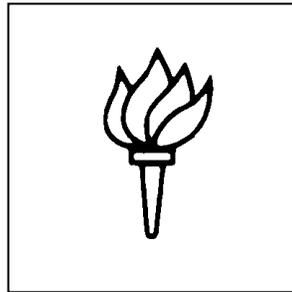


NEW YORK UNIVERSITY

SCHOOL OF LAW

PUBLIC LAW & LEGAL THEORY RESEARCH PAPER SERIES
WORKING PAPER NO. 13-64

LAW & ECONOMICS RESEARCH PAPER SERIES
WORKING PAPER NO. 13-36



**Big Data and Due Process: Toward a Framework to
Redress Predictive Privacy Harms**

Kate Crawford and Jason Schultz

October 2013

Big Data and Due Process: Toward A Framework to Redress Predictive Privacy Harms

By Kate Crawford¹ and Jason Schultz²

Abstract

The rise of “big data” analytics in the private sector poses new challenges for privacy advocates. Unlike previous computational models that exploit personally identifiable information (PII) directly, such as behavioral targeting, big data has exploded the definition of PII to make many more sources of data personally identifiable. By analyzing primarily metadata, such as a set of predictive or aggregated findings without displaying or distributing the originating data, big data approaches often operate outside of current privacy protections (Rubinstein 2013; Tene and Polonetsky 2012), effectively marginalizing regulatory schema. Big data presents substantial privacy concerns – risks of bias or discrimination based on the inappropriate generation of personal data – a risk we call “predictive privacy harm.” Predictive analysis and categorization can pose a genuine threat to individuals, especially when it is performed without their knowledge or consent. While not necessarily a harm that falls within the conventional “invasion of privacy” boundaries, such harms still center on an individual’s relationship with data about her. Big data approaches need not rely on having a person’s PII directly: a combination of techniques from social network analysis, interpreting online behaviors and predictive modeling can create a detailed, intimate picture with a high degree of accuracy. Furthermore, harms can still result when such techniques are done poorly, rendering an inaccurate picture that nonetheless is used to impact on a person’s life and livelihood.

In considering how to respond to evolving big data practices, we began by examining the existing rights that individuals have to see and review records pertaining to them in areas such as health and credit information. But it is clear that these existing systems are inadequate to meet current big data challenges. Fair Information Privacy Practices and other notice-and-choice regimes fail to protect against predictive privacy risks in part because individuals are rarely aware of how their individual data is being used to their detriment, what determinations are being made about them, and because at various points in big data processes, the relationship between predictive privacy harms and originating PII may be complicated by multiple technical processes and the involvement of third parties. Thus, past privacy regulations and rights are ill equipped to face current and future big data challenges.

We propose a new approach to mitigating predictive privacy harms – that of a right to *procedural data due process*. In the Anglo-American legal tradition, procedural due process

¹ Principal Researcher, Microsoft Research; Visiting Professor, MIT Centre for Civic Media; Senior Fellow, NYU Information Law Institute.

² Associate Professor of Clinical Law, NYU School of Law. The authors wish to thank Danielle Citron, Michael Froomkin, Brian Pascal, Brian Covington, and the participants in the 2013 Privacy Law Scholars Conference for their valuable feedback. They also wish to thank Stephen Rushin, Ph.D. for his invaluable research assistance.

prohibits the government from depriving an individual's rights to life, liberty, or property without affording her access to certain basic procedural components of the adjudication process – including the rights to review and contest the evidence at issue, the right to appeal any adverse decision, the right to know the allegations presented and be heard on the issues they raise. Procedural due process also serves as an enforcer of separation of powers, prohibiting those who write laws from also adjudicating them.

While some current privacy regimes offer nominal due process-like mechanisms in relation to closely defined types of data, these rarely include all of the necessary components to guarantee fair outcomes and arguably do not apply to many kinds of big data systems (Terry 2012). A more rigorous framework is needed, particularly given the inherent analytical assumptions and methodological biases built into many big data systems (boyd and Crawford 2012). Building on previous thinking about due process for public administrative computer systems (Steinbock 2005; Citron 2010), we argue that individuals who are privately and often secretly “judged” by big data should have similar rights to those judged by the courts with respect to how their personal data has been used in such adjudications. Using procedural due process principles, we analogize a system of regulation that would provide such rights against private big data actors.

I. Introduction	4
II. Predictive Privacy Harms and the Marginalization of Traditional Privacy Protections.....	5
a. What is Big Data and Why All the Hype?.....	5
b. Big Data’s Predictive Privacy Harms	5
1. Predictive Privacy and Discriminatory Practices.....	7
2. Health Analytics and Personalized Medicine	9
3. Predictive Policing	10
c. Predictive Privacy Harms Threaten to Marginalize Traditional Privacy Protections.....	12
III. Why Big Data Needs Procedural Due Process.....	14
a. Due Process Has Historically Provided a Flexible Approach to Fairness in Adjudication Processes with Large Data Sets and Various Purposes	15
b. Procedural Due Process in the Courts	16
1. Eleven Elements of A Due Process “Hearing”	20
2. The Nature of the Action.....	22
c. The Underlying Values of Due Process.....	22
d. Due Process as Separation of Powers and Systems Management	24
IV. Toward a Model for Data Due Process.....	25
a. Technological Due Process: the Citron Analysis.....	25
b. Procedural Data Due Process.....	28
1. Notice.....	28
2. Opportunity for a Hearing	30
3. Impartial Adjudicator and Judicial Review	30
V. Conclusion.....	31

I. Introduction

“Big data” analytics have been widely hyped in recent years, with many in the business and science worlds focusing on how large datasets can offer new insights into previously intractable problems.³ At the same time, big data is posing new challenges for privacy advocates. Unlike previous computational models that exploit known sources of personally identifiable information (PII) directly, such as behavioral targeting, big data has radically expanded the range of data that can be personally identifying, and uses metadata to manufacture its own PII that does not fall within protected categories. Moreover, by primarily analyzing metadata, such as a set of predictive and aggregated findings or by combining previously discrete data sets, big data approaches often operate outside of current privacy protections.⁴ Existing regulatory schema appear incapable of keeping pace with these advancing business norms and practices.

For example, there are many risks that emerge from the inappropriate inclusion and predictive analysis of an individual’s personal data without their knowledge or express consent. In a well-publicized case from 2012, the retail chain Target was shown to be using data mining techniques to predict which female customers were pregnant, even if they had not announced it publicly, resulting in inappropriate and unauthorized disclosures.⁵ In essence, Target’s predictive analytics “guessed” that a customer was pregnant and disclosed her name to their marketing department, manufacturing PII about her instead of collecting it directly. While there can be little doubt that the customers knew Target was likely collecting data on their individual purchases, it is doubtful that many considered the risk that Target would use data analytics to create models of customers that were so personal and intimate, then send advertising material to homes based on those models. While not necessarily a harm that falls within the conventional “invasion of privacy” boundaries, such harms are still derived from informational collection and use that centers on an individual’s data behaviors. We call these “predictive privacy harms.”

In this article, we confront the tension between the powerful potential benefits of big data and predictive privacy harms. In Part II, we discuss the nature of “big data science” and how personal information can be amassed and analyzed. We then discuss the nature of predictive privacy harms and why traditional privacy protections are insufficient to address the risks posed by big data’s use of personal information. In Part III, we recount the Anglo-American history of procedural due process and the role it has played in both systems of adjudication and separation of powers. We then make the case for why procedural due

³ See, e.g., Strata Conference 2013, <http://strataconf.com/strata2013>; David Brooks, *The Philosophy of Data*, N.Y. TIMES, Feb. 4, 2013, <http://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html>.

⁴ See Ira S. Rubinstein, *Big Data: The End of Privacy or a New Beginning?* N.Y.U. Public Law & Legal Theory Working Papers, Paper No. 357, (2012), http://lsr.nellco.org/nyu_plltwp/357/; Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data: A Time for Big Decisions*, 64 STAN. L. REV. ONLINE 63 (2012), <http://www.stanfordlawreview.org/online/privacy-paradox/big-data>.

⁵ Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. Times, Feb. 19, 2012, <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>.

process principles may be appropriate to draw from to address the risks of predictive privacy harms. In Part IV, we look at the procedural due process literature and suggest ways to analogize a similar framework for private sector big data systems.

II. Predictive Privacy Harms and the Marginalization of Traditional Privacy Protections

a. What is Big Data and Why All the Hype?

*“Knowledge is invariably a matter of degree: you cannot put your finger upon even the simplest datum and say ‘this we know’. In the growth and construction of the world we live in, there is no one stage, and no one aspect, which you can take as the foundation.”*⁶

T.S. Eliot

Big data is a generalized, imprecise term that refers to the use of large data sets in data science and predictive analytics. In practice, it encompasses three types of data magnification and manipulation: first, it refers to technology that maximizes computational power and algorithmic accuracy; second, it describes types of analysis that draw on a range of tools to clean and compare data, and third, it promotes a certain mythology – the belief that large data sets generate results with greater truth, objectivity, and accuracy.⁷ The promise of data at scale has led to profound investment in, consideration of, and hype about the power of big data to solve problems in numerous disciplines, business arenas, and fields.⁸

b. Big Data’s Predictive Privacy Harms

Alongside its great promise, big data presents serious privacy problems. As Omer Tene and Jules Polonetsky write, big data gathers its contents “from online transactions, email, video, images, clickstream, logs, search queries, health records and social networking interactions” as well as “increasingly pervasive sensors deployed in infrastructure such as communications networks, electric grids, global positioning satellites, roads and bridges, a well as in homes,

⁶ T.S. Eliot, from his dissertation on F.H. Bradley.

