活用統計數據！

二零一二年十二月三日
下午二時三十分至四時三十分

香港中央圖書館演講廳

香港大學統計及精算學系副教授　楊良河博士
政府統計處高級統計師　余司帆女士

香港統計學會
Hong Kong Statistical Society

香港特別行政區 政府統計處
Census and Statistics Department
Hong Kong Special Administrative Region

教育局
Education Bureau

活用統計數據 –

做個精明的數據使用者

活用統計數據

- 市民每天也有機會接觸到形形色色的統計數據。
- 作為一個精明的讀者，必須要學懂如何去解讀這些看似很日常生活化，但又很專門的資訊。

抽樣統計調查的代表性

抽樣統計調查的代表性

- 目標總體及抽樣框
- 抽樣方法
- 設計問題

例子一：目標總體及抽樣框清晰嗎？

「文職人士之XXX病誘因普查」：

1. 問卷調查報告

2. 調查方法

由XX會在商場內設置攤位向途經人士蒐集數據・・・並成功向1 322名文職人士蒐集數據・・・

哪些是文職人士？

例子二：目標總體及抽樣框清晰嗎？

- 工會向會員進行統計調查以量度香港的失業情況

目標總體及抽樣框－常見謬誤

- 抽樣框定義不清晰
- 調查範圍不完整

目標總體及抽樣框 – 官方統計的例子

在「綜合住戶統計調查」中，總體的定義是

(1) 全香港(水上除外)

(2) 居於家庭住戶內(不包括公共機構/社團院舍)

(3) 所有居港人口

「綜合住戶統計調查」的抽樣框：屋宇單位框

- 屋宇單位檔案庫
  - 永久性屋宇單位的地址
- 小區檔案庫
  - 未建設地區內的小區的紀錄
  - 每個小區約有10個屋宇單位
- 抽樣框管理
  - 根據行政紀錄不斷更新
  - 經本處人員出勤核實

例子一：坊間的抽樣方法

「青少年就業狀況」問卷調查

抽樣方法

問卷分別在三區街頭以<u>碰巧性抽樣(accidental sampling)</u>形式進行，共收回652份問卷，當中有641份有效問卷，成功率為98.31%・・・

例子二：坊間的抽樣方法

「香港青年聖誕食物消費調查」

抽樣方法

・・・他們亦採用雪球抽樣（Snowball Sampling）收集問卷。這是一種非概率抽樣方法，同學先選擇他們的同學、朋友及家人進行調查，調查之後，再請他們提供另外一些調查對象再進行調查。這一過程會繼續下去，形成一種滾雪球的效果。這方法可大大減低調查的成本，但由於是非概率抽樣，故不能計算抽樣誤差・・・

常見的非概率抽樣方法:

- 偶遇式/隨意抽樣
- 配額/定額抽樣
- 自發式電話調查
- 網上調查

隨機抽樣 ≠ 隨意抽樣

官方統計的抽樣方法的例子

「綜合住戶統計調查」

- 按所屬地區及屋宇單位類別分層
- 分層內使用等距複樣本抽樣法

例子：問題是否有引導性？

「XX大橋工程對香港的經濟效益」調查

1. XX大橋建成後香港前往某市只需30分鐘，較現時的交通工具節省逾一倍時間，你覺得效益大嗎？	大	不大	無意見
	…	…	…
2. 政府預計XX大橋啟用50年內，為本港帶來1,000億元經濟效益，創造20,000個就業職位，你認為興建XX大橋有助刺激經濟嗎？	能夠	不能	無意見
	…	…	…

引導性問題或選項不對稱

- 問卷設計可能影響結果
  - 引導性問題:影響受訪者傾向
  - 選項不對稱:受訪者較大可能給予正面/負面意見

調查結果的可靠性

例子：樣本數目足夠嗎？

調查指基層勞工生活質素偏低

ABC協會在今年四至六月，以抽樣問卷形式，訪問一百名在A區工作或生活的基層勞工，發現勞工生活質素普遍偏低，不少工人長期被僱主剝削．．．

樣本數目太少

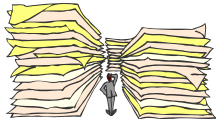
- 未能反映整體情況
- 特定群組的特點不能伸至整體

例子：回應率可以接受嗎？

XX工會在向會員發出近三萬份問卷後，只收回約1 700份問卷，回應率是6%。調查結果是．．．

回應率與統計調查結果的關係

- 回應率較樣本數目更為重要 。



- 請注意，潛在的偏差受以下因素影響：
  1. 回應率；以及
  2. 受訪者與非受訪者的回應差異。

回應率與統計調查結果的關係

假設隨機抽出1 000人作為樣本，詢問他們是否贊同某一項政府政策:

