Computational Social Science

Its Recent Development, Opportunities, and Challenges

Yongjun Zhang, Ph.D.

Department of Sociology
Institute for Advanced Computational Science
State University of New York at Stony Brook
Research Affiliate at New York University

July 14, 2022

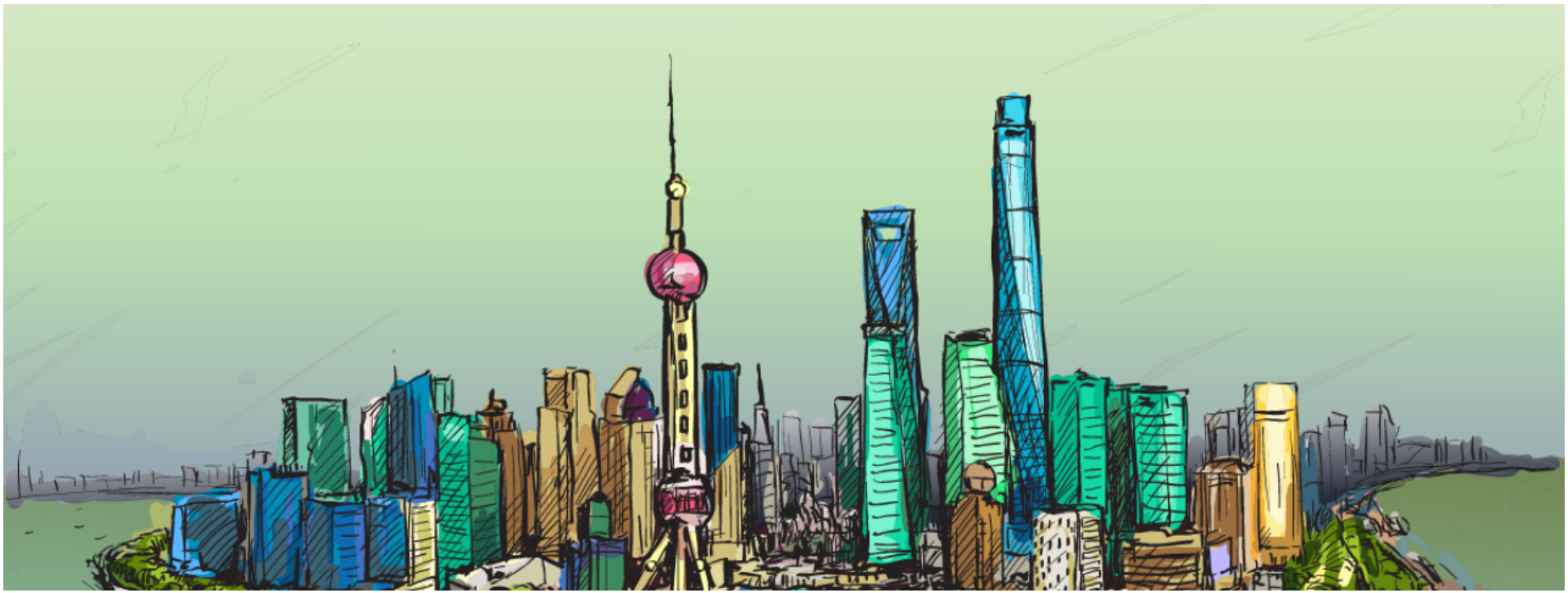
Table of Contents

Welcome and Introduction

Center for

Applied Social and Economic Research

应 用 社 会 经 济 研 究 中 心



Some Logistics

- Course materials: https://yongjunzhang.com/intro2css
- Wechat Group Channel
- Slack Channel: https://join.slack.com/t/css-nyush/shared_invite/zt-1bvm0es8y-FwK5JrP~yE9skyw~eB2_bg
- Research speed date: https://docs.google.com/spreadsheets/d/1huylBZzT3uH5TrdZjiBvpZEpn-ytJaj9pexjtJCycmw/edit?usp=sharing
- My office hour: 8PM-9PM, via Zoom https://stonybrook.zoom.us/j/97533572022?pwd=Z2ZRWFpRbGk0eFB6WTVMNnBkZkZVZz09
- Course recording: We will record the whole sessions

I combine computational, statistical, and network methods with big data to study social, political, and org behavior, including *segregation* and *polarization*.

Yongjun Zhang Data CV ZLab AAPI

Sociologist, Ph.D.

Stony Brook, NY
Twitter
Google Scholar
ORCID

Welcome

I (Chinese: 张勇军audio) am a computational social scientist studying politics, organizations, social movements, and inequality. I received my Sociology PhD in 2020 from the University of Arizona. I am currently working as an Assistant Professor of Sociology and Institute for Advanced Computational Science at Stony Brook University. I am also a research affiliate at New York University.

I combine statistical, network, and computational methods with large-scale datasets to study social, political, and organizational behavior. My past work focuses on the interplay between social movements and social or political changes in the U.S. and the globe. These studies have been published in Journal of Marriage and Family, Demography, Poetics, International Journal of Comparative Sociology, American Journal of Sociology, and PLoS One. I have won the 2020 James Coleman Award from Sociology of Education Section at American Sociological Association and the 2021 SIM Best Paper Submission at Academy of Management.

My ongoing work focuses mainly on understanding mobility, segregation, and polarization in the U.S. I am using big data from FEC with corp data to track the polarization/partisanship trend in corporate elites. I am using population mobility data from SafeGraph and Facebook as well as 190 million voter records and 260 million consumer records to assess the antecedents and consequences of racial/partisan/income/cultural segregation. This human mobility, segregation, and polarization project has been funded by an OVPR seed grant at Stony Brook University. I am also using deep learning methods to detect and monitor anti-AAPI hate speech and incidents (a direct result of polarization and xenophobia) from Twitter since the COVID-19 outbreak. This project has been funded by a seed grant from IACS at Stony Brook University.

I am teaching Intro to Computational Social Science and Research Methods in Sociology at Stony Brook University. I am also guest-editing a special issue on computational social science and Chinese societies for Chinese Sociological Review.

One project focuses on partisanship and polarization in Corporate America using administrative, social media, voter records, and earnings call data.

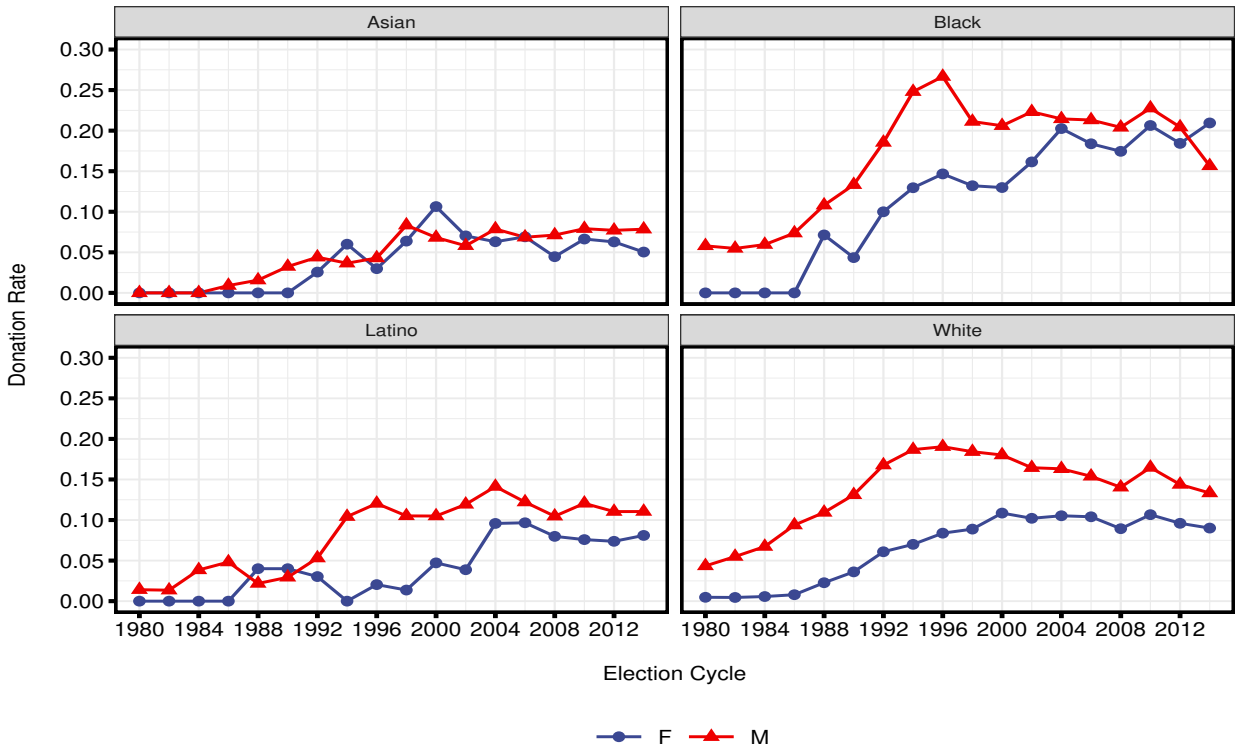


Figure: Political Donation Rate by Gender and Race in Corporate America. We merged billions of FEC donation records with .6 million of BoardEx corporate leaders.

The second project focuses on polarization on social media by monitoring and detecting anti-AAPI hate speech using billions of Tweets.

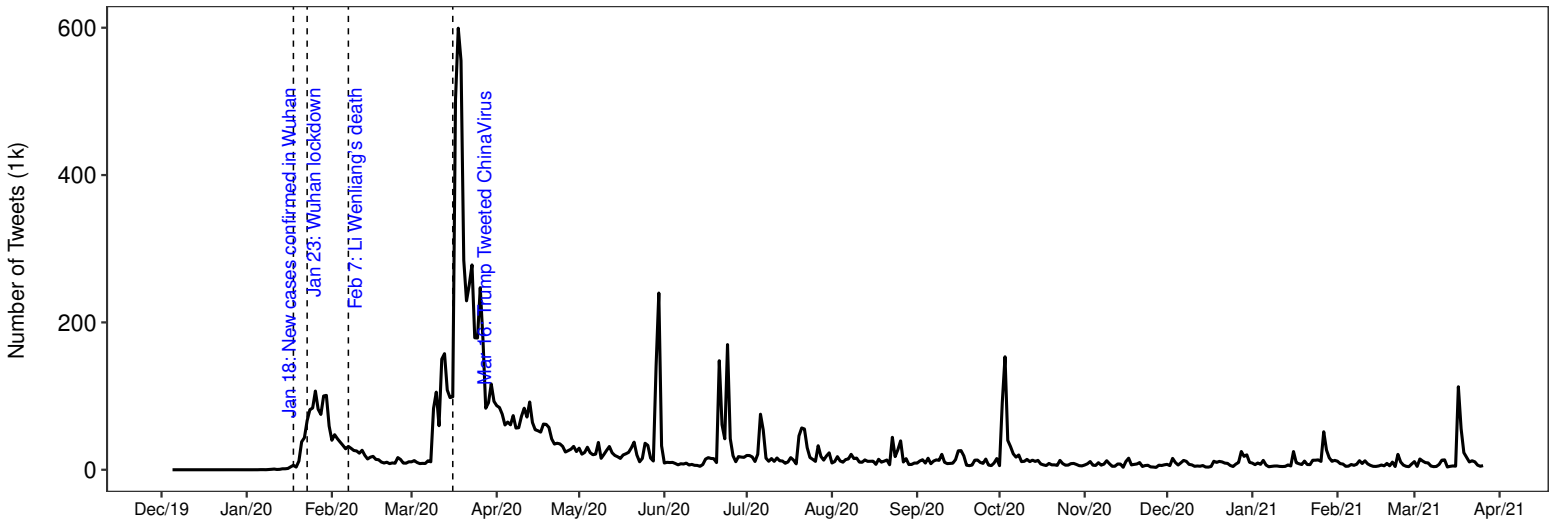


Figure: Daily Trend of Racial Slurs on Twitter. Data were collected using keywords related to ChinaVirus, KungFlu, and CCPVirus in the pandemic. Source: https://yongjunzhang.com/aapi/

The third project focuses on multiplex segregation in the U.S. using 190 million L2 Voter Records and 260 million Infutor Consumer Records as well as Facebook and Safegraph's relational data.

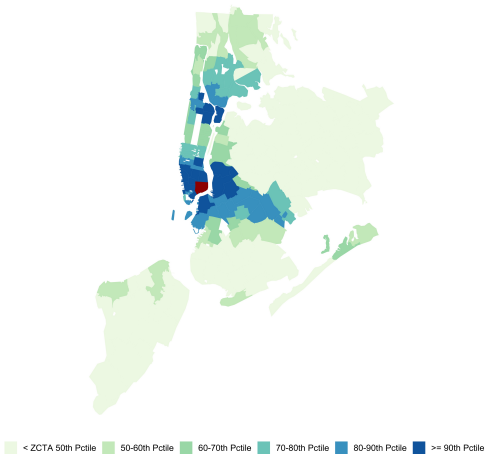


Figure: Facebook SCI Data

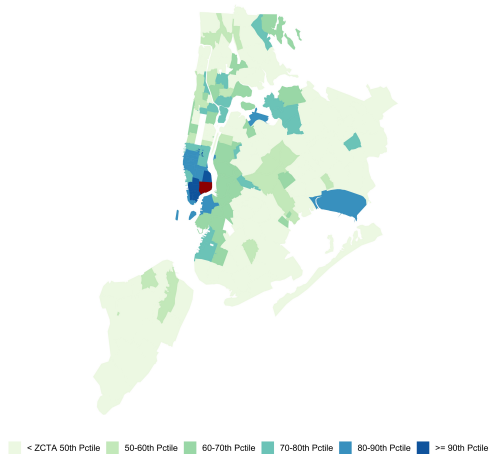


Figure: Safegraph mobility data.

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.

Conventional movement scholars coded New York Times for Protest Events

Dynamics of Collective Action

Published papers

Home

Below is a list of published articles that have used the dataset available on this website.

Earl, Jennifer, Sarah A. Soule, and John D. McCarthy. 2003. "Protest Under Fire? Explaining the Policing of Protest." American Sociological Review 68(4): 581-606. Link

Earl, Jennifer and Sarah A. Soule. 2006. "Seeing Blue: A Police-Centered Explanation of Protest Policing." Mobilization 11(2): 145-164. Link

King, Brayden G., Keith G. Bentele and Sarah A. Soule. 2007. "Protest and Policymaking: Explaining Fluctuation in Congressional Attention to Rights Issues, 1960-1986." Social Forces 86(1):137-164. Link

King, Brayden G. and Sarah A. Soule. 2007. "Social Movements as Extra-Institutional Entrepreneurs: The Effect of Protest on Stock Price Returns." Administrative Science Quarterly 52: 413-42. Link

Menu

- Home
- Data
- Documentation
- Papers
- Contact

Click 'reload' in your browser to see another image above.

Machine-Learning Protest Events Data System

MPEDS - Automated Coding of Protest Event Data

Machine-Learning Protest Event Data System

MPEDS Annotation Interface (MAI)

This is the annotation interface used in creating datasets for the Machine-learning Protest Event Data System (MPEDS). While applied to the specific task of coding for protest events, this can also be used for the development of other types of event datasets.

The MAI is available at https://github.com/mpeds/mpeds-coder.

A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media

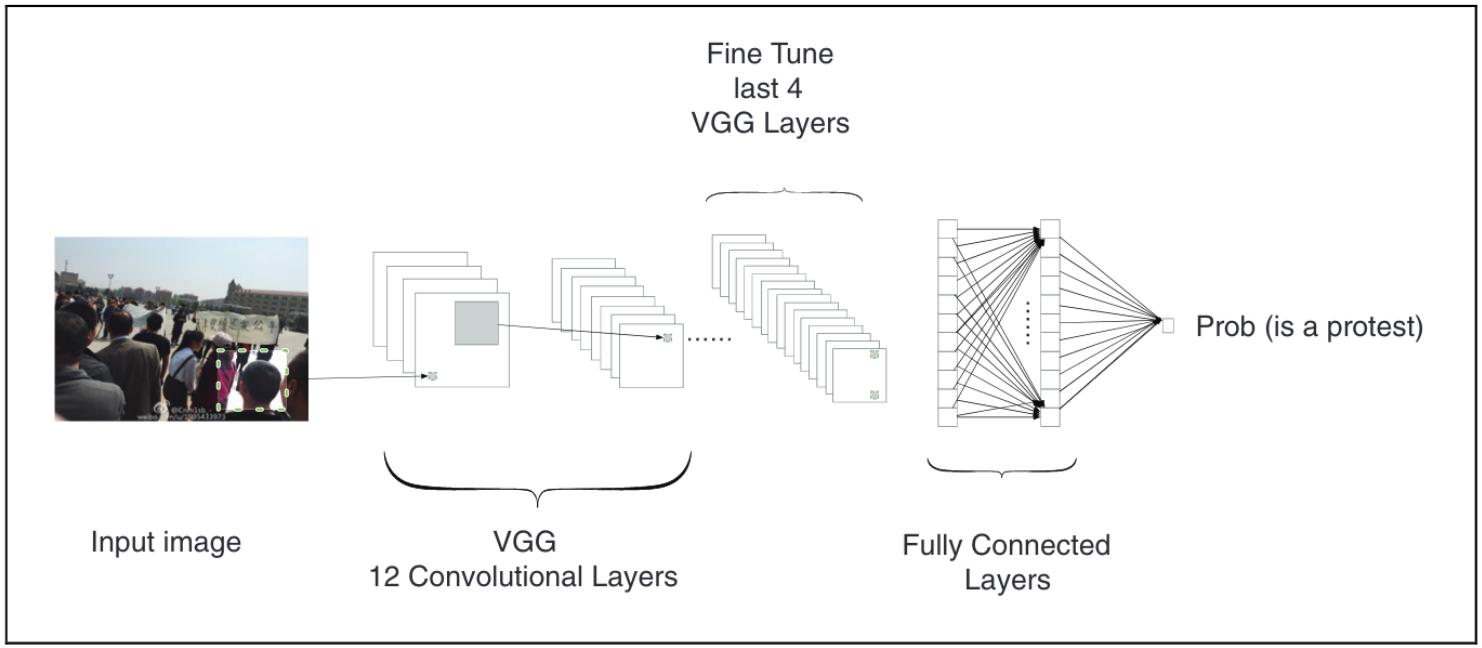
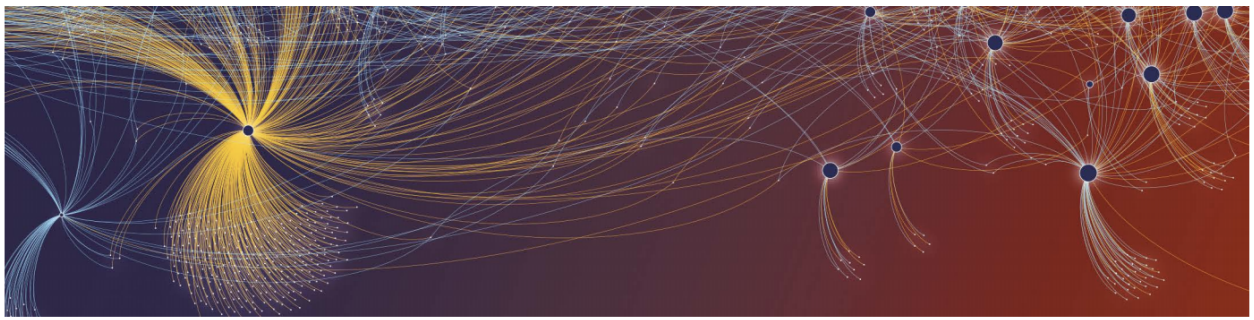


Figure 1. Illustration of our convolutional neural network architecture for image classification.
Note. Input image from Weibo.com.

