2019

The Promises and Perils of Using Big Data to Regulate Nonprofits

Lloyd Histoshi Mayer
Notre Dame Law School, lmayer@nd.edu

THE PROMISES AND PERILS OF USING BIG DATA TO REGULATE NONPROFITS

Lloyd Hitoshi Mayer*

Abstract: For the optimist, government use of "Big Data" involves the careful collection of information from numerous sources. The government then engages in expert analysis of those data to reveal previously undiscovered patterns. Discovering patterns revolutionizes the regulation of criminal behavior, education, health care, and many other areas. For the pessimist, government use of Big Data involves the haphazard seizure of information to generate massive databases. Those databases render privacy an illusion and result in arbitrary and discriminatory computer-generated decisions. The reality is, of course, more complicated. On one hand, government use of Big Data may lead to greater efficiency, effectiveness, and transparency; on the other hand, such use risks inaccurate conclusions, invasions of privacy, unintended discrimination, and increased government power. Until recently, these were theoretical issues for nonprofits because federal and state regulators did not use Big Data to oversee them. But nonprofits can no longer ignore these issues, as the primary federal regulator is now emphasizing "data-driven" methods to guide its audit selection process, and state regulators are moving forward with plans to create a single, online portal to collect required filings. In addition, regulators are making much of the data they collect available in machine-readable form to researchers, journalists, and other members of the public. The question now is whether regulators, researchers, and nonprofits can learn from the Big Data experiences of other agencies and private actors to optimize the use of Big Data with respect to nonprofits. This Article explores the steps that nonprofit regulators have taken toward using Big Data techniques to enhance their ability to oversee the nonprofit sector. It then draws on the Big Data experiences of government regulators and private actors in other areas to identify the potential promises and perils of this approach to regulatory oversight of nonprofits. Finally, it recommends specific steps regulators and others should take to ensure that the promises are achieved and the perils avoided.

---

* Professor, Notre Dame Law School. I am very grateful for helpful comments received on earlier drafts from participants in the International Society of Third Sector Research 13th International Conference, the Notre Dame Faculty Colloquium, the Nonprofit Forum, and the Tulane Intellectual Life Speaker Series. I am also very grateful for research assistance from Thomas Everett and Jasmine Rasheed.

INTRODUCTION

On May 8, 2015, the Commissioner of the Internal Revenue Service's Tax Exempt and Government Entities Division highlighted the new "data-driven approach" the IRS would take to capitalize on the "tremendous amount of information" it collects.[1] She was elaborating on an internal "Program Letter" released late in 2014 that referred to "Data-Driven Decision-Making" but provided almost no explanation of what this term meant other than cryptic commitments to develop "sophisticated analytics," to conduct "analysis to identify opportunities to improves processes," to apply "analytics and research to improve program effectiveness," and to allocate resources "using a data-driven

1. Sunita Lough, Comm'r of the Tax Exempt & Gov't Entities Div., Internal Revenue Serv., Remarks at the Meeting of the Exempt Organizations Committee of the American Bar Association Section of Taxation (May 8, 2015), *in* EO TAX J. 2015-96 (2015).

approach to target existing and emerging high-risk areas."[2] For anyone familiar with recent technological developments, however, these buzzwords signaled that the IRS was moving in a new direction with its oversight of tax-exempt nonprofit organizations in the United States that could be boiled down to two words: Big Data.

Oversimplifying, Big Data refers to the collection and analysis of information that is so large in scope, changes so rapidly, and varies in structure to such an extent that it is not amenable to conventional database techniques of the recent past, but instead requires sophisticated computerized methods to adequately gather and learn from this information.[3] Rapid improvements in storage capacity and computing power have made it possible to both collect and analyze such data. While the term Big Data as it is used today is only a little over twenty years old, it already has developed technological, academic, and societal meanings.[4] At its heart, however, Big Data refers to the rapid accumulation of digital data, from social media posts to cell phone locations, as well as efforts to use such data to discover significant patterns that inform and improve public policy making, business decisions, and personal choices.[5]

A couple of noteworthy examples have highlighted the potential benefits and dangers of Big Data. Starting in 2002, retailer Target used customers' shopping patterns to determine if it was likely they were pregnant.[6] While Target apparently did this only so it could better customize its advertising to those customers, it was wisely wary about how its customers would react if they learned how much Target knew

---

2. SUNITA LOUGH, TAX EXEMPT & GOV'T ENTITIES DIV., INTERNAL REVENUE SERV., TE/GE PROGRAM LETTER FY 2015, at 3 (2014) [hereinafter 2014 Program Letter], https://www.irs.gov/pub/irs-tege/tege_fy15_program_letter.pdf [https://perma.cc/HP2B-AV8P].

3. *See generally* FRANCIS X. DIEBOLD, ON THE ORIGIN(S) AND DEVELOPMENT OF "BIG DATA": THE PHENOMENON, THE TERM, AND THE DISCIPLINE (2018), https://www.sas.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf [https://perma.cc/T59X-D4QW] (discussing the origin of the term "Big Data").

4. *See* Danah Boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFO. COMMC'N & SOC'Y 662, 663 (2012) (defining *Big Data* "as a cultural, technological, and scholarly phenomenon"); DIEBOLD, *supra* note 3, at 3 (discussing the first uses of the term Big Data in its current sense).

5. *See* JAMES MANYIKA ET AL., BIG DATA: THE NEXT FRONTIER FOR INNOVATION, COMPETITION, AND PRODUCTIVITY 1–2 (2011), https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx [https://perma.cc/K6GP-MU7Y] (discussing potential positive uses of Big Data).

6. Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES MAG. (Feb. 16, 2012), https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html [https://perma.cc/4P7H-G3L4] (explaining how private companies use Big Data).

about them (including one instance where a father learned his teenage daughter was pregnant because she received baby-related Target ads in the mail).[7]

Government use of Big Data can be even more unnerving. Edward Snowden famously disclosed that the National Security Agency had been collecting telephone metadata—such as the time and location of calls, as opposed to their content—for every call generated by Verizon customers, so that these data could then be analyzed using various methods to identify certain connections of interest.[8] While Congress reacted to these disclosures by amending the relevant federal statutory provisions, it is unclear whether those amendments prevent continued collection of such data.[9] Other government uses of Big Data methods include environmental protection and Medicare and Medicaid fraud prevention.[10]

The IRS's Big Data move with respect to nonprofits highlights that data about such organizations may have reached a critical tipping point in several respects. First, the IRS's improved analysis of the information it already collects raises the possibility of more efficient, effective, and transparent federal oversight of nonprofits. Second, state regulators are striving to enhance their own oversight capabilities by centralizing their collection of information relating to charitable nonprofits through a "Single Portal" initiative.[11] Third, as a result of long-standing federal

---

7. *Id.*

8. *See* Margaret Hu, *Bulk Biometric Metadata Collection*, 96 N.C. L. REV. 1425, 1452–55 (2018).

9. *Id.* at 1462–68 (explaining the reasons for this lack of clarity and concluding that "under the USA FREEDOM Act, it is unclear whether bulk metadata collection will cease and, thus, whether mass suspicion-less tracking of metadata by the intelligence community will continue in an under-regulated manner").

10. *See* U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-17-710, MEDICARE: CMS FRAUD PREVENTION SYSTEM USES CLAIMS ANALYSIS TO ADDRESS FRAUD 4 (2017) (outlining a congressionally-mandated data analytic system to prevent fraud, waste, and abuse in the Medicare fee-for-service program); Meredith Somers, *EPA Using Data Science to Save Environment, Support Agencies*, FED. NEWS NETWORK (Oct. 18, 2016), https://federalnewsnetwork.com/technology-main/2016/10/epa-using-data-science-save-environment-support-agencies/ [https://perma.cc/32EW-GPJJ] (discussing EPA Big Data initiatives); *see generally* SUBCOMM. ON NETWORKING AND INFORMATIONAL TECH. RESEARCH AND DEV., NAT'L SCI. AND TECH. COUNCIL, SUPPLEMENT TO THE PRESIDENT'S BUDGET FOR FISCAL YEAR 2018, at 46–49 (2017), https://www.nitrd.gov/pubs/2018supplement/FY2018NITRDSupplement.pdf [https://perma.cc/4BYL-LKNP] (describing Big Data agency investments across the federal government).

11. *GuideStar, MRFP, and CityBase Creating Single System for State Charity Registration and Reporting*, GUIDESTAR (Apr. 4, 2018) [hereinafter GUIDESTAR, Press Release], https://learn.guidestar.org/news/news-releases/nasco-guidestar-mrfp-and-citybase-creating-single-system-for-state-charity-registration-and-reporting [https://perma.cc/6KU5-VWRC]; *Registration*

laws, recent litigation, and the efforts of several private organizations, the data collected by the IRS are also generally accessible to the public, including journalists and researchers, with even greater access in the future given the recent enactment of legislation to expand the electronic filing obligations of tax-exempt nonprofits.[12]

The promise of Big Data for government oversight of nonprofits is similar to its promise in other regulatory areas: greater efficiency, effectiveness, and transparency with respect to enforcement of existing legal rules and development of new rules.[13] Additional benefits that could accrue from the public's ability to use such data include better informing donors, customers, members, and nonprofit leaders about the nonprofits of particular interest to them.[14] But as in other areas, Big Data brings possible perils: bad data, bad analysis, over reliance, threats to privacy, discriminatory effects, unchecked government power, and other legal concerns, including constitutional ones.[15]

This Article considers what lessons can be learned from Big Data's application in other areas to inform its emerging application to government and public oversight of nonprofits. While focused primarily on the United States, these lessons are also applicable to other countries, many of which are also moving in the Big Data direction with respect to overseeing nonprofits within their jurisdictions.[16] The ultimate goal is to provide a roadmap for fulfilling the promises of Big Data while avoiding its perils. This roadmap should include measures to ensure awareness of the positive and negative ramifications of Big Data, to rigorously evaluate the improvements to efficiency and effectiveness resulting from a Big Data approach, and to implement precautions to avoid or minimize the perils it presents.

Part I examines the data and analytic tools currently available to the IRS, state regulators, and the public. It also examines the data and tools that may be available to them in the near future, including based on the experience of nonprofit regulators in other countries with Big Data techniques. Part II considers the promise of the Big Data approach for improving government oversight of and public knowledge about nonprofits. Part III explores the perils of this approach for the nonprofit

---

and Filing Portal, Single Portal Initiative, MRFP (2016) [hereinafter MRFP, INC.], http://mrfpinc.org/about.html [https://perma.cc/QKN9-QT2U].

12. *See infra* note 64 and accompanying text; *infra* section I.C.

13. *See infra* section II.A.

14. *See infra* section II.C.

15. *See infra* Part III.

16. *See infra* section I.D.

sector, and ways that regulators and the public can avoid potential missteps.

Part IV concludes with recommendations for ensuring that Big Data as applied to nonprofits ultimately benefits oversight of the nonprofit sector without unduly harming that sector or the public more generally. Those recommendations include the IRS, state regulators, and their independent oversight bodies rigorously evaluating the effectiveness of Big Data approaches and identifying ways to enhance the accuracy of the information they collect and to protect against unnecessary invasions of privacy and prohibited discrimination; researchers being conscious of the accuracy and representational limitations of the data available; and nonprofits being sensitive to the increased visibility of their government filings and other information. These measures will help ensure the promises of Big Data are realized, while minimizing exposure to the perils it creates.

I. BIG DATA AND NONPROFITS

As a brief background on data and datasets generally, Big Data datasets are often characterized as having at least three dimensions: *volume*, measured in bytes; *velocity*, measured by how close creation is to real time; and *variety*, in terms of degree of structure.[17] There are no clear thresholds along any of these dimensions that determine when a dataset becomes Big Data, especially since any such lines would be vulnerable to technological change. That said, volume for Big Data tends to be so large that it is measured in terabytes or petabytes of data.[18] To give a sense of scale, the IRS Director for Research Databases reported in 2013 that the then-volume of data collected by the IRS in its Compliance Data Warehouse was approximately 1.3 petabytes.[19]

---

17. *See* ROB KITCHIN, THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES AND THEIR CONSEQUENCES 68 (2014); DOUG LANEY, 3D DATA MANAGEMENT: CONTROLLING DATA VOLUME, VELOCITY, AND VARIETY 1 (2001), https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [https://perma.cc/4BN7-JBD3].

18. *See* STEVEN FINLAY, PREDICTIVE ANALYTICS, DATA MINING AND BIG DATA: MYTHS MISCONCEPTIONS AND METHODS 13 (2014); MANYIKA ET AL., *supra* note 5, at 1. A terabyte is approximately 1000 gigabytes and a petabyte is approximately 1000 terabytes. *See* FINLAY, *supra*, at 13.

19. Jeff Butler, Presentation at the Government Big Data Symposium 4 (Mar. 5–6, 2013), https://docplayer.net/16174543-Big-data-and-analytics-at-the-irs.html [https://perma.cc/8FTA-YN6W]. A recent new story stated that the Compliance Data Warehouse "has about 40 data sets on taxpayers stretching back more than 30 years." Vidya Kauri, *AI Helping IRS Detect Tax Crimes with Fewer Resources*, LAW360 (Dec. 5, 2018, 8:58 PM EST), https://www.law360.com/articles/1108419/ai-helping-irs-detect-tax-crimes-with-fewer-resources [https://perma.cc/W6E3-AW62].

Velocity for many large datasets now approaches real-time collection, as illustrated by automatic recording of users' searches and other online activities.[20] Variety reflects the extent to which data is organized into fixed fields, like in spreadsheets, versus data that is less fixed, like in emails or tweets.[21]

The process of analyzing expansive datasets is part of the broader field known as "knowledge discovery in databases" or KDD, with the "process of knowledge discovery . . . often subdivided in several steps, such as: (1) capturing and cleansing; (2) aggregating; (3) data mining; and (4) interpreting."[22] All of these steps rely on "machine learning," which is "[a] subspecialty of computer science (within a field historically called 'artificial intelligence') concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data."[23] It is this ability of computer systems to modify their behaviors without human intervention that allows them to process the vast amounts of quickly changing and various types of data that make up Big Data datasets.[24]

This Part examines the extent to which Big Data datasets already exist for nonprofits. It also considers the related issue of to what extent the IRS and others have developed analytical tools, and the potential for the emergence of additional datasets and tools in the near future.

A. The Internal Revenue Service

The Internal Revenue Service (IRS) has been gathering data about taxpayers since its initial creation as the Bureau of Internal Revenue more than 150 years ago.[25] It began computerizing the collection and analysis of its data more than fifty years ago.[26] About fifteen years ago, the IRS started exploring how it could use the data it collects in a

---

20. EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 5 (2014), https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf [https://perma.cc/JD5M-RGUB].

21. *See* MANYIKA ET AL., *supra* note 5, at 33 (discussing how "structured data" resides in fixed fields, as in a spreadsheet, while "unstructured data" does not reside in fixed fields, as in the body of an email).

22. MIREILLE HILDEBRANDT, SMART TECHNOLOGIES AND THE END(S) OF LAW 31–32 (2015).

23. MANYIKA ET AL., *supra* note 5, at 29.

24. *See* KITCHIN, *supra* note 17, at 102–04 (explaining the role of machine learning in the analysis of Big Data).

25. Joseph J. Thorndike, *Reforming the Internal Revenue Service: A Comparative History*, 53 ADMIN. L. REV. 717, 723–25 (2001).

26. *Id.* at 766–67.

comprehensive way to enforce the nation's tax laws.[27] Such use includes not only collecting and analyzing return information provided by taxpayers and other entities, but also collecting and analyzing information obtained from other sources. The IRS now gathers data from many commercial and public data pools—including social media sites like Facebook, Instagram, and Twitter—and plans to gather more information along these lines.[28] Therefore, the 1.3 petabytes of compliance-related data the IRS had gathered as of 2013 reflects not only information provided in tax returns and other filings, which only accumulates at about fifteen-to-twenty terabytes per year, but also a substantial amount of data collected from other sources.[29] This represents an almost ten-fold increase in data during the five years from 2008.[30] The IRS recently estimated that during the ten years from 2007 to 2017 its data volume increased a hundredfold.[31] These efforts are in addition to other financial-related data collection by the federal government.[32] Unfortunately, more recent information regarding the

---

27. David DeBarr & Maury Harwood, *Relational Mining for Compliance Risk*, *in* IRS RESEARCH BULLETIN: RECENT IRS RESEARCH ON TAX ADMINISTRATION AND COMPLIANCE, PUBLICATION 1500, at 175, 175 (2005), https://www.irs.gov/pub/irs-soi/04debarr.pdf [https://perma.cc/32QF-RTYC].

28. *See* Siri Bulusu, *'Doubling Down' on Big Data Upgrades IRS Fraud Investigations*, DAILY TAX REP. (Dec. 5, 2018, 6:22 PM) ("The IRS can currently track data coming from electronic transactions, cell phones, live video feeds, and social media."); Kimberly A. Houser & Debra Sanders, *The Use of Big Data Analytics by the IRS: Efficient Solutions or the End of Privacy as We Know It?*, 19 VAND. J. ENT. & TECH. L. 817, 819 (2017) ("The IRS uses big data analytics to mine commercial and public data pools including social media sites (e.g., Facebook, Instagram, and Twitter)."); Kauri, *supra* note 19 (IRS $99 million, seven-year contract with a private contractor "to sniff out tax cheats by mining data from tax returns, bank reports, property records and even social media posts"); Justin Rohrlich, *The Taxman Browseth: The IRS Wants to Use Social Media to Catch Tax Cheats*, QUARTZ, (Dec. 26, 2018) (citing INTERNAL REVENUE SERVICE, SOCIAL MEDIA RESEARCH REQUEST, (Dec. 18, 2018), https://www.fbo.gov/index?s=opportunity&mode=form&id=e832fdd38233b84f8ad65b7fea26eeba&tab=core&_cview=0 [https://perma.cc/N9JQ-WK62]), https://qz.com/1507962/the-irs-wants-to-use-facebook-and-instagram-to-catch-tax-evaders/ [https://perma.cc/RFC4-98FC].

