Intro to Computational Social Science (CSS)

Yongjun Zhang, Ph.D.

Department of Sociology and IACS, Stony Brook University

2020-08-25

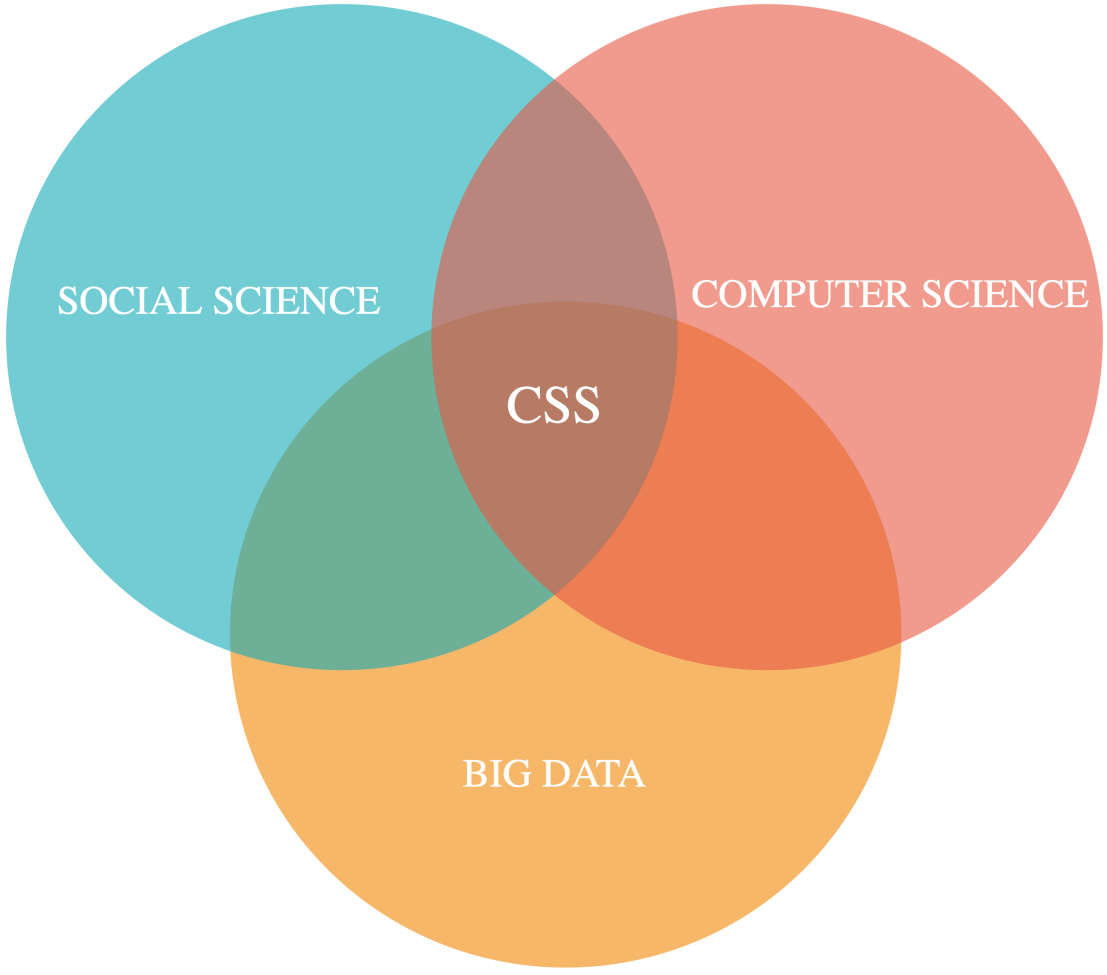
Today's Agenda

- Welcome to Intro to CSS
- A General but Brief Intro
- Syllabus and Lab Issues

What Is Computational Social Science

CSS is 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 (Edelmann et al. 2020:62).

Three Key Elements of CSS: Social Science, Computer Science, and Big Data



The Rise of CSS?