	回應率		
	70%	50%	30%
受訪者的確實人數	700	500	300
不回應的人數	300	500	700

回應率與統計調查結果的關係

	回應率		
	70%	50%	30%
受訪者的確實人數	700	500	300
不回應的人數	300	500	700
假設：在受訪者中，60%表示贊同；40%表示不贊同，即			
贊同的人數	420	300	180
不贊同的人數	280	200	120
統計調查所反映的多數意見	**贊同**	**贊同**	**贊同**

回應率與統計調查結果的關係

	回應率		
	70%	50%	30%
受訪者的確實人數	700	500	300
不回應的人數	300	500	700
假設：在受訪者中，60%表示贊同；40%表示不贊同，即			
贊同的人數	420	300	180
不贊同的人數	280	200	120
統計調查所反映的多數意見	**贊同**	**贊同**	**贊同**
情境1			
假設：在不回應的人士中，40%會贊同有關政策：			
贊同的人數	120	200	280
不贊同的人數	180	300	420
與受訪者的意見一併計算：			
贊同的人數	540	500	460
不贊同的人數	460	500	540
真正的多數意見	**贊同**	**沒有多數**	**不贊同**

回應率與統計調查結果的關係

	回應率		
	70%	50%	30%
受訪者的確實人數	700	500	300
不回應的人數	300	500	700
假設：在受訪者中，60%表示贊同；40%表示不贊同，即			
贊同的人數	420	300	180
不贊同的人數	280	200	120
統計調查所反映的多數意見	**贊同**	**贊同**	**贊同**
情境2			
假設：在不回應的人士中，20%會贊同有關政策：			
贊同的人數	60	100	140
不贊同的人數	240	400	560
與受訪者的意見一併計算：			
贊同的人數	480	400	320
不贊同的人數	520	600	680
真正的多數意見	**不贊同**	**不贊同**	**不贊同**

清晰的圖表展示

回應率與統計調查結果的關係

回應率	受訪者與非受訪者答案的分別	
	微不足道	分別甚大
高	理想	一般
低	尚可接受	誤差頗大

- 問題是，我們不會知道「分別」何在。
- 解決方法：盡量提高回應率 (即減低不回應率) 。

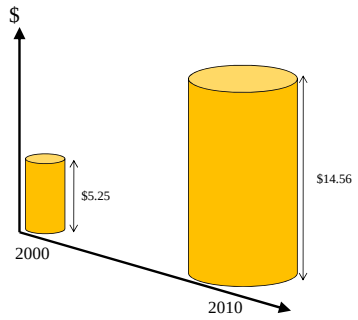
清晰的圖表展示

誤用圖表的典型例子

- 以不合適的維度展示數據
- 誇大比率

圖表展示要注意的地方

三維圖表所引致的錯覺



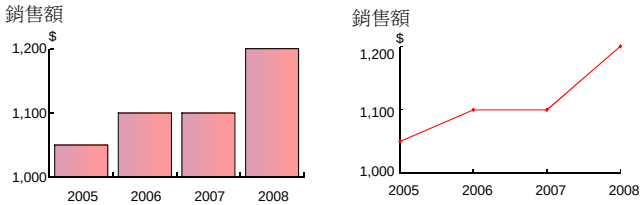
某一貨品的價格升幅:

數據增幅=177%

體積增幅>177%

例子：圖表展示清晰嗎？

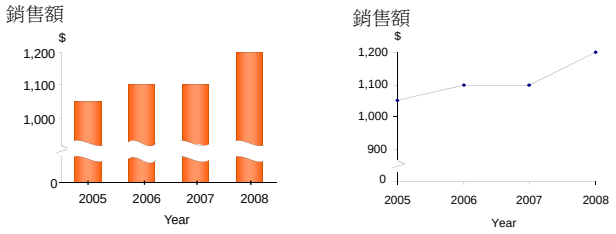
不恰當的縱軸尺度所引致的錯覺



上列圖表有何不妥的地方？

清晰的圖表展示

正確的方法（折斷統計圖）



總結

- 清晰的目標總體及良好的抽樣框
- 科學化抽樣設計
- 設計完善的問卷
- 足夠的樣本規模
- 高的回應率
- 清晰的圖表展示

政府統計處發布的官方統計數據

- http://www.censtatd.gov.hk/

聯絡方法

- **電郵: wsfyue@censtatd.gov.hk**
- **電話: 2582 4955**

政府統計處

- 一般查詢 : **2582 4807**
- 政府統計處 刊物: **2582 3025**

統計數據的發展

二零一二年十二月三日

楊良河博士
香港大學統計及精算學系

香港統計學會 Hong Kong Statistical Society

香港特別行政區 政府統計處 Census and Statistics Department Hong Kong Special Administrative Region

教育局 Education Bureau

統計數據的發展

- 遠古時代：原始人以樹刻計算家畜及財產數量

- 隨著文字的誕生，人類就開始整理數據的工作。歷史記載，世界各地的歷代王朝都有官員負責搜集國民統計數據。
  - 早在公元前2000年的夏朝，中國就進行人口統計。
  - 周朝為了管理統計的工作，設立了司書職位。

統計數據的發展

- 在春秋戰國時代《管子》一書的第24篇《問》中，記載了65個涉及國家的各方面問題：
  - 人之開田而耕者幾何家？
  - 問獨夫寡婦孤寡疾病者幾何人也？
  - 問一民有幾年之食也？

- 漢高祖劉邦認為統計很重要，任命宰相負責管理統計數據。

- 聖經舊約的民數記亦紀錄了約公元前1500年的人口普查(census)的數據，並與及指示摩西進行有關能出去打仗的以色列人數普查。

統計數據的發展

- 統計(statistics)一詞
  - 源於拉丁文的status
  - 意思是各種現象的狀態和狀況
- 最早使用「統計」一詞
  - 由18世紀中的德國學者-阿亨華爾(Gottfried Achenwall)把拉丁文status引申為德語“statistika”（統計學）
  - 意思是主要用文字來記述國家應注意的統計事項的學問，後來傳入英國被譯為statistics。
- 現今的統計工作是人們利用各種科學的統計方法，搜集、整理、分析和提供統計數據工作的總稱。

統計數據的搜集 (Data Collection)

- 普查 (census)
- 抽樣調查 (sample survey)
- 觀察性研究 (observational study)
  - 在不指定任何治療或情況下，研究人員觀察或詢問參加者的意見，行為或表現。
  - 問卷調查只是其中一類的觀察性研究。
- 實驗設計 (experimental design)
  - 隨機分配參加者去參與一個指定治療或情況。

統計數據的搜集 (Data Collection)

- 例：研究人員要比較兩個減肥瘦身方法的成效：
  - 運動
  - 節食/控制飲食

- 如何利用觀察性研究搜集統計數據？
- 如何利用實驗設計搜集統計數據？

Does caffeine affect heart rate?

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Effects of Caffeine on Plasma Renin Activity, Catecholamines and Blood Pressure

David Robertson, M.D., Jürgen C. Frölich, M.D., R. Keith Carr, J. Throck Watson, Ph.D., John W. Hollifield, M.D., David G. Shand, M.D., and John A. Oates, M.D.

N Engl J Med 1978; 298:181-186 | January 26, 1978 | DOI: 10.1056/NEJM197801262980403

Abstract

Using a double-blind, randomized, crossover protocol, we studied the effect of a single dose of oral caffeine on plasma renin activity, catecholamines and cardiovascular control in nine healthy, young, non-coffee drinkers maintained in sodium balance throughout the study period. Caffeine (250 mg) or placebo was administered in a methylxanthine-free beverage to overnight-fasted supine subjects who had had no coffee, tea or cola in the previous three weeks.

