

Rethinking big data: A review on the data quality and usage issues

大数据反思：对数据质量和使用问题的述评

Jianzheng Liu 刘建政

Email: jzliu.100@gmail.com

Web: <http://www.jzliu.net/>



THE UNIVERSITY OF HONG KONG 香港大學
faculty of architecture 建築學院

Centre of Urban Studies and Urban Planning 城市研究及城市規劃中心
Department of Urban Planning and Design 城市規劃及設計系

“Big data is not a substitute for **common sense**, economic **theory**, or the need for careful **research designs**.”

大数据不能替代科学研究中的常识、理论还有严密的研究设计

— Einav, L. and J. Levin (2014) in Science

“Urban big data is ... **just** data. Data **only matters if it is useful**. Much of the current hype about urban big data comes from people who know about data but who are looking for a cause.”
大数据仅仅是个数据而已，数据只有在它有用时才重要...

— Chris Webster

Dean of Faculty of Architecture
(香港大学建筑学院院长)

□ Outline

- ✓ Overview on big data research;
- ✓ “Big-Errors” brought by big data;
- ✓ coping strategies for “big errors”

□ 本文主旨在于

- ✓ 再思考大数据在人文和城市地理学以及城市研究领域的应用;
- ✓ 试图找出可能会困扰大数据在本学科生产可信准确知识的“误差”因素;
- ✓ 提出一些具体可行的对策和措施。

1

1 Overview on big data research 大数据及其在城市研究相关学 科的应用

大数据定义 Defining Big Data

- Linkable information that have large data volumes and complex data structures (Khoury and Ioannidis, 2014)
- “3V” model, where “3V” refers to volume, variety, and velocity (Laney, 2001)
- Data **availability** and **accessibility**提高了研究对象的数据可获取性
- Extraordinary **fine-grained detailed data** in terms of analysis units, spatial, and temporal resolution. 细粒度的详细数据，更小的分析单位。

