When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions

by

Kristen M. Altenburger and Daniel E. Ho*

We make two contributions to understanding the role of algorithms in regulatory enforcement. First, we illustrate how big-data analytics can inadvertently import private biases into public policy. We show that a much-hyped use of predictive analytics – using consumer data to target food-safety enforcement – can disproportionately harm Asian establishments. Second, we study a solution by Pope and Sydnor (2011), which aims to debias predictors via marginalization, while still using information of contested predictors. We find the solution may be limited when protected groups have distinct predictor distributions, due to model extrapolation. Common machine-learning techniques heighten these problems.

Keywords: racial bias, antidiscrimination, predictive targeting, algorithmic fairness

JEL classification code: I18, C53, K23, K42

1 Introduction

While big data holds tremendous promise for social science and public policy (Lazer et al., 2009), one concern lies in whether big data might exacerbate racial and gender bias (Barocas and Selbst, 2016). In 2016, the White House and the U.S. Federal Trade Commission each issued reports warning of the dangers of

* Kristen M. Altenburger: Ph.D. Candidate, Management Science and Engineering, Stanford University, Stanford (CA), USA. Altenburger was supported in part by the Department of Defense through the National Defense Science and Engineering Graduate Fellowship program. Daniel E. Ho (corresponding author): William Benjamin Scott and Luna M. Scott Professor of Law, Professor (by courtesy) of Political Science, Senior Fellow at Stanford Institute for Economic Policy Research, Stanford University, Stanford (CA), USA. We thank Taylor Cranor and Reid Whitaker for valuable research assistance; Jacob Goldin, Sandy Handan-Nader, Daniel Jenson, Gary King, Tim Lytton, Ken Mack, Oluchi Mbonu, Alison Morantz, Johan Ugander, and participants at the Seminar on the New Institutional Economics at the Max Planck Institute for Research on Collective Goods for feedback; Jacob Goldin for suggesting Pope and Sydnor's solution; and Christoph Engel and Krishna Gummadi for valuable discussions of the work. All replication code is available at https://github.com/kaltenburger/Bias.

Journal of Institutional and Theoretical Economics – ISSN 0932-4569
DOI: 10.1628/jite-2019-0001 –

racial discrimination in the use of big data (Executive Office of the President, 2016; Ramirez et al., 2016). The *New York Times* published a op-ed lamenting "Artificial Intelligence's White Guy Problem," citing emerging evidence that the digital economy can perpetuate racial biases (Crawford, 2016).

Online advertisements, for instance, may systematically steer different advertisements to users by race (Sweeney, 2013) and gender (Datta, Tschantz, and Datta, 2015). But because the data and algorithms are typically only available to private companies (King, 2011; Lazer et al., 2009), much less is known about the scope, extent, and mechanism of such bias. In the online-advertisement context, for instance, we do not know whether gender bias (implicit or explicit) is driven by bias of potential employers or because gender groups may differ in click-through patterns (Datta, Tschantz, and Datta, 2015). Altenburger et al. (2017) find that female MBA candidates are less likely to complete unstructured LinkedIn fields, which may affect employment matching algorithms. In addition, while evidence is surfacing of this potential for bias in private economic transactions, much less is known about its influence on public policy. Numerous areas of law and policy are grappling with how to use predictive analytics (GAO, 2004) – most prominently, criminal justice (Berk, 2008), but also taxation (The Department of the Treasury, 2009), health and safety (Kleinberg et al., 2015; Morantz, 2008), and local government (Simon, 2014), to name just a few. And government agencies are increasingly experimenting with using nongovernmental data (e.g., Twitter) for enforcement purposes (see, e.g., Adler, 2016; Grubmüller, Götsch, and Krieger, 2013; Sengupta, 2013; Thompson, 2015).

This article makes two contributions. First, we use the food-safety context to show how "big-data hubris" (Lazer et al., 2014) – the notion that big data can substitute for conventional principles of statistical inference – can allow private bias to migrate into public enforcement. Many have advocated for the use of big-data analytics with consumer data to target regulatory enforcement in food safety (Devinney et al., 2018; Glaeser et al., 2016; Harrison et al., 2014; Kang et al., 2013; Nsoesie, Kluberg, and Brownstein, 2014; Sadilek et al., 2013; Schomberg et al., 2016). As *The Atlantic* provocatively put it, "Yelp might clean up the restaurant industry" (Badger, 2013). While such proposals may seem appealing on the surface, we study whether lay consumers are more likely to issue complaints against Asian than non-Asian establishments. We use (a) New York City inspection data matched with 311 call complaint data and (b) King County (Washington) inspection data matched with Yelp review data to show that complaints and reviews flagging food-safety issues are indeed disproportionately more likely for Asian establishments, holding constant the violation score assigned by food-safety inspectors.

Second, we study a solution proposed recently by Pope and Sydnor (2011) (P&S) that might still allow agencies to deploy such contested predictors (e.g., Yelp reviews), and examine whether this solution may retain predictive accuracy while debiasing regulatory targeting. P&S propose that models use *all* predictors – including socially acceptable predictors (SAPs), contested predictors (CPs) (e.g., Yelp reviews), and socially unacceptable predictors (SUPs) (e.g., race) – but marginal-

ize out SUPs in the predictive step. For instance, in a linear model, a full model with race, complaints, and any other predictors would be fitted, but predicted values would be based on the average race for all units, hence utilizing complaint information orthogonal to race and marginalizing out race. The approach promises "SUP-blind" measures, using information in CPs without proxying for SUP. We study this solution using Monte Carlo simulation and demonstrate important limitations to whether the approach retains predictive accuracy under deviations in the data-generating process. Most importantly, while P&S demonstrated its validity in the linear context with predictors that are independent of SUP group, the approach is more limited when predictor distributions are distinct along SUP lines (e.g., when a predictor is shifted by some mean for one racial group) and when more conventional machine-learning techniques (e.g., decision trees, random forests) are used. Using the New York and King County data, we illustrate how in practice, a restricted random forest model (i.e., one that categorically excludes SUP and CP from the potential feature set and only uses SAPs) can actually outperform P&S's marginalization approach.

Our study makes several key advances over prior work. First, our examination of predictive analytics in food safety is one of the only studies to examine bias in the regulatory usage of algorithms. While extensive scholarship examines these dynamics in criminal justice, much less is known about algorithms and bias in the civil regulatory context. Second, our setting is particularly novel in that inspection data allow us control for food risk at the time complaints or Yelp reviews are made, hence allowing us to credibly assess bias of ordinary consumers. While much research documents racial disparities in algorithmic predictions, it is often much less clear how the alternative (typically, human judgment) fares (see, e.g., Berk and Hyatt, 2015). Third, our study is the first extension beyond P&S to study conditions under which their proposed marginalization can address questions of racial bias outside of the linear (or generalized linear) setting.

We proceed as follows. Section 2 provides background on proposals to target food-safety inspections using consumer data. Section 3 presents tests of whether consumer data (311 calls and Yelp reviews) appear prone to racial bias. Section 4 discusses the solution proposed by P&S, presents Monte Carlo evidence, and applies the approach to New York and King County data. Section 5 concludes with implications. Supplementary material is available online at https://www.mohrsiebeck.com/altenburger-ho.

2 Food Safety and Big Data

The Centers for Disease Control and Prevention estimates that 128,000 Americans annually are hospitalized for foodborne illness (Centers for Disease Control and Prevention, 2011), with the majority of outbreaks attributed to restaurants (Gould et al., 2013). The Food and Drug Administration maintains a model food code, and local jurisdictions bear the principal responsibility for ensuring compliance with

health codes by restaurants. The typical local health department conducts several inspections of each permitted restaurant annually, scoring health-code violations during unannounced visits. Routine inspection frequency is typically determined by permit category, with follow-up inspections for poor performers.

As in other regulatory areas, many have advocated for relying on big data to target health inspection resources. New York City permits reporting of food poisoning in its municipal 311 phone line. The health district of Southern Nevada used Twitter data mentioning "stomach aches" to target inspections (Sadilek et al., 2013). And machine-learning techniques have used consumer review information from Yelp to predict likely violations (Kang et al., 2013; Schomberg et al., 2016). In collaboration with Yelp, New York City's health department used a stream of reviews to conduct investigations (Harrison et al., 2014). Any mention of suspicious terms in reviews (specifically, "sick," "diarrhea," "vomit," and "food poisoning") triggered an investigation by an epidemiology team. Funded in part by Yelp, the City of Boston ran a tournament with Harvard researchers to mine Yelp reviews to predict food-safety violations (Glaeser et al., 2016). And the start-up company "I was poisoned" (https://iwaspoisoned.com/) has attempted to crowd-source food-poisoning complaints.