Computational Methods Have Transformed the Way How Social Scientists Do Research



Scientists studied data from thousands of social-media users to analyse clusters perpetuating extremism.

COMPUTING HUMANITY

How data from Facebook, Twitter and other sources are revolutionizing social science. **By Heidi Ledford**

What Is Computational Social Science?

- CSS was originally used to describe agent-based modeling in social science.

What Is Computational Social Science?

- CSS was originally used to describe agent-based modeling in social science.
- Any study that uses large-scale datasets that describe human behavior in STEM.

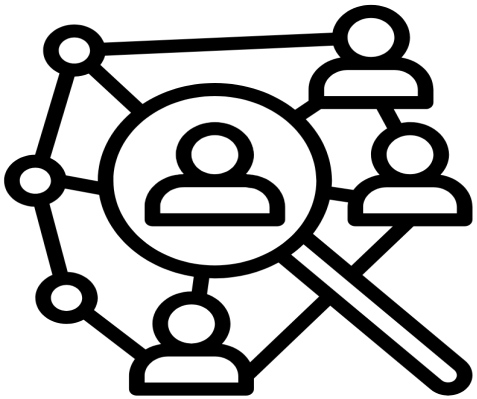
What Is Computational Social Science?

- CSS was originally used to describe agent-based modeling in social science.
- Any study that uses large-scale datasets that describe human behavior in STEM.
- The development and application of computational methods to complex, typically large-scale, human (sometimes simulated) behavioral data.

What Is Computational Social Science?

- CSS was originally used to describe agent-based modeling in social science.
- Any study that uses large-scale datasets that describe human behavior in STEM.
- The development and application of computational methods to complex, typically large-scale, human (sometimes simulated) behavioral data.
- An interdisciplinary field that advances theories of human behavior by applying computational techniques to large datasets from social media sites, the Internet, or other digitized archives such as administrative records.

Three Cores when Defining CSS



(a) Social Science

(b) Computer Science

(c) Big Data

Two Orientations?

- Object-oriented: See computational tools as the study object. The Impact of AI, Big Data, Machine Learning, etc.

Two Orientations?

- Object-oriented: See computational tools as the study object. The Impact of AI, Big Data, Machine Learning, etc.
- Instrument-oriented: See computational tools as a means to advance research. Using Computational Methods to Study Social and Human Behavior

Big Data Surveillance: the Case of Policing

American Sociological Review
2017, Vol. 82(5) 977–1008

DOI: 10.1177/0003122417725865
journals.sagepub.com/home/asr
SAGE

Big Data Surveillance: The Case of Policing

Sarah Brayne[a]

Abstract

This article examines the intersection of two structural developments: the growth of surveillance and the rise of "big data." Drawing on observations and interviews conducted within the Los Angeles Police Department, I offer an empirical account of how the adoption of big data analytics does—and does not—transform police surveillance practices. I argue that the adoption of big data analytics facilitates amplifications of prior surveillance practices and fundamental transformations in surveillance activities. First, discretionary assessments of risk are supplemented and quantified using risk scores. Second, data are used for predictive, rather than reactive or explanatory, purposes. Third, the proliferation of automatic alert systems makes it possible to systematically surveil an unprecedentedly large number of people. Fourth, the threshold for inclusion in law enforcement databases is lower, now including individuals who have not had direct police contact. Fifth, previously separate data systems are merged, facilitating the spread of surveillance into a wide range of institutions. Based on these findings, I develop a theoretical model of big data surveillance that can be applied to institutional domains beyond the criminal justice system. Finally, I highlight the social consequences of big data surveillance for law and social inequality.

Gender Bias in Image Recognition System

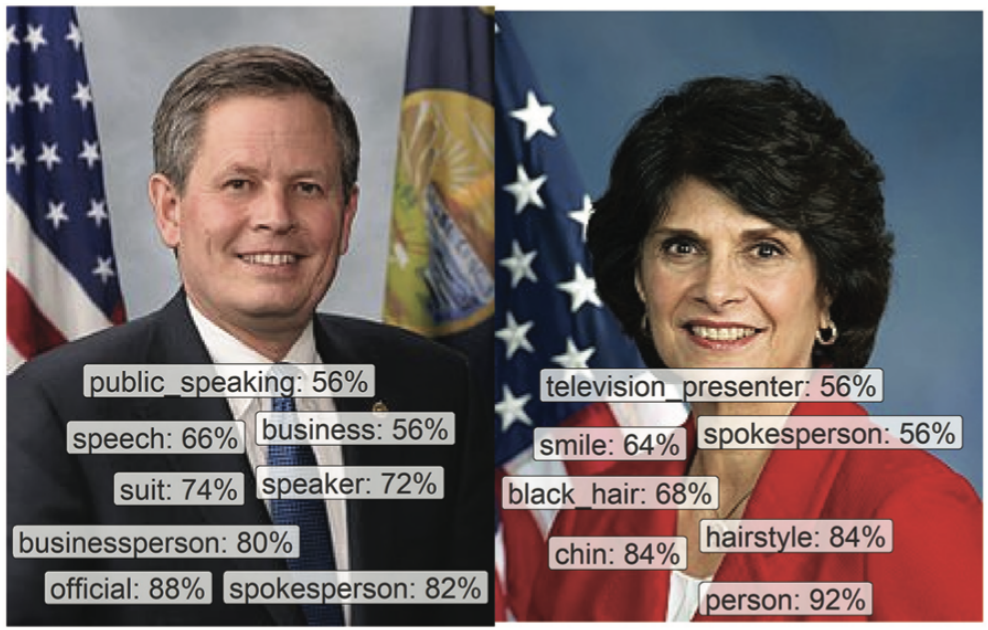


Figure 5. Two images of U.S. members of Congress with their corresponding labels as assigned by Google Cloud Vision. On the left is Steve Daines, a Republican senator from Montana. On the right is Lucille Roybal-Allard, a Democratic representative from California's 40th congressional district. Percentages next to labels denote confidence scores of Google Cloud Vision.

Figure: Schwemmer et al. 2020. Socius.

Racial Bias in Health Care System Related Algorithm

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Algorithm Bias in Hate Speech Detection

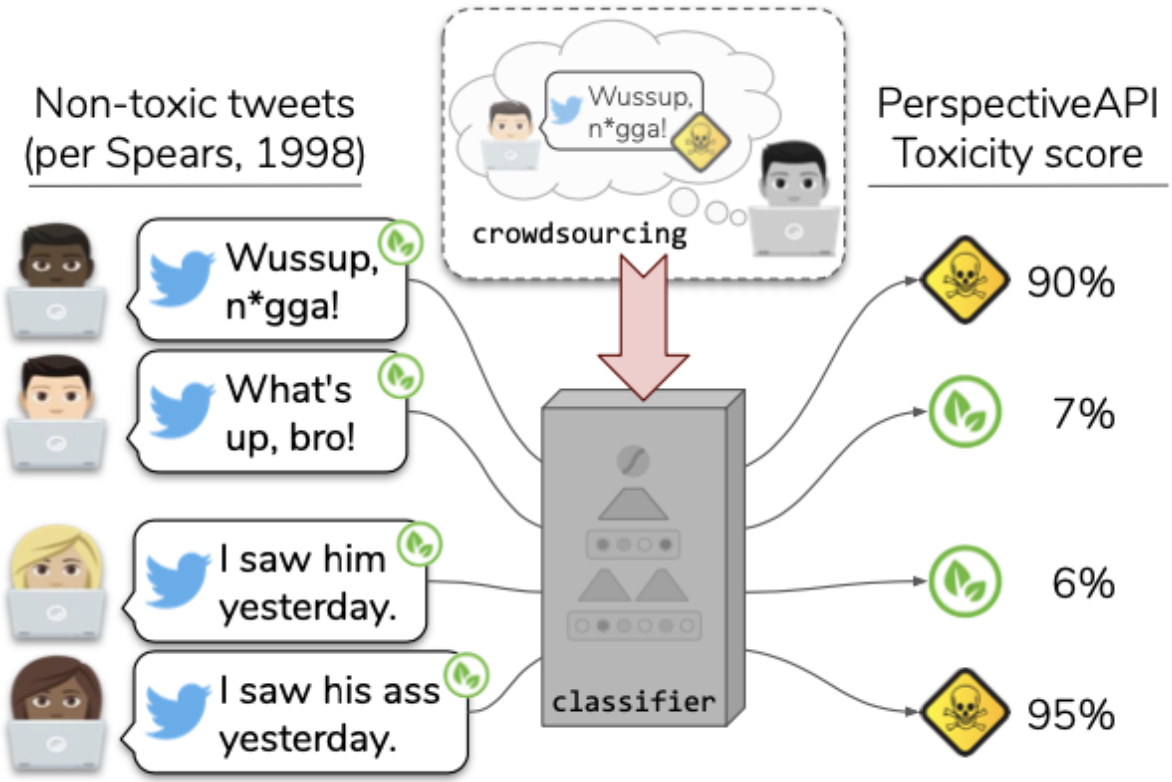


Figure: Maarten Sap et al. 2020

Two Orientations?

- Scholars do not see studies focusing on computational tools per se as CSS.
- CSS scholars tend to see computational tools as a means to tackle big social science problems and advance theory.

Table of Contents

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 in the network. We check our emails 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.

Figure: We are moving toward the METAVERSE!

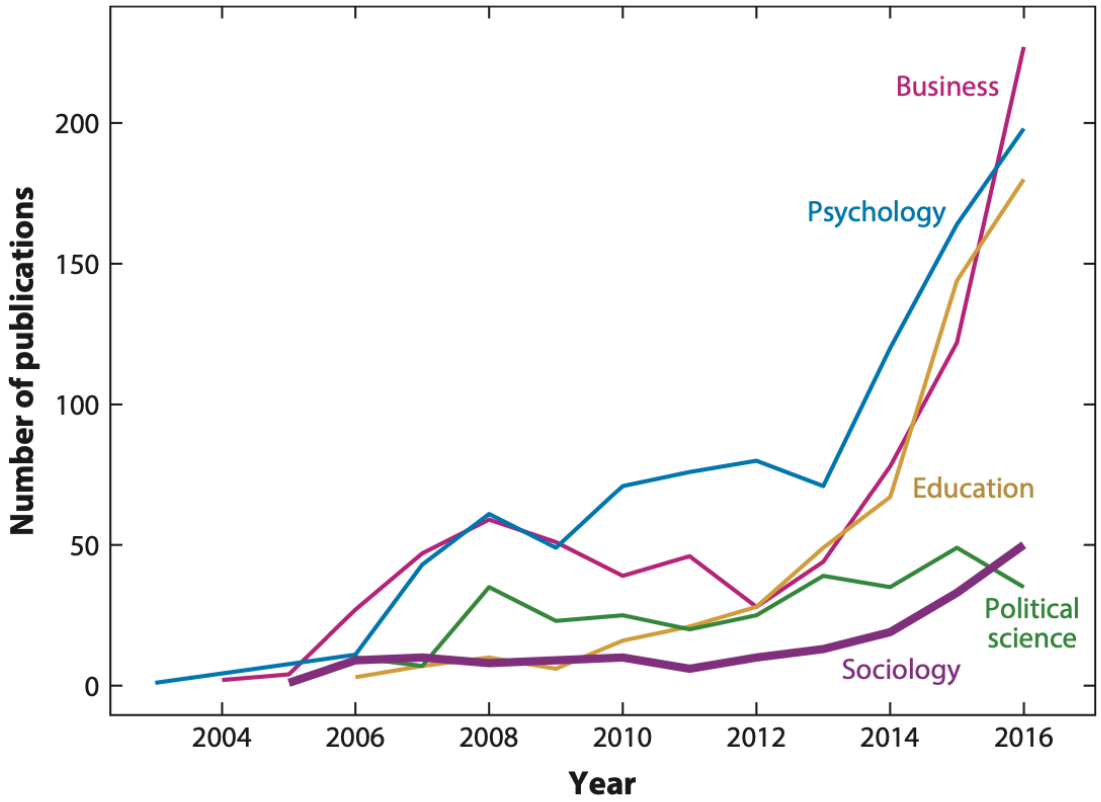


Figure 1

Number of computational social science publications by year—2003–2016—across five scholarly disciplines.

Figure: The development of CSS

The Recent 10 Yrs

- The Access to Large-scale Structured and Unstructured Data
- The Rapid Development of Big Data Analytic
- The Access to Cloud Computing
- The Emergence of CSS Community
- The Funding Priorities

Structured Data

L2
BETTER DATA
BETTER DECISIONS

Figure: 190 Million Voter History Records

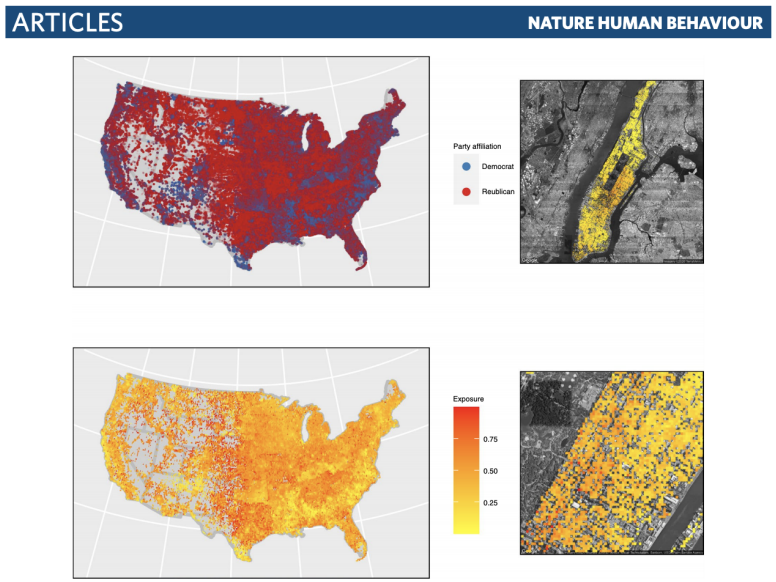


Fig. 2 | Measuring spatial exposure across increasingly small geographies. The exact residential location of every Democrat and Republican in the United States ($n = 180{,}660{,}202$, top left) can be used to measure each Democrat's spatial exposure to Republicans, and this can be averaged across arbitrarily small grid cells for display purposes (1,000 × 1,000 grid, bottom left). Exposure can be examined across any resolution: markedly different residential exposure to Republicans can be seen in Manhattan, NY (500 × 500 grid, top right), with Democrats on the northern and southern extremes of the island having almost no residential exposure to Republicans, whereas Democrats on the Upper East Side (the neighborhood immediately to the right of the lower section of Central Park, which is the long rectangle with no voters in it located in the center of the island) have exposure as high as 0.5 due to the clustering of Republicans in this area. A magnified view of the Upper East Side of Manhattan (75 x 75 grid, bottom right) shows the clustering of Republicans along Central Park and thus Democrats' decreasing exposure to Republicans moving towards the northeast. Map data (righthand figures): Google, TerraMetrics.

Figure: The Measurement of Partisan Sorting of 180 Million Voters

Text as Data

Stanford | SSDS Social Science Data Collection

PROBLEM DOWNLOADING FILES?

Contact Ron Nakao at consult-ssds@lists.stanford.edu.

DATA

COMPARATIVE INCOME TAXATION DATABASE (CITD)

CONGRESSIONAL RECORD FOR THE 43RD-114TH CONGRESSES: PARSED SPEECHES AND PHRASE COUNTS

DATABASE ON IDEOLOGY, MONEY IN POLITICS, AND ELECTIONS: PUBLIC VERSION 2.0

Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts

Submitted by David Michael R... on Wed, 10/25/2017 - 11:40

Abstract:
This dataset contains processed text from the bound and daily editions of the United States Congressional Record, as provided by HeinOnline. The bound edition covers the 43rd to 111th Congresses, and the daily edition covers the 97th to 114th. Each edition includes all text spoken on the floor of each chamber of Congress: the United States House of Representatives and the United States Senate. An automated script parses the text from each session to produce full-text speeches, metadata on speeches and their speakers, and counts of two-word phrases (bigrams) by speaker and party. Text is aggregated over sessions to flag bigrams that relate to congressional procedure or are extremely common or rare. The results of a manual audit of the script and statistics on our rate of matching speeches with members of Congress are included as well.

