「商業資料分析與管理決策」
—資料：競爭優勢

Spring 2023

授課教師：統計系余清祥

2023年3月31日

授課內容：抽樣與問卷

課程下載：csyue.nccu.edu.tw

資料科學家(Data Scientist)

- 統計等同於資料科學(Data Science)嗎？

→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）

- **定義問題**、**溝通能力**！

→**教育**與**資訊**（俾斯麥：**教育**、**交通**）

- 資料科學家需要下列「溝通」能力：

→**與人溝通**：寫作、口語等溝通能力；

→**與資料**（及統計理論）**溝通**：data sense；

→**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**：資訊安全、程式運算。

解決問題(Problem Solving)的流程

定義問題

↓

蒐集資料

↓

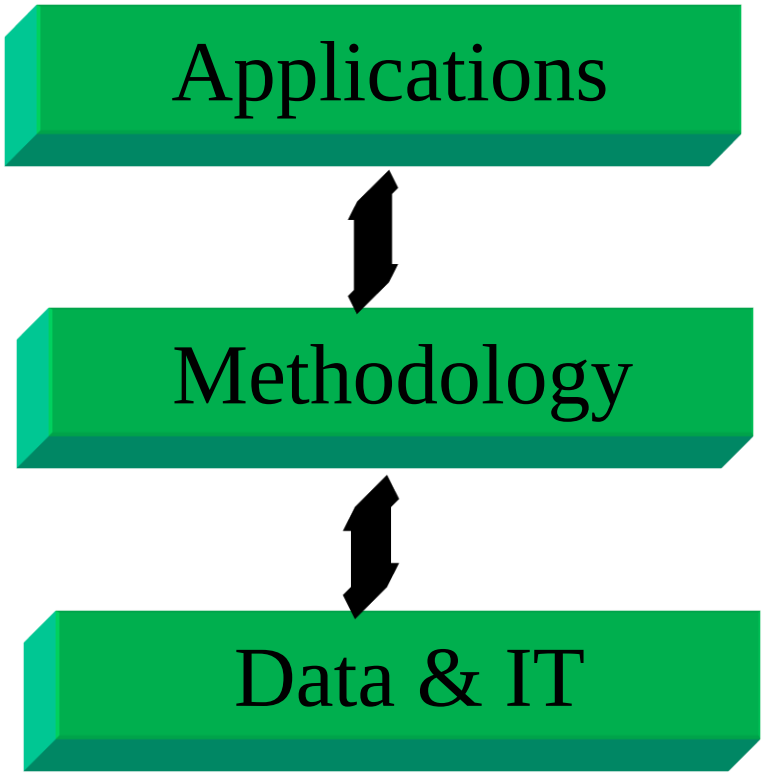
分析資料 ← 絕大多數的教學重心

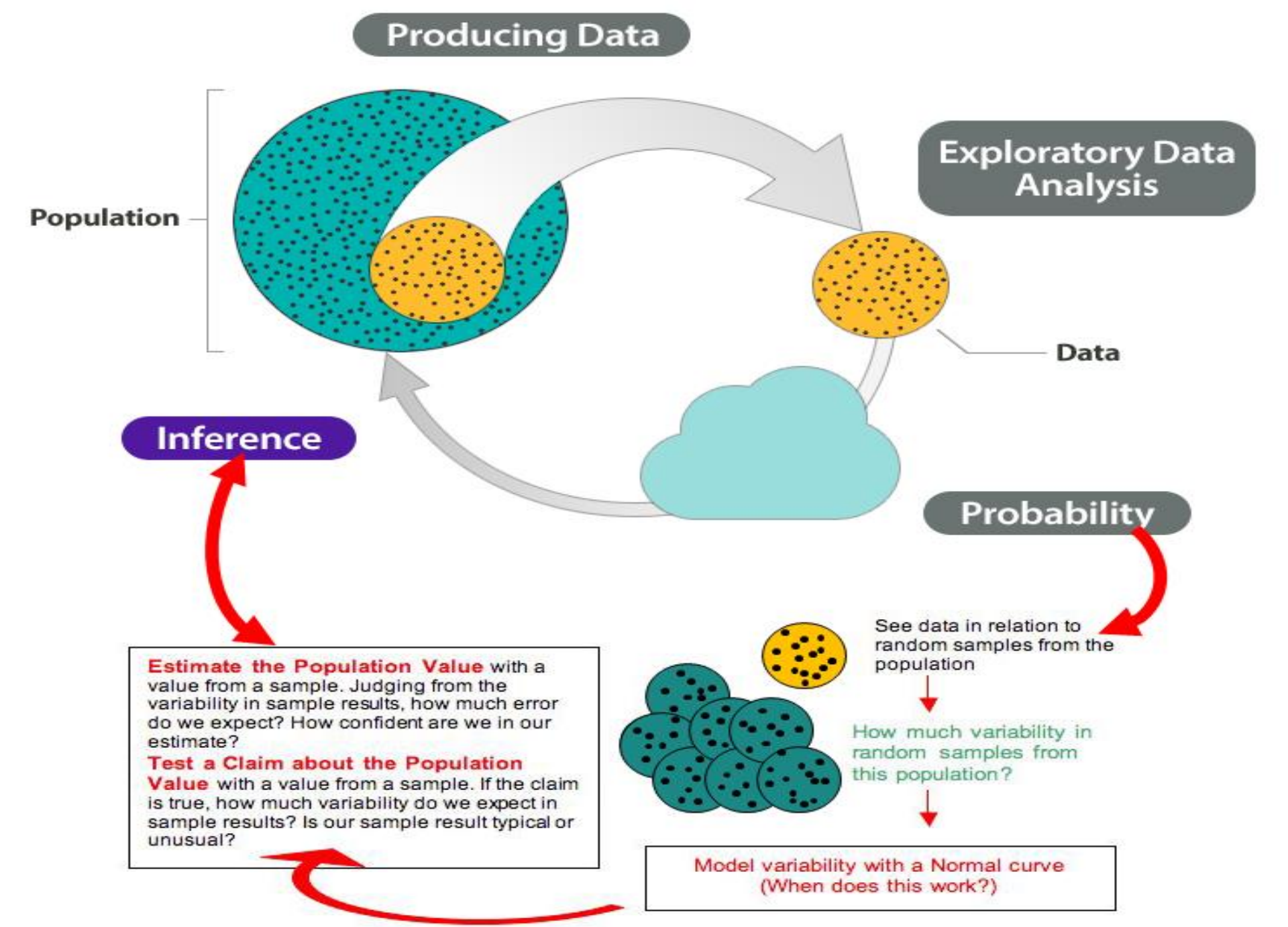
↓

詮釋結果

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！）



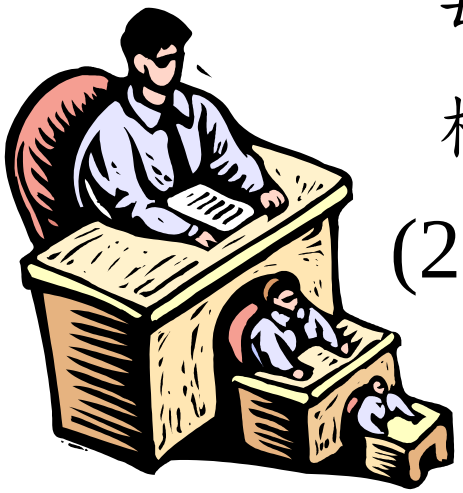


https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/wim-linking-probability-to-statistical-inference/

母體(Population)與樣本(Sample)

- 母體是具有共同特質的個體所組成的群體；樣本是自母體抽出的部分集合。

範例：(1)北市高中學生戶籍在北市的比例



母體→全臺北市的高中學生

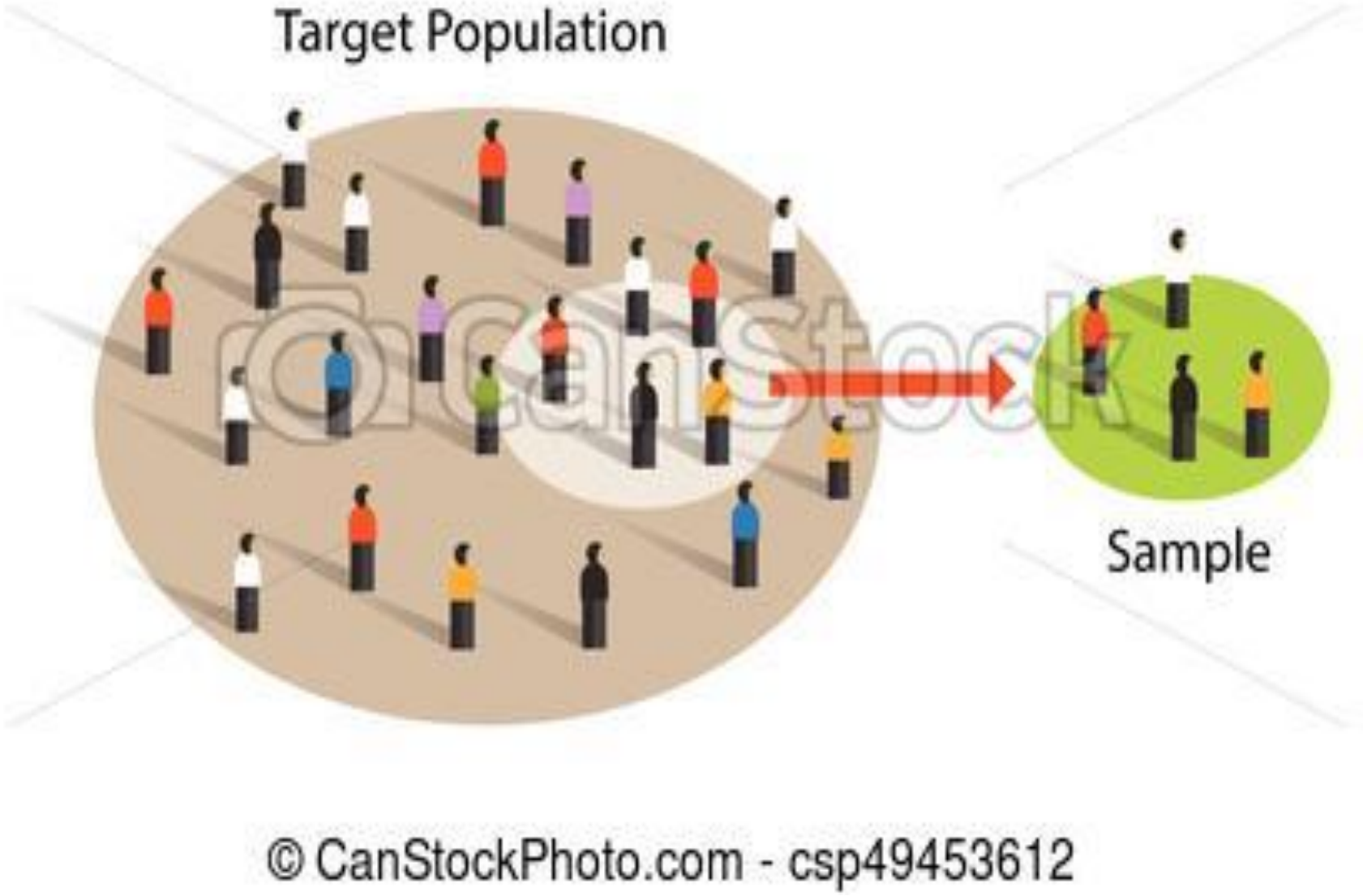
樣本→參加政大課程的高中學生

(2)總統候選人得票率（電話訪問）

母體→全臺灣的合格選民

樣本→被訪談的臺灣居民

統計分析：樣本→母體



https://comps.canstockphoto.com/sample-from-population-statistics-vector-clip-art_csp49453612.jpg

母體參數與估計

- 母體特性通常以參數（Parameter；母數）稱呼，像是成年人的身高、平均壽命、考試通過率。

→一般無法取得母體的資料，只能透過部分成員（亦即樣本）反推全體，統計因此也稱為反向推論(Inverse Inference)。

→但樣本是否能反映全體、分析方法是否適當都是關鍵！

為什麼要抽樣？

- 為什麼只看一部份的母體？

→普查(Census)：逐一檢查母體的所有個體。例如：戶口普查、工商業普查。

→普查需要較長的時間、較多的經費與人力，往往只有政府負擔得起。(政府也是每十年普查一次，其他時間輔以問卷調查、公務統計等等彌補資料的不足。)

→有時抽樣是唯一可行的方法。

常見抽樣的範例

https://orion-pirotehnika.com/wp-content/uploads/2021/08/Fireworks.jpeg

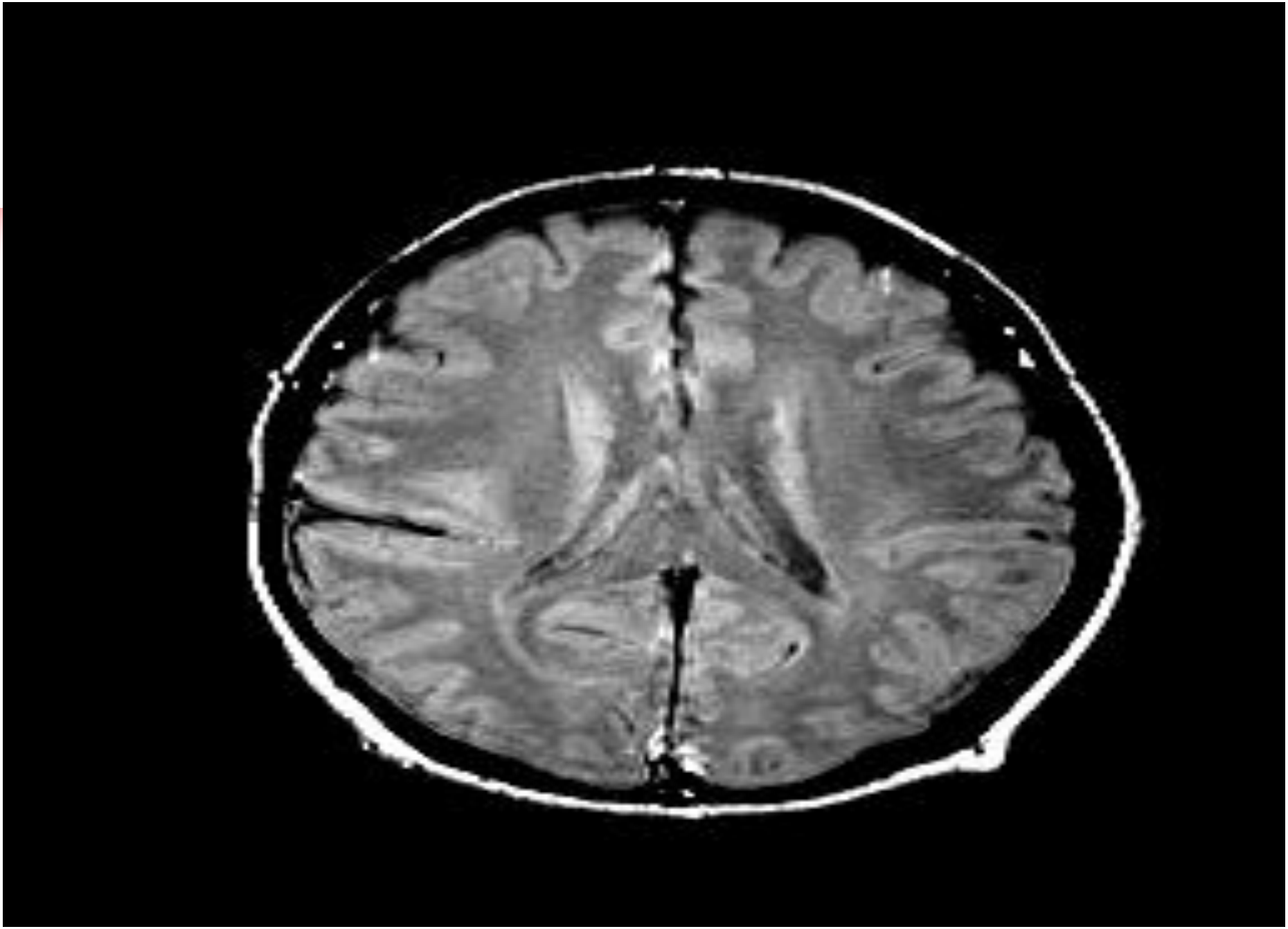
- 品質管制(Quality Control)

