THE LITTLER REPORT

The Big Move Toward Big Data in Employment

August 2015

AUTHORS

Marko Mrkonich
Allan King
Rod Fliegel
Philip Gordon
Harry Jones
Tamsen Leachman
Michael Lotito
Garry Mathiason
Michael McGuire
Natalie Pierce
Paul Weiner
Corinn Jackson
Zoe Argento
Danielle Fuschetti
Shiva Shirazi Davoudian
Chad Kaldor
Elaine Lee
Catherine Losey
Joseph Wientge, Jr.

Littler®

IMPORTANT NOTICE

This publication is not a do-it-yourself guide to resolving employment disputes or handling employment litigation. Nonetheless, employers involved in ongoing disputes and litigation will find the information useful in understanding the issues raised and their legal context. The Littler Report is not a substitute for experienced legal counsel and does not provide legal advice or attempt to address the numerous factual issues that inevitably arise in any employment-related dispute.

Table of Contents

Table of Contents

The Big Move Toward Big Data in Employment

INTRODUCTION TO BIG DATA AND ITS POTENTIAL USES IN THE WORKPLACE

Numbers have always been used to monitor human activity. From primitive tallies etched onto the walls of ancient caves to multi-volume reports generated by computer programs operating "in the cloud," humans have always tried to deploy the power of mathematics and numbers to help understand and guide our behavior. We are not surprised, for example, when actuaries use data regarding life expectancy and health risks to set life insurance premiums. We are also not taken aback when our neighborhood bank or car dealer asks a series of questions and assigns us a credit score before deciding whether to lend us money. And it is expected that an employer will evaluate the performance of its sales force by examining the numbers it generates. Not surprisingly, in today's era of super-computing and digitalized information, our ability to acquire and to use large quantities of data has expanded exponentially, and each day brings new developments in both what we know (or stated more precisely, what we can know if we choose to) and how we can use that data.

The world of Big Data has arrived, and it is beginning to affect employers and their decision-making in ways undreamed of even a few years ago. Employers can access more information about their applicant pool than ever before, and have an ability to correlate data gleaned from the application itself, perhaps supplemented by publicly available social media sources, to determine how long a candidate is likely to stay on a particular job. Similarly, by combing through computerized calendar entries and e-mail headers, Big Data can tell us which employees are likely to leave their employment within the next 12 months. At the same time new tools and methods that rely on concepts of Big Data are becoming part of the daily landscape in human resource departments, employers continue to operate in a legal environment based on precedent and history with few guideposts that translate seamlessly into the world of Big Data. The issues that can arise either are brand new or develop in a context that makes yesterday's compliance paradigm difficult to apply.

The purpose of this white paper is to help provide employers with an introduction to the world of Big Data and what its arrival means for their daily activities. It has become axiomatic to observe that the digitalization of information has resulted in the creation of more data in recent years than in the prior combined history of humankind, and that at the same time we have acquired all of this data, our ability to apply advanced computer-based techniques to use the information has likewise expanded exponentially. It has also become cheaper and more readily accessible to do so for virtually everyone. For employers, these developments create both opportunities and novel issues of concern, and they generate new questions about long-time problems. Big Data potentially affects every aspect of employment decision-making for employers of all size in virtually every industry, from the selection and hiring process, through performance management and promotion decisions, and up to and beyond the time termination decisions are made, whether for performance reasons or as part of a reorganization.

Employers, in essence, need to understand how to balance the opportunities and risks in the brave new world of Big Data. Big Data means that employers can theoretically analyze every aspect of every decision without worrying about a need to rely only on a partial sample, and Big Data allows employers to find (or, in some cases, to disprove) correlations between characteristics and outcomes that may or may not have a seeming connection. As a result, employers need to be able to understand what Big Data means for background checks and employee privacy, to know the implications for the employer's data security obligations, to be able to use Big Data to reduce the risks of traditional discrimination claims without giving rise to new varieties of such claims, and employers need to know how to manage litigation with expanded eDiscovery and new theories of liability and new defenses based on statistical correlations. It is the desire to bring together Littler's most knowledgeable subject matter experts to help employers understand what Big Data means for our shared world that created this white paper.

BIG DATA AND THE ARRIVAL OF ARTIFICIAL INTELLIGENCE IN THE WORKPLACE

In general, applications are still designed to perform predetermined functions or automate business processes, so their designers must plan for every usage scenario and code the logic accordingly. They don't adapt to changes in the data or learn from their experiences. Computers are faster and cheaper, but not much smarter.[1]

The computer makes no decisions; it only carries out orders. It's a total moron, and therein lies its strength.[2]

[1] Judith Hurwitz, Marcia Kaufman, Adrian Bowles, *Cognitive Computing and Big Data Analytics* (Apr. 8 2015), kindle cloud location 289.

[2] Peter Drucker, *Technology, Management, and Society* (Sep. 10, 2012), p. 147.

This is the picture of Big Data analytics before cognitive computing. Big Data exists, but there can be challenges to its accessibility. Computing makes simple analysis faster and easier, but requires substantial human guidance.

When cognitive computing is applied to Big Data, the picture changes: "Acting as partners or collaborators for their human users, [cognitive computing]… systems derive meaning from volumes of natural language text and generate and evaluate hypotheses in seconds based on analysis of more data than a person could absorb in a lifetime."[3] The application of these insights to the workplace has the potential to both create and alleviate legal challenges.

What is Cognitive Computing?

Cognitive computing works by identifying associative connections between data points and reasoning from them.[4] The process by which a cognitive computing system reaches insights about the world is, on a very simple level, similar to that which can enable people to walk into a dark room and intuitively find a light switch.[5] Over time, humans have noted a pattern in the placement of light switches and have extrapolated an insight about their usual location.[6] Cognitive computing systems seek to replicate this process of observation-based learning.

Cognitive computing is hypothesis-driven. This means that a cognitive system can form a hypothesis, test and modify it, and reach an insight about the nature of something in the world. This distinguishes the algorithms grounding cognitive computing from other types of algorithms.[7]

Historically, cognitive computing is an outgrowth of the broader field of Artificial Intelligence.[8] Unfortunately, because AI has been varyingly defined,[9] there are different ways of distinguishing between AI and cognitive computing. One proposed distinction is that AI 'thinks,' while cognitive computers 'learn.'[10] Others characterize cognitive computing as a branch or subset of AI. Depending on how you conceive of AI, cognitive computing is either the branch of AI being adapted to learn from and process Big Data or it is being used instead of AI for this purpose.

How Does Cognitive Computing Enable Big Data Analytics?

Cognitive computing enhances Big Data analysis by unlocking large portions of the world's data and by providing a more sophisticated and dynamic means of analyzing it.

Cognitive computing and Big Data benefit from each other. Cognitive computing systems require large data sets to 'learn.'[11] Big Data can provide these data sets. In turn, the ability of cognitive computing systems to interpret unstructured data, and data from analogous sources, such as articles, videos, photos and human speech, has dramatically increased machine access to reams of unstructured data involved in some of the most important human interactions. Eighty percent of all data is unstructured.[12] Cognitive computing gives computers access to this information.

Allowing machines to access this 80 percent of data gives them the full picture they need to reach accurate insights about the world. For example, streaming and moving data has been traditionally difficult to analyze. Such information includes the movement of a body across a sensor, fluctuations in temperature, video feeds, and the movement of the stock market. Until cognitive computing, there was no effective way for computers to access and interpret this data in real time.

The other advantage of cognitive computing is its ability to learn from this data once it has access. Some potential uses of cognitive computing proposed in *Cognitive Computing and Big Data Analytics* are:

- Providing greater security to jobsites by enabling analysis of movements detected by motion sensors to parse threats from innocuous incidents. For example, distinguishing between a human intruder and a rabbit.

[3] *Supra* note 1.

[4] http://www.techrepublic.com/article/cognitive-computing-leads-to-the-next-level-of-big-data-queries/ (last visited July 14, 2015).

[5] *Id.*

[6] *Id.*

[7] *Supra* note 1 at 585.

[8] *See generally id*. at 503-590.

[9] AI is a very broad field and one where there is not a universally accepted definition, as people continue to discuss and debate exactly what constitutes intelligence. Certainly there is a high degree of overlap between cognitive computing and AI in areas such as machines learning algorithms, knowledge representation, natural language processing and so on.

[10] http://www.computerworld.com.au/article/522302/watson_future_cognitive_computing/ (last visited July 14, 2015).

[11] *Supra* note 1 at 1568-88.

[12] *Id.* at 1652.

- Using sensors on medical instruments to detect malfunctions and alert physicians.
- Interpreting the context of incidents in at-risk physical locations to determine if there is a problem.[13]

Thus, cognitive computing can overcome some of the human limitations on memory and observational capacity while mimicking the aspects of human cognition that permit learning.

Employment Law Consequences of Applying Cognitive Computing to Big Data

Equal Employment Opportunity Issues

If bias is the product of the human mind, must it also be the product of the mechanical mind? Not necessarily.

If cognitive computing algorithms are used to make employment decisions, it is possible that some of those decisions can constitute impermissible discrimination. As discussed below in "Section V. Big Data: Is There A Defense to a Potential Adverse Impact?," in order to prevail under a disparate impact theory of discrimination, a plaintiff must show the algorithm used to make an employment decision adversely impacts a protected group or, if the employer succeeds in establishing legitimate business reasons for using the algorithm, by demonstrating there exists a less discriminatory alternative that is equally efficient at serving the employer's legitimate business needs. Because cognitive computing is algorithm-based, a cognitive computing algorithm could conceivably be the basis of such a claim. However, the algorithms that enable cognitive computers to 'learn' would only have an impact on employees if the insights they derived were used in creating the algorithms later used in making hiring decisions. It would therefore be far more likely that these insight algorithms would be the subject of legal challenges for disparate impact. The associations between high job performance and, for example, visiting a particular website or participation in certain social media, are the kinds of insight algorithms that could be produced by a cognitive computing process.

If relied on for hiring decisions, these insight algorithms could become the basis for disparate impact claims. For example, participation in social media varies based on a number of protected factors. The below chart, created by the Pew Research Center, shows that social media participation varies based on the legally protected categories of age and sex.

WHO USES SOCIAL NETWORKING SITES % OF INTERNET USERS WITHIN EACH GROUP WHO USE SOCIAL NETWORKING SITES	
All internet users	74%
a Men	72
b Women	76
a 18-29	89^{cd}
b 30-49	82^{cd}
c 50-64	65^{d}
d 65+	49
a High school grad or less	72
b Some college	78
c College+	73
a Less than $30,000/yr	79
b $30,000-$49,999	73
c $50,000-$74,999	70
d $75,000+	78

Pew Research Center's Internet Project January Omnibus Survey, January 23-26, 2014.

Note: Percentages marked with a superscript letter (*e.g.,* ^{a}) indicate a statistically significant difference between that row and the row designated by that superscript letter, among categories of each demographic characteristic (*e.g.,* age).

PEW RESEARCH CENTER

[13] *Id.* at 1850.

The likelihood of a cognitive computing process producing an algorithm with an unlawful disparate impact could increase if the data used in creating this algorithm is itself biased. The veracity of Big Data is fundamental to its utility.[14] A recent white paper by management consulting firm McKinsey & Company stressed that "'garbage in, garbage out' applies as much to supercomputers as it did 50 years ago to the IBM System/360."[15]

Bearing this in mind, disparities in the type and volume of data available about different cross-sections of the population could cause cognitive computing systems to produce incorrect hypotheses. For example, relying on the above chart, if a cognitive computing system concluded that individuals with glasses were more successful in a given position, and 18-29-year-olds posted pictures of themselves at the same rate as people over 65, a cognitive computing system could mistakenly conclude that 18-29-year-olds wear glasses almost twice as frequently as persons over 65.[16] The results could be even further skewed if younger social media users post more pictures of themselves on average than those over the age of 65. Conceivably, cognitive computing software could be programmed to evaluate whether the information it receives is skewed and to correct for this, or at least flag it, but its architects would need to be aware of these possibilities to avoid these kinds of results.

Workplace Safety and Automation

Cognitive computing's ability to analyze motion and streaming data to interpret what is happening in the physical world has obvious applications to workplace safety and management. The simplest example is using cognitive computing software to interpret a continuous stream of video and audio to identify illegal activities occurring there. Just as cognitive computers might be used to identify intruders on the jobsite and to distinguish between an unauthorized person and a rabbit, this software may be able to identify dangerous activity, such as failing to wear OSHA-required safety gear, or sexual harassment occurring over email or even in person. This could make it possible for managers to intervene earlier to head off discrimination or injury.

Facilitating Legal Compliance

While cognitive computing could become the focus of discrimination claims, it could also help to reduce workplace discrimination. Cognitive computers could be used to identify disparate impacts resulting from their very own insight algorithms. One interviewee in an article in *Fortune* noted, "If machine learning algorithms working on big data result in racial discrimination, then other algorithms can measure the effect of discrimination."[17]

Cognitive computing could eventually streamline the employment law process by making answers about the law easier for employers to obtain. Students at the University of Toronto recently partnered with IBM's Watson to create a computer capable of providing answers to legal questions. The cognitive computing system, named "Ross," receives a legal question and then "sifts through thousands of legal documents, statutes, and cases to provide an answer [including]… legal citations …articles for further reading, and even…a confidence rating."[18] If Ross and similar systems are able to produce results of comparable quality to those produced by human researchers, they could obviate the role of lawyers as advisors in some instances.

GATHERING AND USING BIG DATA IN THE HIRING AND SELECTION PROCESS

Background Data

The application and scope of the Fair Credit Report Act ("FCRA") is often confusing.[19] Even the Act's name is misleading because the FCRA governs many kinds of background check reports, not just true credit reports from one of the credit bureaus (*e.g.*, Experian, Trans Union and Equifax). Add Big Data to this mix, where online employers or their Big Data companies can have virtually instant access to a wealth of information on employees and applicants, and the FCRA's application becomes even more complicated. When employers use Big Data to obtain information for "employment purposes," the same FCRA strictures may apply along with the same risks, including potential class-action exposure for non-compliance.[20]

[14] *Supra* note 1 at 1588-1610.

[15] Martin Dewhurst and Paul Willmott, "Manager and Machine: The New Leadership Equation" McKinsey & Company (Sept. 2014) available at http://www.mckinsey.com/insights/leading_in_the_21st_century/manager_and_machine (last visited July 14, 2015).

[16] This assumes, very conservatively, that people under the age of 29 wear glasses at the same rates as those over 65.

[17] http://fortune.com/2015/01/15/will-big-data-help-end-discrimination-or-make-it-worse/ (last visited July 14, 2015).

[18] http://www.psfk.com/2015/01/ross-ibm-watson-powered-lawyer-legal-research.html (last visited July 14, 2015).

[19] 15 U.S.C. § 1681 *et seq.*

[20] 15 U.S.C. § 1681a(h) ("The term "employment purposes" . . . means a report used for the purpose of evaluating a consumer for employment, promotion, reassignment or retention as an employee.'").

To summarize, the FCRA is widely known as the federal law that regulates the exchange of consumer credit information between credit bureaus and creditors in connection with mortgage lending and other consumer credit transactions (*e.g.*, true credit reports). By its terms, however, the FCRA also regulates the exchange of information between *employers* and "consumer reporting agencies"[21] (CRAs) that provide "consumer reports"[22] (*i.e.*, background reports). The obligations the FCRA imposes on employers are *not* only triggered when an employer orders a credit report from a CRA; employers must comply with the FCRA when they order virtually any type of consumer report from a CRA, including criminal and motor vehicle records checks.

The FCRA typically does not apply when an employer *itself*, without engaging a CRA, obtains criminal and other background information directly from its primary source, such as when an employer procures publicly-available court records.[23] Following this concept, up until the last few years, it remained uncertain whether employers could perform in-house internet research on applicants and employees without triggering the FCRA. Recent actions by the Federal Trade Commission (FTC)[24] in the Big Data context and court decisions taking expansive views of the FCRA's definitions of "consumer reports" and "CRAs" have suggested that this may no longer be the case (at least in the eyes of the FTC) depending on *how* that research is performed.

Given the swelling tide of FCRA class action litigation against employers and other risks from non-compliance with the FCRA, it is important for employers to understand the potential pitfalls in this area.[25] Even when employers think in-house searches are not subject to the FCRA, they could still inadvertently trigger the Act. The line will continue to be drawn as district courts continue to wrestle with how the FCRA may be implicated in new methods of information-sharing and otherwise. This section summarizes FCRA obligations on employers that use consumer reports, summarizes recent FCRA trends in the Big Data context, and provides practical insights for mitigating the risks that have developed from those trends.

Summary of FCRA Obligations on Employers That Use Consumer Reports

The FCRA imposes requirements on employers who use "consumer reports" or "investigative consumer reports" for employment purposes.[26] A consumer report is known as a credit report or a background report prepared by a CRA, whereas an investigative consumer report is a special type of consumer report whereby the CRA obtains information through *personal* interviews (*e.g.*, an in-depth reference check).[27]

Before an employer may obtain a consumer report from a CRA, typically it must make a "clear and conspicuous" written disclosure to the consumer in a document that consists "solely" of the disclosure.[28] The applicant or employee must provide written permission for the

[21] 15 U.S.C. § 1681a(f) ("The term 'consumer reporting agency' means any person which, for monetary fees, dues, or on a cooperative nonprofit basis, regularly engages in whole or in part in the practice of assembling or evaluating consumer credit information or other information on consumers for the purpose of furnishing consumer reports to third parties, and which uses any means or facility of interstate commerce for the purpose of preparing or furnishing consumer reports.").

[22] 15 U.S.C. § 1681a(d) ("Consumer reports are any written, oral or other communication of any information by a consumer reporting agency bearing on a consumer's credit worthiness, credit standing, credit capacity, character, general reputation, personal characteristics, or mode of living which is used or expected to be used or collected in whole or in part for the purpose of serving as a factor in establishing the consumer's eligibility for … employment purposes."). "The term 'employment purposes' . . . means a report used for the purpose of evaluating a consumer for employment, promotion, reassignment or retention as an employee." 15 U.S.C. § 1681a(h).

[23] *But see* Cal. Civ. Code § 1786.53, which is one of the rare laws that may apply as to certain, defined "public records" even if no CRA is used to assemble the information.

[24] The sometimes controversial Consumer Financial Protection Bureau (CFPB) now shares oversight of the FCRA with the FTC. *See* Rod M. Fliegel and Jennifer Mora, *Employers Must Update FCRA Notices for Their Background Check Programs Before January 1, 2013*, Littler Insight (Sept. 4, 2012) available at http://www.littler.com/employers-must-update-fcra-notices-their-background-check-programs-january-1-2013.

[25] For a detailed discussion of the class action risks employers face under the FCRA, *see* Rod Fliegel, Jennifer Mora, and William Simmons, *The Swelling Tide of Fair Credit Reporting Act (FCRA) Class Actions: Practical Risk-Mitigating Measures for Employers*, Littler Report (Aug. 1, 2014) available at http://www.littler.com/publication-press/publication/swelling-tide-fair-credit-reporting-act-fcra-class-actions-practical-r.

[26] For a detailed discussion of the FCRA's requirements, *see* Rod Fliegel and Jennifer Mora, *The FTC Staff Report on "40 Years of Experience with the Fair Credit Reporting Act" Illuminates Areas of Potential Class Action Exposure for Employers*, Littler Report (Dec. 12, 2011), available at http://www.littler.com/publication-press/publication/ftc-staff-report-40-years-experience-fair-credit-reporting-act-illumin.

[27] 15 U.S.C. §§ 1681a(d) and (e).

[28] 15 U.S.C. § 1681b(b). *But see* 15 U.S.C. § 1681a(y) (related rules for misconduct investigations) and 15 U.S.C. § 1681b(b)(2)(B)(i) (different disclosure requirements for certain commercial drivers regulated by the federal Department of Transportation). If the employer procures an "investigative consumer report," additional disclosures are necessary. The employer must allow the employee to request information about the "nature and scope" of the investigation, and the employer must respond in writing to any such request within five days. 15 U.S.C. § 168 *ld*.

employer to obtain a consumer report.[29] The employer must also make a certification to the CRA regarding its "permissible purpose" for the report and its compliance with relevant FCRA provisions and state and federal equal opportunity laws.[30]

After obtaining the consumer report or investigative consumer report on an employee or applicant, an employer must follow certain requirements *if* it intends to take "adverse action" against the applicant or employee based even in part on the contents of the report.[31] First, *before* the employer implements the adverse action against the applicant or employee, the employer must provide a "pre-adverse action" notice to the individual, which must include a copy of the consumer report and the statutory Summary of Rights.[32] This requirement affords the applicant or employee with an opportunity to discuss the report with the employer before the employer takes adverse action.[33] If the employer ultimately decides to take the adverse action against the applicant or employee, it must then provide to the individual an adverse action notice with certain information specified in the FCRA.[34]

Recent trends in FCRA enforcement in the Big Data context and decisions expansively interpreting the FCRA's definitions of "consumer reports" and "CRAs" have changed the traditional perception of how the foregoing employer obligations may be triggered.