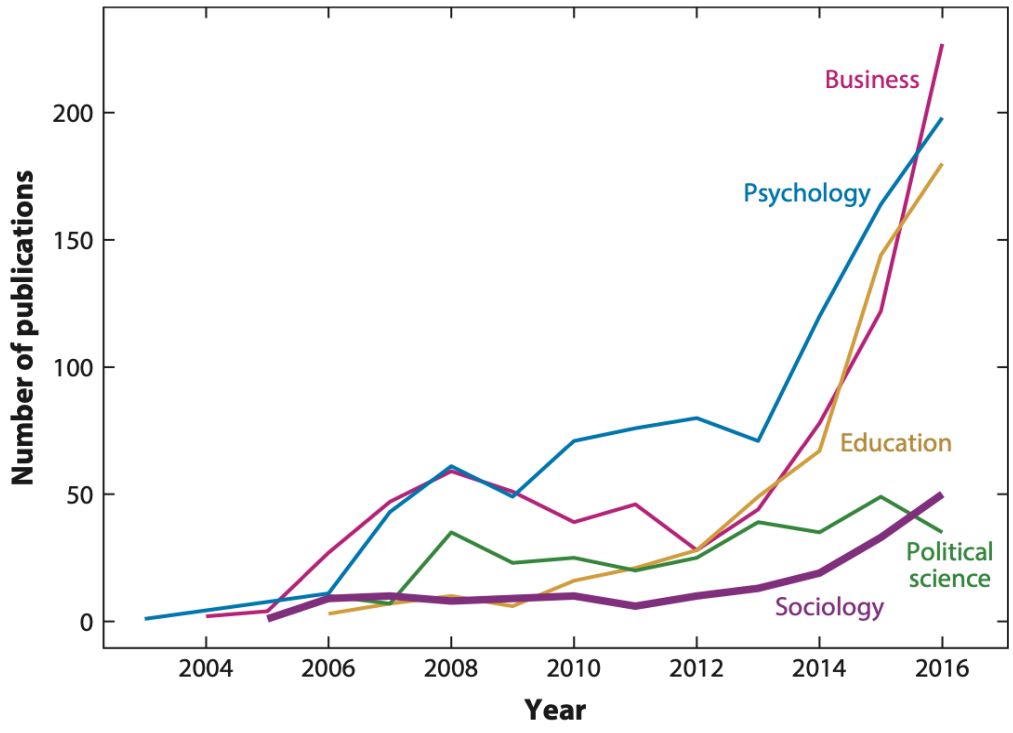


Figure 1

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

Figure 1: Edelmann et al. 2020:64

The State-of-the-Art of Current Computational Sociology

Annual Review of Sociology

Computational Social Science and Sociology

Achim Edelmann,[1,2] Tom Wolff,[3] Danielle Montagne,[3] and Christopher A. Bail[3]

[1]Institute of Sociology, University of Bern, 3012 Bern, Switzerland; email: achim.edelmann@soz.unibe.ch

[2]Department of Sociology, London School of Economics and Political Science, London WC2A 2AE, United Kingdom

[3]Department of Sociology, Duke University, Durham, North Carolina 27708, USA; email: christopher.bail@duke.edu

Abstract

The integration of social science with computer science and engineering fields has produced a new area of study: computational social science. This field applies computational methods to novel sources of digital data such as social media, administrative records, and historical archives to develop theories of human behavior. We review the evolution of this field within sociology via bibliometric analysis and in-depth analysis of the following subfields where this new work is appearing most rapidly: (*a*) social network analysis and group formation; (*b*) collective behavior and political sociology; (*c*) the sociology of knowledge; (*d*) cultural sociology, social psychology, and emotions; (*e*) the production of culture; (*f*) economic sociology and organizations; and (*g*) demography and population studies. Our review reveals that sociologists are not only at the center of cutting-edge research that addresses longstanding questions about human behavior but also developing new lines of inquiry about digital spaces as well. We conclude by discussing challenging new obstacles in the field, calling for increased attention to sociological theory, and identifying new areas where computational social science might be further integrated into mainstream sociology.

https://www.chrisbail.net/post/mapping-computational-social-science

Two Cultures?

- What is the difference between traditional social science and computational social science?

5 Breiman's *'Two Cultures'*, 2001

Leo Breiman, a UC Berkeley statistician who re-entered academia after years as a statistical consultant to a range of organizations, including the Environmental Protection Agency, brought an important new thread into the discussion with his 2001 paper in *Statistical Science*. Titled 'Statistical Modeling: The Two Cultures', Breiman described two cultural outlooks about extracting value from data.

> *Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables x (independent variables) go in one side, and on the other side the response variables y come out. Inside the black box, nature functions to associate the predictor variables with the response variables ...*
>
> *There are two goals in analyzing the data:*
>
> - *Prediction. To be able to predict what the responses are going to be to future input variables;*
> - *[Inference].*[23] *To [infer] how nature is associating the response variables to the input variables.*

Traditional social scientists are searching for statistical significance and intepretable causal mechanisms and are ignoring predictive accuracy as a measure of explanatory power.

- Social scientists ask whether the relationship between Y and X is statistically significant and in the direction predicted by the theory.

Computer scientists are searching for algorithms and ways to improve predictive accuracy, instead of substative interpretability.

A Pragmatist Theory of Social Mechanisms

Neil Gross
University of British Columbia

Some sociologists have recently argued that a major aim of sociological inquiry is to identify the mechanisms by which cause and effect relationships in the social world come about. This article argues that existing accounts of social mechanisms are problematic because they rest on either inadequately developed or questionable understandings of social action. Building on an insight increasingly common among sociological theorists—that action should be conceptualized in terms of social practices—I mobilize ideas from the tradition of classical American pragmatism to develop a more adequate theory of mechanisms. I identify three kinds of analytical problems the theory is especially well poised to address and then lay out an agenda for future research.

