Decision Authority and the Returns to Algorithms

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

April 2021

PRELIMINARY AND NOT FOR DISTRIBUTION

Abstract

Algorithms have the potential to improve managerial decisions—but the returns depend on how decision-makers use them. We explore a pilot run by an Inspectional Services Department that tested a pair of predictive algorithms that varied in their sophistication and inputs. We find that both algorithms provide substantial gains on the department's measure of interest compared to human judgment. But, there is little difference in the average performance between the two algorithmic methods, suggesting that in this context the greatest gains stem from using any data, rather than algorithmic sophistication. Despite measurable gains on the dimension of interest, decision-makers are only half as likely to follow algorithmic recommendations compared to their own judgment. Qualitative and exploratory empirical evidence suggests that this gap appears to be driven at least in part by differing predictions—as inspectors were using intuition about which features were most predictive. Overall, our findings suggest that for algorithms to translate into improvements in managerial decisions, organizations must carefully manage how decision authority is allocated and used, and that while simple rules based on intuition can be helpful, they may also act as an impediment to effective use of algorithms for decision-making.

* Authors are listed alphabetically. Fabian Konig provided excellent research assistance. The authors gratefully acknowledge the helpful comments of Susan Athey, Felipe Csaszar, Avi Goldfarb, Shane Greenstein, Kristina McElheran, Sendhil Mullainathan, Andrei Shleifer, and Mitchell Weiss. Data for this project was provided by the Inspectional Services Department and Yelp. 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, Harvard Business School, the Rappaport Institute for Greater Boston, a Radcliffe Exploratory Seminar Grant, the Alfred P. Sloan Foundation, the Ewing Marion Kauffman Foundation, 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

Organizations are increasingly interested in using algorithms to support decision-making, drawing on insights dating back to at least the 1970s that demonstrate the potential for even simple algorithms to help improve decisions (e.g., Dawes (1979), Grove and Meehl (1996), Kahneman et al (2016)). More recently, the use of data and prediction has been linked directly to performance gains across firms (Brynjolfsson and McElheran (2019), Bajari et al (2019), Camuffo et al (2020)). At the same time, evidence suggests that firms may fail to realize these potential gains (Kim (2020), Brynjolffson, Rock, and Syverson (2021)).

When leveraging data and algorithms, organizations must decide not only whether to use algorithms, but also how to use them. For example, a decision might be completely automated. However, in many managerial contexts, algorithms are instead used to form predictions that decision-makers may take as inputs (Agrawal, Gans, and Goldfarb (2018, 2019), Cowgill (2019), Choudhury, Starr, and Agarwal (2020)). In these cases, managers are given decision authority to use algorithmic recommendations as a decision aid, as opposed to a decision rule. Evaluating how decision-makers use their discretion when faced with data-driven inputs is an important step toward understanding the impact of algorithms in practice (Athey, Bryan, Gans (2020)).

In this paper, we explore the role that decision authority may play in translating the informational gains from algorithms into improvements in decision outcomes, looking at the use of algorithmic predictions in the field. We compare the performance of human judgment with and without algorithmic support, and further compare two algorithms with varying degrees of sophistication. One algorithm is based on simple historical averages, while the other uses a random forest model trained on both historical data from within the organization and additional data from online platforms. We find that algorithms provide substantial gains compared to human judgment. However, the greatest gains in this context stem from simply integrating data into the decision process, rather than from algorithmic sophistication. This suggests that simple heuristics can be valuable in managerial settings—but when driven by data—rather than human intuition alone. Furthermore, to the extent that the algorithm truly captured the department's goals, we find that decision-makers may in some cases use their decision authority to dissipate gains from algorithms by over-rejecting data-driven recommendations in favor of their own judgment. These findings suggest that in some settings, managing decision authority may dominate investments in

algorithmic sophistication or enhanced data collection, at least in these early days of putting algorithms to into practice.

