ANNUAL REVIEWS

Annual Review of Law and Social Science

Tool for Surveillance or Spotlight on Inequality? Big Data and the Law

Rebecca A. Johnson[1] and Tanina Rostain[2]

[1]Quantitative Social Science, Dartmouth College, Hanover, New Hampshire 03755, USA; email: Rebecca.Ann.Johnson@dartmouth.edu

[2]Georgetown University Law Center, Georgetown University, Washington, DC 20001, USA; email: tr238@law.georgetown.edu

Annu. Rev. Law Soc. Sci. 2020. 16:453–72

First published as a Review in Advance on July 28, 2020

The *Annual Review of Law and Social Science* is online at lawsocsci.annualreviews.org

https://doi.org/10.1146/annurev-lawsocsci-061020-050543

Keywords

inequality, surveillance, law, organizations, housing, child welfare

Abstract

The rise of big data and machine learning is a polarizing force among those studying inequality and the law. Big data and tools like predictive modeling may amplify inequalities in the law, subjecting vulnerable individuals to enhanced surveillance. But these data and tools may also serve an opposite function, shining a spotlight on inequality and subjecting powerful institutions to enhanced oversight. We begin with a typology of the role of big data in inequality and the law. The typology asks questions—Which type of individual or institutional actor holds the data? What problem is the actor trying to use the data to solve?—that help situate the use of big data within existing scholarship on law and inequality. We then highlight the dual uses of big data and computational methods—data for surveillance and data as a spotlight—in three areas of law: rental housing, child welfare, and opioid prescribing. Our review highlights asymmetries where the lack of data infrastructure to measure basic facts about inequality within the law has impeded the spotlight function.

1. INTRODUCTION

Big data is a polarizing force among those studying inequality and the law. One strand of research in law and society focuses on the perils of big data: predictive policing that threatens the rights of minorities living in heavily policed communities (e.g., Brayne 2017, 2018; Ferguson 2016, 2019; Selbst 2017), hiring algorithms that threaten the rights of job seekers subject to potentially biased preemployment assessments (Barocas & Selbst 2016, Raghavan et al. 2019), and tenant screening algorithms that threaten the rights of tenants under the Fair Housing Act (Allen 2019, Aronowitz & Golding 2019). This perspective emphasizes multiple ways that big data, and the tools used to extract insight from such data, exacerbates existing inequalities under the law. One problem is bias.[1] First, predictive models get things wrong, and there is growing evidence that models get things wrong more often for vulnerable subgroups. Second are concerns that the opacity of algorithms undermines individuals' ability to remedy these wrongs through various appeals processes (Brauneis & Goodman 2018, Eubanks 2018a, O'Neil 2017, Pasquale 2015). Last are concerns that data's uses subject individuals to surveillance that undermines certain privacy rights (e.g., Ferguson 2019, Vagle 2016). The result—errors made in high-stakes decisions, opacity that undermines individuals' ability to correct those errors, growing surveillance of various domains of life—has sparked growing pessimism about technology as a force for good, as well as growing calls for more oversight and regulation of organizations' use of that technology (e.g., Crawford & Schultz 2014, Whittaker et al. 2018, Zalnieriute et al. 2020).

Amid these legitimate concerns about big data and inequality under the law, a parallel strand of research uses the same computational tools to shine a spotlight on, rather than exacerbate, inequality. Three sets of actors—researchers, advocates, and regulators—have begun to use large-scale data to surveil more powerful institutions for mistreatment of vulnerable individuals, including police (Legal Aid Society 2020, Ouellet et al. 2019), landlords (Desmond et al. 2018, Ye et al. 2019), and employers (Kalev & Dobbin 2006). Yet whereas the perils of big data are well theorized, the prospects of big data as a spotlight for understanding law and inequality have received less theoretical attention.

Our review aims to more closely align these two perspectives—one in which big data is a tool that amplifies existing inequalities in the legal system and another in which big data can be a tool to produce knowledge about inequalities. We begin with a typology of the role of big data in inequality and the law. The typology, focusing on questions of who uses the data and for what, shows that different examples arguably drive polarized views about big data's role in inequality under the law. One use case is powerful institutions' use of big data to engage in surveillance of populations who have faced historical maltreatment, where there is a wealth of scholarship investigating the implications for inequality.

If we flip each element of the example—Who is using the data (a powerful social institution)? What are they using it for (engaging in surveillance)? What is the result (extra monitoring of certain individuals)?—big data has a different relationship to law and inequality. Using examples from rental housing, child welfare, and opioid prescribing, we show how big data and computational methods can shed a spotlight on inequalities that stem from organizations' mistreatment of individuals.

---

[1] Although there are various ways of operationalizing algorithmic bias, most definitions focus on unequal error rates across subgroups. So, for instance, if an algorithm is used to screen for whether a prospective tenant has a criminal record, bias could come in the form of certain prospective tenants having a higher likelihood of false matches—being inappropriately flagged as having a previous conviction when they do not (Dunn & Grabchuk 2010). These include errors in areas of criminal law like criminal justice risk assessments (for reviews, see Berk et al. 2018, Chouldechova 2017) and areas of civil law like benefits eligibility determinations (Eubanks 2018a).

Then, we discuss how new computational methods can complement existing quantitative and qualitative tools in three types of research for this spotlight function: descriptive research to diagnose patterns of inequality, research aimed at causally testing ways to improve these patterns, and predictive models to detect emerging forms of inequality. Yet for big data to play this role, researchers must address large gaps in the data we have on individuals' interactions with the legal system.

In the final section, we show how the wealth of data on interactions with the criminal justice system is accompanied by a paucity of data on interactions with other important legal institutions: civil legal aid providers; administrative hearing systems; and courts focused on topics like custody disputes, debt, and housing. We discuss efforts toward building a data commons that provides researchers interested in studying the three uses—describing patterns of inequality in organizational compliance with law, causally testing interventions aimed at ameliorating those inequalities, and using prediction to catch early-emergent forms of inequality—with the data needed to research them.

2. A TYPOLOGY OF BIG DATA AND THE LAW

2.1. Defining Terms: Big Data and Computational Social Science

The field of computational social science focuses on using large-scale data to shed new insights into social processes (Lazer & Radford 2017, Lazer et al. 2009, Salganik 2019). As Lazer et al. (2009, p. 721) describe, the various interactions individuals have with each other and with social institutions leave "digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies." These trace data are described as having three unique characteristics, known as the three Vs: high variety, high volume, and high velocity (Mooney et al. 2015).

High variety refers to a larger data set's capacity to combine and analyze data sets obtained for different purposes. For instance, a study of racial disparities in how long people need to wait to vote might merge smartphone data that reveal location with census data on neighborhood demographics (Chen et al. 2019). High volume refers to data with orders of magnitude more observations and/or variables per observation than prior data sets in the field. For instance, whereas past studies of voting wait times relied on surveys (Pettigrew 2017) or poll observations, the smartphone location data Chen et al. (2019) discussed offer many more observations.[2] Finally, high velocity refers to a data-generating process in which data are compiled in closer to real time. Whereas survey data on wait times might rely on a survey fielded once every *n* election cycles, and have a several-month lag between when voters are asked about their wait times and when the data are ready to analyze,[3] smartphone data could be made available to monitor wait times in real time.

[2]In this particular case, the smartphone data are "long and narrow," containing many more individuals but considerably less information about each individual than survey data. For instance, the article cannot say that voters of different races have to wait different amounts of time, because they do not know the races/ethnicities of the smartphone owners. Survey data would have both this individual-level information and attitudinal information, such as feelings about various political candidates. In other cases, data are "long and wide," containing both many individuals and large amounts of information about each individual through merging records from different sources.

