BIG DATA'S DISPARATE IMPACT

Solon Barocas[*]
Andrew D. Selbst[†]

Big data claims to be neutral. It isn't.

Advocates of algorithmic techniques like data mining argue that they eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with. Data mining can inherit the prejudices of prior decision-makers or reflect the widespread biases that persist in society at large. Often, the "patterns" it discovers are simply preexisting societal patterns of inequality and exclusion. Unthinking reliance on data mining can deny members of vulnerable groups full participation in society. Worse still, because the resulting discrimination is almost always an unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers, it can be unusually hard to identify the source of the problem or to explain it to a court.

This Article examines these concerns through the lens of American anti-discrimination law—more particularly, through Title VII's prohibition on discrimination in employment. In the absence of a demonstrable intent to discriminate, the best doctrinal hope for data mining's victims would seem to lie in disparate impact doctrine. Case law and the EEOC's Uniform Guidelines, though, hold that a practice can be justified as a business necessity where its outcomes are predictive of future employment outcomes, and data mining is specifically designed to find such statistical correlations. As a result, Title VII would appear to bless its use, even though the correlations it discovers will often reflect historic patterns of prejudice, others' discrimination against members of vulnerable groups, or flaws in the underlying data.

Addressing the sources of this unintentional discrimination and

[*] Postdoctoral Research Associate, Center for Information Technology Policy, Princeton University; Ph.D. 2014, New York University, Department of Media, Culture and Communication.

[†] Law Clerk to the Honorable Jane R. Roth, U.S. Court of Appeals for the Third Circuit; J.D. 2011, University of Michigan Law School. The authors would like to thank Jane Bambauer, James Grimmelmann, Janine Hiller, Chris Hoofnagle, Patrick Ishizuka, Aaron Konopasky, Joshua Kroll, Arvind Narayanan, Paul Ohm, Scott Peppet, Joel Reidenberg, David Robinson, Kathy Strandburg, members of the Privacy Research Group at New York University, and the participants of the 2014 Privacy Law Scholars Conference for their helpful comments. Special thanks also to Helen Nissenbaum and the Information Law Institute at New York University for giving us an interdisciplinary space to share ideas, allowing this paper to come about.

remedying the corresponding deficiencies in the law will be difficult technically, difficult legally, and difficult politically. There are a number of practical limits to what can be accomplished computationally. For example, where the discrimination occurs because the data being mined is itself a result of past intentional discrimination, there is frequently no obvious method to adjust historical data to rid it of this taint. Corrective measures that alter the results of the data mining after it is complete would tread on legally and politically disputed terrain. These challenges for reform throw into stark relief the tension between the two major theories underlying anti-discrimination law: nondiscrimination and anti-subordination. Finding a solution to big data's disparate impact will require more than best efforts to stamp out prejudice and bias; it will require wholesale reexamination of the meanings of "discrimination" and "fairness."

INTRODUCTION

"Big Data" is the buzzword of the decade.[1] Advertisers want data to reach profitable consumers,[2] medical professionals to find side effects of prescription drugs,[3] supply chain operators to optimize their delivery routes,[4] police to determine where to focus resources,[5] and social scientists to study human interactions.[6] Though useful, however, data is not a panacea. Where data is used predictively to assist decisionmaking, it can affect the fortunes of whole classes of people in consistently unfavorable ways. Sorting and selecting for the best or most profitable candidates means generating a model with winners and losers. If data miners are not careful, that sorting might create disproportionately adverse results concentrated within historically disadvantaged groups in ways that look a lot like discrimination.

Despite living in the post-civil rights era, discrimination persists in American society and is stubbornly pervasive in employment, housing, credit, and consumer markets.[7] While discrimination certainly endures in part due to decisionmakers' irrational prejudice, a great deal of modern-day inequality can be attributed to what sociologists call "institutional" discrimination, whereby unconscious, implicit biases and inertia within society's institutions account for a large part of the disparate effects observed, rather than intentional choices.[8] Approached without care, data mining can reproduce existing patterns of discrimination, inherit the

[1] *Contra* Sanjeev Sardana, *Big Data: It's Not A Buzzword, It's A Movement*, FORBES (November 20, 2013), http://www.forbes.com/sites/sanjeevsardana/2013/11/20/bigdata/.

[2] Tanzina Vega, *New Ways Marketers Are Manipulating Data to Influence You*, N.Y. TIMES' BITS (June 19, 2013), http://bits.blogs.nytimes.com/2013/06/19/new-ways-marketers-are-manipulating-data-to-influence-you/.

[3] Nell Greenfieldboyce, *Big Data Peeps At Your Medical Records To Find Drug Problems*, NPR (July 21, 2014), http://www.npr.org/blogs/health/2014/07/21/332290342/big-data-peeps-at-your-medical-records-to-find-drug-problems.

[4] *Business By Numbers*, THE ECONOMIST (Sept. 13 2007), *available at* http://www.economist.com/node/9795140.

[5] Nadya Labi, *Misfortune Teller*, THE ATLANTIC (January/February 2012), *available at* http://www.theatlantic.com/magazine/archive/2012/01/misfortune-teller/308846/.

[6] David Lazer et al., *Computational Social Science*, 323 SCIENCE 721 (2009).

[7] Devah Pager & Hana Shepherd, *The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets*, 34 ANNU. REV. SOCIOL. 181 (2008).

[8] Andrew Grant-Thomas & john a. powell. *Toward a Structural Racism Framework*, 15 POVERTY & RACE 3 (2006)("'Institutional racism' was the designation given in the late 1960s to the recognition that, at very least, racism need not be individualist, essentialist or intentional.")

prejudice of prior decision-makers, or simply reflect the widespread biases that persist in society. It can even have the perverse result of exacerbating existing inequalities by suggesting that historically disadvantaged groups actually deserve less favorable treatment. Algorithmic decision procedures could exhibit these tendencies even if they have not been hand-coded to do so, either by design or by accident. Scholars and policymakers have tended to worry that the inscrutability of algorithms will keep these intentions or mistakes hidden,[9] but discrimination may be an artifact of the data mining process itself, rather than a result of programmers assigning certain factors inappropriate weight.

That the discrimination at issue is unintentional means that even honest attempts to certify the absence of prejudice on the part of those involved in the data mining process may wrongly confer the imprimatur of impartiality on the resulting decisions.[10] Furthermore, because the mechanism through which data mining visits systematic disadvantages upon protected classes is far less obvious in cases of unintentional discrimination, the injustice may be harder to identify and address.

In May 2014, the White House released a report titled *Big Data: Seizing Opportunities, Preserving Values*, the result of President Obama's 90-day Big Data Review (Podesta Report).[11] "A significant finding of th[e] report is that big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace."[12] The report suggests that there may be unintended discriminatory effects from data mining, but does not detail how they might come about.[13] Because the origin of the discriminatory effects remains unexplored, the report's approach does not address the full scope of the problem.

The Podesta Report, as one might expect from the executive branch, seeks to address these effects primarily by finding new ways to enforce

[9] Danielle Keats Citron, *Technological Due Process*, 85 Wash. U.L. Rev. 1249 (2008); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. Rev. 93 (2014); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 Wash. L. Rev. 1 (2014).

[10] Citron, *supra* note 9.

[11] EXECUTIVE OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES (2014), *at* http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf

[12] *Id.* at 1.

[13] *Id.* at 64 ("This combination of circumstances and technology raises difficult questions about how to ensure that discriminatory effects resulting from automated decision processes, whether intended or not, can be detected, measured, and redressed.")

existing law. Its main recommendation in the area of discrimination is that enforcement agencies, such as the Department of Justice, Federal Trade Commission, Consumer Financial Protection Bureau, and Equal Employment Opportunity Commission (EEOC) increase their technical expertise and "develop a plan for investigating and resolving violations of law in such cases."[14]

As this Article demonstrates though, to a large degree, existing law cannot handle these problems. The argument is grounded in Title VII because it has the most developed case law and scholarship of American anti-discrimination jurisprudence and there exists a rapidly emerging field of "work-force science,"[15] for which Title VII will be the primary vehicle for regulation. Under Title VII, it turns out that some, if not most, instances of discriminatory data mining will not generate liability under existing law. While the Article does not show this to be true outside of Title VII itself, the problem is not particular to Title VII, but rather is a problem of our current approach to anti-discrimination jurisprudence, with its focus on procedural fairness. The analysis will likely apply to other traditional areas of discrimination such as housing and education and similar problems will arise in areas that regulate legitimate economic discrimination by restricting access to and use of certain types of information, as in credit and insurance.

This Article proceeds in three Parts. Part I introduces the computer science literature on data mining and proceeds through the various steps of solving a problem this way: defining the target variable, labeling and collecting the training data, feature selection, and making decisions on the basis of the resulting model. Each of these steps creates possibilities for a final result that has a disproportionately adverse impact on protected classes, whether by specifying the problem to be solved in ways that affect classes differently, failing to recognize or address statistical biases, reproducing past prejudice, or considering an insufficiently rich set of factors. Even in situations where data miners are extremely careful, they can still effect discriminatory results with models that, quite unintentionally, pick out proxy variables for protected classes. Finally, Part I notes that data mining poses the additional problem of giving data miners the ability to disguise intentional discrimination as unintentional.

In Part II, the Article reviews Title VII jurisprudence as it applies to data mining. It discusses both disparate treatment and disparate impact, examining which of the various mechanisms identified in Part I will trigger liability under either Title VII theory. At first blush, either theory is viable.

[14] *Id.*

[15] Steve Lohr, "Big Data, Trying to Build Better Workers," *The New York Times*, April 20, 2013, http://www.nytimes.com/2013/04/21/technology/big-data-trying-to-build-better-workers.html.

Disparate treatment is viable because data mining systems treat everyone differently; that is their purpose. Disparate impact should be viable because, as the Article shows, data mining can have various discriminatory effects. But as the Part concludes, data mining combines some well-known problems with discrimination doctrine with new challenges particular to data mining systems, such that liability for discriminatory data mining will be hard to find. Part II concludes with a discussion of the new problems of proof that arise for intentional discrimination in this context.

Finally, Part III addresses the difficulties legal reformers would face in addressing the deficiencies found in Part II. These difficulties take two forms: complications internal to the logic of data mining and political and constitutional difficulties external to the problem that would prevent Title VII reform from addressing its current deficiencies. Internally, the different steps in a data mining problem require constant normative judgments, many of which are hidden, and most of which cannot be solved generally enough to permit of legislative resolution, even assuming consensus on these normative questions were possible. Externally, data mining will force society to explicitly rebalance the two justifications for anti-discrimination law – rooting out intentional discrimination and equalizing the status of historically disadvantaged communities – because methods of proof and remedies for discrimination will often require an explicit commitment to substantive remediation rather than merely procedural remedies. In certain cases, it will be simply impossible to rectify these discriminatory results without engaging with the question of what level of substantive inequality is proper or acceptable in a given context. Given current political realities and trends in constitutional doctrine, legislation enacting a remedy that results from these discussions faces an uphill battle.

To be sure, data mining is a very useful construct. It even has the potential to be a boon to those who would not discriminate, by formalizing decision-making processes and thus limiting the influence of individual bias. But where data mining does perpetuate discrimination, society does not have a ready answer for what to do about it. It's time to have that discussion.

I. HOW DATA MINING DISCRIMINATES

Although commentators have ascribed myriad forms of discrimination to data mining,[16] there remains significant confusion over the precise mechanisms that render data mining discriminatory. This Part

[16] Solon Barocas, *Data Mining and the Discourse on Discrimination* (Proceedings of Data Ethics Workshop, August 24, 2014).

develops a taxonomy that isolates and explicates the specific technical issues that can give rise to models whose use in decision-making may have a disproportionately adverse impact on protected classes.

By definition, data mining is *always* a form of statistical (and therefore seemingly rational) discrimination. Indeed, the very point of data mining is to provide a rational basis upon which to distinguish between individuals and to reliably confer to the individual the qualities possessed by those who seem statistically similar. Nevertheless, data mining holds the potential to unduly discount members of legally protected classes and to place them at systematic relative disadvantage. Unlike more subjective forms of decision-making, data mining's ill effects are often not traceable to human bias, conscious or unconscious. This Part describes five mechanisms by which these disproportionately adverse outcomes might occur, walking through a sequence of key steps in the overall data mining process.

A. Defining the "Target Variable" and "Class Labels"

In contrast to those traditional forms of data analysis that simply return records or summary statistics in response to a specific query, data mining attempts to locate statistical relationships in a dataset. In particular, it automates the process of discovering useful patterns, revealing regularities upon which subsequent decision-making can rely. The accumulated set of discovered relationships is commonly called a "model," and these models can be employed to automate the process of classifying entities or activities of interest, estimating the value of unobserved variables, or predicting future outcomes.[17] Familiar examples of such applications include spam or fraud detection, credit scoring, and insurance pricing. These all involve attempts to determine the status or likely outcome of cases under consideration based solely on access to *correlated* data. Data mining helps identify cases of spam and fraud and anticipate default and poor health by treating these states and outcomes as a function of some other set of observed characteristics. In particular, by exposing so-called "machine learning" algorithms to examples of the cases of interest (previously identified instances of fraud, spam, default, and poor health), the algorithm "learns" which related attributes or activities can serve as potential proxies for those qualities or outcomes of interest. In the machine learning and data mining literature, these states or outcomes of interest are

[17] More formally, classification deals with discrete outcomes, estimation deals with continuous variables, and predictions can deal with both discrete outcomes and continuous variables, but specifically states or values *in the future*. COMMITTEE ON THE ANALYSIS OF MASSIVE DATA, ET AL., FRONTIERS IN MASSIVE DATA ANALYSIS 66-69 (2013), http://www.nap.edu/catalog.php?record_id=18374.

known as "target variables."[18]

The proper specification of the target variable is frequently not obvious, and it is the data miner's task to define it. In doing so, data miners must translate some amorphous problem into a question that can be expressed in more formal terms that computers can parse. In particular, data miners must determine how to solve the problem at hand by translating it into a question about the value of some target variable. The open-endedness that characterizes this part of the process is often described as the "art" of data mining. This initial step requires a data miner to "understand[] the project objectives and requirements from a business perspective [and] then convert[] this knowledge into a data mining problem definition."[19] Through this necessarily subjective process of translation, though, data miners may unintentionally parse the problem and define the target variable in such a way that protected classes happen to be subject to systematically less favorable determinations.

Problem specification is not a wholly arbitrary process, however. Data mining can only address problems that lend themselves to formalization as questions about the state or value of the target variable. Data mining works exceedingly well for dealing with fraud and spam because the question that data mining answers in these cases relies on extant, binary categories. A given instance either is or is not fraud or spam, and the definitions of fraud or spam are, for the most part, uncontroversial. A computer can then flag or refuse transactions or redirect emails according to well-understood categorizations.[20] Data miners can simply rely on these simple, pre-existing categories to define the so-called "class labels": the different classes between which a model should be able to distinguish.

Sometimes, defining the target variable involves the creation of *new* classes. Consider credit scoring, for instance. Although now taken for granted, the predicted likelihood of missing a certain number of loan

[18] The machine learning community refers to classification, estimation, and prediction—the techniques that we discuss in this Article—as 'supervised' learning because analysts must actively specify a target variable of interest. *Id.* Other techniques known as 'unsupervised' learning do not require any such target variables and instead search for general structures in the dataset, rather than patterns that specifically related to some state or outcome. *Id.* Clustering is the most common example of 'unsupervised' learning, in that clustering algorithms simply reveal apparent hot spots when plotting the data in some fashion. *Id.* We limit the discussion to supervised learning because we are primarily concerned with the sorting, ranking, and predictions enabled by data mining.

[19] PETE CHAPMAN, ET AL., CRISP-DM 1.0: STEP-BY-STEP DATA MINING GUIDE 13 (2000).

[20] Though described as a matter of detection, this is really a classification task, where any given transaction or email can belong to one of two possible classes, respectively: fraud or not fraud or spam or not spam.

repayments is not a self-evident answer to the question of how to successfully extend credit to consumers.[21] Unlike fraud or spam, "creditworthy" as a category is an artifact of defining the problem in this way. More to the point, there is no way to directly measure creditworthiness because the very notion of creditworthiness is a function of the particular way the credit industry has constructed the credit issuing and repayment system—one in which an individual's capacity to repay some minimum amount of an outstanding debt on a monthly basis is taken to be a non-arbitrary standard by which to determine in advance and all-at-once whether he is worthy of credit.[22]

Data mining has many uses beyond the four mentioned above. As discussed in the introduction, this Article will focus on the use of data mining for employment decisions to illustrate the broader points. Extending this discussion to employment, then, where employers turn to data mining to develop ways of improving and automating their search for good employees, they face a number of crucial choices.

Like creditworthiness, the definition of a good employee is not a given. "Good" must be defined in ways that correspond to measurable outcomes: relatively higher sales, shorter production time, or longer tenure, for example. When employers use data mining to find good employees, they are, in fact, looking for employees whose observable characteristics suggest, based on the evidence that an employer has assembled, that they would meet or exceed some monthly sales threshold, that they would perform some task in less than a certain amount of time, or that they would remain in their positions for more than a set number of weeks or months. Rather than drawing categorical distinctions along these lines, data mining could also estimate or predict the specific numerical value of sales, production time, or tenure period, enabling employers to rank rather than simply sort employees.

