巨量資料與統計分析

政治大學統計系余清祥

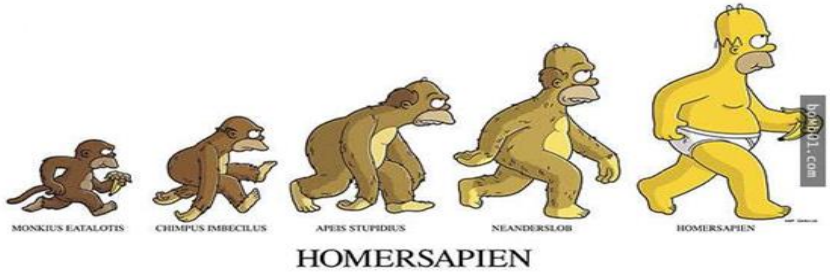
2024年9月24日

第二週：資料科學家

http://csyue.nccu.edu.tw

http://www.bomb01.com/upload/news/original/c29ad198f2f1c77f3b7e7dc8fe636d70.jpg

資訊與決策



- 過去擁有土地、資金等的資本家，因為勞力及資本密集具有優勢；21世紀是知識經濟的時代，擁有及解讀資訊者掌握絕對優勢。(Amazon、Google！)
- 問題：人類歷史（包括石器時代）各年代的生存關鍵是什麼？

註：天下文化《人類大命運》《人類大歷史》。

https://dc-cdn.s3-ap-southeast-1.amazonaws.com/dc-Cover-lnvgt2570vq42r1c2faocrn1v5-20160310202923.Medi.jpeg

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

量化分析與解決問題

- 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，即使政策實施也要和大數據扯上關係。

→透過分析大數據可以窺探現象面以外的事實，充分展露資訊的附加價值。

註：大數據分析的考量因素，包括**如何挖掘出資料的價值**？哪些資料為必要？是否存在訴諸統計的迷思？（例如：從眾效應，Bandwagon effect）

google.org Flu Trends

Google.org home

Dengue Trends

Flu Trends

Home

Select country/region

How does this work?

FAQ

Flu activity

- Intense
- High
- Moderate
- Low
- Minimal

Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

大數據的統計考量重點

- **Veracity**及**Variety**與統計關連較為密切：

→廣義的真實性涵蓋資料品質、相關資料篩選（如抽樣概念）等；

→多樣性則與數量化資料、不同類型資料的整合、選取適當的分析方法等。

- 統計學是研究定義問題、透過資料蒐集、整理、陳示、分析與推論等科學方法，在不確定情況下，做出合理決策的科學。
- 問題：為什麼大家對資料科學家趨之若鶩？

2024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2024年最佳工作排行榜：

I、**大數據相關**（智慧經濟、資訊安全）

- 精算師、資料科學家、資訊安全分析師、軟體設計師、IT經理；

II、**醫療健康**（人本、老化）

- 語言治療師、醫療服務經理、專業護理師、外科醫師助理。

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

2024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	$121,610	118,600(碩士)
2	財務經理人	$139,790	126,600(學士)
3	軟體設計師	$127,260	410,400(學士)
4	IT經理	$164,070	86,000(學士)
5	外科醫師助理	$126,010	39,300(碩士)
6	醫療服務經理	$104,830	144,700(學士)
7	資訊安全分析師	$112,000	53,200(學士)
8	**資料科學家**	**$103,500**	**59,400(學士)**
9	**精算師**	**$113,990**	**7,000(學士)**
10	語言治療師	$84,140	33,100(碩士)

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

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%

https://www.indeed.com/career-advice/news/best-jobs-of-2023

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)**

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

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。

充分完備的資料科學家（個人觀點）

- 資料科學家需要下列「溝通」能力：

→**與人溝通**：寫作、口語表達、溝通能力；

→**與資料**（及統計理論）**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。

- **定義問題、溝通能力！**

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

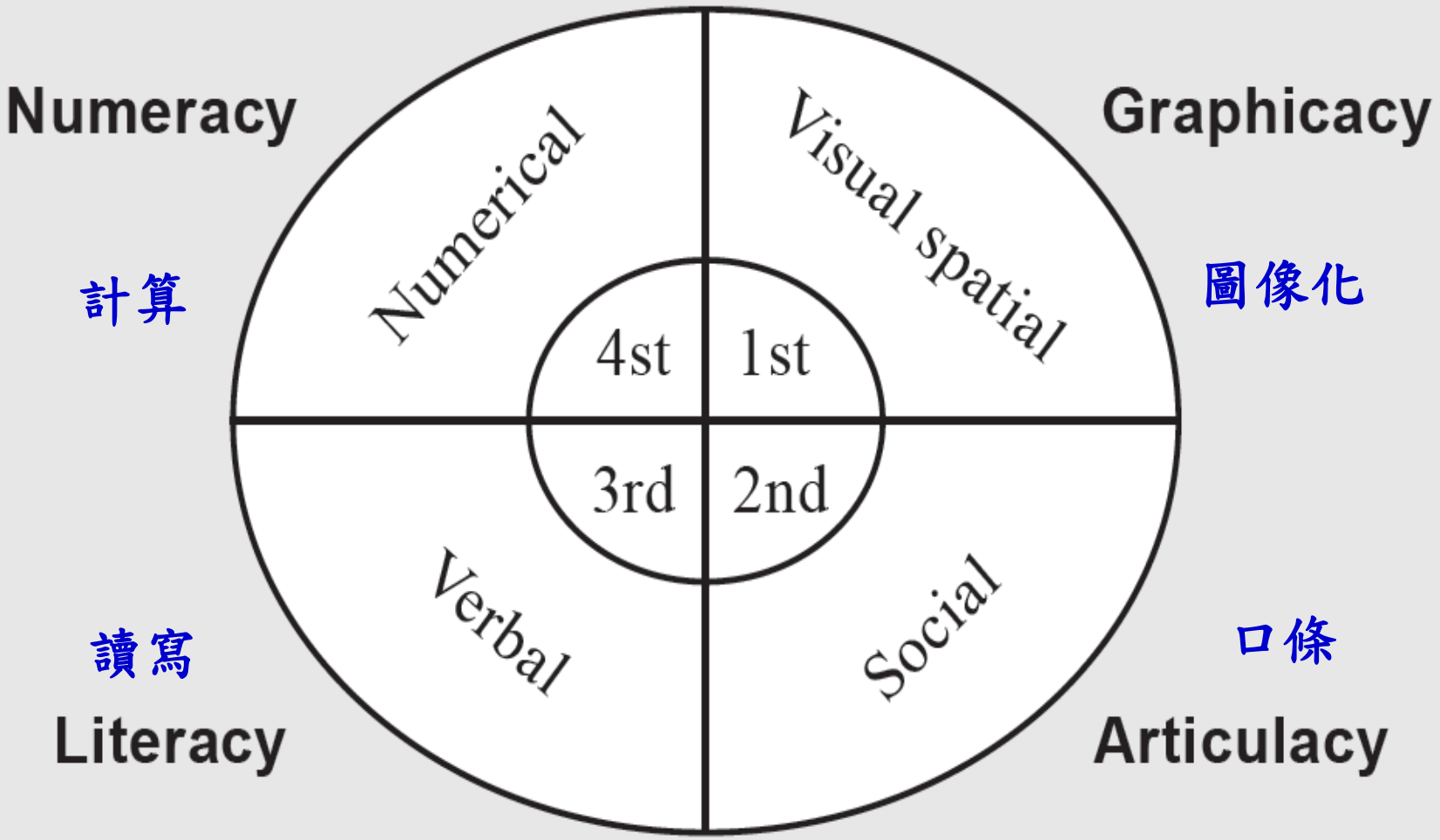
→Articulacy — speaking & listening (people skills)