→毀滅性抽樣(如鞭炮、罐頭等等產品)

- 健康檢查時經常會抽血、驗尿、或萃取某個身體組織附近的樣本，再從檢體中判定是否罹患某種疾病。

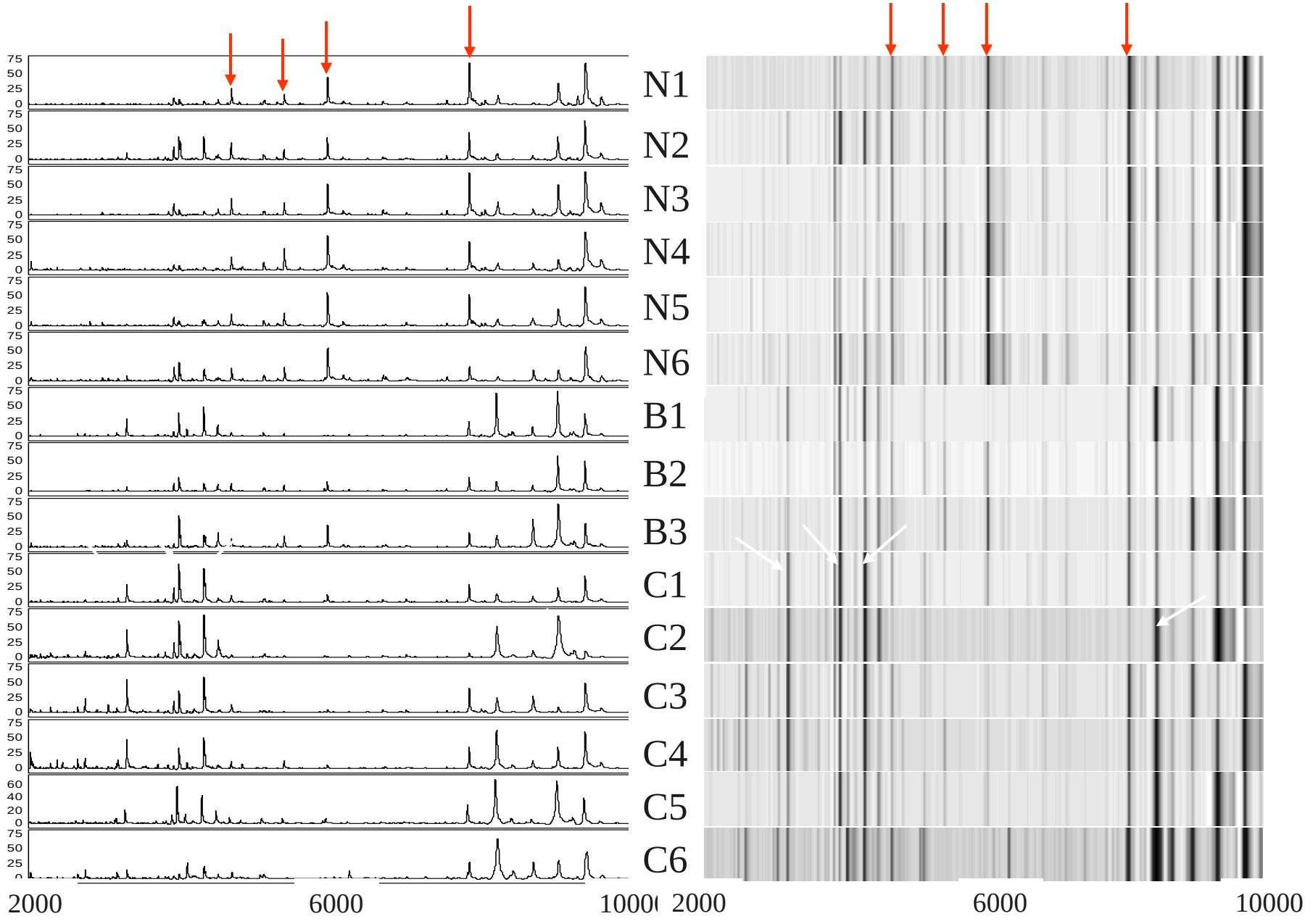
→檢查的結果以圖像表示較為清楚

註：下圖為中風病人的腦部檢查。



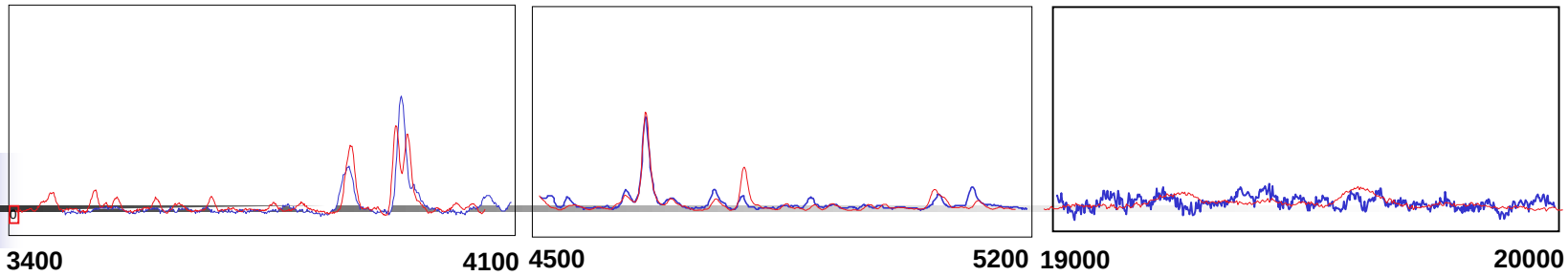
Anatomical Proton-density-weighted Image of Human Brain

SELDI Serum Protein Profile Analysis-Prostate Cancer

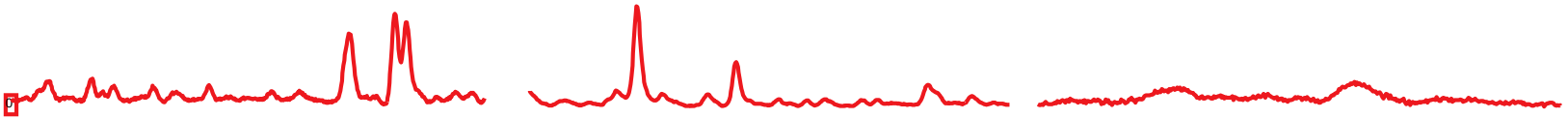


Proteomic Pattern of Sera from Patient

Stage 1 Profile

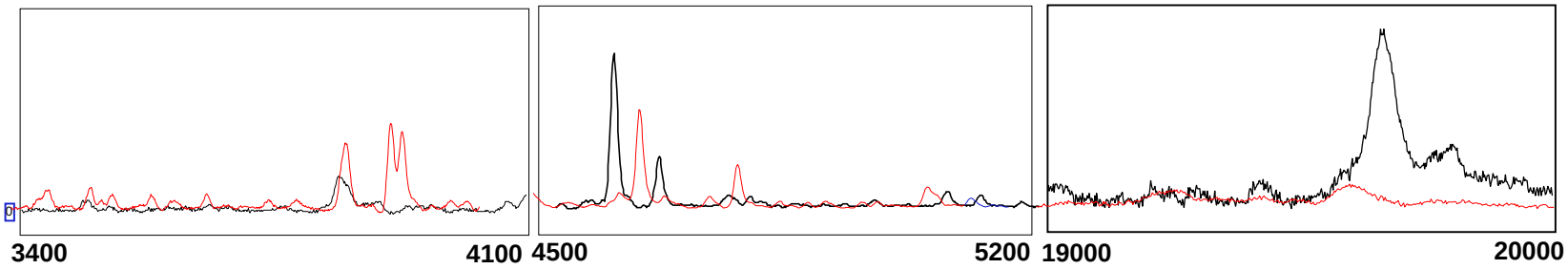


Sample



not a S2 pattern

Stage 2 Profile



協助我們更理性的判斷

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in 99% of the cases in which the disease is actually present, and a correct negative result in 98% of the cases in which the disease is not present. Furthermore, .001 of all people have this cancer.

P(cancer) =	.001	P(~ cancer) =	.999
P(+ | cancer) =	.99	P(– | cancer) =	.01
P(+ | ~cancer) =	.02	P(– | ~cancer) =	.98

$$P(\text{cancer} \mid +) = \frac{P(+ \mid cancer)\,P(cancer)}{P(+)} = \mathbf{.047}$$

計算細節：

- 假設某地區有一百萬人：

  → 999,000人健康，1,000人罹患癌症

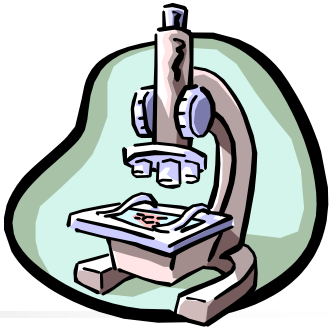
  → 檢查出陽性反應者：

  (1) 健康者中有 $999{,}000 \times 2\% = 19{,}980$

  (2) 癌症患者中有 $1{,}000 \times 99\% = 990$

因此，陽性反應者中罹患癌症的比例：

$$P(cancer \mid +) = \frac{990}{19{,}980 + 990} = \frac{990}{20{,}970} \cong 4.72\%$$



Suppose a second test for the same patient returns a positive result as well. What are the posterior probabilities for cancer?

P(cancer) =	.001	P(~cancer) =	.999
P(+ | cancer) =	.99	P(– | cancer) =	.01
P(+ | ~cancer) =	.02	P(– | ~cancer) =	.98

$$P(\text{cancer} \mid +_1 +_2) = \frac{P(+_1 +_2 \mid cancer)\, P(cancer)}{P(+_1 +_2)} = \mathbf{.710}$$

資料蒐集的方式

- 一般將資料蒐集分類成：

1. 實驗設計(Experimental Design)

→ 包括臨床試驗 (Clinical Trials)，需要較精密計畫，一般分成實驗、對照組，較適合用於推論因果關係的研究。

2. 抽樣調查(Sampling Survey)

→ 設計問卷，藉由調查取得資訊。

- 目標：藉由蒐集的資料推得訊息。

- 另一種常見的資料來源分類，是依據資料產生分成：

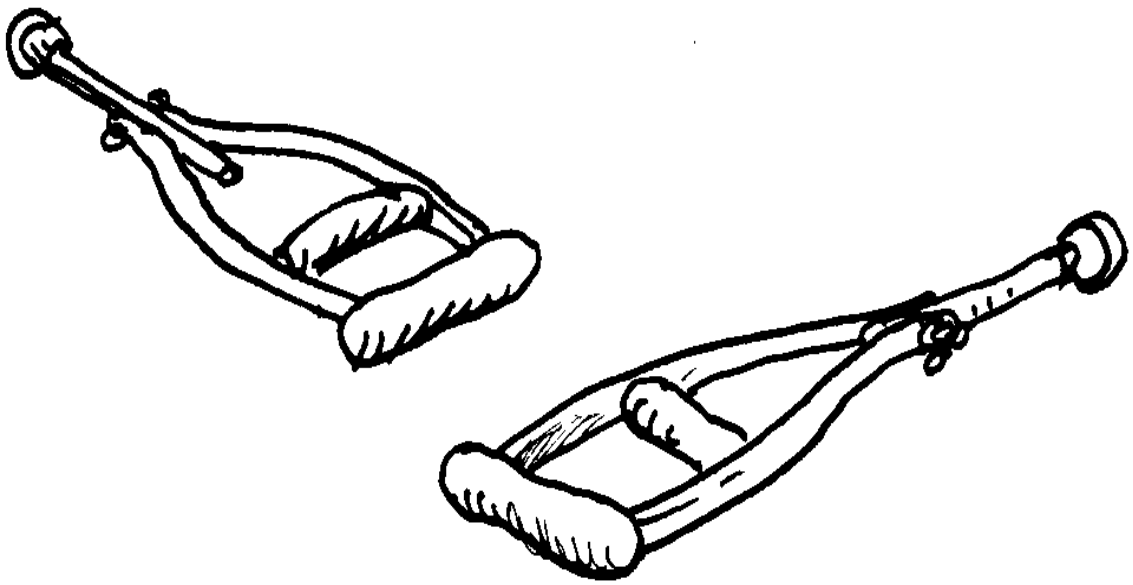
  1. 實驗設計(Experimental Design)
  2. 觀察研究(Observational Study)

→兩者的差異在於資料蒐集者的參與，蒐集資料並不影響觀察研究，像是研究股市、利率、房地產價格，與實驗設計控制變因獲得觀察值不同。

註：實驗設計較為費時費力，而且需要更為縝密的規劃設計，但投資報酬、附加價值通常也大許多。

世界規模最大的醫學實驗(沙克疫苗)

A MORE POSITIVE EXAMPLE IS THE *SALK POLIO VACCINE*. IN 1954, VACCINE TRIALS WERE PERFORMED ON SOME 400,000 CHILDREN, WITH STRICT CONTROLS TO ELIMINATE BIASED RESULTS. GOOD STATISTICAL ANALYSIS OF THE RESULTS FIRMLY ESTABLISHED THE VACCINE'S EFFECTIVENESS, AND TODAY POLIO IS ALMOST UNKNOWN.



