计算社会科学新进展
——方法与应用

孟天广 副教授

清华大学计算社会科学平台 执行主任

清华大学社会科学学院 院长助理

2019年12月12日

内容提要

大数据+社会科学

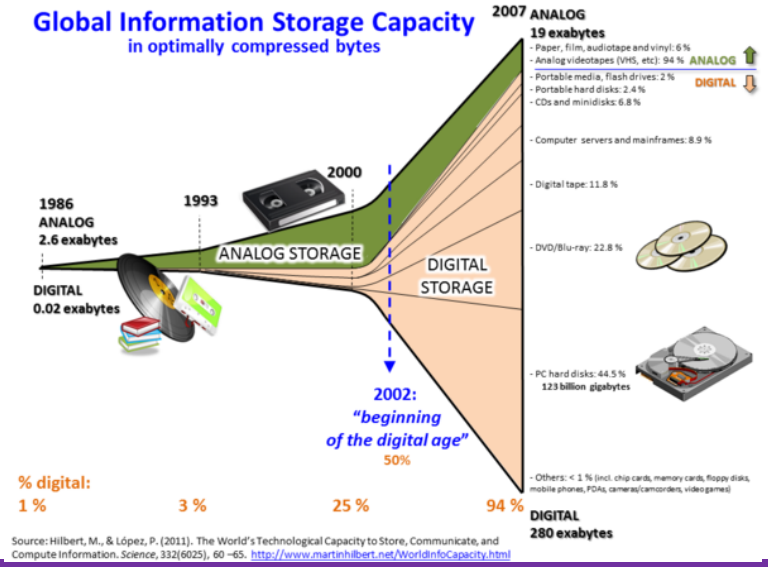
计算社会科学：方法论

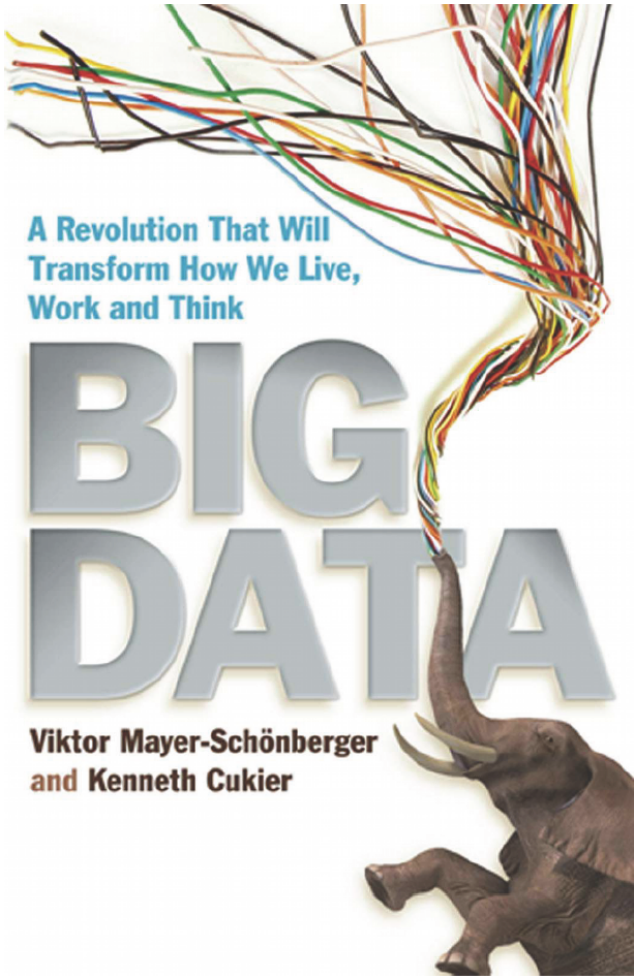
计算社会科学：方法与议题

计算社会科学：前沿应用

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”；





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

What is Big Data

- Data
- Analytics
- ✓Big data is not about Data!
- Industry
- Solution

Why is Big Deal

- Government
- Private Sector
- Science
- Social Science Revolution

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
- 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
- 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

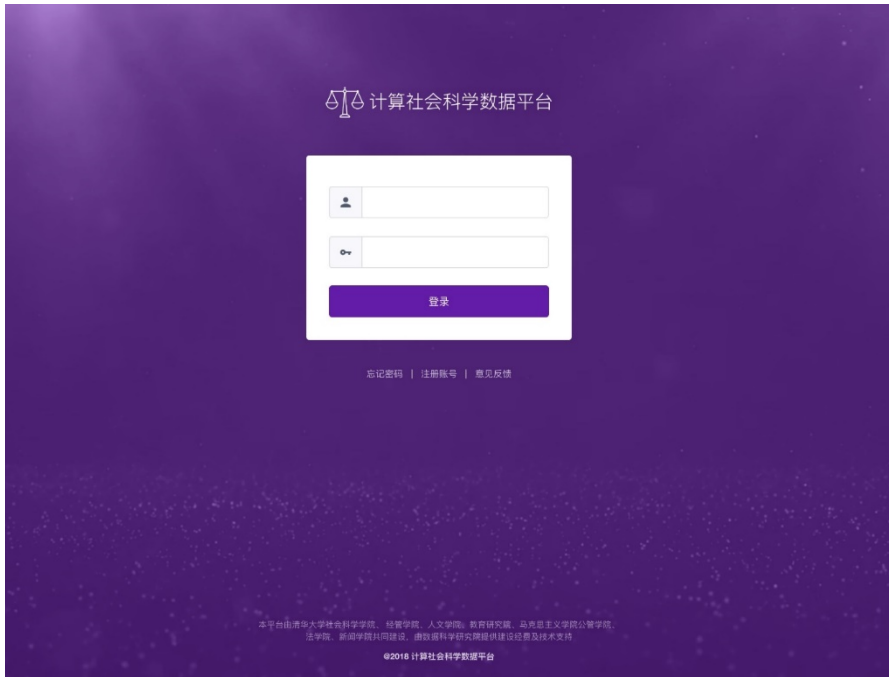
✓ 为中国特色计算社会科学的学科建设、清华高水平社科创新成果、促进重大经济社会问题解决提供理论和应用知识

计算社会科学的清华探索

- 清华大学计算社会科学平台

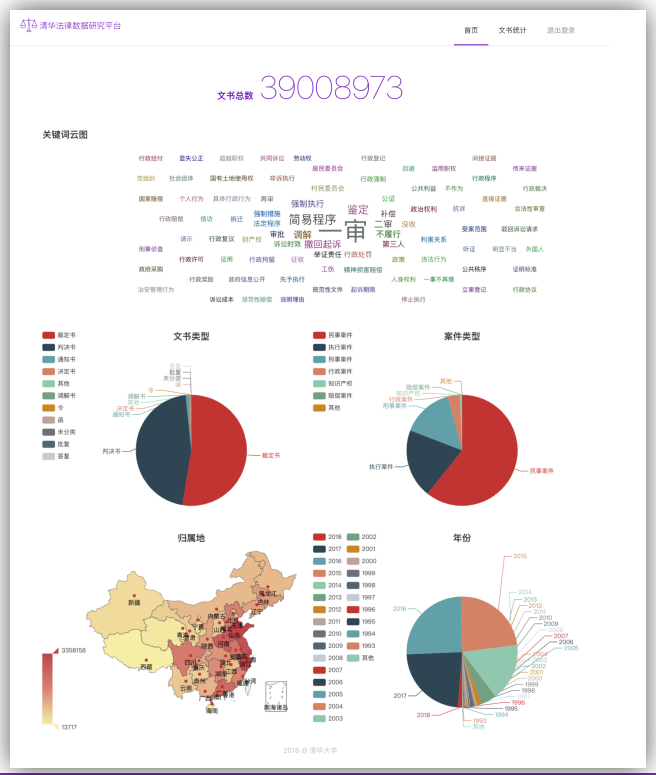
➢法律数据科研条件平台

- 该平台经过一年多的建设，已汇集了8000余万份全国范围内公开的社科内容数据，形成可持续更新的数据库，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/



