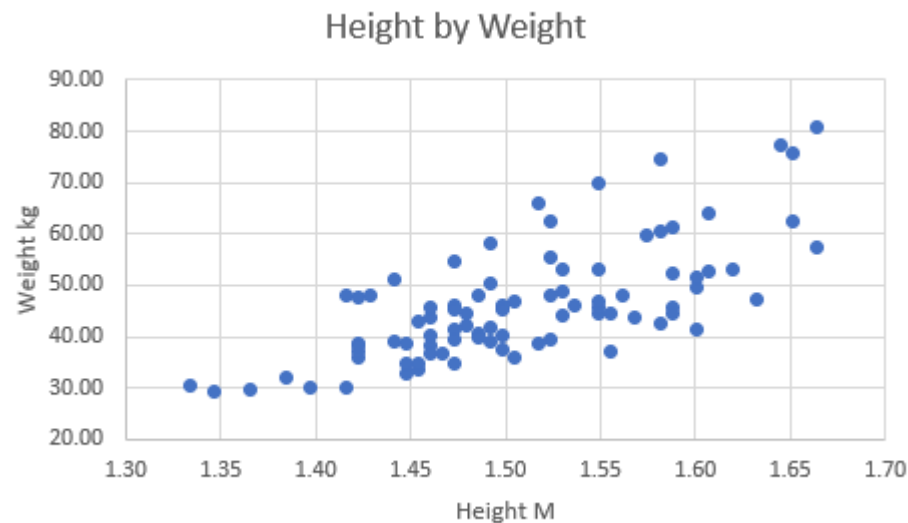
數據科學的兩大運用

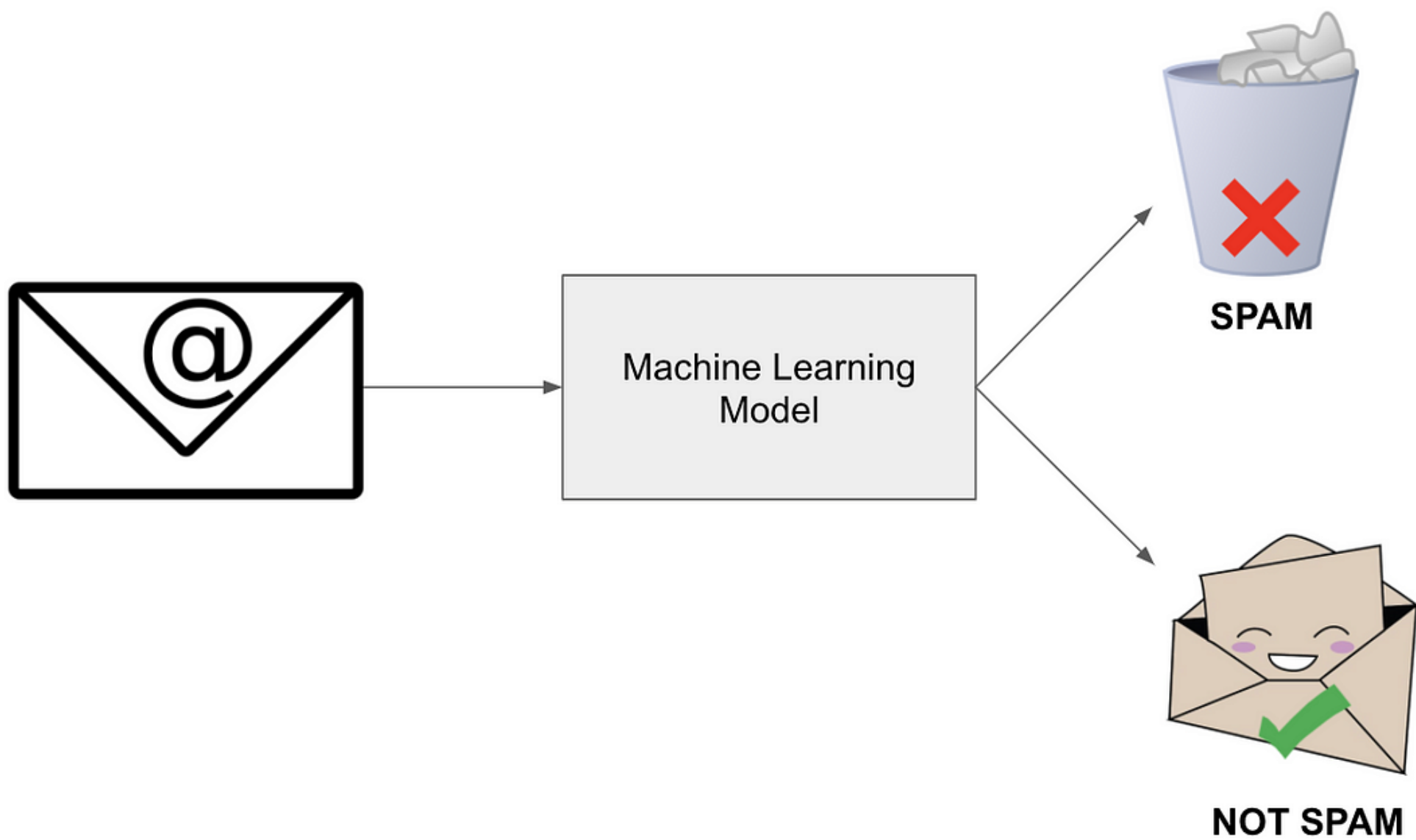
解決回歸問題

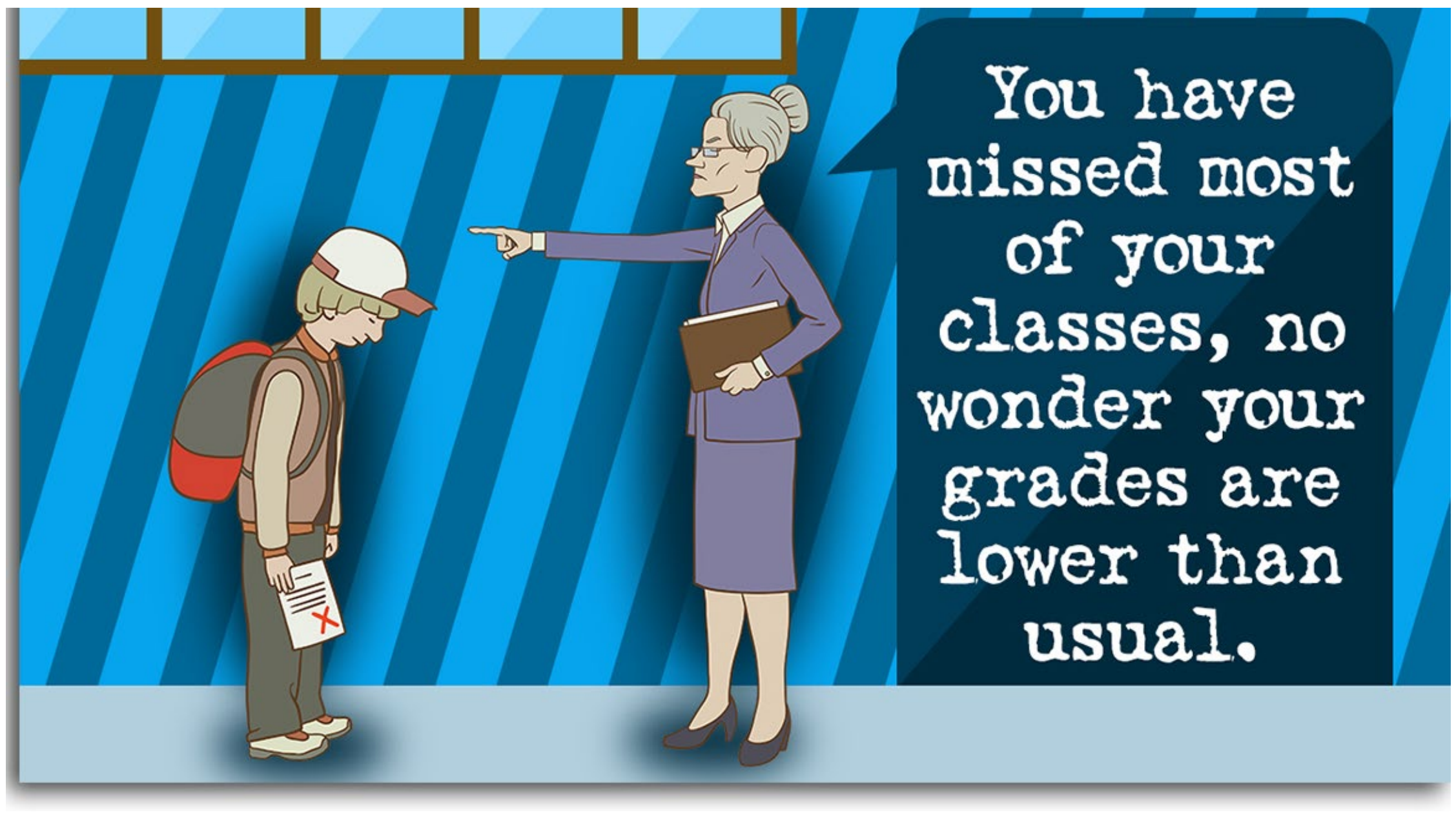
- 找出數據之間的關聯性
- 答案是一個連續數據



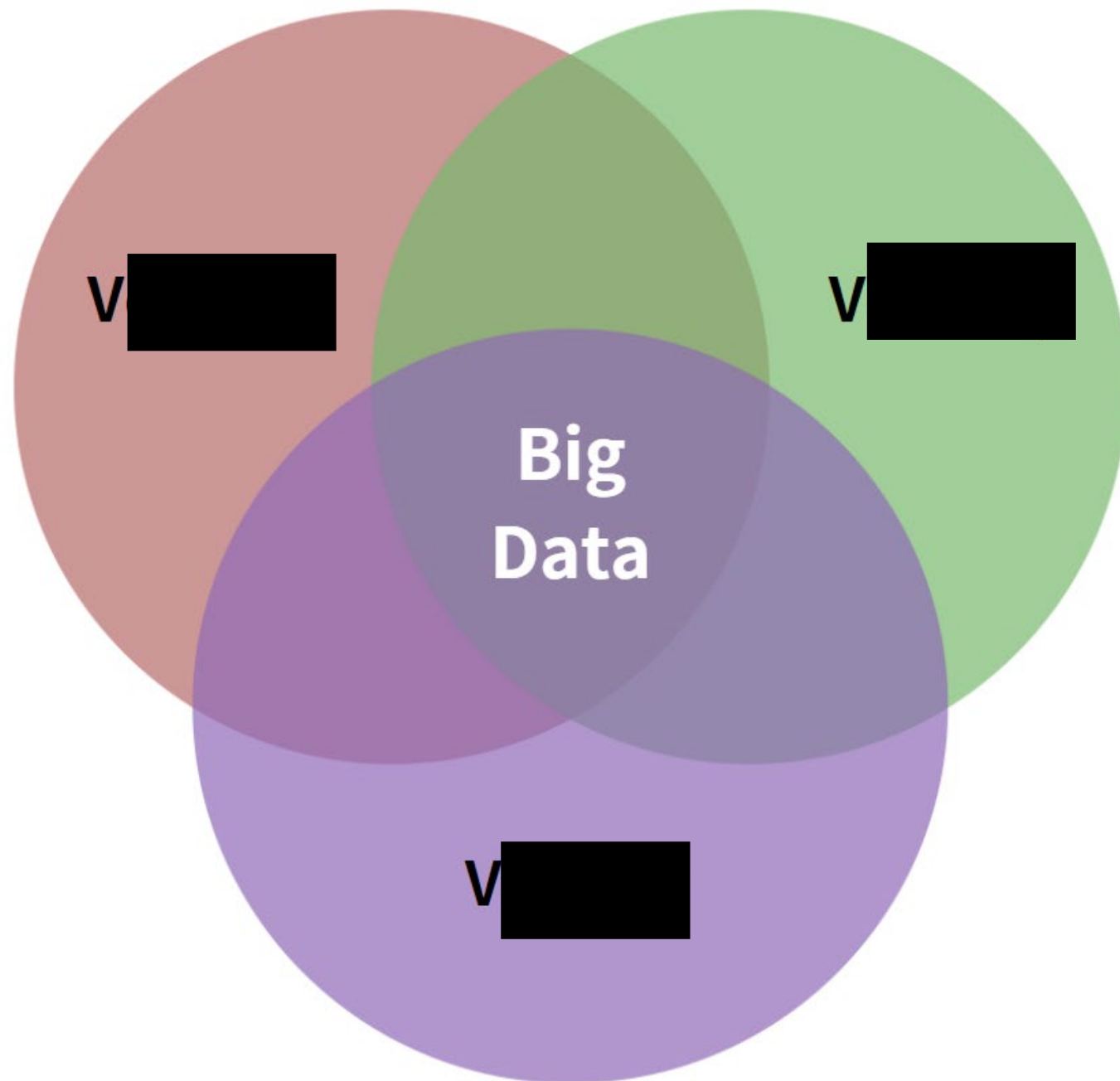
Regression







You have missed most of your classes, no