Principal Investigator:
Matthew Gentzkow
Jesse M. Shapiro
Matt Taddy

Figure: Congressional Speech Data

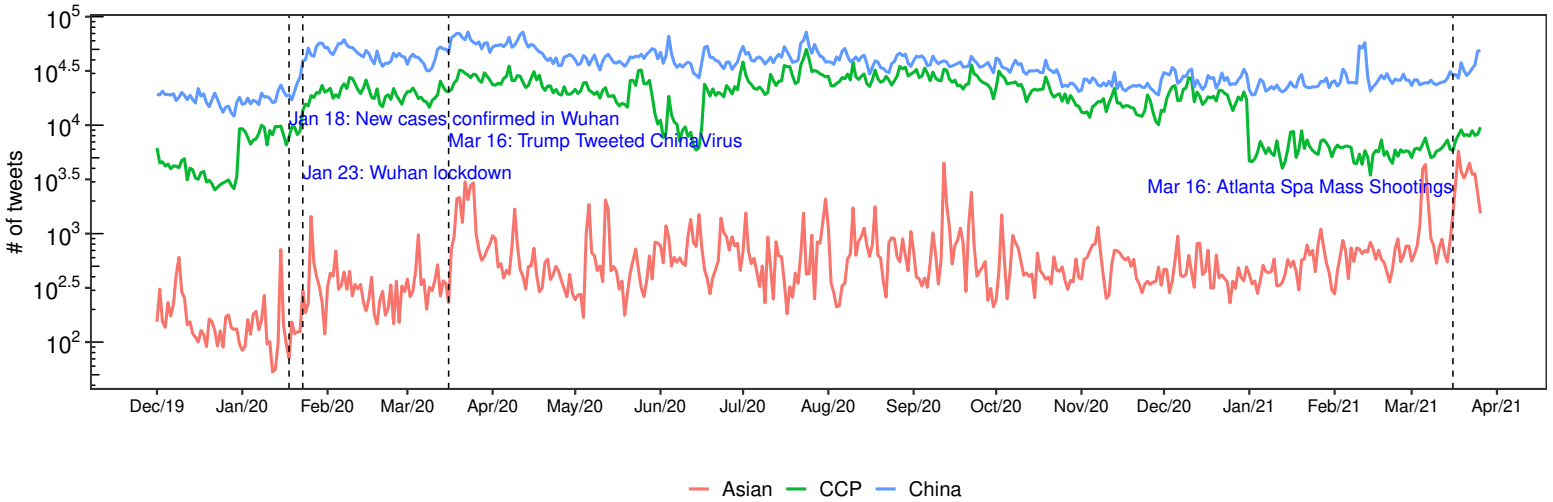


Figure: **Daily Trend of Chinese Tweets mentioning China, Asians/Chinese, and CCP.** Data were collected for both simplified and traditional Chinese tweets from Dec 2020 to April 2021.

Image as data

Figure: The Human Screenome Project

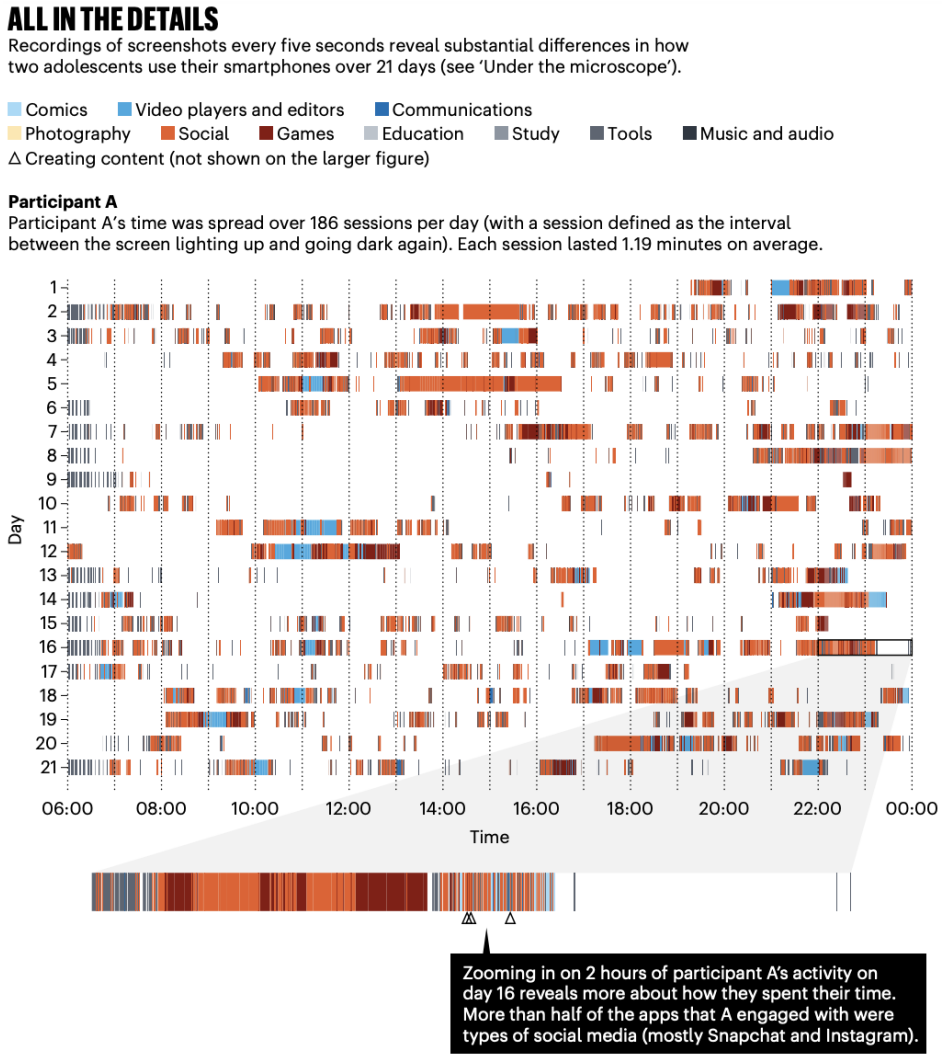
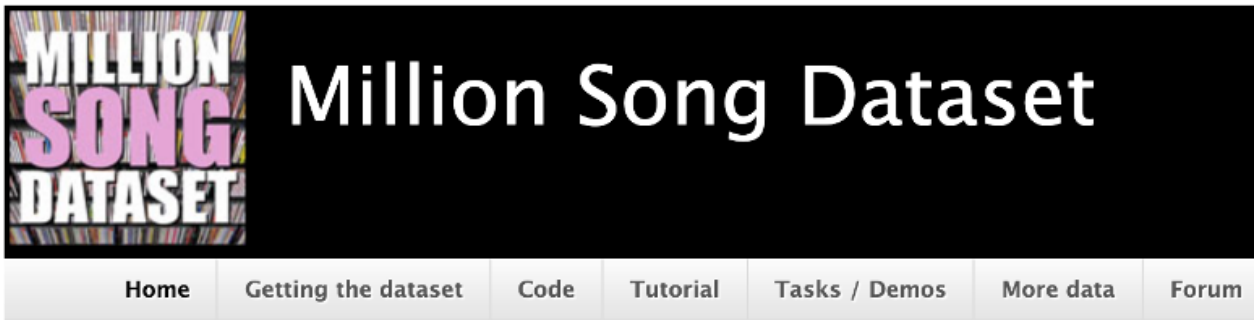


Figure: The Human Screenome Project

Audio as data



Video as data

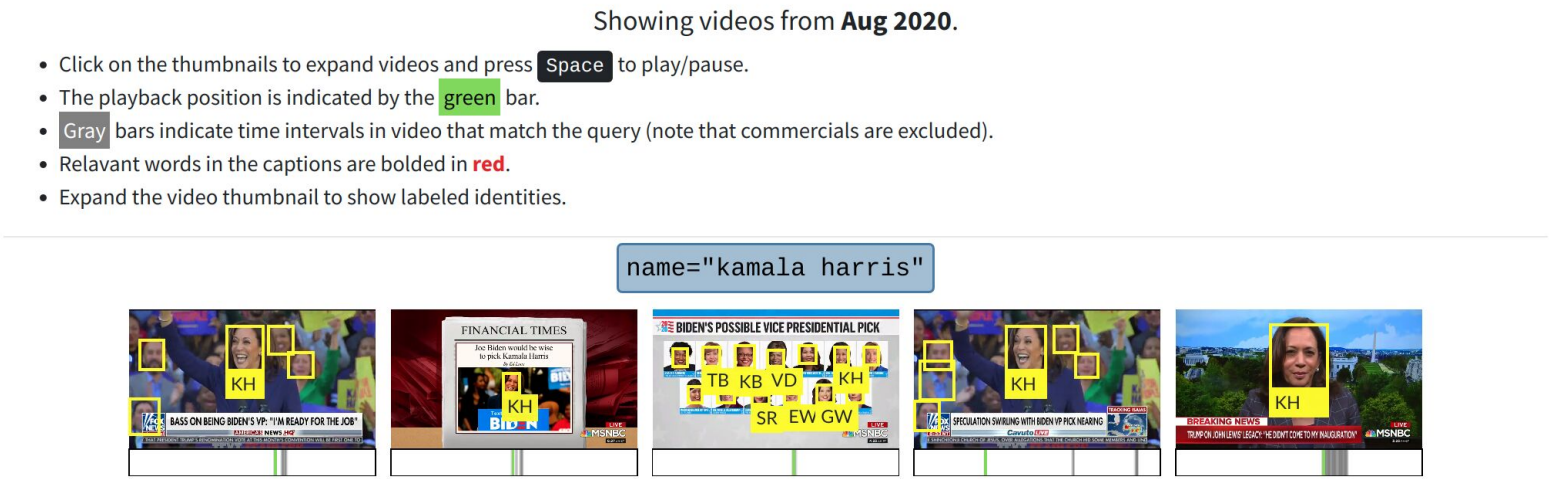


Figure: Stanford Cable TV News Analyzer

https://tvnews.stanford.edu/getting-started
https://tvnews.stanford.edu/data
https://tvnews.stanford.edu/methodology

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-021-01153-1

Check for updates

Banks, alternative institutions and the spatial–temporal ecology of racial inequality in US cities

Mario L. Small[1]✉, Armin Akhavan[2], Mo Torres[1] and Qi Wang[2]

Research has made clear that neighbourhood conditions affect racial inequality. We examine how living in minority neighbourhoods affects ease of access to conventional banks versus alternative financial institutions (AFIs) such as check cashers and payday lenders, which some have called predatory. Based on more than 6 million queries, we compute the difference in the time required to walk, drive or take public transport to the nearest bank versus AFI from the middle of every block in each of 19 of the largest cities in the United States. The results suggest that race is strikingly more important than class: even after numerous conditions are accounted for, the AFI is more often closer than the bank in low-poverty racial/ethnic minority neighbourhoods than in high-poverty white ones. Results are driven not by the absence of banks but by the prevalence of AFIs in minority areas. Gaps appear too large to reflect simple differences in preferences.

Big data is not actually about the data. The revolution is not that there's more data available. The revolution is that we know what to do with it now.

–Gary King, a Harvard political scientist

Topic Modeling

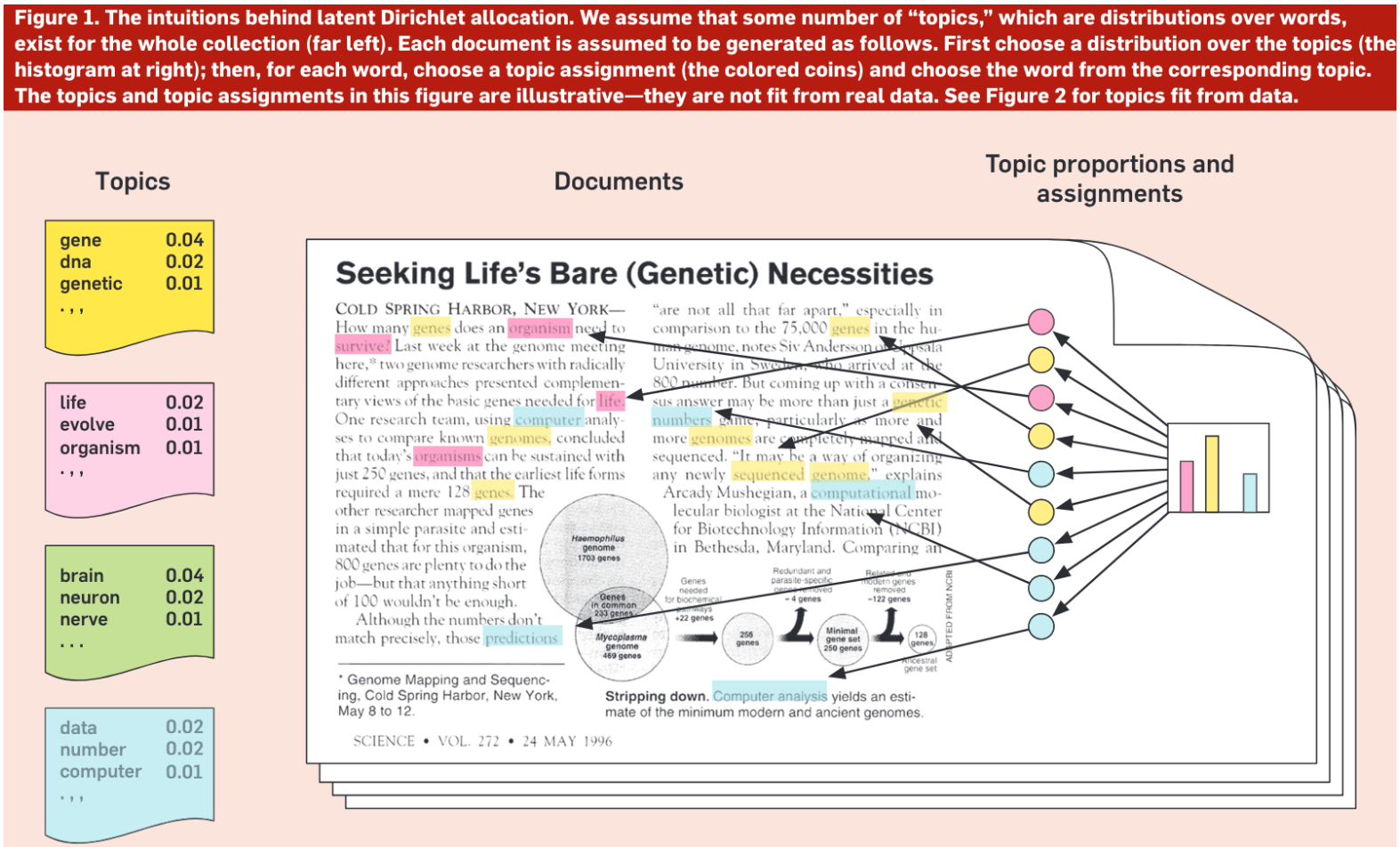


Figure: Probabilistic Topic Models

Structural Topic Models for Open-Ended Survey Responses

Margaret E. Roberts University of California, San Diego
Brandon M. Stewart Harvard University
Dustin Tingley Harvard University
Christopher Lucas Harvard University
Jetson Leder-Luis California Institute of Technology
Shana Kushner Gadarian Syracuse University
Bethany Albertson University of Texas at Austin
David G. Rand Yale University

Collection and especially analysis of open-ended survey responses are relatively rare in the discipline and when conducted are almost exclusively done through human coding. We present an alternative, semiautomated approach, the structural topic model (STM) (Roberts, Stewart, and Airoldi 2013; Roberts et al. 2013), that draws on recent developments in machine learning based analysis of textual data. A crucial contribution of the method is that it incorporates information about the document, such as the author's gender, political affiliation, and treatment assignment (if an experimental study). This article focuses on how the STM is helpful for survey researchers and experimentalists. The STM makes analyzing open-ended responses easier, more revealing, and capable of being used to estimate treatment effects. We illustrate these innovations with analysis of text from surveys and experiments.

Figure: Structural Topic Models

Sentiment Analysis

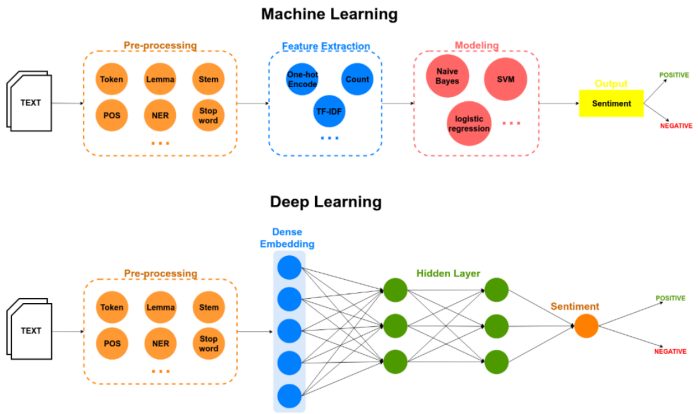


Figure: A Survey of Sentiment Analysis Methods

Firm-Level Political Risk: Measurement and Effects

Tarek A Hassan, Stephan Hollander, Laurence van Lent, Ahmed Tahoun

The Quarterly Journal of Economics, Volume 134, Issue 4, November 2019, Pages 2135–2202, https://doi.org/10.1093/qje/qjz021

Published: 26 August 2019

Cite Permissions Share ▾

Abstract

We adapt simple tools from computational linguistics to construct a new measure of political risk faced by individual U.S. firms: the share of their quarterly earnings conference calls that they devote to political risks. We validate our measure by showing that it correctly identifies calls containing extensive conversations on risks that are political in nature, that it varies intuitively over time and across sectors, and that it correlates with the firm's actions and stock market volatility in a manner that is highly indicative of political risk.

Figure: Using sentiment analysis methods to extract firm-level risks.

