巨量資料與統計分析

政治大學統計系余清祥

2019年11月19日

第十週：非結構資料分析

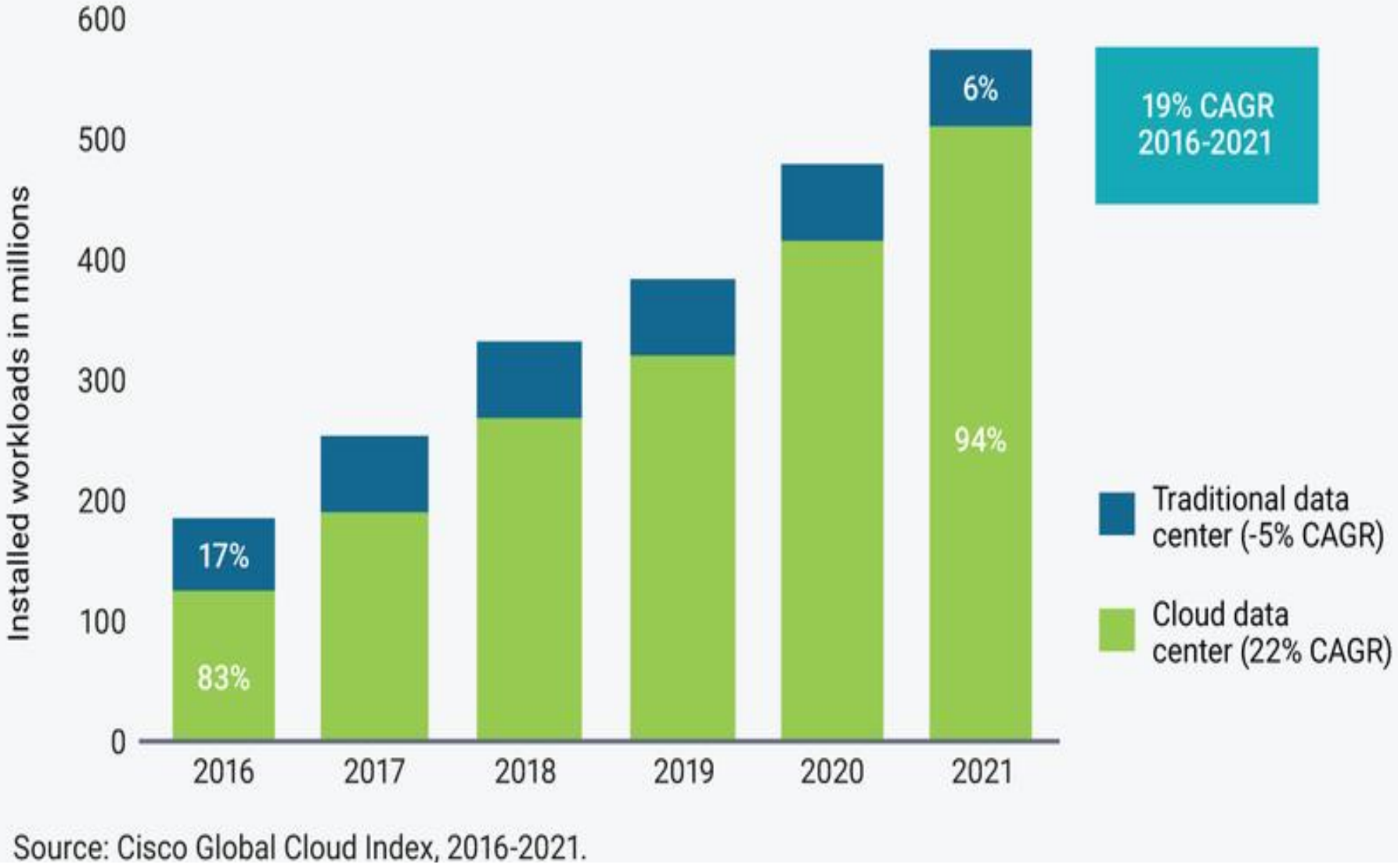
http://csyue.nccu.edu.tw

大數據與非結構資料

- 研究指出產業界產生的資料至少有80%為非結構資料，這個比例將逐年增加，2010年後已經超過90%且繼續攀升。
- 有別於結構資料，非結構資料沒有固定欄位，視問題而量化資料、選擇研究目標，更需結合應用領域的知識，尋求合適的EDA方法。