⁷ See danah boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFORMATION, COMMUNICATION, & SOCIETY 662 (2012). Big data raise numerous critical questions about all three of these shifts. *Id.*

⁸ See Nicholas P. Terry, *Protecting Patient Privacy in the Age of Big Data*, 81 UNIV. OF MISSOURI-KC L. REV. 6 (2012), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2153269 (citing James Manyika et al, Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, May 2011, at 1, (http://www.mckinsey.com/~media/McKinsey/dotcom/Insights_and_pubs/MGI/Research/Technology_and_Innovation/Big_Data/MGI_big_data_full_report.ashx).

clothing, and mobile phones.”⁹ Not only are these data sets sizable, they often contain very intimate aspects of individual lives.¹⁰

Health data is particularly vulnerable, not only because a single breach risks exposing critical information from a multitude of patients’ records,¹¹ but as noted health information law scholar Nicholas Terry observes, data about our online behavior generally – such as buying an e-book about breast cancer survival or liking a disease foundation’s Facebook page – can also reveal information about our health.¹² Even the RFID chips embedded in drug packaging can leave a data “exhaust trail” that links back to information that would normally be considered deeply confidential by health care providers.¹³ When these data sets are cross-referenced with traditional health information, as big data is designed to do, it is possible to generate a detailed picture about a person’s health, including information a person may never have disclosed to a health provider. The combination of data sets and use of predictive analytics can dramatically increase the amount of data that can be considered private.¹⁴ Furthermore, we are seeing users themselves offering up this data directly. For example, in the health context, programs such as the “Blue Button” initiative allow patients to directly download their entire Personal Health Records.¹⁵ Once downloaded, many of these records lose the protections afforded them by federal health privacy statutes such as HIPAA.¹⁶ Other self-help health and fitness approaches, such as Quantified Self and Patients Like Me, are generating data sets that will help identify or predict health attributes.¹⁷

But these privacy problems go beyond just increasing the amount and scope of potentially private information. Big data processes can generate a model of what has a high probability of being PII, essentially *imagining* your data for you. For example, in the Target story about predicting pregnancy, Target had never collected data showing that any particular female customer was pregnant, a fact that most people would almost assuredly consider to be very personal and intimate information. Instead, it predicted it. And as the story showed, the prediction was just as personally sensitive as if it had been collected or shared inappropriately. Target also used the predictive privacy information in a similar manner – exploiting it for marketing purposes. However, because it was not collected from any first or

⁹ Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 240 (2013), *available at* <http://scholarlycommons.law.northwestern.edu/njtip/vol11/iss5/1/>.

¹⁰ See Jay Stanley, *Eight Problems With “Big Data”*, ACLU.ORG, April 25, 2012, <http://www.aclu.org/blog/technology-and-liberty/eight-problems-big-data>.

¹¹ See iHealthBeat, *Report Finds Correlation Between Health Data Breaches, Fraud*, April 30, 2012, <http://www.ihealthbeat.org/articles/2013/4/30/report-finds-correlation-between-health-data-breaches-fraud-cases.aspx>

¹² Terry, *supra* note __ at 12.

¹³ *Id.*

¹⁴ See Paul Schwartz & Dan Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814 (2001) (arguing for a more flexible approach to PII that tracks the “risk of identification” along a spectrum).

¹⁵ See <http://www.healthit.gov/patients-families/blue-button/about-blue-button>.

¹⁶ Terry, *supra* note __, at 10-11.

¹⁷ See Heather Patterson and Helen Nissenbaum, *Context-Dependent Expectations of Privacy in Self-Generated Mobile Health Data* (2013) (on file with author).

third party, there was no need under current privacy regimes to give notice to or gather consent from the customer in the same way that direct collection protocols require. As Terry notes in the context of health information, this is likely to lead to a surge in the use of big data to create “surrogate” health data because of its less regulated nature.¹⁸

Moreover the nature of big data’s dynamic analytical tools is such that the privacy problems of predictive algorithms are often themselves unpredictable, and their effects not even be fully understood by their programmers. As computer scientists have shown, in many contexts, it is *impossible* to guarantee differential privacy when using a learning algorithm that draws data from a continuous distribution.¹⁹ In other words, we cannot know in advance exactly when a learning algorithm will predict PII about an individual; therefore, we cannot predict where and when to assemble privacy protections around that data. When a pregnant teenager is shopping for vitamins, could she predict that any particular visit or purchase would trigger a retailer’s algorithms to flag her as a pregnant customer? At what point would it have been appropriate to give notice and request her consent? How exact does one define this type of privacy problem? Below are further examples of how big data brings predictive privacy harms to the forefront.

1. Predictive Privacy and Discriminatory Practices

For decades, there have been laws prohibiting various discriminatory practices such as advertising real estate sales or rental housing that exclude renters and buyers who fall within racial, gender, or religious categories. One such law, the Fair Housing Act of 1968, prohibits the making, printing, or publication of any “notice, statement, or advertisement . . . with respect to the sale or rental of a dwelling that indicates . . . an intention to make [a] preference, limitation, or discrimination” on the basis of “race, color, religion, sex, familial status, or national origin.”²⁰ These prohibitions work because there are certain moments in the life cycle of housing advertisements that can act as flashpoints for enforcement. Enforcers can monitor published housing advertisements by particular landlords in specific media and check for use of explicit language that calls out the discriminatory intent, such as

¹⁸ Terry, *supra* note ___, at 3. Similar problems exist with computational capacity to “re-identify” personal information or identities that have been stripped away. *See generally* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010); Arvind Narayanan and Vitaly Shmatikov, *Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)*, In Proc. of 29th IEEE Symposium on Security and Privacy, Oakland, CA, May 2008, pp. 111-125. IEEE Computer Society, 2008, available at <http://www.senyt.dk/bilag/netflix2.pdf>.

¹⁹ Chauduri and Hsu, *Sample Complexity Bounds for Differentially Private Learning*, Journal of Machine Learning Research: Workshop and Conference Proceedings 1 (2012), available at <http://cseweb.ucsd.edu/~djhsu/papers/privacy-sample.pdf>.

²⁰ The Fair Housing Act of 1968, 42 U.S.C. § 3604(c); http://portal.hud.gov/hudportal/HUD?src=/program_offices/fair_housing_equal_opp/FHLaws/yourrights. It is worth noting that § 3604(a) prohibits “otherwise mak[ing] unavailable or deny[ing]” housing on the basis of race, color, religion, sex, familiar status, or national origin. Whether using big data to generate profiles that de facto discriminate would violate this provision is unknown and presumably unlikely, given the intent requirement.

“female renters preferred.”²¹ Enforcement cases involve presenting the text of the ad to the adjudicator along with other evidence of intent.

Let us imagine instead landlords and real estate companies shifting away from general advertising in media outlets and toward using big data to determine likely buyers or renters who fit their “ideal” profiles. For those that wish to discriminate, there would be no need to make, print or publish a notice, statement, or advertisement to that effect. Rather, one could design an algorithm to predict the relevant PII of potential buyers and renters and then advertise the properties only to those customers. Such predictions are already occurring. As *Scientific American* observes, companies are drawing on big data models to categorize less credit-attractive users online, and simply not choosing to show them advertisements for loans. If the model indicates a poor credit record, “you won’t even see a credit offer from leading lending institutions, and you won’t realize that loans are available to help you with your current personal or professional priorities.”²² This is despite the fact that it is illegal under federal regulations to discriminate in pricing access to credit based on certain personal attributes.²³

Even very vague signals online, such as liking things on Facebook, can generate a detailed picture. As one University of Cambridge study found, “highly sensitive personal attributes” such as sexual orientation, ethnicity, religious and political views, personality traits, intelligence, use of addictive substances, parental separation, age, and gender were predictable with high degrees of success just from what people liked online.²⁴ Thus, racial discrimination could be applied to prevent some candidates from seeing loans that might be advantageous to them, and housing renters and sellers could potentially use big data to discriminate based on gender, all while circumventing the fair housing laws.²⁵

The connection between big data’s ability to discriminate and maneuver around privacy regulations comes from the particular techniques that big data uses for such determinations.

²¹ See generally *Fair Housing Council of San Fernando Valley v. Roommates.com, LLC*, 521 F.3d 1157 (9th Cir. 2008); *Chicago Lawyer’s Committee for Civil Rights Under Law, Inc. v. Craigslist, Inc.*, 519 F.3d 666 (7th Cir. 2008).

²² Michael Fertik ‘The Rich See a Different Internet to the Poor’, *Scientific American*, February 18, 2013, available at <http://www.scientificamerican.com/article.cfm?id=rich-see-different-internet-than-the-poor>. See also generally Joseph Turow, *THE DAILY YOU* (Yale University Press 2012).

²³ See Equal Credit Opportunity Act, 15 U.S.C. § 1691 et seq.

²⁴ See Michael Kosinski, David Stillwell, and Thore Graepel, *Private traits and attributes are predictable from digital records of human behavior*, PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, available at <http://www.pnas.org/content/early/2013/03/06/1218772110.abstract>.

²⁵ For additional concerns about such practices, see <http://solveforinteresting.com/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it/>; <http://solveforinteresting.com/followup-on-big-data-and-civil-rights/>; *Joint Filing to the Federal Trade Commission of Center for Digital Democracy, U.S. PIRG & World Privacy Forum*, “In the Matter of Real-time Targeting and Auctioning, Data Profiling Optimization, And Economic Loss to Consumers and Privacy”, April 8, 2010, <http://www.centerfordigitaldemocracy.org/sites/default/files/20100407-FTCfiling.pdf>.

Not only can massive amounts of online behavior be collected and assessed to compute the probabilities of a particular demographic characteristic of a given individual, that predictive analysis can then become a form of PII itself. This process can predict highly intimate information, even if none of the *individual pieces* of data could be defined as PII. When functioning well, big data techniques can predict your PII; when not, you can be mislabeled with inaccurate PII. Either way, such processes create a model of possible personal information and associate it with an individual.²⁶ Harms can result both when the model is accurate, and when it is incorrect.

2. Health Analytics and Personalized Medicine

Similar issues arise when big data is used to address health problems. As noted health law scholar Nicholas Terry has written, “Technology, not surprisingly, is viewed as holding the solution [to rising health care costs] because advances have made vast computational power affordable and widely available, while improvements in connectivity have allowed information to be accessible in real time virtually anywhere affording the potential to improve health care by increasing the reach of research knowledge, providing access to clinical records when and where needed, and assisting patients and providers in managing chronic diseases.”²⁷ Some also predict that it will bring forth “personalized medicine” – the use of big data for diagnostic predictions and treatment suggestions.²⁸

Yet in order for such models to function, they need to draw on very detailed personal health information about a patient as well as thousands of patient “profiles” for comparison.²⁹ When algorithms produce predictions, those outputs are again associated with individuals in the same way as PII. As Terry suggests, this threatens the privacy of health information in two ways:

First, our “medical selves” exist outside of the traditional and HIPAA/HITECH-regulated health domain, creating exploitable confusion as health information moves

²⁶ See Cynthia Dwork & Deirdre Mulligan, *It's Not Privacy, and It's Not Fair*, 66 STAN. L. REV. ONLINE 35 (Sept. 3, 2013) at 36-8; Ian Kerr & Jessica Earle, *Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy*, 66 STAN. L. REV. ONLINE 65, 69 (Sept. 3, 2013).

