Predictive Analytics for City Agencies: Lessons from Children's Services

Ravi Shroff, Ph.D.

Center for Urban Science and Progress

New York University

June 6, 2017

Abstract

Many municipal agencies maintain detailed and comprehensive electronic records of their interactions with citizens. These data, in combination with machine learning and statistical techniques, offer the promise of better decision making, and more efficient and equitable service delivery. However, a data scientist employed by an agency to implement these techniques faces numerous and varied choices that cumulatively can have significant real-world consequences. The data scientist, who may be the only person at an agency equipped to understand the technical complexity of a predictive algorithm, therefore bears a good deal of responsibility in making judgments. In this perspective, I use a concrete example from my experience working with New York City's Administration for Children's Services to illustrate the social and technical tradeoffs that can result from choices made in each step of data analysis. Three themes underlie these tradeoffs: the importance of frequent communication between the data scientist, agency leadership, and domain experts; the agency's resources and organizational constraints; and the necessity of an ethical framework to evaluate salient costs and benefits. These themes inform specific recommendations I provide to guide agencies that employ data scientists and rely on their work in designing, testing, and implementing predictive algorithms.

Introduction

City agencies across the United States collect vast amounts of data for record keeping and to understand and improve their operations. Agencies use the data they collect in diverse ways: to compile descriptive reports of criminal activity[1], benchmark building energy usage[2;3], optimize the success rate of fire safety inspections[4], identify suspected terrorists[5], and predict violence[6]. In New York City, the Parks & Recreation Department maintains a database of the health of hundreds of thousands of street trees in the city[7], the Taxi and Limousine Commission authorizes the collection of the time and GPS location of millions of individual taxi trips within the five boroughs[8], and the Department of Transportation has detailed records of current and historical street conditions[9]. Other New York City agencies interact with citizens more directly: the Administration for Children's Services (ACS) investigates alleged abuse or neglect, supervises foster care and juvenile detention, and provides services to support families with children. In the criminal justice system, the Police Department conducts investigative stops, the District Attorney's offices maintain records of court appearances, and the Department of Corrections monitors inmates in city jails. Data recording these interactions usually have common attributes*, such as a date and location, an outcome, characteristics of the interaction, and characteristics of the citizen involved, often including historical interactions with that agency (e.g., a criminal record).

In this article, I examine the role of a data scientist who builds and implements machine learning models to understand agency-citizen interactions, and I provide specific recommendations for agencies that intend to deploy predictive analytics to improve their services.† These issues are illustrated with a detailed example from my own work

*Here, the word "attribute" will refer to characteristics of an individual or interaction in the real world, and "feature" will refer to the numerical or categorical expression of an attribute in data.

†The terms "predictive analytics" and "machine learning" will be used interchangeably to refer to supervised classification tasks. A machine learning algorithm for supervised classification is a set of instructions, implemented by a computer, that determines relationships between features of historical "training" data and known values of a categorical outcome variable. These relationships can be used to make predictions on new data, for which the value of the outcome variable is unknown.

consulting for the Division of Policy, Planning, and Measurement at ACS, but my perspective is also informed by broader research and collaboration experiences with the New York City Police Department (NYPD) and Mayor's Office of Operations, the Manhattan District Attorney's office, and the New York State Office of the Attorney General. The recommendations that conclude this article also highlight forward-thinking practices of ACS leadership that should be adopted by other agencies implementing similar analytic techniques.

