计算社会科学：方法与应用

孟天广 副教授

数据治理研究中心 主任

清华大学政治学系，清华大学苏世民书院

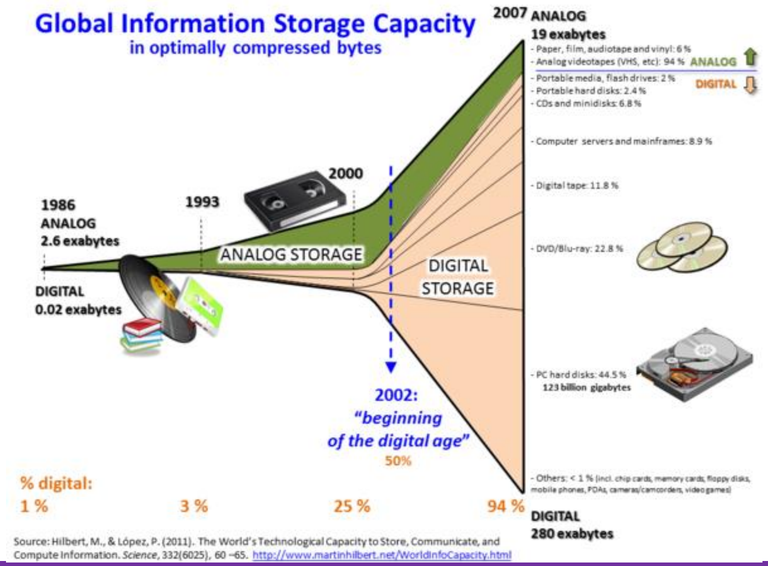
2018年12月27日

内容提要

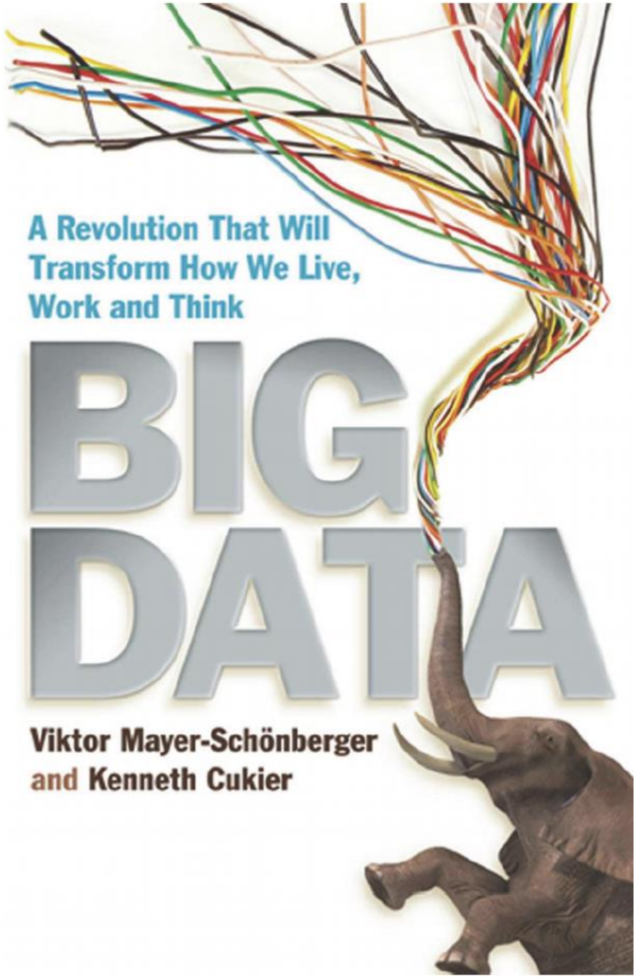
- 大数据+社会科学
- 计算社会科学：方法论
- 计算社会科学：方法与议题
- 计算社会科学：应用举例

大数据时代

- 《科学》2008年提出“大数据”来讨论新信息时代（PB时代）的科学研究；
- 2012年，《纽约时报》刊文宣告“大数据时代已经到来”；



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html



大数据时代

- 大数据是指需要新处理模式才能确保更强的决策力、洞察力和流程优化力的海量、高速增长和多样化的信息财富（Gartner）
- 大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态。（国务院）
- 数据已成为国家基础性战略资源（十三五）

大数据

What is Big Data

- Data
- Analytics
- Industry
- Solution

Why is Big Deal

- Government
- Private Sector
- Science
- Social Science Revolution

大数据的特征

- Volume(体量大):
  - How much data is really relevant to the problem solution? Cost of processing?
  - *So, can you really afford to store and process all that data?*
- Velocity(增速快):
  - Much data coming in at high speed
  - Need for streaming versus block approach to data analysis
  - *So, how to analyze data in-flight and combine with data at-rest*
- Variety(类型多):
  - A small fraction is structured formats, Relational, XML, etc.
  - A fair amount is semi-structured, as web logs, etc.
  - The rest of the data is unstructured text, photographs, etc.
  - *So, no single data model can currently handle the diversity*
- Veracity(真实性):
  - Accuracy, Precision, Reliability, Integrity
  - *So, what is it that you don’t know you don’t know about the data?*
- Value(价值高):
  - How much value is created for each unit of data (whatever it is)?
  - *So, what is the contribution of subsets of the data to the problem solution?*

大数据+社会科学

- 大数据时代的八大机遇

✓海量非结构化数据（信息）

✓“全量数据”而不是“样本数据”

✓丰富、高效的方法工具箱

✓机器学习与预测性分析

✓强时效性数据

✓社会科学知识平民化普及

✓良好的组织环境+充分的社会需求

大数据+社会科学

- 大数据+社会科学

✓数据驱动

✓应用（问题解决）导向

✓中观/微观问题

✓预测性目标

✓诊断性目标

- 从海量数据中利用机器学习抽离出有价值的信息

✓积累海量数据

✓利用统计和数学知识+模式识别技术

✓发现有意义的新关系、新模式或新趋势

计算社会科学的清华探索

- 清华大学计算社会科学平台
- 2017年5月建立
- 成立背景

✓ 服务国家大数据战略

✓ 服务中国特色国家治理结构和治理能力现代化

✓ 服务计算社会科学的学科建设和人才培养

✓ 服务清华社会科学发展，打造清华社科品牌

✓ 国际合作与交流

计算社会科学的清华探索

- 清华大学计算社会科学平台
- 平台定位和发展目标

✓ 国内首家计算社会科学领域的创新研究机构

✓ 营造“创新、共享、开放、合作”的科研环境

✓ 立足清华社科、计算科学和数据科学领域的特色优势和交叉研究基础

✓ 将社科议题、海量数据、大数据方法相结合，促进跨学科融合，开展创新性研究

✓ 集前沿研究、人才培养、科研服务和智库资政于一体

✓ 为中国特色计算社会科学的学科建设、清华高水平社科创新成果、促进重大经济社会问题解决提供理论和应用知识

计算社会科学的清华探索

- 清华大学计算社会科学平台
- 平台功能

✓ 展示清华计算社会科学研究重大突破和创新成果；

✓ 建设数据平台，实现跨院系数据共享与合作；

✓ 数据采集、挖掘与计算技术服务；

✓ 大数据研究方法培训等功能，举办年度性“大数据社会科学讲习班”；

✓ 探索计算社会科学的学科建设、基础理论与方法体系；

✓ 举行“中国计算社会科学高端论坛”等高端学术会议促进学科建设和学术交流。

大数据应用实践举例：计算社会科学数据平台

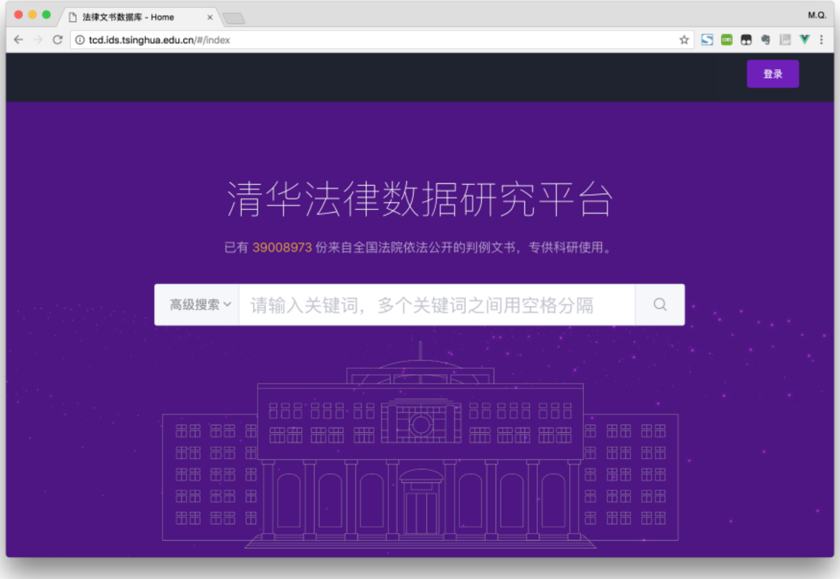
1	司法裁判文书	全国依法公开的司法裁判文书数据，包括民事、刑事、行政、知识产权、执行案件等5000多万份。
2	企业工商基本信息	全国工商登记企业基本信息，2000多万份。
3	企业名录数据	登记在册企业名录信息
4	行政许可	对企业的行政许可信息
5	行政处罚	对企业的行政处罚信息
6	地方领导资料库	地级以上城市主要领导履历资料信息，包括姓名、性别、省份、职位、任职地点、籍贯、出生年月、民族、毕业院校、学历、专业、工作经历等。
7	供地计划	土地供地计划列表信息，包括行政区域、供地计划名称、计划详情链接、发布时间等。
8	供地结果公告	土地使用结果公告列表信息，包括行政区域、土地坐落、总面积、土地用途、供应方式、签订日期、详情url地址等。
9	省市两级政府工作报告	1980年以来有记录的政府工作报告
10	失信人名单	全国公开失信人信息
11	人大代表建议案	各级人大代表的建议案
12	环境领域的运动式治理	专项整治、专项治理信息
	其他更多数据… …	

计算社会科学的清华探索

- 清华大学计算社会科学平台

➢法律数据科研条件平台

- 该平台经过半年多的建设，已汇集了4100余万份全国范围内依法公开的司法判决文书，形成可持续更新的法律数据库，具备全文检索、分类检索、结构化分析、统计分析、可视化报表等在线服务功能。
- 平台链接：http://tcd.ids.tsinghua.edu.cn/


法律文书数据库 - Home
M.Q.
tcd.ids.tsinghua.edu.cn/#/index
登录
高级搜索
请输入关键词，多个关键词之间用空格分隔

清华法律数据研究平台

已有 39008973 份来自全国法院依法公开的判例文书，专供科研使用。

计算社会科学的清华探索

- 清华大学计算社会科学平台

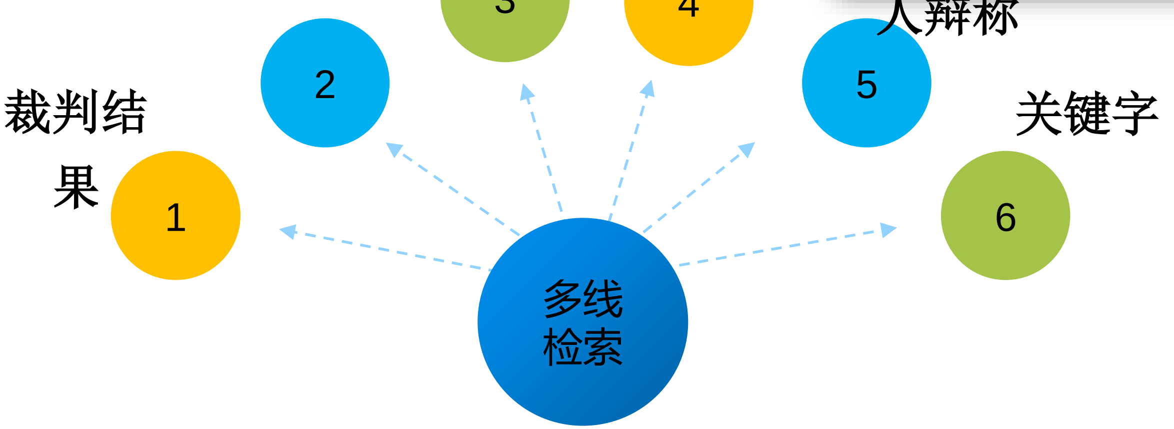
➢法律数据科研条件平台

  - 深层解构文书，帮助用户快速找到有效文书。分类越细，用户交叉搜索的维度越多，结果就越准确，极大提高检索效率。

审理经过

上诉人诉称

本院认为



被上诉人辩称

裁判结果

关键字

1

2

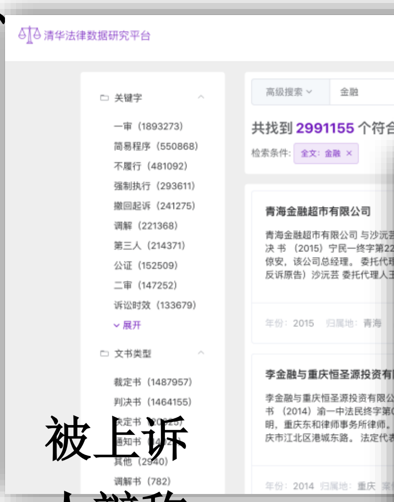
3

4

5

6

多线检索



高级搜索 ⌄ 金融

共找到 2991155 个符合条件的结果 统计搜索结果

检索条件: 全文: 金融 ×

关键字

一审（1893273）
简易程序（550868）
不履行（481092）
强制执行（293611）
撤回起诉（241275）
调解（221368）
第三人（214371）
公证（152509）
二审（147252）
诉讼时效（133679）
⌄ 展开

文书类型

裁定书（1487957）
判决书（1464155）
决定书（20023）
通知书（14?2?）
其他（2940）
调解书（782）

青海金融超市有限公司

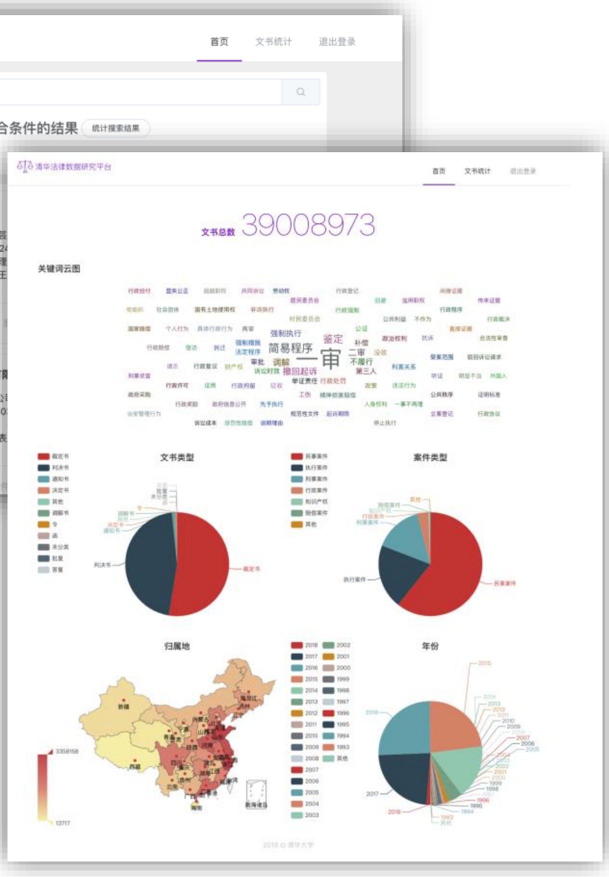
青海金融超市有限公司 与沙沅芸
决 书 （2015）宁民一终字第224
倧安，该公司总经理。委托代理
反诉原告）沙沅芸 委托代理人王

