Oklahoma Law Review

Volume 68 | Number 3

2016

"Big Data" and the Risk of Employment Discrimination

Allan G. King
Littler Mendelson, agking@littler.com

Marko Mrkonich
Littler Mendelson, mmrkonich@littler.com

"BIG DATA" AND THE RISK OF EMPLOYMENT DISCRIMINATION

ALLAN G. KING* & MARKO J. MRKONICH**

I. Introduction

Our digital footprints are massive, and the volume of digitized information is increasing rapidly. As a consequence, an industry has emerged to develop the commercial value of this information and to mine data in ways that assists employers in identifying, recruiting, retaining, and rewarding the most promising employees. This industry is referred to as "Big Data." Its promise lies in its ability to gather and sift through extraordinary amounts of information and arrive at criteria for identifying these sought-after individuals. Moreover, Big Data utilizes methods that largely eliminate discretion, and unconscious bias, from the selection process.[1] Despite these substantial benefits, Big Data also presents risks an employer should consider before adopting its methods for employee selection. This paper identifies the most prominent of these risks and suggests how they might be managed by employers and assessed by the courts as well as the government.

Just as volume of information is the hallmark of Big Data, so too is the industry's insistence that discovering correlations—rather than understanding cause-and-effect relationships—is the most efficient way to address and solve social and scientific questions. For example, Professor Viktor Mayer-Schonberger and Kenneth Cukier, Data Editor at *The Economist*, declare, "Causality won't be discarded, but it is being knocked off its pedestal as the primary fountain of meaning. Big data turbocharges non-causal analyses, often replacing causal investigations."[2]

In place of causation, Big Data relies on "predictive analytics . . . to foresee events before they happen."[3] Based on algorithms derived from vast amounts of data, predictive analytics can be used to identify hit songs,

* Allan G. King is a shareholder in the Austin, Texas, office of Littler Mendelson, P.C. He holds a B.A. from City College of New York, a Master's and Ph.D. from Cornell University, and J. D. from the University of Texas.

** Marko J. Mrkonich is a shareholder in the Minneapolis, Minnesota, office of Littler Mendelson, P.C. He holds a B.A. and J.D. from Harvard University.

1. *But see* Claire Cain Miller, *Algorithms May Echo Human Bias, Studies Find*, N.Y. TIMES, July 13, 2015, at B1.

2. VIKTOR MAYER-SCHONBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 68 (Eamon Dolan ed., 2013).

3. *Id.* at 58.

forecast structural failures in bridges, or merely track the data patterns these events produce. Just as predictive analytics might identify a future hit tune, Big Data claims similar methodologies can spot promising job candidates.

But the correlative methods of Big Data are in tension with Title VII, the Age Discrimination in Employment Act (ADEA), and the Americans with Disabilities Act (ADA) to the extent the correlations those methods discover overlap with protected employee characteristics. Although Big Data may be a potent antidote to intentional discrimination, antidiscrimination laws also prohibit practices that have a disparate impact unless those practices are job-related and consistent with business necessity.[4] The tension arises because Big Data may incorporate information far removed from the workplace, its value lying in the correlation it discerns between non-work-related data and various measures of job performance.[5] Accordingly, the affirmative defense of job-relatedness protects Big Data methodologies only to the extent that courts countenance criteria that are not directly job-related but instead correlate with job performance.

Whether a selection method that produces an adverse impact passes muster under Title VII is often decided with reference to the Uniform Guidelines on Employee Selection Procedures (Uniform Guidelines).[6] The critical inquiry is whether the selection procedure is a "valid" predictor of success on the job.[7] Because the Uniform Guidelines date from 1978, they do not contemplate Big Data's reliance on correlation rather than cause-and-effect relationships. This Article explores the legal significance of the difference between correlative and cause-and-effect methodologies and suggests that the Uniform Guidelines may not serve their intended purpose when confronting the methodology of Big Data.

The Uniform Guidelines require employers to consider whether there are less discriminatory alternatives to any selection procedure, whereas Title VII assigns this burden of proof to the plaintiff.[8] Although Title VII cases

4. 42 U.S.C. §2000e-2(k) (2012). The ADEA requires proof that the practice is a "reasonable factor other than age." 29 C.F.R. § 1625.7 (2015).

5. *See infra* notes 14-17.

6. Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607 (2015).

7. *See* Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

8. *Compare* Ricci v. DeStefano, 557 U.S. 557, 632-33 n.11 (2009) (Ginsburg, J., dissenting) ("Under the Uniform Guidelines on Employee Selection Procedures (Uniform Guidelines), employers must conduct 'an investigation of suitable alternative selection procedures.' 29 CFR § 1607.3(B).") *with* 42 U.S.C. § 2000e-2(k) (2012). *See Ricci*, 557 U.S. at 578 ("[A] plaintiff may still succeed by showing that the employer refuses to adopt an

rarely turn on this element, Big Data may elevate the importance of less discriminatory alternatives. Because Big Data derives its algorithms from vast troves of data, which computers combine and weigh in innumerable ways to select the optimal solution, there will typically be a host of near-optimal alternatives, each differing slightly in terms of its impact on protected groups and its ability to identify superior employees. Courts must then decide whether an algorithm's marginally greater predictive ability is sufficient to justify its greater adverse impact, if indeed the law recognizes any trade-off between the two.

The ADA raises additional issues. Not all segments of the population are equally likely to leave footprints in places searched by Big Data. For example, a Big Data algorithm that tracks online book purchases may misconstrue the reading habits of sight-impaired individuals who may find electronic media less accessible than print. If the algorithm correlates electronic media purchases with positive job performance, this means the algorithm excludes sight-impaired persons and places the often-unknowing employer in jeopardy. Further, the ADA requires employers to modify "examinations, training materials or policies, the provision of qualified readers or interpreters, and other similar accommodations."[9] An employer, however, can only accommodate disabilities of which it is aware. Yet disabled applicants, not knowing the activities and behaviors on which they are being assessed, have no reason to request an accommodation.

Employers who rely on Big Data to assist in hiring, promoting, or otherwise evaluating employees are liable for any disparate impact even though third-party providers may have developed such methods.[10] But these algorithms and their development are typically proprietary and confidential trade secrets the developer is eager to protect.[11] Consequently, the employer

---

available alternative employment practice that has less disparate impact and serves the employer's legitimate needs.") (citing §§ 42 U.S.C. 2000e-2(k)(1)(A)(ii), (C)).

9. 42 U.S.C.A. §§ 12111(9)(B), 12112(a),(b)(5)-(6) (West 2013).

10. *See* 42 U.S.C. § 2000e-2(a)-(c) (2012) (proscribing discriminatory conduct by employers, employment agencies, and labor organizations, not test developers).

11. *See, e.g.*, EEOC v. Aon Consulting, Inc., 149 F. Supp. 2d 601, 608-09 (S.D. Ind. 2001) (ordering a limited protective order where it was important that the "exceptionally sensitive information [of] employment-related tests, be kept confidential from the charging party" and the subpoenaing agency had "not shown that existing statutory and regulatory procedures and protections offer sufficient protection"); EEOC v. C&P Telephone Co., 813 F. Supp. 874, 876 (D.D.C. 1993) (permitting a limited confidentiality agreement to protect employment tests upon "find[ing] that respondents have an extremely strong interest in protecting the subpoenaed information," and "the [agency]'s internal procedures are insufficient" to protect it from improper dissemination).

using Big Data confronts a "black box," about which much is claimed but little is known. The concluding section of this Article discusses how employers might reap the benefits of Big Data while minimizing its risks.

II. What Is "Big Data"?

Merriam-Webster defines "big data" as "an accumulation of data that is too large and complex for processing by traditional database management tools."[12] But Big Data is transcendent in its scope as well as its size. Not only does it include data employers create in the normal course of its business, such as time and attendance, payroll, and performance data, but also *external* databases that supplement what the employer already knows about its employees.[13]

An article in *The Atlantic* describes how one company searched for software engineers proficient in writing computer code.

