Compliance and the Returns to Algorithms

Edward L. Glaeser, Andrew Hillis, Hyunjin Kim, Scott Duke Kominers, and Michael Luca*

December 2020

PRELIMINARY AND NOT FOR DISTRIBUTION

Abstract

Algorithms have the potential to improve managerial decisions—but the returns are limited if decision-makers use discretion to overrule algorithmic recommendations based on their priors. Partnering with Boston's inspectional services department, we compare the performance of different methods to prioritize inspections: (1) human judgment based on inspector discretion (the status quo); (2) a "data-poor" algorithm based on the average number of historical violations; and (3) a "data-rich" algorithm based on a random forest model trained on historical inspections and Yelp data. Both algorithms provide substantial gains compared to inspector discretion, but there is little difference between the two algorithmic methods, suggesting that the greatest gains stem from using data to supplement inspectors' priors rather than from algorithmic sophistication. Despite these gains, inspectors are only half as likely to inspect restaurants based on algorithmic recommendations compared to those based on their own judgment. These findings suggest that if algorithms are to be effective, organizations must be designed to manage discretion and ensure that decision-makers trust algorithms more than their own information.

* We are grateful for the collaboration of the City of Boston (especially Ben Batorsky, Matthew Mayrl, and Commissioner William Christopher), Yelp (Artem Avdacev, Luther Lowe, and Aaron Schur), and DrivenData (Peter Bull and Greg Lipstein). Fabian Konig provided excellent research assistance. Additionally, the authors gratefully acknowledge the helpful comments of Susan Athey, Avi Goldfarb, Shane Greenstein, Sendhil Mullainathan, Andrei Shleifer, and Mitchell Weiss. Data for this project was provided by Yelp and the City of Boston. Kim and Luca have consulted for tech companies, including Yelp. We are grateful for the support of the National Science Foundation (grants CCF-1216095, DGE-1144152, and SES-1459912), the Harvard Milton Fund, Yelp, the Taubman Center for State and Local Government, the Rappaport Institute for Greater Boston, a Radcliffe Exploratory Seminar Grant, the Alfred P. Sloan Foundation, the Ewing Marion Kauffman Foundation, and the Ng Fund and the Mathematics in Economics Research Fund of the Harvard Center of Mathematical Sciences and Applications. All errors are our own.

1. Introduction

With recent reductions in data and computing costs, interest in using predictive algorithms to improve decision-making has grown substantially (Kleinberg et al (2017), Agrawal et al (2018)). In many contexts involving managerial decisions, algorithms provide predictions that decision-makers take as inputs, raising the question of whether and to what extent better prediction from algorithmic sophistication might improve decision-making (Hoffman et al (2017), Cowgill (2018), Choudhury et al (2020)). On the one hand, scholars since the 1970s have highlighted that even the simplest algorithms can reduce bias and increase consistency compared to human decisions (e.g., Dawes (1979), Grove and Meehl (1996), Kahneman et al (2016)). On the other hand, research on heuristics emphasizes the advantages of simple rules based on human intuition over complex analytical tools (e.g., Gigerenzer and Gaissmaier (2011), Sull and Eisenhardt (2015)).

Yet the practical implementation of algorithms and the returns to algorithmic sophistication in managerial settings has been less widely studied (Adner et al (2019)).[1] In particular, managing how decision-makers use algorithmic recommendations is a central concern, since even the most sophisticated algorithms will do little if decision-makers use discretion to follow their own priors.

In this paper, we evaluate the gains from algorithmic sophistication in the field, and explore the extent to which decision-makers comply with algorithmic recommendations. We find that algorithms provide substantial gains in comparison to human judgment. However, the greatest gains stem from simply integrating data into the inspector allocation process, rather than from algorithmic sophistication, per se. This suggests that simple heuristics can go a long way—but when driven by data—rather than human intuition. Despite the large gains to algorithmic recommendations, we find that decision-makers are only half as likely to follow those recommendations compared to their own judgment; they appear to override recommendations to follow their priors. These findings point to the many challenges that organizations may face in using algorithms to support human decision-makers with discretion. They suggest that the returns to algorithms may depend on the strength of priors and the effectiveness of management to incentivize compliance, consistent with prior work on complementarities between information technology investments and basic management practices (Bresnahan et al (2002), Bartel et al (2007), Bloom et al (2012)).

We partnered with the City of Boston's inspectional services department to implement and test algorithms aimed at identifying restaurants at risk of health code violations. Our setting is a natural place to test the power of predictive algorithms, since inspectors' scarce time must be allocated with a fundamentally predictive objective: identifying restaurants that have health code violations. Moreover, the inspectional services department's history of past violations and consumer reviews from platforms like Yelp together yield strong data inputs for a sophisticated prediction algorithm.

We compare three different approaches to allocate inspectors: (1) human judgment relying on inspector discretion ("business-as-usual"); (2) a "data-poor" algorithm based on the average number of historical violations of each restaurant; and (3) a "data-rich" algorithm based on a

[1] Recent empirical studies using regression discontinuity designs shed light on the use of algorithmic tools in judicial decisions, but find conflicting results across different contexts. For example, Stevenson (2017) finds only small, temporary impact of algorithms on pretrial release outcomes, while Berk (2017) finds large reductions in re-arrests. Furthermore, many of these cases involve settings where issues of bias and fairness are central, which confounds the estimation of the returns to using algorithms to improve prediction.

random forest model trained on historical violation data and Yelp reviews. We take the restaurants that have the largest predicted likelihood of violations according to each approach, and provide these lists to inspectors to guide their inspections over four periods of roughly two weeks each.