Prediction and explanation should be viewed as complements, instead of substitutes, in the pursuit of social science knowledge (Hofman et al. 2007)

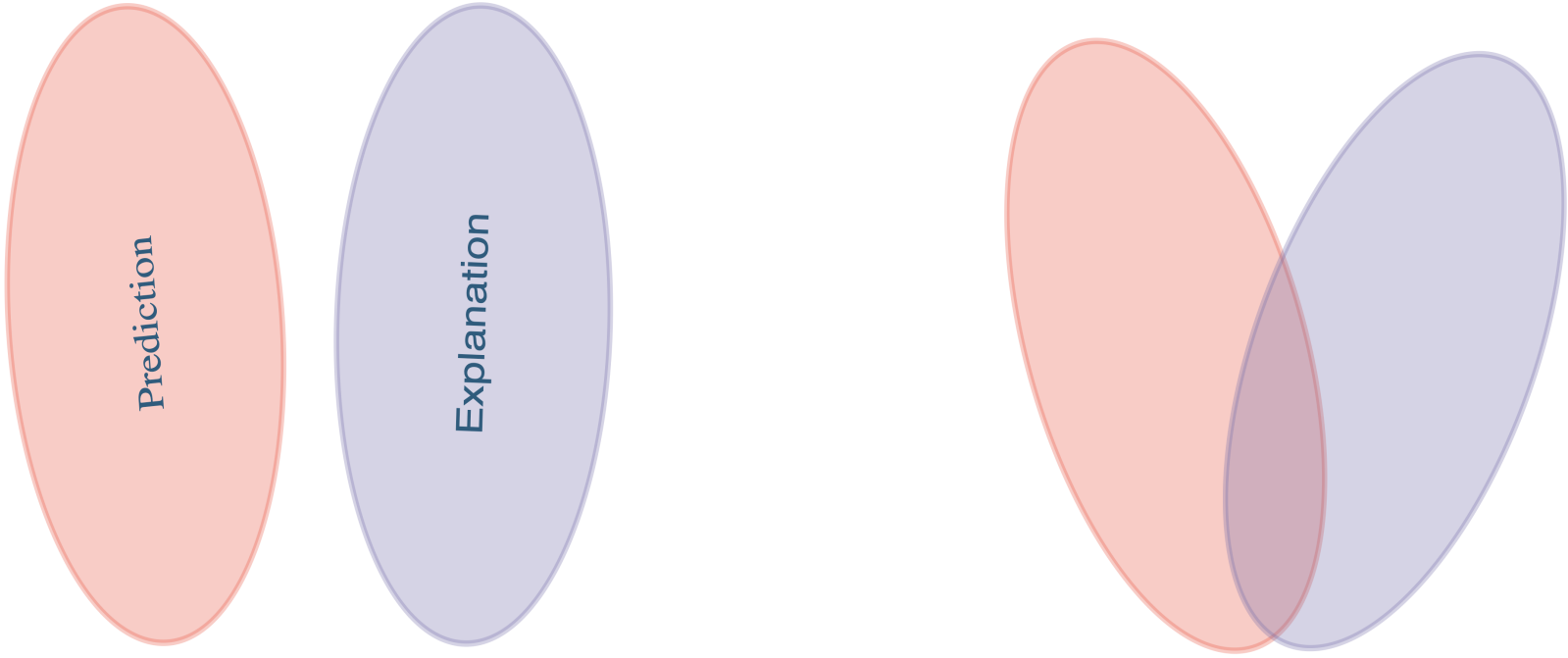


Figure 2: A Convergence of Two Cultures?

But how predictable is human and social behavior?

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)

Figure 3: The Fragile Family Challenge

How predictable is human and social behavior?