²⁷ Terry, *supra* note __, at 2 (citing Institute of Medicine, *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*, Sept. 6, 2012, <http://www.iom.edu/Reports/2012/Best-Care-at-Lower-Cost-The-Path-to-Continuously-Learning-Health-Care-in-America.aspx> at 4-1 to 4-2).

²⁸ <https://www.gov.uk/government/news/30-million-investment-in-health-research-centre-to-tackle-major-diseases>; <http://venturebeat.com/2013/01/27/the-personalized-medicine-revolution-is-almost-here/>. See also Terry, *supra* note __, at 5 (quoting Meaningful Use Of Health IT Stage 2: The Broader Meaning, Health Affairs, March 15th, 2012, <http://healthaffairs.org/blog/2012/03/15/meaningful-use-of-health-it-stage-2-the-broader-meaning/>) (“It will not be long until patient level information is combined with large existing data sets [that] will generate far more accurate predictive modeling, personalization of care, assessment of quality and value for many more conditions, and help providers better manage population health and risk-based reimbursement approaches.”).

²⁹ See, e.g., <http://www.personalgenomes.org/>; <https://www.23andme.com/>.

in and out of protected spaces. Second, big data positions data aggregators and miners to perform an end-run around health care's domain-specific protections by creating medical profiles of individuals in HIPAA-free space. After all, what is the value of HIPAA/HITECH sector-specific protection designed to keep unauthorized data aggregators out of our medical records if big data mining allows the creation of surrogate profiles of our medical selves?³⁰

Thus, even health information – one of the most highly protected types of personal information – will be increasingly vulnerable in the context of big data and predictive analytics.

3. Predictive Policing

Law enforcement agencies around the US are turning to “predictive policing” models of big data in the hopes that they will shine investigative light on unsolved cases or help prevent future crimes.³¹ According to the FBI, the model uses the date, time, type, and location of recent crimes and combines that data with historical crime data to generate “hot spots” as a focus for officer patrols. Again according to the FBI, “no one submits, collects, or uses personal data.”

Yet, for big data, it takes very little to connect time, place, and location with individuals, especially when combined with other data sets.³² Moreover, the predictions that these policing algorithms make – that particular geographic areas are more likely to have crime – will surely produce more arrests in those areas by directing police to patrol them. This, in turn, will generate more “historical crime data” for those areas and increase the likelihood of patrols. For those who live there, these “hot spots” may well become as much PII as other demographic information. Law enforcement also uses similar analytics in their so-called “fusion centers” and other efforts to predict or flag individuals as suspicious or worthy of investigation, search, or detention.³³ As Citron & Gray note:

In one case, Maryland state police exploited their access to fusion centers to conduct surveillance of human rights groups, peace activists, and death penalty opponents over a nineteen-month period wherein fifty-three political activists eventually were classified as “terrorists,” including two Catholic nuns and a Democratic candidate for local office. The fusion center shared these erroneous terrorist classifications with

³⁰ See Terry, *supra* note __, at 5. Terry also argues that HIPAA/HITECH “is known for its function creep and multiple exceptions. As a result a big data-like project such as running data analytics against a hospital’s [Electronic Medical Records] data looking for disease predictors may not be “research” but exempted from regulation as a quality improvement under “health care operations[,]” *Id.* at 27 n. 116, and that de-identified data run through big data might have an even greater risk of re-identification. *Id.* n. 117-120.

³¹ See <http://www.fbi.gov/stats-services/publications/law-enforcement-bulletin/2013/April/predictive-policing-using-technology-to-reduce-crime>.

³² See Yves-Alexandre de Montjoye, Cesar A. Hidalgo, Michel Verleysen & Vincent D. Blondel, *Unique in the Crowd: The privacy bounds of human mobility*, NATURE, Mar. 25, 2013, <http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html>.

³³ Citron & Gray, *Quantitative Privacy* at 4.

federal drug enforcement, law enforcement databases, and the National Security Administration, all without affording the innocent targets any opportunity to know, much less correct, the record.³⁴

When combined and constructed into a composite prediction of a person, such analytics have very serious consequences for privacy. In her concurrence from the recent Supreme Court decision in *United States v. Jones*, Justice Sotomayor expressed serious concerns about invasions of privacy that could result from *direct* collection of massive amounts of personal data:

Awareness that the Government may be watching chills associational and expressive freedoms. And the Government's unrestrained power to assemble data that reveal private aspects of identity is susceptible to abuse. The net result is that GPS monitoring—by making available at a relatively low cost such a substantial quantum of intimate information about any person whom the Government, in its unfettered discretion, chooses to track—may “alter the relationship between citizen and government in a way that is inimical to democratic society.”³⁵

With predictive policing, the government is not only watching and collecting massive amounts of information about us, it is also using predictive analytics to generate “data that reveal private aspects of identity” and is subject to abuse in similar ways.³⁶ The companies designing software for police departments reveal that their systems aren't necessarily right – practically or ethically. The former privacy officer at Intelius, Jim Adler, designed software that predicts at an individual level whether someone will be a felon – using very little data, but considerable predictive guesswork. This goes a step further than neighborhood-based hot spots, and while the software could make some accurate predictions, it also generates false positives. Adler observed that “geeks like me can do stuff like this, we can make stuff work – it's not our job to figure out if it's right or not. We often don't know.”³⁷ This can have particularly harmful impacts not only on racial profiling and other avenues of discrimination but also on programs designed to promote rehabilitation and reincorporation through so-called “Clean Slate” laws,³⁸ which aim to allow non-violent offenders to “clean” their criminal records in order to gain better opportunities for education, employment, and housing.

³⁴ *Id.* at 15.

³⁵ *United States v. Jones*, 132 S.Ct. 945, 956 (2012) (quoting *United States v. Cuevas-Perez*, 640 F.3d 272, 285 (7th Cir. 2011) (Flaum, J., concurring)).

³⁶ See Patricia L. Bellia, *The Memory Gap in Surveillance Law*, 75 U. OF CHICAGO L. REV. 137 (2008); Granick, *Evolving Principles for Regulation of Government Surveillance in the Age of Big Data* (on file with authors).

³⁷ See Jordan Robertson, *How Big Data Could Help Identify the Next Felon or Blame the Wrong Guy*, Bloomberg News, Aug. 14, 2013, <http://www.bloomberg.com/news/2013-08-14/how-big-data-could-help-identify-the-next-felon-or-blame-the-wrong-guy.html>.

³⁸ Lahny Silva, *Clean Slate: Expanding Expungements and Pardons for Non-Violent Federal Offenders*, 79 UNIV. OF CINN. L. REV. 155 (2011).

Questions about the constitutional limits on public data surveillance, such as the GPS tracking in *Jones* continue to test the courts.³⁹ The generative data-making practices of big data will only push these questions well beyond the bounds of traditional Fourth Amendment precedents.⁴⁰ Big data approaches to policing and intelligence may be qualitatively different to the kinds of surveillances approaches in *Jones*, not merely quantitatively different.

c. Predictive Privacy Harms Threaten to Marginalize Traditional Privacy Protections

In light of these predictive privacy harms, it is worth considering what an appropriate set of privacy protections might be to address them. Traditionally, American civil privacy protections have focused on regulating three main activities: information collection, information processing, and information disclosure.⁴¹ For example, the Electronic Communications Privacy Act prohibits the unauthorized collection of communications content.⁴² The Fair Credit Reporting Act prohibits the use of financial records for certain purposes.⁴³ The Video Privacy Protection Act prohibits the disclosure of video rental records.⁴⁴

As noted above, big data approaches have the potential to elude all three of these approaches primarily because of the unpredictable nature of its predictive privacy harms. From the perspective of data collection regulations, when each datum is collected, one cannot assess the predictive privacy risks from a single point of data such as a single tweet, a single “like”, a single search, or a single afternoon drive. Regulating collection is hard, and predictive analytics, impossible. Nor can one necessarily predict when a certain form of information processing will produce predictive privacy harms. Even disclosure regulations become complicated, as again, the data that ends up being personally identifying may not yet exist during the most significant data transfers. For example, predictive policing systems demonstrate situations where numerous data collections and transfers can occur before any predictive private harm comes into existence. In fact, it may only be after all transfers are complete that the predictions occur. For example, if the FBI were to collect crime data from numerous local law enforcement databases and then predict that your street is likely to house sex offenders, there would have been no way to know when the transfers occurred what the resulting harm would be – after all, that is what big data promises to do that was unavailable before. Thus, unless one decides that privacy regulations must govern all data ever collected, processed, or disclosed, deciding where and when to draw lines around these activities becomes extremely difficult with respect to big data information practices.

³⁹ See, e.g., <http://www.aclu.org/blog/technology-and-liberty/aclu-court-today-arguing-gps-tracking-requires-warrant>.

⁴⁰ David C. Gray & Danielle Keats Citron, *The Right to Quantitative Privacy*, 98 MINN L. REV. ___ (2013).

⁴¹ See generally Dan Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477 (2006).

⁴² See Electronic Communications Privacy Act of 1986, Pub. L. No. 99-508, 100 Stat. 1848 (codified as amended at 18 U.S.C. §§ 2510-2520, 2701-2711, 3121-3127 (2000)).

⁴³ 15 U.S.C. § 1681(b).

⁴⁴ 18 U.S.C. § 2710.

Even the anchoring concept of most modern privacy regulations – PII – may fail to provide a sufficient nexus for future big data privacy regulation as the analytics involved are simply too dynamic and unpredictable to know if and when particular information or analyses will become or generate PII. It may only be in hindsight that we will observe the problem, such as with predictive policing or in the Target pregnancy case.

The broader frameworks for privacy may be marginalized by predictive privacy harms. For decades, privacy policymakers have relied on a set of Fair Information Practice Principles (FIPPs) that provide for adapting existing privacy laws and developing new ones, especially in light of new technological developments or information practices.⁴⁵ Various versions of the FIPPs exist, but in general they comprise: (1) Notice/Awareness; (2) Choice/Consent; (3) Access/Participation; (4) Integrity/Security; and (5) Enforcement/Redress.⁴⁶

Recently, The White House released its own FIPPs via a Consumer Privacy Bill of Rights:

- Individual Control: Consumers have a right to exercise control over what personal data companies collect from them and how they use it.
- Transparency: Consumers have a right to easily understandable and accessible information about privacy and security practices.
- Respect for Context: Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data.
- Security: Consumers have a right to secure and responsible handling of personal data.
- Access and Accuracy: Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data is inaccurate.
- Focused Collection: Consumers have a right to reasonable limits on the personal data that companies collect and retain.
- Accountability: Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights.⁴⁷

While somewhat overlapping, The White House version included new additional approaches to information regulation that attempt to expand the scope of FIPPs, especially in terms of “control” over the information at issue by focusing on the user and the category of “personal data.” Yet, even these broadened principles depend on knowing what information is “personal data” and providing notice and choice/control to users *ex ante* any privacy harm.