These may seem like eminently reasonable things for employers to want to predict, but they are, by necessity, only part of an array of possible ways of defining what "good" means. An employer may attempt to define the target variable in a more holistic way—by, for example, relying on the grades that prior employees have received in annual reviews, which are supposed to reflect an overall assessment of performance. These target

[21] MARTHA POON, WHAT LENDERS SEE — A HISTORY OF THE FAIR ISAAC SCORECARD, (2012)(unpublished dissertation), *available at* https://order.proquest.com/OA_HTML/pqdtibeCCtpItmDspRte.jsp?item=3518969&sitex=10020:22372:US&track=DxWeb&dlnow=1&rpath=http%3A%2F%2Fdissexpress.umi.com%2Fdxweb%2Fresults.html%3FQryTxt%3Dmartha%2Bpoon%26By%3D%26Title%3D%26pubnum%3D.

[22] David J. Hand, *Classifier Technology and the Illusion of Progress*, 21 STATISTICAL SCI. 1 (2006).

variable definitions simply inherit the formalizations involved in pre-existing assessment mechanisms, which in the case of human-graded performance reviews, may be far less consistent.[23]

The general lesson to draw from this discussion is that the definition of the target variable and its associated class labels will determine what data mining happens to find. While critics of data mining have tended to focus on inaccurate classifications (false positives and false negatives), as much—if not more—danger resides in the definition of the class label itself and the subsequent labeling of examples from which rules are inferred.[24] While different choices for the target variable and class labels can seem more or less reasonable, valid concerns with discrimination enter at this stage because the different choices may have a greater or lesser adverse impact on protected classes. For example, as later sections will explain, electing to hire on the basis of predicted tenure is much more likely to have a disparate impact on certain protected classes than hiring decisions that turn on some estimate of worker productivity.

B. Training Data

As described above, data mining learns by example. Accordingly, what a model learns depends on the examples to which it has been exposed. The data that function as examples are known as training data: quite literally the data that train the model to behave in a certain way. The character of the training data can have meaningful consequences for the lessons that data mining happens to learn. As computer science scholars explain, discriminatory training data leads to discriminatory models.[25] This can mean two rather different things, though: 1) if data mining treats cases in which prejudice has played some role as valid examples from which to learn a decision-making rule, that rule may simply reproduce the prejudice involved in these earlier cases; and 2) if data mining draws inferences from

[23] Joseph M Stauffer & M Ronald Buckley, *The Existence and Nature of Racial Bias in Supervisory Ratings*, 90 J. OF APPLIED PSYCHOLOGY 586 (2005)(showing evidence of racial bias in performance evaluations). Nevertheless, devising new target variables can have the salutary effect of forcing decisionmakers to think much more concretely about the outcomes that justifiably determine whether someone is a "good" employee. The explicit enumeration demanded of data mining thus also presents an opportunity to make decision-making more consistent, more accountable, and fairer overall. This, however, requires conscious effort and careful thinking and is not a natural consequence of adopting data mining.

[24] *See, infra* Part I.B.

[25] Bart Custers, *Data Dilemmas in the Information Society: Introduction and Overview*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 3, 20 (Bart Custers, et al. ed. 2013).

a biased sample of the populations to which the inferences are expected to generalize, any decisions that rests on these inferences may systematically disadvantage those who are under- or over-represented in the dataset. Both can affect the training data in ways that lead to discrimination, but the mechanisms are sufficiently distinct that they warrant separate treatment.

1. Labeling Examples

Labeling examples is the process by which the training data is separated out into the relevant classifications. In cases of fraud or spam, the data miners draw from examples that come pre-labeled: when individual customers report charges for items that they did not purchase or when users mark a message as spam, they are actually labeling transactions and email for the providers of credit and webmail. Likewise, an employer using grades previously given at performance reviews is also using pre-labeled examples. In certain cases, however, there may not be any labeled data and data miners may have to figure out a way to label examples themselves. This can be a laborious process and it is frequently fraught with peril.[26]

Often the best labels for different classifications will be open to debate. Even where the class labels are uncontested or uncontroversial, they may present a problem because analysts will often face difficult choices in deciding which of the available labels best applies to a particular example. Certain cases may present some, but not all, criteria for inclusion in a particular class.[27] The situation might also work in reverse, where the class labels are insufficiently precise to capture meaningful differences between cases. Such imperfect matches will demand the exercise of judgment.

The unavoidably subjective labeling of examples can skew the resulting findings in such a way that any decisions taken on the basis of those findings will characterize all *future* cases along the same lines, even if such characterizations would seem plainly erroneous to analysts who looked more closely at the individual cases. For all their potential problems, though, the labels applied to the training data must serve as ground truth. The kinds of subtle mischaracterizations that happened during training will be impossible to detect when evaluating the performance of a model, because the training is taken as a given at that point.[28] Thus, decisions taken on the basis of discoveries that rest on haphazardly labeled data or data labeled in a systematically, though unintentionally, biased manner will seem valid according to the customary validation methods employed by data miners. So long as prior decisions affected by some form of prejudice serve

[26] Hand, *supra* note 22.

[27] *Id.*

[28] *Id.*

as examples of *correctly* rendered determinations, data mining will necessarily infer rules that exhibit the same prejudice.

Consider a real-world example from a different context as to how biased data labeling can skew results. St George's Hospital, in the United Kingdom, developed a computer program to help sort medical school applicants on the basis of previous admission decisions.[29] Those admissions decisions, it turns out, had systematically disfavored racial minorities and women with otherwise equal credentials.[30] In drawing rules from biased prior decisions, St. George's Hospital unknowingly devised an automated process that possessed these vary same prejudices. As editors at the *British Medical Journal* noted at the time, "the program was not introducing new bias but merely reflecting that already in the system."[31] Were an employer to undertake a similar plan to automate its hiring decisions by inferring a rule from previous decisions that had been swayed by prejudice, the employer would likewise arrive at a decision procedure that simply reproduces the prejudice of prior decision-makers.[32] Indeed, it would turn the conscious prejudice or implicit bias of individuals involved in previous decision-making into a formalized rule that would systematically discount all applicants in this way.

Not only can data mining inherit *prior* prejudice through the mislabeling of examples, it can also reflect current prejudice through the ongoing behavior of users taken as inputs to data mining. This is what Latanya Sweeney discovered in a study that found that Google queries for black-sounding names were more likely to return contextual (i.e., key-word triggered) advertisements for arrest records than those for white-sounding names.[33] Sweeney confirmed that the companies paying for these ads had not set out to focus on black-sounding names; rather, the fact that black-sounding names were more likely to trigger such advertisements seemed to be an artifact of the algorithmic process that Google employs to determine which advertisements to deliver alongside the results for certain queries. Although the details of the process by which Google computes the so-called "quality score" according to which it ranks advertisers' bids is not fully

[29] Stella Lowry & Gordon Macpherson, *A Blot on the Profession*, 296 BRITISH MEDICAL J. 657 (1988).

[30] *Id.*

[31] *Id*. at 657.

[32] As an example in the traditional hiring context, seniority was often used to both intentionally and unintentionally reproduce past prejudice after the passage of Title VII. *See International Brotherhood of Teamsters v. United States*, 431 U.S. 324, 353 (1977)(noting that seniority reproduced past prejudice, but finding that unless it was used as a pretext, it is not barred by Title VII).

[33] Latanya Sweeney, *Discrimination in Online Ad Delivery,* 56 COMM. OF THE ACM 44 (2013).

known, one important factor is the predicted likelihood, based on historical trends, that users will click on an advertisement.[34] As Sweeney points out, the process "learns over time which ad text gets the most clicks from viewers of the ad" and promotes that advertisement in its rankings accordingly.[35] Sweeney posits that this aspect of the process could result in the differential delivery of advertisements that reflect the kinds of prejudice held by those exposed to the advertisements.[36] In attempting to cater to the preferences of users, Google will unintentionally reproduce the existing prejudices that inform users' choices.

A similar situation could conceivably arise on websites that recommend potential employees to employers, as LinkedIn does through its Talent Match feature. If LinkedIn determines which candidates to recommend on the basis of the demonstrated interest of employers in certain types of candidates, Talent Match will offer recommendations that reflect whatever biases employers happen to exhibit.[37] In particular, if LinkedIn's algorithm observes that employers disfavor certain candidates that are members of a protected class, Talent Match may decrease the rate at which it recommends these types of candidates to employers. The recommendation engine would learn to cater to the prejudicial preferences of employers.

2. Data Collection

How data miners collect examples also matters. Organizations that do not or cannot observe different populations in a consistent way and with equal coverage will amass evidence that fails to reflect the actual incidence and relative proportion of some attribute or activity in the under- or over-observed group. Consequently, decisions that depend on conclusions drawn

[34] Google, *Check and Understand Quality Score*, https://support.google.com/adwords/answer/2454010?hl=en (last visited July 26, 2014).

[35] Sweeney, *supra* note 33

[36] The fact that black people may be convicted of crimes at a higher rate than non-black people does not explain why those who search for black-sounding names would be any more likely to click on advertisements that mention an arrest record than those who see the same exact advertisement when they search for white-sounding names. If the advertisement implies, in both cases, that a person of that particular name has an arrest record, as Sweeney shows, the only reason the advertisements keyed to black-sounding names should receive greater attention is if searchers confer greater significance to the fact of prior arrests when the person happens to be black. Sweeney, *supra* note 33.

[37] Dan Woods, *LinkedIn's Monica Rogati on "What Is a Data Scientist?,"* FORBES (November 27, 2011) *at* http://www.forbes.com/sites/danwoods/2011/11/27/linkedins-monica-rogati-on-what-is-a-data-scientist/. Whether LinkedIn determines its recommendations by looking at the demonstrated interest in certain types of candidates is unclear.

from this data may discriminate against members of these groups.

The data might suffer from a variety of problems: the individual records that a company maintains about a person might have serious mistakes,[38] the records of the entire protected class of which this person is a member might also have similar mistakes at a higher rate than other groups, and the entire set of records may fail to reflect members of protected classes in accurate proportion to others. In other words, the quality and representativeness of records might vary in ways that correlate with class membership (e.g., institutions might maintain systematically less accurate, precise, timely, and complete records). Even a dataset with individual records of consistently high quality can suffer from statistical biases that fail to represent different groups in accurate proportions. Much attention has focused on the harms that might befall individuals whose records in various commercial databases are error-ridden,[39] but far less consideration has been paid to the systematic disadvantage that members of protected classes may suffer from being miscounted and the resulting biases in their representation in the evidence base.

Recent scholarship has begun to stress this point. Jonas Lerman, for example, worries about "the nonrandom, systemic omission of people who live on big data's margins, whether due to poverty, geography, or lifestyle, and whose lives are less 'datafied' than the general population's."[40] Kate Crawford has likewise warned, "because not all data is created or even collected equally, there are 'signal problems' in big-data sets—dark zones or shadows where some citizens and communities are overlooked or underrepresented."[41] Errors of this sort may befall historically disadvantaged groups at higher rates because they are less involved in the formal economy and its data-generating activities, because they have unequal access to and relatively less fluency in the technology necessary to

[38] Data quality is a topic of lively practical and philosophical debate. *See, e.g.*, Richard Y. Wang & Diane M. Strong, *Beyond Accuracy: What Data Quality Means to Data Consumers*, 12 J. MANAGEMENT INFO. SYS. 5 (1996); Luciano Floridi, *Information Quality*, 26 PHIL. & TECH. 1 (2013). The component parts of data quality have been thought to include accuracy, precision, completeness, consistency, validity, and timeliness, though this catalog of features is far from settled. *See* LARRY P. ENGLISH, INFORMATION QUALITY APPLIED: BEST PRACTICES FOR IMPROVING BUSINESS INFORMATION, PROCESSES AND SYSTEMS (2009).

[39] *See, e.g.*, FEDERAL TRADE COMMISSION: REPORT TO CONGRESS UNDER SECTION 319 OF THE FAIR AND ACCURATE CREDIT TRANSACTIONS ACT OF 2003 (2012)(finding that 20% of consumers had an error in one or more of their three credit reports and that 5% of consumers had errors that could result in less favorable loan terms.)

[40] Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55 (2013).

[41] Kate Crawford, *Think Again: Big Data*, FOREIGN POL'Y (May 9, 2013), *available at* http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data.

engage online, or because they are less profitable customers or important constituents and therefore less interesting as targets of observation. Not only will the quality of individual records of members of these groups be poorer as a consequence, but these groups as a whole will also be less well represented in datasets, skewing conclusions that may be drawn from an analysis of the data.

Crawford points to Street Bump, an application for Boston residents that takes advantage of accelerometers built into smart phones to detect when drivers ride over potholes (sudden movement that suggests broken road automatically prompts the phone to report the location to the city).[42] While Crawford praises the cleverness and cost-effectiveness of this passive approach to reporting road problems, she rightly warns that whatever information the city receives from this application will be biased by the uneven distribution of smartphones across populations in different parts of the city.[43] In particular, systematic differences in smartphone ownership will very likely result in the underreporting of road problems in the poorer communities where protected groups disproportionately congregate. If the city were to rely on this data to determine where it should direct its resources, it would only further underserve these communities. Indeed, the city would discriminate against those who lack the capacity to report problems as effectively as wealthier residents with smartphones.[44]

A similar dynamic could easily apply in an employment context if members of protected classes are unable to report their interest in and qualification for jobs listed online as easily or effectively as others due to systematic differences in Internet access. The EEOC has established a program called "Eradicating Racism and Colorism from Employment" (E-RACE) that is, at least in part, aimed at preventing this sort of discrimination from occurring due to an employer's desire for high-tech hiring, such as video resumes.[45] The program clearly attempts to lower the

[42] *Id.*

[43] *Id.*

[44] This is, of course a more general problem with representative democracy. For a host of reasons, the views and interests of the poor are relatively less well represented in the political process. *See, e.g.*, Larry M Bartels, *Economic Inequality and Political Representation*, in THE UNSUSTAINABLE AMERICAN STATE 167, (Lawrence Jacobs & Desmond King, ed. 2009); MARTIN GILENS, AFFLUENCE AND INFLUENCE: ECONOMIC INEQUALITY AND POLITICAL POWER IN AMERICA, (2012). The worry here, as expressed by Crawford, is that, for all its apparent promise, data mining may further obfuscate or legitimate these dynamics rather than overcome them.

[45] Equal Employment Opportunity Commission, *Why Do We Need E-RACE?*, *at* http://www1.eeoc.gov/eeoc/initiatives/e-race/why_e-race.cfm (last visited August 1, 2013). Due to the so-called "digital divide," communities underserved by broadband access rely heavily on mobile phones for internet access and thus often have trouble even uploading and updating traditional resumes. KATHRYN ZICKUHR & AARON SMITH, DIGITAL

barriers that would disproportionately burden applicants who belong to a protected class, but, in so doing, it also ensures that employers do not develop an inaccurate impression of the incidence of qualified and interested candidates from these communities. If employers were to rely on these tallies to direct the focus of their recruiting efforts, for example, any count affected by a reporting bias could have adverse consequences for specific populations systematically under-represented in the dataset. Employers would deny equal attention to those who reside in areas incorrectly pegged as having a relatively lower concentration of qualified candidates.

But these two examples only discuss decisions that depend on raw tallies, rather than datasets from which decisionmakers want to draw generalizations and generate predictions. Additional and even more severe risks may reside in the systematic omission of members of protected classes from such datasets. Data mining is especially sensitive to statistical bias because data mining helps to discover patterns that organizations tend to treat as generalizable findings even though the analyzed data only includes a partial sample from a circumscribed period. To ensure that data mining reveals patterns that obtain for more than the particular sample under analysis, the sample must share the same probability distribution as the data that would be gathered from *all* cases across both time and population.[46] In other words, the sample must be proportionally representative of the entire population, even though the sample, by definition, does not include every case.

If a sample includes a disproportionate representation of a particular class (more or less than its actual incidence in the overall population), the results of an analysis of that sample may skew in favor of or against the over- or under-represented class. While the representativeness of the data is often simply assumed, this assumption is rarely justified, and is "perhaps more often incorrect than correct."[47] Data gathered for routine business purposes tend to lack the rigor of social scientific data collection.[48] As

DIFFERENCES (2012), http://www.pewinternet.org/2012/04/13/digital-differences/ ("Among smartphone owners, young adults, minorities, those with no college experience, and those with lower household income levels are more likely than other groups to say that their phone is their main source of internet access."). This will also lead to reporting bias.

[46] Data mining scholars have devised ways to address this known problem, but applying these techniques is far from trivial. *See* Sinno Jialin Pan & Qiang Yang, *A Survey on Transfer Learning*, 22 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 1345 (2010).

[47] Hand, *supra* note 22, at 7.

[48] David Lazer, *Big Data and Cloning Headless Frogs*, COMPLEXITY AND SOCIAL NETWORKS BLOG, (February 16, 2014), http://blogs.iq.harvard.edu/netgov/2014/02/big_data_and_cloning_headless.html.

Lerman points out, "[b]usinesses may ignore or undervalue the preferences and behaviors of consumers who do not shop in ways that big data tools can easily capture, aggregate, and analyze."[49]

Even where a company performs an analysis of the data from its entire population of employees—avoiding the apparent problem of even having to select a sample—the organization must assume that its future applicant pool will have the same degree of variance as its current employee base. The fact that organizations tend to perform such analyses in order to *change* the composition of their employee base should put the validity of this assumption into immediate doubt. The potential effect of this assumption is the future mistreatment of individuals predicted to behave in accordance with the skewed findings derived from the biased sample. Worse, these results may lead to decision procedures that limit the future contact an organization will have with specific groups, skewing still further the sample upon which subsequent analyses will be performed.[50] Limiting contact with specific populations on the basis of unsound generalizations may deny members of those populations the opportunity to prove that they buck the apparent trend.