wonder your grades are lower than usual.



數據量的重要



4.5 out of 5 stars



4 reviews

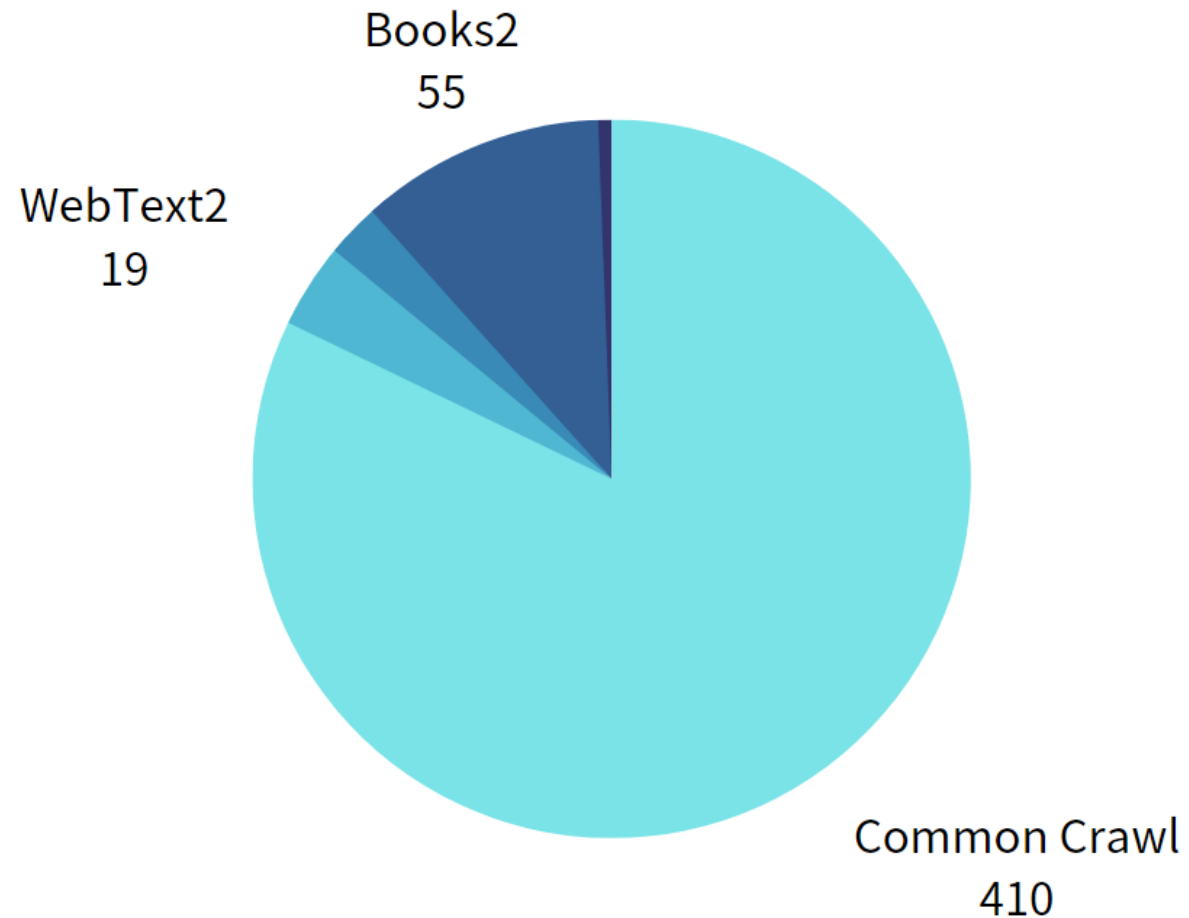


4.2 out of 5 stars

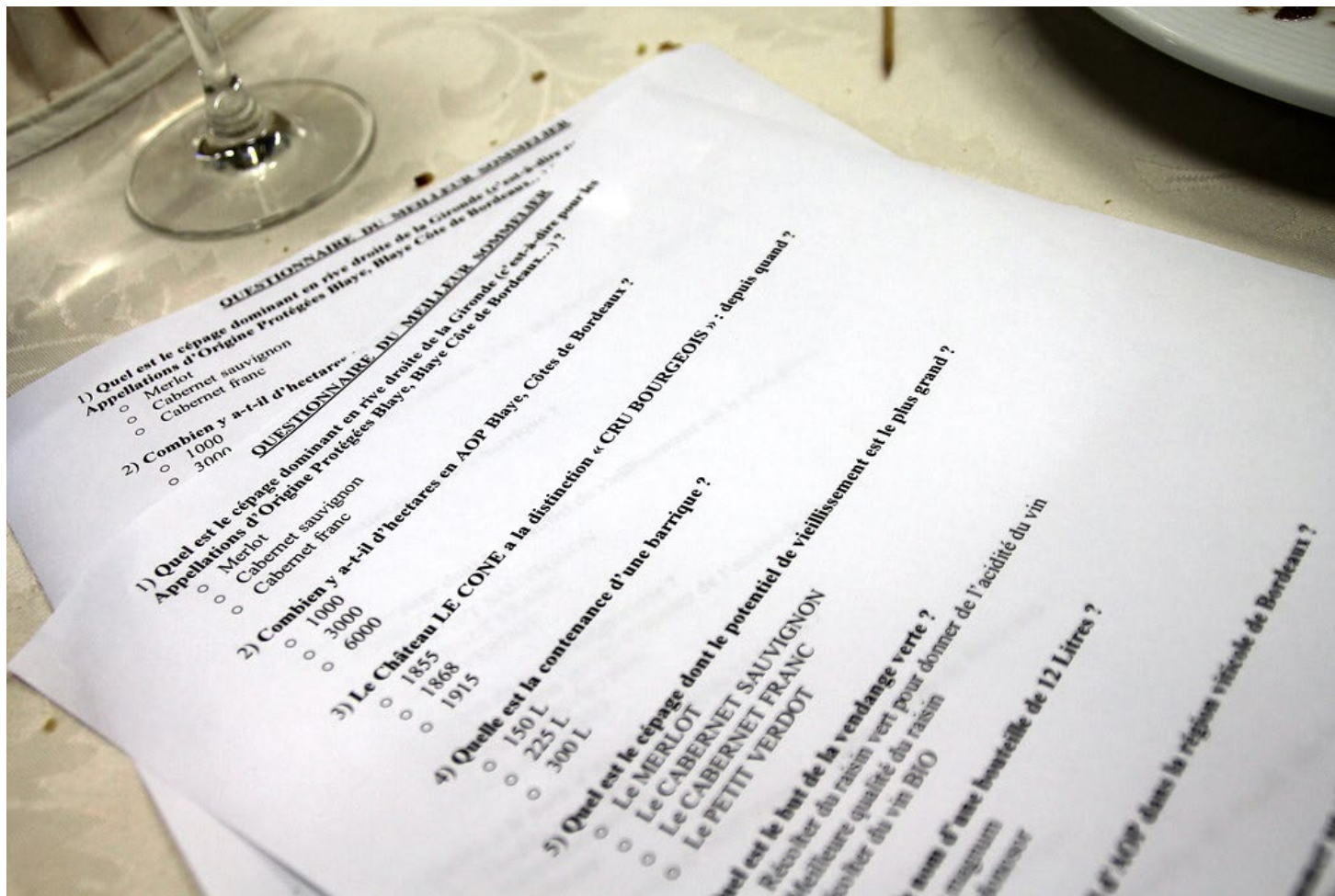


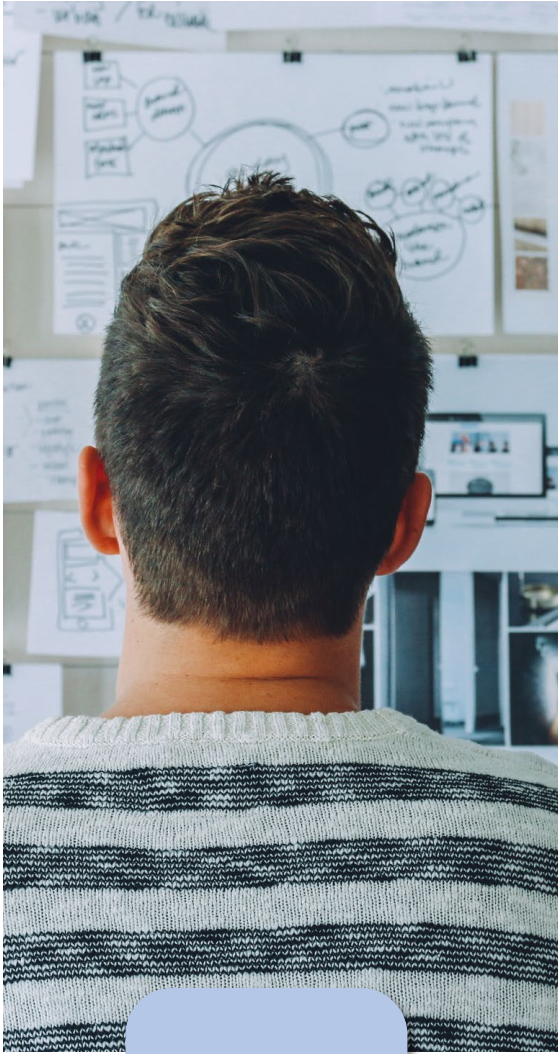
2647 reviews

OpenAI Chat-GPT (GPT-3)