Word Embeddings

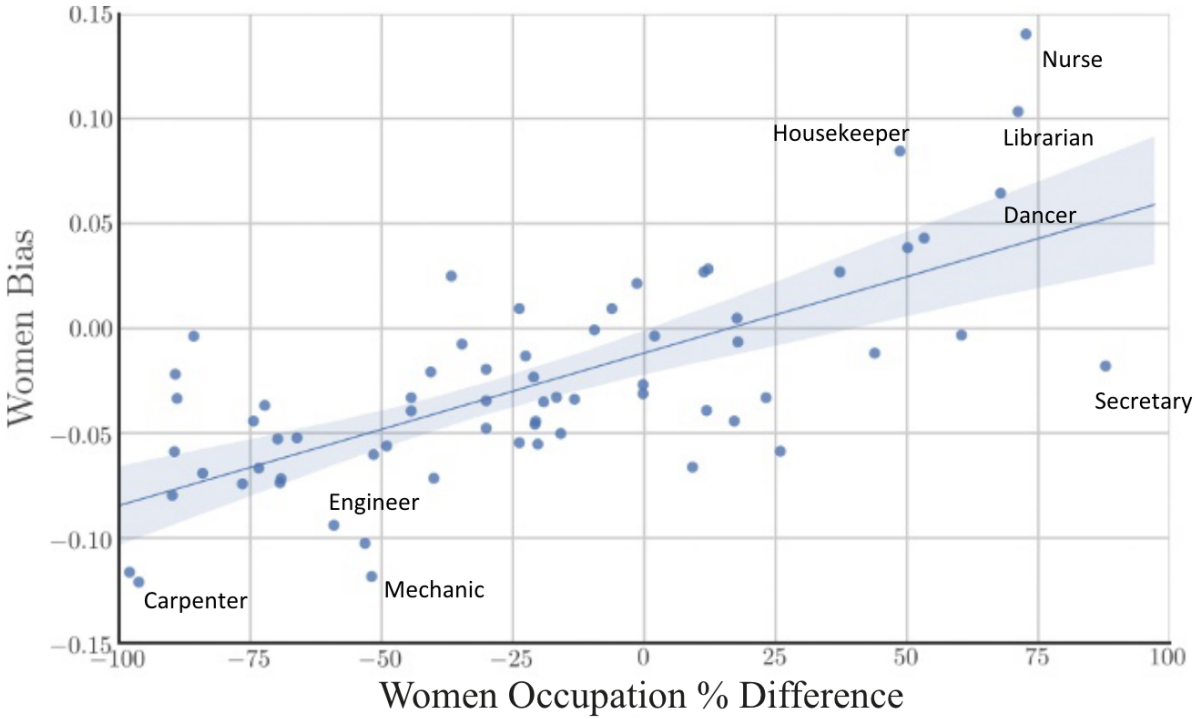


Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.

Figure: Word embeddings quantify 100 years of gender and ethnic stereotypes

Bidirectional Encoder Representations from Transformers

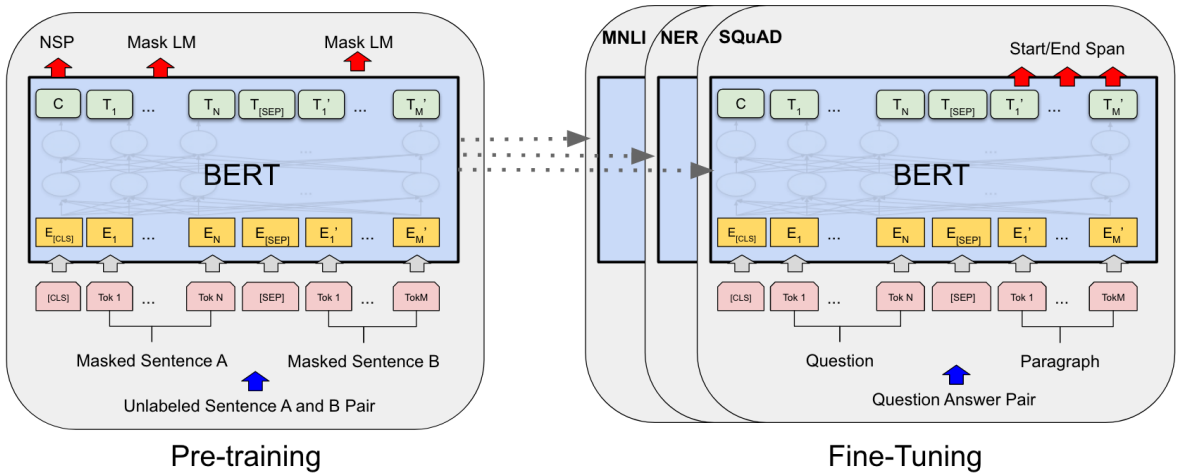


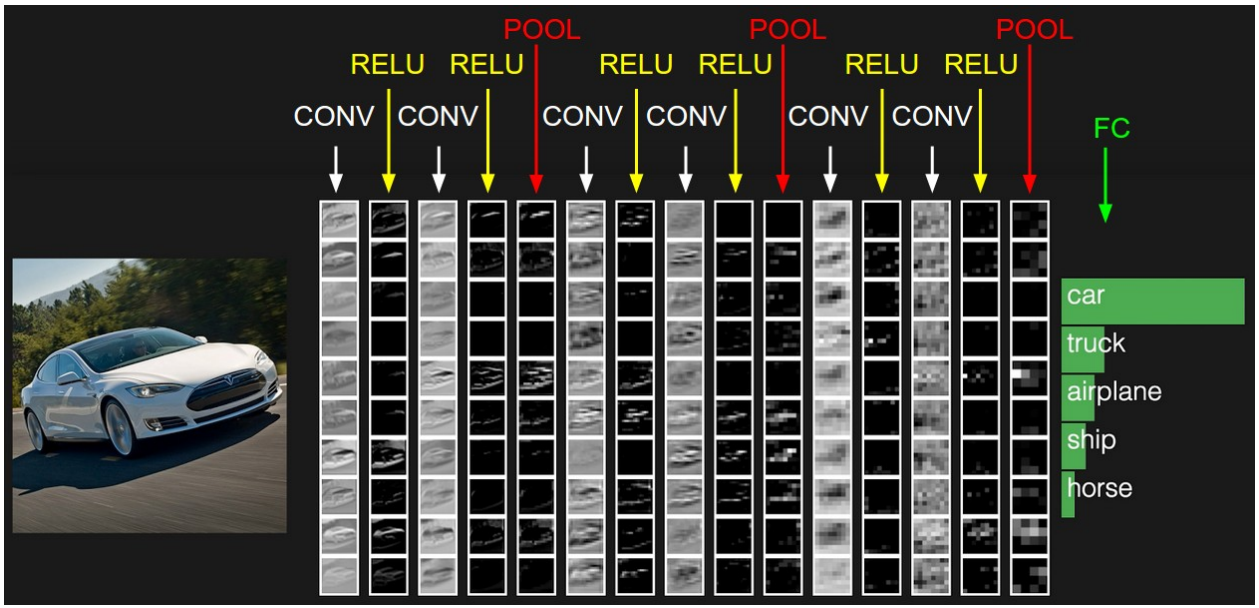
Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

Computer Vision

(a) Examples of results on FDDB

(b) Examples of results on WIDER FACE

Convolutional Neural Network



Deep Neural Networks

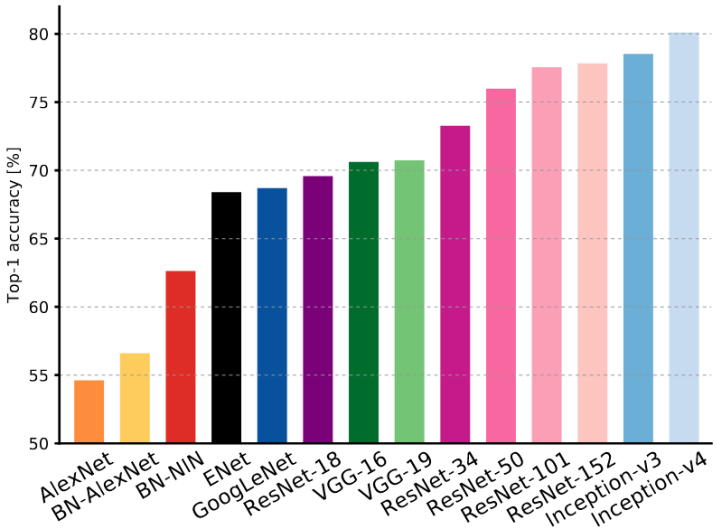


Figure 1: **Top1 *vs.* network.** Single-crop top-1 validation accuracies for top scoring single-model architectures. We introduce with this chart our choice of colour scheme, which will be used throughout this publication to distinguish effectively different architectures and their correspondent authors. Notice that networks of the same group share the same hue, for example ResNet are all variations of pink.

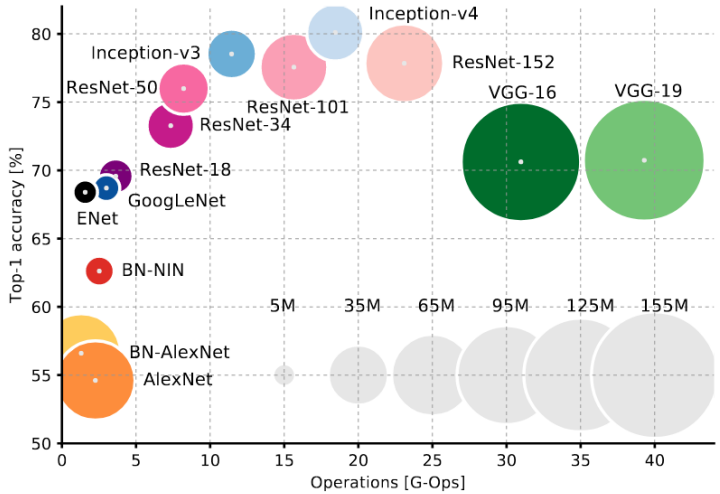


Figure 2: **Top1 *vs.* operations, size $\propto$ parameters.** Top-1 one-crop accuracy versus amount of operations required for a single forward pass. The size of the blobs is proportional to the number of network parameters; a legend is reported in the bottom right corner, spanning from 5×10^6 to 155×10^6 params. Both these figures share the same y-axis, and the grey dots highlight the centre of the blobs.

Swin Transformers

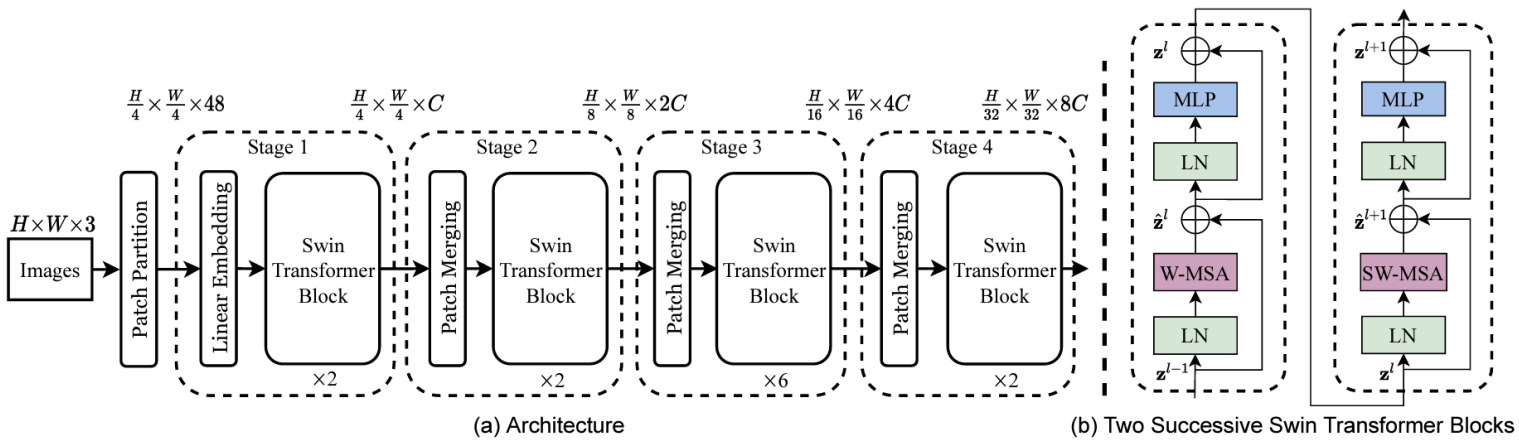
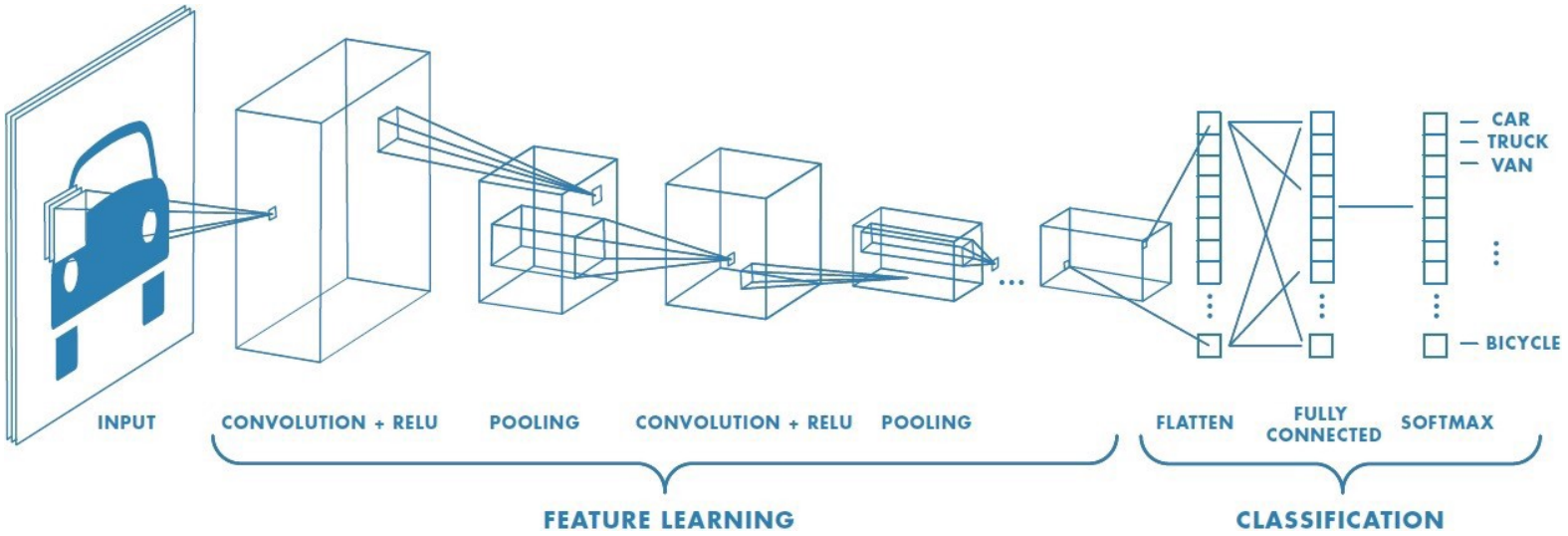


Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.


TensorFlow
K
Keras
PYTORCH

Transfer Learning



RESEARCH ARTICLES

ECONOMICS

Combining satellite imagery and machine learning to predict poverty

Neal Jean,[1,2*] Marshall Burke,[3,4,5*†] Michael Xie,[1] W. Matthew Davis,[4] David B. Lobell,[3,4] Stefano Ermon[1]

Reliable data on economic livelihoods remain scarce in the developing world, hampering efforts to study these outcomes and to design policies that improve them. Here we demonstrate an accurate, inexpensive, and scalable method for estimating consumption expenditure and asset wealth from high-resolution satellite imagery. Using survey and satellite data from five African countries—Nigeria, Tanzania, Uganda, Malawi, and Rwanda—we show how a convolutional neural network can be trained to identify image features that can explain up to 75% of the variation in local-level economic outcomes. Our method, which requires only publicly available data, could transform efforts to track and target poverty in developing countries. It also demonstrates how powerful machine learning techniques can be applied in a setting with limited training data, suggesting broad potential application across many scientific domains.

