

「商業資料分析與管理決策」 ——統計與大數據

Spring 2024

教師：統計系余清祥

內容：統計與生活

下載：<http://csyue.nccu.edu.tw>

Email: csyue@nccu.edu.tw

2024年3月3日



什麼是大數據？



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）

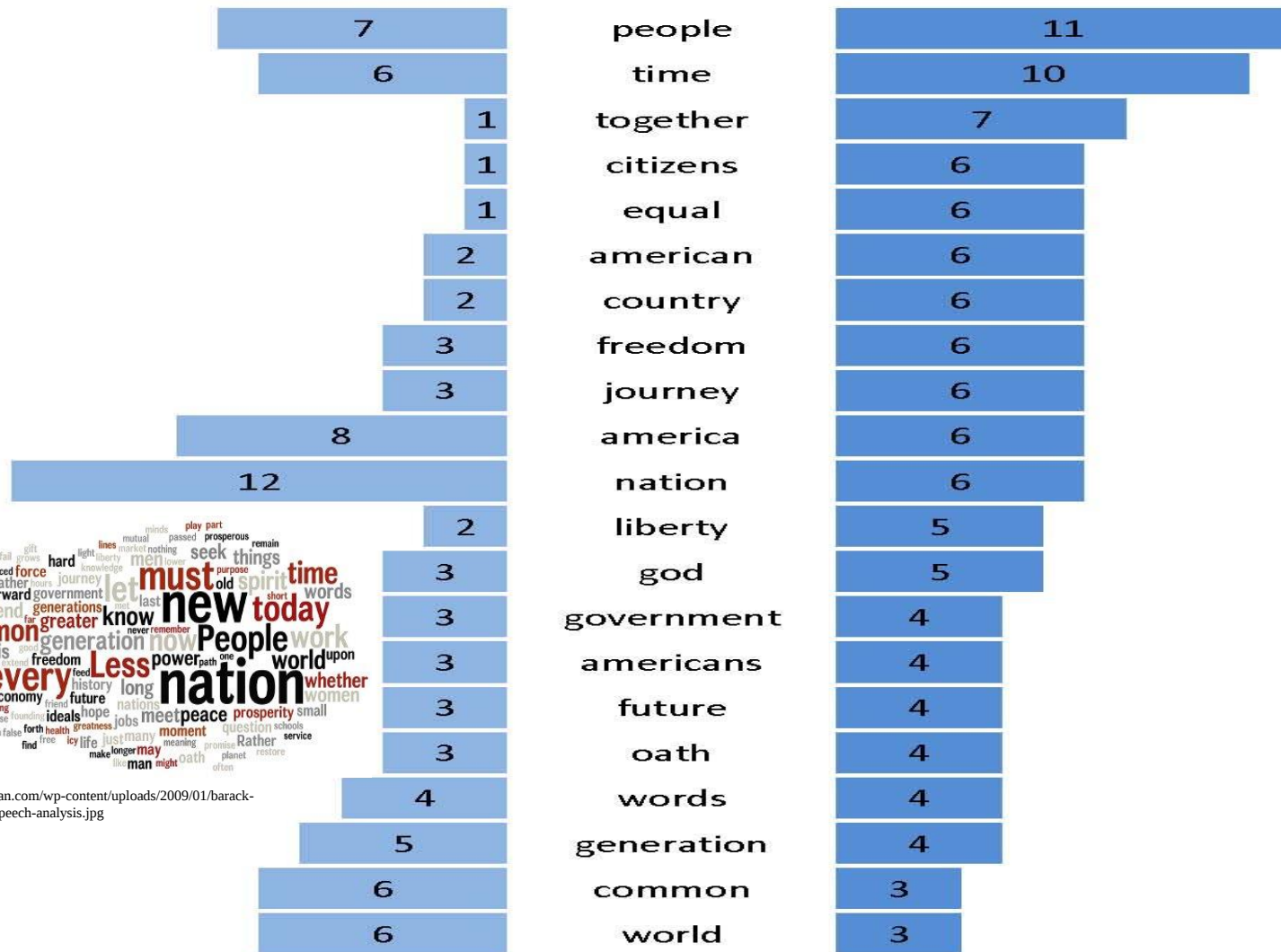
註：加入資料分享與傳遞(Visible)後成為「大、快、雜、疑、轉」；也有人加入價值(Value)成為新的5V。

電腦容量單位的演變（資料爆炸！）

單位	縮寫	意義
Bit	b	1 or 0
Byte	B	8 Bits
Kilobyte	KB	1,024 Bytes
Megabyte	MB	1,024 KB
Gigabyte	GB	1,024 MB
Terabyte	TB	1,024 GB
Petabyte	PB	1,024 TB
Exabyte	EB	1,024 PB
Zettabyte	ZB	1,024 EB
Yottabyte	YB	1,024 ZB

Comparing Inaugural Addresses

■ 2009 ■ 2013



<http://sixminutes.dlugan.com/wp-content/uploads/2009/01/barack-obama-inauguration-speech-analysis.jpg>

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

第14任蔡英文總統就職演講稿最常出現字詞

排序	單字			雙字詞		
	類別	次數	頻率	類別	次數	頻率
1	的	293	5.48%	我們	86	2.012%
2	我	114	2.13%	台灣	41	0.959%
3	們	90	1.68%	政府	37	0.866%
4	一	75	1.40%	國家	32	0.749%
5	會	74	1.38%	一個	29	0.679%
6	是	70	1.31%	新政	27	0.632%
7	個	66	1,23%	經濟	27	0.632%
8	民	63	1.18%	這個	25	0.585%
9	人	59	1.10%	民主	24	0.562%
10	國	59	1.10%	社會	22	0.515%

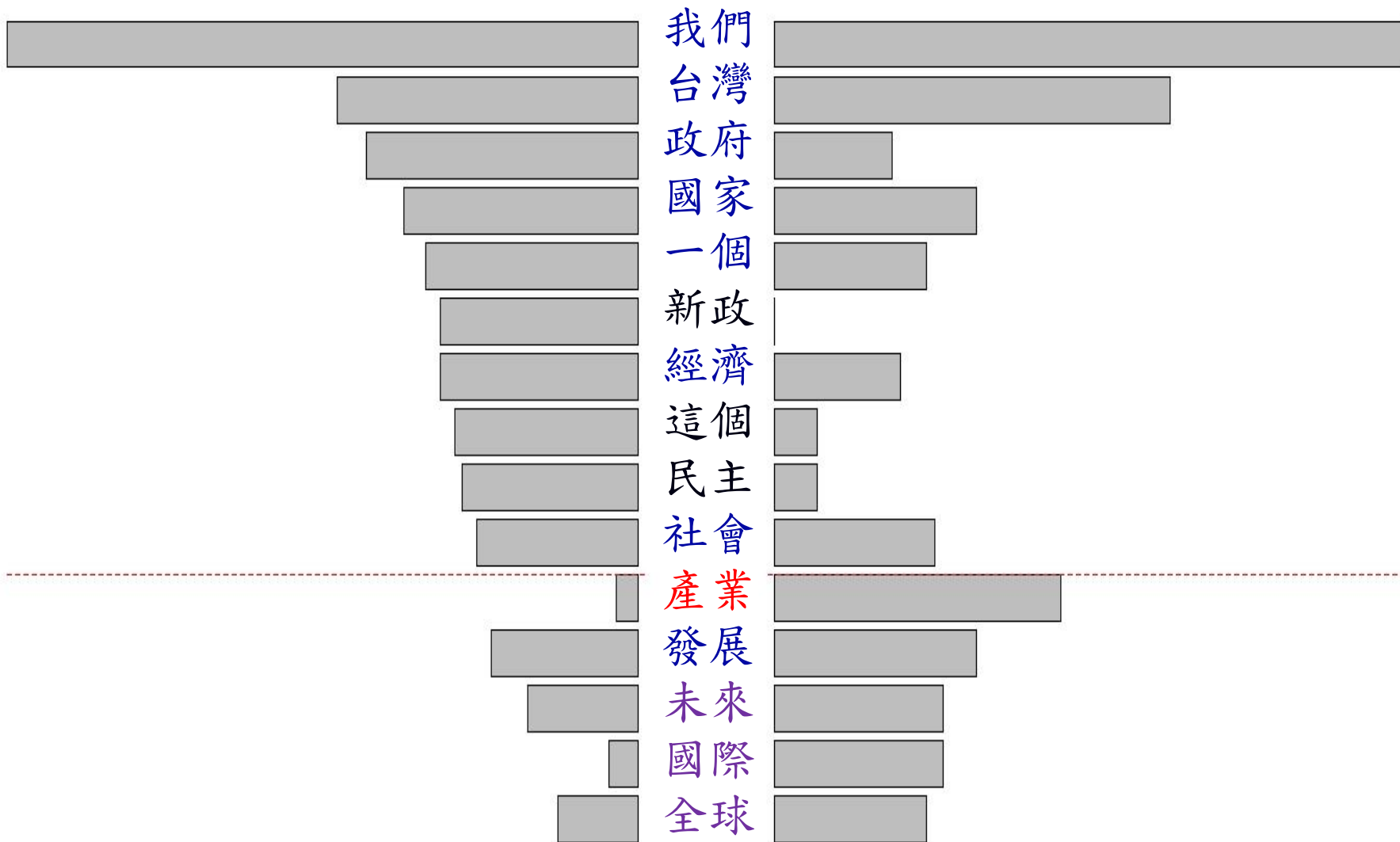
第15任蔡英文總統就職演講稿最常出現字詞

排序	單字			雙字詞		
	類別	次數	頻率	類別	次數	頻率
1	的	257	4.94%	我們	75	2.590%
2	我	114	2.10%	台灣	47	1.620%
3	們	92	1.77%	產業	34	1.170%
4	國	79	1.52%	國家	24	0.830%
5	人	68	1.31%	發展	24	0.830%
6	會	65	1.25%	未來	20	0.690%
7	在	63	1,21%	國際	20	0.690%
8	一	62	1.19%	社會	19	0.660%
9	是	55	1.06%	全球	18	0.620%
10	要	53	1.02%	一個	18	0.620%

蔡英文總統就職演講稿常見雙字詞

第14任

第15任



臺灣報紙頭條新聞的用字比較(2012~2019年)