⁴⁵ Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers* (2012) at 11, <http://www.ftc.gov/os/2012/03/120326privacyreport.pdf>.

⁴⁶ Federal Trade Commission, *Fair Information Practice Principles*, <http://www.ftc.gov/reports/privacy3/fairinfo.shtm>.

⁴⁷ The White House, *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy*, February 2012, <http://www.whitehouse.gov/sites/default/files/privacy-final.pdf>.

Law is obsessed with causality, while big data is generally a tool of correlation. This makes FIPP-style approaches to privacy protection particularly difficult with respect to big data.

But how does one give notice and get consent for innumerable and perhaps even yet-to-be-determined queries that one might run that result the creation of “personal data”? How does one provide consumers with individual control, context, and accountability over such processes? Such difficulties suggest that frameworks like FIPPs, where the focus is on data collection and retention and using a notice-and-consent models as gatekeepers, will fail to successfully regulate for predictive privacy harms.⁴⁸

As big data is versatile, dynamic and unpredictable, traditional notions of privacy that isolate certain categories of information (such as PII) in order to regulate collection, utilization, or disclosure are ill-suited to address these emerging risks.⁴⁹ Even omnibus privacy approaches, such as a “right to be forgotten”, will likely struggle with big data’s ability to re-remember or even *re-imagine* intimate information about a person based on past or present data. These types of privacy problems demand a shift in thinking that can approach the problem with the same dynamic and flexible capacities that big data itself provides.

III. Why Big Data Needs Procedural Due Process

*The heart of the matter is that democracy implies respect for the elementary rights of men, however suspect or unworthy; a democratic government must therefore practice fairness; and fairness can rarely be obtained by secret, one-sided determination of facts decisive of rights.*⁵⁰

As noted above, the power of big data analyses to evade or marginalize traditional privacy protections and frameworks, its drive to bring visibility to the invisible, and its dynamic and unpredictable nature all present challenges to thinking about how privacy and big data can coexist. In response, we propose an alternative approach – *procedural data due process*. Procedural data due process would, rather than attempt regulation of personal data collection, use, or disclosure *ex ante*, regulate the “fairness” of the analytical processes of big data with regard to how they use personal data (or metadata derived from or associated with personal data) in any “adjudicative” process – a process whereby big data is being used to determine attributes or categories for an individual. For example, if a health insurance provider used big data to determine that likelihood that a customer have a certain disease and thus denied coverage on that basis, the customer would have a data due process right

⁴⁸ See Dwork & Mulligan, *supra* note ___ at 36-8; Woodrow Hartzog & Evan Selinger, *Big Data in Small Hands*, STAN. L. REV. ONLINE 81 (Sept. 3, 2013) at 81-83.

⁴⁹ In fact, one notorious data broker, Acxiom, now let’s customers see and change the data it collects about them individually but not the analytics it uses to assess them for sale to marketers, a sign that transparency and regulation of individual data collection is not likely to serve as an effective gate keeping function for controlling privacy harms. *See* <http://www.nytimes.com/2013/09/05/technology/acxiom-lets-consumers-see-data-it-collects.html>.

⁵⁰ *McGrath*, 341 U.S. at 170.

with regard to that determination. Similarly, if a potential employer used big data to predict how “honest” certain job applicants might be.⁵¹

What would such a regulatory process entail? Below we describe some of the history of due process in Anglo-American law and demonstrate why the values embedded within it and the procedures it has traditionally used in courts of law and administrative proceedings may be well-suited for an analogous system regulating private use of big data to mitigate predictive privacy harms. We then discuss what procedural data due process would involve and some possible implementations.

a. Due Process Has Historically Provided a Flexible Approach to Fairness in Adjudication Processes with Large Data Sets and Various Purposes

Throughout the history of Anglo-American law, procedural due process has served as a set of constraints on adjudication – the process of deciding. Adjudications are arguably similar to the type of models and determinations that predictive algorithms create based on massive data sets. Just as information drives big data determinations, so does it drive litigation, legal strategies, and legal outcomes.⁵² Law, much like computer code and data, has its own information rules that are governed by various frameworks, from formal rules like the Code of Civil Procedure to common law and constitutional doctrines such as due process.⁵³

First conceptualized in Magna Carta in 1215, due process was understood to mean that “[n]o free man shall be taken, imprisoned, disseised, outlawed, banished, or in any way destroyed . . . except by the lawful judgment of his peers [or/and] by the law of the land.”⁵⁴ Due process then made its way into the U.S. Constitution as part of the Fifth Amendment, which states “No person shall . . . be deprived of life, liberty, or property, without due process of law[.]”⁵⁵

There are two important components to note here. The first is the prohibition on deprivation. The subjects – life, liberty, or property – are obviously each broad categories that have at times defined the core components of citizenship. They represent, qualitatively, the level of seriousness that the deprivation must constitute in order to invoke due process

⁵¹ See <http://www.nytimes.com/2013/04/21/technology/big-data-trying-to-build-better-workers.html>

⁵² See Fredric M. Bloom, *Information Lost and Found*, 100 CAL. L REV. 636 (2012)

⁵³ *Id.* at 640. See also Lawrence Lessig, CODE: AND OTHER LAWS OF CYBERSPACE (Basic Books, 1999).

⁵⁴ See Nathan S. Chapman & Michael W. McConnell, *Due Process as Separation of Powers*, 121 Yale L.J. 1672, 1682 (2012), available at <http://yalelawjournal.org/images/pdfs/1080.pdf> (quoting MAGNA CARTA ch. 29, reprinted and translated in A.E. DICK HOWARD, MAGNA CARTA: TEXT AND COMMENTARY 43 (1964)). See also *Hurtado v. California*, 110 U.S. 516, 521-25 (1884); *Murray’s Lessee v. Hoboken Land & Improvement Co.*, 59 U.S. (18 How.) 272, 276-78 (1855).

⁵⁵ U.S. Const., Fifth Amend. This prohibition focused on federal state actions. The Fourteenth Amendment also contains a similar clause that extended due process protections to individual state actions.

protection – the type of harm that we wish to prevent. The category of liberty is especially important to consider in the context of privacy and predictive privacy harms. Both John Locke and Lord Blackstone spoke of liberty as a “natural right” “to follow my own Will in all things” that “may only abridged only with the explicit permission of the laws.”⁵⁶ If one considers privacy to be “the right to be let alone”⁵⁷ and to have some freedom⁵⁸ of self-determination and autonomy, then it fits well within the liberty category.⁵⁹ Property and other interests are also implicated, especially as big data analytics are integrated into decisions concerning housing opportunities, employment, and credit provisioning. Thus, predictive privacy harms seem well-suited for due process protection in terms of the type of subject matter covered.

The second component – “without due process of law” – is a means to enforce the probation: a process. But what constitutes this process? What are the underlying values that drive it? How would they fare as a means of regulating big data?

b. Procedural Due Process in the Courts

Today, procedural due process generally describes the constitutional requirement that any government deprivation of a liberty or property right must be preceded, at a minimum, by notice and the opportunity for a hearing on the matter before an impartial adjudicator.⁶⁰

Historically, this conception of due process comes mainly from two landmark Supreme Court cases: *Mathews v. Eldridge*⁶¹ and *Goldberg v. Kelly*.⁶² In *Goldberg*, the Court determined that procedural due process required an evidentiary hearing before the government could

⁵⁶ Chapman & McConnell, *supra* note ___, at 1735 n. 282- 284.

⁵⁷ Warren and Brandeis, *The Right to Privacy*, 4 HARVARD L. REV. 193 (1890).

⁵⁸ See Neil M. Richards, *Intellectual Privacy*, 87 TEX. L. REV. 387 (2008)(arguing that privacy “safeguards the integrity of our intellectual activities by shielding them from the unwanted gaze or interference of others.”).

⁵⁹ See Gray & Citron, *supra* note ___, at 10 (suggesting there is a right to information privacy based on substantive due process), 11-12 (noting scholarship on “the damaging effects of surveillance on projects of self-development” and arguing that “[i]n the face of continuous data collection about our daily activities, individuals cannot make meaningful choices about their activities, preferences, and relations and act on them without fear of embarrassment or recrimination.”), 16 (arguing that “broad programs of indiscriminate surveillance threaten fundamental liberty interests and democratic values.”) (citations omitted), 27-28 (“quoting Anthony Amsterdam: “[t]he insidious, far-reaching and indiscriminate nature of electronic surveillance-and, most important, its capacity to choke off free human discourse that is the hallmark of an open society—makes it almost, although not quite, as destructive of liberty as ‘the kicked-in-door’ ”).

⁶⁰ See *Hamdi v. Rumsfeld*, 542 U.S. 507, 533 (2004) (quoting *Cleveland Board of Education v. Loudermill*, 470 U.S. 532, 542 (1985) (“An essential principle of due process is that a deprivation of life, liberty or property must be preceded by notice and opportunity for hearing appropriate to the nature of the case.”) (internal quotation marks omitted)). See also *Mullane v. Central Hanover Bank & Trust Co.*, 339 U.S. 306, 319 (1950).

⁶¹ 424 U.S. 319 (1976).

⁶² 397 U.S. 254 (1970).

deprive a person of welfare benefits. There, New York City Department of Social Services allowed city case-workers to terminate payments to welfare recipients that the that they deemed ineligible.⁶³ After the City terminated the recipients' welfare payments, the recipients could request a post-termination hearing challenging the decision.⁶⁴ The Court, nonetheless, found that this procedure inadequately protected the welfare recipients' procedural due process rights under the Fourteenth Amendment. Writing for the majority, Justice Brennan explained that, "for qualified recipients, welfare provides the means to obtain essential food, clothing, housing, and medical care."⁶⁵ This means that the revocation of a welfare benefit is a governmentally sanction "grievous loss."⁶⁶ Before the government causes a person such a grievous loss, it must afford the individual certain procedural protections. The state need not resort to a complete judicial or quasi-judicial trial in order to satisfy due process.⁶⁷ Instead, the state must, at minimum, provide the potentially aggrieved party with the opportunity to be heard at a meaningful time and in a meaningful manner, adequate notice, the opportunity to present witnesses, and the ability to present arguments and evidence.⁶⁸ Thus *Goldberg* set a fairly high procedural bar for any action that could deprive an individual a property or liberty interest.