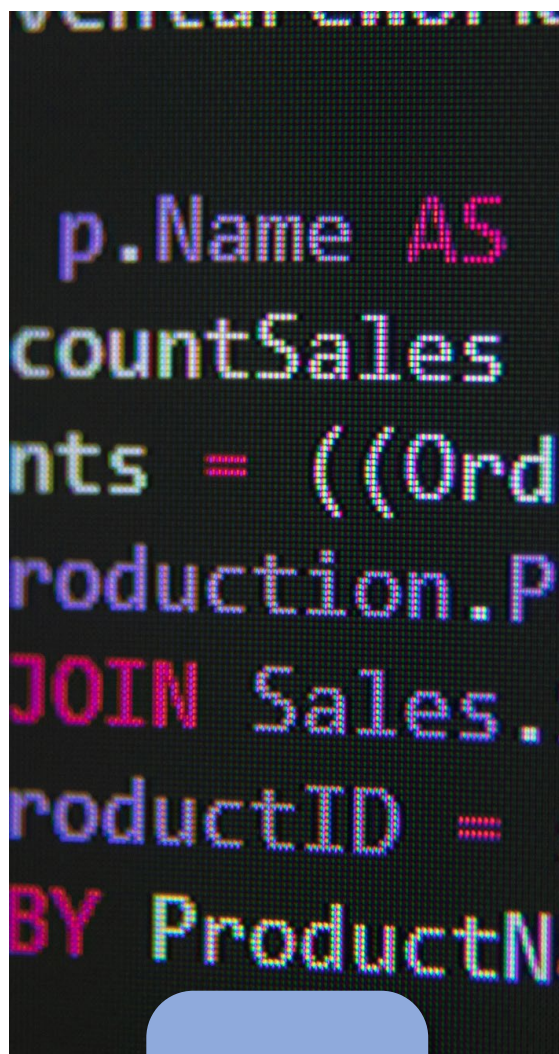


number of tokens (each text correspond to 1-4 tokens) **(in billion)**

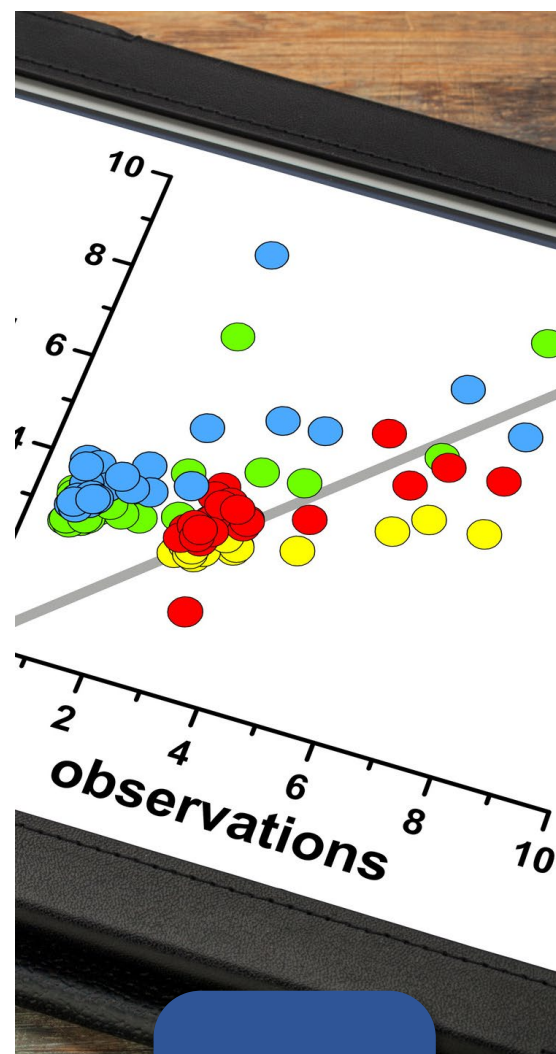




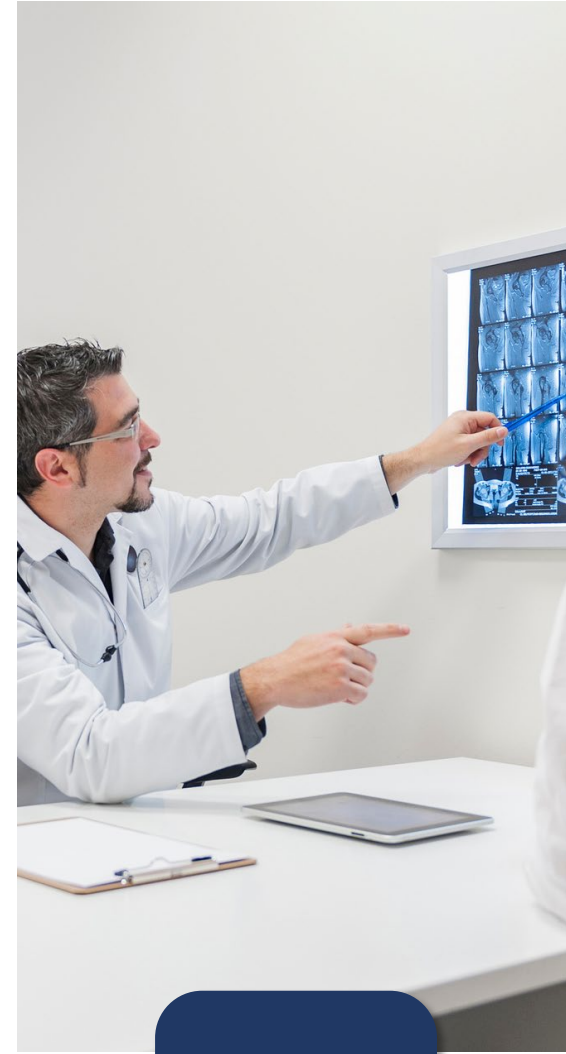
計劃



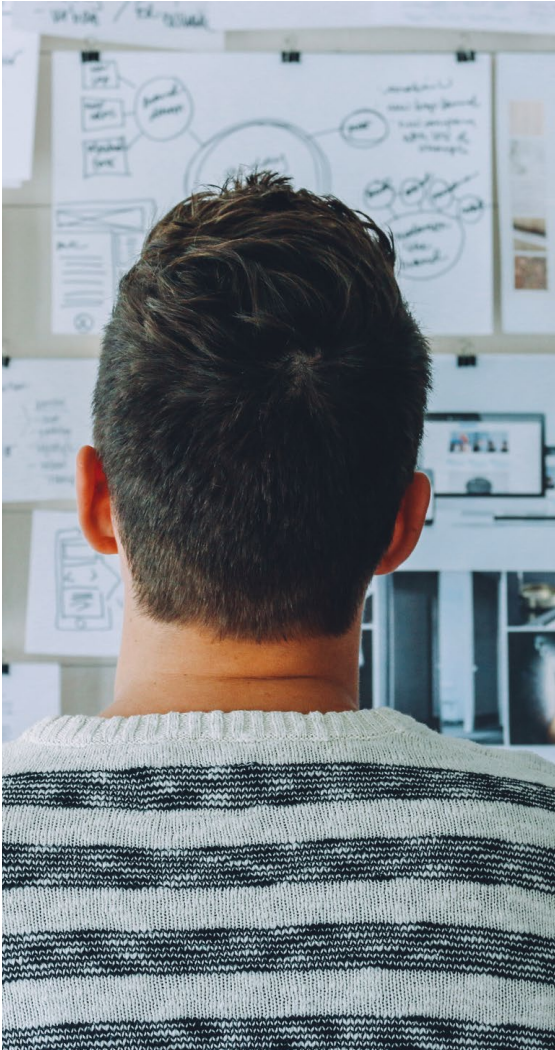
準備數據



模型化



推出及跟進



Planning 計劃

- 訂立計劃目標！
- 準備相關資料及資源



網上平台對吸毒數字的影響

上述分析是偵測本港由2016年至今年6月30日，逾30萬篇毒品相關的帖文及留言。研究發現，近年社交媒體涉及毒品相關的帖文數目及瀏覽量大
幅上升。帖文數目由2016年的1,001篇，上升至今年中的5,237篇，上升比率為423%；估算瀏覽量亦由2016年的440萬次，上升至今年中的
1,734萬次，上升比率為298%。最多人討論及持正面態度的毒品首四位是大麻、啟靈藥（LSD）、迷姦水（GHB）和白瓜子。

Artificial Intelligence Can Now Use Instagram to Predict Substance Abuse Risk

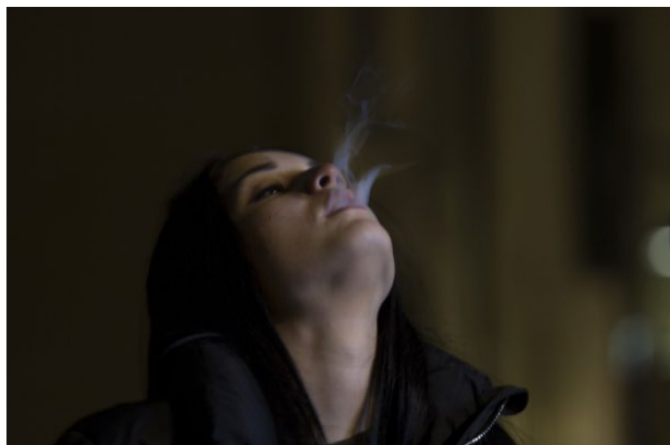
A paper published recently developed a method of predicting someone's risk of substance abuse based on their Instagram accounts ([link](#)). Let's take a look at how they did it.