29. *See* Butler, *supra* note 19, at 4 (explaining that Compliance Data Warehouse information includes third-party data); Eric Lai, *Been Audited Lately? Blame the IRS's Massive, Superfast Data Warehouse*, COMPUTERWORLD (Mar. 22, 2008, 1:00 AM) (discussing how data from one year's worth of tax returns are fifteen-to-twenty terabytes), https://www.computerworld.com/article/2536160/business-intelligence/been-audited-lately—blame-the-irs-s-massive—superfast-data-warehouse.html [https://perma.cc/VDE2-5WC6].

30. *See* Lai, *supra* note 29 (stating that as of 2008, the Compliance Data Warehouse had 150 terabytes of information).

31. *See* INTERNAL REVENUE SERV., STRATEGIC PLAN FY 2018-2022, at 19 (2018) (stating that data volume increased 100 times from 2007 to 2017), https://www.irs.gov/pub/irs-pdf/p3744.pdf [https://perma.cc/4HYL-URXU].

32. *See, e.g.*, OFFICE OF INSPECTOR GEN., DEP'T OF THE TREASURY, AUDIT REPORT, TERRORIST FINANCING/MONEY LAUNDERING: FINCEN'S BSA IT MODERNIZATION PROGRAM IS MEETING MILESTONES, BUT OVERSIGHT REMAINS CRUCIAL 1 (2012),

volume of data the IRS has collected, and from what specific sources, is not readily available.

The Tax-Exempt and Government Entities Division (TE/GE) of the IRS is somewhat late to the party, although there were indications as early as 2003 that it was considering more rigorous use of data.[33] This is not to say that the IRS has not previously sought to collect and computerize data regarding tax-exempt nonprofit organizations. As early as the 1970s, the IRS had already begun to enter certain information into its computer systems from annual returns filed by the financially larger tax-exempt organizations.[34] But the 2014 Program Letter appears to be the first indication that this part of the IRS was taking advantage of Big Data techniques to analyze its collected data.[35]

Being late has the advantage, however, of benefiting from the experience of the Research, Applied Analytics and Statistics Division (RAAS) of the IRS, which implements Big Data projects for the agency.[36] RAAS is the result of a 2016 merger of two previous IRS offices: the Office of Research, Analysis, and Statistics with the Office of Compliance Analytics.[37] The former has been the main provider of statistics about the federal tax system for at least a decade, while the IRS established the latter several years ago to house the IRS's data analytics activities.[38] As of late 2017, RAAS had approximately 350 employees

---

https://www.treasury.gov/about/organizational-structure/ig/Audit%20Reports%20and%20Testimonies/OIG12077.pdf [https://perma.cc/8T75-L2EG] (discussing how in 1990, Congress created the Treasury's Financial Crimes Enforcement Network ("FinCEN"), which administers the Bank Secrecy Act); Steven A. Bercu, *Toward Universal Surveillance in an Information Age Economy: Can We Handle Treasury's New Police Technology?*, 34 JURIMETRICS J. 383, 394–400 (1994) (outlining a FinCEN expert system that analyzes financial data collected from government, private, and foreign databases).

33. *See* J. Christine Harris, *IRS Will Enlist New EO Units to Complement Division's Work, Say Officials*, 42 EXEMPT ORG. TAX REV. 150 (2003) (discussing how the new Data Analysis Unit will assist IRS Exempt Organizations division).

34. John Copeland, *Financial Data from Form 990 Returns for Exempt Charitable, Religious, and Educational Organizations and Private Foundations*, *in* 1 RESEARCH PAPERS SPONSORED BY THE COMMISSION ON PRIVATE PHILANTHROPY AND PUBLIC NEEDS 143 (1977).

35. *See supra* note 2 and accompanying text.

36. TREASURY INSPECTOR GEN. FOR TAX ADMIN., THE RESEARCH, APPLIED ANALYTICS, AND STATISTICS ORGANIZATION PROJECT MANAGEMENT PRACTICES NEED IMPROVEMENT 1–2 (2018), https://www.treasury.gov/tigta/auditreports/2018reports/201810026fr.pdf [https://perma.cc/846Z-9FB3] [hereinafter TIGTA].

37. *See id.* at 1.

38. *See* TREASURY INSPECTOR GEN. FOR TAX ADMIN., THE OFFICE OF RESEARCH, ANALYSIS, AND STATISTICS NEEDS TO ADDRESS COMPUTER SECURITY WEAKNESSES 1 (2008) (explaining the Office of Research, Analysis, and Statistics' role), https://www.treasury.gov/tigta/auditreports/2008reports/200820176fr.pdf [https://perma.cc/DM5V-

and fifty additional personnel detailed from other agencies or student volunteers, plus approximately 300 contractors.[39]

TE/GE expected to have algorithms developed by RAAS in place by April 1, 2017, to select annual information returns filed by tax-exempt nonprofits for audit.[40] It also planned to run the returns filed by the financially largest tax-exempt organizations and by private foundations, which are a subset of charitable tax-exempt organizations, through that selection process.[41] The most recent TE/GE Work Plan indicates that the IRS has not only accomplished these goals but has also expanded these efforts to include the shorter return that tax-exempt nonprofits with relatively modest assets and annual revenues may use.[42] While the latest program letter from TE/GE does not provide much in terms of specifics, it confirms that TE/GE continues to use "data-driven approaches" and to collaborate with RAAS in doing so.[43] And the acting TE/GE Commissioner recently stated that over 50% of TE/GE's examinations that the division closed in fiscal year 2018 were selected using data-driven approaches.[44]

As these plans indicate, the IRS's primary source of information relating to tax-exempt nonprofits are their annual information returns—the Form 990 series—although an IRS official has stated the IRS is also

---

A9YD]; *Nonprofits Decry Wasteful Government Programs*, 55 NO. 6 GOV'T CONTRACTOR ¶ 43 (2013) (establishing Office of Compliance Analytics).

39. TIGTA, *supra* note 36, at 1.

40. Alexander Reid, Outline for News from the IRS and Treasury (Jan. 6, 2017), *in* EO TAX J. 2017-4 (2017).

41. *Id.* at 180.

42. *See* TAX EXEMPT AND GOV'T ENTITIES DIV., FY 2018 WORK PLAN 7 (2017) [hereinafter TE/GE, 2018], https://www.irs.gov/pub/irs-tege/tege_fy2018_work_plan.pdf [https://perma.cc/Q8H7-9MX2] (discussing compliance models developed for Form 990, Form 990-EZ, and Form 990-PF); *Form 990 Series Which Forms Do Exempt Organizations File Filing Phase In*, INTERNAL REVENUE SERV. [hereinafter IRS, *Form 990 Series*], https://www.irs.gov/charities-non-profits/form-990-series-which-forms-do-exempt-organizations-file-filing-phase-in [https://perma.cc/MCE2-66Z6].

43. DAVID W. HORTON & ROBERT S. CHOI, INTERNAL REVENUE SERV., TAX EXEMPT & GOV'T ENTITIES DIV., FISCAL YEAR 2019 PROGRAM LETTER 2, 6 (2018), https://www.irs.gov/pub/irs-pdf/p5313.pdf [https://perma.cc/9M3S-EKJV]; *see also* DAVID W. HORTON & ROBERT S. CHOI, INTERNAL REVENUE SERV., TAX EXEMPT & GOV'T ENTITIES DIV., FISCAL YEAR 2018 ACCOMPLISHMENTS LETTER 5 (2019), https://www.irs.gov/pub/irs-pdf/p5329.pdf [https://perma.cc/EW85-WFD7] (describing "Data-Driven Approaches" used in the most recently completed fiscal year).

44. Margaret Von Lienen, Acting Dir., Exempt Orgs. Div., Internal Revenue Serv., Remarks at the Meeting of the TEGE Exempt Organizations Council (Dec. 11, 2018), *in* EO TAX J. 2018-248 (2018); Wesley Elmore, *TE/GE Eyeing Changes to Exam Process*, TAX NOTES (Dec. 10, 2018), https://www.taxnotes.com/tax-notes-today-federal/audits/te-ge-eyeing-changes-exam-process/2018/12/07/28npd [https://perma.cc/K4KA-SURQ].

using other, unspecified "external data."[45] With the exception of certain religious, governmental, and political organizations, as well as organizations included in group returns that aggregate information for a set of related nonprofits, all tax-exempt nonprofits are required to file a Form 990 series return.[46] These organizations filed more than 1.5 million Form 990 series returns annually in recent years.[47] The breakdown by type of annual return for fiscal year 2015 shows approximately 260,000 Forms 990 and 125,000 shorter Forms 990-EZ filed by the most common types of these organizations, and another approximately 100,000 Forms 990-PF filed by private foundations.[48] A variety of other annual return types comprise the remaining million or so returns filed each year, but the majority of them are likely the very short, ("e-filed") Form 990-N that can be used by financially small tax-exempt nonprofits. For example, more than 600,000 nonprofits filed the Form 990-N for fiscal year 2017.[49] While most tax-exempt nonprofits are also required to file an initial application for recognition of their exempt status, these applications are likely much less useful for purposes of detecting non-compliance because they usually represent uncertain predictions by the

---

45. *See* Sunita Lough, Comm'r of the Tax Exempt & Gov't Entities Div., Internal Revenue Serv., Ernst & Young Webcast (Oct. 11, 2017), *in* EO TAX J. 2017-219 (2017); U.S. GOV'T ACCOUNTABILITY OFF., GAO-15-164, TAX-EXEMPT ORGANIZATIONS: BETTER COMPLIANCE INDICATORS AND DATA, AND MORE COLLABORATION WITH STATE REGULATORS WOULD STRENGTHEN OVERSIGHT OF CHARITABLE ORGANIZATIONS 18 (2014) (describing use of Form 990 data to identify organizations for examination).

46. *See* 26 U.S.C. §§ 6033(a)(1), (a)(3) (2012); 26 C.F.R. §§ 1.6033-2(a)(1), (g)(1) (2018); INTERNAL REVENUE SERV., 2018 INSTRUCTIONS FOR FORM 990 RETURN OF ORGANIZATION EXEMPT FROM INCOME TAX 4 (2018) [hereinafter IRS 2018 INSTRUCTIONS FOR FORM 990], https://www.irs.gov/pub/irs-dft/i990—dft.pdf [https://perma.cc/FM98-PYNP].

47. INTERNAL REVENUE SERV. DATA BOOK 2018 4 (2018) [hereinafter INTERNAL REVENUE SERV. DATA BOOK], https://www.irs.gov/pub/irs-soi/18databk.pdf [https://perma.cc/36UL-24XX].

48. *See* INTERNAL REVENUE SERV. STATISTICS OF INCOME DIV., TABLE 1, DOMESTIC PRIVATE FOUNDATIONS: NUMBER AND SELECTED FINANCIAL DATA, BY TYPE OF FOUNDATION AND SIZE OF FAIR MARKET VALUE OF TOTAL ASSETS, TAX YEAR 2015 (2018), https://www.irs.gov/pub/irs-soi/15pf01ta.xls [https://perma.cc/6BY3-97NH] (Forms 990-PF); INTERNAL REVENUE SERV. STATISTICS OF INCOME DIV., TABLE 2, FORM 990 RETURNS OF 501(C)(3)-(9) ORGANIZATIONS: TOTAL FUNCTIONAL EXPENSES, BY CODE SECTION, TAX YEAR 2015 (2018), https://www.irs.gov/pub/irs-soi/15eo02.xlsx [https://perma.cc/D6BL-RQCQ]; INTERNAL REVENUE SERV. STATISTICS OF INCOME DIV., TABLE 4, FORM 990-EZ RETURNS OF 501(C)(3)-(9) ORGANIZATIONS: SELECTED ITEMS, BY CODE SECTION, TAX YEAR 2015 (2018), https://www.irs.gov/pub/irs-soi/15eo04.xlsx [https://perma.cc/H8WB-E68L].

49. *See* IRS, *Form 990 Series*, *supra* note 42 (stating that the Form 990-N filing threshold is no more than $50,000 in annual gross receipts, on average); Natasha M. Cavanaugh et al., *Recommendations Regarding Incentivizing Universal E-Filing for Form 990*, *in* ADVISORY COMMITTEE ON TAX EXEMPT AND GOVERNMENT ENTITIES 2018 REPORT OF RECOMMENDATIONS 37, 42 (2018), https://www.irs.gov/pub/irs-pdf/p4344.pdf [https://perma.cc/EN49-46WY] (stating there were more than 600,000 Forms 990-N filed in fiscal year 2017).

applicants regarding their future plans, rather than reporting on their already completed financial transactions and other activities, as is reflected in the annual information returns. Perhaps for this reason, it is unclear if TE/GE has included information from those applications in the data it is subjecting to Big Data methods.

The contents of these annual information returns represent a substantial amount of data. The Form 990 asks over 200 questions.[50] The form also has sixteen schedules that a tax-exempt nonprofit may need to complete depending on its specific activities.[51] Nevertheless, there are several reasons to question whether those data actually constitute a "Big Data" dataset as the term is commonly used.

First, the volume of data accumulated annually is likely no more than a couple of terabytes, given that the annual amount of data from all federal tax returns—now approaching 250 million, including almost 190 million income tax returns—is in the low double-digits of terabytes.[52] Of course, this does not take into account the possibility that TE/GE is also pulling information from other tax filings—like individual income tax returns of executives at tax-exempt nonprofits—as part of its data collection efforts, although there are no public indications that this is currently the case. Second, the velocity of these data's accumulation is relatively slow because the annual information returns are not due for four and one-half months after the end of the relevant tax year and, by requesting an automatically granted six-month extension, a nonprofit can push that due date until almost a year after the end of the relevant tax year.[53] In addition, while the proportion of Forms 990, 990-EZ, and 990-PF filed electronically has steadily risen in recent years, there are still a significant number that are filed in hard copy. As a result, adding this information to the IRS's electronic database is further delayed.[54] Third, and finally, there is limited variety; most of the data are "structured" in

50. *See* INTERNAL REVENUE SERV., 2018 FORM 990, https://www.irs.gov/pub/irs-pdf/f990.pdf [https://perma.cc/Y2CK-KUZB].

51. *See id.*

52. *See* INTERNAL REVENUE SERV. DATA BOOK, *supra* note 47, at 4 (stating the number of returns filed with the IRS for fiscal years 2016 and 2017); Lai, *supra* note 29 (noting that data from one year's worth of tax returns are fifteen to twenty terabytes).

53. *See* INTERNAL REVENUE SERV., FORM 8868 (2017), https://www.irs.gov/pub/irs-pdf/f8868.pdf [https://perma.cc/KKS9-XS2F]. The extension is not available for nonprofits planning to file a Form 990-N. *Id.* at 2.

54. *See* Cavanaugh et al., *supra* note 49, at 39 (discussing how "roughly 57 percent of all Forms 990 and 990-EZ were filed electronically" for fiscal year 2017); *IRS 990 E-Filer Database*, GITHUB, https://github.com/Nonprofit-Open-Data-Collective/irs-990-efiler-database [https://perma.cc/3QLL-XLA9] (stating that approximately 60,000 Forms 990-PF were filed electronically for calendar year 2016).

that they track the fields used in the 990 series forms, although the forms do have some questions that require more unstructured, narrative responses.[55] And unlike the more general IRS data collection efforts, it is not clear that TE/GE is currently collecting a significant amount of data from sources other than IRS filings by tax-exempt nonprofits.[56]

It also remains unknown whether the IRS is deploying machine learning techniques to analyze the data, as opposed to human-developed and fixed algorithms. That is not to say that machine learning is necessarily required to analyze every dataset effectively, but to the extent Big Data methods might be fruitful, it is uncertain whether the IRS has begun deploying them in this area, buzzwords notwithstanding.[57] By comparison, as early as 2004 the IRS Statistics of Income Division published a research paper describing its use of a "Support Vector Machine" algorithm that learned over time how to better recognize indicators of tax abuse in business and individual returns.[58]

So TE/GE may not yet be truly using a "Big Data" approach to regulate tax-exempt nonprofits. However, given the use of a Big Data approach by the broader IRS, it likely is only a matter of time before TE/GE begins using the same methods—including harvesting significant amounts of data from sources beyond IRS filings and deploying machine learning techniques for analysis purposes. This application would be consistent with the IRS's increasing reliance on Big Data, broadly; one commentator predicts that over the next twenty-five years, the IRS likely will substantially increase its gathering of information from sources beyond IRS filings in order to reduce the compliance burden on taxpayers and the compliance gap in terms of owed but unpaid taxes.[59] Other commentators have noted that there are various indications the IRS is already gathering outside data, including the agency's purchase of cell phone tracking technology,[60] its disclosure in response to a Freedom of Information Act (FOIA) request that it had been reading taxpayers' private emails without a warrant,[61] and its disclosure in response to

---

55. *See supra* notes 46 and 50.

56. *See supra* note 28 and accompanying text (discussing general IRS external data collection).

57. *See, e.g.*, HORTON & CHOI, *supra* note 43, at 5 (mentioning only "compliance models," risk identification, and collaboration with RAAS).

58. DeBarr & Harwood, *supra* note 27, at 178.

59. Michael Hatfield, *Taxation and Surveillance: An Agenda*, 17 YALE J.L. & TECH. 319, 340 (2015).

60. Houser & Sanders, *supra* note 28, at 822.

61. *Id.* at 823 (noting that the IRS apparently ended this practice after congressional scrutiny).

another FOIA request that it collects information from social media sites such as Facebook using automated "spiders."[62] Given sufficient time and resources, these same techniques could be applied to tax-exempt nonprofits, and also to individuals and other entities associated with them. While tax-exempt nonprofits generally do not owe income taxes, in exchange for the tax benefits they enjoy they may be subject to a number of legal restrictions that the IRS is required to enforce, including a reasonableness requirement for financial transactions with insiders, a more general prohibition on undue private benefit, and limitations on lobbying and election-related activities.[63]

Furthermore, Congress recently enacted legislation that mandates electronic filing for all required annual returns.[64] Previously only relatively large tax-exempt nonprofits were required to file the Form 990 electronically, although the Form 990-N filed by relatively small tax-exempt nonprofits must be filed electronically.[65] A broader e-filing requirement will marginally increase the velocity of the data from annual returns. It also will increase public access to this information, as detailed later in this Part. There has been significant government and nonprofit support for this broader requirement, including from the Advisory Committee for TE/GE.[66]

---

62. *Id.* at 823–24.

