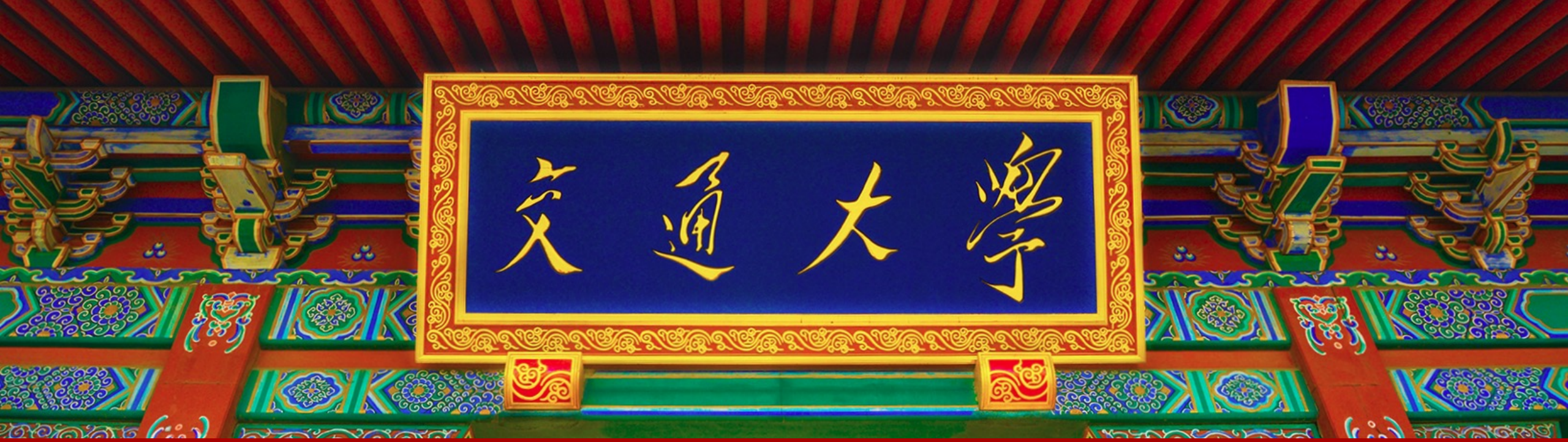



# 计算社会科学导论

## —— 课程简介

金耀辉、许岩岩

2023年2月16日

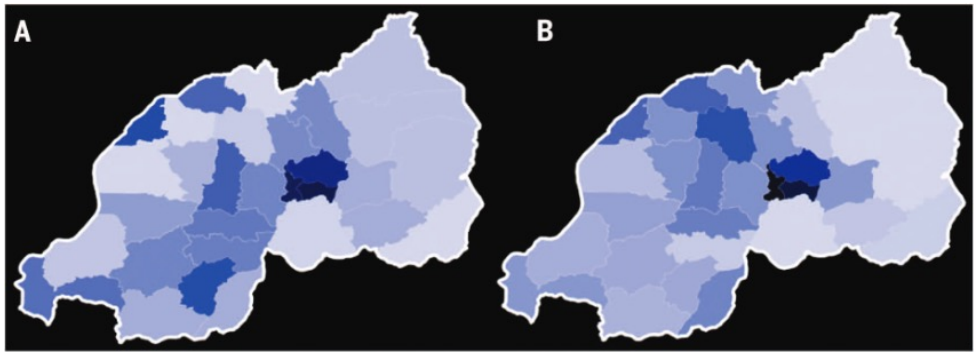
上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

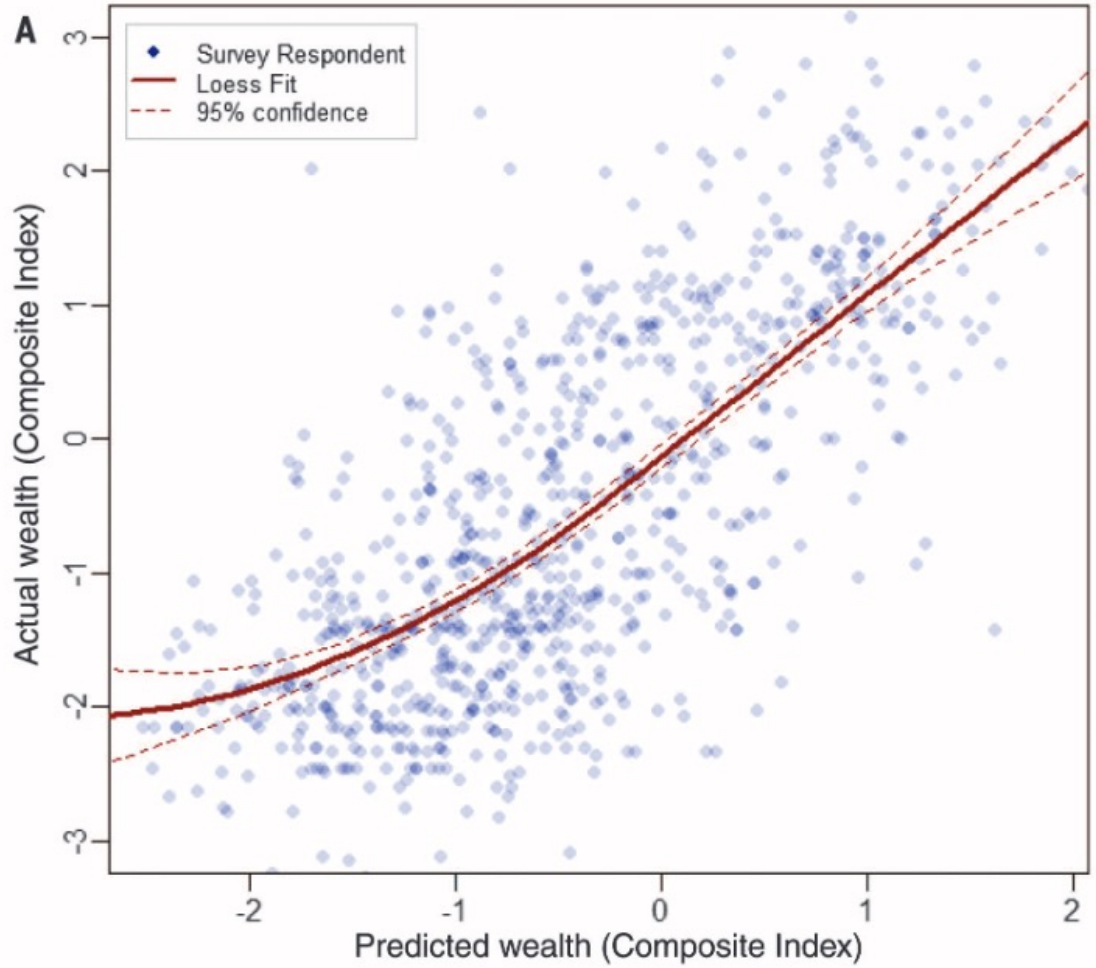
# 传统社会科学研究方法

| 研究方法 | 子类型 | 资料收集方法 | 资料分析方法 | 研究的性质 |
| --- | --- | --- | --- | --- |
| 调查研究 | 普遍调查抽样调查 | 统计报表、自填式问卷、结构式访问 | 统计分析 | 定量 |
| 实验研究 | 实地实验、实验室实验 | 自填式问卷、结构式访问结构式观察、量表测量 | 统计分析 | 定量 |
| 实地研究 | 参与观察、个案研究 | 无结构观察、自由式访问 | 定性分析 | 定性 |
| 文献研究 | 统计资料分析二次分析、内容分析历史比较分析 | 官方统计资料他人原始数据文字声像文献、历史文献 | 统计分析<br>定性分析 | 定量/定性 |

# Intro-计算社会科学

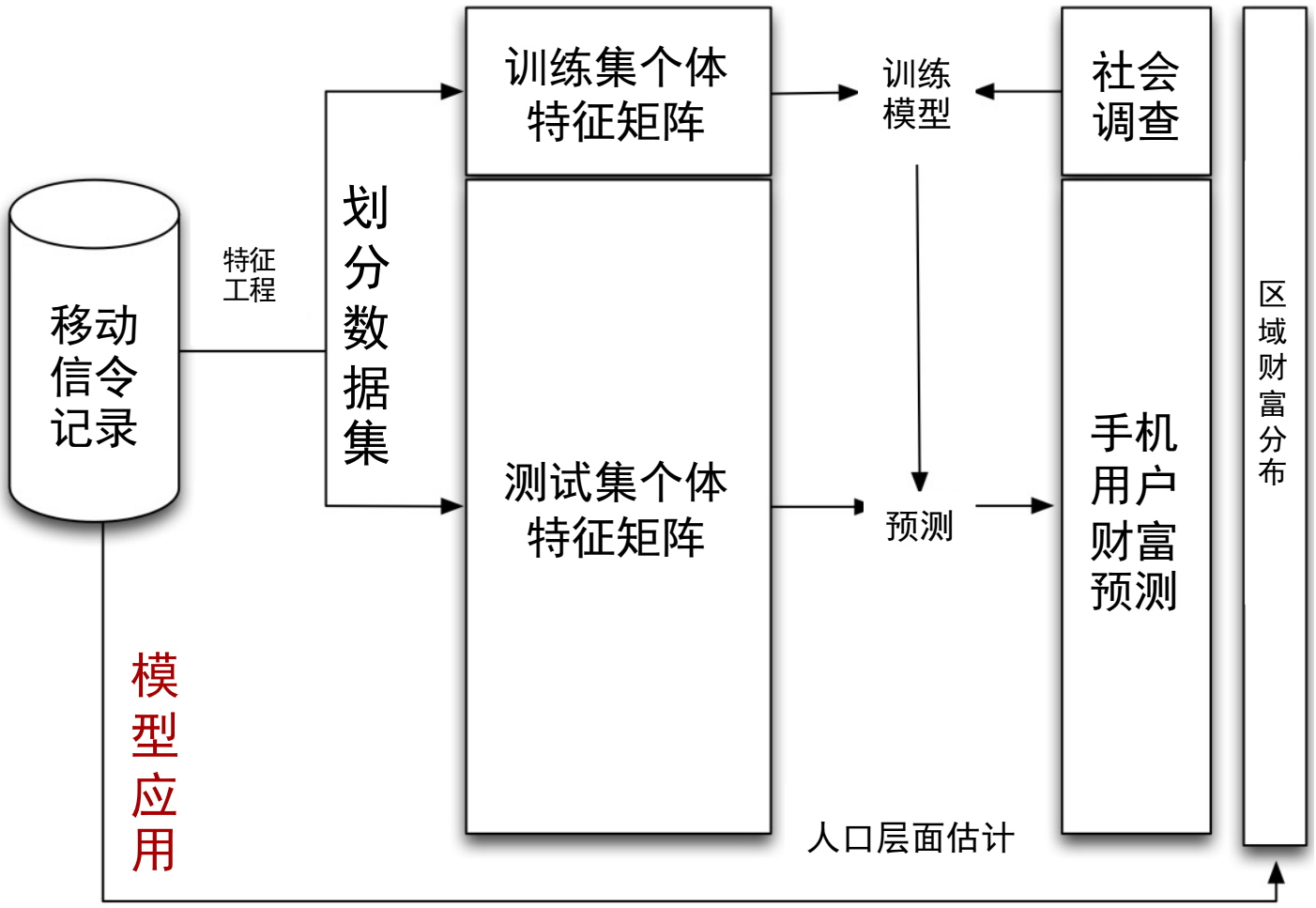
- Blumenstock J, Cadamuro G, On R. Predicting poverty and wealth from mobile phone metadata[J]. Science, 2015, 350(6264): 1073-1076.
- **个体的手机使用记录可以预测其社会经济地位**
- **并且可以用于重建国家或小型地区的资产分布情况**
- **在资源受限地区减少进行大范围普查的成本和时间**




上图中，a:预测财富指数，横轴为预测财富，纵轴为真实财富
左图中，A和B分别表示预测出的区域平均财富指数和调查结果

# Intro-计算社会科学（续）

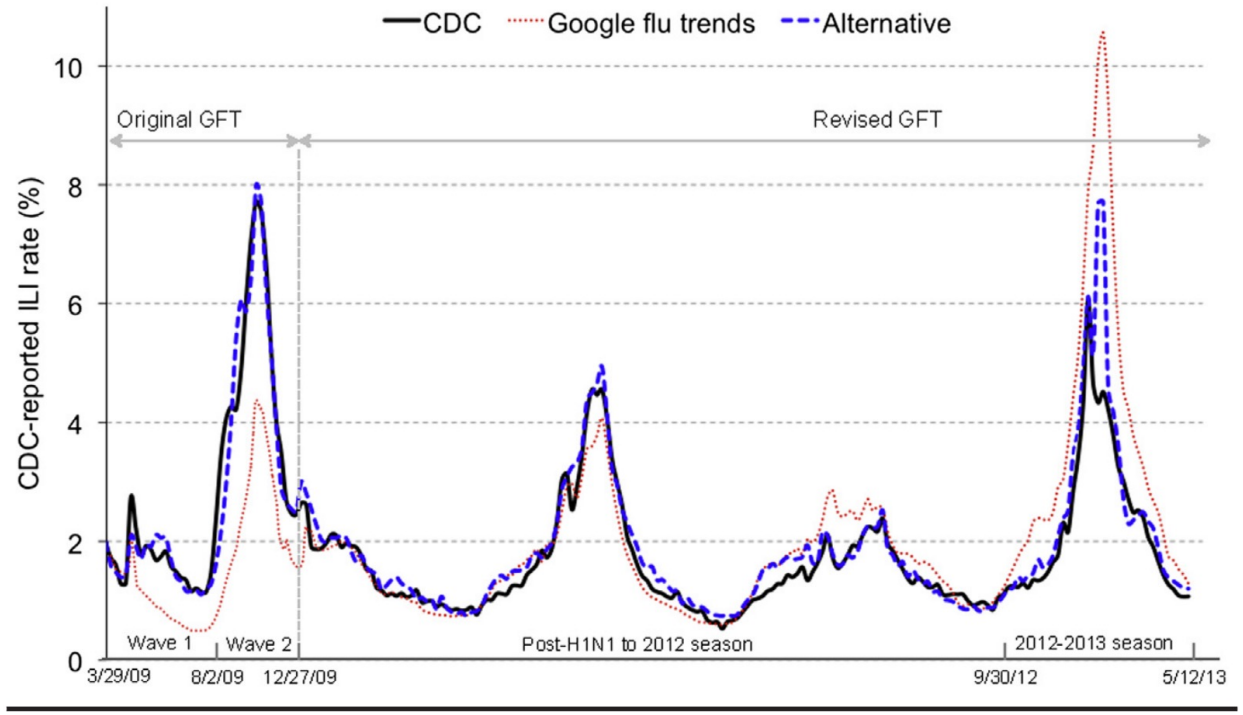



## 特点

- 数据量较少受限于调查成本
  - 基于信息基础设施
- 多样化数据源
- 复杂网络、社会模拟、大容量模型...

Blumenstock J, Cadamuro G, On R. Predicting poverty and wealth from mobile phone metadata[J]. Science, 2015, 350(6264): 1073-1076.

# Google's Flu Trends：成功与争议

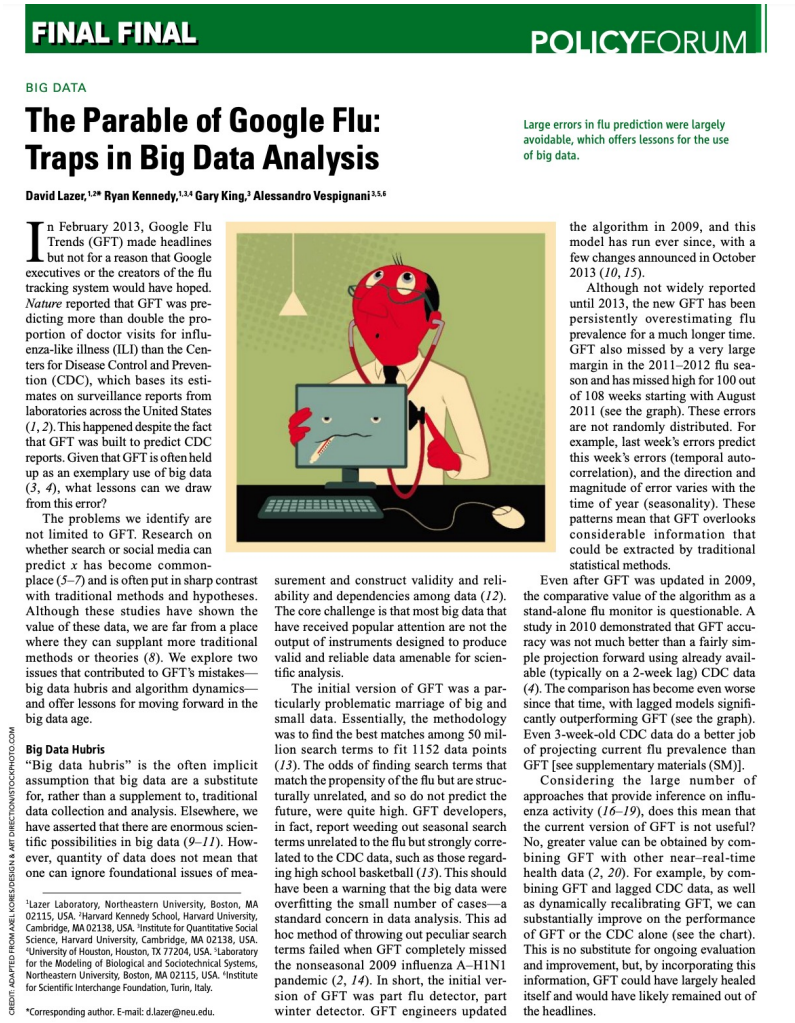



2008年谷歌推出了“谷歌流感趋势”（GFT），这个工具根据汇总的谷歌搜索数据，近乎实时地对全球当前的流感疫情进行估测。

2009年在H1N1爆发几周前，谷歌成功预测了H1N1在全美范围的传播，甚至具体到特定的地区，而且判断非常及时，令公共卫生官员们和计算机科学家们倍感震惊。

2013年1月，美国流感发生率达到峰值，谷歌流感趋势的估计比实际数据高两倍，就是这个不精确性再次引起了媒体的关注。

# 大数据浮夸和算法变化



FINAL FINAL

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[3,5,6]

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 commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

### Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (*9*–*11*). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points (*13*). The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball (*13*). This should have been a warning that the big data were overfitting the small number of cases—a standard concern in data analysis. This ad hoc method of throwing out peculiar search terms failed when GFT completely missed the nonseasonal 2009 influenza A–H1N1 pandemic (*2*, *14*). In short, the initial version of GFT was part flu detector, part winter detector. GFT engineers updated the algorithm in 2009, and this model has 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 in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (*4*). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT [see supplementary materials (SM)].