《蘋果日報》



卡方指標



《中國時報》

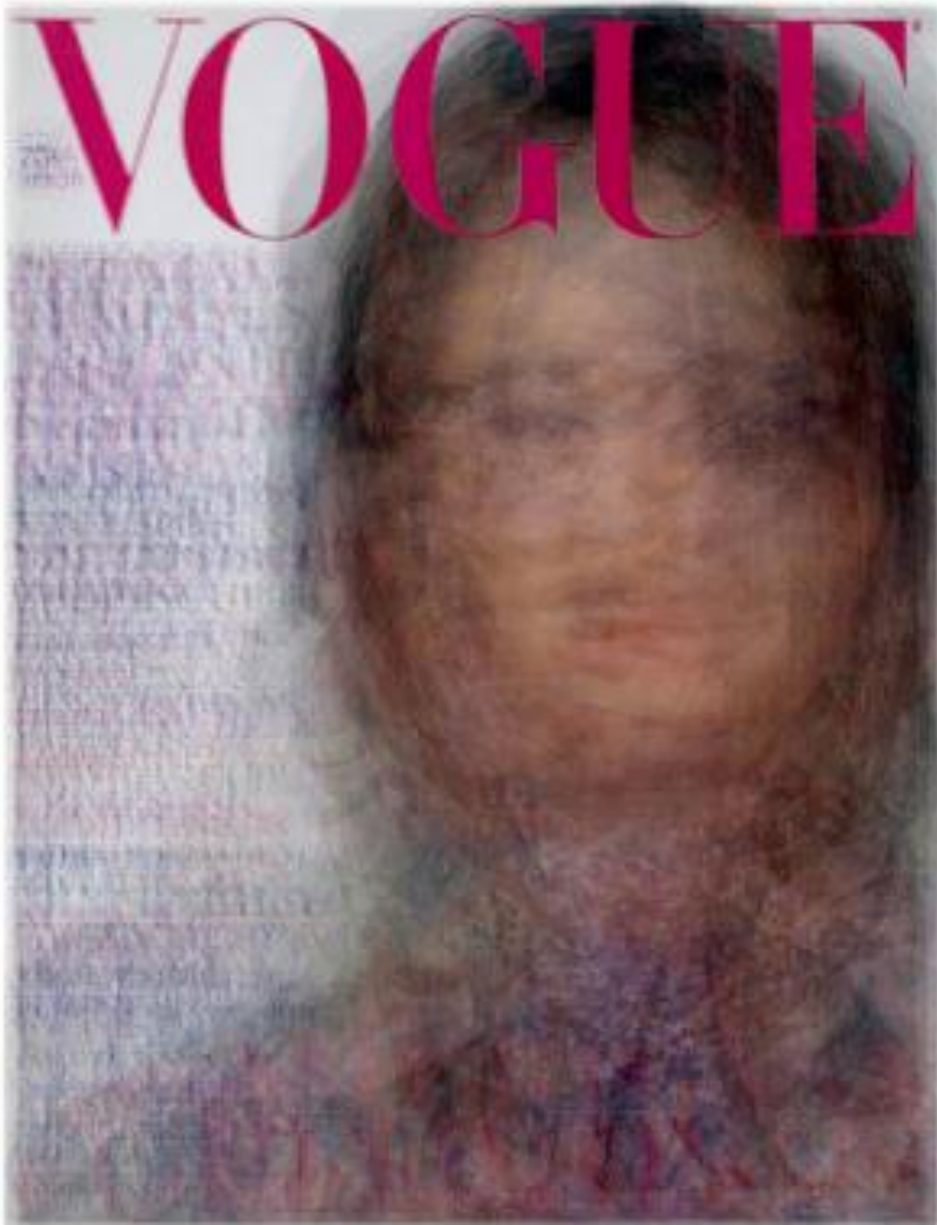


《自由時報》

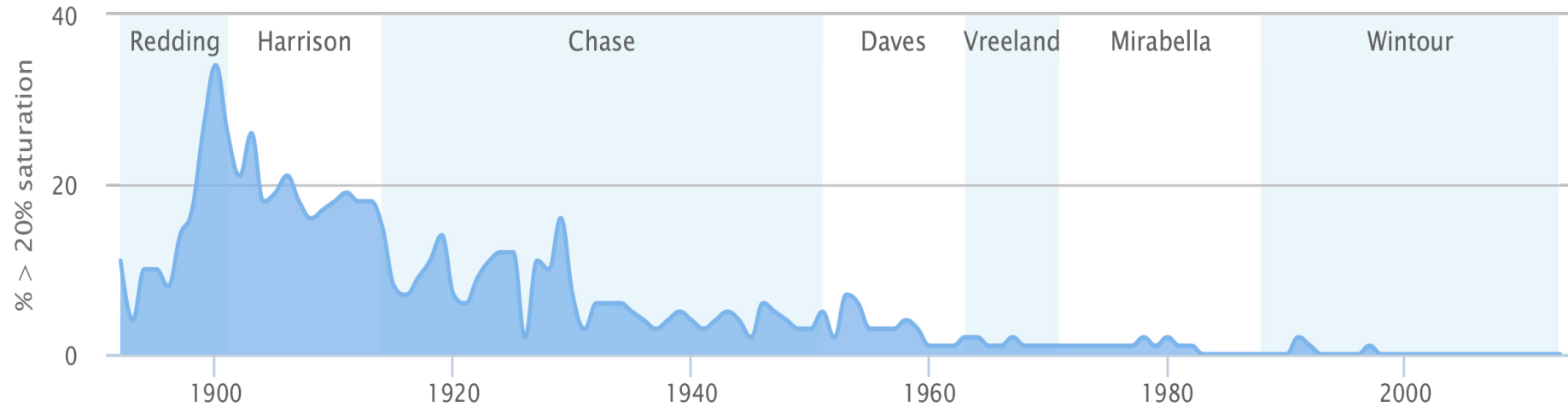
耶魯大學數位人文實驗室「Robots Reading Vogue」

1970

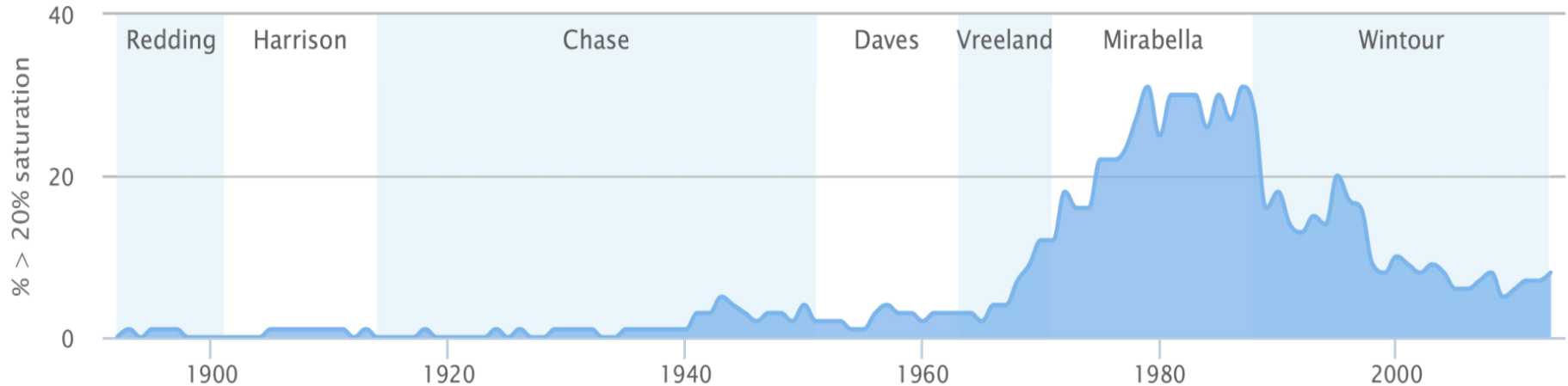
1980



Dressmaking over Time



Women's Health over Time



Vogue雜誌的風格趨勢變化



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

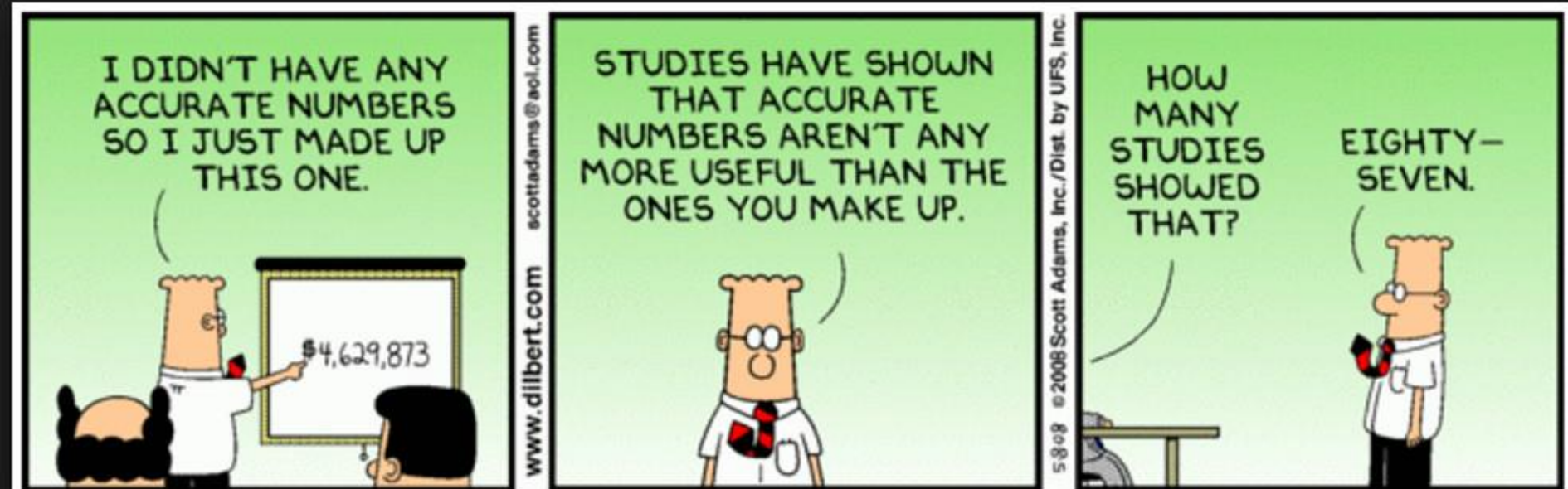


- 深偽技術是「deep learning」和「fake」的混成詞，以深度學習的演算法，從現有影片資料製造出逼真的假影片。（假新聞！）



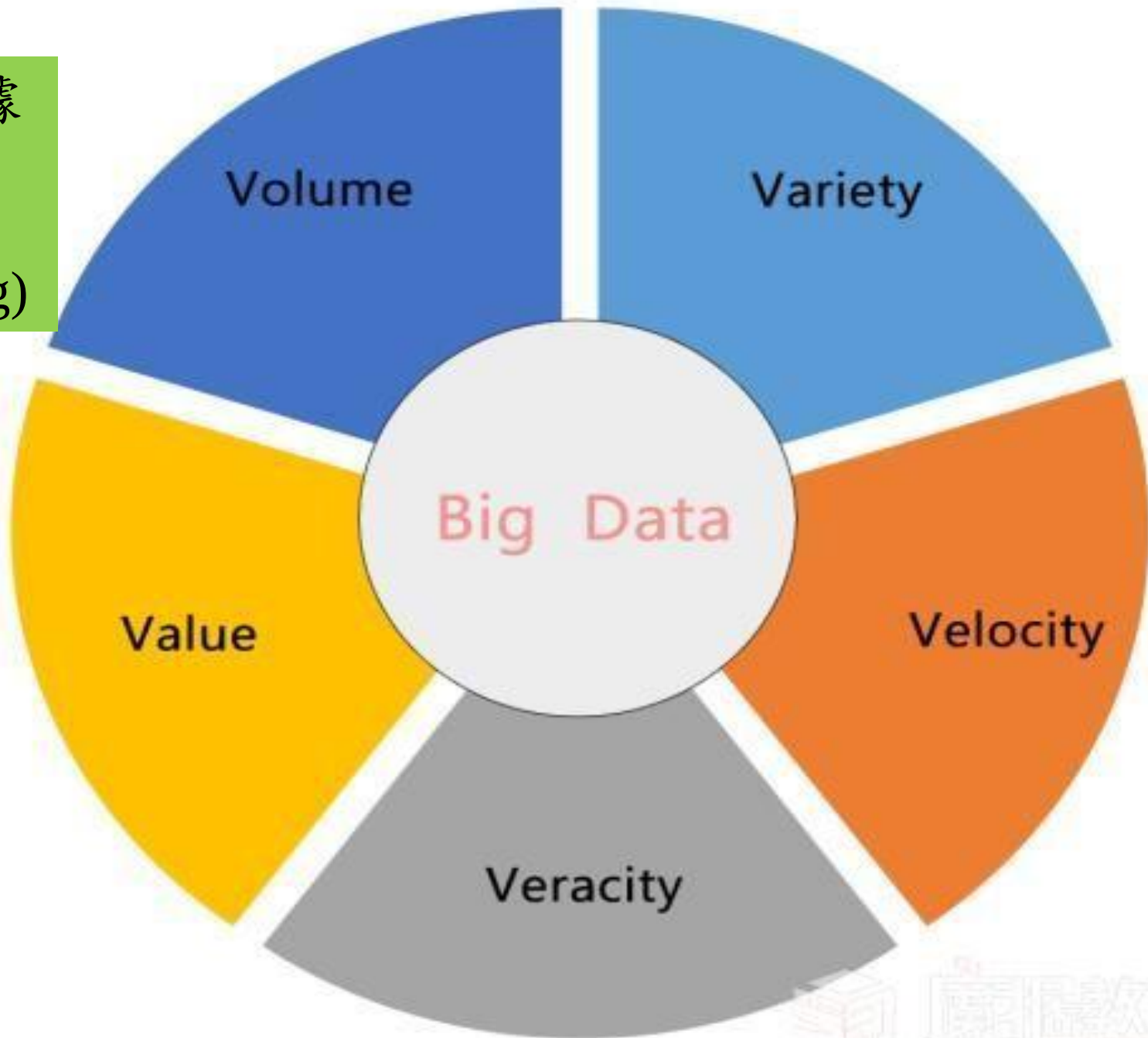


<https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.pinterest.com%2Fpin%2F167688786102533739%2F&psig=AOvVaw191BQGepIMqvj7BKEURwmp&ust=1612364346039000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKCkqrm7y-4CFQAAAAAdAAAAABAD>



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

Value：數據
價值密度相
對較低。
(Datamining)



資訊的價值



- 1990年代後期Amazon網站雇用十幾位書評及編輯，提供推薦閱讀的書單，其銷售量卻比不上以讀者回饋產生的建議，最後解散書評團隊。（現在1/3業績來自於電腦推薦。）
- 2004美國Walmart開始運用歷史交易記錄，發現颶風來臨前，手電筒、小甜點Pop-Tarts銷售量大增。（註：「尿布與啤酒」是另一個知名範例！）

大數據的特性



大數據的幾個基本特性

■ 除了數量龐大，有人認為大數據還有以下幾個特性：

→ 「樣本 = 母體」

→ 不精確且含有雜訊的資料

→ 相關性而非因果關係

註：你/妳認同哪些特性？



關連性vs.因果關係

- 數量分析無法決定因果關係，小學生腳丫大小與拼字能力，可能存在潛在變數(Latent Variable)。
- 巨量資料及電腦科技的進步，使得尋找關連性更為可行。
 - 美國包裹快遞公司(UPS)從2000年開始實施預防性維修(Preventive Maintenance)；
 - 紐約市根據過去人孔蓋維修記錄，預測哪些地方有較高的損壞機會(2010 Wired)。



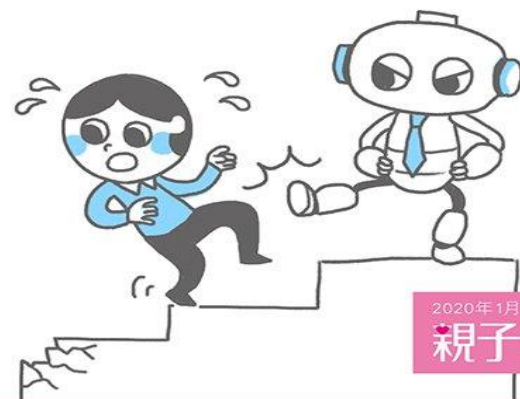
人工智能方向是不是搞错了
本来想是AI帮我们洗碗扫地
我们就有空去写诗画画
现在AI开始写诗画画了
而我们还在洗碗扫地

https://www.popo8.com/host/data/202402/20/18/p1708479682_448type_jpeg_size_660_85_end.jpg_b.jpg

<https://www.parenting.com.tw/article/5080979>

⑦ AI 帶來很多挑戰和反思

許多人擔心AI什麼都會，
下棋贏過世界棋王、會寫詩、
寫新聞、挑咖啡豆。
那以後是不是工作都會被取代？
人還有什麼價值呢？



2020年1月號

親子天下



大數據的分析考量

- 分析大數據需兼顧兩個層面：技術、知識
(Technical Expertise Component & Functional Knowledge Component)，涵蓋以下四個項目：
 - Data Acquisition；（領域相關知識）
 - Data Processing；（資訊量化）
 - Data Management；（SQL等資料管理）
 - Data Mining/Analysis.（資料採礦）

資料分享與傳遞（特性之一）

- 全臺灣有 2172 萬的網路用者，約 91% 人口；活躍的社群平台用戶高達 2135 萬，約 89.4%。
- 每日上網 8 小時 7 分鐘，1/4 用於社群媒體。

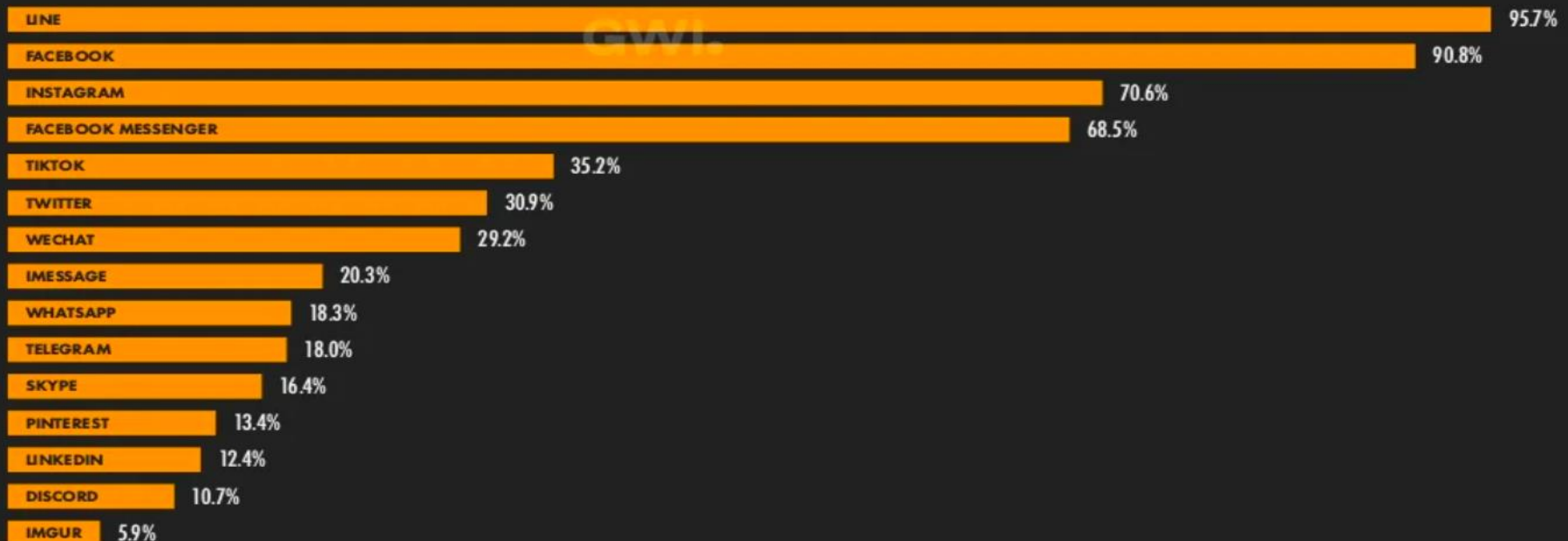
FEB
2022

MOST-USED SOCIAL MEDIA PLATFORMS

PERCENTAGE OF INTERNET USERS AGED 16 TO 64 WHO USE EACH PLATFORM EACH MONTH



TAIWAN



JAN
2023

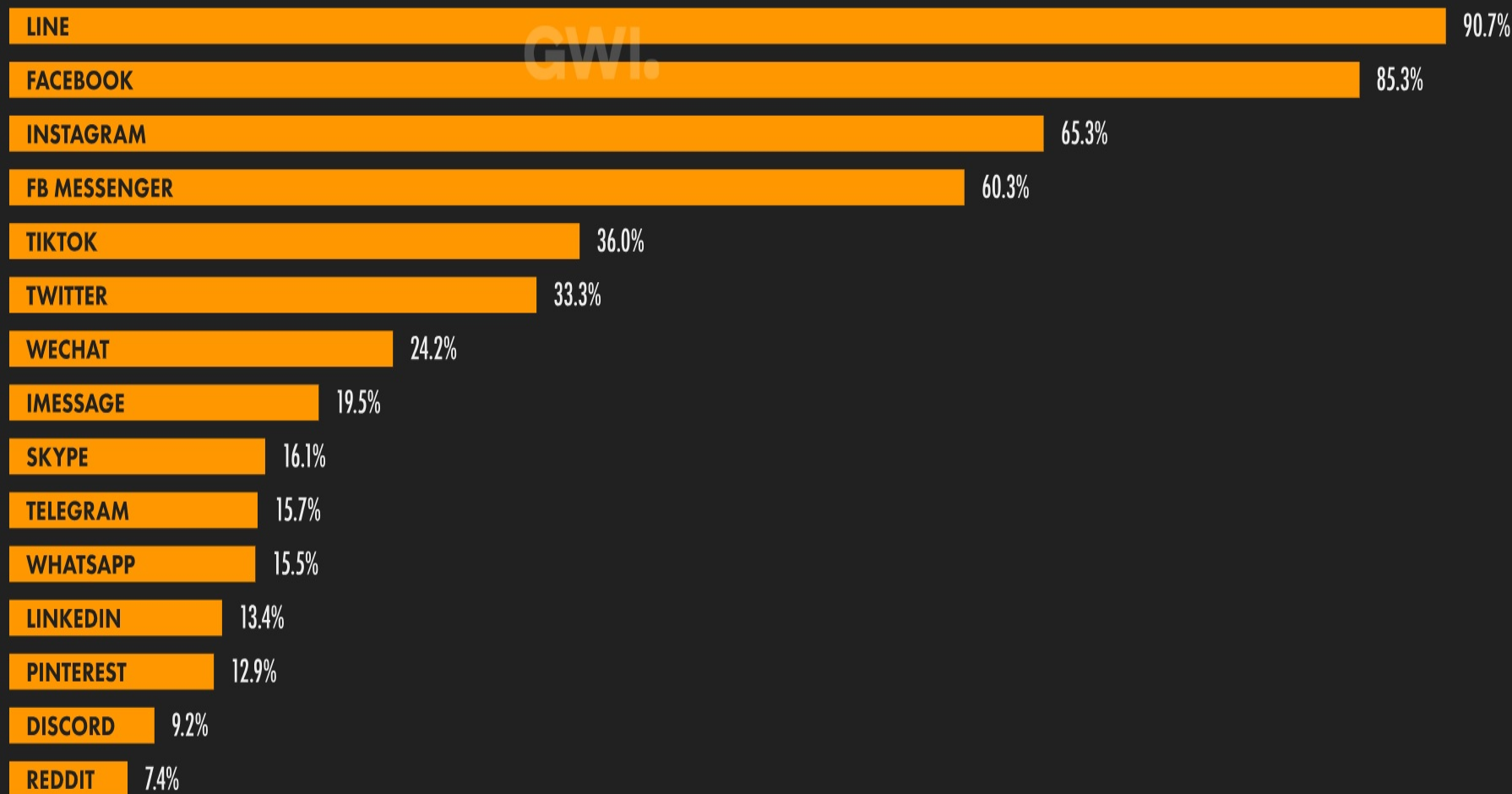
MOST USED SOCIAL MEDIA PLATFORMS

PERCENTAGE OF INTERNET USERS AGED 16 TO 64 WHO USE EACH PLATFORM EACH MONTH

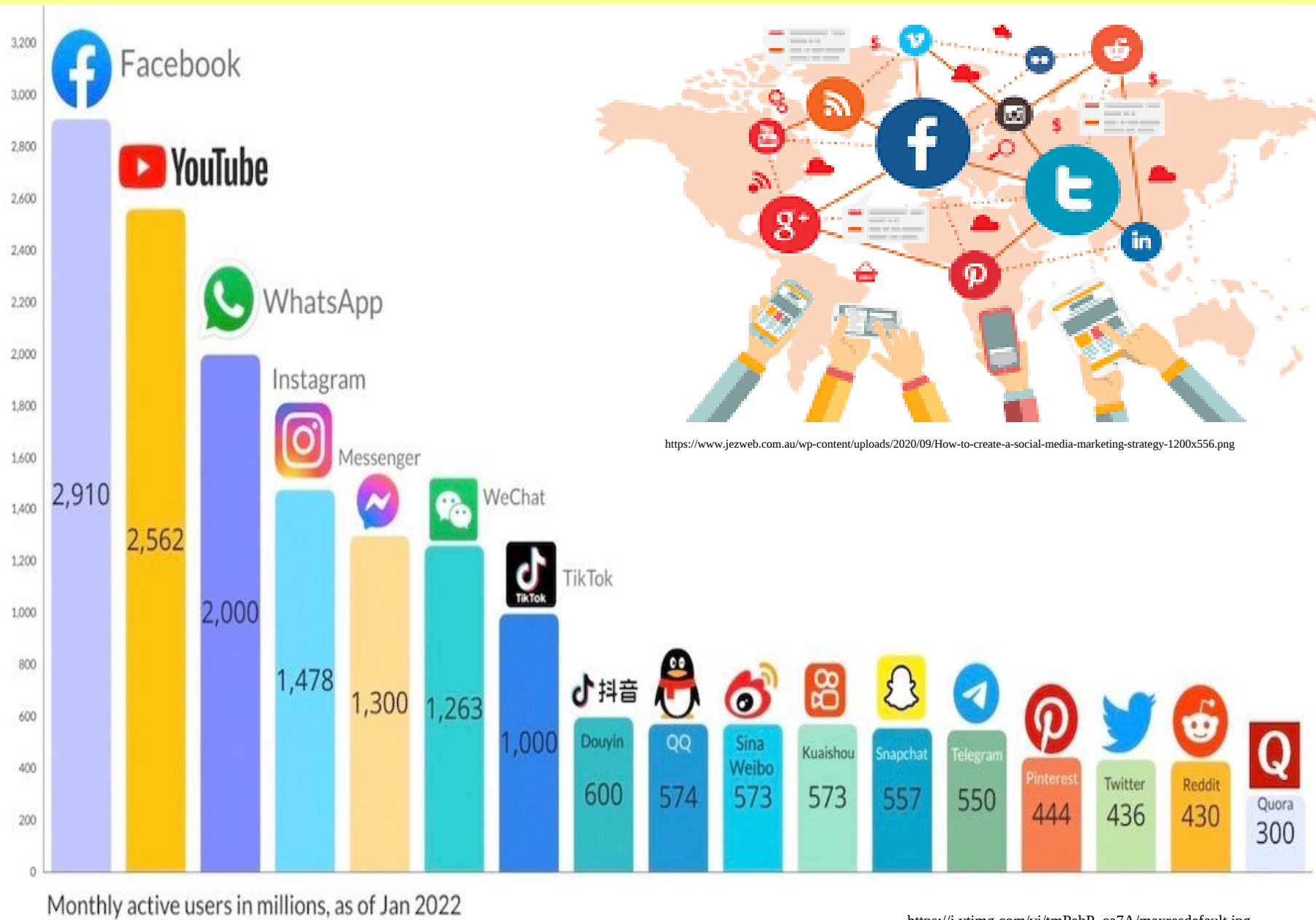
NOTE: YOUTUBE IS NOT OFFERED AS AN ANSWER OPTION FOR THIS QUESTION IN GWI'S SURVEY, SO IT WILL NOT APPEAR IN THIS RANKING



TAIWAN



全球最常使用社群軟體(2022)