63. *See* 26 U.S.C. § 170(c)(2) (2012) (regarding conditions relating to the ability to receive tax-deductible charitable contributions), 26 U.S.C. § 501(c) (regarding conditions relating to exemption from income taxes), 26 U.S.C. § 4911–4912 (regarding excise taxes imposed because of excessive lobbying by charitable, tax-exempt nonprofits), § 4940–4946 (regarding restrictions on private foundations, a subset of charitable, tax-exempt nonprofits), § 4955 (regarding excise taxes imposed because of election-related activities by charitable, tax-exempt nonprofits), § 4958 (regarding excise taxes on "excess benefit transactions" between certain tax-exempt nonprofits and insiders).

64. Taxpayer First Act of 2019, H.R. 3151, 116th Cong. § 3101 (2019). Previous, failed attempts to require electronic filing included Retirement, Savings, and Other Tax Relief Act of 2018, H.R. 88, 115th Cong. § 3101 (Rules Committee Print 115-85, Text of the House Amendment to the Senate Amendment to H.R. 88, 2018); H.R. 5443, 115th Cong. § 1 (2018); H.R. 1, 114th Cong. § 6004 (2014).

65. *See* Cavanaugh et al., *supra* note 49, at 42 (stating total assets of $10 million or more and that file at least 250 returns, such as wage reporting Form W-2, annually).

66. *See id.* at 39 (Advisory Committee support); Lloyd H. Mayer, *"The Better Part of Valour is Discretion": Should the IRS Change or Surrender Its Oversight of Tax-Exempt Organizations?*, 7 COLUM. J. TAX L. 80, 111–12, 111 n.212 (2016) (adding other support); U.S. GOV'T ACCOUNTABILITY OFFICE, *supra* note 45, at 34–36, 41 (recommending Congress consider expanding required e-filing for tax-exempt organizations).

B. State Regulators

In addition to the oversight provided by the IRS, states also oversee the operations of certain nonprofits.[67] More specifically, most states require certain "charitable" nonprofits to register and file regular reports with a state agency if they hold assets for charitable use or solicit donations from state residents.[68] State law usually defines "charitable" broadly so as to encompass not only nonprofits that qualify as charities under federal tax law but also some other types of nonprofits.[69] Each state has its own specific requirements for such filings and its own processes for handling them—including whether electronic filing is required or even available.[70] There is relatively limited information sharing between the states and between the states and the IRS, although the states with filing requirements generally require filers to submit their IRS annual information returns to the state.[71] Furthermore, in almost half the states, authority over charitable nonprofits is split between two state agencies.[72] Local jurisdictions may also have their own registration and reporting requirements for charitable nonprofits that solicit donations from their residents.[73] Finally, the tax treatment of nonprofits—for state

---

67. *See* MARION FREMONT-SMITH, GOVERNING NONPROFIT ORGANIZATIONS: FEDERAL AND STATE LAW AND REGULATION 301–02 (2004); Lloyd H. Mayer, *Fragmented Oversight of Nonprofits in the United States: Does it Work? Can it Work?*, 91 CHI.-KENT L. REV. 937, 938–39 (2016).

68. *See* FREMONT-SMITH, *supra* note 67, at 305–06, 315, 372–73 (discussing charitable assets, registration and reporting, and charitable solicitation, respectively).

69. For an example of the breadth of state law definitions of "charitable," see OHIO ATT'Y GEN., HANDBOOK FOR NONPROFITS: AN OPERATIONAL RESOURCE FOR BOARD MEMBERS OF CHARITABLE ORGANIZATIONS 8 (2015), https://ohio.gov/wps/wcm/connect/gov/c1965cd1-7b72-4c57-a090f88b9f257a2d/AGO_NonprofitHandbook.pdf?MOD=AJPERES&CONVERT_TO=url&CACHEID=ROOTWORKSPACE.Z18_M1HGGIK0N0JO00QO9DDDDM3000-c1965cd1-7b72-4c57-a090-f88b9f257a2d-mm18ZE3 [https://perma.cc/HD5X-VPFN] (stating that nonprofits tax-exempt under both Internal Revenue Code section 501(c)(3) and section 501(c)(4) are generally required to register under the Ohio Charitable Trust Act).

70. CINDY M. LOTT ET AL., URBAN INSTITUTE, STATE REGULATION AND ENFORCEMENT IN THE CHARITABLE SECTOR 14, 26 (2016), https://www.urban.org/research/publication/state-regulation-and-enforcement-charitable-sector [https://perma.cc/GEK4-JLUP].

71. *Id.*

72. CINDY M. LOTT ET AL., URBAN INSTITUTE, BIFURCATION OF STATE REGULATION OF CHARITIES: DIVIDED REGULATORY AUTHORITY OVER CHARITIES AND ITS IMPACT ON CHARITABLE SOLICITATION LAWS 1 (2018), https://www.urban.org/research/publication/bifurcation-state-regulation-charities [https://perma.cc/2QSY-VPHA].

73. *See, e.g.*, CALIFORNIA ATT'Y GEN., GUIDE FOR CHARITIES: BEST PRACTICES FOR NONPROFITS THAT OPERATE OR FUNDRAISE IN CALIFORNIA 73 (2019), https://oag.ca.gov/sites/all/files/agweb/pdfs/charities/publications/guide_for_charities.pdf [https://perma.cc/W44U-SPTP] ("Many local jurisdictions in California also require charitable organizations to obtain a license or permit before soliciting donations from residents and in public spaces."); Joseph W. Mead, *Local Regulation of Charitable Solicitation*, 5 J. PUB. & NONPROFIT

income taxes and also state sales tax and local property taxes—is usually handled by additional state or local agencies.[74]

According to a recent report from the Urban Institute, approximately half the states maintain a database of charitable nonprofits that register and report, which may include information from independent audited financial statements when state law requires the filing of those statements.[75] In general, the more staff a state dedicates to oversight of charitable nonprofits the greater the likelihood that a state agency will maintain a database of registered nonprofits, whether only for internal use or also for public use.[76] However, the use by regulators of these databases appears to be limited to simple searches. For example, an official from California recently remarked that her office uses an algorithm to identify every charitable nonprofit that reports a self-dealing transaction on its state filings.[77] This is a very basic search as it only requires determining which organizations have answered "yes" to a single question on the required annual report. As is the case with the IRS, state regulators are interested in ensuring that nonprofits, particularly charitable nonprofits, comply with their legal obligations. Under state law, those obligations include: ensuring assets remain dedicated to the nonprofit's stated purposes; limiting transactions with insiders to reasonable amounts; complying with registration, reporting, and other requirements associated with charitable solicitation; and, for nonprofit leaders, otherwise fulfilling their fiduciary duties.[78]

State regulators hope, however, to increase their collection of data regarding charitable nonprofits and their ability to use those data. The primary vehicle for these efforts is the "Single Portal Initiative," which aims to create a single website where nonprofits can register with and report to multiple jurisdictions, reducing the compliance burden on the organizations and facilitating access to the registration and reporting

---

AFF. 178 (2019) (surveying charitable solicitation regulations imposed by 49 of the 50 largest cities in the United States).

74. *See* FREMONT-SMITH, *supra* note 67, at 368–70; FRANCES R. HILL & BARBARA L. KIRSCHTEN, FEDERAL AND STATE TAXATION OF EXEMPT ORGANIZATIONS, at 14-4, 14-5 (1994 & Supp. 1996); Evelyn Brody, *All Charities are Property-Tax Exempt, but Some Charities are More Exempt than Others*, 44 NEW ENG. L. REV. 621, 625, 671–732 (2010); Mark J. Cowan, *Nonprofit and the Sales and Use Tax*, 9 FLA. TAX REV. 1077, 1205–45 (2010).

75. LOTT ET AL., *supra* note 70, at 14–15.

76. *Id.* at 25–26.

77. *See* CAL. CORP. CODE § 5233 (2018) (defining self-dealing transaction); Tania Ibanez, Remarks at the Annual Meeting of the TE/GE Joint Councils (February 22, 2018), *in* EO TAX J. 2018-97 (2018).

78. *See* FREMONT-SMITH, *supra* note 67, at 305–06, 370; LOTT ET AL., *supra* note 70, at 1–2.

information for all of the states involved.[79] While this project has been on the drawing board for many years, it recently gained new momentum with the retention of two private vendors to implement the project, and a recent announcement included plans to fully launch it by the end of 2018.[80] However, as of January 2019, it was only active in a pilot phase for two states.[81] State officials also hope to take advantage of increased access to newly released IRS electronic data regarding tax-exempt nonprofits, as detailed below.[82]

C. Nonprofits and the Public

There is an important aspect of the information collected by government regulators regarding nonprofits: theoretically, information relating to nonprofits is also available to nonprofits generally and the public. Furthermore, in practice this information is increasingly available in an easily accessible manner thanks to the efforts of various private actors. This access enables not only potential donors, members, customers, volunteers, and employees to learn more about the particular nonprofits of interest to them, and nonprofits to learn more about other nonprofits, but also permits academic researchers and journalists to access and analyze information about these organizations. In this respect, this information is similar to the data collected by some other federal agencies, such as the Environmental Protection Agency, the Federal Election Commission, and the Securities and Exchange Commission, in that they are available to the stakeholders, the public, and researchers.[83]

The rationale for public disclosure of information related to tax-exempt nonprofit organizations is that it enables the public to ensure that

---

79. *See* MRFP, INC., *supra* note 11.

80. *See* GUIDESTAR, Press Release, *supra* note 11.

81. *See* MFRP ET AL., SINGLEPORTAL.ORG: STATE CHARITY REGISTRATION PORTAL, https://www.nasconet.org/wp-content/uploads/2018/09/SinglePortal_OneSheets-Organizer_v1.pdf [https://perma.cc/QKN9-QT2U].

82. Cindy M. Lott, Remarks at the Annual Meeting of the TE/GE Joint Councils (Feb. 22, 2018), *in* EO TAX J. 2018-96 (2018).

83. *See* METHODS, MODELS, TOOLS, AND DATABASES, ENVIRONMENTAL PROTECTION AGENCY, https://www.epa.gov/research/methods-models-tools-and-databases [https://perma.cc/QHP2-4HGT]; DISCLOSURE DATA CATALOG, FEDERAL ELECTION COMMISSION, https://classic.fec.gov/data/DataCatalog.do [https://perma.cc/V695-L52G]; EDGAR, SEARCH TOOLS, SECURITIES & EXCHANGE COMMISSION, https://www.sec.gov/edgar/searchedgar/webusers.htm [https://perma.cc/9FV4-FWVW]; Rachael E. Salcido, *Reviving the Environmental Justice Agenda*, 91 CHI.-KENT L. REV. 115, 128 (2016) (discussing data made available by the EPA).

these nonprofits deserve the significant tax benefits that they enjoy.[84] A related rationale is that by claiming benefits, nonprofits voluntarily surrender any right to privacy they may have pertaining to information they must then provide to the IRS.[85] Similar considerations presumably are the basis for required public disclosure at the state level relating to charitable nonprofits.[86] While concerns over privacy and potential misuse of disclosed information argue against disclosure, in the context of nonprofits, commentators have found them easily outweighed by these rationales, except possibly when they involve information relating to particular individuals.[87]

More specifically, federal law requires both the IRS and tax-exempt nonprofits to allow any member of the public to access applications for recognition of exemption and recent annual information returns.[88] Requesting information on an organization-by-organization basis may be sufficient for an individual donor interested in a limited number of nonprofits or a journalist working on a story relating to a specific group. However, an organization-by-organization approach would not be very practical for a researcher attempting to develop a broader database of information about such entities. On its website, the IRS provides a search tool to confirm the tax-exempt and charitable, if applicable, status of a given nonprofit and makes certain collected information available for downloading but with limited ease of use.[89]

Several private organizations have successfully worked with—and in one case filed a lawsuit against—the IRS to gain broader access to this information and, to the extent possible, in machine-readable format.[90]

---

84. JOINT COMM. ON TAXATION, STUDY OF PRESENT-LAW TAXPAYER CONFIDENTIALITY AND DISCLOSURE PROVISIONS AS REQUIRED BY SECTION 3802 OF THE INTERNAL REVENUE SERVICE RESTRUCTURING AND REFORM ACT OF 1998, VOLUME II: STUDY OF DISCLOSURE PROVISIONS RELATING TO TAX-EXEMPT ORGANIZATIONS 5–6, 63 (2000), https://www.jct.gov/publications.html?func=startdown&id=2559 [https://perma.cc/JN2P-RJQ7]. However, it remains unclear whether public disclosure actually increases legal compliance by nonprofits. *See* Lloyd H. Mayer, *Nonprofits, Politics, and Privacy*, 62 CASE W. RES. L. REV. 801, 822 n.97 (2012).

85. Mayer, *supra* note 84, at 821.

86. *See* Hugh R. Jones, *The Importance of Transparency in the Governmental Regulation of the Nonprofit Sector: Room for Improvement?*, COLUMBIA | ACAD. COMMONS 2–3, 10–11 (2013), https://academiccommons.columbia.edu/doi/10.7916/D86H4FG7 [https://perma.cc/L8ZM-FULK].

87. *See* JOINT COMM. ON TAXATION, *supra* note 84, at 62–64; Mayer, *supra* note 84, at 821–23, 828–29.

88. *See* 26 U.S.C. § 6104 (2012); 26 C.F.R. §§ 301.6104(a)-1,(d)-1 (2018).

89. *See* INTERNAL REVENUE SERV., SOI TAX STATS – CHARITIES & OTHER TAX-EXEMPT ORGANIZATIONS STATISTICS, https://www.irs.gov/statistics/soi-tax-stats-charities-and-other-tax-exempt-organizations-statistics [https://perma.cc/8XKM-XPZX]; INTERNAL REVENUE SERV., TAX EXEMPT ORGANIZATION SEARCH, https://www.irs.gov/charities-non-profits/tax-exempt-organization-search [https://perma.cc/4GD5-ZRSY].

90. Machine-readable simply means that that the data is in a format that can be automatically read and processed by a computer. *See* OPEN DATA HANDBOOK, GLOSSARY: MACHINE READABLE,

These efforts are part of the larger "open data" movement, which strives to get governments to release information proactively in order to facilitate accountability, transparency, and private-public collaboration.[91]

The first group providing broader access was the National Center for Charitable Statistics (NCCS), originally housed in the national nonprofit umbrella organization Independent Sector in the 1980s, but then transferred in the mid-1990s to its current home at the Urban Institute, itself a charitable nonprofit.[92] NCCS has developed and made publicly available a number of nonprofit databases, based in large part of information obtained from the IRS. NCCS manages collections made available by the IRS: the IRS Exempt Organization Business Master Files contains basic information for all active and filing tax-exempt nonprofits for 1989 and 1995–2016, and the IRS Statistics of Income Sample Data Files contains information from a sample of Forms 990 and 990-EZ for 1982–1983 and 1985–2012.[93] Additionally, NCCS has developed databases containing selected information from all Forms 990, 990-EZ, and 990-PF for 1989–2015.[94] Working with the IRS, NCCS also developed a categorization system for nonprofits known as the National Taxonomy of Exempt Entities (NTEE).[95] NCCS designed its databases primarily by use for researchers, although it is in the process of developing a new data platform that will provide easier access for non-experts.[96]

The second group to come on the scene, but likely the most well-known, is GuideStar, founded in the mid-1990s and itself a charitable

---

http://opendatahandbook.org/glossary/en/terms/machine-readable/ [https://perma.cc/9FKZ-B7MH]. For information about the lawsuit, see *infra* note 107 and accompanying text.

91. *See generally* Beth S. Noveck, *Rights-Based and Tech-Driven: Open Data, Freedom of Information, and the Future of Government Transparency*, 19 YALE HUM. RTS. & DEV. L.J. 1 (2017) (describing the open data movement).

92. *See* WILLIAM G. BOWEN ET AL., THE CHARITABLE NONPROFITS: AN ANALYSIS OF INSTITUTIONAL DYNAMICS AND CHARACTERISTICS 32 n.2 (1994); *About Us*, INDEPENDENTSECTOR.ORG, https://independentsector.org/about-us/ [https://perma.cc/4B36-TUZG] (describing Independent Sector as "the only national membership organization that brings together the charitable community . . . to advance the common good"); *Featured Projects*, NCCS DATA ARCHIVE, https://nccs.urban.org/ [https://perma.cc/KQW9-EUW2] (identifying NCCS as now part of the Urban Institute).

93. *National Center for Charitable Statistics Data Archive*, NCCS DATA ARCHIVE, https://nccs-data.urban.org/index.php [https://perma.cc/9RA5-3wx4].

94. *Id.*

95. *National Taxonomy of Exempt Entities (NTEE) Codes*, NCCS DATA ARCHIVE, http://nccs.urban.org/classification/national-taxonomy-exempt-entities [https://perma.cc/P3HC-59DB].

96. *Nonprofit Data at Your Fingertips*, NCCS DATA ARCHIVE, http://nccs.urban.org/ [https://perma.cc/C5XK-7RNS].

nonprofit.[97] Unlike NCCS, its target audience is the public generally, including nonprofit leaders, donors, and journalists; like NCCS it also relies primarily on information obtained from the IRS, although it also gathers additional information from the nonprofits themselves and other sources.[98] GuideStar has developed a reputation as a source for information about nonprofits because of the breadth of its coverage and its user friendly website.[99] GuideStar also receives approximately 350 variables from Form 990 financial data from the NCCS.[100] It only recently announced that it will merge with the Foundation Center, which collects information on grants, to further enhance its database.[101] In order to finance its operations, however, it requires payment for some search and analysis tools, downloading options, data services, reports, and older information.[102]

The most recent entry into this space is ProPublica, a charitable nonprofit that focuses on investigative journalism.[103] It launched its Nonprofit Explorer website earlier this decade and has continually updated it since that time.[104] The website provides a search engine to access IRS information, including Form 990s, for any given tax-exempt nonprofit beginning with 2013, along with certain federal grant audit

---

97. *GuideStar: A Brief History*, GUIDESTAR, https://learn.guidestar.org/about-us/history [https://perma.cc/W86V-92YQ].