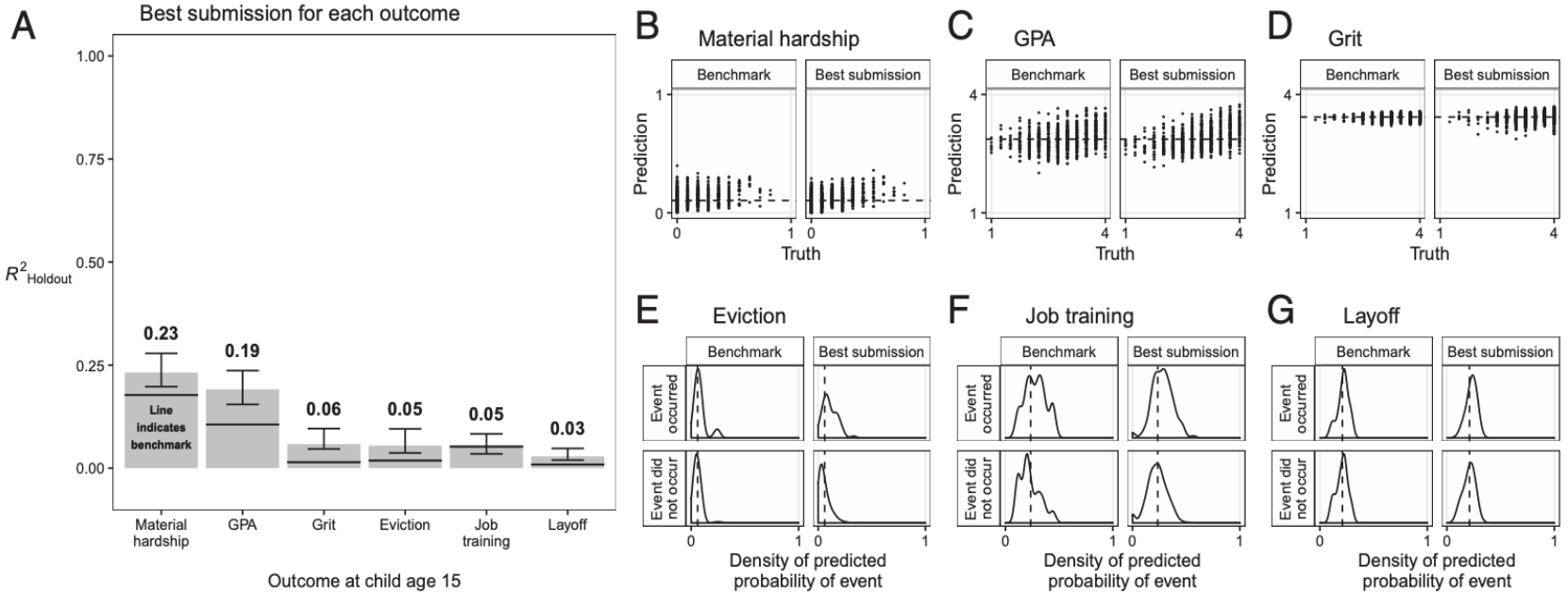


Fig. 3. Performance in the holdout data of the best submissions and a four variable benchmark model (*SI Appendix*, section S2.2). *A* shows the best performance (bars) and a benchmark model (lines). Error bars are 95% confidence intervals (*SI Appendix*, section S2.1). *B–D* compare the predictions 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.

What can CSS Offer?

- Big Data?
- Large-scale Online Experiments?
- Mass Collaboration?
- The State-of-the-Art Architecture?
- Social Theories and Mechanisms?

Some Examples

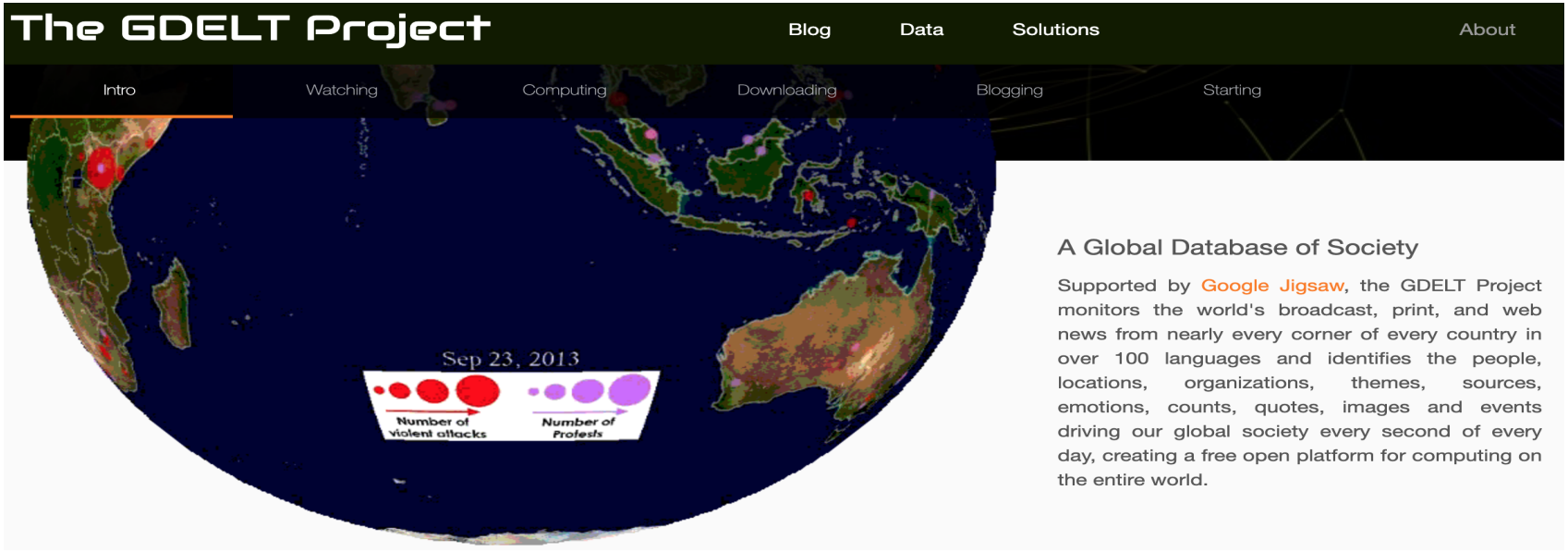


Figure 4: Global Database of Events, Language, and Tone (GDELT)

Some Examples (cont.)