Big Data Enforcement Trends and Potential Impact on Employer Obligations Under the FCRA

The FTC has recently expanded the view of what constitutes a CRA. Traditionally, CRAs were thought of as the major credit bureaus or background screening companies compiling and generating hard copy reports for employers on specific applicants and employees. The FTC has gone beyond that traditional notion, finding that certain online data brokers and even mobile application developers were acting as CRAs without adhering to the strictures of the FCRA. Because at least some of these companies may now be considered CRAs (depending on the range and nature of their product offerings), the information employers obtain from them also may be considered consumer reports.[35] This, in turn, potentially would trigger the employer's obligations under the FCRA for procuring and using such reports for employment purposes.

The FTC filed an administrative complaint against two companies that developed mobile applications allowing users to conduct unlimited searches of criminal history information on individuals.[36] The companies had *specific* disclaimers that the information from the apps should not be considered employment screening tools and were not covered by the FCRA. The FTC found the disclaimers ineffective, noting that the companies also and concurrently had *advertising* suggesting that the apps could be used to screen potential employees.[37] The FTC considered the companies to be CRAs, subject to the FCRA, and in violation of the FCRA.[38]

[29] 15 U.S.C. §§ 1681b(a)(3)(B) and 1681b(b). For DOT-regulated motor carriers, and where the applicant applies for employment by mail, telephone, computer or other similar means, consent may be oral, written or electronic. 15 U.S.C. § 1681b(b)(2)(B)(ii). In addition, the FTC issued an opinion letter in 2001 indicating that it believed that a "consumer's consent is not invalid merely because it is communicated in electronic form," under the FCRA. *See* FTC Opinion Letter May 24, 2001 (Brinckerhoff).

[30] 15 U.S.C. § 1681b.

[31] An adverse action broadly includes "a denial of employment or any other decision for employment purposes that adversely affects any current or prospective employee." 15 U.S.C. § 1681a(k)(l)(B)(ii).

[32] 15 U.S.C. § 1681b(b). If an individual contacts the employer in response to the pre-adverse action notice to say there was a mistake (inaccuracy or incompleteness) in the consumer report, the employer may exercise its discretion whether to move forward with the hiring decision or engagement; the FCRA does not dictate a course of action. DOT-regulated motor carriers are not required to provide a "pre-adverse action" notice to applicants or employees if the applicant applied for employment by mail, telephone, computer or other similar means. 15 U.S.C. § 1681b(b)(3)(B). Rather, motor carriers must provide to the individual, within three days of taking adverse action, an oral, written or electronic notification that adverse action has been taken, which must include the same disclosures required in "adverse action" notices for non-trucking employers. Id.

[33] *Obabueki v. IBM and Choicepoint, Inc.*, 145 F. Supp. 2d 371, 392 (S.D.N.Y. 2001). The text of the FCRA does not dictate the minimum amount of time an employer must wait between mailing the pre-adverse action and adverse action notices. One fairly accepted standard is five business days. *See, e.g., Beverly v. Wal-Mart Stores, Inc.*, 2008 U.S. Dist. LEXIS 2266 (E.D. Va. 2008); *see also Johnson v. ADP Screening and Selection Services*, 768 F. Supp. 2d 979, 983-984 (D. Minn. 2011).

[34] 15 U.S.C. § 1681m(a) (requiring employers to provide: (1) the name, address and telephone number of the CRA that provided the report; (2) a statement the CRA did not make the adverse decision and is not able to explain why the decision was made; (3) a statement setting forth the person's right to obtain a free disclosure of his or her report from the CRA if he or she makes a request for such a disclosure within 60 days; and, (4) a statement setting forth the person's right to dispute directly with the CRA the accuracy or completeness of any information in the report).

[35] Several courts have also adopted expansive views of what constitutes a consumer report under the FCRA, going beyond traditional notions of a paper report compiled and provided by a background screening company. *See Ernst v. Dish Network, LLC*, 12 Civ. 8794 (LGS), 2014 U.S. Dist. LEXIS 132892 (S.D.N.Y. Sept. 22 ,2014) (finding that the report at issue was a consumer report even though the named defendant did not use the report for employment purposes); *Dunford v. American Data Bank, LLC*, No. C 13-03829 WHA, 2014 U.S. Dist. LEXIS 111761 (N.D. Cal. Aug. 12, 2014) (finding that the report at issue was a consumer report under the FCRA even though the CRA provided it only to the consumer herself and not to any prospective employer or other person).

[36] *In the Matter of Filiquarian Publishing, et al.*, FTC Matter/File Number 112 3195 (filed Jan. 10, 2013) available at https://www.ftc.gov/enforcement/cases-proceedings/112-3195/filiquarian-publishing-llc-choice-level-llc-joshua-linsk.

[37] *See* Federal Trade Commission, *Analysis of Proposed Consent Order to Aid Public Comment in the Matter of Filiquarian Publishing, LLC; Choice Level, LLC; and Joshua Linsk, individually and as an officer of the companies,* (File No. 112 3195) available at https://www.ftc.gov/enforcement/cases-proceedings/112-3195/filiquarian-publishing-llc-choice-level-llc-joshua-linsk.

[38] Although the FTC did not impose monetary penalties on these companies, it did require them to follow stringent reporting and records preservation requirements to establish their compliance with the FCRA for several years after the matter was resolved.

In another action, the FTC sued a data broker that compiled information profiles on individuals from internet and social media sources.[39] The FTC alleged that the data broker *marketed* the profiles on a subscription basis to human resources professionals, job recruiters, and others as an employment screening tool. The FTC further asserted the company was a CRA and the profiles were consumer reports.[40] The matter was resolved with the company paying $800,000 to the FTC.[41] This case illustrates how murky this area can be, as the data broker arguably was nothing more than a search engine, like Google. One difference was the data broker provided *targeted* searches of individuals' online identities whereas a person can search for anything on Google.

Although the FTC targeted data brokers in these cases, the agency's actions arguably have significant implications for many employers. Because the FTC considered the data brokers to be CRAs, the information they were providing also would be considered a consumer report. Any employer using that information to make hiring decisions arguably would have been obligated to provide a disclosure to the applicant that it would be seeking the information and obtain the applicant's authorization before viewing the information. The employer also arguably would have had to provide the pre-adverse and adverse notices if it denied the applicant employment based even in part on the information. Most employers may not think that the FCRA would even apply when they obtain information through internet data brokers like the ones discussed above. The FTC's position, as one view of the law, indicates otherwise.

An example further illustrates the blurred line between triggering the FCRA and not triggering the FCRA in the world of Big Data. Suppose an internal recruiter for an employer goes to the Facebook and Instagram pages for an applicant and decides not to hire him or her because of offensive posts and inappropriate Instagram photos. The employer arguably did not trigger the FCRA because it went directly to the separate sources of the information without any third party compiling the information for the employer. Suppose instead that the same internal recruiter had an account to a website that compiled the same Facebook and Instagram pages for the applicant in one place, and the recruiter logged into the account to view the information. The recruiter is viewing the exact same information as in the first example, but this conduct could trigger the employer's FCRA obligations because the information may have been compiled by a CRA under the FTC's expansive view of that term.

Mitigating Measures

There are several practical steps employers can take to mitigate the risks of non-compliance with the FCRA in the Big Data context, including the following:

1. Employers should consider reviewing their current policies and practices regarding employment-purposed internet searches by recruiters and other personnel, including those with direct involvement in the hiring process, such as managers and supervisors.
2. Employers should also consider taking steps to help ensure that they have provided the required disclosure and have a signed authorization from applicants and employees before they obtain background information that may be subject to the FCRA.[42] (Likewise as to efforts to comply with state and local laws, which are beyond the scope of this section.)
3. Employers should consider sending or arranging to send pre-adverse and adverse action notices whenever they take adverse action against job applicants and employees based, in whole or in part, on background information compiled by any third-party.

[39] *United States of America v. Spokeo, Inc.*, U.S. District Court Case No. 12-cv-05001 (C.D. Cal. filed June 7, 2012).

[40] *See id.*; *see also* FTC Staff Closing Letter to Renee Jackson (May 9, 2011) available at https://www.ftc.gov/enforcement/cases-proceedings/closing-letters-and-other-public-statements/staff-closing-letters?title=Social+Intelligence&field_matter_number_value=&field_document_description=&date_filter%5Bmin%5D%5Bdate%5D=&date_filter%5Bmax%5D%5Bdate%5D=&=Apply (finding that a similar data broker that compiled information from social networking sites was a consumer reporting agency).

[41] *See* Federal Trade Commission, *Press Release: Spokeo to Pay $800,000 to Settle FTC Charges Company Allegedly Marketed Information to Employers and Recruiters in Violation of FCRA* (June 12, 2012) available at https://www.ftc.gov/news-events/press-releases/2012/06/spokeo-pay-800000-settle-ftc-charges-company-allegedly-marketed.

[42] An exemption to the FCRA's coverage, added a decade ago in the Fair and Accurate Credit Transactions Act Amendment (FACTA), provides that if a communication from a CRA is made to an employer in connection with an investigation of either "suspected misconduct" or compliance with "Federal, State, or local laws and regulations, the rules of a self-regulatory organization, or any preexisting written policies of the employer" the communication is not a "consumer report." Under this exception, an employer does not have to provide the required disclosure or obtain authorization to obtain a consumer report when conducting these types of investigations. FACTA still requires that employers provide a "summary containing the nature and substance of the communication" after taking an adverse action against an individual based on the communication. For a detailed discussion of the FACTA, *see* Rod Fliegel, Jennifer Mora and William Simmons, *Fair Credit Reporting Act Amendment Offers Important Protections From Lawsuits Targeting Background Check Programs*, Littler Report (Sept. 10, 2013) available at http://www.littler.com/publication-press/publication/fair-credit-reporting-act-amendment-offers-important-protections-lawsu.

In the age of Big Data, employers have instantaneous access to information on employees and applicants. When employers use Big Data for employment purposes, the same FCRA strictures may apply along with the same risks for non-compliance. Employers should be mindful of recent developments expanding the FCRA's application into the world of Big Data, and should consider taking measures to mitigate the risks associated with that expansion.

How Big Data Affects the Numbers: OFCCP and EEOC Implications

Investigations by both the Equal Employment Opportunity Commission ("EEOC") and the Office of Federal Contract Compliance Programs ("OFCCP") frequently are numbers driven, particularly when allegations concern disparities in hiring, promotion, pay, or terminations. Investigators are trained to obtain data regarding the decision-making process that is challenged and to subject that data to elementary statistical analyses, as explained in the agencies' compliance manuals. This often is referred to as "standard deviation analysis," because the statistic on which the decision to accept or reject the null hypothesis of "no adverse impact" is based on the number of "standard deviations" by which the estimated disparity between the protected and favored group differs from zero.[43]

An alternative measure of an adverse impact derives from the Uniform Guidelines on Employee Selection Procedures (UGESP).[44] These guidelines suggest that:

> A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.[45]

These alternatives do not necessarily yield the same result. That is, an employer may select members of a disfavored group at a rate below 80 percent of the favored group, yet the disparity may not be statistically significant. On the other hand, a result may be statistically significant but the selection ratio may be greater than 80 percent. The parties, of course, can be expected to advocate the test that puts their data in the most favorable light. Courts have been less predictable, but it is fair to say that the majority have advocated statistical significance as the litmus test for determining when a disparity is materially different and therefore legally meaningful.[46]

Citations to the line of cases that supply the "two standard deviation" criterion employed by most courts generally begin with the U.S. Supreme Court's decision in *Castaneda v. Partida*.[47] Yet *Castaneda* was not an employment discrimination case, much less a disparate impact case. At issue was whether a South Texas county's method of convening a grand jury unfairly excluded Mexican-Americans, resulting in the discriminatory adjudication of Mexican-American defendants in criminal cases.[48]

In reviewing this claim, the Court compared the percentage of Mexican-Americans among those summoned to serve on the county's grand juries, to Mexican-American representation in the county's eligible population. The Court noted that there were only 339 Mexican-Americans among 870 grand jurors summoned during the relevant timeframe, and that strictly proportional representation would have seated 688 Mexican-American grand jurors. [49] The Court considered this disparity of nearly 100 percent to be material, observing that "if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis . . . would be suspect to a social scientist."[50] Yet in *Castaneda*, the difference between the actual and expected number of Mexican-American grand jurors was approximately 29 standard deviations.[51] Based in part on that comparison, the Court affirmed the district court's finding that Mexican-Americans discriminatorily were excluded from grand jury service.[52]

[43] *Watson v. Ft. Worth Bank & Trust,* 487 U.S. 977, 995 (1988).

[44] 29 C.F.R. § 1607.

[45] 29 C.F.R. § 1607.4(D).

[46] *But see, e.g., Matrixx Initiatives, Inc. v. Siracusano*, 131 S. Ct. 1309, 1321 (2011) (factors other than statistical significance must be considered in determining materiality); and *Clady v. Los Angeles Co.*, 770 F.2d 1421, 1428 (9th Cir. 1985) (rejecting the 80% test), *cert. denied*, 475 U.S. 1109 (1986).

[47] 430 U.S. 482 (1977).

[48] *Id.* at 482.

[49] *Id.* at 496 n.17.

[50] *Id.*

[51] *Id*. "The 'standard deviation' is a unit of measurement that allows statisticians to measure all types of disparities in common terms. Technically, a 'standard deviation' is defined as 'a measure of spread, dispersion, or variability of a group of numbers equal to the square root of the variance of that group of numbers.'" *Palmer v. Shultz*, 815 F.2d 84, 92 n.7 (1987) (quoting David Baldus & James Cole, Statistical Proof of Discrimination 359 (1980)). Case law often erroneously interchanges this term with the more technically-appropriate term, "standard error," which describes the distribution of sample estimators, such as the mean, around its true value. *See* David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, Federal Judicial Center, Reference Manual on Scientific Evidence 174 (2d ed. 2000).

[52] *Castenada*, 430 U.S. at 517.

The Court again referenced the benchmark of "two or three standard deviations" in *Hazelwood School District v. United States*.[53] *Hazelwood* was a pattern and practice suit alleging that a school district engaged in the discriminatory hiring of African-American teachers.[54] The Court compared the percentage of teachers in the district who were African-Americans to the percentage in the relevant labor market. Noting that the disparity exceeded six standard deviations in one year, and five standard deviations in the following year, the Court concluded that the statistical evidence reflected a "gross" disparity that was probative of a pattern and practice of discrimination.[55] Relying in part upon this finding, the Court remanded the case with instructions that the district court craft an acceptable remedy, which was to include injunctive as well as other equitable relief.

In *Watson v. Fort Worth Bank & Trust*, Justice O'Connor reviewed the Court's statistical criteria in employment discrimination cases.[56] Although she acknowledged the prevalence of the *Castaneda-Hazelwood* test of "two or three standard deviations," she noted that the Court never instructed lower courts to apply the standard mechanistically.[57] Rather, courts should evaluate statistical evidence in relation to the disputed issues and determine the appropriateness of such evidence case-by-case. Justice O'Connor observed:

> We have emphasized the useful role that statistical methods can have in Title VII cases, but we have not suggested that any particular number of "standard deviations" can determine whether a plaintiff has made out a *prima facie* case in the complex area of employment discrimination. Nor has a consensus developed around any alternative mathematical standard. Instead courts appear generally to have judged the "significance" or "substantiality" of numerical disparities on a case by case basis. At least at this stage of the law's development, we believe that such a case-by-case approach properly reflects our recognition that statistics "come in infinite variety and . . . their usefulness depends on all of the surrounding facts and circumstances."[58]

Nevertheless, many lower courts have adopted the *Castaneda-Hazelwood* criterion of "two or three standard deviations" as a bright-line rule. In so doing, they have noted that this criterion, when applied to the commonly assumed bell-shaped, normal distribution corresponds to the 0.05 level of "statistical significance" prevalent in the scientific literature.[59] This criterion—the five-percent probability threshold—corresponds, in turn, to the probability of "Type I error," the probability of mistakenly rejecting the null hypothesis of non-discrimination when it is true. Generally, the lower the probability of Type I error, the more confident the researcher is that he or she is not mistakenly claiming a statistical finding to be important. The Seventh Circuit has explained:

> In addition to describing statistical significance in terms of levels of standard deviation, statistical significance also may be expressed as a probability value (P) on a continuous or relative scale ranging from 0 to 1.0. The level of statistical significance rises as the value of the (P) level declines . . . A (P) value below .05 is generally considered to be statistically significant, *i.e.*, when there is less than a 5% probability that the disparity was due to chance. For large samples, statistical significance at a level in the range below 0.05 or 0.01 is "essentially equivalent" to significance at the 2 or 3 standard deviation level.[60]

This reasoning has led many courts to adopt a *per se* rule that statistical evidence failing to meet the 0.05 level of significance is inadmissible.[61]

For example, in *Palmer v. Shultz* the U.S. Court of Appeals for the District of Columbia Circuit extensively considered the rather esoteric question of whether it should apply a one-tailed or two-tailed test of statistical significance.[62] Its decision to apply the two-tailed test ultimately was outcome-determinative and led to rejecting the plaintiff's statistical evidence.[63] Similarly, in *Bennett v. Total Minatome*

[53] 433 U.S. 299, 308 n.14 (1977).

[54] *Id.* at 299.

[55] *Id.* at 308 n.14.

[56] 487 U.S. 977 (1988).

[57] *Id.* at 995 n.3.

[58] *Id.* (internal citations omitted).

[59] In *Castaneda*, the Supreme Court noted that, when dealing with large numbers, social scientists reject the "hypothesis of equality" —that the chances of an event are "equally" likely to result from chance or a proposed cause—if a disparity between actual and expected representation exceeds two or three standard deviations. 430 U.S. at 496 n.17.

[60] *Griffin v. Bd. of Regents of Regency Univs.*, 795 F.2d 1281, 1291 n.19 (7th Cir. 1986) (citing *Coates v. Johnson & Johnson*, 756 F.2d 524, 537 n.13 (7th Cir. 1985)).

[61] *See, e.g., Bennett v. Total Minatome Corp.*, 138 F.3d 1053, 1062 (5th Cir. 1998).

[62] *Palmer*, 815 F.2d 84, 92 (D.C. Cir. 1987).

[63] *Id.* at 94-95.

Corp, the Fifth Circuit explicitly discussed the relationship between the number of standard deviations, and the "p-value"—the probability of Type I error associated with that disparity.[64] It reaffirmed its rule that only statistical results corresponding to a p-value of 0.05 or less are admissible. In the same vein, the Eleventh Circuit has opined:

> The "general rule" is that the disparity must be "greater than two or three standard deviations" before it can be inferred that the employer has engaged in illegal discrimination under Title VII. The Court has also called that sort of imbalance a "gross statistical disparit[y]."[65]

Whether the 80-percent rule or the statistical significance test is likely to favor either party depends in large measure on the number of decisions at issue. Other things the same, the greater the number of decisions, the greater the statistical significance of any disparity:[66]

> For example, if the average wage rate is $10.00 per hour, a wage differential between men and women of $0.10 per hour is likely to be deemed practically insignificant because the differential represents only 1% ($0.10/$10.00) of the average wage rate. That same difference could be statistically significant, however, if a sufficiently large sample of men and women was studied. The reason is that statistical significance is determined, in part, by the number of observations in the data set.[67]

As a result, small employers whose selection disparities are below 80 percent, suggesting a legally meaningful difference, are prone to emphasize that the disparity is not statistically significant. In contrast, a large employer, by virtue of the number of decisions analyzed, is likely to advocate the 80-percent rule, because with enough data even numerically small differences may be statistically significant. Accordingly, as employers have grown and the data available for analysis has increased, plaintiffs and the government have urged that statistical significance is the standard by which disparities should be evaluated, rather than the agencies' own rule of thumb.