> The company's algorithms begin by scouring the Web for any and all open-source code, and for the coders who wrote it. They evaluate the code for its simplicity, elegance, documentation, and several other factors, including the frequency with which it's been adopted by other programmers. For code that was written for paid projects, they look at completion times and other measures of productivity. Then they look at questions and answers on social forums such as Stack Overflow, a popular destination for programmers seeking advice on challenging projects. They consider how popular a given coder's advice is, and how widely that advice ranges.
>
> The algorithms go further still. They assess the way coders use language on social networks from LinkedIn to Twitter; the company has determined that certain phrases and words used in association with one another can distinguish expert programmers from less skilled ones. [The company] knows these phrases and words are associated with good coding because it can correlate them with its evaluation of open-source code, and with the

12. *Big Data*, MERRIAM-WEBSTER, http://www.merriam-webster.com/dictionary/big%20data (last visited Nov. 15, 2015).

13. *See* Steve Lohr, *Big Data, Trying to Build Better Workers*, N.Y. TIMES, Apr. 21, 2013, at BU4 ("Work-force science, in short, is what happens when Big Data meets H.R.").

> language and online behavior of programmers in good positions at prestigious companies.[14]

The article explains that information can be extrapolated to programmers whose code is not available on the Internet by comparing their online histories to those of the best open-source programmers and determining whether their social media footprints are similar. For instance, one company's chief scientist explained that these correlations are "not all obvious, or easy to explain. . . . [O]ne solid predictor of strong coding is an affinity for a particular Japanese manga site."[15]

On January 19, 2015, *The New York Times* reported that new entrants to the consumer-lending industry are utilizing Big Data to predict creditworthiness.[16] "[T]hey may look to see if potential customers use only capital letters when filling out forms, or at the amount of time they spend online reading terms and conditions—and not so much at credit history."[17] Although the Fair Credit Reporting Act regulates the use of credit scores to assess applicants,[18] *correlates* of credit scores are not subject to the same restrictions and may serve as proxies.

A complementary approach elicits responses to particular questions that are not necessarily job-related but highly correlated with behavior relevant to the job. For example, *The Wall Street Journal* reported how Xerox cut attrition in its call centers by twenty percent in a six-month period: it first developed a profile for the ideal employee.[19] "The data say [the ideal employee] lives near the job, has reliable transportation and uses one or more social networks, but not more than four. He or she tends not to be

---

14. Don Peck, *They're Watching You at Work*, ATLANTIC (Dec. 2013), http://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/.

15. *Id.* Manga is "a Japanese comic book or graphic novel." *Manga*, MERRIAM-WEBSTER, http://www.merriam-webster.com/dictionary/manga (last visited Nov. 15, 2015).

16. Steve Lohr, *Creditworthy? Let's Consider Capitalization*, N.Y. TIMES, Jan. 19, 2015, at A1.

17. *Id.*

18. Miller v. Countrywide Bank, N.A., 571 F. Supp. 2d 251 (D. Mass. 2008) (discussing disparate-impact theory applied to lending practices); *see also* EEOC v. Kaplan Higher Educ. Corp., 790 F. Supp. 2d 619 (N.D. Ohio 2011) (discussing employment applicant's disparate-impact claim based upon credit history), aff'd, 748 F.3d 749 (6th Cir. 2014). The Fair Credit Reporting Act (FCRA or Act) requires notice to any consumer subjected to "adverse action . . . based in whole or in part on any information contained in a consumer [credit] report . . . ." 15 U.S.C. § 1681m(a) (2012); *see also* Safeco Ins. Co. of Am. v. Burr, 551 U.S. 47, 53 (2007).

19. Joseph Walker, *Meet the New Boss: Big Data*, WALL ST. J., (Sept. 20, 2012, 11:16 AM), http://www.wsj.com/articles/SB10000872396390443890304578006252019616768.

overly inquisitive or empathetic, but is creative."[20] Xerox then administers a thirty-minute test that presents applicants with situations they might encounter on the job and asks the candidate to respond to statements such as "'I ask more questions than most people do' and 'People tend to trust what I say.'"[21] Based on how an algorithm evaluates the responses, it sorts the applicants into three categories, and the company then seeks those candidates from the top tier.[22]

There are numerous other ways applicant data can be combined with external databases to construct profiles of successful employees, but these few examples are useful in highlighting the employment issues raised by Big Data.

III. Correlation Versus Causation

The methodology underlying the above illustrations is purely, and proudly, correlational. That is, the researcher determines what reasonably accessible information about employees or applicants correlates with the traits of successful employees. This is apparent from the observation that visitors to Japanese manga sites were often good coders.[23] No one is likely to argue that individuals enhance their coding skills by spending time on those sites—surely the majority of visitors have the same coding aptitude before and after their visits. But for whatever reason, those with an aptitude for coding share an appreciation for manga. This is a case in which two traits, an appreciation of manga and coding aptitude, are correlated, but neither causes the other.

This method contrasts with more traditional selection procedures. Each year professional football teams ask the most promising college players to attend a "scouting combine."[24] Players who attend are graded and evaluated regarding speed, strength, and skills that football teams believe translate directly into success on the playing field.[25] Rather than reviewing a prospect's internet search history, the teams assess precisely those skills they believe are most salient.[26] In the coaches' minds, the relationship

20. *Id.*
21. *Id.*
22. *Id.*
23. Peck, *supra* note 14.
24. *See Combine Prep: A Typical Training Day for an NFL Prospect*, NFL (Feb. 20, 2015, 6:02 PM), http://www.nfl.com/news/story/0ap3000000472355/article/combine-prep-a-typical-training-day-for-an-nfl-prospect.
25. *Id.*
26. *Id.*

between performance at this combine and success on the field reflects a causal relationship.

Despite these obvious and straightforward ways to assess athletic ability, some researchers utilize a correlational approach and proclaim its success. On December 26, 2014, *The New York Times* reported one company's efforts to identify an athlete's "emotional DNA" with facial analysis that tracks "which of the 43 muscles in the face are working at any moment."[27] Facial coding, as it is called, claims to identify seven core emotions—happiness, surprise, contempt, disgust, sadness, anger, and fear—it contends are correlated with an athlete's on-field performance.[28] The article recounts how professional sports teams have relied on these categorizations in determining which prospects to draft.[29] Although no one believes facial features affect football performance, the belief is that facial characteristics reflect emotional states, and emotional states affect athletic performance.

The difference between correlation and causation is illustrated by comparing the manga test and the scouting combine. No reasonable applicant believes he or she could improve coding skills by spending more time on manga sites, as opposed to practicing code. On the other hand, an aspiring football player *would* be well advised to practice diligently to increase his speed because more speed *makes* him a better player.

Whether a purely correlational approach is superior to one based on cause and effect is hotly debated in the Big Data industry, but to employers the answer should also reflect legal risks and defenses.

IV. Why Causation Matters

In 2008, the editor of *Wired* declared the scientific method, based on identifying and testing cause-and-effect relationships, was obsolete.[30] In its place, he endorsed "Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required."[31] As an example, consider how "the American retailer, Target, upset a Minneapolis man by knowing more about his teenage daughter's sex life than he did. Target was able to predict his daughter's pregnancy by monitoring her

27. Kevin Randall, *Teams Turn to a Face Reader, Looking for that Winning Smile*, N.Y. TIMES, Dec. 26, 2014, at A1.

28. *Id.*

29. *Id.*

30. Chris Anderson, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, WIRED (June 23, 2008, 12:00 PM), http://www.wired.com/2008/06/pb-theory/.

31. *Id.*

shopping patterns and comparing information to an enormous database detailing billions of dollars of sales."[32] There was no need to understand why pregnant women purchased particular items; it was enough to know that pregnancy and these purchases were correlated.

Indeed, it is unnecessary to understand why two or more things are correlated in many instances. One can rely on a rooster's crowing to know the sun is up without knowing what exactly prompts the rooster to let loose. Some have suggested that cows lie down when inclement weather is approaching, making that behavior a reliable forecast as well.[33] Similarly, we can anticipate that darkening clouds will produce rain, although few have an understanding of why moisture-laden clouds are gray. Because these associations are regular and predictable, we rely on them without understanding why the relationships exist.