Caffeine increased plasma renin activity by 57 per cent, plasma norepinephrine by 75 per cent and plasma epinephrine by 207 per cent. Urinary normetanephrine and metanephrine were increased 52 per cent and 100 per cent respectively. Mean blood pressure rose 14/10 mm Hg one hour after caffeine ingestion. There was a slight fall and then a rise in heart rate.

Plasma caffeine levels were usually maximal one hour after ingestion but there was considerable inter-individual variation. A 20 per cent increase in respiratory rate correlated well with plasma caffeine levels.

Under the conditions of study caffeine was a potent stimulator of plasma renin activity and adrenomedullary secretion. Whether habitual ingestion has similar effects remains to be determined. (N Engl J Med 298:181–186, 1978)

Single-blinded experiment 單盲實驗

Single-blinded experiment 單盲實驗

實驗時間：咖啡因會否影響心跳？

- 隨機抽出7位的現場觀眾
- 其中1位決定把含有咖啡因的可樂倒在紅杯/白杯內，並把紅色/白色咭紙放入封面寫著咖啡因的公文袋。

實驗時間：咖啡因會否影響心跳？

- 其餘6位的現場觀眾，會先量度每人的心跳

- 之後每人飲一杯可樂，然後再量度每人的心跳

實驗時間：咖啡因會否影響心跳？

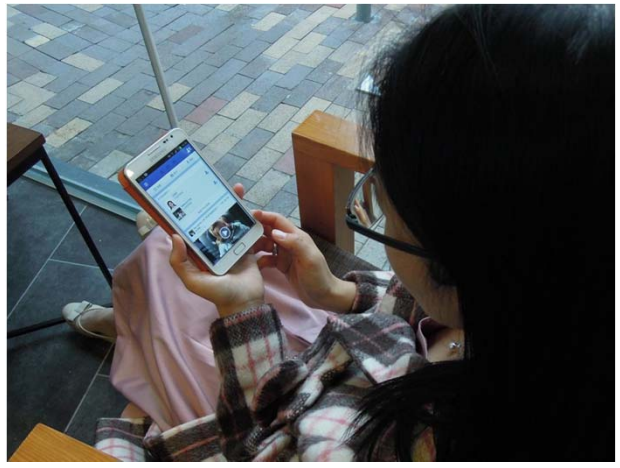
- 統計數據：

	未飲時的心跳	飲了後的心跳
紅杯1		
紅杯2		
紅杯3		
白杯1		
白杯2		
白杯3		

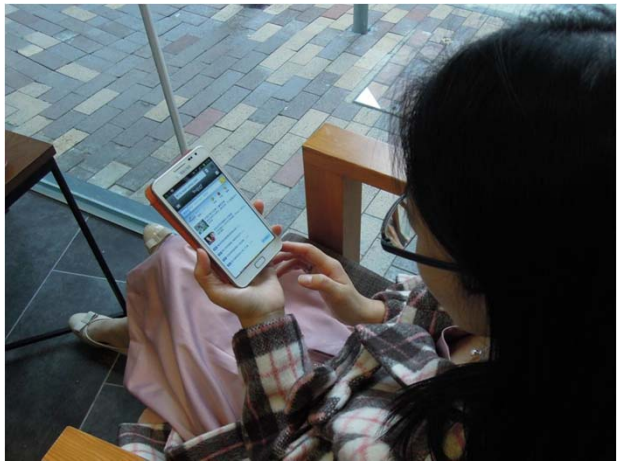
在某日的星巴克裡...

看見一位年輕的女人，手拿著Note...

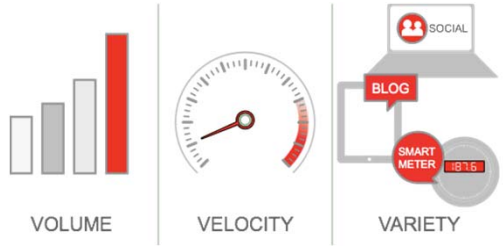
她正在看facebook...



1小時後，她在上網...



3Vs in Big Data



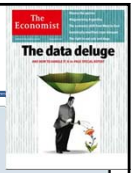
統計數據的近年發展

- 20世紀中，人類發明了電腦。經過短短幾十年，網絡的發展和電腦的速度及儲存能力的迅速提高，令超大規模的數據(Big Data)的不斷出現。

Google Data Center

Volume in Big Data

Economist, 25 Feb 2010



Data inflation

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

Source: *The Economist*

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

http://www.economist.com/node/15579717

Volume in Big Data

- 最近由IDC「數碼宇宙」研究預測，數據量增長高速， 於2011年數據量將達1.8ZB；企業於未來10年將需要管理50倍的數據。

1.8ZB數據 = ?