Over-representation in a dataset can also lead to disproportionately high adverse outcomes for members of protected classes. Consider an example from the workplace: managers may devote disproportionate attention to monitoring the activities of employees who belong to a protected class and consequently observe mistakes and transgressions at systematically higher rates than others, in part because they fail to subject others who behave similarly to the same degree of scrutiny. Not only does this provide managers with justification for their prejudicial suspicions, but it also generates evidence that overstates the relative incidence of offenses by members of these groups. Where subsequent managers who hold no such prejudicial suspicions cannot observe everyone equally, they may rely on this evidence to make predictions about where to focus their attentions in the future and thus further increase the disproportionate scrutiny that they place on protected classes.

C. Feature Selection

[49] Lerman, *supra* note 40, at 59.

[50] Practitioners, particularly those involved in credit scoring, are well aware that they do not know how the person purposefully passed over would have behaved if he had been given the opportunity. Practitioners have developed methods to correct for this bias (which, in the case of credit scoring, they refer to as reject inference). *See, e.g.*, Jonathan Crook & John Banasik, *Does Reject Inference Really Improve the Performance of Application Scoring Models?*, 28 J. BANKING & FINANCE 857 (2004).

Organizations—and the data miners that work for them—also make choices about what attributes they observe and what they subsequently fold into their analyses. Data miners refer to the process of settling on the specific string of input variables as "feature selection."[51] These decisions can also have serious implications for the treatment of protected classes, if those factors that better account for pertinent statistical variation among members of a protected class are not well represented in the set of selected features.[52] Members of protected classes may find that they are subject to systematically less accurate classifications or predictions because the details necessary to achieve equally accurate determinations reside at a level of granularity and coverage that the features fail to achieve.

This problem stems from the fact that data are by necessity reductive representations of an infinitely more specific real-world object or phenomenon.[53] These representations may fail to capture enough detail to allow for the discovery of crucial points of contrast. Increasing the resolution and range of the analysis may still fail to capture the mechanisms that account for different outcomes because such mechanisms may not lend themselves to exhaustive or effective representation in the data, if such representations even exist. As Toon Calders and Indrė Žliobaitė explain, "it is often impossible to collect all the attributes of a subject or take all the environmental factors into account with a model."[54] While these limitations lend credence to the argument that data can never fully encompass the full complexity of the individuals they seek to represent, they do not reveal the inherent inadequacy of representation as such.

At issue, really, is the coarseness and comprehensiveness of the criteria that permit statistical discrimination and the uneven rates at which different groups happen to be subject to erroneous determinations. Crucially, these erroneous and potentially adverse outcomes are artifacts of statistical reasoning rather than prejudice on the part of decision-makers or bias in the composition of the dataset. As Frederick Schauer explains, decision-makers that rely on statistically sound but nonuniversal generalizations "are being simultaneously rational and unfair"[55] because

[51] FEATURE EXTRACTION, CONSTRUCTION AND SELECTION, (Huan Liu & Hiroshi Motoda, eds., 1998).

[52] Toon Calders & Indrė Žliobaitė, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 25, at 43, 46 ("[T]he selection of attributes by which people are described in [a] database may be incomplete.").

[53] Annamaria Carusi, *Data as Representation: Beyond Anonymity in E-Research Ethics*, 1 INT'L J. INTERNET RESEARCH ETHICS 37 2008, *available at* http://ijire.net/issue_1.1/ijire_1.1_carusi.pdf.

[54] Calders & Žliobaitė, *supra* note 52, at 47.

[55] FREDERICK SCHAUER, PROFILES, PROBABILITIES, AND STEREOTYPES 3 (2006).

certain individuals are "actuarially saddled"[56] by statistically sound inferences that are nevertheless inaccurate. Obtaining information that is sufficiently rich to permit precise distinctions can be expensive. Even marginal improvements in accuracy may come at significant practical costs, and may justify a less granular and encompassing analysis.[57]

To take an obvious example, hiring decisions that consider credentials tend to assign enormous weight to the reputation of the college or university from which an applicant has graduated, despite the fact that such credentials may communicate very little about the applicant's job-related skills and competencies.[58] If equally competent members of protected classes happen to graduate from these colleges or universities at disproportionately low rates, decisions that turn on the credentials conferred by these schools, rather than some set of more specific qualities that more accurately sort individuals, will incorrectly and systematically discount these individuals. Even if employers have rational incentive to look beyond credentials and focus on criteria that allow for more precise—and therefore more accurate—determinations, they may continue to favor credentials because they communicate pertinent information at no cost to the employer.[59]

Such was the reasoning that seemed to undergird the practice known as redlining, where financial institutions employed especially general criteria to draw distinctions between sub-populations (i.e., the neighborhood in which individuals happen to reside) because such criteria were easily accessible and cheaply obtained, despite the fact that such distinctions failed to capture significant variation within each sub-population that would have resulted in a different assessment for certain members of these groups. Decision-makers were willing to tolerate higher rates of erroneous determinations for certain groups because the benefits derived from gaining access to data that provided the necessary granularity for more accurate determinations did not seem to justify the costs. Of course, it may be no

[56] *Id.* at 5. Insurance offers the most obvious example of this: the rate that a person pays for car insurance, for instance, is determined by the way other people with similar characteristics happen to drive, even if the person happens to be a better driver than those who resemble him on the statistically pertinent dimensions.

[57] *Id.* at 54 ("[O]btaining information is costly, so it is morally justified, all things considered, to treat people on the basis of statistical generalizations even though one knows that, in effect, this will mean that one will treat some people in ways, for better or worse, that they do not deserve to be treated."). *See also* Brian Dalessandro, Claudia Perlich, & Troy Raeder, *Bigger Is Better, but at What Cost? Estimating the Economic Value of Incremental Data Assets*, 1 BIG DATA 87 (2014).

[58] Matt Ritchel, *How Big Data Is Playing Recruiter for Specialized Workers*, N.Y. TIMES, at BU1 (April 28, 2014).

[59] As one commentator has put it in contemplating data-driven hiring, "Big Data has its own bias. . . . You measure what you can measure." *Id.*

coincidence that such cost-benefit analyses seemed to justify treating groups composed disproportionately of members of protected classes to systematically less accurate determinations, and why redlining is generally seen as illegal.[60]

Cases of so-called rational racism are really just a special instance of this more general phenomenon—one in which race happens to be taken into consideration explicitly. In such cases, decision-makers take membership in a protected class into account, even if they hold no prejudicial views, because such membership seems to communicate relevant information that would be difficult or impossible to obtain otherwise. The persistence of distasteful forms of discrimination may be the result of a lack of information, rather than a continued taste for discrimination.[61] Lior Strahilevitz has argued, for instance, that when employers lack access to criminal records, they may consider race in assessing applicants because statistical differences in the rates at which members of different racial groups have been convicted of crimes provides some basis for sorting applicants according to their relative likelihood of having criminal records. In other words, employers fall back on more immediately available and coarse proxies when they cannot access more specific or verified information.[62] Of course, as Strahilevitz points out, race is a highly imperfect basis upon which to predict an individual's criminal record, despite whatever differences may exist in the rates at which members of different racial groups have been convicted of crimes, because it is too coarse.[63]

D. Proxies

Cases of decision-making that do not artificially introduce

[60] *See* Nationwide Mut. Ins. Co. v. Cisneros, 52 F.3d 1351, 1359 (6th Cir. 1995); NAACP. v. Am. Family Mut. Ins. Co., 978 F.2d 287, 300 (7th Cir. 1992).

[61] Andrea Romei and Salvatore Ruggieri, *Discrimination Data Analysis: A Multi-Disciplinary Bibliography*, in DISCRIMINATION IN THE INFORMATION SOCIETY, *supra* note 25, at 109; *see also* Calders and Žliobaitė *supra* note 52, at 53 ("Inequality may exist between demographic groups even when economic agents (consumers, workers, employers) are rational and non-prejudiced, as stereotypes may be based on the discriminated group's average behavior.").

[62] Lior Jacob Strahilevitz, *Privacy Versus Antidiscrimination*, 75 U. CHI. L. REV. 363 (2008). Of course, this argument assumes that criminal records are relevant to employment, which is often not true. *See* text accompanying *infra* note 152.

[63] *See, infra* Part II.A. The law holds that decision-makers should refrain from considering membership in a protected class even if statistical evidence seems to support certain inferences on that basis. The prohibition does not depend on whether decision-makers can gain (easy or cheap) access to alternative criteria that hold greater predictive value.

discriminatory effects into the data mining process may nevertheless result in systematically less favorable determinations for members of protected classes. Situations of this sort are possible when the criteria that are genuinely relevant in making rational and well-informed decisions also happen to serve as reliable proxies for class membership. In other words, the very same criteria that correctly sort individuals according to their predicted likelihood of excelling at a job—as formalized in some fashion—may also sort individuals according to class membership. In certain cases, there may be an obvious reason for this: just as "mining from historical data may . . . discover traditional prejudices that are endemic in reality (i.e., taste-based discrimination)," so, too, may it "discover patterns of lower performances, skills or capacities of protected-by-law groups."[64] These discoveries reveal the simple fact of inequality, but they also reveal the fact that these are inequalities in which members of protected classes are frequently the groups in the relatively less favorable position. This has rather obvious implications: if features held at a lower rate by members of protected groups nevertheless possess relevance in rendering legitimate decisions, such decisions will necessarily result in systematically less favorable determinations for these individuals.

For example, employers may find, in conferring greater attention and opportunities to employees that they predict will prove most competent at some task, that they subject members of protected groups to consistently disadvantageous treatment because the criteria that determine the attractiveness of employees happen to be held at systematically lower rates by members of these groups.[65] Decision-makers do not necessarily intend this disparate impact because they hold prejudicial beliefs; rather, their reasonable priorities as profit-seekers unintentionally recapitulate the inequality that happens to exist in society. Furthermore, this may occur even if proscribed criteria have been removed from the dataset, the data are free from latent prejudice or bias, the data is especially granular and diverse, and the only goal is to maximize classificatory or predictive accuracy. The problem stems from what researchers call "redundant encodings": cases in which membership in a protected class happens to be encoded in other data.[66] This occurs when a particular piece of data or certain values for that piece of data are highly correlated with membership in specific protected

[64] Romei & Ruggieri, *supra* note 61, at 121.

[65] Faisal Kamiran, Toon Calders, & Mykola Pechenizkiy, *Techniques for Discrimination-Free Predictive Models*, in DISCRIMINATION IN THE INFORMATION SOCIETY, *supra* note 25, at 323, 324.

[66] Cynthia Dwork et al., *Fairness Through Awareness*, (Proceedings of the 3rd Innovations in Theoretical Computer Science Conference 2012). 214, 226 app'x. ("Catalog of Evils").

classes. The fact that these data may hold significant statistical relevance to the decision at hand explains why data mining can result in seemingly discriminatory models even when its only objective is to ensure the greatest possible accuracy for its determinations. If there is a disparate distribution of an attribute, a more precise form of data mining will be more likely to capture it as such. Better data and more features will simply expose the exact extent of inequality.

E. Masking

Data mining could also breathe new life into traditional forms of intentional discrimination because decision-makers with prejudicial views can mask their intentions by exploiting each of the mechanisms enumerated above. Stated simply, any form of discrimination that happens unintentionally can be orchestrated intentionally as well. For instance, decision-makers could knowingly and purposefully bias the collection of data to ensure that mining suggests rules that are less favorable to members of protected classes.[67] They could likewise attempt to preserve the known effects of prejudice in prior decision-making by insisting that such decisions constitute a reliable and impartial set of examples from which to induce a decision-making rule. And decision-makers could intentionally rely on features that only permit coarse-grain distinction-making—distinctions that result in avoidable and higher rates of erroneous determinations for members of a protected class. In denying themselves finer-grained detail, decision-makers would be able to justify writing-off entire groups composed disproportionately of members of protected classes. A form of digital redlining, this decision masks efforts to engage in intentional discrimination by abstracting to a level of analysis that fails to capture lower level variations that might otherwise make certain members of protected classes into more attractive candidates. Here, prejudice rather than some legitimate business reason (e.g., cost) motivates decision-makers to intentionally restrict the particularity of their decision-making to a level that can only paint in avoidably broad strokes, condemning entire groups, composed disproportionately of members of protected classes, to systematically less favorable treatment.

Because data mining holds the potential to infer otherwise unseen attributes, including those traditionally deemed sensitive,[68] it can furnish methods by which to determine indirectly individuals’ membership in protected classes and to unduly discount, penalize, or exclude such people

[67] *Id.*

[68] *See* Solon Barocas, *Leaps and Bounds: Toward a Normative Theory of Inferential Privacy* (in progress).

accordingly. In other words, data mining could grant decision-makers the ability to distinguish and disadvantage members of protected classes without access to explicit information about individuals' class membership. It could instead help to pinpoint reliable proxies for such membership and thus place institutions in the position to automatically sort individuals into their respective class without ever having to learn these facts directly. The most immediate implication is that institutions could employ data mining to circumvent the barriers, both practical and legal, that have helped to withhold from consideration individuals' membership in a protected class.

Additionally, data mining could provide cover for intentional discrimination of this sort because the process would conceal from view that decision-makers had determined and considered the individual's class membership. The worry, then, is not simply that data mining would introduce novel ways for decision-makers to satisfy their taste for illegal discrimination; rather, the worry is that it would mask actual cases of such discrimination.[69] Although scholars, policy-makers, and lawyers have long been aware of the dangers of masking,[70] data mining significantly enhances the capacity to conceal acts of intentional discrimination by finding evermore remote and complex proxies for proscribed criteria.[71]

Intentional discrimination and its masking have so far garnered disproportionate attention in discussions of data mining,[72] often to the exclusion of issues arising from the many forms of unintentional discrimination described above. While data mining certainly introduces novel ways to discriminate intentionally and to conceal those intentions, concerns with this potential should not crowd out careful consideration of the unintentional discrimination that is likely to be more common than the kinds of discrimination that commenters fear could be pursued intentionally.

[69] Data miners who wish to discriminate can do so using relevant or irrelevant criteria. Either way the intent would make the action "masking." If an employer masked using highly relevant data, it is likely that litigation arising from it would be tried under a "mixed-motive" framework, which asks whether the same action would have been taken without the intent to discriminate. *See infra* Part II.A.

[70] Custers, *supra* note 25, at 9-10.

[71] *See* Barocas, *supra* note 68.

[72] *See, e.g.*, Alistair Croll, *Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It*, SOLVE FOR INTERESTING (July 31, 2012) *at* http://solveforinteresting.com/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it/. This post generated significant online chatter immediately upon publication and has become one the canonical texts in the current debate. It has also prompted a number of responses from scholars. *See, e.g.*, Anders Sandberg, *Asking the Right Questions: Big Data and Civil Rights*, PRACTICAL ETHICS (August 16, 2012), *at* http://blog.practicalethics.ox.ac.uk/2012/08/asking-the-right-questions-big-data-and-civil-rights/.

II. TITLE VII LIABILITY FOR DISCRIMINATORY DATA MINING

Current anti-discrimination law is not well equipped to address the various discriminatory features of data mining. This Part analyzes Title VII as a well-developed model of modern anti-discrimination law. Other anti-discrimination laws, such as the Americans with Disabilities Act, will exhibit differences in specific operation, but the main thrust of anti-discrimination law is fairly constant across regimes, and Title VII serves as an illustrative example.[73]

An employer sued under Title VII may be found liable for employment discrimination under one of two theories of liability: disparate treatment and disparate impact. Disparate treatment actually comprises two different strains of discrimination: 1) formal disparate treatment of similarly situated people and 2) intent to discriminate.[74] Disparate impact refers to policies or practices that are facially neutral but have a disproportionately adverse impact on protected classes.[75] Disparate impact is not concerned with the intent or motive for a policy; where it applies, the doctrine first asks whether there is a disparate impact on members of a protected class, then whether there is some business justification for that impact, and finally, whether there were less discriminatory means of achieving the same result.[76]

Liability under Title VII for discriminatory data mining will depend on the particular mechanism by which the inequitable outcomes are generated. This Part explores disparate treatment and disparate impact doctrine and analyzes which decision-points that result in discriminatory data mining could generate liability under each theory.

[73] The biggest difference between the Americans with Disabilities Act and Title VII is the requirement that an employer make "reasonable accommodations" for disabilities. 42 U.S.C. § 12112(b)(5). But some scholars have argued that even this difference is illusory and that accommodations law is functionally similar to Title VII, though worded differently. *See* Samuel R. Bagenstos, "*Rational Discrimination," Accommodation, and the Politics of (Disability) Civil Rights*, 89 VA. L. REV. 825 (2003)(comparing accommodations law to disparate treatment); Christine Jolls, *Antidiscrimination and Accommodation*, 115 HARV. L. REV. 642, 652 (2001)(comparing accommodations law to disparate impact).

[74] Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1351 n.56 (explaining that for historical reasons, disparate treatment became essentially "not disparate impact" and now we rarely notice the two different embedded theories).

[75] Griggs v. Duke Power Co., 401 U.S. 424, 430 (1971).

[76] 42 U.S.C. § 2000e-2(k).

