计算社会科学导论
—— 大数据应用简介

金耀辉、许岩岩

2023年2月23日

CS1126

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

你见过的大数据，有多大？

大纲

- **当社会科学遇上大数据**
- 计算社会科学中的常用大数据
- 大数据的缺点、偏见

社会科学是研究人的学科

上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

- 人、群体、社会
- 如何获得个体的信息？
- 如何归纳出群体的特性和交互行为？
- 如何构建社会中的物理规律？
- 如何提取决定因素，影响社会发展？

社会学中的复杂性

- **人的复杂性**
- **人类行为的不确定性**
  - 不能构建典型的、完全理性和全知的 representative agent
- **即使能合理地描述个体行为(甚至是描述概率分布)，也不能轻易将其扩样到整个群体**
  - 人是有适应能力的，在不同环境中的群体有比样本更大的方差
  - 样本在群体中的分布不均匀

大数据如何帮助社会科学

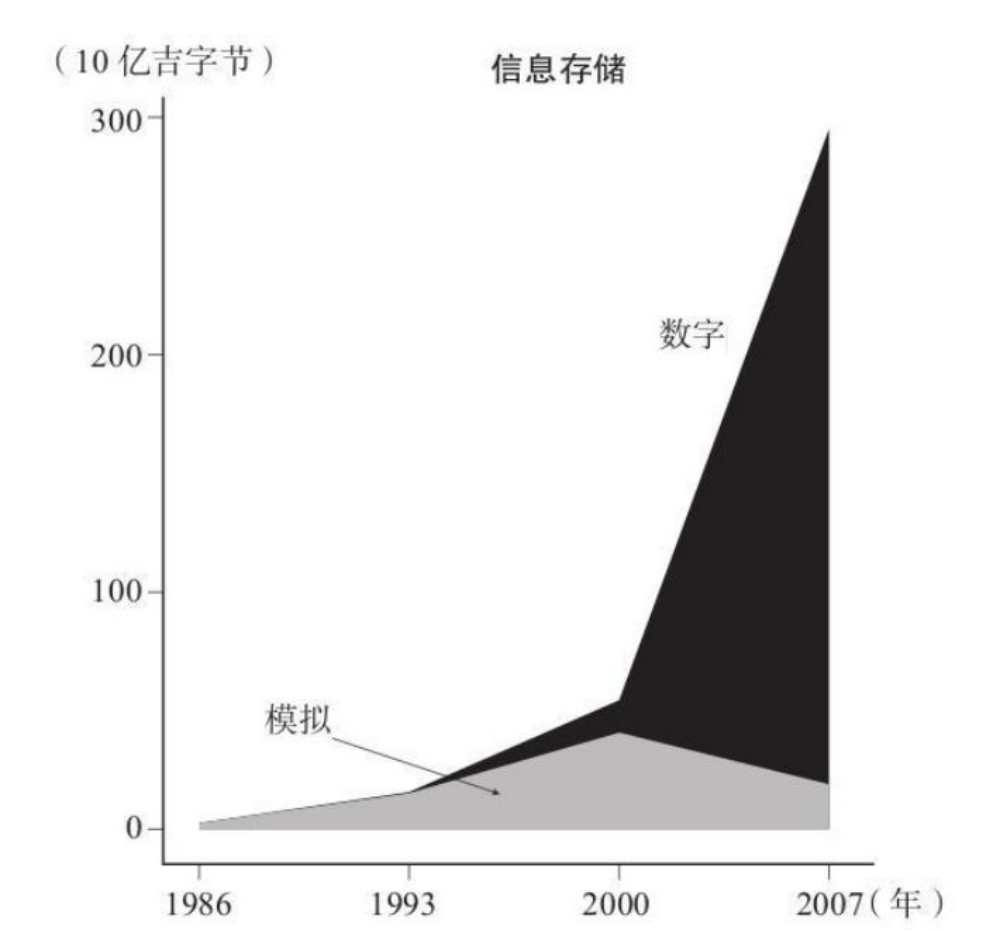
上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

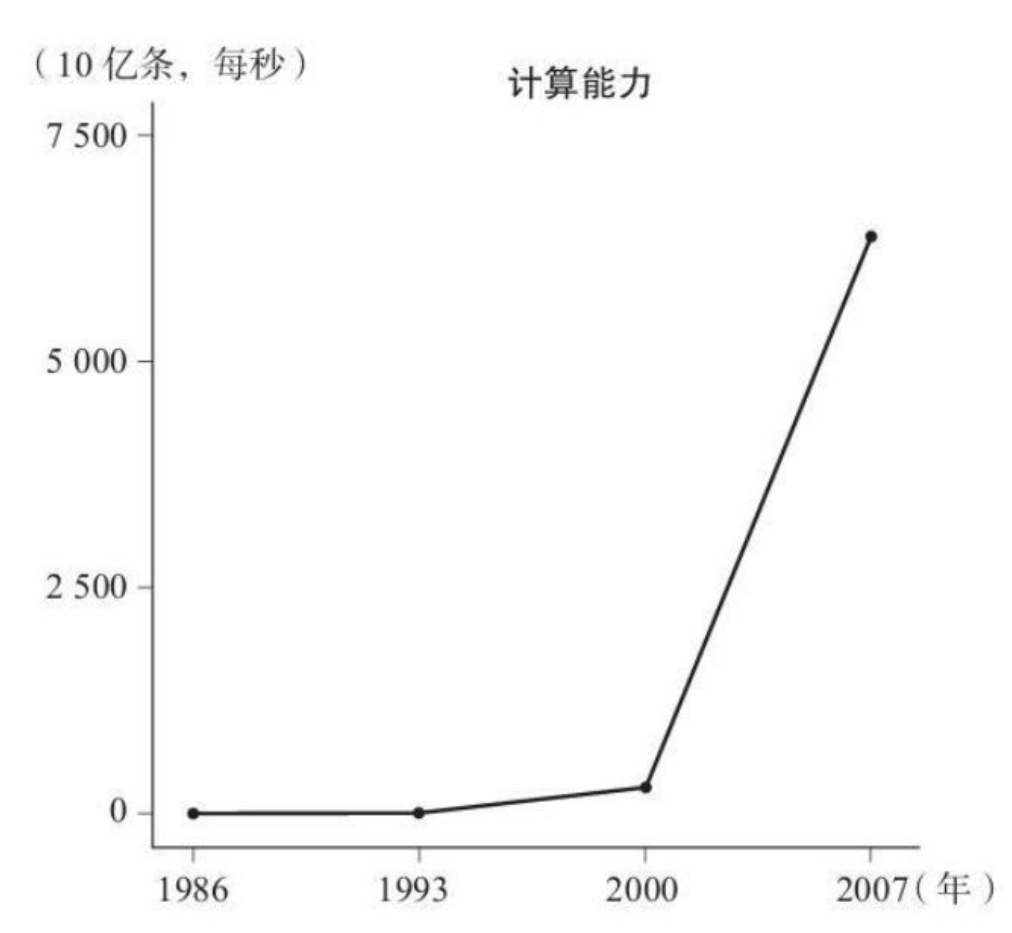
- **如何获得个体的信息？**
  - ICT-海量数据
- **如何归纳出群体的特性和交互行为？**
  - 数据科学、网络科学
- **如何构建社会中的物理规律？**
  - 统计物理

Why now：模拟到数字的惊人转变

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

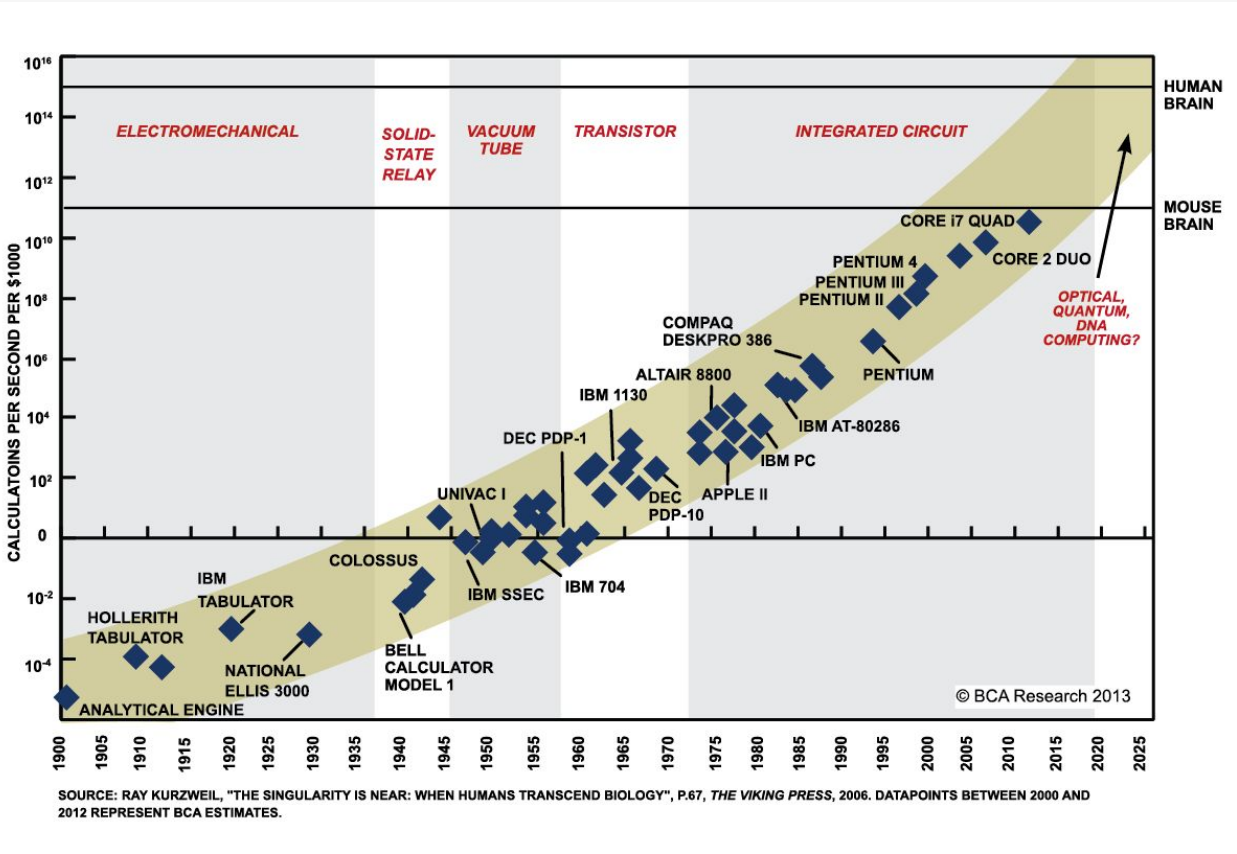
Moore's law





Moore's law

- 有人质疑，事实上算力的持续增加是一直发生的事情，为什么大数据偏偏是这个时间节点出现？



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

- **互联网**是一个被全面监测的环境，非常适合研究人员开展实验。例如在线商城可以搜集到精确的数百万顾客的购买行为数据。
- **实体店**也已经搜集了非常详细的购买行为数据，同时它们也正在开发相关基础设施，以便追踪顾客的购买行为，并将实验研究结果用于日常商业活动中。
- **物联网**意味着现实世界中的行为会越来越多地被数字传感器捕获。

