

統計實務

Spring 2024

授課教師：統計系余清祥

日期：2024年3月6日

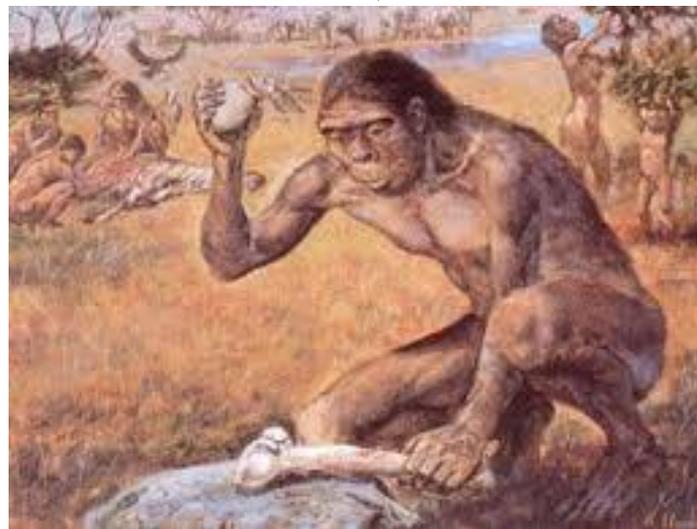
第二週：統計思維



資訊與決策

□ 過去擁有土地、資金等的資本家，透過勞力及資本密集有絕對優勢；21世紀轉變為「知識經濟」，擁有資訊、能夠解讀及應用資訊者具有競爭力。（例如：Google、Facebook！）

→ 問題：以人類歷史（石器時代至今）而言，各個年代掌握生存優勢的關鍵是什麼？



天下文化 遠見

人類大命運

從智人到神人

Homo Deus

A Brief History of Tomorrow

by Yuval Noah Harari

哈拉瑞 著 林俊宏 譯

天下文化

人類大歷史

從野獸到扮演上帝

Sapiens

[From Animals Into Gods]

A Brief History of Humankind

by Yuval Noah Harari

哈拉瑞 著 林俊宏 譯

2021年美國壓力最大(小)的職業

□ 美國求職網站CareerCast從200種工作，評選出2021年最佳和最糟工作排行榜，最佳者：

I、大數據相關（知識經濟、資訊安全）

- 精算師、數學家、統計學家、數據科學家、軟體工程師、電腦系統分析師

II、醫療健康（人本、老化）

- 聽力專家、生物醫學工程師、牙科保健員、職業治療師

2021年美國最佳職業排行榜

排名	職業	年薪中位數 (美元)	就業增長率
1	資料科學家	\$98,230	33%
2	遺傳治療顧問	\$85,700	21%
3	統計學家	\$92,270	35%
4	醫療服務主管	\$104,280	32%
5	數學家	\$110,860	33%
6	大學教授	\$80,790	9%
7	運籌學分析師	\$86,200	25%
8	資訊安全分析師	\$99,730	31%
9	精算師	\$111,030	18%
10	軟體工程師	\$110,140	22%

2022年美國最佳職業排行榜

排名	職業	年薪 (中位數)	職缺人數
1	資訊安全分析師	\$103,590	47,100(學士)
2	專業護理師	\$111,680	114,900(碩士)
3	外科醫師助理	\$115,390	40,100(碩士)
4	醫療服務經理	\$104,280	139,600(學士)
5	軟體工程師	\$110,140	409,500(學士)
6	資料科學家	\$98,230	19,800(學士)
7	財務經理人	\$134,180	118,200(學士)
8	統計學家	\$92,270	14,900(碩士)
9	律師	\$126,930	71,500(博士)
10	語言治療師	\$80,480	45,400(碩士)
20	精算師	\$111,030	6,800(學士)

2023年美國最佳職業排行榜

排名	職業	年薪（平均數）	就業增長率
1	全端工程師	\$129,637	56%
2	資料工程師	\$135,260	80%
3	雲端工程師	\$133,114	65%
4	精神科護理師	\$109,739	45%
5	資深產品經理	\$147,139	44%
6	後端資訊開發人員	\$148,827	81%
7	網站可靠性工程師	\$153,134	121%
8	機器學習工程師	\$153,252	53%
9	心理診所執業護理師	\$134,011	180%
10	產品經理	\$121,363	39%

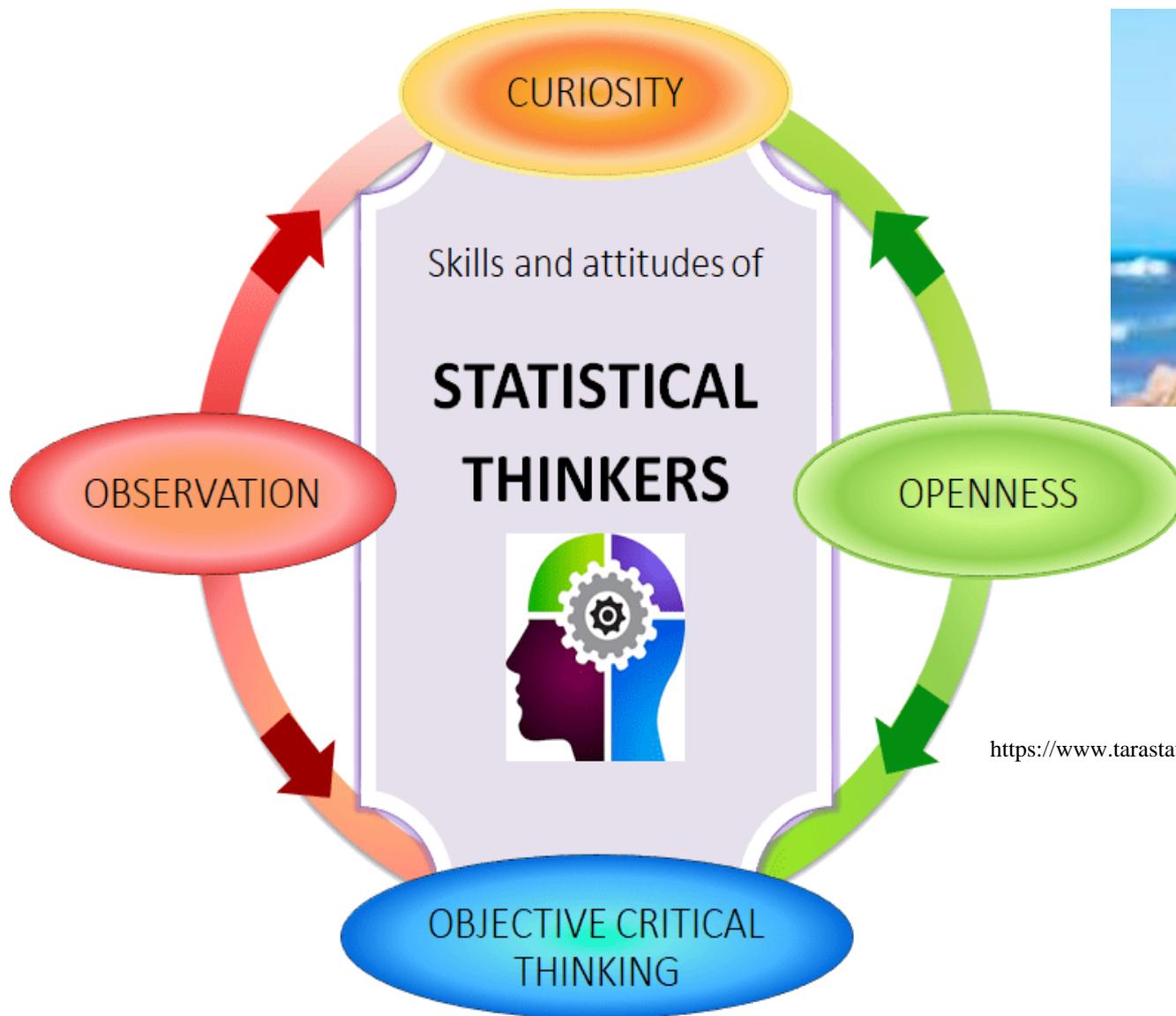
Rank	Job Title	Mean Annual Salary (Above 2021 Nat'l Avg: \$58,260 or higher)	Jobs Per 1M, Period ending 2023-01-01	% change in job share, Period ending 2020-01-01 vs Period ending 2023-01-01	% containing remote + hybrid phrases, Period ending 2023-01-01
1	full stack developer	\$129,637	1,398	56%	51
2	data engineer	\$135,260	879	80%	44
3	cloud engineer	\$133,114	678	65%	42
4	psychiatric nurse	\$109,739	537	45%	18
5	senior product manager	\$147,135	532	44%	45
6	back end developer	\$148,827	429	81%	60
7	site reliability engineer	\$153,134	377	121%	55
8	machine learning engineer	\$153,252	246	53%	37
9	psychiatric mental health nurse practitioner	\$134,011	230	180%	20
10	product designer	\$121,363	213	39%	48

資料科學家(Data Scientist)

- 統計等同於資料科學(Data Science)嗎？
→ 參考Amstat News的文章「The Identity of Statistics in Data Science」。
 - 資料科學家工作內容多元(Multi-disciplinary)，不只熟悉統計分析，也需要與人溝通、報告撰寫、程式軟體、商業智慧與決策等之能力。
- 註：現今學校尚無統合訓練（即使有、人數也不多），人才缺額暫時無法補足。

充分完備的統計/資料科學家

- 統計學家所需技能為下列「溝通」能力：
 - 與人溝通：寫作、口語表達、溝通能力；
 - 與資料（及統計）溝通：data sense、資訊圖像化、特性與趨勢；
 - 與專業溝通：領域知識、問題定義及結果詮釋、附加價值；
 - 與電腦（機器）溝通：資料儲存與更新、資訊安全、程式運算。