- 1.8萬億GB
- 每位香港人連續於107萬年內，每分鐘發送3條微博訊息
- 超過2,000億套高清電影（每套長度為兩小時）(一個人24小時全天候無間斷地把這些電影全部看完需要4,700萬年。)
- 把這些資料量全都裝在32GB版本的iPad裡，這些iPad可以
  - 建起一堵比中國萬里長城平均高出兩倍的iPad長城
  - 堆出一座比富士山高25倍的山

平均每天6100人次到急症室就診

Table 4.4 Number of Accident and Emergency Attendances 2010/11
表 4.4 二零一零/一一年度急症室就診人次

Hospital 醫院	Number of Accident and Emergency attendances 急症就診人次		
	First 首次就診	Follow-up 覆診	Total 合計
Alice Ho Miu Ling Nethersole Hospital	125,656	282	125,938
Caritas Medical Centre	131,307	3,173	134,480
Kwong Wah Hospital	141,841	7,661	149,502
North District Hospital	111,259	4,471	115,730
Pamela Youde Nethersole Eastern Hospital	145,789	9,085	154,874
Pok Oi Hospital	121,682	3,767	125,449
Prince of Wales Hospital	150,638	279	150,917
Princess Margaret Hospital	139,933	8,074	148,007
Queen Elizabeth Hospital	200,695	11,935	212,630
Queen Mary Hospital	124,787	3,611	128,398
Ruttonjee and Tang Shiu Kin Hospitals	84,406	2,285	86,691
St. John Hospital	11,046	0	11,046
Tseung Kwan O Hospital	114,712	4,205	118,917
Tuen Mun Hospital	226,718	6,247	232,965
United Christian Hospital	194,847	3,459	198,306
Yan Chai Hospital	135,932	7,467	143,399
Total 合計	2,161,248	76,001	2,237,249

來源：醫管局統計年報2010-2011

Velocity in Big Data

- Velocity是指資料增加的速度越來越快，諸如行動運算(mobile computing)、社交網路(social network)的風行，使得資料增加的速度比傳統的企業應用程式來得快很多，一旦資料增生速度越快，資料處理、分析的速度也就得跟上

Velocity in Big Data

- 高頻交易 (high frequency trading)

Variety in Big Data

- Variety則是指資料的多樣性
- 我們現在上網不是只看看資訊，同時我們不斷在產出資料：
  - 貼照片
  - 貼影片
  - 這裏按讚、那裏寫個幾句
- 另一方面，各式各樣的監控器、感應器也不停地產出機器資訊，資料的型式已不像過去那麼單純了。

Big Data 數字淘金

- 尿片、啤酒、星期五，這三個名詞相互之間有什麼關係？

- 直至90年代的某一天，一位威名百貨(Walmart, 全美最大零售業者)員工在分析營業數據時偶然發現：
  - 尿片和啤酒竟然常常被放在同一個購物籃中，而且大部分在星期五。

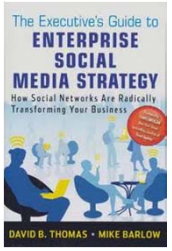
Big Data 數字淘金

- 經過分析後，威名百貨發現，原來美國婦女通常會在星期五請先生下班後替孩子買尿片再回家，而先生在買尿片之餘也想順手買自己晚上要喝的啤酒，這樣的消費習性已經成為常態。
- 威名百貨發現這件事後，改變了貨架的擺放位置：
  - 把啤酒和尿片擺在一起
  - 甚至把比較貴的尿片放在啤酒旁邊（很多男人不看尿片價錢）
- 成功增加了這兩項商品的銷售金額。

另一個可能的原因

Big Data 數字淘金

- A story from the book…
- I'm standing in the queue at Hotel A in Las Vegas, waiting to check in.
- Normally the check-in queues there move pretty fast, but today they aren't.
- Something is clearly wrong at the check-in counter, but no one is telling us anything.