[3]This lag is due less to the inherent nature of the data itself and more to the quality-control process that large-scale survey data go through after fielding and prior to release of data products. This process leads to important corrections for bias—for instance, the survey bureau will generate weights that adjust for certain forms of nonresponse bias (Brick & Tourangeau 2017)—that the companies releasing digital trace data invest

2.2. Two Uses of Big Data in the Criminal Justice System

The three characteristics of big data—the variety of different data that can be combined, the volume of entities whose data are recorded, and the velocity that can lead to more real-time monitoring—make it attractive for various use cases. Yet theory about big data's role in law and inequality tends to focus on one, highly specific case: the criminal justice system's use of big data for policing, bail, and sentencing decisions. Although the next section extends this focus to the civil justice system, here, we use two examples from the more heavily studied criminal justice system to build our framework.

Thinking about the prototype of law enforcement using big data for surveillance, we see first that the organizations using the data to inform the decisions have structural advantages over the individuals impacted by those decisions. Second, the organizations using data have access to substantially more data and more sophisticated tools for processing the data than the individuals affected. For instance, as Richardson et al. (2019) outline, the Chicago Police Department developed a Strategic Subject List that sought to rank Chicago residents based on their risk of becoming either a crime victim or a crime perpetrator. Whereas Richardson et al. (2019) focus on biases in the data behind the list,[4] another issue is asymmetries in the technological sophistication of the police departments that use the data to surveil versus that of the individuals being surveilled. Officers and prosecutors had access to the proprietary data used to create the list: individual-level data, including arrest records and measures of gang affiliation. Yet public defenders representing clients had access to neither the list itself nor the source data behind the list to potentially subject police decisions motivated by a client's placement to more scrutiny. The third feature of this prototypical example is that not only do the two parties—those doing the monitoring and those being monitored—have asymmetric access to data and tools for processing that data, but there are also asymmetric consequences for parties' misuse of the data.[5]

2.2.1. Vulnerable individuals, or advocates or regulators who act on their behalf, using big data to make decisions to constrain powerful institutions. Yet law enforcement's use of big data in ways that perpetuate inequality is not intrinsic to big data and computational methods. If we flip each of the three elements of the predictive policing case, we end up with a very different portrait of how the rise of big data relates to law and inequality. First is that, rather than powerful organizations using big data to make decisions that affect individuals, three sets of actors—researchers, advocates, and regulators—can use big data to monitor powerful actors or organizations. Returning to the policing example, one concern with a police department's use of historical arrest data to generate lists of high-risk residents is biases in these data that stem from inappropriate police behavior like unconstitutional stops, fabrication of data to support certain reporting goals, or the inappropriate use of force in interactions. Yet if we flip the target of the decision, large-scale data can be used to monitor police officers themselves.

---

in much less heavily, if at all. For instance, although geocoded Twitter data are sometimes used to study spatial patterns in social processes, Lippincott & Carrell (2018) show that there are systematic differences between Twitter users who opt in to location sharing and those who opt out. Although standard surveys attempt to adjust for these biases using weighting, the velocity of digital trace data means that it is fed to end users without these adjustments.

[4]Notably, they focus on the fact that historical arrest data reflect potentially biased policing.

[5]In particular, false positives by the police department—placing an individual on the list who does not actually have a high risk of offending in the future—can have extremely negative consequences for the individual being monitored; in contrast, there are few consequences for the monitoring organization of getting that risk classification wrong.

The actors who use the data—researchers, advocates, and regulators—flip each of the three elements of big data's use for decision making. First, the institutions about which data are collected are those at risk of violating the rights of vulnerable individuals, rather than the vulnerable individuals themselves. For instance, the New York Legal Aid Society created a database through the Cop Accountability Project (Legal Aid Society 2020) that reveals potential rights violations by officers. These actors use the data to try to correct structural imbalances between parties—in the first case, structural imbalances between criminal defendants whose arrests might have been preceded by police misconduct, and in the second case, structural imbalances between police officers and residents they patrol.

Second, while the organizations being subject to surveillance continue to have access to better data and more sophisticated tools than those using the data to surveil them, those using big data to monitor powerful organizations have pursued two paths. First, efforts like the Legal Aid Society (2020) have used open records laws to legally compel organizations like police departments to release data that shine a spotlight on inequalities the same organization may perpetuate. Second, efforts like Data Science for Social Good partnerships with police departments—in one case, to use predictive modeling to flag officers at high risk of adverse interactions with members of the public (Carton et al. 2016, Helsby et al. 2018)—use data science capacity for monitoring aimed at reducing inequality. This means that although the powerful organizations being monitored continue to have more data and more sophisticated tools than the less powerful actors doing the monitoring, legal initiatives like open records laws and reforms internal to organizations may help to narrow these gaps.

Finally, although those using data for monitoring may still make erroneous decisions, the advantages of the powerful actor or organization arguably mitigate some of the concerns about these harms. Returning to the example of a predictive model that flags officers at high risk of adverse interactions, the organizational structures of police departments and unions, which emphasize supportive interventions like counseling for errant officers rather than punitive interventions like dismissal, mitigate the consequences of errors on those being monitored.[6]

This example highlights the need for nuance when discussing how big data contributes to law and inequality. When used by powerful organizations to make decisions impacting vulnerable individuals, it can exacerbate inequality. When big data's use is flipped, and when researchers are able to work with advocates and regulators to use it to monitor powerful actors and organizations, the same three Vs of big data—variety, volume, and velocity—have the potential to reduce inequality. Whereas our initial example was parallel cases within criminal justice—police monitoring the public and researchers and advocates monitoring police—in Section 4, we highlight similar parallels across three other legal settings: rental housing, child welfare, and opioid prescribing.

2.3. Connecting Big Data to Long-Standing Preoccupations in Sociolegal Studies with the Law's Role in Inequality

This typology shows how big data, in addition to its well-highlighted role in exacerbating inequality under the law, can also shine a spotlight on inequalities in how the law on the books becomes law in action. From employers' compliance with Civil Rights Act legislation protecting employees from discrimination (Dobbin 2009, Edelman 1992, Munger & Seron 2017) to landlords'

[6]Similarly, with our empirical examples, the targets of the oversight—landlords and property management companies, child welfare agencies, debt collection agencies, and physicians—tend to have more resources to defend themselves against erroneous judgments than more vulnerable individuals.

compliance with Fair Housing Act legislation protecting tenants (Desmond 2016, Desmond & Bell 2015), inequality persists as organizations exploit ambiguities in formal legislation that leave individuals poorly protected (Fuller et al. 2000).

First, at the interpretation stage, big data and computational methods could potentially reveal patterns of attrition in the process behind "naming, blaming, and claiming," that is, whether an individual construes a rights violation as something worthy of a remedy, blames the correct institution, and claims a remedy through either informal or formal channels (Felstiner et al. 1980). Only a subset of those with legitimate grievances against powerful organizations reach court, and some of that attrition might be due to a failure to interpret something troubling in their day-to-day life—a denied request for a raise or a landlord not responding to an inquiry—as a potentially actionable grievance. To the extent that various forms of trace data reveal that these events happened to individuals, this trace data could reveal naming gaps in which patterns of inequality go unrecognized by those impacted.