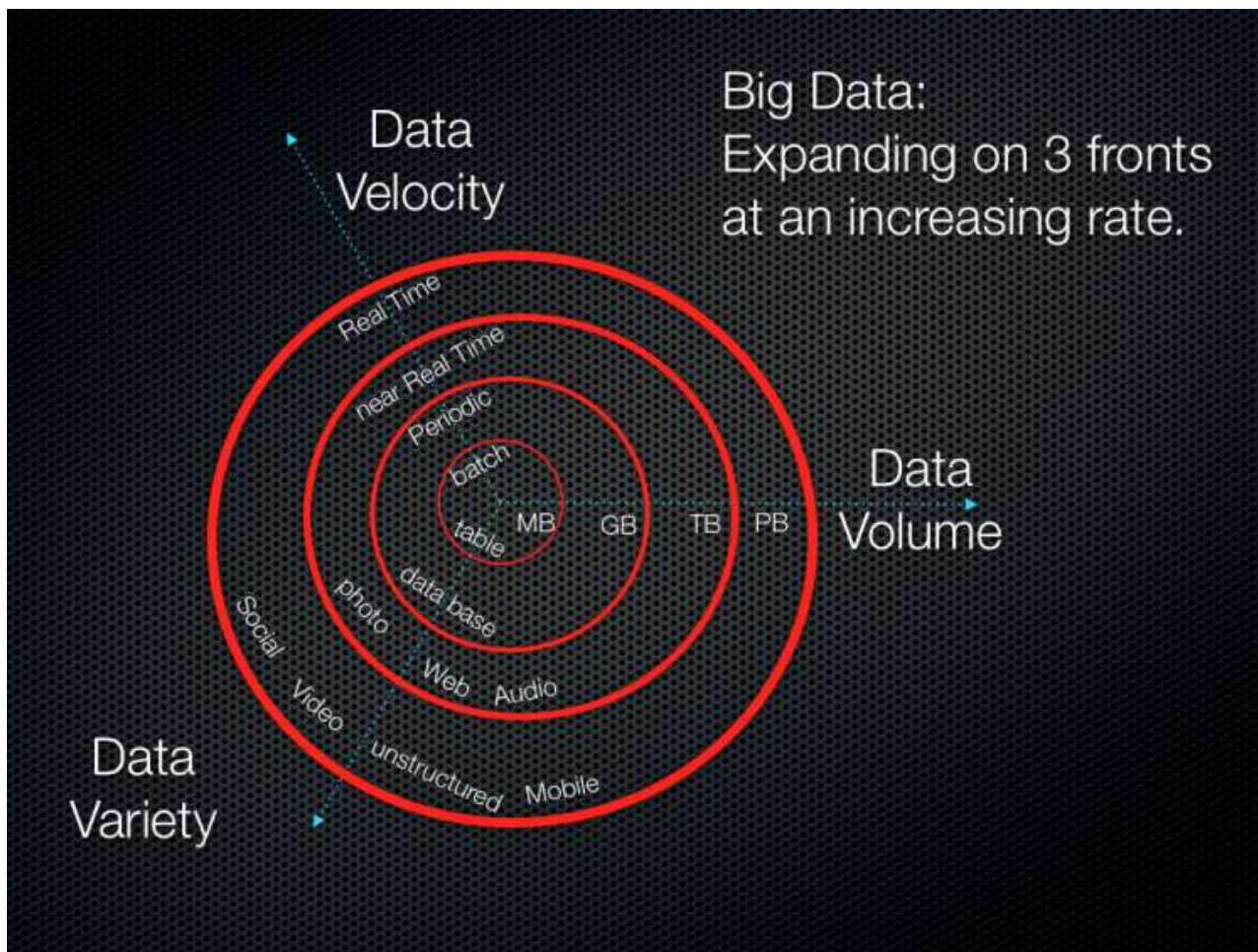


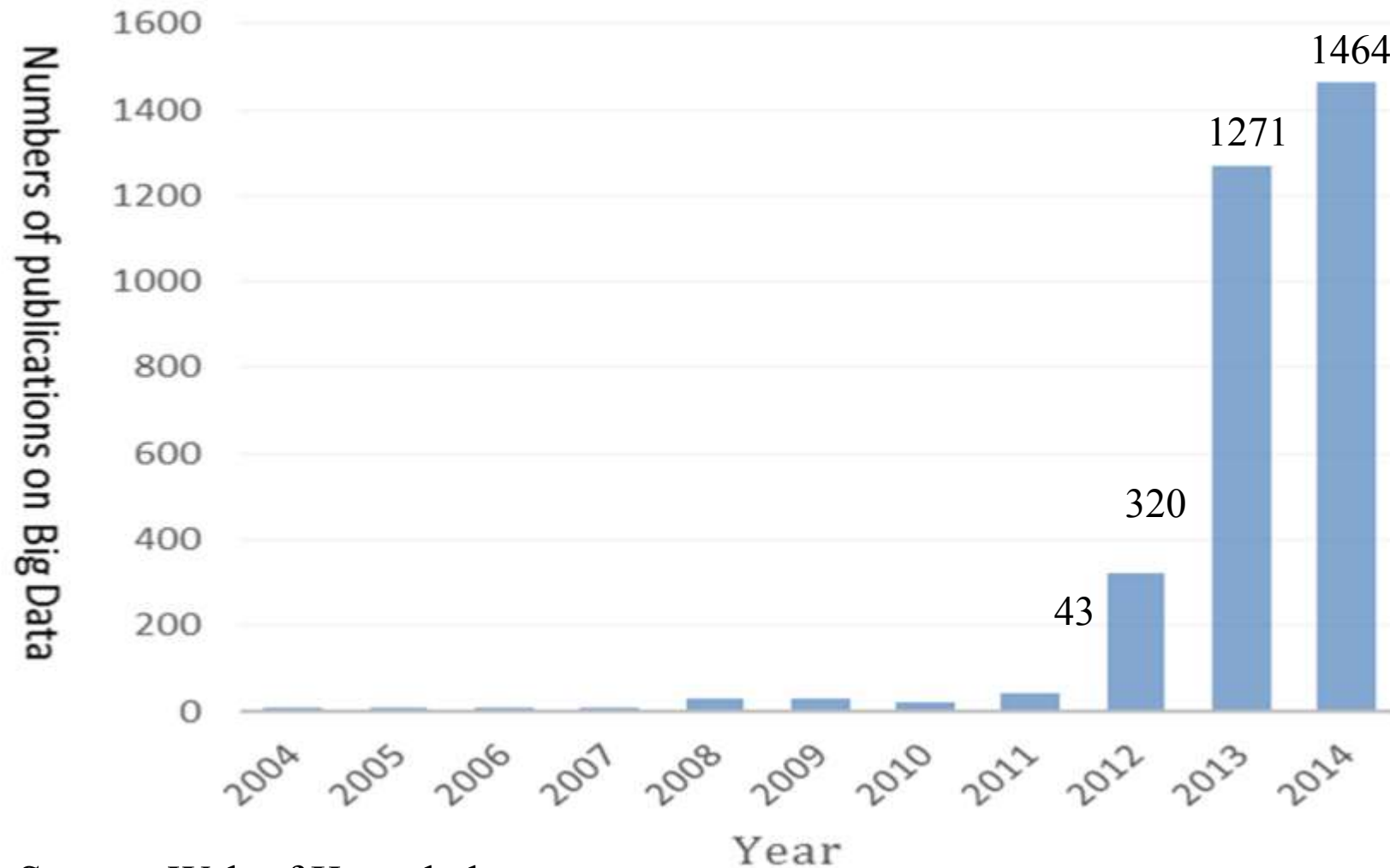
Image Courtesy of Diya Soubra at datasciencecentral.com

“**Fine-scale spatial–temporal data**” will be a more appropriate term to describe the big data in spatial information sciences

在我们城市研究、人文地理领域里面的“大数据”却并不都符合这个3V模型.

