The Promises and Perils of Using Big Data to Regulate Nonprofits.pdf

Lloyd Hitoshi Mayer
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 and expert analysis of those data to reveal previously undiscovered patterns and so revolutionize the regulation of criminal behavior, education, health care, and many other areas. For the pessimist, such use involves the haphazard seizure of information to generate massive databases that render privacy an illusion and result in arbitrary and discriminatory computer-generated decisions. The reality is of course more complicated, with government use of Big Data presenting on one hand the promises of greater efficiency, effectiveness, and transparency, and on the other hand the perils of inaccurate conclusions, invasion of privacy, unintended discrimination, increased government power, and violations of other legal limits on government action.

Until recently, these issues were theoretical for nonprofits in the United States given that the federal and state regulators overseeing them did not use a Big Data approach. 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. And both federal and state regulators are making or plan to make much of the data they collect available in machine-readable form to researchers, journalists, and other members of the public. The question now is therefore whether regulators, researchers, and nonprofits can learn from the Big Data experiences of other agencies and private actors so as to fully realize Big Data’s promises while avoiding the numerous perils it presents.

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 those regulators and others can 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, and the Nonprofit Forum. I am also very grateful for research assistance from Thomas Everett and Jasmine Rasheed.
TABLE OF CONTENTS

Introduction .................................................................................................................. 2  
I. Big Data and Nonprofits ........................................................................................... 6  
   A. The Internal Revenue Service .............................................................................. 7  
   B. State Regulators ................................................................................................ 14  
   C. Nonprofits and the Public ................................................................................... 16  
   D. Other Countries .................................................................................................. 22  
II. The Promises of Using Big Data To Regulate Nonprofits ...................................... 23  
   A. Improved Government Oversight ...................................................................... 24  
   B. Improved Research .............................................................................................. 25  
   C. Improved Information for the Public ................................................................. 27  
III. The Perils of Using Big Data To Regulate Nonprofits ........................................ 28  
   A. Getting It Wrong .................................................................................................. 29  
      1. Data Inaccuracy & Limits: Garbage In, Garbage Out ................................. 29  
      3. Over-Reliance: But the Computer Told Me To Do It ............................... 35  
   B. Larger Concerns ................................................................................................... 36  
      1. Invasion of Privacy: Welcome to the Goldfish Bowl .................................. 36  
      3. Increased Government Power: Orwell and Kafka ..................................... 40  
      4. Other Legal Limits: Building a Plane While You’re Flying ..................... 42  
IV. Recommendations .................................................................................................. 45  
   A. For Regulators ..................................................................................................... 46  
   B. For Researchers .................................................................................................... 48  
   C. For Nonprofits ....................................................................................................... 50  
Conclusion .................................................................................................................. 51  

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 be taking 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 improve processes,” to apply “analytics and research to improve program effectiveness,” and to allocate resources “using a data-driven approach to target existing and emerging high-risk areas.” For anyone familiar with recent technological developments, however, these buzzwords signaled that the IRS was moving in a new direction with respect to 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. 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 20 years old, it already has developed not only a technological meaning but also academic and societal meanings. At its heart, however, it refers to both the rapid accumulation of digital data, ranging from social media posts to financial transactions to cell phone locations, and efforts to use such data to discover significant patterns that can inform and improve public policy making, business decisions, and personal choices.

A couple noteworthy examples have highlighted the potential benefits and dangers of Big Data. In 2012, the N.Y. Times Magazine included in an article about private company use of Big Data the story of how retailer Target started using customers’ shopping patterns to determine if it was likely they were pregnant. While Target apparently did this only so it could better target its advertising to those customers, it was wisely wary about how its customers

---


would react if they learned how much Target knew about them (including one instance when 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. For example, 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 significance of the IRS’ Big Data move with respect to nonprofits is that it highlights that data about such organizations may have reached a critical tipping point in several respects. First, the IRS’ 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 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 the potential for even greater such

---

7 Id.
9 Id. at 1468.
access in the future if the U.S. Congress moves forward with proposals to expand the electronic filing obligations of tax-exempt nonprofits.\textsuperscript{12}

The promise of Big Data with respect to government oversight of nonprofits is similar to its promise in other regulatory areas: greater efficiency, effectiveness, and transparency with respect to both enforcement of existing legal rules and development of new rules.\textsuperscript{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.\textsuperscript{14} But as in other areas, Big Data brings possible perils as well: bad data, bad analysis, over-reliance, threats to privacy, discriminatory effects, unchecked government power, and other legal concerns, including constitutional ones.\textsuperscript{15}

This article considers what lessons can be learned from Big Data’s application in other areas to inform its emerging application with respect 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.\textsuperscript{16} The ultimate goal is to provide a roadmap for fulfilling the promises of Big Data while avoiding its perils, including by ensuring awareness of both the positive and negative ramifications of Big Data, rigorously evaluating the improvements to efficiency and effectiveness (if any) resulting from it, and taking precautions to avoid or minimize the perils it presents.

Part I examines both the data and analytic tools currently available to the IRS, state regulators, and the public, as well as the data and tools that may be available to them in the near future, included based on the use of Big Data techniques by nonprofit regulators in other countries. 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 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

\textsuperscript{12} See infra Section I.C.
\textsuperscript{13} See infra Section II.A.
\textsuperscript{14} See infra Section II.C.
\textsuperscript{15} See infra Part III.
\textsuperscript{16} See infra Section I.D.
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. Such measures can help ensure the promises of Big Data are realized, while minimizing exposure to the perils it creates.

I. BIG DATA AND NONPROFITS

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 bright lines with respect to any of these dimensions for 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 measured in terabytes or petabytes of data (a terabyte is approximately 1000 gigabytes and a petabyte is approximately 1000 terabytes).18 To give a sense of scale, the IRS Director for Research Databases reported in 2013 that the then volume of the data collected by the IRS in its Compliance Data Warehouse was approximately 1.3 petabytes.19 Velocity for many large datasets now approaches real-time collection, as illustrated by automatic recording of users’ searches and other online activities.20 Variety as to structure reflects the extent to which data is organized into fixed fields (such as typical for a spreadsheet) as opposed to lacking such structure (for example, most of the data in emails or tweets).21

---


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.