Big Data pushes this statistical framework to its limits and perhaps beyond. As more and more data is brought to bear on the selection process, disparities between demographic groups are bound to become increasingly significant, in the statistical sense, as a natural consequence of super-sized databases. At the extreme, even differences most would find negligible nevertheless may exceed the "two standard deviation" criterion. A prominent example is the statistical analysis reported in *Wal-Mart Stores, Inc. v. Dukes*, which considers one of the largest data sets to be analyzed in an employment discrimination suit.[68] In his comparison of pay differences between men and women, the plaintiffs' expert reported a standard deviation of one-tenth of one percent. The implication is 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 $9.99 per hour would be judged "statistically significant." When data sets grow to that size, statistical criteria risk trivializing the important question of what may constitute discrimination.[69]

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 "big," Big Data makes it highly likely that any difference between demographic groups in selection rates, be it with respect to promotion, hiring, or termination, will be statistically significant no matter how slight. 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.

[64] *Total Minatome*, 138 F.3d at 1062 (and cases cited therein).

[65] *Peightal v. Metropolitan Dade Co.*, 940 F.2d 1394, 1406 (11th Cir. 1991) (internal citations omitted), cert. denied, 502 U.S. 1073 (1992) (citing: *Casteneda*, 430 U.S. at 497 n.17; *Hazelwood*, 433 U.S. at 308; and *City of Richmond v. J.A. Croson Co.*, 488 U.S. 469, 501 (1989)). *See also Smith v. Xerox Corp.*, 196 F.3d 358, 364-66 (2d Cir. 1999) (finding that a disparity of two or three standard deviations equals a gross statistical disparity); *Ottaviani v. State Univ. of N.Y. at New Paltz*, 875 F.2d 365, 370-74 (2d Cir. 1989) (same), cert. denied, 493 U.S. 1021 (1990); *Palmer*, 815 F.2d at 96-97 (same); *NAACP v. Town of East Haven*, 892 F. Supp. 46, 48, 50-51 (D. Conn. 1995) (same).

[66] "[L]arger sample sizes give more reliable results with greater precision and [statistical] power . . ." *The Importance and Effect of Sample Size*, Select Statistical Services, http://www.select-statistics.co.uk/ article/blog-post/the-importance-and-effect-of-sample-size.

[67] *See* Daniel L. Rubenfeld, *Reference Guide on Multiple Regression*, Federal Judicial Center, Reference Manual on Scientific Evidence 191 (2d ed. 2000).

[68] 131 S. Ct. 2541 (2011). In this suit, the nationwide class consisted of approximately 1.5 million female employees. *Id.* at 2544.

[69] *See, e.g.,* Mark Kelson, *Significantly misleading*, Significance Magazine (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.").

Rules of thumb are common in cases of 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.[70] Analogously, the Eighth Circuit holds that reductions in force that fail to reduce the percentage of the workforce aged 40 and older by more than four percentage points are *per se* not discriminatory.[71] 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 80-percent threshold of the Uniform Guidelines and recognize that in the era of Big Data, statistical significance is the norm and therefore a poor indicator of legal relevance.

Big Data and the Americans with Disabilities Act

The Americans with Disabilities Act of 1990 ("ADA") as amended by the ADA Amendments Act of 2008 ("ADAAA"),[72] poses special challenges for Big Data. Unlike other antidiscrimination laws that merely prohibit certain conduct, the ADA imposes affirmative obligations on employers.[73] Yet, the statute and its regulations reflect the screening and hiring processes as they were configured over 20 years ago. The regulations require employers:

> to 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….[74]

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."[75]

Big Data does not easily fit within this regulation for at least two reasons. First, one of the advantages claimed for Big Data is that the information input into its algorithms is gleaned from activities engaged in voluntarily by individuals, which frequently are unrelated to any work requirement.[76] Thus, Big Data may use visits to particular websites to screen applicants, but that type of activity is not traditionally regarded as a test.

Second, because some of the information relied upon by Big Data is generated by individuals in the normal course of living, they are unaware their extra-curricular activities may be the basis on which their suitability for a position will be judged. Disabled individuals, impaired in the activities monitored by Big Data, cannot request reasonable accommodations if they are unaware how they are being screened. On the other hand, an employer also may not know that an applicant, whose data has been gleaned from the web, has an impairment that might require accommodation. Not only may the employer be unaware the applicant is disabled, it may also be ignorant of the behaviors tracked by Big Data that influence how an applicant is assessed. Although it is unfair to require employers to accommodate unknown disabilities, it is equally unfair to base hiring decisions on criteria that reflect an applicant's disability. However, unless a "test" is construed to include Big Data algorithms and applicants are informed of their elements, disabled applicants may be denied reasonable accommodation in the application process.

[70] *See, e.g., Holowecki v. Fed. Exp. Corp.*, 644 F. Supp. 2d 338, 357-58 (S.D.N.Y. 2009) aff'd, 392 F. App'x 42 (2d Cir. 2010) (vague allegations of preferential treatment to someone three years younger is insufficient to give rise to inference of age discrimination as matter of law); *Grosjean v. First Energy Corp.*, 349 F.3d 332, 339 (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"); *Aliotta v. Bair*, 576 F. Supp. 2d 113, 125 n.6 (D.C. Cir. 2008) (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)).

[71] *See Clark v. Matthews Intern. Corp.*, 639 F.3d 391, 399 (8th Cir. 2011).

[72] 42 U.S.C. § 12101, *et seq.* (2009).

[73] Employers have a duty to engage an employee or applicant in the interactive process to determine whether the employee or applicant with a known disability can perform a position's essential functions with reasonable accommodations. 42 U.S.C. §§ 12111(8), (9), 12112(a) & (b)(5) (2009); 29 C.F.R. §§ 1630.2(o), 1630.9 & Pt. 1630, App. §§ 1630.2(o) & 1630.9; *Humphrey v. Memorial Hosps. Ass'n,* 239 F. 3d 1128, 1137 (9th Cir. 2001) ("Once an employer becomes aware of the need for accommodation, that employer has a mandatory obligation under the ADA to engage in an interactive process with the employee to identify and implement appropriate reasonable accommodations."); *see Equal Employment Opportunity Comm'n v. Sears, Roebuck & Co.,* 417 F. 3d 789, 805–808 (7th Cir. 2005).

[74] 42 U.S.C.§ 12112(b)(7) (2009).

[75] Section 1630.11 Administration of Tests, 29 C.F.R. Pt. 1630.

[76] *See e.g. How Big Data is Taking Recruiters from "I Think" to "I Know."* Theundercoverrecruiter.com, available at http://theundercoverrecruiter.com/big-data-recruiters/ (last visited July 14, 2015).

The ADA provides disabled individuals a cause of action regarding policies and practices that have a disparate impact,[77] but that theory may be unsuited to litigating questions of reasonable accommodation. The disabled are a heterogeneous group and the elements of an employer's Big Data algorithm that affect one disabled applicant 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.[78] In contrast, a disabled applicant is entitled to a reasonable accommodation[79] irrespective of how anyone else is affected by a particular facet of the screening procedure.

A potential solution is to require Big Data to disclose the data input into its algorithms, so disabled applicants have notice of the activities that are monitored. However, these algorithms are proprietary and reflect extensive development efforts by Big Data. Public disclosure, of course, would greatly devalue this intellectual property. Alternatively, employers might be required to disclose that they premise their employment decisions on data gleaned from external sources. This might trigger a dialogue in which a disabled applicant explains his or her physical or mental limitations, and a reasonable alternative may be to evaluate such candidates independently of the Big Data algorithm.

BIG DATA USE IN PERFORMANCE MANAGEMENT AND DISCIPLINE

How Invasive is Too Invasive: Big Data and Privacy in the Workplace

Companies using Big Data in employment (referred to in this section alternately as "Big Data analytics" and "human resources analytics") claim they can increase employee productivity and morale and decrease turnover.[80] While the claims are compelling, employers must address privacy and data security concerns if they engage these services. First, collecting information about employees could potentially violate employees' privacy rights. Second, employers must protect the security of any sensitive information collected about employees. In addition, privacy and data security concerns become quite complex if the employer collects and analyzes the data of international employees.

Privacy

With regard to privacy, federal and state statutes and the common law restrict the information that employers can collect about employees and how they can use it. While these restrictions are particularly onerous when it comes to the collection and use of employees' health information, employers have a surprisingly high degree of latitude with respect to employee data unrelated to health, such as performance and compensation information.

Health Data

Many employers likely would be interested in conducting Big Data analysis on the health of their employees and their employees' family members because employees' and their family members' health can have a significant impact on the employer's bottom line. An employee or an employee's family member with a serious health condition is more expensive to insure. In addition, the employee likely would miss more work due to this condition or his or her performance might suffer. However, employers are effectively precluded from using health information about employees to conduct Big Data analysis, the results of which could be used to make employment decisions.

There are several federal laws designed to protect health information in general, and employee (and their family members') medical information, in particular.[81] These laws recognize that employers may receive health information about employees for a specific purpose and are designed to prevent access to, and use of, that information for a different purpose. The Health Insurance Portability and Accountability Act of 1996 (HIPAA), the Health Information Technology for Economic and Clinical Health (HITECH) Act, and their implementing regulations, are the most comprehensive of these laws. HIPAA regulations govern the use and disclosure of individually identifiable

[77] 42 U.S.C. § 12112(b)(3) (2009); *Raytheon Co. v. Hernandez* , 540 U.S. 44, 52 (2003).

[78] *See e.g. McGregor v. National R.R. Passenger Corp.,* 187 F. 3d 1113, 1116 (9th Cir. 1999) (policy requiring employees to be "100% healed" or "fully healed" in order to return to work after an injury is facially discriminatory and constitutes a *per se* violation of the ADA); *Bates v. United Parcel Service, Inc.,* 511 F. 3d 974, 994-995 (9th Cir. 2007) (hearing standards not otherwise mandated by law constitute *per se* violation of the ADA because the policy screens out hearing-impaired individuals who are otherwise qualified to perform the job. Employer therefore has the burden to establish the affirmative defense of business necessity to show that "performance cannot be established by reasonable accommodation.")

[79] 42 U.S.C. § 12112(a) , (b)(1) & (b)(5) (2009).

[80] Steven Pearlstein, *People analytics: 'Moneyball' for human resources*, Wash. Post, Aug. 1, 2014, available at http://www.washingtonpost.com/business/people-analytics-moneyball-for-human-resources/2014/08/01/3a8fb6ac-1749-11e4-9e3b-7f2f110c6265_story.html (last visited July 14, 2015).

[81] Employers should also be aware of the myriad of state laws that protect employee medical information, such as the California Confidentiality of Medical Information Act, which prohibits employer misuse of medical information. Ca. Civ. Code §§ 56.20-56.245.

health information and significantly restrict access to information and the ways employers can use the information that becomes available. Companies with self-insured health plans are most directly impacted by HIPAA and require the most comprehensive privacy safeguards to ensure that information used to administer health benefit claims is not utilized for other purposes, including making employment decisions.[82]

The Americans with Disabilities Act (ADA) also imposes tight restrictions on the collection and use of employees' health information. The ADA precludes employers from asking applicants and employees about medical conditions or disabilities, with limited exceptions, such as when an employee or applicant is seeking an accommodation. In the case of more intrusive inquiries, such as requiring examinations, the examination must be "job-related and consistent with business necessity."[83] In addition, the results of permitted health exams and other health information collected from employees for purposes of addressing disabilities that impact work must be kept confidential in a file separate from the employees' personnel file and may be disclosed only to very limited categories of individuals within and outside the employer's organization.[84]

The Family Medical Leave Act (FMLA) also has strong privacy protections in that it incorporates, by reference, the ADA's confidentiality language. Consequently, information concerning an employee's request for FMLA leave would be off limits to employers for purposes other than administering leave.[85]

The Genetic Information Nondiscrimination Act (GINA), despite what its name might suggest, protects far more than genetic test results. GINA defines as "genetic information" and protects any information related to the manifestation of a disease or disorder in a family member to the fourth degree.[86] GINA tightly restricts employers' collection and use of genetic information for employment purposes and imposes substantial confidentiality obligations.[87] These restrictions likely would effectively preclude an employer's use of genetic information for Big Data analytics.

Information Unrelated To Health

Putting aside employees' health data, employers are becoming interested in compiling and analyzing information other than health information about employees in order to make better decisions or gain perspective about their current or prospective employees. The following types of data are of particular interest to employers:

- Compensation information
- Performance evaluation
- Job progression
- Tenure
- Business expense reimbursement and compliance with reimbursement and documentation policies

Employers generally can use these categories of information about their own employees to conduct Big Data analytics virtually without legal restriction. However, if the employer maintains a privacy policy for employee data, the employer should confirm that any Big Data analytics using this information complies with the employer's own privacy policy.

As human resources data analytics becomes more prevalent, employers may find themselves receiving more requests for the categories of information listed above about former employees during calls by a prospective employer of an applicant who is a former employee. While these categories of information generally are not protected by statute or otherwise protected as private, employers considering whether to disclose these categories of information to prospective employers of former employees should consider whether disclosure would be consistent with the Company's own policies about the confidentiality of employee information. For example, disclosing personnel information to prospective employers likely would be inconsistent with a policy that describes personnel records as confidential company property and significantly restricts access to the information in the file.

[82] 45 C.F.R. § 164.504(f)(2)(ii)(C).

[83] 42 U.S.C. § 12112(d)(4).

[84] *See* 42 U.S.C. § 12112(d)(3).

[85] 29 C.F.R. § 825.500(g).

[86] 42 U.S.C. § 2000ff(4); 29§ C.F.R. § 1635.3(c).

[87] *See generally* 42 U.S.C. § 2000ff-1, 2000ff-5.

Looking Over the Employee's Shoulder

The past several years have seen massive changes in the rules and attitudes toward privacy and the use of data. While individual privacy may be at an all-time low, the expectations that employers will respect applicants' and employees' personal life are incrementally increasing each year. One illustration of this trend with particular relevant for human resources analytics is the recent wave of "password protection" legislation.

While employers may be interested in using social media content to build profiles of individuals likely to succeed in their workplace, these laws impose substantial limitations on employers' access to online content that is not publicly available. Currently, 22 states have enacted laws aimed at protecting applicants' and employees' personal social media content: Arkansas, California, Colorado, Connecticut, Illinois, Louisiana, Maine, Maryland, Michigan, Montana, Nevada, New Hampshire, New Jersey, New Mexico, Oklahoma, Oregon, Rhode Island, Tennessee, Utah, Virginia, Washington, and Wisconsin.[88] All 22 laws prohibit employers from requesting or requiring that applicants or employees[89] disclose their user name, password, or other information needed to access a personal social media account and most of them impose other restrictions on access by employers to personal online content, such as prohibitions on "shoulder surfing." These laws would effectively prohibit employers from "harvesting" non-public online content of applicants and employees for purposes of conducting data analytics.

Electronic Communication Review

Employers seeking to gather information from employees' electronic communications for purposes of human resources analytics should also be aware of the restrictions imposed by the federal Stored Communications Act ("SCA"). The SCA prohibits accessing electronic communications stored in a "facility through which an electronic communications service [ECS] is provided."[90] The legislative history identifies telephone companies and email providers as examples of providers of ECS.[91] Over the years, courts have not hesitated to apply the term to new forms of service providers, from internet service providers and bulletin board services[92] to Gmail,[93] Skype,[94] and Facebook.[95] As a result, accessing an employee's messages, for example, those stored in an employee's Gmail account, without permission could violate the SCA.

A crucial point for employers is that the SCA creates an exception for *providers* of an ECS to access communications stored on that own service.[96] This means that employers do not violate the SCA when they access communications stored on electronic communication systems they provide themselves, such as emails on their own company's email server.[97]

Location Tracking Devices

Collecting data about employees' location, using Global Positioning System ("GPS") technology, involves risks too. Several states have passed laws prohibiting GPS tracking of vehicles without the consent of the vehicle's owner.[98] In addition, employers face risks of common law invasion of privacy claims. The law in this area is just beginning to emerge. However, the Supreme Court held in *U.S. v. Jones* that the government's use of a location-tracking device to track the vehicle of an individual suspected of drug trafficking for one month was an unreasonable search under the Fourth Amendment because it was conducted without a warrant.[99]

[88] For more discussion of password protection statutes, *see* Philip L. Gordon & Joon Hwang, *Virginia's Password Protection Law Continues the Trend Toward Increasing Legislative Protection of Personal Online Accounts*, Littler Insight (Mar. 30, 2015) *available at* http://www.littler.com/virginias-password-protection-law-continues-trend-toward-increasing-legislative-protection-personal.

[89] The notable exception is New Mexico, which applies the prohibition only to applicants.

[90] 18 U.S.C. § 2701(a)(1).

[91] S. Rep. No. 99-541, at 14.

[92] *Garcia v. City of Laredo*, 702 F.3d 788, 792 (5th Cir. 2012).

[93] *Lazette v. Kulmatycki*, 949 F.Supp.2d. 748 (N.D. Ohio 2013).

[94] *Snyder v. Fantasy Interactive, Inc.*, 2012 U.S. Dist. LEXIS 23087 (S.D.N.Y. 2012).

[95] *Crispin v. Christian Audigier, Inc.*, 717 F. Supp. 2d 965, 981-82 (C.D. Cal. 2010).

[96] 18 U.S.C. § 2701(c)(1).

[97] *See, e.g., Fraser v. Nationwide Mut. Ins. Co.*, 352 F.3d 107, 115 (3d Cir. 2003); *Bohach. v. City of Reno,* 932 F.Supp. 1232, 1236 (D. Nev. 1996).

[98] *See, e.g.*, Cal. Penal Code § 637.7. The California statute, however, creates an exception for a location-tracking device placed by an employer on an employer-owned vehicle. *Id.* at § 637.7(b).

[99] 132 S. Ct. 945 (2012).

The Fourth Amendment does not apply to employee searches conducted by private employers, but the "reasonable expectation of privacy" standard in the common law invasion of privacy tort closely parallels the "reasonable expectation of privacy" standard under the Fourth Amendment.[100] As a result, courts may follow the Supreme Court's decision in *Jones* to find that similar tracking of employees invades their privacy under the common law.

A recent case illustrates how courts may analyze claims based on an employer's use of location-tracking devices. In *Matter of Cunningham v. New York State Dept. of Labor*, the New York Court of Appeals held that agency officials acted unreasonably when they ordered the tracking of an employee's vehicle without his knowledge or consent to investigate whether he was taking unauthorized absences and falsifying time records.[101] After one month of tracking, the agency terminated the employee for misconduct, based, in part, on the GPS data. The employee later sued, claiming the termination was improper and that the GPS data should not have been collected without his consent. The appellate court found that the agency officials' use of GPS technology was reasonable at the inception because there were "ample" grounds to suspect the employee of submitting false time records.[102] However, the court held that use of GPS was unreasonable in scope because "[i]t examined much activity with which the State had no legitimate concern—*i.e.*, it tracked [the employee] on all evenings, on all weekends and on vacation."[103] The court noted that the state removed the GPS device twice to replace it with a new device, but did not bother to remove it when the employee was about to start his annual vacation.[104]

The primary lesson from the *Cunningham* case is that, while GPS tracking can serve as a helpful tool for human resources analytics, it should be used only for an appropriate purpose and within a defined and limited scope, particularly where the employee's work and home life necessarily overlap.[105] Moreover, providing employees with robust notice of the location tracking and obtaining their prior consent should effectively eliminate an employer's exposure to a claim under state laws restricting the use of location-tracking devices under a common law invasion of privacy theory.

Data Security

Multiple federal and state laws and regulations protect the security of personal information. The law in this area tends to focus on information that could be used to commit identify theft—Social Security numbers, driver's license numbers, credit and debit card numbers, and financial account numbers, for example—and on information generally understood to be private, such as health information. Critically, the employer retains responsibility for the data even when the employer outsources the data analysis to a third party. Consequently, the employer may be liable for the missteps of a service provider handling the data on the employer's behalf.

To reduce the risk of outsourcing data analysis, employers should de-identify personal data where possible. "De-identification" refers to the process of removing individually identifying information—name and Social Security number, for example—from a data set. Several studies have thrown into question the extent to which standard de-identification steps actually sever the link between the data and an identifiable individual.[106] Nevertheless, there can be no question that de-identifying personal data at least makes it harder to use the data for identity theft or other harmful purposes. Moreover, de-identifying data can provide a safe harbor under some laws, such as the Health Information Portability and Accountability Act, as long as the de-identification process meets accepted standards.[107] Even where a statutory ore regulatory scheme does not include an express safe harbor, de-identification can effectively create a safe harbor because virtually all data protection and information security laws apply only to information that is individually identifiable.

