ESSAY

Beyond prediction: Using big data for policy problems

Susan Athey

Machine-learning prediction methods have been extremely productive in applications ranging from medicine to allocating fire and health inspectors in cities. However, there are a number of gaps between making a prediction and making a decision, and underlying assumptions need to be understood in order to optimize data-driven decision-making.

A recent explosion of analysis in science, industry, and government seeks to use "big data" for a variety of problems. Increasingly, big-data applications make use of the toolbox from supervised machine learning (SML), in which software programs take as input training data sets and estimate or "learn" parameters that can be used to make predictions on new data. In describing the potential of SML for clinical medicine, Obermeyer *et al.* (*1*) have commented that "Machine learning…approaches problems as a doctor progressing through residency might: by learning rules from data. Starting with patient-level observations, algorithms sift through vast numbers of variables, looking for combinations that reliably predict outcomes…where machine learning shines is in handling enormous numbers of predictors—sometimes, remarkably, more predictors than observations—and combining them in nonlinear and highly interactive ways."

SML techniques emerged primarily from computer science and engineering, and they have diffused widely in engineering applications such as search engines and image classification. More recently, the number of applications of SML to scientific and policy problems outside of computer science and engineering has grown. In the public sector, SML models have been introduced in criminal justice [e.g., (*2*)]; for predicting economic well-being at a granular level using mobile data, satellite imagery, or Google Street View (*3*, *4*, *5*); and for allocating fire and health inspectors in cities (*6*), as well as a variety of other urban applications (*7*). The techniques have also been used to classify the political bias of text (*8*) or the sentiment of reviews (*9*). In medicine, SML-based predictive algorithms have been implemented in hospitals to prioritize patients for medical interventions based on their predicted risk of complications (*10*), and in a wide variety of additional medical applications, including personalized medicine (*1*).

The rapid diffusion of SML methods can be in part attributed to advances in availability of data, computational techniques and resources, data analysis techniques, and open-source software. Another factor is the simplicity of the problems the techniques are designed to solve. Very few assumptions are required for off-the-shelf prediction techniques to work: The environment must be stable, and the units whose behavior is being studied should not interact or "interfere" with one another. In many applications, SML techniques can be successfully applied by data scientists with little knowledge of the problem domain. For example, the company Kaggle hosts prediction competitions (www.kaggle.com/competitions) in which a sponsor provides a data set, and contestants around the world can submit entries, often predicting successfully despite limited context about the problem.

However, much less attention has been paid to the limitations of pure prediction methods. When SML applications are used "off the shelf" without understanding the underlying assumptions or ensuring that conditions like stability are met, then the validity and usefulness of the conclusions can be compromised. A deeper question concerns whether a given problem can be solved using only techniques for prediction, or whether statistical approaches to estimating the causal effect of an intervention are required.

Kleinberg *et al.* (*11*) highlight a case in which off-the-shelf SML techniques can partially, but not fully, address a resource allocation problem in health policy. They consider the problem of deciding which otherwise-eligible patients should not be given hip replacement surgery through Medicare. They use SML to predict the probability that a candidate for joint replacement would die within a year from other causes, and identify patients who are at particularly high risk and should not receive joint replacement surgery. They argue that "benefits accrue over time, so surgery only makes sense if someone lives long enough to enjoy them; joint replacement for someone who dies soon afterward is futile—a waste of money and an unnecessary painful imposition on the last few months of life" (p. 493). In this class of problems, the rationale for focusing on prediction is clear; the average effect of an intervention is known to be negative in certain states of the world (if the patient will die soon), so that predicting the state of the world is sufficient for the decision to forgo the surgery. However, the authors highlight the fact that pure prediction methods do not answer the more complex question of which patients should be given the highest priority to receive surgery, among those who are likely to survive more than a year. The full resource allocation problem requires estimates of heterogeneity in the effect of surgery, for example, because some patients may have higher rates of surgical complications than others. The question of optimally allocating a scarce resource (hip replacement surgery) to the patients for whom the causal effect of the surgery on patient welfare is highest is a much harder problem, one that generally requires answering counterfactual questions: What would happen under a variety of alternative assignment policies, policies that have never been implemented before?