We evaluate the impact of algorithms on decision-making through an intervention implemented by an Inspectional Services department, where inspectors use their judgment to make decisions on which restaurants to inspect. This setting provides compelling attributes for testing the power of predictive algorithms, since inspectors' scarce time must be allocated with an uncertain but at least partially predictive objective: identifying restaurants with health code violations. While inspectors possess experience and insight that inform their judgment, historical administrative records and external data may help to improve predictions (Lehman (2014), Glaeser et al (2016)).

We compare three approaches the department implemented to allocate inspectors: (1) human judgment ("business-as-usual"), (2) a "data-poor" algorithm based on the average number of historical violations for each restaurant; and (3) a "data-rich" algorithm based on a random forest model trained on both historical violations and Yelp data.[1] Restaurants with the highest predicted likelihood of violations according to each approach are randomly sorted and provided as lists to inspectors to guide their inspections over four periods of two weeks each. This design allows us to observe counterfactual inspector judgment and their ultimate decisions, and provide insights into the intensive-margin gains from algorithmic sophistication in the field by comparing "data-poor" and "data-rich" algorithms.

We find substantial gains from predictive algorithms compared to human judgment: algorithmic methods identify restaurants with over 50 percent more violations compared to inspectors. 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 algorithm-recommended restaurants relative to those based on their own judgment—suggesting that managerial discretion may dissipate potential gains from using algorithms. We explore the possibility that selection due to non-compliance in our data could be driving estimated gains from algorithmic inputs, but find little evidence that this can explain the full magnitude of the observed effects.

[1] We use the term "data-rich" in a relative sense to the other algorithm, rather than an absolute sense. One can imagine using a vast set of other data that may yield higher-quality insights, which is beyond the scope of this paper. The motivation behind this treatment was to explore the extent to which richer data modeled in a more sophisticated way adds any marginal gain, given the rising interest and investment in data and advanced technologies.

While our analysis cannot fully pin down the mechanism, we consider a few possible explanations for inspectors' noncompliance with algorithmic recommendations: (1) algorithm aversion, (2) the balancing of another objective such as minimizing geographic distance (i.e., costs of inspection) or the time since last inspection, (3) inspectors' own priors on what drives violations based on their intuition and experience. While not conclusive, anecdotal and exploratory empirical evidence suggests that inspectors rejected algorithmic recommendations when they conflicted with their priors on restaurant attributes that drive violations. These findings suggest that simple rules of thumb developed in the presence of uncertainty may work against introducing algorithms to support decision-making.

In addition to the literature on algorithms and decision-making, our analysis contributes to research on strategy and digitization more broadly. A growing body of work has identified many organizational practices that shape returns to investments in information technology and data (Bresnahan, Brynjolfsson, and Hitt (2002), Bartel, Ichinowski, and Shaw (2007), Bloom et al (2012), Brynjolfsson, Jin, and McElheran (2021)). Our findings point to additional, under-studied challenges that organizations may face in deploying algorithms in practice. While no single context fully generalizes to other settings, our findings suggest that attention to the design and management of decision authority can be more important than sophistication in the data and algorithms themselves.

2. Empirical Context

We evaluate the impact of algorithms on managerial decision-making in the Inspectional Services department at the City of Boston ("the City"), where inspectors use their judgment to decide which restaurants to inspect before carrying out the inspections. The City employed approximately 20-30 inspectors, assigned to at least one of 22 Boston wards or "neighborhoods." Inspectors find a large range of violations across their inspections, ranging from 0 to 60 weighted violations per inspection across 2007-2015 (Appendix Figure 1). Weights are assigned based on the severity of the violation: Level I (1 point) corresponds to non-critical violations such as building defects or standing water. Level II (2 points) are "Critical Violations" such as the presence of fruit flies, which are more likely to create food contamination, illness or environmental hazard.

Level III (5 points) 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.

This context provides several research advantages. First, while the strategy of which restaurants to inspect may be complex, a key component of the strategy involves predicting which businesses will have violations, which raises the potential for algorithms to enhance decision-making (Agrawal et al (2019)). The main objective defined by Boston's Head Inspector is to incapacitate establishments that pose the highest risk to public health.[2] Thus, decision quality depends on inspectors' ability to prioritize restaurants according to their likelihood of violation, flagging restaurants with the highest risk to public health as early as possible. While inspectors are encouraged to conduct inspections in geographic proximity whenever possible, this is considered secondary to inspecting restaurants with greater likelihoods of violation.