However the Court retreated somewhat from this position in *Mathews*. *Mathews* dealt with the Social Security Act's policy for the termination of disability benefits. According to this policy, the disabled worker bore the burden of proving her entitlement to benefits by showing that she was unable to perform her previous work, or any other kind of gainful employment because of a disability.⁶⁹ Local state agencies would review the evidence provided and make continuing determinations as to the worker's eligibility for aid.⁷⁰ If the state agency felt that an aid recipient no longer qualified for disability relief, the agency would inform the recipient and the Social Security Administration (SSA) of the decision and provide both with a summary of the evidence used in making the determination.⁷¹ The SSA would then make a final determination; if the SSA terminated disability benefits, the recipient had the opportunity for a thorough review hearing.⁷²

In many ways, this case was similar to *Goldberg*—in both cases, the state deprived an individual of some government benefit without the opportunity for a pre-termination hearing. But the Supreme Court in *Mathews* found that the termination of disability payments does not require the same pre-termination hearing as the termination of welfare payments. "The private interest that will be adversely affected by an erroneous termination of benefits is likely to be less in the case of a disabled worker than in the case of a welfare recipient."⁷³ This is because "[e]ligibility for disability payments is not based on financial need, and

⁶³ *Id.* at 258-59.

⁶⁴ *Id.* at 259-60.

⁶⁵ *Id.* at 264.

⁶⁶ *Id.* at 263 (quoting *Joint Anti-Fascist Refugee Committee v. McGrath*, 341 U.S. 123, 168 (1951)).

⁶⁷ *Id.* at 266.

⁶⁸ *Id.* at 267-68.

⁶⁹ *Mathews*, 424 U.S. at 319.

⁷⁰ *Id.*

⁷¹ *Id.*

⁷² *Id.* at 319-20.

⁷³ *Id.* at 321.

although hardship may be imposed upon the erroneously terminated disability recipient.”⁷⁴ The countervailing state interest in fiscal prudence and efficiency outweighed the potential for an erroneous and harmful deprivation, making additional procedural protections constitutionally unnecessary. To soften the supposedly stringent requirements of *Goldberg*, the Court in *Mathews* established a test for determining what courts must consider when judging the constitutionality of the deprivation that consisted of balancing three factors: (1) the private interest that will be affected by the official action; (2) the risk of an erroneous deprivation of such interest through the procedures used, and the probable value, if any, of additional or substitute procedural safeguards; and (3), the Government’s interest, including the function involved and the fiscal and administrative burdens that the additional or substitute procedural requirements would entail.⁷⁵

While *Mathews* shows that the level of due process required differs according to the gravity of the deprivation and the magnitude of the countervailing state interest, most cases over time have established four distinct, procedural elements required when the state deprives an individual of a state interest: (1) participatory procedures (i.e. the affected party is present), (2) a neutral arbiter, (3) prior process (i.e. the hearing precedes the adverse action), (4) and continuity (i.e. the hearing rights attach at all stages).⁷⁶

In his seminal 1971 article, *Some Kind of Hearing*, Judge Henry Friendly attempts to articulate the proper elements of procedural due process.⁷⁷ Similar to the balancing test in *Mathews*, Friendly emphasized that there is no specific checklist of procedures required, but rather one might think about a set of procedures that are potentially useful and select an appropriate group of them based on the characteristics of the particular matter, such as the severity of the deprivation and the government interest at stake.⁷⁸ He noted that civil procedural due process had moved beyond regulatory areas such as disability and welfare to nonregulatory ones such as the inclusion of an organization on the Attorney General’s subversive list

⁷⁴ *Id.*

⁷⁵ *Id.* at 335.

⁷⁶ See, e.g., *Cleveland Bd. of Educ. v. Loudermill*, (105 S. Ct. 1487, 1493 (1985) (“An essential principle of due process is that a deprivation of life, liberty, or property ‘be preceded by notice and opportunity for hearing appropriate to the nature of the case.’”)

⁷⁷ Henry J. Friendly, *Some Kind of Hearing*, 123 U. PA. L. REV. 1267 (1975).

⁷⁸ *Id.* at 1270 n. 10:

The common law requirement of a fair procedure does not compel formal proceedings with all the embellishments of a court trial . . . nor adherence to a single mode of process. It may be satisfied by any one of a variety of procedures which afford a fair opportunity for an applicant to present his position. As such, this court should not attempt to fix a rigid procedure that must invariably be observed. Instead, the associations themselves should retain the initial and primary responsibility for devising method which provides an applicant adequate notice of the ‘charges’ against him and a reasonable opportunity to respond.

(quoting *Pinsker v. Pacific Coast Soc’y of Orthodontists*, 12 Cal. 3d 541, 555 (1974) (en banc) (Tobriner, J.) (citations omitted)).

during the McCarthy Era without an opportunity to be heard⁷⁹ and a case where the Court held that a teacher at a public institution may not be dismissed without a hearing if he had a tenure or its reasonable facsimile or the dismissal involved a stigma that would impair his ability to obtain future employment.⁸⁰ Recognition of these “stigmatic” liberty interests was a key turning point in the expansion of civil procedural due process in the American courts and has profound implications for data due process, as predictive privacy harms often have the potential for stigmatic results. In fact, one need only look to the errors and mistaken assumptions that have been revealed about the TSA’s “no fly list” – another product of big data – to see the case for analogizing to the McCarthy Era case for due process.⁸¹

On the other hand, Friendly notes, “it should be realized that procedural requirements entail the expenditure of limited resources, that at some point the benefit to individuals from an additional safeguard is substantially out-weighted by the cost of providing such protection, and that the expense of protecting those likely to be found undeserving will probably come out of the pockets of the deserving.”⁸² Thus, much as we do not want to impose onerous costs on courts that would ultimately slow the process of administration of justice and encourage harassment and game-playing, we do not want to impose similar costs on big data providers. This balance of protection with cost will also itself be dynamic and thus, better considered as a standard than a rule – yet another reason why a due process approach is well-suited for big data.⁸³

So what kind of hearing? Friendly writes: “[A] hearing in its very essence demands that he who is entitled to it shall have the right to support his allegations by argument however brief, and, if need be, by proof, however informal.”⁸⁴ Or as Lord Loreburn put it in the English courts in 1911, “a fair opportunity ... for correcting or contradicting anything prejudicial to

⁷⁹ See *id.* at 1273 n. 33.

⁸⁰ See *id.* at 1274 (citing *Perry v. Sindermann*, 408 U.S. 593 (1972); *Board of Regents v. Roth*, 408 U.S. 564 (1972)).

⁸¹ Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1256 (2008) (noting that every week, approximately 1,500 airline travelers reportedly are mislabeled as terrorists due to errors in the data-matching program known as the “No Fly” list). While some due process cases have held that reputational harms are more appropriate for the province of tort law, when reputation leads to deprivation of liberty or property, due process can apply. Compare *Paul v. Davis*, 424 U.S. 693 (1976) (denying due process claim over stigmatic harm related to future employment opportunities stemming from including in a “flyer” of “active shoplifters”) with *Wisconsin v. Constantineau*, 400 U.S. 433 (1971) (holding that a ban on distributing alcoholic drinks to persons whose names were “posted” as excessive drinkers was a deprivation of liberty because it altered or extinguished a distinct right previously recognized by state law).

⁸² See Friendly, *supra* note ___, at 1276.

⁸³ See Citron, *supra* note ___, at 1301 (discussing the debate over standards vs. rules).

⁸⁴ See Friendly, *supra* note ___, 1277 (citing *Londoner v. Denver*, 210 U.S. 373, 386 (1908)).

Friendly goes on: “In his concurring opinion in *Joint Anti-Fascist Refugee Committee v. McGrath*, still the finest exposition of the need for a ‘hearing,’ Mr. Justice Frankfurter said only, even in the case of ‘a person in jeopardy of serious loss,’ that one must be given ‘notice of the case against him and opportunity to meet it.’” *Id.* (citing *McGrath*, 341 U.S. 123, 171-72 (1950)).

their view[.]”⁸⁵ The required degree of procedural safeguards, Friendly writes, varies directly with the importance of the private interest affected and the need for and usefulness of the particular safeguard in the given circumstances and inversely with the burden and any other adverse consequences of affording it. “Even amplified in this way, such a balancing test is uncertain and subjective, but the more elaborate specification of the relevant factors may help to produce more principled and predictable decisions.”⁸⁶ It is this sense of principle and predictability that inspires us to bring due process to big data and its potential privacy harms.

1. Eleven Elements of A Due Process “Hearing”

Having laid out his general vision of due process hearings, Judge Friendly then goes on to enunciate eleven potential elements of a hearing that may help ensure a fair process. Not all are required, he states, but all are worth consideration depending on the circumstances at issue. They are: (1) an unbiased tribunal;⁸⁷ (2) notice of the proposed action⁸⁸ and the grounds asserted for it;⁸⁹ (3) an opportunity to present reasons why the proposed action should not be taken⁹⁰; (4) the right to call witnesses;⁹¹ (5) the right to know the evidence against one;⁹² (6) the right to have the decision based only on the evidence presented;⁹³ (7)

⁸⁵ *Id.* at 1277 (citing *Board of Educ. v. Rice*, [1911] A.C. 179, 182).

⁸⁶ *Id.* at 1278.

⁸⁷ *Id.* at 1279 (“Although an unbiased tribunal is a necessary element in every case where a hearing is required, sharp disagreement can arise over how much in the way of prior participation constitutes bias.”).

⁸⁸ *Id.* at 1280 (“It is likewise fundamental that notice be given and that it be timely and clearly inform the individual of the proposed action and the grounds for it.”) (citing *Boddie v. Connecticut*, 401 U.S. 371, 378 (1971); *Goldberg v. Kelly*, 397 U.S. at 267-68; *Armstrong v. Manzo*, 380 U.S. 545, 550 (1965); *Mullane v. Central Hanover Bank & Trust Co.*, 339 U.S. 306, 314-15 (1950)). *See id.* at n. 76 (citing *Holland v. Oliver*, 350 F. Supp. 485 E.D. Va. 1972) (written notice of charges one hour before hearing to prisoner did not afford inmate due process even though he had been orally informed of the charges three days previously); *Stewart v. Jozwiak*, 346 F. Supp. 1062, 1064 (E.D. Wis. 1972) (prisoner charged with misconduct is entitled to “reasonable advance notice of such hearing”).

