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

第一週：導論

http://csyue.nccu.edu.tw

資料科學—21世紀的明星產業

- 大數據的興起（IBM, 2010），激盪出新一波知識革命及產業。

→大數據似乎無所不在（無所不能），但其中也存在許多迷思。（**請同學舉例？**）

- 問題：資料科學為什麼應運而起，憑藉哪些特殊本領、創造哪些價值？

→資料科學(統計)專業人員有哪些必要技能與知識，如何因應大數據發展過程的震盪？

關於「巨量資料與統計分析」

- 除了基本觀念外，這門課也提供大數據分析練習，以及相關議題的討論。

→軟體需求：SQL及R(或Python)等軟體。

- 問題：為什麼大數據分析這麼熱門？

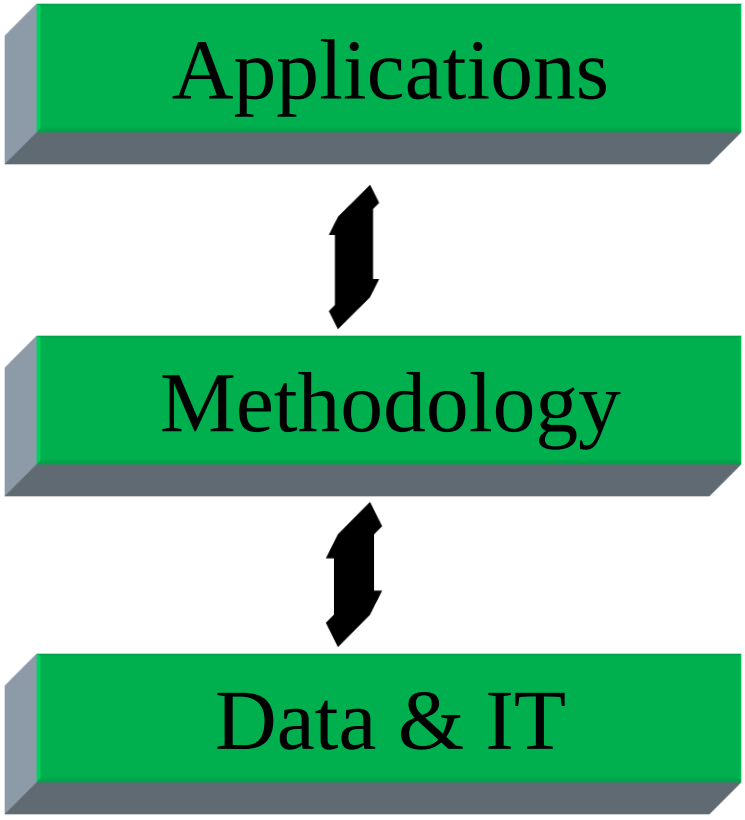
→與職涯發展（就讀科系）的關係：**領域知識！**

→大量資料的分析方法及理念：**統計！**

→需要熟悉的電腦技能及程式撰寫：**電腦！**

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！）



什麼是大數據？

https://www.zdnet.com/article/big-data-2018-cloud-storage-becomes-the-de-facto-data-lake/

大數據：大、快、雜、疑

- 大數據在2010年由IBM 所提出（**4V**）：

→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等資料

→真實性(Veracity)：資料品質(2014年加入)

註： 或可簡稱為「**大、快、雜、疑**」。另外，**資料分享與傳遞(Visible)**也是另一特性。

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

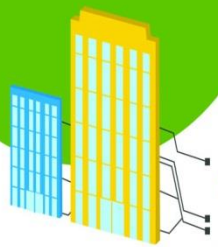
[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION



Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

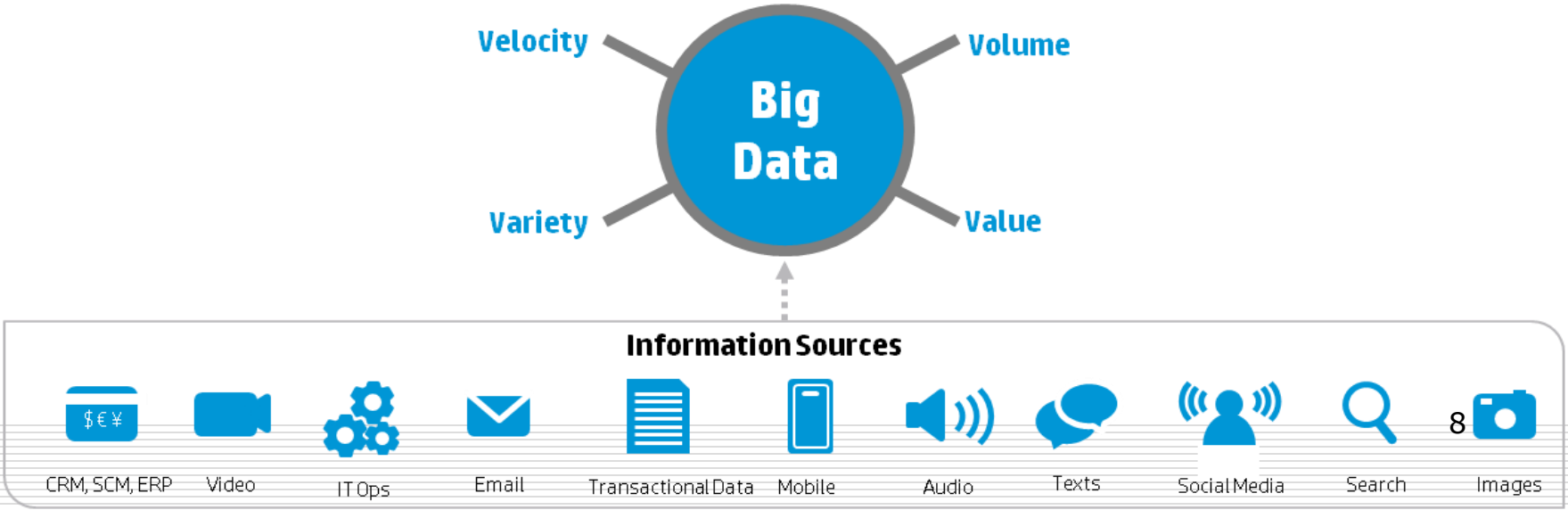
IBM