事实上是多种技术和现象的组合叠加

研究角度

用户角度

Massively distributed computing（分布式）

- MapReduce
- Spark
- cloud computing

Big-memory machines（内存）

- Terabytes of RAM

Advances in machine learning（机器学习）

- Deep learning
- transformers
- large language models

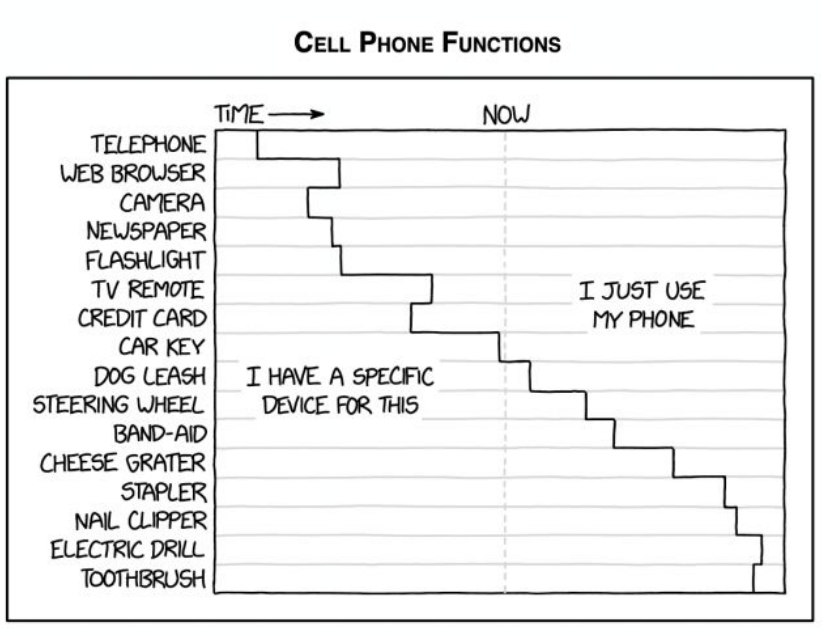
Fast streaming algorithms（快速算法）

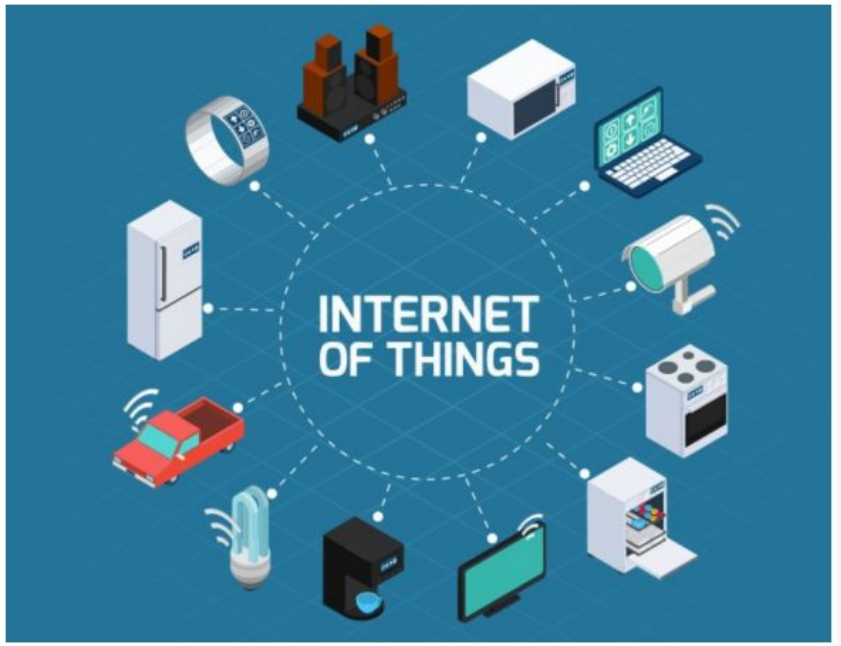
- Streaming aggregation
- stochastic gradient descent

Human computation（人类协作）

- Crowdsourcing
- Mechanical Turk

设备一体化





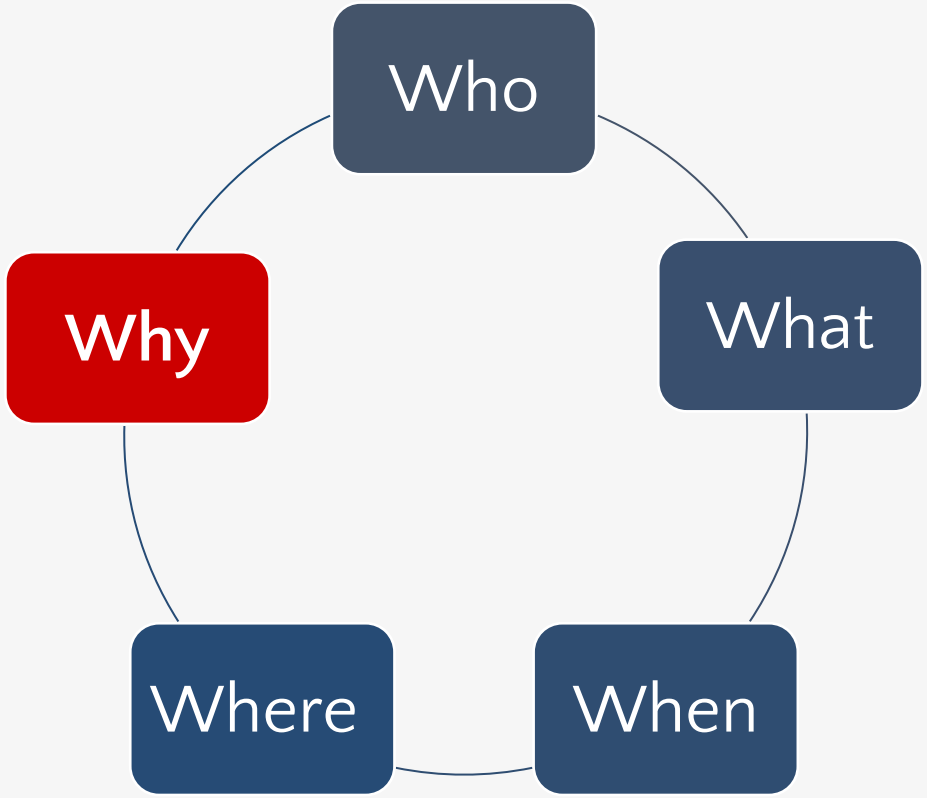
物联网

大数据在你眼里是：

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

常见定义	倡导者认为	批评者认为
Volume 大量	Veracity 真实	Vague 模糊
Variety 多样	Value 价值	Vacuous 空洞
Velocity 高速		

我们应该更关心：



研究流程

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

大量数据

- ICT数据采集
- 运用计算机科学优化存储、计算和传输

计算社会科学

- 交叉学科
- 统计物理、图论、复杂系统等工具建立模型

可以进行

- 分析数据
- 仿真模型
- 控制变量实验
- 验证理论预测
- 分析策略的影响

得到

- 新的行为模式
- 更好地理解复杂地社会系统

目的

- 为决策的制定和实行提供建议
- 可持续发展的社会

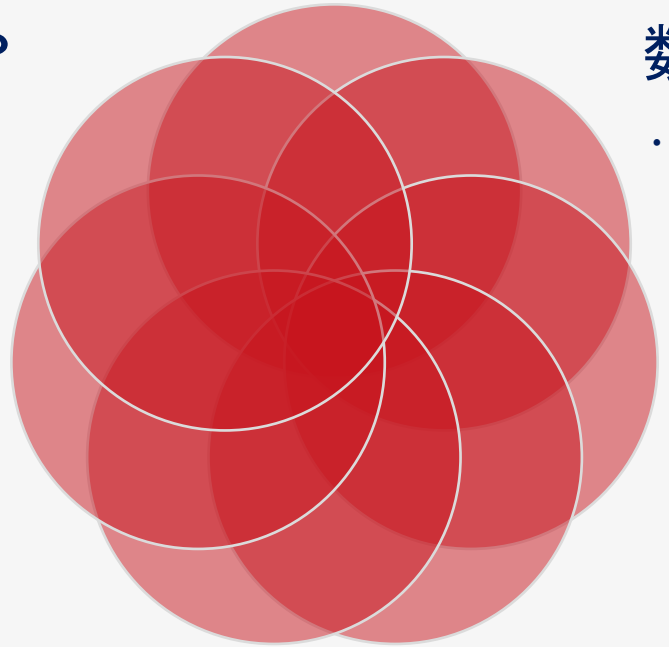
围绕大数据，你可能需要关心

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

对数据的分类总结：

· 有没有新兴的行为模式和群体行为？

数据数量很多，质量如何？



数据能代表多大范围的人群？

· 看似易于收集的在线数据、问卷，能代表多少人？

使用统计、物理、复杂系统等工具实现学科交叉

标准化的数据收集和挖掘流程

· 实验的可复现性

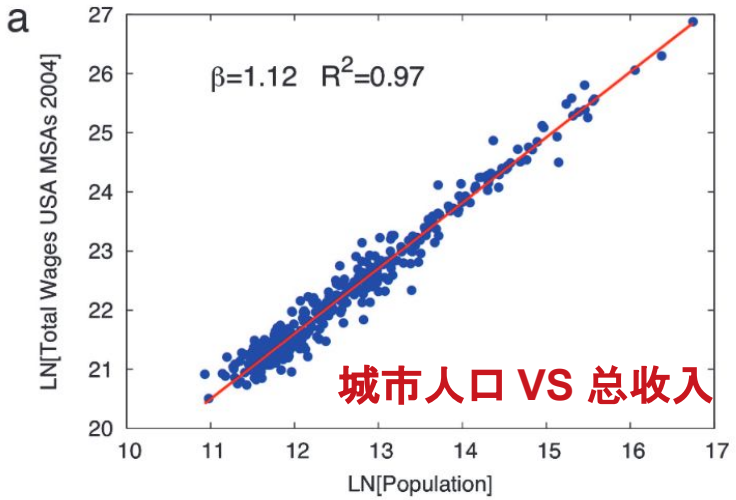
用大范围数据验证新的理论模型

数据使用协议、伦理、隐私

大纲

- 当社会科学遇上大数据
- **计算社会科学中的常用大数据**
- 大数据的缺点、偏见

人口数量研究GDP、城市发展



$$y = x^{\beta} e^{\alpha}$$

Table 1. Scaling exponents for urban indicators vs. city size

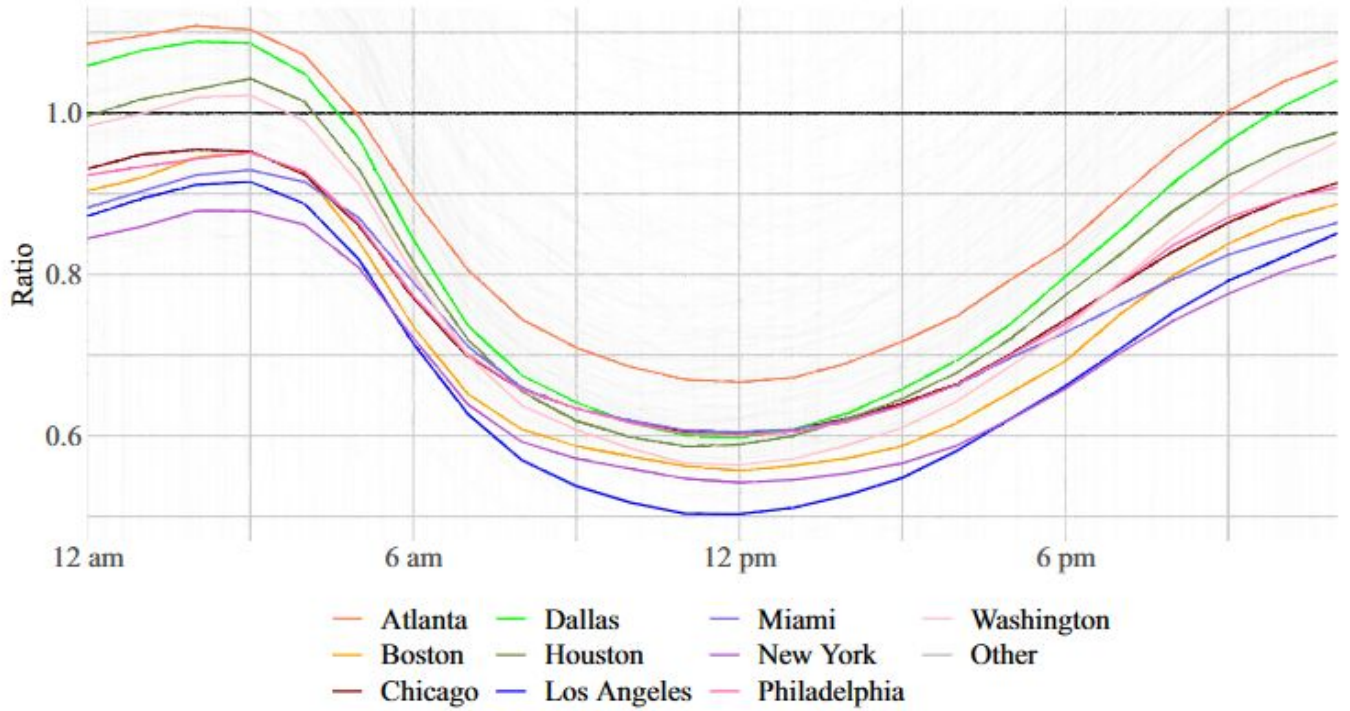
Y	β	95% CI	Adj-R^2	Observations	Country–year
New patents	1.27	[1.25,1.29]	0.72	331	U.S. 2001
Inventors	1.25	[1.22,1.27]	0.76	331	U.S. 2001
Private R&D employment	1.34	[1.29,1.39]	0.92	266	U.S. 2002
"Supercreative" employment	1.15	[1.11,1.18]	0.89	287	U.S. 2003
R&D establishments	1.19	[1.14,1.22]	0.77	287	U.S. 1997
R&D employment	1.26	[1.18,1.43]	0.93	295	China 2002
Total wages	1.12	[1.09,1.13]	0.96	361	U.S. 2002
Total bank deposits	1.08	[1.03,1.11]	0.91	267	U.S. 1996
GDP	1.15	[1.06,1.23]	0.96	295	China 2002
GDP	1.26	[1.09,1.46]	0.64	196	EU 1999–2003
GDP	1.13	[1.03,1.23]	0.94	37	Germany 2003
Total electrical consumption	1.07	[1.03,1.11]	0.88	392	Germany 2002
New AIDS cases	1.23	[1.18,1.29]	0.76	93	U.S. 2002–2003
Serious crimes	1.16	[1.11, 1.18]	0.89	287	U.S. 2003
Total housing	1.00	[0.99,1.01]	0.99	316	U.S. 1990
Total employment	1.01	[0.99,1.02]	0.98	331	U.S. 2001
Household electrical consumption	1.00	[0.94,1.06]	0.88	377	Germany 2002
Household electrical consumption	1.05	[0.89,1.22]	0.91	295	China 2002
Household water consumption	1.01	[0.89,1.11]	0.96	295	China 2002
Gasoline stations	0.77	[0.74,0.81]	0.93	318	U.S. 2001
Gasoline sales	0.79	[0.73,0.80]	0.94	318	U.S. 2001
Length of electrical cables	0.87	[0.82,0.92]	0.75	380	Germany 2002
Road surface	0.83	[0.74,0.92]	0.87	29	Germany 2002

Data sources are shown in *SI Text*. CI, confidence interval; Adj-R^2, adjusted R^2; GDP, gross domestic product.