But sometimes it is important to know why. First, an employer must understand causal relationships in situations where it must effectuate change. For example, federal contractors are subject to Executive Order 11,246, which requires them to engage in affirmative-action practices to reduce and eliminate differences between the utilization and the availability of women and minorities in the relevant labor market.[34] In effect, they must *modify* the correlation between their past practices and the results they produced.

Second, correlations are only useful if they are persistent, in the sense they remain fairly constant from day-to-day or month-to-month. Returning to our example of the manga website favored by ace programmers, visits to that site may have some predictive power—but only until the next hot manga site appears. Without warning, the site *du jour* may change, leaving pedestrian programmers behind after the circle of elite programmers migrates to the trendy new site. If a company is unaware of this change, and continues to recruit based on outdated preferences, it may be surprised to learn the expected correlation has been inverted: hiring visitors to the abandoned manga site now results in a group of coders drawn from the bottom of the barrel.

Moreover, researchers and employers can only determine the effective life of a correlation retrospectively. A flick of a light switch reliably turns on a light hundreds of times until the light bulb burns out. Someone who

32. Mark Graham, *Big Data and the End of Theory?*, GUARDIAN (Mar. 9, 2012, 9:39 AM), http://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory.

33. *Cows Lying Down Before Rain? It Stands Up*, TIMES (May 14, 2013, 12:01 AM), http://www.thetimes.co.uk/tto/weather/article3713025.ece.

34. 41 C.F.R. § 60-2.15 (2014).

relies on correlation alone has no understanding of why the correlation no longer holds or that the correlation can be restored by changing the bulb. On the other hand, someone with just the flimsiest understanding of cause and effect will quickly replace the bulb and solve the problem. In short, the distinction between causation and correlation is the difference between understanding and merely observing.

Finally, much of the legal system is premised on cause-and-effect relationships.[35] "Causation" is an essential element of many causes of action, including, ironically, disparate-impact discrimination.[36] The trilogy of cases that set the standard for admitting scientific evidence was focused on whether the scientific testimony sufficed to prove causation.[37] Defendants are incarcerated, and indeed put to death, because their actions "caused" a particular consequence. Thus, Big Data must win the uphill battle of persuading a judiciary steeped in causation-laden jurisprudence that, in a Big Data world, correlation—not causation—is what ultimately matters.

V. Big Data and the Risk of Employment Discrimination

Given the complexity of amassing and then analyzing vast quantities of information, an employer probably would not reverse engineer the process in order to intentionally discriminate against a protected group.[38] It is far more probable that Big Data will be challenged because it unintentionally

35. David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, *in* FED. JUDICIAL CTR. & NAT'L RESEARCH COUNCIL, REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 211, 262 (3d ed. 2011) [hereinafter REFERENCE MANUAL ON SCIENTIFIC EVIDENCE] ("Researchers—and the courts—are usually more interested in causation. Causation is not the same as association.").

36. Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 994 (1988) ("Once the employment practice at issue has been identified, causation must be proved; that is, the plaintiff must offer statistical evidence of a kind and degree sufficient to show that the practice in question has caused the exclusion of applicants for jobs or promotions because of their membership in a protected group."). Thus, a plaintiff complaining of Big Data must allege that a practice based strictly on correlations, and eschews proof of causation, nevertheless caused her injury. In response, a Big Data defendant may urge dismissal of this claim because the plaintiff has failed to prove her injury was caused by a methodology that disparages the relevance of causation.

37. *See* Daubert v. Merrill Dow Pharm., 509 U.S. 579 (1992); Gen. Elec. Co. v. Joiner, 522 U.S. 136 (1997); Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1998).

38. For example, Clair C. Miller, writing in the *New York Times*, quotes a researcher who studied bias in algorithms and suggests building alogrithms from scratch, to avoid building in bias. Similarly, building in bias also may require starting from scratch because bias is by no means inherent in each and every algorithm. *See supra* note 1.

yields a disparate impact on one or more protected groups. More precisely, a plaintiff or class would allege that the algorithm used for hiring, promotion, or similar purpose adversely impacts one or more protected groups. Let us consider how that case would proceed and the issues that would likely arise.

The plaintiff in a Title VII disparate-impact-discrimination case must (1) identify with particularity the facially neutral practice being challenged, (2) demonstrate that the practice adversely impacts members of the protected group in question, and (3) show that the practice caused the plaintiff to suffer an adverse employment action.[39] If the plaintiff meets that burden, the employer may defend by demonstrating that the practice in question is job-related and consistent with business necessity.[40] When that neutral practice is a test or similar screening device, courts typically require employers to establish the "validity" of the screen.[41] Finally, if the employer establishes this defense, the plaintiff may prevail by proving there is a less discriminatory alternative that similarly serves the employer's needs, but which the employer refuses to adopt.[42]

Cases arising under the ADEA[43] would proceed somewhat differently. As with Title VII, plaintiffs must plead (1) a specific and actionable policy, (2) a disparate impact, and (3) facts raising a sufficient inference of causation.[44] If the plaintiff makes this showing, the employer's burden is to demonstrate that its challenged practice is based on "reasonable factors other than age" (RFOA defense).[45] Accordingly, to avoid liability once an ADEA plaintiff has proved a prima facie case, the employer must establish the reasonableness of its reliance on other neutral criteria.[46]

---

39. 42 U.S.C. §2000e-2(k) (2012); Ricci v. DeStefano, 557 U.S. 557, 578 (2009).

40. *Ricci*, 557 U.S. at 578.

41. *Id.*; *see, e.g.*, Johnson v. City of Memphis, 770 F.3d 464, 478 (6th Cir. 2014) ("The City may meet its . . . burden by showing through 'professionally acceptable methods, [that its testing methodology is] predictive of or 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 Black Law Enf't Officers Ass'n v. City of Akron, 824 F.2d 475, 480 (6th Cir. 1987)).

42. 42 U.S.C. § 2000e-2(k).

43. 29 U.S.C. §§ 621-634 (2012).

44. *See* Smith v. City of Jackson, 544 U.S. 228, 241 (2005).

45. 29 U.S.C. § 623(f)(1).

46. *Smith*, 544 U.S. at 239 ("It is, accordingly, in cases involving disparate-impact claims that the RFOA provision plays its principal role by precluding liability if the adverse impact was attributable to a nonage factor that was 'reasonable.'").

Neither the ADEA nor the Supreme Court has explained what an employer must prove to establish the RFOA defense. The EEOC, however, has elucidated the following:

> To establish the RFOA defense, an employer must show that the employment practice was both reasonably designed to further or achieve a legitimate business purpose and administered in a way that reasonably achieves that purpose in light of the particular facts and circumstances that were known, or should have been known, to the employer.
>
> (2) Considerations that are relevant to whether a practice is based on a reasonable factor other than age include, but are not limited to:
>
> (i) The extent to which the factor is related to the employer's stated business purpose;
>
> (ii) The extent to which the employer defined the factor accurately and applied the factor fairly and accurately, including the extent to which managers and supervisors were given guidance or training about how to apply the factor and avoid discrimination;
>
> (iii) The extent to which the employer limited supervisors' discretion to assess employees subjectively, particularly where the criteria that the supervisors were asked to evaluate are known to be subject to negative age-based stereotypes;
>
> (iv) The extent to which the employer assessed the adverse impact of its employment practice on older workers; and
>
> (v) The degree of the harm to individuals within the protected age group, in terms of both the extent of injury and the numbers of persons adversely affected, and the extent to which the employer took steps to reduce the harm, in light of the burden of undertaking such steps.[47]

Returning to the plaintiff's perspective, under either Title VII or the ADEA, a plaintiff who claims she was rejected because a Big Data algorithm rated her poorly must first identify the algorithm and/or the data

47. 29 C.F.R. § 1625.7(e) (2015).

input into that algorithm.[48] It is probably asking too much, however, to require the plaintiff to identify the specific features of the challenged algorithm that cause the alleged disparate impact. Employers who utilize Big Data might develop their own algorithms; but, more often than not, the employer will contract with a vendor who provides Big Data services—putting those formulas further out of a plaintiff's reach.[49] Moreover, these algorithms are referred to as Big Data precisely because of their complexity. To require the plaintiff to unscramble this complexity and identify an offending bit of data or code would contravene the law surrounding disparate-impact challenges to employment tests, which generally allows plaintiffs to challenge a test as a whole rather than identify specific test questions.[50]

Moreover, correlations can behave oddly when relationships are analyzed piecemeal. For example, two traits, A and B, may be entirely unrelated when the simple correlation between them is measured. On the other hand, when trait C is considered as well, the relationship between A and B may become pronounced and significant.[51] Accordingly, Big Data may provide an instance in which the disparate impact associated with each

48. The distinction between the algorithm and its formula is similar to the difference between a cake's recipe and its ingredients. The difference is aptly illustrated in *Ricci v. DeStefano*, 557 U.S. 557 (2009). At issue in that case was the examination administered to firefighters who were candidates for promotion to lieutenant and captain. The examination consisted of two parts: one written, the other oral. These portions were then weighted 60-40 to arrive at a total. This examination could have been challenged based upon a disparate impact caused by the oral or written portion of the test (the ingredients), or the weights assigned to each portion (the recipe).