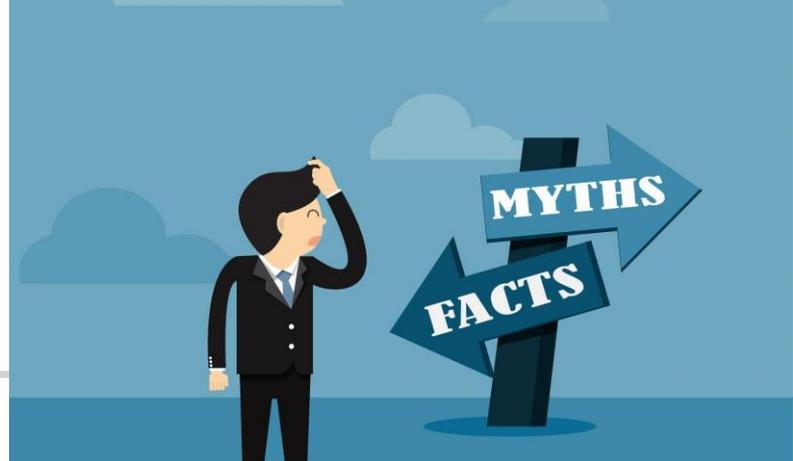


<https://kenwinter.org/blog/what-do-these-stones-mean>

<https://www.tarastats.com/decision-making-8-0/>

統計思維與資料分析

大數據等於一切？



<https://datafloq.com/read/cutting-through-buzz-big-data-myths-debunked/>

□ 無限母體 vs. 有限母體

→ 再多的資料還是樣本！（你們贊成嗎？）

註：曾經有人認為在大數據時代，統計將會滅絕，因為我們可以擁有所有資料。

□ 擁有資料就無所不能？

→ 問題：實務上如何預測未來的天氣？

→ 問題：有人將隨機漫步用於預測股票價格，這代表股市有什麼特性？

英國股票指數的走勢(2024年2月22日)

<https://tw.stock.yahoo.com/>

英國FTSE

7672.61

▲ 10.10

7672.61 ▲ 10.10 (0.13%)

德國DAX

17327.24

▲ 209.12

法國CAC

7871.92

▲ 59.83

荷蘭AEX

855.13

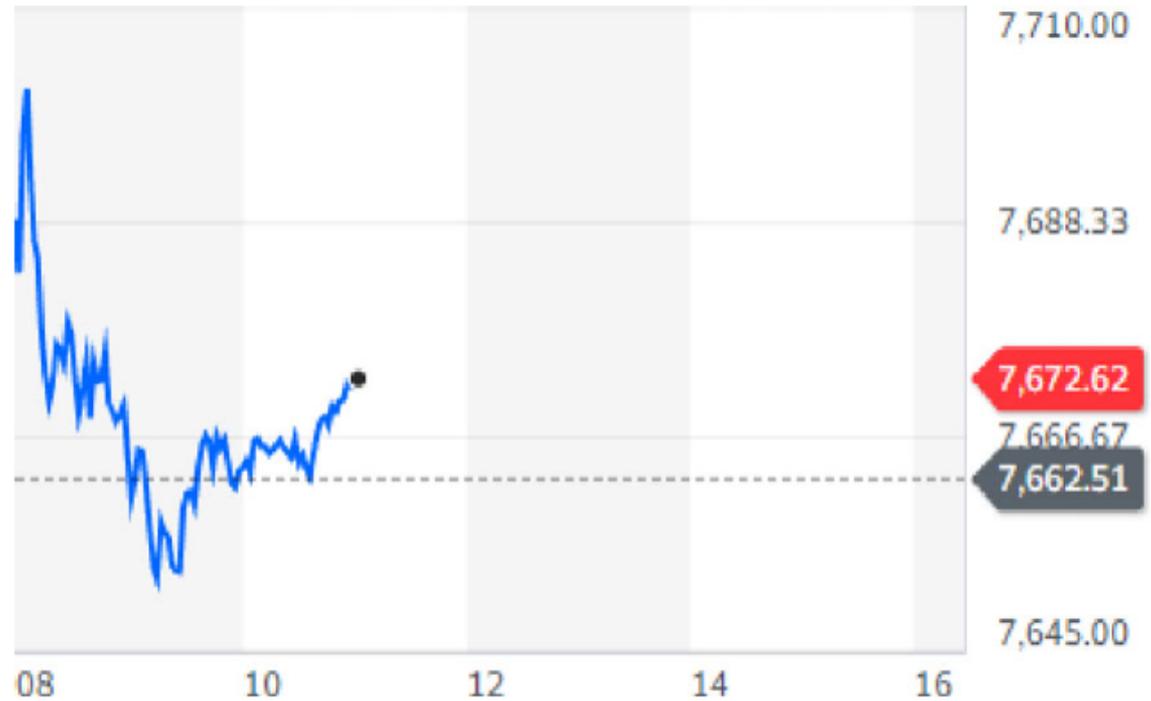
▲ 8.34

開盤
7597.53

最高
7702.86

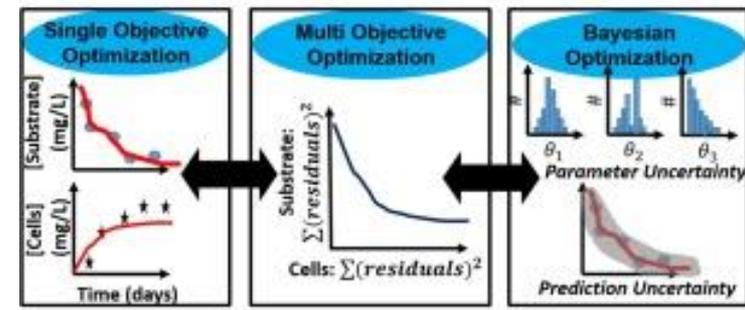
最低
7651.65

昨收
7662.51



問題：接下來指數會往上或往下？

估計及預測母體參數



<https://www.sciencedirect.com/science/article/pii/S2215016119301554>

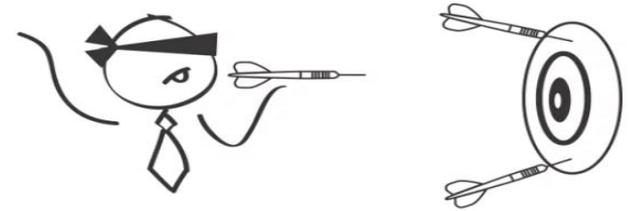
□ 估計 vs. 預測

→ 估計 (Estimation) 由觀察值計算的樣本統計量作為參數估計值；預測 (Prediction) 以過去和現在的觀察值估計未來。

□ 估計需確定觀察值是否足以代表母體

→ 若已有可反映參數的資料，估計方法較為單純；通常會選擇隨機抽樣，但有時只能使用非隨機抽樣。

點估計與區間估計



<https://www.nimblework.com/wp-content/uploads/2019/03/agile-estimation.png>

□ 點估計：最佳估計值

→ 最大概似估計量、最小平方法、動差法、最小變異數不偏估計 (minimum-variance unbiased estimator)。

□ 區間估計：參數的可能範圍

→ 設定容許誤差(或信心水準)，建立可反應抽樣誤差的區間估計，包括Confidence (或Credible) interval。

問題：信賴區間必須是兩側對稱嗎？

預測的不確定性更高



<https://madaboutsience.weebly.com/prediction-vs-hypothesis.html>

□ 估計與預測的差異？

→ 如果未來母體和現在的母體有類似特性，包括：參數相同、或是趨勢接近，則可用現有資訊作為預測依據。

□ 問題：如何確定未來母體的特性？

註：預測相關用詞包括Prediction、Forecast、Projection，三者間有何差異？

英國股票指數的走勢(2024年2月22日中午)

<https://tw.stock.yahoo.com/>

英國FTSE

7678.70

▲ 16.19

7678.70 ▲ 16.19 (0.21%)

開盤

7597.53

最高

7702.86

最低

7651.65

昨收

7662.51

德國DAX

17343.63

▲ 225.51

法國CAC

7890.57

▲ 78.48

荷蘭AEX

857.75

▲ 10.96



■ 你/妳們猜對了嗎？！

英國股票指數的走勢(2024年2月22日收盤)

<https://tw.stock.yahoo.com/>

英國FTSE

7684.49

▲ 21.98

7684.49

▲ 21.98 (0.29%)