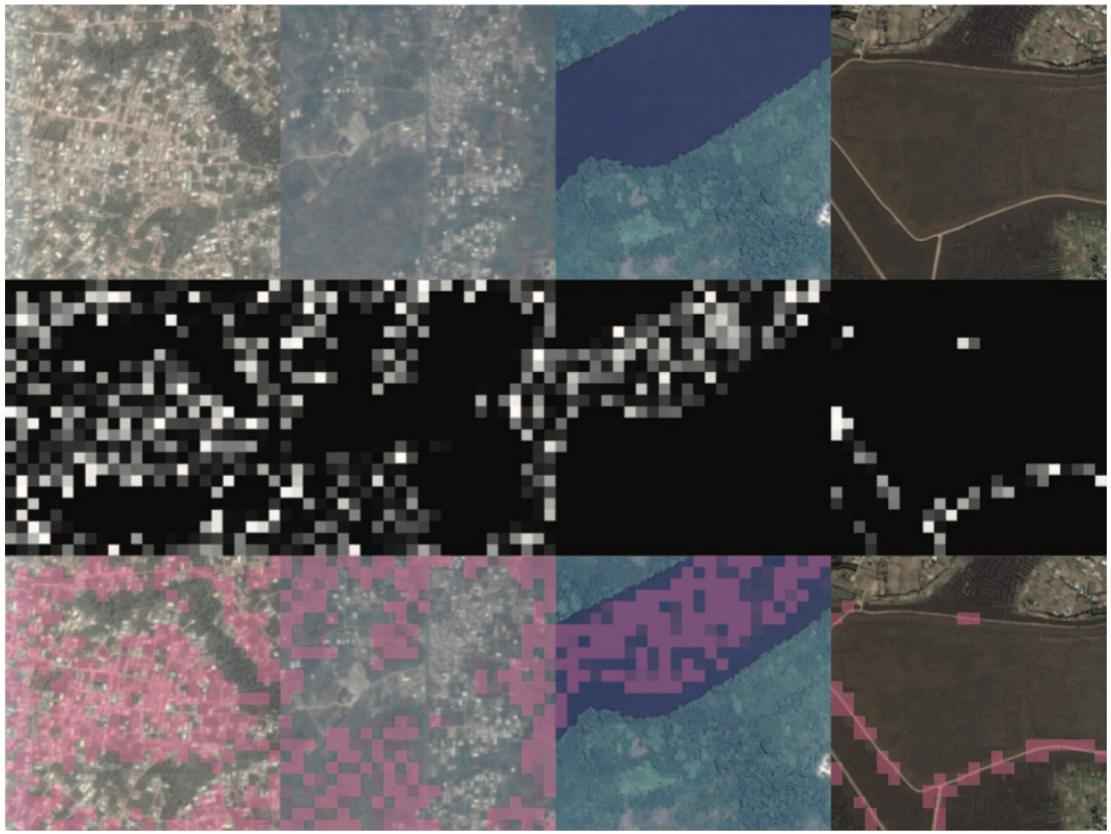


Fig. 2. Visualization of features. By column: Four different convolutional filters (which identify, from left to right, features corresponding to urban areas, nonurban areas, water, and roads) in the convolutional neural network model used for extracting features. Each filter "highlights" the parts of the image that activate it, shown in pink. By row: Original daytime satellite images from Google Static Maps, filter activation maps, and overlay of activation maps onto original images

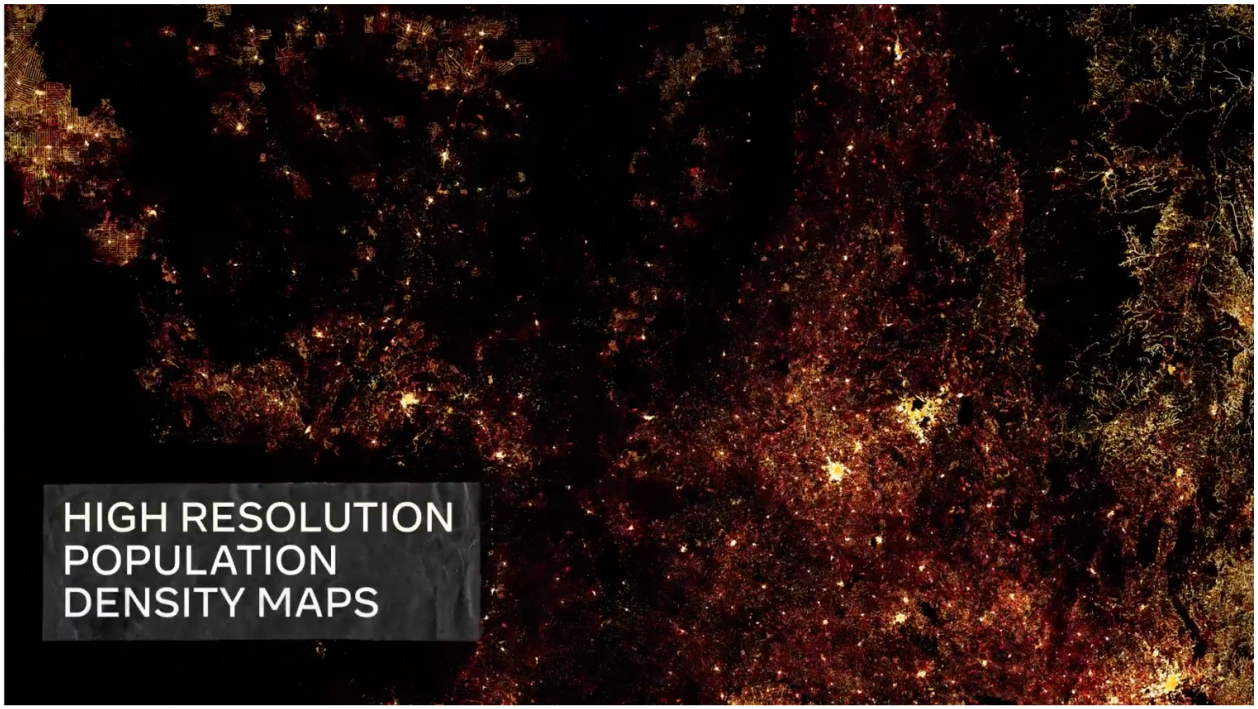


Figure: Facebook Population Density Map

Speech Analysis

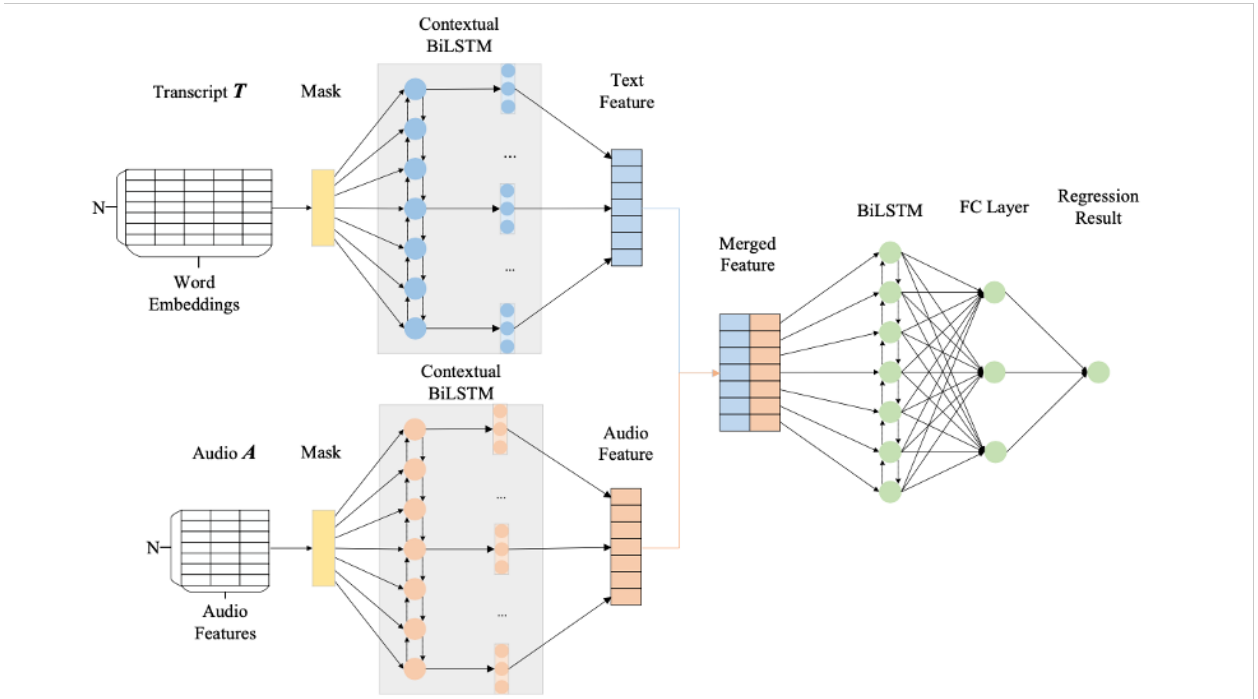


Figure 1: The proposed Multimodal Deep Regression Model (MDRM). The inputs to the model is a company's conference call audio file with correpsonding transcript. Each conference call consists of N sentences. The output variable is a numerical value, i.e., the company's stock price volatility following the conference call.

Figure: Predict Stock

Motion Detection

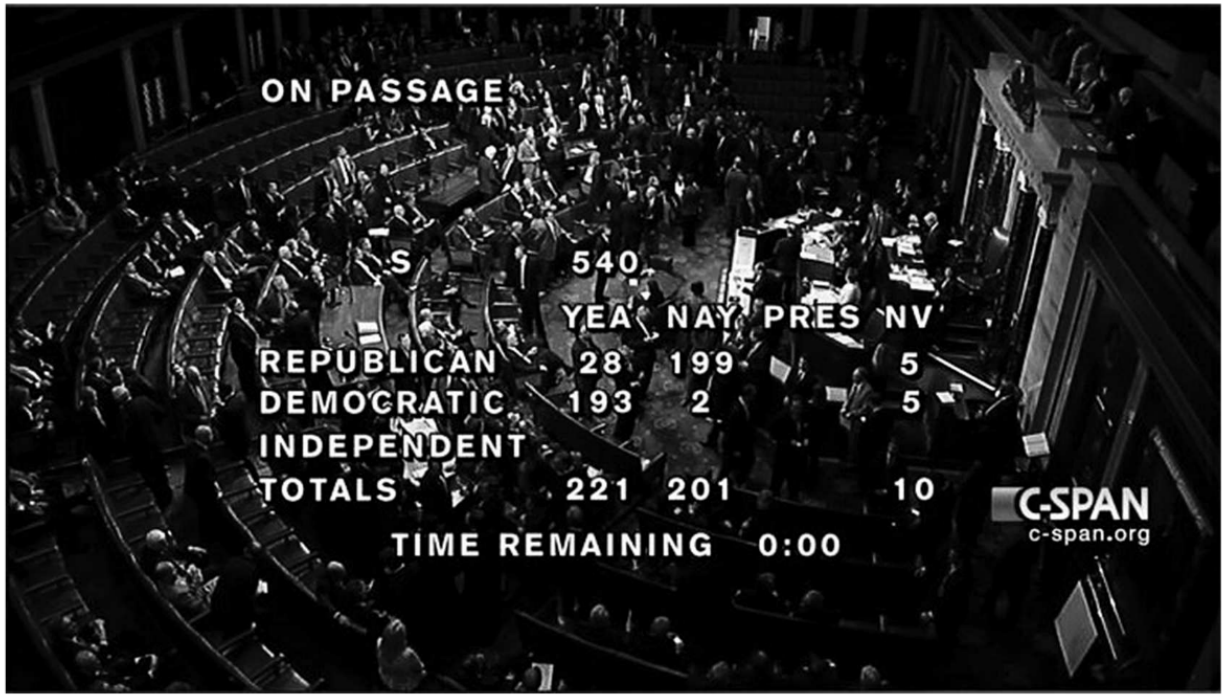


Figure 1. Overhead shot of members of Congress mingling after a roll-call vote. Not only does this shot show all of the social interactions that take place after a floor vote, but it is a quintessential part of C-SPAN coverage. All the analyses presented below consider videos similar to the frame shown here.

Figure: Using Motion Detection to Measure Social Polarization in Congress

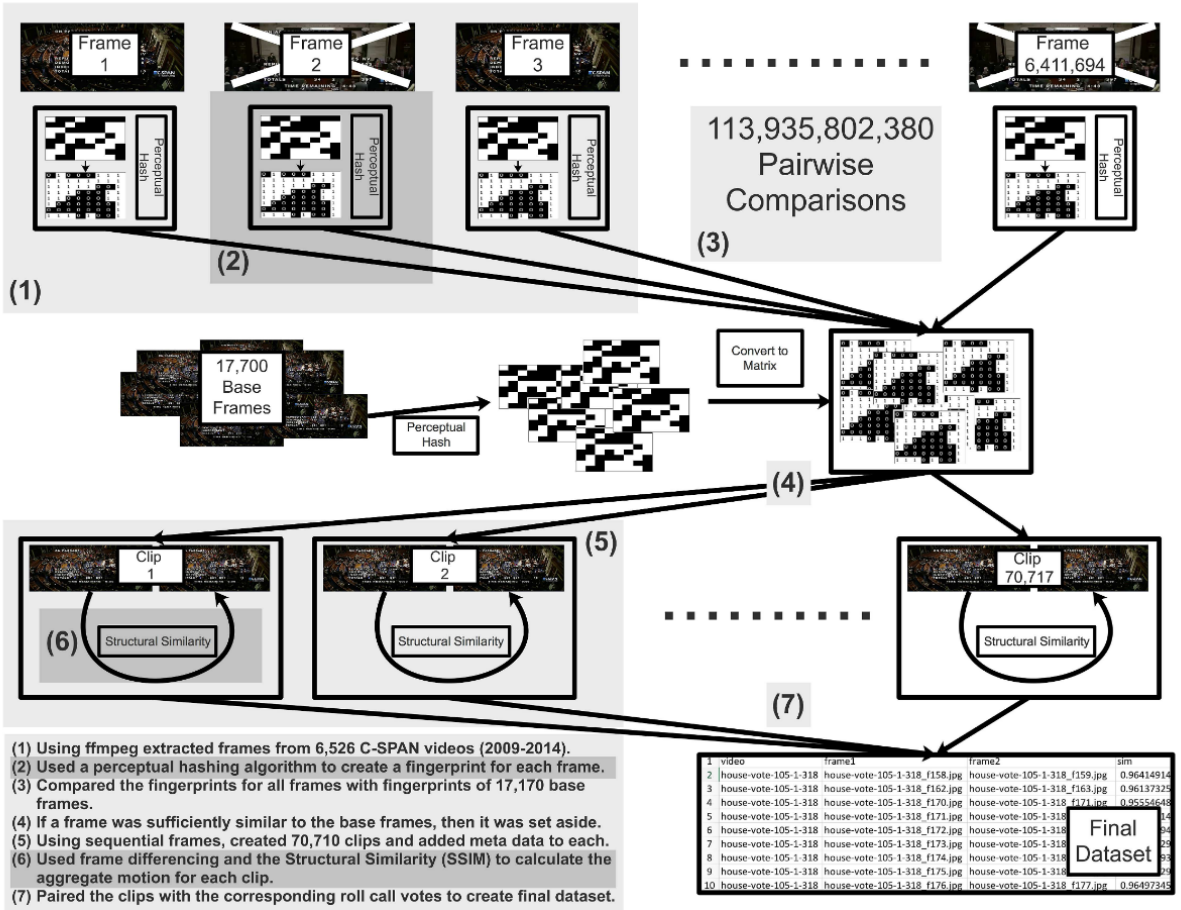


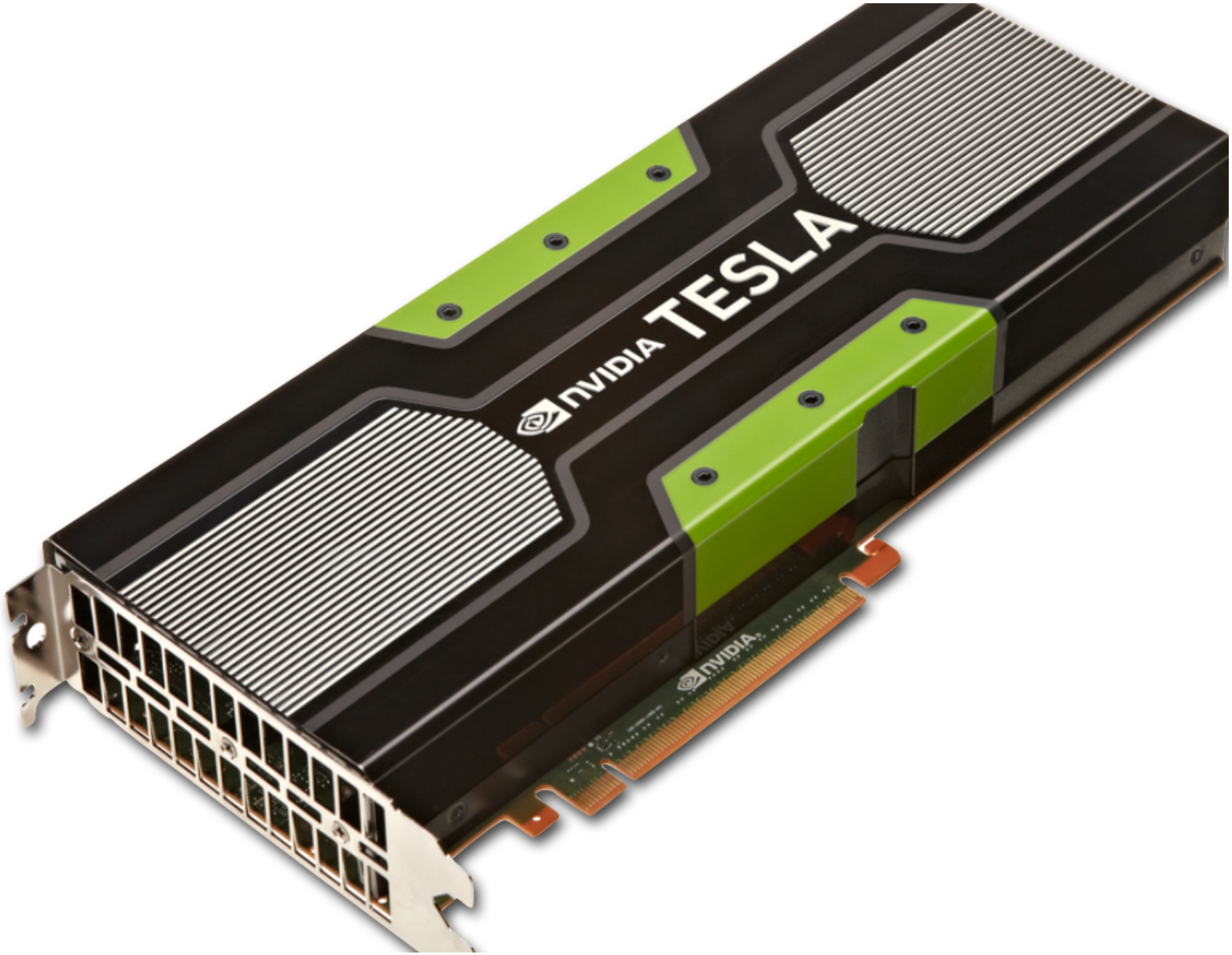
Figure 2. Figure explaining motion detection technique and how the overhead shots were extracted from the C-SPAN videos. Please see Section S1 in the Supplemental Information for more details about how the overhead shots were extracted and video motion was detected.

