Playing with the Data

Paul Ohm[1] and David Lehr[2]

Abstract

Machine Learning is a label that describes a broad, varied, and ever-evolving set of algorithms, processes, and tools used for sophisticated new forms of data analysis. Legal scholars have begun to focus intently on machine learning, which promises to become a key tool of prediction and decisionmaking by industrial actors and governments alike. We think this burgeoning scholarship has tended to treat machine learning too much as a monolith and an abstraction, largely ignoring some of the most consequential stages of machine learning analysis, particularly initial ones. As a result, many potential harms and benefits of automated decisionmaking have not yet been articulated, and policy solutions for addressing those impacts remain underdeveloped.

To fill these gaps in legal scholarship, we begin by providing a rich breakdown of the process of machine learning. We divide this process roughly into nine steps: problem definition, data collection, data cleaning, summary statistics review, data partitioning, model selection, model tuning, validation, and deployment. Far from a straight linear path, most machine learning activity dances back and forth across these steps, whirling through successive passes of model building and refinement.

Simplifying this mapping, we contend that legal scholars should think of machine learning as consisting of two distinct workflows: "playing with the data," which comprises the first eight steps of our breakdown, and "the running model," which describes a machine-learning algorithm deployed and making decisions in the real world. Our core claim is that almost all of the significant legal scholarship to date has focused on the implications of the running model—the predictive policing algorithm directing the deployment of officers, the face recognition system identifying suspects, or the autonomous automobile navigating a left-hand turn—and has neglected the

[1] Professor of Law, Georgetown University Law Center

[2] Research Fellow, Georgetown University Law Center

possibilities of playing with the data. A few notable and important articles pay some attention to playing with the data, but we think even these sophisticated analyses fall short; they largely overlook the key technical details and, more importantly, the policy implications of algorithm development.

This is a fundamental shortcoming of earlier work because the two phases of machine learning give rise to very different issues. The potential harms and benefits (say to fairness or discrimination) that can creep in while playing with the data differ from those of the running model. For example, many have documented the "garbage in-garbage out" problem that can make machine learning models discriminatory, but from the vantage point of the running model, this "garbage" is a static, unavoidable feature of the data. Only one who is attentive to the many ways in which data can be selected and shaped—say during data cleaning or model tuning—will characterize fully the source of the stink. Similarly, a benefit of choosing certain machine learning algorithms is the ability to place weight on particular types of errors over others—for example, to favor false negatives over false positives in criminal justice contexts—but this choice is one that must be made when playing with the data.

Another reason legal scholars in particular need to focus on playing with the data is that combatting harms at the running-model stage is often too little too late. Because playing with the data occurs earlier in time and entails much more human involvement than the running model, this phase provides more opportunities and more behavioral levers for policy prescriptions. As many have documented, a running model is often an inscrutable black box; in contrast, we have opportunities for surveillance (audit trails, keystroke loggers and video cameras can watch data scientists) and mandated interpretability during playing with the data. We can ban certain approaches—deep learning or neural nets, if our concern is opacity—during playing with the data, but with a running model, all we can do is rue the choice that has already been made. These possibilities may be neither necessary nor sufficient to address the many potential harms of machine learning, but they are likely to be missed by those with a single-minded focus on the running model.

Greater attention to playing with the data can also advance several contemporary debates about machine learning. It is common to hear regulation skeptics describe machine learning as "more art than science," but we think this self-serving excuse inappropriately assumes that black-box algorithms have black-box workflows; as we show, the steps of playing with the data are actually quite articulable. We also think greater attention to playing with the data will suggest new arguments for favoring privacy laws that regulate data collection over data use, injecting a new viewpoint into a long-running debate. Finally, many commentators have argued that we must preserve a "human in the loop" of machine learning, but most of them are referring to the running model as the relevant loop. We think there are different—perhaps more imperative—reasons to maintain humans in the underappreciated playing-with-the-data loop as well.

We are not saying that scholars ought to neglect the running model. The best assessments of the promises, perils, and prescriptions for big data will consider both phases of machine learning. But widening the view to encompass earlier stages will be crucial for solving some seemingly intractable problems of our increasingly automated world.

Situating *Playing with the Data* in the Literature

We situate our article in the burgeoning and ever-expanding writing on the impacts to society of big data, artificial intelligence, and machine learning. For present purposes, we focus on how our work will engage legal scholarship, but these topics are inherently interdisciplinary, and we intend to cite specifically and at length to work from outside the law reviews in this article. We are happy to provide the program committee with additional references if it requires it.

Within this ever-expanding universe of work, we intend to engage with: works at the intersection of big data and fairness, such as by Danielle Citron,[3] Frank Pasquale,[4] Deirdre Mulligan and Cynthia Dwork,[5] and Neil Richards;[6] works at the intersection of big data and discrimination, such as by Solon Barocas and Andrew Selbst[7] and

[3] *Technological Due Process; The Scored Society*

[4] *The Scored Society*

[5] *It's Not Privacy and It's Not Fair*

[6] *Big Data Ethics; Three Paradoxes of Big Data*

[7] *Big Data's Disparate Impact*

Pauline Kim;[8] an emerging literature interested in exporting insights from computer science having to do with the correctness of algorithms such as an article written by a gang at Princeton[9] and work by Deven Desai;[10] work on robotics and automation, such as by Ryan Calo,[11] and Woody Hartzog;[12] work rooted more broadly in privacy but touching on these issues, such as by Julie Cohen[13] and Dan Solove;[14] and everything Kate Crawford has written.[15]

Perhaps our most direct and extended engagement will be with scholars investigating the intersection of machine learning and the Fourth Amendment. Much of this work has been previewed at this conference. Representative articles have been published by Steven Bellovin and Renee Hutchins (with computer scientist co-authors)[16] Andrew Ferguson,[17] Elizabeth Joh,[18] Michael Rich,[19] and Andrea Roth.[20] Our thesis will engage at length with this work. To take only one example, Rich notes that errors in algorithms might be suppressed in court only if they were the result of "deliberate, reckless, or grossly negligent misconduct, or of systemic negligence," which could include "the provision of bad data … or mistakes in programming." But what exactly constitutes "bad data" and how algorithms could be mistakenly programmed escape detailed treatment. This is playing with the data, and our research will fill this gap.

[8] *Data-Driven Discrimination at Work*
[9] *Accountable Algorithms*
[10] *Algorithms and the Law*
[11] *Robotics and the Lessons of Cyberlaw*
[12] *Unfair and Deceptive Robots*
[13] *What Privacy is For*
[14] *Privacy Self-Management and the Consent Dilemma*
[15] *E.g. The Anxieties of Big Data*
[16] *When Enough is Enough: Location Tracking, Mosaic Theory, and Machine Learning*
[17] *Policing Predictive Policing; Big Data and Predictive Reasonable Suspicion.*
[18] *The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing; Policing by Numbers: Big Data and the Fourth Amendment.*
[19] *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment.*
[20] *Machine Testimony.*