This design provides a few advantages. By examining both "data-poor" and "data-rich" algorithms, we can assess how differences in algorithmic quality impact inspection efficiency in the field. We also observe counterfactual inspector decisions and (non-)compliance behavior, which are often difficult to measure directly.

We find that our predictive algorithms outperform business-as-usual substantially: they identify restaurants with 50 percent more weighted violations. Most of the gains stem from integrating historical violations data, as the data-poor algorithm results in improvements nearly as large as those from the data-rich algorithm.

Even so, inspectors are only half as likely to inspect restaurants based on the algorithm, relative to those based on their own judgment. Thus, in the context we study, decision-maker compliance issues appear to be first-order relative to algorithm sophistication. As the inspectors may naturally possess private knowledge about likely restaurant violations, non-compliance in principle could have been beneficial. But our findings suggest that at least in the context of restaurant inspections, the gains from incorporating inspectors' private knowledge may not outweigh the gains from using data to guide the allocation process.

Our work shows both the potential and the challenge for algorithms to improve decisions involving prediction in practice. While we cannot claim that any context fully generalizes to other settings, this case study suggests that effective management of decision-makers' discretion and compliance—through incentives as well as clear communication and prioritization of objectives—can be more important than even significant refinements of predictive algorithms.

2. Empirical Context

We partner with the City of Boston to examine restaurant health inspection decisions, where inspectors use their judgment to make decisions on which restaurants to inspect prior to carrying out the inspections. This is a compelling setting in which to evaluate the impact of implementing algorithms for decision-making, for several reasons.

First, a key component of the inspection decision is a prediction problem. The main objective defined by Boston's Head Inspector is to identify and incapacitate establishments that pose the highest risk to public health.[2] As a result, a good decision depends on inspectors' ability to prioritize restaurants according to their likelihood of violation, so that restaurants with the highest risk to public health can be flagged and addressed as early as possible. While inspectors are encouraged to conduct inspections in close geographic proximity to each other when possible, that is considered secondary to inspecting restaurants with a higher likelihood of violation.

Second, improving the targeting of inspections has a direct impact on performance and efficiency, because it reduces the number of inspections needed and better allocates inspectors' scarce time. Inspectors are responsible for inspecting all establishments in their ward two or three times a year: higher-risk facilities like hospitals, nursing homes, day care centers, schools, and

[2] While there are additional possible objectives in our setting, such as deterring restaurants from committing violations in the first place or ensuring fairness in the inspection allocation, our discussions with health departments highlighted the first-order importance of identifying and inspecting restaurants with the highest likelihood of health violations.

caterers require three inspections per year, while restaurant establishments that prepare, cook, and serve most products immediately are aimed to be inspected at least twice a year. However, in practice, inspectors are time-constrained and often unable to make all targeted inspections.

Third, inspections are important to identify violations. Between 2007 and 2015, inspectors found anywhere from 0 to 60 weighted violations per restaurant inspection (Appendix Figure 1). Weights are assigned based on the severity of the violation: Level I (weight of 1) corresponds to non-critical violations such as building defects or standing water. Level II (weight of 2) are "Critical Violations" such as the presence of fruit flies, which are more likely violations to create food contamination, illness or environmental hazard. Level III (weight of 5) are those considered to be "Food-borne Illness Risk Factor[s]" like insufficient refrigeration or a lack of allergen advisories on menus. When critical violations are found in a restaurant, the City temporarily suspends the restaurant's food permit if they pose an imminent public health risk, or reinspects that restaurant within 30 days.

Finally, there are a lot of easily accessible data, both from historical administrative data, as well as from digital platforms like Yelp that provide insights into local restaurants. These datasets raise the possibility of improving prediction substantially, by way of sophisticated algorithms.

At the time of the study, the department employed approximately 20-30 inspectors at any given time, who were assigned to at least one of 22 wards that divided up Boston into neighborhoods. A few wards had multiple inspectors assigned to cover different areas, and inspectors' ward assignments were generally changed every two years.

3. Empirical Design

Between February 1 and March 25, 2016, we partnered with the City of Boston ("the City") to evaluate three methods to predict restaurant violations: (1) human judgment, (2) a "data-rich" algorithm, and (3) a "data-poor" algorithm. While we advised on the empirical design, the City made the final design choices and executed on the implementation.

We compared three methods to predict violations. The first method represented the status quo of relying on inspectors' own judgment to rank restaurants. To obtain these rankings, the Head Inspector asked all inspectors to rank the restaurants in their ward without a mandated priority to inspect, in the order that they intended to inspect them.[3] The second method (a "data-rich" algorithm) ranked restaurants using a random forest model trained on both historical violations and Yelp data—including the number of Yelp reviews, Yelp rating, price range, hours, services available (e.g., wifi, alcohol, take-out, appointments), business ambience (e.g., noise level, children-friendly, ages allowed), and business neighborhood.[4] The third method (a "data-poor

[3] This wording was chosen by the City of Boston as the most natural way to obtain inspector rankings. Restaurants with a mandated priority to inspect include high-risk establishments (e.g., at hospitals and nursing homes), as well as where prior inspections had yielded major violations requiring a re-inspection in 30 days, or restaurants that had been flagged by complaints. These establishments were excluded from our analysis in order to assess how inspectors themselves prioritized restaurants in the non-mandated set.

[4] This method was the second-place winner in a tournament run by the City of Boston to source algorithms for predicting violations (described in Glaeser et al. (2016)). This option provided theoretical efficiency gains of 40% relative to using inspector discretion, and was chosen above the first-place winner because the City felt it would be substantially easier to implement in-house.

algorithm) used the average number of violations across historical inspections to rank restaurants in each ward from most to least likely to have violations.