年份：2015 归属地：青海

李金融与重庆恒圣源投资有限

李金融与重庆恒圣源投资有限公司
书 （2014）渝一中法民终字第0
明，重庆东和律师事务所律师。
庆市江北区港城东路。法定代表

年份：2014 归属地：重庆 案件



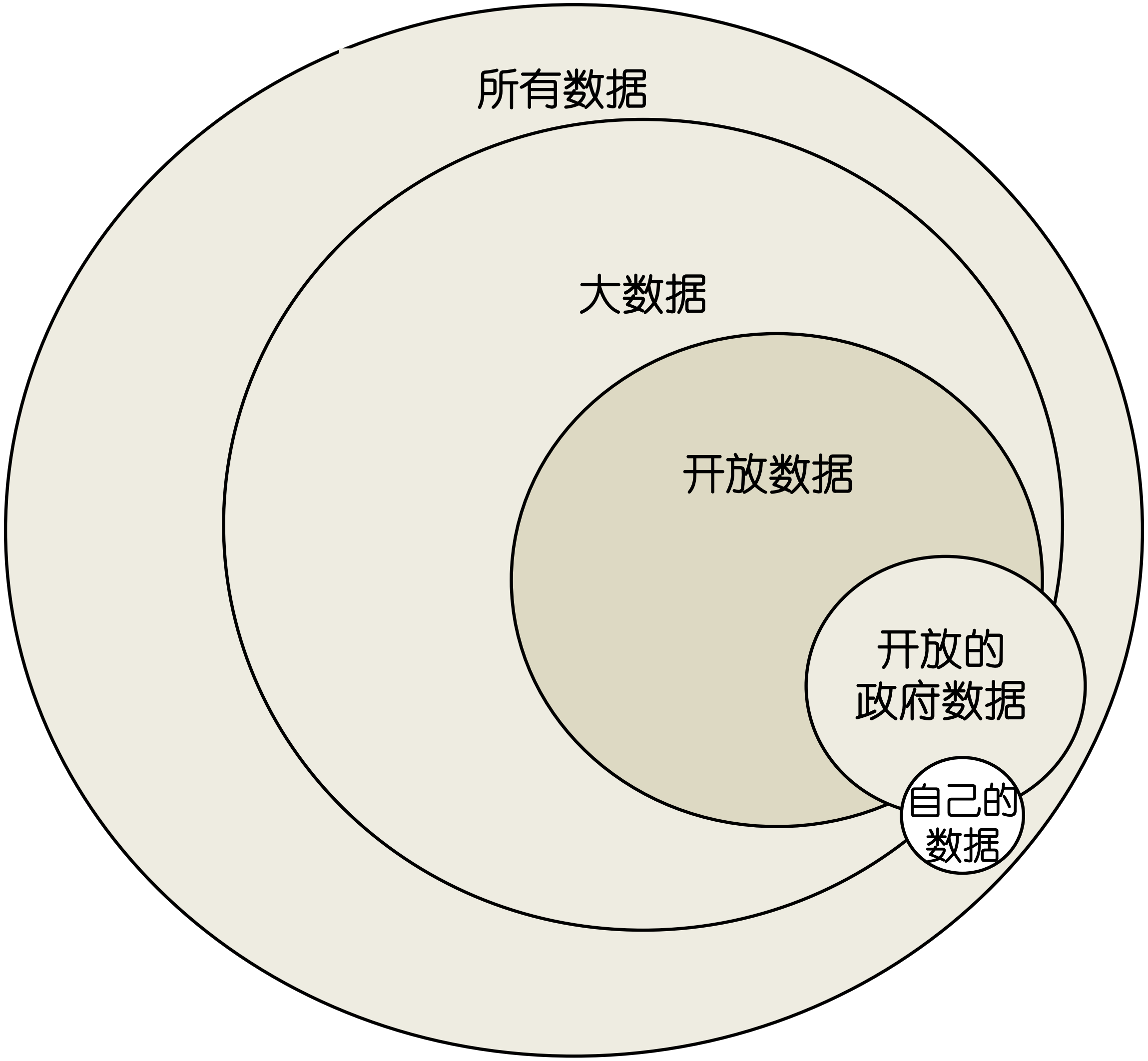
文书总数 39008973

大数据分析的类型

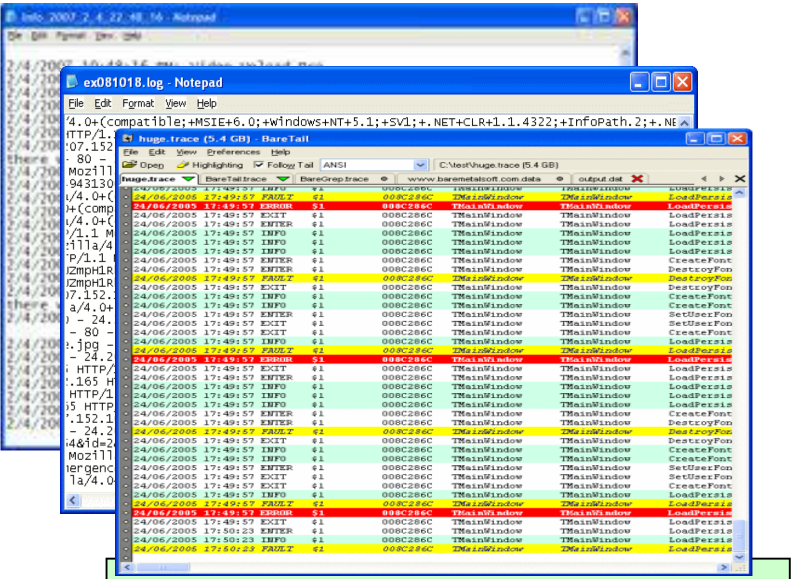
- ***Descriptive***: A set of techniques for reviewing and examining the data set(s) to understand the data and analyze business performance.
- ***Diagnostic***: A set of techniques for determine what has happened and why
- ***Predictive***: A set of techniques that analyze current and historical data to determine what is most likely to (not) happen
- ***Prescriptive***: A set of techniques for computationally developing and analyzing alternatives that can become courses of action – either tactical or strategic – that may discover the unexpected
- ***Decisive***: A set of techniques for visualizing information and recommending courses of action to facilitate human decision-making when presented with a set of alternatives.

	消极的		积极的	
演绎	描述性分析			诊断性分析
归纳	预测性分析			指导性分析

大数据的范畴



用户生成数据


ex081018.log - Notepad

```
4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+1.1.4322;+InfoPath.2;+.NE
```

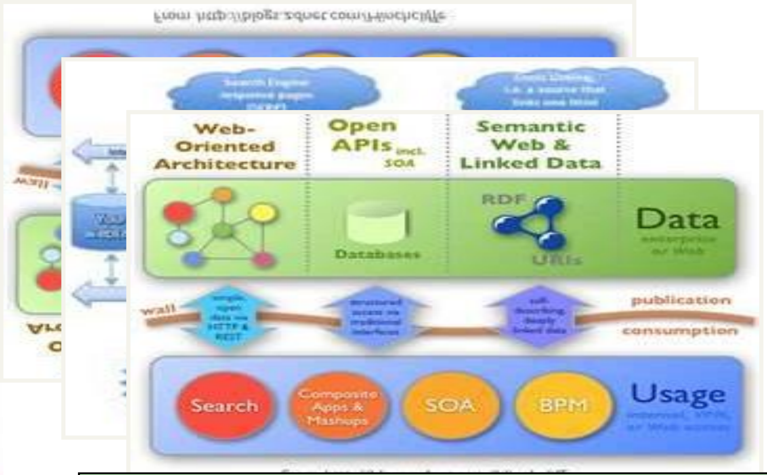
huge.trace (5.4 GB) - BareTail

File Edit View Preferences Help
Open Highlighting ☑ Follow Tail ANSI
C:\test\huge.trace (5.4 GB)
huge.trace
BareTail.trace
BareGrep.trace
www.baremetalsoft.com.data
output.dat

```
24/06/2005 17:49:57 FAULT  $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 ERROR  $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 EXIT   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 ENTER  $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 ENTER  $1   008C286C   TMainWindow   TMainWindow   CreateFont
24/06/2005 17:49:57 ENTER  $1   008C286C   TMainWindow   TMainWindow   DestroyFon
24/06/2005 17:49:57 FAULT  $1   008C286C   TMainWindow   TMainWindow   DestroyFon
24/06/2005 17:49:57 EXIT   $1   008C286C   TMainWindow   TMainWindow   DestroyFon
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   CreateFont
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   CreateFont
24/06/2005 17:49:57 ENTER  $1   008C286C   TMainWindow   TMainWindow   SetUserFon
24/06/2005 17:49:57 EXIT   $1   008C286C   TMainWindow   TMainWindow   SetUserFon
24/06/2005 17:49:57 EXIT   $1   008C286C   TMainWindow   TMainWindow   CreateFont
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 FAULT  $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 ERROR  $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 EXIT   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 ENTER  $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 ENTER  $1   008C286C   TMainWindow   TMainWindow   CreateFont
24/06/2005 17:49:57 ENTER  $1   008C286C   TMainWindow   TMainWindow   DestroyFon
24/06/2005 17:49:57 FAULT  $1   008C286C   TMainWindow   TMainWindow   DestroyFon
24/06/2005 17:49:57 EXIT   $1   008C286C   TMainWindow   TMainWindow   DestroyFon
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   CreateFont
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   CreateFont
24/06/2005 17:49:57 ENTER  $1   008C286C   TMainWindow   TMainWindow   SetUserFon
24/06/2005 17:49:57 EXIT   $1   008C286C   TMainWindow   TMainWindow   SetUserFon
24/06/2005 17:49:57 EXIT   $1   008C286C   TMainWindow   TMainWindow   CreateFont
24/06/2005 17:49:57 INFO   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 FAULT  $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 ERROR  $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:49:57 EXIT   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:50:23 ENTER  $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:50:23 INFO   $1   008C286C   TMainWindow   TMainWindow   LoadPersis
24/06/2005 17:50:23 FAULT  $1   008C286C   TMainWindow   TMainWindow   LoadPersis
```

Deep Web数据

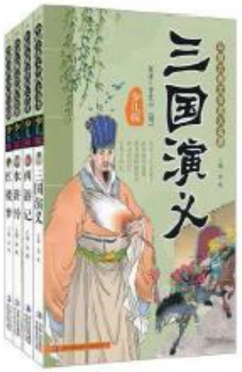
多模态内容数据

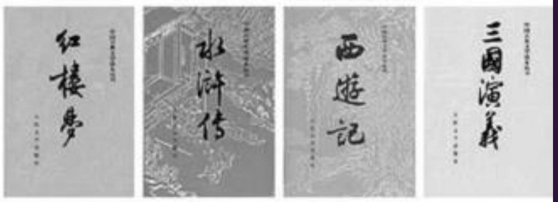


网络与关系数据

文本资料

Text data





中共中央
政治局
群众
维护
中华民族
马克思
结合
生产力
现代化
腐败
反腐倡廉
统筹
上层建筑
保障性
联系
新华社
节约
自信
工程
落实
毛泽东
共产党
研究员
经验
力度
布局
胜利
发展观
综合国力
基本矛盾
良好
反腐败
示范区
继承
来源于
清华大学
科学技术
维权
近年来
意味着
自然
研究所
绿色
艰巨
苦难
阶段性
高新技术
世界大
自主
综合
中国人
应对
行为
获得
城镇化
下决心
带来
血肉
美德
负责人
考验
价值观
决定性
经过
全过程
有利于
海洋权
工程院
主动性
范围
自觉
历史性
多层次

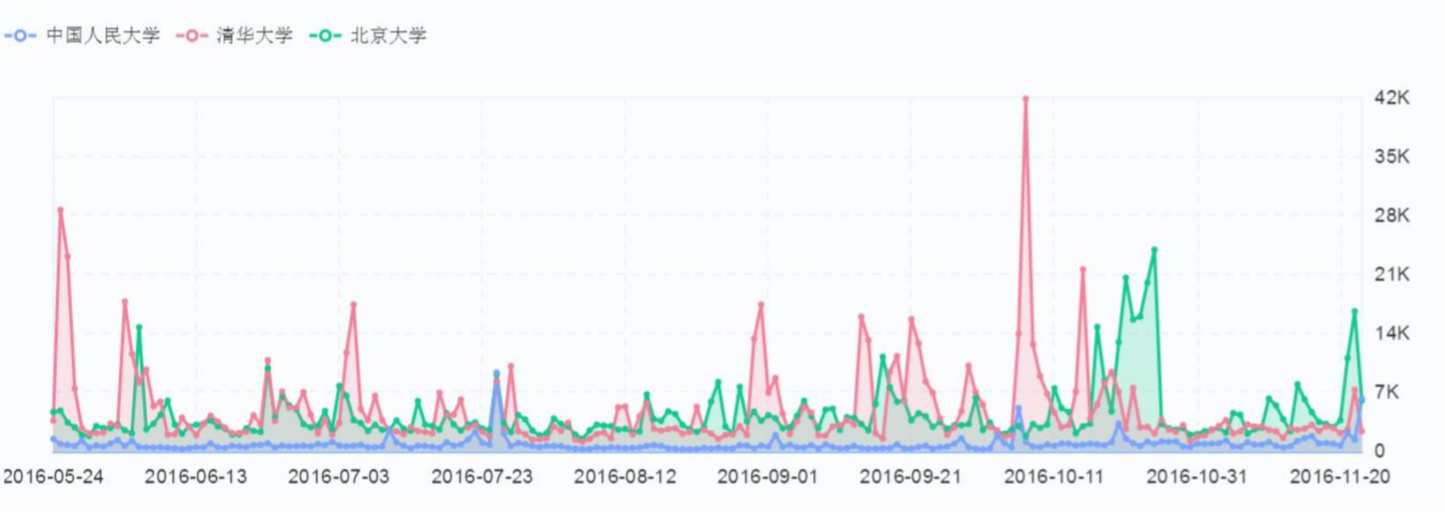
多媒体数据（音频、视频、图片）

Multimedia data

门户网站与新媒体

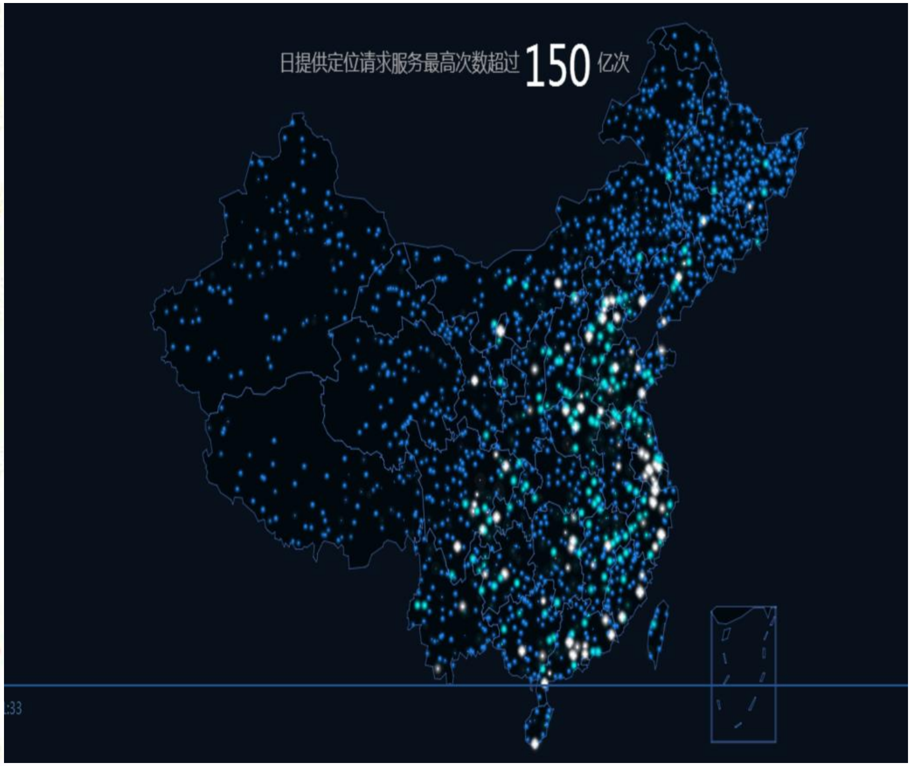
sina新浪网
观看影片
房价点评
买房不亏
图片
新闻
视频
新浪产品

微博
SK-II


中国人民大学
清华大学
北京大学
42K
35K
28K
21K
14K
7K
0
2016-05-24
2016-06-13
2016-07-03
2016-07-23
2016-08-12
2016-09-01
2016-09-21
2016-10-11
2016-10-31
2016-11-20