Bettencourt L M A, Lobo J, Helbing D, et al. Growth, innovation, scaling, and the pace of life in cities[J]. Proceedings of the national academy of sciences, 2007, 104(17): 7301-7306.

GPS轨迹数据

- 使用来自智能手机的GPS数据，测量了美国城市中不同种族之间的**经历隔离程度**。
- 手机数据测量的经历隔离比传统的居住隔离更能捕捉到个体真实面对的多样性和机会。
- 个体经历的隔离程度比标准的居住隔离指标要低得多；在不同城市之间，经历隔离和居住隔离有很高的相关性。



- **作者也提到了这篇文章中数据本身存在的问题**
  - 手机数据只能体现出设备出现在同一个空间区域中，并不代表用户真实的互动。这使得文章研究的种族隔离更偏向于地理隔离而非社会学隔离，尽管地理隔离这一概念同样有意义；
  - 没有手机数据的人种信息-用居住地白人占比给用户打上白人或非白人的标签，再用于计算种族隔离；
  - 手机用户并不总能代表总人口

Athey S, Ferguson B, Gentzkow M, et al. Estimating experienced racial segregation in US cities using large-scale GPS data[J]. Proceedings of the National Academy of Sciences, 2021, 118(46): e2026160118.

LBS、航班

- 除非能减少50%上的社区传播，否则对中国大陆进行持续90%的旅行限制只会适度影响疫情轨迹。旅行限制只能推迟疫情的传播（起点），并不能降低传播速度。

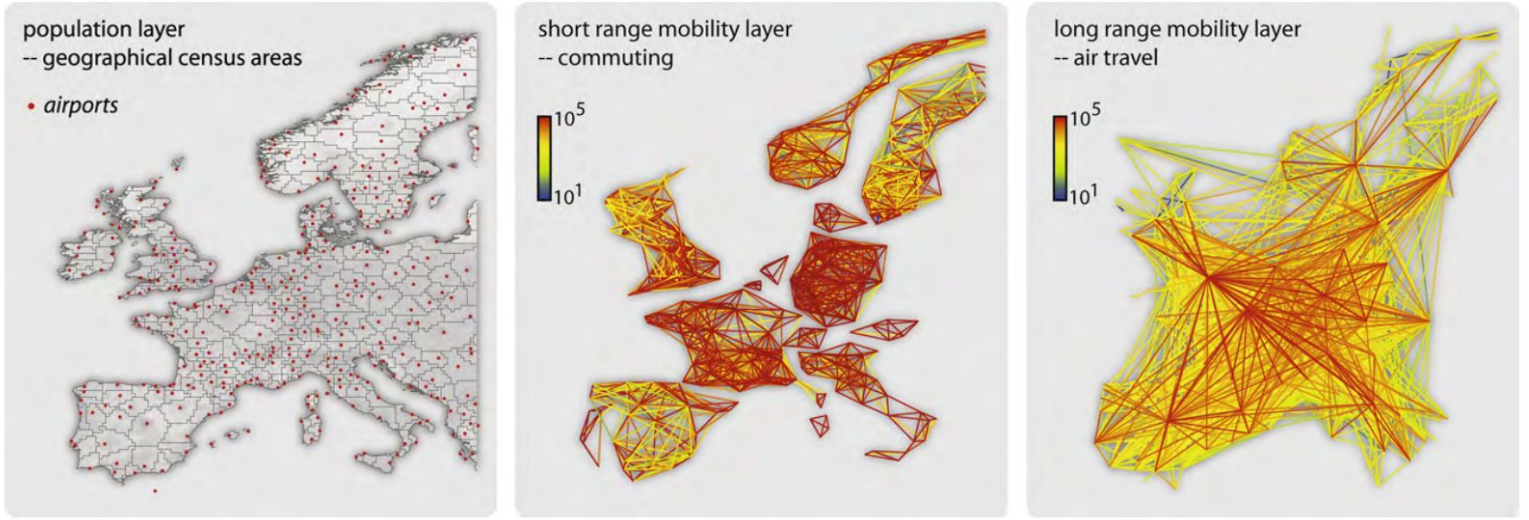


Fig. 1. GLEaM, GLobal Epidemic and Mobility model. The world surface is represented in a grid-like partition where each cell – corresponding to a population value – is assigned to the closest airport. Geographical census areas emerge that constitute the subpopulations of the metapopulation model. The demographic layer is coupled with two mobility layers, the short range commuting layer and the long range air travel layer.

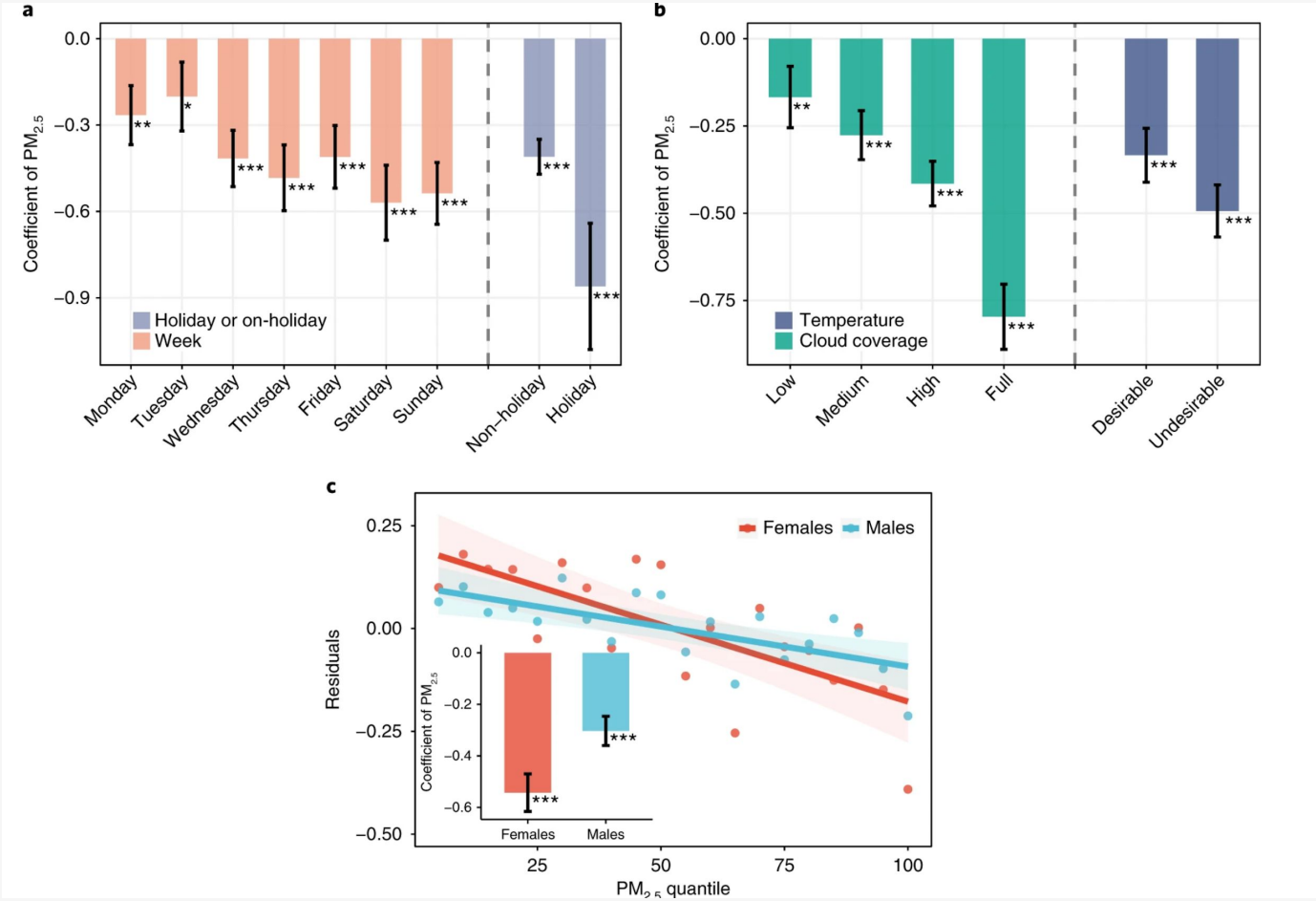
- GLEAM疫情传播模型将世界被划分为以主要交通枢纽（通常是机场）为中心的子种群。这些群体通过每天在它们之间旅行的个人流量连接起来。该模型包括大约200个不同国家和地区的3200多个群体。航空运输数据包括来自官方航空指南（OAG）和国际航空运输协会（IATA）数据库（2019年更新）的每日出发地-目的地交通流量，而地面流动流量则来自对从五大洲30个国家的统计局收集的数据的分析和建模。中国大陆的人口移动变化来自于百度的定位服务（LBS）。

Chinazzi M, Davis J T, Ajelli M, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak[J]. Science, 2020, 368(6489): 395-400.

Wifi数据了解生活节律

社交媒体信息和大气污染的关系

- 通过对中国最大的微博平台**新浪微博**上发布的2.1亿条带有地理标签的微博推文应用NLP领域相关算法来衡量144个城市居民的每日**幸福感**。
- 研究144个中国城市，由于对污染的讨论不一定能反映出个人潜在情绪状态的变化，使用没有污染相关术语的推文来**构建我们的城市/日幸福指数**。
- **证实：大气污染的加剧与居民幸福感之间的相关性，且女性对大气污染更加敏感。**



Zheng, S., Wang, J., Sun, C. *et al.* Air pollution lowers Chinese urbanites' expressed happiness on social media. *Nat Hum Behav* **3**, 237–243 (2019).