⁸⁹ *Id.* at 1280-81 (“Otherwise the individual likely would be unable to marshal evidence and prepare his case so as to benefit from any hearing that was provided[.]”) (citing *In re Gault*, 387 U.S. 1, 33-34 & n.54 (1967)).

⁹⁰ *Id.* at 1281 (citing *Goldberg*, 397 U.S. at 268-69).

⁹¹ *Id.* at 1282 (“Under most conditions there does not seem to be any disposition to deny the right to call witnesses, although the tribunal must be entitled reasonably to limit their number and the scope of examination.”).

⁹² *Id.* at 1283 (“There can likewise be no fair dispute over the right to know the nature of the evidence on which the administrator relies. But with this generalization agreement ends. The most debated issue is the right of confrontation. Since the only provision in the Bill of Rights conferring the right of confrontation is limited to criminal cases, one might think the constitutional right of cross-examination was similarly confined. However, in *Greene v. McElroy*, Chief Justice Warren said that the Court had applied this principle ‘in all types of cases where administrative and regulatory actions were under scrutiny.’” (citing 360 U.S. 474, 497 (1959)).

the right to counsel;⁹⁴ (8) the making of a record;⁹⁵ (9) a statement of reasons;⁹⁶ (10) public attendance;⁹⁷ and (11) judicial review.⁹⁸

Not all of these eleven elements would fit a data due process, of course. The right to call witnesses, for example, would be difficult and potentially cumbersome given how big data systems perform their analytics. However, elements such as “an unbiased tribunal”, “the right to know the evidence against one,” “the making of a record,” and “a statement of reasons” make more sense for data due process. For example, rather than focusing, as FIPPs does, on the right to audit the personal data that has been collected about oneself, data due process would focus on the right to audit the data used to make the determination at issue. Moreover, while both FIPPs and due process describe the concept of “notice” as critical, due process’ notice focuses on the proposed action to be taken against the individual instead of the type and amount of data to be collected or used. Again, because it is hard to predict in advance what processes or queries will be conducted by big data, due process’ notice of a proposed action fits big data’s information practices better than the FIPPs approach to gatekeeping at the collection stage.⁹⁹

An unbiased tribunal and judicial review would also be appropriate for data due process. Algorithmic bias is a serious issue and, as Danielle Citron has noted, there should be means for challenging it.¹⁰⁰ Since predictive privacy harms are often only discernable in hindsight, it may make sense to provide for some agency or judicial oversight when they occur. This oversight would only apply to data due process and not to the actual result. This would

⁹³ *Id.* at 1284 (“While agreeing that these references were wholly appropriate to the witch-hunts of the McCarthy era and that cross-examination is often useful, one must query their universal applicability to the thousands of hearings on welfare, social security and the like which are now held every month”) (citing Davis, *The Requirement of a Trial-Type Hearing*, 70 HARV. L. REV. 193, 213-14 (1956) (need for confrontation and the dangers of “faceless” informers)).

⁹⁴ *Id.* at 1288 (quoting *Powell v. Alabama*: “[t]he right to be heard would be, in many cases, of little avail if it did not comprehend the right to be heard by counsel.) 287 U.S. 45, 68-69 (1932). *Id.* (“Under our adversary system the role of counsel is not to make sure the truth is ascertained but to advance his client's cause by any ethical means.”).

⁹⁵ *Id.* at 1291 (“Americans are as addicted to transcripts as they have become to television; the sheer problem of warehousing these mountains of paper must rival that of storing atomic wastes.”).

⁹⁶ *Id.* (finding a written statement of reasons necessary for purposes of judicial review, to provide for justification as a powerful preventive of wrong decisions, to encourage uniformity across decision-making bodies, and to make decisions somewhat more acceptable to a losing claimant.)

⁹⁷ *Id.* (citing three principal reasons for the right to an open trial as a part of due process: 1) fostering public confidence in the outcome; 2) helping to assure the accuracy of the evidence offered; and 3) pressure on the presiding officials to conduct the proceedings fairly). However, Friendly acknowledges that public attendance can also be disruptive in certain contexts, such as prison disciplinary hearings. *Id.*

⁹⁸ *Id.*

⁹⁹ See Citron, *supra* note __, at 1305-6.

¹⁰⁰ *Id.* at 1262.

ensure that the reviews would be fairly standardized and that the growing expertise of the agency or court performing these reviews would promote efficiency over the long-term.

2. The Nature of the Action

Friendly then goes on to discuss how the nature of the government action should influence the due process requirements. The greater the seriousness of the deprivation to the individual, he argues, the more protections should be in place. For example, taking action against a citizen is far more serious than simply denying a citizen's request.¹⁰¹ Among the deprivations he ranks most highly include revocation of parole or probation, civil commitment, warrants, and revocation of a license to practice a profession. He also suggests that gradations in deprivation also matter: welfare termination is more serious than a reduction; expulsion from public housing is more serious than transfer to a smaller apartment; expulsion from a school is more serious than suspension or loss of credit; and dismissal on a ground carrying a moral stigma is more serious than on one that does not.

In terms of data due process, the type of predictive privacy harm should also influence the due process requirements. The greater the stigma or seriousness of the determination, the more right one should have to question how big data adjudicated that result. For example, health information is among the most precious and protected, so more due process would be afforded. Law enforcement uses would also be among the most subject to scrutiny. On the lesser end might be advertising.¹⁰²

c. The Underlying Values of Due Process

To further assist with assessing the appropriateness of a "due process" approach to big data, it is worth considering the values that underlie many of its rules. As Martin Redish and Larry Marshall have written: "It is possible to devise a model of procedural due process that simultaneously allows the flexibility central to the due process concept as it has evolved, while providing a principled and workable structure. Due process need be flexible mainly in terms of the *specific procedures* that courts require. The *values* that the clause represents, on the other hand, are more enduring."¹⁰³ No matter how ones considers it, application of due process to data requires significant imagination to design the appropriate processes and procedures. As the large computational use of data will be varied and contextual (much like the range of cases that courts consider), a flexible model based more on values and less on specific procedures will likely endure over time.

In their examination of due process, Redish and Marshall set out seven specific enduring sets of values that due process should preserve: (1) accuracy¹⁰⁴; (2) the appearance of fairness¹⁰⁵;

¹⁰¹ See *Friendly*, *supra* note __, at 1295.

¹⁰² Of course, for mixed uses, such as the Target example where it is both advertising and health information, the greater protection should govern.

¹⁰³ Martin H. Redish & Lawrence C. Marshall, *Adjudicator, Independence and the Values of Procedural Due Process*, 95 YALE L.J. 455, 474 (1986).

¹⁰⁴ As the Supreme Court wrote in *Mathews v. Eldridge*, "The rights to notice, hearing, counsel, transcript, and to calling and cross-examining witnesses all relate directly to the accuracy of the adjudicative process." *Id.* at 476.

(3) equality of inputs into the process;¹⁰⁶ (4) predictability, transparency, and rationality;¹⁰⁷ (5) participation;¹⁰⁸ (6) revelation;¹⁰⁹ and (7) privacy-dignity.¹¹⁰

Each of these maps well to our concerns about big data. For big data to deliver the answers we seek, it must be accurate and include all appropriate inputs equally (to overcome any

¹⁰⁵ *Id.* at 483. (“The goal of fostering an appearance of fairness represents the flip side of the accuracy value embodied in the instrumental approach.”), (“It requires that even if a given procedure does not clearly advance the goal of accuracy, it is nonetheless worthwhile insofar as it “generate[s] the feeling, so important to a popular government, that justice has been done.”) (citing *Joint Anti-Fascist Refugee Committee v. McGrath*, 341 U.S. 123, 171-72 (1951) (Frankfurter J., concurring)). *See also id.* (“Describing its prophylactic rules regarding disqualification of judges, the Court explained that “[s]uch a stringent rule may sometimes bar trial by judges who have no actual bias and who would do their very best to weigh the scales of justice equally between contending parties. But to perform its high function in the best way, *justice must satisfy the appearance of justice.*”’) (quoting *In re Murchison*, 349 U.S. 133, 136 (1955)(emphasis added)).

¹⁰⁶ *Id.* at 484 (“[T]he equality value demands that ‘the techniques for making collective decisions not imply that one person’s or group’s contribution (facts, interpretation, policy argument, etc.) is entitled to greater respect than another’s merely because of the identity of the person or group.’”).

¹⁰⁷ *Id.* at 485 (“The values of predictability, transparency and rationality all relate to the ‘participants’ ability to engage in rational planning about their situation, to make informed choices among options.’ ... As Lon Fuller has stated, “[c]ertainly there can be no rational ground for asserting that a man can have a moral obligation to obey a legal rule that does not exist, or is kept secret from him.’” (citing L. FULLER, *THE MORALITY OF LAW*, 39 Rev. ed. 1969)).

¹⁰⁸ *Id.* at 487 n. 130 (“In fact, the Supreme Court has characterized the ‘two central concerns of procedural due process’ as ‘the prevention of unjustified or mistaken deprivations and the promotion of participation and dialogue by affected individuals in the decisionmaking process.’”) (citing *Marshall v. Jerrico, Inc.*, 446 U.S. 238, 242 (1980)). *See also id.* at 487 (“But again a participatory opportunity may also be psychologically important to the individual: to have played a part in, to have made one’s apt contribution to decisions which are about oneself may be counted important even though the decision, as it turns out, is the most unfavorable one imaginable and one’s efforts have not proved influential.”) (quoting Michelman at 127).

¹⁰⁹ *Id.* at 489 “The value of revelation seems to be truly unrelated either to the outcome of the case or to any hope of changing that outcome. As explained by Michelman:

The individual may have various reasons for wanting to be told why, even if he makes no claim to legal protection, and even if no further participation is allowed him. Some of those reasons may pertain to external consequences: the individual may wish to make political use of the information, or use it to help him ward off harm to his reputation. Yet the information may also be wanted for introspective reasons—because, for example, it fills a potentially destructive gap in the individual’s conception of himself.

Michelman at 127.