Jerry Wei

Follow

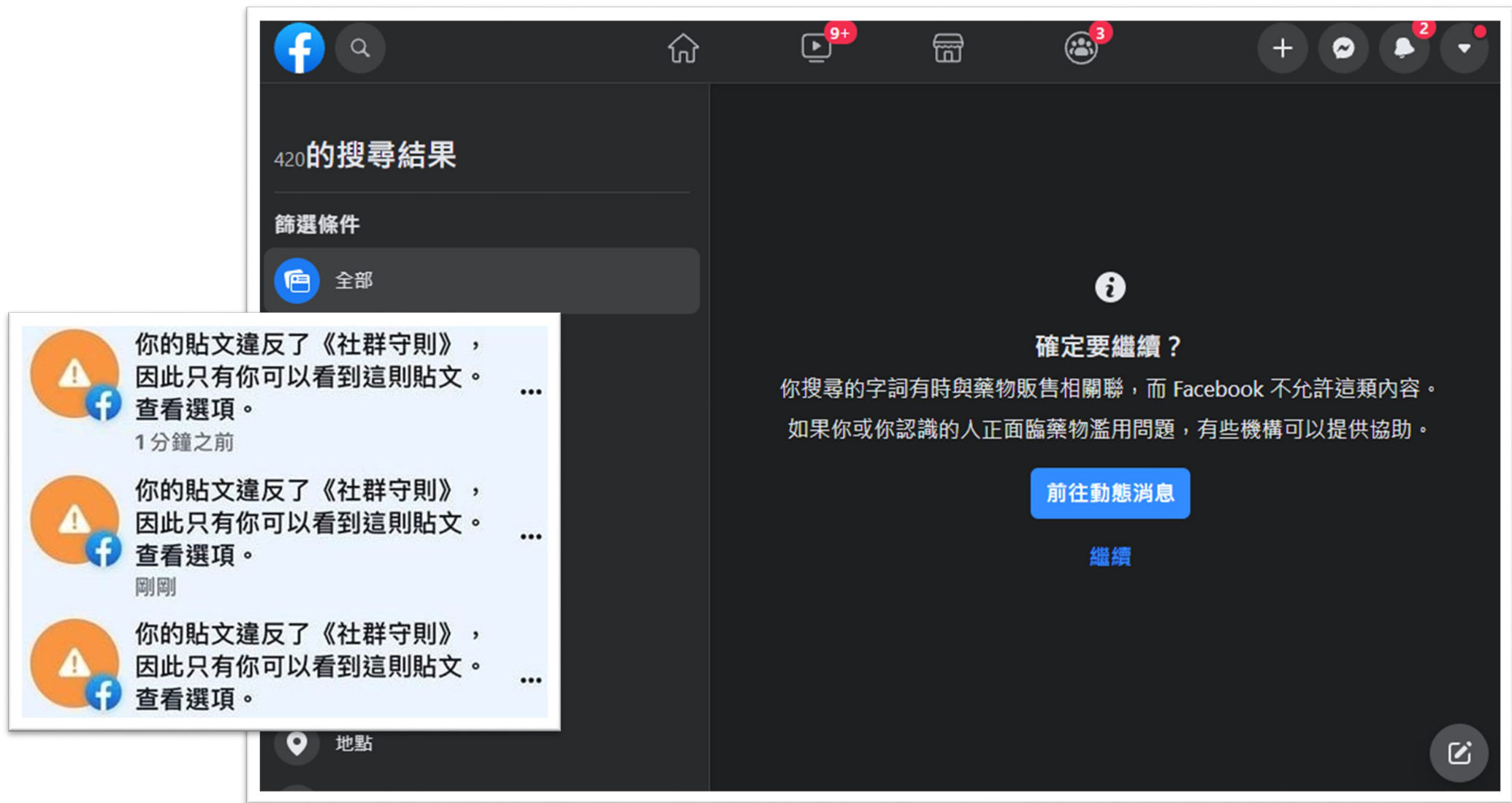
Jun 26, 2019 · 4 min read



Using Artificial Intelligence to Change the Future of Drug Addiction Treatment & Narcotics Detection

Mangalore Today News Network

By Dr. G. Shreekumar Menon



為社交平台設計一個大麻監察的系統

需收集什麼資料？準備什麼資源？

An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning

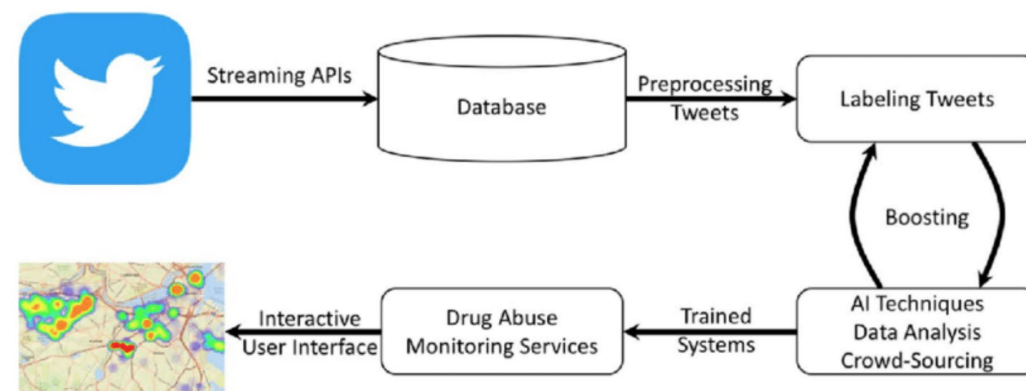
Han Hu, [NhatHai Phan](#) ✉, [Soon A. Chun](#), [James Geller](#), [Huy Vo](#), [Xinyue Ye](#), [Ruoming Jin](#), [Kele Ding](#), [Deric Kenne](#) & [Dejing Dou](#)

[Computational Social Networks](#) 6, Article number: 10 (2019) | [Cite this article](#)

5648 Accesses | 13 Citations | 1 Altmetric | [Metrics](#)

Abstract

Drug abuse continues to accelerate towards becoming the most severe public health problem in the United States. The ability to detect drug-abuse risk behavior at a population scale, such as among the population of Twitter users, can help us to monitor the trend of drug-abuse incidents. Unfortunately, traditional methods do not effectively detect drug-abuse risk behavior, given tweets. This is because: (1) tweets usually are noisy and sparse and (2) the availability of labeled data is limited. To address these challenging problems, we propose a deep self-taught learning system to detect and monitor drug-abuse risk behaviors in the Twitter sphere, by leveraging a large amount of unlabeled data. Our models automatically augment annotated data: (i) to improve the classification performance and (ii) to capture the evolving picture of drug abuse on online social media. Our extensive experiments have been conducted on three million drug-abuse-related tweets with geo-location information. Results show that our approach is highly effective in detecting drug-abuse risk behaviors.



Hu, Han & Phan, Nhat Hai & Chun, Soon & Geller, James & Vo, Huy & ye, Xinyue & Jin, Ruoming & Ding, Kele & Kenne, Deric & Dou, Dejing. (2019). An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning. *Computational Social Networks*. 6. 10.1186/s40649-019-0071-4.

社交平台上有什麼數據？

貼文內容
(文字及emojis)

圖片上的文字

反應 Reaction
(如讚好、嬬嬬)

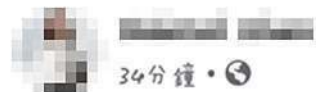
發貼/留言/反應用戶

圖片

留言



圖片的文字



34分鐘 •

好東西需要的懂的加賴

貼文

圖片內容



```
SELECT p.Name AS  
countSales  
nts = ((Ord  
roduction.P  
JOIN Sales.  
roductID =  
BY ProductN
```

Data Preparation 準備數據

- 取得數據
- 資料前處理(Data Preprocessing)

Data collection/ mining

3.1 Mining and Filtering Twitter through Livestream

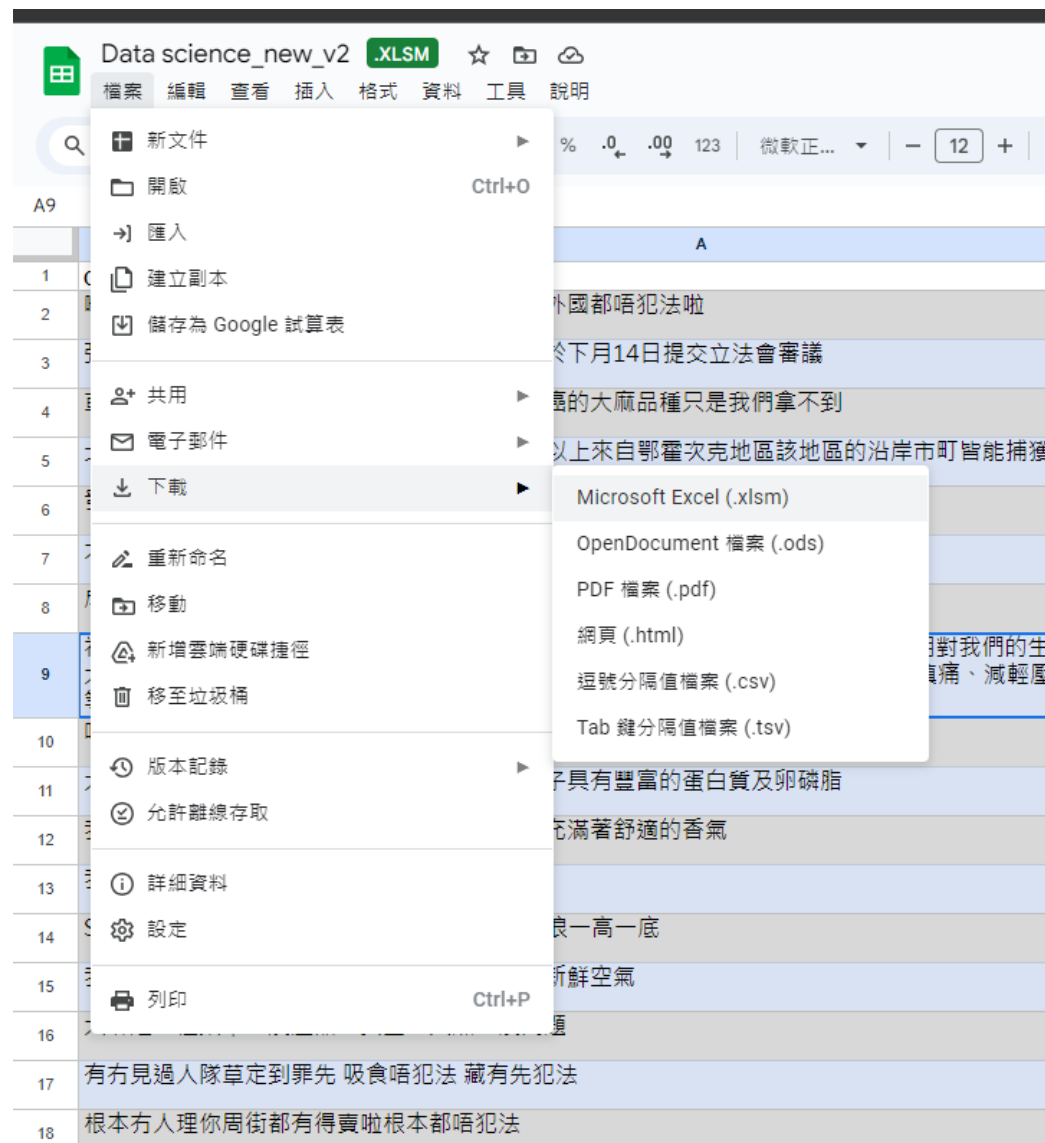
In this first step, tweets are mined from Twitter through a livestream. Tweepy [13], a Python library, was used to access the Twitter streaming API. The stream was mine for 9 days in December for a total of 486,689 tweets. The tweets were then stored in a csv file for bulk processing. The stream was filtered for keywords containing drug names taken from the top most popular drugs from the drugs website [14], totaling 462 different drug names. The drug names from the website were names that were most commonly used instead of the drug's actual scientific name.