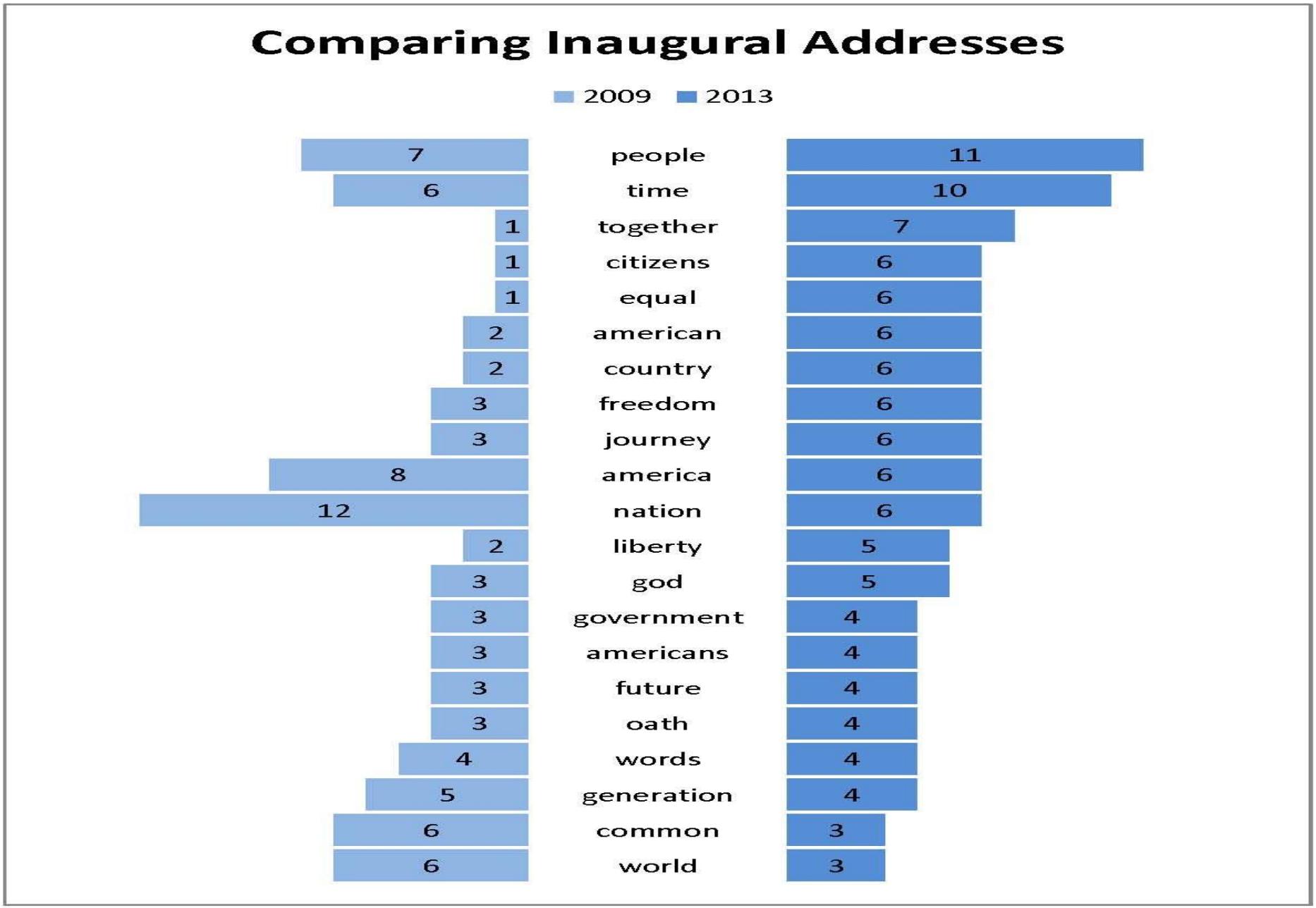
不同類型的大數據？

- 資訊記錄及交流趨於多樣化，分析面向及需求也愈來愈多采多姿。

→文學、音樂、美術等藝術如何分析？

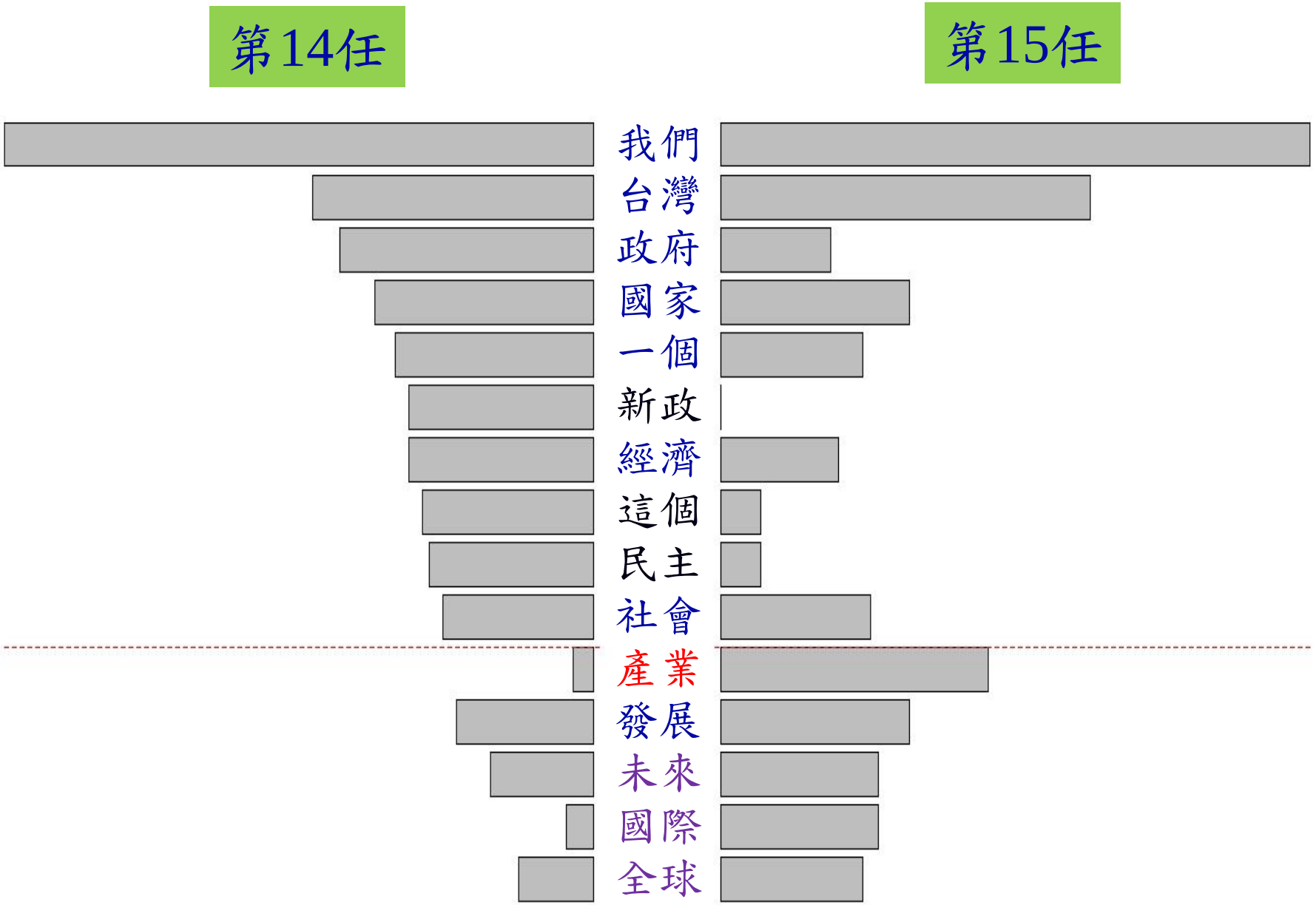