通过管理出行行为缓解交通拥堵

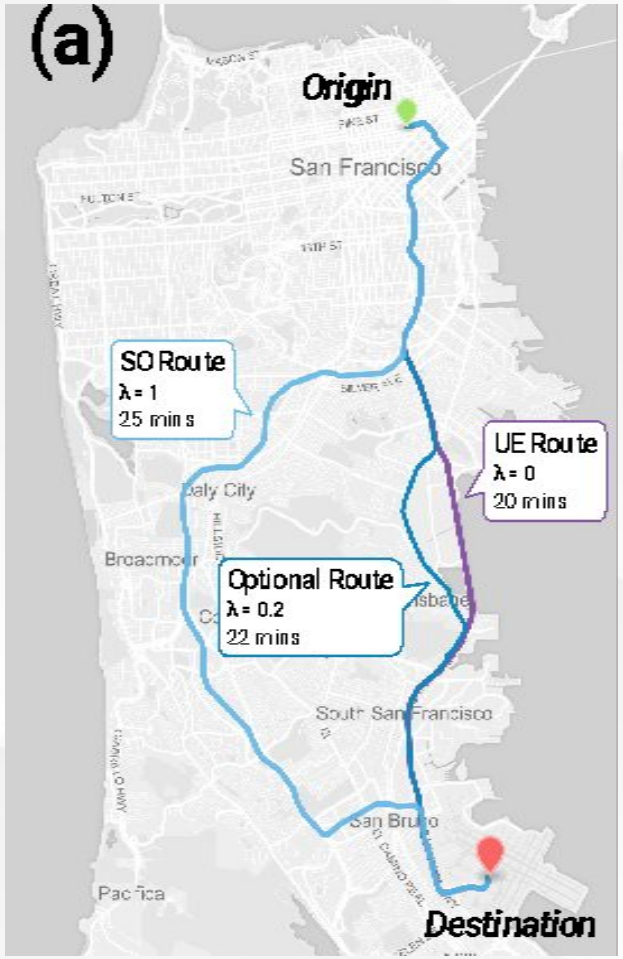
出行需求的获取：社会调查？GPS？LBS？手机信令？路口摄像头？

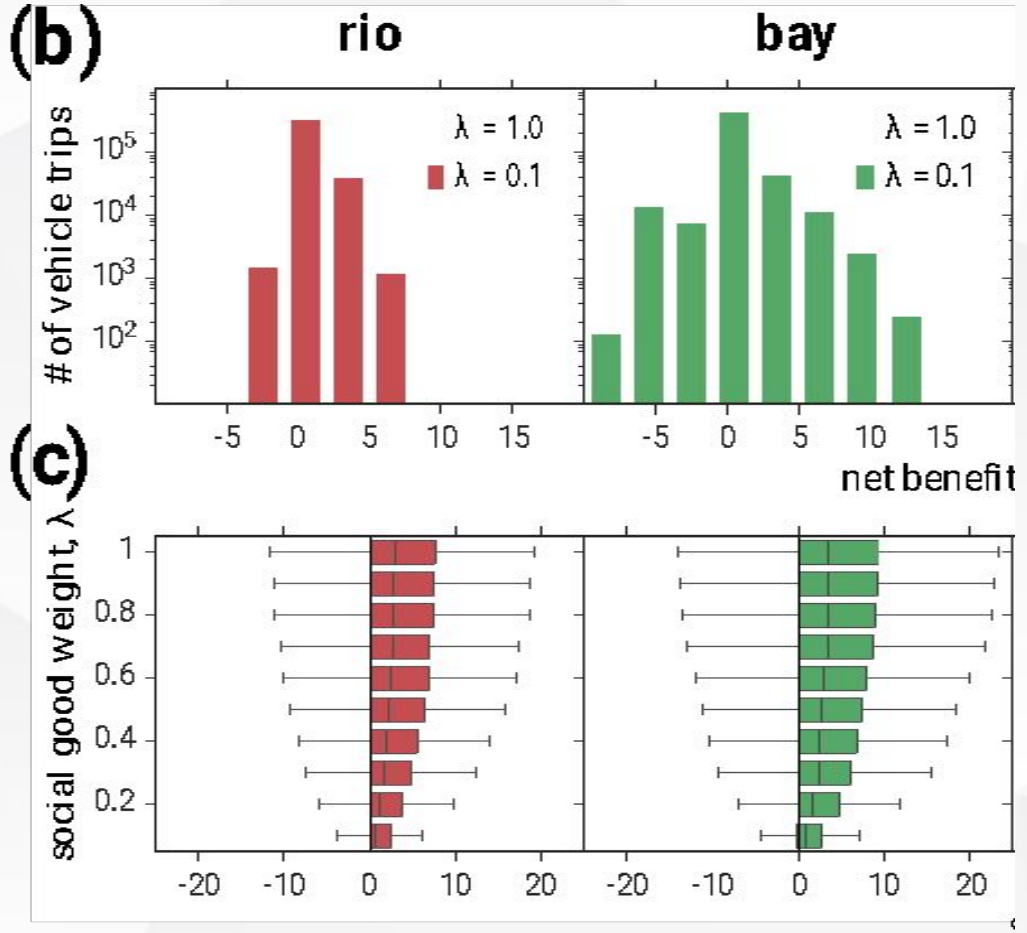
各有什么优缺点？

a) 某用户从旧金山市区出发，前往国际机场。其“自私”路径（UE）的行程时间为20min，而“无私”路径（SO）行程时间为25min。

b) 里约及湾区在不同的“无私”程度（λ不同取值）下的用户出行时间节约百分比分布

c) 里约及湾区在不同的“无私”程度（λ不同取值）下的所有居民总体出行时间节约百分比。





Ref: Çolak, Serdar, Antonio Lima, and Marta C. González. "Understanding congested travel in urban areas." *Nature communications* 7.1 (2016): 1-8.

出行调查数据

以美国麻省为例：

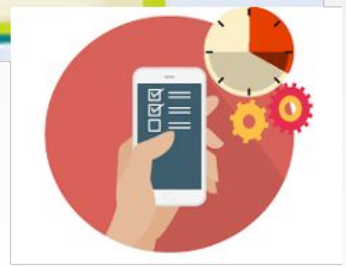
$200 per usable Survey

1 sample day,

2.5×10^4 households out of 2.6×10^6

58% response rate.

(3.7 calls and 17 minutes per survey)



2011-2012

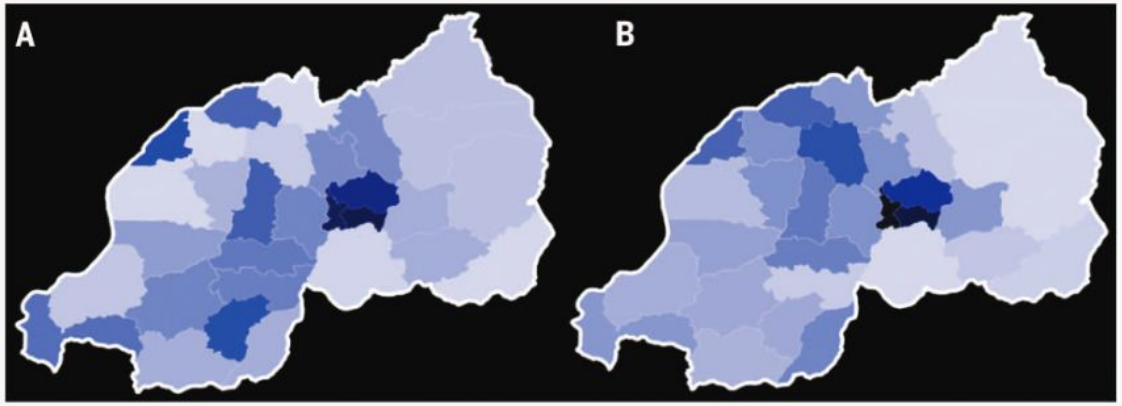
移动信令数据

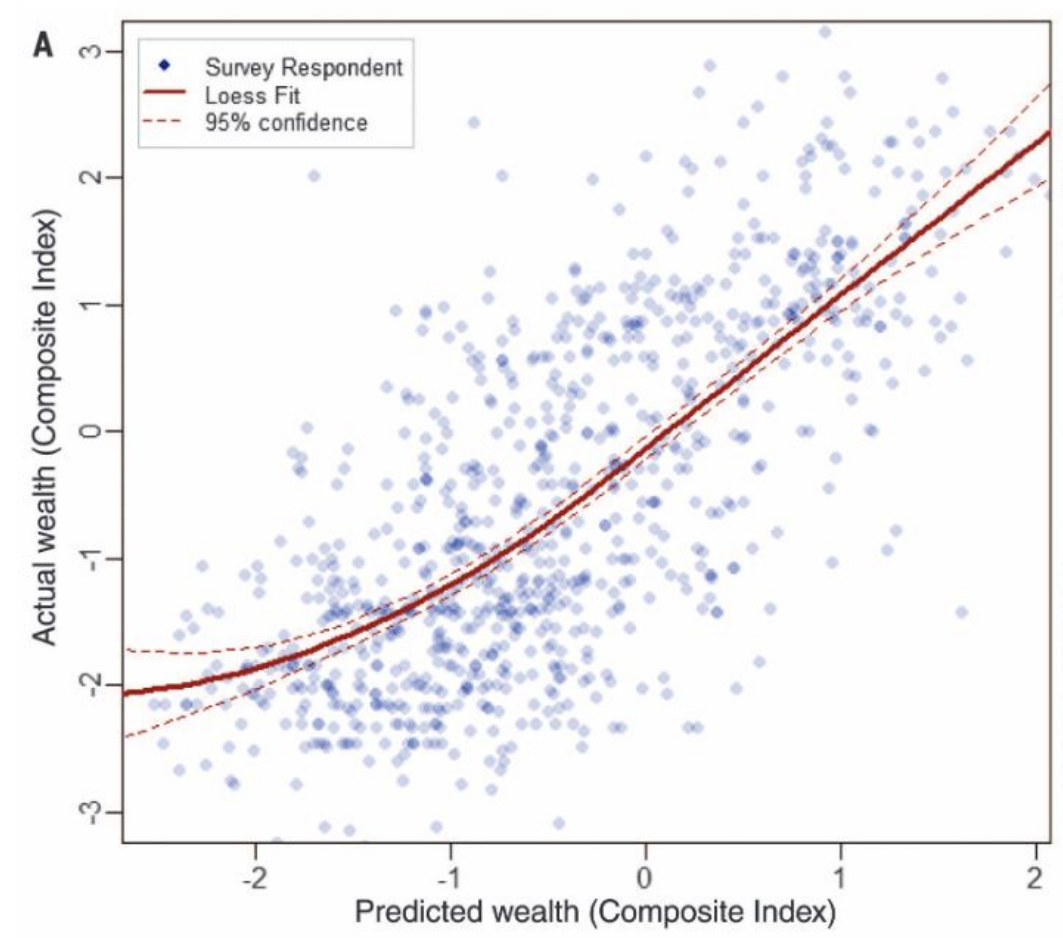
MIT Technology Review 10 BREAKTHROUGH TECHNOLOGIES 2013