计算社会科学的清华探索

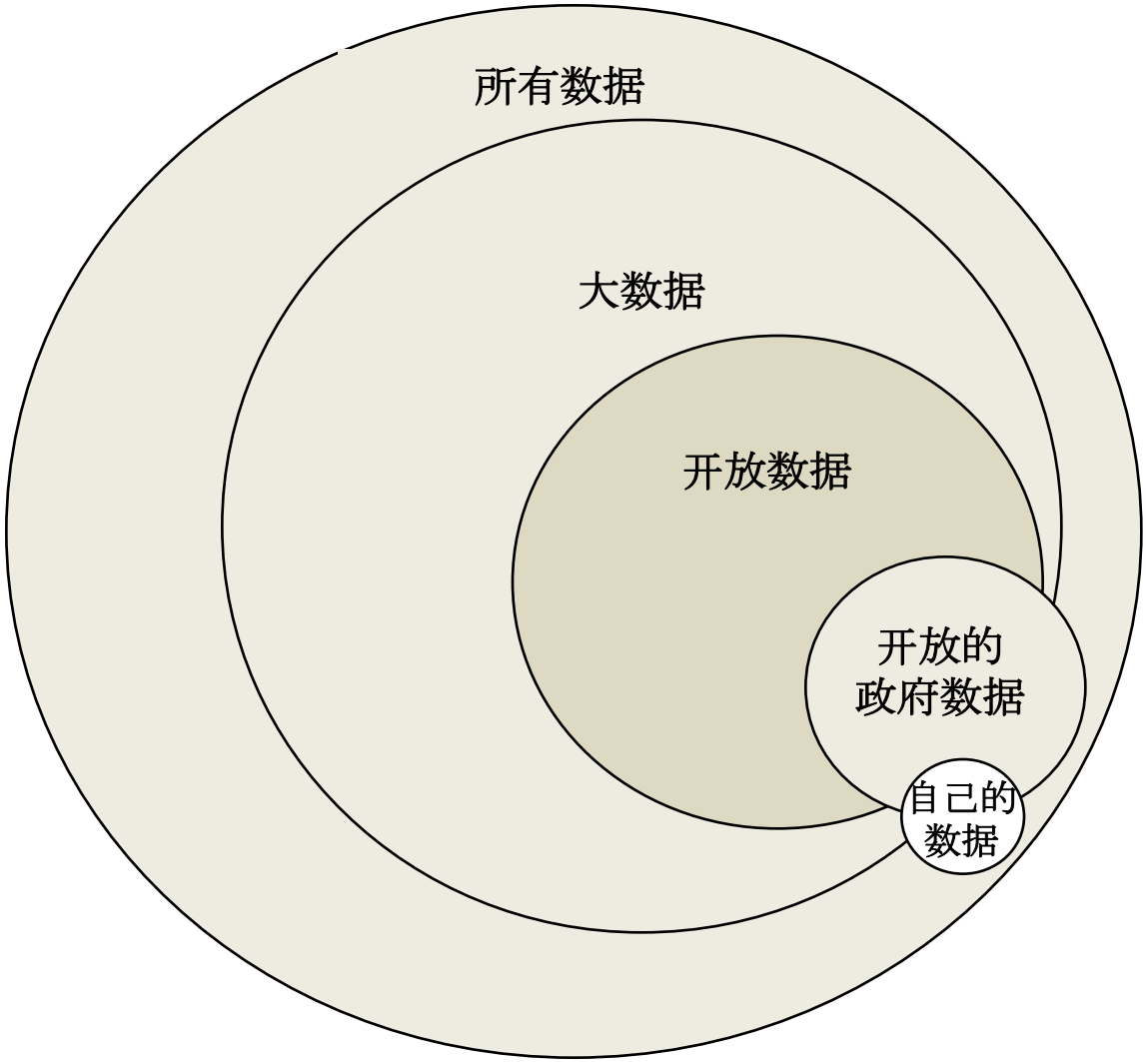
- 2017年11月至今，社科学院与法学院、公管学院、经管学院、人文学院、马克思学院、数据院、计算机系等院系的老师进行过密切的沟通和调研访谈，形成会议记录，明确本平台的建设目的和功能符合师生在社科大数据科研工作的需求。研究并确定基于清华邮箱的账号管理模式属于有效的清华身份认证和账号管理模式，符合要求也方便师生使用。



全局数据纵览

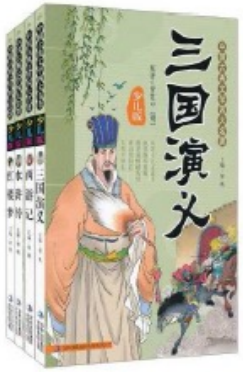
- 对系统内文书总数有充分了解，从可视化图形中纵览各类型文书数据占比、案件类型占比、归属地数据占比及统计年间各个年份数据的占比统计数据，以便用户对系统数据有清楚的全局观。

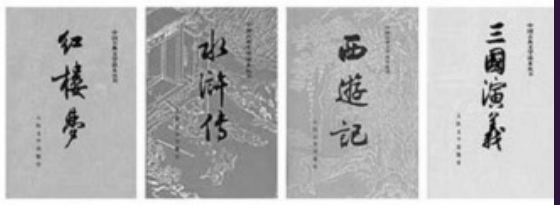
大数据的范畴



文本资料

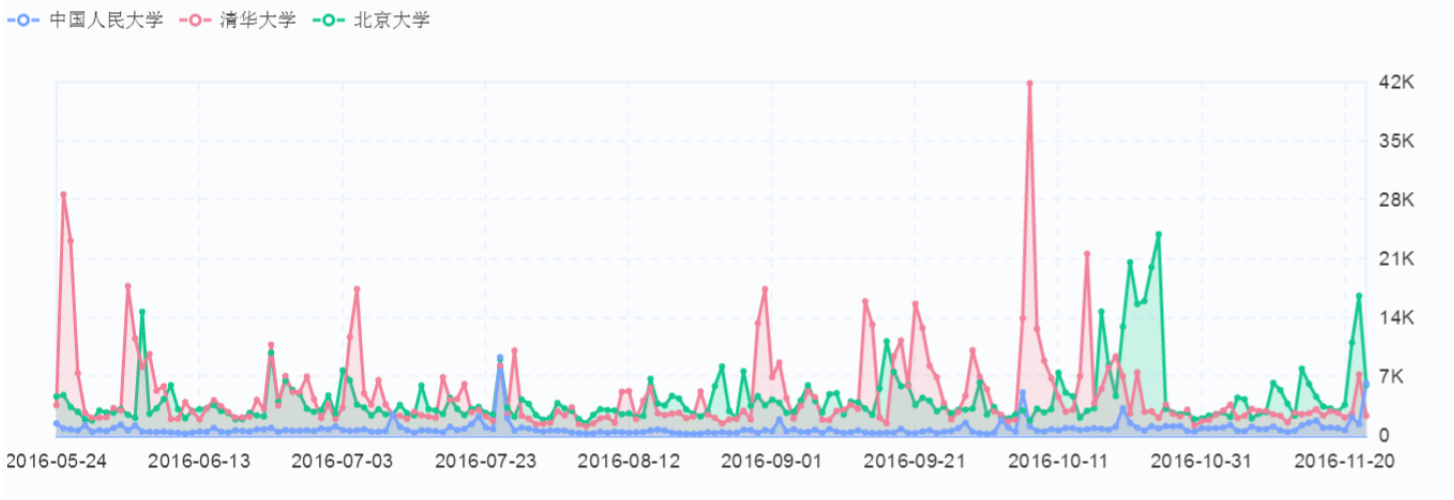
Text data





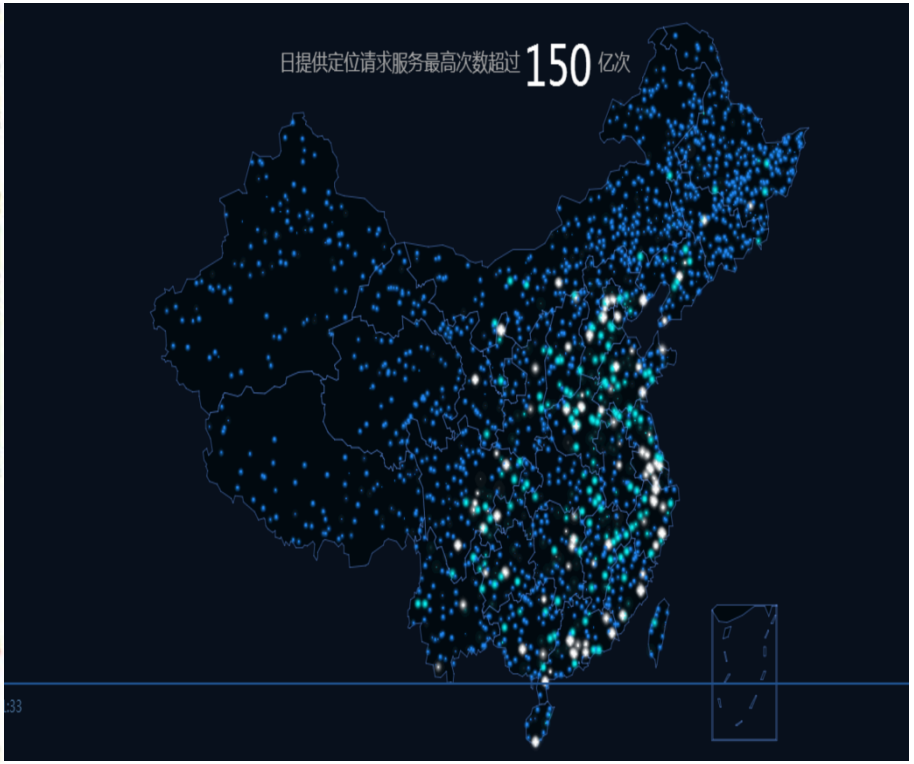
多媒体数据（音频、视频、图片）

门户网站与新媒体


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
小桃园
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)
莲花广场
水上乐园
南京国际展览中心
鼓楼区
紫金塔
冰冻村
河海大学出版社
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
蓝天园
花开四季
清江村
乌龙潭
东宝花园
凤凰西街
小北圩
江东村
明园社区
玄武区
佛心桥
后宰门
山水大酒店
南京博物院
南京航空航天大学
南京博物馆
白下区
中央路
中山路
北京西路
北京东路
华侨路
广州路
珠江路
长白街
户部街
瑞金路