These applications of predictive analytics from consumer data have tremendous popular appeal. They potentially enable the government to efficiently deploy limited enforcement resources, foster public engagement, and remedy underreporting of foodborne illnesses. NPR, the *New York Times*, *Forbes*, *Newsweek*, and many other news outlets provide glowing support for the idea. Yelp's CEO provocatively, if not entirely disinterestedly, claimed that Yelp could "beat 'gold standard' healthcare measures" (Stoppelman, 2016).

Yet the information source for such approaches lies in ordinary consumers, raising basic questions of data validity and reliability (Lazer et al., 2014; Crawford, 2013). Unlike for food-safety inspections, consumers *elect* whether to contact the city. Yelp reviews cover an unrepresentative subset of King County restaurants and customers (Ho, 2017b). Most importantly, most customers are uninformed about the science of microbial food risk (Wilcock et al., 2004). One common misperception, for instance, is that illness was caused by the most recent place eaten at, which can be inconsistent with the long incubation period of many microbial agents. *E. coli* (O157:H7), for instance, typically has an incubation period of 3–4 days, but may take as long as 10 days to become symptomatic. As articulated by Bill Marler, one of the leading food-safety attorneys, "If you don't have stool or blood culture, it is virtually impossible" to attribute food poisoning (Tomky, 2015).

Numerous commentators have conjectured specifically about preconceptions of Asian cuisine. Andy Ricker – the James Beard Award-winning chef of Thai cuisine, who is white – notes a "widespread misperception in this country that restaurants with white owners are somehow cleaner than others" (Lam, 2012). The food writer Naomi Tomky writes, "People love to blame Asian restaurants for food poisoning" (Tomky, 2015). One of the most vivid examples of bias remains the myth of "Chinese restaurant syndrome" due to MSG (Williams and Woessner, 2009). Using

consumer and social media data hence poses the possibility of importing implicit biases, including racial and ethnic biases, of ordinary consumers into public enforcement. Cavallo, Lynch, and Scull (2014), for instance, document substantial demographic disparities in 311 service requests. In San Francisco and New York, census tracts with a higher African-American population issued fewer 311 request. Similarly, Kontokosta, Hong, and Korsberg (2017) show that minority neighborhoods are substantially more likely to underreport heating and water problems via 311 calls. *Slate* reports that 44 of the first 100 Yelp hits for "poisoning" in Los Angeles were for Asian establishments (Simmons, 2014). And one of the leading studies quite directly deploys the Yelp terms "Vietnamese," "Thai," "Japanese," and "Chinese" as predictors for poor safety (Kang et al., 2013).

While the possibility for bias exists, prior evidence also suggests that food-safety inspection scores are in fact worse at "ethnic," and particularly Asian, restaurants (see, e.g., Harris et al., 2015; Ho, 2017b; Kwon et al., 2010; Roberts et al., 2011). Higher complaint rates may hence reflect underlying food risk. Fortunately, in contrast to other areas where racial disparities may be confounded by unobservable differences (e.g., criminality, productivity), inspection scores offers a direct and independent measure of food risk. Inspector assessments of food risk from unannounced inspections hence allow us to construct a simple test of whether lay judgment – relative to expert judgment – exhibits evidence of racial bias.

3 Bias in Regulatory Targeting: Evidence

3.1 New York City

Our first empirical illustration uses data from New York City. Our data come from (a) health-department data for 77,661 (routine) health inspections of 22,096 establishments from 2012 to 2017, and (b) complaints about food poisoning made through New York's 311 phone line from 2010 to 2017. Because the latter data do not contain establishment identifiers, we standardize addresses across the two data sets and merge complaints to establishments based on unique addresses. We manually classify cuisine descriptions into Asian and non-Asian cuisines, with roughly 19 % of establishments classified as Asian. This results in a merged data set of 66,259 inspections and 2,971 establishment-specific food-poisoning complaints made after a routine inspection (and before the next). Because we are interested in inspection scores as control variables to assess racial bias in complaints, we match complaints to the antecedent routine inspection, as our unit of analysis is the inspection cycle (the period from one routine inspection to the next).[1] Inspection scores represent violation points, such that a score of 0 means perfect compliance and higher scores indicate more violations and, in principle, greater food

[1] In contrast, in section 4.3, we predict future scores using past scores, complaints, and other predictors, so that the outcome in those models is the score *after* a complaint is issued.

Table 1
Descriptive Statistics for Asian and Non-Asian Establishments in New York City

		Asian		Non-Asian		
		Mean	SE	Mean	SE	p-value
Inspections	Score	17.37	0.10	15.64	0.04	< 0.01
	Average prior score	18.02	0.08	15.99	0.04	< 0.01
311 Calls	Complaints (%)	4.79	0.18	3.35	0.07	< 0.01
Borough	Bronx	8.55	0.05	10.26	0.10	< 0.01
	Brooklyn	28.34	0.08	26.56	0.15	< 0.01
	Manhattan	31.28	0.09	37.43	0.17	< 0.01
	Queens	29.51	0.09	22.36	0.14	< 0.01
	Staten Island	2.32	0.02	3.40	0.06	< 0.01
ZIP Code	ZIP 10003	2.72	0.03	2.44	0.05	0.06
	ZIP 10013	3.36	0.03	1.83	0.05	< 0.01
	ZIP 10036	1.01	0.02	2.31	0.05	< 0.01
	ZIP 10019	1.45	0.02	2.18	0.05	< 0.01
	ZIP 10002	3.00	0.03	1.51	0.04	< 0.01
	N	13,553		58,547		

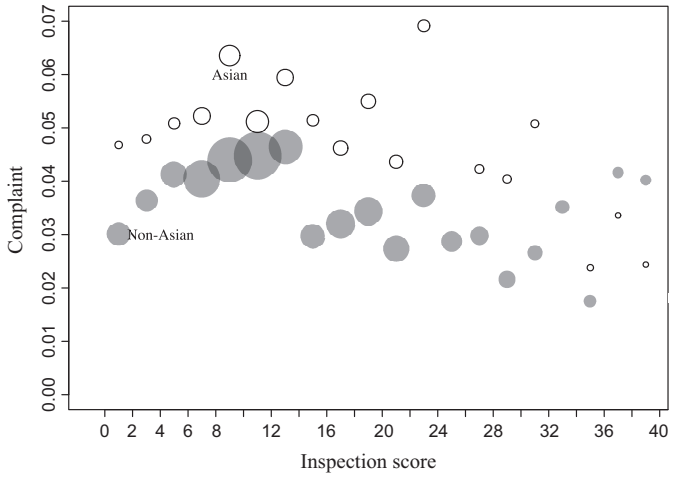
Notes: Unit is the restaurant-inspection cycle. For each geographic region we present the percentages for Asian and non-Asian units in that location. Only the five ZIP codes with more than 1,250 inspections are depicted. Right column presents p-values from a difference-in-means test.

risk. New York uses scores as the basis for letter grades to inform consumers of food-safety risk (Ho, 2012).

Table 1 provides descriptive statistics by Asian and non-Asian establishment of violation scores, average prior violation scores (simple moving average), complaints, boroughs, and ZIP codes. On average, Asian establishments receive more violation points and are subject to more food poisoning complaints. Geographic clustering exists, with a higher percentage of inspections of Asian establishments, for instance, in Queens and ZIP codes in the Lower East Side (e.g., 10002).

To assess whether the complaint disparity is explained by food risk, we examine the conditional probability of a complaint by score, where higher violation scores indicate poorer food-safety performance in the preceding inspection. Figure 1 plots violation scores (binned for visibility) on the x-axis against the frequency of a food poisoning complaint on the y-axis. Gray dots represent non-Asian establishments, and black hollow dots represent Asian establishments, with dots proportional to sample size. Conditional on the same score, Asian establishments are more likely to be targets for complaints. If anything, scores appear negatively correlated with complaints. Table 2 presents regression coefficients from models that sequentially add fixed effects next to controls for inspection scores. Across all models, Asian establishments appear substantively more likely to be subject to complaints. As the

Figure 1
Inspection Score and 311 Complaint



Notes: This figure plots inspection scores (binned for visibility) on the x-axis against the probability of a 311 food poisoning complaint on the y-axis, separately for Asian and non-Asian restaurants in hollow and gray, for New York City. Dots are weighted by sample size, and the x-axis is truncated at 40 for visibility.

baseline rate of complaints is low, the 0.016 increase is large as a relative matter: on average, Asian establishments are subject to 42 % more complaints than non-Asian establishments.