開盤

7597.53

最高

7702.86

最低

7651.65

昨收

7662.51

德國DAX

17370.45

▲ 252.35

法國CAC

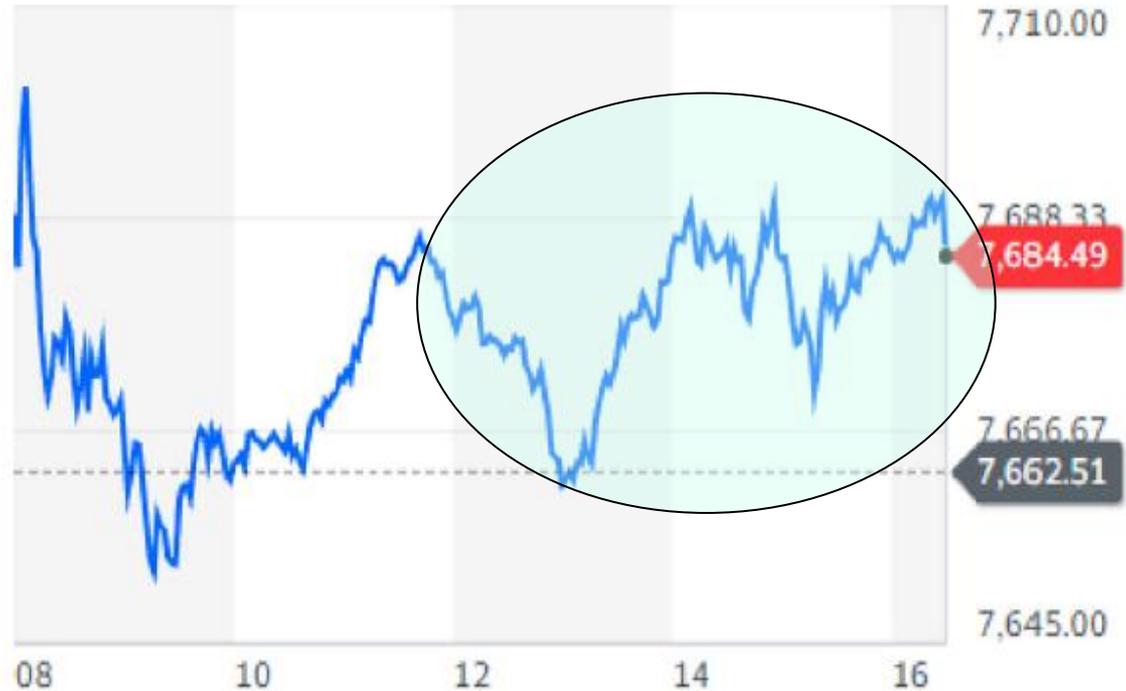
7911.60

▲ 99.51

荷蘭AEX

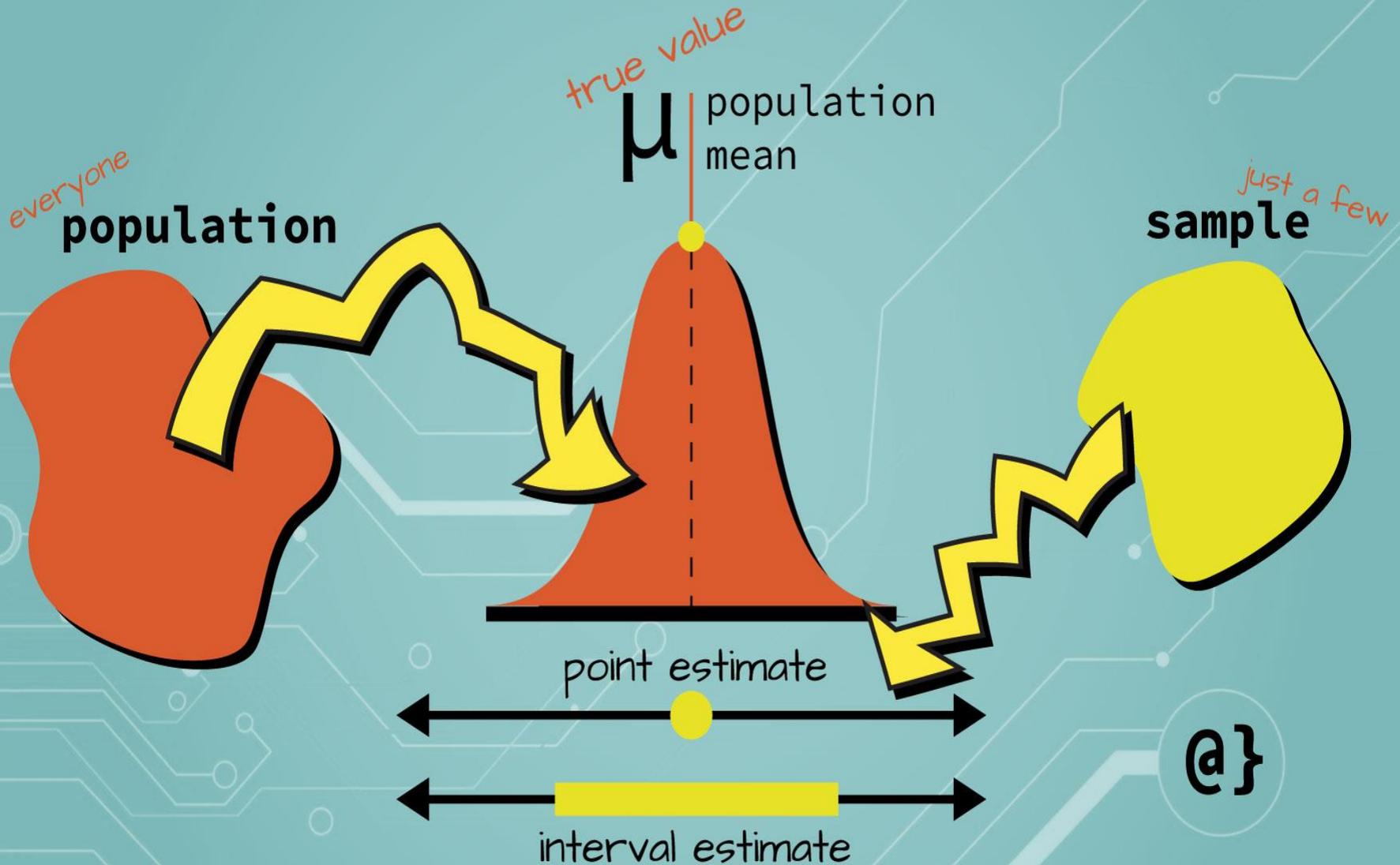
857.51

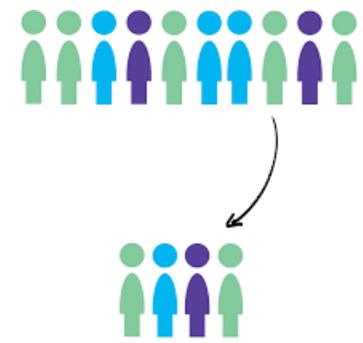
▲ 10.72



■ 從開盤、午盤猜得到收盤嗎？

關於統計估計





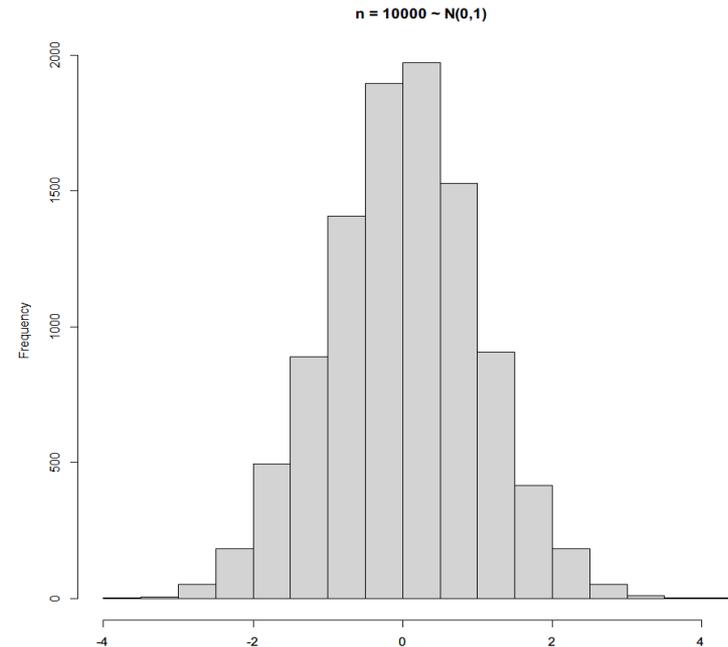
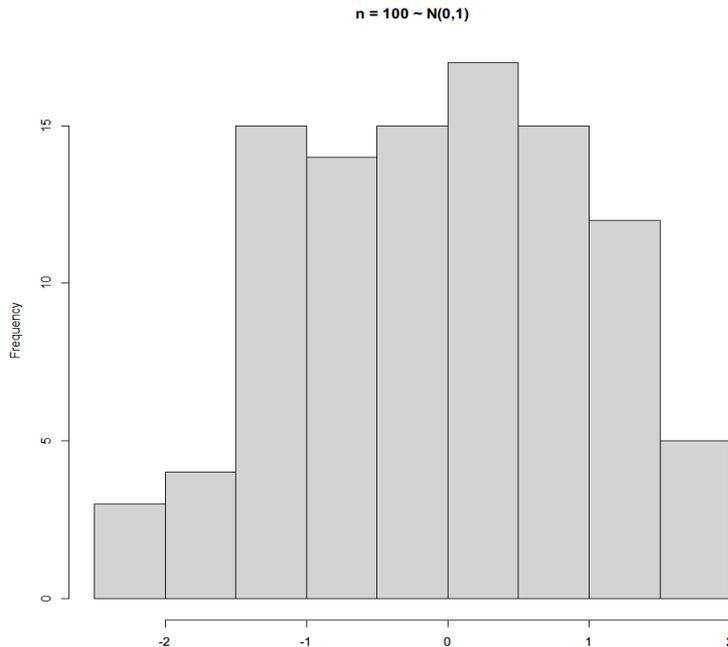
母體與樣本

<https://www.qualtrics.com/experience-management/research/population-vs-sample/>

□ 觀察值反映母體特徵：

$$\text{Observations} = \text{Truth} + \text{Error}$$

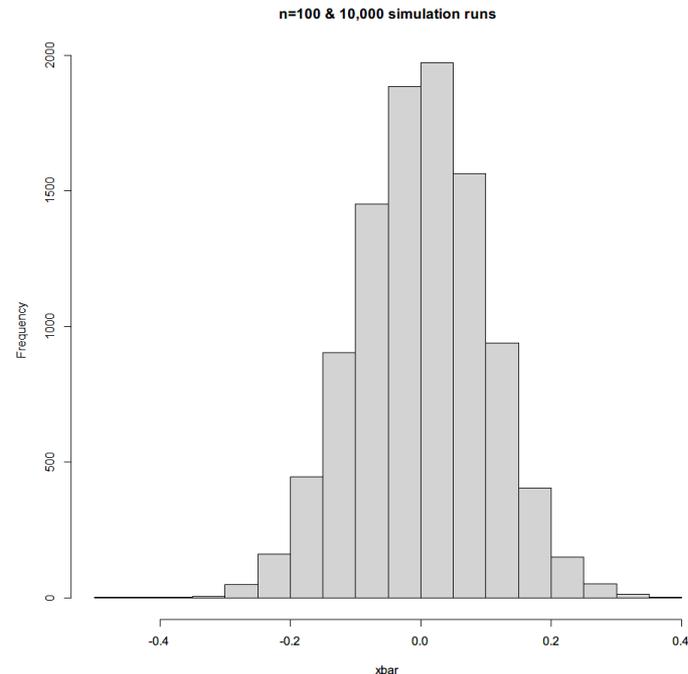
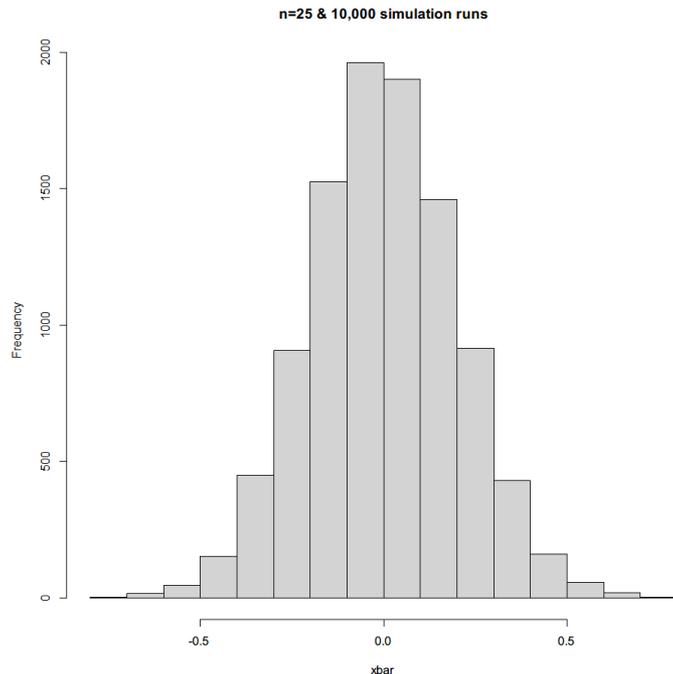
→ 觀察值愈多、愈能瞭解母體特性。