→分析典範？資料科學家的角色？



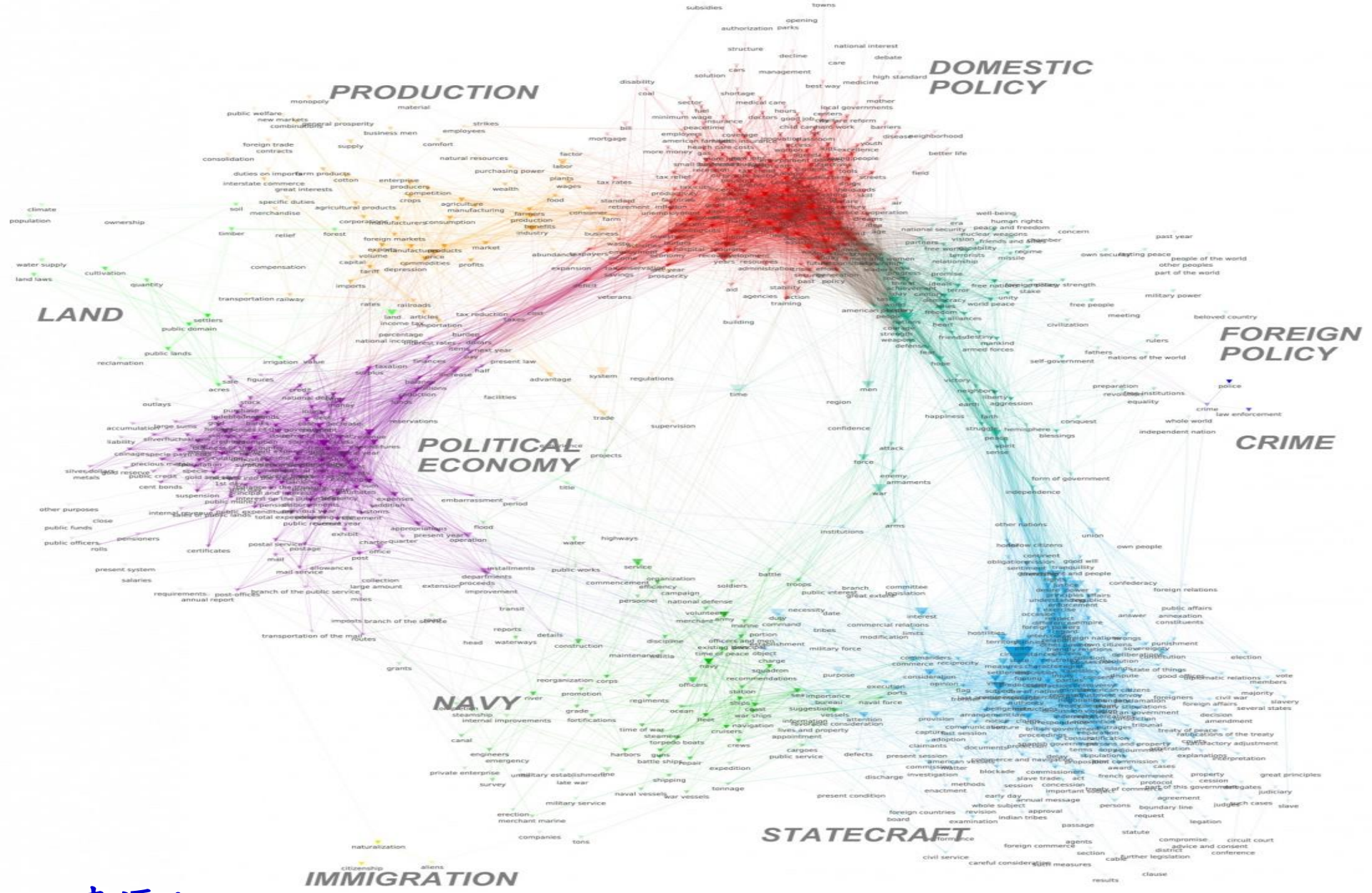


Analyzing the speech of President Obama (Textmining)