Figure: Using Motion Detection to Measure Social Polarization in the U.S. House of Representatives

The Access to Cloud Computing

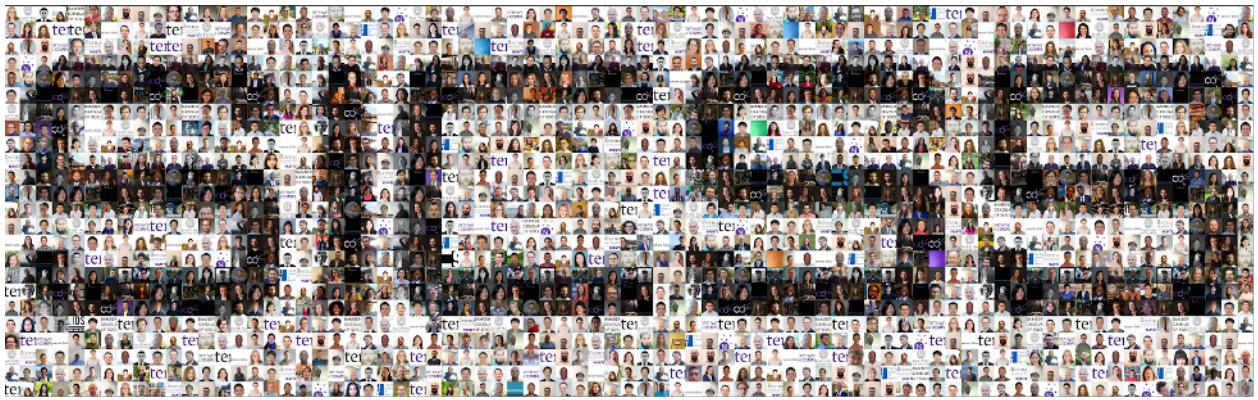
Figure: Google Cloud AI

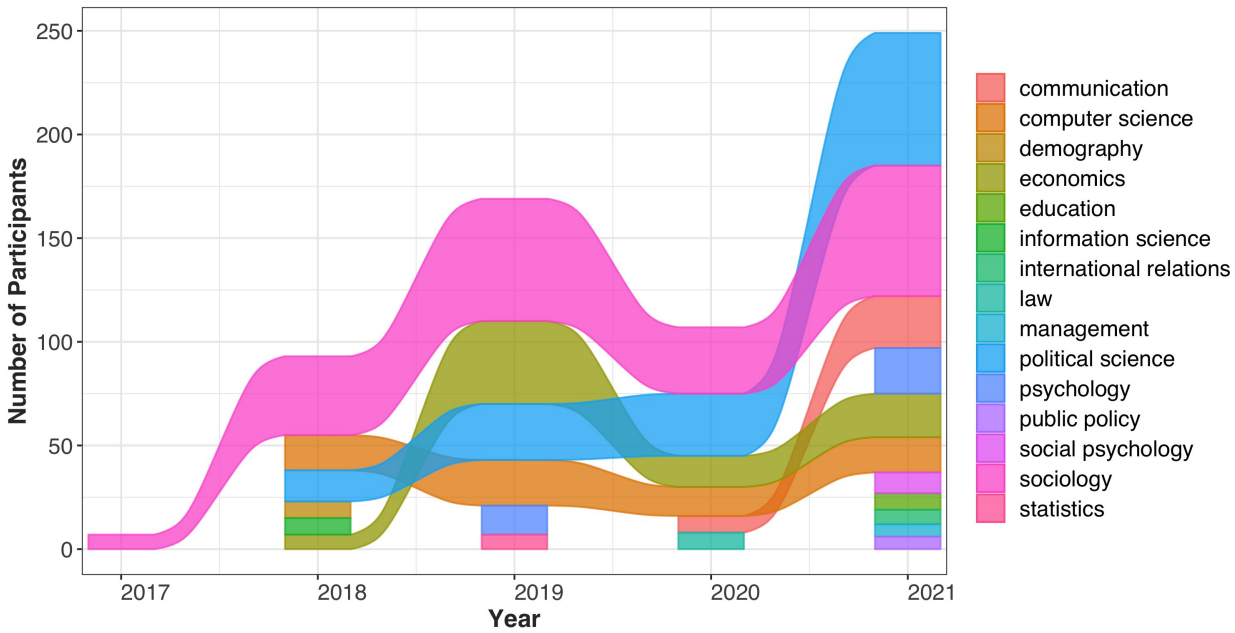
Figure: Seawulf HPC


NVIDIA TESLA
NVIDIA


Use the GeForce, Luke

Summer Institute for Computational Social Science




250
200
150
100
50
0
Number of Participants
2017
2018
2019
2020
2021
Year
communication
computer science
demography
economics
education
information science
international relations
law
management
political science
psychology
public policy
social psychology
sociology
statistics

Knowledge Lab

Yale Institute for Network Science

Human Nature Lab

The Human Nature Lab, directed by Nicholas Christakis, sits within the Yale Institute for Network Science and is currently focused on the relationship between social networks and health.

Inference, Information and Decision Systems Group

The Inference, Information and Decision (I.I.D.) Systems Group is led by Amin Karbasi. The research in I.I.D. is at the intersection of learning theory, large-scale networks, and optimum information processing. We devise new algorithms, build models, analyze the behavior of large and complex...

Polarization lab

The CSS Journals

Funding Opportunities

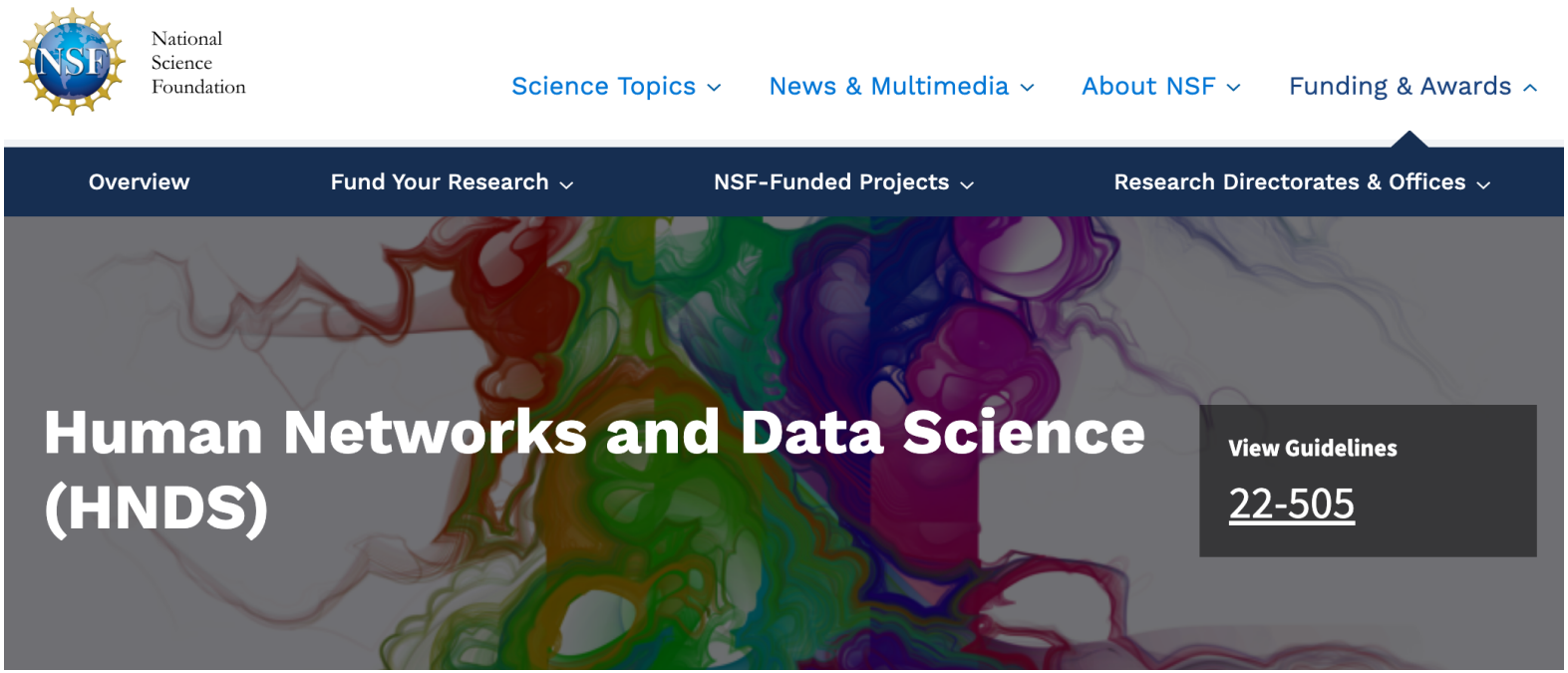
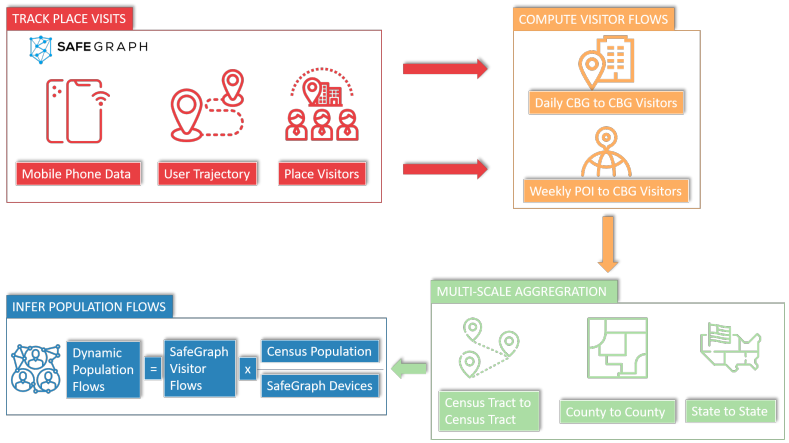


Figure: NSF Human Networks and Data Science Program

Table of Contents

The Collaboration Between Industry and Academia

Population mobility data



FACEBOOK Data for Good

Maps Surveys Insights

Social Connectedness Index

Old Questions With New Data

Racial segregation in the U.S.

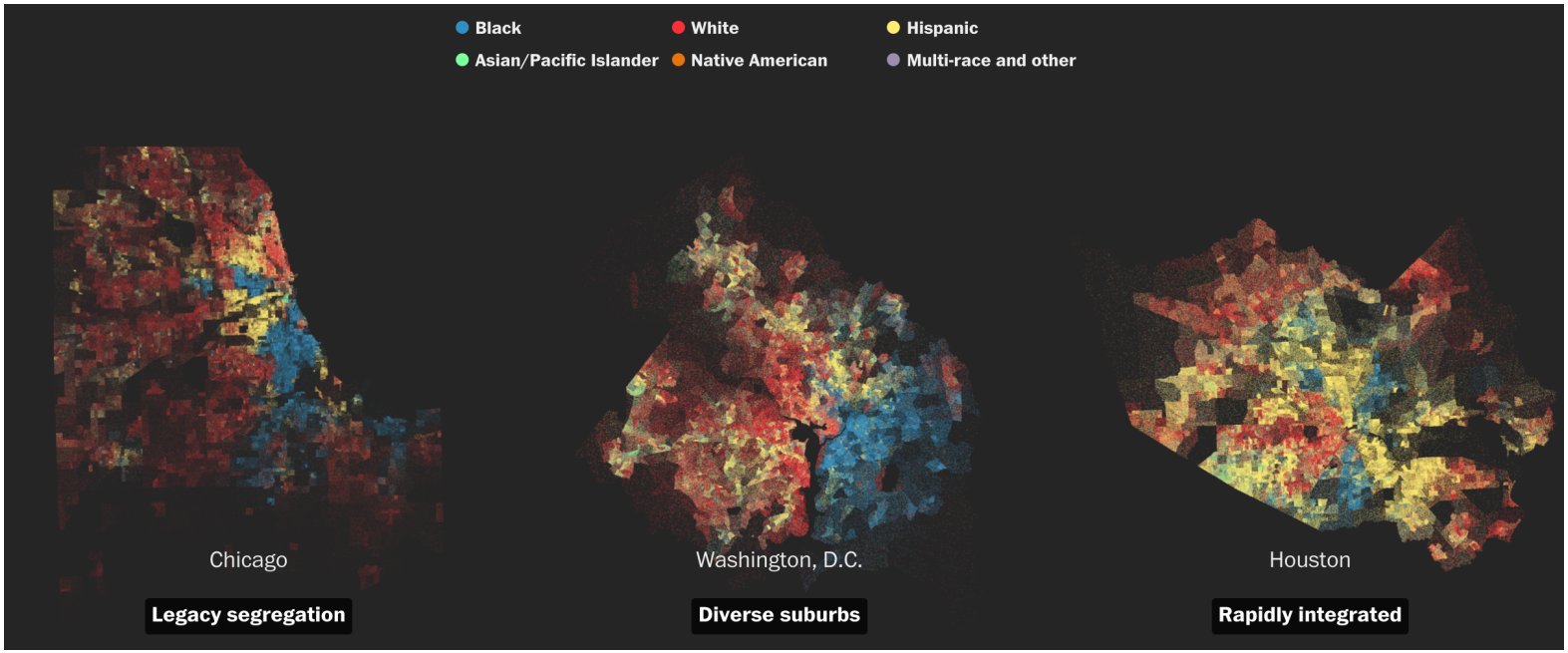


Figure: https://www.washingtonpost.com

Using 180 Million Voter Records to Study Partisan Segregation

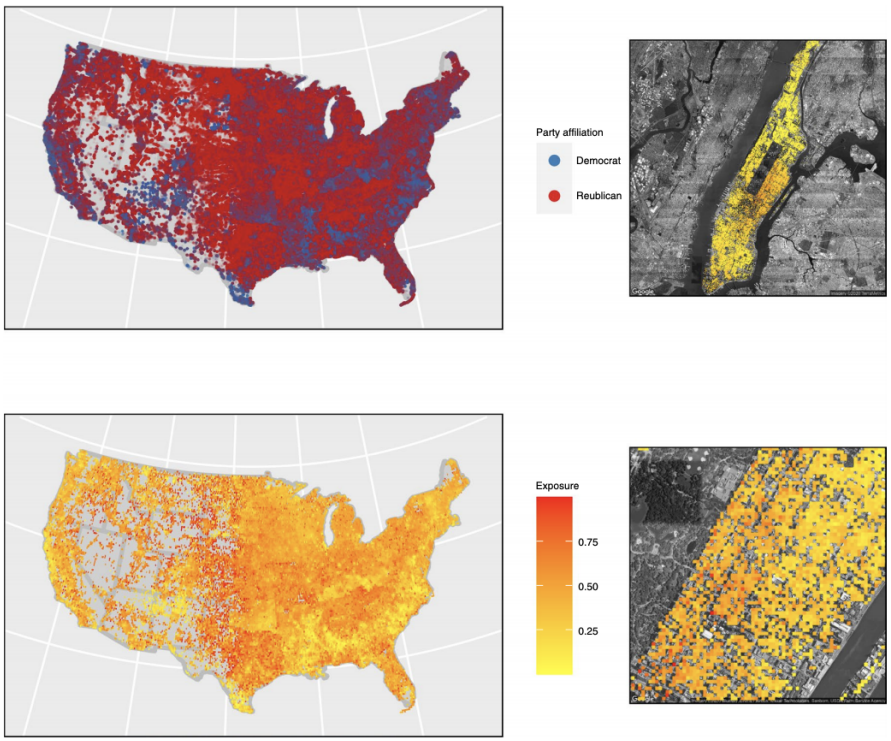


Fig. 2 | Measuring spatial exposure across increasingly small geographies. The exact residential location of every Democrat and Republican in the United States (n = 180,660,202, top left) can be used to measure each Democrat's spatial exposure to Republicans, and this can be averaged across arbitrarily small grid cells for display purposes (1,000 × 1,000 grid, bottom left). Exposure can be examined across any resolution: markedly different residential exposure to Republicans can be seen in Manhattan, NY (500 × 500 grid, top right), with Democrats on the northern and southern extremes of the island having almost no residential exposure to Republicans, whereas Democrats on the Upper East Side (the neighborhood immediately to the right of the lower section of Central Park, which is the long rectangle with no voters in it located in the center of the island) have exposure as high as 0.5 due to the clustering of Republicans in this area. A magnified view of the Upper East Side of Manhattan (75 x 75 grid, bottom right) shows the clustering of Republicans along Central Park and thus Democrats' decreasing exposure to Republicans moving towards the northeast. Map data (righthand figures): Google, TerraMetrics.

Figure: The Measurement of Partisan Sorting of 180 Million Voters

Urban mobility and neighborhood isolation in America's 50 largest cities

Qi Wang[a,1], Nolan Edward Phillips[b], Mario L. Small[b], and Robert J. Sampson[b]

[a]Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115; and [b]Department of Sociology, Harvard University, Cambridge, MA 02138

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved June 6, 2018 (received for review February 10, 2018)

Influential research on the negative effects of living in a disadvantaged neighborhood assumes that its residents are socially isolated from nonpoor or "mainstream" neighborhoods, but the extent and nature of such isolation remain in question. We develop a test of neighborhood isolation that improves on static measures derived from commonly used census reports by leveraging fine-grained dynamic data on the everyday movement of residents in America's 50 largest cities. We analyze 650 million geocoded Twitter messages to estimate the home locations and travel patterns of almost 400,000 residents over 18 mo. We find surprisingly high consistency across neighborhoods of different race and income characteristics in the average travel distance (radius) and number of neighborhoods traveled to (spread) in the metropolitan region; however, we uncover notable differences in the composition of the neighborhoods visited. Residents of primarily black and Hispanic neighborhoods—whether poor or not—are far less exposed to either nonpoor or white middle-class neighborhoods than residents of primarily white neighborhoods. These large racial differences are notable given recent declines in segregation and the increasing diversity of American cities. We also find that white poor neighborhoods are substantially isolated from nonpoor white neighborhoods. The results suggest that even though residents of disadvantaged neighborhoods travel far and wide, their relative isolation and segregation persist.

commuting ties, which focuses on adults' travel between home and work (12, 16). However, commuting does not include neighborhoods experienced through leisure, errand activities, or visits to friends and family, all of which affect the extent of isolation. Second, several studies have used travel diaries collected by volunteers (15, 17, 18). While such methods produce rich data on the multiple locations visited by respondents, they are typically limited to one city and constrained by sample size limitations, given the onerous demands placed on study participants. These constraints are especially important given potential differences between cities. For example, travel patterns in cities with expansive public transit systems (e.g., New York City or Chicago) may differ from those in cities where driving is the primary mode of transportation (e.g., Houston or Los Angeles). These differences may also exacerbate inequalities in neighborhood isolation across race and class lines. Third, a few studies have examined the differences in mobility patterns among different social groups (19, 20), as well as their geographical interactions (21), using geolocation records from cell phones and social media platforms. However, only a few of these studies have examined race or class differences in mobility and none have done so across a large sample of cities.