Second, there are readily accessible data one might use to potentially improve these predictions – such as historical data the City already has access to, and external data (e.g. from platforms such as Yelp, Twitter, or TripAdvisor). The data used in this implementation are similar to what was used in Glaeser et al (2016), which found that algorithms could in principle help to identify a much larger number of violations.

Third, inspectors both possess experience and insight to inform their decisions and are motivated to prioritize higher-risk restaurants—providing a meaningful estimate for human judgment. Many inspectors have been working with the City for several years and have relevant expertise. They are assigned to a particular ward for approximately two years, which balances learning about restaurants with reducing possibilities for strategic behavior that may lead to regulatory capture. Furthermore, as complaints about unsanitary conditions or illness require inspections within a specific time period, inspectors can prevent uncompensated increases in their workload by prioritizing inspections with greater likelihoods of violation. The quality of their inspections can also play a role in career opportunities.

[2] While there are additional possible objectives, such as deterring restaurants from committing violations or ensuring fairness in the inspection allocation, our discussions with the department highlighted the primary importance of identifying restaurants with the highest likelihood of health violations.

Fourth, improving the targeting of inspections has a direct impact on organizational performance. Inspectors are responsible for inspecting all establishments in their ward multiple times a year: higher-risk facilities like hospitals and schools require three inspections per year, while restaurants 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. Thus, better prioritizing inspections can improve the allocation of inspectors' scarce time.

3. Empirical Design

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

The empirical design 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-poor algorithm) used the average number of violations across historical inspections to rank restaurants in each ward from most to least likely to have violations. The third 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., reservations), business ambience (e.g., children-friendly), and neighborhood.[4] While there are certainly more sophisticated approaches that might yield higher-quality insights, this algorithm bundled a comparatively more sophisticated model and richer data

[3] This wording was chosen by the City as the most natural way to obtain inspector rankings. Restaurants with a mandated priority to inspect include high-risk establishments (e.g., hospitals and nursing homes), re-inspections, and restaurants flagged by complaints. We excluded these establishments to assess how inspectors prioritized restaurants.

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

than the "data-poor" algorithm, as one way to emulate common practices by firms to invest in more complex technologies and data.

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 to list 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 docket.[6]

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

Because inspectors were asked to first rank their own choices, it is easier to understand what inspectors' counterfactual decisions would have been without algorithms. Moreover, the variation in the degree of algorithmic sophistication sheds light on how features of different algorithms impact outcomes. Last, randomizing the order of restaurants on the docket makes it possible to identify whether algorithmic methods effectively identified restaurants with a higher number of violations.

3.1 Data and Empirical Approach

The main data we observe is anonymized data on rankings and inspection results. However, several important implementation issues led to empirical challenges.

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

[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 include all restaurants inspectors had prioritized.

Second, the City modified the docket generation process for the last two periods, after observing that inspectors could not complete the dockets. For these 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.

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 only 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 data from historical inspections or Yelp.

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-ranked restaurants provides insight into how the top-ranked restaurants under each of the three methods differ, and whether restaurants that were ranked highly by algorithms have a higher number of violations.

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 substantial overlaps between the methods, especially those using algorithms, with 176 restaurants (26%) ranked in the top 20 by at least two methods. 108 (16%) are ranked in the top 20 by the data-rich algorithm alone, 97 (14%) by the 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; δ accounts for overlaps between methods and represents the mean expected difference in weighted violations for a restaurant ranked by multiple methods.

We explore robustness of the results across alternative subsamples. We vary the threshold of the top 20 across 10-30, as well as the full sample of inspected restaurants. We also account for changes in the docket-generation process by restricting our sample to the first two periods before the modification occurred.