In another resource allocation example, it is common in industry to use SML to predict the probability of customer "churn," in which a customer abandons a company or service, and the company responds by allocating interventions

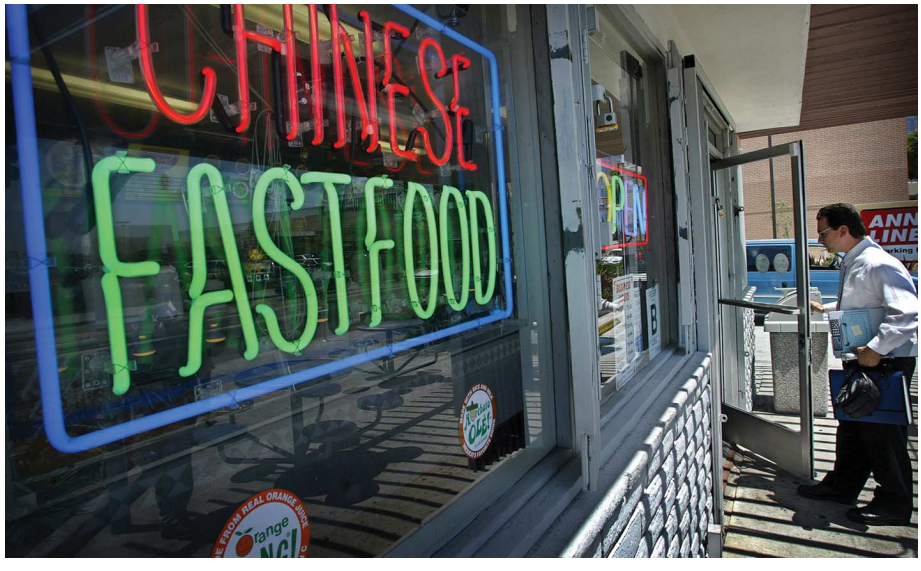

PHOTO: RICARDO DEARATANHA/*LOS ANGELES TIMES*

Fig. 1. What is the best way to allocate food-safety inspectors?

Graduate School of Business, Stanford University, Stanford, CA, USA.
Email: athey@stanford.edu

(such as outreach by salespeople) to those customers at highest risk of churn. Ascarza (*12*) documented firms following this type of practice, and then used methods from the causal inference literature to provide empirical evidence that allocating resources according to a simplistic predictive model is not optimal. The overlap between the group with highest risk of churning and the group who would respond most to interventions was only 50%. Thus, treating the problem of retaining customers as if it were a prediction problem yielded lower payoffs to the firm.

A public-sector resource allocation problem is the question of how a city should allocate building inspectors optimally to minimize safety or health violations. New York City's Firecast algorithm allocates fire inspectors according to the predicted probability of a violation being detected upon inspection, and Glaeser *et al.* (*6*) developed and implemented a similar system for allocating health inspectors to restaurants in Boston, with preliminary estimates showing a 30 to 50% improvement in the number of violations found per inspection.

The decision problem of how to optimally allocate inspectors would fall squarely in the prediction domain if the following simplifying assumptions were true: (i) The behavior of the individual establishments being inspected is fixed; and (ii) when problems are identified, they can be immediately fixed at a low cost that does not vary across units. Knowing which establishments are more likely to have violations would be equivalent to knowing which ones should be inspected. However, a more realistic setting incorporates heterogeneity across units: A building may be at higher risk of fire due to old wiring, but other considerations make it difficult to replace the wiring. Other units may have lower predicted risk, but it may be easy and inexpensive to make substantial improvements. Another consideration is responsiveness; if violations entail fines, some firms may be more sensitive to the prospect of fines than others. Overall, solving the city's inspection allocation problem involves estimating the causal effect of inspection policies: What is the expected improvement in overall quality of units (e.g., food poisoning rates) in the city under a new inspector allocation regime?

Thus, prediction and causal inference are distinct (though closely related) problems. Outside of randomized experiments, causal inference is only possible when the analyst makes assumptions beyond those required for prediction methods, assumptions that typically are not directly testable and thus require domain expertise to verify. A large literature in causal inference that spans multiple disciplines (social science, computer science, medicine, statistics, epidemiology, and engineering) has emerged to analyze this type of problem [see Imbens and Rubin (*13*) for a review]. One approach to estimating causal effects using data that were not generated from a randomized experiment is to adjust for factors that led to differential inspection probabilities in the past, and then to estimate the effect of inspection on restaurant-specific health outcomes (perhaps using audits). Recent methodological advances focus on adjusting for observed confounders in big-data applications [e.g., (*14*–*16*)]. A theme in this literature is that off-the-shelf prediction methods from SML lead to biased estimates of causal effects, but that consistent and efficient estimation of causal effects can be achieved by modifying SML techniques.