- 若以時間來區分，資料可分成：

  1. 縱向資料(Longitudinal Data)
  2. 橫向資料(Cross-sectional Data)

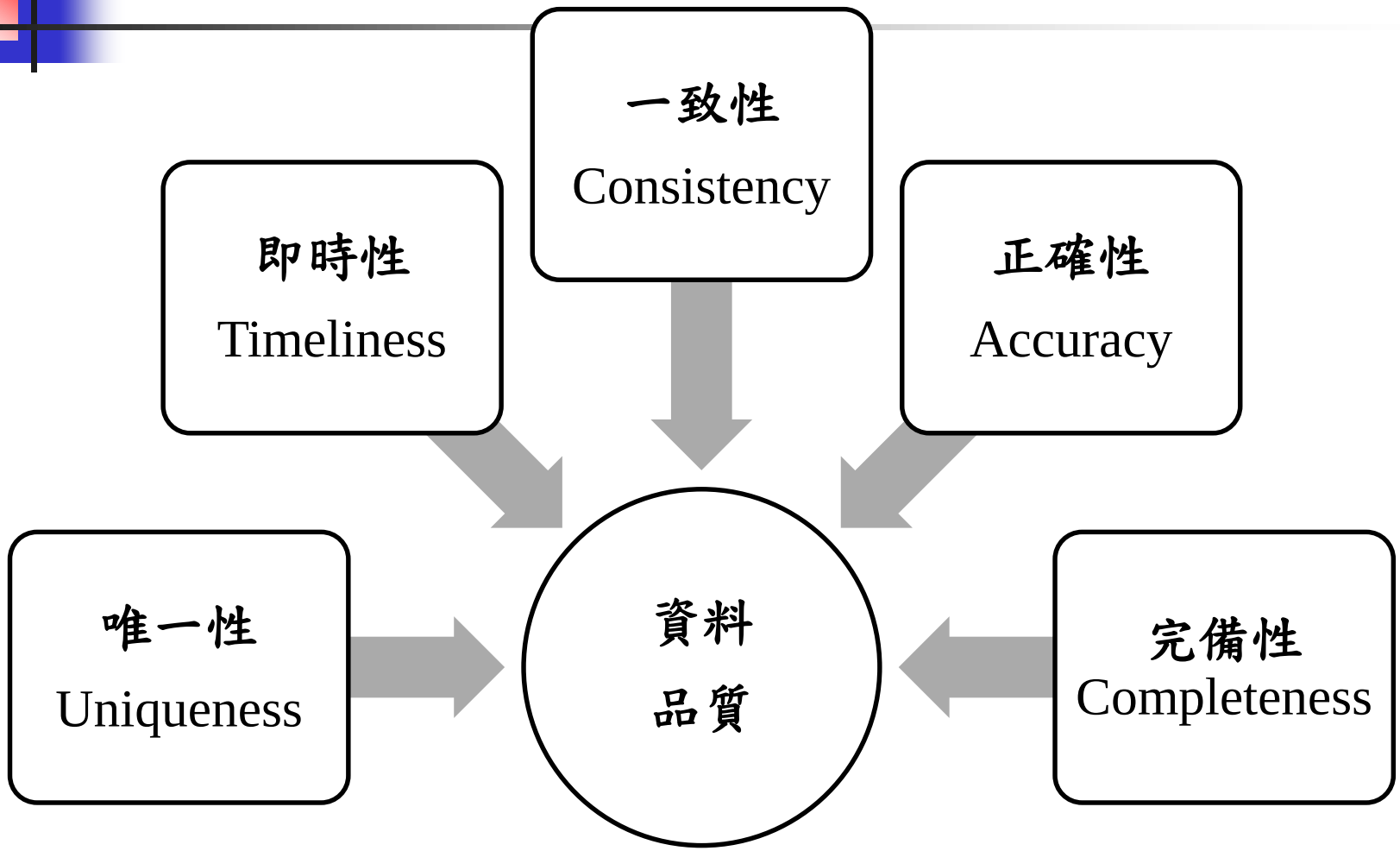
→縱向資料又稱為長期追蹤(Panel)資料，研究相同個體在各時間的變動趨勢，也稱為世代(Cohort)資料。橫向資料則研究某個時間點的母體，但不同時間點的資料未必可互相比較。

註：國內外較知名的縱向資料包括「華人家庭動態研究」與PSID(Panel Study of Income Dynamics)。

資料品質

http://blog.backupify.com/2013/04/01/the-5-but-really-the-6-mistakes-made-in-big-data/00009-08_sept253-jpg/

資料品質的傳統定義



來源：Guide for Checking the Quality of Syndromic Data, International Society for Disease Surveillance

抽樣與資料品質

- 除資料品質外，**選取適合資料（包括抽樣）**也要謹慎的考慮。

→統計是由觀察值（或現象）反推出發生原因，如何選取樣本非常重要，而足夠觀察值可看出母體原貌（三人成虎）。

→ 但為了避免「瞎子摸象」及「以偏概全」的問題，檢查***樣本代表性***是資料分析時必須考慮的步驟。

資料品質！（錯覺或是瑕疵？）

https://img.ltn.com.tw/Upload/news/300/2017/12/13/phpZWbIuX.jpg

https://i2.wp.com/cdn.shortpixel.ai/client/to_webp,q_glossy,ret_img,w_1000,h_400/https://www.dataquest.io/wp-content/uploads/2019/08/garbage-in-garbage-out.jpg?zoom=2.5&w=450&ssl=1

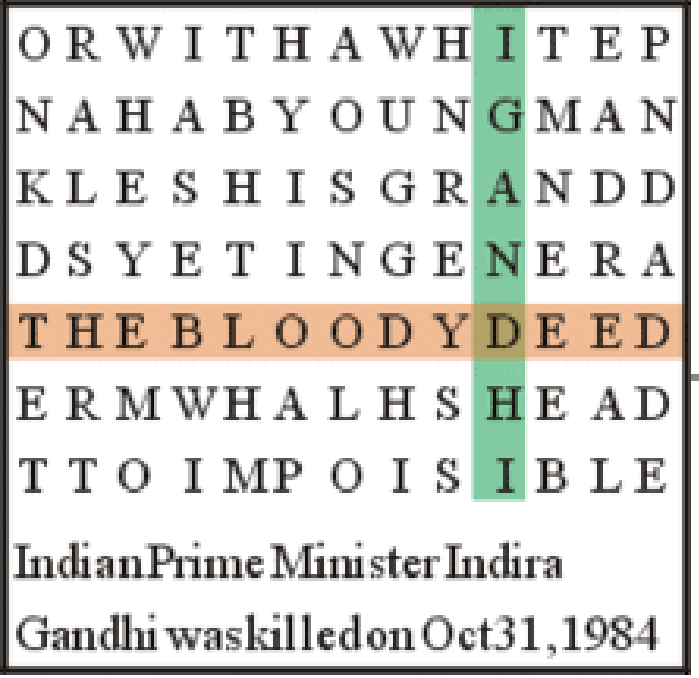
https://i1.kknews.cc/UUCVcttsu9IA2ikKzpks8sQe8zhiDd5psO6Z4XqRPwo/0.jpg

資料品質(Data Quality)

→Garbage in, garbage out!

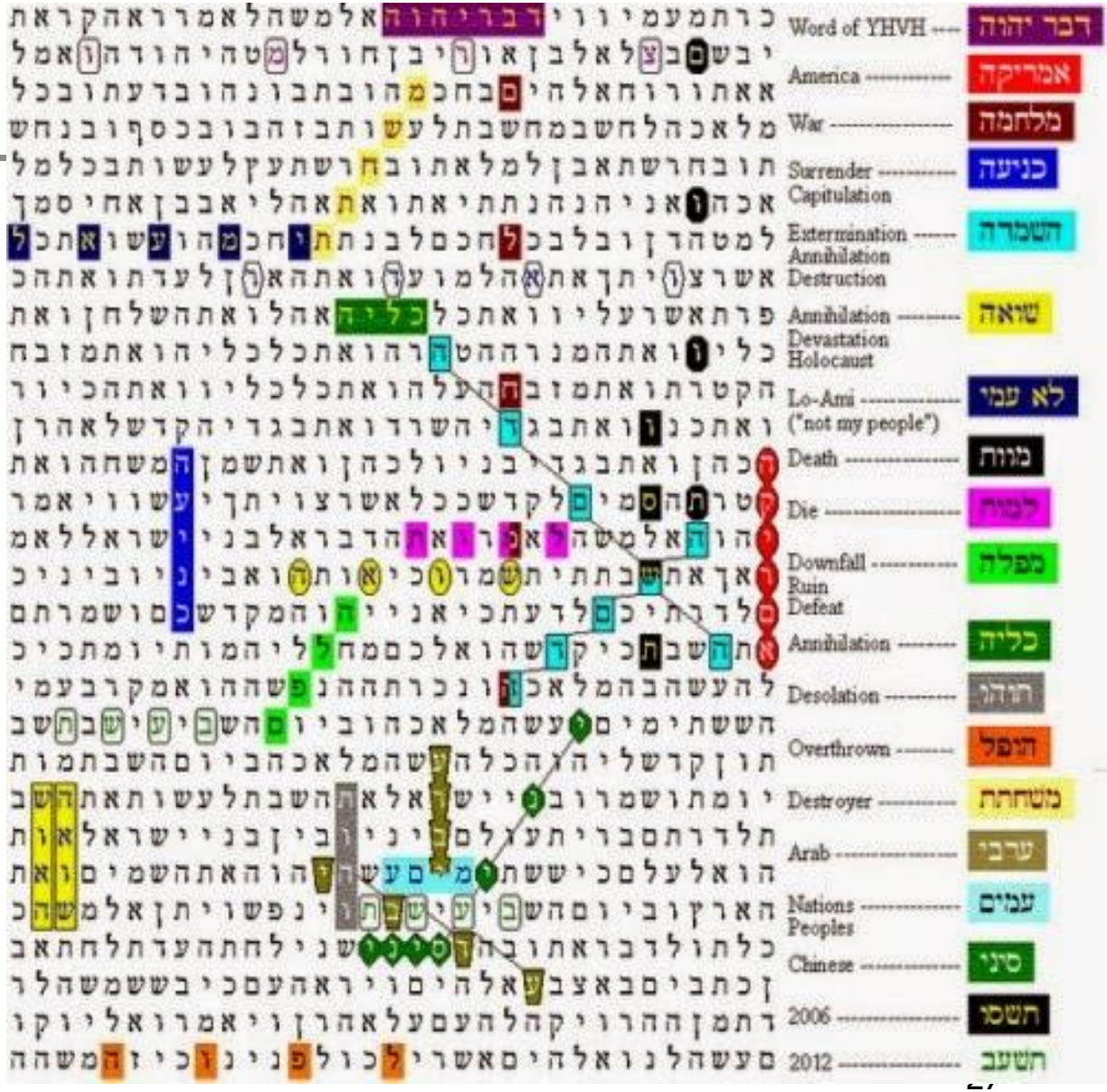
https://pbs.twimg.com/media/CCfGxAqW0AEDr8-.jpg

The Bible Code



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

聖經密碼（Torah密碼）



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

博恩夜夜秀 第三季第七集

行前特別聲明

1. 本錄影為商業售票演出，敬請各位媒體朋友尊重智慧產權，切勿觸法。
2. 官方高畫質原音媒體素材免費商業授權請洽：press@strnetwork.cc
3. 在不妨礙現場觀賞體驗與錄影作業的前提下，歡迎觀眾自備應援道具。
4. 韓國瑜先生及團隊約定19:30抵達攝影棚，橋段20:15開始，敬請期待。
5. 假使內容因來賓行程有變，本節目無需亦無法負責，造成不便請見諒。

備註： 活動過程中，發生人員或道具影響錄影進行之情事，為維護購票觀眾之權益，請遵照維安人員指示離場。

敬請各界朋友拿出民主法治國家的素養
讓博恩與韓市長為大家帶來難忘的夜晚
期待與大家見面

製作人 Hauer

記得要準時

16:45 開始驗票
17:15 開放入場
18:00 正式開始

遲到需等待中場休息才能入場

千萬要帶錢

現場觀眾限定的扭蛋
記得攜帶足夠現金
可使用Linepay

扭蛋四款賣完就沒了

看秀五支箭

進場掃描ＱＲcode，結束會發大合照
請勿於官方上傳前爆雷
請依照現場導播指示
棚內全程禁止飲食拍照錄影錄音
將手機調整為靜音或振動

如何解讀資訊？（推背圖、燒餅歌）

北冥有魚，其名為鯤，鯤之身長過一八○，其重越百斤，一日，化而為巨鳥，其名為鵬，鵬之俊俏，堪稱鳥中之美女，鵬之大，無可比擬，故名之曰：鳥王。鳥王生活無目的可言，故常作無聊狀，天帝見其狀，甚怒，命其改過自新，並予一卵，令鵬孵之，十年歲月，孵出一人，其貌似書呆子，鳥王甚愛之，命名為莊子。某日，莊子為惡獸所包圍，莊子面不改色，引吭高歌，狀甚瀟灑，惡獸懼，皆沒入林中逃遁，莊子大笑，返以告鵬，鵬弗信，以為此乃水中撈月之事，絕無僅有，遂怒逐莊子，莊子甚感悲哀，準備遁世，後受仙人感召，出山林，作逍遙遊以告世人。

執編：李文堯 彭日榮 郭李同

逍遙之音

逍遙遊發刊詞

北冥有魚，其名爲鯤，鯤之身長過一八〇，其重越百斤，一日，化而爲巨鳥，其名爲鵬，鵬之俊俏，堪稱鳥中之美女，鵬之大，無可比擬，故名之曰：鳥王。鳥王生活無目的可言，故常作無聊狀，天帝見其狀，甚怒，命其改過自新，並予一卵，令鵬孵之，十年歲月，孵出一人，其貌似書呆子，鳥王甚愛之，命名爲莊子。某日，莊子爲惡獸所包圍，莊子面不改色，引吭高歌，狀甚瀟灑，惡獸懼，皆沒入林中逃遁，莊子大笑，返以告鵬，鵬弗信，以爲此乃水中撈月之事，絕無僅有，遂怒逐莊子，莊子甚感悲哀，準備遁世，後受仙人感召，出山林，作逍遙遊以告世人。

執編：李文堯
彭日燊
郭李同

藏頭詩「北一女的新書包沒水準」

北冥有魚，其名為鯤，鯤之身長過一八○，其重越百斤，一日，化而為巨鳥，其名為鵬，鵬之俊俏，堪稱鳥中之美女，鵬之大，無可比擬，故名之曰：鳥王。鳥王生活無目的可言，故常作無聊狀，天帝見其狀，甚怒，命其改過自新，並予一卵，令鵬孵之，十年歲月，孵出一人，其貌似書呆子，鳥王甚愛之，命名為莊子。某日，莊子為惡獸所包圍，莊子面不改色，引吭高歌，狀甚瀟灑，惡獸懼，皆沒入林中逃遁，莊子大笑，返以告鵬，鵬弗信，以為此乃水中撈月之事，絕無僅有，遂怒逐莊子，莊子甚感悲哀，準備遁世，後受仙人感召，出山林，作逍遙遊以告世人。

執編：李文堯 彭日樂 郭李同

EPOCHTIMES.COM

大紀元 - 台「北一女書包沒水準」楊照憶年少輕狂歲月

(大紀元記者江禹嬋台北報導) 17歲的青春歲月，該如何尋找自我？知名媒體人楊照說：「高中是擁有自我的開始，但卻還得活在別人給你的框架之中」所以他想盡辦法打破現有的規定，甚至在校刊上嘲弄隔鄰的女校，在文中嵌入「北一女的新書包沒水準」文字，他憶起，「那是我們

對樣本的要求

- 因為我們將從樣本推測出母體的原貌，抽出的部分必須能反映全體的特性，也就是說樣本需能代表母體。

→ ***樣本代表性！！！***

→ 最忌諱「瞎子摸象」

大數據層級的資料蒐集

- 產業界（如人力資源）大多仍以傳統問卷形式蒐集資料，很容易造成流失及扭曲。

→**冷備份**vs.**熱備份**（高成本！）

→次級資料vs.原始資料（自由心證！）

→樣本vs.母體（抽樣偏差！）

- 註：誘導性文字（「額外資訊」）、敏感性議題（「收入」）等也會有品質問題。

目標母體與實際母體

- 無論是實驗設計或是觀察研究，抽取樣本需要謹慎規劃，確保目標與實際兩者一致。

→例如：藉由民意調查獲取台北市長的施政滿意度，先確定受訪者為台北市民，可先詢問受訪者是否為「居住」在台北市的市民。

註：「戶籍人口」vs.「常住人口」

表3 匹茲堡睡眠品質量表(13)