→Literacy — writing & reading

→Graphicacy — producing & understanding graphics

參考文獻：Balchin WGV (1976). Graphicacy. American Cartographer, 3, 33-38.

統計學家需具備的四大能力



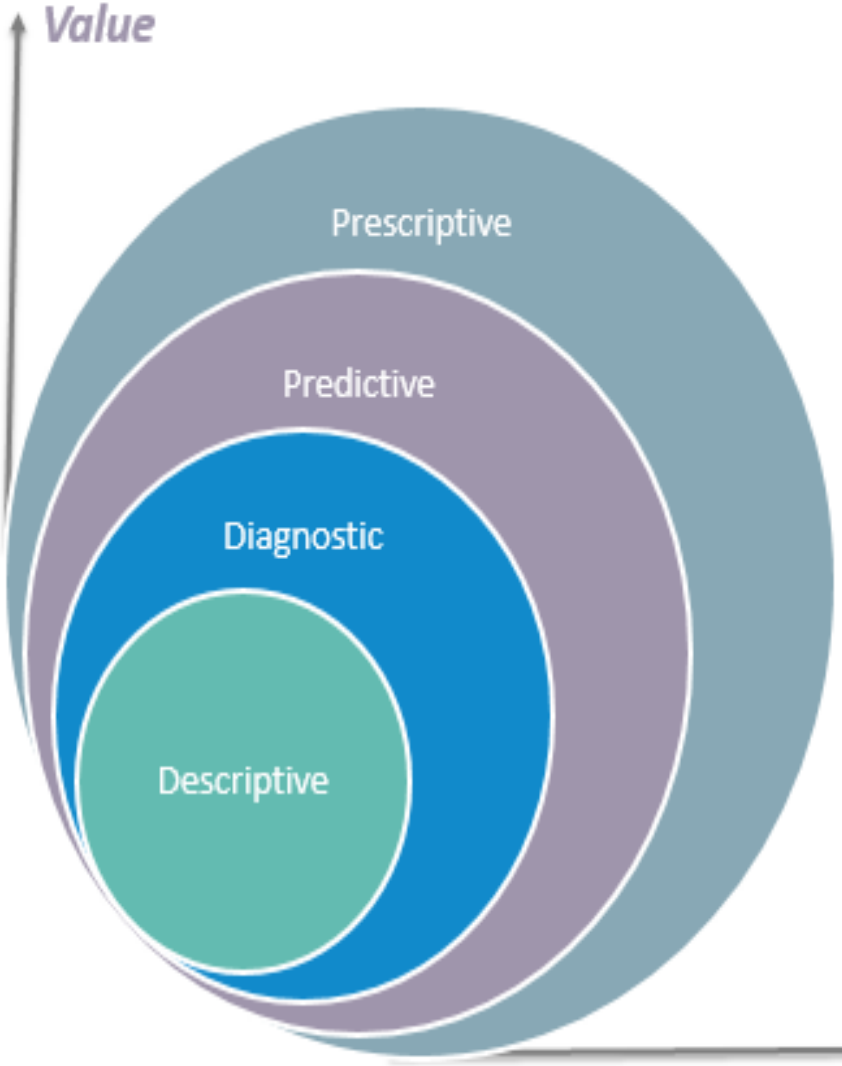
資料科學家的角色

- 資料科學家必須有良好的統計訓練，有助於面對問題時，能夠利用科學方法釐清問題關鍵，並尋求可行及有效的解決策略。
- 若參考「Career Guide for Statistician」的建議，其中列出了十項與統計學家相關的技能，包括(Skills)、知識(Knowledge)、 能力(Abilities)、工作(Tasks)等各個面向。

→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

資料科學家的其他觀點

- 優秀的資料科學家，透過資料分析、說故事，用資訊圖表、聽得懂的人話發揮數據影響力：

→數據獲取能力（程式語言、Python、Java）；

→處理能力（大數據開發技術、Hadoop）；

→分析能力（演算法開發、資料探勘技術）；

→視覺化能力和企業業務領域知識；

→故事敘說能力。

https://medium.com/simple-is-power/%E8%AA%AA%E8%B3%87%E6%96%99%E7%9A%84%E6%95%85%E4%BA%8B-%E8%B3%87%E6%96%99%E7%A7%91%E5%AD%B8%E5%AE%B6%E4%B9%9F%E6%98%AF-%E8%B3%87%E6%96%99%E6%95%85%E4%BA%8B%E5%AE%B6-b05372264c73

https://www.bfa.com.tw/blog/datasotry

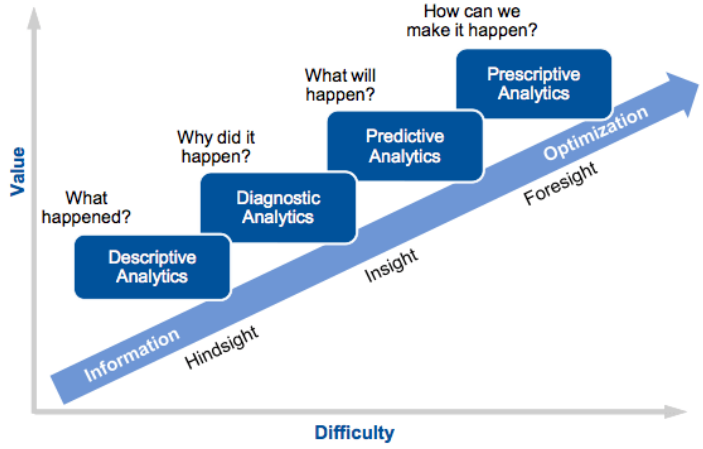
資料科學家的其他觀點(續)