大数据定义为**细尺度时空大数据(fine-scale spatial–temporal data)**更为契合。

Number of publications on 'big data'

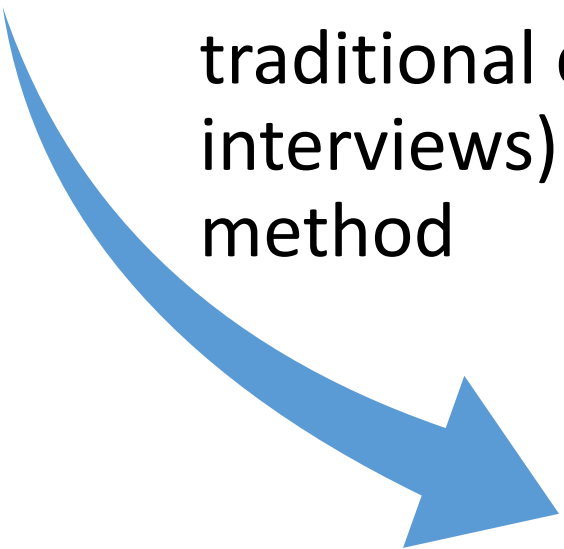


Source: Web of Knowledge

Entry	Study focus	Big data	Method	Result
(Gonzalez et al., 2008)	Human mobility	Mobile phone data	Statistical fitting	Human trajectories show a high degree of temporal and spatial regularity
(Roth et al., 2011)	Human mobility	London subway “Oyster” card data	Null model	A polycentric structure composed of large flows organized around a limited number of activity centers.
(Krings et al., 2009)	Spatial interaction	Mobile phone data	Gravity model	Communication intensity between two cities is proportional to the product of their sizes divided by the square of their distance.
(Zheng et al., 2013)	Urban computing	Air monitoring data and points of interest	A semi-supervised learning approach based on the artificial neural network and conditional random field.	Advantages of the proposed method over four categories of baselines such as decision tree, CRF, and ANN.
(Fu and Chau, 2013)	Data quality	Social media (Sina Weibo)	Random sampling approach	Representative and reliable statistics on Chinese micro-bloggers are limited.
(Haklay, 2010)	Data quality	OpenStreetMap	Comparison with Ordnance Survey datasets	OSM information can be fairly accurate

How does big data change current research

- Big data exert their impact on spatial information sciences and related fields in three aspects, namely, **data collection**, **data processing**, and **data analysis**.
- Data collection approach has been transformed from traditional data collection methods (e.g., questionnaires and interviews) into a fast and powerful ICT-based data collection method



“Urban big data is ... **just** data. Data **only matters if it is useful**. Much of the current hype about urban big data comes from people who know about data but who are looking for a cause.”
大数据仅仅是个数据而已，数据只有在它有用时才重要...

— Chris Webster
(香港大学建筑学院院长)

Evaluating current big data research

评价当前的大数据研究

- Some of the existing big data studies basically follow a research paradigm of combining “new approaches based on new data” with old topics
- 现有的部分大数据研究基本上是在一个“基于新数据的新方法” + “旧话题”的创新模式上进行

	New phenomenon / problem /topic	New method	New data/context
Old phenomenon/ problem/topic	Nil	Good (Methodological study)	Problematic if without new insights
Old method	Good (New areas)	Nil	Problematic if without new insights
Old data/context	Good (New areas)	Good	Nil

2

2 “Big errors” 数据质量和使用问题

Big Data may bring “Big Error”

大数据之大不仅在于数据容量之大、
数据结构之复杂、数据实时性之好，
也在于数据误差之大。

Inauthentic data collection 大数据收集的问题

- The collection of Big data lacks authenticity and credibility 众多的大数据的收集在权威性和公信力这方面存在问题
- **Commercial** companies that are not established for scientific research purposes but are business platforms that **pursue profits**. 以逐利为目的的商业化平台
 - **neither adopt scientific** data collection procedures **nor follow solid** and scientific data processing procedures. **Unkown** sampling method and processing algorithms behind the web service
 - can change the sampling methods and processing algorithms at any time **without any notice**
 - **no obligation or motivation** to ensure the authenticity and validity of the data

Google Flu Trends Index 谷歌流感趋势指数

- Google often changes the algorithms and make the prediction unstable (Lazer, D. M., et al., 2014)。

google.org Flu Trends

[Google.org home](#)

[Dengue Trends](#)

Flu Trends

[Home](#)

United States ▼

National ▼

[Download data](#)

[How does this work?](#)