Second, at the enforcement stage, the spotlight function of big data could shed light on two types of enforcement problems. As Fuller et al. (2000, p. 207) highlight, although laws "authorize administrative agencies to monitor compliance, enforcement depends primarily on the extent to which individuals who perceive rights violations take action to make use of, or mobilize, their legal rights." The first spotlight function is to understand why so few individuals harmed by powerful institutions enforce their legal rights (Sandefur 2008). The second function is to empower administrative agencies to use data for enforcement, enforcement that might rely on data shining a spotlight on legal violations, even if the individuals impacted by those violations are unaware of them. We now move to examples of the dual role of big data and computational methods: as a tool for surveillance by powerful institutions and as a spotlight on powerful institutions whose on-the-ground treatment of individuals deviates from what the law requires.

3. CASE STUDIES: DATA FOR SURVEILLANCE AND DATA AS A SPOTLIGHT

Here we focus on case studies of big data's role in the areas of rental housing, child welfare, and medical prescribing.

3.1. Housing: Data to Monitor Tenants Versus Data to Monitor Landlords

Enacted in 1968, the Fair Housing Act aimed to reduce discrimination in housing markets to decrease residential segregation. Despite its ambitions, the legislation contained weak enforcement mechanisms (Massey 2015). The primary administrative agency, the Department of Housing and Urban Development, can pursue enforcement only following a complaint of discrimination and has limited tools at its disposal to sanction landlords and sellers who are out of compliance. Similarly, the "aggrieved individuals" meant to shoulder the burden of enforcement face tight timelines and limited remedies (Massey 2015, Metcalf 1988), with recent rulemaking significantly increasing these burdens (Rose 2019). Although formal laws protect tenants from discrimination on the basis of attributes like race or disability status, landlords who violate the law face few consequences. What role can big data and computational methods play in either exacerbating or mitigating power disparities between landlords and tenants?

3.1.1. Big data to screen tenants. First is the case of powerful institutions using big data for surveillance of vulnerable individuals. Landlords rely on a variety of data sources and tools to screen prospective tenants. The potential for errors in this screening, coupled with the opacity of

the screening process, makes it difficult for tenants and advocates to monitor whether the screening violates antidiscrimination statutes.

Four components go into a tenant screening report: the tenant's residential history; his or her credit report; a criminal background check; and a civil litigation history, with a focus on eviction judgments, but including information on consumer or medical debt (Dunn & Grabchuk 2010). The information is then packaged into a tenant screening score, a numerical summary of tenant risk that landlords can use to make decisions. For instance, TransUnion releases a ResidentScore and emphasizes that, "with ResidentScore, you aren't stuck basing your decisions on the same algorithms that a bank would use. Instead, you're using a score that specifically analyzes predictors of a bad rental outcome" (Collatz 2019). These scores may perpetuate on-the-ground inequalities within formal law through two pathways.

The first is through the variety of data that feed the score. Most detrimental are false-positive matches between tenant identifiers—name and date of birth—and identifiers in criminal records or eviction databases. This might mean that a stigmatizing mark is attributed to a prospective tenant based on a false match. Although various reports document these false positives using a small number of examples (Desmond & Bell 2015, Dunn & Grabchuk 2010, Kleysteuber 2006), the opacity of tenant screening algorithms makes it difficult to systematically study bias. As a result, academics are unable to investigate whether the screening practices align with computational advances in record linkage methods that might reduce false positives, like avoiding too much name standardization (Foster et al. 2016, Harron et al. 2017, Randall et al. 2013).

Even if the score is based on sound linkage practices, the volume and velocity of the data can put tenants at a disadvantage. First, for volume, tenant screening that might have relied on data from a single local court now relies on a high volume of data from nationwide databases. Yet lack of transparency about the inputs to the score makes it difficult for tenants to know why they are being rejected by landlords and to clarify misleading information, for instance, that they prevailed in a housing court claim against them, which therefore should not be factored into their score (Lebovits & Addonizio 2015).

Whereas the first issue comes from a high volume of data, the second issue comes from insufficient velocity. Companies tend to use more automated feeds of data to produce credit scores, but tenant screening scores more often involve one-time requests to furnishers of data, such as court records, rather than an automated feed (Dunn & Grabchuk 2010). Therefore, after a report is produced in response to a specific request, the company may delete the report, making it difficult for a tenant challenging a score to receive the underlying data. This perpetuates inequality, for instance, when tenants are unaware that undergoing a laborious criminal records expungement process (Prescott & Starr 2020) might improve their score. It also impacts the tenant's privacy rights, as he or she may be unaware that, by submitting a rental application to a particular property, they are allowing the landlord access to information about their debt or legal issues.

In sum, in the first stylized use case of big data in rental housing, the three Vs of big data widen gaps between the law on the books and law on the ground. Tenants have rights when searching for rental housing. Landlords' and property managers' use of screening tools may violate these rights. The variety of data means that the scores rely on linkages based on features like names that can produce false positives; volume makes it difficult for tenants to figure out where a black mark is coming from; and velocity makes it difficult for tenants to retrieve the underlying inputs to the score. Yet these on-the-ground violations are difficult to detect and challenge.

3.1.2. Big data to monitor landlords. Flipping the elements of the first use case reveals a different role for data in inequality under the law: Researchers use big data and computational methods to monitor the behavior of those same landlords and property managers, as well as to

potentially inform the enforcement activities of legal aid providers who represent tenants and regulators like city-level housing code enforcement agencies.

Princeton's Eviction Lab National Database uses the same underlying eviction data as the tenant screening reports but flips the target of surveillance from tenants to landlords and housing courts (Desmond et al. 2018).[7] Whereas the tenant screening reports use eviction filings to add a black mark to prospective tenants, early applications of the Eviction Lab Database highlight a flipped use: Researchers study how powerful institutions engage in aggressive filing behavior against tenants, as well as asymmetries in tenants' likelihood of showing up in court and fighting back.[8] Heightened scrutiny of the extent to which different types of landlords—individuals, small versus large property managers, or federally funded public housing authorities—act in line with laws on the books gives tenant advocates evidence to focus on the most egregious legal violations.

Whereas this first example is one in which the spotlight function is one step removed from action—researchers illuminate patterns of inequality in landlord behavior, and legal aid organizations and government regulators may or may not respond—other examples we discuss in greater detail in Section 4 use the spotlight in more immediately actionable ways. Ye et al. (2019) work with a New York City government agency trying to decide which at-risk tenants to extend outreach to first, and the McCourt School of Public Policy (2020) works with the Washington, DC, Housing Inspections Agency (more formally, the Department of Consumer and Regulatory Affairs) to prioritize inspections of housing owned by the most egregious violators. Thus, in the same way that landlords use predictive models to inform immediate decisions—Should I rent to this applicant? Whom should I be lenient with if behind on rent?—models can inform immediate decisions by government agencies charged with oversight.

3.2. Child Welfare: Data to Monitor Risky Families Versus Data to Monitor Child Welfare Agencies

Just as housing law features two stylized use cases—big data for surveillance of tenants and big data as a spotlight on inequalities that tenants face despite formal legal protection—the growing use of large-scale data and predictive modeling in child welfare has begun to follow a similar pattern. However, in contrast to the rental housing case, in which research and policy to monitor landlords have blossomed, the bulk of work in child welfare still focuses on data to monitor families rather than data to monitor the more powerful child welfare agencies.

3.2.1. Big data to monitor families. First is the stylized use case of powerful institutions using big data for surveillance of vulnerable individuals. Landlords have financial incentives to use large-scale data to screen tenants, as well as power imbalances that stem from a scarce stock of affordable housing.[9] Meanwhile, child welfare agencies have legal mandates that focus on protecting the rights of the child and power imbalances that stem from the high stakes of the

[7] The project collects all eviction filings from 2000 to 2016. Whereas some of the database is based on records requests to courts, a major source of data comes from private aggregators that collect and sell filings. In particular, the database uses filings from LexisNexis and American Information Research Services.