→策略性思維、快速學習與適應能力、問題分析與解決能力、數據的敏銳度及分析能力——遠傳總經理李彬。

→分析資料能力以及定義問題能力是決定數據應用及分析成敗最大的關鍵—台灣大哥大。

→數據分析師的學習重點：專業技能、產業知識、批判性思。

https://tw.alphacamp.co/blog/data-analysis-process



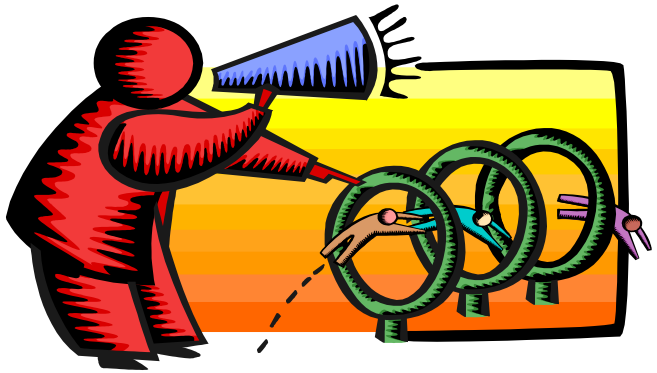
Source: Gartner (March 2012)

如何分析大數據？

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

大數據與學術研究

- 過去統計研究多仰賴理論推導、電腦模擬，大數據讓實證分析的角色更多元，解除資料大小及範圍的限制。
- **資料驅動**（**Data Driven**；讓資料說話）提供另一種角度的思維，有別於由專家意見導引研究方向，藉由基本資料分析篩選出資料的重要特質。



https://rosacobos.wordpress.com/2010/02/21/no-quiero-que-mi-jefe-me-haga-coaching/#more-1071

數據轉化為知識的過程

數 據 (Data)

資 訊 (Information)

事 實 (Fact)

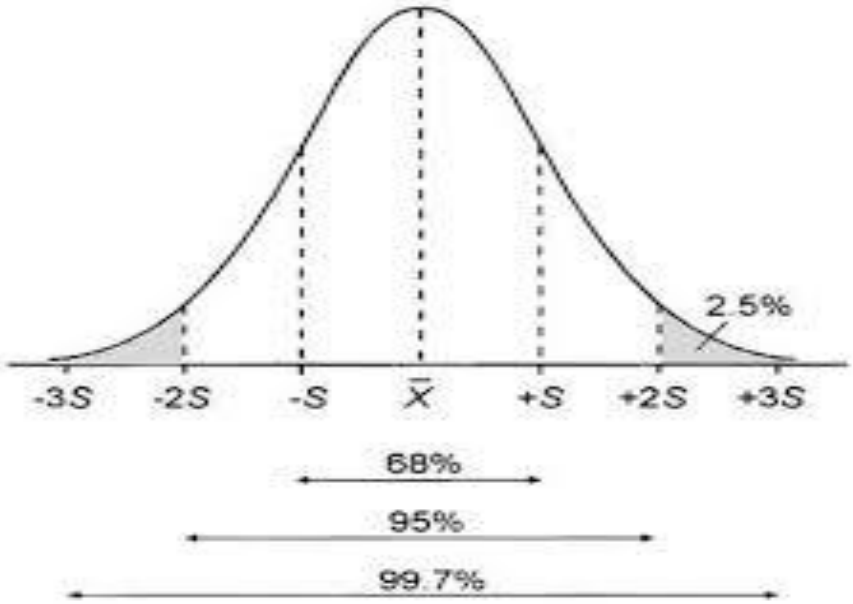
知 識 (Knowledge)