日提供定位请求服务最高次数超过 150 亿次

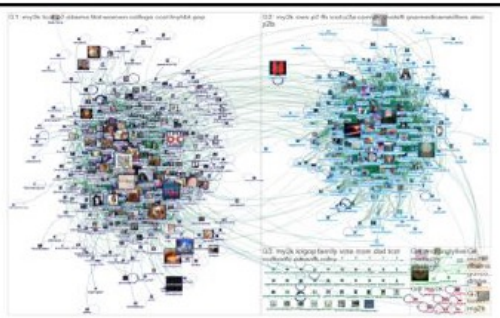
新闻联播中的“国际”与“国内”地域关注

国际地域关注

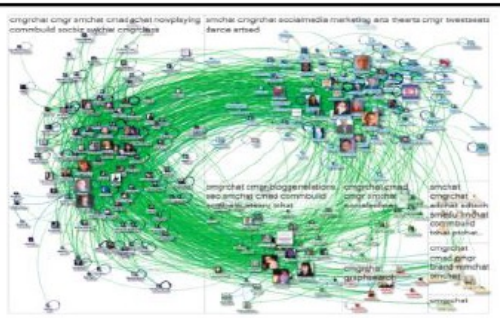
国内地域关注

网络关系数据

6 kinds of Twitter social media networks

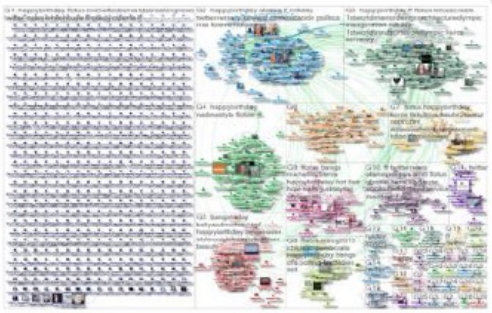


Polarized: two dense clusters with little interconnection

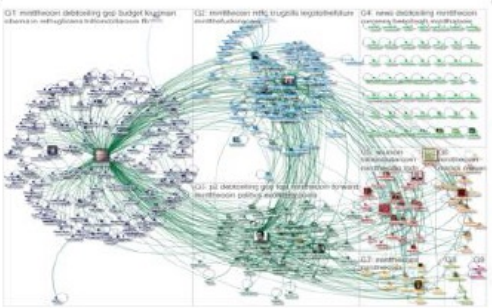


In-group: few disconnected isolates, many connections

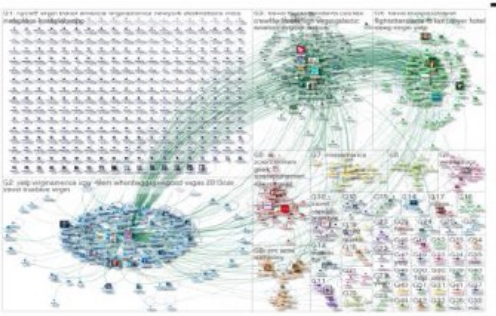
Brand/Public Topic: many disconnected isolates, some small groups



Bazaar: many medium sized groups, some isolates



Broadcast: a hub which is retweeted by many disconnected users

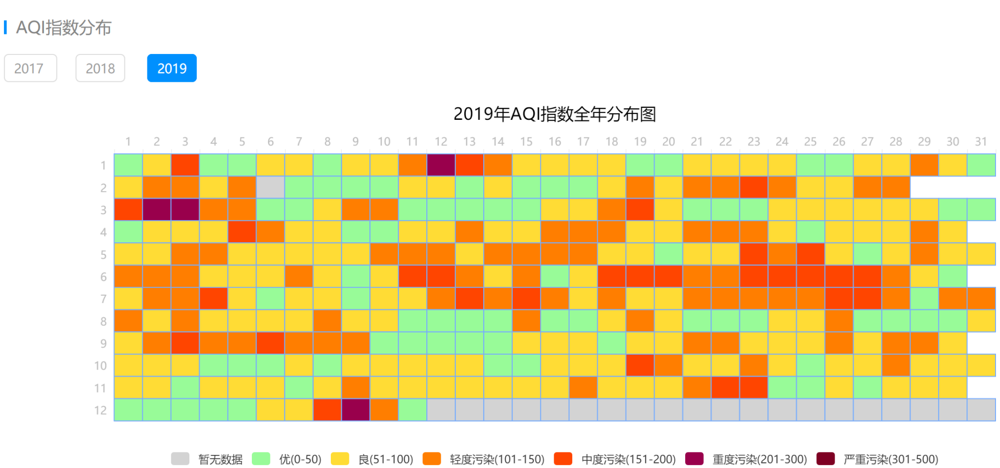


Support: a hub which replies to many disconnected users

行政数据


登录 注册
中国裁判文书网
China Judgements Online
首页 刑事案件 民事案件 行政案件 赔偿案件 执行案件 其他案件 民族语言文书
高级检索 输入案由、关键词、法院、当事人、律师 搜索 ?
开放 动态 透明 便民
今日新增 篇
文书总量 篇
访问总量 次
民事文书 篇
赔偿文书 篇
刑事文书 篇
执行文书 篇
行政文书 篇
其他文书 篇


cnki 中国知网 cnki.net
政府文件
发布机关
全选 清除
中央单位
全国人大
法院、检察院
国务院及所属各部门
党中央及其所属机构
国家监察委员会
军事机构
其他机构
地方单位
北京市
天津市
河北省
山西省
内蒙古自治区
辽宁省
吉林省
黑龙江省
上海市
江苏省
浙江省
安徽省
福建省
江西省
山东省
河南省
高级检索 专业检索
标题 词频 并含 词频
并且 主题 词频 并含 词频
发布日期：从 到
中英文扩展 同义词扩展 检索
排序： 相关度 发布日期 每页显示： 10 20 50 切换到摘要模式
找到335404条结果 1/6709
标题 发布日期
1 中国人民银行营业管理部北京市民政局关于规范全市社会组织开立临时存款账户有关事项的通知 2016-06-21
2 北京市人民政府房改办公室、北京市国土资源和房屋管理局、北京市物价局、中央国家机关房改办、中共中央直属机关房改办关于2003年向职工出售公有住房的价格及政策的通知 2003
3 中国人民银行营业管理部北京市民政局关于规范全市社会组织开立临时存款账户有关事项的通知 2016-06-21
4 文化部、财政部、教育部、国家广电总局、共青团中央、中国文联、北京市人民政府关于开展少儿歌曲创作推广工作的通知 2005-09-14
5 国家统计局北京调查总队北京市统计局关于印发2016年统计年报和2017年定期统计报表制度的通知 2016-11-08
6 中宣部司法部国家体育总局北京奥组委全国普法办关于加强奥运法制宣传的通知 2007-08-30
7 国家统计局 北京市统计局关于发布2017年京津冀区域发展指数的通告 2018-08-01


AQI指数分布
2017
2018
2019
2019年AQI指数全年分布图
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
1 2 3 4 5 6 7 8 9 10 11 12
暂无数据 优(0-50) 良(51-100) 轻度污染(101-150) 中度污染(151-200) 重度污染(201-300) 严重污染(301-500)


企业 人员 风险 知识产权 投融资 公告 研报 新闻 全球企业
小查为你找到 167 位同名老板
王思聪
共任职 54 家企业，主要分布：
上海（共21家） 飞凡电子商务有限公司...等
北京（共10家） 北京普思投资有限公司...等
其他（共23家） 珺娱（湖州）文化发展中心...等
王
王思聪
共任职 24 家企业，主要分布：
云南（共24家） 昆明扑通扑通商贸有限公司...等
查看全部
167 位王思聪
查找范围 企业名 法人或股东 高管 品牌/产品 联系方式 经营范围 多选搜索 NEW
省份地区 广东 (33) 北京 (48) 江苏 (24) 上海 (37) 浙江 (34) 更多
行业门类 批发和零售业 (181) 租赁和商务服务业 (99) 科学研究和技术服务业 (35) 制造业 (32) 更多
更多筛选 成立日期 企业状态 注册资本 资本类型 组织机构 企业类型
VIP筛选 联系电话 手机号码 联系邮箱 网址信息 纳税信用 商标信息 专利信息
融资信息 上市状态 失信信息 作品著作权 软件著作权 参保人数 高新企业
招投标 进出口信用 动产抵押 清算信息

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；

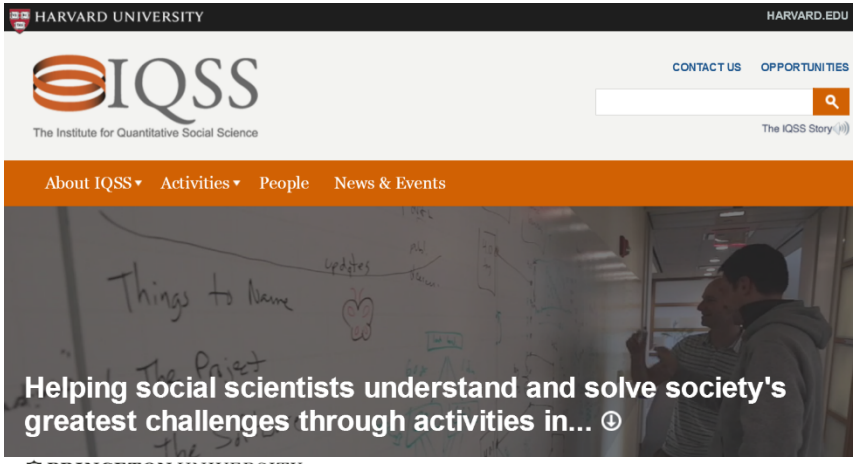
✓2009 年，哈佛大学David Lazer等15 位美国学者在《Science》上联合发表了一篇具有里程碑意义的文章“Computational Social Science”；

✓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。

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

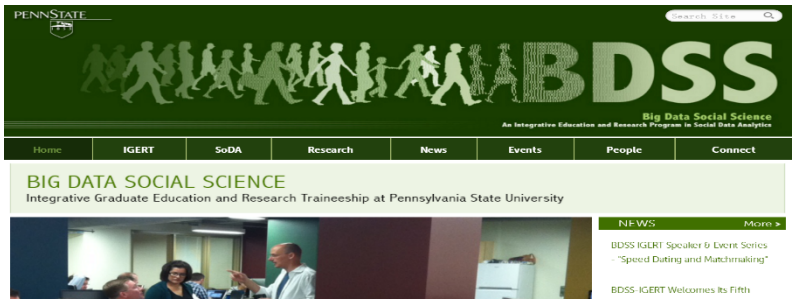
Big Data and Social Analytics certificate course