[8] This spotlight reveals new forms of inequality under the law. For instance, although many city-level efforts to protect tenants from eviction focus on private market rentals, a conference presentation shows the outsized role of public housing authorities in evicting tenants (Hepburn et al. 2019).

[9] Of course, landlords balance their desire to screen out inappropriate tenants with financial pressures to make sure that units are occupied. This can change the power imbalances between landlords and tenants and might also lead landlords to adopt more lenient screening policies out of a need to make sure units are occupied.

decision. To pursue these mandates, some agencies have shifted from standard screening methods to use of large-scale data to decide which reports of suspected abuse or neglect warrant extra staff attention.

Eubanks (2018a) documents how intake workers in Allegheny County, Pennsylvania, use a screening tool to help guide which referrals a caseworker should investigate in greater depth.[10] Yet limits stemming from inadequacies in the three Vs—lack of variety, too small a volume, and too slow a velocity—mean that the tool, and similar tools like it tested in jurisdictions like Illinois and Los Angeles (Brown et al. 2019), can lead to inequality.

First is inadequate variety. A key feature of Allegheny County's model is that it draws together administrative data on county residents from a variety of sources, including Head Start, the local housing authority, Pittsburgh public schools, and Medicaid claims (Eubanks 2018a, Vaithianathan et al. 2019). Drawing on data from different sources supposedly leads to higher accuracy than models trained solely on information from within the child welfare agency. Yet limitations on this variety persist (Eubanks 2018a). In a summary, Eubanks (2018b) points out how the reliance on Medicaid claims data results in heightened surveillance of low-income households, whereas higher-income households that "use private insurance or pay out of pocket for mental health or addiction treatment" face less scrutiny. Thus, one culprit behind big data's role in exacerbating inequality within the law is lack of variety. The predictive model involves merging data from a range of sources. But these sources, although much more varied than internal child welfare records, have important gaps for residents who seek help in settings that leave fewer traces in administrative data.

Although the child welfare case aligns with big data's first use case—a powerful state agency uses data to surveil county residents—we see that inequality within the law stems not from the use of the data but in the specifics of how they were used. And researchers concerned with inequality were able to closely scrutinize these specifics because, in contrast to the tenant screening algorithms discussed in Section 3.1.1, researchers, rather than private firms, developed the algorithm and released extensive details about its methodology (Vaithianathan et al. 2019). This reveals enough detail to show that the culprit was not replacing discretion by mandated reporters and caseworkers, which other research argues results in persistent biases (e.g., Dettlaff et al. 2011, Hampton & Newberger 1985, Roberts 2003). Instead, the culprit was "big-enough" data—data that could be used to build a model with a reasonable degree of accuracy for the proxy outcomes it was trained on but not big enough to ensure that families from different socioeconomic backgrounds had similar odds of being subject to investigation.[11]

3.2.2. Big data to monitor organizational referrals to child welfare agencies. The high stakes of child welfare decisions come from two types of errors. First are false-positive investigations of families, which can cause stress and undue scrutiny. Second are false negatives, or families who should have had a referral followed by an in-depth investigation but who are either never referred or referred but not investigated. Other false negatives come from children who are prematurely reunited with their biological families in ways that can harm their long-term outcomes (Font et al. 2018).

[10]As the county noted in a letter in response to the coverage of their model, the score is one input into the decision to investigate, but high scores do not automatically trigger an investigation. Instead, the score helps flag high-risk cases, and supervisory caseworkers decide whether or not the case should actually be investigated.
[11]The authors show, however, that the model has similar error rates, defined using area under the curve, across Black and non-Black children (Vaithianathan et al. 2019).

These high stakes mean that another role for data is as a tool for overseeing these decisions. And the same team of researchers who built the predictive model to guide oversight of families also analyzed historical data from the child welfare agency. Those data showed large racial disparities in referrals to the child welfare agency, with Black children referred at a significantly higher rate than White children (Vaithianathan et al. 2019). This suggests that computational methods, rather than solely being used to predict risk that families pose to children, can also be used to oversee how organizations like schools and hospitals may exhibit biases in the referral process. For instance, Putnam-Hornstein et al. (2016) use vital records data on more than 400,000 births in California, merged with medical records data on prenatal exposure to alcohol and drugs and child welfare system data (Child Protective Services), to show that among infants with evidence of prenatal exposure to substances, Black and Hispanic infants had a higher risk of a Child Protective Services referral. Although it is unclear whether the California hospitals should have been referring White parents of children with prenatal substance exposures at a higher rate or referring non-White parents at a lower rate, the example shows that large-scale data can be used to oversee organizations' referral practices.[12]

Yet, as in the rental housing case, power gaps remain between organizations using data to monitor vulnerable individuals and researchers using data to enhance oversight of organizations. Whereas the child welfare agency can take immediate action based on the predictive model—they flag some families for more scrutiny—researchers using the same methods as a spotlight on inequalities in referral processes rely on others to use the findings to enhance oversight.

3.3. Prescription Drug Misuse: Data to Target Marketing to Prescribers and Data to Target Inappropriate Prescribing

The final case involves inequality within laws that grant physicians significant professional discretion to prescribe as they see fit. Van Zee (2009) describes how large-scale data formed the backbone of Purdue Pharma's marketing plan for OxyContin. Using publicly available documents on marketing strategies, shows how the company built "prescriber profiles" that sought to "identify the highest and lowest prescribers of particular drugs in a single zip code, county, state, or the entire country." The company then targeted gifts and in-person marketing visits to high prescribers.

Just as the same eviction filings data can be used by tenant screening firms to place a black mark on the record of a tenant or by researchers to study unequal practices, the same large-scale prescribing data from sources like Medicare claims can be used either by firms to target marketing or by researchers and regulators to target oversight. First is oversight of prescribers. The Center for Medicare and Medicaid Services (CMS) used data on Medicare Part D prescribing patterns to identify "outlier prescribers" of opioids and other Schedule II substances, or prescribers whose behavior relative to peers in the same specialty/state likely indicated inappropriate behavior (rather than just a high-need patient pool), testing an intervention that warned the physicians

[12] Indeed, Heimer & Staffen (1995), studying how professionals in a neonatal intensive care unit label parents as inappropriate caregivers, show that, contrary to predictions, disadvantaged parents do not necessarily face increased risks of a negative label. Because hospitals needed to invest resources in improving the behavior of caregivers labeled inappropriate, resource scarcity could mean some parents' behavior went unlabeled. In contrast, in Putnam-Hornstein et al.'s (2016) example, it is relatively costless for a hospital to make an outside referral. These cases suggest that part of the oversight function could focus on incentives that child welfare agencies have toward false negatives when investigating parents requires resources the agency lacks.

of their outlier status compared to their peers (Sacarny et al. 2016).[13] Second is oversight of the pharmaceutical companies themselves, more powerful given that some inappropriate prescribing stemmed from the misleading information pharmaceutical companies gave to physicians. Although civil litigation against opioid manufacturers has faced legal hurdles, Gluck et al. (2018) argue that, similar to the way that litigation against tobacco companies showed inequalities in their marketing practices, litigation against opioid manufacturers has led to the disclosure of new data used for oversight. Yet whereas the companies' use of data for targeting was proactive, these data are largely reactive—uncovering harms after they have already transpired. This shows that, despite data's dual use, the spotlight function often comes too late.

