統計諮詢

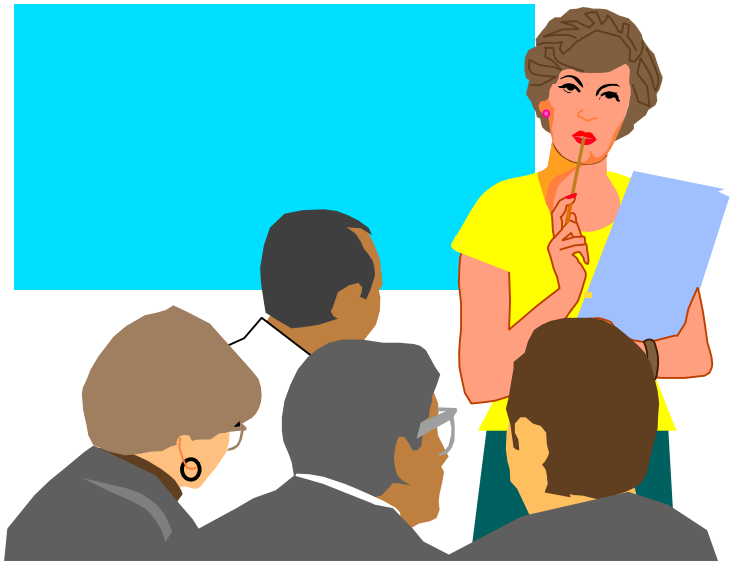
Fall 2025

授課教師：統計系余清祥

日期：2025年9月11日

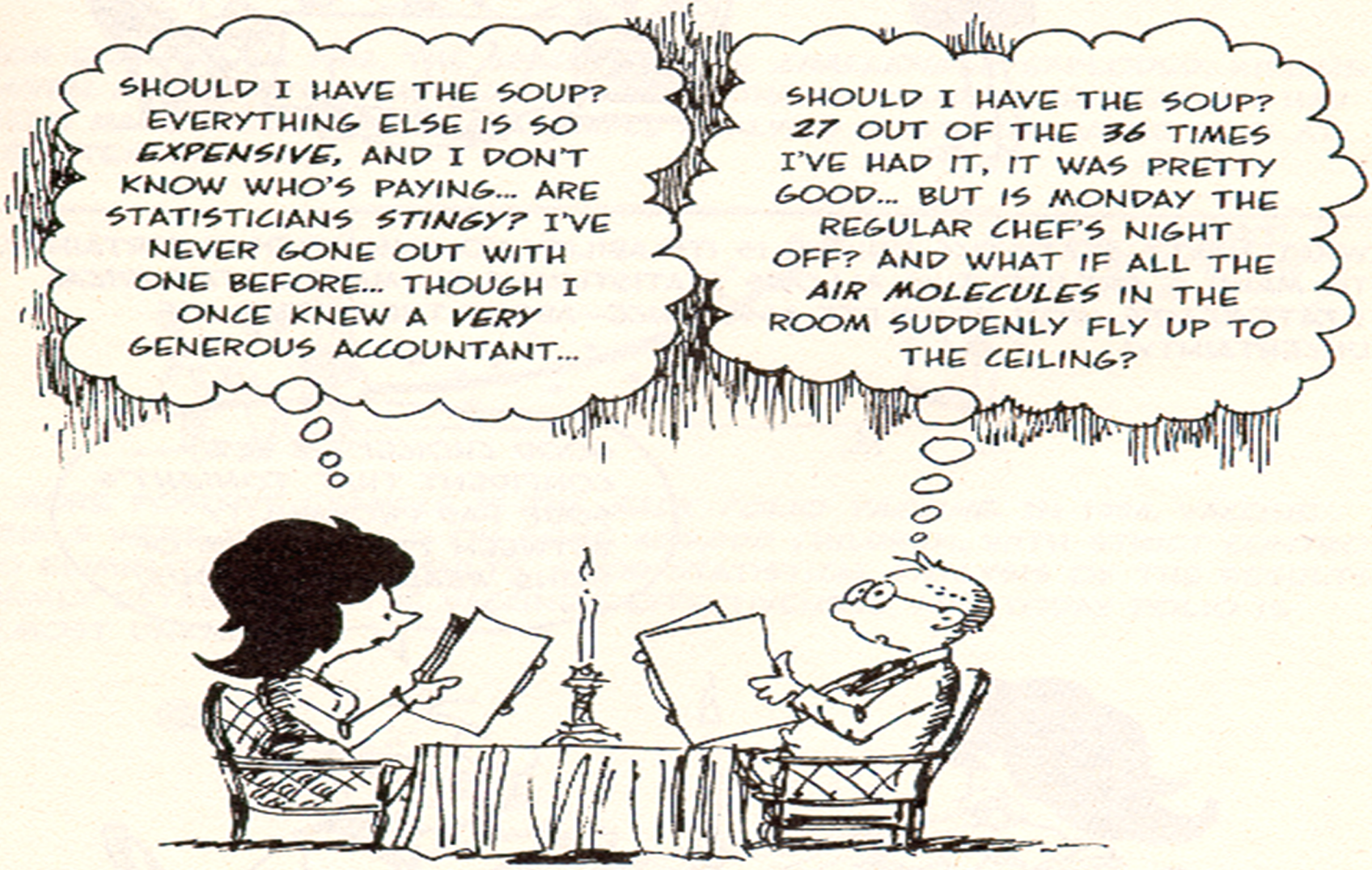
第二週：統計諮詢師

(資料科學家？)



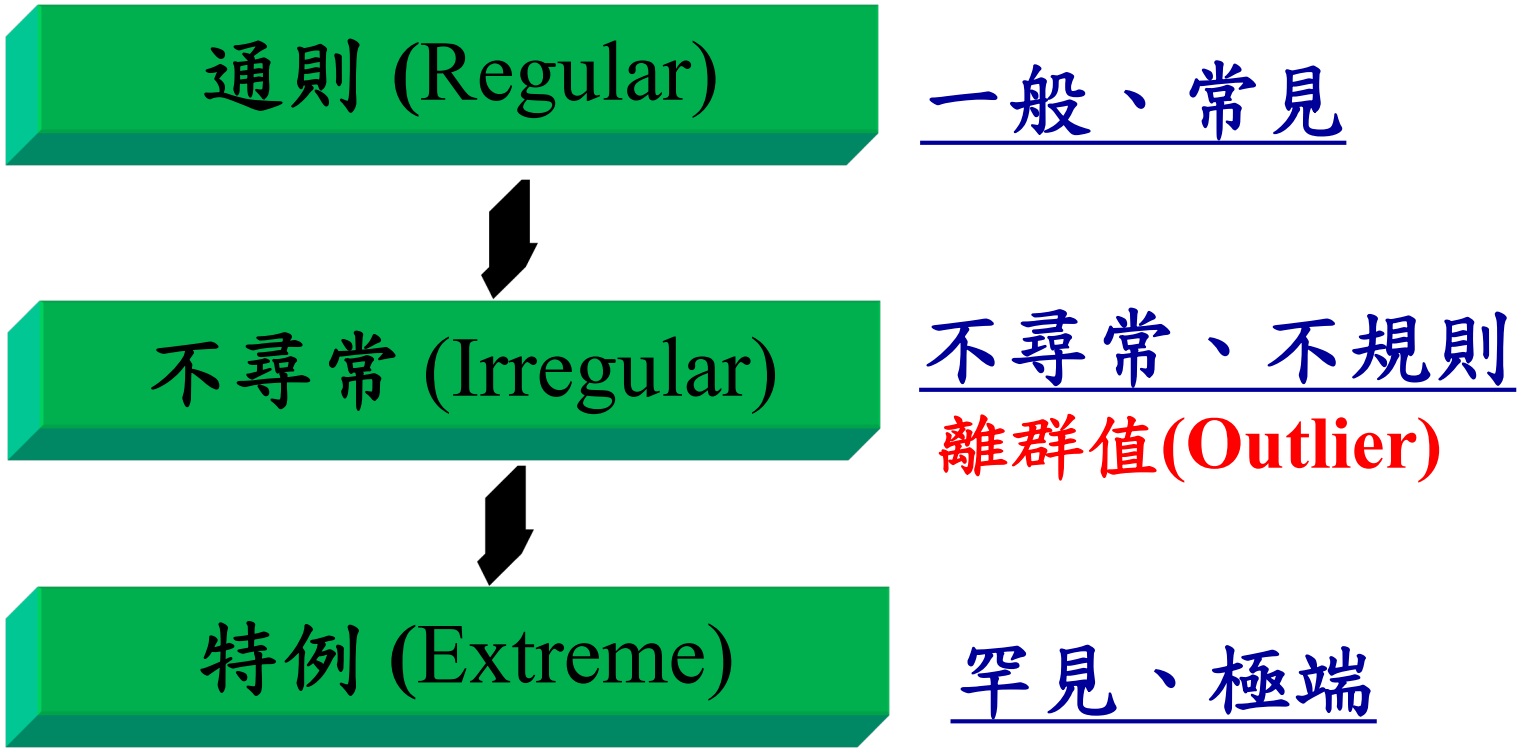
WHAT IS STATISTICS?

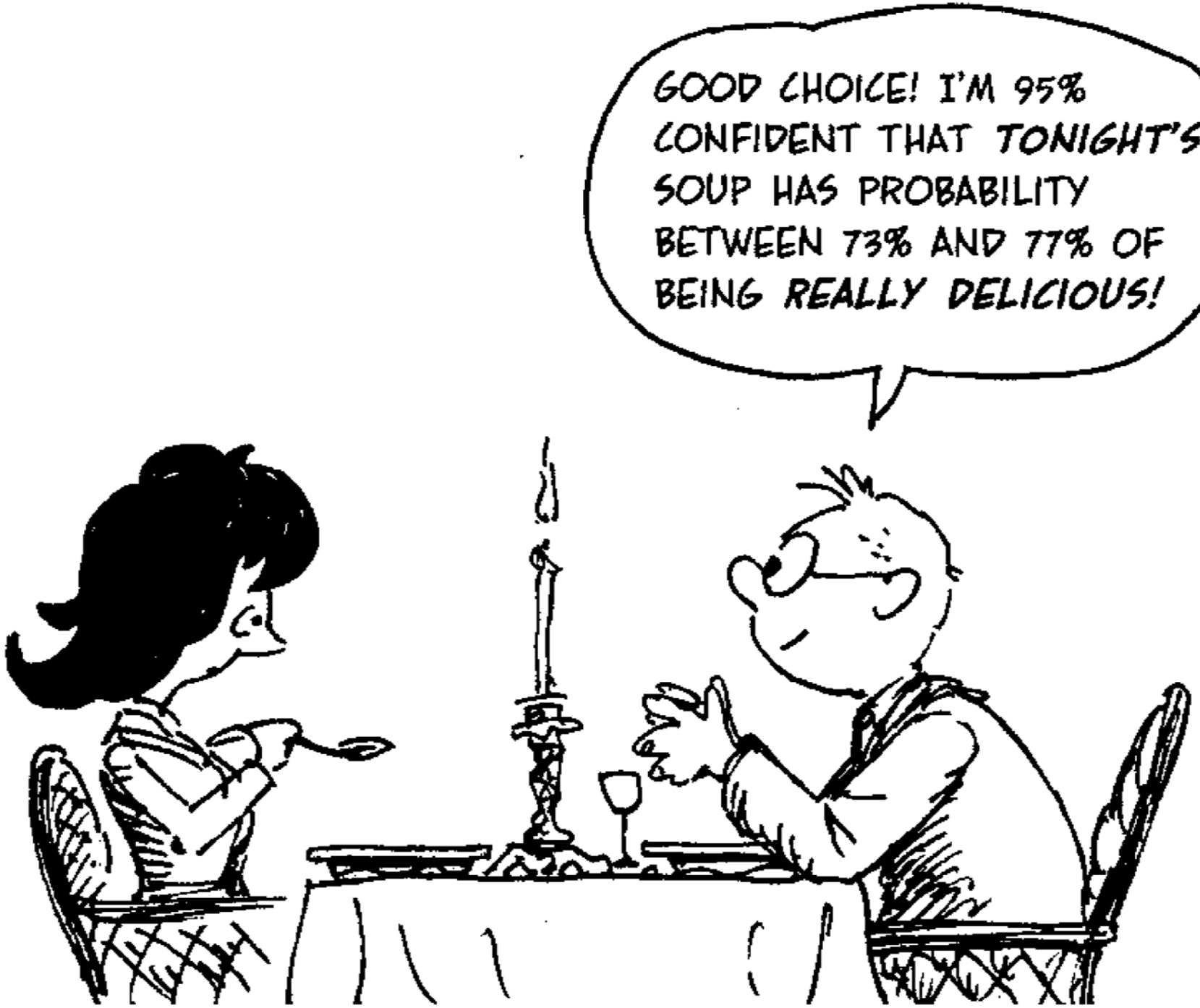
WE MUDDLE THROUGH LIFE MAKING CHOICES BASED ON *INCOMPLETE INFORMATION*...