統計與知識

- **統計**（與其他數量分析工具）可視為**歸納法**（馭繁為簡）的一種，協助我們區隔哪些情形為常見、哪些為罕見。

→以統計的角度思考，

p-value測量觀察值的罕見程度，並以此作為檢定的判斷依據。

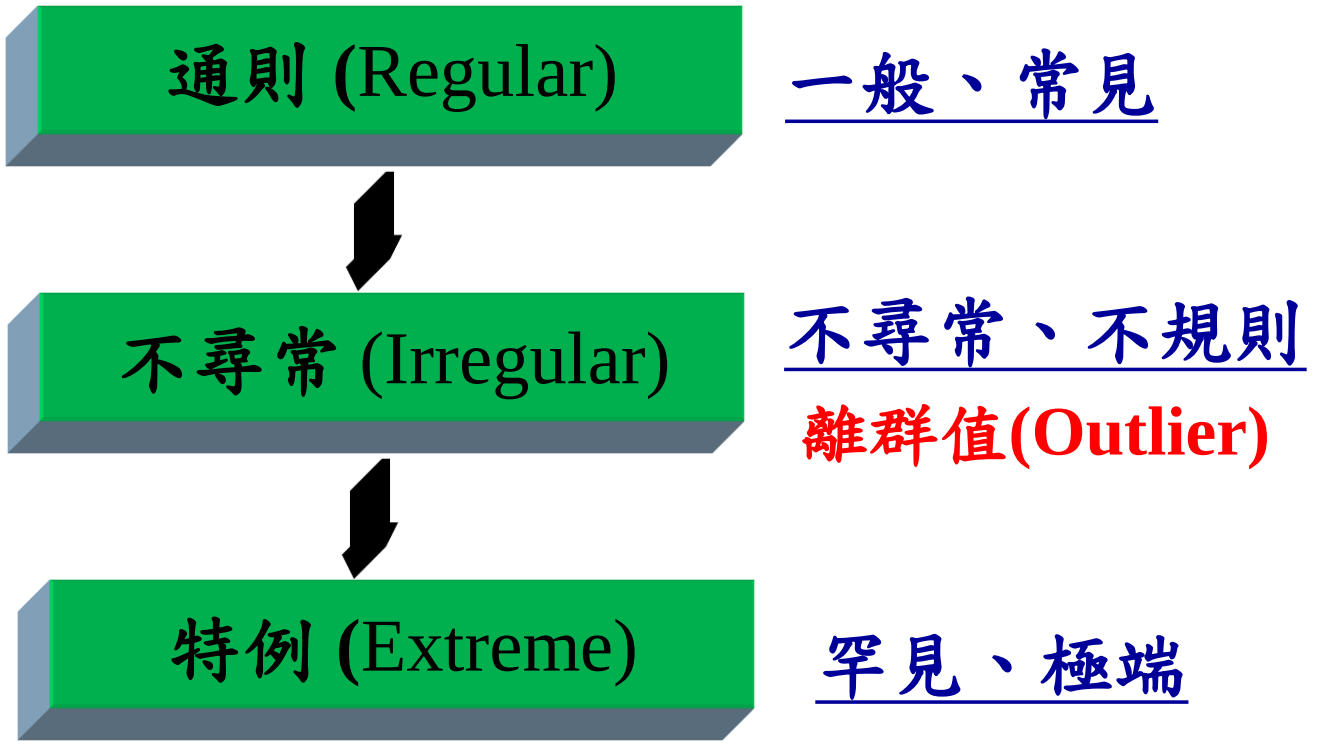


註：R軟體教學影片

https://www.youtube.com/watch?v=fcd6zSk0yd8&feature=youtu.be

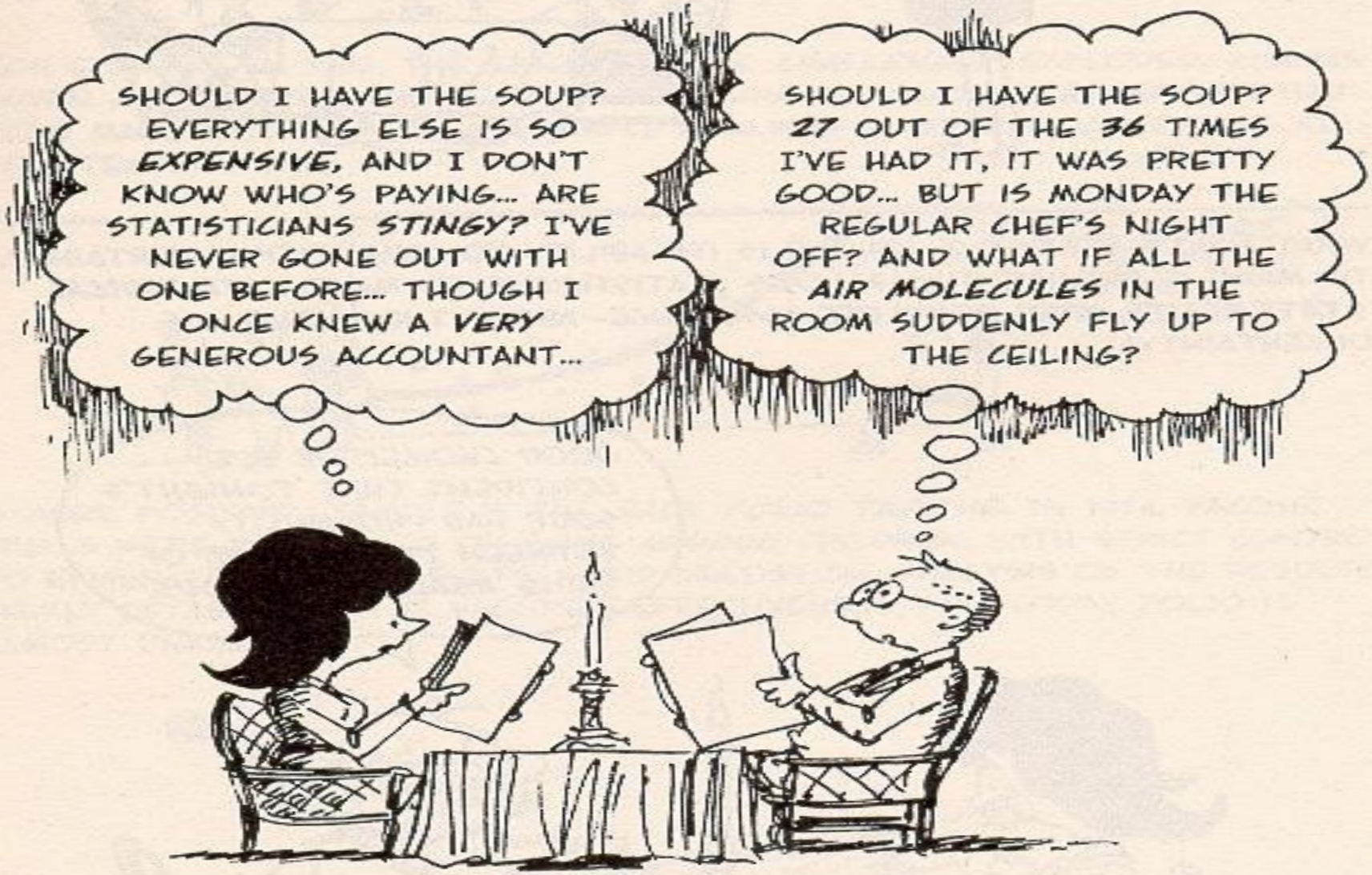
歸納法

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



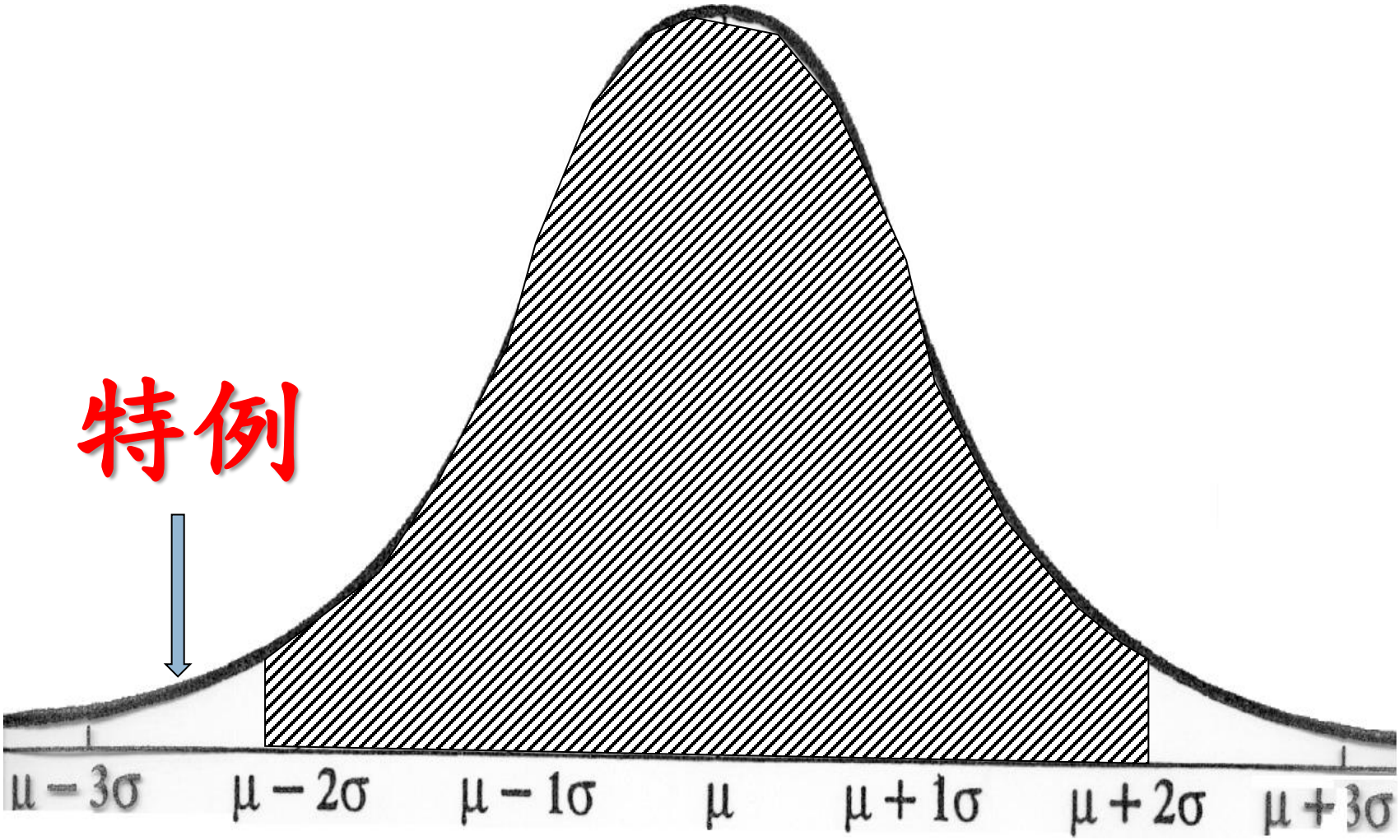
WHAT IS STATISTICS?