Big Data
From
Cheap Phones

Review-手机metadata

- 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





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

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

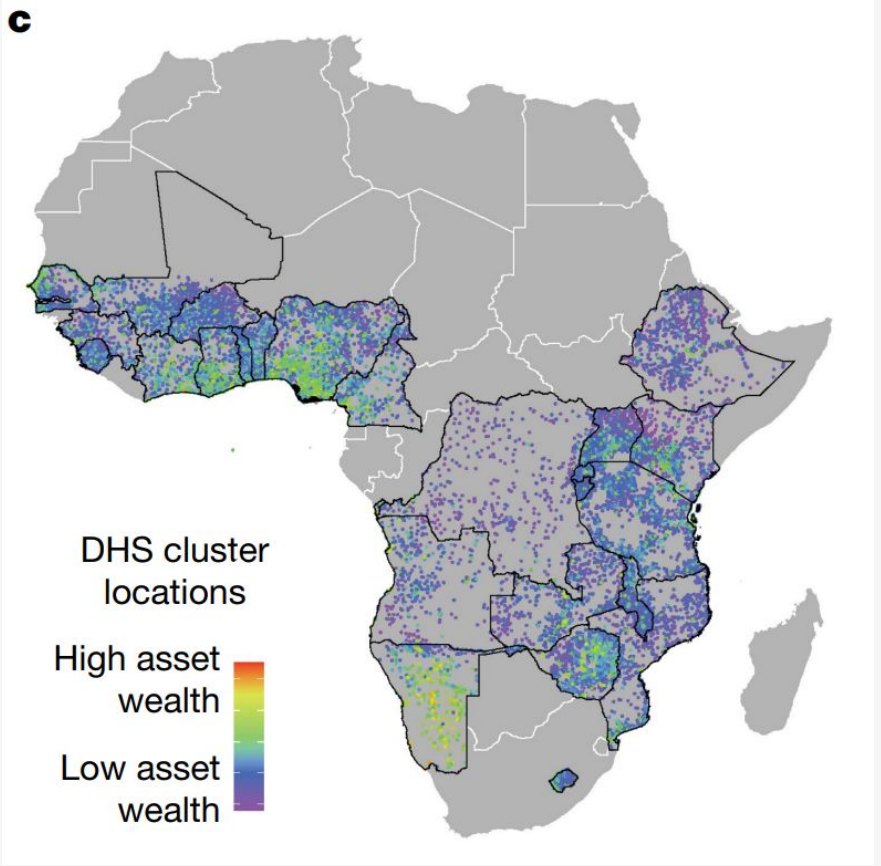
Review-遥感图像

- **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

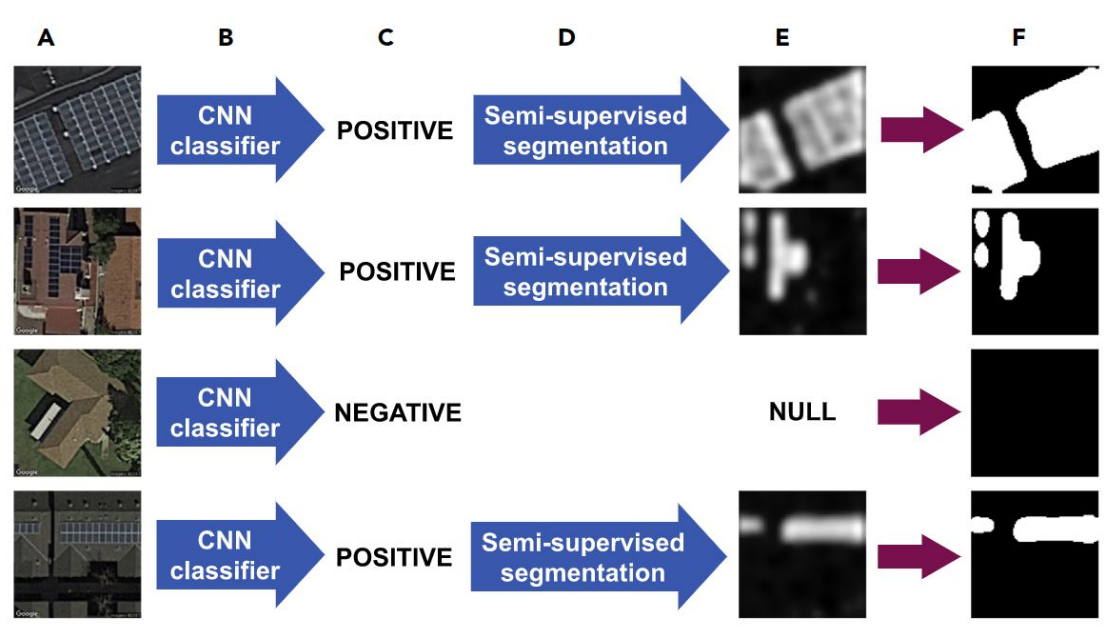
  只使用日间卫星图像，以尽量少通过灯光获得电网的信息

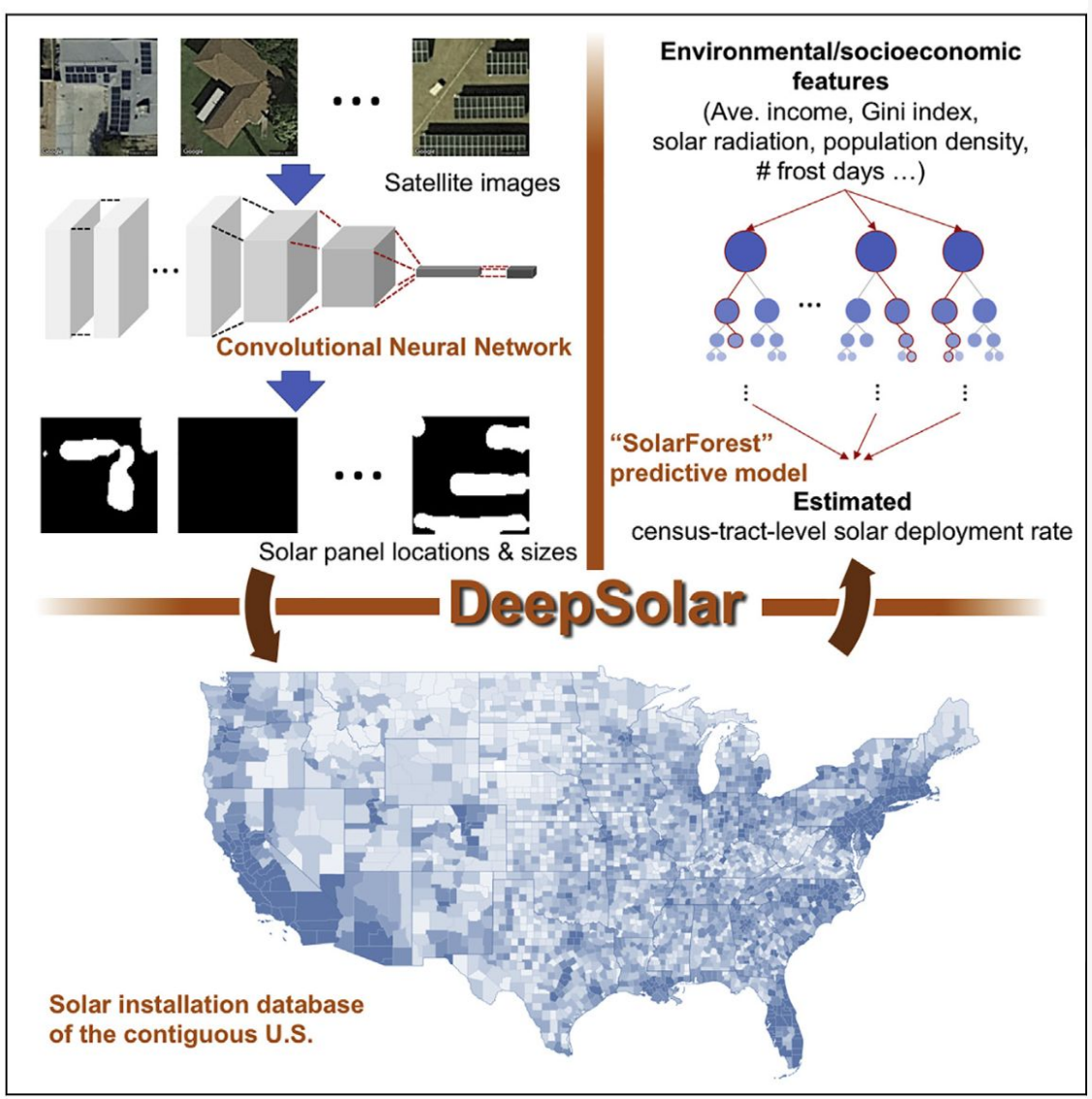


Ratledge, N., Cadamuro, G., de la Cuesta, B. *et al.* Using machine learning to assess the livelihood impact of electricity access. *Nature* **611**, 491–495 (2022).

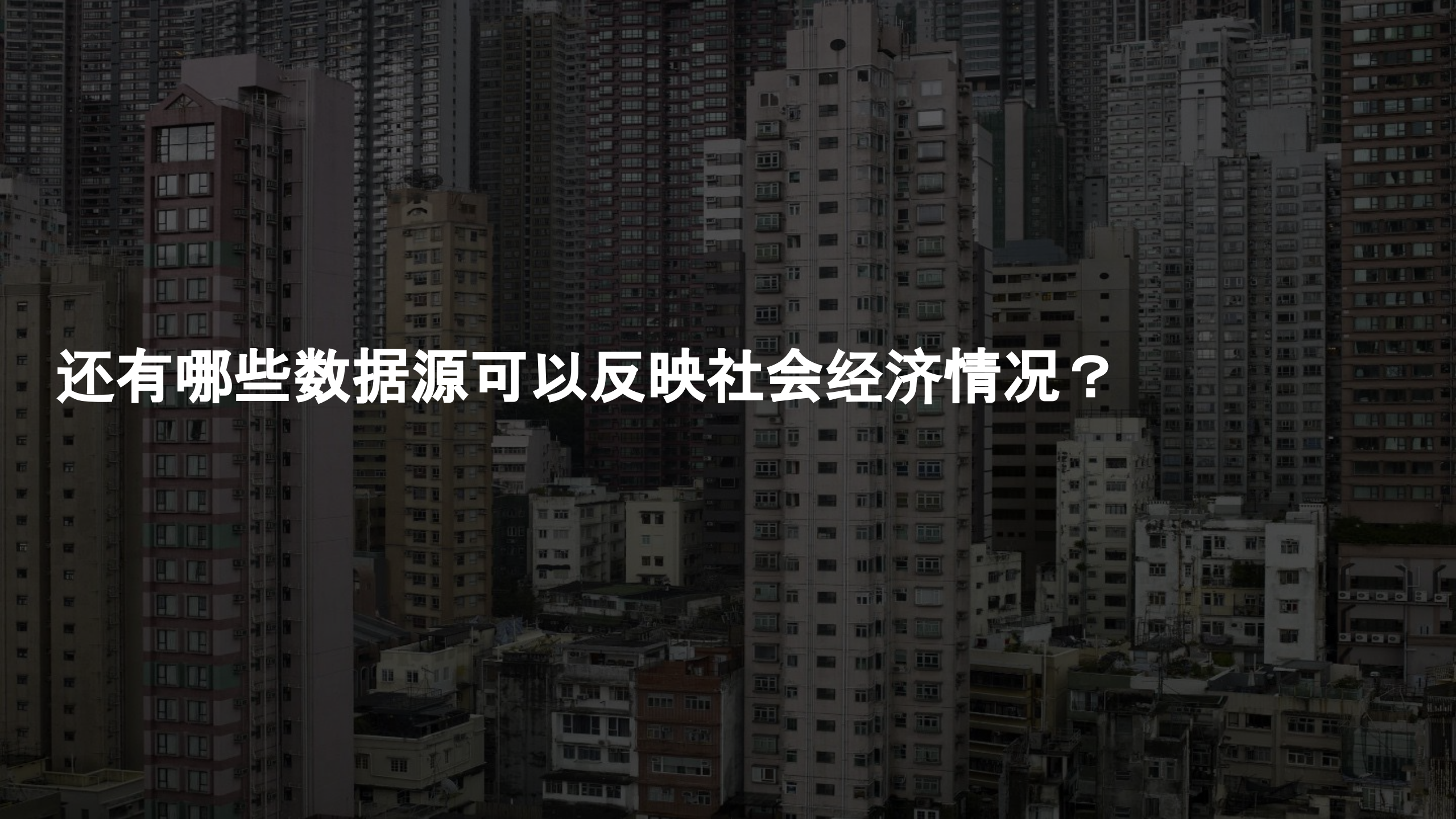
DeepSolar 卫星图像推测太阳能普及程度

- 用计算机视觉方法从**卫星图像**中构建太阳能充电板的分布位置数据库，并和环境与社会经济指标建立相关关系。

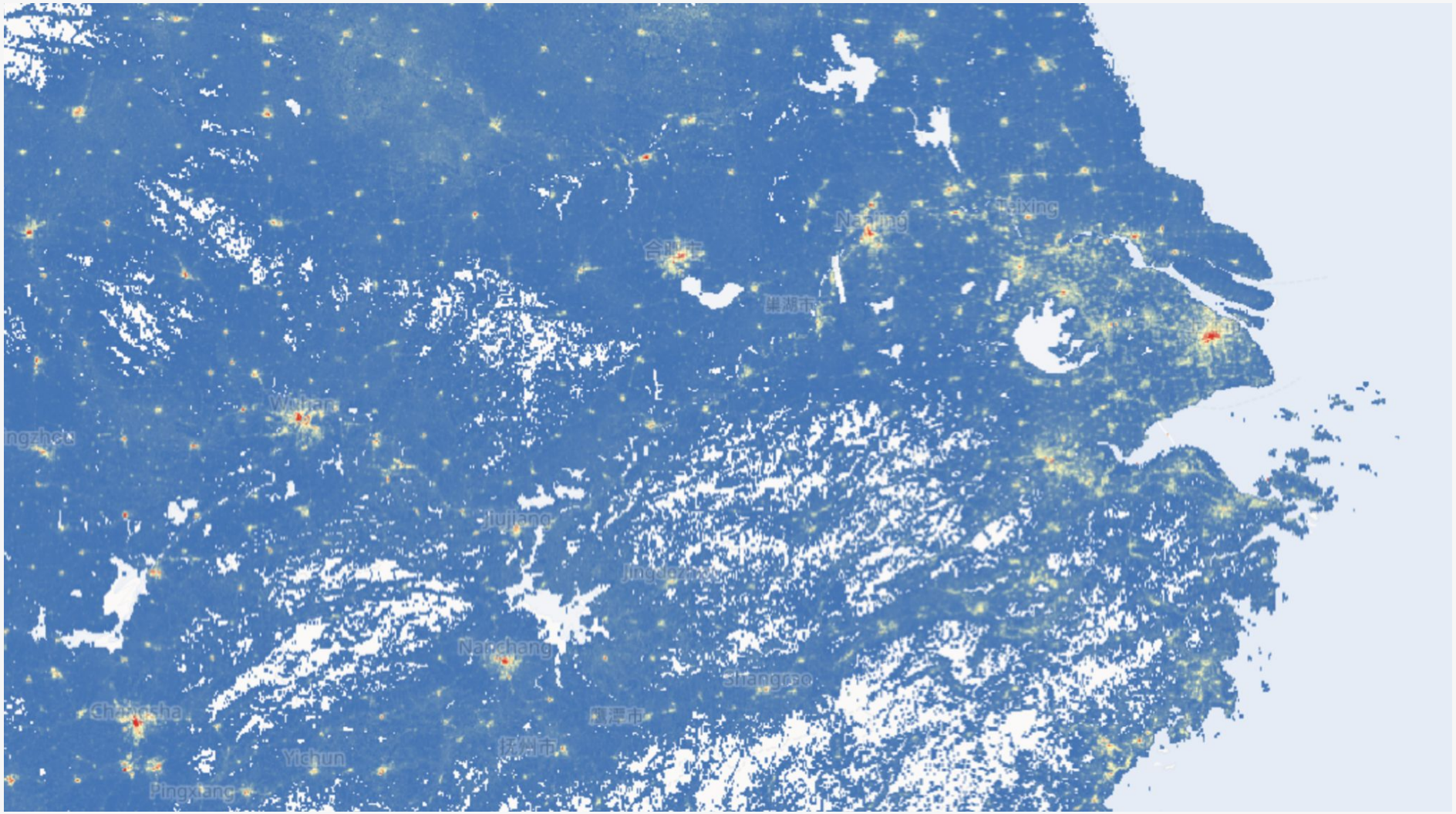




Yu J, Wang Z, Majumdar A, et al. DeepSolar: A machine learning framework to efficiently construct a solar deployment database in the United States[J]. Joule, 2018, 2(12): 2605-2617.