3.2 King County

Our second illustration uses data from King County, the most populous county in Washington state, home to Seattle. We study a publicly available data set of food-safety inspections of 1,756 Seattle restaurants, matched with 13,299 professional food-safety inspections, and 152,153 Yelp reviews from 2006–2013 (Kang et al., 2013). We again hand-code cuisines as Asian and non-Asian, resulting in roughly 28 % classified as Asian establishments.

Table 3 presents descriptive statistics by type of establishments. As before, the unit of analysis is the restaurant-inspection cycle. Asian establishments fare substantially worse on health inspections, are less likely to have Yelp reviews written, and receive slightly lower Yelp evaluations. We again see evidence of geographic clustering, with inspections of Asian restaurants more likely, for instance, in Seattle's International District (98104). On average, there is a 6.7-point differential across inspections between Asian and non-Asian establishments (p-value < 0.001). The difference in New York is substantively smaller (see the online appendix for

Table 2
Linear Models of Number of Complaints as Dependent Variable in New York City

	Model 1	Model 2	Model 3	Model 4	Model 5
Asian	0.016***	0.016***	0.016***	0.020***	0.020***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Score		−0.043**	−0.032*	−0.039**	−0.038**
		(0.013)	(0.013)	(0.013)	(0.013)
Average prior score		0.032	0.023	0.025	0.026
		(0.017)	(0.017)	(0.016)	(0.016)
N	72,100	72,100	72,100	72,100	72,100
Year FE	no	no	yes	yes	yes
Borough FE	no	no	no	yes	no
ZIP FE	no	no	no	no	yes
Parameters	2	4	9	13	218
R^2	0.001	0.001	0.033	0.042	0.047

Notes: Score and average-prior-score coefficients represent increase associated with 100-point increase, for readability. FE indicates fixed effects. Cluster-robust standard errors to account for dependence within establishment are presented in parentheses. Results are comparable using a count model. *, **, *** indicate statistical significance at 0.05, 0.01, and 0.001.

quantile–quantile plots). King County's ethnic difference is a matter of considerable controversy (Ho, 2017b).

To construct a test of bias, we deploy the same search terms that New York City identified in Yelp reviews to target health-department resources ("food poisoning," "vomit," "sick," and "diarrhea"). Replicating New York's targeting strategy with King County allows us to examine the implications of an actual public-health intervention, avoiding difficulties with search-term selection. Table 3 shows Asian establishments are likelier targets for suspicious terms (see the online appendix for counts of suspicious search terms).

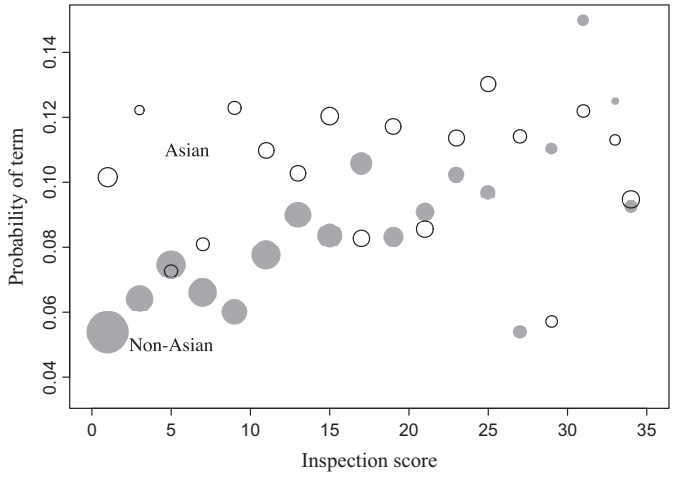
Figure 2 displays the data. The x-axis presents the inspection score, and the y-axis presents the probability of a suspicious search term. For visibility, data points are binned in two-point intervals, weighted by the number of observations, in gray for non-Asian and black hollow for Asian establishments. For instance, the leftmost gray dot represents 1,818 non-Asian establishment-inspection cycles with average violation scores between 0 and 2 points, with 5.4 % subjected to suspicious search terms. Conditional on the same inspection history, Asian establishments are systematically more likely to be subjected to suspicious reviews. A non-Asian establishment with an average inspection score has a 7.7 % probability of a suspicious term, but that rate increases to 9.8 % for Asian establishments with the same inspection score (p-value < 0.001). This 2.1-percentage-point increase represents a substantial increase of 28 % relative to baseline.

Table 3
Descriptive Statistics for Asian and Non-Asian Establishments in King County

		Asian		Non-Asian		
		Mean	SE	Mean	SE	p-value
Inspections	Score	16.56	0.29	9.84	0.14	< 0.01
	Average prior score	17.98	0.18	10.67	0.10	< 0.01
Yelp	Number of reviews	10.17	0.26	12.00	0.20	< 0.01
	Five-star rating	3.60	0.01	3.64	0.01	0.03
	Suspicious terms	9.90	0.47	7.40	0.27	< 0.01
ZIP Code	ZIP 98101	7.81	0.13	12.71	0.25	< 0.01
	ZIP 98104	21.23	0.21	6.65	0.18	< 0.01
	ZIP 98105	10.76	0.15	7.86	0.20	< 0.01
	ZIP 98103	7.00	0.13	9.30	0.21	< 0.01
	N	4060		9239		

Notes: Only ZIP codes with more than 1,000 units are listed.

Figure 2
Association between Food Inspection Violation Score and Probability of Terms Indicative of Food Poisoning in King County



Notes: The x-axis represents the average inspection score received by an establishment prior to a Yelp review, with higher values indicating more violations. The y-axis represents the probability of a suspicious term. Dots are binned at 2-point intervals and weighted by the number of observations. Gray dots represent non-Asian establishments and hollow dots present Asian establishments. For visibility, the x-axis is truncated at 35 points.

Table 4
Logistic Regression Estimates with Dependent Variable of whether Suspicious Search Term Appeared in Review for that Inspection Cycle in King County

	Model 1	Model 2	Model 3	Model 4	Model 5
Asian	0.32***	0.44***	0.38***	0.40***	0.39***
	(0.09)	(0.08)	(0.09)	(0.09)	(0.09)
Prior score		0.00			
		(0.00)			
Avg. prior score			0.01***	0.01***	0.01***
			(0.00)	(0.00)	(0.00)
Review count		0.03***	0.03***	0.03***	0.03***
		(0.00)	(0.00)	(0.00)	(0.00)
Review rating				−0.49***	−0.48***
				(0.04)	(0.04)
Year FE	no	yes	yes	yes	yes
ZIP FE	no	no	no	no	yes
Parameters	2	11	11	12	41
N	13,299	13,299	13,299	13,299	13,299

Notes: Standard errors (using a pairs-cluster-bootstrapped t-statistic with 1,000 bootstrapped replicates to allow for dependence within establishment) are presented in parentheses (Cameron, Gelbach, and Miller, 2008; Esarey and Menger, 2018). *** indicates statistical significance at the 0.001 level. FE indicates fixed effects.

To formally test whether Asian establishments are subjected to more suspicion, Table 4 presents logistic regressions of the probability of a suspicious term submitted during an inspection cycle. Model 1 tests for the raw difference, and models 2–5 sequentially add controls for inspection history, Yelp attributes, and year/ZIP code fixed effects. Models 3–5 confirm that suspicious terms are positively correlated with more violations, but Asian establishments are persistently subjected to a higher frequency of suspicious terms. In sum, Asian establishments are targets for more suspicious terms, holding constant inspection violation scores prior to when reviews were written.

3.3 Interpretation

We now discuss several points of interpretation. First, in principle our evidence could also suggest that health inspectors are biased *in favor* of Asian restaurants. If so, consumer complaints would be useful not just for targeting public resources, but also for identifying potential sources of bias of government officials. This mechanism, however, seems unlikely. Inspectors are full-time employees trained in food-safety principles and observe actual risk factors in the kitchen, such as temperature controls, evidence of cross-contamination, and employee hand-washing

practices. While substantial inter-inspector variability exists (Ho, 2012) and King County inspectors have substantial disagreements around the inspection of "ethnic" establishments (Ho, 2017b), as a relative matter, consumers are undoubtedly less versed in food safety. For one thing, ordinary consumers harbor substantial misconceptions about attributing foodborne illness (Wilcock et al., 2004). Moreover, customers typically observe only conditions in the dining room, but food risk emanates predominantly in the kitchen. Indeed, the reliance on complaint-initiated inspections is controversial, precisely because of the weak informational basis of consumer complaints (Goodin and Klontz, 2007, finding evidence that consumer complaints were *negatively* correlated with critical violations).