空间数据

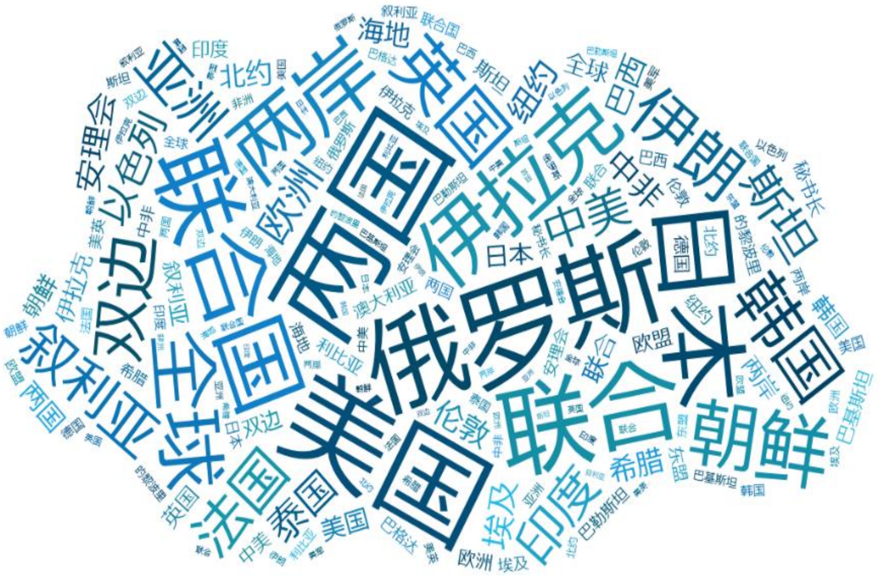
南京站
地铁1号线
四川路
许府巷
护城河
丁山花园大酒店
南京邮电大学
南京工业大学(新模范马路)
水上乐园
南京国际展览中心
莲花广场
梁洲
樱洲
模范西路
湖南路
鼓楼区
紫金城
中央路
冰凉村
河海大学出版社
北京西路
北极阁公园
北京东路
南京大学
东南大学(四牌楼校区)
南京师范大学
南京古南都饭店
中山路
玄武区
佛心桥
后宰门
广州路
乌龙潭
华侨路
珠江路
总统府
山水大酒店
南京博物院
凤凰西街
小北圩
江东村
清江村
东宝花园
花开四季
南京航空航天大学
南京博物馆
中山南路
长白街
瑞金路
明园社区
白下区
建邺区

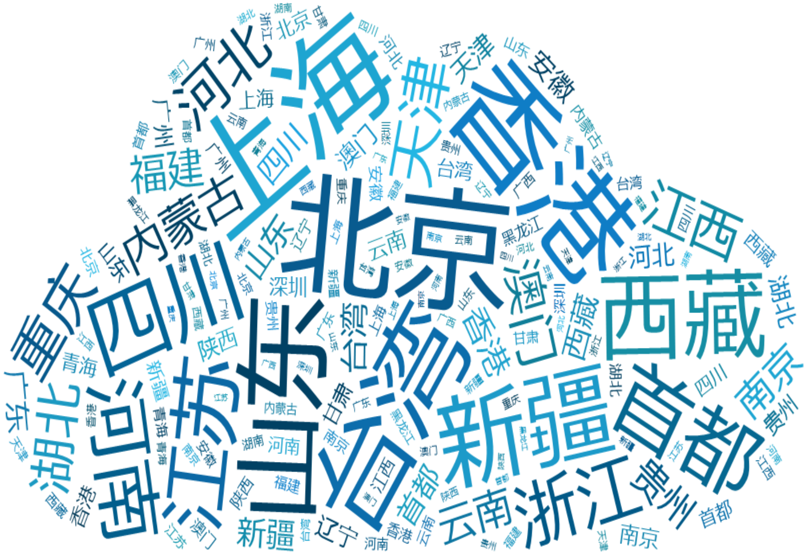

日提供定位请求服务最高次数超过 150 亿次

新闻联播中的“国际”与“国内”地域关注

国际地域关注

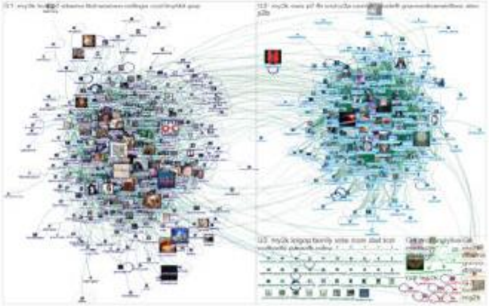
国内地域关注



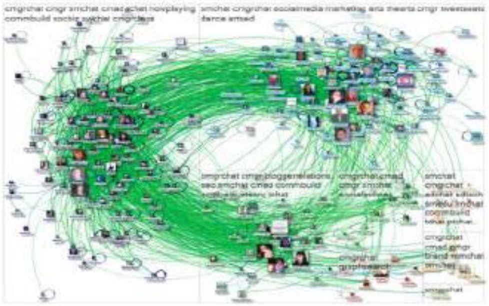


网络关系数据

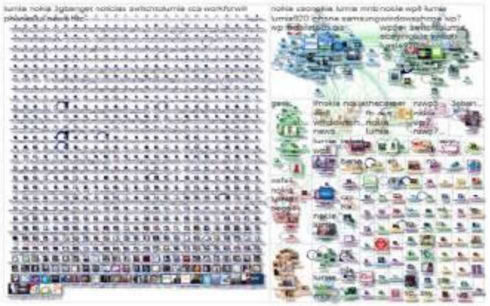
6 kinds of Twitter social media networks



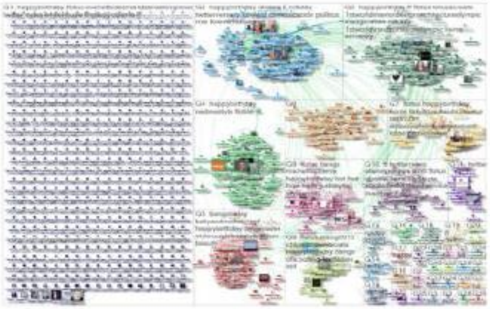
Polarized: two dense clusters with little interconnection



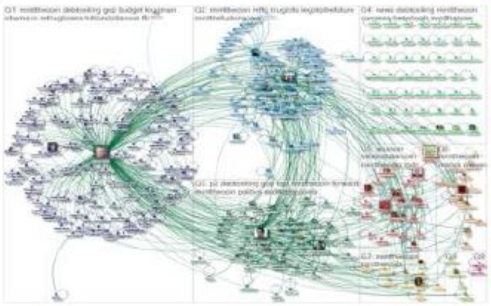
In-group: few disconnected isolates, many connections



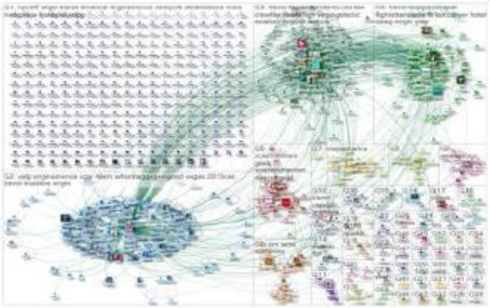
Brand/Public Topic: many disconnected isolates, some small groups



Bazaar: many medium sized groups, some isolates



Broadcast: a hub which is retweeted by many disconnected users



Support: a hub which replies to many disconnected users

计算社会科学：方法论

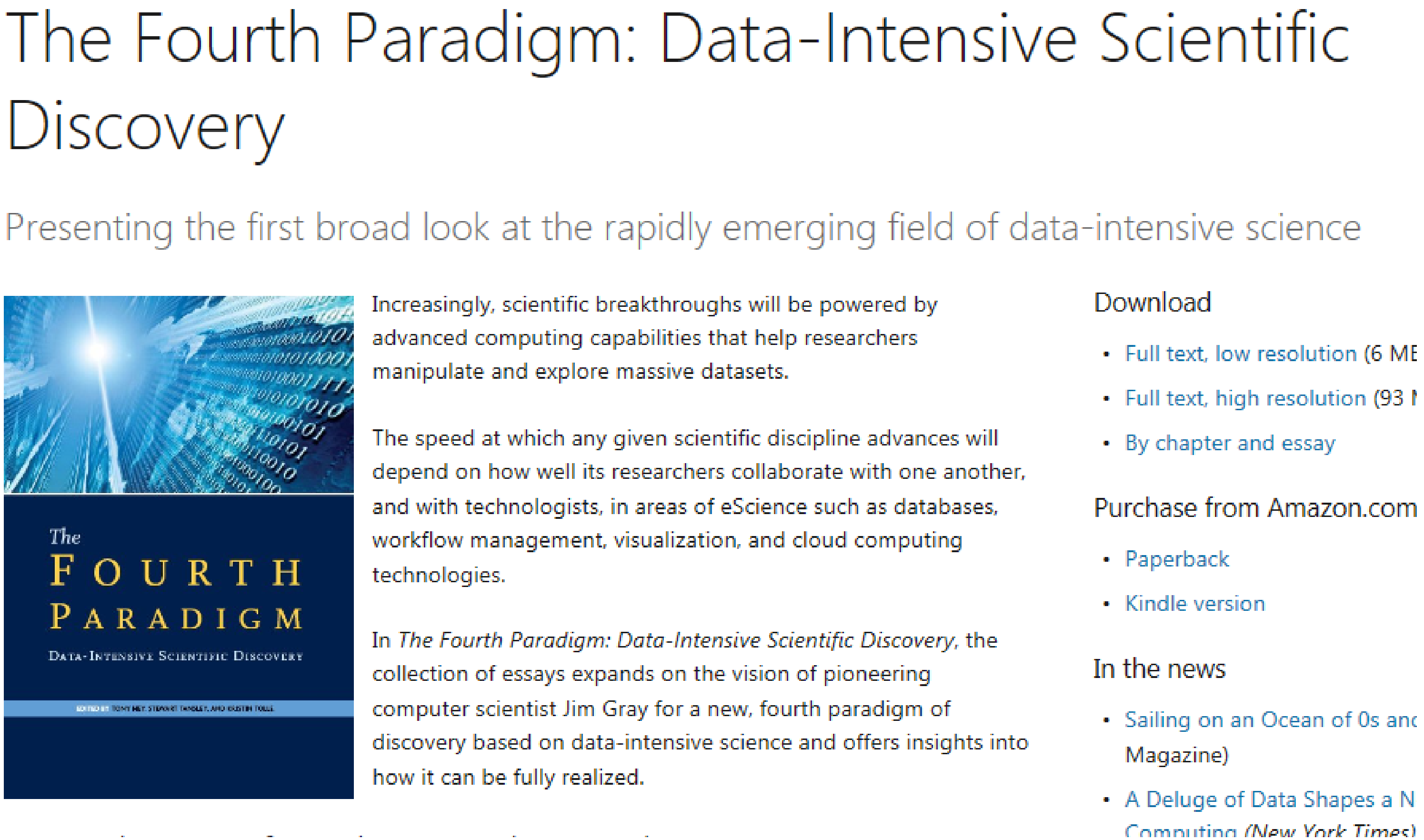
- 计算社会科学：大数据+社会科学

✓图灵奖得主J. Gray（2010）：大数据时代将形成数据密集型科学研究“第四范式”。大数据时代的科学研究将不再需要模型和假设，而是利用超级计算能力直接分析海量数据发现相关关系即可获得新知识；

✓2009 年，哈佛大学David Lazer等15 位美国学者在《Science》上联合发表了一篇具有里程碑意义的文章 “Computational Social Science”；

✓2014年，哈佛大学Gary King认为大数据方法将终结传统的定量、定性方法分野。

第四范式



The Fourth Paradigm: Data-Intensive Scientific Discovery

Presenting the first broad look at the rapidly emerging field of data-intensive science

The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.

The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud computing technologies.

In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, the collection of essays expands on the vision of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-intensive science and offers insights into how it can be fully realized.

Download

- Full text, low resolution (6 ME
- Full text, high resolution (93 I
- By chapter and essay

Purchase from Amazon.com

- Paperback
- Kindle version

In the news

- Sailing on an Ocean of 0s anc Magazine)
- A Deluge of Data Shapes a N Computing (New York Times)

第四范式

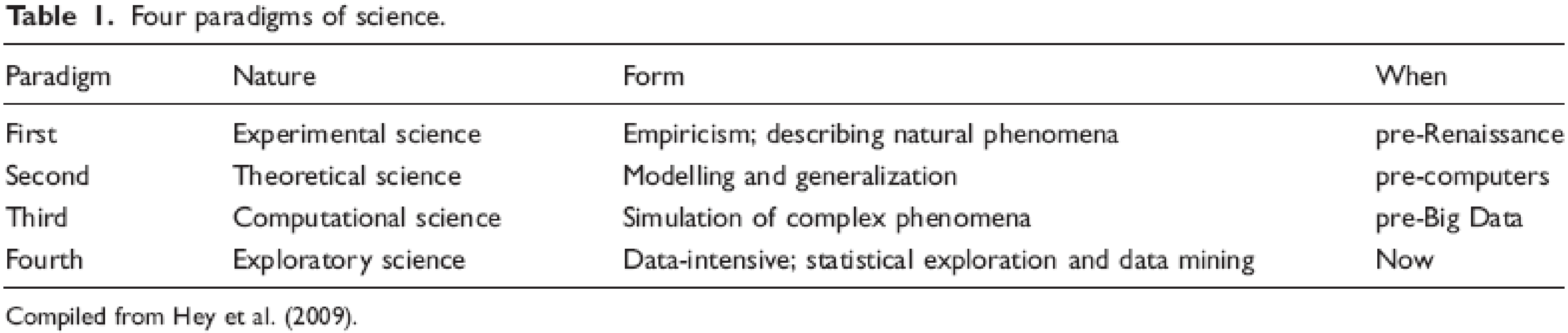


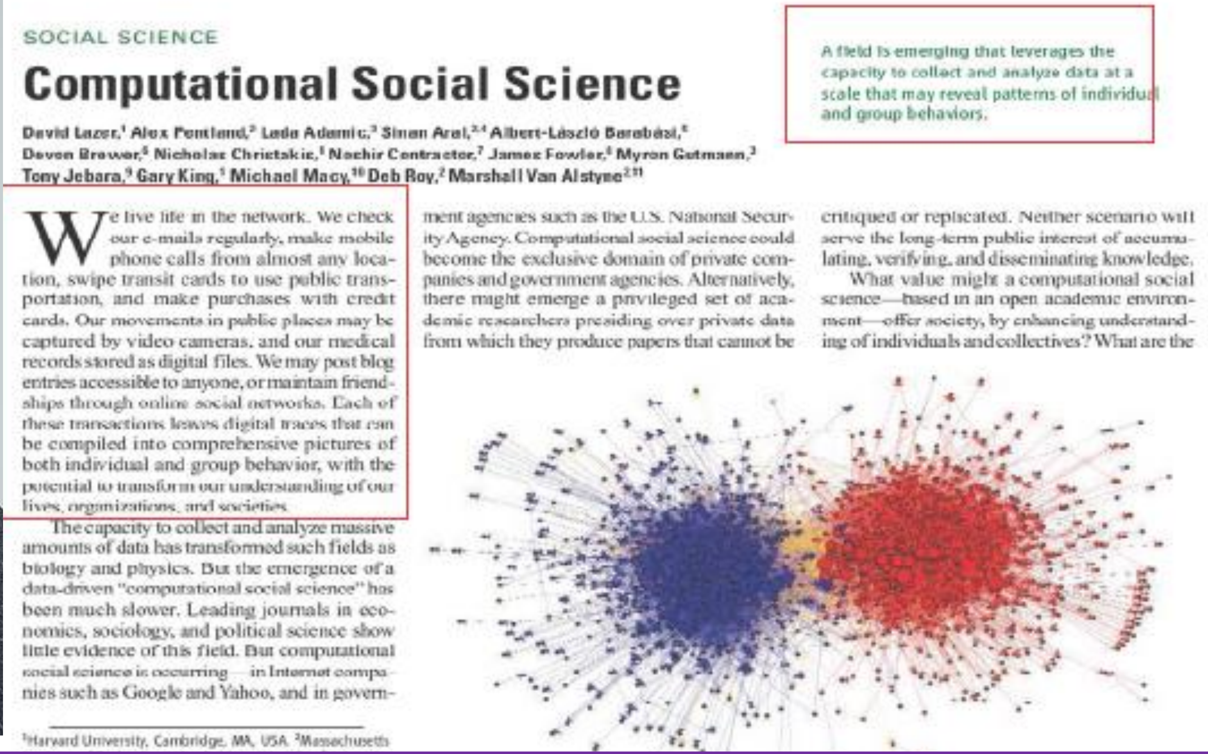
Table 1. Four paradigms of science.