[100] *See, e.g.*, *Smyth v. Pillsbury Co.*, 914 F. Supp. 97, 101 (E.D. Pa. 1996) (holding that a claim of invasion of privacy under the tort of intrusion upon seclusion against an employer requires a showing that the employer invaded a reasonable expectation of privacy).

[101] 2013 N.Y. LEXIS 1729 (May 29, 2013).

[102] 2013 N.Y. LEXIS 1729, at **8–9.

[103] 2013 N.Y. LEXIS 1729, at **9–10.

[104] 2013 N.Y. LEXIS 1729, at *10.

[105] Justice Sotomayor expressed a similar sentiment in her 2012 concurring opinion in *U.S. v. Jones*, 565 U.S. 945, 957 (2012) (Sotomayor Concurring), "[I]t may be necessary to reconsider the premise that an individual has no reasonable expectation of privacy in information voluntarily disclosed to third parties[.]" "This approach is ill suited to the digital age, in which people reveal a great deal of information about themselves to third parties in the course of carrying out mundane tasks."

[106] *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701, 1716-23 (2010).

[107] 45 C.F.R. 164.514(a).

If the service provider must use identified data, the employer should conduct due diligence on the service provider to confirm that it can protect the data and obtain written assurances that the service provider will provide reasonable safeguards for that information. In some circumstances, the employer may be legally required to address information security in the service agreement. California, for example, has enacted a statute requiring any business that owns or licenses personal information regarding a California resident and that shares such information with a third party to "require by contract that the third party implement and maintain reasonable security procedures and practices appropriate to the nature of the information, to protect the personal information from unauthorized access, destruction, use, modification, or disclosure.[108] Massachusetts and Oregon have enacted similar statutes regarding their residents.[109]

Employers also should note that HIPAA mandates data security provisions regarding protected health information in contracts between any employer covered by HIPAA and a third-party service provider, a "business associate" in HIPAA parlance, which will handle the employer's protected health information.[110] Such contracts must establish the permitted and required uses and disclosures of the protected health information, including appropriate safeguards to prevent unauthorized disclosures, and reporting requirements in the event of any such unauthorized disclosure.[111]

For those employers that are not required by law to address information security in service agreements with data service providers, a good practice is to require that service providers safeguard data by contract anyway. One third of states require companies to implement general safeguards to protect some forms of personal information.[112] Forty-seven states require notification of a breach of security, including when the personal information is held by a service provider.[113] Consequently, employers in many states could be liable not only for failing to report a breach, but also for the absence of safeguards that led to the breach, even if the breach was caused by a service provider.

Indeed, a company could theoretically face penalties from government regulators for a service provider's failure to apply reasonable security measures even without a breach. In practice, government regulators are unlikely to crack down on a company for its service providers' failure to maintain the required safeguards. If a company fails to conduct due diligence, however, and the service provider suffers a data breach, the company may find itself facing an enforcement action. The company may also become the target of class actions alleging negligence. Although negligence claims in cases of data breaches have generally foundered on the elements of harm and causation, defending against the claims can be costly.[114] Leaving aside legal liability, a breach may be embarrassing for the company and undermine employee morale.[115] Security breaches also can be costly due to the expenses involved in providing notification in accordance with breach notification laws.[116]

The company can delegate breach notification to the service provider by contract, but the company retains the responsibility under breach notification laws to ensure that notifications are provided to affected individuals.[117] Therefore, in addition to requiring that the service provider implement reasonable data security safeguards, the employer should require the service provider to promptly report any data breach to the employer and to cooperate with the employer in the data breach investigation. The contract should also provide that the service provider will reimburse the employer for all costs incurred by the employer when responding to a security breach involving personal information in the service provider's possession and indemnify the employer from any third-party claims arising out of the security breach. Finally, the contract should clearly indicate which party will provide breach notifications and give the employer the right to supervise the notification process.

[108] *See* Cal. Civ. Code 1798.81.5(c).

[109] M.G.L. c. 93H as implemented by 201 C.M.R. 17.00; Or. Rev. Stat. 646A.622(2)(d).

[110] 45 C.F.R. 164.504(e).

[111] *Id.*

[112] *See, e.g.,* Fla. Stat. § 501.171(2); Tex. Bus. & Com. Code Ann. § 521.052(a).

[113] *See, e.g.,* 815 Ill. Comp. Stat. 530/5 *et seq.*; Mich. Comp. Laws § 445.72; N.Y. Bus. Law § 899-aa; Ohio Rev. Code § 1349.19.

[114] Douglas H. Meal & David T. Cohen, *Private Data Security Litigation in the United States*, in Inside the Minds: Privacy and Surveillance Legal Issues (Aspatore 2014).

[115] According to the Ponemon Institute's 2014 survey on data breach costs, businesses lost over three million dollars' worth of business on average after experiencing a data breach. Ponemon Institute, *2014 Cost of Data Breach Study: Global Analysis*, 16 (May 2014).

[116] The Ponemon Institute also estimated that the average cost of breach notification in the United States in 2014 was over $500,000. *Id.* at 15.

[117] For example, New York, like many other states, imposes a duty to notify affected individuals on the party that "owns or licenses" the data. N.Y. Gen. Bus. Law § 899-aa(2). A party that "maintains" the data only has the obligation to notify the data's owner or licensor. *Id.* at § 899-aa(3).

International Data Protection

Although a full discussion of international data protection regimes is beyond the scope of this paper, employers should exercise particular caution when analyzing the data of non-U.S. employees, especially employees who reside in the European Union. All countries in the European Union have enacted laws to implement the European Union Data Protection Directive, which tightly regulates the processing of personal data.[118] Many other countries have adopted data protection laws similar to those of the European Union.[119]

The Data Protection Directive is broad. For instance, "personal data" is defined as: "any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity."[120] This definition covers a much wider array of information than most U.S. legal definitions of personal information, which limit protection to specific data points, such as Social Security numbers or protected health information subject to HIPAA. In the context of data analytics, the broader definition is particularly noteworthy because categories of information that generally are not protected under U.S. law, such as performance appraisals, records of discipline, and compensation information, are protected under the Directive.

Critically, when discussing data analytics, the E.U. Data Protection Directive forbids decisions based on the automated processing of personal data except in certain circumstances.[121] Making decisions based on data-driven analysis of employee personal data could potentially violate this prohibition. As a result, "Moneyball"-like techniques[122] may court considerable risk in the E.U. even if all the other requirements that we discuss below are met.

Among other points, the E.U. Data Protection Directive requires notice regarding the processing of personal data, and unless an exception applies, also requires consent.[123] This means that employers must provide notice to employees regarding how their personal data is processed, including the processing of their personal data to conduct data analytics. Employers would have to notify their employees about the purpose of the analysis and whether a third party was conducting the analysis.[124] Employees generally have the right to object to data processing, and could refuse to let the employer conduct the analysis.[125]

In addition to granting the subjects of personal data the right to object to processing, the E.U. Data Protection Directive grants them the right to access, correct, and delete their personal data.[126] The Directive also requires reasonable security for personal data.[127] Consequently, just as it should when sharing the personal data of U.S. employees with a service provider, the employer should conduct due diligence on service providers and execute agreements containing data security provisions and provisions addressing these obligations before disclosing the personal data of E.U.-based employees to a service provider.

Employers cannot simply bypass the E.U.'s restrictions by processing the data on U.S. territory. The E.U. Data Protection Directive forbids the transfer of E.U. residents' data to countries where the local law does not provide "an adequate level of protection."[128] As of now, the E.U. has determined that the U.S. law generally does not ensure an adequate level of protection.[129] Companies may, however, may provide an adequate level of protection for data transferred to the U.S. by implementing one of three mechanisms to assure that the personal data will

[118] *See* Commission Directive 95/46/EC, of the European Parliament and the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and the Free Movement of Such Data, 1995 O.J. (L281) 31 [hereinafter E.U. Data Protection Directive].

[119] A few examples of non-E.U. countries with broad data protection laws are Australia (The Privacy Act 1988 (Cth)), India (Information Technology Act, 2000, No. 21 of 2000, as amended by Information Technology (Amendment) Act, 2008 and Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, G.S.R. 313(E) (Apr. 11, 2011)), Mexico (Ley Federal de Protección de Datos Personales en Posesión de los Particulares, 5 de Julio de 2010), and South Korea (Personal Information Protection Act, Act No. 10465, Mar. 29, 2011).

[120] E.U. Data Protection Directive, ch. I, art. 2(a).

[121] Id., ch. II, art. 15.

[122] In an approach made famous by the book "Moneyball" and the movie of the same name, the Oakland Athletics baseball team used data-driven analytical techniques to determine what quantifiable measures made baseball players successful and then recruited players based on these statistics.

[123] *Id.*, ch. II, arts. 10, 14.

[124] *See id.*, ch. II, arts. 10(a), (b).

[125] *See id.*, ch. II, art. 14.

[126] *Id.*, ch.II, art. 12.

[127] *Id.*, ch. II, art. 17.

[128] *Id.*, ch. IV, art. 25(1).

[129] The E.U. Commission's list of countries with an adequate level of protection is available here: http://ec.europa.eu/justice/data-protection/document/international-transfers/adequacy/index_en.htm.

continue to be subject to E.U.-like protections once transferred to the U.S.: model data transfer agreements;[130] Binding Corporate Rules;[131] or certifying to the U.S.-E.U. Safe Harbor.[132] The most commonly used of these approaches is the U.S.-E.U. Safe Harbor. Certifying to the U.S.-E.U. Safe Harbor effectively requires the certifying company to implement a privacy framework for transferred personal data and to implement information security safeguards that mirror those required by the Data Protection Directive.[133]

The Federal Trade Commission (FTC) enforces the Safe Harbor.[134] Until now, the FTC's enforcement has been relatively lax; the agency has focused principally on companies that represent that they are still Safe Harbor-certified after their certification has expired. Recently, the FTC has come under pressure from European regulators for inadequately enforcing the Safe Harbor. As a result, the FTC may step up enforcement efforts and target more substantive violations.

The European Union is also likely to tighten rules on processing personal data. The European Commission has proposed comprehensive reform of the Data Protection Directive to bolster privacy protections for personal data.[135] Among other points, the latest proposal calls for fines of 2-5% of a noncompliant company's global annual turnover, a right of consumers to have unnecessary data deleted (the "right to be forgotten"), and mandatory reporting of data breaches to state authorities.[136] The Commission opted for a Regulation instead of a Directive, because no transposition into local law will be required, and the Regulation will directly and equally apply in all Member States.[137]

The exact wording of the draft Regulation is being negotiated by the members of the European Parliament and the member state delegations. A final text is expected by the end of 2015.[138] After official publication, a two year transitional period will apply.[139] The new data protection regime likely will have a significant impact on employers' ability to use the personal data of E.U.-based employees to conduct data analytics.

BIG DATA: IS THERE A DEFENSE TO A POTENTIAL ADVERSE IMPACT?

Title VII of the Civil Rights Act of 1964, as amended,[140] the Age Discrimination in Employment Act,[141] and the Americans with Disabilities Act,[142] all prohibit disparate impact discrimination. The gist of this claim is the same under each statute—a *prima facie* case requires a plaintiff to (a) identify a facially neutral policy or practice, (b) prove that this policy or practice adversely impacts members of a specific protected group, and then (c) demonstrate that this causes an adverse employment action affecting the plaintiff.[143] Big Data lends itself to this proof.

One of the consequences of Big Data is that small differences between groups may be "statistically significant." "Statistical significance" is the criterion many courts use to assess proof of an adverse impact.[144] However, statistical significance is a flexible yardstick and, other

[130] Commission Decision 2001/497/EC Controller to Controller Transfers (amended by Commission Decision C(2004) 5271).

[131] Working Document: Transfers of personal data to third countries: Applying Article 26 (2) of the EU Data Protection Directive to Binding Corporate Rules for International Data Transfers (Adopted June 3, 2003).

[132] Commission Decision 2000/520/EC of 26 July 2000 Pursuant to Directive 95/46/EC of the European Parliament and of the Council on the Adequacy of the Protection Provided by the Safe Harbor Privacy Principles and Related Frequently Asked Questions Issued by the US Department of Commerce, 2000 O.J. (L 215) 7-47.

[133] *See id.* at Art. 1, § 3.

[134] *Id.*, Annex VII.

[135] European Commission, *Data Protection Day 2015: Concluding the EU Data Protection Reform essential for the Digital Single Market* (Jan. 28, 2015) *available here* http://europa.eu/rapid/press-release_MEMO-15-3802_en.htm.

[136] *Id.*

[137] *Id.*

[138] On June 15, 2015, the EU Council agreed on a "general approach" to proposing a final draft of the Regulation. http://www.consilium.europa.eu/en/press/press-releases/2015/06/15-jha-data-protection/ (last visited July 14, 2015).

[139] European Parliament, *Q&A on EU data protection reform*, Mar., 3, 2014, *available at* http://www.europarl.europa.eu/news/en/news-room/content/20130502BKG07917/html/QA-on-EU-data-protection-reform.

[140] Title VII of The Civil Rights Act, 42 U.S.C. § 2000e, *et seq.*

[141] Age Discrimination in Employment Act of 1967 ("ADEA"), 29 U.S.C. §§ 621-634 (2000).

[142] Americans with Disabilities Act of 1990 ("ADA"), 42 U.S.C. §§ 12101-12213 (2000).

[143] Courts apply the burden-shifting framework to claims under each of these statutory frameworks. *See e.g.,* 42 U.S. C. § 2000e-2(k) (Title VII of The Civil Rights Act); *Shelley v. Geren*, 666 F.3d 599, 607-608 (9th Cir. 2012) (ADEA); *Roggenbach v. Touro College of Osteopathic Medicine*, 7 F. Supp. 3d 338, 343-44 (2014) (ADA).

[144] *See, e.g., Contreras v. City of Los Angeles*, 656 F.2d 1267 (9th Cir. 1981); *Waison v. Port Authority*, 948 F.2d 1370, 1376 (2d Cir. 1991); *Ottaviani v. State Univ. of N.Y. at New Paltz*, 875 F.3d 365, 370-371 (1989); *Sobel v. Yeshiva Univ.*, 839 F.2d 18 (1988).

things equal, any disparity between two groups will increase in statistical significance the larger the sample on which the analysis is based.[145] Thus, Big Data may identify selection criteria that are statistically significant, although in practical terms the difference between success and failure may be quite small. In the landmark case of *Wal-Mart Stores, Inc. v. Dukes,*[146] the large number of promotions would cause a disparity in the number of men and women promoted of just seven-tenths of one percent to be judged "statistically significant."[147]

Once the adverse impact of the selection criterion is established, the plaintiff next must prove the algorithm caused her to suffer an adverse employment action.[148] The question is whether, if the algorithm had valued this candidate more highly, the plaintiff would have been more likely to be selected? This too can be established statistically, by comparing the selection rate among those who score more favorably than the plaintiff with those who score no more favorably. If a statistically-significant difference exists, a fact-finder reasonably may conclude that the algorithm caused an adverse employment action.

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.[149] This may be Big Data's greatest challenge. Some of its most vocal advocates contend Big Data is valuable precisely because it crunches data that are ubiquitous and *not* job-related. 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 the information gleaned from any all sources and job performance. The legal question is whether an employer can meet its burden to prove job-relatedness on the basis of evidence that is strictly correlational.

The Uniform Guidelines on Employee Selection Procedures, although published in 1978, continue to inform how courts view validation.[150] An overarching principle is that an employer generally will not be able to establish validity based upon the job performance of employees who work elsewhere, except in particular circumstances. In order to "transport" statistical findings from one workplace to another, the employer must demonstrate 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."[151] The regulations contemplate a comparison between the job duties of the subjects of the validation study and the job duties of those the selection procedure will screen, which should be similar in material respects.[152]

The Uniform Guidelines approve three types of validation studies: criterion, content, and construct validity studies.[153] 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 that job.[154] The typing test given to a prospective typist is the paradigm, although even here it would be important to demonstrate that the text on which the examination is given is similar to the text that must be 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 performance. On its face, the data relied upon by the algorithm, and the algorithm itself, are likely to be far-removed from the tasks the job requires. Thus, content validity is an unlikely method for validating Big Data.

Construct and criterion validity are closely related. Construct validity measures the degree to which candidates have "identifiable characteristics which have been determined to be important for successful job performance."[155] Sometimes these traits may be apparent. Other things the same, speed is likely to be a substantial asset to a football player. In other settings, identifying the salient traits is more

[145] "[L]arger sample sizes give more reliable results with greater precision and [statistical] power . . ." *The Importance and Effect of Sample Size*, Select Statistical Services, http://www.select-statistics.co.uk/article/blog-post/the-importance-and-effect-of-sample-size.

[146] 131 S. Ct. 2541 (2011).

[147] Allan G. King, *"Gross Statistical Disparities" as Evidence of a Pattern and Practice of Discrimination: Statistical versus Legal Significance*, 22 The Labor Lawyer 271, 280 (2007).

[148] *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 802-804 (1973).

[149] *Id.*; 42 U.S.C. § 2000e-2(k)(1)(A)(i).

[150] "There are two sources of expertise upon which the courts often rely in deciding whether a test has been properly validated . . . Perhaps the most important source of guidance is the Equal Employment Opportunity Commission's 'Uniform Guidelines on Employment Selection Procedures' . . ." *Gulina v. N.Y. State Educ. Dept.*, 460 F.3d 361, 383 (2d Cir. 2006).

[151] 29 C.F.R. § 1607.7(B)(2).

[152] *Id.*

[153] *See generally* 29 C.F.R. § 1607.14.

[154] 29 C.F.R. § 1607.14(C).

[155] 29 C.F.R. § 1607.16(E).

challenging. Accordingly, the Uniform Guidelines caution employers that "[t]he 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 . . ."[156] Thus, an employer first must 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.[157]

Criterion validity, or predictive validity as it sometimes is known, differs in that the objective is to predict ultimate success on the job, rather than traits believed to lead to success. Because this method of validation was long-standing when the Uniform Guidelines were formulated, the regulations governing criterion validity are more detailed than those pertaining to construct validity. The Guidelines list several steps they deem "essential," many of which pertain to a "job analysis."[158] A job analysis should identify the behaviors or outcomes that are critically important, the proportion of time spent on each, their difficulty, the consequences of errors, and the frequency with which various tasks are performed.[159] The purpose in systematizing this information is to determine which jobs reasonably may be grouped to rate employees for their proficiency and identify a common test or screen for selecting them.[160] Employers must also explain the bases for selecting the success measures and the means by which they were observed, recorded, evaluated, and quantified.[161]

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 procedure and performance on the criterion measure is statistically significant at the 0.05 level of significance."[162] Generally, there are two methods of establishing either construct or criterion validity. One is "concurrent validity"; the other is "predictive validity."[163] 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. For example, an incumbent workforce, whose job performance can be evaluated, 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 but not used in hiring decisions. Some portion of the applicant pool is hired, and subsequently evaluated in terms of on-the-job performance. The selection scores are then correlated with measures of performance to assess whether selection scores predicted performance accurately.[164]

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 will be explained), there is no reason the correlations that underlie Big Data solutions should persist beyond the sample period. Because concurrent validity is based upon information regarding incumbent employees, the correlations uncovered regarding these individuals will be relevant to the applicant pool only if incumbents and applicants, are similar in the 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 a generation of *incumbents* 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,[165] will be relevant only if patterns observed in the past continue to be relevant to job performance. 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.

[156] 29 C.F.R. § 1607.14(D).

[157] *Id.*

[158] 29 C.F.R. § 1607.15(B)(3).

[159] *Id.*

[160] *Id.*

[161] 29 C.F.R. § 1607.15(B)(5).

[162] 29 C.F.R. § 1607.14B(5).

[163] 29 C.F.R. § 1607.14(B)(4).

[164] Richard Jeanneret, *Professional and Technical Authorities and Guidelines*, in Employment Discrimination Litigation: Behavioral, Quantitative, and Legal Perspectives 47, 58 (Frank Landy, 2005).