EXperimental Learning



PRINCETON UNIVERSITY

CENTER FOR STATISTICS AND MACHINE LEARNING



LAZER LAB

HOME PEOPLE PROJECTS PUBLICATIONS DATA EVENTS

TOP NEWS

VOLUNTEE SCIENCE

Volunteer Science

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

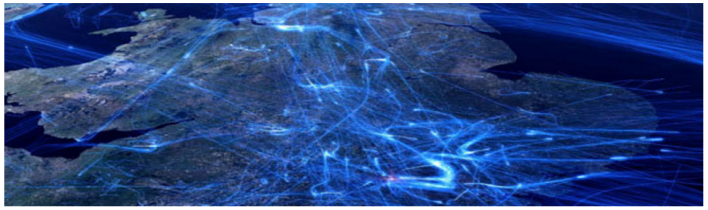
WZB

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

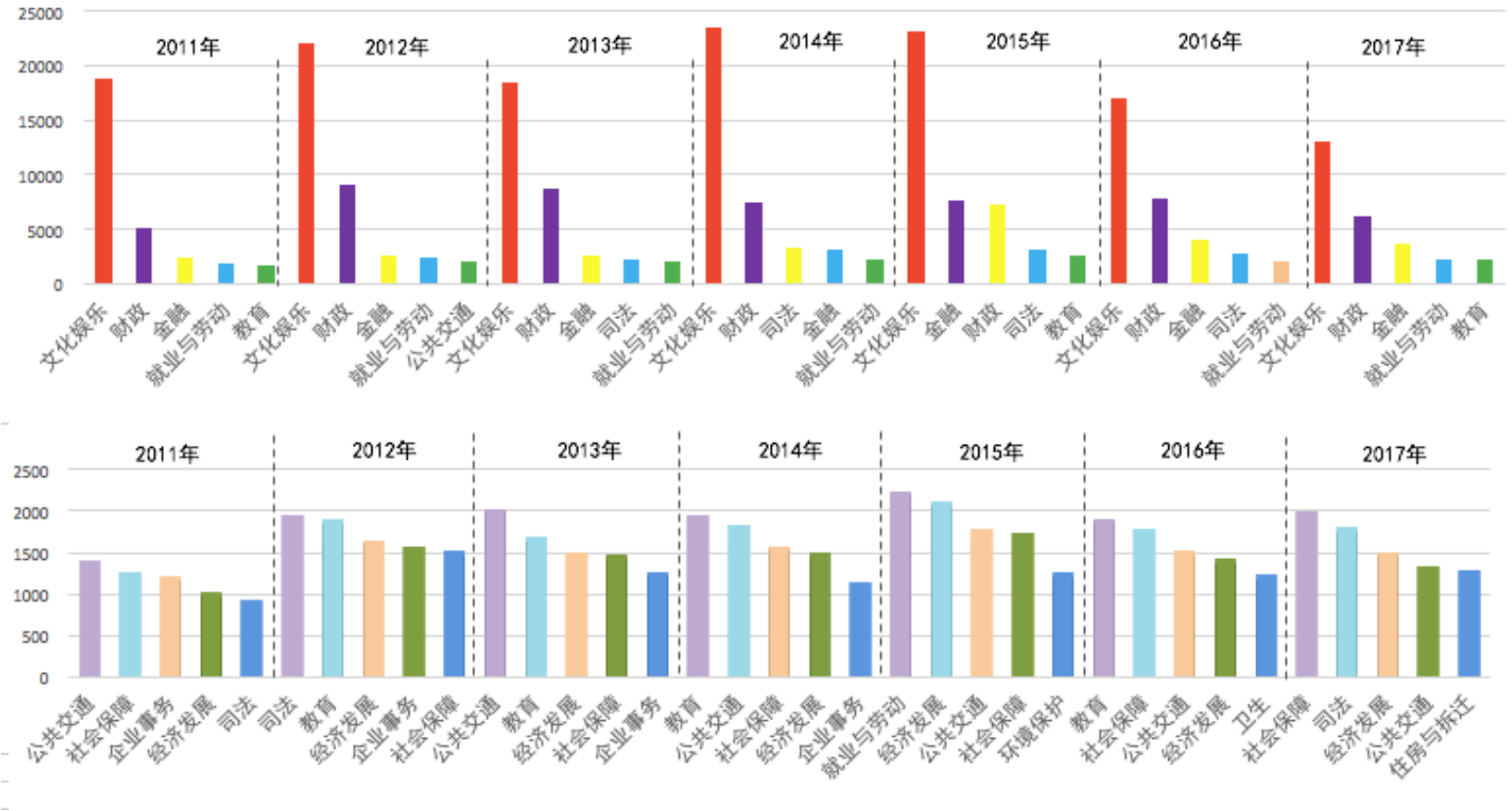
- 网络爬虫
- 对搜索引擎搜索记录的分析
- 自动文本分析
- 视频/图片分析
- 社会网络分析
- 空间/时间分析
- 可视化

- 机器学习
- 自然语言过程
- 统计分析

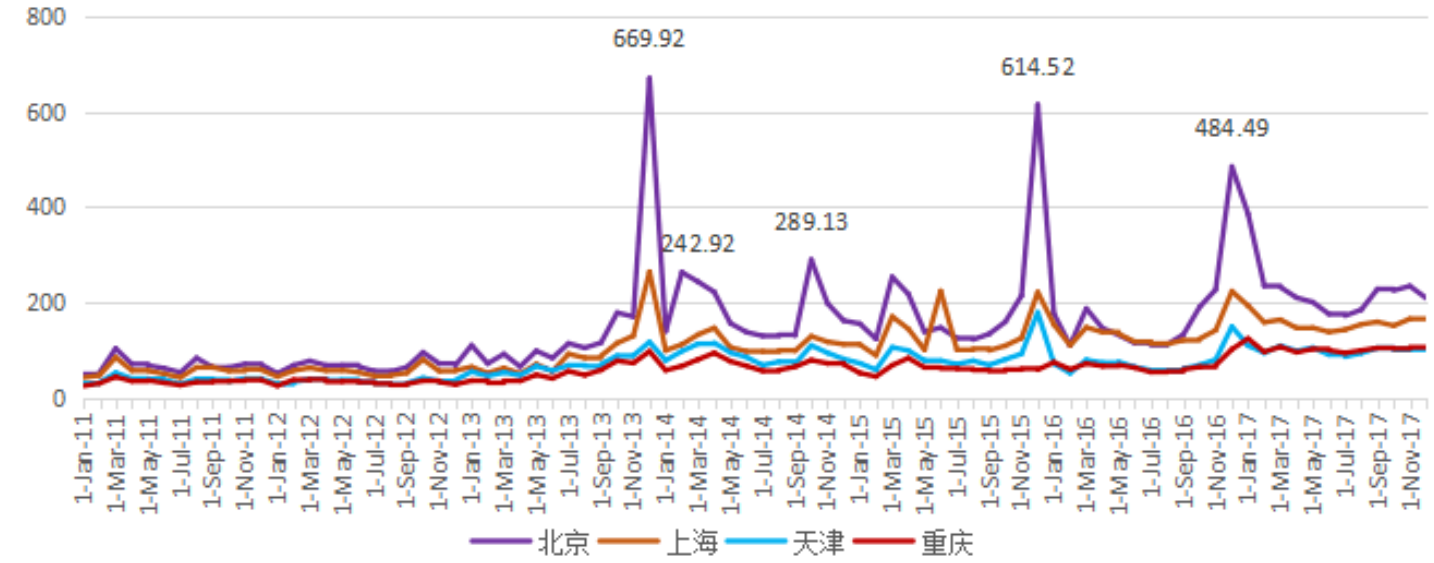
社科研究中的大数据：方法功能

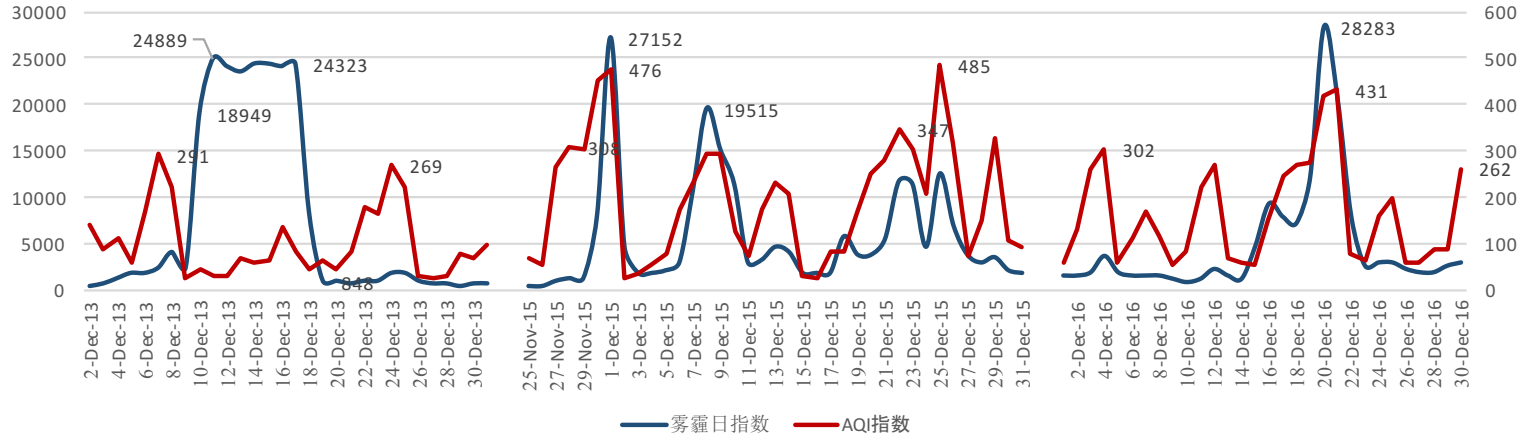
- 作为研究方法的大数据分析
  - 数据采集与管理
  - 测量工具
  - 分类与聚类
  - 关联分析
  - 因果推论（回归分析）
  - 信息呈现（可视化）