The use of interactions data by municipal agencies to build predictive models differs in several key ways from analogous uses in the private sector (particularly the technology sector, which has led the way in the research and application of new machine learning methods). Instead of the common industry goal of encouraging repeat business, public sector implementations of predictive analytics are often motivated by reducing the likelihood of another interaction, such as an application for emergency housing, or the re-entry of a child into foster care. At a structural level, the organizational arrangements and personnel responsible for analytics work vary from agency to agency; the job of "data scientist" may be done by employees with the title of city research scientist, director of management and outcomes, or assistant commissioner of data analytics. The diversity of titles reflects both the amorphous responsibilities of a municipal data scientist—and the fact that such responsibilities may often not belong to just one person—as well as the diffusion of expertise across existing positions. The scale of available data is also relatively small; for example, each year ACS conducts roughly 55,000 investigations of alleged abuse or neglect, and New York City receives several million 911 calls (Google or Facebook, on the other hand, analyze datasets that are many orders of magnitude larger). Moreover, decisions made by city agencies often have high stakes: there are significant and immediate human costs associated with using an algorithm to guide a police officer's decision to stop a citizen, or to allocate preventive services to one family rather than another. By automating and augmenting aspects of decision-making, predictive

algorithms can redirect and increase the scale of real-world interventions.

Given these high stakes, it is particularly important that a data scientist "gets things right" when developing a predictive model. However, designing, testing, and implementing a model involves a myriad of decisions, and I argue that balancing the social and technical tradeoffs of these decisions should be facilitated through agency practices that address three themes: first, the communication between data scientists and other agency employees; second, constraints on resources and organizational practices; and third, the necessity of an ethical framework for evaluating costs and benefits. These practices are essential to "getting things right" (and defining what "right" actually means).

The following sections illustrate a selection of choices and tradeoffs involved in the main stages of a data analysis pipeline, using an example from my work with ACS* predicting *repeat reports of abuse or neglect*. Specifically, I have been applying machine learning methods to estimate the likelihood that a child currently in an investigation of alleged abuse or neglect will have a new report of abuse or neglect made within six months. These methods are trained using hundreds of thousands of records in ACS databases of current and historical investigations, and information about all individuals, child and adult, involved in these investigations. The resulting model is intended to improve operations in several ways, including:

- Providing information and resources to supervisors and staff dealing with investigations, case planning, and case closures;
- Helping match families with services to encourage positive outcomes and mitigate negative outcomes, and identifying gaps in the existing array of services;
- Adjusting performance metrics to account for differences in case difficulty.

The second half of this article provides recommendations—informed, in part, by current and planned ACS practices—that will enable the navigation of social, technical, and

*This work also involves collaboration with domain experts from the City University of New York.

ethical concerns involved in making these choices.

Predicting Repeat Reports

Formalizing the Problem

Applying machine learning techniques to estimate the likelihood of a repeat report requires constructing a training dataset, which includes identifying the sample, features, and outcome variable to be used. Some specifications were provided by a group of domain specialists, and other restrictions were imposed by aspects of existing data, but many decisions required judgment calls.

First, it was necessary to decide what real-world entity each data point represented. Since a given child could be involved in multiple investigations, and a given investigation might involve multiple children, a natural choice was for each row in the data to represent a child-investigation pair. Determining the temporal and geographic extent of the training data came next. Since ACS maintains decades of historical investigations, choosing a time range required balancing computational expense and accuracy. Also, certain features only appeared in the data after the introduction of a new methodology for measuring child safety risk. Weighing the benefits of a richer set of features with the costs of discarding data also influenced decisions on how to restrict the time period under consideration.

The choice of features depends on more than just availability. Many machine learning algorithms can quickly process data with thousands of features[10], and a common philosophy is to use as many features as possible. However, the task of transforming the investigation history of a child into a set of features could be accomplished in a variety of reasonable ways, with no clear best option. For instance, the history could be encoded as a simple count of all past investigations, or of all past investigations satisfying certain criteria (e.g., those with a substantiated allegation of educational neglect), or as

the lengths of time between past investigations, or all of the above. On the other hand, there are a variety of methods for automatically constructing features from data[11]. Such features, while perhaps optimized for predictive accuracy, do not always lend themselves to real-world explanations. Therefore, using these methods may require sacrificing their interpretability for a human audience.