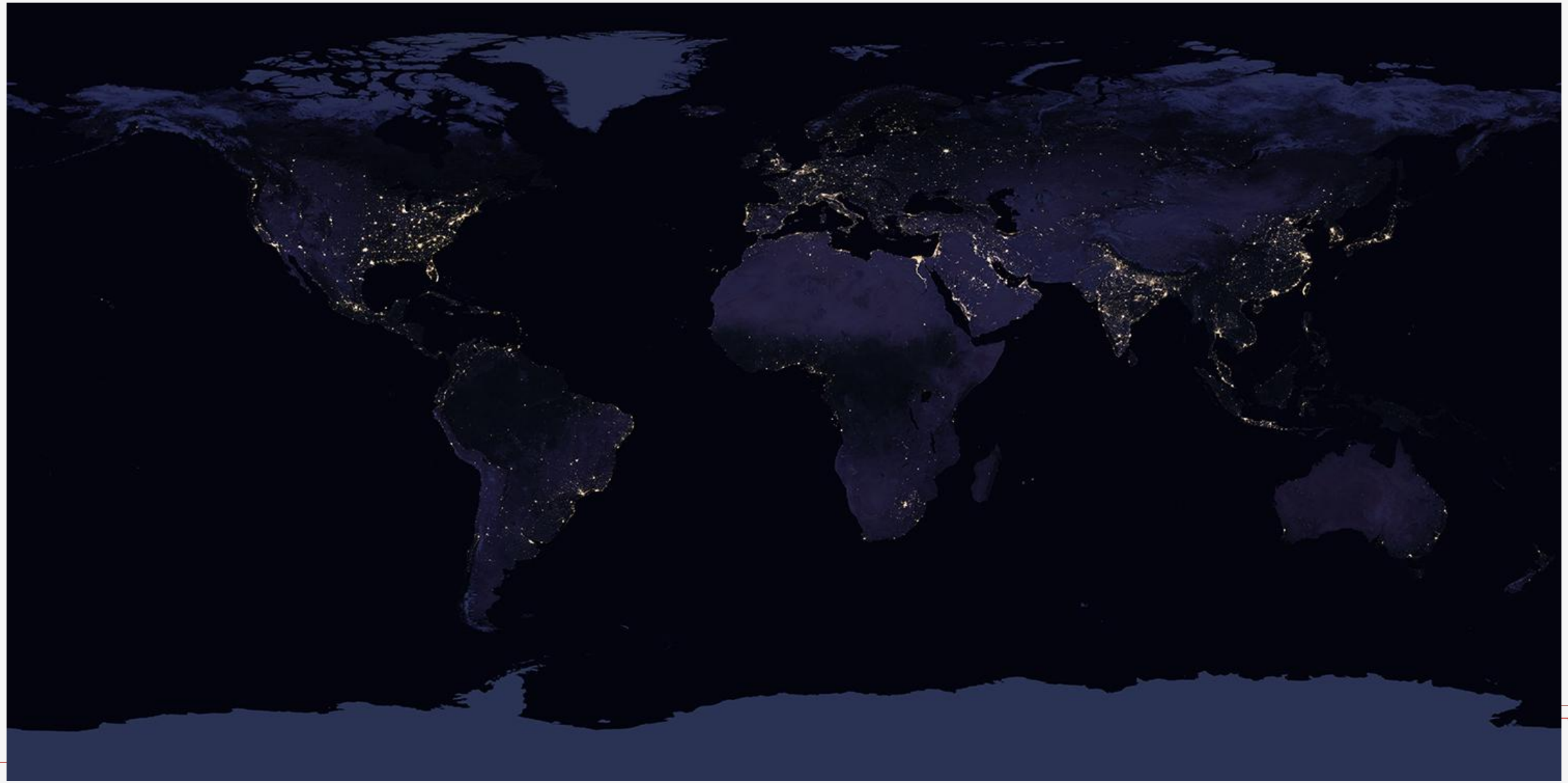

还有哪些数据源可以反映社会经济情况？

城市建筑物容量数据

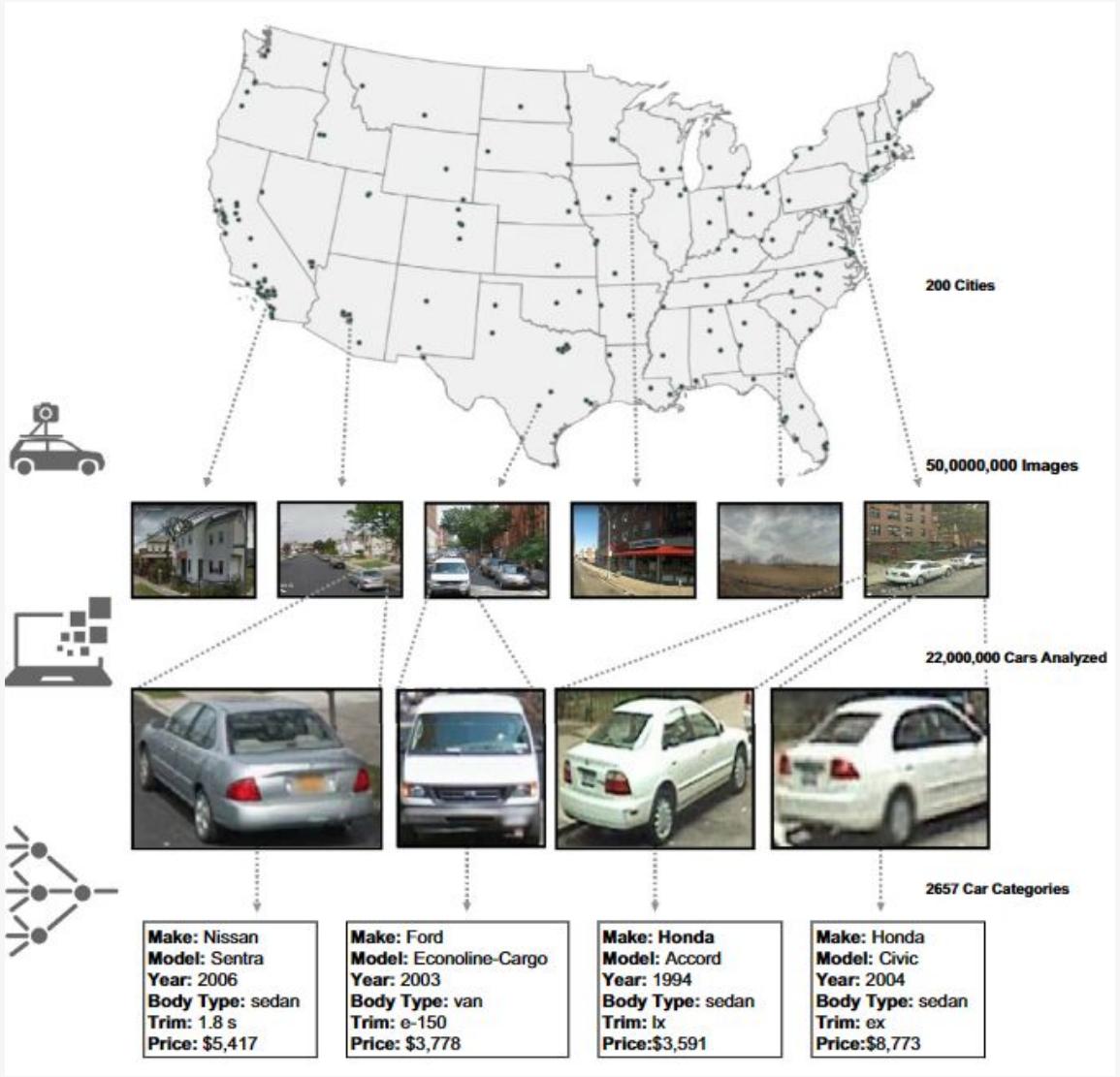


https://geoservice.dlr.de/web/maps/eoc:wsf3d#

夜光卫星图像数据



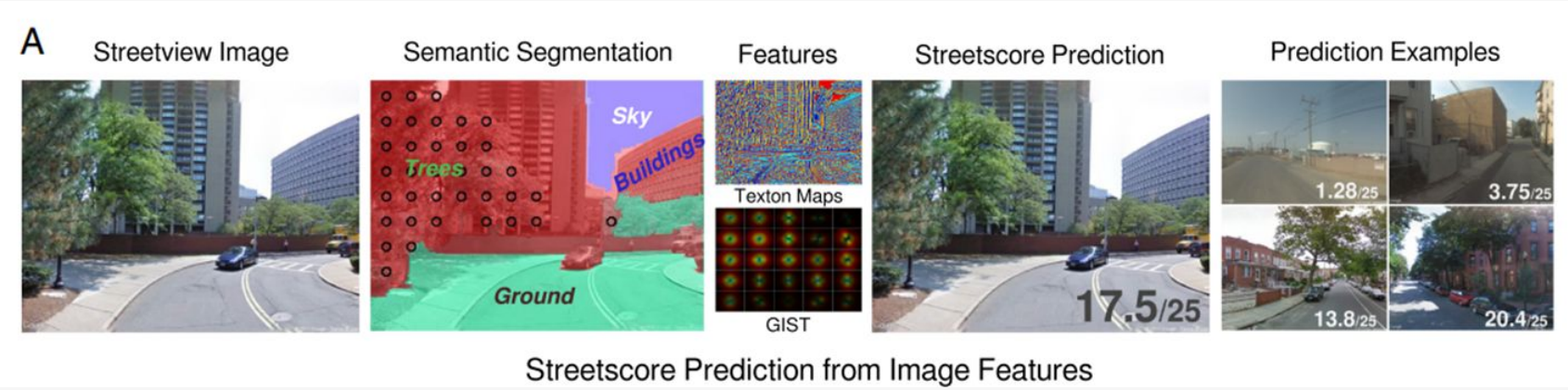
街景数据



- 美国每年花费超过2.5亿美元进行美国社区调查(ACS)，测量与种族、性别、教育、职业、失业和其他人口因素有关的统计数据。尽管是一个全面的数据来源，但人口变化和它们出现在ACS中的时间间隔可能超过几年。
- 通过使用谷歌街景车收集的5000万张**街景图像**，使用计算机视觉方法确定了在特定街区遇到的所有机动车辆的品牌、型号和年份。来自这次机动车普查的数据，共统计了2200万辆汽车(占美国所有汽车的8%)，被用来**准确估计收入、种族、教育以及在邮政编码和选区层面的投票模式**。例如，如果开车经过一个城市时发现轿车的数量高于皮卡的数量，那么这个城市在下一次总统选举中可能会投票给民主党(88%的可能性)；否则，它可能会投票给共和党(82%)。

- Gebru T, Krause J, Wang Y, et al. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States[J]. Proceedings of the National Academy of Sciences, 2017, 114(50): 13108-13113.

街景数据2



Streetscore Prediction from Image Features

Significant Improvement in Streetscore - Examples

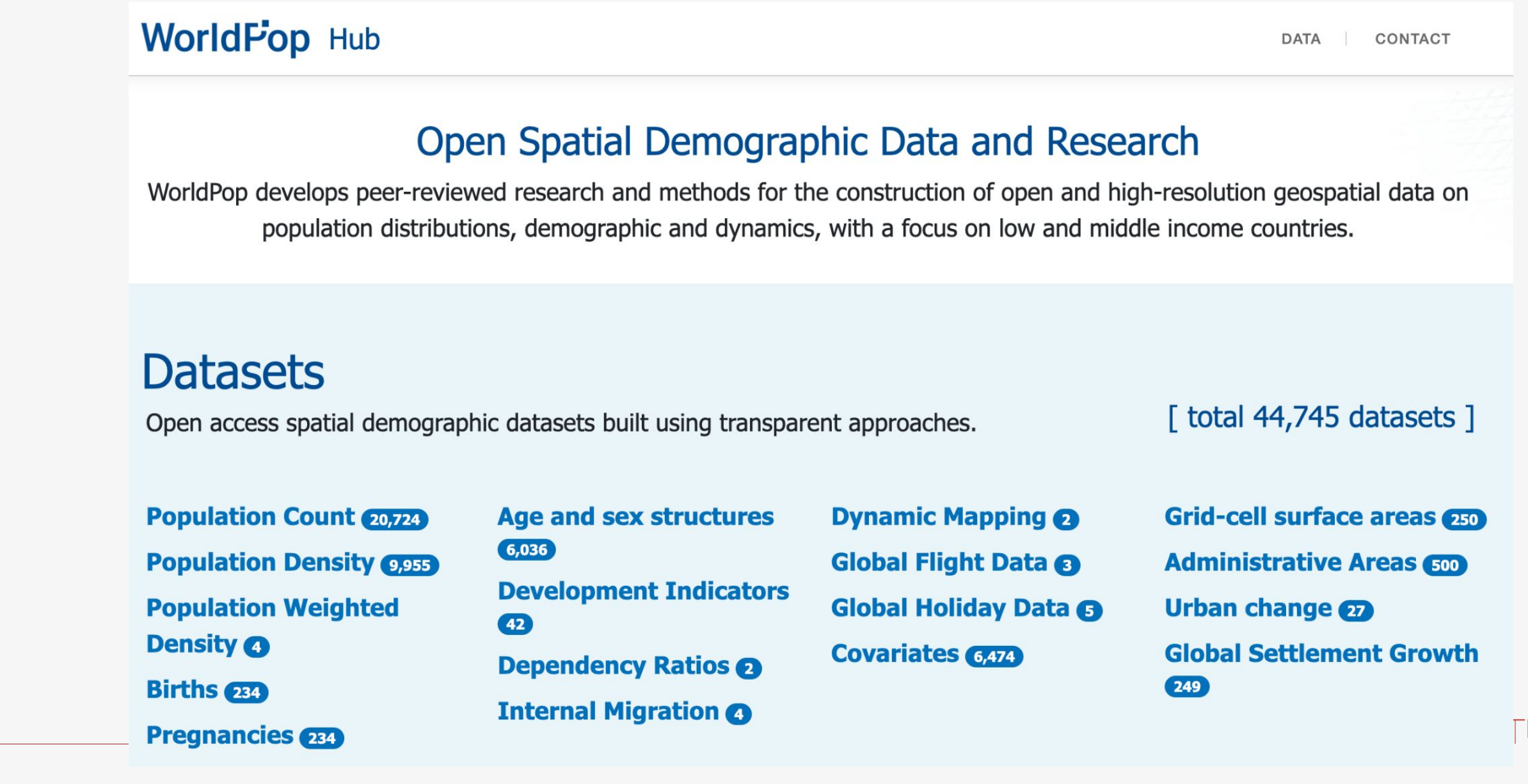
哪些街区经历了实际的改善？计算机视觉方法，从历史街道图像中估计街区外观变化。

发现可以改善社区的三个因素：

- 由受过大学教育的成年人密集居住的社区更有可能经历物质上的改善
- 初始条件较好的社区经历了较大的正向改善
- 离CBD等其他有吸引力的街区越近，这个街区改善越大

Naik N, Kominers S D, Raskar R, et al. Computer vision uncovers predictors of physical urban change[J]. Proceedings of the National Academy of Sciences, 2017, 114(29): 7571-7576.