歸納法

- 歸納法(Induction)，從龐雜的資料找出共同趨勢，並區分資料的特性。




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

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」。

- 資料科學家不只熟悉統計分析，本身的工作內容非常多元(Multi-disciplinary)，需要具有與人溝通、報告撰寫、程式軟體、商業智慧與決策等之能力。

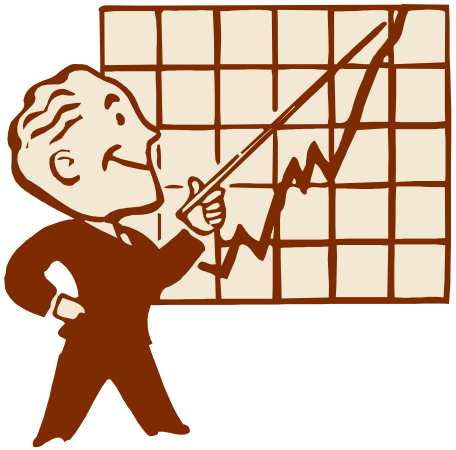
註：現今學校尚無統合訓練（即使有、人數也不多），人才缺額暫時無法補足。

數 據 (Data)

資 訊 (Information)

事 實 (Fact)

知 識 (Knowledge)



統計系畢業後的出路？

- 學長姐畢業後多半從事哪些行業的工作？

→為什麼這些工作比較熱門？

- 以近年電腦科技發展的速度評估，哪種類型的工作未來會比較吃香，這些工作會與統計有關嗎？

→如果有關，你/妳需要具備哪些技能才能提高競爭力？

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)**：文字、視影音等資訊

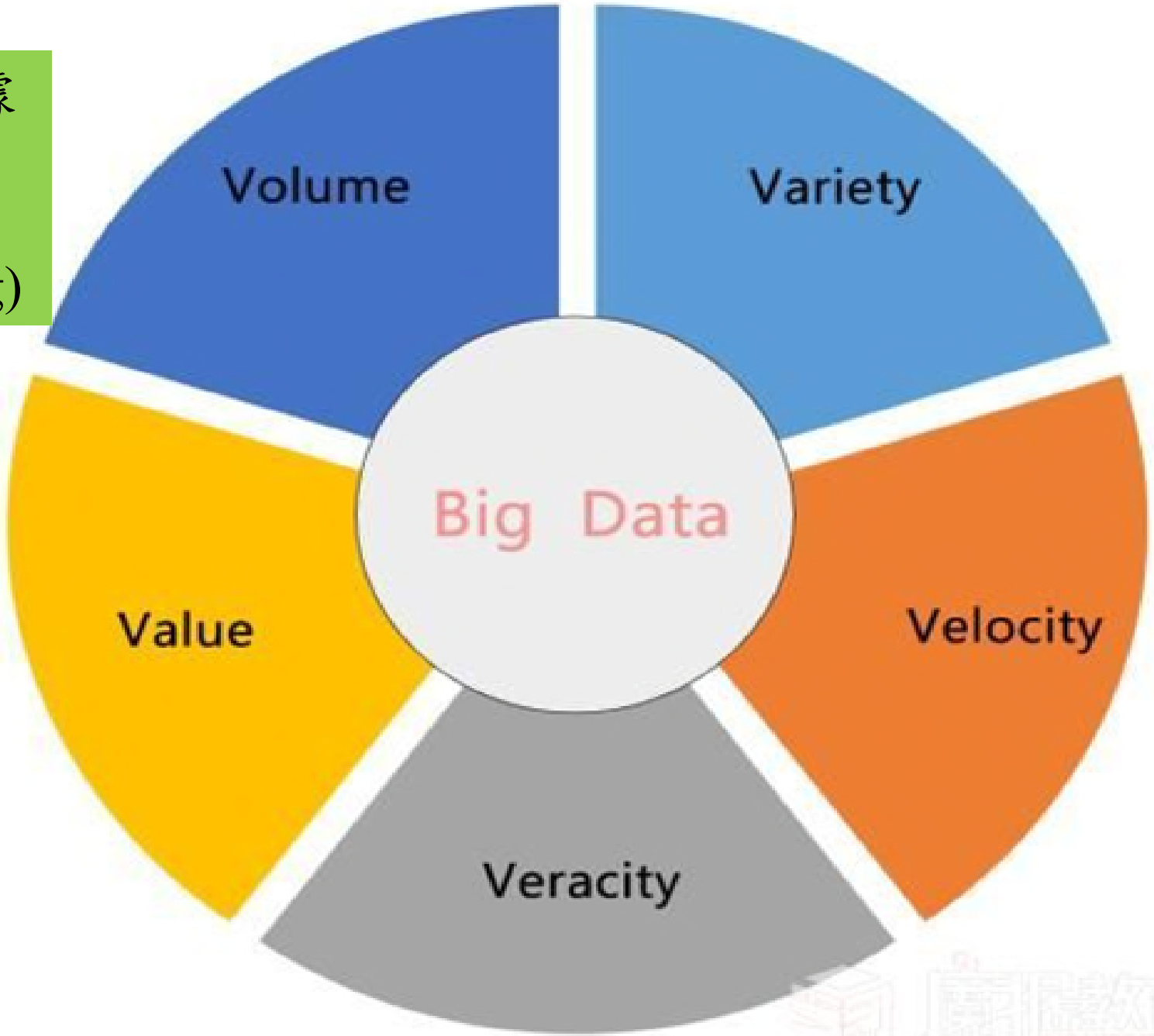
→**真實性(Veracity)**：資料品質（2014年提出）

註：加入**資料分享與傳遞(Visible)** 後成為「**大、快、雜、疑、轉**」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

Value：數據價值密度相對較低。(Datamining)



http://5b0988e595225.cdn.sohucs.com/images/20171205/8cb978ed1d064f27a3bd919e95ba1153.jpeg

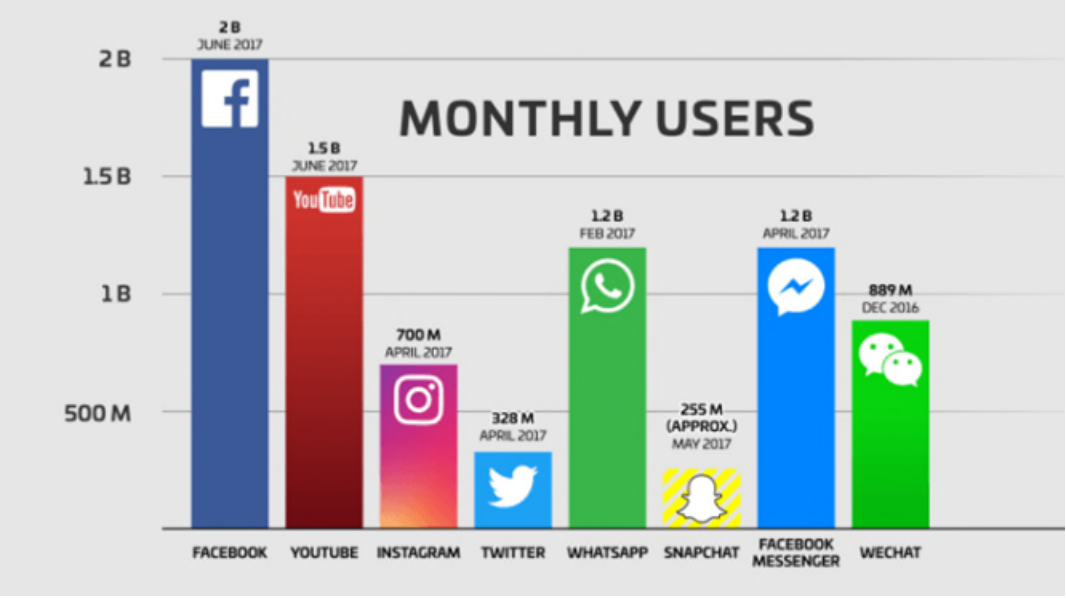
資訊的價值

- 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！）

資料分享與傳遞(另一個特性！)

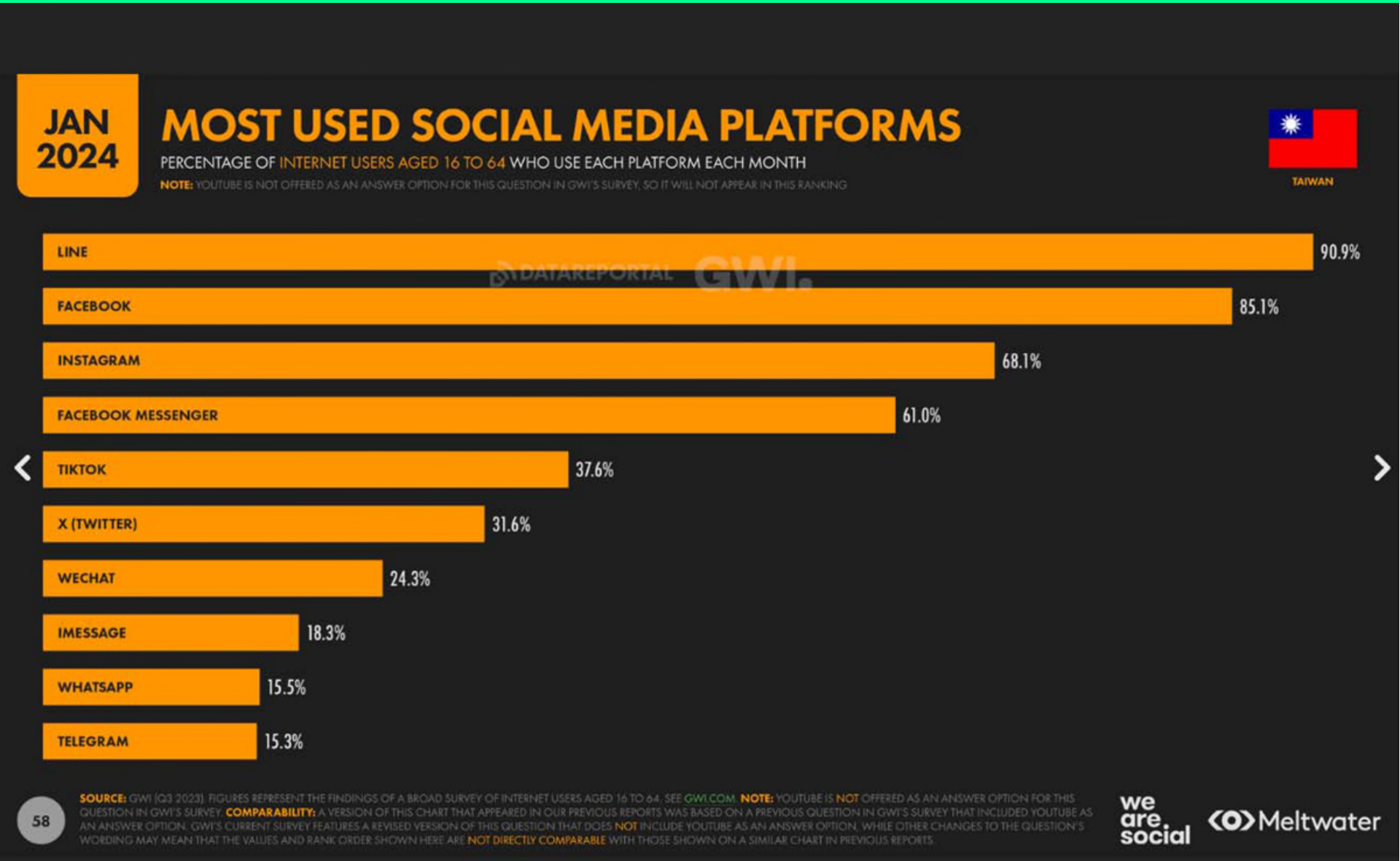
- Facebook成長快速，每月活躍用戶已達20億，每天在Facebook按讚人數高達8億，使用愛心符號回應有1.75億，每月超過10億人使用社團功能。