Ethical and legal considerations also influence feature selection. Even if available in the data, regulations may prohibit using a juvenile's criminal record, or require allegations of abuse or neglect to be expunged*. Even if available and legally permissible, features such as a citizen's race or ethnicity can be ethically troubling to use in an algorithm (we did not use these features), as well as location, which can act as a proxy for race[13]. There are also ethical issues around the use of co-occurring interactions. Our data includes the investigative history of a child's siblings, but in the context of predicting repeat investigations of child abuse or neglect, using these data seems less troubling than, for example, using the criminal history of a co-arrested defendant in the context of police decisions about whom to put under surveillance[14].

The outcome variable can also be determined in several ways. In this example, domain experts and leadership from several ACS divisions specified that the appropriate outcome variable was the occurrence of a new report within six months. Other possible outcome variables could be the occurrence of two new reports within six months, one new report within one year, or one new substantiated report of maltreatment in one year[15]. The agency might specify an outcome variable, but the data scientist should still provide input if the prediction task becomes difficult for statistical or computational reasons.

*In New York State, "The record of the report to the central register shall be expunged 10 years after the 18th birthday of the youngest child named in the report. In the case of a child in residential care, the record of the report to the central register shall be expunged 10 years after the reported child's 18th birthday." [12]

Cleaning and Processing Data

Preparing a training dataset for a predictive algorithm often requires dealing with ambiguous or missing data. At times, when an ACS caseworker conducts an investigation involving alleged abuse or neglect of a child, another investigation involving the same child is ongoing. Through consultation with ACS employees, it was determined that the appropriate unit of analysis should be an investigation "window," formed by consolidating overlapping investigations. However, this consolidation introduced multiple possible values for the same feature. For instance, an allegation of educational neglect may be unsubstantiated in the first investigation, and subsequently substantiated in the second investigation. A child's gender might be marked "male" in the first investigation and "female" in subsequent overlapping investigations. At the level of the investigation window, was the allegation of educational neglect substantiated? And what value should be chosen for the child's gender? A natural choice for resolving the ambiguity in gender could be the value that occurs with highest frequency, while allegations could be resolved by choosing the most "severe" value (in this case, substantiated).

Missing or incomplete records also occur in ACS databases. Every investigation should receive an initial Risk Assessment Profile (RAP) score of "low," "moderate," "high," or "very high," but sometimes these values do not appear in the data. Possible solutions would be to exclude the RAP score as a feature in the dataset, to remove the investigations with missing RAP score, or use imputation techniques, with each option affecting the final output. Excluding the RAP score might ultimately lower model performance, but entirely removing investigations with a missing RAP score could introduce bias into predictions if, for example, certain investigations are systematically less likely to have a RAP score recorded.

Non-representative samples of data are also problematic for applying predictive methods. It can be misleading to apply an algorithm trained on one population to a population with a different distribution of attributes. For example, ACS leadership has been clear

in stating that the repeat reports model described in this article will only be used to predict the likelihood that a child *already in their system* will have another investigation of abuse or neglect. In particular, even if required data were available, the model will not be used to make predictions on the general population of children in New York City. It is important to note that in general, if data collection procedures are strongly biased in some systematic way, the only option may be to devise strategies to improve those procedures.

Algorithm Selection and Evaluation

Choosing a specific predictive algorithm depends on numerous factors, from the particular formalization of the problem, to the available computational resources and the data scientist's training. An important tradeoff discussed in the context of the repeat reports model at ACS was the desired balance between model performance and interpretability. An interpretable model may be more likely to be implemented by an agency, and such models are easy to audit and explain to nontechnical audiences[16;17;18]. On the other hand, accuracy matters, as there are real costs to false negatives, in this case, a child who does not receive appropriate services, and false positives, like a costly service directed to a child who will not benefit from it[19]. After a series of conversations with ACS leadership, we adopted a random forest model, and decided that the relative performance gain outweighed the cost to interpretability. In the discussion section below, I provide additional thoughts on how to effectively conduct these conversations.