Considering the large number of approaches that provide inference on influenza activity (*16*–*19*), does this mean that the current version of GFT is not useful? No, greater value can be obtained by combining GFT with other near–real-time health data (*2*, *20*). For example, by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, we can substantially improve on the performance of GFT or the CDC alone (see the chart). This is no substitute for ongoing evaluation and improvement, but, by incorporating this information, GFT could have largely healed itself and would have likely remained out of the headlines.

[1]Lazer Laboratory, Northeastern University, Boston, MA 02115, USA. [2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [4]University of Houston, Houston, TX 77204, USA. [5]Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. [6]Institute for Scientific Interchange Foundation, Turin, Italy.

*Corresponding author. E-mail: d.lazer@neu.edu.





D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis", Science, vol. 343, Mar. 14, 2914, pp. 1203-1205

（1）大数据是传统的数据收集和分析的替代品，而不是补充。我们断言大数据有巨大的科学可能性，但是，数据的体量并不意味着人们可以忽略测量的基本问题，构造效度和信度以及数据间的依赖关系。

（2）在谷歌为改善其服务中，也改变了数据生成过程。这些调整有可能人为推高了一些搜索，并导致谷歌的高估。

# 计算社会科学: 我们准备好了吗？

- 罗伯特·默顿（Robert Merton，1997年诺贝尔经济学奖得主）的名言：“**也许社会学还没有为爱因斯坦做好准备，因为它还没有找到它的开普勒……**”[Merton, R. K. in Social Theory and Social Structure 39–72 (Free Press, 1968).]
- 默顿这句挑衅性的话是指，社会学还没有建立伟大理论的经验基础。对此，邓肯·瓦茨（Duncan Watts，计算社会科学家）在43年后回应道：“**……通过将不可测量的变量变得可测量，移动、网络和互联网通信领域的技术革命有可能彻底改变我们对自己和我们与世界如何互动的理解**”[Watts, D. J. Everything Is Obvious: Once You Know the Answer (Crown Business, 2011)]。默顿是对的：社会科学仍然没有找到它的开普勒。

# 任课教师简介

## 金耀辉，上海交通大学·人工智能研究院 长聘教授

- 上海交通大学智慧法院研究院副院长、人工智能研究院总工程师
- 提升政府治理能力大数据应用技术国家工程实验室 专家委员会委员
- 木兰开源社区技术委员会委员
- 上海市公共数据开放专家委员会成员

---

**研究领域**：长期从事**人工智能和通信网络**应用研究，近年来专注数据治理与共享架构、 自然语言理解与深度学习、时空数据挖掘与应用，特别是人工智能与法律人文、公共管理**交叉研究**

**项目经验**：科技部863计划“高性能宽带通信网”重大专项任务专家组副组长，**科技部重点研发计划“全流程管控的精细化执行技术及装备研究”**（2018-2021）项目负责人，**科技部重点研发计划“法检司协同分布式大数据融合关键技术研究”**（2023-2026）项目负责人，**上海市科技进步一等奖**（2007）

**主要成果**：近5年在AAAI、IJCAI、MM、ACL、EMNLP、COLING等**人工智能领域顶会**上发表论文20+篇，拥有发明专利18项

# 任课教师简介

## 许岩岩，上海交通大学·人工智能研究院 长聘教轨副教授

- 博士后研究员 / UC Berkeley (2018-2020)
- 博士后研究员 / MIT (2015-2018)
- 客座博士后研究员 / 劳伦斯伯克利国家实验室 (2017-2018)

**研究领域：人工智能交叉应用、人类动力学、复杂系统建模与优化、计算城市科学**

### 科研课题：

- **国家海外优秀青年基金 (2021年)**
- 上海市海外领军人才（2021）
- 上海市浦江人才计划（2021）
- 国家自然基金青年项目（2021）
- 科技部重点研发子课题（2022）
- 科技合作：工商银行、eBay、联通数科、上海电信

### 学术成果：

- 发表SCI及人工智能会议学术论文30余篇
- **自然子刊 Nature Energy（IF=67）**
- **科学子刊 Science Advances（IF=14）**
- 人工智能顶级会议 IJCAI等
- J. R. Soc. Interface（英国皇家学会旗舰综合期刊）

nature energy

ScienceAdvances

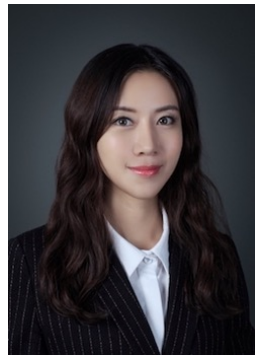
# 课程学术顾问

Prof. Marta Gonzalez
DCRP & CEE
UC Berkeley
研究方向：人类动力学、计算社会科学、复杂系统

杨力 长聘教授
凯原法学院
智慧法院研究院
研究方向：智慧司法、社会治理

何浩 副教授
电子信息与电气工程学院
研究方向：自然语言处理、法律科技

史冬波 副研究员
国际与公共事务学院
研究方向：创新经济学，战略管理



李春晓 助理教授
安泰经济与管理学院
研究方向：信息系统、普惠金融

# 课程助教

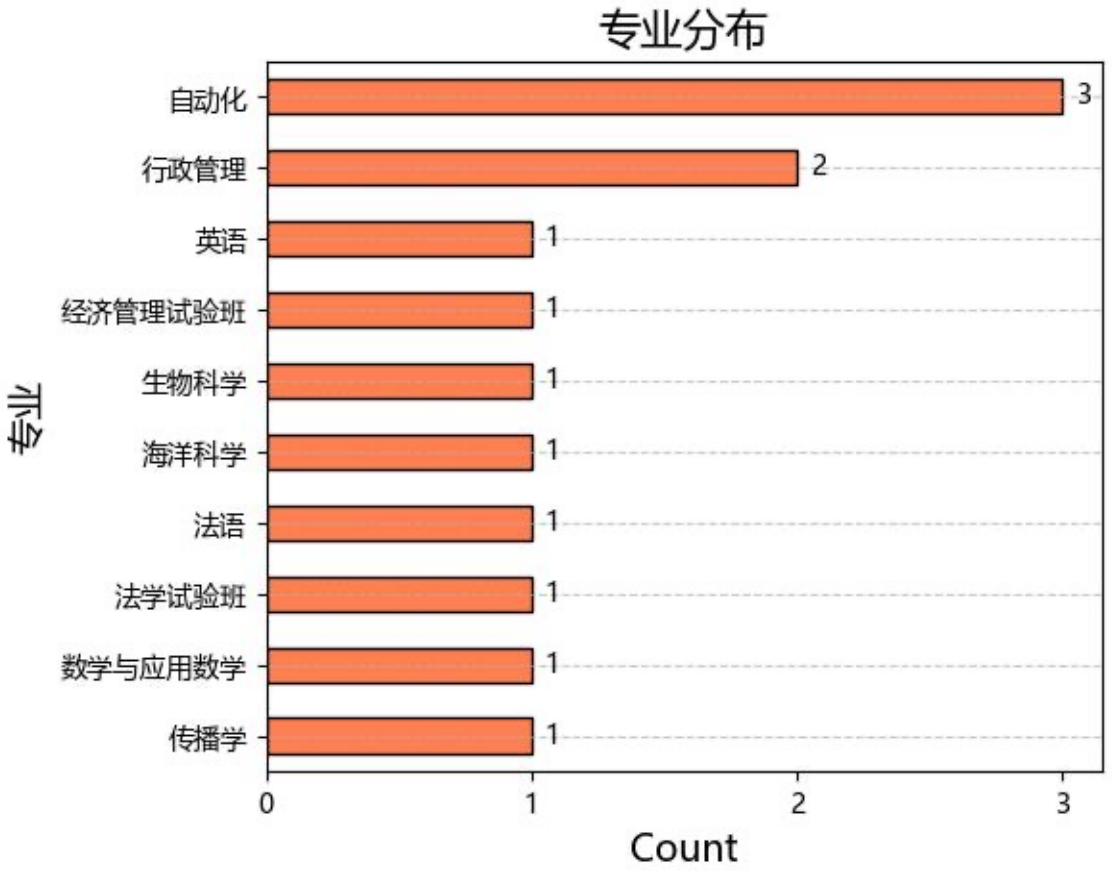
赵帅
人工智能研究院 博士后
凯原法学院 法学博士
研究方向：计算法学
shuaizhao@sjtu.edu.cn

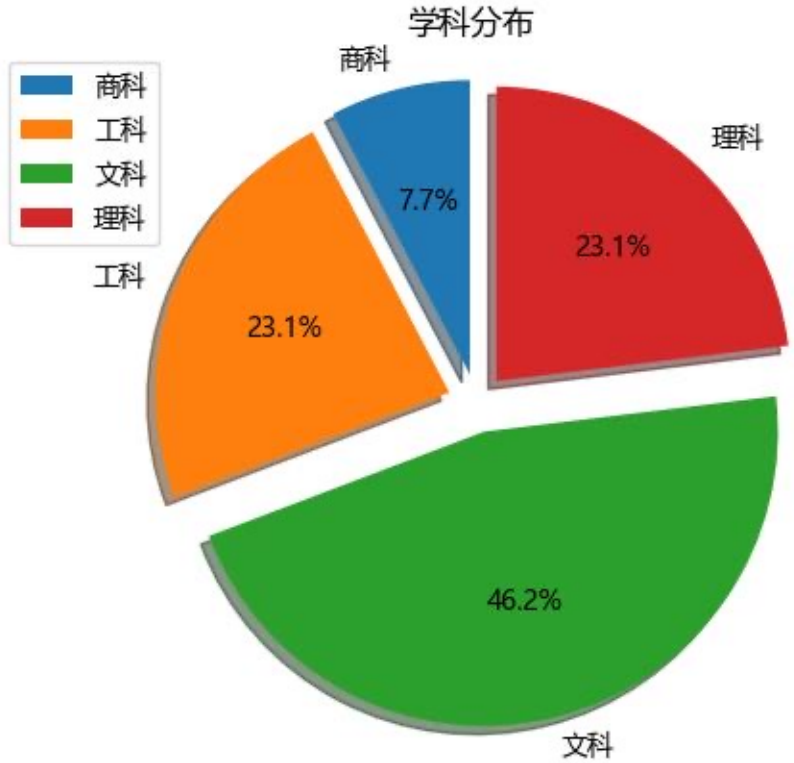
黄劭煜
人工智能研究院
2022级博士研究生
研究方向：计算城市科学
leak_ish@sjtu.edu.cn

胡钊萍
人工智能研究院
2021级硕士研究生
研究方向：时空数据挖掘
zhaopinghu@sjtu.edu.cn

袁凡露
凯原法学院
2022级硕士研究生
研究方向：智慧司法
fanlu_yuan@sjtu.edu.cn

# 关于你们

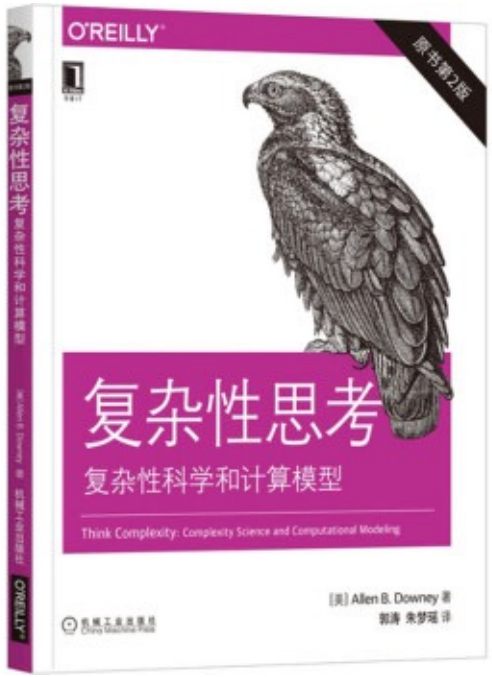

专业分布
自动化
行政管理
英语
经济管理试验班
生物科学
海洋科学
法语
法学试验班
数学与应用数学
传播学
3
2
1
1
1
1
1
1
1
0
1
2
3
Count


学科分布
商科
工科
文科
理科
7.7%
23.1%
23.1%
46.2%

# 参考教材

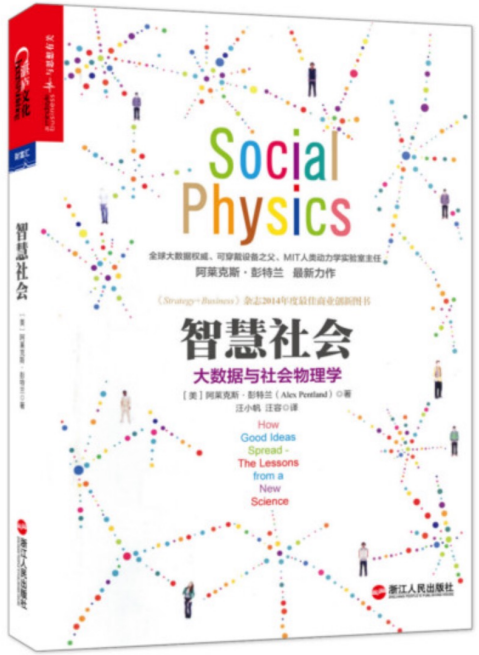

**计算社会科学：数字时代社会研究的新方法**

郝龙
南京航空航天大学人文与社会科学学院讲师




**复杂性思考：复杂性科学和计算模型**

Allen B．Downey
富兰克林欧林工程学院的计算机科学副教授
*Think Python, Think Bayes, Think Stats*等书的作者




**智慧社会：大数据与社会物理学**

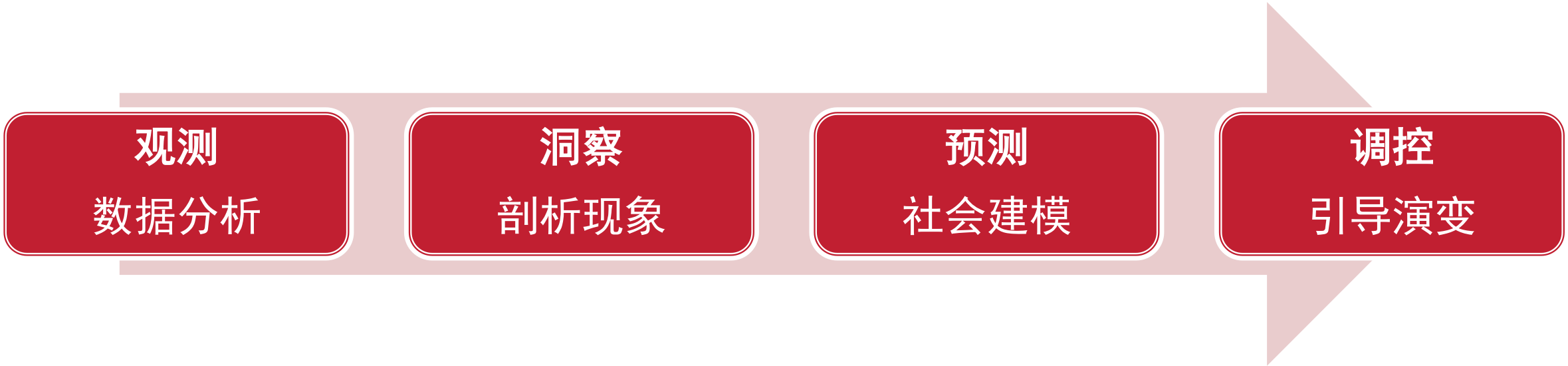
阿莱克斯•彭特兰（Alex Pentland）
MIT人类动力学实验室主任
Media Lab联合创始人
全球大数据专业、可穿戴设备之父


**计算社会学：数据时代的社会研究**

马修·萨尔加尼克（Matthew J. Salganik）普林斯顿大学社会学教授，《科学》杂志评价他是“纯然的计算社会学家”。

# 课程简介与课程目标

- 随着大数据与人工智能技术的发展，科学家可以利用数据驱动和自然科学的方法量化人类行为与建模复杂的社会现象，提升人类社会福祉，也催生了理、工、文交叉的当代社会科学研究新范式——**计算社会科学**。
- 本课程将**介绍计算社会科学研究问题与研究方法，引导学生从交叉学科、系统科学的视角探究社会发展进程中的重要问题及学习潜在解决方案。**