搜索指数：Google Trends、百度指数



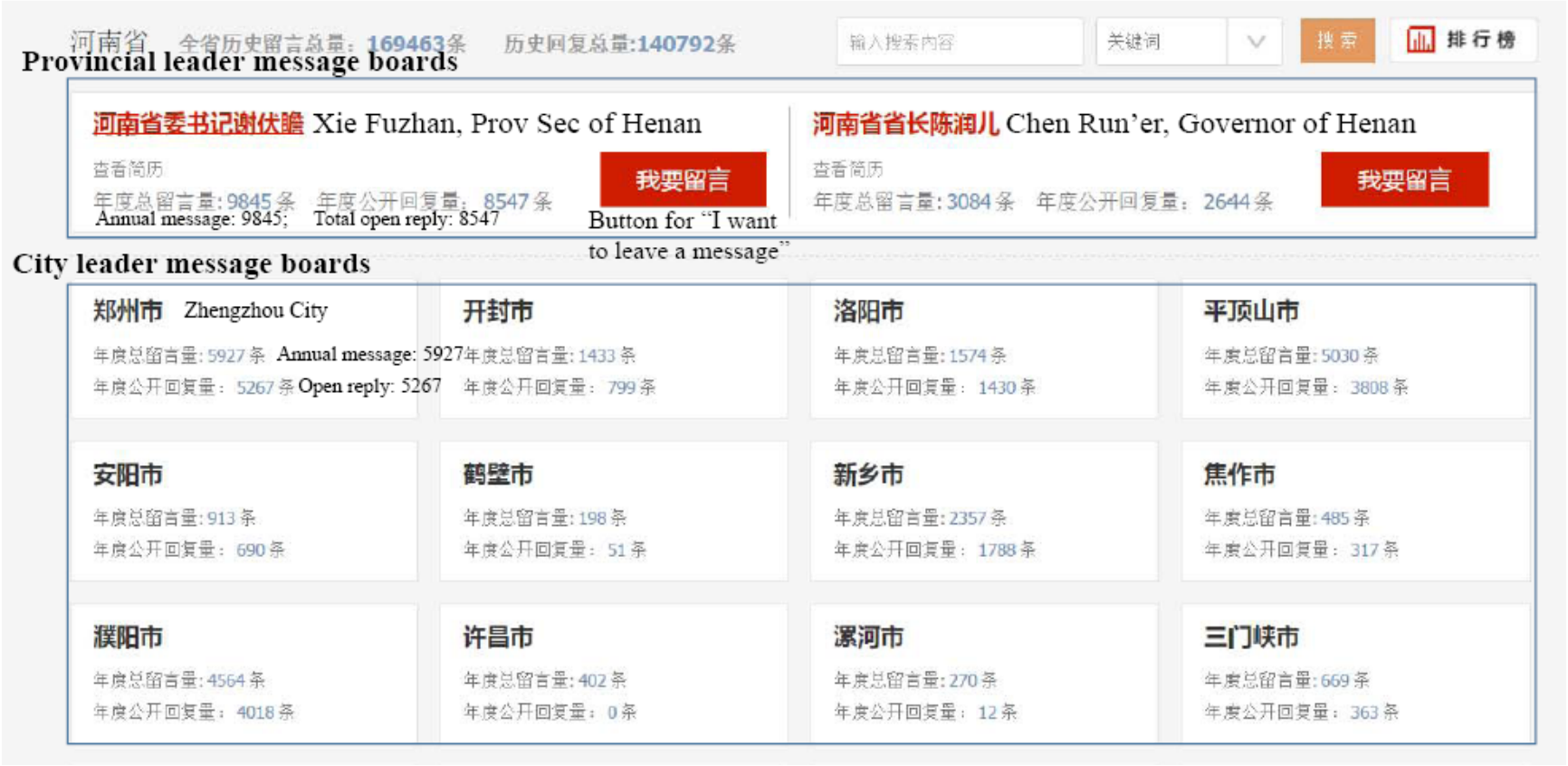
搜索指数：测量公共关注度





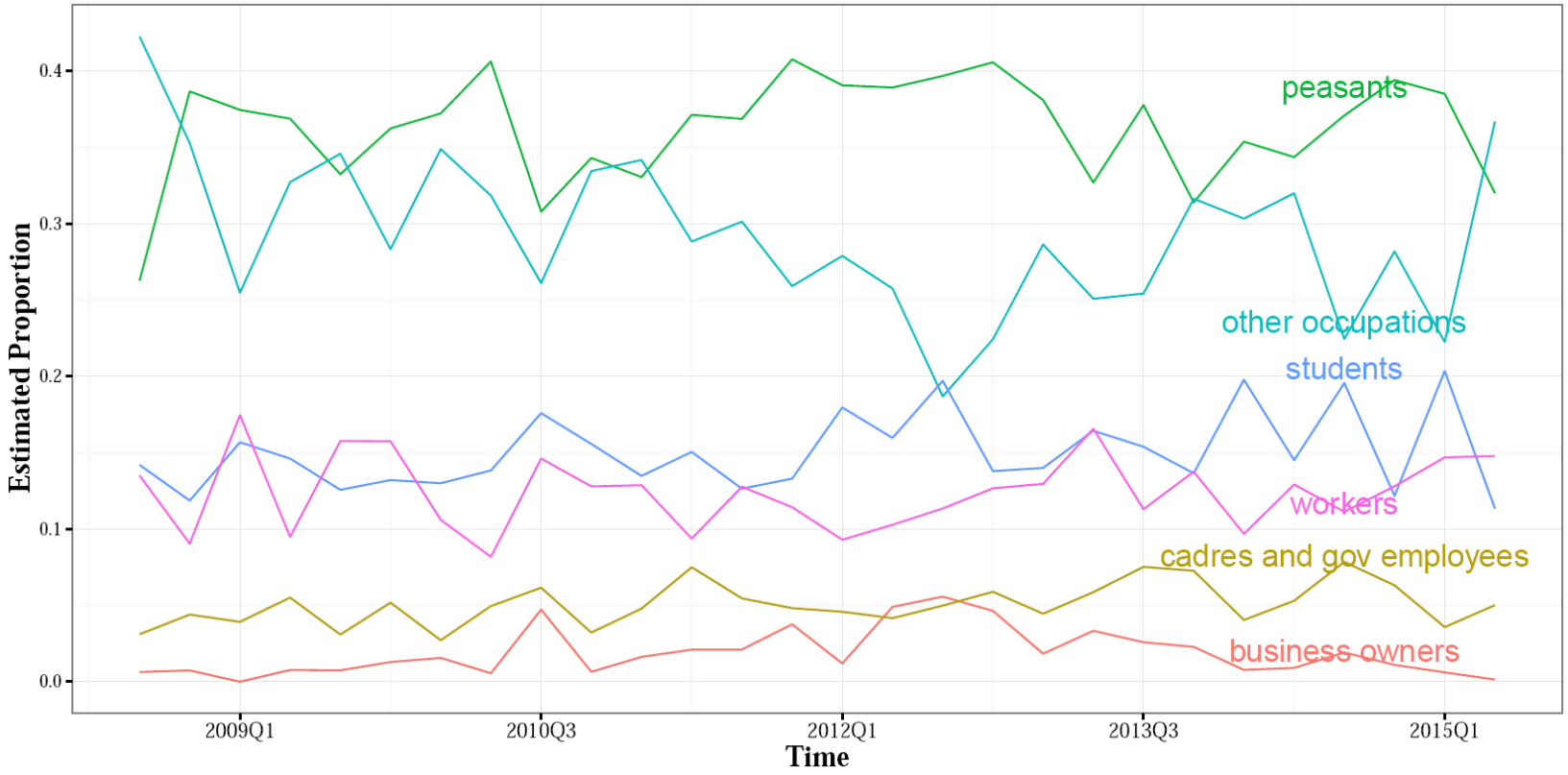
The Local Leader Message Board

The Interface

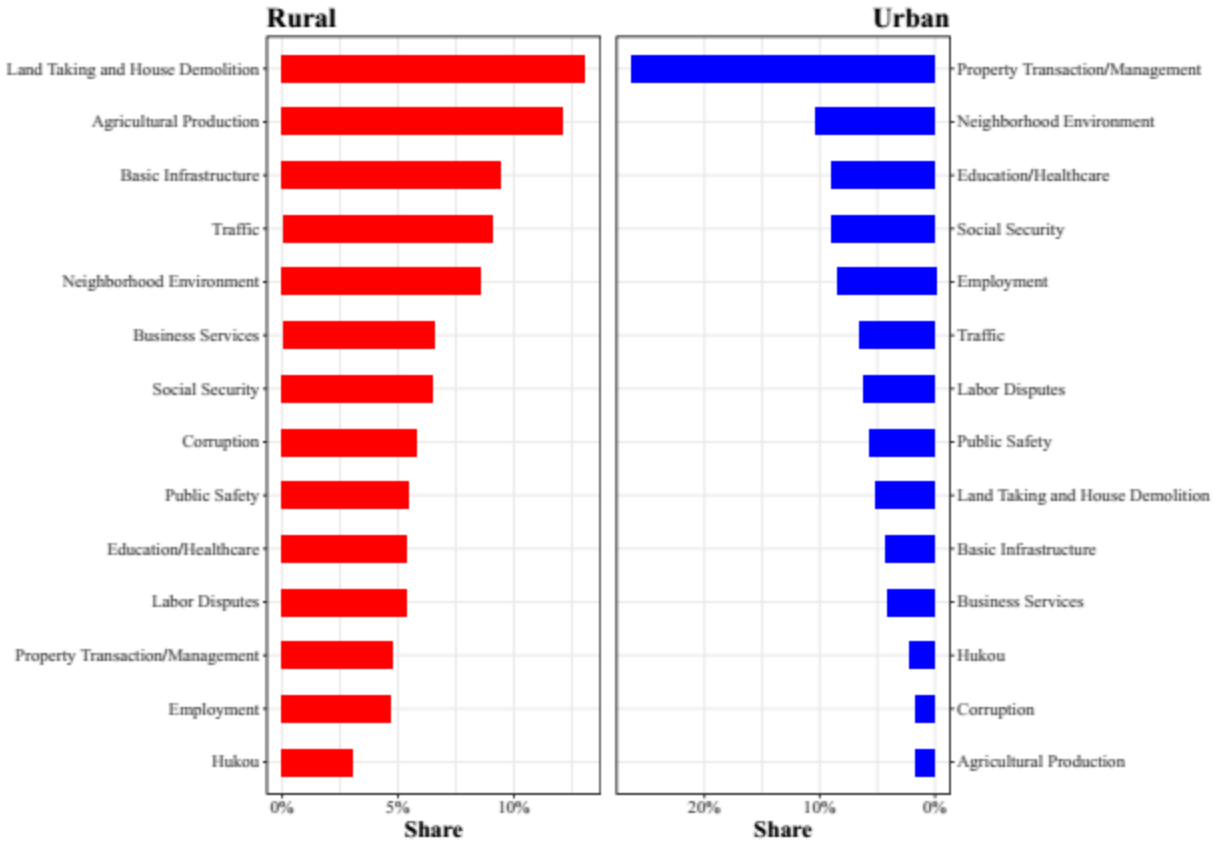