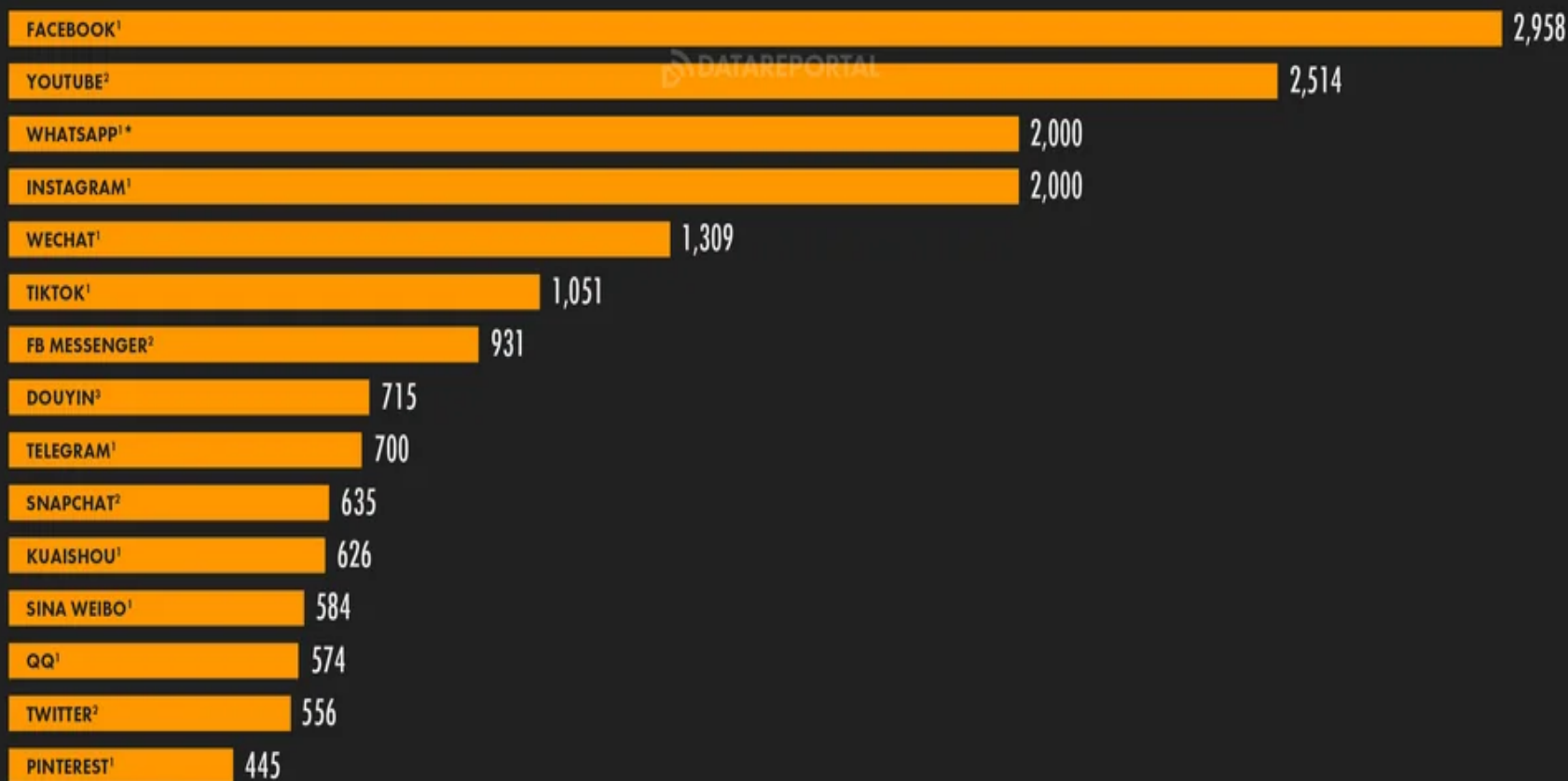
JAN
2023

THE WORLD'S MOST USED SOCIAL PLATFORMS

RANKING OF SOCIAL MEDIA PLATFORMS BY GLOBAL ACTIVE USER FIGURES (IN MILLIONS)



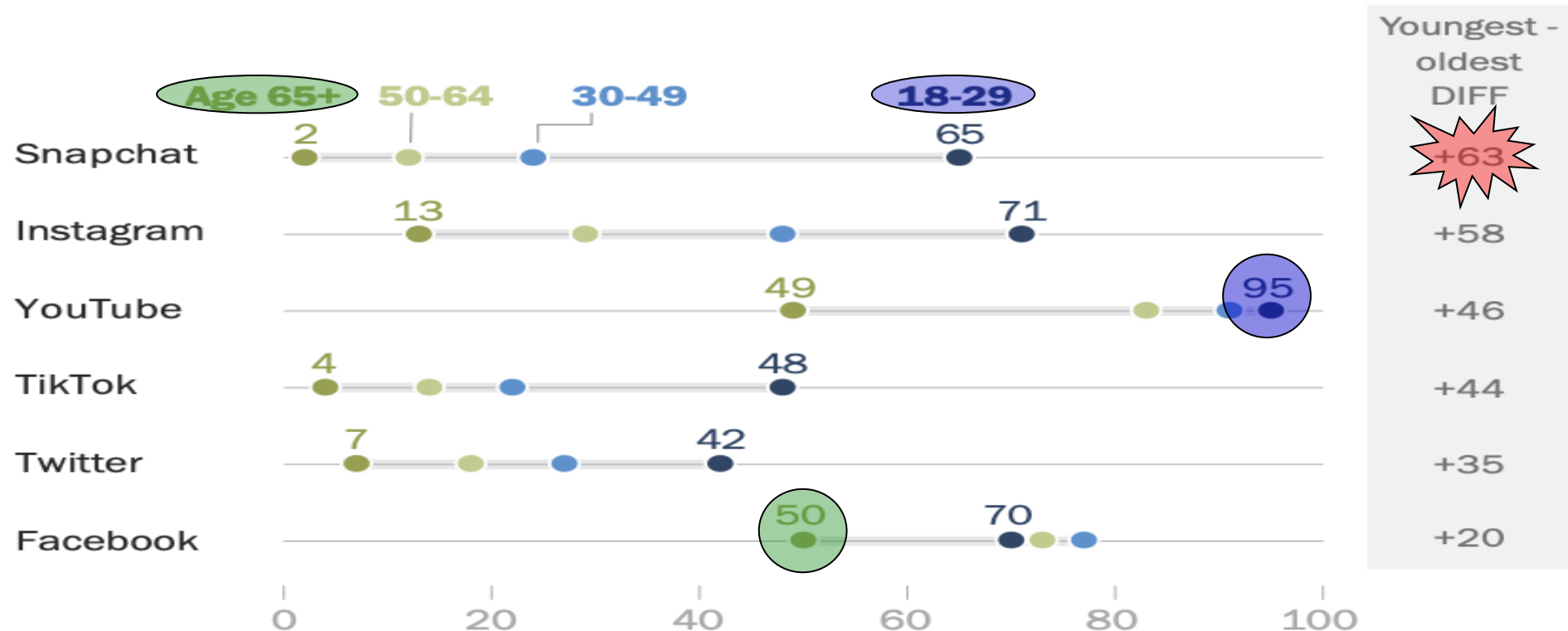
GLOBAL OVERVIEW



社群軟體也有代溝？

Age gaps in Snapchat, Instagram use are particularly wide, less so for Facebook

% of U.S. adults in each age group who say they ever use ...



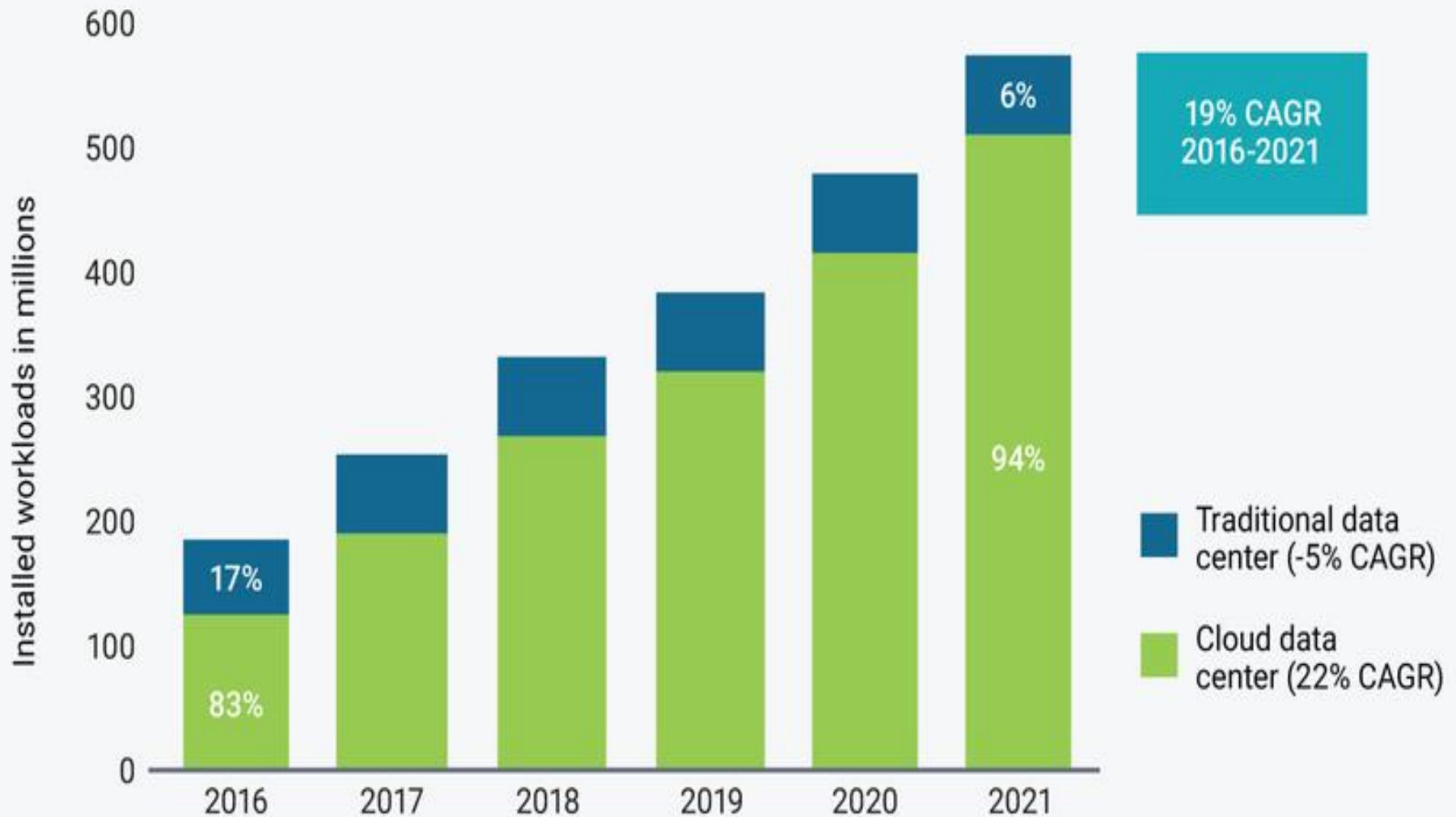
Note: All differences shown in DIFF column are statistically significant. The DIFF values shown are based on subtracting the rounded values in the chart. Respondents who did not give an answer are not shown.

Source: Survey of U.S. adults conducted Jan. 25-Feb. 8, 2021.

"Social Media Use in 2021"

PEW RESEARCH CENTER

雲端儲存資料更為普遍

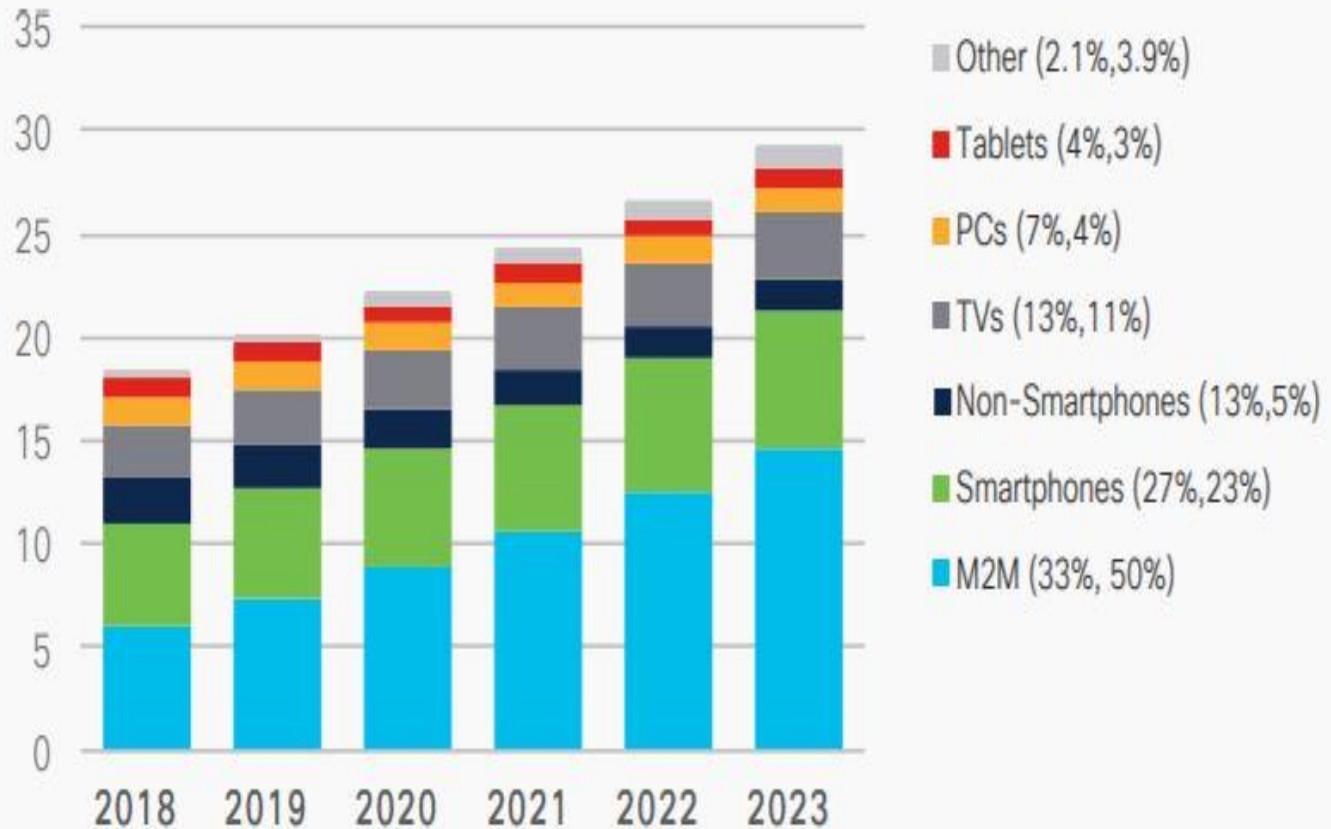


Source: Cisco Global Cloud Index, 2016-2021.

未來趨勢：機器對機器。。

10% CAGR
2018-2023

Billions of
Devices



* Figures (n) refer to 2018, 2023 device share

機器對機器 (Machine to machine, M2M)

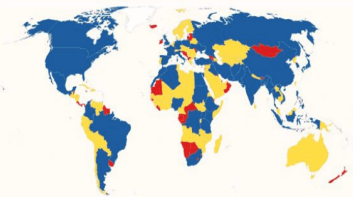
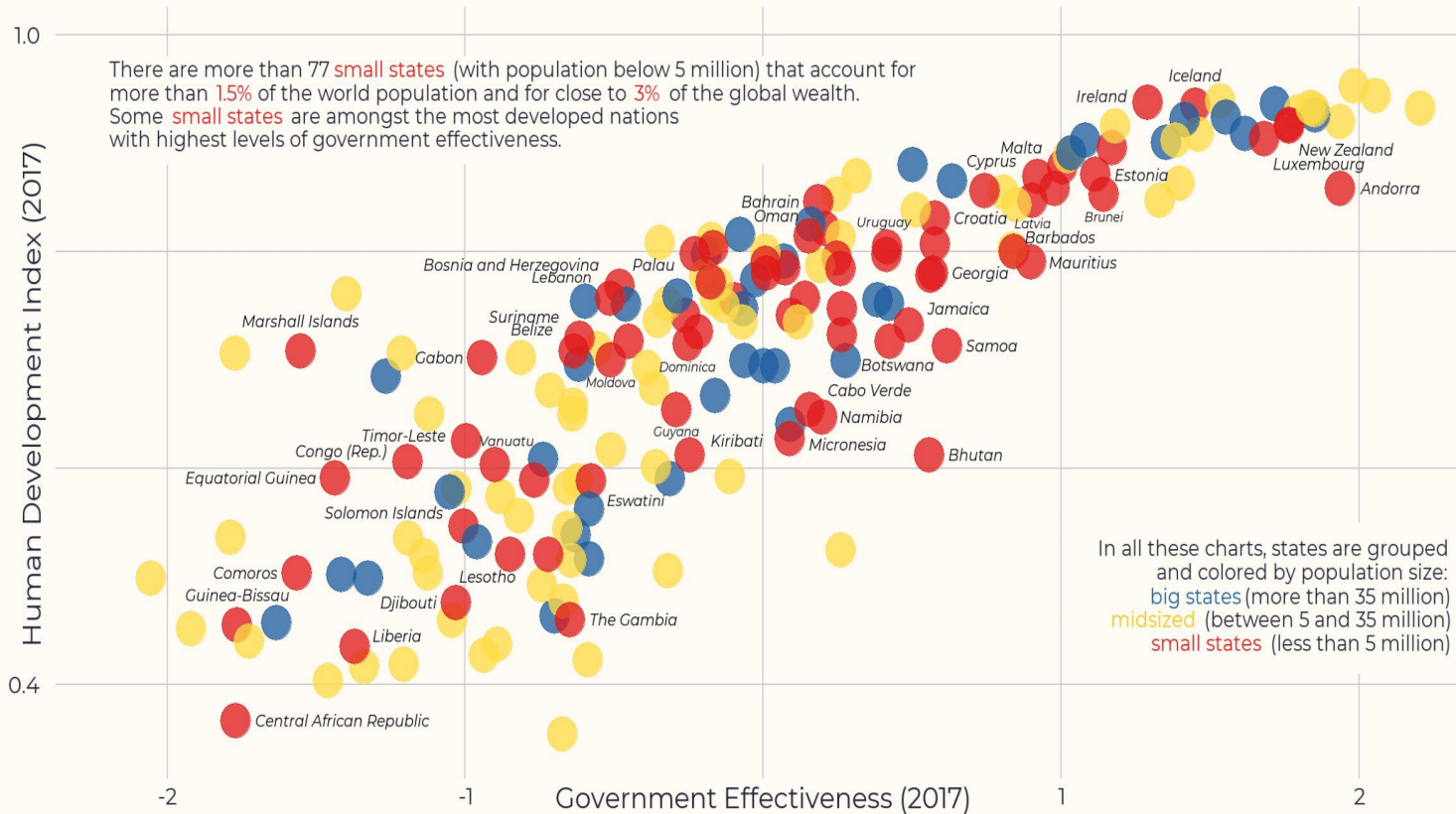
機器裝置間在無需人為干預下，直接透過網路溝通而自行完成任務。

Source: Cisco Annual Internet Report, 2018-2023

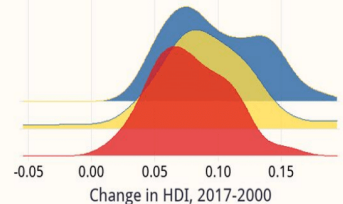
<https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>

第五個V：資料視覺化(Visualization)

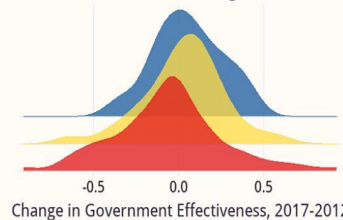
Small States Can Be Big Players in Development and Good Governance



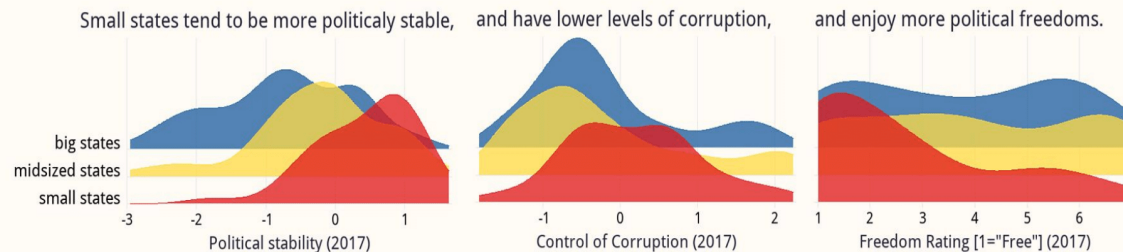
But small states have improved less



and have even lost some ground.