- 從社交平台收集包含「草」或「大麻」的留言

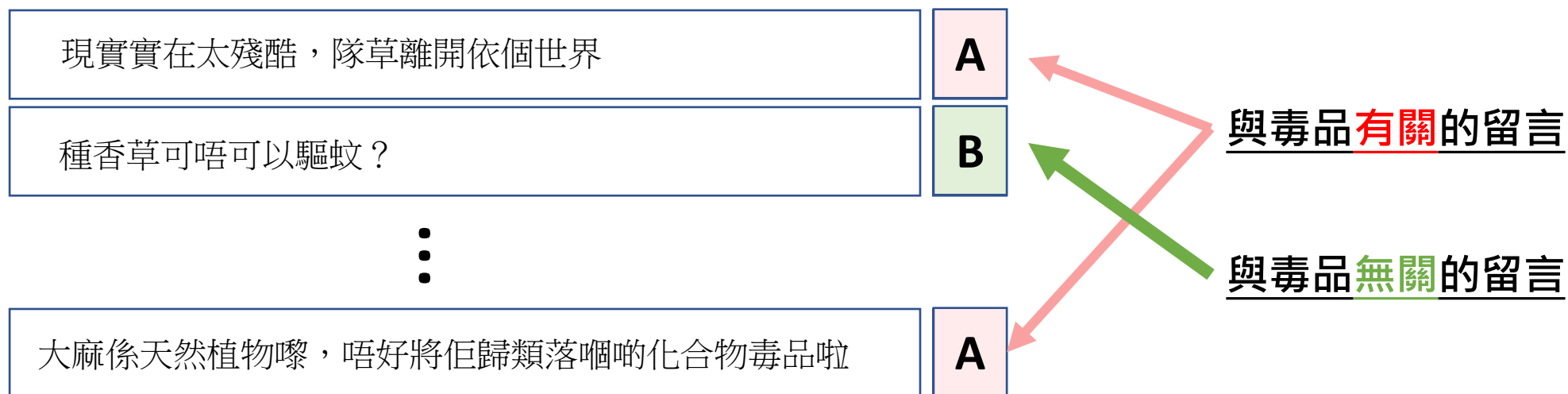
設計大麻討論監測系統

https://bit.ly/BDF_datasci

檔案-下載-Excel



Data pre-processing (Labelling)



- 與毒品有關的留言中有什麼常見的誤解？

問：以下哪一項是會令人出現幻覺的大麻素？

A. THC

B. CBD

C. MDMA

D. CNS

同樣是大麻素，但...

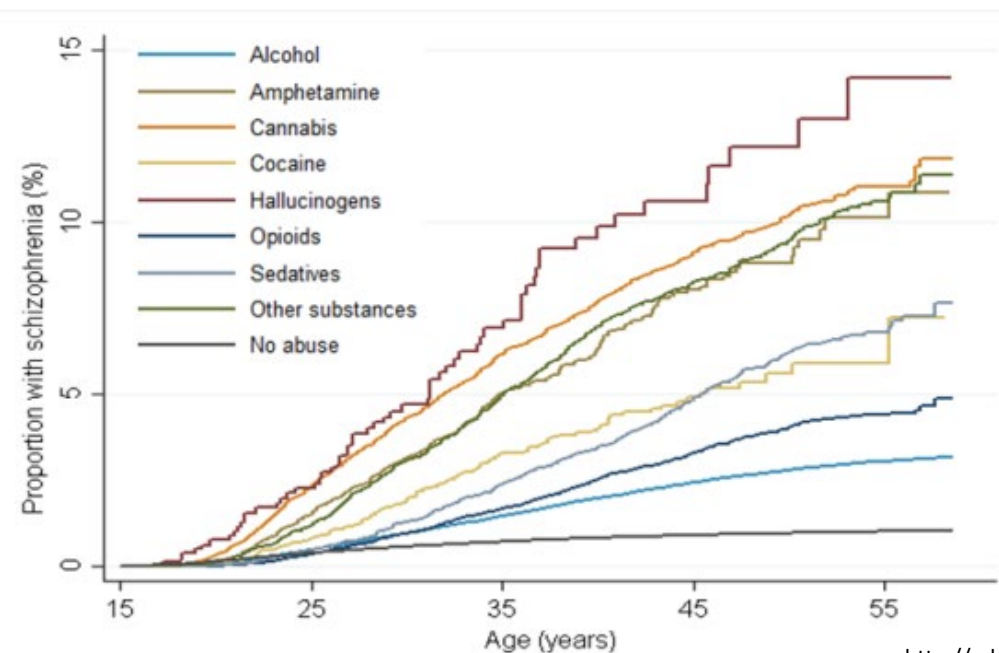
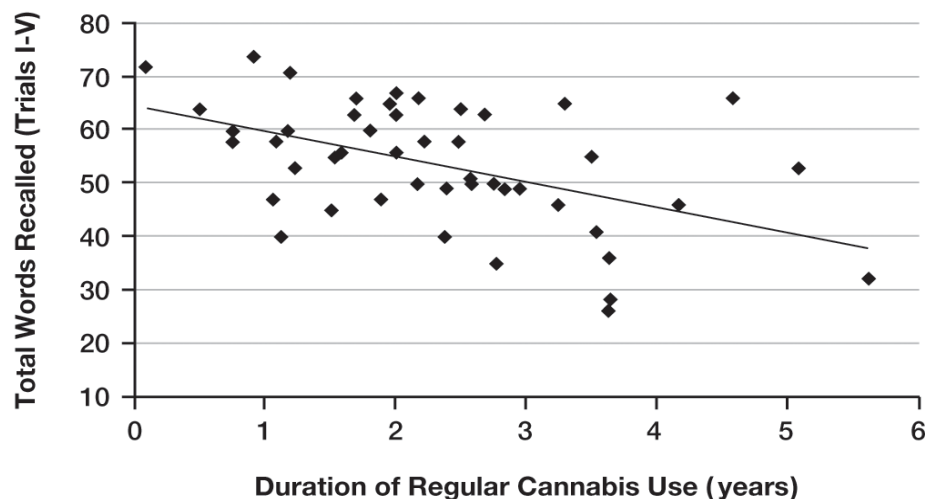


同樣是大麻素



1. 在香港擁有THC會屬刑事罪行
2. 有迷幻的感覺
3. 影響行為和認知能力
4. 長期使用能導致幻聽或精神分裂

Figure 1: Risk of Schizophrenia for All Substance Abuse Disorders

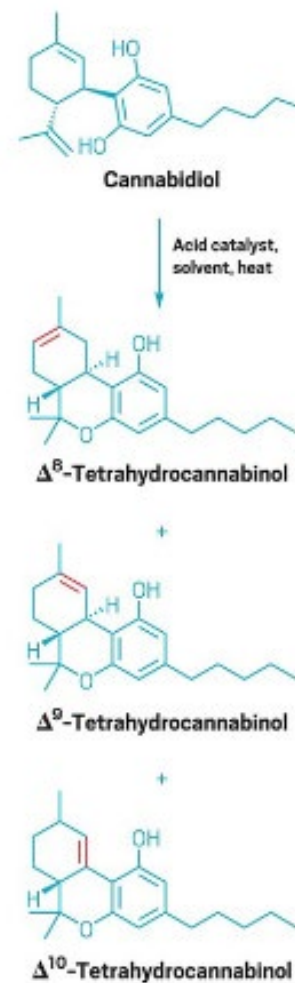


同樣是大麻素

- 聲稱能有助舒緩痛症及情緒，但**暫時未有完善的科學實證**
- 《危險藥物條例》已於2023年2月1日修訂，將CBD納入管制藥物
- 提煉時仍有機會有THC殘留，並會觸犯有關法例
- 研究指出有化學方法將CBD轉化成THC，或會有殘餘強酸及重金屬



香港海關突擊搜查了七間售賣 CBD 產品的零售店，包括大型連鎖零售商萬寧旗下三間分店，還有位於長沙灣和葵涌的兩個倉庫，搜獲25,000件懷疑含有 THC 的 CBD產品，包括 CBD 油、護膚品和寵物零食。



2021年12月1日星期三

下一篇

大麻二酚產品含禁藥 拘8人 標榜健康功效 警揭煙油咖啡粉含致癮四氫大麻酚



讚好 5

A⁺ A⁻



問：以下哪一種植物對人體來說是有毒的？

A. 夾竹桃

B. 水仙

C. 一品紅

D. 曼陀羅

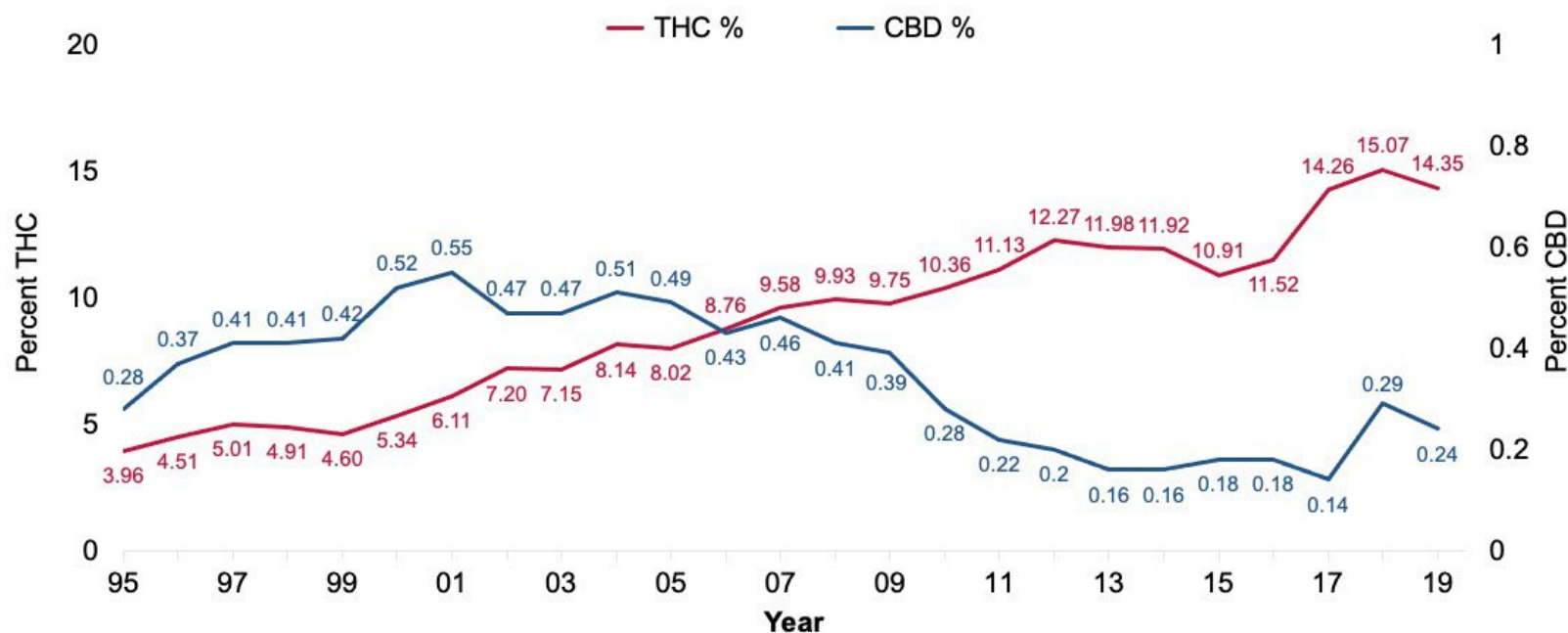
大麻是天然就是對人無害？



再者...