98. *See Hi! We're GuideStar*, GUIDESTAR, https://learn.guidestar.org/about-us [https://perma.cc/5HCP-G8ER] (data sources described under "What We Do").

99. *See GuideStar: A Brief History*, GUIDESTAR, *supra* note 97 (noting that Forms 990 and 990-EZ were first posted in 1999; Form 990-PF in 2000; and the database was expanded to include all tax-exempt organizations in 2005).

100. *Getting Started with NCCS Data*, NCCS DATA ARCHIVE, http://nccs.urban.org/database/overview-nccs-data-files [https://perma.cc/5AE8-MW9Y] (describing NCCS-GuideStar National Nonprofit Research Database); GUIDE TO USING NCCS DATA, NAT'L CTR. FOR CHARITABLE STATISTICS 3, https://nccs-data.urban.org/NCCS-data-guide.pdf [https://perma.cc/XKP7-VQZS].

101. *See* Marc Gunther, *2 Information Giants, Foundation Center and GuideStar, to Join Forces*, THE CHRONICLE OF PHILANTHROPY (Feb. 5, 2019), https://philanthropy.com/article/2-Information-Giants/245608 [https://perma.cc/N7Z5-EKZS]; Press Release, Foundation Ctr., Foundation Center and GuideStar Join Forces to Become a New Nonprofit Entity Named Candid (Feb. 5, 2019), https://foundationcenter.org/about-us/press-room/archive/foundation-center-and-guidestar-join-forces-to-become-a-new-nonprofit-entity-named-candid [https://perma.cc/ED32-68XU].

102. *See GuideStar Products & Services*, GUIDESTAR, https://learn.guidestar.org/products/ [https://perma.cc/ W86V-92YQ].

103. *See The Mission, About Us*, PROPUBLICA, https://www.propublica.org/about/ [https://perma.cc/QSF7-XYGK].

104. *See* Mike Tigas, *Announcing the Nonprofit Explorer API*, PROPUBLICA (Sept. 26, 2013, 10:20 AM), https://www.propublica.org/nerds/announcing-the-nonprofit-explorer-api [https://perma.cc/X9DR-WM27].

information since 2016.[105] It also provides an application-programming interface that allows more robust searches.[106]

This machine-readable data became public in 2017 when an open data promoting organization challenged the IRS.[107] While the organization only sought access to machine-readable data from the annual information returns filed by certain tax-exempt nonprofits, in the wake of a federal trial court ruling against it, the IRS decided to publicly release the machine-readable data for all e-filed annual information returns from 2011 forward as an Amazon Web Services public dataset.[108] Congress also recently codified a requirement that the IRS release these data in machine-readable form.[109] Finally, there have been some efforts to create more specialized databases of nonprofit information. For example, a Vermont nonprofit launched a database focused on nonprofits from that state.[110]

As the introduction to this Part indicates, the mere release of data does not mean that those data are immediately usable by researchers and others. It is taking a concerted effort by a number of academics and private organizations working together as the Nonprofit Open Data Collective to clean up those data in order to make them usable.[111] For

---

105. *See Nonprofit Explorer*, PROPUBLICA, https://projects.propublica.org/nonprofits/ [https://perma.cc/X3D3-J42S] (referencing the "About This Data" section).

106. *Id.*

107. *See* Public Resource.org v. IRS, 78 F. Supp. 3d 1262, 1263 (N.D. Cal. 2015); Alec Glassford, *Nonprofit Explorer Update: Full Text of 1.9 Million Records*, PROPUBLICA (Aug. 10, 2017, 11:38 AM), https://www.propublica.org/nerds/nonprofit-explorer-update-full-text-of-nearly-two-million-records [https://perma.cc/53GG-X3RJ].

108. *See id.*; *IRS 990 Filing Data Now Available as an AWS Public Data Set*, AMAZON WEB SERVICES: AWS GOV'T, EDUC., & NONPROFITS BLOG (June 15, 2016), https://aws.amazon.com/blogs/publicsector/irs-990-filing-data-now-available-as-an-aws-public-data-set/ [https://perma.cc/WS9G-KKPJ]; Suzanne Perry, *IRS Plans to Begin Releasing Electronic Nonprofit Tax Forms Next Year*, THE CHRONICLE OF PHILANTHROPY (June 30, 2015), https://www.philanthropy.com/article/IRS-Plans-to-Begin-Releasing/231265 [https://perma.cc/UU6S-EZFM].

109. *See* Taxpayer First Act of 2019, H.R. 1957, 116th Cong. § 3101(c) (2019).

110. *See* Andrea Suozzo, *Explore the Organizations and People that Power Vermont's $6.8 Billion Nonprofit Economy*, SEVEN DAYS: VERMONT NONPROFIT NAVIGATOR, https://nonprofits.sevendaysvt.com [https://perma.cc/U3UF-3WZG]; Matthew Roy, Seven Days *Launches Vermont Nonprofit Navigator Database*, SEVEN DAYS: OFF MESSAGE (June 28, 2018, 4:00 AM), https://www.sevendaysvt.com/OffMessage/archives/2018/06/28/seven-days-launches-vermont-nonprofit-navigator-database [https://perma.cc/JKB5-F79W].

111. *See How the Nonprofit Open Data Collective Came Together to Work on IRS 990 Data in the Cloud*, AMAZON WEB SERVICES: AWS GOV'T, EDUC., & NONPROFITS BLOG (Oct. 23, 2017), https://aws.amazon.com/blogs/publicsector/how-the-nonprofit-open-data-collective-came-together-to-work-on-irs-990-data-in-the-cloud/ [https://perma.cc/3PEX-69HZ]; Stephanie Román, *The Aspen Institute's Program on Philanthropy and Social Innovation (PSI) Hosts Nonprofit Datathon*, ASPEN

example, the dataset reflected dozens of different electronic formats for the Form 990 over the time period at issue, which had to be reconciled sufficiently to allow for meaningful analysis of those data in the aggregate.[112] The cleaned up data are now available on GitHub, an online platform recently purchased by Microsoft.[113] These data represent a major increase in the computerized data available from the Form 990, since it includes thousands of fields as compared to the hundreds of fields previously digitized by the NCCS.[114] Cleaning work is continuing on the initially released data, and will have to continue indefinitely as the IRS continues to release data from newly filed annual information returns. But these data only reflect e-filed returns, and so currently provide only a partial picture of Form 990, Form 990-EZ, and Form 990-PF filers.[115]

Nonprofits and the public also have access to the state databases for charitable nonprofits in some states, although the extent of access varies.[116] It appears that public access comes with, at most, simple search capabilities that, while perhaps sufficient to locate information regarding a particular charitable nonprofit, do not facilitate consideration of the data in the aggregate.[117] Finally, a variety of private organizations, including watchdog groups such as Charity Navigator and the Better Business Bureau's Wise Giving Alliance and umbrella organizations such as the Foundation Center, have complied their own databases and analyses of selected nonprofits drawing from a range of available

---

INSTITUTE (June 6, 2017), https://www.aspeninstitute.org/blog-posts/aspen-institutes-program-philanthropy-social-innovation-psi-hosts-nonprofit-datathon/ [https://perma.cc/K986-NYT9].

112. Román, *supra* note 111.

113. *See* Nonprofit Open Data Collective, *supra* note 54; *Microsoft completes GitHub acquisition*, OFFICIAL MICROSOFT BLOG (Oct. 26, 2018), https://blogs.microsoft.com/blog/2018/10/26/microsoft-completes-github-acquisition/ [https://perma.cc/ECP5-65Z7].

114. *See supra* notes 93, 100 and accompanying text.

115. *See* Nonprofit Open Data Collective, *supra* note 54 (referencing the "IRS 990 E-Filer Data" section).

116. LOTT ET AL., *supra* note 70, at 25.

117. *See, e.g., Registry Verification Search*, STATE OF CAL. DEP'T OF JUSTICE: OFF. OF THE ATT'Y GEN., http://rct.doj.ca.gov/Verification/Web/Search.aspx?facility=Y [https://perma.cc/C9W7-PKBN] (stating the search tool is designed to obtain information with respect to a particular charitable organization or fundraiser); *Charitable Trust Database Search*, ILL. ATT'Y GEN., http://charitableviewer.ilattorneygeneral.net/ [https://perma.cc/QY4K-8V6U] (stating that the database provides financial information about individual charities); *Welcome to the Charities Bureau Registry Search*, CHARITIESNYS.COM, NEW YORK STATE OFFICE OF THE ATT'Y GEN., https://www.charitiesnys.com/RegistrySearch/search_charities.jsp [https://perma.cc/527N-6CAK] (stating the search tool is designed to search for a specific organization).

sources, but these databases tend to be relatively limited in scope, accessibility, and searchability.[118]

These government-related efforts do not reflect the only potential use of Big Data approaches by nonprofits, of course. For example, the consulting firm M+R offers a variety of Big Data-related services to nonprofits, promoted in part through its annual Benchmarks Study.[119] Its most recent Benchmarks Study focused on over five billion digital interactions with supporters of 154 participating nonprofits.[120] However, these purely private efforts do not relate directly to government use of Big Data to regulate nonprofits. They also do not make information available to the public, except either in aggregate form or with the consent of the particular nonprofit involved. They are therefore not the focus of this Article.

D. Other Countries

Before considering how trends in the collection and analysis of data regarding nonprofits in the United States may benefit or hurt the oversight of the nonprofit sector here, it is worth noting that in most other countries government regulators are also moving in a similar direction. The Canada Revenue Agency is expected to begin using Big Data methods to analyze its nonprofit database in the near future.[121] Since 2013, the agency has made all of its datasets dating from 2000 forward available through the government's data portal.[122] The Charity Commission for England and Wales similarly provides the public with

---

118. *See, e.g.*, *Which Charity Report Are You Looking For?*, BBB WISE GIVING ALLIANCE, https://www.give.org/ [https://perma.cc/4M5V-PHJ3] (allowing searches by charity name); *Advanced Search*, CHARITY NAVIGATOR, https://www.charitynavigator.org/index.cfm?bay=search.advanced [https://perma.cc/PQ7V-GMHY] (allowing searches by a limited set of criteria); *990 Finder*, FOUNDATION CENTER, https://foundationcenter.org/find-funding/990-finder [https://perma.cc/NDR6-G7KB] (same).

119. *See Services*, M+R, https://www.mrss.com/work/ [https://perma.cc/WDV4-UAFS]; M+R, *The 2019 Benchmarks will be mind-boggling. Join us for the journey!*, M+R BLOG (Oct. 18, 2018), https://www.mrss.com/lab/the-2019-benchmarks-will-be-mind-boggling-join-us-for-the-journey/ [https://perma.cc/YZR2-UWGU].

120. M+R, BENCHMARKS 2018, https://namp.americansforthearts.org/sites/default/files/2018_Benchmarks_Study.pdf [https://perma.cc/S6E7-Y8DX].

121. *See* C. Yvonne Chenier, *CRA and the Champ*, DRAKE APTOWITZER LLP, (Sept. 14, 2017), http://drache.ca/articles/cra-and-the-champ/ [https://perma.cc/844A-69WY].

122. JULIET MCMURREN ET AL., OPEN DATA'S IMPACT: OPENING CANADA'S T3010 CHARITY INFORMATION RETURN DATA: ACCOUNTABILITY OF CHARITIES THROUGH OPEN DATA 3 (2016), http://odimpact.org/case-opening-canadas-t3010-charity-information-return-data.html [https://perma.cc/KZG3-SRE3].

the ability to electronically extract the data from its registry of charities for all registered charities.[123] The Australian Charities and Not-for-Profits Commission recently made available the information submitted by charities in that country on their 2015 Annual Information Statements.[124] The New Zealand Charities Services agency provides a robust search engine and the ability to download data in mass on its website.[125] The charity regulators in Ireland and Scotland also allow public, online access to their charity registers.[126] It appears that access is limited to only the most basic facts regarding registered charities, as opposed to the entirety of the information filed by those charities.[127] But in Ireland, a private partner has stepped up to create a more robust, publicly available database of information for nonprofit organizations in that country, similar to what has happened in the United States.[128] This incomplete list only provides a sample of the data available outside the United States, as there are other private and government databases.[129]

These developments indicate that while the United States may be at or close to the leading edge of the Big Data approach to information regarding nonprofits, it is far from alone in moving in this direction. This is particularly true because, as will be seen, most aspects of both the promises and perils are not country-dependent but apply universally. The one major exception is legal limits, which of course are country-dependent.

---

123. *See Registered Charities in England and Wales: Data for Registered and Removed Charities in England and Wales*, CHARITY COMMISSION, http://data.charitycommission.gov.uk/data-definition.aspx [https://perma.cc/NQG4-U2FF].

124. *See New Charity Data Available on Data.gov.au,* AUSTRALIAN CHARITIES AND NOT-FOR-PROFITS COMMISSION: NEWS (Jan. 17, 2017), https://www.acnc.gov.au/media/news/new-charity-data-available-datagovau [https://perma.cc/8E48-V8A2].

125. *See The Charities Register,* CHARITIES SERVICES, https://www.charities.govt.nz/charities-in-new-zealand/the-charities-register/ [https://perma.cc/T7QZ-UDFF].

126. *See Information for the Public*, CHARITIES REGULATOR, http://www.charitiesregulatoryauthority.ie/en/CRA/Pages/WP16000072 [https://perma.cc/S4B5-S88Z]; *Charity Register Download*, SCOTTISH CHARITY REGULATOR, https://www.oscr.org.uk/about-charities/search-the-register/charity-register-download [https://perma.cc/S85U-64T6].

127. *See supra* note 26.

128. *See What is Benefacts?*, BENEFACTS, https://en.benefacts.ie/about/what-is-benefacts/ [https://perma.cc/QBM2-K57F]; Patricia Quinn, *Benefacts: Building a Data Infrastructure to Support Public Policy for Nonprofits in Ireland* (2018) (unpublished manuscript) (on file with author).

129. *See, e.g.*, THE CHARITY VAULT, https://web.archive.org/web/20180224033949/http://charity-charities.org/ [https://perma.cc/GKS7-N6XN] (referencing a private directory of more than 100,000 organizations around the world); ISRAEL GIVES, https://www.israelgives.org/pages/index [https://perma.cc/QW7B-YSBG] (referencing a private database of "all of Israel's 42,000 non-profits and charities").

II. THE PROMISES OF USING BIG DATA TO REGULATE NONPROFITS

The promises of Big Data are based on the technology-driven growth in the amount of data collected electronically, particularly through the Internet, and the technology-enabled development of advanced analytics to analyze those data in order "to uncover patterns, correlations, anomalies, outliers, and other insights not suggested by *a priori* hypotheses or explicit assumptions" that can then be used to improve actions or decisions.[130] In the context of government regulation, the potential exists both to more efficiently enforce existing laws and to better design rules to achieve desired outcomes.[131] For example, in 2011, New York City launched a new data-based system for determining what buildings to inspect for "illegal conversions"—dividing up a property so it can house many more people that it was designed to do so safely—and overnight inspector visits found unsafe conditions meriting orders to vacate in over 70% of cases as opposed to the previous rate of 13%.[132] In addition, to the extent government collected data are made available to the public, that availability could increase government accountability and transparency, spur private research that aids government goals, and provide improved information to the public.[133] This Part addresses whether and to what extent these benefits might be found in the context of government regulation of nonprofits.

A. Improved Government Oversight

The regulatory goal of the IRS and state agencies is to ensure compliance with the applicable laws in a fair manner. With respect to regulating nonprofits, this goal is easy to state, but difficult to implement.[134] This task is complicated by the very limited resources that

---

130. U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-16-659SP, DATA AND ANALYTICS INNOVATION: EMERGING OPPORTUNITIES AND CHALLENGES 11 (2016).

131. *See* KITCHIN, *supra* note 17, at 115–16; MANYIKA ET AL., *supra* note 5, at 61–62.

132. VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 186–88 (2013).

133. *See supra* note 91 and accompanying text.

134. *See Tax Exempt & Government Entities Division At-a-Glance*, INTERNAL REVENUE SERV., https://www.irs.gov/government-entities/tax-exempt-government-entities-division-at-a-glance [https://perma.cc/4N3Y-2KPQ] ("Our mission is to provide our customers top quality service by helping you understand and comply with applicable tax laws and to protect the public interest by applying the tax law with integrity and fairness to all."); NATIONAL ASSOC. OF ST. CHARITY OFFICIALS, https://www.nasconet.org/ [https://perma.cc/XH8W-NDEL] (an association of "state offices . . . charged with oversight of charitable organizations and charitable solicitation in the United States").

both the IRS and state regulators have available for regulating nonprofits in the United States compared to the number, financial size, and variety of these nonprofits.[135] These limited resources have, in turn, led to a federal audit rate of less than half-a-percent for the annual information returns filed by tax-exempt nonprofits.[136] No data is available regarding the overall level of enforcement activities by state regulators, but anecdotal evidence indicates a similarly low level of enforcement activity.[137]

In its most recent Work Plans, TE/GE has emphasized improving transparency, efficiency, and effectiveness, with the increased reliance on data-driven methods particularly related to the latter two goals.[138] For example, in 2016 TE/GE noted that testing case selection modeling techniques resulted in a change rate of 85% for closed audits, which appears to mean that in 85% of cases, the IRS successfully asserted that the filing organization had to change its activities or reporting.[139] More recent information released by the IRS indicated a change rate of slightly over 80%, which exceeds the IRS's usual target.[140] Increased change rates for audits would indicate improvements to efficiency—the greater the proportion of audits that result in changes, the more efficient the audit selection process is, as agents are not wasting their time and the audited organization's time by checking organizations that are in full compliance with the law. Increased change rates would also indicate improvements to effectiveness, because a higher change rate indicates that fewer legal violations are going undetected and uncorrected.

---

135. *See* LOTT ET AL., *supra* note 70, at 8; Mayer, *supra* note 66, at 83.