[165] Dr. Jeanneret recommends assessing performance no sooner than six months after hire. *Id.*

Underlying each validation method approved by the Uniform Guidelines is the requirement of a job analysis.[166] This reflects the common-sense view that to design a test or selection instrument that distinguishes those best able to perform a job from those who are least able requires 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."[167]

Big Data begins from the opposite perspective. 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 (the information available to it) the best distinguishes these groups. Consequently, the tests of statistical significance that ensure the ultimate validity of conventionally developed tests are far less relevant to Big Data. Well-conceived algorithms will eliminate every alternative that is not significantly related to job performance—that is the cornerstone of the methodology. As a result, applying conventional tests of validity to Big Data makes little sense because, by design, its algorithms identify the correlates that best fit the (job performance) data, without regard to why they are related.

This is the salient difference between predictions based upon theories of causation versus those based solely on correlations. The logic of "cause and effect" is that the associated theory is based on something believed to be fundamental to the relationship in question. For example, the logic that suggests faster athletes make better football players reflects the assumption that the faster a football player can run towards or away from an opponent, the more likely that player is to be in the right place at the right time. As long as the rules of the game remain the same, this proposition should remain true.

In contrast, suppose these players were selected on the basis of social media profiles and Internet activity. It well may be possible to achieve higher correlations using this information than relying on the player's speed, but there is no telling how long those correlations will be reliable. If a new fad sweeps college campuses, today's high correlation may quickly become tomorrow's zero correlation. Consequently, the correlation demanded by the Uniform Guidelines may be a poor criterion by which to judge the performance of Big Data.

Thus, Big Data effects a shift from selection criteria distilled from job-related knowledge, skills, and ability, which uses correlation to establish that the correct criteria were identified, to one in which correlation is established at the outset, independently of knowledge, skills, and ability, and leaves the duration of that correlation in question. Accordingly, rather than assessing Big Data in terms of correlation, which it will pass with flying colors, the relevant question is the probable duration of the correlations on which its algorithms are based. In terms of validation, this translates into a comparison between the time elapsed since the algorithm was first calibrated and the time it was applied to the plaintiff, relative to the probable 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 correlation should be greatest when the algorithm initially is calibrated and should decay as time passes. But how much decay is tolerable before the algorithm is deemed too unreliable to pass legal scrutiny? An answer is suggested by the Uniform Guidelines: "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 0.05 level of significance . . ."[168] 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 0.05 level. This lifespan therefore is the critical inquiry in assessing how long a Big Data algorithm lawfully is applied to an employer's workforce.

LITIGATION IN A WORLD OF BIG DATA

Class Action Risks and Exposure

Class actions pose significant risks to employers. On March 17, 2015, it was reported that "[t]op legal counsel at nearly 350 companies managed on average five class actions in 2014 and spent $2 billion on class actions."[169] The same survey reported that 23 percent of these cases are employment class actions.[170] Thus, whether Big Data adds to these risks should be an important inquiry for in-house counsel and employers. An important consideration, therefore, is whether classes challenging Big Data are more likely to be certified.

[166] *See generally* 29 C.F.R. § 1607.15.

[167] 29 C.F.R. § 1607.14(A).

[168] 29 C.F.R. § 1607.14(B)(5).

[169] Melissa Maleske, *GCs Facing More Class Actions, Higher-Exposure Cases*, Law 360 (Mar. 17, 2015) http://www.law360.com/articles/632403/gcs-facing-more-class-actions-higher-exposure-cases (last visited July 14, 2015).

[170] *Id.*

Class actions are driven by common questions with common answers.[171] The Supreme Court reversed class certification, in *Wal-Mart Stores, Inc. v. Dukes*, because the decisions that were challenged as discriminatory were decentralized, and did not reflect the effects of a policy that was commonly applied to class members.[172] Consequently, resolving these claims would have required determining whether each class member was treated in a discriminatory manner.[173]

The implication of *Dukes* is that policies that are more centralized, and applied uniformly, enhance the likelihood a class will be certified. This suggests that a Big Data algorithm, applied uniformly and consistently throughout an employer's workforce, potentially provides the "glue" missing from *Dukes*. However, that conclusion may be too facile.

Big Data is a methodology rather than a particular selection device. It is a data-intensive means of distinguishing promising from unpromising employees, based upon correlations between measures of success and other information a developer can glean from various sources. If Big Data is the method, then the algorithm is the predictive model that is designed to select the most promising employees. As opposed to the Big Data methodology, the algorithm essentially is the equation that sorts applicants or incumbent employees into groups with greater or lesser promise. The algorithm therefore is a legitimate target of a discrimination lawsuit.[174]

In terms of class action litigation, and the critical question of class certification, the corresponding issue is whether decision-making by algorithms enhances the risk of class certification. Superficially, it may seem that the discriminatory impact of a commonly-applied algorithm is precisely the question amenable to a common answer, and a prime target of class action litigation. However, the "algorithm" may vary by day, week, or month, depending upon how long any one remains "best," and may have as many variations as the individualized decisions that caused *Dukes* to be decertified.

But first let's consider the simplistic case in which an employer adopts a single, unchanging algorithm it applies across-the-board to all applicants and those seeking promotion. Subject to a plaintiff discovering the demographic characteristics of applicants, this appears to be a practice that lends itself to class litigation.[175] Although an employer might argue that it differentially impacts applicants for various positions, *e.g.*, engineers and accountants, this is likely to be viewed as an argument for subclasses rather than one that defeats certification.[176] Indeed, this scenario is very close to a common fact pattern in which an employer administers a standardized test to its applicants, and the test has an adverse impact on a protected group. Cases of that type frequently are certified.[177]

However, this hypothetical puts Big Data in an artificial straightjacket. One of the advantages of Big Data is the relatively low cost in mining data once it is harvested from the sources that feed it. Consequently, the cost of producing job-specific algorithms, which are likely to be more accurate than one "standard" model, may be just marginally greater than the cost of developing one grand algorithm. Thus, a class action plaintiff may be faced with challenging numerous algorithms.

The problem this creates is that no one plaintiff is likely to have been affected by each of the algorithms in use. As a result, that plaintiff may not have standing to challenge algorithms by which he or she was unaffected. If a court determines the scope of a class or subclass must be limited to those affected by the same algorithm, it may be infeasible to bring large class actions against employers whose decisions reflect an array of algorithms that differ by position.

[171] *See, e.g.*, Fed. R. Civ. P. 23(a)(2); *Brinker Restaurant Corp. v. Super. Ct.*, 53 Cal. 4th 1004, 1022 n.5 (2012).

[172] *Wal-Mart Stores, Inc. v. Dukes*, 131 S. Ct. 2541, 2554-55 (2011). In this suit, the U.S. Supreme Court reversed the class certification of a gender discrimination claim that included a class of 1.6 million women who currently worked or had worked for Wal-Mart stores. In particular, the plaintiff alleged that Wal-Mart engaged in discriminatory pay and promotion policy and practices.

[173] *Id.* at 2552 ("Without some glue holding the alleged reasons for all those decisions together, it will be impossible to say that examination of all the class members' claims for relief will produce a common answer to the crucial question why was I disfavored."). *See also, e.g., Duran v. US Bank Nat'l Ass'n*, 59 Cal. 4th 1, 25 (2014) ("Faced with the potential difficulties of managing individual issues . . . many trial courts have denied certification or decertified the class before trial . . . [S]uch decisions have been routinely upheld.").

[174] *See, e.g., Griggs v. Duke Power Co.*, 401 U.S. 424, 436 (1971) ("Nothing in the [Civil Rights] Act precludes the use of testing or measuring procedures; obviously they are useful. What Congress has forbidden is giving these devices and mechanisms controlling force unless they are demonstrably a reasonable measure of job performance.")

[175] "Claims alleging that a uniform policy consistently applied to a group of employees is in violation of the wage and hour laws are of the sort routinely, and properly, found suitable for class treatment." *Brinker*, 53 Cal. 4th at 531.

[176] "When appropriate . . . a class may be divided into subclasses and each subclass treated as a class, and the provisions of this rule [23(c)(4)(B)] shall then be construed and applied accordingly." Fed. R. Civ. P. 23(c)(4)(B).

[177] *See, e.g., Griggs*, 401 U.S. 424 (certifying class action against employer that required a high school education or passing a standardized general intelligence test as condition of employment, when neither was significantly related to job performance and both disqualified black applicants at a substantially higher rate than white applicants).

Apart from algorithms specific to particular jobs, algorithms may, and probably should, change periodically. As a result, employees hired or promoted at different times may have been selected by different algorithms. Developers may find it useful to update algorithms—not to avoid litigation, but because Big Data algorithms are based on correlation and not causation.[178] This means that the algorithms they create are useful only as long as the correlations on which they are based remain substantial.

For example, an article in the *Atlantic* describes how one company searched for software engineers proficient in writing computer code:

> 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 language and online behavior of programmers in good positions at prestigious companies.[179]

The article explains this information then can identify promising programmers whose code is not available on the Internet by determining whether their social media footprint is similar to that of the best open-source programmers, by comparing their on-line histories.[180] The chief scientist with this company explained, "They're not all obvious, or easy to explain . . . [O]ne solid predictor of strong coding is an affinity for a particular Japanese manga site."[181] However, any website ultimately may be of passing interest to a group that once found it an obsession.

If correlations unearthed by Big Data are ephemeral, then their algorithms must be dynamic to maintain validity. This requires constantly updating the profiles of individuals best qualified for various positions, which may lead to new algorithms. Consequently, an employer's hiring decisions might be predicated on an ever-changing array of algorithms, with only minimal commonality. As a result, the likelihood of common answers is drastically diminished because each version of the algorithm may have different properties, in terms of both adverse impact and its degree of validity. In principle, one version might impact neutrally and another might be highly predictive, although its adverse impact is greater. Thus, the dynamic differences in Big Data may be the contemporary counterpart of the store-to-store differences that precluded class certification in *Dukes*.[182]

eDiscovery and Big Data Algorithms

The Big Data Deluge Springboards eDiscovery to the Spotlight

One of the most immediate impacts Big Data has had on employers is on eDiscovery in litigation. With continued advancements in technology and the ongoing digitization of the global workplace, electronic records—whether they are "born digital" or "born analog,"[183]—are more than ever at the center of lawsuits. As examples, in recent submissions to the Civil Rules Committee on Rules of Practice and Procedure of the Judicial Conference of the United States in support of proposed amendments to the Federal Rules of Civil Procedure to address eDiscovery preservation and sanctions:[184]

- A large technology company reported the following statistics regarding electronic data for its "average" case:[185]
    i. it preserves the equivalent of over 48 million pages of written documents;

[178] Chris Taylor, *Big Data's Slippery Issue of Causation vs. Correlation*, Wired (Jul. 15, 2013), available at http://insights.wired.com/profiles/blogs/big-data-s-slippery-issue-of-causation-versus-correlation#axzz3WY6JsHAy (last visited July 14, 2015).

[179] Don Peck, *They're Watching You at Work*, The Atlantic (Nov. 20, 2013), http://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/ (last visited July 14, 2015).

[180] *Id*.

[181] *Id*. Manga are comics created in Japan, or by Japanese creators in the Japanese language, conforming to a style developed in Japan in the late 19th century. http://en.wikipedia.org/wiki/Manga (last visited July 14, 2015).

[182] Some of the store-to-store differences cited in *Dukes* include the availability of women, qualified women, or interested women, as well as the nature and effects of sex-neutral, performance-based criteria in each specific location. *Dukes*, 131 S. Ct. at 2555.

[183] *See* Big Data: Seizing Opportunities, Preserving Values, Executive Office of the President, May 2014, p. 4, which notes (footnotes omitted):

> There is not only more data, but it also comes from a wider variety of sources and formats. As described in the report by the President's Council of Advisors of Science & Technology, some data is "born digital," meaning that it is created specifically for digital use by a computer or data processing system. Examples include email, web browsing, or GPS location. Other data is "born analog," meaning that it emanates from the physical world, but increasingly can be converted into digital format. Examples of analog data include voice or visual information captured by phones, cameras or video recorders, or physical activity data, such as heart rate or perspiration monitored by wearable devices.

[184] Each of these submissions is available on the "Regulations.gov" website that can be accessed via http://www.uscourts.gov, and in particular the "Submit or Review Comments on the Proposed Amendments to the Federal Rules of Civil Procedure," link at http://www.uscourts.gov/RulesAndPolicies/rules/proposed-amendments.aspx.

[185] Microsoft Corp. Comment, Aug. 31, 2011, pp. 4-5.

ii. it collects the equivalent of about 13 million pages of written documents; and

iii. it reviews the equivalent of over 645,000 pages of written document.

- A global pharmaceutical company reported that in one products liability multidistrict litigation ("MDL"), it produced over 90 million pages of data/documents, while in another MDL matter it produced over 50 million pages of data/documents.[186]
- Another global pharmaceutical company reported that for the 12-month period ending October 1, 2013, it collected roughly 1 billion pages from 3,000 custodians in connection with as many as 60 ongoing litigation matters. Of the 1 billion pages collected, approximately 140 million pages were identified as potentially responsive to discovery requests, roughly 25 million pages were produced, of which 5.5 million required at least one redaction.[187]

Coupled with the staggering volumes of data created every day,[188] and the Judiciary's lower tolerance for eDiscovery ignorance or misconduct when handling electronically stored information ("ESI") in litigation,[189] the impact of Big Data on 21st century litigation will only continue to grow.

In direct response to the Big Data phenomenon, new technologies, sometimes referred to as "Predictive Coding," are emerging for "finding the needle in the haystack"[190] to respond to discovery in litigation. However, as discussed more fully below, employers must understand the legal risks and limitations that are presented by those tools until a more settled legal framework emerges.

The Big Data phenomenon has also led to changes in the ethical framework for lawyers along with amendments to the Federal Rules of Civil Procedure that are designed to align those standards with modern technological advances and the massive volumes of data that are now commonplace.

The Next Horizon for Analyzing Big Data in Litigation: TAR/Predictive Coding

Predictive Coding, also referred to as Technology Assisted Review ("TAR"), has been widely touted as a solution for analyzing Big Data in litigation. In short, TAR involves reviewing a relatively *small percentage* of documents and then—using technology—extrapolating those results to *all of the documents* in a data set (*e.g.*, in a case that has 1,000,000 documents, reviewing a sample of 30,000 for responsiveness and then using a computer algorithm to apply those results across the remaining 970,000 documents).

[186] Bayer Corp. Comment, Oct. 25, 2013, pp. 2–3.

[187] Pfizer Inc. Comment, Nov. 5, 2013, p. 4.

[188] It was estimated that 3.5 Zettabytes (3.5x10,21 or 3,500,000,000,000,000,000,000 bytes) of unique information would be created worldwide in 2014, which is the equivalent of all of the words spoken by all of humanity since the beginning of time—times 200. *See* "Business Increasingly Relying on Big Data Analytics," New Horizons Computer Learning Centers, June 2, 2014 (available athttp://www.newhorizons.com/IT-Career-Development-News/627869/Business-increasingly-rely-on-big-data-analytics/)) (last visited July 14, 2015); Mark Lieberman, "Zettascale Linguistics," Nov. 17, 2003 (available at http://itre.cis.upenn.edu/myl/languagelog/archives/000087.html) (last visited July 14, 2015) (estimating that all of the words spoken by human beings is the equivalent of 5 Exabytes of data).

It has also been estimated that 90% of the world's data has been generated in the past two years. *See* "Big Data, for better or worse: 90% of world's data generated over last two years," Sci. Daily, May 22, 2013, available at http://www.sciencedaily.com/releases/2013/05/130522085217.htm (last visited July 14, 2015).

[189] *See e.g., Brown v. Tellermate Holdings, Ltd.*, 2014 WL 2987051 (S.D. Ohio July 1, 2014), in which the court instructed:

> Over the past decade, much discussion has been devoted to the topic of how the prevalence of electronically stored information (ESI) either has impacted, or should impact, discovery in civil actions filed in state and federal courts. While the preservation, review, and production of ESI often involves procedures and techniques which do not have direct parallels to discovery involving paper documents, the underlying principles governing discovery do not change just because ESI is involved. Counsel still have a duty (perhaps even a heightened duty) to cooperate in the discovery process; to be transparent about what information exists, how it is maintained, and whether and how it can be retrieved; and, above all, to exercise sufficient diligence (even when venturing into unfamiliar territory like ESI) to ensure that all representations made to opposing parties and to the Court are truthful and are based upon a reasonable investigation of the facts.

Courts have also held that technical ignorance or incompetence of counsel do not excuse eDiscovery mistakes. *See, e.g., Small v. Univ. Medical Center of Southern Nevada,* 2014 WL 4079507 (D. Nev. Aug. 18, 2014) ("Ignorance of technology does not excuse counsel or clients from carrying out their duties to preserve and produce ESI. … Today, ignorance of technology is simply an inadequate excuse for failure to properly carry out discovery obligations."); *Garcia v. Berkshire Life Ins. Co. of America,* 2007 U.S. Dist. LEXIS 86639 (D. Colo. Nov. 13, 2007) (holding that technical incompetence, mistake or ignorance of counsel is not a good faith defense to a motion to compel, and rejecting argument that because "Plaintiff's counsel does not employ a full time computer technician, occasionally a technology issue arises which exceeds Plaintiff's computer expertise."); *Martin v. Northwestern Mutual Life Ins. Co.,* 2006 W.L. 14899 (S.D. Fla. Jan. 19, 2006) (claim that counsel is computer illiterate and therefore incapable of retrieving emails "is frankly ludicrous" and does not serve as a good faith defense to sanctions motion).

[190] *See* Big Data: Seizing Opportunities, Preserving Values, Executive Office of the President, May 2014, p. 6 ("Computational capabilities now make 'finding a needle in a haystack' not only possible, but practical. … Big data analytics enable data scientists to amass lots of data, including unstructured data, and find anomalies or patterns.")

Yet, as discussed more fully below, TAR tools and methodologies are not akin to hitting an "Easy Button®"[191] whereby a large amount of data is dumped into the tool on the front end, and a nicely-packaged, complete, smaller volume of relevant and responsive data is quickly and simply exported on the back end.[192] Unfortunately, many in the eDiscovery industry have created such an impression in the market by heavily promoting TAR as a turnkey, budget-slashing "solution" to meet the challenges of Big Data in litigation for all cases large and small.

While TAR has the potential for significant cost savings and substantive benefits in litigation, TAR is very new and the legal landscape concerning the use of TAR is still developing. When deciding whether TAR is a fit for a particular case, there are a host of issues that need to be considered including the case law in the applicable jurisdiction about how and when a party can use TAR, decisions about several statistical measures that underlie any TAR methodology, whether and to what extent specifics about your TAR methodology must be disclosed to your adversary, and how involved experts will need to be in the process.

a. An Overview of TAR Methodologies

i. Threshold Considerations

While there are differences among the tools being offered by different vendors (and those differences are significant), at their core, all of these tools apply a combination of sophisticated search capabilities, statistical sampling techniques and a defined workflow to change the way lawyers approach the task of reviewing documents. The search method underlying TAR is sometimes referred to as "conceptual search." While often called "new," the technology has in fact been used in a variety of ways for many years (*e.g.*, in SPAM filters, Google searches). What is new is how such technology is now being used to help attorneys make relevance determinations in litigation on a broad set of documents,[193] based upon the review of a limited portion of that set. Because relevancy determinations in litigation are controlled by obligations imposed by the Federal Rules of Civil Procedure (or a state equivalent) and the rules of ethics, using TAR in litigation to identify responsive documents in discovery is very different than conducting a run-of-the-mill search in Google, or as one influential Judge observed: "Searching for *an* answer on Google (or Westlaw or Lexis) is very different from searching for *all* responsive documents in the … eDiscovery context. "[194]

Unlike a traditional or linear document review, statistics are an integral part of any TAR process.[195] In every case involving TAR, choices must be made about the "confidence level" and "margin of error" used in defining the initial small "seed set" of documents that humans review. Because those choices will dictate the size of the samples, a "confidence level" and a "margin of error" to be applied at various points in the process must be chosen. Part of that analysis should also include issues surrounding "precision," "recall" and "F-Scores," which are statistical calculations that can be used to measure the success of the TAR process.[196] Those decisions allow a case team to review and code a

[191] "Easy Button®" is a registered trademark used for Retail Store Services, Mail Order Catalog Services, and Computerized Online Retail Store Services Featuring Office Supplies, Office Equipment, Including Computer Hardware, Copiers and Telephones, and Office Furniture and owned by Stapes the Office Superstore, LLC.