Big Data 數字淘金

- I check Twitter on my iPhone and I can see that people are tweeting from the queue I'm standing in.
- Then I see a tweet from someone at the Hotel Z, a couple of blocks down the Strip. The tweet says something like "Hey, if you're tired of standing on line at Hotel A, come on over to Hotel Z. We'll set you up with a room---at 50% off the regular price."
- Then I see a couple of people pick up their bags and head out for the taxi stand. Wow, I thought, somebody at Hotel Z is on the ball.

Big Data 很熱爆

- 根據於2011年5月McKinsey(麥肯錫)發表的報告"大數據：創新，競爭和生產力"，到2018年，美國將缺乏140,000-190,000擁有相關的數據分析能力的人才。
- 2012年3月29日，奧巴馬政府宣布了大數據的研究和發展計劃。
- 多個聯邦機構(包括NSF, NIH, DOD, etc)承諾數百萬美元的額外基金，資助有關大數據的研究項目。
- 例：天文學裡的big data、社會網絡的big data

Physica A 329 (2003) 473–483

Information categorization approach to literary authorship disputes

Albert C.-C. Yang[a,b,c,*], C.-K. Peng[a], H.-W. Yien[b,c], Ary L. Goldberger[a]

[a] *Cardiovascular Division and Margret and H.A Rey Institute for Nonlinear Dynamics in Medicine, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, MA 02215, USA*
[b] *School of Medicine, National Yang-Ming University, Taipei, Taiwan*
[c] *Taipei Veterans General Hospital, Taipei, Taiwan*

程甲本《紅樓夢》程偉元序

《紅樓夢》小說本名《石頭記》，作者相傳不一，究未知出自何人，惟書內記曹雪芹先生刪改數過。好事者每傳抄一部，置廟市中，昂其值，得數十金，可謂不脛而走者矣。然原目一百廿卷，今所傳只八十卷，殊非全本。即間稱有全部者，及檢閱，仍只八十卷，讀者頗以為憾。不佞以是書既有百廿卷之目，豈無全璧？爰為竭力收羅，自藏書家甚至故紙堆中無不留心，數年以來，僅積有廿余卷。一日偶於鼓擔上得十餘卷，遂重價購之，欣然繙閱，見其前後起伏，尚屬接筍，然漶漫不可收拾。乃同友人細加厘剔，截長補短，抄成全部，復為鐫板，以公同好，《紅樓夢》全書始自是告成矣。書成，因並志其緣起，以告海內君子。凡我同人，或亦先睹為快者歟？小泉程偉元識

Collier et al. *Journal of Biomedical Semantics* 2011, 2(Suppl 5):S9
http://www.jbiomedsem.com/content/2/S5/S9

JOURNAL OF BIOMEDICAL SEMANTICS

RESEARCH Open Access

OMG U got flu? Analysis of shared health messages for bio-surveillance

Nigel Collier[1,2*], Nguyen Truong Son[3], Ngoc Mai Nguyen[3]

From Fourth International Symposium on Semantic Mining in Biomedicine (SMBM)
Hinxton, UK. 25-26 October 2010

* Correspondence: collier@nii.ac.jp
[1]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku,Tokyo, Japan

Abstract

Background: Micro-blogging services such as Twitter offer the potential to crowdsource epidemics in real-time. However, Twitter posts ('tweets') are often ambiguous and reactive to media trends. In order to ground user messages in epidemic response we focused on tracking reports of self-protective behaviour such as avoiding public gatherings or increased sanitation as the basis for further risk analysis.

Results: We created guidelines for tagging self protective behaviour based on Jones and Salathé (2009)'s behaviour response survey. Applying the guidelines to a corpus of 5283 Twitter messages related to influenza like illness showed a high level of inter-annotator agreement (kappa 0.86). We employed supervised learning using unigrams, bigrams and regular expressions as features with two supervised classifiers (SVM and Naive Bayes) to classify tweets into 4 self-reported protective behaviour categories plus a self-reported diagnosis. In addition to classification performance we report moderately strong Spearman's Rho correlation by comparing classifier output against WHO/NREVSS laboratory data for A(H1N1) in the USA during the 2009-2010 influenza season.