21 MANYIKA ET AL., supra note 5, at 33.
The promises and perils of using big data

The process of analyzing Big Data datasets is part of the broader field known as “knowledge discovery in databases” or KDD, with “[t]he 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 machine learning – the ability of computer systems to modify their behaviors without human intervention – that allows for the processing of 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 with respect to nonprofits and the related issue of to what extent the IRS and others have developed tools to analyze such datasets, as well as the potential for the emergence of such datasets and tools in the near future.

A. The Internal Revenue Service

The 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 those data more than 50 years ago. 26 And approximately 20 years ago the IRS started exploring how it could use the now vast amount of data it collects in a 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. For example, the IRS now gathers data from many commercial and public data pools, including social media sites such as Facebook, Instagram, and Twitter, and plans to do more along these lines. 28 The 1.3 petabytes of compliance-

23 MANYIKA ET AL., supra note 5, at 29.
24 See KITCHIN, supra note 17, at 101-02 (role of machine learning in the analysis of Big Data).
26 Id. at 766-67.
28 See 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).”); Siri Bulusu,
related data the IRS had gathered as of 2013 therefore reflects not only
information provided in tax returns, which only accumulates at about 15 to
20 terabytes per year, but also a substantial amount of data collected from
other sources. It also represents an almost ten-fold increase in such data
during the five years from 2008, and the IRS recently estimated that during
the 10 years from 2007 to 2017 its data volume increased a hundred fold. These
efforts are in addition to other financial-related data collection by
the federal government, such as by the Treasury’s Financial Crimes Enforcement
Network (created in 1990 and which administers the Bank Secrecy Act).
Unfortunately, more recent information regarding the volume of information
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 therefore somewhat late to the party, although there indications as early as
2003 that it was considering a more rigorous use of data. This is not to say
that the IRS has not previously sought to collect and computerize data
regarding tax-exempt nonprofit organizations. For example, as early as the

6, 2018 (“The IRS can currently track data coming from electronic transactions, cell phones,
live video feeds, and social media.”); 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, available at
https://www.fbo.gov/index?s=opportunity&mode=form&id=e832fdd38233b84f8ad65b7fe
a26eeba&tab=core&cview=0), https://qz.com/1507962/the-irs-wants-to-use-facebook-
and-instagram-to-catch-tax-evaders/.

29 See Butler, supra note 19, at 4 (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 (data from one year’s worth of tax returns are
15 to 20 terabytes). https://www.computerworld.com/article/2536160/business-

30 See INTERNAL REVENUE SERVICE (IRS), STRATEGIC PLAN FY2018-2022, at 19 (2018)
(data volume increased 100 times from 2007 to 2017), https://www.irs.gov/pub/irs-
pdf/p3744.pdf.

31 See OFFICE OF INSPECTOR GENERAL, DEP’T OF THE TREASURY, AUDIT REPORT,
TERRORIST FINANCING/MONEY LAUNDERING: FinCEN’S BSA IT MODERNIZATION
PROGRAM IS MEETING MILESTONES, BUT OVERSIGHT REMAINS CRITICAL 1 (2012) (role of
FinCEN), https://www.treasury.gov/about/organizational-
structure/ig/Audit%20Reports%20and%20Testimonies/OIG12077.pdf; 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) (FinCEN expert system
that analyzes financial data collected from government, private, and foreign databases).

32 See J. Christine Harris, IRS Will Enlist New EO Units to Complement Division’s Work,
Say Officials, 42 EXEMPT ORG. TAX REV. 150 (2003) (new Data Analysis Unit to assist IRS
Exempt Organizations division).
1970s the IRS had already begun to enter certain information from the annual information returns filed by the larger (financially) of such organizations into its computer systems. But the 2014 Program Letter appears to be the first indication that this part of the IRS was taking advantage of Big Data techniques when analyzing these data.

Being late has the advantage, however, of being able to benefit from the experience of the Research, Applied Analytics and Statistics Division (RAAS) of the IRS, which implements Big Data projects for the agency. RAAS is the result of a 2016 merger of two previous IRS offices, the Office of Research, Analysis, and Statistics and the Office of Compliance Analytics. 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’ data analytics activities. As of late 2017, RAAS had approximately 350 employees plus an additional 50 personnel detailed from other agencies or student volunteers, as well as approximately 300 contractors.

TE/GE expected to have algorithms developed by RAAS in place by April 1, 2017 to select annual information returns (Form 990 series) filed by tax-exempt nonprofits for audit. It also planned to run every Form 990 (the longest return, filed by the financially largest tax-exempt organizations) and every Form 990-PF (the return filed by all private foundations, which are subset of charitable tax-exempt organizations) through that selection process. The most recent TE/GE Work Plan indicates that the IRS has now expanded these efforts to also include the Form 990-EZ, a shorter version of

---


34 See *supra* note 2 and accompanying text.


36 See *id.* at 1.


the Form 990 that may be used in place of the Form 990 by tax-exempt nonprofits (other than private foundations) with relatively modest assets and annual revenues (annual gross receipts of less than $200,000 and total assets of less than $500,000).\textsuperscript{41} 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.\textsuperscript{42} And the acting TE/GE Commissioner recently stated that over 50 percent of TE/GE’s examinations that the division closed in fiscal year 2018 were selected using data-driven approaches.\textsuperscript{43}

As these plans indicate, the IRS’ primary source of information relating to tax-exempt nonprofits is the annual information returns filed by such organizations, although an IRS official has stated the IRS is using other, unspecified “external data” as well.\textsuperscript{44} 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 such a return.\textsuperscript{45} These organizations filed a total of approximately 1.5 million such returns annually in recent years.\textsuperscript{46} The breakdown by type of annual return for fiscal year 2015 shows approximately 260,000 Forms 990 and 125,000 Forms 990-EZ filed by the most common types of such organizations, and another approximately 100,000 Forms 990-PF filed by private foundations.\textsuperscript{47} A variety of other


\textsuperscript{44} See Sunita Lough, Commissioner of the Tax Exempt & Gov’t Entities Div., I.R.S., Ernst & Young Webcast (Oct. 11, 2017), in EO Tax J., 2017-219, Nov. 9, 2017.