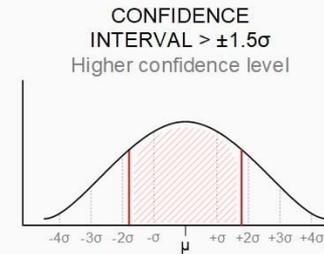
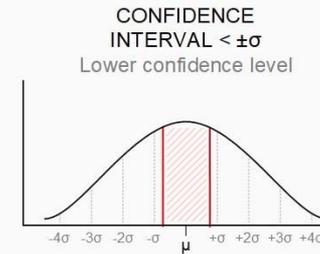
樣本數與估計誤差

□ 如果估計期望值，觀察值服從常態分配，樣本數愈多、平均數分佈愈集中！

變異數為0.0401 & 0.0098 (約4倍)



抽樣誤差與區間估計



<https://vru.vibrationresearch.com/lesson/confidence-intervals/>

□ 抽樣誤差伴隨的不確定性，通常以區間(範圍)估計因應，精確度要求愈高、愈無法達成。

→ 信心係數愈大、信賴區間也愈寬！

□ 如何詮釋信賴區間？

→ 95% 不代表建立信賴區間後，涵蓋真實參數值的機率！！（註：參數值未知且固定，只有落在信賴區間之內或之外兩種結果。）

註：貝氏統計對參數的認知略有不同。

生活中關於機率的案例

□ 預測臺北市明天下雨機率為30%，如何詮釋？

→ 過去有100次和明天類似的氣候環境，平均有30次下雨。(問題：氣象署的下雨定義？)

→ 明天臺北市有30%的面積會下雨。

□ 某位棒球選手打擊率為三成，如何詮釋？

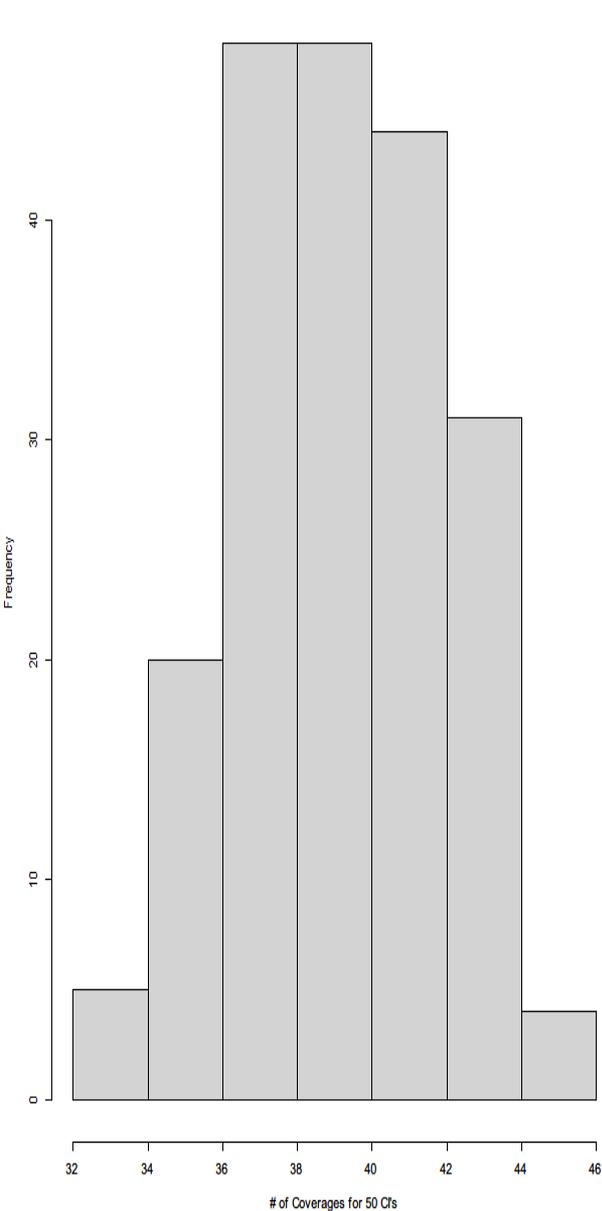
→ 選手維持三成打擊率，每次上場打擊時有30%出現安打，並非預測下次擊出安打的正確性為三成。(隨機猜中選擇題答案也類似。)

如何詮釋信心係數？

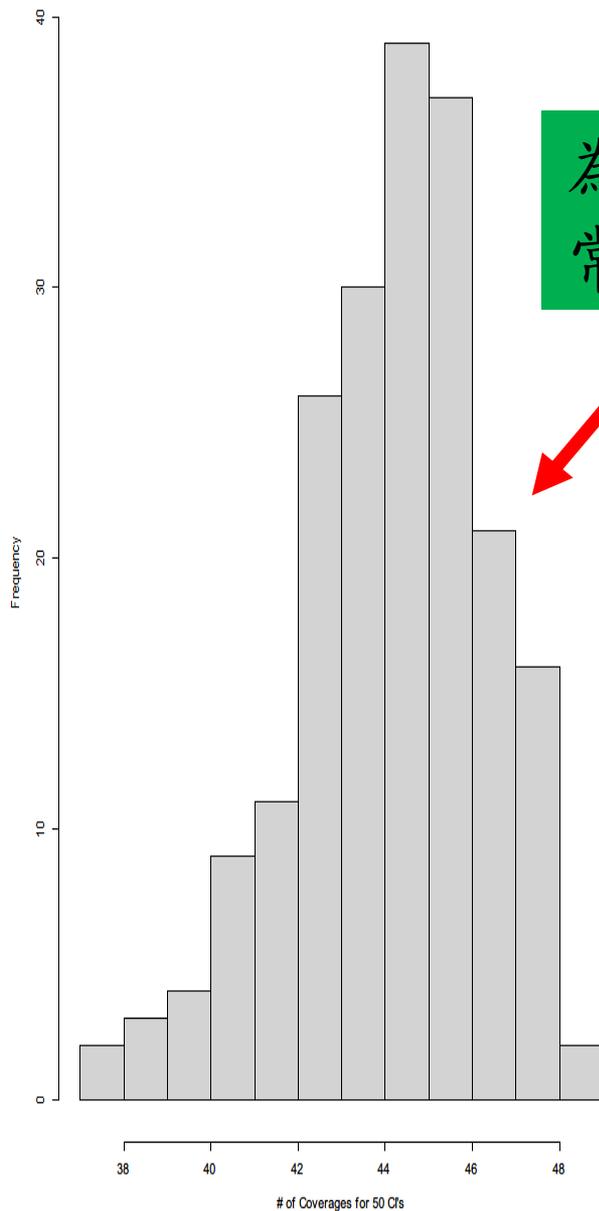
- 信賴區間為程序機率，不用於單次解釋。
→ 一旦取得樣本資料，點估計、區間估計就已確定，沒有所謂的機率、或是信心水準。
- 再次以電腦模擬驗證信心係數，隨機產生100個標準常態分配的亂數，80%、90%、95%對應臨界值為1.281552、1.644854、1.959964，重複一萬次電腦模擬，涵蓋真實期望值的比例分別為79.33%、89.30%、94.64%。

每50個信賴區間涵蓋真實值的個數(200次)

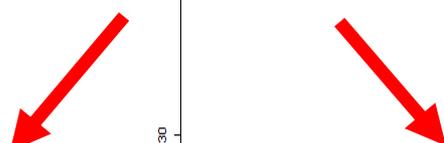
80% Confidence Interval



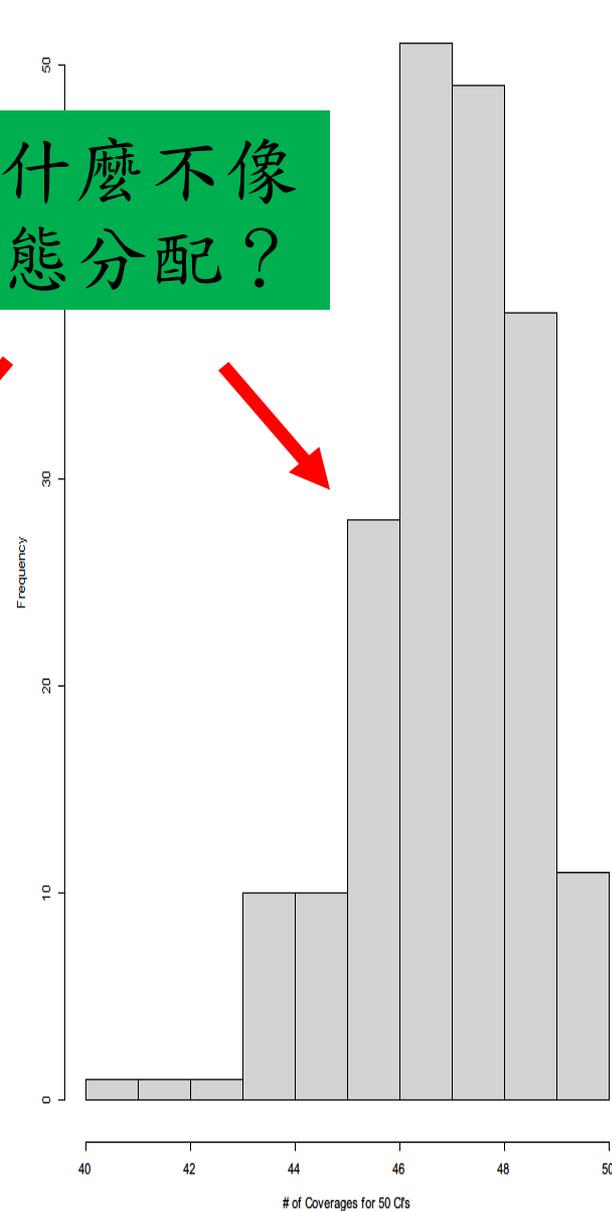
90% Confidence Interval



為什麼不像
常態分配？



95% Confidence Interval



解決問題的關鍵



<https://web.colby.edu/cogblog/2017/04/05/are-you-a-victim-of-the-bandwagon-effect/>

□ 資料品質仍是主要關鍵！

→ 人工智慧仰賴好的訓練資料，尋找背後的規則及趨勢（問題：如果只給予AlphaGo初學者的範例，結果又會如何？）。

→ Google Flu Trends以搜尋引擎的「關鍵詞」，比美國疾管局（Center of Disease Control; CDC）更早偵測出流感盛行。（問題：這種方法有什麼瑕疵？）

註：什麼是從眾效應(Bandwagon effect)？

Home

Select country/region

[How does this work?](#)

[FAQ](#)

Flu activity



[Download world flu activity data](#) - [Animated flu trends for Google Earth](#) - [Compare flu trends across regions in Public Data Explorer](#)

數據依賴(Data Dependency)