# 课程结构

| 章节 | 教学内容 | 学时 | 学习目标 |
| --- | --- | --- | --- |
| 第一章 | 介绍计算社会科学的发展历程、基本概念和问题 | 2 | |
| 第二章 | 大数据视角下的计算社会科学典型研究方法 | 2 | |
| 第三章 | 社会数据处理与分析方法 | 4 | 使用Python语言完成经济数据处理与可视化 |
| 第四章 | 计算社会科学典型研究场景 | 4 | 寻找现实生活可利用数据研究的社会现象或场景，并给出解决思路 |
| 第五章 | 复杂网络科学 | 4 | |
| 第六章 | 社会系统建模方法 | 6 | 使用Python建立简易的ABM模型，实现传染病传播模拟 |
| 第七章 | 社会数据挖掘与人工智能分析方法 | 6 | 使用Python编写AI模型，实现对居民出行需求的预测 |
| 第八章 | 经典研究工作解读 | 2 | |
| 第九章 | 小组课程项目汇报 | 2 | 完成项目答辩，提交课程项目论文 |

# 考核方式

（1）出勤与课堂表现：10%

（2）平时作业: 30%

（3）课程项目 60%，其中课程项目演讲30%，课程论文30%。注：每个项目小组由3名同学组成，1名老师或博士生单独指导，**课程论文鼓励投稿发表**。

**学术规范**

- 在规定的时间之前提交作业
- 鼓励交流讨论，但是作业要独立完成

# Office Hour

金耀辉：

Email: jinyh@sjtu.edu.cn

许岩岩：

Email: yanyanxu@sjtu.edu.cn


Group: CS1126 计算社会科学导论

Valid until 2/19 and will update upon joining group

Office Hour:

周五下午 3:00-5:00，软件大楼5-512

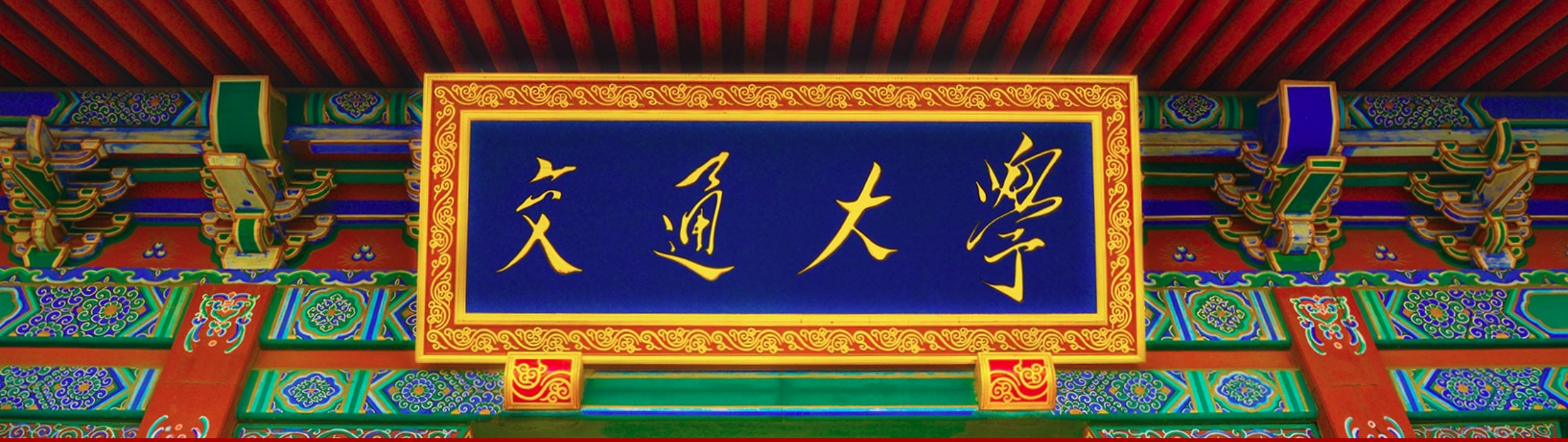
在线答疑：微信群，canvas讨论。 课程网址：https://urbanmobility.github.io/CompSocSci/

# 课前调查

CS1126 计算社会科学导论 课前调查

问卷仅为了解同学们的相关技能掌握情况，便于后续合理安排课程内容，

腾讯问卷




# 计算社会科学导论
# —— 计算社会科学发展简介

金耀辉、许岩岩

2023年2月16日

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

# 什么是计算社会科学

- **From Wikipedia:**
- **Computational sociology** is a branch of sociology that uses computationally intensive methods to analyze and model social phenomena. Using computer simulations, artificial intelligence, complex statistical methods, and analytic approaches like social network analysis, computational sociology develops and tests theories of complex social processes through bottom-up modeling of social interactions.

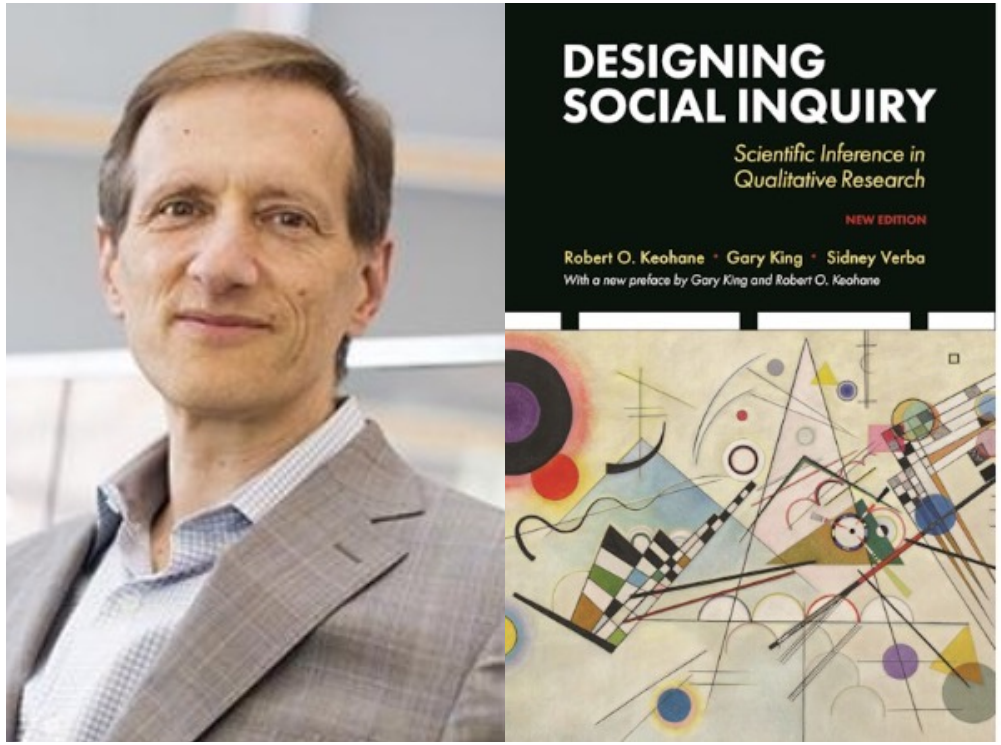
计算社会学是社会学的一个分支，它使用计算密集型方法来分析和模拟社会现象。使用计算机模拟、人工智能、复杂的统计方法和社会网络分析等分析方法，计算社会学通过对社会互动进行自下而上的建模，发展和测试复杂社会过程的理论。

# 计算社会科学发展历史-缘起

计算社会科学的“前身”：定量社会科学的发展。

按照哈佛大学定量社会科学研究所主任、著名方法论学者加里·金（Gary King）的解释，**定量社会科学**不仅是指长期以来社会科学发展历程中相对于质化研究而言的量化研究方法，而且是经历了近十年的突破发展而引发当代社会科学转型的一股浪潮。

King G, Keohane R O, Verba S. Designing Social Inquiry: Scientific Inference in Qualitative Research[J]. 1994.

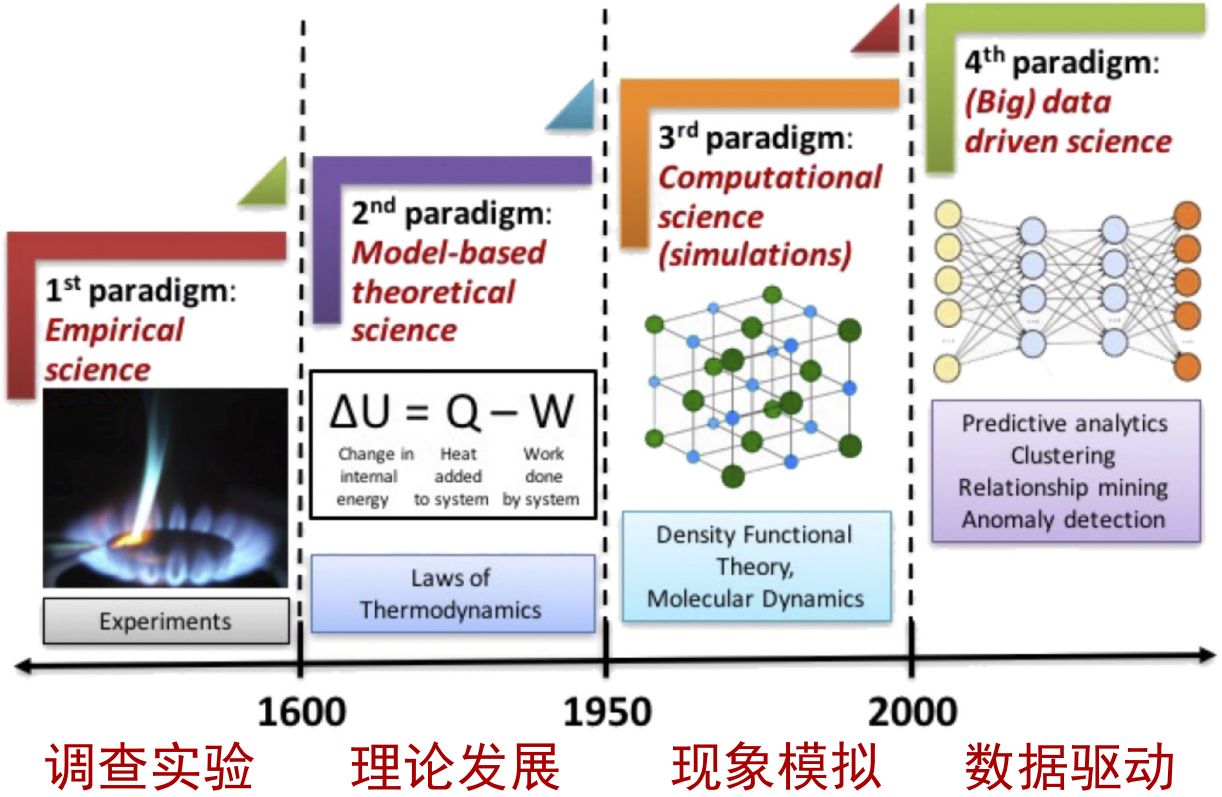



Gary King（1958 - ）

# 计算社会科学发展历史-缘起

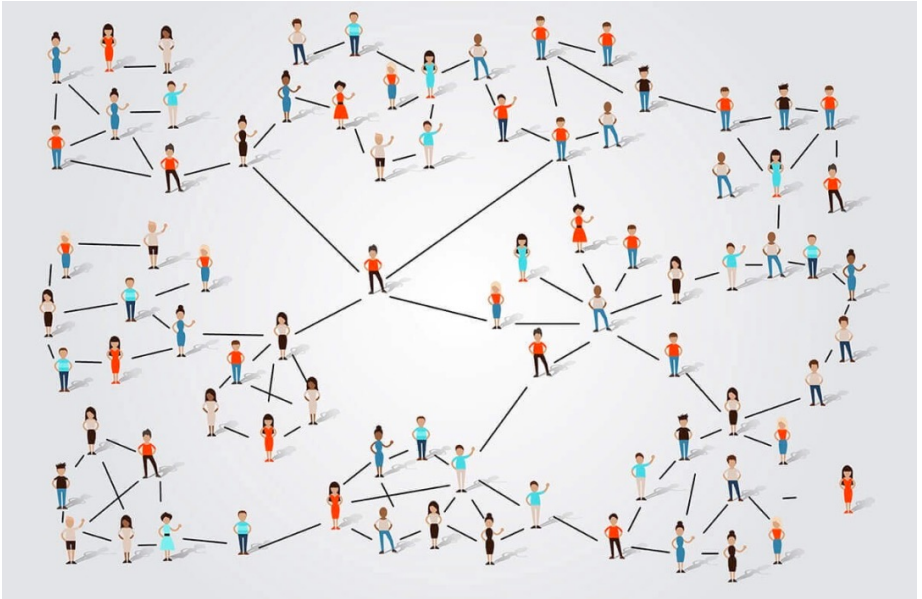
计算社会科学的创生：新兴学科本体论的探讨。

20 世纪后半叶以来，社会科学发展的趋势正是研究者自发地使用海量数据开展以纯理论或应用为目的研究。



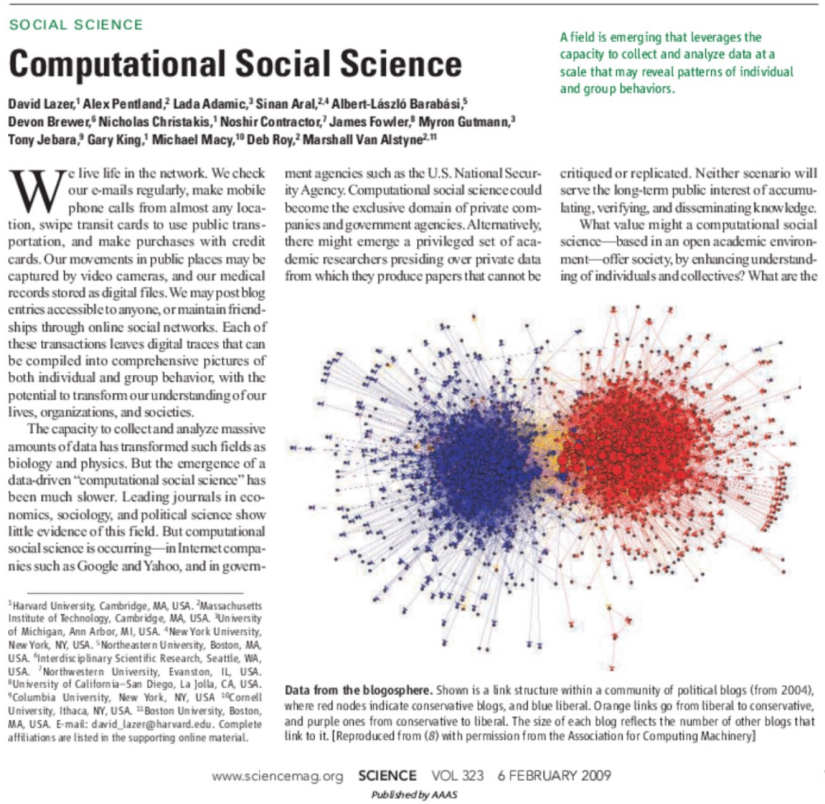
# 计算社会科学的发展蓄势 —— 网络科学

- 社会系统涉及到大量异构实体的相互作用，**网络科学**迅速发展为社会分析提供了强有力的分析手段。
- 2007年，“小世界网络之父”奠基人邓肯·瓦茨在 Nature 发表了题为《A twenty-first century science》的文章，这成为计算社会科学时代即将来临的标志之一。这篇文章采用了网络分析的方法来分析社会现象中的网络偏好以及个体选择的问题。



小世界网络，源自对六度空间理论的类比，是一种介于随机网络和规则网络间的模型

# 计算社会科学 —— 社会科学的全新研究范式



SOCIAL SCIENCE

## Computational Social Science

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.

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,[3] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

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 Institute of Technology, Cambridge, MA, USA. [3]University of Michigan, Ann Arbor, MI, USA. [4]New York University, New York, NY, USA. [5]Northeastern University, Boston, MA, USA. [6]Interdisciplinary Scientific Research, Seattle, WA, USA. [7]Northwestern University, Evanston, IL, USA. [8]University of California–San Diego, La Jolla, CA, USA. [9]Columbia University, New York, NY, USA. [10]Cornell University, Ithaca, NY, USA. [11]Boston University, Boston, MA, USA. E-mail: david_lazer@harvard.edu. Complete affiliations are listed in the supporting online material.

**Data from the blogosphere.** Shown is a link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. [Reproduced from (*8*) with permission from the Association for Computing Machinery]



D. Lazer et al., 2009, Science

Science

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.

**新的研究领域正在浮现，它利用大规模数据采集和分析能力，试图揭示社会系统中的个人和群体行为模式。**

# 计算社会科学 —— 大数据带来的机遇

- **在2012年，R. Conte，C. Cioffi-Revilla等14位欧美学者在《The European Physical Journal Special Topics》（第1期）上联合发布了一份《计算社会科学宣言》，力图呼唤一场社会科学革命。**
- The increasing integration of technology into our lives has created unprecedented volumes of data on society' s everyday behaviour. Such data opens up exciting new opportunities to work towards a quantitative understanding of our complex social systems, within the realms of a new discipline known as Computational Social Science. Against a background of financial crises, riots and international epidemics, the urgent need for a greater comprehension of the complexity of our interconnected global society and an ability to apply such insights in policy decisions is clear.