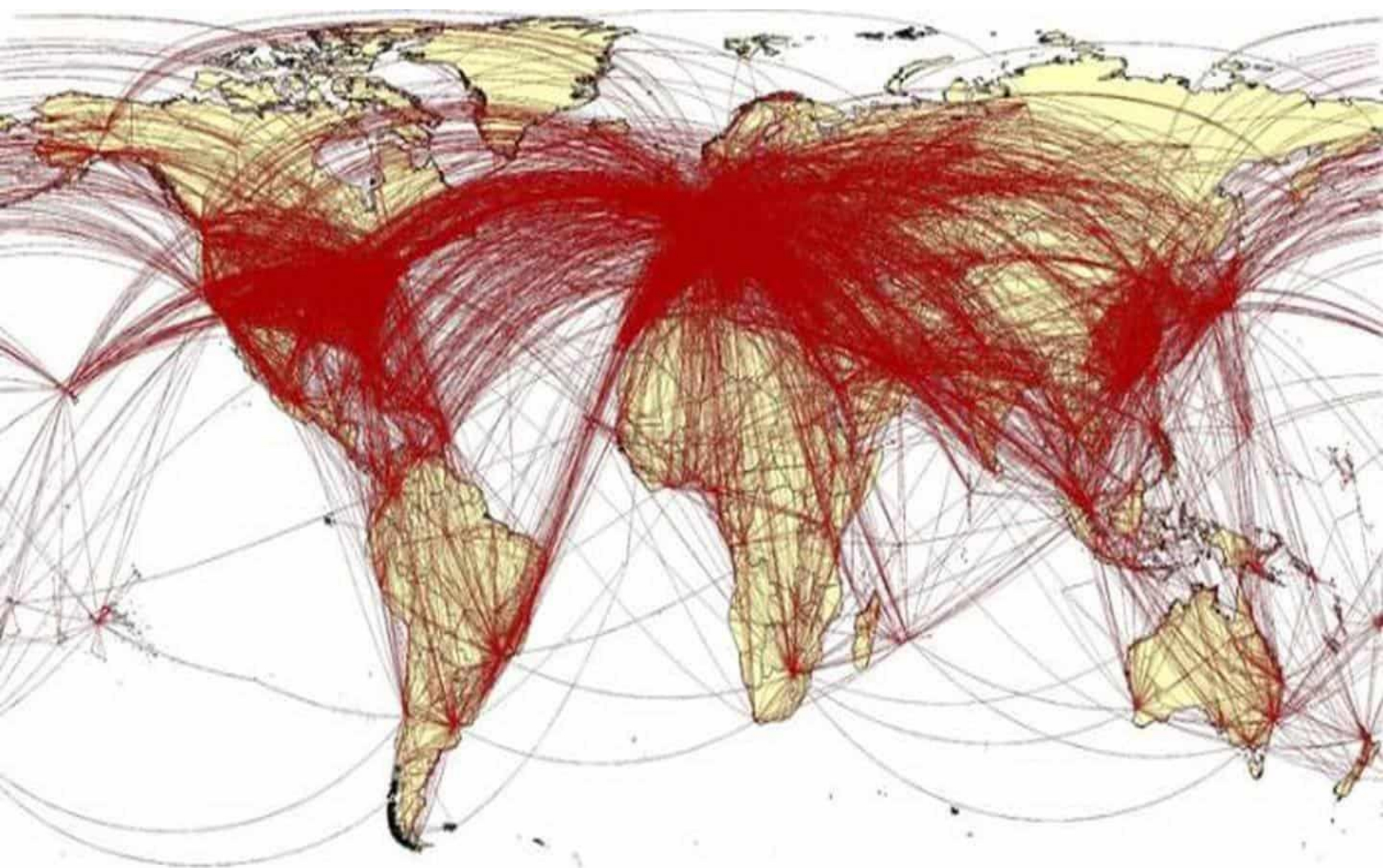
These charts show the distribution densities of different variables by three groups based on state size:



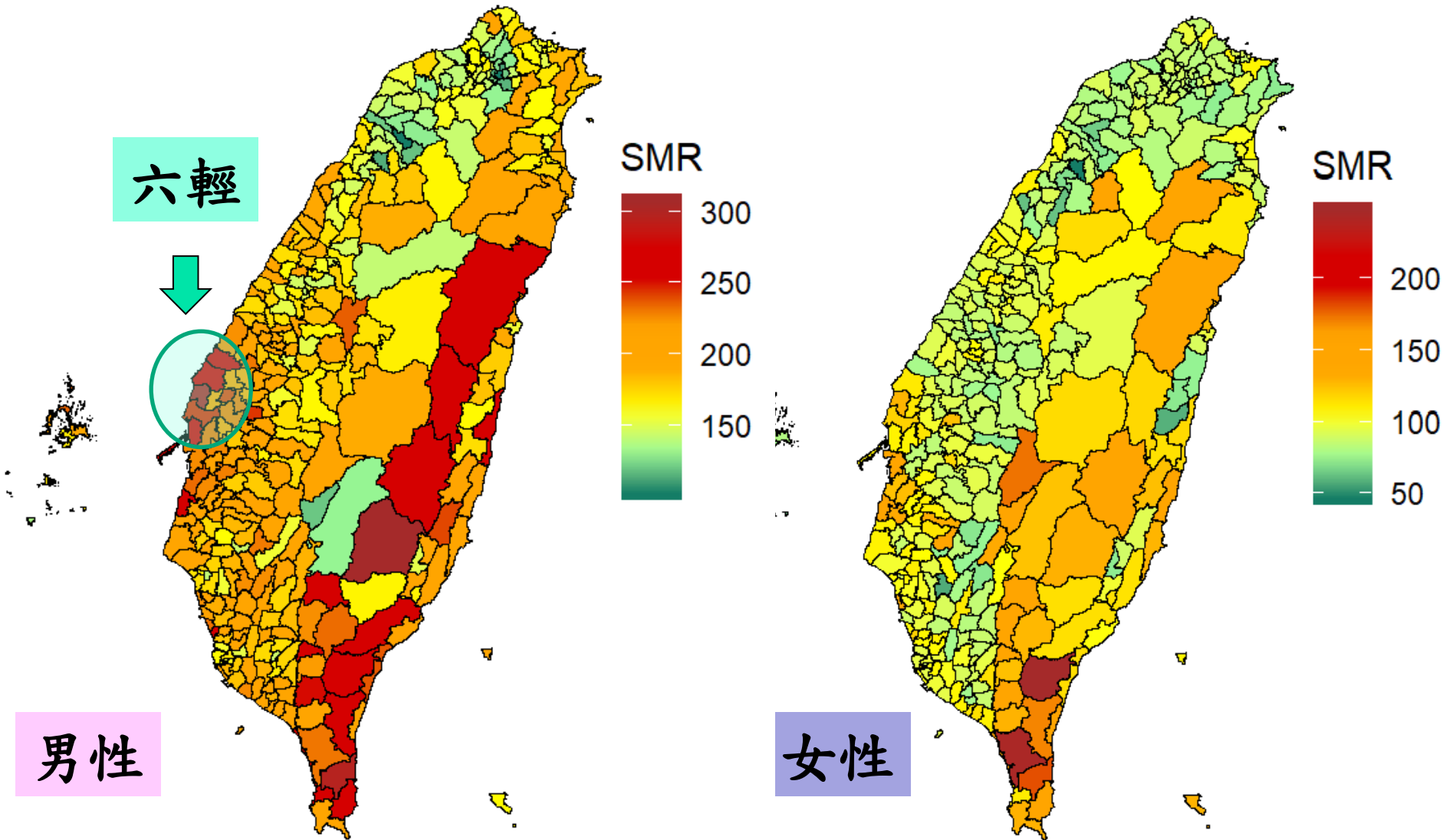
الجائزة العالمية
لفن عرض للبيانات
WORLD DATA
VISUALIZATION PRIZE

Data: Human Development Index from Human Development Reports. Government Effectiveness, Control of Corruption, and Political Stability from Worldwide Governance Indicators. Freedom Rating from Freedom House.

武漢大逃亡 Worldpop 分析路線圖

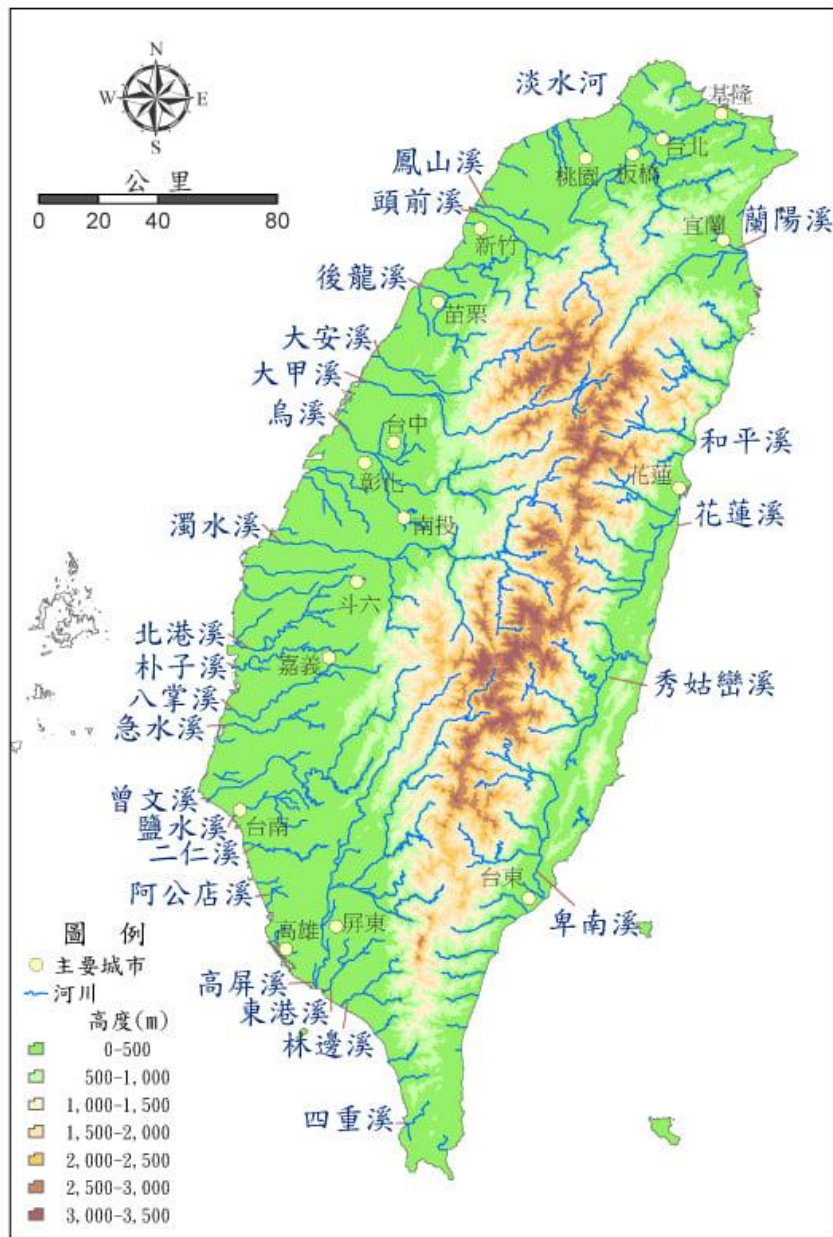


台灣2014~18年鄉鎮市區癌症死亡SMR



註：標準死亡比(SMR, Standard Mortality Ratio)可視為死亡率比值。

台灣主要河川分布圖



資料來源:經濟部水利署
地圖製作:余紀忠文教基金會 林書楷
中華民國100年9月

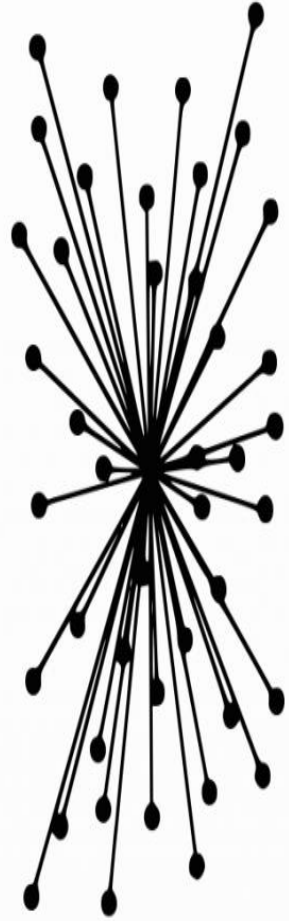
癌症地圖

癌症集中在河流下游

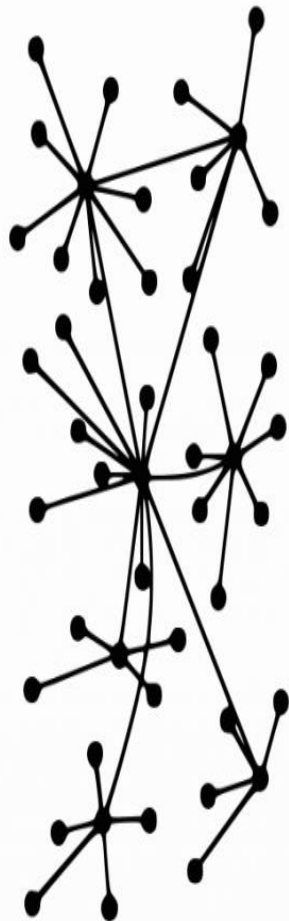
- 肝癌
- 胃癌
- 肺癌
- 腎臟癌
- 膀胱癌
- 子宮頸癌
- 乳癌
- 血癌
- 結腸直腸癌
- 攝護腺癌



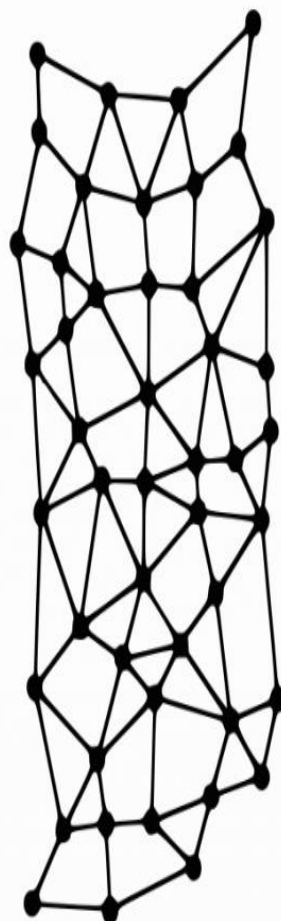
<https://www.masters.tw/270/%E7%99%8C%E7%97%87%E6%82%A3%E8%80%85%E7%82%BA%E4%BD%95%E9%9B%86%E4%B8%AD%E5%9C%A8%E6%B2%B3%E6%B5%81%E4%B8%8B%E6%B8%B8>



Centralized

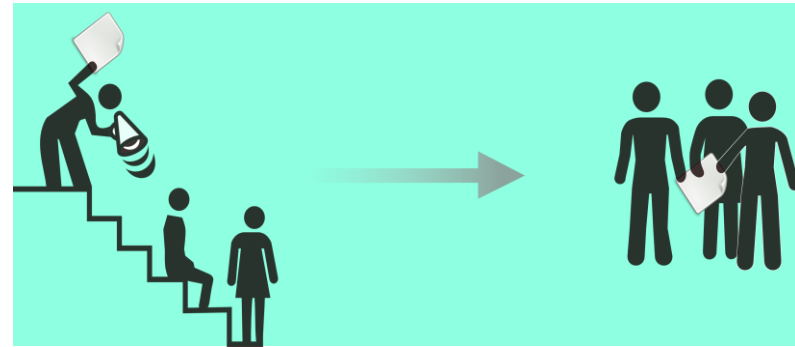


Decentralized



Distributed

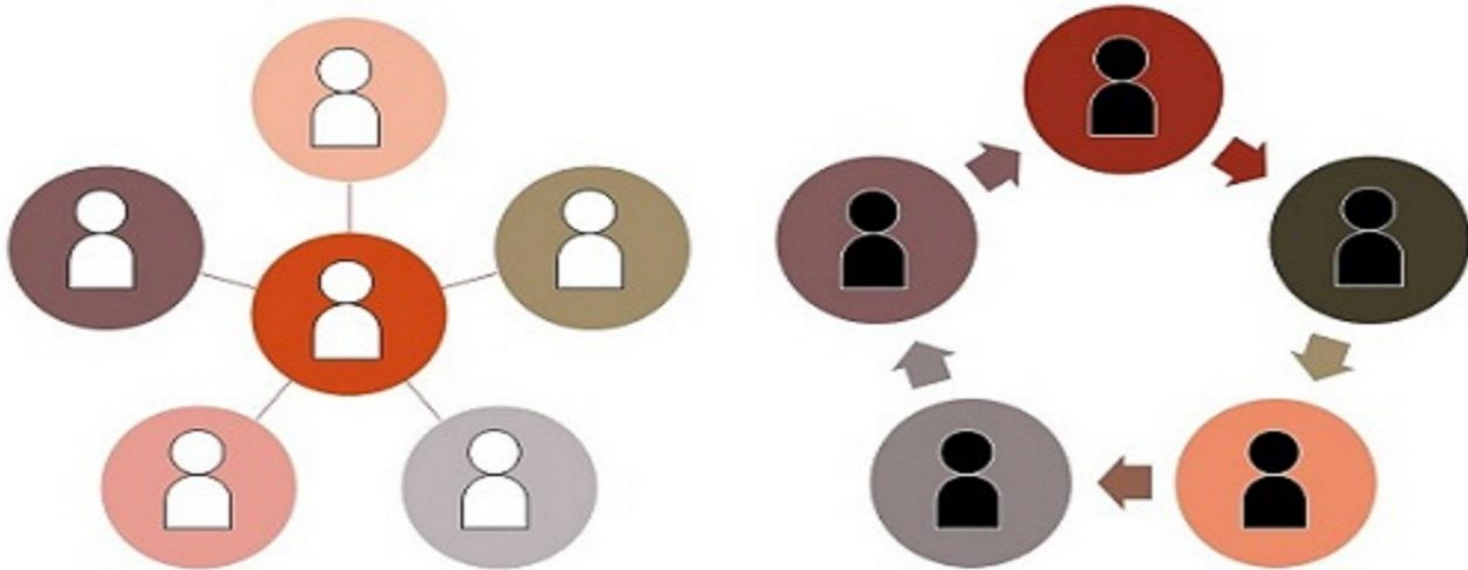
去中心化 (Decentralization)



https://upload.wikimedia.org/wikipedia/commons/b/ba/From_hierarchy_to_decentralisation_400px.png

另一個特性：去中心化

- 去中心化 (Decentralization) 是一種現象或結構，出現在眾多用戶或眾多節點的系統，每個用戶都可連接並影響其他節點。



CENTRALIZATION VS DECENTRALIZATION

金融交易是「中心化」的典範

■ 關於金融交易的幾個問題：

→ 誰賦予貨幣(如：台幣1000元)的價值？

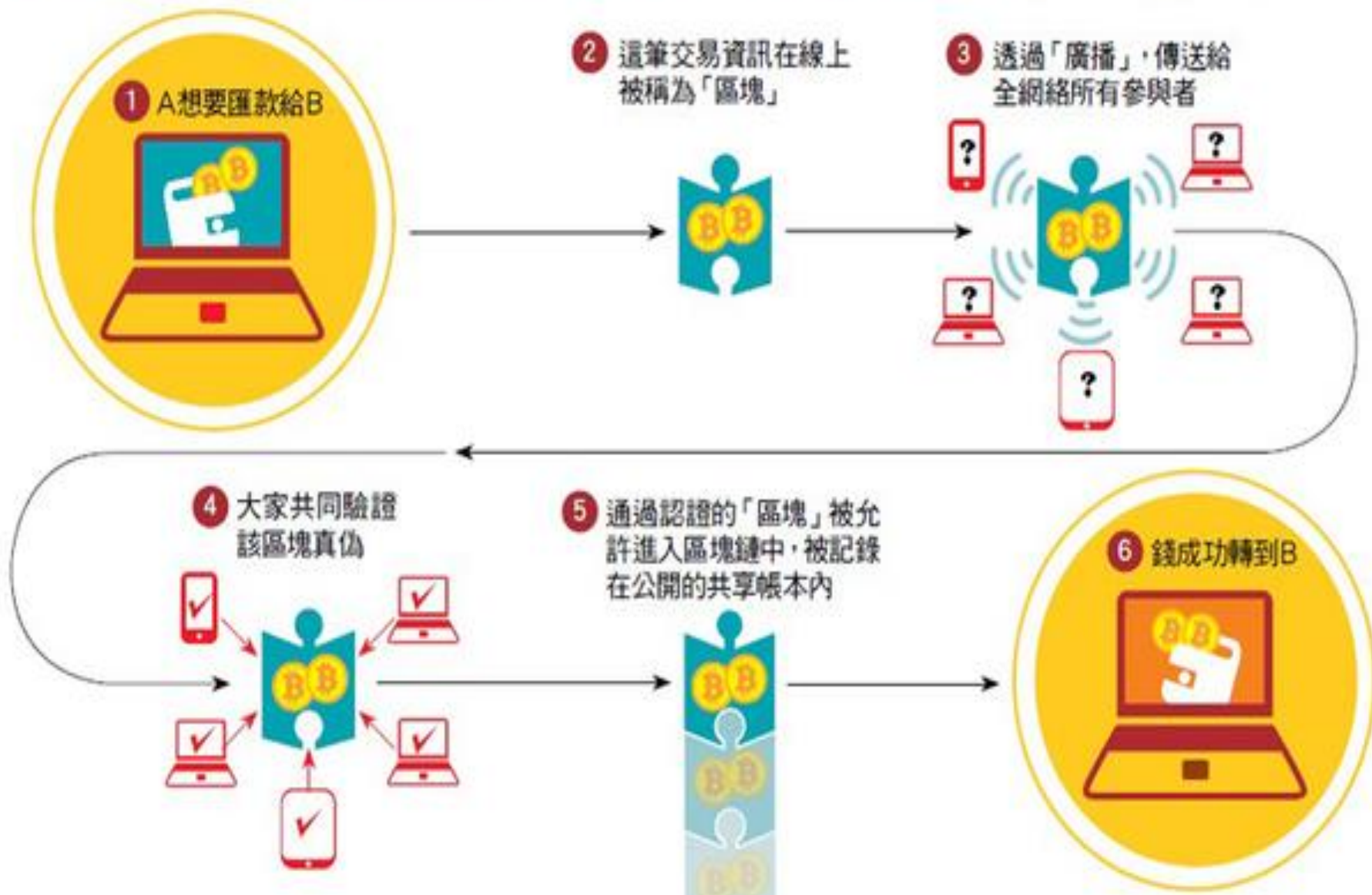
→ 貨幣價值由誰保障？

單筆付款流程 100100



區塊鏈這樣運作——以金融轉帳流程為例

資料來源：高盛、微拓、金融時報





去中心化與金融保險的發展

■ 金融科技(Fintech)與保險科技(Insurtech)