\textsuperscript{47} See Internal Revenue Service Statistics of Income Division (IRS SOI), Table 1, Domestic Private Foundations: Number and Selected Financial Data, by Type of Foundation and Size of End-of-year Fair Market Value of Total Assets, Tax Year 2015 (2018) (Forms 990-PF), https://www.irs.gov/pub/irs-soi/15pf01ta.xls; Internal Revenue Service Statistics
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 (no more than $50,000 in annual gross receipts, on average) tax-exempt nonprofits; for example, more than 600,000 nonprofits filed the Form 990-N for fiscal year 2017. 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 predications by the applicants regarding their future plans, not reporting about 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; for example, the Form 990 asks over 200 questions (most requiring a yes/no or single numerical answer, but some requiring narrative responses, and many with subparts) and also has 16 schedules that a tax-exempt nonprofit may need to complete depending on its specific activities. Nevertheless, it is important to note that there are reasons to question whether those data actually constitute a “Big Data” dataset as that term is commonly used for several reasons.

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. Of course this does not take into account the possibility that TE/GE is also pulling information from other tax filings – such as the 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 Income Division (IRS SOI), 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; Internal Revenue Service Statistics of Income Division (IRS SOI), 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.


50 See IRS, supra note 46, at 4 (number of returns filed with the IRS for fiscal years 2016 and 2017); Lai, supra note 29 (data from one year’s worth of tax returns are 15 to 20 terabytes).
of these data’s accumulation is relatively slow because the annual information returns are not due for four and a-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 (although this extension is not available for nonprofits planning to file a Form 990-N).\footnote{See \textsc{Internal Revenue Service} (IRS), \textit{Form 8868} (2017), \url{https://www.irs.gov/pub/irs-pdf/f8868.pdf}.} 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 (over 40% of Forms 990 and Forms 990-EZ, and a similar proportion of Forms 990-PF) that are filed in hard copy and so addition of their information to the IRS’ electronic database is further delayed.\footnote{See Cavanaugh et al., \textit{supra} note 48, at 39 (roughly 57 percent of all Forms 990 and 990-EZ were filed electronically for fiscal year 2017); Nonprofit Open Data Collective, IRS 990 efiler database (2018), \url{https://github.com/Nonprofit-Open-Data-Collective/irs-990-efiler-database} (approximately 60,000 Forms 990-PF filed electronically for calendar year 2016).} Third and finally, there is limited variety to these data; most of these data are “structured” in that they track the fields used in the 990 series forms, although the forms do have some questions that require more unstructured, narrative responses.\footnote{See IRS, \textit{supra} note 45; IRS, \textit{supra} note 49.} And unlike the more general IRS date collection efforts, it is not clear that TE/GE is currently collecting a significant amount of data from sources other than IRS filings with respect to tax-exempt nonprofits.\footnote{See supra note 28 and accompanying text (general IRS external data collection).}

It also is not clear to what extent the IRS is deploying machine learning techniques to analyze these 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 not clear the IRS has begun deploying them in this area (buzzwords notwithstanding).\footnote{See, \textit{e.g.}, TE/GE, \textit{supra} note 41, at 6 (mentioning only “compliance models,” risk identification, and collaboration with RAAS).} 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.\footnote{DeBarr & Harwood, \textit{supra} note 27, at 178.}

So TE/GE may not yet be truly using a “Big Data” approach to regulate tax-exempt nonprofits (and TE/GE officials have not actually used that term in their pronouncements). Given, however, the use of such an approach by the broader IRS it likely is only a matter of time before the same methods – including harvesting significant amounts of data from sources beyond IRS
filings and deploying machine learning techniques to analyze such data – are also used by TE/GE. This would be consistent with the increasing reliance of the IRS generally on such an approach; one commentator predicts that over the next 25 years the IRS likely will substantially increase its gathering of information from sources beyond IRS filings in order to reduce both the compliance burden on taxpayers and the compliance gap in terms of owed but unpaid taxes, although there are some technological and political barriers to doing so.\textsuperscript{57} Other commentators have noted that there are various indications the IRS is already gathering such data, including the agency’s purchase of cell phone tracking technology, its disclosure in response to a Freedom of Information Act (FOIA) request that it had been reading taxpayers’ private emails without a warrant (although the IRS apparently ended this practice after congressional scrutiny), and its disclosure in response to another FOIA request that it collects information from social media sites such as Facebook (apparently using automated computer programs sometimes referred to as “spiders”).\textsuperscript{58} Given sufficient time and resources, these same techniques could be applied to tax-exempt nonprofits, as well as 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.\textsuperscript{59}

Furthermore, Congress has repeatedly considered legislation that would require all tax-exempts organizations required to file annual returns to do so electronically.\textsuperscript{60} Currently only relatively large tax-exempt nonprofits (total assets of $10 million or more and that file at least 250 returns, such as wage reporting Form W-2, annually) are required to file the Form 990 electronically, although the Form 990-N filed by relatively small tax-exempt nonprofits must be filed electronically.\textsuperscript{61} A broader e-filing requirement

\textsuperscript{57} Michael Hatfield, \textit{Taxation and Surveillance: An Agenda}, 17 \textit{YALE J. L. \& TECH.} 319, 340 (2015).

\textsuperscript{58} Houser & Sanders, \textit{supra} note 28, at 822-24.