Paradigm	Nature	Form	When
First	Experimental science	Empiricism; describing natural phenomena	pre-Renaissance
Second	Theoretical science	Modelling and generalization	pre-computers
Third	Computational science	Simulation of complex phenomena	pre-Big Data
Fourth	Exploratory science	Data-intensive; statistical exploration and data mining	Now

Compiled from Hey et al. (2009).

计算社会科学

- “计算社会科学”这一研究领域正在兴起，人们将在前所未有的深度和广度上自动地收集和利用数据，为社会科学的研究服务。社会计算是指社会行为和计算系统交叉融合而成的一个研究领域，研究的是如何利用计算系统帮助人们进行沟通与协作，如何利用计算技术研究社会运行的规律与发展趋势。



SOCIAL SCIENCE

Computational Social Science

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

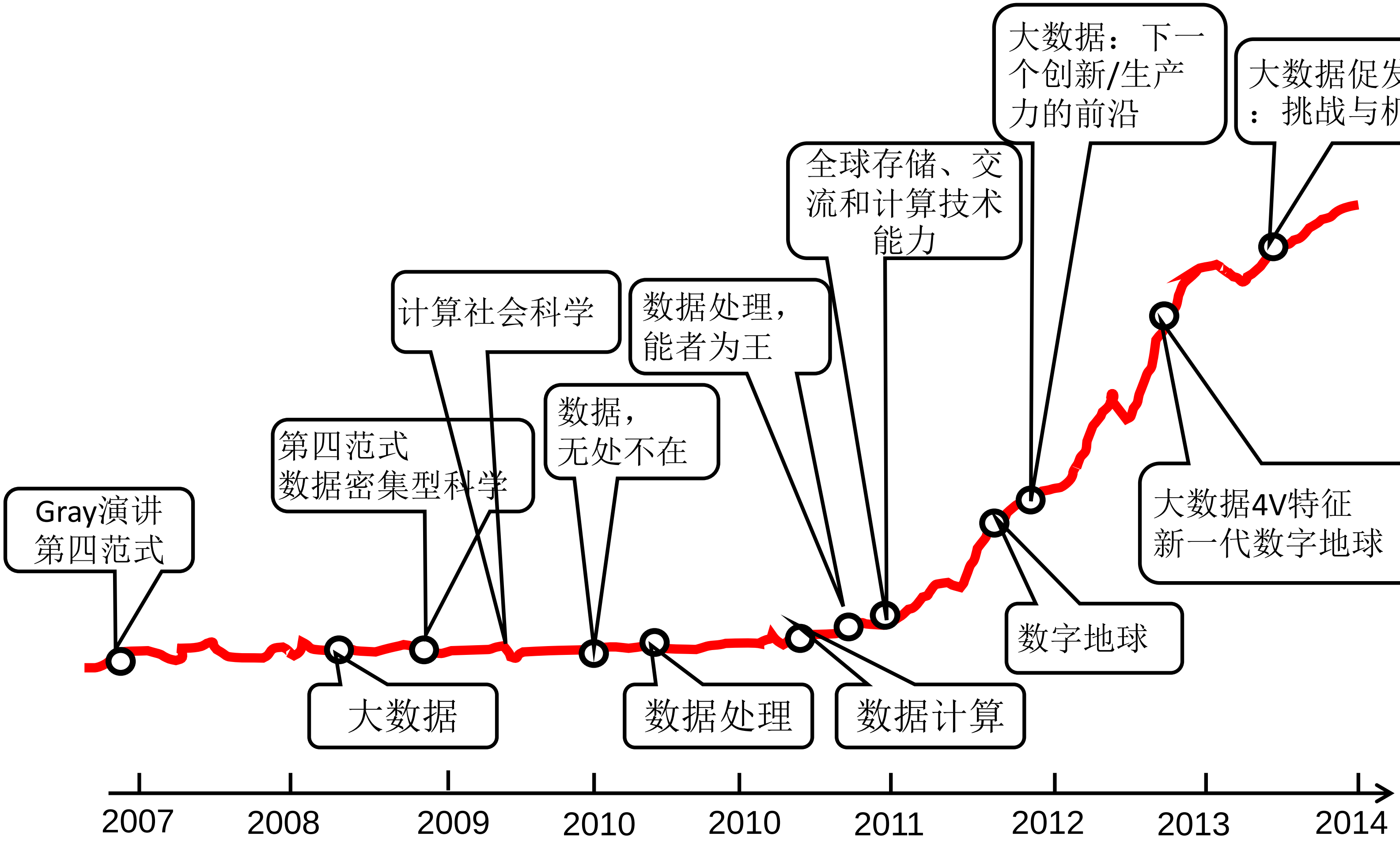
We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the

[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts

大数据的新进展


Gray演讲
第四范式
第四范式
数据密集型科学
计算社会科学
数据，
无处不在
数据处理，
能者为王
全球存储、交流和计算技术能力
大数据：下一个创新/生产力的前沿
大数据促发展：挑战与机遇
大数据4V特征
新一代数字地球
数字地球
大数据
数据处理
数据计算
2007
2008
2009
2010
2010
2011
2012
2013
2014

计算社会科学

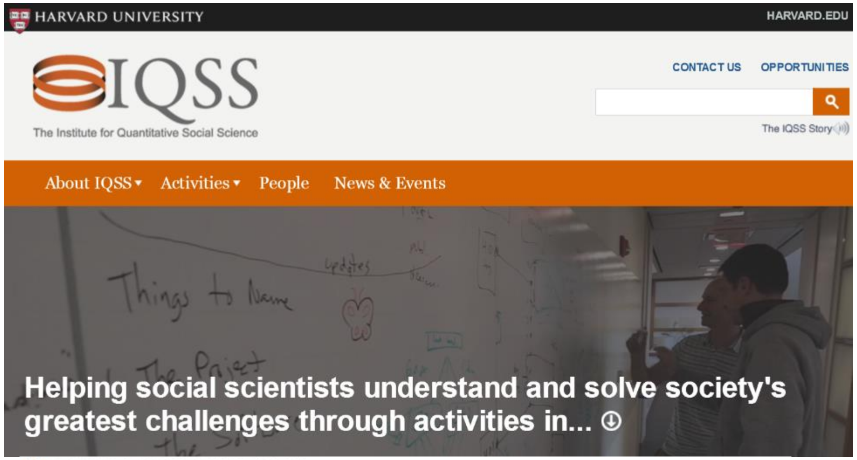
- 数据驱动的社会科学（Data-driven social science）
  - 计算社会科学寻求从海量数据挖掘中获取社会现象的有价知识
  - 综合应用统计、数学和计算科学来挖掘数据
  - 嫁接定量分析和定性分析
  - 应用导向的研究

Evaluate Public Policy Programs
Understand What Billions of Social Media Posts Say
Estimate The Cause of death in the Developing World
Ensure fair Legislative Redistricting
Reverse Engineer The Chinese Censorship Program
Forecast Presidential elections & International conflict
AT HARVARD, I DISCOVERED...
Methods of analyzing Big Data to...

大数据社会科学？方法价值与学科价值

- R. Chang（2013）：大数据带来的社会科学范式转换
- 大数据推动社会科学范式转换的过程中，技术的进步、学科间的融合、比较战略和新的数据分析技巧的使用、新的商业和组织环境将会使得这种范式的转换成为可能并加速进行。
- 涉及若干方面：研究视角涉及学科间在研究方法、研究理论及测量方法上的整合；
- 大数据：更便捷的数据收集技术；社会科学与计算科学、网络科学相结合；“计算社会科学”（computational social science）和“网络社会科学”（e-social science）的方向转变。

Big Data and Social Analytics certificate course



Helping social scientists understand and solve society's greatest challenges through activities in...

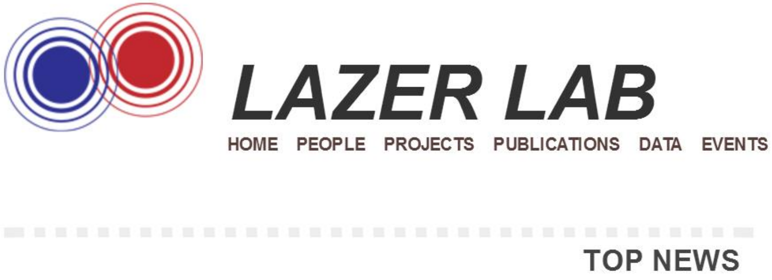
CENTER FOR STATISTICS AND MACHINE LEARNING

BIG DATA SOCIAL SCIENCE

Integrative Graduate Education and Research Traineeship at Pennsylvania State University

BDSS IGERT Speaker & Event Series - "Speed Dating and Matchmaking"

BDSS-IGERT Welcomes Its Fifth



Volunteer Science

Volunteer Science offers people around the world to become involved in scientific projects on everything from... more

WZB
Berlin Social Science Center
Home | Publications | WZB-Mitteilungen
Publications

Big Data in the social sciences

Alexandros Tokhi and Christian Rauh

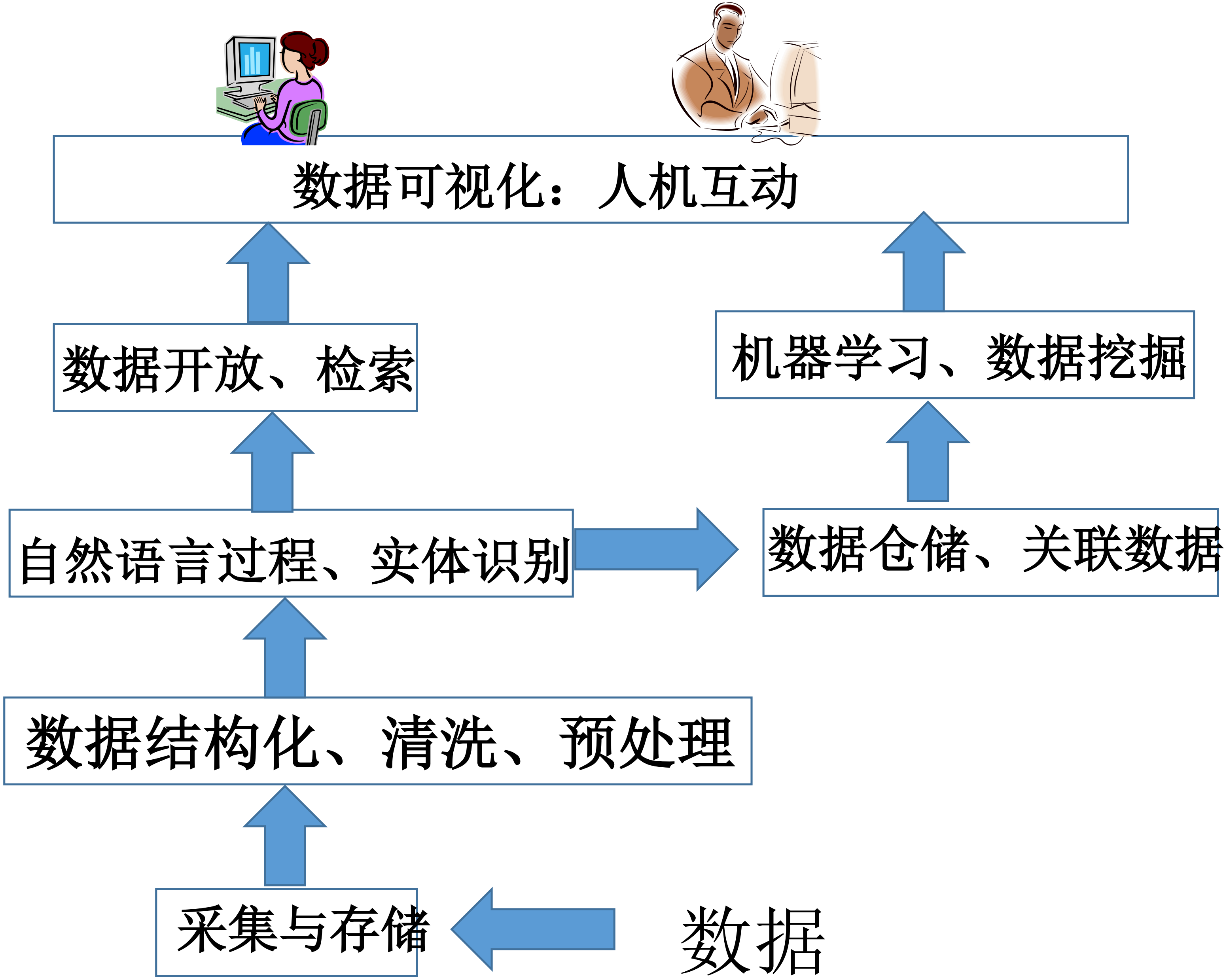
Big Data has become an ubiquitous buzzword. Social networks, smartphones, and various online applications and websites constantly produce and provide information on an unprecedented scale and level of detail. Data storage is cheap and ever-improving analytical

计算社会科学：研究方法

- ➢网络爬虫
- ➢对搜索引擎搜索记录的分析
- ➢自动文本分析
- ➢视频/图片分析
- ➢社会网络分析
- ➢空间/时间分析
- ➢可视化

- 机器学习
- 自然语言过程
- 统计分析

大数据分析全过程



数据可视化：人机互动

数据开放、检索

机器学习、数据挖掘

自然语言过程、实体识别

数据仓储、关联数据

数据结构化、清洗、预处理

采集与存储

数据

Big Data in Political Science

- 文本分析

Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate
Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict
The Genealogy of Law
Finding Jumps in Otherwise Smooth Curves: Identifying Critical Events in Political Processes

- 大规模实验研究

Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk

- **GIS**与空间分析

Reaching Migrants in Survey Research: The Use of the Global Positioning System to Reduce Coverage Bias in China

- 网络分析

Inferential Network Analysis with Exponential Graph Models

- 计算机算法

An Introduction to Bayesian Inference via Variational Approximations
Improving Predictions Using Ensemble Bayesian Model Averaging
Bayesian Metric Multidimensional Scaling

