

Humans in the Loop: Incorporating Expert and Crowd-Sourced Knowledge for Predictions Using Survey Data

Socius: Sociological Research for a Dynamic World
 Volume 5: 1–15
 © The Author(s) 2019
 Article reuse guidelines:
sagepub.com/journals-permissions
 DOI: 10.1177/2378023118820157
srd.sagepub.com
SAGE

Anna Filippova¹, Connor Gilroy², Ridhi Kashyap³,
 Antje Kirchner^{4,5}, Allison C. Morgan⁶, Kivan Polimis⁷ ,
 Adaner Usmani⁸, and Tong Wang⁹

Abstract

Survey data sets are often wider than they are long. This high ratio of variables to observations raises concerns about overfitting during prediction, making informed variable selection important. Recent applications in computer science have sought to incorporate human knowledge into machine-learning methods to address these problems. The authors implement such a “human-in-the-loop” approach in the Fragile Families Challenge. The authors use surveys to elicit knowledge from experts and laypeople about the importance of different variables to different outcomes. This strategy offers the option to subset the data before prediction or to incorporate human knowledge as scores in prediction models, or both together. The authors find that human intervention is not obviously helpful. Human-informed subsetting reduces predictive performance, and considered alone, approaches incorporating scores perform marginally worse than approaches that do not. However, incorporating human knowledge may still improve predictive performance, and future research should consider new ways of doing so.

Keywords

Fragile Families Challenge, machine learning, surveys, prediction, missing data

Social science survey data sets are often wider than they are long. Resource limitations demand that surveys ask many questions of the minimum number of respondents needed for statistical analyses. Moreover, social scientists are often interested in hard-to-reach populations, accentuating the need to ask many questions of few respondents. These difficulties characterize the Fragile Families and Child Wellbeing Study (FFCWS), which follows a cohort of nearly 5,000 children born in large U.S. cities between 1998 and 2000, roughly three quarters of them to unmarried parents (Reichman et al. 2001). The study collects a wealth of information about this disadvantaged group, including children’s physical and mental health, cognitive function, schooling, and living and family conditions. Overall, the FFCWS data set contains nearly 13,000 variables.

The breadth of the variables contained in the FFCWS data set presents opportunities for a prediction task such as the Fragile Families Challenge (FFC). The FFC asked participants to use a data set containing variables collected from the child’s birth until year 9, and some training data from year 15, to predict six outcomes in the year 15 data:

grade point average (GPA) and grit of the child,¹ material hardship and eviction of the family, layoff of the primary caregiver, and whether the primary caregiver participated in a job skills program. Although there is considerable information on each child, there are few children in the data

¹The FFCWS defines grit as a measure of passion and perseverance.

¹GitHub, Carnegie Mellon University, San Francisco, CA, USA

²University of Washington, Seattle, WA, USA

³University of Oxford, Oxford, UK

⁴RTI International, Research Triangle Park, NC, USA

⁵University of Nebraska–Lincoln, Lincoln, NE, USA

⁶University of Colorado, Boulder, CO, USA

⁷Donde Centre, Università Bocconi, Bocconi Institute for Data Science and Analytics, Milan, Italy

⁸Brown University, Providence, RI, USA

⁹University of Iowa, Iowa City, IA, USA

Corresponding Author:

Connor Gilroy, University of Washington, Box 353340, Seattle, WA 98195, USA
 Email: cgilroy@uw.edu



set. As a result, new problems arise. Specifically, the high ratio of variables to observations increases the possibility of overfitting, that is, of fitting a complex model to statistical noise in a way that yields less useful out-of-sample predictions. In this article, we explore whether human-informed variable selection and parameter tuning can help solve this problem.

Machine-learning (ML) methods have been increasingly applied to data with a high ratio of variables to observations to help with these same tasks (so-called feature selection). They provide ways to effectively use vast amounts of information contained in high-dimensional data sets (Donoho 2017). In contrast to substantive social science approaches, ML methods are less concerned with theoretical informativeness and favor data-driven predictive performance. Social scientists, on the other hand, usually draw on knowledge about the underlying data-generating process linking variables to outcomes.

Increasingly, a number of applications in computer science have sought to incorporate human knowledge into ML methods (e.g., Branson et al. 2010). However, applications of these “human-in-the-loop” approaches are rare in the social sciences. In this article, we implement a human-in-the-loop approach to the FFCWS’s prediction tasks. We surveyed a scholarly community of social scientists as well as an anonymous community of laypeople to elicit their beliefs about which variables in the FFCWS data set would best predict each of the six outcomes. We used the information from these surveys in different ways. First, we subsetting the FFCWS data set preemptively, using either the variables identified by these surveys or a preexisting set of variables identified by the Fragile Families team. Second, we used information on scores assigned to particular variables to assign weights in the ML method. In effect, our ML approach was more likely to use variables with higher scores. We contrasted these human-in-the-loop approaches to a data-driven ML approach making use of the full data set of nearly 13,000 variables.

The article proceeds as follows. First we outline how we elicited scholarly expertise and lay judgments. To use the extensive collection of variables in the FFCWS for our modeling approaches, we needed to address the issue of missing values in the data set. Next we describe how we addressed missingness. Thereafter we describe the models used, present results, and conclude.

Using Expert and Crowd-Sourced Knowledge

There might be several ways to collect knowledge about the predictors of the outcomes in the FFC. One could screen publications or conduct interviews with individuals familiar with the FFCWS. We leveraged computational tools to retrieve insights from scholars in three steps. First, we used Amazon Mechanical Turk (MTurk) to retrieve the contact

information on every author who had published using the FFCWS (786 authors). Then, we administered online surveys to each author to identify relevant predictors of each outcome. Expert surveys have been used for a variety of predictive or forecasting tasks, from projections of fertility, mortality, and immigration (Billari, Graziani, and Melilli 2012; Bijak and Wiśniowski 2010) to measuring the quality of democracy (Pemstein et al. 2015) and to school planning (Raftery et al. 2012). Experienced researchers carry a wealth of knowledge about the relationships between variables and outcomes in these data, not all of which is published. By surveying researchers, we hoped to recover knowledge that was otherwise inaccessible at relatively low cost and over no more than few days. We also fielded the same survey to a comparison sample of laypeople that we crowd-sourced using MTurk.

To elicit expert and lay beliefs, we used a wiki survey. We chose this to maximize accessibility, efficiency, and openness to new knowledge (Salganik and Levy 2015). We asked participants to choose which of two randomly selected predictors were likely to best predict a given outcome. These predictors were initially drawn from a list of 27 predictors suggested by a group of researchers familiar with the FFC, but participants were given the opportunity to add candidate predictors to the list (which would then be voted on by subsequent participants). As we explain in the Appendix, these predictors were higher-level concepts rather than specific variables. We used the data from the online surveys to generate an ordered list of candidate predictors; we scored each variable as the number of times it was voted for divided by the number of times it appeared in a pair. Further details about the surveys are included in Appendix A.

Overall, 104 of 786 sampled experts participated, generating 2,651 votes. Seven hundred laypeople participated in our MTurk surveys, generating 27,221 votes. We used the variables identified through the expert and MTurk surveys in two different ways for our predictions. First, we used it to subset the data. Together, the expert and MTurk surveys yielded 68 higher-level concepts, which we associated with 271 variables from the FFCWS data set. We took these 271 variables as a single, wiki survey-generated subset.² Second, we used the rankings generated by the expert and MTurk surveys directly, as information passed to an ML algorithm. In this case, this yielded two approaches rather than one: one that used expert scores and one that used lay scores. Details are provided in the section “Models.”

Imputation

Because most ML approaches require a numeric and complete data set, processing the FFCWS data to handle

²Alternatively, it would have been possible consider two subsets and thus two approaches: the set of variables voted on by experts and the set of variables identified by laypeople.