大麻真的是全天然嗎？

Percentage of THC and CBD in cannabis samples seized by the DEA from 1995-2019



根據美國緝毒局(DEA) 按年分析非法大麻中的THC及CBD的含量
→ 人為因素令THC 含量顯著上升

Text normalization

I think marijuana can be a great way to bond w/ frds! 🐼 #chill



i think marijuana can be a great way to bond **with friend chill**



think

marijuana

can

great

way

bond

friends

chill

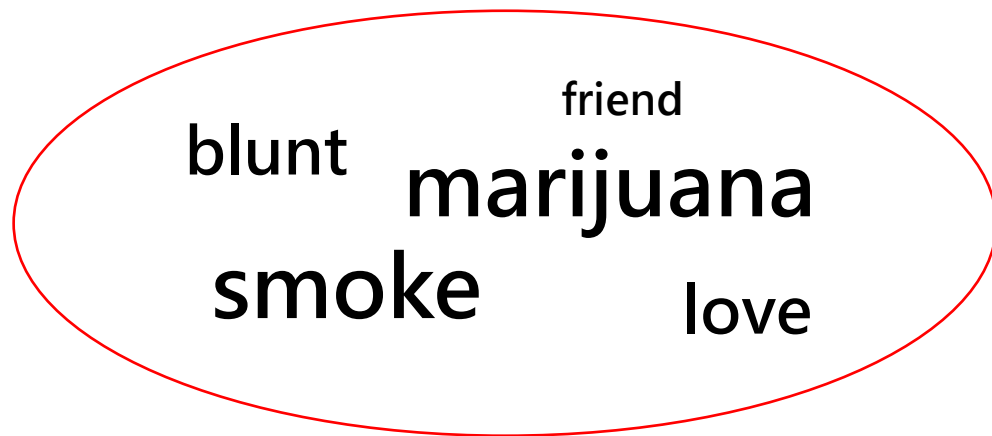
Normalized text

Bag-of-words

smoke marijuana friend

love marijuana

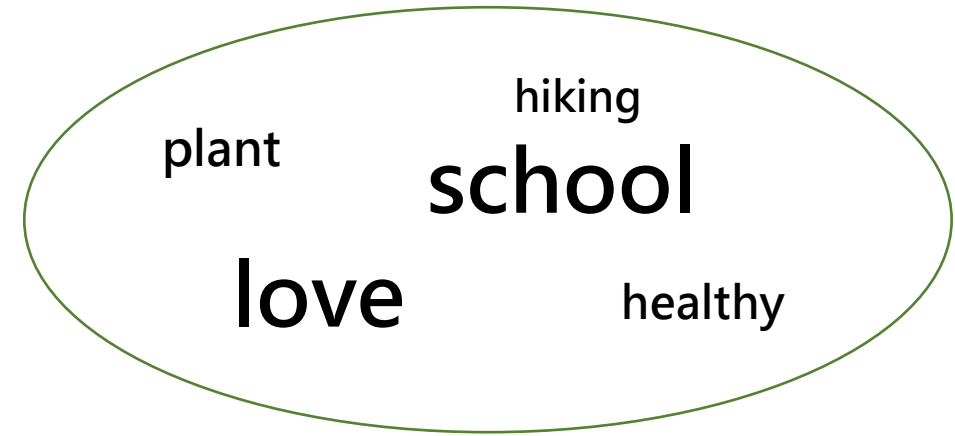
smoke blunt healthy



love plant school

love hiking

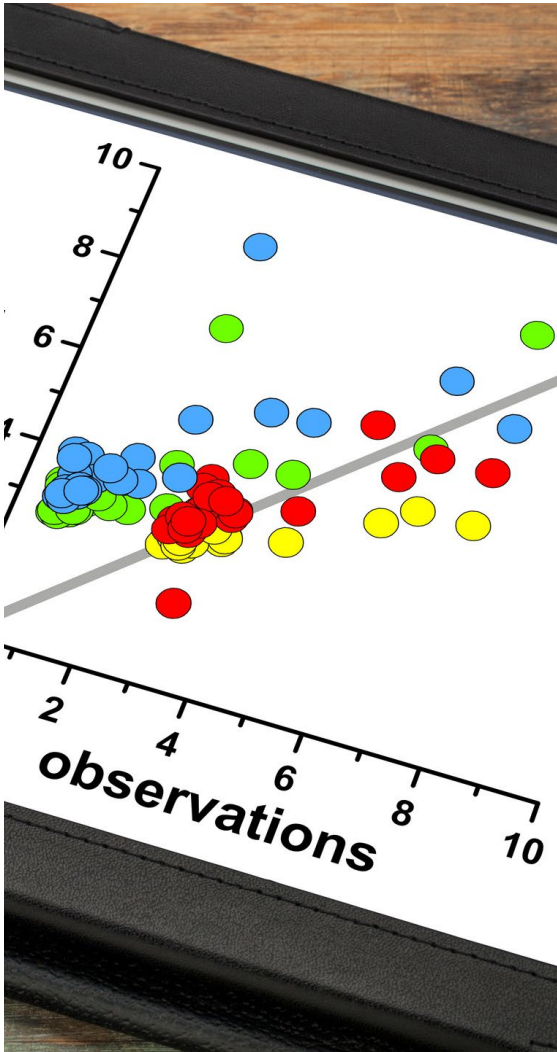
go school healthy



第3步:資料視覺化及運用:文字雲 (Word Cloud)

- <https://wordcloud.timdream.org/>





Modelling 製作模型

- 找出數據之間的關係
- 選用適合的演算法製作模型

提取有用的資訊

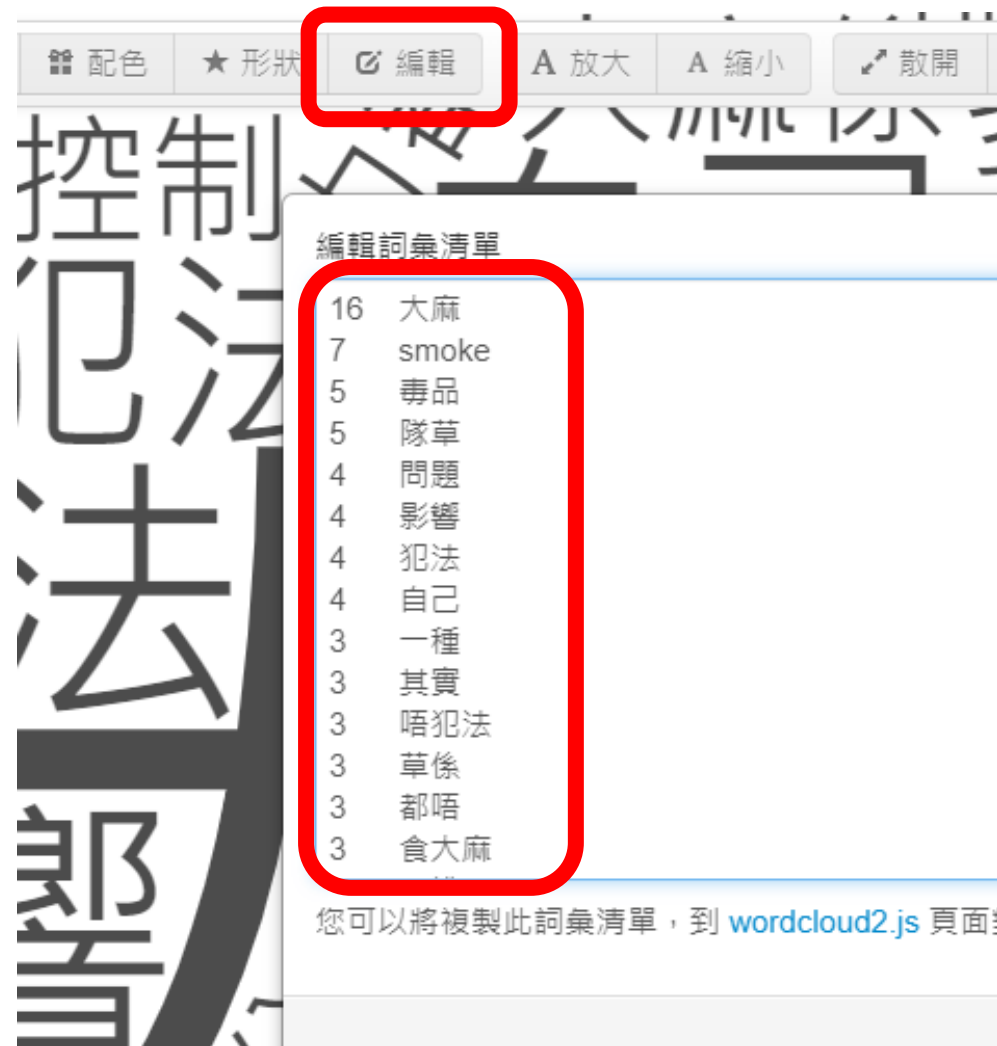
與濫毒有關的留言中
有什麼關鍵詞？

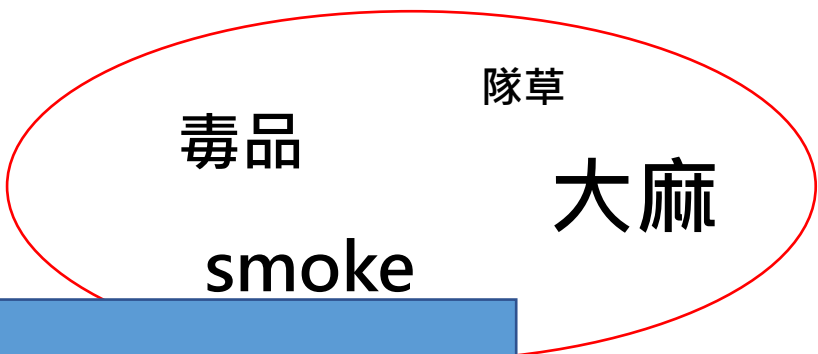
草/weed/Blunt/Hash

Smoke/stone/420



.....





分數越高，代表出現的次數越多

關鍵字的關聯越大

Class A

Class B

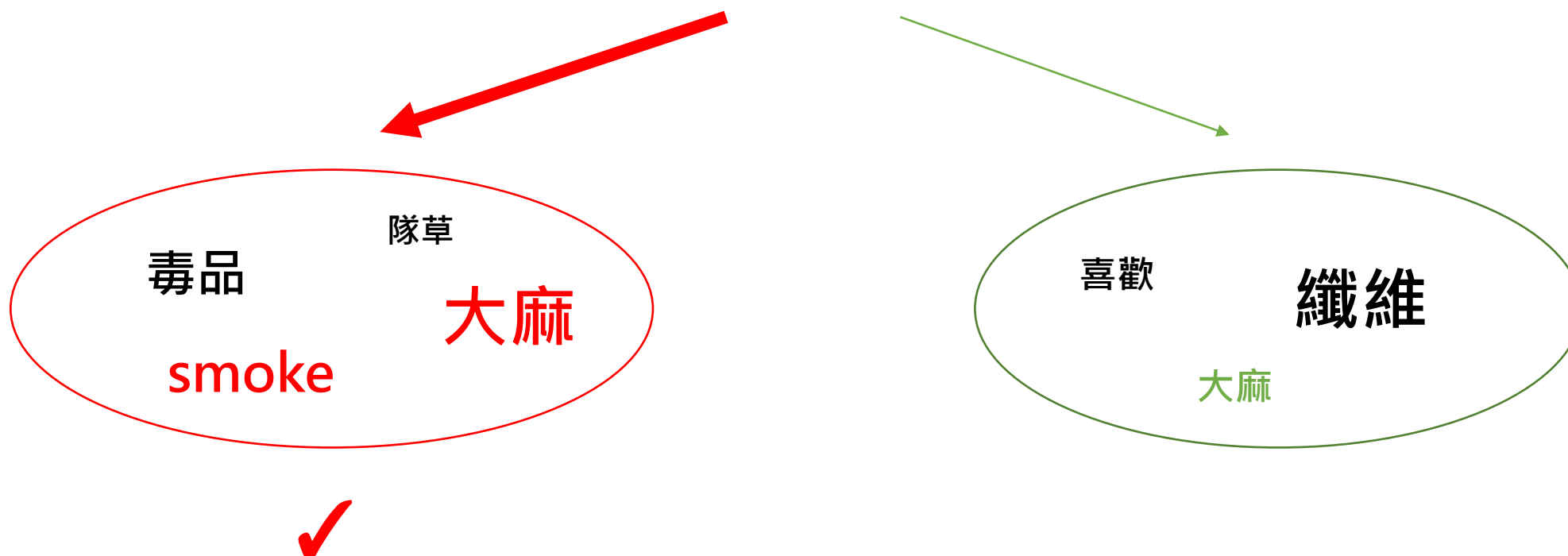
大麻	0.13	0.01
纖維	0.01	0.10
smoke	0.33	0.01
隊草	0.04	0.01
喜歡	0.01	0.05