蔡英文總統就職演講稿常見雙字詞



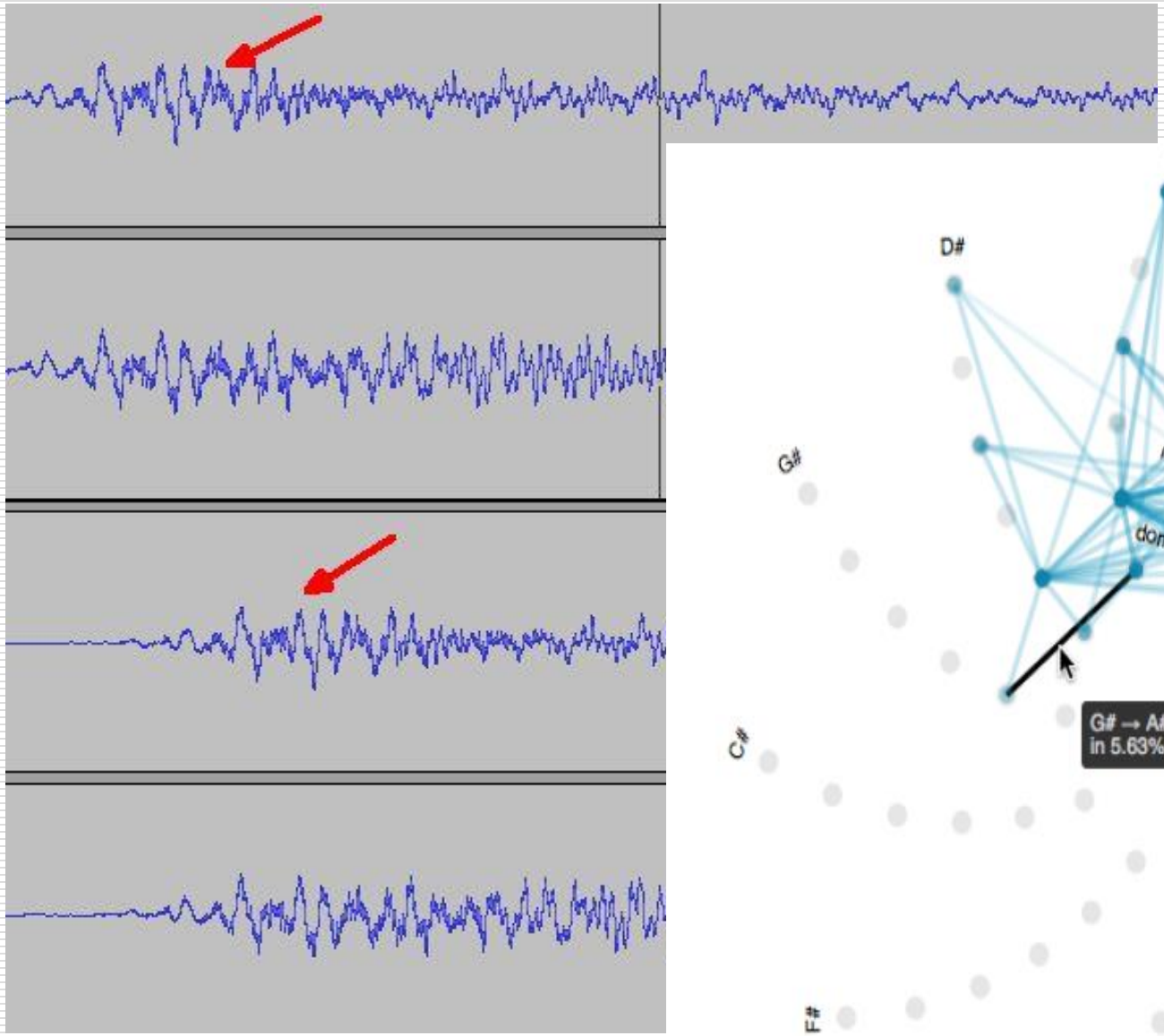
美國總統國情咨文(State of Union;1790年至今)