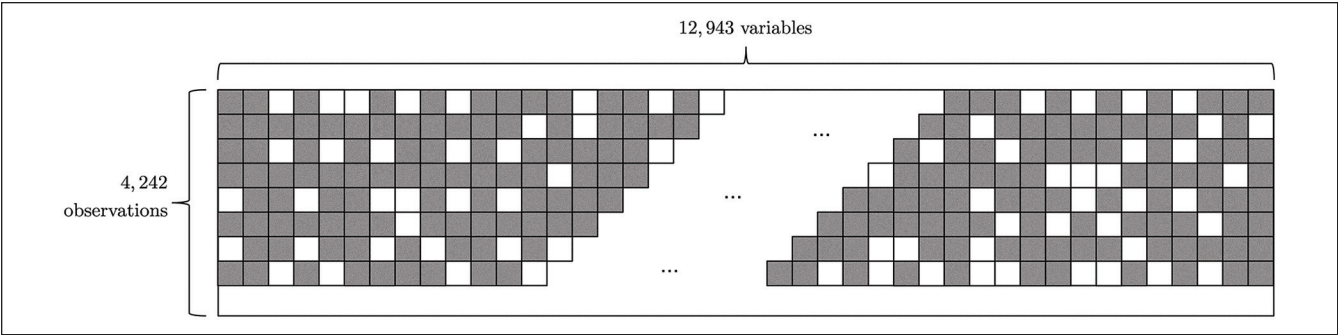


Figure 1. Missing data.
Note: The Fragile Families and Child Wellbeing Study data set includes 12,943 variables for 4,232 observations. Within this data set, 74 percent of our observations are missing completely at random, missing at random, or missing not at random.

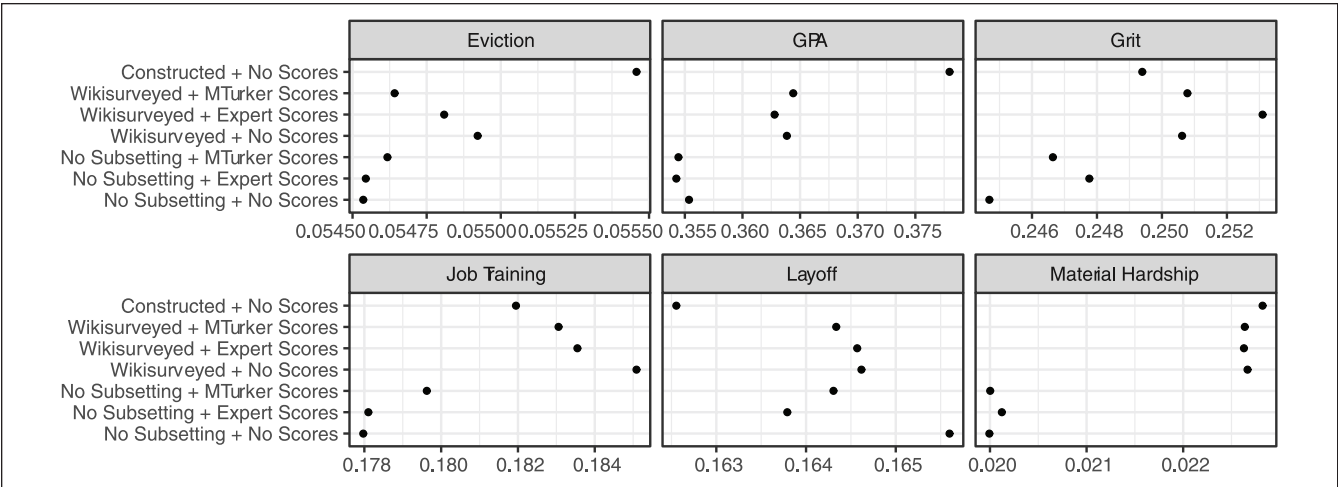


Figure 2. MSE's from approaches relevant to human-in-the-loop rankings.
Note: The mean squared errors (MSEs) from the approaches when evaluating which “human-in-the-loop” strategy performed best are shown. As explained in “Results,” because the permutation space is not filled, this comparison must be conducted on a restricted set of seven approaches (i.e., each “human-in-the-loop” approach fit to a data set imputed by linear regression accounting for variable types). Here we plot the scores from that restricted set by outcome.

missingness was a crucial step in preparing variables for modeling. To appreciate the extent of this problem, note that all observations had some missingness on some variables, which implies that there would have been no observations left with listwise deletion. Data were missing for different reasons, including unwillingness to respond, “don’t know” responses, logical skips, panel attrition, anonymization of sensitive information, and error. Roughly 74 percent of the data were missing in a way that posed problems for prediction (Figure 1). In a complex study such as this, the problems posed by missingness are particularly acute. We thus explored different imputation approaches with trade-offs in terms of efficiency and effectiveness (Appendix B). Because our different imputation strategies make different assumptions, we produced five distinctly imputed data sets on the basis of three unique approaches.

Models

We modeled the six outcomes with regularized regression. Regularization is an ML technique that can improve prediction on new data by avoiding overfitting on training data (James et al. 2017). Regularized models can be fit with large numbers of variables and relatively few observations. Regularized regression biases or shrinks model coefficients toward zero, relative to their maximum likelihood estimators, by applying a penalty to the likelihood function. Each nonzero coefficient has an associated cost. Absent other information, this cost is the same for every variable. If outside information warrants, however, the penalty can be relaxed for specific variables. The human knowledge of variable rankings captured through the scores from our survey is precisely this kind of information, and we drew on these scores to relax the penalties for the associated variables

to differing degrees. For each scored variable, the global shrinkage parameter λ , which determines the overall degree of regularization, was multiplied by a local, variable-specific *penalty factor* ranging between 0 and 1. The wiki survey scores, ranging from 0 to 100, were mapped onto penalty factors in an inverse linear fashion to determine an appropriate local penalty factor for each variable. For instance, a score of 100 mapped to a penalty factor of 0, producing an unpenalized coefficient, while a score of 0 mapped to a penalty factor of 1 and full application of the global shrinkage parameter λ . Depending on the model, variables without scores were either treated as having a wiki survey score of 0 or else excluded entirely. Although simpler, this approach takes inspiration from Bayesian approaches to global and local shrinkage (Carvalho, Polson, and Scott 2009; Lee et al. 2010; Pironen and Vehtari 2016).

We fit linear regressions for the continuous outcomes (GPA, grit, and material hardship) and logistic regressions for the binary outcomes (eviction, layoff, and job training). We used the implementation of regularized regression, with an “elasticnet” penalty, from the glmnet R package (Friedman, Hastie, and Tibshirani 2010). Appendix C describes the statistical and mathematical details of our models.

Results

In sum, we explored a total of 25 different approaches to prediction, distinguished by choices made at the following stages: (1) how we imputed missing observations, (2) whether we subsetted the data set prior to prediction and in what way, and (3) whether we incorporated outside knowledge into our modeling and in what way. As discussed, we considered five types of imputed data sets, three approaches to subsetting (no subsetting, subsetting to the variables identified by our wiki survey, and subsetting to the constructed variables identified by the Fragile Families team³), and three approaches to incorporating scores (expert scores, MTurk scores, and no scores). There were thus 45 possible permutations across these methods; of these, we focused on 25. Limitations of time and other resources narrowed the models we could run. For instance, the multiple imputation (MI) method we chose could not be run on the full data set of 13,000 variables using available computational resources.

These 25 approaches can be compared in terms of mean squared error (Figure 2). However, because we did not fill the permutation space, it is complicated to rank the performance of choices at any given stage. In an unfilled permutation space, an unrestricted comparison of any set of choices does not hold all other strategies constant. For example, the fact that we used

mean imputation with six subsetting and scoring approaches, but MI with only three, skews any comparison of the five imputation choices. Because the analytic choices we made affect our predictions, this kind of comparison is invalid. Therefore, when considering the best strategy in any given dimension, we restrict ourselves to that part of the permutation space in which we can compare across the relevant choices (Figure 3). We identify the best approach as the choice which minimizes the average or median mean squared error (MSE) across all other approaches and outcomes (Figure 4). This illustrates the relative rankings of these approaches, but the differences in performance also vary in magnitude. Therefore, we also illustrate the improvement made by any given approach, which we calculate as the average percentage improvement in MSE relative to the outcome-specific baseline MSE (Figure 5).^{4,5}

In what follows we consider what our results suggest for four different questions: (1) how to impute, (2) whether to subset, (3) whether to incorporate scores, and finally (4) whether it makes sense to include humans in the loop at all (whether by informed subsetting, or scoring, or both).⁶

Imputation

How should researchers approach issues of missingness? Overall, our results suggest that MI is best. If researchers have the computational power to pursue this approach, they should. Note, though, that by the metric of average MSE, the next best strategy is simple mean imputation and that the dividends to MI are not obviously enormous (Figure 4a). MI results in a 4.94 percent reduction in MSE relative to baseline, on average, whereas mean imputation results in a 4.61 percent reduction (Figure 5a). So, where resource constraints are an issue, mean imputation may be a viable alternative.⁷ Also, regression-based imputation methods do not clearly outperform simple mean imputation, which is noteworthy given their additional computational costs.