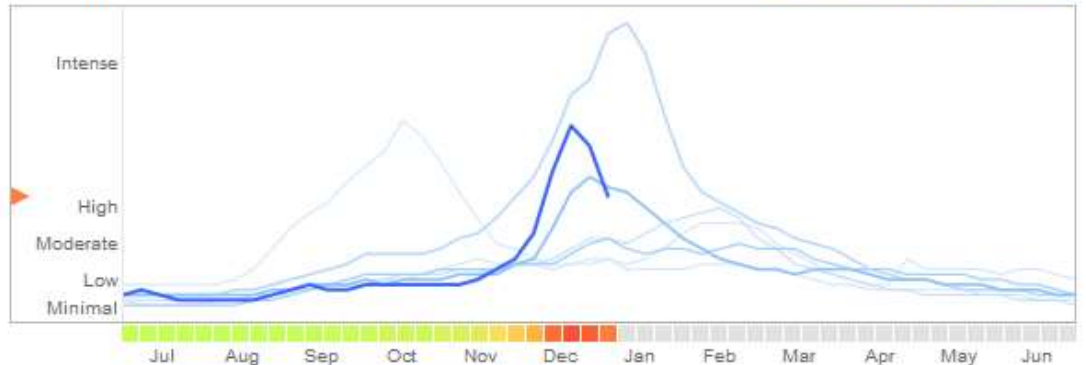
[FAQ](#)

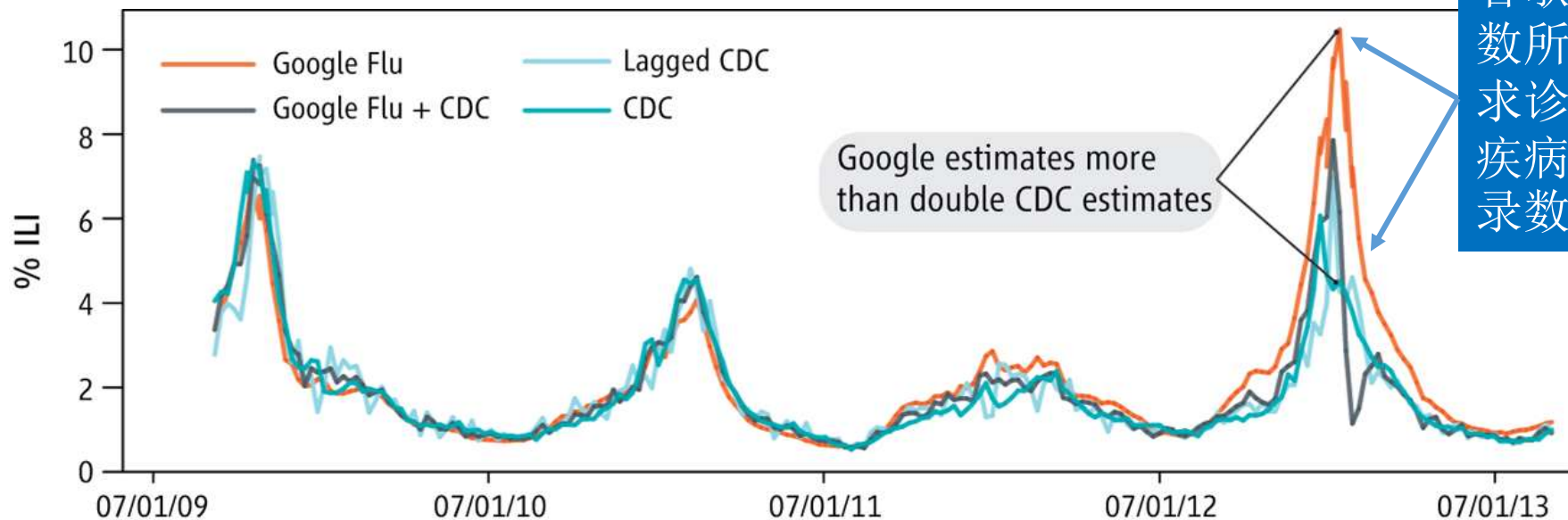
Explore flu trends - United States

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 »](#)