136. Mayer, *supra* note 66, at 91.

137. *See* SHIRLEY ADELSTEIN & ELIZABETH T. BORIS, STATE REGULATION OF THE CHARITABLE SECTOR: ENFORCEMENT, OUTREACH, STRUCTURE, AND STAFFING 2 (2018), https://www.urban.org/research/publication/state-regulation-charitable-sector [https://perma.cc/2A2G-KYPF]; Mayer, *supra* note 67, at 940.

138. TE/GE, 2018, *supra* note 42, at 6–8; TAX EXEMPT AND GOVERNMENT ENTITIES (TE/GE), FY 2017 WORK PLAN 3–8 (2016) [hereinafter TE/GE, 2017], https://www.irs.gov/pub/irs-tege/tege_fy2017_work_plan.pdf [https://perma.cc/KM5D-WRD2]. This parallels similar efforts in the broader IRS, driven in significant part by the need to maximize the use of limited resources. *See* Bulusu, *supra* note 28; Sony Kassam, *IRS Catches $10 Billion in Tax Fraud in 2018*, DAILY TAX REP., BLOOMBERG TAX (Nov. 14, 2018, 5:25 PM), https://news.bloombergtax.com/daily-tax-report/irs-catches-10-billion-in-tax-fraud-in-2018-1 [https://perma.cc/J4FF-RC8C] (crediting data analytics for a four-fold increase in the amount of tax fraud discovered).

139. *See* TE/GE, 2017, *supra* note 138, at 6–7. A greater amount of data and a greater ability to use those data could also benefit efforts to design new rules for nonprofits. *See* Ruth Madrigal, Attorney-Advisor, Department of the Treasury, Remarks at the Urban Institute (Dec. 1, 2015), *in* EO TAX J. 2016-6 (Jan. 11, 2016).

140. Von Lienen, *supra* note 44, at 7.

While touted primarily as a means of reducing the compliance burden on charitable nonprofits, the states' Single Portal initiative also holds the promise of improving the efficiency and effectiveness of state oversight by creating an electronic database of information that state regulators can use.[141] The states anticipate that the database will "maximize efficiency, data transparency, and information sharing," including by helping regulators connect Form 990 data and state registration data and by enabling them to use analytics "to better understand charitable resources and solicitations, to better focus law enforcement and fraud prevention resources, and [to] enable better policy making for protection of charitable resources."[142] In other words, improved efficiency and effectiveness. Of course, it remains to be seen whether the promise of better government oversight is actually fulfilled by these technological developments. As for improved accountability and transparency, the regulators in this area are hindered by the need to not unduly tip their hand to potential bad actors.[143] That said, the data they have made available to outside researchers may also lead to improvements in this area.

B. Improved Research

Another often cited benefit of these Big Data moves relating to nonprofit organizations is greater access to information by academic researchers, who presumably in turn will conduct analyses that then aid policy makers and the public.[144] The Single Portal initiative is touted in part because "[a]cademics, policy makers and the public will be able to conduct their own inquiries or download data in machine-readable format."[145] The IRS has not specifically promoted this research benefit for the data it makes available, but the effect of its decision to release in machine-readable format all recent e-filed Forms 990, when combined with the efforts of private actors to clean those data to make them usable by researchers, has now made these data accessible for private research.[146]

---

141. *See* GUIDESTAR, Press Release, *supra* note 11.

142. MRFP, INC., *supra* note 11.

143. *See* Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503, 1512 (2013) ("The IRS maintains full secrecy as to the selective auditing schemes it applies").

144. *See* BETH S. NOVECK & DANIEL L. GOROFF, INFORMATION FOR IMPACT: LIBERATING NONPROFIT SECTOR DATA (2013), https://assets.aspeninstitute.org/content/uploads/files/content/docs/psi/psi_Information-for-Impact.pdf [https://perma.cc/X7X7-6QUX].

145. MRFP, INC., *supra* note 11.

146. *See supra* notes 107–111 and accompanying text.

Recent work by Professors Brian Galle and Terri Helge hints at the potential of greater access to data about nonprofits. In two articles, Galle drew upon the private foundation data compiled by NCSS and other information to provide new analyses of (1) private foundation investment returns and their relationship to the rate of distributions to support current charitable activities, and (2) how differences in state law standing rules correlated with private foundation administrative cost ratios.[147] His research has important ramifications for the law relating to restricted gifts, particularly for the appropriate legally required distribution rate (the "payout" rate) for private foundations, and also for developments with respect to state standing law. Helge comprehensively reviewed all of the letters from IRS denying applicants for recognition of exemption as charities from 2004 through early 2017 to determine the grounds for denials and their relative frequency.[148] Her research is relevant to whether recent changes in the application process may result in a significant number of unqualified organizations nevertheless being recognized by the IRS as tax-exempt charities, and also may help identify the common pitfalls for such organizations when they are first starting out. While less directly applicable to legal compliance, researchers in other fields have also begun to take advantage of the increasing amount of available nonprofit data. For example, Professors Nathan Grasse and Jesse Lecy and their co-authors have used nonprofit data to explore a variety of topics, including government grants to nonprofits and issues relating to nonprofit compensation.[149] Researchers in other countries are also exploring possible uses of nonprofit data newly available in their home countries.[150]

---

147. *See* Brian Galle, *Valuing the Right to Sue: An Empirical Examination of Nonprofit Agency Costs*, 60 J.L. & ECON. 413 (2017); Brian Galle, *Pay It Forward? Law and the Problem of Restricted-Spending Philanthropy*, 93 WASH. U. L. REV. 1143, 1143–44 (2016).

148. Terri L. Helge, *Rejecting Charity: Why the IRS Denies Tax Exemption to 501(c)(3) Applicants*, 14 PITT. TAX REV. 1, 3 (2016).

149. *See, e.g.*, Nathan Grasse et al., *Understanding the Compensation of Nonprofit Executive Directors: Examining the Influence of Performance and Organizational Characteristics*, 24 NONPROFIT MGMT. & LEADERSHIP 377, 378 (2014) (using Form 990 data from the NCCS); Jesse Lecy & Jeremy Thornton, *What Big Data Can Tell Us About Government Awards to the Nonprofit Sector: Using the FAADS*, 45 NONPROFIT & VOLUNTARY SECTOR Q. 1052, 1060–62 (2016) (same).

150. *See, e.g.*, Natasha Cortis, Linking Non-Profit Data Across Government: An Australian Example 2 (July 2018) (unpublished manuscript) (on file with author) (proposing linking data from the Australian Charities Non-for-profits Commission dataset to the Workplace Gender Equality Agency dataset to explore gender equality issues in the charity workforce of Australia); Megan LePere-Schloop et al., NGO Classification From the Bottom-Up: Using Self-Reported Data and Machine Learning to Generate Categories of NGOs in Ghana 4 (July 2018) (unpublished manuscript) (on file with author) (using digitized data from the annual reports of nongovernmental organizations (NGOs) operating in Gahan to develop a database of these NGOs).

Of course, the potential for new research resulting from the greater public access to these government-collected datasets will depend on the level of funding available and the related interest of academics in pursuing this research. But the studies noted above and the existence and support for the efforts of NCSS, GuideStar, ProPublica, and the Nonprofit Open Data Collaborative indicate that both of these critical factors exist to some extent. Indeed, two major foundations, the Charles Stewart Mott Foundation and the Bill & Melinda Gates Foundation, helped support the latter effort.[151]

C. *Improved Information for the Public*

The government-collected datasets also hold the promise of greater information reaching the public in the form of academic research, direct public access, and public dissemination through intermediaries. In terms of direct access at the federal level, the IRS has recently revamped its now-renamed "Tax Exempt Organization Search" function to ease public access to its database for tax-exempt nonprofits and to expand the amount of information available to the public through it, although it has acknowledged some glitches in its implementation.[152] At the state level, the Single Portal initiative will include a search feature to allow public access to the data collected.[153]

Intermediaries who will likely benefit from the increase public access to information include journalists and watchdog organizations. A Wall Street Journal reporter recently tapped into the newly available e-filed annual information returns data to identify all of the charity employees receiving seven-figure annual compensation amounts.[154] Watchdog organizations such as Charity Navigator, MinistryWatch, and Wise Giving Alliance presumably could also benefit from increased access to information about the charities that they evaluate or otherwise provide

---

151. *See Nonprofit Data Project Updates,* ASPEN INST., https://www.aspeninstitute.org/programs/program-on-philanthropy-and-social-innovation-psi/nonprofit-data-project-updates/ [https://perma.cc/UDH9-98QR] (identifying these two foundations as the funders for the project).

152. Press Release, Internal Rev. Serv., New IRS Online Tool Offers Expanded Access to Information on Tax-Exempt Organizations; Newly-Filed Data Available to Public for First Time (May 7, 2018), https://www.irs.gov/newsroom/new-irs-online-tool-offers-expanded-access-to-information-on-tax-exempt-organizations-newly-filed-data-available-to-public-for-first-time [https://perma.cc/Z7EP-2RAY]; Von Lienen, *supra* note 44 ("We had a couple of little glitches with [the Tax-Exempt Organization Search] that have popped up.").

153. MRFP, INC., *supra* note 11.

154. Andrea Fuller, *Million Dollar Paydays Jump for Officials at Charities*, WALL STREET J., Mar. 6, 2017, at A1.

information regarding.[155] For example, while Charity Navigator only rates approximately 10,000 charities, it provides basic information for all charities registered with the IRS; the amount of such basic information available could dramatically increase if Charity Navigator takes advantage of access to e-filed annual information returns.[156]

III. THE PERILS OF USING BIG DATA TO REGULATE NONPROFITS

The promises of applying Big Data methods to nonprofits, both through direct effects on government oversight and indirect effects on academic research and information available to the public, are therefore significant but as yet mostly unrealized. There are, however, a number of significant perils associated with such methods that commentators have identified generally and all of which apply at least to some extent in the context of nonprofits. Some of these concerns arise with respect to greater access to data of all types, while some are particular to the Big Data context. These perils include faulty results stemming from data inaccuracies, analysis inaccuracies, or over-reliance on analysis results. They also include larger concerns, such as invasion of privacy, improper discrimination, increased government power, and violations of constitutional or statutory protections more generally.[157]

A. Getting It Wrong

One of the greatest concerns raised by commentators regarding Big Data is the tendency of some Big Data supporters to oversell what it can accomplish. Chris Anderson, then the Editor-in-Chief of *Wired*

---

155. *See More About Us*, BBB WISE GIVING ALLIANCE, https://www.give.org/about-bbb-wga/more-about-us/ [https://perma.cc/LT85-5JQY]; *About Us*, CHARITY NAVIGATOR, https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=1653 [https://perma.cc/739D-XHLR]; *About MinistryWatch*, MINISTRYWATCH, https://www.ministrywatch.com/about.php [https://perma.cc/7S2J-TWYV].

156. *See* CHARITY NAVIGATOR, *About Us*, *supra* note 155.

157. *See generally* DAVID BOLLIER, THE PROMISE AND PERIL OF BIG DATA (2010), https://assets.aspeninstitute.org/content/uploads/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf [https://perma.cc/M3DV-V4VK] (summary of benefits and concerns arising from Big Data expressed by participants in an Aspen Institute Roundtable on Information Technology); FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION?: UNDERSTANDING THE ISSUES 8–11 (2016), https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf [https://perma.cc/RK6M-STVR] (discussing concerns relating to private company use of Big Data techniques); Nathan Cortez, *Regulation by Database*, 89 COLO. L. REV. 1, 29–37 (2018) (discussing the potential downsides of publicly disclosed, government-compiled databases).

magazine, once argued that the emergence of Big Data was making the scientific method obsolete.[158] But in reality, Big Data methods have numerous limitations relating to the data on which they rely, the choices regarding how to analyze those data, and the interpretations of the results.[159] All of these limitations can be found in the nonprofit area and so users of Big Data methods with respect to nonprofits—whether government regulators, researchers, or nonprofits and the public more generally—have to be conscious of them in order to avoid reaching incorrect conclusions or incorrectly applying results.

1. Data Inaccuracy and Limits: Garbage In, Garbage Out

At least currently, both the IRS and the major private databases rely primarily on information contained in the annual returns filed by tax-exempt nonprofits. However, the return information is not necessarily accurate. Furthermore, the limited reliance on data from other sources means that any inaccuracies may remain undetected absent individualized consideration of a particular nonprofit, whether through an IRS audit or a private party's investigation. And, as already noted, datasets may require careful cleaning—for example, matching annual information return fields over multiple versions of the form over time—to be rendered into a format that is amenable to accurate analysis.[160] Finally, financially smaller tax-exempt nonprofits can file shorter returns that contain significantly less information while a large number of tax-exempt nonprofits—particularly churches and certain church-related entities—are not required to file a return.[161]

While research on the accuracy of annual information returns is limited, the research that exists indicates that inaccuracies are widespread for at least some types of information. In 2000, Professors Karen Froelich, Terry Knoepfle, and Thomas Pollalk compared the annual information returns for a sample of 350 nonprofit organizations

158. Chris Anderson, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, WIRED (June 23, 2008, 12:00 PM), https://www.wired.com/2008/06/pb-theory/ [https://perma.cc/EZ4W-J97E].

159. Boyd & Crawford, *supra* note 4, at 666; D. Brent Edwards Jr., *Big Data? Big Deal: The Inability of Big Data to Escape the Limitations of Impact Evaluations*, NORRAG: BLOG (Mar. 12, 2019), https://www.norrag.org/big-data-big-deal-the-inability-of-big-data-to-escape-the-limitations-of-impact-evaluation-by-d-brent-edwards-jr/ [https://perma.cc/A4GS-3UJL].

160. *See* Nonprofit Open Data Collective, *supra* note 54; *supra* note 111 and accompanying text.

161. *See supra* notes 42, 46 and accompanying text.

with their audited financial statements.[162] They found that disparities between these two types of documents varied depending on the type of information at issue, concluding that "[v]ariables outside the primary focus of nonprofit organizations, such as net rental income and gross profit from sales, and to a lesser extent, management expenses, should be used with caution and interpreted carefully."[163] In 2004, Thomas Pollak at the Urban Institute reported that he had found a substantial portion of tax-exempt nonprofits listed significant contributions ($50,000 or more) but no fundraising costs on their returns, indicating significant underreporting of such costs.[164] A more recent report by the Scripps Howard News Service found that 41% of the almost 38,000 tax-exempt nonprofits that reported at least $1 million in contributions also reported no fundraising costs on their returns.[165] It is therefore likely that fundraising costs are often underreported, particularly given that some watchdog groups have relied heavily on the ratio of programmatic to administration and fundraising costs as a measure of nonprofit effectiveness and some state regulators have also emphasized this ratio.[166]

Such inaccurate reporting may extend to other areas that could draw negative public or regulator attention. In 2015, Yahoo News reported that the National Rifle Association (NRA) had spent millions on election-related activities over a number of years but failed to report any such expenditures on its IRS annual return as required, an error the NRA

---

162. Karen A. Froelich et al., *Financial Measures in Nonprofit Organization Research: Comparing IRS 990 Return and Audited Financial Statement Data*, 29 NONPROFIT & VOLUNTARY SECTOR Q. 232, 232 (2000).

163. *Id.* at 251.

164. MARK A. HAGER ET AL., URBAN INSTITUTE, WHAT WE KNOW ABOUT OVERHEAD COSTS IN THE NONPROFIT SECTOR 1 (2004), https://www.urban.org/sites/default/files/publication/57576/310930-What-We-Know-about-Overhead-Costs-in-the-Nonprofit-Sector.PDF [https://perma.cc/99HT-5LU5].

165. Thomas Hargrove & Waqas Naeem, MONEY FOR NOTHING: WHEN NONPROFITS MISLEAD, SCRIPPS HOWARD NEWS SERVICE 4 (2012), https://www.scribd.com/document/111941936/Money-for-nothing-When-nonprofits-mislead [https://perma.cc/HC7P-VHRY].

166. *See, e.g.*, *Criteria & Methodology*, CHARITY WATCH, https://www.charitywatch.org/about-charitywatch/criteria-methodology/3113/3147 [https://perma.cc/G79V-38J9] (grading charities based on the percent of total expenses a charity spent on its programs and the cost to the charity of bringing in a $100 of cash donations from the public); N.Y. STATE OFFICE OF THE ATT'Y GEN., PENNIES FOR CHARITY: WHERE YOUR MONEY GOES: FUNDRAISING BY PROFESSIONAL FUNDRAISERS 2 (2018), https://www.charitiesnys.com/pdfs/pennies-for-charity-2018.pdf [https://perma.cc/7EZ9-N3P4] (focusing on the dollar amount and percentage of total donations to a given charity that go toward the expense of professional fundraisers).

acknowledged in response.[167] The National Taxpayer Advocate and others have also identified numerous errors on the recently created, streamlined tax exemption application form (Form 1023-EZ) for smaller, charitable nonprofits, which do not bode well for how accurate the annual returns filed by these organizations will be.[168]

One of the potential benefits of Big Data is it can identify possible inaccuracies by networking and comparing multiple datasets, in effect doing on a larger scale and more automatically what the researchers for the 2000 study did with a small sample of nonprofits.[169] It is possible that the launch of the Single Portal initiative, and with it the ability to compare federal information returns with state filings, could take advantage of this potential benefit of Big Data. The IRS has also stated it will begin using unidentified external data sources to help it identify tax-exempt nonprofits at risk of certain legal violations involving prohibited financial or other benefits for insiders or others, presumably in a manner similar to what the IRS already does for taxable individuals and entities.[170] It remains to be seen, however, the extent to which this networking may identify and help correct inaccuracies in those returns. It is of course also possible that networking could inaccurately flag entries that are actually correct. For example, Facebook recently began labeling a broad range of ads as "political"; if the IRS were to use that labeling as an indication of possible prohibited political campaign intervention expenditures by charitable nonprofits that purchase Facebook ads, it could lead to many false positives.[171]

---

167. Alan Berlow, *The NRA's brazen shell game with donations: A Yahoo News investigation*, YAHOO POLITICS (Apr. 21, 2015), https://www.yahoo.com/news/the-nras-brazen-shell-game-with-donations-a-116744915796.html [https://perma.cc/F95W-PH86].