→ 金融科技：創造新的模式、業務、流程與產品，包括前端產業及後臺技術。例如：
互聯網和移動支付、網路信貸、區塊鏈。

→ 保險科技：產品、銷售通路、核保、理賠、後台作業與客服等傳統價值鏈，都將被保險科技帶來的創新徹底顛覆。

資訊解讀的價值



- Data Mining
- Big Data
- Predictive Analytics
- Artificial Intelligence
- Machine Learning

<https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwiBkpu7v5XiAhVLEqYKHRnkCLAQjRx6BAgBEAU&url=https%3A%2F%2Fwww.datanami.com%2F2016%2F09%2F15%2Fhow-do-you-value-information%2F&psig=AOvVaw20ofNnCNEO0Ay9m6JOzOUm&ust=1557733408183241>

資訊與知識的價值(資料採礦)

- 資料挖掘(Data Mining)的範例：\$ \$ \$ \$
→ 協助超級市場促銷及陳設商品。

Milk, eggs, sugar, bread



Customer1

Milk, eggs, cereal, bread



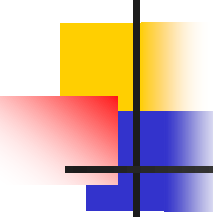
Customer2

Eggs, sugar



Customer3

數據分析的嚆頭！？



沃馬特量販店(Wal-Mart)

- 沃馬特最先蒐集、分析顧客資料，並以整理所得的資訊，提高銷售業績。
 - 分析發現美國消費者在週末購物時，許多人會同時購買尿布及啤酒。
 - 問題：為什麼這兩種商品會一起購買？又如何將這份資訊轉變為業績？
- 註：沃馬特從美國西南部發跡，剛開始只是一家五金行，現在已是全美最大的百貨零售業者。

「尿布與啤酒」的延伸價值

- 尿布與啤酒屬於關連性(Association)的關係，與常見的因果關係(Causality)不同。
- 關連性的價值未必低於因果關係，像是尿布與啤酒的關連，可用於：
 - 商品定價與促銷；
 - 商品擺設（商場動線）；
 - 商品倉儲。





賣場如何應用資訊？

- 商品定價與促銷

- 尿布及啤酒的定價（如：打折）

- 如何促銷尿布及啤酒以外的商品

- 商品擺設（商場動線）

- 尿布、啤酒兩種商品的相對位置

- 熱銷商品如何擺放

- 商品倉儲

- 何時進貨（購買資訊？）

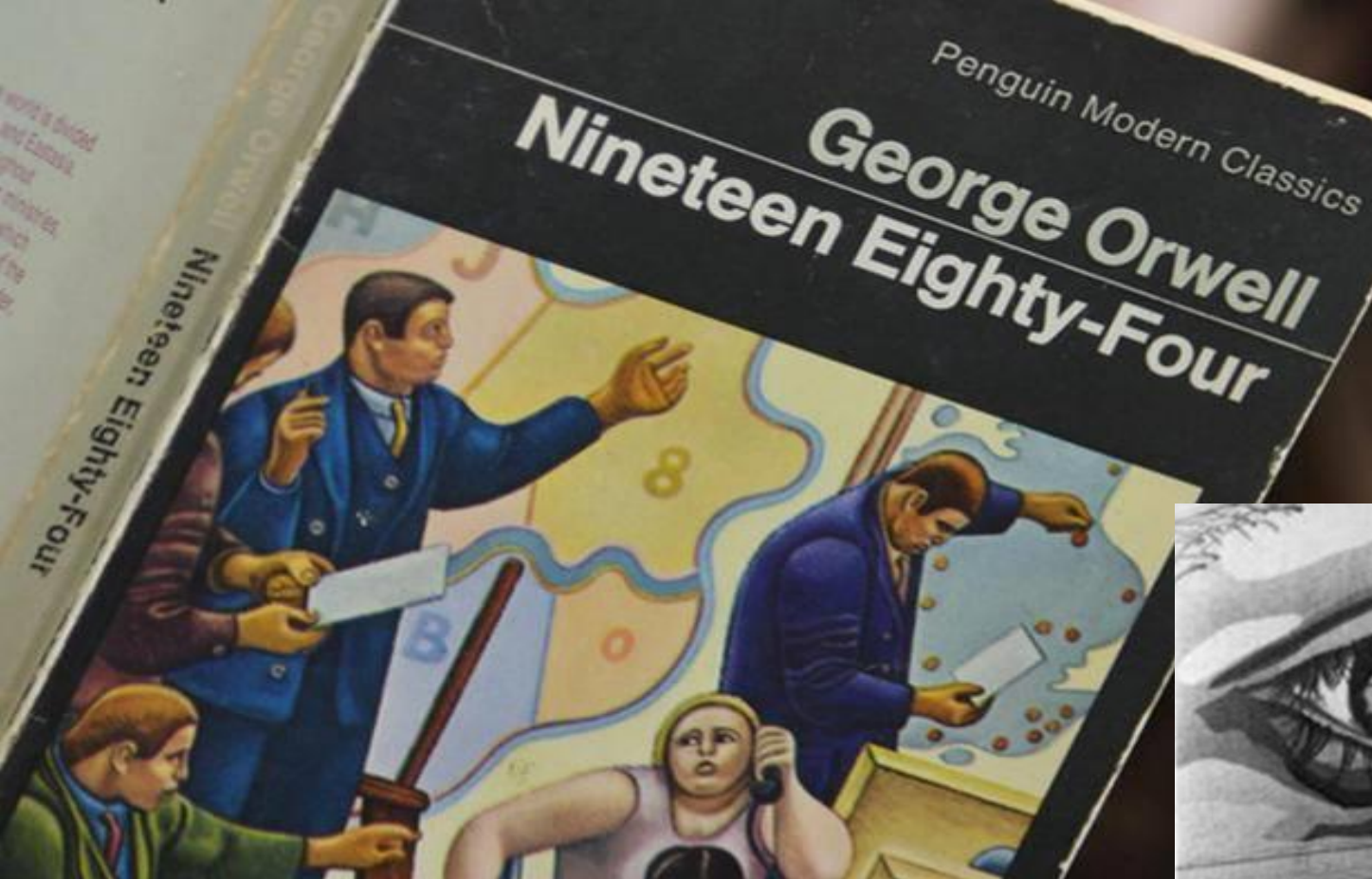
寶可夢旋風也能帶來商機嗎？



參考資料：<https://tw.news.yahoo.com/%E5%8F%B0%E5%8D%97%E5%A1%9E%E7%88%86-%E5%B0%8F%E9%BB%83%E7%B4%99%E6%A2%9D%E6%9B%9D-%E9%BE%9C%E9%80%9F%E5%8E%9F%E5%9B%A0-%E7%B6%B2%E7%AC%91-%E4%B8%8D%E5%8F%AD%E4%BD%A0%E4%BA%86-143342818.html>

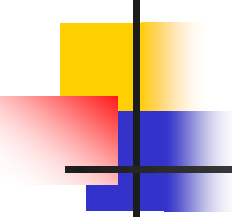
Pokémon GO Safari Zone in Tainan (寶可夢台南狩獵區)：估計主場都會公園奇美博物館有8萬人，大台南全區16萬人，連續五天活動總計主場有56萬人次，台南全區100萬人次。 六億商機！！
今日新聞NOWnews 記者陳聖璋 2018年11月5日





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





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，因為迷戀資料分析、濫用、以致於形成「資料獨裁」，或是「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...



使用大數據的建議

- 監管大數據的三個策略：
 - 個資保護由「個人同意制」，轉成資料使用者的「使用責任制」；
 - 運用大數據時，需尊重個人的「能動性」
（Human Agency; 政府對人民的判斷是基於實際的行動，而非大數據分析的預測）；
 - 培養「演算學家」以處理大數據的審計師。
註：除了現有量化分析人員，公司也需積極培養大數據的專家（IT、分析師等）。

- 不確定性 : It's Difficult to Make Predictions, Especially About the Future.

Predictive Tools



• Prediction is very difficult, especially about the future.'

- Niels Bohr – Physicist (1885-1962)

"It is tough to make predictions, especially about the future."

- Yogi Berra, Baseball Savant





資料分享的副作用

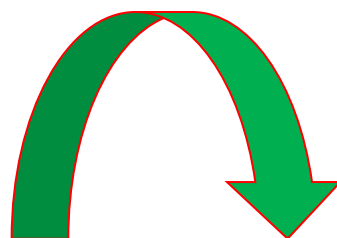
- Facebook社群軟體盛行也帶來負面效應，除了個人資料安全疑慮，害怕與人接觸的社交恐懼症 (sociophobia) 日趨嚴重。
 - 擔心自己遭受社會遺忘或淘汰；
「滿紙荒唐言，一把辛酸淚！都云作者痴，誰解其中味？」
 - 人際關係的疏離（虛擬vs.實際）；
「假作真時真亦假，無為有處有還無」
 - 私刑（人肉搜索）、扮演法官。

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





大數據等於一切？

■ 無限母體 vs. 有限母體

→ 再多的資料還是樣本！（如消費者行為）

■ 資料品質仍是主要關鍵！

→ 人工智慧（AI）仰賴好的訓練資料，尋找背後的規則及趨勢（如 alphago）。

→ Google Flu Trends透過搜尋引擎的「關鍵詞」，比美國疾管局（Center of Disease Control; CDC）更早偵測出流感盛行。

→ 族群及關鍵詞、從眾效應(Bandwagon effect)



大數據研究的建議

- 統計學家（&資料科學家）的專業技能

→ 與人溝通、與資料溝通(data sense)、與專業溝通、與電腦（機器）溝通。

- 增加專業領域間的溝通平台與機會

→ 如何結合人與機器？

→ 「多樣性」（物種、知識）是重要資產，
不少發現源自其他角度的思維。

（馬爾薩斯《人口論》與達爾文、馬克斯）

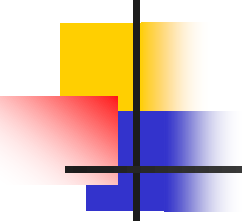
什麼是統計？



http://farm2.static.flickr.com/1281/4669551487_34496c71aa.jpg



<https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.canstockphoto.com%2Fbright-statistics-illustration-35101794.html&psig=AOvVaw10rqxE16qvmWwyShbVj6ml&ust=1672757240051000&source=images&cd=vfe&ved=0CBAQjRxqFwoTCPgagpCRqfwCFQAAAAAdAAAAABAJ>



商學院有3G（三「計」）

會計——很快忘記！

經濟——經常忘記！！

統計——通通忘記！！！！

註：為什麼這三門課特別棘手，統計
似乎最難上手？

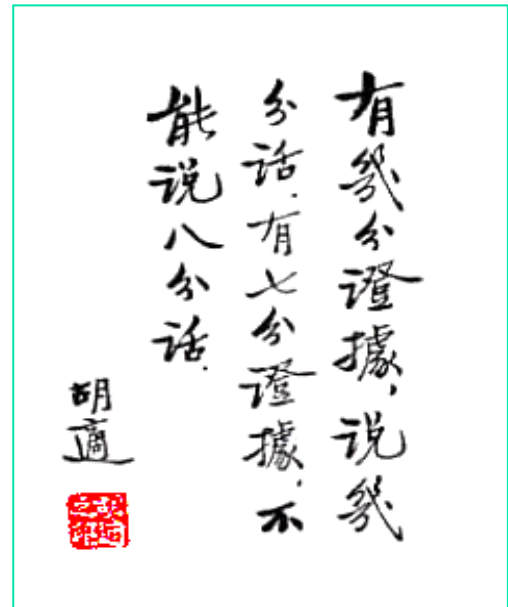


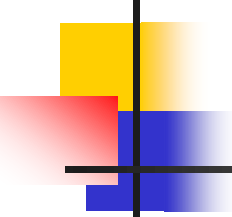
什麼是統計？

- 統計學是研究定義問題、運用資料蒐集、整理、陳示、分析與推論等科學方法，在不確定(Uncertainty)情況下，做出合理決策的科學。



https://en.wikipedia.org/wiki/File:Hu_Shih_1960_color.jpg





學習統計的幾個要素

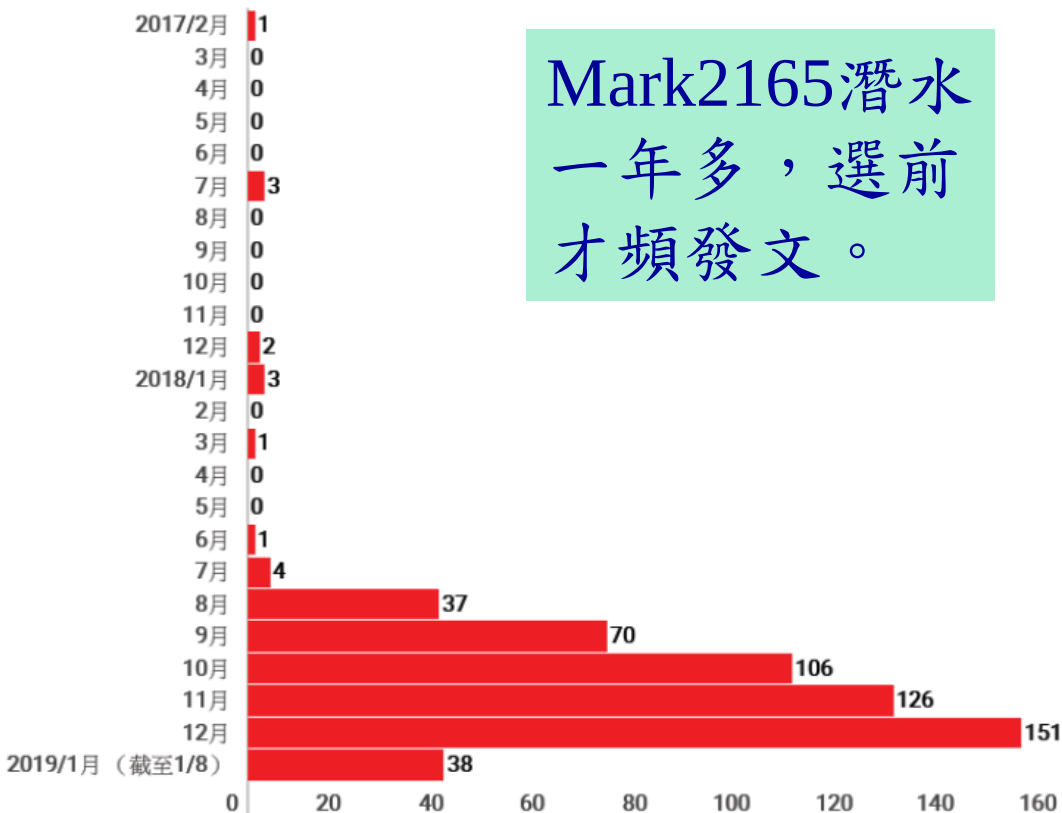
大胆的假设
小心的求證
通

- 解決問題大致有以下幾個要素：
 - 如何定義、測量？(e.g. Variable Format, Data Collection)
 - 如何判斷、取捨？(e.g. Estimation, Prediction, Testing)
 - 如何詮釋、增加附加價值？(e.g. Utility, Decision)

如何從龐雜訊息中篩選資訊？

- 訊息產生及傳播方式愈趨大量多元，如何篩選、整理、分析變得更加困難。

Mark2165潛水
一年多，選前
才頻發文。



<https://www.wittenberg.edu/sites/default/files/styles/boxlarge2x/public/2020-05/Insta%20Fake%20News.png?itok=16pf5u7H>

<https://www.cw.com.tw/article/5093610>



報紙上常看到的敘述！

- 2004年九月底某學校公布使用手機的肇事率，發現在車內用免持聽筒打電話，比用手持聽筒還要危險。
 - 由於民眾以為使用免持聽筒就能夠一面開車、一面通話，但往往更容易分心，肇事率比用手持聽筒高出將近1倍。
 - 你/妳覺得這裡有甚麼疑問？

解決問題(Problem Solving)流程



<https://www.oneeducation.org.uk/course/creativity-and-problem-solving-skills/>

定義問題



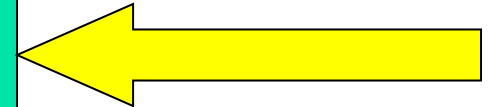
蒐集資料



分析資料



詮釋結果



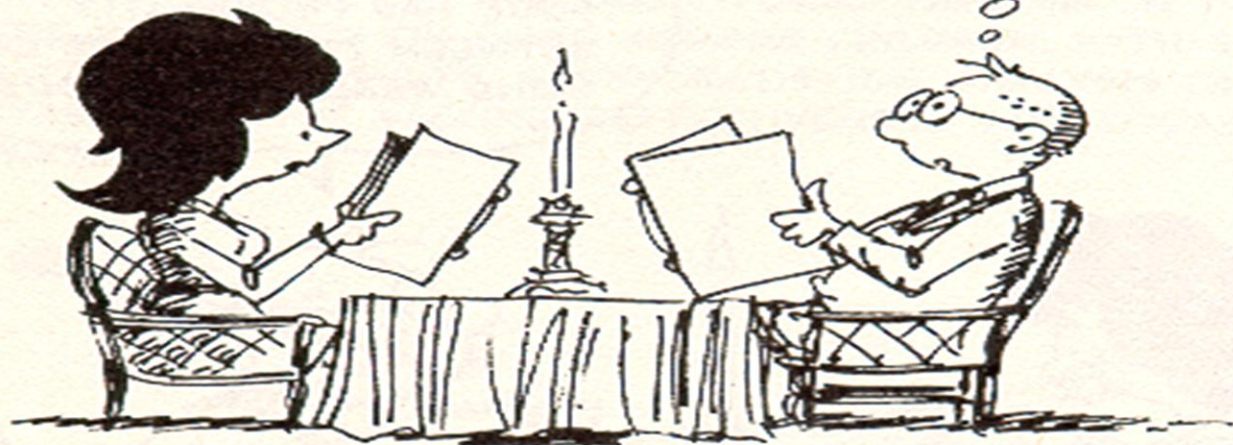
絕大多數的
資料分析重心

WHAT IS STATISTICS?