49. *See, e.g.*, EEOC v. Kronos Inc., 694 F.3d 351 (3d Cir. 2012).

50. *Cf.* 42 U.S.C. § 2000e-2(k)(1)(B)(i) (2012) (permitting the elements of an "employment practice" to be analyzed as one, if they are incapable of separation for analysis); Briscoe v. City of New Haven, No. 3:09-cv-1642 (CSH), 2013 U.S. Dist. LEXIS 162116 (D. Conn. Nov. 14, 2013) (discussing amended complaint challenging disparate impact caused by test as a whole). *But see* Nash v. Consol. City of Jacksonville, 837 F.2d 1534, 1539 (11th Cir. 1988) ("The City's expert admitted that 27 of the 97 questions on the 1981 examination had an adverse impact on black applicants.").

51. Kaye & Freedman, *supra* note 35, at 262-63 ("The association between two variables may be driven by a lurking variable that has been omitted from the analysis. For an easy example, there is an association between shoe size and vocabulary among schoolchildren. However, learning more words does not cause the feet to get bigger, and swollen feet do not make children more articulate. In this case, the lurking variable is easy to spot—age. In more realistic examples, the lurking variable is harder to identify.") (citation omitted).

challenged practice cannot be separated for analysis because of this contextual effect.[52]

Once the algorithm is identified, the plaintiff then must prove this algorithm adversely impacts the protected group to which the plaintiff belongs.[53] Typically, this proof is made by determining whether the algorithm produces a significantly lower "pass-rate" for the protected group than for the majority.[54] The "pass-rate" is usually based on the "score" the algorithm assigns to applicants or candidates for promotion from each demographic group, or the rate at which each group exceeds the minimum "cut-score."[55]

But disparate impact cannot be gauged by merely studying the demographics of those hired or selected, i.e., the "bottom line."[56] Rather, the plaintiff must prove that the algorithm's output (e.g., its scores or categorizations) adversely impacted her protected group. This may require extensive discovery because an applicant may not at first know that Big Data was in the picture. Unless aspects of the data input into the algorithm qualify as a "consumer credit report" under the FCRA or constitute a "medical examination" subject to the ADA, an employer will probably not be required to disclose the specific reasons an applicant was rejected.[57] As a

---

52. *Cf.* 42 U.S.C. § 2000e-2(k)(1)(B).

53. Briscoe v. City of New Haven, 967 F. Supp. 2d 563, 590 (D. Conn. 2013) (dismissing disparate-impact claim because challenged policy did not adversely impact the protected group).

54. Ricci v. DeStefano, 557 U.S. 557, 587 (2009) (comparing pass-rate among black and white firefighters seeking promotion to captain).

55. Lanning v. SEPTA, 181 F.3d 478, 481 (3d Cir. 1999) ("[A] discriminatory cutoff score on an entry level employment examination must be shown to measure the minimum qualifications necessary for successful performance of the job in question . . . ."); United States v. Delaware, No. Civ. A. 01-020-KAJ, 2004 WL 609331, at *24 (D. Del. Mar. 22, 2004) (explaining that "minimum qualifications necessary" means "likely to be able to do the job"). As interpreted by the Seventh Circuit, this means that a cut score may satisfy the business necessity requirement if it is based on "a professional estimate of the requisite ability levels, or, at the very least by analyzing the test results to locate a logical 'break-point' in the distribution of scores." Gillespie v. Wisconsin, 771 F.2d 1035, 1045 (7th Cir. 1985).

56. Connecticut v. Teal, 457 U.S. 440, 452 (1981) ("In sum, respondents' claim of disparate impact from the examination, a pass-fail barrier to employment opportunity, states a prima facie case of employment discrimination under § 703(a)(2), despite their employer's nondiscriminatory 'bottom line,' and that 'bottom line' is no defense to this prima facie case under § 703(h).").

57. *See* 15 U.S.C. § 1681b (2012); 42 U.S.C. § 12112(d) (2012); *see, e.g.*, Leonel v. Am. Airlines, Inc., 400 F.3d 702, 709 (9th Cir. 2005) ("The ADA recognizes that employers may need to conduct medical examinations to determine if an applicant can perform certain

result, only through discovery may a plaintiff learn the true cause of an employment decision.

Although Big Data draws on a variety of data sources and algorithms, to the extent it relies on data created outside the workplace, its methodology seems prone to adversely impact *some* protected group. Individuals who share a common aptitude or skill may differ significantly in other aspects of their lives. Hypothetically, an algorithm that relies on a correlation between fast programmers and those that drive fast cars will reject applicants who are too poor to own a car, who live within walking distance of their usual haunts, or who have a disability that prevents them from driving. Nevertheless, the correlation between programming speed and fast cars could be substantial, based upon the behavior of the majority of that field. Therefore, should an employer rely on the algorithm to find speedy applicants, it does so to the actionable detriment of the disabled, yet similarly qualified, programmers.

Obvious differences also exist among age groups. Social media use, family responsibilities, or a working spouse may influence how equally proficient employees spend their incomes and leisure.[58] Correlational studies, however, are about the "rule" rather than the exceptions that might characterize one or more demographic groups. Indeed, statisticians dismiss those who do not fit the norm as "outliers."[59] Accordingly, a diverse population may consist of groups that differ markedly in their traits tracked by Big Data, yet are similar in their capacity to perform a job.

In practice, disparate impact comes down to statistics. A plaintiff establishes an adverse impact by demonstrating there is a statistically significant difference between a protected group and the majority regarding their scores or pass-rates on a test or screen.[60] However, this yardstick is

---

jobs effectively and safely. The ADA requires only that such examinations be conducted as a separate, second step of the selection process, after an individual has met all other job prerequisites . . . . When employers rescind offers made conditional on both non-medical and medical contingencies, applicants cannot easily discern or challenge the grounds for rescission. When medical considerations are isolated, however, applicants know when they have been denied employment on medical grounds and can challenge an allegedly unlawful denial.") (citation omitted).

58. *See, e.g.*, Adele Atkinson & David Hayes, *Consumption Patterns Among Older Consumers*, INT'L LONGEVITY CTR.—UK (Nov. 30, 2010), tbl. 3, http://www.ilcuk.org.uk/index.php/publications/publication_details/consumption_patterns_among_older_consumers_-_statistical_analysis.

59. Kaye & Freedman, *supra* note 35, at 262 ("The correlation coefficient can be distorted by outliers—a few points that are far removed from the bulk of the data.").

60. Jones v. City of Boston, 752 F.3d 38, 47 n.9 (1st Cir. 2014) (collecting cases).

flexible and, other things being equal, any disparity between two groups will increase in statistical significance the larger the sample on which the analysis is based.[61] Thus, Big Data may identify criteria that are predictive of success in terms of statistical significance; although, in practical terms, the difference between success and failure may be quite small.