168. *See* Nina E. Olson, *A Problem of the IRS's Own Making—Automatic Revocations of 1023-EZ Exempt Organizations*, TAX PAYER ADVOC. SERV.: NTA BLOG (May 16, 2018) https://taxpayeradvocate.irs.gov/news/nta-blog-automatic-revocation-of-1023EZ-exempt-organizations [https://perma.cc/GQ6J-F89H] (summarizing concerns relating to the Form 1023-EZ, including the relatively high rate of automatic revocation for failure to file required annual returns).

169. *See* Froelich et al., *supra* note 162 and accompanying text.

170. *See* Marcus Owens, Materials for "The Current Climate for Nonprofit Enforcement," *in* EO TAX J. 2017-94 (May 12, 2017); *supra* note 45 and accompanying text (discussing how data regarding federal grants could be used to check whether tax-exempt nonprofits receiving such grants are accurately reporting them); JENNIFER TEEFY, CONG. RESEARCH SERV., R44027, TRACKING FEDERAL AWARDS: USASPENDING.GOV AND OTHER DATA SOURCES (2019).

171. *Compare Requiring Authorization and Labeling for Ads with Political Content*, FACEBOOK BUSINESS, https://www.facebook.com/business/news/requiring-authorization-and-labeling-for-ads-with-political-content [https://perma.cc/S79X-94LG] (identifying as "political" ads relating to "any national legislative issue of public importance"), *with The Restriction of Political Campaign Intervention by Section 501(c)(3) Tax-Exempt Organizations*, IRS, https://www.irs.gov/charities-non-profits/charitable-organizations/the-restriction-of-political-campaign-intervention-by-section-

As for cleaning data, the efforts to render the recently released e-filed information returns data illustrate the challenges in this area.[172] The Aspen Institute's Program on Philanthropy and Social Innovation has now hosted two multiple-day events, involving dozens of experts, to resolve issues with these data, including reconciling the various versions of the returns reflected in the dataset, developing a unified set of standards, and checking numerous variables.[173] Yet their work is far from complete with respect to the initially available dataset, much less subsequent data releases. And, of course, until electronic filing becomes mandatory for all tax-exempt nonprofits, this dataset lacks information from the significant proportion of such organizations that do not file electronically.

While it may be acceptable to have less information on financially smaller organizations since the likely impact of any legal violations by such groups is proportionately less as compared to financially larger organizations there are a large number of tax-exempt nonprofits that are not required to file annual information returns or state registration forms and annual reports. Most notably, this exempt group includes all churches and certain church-related entities, regardless of their financial size or scope of activities, unless they have unrelated trade or businesses.[174] By most estimates there are between 300,000 and 450,000 churches of all faiths in the United States, ranging from tiny neighborhood congregations to enormous megachurches and denominational bodies.[175] While figures are difficult to determine in this area because of the lack of government filings, one recent study estimated these entities had at least $100 billion in annual gross revenue and $600 billion in real property assets.[176] By comparison, for tax year 2015, tax-exempt charities that file Form 990 annual information returns

---

501c3-tax-exempt-organizations [https://perma.cc//APL2-657T] (identifying as "political" only communications or actions relating to candidates for elective public office).

172. *See* Nonprofit Open Data Collective, *supra* note 54; *supra* note 111 and accompanying text.

173. Cinthia S. Ottinger, *Aspen Hosts 990 "Vali-Datathon" as Part of Philanthropy's Data Revolution*, ASPEN INSTITUTE: BLOG (Jan. 17, 2018), https://www.aspeninstitute.org/blog-posts/aspen-hosts-990-vali-datathon-part-philanthropys-data-revolution/ [https://perma.cc/59TH-8RGF].

174. *See supra* note 46 and accompanying text; INTERNAL REVENUE SERV., 2017 INSTRUCTIONS FOR FORM 990T EXEMPT ORGANIZATION BUSINESS INCOME TAX RETURN (AND PROXY TAX UNDER SECTION 6033(E)) 2 (2018), https://www.irs.gov/pub/irs-prior/i990t—2017.pdf [https://perma.cc/VL66-65W9] ("Who Must File" section).

175. *See* Simon G. Brauer, *How Many Congregations Are There? Updating a Survey-Based Estimate*, 56 J. SCI. STUDY RELIGION 438, 439, 445 (2017).

176. Ryan T. Cragun et al., *Research Report: How Secular Humanists (and Everyone Else) Subsidize Religion in the United States*, FREE INQUIRY, June–July 2012, at 39, 41–42.

reported $2.0 trillion in annual revenue and $3.8 trillion in assets, the latter partly offset by $1.5 trillion in liabilities.[177] Any use by government regulators or private parties of Big Data methods will miss this significant part of the nonprofit universe.

Lastly, there is a lot of important information relating to nonprofits that the IRS understandably does not ask for because it is not directly relevant to compliance with the applicable tax laws. Perhaps the most obvious area is information on how effective or impactful a nonprofit is in achieving its stated mission or goals. Even for the largest nonprofits with the broadest range of activities, all that the form requires are brief statements relating to program service accomplishments in the nonprofit's largest program areas.[178] While the Form 990 instructions purport to require specific information, such as clients served, in practice many if not most organizations provide vague, self-serving statements.[179]

2. *Analysis Inaccuracy and Limits: The Ghost in the Machine*

The sophisticated computational tools that can be applied to Big Data datasets, particularly ones that self correct as a result of machine learning, hold out the promise of finding the proverbial needle in the haystack in the form of previously undiscovered patterns and relationships. At the same time, "the information revealed by big data analysis isn't necessarily perfect. Identifying a pattern doesn't establish whether that pattern is significant. Correlation still doesn't equal causation . . . . In big data, as with all data, interpretation is always important."[180]

It is difficult to judge the methods used by TE/GE as part of its new data-driven approach to select annual information returns for audit since TE/GE understandably does not share the details of its algorithms in order to avoid providing a roadmap for intentional wrongdoers to avoid

177. STATISTICS OF INCOME DIVISION, IRS, TBL. 1, FORM 990 RETURNS OF 501(C)(3) ORGANIZATIONS: BALANCE SHEET AND INCOME STATEMENT ITEMS, BY ASSET SIZE, TAX YEAR 2015 (2018), https://www.irs.gov/pub/irs-soi/15eo01.xlsx (last visited Aug. 9, 2019).

178. *See* IRS 2018 INSTRUCTIONS FOR FORM 990, *supra* note 46, at 11–12.

179. *See, e.g.*, President and Fellows of Harvard College, Form 990, at 2 (2017) (listing two specific program service accomplishments described in two, relatively short paragraphs: one "education and other institutional activities" involving "approximately 7,330 undergraduate and 14,460 graduate students," and the other "sponsored activities, including sponsored research" for which the University "incurred $806,228,926 in expenses funded by sponsored awards").

180. EXEC. OFFICE OF THE PRESIDENT, *supra* note 20, at 7.

IRS scrutiny.[181] While the initial report of a high change rate for audits that results from this process is encouraging, without knowing how significant the changes were that were discovered or a rigorous comparison with the change rate resulting from other selection methods, it is impossible to evaluate how effective the new process has been as compared to other methods.[182] That said, the IRS is evaluating its methods; the acting TE/GE Commissioner recently stated that when it evaluated data-driven attempts to identify noncompliance among private foundations, no action resulted in 48% of identified cases, indicating this might not be a fruitful approach.[183] To address this gap further, studies by both the IRS and other internal government offices that are allowed access to the details of such methods are needed, along the lines of the 2015 Government Accountability Office study that highlighted the need to strength internal controls for selection processes then employed by TE/GE for tax-exempt nonprofits.[184]

External researchers could aid in this process by publicly developing and testing their own algorithms and other methods for analyzing the data released by the IRS and the states, and also the data available from other sources regarding nonprofits. In doing so, they need to be especially conscious of the fact that in general these methods reveal correlations but do not demonstrate causation; any correlations revealed represent, at best, highly probable relationships but not absolute certainty. In other words, outside researchers and the IRS need to recognize that Big Data methods may identify when there is a greater likelihood of illegal behavior, but they rarely if ever generate 100% accurate results.[185] This is particularly true when Big Data methods are used to predict future behavior of individuals or organizations, the risks of which were made famous by the short story and movie *The Minority Report*.[186] However, in the nonprofit context, no one is advocating

181. *See* Zarsky, *supra* note 143, at 1512.

182. *See supra* note 139 and accompanying text.

183. Von Lienen, *supra* note 44 (reporting the 48% no change rate as an improvement over the previous 65% no change rate).

184. U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-15-514, IRS EXAMINATION SELECTION: INTERNAL CONTROLS FOR EXEMPT ORGANIZATION SELECTION SHOULD BE STRENGTHENED 52–53 (2015); *see also* TIGTA, *supra* note 36 (evaluating RAAS project management practices).

185. *See* MAYER-SCHÖNBERGER & CUKIER, *supra* note 132, at 50–72 (correlation and Big Data).

186. *See* Philip K. Dick, *The Minority Report*, *in* SELECTED STORIES OF PHILIP K. DICK 227 (2002); THE MINORITY REPORT (20th Century Fox 2002). This field is formally known as "predictive analytics." *See* Zarsky, *supra* note 143, at 1505–06.

letting the computers find a violation of even civil law, much less criminal law, without a human-conducted, case-by-case analysis.[187]

For example, say that an analysis reveals a significant correlation between a nonprofit reporting that it lacks a conflict of interest policy and a nonprofit later reporting an excess benefit transaction.[188] This result would indicate that nonprofits reporting no conflict of interest policy are at greater risk of excess benefit transactions. This result would not indicate, however, that requiring nonprofits to adopt a conflict of interest policy would reduce this risk, because it is unknown whether the absence of this policy leads to the excess benefit transaction or whether there might be a third, unmeasured factor that may cause both the lack of such a policy and an excess benefit transaction. The lack of a conflict of interest policy itself would not be a full proof predictor of a future excess benefit transaction, only an indication of a higher risk for such a transaction. Of course, being selected for an audit is not the equivalent of being found guilty of a civil or criminal violation—that would still require a human-conducted, individualized analysis—but such selection is also not costless to the affected nonprofit or the government, so any supposed correlations should be subject to rigorous testing before being relied upon for audit selection purposes.

3. *Over-Reliance: But the Computer Told Me to Do It*

Big Data methods can lead to excessive claims of accuracy and objectivity that in turn lead to over-reliance on results.[189] Simply because certain results are drawn from Big Data datasets using computerized analytics does not mean they provide an accurate or complete picture of the legal compliance of nonprofits such that other methods for measuring compliance are now obsolete and unreliable in comparison. As noted by David Solove, one problem with reliance on databases "is that such records often fail to tell the entire story."[190] Inaccuracy may

187. But such reliance on computers is starting to occur in other legal contexts; for example, Estonia is in the process of designing a "robot judge" to resolve small claims disputes. *See* Eric Niiler, *Can AI Be a Fair Judge in Court? Estonia Thinks So*, WIRED (Mar. 25, 2019, 7:00 AM), https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/ [https://perma.cc/S44Y-WHFG].

188. *See* 26 U.S.C. § 4958 (2012) (defining "excess benefit" as the provision of a legally prohibited, undue economic benefit to an insider). Taxes are imposed on persons who benefit from an excess benefit transaction and managers who approve such a transaction. *Id.*

189. Boyd & Crawford, *supra* note 4, at 666–68.

190. David J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393, 1424 (2001).

arise from problems with the underlying data or the methods used to analyze it, as already noted.[191] But even when those problems are adequately addressed, the limits on Big Data necessarily mean that it cannot fully replace other methods of discovering legal noncompliance, such as individual complaints, new stories, and whistleblowers. A focus on Big Data also could lead to the IRS and state agencies devoting even fewer resources to providing needed guidance regarding the legal rules applicable to nonprofits, which—particularly at the federal level—has already declined significantly in recent years.[192]

B. Larger Concerns

Beyond inaccuracy and over-reliance concerns, other Big Data issues raised loom quite large. These include risks to privacy, prohibited discriminatory treatment, enhanced government power, and other conflicts with constitutional and statutory limits on government power.

1. Invasion of Privacy: Welcome to the Goldfish Bowl

Arguably, the greatest risk of Big Data analytics—loss of privacy—is inapplicable to nonprofits because they are already required to publicly disclose private information.[193] However, there are at least two privacy-related issues. The first is the inadvertent disclosure of personal or other sensitive information. The second is the unintended consequences of greater access to information that comes with Big Data, including its possible use with other data to provide a more complete profile of individuals and other entities associated with nonprofits.[194]

Privacy "encompasses not only avoiding observation, or keeping one's personal matters and relationships secret, but also the ability to share information selectively but not publicly."[195] The risk of Big Data is both that nonprofits lose the ability to control data about themselves and that no one is ultimately controlling the use of those data since it is

191. *See supra* Part III.

192. *See* Mayer, *supra* note 66, at 93–94.

193. *See* Alfred Ng, *Tech's invasion of our privacy made us more paranoid in 2018*, CNET (Dec. 7, 2018, 5:01 AM), https://www.cnet.com/news/techs-invasion-of-our-privacy-made-us-more-paranoid-in-2018/ [https://perma.cc/Q6KC-5EB3]; *supra* notes 88, 116 and accompanying text (public disclosure of nonprofit government filings).

194. *See* EXEC. OFFICE OF THE PRESIDENT, BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE (2014), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf [https://perma.cc/58EX-B5Q6] (identifying such use as "data fusion" that "occurs when data from different sources are brought into contact and new facts emerge").

195. *Id.* at 2.

readily accessible to anyone for almost any purpose.[196] Much of the recent outrage against Facebook did not involve its use of data regarding its users--which was presumably expected by them--but the use of its data by others, such as Cambridge Analytica, in ways that were unexpected.[197]

As for inadvertent disclosure of personal information, this can happen either because a nonprofit provides such information on its government filings even though not required to do so or because a nonprofit provides required personal information that is not subject to public disclosure but somehow that information is in fact publicly disclosed. The first situation has included listing the home addresses of directors, officers, employees, grantees, or others affiliated in some way with the nonprofit, even though the organization's address may now be provided instead, and even listing the Social Security numbers of such individuals.[198] This of course is an unforced error by the nonprofits involved and therefore could be at least partially addressed by better IRS instructions and public education.[199]

The more troubling situation is when a nonprofit is required to provide certain personal information to the IRS and that information is publicly disclosed even though it is not supposed to be. This situation has arisen for donor information that is not subject to public disclosure for most tax-exempt nonprofits, with public release arising from both IRS and nonprofit errors.[200] Some state authorities have also begun

196. *See* Solove, *supra* note 190, at 1428 ("'[C]ontrol out of control'—a situation where nobody is exercising meaningful control over the information.").

197. Ian Sherr, *Facebook, Cambridge Analytica and data mining: What you need to know*, CNET (Apr. 18, 2018, 5:10 PM), https://www.cnet.com/news/facebook-cambridge-analytica-data-mining-and-trump-what-you-need-to-know/ [https://perma.cc/NX6Q-A58Y].

198. *See* Noelle Barton, *Many Nonprofits Include Social Security Numbers on Public Documents, Study Finds*, CHRON. PHILANTHROPY (Apr. 11, 2012), https://www.philanthropy.com/article/Many-Nonprofit-Documents/156741 [https://perma.cc/R9BU-XWBU]. A study of Form 990 filings from 2001 to 2006 found that more than 132,000 nonprofits had included at least one social security number on that form. *Id.*; Christopher Quay, *Changes, New Schedule to Draft Redesign Form 990 Coming, Official Says*, TAX NOTES TODAY 1 (Nov. 19, 2007).

199. *See, e.g.*, IRS, *supra* note 46, at 2 ("Reminder: Don't Include Social Security Numbers on Publicly Disclosed Forms") (emphasis omitted).

200. *See, e.g.*, Mark J. Fitzgibbons, *Oklahoma Secretary of State "Doxxes" O'Keefe Donors*, THE DAILY CALLER (Dec. 19, 2017, 2:48 AM), http://dailycaller.com/2017/12/19/oklahoma-secretary-of-state-doxxes-okeefe-donors/ [https://perma.cc/J9FC-DF2Q] (reporting that a nonprofit failed to redact donor information when it filed a copy of its IRS return with a state agency that released all nonprofit filings to the public); Tom Metcalf, *A Peek Into Goldman's Black Box Charity Reveals Tech Billionaires*, BLOOMBERG (Mar. 14, 2018, 1:00 AM), https://www.bloomberg.com/news/articles/2018-03-14/peek-into-goldman-s-black-box-charity-reveals-tech-billionaires [https://perma.cc/DA28-LWRC] (reporting that the IRS failed to redact donor information from a filing provided to GuideStar); Sam Stein, *Mitt Romney's PAC Funded*

asking for unredacted copies of the federal schedule that lists donor information; while those authorities usually commit to keep the schedules private, such requests add another possible source of inadvertent disclosures.[201] The few press reports describing such situations indicate that they are rare although they have attracted congressional attention.[202] In fact, the IRS is planning to eliminate the requirement that non-charitable, tax-exempt nonprofits provide donor information to the IRS in part because of these concerns.[203]

The increasing ease of access to return and other information relating to nonprofits raises a subtler, but potentially greater, privacy concern. One aspect of Big Data is the growing ability to connect multiple datasets in order to discover patterns or other information that is not apparent by analyzing only one of the datasets in isolation. Say a nonprofit reports—as it is required to do—its compensation for various officers and other highly paid employees. Most people, at least in the United States, view compensation information as private and only to be shared selectively with others. Senior employees and officers of tax-exempt nonprofits hopefully know that as a consequence of their employer's favorable tax status their compensation will be publicly revealed. But what they might not expect is that a private company could obtain this publicly available data and connect those data with its existing database of potential customer profiles, for which compensation

---

Anti-Gay Marriage Group Under The Radar, HUFFPOST: POL. BLOG (Mar. 30, 2012, 1:59 PM), https://www.huffingtonpost.com/2012/03/30/mitt-romney-gay-marriage_n_1391867.html [https://perma.cc/QL85-78SP] (reporting that donor information became public because of a leaked IRS filing).