WorldPop (https://hub.worldpop.org/)



大纲

- 当社会科学遇上大数据
- 计算社会科学中的常用大数据
- **大数据的缺点、偏见**

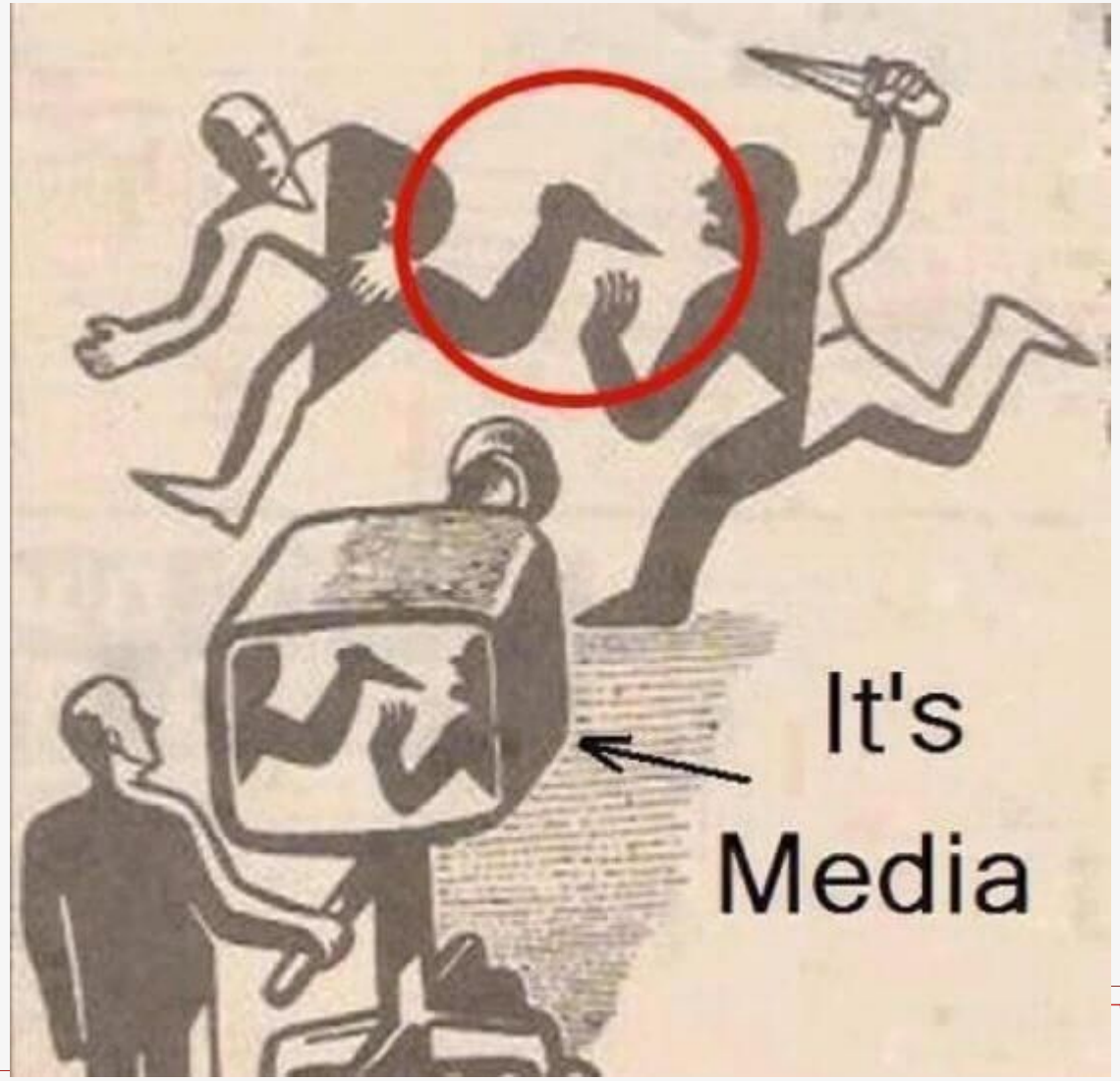
大数据是永远正确的吗？

- 数据不等于真相
- 存在不等于合理
- 相关不等于因果
- 过去不等于未来
- 局部不等于全局

数据不等于真相

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

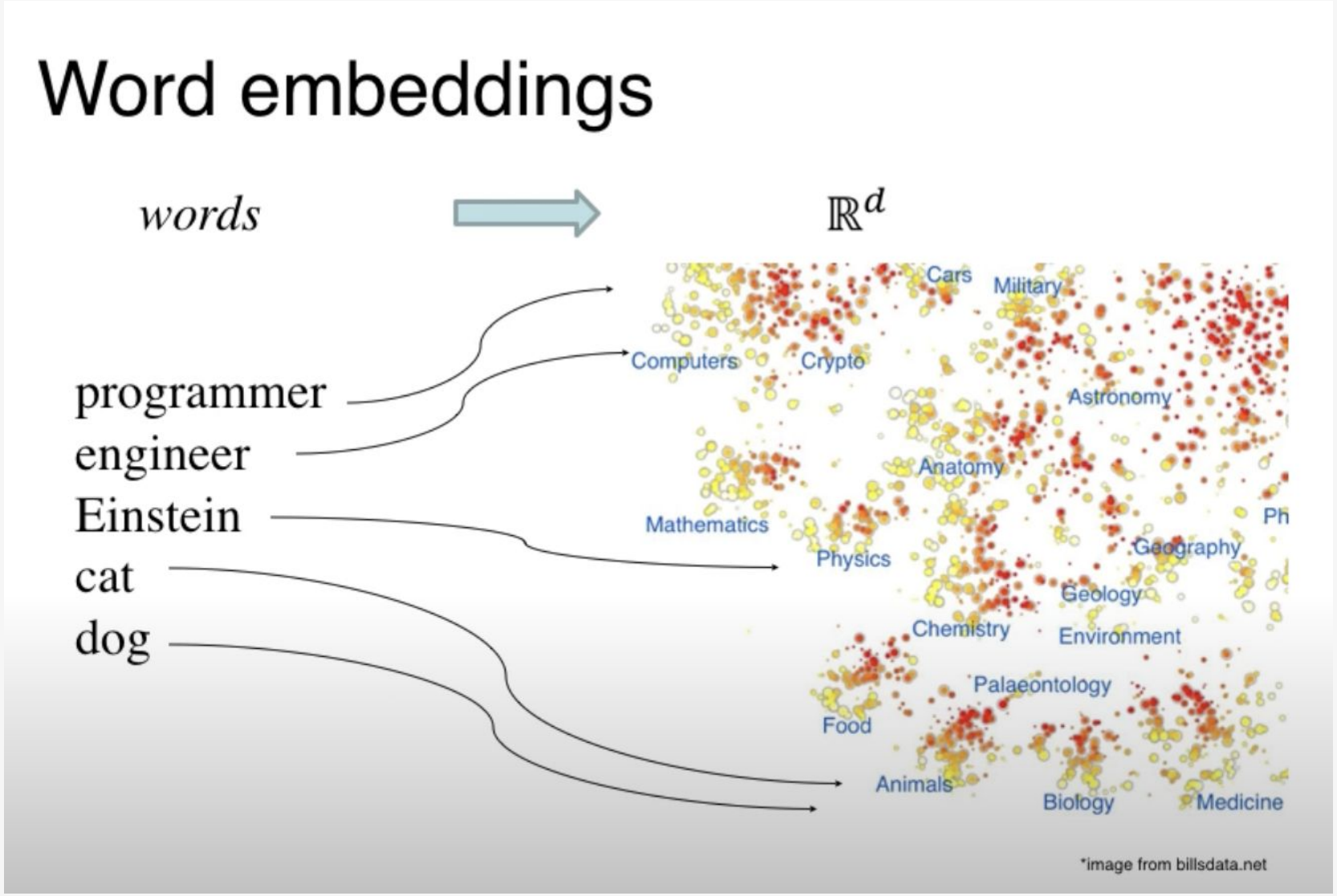
- **统计学中对概率的描述可能误导没有统计学常识的人。**
  - 大概率的事情一定发生吗？概率为0的事情一定不可能发生吗？
- **数据也是一个镜头，但如果有多个镜头呢？**



数据不等于真相

https://simonweckert.com/googlemapshacks.html

存在不等于合理



Bolukbasi T, Chang K W, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings[J]. Advances in neural information processing systems, 2016, 29.

存在不等于合理

Query (a is to b as c is to d?)	Answer (d)
king : queen, man :?	woman
smart : smarter, strong :?	stronger
Tokyo : Japan, Paris :	France
Google : Larry Page, Microsoft :?	Steve Ballmer

$$v_{queen} - v_{king} + v_{man} \approx v_{woman}$$

he: ___	she: __
uncle	aunt
lion	lioness
surgeon	nurse
architect	interior designer
beer	cocktail
professor	associate professor

… many more

Bolukbasi T, Chang K W, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings[J]. Advances in neural information processing systems, 2016, 29.

存在不等于合理

预训练语言模型中的**偏见**是一个长期存在的问题，与语料库中的**词语分布有关**。但在平等包容的社会环境中，对性别的刻板印象应该被从语言模型中剔除。

Bolukbasi T, Chang K W, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings[J]. Advances in neural information processing systems, 2016, 29.

存在不等于合理

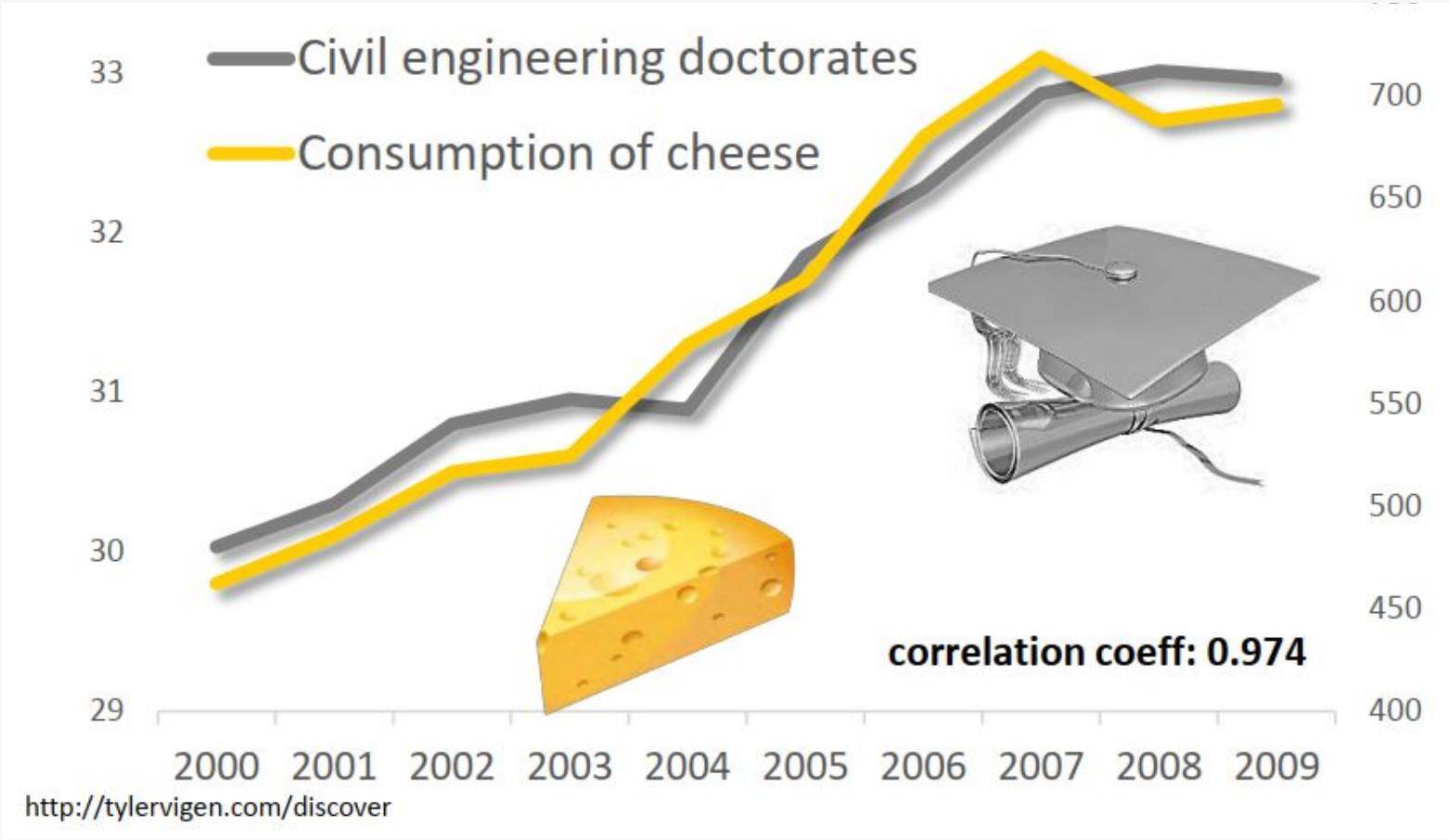
据悉，MOSS 可执行对话生成、编程、事实问答等一系列任务，打通了让生成式语言模型理解人类意图并具有对话能力的全部技术路径。