[192] In reality, implementation of TAR workflows and tools requires a significant investment of human capital and other resources, and discussed below. Indeed, a recent study conducted by Oracle, several Stanford University professors and the non-profit Electronic Discovery Institute concluded (emphasis supplied):

> The first phase of the highly-anticipated Oracle/Electronic Discovery Institute joint research project has been completed, and confirms what many advocates have been preaching about technology-assisted review (aka predictive coding)—that spending more money doesn't correlate with greater quality; that senior attorneys know what they are doing; and *that you can't turn discovery over to robots—humans are still the most vital component of the project*.

Monica Bay, *EDI-Oracle Study: Humans are Still Essential in eDiscovery*, L. Tech. News, Nov. 20, 2013 (emphasis in original). *See also*, Steve Green & Mark Yacano, *Pitting Computers Against Humans in Document Review*, L. Tech. News, Oct. 31, 2012 ("In the past few years, document review has begun to change from a heavily staffed people effort to an approach that optimizes both technology and high-level human involvement. Sometimes, though, the description and business case is pushed too far, as some eDiscovery professionals advocate for futuristic, technical methods and minimize what skilled document reviewers bring to the process…. There are reasons to doubt that the [results of a leading study cited in favor of the use of Predictive Coding over human review] can lead to generalized conclusions about the efficacy of computer-assisted methods.… Every eDiscovery tool currently available, even the most technically advanced, still depends entirely on the very 'manual' work of its operators.")

[193] TAR also has many uses beyond making responsive/non-responsive decisions in litigation. For example, TAR can be used to perform early case assessment, prioritize the order in which documents are reviewed, and to quickly analyze a large incoming production.

[194] *National Day Laborer Org. Network v. United States Immigration & Customs Enforcement Agency*, 2012 U.S. Dist. LEXIS 97863 (S.D.N.Y. July 13, 2012) (Scheindlin, J.) (emphases in original).

[195] *See* fn. 18.

[196] *See generally* Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in eDiscovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII Rich. J.L. & Tech. 11 (2011) (scholarly article that generally discusses the statistics underlying TAR methodologies and concepts such as "precision," "recall," and "F-Score"). *See also* Bill Dimm, *Predictive Coding Performance and the Silly F1 Score*, www.blog.cluster-text.com, May 5, 2013 (discussing the F-Score measure of the performance of predictive coding algorithms), available at http://blog.cluster-text.com/2013/05/08/predictive-coding-performance-and-the-silly-f1-score/ (last visited July 14, 2015).

statistically significant sample of documents, and then apply those results to the larger dataset.

TAR also requires a documented workflow that must be rigorously followed by case teams. A TAR workflow often is the inverse of a normal document review process, requiring "front loading" of high-level trial team resources (*i.e.*, Shareholder or Senior Associate level) to review and code sample documents for responsiveness.[197] The sample may be quite large, requiring a significant investment of time by senior team members. Lower-cost reviewers (*i.e.*, Junior Associates or managed contract attorneys) are only used to make substantive calls later in the process. Despite this early investment of time, the process has the potential to significantly reduce review costs in appropriate cases.[198]

ii. The TAR Process

TAR tools generally start the process by selecting a random sample of documents[199] and presenting these to a Subject Matter Expert ("SME"). The SME must be someone who knows the factual background of the case and trial strategy, and is capable of making final, binding decisions about which documents are relevant to the case. One of the biggest pitfalls of this type of process is inconsistent determinations during the learning rounds that "confuse" the TAR tool. Because of the number of rounds of learning, the SME—who is often a very senior member of a case team—may need to dedicated weeks of time to a review.

After the SME has reviewed and coded the sample documents as "Responsive" or "Not Responsive,"[200] the TAR tool uses its conceptual search feature to compare the two sets of sample documents (those coded as "Responsive" and "Not Responsive") against all the other documents in the case. The TAR tool then scores each document in the dataset based on its similarity to the documents coded by the SME as "Responsive" and "Not Responsive." The TAR tool uses those scores to predict how the SME would code the remainder of the documents in the case, without having the SME review the remaining documents. If the TAR tool cannot score a document, it is categorized as "Undetermined."

The process of the SME reviewing a sample to code documents as Responsive and Non-Responsive (called a "round") is repeated several times with multiple batches of randomly selected documents. As each round is completed, the tool "learns" how the SME would code the un-reviewed documents in the dataset and analyzes the accuracy of its predictions from earlier rounds. After each round, the review team analyzes the accuracy of the tool's predictions and makes decisions about whether additional learning rounds are necessary. The number of rounds, and the size of the sample datasets involved, varies among vendors and tools.[201] Deciding when the learning rounds are "done" is complicated and depends on a number of factors. In some cases, parties will be guided by the precision, recall, and F Score statistics

[197] *See generally*, Magistrate Judge Andrew Peck (S.D.N.Y.), "Search, Forward: Will Manual Document Review and Keyboard Searches be Replaced by Computer-Assisted Coding?", L. Tech. News (Oct. 2011):

> Unlike manual review, where the review is done by the most junior staff, computer-assisted coding involves a senior partner (or team) who review and code a "seed set" of documents. The computer identifies properties of those documents that it uses to code other documents. As the senior reviewer continues to code more sample documents, the computer predicts the reviewer's coding. (Or, the computer codes some documents and asks the senior reviewer for feedback.)

[198] On the other hand, because of the way the TAR tools work, many cases are not suitable for a TAR approach. For example, cases with a relatively small number of documents are probably ill suited for TAR because by the time the senior team member finishes training the TAR tool, he or she may have reviewed more documents than would have been reviewed by lower-level resources using a traditional keyword search process. A case that is likely to settle early may also not be a good candidate for TAR as front-loading expensive senior SME resources may disproportionately impact a case budget. And because the amount of review required is open ended, managing TAR on a fixed budget can also be problematic. TAR is also difficult to use where the scope of responsiveness is fluid and likely to shift, and for the reasons discussed more fully below, it is often important to assess the degree of disclosure that will be required, including through the use of expert testimony or disclosures, with the use of a TAR methodology.

[199] In a recent Predictive Coding protocol that was approved by a court, the parties defined the term "Statistically Valid Sample" to include:

> [A] random sample of sufficient size and composition to permit statistical extrapolation with a margin of error of +/- 2% at the 95% confidence level … The size of the sample will vary depending upon several factors, and shall be calculated using the formula
>
> n = X2 * N * P * (1- P)______
>
> (ME2 *(N—1))—(X2 * P * (1- P)),
>
> where ME is the margin of error; X is the Confidence Level (1.96 for a 95% Confidence Level); P is judgment of richness, N is the population and n is sample size. Where richness is not reasonably estimable, 0.5 may be used. Based on a Confidence Level of 95%, richness of 0.5, a Population of 1,000,000, and a margin of error of 2%, the resulting sample size is 2,395 documents.

Rio Tinto PLC v. Vale S.A., 2015 WL 872294 (S.D.N.Y. Mar. 2, 2015) (Peck, M.J.).

[200] At this stage in their development, most TAR tools are designed to identify documents as "Responsive" or "Not Responsive." Thus, most TAR tools do not predict for responsiveness to specific issues or for privilege.

[201] Some vendors are developing tools that continue the "learning" process once the initial learning rounds are completed.

discussed earlier. In other situations, principles of proportionality will be a better guide to when the review is completed.

Regardless of the measure used, once the case team concludes the learning rounds are completed and the TAR tool is adequately "trained," the TAR tools applies that learning by applying decisions that have been made during the "training" rounds on the seed sets of documents to the entire data set (*e.g.*, if the data set comprises 1 million documents and 30,000 were reviewed/included in the seed set of documents to "train" the TAR tool, the tool then uses an algorithm to apply those decisions against the remaining 970,000 documents). After this is done, all the documents in the data set will be grouped into one of three categories:

1. "Responsive;"
2. "Not Responsive;" and
3. "Undetermined."

At this point, a decision must be made about what to do with each group, and there is no clear industry consensus about how to proceed. The most typical approach is:

- First, the "Responsive" documents are reviewed by junior case team members subject to an appropriate review protocol to confirm their "Responsive" nature, identify and tag privileged documents, and code for confidentiality per the applicable Protective Order in the case.
- Next, the "Undetermined" documents must be reviewed in some manner, as the TAR tool cannot suggest the relevance of the group. This group of documents includes any documents that were excluded from the TAR process at the outset.[202] This activity is generally performed by junior case team members using a traditional search term approach, or by using other techniques such as clustering, similar document identification ("find similar"), filtering based on metadata properties, file type analysis, domain analysis and/or sampling to identify potentially "Responsive" documents.
- Finally, developing case law suggests that the "Not Responsive" group should be sampled using the same statistical sampling techniques as discussed above, and reviewed (subject to an appropriate review protocol) to determine if "Responsive" documents were miscategorized as "Not Responsive" by the TAR tool.

Eventually, all documents that have been tagged as responsive by the trial team (including those that are confirmed as responsive from their review of documents the TAR tool identified as "Responsive," as well those that are subsequently included, if any, via a review of the documents the TAR tool had identified as "Undetermined" and "Not Responsive" following the protocols discussed above) and that are not privileged are produced to your adversary (along with appropriate confidentiality designations, redactions and bates designations).

[202] Most TAR tools have an important limitation: they cannot analyze all of the files in a dataset. For example, spreadsheets that consist mainly of numbers, scanned documents or image files with no searchable text, and extraordinarily short or long documents are excluded from some vendors' processes. In addition, because of the way the conceptual search tools analyze documents, some of the content of documents may not be included in the analysis. For example, when indexing documents for a TAR analysis (which is a foundational requirement), some tools do not index basic things like the "To," "From" and "Date" fields of emails. Those fields of data may be critical to determining responsiveness or otherwise analyzing the merits of a case and the inability to index them may impact the efficacy of a TAR process. For this reason, it is critical to have a deep understanding of the functionality (and limitations) of the TAR tool that is used in a particular case.

b. The Legal Landscape with Respect to TAR is Far From Settled, Yet Will Develop Rapidly Over the Next Several Years

At the time of this publication, only a few cases[203] have specifically addressed the use of TAR, a nascent area of the law that is still developing. The level of transparency that must accompany a party's use of TAR and whether a responding party can be forced to use TAR over its objection are two areas, among others, that need to be considered when deciding on whether TAR is appropriate for a particular case.

i. Transparency and TAR

Under the Federal Rules of Civil Procedure, discovery is self-executing and a party generally does not have to defend what it has done to locate and produce responsive documents or "prove" the reasonableness of its production.[204] Those rules also are the foundation for T*he Sedona Principles: Best Practices, Recommendations & Principles for Addressing Electronic Document Production*[205] Principle 6 ("Principle 6"), which provides "[r]esponding parties are best situated to evaluate the procedures, methodologies, and technologies appropriate for preserving and producing their own electronically stored information."[206] The comments to Sedona Principle 6 explain it is the producing party's responsibility to identify information responsive to discovery requests and to produce relevant, non-privileged information.[207] The

[203] In addition to the cases discussed *infra*, the following cases have also addressed TAR:

- *Global Aerospace, Inc. v. Landow Aviation, L.P.*, 2012 Va. Cir. LEXIS 50, at *2 (Va. Cir. Ct. Apr. 23, 2012) (In response to defendants' motion requesting use of predictive coding or requiring plaintiff to pay for linear human review, that was opposed by plaintiffs because "[n]o computer program is an adequate substitute for having human beings review and sort the documents," the court permitted predictive coding, but gave no directives on how to implement it. Specifically, the court ordered that defendants were permitted to use predictive coding for "processing and production of electronically stored information … without prejudice to a receiving party raising with the Court an issue as to completeness or the contents of a production or the ongoing use of predictive coding."). *See also Predictive Coding—A Dispatch From the Front Lines of eDiscovery*, available at http://www.jonesday.com/predictive_coding/ (As discovery deadline was fast approaching, defendants in *Global Aerospace*, who had aggressively pushed for predictive coding, sought extension because "[they] were seriously delayed in commencing predictive coding due to difficulties in processing such a large amount of data." The production never became an issue because the case settled shortly thereafter);

- *In re Actos (Pioglitazone) Prods. Liab. Litig.*, 2012 U.S. Dist. LEXIS 187519 (W.D. La. July 27, 2012) (entering case management order pursuant to agreement of the parties that included a detailed "Search Methodology Proof of Concept" that involved the parties meeting and conferring to "evaluate the potential utility of advanced analytics as a document identification mechanism," including specifics about the "Control" and "Training" sets that would be used to reach "Stability" of the mutually agreed upon vendor's TAR software and coding process, but noting that "[w] hile the Parties agree to explore the use of advanced analytics as a technique to ensure appropriate responses to discovery requests, the Parties agree that Defendants retain the right to review documents after predictive coding but prior to production for relevance, confidentiality, and privilege.");

- *EORHB, Inc. v. HOA Holdings L.L.C.*, C.A. No. 7409-VCL (Del. Ch.) (transcript dated Oct. 15, 2012) (court *sua sponte* orders: "This seems to me to be an ideal non-expedited case in which the parties would benefit from using predictive coding. I would like you all, if you do not want to use predictive coding, to show cause why this is not a case where predictive coding is the way to go."); 2013 WL 1960621 (Del. Ch. May 6, 2013) ("WHEREAS, the parties have agreed that, based on the low volume of relevant documents expected to be produced in discovery … the cost of using predictive coding assistance would likely be outweighed by any practical benefits of its use … THEREFORE [it is] ORDERED that … Plaintiffs may conduct document review using traditional methods");

- *Fosamax Alendronate Sodium Drug Cases*, No. JCCP 4644 (Orange Co. Cal. Sup. Ct.) (Apr. 18, 2013) (denying plaintiffs' motion to compel defendants to use predictive coding because: "Defendant has established that it is possible to obtain the requested documents from a less burdensome and less expensive source … [and] Plaintiffs have not been able to show that the likely benefit received from production of the documents in native format using predictive coding outweighs the burden on defendant ….");

- *Fed. Deposit Ins. Corp. v. Bowden*, 2014 WL 2548137 (S.D. Ga. June 6, 2014) (holding that if the parties had any further ESI related disputes they should attempt to create a Joint ESI Protocol or Consent Order and consider the use of predictive coding);

- *In re Bridgepoint Education, Inc.*, 2014 WL 3867495 (S.D. Cal. Aug. 6, 2014) (denying plaintiff's motion to compel production of additional documents as unduly burdensome and taking into account defendant's argument that predictive coding software is not foolproof and additional costs for attorney review would be incurred even if predictive coding software was used and denying plaintiff's request for documents already produced to be run through the predictive coding software after they were already located via keyword searching as approved by the court); and

- *Green v. Am. Modern Home Ins. Co.*, 2014 WL 6668422 (W.D. Ark. Nov. 24, 2014) (entering the parties' Joint Motion for Entry of Agreed Order Establishing Protocol for Production of Electronically Stored Information, which stated that "[i]n lieu of identifying responsive ESI using the search terms and custodians/electronic systems as described in Sections II.C & II.D above, a party may use a technology assisted review platform to identify potentially relevant documents and ESI.").

[204] *Cf.* Fed. R. Civ. P. 26(g)(1)(B) (counsel's certification of discovery requests and responses); 34(b)(2)(A) (parties directed to respond to discovery requests).

[205] Available at https://thesedonaconference.org/publications (last visited July 14, 2015). The Sedona Principles were first published in March 2003. Since that time, several updates to the Commentaries to the Principles have been issued to address developments in the case law, as well as amendments to the Federal Rules of Civil Procedure and several state civil procedure rules. Principle 6 has, however, remained fundamentally unchanged.

[206] *The Sedona Principles: Best Practices, Recommendations & Principles for Addressing Electronic Document Production* at 38 (2d ed. June 2007).

[207] *Id*.

comments also caution "[a] producing party should not be required to undertake more heroic efforts merely because the party seeking discovery is suspicious of the efforts undertaken by the producing party."[208] Sedona Principle 6 also makes clear that responding parties—not their adversaries or the courts—are in the best position to choose the procedures, methodologies and technologies for preserving, searching and producing their own ESI.[209] Sedona Principle 6 has been cited with approval by numerous courts.[210]

Consistent with Sedona Principle 6, well-established case law makes clear that an adversary does not have the right to challenge decisions a party has made about how it will respond or has responded to discovery (including the procedures it will follow or has followed), unless and until the adversary can demonstrate a deficiency in a discovery production.[211]

Despite those well-settled rules, case law and principles, some requesting parties have claimed that when a producing party uses TAR, that party *must* disclose to their adversary the fact it is using TAR in the first instance, the specifics of its methodology (including the confidence level and margin of error), and in some instances share the "seed sets" used to train the TAR tool (including non-responsive documents). Those arguments are often premised on a misapplication of the holdings of the few reported TAR cases. Indeed, the seminal TAR case of *Da Silva Moore v. Publicis Groupe & MSL Group*[212] is often mis-cited for this proposition.

In *Da Silva Moore*, Magistrate Judge Peck judicially endorsed a *party-negotiated* TAR protocol. Many litigants and judges have mis-cited *Da Silva Moore* for the proposition the protocol in that case is one that parties must follow when using TAR, or that courts can or should order the parties to follow when using TAR, including the exchange of training or non-responsive documents.[213] If parties in a case

[208] *Id*.

[209] *See* Hon. James C. Francis, IV, *Judicial Modesty: The Case for Jurist Restraint in the New Electronic Age*, Law Technology News (Feb. 2013) (no Federal Rule "has given judges the authority . . . to dictate to the parties how or where to search for documents.").

[210] *See, e.g., Rio Tinto PLC v. Vale S.A.*, 2015 WL 872294, at * 6 (when specifically discussing the legal landscape with respect to the use of TAR, citing Sedona Principle 6 for the proposition that responding parties are in the best position to choose how to respond to discovery requests); *Kleen Products L.L.C. v. Packaging Corp. of America*, 2012 WL 4498465, at *5 (N.D. Ill. Sept. 28, 2012) (observing that under Sedona Principle 6 "[r]esponding parties are best situated to evaluate the procedures, methodologies, and techniques appropriate for preserving and producing their own electronically stored information."); *Ford Motor Co. v. Edgewood Properties, Inc.*, 257 F.R.D. 418, 427 (D.N.J. 2009) ("The Sedona Principles wisely state that it is, in fact, the producing party who is the best position[ed] to determine the method by which they will collect documents. . . . absent an agreement or timely objection, the choice is clearly within the producing party's sound discretion."); *Cache La Poudre Feeds, L.L.C. v. Land O'Lakes, Inc.*, 244 F.R.D. 614, 628 (D. Colo. 2007) ("in the typical case, '[r] esponding parties are best situated to evaluate the procedures, methodologies, and technologies appropriate for preserving and producing their own electronic data and documents.'").

[211] *See, e.g., In Re Ford Motor Co.*, 345 F.3d 1315, 1317 (11th Cir. 2003) (vacating order allowing discovery of certain databases where there was no factual finding of "some non-compliance with discovery rules by [the responding party]"); *Orillaneda v. French Culinary Inst.*, 2011 U.S. Dist. LEXIS 105793, at *27 (S.D.N.Y. Sep. 19, 2011) (defendant moved for a protective order against plaintiff's discovery requests, which sought production of a list of "backup sets," litigation hold notices, organizational charts for defendant's IT department, a graphic representation of defendant's computer network, disaster recovery and document retention plans, descriptions of locations used to store emails and word processing documents and information concerning computer hardware and applications used by certain individuals. The court granted the protective order explaining: "[B]ased on the present record, I find that plaintiff's requests for discovery of defendant's search procedures and information systems do not seek relevant information. Discovery concerning these areas may be appropriate in certain circumstances, but it is not appropriate in this case unless and until plaintiff makes a specific showing that defendant's production is deficient."); *Steuben Foods, Inc. v. Country Gourmet Foods, L.L.C.*, 2011 U.S. Dist. LEXIS 43145, at **19-20 (W.D.N.Y. Apr. 21, 2011) ("[G]iven that [Defendant] has failed to establish that Plaintiff destroyed any relevant evidence even in the absence of a written litigation hold, [Defendant's] request for sanctions based on spoliation is unwarranted. Nor will the court grant [Defendant's] alternative request to conduct discovery directed to Plaintiff's document preservation actions in this case. Given the lack of colorable factual basis for [Defendant's] spoliation motion, such request amounts to one seeking to initiate a 'fishing expedition' based on mere speculation."); *Hubbard v. Potter*, 2008 WL 43867, at *4 (D.D.C. Jan. 3, 2008) (Facciola, M.J.) (rejecting a request for additional discovery because speculation that other electronic documents existed does not overcome a Rule 26(g) certification); *Scotts Co., L.L.C. v. Liberty Mut. Ins. Co.*, 2007 WL 1723509 (S.D. Ohio June 12, 2007) (mere suspicion that defendant was withholding electronically stored information is an insufficient basis to permit discovery on discovery, including forensic searches of defendant's computer systems, network servers, and databases). *See also, Hanan v. Corso*, 1998 U.S. Dist. LEXIS 11877, at *23 (D.D.C. April 24, 1998) (Facciola, M.J.) ("Plaintiff seeks 'all documents relating to [defendant company's] previous efforts to respond to [plaintiff's] request for production in this case.' Plaintiff therefore wants discovery about discovery.… [P]laintiff cites no authority for the proposition that the Federal Rules of Civil Procedure contemplate that discovery is itself a fit subject for discovery. To the contrary, discovery is only permitted of information which is either relevant or likely to lead to admissible evidence. Fed. R. Civ. P. 26(b)(1). Plaintiff never explains why discovery about discovery meets that standard, no matter how liberally it is construed, nor any legal authority for the proposition that the federal courts deem the discovery process itself a fit subject for additional discovery.").