A. Disparate Treatment

Disparate treatment doctrine recognizes liability for both explicit formal classification and intentional discrimination. Formal discrimination, in which membership in a protected class is used as an input to the model, corresponds to an employer classifying employees or potential hires according to membership in a protected class. Because classification itself is a legal harm, irrespective of the effect,[77] the same should be true of using protected class as an input to a system for which the entire purpose is to build a classificatory model.[78] Formal liability does not correspond to any particular discrimination mechanism; it can occur equally well in any of them. The traditional focus of formal discrimination has been the use of rational racism. In traditional contexts, rational racism is considered rational because there are cases in which its users believe it is an accurate, if coarse grained, proxy, or at least the best available one in a given situation. In the world of data mining, though, that need not be the case. Even if membership in a protected class were specified as an input, the eventual model that emerges could see it as the least significant feature. In that case, there would be no discriminatory effect, but there would be a disparate treatment violation, because considering membership in a protected class as a potential proxy is a legal classificatory harm in itself.

The irony is that the use of protected class as an input is usually irrelevant to the outcome in terms of discriminatory effect, at least given a large enough number of input features. The target variable will, in reality, be correlated to the membership in a protected class somewhere between 0% and 100%. If the trait is perfectly uncorrelated, including membership in the protected class as an input will not change the output, and there will be no discriminatory effect.[79] On the other end of the spectrum, where membership in the protected class is perfectly predictive of the target variable, the fact will be redundantly encoded in the other data. The only way using membership in the protected class as an explicit feature will change the outcome is if the information is otherwise not rich enough to detect such membership. Membership in the protected class will show up as relevant to the exact extent it is already redundantly encoded. Given a rich

[77] Jed Rubenfeld, *Affirmative Action*, 107 YALE L.J. 427, 433 (discussing “classificationism”); Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 Harv. L. Rev. 493, 504, 567-68 (2003)(discussing expressive harms).

[78] Membership in a protected class is still a permissible input to a holistic determination when the focus is diversity, but where classification is the goal, such as here, it is not. *See* Grutter v. Bollinger, 539 U.S. 306, 325 (2003)(diversity is a compelling state interest that can survive strict scrutiny).

[79] That is, not counting any expressive harm that might come from classification by protected class.

enough set of features, the chance that such membership is redundantly encoded approaches certainty. Thus, a data mining model with a large number of variables will determine the extent to which membership in a protected class is relevant to the sought-after trait whether or not that information is an input. Formal discrimination therefore has no bearing whatsoever on the outcome of the model. Additionally, by analyzing the data, it would be possible to probabilistically determine membership in that same protected class if an employer did indeed want to know.

To analyze intentional discrimination other than mere formal discrimination, a brief description of disparate treatment doctrine is necessary. A Title VII disparate treatment case will generally proceed under either the *McDonnell-Douglas* burden-shifting scheme or the *Price-Waterhouse* regime, for a "mixed-motive" case.[80] In the *McDonnell-Douglas* framework, the plaintiff who has suffered an adverse employment action has the initial responsibility to establish a prima facie case of discrimination by demonstrating that a similarly situated person who is not a member of a protected class would not have suffered the same fate.[81] This can be shown with circumstantial evidence, such as disparaging remarks made by the employer or procedural irregularities in promotion or hiring that suggest an intent to discriminate; only very rarely will an employer openly admit to discriminatory conduct. If the plaintiff successfully demonstrates that the adverse action treated protected class members differently, then the burden shifts to the defendant-employer to offer a legitimate, non-discriminatory basis for the decision. The defendant need not prove the reason is true; his is only a burden of production.[82] Once the defendant has offered a non-discriminatory alternative, the ultimate burden of persuasion falls to the plaintiff to demonstrate the proffered reason is pretextual.[83]

In the data mining context, this makes masking an easy case as a theoretical matter, no matter which mechanism for discrimination is employed. The fact that it is accomplished algorithmically does not make it less of a disparate treatment violation, as the entire idea of masking is pretextual. In fact, in the traditional, non-data mining context, the word

[80] McDonnell Douglas Corp. v. Green, 411 U.S. 792 (1973); Price Waterhouse v. Hopkins, 490 U.S. 228 (1989).

[81] This is similar to the computer science definition of discrimination. Calders & Žliobaitė, *supra* note 52, at 49. ("A classifier discriminates with respect to a sensitive attribute, e.g. gender, if for two persons which only differ by their gender (and maybe some characteristics irrelevant for the classification problem at hand) that classifier predicts different labels.")

[82] St. Mary's Honor Ctr. v. Hicks, 509 U.S. 502, 507 (1993).

[83] *Id.*

masking has occasionally been used to refer to pretext.[84] Like any disparate treatment case, however, proof will be difficult to come by, something even truer for masking.[85]

The *McDonnell-Douglas* framework operates on a presumption that if the rationale that the employer has given is found to be untrue, the employer must be hiding his "true" discriminatory motive and it is the job of the plaintiff at that point to prove it.[86] Because the focus of the *McDonnell-Douglas* framework is on pretext and coverup, it can only address conscious, willful discrimination.[87] Under the *McDonnell-Douglas* framework, a court must find either that the employer *intended* to discriminate or did not discriminate at all.[88] Thus, unintentional discrimination will not lead to liability.

A case can also be tried under the mixed-motive framework, first recognized in *Price-Waterhouse v. Hopkins*,[89] and most recently modified by *Desert Palace v Costa*.[90] In the mixed-motive framework, a plaintiff

[84] See Custers, supra note 25; Megan Whitehill, *Better Safe Than Subjective: The Problematic Intersection of Pre-Hire Social Networking Checks and Title VII Employment Discrimination*, 85 TEMP. L. REV. 229, 250 (2012)(referring to "masking pretext" in the third stage of McDonnell-Douglas framework); Keyes v. Sec'y of the Navy, 853 F.2d 1016, 1026 (1st Cir. 1988)(plaintiff's burden to show that the proffered reasons for hiring an alternative were "pretexts aimed at masking sex or race discrimination")

[85] *See,* Part I.E, *supra.* This is a familiar problem to anti-discrimination law, and is often cited as one of the rationales for disparate impact liability in the first place – to "smoke out" intentional invidious discrimination. *See infra* Part III.B.

[86] McDonnell Douglas Corp. v. Green, 411 U.S. 792, 805 (1973)(The plaintiff "must be given a full and fair opportunity to demonstrate by competent evidence that the presumptively valid reasons for his rejection were in fact a coverup for a racially discriminatory decision.") While, as a theoretical matter, the plaintiff must prove that the employer's reason was a pretext for discrimination specifically, the Supreme Court has held that a jury can reasonably find that the fact that an employer had only a pretextual reason to fall back on is itself circumstantial evidence of discrimination. *Hicks*, 509 U.S. at 511 ("The factfinder's disbelief of the reasons put forward by the defendant (particularly if disbelief is accompanied by a suspicion of mendacity) may, together with the elements of the prima facie case, suffice to show intentional discrimination.").

[87] Tristin K. Green, *Discrimination in Workplace Dynamics: Toward a Structural Account of Disparate Treatment Theory*, 38 HARV. C.R.-C.L. L. REV. 91, 91 (2003); Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458, 458 (2001); *see also* Melissa Hart, *Subjective Decisionmaking and Unconscious Discrimination*, 56 ALA. L. REV. 741, 749-50 (2005)(critiquing the courts' requirement of proving employer "dishonesty," but suggesting that, absent this requirement, Title VII could handle unconscious discrimination without altering the law).

[88] Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161, 1170 (1995).

[89] 490 U.S. 228 (1989).

[90] 539 U.S. 90 (2003).

need not demonstrate that the employer's non-discriminatory rationale was pretextual, but merely that discrimination was a "motivating factor" in the adverse employment action.[91] As a practical matter, this means that the plaintiff must show that the same action would not have been taken absent the discriminatory motive.[92] As several commentators have pointed out, motive and intent are not necessarily synonymous.[93] Motive could be read more broadly to include unconscious discrimination, including anything that influences a person to act, such as emotions or desires.[94] Nonetheless, courts have conflated the meanings of motive and intent such that the phrase "motive or intent" has come to refer only to conscious choices.[95] Thus, while most individual decisionmaking probably belongs in a mixed-motive framework, as each decision a person makes comprises a complicated mix of motivations,[96] the mixed-motive framework will be no better than the pretext framework at addressing bias that occurs absent conscious intent.[97]

Except for masking, discriminatory data mining is by stipulation unintentional. Unintentional disparate treatment is not a problem that is new to data mining. A vast scholarly literature has developed regarding the law's treatment of unconscious, implicit bias.[98] Such treatment can occur when an

[91] 42 U.S.C. § 2000e-2(m) (2000); *Desert Palace*, 539 U.S. at 101 ("In order to obtain [a mixed-motive jury instruction], a plaintiff need only present sufficient evidence for a reasonable jury to conclude, by a preponderance of the evidence, that 'race, color, religion, sex, or national origin was a motivating factor for any employment practice.'"

[92] Charles A. Sullivan, *Disparate Impact: Looking Past the* Desert Palace *Mirage*, 47 WM. & MARY L. REV. 911, 914-916 & n.20 (2005); see also D. Don Welch, *Removing Discriminatory Barriers: Basing Disparate Treatment Analysis on Motive Rather Than Intent*, 60 S. CAL. L. REV. 733, 740 (1987); Krieger, supra note 88, at 1170-72.

[93] Sullivan, *supra* note 92, at 915; Krieger, *supra* note 88, at 1243.

[94] Sullivan, *supra* note 92, at 915 n.18 (quoting OXFORD ENGLISH DICTIONARY 698 (1933)); Krieger, *supra* note 88, at 1243.

[95] Sullivan, *supra* note 92, at 914-916 & n.20.

[96] Amy L. Wax, *Discrimination as Accident*, 74 IND. L.J. 1129, 1149 & n.21 (1999); Krieger, *supra* note 88 at 1223. In fact, after the Supreme Court decided *Desert Palace*, many scholars thought that it *had* effectively overruled the *McDonnell-Douglas* framework, forcing all disparate treatment cases into a mixed-motive framework. *See, e.g.* Sullivan, *supra* note 92, at 933-36 (discussing the then-emerging scholarly consensus). This has not played out so far, with courts and scholars split on the matter. *See, e.g.*, Kendall D. Isaac, *Is It "A" or Is It "The"? Deciphering the Motivating-Factor Standard in Employment Discrimination and Retaliation Cases*, 1 TEX. A&M L. REV. 55, 74 (2013)("*McDonnell Douglas* has never been overruled and remains widely utilized."); Barrett S. Moore, *Shifting the Burden: Genuine Disputes and Employment Discrimination Standards of Proof*, 35 U. ARK. LITTLE ROCK L. REV. 113, 128 & n.146 (2012)(noting a circuit split on the issue).

[97] Krieger, *supra* note 88, at 1182-83.

[98] *See, e.g.* Christine Jolls & Cass R. Sunstein, *The Law of Implicit Bias*, 94 CAL. L. REV. 969, 978 n.45 (2006)(collecting sources); Linda Hamilton Krieger & Susan T. Fiske,

employer has internalized some racial stereotype and applies it, or without realizing it, monitors an employee more closely until he finds a violation.[99] The employee is clearly treated differently, but it is not intentional, and the employer is unaware of it. As Samuel Bagenstos summarized, at this point, "it may be difficult, if not impossible, for a court to go back and reconstruct the numerous biased evaluations and perceptions that ultimately resulted in an adverse employment decision."[100] Within the scholarly literature, there is "surprising unanimity" that the law does not adequately address unconscious disparate treatment.[101]

There are a few possible ways to analogize discriminatory data mining to unintentional disparate treatment in the traditional context, based on where one believes the "treatment" lies. Either the disparate treatment occurs at the decision to apply a predictive model that will treat members of a protected class differently, or it occurs when the disparate result of the model is used in the ultimate hiring decision. In the first scenario, the intent at issue is the decision to apply a predictive model with known disparate impact. In the second, the disparate treatment occurs if, after the employer sees the disparate result, he proceeds anyway. If the employer continues *because* he liked the discrimination produced in either scenario, then intent is clear. If not, then this just devolves into a standard disparate impact scenario. Where most disparate impact cases involve placement exams, such an exam is just another form of a sorting algorithm, and the decision to apply it after noticing the discriminatory effect does not lead to disparate treatment liability.[102]

Another possibility is to imagine the *model* as the decisionmaker exhibiting implicit bias. That is, because of biases hidden to the predictive model such as non-representative data or mislabeled examples, it reaches a

Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate Treatment, 94 CAL. L. REV. 997, 1003 n.21 (2006)(same).

[99] This example can be ported directly to data mining as over-representation in data collection. *See supra* Part I.B.2.

[100] Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CAL. L. REV. 1, 9 (2006).

[101] Sullivan, *supra* note 92, at 1000. There is, however, no general agreement whether the law should treat such discrimination as disparate treatment or disparate impact. *Compare* Krieger, *supra* note 88, at 1231 (explaining that because the bias causes employers to *treat* people differently, it should be considered a disparate treatment violation) *with* Sullivan, *supra* at 969-71 (arguing that the purpose of disparate impact is a catchall provision to address those types of bias that disparate treatment cannot reach). This disagreement is important and even more pronounced in the case of data mining. *See infra* Part III. For now, we assume each case can be analyzed separately.

[102] In fact, after *Ricci v. DeStefano*, 557 U.S. 557 (2009), deciding *not* to apply it after noticing the discriminatory effect may give rise to a disparate treatment claim in the other direction.

discriminatory result. This analogy turns every mechanism except proxy discrimination into the equivalent of implicit bias exhibited by individual decisionmakers. The effect of bias is one factor among the many different factors that go into the model-controlled decision, just like in an individual's adverse employment decision.[103] Would a more expansive definition of motive fix this scenario?

Because the doctrine is focused on *human* decisionmakers as discriminators, the answer is no. Even if disparate treatment doctrine could capture unintentional discrimination, it would only address such discrimination stemming from human bias. The person who came up with the idea for Street Bump devised a system that suffers from reporting bias,[104] but it was not because he or she was implicitly employing some racial stereotype. Rather, it was simply inattentiveness to problems with the sampling frame. This is not to say that his or her own bias had nothing to do with it—the person likely owned a smartphone and thus did not think about the people who do not—but no one would say that it was even implicit bias against protected classes that motivated the decision, even under the expansive definition of the word "motive."[105]

The only possible analogy relevant to disparate treatment, then, is to those mechanisms of unintentional discrimination that reflect a real person's bias—something like LinkedIn's Talent Match recommendation engine, which relies on potentially prejudiced human assessments of employees.[106] As a general rule, an employer may not avoid disparate treatment liability by encoding third-party preferences as a rationale for a hiring decision.[107] But once again, to find liability under current doctrine, the employer would likely both have to at least know that this is the specific failure mechanism of the model and choose it based on this fact.

There is one other interesting question regarding disparate treatment doctrine: whether the intent standard includes knowledge. This is not a problem that arises often when a human is making a single employment determination. Assuming disparate treatment occurs in a given case, it is generally either intended or unconscious. What would it mean in the individual case to have an employer *know* that he was treating an employee

[103] Krieger, *supra* note 88; Bagenstos, *supra* note 100, at 9.

[104] *See* Crawford, *supra* note 41 and accompanying text.

[105] Of course, the very presumption of a design's neutrality is itself a bias that may work against certain people. *See* Langdon Winner, *Do Artifacts Have Politics?* 109 Daedalus 121, 125 (1980). But, as this is a second-order effect, we need not address it here.

[106] Woods, *supra* note 37

[107] *See* 29 C.F.R. § 1604.2(a)(1)(iii)(stating the EEOC's position that "the preferences of coworkers, the employer, clients or customers" cannot be used to justify disparate treatment); *see also* Fernandez v. Wynn Oil Co., 653 F.2d 1273, 1276-77 (9th Cir. 1981); Diaz v. Pan Am. World Airways, Inc., 442 F.2d 385, 389 (5th Cir. 1971).

differently, but still take the action he had always planned to take without *intent* to treat the employee differently? It seems like an impossible line to draw.

With data mining, though, unlike unconscious bias, it is possible to audit the resulting model and inform an employer that he is or will be treating individuals differently before he does so. If an employer *intends* to employ the model, but *knows* it will produce a disparate impact, does she intend to discriminate? This is a more realistic parsing of intent and knowledge than in the case on an individual, non-systematic employment decision. Under current doctrine, there is neither pretext nor motive, and throughout civil and criminal law, "knowledge" and "intent" are considered distinct states of mind, so there would likely be no liability. On the other hand, some courts have used knowledge of discrimination as evidence to find intent. And while the statute's language only covers intentional discrimination,[108] a broad definition of intent could include knowledge or substantial certainty of the result.[109] Because the situation has not come up often, the extent of the "intent" required is as yet unknown.[110]

In sum, aside from rational racism and masking (with some difficulties), it does not appear that disparate treatment doctrine does much to regulate discriminatory data mining.

B. Disparate Impact

Where there is no intent, disparate impact doctrine should be better suited to the job. In a disparate impact case, a plaintiff must show that a

[108] 42 U.S.C. § 2000e-2(h).