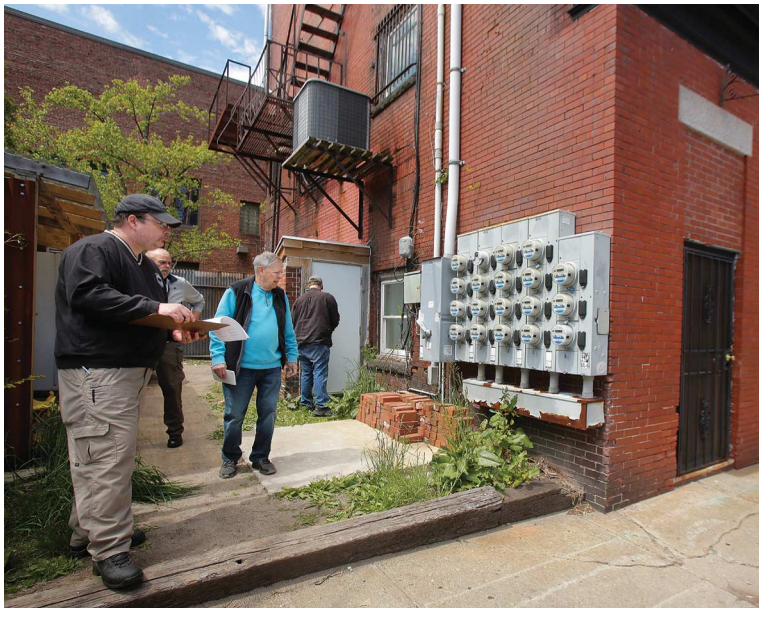

Fig. 2. What is the causal effect of inspections?

Another approach to estimating causal effects is to make use of designed experiments. Blake *et al.* (*17*) used a city-based difference-in-difference methodology (implementing a new policy in a randomly selected set of "treatment" cities while controlling for time trends by also measuring outcomes in a set of "control" cities) to evaluate the causal effect of search advertising for eBay. Like many search advertisers, eBay relied on historical data to measure the benefit of search advertising, but it did not attempt to separate correlation from causality. Rather, eBay measured advertising effectiveness with a simple predictive model in which clicks were used to predict sales, finding that the return on investment for advertising clicks (that is, the ratio of eBay sales attributed to clicks to the cost of the advertising clicks) was about 1400%. Using the experimental data to measure the causal effect of the advertisements, the authors found that the true return on investment was −63%. Part of the gap between the naïve analysis and the results from the experiment arose because many people who clicked on eBay search advertisements would have purchased items from eBay, anyway. Although a click on an eBay ad was a strong predictor of a sale—consumers typically purchased right after clicking—the experiment revealed that a click did not have nearly as large a causal effect, because the consumers who clicked were likely to purchase, anyway.

Beyond resource allocation problems, the distinction between pure prediction and causal inference has been the subject of decades of odological and empirical research in many disciplines. Economics has placed particular focus on this distinction, perhaps because some of the most fundamental economic questions, such as how consumer demand varies with price, cannot be answered with purely predictive models. For example, how much of a product would consumers buy at different (hypothetical) price levels? Although it might seem straightforward to use off-the-shelf SML to predict the outcome "quantity sold" with the price level as an explanatory "feature," in practice, this approach fails badly if it is used as a method to estimate the causal effect of price on quantity sold. Suppose that an analyst has historical data from hotel prices and occupancy rates. Typically, prices and occupancy are positively correlated because the existing pricing policy for hotels (often implemented through yield management software) specifies that hotels raise their prices as they become more fully booked. Off-the-shelf applications of SML techniques are designed to answer the following type of question: If an analyst is told that on a particular day, prices were unusually high, what is the best prediction of occupancy on that day? The correct answer is that occupancy is likely to be high. By contrast, the question of the effect of changing the pricing policy is a causal question, and common experience indicates that if the firm implemented a new policy to systematically raise prices by 5% everywhere, it would be unlikely to sell more hotel rooms. A different set of statistical techniques is required to answer this question, perhaps exploiting "natural experiments" in the data or an approach known as "instrumental variables" [see (*13*) for a review of these techniques]. Recently, several authors have combined advances from SML with this traditionally "small data" set of methods, both for estimating average causal effects (*18*) and for personalized estimates of causal effects (*19*).

Beyond the distinction between prediction and causal inference, methods optimized solely

for prediction also do not account for other factors that may be important in data-driven policy analysis or resource allocation. For example, incentives and manipulability can be important. If a building or restaurant owner anticipates a low probability of being inspected based on these characteristics, he or she may reduce efforts for safety.

In an example of data-driven policy where manipulability played a role, the market pricing system (MPS) of British Columbia is used to set prices for harvest of timber from government-owned land that has been allocated to timber companies under long-term leases. The MPS builds a predictive model using data from timber sold at auctions to predict the prices that would have been obtained if a tract harvested under a long-term lease had instead been sold via auction. However, a lease-holder could potentially have an incentive to bid artificially low in auctions in order to influence the predicted prices for timber harvested under the long-term lease and thus lower their costs of harvesting from long-term leases. As part of the model selection process, the predictive model for MPS was subject to simulations to assess its manipulability by any single large timber company (*20*). The model implemented was not the model with the best possible predictive power to achieve the desired robustness against manipulability.