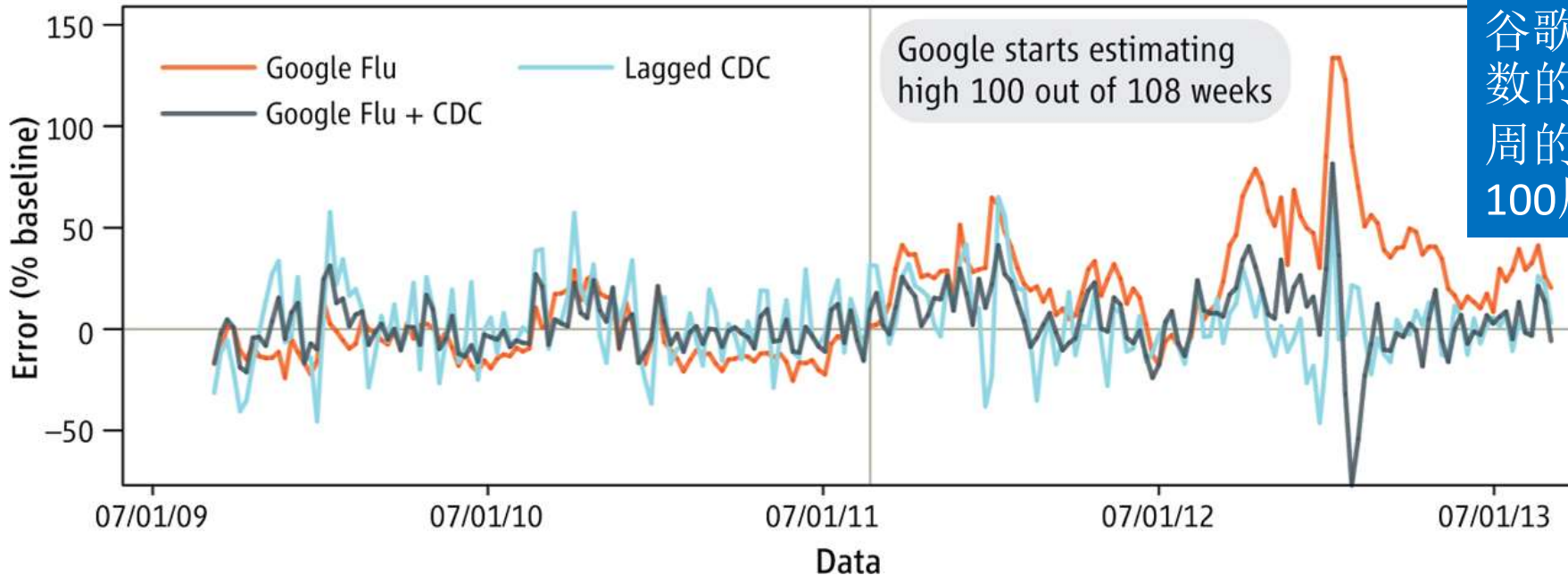
National

● 2014-2015 ● Past years ▼





谷歌流感趋势指数所估计的流感求诊次数是美国疾病预防控制中心记录数据的两倍



谷歌流感趋势指数的估计在108周的统计中有100周都过高

Information incompleteness

大数据的信息片面性

- Mobile phone data: **Lack of socio-economic attributes** which restrain its applications
- Cannot reflect the **differences** among respondents, or describe residents' **behavior characteristics** and other essential information that interests researchers and readers.



Image Courtesy at express.co.uk & theguardian.com

Information incompleteness

大数据的信息片面性

- Geo-location information in mobile phone data is **not the exact location** of phone calling activities
- Can identify **only working and residential activities** during weekdays and recreation activities during weekends.
- Only record the moving patterns of people, which is **only a small part of the daily life of people, but ignore most of the time** spent in the office or at home.



Information incompleteness

大数据的信息片面性

- **Smart card data** have similar information incompleteness problems with mobile phone data.
- Several researchers have adopted a winding approach to obtain the socio-economic data of smart card holders by combining traditional resident travel survey and land use data (Long and Thill, 2013).
- The results may have a **certain accuracy** level, but introduce **huge errors** because of many uncertainties



Image Courtesy of www.leiphone.com

Representativeness problems

大数据的样本片面性

- The large size and volume of big data do not necessarily mean that the data is random and representative (Boyd and Crawford, 2012)