沒有出現的字詞，分數極少
關聯很小

Naïve-Bayes Classifier

- 根據關鍵詞出現的頻率，計算與哪個分類的關聯性較高

大麻唔係毒品，我日日smoke都冇事



Naïve-Bayes Classifier

我 喜歡 穿 大麻 纖維 的 衣服 喜歡 質感

喜歡

大麻

纖維

喜歡

2

1

1

Classifier	Class A	Class B	Test
大麻	0.13	0.01	1
纖維	0.01	0.10	1
smoke	0.33	0.01	
隊草	0.04	0.01	
喜歡	0.01	0.05	2

$$\text{Score (Class A)} = p(\text{Class A}) \times 0.01^2 \times 0.13^1 \times 0.01^1 = 0.000001$$

$$\text{Score (Class B)} = p(\text{Class B}) \times 0.05^2 \times 0.01^1 \times 0.10^1 = 0.000180$$

Testing your algorithms

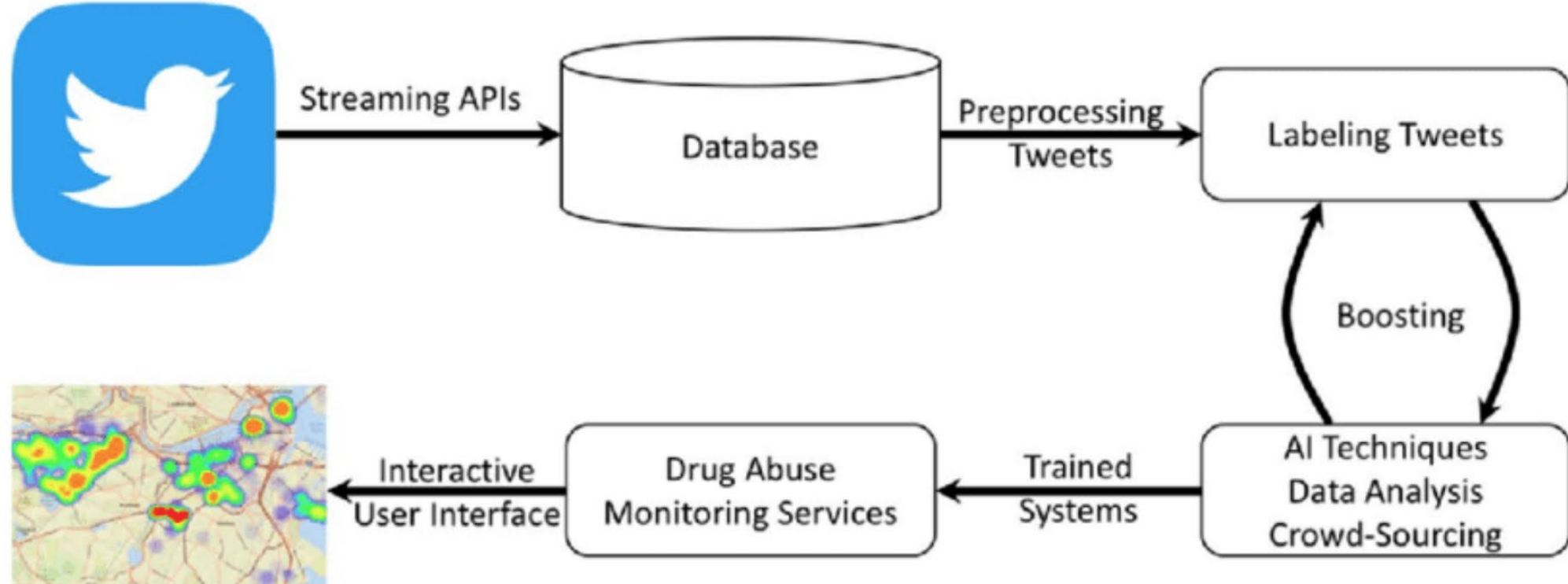
- 大麻都唔會上癮，我返工放break都會smoke，啲人控制唔到就話毒品，有冇了解過架

A

- 大麻纖維結構穩定，大分子排列取向度高，產生靜電能力極低，可用作製作衣物。

B

**這個演算法有什麼限制？
有什麼方法可以提升系統的準確度？**

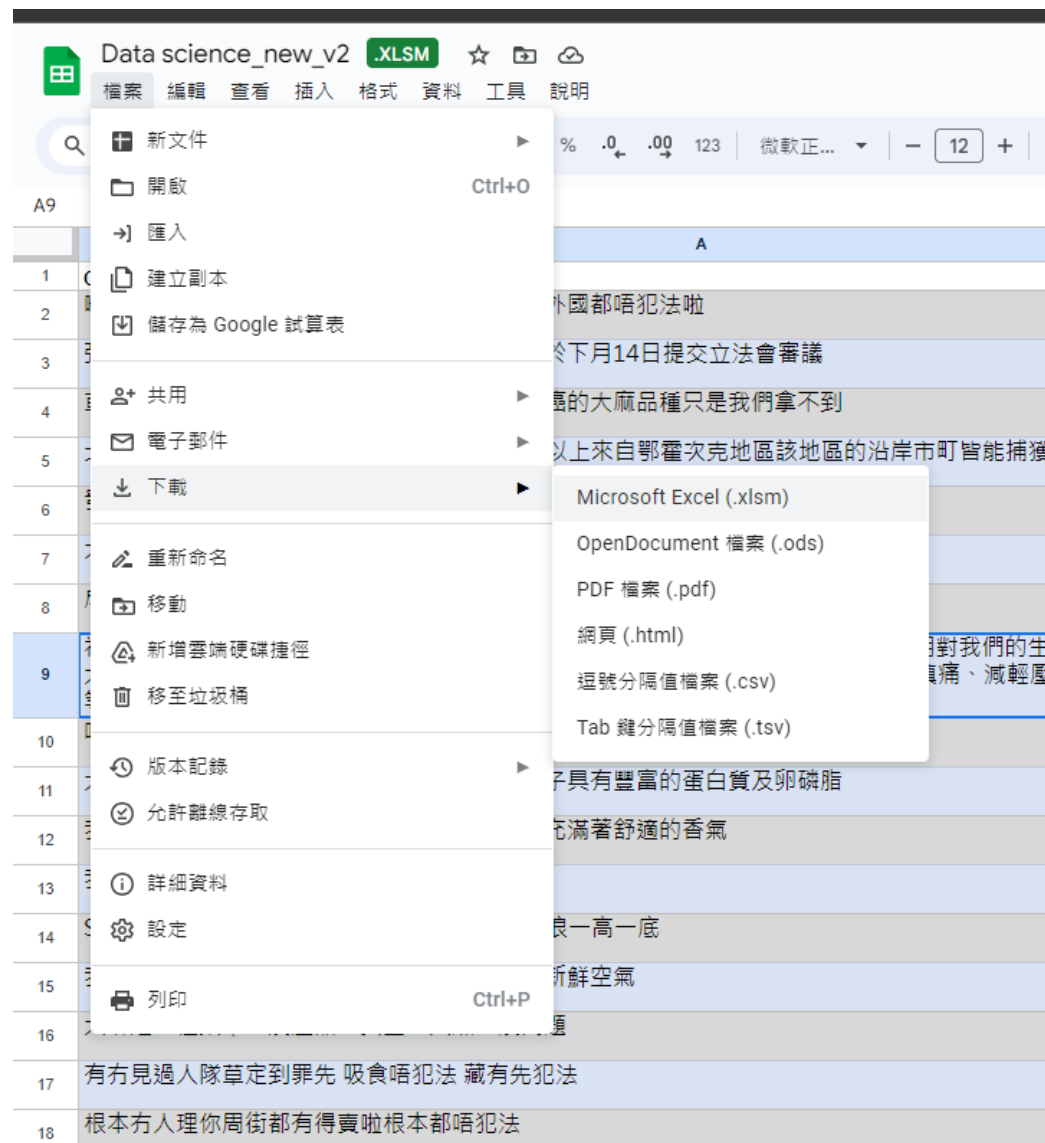


Hu, Han & Phan, Nhat Hai & Chun, Soon & Geller, James & Vo, Huy & ye, Xinyue & Jin, Ruoming & Ding, Kele & Kenne, Deric & Dou, Dejing. (2019). An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning. Computational Social Networks. 6. 10.1186/s40649-019-0071-4.

分析吸食大麻與記憶力之影響

https://bit.ly/BDF_datasci

檔案-下載-Excel

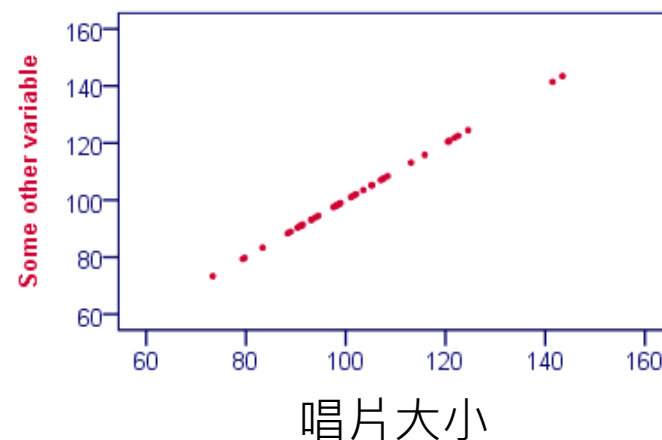


透過分析數據展示兩者的關係

- 利用散佈圖去展示每個個別的數據
- 在散佈圖上畫上線性回歸線理解兩者的關係
- 例子：問卷調查結果（第1及2題）
 - 插入圖表（散佈圖）
 - 加上趨勢線
 - 顯示R平方值