→非結構資料的EDA尚無統一的進行方式，經常隨著研究者興趣而不同。

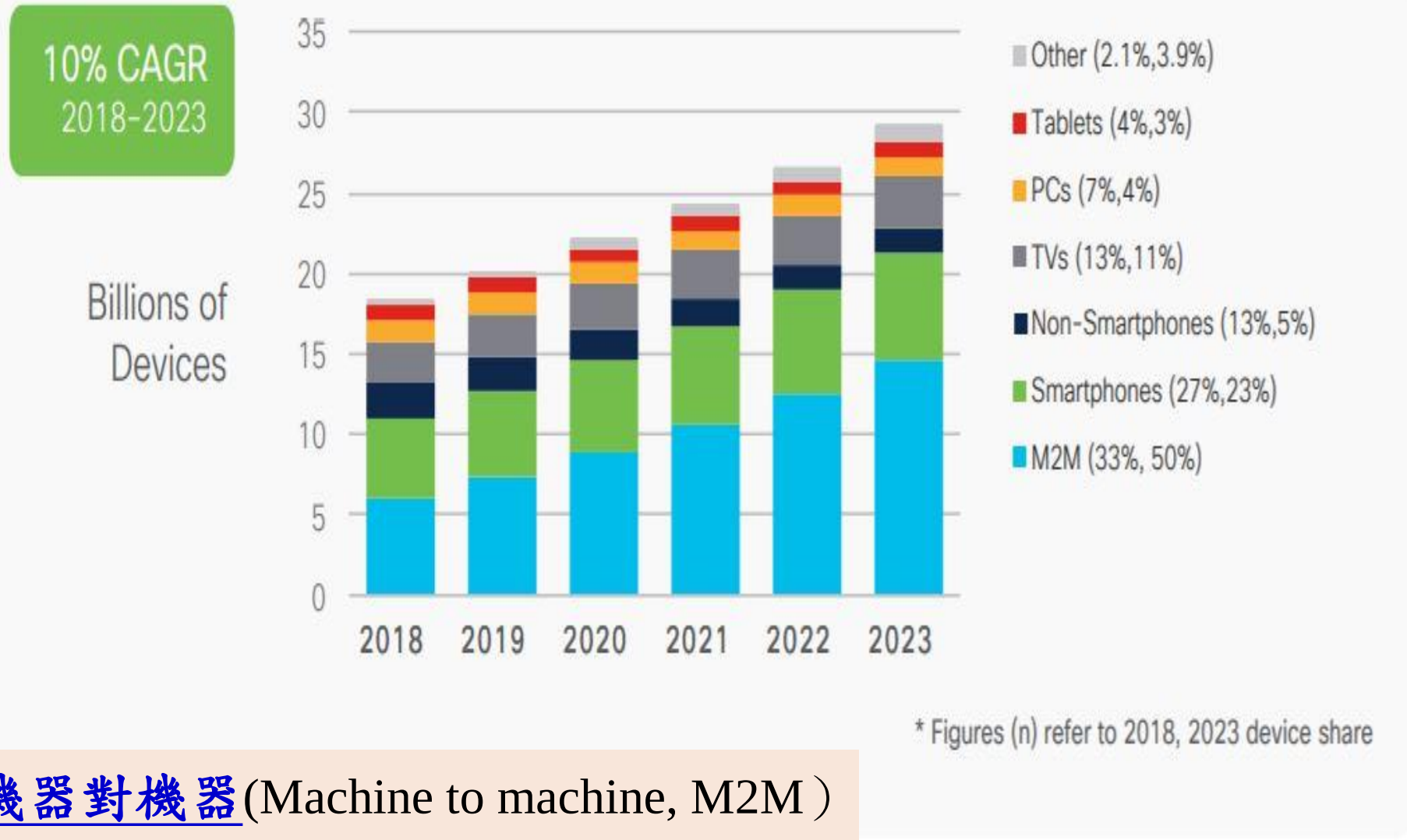
雲端儲存資料更為普遍



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

未來趨勢：機器對機器。。。



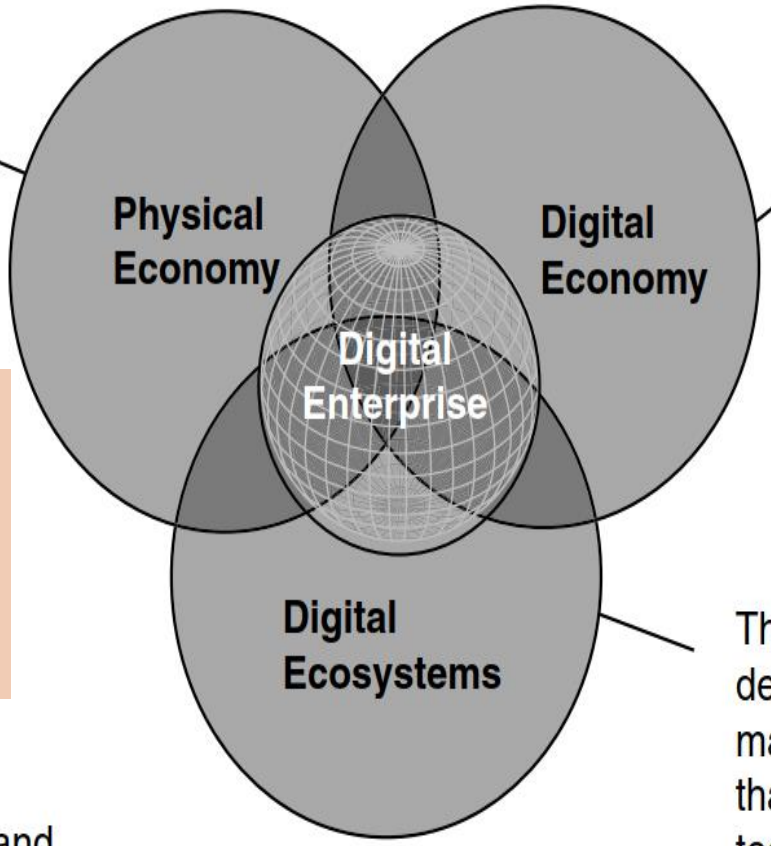
機器對機器(Machine to machine, M2M）機器裝置間在無需人爲干預下，直接透過網路溝通而自行完成任務。

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

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

Predictive analytics, data mining and big data myths, misconceptions and methods (2014)

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

商管學院

財金產業

1.風險控管(Risk Control)

→信用評等、信用卡盜刷、貸款審核與違約預警

2.**金融科技**(**Fintech**；Financial Technology)

→第三方支付單位(PayPal、Apple Pay、支付寶等)網路收款及付款服務，保障買賣雙方權益。

→網路銀行提供線上匯款、金融交易與投資理財功能(美國銀行、摩根、大通等)

圖片來源：http://www.sbs.ox.ac.uk/faculty-research/entrepreneurship-centre/events/fintech-founders-perspective
https://mattermark.com/sizing-the-fintech-opportunity/

- 人力資源、公司績效(如：業績、滿意度、升遷率、離職率等)。

→業務人員離職率(流動率、忠誠度、獎勵措施)

→延長營業時間與人力運用的配置

- 品管：

→以千萬筆資料測試積體電路電流的穩定性。

圖片來源：http://www.eazystock.com/blog/2015/09/01/4-ways-big-data-is-changing-how-companies-manage-inventory/

大數據與金融保險的發展

- 金融科技(Fintech)與保險科技(Insurtech)

→**金融科技**指技術帶來的金融創新，它能創造新的**模式**、**業務**、**流程**與**產品**，既可以包括前端產業也包含後臺技術。例如：**互聯網和移動支付**、**網路信貸**、**區塊鏈**。

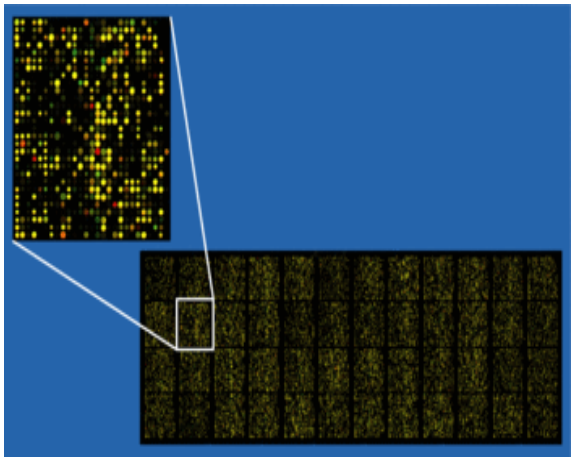
→**保險科技**是科技進步帶來的保險創新。無論是**產品**、**銷售通路**、**核保**、**理賠**、**後台作業**與**客服**等傳統價值鏈，都將被保險科技帶來的創新徹底顛覆。

醫學、生物科技學院

- 全民健保資料庫

→長期照護計畫政策與費用，可藉由死亡率、罹病率、失能率等相關數據估算。

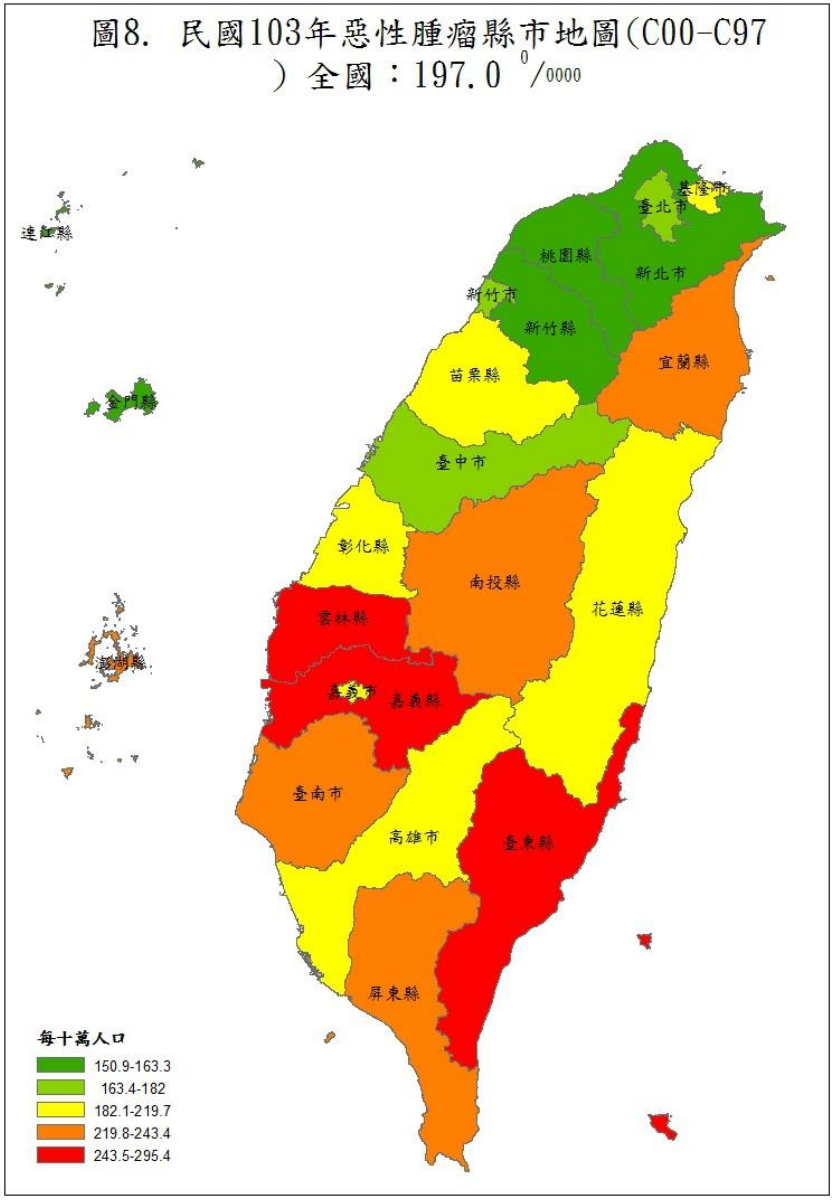
- 生物晶片：遺傳性疾病診斷、癌症診斷與預防
- 疾病地圖(Disease Mapping)