¹¹⁰ *Id.* at 482 (citing cases condemning coerced confessions as exemplifying this principle, such as *Brown v. Mississippi*, 297 U.S. 278, 287 (1936)).

signal problems); otherwise its outputs will lead us astray as Google Flu Trends did this year. Before there can be greater social acceptance of big data's role in decision-making, especially within government, it must also appear fair, and have an acceptable degree of predictability, transparency, and rationality. Without these characteristics, we cannot trust big data to be part of governance.¹¹¹ Finally, while perhaps not necessary in the same way, participation, revelation, and privacy-dignity would help optimize big data's role in public decision-making for the same reasons that they help with optimizing the judicial or administrative process – bringing legitimacy. Or put another way: “citizens care enormously about the process by which outcomes are reached—even unfavorable outcomes.”¹¹²

Redish and Marshall also raise two cautionary concerns about what they consider to be the centerpiece of any due process framework – the independent adjudicator.¹¹³ Specifically, they highlight the dangers that arise when an adjudicator has either a direct financial interest¹¹⁴ in the outcome of the proceeding or an inherent personal bias.¹¹⁵ As we note in another article, issues of bias also exist within big data's algorithms and data sets, despite their appearance of objectivity.¹¹⁶ Further, there can be no doubt that for-profit providers of big data analytics have direct financial interests in some of the outputs they produce. These present two additional reasons to apply due process to these data regimes.

d. Due Process as Separation of Powers and Systems Management

Another favorable reason to consider due process as a mechanism for addressing how big data handles personal information is the role it has historically played as a means of separating powers among governments. In other words, due process has ensured that those who pass general laws are kept separate from those who are called upon to enforce them in specific circumstances and those who judge whether or not those cases have merit. As McConnell writes, this protects citizens against directed executive punishment in the form of adjudication. Congress may pass laws affecting our lives, liberty, and property and the President may sign them, but their enforcement requires a fair process overseen by a neutral

¹¹¹ By the term “governance”, we are referring primarily to networked or technological governance, which involve both governmental aspects as well as private and individual ones. See Kate Crawford and Catherine Lumby, *Networks of Governance: Users, Platforms, and the Challenges of Networked Media Regulation*, 2 INTERNATIONAL JOURNAL OF TECHNOLOGY POLICY AND LAW ___ (Forthcoming 2013). Available at <http://ssrn.com/abstract=2246772>.

¹¹² Robert J. MacCoun, *Voice, Control, and Belonging: The Double-Edged Sword of Procedural Justice*, 1 Annual Rev. Soc. Sci. 171, 171 (2005).

¹¹³ Redish & Marshall, *supra* note ___, at 494.

¹¹⁴ See *Dr. Bonham's Case*, 77 Eng. Rep. 646, 8 Coke 114a C.P. (1610). See also *Tumey v. Ohio*, 273 U.S. 510 (1927) (holding Judge-Mayor's direct financial interest in the case's outcome violated due process because it served as a “possible temptation to the average man as a judge” even absent any finding of actual partiality, noting that actual influence could rarely be proven).

¹¹⁵ Redish & Marshall, *supra* note ___, at 500.

¹¹⁶ See Kate Crawford, “The Hidden Biases in Big Data,” HARVARD BUSINESS REVIEW, Apr. 1, 2013, http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html; Kate Crawford, “Think Again: Big Data”, FOREIGN POLICY, May 9, 2013, http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data.

arbitrator.¹¹⁷ Thus, a core component of due process is separating out those who write the legal code from those to adjudicate using it.¹¹⁸

With many big data determinations, there is little or no regulation of the interactions between the designer of the algorithm (the lawmaker), the persons oversees the queries (the executive), and the adjudicator (the computational output). In other words, there is no system of checks and balances to ensure that biases are not present in the system, especially a system of enforcement. “Chief Justice Marshall put it this way: “It is the peculiar province of the legislature to prescribe general rules for the government of society; the application of those rules to individuals in society would seem to be the duty of other departments.”¹¹⁹ Due process would help ensure that big data does not blur its processes with its provinces.

Various due process scholars have also conceptualized the doctrine as a form of “systematic management technique” that should focus less on any individual harm and more on discovering errors, identifying their causes, and implementing corrective actions.¹²⁰ Or, as Richard Fallon suggests, to include injustices in individual level but also look beyond them to the managerial level, creating schemes and incentives to normatively circumscribe government actions within the bounds of law.¹²¹ Similarly, due process can help as a systematic management technique for big data, uncovering errors and identifying their causes and providing schemes and incentives to correct them in keeping within the bounds of privacy laws and norms.

IV. Toward a Model for Data Due Process

a. Technological Due Process: the Citron Analysis

The idea of applying due process to automated systems generally is not new. In her 2010 article *Technological Due Process*,¹²² Danielle Citron examined the use of automated systems in governmental administrative proceedings, the risks they posed to deprivations of liberty and

¹¹⁷ Chapman & McConnell, *supra* note ___, at 1677 (“Legislative acts violated due process not because they were unreasonable or in violation of higher law, but because they exercised judicial power or abrogated common law procedural protections”). *Id.* at 1684-4 (“A 1368 statute, for example, provided that ‘no Man [shall] be put to answer without Presentment before Justices, or Matter of Record, or by due Process and Writ original, according to the old Law of the Land; the law expressly forbade adjudications by the King’s councils instead of the common law courts.”). *Id.* at 1716 (“Hamilton, relying on Coke, maintained that the law of the land requires certain procedural safeguards before someone may be deprived of his rights. The legislature is inherently incapable of providing those safeguards, and thus the deprivation of rights must be left to that branch of government capable of doing so.”).

¹¹⁸ See Lessig, *supra* note ___.

¹¹⁹ See Chapman & McConnell, *supra* note ___, at 1733.

¹²⁰ See Jerry L. Mashaw, *The Managerial Side of Due Process: Some Theoretical and Litigation Notes on the Assurance of Accuracy, Fairness, and Timeliness in the Adjudication of Social Welfare Claims*, 59 CORNELL L. REV. 722 (1974).

¹²¹ Richard H. Fallon, Jr., *Some Confusion about Due Process, Judicial Review, and Constitutional Remedies*, 93 COLUM. L. REV. 309 (1993).

¹²² 85 WASH U. LAW REV. 1249 (2010).

property, and how a reinvigorated approach to due process could help mitigate and address these problems.¹²³ There, she identifies several issues with these systems that could be expanded to address the predictive privacy harms of big data.

First, she identifies various automated systems that government administrative officials use to adjudicate individual liberty or property interests, including systems designed to decide termination of Medicaid, food stamp, and other welfare benefits; targeting of people for exclusion from air travel; parents who neglected child support payments, voters to be purged from rolls without notice, and small businesses deemed ineligible for federal contracts.¹²⁴ She also notes that most of these systems failed to give adequate notice to individuals whose interests were at stake,¹²⁵ any opportunity to be heard before a decision was rendered,¹²⁶ and often adjudicated their case in secret or without leaving any record for audits or judicial review.¹²⁷

In particular, she notes that automatic systems generally fail to give adequate notice to individuals when their liberty or property interests are being algorithmically adjudicated. In administrative proceedings, notice of an action against one's liberty interest be "reasonably calculated" to inform one of the issues to be decided, the evidence supporting the case, and the decisional process that agency will use. As Citron writes, clear notice is meant to "decrease the likelihood that agency action will rest upon 'incorrect or misleading factual premises or on the misapplication of rules.'"¹²⁸ As a result, "affected individuals lack the information they would need to effectively respond to an agency's claim."¹²⁹ To counteract this failure to give notice, Citron argues that automated administrative systems must include audit trails that record the facts and rules supporting each decision.¹³⁰ This trail can then be compiled into some form of sufficient notice when a decision is made and transmitted to the subject of the decision.

¹²³ *Id.* See also Ian Kerr, *Prediction, Preemption, Presumption: The Path of Law After the Computational Turn*, forthcoming in *PRIVACY, DUE PROCESS AND THE COMPUTATIONAL TURN: THE PHILOSOPHY OF LAW MEETS THE PHILOSOPHY OF TECHNOLOGY*, eds. Mireille Hildebrandt & Ekaterina De Vries, available at <http://iankerr.ca/wp-content/uploads/2011/08/SUBMISSION-Prediction-Presumption-Preemption-CFS.-August-2011-Website-Template-Edits.pdf> (noting that "from a broad legal and ethical perspective, problems are sure to arise when anticipatory algorithms and other computational systems import norms that undermine the due process otherwise afforded to citizens by law."); Daniel J. Steinbock, *Data Matching, Data Mining, and Due Process*, 40 Ga. L. Rev. 1 (2005) (addressing the due process effects of using data matching and mining to identify persons against whom official action is taken, such as in use of air passenger screening and the maintenance of watch lists).

¹²⁴ *Id.* at 1252.

¹²⁵ *Id.* at 1281.

¹²⁶ *Id.* at 1283.

¹²⁷ *Id.* at 1253 n. 22, 1279.

¹²⁸ *Id.* at 1282 (citing *Goldberg*, 397 U.S. at 268). See also *Big Data for All*, supra note ___, at 271. ("Fairness and due process mandate that individuals are informed of the basis for decisions affecting their lives, particularly those made by machines operating under opaque criteria.")

¹²⁹ *Id.* at 1282 (citing *Cosby v. Ward*, 843 F.2d 967, 984 (7th Cir. 1988)).

¹³⁰ *Id.* at 1305.

Big data systems suffer from many of the same weaknesses as government administration systems regarding notice. Individuals or groups that are subjected to predictive privacy harms rarely receive any meaningful notice of the predictions before they occur or are implemented, and even then, providers are unlikely to share the evidence and reasoning for the predictions that were made. Certainly, there is currently no legal requirement that the provider archive any audit trail or retain any record of the basis of the prediction.