Springer Link

Regular Article | Open Access | Published: 05 December 2012

Manifesto of computational social science

R. Conte, N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J. -P. Nadal, A. Sanchez, A. Nowak, A. Flache, M. San Miguel & D. Helbing

*The European Physical Journal Special Topics* **214**, 325–346 (2012) | Cite this article

**11k** Accesses | **217** Citations | **29** Altmetric | Metrics

技术越来越多地融入我们的生活，为社会的日常行为创造了前所未有的大量数据。这些数据为我们提供了令人兴奋的新机会，使我们能够在一个被称为 "计算社会科学 "的新学科范围内，对我们复杂的社会系统进行量化理解。**在金融危机、暴乱和国际流行病的背景下，显然迫切需要对我们相互关联的全球社会的复杂性有更多的了解，并有能力在政策决定中应用这些见解。**

Conte, R. *et al. Eur. Phys. J. Spec. Top.2012*

# 计算社会科学 —— 社会科学的全新研究范式



POLICY FORUM

SOCIAL SCIENCE

## Computational social science: Obstacles and opportunities

Data sharing, research ethics, and incentives must improve

By David M. J. Lazer[1,2], Alex Pentland[3], Duncan J. Watts[4], Sinan Aral[3], Susan Athey[5], Noshir Contractor[6], Deen Freelon[7], Sandra Gonzalez-Bailon[4], Gary King[2], Helen Margetts[8,9], Alondra Nelson[10,11], Matthew J. Salganik[12], Markus Strohmaier[13,14], Alessandro Vespignani[1], Claudia Wagner[14,15]

The field of computational social science (CSS) has exploded in prominence over the past decade, with thousands of papers published using observational data, experimental designs, and large-scale simulations that were once unfeasible or unavailable to researchers. These studies have greatly improved our understanding of important phenomena, ranging from social inequality to the spread of infectious diseases. The institutions supporting CSS in the academy have also grown substantially, as evidenced by the proliferation of conferences, workshops, and summer schools across the globe, across disciplines, and across sources of data. But the field has also fallen short in important ways. Many institutional structures around the field—including research ethics, pedagogy, and data infrastructure—are still nascent. We suggest opportunities to address these issues, especially in improving the alignment between the organization of the 20th-century university and the intellectual requirements of the field.

We define CSS as the development and application of computational methods to complex, typically large-scale, human (sometimes simulated) behavioral data (*1*). Its intellectual antecedents include research on spatial data, social networks, and human coding of text and images. Whereas traditional quantitative social science has focused on rows of cases and columns of variables, typically with assumptions of independence among observations, CSS encompasses language, location and movement, networks, images, and video, with the application of statistical models that capture multifarious dependencies within data. A loosely connected intellectual community of social scientists, computer scientists, statistical physicists, and others has coalesced under this umbrella phrase.

### MISALIGNMENT OF UNIVERSITIES

Generally, incentives and structures at most universities are poorly aligned for this kind of multidisciplinary endeavor. Training tends to be siloed. Integrating computational training directly into social science (e.g., teaching social scientists how to code) and social science into computational disciplines (e.g., teaching computer scientists research design) has been slow. Collaboration is often not encouraged, and too often is discouraged. Computational researchers and social scientists tend to be in different units in distinct corners of the university, and there are few mechanisms to bring them together. Decentralized budgeting models discourage collaboration across units, often producing inefficient duplication.

Research evaluation exercises such as the United Kingdom's Research Excellence Framework, which allocate research funding, typically focus within disciplines, meaning that multidisciplinary research may be less well recognized and rewarded. Similarly, university promotion procedures tend to underappreciate multidisciplinary scholars. Computational research infrastructures at universities too often cannot fully support analysis of large-scale, sensitive data sets, with the requirements of security, access to a large number of researchers, and requisite computational power. To the extent these issues have been partially resolved in the academy (e.g., with genomic data), lessons have not fully made their way into practice in CSS.

### INADEQUATE DATA-SHARING PARADIGMS

Current paradigms for sharing the kinds of large-scale, sensitive data used in CSS offer a mixed bag. There have been successes built on partnerships with government, especially in economics, from the study of inequality (*2*) to the dynamics of labor markets (*3*). There are emerging, well-resourced models of administrative data research facilities serving as platforms for analyzing microlevel data while preserving privacy (*4*). These offer important lessons for potential collaboration with private companies, including the development of methodologies to keep sensitive data secure, yet accessible for analyses (e.g., innovations in differential privacy).

The value proposition for private companies is different and there has been predictably less progress. Data possessed by government agencies are held in trust for the public, whereas data held by companies are typically seen as a key proprietary asset. Public accountability inherent in sharing data is likely seen as a positive for the relevant stakeholders for government agencies, but generally, far less so for shareholders for private companies. Access to data from private companies is thus rarely available to academics, and when it is, it is typically granted through a patchwork system in which some data are available through public application programming interfaces (APIs), other data only by working with (and often physically in) the company in question, and still other data through personal connections and one-off arrangements, often governed by nondisclosure agreements and subject to potential conflicts of interest. An alternative has been to use proprietary data collected for market research (e.g., Comscore, Nielsen), with methods that are sometimes opaque and a pricing structure that is prohibitive to most researchers.

We believe that this approach is no longer acceptable as the mainstay of CSS, as pragmatic as it might seem in light of the apparent abundance of such data and limited resources available to a research community in its infancy. We have two broad concerns about data availability and access.

First, many companies have been steadily cutting back data that can be pulled from their platforms (*5*). This is sometimes for good reasons—regulatory mandates (e.g., the European Union General Data Protection Regulation), corporate scandal (Cambridge Analytica and Facebook)—however, a side effect is often to shut down avenues of potentially valuable research. The susceptibility of data availability to arbitrary and unpredictable changes by private actors, whose cooperation with scientists is strictly voluntary, renders this system intrinsically unreliable and potentially biased in the science it produces.

[1]Northeastern University, Boston, MA, USA. [2]Harvard University, Cambridge, MA, USA. [3]Massachusetts Institute of Technology, Cambridge, MA, USA. [4]University of Pennsylvania, Philadelphia, PA, USA. [5]Stanford University, Stanford, CA, USA. [6]Northwestern University, Evanston, IL, USA. [7]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [8]University of Oxford, Oxford, UK. [9]The Alan Turing Institute, London, UK. [10]Institute for Advanced Study, Princeton, NJ, USA. [11]Social Science Research Council, New York, NY, USA. [12]Princeton University, Princeton, NJ, USA. [13]RWTH Aachen University, Aachen, Germany. [14]GESIS–Leibniz Institute for the Social Sciences, Cologne, Germany. [15]University of Koblenz-Landau, Landau, Germany. Email: d.lazer@northeastern.edu





## 障碍

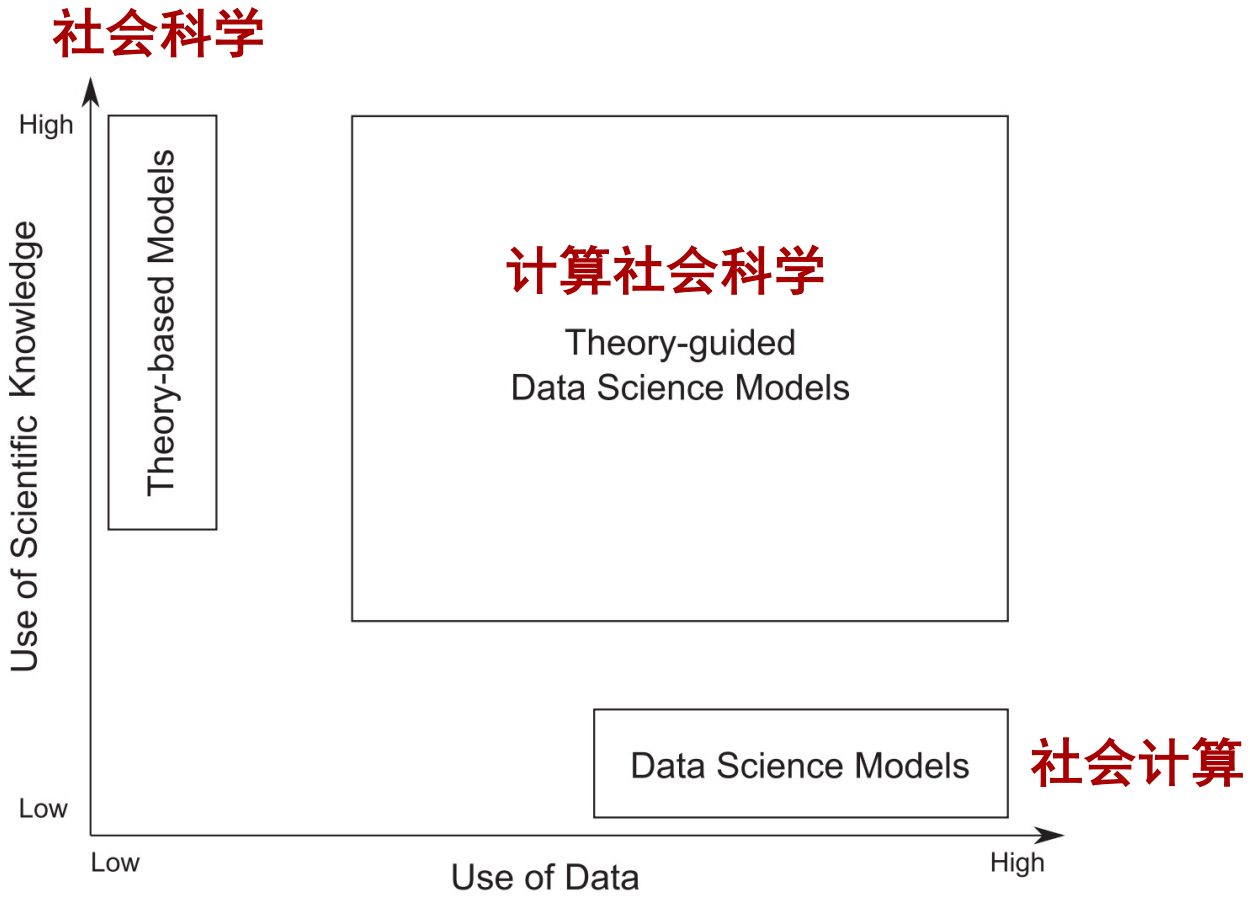
- 大学教育和研究缺位
- 不充分的数据共享范式
- 隐私数据规则不完善

## 建议

- 加强研究界和工业界的合作
- 新的数据基础设施
- 伦理、法律和社会影响
- 大学加强跨专业交流
- 解决现实世界的问题

D. Lazer et al., 2020, Science

# 计算社会科学 VS. 社会计算




- **社会科学：**

  重理论、轻数据

- **社会计算：**

  重数据、轻理论

- **计算社会科学：**

  理论与数据融合

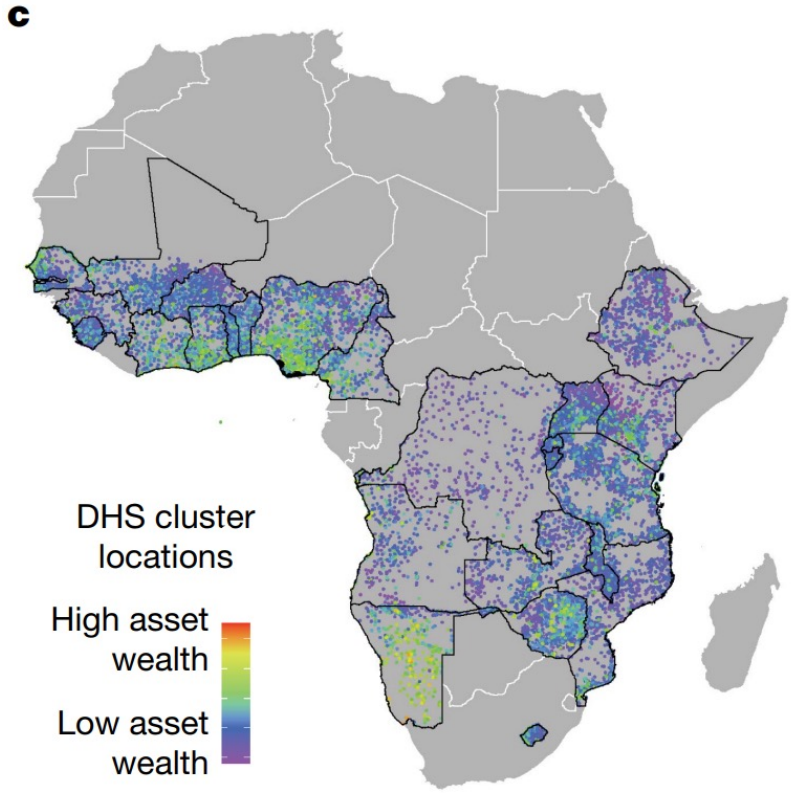
Ref: · Karpatne et al. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Transactions on Knowledge and Data Engineering, 2017

# 计算社会科学典型案例一

- **机器学习携手卫星影像，理解电力设施与经济财富的因果关系（Nature 封面文章）**
- 背景：乌干达在2010-2019年将电气覆盖率从12%提升到了41%
- 问题：电网扩张如何影响低收入地区经济产出？缺少不同时/空的统计数据？因果证据？
- 利用遥感和机器学习模型预测精细空间粒度下的乌干达资产财富指标；

  利用多光谱遥感数据为输入，基于**撒哈拉以南非洲地区（SSA）**27000个村庄的统计调查数据中的多项相关指标，构造资产财富指数作为标签，使用深度学习估计25个非洲国家2005-2018年的资产财富状况；

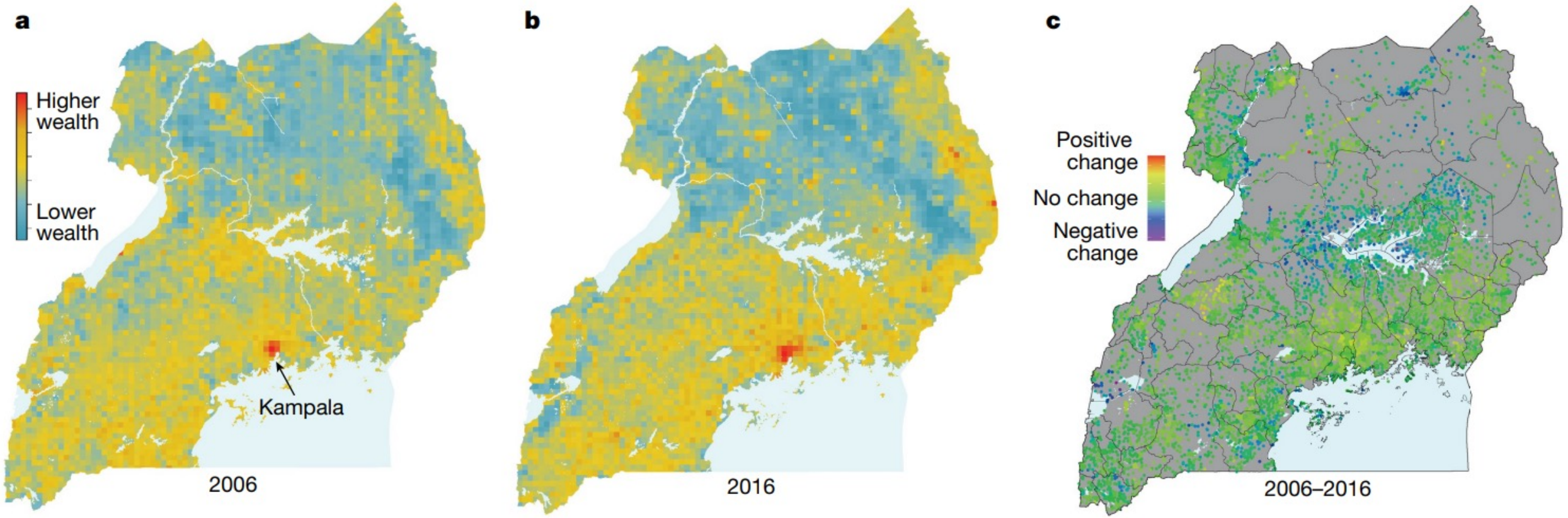
  **填补了超过已有数据10倍以上的空白数据**




Ratledge, N., Cadamuro, G., de la Cuesta, B. *et al.* Using machine learning to assess the livelihood impact of electricity access. *Nature* **611**, 491–495 (2022).