社科研究中的大数据：方法功能

- 作为研究方法的大数据分析

➢数据获取与管理

➢测量工具

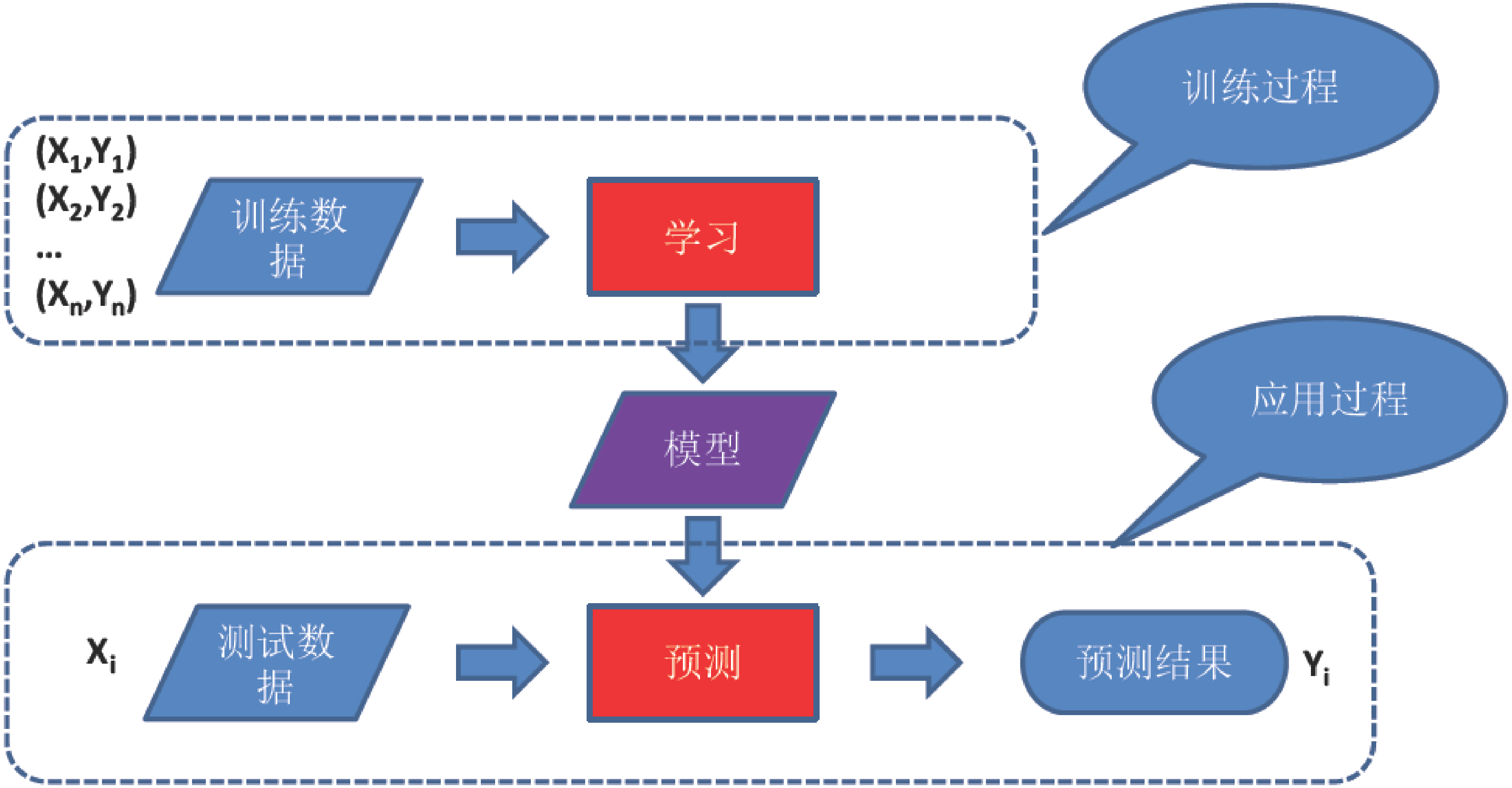
➢分类与聚类

➢关联分析

➢因果推论（回归分析）

➢信息呈现（可视化）

机器学习与预测分析



机器学习

任务

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]
- Collaborative Filter [Predictive]

算法

- 分类（贝叶斯分类器）
- *k* 临近
- 回归
- 聚类
- LDA
- 神经网络
- 支持向量机
- 决策树

Big Data in Political Science

- 文本分析

Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate
Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict
The Genealogy of Law
Finding Jumps in Otherwise Smooth Curves: Identifying Critical Events in Political Processes

- 大规模实验研究

Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk

- **GIS**与空间分析

Reaching Migrants in Survey Research: The Use of the Global Positioning System to Reduce Coverage Bias in China

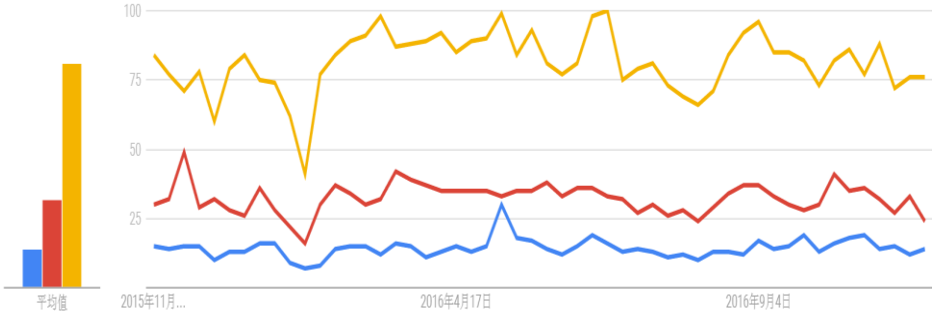
- 网络分析

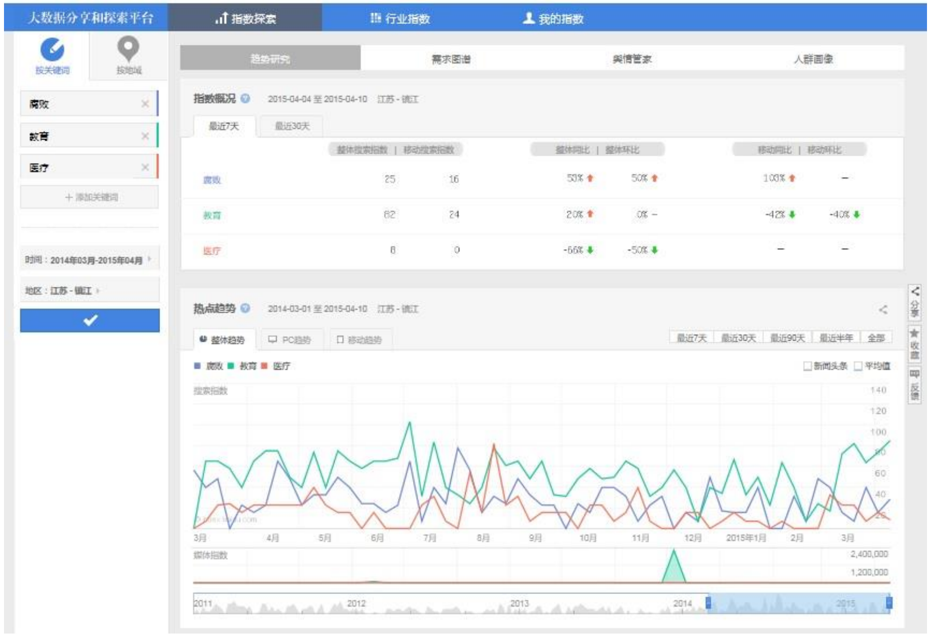
Inferential Network Analysis with Exponential Graph Models

- 计算机算法

An Introduction to Bayesian Inference via Variational Approximations
Improving Predictions Using Ensemble Bayesian Model Averaging
Bayesian Metric Multidimensional Scaling

搜索指数：Google Trends、百度指数





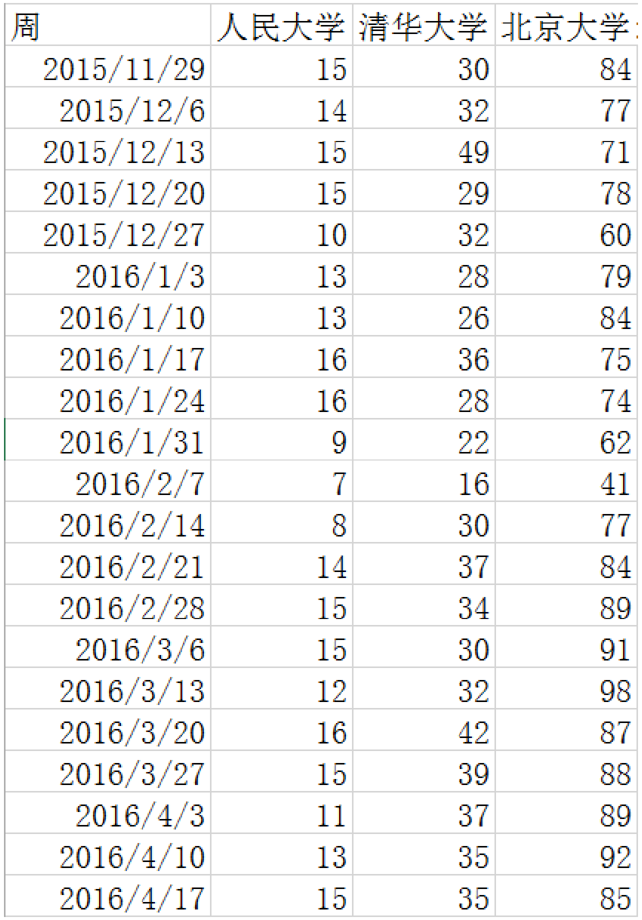
时间：2014年03月-2015年04月

地区：江苏 - 镇江

指数概况 2015-04-04 至 2015-04-10 江苏 - 镇江

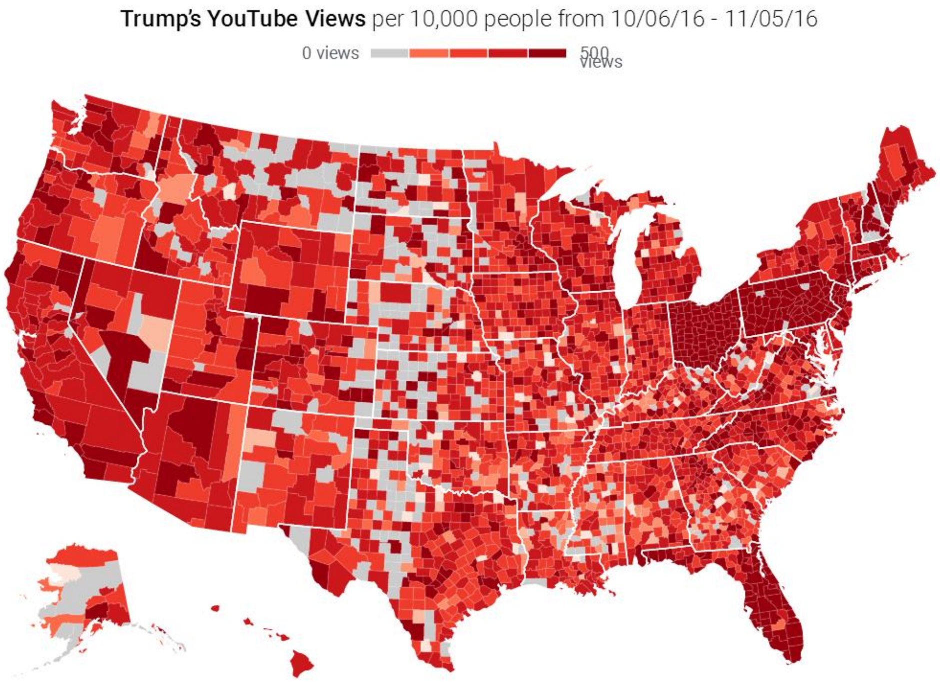
	整体搜索指数	移动搜索指数	整体同比	整体环比	移动同比	移动环比
腐败	25	16	50% ↑	50% ↑	100% ↑	–
教育	62	24	20% ↑	0% –	-42% ↓	-40% ↓
医疗	8	0	-56% ↓	-50% ↓	–	–

热点趋势 2014-03-01 至 2015-04-10 江苏 - 镇江

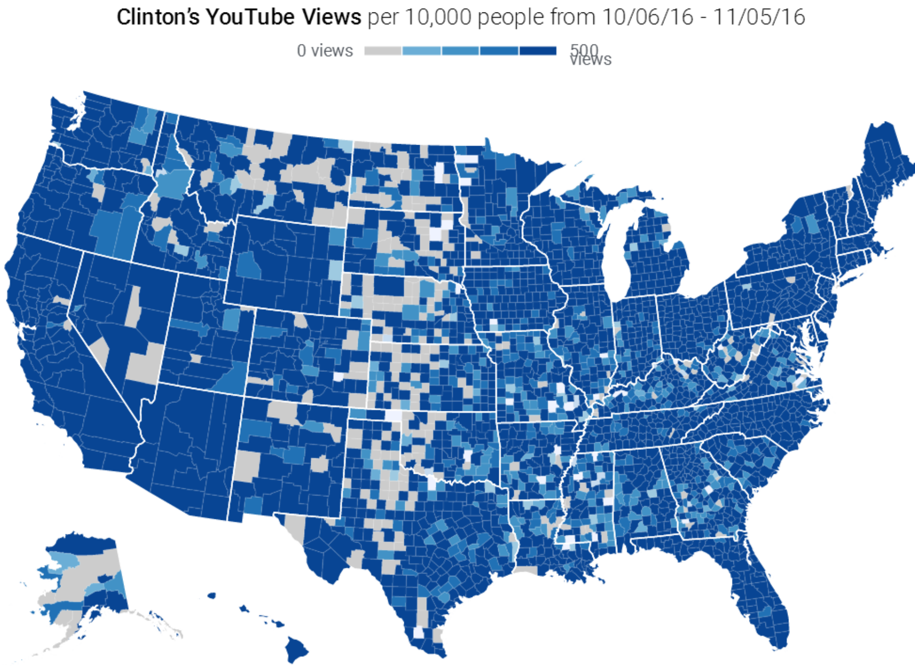


周	人民大学	清华大学	北京大学
2015/11/29	15	30	84
2015/12/6	14	32	77
2015/12/13	15	49	71
2015/12/20	15	29	78
2015/12/27	10	32	60
2016/1/3	13	28	79
2016/1/10	13	26	84
2016/1/17	16	36	75
2016/1/24	16	28	74
2016/1/31	9	22	62
2016/2/7	7	16	41
2016/2/14	8	30	77
2016/2/21	14	37	84
2016/2/28	15	34	89
2016/3/6	15	30	91
2016/3/13	12	32	98
2016/3/20	16	42	87
2016/3/27	15	39	88
2016/4/3	11	37	89
2016/4/10	13	35	92
2016/4/17	15	35	85