Figure 1. Picking an inappropriate sample of your population

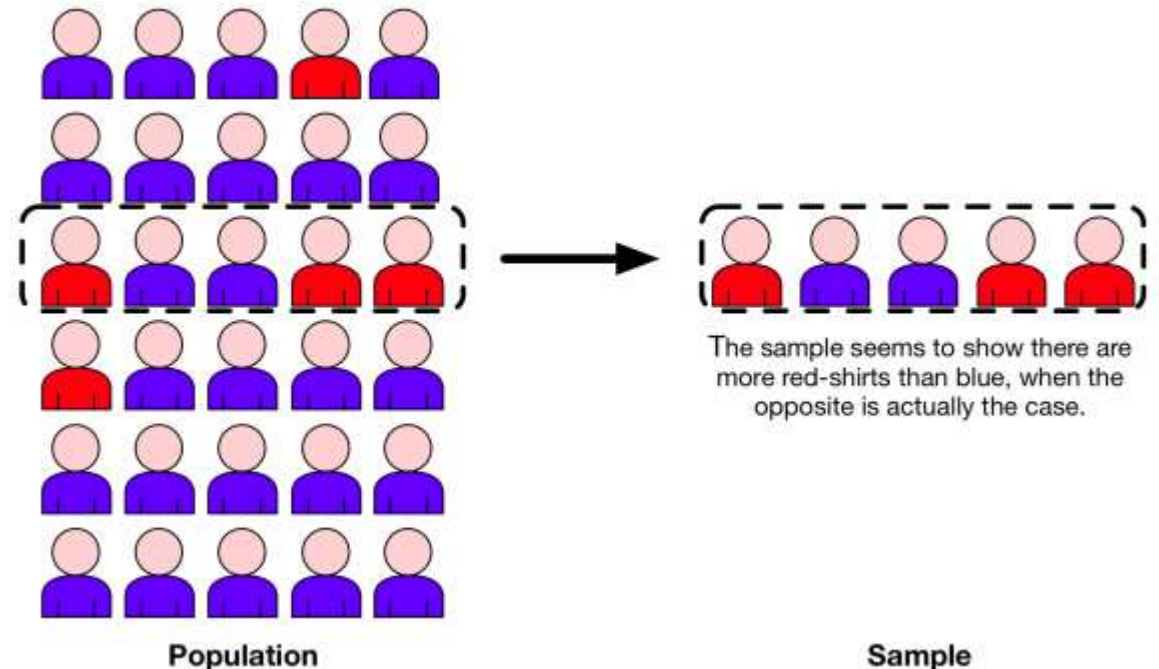
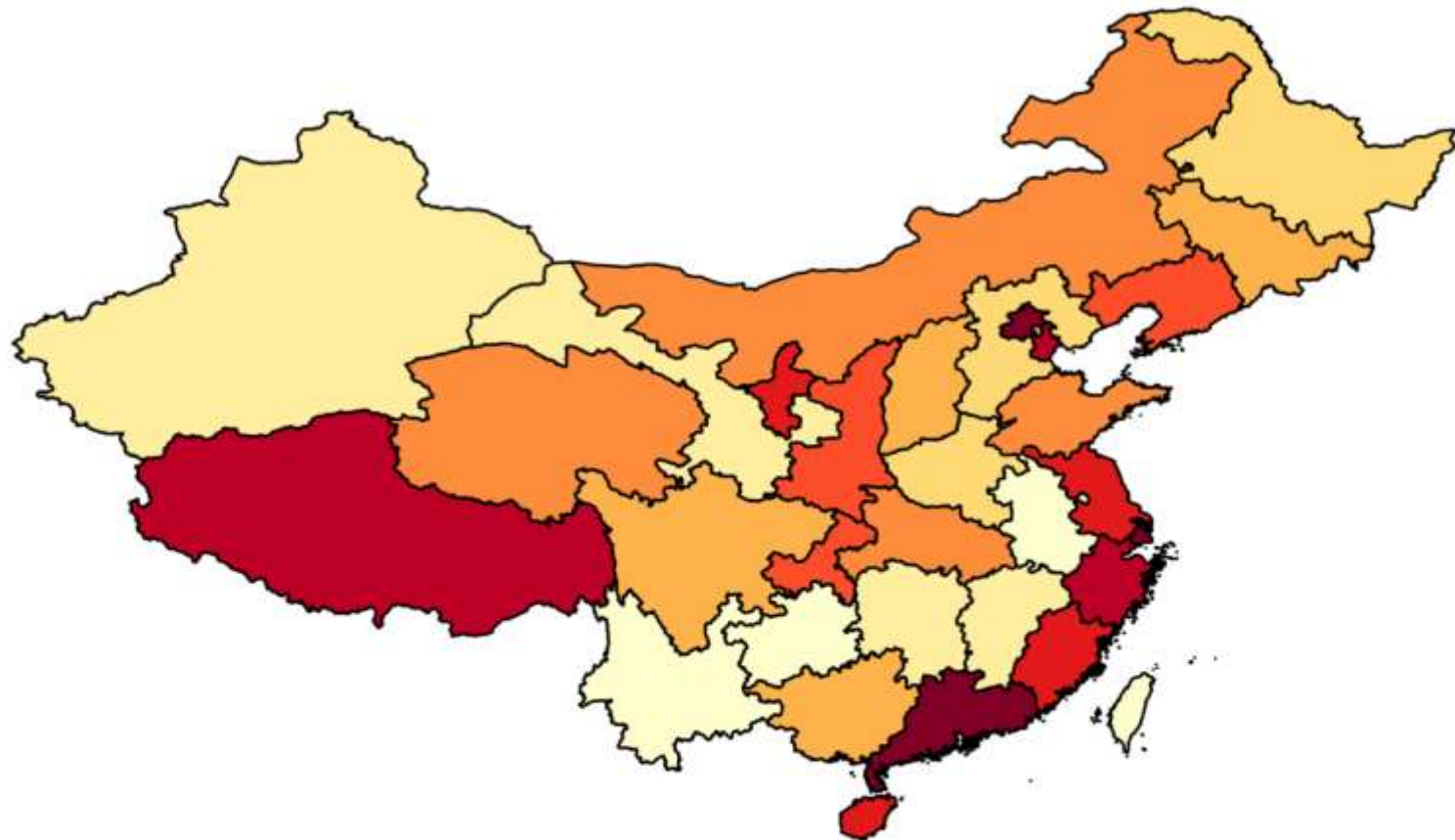


Image Courtesy of arrestingthoughts.org



Ratio of observed / expected # of microbloggers



More than one quarter of microblogging users are located in well-developed regions, including Guangdong, Beijing, and Shanghai where the Internet users of these three places only account for 9% of the total population of China's Internet users

超过四分之一的微博用户集中在广东、北京和上海，而这三个地方的网民却只占全国网民数量的9%

(Fu & Chau, 2013)

The Landscape of Social Media Users

	% of internet users who....	The service is especially appealing to ...
Use Any Social Networking Site	67%	Adults ages 18-29, women
Use Facebook	67	Women, adults ages 18-29
Use Twitter	16	Adults ages 18-29, African-Americans, urban residents
Use Pinterest	15	Women, adults under 50, whites, those with some college education
Use Instagram	13	Adults ages 18-29, African-Americans, Latinos, women, urban residents
Use Tumblr	6	Adults ages 18-29

Source: Pew Research Center's Internet & American Life Project Post-Election Survey, November 14 – December 09, 2012. N=1,802 internet users. Interviews were conducted in English and Spanish and on landline and cell phones. Margin of error is +/- 2.6 percentage points for results based on internet users. Facebook figures are based on Pew Research Center's Internet & American Life Project Omnibus Survey, December 13-16, 2012. Margin of error for Facebook data is +/- 2.9 percentage points for results based on internet users (n=860).