Subsetting

Does it make sense to preemptively subset the data before modeling? Most social science researchers who use these data no doubt do, because it is impossible for humans to make much sense of thousands of variables. It is thus tempting to do the same in a prediction exercise of this kind. Yet our results

³These 600 variables were “constructed” by the Fragile Families research staff to help future researchers on the basis of multiple reports in order to reduce missing data. These variables represent constructs that social scientists consider meaningful and can also be considered a type of approach informed by substantive human knowledge.

⁴Baseline MSE is the MSE obtained on the holdout data when each observation’s outcome is predicted by the sample mean for that outcome.

⁵Note that the facets in Figures 4 and 5 are ordered to correspond to the four sections under “Results”: “Imputation,” “Subsetting,” “Scoring,” and “Humans in the Loop?”

⁶Note that all results discussed in this paper refer to MSEs obtained from the holdout data set after the FFC closed. We requested these scores after rewriting our code to be reproducible and to explore areas of the permutation space that we had not yet explored. Table D2 in Appendix D shows the original MSEs.

⁷Note, however, that variance estimates may be underestimated.



Figure 3. Permutation space of possible and relevant approaches.

Note: The choice of five different imputation strategies, three different subsetting strategies, and three different score incorporation strategies yields a permutation space of 45 different approaches. Computational resources were available to explore only 25. Moreover, as explained in “Results,” the fact that not all 45 approaches were explored made it necessary to restrict to a subset of the explored approaches when comparing choices made in one or more dimensions. This figure shows the relevant set to which the analysis was restricted when addressing one of the four questions asked in this article. Those questions are given in the facet titles on the right-hand side of the figure.

suggest that human-informed subsetting *worsens* rather than improves predictive performance. Of all of the results in this article, this is the clearest: human-informed subsetting discards useful information. In this domain, human loses to machine.

Interestingly, the two strategies that involve subsetting are not clearly distinguishable in terms of their predictive performance. By average MSE, it seems preferable to subset to the variables from our wiki survey, but by median MSE, the constructed variables fare better. In one sense, this is as encouraging as it is surprising. The constructed variables represent the considered judgment of people with experience in the field and with the FFCWS, whereas the wiki survey variables were selected in a few days and at low cost by an anonymous community of experts and laypeople. Of course, the wiki survey was fielded within the context of the FFC with the clearly assigned task of identifying predictors for the outcomes, whereas the constructed variables were not generated explicitly for this prediction task. Nevertheless, we find there is not much to distinguish them, and if anything, the wiki survey variables perform better (Figure 5b).

Scoring

Is it useful to incorporate human knowledge into the modeling process, as described earlier? Not really, according to either of the metrics we use to rank approaches. Whether measured by average or median MSE, approaches that ignore scores altogether outperform approaches that use expert or lay scores. For advocates of an approach that marries the powers of machines to human wisdom, this is disheartening. However, there are at least two caveats. First, the differences in performance are very small. On average, as Figure 5c shows, approaches that do not use scores reduce MSE relative to baseline by about 5.39 percent, compared with 5.29 percent and 5.25 percent for experts and MTurk users, respectively. Second, as we argue below, our approach to knowledge incorporation was ad hoc. As long as it is possible to imagine better ways of incorporating human knowledge into the loop, future research should consider them.

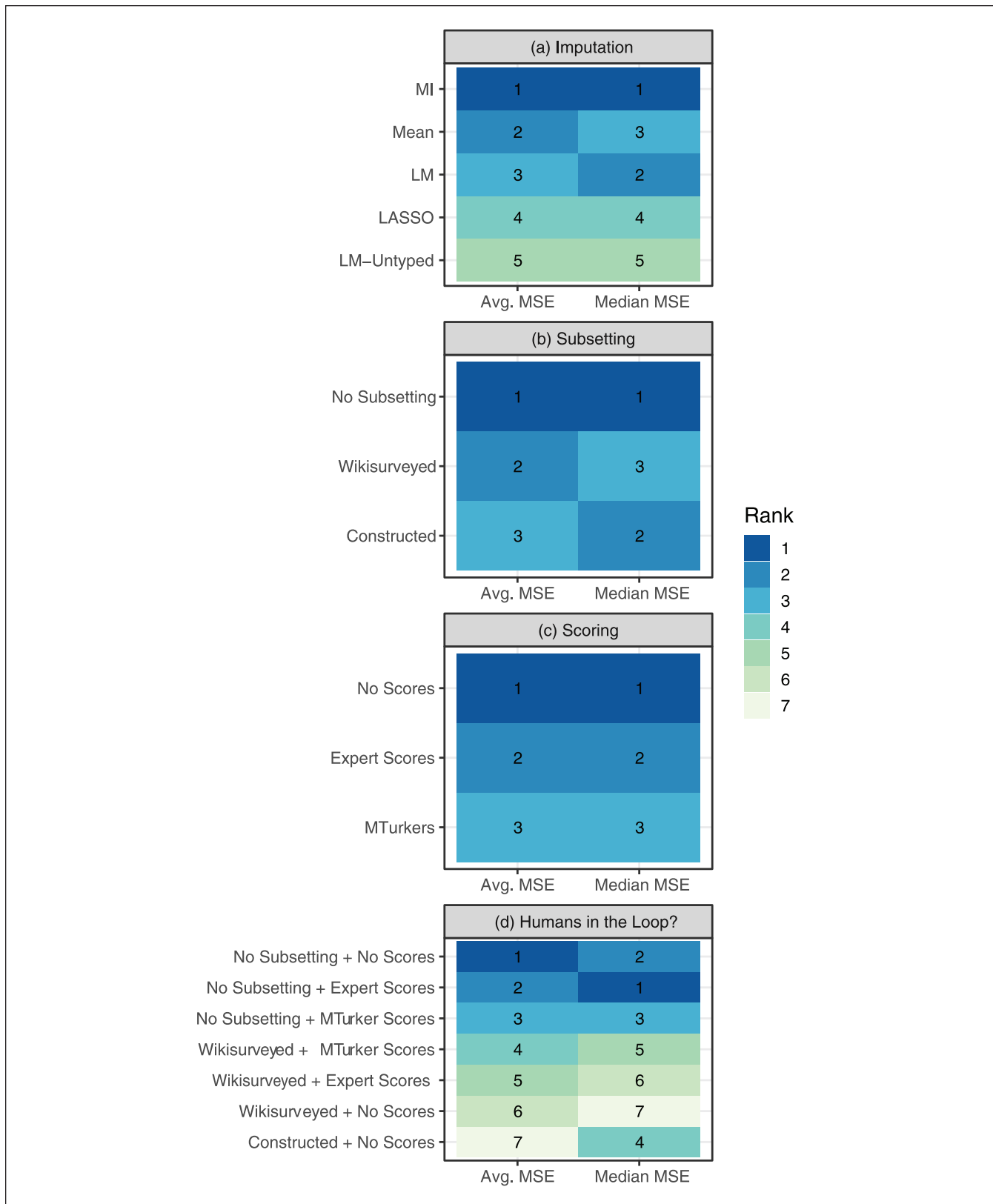


Figure 4. Rankings by lowest average and median MSE.

Note: The figure illustrates the lessons the results yield when researchers are confronted with any one of four questions: how to impute, whether to subset, whether to score, and whether to involve humans in the loop. Approaches are ranked by both average and median mean squared error (MSE), across outcomes. See “Results” for a complete discussion.