3.4. Similarities and Differences

The three cases show that the same data that can be used in ways that perpetuate inequality within the law can also be used to shed a spotlight on, and potentially mitigate, these same forms of inequality. Yet the three cases vary in which actors are able to use data for that spotlight function, and in the potency of that spotlight. In the child welfare case, the spotlight remains relatively removed from concrete oversight actions. In the rental housing case, some research remains removed. Other efforts are embedded directly in city-level government agencies that assist vulnerable tenants and that use the research to modify their operations, though in two liberal jurisdictions (New York City and Washington, DC), rather than nationwide. In the prescribing case, the research was undertaken jointly between researchers and CMS, a federal agency with both direct access to and powerful enforcement power over prescribing.

The cases show that although the spotlight function of big data can reduce some power disparities, two conditions greatly enhance the potency of the spotlight. First are organizations that have both direct access to nationwide data on the behavior of powerful actors and mandates that involve using that data for oversight. CMS, as the reimbursement source for Medicare prescriptions, has access to standardized nationwide data on prescribing. The Department of Housing and Urban Development has access to some data on both rental behavior and housing code issues within the voucher and public housing system but less of a focus on oversight. Meanwhile, local governments that have data on the child welfare system have a mandate more focused on oversight of families than on oversight of how placement decisions affect long-run outcomes.

Second are partnerships between researchers who can shine a spotlight on inequality and civil society and government organizations that can use that spotlight to take action. An algorithm to find landlords who are serial evictors or who perpetuate unsafe housing conditions becomes useful only if those landlords face sanctions; that translation requires partnerships of the sort we discuss in Section 5.

4. THREE TYPES OF SPOTLIGHTS: DESCRIBING INEQUALITY, TESTING INTERVENTIONS, AND PREDICTING FUTURE HARMS

The previous sections show that the focus on one stylized use case of big data—powerful actors like law enforcement using big data and computational methods to surveil vulnerable individuals—offers an incomplete picture of big data's role in law and inequality. A second stylized use case, illustrated in our examples from rental housing, child welfare, and prescribing, involves flipping

[13] Although that particular intervention had no impact, interventions with stronger warnings significantly reduced inappropriate prescribing (Sacarny et al. 2018).

the use of big data and predictive modeling to shine a spotlight on inequalities within the law. These efforts involve researchers, advocates, and regulators.

The case studies also highlight a mix of methods through which data become a spotlight on inequality. The researchers building the evictions database used computational tools to generate usable data from the text of raw case records (Desmond et al. 2018), for instance, using dictionary-based text mining to exclude cases filed against commercial defendants and Levenshtein edit distance to standardize tenant names. The use of these tools facilitated description on a much larger scale than past, single-city studies of eviction.

The example shows that although randomized controlled trials (RCTs) have long been the gold-standard quantitative methodology to study inequality and the law, computational social science may elevate the role of description in studying law and inequality. Here, we review how computational social science can augment two existing quantitative methods—surveys and RCTs—to shine a new spotlight on patterns of law and inequality.[14]

4.1. Describing Inequalities Within the Law

The case studies showed an important descriptive role for computational social science. This descriptive role often involves repurposing data developed for one use—housing court records, inspection agency complaints, child welfare administrative data, or prescription data—to yield new descriptive insights.

4.1.1. Surveys and big data. Surveys have helped illuminate inequalities within the law, complementing methods like ethnography and interviews to show how patterns revealed on a smaller scale are borne out on a population level. One important strand of survey research focuses on the legal problems people experience that may not make it to the formal justice system. Surveys can be used to ask about problems or situations people have experienced—problems with a landlord, unpaid debt, or an insurer rejecting a request for reimbursement—and about what help they sought, if any, to resolve those problems (Sandefur 2015). As Sandefur (2014) shows, these problems and their consequences place a disproportionate burden on people living in poverty and, particularly, on poor African American and Latino households.

How can big data augment survey-based diagnoses of inequality within the law? The advantages come from the three Vs. Variety can mean studying the same issue—e.g., a biased insurance denial that should have been appealed—from multiple vantage points. A survey allows researchers to quantify the prevalence of people who self-report an insurance denial as a deprivation of justice. But surveys may underestimate inequalities within the law for situations in which the main form of injustice is that people whose rights have been violated fail to construe the situation as such—in the present example, someone who views an insurer's denial of their claim as appropriate.

Imagine augmenting the same survey question with data from Medicaid or a private insurer on (*a*) reimbursement denials among beneficiaries, (*b*) which beneficiaries appealed the denial, (*c*) the text of the appeal, and (*d*) the insurer's final determination. Comparing these data with the survey responses could help researchers investigate questions like who had baseless denials but did not report them as unfair on a survey, or how the sophistication of an appeal that a beneficiary filed affected the likelihood that the reimbursement was granted. Answering these questions involves

[14]Although these methods can also complement qualitative methods like ethnography and interviews, those are beyond the scope of this review.

various computational methods; for instance, how should one operationalize "sophistication" in an appeal? Yet, as a recent review of computational legal analysis highlights (Frankenreiter & Livermore 2020), these methods have been applied to other sets of legal documents: Supreme Court decisions (Black & Spriggs 2013, Black et al. 2016b, Bommarito et al. 2009, Carlson et al. 2015), Supreme Court amicus briefs (Black et al. 2016a), and federal appellate court decisions (Chen & Ash 2020). Using these tools on documents that record more everyday forms of adversity can supplement these surveys. Big data is not a magic bullet but a way to triangulate our understanding of legal inequality through revealing new dimensions of that inequality.

4.2. Testing Interventions

The previous section shows that computational methods, working in tandem with existing quantitative methods in law and society research, can help elevate the role of descriptive research as a spotlight on inequalities within the law. Yet suppose researchers are interested in causally examining the impact of either help with justice issues (e.g., the role of counsel) or the justice issues themselves (e.g., having a debt or eviction filing). Just as large-scale data and computational methods can augment surveys and focused ethnography to improve our descriptive understanding of inequality within the law, these methods can augment RCTs aimed at causally testing interventions to improve outcomes.

Greiner & Matthews (2016) review the rise of RCTs to study inequalities within access to justice. For instance, Greiner & Pattanayak (2011), in their groundbreaking study of how representation affects claims for unemployment insurance, randomize people who call the Harvard Legal Aid Bureau asking for assistance and measure the effect of the assistance offers on how an administrative law judge ruled on the claim.

Albiston & Sandefur (2013) describe some limitations of this framework that large-scale data and computational methods can help address.[15] The main limitation that data with increased variety can help address is what Albiston & Sandefur (2013, p. 107) call a black box of mechanisms linking cause and effect:

> For example, does representation make a difference because lawyers present the case effectively in court, because lawyers provide detailed legal information that enables clients to obtain favorable settlements in their own negotiations with the other side, or because lawyers understand how to navigate informal relationships in the court system that help produce smoother case processing and resolution? These mechanisms remain an unspecified black box in most outcome studies.

Just as methods like computational text analysis of transcripts from legal hearings can shed light on descriptive inequalities, these methods, when combined with RCTs, can shed light on potential mechanisms linking cause to effect. Although not all elements of effective counsel leave a mark in this "trace data,"[16] we still might see evidence of these elements through taking audio recordings from two cases—one in which the defendant had counsel and the other in which they did not—and comparing features like the landlord's tone of voice when speaking to the tenant in

[15] The methods do not address all limitations the authors outline in their article. Most notably, aspects like the need to expand the focus beyond individuals living in poverty are more about what substantive questions RCTs should pose, rather than about what data and methods they have at their disposal to answer those questions. Yet Albiston & Sandefur (2013) outline other shortcomings that larger-scale administrative data could help address, including questions about who selects into an RCT measuring the impact of legal assistance and data on long-term outcomes beyond a judgment on a specific case.