The more likely mechanism is that 311 callers and Yelp reviewers exhibit forms of implicit bias. Reviews corroborate this mechanism. Some attributed food poisoning to Asian restaurants inconsistently with incubation periods (e.g., "I got food poisoning right after I ate"). Wrote another: "I usually have such a difficult time digesting Chinese food, however, the food here was different. It was edible and it was good […] I had been looking for a place that served 1. Americanizedish chinese food and 2. didn't make me feel sick." One reviewer opined, "a Mongolian grill […] can also be a breeding ground for food poisoning." Some bias was more express: "I expect all restaurants in Chinatown to be dirty." And droves of other Yelp reviews confirm race-based conceptions in reviews: "The staff was also pretty friendly for an Asian restaurant. Not to sound racist, but there is a reputation for very cold service at times from these places." "[T]he service in general was slow and inattentive (like most Asian restaurants I've visited – sorry if this comes across as borderline racist, but my experience is my experience)." These results are consistent with previous evidence that Yelp reviewers can exhibit racist views (Zukin, Lindeman, and Hurson, 2015) and that 311 calls exhibit racial bias when divergent cultures collide (Legewie and Schaeffer, 2016).

Second, the presence of Yelp search terms does not necessarily mean the reviewer subjectively believed she received food poisoning from the establishment. However, even the mention of sanitation as a concern can be indicative of implicit bias (e.g., one review volunteered that "i was afraid that i would get food poisoning"). And recall that we deployed the exact same search terms that New York used to trigger investigations (Harrison et al., 2014). The false-positive rate and subjectivity of such terms should give one pause about naive deployment of big data in public enforcement.

Third, we do not observe which establishments may be engaged in more actively "managing" online reputations (e.g., soliciting favorable reviews, hiring an online reputation management company). The disparity may result from Asian establishments disproportionately being less willing to engage in such reputational management. Even so, differential ability to manage online reputations calls into question the use of social media data for public enforcement.

Last, Lehman, Kovács, and Carroll (2014) find that "authenticity," particularly of "ethnic" cuisine, may cause food-safety concerns to recede in the minds of consumers. The mechanism might imply that we should be less likely to be able to

detect an effect, as consumers are more likely to decline to comment on food safety of "authentic" cuisine. Our findings, however, suggest that racial bias at the very least has a more dominant effect on reviews. As a policy matter, the implications of an authenticity norm are complicated: authenticity may weaken incentives for establishments to take remedial measures, therefore undercutting the efficacy of proposals for algorithmic targeting.

Regardless of the exact mechanism, our results show that targeting inspections based on complaints would disproportionately burden Asian establishments.

4 A Potential Statistical Debiasing Solution

4.1 Marginalizing SUP

Given the pervasiveness of concerns of racial and gender bias, are there available statistical methods to nonetheless deploy algorithmic targeting in regulatory enforcement? Many solutions have been proposed (Feldman et al., 2015; Hardt, Price, and Srebro, 2016; Kamiran, Žliobaitė, and Calders, 2013; Zafar et al., 2017). We study one proposal in the economics literature from Pope and Sydnor (P&S) that addresses the specific problem of contentious predictors, such as consumer complaints, that may partially proxy for race. P&S first consider predictors to fall into three classes, as prespecified by the researcher or government agency: (1) "socially acceptable predictors" (SAPs) that are uncontroversial and hence socially acceptable (e.g., violation score); (2) "socially unacceptable predictors" (SUPs) that are unacceptable for legal or moral reasons (e.g., race); and (3) "contentious predictors" (CPs) that may contain valuable information to improve accuracy, but that may partially proxy for SUPs (e.g., Yelp complaints). For the moment, we note that characterizing predictors as SAPs, SUPs, or CPs may not be obvious; we spell this out further in section 5.

P&S then posit the following data-generating process (DGP), meeting the usual assumptions of ordinary least squares (OLS):

$$y_i = \beta_0 + \beta_1 \mathrm{X}_i^{\mathrm{SAP}} + \beta_2 \mathrm{X}_i^{\mathrm{CP}} + \beta_3 \mathrm{X}_i^{\mathrm{SUP}} + \varepsilon_i,$$

where y_i is the observed outcome, $\mathrm{X}_i^{\mathrm{SAP}}$ is the vector of SAPs, $\mathrm{X}_i^{\mathrm{CP}}$ is the vector of CPs, $\mathrm{X}_i^{\mathrm{SUP}}$ is the vector of SUPs for unit i, and $\varepsilon_i \sim N(0,\sigma)$. SAPs are assumed independent of SUPs and CPs, but SUPs are correlated with CPs:

$$\mathrm{X}_i^{\mathrm{SUP}} = \delta_0 + \delta_{\mathrm{CP}} \mathrm{X}_i^{\mathrm{CP}} + \mathrm{v}_i$$

with $\mathrm{v}_i \sim N(0,\tau)$. P&S examine the predictive accuracy of four OLS approaches to dealing with CPs: (1) the full model that controls for SAP, CP, and SUP; (2) a "common" model that controls only for SAP and CP; (3) a "restricted" model that controls only for SAP; and (4) a "proposed" method that fits the full model, but uses the average SUP value to calculate predictions, hence marginalizing out SUPs. For the

remainder of this paper, we will refer to this "proposed" approach as the *P&S approach* or *marginalization.* In the P&S approach, fitting the full model would eliminate the influence of SUPs as proxied by CPs (as in the common model), but would retain predictive power of CP information orthogonal to SUPs. P&S prove that in the OLS case, marginalization provides predictive accuracy that is higher than the restricted method (but lower than the common method).

P&S also provide simulations to support extending marginalization to the nonlinear context using a probit DGP. Their simulations vary the strength of δ_{CP}, β_1, and β_2, calculating prediction error across modeling approaches. They find that predictive accuracy remains higher with marginalization than the restricted approach, and hence conclude that marginalization "can be used mechanically with a range of estimation models."

We address whether marginalization is generalizable to more common machine-learning algorithms. If so, the approach could have important consequences for a wide range of predictive problems in public policy, including the food-safety context. It is important to acknowledge that our analysis adheres to P&S's focus on predictive accuracy.[2] Deep normative questions can be raised about the goal of retaining predictive accuracy with CPs. If outcomes themselves exhibit bias, any improvement in predictive accuracy will increase discrimination (Barocas and Selbst, 2016, p. 720–722). These are important questions, but we focus here on whether the P&S approach generalizes on its own terms.

4.2 Monte Carlo Evidence of Extrapolation

We investigate the applicability of P&S's approach to a typical situation in predictive analytics. First, while P&S's approach assumes that SAPs are statistically independent of SUPs and CPs, in practice this is unlikely to be the case. A commonsense classification of "socially acceptable" does not necessarily imply statistical independence. Many predictors that may superficially seem "socially acceptable" are in fact highly correlated with race (see, e.g., Barocas and Selbst, 2016, pp. 695 and 721; Corbett-Davies et al., 2017; Kamiran, Žliobaitė, and Calders, 2013, § 2.2; Dettling et al., 2017, reporting a mean household income of $123k for white respondents and $54k for black respondents; Board of Governors of the Federal Reserve System, 2007, documenting average TransUnion credit scores of 54 for white individuals and 25.6 for black individuals). Indeed, the problem is so common that P&S's own SAPs exhibit sharp distributive differences: white individuals were twice as likely to earn college degrees and three times as likely to earn graduate degrees as black individuals; and men were seven times as likely to be in construction or agriculture and half as likely to be in public administration as women (Pope and Sydnor, 2011, p. 220). Predictors that are highly correlated with SUP are present too in food safety, as shown by substantial differences between Asian and non-Asian establishments in Tables 1 and 3. Even when marginal

[2] To be sure, P&S are also concerned about omitted-variable bias with the common approach.

distributions are not that distinct by SUP group, imbalance and nonidentical support can emerge in higher dimensions of the predictor space. If such distributional differences imply that these are not SAPs, the P&S approach might have limited practical application.[3]

Second, we use models that are dominant in predictive analytics, such as decision trees and random forests. P&S's approach assumes that the researcher has correctly specified the linear (or generalized linear) model. Yet one appeal of dominant machine-learning algorithms is that they weaken functional-form assumptions (e.g., by permitting feature selection, nonlinearities, and higher-order interactions). In practice, nonlinear methods such as decision trees and random forests are much more commonly deployed for predictive analytics.