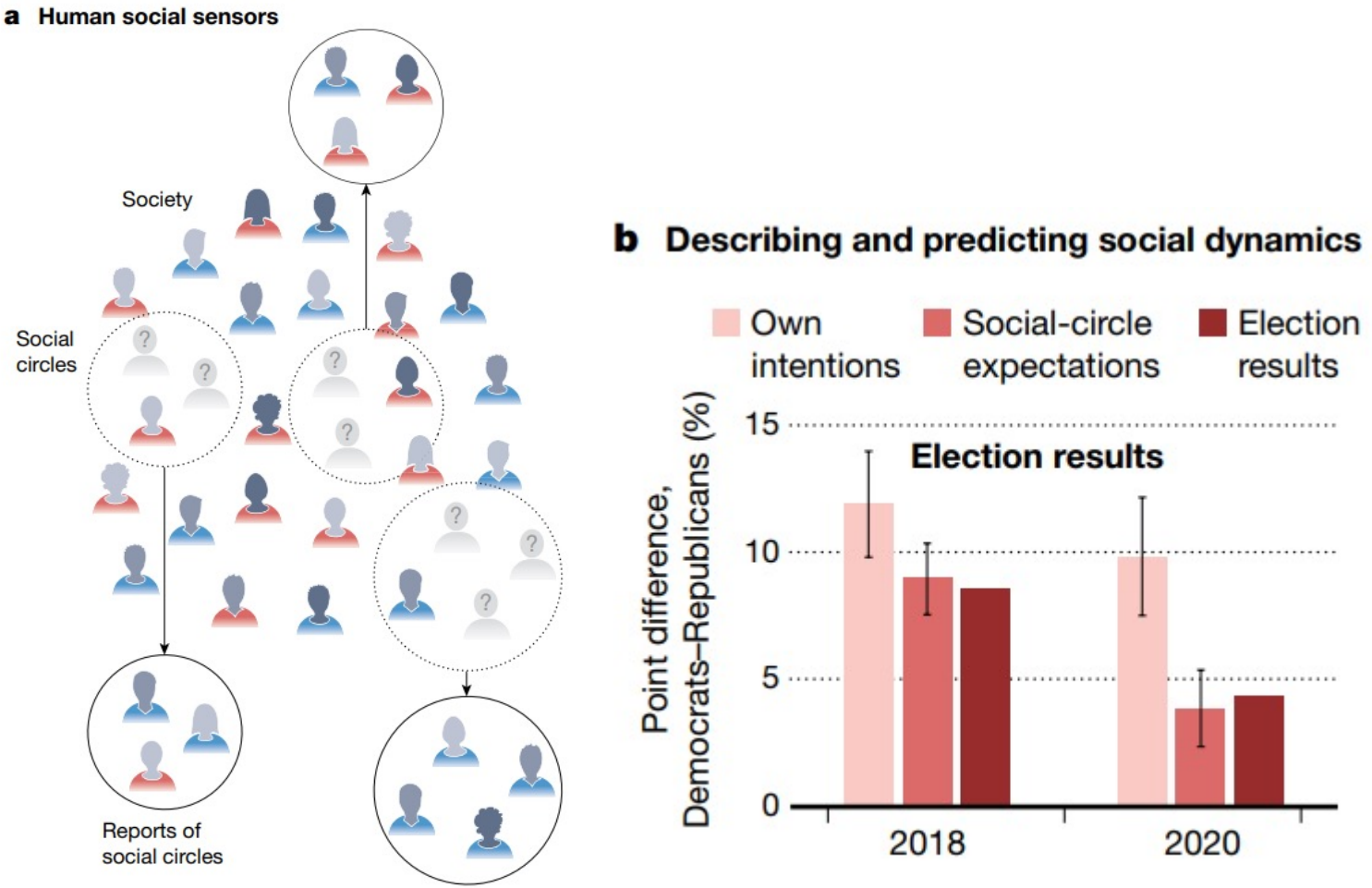
# 计算社会科学典型案例一

- 资产财富预测模型展现出十年间的区域财富增长情况




- a,b:2006和2016年的模型预测结果；c:6900个人口密集村庄的10年间财富差异

# 计算社会科学典型案例二




- Human social sensing
- 人类社会感知可以为研究信仰和行为（beliefs and behaviors）提供更有效的信息。
- 例如2018和2020美国选举中，对社交圈的描述（properties of their immediate social environments）可以比询问自身倾向的抽样调查更好地预测选举结果。

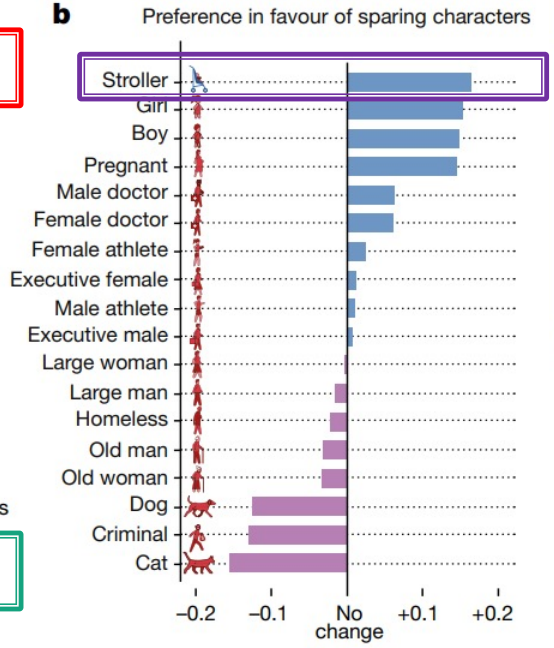
你的朋友可能比你更了解你！

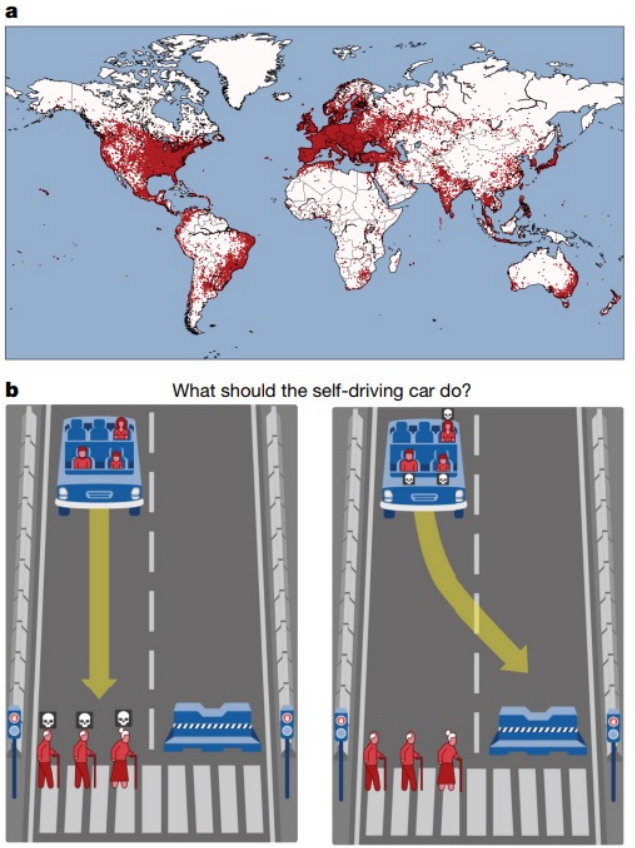
Galesic, M., Bruine de Bruin, W., Dalege, J. *et al.* Human social sensing is an untapped resource for computational social science. *Nature* **595**, 214–222 (2021).

# 计算社会科学典型案例三

- The Moral Machine experiment，一项全球性的道德实验
- 面临电车难题时，自动驾驶应该遵循什么样的道德准则？
- 全球结果中的道德偏好：

Awad, E., Dsouza, S., Kim, R. *et al.* The Moral Machine experiment. *Nature* **563**, 59–64

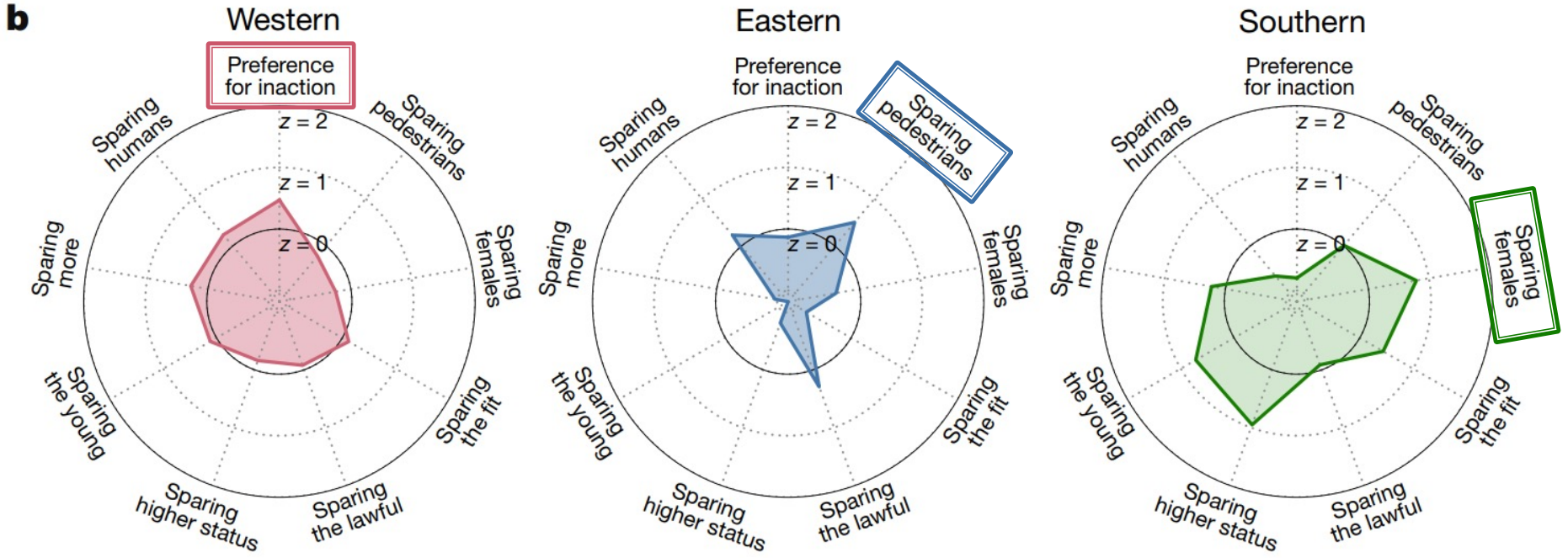



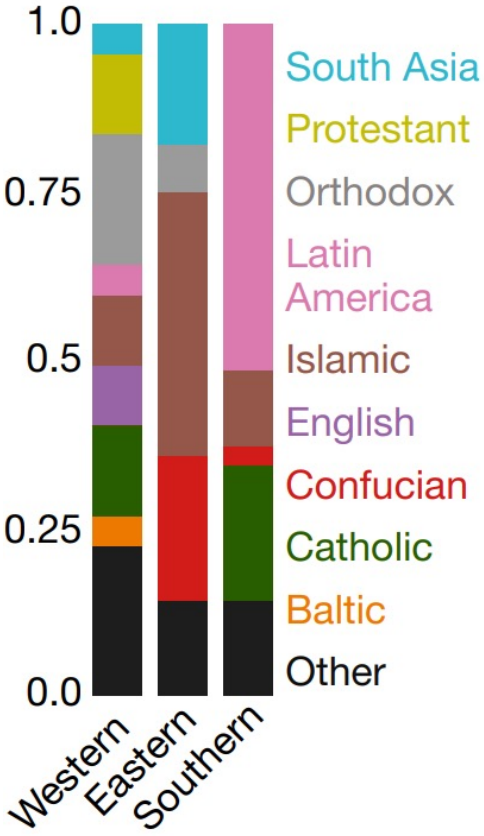




- 不转弯比转弯的概率大0.05；救人类的概率比救宠物高约0.6
- 相较于拯救一个人类，救下一个婴儿车的概率高了约0.15;

# 计算社会科学典型案例三

- The Moral Machine experiment，一项全球性的道德实验
- 面临电车难题时，自动驾驶应该遵循什么样的道德准则？
- 例如西方人可能更相信命运，选择不做出动作的更多；东方国家在拯救行人方面高于其他两个区域；南方国家的选择更倾向于拯救女性...




国家分类，东方以伊斯兰、南亚和孔子文明国家为主

wad, E., Dsouza, S., Kim, R. *et al.* The Moral Machine experiment. *Nature* **563**, 59–64

# 交大案例

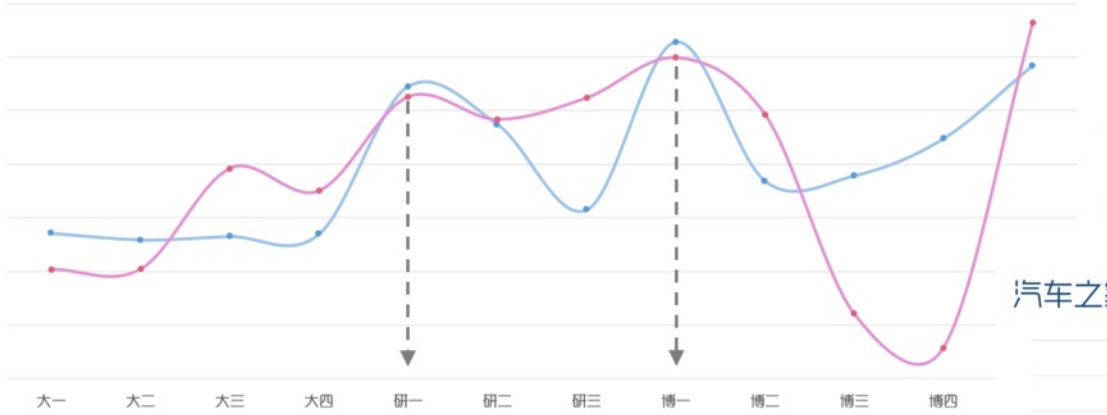

EMC杯 智慧校园开放数据大赛
2015年4月1日～5月20日 2~5人一组
宣讲会：4月9日19:00，网络中心8楼
想了解交大学生的饮食习惯吗?
一二三四五六餐，哪家是小伙伴们的最爱?
总是面对茫茫人海，怎样错峰选择吃饭时间?
想知道交大学生喜欢上哪些网站吗?
哪个时间段是网络流量的最高峰?
中国知网VS某宝，谁的点击率会更高?
奖项众多，总有一款适合你!
更有￥25000奖金等你来拿!
最佳可视化设计
最佳政策建议
最佳R语言应用
最佳Excel应用
最佳数据工程
全能综合大奖
赛事评委
金耀辉
电子信息与电气工程学院, 教授
网络信息中心副主任
博士生导师
张鹏翥
安泰经济与管理学院, 教授
管理信息系统专业主任
博士生导师
韩东
数学系, 系主任, 教授
博士生导师
郑磊
复旦数字与移动治理实验室, 主任
国际关系与公共事务学院院长助理
副教授
韩挺
媒体与设计学院, 副院长
工业设计专业主任
张娅
电子信息与电气工程学院, 研究员
博士生导师
符冰
网络信息中心
服务部主管
蔡远进
后勤集团
财务总监
陈春曦
EMC中国卓越研发集团
上海公司总经理
主办单位：上海交大网络信息中心
指导单位：上海交大学生工作指导委员会 赞助单位：EMC卓越研发集团
承办单位：Kesci校园比赛平台 协办单位：上海交大电院学生会、上海交大数据分析俱乐部
合作媒体：
大数据文摘
大数据
数据科学家联盟
统计之都
更多详情
扫描二维码

# 交大案例：数据集

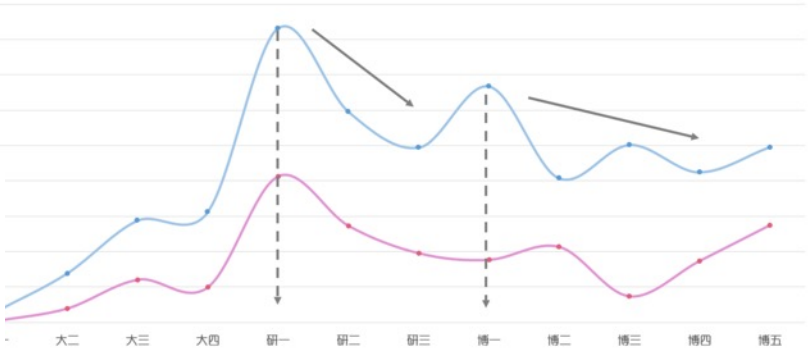
- 一卡通消费记录
  - 用户信息表（总数30800）
    - 卡号(X)、学号(X)、性别、出生年、入学年、本硕博[X表示做过去隐私处理]
  - 商户信息表（总数135，闵行校区）
    - 代码、所属食堂、窗口名称、位置、开户时间
  - 交易流水（总数420万条，时间跨度20140801-20150131）
    - 卡号、代码、时间、金额
- WiFI上网记录
  - 用户信息（总数20000+）
    - ID(X)，性别、年龄、入学年
  - 上网记录（总数1200万条，时间跨度20140901-20150131）
    - ID、上网地点、开始时间、结束时间、发送HTTP请求数、通信字节数、服务提供商、服务类型、域名
    - 1000, 东上院, 1412229603742, 1412229611551, 4, 11656, 腾讯微信, 即时通讯, qq.com
- 气象记录（时间跨度20140815-20150325，每10分钟一个采样点）
    - 气温 降水 瞬时风向 瞬时风速 瞬时能见度 相对湿度 瞬时气压 一小时最高气温 一小时最低气温 一小时最大风速 一小时最大风速时风向