→Youtube也有10億以上的使用者。



https://techcrunch.com/2017/06/27/facebook-2-billion-users/

資料分享與傳遞（特性之一）



https://www.applemint.tech/en/blog/2024_taiwan_sns/

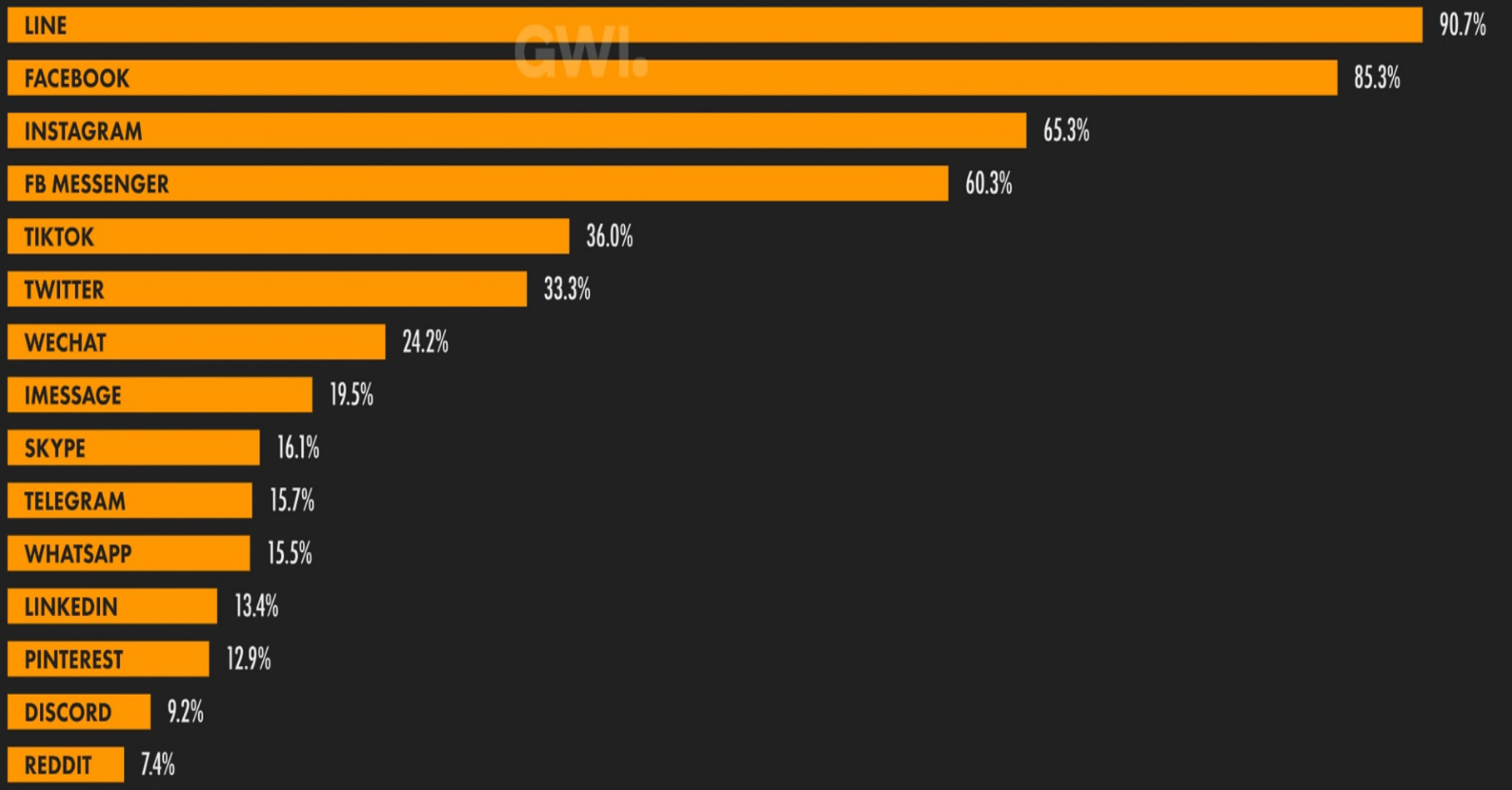
JAN 2023

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



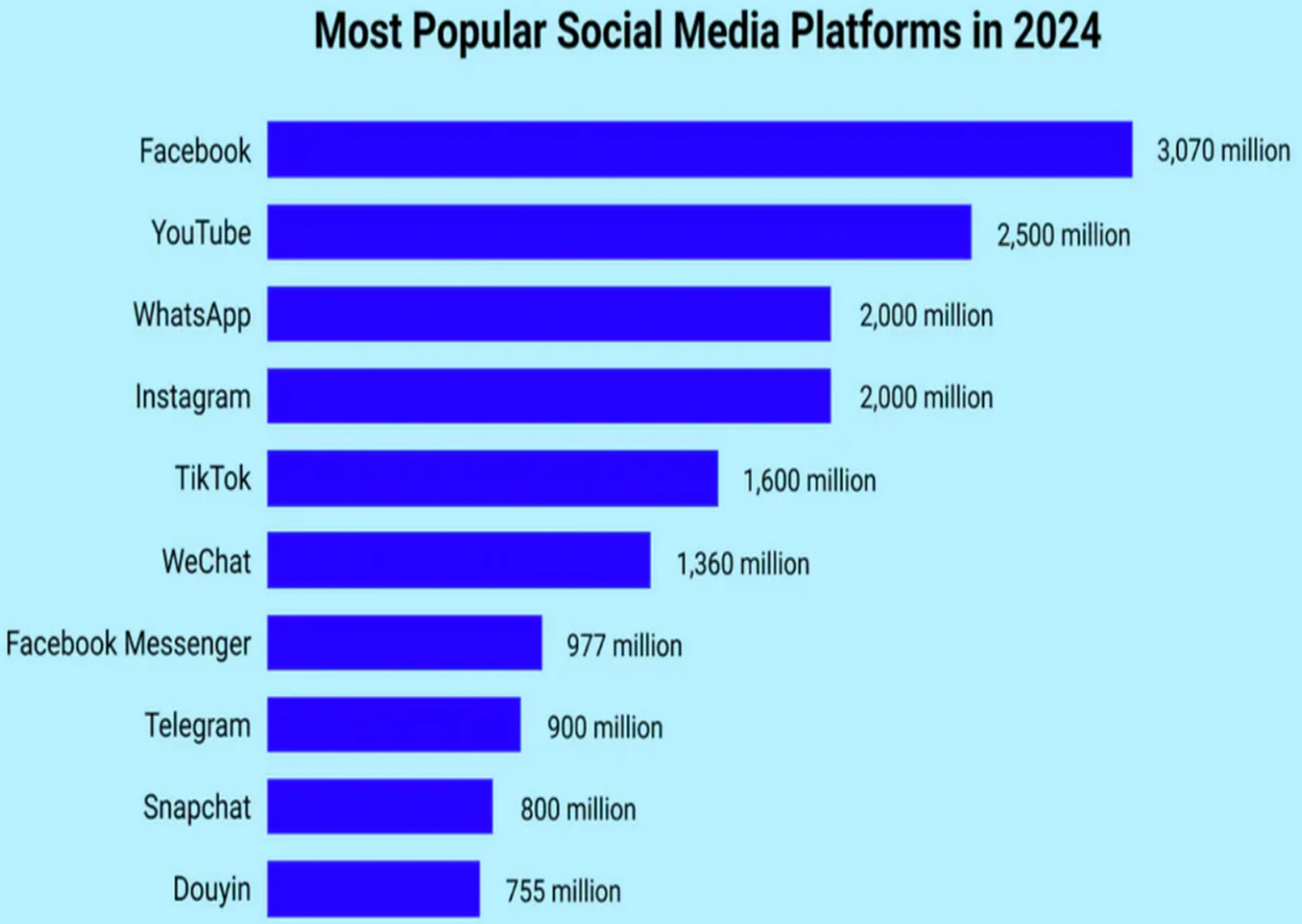
SOURCE: GWI (Q3 2022). FIGURES REPRESENT THE FINDINGS OF A BROAD GLOBAL SURVEY OF INTERNET USERS AGED 16 TO 64. SEE GWI.COM FOR FULL DETAILS. **NOTE:** YOUTUBE IS NOT OFFERED AS AN ANSWER OPTION FOR THIS QUESTION IN GWI'S SURVEY. **COMPARABILITY:** A VERSION OF THIS CHART THAT APPEARED IN OUR PREVIOUS REPORTS WAS BASED ON A PREVIOUS QUESTION IN GWI'S SURVEY THAT INCLUDED YOUTUBE AS AN ANSWER OPTION. GWI'S CURRENT SURVEY FEATURES A REVISED VERSION OF THIS QUESTION THAT DOES NOT INCLUDE YOUTUBE AS AN ANSWER OPTION, WHILE OTHER CHANGES TO THE QUESTION'S WORDING MAY MEAN THAT THE VALUES AND RANK ORDER SHOWN HERE ARE NOT DIRECTLY COMPARABLE WITH THOSE SHOWN ON A SIMILAR CHART IN PREVIOUS REPORTS.

we are social

Meltwater

https://socialgenie.s3.ap-northeast-1.amazonaws.com/wp-content/uploads/2023/09/most-used-social-media-platforms-taiwan.jpg

全球最常使用社群軟體(2024年10月)