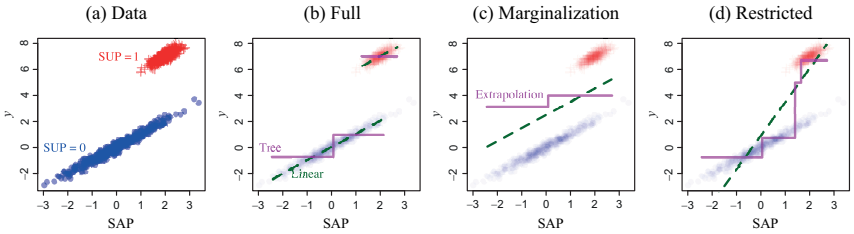
We consider variations on the P&S setup, principally by evaluating the consequence of (a) differences in SAP support between SUP classes (i.e., when SAP is not independent of SUP), and (b) more complex functional forms (leading to model misspecification in the linear context). We analyze and compare the predictive performance for OLS and random forest (RF) models (Breiman, 2001). We first illustrate the problem of extrapolation with a simple example. We then generalize to a more comprehensive simulation to understand the relative accuracy of marginalization for OLS and RFs under different conditions.

Extrapolation with OLS and Decision Trees. Figure 3 demonstrates performance differences when there is a lack of overlap in the SAP support, using simulated data. The left panel plots data, with crosses and dots distinguishing two SUP groups. The data meet all of the P&S DGP assumptions (linearity, additivity, parallel slopes), except for one: the support of the SAP distribution differs by SUP group, with no SAP values below 1 when SUP $= 1$. This simulates the common setting where sharp preexisting demographic differences may exist across groups.

The second panel from the left plots predicted values using OLS (dashed lines) and a decision tree (solid lines). As expected, the regression lines fit well and the tree approximates the regression lines with a step function. The third panel from the left plots the predicted values marginalizing out SUP. In the OLS context, one can marginalize out the SUP by simply predicting with the population average of SUP. In the nonlinear decision-tree context, marginalization happens by predicting values for each unit at the actual and the counterfactual SUP value, and then using a weighted average based on the population mean SUP. The dashed line plots regression-predicted values for the mean SUP value, which, expectedly, fall between the SUP groups. The decision tree, however, exhibits poorer performance: because the full tree immediately splits on SUP, it does not differentiate outcomes along SAP for the group with SUP $= 1$. This means that all counterfactual predictions for dots (observed SUP $= 0$) are the (high) average outcome values when SUP $= 1$. As SAP values decrease, the marginalization extrapolates more. In con-

[3] If all SAPs were classified as CPs, one might fit a model with SUPs and CPs only and marginalize out SUPs. No restricted model would be available.

Figure 3
Illustration of How Decision Tree Can Magnify Extrapolation



Notes: The left panel presents data (dots for SUP = 0 and crosses for SUP = 1), and each of the subsequent panels plots predicted values based on a linear model (dashed) and decision tree (solid), with the full model, P&S's marginalization, and the restricted approach.

trast, OLS relies on linearity and additivity to make counterfactual predictions for observations with lower SAP values, so the extrapolation is less severe.

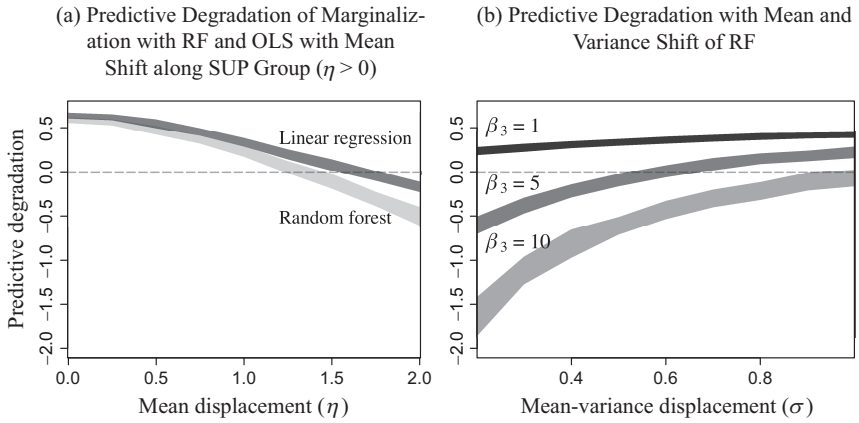
The rightmost panel plots predicted values for restricted linear and tree models that utilize only SAP values. While the linear model is misspecified from a parameter estimation perspective, its predictive accuracy is higher than with marginalization. The tree step function similarly predicts values much closer to the observed values.

In sum, this simple example demonstrates two points. First, when there is nonidentical support – even when linearity, additivity, and all other features of the P&S DGP hold – the restricted model can outperform marginalization. Second, machine-learning techniques can exacerbate extrapolation with marginalization. Nonparametric techniques have the virtue of being able to detect higher-order interactions, but have the liability of potentially generating more error when predicting counterfactual values, precisely because outcomes are distinct along SUP lines.

Extrapolation with Random Forests. We now examine more general conditions under which the restricted model can outperform the P&S approach. We start with the same DGP as P&S (with scalar SAP, CP, and SUP) but allow for (a) nonidentical support with mean and variance shift parameters η and σ, and (b) nonlinearity in β_4. These parameters are shown in boxes below to illustrate how our DGP differs from that of P&S:

$$
\begin{aligned}
X_i^{\mathrm{CP}} &\sim N(0,1);\\
X_i^{*\mathrm{SUP}} &= \frac{1}{1+\exp[-(\delta_0+\delta_{\mathrm{CP}}X_i^{\mathrm{CP}}+v_i)]};\\
X_i^{\mathrm{SUP}} &= \begin{cases} 1 & \text{if } X_i^{*\mathrm{SUP}} > 0.5,\\ 0 & \text{otherwise;} \end{cases}\\
X_i^{\mathrm{SAP}} &\sim N\left(\boxed{\eta}\times X_i^{\mathrm{SUP}}, \boxed{\sigma}\times X_i^{\mathrm{SUP}} + (1-X_i^{\mathrm{SUP}})\right);
\end{aligned}
$$

Figure 4



Notes: Baseline parameter values are $\beta_1 = 1, \beta_2 = 1, \beta_3 = 5, \sigma = 1$ (left), $\delta_0 = 0, \delta_{\text{CP}} = 1$, and $\eta = 1$.

$$y_i = \beta_0 + \beta_1 X_i^{\text{SAP}} + \beta_2 X_i^{\text{CP}} + \beta_3 X_i^{\text{SUP}} + \boxed{\beta_4}\left[X_i^{\text{SUP}} \times \left((X_i^{\text{SAP}} + 2)^2\right)\right] + \varepsilon_i$$

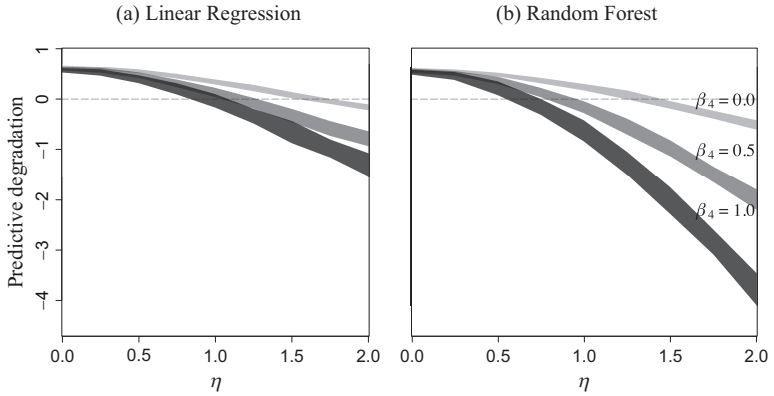
with $\varepsilon_i \sim N(0,1)$ and $v_i \sim N(0,1)$. For each set of parameter values, we draw 100 simulated data sets ($N = 10{,}000$), tune hyperparameters on the training data via cross-validation (see the online appendix), fit the model on a random 80 % subset of observations, and report results on the remaining 20 % test set. For all simulations, we report the predictive performance difference between the restricted and P&S approaches, measured by the difference in root mean squared error (RMSE). As shorthand, we refer to this difference as the *predictive degradation,* where < 0 means that the restricted outperforms the P&S approach.[4]

The left panel of Figure 4 shows that as the mean for one SUP group increases (as η departs from 0), the restricted model outperforms the P&S approach for both RF and OLS. This performance gap is slightly worse with the RF model than with OLS as η increases. The right panel shows that the restricted model also outperforms the P&S model as the variance for one SUP group (σ) decreases. As SUP feature importance increases ($\beta_3 > 0$), the performance gap between the restricted and P&S approaches increases.