# 精彩的结果（1）

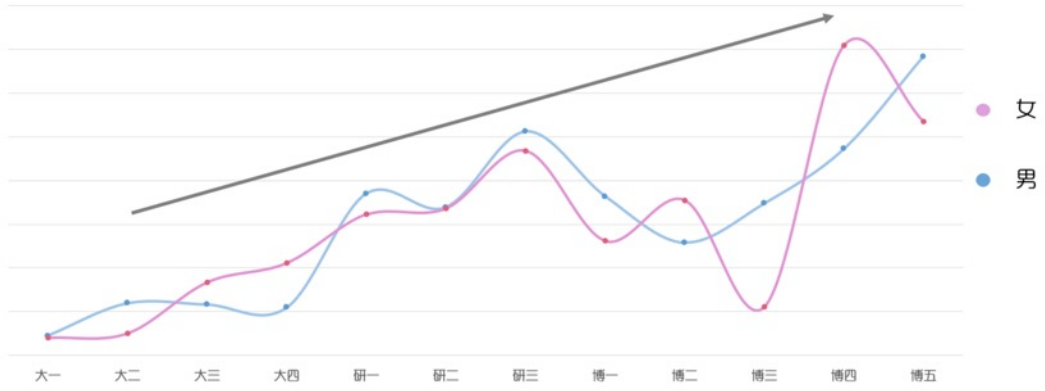
世纪佳缘告诉我们：




汽车之家告诉我们：

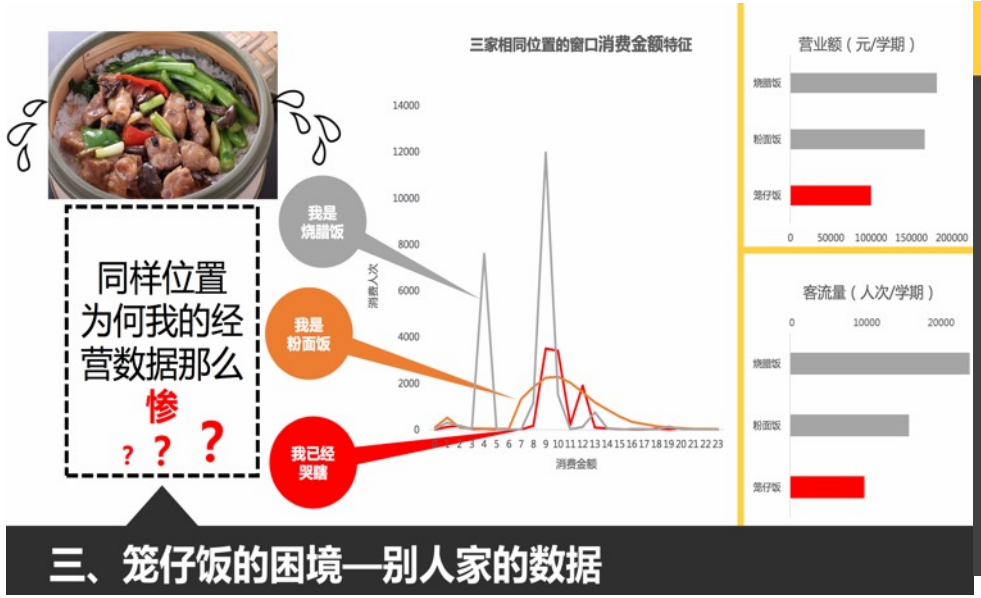


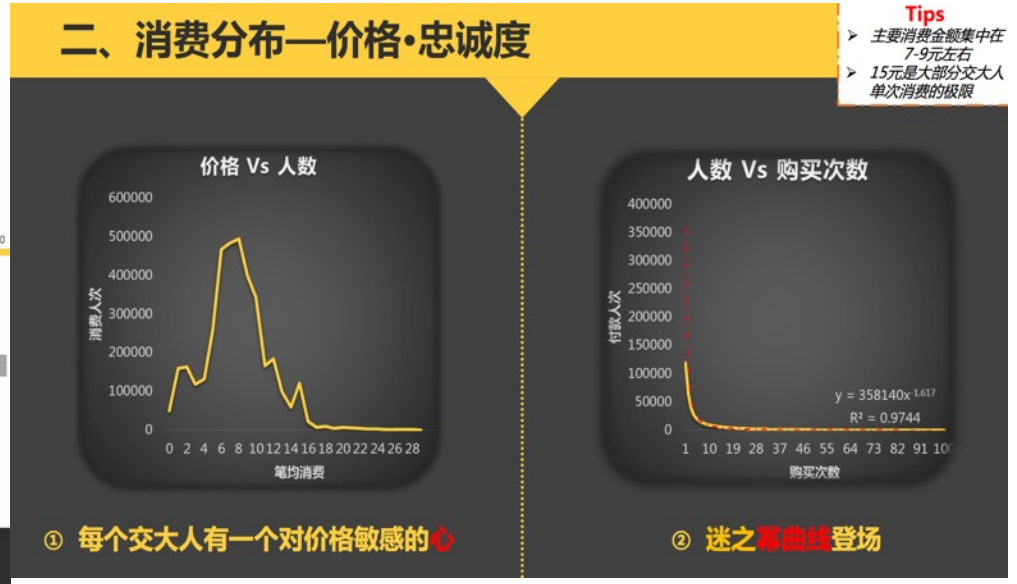

安居客告诉我们：

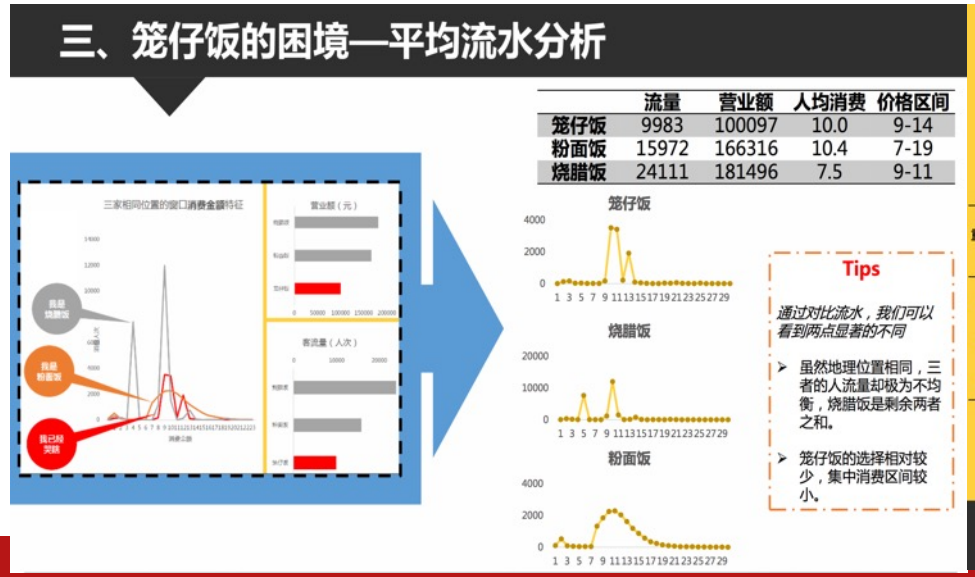


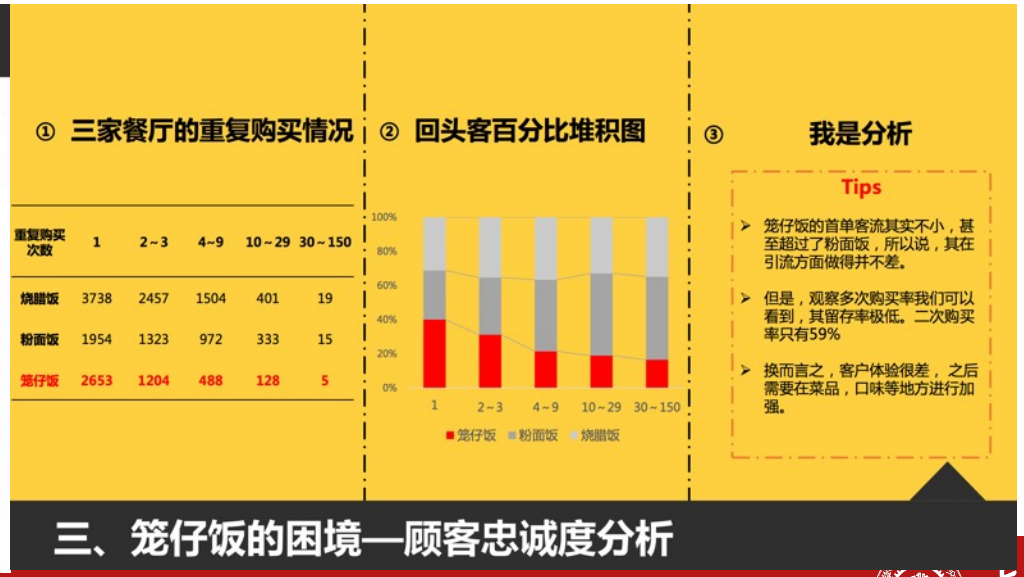

不同年级性别关注不同类别网站的分析

# 精彩的结果（2）



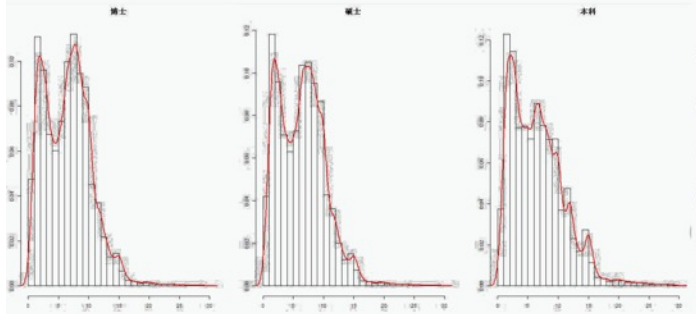



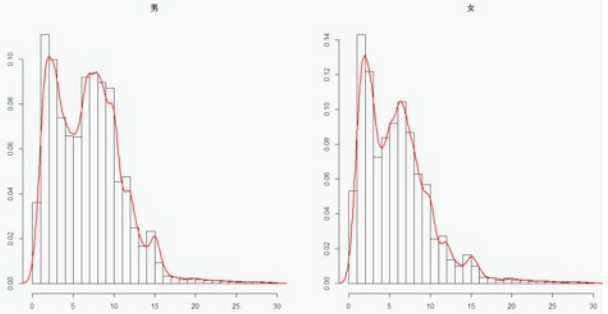
# 精彩的结果（3）

## 6. 每餐要花多少钱？

不同类型同学单次刷卡金额频率图



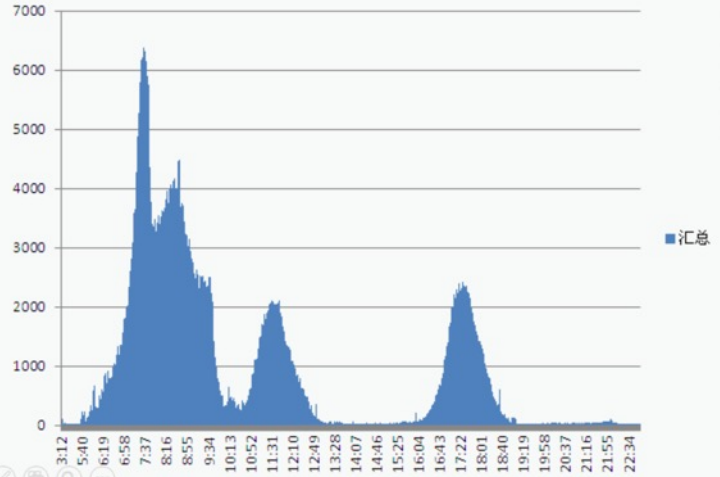
不同性别同学单次刷卡金额频率图



●焦点1：同学们每餐刷卡金额存在两个峰值，第一个为2.5元，第二个为6-9元，可推断，第一个峰是早餐，第二个峰是正餐，研究生峰值明显，研究生正餐花更多钱（7元以上）的频率大于本科生。
●焦点2：男生正餐花更多钱的频率大于女生。
●焦点3：各个食堂也会比较明显地呈现出这个趋势，而如五餐这样早餐规模较小的食堂则没有明显的第一个峰。

## 6.1 有趣的2.5元

各时段校园卡单次消费2.5元以下的频数




●焦点1：单次消费2.5元以下的情况明显出现了四个峰，第一次7:20是本科生吃早饭的时间，第二次在8:30是研究生吃早饭的时间。
●焦点2：午饭和晚饭期间分别出现了一次峰值，推测为正餐之外加了粥、包子、饼之类的辅食，加餐的频数远远低于每天吃早饭的频数。

# 精彩的结果（4）

## 人群聚类

**CLASS 1：女神or男神**

女生

**CLA：潜力股or学术帝**

男生 博士比例最高，相对于CLASS 1生 更规律、去
食堂 更多

**CLA：土豪**

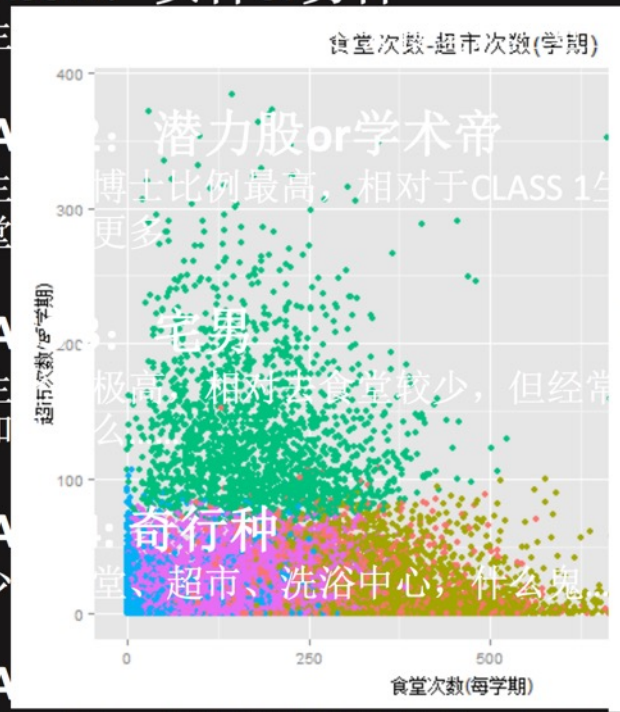
男生 极高，相对去食堂较少，但经常 超市，平均
面和 么

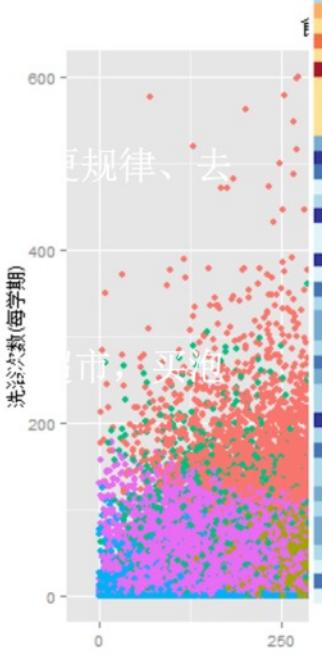
**CLA：奇行种**

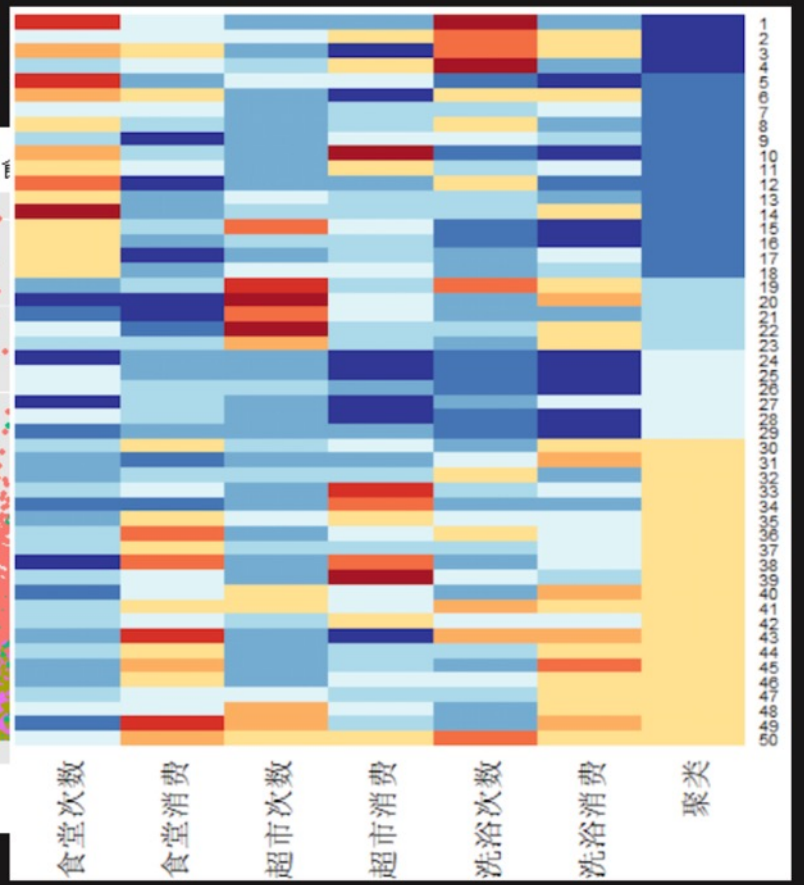
很少 堂、超市、洗浴中心，什么鬼

**CLA**

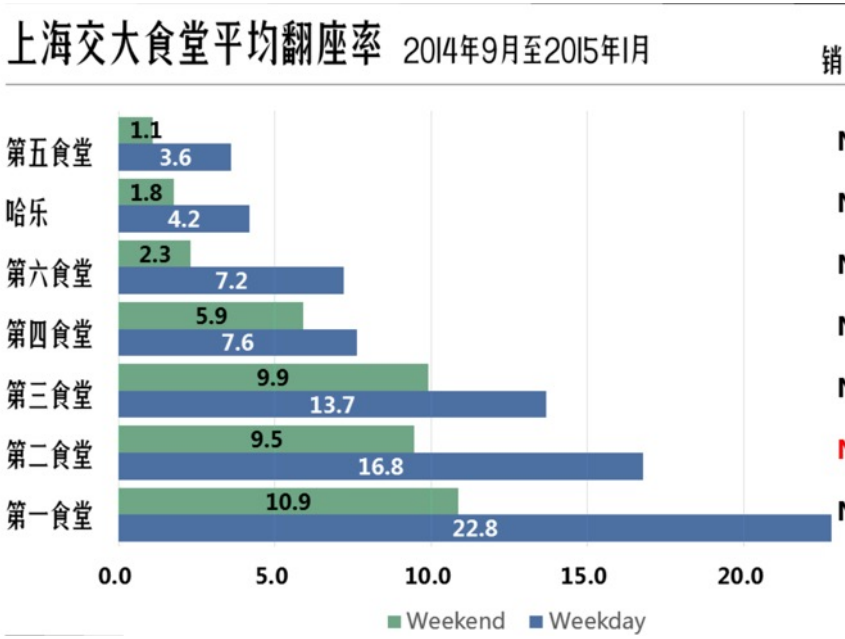
食堂、超市、澡堂次数比较均衡，人民群众中的一员



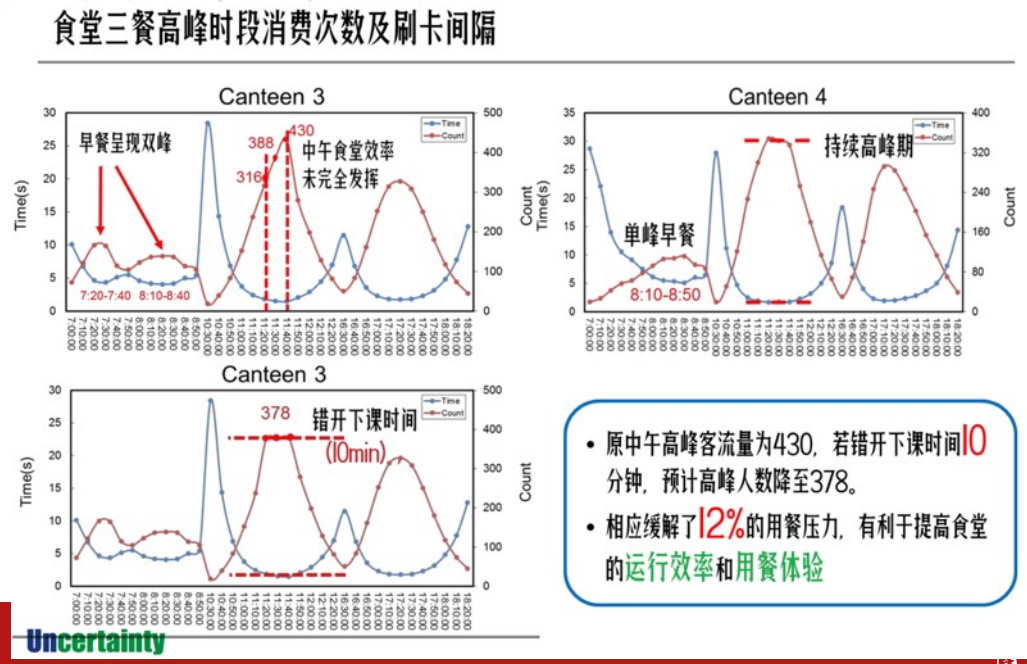



# 精彩的结果（5）




| 销售排名 | 座位数 | 商户数 | 入围美食 |
|---|---|---|---|
| No.5 | 796 | 4 | 2 |
| No.6 | 356 | 1 | 2 |
| No.7 | 232 | 3 | 3 |
| No.4 | 1541 | 21 | 5 |
| No.2 | 1120 | 4 | 4 |
| No.1 | 2034 | 10 | 9 |
| No.3 | 856 | 14 | 8 |