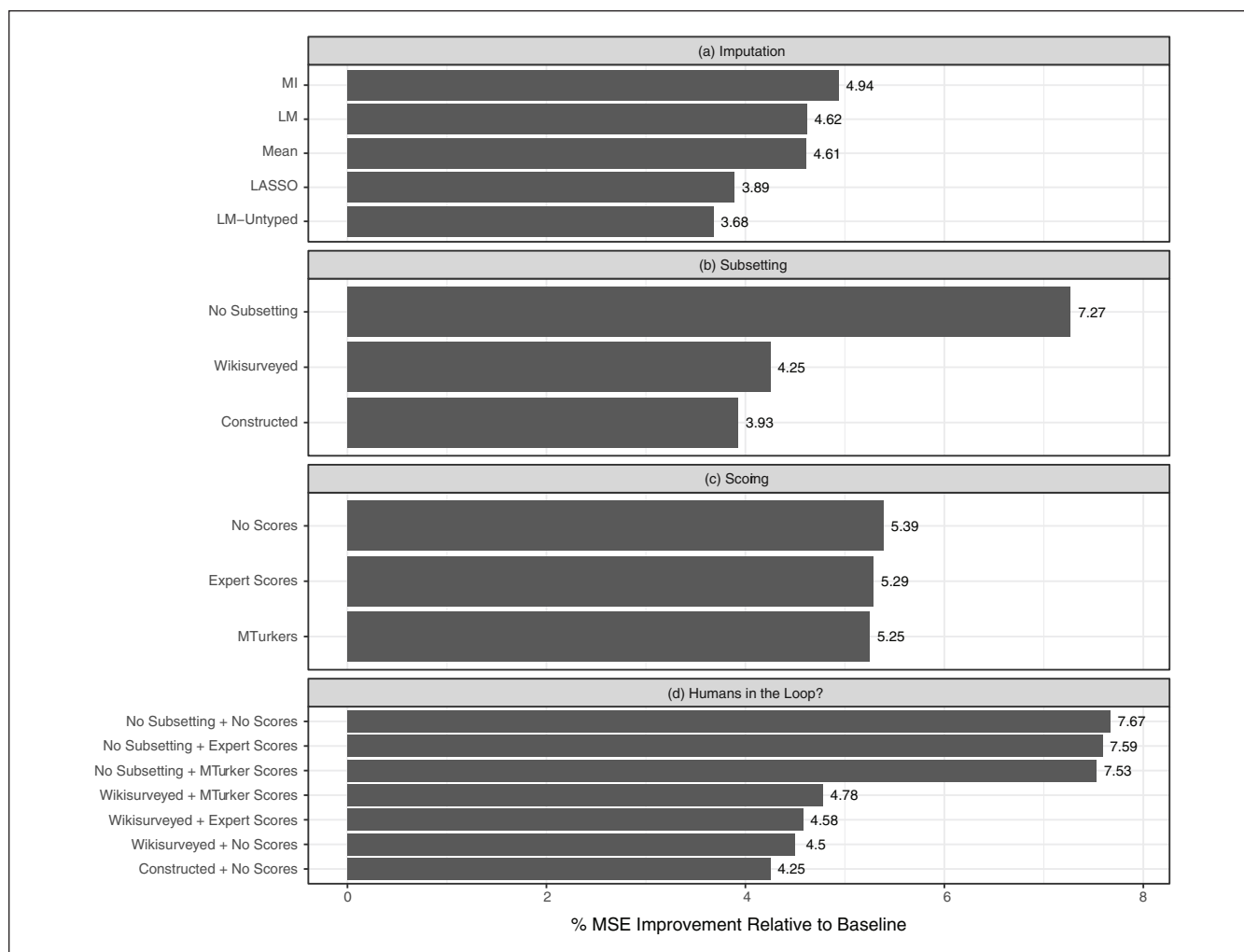


Figure 5. Average percentage reduction in MSE.

Note: Figure 4 gives the relevant rankings, but it does not convey how much these choices matter. Here, the average percentage reduction in mean squared error (MSE) is plotted relative to baseline, across outcomes, for the same four questions that surface in the main article.

Humans in the Loop?

Does all this suggest that there is no role for humans in the loop? Not entirely. By average MSE, the best approach overall is one that does not subset and does not incorporate outside knowledge. Yet, again, the differences between this and the next best (and, indeed, the third best approach) are slight: a reduction of 7.67 percent versus 7.59 percent. Furthermore, if ranked by lowest median MSE, our best performing approach does enlist humans: one that incorporates expert scores while not subsetting the data set (Figure 4d). The discrepancy between average and median MSE rankings is explained by the very poor performance of a no subsetting and no scores approach in predicting the layoff of a child's primary caregiver. This may suggest that outside information is useful for some outcomes but not others. Furthermore, one possible interpretation of this result is that strategies using expert scores are more robust to bad performance in a single outcome.

What is clear from our results is that if humans are to enter the loop, it ought not to be by preemptively subsetting the

data but rather by incorporating their wisdom into an approach that still leverages ML to extract information from the full data set. Making use of the full data set may not always be possible, as exemplified by the computational constraints we faced for generating a full data set with MI. However when it is possible, it can usefully augment prediction. Our approach incorporating scores on the basis of expert surveys fared better as a human-in-the-loop strategy. Although neither our approach to generating scores from the wiki surveys nor incorporating them in the models is dispositive, we believe that such approaches with further refinements may hold promise for human-in-the-loop strategies. In short, although there is obviously important information that only machines pick up, strategies that incorporate human knowledge to tune parameters in a model merit further exploration.⁸

⁸We trialed Bayesian rule set models on the binary outcomes in the FFC. These models incorporate the expert and MTurk scores as explicit priors. However, the predictive performance of these was poor.

Conclusions

In this article, we considered different ways of tackling a difficulty faced by researchers seeking to use survey data sets for prediction, namely, that the large ratio of variables to observations makes informed variable selection difficult. To tackle this problem, we proposed a low-cost way to mine a scholarly community for insights. We considered ways to use this information to subset a data set preemptively or at the modeling stage (or both together).

What did we find? First, our results do not recommend preemptively subsetting the data. This is common practice in social science research, which is understandable, because social scientists are often more concerned with description and explanation rather than prediction, and humans cannot make any theoretical sense of thousands of variables. But for prediction purposes, this approach obviously discards useful information. Approaches that relied on this strategy fared worse than approaches that did not. Second, we find some evidence that human insight *may* be useful if fed into ML approaches at the modeling stage, but we have not demonstrated this beyond doubt. By average MSE, our best performing approach is one that neither subsets nor incorporates scores. But it does not perform much better than one that incorporates expert scores, and this approach actually ranks best by median MSE across outcomes.

What, then, is the future of humans in the loop? We believe that future research should consider at least two types of improvements to our approach. First, the response rate of our expert survey was low: improving this would make it much easier to compare the dividends of surveying experts rather than laypeople. We expect that experts bring knowledge that laypeople do not, but our results do not clearly demonstrate this. Second, future work should consider alternative ways to incorporate human knowledge into ML models. We did so in an ad hoc way, but better formalization of our intuition and better use of the scores in modeling will surely help in deciding the place of humans in the loop, going forward.

In closing, this project considered whether approaches from the tradition of informative, human-centered modeling can be usefully combined with ML techniques. We found that their combination is not always profitable but also that their judicious combination may yet be useful.

Appendix A: Surveying the Scholarly Community

The administrators of the FFCWS maintain a database of all published work and authors. In all, there were roughly 786 authors in this database. To obtain contact information for each FFCWS author, we crowdsourced via MTurk.⁹ Using

⁹MTurk users were paid \$0.12 to click on a Google query corresponding to an author in a database. In all, this took only a few hours and cost us less than \$100. Roughly 94 percent of these e-mail addresses

their e-mail addresses, we invited these researchers to identify causes of the six outcomes of interest: eviction, GPA, grit, job training, layoffs, and material hardship.¹⁰ We supplemented this with targeted e-mails to relevant communities of experts whom we identified manually. At insignificant cost to ourselves, we were able to contact a scholarly community spanning several disciplines.

We could not possibly survey each author about each variable for each outcome, so we proceeded differently. We began with a list of candidate predictors. These were not variables from the data set but higher-level concepts (e.g., school quality) that might correspond to several variables at different points in time (e.g., a parent-teacher association at a child's school, the type of school the child attended, a gifted and talented program at the school). Surveys were seeded with 27 predictors that were suggested by group of researchers familiar with the FFC. We administered a rolling set of comparisons between two randomly selected predictors on our list using a wiki survey (Salganik and Levy 2015). Participants in a given survey (for a given outcome) chose the most relevant predictor of the pair shown.¹¹ Importantly, wiki surveys also give participants the option of adding candidate causes, which are then voted on by subsequent participants. We invited expert academics to respond to surveys pertaining to any (or all) of the six outcome variables. In addition, we administered the same surveys to MTurk users to furnish a comparison set for the experts: do experts outperform lay wisdom?¹²

Last, we generated scores for each of the variables included in our surveys. Each participant was allowed to answer an unlimited amount of questions for a given outcome and a minimum of one. The resulting data were used to generate an ordered list of candidate predictors (i.e., ranked by their estimated relevance to the outcome). We used the scores from this procedure—roughly, the number of wins divided by the number of wins plus losses—for variable selection (Salganik and Levy 2015). We manually associated each of the 68 predictors from both the expert and lay surveys with a variable (or set of variables) in the FFCWS data set and assigned to the 271 resulting variables the prior of its associated cause (Table A1).¹³ Overall, 104 of 786

were accurate; very rarely did MTurk users fail to find an author's contact information, except when authors were genuinely unlisted.