[109] *See, e.g.,* Restatement (Second) of Torts § 8A cmt. b (1965)("Intent is not . . . limited to consequences which are desired. If the actor knows that the consequences are certain, or substantially certain, to result from his act, and still goes ahead, he is treated by the law as if he had in fact desired to produce the result.") *Id.*

[110] Determining that a model is discriminatory is also like trying and failing to validate a test under disparate impact doctrine. *See infra* Part II.B. If a test fails validation, the employer using it would know that he is discriminating if he applies it, but that does not imply that he is subject to disparate treatment liability. Nonetheless, validation is part of the business necessity defense, and that defense is not available against disparate treatment claims, so the analysis does not necessarily have the same result. 42 U.S.C. § 2000-e-2(k)(2). One commentator has argued that including knowledge as a state of mind leading to disparate treatment liability would effectively collapse disparate impact and disparate treatment because it would conflate intent and effect. Jessie Allen, *A Possible Remedy for Unthinking Discrimination*, 61 BROOK. L. REV. 1299, 1314 (1995). But others still have noted that with respect to knowledge, a claim is still about the *treatment* of an individual, not the incidental disparate impact of a neutral policy. Carin Ann Clauss, *Comparable Worth—The Theory, Its Legal Foundation, and the Feasibility of Implementation*, 20 U. MICH. J.L. REFORM 7, 62 (1986).

particular facially neutral employment practice causes a disparate impact with respect to a protected class.[111] After that showing, the defendant-employer may "demonstrate that the challenged practice is job related for the position in question and consistent with business necessity."[112] Finally, if the defendant makes a successful showing to that effect, the plaintiff may show that the employer could have used an "alternative employment practice" with less discriminatory results.[113]

The statute is unclear as to the required showing for essentially every single element of a disparate impact claim. The first unclear point is how much disparate impact is needed to make out a prima facie case.[114] The EEOC, charged with enforcing Title VII's mandate, has created the so-called "four-fifths rule" as a presumption of adverse impact: "A selection rate for any race, sex, or ethnic group which is less than four-fifths . . . of the rate for the group with the highest rate will generally be regarded . . . as evidence of adverse impact."[115] The Uniform Guidelines on Employment Selection Procedures (Guidelines) also state, however, that smaller differences can constitute adverse impact and greater differences may not, depending on circumstances, so the four-fifths rule is truly just a guideline.[116] For the purposes of this Part, it is worthwhile to just assume that the discriminatory effects are prominent enough to establish disparate impact as an initial matter.[117]

The next step in the litigation is the "business necessity" defense. This defense is, in a very real sense, the crux of disparate impact analysis, weighing Title VII's competing goals of limiting the effects of discrimination while allowing employers discretion to advance important business goals. Established in *Griggs v. Duke Power Co.*[118] alongside disparate impact doctrine itself, the business necessity defense had several different articulations in the same case: "A challenged employment practice must be 'shown to be related to job performance,' have a 'manifest relationship to the employment in question,' be 'demonstrably a reasonable

[111] 42 U.S.C. § 2000e-2(k)(1)(A).

[112] *Id.*

[113] *Id.*

[114] The statute does not define the requirement and Supreme Court has never addressed the issue. *See, e.g.,* Sullivan, *supra* note 92, at 954 & nn. 153-54. For a brief discussion on the issue, see Pamela L. Perry, *Two Faces of Disparate Impact Discrimination*, 59 FORDHAM L. REV. 523, 570-74 (1991).

[115] Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607.4(D)[hereinafter "Guidelines"].

[116] *Id.*

[117] We will return to this when discussing the need to grapple with substantive fairness. *See infra* Part III.B.

[118] 401 U.S. 424 (1971).

measure of job performance,' bear some 'relationship to job-performance ability,' and/or 'must measure the person for the job and not the person in the abstract.'" [119] The Court was not clear on what, if any, difference existed between job-relatedness and business necessity, at one point seeming to use the terms interchangeably: "The touchstone is business necessity. If an employment practice which operates to exclude Negroes cannot be shown to be related to job performance, the practice is prohibited."[120] The focus of the Court was clearly on future job performance, and the term "job-related" has come to mean a practice that is predictive of job performance.[121] Because the definitions of job-relatedness and business necessity have never been clear, there is some disparity in the courts as to how to apply the doctrine and find the appropriate balance.[122]

Originally, the defense seemed to apply narrowly. In *Griggs*, Duke Power had instituted new requirements of a high school diploma and success on a "general intelligence" test in order to be hired for its divisions where previously only white workers had been employed, and no such requirements where it had previously hired black employees. The Court ruled that the new requirements were not a business necessity, because "employees who have not completed high school or taken the tests have continued to perform satisfactorily and make progress in departments for which the high school and test criteria are now used" and the requirements were implemented without any study of their future effect.[123] The Court also rejected the argument that the requirements would improve the overall quality of the workforce.[124]

By 1979, the Court began treating business necessity as a much looser standard.[125] In *New York City Transit Authority v. Beazer*,[126] the transit authority had implemented a rule barring drug users from employment, including current users of methadone, otherwise known as *recovering* heroin addicts. In dicta, the Court stated that a "narcotics rule," which "significantly serves" the "legitimate employment goals of safety and

[119] Linda Lye, Comment, *Title VII's Tangled Tale: The Erosion and Confusion of Disparate Impact and the Business Necessity Defense*, 19 BERKELEY J. EMP. & LAB. L. 315, 321 (1998) (quoting Griggs v. Duke Power Co., 401 U.S. 424, 431-36 (1971)).

[120] *Griggs,* 401 U.S. at 431; *see also* Lye, *supra* note 119, at 321;

[121] Lye, *supra* note 119, at 355 & n.206.

[122] *Id.* at 319-20, 348-53; Amy L. Wax, *Disparate Impact Realism*, 53 WM & MARY L. REV. 621, 633-34 (2011).

[123] *Griggs* 401 U.S. at 431-32.

[124] Id.

[125] *See* Lye, *supra* note 119, at 328; Nicole J. DeSario, *Reconceptualizing Meritocracy: The Decline of Disparate Impact Discrimination Law*, 38 HARV. C.R.-C.L. L. REV. 479, 495-96 (2003).

[126] 440 U.S. 568 (1979).

efficiency" was "assuredly" job-related.[127] This was the entire analysis of the business necessity defense in the case. Moreover, the rationale was acceptable as applied to the entire transit authority, despite seventy-five percent of the jobs presenting no safety concern whatsoever.[128] Ten years later, the Court made business necessity doctrine even more defendant-friendly in *Wards Cove Packing Co. v. Antonio*.[129] On the substance, after *Wards Cove*, business necessity meant a court should engage in a "a reasoned review of the employer's justification for his use of the challenged practice . . . . [T]here is no requirement that the challenged practice be 'essential' or 'indispensable' to the employer's business for it to pass muster."[130] The Court also reallocated the burden to plaintiffs to prove that business necessity was lacking. The Court even referred to the defense as the "business justification" phase of the analysis, rather than business necessity.[131] The *Wards Cove* Court went so far that Congress directly addressed the decision in the Civil Rights Act of 1991, which codified disparate impact and reset the standards to the day before *Wards Cove* as decided.[132]

Because the substantive standards for job-relatedness or business necessity were uncertain before *Wards Cove*, however, the confusion persisted.[133] After the passage of the 1991 Act, both sides—civil rights groups and the Bush administration, proponents of a strong and weak business necessity defense respectively—declared victory.[134] Until *Wards Cove*, all the Court's language loosening the standards for the defense was arguably in dicta or in plurality opinions,[135] and after passage of the Act, the civil rights groups, believing that *Wards Cove* simply hijacked the *Griggs* standard, claimed that strict business necessity was again law of the land.[136]

[127] *Id.* at 587 & n.31.

[128] *Id.*

[129] 490 U.S. 642 (1989).

[130] *Id.* at 659.

[131] *Id.*

[132] 42 U.S.C. § 2000e-2(k)(1)(C).

[133] Legislative history was no help either. The sole piece of legislative history is an "interpretive memorandum" that specifies that the standards were to revert to before *Wards Cove*, coupled with an explicit instruction in the Act to ignore any other legislative history regarding business necessity. Susan S. Grover, *The Business Necessity Defense in Disparate Impact Discrimination Cases*, 30 GA. L. REV. 387, 392 (1996).

[134] Andrew C. Spiropoulos, *Defining the Business Necessity Defense to the Disparate Impact Cause of Action: Finding the Golden Mean*, 74 N.C. L. Rev. 1479, 1484 (1996).

[135] In *Watson v. Fort Worth Bank & Trust*, the culmination of pre-*Wards Cove* business necessity jurisprudence, Justice O'Connor's plurality opinion read the defendant's burden as "producing evidence that its employment practices are based on legitimate business reasons." 487 U.S. 977, 998 (1988).

[136] *Id*. at 1504-16.

The Bush administration, however, saw *Wards Cove* as a "natural result of the Court's continuing attempt over many years to construct a workable structure for the disparate impact cause of action."[137] The administration argued that the 1991 Act merely overturned the *Wards Cove* holding as to the allocation of the burden of proof, but not as to the substance of the business necessity defense, for which the act is ambiguous.[138]

Since then, courts have recognized that business necessity lies somewhere in the middle of the extremes and generally state the standard accordingly. Some courts require that the trait bear a "manifest relationship"[139] to the employment in question or that the trait be "significantly correlated" to job performance.[140] The Third Circuit was briefly an outlier, holding "that hiring criteria must effectively measure the 'minimum qualifications for successful performance of the job in question,'" to distinguish between "business necessity" and "business convenience or some weaker term."[141] This tougher standard would, as a practical matter, ban general aptitude tests with any disparate impact because a particular cutoff score could not be shown to divide those able to do the work from those completely unable.[142] Other unmeasured skills and abilities could theoretically compensate for the lower score on an aptitude test, rendering a certain minimum score not "necessary" if it does not measure minimum qualifications.[143] But, in a later case, the Third Circuit recognized that Title VII does not require an employer to choose someone "less qualified" (as opposed to unqualified) in the name of non-discrimination, and noted that aptitude tests "are legitimate and useful hiring tools so long as they accurately measure a person's qualifications."[144] The court concluded: "Putting these standards together, then, we require

[137] *Id.* at 1516.

[138] *Id.* at 1516-20.

[139] *See, e.g.,* Gallagher v. Magner, 619 F.3d 823, 834 (8th Cir. 2010); Anderson v. Westinghouse Savannah River Co., 406 F.3d 248, 265 (4th Cir. 2005).

[140] Gulino v. New York State Educ. Dep't, 460 F.3d 361, 383 (2d Cir. 2006)("significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which candidates are being evaluated" (quoting *Albemarle,* 422 U.S. at 431 (internal quotation marks omitted))).

[141] El v. Se. Pennsylvania Transp. Auth. (SEPTA), 479 F.3d 232, 242 (3d Cir. 2007) (quoting Lanning v. Se. Pennsylvania Transp. Auth. (SEPTA), 181 F.3d 478, 481 (3d Cir. 1999)).

[142] Michael T. Kirkpatrick, *Employment Testing: Trends and Tactics*, 10 EMP. RTS. & EMP. POL'Y J. 623, 633 (2006).

[143] *Id.* Note, though, that this is similar to arguing that there is a less discriminatory alternative employment practice. This argument, then, would place the burden of the alternative employment practice prong on the defendant, contravening the burden-shifting scheme in the statute. *See infra* notes 147-151 and accompanying text.

[144] *El*, 479 F.3d at 242.

that employers show that a discriminatory hiring policy accurately—but not perfectly—ascertains an applicant's ability to perform successfully the job in question. In addition, Title VII allows the employer to hire the applicant most likely to perform the job successfully over others less likely to do so."[145] Thus, all circuits seem to accept varying levels of job-relatedness rather than strict business necessity, but it may be worth considering how data mining would fare under the strict standard as well.[146]

The last piece of disparate impact doctrine is the "alternative employment practice" prong. Shortly after *Griggs*, the Supreme Court decided *Albemarle Paper Co. v. Moody*, in which the Court held in part that "[i]f an employer does then meet the burden of proving that its tests are 'job related,' it remains open to the complaining party to show that other tests or selection devices, without a similarly undesirable racial effect, would also serve the employer's legitimate interest in 'efficient and trustworthy workmanship.'"[147] This burden-shifting scheme was then codified in the 1991 Act as the "alternative employment practice" requirement.[148] Congress

[145] *Id.*

[146] Interestingly, it seems that many courts read identical business necessity language in the Americans with Disabilities Act to refer to a minimum qualification standard. *See, e.g.*, Sullivan v. River Valley Sch. Dist., 197 F.3d 804, 811 (6th Cir. 1999)("There must be significant evidence that could cause a reasonable person to inquire as to whether an employee is still capable of performing his job. An employee's behavior cannot be merely annoying or inefficient to justify an examination; rather, there must be genuine reason to doubt whether that employee can 'perform job-related functions.'"). Presumably, this is because disabilities more immediately raise the question of whether a person is capable at a minimum of performing a job than race or sex does, but, ironically, it means that disparate impact will be *more* tolerated where it is less likely to be obviously justified. Christine Jolls has in fact argued that disparate impact is, to a degree, functionally equivalent to accommodations law. Jolls, *supra* note 73, at 652.

[147] 422 U.S. 405, 425 (1975).

[148] 42 U.S.C. § 2000e-2(k)(1)(A). The "alternative employment practice" test has not always been treated as a separate step. In *Dothard v. Rawlinson*, the Court rejected a business necessity defense that physical strength was necessary to the job because the job requirements at issue involved height and weight, rather than measuring strength directly. 433 U.S. 321, 332 (1977). Thus, rather than a separate prong, the *Dothard* Court treated the alternative employment practice test as a narrow tailoring requirement for the business necessity defense. Similarly, the *Wards Cove* Court treated the test as part of the "business justification" phase. *Wards Cove*, 490 U.S. at 659. The *Albemarle* Court, though creating a surrebuttal and thus empowering plaintiffs, seemed to regard the purpose of disparate impact as merely smoking out pretexts for intentional discrimination. 422 U.S. at 425; *see also* Primus, *supra* note 77, at 537 (2003). If this is so, treatment of the alternative employment practice requirement as a narrow tailoring requirement does make sense, much as the narrow tailoring requirement of strict scrutiny in equal protection serves the function of smoking out invidious purpose. Rubenfeld, *supra* note 77, at 428; City of Richmond v. J.A. Croson Co., 488 U.S. 469, 493, 109 S. Ct. 706, 721, 102 L. Ed. 2d 854 (1989).

Every circuit to address the question, though, has held that the 1991 Act returned the did not define the phrase, and its substantive meaning remains uncertain. *Wards Cove* was the first case to use the specific phrase, so Congress's instruction to reset the law to the pre-*Wards Cove* standard is particularly perplexing.[149] The best interpretation is probably to reach back to *Albemarle*'s reference to "other tests or selection devices without a similarly undesirable racial effect."[150] But this interpretation is slightly odd because in *Albemarle*, business necessity was still somewhat strict, and it is hard to imagine a business practice that is "necessary" while there exists a less discriminatory alternative that is just as effective.[151] If business necessity/job-relatedness is a less stringent requirement, though, then the presence of the alternative employment practice requirement does at least give it some teeth.

Now return to data mining. For now, assume a court does not apply the strict necessity standard, but has some variation of "job-related" in mind (as all circuit courts do today).[152] The threshold issue is clearly whether the sought-after trait—the target variable—is job-related, regardless of the machinery used to predict it. If the target variable is not sufficiently job-related, a business necessity defense will fail, regardless of the fact that the decision was made by algorithm. Thus, disparate impact liability can be found for improper care in target variable definition. For example, it would be difficult for an employer to justify an adverse determination triggered by the appearance of an advertisement suggesting a criminal record alongside the search results for a candidate's name. Sweeney found such a search to

doctrine to the *Albemarle* burden-shifting scheme. Jones v. City of Boston, 752 F.3d 38, 54 (1st Cir. 2014); Howe v. City of Akron, 723 F.3d 651, 658 (6th Cir. 2013); Tabor v. Hilti, Inc., 703 F.3d 1206, 1220 (10th Cir. 2013); Puffer v. Allstate Ins. Co., 675 F.3d 709, 717 (7th Cir. 2012); Gallagher v. Magner, 619 F.3d 823, 833 (8th Cir. 2010); Gulino v. N.Y. State Educ. Dep't, 460 F.3d 361, 382 (2d Cir. 2006); Int'l Bhd. of Elec. Workers, AFL-CIO, Local Unions Nos. 605 & 985 v. Mississippi Power & Light Co., 442 F.3d 313, 318 (5th Cir. 2006); Anderson v. Westinghouse Savannah River Co., 406 F.3d 248, 277 (4th Cir. 2005); Ass'n of Mexican-Am. Educators v. State of California, 231 F.3d 572, 584 (9th Cir. 2000); EEOC v. Joe's Stone Crab, Inc., 220 F.3d 1263, 1275 (11th Cir. 2000); Lanning v. Se. Pennsylvania Transp. Auth. (SEPTA), 181 F.3d 478, 485 (3d Cir. 1999). The D.C. Circuit has not explicitly observed that a burden-shifting framework exists.

[149] Sullivan, *supra* note 92, at 964; Michael J. Zimmer, *Individual Disparate Impact Law: On the Plain Meaning of the 1991 Civil Rights Act*, 30 LOY. U. CHI. L.J. 473, 485 (1999).

[150] *See, e.g.*, Jones, 752 F.3d at 53 (citing *Albemarle* to find meaning in the 1991 Act's text); *Allen v. City of Chicago*, 351 F.3d 306, 312 (7th Cir. 2003)(same, but with a "see also" signal).

[151] William R. Corbett, *Fixing Employment Discrimination Law*, 62 SMU L. REV. 81, 92 (2009).