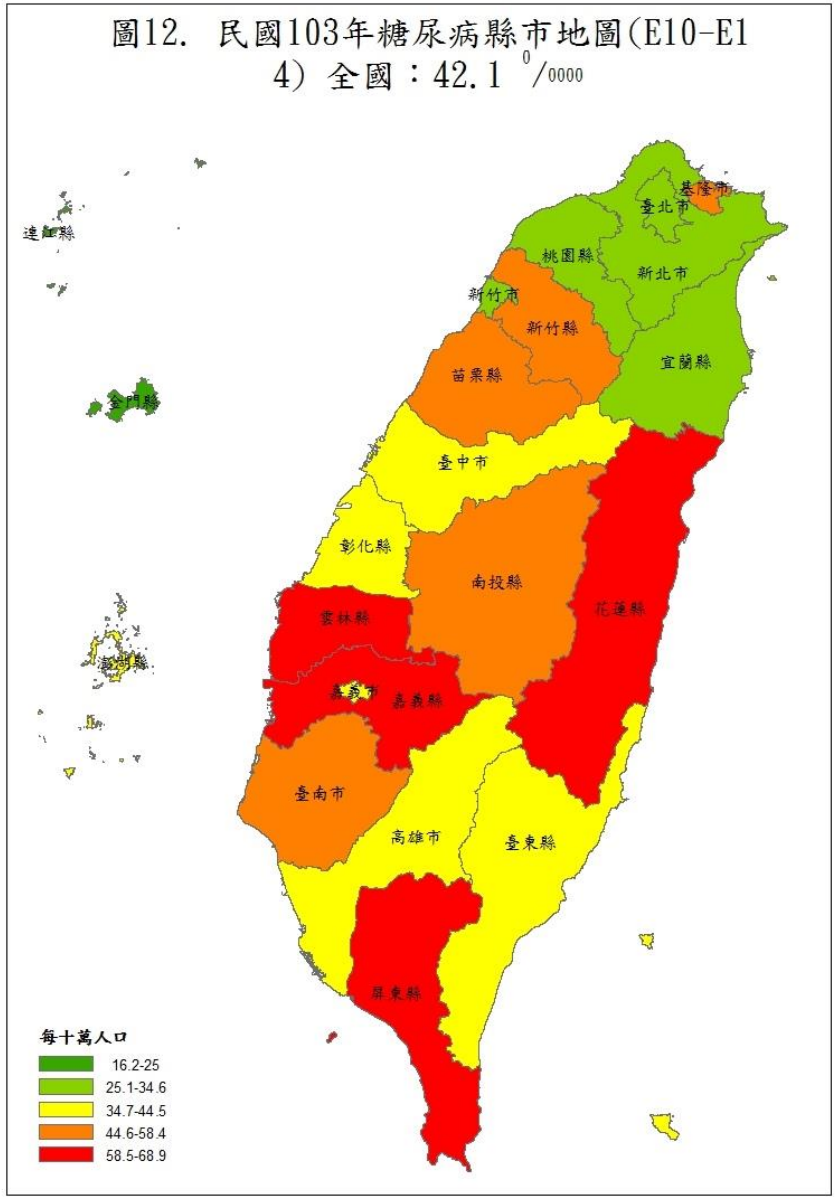


Zika Virus MAP

圖片來源：https://zh.wikipedia.org/wiki/DNA微陣列, http://id.mhsblogs.com/blogs/2016/02/04/zika-virus-what-you-need-to-know/

台灣2014年疾病地圖（十大死因）





來源：http://fd2016.pixnet.net/album/photo/21391582

運動與休閒學院

- 美國職棒MLB進行大數據分析

→從打擊率、打點、全壘打、盜壘、勝場數、救援成功等數據，作為球員薪水談判籌碼的參考。

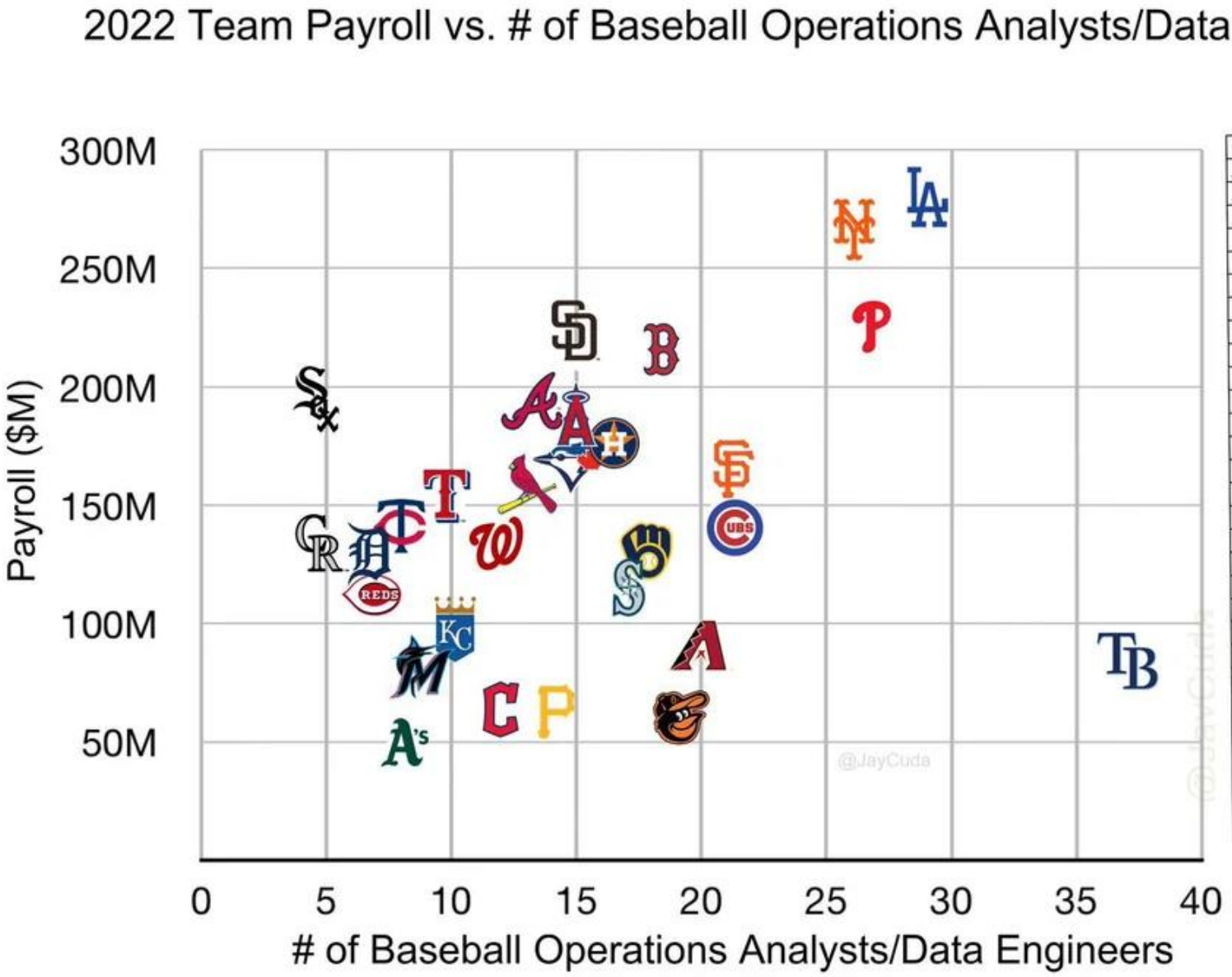
- 訓練運動選手

→分析肌力、耐力、爆發力、呼吸、步調與姿勢等特質，以提高體能與運動成績。

- 航空公司調整航班、航線。

圖片來源：
https://en.wikipedia.org/wiki/Uncle_Sam

美國職棒各隊薪資與資料分析師



Org.	Payroll ($M)	Analysts
A's	$49	8
Angels	$182	15
Astros	$174	16
Blue Jays	$171	15
Braves	$185	14
Brewers	$132	18
Cardinals	$158	13
Cubs	$146	21
D'backs	$93	20
Dodgers	$285	29
Giants	$158	21
Guardians	$60	12
Mariners	$105	17
Marlins	$80	9
Mets	$266	26
Nationals	$135	12
Orioles	$58	19
Padres	$212	15
Phillies	$224	27
Pirates	$62	14
Rangers	$152	10
Rays	$84	37
Red Sox	$212	18
Reds	$113	7
Rockies	$133	5
Royals	$95	10
Tigers	$130	7
Twins	$137	8
White Sox	$190	5
Yankees	$237	

https://www.reddit.com/r/chelseafc/comments/xqgk14/thought_folks_might_be_interested_in_this_from/?rdt=48367

奧運金牌運動員背後的科技秘器(Swim Analyzer)