¹⁰Features, variables, and causes are used similarly throughout the text. Feature is preferred when discussing ML models, variables are used to refer to the FFCWS data set, and causes are used in the wiki survey framework.

¹¹The exact question text was "Which variable is a better predictor of . . . ?"

¹²These workers were paid \$0.38 per assignment and given up to 10 minutes to complete the survey.

¹³The modal predictor was associated with two variables, and on average, each predictor was associated with about five.

sampled experts participated, generating more than 2,600 votes. Far more rankings were collected using MTurk, roughly 700 voters, together contributing more than 27,000 votes (see Table A2).

Table A1. Predictors and Resulting Variables.

Predictor	Variables
Number of times father has missed work	{f2b30a, f3b22}
Child's IQ	{hv5_ppvtp}
Foreign-born mother*	{m3h1b, m3h1a}
Parents are in cohabiting partnership*	{cf2coh}
Private school	{p5i1a, p5i23}
Parents' substance abuse	{cm3drug_case, cf3alc_case, cm3alc_case, cf3drug_case, m4j2l, m5g20, f5g20, f4j2l}
Number of books at home	{f5k14e, m4b27, f4b27, m5k14e}
Existing number of siblings*	{cm3kids, cf4kids, cf3kids, cm1kids, cm4kids, cm5kids, cf5kids, cf2kids, cf1kids, cm2kids}
Father's interest in sports or entertainment	{f5k14b}
Home schooling	{p5i1a}
Mother's mental health	{f4c38, f5b31x}
Child's perseverance	{k5g1e, k5g1d, k5g1a, k5g1b}
Father's nonstandard work hours*	{f4k16a, f5k17l, f3k17a, f5i16a, f2k18a}
Child's birth weight*	{cm1lbw}
Families on block known well	{m4i0, m4i0l, f4i0l, p5m1, m4i0n2, m4i0n3}
Child's physical disability	{hv3a2}
Number of child's emergency room visits	{p5h10, f2b8, m2b8, hv3a9, hv4a14}
Domestic violence	{m4a8b_7, f3d7n1, m3d7n1, f5f26b2_10, f4a8b_7, m3a8b_7, m3d7p, f5f28a_10, m5f28a_10, m3d7m, m3d7n, m3d7o, m3e23q, m3d9p1, m3e23o, m3d9p, f3d9i, f3d9h, f3d9m, f3d9o, f3d9n, f3d9p, m5b30a1_10, f3a8b_7, m3d9n, m3d9o, m3d9m, m3e23p, m3d9i, m3d9n1, m5f26b2_10, f3d9n1, f3d7o, f3d7n, f3d7m, m3d7p1, m3e23p1, f5b26x_10}
Household income*	{cf1hhinc, cm5hhinc, cm2hhinc, cf5hhinc, cf3hhinc, cm1hhinc, cf4hhincb, cf2hhinc, cf5hhincb, cf3hhincb, cf4hhinc, cf2hhincb, cm4hhinc, cm3hhinc}
Child's gender*	{cm1bsex, hv4sex_child}
Mother's employment	{m2k12, m2k8, m3k15, m4k15}
Income-to-poverty ratio*	{cf3povco, cf1inpov, cf4povcob, cm3povco, cf3p...
Mother's education*	{cm5edu}
Teacher quality*	{t5g7}
Child's participation in sports	{p5i1b, m5k14b}
Grandparents are present in household	{cm3gmom, cm3gdad}
Number of parental romantic relationships*	{m3a13, f3a13, m5a10, f5a10l, f5a10, m5a10l}
Child's exposure to someone smoking	{hv4a24, p5q3cr, p5h15c}
Mother's incarceration*	{m3i29}
Child's race*	{m1h3, f1h3}
Mother's substance abuse	{cm3drug_case, cm3alc_case}
Mother's age at childbirth*	{cm2fbir, cm1age}
Child's participation in chores	{f3b32d, f3b4d, f3c3d, f3e18d, f5k14a, m3b32d, m3b4d, m3c3d, m3e18d, m5k14a, p5i1a, p5i31a, p5i40a}
Teacher says child works independently	{t5b2c}
Domestic abuse in family	{m4a8b_7, f3d7n1, m3d7n1, f5f26b2_10, f4a8b_7, m3a8b_7, m3d7p, f5f28a_10, m5f28a_10, m3d7m, m3d7n, m3d7o, m3e23q, m3d9p1, m3e23o, m3d9p, f3d9i, f3d9h, f3d9m, f3d9o, f3d9n, f3d9p, m5b30a1_10, f3a8b_7, m3d9n, m3d9o, m3d9m, m3e23p, m3d9i, m3d9n1, m5f26b2_10, f3d9n1, f3d7o, f3d7n, f3d7m, m3d7p1, m3e23p1, f5b26x_10}
Divorce or separation*	{m2a8c}
Father's age at childbirth*	{cf1age}
Father's education*	{cf5edu}
Mother has chronic illness	{m5g2a_107}
Parent's mental health	{m5g2a_101}

(continued)

Table A1. (continued)

Predictor	Variables
Amount of parental involvement in school	{m4i0d, f4i0d, m4i0}
Mother's nonstandard work hours*	{m3k16a, m2k13a, m5k17l, m4k16a, m5i16a}
Parent's chronic illness	{f5a3a1_9, m5g2a_107}
Parent impulsivity	{p5q3an}
Number of books in the home	{f5k14e, m4b27, f4b27, m5k14e}
School quality*	{t5g4_104, m4i0, m4i0b, f5k5d, p5l1a, f5k5e, p5l13f, t5c7f}
Foreign-born father*	{f3h1a, f3h1b}
Father's sense of familial responsibility	{p5i37, p5i32b, f5k15, m2e4b, n5c3f, m2c3b, f2b17a, f2b17b, f2b17c, f2b17d, f2b17e, f2b17f, f2b17g, f2b17h, p5i32a, p5i32c, n5c3e}
Availability of extended family	{cf4gdad, cflgdad, cm1gmom, cm2gdad, cm5gmom, cm4gmom, cf5gmom, cf3gmom, cf3gdad, cm5gdad, cf5gdad, cm4gdad, cm3gmom, cflgmom, cm1gdad, cm3gdad, cf4gmom, cm2gmom, cf2gmom, cf2gdad}
Father's incarceration*	{f3i29}
Mother's prenatal smoking*	{m1g4}
Family on welfare	{m3i8cl, f3i8cl, m5f8cl, m4i8cl, m5f8bl, m3i8bl, m4i8bl, m2h9cl, m2h9bl, f5f8cl, f1k2a, m1j2b, f2h8cl, f4i8cl}
Child's learning disability	{kind_a13}
Child makes friends easily	{t5b1h}
Mother's multiple job holding*	{m3k18, m3k17, m2k14a}
Father's substance abuse	{cf3alc_case, cf3drug_case}
Parents have savings account	{m5j6h}
Father's multiple job holding*	{f2k19a, f3k18}
Father absent at time of birth*	{m1a6}
Household size*	{cf5adult, cm4adult, cm3kids, cm5kids, cfladul...
Parent's religion	{f3r1, m3r1}
Neighborhood crime	{t5f4a}
Father's unemployment	{m2c33, m3c41}
Childcare center enrollment	{m4b13}
Never-married mother*	{cm5relf}
Child's health	{f5a3i_10, f5a6g02_10, m5a3a1_10, f5a3a1_10, m5a6g01_10, m5a6g03_10, f5a6g03_10, m5a3i_10, m5a6g02_10, f5a6g01_10}
Grandparents in the household	{cm3gmom, cm3gdad}
Multigenerational household	{cf4gdad, cm4gdad, cf4gmom, cm4gmom}

* Ideas with which each survey was seeded.

In Table A1, asterisks are used to denote the 27 ideas with which we seeded each survey. Therefore, of the 68

total predictors considered, 60 percent were respondent generated.

Table A2. Participation in Wiki Survey by Outcome.