\textsuperscript{59} See 26 U.S.C. §§ 170(c)(2) (conditions relating to the ability to receive tax-deductible charitable contributions), 501(c) (conditions relating to exemption from income taxes), 4911-4912 (excise taxes imposed because of excessive lobbying by charitable, tax-exempt nonprofits), 4940-4946 (restrictions on private foundations, a subset of charitable, tax-exempt nonprofits), 4955 (excise taxes imposed because of election-related activities by charitable, tax-exempt nonprofits), 4958 (excise taxes on “excess benefit transactions” between certain tax-exempt nonprofits and insiders).

\textsuperscript{60} See, e.g., Retirement, Savings, and Other Tax Relief Act of 2018, Rules Committee Print 115-85, Text of the House Amendment to the Senate Amendment to H.R. 88, § 3101 (2018); H.R. 5443, 115th Cong. § 1 (2018); H.R. 1, 114th Cong. § 6004 (2014).

\textsuperscript{61} \textit{INTERNAL REVENUE SERVICE (IRS), E-FILE FOR CHARITIES AND NON-PROFITS},
would marginally increase the velocity of the data from annual returns. It also would increase public access to this information, as detailed later in this part. There is significant government and nonprofit support for such a broader requirement, including from the Advisory Committee for TE/GE. It is unclear if and when Congress will pass such legislation, however, especially since members of Congress have introduced this proposal multiple times in recent years without success.

B. State Regulators

In addition to the oversight provided by the IRS, states also oversee the operations of certain nonprofits. More specifically, most although not all 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 (or both), with “charitable” usually defined broadly so as to encompass not only nonprofits that qualify as charities under federal tax law (and so tax-exempt under Internal Revenue Code section 501(c)(3)) but also some other types of nonprofits. 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 – with relatively limited information sharing both 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 as well. Furthermore, in almost half the states authority over charitable


64. See FREMONT-SMITH, supra note 63, at 305-06 (charitable assets), 315 (registration and reporting), 372-73 (charitable solicitation). For an example of the breadth of state law definitions of “charitable,” see OHIO ATTORNEY GENERAL, HANDBOOK FOR NONPROFITS: AN OPERATIONAL RESOURCE FOR BOARD MEMBERS OF CHARITABLE ORGANIZATIONS 8 (2015). https://www.ohioattorneygeneral.gov/Files/Publications-Files/Publications-for-Non-Profits/NonprofitHandbook (nonprofits tax-exempt under both Internal Revenue Code section 501(c)(3) and section 501(c)(4) generally required to register under the Ohio Charitable Trust Act).

nonprofits is split between two state agencies.\textsuperscript{66} Local jurisdictions may also have their own registration and reporting requirements for charitable nonprofits that solicit donations from their residents.\textsuperscript{67} Finally, the tax treatment of nonprofits, including with respect to not only state income taxes but also state sales tax and local property taxes, is usually handled by additional state or local agencies.\textsuperscript{68}

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 such statements.\textsuperscript{69} In general, the more staff a state dedicates to oversight of charitable nonprofits the greater the likelihood that a state agency will maintain such a database, whether only for internal use or also for public use.\textsuperscript{70} The use by regulators of such databases appears to be limited to relatively simple searches, however. For example, an official from the largest state by population (California) recently remarked that her office uses an algorithm to identify every charitable nonprofit that reports a self-dealing transaction—generally a transaction to which the nonprofit is a party and in which a director of the nonprofit has a material financial interest—on its state filings; this is a very basic search as it only requires determining which such organizations have answered “yes” to a single question on the required annual report.\textsuperscript{71} As is the case with the IRS, state regulators generally are interested in ensuring that nonprofits, and particularly charitable nonprofits, and their leaders comply with their legal obligations, which under state law include ensuring assets remain dedicated to the nonprofit’s stated purposes, limiting transactions with

\begin{thebibliography}{99}
\bibitem{67} See, e.g., CALIFORNIA ATTORNEY GENERAL, GUIDE FOR CHARITIES: BEST PRACTICES FOR NONPROFITS THAT OPERATE OR FUNDRAISE IN CALIFORNIA 73 (2017), https://oag.ca.gov/sites/all/files/agweb/pdfs/charities/publications/guide_for_charities.pdf (“Many local jurisdictions in California also require charitable organizations to obtain a license or permit before soliciting donations from residents and in public spaces.”).
\bibitem{69} LOTT ET AL., supra note 65, at 14-15.
\bibitem{70} Id. at 25-26.
\end{thebibliography}
insiders to reasonable amounts, complying with registration, reporting, and other requirements associated with charitable solicitation, and, for nonprofit leaders, otherwise fulfilling their fiduciary duties.\textsuperscript{72}

State regulators hope, however, to increase both 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 is an attempt to create a single website where such organizations can register with and report to multiple jurisdictions, reducing both the compliance burden on the organizations and facilitating access to the registration and reporting information for all of the states involved.\textsuperscript{73} 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, although as of January 2019 it was only active in a pilot phase for two states.\textsuperscript{74} State officials also hope to take advantage of increased access to newly released IRS electronic data regarding tax-exempt nonprofits, as detailed below.\textsuperscript{75}

C. Nonprofits and the Public

There is an important aspect of the information collected by government regulators regarding nonprofits: information relating to nonprofits is in theory also available to both 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 such organizations. In this respect this information is therefore 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.\textsuperscript{76}

\textsuperscript{72} See FREMONT-SMITH, \textit{supra} note 63, at 305-06, 370; LOTT ET AL., \textit{supra} note 65, at 1-2.

\textsuperscript{73} See MRFP, INC., \textit{supra} note 11.


\textsuperscript{76} See \textit{Environmental Protection Agency (EPA), Methods, Models, Tools and Databases}, https://www.epa.gov/research/methods-models-tools-and-databases (last
The rationale for public disclosure of information related to tax-exempt nonprofit organizations is that it enables the public to ensure that such organizations deserve the significant tax benefits that they enjoy. Whether public disclosure actually increases legal compliance by nonprofits is unclear, however. A related rationale is that by claiming such benefits nonprofits voluntarily surrender any right to privacy they may have with respect to information they must then provide to the IRS. Similar considerations presumably are the basis for required public disclosure at the state level relating to charitable nonprofits. While privacy and potential misuse of disclosed information concerns generally 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 specific individuals.