來源：加拿大游泳選手於2012倫敦奧運

http://wired.tw/posts/bdma_8_1_london_olympics_2

什麼是 Sports Science 精準運動科學

使用現代化的科技，讓運動員擁有事半功倍的訓練效果、降低運動傷害、創造最佳的運動成績。

智慧化訓練

智慧桌球球拍、舉重使用的避震減噪地墊

選手選拔、訓練歷程紀錄、疲勞監控、心智訓練、營養控制。

貼心化照護

建立運動傷害監測系統，提供傷害預防策略。

情報化分析

戴資穎、周天成及王子維都有使用

開發戰情技術分析、進行戰情資料蒐集與分析。

https://2030.tw/article/digi-plus-2021-06-smart-recreation

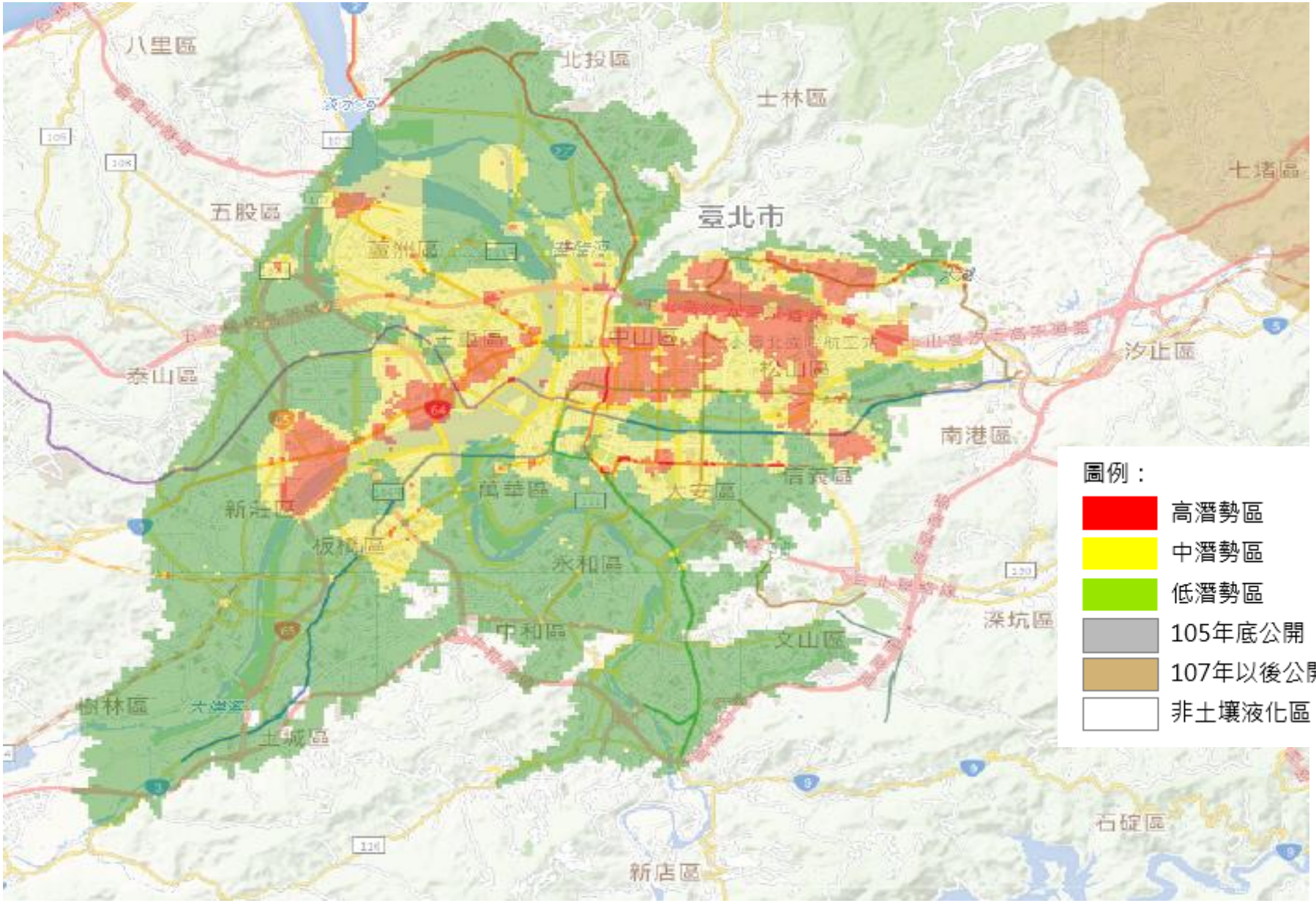
社會科學學院

- 預警性警務軟體PredPol找出犯罪熱點。
- 學校根據課程、授課教師與歷年選課資料，可預測修課人數、調整教室與時段。

- 日本311大地震救命大數據
- 預估房地產價格，作為課徵土地稅與房屋稅的參考。
- 高屏於88風災後開始「養水種電」，屏東縣於2013年獲選全球智慧城市大挑戰名單

圖片來源：
http://tech.sina.com.cn/d/2012-07-17/08017395390.shtml http://www.earthday.org.tw/column/greencitycase/5779

土壤液化潛勢查詢系統 - TGOS Maps



http://map.tgos.tw/TGOSimpleViewer2/Web/Map/TGOSimpleViewer_Map.aspx

智慧運輸系統(Intelligent Transportation System)

https://transport-curation.nat.gov.tw/museum-ITS2020/ITS.html

倫敦(2012)利用Big Data智慧網解決奧運交通混亂

http://wired.tw/2013/05/20/bdma_8_1_london-olympics/index.html#

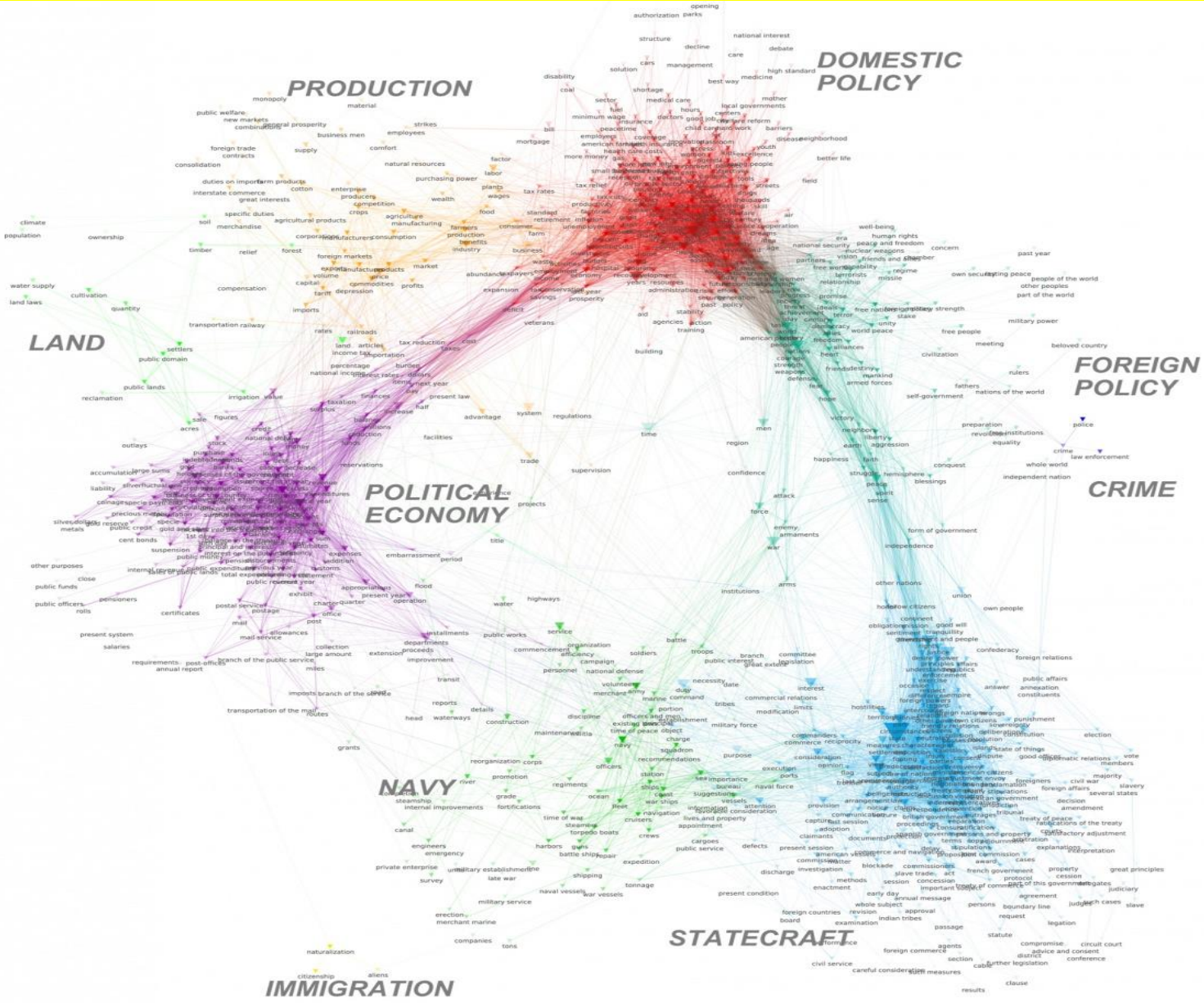
藝術人文學院

- 除文字分析，尚有音樂、繪畫等類型資料。

→耶魯大學數位人文實驗室進行「Robots Reading Vogue」的研究，分析Vogue雜誌的封面大小與位置，同時考量主題模型(Topic Modelling)分析。