[152] The difference would be whether mining for a single job-related trait, rather than an holistic ranking of "good employees" is permissible at all. *See* discussion surrounding notes 171-173, *infra*.

have a disparate impact, and the EEOC and several federal courts have interpreted Title VII to prohibit discrimination on the sole basis of criminal record, unless there is a specific reason the particular conviction is related to the job.[153] This is true independent of the fact that it is an artifact of third-party bias; it is just a matter of whether the target variable is job-related. In the end, though, because determining that a business practice is not job-related actually requires a normative determination that it is instead discriminatory, courts tend to accept most common business practices for which an employer has a plausible story.[154]

Once a target variable is established as job related, the next two questions are whether the model is predictive of that trait in the future and whether it is accurate enough. The nature of data mining suggests that both will be the case. First, data mining is designed entirely to predict future outcomes, and, if seeking a job-related trait, future job performance. One commentator who reviewed the trend of disparate impact cases lamented the weakening of disparate impact law since its inception, noting that "[f]ederal case law has shifted from a prospective view of meritocracy to a retrospective view, thereby weakening disparate impact law."[155] What the author meant was that, in *Griggs*, the Court recognized that education and other external factors were unequal, and discounted a measure of meritocracy that looked to past achievements, in favor of comparing the likelihood of future ones. On the other hand, by the time *Wards Cove* was decided, the Court had shifted to a model of retrospective meritocracy that presumed the legitimacy of past credentials, thus upholding the status quo.[156] While data mining must take the past—represented by the training data—as given, the predictions are much more accurate than those past credentials that disparate impact doctrine has come to accept.[157] In a

[153] EQUAL EMP'T OPPORTUNITY COMM'N, CONSIDERATION OF ARREST AND CONVICTION RECORDS IN EMPLOYMENT DECISIONS UNDER TITLE VII OF THE CIVIL RIGHTS ACT OF 1964, *available at* http://www.eeoc.gov/laws/guidance/upload/arrest_conviction.pdf; Michael Connett, *Employer Discrimination Against Individuals with A Criminal Record: The Unfulfilled Role of State Fair Employment Agencies*, 83 TEMP. L. REV. 1007, 1017 & nn. 82-83 (2011)(citing EQUAL EMP'T OPPORTUNITY COMM'N, POLICY STATEMENT ON THE ISSUE OF CONVICTION RECORDS UNDER TITLE VII OF THE CIVIL RIGHTS ACT OF 1964 (Feb. 4, 1987), *available at* http://www.eeoc.gov/policy/docs/convict1.html and several circuit court cases).

[154] Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. Rev. 701, 753 (2006).

[155] DeSario, *supra* note 125, at 481.

[156] *Id.* at 493. *See also* Conclusion, *infra*.

[157] *See* Don Peck, *They're Watching You at Work*, THE ATLANTIC (Nov. 20, 2013), *available at* http://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/ (discussing Google's abandonment of traditional hiring metrics such as

hypothetical perfect case of data mining, the available information would be rich enough that no difference would exist between the reliance on the past information and the possibilities for the future. Thus, data mining would likely have satisfied even the *Griggs* Court that the models are looking toward future job performance, not merely past credentials.

Second, the doctrine will ask whether the model adequately enough predicts what it is supposed to. In the traditional context, this type of question arises in the case of general aptitude tests that might end up measuring unrelated elements of cultural awareness better than intelligence.[158] This is where the different mechanisms for discriminatory effects matter. Part I posited that proxy discrimination optimizes correctly, so if it evidences a disparate impact, it reflects unequal distribution of relevant traits in the real world. Therefore, it will be as good a job predictor as possible given the current shape of society. Models trained on biased samples and mislabeled examples, on the other hand, will result in correspondingly skewed assessments, rather than reflecting real-world disparities. The same effect may be present in models that rely on insufficiently rich or insufficiently granular datasets: by designation they do not reflect reality. These models might or might not be considered job-related, depending on whether the errors distort the outcome enough to make them no longer good job performance predictors.

The Guidelines have set forth validation procedures intended to create a standard for whether selection procedures are considered job-related. Quantifiable tests that have a disparate impact must be validated according to the procedures in the Guidelines (if possible), or a presumption arises that they are not job-related.[159] Under the Guidelines, a showing of validity takes one of three forms: criterion-related, content, or construct.[160] Criterion-related validity "consist[s] of empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance."[161] The "relationship between

brain-teasers and GPA for those more than two years out of school because they are not good predictors of performance).

[158] *See, e.g., Griggs v. Duke Power Co.*, 420 F.2d 1225, 1240 (4th Cir. 1970) *rev'd*, 401 U.S. 424 (1971)("Since for generations blacks have been afforded inadequate educational opportunities and have been culturally segregated from white society, it is no more surprising that their performance on 'intelligence' tests is significantly different than whites' than it is that fewer blacks have high school diplomas.")

[159] 29 C.F.R. § 1607.5. The Guidelines also cite two categories of practices that are unsuitable for validation: informal and unscored practices, and technical infeasibility. *Id.* § 1607.6(B). For the latter case, the Guidelines state that the selection procedure still should be justified somehow or another option should be chosen.

[160] *Id.* § 1607.5(B)

[161] *Id.*

performance on the procedure and performance on the criterion measure is statistically significant at the 0.05 level of significance."[162] Content validity refers to testing skills or abilities that generally are or have been learned on the job, though not those that could be acquired in a "brief orientation."[163] Construct validity refers to a test designed to measure some innate human trait such as honesty. A user of a construct "should show by empirical evidence that the selection procedure is validly related to the construct and that the construct is validly related to the performance of critical or important work behavior(s)."[164]

As a statistical predictive measure, a data mining model could be validated by either criterion-related or construct validity, depending on the trait being sought. Either way, the important part is that there be statistical significance showing that the result of the model correlates to the trait (which was already determined to be an important element of job performance). This is an exceedingly low bar for data mining, because data mining's predictions necessarily rest on demonstrated statistical relationships. If data mining actually is used in the selection procedure, it will have been because it was predictive of *something*. So the question solely comes down to whether the trait sought is important enough to job performance to justify its use in any context.

Even assuming the Guidelines were a hurdle for data mining, some courts ignore the Guidelines' recommendation that an unvalidated procedure be rejected, preferring to rely on "common sense" or finding a "manifest relationship" between the criteria and successful job performance.[165] Moreover, it is possible that the Supreme Court inadvertently overruled the Guidelines in 2009. In *Ricci v. Destefano*, a case that will be discussed in more detail in Part III.B, the Court found no genuine dispute that the tests at issue met the job-related and business necessity standards despite not having been validated under the Guidelines and despite the employer *actively denying* that they could be validated.[166] While the business necessity defense was not directly at issue in *Ricci*, "[o]n the spectrum between heavier and lighter burdens of justification, the Court

[162] *Id.* § 1607.14(B)(5).

[163] *Id.* §§ 1607.5(F) &1607.14(C).

[164] *Id.* § 1607.14(D)(3).

[165] Wax, *supra* note 119, at 633-34.

[166] David A. Drachsler, Assessing the Practical Repercussions of Ricci, AMERICAN CONSTITUTION SOCIETY BLOG (July 27, 2009), http://www.acslaw.org/node/13829 (observing that the Court in *Ricci v. DeStefano* found no genuine dispute that the tests at issue met the job-related and business necessity standards despite not being validated under the Guidelines, and the Guidelines creating a presumption of invalidity for non-validated tests that are discriminatory).

came down decidedly in favor of a lighter burden."[167]

Thus, there is good reason to believe that any or all of the data mining models predicated on legitimately job-related traits pass muster under the business necessity defense. Models trained on biased samples, mislabeled examples, and with limited features, however, might trigger liability under the alternative employment practice prong. If it can be shown that an alternative, less discriminatory practice that accomplishes the same goals exists, and that the employer "refuses" to use it, he can be found liable. In this case, the obvious alternative employment practice would be to fix the problems with the models.

Fixing the models is not a trivial task. For example, in the LinkedIn hypothetical, where the demonstrated interest in different kinds of employees reflects employers' prejudice, LinkedIn is the party that determines the algorithm by which the discrimination occurs (in this case, based on reacting to third-party preferences). If an employer were to act on the recommendations suggested by the LinkedIn engine, there would not be much he could do to make it less reflective of third-party prejudice, aside from calling LinkedIn and asking nicely. Thus, it could not really be said that the employer "refuses" to use an alternative employment practice. The employer can either use the third-party tool or not. Similarly, a case suffering from reporting bias, like Street Bump, might be possible to fix, but the employer would need access to the raw data in order to do so.[168] In the case of insufficiently rich or granular features, the employer would need to collect more data in order to make the model more discerning. But collecting more data can be time consuming and costly, if not impossible to do for legal or technical reasons.

Moreover, the under- and over-representation of members of protected classes is not always evident. Nor is the mechanism by which such under- or over- representation occurs. The idea that the representation of different social groups in the dataset can be brought into proportions that better match those in the real world presumes that analysts have some independent mechanism for determining these proportions. Thus, in cases where the employer created or has access to the model, can discover that there is discriminatory effect, and can discover the particular mechanism by which that effect operates, and all this can be proven by a plaintiff, the employer might be held liable under disparate impact because there is a less discriminatory alternative. The same can be said for models with insufficiently rich feature sets. Clearly there are times when more features would improve a discriminatory outcome. But it is, almost by definition,

[167] George Rutherglen, Ricci v. Destefano*: Affirmative Action and the Lessons of Adversity*, 2009 SUP. CT. REV. 83, 107.

[168] *See infra* Part III.B.1

hard to know which features are going to make the model more or less discriminatory. Indeed, it is often impossible to know which features are missing because data miners do not operate with a causal model in mind. So while theoretically a less discriminatory alternative would nearly always exist, proving it is a difficult proposition.

There is yet one more hurdle. Neither Congress nor courts have specified what it means for an employer to "refuse" to adopt the less discriminatory procedure. Scholars have suggested that perhaps the employer cannot be held liable until it has considered the alternative and rejected it.[169] Thus, if the employer has run an expensive data collection and analysis operation without ever being made aware its discriminatory tendencies, and it cannot afford to re-run the entire operation, is the employer "refusing" to use a less discriminatory alternative, or does one simply not exist? How much would the error correction have to cost an employer before it is not seen as a refusal to use the procedure?[170] Should the statute actually be interpreted to mean that an employer "unreasonably refuses" to use an alternative employment practice? These are all difficult questions, but suffice it to say, the prospect of winning a data mining discrimination case on alternative employment practice grounds seems slim.

The final consideration for disparate impact liability is the possibility that a court or Congress will reinvigorate strict business necessity.[171] In that case, things look a little better. Where an employer models job tenure,[172] for example, a court may be inclined to hold that it is job-related because it is a "legitimate, non discriminatory business objective."[173] But it is clearly not necessary to the job. The same reasoning applies to mining for any single trait that is job-related – the practice of data mining is not focused on discovering make-or-break skills. Unless the employer can show that below the cut score, employees cannot do the work, then the strict business necessity defense will fail. Thus, disparate impact

[169] Sullivan, *supra* note 92, at 964; Zimmer, *supra* note 149, at 506.

[170] For a discussion of courts using cost as a rationale here, see Ernest F. Lidge III, *Financial Costs As A Defense to an Employment Discrimination Claim*, 58 Ark. L. Rev. 1, 32-37 (2005).

[171] This would likely require Congressional action because strict business necessity essentially transfers the burden to prove a lack of an alternative employment practice to the defense. By implication, if a practice is "necessary," there cannot be alternatives. The statute, as it reads now, clearly states that the plaintiff has the burden for that prong. 42 U.S.C. § 2000e-2(k)(1)(A)(ii).

[172] An increasingly common practice in low-wage, high-turnover jobs. Peck, *supra* note 157.

[173] EEOC v. Joe's Stone Crab, Inc., 220 F.3d 1263, 1275 (11th Cir. 2000); *see also* Gallagher v. Magner, 619 F.3d 823, 834 (8th Cir. 2010)("legitimate, non discriminatory policy objective").

that occurs as an artifact of the problem specification stage can potentially be addressed by strict business necessity.

This reasoning is undermined, though, where employers do not mine for a single trait, but automate their decision process by modeling job performance on a holistic measure of what makes good employees. If employers determine traits of a good employee by simple ratings, and use data mining to appropriately divine good employees' characteristics among several different variables, then the argument that the model does not account for certain skills that could compensate for the employee's failings on that metric loses its force. Taken to an extreme, while an 8000-feature holistic determination of a "good employee" would still not be strictly "necessary," holding a business to a standard that does not allow it to use such a determination would simply be forbidding business to rank candidates at all if there is any disparate impact that results. Thus, while the strict business necessity defense could prevent myopic employers from creating disparate impacts by their choice of target variable, it would still not address forms of data mining that model general job performance rather than predict specific traits.

C. Masking and Problems of Proof

Last, it is necessary to examine masking. As discussed earlier, there is no theoretical problem with finding liability for masking.[174] It is a disparate treatment violation as clear as any. But like traditional forms of intentional discrimination, it suffers from difficulties of proof. While the traditional difficulties finding intent from stray remarks or other circumstantial evidence are surely present, masking presents additional complications with detection.

Data mining allows employers who wish to discriminate on the basis of a protected class to disclaim any knowledge of the protected class in the first instance, while simultaneously inferring such knowledge from the data. An employer may want to discriminate by using proxies for protected classes, such as in the case of redlining.[175] Due to housing segregation, neighborhood is a good proxy for race, and can be used to redline candidates without reference to race. This is a relatively unsophisticated example, however. It is possible that some combination of musical tastes,[176]

[174] *See*, text accompanying *supra* notes 84-85.

[175] *See supra* Part I.E.

[176] Alistair Croll, "Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It," *Solve for Interesting*, July 31, 2012, http://solveforinteresting.com/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it/.

stored "likes" on Facebook,[177] and network of friends[178] will reliably predict membership in protected classes, and the employer can use these to discriminate by setting up future models to sort by those items.

More generally, as discussed in Part I, any of the mechanisms by which unintentional discrimination can occur can also be employed intentionally. The example described above is intentional discrimination by proxy, but it is also possible to intentionally bias the data collection process, purposefully mislabel examples, or deliberately use an insufficiently rich set of features, though some of these would probably require a great deal of sophistication.[179] These methods of intentional discrimination will look, for all intents and purposes, identical to the unintentional discrimination that can result from data mining. Therefore, detection of the discrimination in the first instance will require the same techniques as detecting unintentional discrimination, namely a disparate impact analysis. Further, assuming there is no circumstantial evidence like stray remarks with which to prove intent, one might attempt to prove intent by demonstrating that the employer is using less representative data, poorer examples, or fewer and less granular features than he might otherwise use were he interested in the best possible candidate. That is, one could show that the neutral employment practice is a pretext by demonstrating that there is a more predictive alternative.

This sounds like disparate impact analysis: a plaintiff asks the same question as in the "alternative employment practice" prong: whether there were more relevant measures the employer could have used.[180] But the business necessity defense is not available in a disparate treatment case,[181] so alternative employment practice is not the appropriate analysis. Scholars have noted, though, that the line between disparate treatment and disparate impact in traditional cases is not always clear,[182] and sometimes employer actions can be legitimately categorized as either or both.[183] As George Rutherglen has pointed out, "[c]oncrete issues of proof, more than any

[177] Michal Kosinski, David Stillwell, and Thore Graepel, "Private Traits and Attributes Are Predictable From Digital Records of Human Behavior," *Proceedings of the National Academy of Sciences* 110, no. 15 (April 9, 2013): 5802–5, doi:10.1073/pnas.1218772110.

[178] Carter Jernigan and Behram F T Mistree, "Gaydar: Facebook Friendships Expose Sexual Orientation," *First Monday* 14, no. 10 (September 25, 2009), doi:10.5210/fm.v14i10.2611.

[179] *See* Dwork, et al., *supra* note 66, app'x. ("Catalog of Evils").

[180] *Cf.* Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975)(creating alternative employment practice prong for the purpose of rooting out pretext).

[181] 42 U.S.C. § 2000e-2(k)(2)

[182] George Rutherglen, *Disparate Impact, Discrimination, and the Essentially Contested Concept of Equality*, 74 FORDHAM L. REV. 2313, 2313; Stacy E. Seicshnaydre, *Is The Road To Disparate Impact Paved With Good Intentions?: Stuck On State Of Mind In Antidiscrimination Law*, 42 WAKE FOREST L. REV. 1141, 1142-43 (2007).

[183] Rutherglen, *supra* note 182, at 2320-21.

abstract theory, reveal the fundamental similarity between claims of intentional discrimination and those of disparate impact. The evidence submitted to prove one kind of claim invariably can be used to support the other."[184] Rutherglen's point is exactly what must happen in the data mining context. Disparate treatment and disparate impact become essentially the same thing from an evidentiary perspective.

To the extent that disparate impact and treatment are, in reality, different theories, they are often confused for each other. Plaintiffs will raise both types of claims as a catchall because they cannot be sure on which theory they might win, so both theories will be in play in a given case.[185] As a result, courts often seek evidence of state of mind in disparate impact cases,[186] and objective, statistical evidence in disparate treatment.[187] Assuming the two theories are not functionally the same, the need to use the same evidence for disparate treatment and disparate impact will only lead to more confusion, and as a result, uncertainty within the courts. This uncertainty puts masking on less solid footing than one might think initially, despite its clear nature as a theoretical violation.

A final point is that traditionally, employers who do *not* want to discriminate go to great lengths to avoid raising the prospect that they have violated the law. Thus they tend to avoid collecting information about attributes that reveal an individual's membership in a protected class. Employers even pay third parties to collect relatively easy-to-find information on job applicants, such as professional honors and awards, as well as compromising photos, videos, or membership in online groups, so that the third party can send back a version of the report that "remove[s] references to a person's religion, race, marital status, disability and other information protected under federal employment laws."[188] This allows employers to honestly disclaim any knowledge of the protected information. Nonetheless, in a data mining regime, if an employer seeks to discriminate according to protected classes, it will often be possible to infer the information from the data. Thus, employers' old defense to suspicion of discrimination—that they did not even see the information—is no longer adequate to separate would-be intentional discriminators from employers that would not do so.