Because there were only 18 inspectors who conducted these inspections, each inspector was assigned to inspect equal numbers of restaurants ranked by every method, rather than being randomly assigned to one of the three methods. Each inspector received a docket of restaurants to inspect in each period, which listed the top-ranked restaurants from each of the methods in randomly sorted order.[5] The City determined the number of restaurants listed on each docket based on the number of restaurants that each inspector ranked for that period, which typically ranged from 15-21. Based on this number, the City's Data team sourced equal numbers of the highest rankings from the other two methods, removed any duplicates, and randomly sorted all restaurants to create a given docket.[6]

These dockets were presented as a "new way of doing inspections" to guide inspector decisions, and inspectors were explicitly informed that these dockets supplemented the list of restaurants that they had ranked with those that were prioritized using data that the City's data team had processed.[7] They were asked to go down the docket in each inspection period.

The rationale underlying this design was to address a few challenges in evaluating the effect of algorithms in practice. First, by asking inspectors to first rank their own choices, we could better observe what inspectors' counterfactual decisions would have been without the algorithmic recommendations. Second, we wanted to explicitly vary the degree of algorithmic sophistication, to understand how the quality of the algorithm being tested impacts the gains. Last, we sought to exogenously influence which restaurant was inspected when by randomizing the order of restaurants on the docket, in order to identify whether algorithmic sophistication identified restaurants with a higher number of violations.

3.1 Data and Empirical Approach

This design provided us with restaurant rankings by each method, dockets that randomly varied the sequence of inspections, and the resulting inspections data from the City to conduct our analysis. However, a few design and implementation issues led to empirical challenges.

First, inspectors in practice inspected substantially fewer restaurants than those assigned on the dockets. The total number of unique restaurants listed across all dockets over this period was 1,042. However, inspectors were only able to inspect 361 restaurants, averaging to approximately 20 restaurants per inspector during the study.

Second, the City modified the docket generation process for the last two periods, after observing that inspectors could not complete the dockets. For the latter two periods, dockets were filled by listing restaurants that had not yet been inspected from previous dockets. While dockets were still capped at a maximum of 47 restaurants, this change meant that each docket no longer sourced an equal number of restaurants from each method if inspectors had completed an imbalanced number of restaurants across methods in prior weeks.

---

[5] Each inspection period covered approximately 2 weeks, and rankings were processed prior to the inspection periods.

[6] The City made this decision in order to accommodate inspectors and include all the restaurants that they themselves had prioritized.

[7] The Health Commissioner chose this wording in order to avoid calling attention to the new docket practice.

Lastly, rankings from all three methods were not available for all restaurants. Inspectors ranked only their highest-ranked restaurants in each period, so restaurants that were listed on the dockets because they were ranked highly by algorithmic methods did not have an inspector ranking. There were also some restaurants ranked highly by inspectors that lacked rankings from algorithmic methods if there were no historical inspections or Yelp data.

To address these issues, we take the following steps. First, we focus our main analyses on evaluating whether inspected restaurants ranked in the top 20 by algorithms have a higher number of violations than those ranked in the top 20 by inspectors. Restricting to this subsample ensures a more consistent availability of rankings, and allows us to compare inspection outcomes across comparable rankings in each method. Furthermore, since inspectors ranked their highest-priority restaurants, comparing the top 20 restaurants provides insight into how the top-ranked restaurants under each of the three methods differ, and whether restaurants that were ranked highly by the inspectors versus the algorithmic methods have a higher number of violations. In this analysis, any missing rankings from algorithmic methods should lead to a conservative estimate of any gains from algorithmic methods.

This subsample consists of 280 restaurants out of the full set of 361 that were inspected, and represents a subset of all 674 restaurants that were ranked in the top 20 by any method. We find overlaps between the methods, especially those using algorithms, with 176 restaurants (26%) ranked by at least two if not all three methods to be in the top 20. 108 (16%) are ranked in the top 20 by data-rich algorithm alone, 97 (14%) by data-poor algorithm alone, and 293 (43%) by inspectors alone.

Based on this data, we assess the gains from using algorithms by examining the number of violations found across restaurants ranked in the top 20 by algorithmic methods compared to those ranked by inspectors. We use the following model as our main specification for restaurant i:

$$Total\ Violations_i = \alpha + \beta DataRich_i + \gamma DataPoor_i + \delta MultipleMethods_i + \varepsilon_i \qquad (1)$$

Here, α represents the mean number of weighted violations for restaurants ranked in the top 20 by inspectors; β and γ represent the mean expected difference in weighted violations for a restaurant ranked by the data-poor and data-rich algorithms relative to a restaurant ranked by inspectors, respectively; meanwhile, δ accounts for overlaps between methods and represents the mean expected difference in weighted violations for a restaurant ranked by multiple methods.

We explore the robustness of the results across alternative subsamples. We vary the threshold of the top 20 and show that results are robust to taking different thresholds ranging from 10-30, as well as the full sample of inspected restaurants. We also account for changes in the docket-generation process by evaluating whether our results are robust to restricting to the first two periods before the modification occurred.

Lastly, we evaluate selection issues in terms of which restaurants were inspected and how they might impact our estimates of the gains from algorithms.

4. Results

We find large gains from using algorithms: algorithmic methods identify restaurants with over 50% more violations on average compared to those prioritized by inspectors. The largest gains stem from using *any data at all*, rather than algorithmic sophistication.

Yet despite these gains, we find that inspectors were half as likely to follow algorithmic recommendations—instead, they chose to prioritize the restaurants they themselves ranked. This non-compliance poses a selection issue that could be driving the gains from algorithms that we identify, but we find little evidence that selection is likely to explain the full magnitude of the effects we observe.

Meanwhile, we find suggestive evidence that non-compliance may be driven by inspectors' strong priors on predictors of restaurant violations, suggesting that human discretion may prevent organizations from realizing gains from algorithms in the presence of strong mental models.