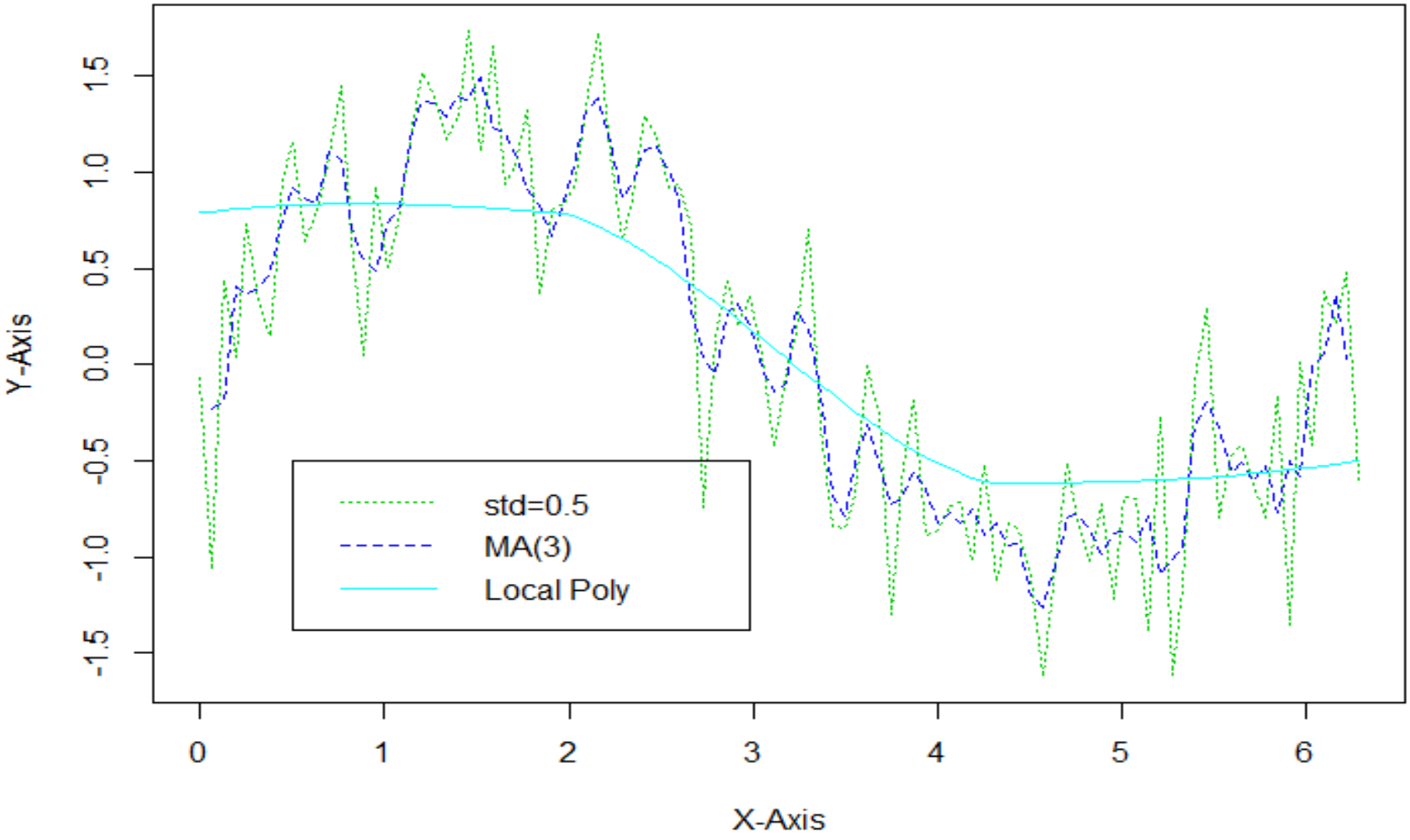
WE MUDDLE THROUGH LIFE MAKING CHOICES BASED ON *INCOMPLETE INFORMATION*...



原則!!