Source: DataReportal

shopify

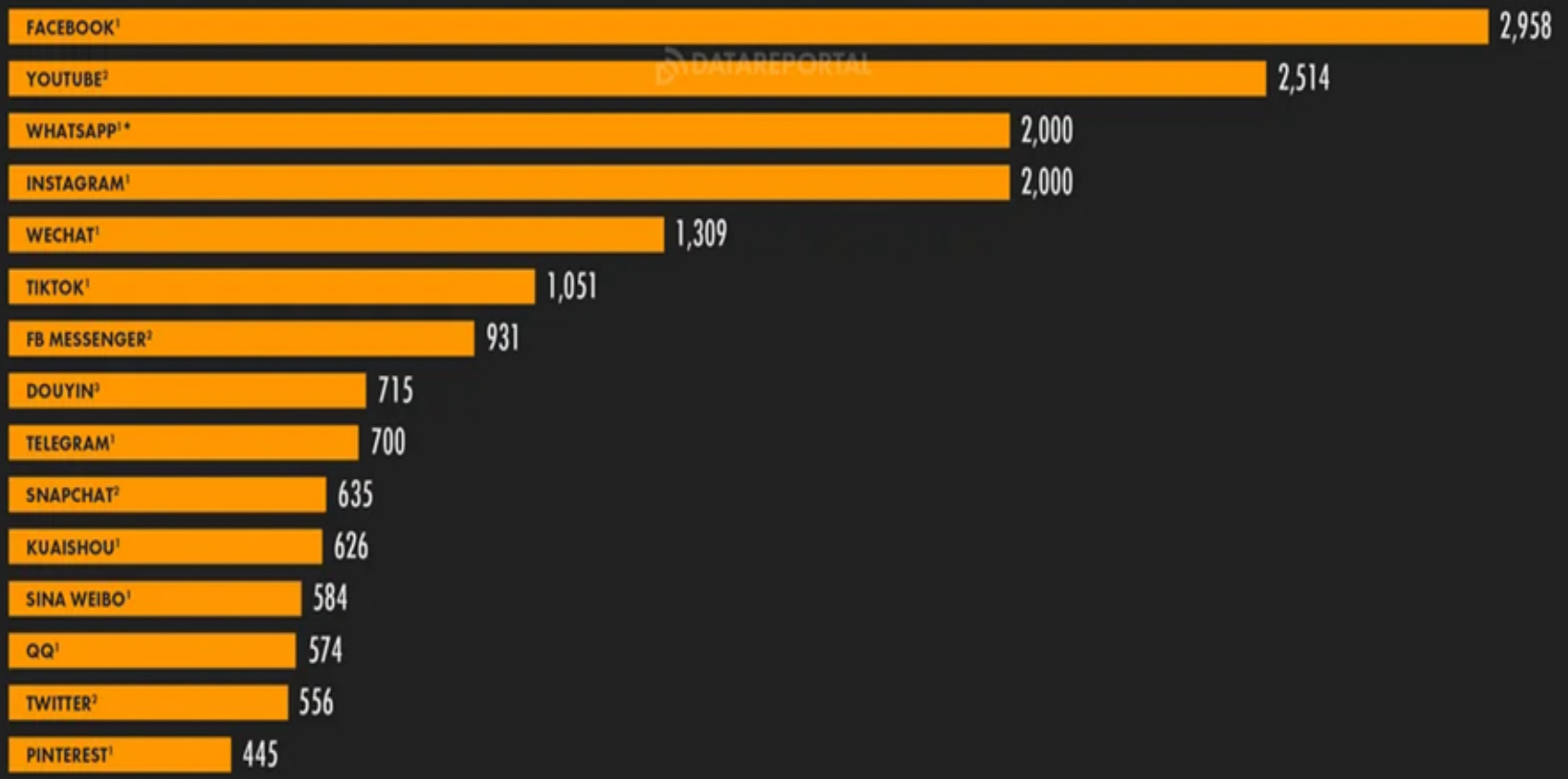
https://www.shopify.com/zh/blog/most-popular-social-media-platforms

we are social

JAN 2023 THE WORLD'S MOST USED SOCIAL PLATFORMS

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

GLOBAL OVERVIEW



182

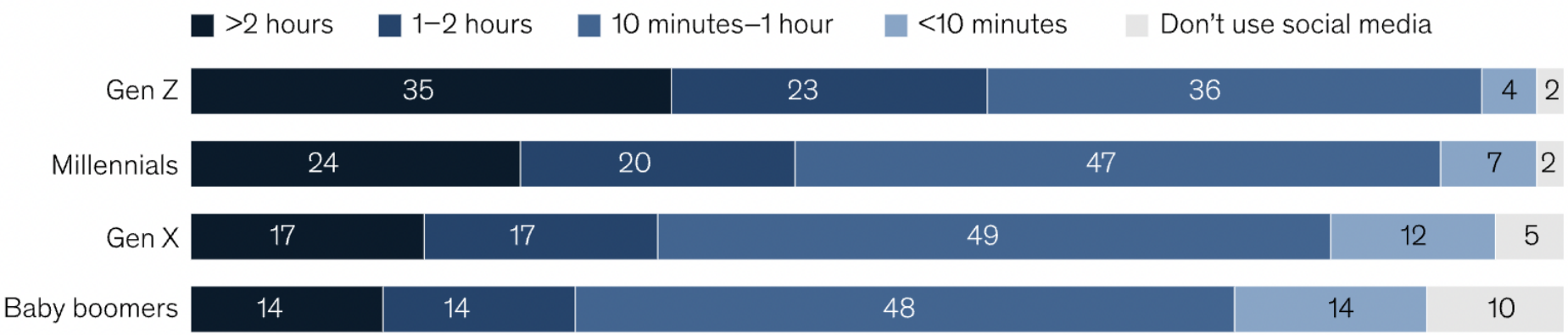
SOURCES: KEPIOS ANALYSIS OF [1] COMPANY ANNOUNCEMENTS OF MONTHLY ACTIVE USERS; [2] PLATFORMS' SELF-SERVICE ADVERTISING RESOURCES; [3] ANALYSYS.CN. **ADVISORY:** USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. **COMPARABILITY:** PLATFORMS IDENTIFIED BY [*] HAVE NOT PUBLISHED UPDATED USER FIGURES IN THE PAST 12 MONTHS, SO FIGURES ARE LESS REPRESENTATIVE. BASE CHANGES AND METHODOLOGY CHANGES IN SOURCE DATA MEAN THAT FIGURES MAY NOT BE DIRECTLY COMPARABLE WITH THOSE PUBLISHED IN PREVIOUS REPORTS.

we are social Meltwater

年齡與社群軟體使用時間(2024)

Time spent on social media daily,[1] % of respondents (n = 41,960)

https://sproutsocial.com/insights/new-social-media-demographics/



- 18-29 years – Snapchat (41%), TikTok (35%), Instagram (32%)
- 30-39 years – LinkedIn (34%), X/Twitter (34%), Snapchat (33%), Instagram (32%)
- 40-49 years – LinkedIn (25%), Facebook (22%), X/Twitter (21%)
- 50-59 years – Facebook (29%), LinkedIn (24%), Pinterest (24%)

2024年
1月

社群媒體使用一覽

社群媒體採用和使用的主要數據（註：一個用戶身分未必代表唯一的個體）

社群媒體用戶總數	上季社群媒體用戶數變化	去年同期社群媒體用戶數變化	平均每日使用社群媒體的時間	每個月平均使用的社群媒體平台數量
1920萬	0% [沒有改變]	-5.0% 減少100萬	1小時59分鐘 較去年同期少7分鐘	6.1

社群媒體用戶數 vs. 總人口	18歲以上社群媒體使用者 vs. 18歲以上人口	社群媒體用戶數 vs. 網路使用者人數	女性社群媒體用戶數 vs.社群媒體用戶總數	男性社群媒體用戶數 vs.社群媒體用戶總數
80.2%	85.0%	88.5%	50.4%	49.6%

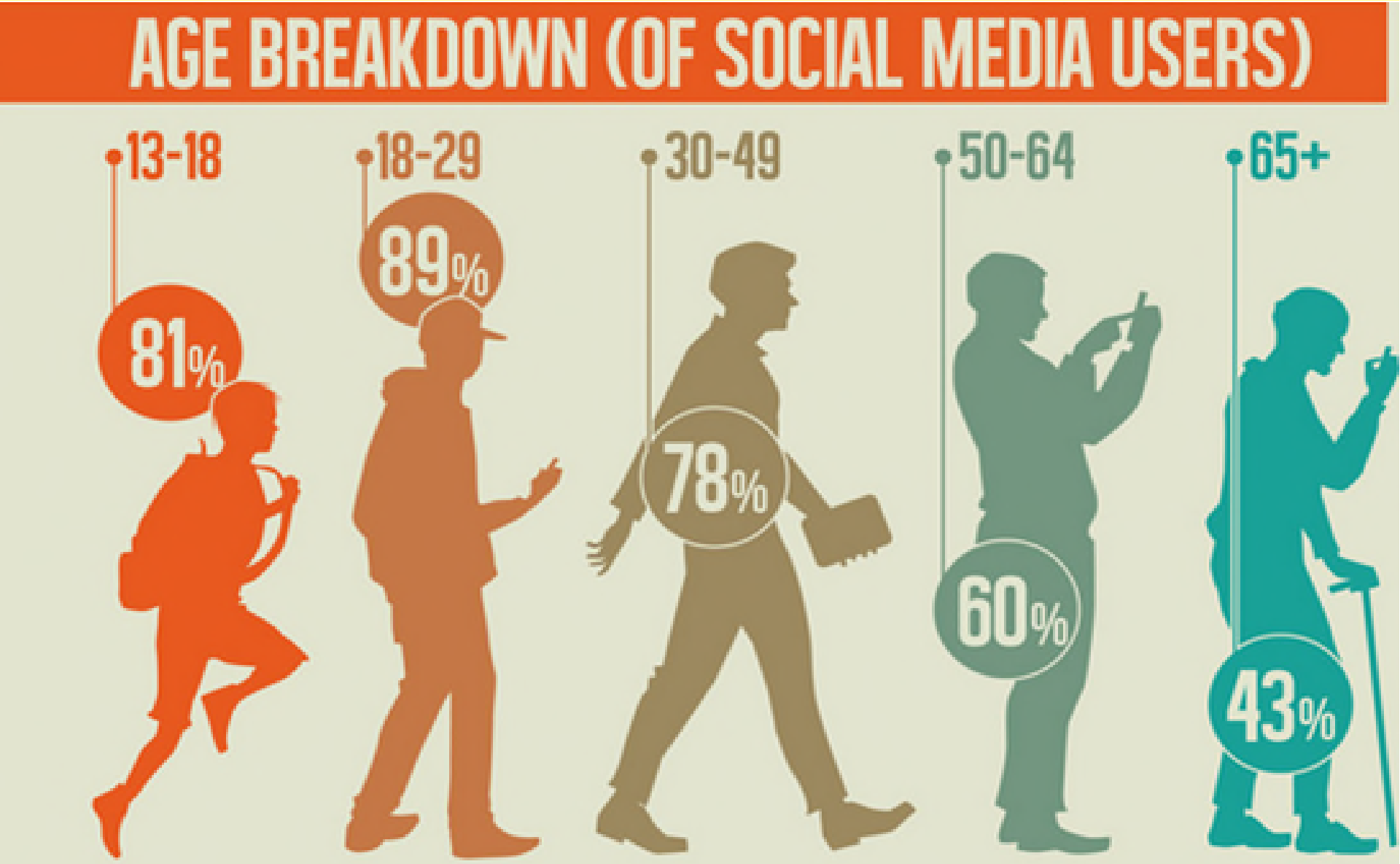
55

SOURCES: KEPIOS ANALYSIS; COMPANY ADVERTISING RESOURCES; CNNIC; BETA RESEARCH CENTER; OCDH; U.N.; GWI (Q3 2023). NOTE: AVERAGE NUMBER OF PLATFORMS INCLUDES DATA FOR YOUTUBE. ADVISORY: SOCIAL MEDIA USER IDENTITIES MAY NOT REPRESENT UNIQUE INDIVIDUALS. COMPARISONS WITH POPULATION AND INTERNET USERS MAY EXCEED 100% DUE TO DUPLICATE AND FAKE ACCOUNTS, USER AGE MISSTATEMENTS, DIFFERENT REPORTING PERIODS, AND DIFFERENCES BETWEEN CENSUS COUNTS AND RESIDENT POPULATIONS. COMPARABILITY: SOURCE AND METHODOLOGY CHANGES; BASE REVISIONS. SEE NOTES ON DATA.