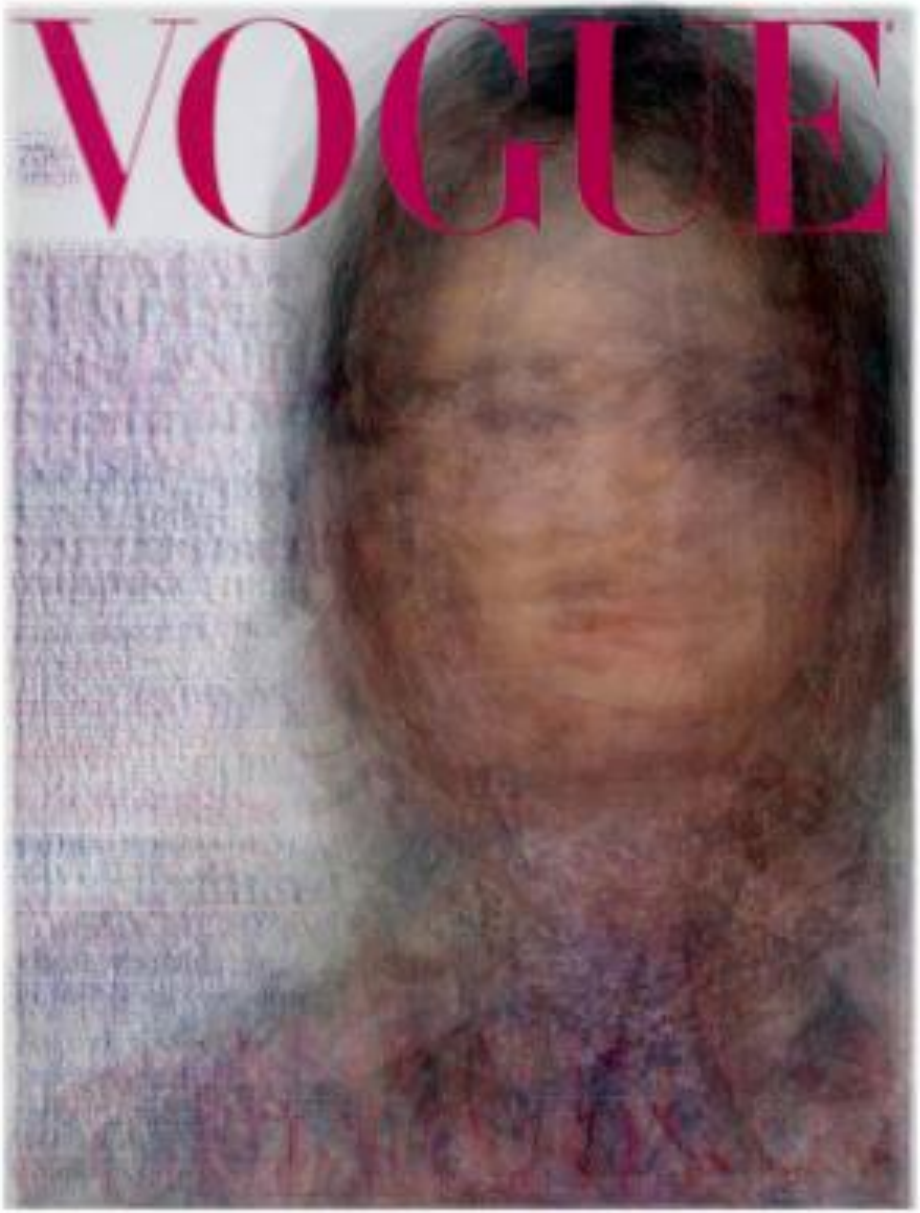
→Spotify等業者利用串流平台蒐集個人歌曲清單，或由物聯網的收聽紀錄找出歌手與作者的歡迎程度，以及收聽族群特徵。

美國總統國情咨文(State of Union;1790年至今)



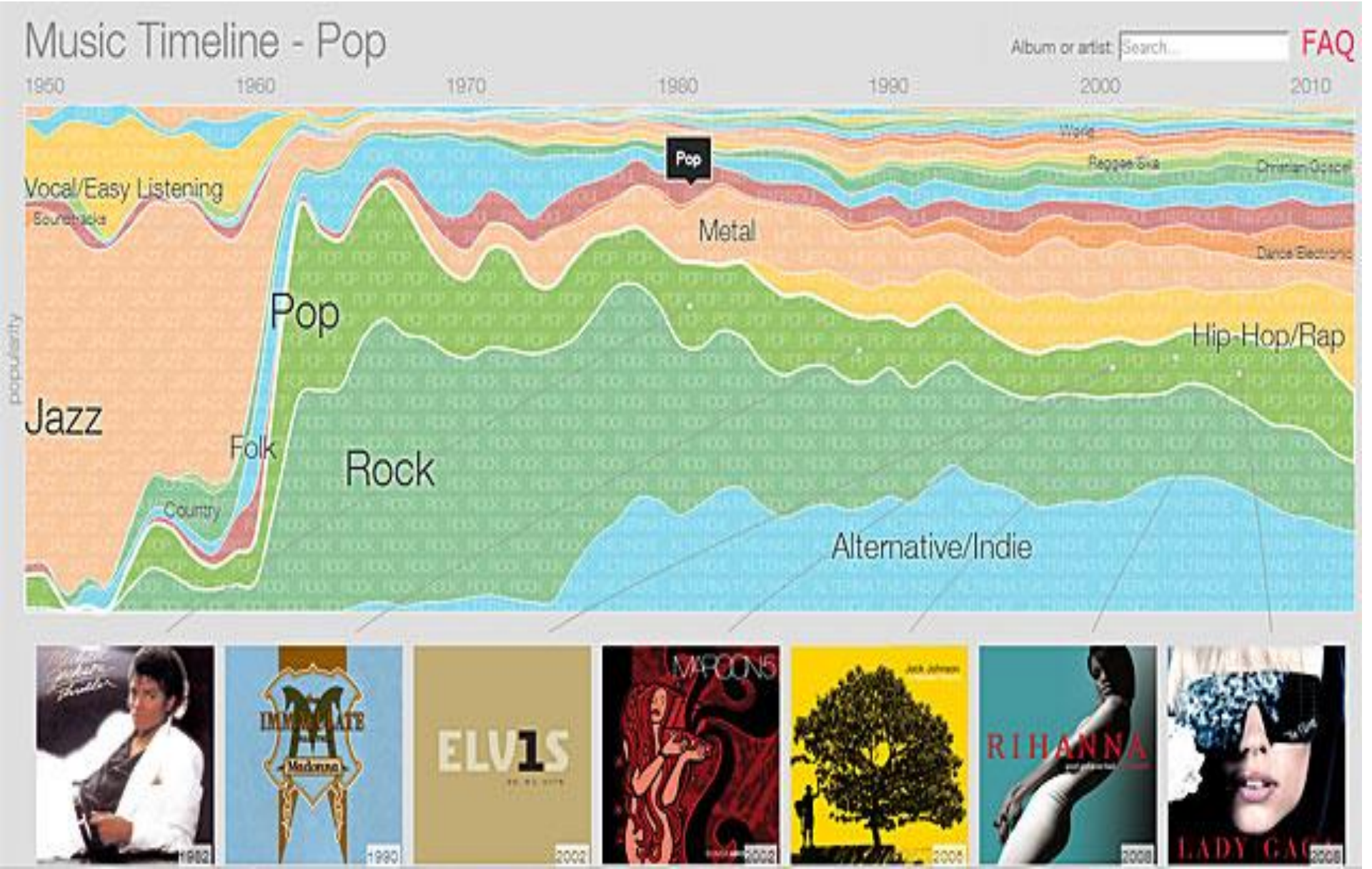
來源：https://sciencenode.org/feature/text-mining-strikes-gold-in-political-discourse.php