特例





問題：過與不及？如何選擇合適的方法或模型？

試誤法(Trial and Error)

- 因為大數據的龐雜，不易規範資料分析的標準操作程序(SOP)，較為可行的相關性分析或許必須透過試誤法。

→卡古(Kaggle)公司常舉辦資料挖掘比賽，2012年對二手車的相關分析中，發現橘色烤漆的車故障率較低。

→以2008年紐約人紀錄預測2009年危險人孔蓋名單，高達44%正確，其中最重要的指標是「電纜年份」。

定義問題

蒐集資料

分析資料

詮釋結果

大數據分析的特徵

- 分析大數據主要可分為兩部分：

→如何儲存資料？(Storage)

→如何處理資料？(Processing)

- 資料分析軟體（例如：Excel）通常有處理容量的限制，單一軟體無法因應需求，多半得結合數種資訊處理工具。

→1990年代曾處理過房地產資料，其大小相當於半個硬碟的容量。

因果謬誤（Causal Fallacy）

- 因果謬誤（Causal Fallacy）是一種非形式誤謬，泛指各種未有充分證據便輕率斷定因果關係的不當推論。（來源：維基百科）

→A 和 B 之間的關係有五種可能性：

1. A（因）導致 B（果）；
2. B（因）導致 A（果）；
3. A 和 B（因）互相導致對方出現（果）；
4. A 和 B（因）一起導致 C（果）；
5. 觀察的關係純屬偶然（沒有因果關係）。

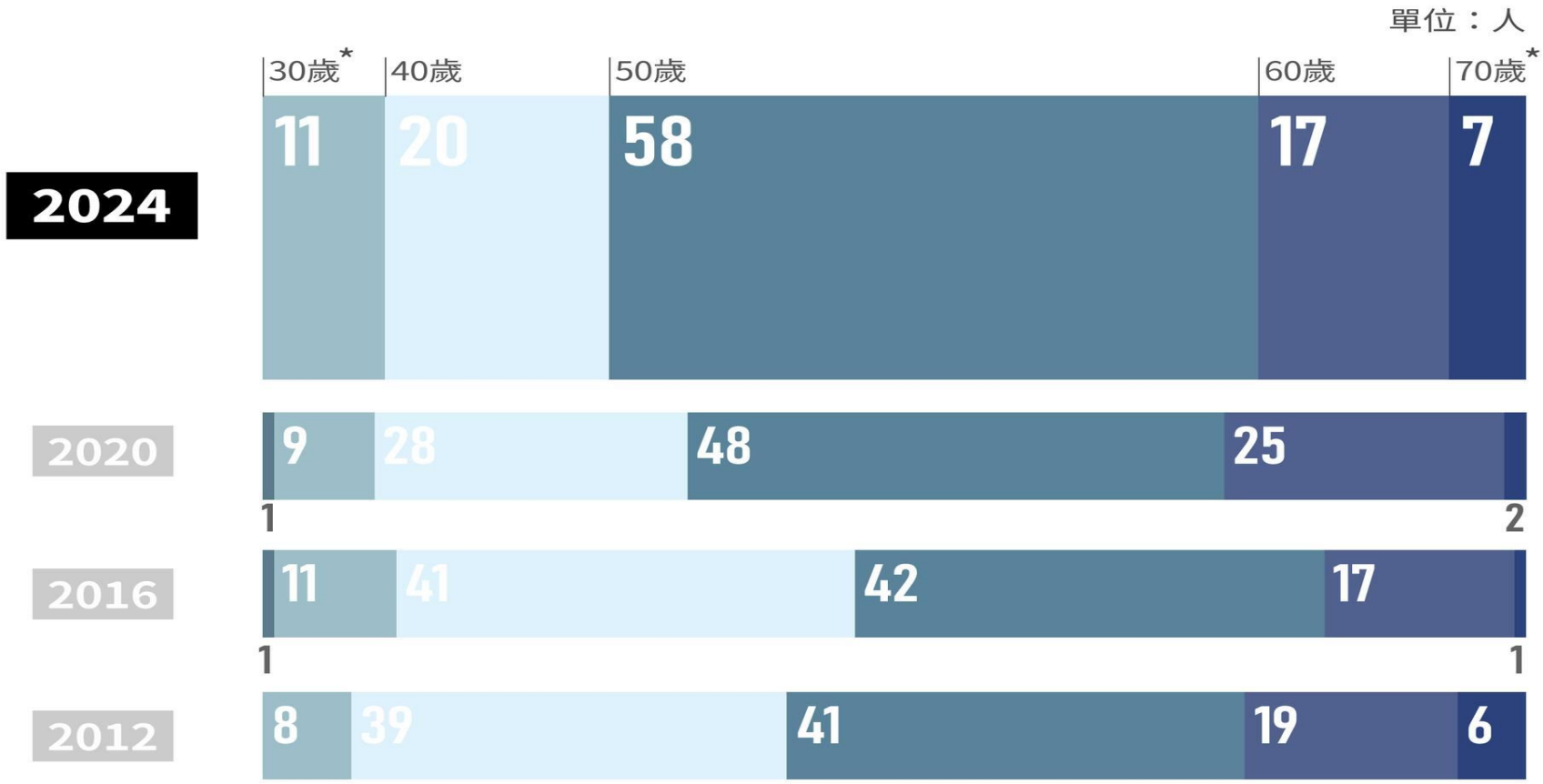
2015年新北市離婚最多的星座組合

排名	星座組合	申請件數量
1	處女座 vs. 天秤座	196
2	獅子座 vs. 摩羯座	164
3	處女座 vs. 射手座	161
4	天秤座 vs. 天蠍座	159
5	雙子座 vs. 處女座	157
6	摩羯座 vs. 水瓶座	156
7	天蠍座 vs. 雙魚座	155
8	巨蟹座 vs. 處女座	153
9	白羊座 vs. 獅子座	153
10	處女座 vs. 摩羯座	153

· 資料統計日期: 2015/01/01~2015/12/31
· 資料來源：新北市政府民政局提供

網路溫度計

第11屆立委年齡分布



*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)