請針對您過去一個月內夜間睡眠情形之大概狀況，回答最適合您情況的答案

1.過去一個月來，你通常何時上床？ _____時_____分

2.過去一個月來，你通常多久才能入睡？ _____分鐘

3.過去一個月來，你早上通常何時起來？ _____時_____分

4.過去一個月來，你實際每晚可以入睡幾小時？ _____時_____分

以下5、6、7、8題計分方式如下

0分：從來沒有 1分：一週少於一次 2分：一週兩次 3分：一週超過三次以上

5.過去一個月來，您睡眠問題被以下情況所干擾的次數如何？

(1)無法在30分鐘內入睡 ____ (2)半夜或凌晨便清醒 ____

(3)必須起來上廁所 ____ (4)覺得呼吸不順暢 ____

(5)大聲打鼾或咳嗽 ____ (6)會覺得冷 ____

(7)覺得躁熱 ____ (8)作惡夢 ____

(9)身上有疼痛 ____ (10)其他(請說明) ____

由受訪者填答可能衍生的問題？

表4 睡眠日誌(19)

日期	晚				午夜					早上							中午				下午					
Time	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6
1/1	E				◎	X	X	X	X	X	X	X	○	C					S	+				
1/2																								
1/3																								

◎ 熄燈或躺在床上試圖睡著 XXXX睡著的時段 ○ 開燈或起床 ++++半夢半醒 C 飲用咖啡因的飲料(咖啡、汽水或茶)

6.(A)請問您喝紹興酒或陳紹，那一種為主？□紹興酒為主 □陳紹為主 □紹興／陳紹一樣 □不喝

(B)您對下列酒的整體印象是：

	紹興／陳紹	紅葡萄酒	白葡萄酒	啤酒	白蘭地／威士忌	高粱酒／白酒
	非常贊成 贊成 沒意見 反對 非常反對	非常贊成 贊成 沒意見 反對 非常反對	非常贊成 贊成 沒意見 反對 非常反對	非常贊成 贊成 沒意見 反對 非常反對	非常贊成 贊成 沒意見 反對 非常反對	非常贊成 贊成 沒意見 反對 非常反對
健康	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
浪漫	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
休閒	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
高貴	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
實惠	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
活力	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
青春	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
豪爽	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
新潮	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
傳統/古板	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
帥氣	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
典雅	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
和諧	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
補身體	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□
料理酒	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□	□□□□□

7.(A)您想購買紹興／陳紹，下列那些地方曾讓您買不到？(可複選)

□沒買過 □餐廳 □路邊攤 □便利商店 □雜貨店 □KTV □酒廊

□PUB □超市 □量販店 □洋酒專賣店 □去買的地方，都買到 □其他_____

(B)上個月中您本人喝或用紹興酒／陳紹的場合為何？

	次數	百分比(飲用量／紹興/陳紹總用量)
婚喪大宴	_____	_____
KTV、酒廊、PUB	_____	_____
平常宴客	_____	_____
聚餐吃飯	_____	_____
自己小飲	_____	_____
※料理食物(作菜者才填)	_____	_____
其他_____	_____	_____

8.(A)您目前的行業是：(單選)

□軍公教 □農林漁牧業 □礦產及土石採取業 □製造業 □水電燃氣

□營造、建築業 □交通、運輸及倉儲業 □金融保險不動產業 □資訊、通信業(製造除外)

□貿易 □顧問、公關公司 □文化傳播娛樂,出版業 □餐飲業 □家庭主婦

□學生 □自由業 □服務業(商店,百貨....) □其他_____

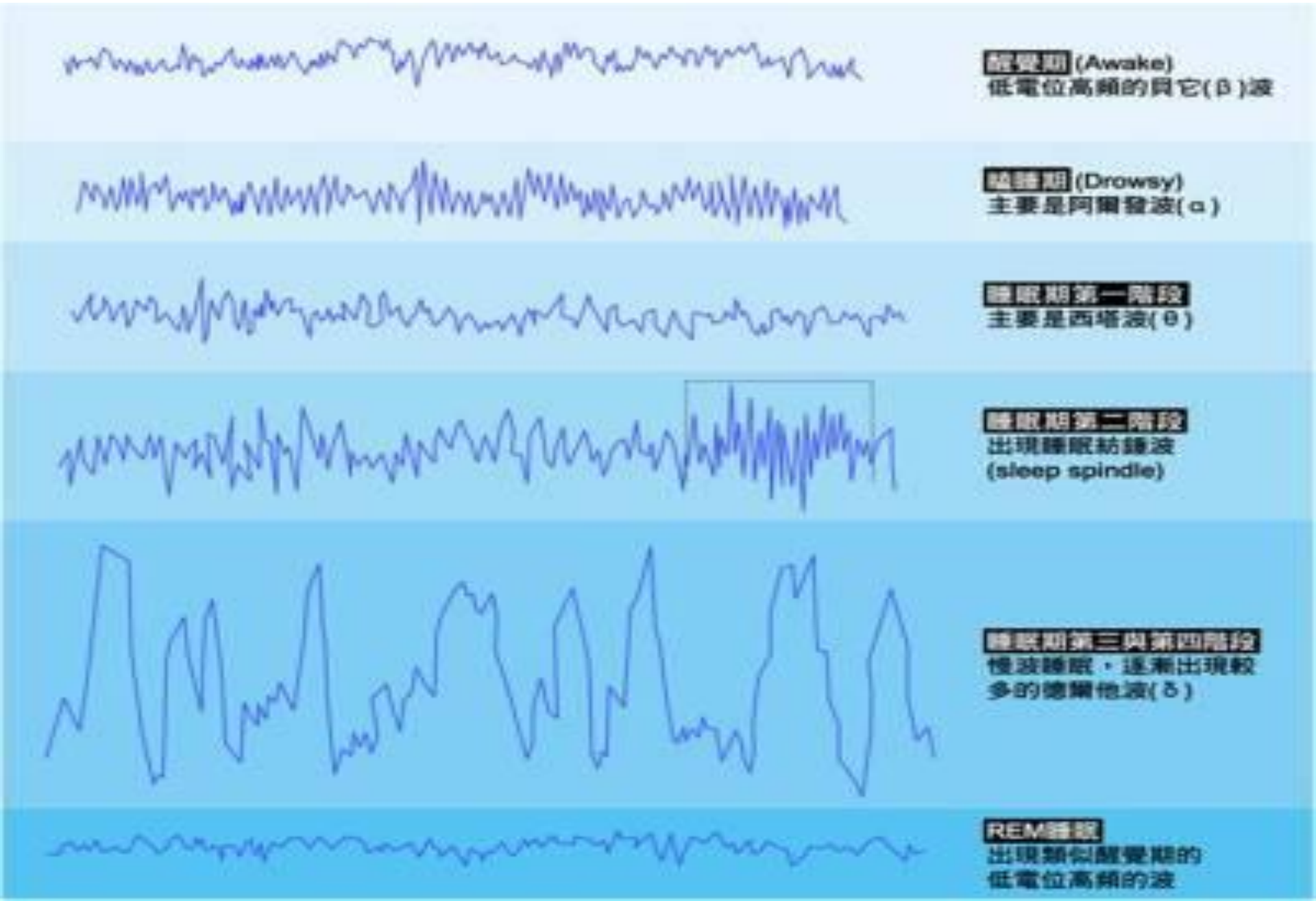
(B)您的工作型態為：(單選)

□業務人員 □行政事務人員 □勞務人員 □服務職 □知識性工作

□家庭主婦 □學生 □享清福

若您對本問卷有任何建議，請寫在背面空白處。

儀器量測更為精確，但資料蒐集不易！



抽樣方法的分類

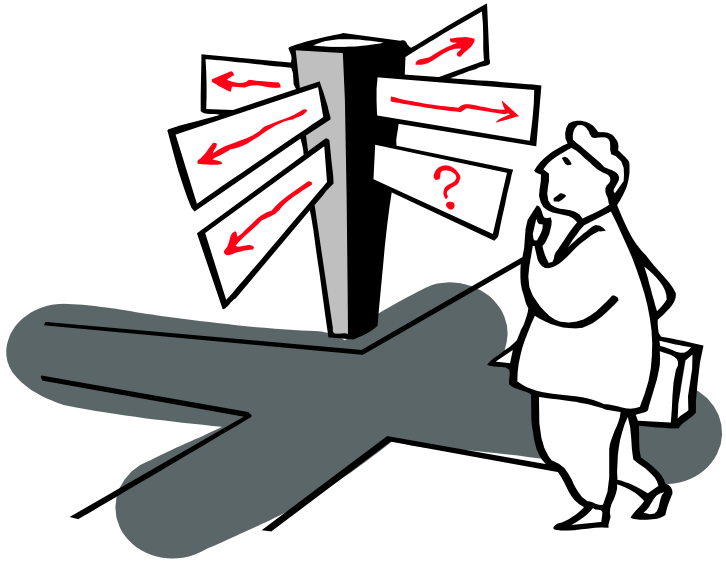
- 抽樣方法可分為隨機抽樣(或機率性抽樣，Random Sampling)及非隨機抽樣，前者不加人為意志，僅以隨機抽取樣本；後者按人為建議選取具有典型代表性樣本。

→隨機抽樣法因樣本以隨機抽出，較具代表性，但需要較完備的規劃，通常衍生的費用也較高。

註：**市話**、**手機**抽樣的差異、進行方式？

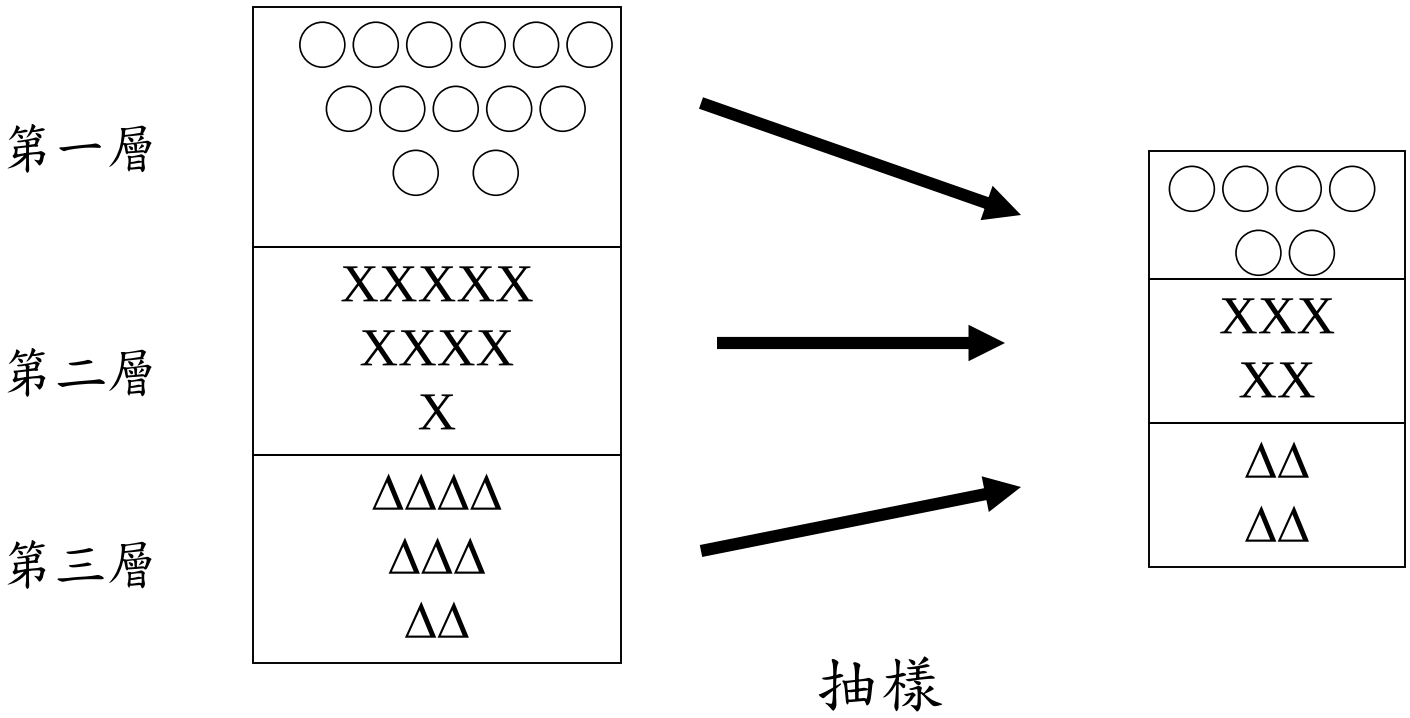
較常見的隨機抽樣方法

- 簡單隨機抽樣(Simple Random Sampling)
- 分層隨機抽樣
- 集體隨機抽樣
- 系統抽樣
- 兩段抽樣



→簡單隨機抽樣如同摸彩，將所有的個體逐一編號再抽出。

分層隨機抽樣(Stratified Random Sampling)



較常見的非隨機抽樣法

- 立意抽樣：不依隨機原則抽取樣本，由專家選取部份具有典型代表樣本。(e.g. 專家意見)
- 便利抽樣：事先不預定樣本，碰到即問或樣本自動回答。(e.g. 街頭調查)
- 滾球抽樣：利用樣本尋找樣本，對於特定族群樣本取得不易時採用。(e.g. 愛滋病的罹病人數)
- 配額抽樣：規定具有某種特性的樣本比例，類似分層隨機抽樣。

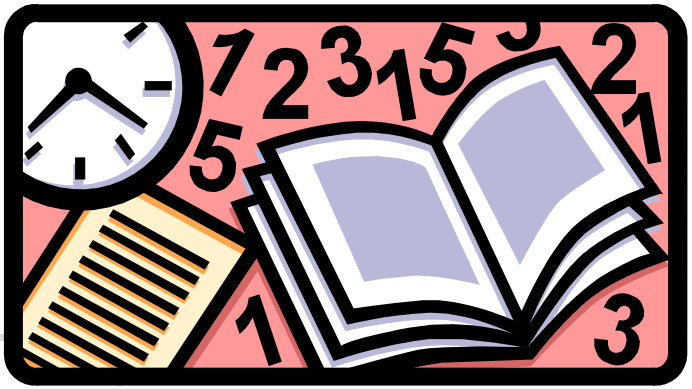
隨機抽樣 vs. 非隨機抽樣

- 隨機抽樣不代表每個個體被抽到的機會都相同，而是樣本選取不受人為因素影響！
- 問題：隨機抽樣有什麼優點？何時會選用非隨機抽樣？

→「隨機」意謂樣本選取與某個機率分配有關，估計及推論較有依據。

→對於罕見事件、蒐集未知領域的資訊，非隨機抽樣往往更合適。(例如：調適型抽樣，Adaptive Sampling)

問卷調查的步驟



- 定義問題、確定抽樣方法
- 問卷設計(Questionnaire Design)
- 問卷預試(Pretest)、訪員訓練
- 修訂問卷
- 正式訪問(發出問卷)
- 收回問卷、資料偵錯、資料輸入與整理

調查方法

- 調查方法通常可因獲取資料方法之異，通常分為：

(1)人員調查法(Personal Survey)

(2)電話調查法(Telephone Survey)

(3)郵件調查法(Mail Survey)

(4)網路調查法(Internet Survey)

問卷種類

- 問卷調查的題目通常分成三類：

1. 開放式：不列出可能答案，由被調查者自由作答。
2. 封閉式：(1)是否式(2)選擇式(3)排列式(4)填入式(5)尺度式
3. 半封閉式：封閉式為主體；若選項皆非填答者的選擇，則自由作答。

問卷題目範例

(1) 請問您本次購買的機車是

什麼廠牌_____ 汽缸大小_____c.c.