The planned use of the algorithm's output also affects the choice of model[20]. ACS wants to eventually use the repeat reports model to not only measure which children have a high probability of an adverse outcome, but also to understand features important in the calculation of these probabilities. Whereas linear models provide one natural way to produce these important features (by standardizing the data and examining the relative magnitudes of associated coefficients), we used Gini tests and permutation tests to select

the important features in the random forest.

Choosing appropriate evaluation metrics was also done in conversation with ACS leadership. Domain experts weighed in on the relative importance of model precision (the proportion of children who actually had a repeat report, of those predicted to have a repeat report) and model recall (the proportion of children predicted to have a repeat report, of those who actually had a repeat report). These discussions covered the relative benefits of minimizing the misallocation of scarce resources (e.g., caseworker time), versus detecting every child likely to have another report. Additionally, the amount of data played a role in determining how models were evaluated. Since lots of data were available, model performance was checked with a simple train/test split, but with less data, cross validation could have been a better choice[21].

Presentation and Implementation

When presenting the repeat reports analysis to agency leadership, it was important to show these results in a manner that was both technically sound, yet understandable to members of the audience without any knowledge of predictive analytics. An early presentation included a graphical plot that displayed several different modeling strategies. Afterward, agency leadership chose the model that met their needs for performance and interpretability. In subsequent presentations, results were presented as a simple table of model scores, and the number of investigations, precision, and recall for different decision thresholds.

Although the repeat reports model described here has yet to be implemented, I participated in many planning conversations with ACS employees about how to use the model to inform caseworker decisions. One strategy we discussed was to use an automated visual "dashboard" to highlight investigations in which a child has a high probability of a repeat report, and to alert relevant supervisors. However, this raised concerns about ensuring that caseworkers and supervisors correctly interpreted this dashboard alert. An

alternative approach focuses on matching families to appropriate preventive services: a caseworker would be informed of a recommended service—this recommendation would be linked to the numerical likelihood of a repeat report—but would not see a number or risk category. However, caution must be taken when selecting any numerical boundary (e.g., if a probability of .75 or above is required before certain services are recommended, an individual scoring .74 might receive fewer resources than if a lower boundary were chosen). A responsible agency should check if different choices of category boundaries result in, for example, disparate impacts on vulnerable or otherwise sensitive subpopulations.

Discussion and Recommendations

The detailed example above illustrates three common themes that govern social and technical tradeoffs associated with predictive techniques in the public sector: the frequency of communication between data scientists, agency leadership, and domain experts; the organizational and resource capacities of the agency; and the presence of a framework for making ethical judgment calls. For each of these three themes, I provide recommendations—particularly for those who employ data scientists and rely on their work—to guide the process of implementing machine learning models to understand and utilize data. The recommendations below are based in part on good practices that are either already implemented or in the process of being implemented by ACS and other city and state agencies, but may also be relevant in industry. On the other hand, it is important to recognize that no single solution or set of practices will cover all issues that can arise.

Communication

Although at a high level, predictive modeling tasks may appear similar between agencies, available data will vary in type, quality, and generating process, and applications of

analytics will be regulated by different laws, and raise different ethical questions. It is rare that a data scientist alone (particularly a consultant or new employee) possesses both the domain knowledge and technical expertise to understand and implement useful machine learning techniques from end to end. It is crucial for data scientists to work in tandem with agency leadership, who can define goals and set priorities, along with veteran "on-the-ground" employees, to understand the constraints of the data-generating process and data recording procedures.