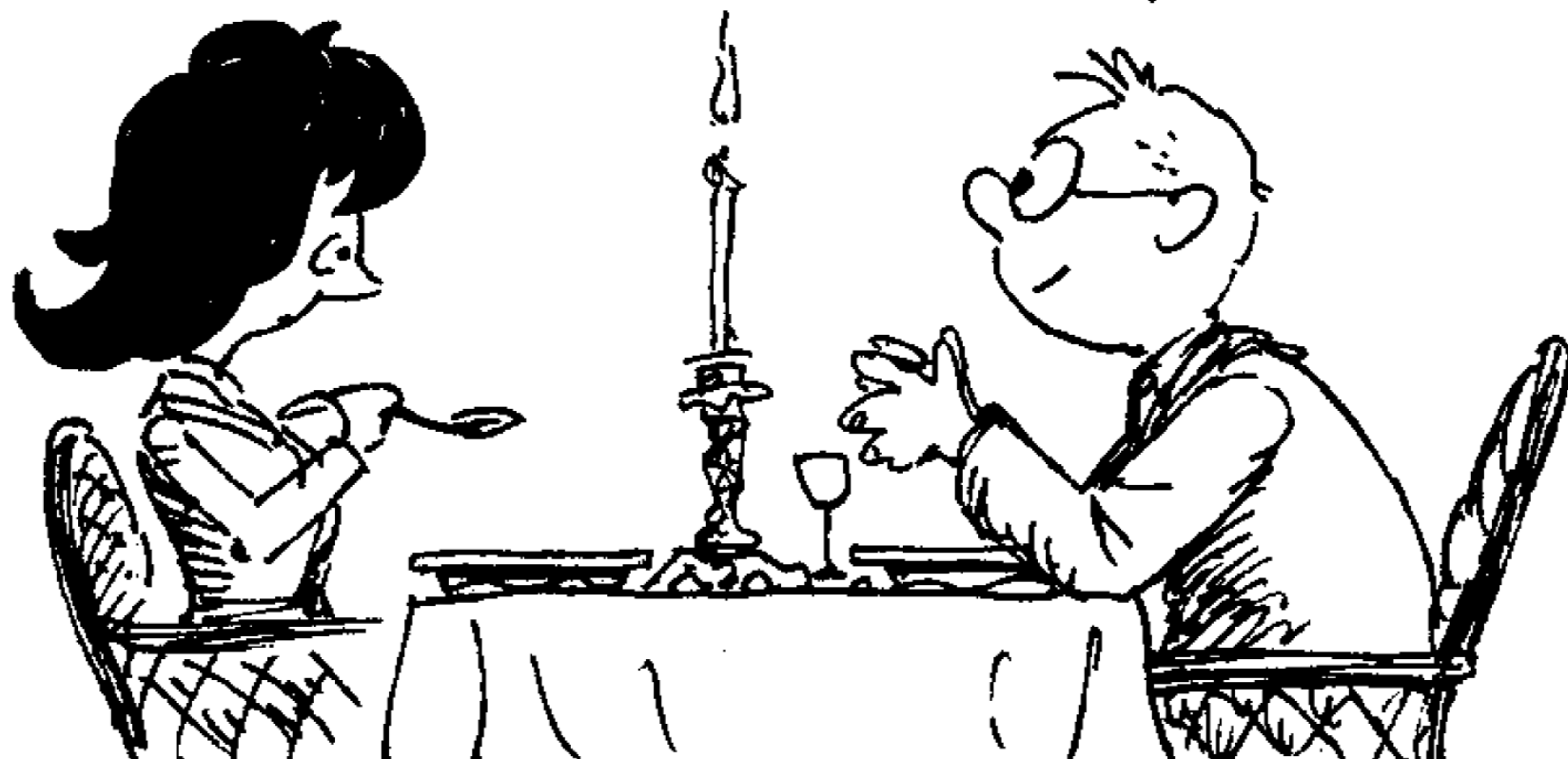
WE MUDDLE THROUGH LIFE MAKING CHOICES
BASED ON INCOMPLETE INFORMATION...

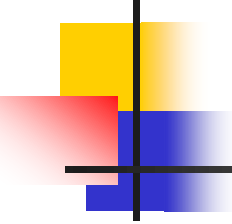
SHOULD I HAVE THE SOUP?
EVERYTHING ELSE IS SO
EXPENSIVE, AND I DON'T
KNOW WHO'S PAYING... ARE
STATISTICIANS STINGY? I'VE
NEVER GONE OUT WITH
ONE BEFORE... THOUGH I
ONCE KNEW A VERY
GENEROUS ACCOUNTANT...

SHOULD I HAVE THE SOUP?
27 OUT OF THE 36 TIMES
I'VE HAD IT, IT WAS PRETTY
GOOD... BUT IS MONDAY THE
REGULAR CHEF'S NIGHT
OFF? AND WHAT IF ALL THE
AIR MOLECULES IN THE
ROOM SUDDENLY FLY UP TO
THE CEILING?



GOOD CHOICE! I'M 95%
CONFIDENT THAT TONIGHT'S
SOUP HAS PROBABILITY
BETWEEN 73% AND 77% OF
BEING REALLY DELICIOUS!





馬克吐溫對統計的想法

There are three kinds of lies:

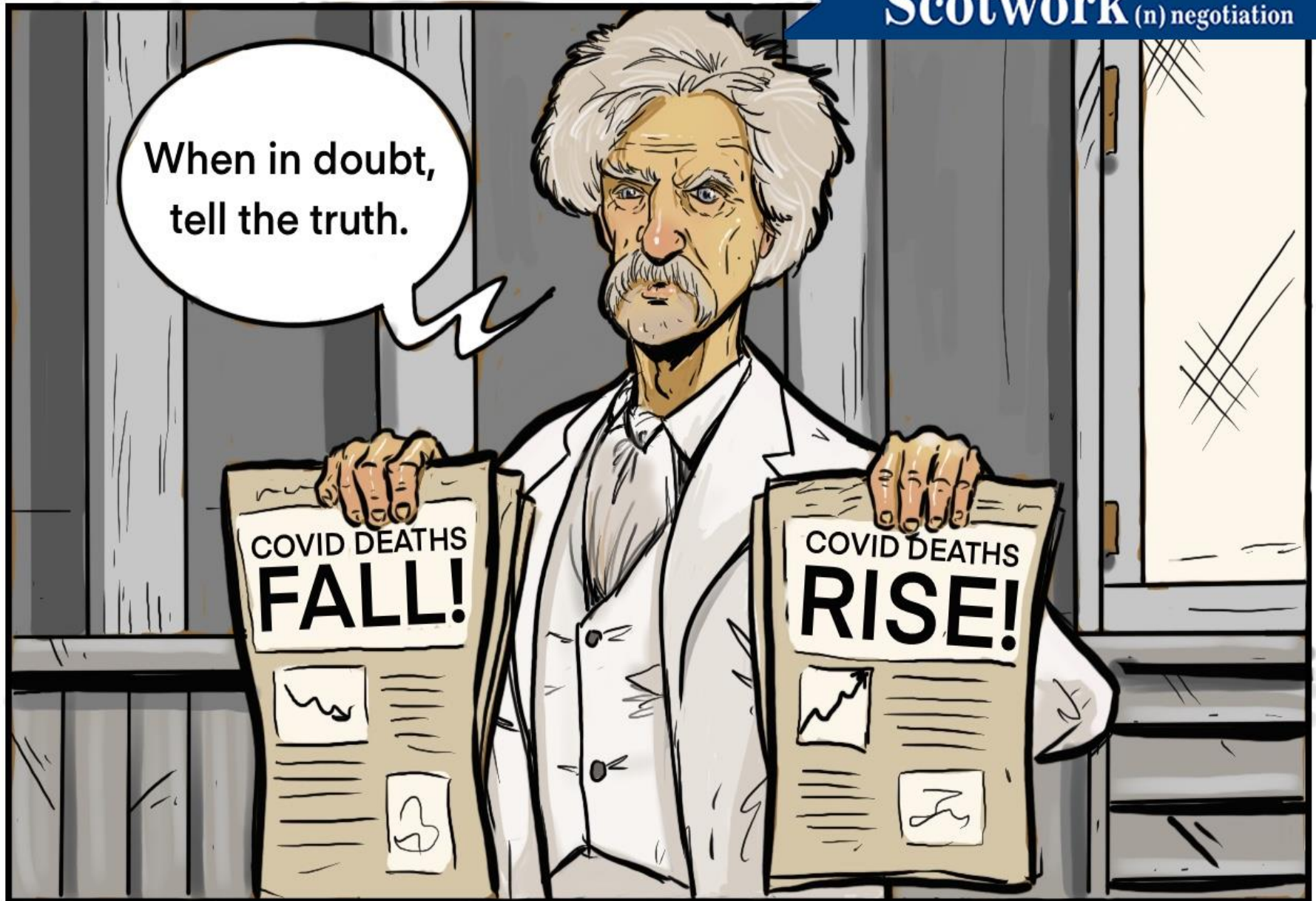
Lies,

Damned lies,

and **Statistics!!**

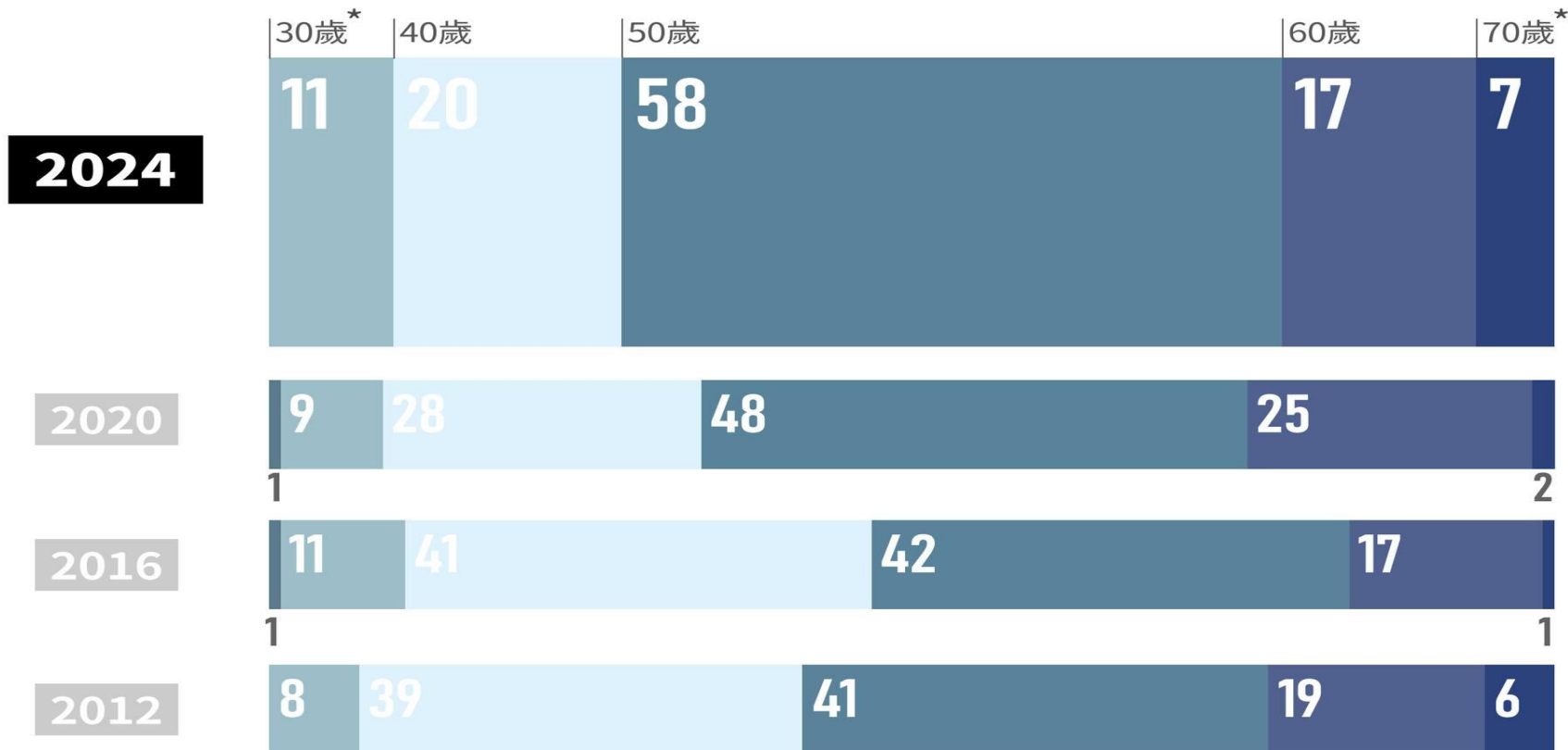
by Benjamin Disraeli (英國首相)





第11屆立委年齡分布

單位：人



*註：70歲以上 - 游錫堃(75)、陳永康(73)、柯建銘(72)、賴士葆(72)、陳超明(72)
陳雪生(72)、林德福(70)

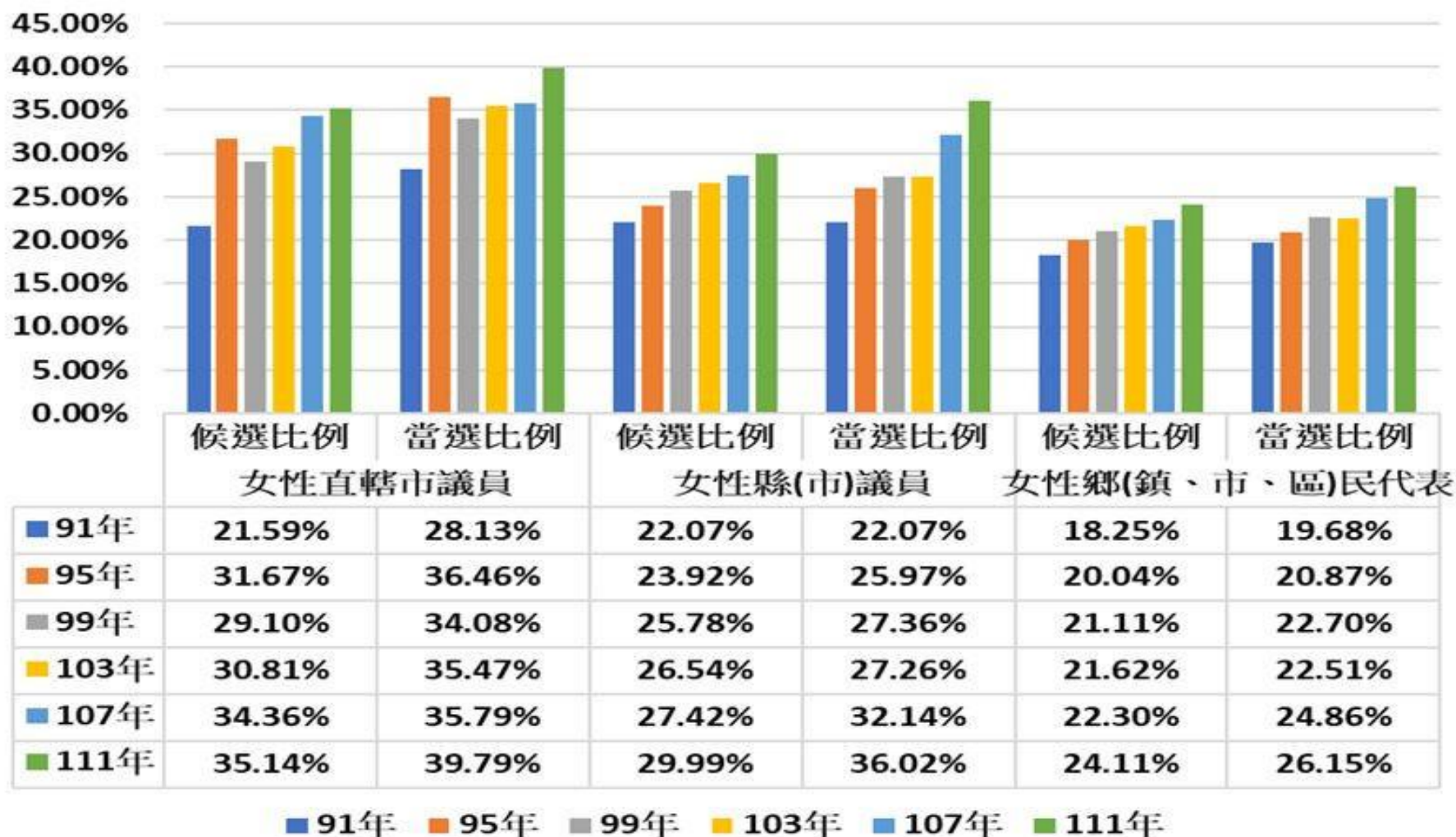
30至39歲 - 黃捷(30)、廖偉翔(33)、牛煦庭(33)、羅廷瑋(34)、徐巧芯(34)
廖先翔(35)、黃健豪(35)、邱若華(35)、吳沛憶(36)、陳冠廷(37)、洪申翰(39)

資料來源：中選會、立法院

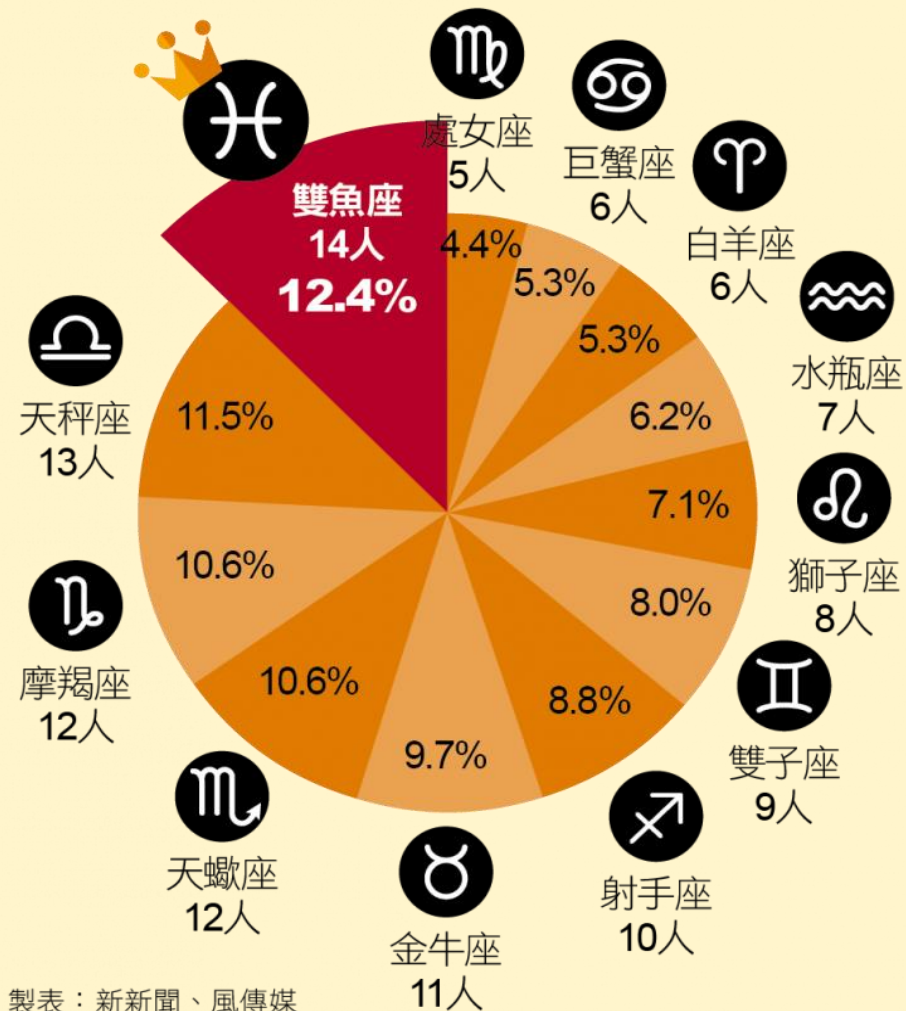
https://scontent-tpe1-1.xx.fbcdn.net/v/t39.30808-6/425288557_770454721779762_4629537348715970133_n.jpg?_nc_cat=110&ccb=1-7&_nc_sid=3635dc&_nc_ohc=nK9p5qScxDsAX_3BIDr&_nc_ht=scontent-tpe1-1.xx&oh=00_AfD6Gfg9MnO1WlNiv4RLSqkS5najZlDyN09IUmy-hZ2VLA&oe=65E036E2