(2) 請問上一部機車行駛多少公里？

__15,000公里以下 __15,001~30,000公里

__30,001~45,000公里 __45,001~60,000公里

__60,001公里以上

（3）請問您打算幾年後換購新機車？

__1年以下 __1-2年 __3-4年

__5年以上 __其他(請說明)

一、基本資料：

捐血日期：_____　　捐血地點：_____

性別：□男　□女

年齡：□17-20 歲□21-30 歲□31-40 歲□41-50 歲□51-65 歲

教育程度：□國小□國中□高中/職□大學/專□研究所以上

職業：□工商□製造業□服務業□軍公教□學生□自由業□其他

二、引發您此次捐血的動機是：(可複選)□師長朋友鼓勵□醫療人員建議□相關宣導活動影響□完全自發性□其他_____

三、您對第一次捐血的感受是：(可複選)□有點緊張□非常緊張□有點不舒服□非常不舒服□很好□沒任何感覺□其他_____

四、您往後會持續捐血？□是□否□不確定；會鼓勵您的親朋好友捐血？□是□否□不確定

五、請就以下項目對捐血人的重要性及您對本次捐血過程的滿意度，加以勾選

項目	很不重要	不重要	普通	重要	很重要	很不滿意	不滿意	尚可	滿意	很滿意
1. 捐血地點是否容易尋找	□	□	□	□	□	□	□	□	□	□
2. 捐血步驟及注意事項於捐血地點是否有清楚的標示	□	□	□	□	□	□	□	□	□	□
3. 捐血環境是否整潔舒適	□	□	□	□	□	□	□	□	□	□
4. 進入捐血地點時，是否馬上有人員招呼	□	□	□	□	□	□	□	□	□	□

案例討論：資料輸入與變數格式

- 請至TVBS的網站(http://www.tvbs.com.tw)，點選「TVBS民調」，再點選「**93年06月18日**」的「七年級生愛情態度(1)-愛情合約」

→這個調查總共有三個檔案，分別是問卷及編碼、原始資料、調查報告。

→資料輸入可使用Excel，建議大家動手試試。

→資料編碼非常重要，瞭解變數格式及意義。

抽樣調查的浮濫與誤用

『在超市與藥房，私人贊助的調查取代了醫師、父母與藥劑師的地位；在法庭上，各類調查已取代律師的功能；在立法院，民意調查是人民的代言人；市調更是廣告與促銷最有用的利器。市調與民意的關係是一種詭異的循環，個人的信念被千百名陌生人的信念左右』

--- 《真實的謊言》，時報文化

問卷設計的步驟

- 決定研究目的、建立分析架構
- 決定問卷的形式
- 編擬問卷初稿
- 專家檢視與問卷修訂
- 問卷預試
- 問卷定稿、訂定使用說明

問卷設計的原則

(1) 確定整份問卷的目標

(2) 每個問題定義明確、
用字簡明易懂

(3) 問卷內容合乎邏輯

(4) 避免誘導、假設、敏感
的內容

(5) 與事後編碼及分析配合

問卷設計的原則(一)

(1) 確定整份問卷的目標

- 每個問題是否與整體目標相關

→避免東問一題、西問一題。不妨以到一個陌生的地方問路為例，你/妳的目標是找到某個地標或建築物。

- 設計問題也可以拼圖為例，每一塊小拼圖的個別角色、幾塊拼圖合成之後的角色又如何。

問卷設計的原則(二)

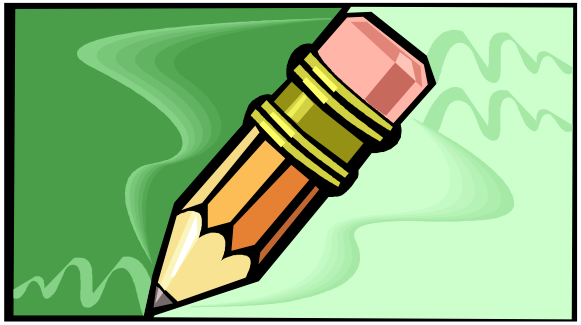
(2) 每個問題定義明確、用字簡明易懂

- 使用的文字、敘述方式以出現在日常生活為原則，不使用過於艱澀、但也避免過於口語化的用詞，也避免一詞多義。

例如：請問你/妳常接觸的傳播媒體？

→或可改為「請問你/妳最常接受新聞的來源」。

問卷設計的用字



- 詞意統一(Uniformity of meaning)
- 詞意明確(Preciseness of meaning)
- 避免偏見及誤導(Freedom from undue influence of prejudice or bias)
- 避免非理性及情緒化反應(Freedom from tendency to arouse irrational or extremely emotional response)
- 避免雙重否定的用詞

問卷範例

- *請問您對本郵局的滿意程度為何？*

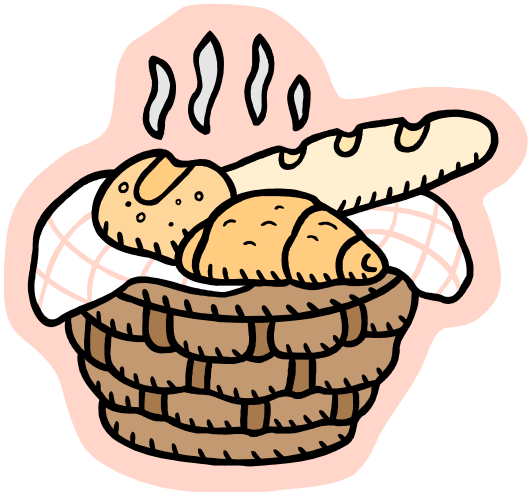
  □ 1.滿意　　□ 2.普通　　□ 3. 不滿意

→ 銀行郵局項目不只一項，例如可分為郵務、儲蓄、劃撥等項目，以便受訪者就各項目填寫意見。

- *請問您是否為素食者？*

  □ 1.是　　□ 2.否

→ 素食的定義不明！



問卷範例(續)

- *您家中共有幾人工作？*

→ 未註明調查所指的時間，也未說明是否也包含受訪者。

→「工作」指的是全職、兼職？

可能的修改方式：

- 今年9月30日您家中共有幾人(包括您自己)擁有經常性(每週平均20小時以上)的工作？

問卷範例(續)

https://score-more.com/ENG/wp-content/uploads/2021/12/different-types-of-sports-June32020-1-min.jpg

- *您經常做什麼運動？*

→ 每一個人對運動的定義不同。

- *您最近是否頭痛或生病？*
- *您是否會想藉由網路來獲得醫藥服務？例如：保健資訊，線上購藥，藥物諮詢。*

→ 項目太多，不知指的是哪一項。

https://www.liberty.edu/champion/2020/03/sports-are-canceled-and-im-not-ok-five-alternatives-for-fans-amid-coronavirus/

問卷設計的原則(三)

(3) 問卷內容合乎邏輯

- 問題順序應合乎邏輯，無論是以事件發生的先後順序，或是對受訪者的重要程度。

例如：

→請問你/妳有幾個小孩？

→請問你/妳結過婚嗎？

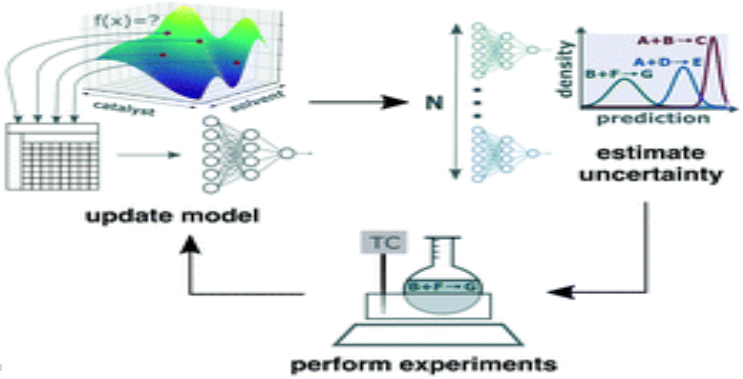
問卷設計的原則(四)

(4) 避免誘導、假設、敏感的內容

- 問卷文字忌諱引導受訪者，使之依照問卷設計者的意思回答，因此意見應包括正反兩方意見。(例如：贊成 vs. 反對)
- 另外，除非沒有更合適的替代方案，避免假設性或是令受訪者困窘的題目。

  (例如：請問你/妳是否曾有過婚外情。)

https://pubs.rsc.org/en/content/articlelanding/2020/re/d0re00232a



實驗設計

- 問卷調查蒐集的資料絕大多數屬於觀察研究(Observational Study)，經常無法確定觀察出的結果之成因。

  → 例如：研究發現國小學童中腳較大者，拼字能力也較強。(腳的大小影響拼字？)
- 實驗設計控制外在環境，只容許有興趣的部分(稱為「處理」；Treatment)變動，藉以分離出影響結果的原因。

實驗設計與因果關係

- 由實驗設計應可推論出較精確的結果，但實驗設計的人力、金錢、時間的需求較高，且需更為精密的事前規劃。然而，實驗設計也無法使用於所有情形，有時問卷調查是唯一可能獲得資料的方法。

- 原則上而言，實驗可對因果關係提供好的證據。

關於臨床試驗

https://www.quora.com/What-is-a-double-blind-procedure

- 實驗組→處理；處方

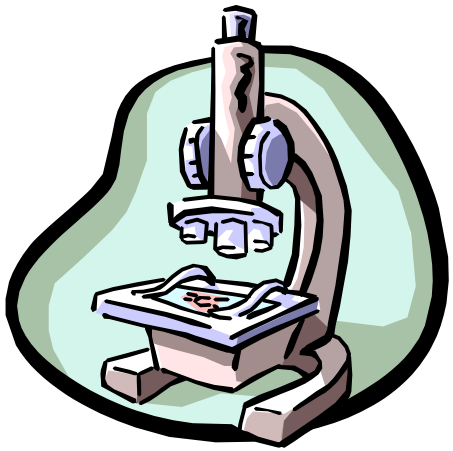
  對照組→安慰劑(Placebo)

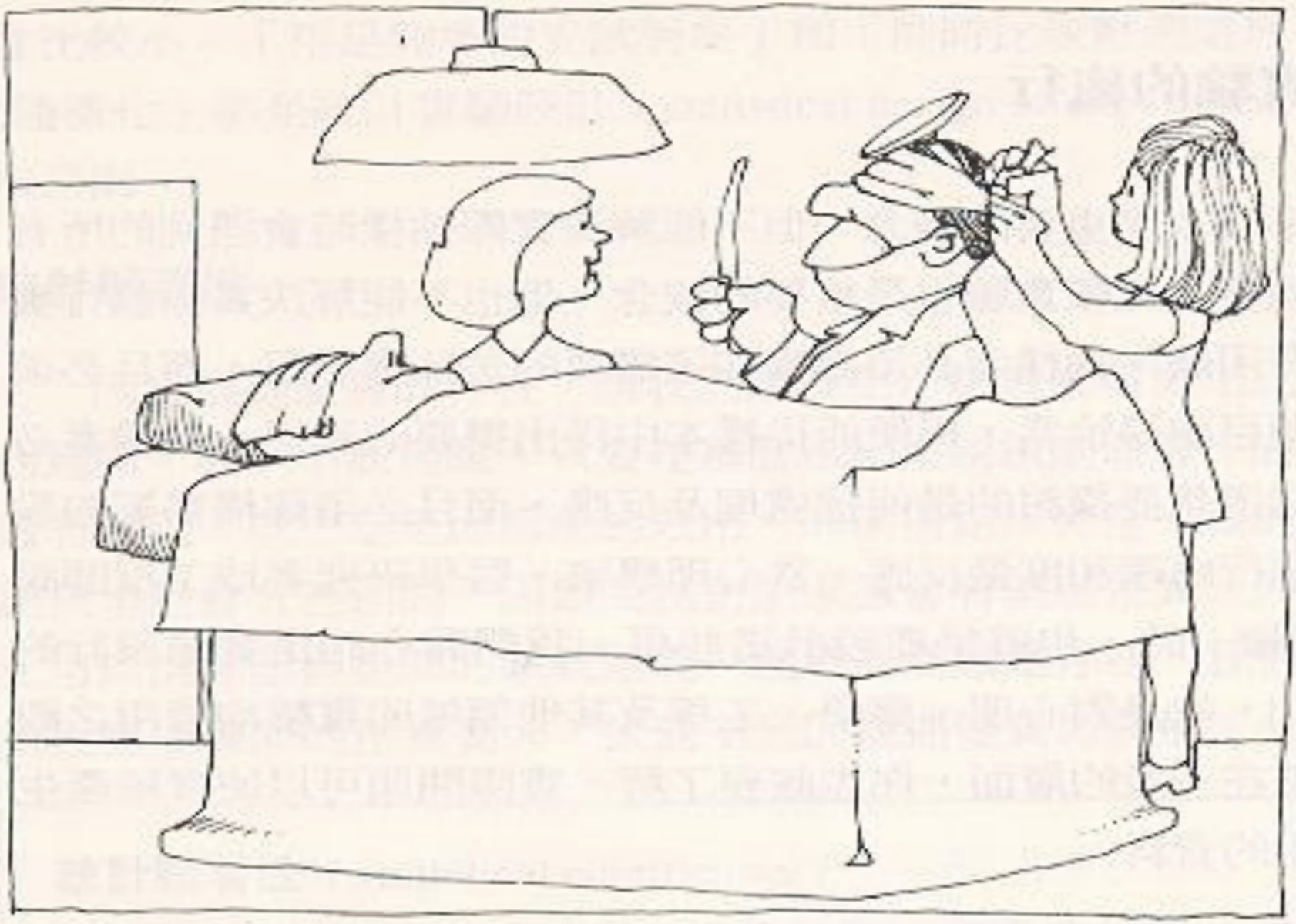
- 單盲與雙盲實驗：

→ 單盲：只有受試者不知道自己的處方

→ 雙盲：醫生與受試者都不知道

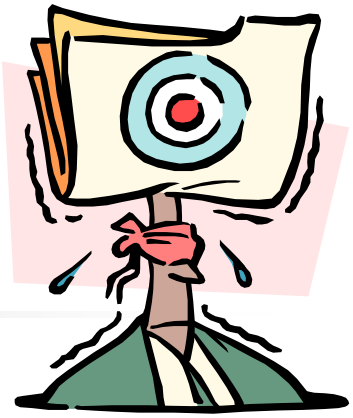
處方的分配





「伯恩斯醫師，您確定統計學家說的雙盲實驗是這個意思嗎？」

設計實驗的成本以外考量



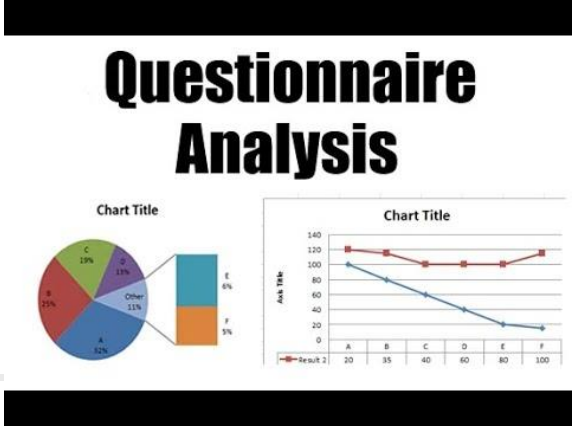
- 道德因素(Ethical factor)
  →讓重病病患使用可能較差的處方(或服用安慰劑)，雖然可證明實驗處方較佳，但也因此令病人縮短壽命(Patients' Right)。
- 公共政策的實驗
  →新的福利制度、健康保險等等公共政策的制訂，經常根據很多想像與很少資訊。對問題較小的政策且需比較的處理明確，通常較容易成功。