Traditional studies that examined neighborhood isolation using surveys, field experiments, or tax records do not track everyday mobility for large populations with sufficient detail for

Experienced Segregation in the U.S.

Estimating experienced racial segregation in US cities using large-scale GPS data

Susan Athey[a,b,1,2], Billy Ferguson[c,1], Matthew Gentzkow[a,b,1], and Tobias Schmidt[1]

[a]Department of Economics, Stanford University, Stanford, CA 94305; [b]National Bureau of Economic Research, Cambridge, MA 02138; and [c]Stanford Graduate School of Business, Stanford University, Stanford, CA 94305

Contributed by Susan Athey, May 6, 2021 (sent for review December 19, 2020; reviewed by Keith Chen and Jessie Handbury)

We estimate a measure of segregation, experienced isolation, that captures individuals' exposure to diverse others in the places they visit over the course of their days. Using Global Positioning System (GPS) data collected from smartphones, we measure experienced isolation by race. We find that the isolation individuals experience is substantially lower than standard residential isolation measures would suggest but that experienced isolation and residential isolation are highly correlated across cities. Experienced isolation is lower relative to residential isolation in denser, wealthier, more educated cities with high levels of public transit use and is also negatively correlated with income mobility.

racial segregation | isolation | mobility

of 2017. The data are obtained from a company that aggregates anonymous pings from a range of smartphone apps. We observe each device's home location as well as the location of every ping by the device recorded in the data. We map these locations to a grid of geographic units ∼500 ft × 500 ft, known as geohash7s. The sample of individuals is not random but is reasonably close to representative along a number of dimensions, and it has sufficient coverage that we can correct for deviations from representativeness using sample weights. We use the movement patterns we observe to compute experienced racial isolation.

Because we do not observe an individual's race directly, we define the two types whose segregation we study as individuals with homes in majority White geohash7s and individuals with

Using Population Mobility Data to Measure Black-White Residential Segregation in the COVID-19 Pandemic

Yongjun Zhang[a,1]

[a]Department of Sociology and Institute for Advanced Computational Science, State University of New York at Stony Brook, 100 Nicolls Road, Stony Brook, New York 11794; ORCiD: https://orcid.org/0000-0002-8265-925X

June 23, 2021

Racial and ethnic residential segregation has long been the central focus of stratification and inequality research, and it is a linchpin of racial stratification in the U.S. Sociologists and demographers have developed a series of spacial or aspatial measures to capture distinct aspects of segregation. Although the recent development of segregation measures, for instance, spacial exposure, accounts for spacial proximity among different groups, it is static and ignores the social connectedness dimension. This article uses population mobility across communities to correct the potential bias in spacial segregation measures. As population mobility is highly racially segregated, we modify the conventional spatial isolation index by adding an extra layer of social connectedness between communities to create a socially and spatially weighted segregation measure. We then use this spatial and social segregation measure to quantify the level of blacks' isolation with whites in the local neighboring communities.

residential segregation | population mobility | big data | COVID19

proximity. But this assumes that different social groups living in proximate geographic areas have a greater likelihood to interact with each other. This underlying assumption is not necessarily true, given that two proximate geographic units might not be tightly connected due to various reasons (e.g., rivers, highways).

To address these shortcomings, Echenique and Fryer took a social network approach to develop a measure of segregation based on social interactions (16). The rationale is that an individual is more segregated when interacting more with other segregated agents in a community. They also highlight that the measure of segregation should disaggregate to the individual-level. Yet, their measure receives less attention due to lack of large-scale social interaction data across different social groups. A few notable exceptions are Wang, Candipan, and their collegue's seminal work on using geocoded twitter users'

Relational Segregation in the U.S. Metro Areas

Based on Facebook and Safegraph Data

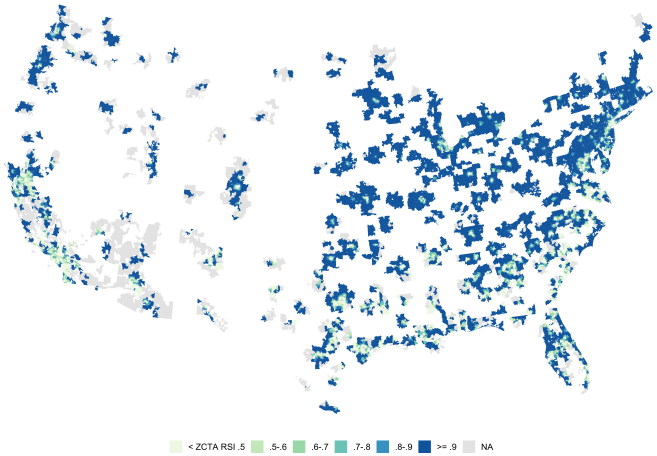


Figure: Facebook SCI

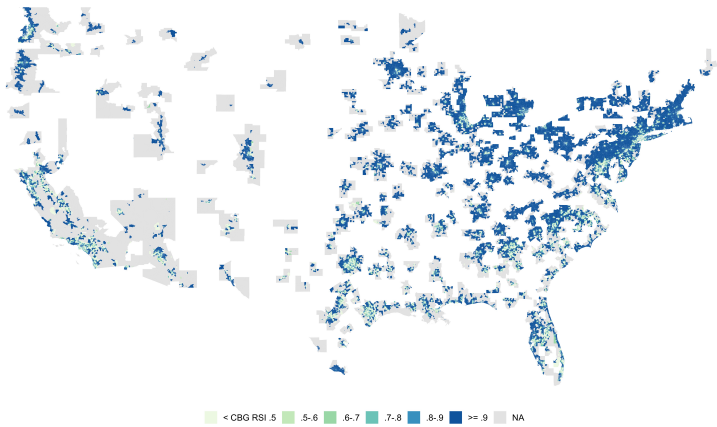


Figure: Safegraph mobility

Interdisciplinary Collaboration

Working with scholars from different disciplines and industries

Stony Brook University | Institute for Advanced Computational Science

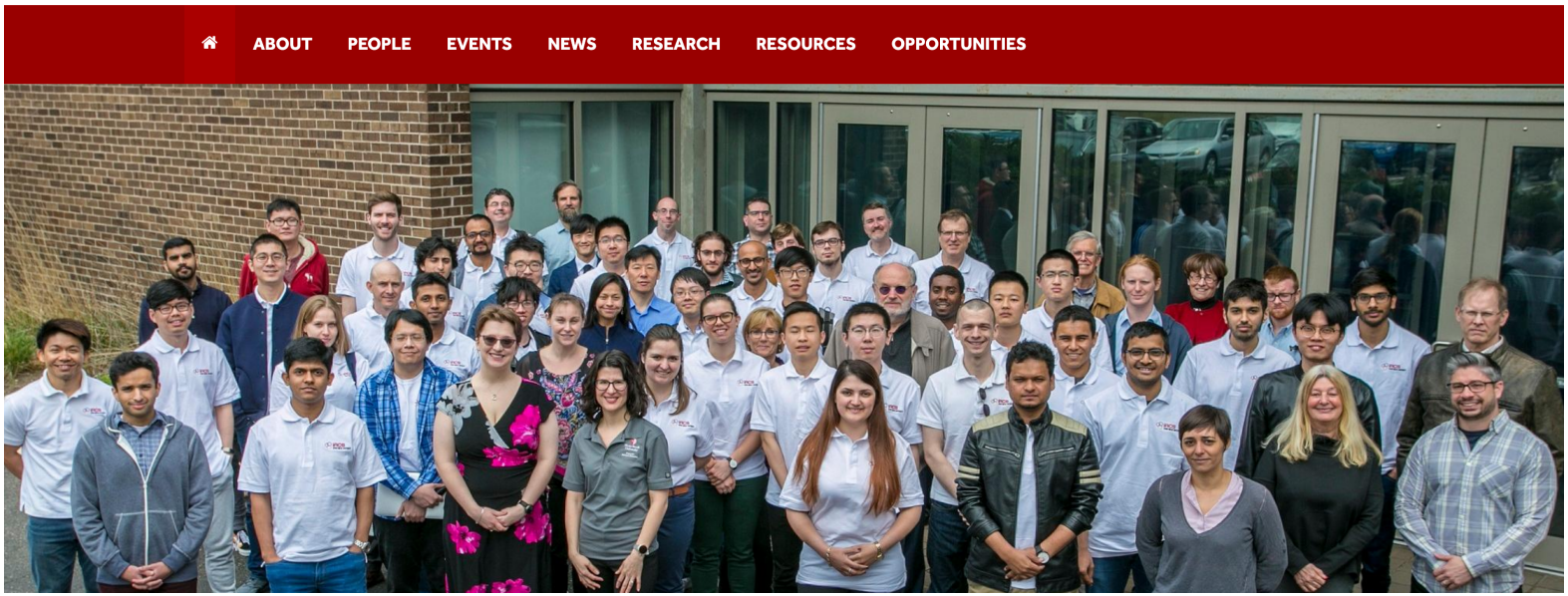


Table of Contents

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]**

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

Figure: Cambridge Analytica Scandal

Research Ethics in Digital Age

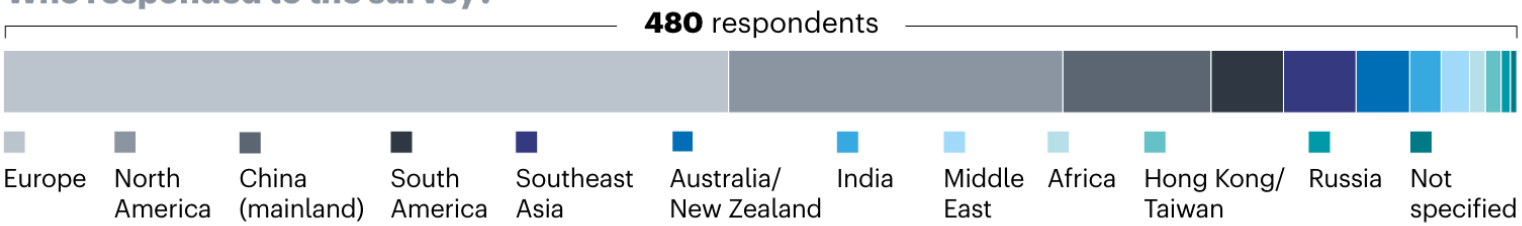
Data without Consent?

Figure: MegaFace

FACIAL RECOGNITION: A SURVEY ON ETHICS

Nature surveyed* nearly 500 researchers who work in facial recognition, computer vision and artificial intelligence about ethical issues relating to facial-recognition research. They are split on whether certain types of this research are ethically problematic and what should be done about concerns.

Who responded to the survey?



Restrictions on image use

Question: Researchers use large data sets of images of people's faces — often scraped from the Internet — to train and test facial-recognition algorithms. What kind of permissions do researchers need to use such images?

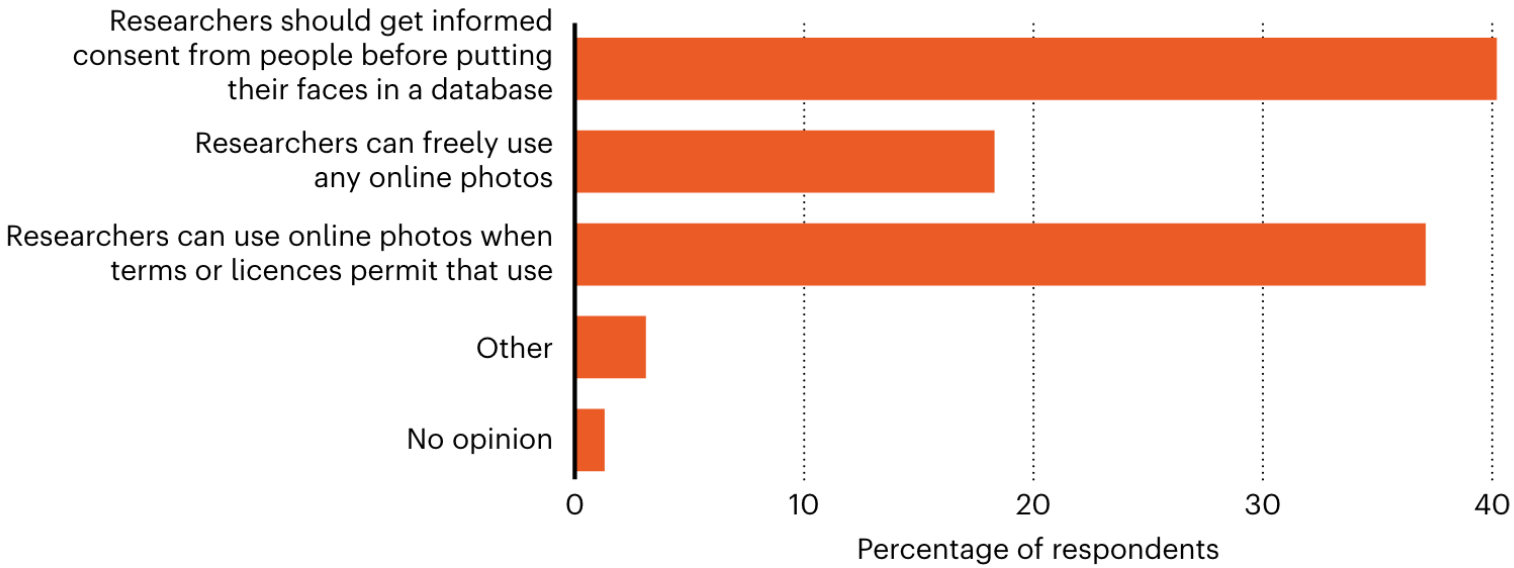


Figure: Nature Surveying 480 researchers

Research Ethics in Digital Age

Data with legal risks?

JONATHAN H. BLAVIN (State Bar No. 230269)
jonathan.blavin@mto.com
NICHOLAS D. FRAM (State Bar No. 288293)
nicholas.fram@mto.com
MUNGER, TOLLES & OLSON LLP
560 Mission Street
Twenty-Seventh Floor
San Francisco, California 94105-2907
Telephone: (415) 512-4000
Facsimile: (415) 512-4077

Attorneys for LinkedIn Corporation

UNITED STATES DISTRICT COURT

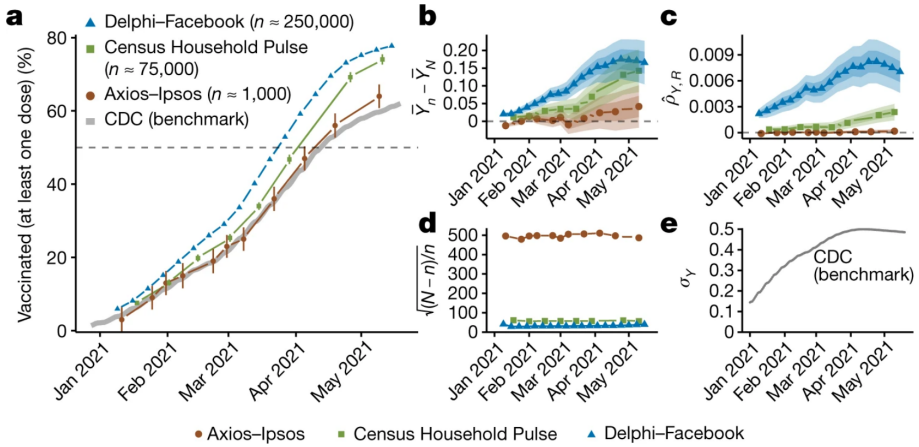
NORTHERN DISTRICT OF CALIFORNIA

LinkedIn Corporation, Plaintiff, vs. Does, 1 through 100 inclusive, Defendants.	Case No. 5:16-cv-4463 **Complaint For:** **(1) VIOLATION OF THE COMPUTER FRAUD AND ABUSE ACT, 18 U.S.C. §§ 1030 ET SEQ.;** **(2) VIOLATION OF CALIFORNIA PENAL CODE §§ 502 ET SEQ.;** **(3) VIOLATION OF THE DIGITAL MILLENNIUM COPYRIGHT ACT, 17 U.S.C. §§ 1201 ET SEQ.;** **(4) BREACH OF CONTRACT;** **(5) TRESPASS; AND** **(6) MISAPPROPRIATION** **DEMAND FOR JURY TRIAL**

Figure: Linkedin vs. Hiq Labs

Garbage In, Garbage Out

1% survey vs. 80% pop-level big data, which one is better?



a, Estimates of vaccine uptake for US adults in 2021 compared to CDC benchmark data, plotted by the end date of each survey wave. Points indicate each study's weighted estimate of first-dose vaccine uptake, and intervals are 95% confidence intervals using reported standard errors and design effects. Delphi–Facebook has $n = 4{,}525{,}633$ across 19 waves, Census Household Pulse has $n = 606{,}615$ across 8 waves and Axios–Ipsos has $n = 11{,}421$ across 11 waves. Delphi–Facebook's confidence intervals are too small to be visible. **b**, Total error $\bar{Y}_n - \bar{Y}_N$. **c**, Data defect correlation $\hat{\rho}_{Y,R}$. **d**, Data scarcity $\sqrt{(N-n)/n}$. **e**, Inherent problem difficulty σ_Y. Shaded bands represent scenarios of ±5% (darker) and ±10% (lighter) imprecision in the CDC benchmark relative to reported values (points). **b**–**e** comprise the decomposition in equation (1).