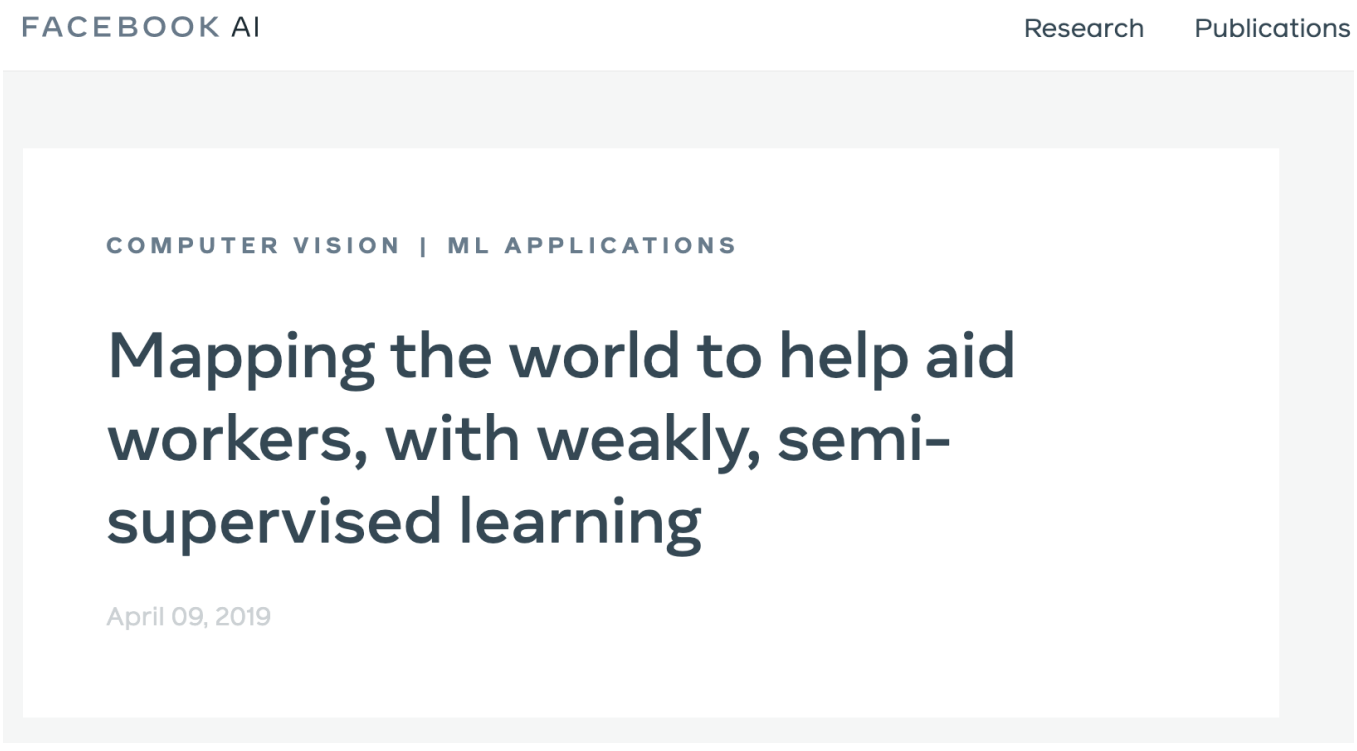


Figure 5: Facebook Population Density Project

Some Examples (cont)