Big Data pushes these statistical criteria to their limits and perhaps beyond. As more data is brought to bear on the selection process, disparities between demographic groups are bound to become increasingly statistically significant.[62] At the extreme, even differences most would find negligible may nevertheless exceed the "two standard deviation" criterion. One prominent example is the statistical analysis employed in *Wal-Mart Stores, Inc. v. Dukes*, which considered one of the largest data sets ever analyzed in an employment discrimination suit.[63] In his comparison of pay differences between men and women, the plaintiffs' expert reported that a standard deviation equaled just one-tenth of one percent.[64] The implication was that a gender difference in pay of just two-tenths of a percent—the difference between a male employee paid $10 per hour and a female paid $9.98 per hour—would be judged "statistically significant" in terms of discriminatory impact. When data sets grow to that size, statistical criteria risk trivializing the important question of what constitutes discrimination.[65]

After decades of increasing comfort and growing sophistication with statistical criteria, courts now have to confront the problem that the criteria for identifying discrimination honed in a small-data world may be unhelpful in a world of Big Data. Precisely because it is so "Big," Big Data makes it highly likely that any difference between demographic groups—no matter

---

61. Daniel L. Rubinfeld, *Reference Guide on Multiple Regression*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE, *supra* note 35, at 303, 318-19.

62. *Id.*

63. 131 S. Ct. 2541 (2011). In this suit, the nationwide class consisted of approximately 1.5 million female employees. *Id.* at 2544.

64. *Id.*

65. *See, e.g.*, Mark Kelson, *Significantly Misleading*, SIGNIFICANCE MAG. (Oct. 22, 2013), http://www.statslife.org.uk/the-statistics-dictionary/1000-the-statistics-dictionary-significantly-misleading ("Imagine if an environmentalist said that oil contamination was detectable in a sample of water from a protected coral reef. The importance of that statement would change drastically depending on whether they were referring to a naked-eye assessment of a water sample or an electron microscope examination. The smaller the amount of oil, the harder we would have to look. The same is true for a clinical study that detects a statistically significant treatment effect. If the study is huge, then issues of statistical significance become unimportant, since even tiny and clinically unimportant differences can be found to be statistically significant.").

how slight—will be statistically significant. A reasonable response by the courts may be to resurrect a rule of thumb—an arbitrary, but reasonable, threshold for determining when a disparity is of legal consequence.

Rules of thumb are common in age-discrimination litigation. For example, several circuits have declared that disparities in the treatment of employees who differ by less than five, six, or even eight years are not probative of discrimination.[66] Similarly, the Eighth Circuit has held that reductions in workforce that fail to reduce the percentage of workers aged forty and older by more than four percentage points are per se not discriminatory.[67] Although courts are receptive to statistical proof beyond those thresholds, these standards of proof reflect the view of many courts that, notwithstanding statistical significance, minimal differences lack probative value and should be ignored. More generally, perhaps it is time to revive the eighty-percent threshold of the Uniform Guidelines and recognize that, in the era of Big Data, statistical significance is the norm and thus a poor indicator of legal relevance.

Once the plaintiff establishes the adverse impact of the selection criterion, she must next prove the algorithm caused her to suffer an adverse employment action.[68] The question is whether, if the algorithm had valued this candidate more highly, the plaintiff would have been more likely to be selected.[69] This too can be established statistically by comparing the selection rate among those who score more favorably than the plaintiff with those who do not score more favorably.[70] If a statistically significant difference exists, a fact-finder may reasonably conclude that the algorithm caused an adverse employment action.[71]

---

66. *See, e.g.*, Aliotta v. Bair, 576 F. Supp. 2d 113, 125 n.6 (D.C. Cir. 2008) (stating age difference of seven years insignificant without further evidence showing age was a determining factor) (citing Dunaway v. Int'l Bhd. of Teamsters, 310 F.3d 758, 767 (D.C. Cir. 2002)); Grosjean v. First Energy Corp., 349 F.3d 332, 340 (6th Cir. 2003) (adopting bright-line rule that "in the absence of direct evidence that the employer considered age to be significant, an age difference of six years or less between an employee and a replacement is not significant"); Holowecki v. Fed. Express Corp., 644 F. Supp. 2d 338, 357-58 (S.D.N.Y. 2009), *aff'd*, 382 F. App'x 42 (2d Cir. 2010) (stating vague allegations of preferential treatment to someone three years younger is insufficient to give rise to inference of age discrimination as matter of law).

67. *See* Clark v. Matthews Int'l Corp., 639 F.3d 391, 399 (8th Cir. 2011).

68. *See* Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 995 n.3 (1988).

69. *Id.*

70. *Id.*

71. *Id.*

If a plaintiff makes this proof, the burden shifts to the employer to prove that the challenged algorithm is job-related for that position and consistent with business necessity.[72] Satisfying this burden may be Big Data's greatest challenge. Some of Big Data's most vocal advocates contend Big Data is valuable precisely because it crunches data that are ubiquitous and *not* directly job-related. Returning to our coders, one expects them to visit their preferred manga site when they are away from work, but for reasons unrelated to their jobs. The employer's reliance on the algorithm may be job-related, but the algorithm itself is measuring and tracking behavior that has no direct relationship to job performance. Its value derives solely from a correlation between this recreational behavior and job performance. The legal question is whether an employer can meet its burden of proving job-relatedness with evidence that is strictly correlational.

The Uniform Guidelines, although published in 1978, continue to inform how courts view validation. According to the guidelines, an employer will generally not be able to establish that a selection procedure is valid based on the job performance of employees working elsewhere. In order to "transport" statistical findings established elsewhere into its own workplace, that employer must demonstrate that its own employees, and those who are the subject of the validation study, "perform substantially the same work behaviors, as shown by appropriate job analyses."[73] In sum, the regulations require the job duties of the subjects of the validation study, and those of the employer who relies on that study, to be substantially similar.

This "transportability" principle has implications for the Big Data correlations discussed above. For instance, suppose an algorithm that predicts coding ability based upon social-media use has been validated by an employer whose programmers write code in a particular computer language. But is the ability to code in that language similar enough to the skills required to code in a different language, so that the same algorithm yields accurate predictions for both? That is, are programming skills so generic that programmers proficient in one language will be proficient in another? Or are programming languages like spoken languages, in which facility in one's native language may be a poor indicator of one's ability to master a second? In either case, a methodology that first discerns what general aptitudes good programmers have in common, and then determines how to measure those aptitudes, would likely fare better than one that merely maximizes the correlation in a given employment setting. In the

72. 42 U.S.C. §2000e-2(k)(1)(A)(i) (2012).

73. 29 C.F.R. § 1607.7B(2) (2015).

latter case, unless there is evidence that an algorithm correlated with proficiency in one programming language also correlates with proficiency in the second language, an employer may not "transport" the algorithm for use with the second programming language.[74]

Assuming transport validity is out, how can an employer validate its own selection procedure? The Uniform Guidelines condone three types of validation studies: criterion, content, and construct.[75] Content validity is the most straightforward, but the least relevant to Big Data. It relies on a close correspondence between the skills tested and those required to succeed in the job.[76] The typing test given to a prospective typist is the paradigm; although even here the text on which the examination is given must be similar to the text typed by a proficient employee.

But this close correspondence is anathema to Big Data. The contribution claimed for Big Data is that the information fed to the algorithm may be entirely unrelated to the job requirements, so long as it is predictive of job

---

74. This principle is reinforced by the Guidelines' caution that the general reputation of a test, its author, or its publisher, or casual evidence of validity will be accepted in lieu of statistical evidence. Further, specifically ruled out are assumptions based on a procedure's name, descriptive labeling, promotion literature, testimonial statements, or the frequency of a procedures usage. *Id.* § 1607.9A.

75. *See generally id.* § 1607.14.

These methods have been concisely described as follows:

> A criterion-related validation study determines whether the test is adequately correlated with the applicant's future job performance. Criterion-related tests are constructed to measure certain traits or characteristics thought to be relevant to future job performance. An example of an employment test that would be validated by the criterion-related validation method is an intelligence test. The content validation strategy is utilized when a test purports to measure existing job skills, knowledge or behaviors. "The purpose of content validity is to show that the test measures the job or adequately reflects the skills or knowledge required by the job." For example, a typing test given to prospective typists would be validated by the content validation method. Construct validity is used to determine the extent to which a test may be said to measure a theoretical construct or trait. For example, if a psychologist gave vocabulary, analogies, opposites and sentence completion tests to a group of subjects and found that the tests have a high correlation with one another, he might infer the presence of a construct—a verbal comprehension factor.