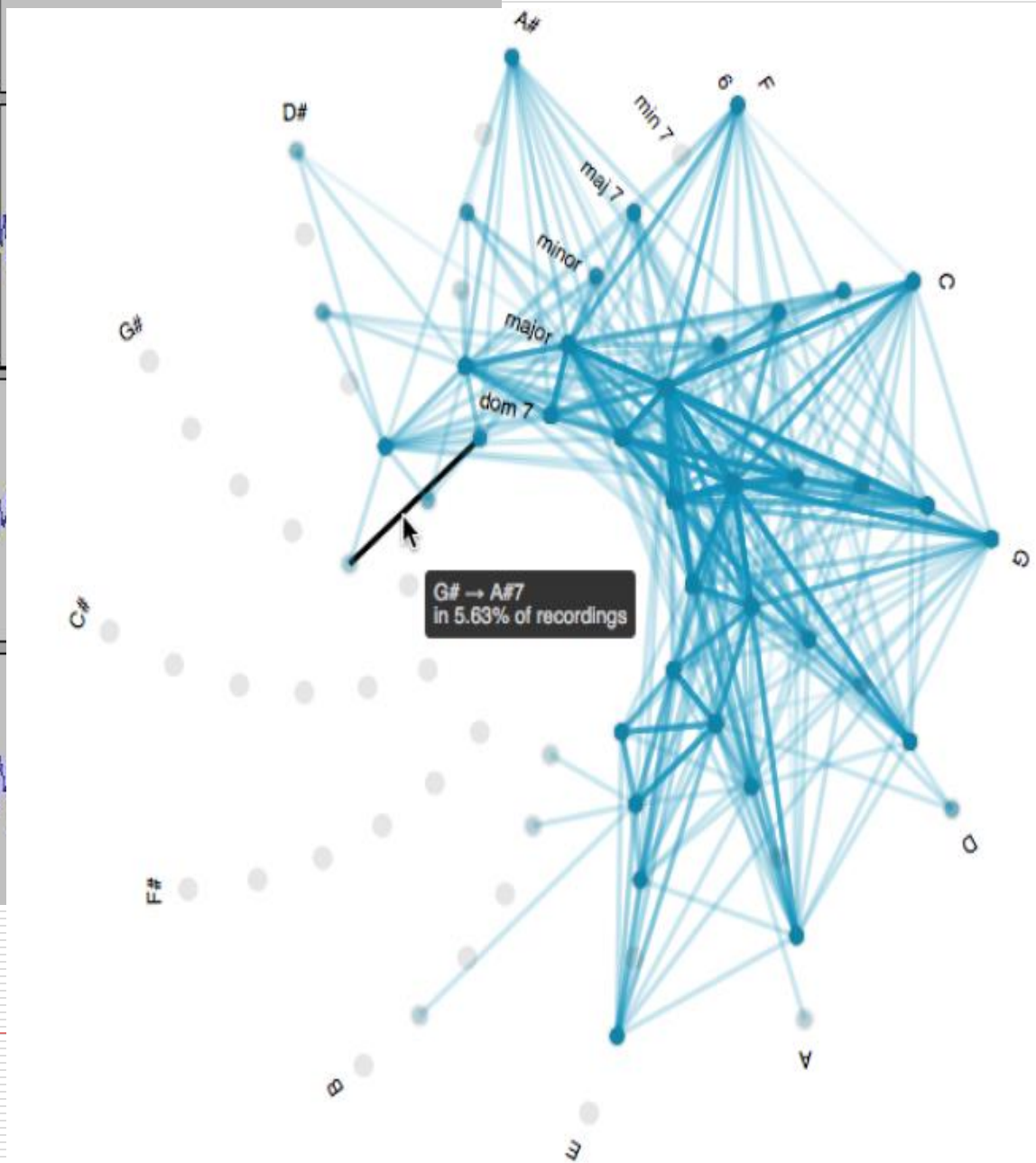


來源：https://sciencenode.org/feature/text-mining-strikes-gold-in-political-discourse.php

True/Fake

True/Fake





Music and Big Data

大數據的特性

http://www.abc.net.au/reslib/201204/r926508_9670000.jpg

大數據的幾個基本特性？

- 數量龐大外，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

http://mmdays.com/wp-content/uploads/2013/03/Bigdata.jpg

「樣本=母體」（n = All?）

- 大數據強調完整的資料，不只依賴由樣本推論出整個母體的特性。

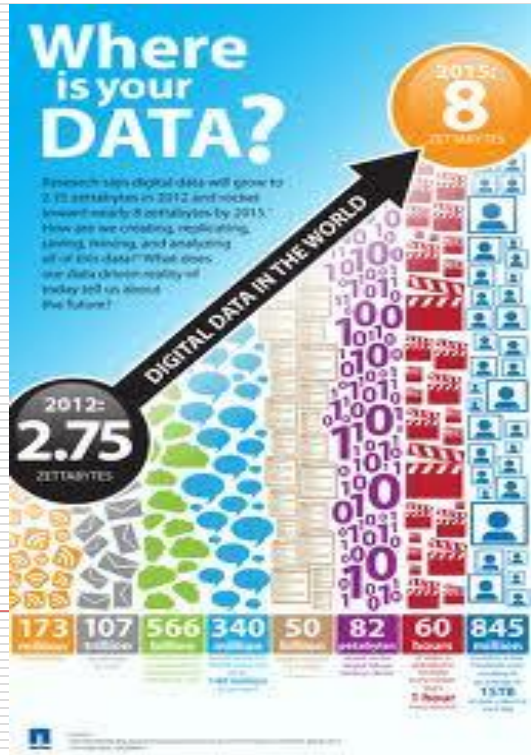
→即時反映資料特性（傳統普查）；

→去除抽樣造成的偏誤；

→樣本數不足（局部解析度）。

- 大數據也需有配套措施，像是更新、儲存、分析等問題。

註：資料蒐集者也會造成偏差！



http://www.smartercomputingblog.com/wp-content/uploads/2012/06/where-is-your-data.png

大數據也會有Sampling Bias!!