Source: <http://www.pewinternet.org>

Consistency and reliability problems

大数据的可靠性问题

- Some big data fails in genuinely representing the facts and information on research subjects
- Result based on some big data are not consistent and stable

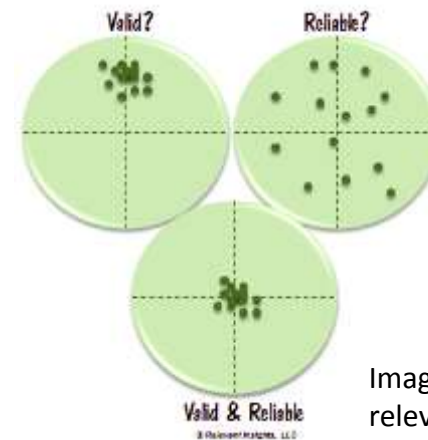
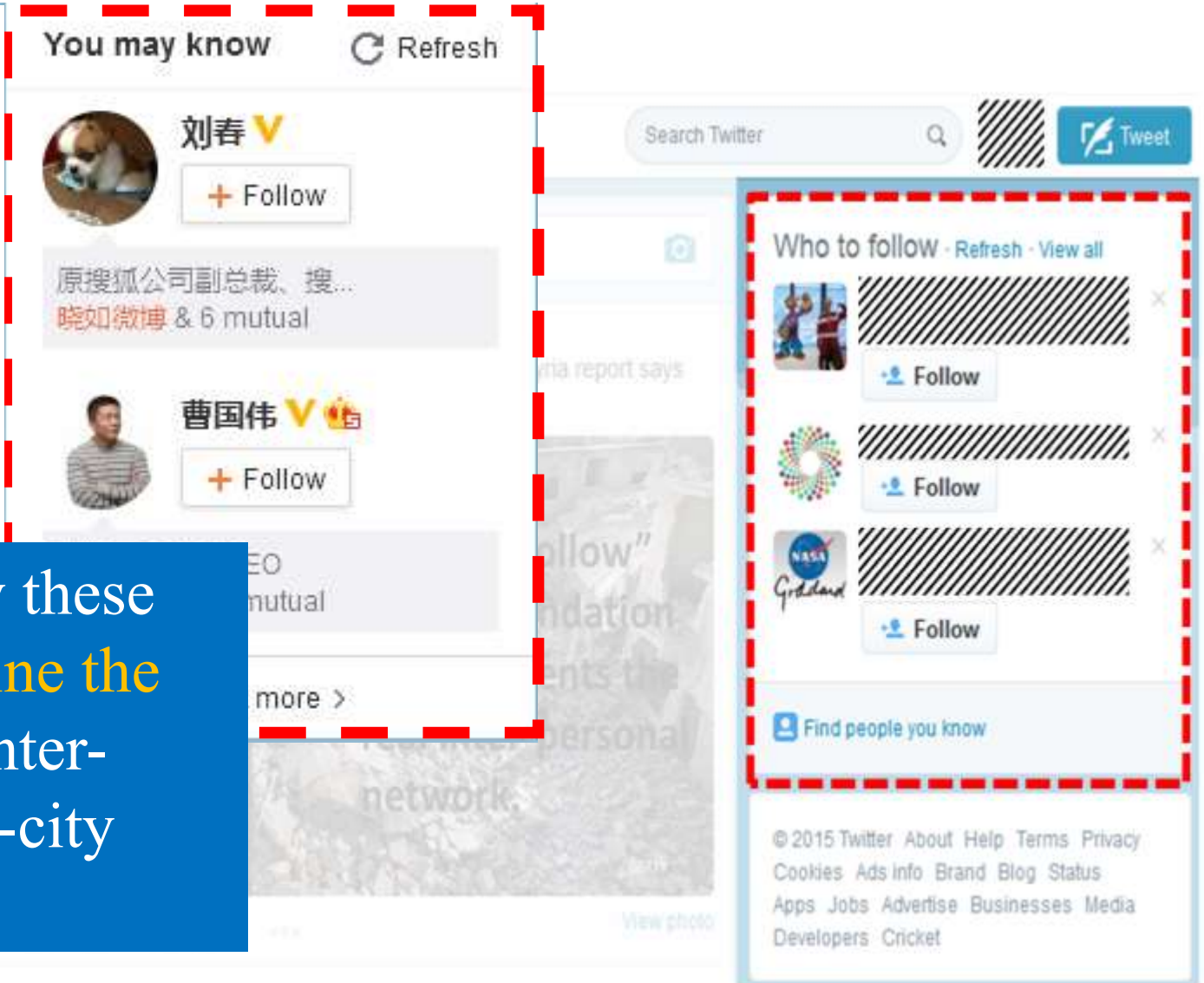


Image Courtesy of
relevantinsights.com

Case 案例

- when studying the inter-personal communication links between cities, researchers are likely to overlook the fact that the social media services itself **alters the true inter-personal networks** by recommending friends on the basis of the user's place of birth, gender, education background, and other attributes.