Lastly, given the selection present in our data due to only a subset of restaurants being inspected, we evaluate how selection bias might impact our estimates of the gains from algorithms.

4. Results

We find large gains in identifying violations from using algorithms: algorithms identify restaurants with over 50% more violations on average compared to those prioritized by inspectors. The largest gains stem from using any data, rather than algorithmic sophistication. Despite these gains, we find that inspectors were half as likely to follow algorithmic recommendations compared to restaurants that they ranked themselves.

4.1 The gains from algorithms

We find that algorithms identify restaurants with more violations than those prioritized by inspectors. Table 1 Column 1 shows a comparison of weighted violations by restaurants ranked by one of the methods alone or by multiple methods. Restaurants ranked by inspectors alone have 6.8 violations on average, which is equivalent to having a Level II and a Level III violation. Our estimates of the gains from algorithms over human judgment, β and γ, are 5.03 and 4.88 respectively, which represents a difference of targeting a restaurant with one more Level III violation. The coefficients on the two algorithmic methods are not statistically different, although the data-rich algorithm used both richer data and a more sophisticated algorithm.

In Column 2 of Table 1, we explore a specification that accounts for restaurants ranked by inspectors that were also ranked by one of the algorithms. The constant term shows the mean number of weighted violations for both restaurants 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. We also separate out restaurants ranked by both algorithms and all three methods, and find that these increase the violations found by twofold.

These results 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 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 inspectors may not have been prioritizing restaurants with more violations. In our interpretations, we assume this method represents inspector judgment on restaurants with the highest likelihood of violation, as the wording was chosen by the City as the most natural way to obtain inspector rankings. As described in Section 2, inspectors were trained to prioritize restaurants by their likelihood of violations, and had some incentive to do so as any high-risk restaurants later flagged through complaints would increase their workload.

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

rather than 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 strategic behavior that might lead to regulatory capture, eventually reducing the efficacy of this approach. Lastly, while predicting violations are part of the managerial problem, they are unlikely to be the full problem. To the extent that inspections are meant to do more than help rectify existing problems, one may not want to prioritize solely based on these predictions.

4.2 Decision authority and non-compliance

Despite these gains from algorithms, inspectors were less likely to inspect algorithmically-ranked restaurants compared to those based on their own judgment.

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. Mapping these numbers to 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 58% of the 293 restaurants that they alone ranked in the top 20, but only inspected 27% and 29% of the 108 and 97 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 inspector, with the red line indicating what the percentage breakdown would have been if the inspector had followed the dockets. While we observe some heterogeneity, nearly all inspectors inspected more restaurants prioritized by their own judgment compared to those ranked by algorithms.

While this non-compliance raises an important challenge for organizations in realizing gains from algorithms in practice, 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 selected algorithm-ranked restaurants with higher likelihoods 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 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 should 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 judgment 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 inspectors 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 uninspected restaurants are much cleaner than the inspected ones, which would affect the magnitude of gains from algorithms one could expect 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 other objectives such as minimizing geographic distance or inspecting restaurants that have gone longer without an inspection, (3) inspectors' priors on what predicts violations. We find the most supportive evidence for this last explanation.

One potential explanation is algorithmic aversion, which has been shown to play a role in some settings (e.g., Dietvorst et al. 2015). However, the department chose to not explicitly communicate that these recommendations were driven by algorithms—only stating that they supplemented inspectors' lists with restaurants prioritized using data, at which inspectors expressed enthusiasm. If the department had, it could have led to different responses, and may have even increased rates of non-compliance. But, as algorithms in this setting had the effect of simply suggesting a different set of restaurants to inspect, algorithm aversion may be less likely to explain the effects we observe.

Another explanation may be that inspectors were balancing another objective than the number of violations, such as geographic distance, thus sacrificing targeting restaurants with higher violations to reduce the distance that they traveled. To explore this, we evaluate the 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 (Figure 2).