Hearn v. City of Jackson, 340 F. Supp. 2d 728, 733 (S.D. Miss. 2003), *aff'd*, 2004 U.S. App. LEXIS 21338 (5th Cir. 2004) (quoting Gillespie v. State of Wisconsin, 771 F.2d 1035, 1040 n.3 (7th Cir. 1985)).

76. Washington v. Davis, 426 U.S. 229, 247 n.13 (1975) (content validity is demonstrated by tests whose content closely approximates tasks to be performed on the job by the applicant).

performance. On its face, the data relied upon by the algorithm, and the algorithm itself, are likely to appear far removed from the tasks the job requires. Thus, content validity can be dismissed more or less out of hand as a method for validating Big Data.

Construct and criterion validity are closely related. A construct is "identifiable characteristics which have been determined to be important for successful job performance."[77] Sometimes these traits may be apparent. All else being equal, speed is a substantial asset to a football player. In other settings, however, identifying the salient traits is more challenging. Accordingly, the Uniform Guidelines caution employers that "the user should be aware that the effort to obtain sufficient empirical support for construct validity is both an extensive and arduous effort involving a series of research studies . . . ."[78] Thus, an employer must first establish that the construct in question contributes significantly to success on the particular job, and that the procedure or test accurately identifies those who possess that construct.[79]

Criterion validity, or predictive validity as it is sometimes known, differs in that the objective is to predict ultimate success on the job rather than traits believed to lead to success.[80] The regulations governing criterion validity are more detailed than those pertaining to construct validity. The Uniform Guidelines list several steps deemed "essential," many of which pertain to a "job analysis." A job analysis should identify the behaviors or outcomes that are critically important; the proportion of time spent on each such behavior or outcome; the difficulty of accomplishing these behaviors or outcomes; the consequences of errors in those regards; and the frequency with which various tasks are performed.[81] The purpose in systematizing this information is to determine which jobs may be reasonably grouped to rate employees for their proficiency and to identify a common test or screen for selecting them.[82] Employers must also explain the bases for selecting the success measures and the means by which they were observed, recorded, evaluated, and quantified.[83] The Uniform Guidelines provide that "a selection procedure is considered related to the criterion, for the purposes of these guidelines, when the relationship between performance on the

77. 29 C.F.R. § 1607.16E.
78. *Id.* § 1607.14D.
79. *Id.*
80. *Id.* § 1607.14B.
81. *Id.* § 1607.15B
82. *Id.*
83. *Id.*

procedure and performance on the criterion measure is statistically significant at the .05 level of significance."[84]

There are generally two methods to establish construct or criterion validity. One is "concurrent validity"; the other is "predictive validity."[85] In a concurrent study, both the selection-procedure score (e.g., a test score) and the performance score it is intended to predict are collected at the same time.[86] For example, an incumbent workforce, whose job performance can be rated, may be administered a proposed test to see if test scores are correlated with a measure of on-the-job performance. In a predictive validity study, selection scores are obtained for a group of applicants.[87] The group that is hired is subsequently evaluated in terms of on-the-job performance.[88] The selection scores are then correlated with measures of performance to assess whether they predicted performance accurately.[89]

Both methods of validation pose challenges for Big Data solutions based largely on correlations. Because the relationships relied upon by Big Data are entirely empirical and both concurrent and predictive validity are time dependent (as explained below), there is no reason the correlations that underlie Big Data solutions should persist beyond the sample period. Because concurrent validity is based upon information from incumbent employees, the correlations regarding these individuals will be relevant to the applicant pool only if incumbents and applicants are similar in the many dimensions measured by Big Data. For example, if incumbents are older than applicants, then the social-media profile of this older group may differ markedly from that of younger job applicants. Accordingly, an algorithm highly accurate in sorting *incumbents* for their proficiency may yield *applicants* notable only for their "retro" tastes and lifestyles.

Similarly, a predictive validity study, in which applicants first are screened in the dimensions relevant to Big Data and then have their job performance assessed after they are employed for a reasonable time,[90] will be relevant only if patterns observed in the past continue to be relevant to

84. *Id.* § 1607.14B(5).

85. Brunet v. City of Columbus, 642 F. Supp. 1214, 1242 (S.D. Ohio 1986), *rev'd*, 1 F.3d 390 (6th Cir. 1993).

86. *Id.*

87. *Id.*

88. *Id.*

89. Richard Jeanneret, *Professional and Technical Authorities and Guidelines*, *in* EMPLOYMENT DISCRIMINATION LITIGATION: BEHAVIORAL QUANTITATIVE AND LEGAL PERSPECTIVES 47, 58 (Frank J. Landy ed., 1st ed. 2005).

90. Dr. Jeanneret recommends assessing performance no sooner than six months after hire. *Id.*

job performance. This is the case to which the manga example applies. If in January the best programmers have flocked to a particular website, but by July a different website is the hottest draw, an algorithm that continues to rely on visits to the first website may be mistaking the very best applicants. Thus, the gold standard is not mere correlations but *stable* correlations that yield reliable predictions over a relatively long time.

Correlations that reflect causation as opposed to happenstance are much more likely to yield persistent results. Because doctors understand how penicillin cures infections, they can be rather certain that a shot of penicillin administered this week will eradicate an infection—just as it did for weeks, months, and years before. But if correlations unearthed by Big Data are ephemeral, then their algorithms must be updated regularly to maintain validity. There may, however, be a minimum lag inherent in the methodology. If newly hired employees can be reasonably assessed only after six months of employment, that period will define the minimum time a correlation must persist to yield meaningful results. Although concurrent validity is not plagued by the same type of time lag, the behavioral differences between younger applicants and older incumbents may be even harder to account for.

Besides validating Big Data methods, the Uniform Guidelines require employers to assess the "fairness" of any selection procedure that produces an adverse impact on a protected group.[91] In this context, "fairness" addresses whether an assessment measure, valid for employees and applicants, rates members of the majority and protected groups equally.[92] Returning to our example of the manga website, suppose visits to this site reliably predict coding ability for the great majority of persons, but Hispanic coders—even the best—are not interested in this variety of literature. Although the better Hispanic coders may score higher on this measure than poorer Hispanic coders, indicating this measure is valid among Hispanics, Hispanics as a group may fare poorer on this measure relative to majority group members. Validity, standing alone, would seem to justify use of this selection criterion; only by distinguishing the correlation among groups is it apparent that this measure is a poor indicator of coding proficiency between Hispanics and other groups. As a result, general validity would not provide a complete defense to a disparate-impact claim brought by a Hispanic applicant who was poorly rated by a Big Data algorithm.

91. 29 C.F.R. § 1607.14B(8) (2015).
92. *See id.* § 1607.14B.

Fairness has wider implications than this example illustrates. The Uniform Guidelines define "unfairness" as a condition in which "members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance."[93] Recall that one of the consequences of big data is that all disparities are statistically significant—the criterion many courts use for materiality.[94] As a result, every algorithm will calibrate differently for each protected group for which it is estimated. In other words, the algorithm that accurately predicts the success of female candidates is likely to differ, however slightly, from the algorithm that predicts the success of male candidates. If one algorithm is used for both groups, it will necessarily be less accurate than a female-specific algorithm. Moreover, if it selects female applicants at a lower rate, it will be "unfair."[95]

VI. Beyond the Guidelines

The Uniform Guidelines reflect a regime in which selections were based on scored test results. The paradigm is one in which test developers first discern the critical aspects of a particular job, deduce the skills necessary to perform those tasks, devise tests to assess those skills, and then validate those tests either by mirroring the content of the job or by demonstrating the test successfully distinguishes good from bad employees. In that world, a statistically significant correlation confirms that the test designer's assessment of the job, and design or choice of the test, accurately captured the knowledge, skills, and abilities required for success on the job.

A recent Sixth Circuit opinion illustrates how a court applied the Uniform Guidelines to assess a test administered to candidates for police sergeant.