Since the process of formalizing a problem, in particular, is one of continuous refinement, for a data scientist to make the best possible choices when developing a predictive analytics solution, he or she must be in constant communication—from start to finish—with both agency leadership and domain experts. At ACS, I participated in weekly calls with associate and assistant commissioners, as well as regular face-to-face discussions with directors who had decades of child welfare experience. The commissioners helped define the modeling task and outcome variable (e.g., to define a repeat report as one which occurs within six months), and the directors helped me understand the best ways to resolve ambiguities when merging overlapping investigations, among other things. Therefore, my first recommendation is that **agencies schedule frequent, recurring meetings that emphasize three-way communication between senior leadership, domain experts, and data scientists**.

This communication must go both ways. A data scientist not only needs to learn the answers to domain-related questions and agency priorities, but also needs to communicate details of the analytic approach in such a manner that agency employees with a limited technical background can appreciate the associated benefits and drawbacks. Technical expertise in public agencies often diminishes as one goes up the chain of command, to the point that agency leadership, who need to provide critical direction to the data scientist, may not be in a position to understand the issues that need to be resolved. The data scientist in fact might be the only person in the organization who understands how a

given machine learning model works.

Specific examples of issues that I needed to communicate to ACS leadership were:

- The relative difficulty of implementing different models from a statistical standpoint, e.g., a model that constantly updates in response to new feature values versus one that makes predictions at a fixed point in time.
- Whether sufficient data existed to predict the phenomenon of interest with a specified level of performance.
- A non-technical description of the random forest algorithm.
- The difference between measuring feature importance for predictive performance, measuring the correlation between a feature and the outcome variable, and making real-world decisions based on a particular feature.

On the other hand, agency leadership may need to communicate issues like sensitivity to language (at ACS, we were cautioned against using the word "risk" because of negative connotations), the structure of an agency employee's day, and how an employee would use a tool that incorporates the results of an algorithm. At the end of the day, it is often the responsibility of the data scientist to alert agency leadership to possible tradeoffs since they may not be aware of these ahead of time. Accordingly, my second recommendation is that **agencies should prioritize communication skills when hiring data scientists, specifically the ability to accurately and concisely explain technical concepts to those with differing backgrounds and expertise**.

In order to facilitate a review of a data scientist's models, their work should be reproducible, implying in part that any choices made are clearly documented and can be replicated exactly. Besides reproducibility, model simplicity and transparency is another worthy goal when choosing a predictive algorithm. As explained in the example above, we decided to, at least for the time being, adopt a more opaque random forest algorithm for the ACS repeat reports model, but in some settings, simpler models

such as weighted checklists can be adopted with little to no drop in performance[22]. My next recommendation would therefore be to **use simple, interpretable models when possible—subject to performance requirements—rather than complicated black-box models**, for their transparency and ease of communication.

Capacity

Many New York City agencies have large budgets (ACS has a budget of close to three billion dollars[23]), as well as IT departments with the engineering capability to automate data ingestion and processing pipelines. However, unlike similarly-sized companies in the technology sector, few agencies currently have dedicated teams of data scientists, and those teams that exist are often small. Designing, building, testing, and deploying an end-to-end predictive analytics process involves important computational and data-related concerns, and these concerns can be addressed by borrowing existing best practices from industry. Regarding the connections between IT departments and data scientists, legacy systems should be updated, the latest software for statistical modeling should be installed, and processes such as code reviews should be implemented. Internal clusters or cloud computing resources should be funded, as available computational resources guide algorithm selection. Data privacy concerns are also paramount here; it is hard to think of data more sensitive than the personal information of the vulnerable children and families that ACS serves, and information security practices should be a part of training.