Who Participate Online?

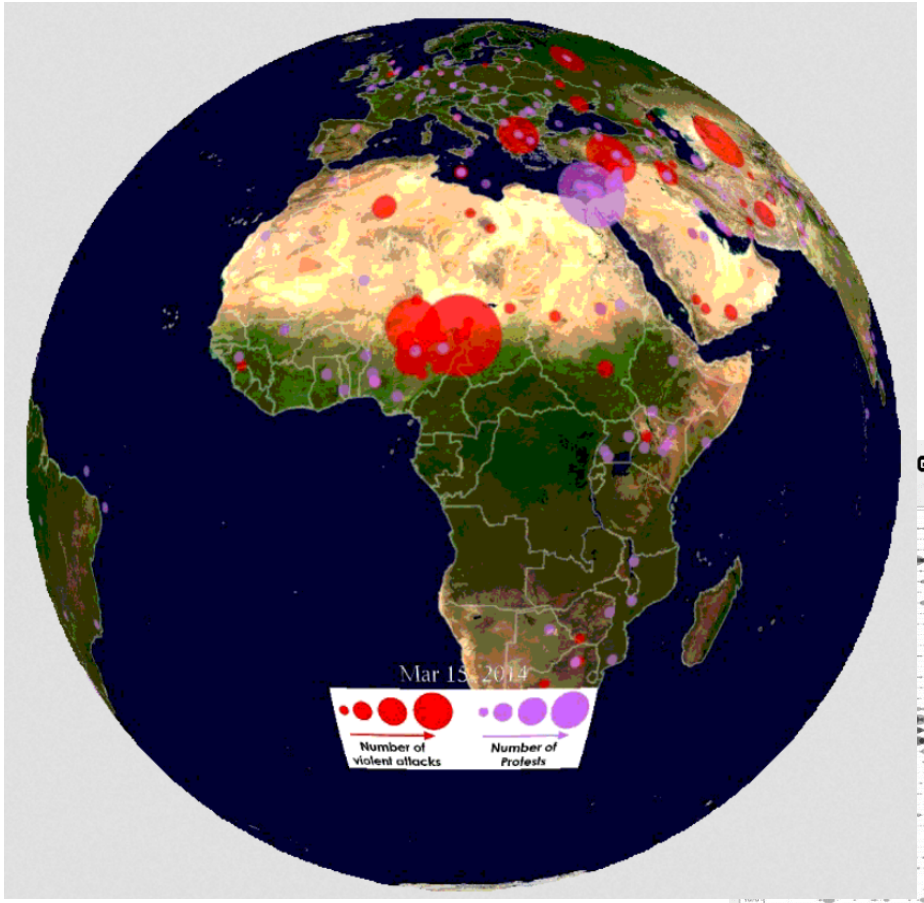
- Supervised Learning with readme (Hopkins and King, 2010)

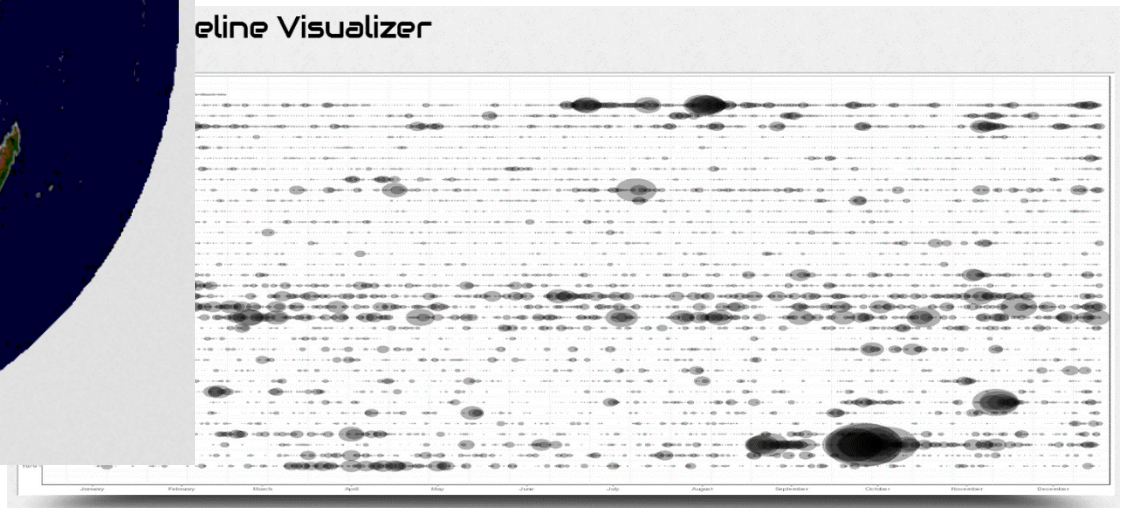


Participate for What?