統合分析(Meta-analysis)

- 因為成本、時間及其他因素，研究可能需要合併不同研究的資料，這些資料可能有來不同來源、蒐集時間不一樣、或甚至有不同母體，如何因應問題需要結合資料，也是近年來另一種資料蒐集的方法。

→例如：各國蒐集該國罹患SARS、AIDS等疾病，希望找到共同的特性；選舉研究如何整合不同地區及時間得出的電訪結果，以獲得當前選舉的趨勢。

https://i.ytimg.com/vi/SEZT78XYhQw/hqdefault.jpg



問卷設計的統計分析

- 敘述性統計量(Descriptive Statistics)
- 相關性分析
- 卡方檢定(Chi-Square Test)
- 因素分析(Factor Analysis)
- 其他方法(例如：迴歸、時間數列、存活分析、類別資料分析)

樣本代表性

- 樣本代表性意指抽出的樣本，其特性與母體非常類似，足以由樣本來代表整個母體。

→ 檢查樣本代表性是分析問卷資料的首要步驟，若樣本與母體差異過大，則以樣本的資訊推測母體的特性，將顯得不合適。

常見的樣本代表性檢查項目

- 通常用於檢查樣本代表性的問項：

1.性別比例

2.年齡結構

3.居住地區

4.教育程度

5.職業別

6.婚姻狀態

7.其他因素

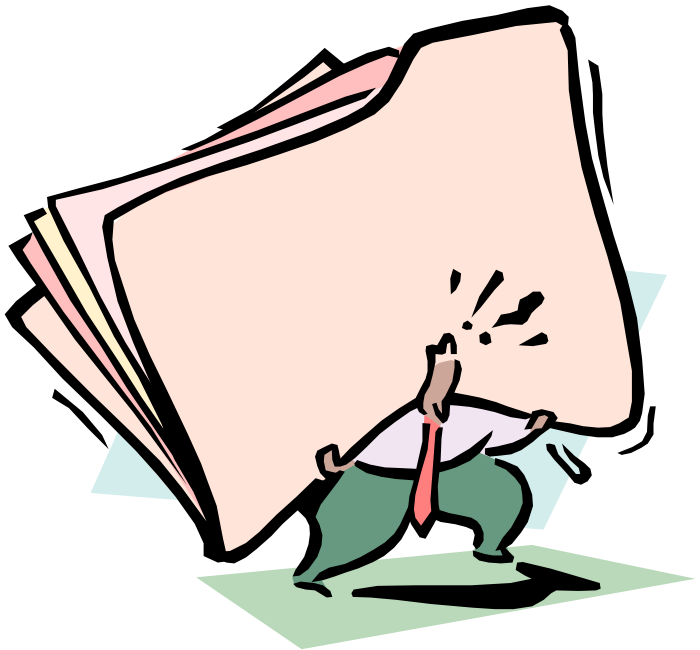
樣本代表性(續)

- 台灣地區的人口統計資料可到內政部統計處查詢，網址 https://www.moi.gov.tw/cp.aspx?n=5590

→內政統計年報

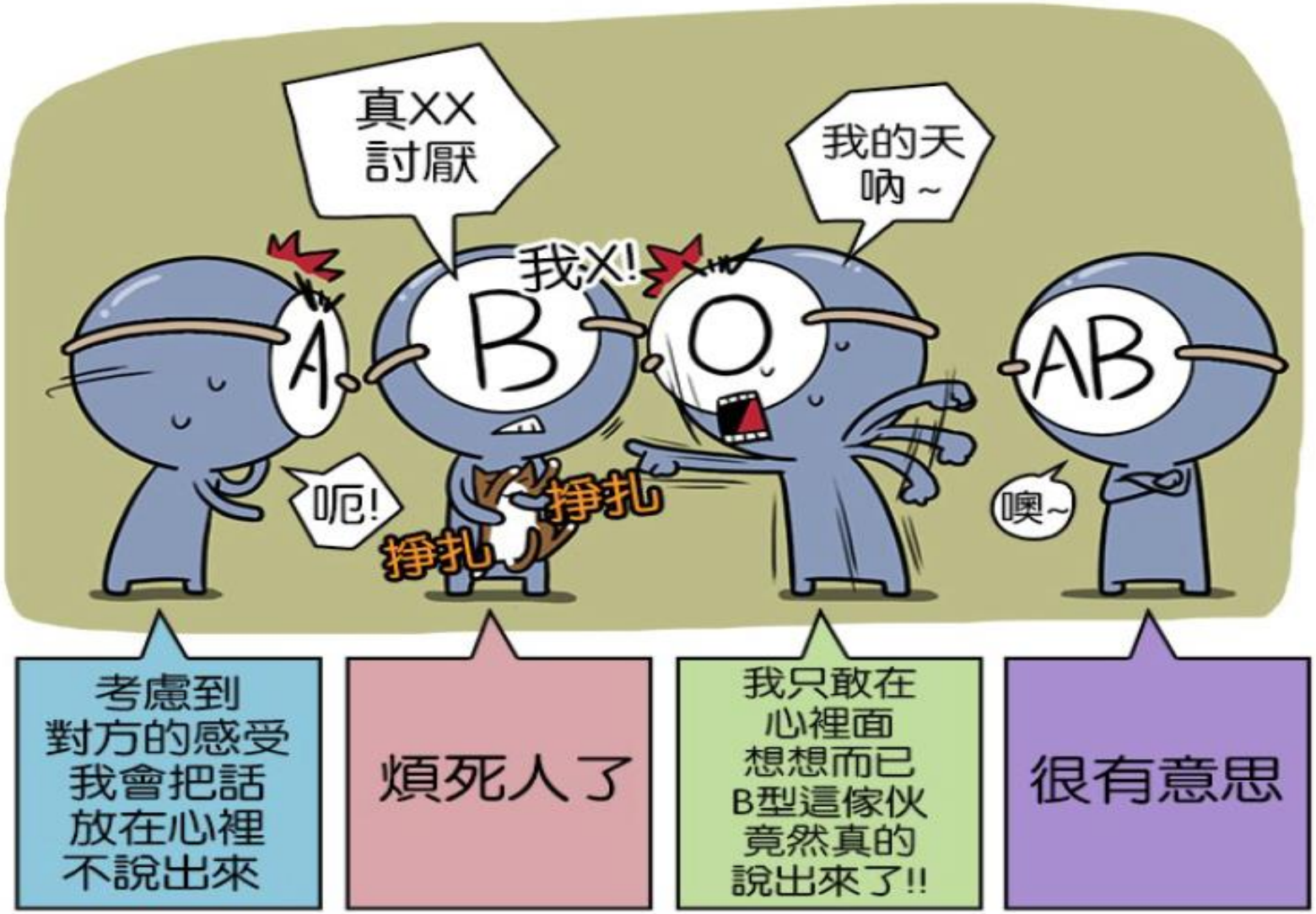
→性別統計資料

→重要參考指標



不同血型遇到相同狀況時。。。

每個人表現出來的性格都不同



樣本代表性的實例

- 美國1930年代調查業巨人《文摘》(Literary Digest)，1936年發出一千萬問卷，預言共和黨總統候選人Alfred Landon將大獲全勝。

→調查對象為擁有汽車與電話的家庭。

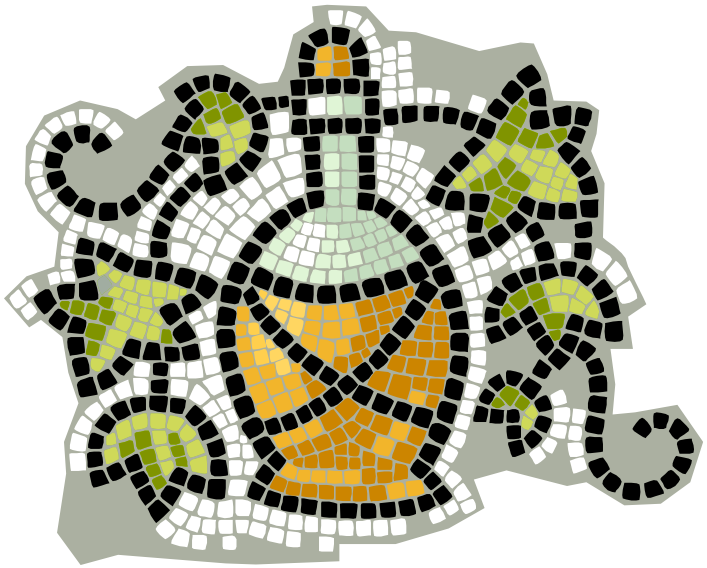
- 但蓋洛普等三大業者1948年也錯誤預測共和黨杜依會獲勝。

→使用配額抽樣法，訪問五百位家庭主婦、二十位農夫與三百位老人。

註：資料來源【真實的謊言】

樣本代表性分析範例：

- 菸酒公賣局(現在的台灣菸酒股份公司) 委託台灣大學管理學院辦理的民眾飲用紹興酒的習慣調查。



https://cdn2.ettoday.net/images/4903/d4903906.jpg

1.(A)請問您喝酒的經驗是？(可複選)　□紹興酒／陳紹　□其他酒

(B)最近 30 天內您共喝了_____瓶酒，其中紹興酒／陳紹佔_____%。

(C)如果您喝紹興／陳紹，今年喝的量與去年相比，預計會

□增加　□減少　□一樣　□不喝

2.您曾經買或喝紹興／陳紹，下列因素的重要性為何？(從未喝過或未買者可免填)

	極強	強	中	弱	無
別人請客	□	□	□	□	□
請客人	□	□	□	□	□
送禮	□	□	□	□	□
口感好	□	□	□	□	□
習慣	□	□	□	□	□
價格適中	□	□	□	□	□
購買方便	□	□	□	□	□
解悶	□	□	□	□	□
強身	□	□	□	□	□
流行	□	□	□	□	□

3.如果您減少喝或不喜歡紹興／陳紹，下列因素的重要性為何？(多喝者可免填)

	極強	強	中	弱	無
別人請客時，減少提供紹興／陳紹	□	□	□	□	□
自己請客，減少提供紹興／陳紹	□	□	□	□	□
收到紹興／陳紹禮品的機會減少	□	□	□	□	□
酒的選擇增多	□	□	□	□	□
知名度不足(媒體少報導)	□	□	□	□	□
身體狀況不允許	□	□	□	□	□
飲用後，會頭痛，打嗝難聞	□	□	□	□	□
品質不穩定	□	□	□	□	□
口感不好，氣味不好	□	□	□	□	□

4.下列的酒會取代您喝紹興／陳紹的意願為何？(每一項單獨與紹興／陳紹相比)

	極強	強	中	弱	無
紅葡萄酒	□	□	□	□	□
白葡萄酒	□	□	□	□	□
啤酒	□	□	□	□	□
白蘭地(含 XO)/威士忌	□	□	□	□	□
高梁酒/白酒	□	□	□	□	□
水果酒/清酒	□	□	□	□	□

5.要讓紹興類酒銷路更好，您認為下列方法有效程度為何？

	極強	強	中	弱	無
創造流行，塑造形象	□	□	□	□	□
大力廣告	□	□	□	□	□
贈品	□	□	□	□	□
折價	□	□	□	□	□
增加銷售點，方便購買	□	□	□	□	□
改變口感	□	□	□	□	□
改變包裝	□	□	□	□	□
其他_____	□	□	□	□	□

6.(A)請問您喝紹興酒或陳紹，那一種為主？□紹興酒為主 □陳紹為主 □紹興／陳紹一樣 □不喝

(B)您對下列酒的整體印象是：

	紹興／陳紹					紅葡萄酒					白葡萄酒					啤酒					白蘭地／威士忌					高粱酒／白酒				
	非常贊成	沒意見	反對	非常反對		非常贊成	沒意見	反對	非常反對		非常贊成	沒意見	反對	非常反對		非常贊成	沒意見	反對	非常反對		非常贊成	沒意見	反對	非常反對		非常贊成	沒意見	反對	非常反對	
健康	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
浪漫	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
休閒	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
高貴	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
實惠	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
活力	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
青春	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
豪爽	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
新潮	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
傳統/古板	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
帥氣	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
典雅	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
和諧	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
補身體	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
料理酒	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□

7.(A)您想購買紹興／陳紹，下列那些地方曾讓您買不到？(可複選)

□沒買過 □餐廳 □路邊攤 □便利商店 □雜貨店 □KTV □酒廊

□PUB □超市 □量販店 □洋酒專賣店 □去買的地方，都買到 □其他_____

(B)上個月中您本人喝或用紹興酒／陳紹的場合為何？

	次數	百分比(飲用量／紹興/陳紹總用量)
婚喪大宴	_____	_____
KTV、酒廊、PUB	_____	_____
平常宴客	_____	_____
聚餐吃飯	_____	_____
自己小飲	_____	_____
※料理食物(作菜者才填)	_____	_____
其他_____	_____	_____

8.(A)您目前的行業是：(單選)

□軍公教 □農林漁牧業 □礦產及土石採取業 □製造業 □水電燃氣

□營造、建築業 □交通、運輸及倉儲業 □金融保險不動產業 □資訊、通信業(製造除外)

□貿易 □顧問、公關公司 □文化傳播娛樂,出版業 □餐飲業 □家庭主婦

□學生 □自由業 □服務業(商店,百貨....) □其他_____

(B)您的工作型態為：(單選)

□業務人員 □行政事務人員 □勞務人員 □服務職 □知識性工作

□家庭主婦 □學生 □享清福

若您對本問卷有任何建議，請寫在背面空白處。

表一 樣本基本資料表

n = 1,400

樣本分佈	樣本數	比例
台北	536	38.3
桃園	80	5.7
台中	162	11.6
嘉義	38	2.7
台南	84	6.0
高雄	233	16.6
屏東	36	2.6
基隆	16	1.1
新竹	24	1.7
苗栗	25	1.8
花蓮	11	0.8
宜蘭	29	2.1
彰化	60	4.3
雲林	30	2.1
南投	26	1.9
台東	10	0.7