[212] 287 F.R.D. 182, 185 (S.D.N.Y. 2012) (Peck, M.J.).

[213] In another opinion that Judge Peck issued three years after *Da Silva Moore*, in which he captioned his Order "Predictive Coding a.k.a. Computer Assisted Review a.k.a. Technology Assisted Review (TAR)—*Da Silva Moore* Revisited," Judge Peck specifically stated:

> The Court is approving the parties' TAR protocol but notes that it is the result of the parties' agreement, not Court order. And as in *Da Silva Moore*, the Court's approval "does not mean … that the exact ESI protocol approved here will be appropriate in all [or any] future cases that utilize [TAR]. Nor does this Opinion endorse any vendor …, nor any particular [TAR] tool."

Rio Tinto PLC v. Vale S.A., 2015 WL 872294 at *7 (*citing Da Silva Moore*, 287 F.R.D. at 193).

cooperatively agree to such a protocol and seek the court's approval of it, *Da Silva Moore* encourages judicial endorsement of such a party-agreed protocol. However, if parties cannot agree on a protocol, or whether TAR is an appropriate procedure for one or both of the parties to use to respond to discovery, *Da Silva Moore* does not stand for the proposition that a court can require a party to follow a certain procedure (including the exchange of training or non-responsive documents) and/or to use TAR to respond to discovery, especially if the party is objecting to such a procedure.[214]

Stated another way, while a producing party may choose to disclose the fact that it plans to use a TAR methodology, the workflow and related intricacies of the methodology itself, and/or in limited instances seed sets of documents, there is no legal requirement under the Federal Rules of Civil Procedure or otherwise mandating such disclosure. For example, in I*n re Biomet M2a Magnum Hip Implant Products Liability Litigation*,[215] the defendant used key word searches to narrow the initial universe of 19.5 million documents, and then applied TAR to the remaining, much smaller set of documents. The plaintiffs, who sought a more active role in the defendants' production methodology, objected to that approach and wanted the defendant to re-run TAR against all of the original 19.5 million documents. Plaintiffs also demanded they work jointly with the defendants to train the TAR tool. The court rejected plaintiffs' request, holding:

> The issue before me today isn't whether predictive coding[/TAR] is a better way of doing things than keyword searching prior to predictive coding[/TAR]. I must decide whether [defendant's] procedure satisfies its discovery obligations and, if so, whether it must also do what [plaintiffs] seek. What [defendant] has done complies fully with the requirements of Federal Rules of Civil Procedure 26(b) and 34(b)(2). I don't see anything inconsistent with the Seventh Circuit Principles Relating to the Discovery of Electronically Stored Information. Principle 1.02 requires cooperation, but I don't read it as requiring counsel from both sides to sit in adjoining seats while rummaging through millions of files that haven't been reviewed for confidentiality or privilege.[216]

In a subsequent opinion in the same case,[217] plaintiffs sought to compel defendant to identify "seed" documents used to train the predictive coding tool. The court held that Federal Rule 26(b)(1) did not require such disclosure, reasoning:

> The [plaintiffs] want the whole seed set [defendant] used for the algorithm's initial training. That request reaches well beyond the scope of any permissible discovery by seeking irrelevant or privileged documents used to tell the algorithm what not to find. That [plaintiffs] have no right to discover irrelevant or privileged documents seems self-evident.
>
> ….
>
> The only authority the [plaintiff] cites is a report of the Sedona Conference that has had a significant, salutary, and persuasive impact on federal discovery practice in the age of electronically stored information. Sedona Conference Cooperation Proclamation, 10 Sedona Conf. J. 331 (Fall Supp. 2009). [Defendant], the [plaintiff] says, isn't proceeding in the cooperative spirit endorsed by the Sedona Conference and the corresponding Seventh Circuit project. But neither the Sedona Conference nor the Seventh Circuit project expands a federal district court's powers, so they can't provide me with authority to compel discovery of information not made discoverable by the Federal Rules.[218]

ii. Forced Use of TAR

Incredibly, in contravention of well-established rules and case law that allow a producing party the discretion to select the best way to respond to discovery requests, some litigants have attempted to argue that a responding party should be *forced* to use TAR (irrespective of

[214] *Accord* Christopher Boehning & Daniel Toal, *No Disclosure: Why Search Terms Are Worthy of Court's Protection,* Law Technology News (Dec. 9, 2013) ("Forced cooperation, in the form of directing a party to turn over . . . search terms or information related to predictive coding seeding [] is . . . a dangerous dance . . . ."); Sean Grammel, *Protecting Search Terms as Opinion Work Product: Applying the Work Product Doctrine to Electronic Discovery*, 161 U. Pa. L. Rev. 2063 (2013) (scholarly analysis of why search terms deserve protection from compelled disclosure as opinion work product).

[215] 2013 WL 1729682, at *2 (N.D. Ind. Apr. 18, 2013).

[216] *Id* at *4.

[217] 2013 WL 6405156 (N.D. Ind. Aug. 21, 2013).

[218] In a recent case, Judge Peck (who wrote *Da Silva Moore*), noted that at least one federal judge had ordered in a decision from the bench that the defendant had to provide "full access to the seed set's responsive and non-responsive documents (except privileged)," and that the "debate in the discovery literature is robust" on whether seed set documents need to be disclosed to an adversary. *Rio Tinto PLC v. Vale S.A*., 2015 WL 872294 at *7 (*citing Fed. Hous. Fin. Agency v. HSBC N.A. Holdings, Inc*., 11 Civ. 6189 (S.D.N.Y. July 24, 2012) and John M. Facciola & Philip J. Favro, Safeguarding the Seed Set: Why Seed Set Documents May Be Entitled to Work Product Protection, 8 Fed. Cts. L. Rev. 1 (2015)).

the costs of using such a methodology). For example, in *Gordon v. Kaleida Health,*[219] the plaintiffs sought an order compelling "defendants to meet and confer with respect to establishing an agreed protocol for implementing the use of predictive coding software." Plaintiffs (mis)cited *Da Silva Moore* for the proposition that "it is necessary that the parties negotiate a protocol to guide the use of predictive coding software for the case." The plaintiffs also claimed that "cooperation" required "a negotiated ESI protocol." The defendants responded that the general rule is that ESI production is within the "sound discretion" of the producing party. The defendants also pointed out that in *Da Silva Moore*, the court did not *direct* defendants to provide plaintiffs with the seed-set documents, rather, defendants *volunteered* to provide such information. While the court ultimately determined it was unnecessary to address the merits of the plaintiffs' motion because the parties had agreed to meet-and-confer on discovery issues, [220] the fact that the plaintiffs would claim that they could somehow dictate the defendant's discovery process in-and-of-itself is troublesome.

In another high-profile TAR case, *Kleen Prods. L.L.C. v. Packaging Corp. of Am.*,[221] after holding multiple evidentiary hearings at which predictive coding experts testified for both parties about the sufficiency of defendants' search to determine whether the defendants could be compelled to use TAR over their objection, the court held that, absent agreement or evidence of a deficient search, consistent with Sedona Principle 6, the responding party is in the best position to evaluate and select the appropriate tools and techniques for document collection and production, and instructed the parties to collaborate and work toward a mutually-agreeable search protocol.[222] As a postscript, the plaintiffs eventually withdrew their demand for defendants to use TAR and the parties agreed upon search terms in that case.[223]

Similarly, in *Fosamax Alendronate Sodium Drug Cases,*[224] the court denied the plaintiffs' motion to compel the defendants to use TAR because: "Defendant has established that it is possible to obtain the requested documents from a less burdensome and less expensive source … [and] Plaintiffs have not been able to show that the likely benefit received from production of the documents in native format using predictive coding outweighs the burden on defendant … ."

Cases like *Gordon, Hiterbeger, Kleen Products* and *Fosamax* are also distinguishable from those where a responding party *chooses* to use a TAR methodology to respond to discovery, and seeks court approval to do so.[225]

Such arguments that a responding party could be *forced* to use TAR over its objection are also squarely at odds with the Federal Rules framework that discovery is self-executing and the mandates of Sedona Principle 6 that allow a responding party to choose how to respond to discovery served by its adversary. Indeed, one influential jurist in the eDiscovery space observed that judges should be cautious about ordering (versus allowing parties to choose) the use of TAR in any particular case, instructing:

> The judiciary's lack of technical expertise is even more pronounced when it comes to technology-assisted review tools that do not depend on familiar keyword strategies. Few vendors are likely to be willing to share the details of their technology with a court for fear of divulging proprietary information to competitors, but even if they did, there is little chance that the judge could comprehend the algorithms. While it is no doubt entirely appropriate

[219] 2013 WL 2250579 (W.D.N.Y. May 21, 2013).

[220] *See also Hiterberger v. Catholic Health Sys., Inc.*, 2013 U.S. Dist. LEXIS 73141 (W.D.N.Y. May 21, 2013) (same facts and holding as *Gordon*).

[221] 2012 U.S. Dist. LEXIS 139632 (N.D. Ill. Sept. 28, 2012).

[222] *Id.*

[223] Matthew Nelson, "*Kleen Products* Predictive Coding Update—Judge Nolan: "I am a believer of principle 6 of Sedona," Symantec eDiscovery Blog, June 5, 2012 (reporting that at a March 28, 2012 hearing, Judge Nolan stated: "[T]he defendant under Sedona [Principle] 6 has the right to pick the [eDiscovery] method") (available at http://www.symantec.com/connect/blogs/kleen-products-predictive-coding-update-judge-nolan-i-am-believer-principle-6-sedona). (last visiteded July 14, 2015).

[224] No. JCCP 4644 (Orange Co. Cal. Sup. Ct.) (Apr. 18, 2013).

[225] *See e.g., Rio Tinto PLC v. Vale S.A.*, 2015 WL 872294 at * 8 ("In the three years since *Da Silva Moore*, the case law has developed to the point that it is now black letter law that where the producing party wants to utilize TAR for document review, the courts will permit it."); *Bridgestone Americas, Inc. v. Int. Bus. Machs. Corp.*, 2014 WL 4923014 (M.D. Tenn. July 22, 2014) (court allowed Plaintiff to switch from an agreed-upon keyword protocol to a TAR protocol over the defendants' objection, or as the court phrased it, "allow[ed] Plaintiff to switch horses in midstream."); *Dynamo Holdings Ltd. Partnership,* 2014 U.S. Tax. Ct. LEXIS 40 (Sept. 17, 2014) (denying petitioner's motion to preclude production of ESI contained on two backup tapes but granting petitioner's alternative motion to use TAR to identify non-privileged information responsive to respondent's discovery requests, including because petitioner presented expert testimony that its TAR methodology could reduce the number of documents subject to attorney review from 3.5—7 million to 200—400,000, thus saving between $420,000 and $465,000 in attorney review costs). *But see, Progressive Cas. Ins. Co. v. Delaney,* 2014 WL 3563467 (D. Nev. July 18, 2014) (After the parties initially agreed to a keyword searching protocol that reduced the ESI at issue from 1.8 million documents to 565,000 documents, the court would not allow producing party to use TAR to review the resultant 565,000 documents, including because the producing party was unwilling to engage in the type of "cooperation and transparency" that the court believed was necessary to switch a TAR procedure).

> to permit a party to choose predictive coding or any other form of technology-assisted review to collect and review data, whether the tool selected will ultimately produce reliable results is a determination that a judge is ill-equipped to make in advance.[226]

iii. Additional Issues to Consider

TAR may also require the use of experts to support its use, including expert testimony to support the specific TAR methodology, specifics about the particular TAR tool in use, and the statistics that were applied during the TAR process.[227]

> Additional questions also remain open about whether a TAR tool that applies proprietary and "black-box"-type algorithms can be subject to a *Daubert* challenge[228] and whether TAR is appropriate for making privilege designations.[229]
>
> * * * * *
>
> Simply stated, the legal standards on whether, how and when TAR can be used defensibly in litigation are still being developed. Based upon this legal landscape, litigants must consider the case law in the applicable jurisdiction about how and when a party can use TAR, how statistics will be applied during the process, what level of "transparency" must accompany the process (including whether "seed sets" must be shared with an adversary), and whether expert testimony will be required to defend the process.

Having said that, in the next few years, additional advances and standardization of TAR technologies, coupled with need to address an increasing volumes of information in the age of Big Data, will lead to a greater acceptance of TAR tools and methodologies by litigants and the courts.

The Rules of Ethics and Civil Procedure are Adapting to Keep Pace with Big Data

c. The Ethical Landscape

In August 2012, The American Bar Association ("ABA") amended Rule 1.1 of the ABA Model Rules of Professional Conduct ("ABA Model Rules") to require lawyers to keep pace with "relevant technology" to comply with their obligation to competently represent clients.

[226] Hon. James C. Francis IV, *Judicial Modesty: The Case for Jurist Restraint in the New Electronic Age*, L. Tech. News, Feb. 2013. *See also, Rio Tinto PLC v. Vale S.A.*, 2015 WL 872294 at *8 ("In the three years since *Da Silva Moore*, the case law has developed to the point that … where the requesting party has sought to force the producing party to use TAR, the courts have refused.") and *Dynamo Holdings Ltd. P'ship v. Comm'r of Internal Revenue*, 2014 U.S. Tax. Ct. LEXIS 40 at * 7 (Sept. 17, 2014), where the court stated:

> And although it is a proper role of the Court to supervise the discovery process and intervene when it is abused by the parties, the Court is not normally in the business of dictating to parties the process that they should use when responding to discovery. If our focus were on paper discovery we would not (for example) be dictating to a party the manner in which it should review documents for responsiveness or privilege, such as whether that review should be done by a paralegal, a junior attorney, or a senior attorney.

[227] *See, e.g., Kleen Prods. L.L.C. v. Packaging Corp. of Am.*, 2012 U.S. Dist. LEXIS 139632 (N.D. Ill. Sept. 28, 2012), discussed *infra*. *See also Dynamo Holdings Ltd. P'ship v. Comm'r of Internal Revenue*, 2014 U.S. Tax. Ct. LEXIS 40 (Sept. 17, 2014) (analyzing testimony that was presented at an evidentiary hearing by competing predictive coding experts about predictive coding protocols as well as a comparison of volumes of data and associated review costs using traditional linear review vs. TAR approach); *Aurora Coop. Elevator Co. v. Aventine Renewable Energy—Aurora W. LLC.*, No. 12-CIV-00230, Dkt. No. 147 (D. Neb. Mar. 10, 2014) (granting plaintiff's motion to compel and holding that "[f]or at least this first discovery stage, the parties shall consult with a computer forensic expert to create search protocols, including predictive coding as needed, for a computerized review of the parties' electronic records.")

[228] *See* Hon. James C. Francis IV, *Judicial Modesty: The Case for Jurist Restraint in the New Electronic Age*, L. Tech. News, Feb. 2013:

> We do know, however, that the collateral proceedings required to obtain a judicial determination on a technical matter can be substantial. In one recent case, a judge devoted two full days of hearings to a dispute over search methodology, at the end of which she encouraged the parties to reach agreement, which they did (after numerous additional conferences with the court). *Kleen Products LLC v. Packaging Corp. of America*, No. 10 C 5711, 2012 WL 4498465, at *5 (N.D. Ill. Sept. 28, 2012).
>
> That the parties were required to devote substantial resources to this dispute is not surprising. The judge had to be educated about the technologies at issue, and courts rightly demand expert testimony in such cases rather than relying upon the representations of counsel. *Compare Da Silva Moore*, 2012 WL 607412, at *2 (technology-assisted review process requires validation but is not subject to *Daubert* standard) *with [United States v.] O'Keefe*, 537 F. Supp. 2d [14, 24 (D.D.C. 2008)] (evidence challenging search terms required to meet standard of Fed. R. Evid. 702).

[229] For example, while TAR may be able to identify that an email was sent from an in-house lawyer to a client, it is questionable whether that technology can discern whether the lawyer was acting in a business role or in a legal capacity when sending that communication, which could impact the privileged nature of the document. *See also* The Sedona Conference Commentary on Protection of Privileged ESI (Public Comment Version Nov. 2014) at 32 (available at https://thesedonaconference.org/publications). (last visited July 14, 2015) (analyzing use of TAR in privilege identification and noting that it is too early to say how useful TAR will be in privilege review).

This amendment was proposed by the ABA's Commission on Ethics 20/20, which was created by then ABA President Carolyn B. Lamm to perform a thorough review of the ABA Model Rules of Professional Conduct and the U.S. system of lawyer regulation in the context of advances in technology and global legal practice developments.[230]

In particular, Model Rule 1.1 on "Competence," provides:

> A lawyer shall provide competent representation to a client. Competent representation requires the legal knowledge, skill, thoroughness and preparation reasonably necessary for the representation.

The August 2012 amendment concerned Comment 8 to Rule 1.1, which states (amended language underlined):

> i. Maintaining Competence
>
> [8] To maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology, engage in continuing study and education and comply with all continuing legal education requirements to which the lawyer is subject.

In its report to support this amendment,[231] the Commission explained:

> Advances in technology have enabled lawyers in all practice settings to provide more efficient and effective legal services. Some forms of technology, however, present certain risks ….
>
> [T]he Commission concluded that competent lawyers must have some awareness of basic features of technology.
>
> … The Commission concluded that, in order to keep abreast of changes in law practice in a digital age, lawyers necessarily need to understand basic features of relevant technology and that this aspect of competence should be expressed in the Comment. For example, a lawyer would have difficulty providing competent legal services in today's environment without knowing how to use email or create an electronic document.
>
> Comment [8] already encompasses an obligation to remain aware of changes in technology that affect law practice, but the Commission concluded that making this explicit, by addition of the phrase "including the benefits and risks associated with relevant technology," would offer greater clarity in this area and emphasize the importance of technology to modern law practice. The proposed amendment, which appears in a Comment, does not impose any new obligations on lawyers. Rather, the amendment is intended to serve as a reminder to lawyers that they should remain aware of technology, including the benefits and risks associated with it, as part of a lawyer's general ethical duty to remain competent.[232]

In its recent Opinion, the State Bar of California Standing Committee on Professional Responsibility and Conduct ("CA Ethics Opinion") goes further, stating:

> Competency may require even a highly experienced attorney seek assistance in some litigation matters involving ESI. An attorney lacking the required competence for eDiscovery issues has three options: (1) acquire sufficient learning and skill before performance is required; (3) associate with or consult technical consultants or competent counsel; or (3) decline the client representation.[233]

The CA Ethics Opinion squarely addressed "What are an attorney's ethical duties in the handling of discovery of electronically stored information," in the context of a fact pattern whereby confidential and privileged information was turned over to a client's chief competitor, based upon a deficient search term and clawback protocol, as well as the deletion of potentially relevant information based upon the attorney's failure to stop the client's normal document retention policy. In reaching its decision, the state bar explained:

> The ethical duty of competence requires an attorney to assess at the outset of each case what electronic discovery issues might arise during the litigation, including the likelihood that eDiscovery will or should be sought by either side. If eDiscovery will probably be sought, the duty of competence requires an attorney to assess his or her own eDiscovery skills and resources as part of the attorney's duty to provide the client with competent

[230] *See* http://www.americanbar.org/groups/professional_responsibility/aba_commission_on_ethics_20_20/about_us.html (last visited July 14, 2015).