https://www.i-buzz.com.tw/article/article?atype=community&id=775

來源：《Digital 2024: Taiwan》 FANS FEED 編譯

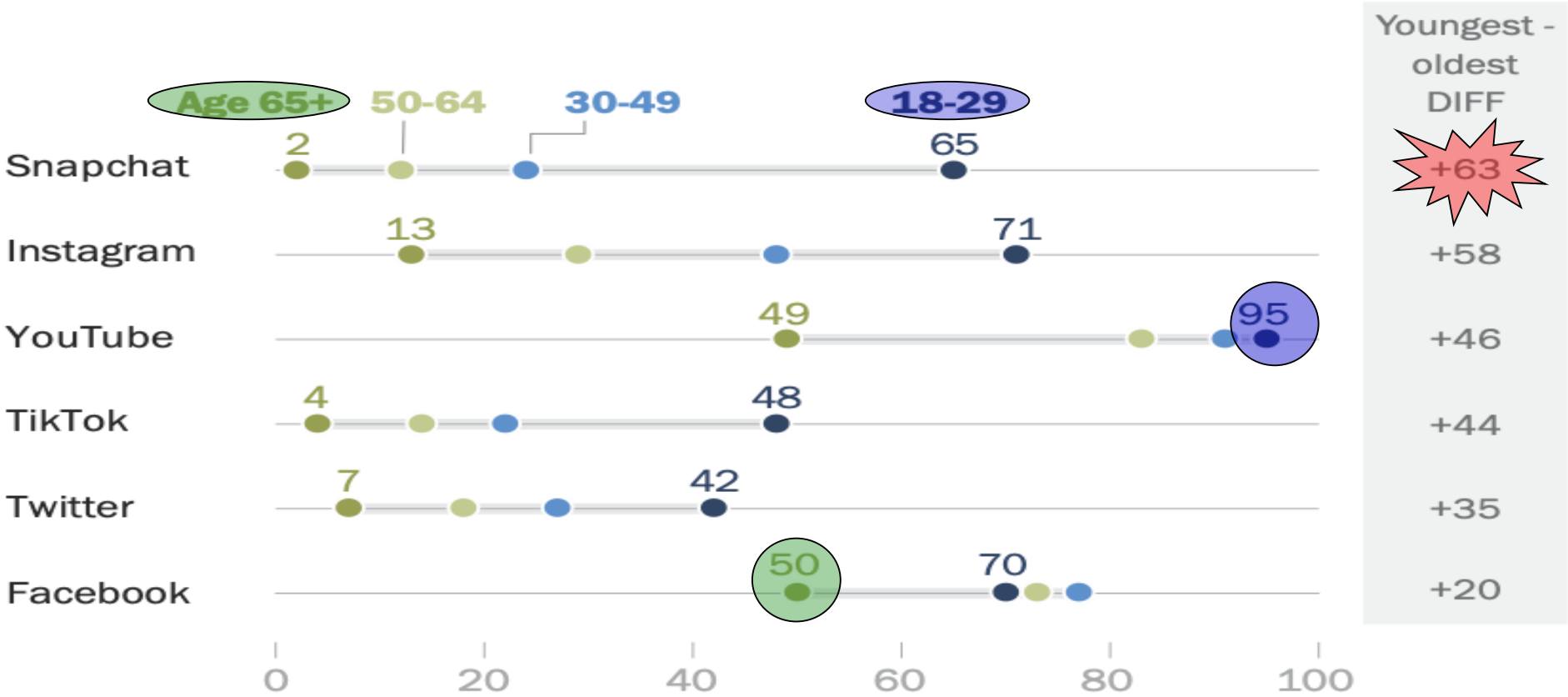
社群軟體也有代溝？(2024)



https://www.business2community.com/social-media-articles/bridging-social-media-marketing-generation-gap-01339396

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.

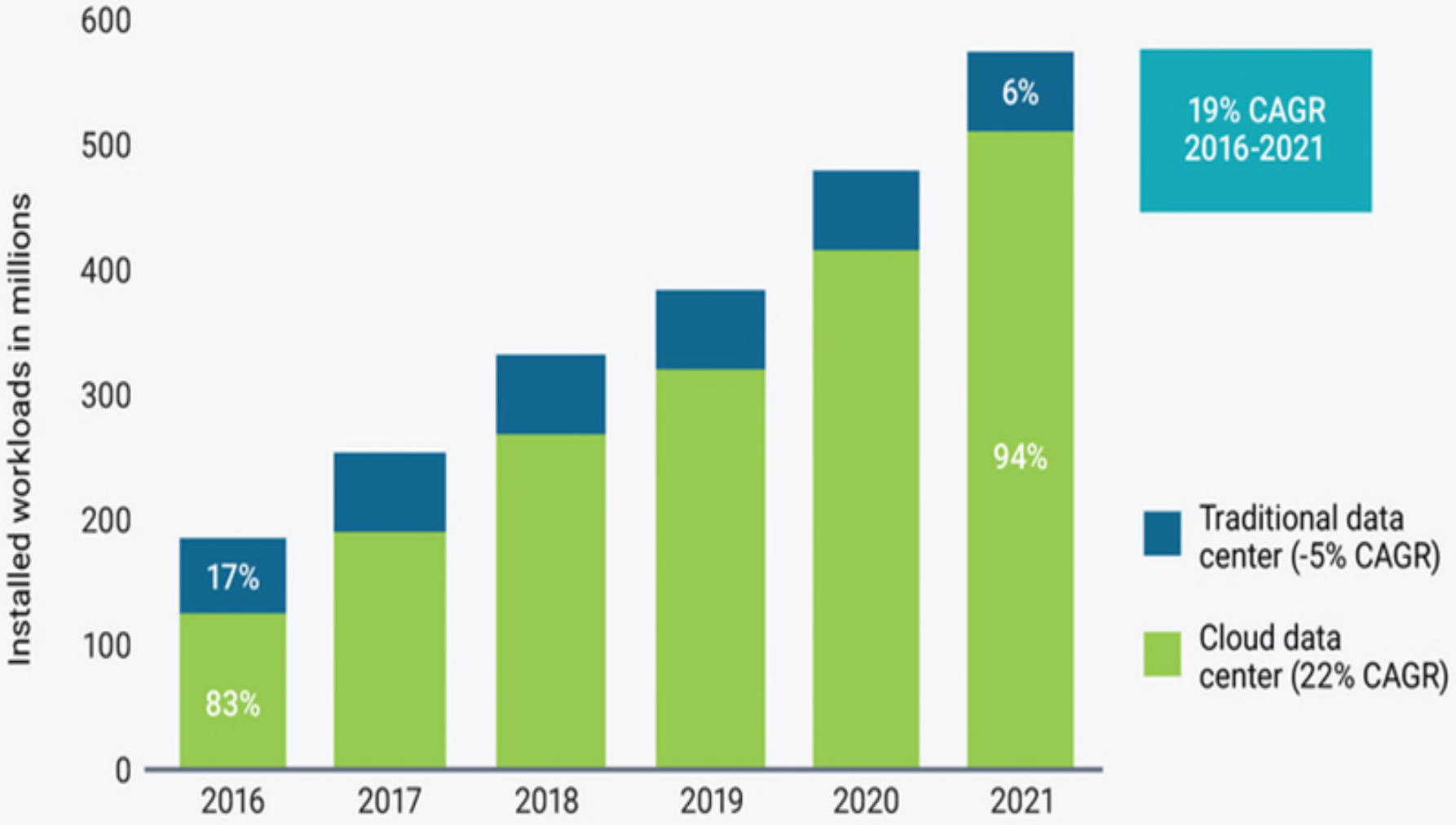
Source: Survey of U.S. adults conducted Jan. 25-Feb. 8, 2021.

"Social Media Use in 2021"

PEW RESEARCH CENTER

https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2021/04/PI_2021.04.07_social-media_0-02a.png

雲端儲存資料更為普遍



Source: Cisco Global Cloud Index, 2016-2021.

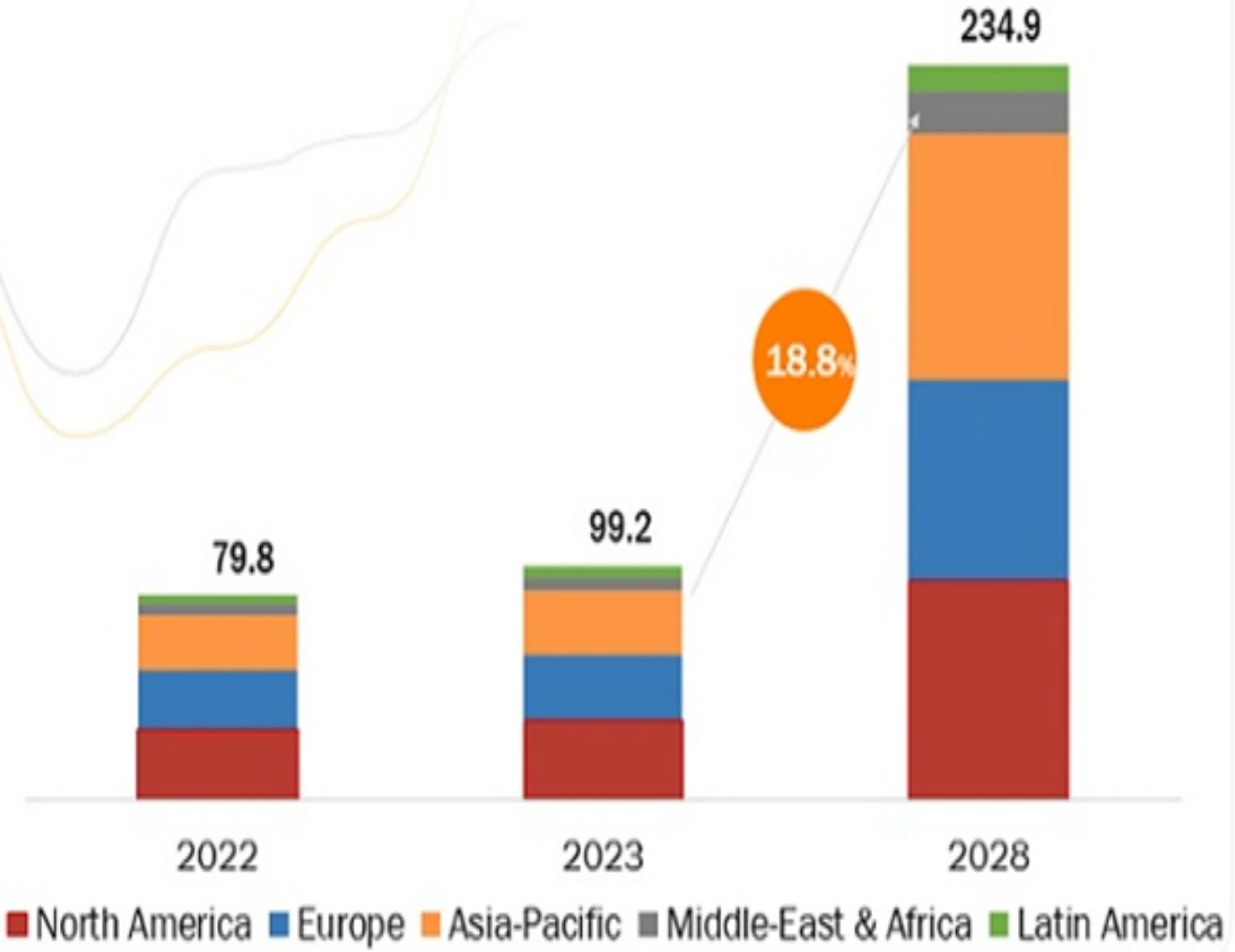
https://www.synopsys.com/content/dam/synopsys/designware-ip/diagrams/converg-cloud-comp-fig1.jpg.imgw.850.x.jpg

CLOUD STORAGE MARKET GLOBAL FORECAST TO 2028 (USD BN)