年齡	樣本數	比例
15-19	87	6.2
20-24	251	17.9
25-29	279	19.9
30-34	219	15.6
35-39	165	11.8
40-44	132	9.4
45-49	86	6.1
50-54	60	4.3
55-59	27	1.9
60-64	21	1.5
65+	35	2.5
未填	38	2.7

性別	樣本數	比例
男	939	67.1
女	433	30.9
未填	28	2

教育程度	樣本數	比例
無	12	0.9
小學	64	4.6
初中	107	7.6
高中	436	31.1
專科	352	25.1
大學以上	394	28.1
未填	35	2.5

行業	樣本數	比例
軍公教	162	11.6
農林漁牧業	28	2
礦產及土石採取業	10	0.7
製造業	109	7.8
水電燃氣	15	1.1
營造建築	53	3.8
交通運輸及倉儲業	38	2.7
金融保險不動產業	62	4.4
資訊/通信業	56	4
貿易	58	4.1
顧問/公關公司	12	0.9
文化傳播娛樂出版業	20	1.4
餐飲業	43	3.1
家庭主婦	61	4.4
學生	229	16.4
自由業	104	7.4
服務業	212	15.1
其他	98	7
未填	30	2.1

工作型態	樣本數	比例
業務人員	158	11.3
行政事務人員	194	13.9
勞務人員	207	14.8
服務職	309	22.1
知識性工作	132	9.4
家庭主婦	66	4.7
學生	224	16
享清福	63	4.5
未填	47	3.3

表二 樣本與母體之地域比較

	樣本(n=1,400)		母體(N=21,592,147)	
	人數	比例(%)	人數	比例(%)
台北	536	38.3	6019028	27.9
桃園	80	5.7	1614471	7.5
台中	162	11.6	2349722	10.9
嘉義	38	2.7	830517	3.8
台南	84	6.0	1814062	8.4
高雄	233	16.6	2663302	12.3
屏東	36	2.6	913764	4.2
基隆	16	1.1	379370	1.8
新竹	24	1.7	773521	3.6
苗栗	25	1.8	560344	2.6
花蓮	11	0.8	358077	1.7
宜蘭	29	2.1	466603	2.2
彰化	60	4.3	1297744	6.0
雲林	30	2.1	751913	3.5
南投	26	1.9	546707	2.5
台東	10	0.7	253002	1.2

註一: Pearson 卡方 = 7.617, 自由度 = 15, p = .938。
註二: 本表並未將"未填"者列入分析。

- 分析的可能瑕疵：

→台北地區的樣本數536人(38.3%)，已遠比母體中的比例值 27.9%高，單就是否屬於台北地區來分類，可得：

$$\chi^2 = \frac{(391-536)^2}{391} + \frac{(1009-864)^2}{1009} = 74.61 > 3.841 = \chi_1^2(0.05)$$

抽到的樣本有過多屬於台北地區，因此樣本與母體就居住地區而言，兩者已有顯著差異。若以樣本的分析結果推論母體也具有這些結果，本身存有疑問。

資料分析vs.資料品質

- 許多人宣稱**資料量多寡**比**資料品質**重要，但「Garbage in, garbage out」，偏頗資料會扭曲我們的判斷（如：何不食肉糜）。
- 資料科學家**分析前應先確認資料來源可信度**，檢查資料品質是否有重大瑕疵。

→網路有名的「世界四大不能信」：英國研究、臺灣報導、中國製造、韓國發源。

https://3.bp.blogspot.com/-WIlANRKP6M0/UX78_Egp_NI/AAAAAAAAAD4/SZNirCpqejk/s1600/social+bias+pic+5.jpg

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。

→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

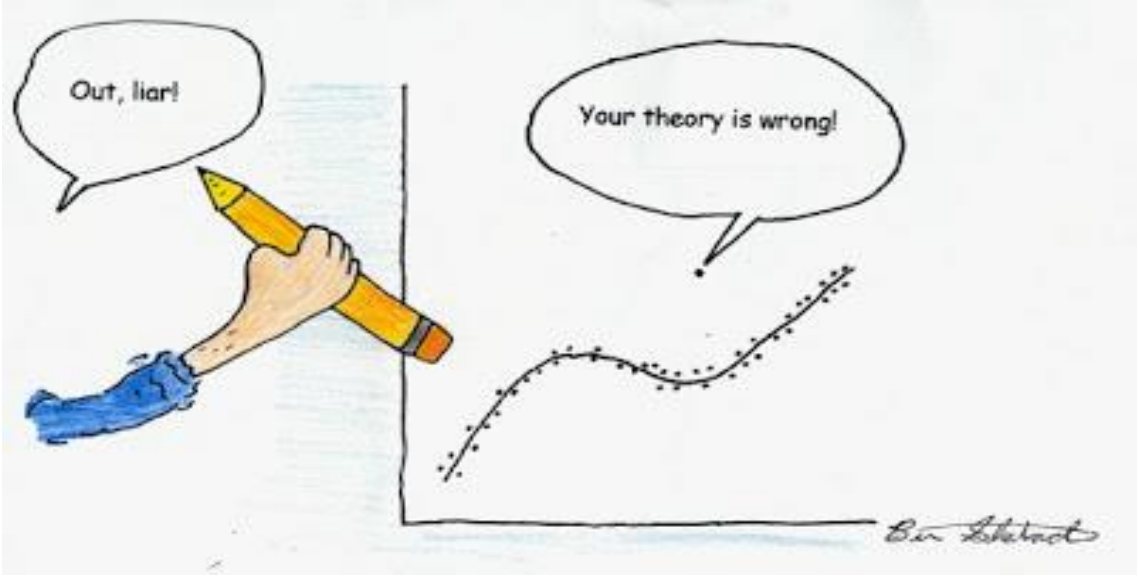
- Intense
- High
- Moderate
- Low
- Minimal

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 »

倖存者偏差(Survivorship Bias)

- 有時無法判斷蒐集到局部或全體的資料，檢查是否存在倖存者偏差（例如：谷歌流感趨勢預測）；不少人為了省事與得到漂亮結果，而移除離群值！



樣本好壞的判斷

- 母體的特性→參數(Parameter)
- 樣本中用以推測母體特性的估計值
  → 統計量(Statistic)

- 對統計量的要求：

(1) 不偏(Unbiased): E(統計量) = 參數

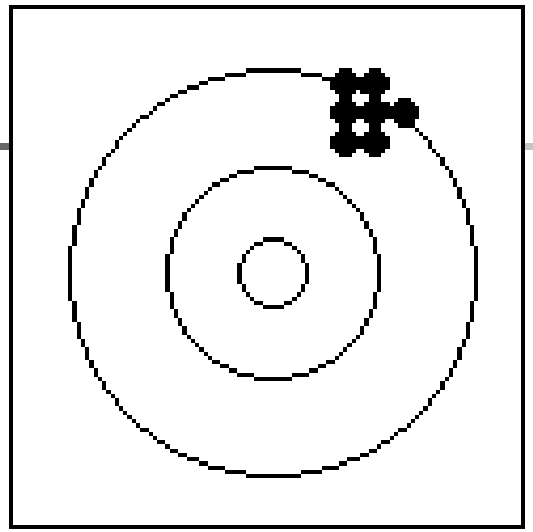
(2) 變異數(Variance)愈小愈好

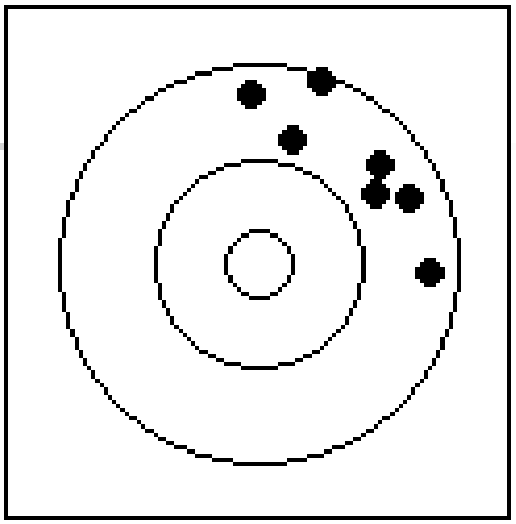
→ 變異數與風險(Risk)有相似的涵意

Precise

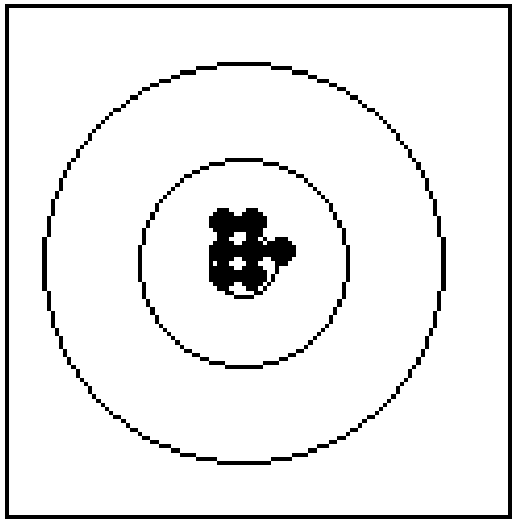
Imprecise

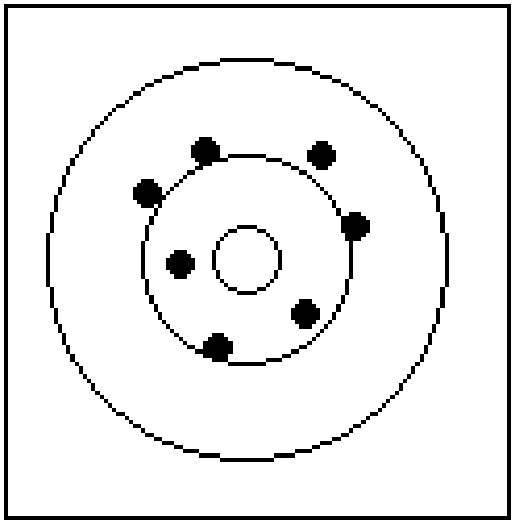
Biased





Unbiased



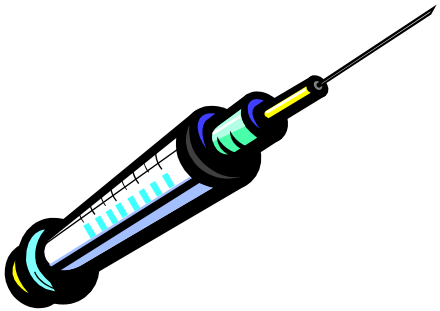


多少樣本才足夠？

- 抽樣時常見的迷思：

→樣本數必須達到母體的一定比例？

範例(1)一般抽血不多於 10 c.c.，不論大人或小孩。



範例(2)台灣與美國人口數差了10倍以上，但民意調查多半只抽1,000份左右。

抽取1,000份樣本的原因

- 民意、市場調查的多為封閉問卷，有興趣的多為某個問項佔的比例，例如：某位候選人的支持程度→二項分配。
- 在信心水準為95%及最大誤差不大於 3% 的要求下：

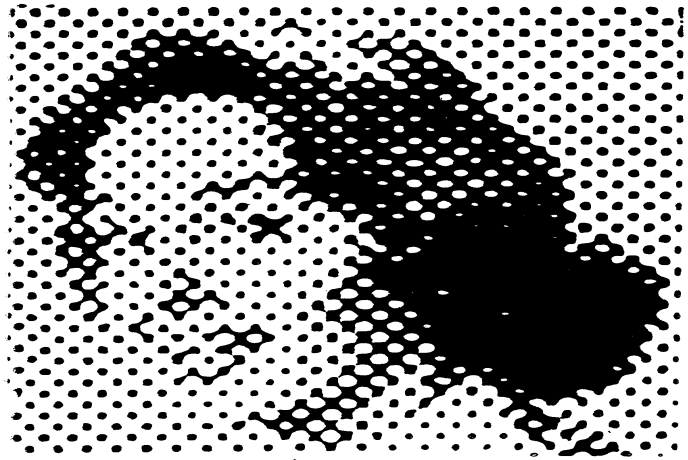
$$1.96\times\sqrt{\frac{p(1-p)}{n}}\le 0.03$$

$$\Leftrightarrow \sqrt{n}\ge \frac{1.96\sqrt{p(1-p)}}{0.03}\cong \frac{1.96\times 1/2}{0.03}$$

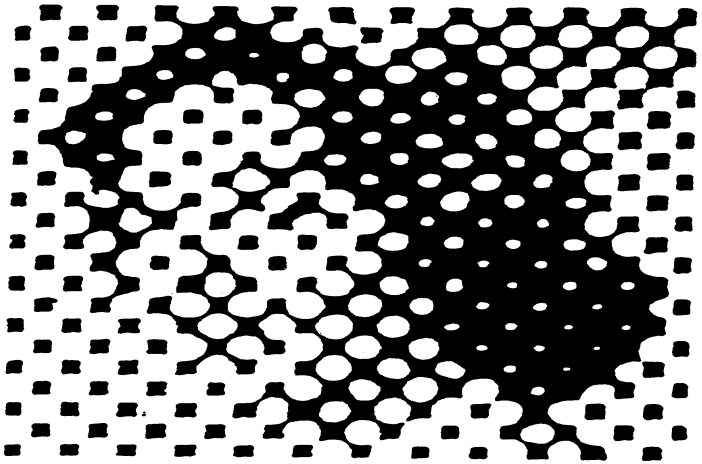
$$\Leftrightarrow \quad n\ge 1{,}067$$

500,000

2,000



1,000

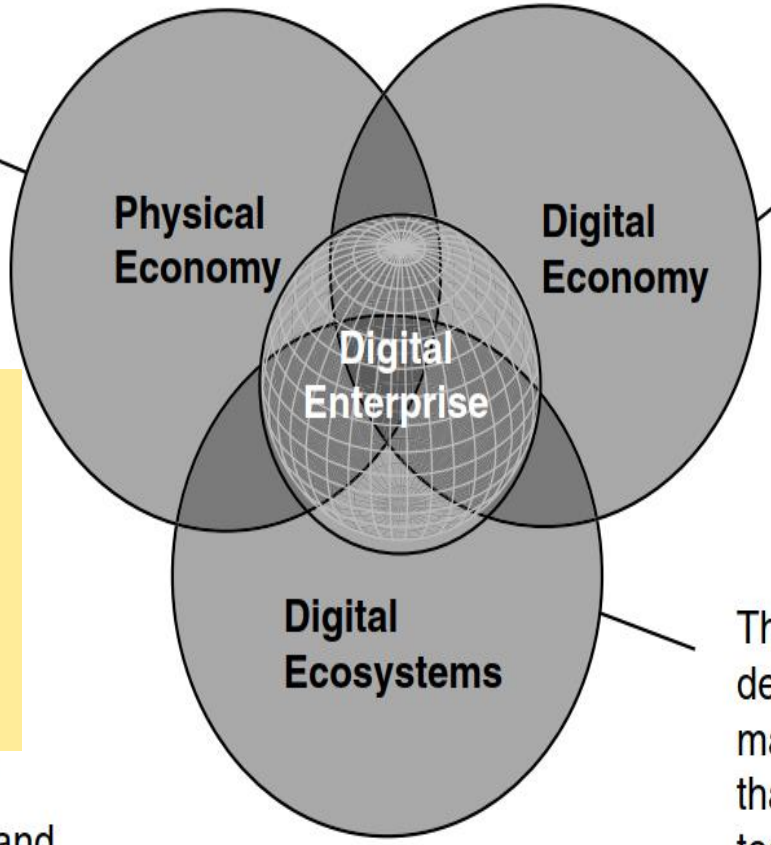


250

樣本對母體之代表性

大數據時代的經濟架構

Physical markets, companies, resources, and services that contribute to GDP and net worth