[231] Available at http://www.americanbar.org/content/dam/aba/administrative/ethics_2020/2012_hod_annual_meeting_105a_filed_may_2012.pdf.

[232] *Id.*

[233] *The State Bar of California Standing Committee on Professional Responsibility and Conduct*, Formal Opinion Interim No. 11-0004 (Jan. 7, 2015).

representation. If an attorney lacks such skills and/or resources, the attorney must take try to acquire sufficient learning and skill, or associate or consult with someone with appropriate expertise to assist. Rule 3-110(C). Attorneys handling eDiscovery should have the requisite level of familiarity and skill to, among other things, be able to perform (either by themselves or in association with competent co-counsel or expert consultants) the following:

- initially assess eDiscovery needs and issues, if any;
- implement/cause to implement appropriate ESI preservation procedures;
- analyze and understand a client's ESI systems and storage;
- identify custodians of relevant ESI;
- perform data searches;
- collect responsive ESI in a manner that preserves the integrity of that ESI;
- advise the client on available options for collection and preservation of ESI;
- engage in competent and meaningful meet and confer with opposing counsel concerning an eDiscovery plan; and
- produce responsive ESI in a recognized and appropriate manner.[234]

Lawyers and other professionals whose practice concentrates in this area can provide focused guidance and expertise on all aspects of eDiscovery in litigation, including developing strategies for efficient and effective data harvesting, review and production of large volumes of data, and specialized ontologies and protocols for protecting against the inadvertent disclosure of confidential and privileged information when dealing with large sets of data.

As Big Data continues to proliferate, those skills will become more integral to the litigation process both from an ethical and a practical standpoint.

ii. Amended Rules of Civil Procedure

In December 2006, the Federal Rules of Civil Procedure were amended to specifically address eDiscovery and the explosive growth in information and emergent technologies. To put things in perspective, the Apple iPhone had not yet been introduced at that time.[235] In contrast, about 40 million Apple iPhones were sold in a single quarter after the iPhone 6 was released on September 19, 2014.[236] That equates to the sale of over 34,000 iPhone 6s every hour, 24 hours a day, during the quarter.[237] Not surprisingly, the Federal Rules need to adapt to keep up with the pace of technology.

In September 2013, the Judicial Conference of the United States approved changes to the Federal Rules of Civil Procedure affecting discovery, which will go to the U.S. Supreme Court for adoption and could take effect on December 1, 2015. The proposed amendments to the Rules represent a multi-year effort by the Advisory Committee on Federal Rules of Civil Procedure of the Judicial Conference of the United States that started in 2010. After holding several conferences to develop rules proposals, a package of proposed amendments was released for public comment in August 2013. In response, the Advisory Committee received over 2,300 written comments.[238] The Advisory Committee also held three public hearings in Washington, D.C., Phoenix, Arizona[239] and Dallas, Texas, during which over 120 witnesses testified about the proposed amendments. After the close of the public comment period, the proposed amendments were further revised. In April 2014, after additional changes were made to the proposed amendments, the Advisory Committee ultimately adopted and approved its

[234] *Id.*

[235] Wikipedia—iPhone (available at http://en.wikipedia.org/wiki/Iphone). (last visited July 14, 2015).

[236] http://techcrunch.com/2014/10/20/apple-hardware-sales-q4-2014/ (last visited July 14, 2015).

[237] James B. Stewart, *How, and Why, Apple Overtook Microsoft*, The New York Times, Jan. 29, 2015, available at http://mobile.nytimes.com/2015/01/30/business/how-and-why-apple-overtook-microsoft.html?_r=0 (last visited July 14, 2015).

[238] On February 3, 2014, Littler submitted written comments to the Advisory Committee in support of the proposed amendments (available at: http://www.uscourts.gov/RulesAndPolicies/rules/proposed-amendments.aspx) (last visited July 14, 2015).

[239] On January 9, 2014, a representative of Littler testified before the Advisory Committee in support of the proposed amendments. The transcript of that hearing is available at: http://www.uscourts.gov/RulesAndPolicies/rules/proposed-amendments.aspx.

final version of proposed amendments. In May 2014, the Standing Committee on Rules of Practice and Procedure, and in September 2014, the Judicial Conference of the United States, respectively approved the proposed amendments.[240] They have now been forwarded to the U.S. Supreme Court and Congress, who will review them in 2015, and if the Court or Congress do not take action, the amendments will become effective on December 1, 2015.

Even though the words "Big Data" are not explicitly used, the drafters of the proposed amendments recognized throughout the process Big Data considerations are a significant driver of the new rules, stating:

- "The explosion of ESI in recent years has affected all aspects of civil litigation."[241]
- "[T]he explosion of ESI in recent years has presented new and unprecedented challenges in civil litigation. This is the primary fact motivating an amendment to Rule 37(e). … [T]he remarkable growth of ESI will continue and even accelerate. One industry expert reported to the Committee that there will be some 26 billion devices on the Internet in six years—more than three for every person on earth. Significant amounts of ESI will be created and stored not only by sophisticated entities with large IT departments, but also by unsophisticated persons whose lives are recorded on their phones, tablets, cars, social media pages, and tools not even presently foreseen. Most of this information will be stored somewhere on remote servers, often referred to as the 'cloud," complicating the preservation task. Thus, the litigation challenges created by ESI and its loss will increase, not decrease, and will affect unsophisticated as well as sophisticated litigants."[242]
- "The 1993 Committee Note further observed that '[t]he information explosion of recent decades has greatly increased both the potential cost of wide-ranging discovery and the potential for discovery to be used as an instrument for delay or oppression.' What seemed like an explosion in 1993 has been exacerbated by the advent of eDiscovery."[243]
- "The [rule adopted in December 2006] has not adequately addressed the serious problems resulting from the continued exponential growth in the volume of [electronically stored] information….New Rule 37(e) replaces the 2006 rule."[244]

While the proposed amendments cover a wide array of issues,[245] the two biggest changes of the proposed rules in terms of eDiscovery are:

- First: the "proportionality" factors currently contained in Rule 26(b)(2)(C)(iii) have been moved up into the definition of the scope of discovery in Rule 26(b)(1), to clarify proportionality is a fundamental consideration in all aspects of modern litigation. Moving the proportionality factors to Rule 26(b)(1) where the scope of discovery is defined should help achieve Rule 1's objective of the "just, speedy, and inexpensive determination of every action."
- Second: Proposed Rule 37(e) provides a new paradigm for awarding sanctions—or "curative measures"—for the loss of relevant ESI. Not only does it provide a uniform, national framework for adjudicating those issues,[246] but it also raises the bar and establishes a very high culpability standard (*i.e.*, a party must "act[] with the intent to deprive another party of the information's use in the litigation") for when severe measures—like adverse inference instructions or default judgments—can be ordered based upon the loss of ESI. The new rule also sets forth certain threshold factors that must be met before less severe "curative measures" can be ordered for the loss of

[240] A copy of the proposed amendments and the Committee Notes can be found here: http://www.uscourts.gov/RulesAndPolicies/rules/pending-rules.aspx. (last visited July 14, 2015).

[241] *See Memorandum to Judge Jeffrey Sutton, Chair, Standing Committee on Rules of Practice and Procedure, from Judge David G. Campbell, Chair, Advisory Committee on Federal Rules of Civil Procedure, re: Proposed Amendments to the Federal Rules of Civil Procedure*, dated June 12, 2014, available at www.uscourts.gov/file/14140/download?token=McTrl8L0. (last visited July 14, 2015) (hereinafter the "*Advisory Committee Report*"), p. 14.

[242] *Id.*, p. 15.

[243] Committee Note to proposed Amended Rule 26(b)(1).

[244] Committee Note to proposed Amended Rule 37(e).

[245] For a more thorough discussion of the proposed amendments, *see* Paul Weiner, "Amended Rules of Civil Procedure Address eDiscovery Preservation and Sanctions, Among Other Areas," Bloomberg BNA Corporate Accountability Report, Jan. 16, 2015.

[246] During the public comment period, significant comments and testimony were also provided to the Advisory Committee about:

- the split among circuits regarding when it is appropriate to award serious sanctions like adverse inference instructions, with some circuits imposing them for the negligent loss of ESI while others required a showing of bad faith; and
- Large companies with national footprints spending hundreds of millions of dollars to over-preserve ESI out of fear their actions might in hindsight be viewed as negligent and result in serious—indeed, case ending—sanctions if they were sued in a circuit that permits adverse inference instructions on the basis of negligence.

> ESI. The new rule is limited to ESI, including because, according to the drafters, "there are some clear practical distinctions between ESI and other kinds of evidence."[247]

While the proposed rules go a long way towards bringing federal court standards in line with the realities of litigating cases in today's digital age, it is anticipated that new rules or additional amendments will be required to keep pace with the massive volumes of data being created daily and the exponential developments in technology because of the Big Data phenomenon.

A Whole New World of Experts

Big Data changes the testimony necessary to admit expert evidence at trial and establish the business necessity of using a particular Big Data algorithm. To most employers, Big Data algorithms are a "black box," the inner workings of which are known largely by the developer. Although the employer may be in the dark about the algorithm, it nevertheless may be liable if the algorithm screens out members of a protected group and the employer is unable to explain how or why its results should be trusted. In order to prevail under a disparate impact theory of discrimination, a plaintiff must show the algorithm adversely impacts a protected group or, if the employer succeeds in establishing the business reasons for using the algorithm, by demonstrating there exists a less discriminatory alternative that is equally efficient at serving the employer's legitimate business needs.

A Big Data algorithm is created by gathering data regarding the various dimensions of employee performance, and correlating them with any and all information regarding those employees.[248] Because Big Data algorithms are entirely empirical—a case of measurement without theory—all information regarding each employee potentially is relevant until shown to add nothing to the algorithm's ability to distinguish between good and bad performers. After searching among all possible combinations of data, and allowing the search process to consider each item of information to maximum effect, the algorithm identifies the optimal combination. By design, the algorithm selected must bear a significant statistical relationship to the measure(s) of employee performance, otherwise it would have been discarded in favor of a different algorithm that is a better predictor.

Given Big Data's dependence on information gathered from sources external to the employer, and indirectly from applicants and employees, a variety of novel issues arise. Particularly with respect to applicants, potential plaintiffs are confronted with the problem of identifying the applicants who have been evaluated unfavorably by the Big Data algorithm. The difficulty this presents is that the gender, race, and ethnicity of these rejected applicants may not be documented by employers, nor is it likely to be known how any algorithm impacts more generally. For example, suppose an algorithm returns the result that applicants who drive standard-shift automobiles are the best and brightest employees—how is one to learn whether that criterion impacts a protected group adversely?

Litigants have tried various methods to identify the race of applicants who do not self-identify, sometimes with disastrous results. Perhaps the most notorious was the EEOC's efforts to use "race panels" to determine the race of applicants from driver's license photographs. In *EEOC v. Kaplan Higher Education Corporation*,[249] the Sixth Circuit affirmed the district court's decision striking the expert who relied on

this methodology, in rather harsh terms. "The EEOC brought this case on the basis of a homemade methodology, crafted by a witness with

[247] On this issue, the drafters noted:

> ESI is created in volumes previously unheard of and often is duplicated in many places. The potential consequences of its loss in one location often will be less severe than the consequences of the loss of tangible evidence. ESI also is deleted or modified on a regular basis, frequently with no conscious action on the part of the person or entity that created it. These practical distinctions, the difficulty of writing a rule that covers all forms of evidence, as well as an appropriate respect for the spoliation law that has developed over centuries to deal with the loss of tangible evidence, all persuaded the Advisory Committee that the new Rule 37(e) should be limited to ESI.

Advisory Committee Report, p. 16.

[248] In some instances, the information has nothing to do with an employee's potential to perform a specific job. For example, Gild's Big Data algorithms found that one solid predictor of strong computer coding was an affinity for a specific Japanese manga site. "'Obviously, it's not a causal relationship' . . . But Gild does have 6 million programmers in its database, [Vivienne Ming, Gild's chief scientist] said, and the correlation, even if inexplicable, is quite clear." 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/ (last visited July 14, 2015). Similarly, a different Big Data algorithm found that job applicants who completed online job applications using browsers that did not come with the computer (*e.g.*, Microsoft's Internet Explorer on a Windows OC) and instead had to be installed (*e.g.*, Firefox or Google Chrome) performed better at their jobs and have a lower level of turnover. *Robert recruiters: How software helps firms hire workers more efficiently*, The Economist, Apr. 6, 2013, available at http://www.economist.com/news/business/21575820-how-software-helps-firms-hire-workers-more-efficiently-robot-recruiters (last visited July 14, 2015).

[249] 748 F.3d 749 (6th Cir. 2014).

no particular expertise to craft it, administered by persons with no particular expertise to administer it, tested by no one, and accepted only by the witness himself."[250]

However, other methods of imputing race have received a more favorable reception.[251] "Geocoding" is a methodology that relies on an individual's place of residence, and the racial or ethnic composition of others in that same location, to impute the racial and ethnic identity of any particular individual. For instance, if it is known that an applicant resides in a census block that is 50 percent Hispanic, that applicant is assigned a 50 percent probability of being Hispanic. By summing these probabilities over all applicants, the researcher arrives at an estimate of the composition of the entire group of applicants.[252]

Plaintiffs also have surveyed those identified as applicants in order to characterize the demographics of the applicant pool. Private survey research companies abound, and have been retained to provide expert testimony regarding the make-up of a pool of applicants. When survey methodologies conform to generally accepted principles, such as those followed by the U.S. Bureau of the Census, these surveys have been admitted into evidence.[253]

Accordingly, as Big Data relies increasingly on data sources that are external to the employer, employers who are not obligated to collect demographic data regarding their applicants may find themselves contesting with experts who purport to impute racial, ethnic, and gender characteristics to their applicant flow, using methodologies that may or may not withstand scientific scrutiny.

Just as Big Data elevates trivial differences to statistical significance, so too does it trivialize the concept of "validity," as conventionally defined. In traditional cases in which the challenged screening mechanism takes the form of a test, the standard controversy concerns whether the test is "valid," usually as that is defined by the Uniform Guidelines on Employee Selection Procedures.[254]

Instead, of disputing validity, which Big Data should pass with flying colors because algorithms are selected precisely because they maximize correlations, we anticipate that much of the battle will be fought over an issue that rarely surfaces in litigation over conventional selection methods. Title VII provides that a plaintiff may prevail, despite the employer's proof that its selection criterion is valid, if it can demonstrate that there exists an alternative selection criterion that meets the employer's business needs, but is less discriminatory in its impact.[255] But selecting among alternatives is precisely what Big Data is about.

The procedure for developing a Big Data algorithm is to assess the performance of all possible algorithms, within the limitations of the data available, and select the one that works best. The plaintiff's rebuttal to the employer's defense of validity requires searching among those same algorithms and determining if any impacts the protected group less harshly. This sets up a contest among algorithms, necessitating experts skilled in developing and measuring how they perform. Although these comparisons may be assessed in statistical terms, the experts who will be in demand must be skilled in constructing and assessing algorithms, as well as applying the statistical tests by which this contest may be decided.

PREPARING FOR THIS BRAVE NEW WORLD

Employers should prepare for a new human resources world dominated by data sets, analytics, and statistical correlations. Depending on the employer, that world either has already arrived or is in the process of arriving quickly and with momentum. Big Data is here to stay and will not easily be separated from effective human resources management techniques or from the legal world in which employers operate. Whether addressing the laws that govern the gathering and storage of information about candidates and employees or the tests used to determine whether illegal discrimination has occurred, or examining the ways in which parties manage data in litigation, employers need to know and understand the interplay between Big Data and the human resources laws that dictate what can and cannot be done.

The challenge for employers is to find a way to embrace the strengths of Big Data without losing sight of their own business goals and

[250] *Id*. at 754.

[251] *See, e.g., Israel v. United States*, No. 09-CF-687, Dist. of Columbia Ct. of Appeals (Nov. 26, 2014) (citing geogcoded estimates of potential jurors).

[252] *See, e.g., United States v. Reyes*, 934 F. Supp. 553, 560 (S.D.N.Y. 1996).

[253] *EEOC v. FAPS, Inc.*, Civil No. 10-3095 (JAP)(DEA) (D.N.J. Sept. 26, 2014) (unpublished op. citing cases admitting survey evidence).

[254] *See generally* 29 C.F.R. § 1607.

[255] *See, e.g., Albemarle Paper Co. v. Moody*, 422 U.S. 405, 425 (1975); *Contreras v. City of Los Angeles*, 656 F.2d 1267, 1284-85 (1981).

culture amidst potential legal risks. An important part of this process is finding and working with key business partners to assist in Big Data efforts and developing strategies that have the potential to make the workplace function more effectively for everyone. Employers also need to work with those business partners to establish a clear understanding regarding who is responsible for managing which risks and who bears the responsibility for any legal action that might arise.

Responsible employers will be mindful of the risks attendant upon collecting and storing Big Data from the perspectives of data collectors, privacy laws, and data security custodians, while simultaneously using Big Data to achieve key business goals and create a more cohesive and collegial working environment. To do that well, it is vital that human resources professionals and their lawyers have a seat at the table when business decisions are made regarding how and when to use Big Data. The first step in securing that place in the decision-making process is to better understand what Big Data is and how it relates to the current legal system and human resources framework. Our goal in this paper has been to help employers achieve that improved understanding.

U.S. Office Locations

Albuquerque, NM
505.244.3115

Anchorage, AK
907.561.1214

Atlanta, GA
404.233.0330

Birmingham, AL
205.421.4700

Boston, MA
617.378.6000

Charlotte, NC
704.972.7000

Chicago, IL
312.372.5520

Cleveland, OH
216.696.7600

Columbia, SC
803.231.2500

Columbus, OH
614.463.4201

Dallas, TX
214.880.8100

Denver, CO
303.629.6200

Detroit, MI*
313.446.6400

Fayetteville, AR
479.582.6100

Fresno, CA
559.244.7500

Houston, TX
713.951.9400

Indianapolis, IN
317.287.3600

Irvine, CA
949.705.3000

Kansas City, MO
816.627.4400

Las Vegas, NV
702.862.8800

Lexington, KY
859.317.7970

Long Island, NY
631.247.4700

Los Angeles, CA
Century City
310.553.0308

Los Angeles, CA
Downtown
213.443.4300

Memphis, TN
901.795.6695

Miami, FL
305.400.7500

Milwaukee, WI
414.291.5536

Minneapolis, MN
612.630.1000

Mobile, AL
251.432.2477

Morgantown, WV
304.599.4600

Nashville, TN
615.383.3033

New Haven, CT
203.974.8700

New York, NY
212.583.9600

Newark, NJ
973.848.4700

Orlando, FL
407.393.2900

Overland Park, KS
913.814.3888

Philadelphia, PA
267.402.3000

Phoenix, AZ
602.474.3600

Pittsburgh, PA
412.201.7600

Portland, OR
503.221.0309

Portland, ME
207.774.6001

Providence, RI
401.824.2500

Reno, NV
775.348.4888

Rochester, NY
585.203.3400

Sacramento, CA
916.830.7200

San Diego, CA
619.232.0441

San Francisco, CA
415.433.1940

San Jose, CA
408.998.4150

San Juan, Puerto Rico
787.765.4646

Santa Maria, CA
805.934.5770

Seattle, WA
206.623.3300

St. Louis, MO
314.659.2000

Tysons Corner, VA
703.442.8425

Walnut Creek, CA
925.932.2468

Washington, D.C.
202.842.3400

*In Detroit, Littler Mendelson, PLC and in Lexington, Littler Mendelson, P.S.C., both are wholly-owned subsidiaries of Littler Mendelson, P.C.

Global Office Locations

Toronto, Canada
416.865.0504

Barranquilla, Colombia
57.5.385.6071

Bogotá, Colombia
57.1.317.4628

San José, Costa Rica
506.2545.3600

Santo Domingo, Dominican Republic
809.472.4202

San Salvador, El Salvador
503.2296.9500

Guatemala City, Guatemala
506.2545.3651

San Pedro Sula, Honduras
504.2516.1133

Mexico City, Mexico
52.55.5955.4500

Monterrey, Mexico
52.81.8851.1200

Managua, Nicaragua
506.2545.3651

Panama City, Panama
507.830.6552

Lima, Peru
511.226.1600

Caracas, Venezuela
58.212.610.5450

Valencia, Venezuela
58.241.824.4322

Littler

Employment & Labor Law Solutions Worldwide®

www.littler.com • Littler Mendelson, P.C.