Figure: Errors in estimates of vaccine uptake. From: Unrepresentative big surveys significantly overestimated US vaccine uptake

Garbage In, Garbage Out

Data with noises?

Figure: Social-media bots. Credit: Omer Messinger

Garbage In, Garbage Out

Measurement with biases?

Perspective

Measuring algorithmically infused societies

https://doi.org/10.1038/s41586-021-03666-1

Received: 5 March 2021

Accepted: 21 May 2021

Published online: 30 June 2021

Check for updates

Claudia Wagner[1,2,3]✉, Markus Strohmaier[1,2,3], Alexandra Olteanu[4,5], Emre Kıcıman[6], Noshir Contractor[7] & Tina Eliassi-Rad[8]

It has been the historic responsibility of the social sciences to investigate human societies. Fulfilling this responsibility requires social theories, measurement models and social data. Most existing theories and measurement models in the social sciences were not developed with the deep societal reach of algorithms in mind. The emergence of 'algorithmically infused societies'—societies whose very fabric is co-shaped by algorithmic and human behaviour—raises three key challenges: the insufficient quality of measurements, the complex consequences of (mis)measurements, and the limits of existing social theories. Here we argue that tackling these challenges requires new social theories that account for the impact of algorithmic systems on social realities. To develop such theories, we need new methodologies for integrating data and measurements into theory construction. Given the scale at which measurements can be applied, we believe measurement models should be trustworthy, auditable and just. To achieve this, the development of measurements should be transparent and participatory, and include mechanisms to ensure measurement quality and identify possible harms. We argue that computational social scientists should rethink what aspects of algorithmically infused societies should be measured, how they should be measured, and the consequences of doing so.

Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew J. Salganik[a,1], Ian Lundberg[a], Alexander T. Kindel[a], Caitlin E. Ahearn[b], Khaled Al-Ghoneim[c], Abdullah Almaatouq[d,e], Drew M. Altschul[f], Jennie E. Brand[b,g], Nicole Bohme Carnegie[h], Ryan James Compton[i], Debanjan Datta[j], Thomas Davidson[k], Anna Filippova[l], Connor Gilroy[m], Brian J. Goode[n], Eaman Jahani[o], Ridhi Kashyap[p,q,r], Antje Kirchner[s], Stephen McKay[t], Allison C. Morgan[u], Alex Pentland[e], Kivan Polimis[v], Louis Raes[w], Daniel E. Rigobon[x], Claudia V. Roberts[y], Diana M. Stanescu[z], Yoshihiko Suhara[e], Adaner Usmani[aa], Erik H. Wang[z], Muna Adem[bb], Abdulla Alhajri[cc], Bedoor AlShebli[dd], Redwane Amin[ee], Ryan B. Amos[y], Lisa P. Argyle[ff], Livia Baer-Bositis[gg], Moritz Büchi[hh], Bo-Ryehn Chung[ii], William Eggert[jj], Gregory Faletto[kk], Zhilin Fan[ll], Jeremy Freese[gg], Tejomay Gadgil[mm], Josh Gagné[gg], Yue Gao[nn], Andrew Halpern-Manners[bb], Sonia P. Hashim[y], Sonia Hausen[gg], Guanhua He[oo], Kimberly Higuera[gg], Bernie Hogan[pp], Ilana M. Horwitz[qq], Lisa M. Hummel[gg], Naman Jain[x], Kun Jin[rr], David Jurgens[ss], Patrick Kaminski[bb,tt], Areg Karapetyan[uu,vv], E. H. Kim[gg], Ben Leizman[y], Naijia Liu[z], Malte Möser[y], Andrew E. Mack[z], Mayank Mahajan[y], Noah Mandell[ww], Helge Marahrens[bb], Diana Mercado-Garcia[qq], Viola Mocz[xx], Katariina Mueller-Gastell[gg], Ahmed Musse[yy], Qiankun Niu[ee], William Nowak[zz], Hamidreza Omidvar[aaa], Andrew Or[y], Karen Ouyang[y], Katy M. Pinto[bbb], Ethan Porter[ccc], Kristin E. Porter[ddd], Crystal Qian[y], Tamkinat Rauf[gg], Anahit Sargsyan[eee], Thomas Schaffner[y], Landon Schnabel[gg], Bryan Schonfeld[z], Ben Sender[fff], Jonathan D. Tang[y], Emma Tsurkov[gg], Austin van Loon[gg], Onur Varol[ggg,hhh], Xiafei Wang[iii], Zhi Wang[hhh,jjj], Julia Wang[y], Flora Wang[fff], Samantha Weissman[y], Kirstie Whitaker[kkk,lll], Maria K. Wolters[mmm], Wei Lee Woon[nnn], James Wu[ooo], Catherine Wu[y], Kengran Yang[aaa], Jingwen Yin[ll], Bingyu Zhao[ppp], Chenyun Zhu[ll], Jeanne Brooks-Gunn[qqq,rrr], Barbara E. Engelhardt[y,ii], Moritz Hardt[sss], Dean Knox[z], Karen Levy[ttt], Arvind Narayanan[y], Brandon M. Stewart[a], Duncan J. Watts[uuu,vvv,www], and Sara McLanahan[a,1]

Contributed by Sara McLanahan, January 24, 2020 (sent for review October 1, 2019; reviewed by Sendhil Mullainathan and Brian Uzzi)

Garbage In, Garbage Out

Social predictability?

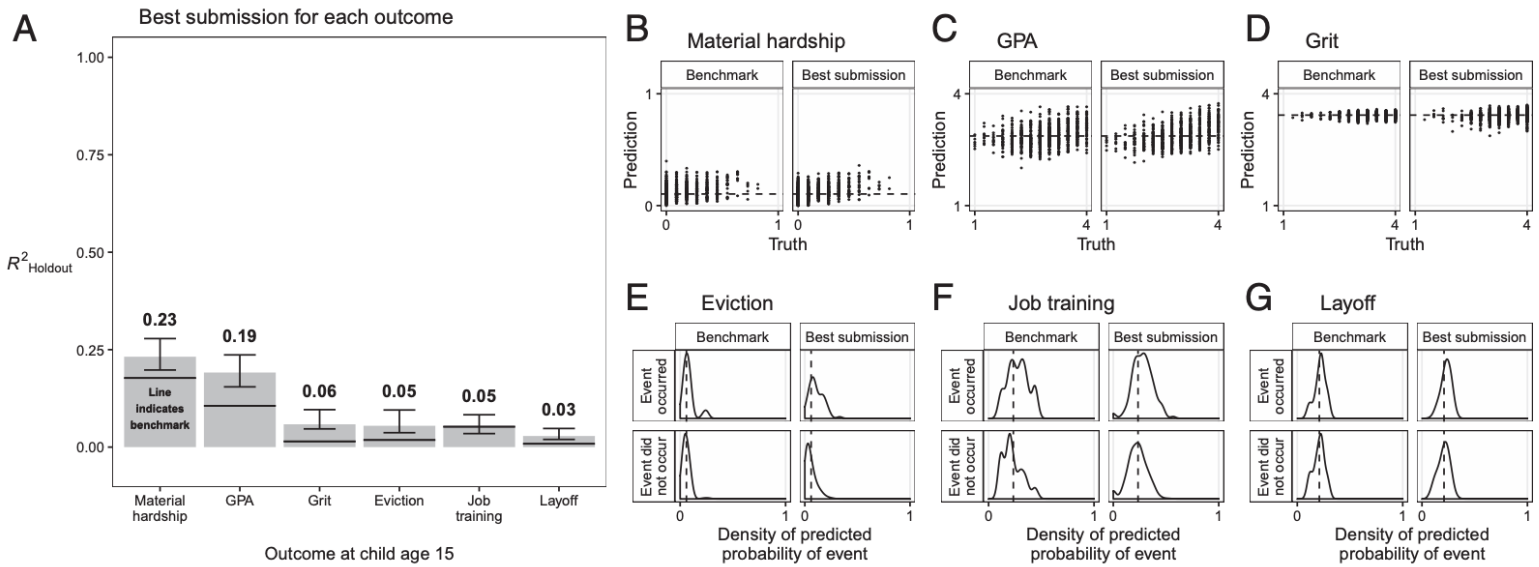


Fig. 3. Performance in the holdout data of the best submissions and a four variable benchmark model (*SI Appendix*, section S2.2). *A* shows the best performance (bars) and a benchmark model (lines). Error bars are 95% confidence intervals (*SI Appendix*, section S2.1). *B–D* compare the predictions and the truth; perfect predictions would lie along the diagonal. *E–G* show the predicted probabilities for cases where the event happened and where the event did not happen. In *B–G*, the dashed line is the mean of the training data for that outcome.

Divergence in Explanation and Prediction Modeling?

Table 1 | A schematic for organizing empirical modelling along two dimensions, representing the different levels of emphasis placed on prediction and explanation

	No intervention or distributional changes	Under interventions or distributional changes
Focus on specific features or effects	Quadrant 1: Descriptive modelling Describe situations in the past or present (but neither causal nor predictive)	Quadrant 2: Explanatory modelling Estimate effects of changing a situation (but many effects are small)
Focus on predicting outcomes	Quadrant 3: Predictive modelling Forecast outcomes for similar situations in the future (but can break under changes)	Quadrant 4: Integrative modelling Predict outcomes and estimate effects in as yet unseen situations

The rows highlight where we focus our attention (on either specific features that might affect an outcome of interest, or directly on the outcome itself), whereas the columns specify what types of situations we are modelling (a 'fixed' world in which no changes or interventions take place, or one in which features or inputs are actively manipulated or change owing to other uncontrolled forces).

The End of Social Theory?

All models are wrong, but some are useful; or increasingly you succeed without them

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS MORE SIGN IN SUBSCRIBE

CORONAVIRUS HOW TO GET A VACCINE APPOINTMENT BEST FACE MASKS COVID-19 FAQ NEWSLETTER

CHRIS ANDERSON SCIENCE 06.23.2008 12:00 PM

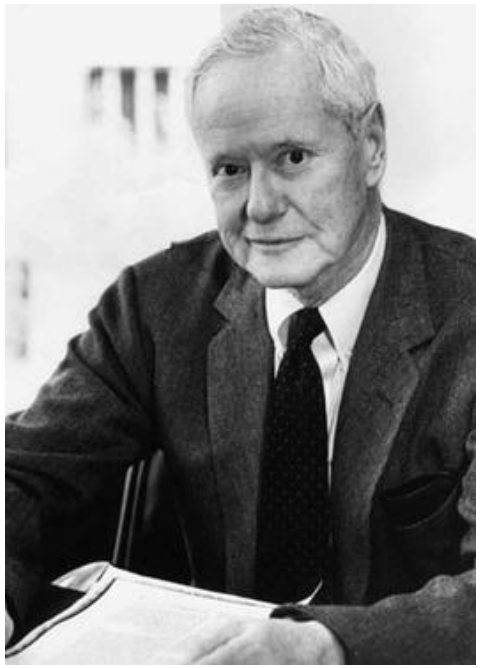
The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

Illustration: Marian Bantjes "All models are wrong, but some are useful." So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies [...]

There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

Table of Contents

Robert Merton famously wrote,
"Perhaps sociology is not yet ready for its Einstein because it has not yet found its Kepler...."



Duncan Watts, in response, writes 62 years later, "...by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves and how we interact. Merton was right: social science still has not found its Kepler. But three hundred years after Alexander Pope argued that the proper study of mankind should lie not in the heavens but in ourselves, we have finally found our telescope."

Theory In, Theory Out

Formal theory is useful not only in generating hypotheses, but also in selecting an appropriate way of measuring constructs with big data.

Computational Grounded Theory

Article

Sociological Methods & Research
1-40

DOI: 10.1177/0049124117729703
journals.sagepub.com/home/smr
SAGE

Computational Grounded Theory: A Methodological Framework

Laura K. Nelson[1]

Abstract

This article proposes a three-step methodological framework called computational grounded theory, which combines expert human knowledge and hermeneutic skills with the processing power and pattern recognition of computers, producing a more methodologically rigorous but interpretive approach to content analysis. The first, pattern detection step, involves inductive computational exploration of text, using techniques such as unsupervised machine learning and word scores to help researchers to see novel patterns in their data. The second, pattern refinement step, returns to an interpretive engagement with the data through qualitative deep reading or further exploration of the data. The third, pattern confirmation step, assesses the inductively identified patterns using further computational and natural language processing techniques. The result is an efficient, rigorous, and fully reproducible computational grounded theory. This framework can be applied to any qualitative text as data, including transcribed speeches, interviews, open-ended survey data, or ethnographic field notes, and can address many potential research questions.

Figure

Computational Grounded Theory: 3 steps

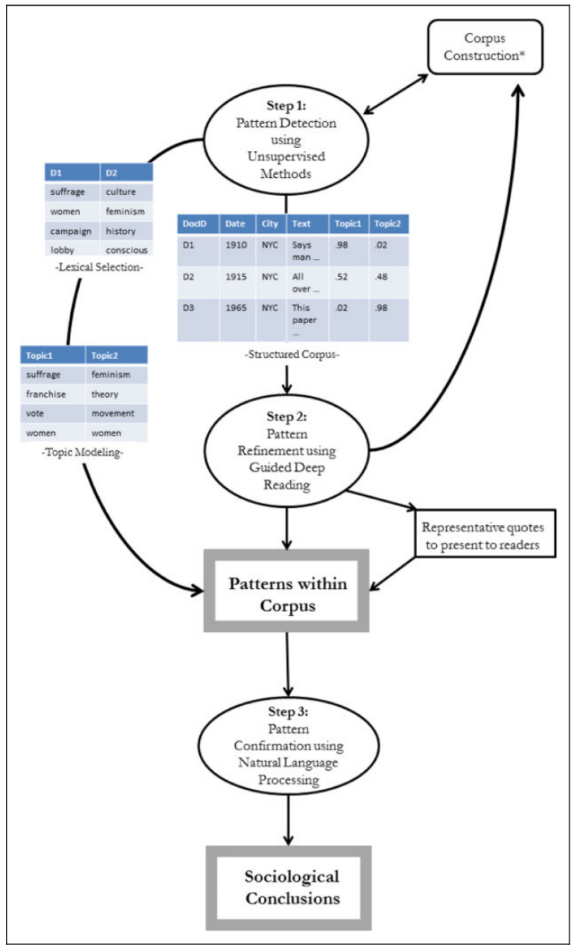
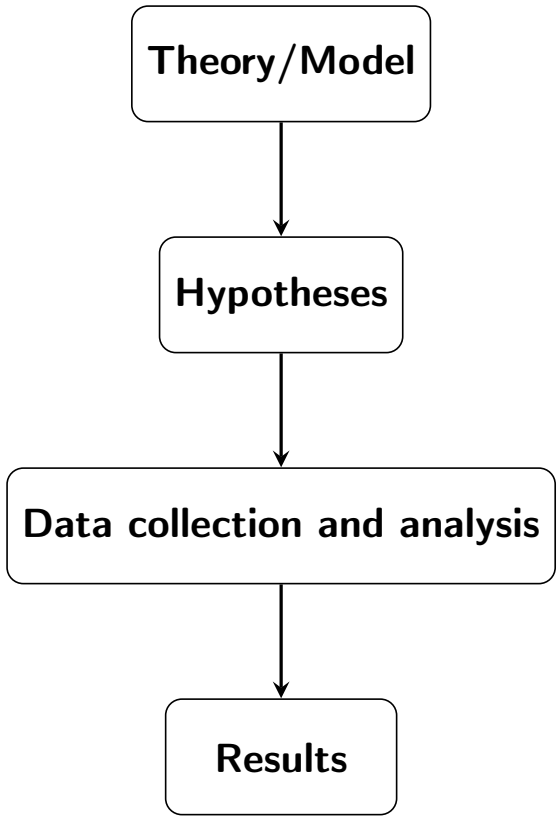
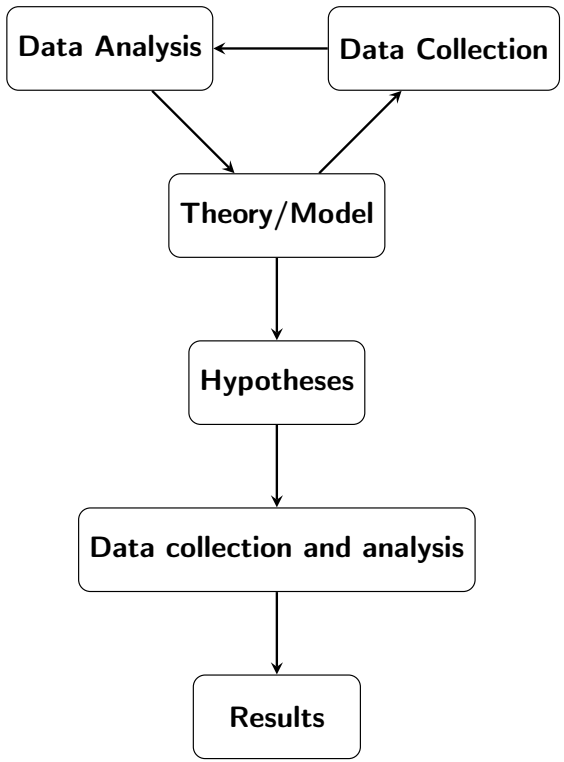


Figure 2. Three-step computational grounded theory framework: From dataframe to conclusion. This figure graphically represents the three-step computational grounded theory process. Step 1 serves two purposes: It outputs interpretable lists of

Moving from deductive to agnostic: Social science as an iterative and cumulative process



Moving from deductive to agnostic: Social science as an iterative and cumulative process



YongjunZhang.com | @DrJoshZhang

Thank you!

Following us on Twitter: @SBU_Sociology;@IACSComputes

Acknowledgment: I wish to thank IACS for the access to Seawulf High Performance Computing System at Stony Brook University.