> Here, in deeming the 2002 process's testing methods valid, the district court detailed Dr. Jeanneret's "comprehensive job analysis," on behalf of the City, to identify the most important knowledge, skills, abilities, and personal characteristics (KSAPs) for the sergeant position.
>
> Jeanneret & Associates sought to assess all 44 of the important KSAPs identified in the job analysis and designed the

93. *Id.* § 1607.14B(a).
94. Rubinfeld, *supra* note 61.
95. *See* 29 C.F.R. § 1607.6V.

> test questions to meet the content validity requirements for the assessment. The investigative forms and other materials used in the investigative logic test and oral component were very similar to the actual materials used on the job and clearly simulated critical job duties. Additionally, all of the items on the job knowledge test were developed using the same reference materials used by MPD sergeants on the job. The investigative logic test involved realistic scenarios that were designed to simulate situations encountered and investigative activities performed by sergeants on the job. Likewise, the application of knowledge test was designed to evaluate how a candidate would respond to common situations encountered on the job. The [video-based] oral component also involved realistic scenarios designed to simulate situations in which a sergeant would be expected to use oral communication skills in responding to a superior officer, responding to the mother of a victim, and responding to a new partner.[96]

Underlying each validation method approved by the Uniform Guidelines is the requirement of a job analysis.[97] This reflects the commonsense view that, to design a test or selection instrument that distinguishes those best able to perform a job from those who are least able, the test designer must have some understanding of what the job entails. The Guidelines' technical standards require at a minimum that "[a]ny validity study should be based upon a review of information about the job for which the selection procedure is to be used.'"[98]

Big Data begins from the opposite perspective—it searches first for correlations. The algorithm is uninterested in what any employee actually does, so long as the employer can identify who does it well and who does it poorly. The algorithm will identify the set of variables (from the information available to it) that best distinguishes these groups. Consequently, the tests of statistical significance that ensure the ultimate validity of conventionally developed tests have far less relevance to Big Data because well-conceived algorithms will eliminate every alternative that is not significantly related to job performance—this is the cornerstone of the methodology. As a result, applying simple tests of validity to Big Data makes little sense because its algorithms are derived using those

96. Johnson v. City of Memphis, 770 F.3d 464, 478-79 (6th Cir. 2014).
97. 29 C.F.R. § 1607.15.
98. 41 C.F.R. § 60-3.14(A) (2014).

criteria, without knowing the details of what any employee does. The algorithm simply provides the answer that best fits the (job performance) data, no matter how any employee achieved that performance.[99]

The problem with Big Data is that there is no reason the algorithm that best fits the data on Monday will do so on Tuesday. This is the Achilles heel of purely correlation-based methods.[100] Because there is no understanding of *why* the correlation exists, there is no basis for surmising how long it will persist. This contrasts with the longevity courts attribute to conventional job analyses and the associated validity studies, which are premised on cause-and-effect relationships. "There is no requirement in the industry or in the law that a new job analysis be prepared for each successive selection procedure, and an earlier-developed job analysis may appropriately be used so long as it is established that the job analysis remains relevant and accurate."[101] Additionally, expert testimony has stated that "conventional wisdom places the shelf-life of a job analysis for [certain positions] at 'five plus years,' and up to ten years more."[102]

Big Data effects a shift from selection criteria distilled from job-related knowledge, skills, and abilities, leaving correlation to be established empirically, to one in which correlation is first established empirically—independently of knowledge, skills, and ability—and leaves the duration of that correlation in question. Accordingly, rather than assess Big Data in terms of correlation, which it should pass with flying colors, courts and employers should ask how long the underlying correlations will endure. In terms of validation, this translates into determining the time elapsed since the algorithm was first calibrated and the time it was applied to the plaintiff, relative to the expected duration of the correlation.

Because Big Data algorithms, by design, maximize the correlation between Big Data variables and some measure(s) of job performance, the

---

99. This means that Big Data algorithms must be evaluated relative to "out-of-sample" observations. That is, the same sample that was used to develop the Big Data algorithm will obviously validate the algorithms predictions, because that is how the algorithm was developed. A more telling comparison can be made by splitting that sample in two—using one half to develop the algorithm and the other half to test it. This is one among several methods known as "cross-validation." *See generally* Rob J. Hyndman, *Why Every Statistician Should Know About Cross-Validation*, HYNDSIGHT BLOG (Oct. 4, 2010), http://robjhyndman.com/hyndsight/crossvalidation/.

100. *See, e.g.*, Michael J. Mauboussin, *The True Measure of Success*, HARV. BUS. REV. (Oct. 2012), https://hbr.org/2012/10/the-true-measures-of-success (arguing for importance of distinguishing cause-and-effect relationships).

101. Hearn v. City of Jackson, 340 F. Supp. 2d 728, 738-39 (2003) (citation omitted).

102. *Id.* at 738 n.10.

correlation should be greatest when the algorithm is initially calibrated and should decay as time passes. But how much decay is tolerable before the algorithm is too unreliable to pass legal scrutiny? The Uniform Guidelines suggest the following: "Generally, a selection procedure is considered related to the criterion, for the purpose of these guidelines, when the relationship between performance on the procedure and performance on the criterion measure is statistically significant at the .05 level of significance."[103] This suggests that the useful life of an algorithm should be measured by the time elapsed before the correlation is reduced in significance to the .05 level. This definition fails, however, to consider how long the algorithm remains superior to less discriminatory alternatives. Therefore, determining how long the correlation persists, in both dimensions, is the critical inquiry in assessing whether a Big Data algorithm is lawfully applied to the employer's workforce.

VII. A Less Discriminatory Alternative

Title VII provides that a plaintiff may overcome an employer's proof that the challenged practice is job-related and consistent with business by demonstrating there exists an alternative practice—one which is less impactful on the protected group, yet as effective in meeting the employer's business needs—which the employer refuses to adopt.[104] Commentators suggest this formulation derives from *Albemarle Paper Co. v. Moody*,[105] in which the Supreme Court observed that Title VII liability exists when the "complaining party . . . show[s] that other tests or selection devices, without a similarly undesirable racial effect, would . . . serve the employer's legitimate interest in 'efficient and trustworthy workmanship.'"[106]

Solon Barocas and Andrew D. Selbst observe that an obvious starting point in devising an alternative begins with the algorithm that creates the adverse impact.[107] If tweaking that algorithm, perhaps by rectifying any errors or eliminating any biases, reduces the adverse impact, then the modified algorithm should be an acceptable alternative with a less discriminatory impact.[108]

---

103. 29 C.F.R. § 1607.14B(5) (2015).

104. 42 U.S.C. §2000e-2(k)(1)(A)(ii) (2012).

105. 422 U.S. 405 (1975).

106. *Id.* at 425 (quoting McDonnell Douglas Corp. v. Green, 411 U.S. 792, 801 (1973)).

107. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. (forthcoming Feb. 2016) (manuscript at 36), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

108. *Id.* (manuscript at 40).

This strategy raises questions that are fundamental to Big Data. First, the plaintiff must obtain that algorithm and the data with which it was estimated, as well as the criterion or construct data by which "success" on the job was measured. Often, third-party Big Data companies possess this information, and thus obtaining it may require a protracted discovery battle, as exemplified by *EEOC v. Kronos, Inc.* In such a scenario, a Big Data company's investment in developing the algorithms—its primary products—may be at risk.

Next, a plaintiff must devise an alternative algorithm with a less discriminatory impact. What constitutes "less," however, is unclear. In a Big Data world, almost any improvement, no matter how slight, in the proportion of a protected group that passes a screen will be deemed "statistically significant" yet negligible in a practical sense. Will a court order a company to abandon a product in which it has invested heavily, in order to increase the pass-rate of a protected group by a statistically significant fraction of a percent?

Further, there is a "whack-a-mole" aspect to this process. Suppose a female plaintiff undertakes the expense required to re-engineer the company's algorithm and finds a version that reduces the adverse impact on women. As a result, she persuades the employer to adopt this alternative. Subsequently, and unintentionally, the new algorithm enhances the adverse impact against African Americans. An African American plaintiff now sues and suggests an alternative that minimizes the adverse impact on his protected group but inadvertently enhances the adverse impact on Hispanics. The employer finds itself in the center of a game that ends only if there is a solution that minimizes the algorithm's disparate impact on *every* protected group.