201. *See, e.g.*, Am. for Prosperity Found. v. Harris, 809 F.3d 536, 538 (9th Cir. 2015) (discussing the California requirement that charities submit to the Attorney General a copy of the (nonpublic) donor information submitted to the IRS). The Single Portal could help states ensure such information remains private by establishing a uniform process for handling the donor schedules.

202. *See, e.g.*, Letter from Representative Trey Gowdy et al., to David J. Kautter, Assistant Sec'y for Tax Policy (June 27, 2018), https://oversight.house.gov/wp-content/uploads/2018/06/2018-06-27-TG-JJ-MM-to-Kautter-IRS-Schedule-B-Briefing.pdf [https://perma.cc/5H6A-H2RQ] (discussing how members of Congress objected to the requirement that tax-exempt nonprofits report the names and addresses of all substantial contributors to the IRS); *see also* Cavanaugh et al., *supra* note 49, at 50–51 (outlining a TE/GE Advisory Committee recommendation to consider eliminating the required reporting of donor information as part of expanding e-filing).

203. *See* Guidance Under Section 6033 Regarding the Reporting Requirements of Exempt Organizations, 84 Fed. Reg. 47,447, 47,451–52 (proposed Sept. 10, 2019) (to be codified at 26 C.F.R. pt. 1) (proposing regulations that in relevant part would eliminate the requirement that tax-exempt nonprofits report the names and addresses of their contributors on their annual information returns, other than for section 501(c)(3) organizations). The IRS could not eliminate the requirement for charitable, tax-exempt nonprofits because that requirement is statutory. *See* 26 U.S.C. § 6033(b)(5) (2012) (stating how section 501(c)(3) organizations must include in their annual information return "the names and addresses of all substantial contributors").

level may be an important element. Such linkages may actually be harmful to the individuals involved if, for example, their compensation information is used to adjust upward the prices charged to them based on their income. However, using data in this way does not appear to be widespread at this point, and, fortunately, there is little evidence of more nefarious use of such data.[204] Nevertheless, the greater availability of what many would consider sensitive, personal information may raise concerns with those who consider working for a nonprofit.[205]

2. *Prohibited Discrimination: Can Computers Be Racist?*

Another prominent Big Data concern is that it could lead to discrimination based on race, gender, or other prohibited grounds, including inadvertently through the reproduction of existing patterns of discrimination.[206] Countering such discrimination can be tricky, in part, because adjusting the results of Big Data analysis to compensate for such impacts may be technically difficult or may itself raise discrimination concerns.[207] Obvious areas in which discrimination concerns can arise are the criminal justice and employment contexts. For example, a recent study found that ads for STEM jobs were less likely to be shown to women than to men on Facebook as ads targeting women on Facebook are more expensive.[208] But discrimination concerns can

---

204. *See* EXEC. OFFICE OF THE PRESIDENT, BIG DATA AND DIFFERENTIAL PRICING 8–13 (2015), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf [https://perma.cc/9M9K-5TU7].

205. For similar concerns relating to the use of personal data to target political advertisements, *see* Bethany Shiner, *Big Data, Small Law: How Gaps in Regulation are Affecting Political Campaigning Methods and the Need for Fundamental Reform*, Abstract (Oct. 28, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3274212 [https://perma.cc/UCV7-F6MV] ("[T]he law cannot adequately deal with the issues posed by the collection and use of personal data for the design and deployment of targeted social media political campaign advertisements.").

206. *See* EXEC. OFFICE OF THE PRESIDENT, BIG DATA: A REPORT ON ALGORITHMIC SYSTEMS, OPPORTUNITY, AND CIVIL RIGHTS 5–6 (May 2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf [https://perma.cc/MC2E-NDXS]; FTC, *supra* note 157; Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 673–74 (2016); Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy, and It's Not Fair*, 66 STAN. L. REV. ONLINE 35, 35–36 (2013).

207. *See* Barocas & Selbst, *supra* note 206, at 714–15.

208. Anja Lambrecht & Catherine E. Tucker, *Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads*, MARKETING SCI. FRONTIERS (forthcoming) (manuscript at 1), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260 [https://perma.cc/4PXD-CED5].

also arise in many other situations, such as access to credit and higher education.[209]

Consider the recent Tea Party application controversy at the IRS. While political smoke has made discerning the underlying facts difficult, the most innocuous description of what occurred is that IRS employees, in a good faith effort to do their jobs efficiently and effectively, chose certain terms for their Be On the Lookout (BOLO) list that, unintentionally, disparately impacted politically conservative applicants by subjecting them to greater scrutiny than some other applicants.[210] The same risk exists for the use of Big Data to select which tax-exempt nonprofits to audit in that some algorithms might have a disparate impact on certain types of nonprofits. For example, what if a correlation is found between the risk of legal noncompliance and location in certain zip codes, which happen to correspond to low-income areas? Using that correlation to select nonprofits for audit could, unintentionally, disparately impact nonprofits led by members of minority racial or ethnic groups, if groups with such leaders are disproportionately based in low income areas. Another example might be if compensation information for nonprofit executives was used to get around laws prohibiting inquiring about compensation history.[211] These laws are usually justified as seeking to reduce gender disparities in compensation.[212]

Computer scientists are well aware of these possible issues and have engaged in extensive research on possible technical fixes to address them.[213] A number of algorithm developers have made their initial algorithms open source so that anyone can analyze them and detect potential discrimination issues, among other problems.[214] While this may be less possible in the nonprofit context given the enforcement-focused

---

209. *See* EXEC. OFFICE OF THE PRESIDENT, BIG DATA AND DIFFERENTIAL PRICING, *supra* note 206, at 4.

210. *See* Leandra Lederman, *IRS Reform: Politics As Usual?*, 7 COLUM. J. TAX L. 36, 50–52 (2016).

211. *See* Áine Cain et al., *9 places in the US where job candidates may never have to answer the dreaded salary question again*, BUSINESS INSIDER (Apr. 10, 2018, 9:08 AM), https://www.businessinsider.com/places-where-salary-question-banned-us-2017-10 [https://perma.cc/2G6J-C7WS].

212. *Id.*

213. *See* FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY IN MACHINE LEARNING (FAT/ML), SCHOLARSHIP, http://www.fatml.org/resources/relevant-scholarship [https://perma.cc/U3Y6-X4A7].

214. *See, e.g.*, Justin Sherman, *How Open Source Can Fight Algorithmic Bias*, THE STARTUP (Jan. 18, 2018), https://medium.com/swlh/how-open-source-can-fight-algorithmic-bias-836a10e409f9 [https://perma.cc/8ABV-8SW4] (describing the use of open source technology and methodologies to combat data bias).

IRS and state regulator role, the key initial requirement for addressing such concerns is awareness that these issues can arise and a willingness to consider how otherwise acceptable data collection and analysis methods may, unintentionally, result in objectionable disparate impacts. Once identified, the necessary resources must be committed to eliminating such impacts to the degree possible.

3. Increased Government Power: Orwell and Kafka

With new technology comes the risk that governments may deploy that technology to increase their power vis-à-vis the governed, especially if existing limits on governmental power fit poorly with such advances.[215] Big Data, with its mass collection and analysis of information relating to the governed and their organizations, certainly raises such concerns.[216] This concern is particularly pronounced when government makes what Professors Ian Kerr and Jessica Earle label "preemptive predictions;" a well-known example would be placing an individual on the no-fly list because Big Data finds a correlation between her attributes and the likelihood of engaging in terrorist activity on an airplane.[217]

This concern is less pronounced in the nonprofit context, however, for at least two reasons. First, both the IRS and state authorities have authority to open an audit or investigation of any nonprofit in their jurisdiction for any rational reason other than reasons that are constitutionally or otherwise legally prohibited. The one exception for the IRS is churches, for which Congress requires specific findings to commence an inquiry or audit.[218] But since the IRS lacks annual return data for churches the application by the IRS of a Big Data approach to them is not feasible at this point under any conditions.[219] Second, the data collected and analyzed for nonprofits are usually viewed as reasonably related to their compliance with applicable laws, although there are occasional debates on this point with respect to certain specific information collected.[220]

---

215. *See* Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296, 2297–99 (2014).

216. *See* EXEC. OFFICE OF THE PRESIDENT, *supra* note 20, at 22.

217. Ian Kerr & Jessica Earle, *Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy*, 66 STAN. L. REV. ONLINE 65, 67 (2013).

218. *See* 26 U.S.C. § 7611 (2012).

219. *See supra* note 46 and accompanying text.

220. *See, e.g.*, Lloyd H. Mayer & Brendan M. Wilson, *Regulating Charities in the Twenty-First Century: An Institutional Choice Analysis*, 85 CHI.-KENT L. REV. 479, 479–80 & nn.1–3 (2010)

There is, however, one issue in this regard. By asking certain questions on the required annual information return, the IRS may influence nonprofit behavior in a way that goes beyond what the law requires. For example, a little over ten years ago the IRS added return questions asking whether the nonprofit completing the return had certain governance practices, even though federal law does not require those specific practices.[221] Many commentators, including the TE/GE's own Advisory Committee, worried that the IRS asking these questions would put pressure on all nonprofits to adopt these practices even it might be unwise for them to do so.[222] The increasing ability of both the IRS and outside parties to search return information and identify correlations between the answers to these governance-related questions and compliance with federal tax law increases the risk that tax-exempt nonprofits may feel pressured to adopt such practices not because of their inherent value but because the lack of them might be perceived as increasing the risk of IRS audit or adverse public attention. Similarly, Congress has chosen to focus greater attention on charitable, tax-exempt hospitals by requiring them to provide information regarding their provision of certain community benefits and adoption of policies relating to such benefits, even though current law only requires such hospitals to provide community benefits generally, without mandating any specifics.[223] This underlines the dynamic nature of correlations; just asking certain questions might cause a change in behavior, which might in turn be correlated with greater, or less, legal compliance.

If asking certain questions on the Form 990 does in fact lead to a change in behavior, that fact could tempt the IRS or Congress to add further questions relating to practices they deem desirable in order to push nonprofits toward adopting them, even though such practices are not legally required. States could engage in similar encouragement through questions on their required registration and reporting forms. The only apparent way to stop this increase in government power would be for the government actors to restrain themselves from exercising their form-drafting discretion in this manner or, to the extent this restraint is

---

(discussing a disagreement over federal role in nonprofit governance, including collecting information regarding governance practices); Fred Stokeld, *IRS Drops Donor Disclosure Requirements*, 82 EXEMPT ORG. TAX REV. 113 (2018) (disagreement over need to collect donor information).

221. *See* Bonnie S. Brier et al., *The Appropriate Role of the Internal Revenue Service With Respect To Tax-Exempt Organization Good Governance Issues*, *in* ADVISORY COMMITTEE ON TAX EXEMPT AND GOVERNMENT ENTITIES (ACT) REPORT OF RECOMMENDATIONS 3, 46 (2008), https://www.irs.gov/pub/irs-prior/p4344—2008.pdf [https://perma.cc/6US2-4WZ7].

222. *See id.*; Mayer & Wilson, *supra* note 220, at 480 n.3 (other critics).

223. *See* 26 U.S.C. § 501(r) (2012); Rev. Rul. 69-545, 1969-2 C.B. 117.

lacking at the agency level, for the relevant legislature to forbid them from doing so.

4. *Other Legal Limits: Building a Plane While You Are Flying*

Commonly cited constitutional limitations on government deployment of Big Data are the free speech clause of the First Amendment, the Fourth Amendment's limits on searches and seizures, and the due process obligations under the Fifth and Fourteenth Amendments.[224] The protections provided by these provisions are related in part, but not entirely, to the privacy concerns discussed above.[225] There are few, if any, statutory or other legal limitations specifically targeting government use of Big Data methods.[226] In the federal tax context, the IRS is limited by generally applicable taxpayer privacy requirements, except for information Congress has explicitly made subject to public disclosure.[227] It is also limited by third party contact rules.[228]

Government use of Big Data implicates the First Amendment if such use places burdens on free speech, including but not limited to impairing the ability of individuals to speak anonymously, to associate with others without public disclosure, or to make intimate decisions without government interference, all of which are aspects of privacy.[229] It also implicates the Fourth Amendment to the extent the collection and use of personal data without the knowledge and consent of the affected

---

224. *See* MARTIN KUHN, FEDERAL DATAVEILLANCE: IMPLICATIONS FOR CONSTITUTIONAL PRIVACY PROTECTIONS 7–11 (2007); Jack M. Balkin, *The Constitution in the National Surveillance State*, 93 MINN. L. REV. 1, 19–23 (2008); Ian Kerr, *Predication, Preemption, Presumption: The Path of Law After the Computational Turn*, *in* PRIVACY, DUE PROCESS AND THE COMPUTATIONAL TURN: THE PHILOSOPHY OF LAW MEETS THE PHILOSOPHY OF TECHNOLOGY 91, 108 (Mireille Hildebrandt & Katja de Vries eds., 2013).

225. *See* KUHN, *supra* note 224, at 11–12 (different conceptualizations of privacy).

226. The most prominent federal guidance appears to be the Federal Trade Commission's non-binding "Fair Information Practices in the Electronic Marketplace" set of guidelines relating to consumer-oriented commercial Internet sites. *See* FED. TRADE COMM'N, PRIVACY ONLINE: FAIR INFORMATION PRACTICES IN THE ELECTRONIC MARKETPLACE, A REPORT TO CONGRESS (2000), https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission-report/privacy2000.pdf [https://perma.cc/G5TA-Y8K7].

227. *See* 26 U.S.C. § 6103 (2012).

228. *See* 26 U.S.C. § 7602 (2012).

229. *See* McIntyre v. Ohio Elections Comm'n, 514 U.S. 334, 341–42 (1995); Griswold v. Connecticut, 381 U.S. 479, 486 (1965); NAACP v. Alabama, 357 U.S. 449, 466 (1958); KUHN, *supra* note 224, at 51–76.

individuals is a search or seizure in the Fourth Amendment sense.[230] Finally such use implicates the due process requirements of the Fifth and Fourteenth Amendments to the extent that Big Data methods infringe on the life, liberty (including with respect to privacy), or property of individuals and their associations.[231]

In the nonprofit context, these constitutional concerns are muted but not completely absent. As already noted, privacy is not as obviously a pressing concern for nonprofits as it is in most other areas because almost all of the data at issue are already required by law to be publicly disclosed.[232] But the possible use of Big Data methods to reveal connections involving individuals still has privacy implications, and also Big Data methods could lead to disparate impacts based on the views expressed by nonprofits and therefore raise First Amendment concerns, especially since it could be difficult to determine if such disparate impacts were intentional or inadvertent.

For the data drawn from the annual information returns there are usually no Fourth Amendment concerns because the nonprofits themselves provide such data, as one of the conditions on the tax benefits they receive. In the future, however, Fourth Amendment concerns could arise for the IRS and state officials obtaining data from other sources without the knowledge or consent of the nonprofits involved, such as through mining social media or reading email, text, or other communication traffic. For example, the Supreme Court recently held that an individual has "a legitimate expectation of privacy in the record of his physical movements as captured through [cell-site location data]" and therefore this information is protected by the Fourth Amendment.[233] Finally, and as Professor Danielle Keats Citron has highlighted, any automated decisionmaking process—such as one used to select nonprofits to audit or otherwise investigate—raises significant due process concerns with respect to notice and the opportunity to heard.[234] Due process concerns are heightened if an automated decision-

230. *See, e.g.*, Carpenter v. United States, 585 U.S.__, 138 S. Ct. 2206, 2217–18 (2018) (applying the Fourth Amendment to collection of cell phone location records); United States v. Jones, 565 U.S. 400, 404 (2012) (applying the Fourth Amendment to attachment of a GPS device to a vehicle and use of that device to monitor the vehicle's movements).

231. *See* KUHN, *supra* note 224, at 200–01 (discussing due process and privacy as knowledge control).

232. *See supra* notes 88, 116 and accompanying text (discussing public disclosure of nonprofit government filings). Improper disclosure of the limited information that is not supposed to be subject to public disclosure appears to be relatively rare. *See supra* note 200 (outlining instances of improper disclosure).

233. *Carpenter*, 138 S. Ct. at 2217.

234. Danielle K. Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1281–88 (2008).

making process results in reflexive imposition of some type of penalty, such as revocation of tax-exempt status or a cease-and-desist order relating to charitable solicitation. At this point, however, no government regulator of nonprofits appears be going beyond audit selection based solely on algorithmic results.

While statutory or other legal limitations targeted specifically at Big Data methods do not exist at either the federal or state level, the IRS is subject to general taxpayer privacy and third party contact rules that may be implicated by the collection and analysis of data.[235] To maintain taxpayer privacy, the IRS is prohibited from publicly disclosing which specific tax-exempt nonprofits it is auditing or even the results of such audits unless the audit leads to revocation of tax exemption.[236] This limitation hinders the ability of outside parties to evaluate the effectiveness of the IRS's selection processes for audits, including any Big Data methods used. As for third party contacts where the IRS is inquiring about a specific individual or other entity to learn information relevant to tax liability, the IRS as a matter of practice seeks to obtain the desired information from the audit target, such as a tax-exempt nonprofit, before contacting a third party.[237] It is not clear, however, that this limitation would apply to a general request for information relating to multiple tax-exempt nonprofits, such as a broad request for social media data. This is because the IRS Internal Revenue Manual provides that the definition of third-party contacts does *not* include "[s]earches made on computer databases that do not require any personal involvement on the other end."[238] This exception might apply to at least some broad inquiries because the IRS may be able to draw information from some commercial databases to which it gains access through inquiries that are mediated solely by the computers that store the databases and so do not involve any human beings outside of the agency.

The uncertainty regarding whether and how the constitutional limitations and the applicable statutory limitations apply to Big Data collection and analysis mean that nonprofits subject to Big Data methods may face a costly and risky battle if they choose to challenge the actions of the IRS based on any of these other legal limits. It took congressional attention to end the IRS tactic of reading emails of taxpayers without

---

235. *See supra* notes 225–228 and accompanying text.