Outcome	Experts		MTurk Users	
	Votes	Voters	Votes	Voters
Grade point average	530	24	5,130	137
Grit	299	9	4,419	110
Material Hardship	777	31	5,533	127
Eviction	980	32	3,741	113
Layoff	32	4	3,964	115
Job Training	33	4	4,434	129

Appendix B: Methods of Imputation

To handle missing data, we implemented three different imputation techniques:

1. Mean imputation: For each variable, we imputed the mean (continuous) or modal (categorical) value. This served as a baseline to compare our more sophisticated imputation methods.¹⁴
2. Regression-based approaches: We considered each variable in isolation as the outcome of three to five of the best available predictors of the variable.¹⁵ We developed three variations on this approach:
 - a. linear (ordinary least squares) regression imputation, treating all variables as continuous;
 - b. linear regression imputation, treating continuous variables as continuous and categorical variables as categorical; and
 - c. regularized (least absolute shrinkage and selection operator [LASSO] based) regression imputation (Kenkel and Signorino 2014), accounting for variable types.
3. MI: There are many approaches to multiply imputing missing data. We used Amelia (Honaker, King, and Blackwell 2011), which generates missing values from a multivariate normal distribution, with appropriate transformations for categorical variables. For computational reasons we applied this method to subsets of variables rather than the whole data set.

Additionally, we had to distinguish between categorical and continuous variables in our data, a classification that is not obvious from the FFCWS codebook; missing information regarding variable type is another characteristic of data sets with high dimensionality. We developed metadata-based heuristics for automated classification, followed by a manual reclassification of 230 variables from categorical to continuous.¹⁶ We made several simplifying assumptions to keep our imputation strategies parsimonious. We did not use additional data, interactions or nonlinear effects for imputation. Finally, for the regression-based imputation, missingness was extensive enough to render an initial round of mean imputation necessary.

Appendix C: Regularized Regression

Regularized regression is a statistical learning method for addressing cases in which the number of parameters to be

estimated is large relative to the number of observations available (Friedman et al. 2010). By imposing a constraint on model coefficients, it reduces overfitting (James et al. 2017). In statistical terms, regularized regression trades increased bias for reduced variance, and it often achieves better predictive performance than maximum likelihood estimation (MLE). Here we describe the technical details of our regularized regression models.

Regularized regression differs from MLE through addition of a penalty term. For continuous outcomes, elasticnet regularized regression estimates model coefficients β as follows:

$$\underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2n} \|y - X\beta\|_2^2 + \lambda * \text{penalty} \right),$$

where λ is a global shrinkage parameter, described below. For a given variable j , the penalty term is

$$s_j \left[\frac{(1-\alpha)}{2} \|\beta_j\|_2^2 + \alpha \|\beta_j\|_1 \right],$$

where α is a mixing parameter, described below. The local penalty factor s_j ranges from 0 to 1 and is mapped onto scores ranging from 0 to 100 as follows:

$$s_j = \left(1 - \frac{\text{score}}{100} \right).$$

In glmnet (Friedman et al. 2010), these penalty factors s are rescaled internally to sum to J , where J is the number of variables. The parameters λ and α control the degree and type of regularization.

First, how much should coefficient estimates be regularized? The overall degree of regularization is determined by the global shrinkage parameter λ . High values of λ force all coefficients closer to zero; by contrast, as λ itself goes to zero, regularized regression approaches traditional MLE. In our models, λ was *tuned* for each model separately; the optimal λ value for every model was chosen through cross-validation.

Second, what kind of regularization should be used? The two approaches to penalization are summing absolute values of coefficients (called LASSO regression) and summing squared values (called ridge regression). Elasticnet regression combines these approaches, using the mixing parameter α to control the weight given to each form of regularization. The parameter α ranges from 0 to 1, with 0 corresponding to ridge regression and 1 corresponding to LASSO regression. In our models, we tuned α values a single time for each outcome using a grid search and then held these values fixed for subsequent model runs. Our tuned α values ranged from 0.025 to 0.15. Because different α values did not produce large differences in cross-validated model fits on the training data, we did not pursue further optimization of this parameter.

Finally, how can human knowledge be incorporated into regularized regression? We used wiki surveys to produce a ranking of concepts relevant to our six outcomes, then translated these into variables in the FFCWS. Each ranked

¹⁴Missingness in the data set is sufficiently high that listwise deletion is not reasonable.

¹⁵The predictors were automatically selected from other variables in the FFCWS data set that correlated highly with the variable in question (e.g., the same measurement across different waves). Greater detail on this methodology and open source code is available at <https://github.com/annafil/FFCRegressionImputation>.

¹⁶A vignette describing this process is available at <http://bit.ly/2yMUrPd>.

variable received a score from 0 to 100. We mapped these scores onto variable-specific penalty factors, s_j , ranging from 0 to 1, in an inverse linear fashion. As shown in the equations above, these penalty factors were multiplied by λ , meaning that a larger penalty factor would result in stronger

regularization for that coefficient. We considered alternative nonlinear mappings of scores onto values of s_j , but differences in model performance appeared minor. Future work could give this problem a more formal mathematical treatment, potentially within a Bayesian framework.

Appendix D: Results

Table D1. Holdout Scores.

Imputation	Subsetting	Scores	Eviction	GPA	Grit	Job Training	Layoff	Material Hardship
MI	Wiki surveyed	Experts	0.05443	0.36480	0.25285	0.18084	0.16320	0.02268
MI	Wiki surveyed	MTurk users	0.05471	0.37015	0.24998	0.18114	0.16334	0.02265
MI	Wiki surveyed	No scores	0.05465	0.36588	0.25013	0.18206	0.16360	0.02256
MI	Constructed	No scores	0.05533	0.37728	0.24839	0.18237	0.16265	0.02265
LASSO	No subsetting	No scores	0.05491	0.35455	0.24754	0.17909	0.16611	0.02056
LASSO	Wiki surveyed	Experts	0.05497	0.37064	0.25509	0.18220	0.16399	0.02309
LASSO	Wiki surveyed	MTurk users	0.05546	0.37512	0.25064	0.18209	0.16574	0.02318
LASSO	Wiki surveyed	No scores	0.05539	0.36750	0.25139	0.18318	0.16570	0.02300
LASSO	Constructed	No scores	0.05546	0.37793	0.24976	0.18367	0.16467	0.02320
LM-untyped	Wiki surveyed	Experts	0.05507	0.37089	0.25292	0.18246	0.16558	0.02329
LM-untyped	Wiki surveyed	MTurk users	0.05546	0.37540	0.25041	0.18278	0.16704	0.02331
LM-untyped	Wiki surveyed	No scores	0.05537	0.36781	0.25171	0.18377	0.16573	0.02310
LM	No subsetting	Experts	0.05454	0.35425	0.24777	0.17810	0.16379	0.02012
LM	No subsetting	MTurk users	0.05462	0.35444	0.24664	0.17962	0.16431	0.02000
LM	No subsetting	No scores	0.05454	0.35537	0.24468	0.17797	0.16560	0.01999
LM	Wiki surveyed	Experts	0.05481	0.36279	0.25310	0.18355	0.16457	0.02263
LM	Wiki surveyed	MTurk users	0.05464	0.36439	0.25078	0.18306	0.16434	0.02264
LM	Wiki surveyed	No scores	0.05492	0.36387	0.25062	0.18509	0.16462	0.02267
LM	Constructed	No scores	0.05546	0.37797	0.24939	0.18195	0.16255	0.02282
Mean	No subsetting	Experts	0.05446	0.35449	0.24913	0.17771	0.16375	0.02000
Mean	No subsetting	MTurk users	0.05448	0.35258	0.24629	0.17909	0.16427	0.01993
Mean	No subsetting	No scores	0.05424	0.35182	0.24566	0.17704	0.16473	0.01985
Mean	Wiki surveyed	Experts	0.05493	0.36457	0.25324	0.18290	0.16426	0.02267
Mean	Wiki surveyed	MTurk users	0.05477	0.36477	0.25064	0.18311	0.16444	0.02268
Mean	Wiki surveyed	No scores	0.05498	0.36353	0.25033	0.18371	0.16481	0.02269

Note: Each row displays mean squared errors from 1 of the 25 different strategies we used. The first column describes the imputation strategy used. The second describes whether we subsetted the data and, if so, to which set. The third describes which set of scores were used, if any. GPA = grade point average; LASSO = least absolute shrinkage and selection operator; LM = linear model; MI = multiple imputation.

Table D2. Original Holdout Scores.