Next, we examine the effect of the simple polynomial term ($\beta_4 > 0$), as one basic rationale for using nonparametric techniques is to account for nonlinearities by SUP class. While the left panel of Figure 4 suggests that nonoverlap has to be substantial (η between 1.5 and 2) to see performance degradation for marginalization, the left panel of Figure 5 shows that nonlinearities cause such degradation at lower

[4] Predictive degradation $:= \text{RMSE}_{\text{Restricted}} - \text{RMSE}_{\text{Marginalization}}$.

Figure 5
Effect of Mean-Displacement (η) across SAP–SUP Interactions ($\beta_4 > 0$)



levels of displacement. We observe that when η is between 0.5 and 1, the P&S approach can fare worse than the restricted approach.

The online appendix presents other simulations showing that (a) as the correlation between SUP and CP increases (δ_{CP}), the performance gap between the restricted and P&S approaches decreases, and (b) class imbalance, when $|\delta_0| > 0$, generally improves the accuracy of marginalization.

To summarize, our simulations offer several lessons for the P&S approach. First, when the influence of SUP is strong and there are interaction terms, the marginalization solution is less likely to work for standard machine-learning algorithms. This limitation is important, because the questions of bias are typically of most concern when the effect of SUP is strong (e.g., criminal justice, employment, regulatory enforcement) and predictive analytics frequently grapple with complex interactions. Second, and relatedly, substantial distributional differences along SAP dimensions may erode the performance of marginalization. This problem of extrapolation, which can be exacerbated with nonparametric machine-learning techniques, can be acute, as substantial racial differences, for instance, may exist across many potential predictors. Third, if the marginal predictive power of CP in the full model is low, as is the case in both the King County and New York data, the restricted approach may actually be preferred. The ideal situation for marginalizing out SUP is when (a) the effect of SUP is low, (b) the effect of CP is high, and (c) the correlation between SUP and CP is low.

4.3 Application to New York and King County

We now examine how marginalization performs with New York and King County to potentially enable the use of 311 call and Yelp data, while ensuring that predictions are "SUP-blind." Consistent with our simulations, we find that when SUP has

a strong association with outcomes, as is the case in King County, marginalization may generate lower predictive accuracy than the restricted approach.

Our goal is to predict future inspection scores with high accuracy, but without importing private biases in 311 calls and Yelp reviews. For New York, we consider previous inspection scores and borough/ZIP code as SAPs,[5] whether an establishment is Asian establishment as an SUP, and 311 calls as a CP (per the analysis in section 3). For King County, we similarly categorize SAPs and SUP, adding Yelp review count and rating as SAPs,[6] and treat a Yelp review mentioning suspicious terms ("sick," "vomit," "diarrhea," or "food poisoning") as a CP. Other studies advocating for targeting with Yelp have similarly used cuisine, Yelp reviews, and inspection history as predictors (e.g., Kang et al., 2013).[7] Controlling for inspection history may be motivated from the perspective that follow-up inspections are already scheduled in most jurisdictions based on poor routine inspection results.

We apply the P&S approach for both OLS and RF models, and compare the predictive accuracy with marginalization (model fit on all predictors and predictions based on marginalizing SUP) and restricted (SAP-only) approaches. To avoid overfitting, for each of 100 iterations, we randomly select 80 % of the data, tune hyperparameters via cross-validation, fit the model, and report results on the 20 % test data set.

As expected, RF models improve predictive accuracy substantially relative to OLS (see the online appendix). Figure 6 plots performance degradation (restrictive RMSE versus RMSE from marginalization) on the testing data for 100 simulations for New York (left) and King County (right). Performance is comparable for New York, with a mean RMSE difference between marginalization and the restricted approach of 0.003. For King County, however, the restricted model on average outperforms marginalization, with an average RMSE difference of -0.017.

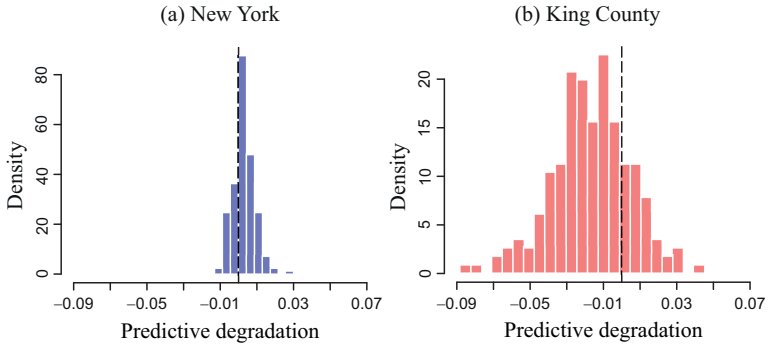
To understand this difference, we calculate feature importance for each application on a single train–test split (see the online appendix). In both applications, inspection history is by far the most important predictor. Asian splits can occur early in the decision tree, which heightens problems of extrapolation. The relative feature importance of consumer 311 and Yelp complaints also differs. In King County, although they are in fact used as predictors in RF trees, the Yelp terms provide very relatively little predictive power. In New York, the presence of a 311 complaint appears more substantial, ranking within the borough identifiers in predictive power. If anything, these findings seem to suggest that notwithstanding all of the media hype and scholarly attention, Yelp reviews add very little usable information.

[5] As we note below, ZIP codes could also plausibly be considered CPs, which illustrates the difficulty of classifying predictors by SAP, CP, and SUP.

[6] Ratings may obviously be considered CPs as well.

[7] Other studies use establishment neighborhood and ZIP code as a predictor. See, e.g., https://www.drivendata.org/competitions/5/keeping-it-fresh-predict-restaurant-inspections/page/28/. When inspectors are assigned by area, these may largely proxy for inspection history, as inspector differences explain much of the variability in violation scores (Ho, 2017a).

Figure 6
Predictive Degradation for RF Models in New York and King County



Note: Dashed vertical line at 0 represents comparable predictive performance between marginalization and the restricted approach.

5 Implications

We close with several implications of our study. First, as a policy matter, our evidence does not provide support for the popular enthusiasm for crowdsourcing food safety from social media. Our analysis in section 4.3 shows that the marginal predictive power of 311 calls and Yelp review terms is trivially small. The average relative percentage decrease in RMSE is essentially 0 % for adding Yelp terms in King County and 0.02 % for adding complaints in New York.[8] Moreover, the notion of "accuracy," as measured by subsequent inspection scores, is not straightforward. In King County, inspectors are principally assigned by ZIP code and there are substantial differences in stringency across inspectors. As a result, allocating more inspections to areas where more violations are scored solely because the inspector is strict may not be deemed "accurate." More generally, to the extent that there is racial bias in inspection scores, higher accuracy may simply replicate that bias.

Second, the food-safety context is particularly novel, at least in contrast to other applications of predictive analytics, because formal equality exists as a baseline before the algorithm is introduced. Typically, by law, establishments within the same permit category are required to be visited at the same frequency.[9] This stands in sharp contrast to many other areas, where human discretion pervades business-as-usual (e.g., judicial discretion in criminal justice). Because the predictive gain

[8] We evaluate the relative percentage decrease in RMSE from a model that includes all predictors except the CPs (i.e., Yelp terms for King County and complaints for New York) and compare RMSE across 50 train–test splits.

[9] Establishment permit categories may be based on the level of food preparation, which determines the inspection frequency. Due to limited on-site food preparation, Starbucks, for instance, may be in a different permit category than a full restaurant. For full food preparation, the baseline frequency is typically constant. In King County, for instance, all food establishments with full food preparation are classified as "risk III" and

is so low relative to a baseline of formal equality, algorithmic targeting may not warrant the cost in disparate impact.

Third, while governments should be applauded for bringing data science and predictive analytics to regulation, our findings demonstrate that naive analytics using consumer complaints can import private bias into public policy. Our findings underscore the need for predictive analytics to grapple more seriously with the institutions whose problems they purport to solve. Our study also underscores the need for greater transparency and data-sharing policies for scientists and government agencies to be able to study, understand, and remedy the distributive implications of predictive analytics (King, 2011; Lazer et al., 2009; Pasquale, 2015).

Fourth, a serious case can be made that if private data are used for public enforcement, the same scrutiny that attaches to government records should apply (but see *Loomis v. Wisconsin*, 881 N.W.2d 749 (Wis. 2016), cert. denied 137 S.Ct. 2290 (2017)). For public enforcement, validity and reliability of inputs are particularly important. Governments and scientists must address conventional questions of statistical inference: Are the data representative of the target population? Is the information valid and reliable? What biases might exist? Companies like Yelp or Twitter can of course serve only a subset of New York with a proprietary algorithm; but the New York government cannot. Conflicts of interest may be acute when an entity like Yelp controls the data and funds studies to promote its usage in public enforcement.