耶魯大學數位人文實驗室「Robots Reading Vogue」



圖片來源：美國耶魯大學Digital Humanities Lab

- 根據全球樂迷的喜好、播放次數、選購與評分，Google分析出世界音樂發展年表Music Timeline。



文字分析

- 非結構資料的類型廣泛，僅以文字為例示範資料分析。（文字採礦，Text Mining）

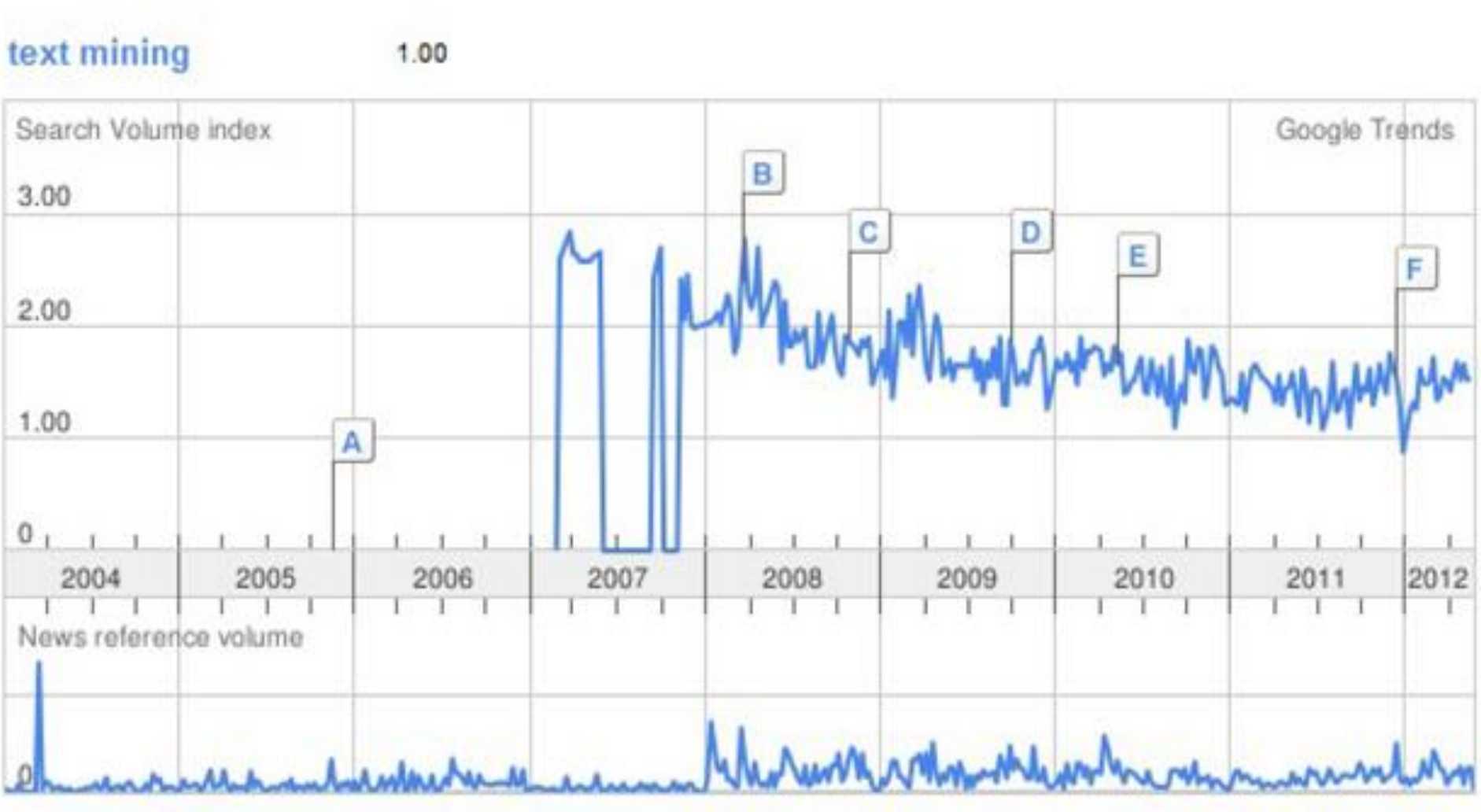
→R軟體有不少關於文字採礦的介紹及分析模組，較為重要者包括「twitteR」、「tm」、「wordcloud」。

→參考*Text Mining with R: Twitter Data Analysis*

- 文字分析非常需要領域知識，一知半解地代入模型和方法，容易得出似是而非的結果。

→例如：中文、英文的分析差異很大。

使用Google搜尋Text Mining趨勢圖



來源：*Text Mining and Analysis, Practical Methods, Examples and Case Studies (2013)*

量化非結構資料

- 非結構資料的分析，通常會藉由自然語言處理(Natural Language Processing；NLP)之類的統計、機器學習方法，先將文字轉化為具有一定格式的資料（亦即「**結構化**」）。

→接著可套用結構化資料的分析方法。

- 給予非結構化資料「結構」最為關鍵！

→然而結構化並無一定做法，往往與研究目的、研究素材有關。

文字分析歷史悠久

- Efron and Thisted (1976)

→估計莎士比亞(Shakespeare)的字彙總數；

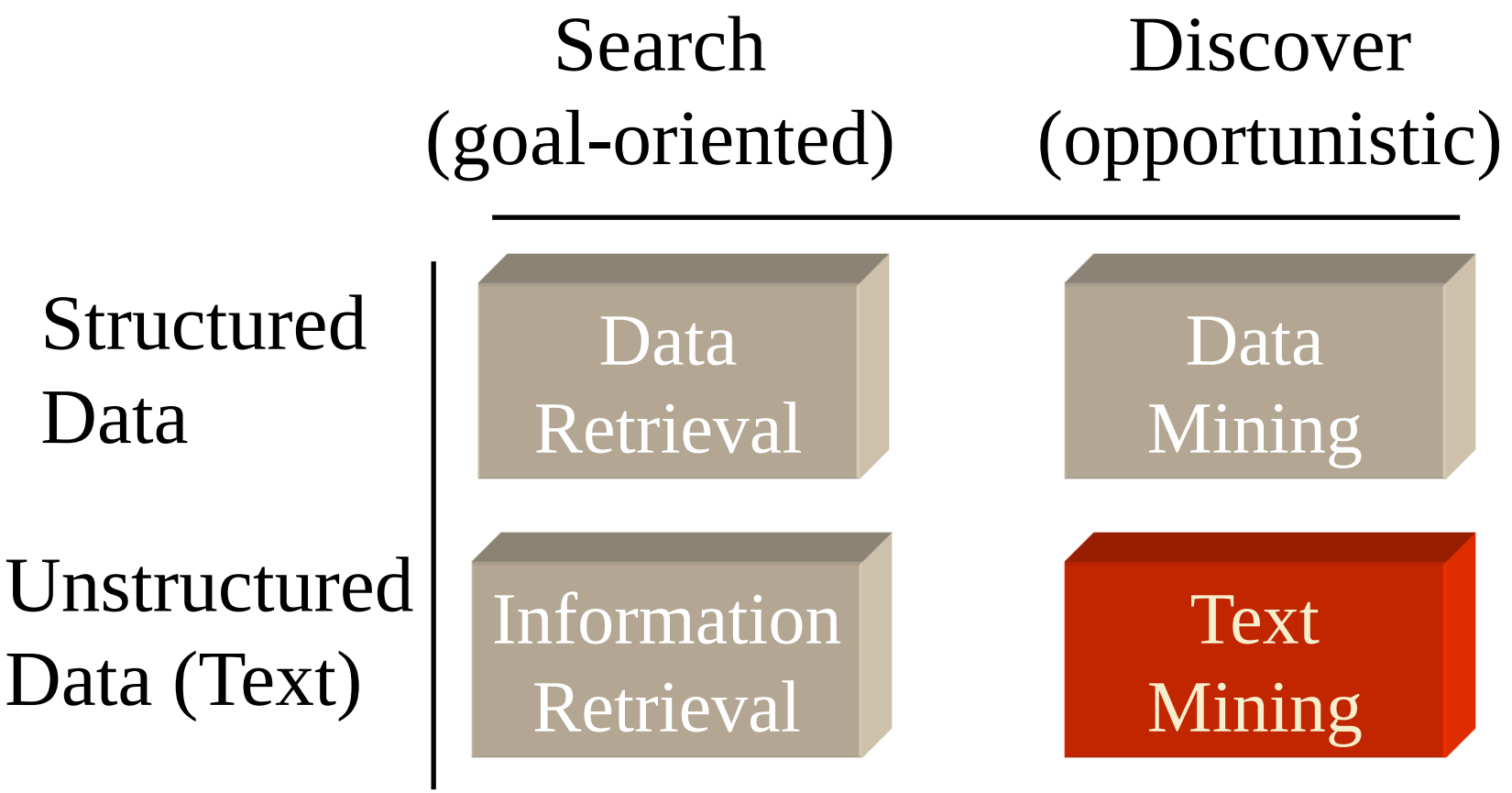
→用Poisson Process估計字彙，估計方法與Alan Turing有關(Turing's Estimate)；