□ 模型分析因為資料而產生不同結果。

→ 世界衛生組織計算標準體重之方法：

男性： $(\text{身高cm} - 80) \times 70\% = \text{標準體重}$

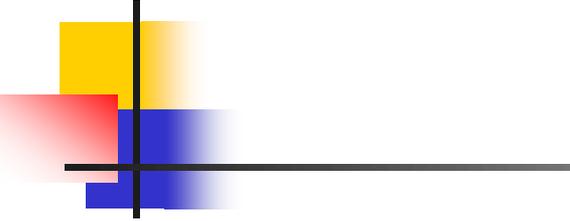
女性： $(\text{身高cm} - 70) \times 60\% = \text{標準體重}$

問題：如何確定不同母體的差異？

□ Data Dependency有幾種定義，前後程式、不同變數之間，代表意義也非常不同。

→ 迴歸分析會考慮變數間是否相關，避免線性相依(Linear Dependency)造成的共線性問題。

解決問題(Problem Solving)的流程



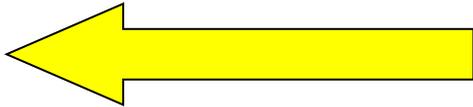
定義問題

蒐集資料



分析資料

詮釋結果



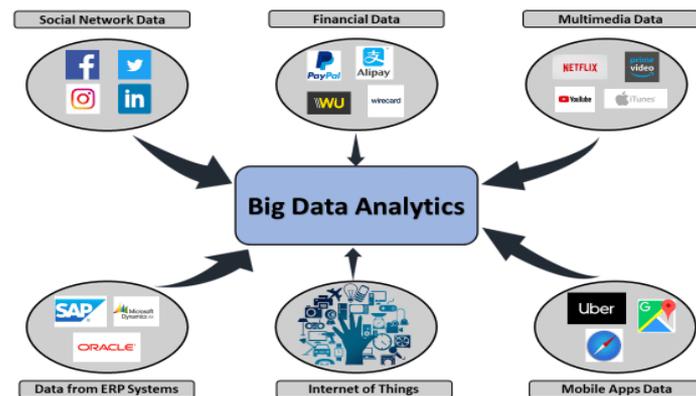
絕大多數的
統計教學重心

資訊來源

□ 拜網際網路之賜，今日資料來源更為多元，除了政府或當局的官方網站，也有個人分享經驗、提供建議。

→ 如何判斷資訊來源較為可信？

註：透過搜尋引擎及關鍵字詞、各種網路資源、或AI軟體獲得即時資訊。



資訊爆炸 vs. 分析解讀

□ 日常生活中到處充斥「資訊」，哪些真正需要的關鍵因素？

→ 哪些是必要資訊？

→ 資料品質、判斷資料真偽？（「Garbage in, garbage out」）

→ 如何根據既有資訊判斷？（哪些決策的風險較高、如何降低風險？）

什麼是大數據？

□ 大數據2010年由IBM所提出，涵蓋四個V：

→ 大量化(Volume)：至少TB及PB以上

→ 快速化(Velocity)：即時處理

→ 多樣化(Variety)：視頻、GIS等多樣性

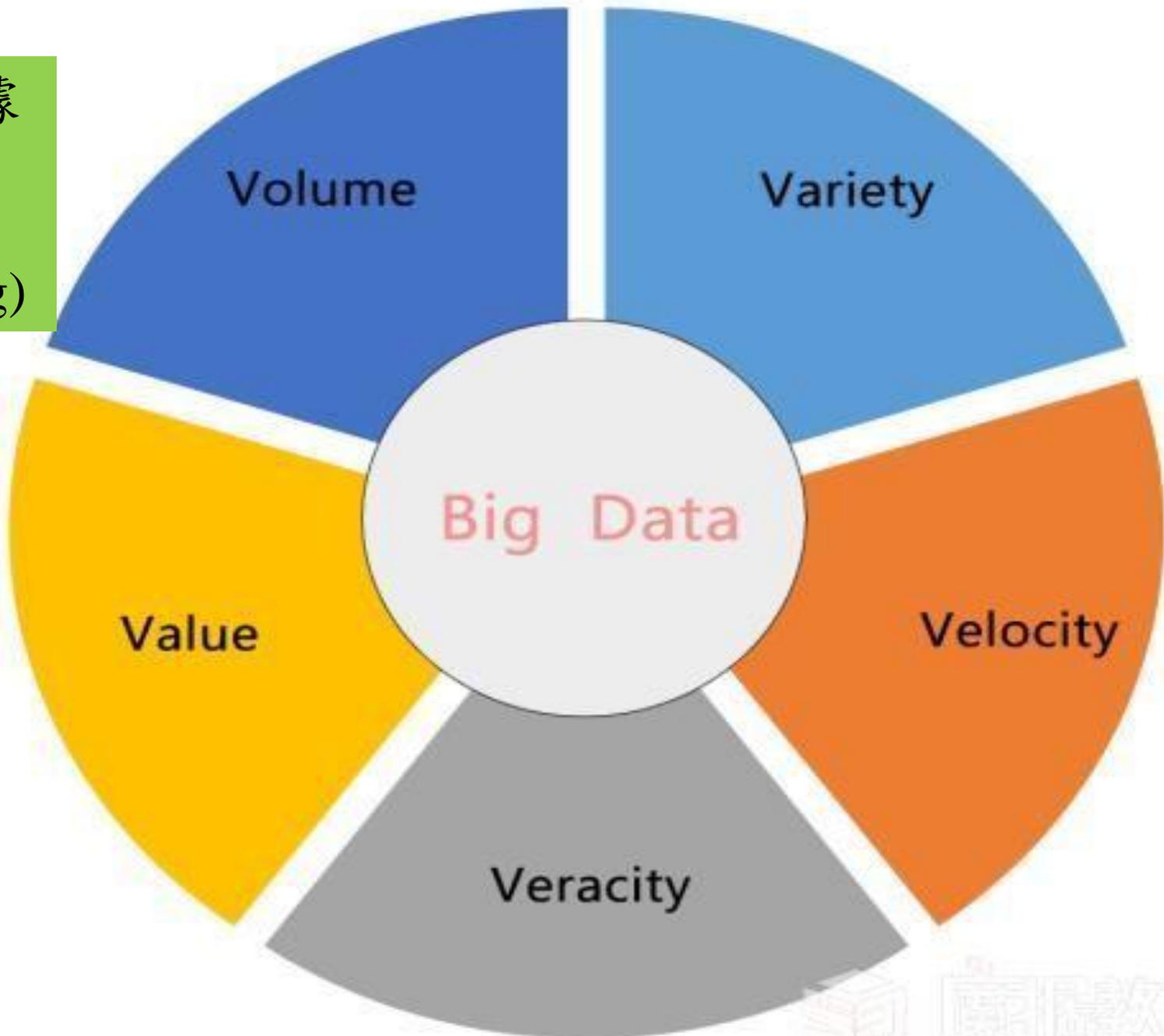
→ 真實性(Veracity)：資料品質（2014年提出）

註：加入資料分享與傳遞(Visible)後成為「大、

快、雜、疑、轉」；也有人加入價值(Value)

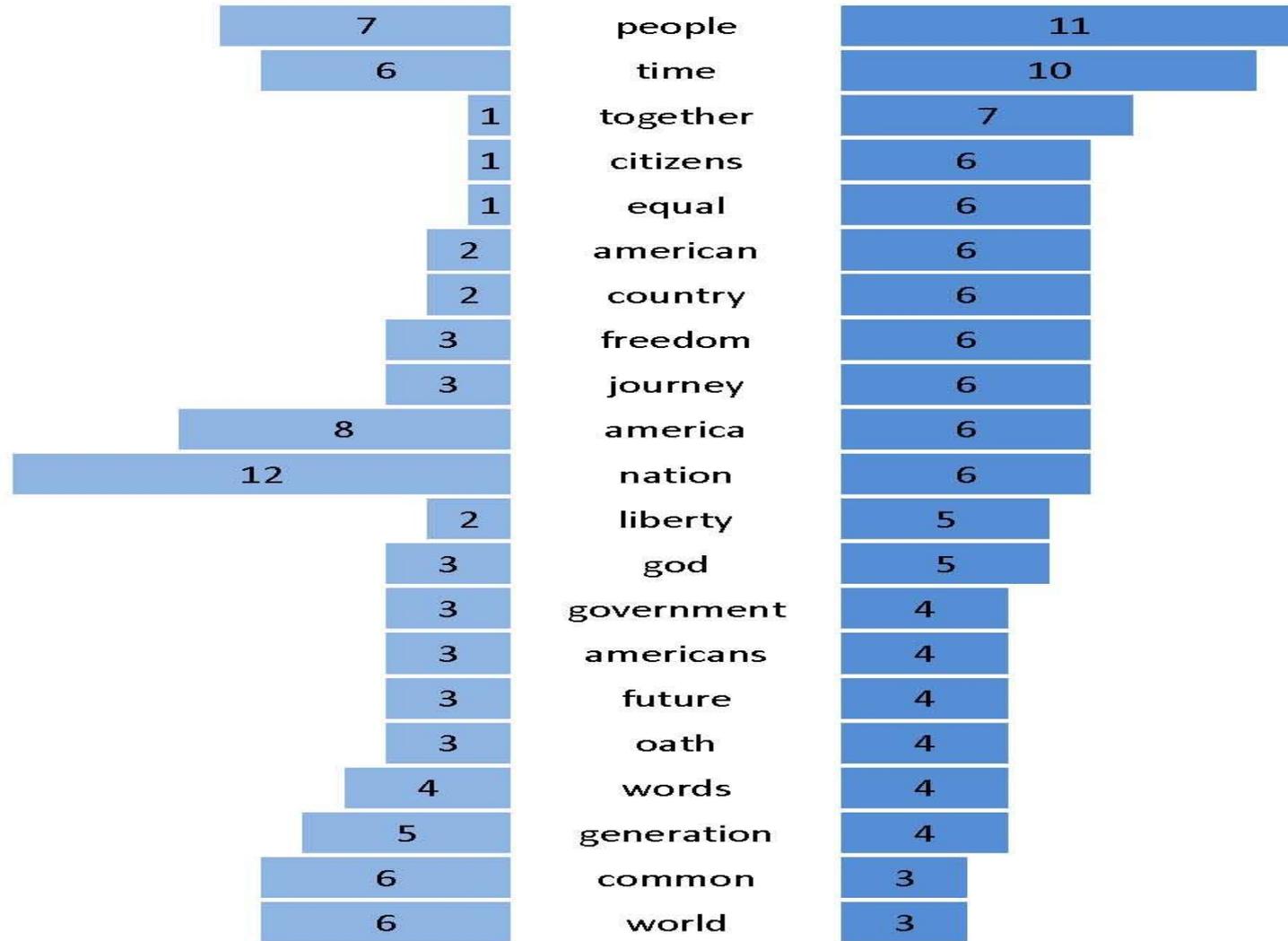
成為新的5V。

Value：數據
價值密度相
對較低。
(Datamining)