[184] *Id.* at 1320.

[185] Seicshnaydre, *supra* note 182, at 1147-48.

[186] *Id.*, *passim*.

[187] Rutherglen, *supra* note 182, at 2321-22.

[188] Jennifer Prestion, *Social Media History Becomes a New Job Hurdle*, N.Y. TIMES, at B1 (July 25, 2011), *available at* http://www.nytimes.com/2011/07/21/technology/social-media-history-becomes-a-new-job-hurdle.html.

III. The Difficulty for Reforms

While each of the mechanisms for discrimination in data mining present difficulties for Title VII as currently written, there will also be certain impediments to reforming Title VII to address the resulting problems. Computer scientists and others are working on technical remedies,[189] so to say that there are problems with legal remedies is not to suggest that the problems with discrimination in data mining cannot be solved at all. Nonetheless, this Part focuses on the legal aspects, and as it illustrates, even assuming that the political will to reform Title VII exists, potential legal solutions are not straightforward.

This Part discusses two types of difficulties with reforming Title VII. First, at each stage of the data mining process, there are issues internal to its operation that make legal reform difficult. For example, there is a certain amount of subjectivity and latitude for employers required in defining a "good employee," but, at the same time, some answers are clearly better than others. How does one draw that line? Can employers gain access to the additional data necessary to correct for collection bias? How much will it cost them to find it? How do we identify the "correct" baseline historical data to avoid reproducing past prejudice or the "correct" level of detail and granularity in a dataset? Before laws can be reformed, policy-level answers to these basic technical, philosophical, and economic questions need to be addressed at least to some degree.

Second, reform will face political and constitutional constraints external to the logic of data mining that will affect how Title VII can be permissibly reformed to address it. Not all of the mechanisms for discrimination seem to be amenable to procedural remedies. If that holds true, only after-the-fact reweighting of results may be able to compensate for the discriminatory outcomes. This is not a matter of missing legislation; it's a matter of practical realities. Unfortunately, while in many cases no procedural remedy will be sufficient, any attempt to design a legislative or judicial remedy premised on reallocation of employment outcomes will not survive long in the current political or constitutional climate, as it raises the specter of affirmative action. Politically, anything that even hints at affirmative action is a nonstarter today, and to the extent that it is permissible to enact such policies, their future constitutionality is in doubt.

A. Internal Difficulties

[189] *See generally,* Discrimination IN THE INFORMATION SOCIETY, *supra* note 25,

1. Defining the Target Variable

Settling on a target variable is a necessarily subjective exercise. Disputes over the superiority of competing definitions are often insoluble because the target variables are themselves incommensurable. There are, of course, easier cases, where prejudice or carelessness leads to definitions that subject members of protected classes to avoidably high rates of adverse determinations. But most cases are likely to involve genuine business disagreements over ideal definitions, with each having a potentially greater or lesser impact on protected classes. There is no stable ground upon which to judge the relative merits of definitions because they often reflect competing ideas about the very nature of the problem at issue.[190] As Oscar Gandy has argued, "certain kind of biases are inherent in the selection of the goals or objective functions that automated systems will be designed to support."[191] There is no escape from this situation; a target variable *must* reflect judgments about what really is the problem at issue in making hiring decisions. For certain employers, it might be rather obvious that the problem is one of reducing the administrative costs associated with turnover and training; for others, it might be improving sales; for still others, it might be increasing innovation. Any argument for the superiority of one over the other will simply make appeals to competing and incommensurate values.

For these same reasons, however, defining the target variable also offers an opportunity for creative thinking about the potentially infinite number of ways of making sound hiring decisions. Data miners can experiment with multiple definitions that each seem to serve the same goal, even if these fall short of what they themselves consider ideal. In principle, employers should rely on proxies that are maximally proximate to the actual skills demanded of the job, at the most extreme assessing how well an applicant performs the actual task to which he or she will be assigned. While there should be a tight nexus between the sought-after features and the skills demanded of the job, this may not be possible for practical and economic reasons. Which leaves data miners in a position to dream up many different non-ideal ways to make hiring decisions that may have a greater or less adverse impact on protected classes.

The Second Circuit considered such an approach in *Hayden v. County of Nassau*.[192] In *Hayden*, the county's goal was to find a test that

[190] David J Hand, *Deconstructing Statistical Questions*, 157 J. ROYAL STATISTICAL SOC'Y. SERIES A (STATISTICS IN SOCIETY) 317, 318-320 (1994).

[191] Oscar H. Gandy Jr., *Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems*, 12 ETHICS & INFORMATION TECH. 29, 39 (2009).

[192] 180 F.3d 42, 46 (2d Cir. 1999).

was "valid, yet minimized the adverse impact on minority applicants."[193] The county thus administered a police entrance exam with twenty-five parts that could be scored independently, but for which a statistically valid result could be achieved by one of several configurations that counted only a portion of the test sections, without requiring all of them.[194] The county ended up using nine of the sections as a compromise, after rejecting one configuration that was more advantageous to minority applicants but less statistically sound.[195] This was a clear example of defining a problem in such a way that it becomes possible to reduce the disparate impact without compromising the accuracy of the assessment mechanism.

2. Training Data

a. Labeling Examples

A rule that forbids employers from modeling decisions based on historical examples tainted by prejudice would, in theory, address the problem of improper labeling. But if the only examples an employer has to draw on are those of past employees whose representativeness has been skewed by prior discrimination, all learned rules will recapitulate this discrimination. Therefore, any solution must be a compromise between a rule that forbids employers from relying on past discrimination and one that allows them to base hiring decisions on historical examples of good employees.

Title VII has always had to balance its mandate to eliminate discrimination in the workplace with employers' legitimate discretion. For example, one of the most common selection procedures that explicitly reproduced past discrimination was seniority.[196] Seniority was and is still often a legitimate metric for promotion, and is especially important in collective bargaining, but, after the passage of Title VII, it was also often used to keep black people from advancing to better jobs because they had not been hired until Title VII forced employers to hire them.[197] Despite this obvious problem with seniority, Title VII contains an explicit carve-out for "bona fide seniority or merit system[s],"[198] and, as a result, the Supreme Court has held that "absent a discriminatory purpose, the operation of a

[193] *Id.*

[194] *Id.*

[195] *Id.*

[196] Selmi, *supra* note 154 at 715.

[197] Albemarle Paper Co. v. Moody, 422 U.S. 405 (1975)("The basis of Albemarle's liability was that its seniority system perpetuated the effects of past discrimination.").

[198] 42 U.S.C. § 2000e-2(h).

seniority system cannot be an unlawful employment practice even if the system has some discriminatory consequences."[199] Given the inherent tension between ensuring that past discrimination is not reproduced in future decisions and permitting employers legitimate discretion, it should be unsurprising that, when translated to data mining, the problem is not amenable of a clear solution.

The difficulty is even more central to data mining, though. Data miners who attempt to remove the influence that prejudice had on prior decisions by recoding or relabeling examples may find that they cannot easily resolve what the non-prejudicial determination would have been. As Calders and Zliobaite point out, "the notion of what is the correct label is fuzzy."[200] Employers are unlikely to have perfectly objective and exhaustive standards for hiring; indeed, part of the hiring process is purposefully subjective. At the same time, employers are unlikely to have discriminated so completely in the past that the only explanation for rejecting an applicant is whether they happened to be a member of one of these protected classes. This leaves data miners who have been tasked with correcting for prior prejudice with the impossible challenge of determining what the correct subjective employment decision would have been absent prejudice. Undoing the imprint of prejudice on the data may demand a complete re-rendering of the biased decisions rather than simply adjusting those decisions according to some fixed statistical measure.

b. Data Collection

When datasets are very obviously skewed, as in the case of Street Bump or an employer gauging interest and ability based on information available only to those with easy access to the internet, identifying and correcting for the reporting biases might be a relatively straightforward affair. Boston's Office of New Urban Mechanics, for instance, has already partnered with "a range of academics to take into account issues of equitable access and digital divides."[201] Often, however, the source and/or degree of the bias might not be as immediately apparent.[202] Analysts can only determine the extent of, and correct for, unintentional discrimination

[199] Trans World Airlines, Inc. v. Hardison, 432 U.S. 63, 82 (1977).

[200] Calders & Žliobaitė, *supra* note 52.

[201] Crawford, *supra*, note 41. Such techniques would also address the concerns raised in Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55 (2013).

[202] For example, establishing whether and to what extent crime statistics misrepresent the relative proportion of offenses committed by different social groups, especially those crimes that are more likely to go under- or un-reported if not directly observed by the police, is not an easy task. *See* BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE (2007).

that results from reporting, sampling, and selection biases if the analyst has access to information that somehow reveals misrepresentations of members of—and matters concerning—protected classes in the dataset.

Any attempt to correct for collection bias confronts an immediate problem. An employer must first recognize the specific type of bias that is producing disparate results. Then, in order to correct for it, an employer must have access to the underlying data, and often an ability to collect more. Where more data is clearly not accessible, some of the error can be mitigated by oversampling underrepresented communities to proactively compensate for the likelihood of disparate impact.[203]

If the employer fails to be proactive, however, or tries and fails to detect the bias that causes the disparate impact, whether he will face any liability is an open question, as discussed in Part II.B, partly based on how liberally a court interprets the requirement that an employer "refuses" to use an alternative scheme. Even a liberal interpretation, though, would require evidence of the particular type of discrimination at issue, coupled with evidence that such an alternative scheme exists, so finding liability seems unlikely. But there is an equally difficult question about what to do about it.

To address collection bias directly, an employer or an auditor must have access to the underlying data and the ability to adjust the model. Congress could require this directly of any employer using data mining techniques. Some employers are investing in their own data now, so for them it would not be a problem.[204] But employers also seem happy to rely on models developed and administered by third parties, who may have a far greater set of example and far richer data than any individual company.[205] Furthermore, due to economies of scale that are especially important in data analysis, one can imagine that third parties specializing in work-force science will be able to offer employers this service much less expensively than they could manage it themselves. Congress would face strong resistance from the ever-growing data analysis industry if it attempted to demand that employers have access to the data, whose business depends on the proprietary nature of the amassed information. More likely Congress could require audits by a third party like the EEOC or a private auditor, in order to protect trade secrets, but this still seems a tall task. Ultimately,

[203] Faisal Kamiran & Toon Calders, *Data Preprocessing Techniques for Classification without Discrimination*, 33 Knowl Inf Syst 1, 3 (2011).

[204] Peck, *supra* note 157.

[205] Matt Richtel, *How Big Data Is Playing Recruiter for Specialized Workers*, N.Y. TIMES, at BU1 (Apr. 28, 2013), *available at* http://www.nytimes.com/2013/04/28/technology/how-big-data-is-playing-recruiter-for-specialized-workers.html.

though, because proactive oversampling and retroactive data correction are at least possible, collection bias has the most promising prospects for a workable remedy of any of the identified mechanisms.

3. Feature Selection

Even in the absence of prejudice or bias, determining the proper degree of precision in the distinctions drawn through data mining can be extremely difficult. Under formal disparate treatment, this is straightforward: any decision that expressly classifies by membership in a protected class is one that draws distinctions on illegitimate grounds. But far less clear is the degree of individuation that confers legitimacy on statistical discrimination when it does not draw distinctions on the basis of proscribed criteria. In these cases, the perceived legitimacy seems to depend on a number of factors: (1) whether the errors seem avoidable because (2) gaining access to additional or more granular data would be trivial (3) or would not involve costs that (4) outweigh the benefits. This seems to suggest that the task of evaluating the legitimacy of data mining can be reduced to a rather straightforward cost-benefit analysis, in which companies have an obligation to pursue ever more—and more granular—data until the costs of gathering that data exceed the benefits conferred by the marginal improvements in accuracy.

Unfortunately, as is often the case with cost-benefit analyses, this approach fails to consider how different actors will perceive the value of the supposed benefits as well as the costs associated with errors. The obvious version of this objection is that victims of erroneous determinations may find cold comfort in the fact that certain decisions are rendered more reliably overall when decision-makers employ data mining.[206] A more sophisticated version of this criticism focuses on the way such errors assign benefits and costs to different actors at systematically different rates. A model with any error rate that continues to turn a profit may be acceptable to decision-makers at a company, no matter the costs or inconvenience to specific customers.[207] Even when companies are subject to market pressures that would force them to compete by lowering these error rates, the companies may find that there is simply no reason to invest in efforts that do so if the errors happen to fall disproportionately on especially

[206] As Schauer explains, perfectly particularized decisions are, of course, a logical impossibility. Accepting this inherent limitation introduces a different sort of procedural concern: occasional errors might be tolerable if they are easy to detect and rectify, which is why, among other things, the perceived legitimacy of decisions often also depends on due process. See SCHAUER, *supra* note 55, at 172; *see also* Citron, *supra* note 9,

[207] Gandy, *supra* note 191.

unprofitable groups of consumers. Furthermore, assessing data mining as a matter of balancing costs and benefits leaves no room to consider morally salient disparities in the degree to which the costs are borne by different social groups. This raises the prospect that there might be systematic differences in the rates at which members of protected classes are subject to erroneous determinations. Condemning these groups to bear the disproportionate burden of erroneous determinations, even if this means that the majority enjoys greater accuracy in decision-making, would strike many as highly objectionable. Indeed, simply accepting these cost differences as a given would subject those already in less favorable circumstances to less accurate determinations.

Even if companies assumed the responsibility for ensuring that members of protected classes do not fall victim to erroneous determinations at systematically higher rates, they could find that increasing the resolution and range of their analysis still fails to capture the causal mechanisms that account for different outcomes because those mechanisms are not easily represented in data.[208] In such cases, rather than reducing the error rate for those in protected classes, data miners could structure their analysis in such a way that it minimizes the difference in error rates between groups. This solution may involve some unattractive tradeoffs, however: in reducing the disparate impact of errors, it may increase the overall amount of errors. In other words, generating a model that is equally unfair for protected and unprotected classes might increase the overall amount of unfairness.

4. Proxies

Computer scientists have been unsure how to deal with redundant encodings in datasets. Simply withholding these variables from the data mining exercise often removes criteria that hold demonstrable and justifiable relevance to the decision at hand. As Calders and Žliobaitė note, "it is problematic [to remove correlated attributes] if the attribute to be removed also carries some objective information about the label [quality of interest]."[209] Part of the problem seems to be that there is no obvious way to determine *how* correlated a relevant attribute must be with class membership to be worrisome. Nor is there a self-evident way to determine when an attribute is sufficiently relevant to justify its consideration, despite the fact that it is highly correlated with class membership. As Devin Pope and Justin Sydnor note, "variables are likely neither solely predictive nor purely proxies for omitted characteristics."[210]

[208] *See supra* note 52 at accompanying text.

[209] Id.

[210] Devin G Pope and Justin R Sydnor, *Implementing Anti-Discrimination Policies in*

But there is a bigger problem here: attempting to ensure fairly rendered decisions by excising highly correlated criteria only makes sense if the disparate impact happens to be an *avoidable* artifact of a particular way of rendering decisions. And yet, even when denied access to these highly correlated criteria, data mining may suggest alternative methods for rendering decisions that still result in the same disparate impact. This might seem to imply that focusing on isolated data points is a mistake because class membership may be encoded in more than one specific and highly correlated criterion; indeed, it is very likely that class membership is reflected across a number of interrelated data points.[211] But such outcomes might instead demonstrate something more unsettling: that *other* relevant criteria, whatever they are, happen to be possessed at different rates by members of protected classes. This explains why, for instance, champions of predictive policing have responded to critics by arguing that "[i]f you wanted to remove everything correlated with race, you couldn't use anything. That's the reality of life in America."[212] Making accurate determinations means considering factors that are somehow correlated with proscribed features.

Computer scientists have even shown that "[r]emoving all such correlated attributes before training does remove discrimination, but with a high cost in classifier accuracy."[213] This reveals a rather uncomfortable truth: the current distribution of relevant attributes—attributes that can and should be taken into consideration in apportioning opportunities fairly—are demonstrably correlated with sensitive attributes because the sensitive attributes have meaningfully conditioned what relevant attributes individuals happen to possess.[214] As such, attempts to ensure procedural fairness by excluding certain criteria from consideration may be in conflict with the imperative to ensure accurate determinations. The only way to ensure that decisions do not place at systematic relative disadvantage members of protected classes is to reduce the overall accuracy of all determinations. As Dwork et al. remark, these results "demonstrate a

Statistical Profiling Models, 3 AM. ECON. J.: ECON. 206, 206 (2011).

[211] Discussion accompanying *supra* note 79.

[212] Labi, *supra* note 5 (quoting Ellen Kurtz, Director of Research for Philadelphia's Adult Probation and Parole Department).

[213] Toon Calders and Sicco Verwer, *Three Naïve Bayes Approaches for Discrimination-Free Classification*, at 9 (presented at the The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Barcelona, Spain, 2010), *available at* http://wwwis.win.tue.nl/~tcalders/pubs/dami2010.pdf.