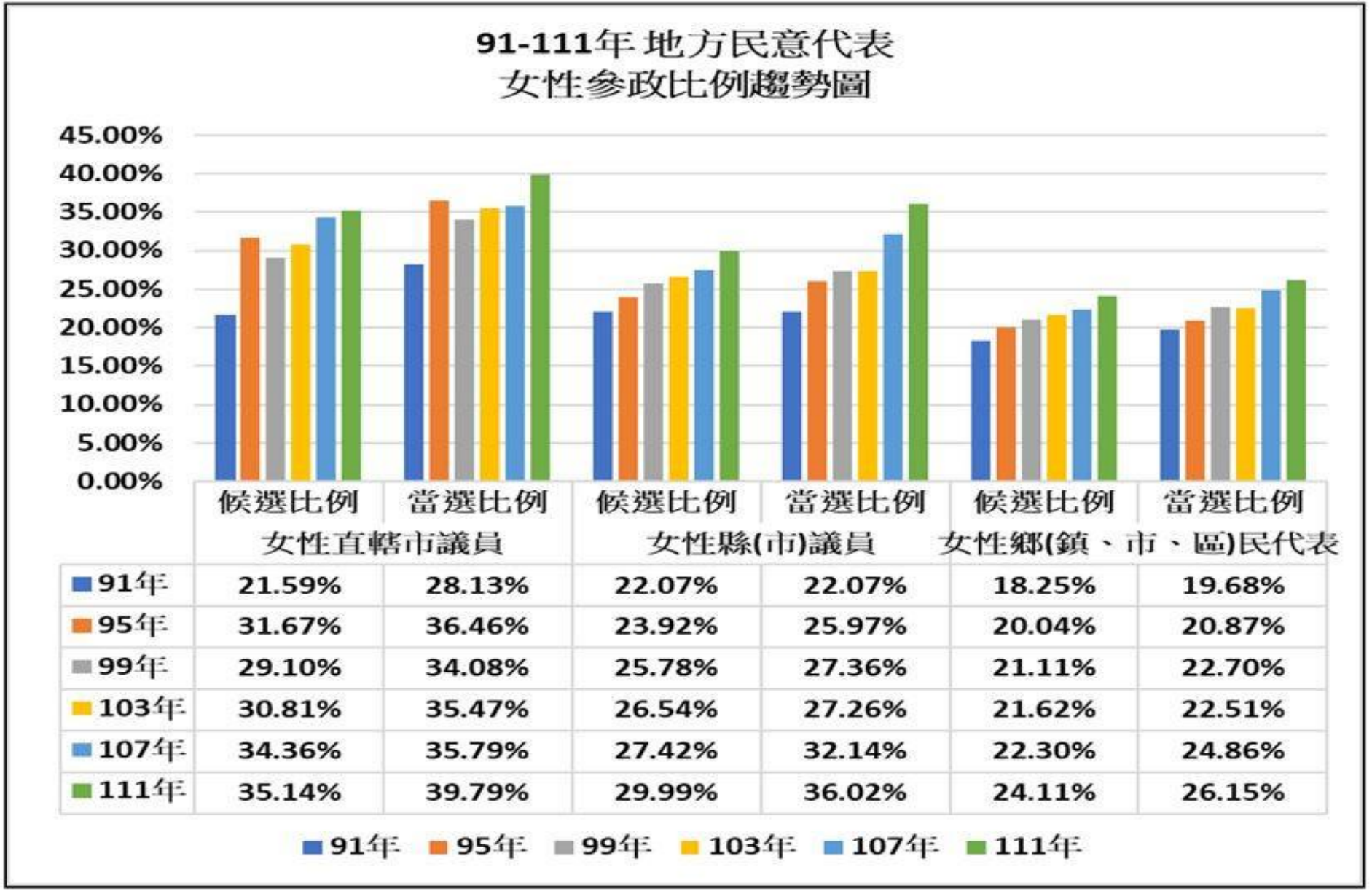
30至39歲 - 黃捷(30)、廖偉翔(33)、牛煦庭(33)、羅廷瑋(34)、徐巧芯(34)
廖先翔(35)、黃健豪(35)、邱若華(35)、吳沛憶(36)、陳冠廷(37)、洪申翰(39)

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_AfD6Gfg9MnO1WINiv4RLSqkS5najZlDyN09IUmy-hZ2VLA&oe=65E036E2

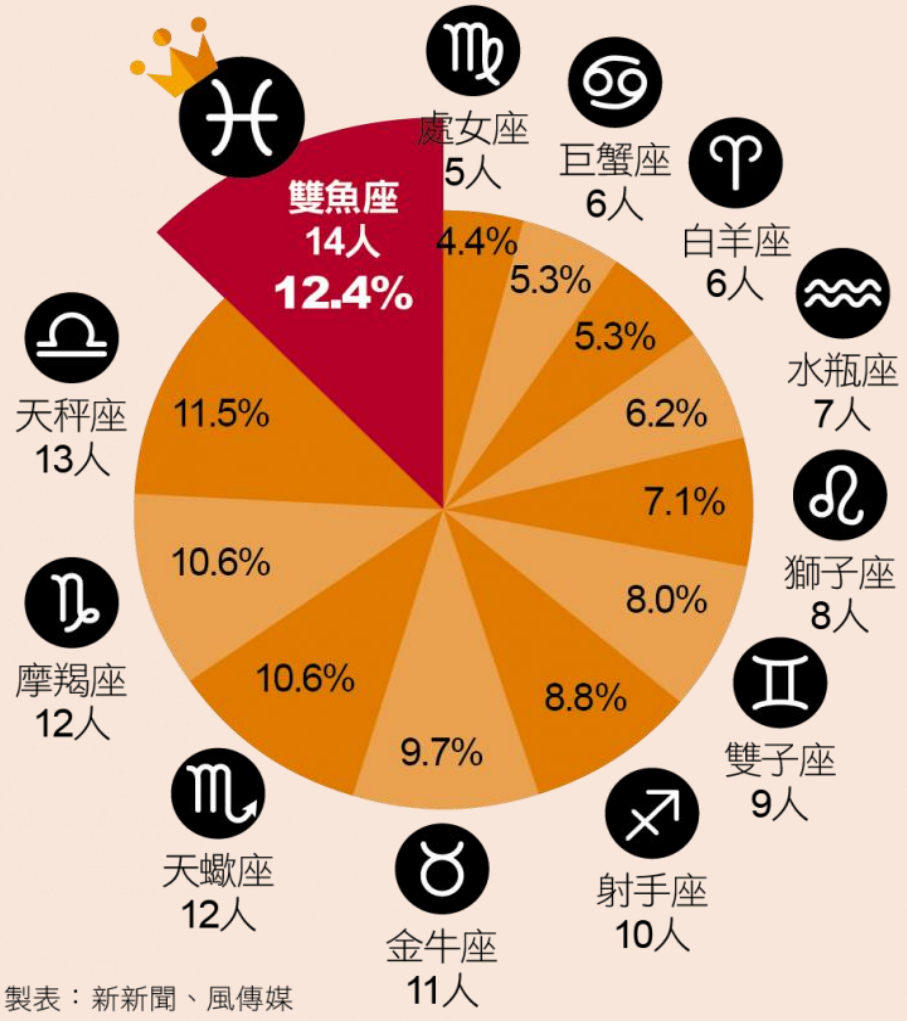
中央通訊社

表 2：近 20 年(六屆) 地方民意代表選舉結果性別統計資料



資料來源：中央選舉委員會選舉資料庫 https://www.taiwanhot.net/news/1025044/%E5%A5%B3%E5%8A%9B%E5%B4%9B%E8%B5%B7+%E5%8F%B0%E7%81%A3%E6%B0%91%E6%84%8F%E4%BB%A3%E8%A1%A8%E5%A5%B3%E6%80%A7%E5%8F%83%E6%94%BF%E6%AF%94%E4%BE%8B%E5%89%B520%E5%B9%B4%E6%96%B0%E9%AB%98