More specifically, federal law generally requires both the IRS and tax-exempt nonprofits to allow any member of the public to access both applications for recognition of exemption (including all related correspondence) and recent annual information returns. While requesting such 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, such an approach would not be very practical for a researcher attempting to develop a broader database of information about such entities, however. On its website the IRS both 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.


79 Id. at 821.

80 See JCT, supra note 77, at 62-64; Mayer, supra note 78, at 821-23, 828-29.


Several private organizations have therefore successfully worked with – or 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. Machine-readable simply means that that the data is in a format that can be automatically read and processed by a computer. 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.

The first such group was the National Center for Charitable Statistics (NCCS), originally housed at 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. NCCS has developed and made publicly available a number of nonprofit databases, based in large part of information obtained from the IRS. Its current databases include both the data collections now also made available by the IRS on its website – the IRS Exempt Organizations Business Master Files (listing basic information for all active tax-exempt nonprofits that had filed as such with the IRS) for 1989 and 1995-2016 and IRS Statistics of Income Sample Data Files (information from a sample of Forms 990 and Form 990-EZ) for 1982-1983 and 1985-2012 – and NCCS developed databases containing selected information from all Forms 990, 990-EZ, and 990-PF for 1989-2015. NCCS also shares approximately 200 fields of financial data from the Form 990 with the nonprofit data website GuideStar. Working with the IRS, NCCS also developed a categorization system for nonprofits known as the National Taxonomy of Exempt Entities (NTEE). 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 to them for non-experts.\footnote{URBAN INSTITUTE, NONPROFIT DATA AT YOUR FINGERTIPS, http://nccs.urban.org/ (last visited Dec. 6, 2018).}

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 nonprofit.\footnote{GUIDESTAR, GUIDESTAR: A BRIEF HISTORY, https://learn.guidestar.org/about-us/history (last visited Dec. 6, 2018).} Unlike NCCS, its target audience is the public generally, including nonprofit leaders, donors, and journalists, but 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.\footnote{See GUIDESTAR, Hi! WE'RE GUIDESTAR, https://learn.guidestar.org/about-us (last visited Dec. 6, 2018) (data sources described under “What We Do”).} GuideStar likely has become the most well-known source for information about nonprofits because of both the breadth of its coverage – all tax-exempt nonprofits for which the IRS has information, including annual information returns now going back almost 20 years for charities and more than 10 years for other types of tax-exempt nonprofits – and its user friendly website.\footnote{See GUIDESTAR, GUIDESTAR: A BRIEF HISTORY, https://learn.guidestar.org/about-us/history (last visited Dec. 13, 2018) (Forms 990 and 990-EZ first posted in 1999; Form 990-PF in 2000; expansion of database to include all tax-exempt organizations in 2005).} In order to finance its operations, however, it requires payment for access to some search and analysis tools, downloading options, data services, reports, and older information.\footnote{See GUIDESTAR, GUIDESTAR PRODUCTS & SERVICES, https://learn.guidestar.org/products/ (last visited Dec. 6, 2018).}

The most recent entry into this space is ProPublica, a charitable nonprofit that focuses on investigative journalism.\footnote{See PROPUBLICA, ABOUT US, THE MISSION, https://www.propublica.org/about/ (last visited Dec. 6, 2018).} It launched its Nonprofit Explorer website earlier this decade, and has continually updated it since that time.\footnote{See Mike Tigas, Announcing the Nonprofit Explorer API (Sept. 26, 2013, 10:20 AM), https://www.propublica.org/nerds/announcing-the-nonprofit-explorer-api (last visited Dec. 6, 2018).} The website provides a search engine to gain access to IRS information, including Form 990s, for any given tax-exempt nonprofit beginning with 2013, along with certain federal grant audit information since 2016.\footnote{See PROPUBLICA, NONPROFIT EXPLORER, https://projects.propublica.org/nonprofits/ (last visited Dec. 13, 2018) (“About This Data” section).} It also provides an application programming interface that allows more robust searches and includes access to machine-readable data from e-filed Forms 990 filed from 2011 to the present.\footnote{See PROPUBLICA, NONPROFIT EXPLORER API, https://projects.propublica.org/nonprofits/api (last visited Dec. 6, 2018); [CITE NEEDED for latter statement].}
These machine-readable data became public in 2017 because of a lawsuit brought by an open data promoting organization. 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. Finally, there have been some efforts to create more specialized databases of nonprofit information. For example, a Vermont nonprofit has launched a database focused on nonprofits from that state.

As the introduction to this Part indicates, however, the mere release of such 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, including the Urban Institute and GuideStar, working together as the Nonprofit Open Data Collective to clean up those data in order to make them usable. For 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. The cleaned up data are now available on GitHub, an online platform (recently purchased by Microsoft). These data represent a major


102 Román, supra note 101.

103 Nonprofit Open Data Collective, supra note 52.
increase in the computerized data available from the Form 990, since it includes thousands of fields as compared to the approximately 200 fields previously digitized by the NCCS.\(^\text{104}\) Such cleaning work is continuing on the initially released data, and will of course have to continue indefinitely as the IRS continues to release data from newly filed annual information returns. It is important to note that these data only reflect e-filed returns, and so provide only a partial picture of Form 990, Form 990-EZ, and Form 990-PF filers.\(^\text{105}\)

Nonprofits and the public also have access to the state databases for charitable nonprofits in some of the approximately half of states that have such databases, although the extent of access varies.\(^\text{106}\) As with state use of such databases, it appears that public access comes with at most relatively simple search capabilities that while perhaps sufficient to locate information regarding a particular charitable nonprofit do not facilitate consideration of the data contained in such databases in the aggregate.\(^\text{107}\) 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 compiled their own databases and analyses of selected nonprofits drawing from a range of available sources, but these databases tend to be relatively limited in scope, accessibility, and searchability.\(^\text{108}\)

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.\(^\text{109}\) Its most recent

\(^{104}\) See supra note 87 and accompanying text.