Fifth, our analysis has shown that P&S offers a way forward for governments to purge contentious predictors of bias. The formalization by P&S clarifies assumptions under which predictors can be used. On the other hand, our analysis reveals limitations: when groups are distinct along SUP lines, predictive accuracy may not in fact improve with marginalization. When the distribution of SAPs is distinct along SUP lines – as is very common when examining predictor differences along gender or racial lines (see, e.g., Tables 1 and 3) – SAPs themselves may proxy for race. Substantively, the application of P&S requires classification of predictors into either "socially acceptable," "socially unacceptable," or "contentious" predictors. Yet such classification can be highly contested.[10] Are inspection scores in fact "socially acceptable" when inspectors themselves may be affected by implicit bias? Are ZIP codes SAPs, CPs, or SUPs when a ZIP code may comprise Chinatown? The point here is much the same as that made about causal inference with immutable characteristics (Holland, 1986): because race and gender may affect ev-

receive three visits per year. To be sure, front-line inspectors may not be able to make all visits in a year and hence exercise some discretion in visits, but as a formal regulatory matter, routine inspection frequency between comparable establishments is typically the same.

[10] For a proposal to resolve these questions by asking survey respondents about the fairness of feature selection, see Grgić-Hlača et al. (2016). Kontokosta, Hong, and Korsberg (2017) posit a notion of counterfactual fairness that draws on a causal model of race/gender. Causal graph models can help to make assumptions explicit; they also acknowledge how strong and contested those assumptions can be.

erything, settling on pretreatment covariates (or socially acceptable predictors) is challenging to say the least.

In that situation, one should be cautious about purely technical solutions to debias algorithms (Campolo et al., 2017), many of which rely on a determination of "legitimate" predictors (see, e.g., Corbett-Davies et al., 2017; Hardt, Price, and Srebro, 2016; Kamiran, Žliobaitė, and Calders, 2013). Are prior convictions "legitimate" predictors for recidivism (Corbett-Davies et al., 2017), when others argue that prior criminal history proxies for race (see, e.g., Harcourt, 2015)? Adjusting for bias requires understanding the theory of discrimination and the mechanism of bias in the specific domain (Ho, 2018). While statistical debiasing solutions are critical in this debate, bias cannot be solved by an algorithm alone.

References

Adler, Laura (2016), "Learning from Location," March 24, https://datasmart.ash.harvard.edu/news/article/learning-from-location-806, accessed August 20, 2018.

Altenburger, Kristen M., Rajlakshmi De, Kaylyn Frazier, Nikolai Avteniev, and Jim Hamilton (2017), "Are there Gender Differences in Professional Self-Promotion? An Empirical Case Study of LinkedIn Profiles among Recent MBA Graduates," in: Derek Ruths (ed.), *Proceedings of the Eleventh AAAI Conference on Web and Social Media (ICWSM 2017), Montréal, Québec, Canada, May 15–18,* The AAAI Press, Palo Alto (CA), pp. 460–463.

Badger, Emily (2013), "How Yelp Might Clean Up the Restaurant Industry," *The Atlantic*, July/August issue, https://www.theatlantic.com/magazine/archive/2013/07/youll-never-throw-up-in-this-town-again/309383/, accessed August 20, 2018.

Barocas, Solon, and Andrew D. Selbst (2016), "Big Data's Disparate Impact," *California Law Review*, 104(3), 671–732.

Berk, Richard (2008), "Forecasting Methods in Crime and Justice," *Annual Review of Law and Social Science*, 4, 219–238.

— and Jordan Hyatt (2015), "Machine Learning Forecasts of Risk to Inform Sentencing Decisions," *Federal Sentencing Reporter*, 27(4), 222–228.

Board of Governors of the Federal Reserve System (2007), "Report to the Congress on Credit Scoring and its Effects on the Availability and Affordability of Credit," https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf, accessed August 20, 2018.

Breiman, Leo (2001), "Random Forests," *Machine Learning*, 45(1), 5–32.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2008), "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 90(3), 414–427.

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford (2017), "AI Now 2017 Report," AI Now Institute, New York University.

Cavallo, Sara, Joann Lynch, and Peter Scull (2014), "The Digital Divide in Citizen-Initiated Government Contacts: A GIS Approach," *Journal of Urban Technology*, 21(4), 77–93.

Centers for Disease Control and Prevention (2011), "CDC Estimates of Foodborne Illness in the United States," https://www.cdc.gov/foodborneburden/pdfs/factsheet_a_findings_updated4-13.pdf, accessed August 20, 2018.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq (2017), "Algorithmic Decision Making and the Cost of Fairness," in: *KDD'17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Min-*

ing, Halifax, Nova Scotia, Canada, August 13–17, The Association for Computing Machinery, New York, pp. 797–806.

Crawford, Kate (2013), "The Hidden Biases in Big Data," *Harvard Business Review*, April 1, https://hbr.org/2013/04/the-hidden-biases-in-big-data, accessed August 20, 2018.

— (2016), "Artificial Intelligence's White Guy Problem," *The New York Times*, June 25, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html, accessed August 20, 2018.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta (2015), "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination," *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2015(1), 92–112.

Dettling, Lisa J., Joanne W. Hsu, Lindsay Jacobs, Kevin B. Moore, and Jeffrey P. Thompson (2017), "Recent Trends in Wealth-Holding by Race and Ethnicity: Evidence from the Survey of Consumer Finances," *FEDS Notes*, September 27, https://www.federalreserve.gov/econres/notes/feds-notes/recent-trends-in-wealth-holding-by-race-and-ethnicity-evidence-from-the-survey-of-consumer-finances-2017 0927.htm, accessed August 20, 2018.

Devinney, Katelynn, Adile Bekbay, Thomas Effland, Luis Gravano, David Howell, et al. (2018), "Evaluating Twitter for Foodborne Illness Outbreak Detection in New York City," *Online Journal of Public Health Informatics (OJPHI)*, 10(1), http://ojphi.org/ojs/index.php/ojphi/article/download/8894/7343, accessed August 20, 2018.

Esarey, Justin, and Andrew Menger (2018), "Practical and Effective Approaches to Dealing with Clustered Data," *Political Science Research and Methods*, January 19, DOI: 10.1017/psrm.2017.42.

Executive Office of the President (2016), "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights," The White House, Washington (DC).

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian (2015), "Certifying and Removing Disparate Impact," in: *KDD'15: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, August 10–13*, Association for Computing Machinery, New York, pp. 259–268.

GAO (2004), "Data Mining: Federal Efforts Cover a Wide Range of Uses," Report GAO-04-548, U.S. General Accounting Office, Washington (DC).

Glaeser, Edward L., Andrew Hillis, Scott Duke Kominers, and Michael Luca (2016), "Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy," *The American Economic Review*, Papers and Proceedings, https://pubs.aeaweb.org/doi/pdfplus/10.1257/aer.p20161027.

Goodin, Amanda Kate, and Karl C. Klontz (2007), "Do Customer Complaints Predict Poor Restaurant Inspection Scores? The Experience in Alexandria, Virginia, 2004–2005," *Journal of Food Safety*, 27(1), 102–110.

Gould, L. Hannah, Kelly A. Walsh, Antonio R. Vieira, Karen Herman, Ian T. Williams, et al. (2013), "Surveillance for Foodborne Disease Outbreaks – United States, 1998–2008," *Morbidity and Mortality Weekly Report (MMWR): Surveillance Summaries*, 62(2), 1–34.

Grgić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller (2016), "The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making," in: *Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems, Barcelona, Spain, December 3–8*, http://www.mlandthelaw.org/papers/grgic.pdf, accessed August 20, 2018.

Grubmüller, Verena, Katharina Götsch, and Bernhard Krieger (2013), "Social Media Analytics for Future Oriented Policy Making," *European Journal of Futures Research*, 1(1), 1–20.

Harcourt, Bernard E. (2015), "Risk as a Proxy for Race: The Dangers of Risk Assessment," *Federal Sentencing Reporter (FSR)*, 27(4), 237–243.

Hardt, Moritz, Eric Price, and Nati Srebro (2016), "Equality of Opportunity in Supervised Learning," in: Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon,

and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: 30th Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain, December 5–10*, Curran Associates, Red Hook (NY), pp. 3315–3323.