We also explore whether inspectors may have been more sensitive to overdue inspections, by examining the number of days elapsed since the last inspection. We find that inspectors tended to slightly prioritize restaurants that were more overdue, suggesting that objectives may not have been fully aligned. However, differences in the number of days elapsed across inspected versus non-inspected restaurants, or restaurants ranked highly by the inspectors compared to those ranked by the algorithms, are not statistically significant (Table 4).

Finally, anecdotal evidence on inspector decisions seems to broadly suggest the presence of strong priors. Discussions with individuals involved with the implementation suggest that inspectors might have prioritized restaurants with certain features that they viewed as being correlated with violations, such as chains, lower-end businesses, and seafood restaurants—simple rules based on intuition, which may have helped their decision-making prior to using algorithms

(Sull and Eisenhardt (2015)). This raises the potential that inspectors appear to have overridden algorithmic recommendations when they conflicted with this intuition. We find suggestive evidence consistent with this interpretation in the summary statistics of restaurants ranked by each method and inspected (Table 4).

While this analysis provides little conclusive evidence on mechanisms, it raises the possibility that allocating decision authority to decision-makers may prevent organizations from realizing gains from algorithms in decision-making. In this case, simple rules of thumb, which may have provided advantages for decision-making in the past, may have ended up as an impediment when using algorithms for decision-making. This is consistent with evidence found by Hoffman et al (2018), where managers who appear to hire against job test scores ended up with worse average hires. As theorized by Athey et al (2020), whether to allocate decision authority to decision-makers compared to algorithms 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, how biased they may be, and how well they can predict compared to algorithms. 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 able to exercise discretion more carefully.

However, these findings—and the extent to which decision-makers may use their discretion to reject algorithmic recommendations—may also crucially depend on the organizational context and practices. One key choice variable may be how to communicate about the algorithm being implemented, as further explanations about the algorithm and the motivation for using it may help decision-makers better apply their discretion. Similarly, clarity on organizational objectives and higher-powered incentives may help better align decision-makers.

5. Discussion

Our results show a clear role for algorithms in improving decisions, but also highlight that managing decision authority is an important issue. Even a simple algorithm based on internal historical data has the potential to better prioritize restaurants relative to human judgment. Moreover, much of the gains stemmed from simply integrating data into the decision process,

rather than a more sophisticated algorithm. Yet despite these gains, inspectors frequently chose to prioritize restaurants based on their own judgment rather than algorithms.

Our analysis has important limitations. First, our analysis takes the goal of the department as given. While the department's goal in this context was to prioritize based solely on the severity of violations, in practice, one can imagine a variety of other goals that departments might want to incorporate—such as adjusting for travel distance, recency of inspection, or more easily rectifiable violations, and a simple predictive algorithm may not be completely aligned with their objectives. More broadly, if inspections also deter future violations, then this would suggest that a department may want to change its approach to prioritization. Furthermore, to the extent that behavior changes over time (whether through deterrence or other mechanisms), the effect of implementing different targeting strategies could vary. Second, our analysis assumes that inspections accurately capture true violations. To the extent that violations are inaccurate or biased, then predictions based on those would also be biased. Third, we examine one specific data set, within one particular context. Other datasets or algorithms might be more productive than these approaches, and organizations need to carefully consider the quality of their data, and the noise and bias present. Similarly, the compliance patterns we observe may not generalize to other settings with different communication and organizational dynamics.

Stepping back from this application, organizations are increasingly investing in technologies to support their decision-making, and our findings speak to some of the promise and challenge involved in implementing such approaches at scale. While we cannot claim that these findings fully generalize to other settings, they highlight the importance of carefully considering how decision authority should be allocated and managed. This particular decision context is characterized by moderate complexity, with higher costs for mistakes that make some degree of human supervision valuable, and our findings may be most generalizable to similar contexts. In settings with higher complexity and richer data, the benefits of algorithmic inputs to decision-making may be far higher than what we find, suggesting that decision authority may be better allocated to algorithms.

However, the solution may rarely be as simple as removing decision authority from human decision-makers. For many decisions in managerial contexts, removing humans from the decision process may involve substantial risks, and some degree of human supervision may remain necessary for edge cases. Much work remains to be done to further understand how organizations

can effectively implement algorithms for decision-making without removing managerial discretion. Some promising directions for future work might explore how decision processes can be redesigned and what organizational practices may help decision-makers learn to better inform their decisions using data.