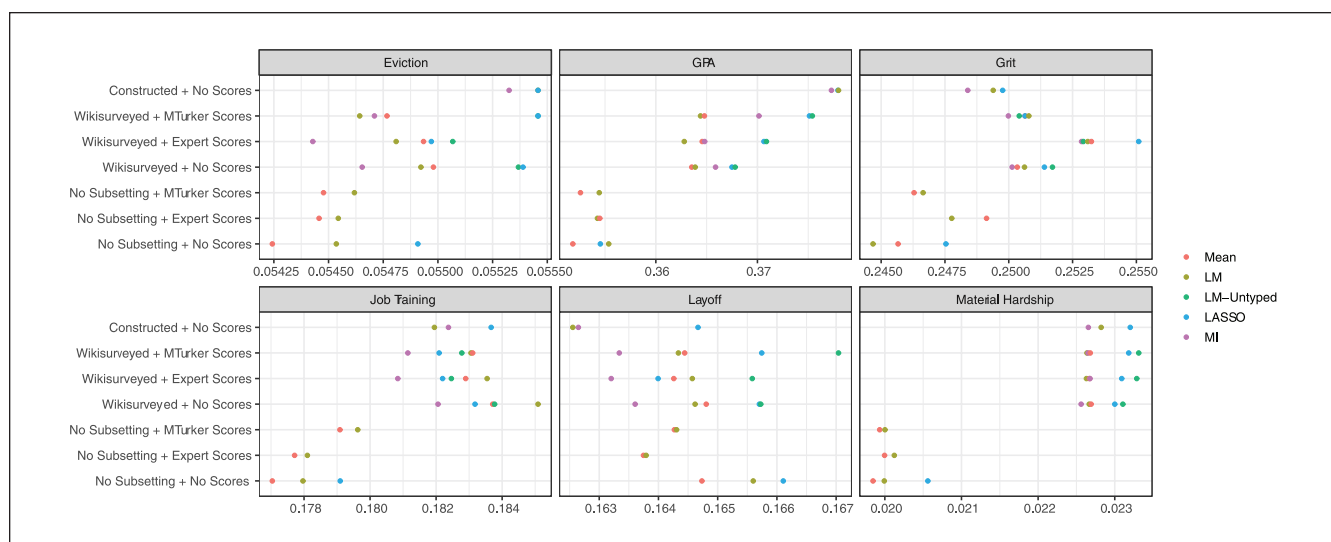
Imputation	Subsetting	Scores	Eviction	GPA	Grit	Job Training	Layoff	Material Hardship
MI	Wiki surveyed	Experts	0.0545	0.3642	0.2522	0.1817	0.1637	0.0226
MI	Wiki surveyed	MTurk users	0.0547	0.3700	0.2498	0.1813	0.1643	0.0226
MI	Wiki surveyed	No scores	0.0547	0.3645	0.2506	0.1827	0.1644	0.0225
MI	Constructed	No scores	0.0548	0.3761	0.2480	0.1826	0.1627	0.0227
LASSO	No subsetting	No scores	0.0543	0.3587	0.2453	0.1795	0.1672	0.0203
LASSO	Wiki surveyed	Experts	0.0555	0.3779	0.2533	0.1825	0.1672	0.0249
LASSO	Wiki surveyed	MTurk users	0.0555	0.3729	0.2537	0.1822	0.1670	0.0248
LASSO	Wiki surveyed	No scores	0.0554	0.3721	0.2517	0.1832	0.1668	0.0241
LASSO	Constructed	No scores	0.0555	0.3803	0.2516	0.1840	0.1662	0.0236
LM-untyped	Wiki surveyed	Experts	0.0550	0.3734	0.2533	0.1822	0.1642	0.0231

(continued)

Table D2. (continued)

Imputation	Subsetting	Scores	Eviction	GPA	Grit	Job Training	Layoff	Material Hardship
LM-untyped	Wiki surveyed	MTurk users	0.0555	0.3804	0.2505	0.1820	0.1656	0.0232
LM-untyped	Wiki surveyed	No scores	0.0554	0.3718	0.2504	0.1834	0.1658	0.0230
LM	No subsetting	Experts						
LM	No subsetting	MTurk users						
LM	No subsetting	No scores	0.0545	0.3551	0.2445	0.1779	0.1654	0.0200
LM	Wiki surveyed	Experts	0.0549	0.3633	0.2531	0.1824	0.1649	0.0226
LM	Wiki surveyed	MTurk users	0.0547	0.3644	0.2509	0.1831	0.1643	0.0226
LM	Wiki surveyed	No scores	0.0549	0.3639	0.2506	0.1851	0.1646	0.0226
LM	Constructed	No scores	0.0555	0.3764	0.2495	0.1818	0.1635	0.0228
Mean	No subsetting	Experts						
Mean	No subsetting	MTurk users						
Mean	No subsetting	No scores	0.0548	0.3521	0.2471	0.1773	0.1648	0.0199
Mean	Wiki surveyed	Experts						
Mean	Wiki surveyed	MTurk users	0.0548	0.3708	0.2509	0.1838	0.1646	0.0228
Mean	Wiki surveyed	No scores	0.0550	0.3658	0.2507	0.1846	0.1650	0.0227

Note: Each row displays mean squared errors (MSEs) from our original submission to the Fragile Families Challenge. To explore an additional four strategies (which are blank here), we obtained new holdout MSEs after the original challenge had closed. These are shown in Table D1. We include these original MSEs as a reference for the interested reader. GPA = grade point average; LASSO = least absolute shrinkage and selection operator; LM = linear model; MI = multiple imputation.

**Figure D1. Raw MSEs from all approaches.**

Note: Mean squared errors (MSEs) for all outcomes and all approaches considered. The best MSE values achieved from the Fragile Families Challenge are 0.052424 for eviction, 0.351820 for grade point average, 0.244684 for grit, 0.177041 for job training, 0.162553 for layoff, and 0.019847 for material hardship.

Acknowledgments

The results in this article were created with software written in R 3.4.3 (R Core Team 2017) using the following packages: glmnet 2.0.13 (Friedman et al. 2010), Amelia 1.7.4 (Honaker et al. 2011), caret 6.0.78 (Kuhn 2017), polywog 0.4.0 (Kenkel and Signorino 2014), Matrix 1.2.12 (Bates and Maechler 2017), doParallel 1.0.11 (Microsoft Corporation and Weston 2017), parallel 3.4.3 (R Core Team 2017), dplyr 0.7.4 (Wickham, François, et al. 2017), forcats

0.3.0 (Wickham 2018a), haven 1.1.0 (Wickham and Miller (2017), labelled 1.0.1 (Larmarange 2017), purrr 0.2.4 (Henry and Wickham 2017), readr 1.1.1 (Wickham, Hester, and François 2017), stringr 1.3.0 (Wickham 2018b), tidyr 0.8.0 (Wickham and Henry 2018), devtools 1.13.5 (Wickham, Hester, and Chang 2018), rmarkdown 1.9 (Allaire et al. 2018), rprojroot 1.3-2 (Müller 2018), ggplot2 2.2.1 (Wickham 2009), plyr 1.8.4 (Wickham 2011), and data.table 1.10.4-3 (Dowle and Srinivasan 2017).

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for the FFCWS was provided by the Eunice Kennedy Shriver National Institute of Child Health and Human Development through grants R01HD36916, R01HD39135, and R01HD40421 and by a consortium of private foundations, including the Robert Wood Johnson Foundation. Funding for the FFC, and for the MTurk fees associated with this project, was provided by the Russell Sage Foundation. Support for the computational resources for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant (P2C HD042828) to the Center for Studies in Demography and Ecology at the University of Washington.

ORCID iD

Kivan Polimis  <https://orcid.org/0000-0002-3498-0479>

Supplemental Material

Supplemental material for this article is available with the manuscript on the *Socius* website.