好友推荐！ friend recommendations



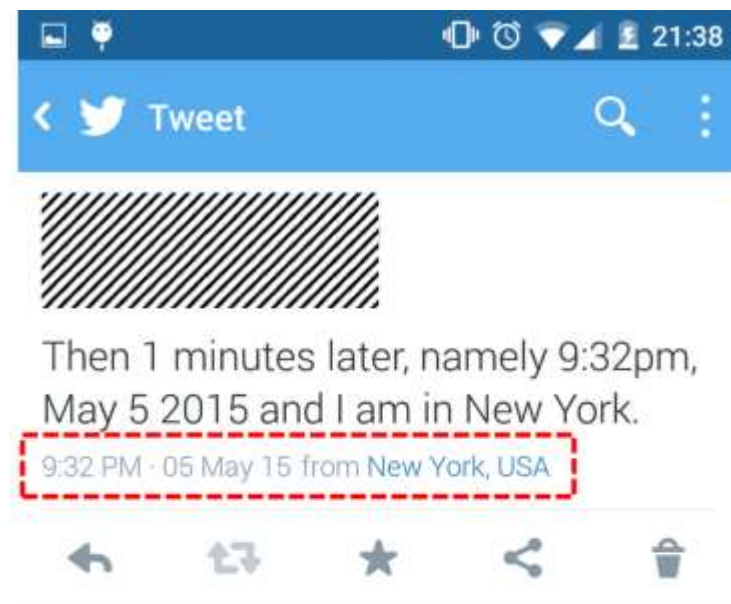
This hidden influence exerted by these social media companies **undermine the credibility** of the conclusion of inter-personal communication or inter-city cyberspace analyses

Case 案例

- The geo-location of the geo-tagged tweets is probably inaccurate and wrong because Twitter users can **post geo-tagged tweets literally at “any place”** on Earth.
- Search engine data are another example of the **distortion of users’ behavior and activities** (Ruths & Pfeffer, 2014)。

Geo-tagged tweets
can be sent **without
verification** on the
geo-location.

Twitter可以随意在
任何地方签“**到**”



Free and without-verification geo-tagged tweet postings

可以随意在任何地方签“到”



自动完成(auto-complete)

谷歌更新了搜索引擎中的自动完成功能

发布时间: 2011. 04. 22 07:29

来源: 赛迪网

作者: 文良

【赛迪网讯】4月22日消息，据国外媒体报道，谷歌近日更新了搜索引擎中的自动完成（Autocomplete）功能。现在的自动完成功能可以预测搜索词条即使这个词条之前并没有用户输入过。一般情况下，自动完成功能依赖于之前的搜索。但是如果这是一个新词条，自动完成功能往往就会失效。自动完成功能在进行词条预测时，仍然关注的是已输入的字词，哪怕只有一个字，所以与以往的自动完成功能相比此次更新能够预测更多。

谷歌软件工程师巴特罗密-涅奇威（Bartłomiej Niechwiej）称：“现在我们正在提高自动完成功能的预测能力，它可以帮助你搜索这个世界上还没有人搜索过的东西。自动完成功能工作的主要方式之一就是通过寻找最普遍的搜索词条，比如，大多数人输入wea想要搜索weather的时候，谷歌就可以做出这样的预测。”

涅奇威补充道：“棘手的是相当比例的搜索请求是我们几乎没有输入过的。所以这就让我们基于普遍性的词条预测变得很难。目前该功能还仅限于在Google.com用英语搜索，不久还会适用于其他语言。



北京|

北京樓

北京天氣

北京同仁堂

北京

北京道一號

北京老家

北京愛情故事

北京人家

北京地鐵

北京遇上西雅圖

Google Search

I'm Feeling Lucky

Ethical issues 大数据的隐私问题

- Just because it is accessible does not make it ethical
- ignoring the ethical evaluation of research is problematic for researchers because the data are seemingly public (Boyd and Crawford, 2012)



Image Courtesy of research.unsw.edu.au

Ethical issues 大数据的隐私问题

- For studies that look into the research topics involved with user privacies,
 - Does the researcher collect sensitive data regarding social network users' political ideas or actions?
 - Is the research data stored properly?
 - Is there any risk of data leakage?
 - Are the user IDs and names replaced with pseudo codes?
 - Have measures been taken to minimize the minimum risk emphasized in the international research ethics?

Big Data may brings “Big Error”

大数据之大不仅在于数据容量之大、数据结构之复杂、数据实时性之好，也在于数据误差之大。

Thank you!



www.jzliu.net