4.1 The gains from algorithms and algorithmic sophistication

We begin by comparing the performance of the three methods. Table 1 presents our estimates of $\alpha, \beta,$ and γ under two specifications. Column 1 shows a comparison of mean weighted violations by restaurants ranked by one of the methods alone or by multiple methods. The mean number of weighted violations for restaurants ranked by human judgment alone is 6.8. This is equivalent to having a Level II and a Level III violation. Our estimates of the returns to algorithms over human judgment, β and γ, are 5.03 and 4.88 respectively, which represents a difference of targeting a restaurant with on average one more Level III violation. The coefficients on the two algorithmic methods are not statistically distinguishable, although the data-rich algorithm used both far richer data and a more sophisticated algorithm.

In Column 2 of Table 1, we explore a slightly different specification that accounts for restaurants ranked by inspectors that were also ranked by one of the algorithms. The constant term here shows the mean number of weighted violations for restaurants ranked by inspectors—accounting for both restaurants that were ranked by inspectors alone and those that overlapped with one of the algorithms. Accounting for these overlaps increases the average number of violations found at inspector-ranked restaurants to 7.4 compared to 6.8. We also separate out restaurants that were ranked by both algorithms and all three methods, and find that restaurants ranked by both algorithms or all three methods increase this number by twofold.

The results just described are robust across alternative subsamples that vary the threshold of top-ranked restaurants (Appendix Table 1), as well as subsamples that restrict to the first one or two inspection periods that occurred prior to the modification in the docket generation process (Appendix Table 2).

One key consideration in interpreting these results is what the inspector-ranked method represents. Inspectors were asked to rank the restaurants in the order they intended to inspect them, raising the possibility that this status quo may not be measuring inspector judgment if inspectors were not prioritizing restaurants with a high number of violations. In our interpretations, we view this method as likely to represent inspector judgment, as the wording was chosen by the City as the most natural way to obtain inspector rankings—inspectors were trained to prioritize restaurants with the highest number of violations, and incentivized to do so as any high-risk restaurants that were later flagged through complaints would affect their work.

Based on these results, we draw two conclusions. First, the data-poor and data-rich algorithms outperform human judgment in predicting violations, and these performance improvements are on the order of over 50% and statistically significant.

Second, the performance of the data-poor and data-rich algorithms are statistically indistinguishable, suggesting that the marginal benefit of additional data may be limited in this case. This is consistent with findings in similar applications to problems with representative

datasets, especially when the scale of the dataset is smaller (Ng 2018). This result suggests that in some cases, algorithmic sophistication may not lead to substantially larger gains in decision-making, and reinforces that simple heuristics can go a long way—but when driven by data, rather than human decision-makers with discretion.

While these results suggest that prior violations play an important role in predicting current violations, one can imagine important reasons why a city might not want to use them to guide inspection decisions. For example, if heterogeneity is driven by variation in inspector stringency as opposed to true variation in violations, as found in Jin and Lee (2018), we may be concerned about relying heavily on past data. Furthermore, as with any simple algorithm, using historical violations to guide decisions may facilitate gaming, eventually reducing the efficacy of this approach.

4.2 Compliance and Selection

Despite the gains from algorithms, we find that inspectors were only half as likely to inspect restaurants recommended by algorithms, compared to those based on their own judgment. This non-compliance issue highlights an important implementation concern: in practice, effective management of compliance may be important for organizations hoping to realize gains from algorithms. At the same time, compliance presents a potential problem of selection bias: we are able to observe inspection results for only a small subset of the restaurants ranked in the top 20, and the results described in the prior section depend upon having a representative sample from each method's rankings.

Table 2 shows the extent of the non-compliance we observe. Inspector-only ranked restaurants accounted for 61% of all inspected restaurants, whereas either of the algorithm-only ranked restaurants accounted for only 10% each of all inspected restaurants. Comparing these percentages to the percentages across all top-20 ranked restaurants detailed in Section 3.1, we find that inspectors were only half as likely to inspect restaurants based on algorithms relative to their own judgment. They inspected 171 out of the full set of 293 restaurants that they alone ranked in the top 20 (58%), but only inspected 29 out of 108 (27%) and 28 out of 97 (29%) restaurants that the data-rich or data-poor algorithm alone ranked.

Figure 1 examines heterogeneity across inspectors, plotting the percentage of restaurants inspected by each method, with the red line plotting what the percentage breakdown would have been if the inspector had followed the dockets. While we observe some heterogeneity across inspectors, almost all inspectors appear to have inspected more restaurants prioritized by their own judgment compared to those that were algorithmically ranked.

While this non-compliance result is interesting to observe, it also poses a potential threat to our results, because we observe inspection results for only a subset of the restaurants on each docket. In particular, it raises the concern that inspectors may have been more likely to inspect restaurants on the algorithm-generated lists with a higher likelihood of violation. The performance differences we observe across methods could then be driven primarily by a selection effect of not observing outcomes for restaurants ranked lower by the data-rich and data-poor algorithms, rather than a treatment effect.

We test this concern in Column 1 of Table 3 by looking for differences in average ranking by method for inspected restaurants, excluding any that were ranked by multiple methods. If inspectors inspected higher-ranked restaurants on algorithmic lists, then the average ranking of restaurants on algorithmic lists would be higher than those on the inspector-generated list.

The point estimates suggest that there is a slight bias in the opposite direction, with restaurants ranked by inspectors alone occupying higher ranking positions compared to those ranked by the data-rich and data-poor algorithms, although differences are small and statistically insignificant. This suggests that the results are unlikely to be driven by observing different parts of the ranking distribution for each method and misattributing these differences.