- 原中午高峰客流量为430，若错开下课时间10分钟，预计高峰人数降至378。
- 相应缓解了12%的用餐压力，有利于提高食堂的运行效率和用餐体验

# 上海疫情数据公开及可视化

人工智能研究院
Artificial Intelligence Institute

**平台主要显示：**

**小区级别：**

- 小区动态信息
- 14天内通报情况

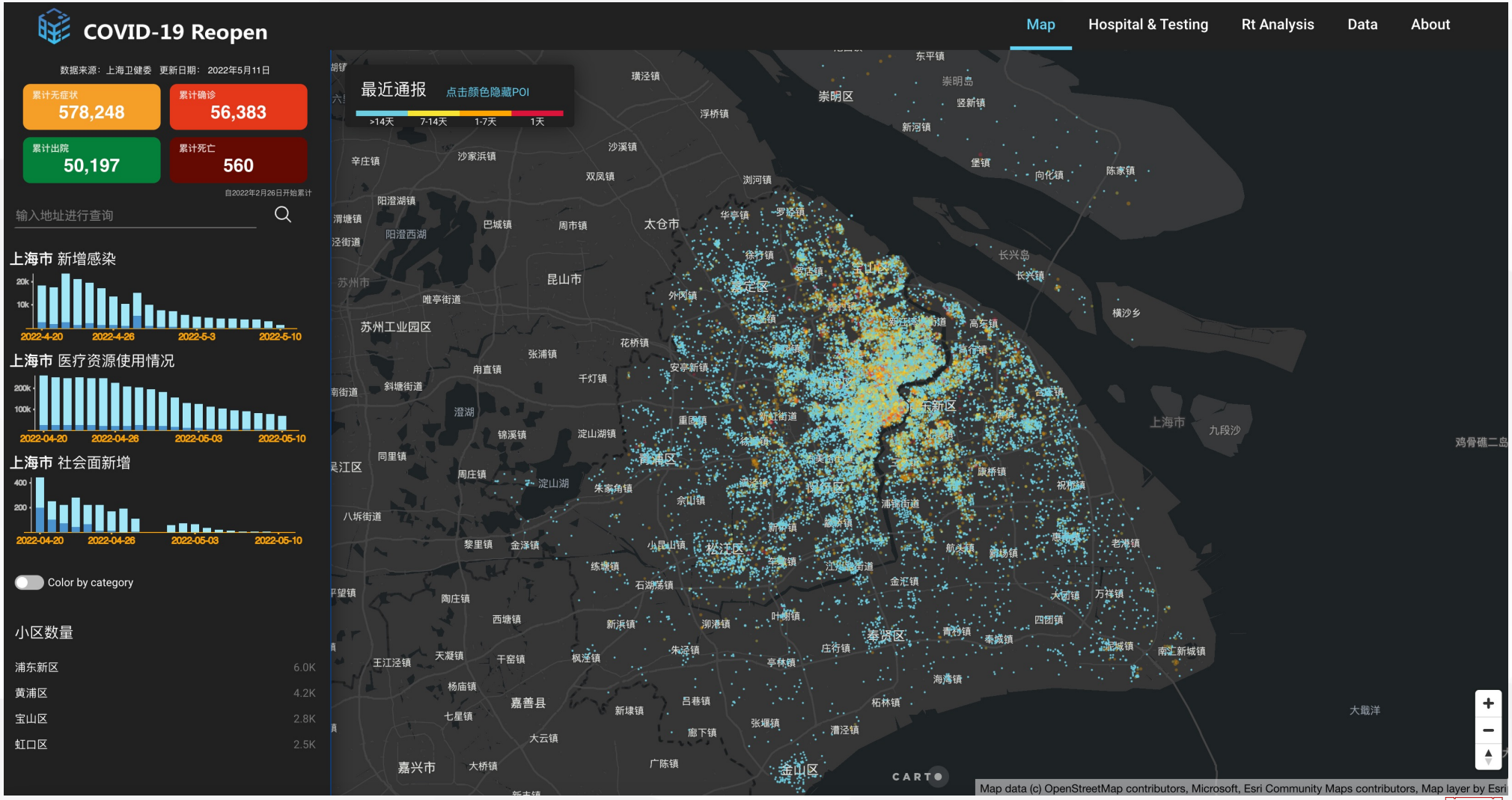
**市、区级：**

- 每日新增
- 社会面新增
- 确诊存量
- 无症状感染存量

**最新医疗/核酸资源**

**疫情走势分析**

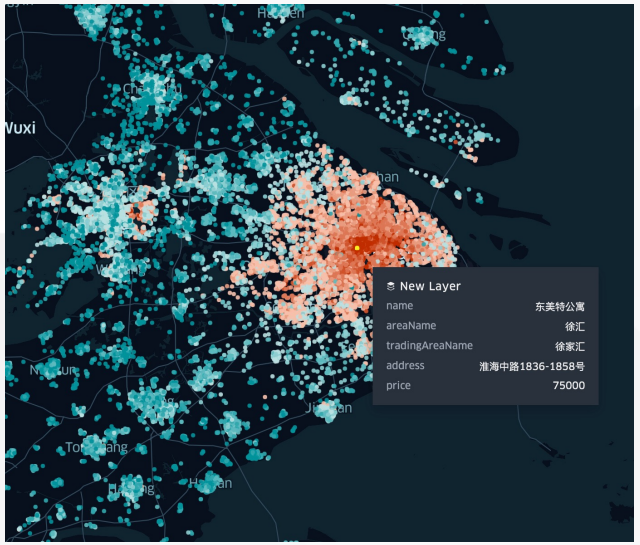
**疫情仿真模拟**

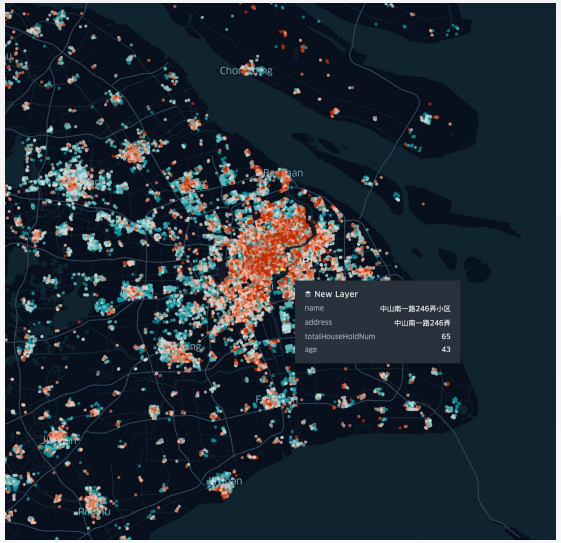


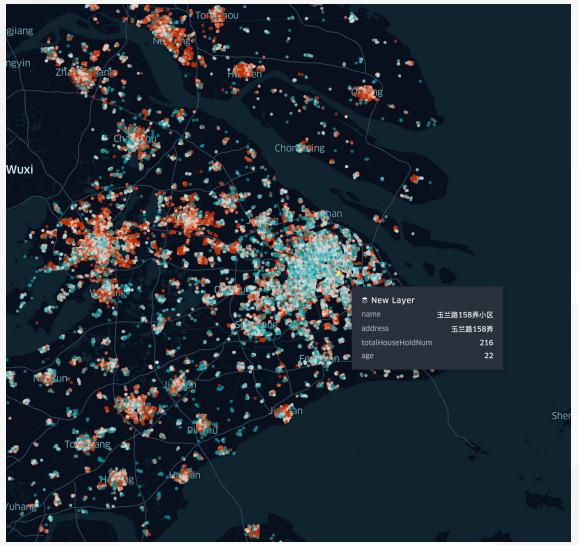

http://reopen.baiyulan.org.cn/#

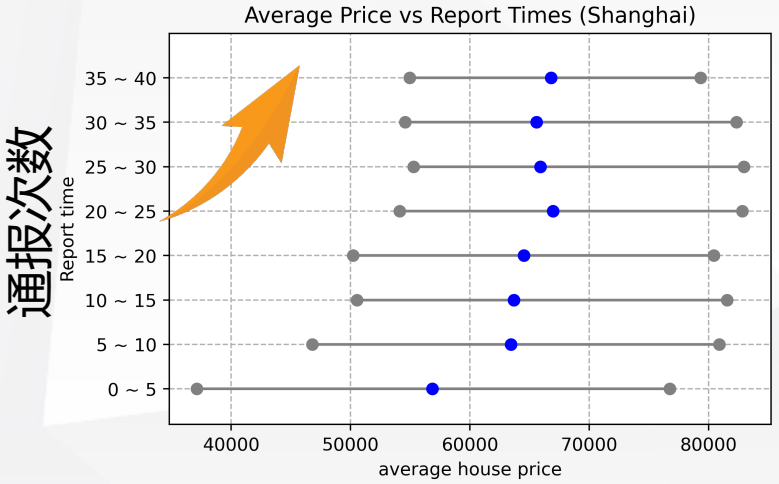
# 小区疫情风险分析（融合安居客房源信息）

人工智能研究院
Artificial Intelligence Institute

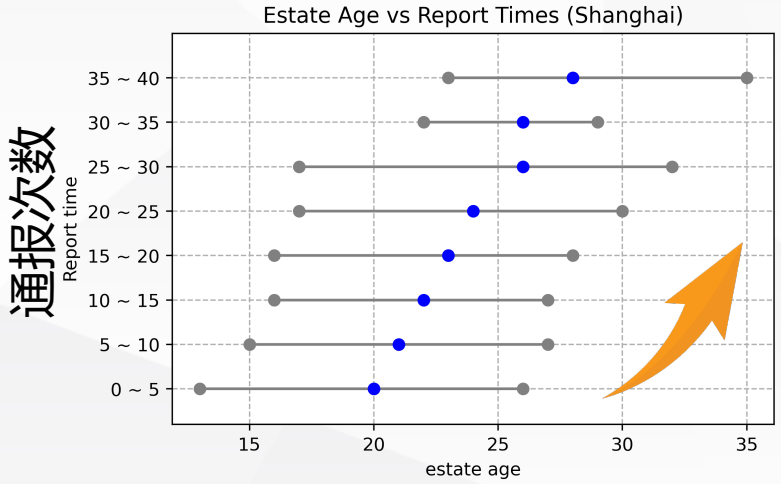


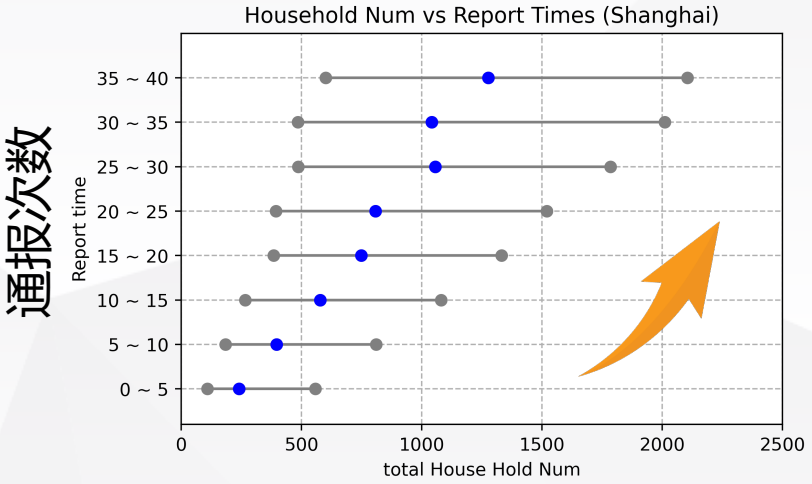




二手房均价



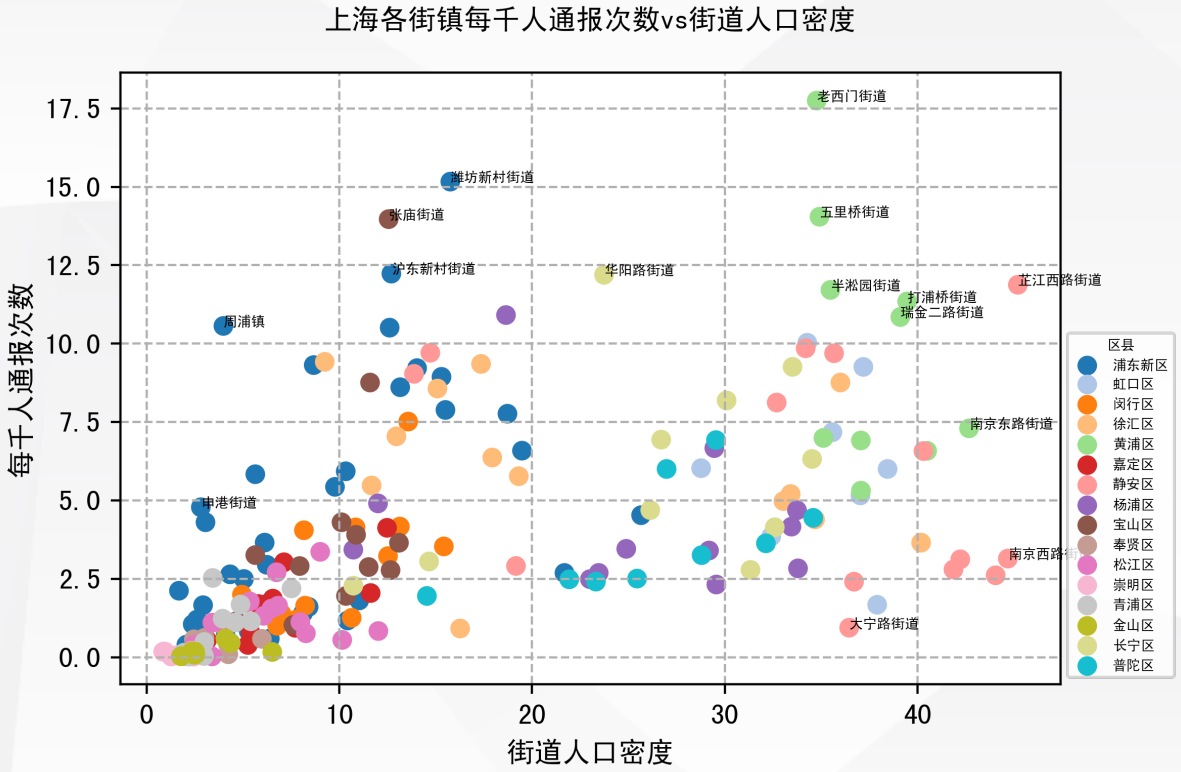

小区房龄

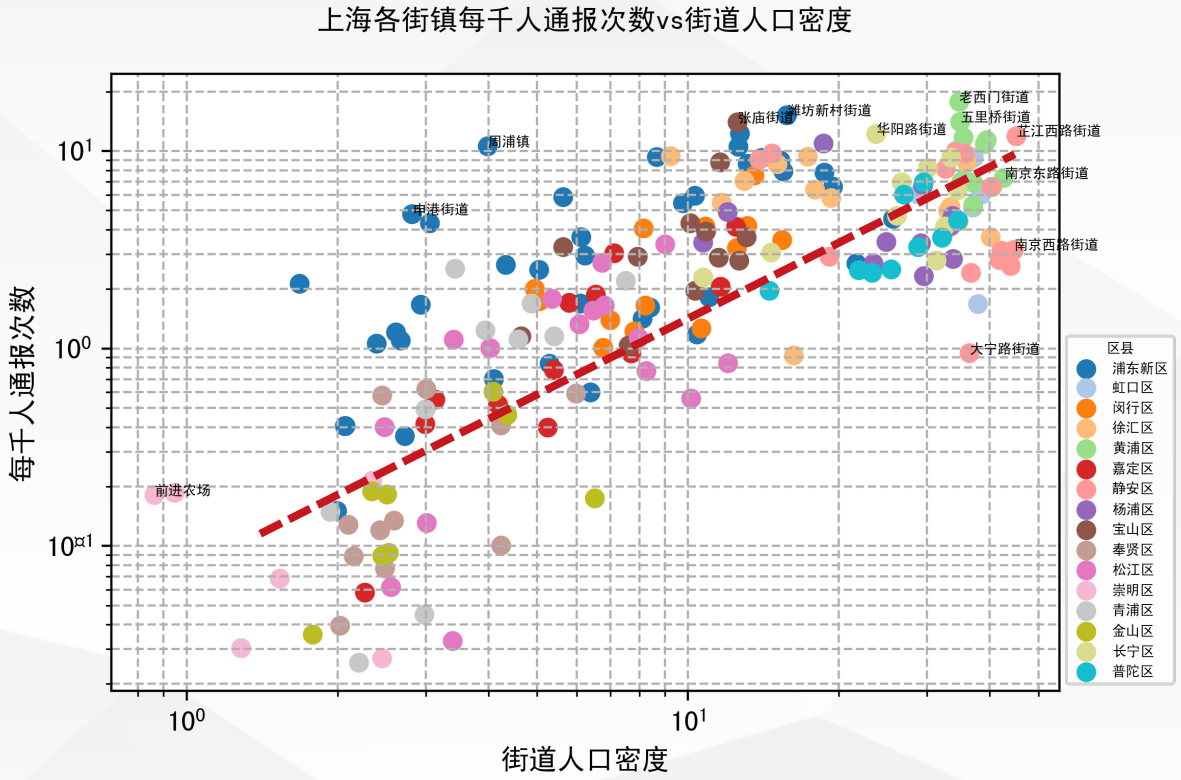



小区总户数

# 街道尺度疫情分析

人工智能研究院
Artificial Intelligence Institute




- 街道人口密度越大，疫情**风险越高，且**关系呈现出**幂律分布**