[214] In a sense, computer scientists have unwittingly furnished the kind of evidence that social scientists routinely seek: the particular contours of inequality. *See, e.g.*, SOCIAL INEQUALITY, (Kathryn M. Neckerman, ed., 2004).

quantitative trade-off between fairness and utility."[215]

In certain contexts, data miners will never be able to fully disentangle legitimate and proscribed criteria. Take, for example, the fact, discovered by Evolv, a "workforce optimization" consultancy, that "distance between home and work . . . is strongly associated with employee engagement and retention." Despite the strength of this finding, Evolv "never factor[s it] into the score given each applicant . . . because different neighborhoods and towns can have different racial profiles, which means that scoring distance from work could violate equal-employment-opportunity standards.[216] Scholars have taken these cases as a sign that the "major challenge is how to find out which part of information carried by a sensitive (or correlated) attribute is sensitive and which is objective."[217] While researchers are well aware that this may not be easy to resolve, let alone formalize into a computable problem, there is a bigger challenge from a legal perspective: any such undertaking would necessarily wade into the highly charged debate over the degree to which the relatively less favorable position in which members of protected classes find themselves warrants the protection of anti-discrimination law in the first instance.

B. External Difficulties

Two competing principles have always undergirded anti-discrimination law: nondiscrimination and antisubordination.[218] Nondiscrimination is the narrower of the two, holding that the responsibility of the law is to eliminate the unfairness individuals experience at the hands of decisionmakers' choices due to membership in certain protected classes. Antisubordination theory, in contrast, holds that the goal of anti-discrimination law is, or at least should be, to eliminate status-based inequality due to membership in those classes, not as a matter of procedure, but substance.

Different mitigation policies effectuate different rationales. Disparate treatment doctrine arose first, clearly aligning with the nondiscrimination principle by proscribing intentional discrimination,

[215] Dwork, et al., *supra* note 66, at 215. *Cf.* Wax, *supra*, note 122, at 711 (noting intractable problems due to a "validity-diversity tradeoff" in employment metrics).

[216] Peck, *supra* note 157.

[217] Calders & Žliobaitė, *supra* note 52, at 56.

[218] *See, e.g.*, Bagenstos, *supra* note 100, at 40-42 & nn. 214-15 (collecting sources); Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107, 157 (1976). What we refer to here as the "nondiscrimination" principle is often referred to as the "antidiscrimination" principle, especially when contrasted with antisubordination. We use the term "nondiscrimination" to avoid confusion with our references throughout this Article to anti-discrimination law as the entire body of jurisprudence.

whether that discrimination comes in the form of explicit singling out of protected classes for harm, or masked intentional discrimination. From the time when disparate impact developed, however, there has never been clarity as to which of the principles it is designed to effectuate. On the one hand, disparate impact doctrine serves nondiscrimination by being an "evidentiary dragnet" used to "smoke out" well hidden disparate treatment.[219] On the other hand, as an effects-based doctrine, there is good reason to believe it was intended to address substantive inequality. In this sense, "business necessity" is a necessary backstop so that members of traditionally disadvantaged groups could not simply force their way in regardless of the possibility that they lack necessary skill or ability.

The mapping from nondiscrimination and antisubordination to disparate treatment and disparate impact was thus never clean, but disparate treatment was definitely *limited* to the nondiscrimination purpose. Early critics of civil rights laws actually complained that proscribing consideration of protected class was a subsidy to black people, but this argument quickly gave way in the face of the rising importance of the nondiscrimination norm.[220] Over the years, the nondiscrimination principle has come to dominate the landscape so thoroughly that a portion of the populace thinks (as do a few Justices on the Supreme Court) that it is the only valid rationale for anti-discrimination law.[221]

The move away from antisubordination began only five years after disparate impact was established in *Griggs*. In *Washington v. Davis*,[222] the Court held that disparate impact could not apply to constitutional claims because equal protection only prohibited intentional discrimination. Since then, the various affirmative action cases have overwritten the distinction between benign and harmful categorizations of race, in favor of a formalistic nondiscrimination principle, removed from its origins as a tool to help members of historically disadvantaged groups.[223] White men can now bring disparate treatment claims.[224] Where anti-discrimination law is no longer thought to serve the purpose of improving the relative conditions of traditionally disadvantaged groups, antisubordination is not part of the equation.

While the Court has clearly established that antisubordination is not part of equal protection doctrine, that does not mean that it cannot animate statutory anti-discrimination law. The two principles came into sharp

[219] Primus, *supra* note 77, at 520-23.
[220] *Id.*
[221] *See* Bagenstos, *supra* note 100, at 41.
[222] 426 U.S. 229 (1976).
[223] Rubenfeld, *supra* note 77, at 428, 433-36.
[224] Ricci v. DeStefano, 557 U.S. 557 (2009).

conflict in *Ricci v. Destefano*, a 2009 case in which the City of New Haven refused to certify a promotion exam given to its firefighters on the grounds that it would have produced a disparate impact based on its results.[225] The Supreme Court held that the refusal to certify the test, a facially race-neutral attempt to correct for perceived disparate impact, was in fact a race-conscious remedy that constituted disparate treatment of the majority-white firefighters who would have been promoted based on the exam's results.[226] The Court held that disparate treatment cannot be a remedy for disparate impact without a "strong basis in evidence" that the results would lead to actual disparate treatment liability.[227]

This was the first indication at the Supreme Court that disparate impact doctrine could be in conflict with disparate treatment.[228] The Court had previously ruled in essence that the antisubordination principle could not motivate a constitutional decision,[229] but had not suggested that law effectuating that principle could itself be discriminatory against the dominant groups. That has changed now.[230] While the decision was formally about Title VII only, and thus amenable to statutory resolution, the reasoning applies equally well to a future equal protection claim,[231] making the future of disparate impact itself uncertain.[232] Justice Scalia stated as much in his concurrence.[233]

This has two main consequences for data mining. First, where the internal difficulties described above can be resolved, legislation that requires or enables such resolution may run afoul of *Ricci*. Suppose Congress amended Title VII to require that employers make their training data and models auditable. In order to correct for detected biases in the

[225] *Id.*

[226] *Id.*

[227] *Id.*

[228] Primus, *supra* note 74, at 1343; Lawrence Rosenthal, *Saving Disparate Impact*, 34 CARDOZO L. REV. 2157, 2162-63 (2013).

[229] *See* Washington v. Davis, 426 U.S. 229, 239 (holding that discriminatory purpose is necessary to finding a violation of equal protection).

[230] Primus, *supra* note 74, at 1343.

[231] *Id.* at 1354-55 ("Despite the Court's professed intention to avoid equal protection issues, the *Ricci* premise is properly understood as a constitutional proposition as well as a statutory one. . . . [T]he problem is squarely put. If administering the disparate impact doctrine would be a disparate treatment problem but for the statutory carve-out, it is also an equal protection problem.").

[232] *Id.* at 1385-87.

[233] *Ricci* 557 U.S. at 594 (Scalia, J. concurring)("I . . . write separately to observe that its resolution of this dispute merely postpones the evil day on which the Court will have to confront the question: Whether, or to what extent, are the disparate-impact provisions of Title VII of the Civil Rights Act of 1964 consistent with the Constitution's guarantee of equal protection?")

training data that result in a model with a disparate impact, for example, the employer would first have to make membership in the protected class a consideration in the analysis. The remedy is inherently race-conscious. The *Ricci* Court did hold that an employer may tweak a test during the "test-design stage," however.[234] So, as a matter of timing, data mining might not formally run into *Ricci* if the bias resulting in a disparate impact is corrected before applied to individual candidates. After an employer begins to use the model to make hiring decisions, only a strong basis in evidence that the employer will be successfully sued for disparate impact will permit corrective action. Of course, unless every single model used by an employer is subject to a pre-screening audit (an idea that seems so resource-intensive that it is effectively impossible[235]), the disparate impact will be discovered only when the employer faces complaints. Additionally, while *Ricci*'s holding was limited in scope, the "strong basis in evidence" standard did not seem to be dictated by the logic of the opinion, which illustrated a more general conflict between disparate treatment and disparate impact, likely to come to a head in a later case.

Second, where the internal difficulties *cannot* be overcome, there is likely no way to correct for the discriminatory outcomes aside from some results-focused balancing, and requiring this will pose constitutional problems. For those who adhere to the nondiscrimination principle alone, such an impasse may be perfectly acceptable. They might say that as long as employers are not intentionally discriminating on the basis of explicitly proscribed criteria, the chips should fall where they may. To those that believe some measure of substantive equality is important over and above procedural equality, this result will be deeply unsatisfying.

Another answer is to reexamine the purpose of anti-discrimination law. The major justification for reliance on formal disparate treatment is that prejudice is simply irrational and thus unfair. But if an employer knows that his model has a disparate impact, but it is also his most predictive, the argument that the discrimination is irrational loses any force. Thus, data mining may require us to reevaluate why and whether we care about not discriminating.

[234] *Id.* at 585 (majority opinion)("Title VII does not prohibit an employer from considering, before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race. And when, during the test-design stage, an employer invites comments to ensure the test is fair, that process can provide a common ground for open discussions toward that end.")

[235] If the EEOC were the auditor, it would itself need a great deal of resources. If the EEOC relied on private auditors, someone would have to audit the auditors to make sure their methods for detecting disparate impact were sound. Audits are complicated, even leaving aside the difficulties of determining what a sound analysis would be in the first instance.

Consider a another example involving tenure predictions, one in which an employer ranks potential employees with the goal of hiring only those applicants that the company expects to retain for longer periods of time. In optimizing its selection of applicants in this manner, the employer may unknowingly discriminate against women if the historical data demonstrates that they leave their positions after fewer years than their male counterparts do. If sex accounts for a sufficiently significant difference in the tenure of employees, data mining will generate a model that simply discriminates on the basis of sex or those criteria that happen to be proxies for sex. Although selecting applicants with an eye to retention might seem both rational and reasonable, granting significance to predicted tenure would nevertheless result in women being subject to systematic disadvantage if sex accounts for a good deal of the difference in tenure. If that is the case, any data mining exercise that attempts to predict tenure will invariably rediscover this relationship. Which raises the obvious question: should the law permit a company to hire no women at all—or none that it correctly predicts will depart following the birth of a child—because it is the most rational choice according to their model?

If not, why not? On what basis should the law object to rational decisions, taken on the basis of seemingly legitimate criteria, which place members of protected classes at systematic disadvantage? Here, shielding members of protected classes from less favorable treatment is not justified on the basis of combatting prejudice or stereotyping. The only escape from this situation is one in which the relevance of sex is purposefully ignored and all factors correlated with sex are suppressed, the outcome of which is a necessarily less accurate model. The justification for placing restrictions on employers, and limiting the effectiveness of their data mining, would have to depend on an entirely different set of arguments than those advanced to explain the wrongfulness of biased data collection, poorly labeled examples, or an impoverished set of features. In other words, any prohibition in this case could not rest on a procedural commitment to ensuring ever more accurate determinations; instead, the prohibition would have to rest on a substantive commitment to equal representation of women in the workplace.

One solution to cases involving tenure predictions could be for Congress to amend Title VII to reinvigorate strict business necessity.[236] This would allow a court to accept that relying on tenure is rational but not strictly "necessary," and that perhaps other factors could make up for the lack of predicted tenure. Unfortunately, the normative justification for any

[236] Remember that if there is disparate impact, but no liability, it is because the goal was deemed job-related or satisfied business necessity.

such solution would have to depend on the antisubordination principle, as there is still no prejudice to combat, by stipulation. And the more disparate impact doctrine is thought to embody the antisubordination principle—as opposed to the "evidentiary dragnet" in service of the nondiscrimination norm—the more vulnerable it will become to constitutional challenge.[237] Of course, the farther the doctrine gets from substantive remediation, the less utility it has in remedying these kinds of discriminatory effects.[238]

This raises a point about disparate *treatment* and data mining. Within data mining, the prohibition on the use of certain information exists on a spectrum of two extremes. On one end, the prohibition has no effect because either the information is redundantly encoded or the results do not vary along lines of protected class. On the other end, the prohibition reduces the accuracy of the models. That is, if protected class were included as a feature, it would alter the results, presumably by making members of protected classes worse off. Thus, as a natural consequence of data mining, a command to ignore certain data has either no effect, or the effect of raising the fortunes of those protected classes in substantive ways.[239] Therefore with respect to data mining, due to the zero-sum nature of a ranking system, even disparate treatment doctrine is a reallocative remedy similar to affirmative action.[240] Once again, this erodes the legitimate rationale for supporting a nondiscrimination principle but holding fast against antisubordination in this context. They accomplish the same thing, but one is less effective at achieving substantive equality.

This reveals that the pressing challenge does not lie with ensuring procedural fairness through a more thorough stamping out of prejudice and bias, but rather with developing ways of reasoning that can help adjudicate when and what amount of disparate impact is tolerable. Abandoning a belief in the efficacy of procedural solutions leaves policymakers in an awkward position because there is no definite or consensus answer to questions about the fairness of specific outcomes. These need to be worked out on the basis of different normative principles. At some point, society will be forced to acknowledge that this is really a discussion about what constitutes a tolerable level of disparate impact in employment. Under the current constitutional order and in the political climate, it is tough to even imagine having such a conversation. But, until that happens, data mining will be permitted to exacerbate existing inequalities in difficult-to-counter ways.

[237] Primus, *supra* note 77, at 536-37;

[238] *Id.* at 537.

[239] *See* text accompanying *infra* note 79.

[240] For an argument that this is true more generally, see Bagenstos, *supra* note 73.

CONCLUSION

This Article has identified two types of discriminatory outcomes from data mining: a family of outcomes where data mining goes "wrong," and one outcome where it goes too "right." Data mining can go wrong in any number of ways: by choosing a target variable that correlates to protected class more than others would, by injecting current or past prejudice into the decision about what makes a good training example, by choosing too small a feature set, or by not diving deep enough into each feature. Each of these potential errors is marked by two facts: an ex-post determination has been made that the overall outcome is unfair and at least one seemingly nondiscriminatory choice made by a data miner has created a disparate impact. Where data mining goes "right," the data miners could not have been any more accurate given the starting point of the process; it is that very accuracy, exposing an uneven distribution of the attributes that predict the target variable, that gives such a result its disparate impact.

By now it should be clear that Title VII, and very likely other similarly process-oriented civil rights laws, cannot effectively address this situation. This means something different for the two families, and it should be slightly more surprising for the former. At a high level of abstraction, where a decision process goes "wrong" and this wrongness creates a disparate impact, Title VII and similar civil rights laws should be up to the task of solving the problem; that is ostensibly their entire purpose. But aside from a few more obvious cases like Street Bump, it is quite difficult to determine ahead of time what "correct" data mining looks like. As explained earlier, a decisionmaker can rarely discover that a particular choice of target variable or features is more discriminatory than other choices would have been until after the fact. And even though ensuring representative samples before training the model is a possibility, the data may never be perfect, and determining whether the marginal cost for more data is worthwhile also requires an ex-post analysis of discriminatory results. Thus, even at this level of abstraction, it becomes clear that holding the decisionmakers responsible for these disparate impacts is at least partly troubling from a due process perspective, and may counsel against using data mining in the first place.

If liability for getting things "wrong" is difficult to imagine, how does liability for getting things "right" make any more sense? That proxy discrimination largely rediscovers pre-existing inequalities suggests that perhaps Title VII is not the appropriate vehicle for addressing it. If what is at stake are the results of decades of historical discrimination and wealth concentration that have created profound inequality in society, is that not too big a problem to remedy through individual lawsuits, assuming affirmative action and similar policies are off the table? Thus, perfect data mining forces the question: If employers can say with certainty that, given the status quo,[241] candidates from protected classes are on average less ready for certain jobs than more privileged candidates, why should employers specifically be penalized for hiring them in lesser proportions?

Assuming, then, that Title VII alone cannot solve these problems, where should society look for answers? Well, the first answer is based on the caveat in the last paragraph: "given the status quo." Data mining takes the existing state of the world as a given, and ranks candidates according to their predicted attributes in *that* world. Data mining, by its very nature, treats the target variable as the only item that employers are in a position to alter; everything else under consideration that happens to correlate with different values for the target variable are assumed stable. But there are many reasons to question these background conditions. Sorting and selecting individuals on the basis of their apparent qualities hides the fact that the predicted effect of possessing these qualities with respect to a specific outcome is also a function of the conditions under which these decisions are made. Recall the tenure example from Part III.B. In approaching appropriate hiring practices as a matter of selecting the "right" candidates at the outset, an employer will fail to recognize potential changes that he could make to workplace conditions. A more family-friendly workplace, greater on-the-job training, or a workplace culture more welcoming to historically under-represented groups could affect the course of employees' tenure and their long-term success in ways that undermine the seemingly prophetic nature of data mining's predictions.

Society should also explore other nonlegal solutions. As a problem of technology, perhaps there are technical fixes. The first family of problems is rightly the subject of ongoing research in this area.[242] Additionally, technical solutions will have legal implications and future legal solutions may require developments in the technology, such as easily auditable structures. Researchers should be thinking jointly about applying a combination of law and technology to these problems. As to the second family, because the problem is at root one of inequality, social welfare policy seems the most salient solution. Here too, data can help. For example, comparing the performance of equally qualified candidates across workplaces might allow a reformer to use data to discover the effect of workplace policies so that they can be reformed.

Ideally, institutions can find a way to trust and use data mining for

[241] We cannot stress enough the import of these caveats. Certainty is a strong and unlikely precondition and the status quo should not be taken as a given outside of data mining.

[242] Kamiran, Calders, & Pechenizkiy, *supra* note 65.

the good it brings in terms of generating new knowledge and formalizing decisionmaking. There are surely good uses for data mining in the service of increasing equality. But where data mining is only brought to bear for the purpose of optimizing selection procedures, the potential to exacerbate inequality cannot be ignored.