视频分析：Youtobe选情

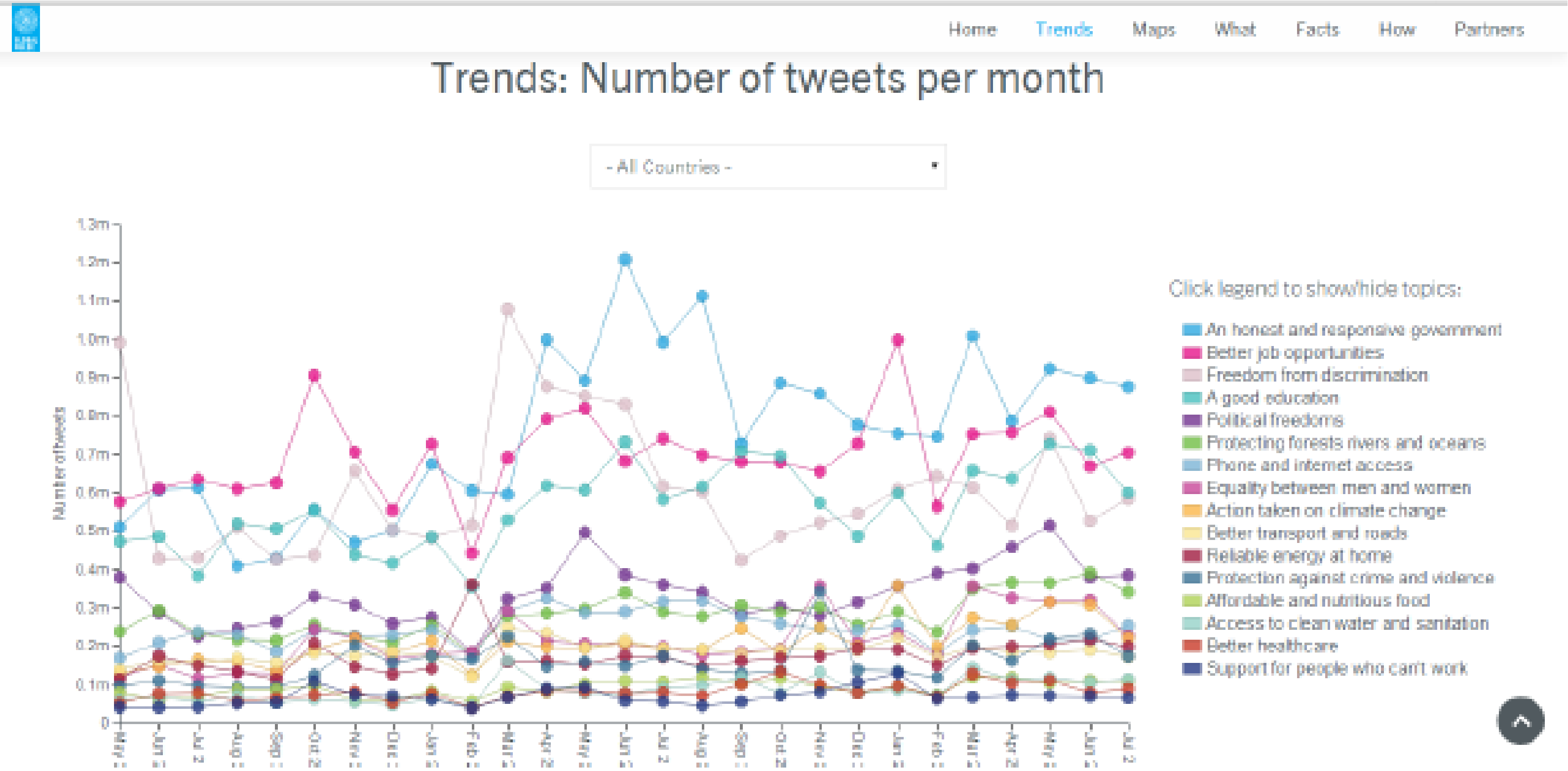


Trump's YouTube Views per 10,000 people from 10/06/16 - 11/05/16



Clinton's YouTube Views per 10,000 people from 10/06/16 - 11/05/16

文本挖掘：Global Pulse

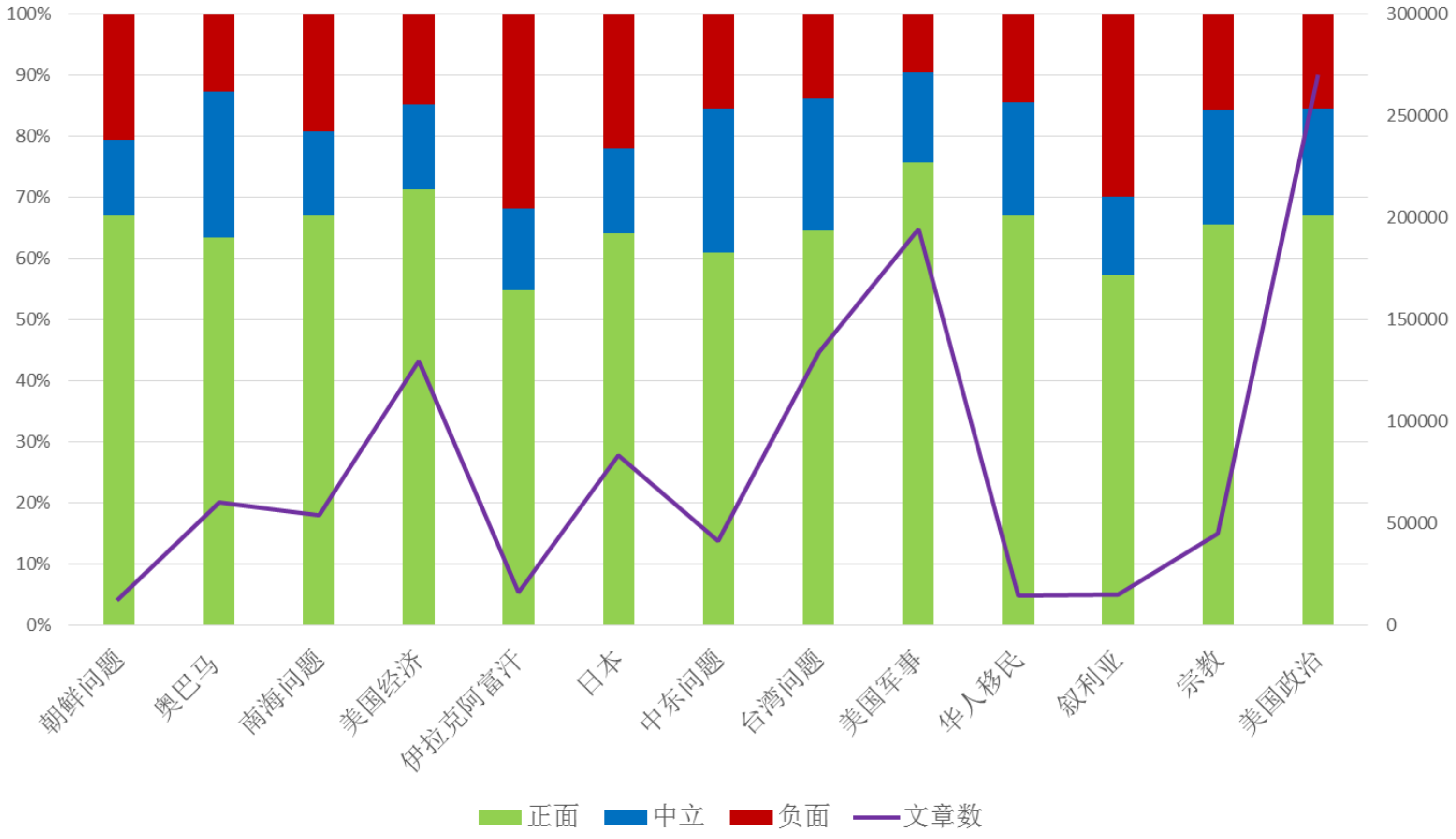


Home
Trends
Maps
What
Facts
How
Partners

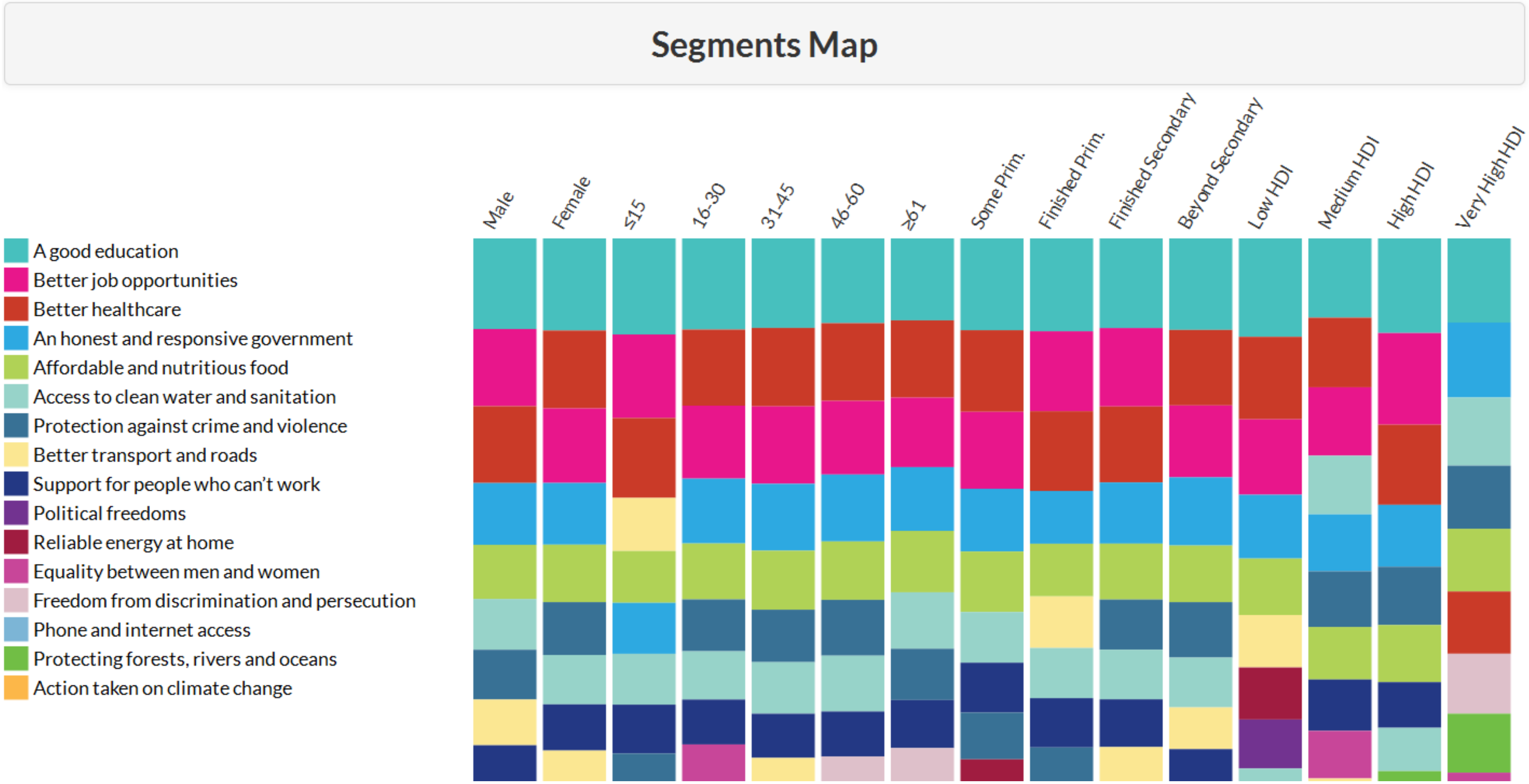
Trends: Number of tweets per month

- All Countries -
Number of tweets
1.3m
1.2m
1.1m
1.0m
0.9m
0.8m
0.7m
0.6m
0.5m
0.4m
0.3m
0.2m
0.1m
0
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Click legend to show/hide topics:
An honest and responsive government
Better job opportunities
Freedom from discrimination
A good education
Political freedoms
Protecting forests rivers and oceans
Phone and internet access
Equality between men and women
Action taken on climate change
Better transport and roads
Reliable energy at home
Protection against crime and violence
Affordable and nutritious food
Access to clean water and sanitation
Better healthcare
Support for people who can't work