Comparing Inaugural Addresses

■ 2009 ■ 2013

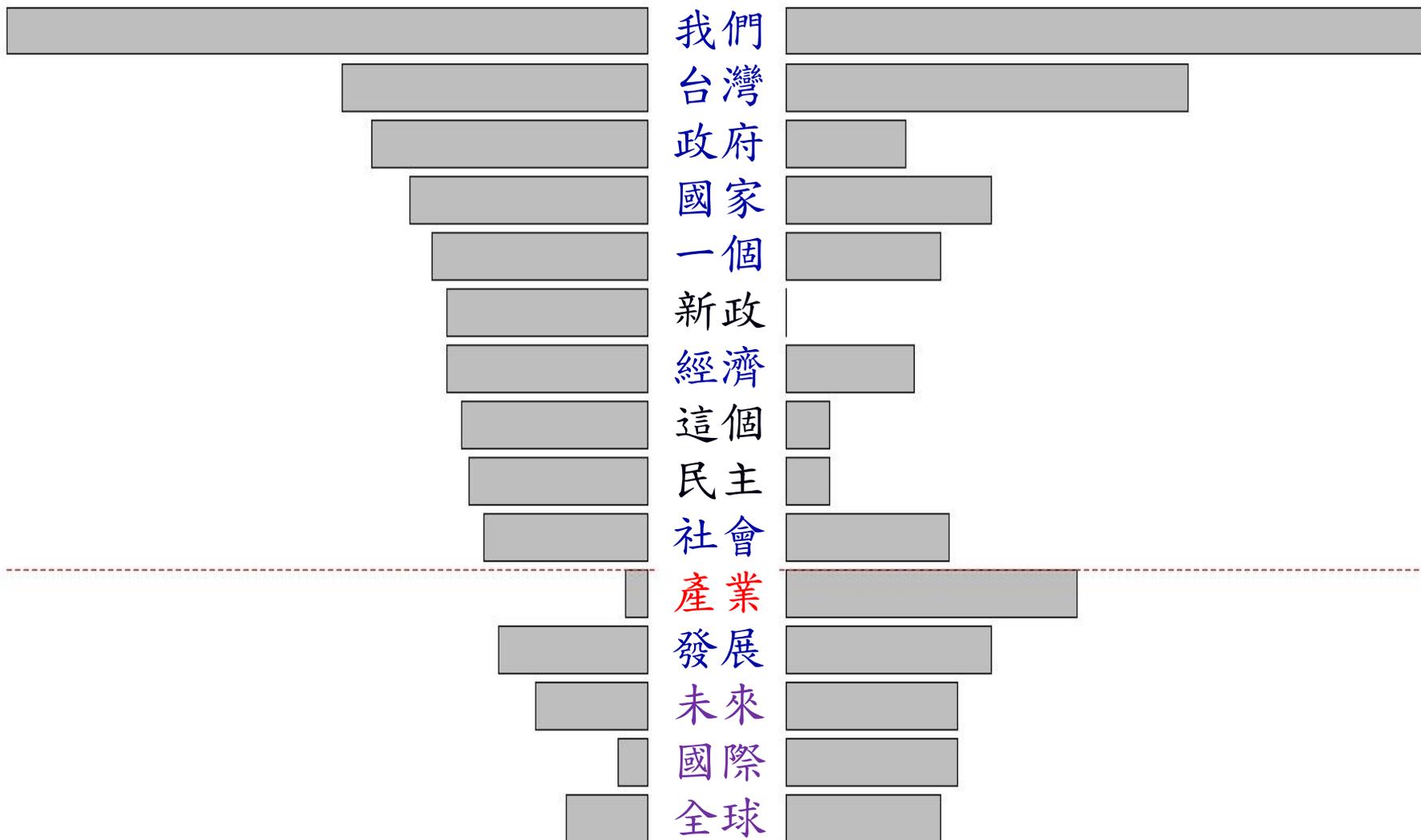


Analyzing the speech of President Obama (**Textmining**)

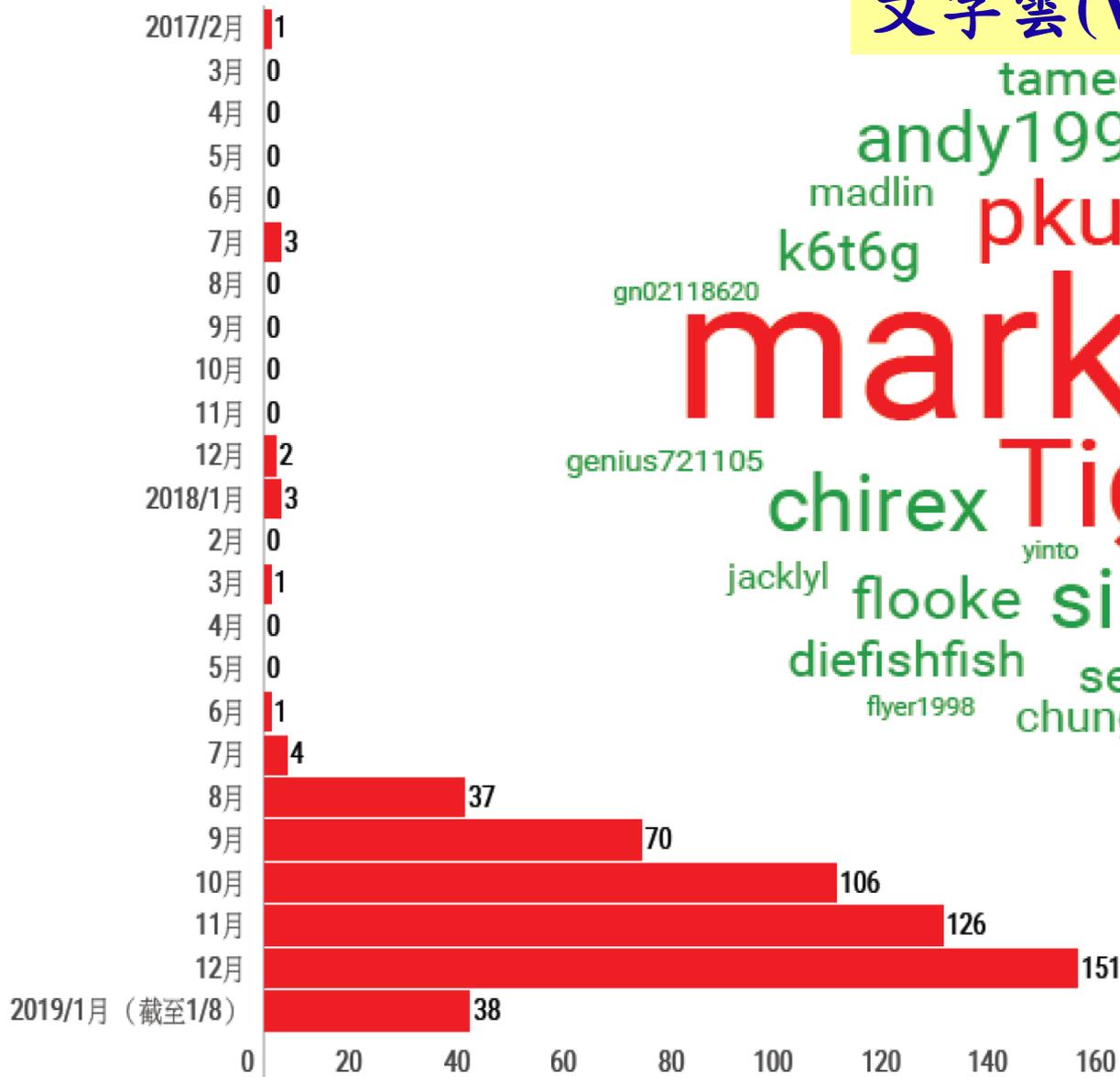
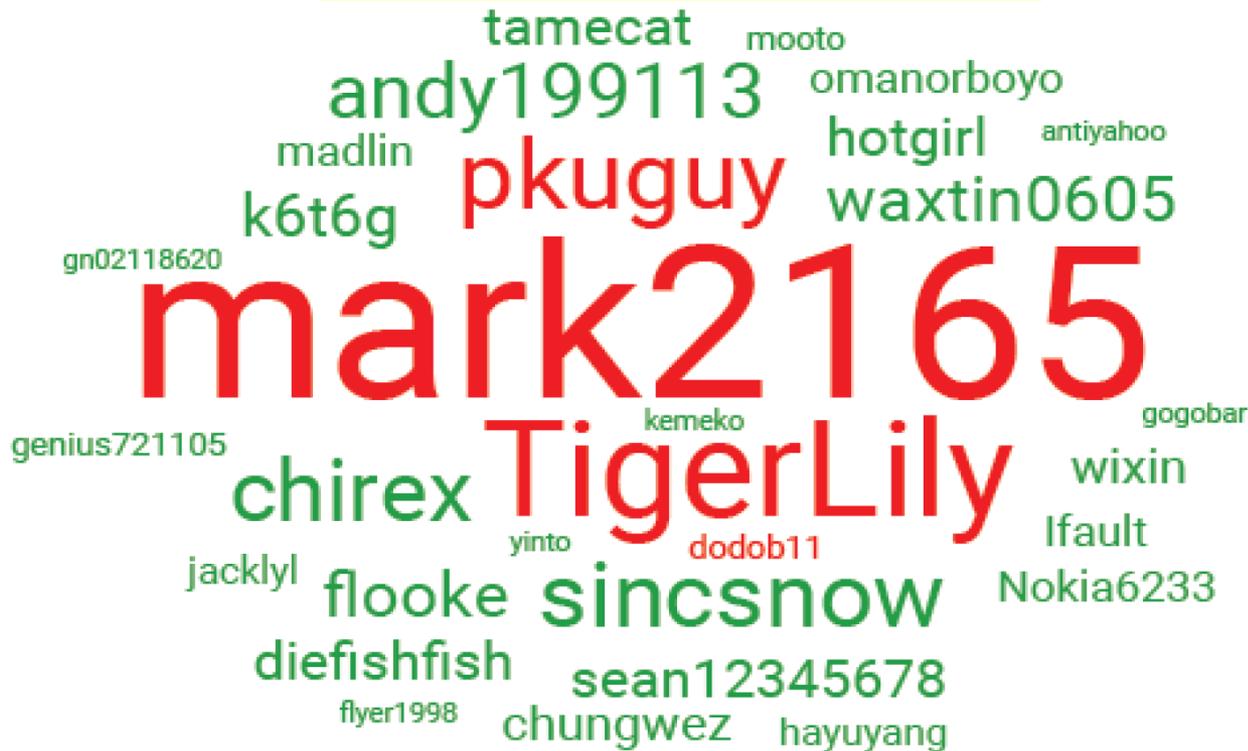
蔡英文總統就職演講稿常見雙字詞