CAGR OF

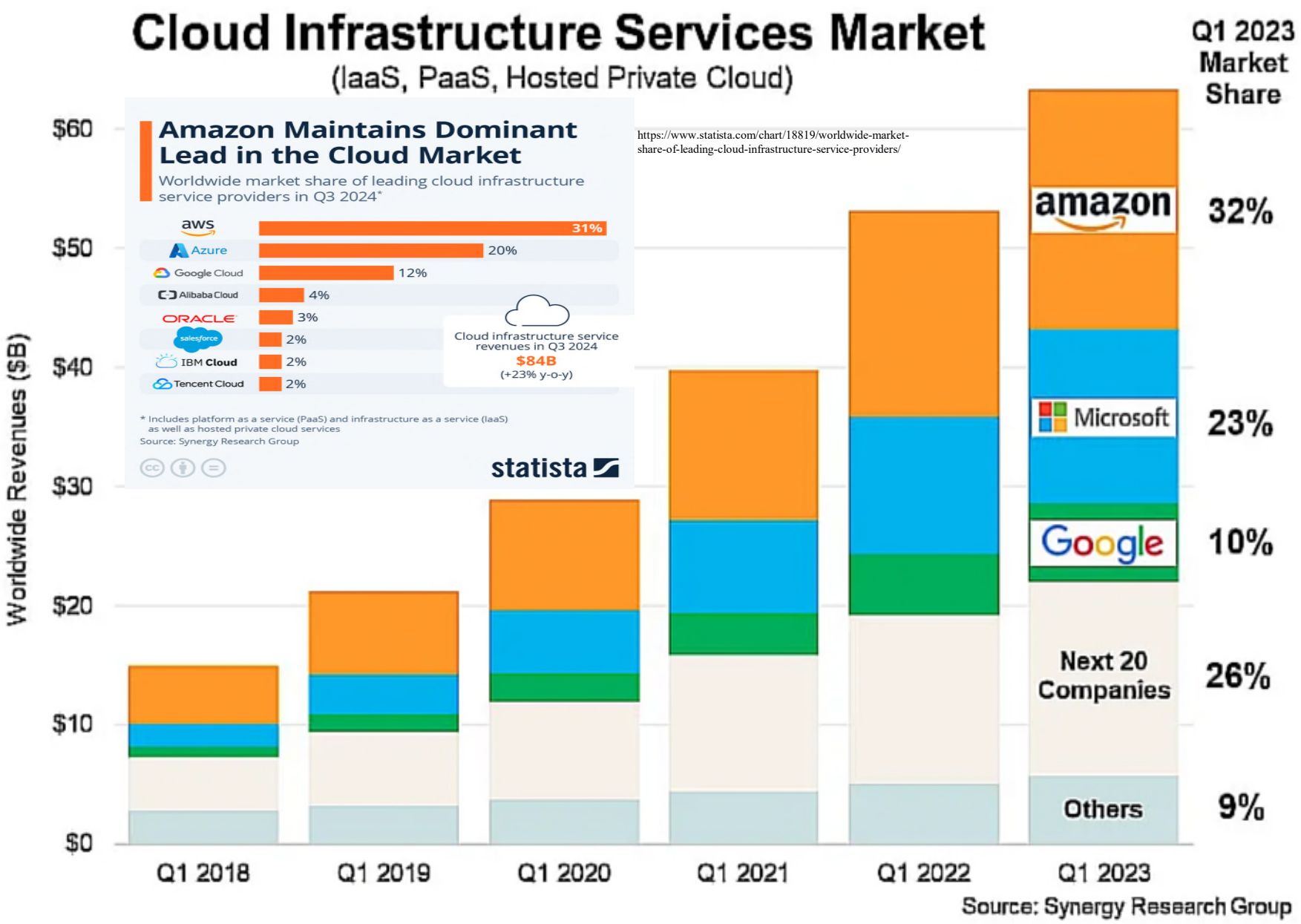
18.8%

The global Cloud Storage market is expected to be worth USD 234.9 billion by 2028, growing at a CAGR of 18.8% during the forecast period.



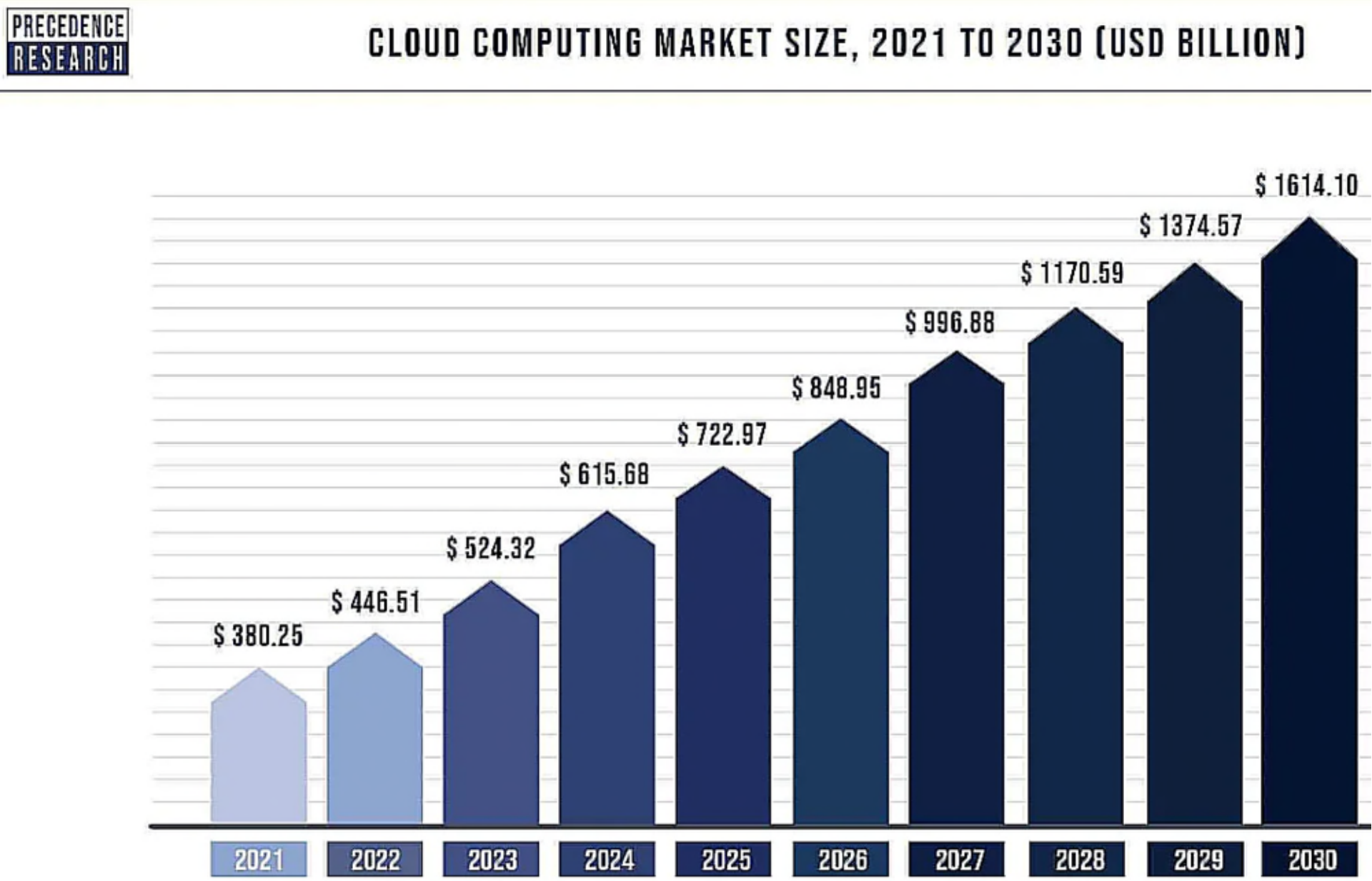
https://www.marketsandmarkets.com/Market-Reports/cloud-storage-market-902.html

雲端基礎設施服務的市場



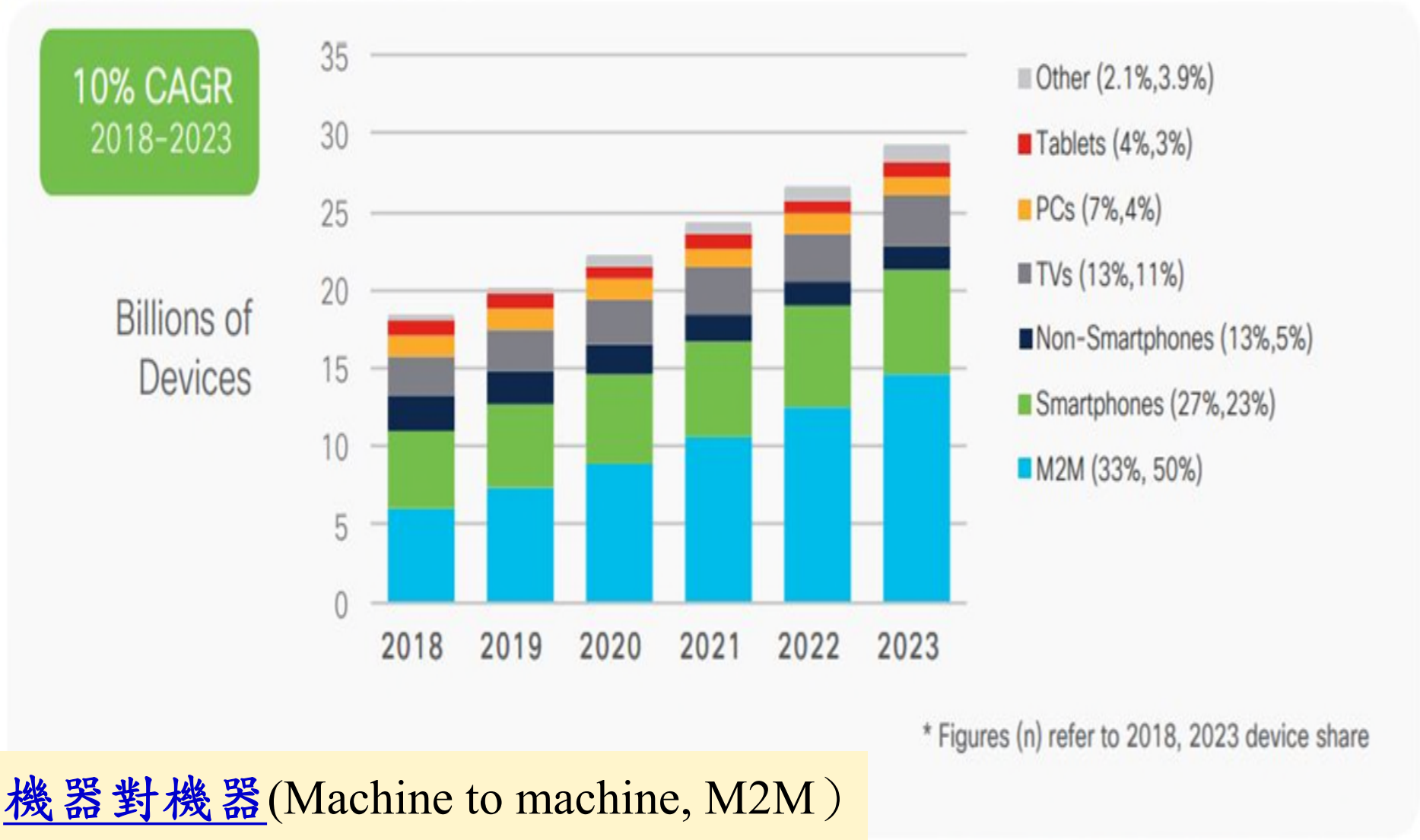
https://www.cloudzero.com/blog/cloud-computing-market-size/