Harris, Kimberly J., Kevin S. Murphy, Robin B. DiPietro, and Gretchen L. Rivera (2015), "Food Safety Inspections Results: A Comparison of Ethnic-Operated Restaurants to Non-Ethnic-Operated Restaurants," *International Journal of Hospitality Management*, 46, 190–199.

Harrison, Cassandra, Mohip Jorder, Henri Stern, Faina Stavinsky, Vasudha Reddy, et al. (2014), "Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness – New York City, 2012–2013," *Morbidity and Mortality Weekly Report (MMWI)*, 63(20), 441–445.

Ho, Daniel E. (2012), "Fudging the Nudge: Information Disclosure and Restaurant Grading," *The Yale Law Journal*, 122(3), 574–688.

— (2017a), "Does Peer Review Work?: An Experiment of Experimentalism," *Stanford Law Review*, 69(1), 1–119.

— (2017b), "Equity in the Bureaucracy," *UC Irvine Law Review*, 7(2), 401–451.

— (2018), "Judging Statistical Criticism," Comment, *Observational Studies*, 4, 42–56.

Holland, Paul W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81(396), 945–960.

Kamiran, Faisal, Indrė Žliobaitė, and Toon Calders (2013), "Quantifying Explainable Discrimination and Removing Illegal Discrimination in Automated Decision Making," *Knowledge and Information Systems*, 35(3), 613–644.

Kang, Jun Seok, Polina Kuznetsova, Michael Luca, and Yejin Choi (2013), "Where *Not* to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews," in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, October 18–21,* Association for Computational Linguistics, Stroudsburg (PA), pp. 1443–1448.

King, Gary (2011), "Ensuring the Data-Rich Future of the Social Sciences," *Science*, 331(6018), 719–721.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015), "Prediction Policy Problems," *The American Economic Review*, 105(5), 491–495.

Kontokosta, Constantine, Boyeong Hong, and Kristi Korsberg (2017), "Equity in 311 Reporting: Understanding Socio-Spatial Differentials in the Propensity to Complain," Electronic Preprint, Cornell University Library, arXiv:1710.02452.

Kwon, Junehee, Kevin R. Roberts, Carol W. Shanklin, Pei Liu, and Wen S. F. Yen (2010), "Food Safety Training Needs Assessment for Independent Ethnic Restaurants: Review of Health Inspection Data in Kansas," *Food Protection Trends*, 30(7), 412–421.

Lam, Francis (2012), "Cuisines Mastered as Acquired Tastes," *The New York Times*, https://www.nytimes.com/2012/05/30/dining/masters-of-a-cuisine-by-calling-not-roots.html, accessed August 20, 2018.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani (2014), "The Parable of Google Flu: Traps in Big Data Analysis," *Science*, 343(6176), 1203–1205.

—, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, et al. (2009), "Computational Social Science," *Science*, 323(5915), 721–723.

Legewie, Joscha, and Merlin Schaeffer (2016), "Contested Boundaries: Explaining where Ethnoracial Diversity Provokes Neighborhood Conflict," *American Journal of Sociology (AJS)*, 122(1), 125–161.

Lehman, David W., Balázs Kovács, and Glenn R. Carroll (2014), "Conflicting Social Codes and Organizations: Hygiene and Authenticity in Consumer Evaluations of Restaurants," *Management Science*, 60(10), 2602–2617.

Louppe, Gilles, Louis Wehenkel, Antonio Sutera, and Pierre Geurts (2013), "Understanding Variable Importances in Forests of Randomized Trees," in: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.), *Advances in Neural Information*

Processing Systems, 26th Annual Conference on Neural Information Processing Systems 2013, Lake Tahoe, Nevada, December 5–10, Vol. 1, Curran Associates, Red Hook (NY), pp. 431–439.

Morantz, Alison (2008), "Mining Mining Data: Bringing Empirical Analysis to Bear on the Regulation of Safety and Health in U.S. Mining," *West Virginia Law Review*, 111, 45–74.

Nsoesie, Elaine O., Sheryl A. Kluberg, and John S. Brownstein (2014), "Online Reports of Foodborne Illness Capture Foods Implicated in Official Foodborne Outbreak Reports," *Preventive Medicine (PM)*, 67, 264–269.

Pasquale, Frank (2015), *The Black Box Society: The Secret Algorithms that Control Money and Information*, Harvard University Press, Cambridge (MA).

Pope, Devin G., and Justin R. Sydnor (2011), "Implementing Anti-Discrimination Policies in Statistical Profiling Models," *American Economic Journal: Economic Policy*, 3(3), 206–231.

Ramirez, Edith, Julie Brill, Maureen K. Ohlhausen, and Terrell McSweeney (2016), "Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues," Report, Federal Trade Commission, Washington (DC), https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf, accessed August 20, 2018.

Roberts, Kevin, Junehee Kwon, Carol Shanklin, Pei Liu, and Wen-Shen Yen (2011), "Food Safety Practices Lacking in Independent Ethnic Restaurants," *Journal of Culinary Science & Technology*, 9(1), 1–16.

Sadilek, Adam, Sean Brennan, Henry Kautz, and Vincent Silenzio (2013), "nEmesis: Which Restaurants Should you Avoid Today?" in: *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing, Palm Springs, California, USA, November 7–9*, Association for the Advancement of Artificial Intelligence, Palo Alto (CA), pp. 138–146.

Schomberg, John P., Oliver L. Haimson, Gillian R. Hayes, and Hoda Anton-Culver (2016), "Supplementing Public Health Inspection via Social Media," *PLoS ONE*, 11(3), 1–21.

Sengupta, Somini (2013), "In Hot Pursuit of Numbers to Ward Off Crime," *The New York Times*, June 19, https://bits.blogs.nytimes.com/2013/06/19/in-hot-pursuit-of-numbers-to-ward-off-crime/, accessed August 20, 2018.

Simmons, Andrew (2014), "Gastronomic Bigotry," *Slate*, June 6, http://www.slate.com/articles/life/food/2014/06/ethnic_restaurants_and_food_poisoning_the_subtle_racism_of_saying_chinese.html?via=gdpr-consent, accessed August 20, 2018.

Simon, Phil (2014), "Potholes and Big Data: Crowdsourcing our Way to Better Government," *Wired*, March, https://www.wired.com/insights/2014/03/potholes-big-data-crowdsourcing-way-better-government/, accessed August 20, 2018.

Stoppelman, Jeremy (2016), "Yelp CEO Says Online Reviews Could Beat 'Gold Standard' Healthcare Measures," *Modern Healthcare*, March 31, http://www.modernhealthcare.com/article/20160531/NEWS/160539986, accessed August 20, 2018.

Sweeney, Latanya (2013), "Discrimination in Online Ad Delivery," *ACM Queue*, 11(3), 1–19.

The Department of the Treasury (2009), "Federal Agency Data Mining Report (2008)," Report to Congress, https://www.treasury.gov/privacy/annual-reports/Documents/FY 2008 Data Mining Report.pdf, accessed August 20, 2018.

Thompson, Tim (2015), "How our Cities are Using Social Data," *IBM Big Data & Analytics Hub*, May 12, https://www.ibmbigdatahub.com/blog/how-our-cities-are-using-social-data, accessed August 20, 2018.

Tomky, Naomi (2015), "How Not to Be a Restaurant Racist," *CityLab*, October 14, https://www.citylab.com/life/2015/10/how-not-to-be-a-restaurant-racist/410482/, accessed August 20, 2018.

Wilcock, Anne, Maria Pun, Joseph Khanona, and May Aung (2004), "Consumer Attitudes, Knowledge and Behaviour: A Review of Food Safety Issues," *Trends in Food Science & Technology*, 15(2), 56–66.

Williams, A. N., and K. M. Woessner (2009), "Monosodium Glutamate 'Allergy': Menace or Myth?" *Clinical & Experimental Allergy*, 39(5), 640–646.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi (2017), "Fairness Constraints: Mechanisms for Fair Classification," in: Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics 2017, Fort Lauderdale, Florida, USA, April 20–22,* pp. 962–970.

Zukin, Sharon, Scarlett Lindeman, and Laurie Hurson (2015), "The Omnivore's Neighborhood? Online Restaurant Reviews, Race, and Gentrification," *Journal of Consumer Culture*, 17(3), 459–479.

Kristen M. Altenburger
Management Science and Engineering
Stanford University
475 Via Ortega Avenue
Stanford, CA 94305-6015
USA
kaltenb@stanford.edu

Daniel E. Ho
Stanford Law School
Stanford University
559 Nathan Abbott Way
Stanford, CA 94305
USA
dho@law.stanford.edu