→推論1985年在莎翁故居附近發現一首詩，應是莎士比亞所作(1987)。

文字資料的分析圖例

Text Analytics				
Search (Information Organization and Access)			Descriptive and Predictive Analysis (Discovering Trends, Patterns, and Modeling)	
Information Retrieval	Content Categorization	Ontology Management	Text Mining	Sentiment Analysis

“Search” versus “Discover”



文字資料的前置處理

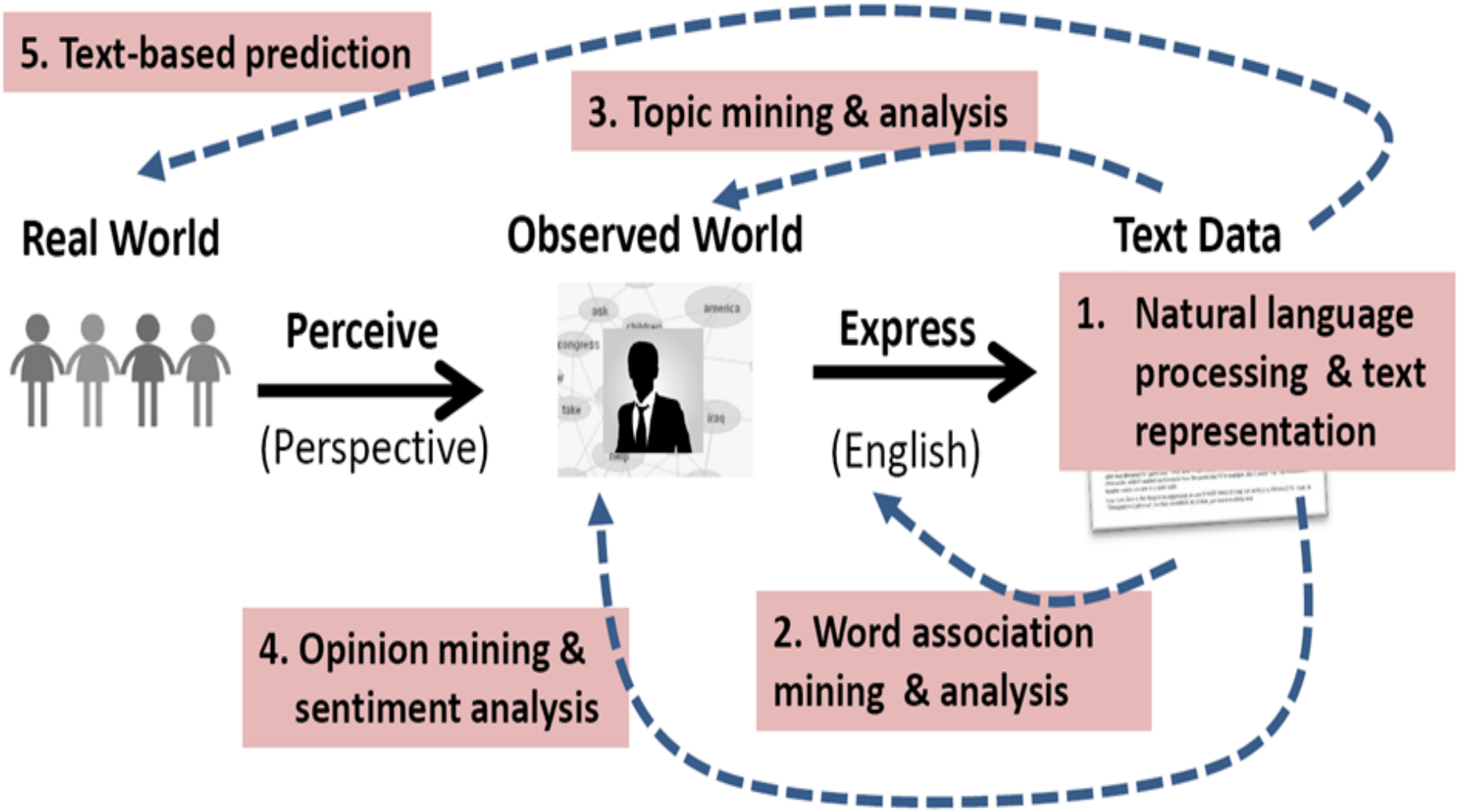
- 結構化文字資料之前，多半需要不少時間資料清理，尤其是英文之類由拼字構成的語言，標點及空格、時態、單複數都要先處理。
- 中文屬於方塊字，可將每個字視為有相同測量單位的觀察值，但仍應謹慎處理標點符號及斷句，確定詞彙及其字數（雙字詞？）。

→另外，方塊字也有同義字（異體字）、簡體字等之變化，和處理英文資料頗為類似。

Converting Text into Structured Data

- A huge amount of ***preprocessing*** is required to convert text.
  - Cleaning up 'dirty' texts
    - Remove mark-up tags from web documents, encrypted symbols such as emoticons/emoji's, extraneous strings such as "AHHHHHHHHHHHHHHHHHHHH"；Correct misspelled words..
  - Tokenization
    - Remove punctuations, normalizing upper/lower cases, etc.
  - Sentence splitting
  - Identifying multi-word expressions (e.g. "as well as", "radio wave") and Named Entities (e.g. "Allied Waste", "Super Mario Bros.")
- Adding other **linguistic** information
  - Parts-of-speech (e.g. noun, verb, adjective, adverb, preposition)
- Filtering non-significant/irrelevant words – to reduce dimensions
  - Filtering non-content words using a *stop-list* (e.g. "the", "a", "an", "and")
  - Combining tokens by stemming/lemmatizing or using synonyms
- Other NLP features/techniques, e.g. n-grams, syntax trees