文本挖掘：中国人的美国观

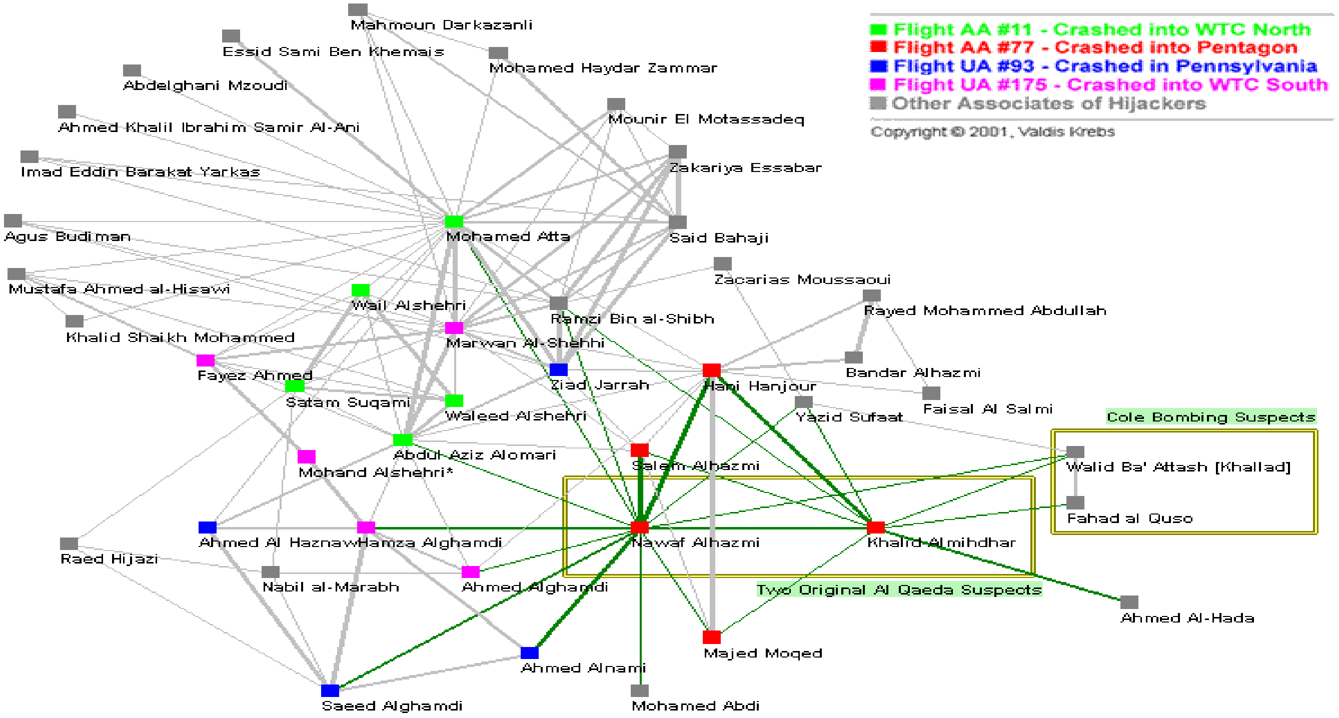


Myworld



Segments Map

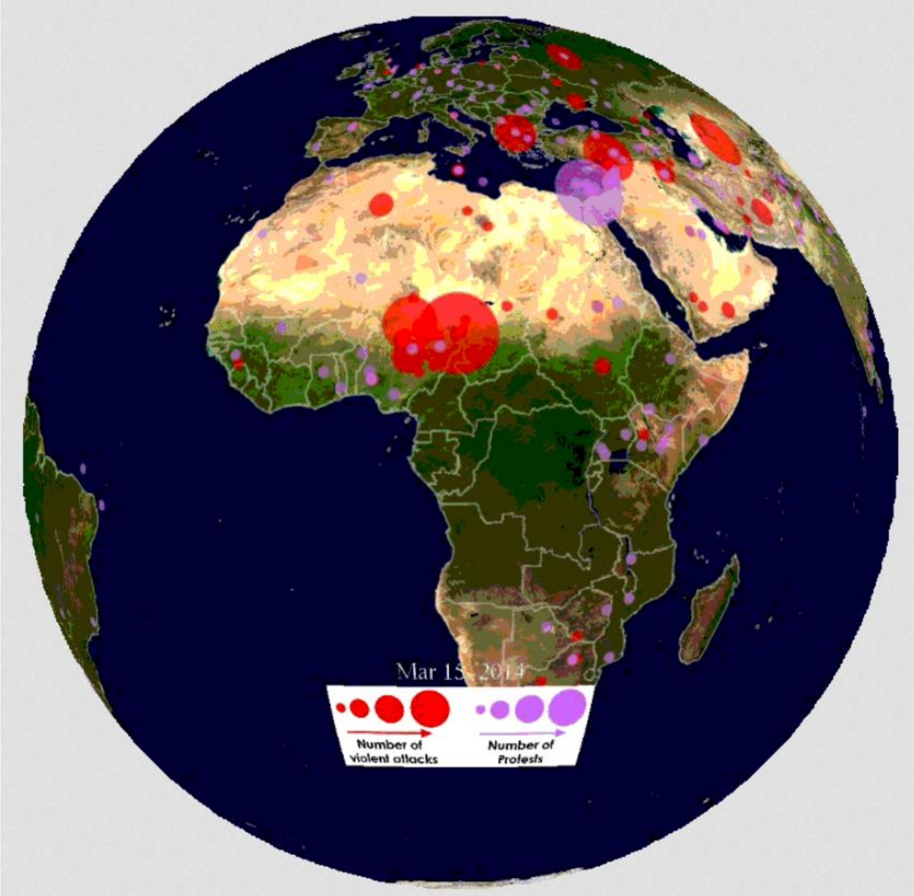
社交网络分析：9/11 基地组织网络

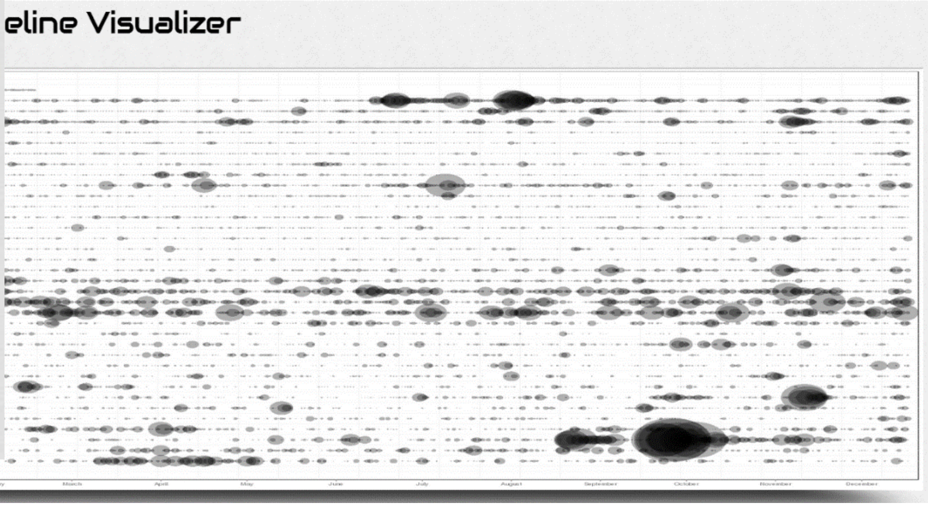


Copyright © 2001, Valdis Krebs

Figure 3 - All Nodes within 2 steps / degrees of original suspects

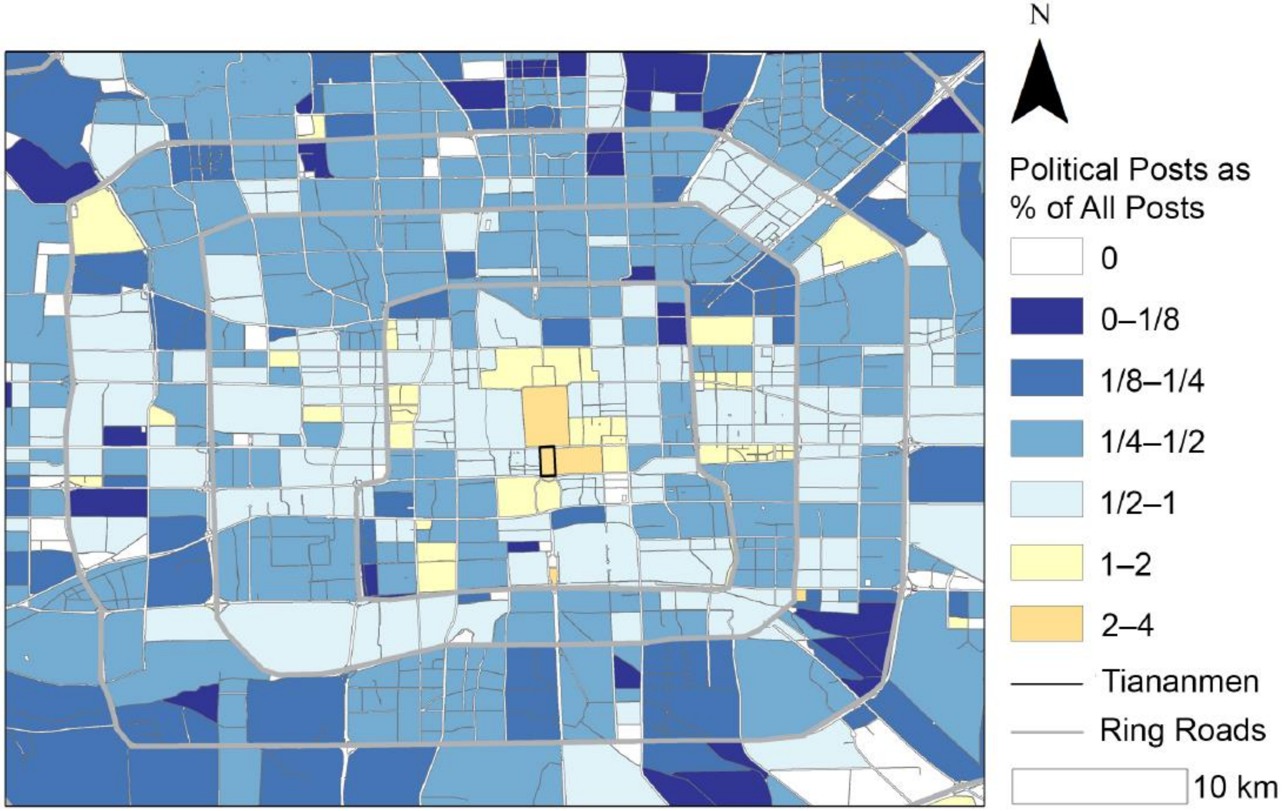
时间序列分析：全球暴力活动（GDELT）



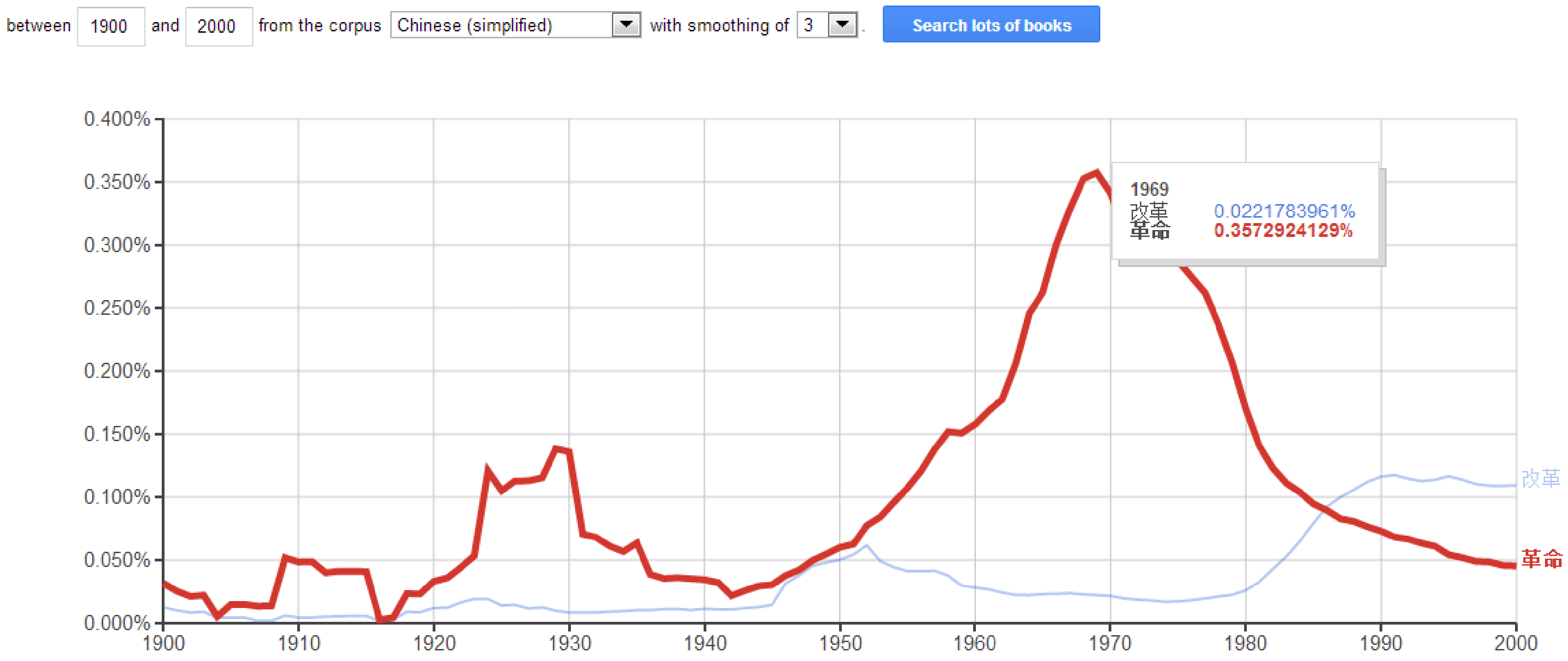


空间分析：空间与政治话题

Figure 2. Spatial Distribution of Political Posts, Beijing, 1 July 2014–15 June 2015



开源大数据分析工具：Google books



between 1900 and 2000 from the corpus Chinese (simplified) with smoothing of 3.

Search lots of books

计算社会科学：研究议题

学术性研究

- 政治传播
- 战争、反恐与社会运动
- 选举和投票
- 议会（精英）政治
- 国际事务与外交
- 公共外交(软实力)
- 公共政策
- 政治学方法论

应用性研究

- 公共卫生
- 公共安全与社会治安
- 交通治理
- 反腐败
- 网络舆情
- 反恐与应急管理
- ……

Bright, J. (2012). The Dynamics of Parliamentary Discourse in the UK: 1936-2011.

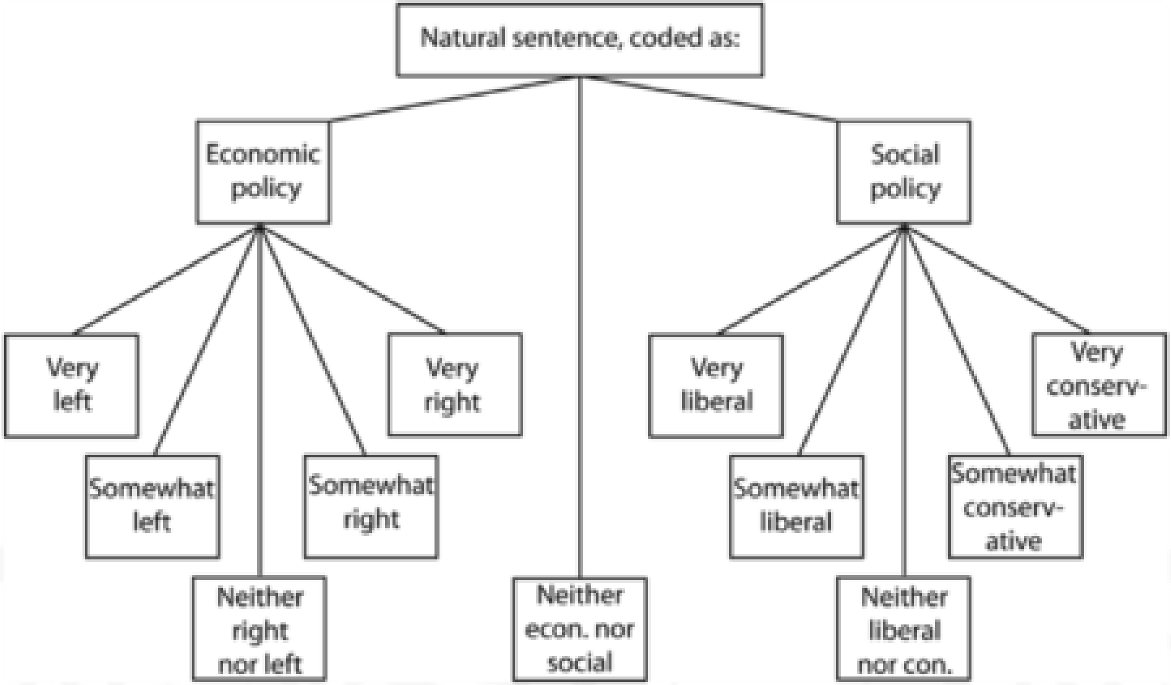
- 利用议会文本记录分析英国议会中议会争论的发展特点：
- 利用英国议会解析网站提供的议会资料构建了1936-2011年间英国国会下院发布的由7.4亿单词所构成的数据库
- 利用自动编码技术对法律、国防、环境、卫生、就业、权利、教育、农业、经济等关键词在75年间的出现频率进行了描绘
- 发现这些关键词的出现频率具有一定稳定性，但也存在很大变化。例如，争论变得更加激烈；环境议题更为突出，而农业等议题则逐渐衰落；女性议员倾向于较长的发言时间；而贵族议员被打断的频率更高一些。

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases.

- 利用文本分析方法研究美国参议员与选民的政治沟通；
- 采用了贝叶斯分层主题模型；
- 搜集美国参议院2007年来发布的24000余份新闻通告，利用无监督学习法（unsupervised learning methods）进行文本分析；
- 发现：每个参议员的关注重点与其他参议员的关注事件之间存在显著相关性；关注重点的地域分布具有一定的集聚性；议员对挪用的关注程度与他们对禁止挪用法案的投票呈现出正相关关系。

Benoit et al. (2016). “**Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data”. American Political Science Review.**

FIGURE 1. Hierarchical Coding Scheme for Two Policy Domains with Ordinal Positioning



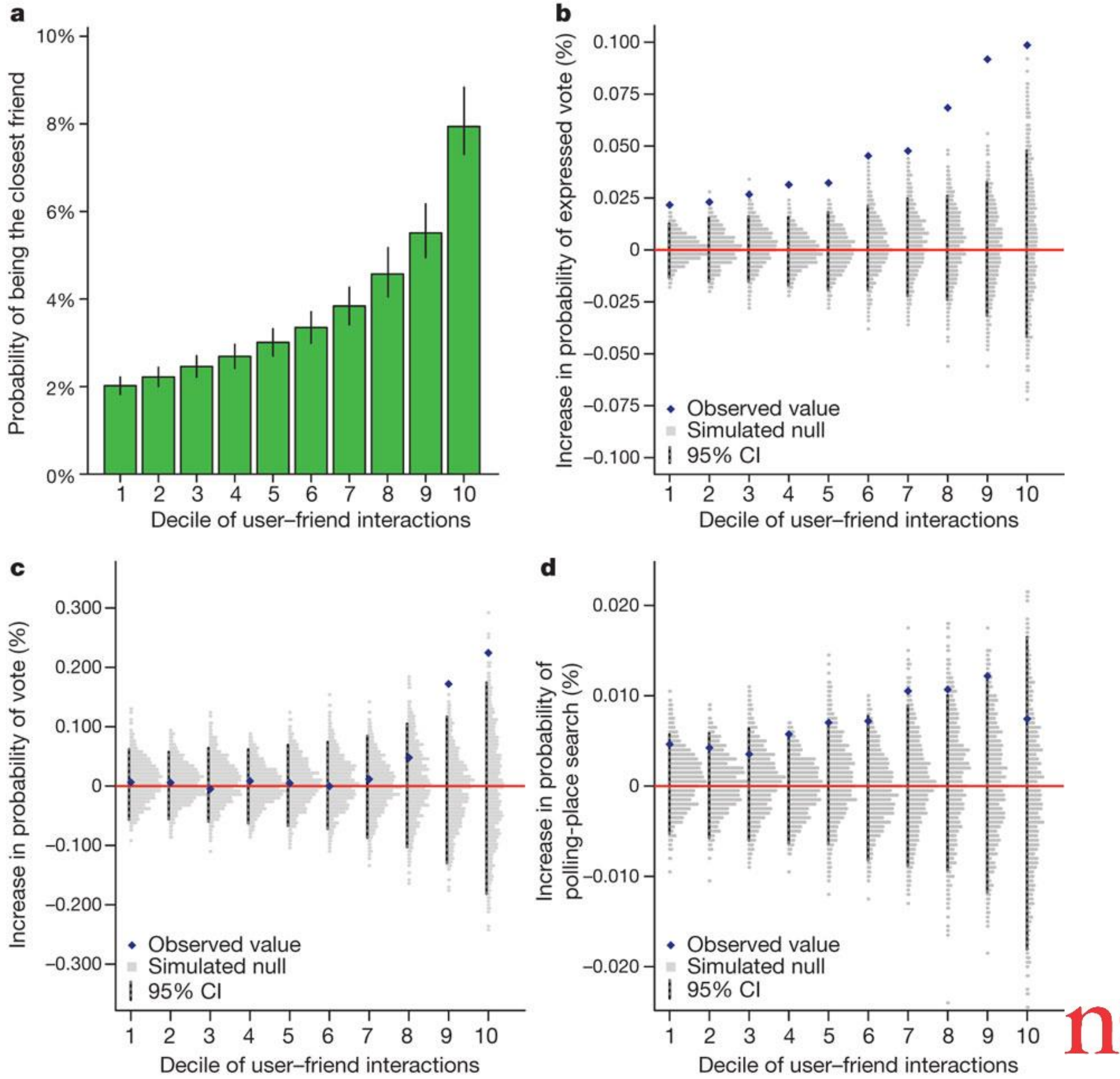
English etc al. (2011). "YouTube-ification of Political Talk: An Examination of Persuasion Appeals in Viral Video."*American Behavioral Scientist*.