236. *See* 26 U.S.C. §§ 6103(a), (b)(2)(A) (2012); *Revocations of 501(c)(3) Determinations*, INTERNAL REVENUE SERV., https://www.irs.gov/charities-non-profits/charitable-organizations/revocations-of-501c3-determinations [https://perma.cc/7BXX-RYFF].

237. I.R.S. Notification Requirements, IRM 25.27.1.3(1) (Oct. 19, 2017).

238. *Id.* at 25.27.1.2(2)(a).

consent, notice, or court permission.[239] It may therefore take concerted action by umbrella organizations, including through calling for congressional action when warranted, to fully apply these limits to the use of Big Data methods with respect to nonprofits.

IV. RECOMMENDATIONS

Considering the steps that government agencies and private parties have taken to apply Big Data methods to overseeing and studying nonprofits, along with the promises and perils that doing so present, there are several recommendations that regulators, researchers, and nonprofits should consider. For agencies and legislatures, continued pursuit of these methods is clearly desirable, but specific steps must be taken to rigorously evaluate their effectiveness and to avoid the potential perils they present. For researchers, the easier access to much greater amounts of data relating to nonprofits must be tempered by recognition of those data's limitations and the possible effects such research could have on nonprofits. And for nonprofits, they must complete their government filings with the knowledge that ever-greater numbers of people will be accessing them and also the awareness that other, publicly accessible information such as their websites and social media presences may be compared with those filings for consistency and completeness.

A. For Regulators

Given the acknowledged resource limitations faced by the IRS and state agencies when it comes to overseeing nonprofits, it is imperative that they continue to leverage the technological advantages that a Big Data approach to such oversight provides if they can do so cost effectively.[240] At the same time, and as the IRS has acknowledged, not every algorithm will prove effective and efficient in identifying

---

239. *See* Houser & Sanders, *supra* note 28, at 823, 823 n.31 (citing Tim Sampson, *IRS Reverses Course on Warrantless Email Snooping*, DAILY DOT, (Apr. 17, 2013), https://www.dailydot.com/news/irs-email-warrantless-snooping-reversal/ [https://perma.cc/2UTK-UQNE]).

240. *See supra* section II.A (resource limitations). This recommendation assumes that a Big Data approach would allow regulators to do more with their existing, limited resources even given the technological challenges of implementing this approach, which assumption is generally consistent with the views of commentators. *See, e.g.*, EXEC. OFFICE OF THE PRESIDENT, *supra* note 20, at 37–38 ("Big data holds enormous power to make the provision of services more efficient across the entire spectrum of government activity . . . ."); MANYIKA ET AL., *supra* note 5, at 25 (predicting "that we will, at some point, see investments in big-data-related capital deepening pay off in the form of productivity gains").

nonprofits that are likely out of compliance with the applicable law.[241] This is particularly true given the possible accuracy issues with both the underlying data and the tools used to analysis them.[242] It is therefore important that the agencies themselves rigorously evaluate whether the specific methods chosen are significantly more effective at detecting noncompliance as compared to other, non-Big Data approaches, and that independent oversight entities such as the Treasury Inspector General for Tax Administration also do so.[243] The IRS has stated that it continues to rely heavily on its "robust referral process" to identify noncompliance, along with more data-driven methods.[244] There may also be a role for outside researchers in this regard, as detailed in the next section, although that role is necessarily limited by the need for such agencies to limit public access to the criteria used to identify nonprofits that are likely to be violating the applicable laws.[245]

Regulators may also have to consider making more efforts to ensure that government filings are accurate. In theory, inaccuracies can result in an IRS return being treated as incomplete, triggering late filing penalties.[246] Inaccuracies can also theoretically result in a perjury conviction for the nonprofit leader who signs the return.[247] However, it is, at best, unclear how often the IRS imposes penalties for inaccurate information.[248] While in early 2018 the IRS implemented a new process for rejecting incomplete or incorrect paper-filed Form 990s, leading to an over 10% rejection rate, it is unclear what other steps the IRS has

241. *See supra* note 183 and accompanying text (discussing the ineffectiveness of approach used to try to identify noncompliant private foundations).

242. *See supra* sections III.A.1, III.A.2.

243. *See About TIGTA*, TREASURY INSPECTOR GEN. FOR TAX ADMIN., https://www.treasury.gov/tigta/about.shtml [https://perma.cc/JF5W-8A7Z] (noting their mission to "[p]rovide quality, professional audit, investigative, and inspection and evaluation services that promote integrity, economy, and efficiency in the administration of the Nation's tax system").

244. Von Lienen, *supra* note 44.

245. *See supra* note 143 and accompanying text.

246. *See* 26 U.S.C. § 6652(c)(1)(ii) (2012) (stating that an excise tax can be imposed on a tax-exempt nonprofit for "a failure to include any of the information required to be shown on [an annual information return] . . . or to show the correct information").

247. *See* 26 U.S.C. § 7206 (creating a felony for willfully making and subscribing any return that contains a written declaration that it is made under the penalties of perjury if the person doing so does not believe the return to be true and correct as to every material matter); IRS, 2018 FORM 990 *supra* note 50, at 1 (stating that a signature is made "[u]nder penalties of perjury").

248. *See* IRS, *supra* note 47, at 42–43 & n.14 (reporting the assessment of approximately 54,000 "daily delinquency" penalties, totaling approximately $177 million, but indicating such penalties are primarily for failures to file).

taken to improve return accuracy.[249] To the extent the IRS's Big Data approaches rely primarily if not almost exclusively on annual information return data, the IRS should consider increasing the frequency of penalties when it discovers significant inaccuracies in such returns. State agencies should also consider similar measures for their required filings.

These points are of course obvious and appear to have already been embraced at least in part by at least the IRS. What is less obvious, but also should be considered by these agencies, is the extent to which Big Data approaches implicate the larger concerns identified above.[250] More specifically, for privacy, IRS and state agencies need to consider what if any steps they should take to prevent personal information that is not required to be included in government filings, but that nonprofits inadvertently provide, from becoming publicly available. The IRS has in fact already taken at least one step in this regard, in that it removes social security numbers from e-filed Forms 990 before providing that information to Amazon Web Services.[251] However, it acknowledged that it does not identify or redact other personally identifiable information such as home addresses and the ages of students receiving scholarships.[252]

The IRS and state agencies also have to be sensitive to the possibility that methods chosen may have unintentional but undesirable discriminatory effects. The Federal Trade Commission (FTC) has issued a report highlighting these concerns for private actors and Big Data, noting that to address this potential requires considering whether a data set is representative, whether a data model accounts for possible improper biases, how accurate predications based on Big Data actually are, and the appropriate balance between predictive accuracy and fairness considerations.[253] Similar considerations need to reflected in evaluations of Big Data methods adopted to help oversee nonprofits. The last point is especially important, because even an algorithm that is highly successful in detecting legal noncompliance may have to be set aside if its results also are fundamentally unfair in that they

249. *See* Von Lienen, *supra* note 44.

250. *See supra* section III.B.

251. Transcript of Webcast with Sunita Lough, Insights into the IRS's New Data-Driven Approach to Examining Tax-Exempt Organizations (Part 1) (Oct. 11, 2017), *in* EO TAX J. 2017-220 (Nov. 10, 2017).

252. *Id.*

253. FTC, *supra* note 157, at iv–v.

disproportionately identify noncompliance within, for example, organizations with a particular ideological bent.

Concerns about government overreach, including possibly implicating constitutional and statutory limitations, should also temper proposals either to expand the information collected on required government filings or to gather information from external sources. The latter is of particular concern, given that the Congress has already chastised the IRS for obtaining information such as emails in possible violation of constitutional limitations.[254] The IRS and state agencies should therefore be sure to measure such data gathering against such legal limitations.

Congress and state legislatures also have a role that goes beyond overseeing the IRS and state agencies, respectively, in the areas already detailed. First and most obviously, they can mandate electronic filing of government forms required from nonprofits, thereby easing access to the information on such forms for the agencies and, if released, researchers and other private parties. This is particularly important given that organizations seeking to avoid regulatory and public scrutiny may increasingly avoid e-filing as access to e-filed returns increases, unless e-filing is mandatory. Congress has in fact recently imposed mandatory e-filing for tax-exempt nonprofits.[255]

Congress and state legislatures also could take steps to address some of the larger concerns identified above. One step Congress should take is to enact a prohibition on private parties using individual information—such as compensation—provided on nonprofit returns for commercial purposes, to address the privacy concern raised by the reporting and disclosure of such information.[256] This is not unprecedented, as in other contexts where individuals are subject to public financial disclosures because of their affiliation with a particular type of organization or government position, Congress has prohibited the sale or use of information regarding these individuals for any commercial purpose.[257] At least one such rule has survived constitutional challenge.[258] Congress

254. *See supra* note 239 and accompanying text.

255. *See supra* note 64 and accompanying text.

256. *See supra* note 205 and accompanying text.

257. *See* 5 U.S.C. app. 4 § 105(c)(1)(B) (2012) (information for senior government officials contained in ethics reports); 52 U.S.C. § 30111(a)(4) (2012) (information for contributors to candidates, political parties, and political committee); 11 C.F.R. § 104.15 (2019) (same); *Sale or Use of Contributor Information*, FEDERAL ELECTION COMM'N, https://www.fec.gov/updates/sale-or-use-contributor-information/ [https://perma.cc/ND7Z-DJF2] (same).

258. *See* Fed. Election Comm'n v. Int'l Funding Inst., 969 F.2d 1110 (D.C. Cir. 1992) (en banc) (rejecting both a facial and an as applied challenge to 52 U.S.C. § 30111(a)(4) under the First Amendment).

and state legislatures could also be the appropriate venues for considering whether certain agency practices may come too close to or even cross constitutional and other legal limits, thereby possibly avoiding lengthy and expensive litigation.

B. For Researchers

Some researchers, both in the law and in other disciplines, have already started working with the existing and growing databases of government-collected information relating to nonprofits.[259] As the amount and visibility of these data grows, presumably this group will also grow. This is a welcome development, as the resources and interests of government agencies are very limited, while private researchers should be able to tap additional funding sources and analyze important issues that may not attract agency attention.[260] In fact, at least one IRS official has encouraged feedback from private researchers.[261]

That said, researchers must be wary of at least two significant pitfalls. The first and most obvious one is the accuracy of the underlying data. As noted above, there are indications that at least the annual information returns filed with the IRS may contain significant inaccuracies for many nonprofits, particularly with respect to information that has public relations or rating implications.[262] While it may be difficult to identify, much less correct, such inaccuracies, researchers should be sensitive to how such inaccuracies may affect their results. When possible, they also should seek other sources of information to corroborate and correct IRS-provided data as necessary.

The second and less obvious one is the so-called "streetlight effect," based on the story of the drunk looking for their keys under the streetlight even though they lost them elsewhere because "the light is better."[263] The currently available machine-readable data is limited in at

---

259. *See supra* notes 147–149 and accompanying text.

260. Jesse Lecy, Nathan Grasse & Leonor Camarena, Do Female Board Members Reduce the Gender Pay Gap for Nonprofit Managers?, Panel at the Int'l Soc'y for Third-Sector Res. Conference (July 10–13, 2018), https://cdn.ymaws.com/www.istr.org/resource/resmgr/amsterdam/program.final_6.12.18.pdf [https://perma.cc/G2MH-MTHK] (discussing gender differences in compensation).

261. *See* Lough, *supra* note 45 (inviting the public to look at the IRS-related e-filing data and "tell us what they're seeing").

262. *See supra* notes 162–168 and accompanying text.

263. *See* Jake Linford, *Datamining the Meaning(s) of Progress*, 2017 B.Y.U. L. Rev. 1531, 1554 n.121 ("[u]sing inapplicable but easily accessible data can contribute to behavior based on bad information, triggering a 'streetlight' effect"); David H. Freedman, *Why Scientific Studies Are So Often Wrong: The Streetlight Effect*, DISCOVER (July-Aug. 2010), http://discovermagazine.com/2010/jul-aug/29-why-scientific-studies-often-wrong-streetlight-effect [https://perma.cc/3WES-MJ2F].

least three important ways, each of which could cause researchers to unduly focus on certain aspects of the nonprofit sector or try to draw inappropriate conclusions from these limited data. One limitation is that the comprehensive IRS annual information return data only includes recently e-filed returns, although the NCCS databases provide older data for a limited set of fields from all returns.[264] Use of these data therefore is not fully representative of the nonprofit sector of a whole and also only provides a limited historical perspective.

Another limitation is that the annual information returns understandably focus primarily on financial data and on certain information relating to specific legal requirements. This makes them a rich source of information for legal and financial researchers but limits their usefulness for other purposes. For example, researchers interested in exploring the relative impacts of various nonprofits will likely find little information to assist in their efforts. Or worse, they may try to draw inappropriate conclusions from the data available—as is often the criticism made of rating agencies and others that rely heavily on administrative and fundraising to programmatic cost ratios.[265]

A third limitation involves organizations that either are permitted to file less lengthy annual information returns or are entirely exempted from filing any such returns. The former are usually limited to nonprofits that are smaller financially, which may make them less important for some but not all research purposes.[266] The latter include most notably churches and church-related entities, but also include, among other entities, public colleges and universities.[267] So a researcher interested in higher education would find the data they obtained from the IRS filings to be missing this important portion of that sector.

None of these limitations necessarily undermine the usefulness of the increasingly available data regarding nonprofits, but they must be considered when designing research projects in order to avoid both misleading conclusions and neglect of certain aspects of the nonprofit sector. Because of these limitations, researchers who want to analyze issues not fully supported by the existing data will need to consider gathering data from other sources to address these representational and

---

264. *See supra* notes 54, 94 and accompanying text.

265. *See Overhead Ratios Are Essential for Informed Giving*, CHARITY WATCH, https://www.charitywatch.org/charitywatch-articles/overhead-ratios-are-essential-for-informed-giving/133 [https://perma.cc/6RQB-WQG9] (summarizing the debate relating to using such ratios to evaluate charities while defending their use).

266. *See supra* notes 42–49 and accompanying text.

267. *See supra* note 46 and accompanying text.

other concerns. A related concern is that by focusing on certain reported information, such as fundraising ratios, researchers may actually influence nonprofit behavior, at least for those nonprofits sensitive to how they are portrayed by researchers.

C. For Nonprofits

Finally, the nonprofits subject to the increasingly Big Data focused scrutiny of regulators and researchers need to consider how best to adapt to this change. First and most obviously, while the public nature of IRS annual information returns has long made them public relations documents as well as government filings, the increased access those returns and the data they contain only further emphasizes their public relations role. Particularly for nonprofits that depend on having a strong public reputation, ensuring that their returns both accurately and positively reflect their finances and activities will become even more important. This is in addition to ensuring that information that is sensitive and not required to be provided, such as home addresses of board members or social security numbers of employees, is not included in the returns.

But even for nonprofits that are less sensitive to public perceptions, perhaps because they depend primarily on government contracts or a single donor for revenues, the increased access to their government filings could become problematic if those filings are not consistent with information in other databases, such as federal grant audit results, or other publicly available information sources, such as social media. For example, a state political organization filing revealed a donation by the Donald J. Trump Foundation to that organization even though the Foundation was prohibited from making such contributions by federal tax law and had, inaccurately, reported the grant as having been made to a similarly named but different, non-political organization on the Foundation's IRS annual information return.[268] While that revelation did not prevent Donald Trump's election, it did create some negative publicity and forced him to both correct the expenditure, presumably by reimbursing the Foundation, and pay a penalty tax to the IRS on behalf

268. Letter from Noah Bookbinder, Exec. Dir., Citizens for Responsibility and Ethics in Washington, to John A. Koskinen, IRS Comm'r (Mar. 21, 2016) (discussing how a Florida campaign finance filing by recipient political organization revealed donation from Foundation); *see* Allison Graves, *Donald Trump, Pam Bondi and $25K: Was it pay to play?*, POLITIFACT FLA. (Sept. 21, 2016, 5:53 PM), https://www.politifact.com/florida/article/2016/sep/21/donald-trump-pam-bondi-and-25k-was-it-pay-play/ [https://perma.cc/P2JW-V82S] (discussing the inaccurate reporting of donation on Foundation IRS annual return).

of the Foundation.[269] As it becomes easier for not just regulators and academic researchers but also reporters, critics, and other members of the public to connect the dots of publicly available nonprofit information, ensuring that such issues do not arise will become increasingly important for most nonprofits.

CONCLUSION

In the United States and elsewhere, government regulators and interested private parties are starting to turn toward Big Data—the collection of enormous amounts of digital data analyzed with sophisticated machine learning tools—to better oversee and understand nonprofits. One of the advantages of being relatively late to adopt a Big Data approach is the ability to take advantage of the now significant amount of thought given to how such an approach can best enhance such oversight and what risks such an approach may present.

Big Data may allow government regulators to deploy their limited resources in a much more efficient manner and to more effectively fulfill their oversight roles, which is very much needed given longstanding resource constraints. That said, to fully realize these promises, regulators and others must rigorously test the methods applied to determine whether and to what extent they improve efficiency and effectiveness of government oversight. It is not yet clear whether either the IRS or state regulators are committed to such rigorous review, although the IRS has indicated it has taken some steps along these lines. To the extent legally and pragmatically permissible, regulators should also make their methods visible to internal and external evaluators so as to enhance this review and at the same time minimize the perils posed by using Big Data techniques.

As for those perils, regulators and private parties must be conscious of the limitations on the data available to them, in terms of accuracy and comprehensiveness, and on the analytical tools they employ. They must not forego the continued use of other methods to help compensate for these limitations. However, they must also be alert to the possibility that Big Data may lead to unintentional invasions of privacy, have prohibited disparate impacts, create temptations to increase government power beyond legislative mandates or infringe other legal limits. They must therefore establish procedures to detect and counter such possibilities if they arise. Fortunately, there is time to put such procedures in place at

269. *See* Graves, *supra* note 268.

both the federal and state level, but this task must be considered an integral part of establishing a Big Data approach and not just an afterthought. If it is, then the perils of Big Data are likely to be avoided while its promises are fulfilled for the regulation of nonprofits.