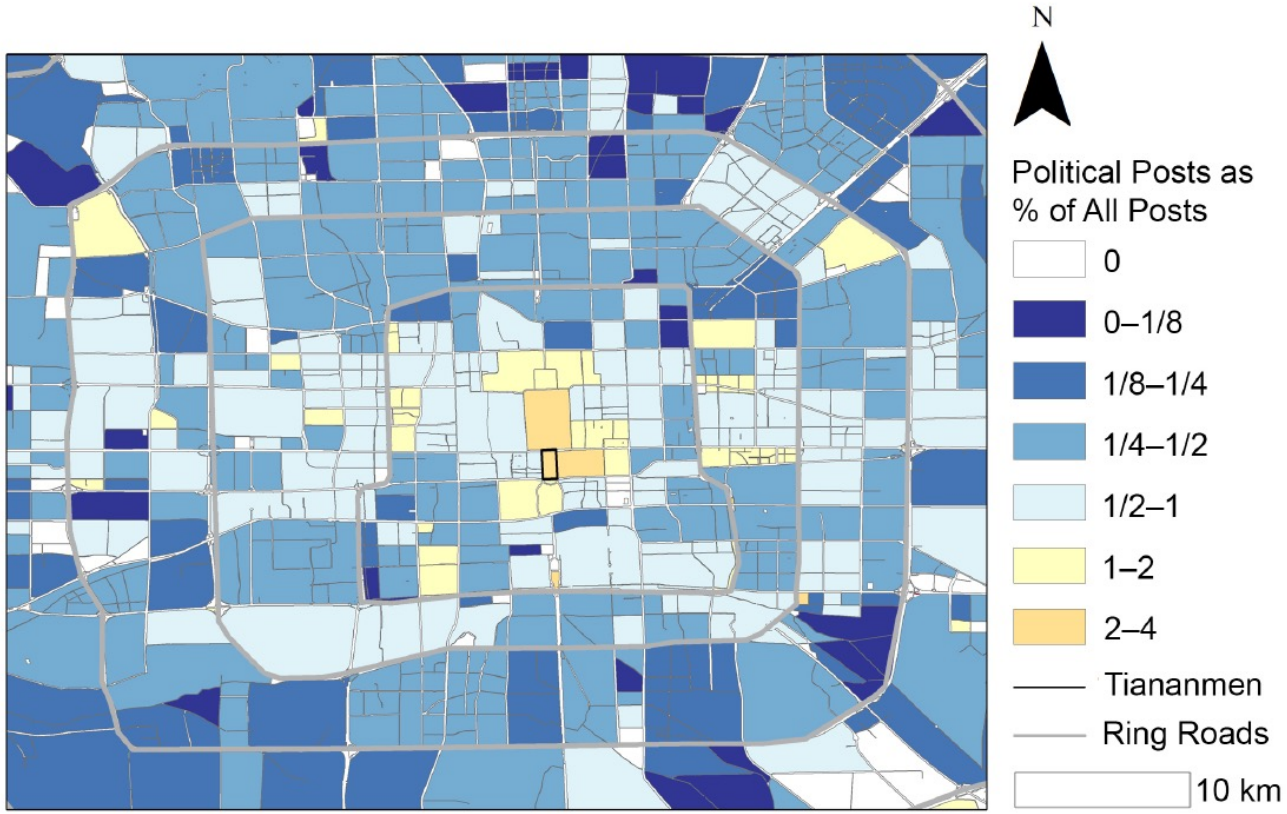
时间序列分析：全球事件、语调与语言数据库（GDELT）





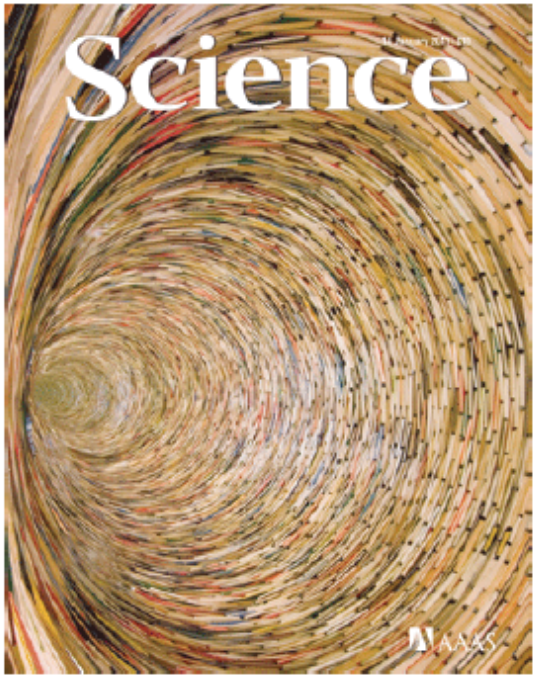
空间分析：空间与政治

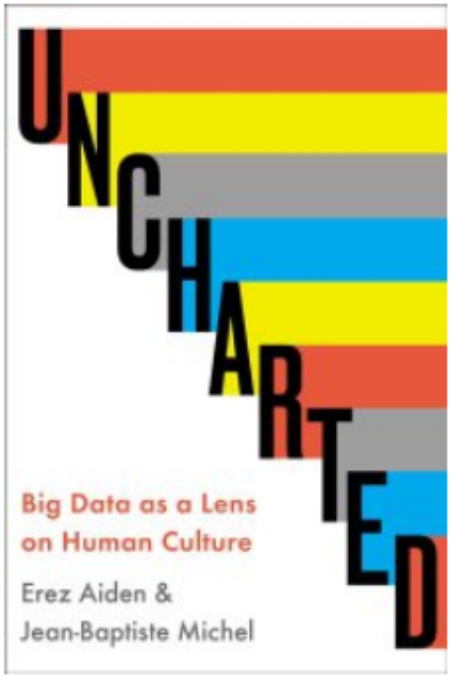
Figure 2. Spatial Distribution of Political Posts, Beijing, 1 July 2014–15 June 2015



文本分析：基于关键词的文化组学

- 文化组学(culturomics) http://www.culturomics.org/
- 哈佛大学研究团队利用Google Books 收集并扫描识别了1800-2000年之间的500万种出版物，通过关键词使用频率随时间变化的趋势，分析人类文化演进特点
- Google Book N-grams (https://books.google.com/ngrams)



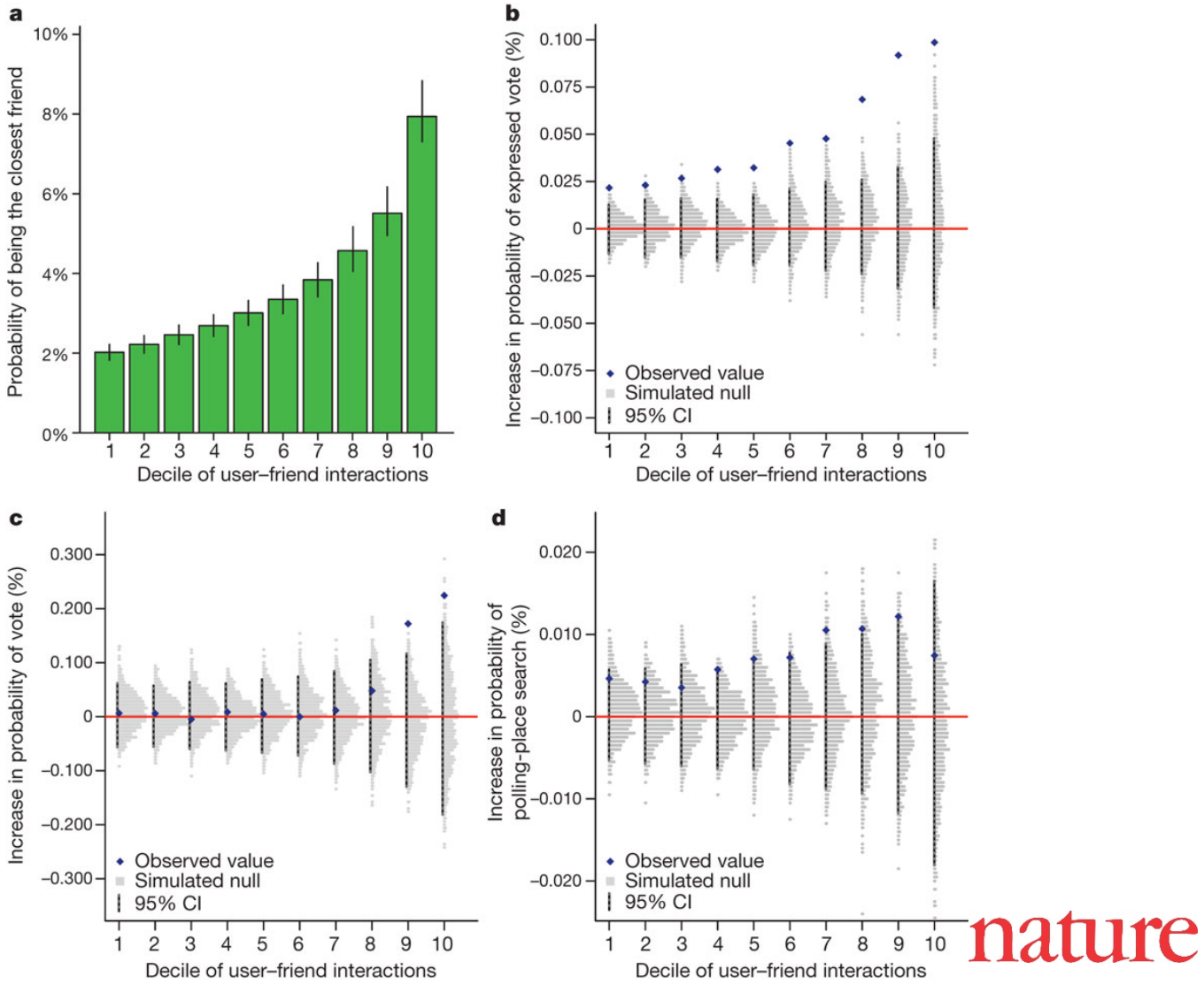


开源文本分析：Google Ngram

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

King, G. , Schneer, B. , & White, A. . (2017). How the news media activate public expression and influence national agendas. Science, 358(6364), 776-780.