[16] For instance, lawyers' informal relationships established outside the court room are unlikely to leave traces.

each case. New methods for automatically detecting features like tone and sentiment from audio data (Knox & Lucas 2019), in combination with a data infrastructure that supports the sharing of audio recordings from public hearings, could help with research that peers into the black box of mechanisms.

Beyond RCTs, volume and variety also allow researchers to conduct natural experiments to study the effect of legal events on life outcomes that it would be illegal to randomize, for instance, the effect of eviction. Whereas RCTs often study the impact of providing help with legal events, quasi-experimental methods often study the consequences of legal events in the absence of help. For instance, Humphries et al. (2019) leverage variation in the leniency of Cook County, Illinois, housing court judges to examine the effect of eviction on household economic outcomes. Exploiting this variation requires hundreds of thousands of court records (volume). A key finding was that the causal effect was small because, in the months preceding an eviction filing, households had rapidly declining credit scores and financial outcomes; whether or not the judge then ruled in their favor made only a small dent in this downward trajectory. Thus, the velocity of data—repeated observations of the same individual for many months—offers a different descriptive picture than an RCT that first observes people when they are in the throes of a legal event.

4.3. Predicting Harms

The final spotlight function of big data and computational methods is to predict small harms before they progress to justiciable ones. Returning to the housing law example, Greiner et al. (2012) study tenants who seek help from lawyers in addressing substantial legal issues with their landlords. The model of access to justice that RCTs use is thus one of reactive rights enforcement, where the legal system relies on tenants to (*a*) recognize that a rights violation has occurred, (*b*) decide to report the violation, and (*c*) navigate reporting bureaucracies, with RCTs then studying outcomes among those who seek help.

The final spotlight function of large-scale data and computational methods is to help advocates and legal services providers shift to proactive rights enforcement. Proactive rights enforcement is meant to address attrition through the "naming, blaming, and claiming" process—organizations can use large-scale data to help find employees who might not recognize that surprisingly low wages one month might constitute wage theft or recipients of public benefits who might not recognize a benefits discontinuation as within their rights. Once these individuals are found, the organizations can provide them help if the issue is a rights violation.

In one example, Ye et al. (2019) partner with a New York City government agency that sends outreach workers to canvass in rapidly changing neighborhoods. The outreach workers have limited time, and so want to target their canvassing to buildings where they are most likely to find tenants with serious issues with their landlords that, if left unaddressed, could progress to eviction and other outcomes that would bring each to court. Using supervised machine learning and a variety of large-scale data sources, i.e., details of the canvassing in past years and housing code violations data, the researchers find that, relative to legal outreach guided by community worker judgment, machine learning–guided outreach both helps find more cases of serious housing issues and distributes that help equitably across neighborhoods. In a similar project, data scientists partnered with the Washington, DC, housing code enforcement agency to target inspections to the highest-risk units (McCourt School of Public Policy 2020). The spotlight function of these projects was limited in scope compared with data for surveillance efforts like predictive policing and tenant screening: Rather than nationwide, the projects focused on two cities; rather than being undertaken by a private firm and then sold to cities or landlords, the projects were pro bono efforts by part-time data scientists. But with a better infrastructure in place, examples like these

show that the three Vs can guide predictive models aimed at addressing inequalities in who asks for help.

5. A DATA INFRASTRUCTURE FOR THAT SPOTLIGHT

As our examples illustrate, computational social science provides a methodology to document a range of ways in which law acts—and fails—on the ground, including the exploitation of enforcement gaps by powerful actors and flaws in enforcement strategies. Although these use cases show big data's potential to study the operation of legal institutions, there remains a paucity of computational research on how the formal justice system operates, or what the consequences are for people who become caught in that system. Courts and other civil justice institutions hold out the promise of levelling inequalities but, as sociolegal research has long demonstrated, often function to entrench or worsen power disparities (Albiston 1999, Galanter 1974). In the United States, tens of millions of people a year interact with formal legal institutions that address their housing, debt, and public benefits issues, creating vast data flows, yet few of these data are easily available for research.

These data flows could be analyzed to help courts, legal service providers, and administrative agencies to understand their operations, allocate their resources more efficiently, and increase the fairness of their decision-making processes and service delivery models. Yet these data are not easily available. Currently, data from courts, legal service providers, and administrative agencies are collected in multiple formats under a variety of incompatible taxonomies—for instance, different ways of categorizing which individuals are living in poverty—and housed among multiple institutions. Access to these data, moreover, is governed by a hodgepodge of statutes, regulations, court rules, and policies. These technical, regulatory, and public policy barriers have hindered both research efforts to understand the role of law in the lives of community members and policy efforts to make civil justice institutions more equitable.

One promising approach to making data about these institutions available for research and public policy debates is the creation of civil justice data repositories to collect, harmonize, and make available court, legal services, and agency data to researchers and policy makers under a joint governance regime. The goal of a civil justice data repository would be to extend social science knowledge about what happens to individuals before and after civil justice problems beyond the limited issue areas where researchers have begun to build large-scale data sets [e.g., for evictions (Desmond et al. 2018, Humphries et al. 2019) or human services (Font et al. 2018)]. For example, though an estimated 30 million unrepresented people appear in civil court every year, we have little understanding of (*a*) what precipitated these individual's issues (e.g., a parent struggles to collect child support payment and then faces medical debt collection for an unpaid bill), (*b*) why they did not receive legal representation (e.g., they contacted a legal service provider that was understaffed and unable to offer help or never contacted a legal service provider), and (*c*) the effects of that civil justice contact on short- and long-term outcomes (e.g., whether the marred credit score from the unpaid medical debt precipitated a move to a higher-poverty neighborhood) (Sandefur 2015, Self-Represent. Litig. Netw. 2019). In other domains, data repositories ease the sharing of data, and computational science produces insights relevant to public policy (Verhulst & Sangokoya 2015). There has been no similar surge of computational science research in the civil justice field, in great part because access to usable data is lacking.

Finally, once the data are available, the case studies in Section 3 show how to make the insights more immediately actionable for oversight of powerful organizations. A more continuous stream of evictions data could help local legal services providers predict where their help is needed, data on child welfare referral practices could help the Department of Health and Human Services's

Office for Civil Rights monitor illegal biases, and CMS could continue to monitor pharmaceutical marketing and physician prescribing practices. Each of these depends not only on data availability but also on trained researchers to help watchdog organizations effectively use those data, with others discussing ways to match "academics seeking opportunities to study important problems and policymakers seeking analytical muscle" (Chien & Sukhatme 2019).

6. CONCLUSION

In this review, we outline two polarized perspectives on the roles that big data and computational methods play in law and inequality. In one perspective, big data and tools like predictive modeling amplify inequalities in the law, subjecting vulnerable individuals to enhanced surveillance. In another, these data and tools serve an opposite function, shining a spotlight on inequality and subjecting powerful institutions to enhanced oversight.

We highlight the dual uses—data for surveillance and data as a spotlight—in three areas of law: housing, child welfare, and medical prescribing. Our review highlights asymmetries in which the lack of a data infrastructure to measure basic facts about inequality within the law—Which hospitals engage in aggressive debt collection? Which insurers are racially biased in their claims denials? Which states' hearing systems judge people protesting benefits denials most harshly?—has impeded the spotlight function. Future work should focus both on improving this data infrastructure and on using computational methods to shine a spotlight on inequalities within the law.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This collaboration grew out of a workshop at Georgetown Law University Center on Computing, Data Science, and Access to Justice, in June 2019, sponsored by National Science Foundation Grant Award 7773294.