表 2：近 20 年(六屆) 地方民意代表選舉結果性別統計資料

91-111年 地方民意代表 女性參政比例趨勢圖



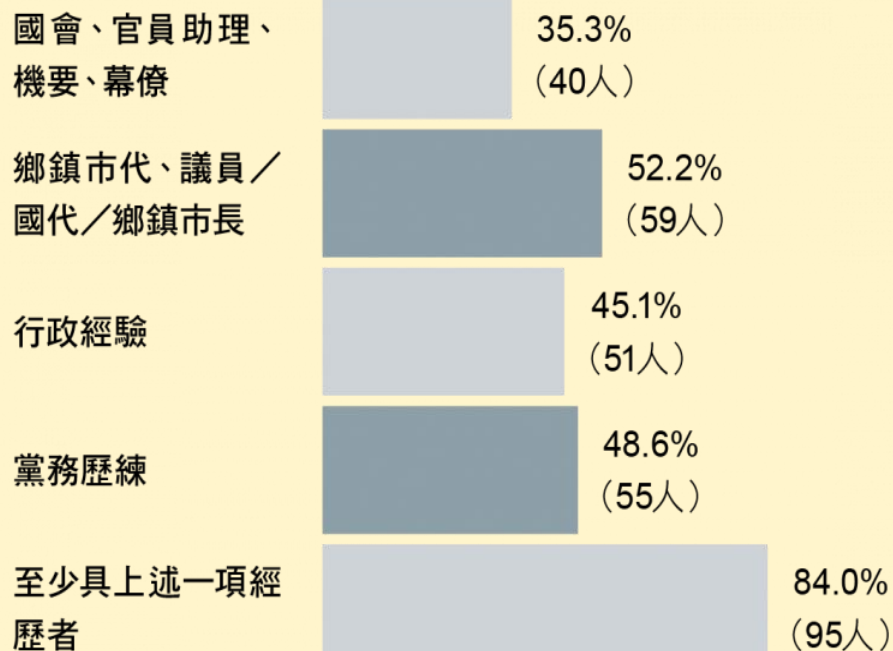
雙魚座最適合當立委？



- 適合度的定義？
- 分布均勻嗎？(0.4678)

選立委需要具備哪些履歷？

■ 本屆立委歷任》



註1.統整立委當選前的經歷。連任立委以統計選上首任立委前的經歷為主；屆期不連續者，以選上最近一任之前的經歷為主。

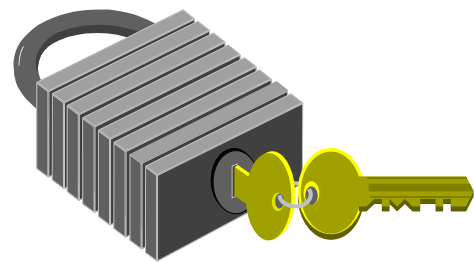
註2：黨務經歷不計國民黨、民進黨全國黨代表、國民黨中央委員

資料來源：選舉公報、立委網頁自介、維基百科

資料整理：新新聞、風傳媒

https://image.cache.storm.mg/styles/smg-800xauto-er/s3/media/image/2020/03/25/2020022724_U8719_M601088_bd60.png?itok=xcecF5tv

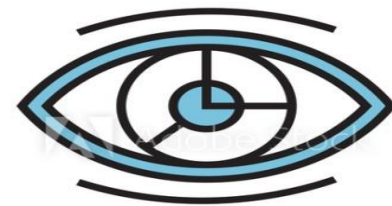
https://image.cache.storm.mg/styles/smg-800xauto-er/s3/media/image/2020/03/25/20200325-022724_U8719_M601089_98d1.png?itok=UF0kvWz



什麼是統計？

- 統計學是研究定義問題、運用資料蒐集、整理、陳示、分析與推論等科學方法，在不確定(Uncertainty)情況下，做出合理決策的科學。

註：通常由部分資料（「樣本」 Sample）推敲出整體（「母體」 Population）的特性。



Visualization

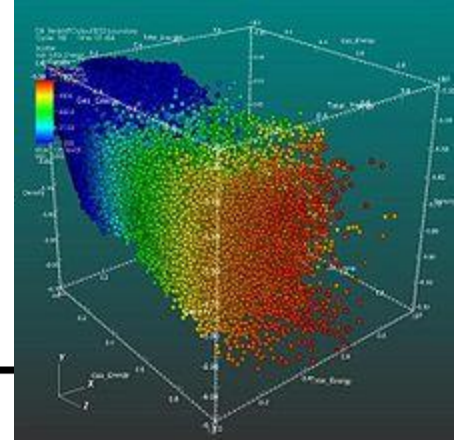
#212488495

文字分析：觀念的視覺化

為什麼需要文學？了解文學、接近文學，對我們形成價值判斷有什麼關係？如果說，文學有一百種所謂「功能」而我必須選擇一種最重要的，我的答案是：德文有一個很精確的說法「macht sichtbar」，意思是「**使看不見的東西被看見**」。在我自己的體認中，這就是文學跟藝術最重要、最實質、最核心的一個作用。

——龍應台

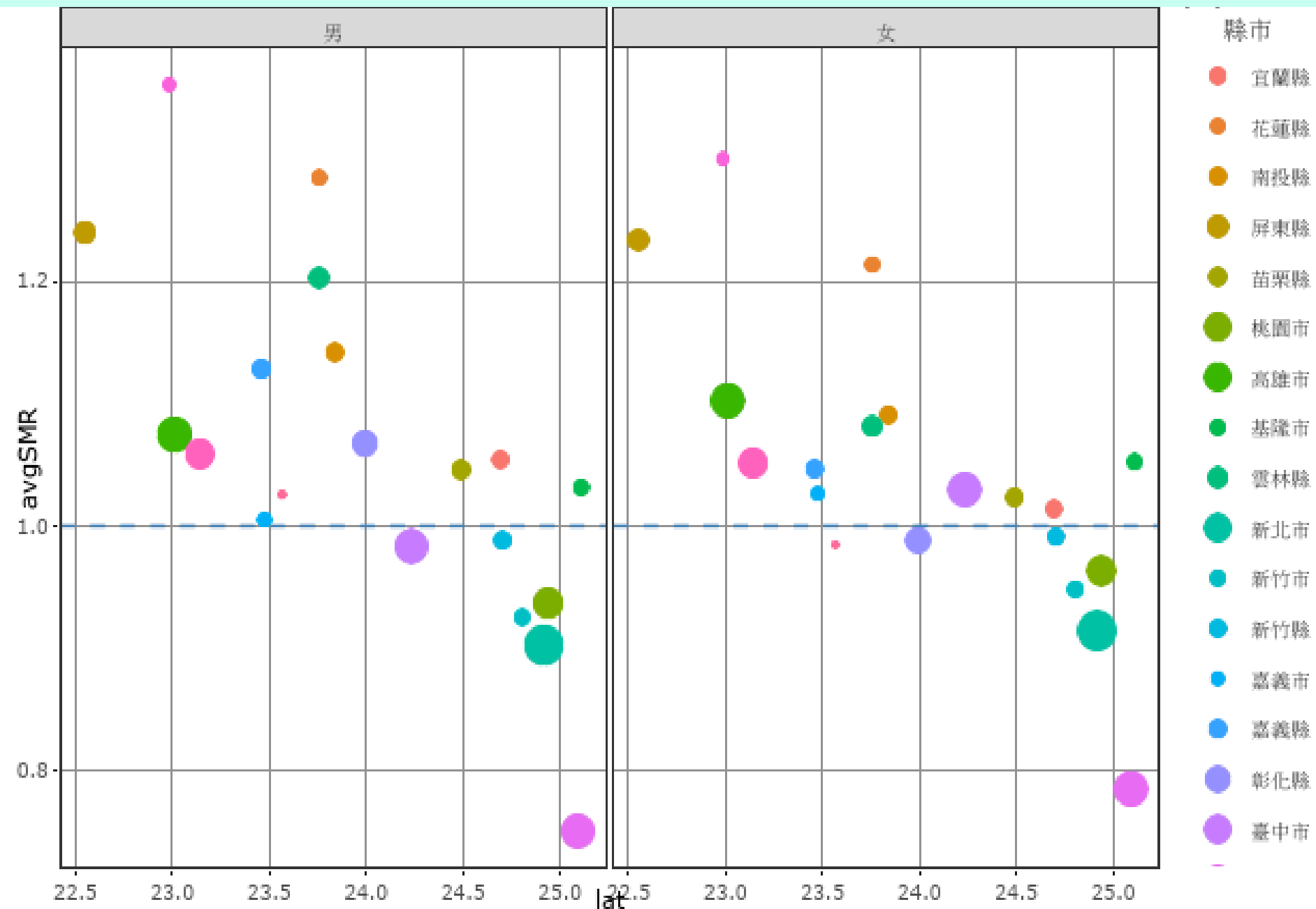
Data visualization



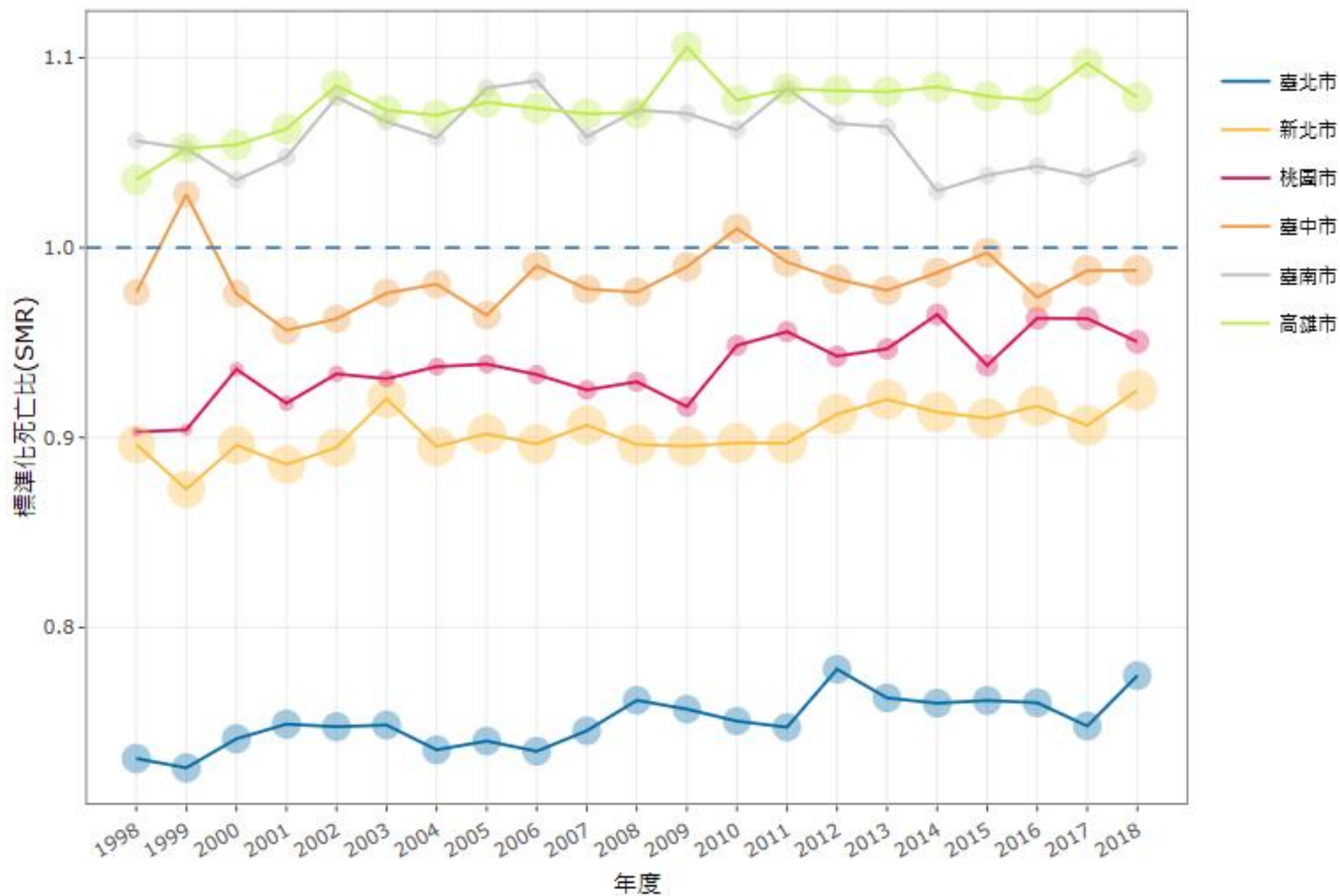
Data visualization is the graphic representation of data. It involves producing images that communicate relationships among the represented data to viewers of the images. This communication is achieved through the use of a systematic mapping between graphic marks and data values in the creation of the visualization. This mapping establishes how data values will be represented visually, determining how and to what extent a property of a graphic mark, such as size or color, will change to reflect changes in the value of a datum.

To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message.^[1] Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

緯度與死亡率的關係（台灣SMR）



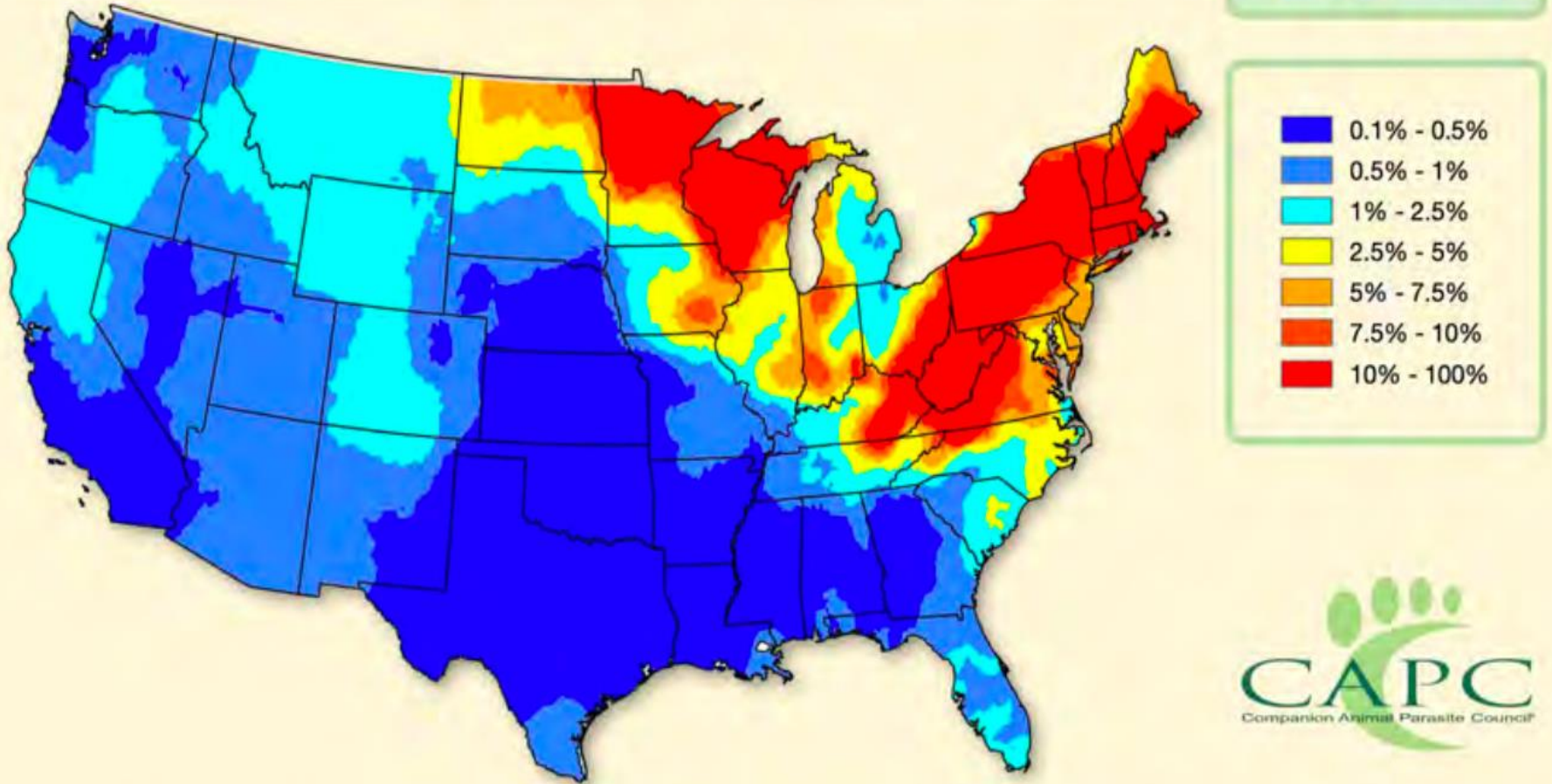
六都1998-2018年男性標準死亡比



美國2022年萊姆病預測(Lyme disease, 壁蝨)

Forecasted Prevalence of Lyme Disease

2022



John Snow & Spatial Analysis

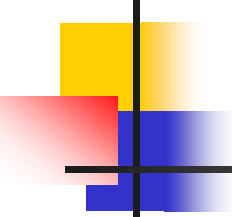


巴黎48個行政區的霍亂熱區圖

Charles Picquet

https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcSh_go8YVCyyZvQsr7UBFWtj66b1GW2cN0Rlg&usqp=CAU

<https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTUhpPXin7G9FYiZDkwcle2ZV1YIoqMCEm2g&usqp=CAU>



圖形與表格

- 除了基本的敘述統計量外，圖形與表格可以輔助判斷資料的特性。

→ 常見的圖形：Boxplot、Histogram

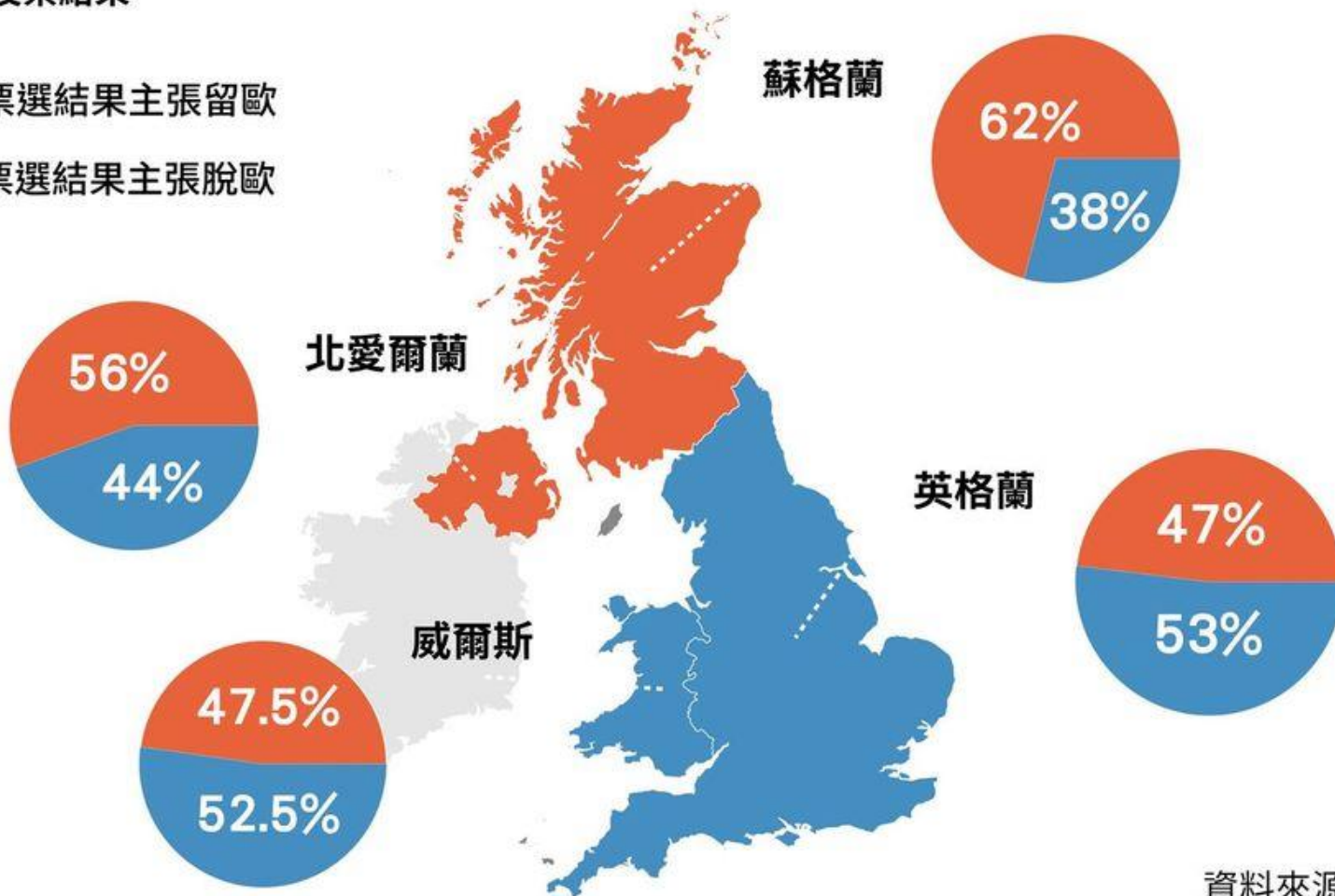
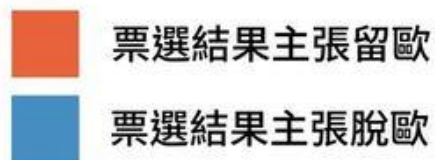
- 這些圖表看似簡單，但仔細判讀仍可發現重要訊息，甚至不需進階統計分析，即能約略猜出分析的結論。

註：A picture is worth a thousand words!