Furthermore, we find little evidence that the gains from algorithms are emerging from a particular part of the ranking distribution. In Column 2 of Table 3, we explore whether the gains from algorithms vary across rank. The gains from algorithms appear to be spread across the ranking distribution, as the coefficients on interactions with rank are both small and relatively precise around 0.

These results, in context of our broader findings, suggest that the performance differences we observe between the algorithmic approaches and inspector discretion are unlikely to be fully explained by selection alone. First, while there may be selection in the restaurants that inspectors choose to inspect, inspectors do not appear to choose substantially dirtier restaurants from the algorithmic approaches compared to their own list. This suggests that inspectors may not be making sophisticated tradeoffs, and makes it difficult to construct a clear alternative story driven by selection. Second, the magnitude of the differences we observe between algorithmic approaches and inspector discretion is quite large, and does not differ significantly across rankings. Given this, it seems unlikely that selection would change these results directionally.

However, one key limitation to our analysis is that we do not know how clean the restaurants that inspectors do not visit may be. Although inspectors are not systematically prioritizing restaurants predicted to have the most violations, it remains possible that the uninspected restaurants are much cleaner than the inspected ones, which would affect the magnitude of gains from algorithms in a higher compliance world.

4.3 What drives non-compliance?

These findings also raise the question of what drives inspectors' non-compliance. While our analysis cannot fully pin down the mechanism, we consider a few possible explanations: (1) algorithmic aversion, (2) balancing another objective like minimizing geographic distance, (3) inspectors' strong priors on what predicts violations.

One explanation is algorithmic aversion, which has been shown to play a role in some settings (e.g., Dietvorst et al. 2015). This is possible, as inspectors were informed that the dockets were supplemented with restaurants prioritized by data. But, inspectors were excited by the prospect of using data, and were not explicitly informed of algorithms, suggesting that this explanation may be less likely.

Another explanation may be that inspectors were balancing and prioritizing another objective than the number of violations, such as geographic distance, and thus sacrificing targeting restaurants with higher violations to reduce the distance that they traveled. To explore this, we compare the geographic distance inspectors traveled to their next restaurant compared to the distance from the closest algorithm-ranked restaurant that they did not inspect. However, we find the latter to be a subset of the first—suggesting that inspectors often had an algorithmically-ranked restaurant in closer proximity than the next restaurant they did travel to (see Figure 2).

Finally, anecdotal evidence from inspectors seems to broadly suggest the presence of strong priors. Inspectors appear to have overridden algorithmic recommendations when they conflicted with their priors, rather than using them to update and improve the accuracy of their own mental

models. In particular, inspectors had their own simple heuristics that drove predictions. We find some suggestive evidence consistent with this anecdotal evidence in the summary statistics of restaurants ranked by each of the methods, which show that, compared to algorithms, inspectors were more likely to prioritize chains, seafood restaurants, older, and lower-end businesses (Table 4).

While this provides little conclusive evidence on mechanisms, it raises the possibility that human discretion may prevent organizations from realizing gains from algorithms in decision-making. In this case, simple rules developed by human intuition, which may have provided advantages for decision-making in the past, may have ended up as a crutch when using algorithms for decision-making. This is consistent with evidence found by Hoffman et al (2018) in hiring decisions, where managers who appear to hire against job test scores ended up with worse average hires. The value of discretion likely depends on a number of factors, including how much private information decision-makers have, how aligned their incentives are with the objective at hand, and how biased (or unbiased) they may be. Furthermore, the value of discretion may be highly dynamic, if decision-makers become more likely to rely on algorithms as they observe their performance and are able to exercise discretion more carefully.

5. Discussion

Our study shows that in the case of restaurant inspections, using predictive algorithms can significantly improve decisions. Even a simple algorithm based on internal historical data better prioritized restaurants relative to inspector discretion. Moreover, a simple algorithm provided gains nearly as large as those from a more data-rich algorithm. But, we also find compliance to be a first-order issue: inspectors frequently chose to prioritize restaurants based on their own judgment, rather than those based on algorithms.

Chicago and New York City have run pilots using similar inspection targeting technologies (Ravindranath (2014)), and our findings speak to the promise and challenge involved in implementing such approaches at scale. Our results show a clear role for algorithms, but also highlight that designing the decision-making process to enable compliance and learning may be a first-order issue—rather than improving algorithm sophistication and/or expanding the range of input data. In particular, decision-makers with discretion may not comply with algorithms that ignore their priors, which means that algorithms must either cater to those preferences, or the organization needs to better guide them to follow algorithmic recommendations.

References

Adner, R., Puranam, P., & Zhu, F. (2019). What Is Different About Digital Strategy? From Quantitative to Qualitative Change. *Strategy Science*, *4*(4), 253-261.

Agrawal, A., Gans, J., and Goldfarb, A (2018). Prediction Machines: The Simple Economics of Artificial Intelligence.

Bartel, A., C. Ichniowski, K. Shaw. 2007. How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills. Quarterly Journal of Economics 122(4): 1721-17

Berk, Richard (2017). "An impact assessment of machine learning risk forecasts on parole board decisions and recidivism," Journal of Experimental Criminology, 13 (2), 193–216.

Bloom, N., R. Sadun, J. Van Reenen. 2012. Americans do IT better: US multinationals and the productivity miracle. Amer. Econom. Rev. 102(1) 167-201.

Bresnahan, T., E. Brynjolfsson, L. M. Hitt. 2002. Information technology, workplace organization and the demand for skilled labor: Firm-level evidence. Quart. J. Econom. 117(1) 339-376

Choudhury, Prithwiraj, Evan Starr, and Rajshree Agarwal (2020). *Machine learning and human capital: Experimental evidence on productivity complementarities*. Strategic Management Journal, forthcoming.