LITERATURE CITED

Albiston C. 1999. The rule of law and the litigation process: the paradox of losing by winning. *Law Soc. Rev.* 33(4):869–910

Albiston CR, Sandefur RL. 2013. Expanding the empirical study of access to justice. *Wis. Law Rev.* 2013:101–20

Allen JA. 2019. The color of algorithms: an analysis and proposed research agenda for deterring algorithmic redlining. *Fordham Urban Law J.* 46:219–70

Aronowitz M, Golding E. 2019. *HUD's proposal to revise the disparate impact standard will impede efforts to close the homeownership gap*. Publ., Hous. Financ. Policy Cent., Urban Inst., Washington, DC. **https://www.urban.org/sites/default/files/publication/101015/huds_proposal_to_revise_the_disparate_impact_standard_0.pdf**

Barocas S, Selbst AD. 2016. Big data's disparate impact. *Calif. Law Rev.* 104:671–732

Berk R, Heidari H, Jabbari S, Kearns M, Roth A. 2018. Fairness in criminal justice risk assessments: the state of the art. *Sociol. Methods Res.* **https://doi.org/10.1177/0049124118782533**

Black RC, Hall ME, Owens RJ, Ringsmuth EM. 2016a. The role of emotional language in briefs before the US Supreme Court. *J. Law Courts* 4:377–407

Black RC, Owens RJ, Wedeking J, Wohlfarth PC. 2016b. The influence of public sentiment on Supreme Court opinion clarity. *Law Soc. Rev.* 50:703–32

Black RC, Spriggs JF. 2013. The citation and depreciation of US Supreme Court precedent. *J. Empir. Legal Stud.* 10:325–58

Bommarito MJ II, Katz D, Zelner J. 2009. Law as a seamless web? Comparison of various network representations of the United States Supreme Court corpus (1791–2005). In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pp. 234–35. New York: Assoc. Comput. Mach.

Brauneis R, Goodman EP. 2018. Algorithmic transparency for the smart city. *Yale J. Law Technol.* 20:103–76

Brayne S. 2017. Big data surveillance: the case of policing. *Am. Sociol. Rev.* 82:977–1008

Brayne S. 2018. The criminal law and law enforcement implications of big data. *Annu. Rev. Law Soc. Sci.* 14:293–308

Brick JM, Tourangeau R. 2017. Responsive survey designs for reducing nonresponse bias. *J. Off. Stat.* 33:735–52

Brown A, Chouldechova A, Putnam-Hornstein E, Tobin A, Vaithianathan R. 2019. Toward algorithmic accountability in public services: a qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. New York: Assoc. Comput. Mach.

Carlson K, Livermore MA, Rockmore D. 2015. A quantitative analysis of writing style on the U.S. Supreme Court. *Wash. Univ. Law Rev.* 93:1461–510

Carton S, Helsby J, Joseph K, Mahmud A, Park Y, et al. 2016. Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 67–76. New York: Assoc. Comput. Mach.

Chen DL, Ash E. 2020. Case vectors: spatial representations of the law using document embeddings. *Comput. Anal. Law*. In press

Chen MK, Haggag K, Pope DG, Rohla R. 2019. *Racial disparities in voting wait times: evidence from smartphone data*. Work. Pap. 26487, Natl. Bur. Econ. Res., Cambridge, MA

Chien CV, Sukhatme NE. 2019. A proposal for policypilots.gov. *Regulatory Review*, Nov. 19. **https://www.theregreview.org/2019/11/19/chien-sukhatme-proposal-policypilots-gov/**

Chouldechova A. 2017. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 5:153–63

Collatz A. 2019. SmartMove's ResidentScore vs. a typical credit score: Which is better? *SmartMove*, April 24. **https://www.mysmartmove.com/SmartMove/blog/residentscore-tailored-tenant-screening.page**

Crawford K, Schultz J. 2014. Big data and due process: toward a framework to redress predictive privacy harms. *Boston Coll. Law Rev.* 55:93–128

Desmond M. 2016. *Evicted: Poverty and Profit in the American City*. New York: Broadway Books

Desmond M, Bell M. 2015. Housing, poverty, and the law. *Annu. Rev. Law Soc. Sci.* 11:15–35

Desmond M, Gromis A, Edmonds L, Hendrickson J, Krywokulski K, et al. 2018. *Eviction Lab Methodology Report: Version 1.0*. Princeton, NJ: Princeton Univ. Press

Dettlaff AJ, Rivaux SL, Baumann DJ, Fluke JD, Rycraft JR, James J. 2011. Disentangling substantiation: the influence of race, income, and risk on the substantiation decision in child welfare. *Child. Youth Serv. Rev.* 33:1630–37

Dobbin F. 2009. *Inventing Equal Opportunity*. Princeton, NJ: Princeton Univ. Press

Dunn E, Grabchuk M. 2010. Background checks and social effects: contemporary residential tenant-screening problems in Washington State. *Seattle J. Soc. Just.* 9:319–99

Edelman LB. 1992. Legal ambiguity and symbolic structures: organizational mediation of civil rights law. *Am. J. Sociol.* 97:1531–76

Eubanks V. 2018a. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's

Eubanks V. 2018b. A child abuse prediction model fails poor families. *Wired*, Jan. 15. **https://www.wired.com/story/excerpt-from-automating-inequality/**

Felstiner WLF, Abel RL, Sarat A. 1980. The emergence and transformation of disputes: naming, blaming, claiming. . . *Law Soc. Rev.* 15:631–54

Ferguson AG. 2016. Policing predictive policing. *Wash. Univ. Law Rev.* 94:1109–89

Ferguson AG. 2019. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. New York: N.Y. Univ. Press

Font SA, Berger LM, Cancian M, Noyes JL. 2018. Permanency and the educational and economic attainment of former foster children in early adulthood. *Am. Sociol. Rev.* 83:716–43

Foster I, Ghani R, Jarmin RS, Kreuter F, Lane J. 2016. *Big Data and Social Science: A Practical Guide to Methods and Tools*. Boca Raton, FL: CRC

Frankenreiter J, Livermore M. 2020. Computational methods in legal analysis. *Annu. Rev. Law Soc. Sci.* 16:39–57

Fuller SR, Edelman LB, Matusik SF. 2000. Legal readings: employee interpretation and mobilization of law. *Acad. Manag. Rev.* 25:200–16

Galanter M. 1974. Why the "haves" come out ahead: speculations on the limits of legal change. *Law Soc. Rev.* 9:95–160

Gluck AR, Hall A, Curfman G. 2018. Civil litigation and the opioid epidemic: the role of courts in a national health crisis. *J. Law Med. Ethics* 46:351–66

Greiner DJ, Matthews A. 2016. Randomized control trials in the United States legal profession. *Annu. Rev. Law Soc. Sci.* 12:295–312

Greiner DJ, Pattanayak CW. 2011. Randomized evaluation in legal assistance: What difference does representation (offer and actual use) make? *Yale Law J.* 121:2118–214

Greiner DJ, Pattanayak CW, Hennessy J. 2012. The limits of unbundled legal assistance: a randomized study in a Massachusetts district court and prospects for the future. *Harvard Law Rev.* 126:901–89

Hampton RL, Newberger EH. 1985. Child abuse incidence and reporting by hospitals: significance of severity, class, and race. *Am. J. Public Health* 75:56–60

Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, et al. 2017. Challenges in administrative data linkage for research. *Big Data Soc.* 4. **https://doi.org/10.1177/2053951717745678**