英國四大構成 脫歐留歐各有所好

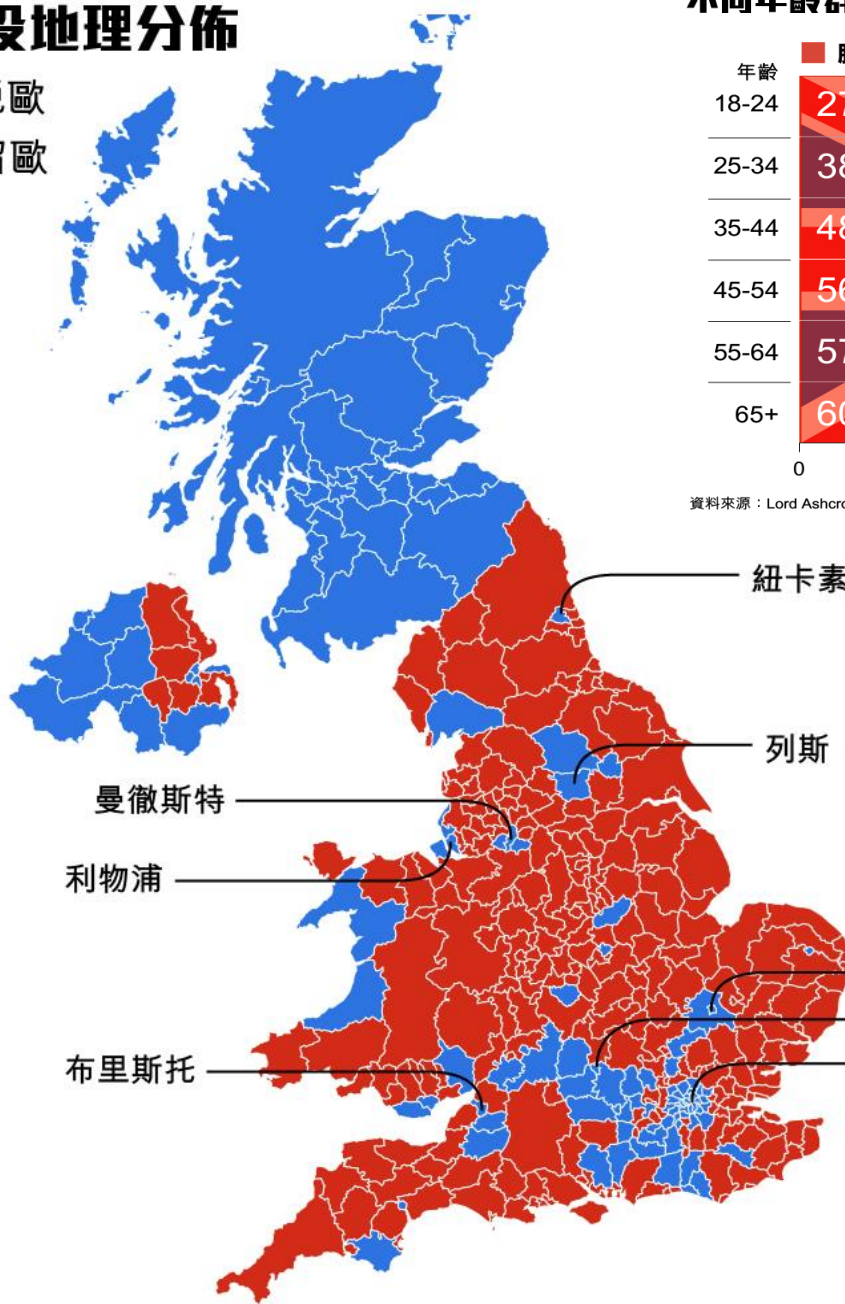
各構成國投票結果



資料來源: BBC

脫歐選舉公投地理分佈

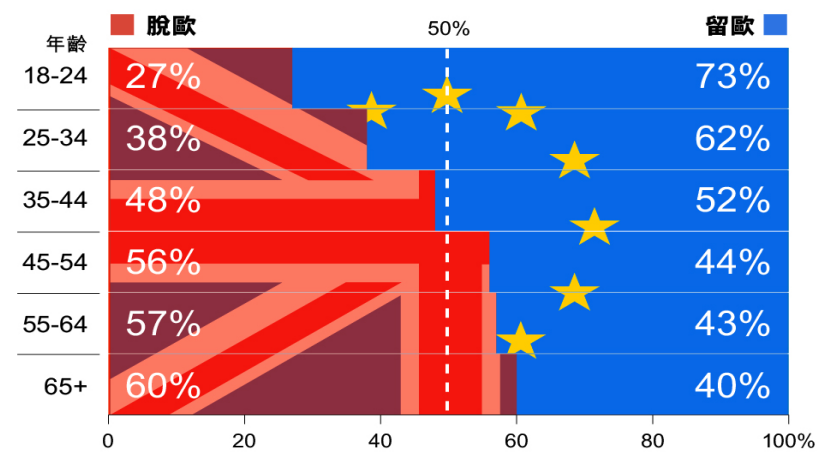
■ 大多數支持脫歐
■ 大多數支持留歐



曼徹斯特
 利物浦
 布里斯托
 劍橋
 牛津
 倫敦
 紐卡素 (台譯新堡)
 列斯 (台譯里茲)



不同年齡群投票結果



資料來源：Lord Ashcroft Polls、BBC

<https://theinitium.com/article/20160627-opinion-britishandeu/>

圖形有時描述地更傳神！

A Picture



is worth
=

Creamy, delicious, yummy,
fudge ice cream, smooth,
chocolate-chip mint ice
cream, strawberry ice
cream with real chunks of
strawberry, colored sugar
sprinkles, waffle sugar
cone, sweet, wonderful,
tastes great, cold, nice to
eat, dessert, good yummy
toppings, chocolate
sprinkles, comforting,
good, fun, dipping, terrific,

A thousand words.

©2003 E. Aoyama



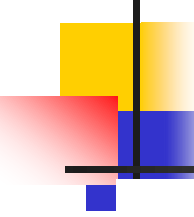


<https://corticare.com/wp-content/uploads/revslider/home-new/clinical-trials-bg.jpg>

如何建立模型？



<https://www.bigskyassociates.com/hs-fs/hub/283820/file-2615489650-jpg/statistical-modeling-methods-for-security-data-analysis.jpg.jpg?width=217&name=statistical-modeling-methods-for-security-data-analysis.jpg.jpg>



量化模型

- 透過量化模型描述觀察結果：

$$\text{觀察現象} = \text{模型} + \text{誤差}$$

或是

$$y = f(x) + \text{error} ; \text{觀察值} = \text{訊號} + \text{雜訊}$$

- 數量化模型的關鍵：

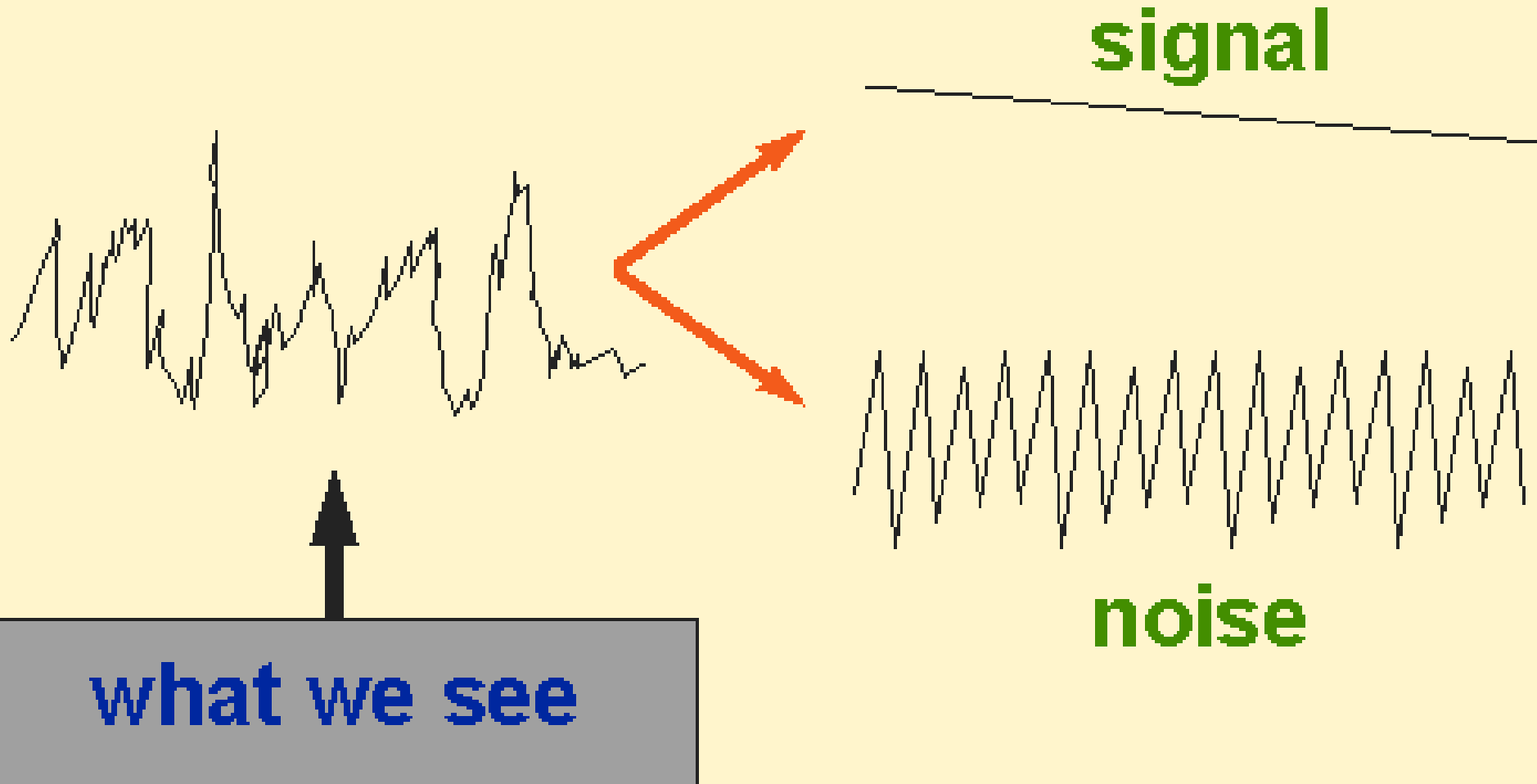
→ 量化目標值 y ：定義問題！

→ 選取關鍵變數： $x_1, x_2, \dots, x_p$

→ 建立量化模型：統計學習、機器學習。

觀察現象 = 模型 + 誤差

What we observe can be divided into:



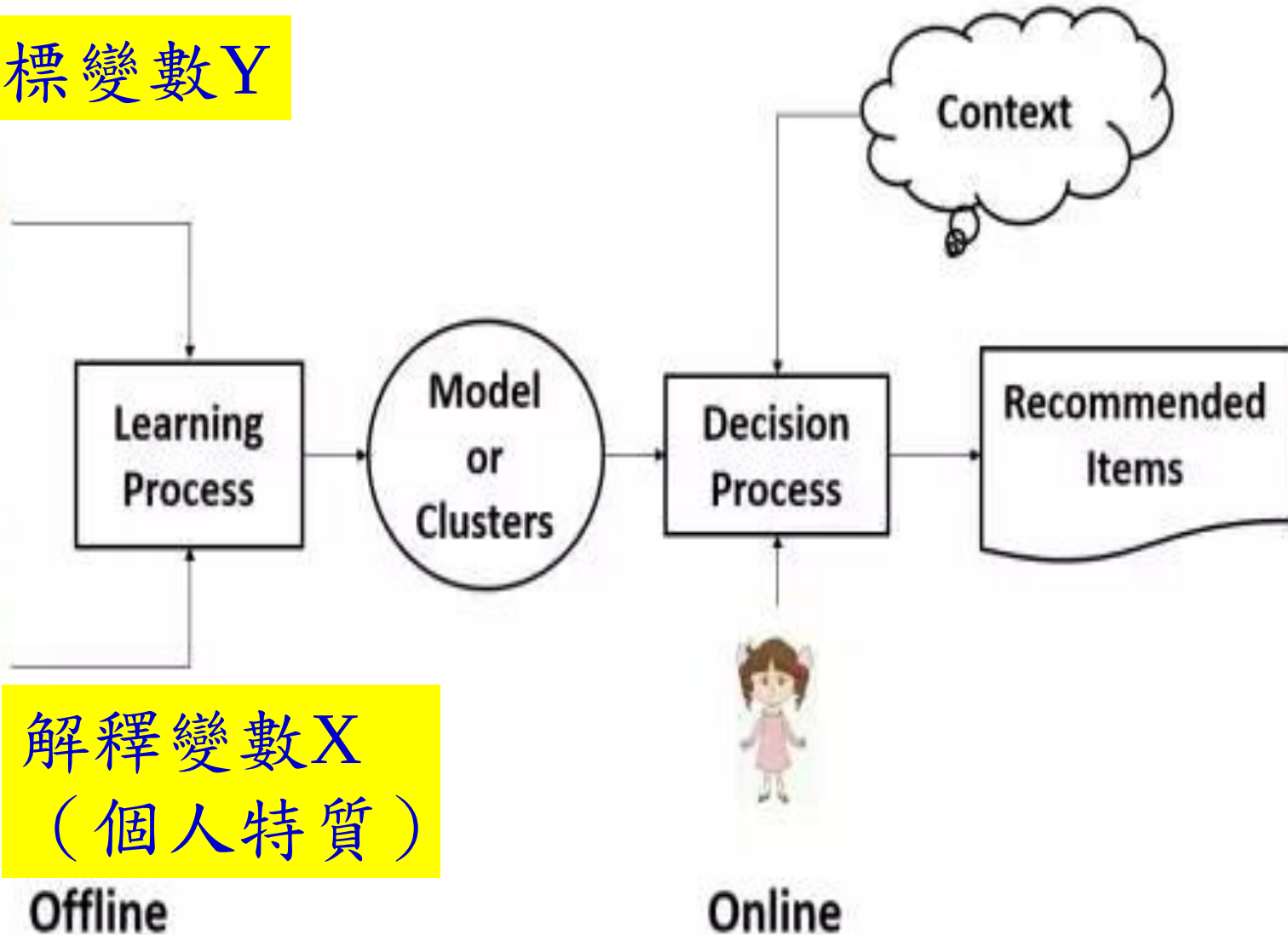
推薦系統(Recommendation System)分析流程

目標變數Y

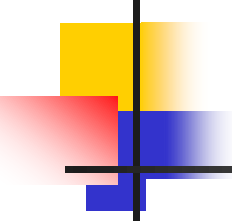


解釋變數X
(個人特質)

Offline



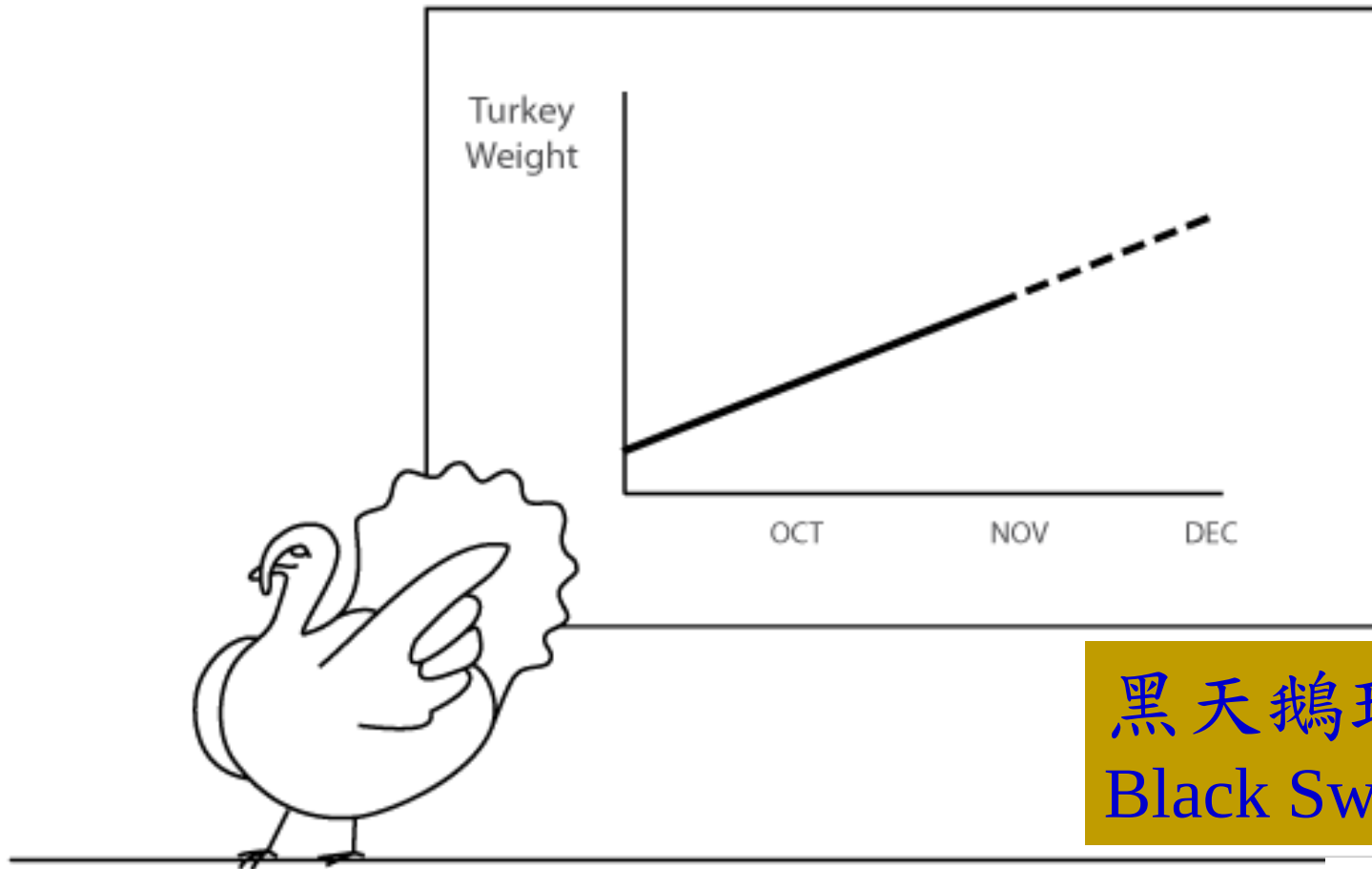
Online



解決問題的考量因素？

- 生活上遇到問題時，解決過程大致包括：
 - 確定問題目標及需求 Y ；
 - 蒐集相關資訊 X （資料來源、資料可信度、資料量及其品質）；
 - 資料整理、分析比對；
 - 決策評估標準 $f(X)$ （不同決策衍生出的優勢及劣勢、短期及中長期的影響）。

THANKSGIVING PREDICTIVE ANALYTICS



黑天鵝理論
Black Swan Theory

"I see no reason why excellent growth shouldn't continue.."

<https://i0.wp.com/timoelliott.com/blog/wp-content/uploads/2013/07/turkey-predictive-analytics.png?ssl=1>

Q：灰犀牛理論？



<https://cdn.technasia.com/wp-content/uploads/2016/10/bs1.jpg>

祝大家有個
愉快的學期！