Cowgill, B. (2018). Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening. Working Paper.

Dawes, R. M. (1979). The Robust Beauty of Improper Linear Models in Decision Making. American Psychologist, 34 (7), 571–582.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. Journal of Experimental Psychology: General, 144 (1), 114–126.

Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li. "Discretion in hiring." *The Quarterly Journal of Economics* 133, no. 2 (2018): 765-800.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, *62*, 451-482.

Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. (2016). Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. American Economic Review, 106 (5), 114–118.

Grove, W., & Meehl, P. (1996). Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy. Psychol Public Policy Law, 2 (2), 293–323.

Jin, Ginger Zhe, and Jungmin Lee (2018). A Tale of Repetition: Lessons from Florida Restaurant Inspections. Working Paper.

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). NOISE: How to overcome the high, hidden cost of inconsistent decision making. *Harvard business review*, *94*(10), 38-46

Kleinberg, J., Lakkaraj, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human Decisions and Machine Predictions. NBER Working Paper No. 23180.

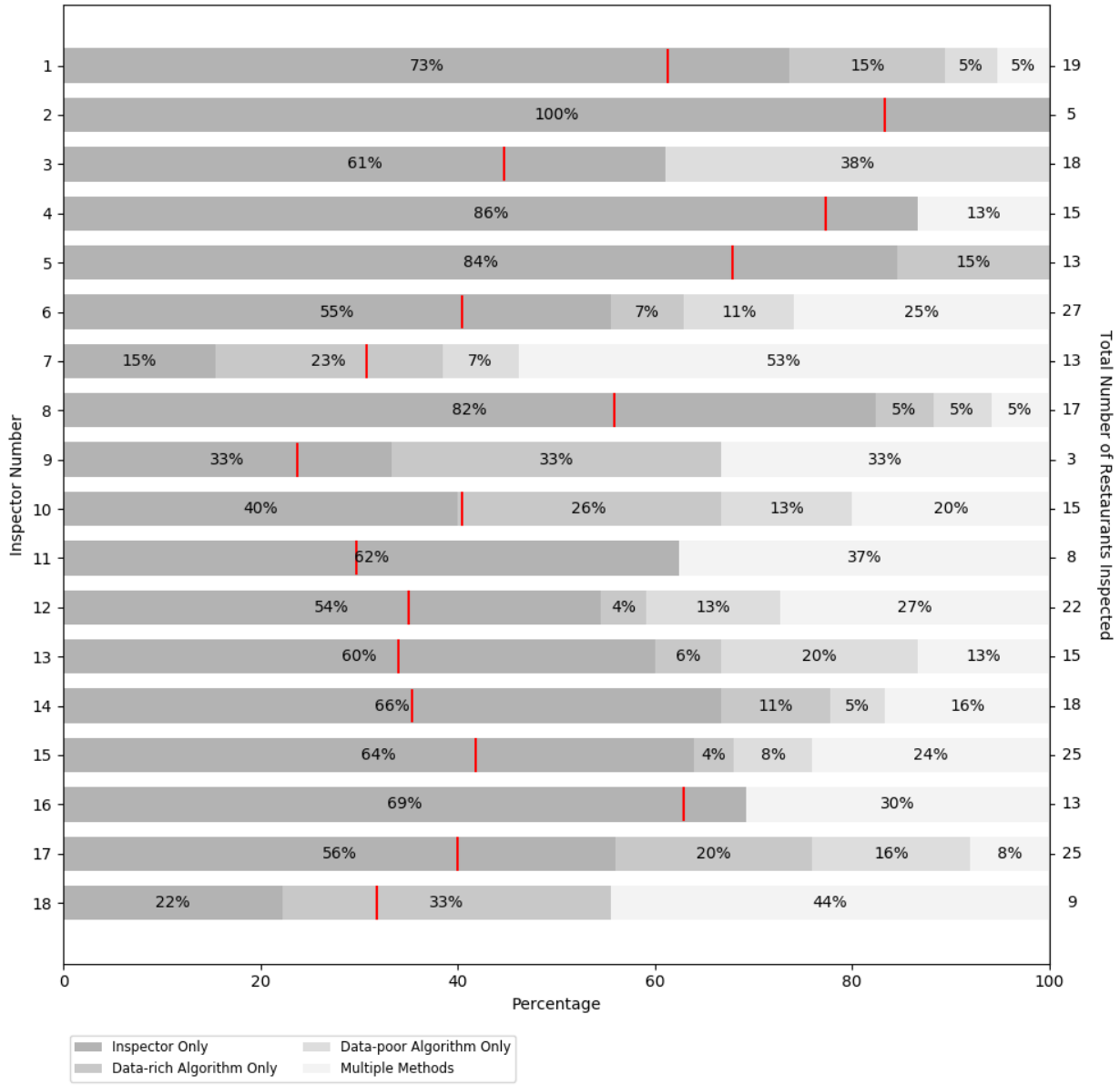
Ng, Andrew (2018). Machine Learning Yearning: Technical Strategy for AI Engineers, In the Era of Deep Learning. Draft Version.

Ravindranath, Mohana (2014). "In Chicago, food inspectors are guided by big data." The Washington Post, September 28, 2014. https://www.washingtonpost.com/business/on-it/in-chicago-food-inspectors-are-guided-by-big-data/2014/09/27/96be8c68-44e0-11e4-b47c-f5889e061e5f_story.html

Stevenson, Megan (2017). "Assessing Risk Assessment in Action," George Mason Law & Economics Research Paper, (17-36), 4.