References

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

------ (2019). Artificial intelligence: the ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives*, *33*(2):31-50.

Athey, S., Bryan, K., & Gans, J. (2020). The allocation of decision authority to human and artificial intelligence. *AEA Papers and Proceedings* (Vol.110:80-84).

Bajari, P., Chernozhukov, V., Hortaçsu, A., & Suzuki, J. (2019, May). The impact of big data on firm performance: An empirical investigation. In *AEA Papers and Proceedings* (Vol. 109, pp. 33-37).

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

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

Brynjolfsson, E., & McElheran, K. (2019). Data in action: data-driven decision making and predictive analytics in US manufacturing. *Rotman School of Management Working Paper*, (3422397).

Brynjolfsson, E., Jin, W., & McElheran, K. (2021). The Power of Prediction. Working Paper.

Camuffo, A., Cordova, A., Gambardella, A., & Spina, C. (2020). A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial. *Management Science*, *66*(2):564-586.

Choudhury, P., Starr, E., Agarwal, R. (2020). *Machine learning and human capital: Experimental evidence on productivity complementarities*. Strategic Management Journal.

Cowgill, B. (2019). Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening. Working Paper.

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

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

Hoffman, M., Kahn, L., Li, D. (2018). "Discretion in hiring." *The Quarterly Journal of Economics* 133(2):765-800.

Glaeser, E., Hillis, A., Kominers, S., & 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, G. & Lee, J. (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.

Kim, H. (2020). The Value of Competitor Information: Evidence from a Field Experiment. Working Paper.

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

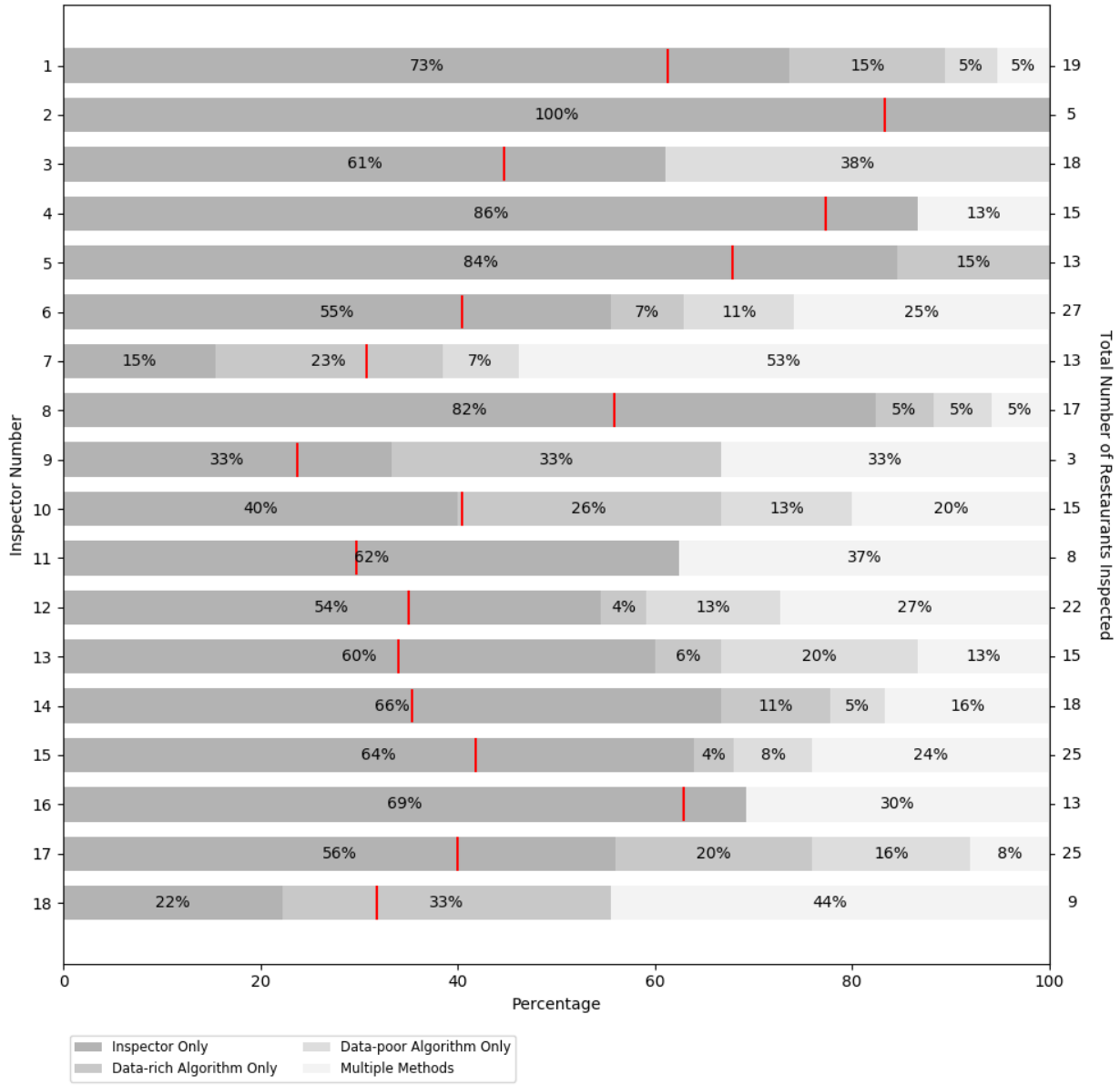
McAfee, A., Brynjolfsson (2012). Big data: the management revolution. *Harvard Business Review*, *90*(10):60-68.