- King等（2017）利用48个社交媒体开展（五年期）田野实验

✓在真实媒体环境中设计并随机化分配媒体资讯（报道）

✓识别媒体报道对个体公共意见（政治知识）的效应（Individual Effect）

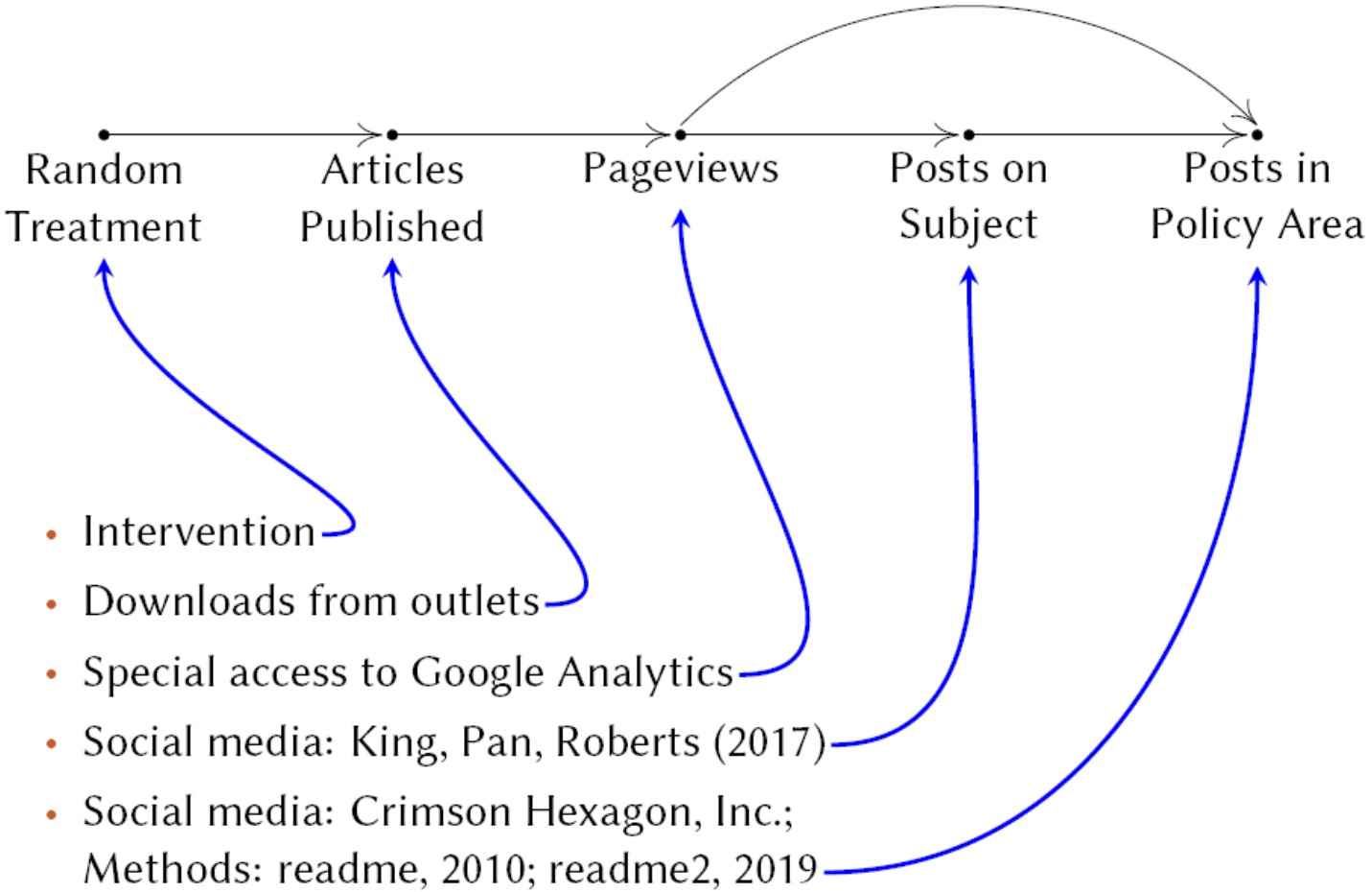
✓识别媒体报道对国家政策议程的效应（Collective Effect）

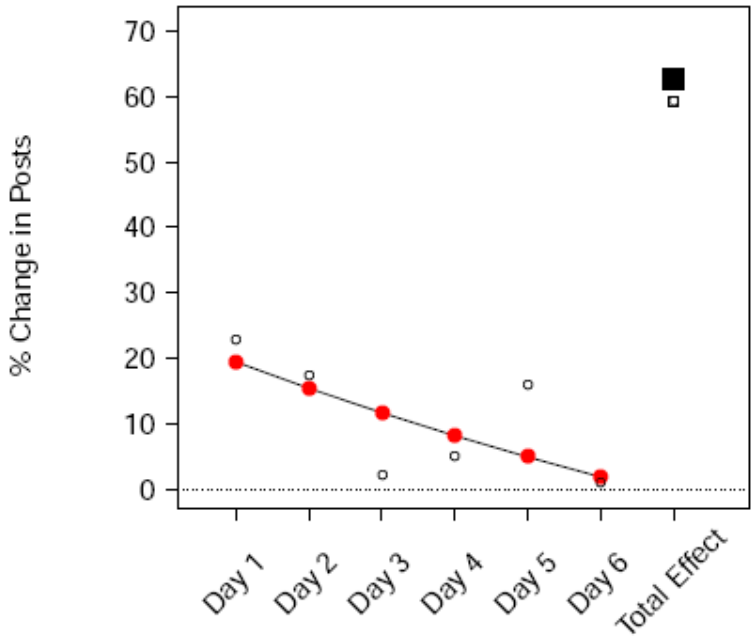
King, G. , Schneer, B. , & White, A. . (2017). How the news media activate public expression and influence national agendas. Science, 358(6364), 776-780.

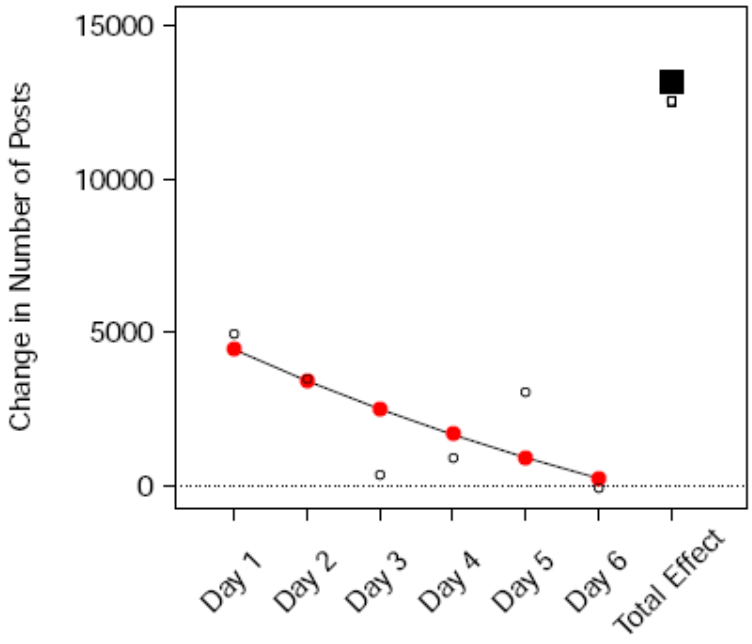
- Signup 48 small media outlets (& > 12 others just for info)
  - 17 for trial runs, 33 in experiment, 2 in both
  - Median size: *The Progressive*, 50,000 subscribers
  - Examples:

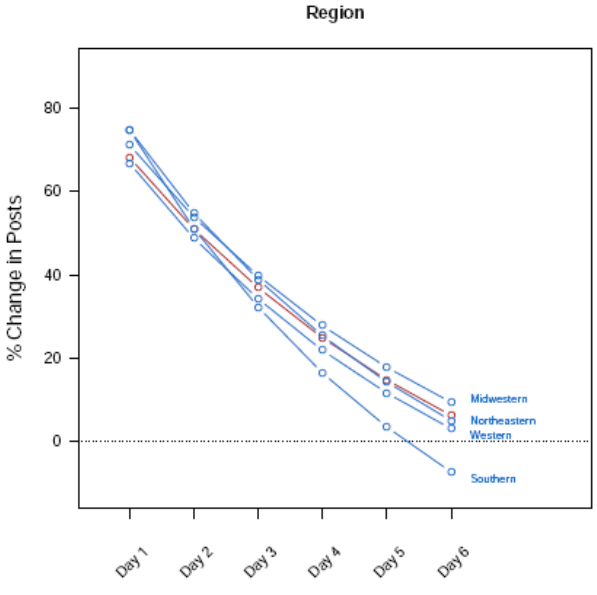
- Establish 11 broad *policy areas*
  - Rules: (a) major national importance; (b) interest to outlets
  - race, immigration, jobs, abortion, climate, food policy, water, education policy, refugees, domestic energy production, and reproductive rights

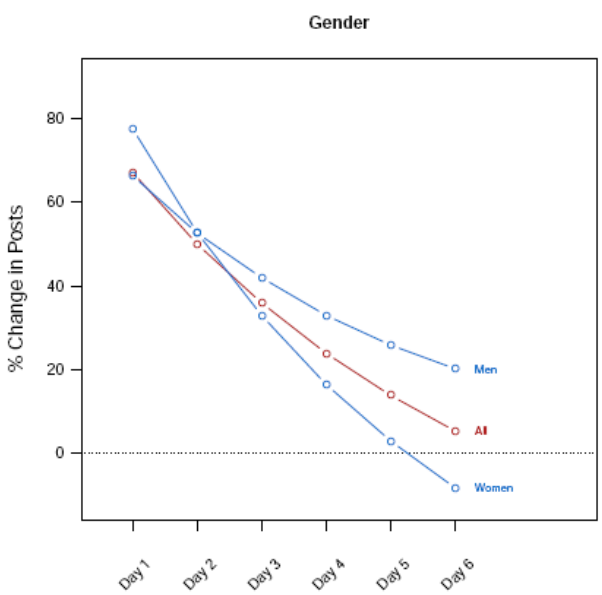
King, G. , Schneer, B. , & White, A. . (2017). How the news media activate public expression and influence national agendas. Science, 358(6364), 776-780.

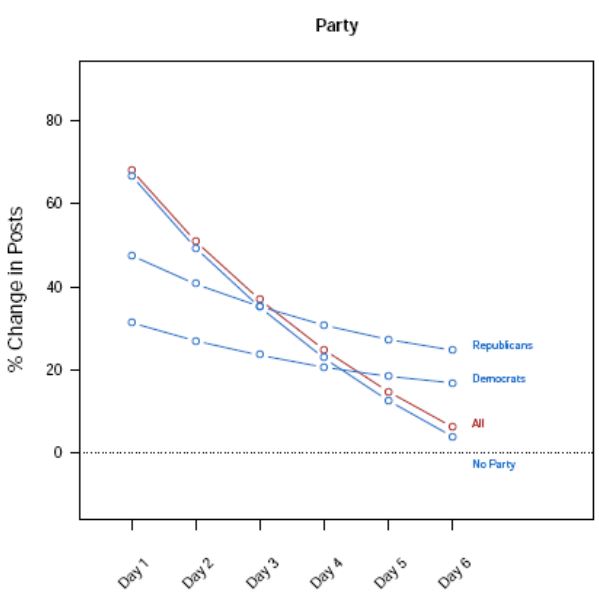



% Change in Posts
70
60
50
40
30
20
10
0
Day 1
Day 2
Day 3
Day 4
Day 5
Day 6
Total Effect


Change in Number of Posts
15000
10000
5000
0
Day 1
Day 2
Day 3
Day 4
Day 5
Day 6
Total Effect


Region
% Change in Posts
80
60
40
20
0
Midwestern
Northeastern
Western
Southern
Day 1
Day 2
Day 3
Day 4
Day 5
Day 6


Gender
% Change in Posts
80
60
40
20
0
Men
All
Women
Day 1
Day 2
Day 3
Day 4
Day 5
Day 6


Party
% Change in Posts
80
60
40
20
0
Republicans
Democrats
All
No Party
Day 1
Day 2
Day 3
Day 4
Day 5
Day 6

大数据方法：支持VS批评

- 大数据方法的优势
  - 数据：“全量数据”、“消极数据”、“大样本小概率事件”、高维数据
  - 方法：机器学习、预测
  - 经济性/可行性：低成本、实效性、高效率
  - 影响：知识平民化传播

大数据方法的批评

- 大数据方法的局限性
  - 数据：“有偏数据”、“分析单位”、“假数据”
  - 方法：效度与信度、技术门槛
  - 可行性：数据不开放、技术门槛
  - 伦理：数据（隐私）权利、社会实验的伦理困境

Q&A