文字採礦的分析思維


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

英文資料的前置分析

- 以英文資料為例，統計分析包含以下步驟：
- Data Collection
- Text Parsing and Transformation

→摘錄、清理、由NLP定義變數等，包括斷句、篩選相關資料段落、定義關鍵字詞。

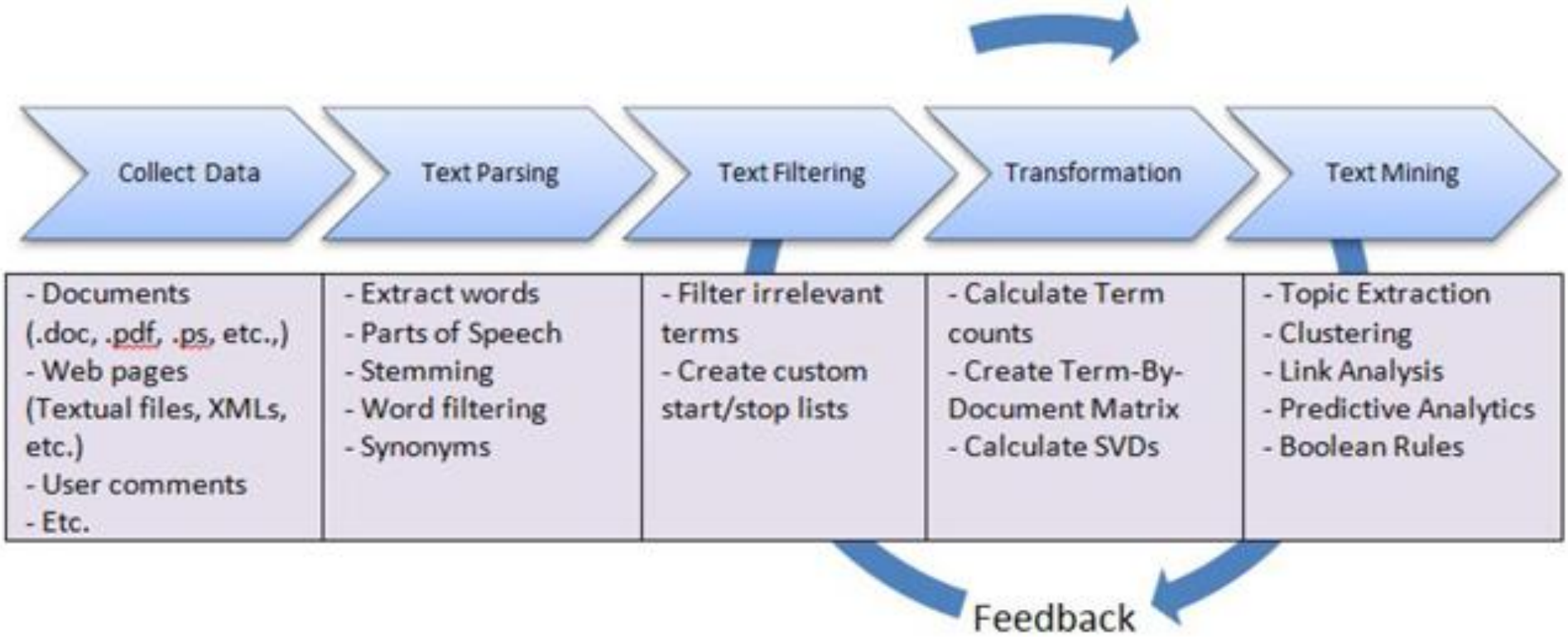
- Text Filtering

→挑選合適關鍵字詞。

- Text Mining

→Clustering, Classification, Association, and Link Analysis.

英文資料的處理流程範例

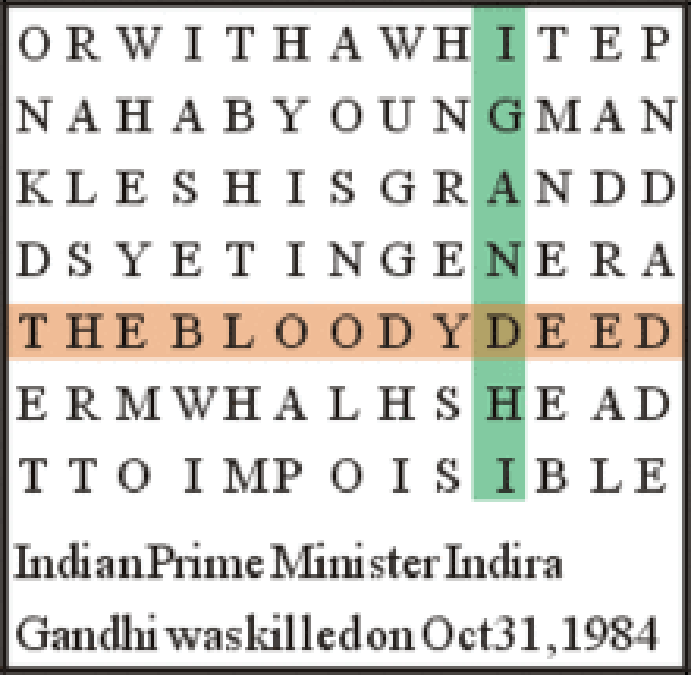


- Process is essentially a linear pipeline.
- Feedback from the results of Text Mining might affect earlier preprocessing (to Parsing, or even data collection)…

- Doc. 1: *I am an avid fan of this sport book. I like this book.*
- Doc. 2: *This book is a must for athletes and sportsmen.*
- Doc. 3: *This book tells how to command the sport.*

Term/Document	Document 1	Document 2	Document 3
the	0	0	1
I	2	0	0
am	1	0	0
avid	1	0	0
fan	1	0	0
this	2	1	1
book	2	1	1
athletes	0	1	0
sportsmen	0	1	0
sport	1	0	1
command	0	0	1
tells	0	0	1
for	0	1	0
how	0	0	1
love	1	0	0
an	1	0	0
of	1	0	0
is	0	1	0
a	0	1	0
must	0	1	0
and	0	1	0
to	0	0	1

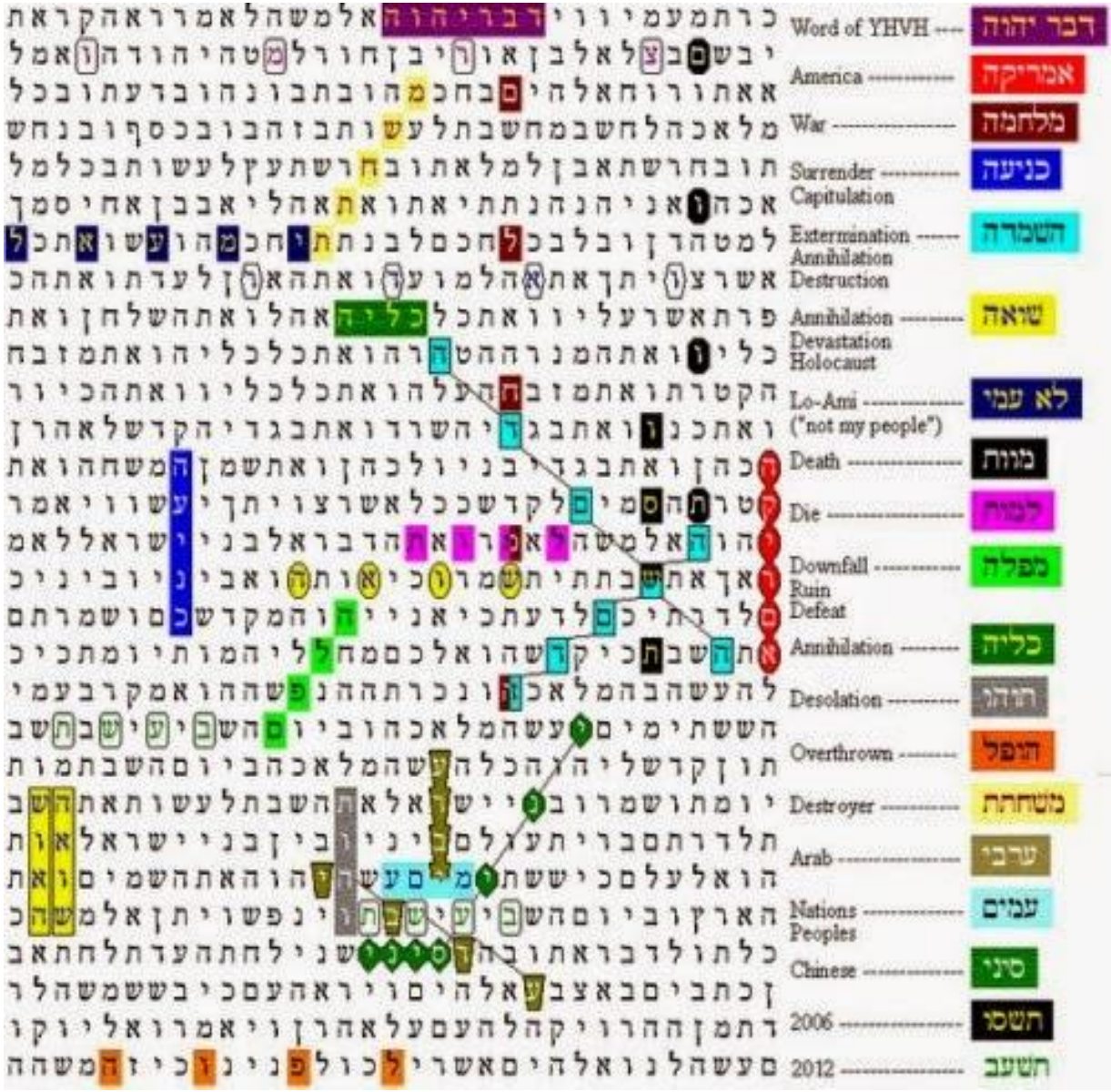
The Bible Code



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

聖經密碼（Torah密碼）

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



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 彭日榮 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

臺灣四大報的頭版頭條比較之一

臺灣四大報的頭版頭條比較之二

臺灣四大報的頭版頭條比較之三

臺灣四大報的頭版頭條比較之四