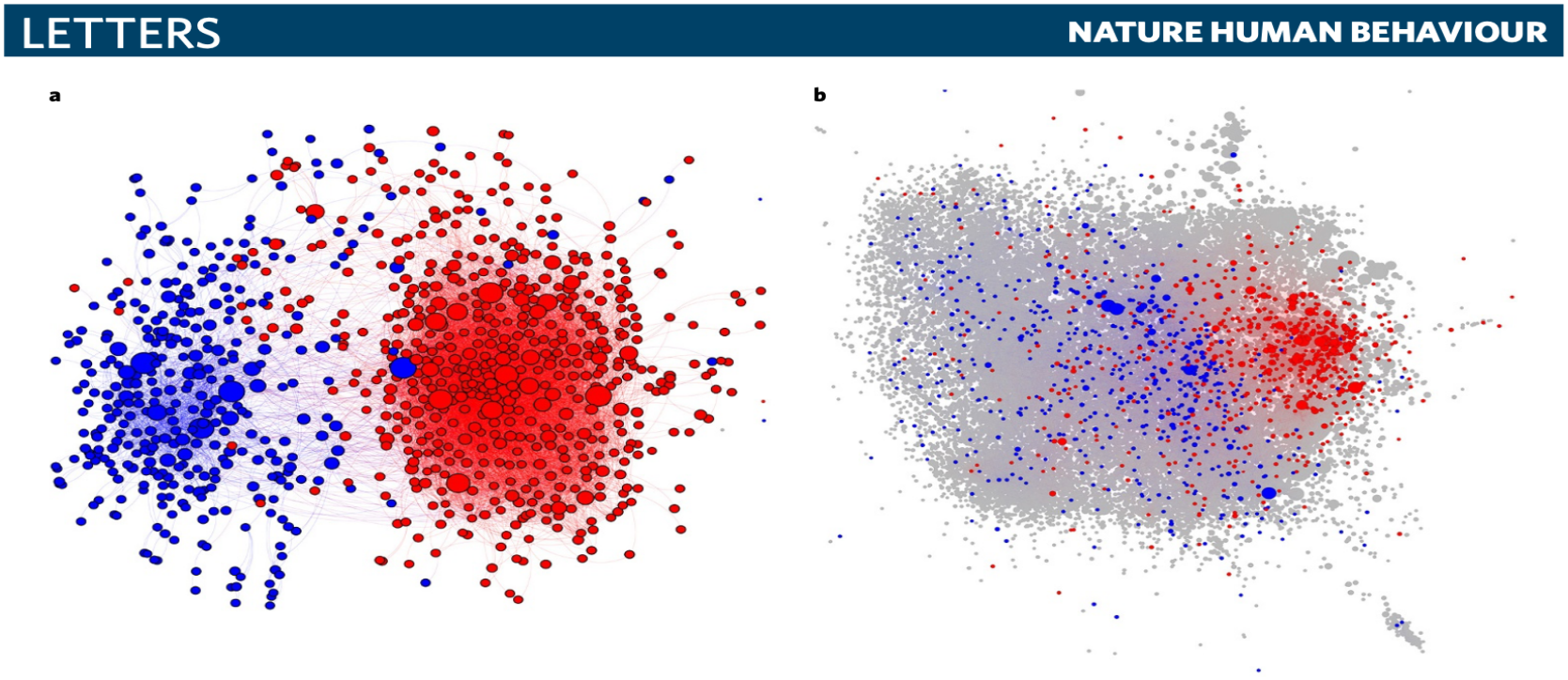


Figure 1 | Visualization of the co-purchase network among liberal, conservative and scientific books. a, Links between 583 liberal (blue) and 673 conservative (red) books. **b**, Links between these books and science (grey) books. As shown in **a**, 97.2% of red books linked to other reds and 93.7% of blue books linked to other blues. A small number of books were more likely to be co-purchased with books of a different colour. We subjected these blinded books to additional judges and found that the original codings were nearly all correct. A number of red 'orphans' were written by moderate Republicans critical of the religious right while blue 'orphans' were written by progressive community organizers such as Saul Alinsky (*Rules for Radicals*[40]), later rediscovered by the Tea Party who reference their effective ethnic blue-collar organizing tactics. In **b**, the broader distribution of blue-linked science books indicates that readers of blue books have broader disciplinary interest and co-purchased books more centrally located in the network of co-purchased science books.

Figure 6: Millions of online book co-purchases reveal partisan differences in the consumption of science

Some Examples (cont)

American Political Science Review Vol. 108, No. 3 August 2014
doi:10.1017/S0003055414000306

An Empirical Evaluation of Explanations for State Repression

DANIEL W. HILL, JR. *University of Georgia*
ZACHARY M. JONES *Pennsylvania State University*

The empirical literature that examines cross-national patterns of state repression seeks to discover a set of political, economic, and social conditions that are consistently associated with government violations of human rights. Null hypothesis significance testing is the most common way of examining the relationship between repression and concepts of interest, but we argue that it is inadequate for this goal, and has produced potentially misleading results. To remedy this deficiency in the literature we use cross-validation and random forests to determine the predictive *power of measures of concepts the literature identifies as important causes of repression. We find that few of these measures are able to substantially improve the predictive power of statistical models of repression. Further, the most studied concept in the literature, democratic political institutions, predicts certain kinds of repression much more accurately than others. We argue that this is due to conceptual and operational overlap between democracy and certain kinds of state repression. Finally, we argue that the impressive performance of certain features of domestic legal systems, as well as some economic and demographic factors, justifies a stronger focus on these concepts in future studies of repression.*

Some Examples (cont)

Journal of Child and Family Studies
https://doi.org/10.1007/s10826-020-01775-5

ORIGINAL PAPER

Adolescent Family Experiences Predict Young Adult Educational Attainment: A Data-Based Cross-Study Synthesis With Machine Learning

Xiaoran Sun[1] · Nilam Ram[1] · Susan M. McHale[1]

Abstract
Grounded in theory and research on the role of adolescent family experiences in young adult educational attainment, this study took the novel step of synthesizing results from prior studies and using a machine learning (ML) approach to address three questions: (1) By incorporating adolescent family experience factors examined across prior studies in a single analysis, how accurately can we predict young adult educational attainment? (2) Which family experience factors are the best predictors of young adult educational attainment? (3) What complex patterns among family experience predictors merit further examination? Based on a review of 101 publications that used National Longitudinal Study of Adolescent Health data to investigate links between adolescent family experiences and young adult attainment, we identified 53 family experience independent variables. We used an ML-based approach to train and test models with these 53 Wave I family variables (adolescent in Grade 7–12) as predictors of both college enrollment ($N = 4598$) and graduation ($N = 4180$) at Wave IV (young adult mean age = 28.88, $SD = 1.76$). Our models (1) obtained prediction accuracies of 73.43% and 72.33% for college enrollment, and 79.10% and 79.07% for college graduation, (2) identified the best predictors of college enrollment and graduation, including family socioeconomic characteristics and parent educational expectations, and (3) highlight nonlinear patterns for further examination. This study advanced understanding of how adolescent family experiences may influence educational attainment and provided a paradigm for developmental research to synthesize existing findings into novel discoveries with large-scale datasets.