- 2008年，美国网民在YouTube上观看140亿个视频
- 2008总统竞选中YouTube视频成为三大最受欢迎的政治动员
- 通过三组后测实验发现，不同的政治游说叙事有迥异效果：精神说服最有效，然后是理性逻辑说服，效果最弱的是悲情说服
- YouTube用户会抵抗被情绪所动摇，更为关注信息源可信度

Robert M. Bond, et al, "A 61-million-person experiment in social influence and political mobilization", *Nature,* Vol.489, No.7415, 2012, pp.295-298.

- Bond等（2012）比较了线上社交网络和面对面社交网络影响政治行为的路径。
- 2010年美国国会大选时对6100万Facebook用户实施发送政治动员消息的随机控制实验；
- 政治动员消息直接影响网民的政治自我表达、信息搜寻和现实投票行为;
- 政治动员消息不仅影响了接受者，还影响了接受者的网友、网友的网友，而这种社会传递效应对投票行为的影响要强于直接动员效应；
- 信息传播更容易发生在具有见面关系的关系密切的朋友中。表明强关系有助于社交网络中对于在线和现实生活中的政治动员。

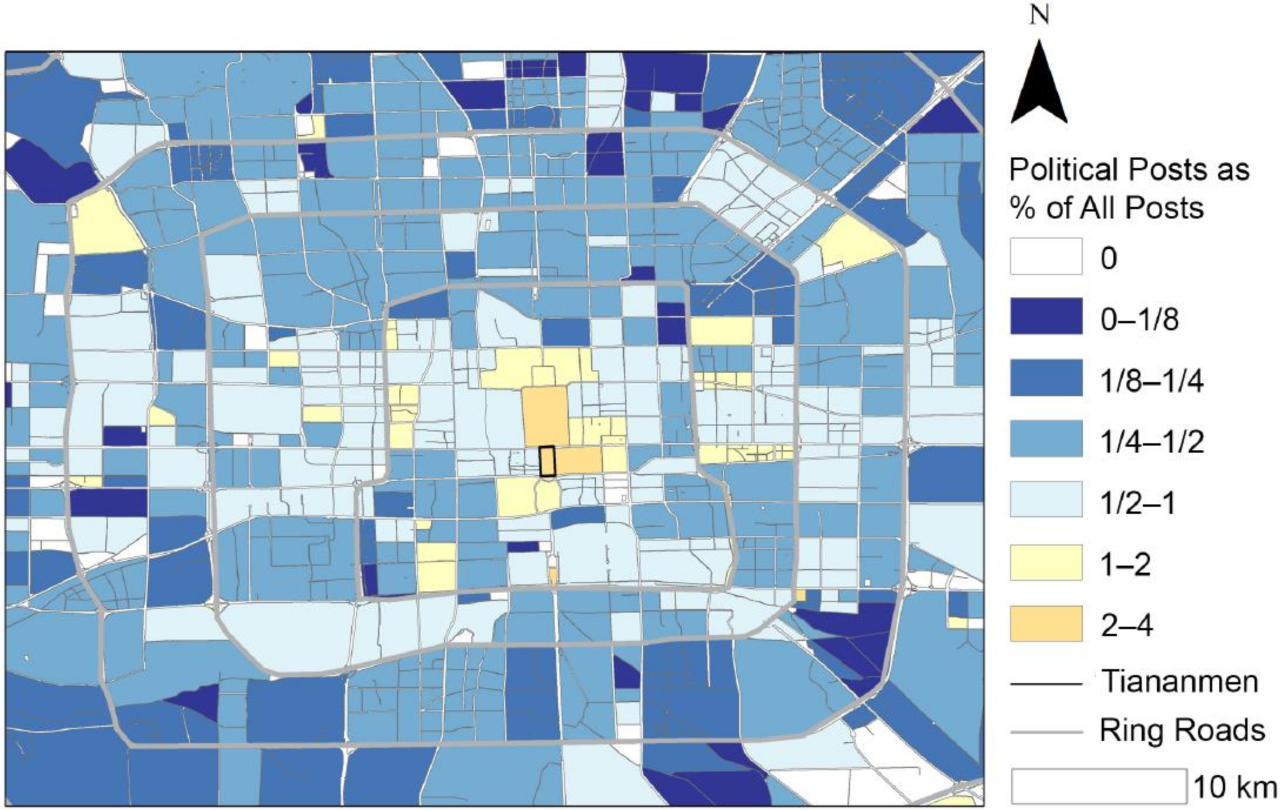
The effect of mobilization treatment that a friend received on a user’s behaviour.



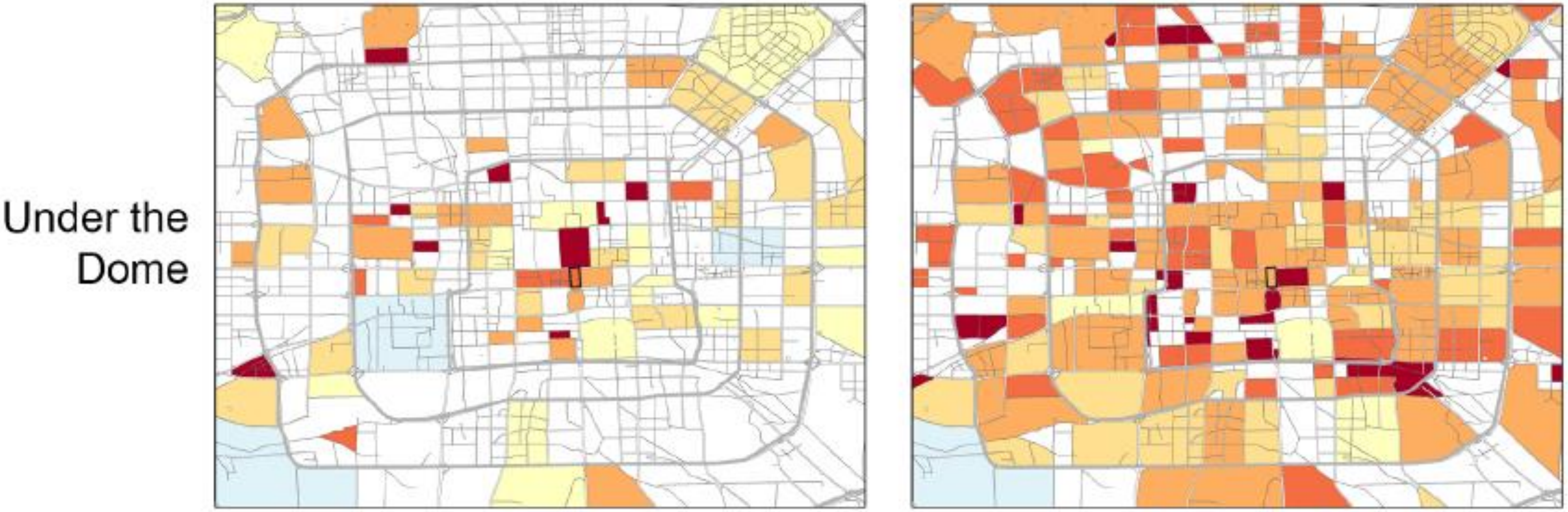
RM Bond et al. Nature **489**, 295-298 (2012) doi:10.1038/nature11421

空间分析：空间与政治话题

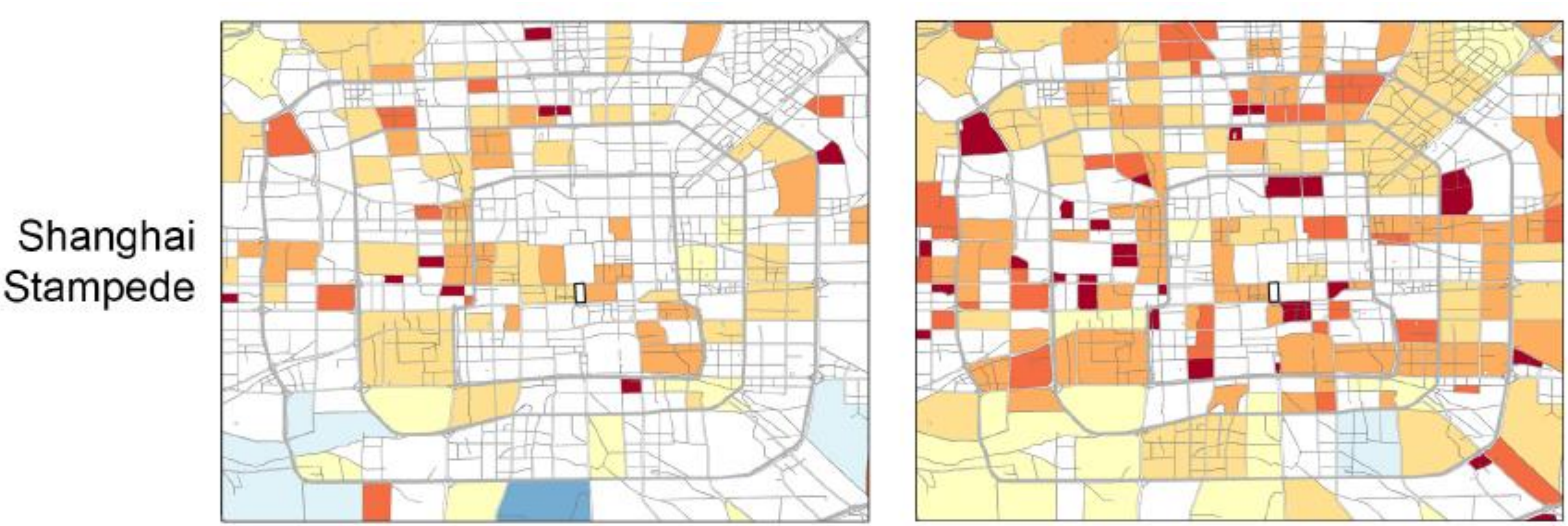
Figure 2. Spatial Distribution of Political Posts, Beijing, 1 July 2014–15 June 2015

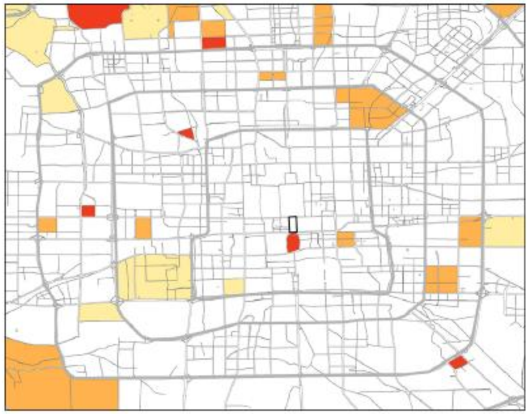


《穹顶之下》播出前后

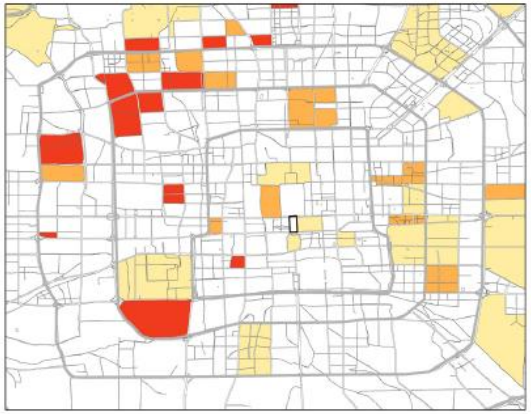


上海踩踏事件

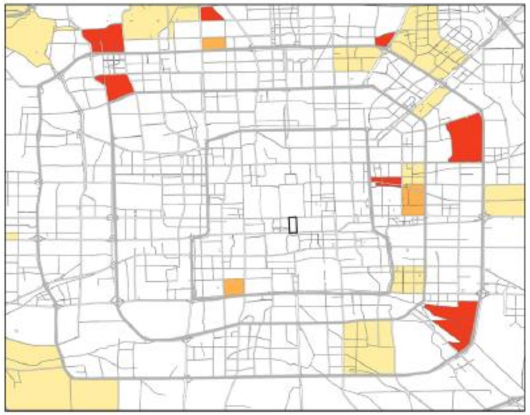




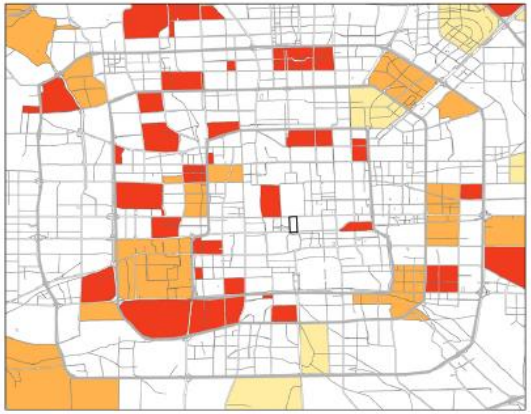
Shanghai Stampede



Under the Dome



Nanjing Massacre Memorial Day



Mega-Tiger Zhou Investigation

大数据方法：支持VS批评

- 大数据方法的优势

➢数据

➢方法

➢经济性/可行性

➢影响

大数据方法的优势

➢数据优势

✓数据模态多元化（结构化数据——非结构化数据）

✓“全量数据”而不是“样本数据”

✓“真实数据”而不是“设计的数据”

✓“大样本数据”为小概率事件分析提供可能

✓数据蕴含丰富的时空信息（Spatial and Time Dynamics）

大数据方法的优势

➢方法优势

✓更为丰富的方法工具箱

✓为定性/定量方法分野创建桥梁

✓机器学习提升分析效率

✓探究相关关系

✓预测能力和方法

大数据方法的优势

➢经济性/可行性优势

✓低成本

✓时效性

✓高效率

大数据方法的优势

➢学术影响优势

✓与互联网无缝对接

✓社科知识平民化普及

✓沟通“象牙塔”与“普罗大众”

✓良好的组织环境+充分的社会需求

大数据方法的批评

- 大数据方法的局限性

➢数据

➢方法

➢可行性

➢伦理

大数据方法的批评

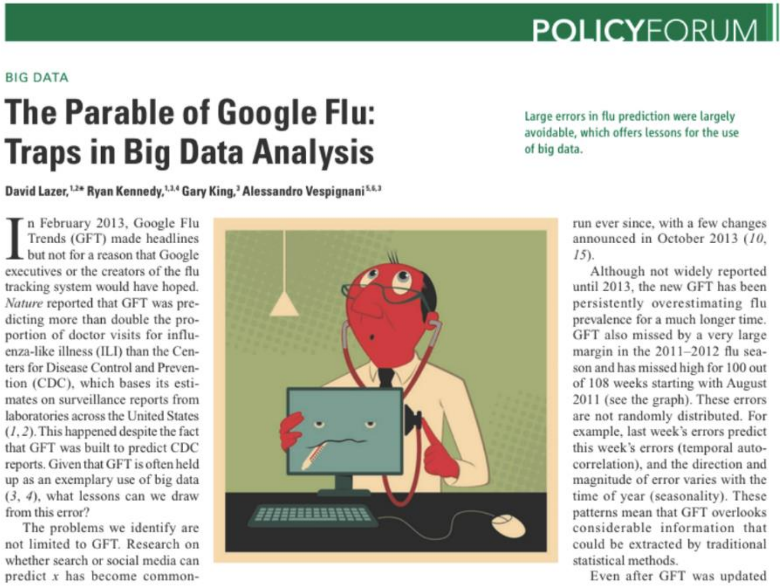
➢数据-批评

✓“有偏数据”而非“全量数据”

✓“行为记录”难以替代态度价值数据

✓“大数据的傲慢”——测量信度和效度

✓网络“伪信息”（虚假信息、机器人、谣言）



POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become common-

run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated

大数据方法的批评

➢方法-批评

✓探究“相关关系”而非“因果关系”

✓机器学习的效度

✓技术门槛高

✓算法不透明

✓大数据方法“嫁接”统计方法

大数据方法的批评

- ➢可行性-批评
- ✓数据开放程度低
- ✓存在技术壁垒
- ✓软硬件要求高

大数据方法的批评

➢伦理-批评

✓个人隐私保护

✓数据（所有、利用和处置等）权利保障

✓社会实验的伦理困境

Q&A