不精確且含有雜訊的資料

- 資料量愈多、愈有可能不正確，除此之外，大數據背後還隱藏其他類型的「雜亂」：

→結合不同源頭、不同類型的資料，產生的相容性問題。(Meta Analysis；整合型分析)

- 資料數量比資料品質重要！

→CPI官方統計vs.網頁五十萬項商品價格(雷曼破產後通貨緊縮)

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

相關性而非因果關係

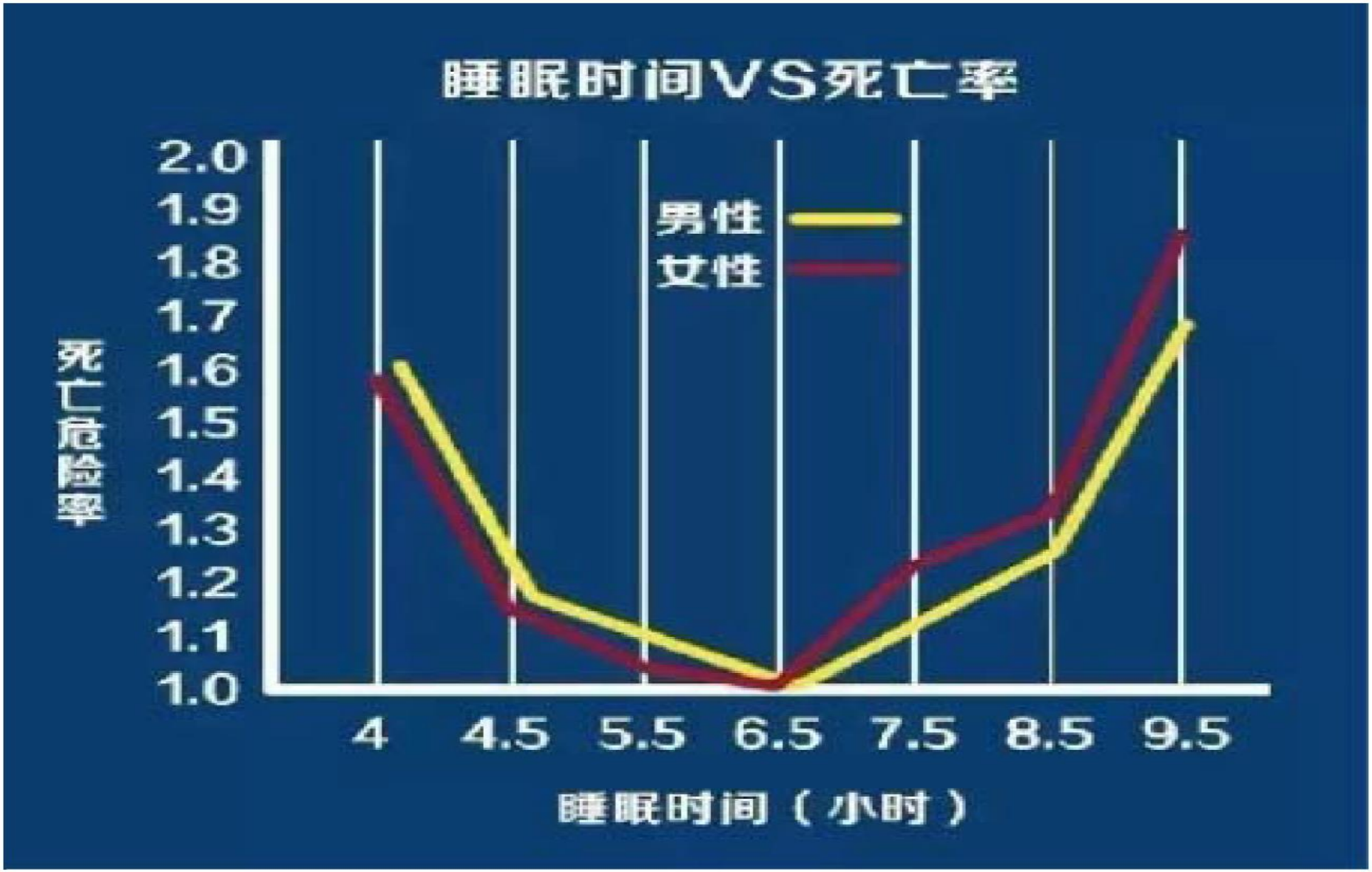
- 因果關係未必可由資料分析、模型中看出，像是小學生的拼字能力與腳丫大小，背後可能還有其他原因。