\(^{105}\) See Nonprofit Open Data Collective, supra note 52 (“IRS 990 E-Filer Data” section).

\(^{106}\) LOTT ET AL., supra note 65, at 25.


Benchmarks Study focused on over 5 billion digital interactions with supporters of 154 participating nonprofits. But since these purely private efforts do not relate directly to government use of Big Data to regulate nonprofits, they are 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. For example, the Canada Revenue Agency both is expected to begin using Big Data methods to analyze its nonprofit database in the near future and since 2013 has made all of its datasets dating from 2000 forward available through the government’s data portal. The Charity Commission for England and Wales similarly provides the public with the ability to electronically extract the data from its registry of charities for all registered charities, 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, and the New Zealand Charities Services agency provides both a relatively robust search engine and the ability to download data in mass on its website. The charity regulators in Ireland and Scotland also allow public, online access to their charity registers, although 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 with these regulators. 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.\footnote{See Benefacts, What is Benefacts?, https://en.benefacts.ie/about/what-is-benefacts/ (last visited Dec. 6, 2018); Patricia Quinn, Benefacts: building a data infrastructure to support public policy for nonprofits in Ireland (July 2018) (unpublished manuscript) (on file with author).} This incomplete list only provides a sample of the data available outside the United States, as there are other private and government databases.\footnote{See, e.g., The Charity Vault, http://www.charity-charities.org/index.htm# (last visited Dec. 6, 2018) (private directory of more than 100,000 organizations around the world); IsraelGives, https://www.israelgives.org/pages/index (last visited Dec. 6, 2018) (private database of “all of Israel’s 42,000 non-profits and charities”).}

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. So the following discussion of the promises and perils of Big Data for regulating nonprofits is not only applicable to the United States, but to the many other countries considering or already moving into this area. This is particularly true because, as will be seen, most aspects of both promises and perils are not country-dependent but apply universally. The one major exception is with respect to legal limits, which of course are country-dependent.

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 \textit{a priori} hypotheses or explicit assumptions” that can then be used to improve actions or decisions.\footnote{U.S. Gov’t Accountability Off., GAO-16-659SP, Data and Analytics Innovation: Emerging Opportunities and Challenges 11 (2016).} 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.\footnote{See Kitchin, supra note 17, at 115-16; Manyika et al., supra note 5, at 61-62.} 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 percent of cases as opposed to the previous rate of...
13 percent. 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. 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 that oversee nonprofits is relatively easy to state if difficult to implement: to ensure compliance with the applicable laws (federal tax law for the IRS and state fiduciary duty and other laws for state agencies) in a fair manner. This task is complicated by the relatively limited resources that both the IRS and state regulators have when it comes to regulating nonprofits in the United States when compared to the number, financial size, and variety of those organizations, which limits are long-standing and well documented. 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. 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.

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. For

119 See supra note 84 and accompanying text.
121 See LOTT ET AL., supra note 65, at 8; Mayer, supra note 62, at 83.
122 Mayer, supra note 62, at 91.
124 TE/GE, supra note 41, at 6-8; TAX EXEMPT AND GOVERNMENT ENTITIES (TE/GE), FY 2017 WORK PLAN 3-8 (2016), https://www.irs.gov/pub/irs-tege/tege_fy2017_work_plan.pdf. 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
example, TE/GE noted that testing case selection modeling techniques resulted in a change rate of 90% for closed audits (presumably this means that in 90% of cases the IRS successfully asserted that the filing organization had to change its activities or reporting, although TE/GE did not provide a specific definition of “change”). More recent information released by the IRS indicated a change rate of slightly over 80%, which exceeds the IRS’ usual target. Increased change rates for audits would indicate improvements to both 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 (and the audited organization’s) time checking organizations that are in full compliance with the law – and effectiveness – given the limited number of audits the IRS can conduct because of resource constraints, a higher change rate indicates that fewer legal violations are going undetected and uncorrected.

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. For example, 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 enabling them to use analytics “to better understand charitable resources and solicitations, to better focus law enforcement and fraud prevention resources, and enable better policy making for protection of charitable resources.” 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 somewhat hindered by the need to not unduly tip their hand to potential bad actors. That said, the data they have made available to outside researchers may lead to improvements in this area as well.

**B. Improved Research**

28; Sony Kassam, *IRS Catches $10 Billion in Tax Fraud in 2018*, DAILY TAX REP., Nov. 14, 2018 (crediting a four-fold increase in the amount of tax fraud discovered to a focus on data analytics).  

125 See TE/GE, * supra* note 124, at 6-7.  
126 Von Lienen, * supra* note 43.  
128 MRFP, INC., * supra* note 11.  
129 See Tal Z. Zarsky, Transparent Predictions, 2013 U. Ill. L. Rev. 1503, 1512 (2013) (“[t]he IRS maintain full secrecy as to the selective auditing schemes it applies”).
Another often cited benefit of these Big Data moves with respect 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. For example, 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.” The IRS has not specifically promoted this research benefit with respect to the data it makes available, but the effect of its decision (under duress) 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 to researchers.

Recent research by Brian Galle and Terri Helge hints at the potential of such 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 analyzes 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. His research has important ramifications for the law relating to restricted gifts, and particularly with respect to 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 2014 thru early 2017 to determine the grounds for such denials and their relative frequency. 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, as well as helping to identify the common pitfalls for such organizations when they are first starting out.

131 MRFP, INC., supra note 11.
132 See supra notes 98-101 and accompanying text.
to legal compliance, researchers in other fields have also begun to take advantage of the increasing amount of available nonprofit data. For example, Nathan Grasse, Jesse Lecy, and their co-authors have been using nonprofit data to explore various topics, including government grants to nonprofits and issues relating to nonprofit compensation. Researchers in other countries are also exploring possible uses of nonprofit data newly available in their home countries.