Conclusions: The study adds to evidence supporting a high degree of correlation between pre-diagnostic social media signals and diagnostic influenza case data, pointing the way towards low cost sensor networks. We believe that the signals we have modelled may be applicable to a wide range of diseases.

Rank	1-40		41-80		81-120	
	Word	Frequency	Word	Frequency	Word	Frequency
1	了	6250	了	8301	了	6946
2	不	4505	不	5676	的	5499
3	的	4010	的	5539	不	5009
4	一	3891	一	4942	來	3944
5	道	3683	來	4097	道	3756
6	來	3563	人	3892	是	3741
7	人	3139	我	3769	人	3644
8	我	2843	是	3720	一	3461
9	是	2833	道	3683	說	3391
10	說	2805	說	3637	我	2743

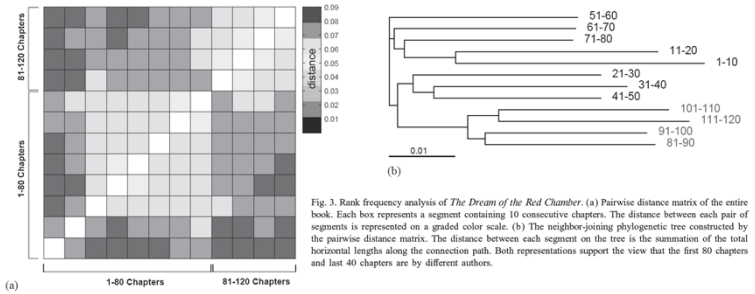
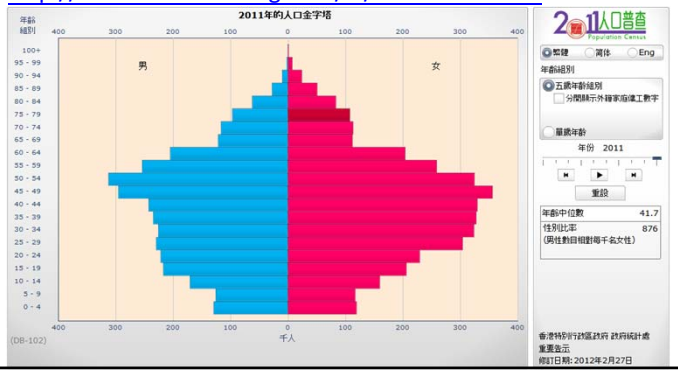


Fig. 3. Rank frequency analysis of *The Dream of the Red Chamber*. (a) Pairwise distance matrix of the entire book. Each box represents a segment containing 10 consecutive chapters. The distance between each pair of segments is represented on a graded color scale. (b) The neighbor-joining phylogenetic tree constructed by the pairwise distance matrix. The distance between each segment on the tree is the summation of the total horizontal lengths along the connection path. Both representations support the view that the first 80 chapters and last 40 chapters are by different authors.

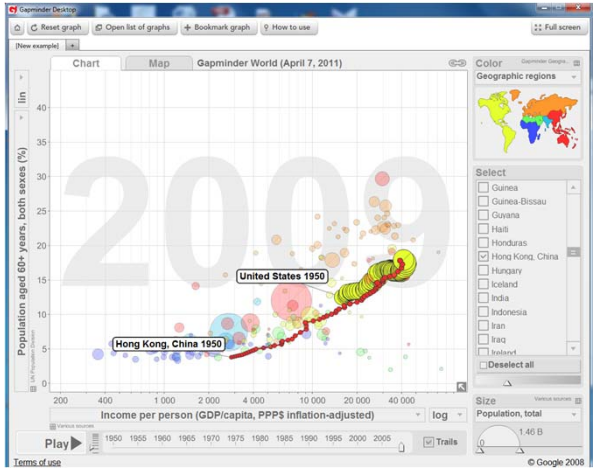
互動人口金字塔

- 了解過去50年人口結構的轉變。

http://www.census2011.gov.hk/tc/interactive-visualisations.html
http://www.census2011.gov.hk/tc/main-table.html



Gapminder http://www.gapminder.org/



37

多謝

Statistics – Dream job of the next decade

"I keep saying that the sexy job in the next 10 years will be statisticians."

Google's Chief Economist, Hal Varian, interviewed by McKinsey Quarterly in January 2009