Heimer CA, Staffen LR. 1995. Interdependence and reintegrative social control: labeling and reforming "inappropriate" parents in neonatal intensive care units. *Am. Sociol. Rev.* 60:635–54

Helsby J, Carton S, Joseph K, Mahmud A, Park Y, et al. 2018. Early intervention systems: predicting adverse interactions between police and the public. *Crim. Justice Policy Rev.* 29:190–209

Hepburn P, Faber J, Kneebone E, Hendrickson J, Thomas T, et al. 2019. *Super session: causes and consequences of eviction*. Presented at the 41st Annual Fall Research Conference, Association for Public Policy Analysis and Management, Denver, CO, Nov. 7–9

Humphries JE, Mader NS, Tannenbaum DI, Van Dijk WL. 2019. *Does eviction cause poverty? Quasi-experimental evidence from Cook County, IL*. Work. Pap. 26139, Natl. Bur. Econ. Res., Washington, DC

Kalev A, Dobbin F. 2006. Enforcement of civil rights law in private workplaces: the effects of compliance reviews and lawsuits over time. *Law Soc. Inq.* 31:855–903

Kleysteuber R. 2006. Tenant screening thirty years later: a statutory proposal to protect public records. *Yale Law J.* 116:1344–88

Knox D, Lucas C. 2019. *A dynamic model of speech for the social sciences*. Work. Pap., Princeton Univ., Princeton, NJ. **https://asiapolmeth.princeton.edu/sites/default/files/polmeth/files/deanknox.pdf**

Lazer D, Pentland A, Adamic L, Aral S, Barabási AL, et al. 2009. Computational social science. *Science* 323:721–23

Lazer D, Radford J. 2017. Data ex machina: introduction to big data. *Annu. Rev. Sociol.* 43:19–39

Lebovits G, Addonizio JM. 2015. *The use of tenant screening reports and tenant blacklisting*. LEGALEase Pamphlet, N.Y. State Bar Assoc., Albany, NY

Legal Aid Society. 2020. *Cop Accountability Project (Cap)*. **https://legalaidnyc.org/programs-projects-units/the-cop-accountability-project/**

Lippincott T, Carrell A. 2018. Observational comparison of geo-tagged and randomly-drawn tweets. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, ed. M Nissim, V Patti, B Plank, pp. 50–55. Stroudsburg, PA: Assoc. Comput. Linguist.

Massey DS. 2015. The legacy of the 1968 Fair Housing Act. *Sociol. Forum* 30:571–88

McCourt School of Public Policy. 2020. McCourt students use data to improve housing inspections in DC. *News*, Jan. 13. **https://mccourt.georgetown.edu/news/using-data-to-improve-housing-inspections-in-dc/**

Metcalf GR. 1988. *Fair Housing Comes of Age*. Contrib. Political Sci. 198. Santa Barbara, CA: Praeger
Mooney SJ, Westreich DJ, El-Sayed AM. 2015. Epidemiology in the era of big data. *Epidemiology* 26:390–94
Munger FW, Seron C. 2017. Race, law, and inequality, 50 years after the Civil Rights era. *Annu. Rev. Law Soc. Sci.* 13:331–50
O'Neil C. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown
Ouellet M, Hashimi S, Gravel J, Papachristos AV. 2019. Network exposure and excessive use of force: investigating the social transmission of police misconduct. *Criminol. Public Policy* 18:675–704
Pasquale F. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard Univ. Press
Pettigrew S. 2017. The racial gap in wait times: why minority precincts are underserved by local election officials. *Political Sci. Q.* 132:527–47
Prescott JJ, Starr SB. 2020. Expungement of criminal convictions: an empirical study. *Harvard Law Rev.* 133:2460–555
Putnam-Hornstein E, Prindle JJ, Leventhal JM. 2016. Prenatal substance exposure and reporting of child maltreatment by race and ethnicity. *Pediatrics* 138:e20161273
Raghavan M, Barocas S, Kleinberg J, Levy K. 2019. Mitigating bias in algorithmic employment screening: Evaluating claims and practices. arXiv:1906.09208 [cs.CY]
Randall SM, Ferrante AM, Boyd JH, Semmens JB. 2013. The effect of data cleaning on record linkage quality. *BMC Med. Inform. Decis. Making* 13:64
Richardson R, Schultz J, Crawford K. 2019. Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. *N.Y. Univ. Law Rev. Online* 94:192–233
Roberts DE. 2003. Child welfare and civil rights. *Univ. Ill. Law Rev.* 2003:171–82
Rose H. 2019. *How the Trump administration's plan to limit disparate impact liability would undermine the Fair Housing Act's goal of promoting residential integration*. Work. Pap., Law School, Loyola Univ., Chicago, IL. **https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3464555**
Sacarny A, Barnett ML, Le J, Tetkoski F, Yokum D, Agrawal S. 2018. Effect of peer comparison letters for high-volume primary care prescribers of quetiapine in older and disabled adults: a randomized clinical trial. *JAMA Psychiatry* 75:1003–11
Sacarny A, Yokum D, Finkelstein A, Agrawal S. 2016. Medicare letters to curb overprescribing of controlled substances had no detectable effect on providers. *Health Aff.* 35:471–79
Salganik M. 2019. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton Univ. Press
Sandefur RL. 2008. Access to civil justice and race, class, and gender inequality. *Annu. Rev. Sociol.* 34:339–58
Sandefur RL. 2014. *Accessing justice in the contemporary USA: findings from the community needs and services study*. Doc., Am. Bar Found., Chicago. **http://www.americanbarfoundation.org/uploads/cms/documents/sandefur_accessing_justice_in_the_contemporary_usa._aug._2014.pdf**
Sandefur RL. 2015. What we know and need to know about the legal needs of the public. *S. C. Law Rev.* 67:443–60
Selbst AD. 2017. Disparate impact in big data policing. *Ga. Law Rev.* 52:109–95
Self-Represent. Litig. Netw. 2019. *SRLN brief: How many SRLs?* **https://www.srln.org/node/548/srln-brief-how-many-srls-srln-2015**
Vagle JL. 2016. Tightening the OODA loop: police militarization, race, and algorithmic surveillance. *Mich. J. Race Law* 22:101–37
Vaithianathan R, Kulick E, Putnam-Hornstein E, Benavides Prado D. 2019. *Allegheny Family Screening Tool: Methodology, Version 2*. Pittsburgh, PA: Allegheny County
Van Zee A. 2009. The promotion and marketing of OxyContin: commercial triumph, public health tragedy. *Am. J. Public Health* 99:221–27
Verhulst S, Sangokoya D. 2015. *Data collaboratives: exchanging data to improve people's lives. Medium*, April 22. **https://medium.com/@sverhulst/data-collaboratives-exchanging-data-to-improve-people-s-lives-d0fcfc1bdd9a**
Whittaker M, Crawford K, Dobbe R, Fried G, Kaziunas E, et al. 2018. *AI Now Report 2018*. Rep., AI Now Inst., N.Y. Univ., New York

Ye T, Johnson R, Fu S, Copeny J, Donnelly B, et al. 2019. Using machine learning to help vulnerable tenants in New York City. In *Proceedings of the Conference on Computing & Sustainable Societies*, pp. 248–58. New York: Assoc. Comput. Mach.

Zalnieriute M, Burton L, Boughey J, Bennett Moses L, Logan S. 2020. From rule of law to statute drafting: legal issues for algorithms in government decision-making. In *Cambridge Handbook of the Law of Algorithms*, W Barfield, pp. 19–30. Cambridge, UK: Cambridge Univ. Press. In press