Of course the potential for such research resulting from the greater public access to these government-collected datasets will depend on both the level of funding available and the related interest of academics in pursuing such 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.

C. Improved Information for the Public

The government-collected datasets also hold the promise of greater information reaching the public not only in the form of academic research results but also in the form of both direct access and access mediated by other intermediaries. In terms of direct access at the federal level, the IRS has recently revamped its now-renamed “Tax Exempt Organization Search” function to both 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. At the

---


136 See, e.g., Natasha Cortis, Linking non-profit data across government: an Australian example (July 2018) (unpublished manuscript) (on file with author); 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 (July 2018) (unpublished manuscript) (on file with author).


138 Press Release, Internal Revenue Service (IRS), 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-
state level, the Single Portal initiative will include a search feature to allow public access to the data collected.\textsuperscript{139}

Intermediaries who will likely benefit from the increase public access to information include both journalists and watchdog organizations. For example, a \textit{Wall Street Journal} reporter tapped into the newly available e-filed annual information returns data to identify all of the charity employees receiving seven-figure annual compensation amounts.\textsuperscript{140} 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 information regarding.\textsuperscript{141} 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.\textsuperscript{142}

\section*{III. The Perils of Using Big Data to Regulate Nonprofits}

The promises of applying Big Data methods to nonprofits, both with respect to direct effects on government oversight and indirect effects on academic research and information available to the public, are therefore potentially substantial although 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 generally, 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.\textsuperscript{143}

\begin{flushright}
to-public-for-first-time; Von Lienen, \textit{supra} note 43 (“We had a couple of little glitches with [the Tax-Exempt Organization Search] that have popped up.”).
\end{flushright}

\begin{flushright}
\textsuperscript{139}MRFP, I\textsc{nc.}, \textit{supra} note 11.
\end{flushright}

\begin{flushright}
\textsuperscript{140}Andrea Fuller, \textit{Charity Officials Are Increasingly Receiving Million-Dollar Paydays}, \textsc{Wall St. J.}, Mar. 6, 2017.
\end{flushright}

\begin{flushright}
\end{flushright}

\begin{flushright}
\textsuperscript{142}See \textsc{Charity Navigator, supra} note 141.
\end{flushright}

\begin{flushright}
\textsuperscript{143}See generally \textsc{David Bollier, The Promise and Peril of Big Data} (2010),
\end{flushright}
A. Getting It Wrong

One of the greatest concerns raised by commentators with respect to Big Data is the tendency of some Big Data supporters to oversell what it can accomplish. For example, one such supporter – Chris Anderson, then the Editor-in-Chief of Wired magazine – argued in 2008 that the emergence of Big Data was making the scientific method obsolete. 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 interpretations of the results of such analysis. 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 & 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. The information contained in such returns is not necessarily accurate, however. 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. Finally, financially smaller tax-exempt nonprofits can file shorter returns that contain significantly less information, and a large number of tax-exempt nonprofits –

---

145 boyd & Crawford, supra note 4, at 666.
146 See supra notes 101-102, 105 and accompanying text.
particularly churches and certain church-related entities— are not required to file such returns at all.\footnote{147}

While research on the accuracy of annual information returns is limited, the research that exists indicates that inaccuracies are relatively widespread with respect to at least some types of information. For example, in 2000 Karen Froelich, Terry Knoepfle, and Thomas Pollak compared the annual information returns for a sample of 350 nonprofit organizations with their audited financial statements.\footnote{148} 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.”\footnote{149} 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.\footnote{150} A more recent report by the Scrips Howard News Service found that 41 percent of the almost 38,000 tax-exempt nonprofits that reported at least $1 million in contributions also reported no fundraising costs on their returns.\footnote{151} It is therefore likely that such 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.\footnote{152}

Such inaccurate reporting may extend to other areas that may draw negative public or regulator attention. For example, 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 acknowledged in response.\textsuperscript{153} 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 such organizations will be.\textsuperscript{154}

One of the potential benefits of Big Data is to 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 relatively small sample of nonprofits.\textsuperscript{155} It is possible that the launch of the Single Portal initiative, and with it the ability to compare federal information returns with state filings (including possibly audited financial statements), 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 with respect to taxable individuals and entities.\textsuperscript{156} It remains to be seen, however, the extent to which such networking may identify and help correct inaccuracies in those returns. It is of course also possible that such 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 such ads, it could lead to many false positives.\textsuperscript{157}


\textsuperscript{154} See A Problem of the IRS’s Own Making – Automatic Revocations of 1023-EZ Exempt Organizations, NTA BLOG (May 16, 2018) (summarizing concerns relating to the Form 1023-EZ, including the relatively high rate of automatic revocation for failure to file required annual returns).

\textsuperscript{155} See supra note 148 and accompanying text.


With respect to cleaning data, the efforts to render the recently released e-filed information returns data illustrate the challenges in this area. 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, and their work is far from complete with respect to the initially available dataset, much less subsequent data releases. And of course until such time as 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 probably proportionately less as compared to financially larger organizations – at least if the smaller organizations are accurately identifying themselves as such – 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. This exempted group includes most notably all churches and certain church-related entities, regardless of their financial size or scope of activities (unless they have unrelated trade or businesses, in which case reporting with respect to such activities is generally required). 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. 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 revenue (before deducting any expenses) and $600 billion in real property assets. By comparison, for tax year 2015 tax-exempt charities that file Form 990 annual information returns reported $2.0 trillion in annual revenue.

organizations/the-restriction-of-political-campaign-intervention-by-section-501c3-tax-exempt-organizations (last visited Dec. 7, 2018) (identifying as “political” only communications or actions relating to candidates for elective public office).

See supra notes 101-102, 105 and accompanying text.


and $3.8 trillion in assets (partly offset by $1.5 trillion in liabilities).\textsuperscript{163} Any use by government regulators or private parties of Big Data methods therefore will miss this significant part of the nonprofit universe.