Keywords Adolescent family experiences · Young adult educational attainment · Machine learning · Cross-study synthesis · National Longitudinal Study of Adolescent Health

Some Examples (cont)

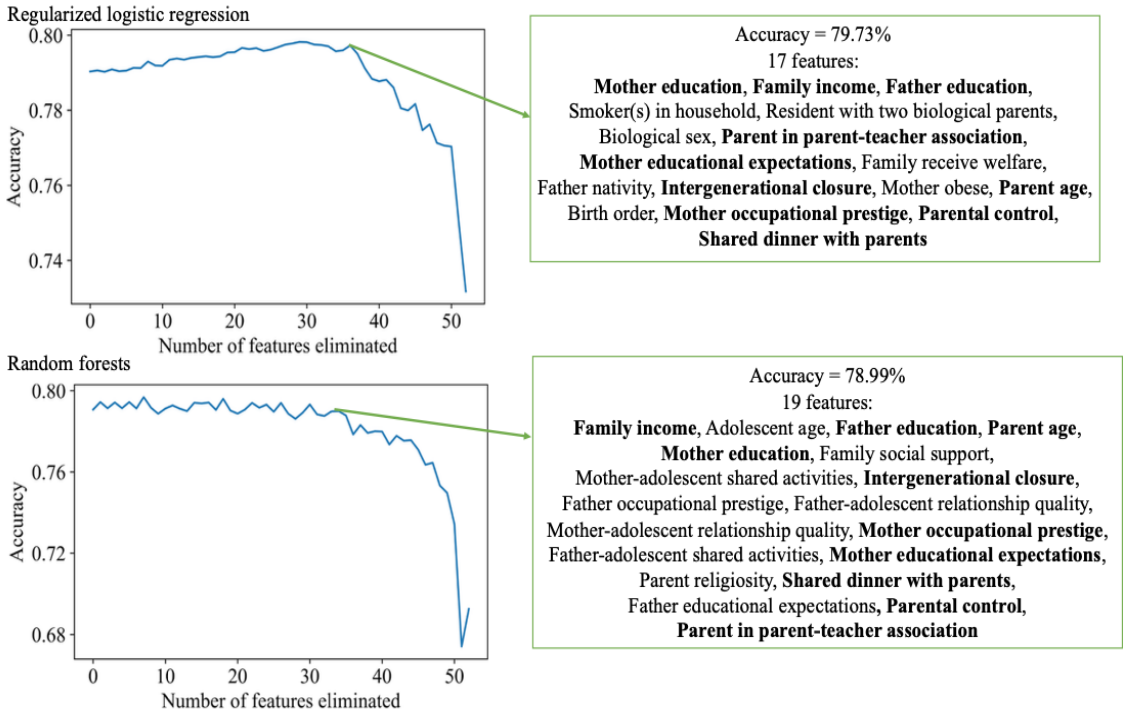


Fig. 2 Recursive feature elimination results for regularized logistic regression (upper) and random forests (lower) models predicting college graduation. The line indicates accuracy of models where the least useful predictors (between 0 and 52) were eliminated from the model. The box indicates accuracy of the most parsimonious model that performed well, along with a list of the features that were identified as important for prediction. Features that were important in both the logistic regression and random forest models are indicated in bold

Some Examples (cont)

RESEARCH

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.

Some Examples (cont)