References

- Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. 2018. "rmarkdown: Dynamic Documents for R." R package version 1.9. (<https://CRAN.R-project.org/package=rmarkdown>).
- Bates, Douglas, and Martin Maechler. 2017. "Matrix: Sparse and Dense Matrix Classes and Methods." R package version 1.2.12. (<https://CRAN.R-project.org/package=Matrix>).
- Bijak, Jakub, and Arkadiusz Wiśniowski. 2010. "Bayesian Forecasting of Immigration to Selected European Countries by Using Expert Knowledge." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173(4):775–96.
- Billari, Francesco C., Rebecca Graziani, and Eugenio Melilli. 2012. "Stochastic Population Forecasts Based on Conditional Expert Opinions." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175(2):491–511.
- Branson, Steve, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. 2010. "Visual Recognition with Humans in the Loop." Retrieved December 8, 2018 (<https://pdfs.semanticscholar.org/b42c/4b804d69a031aac797346acc337f486e4a09.pdf>).
- Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott. 2009. "Handling Sparsity via the Horseshoe." Pp. 73–80 in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*. Retrieved December 8, 2018 (<http://proceedings.mlr.press/v5/carvalho09a/carvalho09a.pdf>).
- Donoho, David. 2017. "50 Years of Data Science." *Journal of Computational and Graphical Statistics* 26(4):745–66.
- Dowle, Matt, and Arun Srinivasan. 2017. "data.table: Extension of 'data.frame.'" R package version 1.10.4-3. (<https://CRAN.R-project.org/package=data.table>).
- Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33(1):1–22.
- Henry, Lionel, and Hadley Wickham. 2017. "purrr: Functional Programming Tools." R package version 0.2.4. (<https://CRAN.R-project.org/package=purrr>).
- Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45(7):1–47.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2017. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.
- Kenkel, Brenton, and Curtis S. Signorino. 2014. "Package 'polywog.'" Retrieved December 8, 2018 (<https://cran.r-project.org/web/packages/polywog/index.html>).
- Kuhn, Max. 2017. "caret: Classification and Regression Training." R package version 6.0.78. (<https://CRAN.R-project.org/package=caret>).
- Larmarange, Joseph. 2017. "labelled: Manipulating Labelled Data." R package version 1.0.1. (<https://CRAN.R-project.org/package=labelled>).
- Lee, Anthony, Francois Caron, Arnaud Doucet, and Chris Holmes. 2010. "A Hierarchical Bayesian Framework for Constructing Sparsity-inducing Priors." *arXiv*. Retrieved December 8, 2018 (<https://arxiv.org/abs/1009.1914>).
- Microsoft Corporation, and Steve Weston. 2017. "doParallel: Foreach Parallel Adaptor for the 'parallel' Package." R package version 1.0.11. (<https://CRAN.R-project.org/package=doParallel>).
- Müller, Kirill. 2018. "rprojroot: Finding Files in Project Sub-directories." R package version 1.3-2. (<https://CRAN.R-project.org/package=rprojroot>).
- Pemstein, Daniel, Kyle L. Marquardt, Eitan Tzelgov, Yi-ting Wang, and Farhad Miri. 2015. "The V-Dem Measurement Model: Latent Variable Analysis for Cross-national and Cross-temporal Expert-coded Data." Varieties of Democracy Institute Working Paper No. 21. Retrieved (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2704787).
- Piironen, Juho, and Aki Vehtari. 2016. "On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior." *arXiv*. Retrieved December 8, 2018 (<https://arxiv.org/abs/1610.05559>).
- Raftery, Adrian E., Nan Li, Hana Ševčíková, Patrick Gerland, and Gerhard K. Heilig. 2012. "Bayesian Probabilistic Population Projections for All Countries." *Proceedings of the National Academy of Sciences* 109(35):13915–21.
- R Core Team. 2017. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. (<https://www.R-project.org/>).
- Reichman, Nancy E., Julien O. Teitler, Irwin Garfinkel, and Sara S. McLanahan. 2001. "Fragile Families: Sample and Design." *Children and Youth Services Review* 23(4–5):303–26.
- Salganik, Matthew J., and Karen E. C. Levy. 2015. "Wiki Surveys: Open and Quantifiable Social Data Collection." *PLoS ONE* 10(5):e0123483.
- Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.
- Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40(1):1–29.
- Wickham, Hadley. 2018a. "forcats: Tools for Working with Categorical Variables (Factors)." R package version 0.3.0. (<https://CRAN.R-project.org/package=forcats>).
- Wickham, Hadley. 2018b. "stringr: Simple, Consistent Wrappers for Common String Operations." R package version 1.3.0. (<https://CRAN.R-project.org/package=stringr>).

- Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2017. “dplyr: A Grammar of Data Manipulation.” R package version 0.7.4. (<https://CRAN.R-project.org/package=dplyr>).
- Wickham, Hadley, and Lionel Henry. 2018. “tidyr: Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions.” R package version 0.8.0. (<https://CRAN.R-project.org/package=tidyr>).
- Wickham, Hadley, Jim Hester, and Winston Chang. 2018. “devtools: Tools to Make Developing R Packages Easier.” R package version 1.13.5. (<https://CRAN.R-project.org/package=devtools>).
- Wickham, Hadley, Jim Hester, and Romain François. 2017. “readr: Read Rectangular Text Data.” R package version 1.1.1. (<https://CRAN.R-project.org/package=readr>).
- Wickham, Hadley, and Evan Miller. 2017. “haven: Import and Export ‘SPSS’, ‘Stata’ and ‘SAS’ Files.” R package version 1.1.0. (<https://CRAN.R-project.org/package=haven>).

Author Biographies

Anna Filippova is a postdoctoral researcher with the Institute for Software Research at Carnegie Mellon University, where she works toward supporting sustainable open collaborative community development, particularly in the context of free and open-source software and Wikipedia communities. She received her PhD from the National University of Singapore. Her research interests include social norms and conflict in virtual environments, inclusive group processes in diverse teams, and the role of face-to-face events in supporting the development of online peer production communities. She has also been involved in organizing free and open-source community events, such as the Abstractions conference and Ruby monthly meet-ups.

Connor Gilroy is a PhD student in sociology at the University of Washington. He studies LGBTQ communities and populations to understand social processes of visibility, acceptance, and assimilation. His current research investigates patterns of sociodemographic change in gay neighborhoods. Additionally, he has projects on improving demographic estimates of queer populations with social media data and on using agent-based models to explore the macro-level impacts of the interpersonal process of coming out as LGBTQ.

Ridhi Kashyap is an associate professor of social demography and fellow of Nuffield College at the University of Oxford. She finished her DPhil in sociology jointly affiliated with the University of Oxford and the Max Planck Institute for Demographic Research in 2017. Her research spans a number of substantive areas in demography and sociology, including gender, mortality and health, the diversification of family forms, and ethnicity and migration. Her work has sought to adopt computational innovations both in terms of modeling approaches such as agent-based models and digital trace data from Web and social media platforms to study social and demographic processes. She is currently leading a Data2X and UN Foundation-supported project that uses big data from the Web, in particular large-scale online advertising data that provide

information on the aggregate numbers of users of online platforms by demographic characteristics, to measure sustainable development and gender inequality indicators.

Antje Kirchner is a research survey methodologist at RTI International and an adjunct research assistant professor at the University of Nebraska–Lincoln. Her research addresses challenges in survey methodology, including ways to examine nonresponse bias using ML techniques, adaptive and responsive design, assessing the quality of survey and administrative data, eliciting and analyzing answers to sensitive questions, detecting problems in the respondent-interviewer interaction, and how to improve response quality in Web surveys using paradata. Her research has been published in journals such as *Public Opinion Quarterly*, the *Journal of Survey Statistics and Methodology*, and *Journal of the American Statistical Association*.

Allison C. Morgan is pursuing her PhD in computer science at the University of Colorado Boulder. She is interested in using data mining, ML, and social network analysis to develop and test hypotheses about the origins and effects of gender imbalance within academia. She is supported by the National Science Foundation’s Graduate Research Fellowship. Prior to graduate school, Allison worked as a data scientist for two years at a small tech startup in Portland, Oregon. She earned her BA in physics from Reed College.

Kivan Polimis is a data scientist at Maana in Houston, Texas. Kivan is interested in structural inequality, natural language processing, and developing programming solutions to social problems. He was previously with Bocconi University’s Center for Social Dynamics and Public Policy and Institute for Data Science and Analytics as a postdoctoral researcher and affiliate. Prior to Bocconi, Kivan was the program coordinator for the University of Washington’s Data Science for Social Good program and a civic technology fellow with Microsoft. Kivan is originally from Saint Lucia and holds degrees in sociology from Princeton University (BA), the University of North Carolina at Chapel Hill (MA), and the University of Washington (PhD).

Adaner Usmani is a postdoctoral fellow at the Watson Institute for International and Public Affairs at Brown University. His dissertation examines the rise and fall of labor movements over the twentieth and early twenty-first centuries and considers the effects of these facts on politics and public opinion. In other work, he has written about American mass incarceration, with an eye on the racial politics of its origins and reproduction.

Tong Wang is an assistant professor of management sciences at the Tippie College of Business, University of Iowa. She received her PhD in computer science from the Massachusetts Institute of Technology in 2016. Her general research interests include interpretable ML and applied data mining, with its application in computational criminology, health care, social marketing, and other areas. Her research on crime data mining is the second place winner in “Doing Good with Good OR” at INFORMS 2015. Her work on crime data mining has been reported in multiple media, including Wikipedia.