Lehman, S. (2014). Twitter helps Chicago find sources of food poisoning. Reuters Health. https://www.reuters.com/article/us-chicago-twitter-food-poisoning/twitter-helps-chicago-find-sources-of-food-poisoning-idUSKBN0GQ25820140826

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

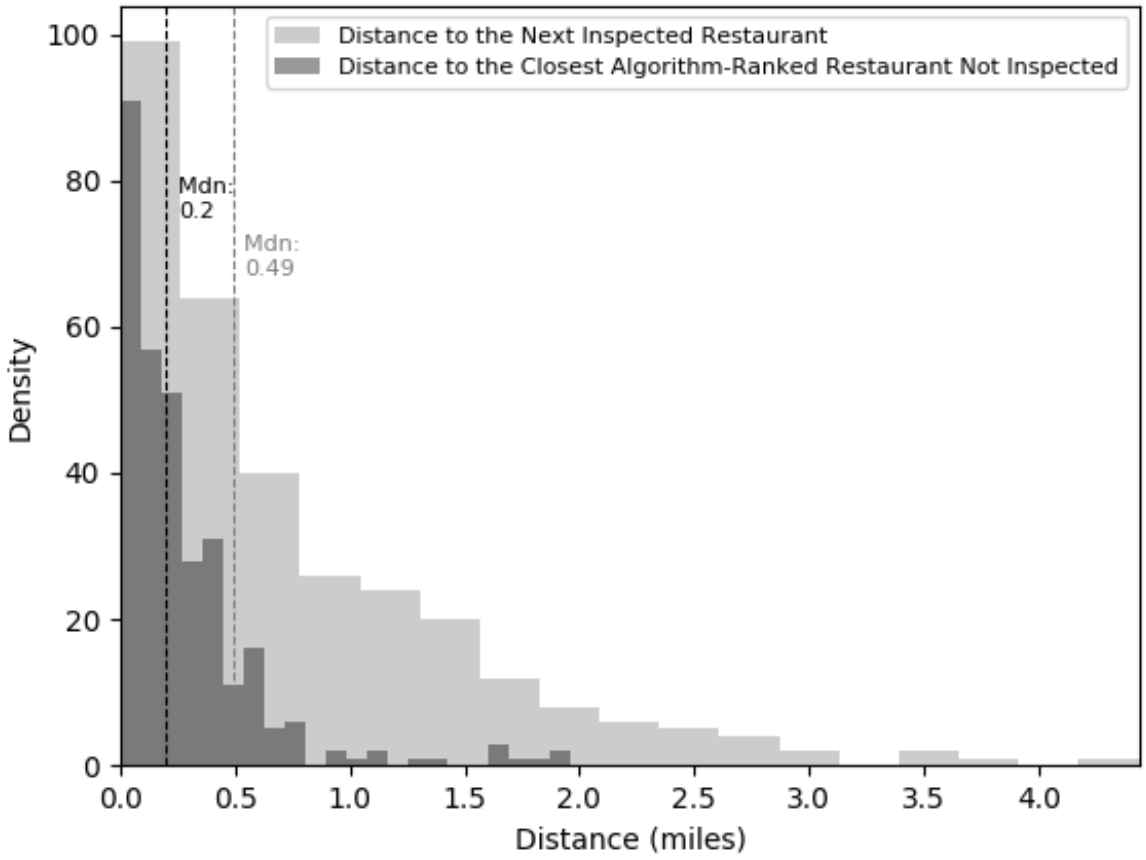
Sull, D. and Eisenhardt, K. (2015). Simple rules: How to thrive in a complex world. Houghton Mifflin Harcourt.