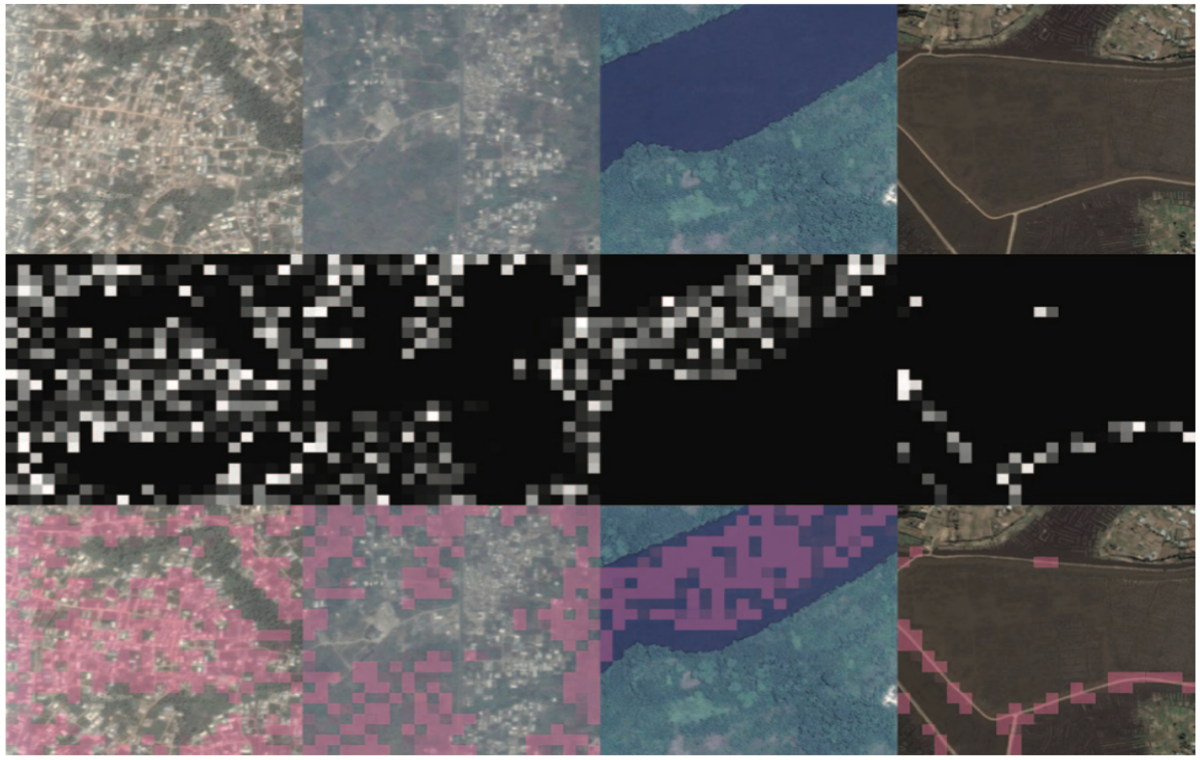


Fig. 2. Visualization of features. By column: Four different convolutional filters (which identify, from left to right, features corresponding to urban areas, nonurban areas, water, and roads) in the convolutional neural network model used for extracting features. Each filter "highlights" the parts of 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

Some Examples (cont)

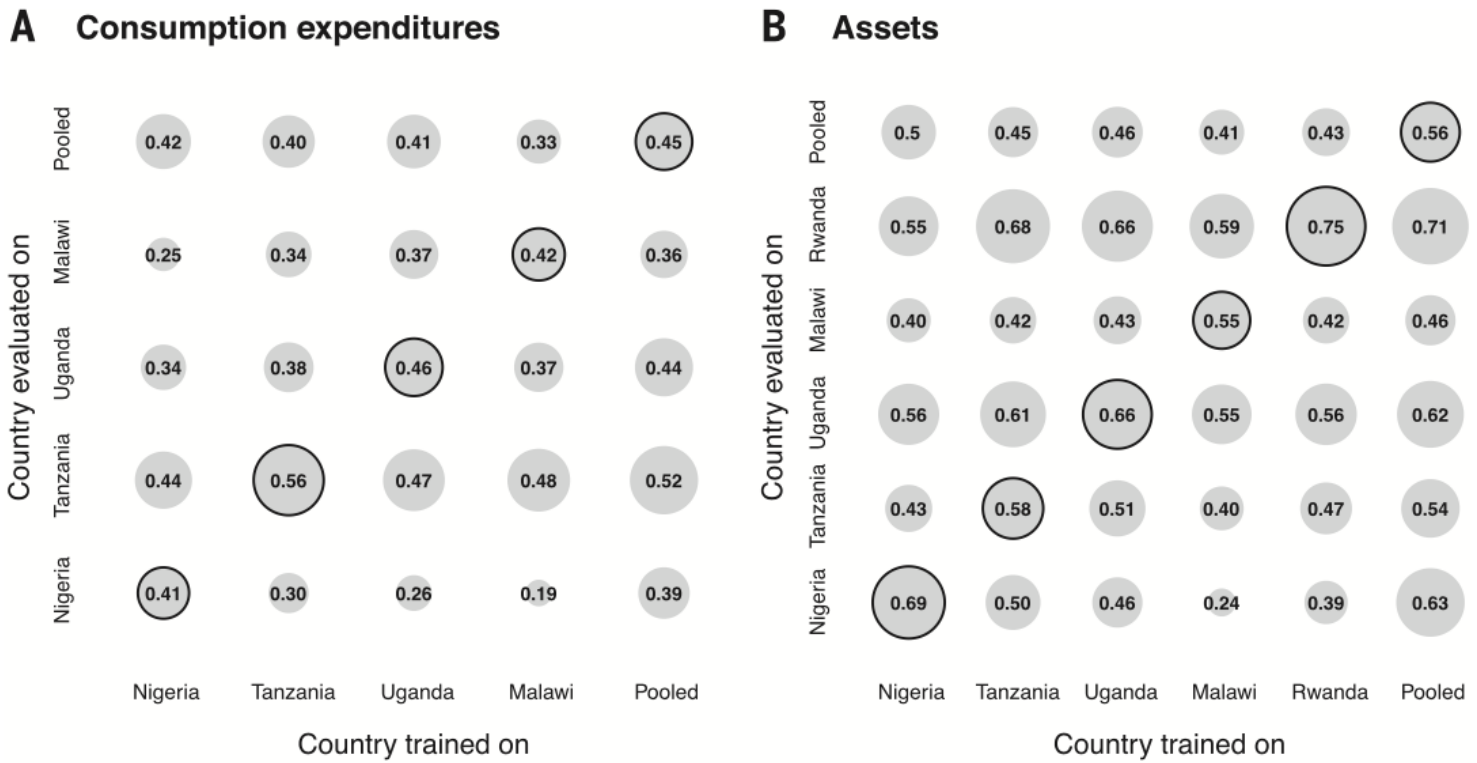


Fig. 5. Cross-border model generalization. (**A**) Cross-validated r^2 values for consumption predictions for models trained in one country and applied in other countries. Countries on *x* axis indicate where model was trained, countries on *y* axis where model was evaluated. Reported r^2 values are averaged over 100 folds (10 trials, 10 folds each). (**B**) Same as in (A), but for assets.

Walk through our Syllabus

Please click here for soc591.s2 syllabus.

Thank You

Yongjun Zhang, Ph.D.

Yongjun.Zhang@stonybrook.edu

https://yongjunzhang.com