Opportunities to be heard also present problems for automated systems and due process.¹³¹ Citron posits that an “opportunity to be heard” in this context would involve access to an automated program’s source code or a hearing on the logic of a computer program’s decision, and would often be found far too expensive under the *Mathews* balancing test.¹³² She notes, however, that because such challenges would further fairness for future cases, it might be worth it even under *Mathews*.¹³³ There are, however, potential problems with this approach. Citron acknowledges that certain automated systems, such as the “No Fly” list, involve state secrets and would rarely be subjected to scrutiny.¹³⁴

In response, Citron suggests first that instead of subjecting every automated system to cross-examination, one could at a minimum invest in extra education about the biases and fallacies of automation for government personnel who use the systems to make administrative decisions.¹³⁵ Educating these individuals about the systems’ flaws could help them scrutinize their outputs more fairly.¹³⁶ Second, she suggests that agencies should require hearing officers “to explain, in detail, their reliance on an automated system’s decision” including any computer-generated facts or legal findings.¹³⁷ Third, she suggests that agencies should be required to regularly test their system’s software for bias and other errors.¹³⁸

For big data adjudications, many of these same problems exist with algorithmic biases and the potential to inaccurately predict PII about individuals. Thus, similar “opportunities to be heard” may well be appropriate, especially with respect to educating data scientists about the biases of big data, requiring those who use big data for significant decisions concerning individuals to disclose which data sets were used, and also to require testing of predictive analytics to assess how accurate a given system can be.

Finally, Citron discusses what meaningful judicial review for automated systems might entail and why most of these systems evade it.¹³⁹ Specifically, she critiques automated administrative systems because they often fail to retain any audit record of how they made

¹³¹ *Id.* at 1283 (noting that *Mathews* aspires to provide an opportunity to be heard “at a meaningful time and in a meaningful manner” but only if it is cost-efficient).

¹³² *Id.* at 1284.

¹³³ *Id.*

¹³⁴ *Id.* at 1286.

¹³⁵ *Id.* at 1306.

¹³⁶ *Id.* (noting the success of special workshops on scientific theory and methodology with federal district court judges who need to assess the reliability of expert testimony).

¹³⁷ *Id.* at 1307.

¹³⁸ *Id.* at 1310.

¹³⁹ *Id.* at 1298.

the decisions at issue or upon what data the decision was based.¹⁴⁰ Again, similar to the need for notice, an audit trail for big data may provide similar reassurance and increase accuracy. It would also allow individuals to raise specific objections to how and when their data is being used in various processes.

b. Procedural Data Due Process

As noted above, procedural due process generally describes the constitutional requirement that any government deprivation of a liberty or property right must be preceded, at a minimum, by notice and the opportunity for a hearing on the matter before an impartial adjudicator. In thinking about procedural data due process, we will use these same three elements as well as occasionally drawing additional ones from Judge Friendly's list of eleven and from the values of due process outlined by Redish and Marshall.¹⁴¹

To begin, we note that some uses of big data will be difficult to fit in the mold of individualized due process adjudication, such as the opportunities individuals were not selected for, the advantageous insurance advertising offers that did not appear in their search results, the jobs they never knew existed because they didn't fit the desired profile of a marketer. But when individuals are aware of or involved directly in processes where big data is used as part of the outcome, such as when it is used to identify top candidates from a given pool of applicants, individualized due process approaches will seem most appropriate. For the more opaque predictive problems, such as a real estate offer one never sees because big data might have judged one unworthy, a more structural due process approach might be better, with oversight and auditing primarily driven by public agencies.

Another key question arises as to when due process should attach to a particular decision. While the exact moment for any particular decision is too specific for this article to discuss, one can imagine that certain determinations will be more regularized and repetitive (such as a determination of employment eligibility),¹⁴² and thus due process could attach from the moment the decision is made to determine the specific eligibility of a particular set of individuals. The moment of attachment could also be triggered sooner as the generated data approaches the equivalent of PII. The closer to PII-type information, the stronger the case for procedural data due process to attach. In addition, the greater the seriousness of the decision, the more big data due process is afforded.

With these questions in mind, we turn to principles for implementation.

1. Notice

¹⁴⁰ *Id.* at 1300.

¹⁴¹ Chapman & McConnell, *supra* note __, at 1774 (“In keeping with arguments advanced by lawyers and courts from the earliest days of the Republic, the Supreme Court declared in *Murray’s Lessee v. Hoboken Land & Improvement Co.* that to comply with due process, statutes must either provide for the use of common law procedures or, if they do not, employ alternative procedures that the courts would regard as equivalently fair and appropriate.”)

¹⁴² *See, e.g.,*

<http://online.wsj.com/article/SB10000872396390443890304578006252019616768.html>.

Our conception of notice for procedural data due process centers on providing those who may suffer from predictive privacy harm an opportunity to intervene in the predictive process to ensure fairness with respect to the processes by which their interests are being affected, either individually or structurally. This would require those who use big data to “adjudicate” others, i.e. make categorical or attributive determinations, to post some form of notice disclosing not only the type of predictions they are attempting, but also the general sources of data that they are drawing upon as inputs, including a means whereby those whose personal data is included can learn of that fact.

One could also imagine a variety of notice “rights” and obligations that would enable consumers to petition big data providers to check and see if their data was being included or used in any predictive adjudications, and whether that data was accurate. Similar laws are available for data collection.¹⁴³ These models could be expanded to include data processing for big data so that just as privacy policies disclose data collection practices, so too would they disclose data prediction practices “reasonably calculated” to inform individuals of the risks to which they may be exposed in terms of predictive privacy harms. Moreover, when a particular set of predictions about an individual or discrete group has been queried (or “adjudicated”), notice would be sent out that is “reasonably calculated” to inform those affected of, at a minimum, the issues that were predicted if not also the data considered and the methodology employed. At a minimum, this notice should provide for a mechanism to access the audit trail or record created in the predictive process.¹⁴⁴

For example, if a company were to license search query data from Google and Bing in order to predict which job applicants would be best suited for a particular position, it would have to disclose to all applicants who apply that it uses search queries for predictive analytics related to their candidacy. Or in the case of predictive policing, the government would have to notify citizens that it was using predictive analytics and the particular sets of public records to determine which areas of a city they have marked out as “hot spots” as well as the capacity to determine if one lived or worked within the actual hot spots.

Another example would be around the issue of fair housing. If landlords and real estate companies were to shift away from general advertising in media outlets and toward using big data to determine likely buyers or renters who fit their “ideal” profiles, again we could require them to disclose this practice. Depending on the specifics of the practice, one could imagine the notice either on an individual level (to those who knew of their inclusion

¹⁴³ See, e.g., Rainey Reitman, “New California ‘Right to Know’ Act Would Let Consumers Find Out Who Has Their Personal Data -- And Get a Copy of It”, EFF.org, April 2, 2013, <https://www.eff.org/deeplinks/2013/04/new-california-right-know-act-would-let-consumers-find-out-who-has-their-personal>; California Online Privacy Protection Act, Cal. Bus. & Prof. Code §§ 22575–22579 (requiring commercial operators of online services to conspicuously post privacy policies informing users of what personally identifiable information is being collected and how it will be used).

¹⁴⁴ See Citron, *supra* note __ (discussing audit of code and data for administrative technological systems used by government); Dwork & Mulligan, *supra* note __ at 38-39 (suggesting the use of test files to analyze the fairness of big data as a socio-technical system).

in the predictions) or on a structural level (if the predictions were for a large set of a given population).

2. Opportunity for a Hearing

Once notice is available, the question then becomes how one might challenge the fairness of the predictive process employed. We believe that the most robust mechanism for this is the opportunity to be heard, and if necessary, correct the record. This would include examining the evidence used including both the data input and the algorithmic logic applied. In contexts where security and proprietary concerns arise, or in more structural situations, this role could be given to a trusted third party, a neutral data arbitrator, who could routinely examine big data providers who adjudicate in ways where predictive privacy harms occur. For example, the Federal Trade Commission which current addresses many privacy harms involving technology and has recently hired technologists to assist its investigations and enforcement actions,¹⁴⁵ could investigate complaints based on predictive privacy harms and in the process of those complaints investigate the basis of the predictions.¹⁴⁶

The presence of a neutral data arbiter would not only provide for the public to have an opportunity to be heard and examine and challenge the evidence used in adjudicative predictions, but it would also comport with several of the underlying values of due process: accuracy of the determination, appearance of fairness, predictability, transparency, and rationality, participation, and revelation. In particular, because big data generally excludes any user participation in its decision-making, a neutral data arbiter would be especially important to ensure that there was a “meaningful” hearing for public concerns.

3. Impartial Adjudicator and Judicial Review

One of the primary myths about big data is that it produces outputs that are somehow free from bias and closer to objective truth than other forms of knowledge.¹⁴⁷ Due process requires that those who deprive individuals of a liberty interest do so without unwarranted bias or a direct financial interest in the outcome. Therefore procedural data due process can also serve as a valuable framework for ensuring greater fairness with predictive analytics. A neutral data arbiter could field complaints and then investigate sufficient allegations of bias or financial interest that might render the adjudication unfair. In particular, drawing on the literature exploring due process as a function of separation of powers, the arbiter could examine the relationship between those who designed the analytics and those who run the individual processes to make sure that their roles are appropriate and distinct. This would require some form of audit trail that recorded the basis of predictive decisions, both in terms

¹⁴⁵ See Federal Trade Commission, *FTC Names Edward W. Felten as Agency’s Chief Technologist; Eileen Harrington as Executive Director*, Nov. 4, 2010, <http://www.ftc.gov/opa/2010/11/cted.shtm>.

¹⁴⁶ This would also address concerns about standing, where a single plaintiff might not have sufficient evidence to show individual concrete harm in their particular case without gathering evidence through the litigation discovery process.

¹⁴⁷ boyd & Crawford, *supra* note ___, at 667.

of data used and the algorithm employed. Such audits are already in use in various data-mining contexts and, thus, would not be unreasonable to require.¹⁴⁸

V. Conclusion

In concluding his article on hearings, Judge Friendly wrote “We have traveled over wide areas—from termination of welfare payments to the establishment of incentive per diem for freight cars, from student and prison discipline to rates for natural gas. Yet the problem is always the same—to devise procedures that are both fair and feasible.”¹⁴⁹ We end here with the same observation. Big data presents many challenges for privacy to which we believe a model of procedural data due process can respond. How exactly it responds to each may vary, but ultimately it will succeed if it can ensure protections that are both fair and feasible for those that are risk from this new form of privacy harm.

¹⁴⁸ See Gray & Citron, *supra* note ___, at 38-9 (noting that the predictive analytic systems of New York City’s DAS and Palantir both provide mechanisms for outside review). On the question of remedies, with procedural safeguards, one can imagine several common remedies for curing a violation. For example, in the judicial system, a specific proceeding or determination might be invalidated, thereby forcing the agency or adjudicator to revisit the determination using proper processes. In the privacy context, there is some precedent for this in France.

¹⁴⁹ Friendly, *supra* note ___, at 1315.