While an initial solution may be to hire an external consultant, or to purchase a ready-to-use product from the private sector, city agencies in New York must also be extremely responsive to demands by the Mayor's Office. In addition, as described above, the process of building predictive models is an iterative one, requiring repeated modification and robustness checks. Therefore, these temporary options are ultimately not as sustainable as building a dedicated internal team who can customize, maintain, and update predictive

models as needed. Besides engineering and programming expertise, statistical expertise is important. Relationships should be forged with trusted statisticians to conduct external methodology reviews, but ideally statistical knowledge should be developed in-house. ACS has emphasized building this internal capacity, and has focused on training and knowledge transfer to transition the repeat reports model (and other predictive models I developed) to their in-house team. Along these lines, I suggest that **agencies focus on building internal capacity by hiring dedicated data scientists and adopting industry-standard engineering and security practices.**

The deployment of predictive models should be accompanied by their rigorous evaluation, when possible. Ideally, a controlled experiment could be performed, with the algorithm's output only available to a randomly chosen subset of agency employees. However, the imperatives of quickly rolling out a working tool (and minimizing social cost) might force an agency not to conduct a proper evaluation. Of course, an algorithm may be well designed in terms of predicting accurately, but the chosen implementation may still be ineffective. Moreover, it is important to recognize that just as data is used to train predictive models, careful ex-post evaluation of these models can suggest better ways to measure and collect new data. Therefore, **agencies should emphasize model evaluation and improve data collection practices.**

Ethical Considerations

The detailed example from ACS described above illustrates the variety of technical choices that can be made by data scientists at city agencies, and the ethical implications of these choices. Note that similar applications of predictive methods by other agencies can raise very different ethical questions. For example, the Chicago Police Department's "heat list"[24] identifies individuals likely to be involved in future gun violence, with the goal of monitoring these individuals' whereabouts and activities. One obvious way these two examples differ is that the goal of the repeat reports model is to improve the allocation

of benefits (preventive services and case reviews) to vulnerable citizens, whereas the heat list aims to allocate sanctions (surveillance and punishment) to "risky" individuals.

Many ethical questions that arise in the design and use of predictive analytics have no easy answers. One approach to tackling these questions is by bringing together diverse stakeholders and creating organizational structures tasked specifically with addressing ethical concerns. ACS is currently using an approach involving an external ethics advisory group that reviews proposals for predictive models in conjunction with an internal committee composed of senior leadership. The external group, consisting of racially and professionally diverse stakeholders—including parents and youth involved with ACS (and their legal advocates), contracted service providers, and industry and academic experts—will provide valuable advice on how to balance the short and long-term costs and benefits of the myriad decisions made when implementing machine learning models. This ethics review group will also identify opportunities to engage other external stakeholders, and will commit to regular contacts with ACS through meetings and phone calls. Following the example of ACS, I advise that **agencies set up internal and external boards to conduct regular ethical reviews of planned and ongoing data science work**.

A particular issue of concern associated with predictive techniques applied to vulnerable populations concerns systematic biases in the data. In the context of the criminal justice system, a common attribute used in models is a suspect's arrest history, and rearrest is a common choice of outcome variable to measure recidivism. Due to implicit, explicit, and structural biases, a suspect of color may be more likely to have a longer arrest record than a similar white suspect. Moreover, a suspect of color may be more likely to be rearrested than a similar white suspect[25]. Ongoing debates by journalists, academics, and the general public have focused on issues of fairness involved in using algorithms to make decisions, when those algorithms could potentially institutionalize biases reflected in historical data[26]. Again, there is not always an easy answer to whether or not a predictive algorithm should be used in a given context, and there is no statistical

"silver bullet" to design a fair algorithm[27;28]. While this is an active area of research, I recommend that for the time being, in addition to constructing an advisory board, as outlined above, that agencies deal with this issue by **estimating—to the best extent possible—the effects of a planned algorithm on vulnerable populations, paying particular attention to protected classes such as race and gender. After deployment, changes resulting from the use of predictive methods should be measured to detect any unintended consequences.**

The use of predictive analytics by city agencies offers the promise of optimized service delivery, and appropriate care and foresight can minimize potential harms. An appreciation of the numerous and varied decisions faced by a data scientist applying machine learning models to municipal agency data is important, as the use of these techniques in local government is becoming more and more common. My experiences with ACS have informed the recommendations listed above, and I believe that other organizations in the public sector should adopt some of ACS's more forward-looking practices. Although fully implementing all of the recommendations may be challenging, even addressing a few of the suggestions would be a valuable investment for an agency to make to guide the ethical and optimal use of computational techniques to improve city operations and the quality of life for residents.