雙魚座最適合當立委？

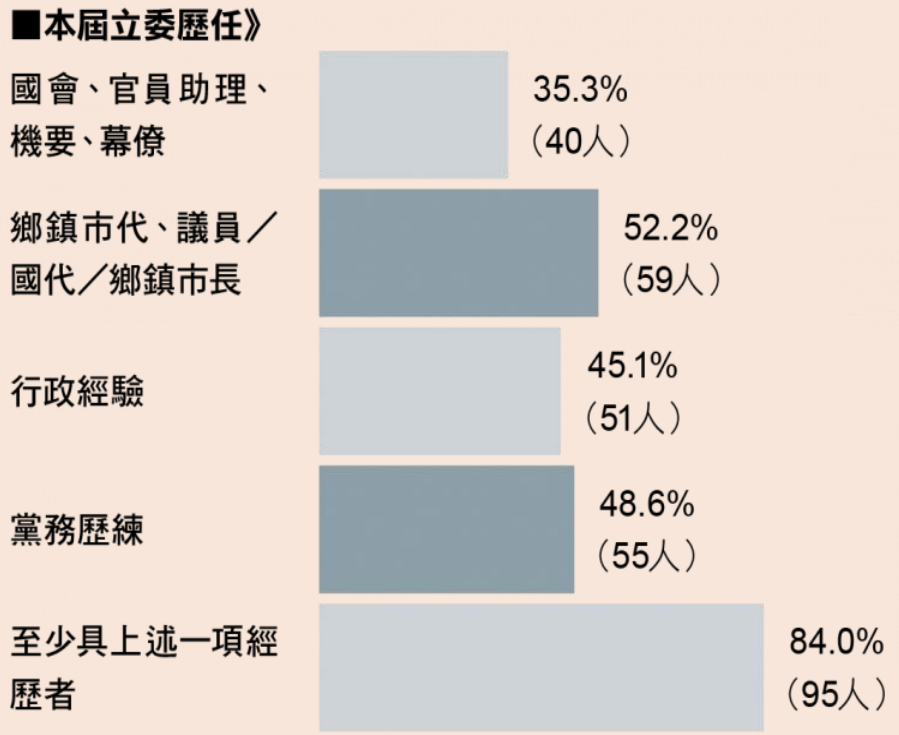


製表：新新聞、風傳媒

https://image.cache.storm.mg/styles/smg-800xauto-er/s3/media/image/2020/03/25/20200325-022724_U8719_M601088_bd60.png?itok=xcecF5tv

- 適合度的定義？
- 分布均勻嗎？(0.4678)

選立委需要具備哪些履歷？



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/2020/03/25/20200325-022724_U8719_M601089_98d1.png?itok=UFe0kvWz

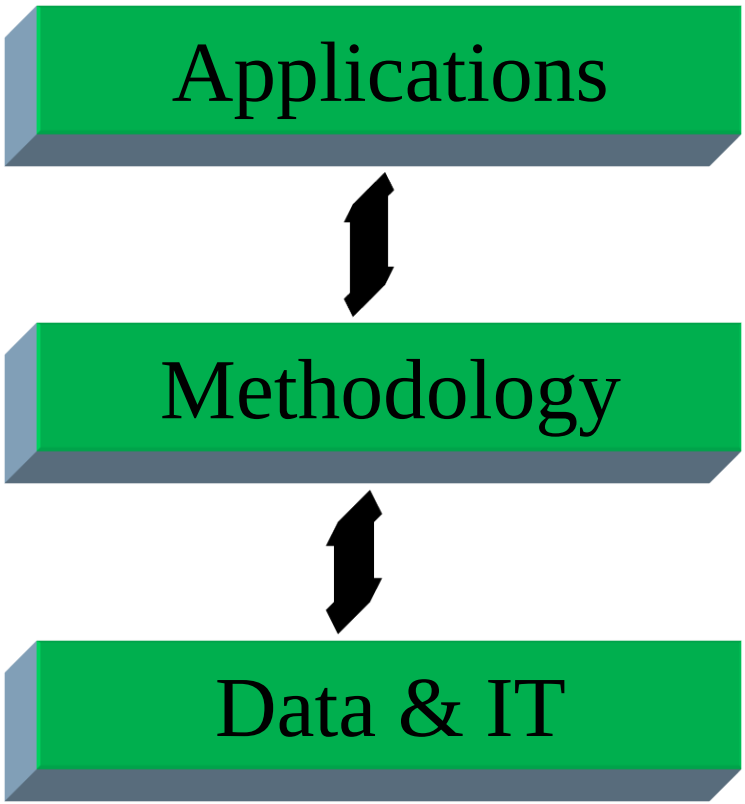
如何定義相似性（亮點）？

Q版畫像的要訣？

音樂也有類似狀況（國樂、日本）

大數據分析為跨領域合作

- 透過數量化分析，篩選出應用領域所需的重要訊息及知識。（三者配合！）



增加大數據的價值

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

發展大數據的潛在優勢

- 教育普及、全球化、電腦科技進步等因素，使得大數據的研究更加可行：

→人類基因密碼重組歷經十年，才完成三十億個鹼基對的定序，現在只需一天；

→教育、資訊普及，外加臉書及推特等工具使得使用者回饋、資訊分享更為可行（Google、「群的智慧」—Smart Swarm、人肉搜索）；

註：正面/負面表列找出特定需求

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美國沃爾瑪賣場(Walmart)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。（臺灣則是泡麵、瓶裝水）

→尿布與啤酒是Walmart另一個知名範例！

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。

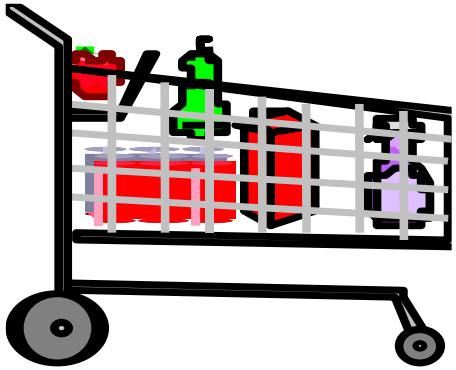
Milk, eggs, sugar, bread

Customer1

Milk, eggs, cereal, bread

Customer2

Eggs, sugar



Customer3

數據分析的噱頭！？

Walmart賣場—尿布與啤酒

- 沃馬特蒐集與分析顧客資料，並以整理所得的資訊，提高銷售業績，成為全球最大的連鎖零售業者
- 美國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？如何將這份資訊轉變為業績？

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，現今已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。

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

「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