A number of additional considerations arise when using statistical models in practice. It is sometimes important for stakeholders to understand the reason that a decision has been made, or decision-makers may need to commit a decision rule to memory (e.g., doctors). Transparency and interpretability considerations might lead analysts to sacrifice predictive power in favor of simplicity of a model. Another consideration is fairness, or discrimination. Consumer protection laws for lending in the United States prohibit practices that discriminate on the basis of race. Firms might wish to use SML methods to select among job applicants for interviews; but they might wish to incorporate diversity objectives in the algorithm, or at least prevent inequities by gender or race. These issues have received recent attention in the literature on SML [e.g., (*21*)].

Overall, for big data to achieve its full potential in business, science, and policy, multidisciplinary approaches are needed that build on new computational algorithms from the SML literature, but also that bring in the methods and practical learning from decades of multidisciplinary research using empirical evidence to inform policy. A nascent but rapidly growing body of research takes this approach: For example, the International Conference on Machine Learning (ICML) in 2016 held separate workshops on causal inference, interpretability, and reliability of SML methods, while multidisciplinary research teams at Google (*22*), Facebook (*23*), and Microsoft (*24*) have made available toolkits with scalable algorithms for causal inference, experimental design, and the estimation of optimal resource allocation policies. As the SML research community and other disciplines continue to join together in pursuit of solutions to real-world policy problems using big data, we expect that there will be even greater opportunities for methodological advances, as well as successful implementations, of data-driven policy.

REFERENCES AND NOTES

1. Z. Obermeyer, E. J. Emanuel, *N. Engl. J. Med.* **375**, 1216–1219 (2016).
2. R. Berk, *Criminal Justice Forecasts of Risk: A Machine Learning Approach* (Springer Briefs in Computer Science, 2012).
3. N. Naik, R. Raskar, C. A. Hidalgo, *Am. Econ. Rev.* **106**, 128–132 (2016).
4. J. Blumenstock, G. Cadamuro, R. On, *Science* **350**, 1073–1076 (2015).
5. R. Engstrom, J. Hersh, D. Newhouse, "Poverty in HD: What Does High Resolution Satellite Imagery Reveal about Economic Welfare?" (2016).
6. E. L. Glaeser, A. Hillis, S. D. Kominers, M. Luca, *Am. Econ. Rev.* **106**, 114–118 (2016).
7. E. L. Glaeser, S. D. Kominers, M. Luca, N. Naik, Big data and big cities: The promises and limitations of improved measures of urban life (Technical Report, National Bureau of Economic Research, 2015).
8. J. Grimmer, B. M. Stewart, *Polit. Anal.* **21**, 267–297 (2013).
9. J. S. Kang, P. Kuznetsova, M. Luca, Y. Choi, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2013), pp. 1443–1448.
10. M. Bayati *et al.*, *PLOS ONE* **9**, e109264 (2014).
11. J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, *Am. Econ. Rev.* **105**, 491–495 (2015).
12. E. Ascarza, Retention futility: Targeting high risk customers might be ineffective (2016); available at SSRN.
13. G. W. Imbens, D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge Univ. Press, 2015).
14. M. Dudık, J. Langford, L. Li, in *Proceedings of the 28th International Conference on Machine Learning* (ICML, 2011), pp. 1097–1104.
15. A. Belloni, V. Chernozhukov, C. Hansen, *Rev. Econ. Stud.* **81**, 608–650 (2014).
16. S. Athey, G. Imbens, S. Wager, Approximate residual balancing: De-biased inference of average treatment effects in high dimensions; https://arxiv.org/abs/1604.07125 (2016).
17. T. Blake, C. Nosko, S. Tadelis, *Econometrica* **83**, 155–174 (2015).
18. A. Belloni, D. Chen, V. Chernozhukov, C. Hansen, *Econometrica* **80**, 2369–2429 (2012).
19. S. Athey, J. Tibshirani, S. Wager, Solving heterogeneous estimating equations with gradient forests; https://arxiv.org/abs/1610.01271 (2016).
20. S. Athey, P. Cramton, A. Ingraham, Auction based timber pricing and complementary market reforms in British Columbia, White Paper; ftp://www.cramton.umd.edu/papers2000-2004/ (2002).
21. T. Kamishima *et al.*, "Fairness-aware classifier with prejudice remover regularizer," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Springer, 2012).
22. S. L. Scott, *Appl. Stochastic Models Data Anal.* **31**, 37–45 (2015).
23. E. Bakshy, D. Eckles, M. S. Bernstein, "Designing and deploying online field experiments," *Proceedings of the 23rd International Conference on World Wide Web* (Association for Computing Machinery, 2014).
24. A. Agarwal *et al.*, https://arxiv.org/abs/1606.03966 (2016).

ACKNOWLEDGMENTS

I am grateful to G. Imbens, S. Mullainathan, M. Luca, and S. Wager for helpful conversations.

10.1126/science.aal4321