→多數車禍發生在時速40～60公里，僅有少數在車速超過100公里。開快車比較安全？

- 蒐集資料也決定了如何詮釋！

→中國2001年百歲人瑞調查，發現老大居多。

問題：睡得多、睡得少的死亡率比較高？



睡眠時間與死亡率的關聯圖表。取自生命時報

直覺、理性：什麼是事實？

- 除了樣本與母體的差異，資料品質及可測量性也是考量重點。

→以武漢肺炎的檢測為例，試劑必然存有偽陰性及偽陽性，如何判斷蒐集的訊息？

- 眼見為憑？

→當資料無法測量、或是觀察結果存有高度誤差時，又該如何因應？

	有病者	無病者
檢驗結果 陽性 +	真陽性 a	偽陽性 c
檢驗結果 陰性 -	偽陰性 b	真陰性 d

檢測的不確定性！

https://epaper.ntuh.gov.tw/health/201606/images/health_5_clip_image002.jpg

敏感性 $= \frac{a}{a+b}$ · 真陽性率：有病者檢驗結果為陽性的比率

特異性 $= \frac{d}{c+d}$ · 真陰性率：無病者檢驗結果為陰性的比率

- 型一誤差(Type-1 Error)等於P(拒絕H_0|H_0為真)
- 型二誤差(Type-2 Error)等於P(不拒絕H_0|H_0不為真)

直覺與理性判斷：醫學檢驗

- 假設某城市癌症盛行率0.1%，經過快篩試劑的測試，99%癌症患者被測出陽性反應，98%健康者測出陰性反應。

→某甲陽性反應、他是癌症患者的機率？

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

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

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？

→資料品質？(「Garbage in, garbage out」)

→哪些是重要（或必要）資訊？

→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？）

- 資料愈多愈好嗎？(Limit of big data?)

資料科學家（統計觀點）

- 分析大數據常屬於跨領域議題，一人、或一個領域很難面面俱到，需要講求團隊合作。
- 4V與統計間的關連強弱，依次真實性（**資料品質**）、多樣化（**不同來源及類型**）、大量化、快速化，可借助統計方法與理論。

→如何量化資料經常是分析大數據的第一個課題，需要集結各領域的知識及理念。

本課程設計與規劃

- 本課程從統計角度考量大數據的分析：

→以解決問題為導向(Problem-based)，根據問題需要攫取資料、定義變數，強調資料的多樣化、真實性（資料品質）。

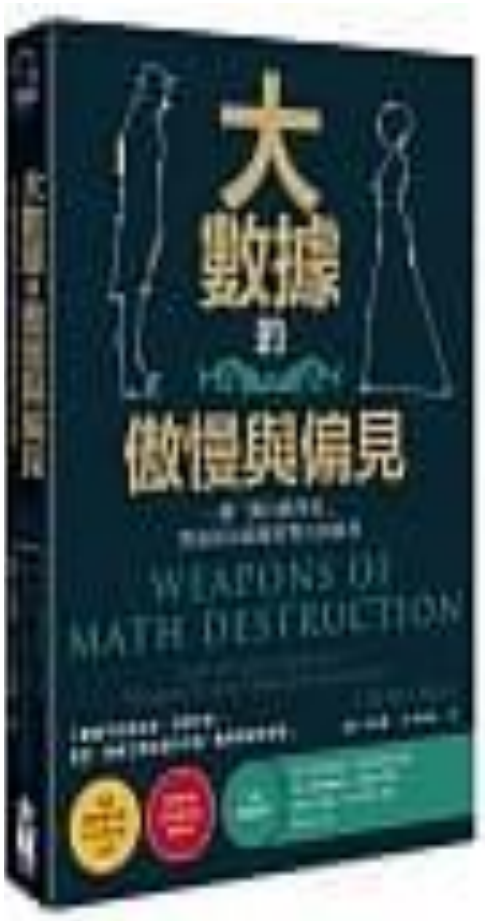
→根據資料屬性(Hard data vs. Soft data)介紹資料分析方法。

- 資料分析練習包括作業、期末報告，採兩人一組的方式。

大數據相關參考書


滄海圖書
大數據
知識經濟與實務應用
the Data Wisdom and Knowledge-based Economics
Big Data
余清祥、顏貝珊 著
大數據
知識經濟與實務應用
余清祥、顏貝珊 著
MA0373


大數據
BIG DATA


大數據的
傲慢與偏見
WEAPONS OF MATH DESTRUCTION

祝大家學期順利！

Q & A