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: The Informational Gains from Algorithms

	(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.

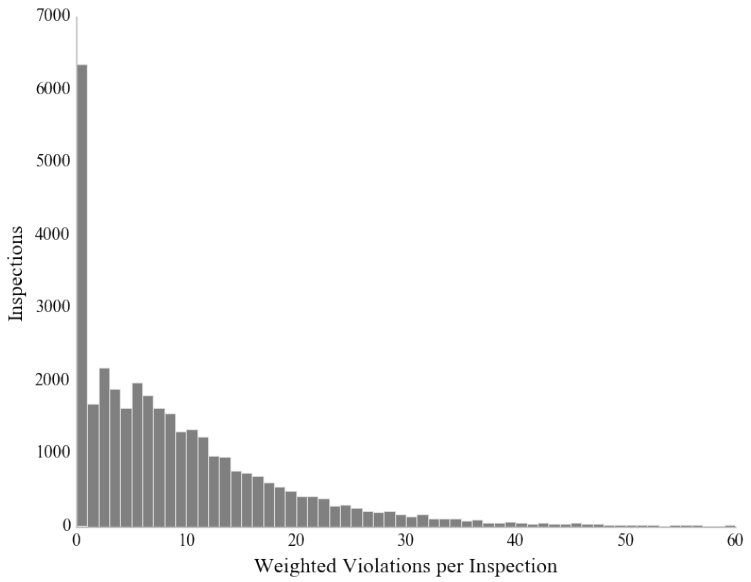
Table 4: Characteristics of Ranked and Inspected Restaurants

Panel A: Restaurant Ranked in the Top 20 by Each Method					
	(1) Data-Rich Algorithm Only	*(2) Data-Poor Algorithm Only*	*(3) 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
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
Days Since Last Inspection	153.43	164.91	181.21	0.23	0.39

Panel B: Inspected vs. Non-Inspected Restaurants		
	Not Inspected	*Inspected*
Chain	0.047	0.06
	(0.011)	(0.02)
Yelp Rating	2.94	2.93
	(0.07)	(0.11)
Review Count	158.3	140.9
	(13.46)	(16.16)
Seafood	0.06	0.04
	(0.01)	(0.01)
Restaurant Age	4.34	5.47
	(0.58)	(1.17)
Price Range	1.31	1.23
	(0.05)	(0.06)
Accepts Reservations	0.28	0.16***
	(0.02)	(0.03)
Table Service	0.42	0.32***
	(0.03)	(0.03)
Days since Last Inspection	172.46	180.74
	(9.08)	(6.68)

Panel A 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. Panel B compares the attributes of inspected and non-inspected restaurants.

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.