雲端計算市場發展迅速(98%企業使用)



https://www.cloudzero.com/blog/cloud-computing-market-size/

未來趨勢：機器對機器。。。

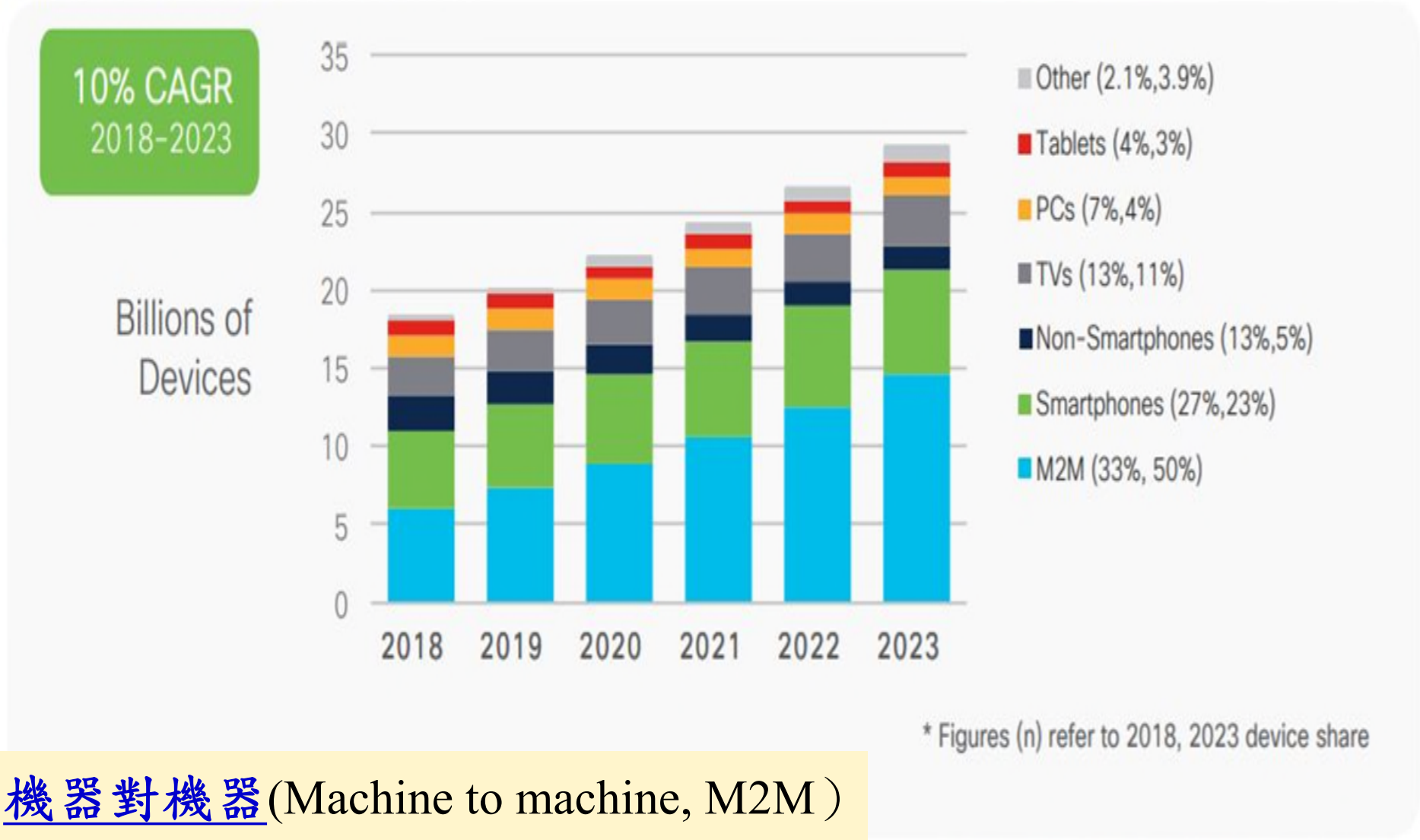


機器對機器(Machine to machine, M2M）機器裝置間在無需人爲干預下，直接透過網路溝通而自行完成任務。

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

未來趨勢：機器對機器。。。



機器對機器(Machine to machine, M2M）機器裝置間在無需人爲干預下，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

2025年美國最佳職業排行榜

排名	職業	年薪（中位數）	職缺人數
1	專業護理師	$126,260	135,500(碩士)
2	IT經理人	$165,510	106,900(學士)
3	外科醫師助理	$130,020	43,700(碩士)
4	財務經理人	$156,100	138,300(學士)
5	軟體工程師	$132,270	303,700(學士)
6	資訊安全分析師	$120,360	59,100(學士)
7	醫療服務經理	$110,680	160,600(學士)
8	**資料科學家**	**$108,020**	**73,100(學士)**
9	語言治療師	$89,290	33,300(碩士)
10	**精算師**	**$120,480**	**6,600(學士)**

https://money.usnews.com/careers/best-jobs/rankings/the-100-best-jobs

統計在大數據的角色

- 數量化資料透過統計分析（經過確認目標、篩選變數後），可獲得關鍵資訊以協助決策及方向。

→如果套入解決問題的流程，統計從業人員需要哪些專業訓練與哪些專業人士合作？

- 問題：以賣場銷售為例，請同學提供可行的操作模式（加上人力、物力等配合），發掘出潛在的價值與商機！

Walmart賣場—尿布與啤酒

- 沃馬特蒐集與分析顧客資料，並以既有資訊提高銷售業績，成為全球最大的連鎖零售業者。其中發現消費者在週末購物時，許多人會同時購買尿布及啤酒。

→為什麼這兩種商品會一起購買？

http://www.bella-consults.com/deceiving-with-diapers

沃爾瑪於1982年發跡於美國西南部，剛開始只是一家小鎮上的雜貨店，因擅於資料分析，現今是全球最大連鎖零售業者，擁有八千多家分店與超過200萬名的員工。

「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：

→商品定價與促銷活動；

→商品擺設與賣場動線；

→物流倉儲與人員管理。

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

「Pokémon GO Safari Zone in Tainan」今（1）日早上10點半開始將連續5天展開，而這場寶可夢盛會預計帶來的人潮預估將有20萬，各大飯店民宿住房率超過8成，甚至有業者開心表示，「比陸客來台還有用！」

三立新聞網 2018年11月1日 上午11:06

廣告

Pokémon GO CITY SAFARI

2024 年 3 月 9-10 日

City Safari 台南

臺南 TAINAN

龍耀臺南 GLORIOUS TAINAN

https://cpok.tw/47201

https://www.youtube.com/watch?app=desktop&v=wPQxZq2s0oE&ab_channel=%E5%8F%B0%E8%A6%96%E6%96%B0%E8%81%9ETTVNEWS

台南週末訂房滿 網哀嚎沒飯店

15萬人湧入抓寶！

台南寶可夢大爆發

週六高鐵票已賣光

統計數據與解決問題

- 隨著科技發達，使得蒐集及儲存資料的成本降低，到處都充斥著統計數字，即使政策實施也搭配民意調查，似乎統計已成為生活的一部份。

→但統計數字真的可以反映事實，或甚至是解決問題不可或缺的要件嗎？

註：是否存在訴諸統計的迷思？

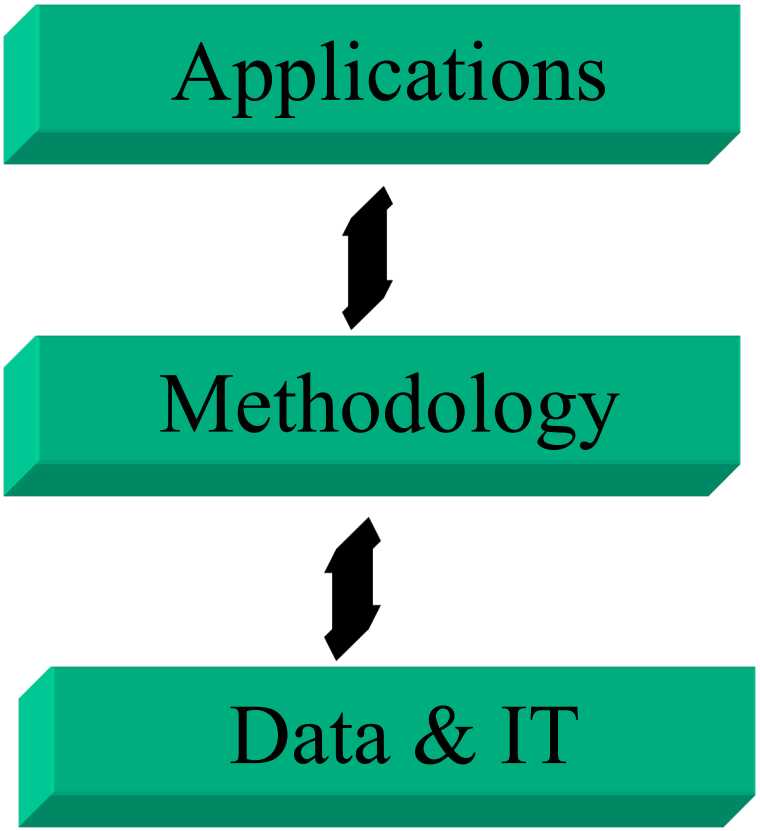
決策優勢與資訊能力

- 過去擁有土地、資金的資本家，透過勞力及資本密集而有絕對優勢；21世紀轉變為知識經濟時代，擁有充分資訊的人有絕對優勢。(Amazon、Google！)

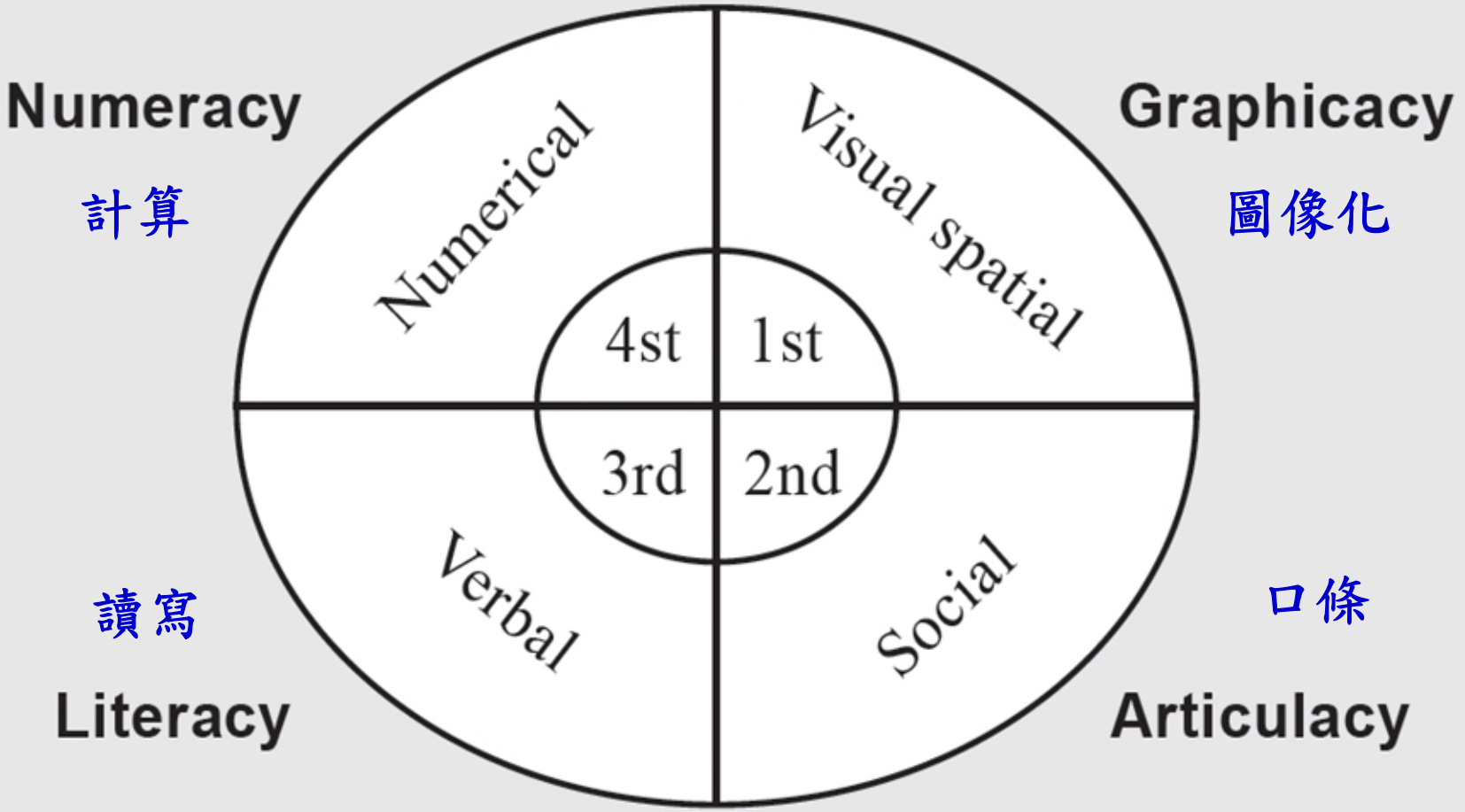
→未來優勢為資料、處理及分析資料的能力（軟實力），因為被複製的可能性高，需要有法律保障（智慧財產權、版權）。

統計與實務應用

- 統計分析與方法論(Methodology或是Model)有關，需要配合應用領域的專業知識以及相關的資料庫。



統計學家需具備的能力



Source: The Completely Sufficient Statistician

充分完備的統計學家

- W.G.V. Balchin (The American Cartographer) illustrated in 1976 that humans evolved by first developing keen **visual spatial** skills, then **social** skills, **verbal** skills, and **numerical** skills.