There is a similar lack of precision in how well the alternative must perform relative to the original model in order to serve the employer's legitimate interest in "efficient and trustworthy workmanship."[109] Predictive analytics is engineered to select the "best" predictor of the success metric, in the sense that no other combination of data will be more accurate.[110] Accuracy, however, will likely decay as time passes. Therefore, an alternative that was less accurate when the original algorithm was adopted might become more accurate as the original correlations decay. Must an

109. *Moody*, 422 U.S. at 425 (quoting *Green*, 411 U.S. at 801).

110. *See, e.g.*, the description of how correlation-derived algorithms helped solve New York City's problem of identifying which of 51,000 manholes were most likely to catch fire, in Viktor Mayer-Schonberger & Kenneth Cukier, *Big Data* at 94-97.

employer shuffle between algorithms although the predictive power of the original remains satisfactory although inferior to other algorithms? A literal reading of *Albemarle* suggests that conclusion.

VIII. The Special Case of the ADA

The ADA poses special challenges for Big Data. Unlike other antidiscrimination laws that merely prohibit certain conduct, the ADA imposes affirmative obligations on employers. Yet the statute and its regulations reflect the screening and hiring processes as they were configured over twenty years ago. The regulations require employers:

> [T]o select and administer tests concerning employment in the most effective manner to ensure that, when a test is administered to a job applicant or employee who has a disability that impairs, sensory, manual or speaking skills, the test results accurately reflect the skills, aptitude or whatever other factor of the applicant or employee that the test purports measure, rather than reflecting the impaired sensory, manual, or speaking skills of such employee or applicant . . . .[111]

The Interpretive Guidance explains:

> The intent of this provision is to further emphasize that individuals with disabilities are not to be excluded from jobs that they can actually perform merely because a disability prevents them from taking a test, or negatively influences the results of a test, that is a prerequisite to the job.[112]

Big Data does not easily fit within this regulation for at least two reasons. First, one of the advantages of Big Data is that the information fed into its algorithms is gleaned from activities that are frequently unrelated to any work requirements (think manga websites). Thus, Big Data may use

---

111. 29 C.F.R. § 1630.11 (2015).

112. 29 C.F.R. pt. 1630, app. section 1630.11 (2011). The appendix the EEOC added to the ADA regulations contains "the Commission's interpretive guidance to the ADA." Smith v. Midland Brake, 180 F.3d 1154, 1166 n.5 (10th Cir. 1999). "As administrative interpretations of the ADA . . . these guidances are 'not controlling upon the courts by reason of their authority,' but they 'do constitute a body of experience and informed judgment to which courts and litigants may properly resort for guidance.'" *Id.* (quoting Meritor Sav. Bank v. Vinson, 477 U.S. 57, 65 (1986)). Additionally, if the interpretation is of the EEOC's own regulations, then the interpretation is entitled to greater deference. *Id.*

visits to a manga site to screen applicants, although that type of activity is not traditionally regarded as a test.

Second, because the information relied upon by Big Data may be generated in the normal course of living, applicants are unaware their extra-curricular activities may be the basis on which their suitability for a position will be judged. Practically, this means that disabled individuals—unaware that Big Data is monitoring their personal habits—are unlikely to request reasonable accommodation. From the other perspective, this means that an employer may have no reason to know that an applicant, whose data has been gleaned from the web, has an impairment that requires accommodation.[113] Not only may the employer be unaware of the applicant's disability, but it may also be ignorant of the behaviors Big Data tracks. Although it is unfair to require employers to accommodate unknown disabilities, particularly when the employer does not know the specifics of how applicants are screened, it is equally unfair to base hiring decisions on criteria that prejudice an applicant's disability. However, unless a "test" is construed to include Big Data algorithms, and unless applicants are informed of the test's elements, disabled applicants may be denied reasonable accommodation in the application process.[114]

The ADA offers disabled individuals a cause of action when policies and practices have a disparate impact, but that is not the same as requiring employers to provide reasonable accommodations.[115] The disabled are a heterogeneous group and the elements of an employer's Big Data algorithm that affect one applicant with a disability may have no impact on other disabled applicants. As a result, the paucity of numbers might not permit a disabled applicant to prove a class-wide impact. Indeed, there are few reported cases of a successful disparate-impact claim under the ADA.[116] In contrast, a disabled applicant is entitled to reasonable accommodations

113. *See, e.g.*, 42 U.S.C. §12112(b)(5)(A) (2012) (requiring "reasonable accommodation to the known physical or mental limitations of an otherwise qualified individual with a disability who is an applicant").

114. 29 C.F.R. §1630.11. *See generally* Rawdin v. Am. Bd. of Pediatrics, 985 F. Supp. 2d 636 (E.D. Pa. 2013), *aff'd*, 582 F. App'x 114 (3d Cir. 2014) (pediatric medical exam); Bartlett v. N.Y. State Bd. of Law Exam'rs, 970 F. Supp. 1094 (S.D.N.Y. 1997) (state bar exam).

115. 42 U.S.C. §12112(b)(5)(a).

116. The regulation pertaining to the selection and administration of tests to disabled applicants, is relevant to only two opinions available on Lexis, as of January 21, 2015, and each pertains to a formal test. *See generally Rawdin*, 985 F. Supp. 2d 636; *Bartlett*, 970 F. Supp. 1094 (state bar exam).

irrespective of how anyone else is affected by a particular screening procedure.[117]

IX. Minimizing the Employer's Exposure to the Risks of Big Data

The complexity of the data mining that underlies Big Data means that most employers do not have the capacity to learn how these algorithms were constructed or assess their limitations. Additionally, because the algorithms are Big Data's intellectual property, they will probably not be shared with a prospective client who wishes to understand how they work. Nevertheless, an employer who subscribes to Big Data is liable for the consequences, notwithstanding a lack of discriminatory intent or knowledge of the algorithm. The issue facing the employer is how to proceed in an environment that promises substantial benefits, but whose inner workings may be beyond understanding.

An efficient solution may be for the developer to indemnify the employer. Generally, the risk associated with a black-box solution is something the employer would pay to avoid. How much the employer is willing to pay depends on the risk it perceives and its capacity to absorb that risk. A developer, however, is likely to have a more accurate perception of the risk, since it developed the product. Thus, it may be more efficient for the developer to indemnify the employer at a price that reflects the true risk of the Big Data solution.

Suppose an employer who is unable to evaluate Big Data's risks is reluctant to subscribe to the product because its price plus its perceived risk exceeds its expected benefits. Suppose further that the developer, who completely understands the product, is certain that the risk the employer perceives is unfounded and will not materialize. Under those circumstances, it makes sense for the developer to indemnify the employer (which is costless if the product is truly riskless), thereby lowering the total cost to the employer and increasing the likelihood of purchase.

Because no product is fail-proof, and start-up companies may be ill equipped to bear the risk that a product sold to a major employer could result in liability to a large class of employees, allocating and protecting against risk is more complicated than this example portrays. But if this technology is to flourish in the employment realm, Big Data may need to solve the question of how this associated risk will be borne.

---

117. *Cf.* Eckles v. Consol. Rail Co., 94 F.3d 1041, 1046 n.8 (7th Cir. 1996) (reviewing cases holding unions cannot waive individual rights under various antidiscrimination laws).

X. Conclusion

Big Data holds the promise of more accurately and more cheaply identifying the most promising employees—and greatly reducing the likelihood of intentional discrimination in the process. Yet how these algorithms may adversely impact protected groups is unknown, and traditional legal standards are not readily applied to Big Data methods.

Rather than discover cause-and-effect relationships—which are fundamental and long lasting—Big Data identifies correlations, which may be context specific and of unknown duration. These features raise difficult questions for employers, and if each employer undertakes an independent search for answers, the hoped-for efficiencies may disappear.

Instead, the risk of noncompliance with the antidiscrimination laws may best be placed on Big Data, which is positioned to assess its algorithms and insure against those risks. Accordingly, one solution may be for Big Data to indemnify employers who purchase its services, which should motivate Big Data vendors to identify, minimize, and insure against potential risks most efficiently.