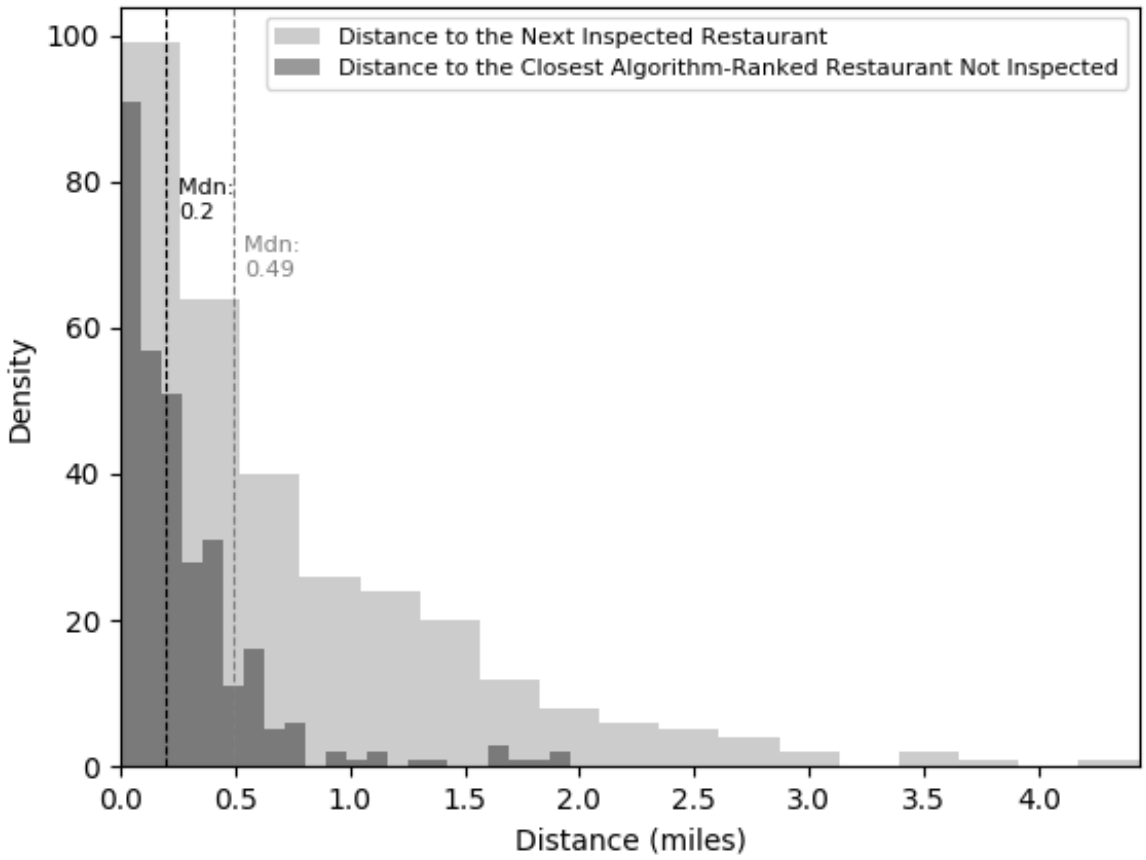
Sull, D. and Eisenhardt, K. (2015). *Simple rules: How to thrive in a complex world*. Houghton Mifflin Harcourt.

Figure 1: Percentage Inspected by Method across Inspectors



This figure plots the percentage of inspected restaurants by ranked method for each inspector. Each bar represents a single inspector, where the left axis indicates the inspector, and the right axis shows the number of restaurants that the inspector inspected. The red line indicates the percentage of inspector-only ranked restaurants in the full sample of top 20-ranked restaurants, which is where the Inspector-Only bar (in dark grey) should have ended if inspectors had fully complied.

Figure 2: Comparison of the distance inspectors travelled versus the closest algorithm-ranked restaurant not inspected



This figure plots the distribution of the distance inspectors travelled to their next restaurant, compared with the distance to the closest algorithm-ranked restaurant on the docket that was not inspected.

Table 1: Performance of Rankings

	(1)	(2)
Outcome:	Total Violations	Total Violations
	b/se	b/se
Data-rich Algorithm Only	5.03***	4.43***
	(1.13)	(1.16)
Data-poor Algorithm Only	4.88***	4.28***
	(1.36)	(1.33)
Multiple Methods	7.66***	
	(1.27)	
Both Algorithms		7.46***
		(1.45)
All Methods		8.37**
		(3.16)
Constant	6.80***	7.40***
	(0.61)	(0.80)
Observations	280	280
Including Ranking Up To:	20	20

Total violations is a weighted sum of one, two, and three star violations. *Data-rich Algorithm Only* and *Data-poor Algorithm Only* are binary variables indicating restaurants that were ranked in the top 20 by the data-rich algorithm or the data-poor algorithm only, respectively. *Multiple Methods* indicates restaurants that were ranked in the top 20 by at least two or all three methods. *Both Algorithms* indicates restaurants ranked in the top 20 by both data-rich and data-poor algorithms, but not the inspectors. *All Methods* indicates restaurants ranked in the top 20 by all three methods. Standard errors are clustered at the inspector level.

Table 2: Inspector Compliance

	(1)	(2)	(3)
	Number of Restaurants Inspected	(%)	% of Restaurants Inspected Out of All Top-20 Ranked Restaurants
Data-rich Algorithm Only	29	10.36	26.85
Data-poor Algorithm Only	28	10	28.87
Inspector Only	171	61.07	58.36
Multiple Lists	52	18.57	29.55
Total	*280*	*100%*	*100%*

This table shows the breakdown of inspected restaurants by ranking method. Column (1) and (2) respectively show the number of restaurants that were inspected in each category and the corresponding percentages. Column (3) shows the percentage of restaurants inspected out of all top-20 ranked restaurants in that category.