Virtual resources and digital transactions in markets, companies, resources, and services that contribute to GDP and net worth

The physical economy, digital economy, and digital ecosystem

Technological Ecosystems are shaping the supply chain and define the digital enterprise

The digital ecosystem is a described boundary of a market and business activity that is using connected technologies to enable a new kind of market and business performance and user experience

如何分析非結構資料？

- 生活中大多數為非結構化資料，但如何分析尚無定論。
- 蒐集方法、測量值都是關鍵！

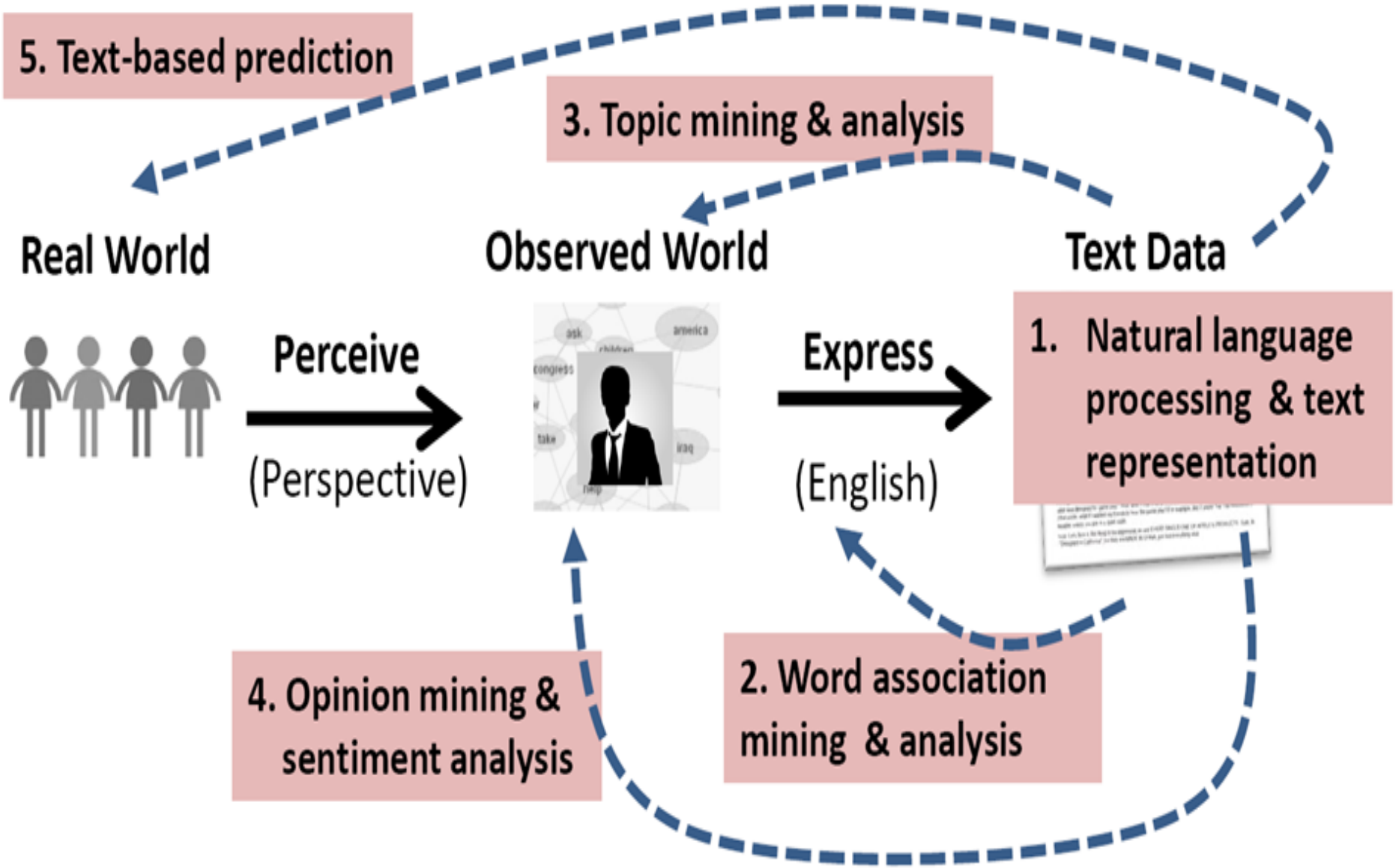
→日本311大地震「救命大數據」。

https://image.cache.storm.mg/styles/smg-800xauto-er/s3/media/image/2016/03/09/20160309-060650_U3314_M135842_f21c.jpg?itok=bU042Js3

www.shutterstock.com · 1654013521

https://image.shutterstock.com/image-vector/cloud-on-premise-etl-elt-600w-1654013521.jpg

文字分析的可能步驟


5. Text-based prediction
3. Topic mining & analysis
Real World
Observed World
Text Data
Perceive
(Perspective)
Express
(English)
1. Natural language processing & text representation
4. Opinion mining & sentiment analysis
2. Word association mining & analysis

文字資料的前置分析

文字資料的統計分析包含以下步驟：

- Data Collection
- Text Parsing and Transformation

→包括斷句、篩選相關資料段落、定義關鍵字詞。

- Text Filtering

→挑選合適關鍵字詞。

- Text Mining

→Clustering, Classification, Association, and Link Analysis.

斷詞、關鍵詞與統計分析

- 關鍵詞(Keywords)猶如統計模型的變數，根據研究目的及問題定義，挑選適當變數以提分析的效率和準確性。

→傳統統計模型也注重解釋性。

- 中文分析先經過斷詞，接著再從中挑選出重要關鍵詞。

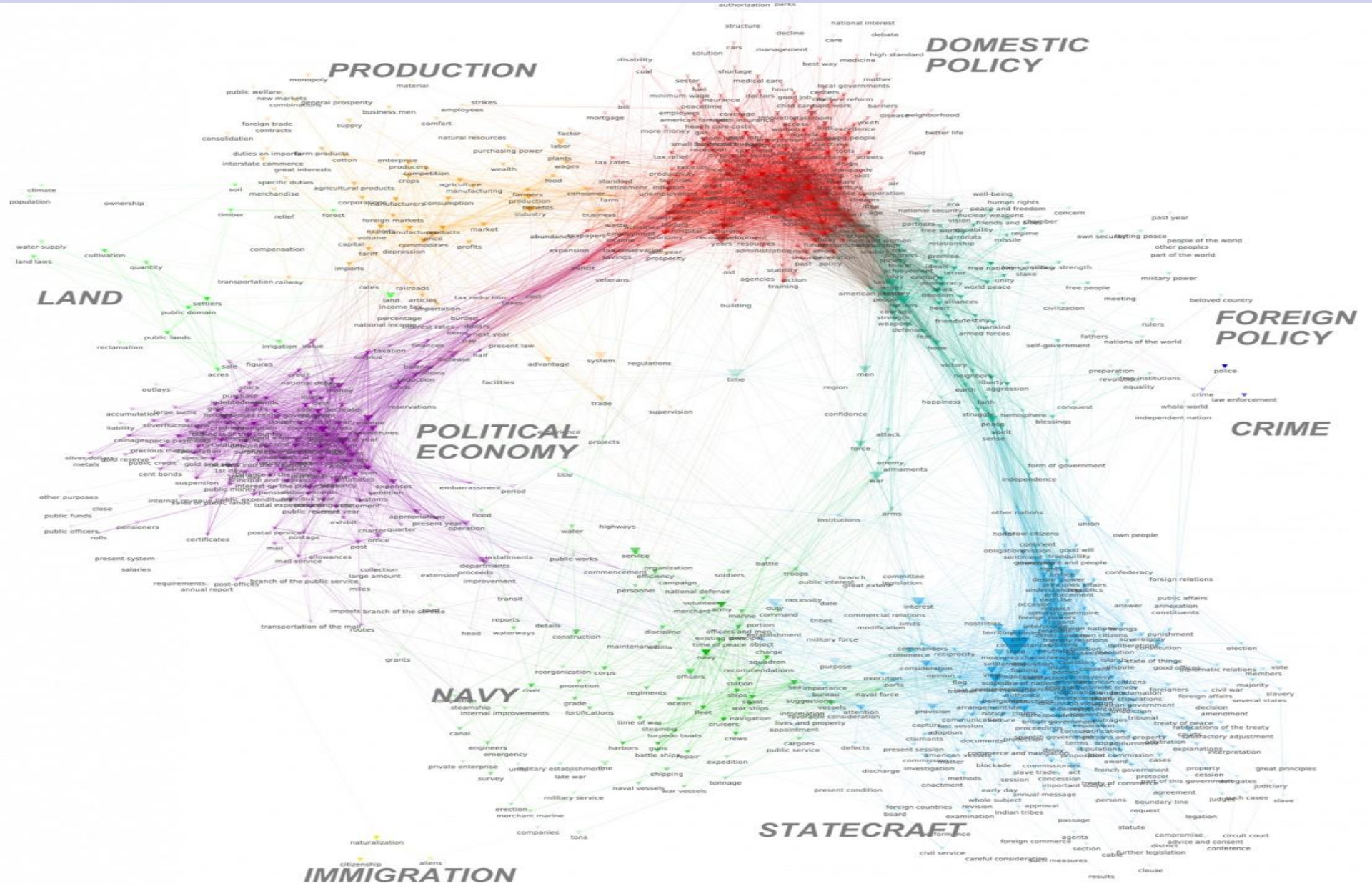
→白話文多以雙字詞（多字詞）表達觀念。

（註：字→詞→有意義的詞）

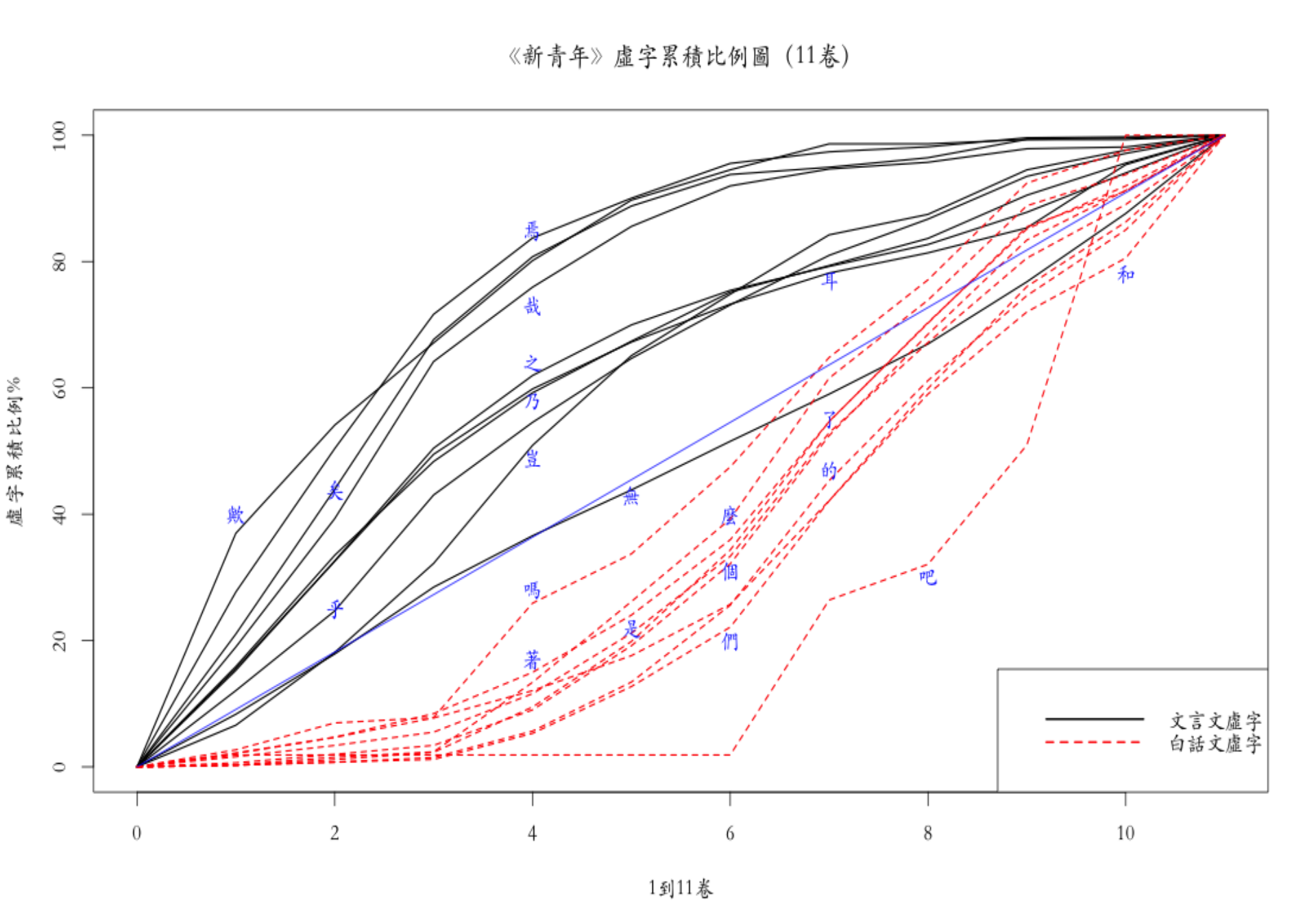
文言文、白話文的比較（維基百科）

比較	文言文	白話文/現代中國語文
長短	言簡意賅	較長篇
出處用法	書面語為主	「我手寫我口」為主，亦經修飾
語感	古雅精煉	通俗易明
文法詞組次序	彈性較大	次序明確
用詞1	單字已有獨立意思	二字詞為主
用詞2	一字多用	異字異用
用詞3	之	的
句末助語詞	已、矣、乎、也.....	了、吧、啊、嗎.....
標點	標點少而簡，句讀為主	標點繁多
經典例	《桃花源記》、《醉翁亭記》、《庖丁解牛》、《出師表》、《六國論》.....	魯迅《吶喊》自序、朱自清《綠》、冰心《紙船》、舒乙《香港：最貴的一棵樹》.....
流傳	限於曾學習文言的人，須有一定傳統文學修養，但可於東亞通行	一般中小學生也能看懂，廣傳於華文世界
習法	背誦為主，輔以字詞拆解	字詞拆解為主，文法分析輔助

美國總統國情咨文(State of Union,1790~)



來源：https://sciencenode.org/feature/text-mining-strikes-gold-in-political-discourse.php



文言、白話虛詞的趨勢變化

雙字詞的關連性

《人民日報》的常見雙字詞組合

First	1st	2nd	3rd	4th
人民	反对	服务	群众	群众
群众	意見	一起	意見	关心
发展	农业	速度	战略	道路
国家	之间	节约	主席	主席

First	文本	Most Likely Phrase
中國	People's Daily	人民、大使、特色、特色
	United Daily News	代表、歷史、原則、崛起