第14任

第15任



文字雲(Word Cloud)



「韓流」現象
→何時開始、
人為操縱？

The Bible Code

OR WITH A WHITE P
 NAH A B Y O U N G M A N
 K L E S H I S G R A N D D
 D S Y E T I N G E N E R A
T H E B L O O D Y D E E D
 E R M W H A L H S H E A D
 T T O I M P O I S I B L E

Indian Prime Minister Indira
 Gandhi was killed on Oct 31, 1984

<https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRbn2zY1y0w1ND8ys4Q5BUJlfzg3QGtE02qG4gGaaeIoIQMNT0KEw>

聖經密碼
 (Torah 密碼)

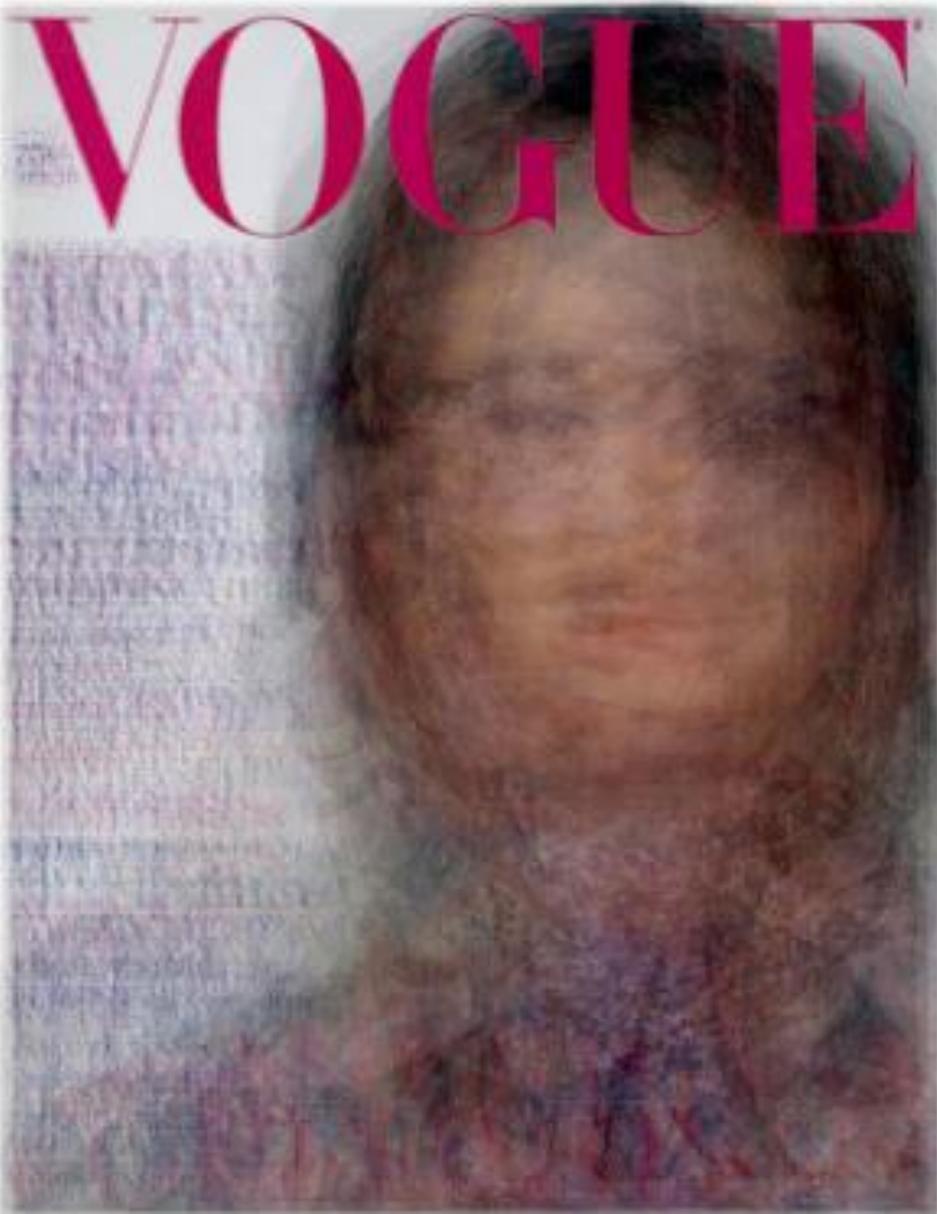
<http://hidupgila01.blogspot.tw/2015/04/>

Word of YHVH	דבר יהוה
America	אמריקה
War	מלחמה
Surrender Capitulation	כניעה
Extermination Annihilation Destruction	השמדה
Annihilation Devastation Holocaust	שואה
Lo-Ami ("not my people")	לא עמי
Death	מוות
Die	למות
Downfall Ruin Defeat	כשלה
Annihilation	כליה
Desolation	חרוה
Overthrown	היפול
Destroyer	מטחחח
Arab	ערבי
Nations Peoples	עמים
Chinese	סיני
2006	השסו
2012	השעב

耶魯大學數位人文實驗室「Robots Reading Vogue」

1970

1980





https://miro.medium.com/max/3778/1*zdoQ-oKnWAPBKbUMYYL--w.jpeg



Nov 1, 1977 Patrick Demarchelier	Oct 1, 1977 Arthur Elgort	Sep 1, 1977 Patrick Demarchelier	Aug 1, 1977 Albert Watson	Jul 1, 1977 Arthur Elgort	Jun 1, 1977 Albert Watson	May 1, 1977 Arthur Elgort	Apr 1, 1977 Richard Avedon	Mar 1, 1977 Arthur Elgort	Feb 1, 1977 Albert Watson	Jan 1, 1977 Arthur Elgort
-------------------------------------	------------------------------	-------------------------------------	------------------------------	------------------------------	------------------------------	------------------------------	-------------------------------	------------------------------	------------------------------	------------------------------



Dec 1, 1976 Arthur Elgort	Nov 1, 1976 Arthur Elgort	Oct 1, 1976 Chris von Wangenheim	Sep 1, 1976 Francesco Scavullo	Aug 1, 1976 Arthur Elgort	Jul 1, 1976 Arthur Elgort	Jun 1, 1976 Francesco Scavullo	May 1, 1976 Francesco Scavullo	Apr 1, 1976 Francesco Scavullo	Mar 1, 1976 Francesco Scavullo	Feb 1, 1976 Arthur Elgort
------------------------------	------------------------------	-------------------------------------	-----------------------------------	------------------------------	------------------------------	-----------------------------------	-----------------------------------	-----------------------------------	-----------------------------------	------------------------------



聊天機器人Chat GPT

□ 聊天機器人Chat GPT可生成類似人類風格的文字。(Generative Pre-Trained Transformer)

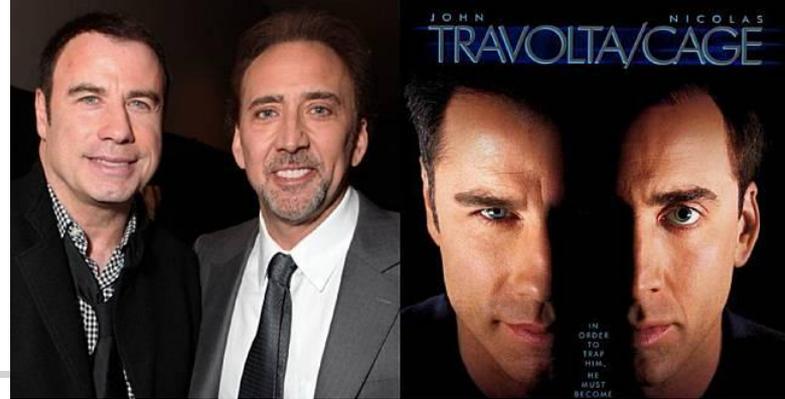
→ 「假作真時真亦假，無為有處有還無」！

商周

可以寫一段推薦《商業周刊》的話嗎？



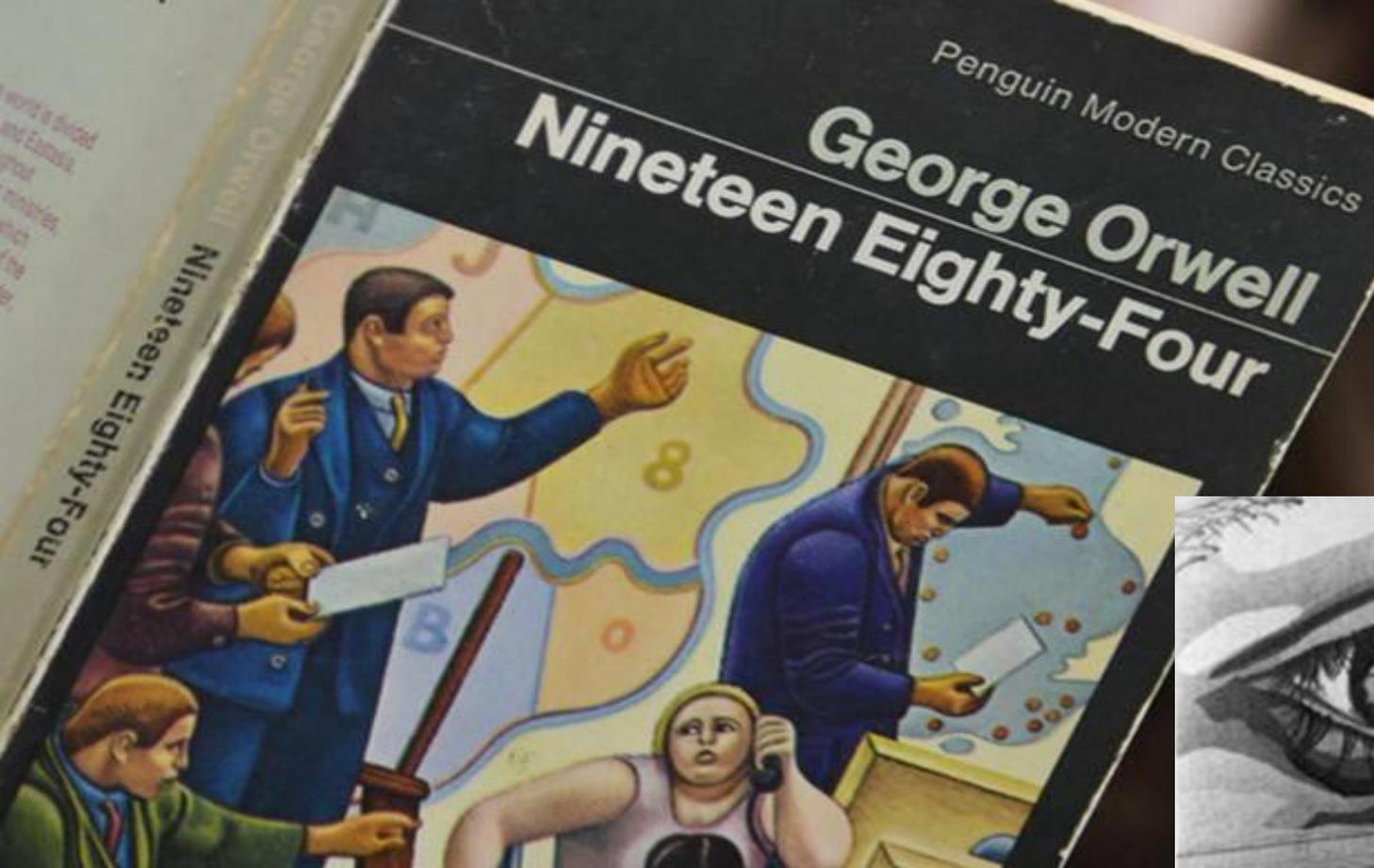
當然可以。《商業周刊》是一份權威的商業新聞雜誌，提供最新的商業資訊和分析。它的內容涵蓋了金融、科技、消費品、能源和基礎設施等領域，為讀者提供全面而及時的商業資訊。此外，《商業周刊》還有一個強大的團隊，由一流的記者和專家組成，為讀者提供深入而獨特的視角。如果您對商業新聞感興趣，那麼《商業周刊》絕對值得您一試。