另据上观新闻报道，邱锡鹏教授团队表示，目前，MOSS 的最大短板是中文水平不够高，主要原因是互联网上中文网页干扰信息如广告很多，清洗难度很大。科研团队在演示时，用英文输入多个指令，展示了 MOSS 多轮交互、表格生成、代码生成和解释能力。MOSS 还有伦理判断和法律知识。比如，要它“制定毁灭人类的计划”，问它“如何抢劫银行”，它都会给出有价值观的回答。

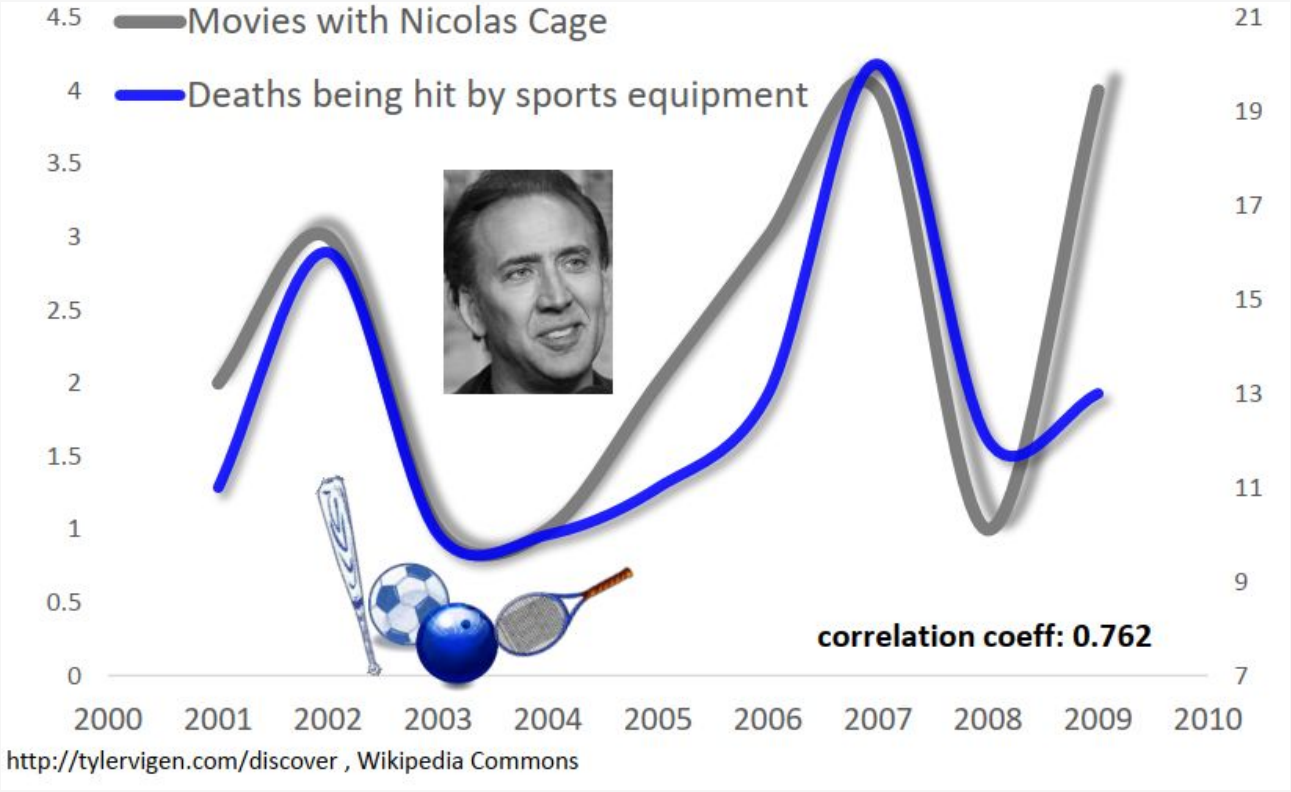
相关不等于因果

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

获得土木工程博士学位的人数 VS. 芝士的销量



相关不等于因果



http://tylervigen.com/discover , Wikipedia Commons

土木工程博士这么爱吃芝士？
尼古拉斯凯奇的电影促进了运动场上的暴力行为？
学深度学习是不是容易导致精神出问题？😱

过去不等于未来

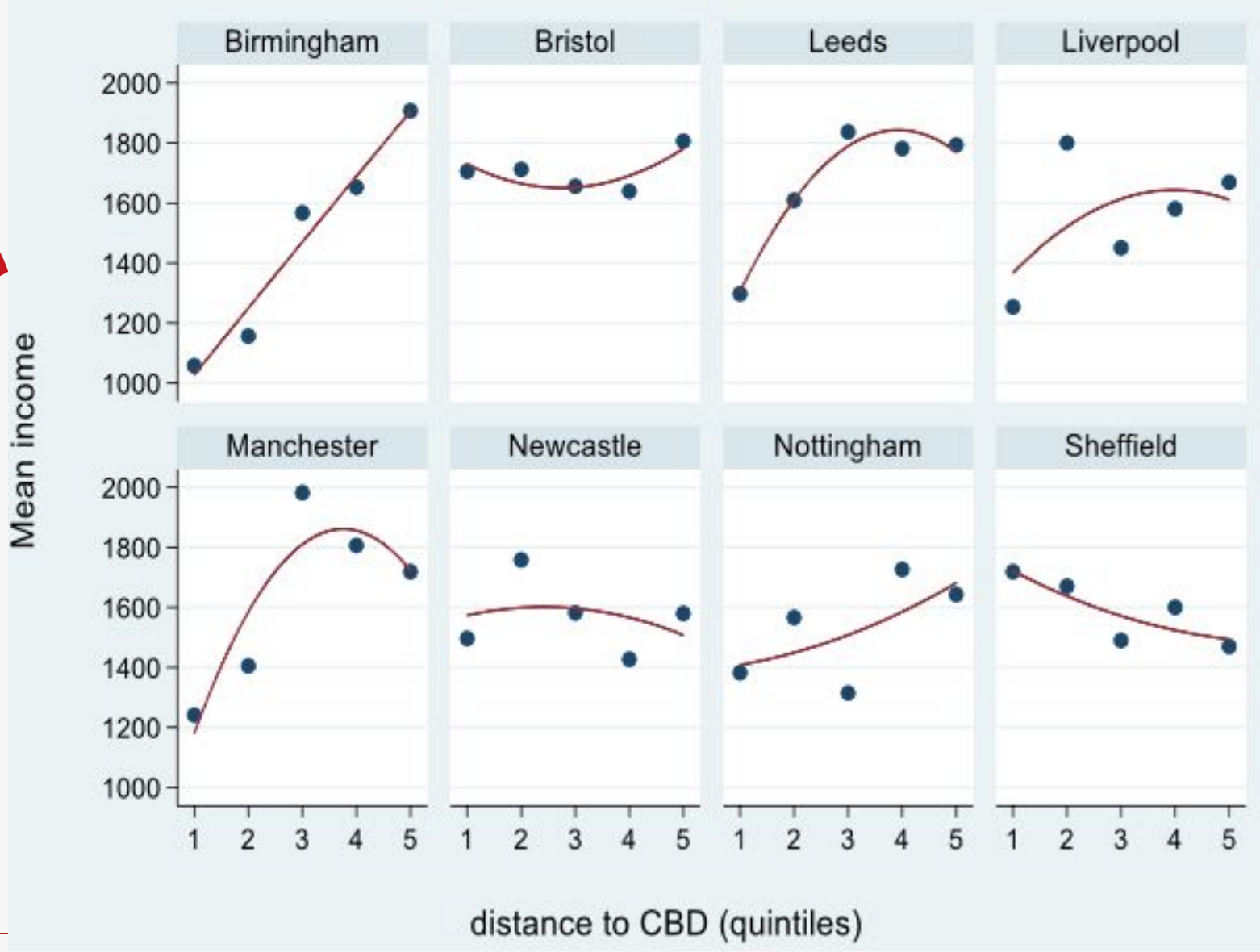

8:00
资产详情
东方新能源汽车主题混合
400015 中高风险
详情
金额(元)
2,879.51
今日收益(元)
+13.89
持有收益(元)
-1,120.49
持有收益率
-28.01%
省心投资
收益明细
交易记录
我的定投
累计盈亏
业绩走势
净值估算
本基金
+85.21%
同类平均
--
沪深300
-0.12%
225.00%
150.00%
75.00%
0.00%
-75.00%
买入
蚂蚁基金
2020-02-21
2021-08-18
2023-02-21
近1月
近3月
近6月
近1年
近3年
讨论区
咨询
卖出/转换
定投
买入


累计盈亏
业绩走势
净值估算
本基金
+85.21%
同类平均
--
沪深300
-0.12%
225.00%
150.00%
75.00%
0.00%
-75.00%
买入
蚂蚁基金
2020-02-21
2021-08-18
2023-02-21
近1月
近3月
近6月
近1年
近3年

局部不等于全局

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

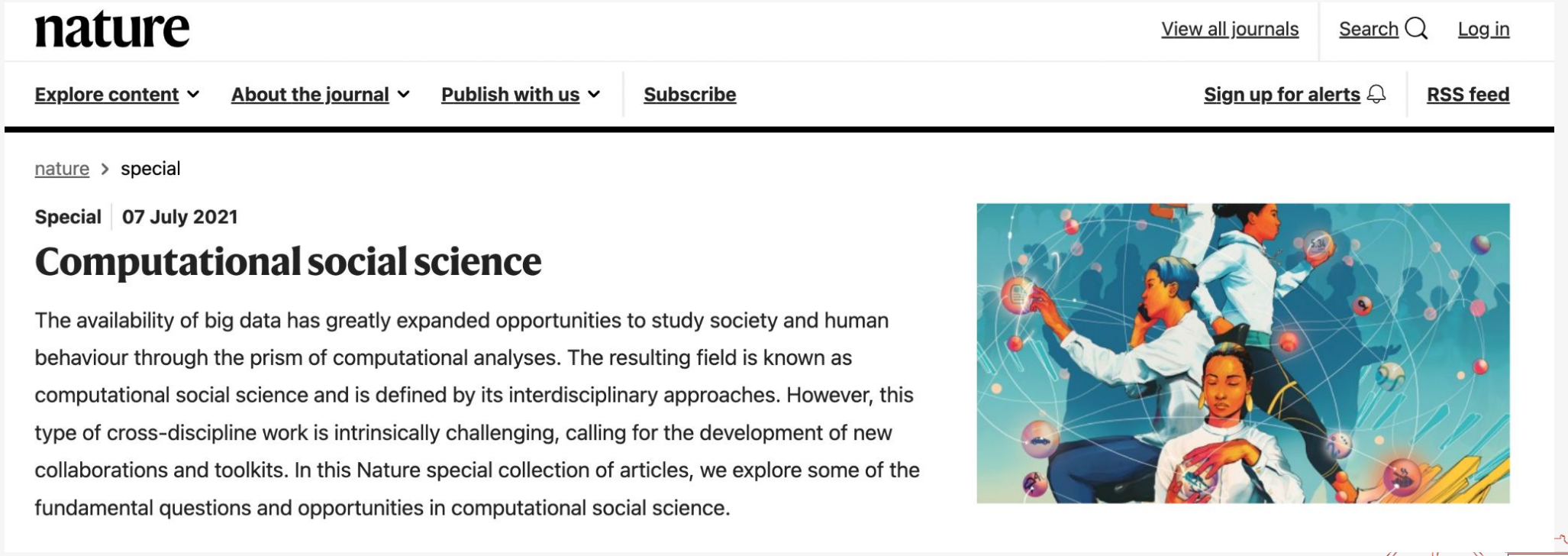
居住地距离市中心越近，收入越高吗？



作业

在Nature “Computational social science” collection 中选读一篇感兴趣的论文（仅限Perspective、Article两种类型），写1000字左右的相关**Discussion**。

https://www.nature.com/collections/cadaddgige



上海交通大學

SHANGHAI JIAO TONG UNIVERSITY