Table 3: Differences in Rankings and Performance across the Ranking Distribution

	(1)	(2)
Outcome:	Rank	Total Violations
	b/se	b/se
Data-rich Algorithm Only	1.29	4.55*
	(1.07)	(2.55)
Data-poor Algorithm Only	1.06	3.28
	(1.09)	(2.84)
Data-Rich Algorithm x Rank		0.04
		(0.19)
Data-Poor Algorithm x Rank		0.14
		(0.22)
Rank		-0.03
		(0.06)
Constant	10.44***	7.15***
	(0.41)	(0.89)
Observations	228	228

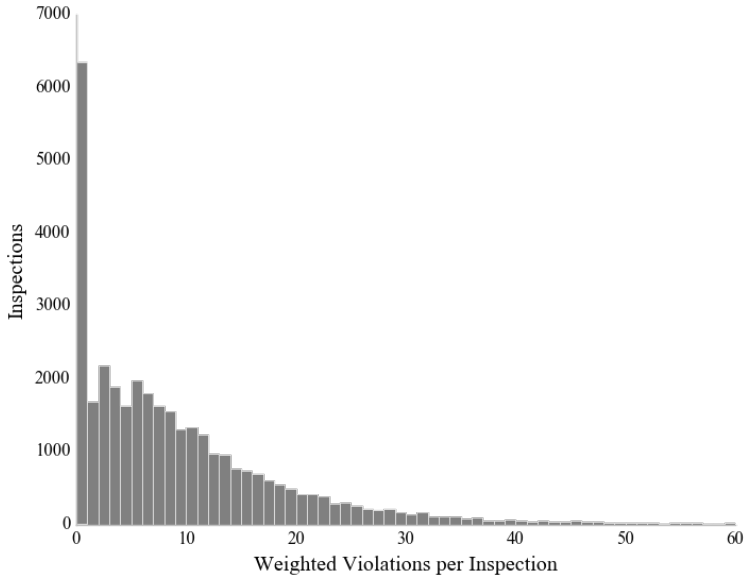
These regressions are run across the subsample of restaurants ranked in the top 20 by one of the methods alone, excluding any restaurants ranked by multiple methods. Column (1) analyzes differences in rankings across inspected restaurants, where *Rank* indicates the ranking position using the method that ranked the restaurant in the top 20. Column (2) analyzes whether the performance of algorithmic methods differs depending on the ranking position, where *Total violations* is a weighted sum of one, two, and three star violations. Standard errors are clustered at the inspector level.

Table 4: Characteristics of Top 20-Ranked Restaurants Across Methods

	(1)	(2)	(3)		
	Data-Rich Algorithm Only	Data-Poor Algorithm Only	Inspector Only	p-value 1=3	p-value 2=3
Chain	0***	0.04*	0.1	<0.001	0.07
Yelp Rating	3.14	2.6	2.97	0.33	0.17
Yelp Reviews	119.9	144.41	154.28	0.19	0.74
Ethnic Cuisine	0.46***	0.49***	0.25	0.01	<0.001
Seafood	0***	0.05	0.06	0.004	0.66
Restaurant Age	1.69***	3.18**	7.27	0.003	0.05
Price Range	1.4	1.14	1.27	0.17	0.4
Accepts Reservations	0.27	0.22	0.21	0.2	0.84
Table Service	0.46**	0.38	0.32	0.03	0.34

This table compares the attributes of restaurants ranked in the top 20 by each method, excluding any restaurants included across multiple methods. Columns (1)-(3) show means of each variable, and the last two columns display the p-value of the difference between restaurants ranked in the top 20 by the *Data-Rich Algorithm Only* and *Inspector Only*, and the *Data-Poor Algorithm Only* and *Inspector Only*, respectively, from a regression of the restaurant attribute on an indicator for being ranked by one of the algorithmic methods.

Appendix Figure 1: Distribution of Violations



This figure shows the distribution of weighted violations across inspections from January 2007 through June 2015.

Appendix Table 1: Robustness across sample restrictions

	(1)	(2)	(3)	(4)	(5)
Outcome:	Total Violations	Total Violations	Total Violations	Total Violations	Total Violations
	b/se	b/se	b/se	b/se	b/se
Data-rich Algorithm Only	4.15**	4.87***	4.59***	4.63***	4.28***
	(1.79)	(1.19)	(1.22)	(1.20)	(1.30)
Data-poor Algorithm Only	4.03*	4.30**	4.44***	4.48***	4.13***
	(2.22)	(1.63)	(1.35)	(1.37)	(1.27)
Multiple Methods	7.82***	7.68***	7.22***	7.26***	6.91***
	(1.75)	(1.33)	(1.26)	(1.29)	(1.30)
Constant	7.10***	7.02***	7.24***	7.20***	7.55***
	(0.80)	(0.64)	(0.64)	(0.57)	(0.55)
Observations	155	220	312	337	361
Including Ranking Up To:	10	15	25	30	All

Only restaurants ranked within the top 10 by any condition are included. *Total violations* is a weighted sum of one, two, and three star violations.

Appendix Table 2: Robustness across time periods

	(1)	(2)
Outcome:	Total Violations	Total Violations
	b/se	b/se
Data-rich Algorithm Only	3.76**	4.39***
	(1.37)	(1.33)
Data-poor Algorithm Only	4.44**	4.23**
	(1.86)	(1.73)
Multiple Methods	8.73***	7.98***
	(1.79)	(1.49)
Constant	6.85***	6.97***
	(0.53)	(0.51)
Observations	200	220
Including Periods Up To:	1	2

Only restaurants ranked within the top 20 by any condition are included. *Total violations* are a weighted sum of one, two, and three star violations.