深偽技術 Deepfake

- 深偽技術是「deep learning」和「fake」的混成詞，以深度學習的演算法，從現有影片資料製造出逼真的假影片。（假新聞！）



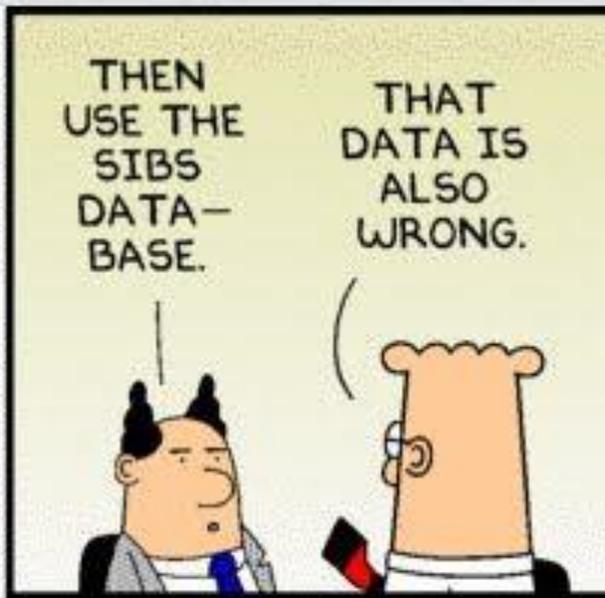


「老大哥在看著你」
(Big Brother is Watching You)





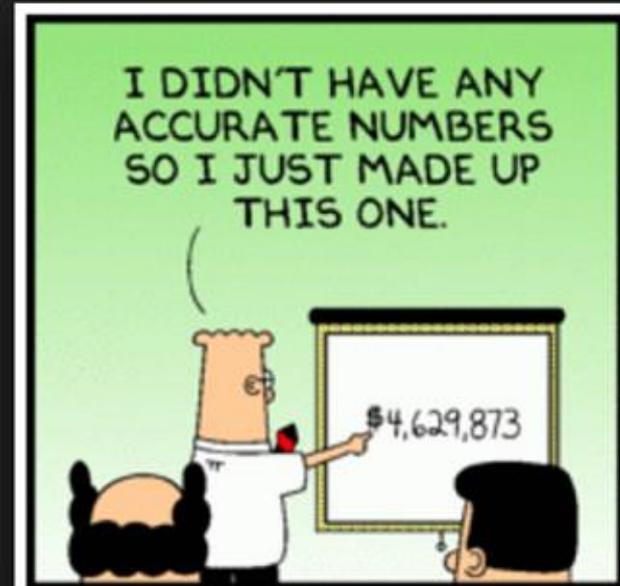
www.dilbert.com
scottadams@aol.com



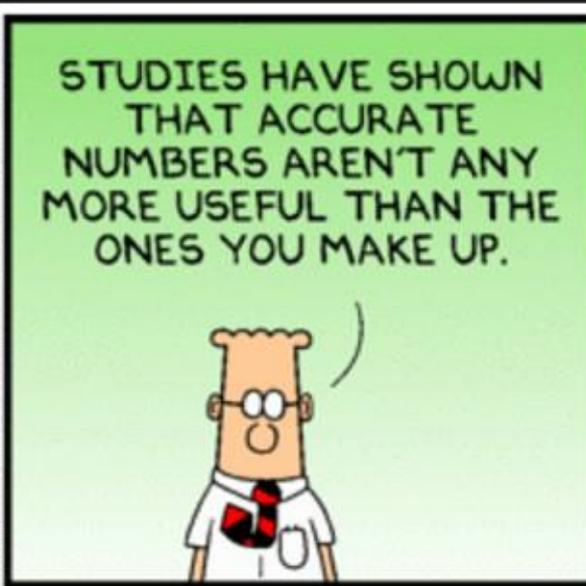
5-7-08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.



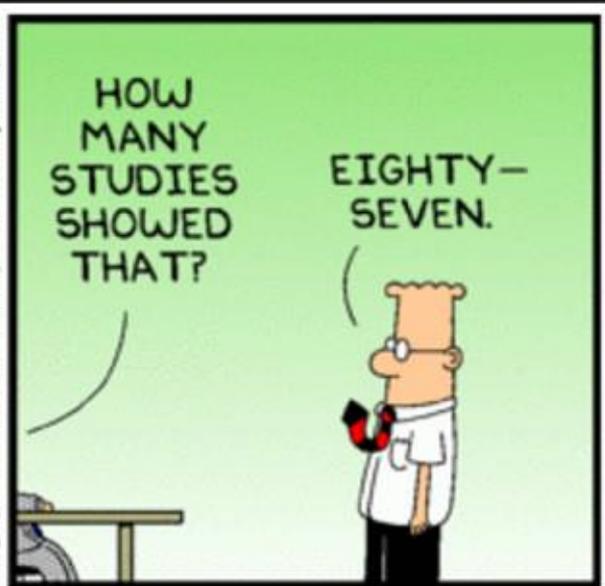
<https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.pinterest.com%2Fpin%2F167688786102533739%2F&psig=AOvVaw191BQGepIMqvj7BKEURwmp&ust=1612364346039000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKCkqrm7y-4CFQAAAAAdAAAAABAD>



www.dilbert.com
scottadams@aol.com



5-8-08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.



https://media-exp1.licdn.com/dms/image/C4E12AQEZaBITJLladA/article-cover_image-shrink_600_2000/0/1520178855524?e=1616025600&v=beta&t=VsCZNHDKnmoUy6sdD0t86DirLkdqM7rNLTGatwMZoSc

大數據與人性

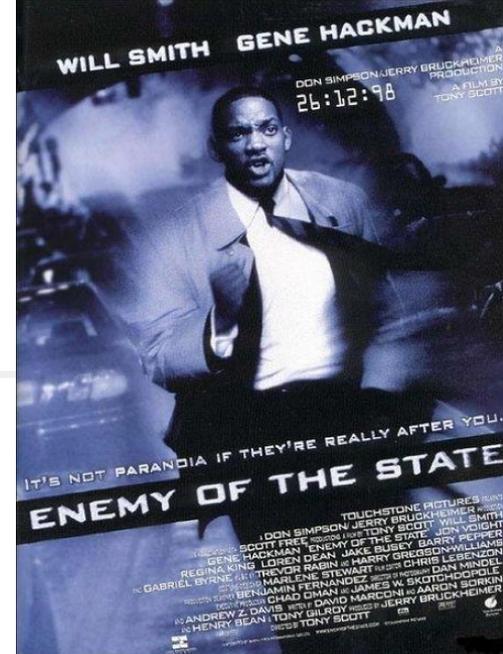


<http://cdn.marketingtechblog.com/wp-content/uploads/2013/05/Screenshot-2013-05-28-at-11.22.05-AM.png>

大數據也有黑暗面

- ❑ 資訊蒐集對隱私造成威脅（類似歐威爾的小說「1984」）
- ❑ 避免迷信因果關係而做出「預防性的懲罰」（例如：「關鍵報告」）
- ❑ 類似「訴諸權威」的誤謬，以致於形成「資料獨裁」迷戀資料分析而濫用或「大數據的傲慢與偏見」。

→ 「考試成績並不等於真正的學識」！





資料獨裁 vs. 資料驅動

To be or not to be, that is the question...



資料分享的副作用

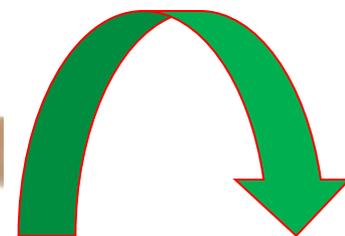
- Facebook社群軟體盛行也帶來負面效應，除了個人資料安全疑慮，害怕與人接觸的社交恐懼症 (sociophobia)日趨嚴重。
 - 擔心自己遭受社會遺忘或淘汰；
「滿紙荒唐言，一把辛酸淚！都云作者痴，誰解其中味？」
 - 人際關係的疏離（虛擬vs.實際）；
「假作真時真亦假，無為有處有還無」
 - 私刑（人肉搜索）、扮演法官。

科技來自於或是改變人性？

- 「科技始終來自於人性」是Motorola知名的廣告用詞，強調其產品根據人性因素設計。
- 科技也改變了我們的生活，但是否會改變人們的思維模式、甚至人性？



NOKIA
Connecting People



https://scontent-lga3-1.cdninstagram.com/v/t51.2885-15/sh0.08/e35/s640x640/70025765_480213932560518_6084948241454488487_n.jpg?_nc_ht=scontent-lga3-1.cdninstagram.com&_nc_cat=106&_nc_ohc=nGAHr_5D4wEAX-ZkqE8&oh=c1b19989c814908e98200e4b0759d6c5&oe=5EA54259