Acknowledgements

The author would like to thank Martin Jankowiak, Alex Chohlas-Wood, Thomas Laetsch, Hari Shroff, Arya Tafvizi, Jenny Xie, and the reviewers for their valuable suggestions.

Author Disclosure Statement

The author is partially supported by New York City's Administration for Children's Services.

References

[1] New York City Police Department. Crime statistics. `http://www.nyc.gov/html/nypd/html/crime_prevention/crime_statistics.shtml`, September 2016.

[2] John Lee, Donna Hope, Stacy Lee, et al. New York City Local Law 84 benchmarking report. `http://www.nyc.gov/html/planyc/downloads/pdf/publications/2014_nyc_ll84_benchmarking_report.pdf`, September 2014.

[3] Daniel E. Marasco and Constantine E. Kontokosta. Applications of machine learning methods to identifying and predicting building retrofit opportunities. *Energy and Buildings*, 128:431 – 441, 2016.

[4] Brian Heaton. New York City fights fire with data. `http://www.govtech.com/public-safety/New-York-City-Fights-Fire-with-Data.html`, May 2015.

[5] Neal Ungerleider. NYPD, Microsoft launch all-seeing "Domain Awareness System" with real-time CCTV, license plate monitoring. `https://www.fastcompany.com/3000272/nypd-microsoft-launch-all-seeing-domain-awareness-system-real-time-cctv-license-pl` August 2012.

[6] Anthony A. Braga and David L. Weisburd. *Policing Problem Places*. Oxford University Press, 2010.

[7] New York City Department of Parks and Recreation. 2015 street tree census - tree data. `https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh`, June 2016.

[8] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, Dec 2013.

[9] New York City Department of Transportation. New York City Department of Transportation data feeds. `http://www.nyc.gov/html/dot/html/about/datafeeds.shtml`, September 2016.

[10] Alekh Agarwal, Oliveier Chapelle, Miroslav Dudík, and John Langford. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15:1111–1133, 2014.

[11] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[12] U.S. Department of Health and Children's Bureau Human Services. Child welfare state statutes. `https://www.childwelfare.gov/pubPDFs/registry.pdf`, July 2014.

[13] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104:671–732, 2016.

[14] Jeremy Gorner. Chicago police use 'heat list' as strategy to prevent violence. `http://articles.chicagotribune.com/2013-08-21/news/ct-met-heat-list-20130821_1_chicago-police-commander-andrew-papachristos-heat-list`, August 2013.

[15] Emily Keddell. Substantiation, decision-making and risk prediction in child protection systems. *Policy Quarterly*, 2016.

[16] Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.

[17] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.*, 9(3):1350–1371, 09 2015.

[18] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2016.

[19] Jesse Russell. Predictive analytics and child protection: Constraints and opportunities. *Child Abuse & Neglect*, 46:182–189, 2015.

[20] Kiri L Wagstaff. Machine learning that matters. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[21] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2 edition, 2009.

[22] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. Simple rules for complex decisions. `https://ssrn.com/abstract=2919024`, 2017.

[23] Office of Management and Budget. The City of New York executive budget, fiscal year 2016. budget summary. `http://www.nyc.gov/html/omb/downloads/pdf/sum5_15.pdf?epi-content=GENERIC`, June 2015.

[24] Jessica Saunders, Priscillia Hunt, and John S. Hollywood. Predictions put into practice: a quasi-experimental evaluation of chicago's predictive policing pilot. *Journal of Experimental Criminology*, 12(3):347–371, 2016.

[25] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[26] Julia Angwin, Jeff Larsen, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Propublica*, 2016.

[27] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2017.

[28] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *arXiv preprint arXiv:1701.08230*, 2017.