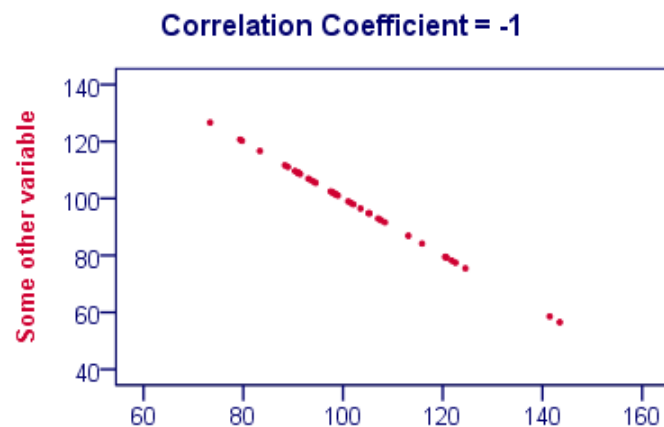


總播放時間

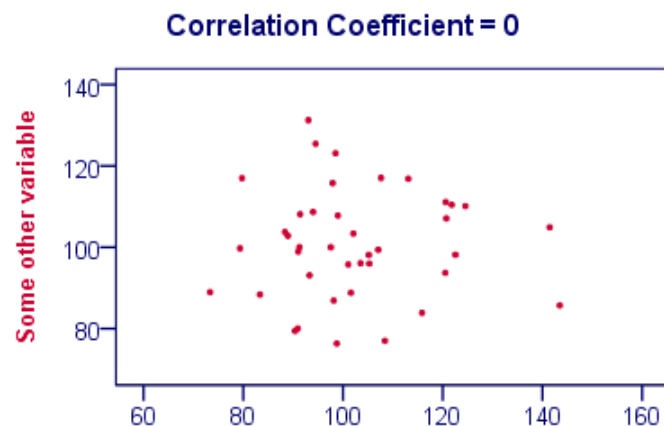


利用線性回歸理解兩者的關係

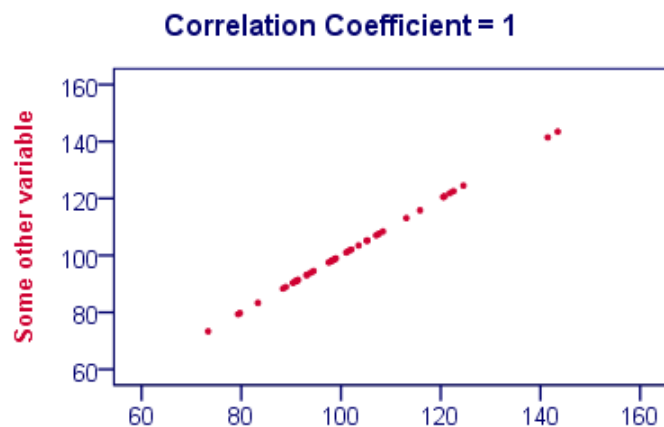
- 相關係數 (R) 反映兩個變數之間的相互關係及其相關方向
- 由-1至1



強烈負相關的關係



沒有明顯的關係

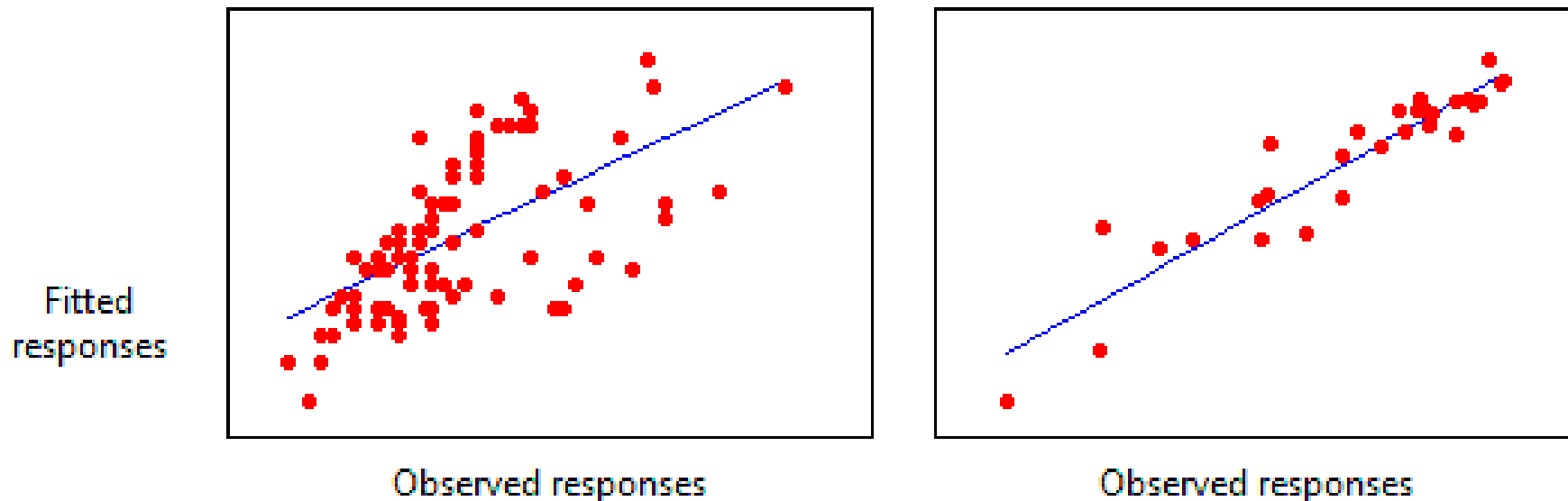


強烈正相關的關係

利用線性回歸理解兩者的關係 (2)

- 決定系數 (R^2) 反映判斷回歸模型的解釋力
- 由0至1表達數據與回歸模型之吻合度

Plots of Observed Responses Versus Fitted Responses for Two Regression Models



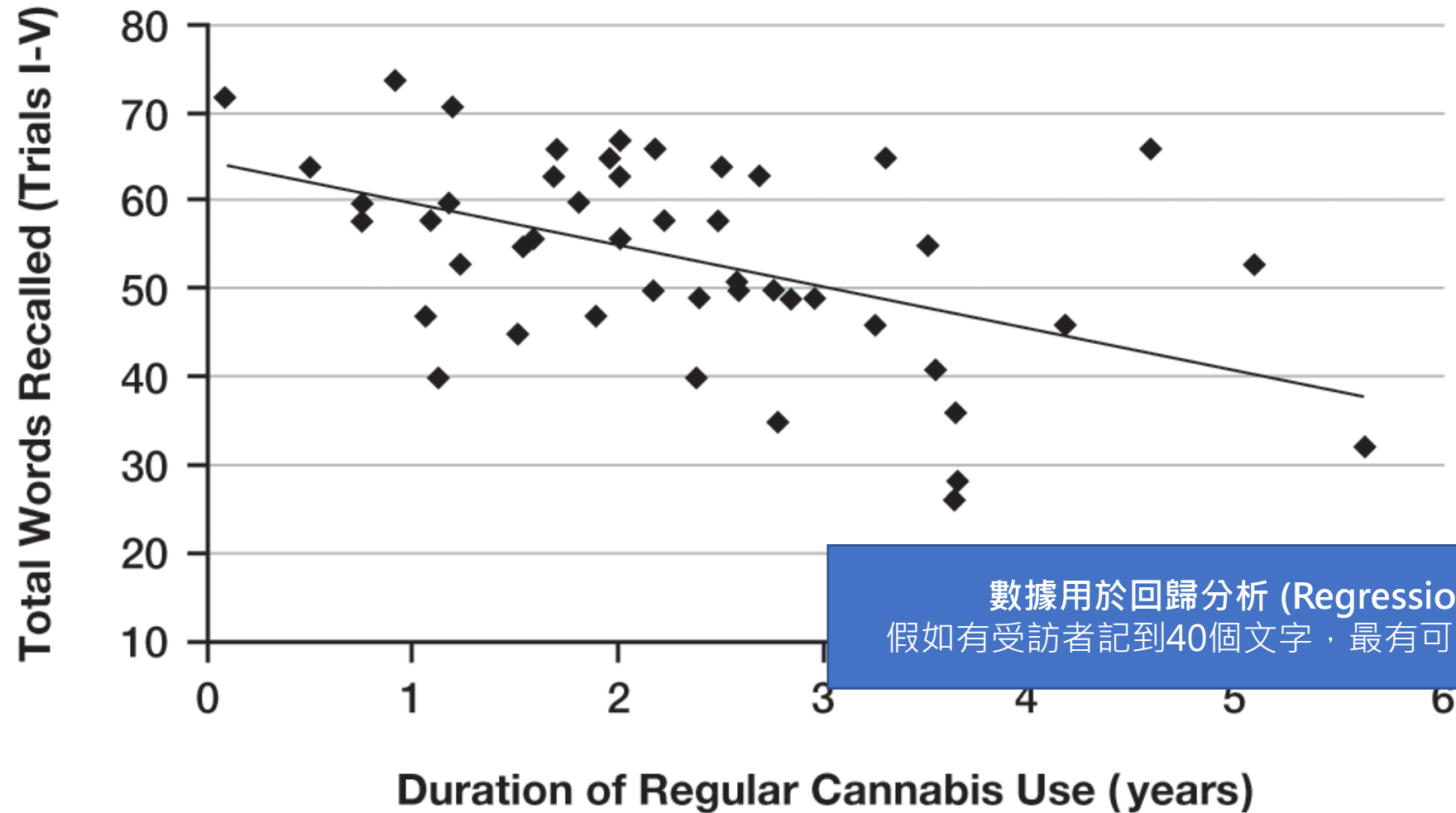
運用數據模型 分析吸食大麻與記憶力的關係？(2)

- 利用線性回歸分析吸食大麻的年資及記憶力的關係
- 47個曾經吸食大麻的青少年
- 吸食大麻的年資及能夠記起文字的數目之關係

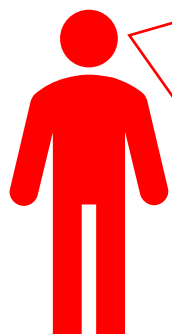
No.	Codename	Duration of Regular Cannabis (Months)	Total Words Recalled (Trials I-V)
1	AA	1	72
2	AB	7	64
3	AC	10	60
4	AD	10	57
5	AE	12	74
6	AF	15	71
7	AG	13	59
8	AH	14	61
9	AI	15	53
10	AJ	13	47
11	AK	14	40
12	AL	19	55
13	AM	20	56
14	AN	19	46
15	AO	22	65
16	AP	22	63
17	AQ	23	59
18	AR	23	57
19	AS	24	65
20	AT	25	67
21	AU	25	63
22	AV	25	57

25	AY	26	50
26	AZ	29	49
27	BA	28	40
28	BB	30	63
29	BC	30	58
30	BD	31	50
31	BE	31	51
32	BF	33	63
33	BG	34	36
34	BH	34	51
35	BI	35	50
36	BJ	35	50
37	BK	40	65
38	BL	43	55
39	BM	39	48
40	BN	43	42
41	BO	44	36
42	BP	44	28
43	BQ	43	26
44	BR	50	46
45	BS	56	66
46	BT	61	51
47	BU	69	31

利用線性回歸分析吸食大麻毒齡及記憶力的關係



數據的重要性



• 剛剛



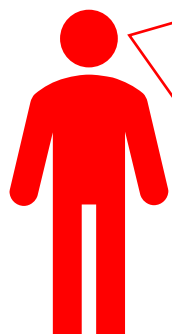
顯示 #1

食少少，最多咪瞓覺怕咩啫？我食咁耐都冇事。



網上留言的數據真確性成疑

數據的重要性



• 剛剛



顯示 #1

食少少，最多咪瞓覺怕咩啫？我食咁耐都冇事。



數據真確性及代表性