- 人口密度视角偏离分布的街镇：

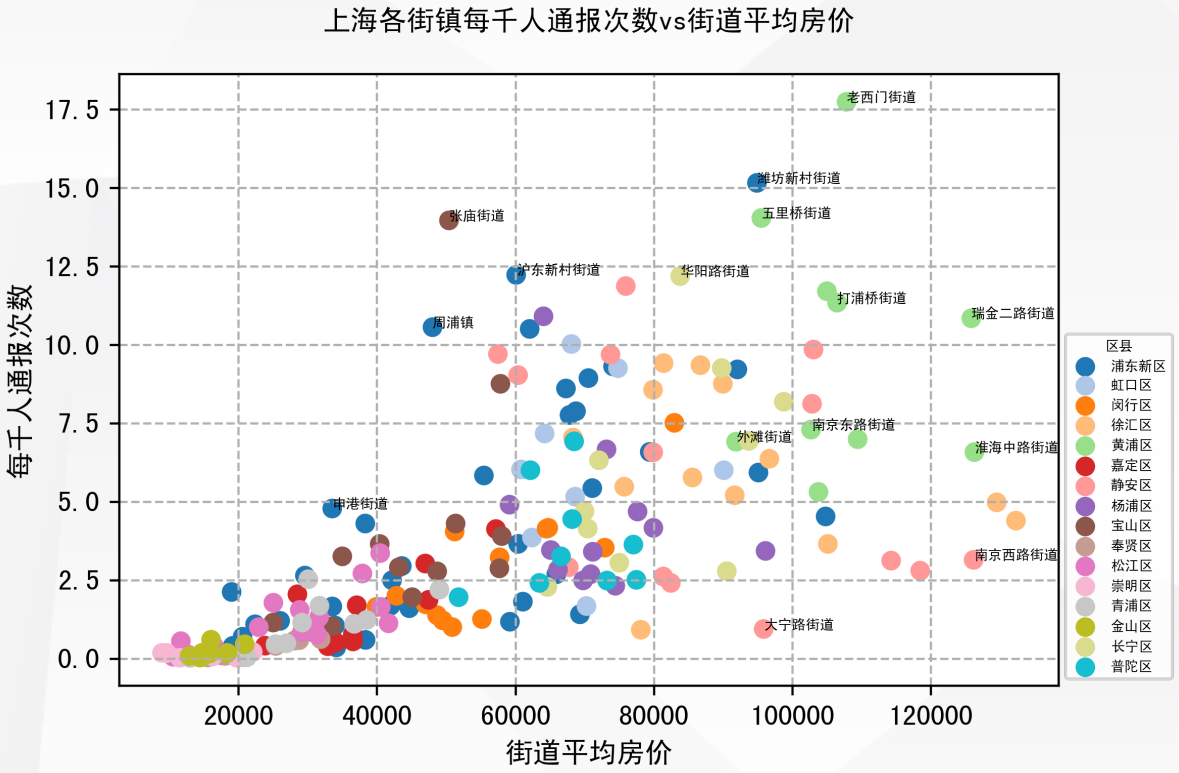
  向上偏离（风险偏大）：**周浦街道、申港街道等**

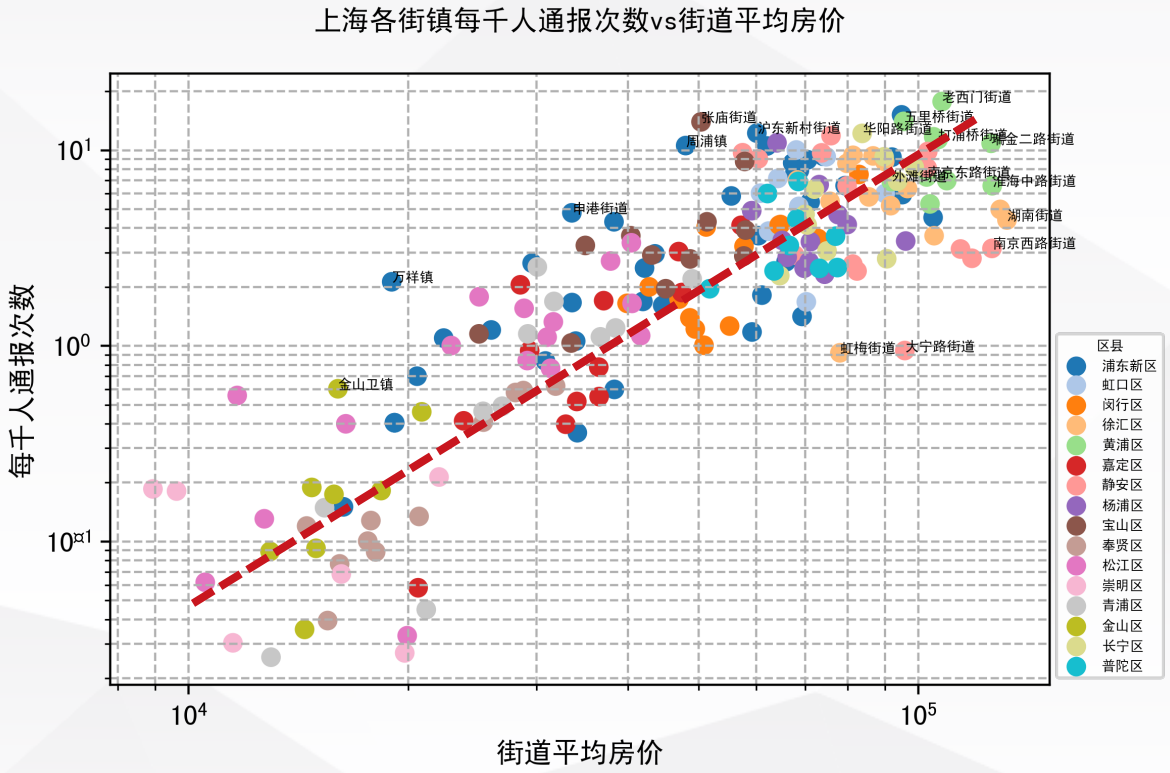
  向下偏离（风险低于预期）：**大宁路街道等**

百米网格人口数据来源：https://www.worldpop.org/

WorldPop

# 街道尺度疫情分析

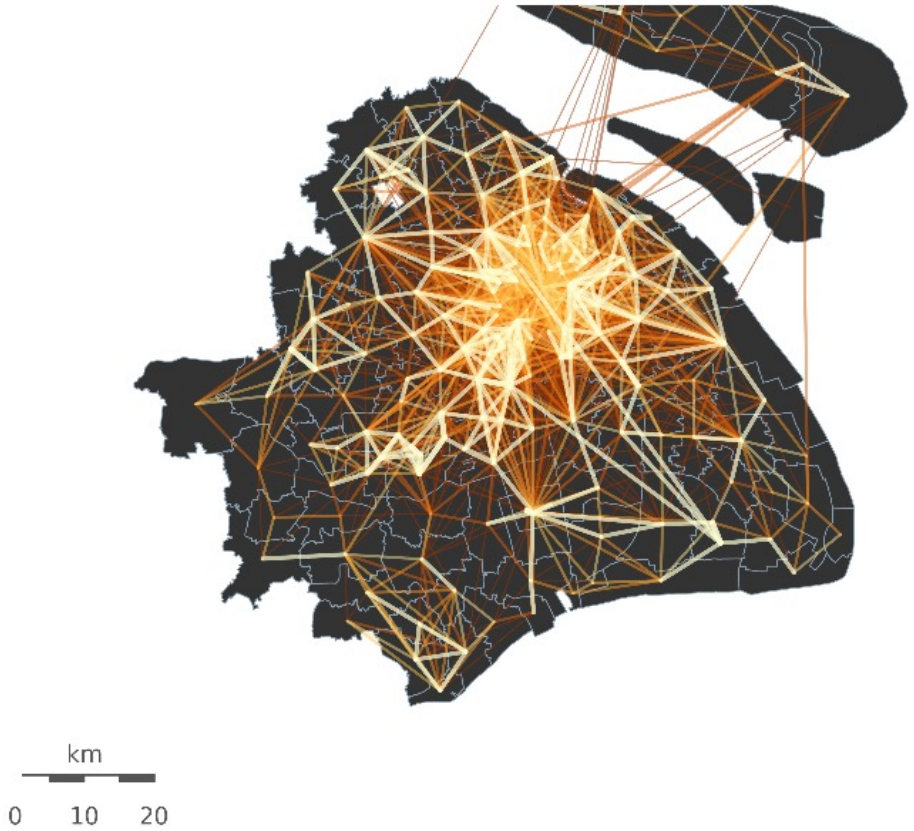



- 房价与疫情风险之间呈现出了正相关的幂律关系，房价越高，疫情**风险越大**
- 房价视角表现较好的街道：**大宁路街道、虹梅街道、南京西路街道等**

# 上海疫情数据公开及可视化

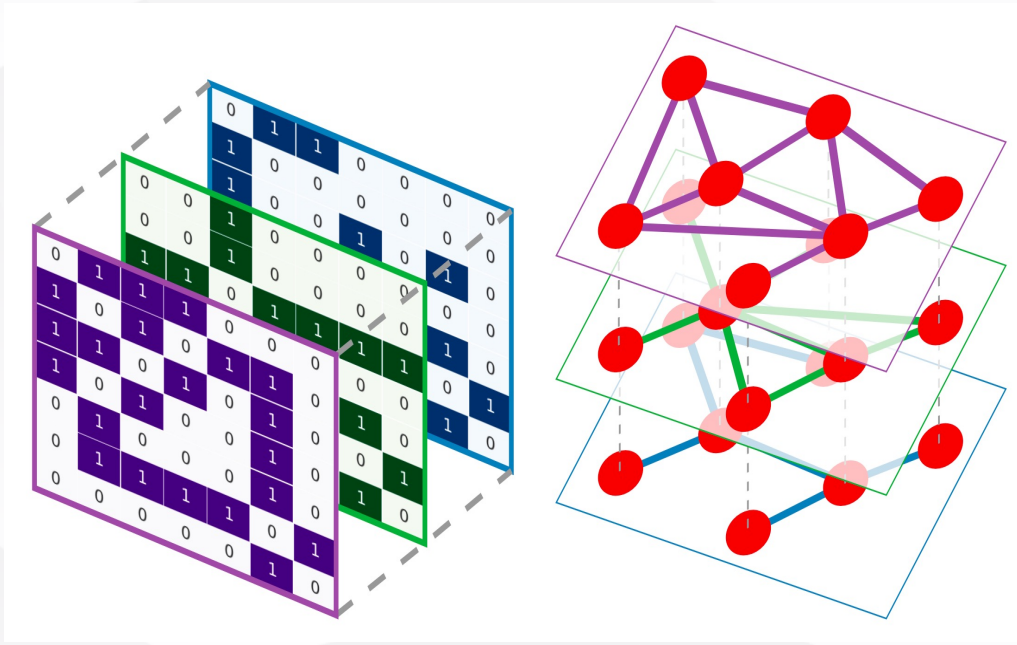
**基于历史数据的初步建模：**




上海市2800万人口移动行为

ABM模型

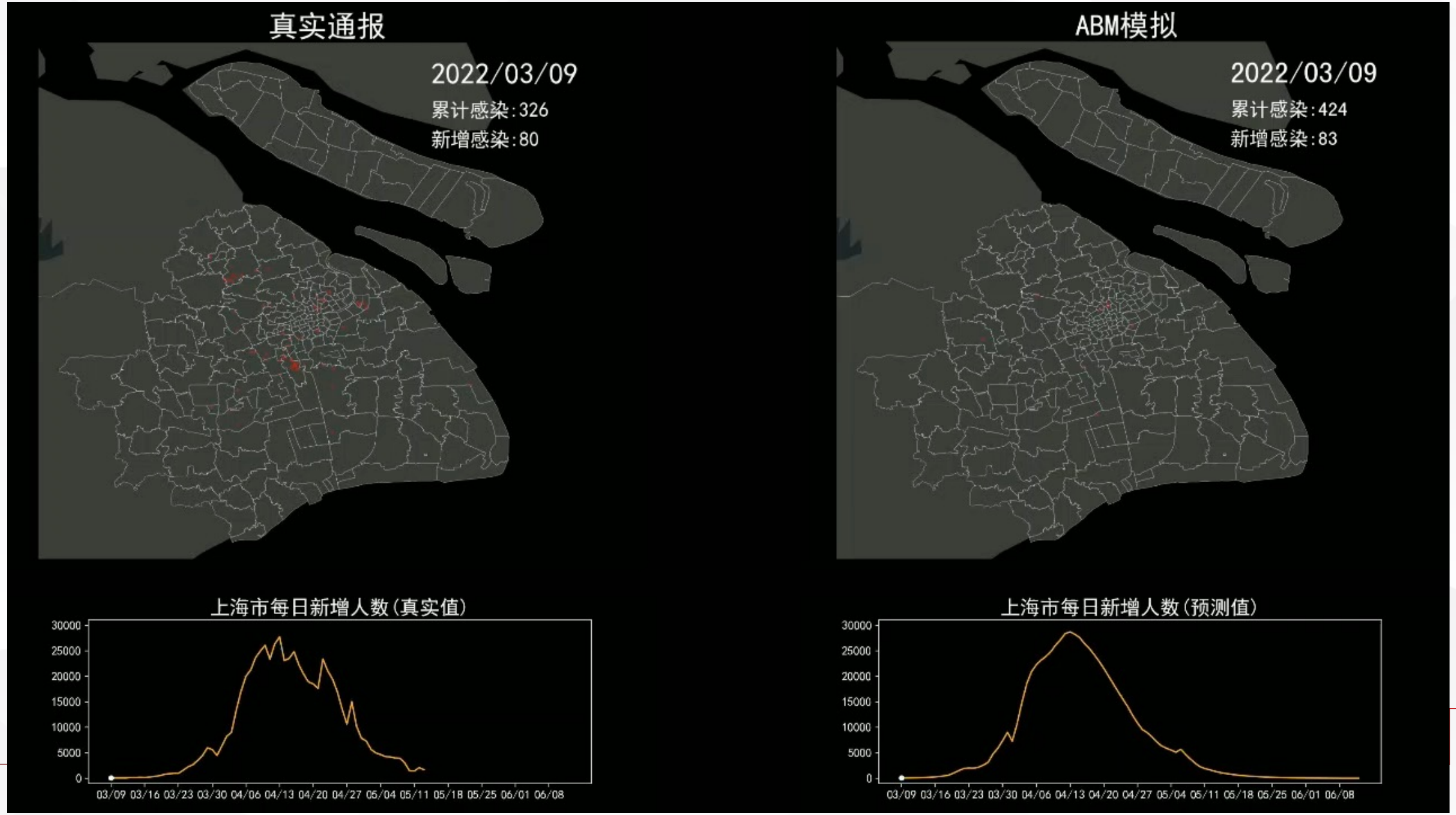
**由个体到区域的精准推演**



100m网格为基本空间单元的每小时疫情扩散模拟

# 上海市疫情发展模拟与推演

人工智能研究院
Artificial Intelligence Institute



# 疫情管控策略评估

## 以闵行区为例，社会活动完全开放，不同强度的管控措施，如何影响疫情走势？

## ABM模拟60天疫情走势



# 作业

- https://www.bilibili.com/video/BV1ce4y1w7Cu
- 观看视频后，写一篇500字左右的短文，探讨人工智能、元宇宙等对你的专业和未来职业可能的影响。
- 提交于canvas平台
- 截止时间：2月23日23:59
- 视频链接二维码：


# 上海交通大學

SHANGHAI JIAO TONG UNIVERSITY