→ Numeracy — formulating & solving problems using mathematics and computing

→Articulacy — speaking & listening (people skills)

→Literacy — writing & reading

→Graphicacy — producing & understanding graphics

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。

統計學家的角色

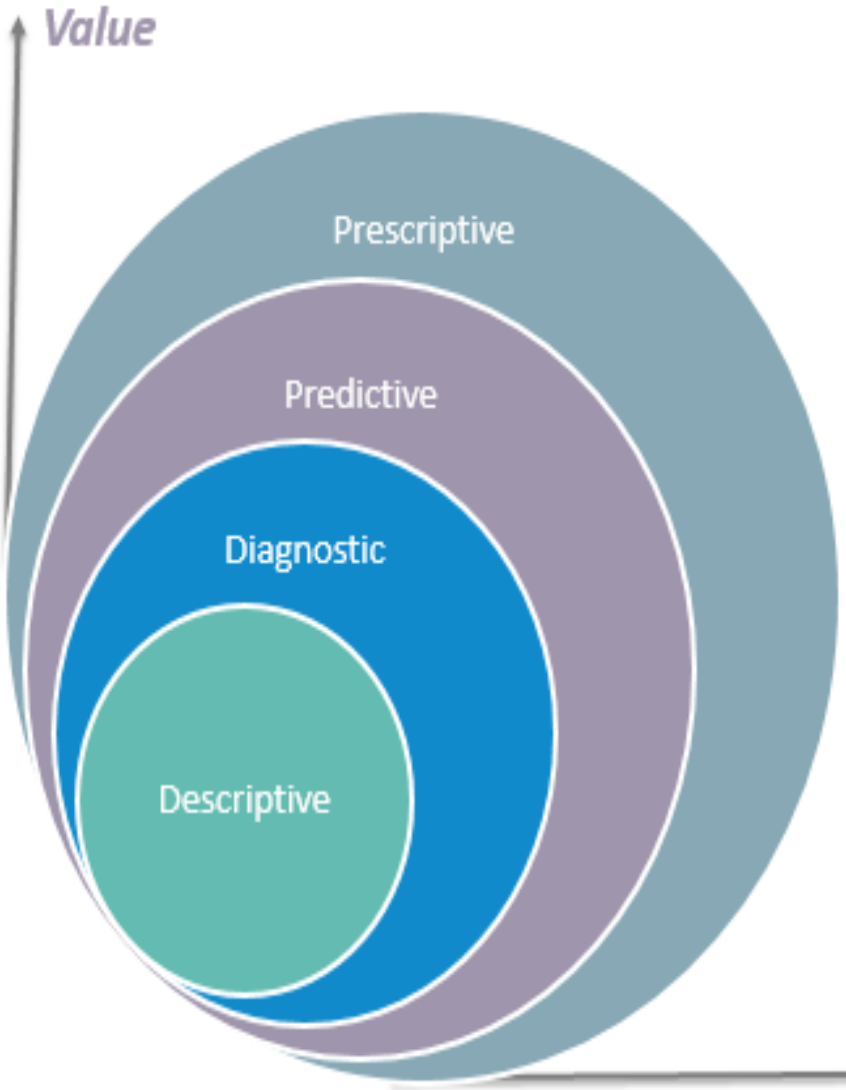
- 我們可以自詡為統計學家，覺得自己應該具備哪些特質？（或是畢業時具有哪些專長。）

→參考「Career Guide for Statistician.htm」的建議，其中列出了十項與統計學家相關的技巧(SKILLS)、知識(KNOWLEDGE)、 能力(ABILITIES)、工作(TASKS)。

註：你/妳覺得自己最需要（或缺乏）哪一方面的訓練？

	統計學家應具備的10項技能(Career Guide for Statistician)
1	選擇正確的數學方法去解決問題 Choose the right mathematical methods or formulas to solve a problem
2	能夠快速且正確計算加減乘除的數字能力 Add, subtract, multiply, or divide quickly and correctly
3	對於特定問題，能夠透過一般的理論給出易懂的答案 Apply general rules to specific problems to produce answers that make sense
4	能夠透過書面表達來清楚溝通自己的想法 Communicate information and ideas in writing so others will understand.
5	能夠將零散的訊息整合為通則或結論 Combine pieces of information to form general rules or conclusions (includes finding a relationship among seemingly unrelated events)
6	觀察細節的能力 See details at close range (within a few feet of the observer)
7	擁有閱讀與解讀書面資訊的能力 Read and understand information and ideas presented in writing.
8	能夠將事物或行為以特定順序或規則進行安排 Arrange things or actions in a certain order or pattern according to a specific rule or set of rules (e.g., patterns of numbers, letters, words, pictures, mathematical operations.
9	能夠快速地將理解、整合與組織訊息，使其為有意義的知識 Quickly make sense of, combine, and organize information into meaningful patterns.
10	擁有歸納與推理的能力 Generate or use different sets of rules for combining or grouping things in different ways

4 types of Data Analytics



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

http://www.kdnuggets.com/2017/07/4-types-data-analytics.html

分析方法的類型

- 統計觀點可分為兩類：

→探索性資料分析(Exploratory Data Analysis)

→驗證性資料分析(Confirmatory Data Analysis)

- 機器學習觀點分為三類：

→敘述性分析(Descriptive Analytics)、預測性分析(Predictive Analytics)、建議性分析(Prescriptive Analytics)；

→「發生了什麼事」(What has happened)、「未來會如何」(What would happen)、「我們如何調整」(What should we do)。

統計與解讀資訊

http://h30507.www3.hp.com/t5/Journey-through-Enterprise-IT/Analyze-This-Big-Data-is-insurance-against-losing-a-competitive/ba-p/143577#.UgZmpLQVEqQ

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？（資料愈多愈好？）

→資料品質？(「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？）

統計諮詢的價值

- 統計學家必定非常注重實驗之邏輯性。

→統計學家承傳了科學研究方法，因為他們所關心、所寫、所做的皆是關於科學研究方法的事物。(Hooke, 1980)

→統計學家終將承擔所有研究發明在邏輯方面的重擔，必須對此一角色認同，且做好面對一堆看不出資訊的原始資料的準備。(Price, 1982)

Table 1. The Old View of the Statistician's Role.

Statistician	Client	The role of the statistician:
Passive	Passive	None
Active	Passive	Crusader
Passive	Active	Helper
Active	Active	Colleague

From Snee (1991); adapted from Hunter (1981).

Table 2. The New View of the Statistician's Role.

If the statistician is:	And the client is:	And the organization is:	The statistician's role is:
Passive	Passive	Passive	None
Active	Passive	Passive	Crusader
Passive	Active	Passive	Helper
Active	Active	Passive	Colleague
Passive	Passive	Active	Teacher
Active	Passive	Active	Leader
Passive	Active	Active	Data blesser
Active	Active	Active	Collaborator

From Snee (1991).

統計諮詢的進行步驟

- Make sure you understand the problem.
- Plan the investigation and collect the data in an appropriate way.
- Assess the structure and quality of the data.
- Examine initially the data, obtaining summary statistics, graphs, and tables.
- Carry out an appropriate statistical analysis.
- Compare the findings with previous results.
- Interpret and communicate the results.

統計諮詢的進行步驟(續)

- 建立共識(establishing rapport)
- 確認研究的問題(identifying the research problem)
- 設定目標(setting goals)
- 確立責任分工(agreeing on a division of responsibility)
- 總結(summing-up stage)。

統計諮詢中的常用分析

- 生活上的統計應用，不見得會用到非常複雜的分析方法，重點在於按部就班，找出資料的特性與關鍵，而不是譁眾取寵的使用最熱門的分析。

→建議先使用初步或探索性資料分析(Initial Data or Exploratory Data Analysis)，例如：資料的敘述性統計量、以圖形或表格整理資料特性。(EDA vs. CDA! Confirmatory DA)

常見的統計分析

- 敘述性統計量(Descriptive Statistics)
- 相關性分析
- 卡方檢定(Chi-Square Test)
- 多變量分析(Multivariate Analysis)
- 其他方法（例如：迴歸及類別資料分析、時間數列、存活分析；廣義線性模型，Generalized Linear Models）

Predictive Analytics (PA) Exam

- Predictive Analytics exam requires analysis of a data set in the context of a business problem and a written report.

→Candidates are expected to be have knowledge of probability, mathematical statistics, and selected analytical techniques, as well as ability of R within the R-Studio environment.

註：PA要求精算師具有<u>解決問題的能力</u>。

2. **Topic: Problem Definition**

Learning Objectives

The Candidate will be able to identify the business problem, how the available data relates to possible analyses, and use the information to propose models.

3. **Topic: Data Visualization**

5. **Topic: Data Issues and Resolutions**

6. **Topic: Generalized Linear Models**

Learning Objectives

The Candidate will be able to describe and select a Generalized Linear Model (GLM) for a given data set and regression or classification problem.

9. **Topic: Communication**

Learning Objectives

The Candidate will be able to effectively communicate the results of applying predictive analytics to solve a business problem.

Lesson Goals

- Understand the basic concepts of the learning problem and why/how machine learning methods are used to learn from data to find underlying patterns for prediction and decision-making.
- Understand the learning algorithm trade-offs, balancing performance within training data and robustness on unobserved test data.
- Differentiate between supervised and unsupervised learning methods as well as regression versus classification methods.
- Understand the basic concepts of assessing model accuracy and the bias-variance trade-off.
- Become familiar with using the R statistical programming language and practice by exploring data using basic statistical analysis.

Ethical and Responsible Use of Data and Predictive Models Certificate Program

Learning Objectives

Module 1 – Ethical and Responsible Use of Big Data and Predictive Models

Learning Objectives

At the conclusion of Module 1 the participant will be able to:

1. Understand the ethical concerns around using data and predictive models.
2. Get motivated through examples in the insurance industry as well as non-insurance industries.
3. Understand context specificity to insurance.

Module 2 – Regulation & Ethical Framework

Learning Objectives

At the conclusion of Module 2 the participant will be able to:

1. Distinguish between ethical and legal issues
2. Understand how market and regulatory context affects models (in the sense that what is permissible varies by jurisdiction and product)
3. Understand a general ethical framework for evaluating data and analytics use cases

References

- An Introduction to Statistical Learning, with Applications in R (2013), by G. James, D. Witten, T. Hastie, and R. Tibshirani.
- The Elements of Statistical Learning (2009), by T. Hastie, R. Tibshirani, and J. Friedman.
- Learning from Data: A Short Course (2012), by Y. Abu-Mostafa, M. Magdon-Ismail, and H. Lin.
- Machine Learning: A Probabilistic Perspective (2012), by K. Murphy
- R and Data Mining: Examples and Case Studies (2013), Y. Zhao.