Lastly, it is important to recognize that 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 such area is information relating to how effective or impactful a nonprofit is with respect to 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; 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.\textsuperscript{164}

2. Analysis Inaccuracy & 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. Finding a correlation with big data techniques may not be an appropriate basis for predicting outcomes or behavior, or rendering judgments on individuals. In big data, as with all data, interpretation is always important.”\textsuperscript{165}

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


\textsuperscript{164} See IRS, supra note 45, at 11-12; 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”).

\textsuperscript{165} EXEC. OFF. OF THE PRESIDENT, supra note 20, at 7.
IRS scrutiny. 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 using other selection methods it is impossible to evaluate how effective the new process has been as compared to other methods. That said, the IRS is evaluating its methods; for example, 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 percent of identified cases, indicating this might not be a fruitful approach, although that was an improvement over a previous 65 no change rate. 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 strengthen internal controls with respect to selection processes then employed by TE/GE for tax-exempt nonprofits.

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 (eventually) the states, as well as the data available from other sources with respect to nonprofits. In doing so, they need to be especially conscious of the fact that in general such methods reveal correlations but do not demonstrate causation. And 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, for example, illegal behavior, but they rarely if ever generate one hundred percent accurate results. This is particularly true when Big Data methods are used to predict future behavior of individuals or organizations (a field known as predictive analytics), the risks of which were made famous by the short story and movie *The Minority Report*, although in the nonprofit context no one is advocating letting the computers find a violation of even civil law, much less criminal law, without a human-conducted, case-by-case analysis.

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 (the provision of a legally

---
166 See Zarsky, *supra* note 129.
167 See *supra* note 125 and accompanying text.
170 See MAYER-SCHÖNBERGER & CUKIER, *supra* note 118, at 50-72 (correlation and Big Data).
171 See Zarsky, *supra* note 129, at 1505-06.
prohibited, undue economic benefit to an insider). This result would indicate that nonprofits reporting no conflict of interest policy are at greater risk of (later discovered and reported) 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 lack of such a 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. And 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, especially given the government’s stretched resources with respect to regulating nonprofit), 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 not only accuracy but objectivity that lead to over-reliance on results. Simply because certain results are drawn from Big Data datasets using computerized analytics does not mean they provide an accurate or complete picture of, for example, the legal compliance of nonprofits such that other methods for measuring such 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.” Inaccuracy may arise from problems with the underlying data or the methods used to analyze it, as already noted. 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.

173 boyd & Crawford, supra note 4, at 666-68.
175 See Mayer, supra note 62, at 93-94.
B. Larger Concerns

Beyond possible inaccuracy and over-reliance concerns loom larger issues raised by Big Data generally. These include the risks to privacy it creates, the possibility it may lead to prohibited discriminatory treatment, the potential it has to enhance government power to the detriment of the governed, and other conflicts with constitutional and statutory limits on government power. While these concerns have broad applicability, they apply in specific ways to the application of Big Data to government regulation of nonprofits.

1. Invasion of Privacy: Welcome to the Goldfish Bowl

While loss of privacy is often considered the greatest risk of Big Data – think of Facebook and its vendors gathering and analyzing social media activity – at first glance this issue seems to be a non-concern with respect to nonprofits because information regarding them is already required to be publicly disclosed. There are, however, at least two privacy-related issues that arise in this context. 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 (“data fusion”) to provide a more complete profile of individuals and other entities associated with nonprofits.

Privacy in this context “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.” The risk of Big Data in this context 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 (or soon will be) readily accessible to anyone for (almost) any purpose. For example, much of the recent outrage against Facebook did not involve its

---


177 See EXEC. OFF. 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 (“[d]ata fusion occurs when data from different sources are brought into contact and new facts emerge”).

178 Id. at 2.

179 See Solove, supra note 174, at 1428 (“control out of control’ – a situation where nobody is exercising meaningful control over the information”)
use of data regarding its users, which use was presumably expected by them, but the use of its data by others such as Cambridge Analytica in ways that were unexpected (in some respects possibly even to Facebook).  

With respect to 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, for example, 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 (a study of Form 990 filings from 2001 to 2006 found that more than 132,000 nonprofits had included at least one such number on that form). 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.

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 with respect to donor information that is generally not subject to public disclosure for most tax-exempt nonprofits, with public release arising from both IRS and nonprofit errors (the latter including inadvertently submitting such information to state regulators who then make it publicly available). Some state authorities have also begun asking for unredacted copies of the federal schedule that lists donor information; while those authorities generally

---


182 See, e.g., IRS, supra note 45, at 2 (“Reminder: Don’t Include Social Security Numbers on Publicly Disclosed Forms”) (emphasis omitted).

commit to keep the schedules private, such requests add another possible source of inadvertent disclosures. The relatively few press reports describing such situations indicate that they are relatively rare, however, although they have attracted congressional attention. In fact, the IRS recently eliminated the requirement that non-charitable, tax-exempt nonprofits provide donor information to the IRS in part because of these concerns (it could not eliminate the requirement for charitable, tax-exempt nonprofits because that requirement is statutory).

There is also a more subtle but potentially greater privacy concern raised by the increasing ease of access to return and other information relating to nonprofits. 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. For example, say a nonprofit reports – as it is required to do – the compensation paid to its various officers and other highly compensated employees. Most people, at least in the United States, view compensation information as private and so generally only to be shared selectively with others. Senior employees and officers of tax-exempt nonprofits hopefully know, however, 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 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, although such practices do not appear to be widespread at this point.

---

184 See, e.g., Americans for Prosperity Foundation v. Harris, 809 F.3d 536, 538 (9th Cir. 2015) (California requirement that charities submit to the Attorney General a copy of the (nonpublic) donor information submitted to the IRS).


186 See 26 U.S.C. § 6033(b)(5) (section 501(c)(3) organizations must include in their annual information return “the names and addresses of all substantial contributors”); Rev. Proc. 2018-38, 2018-31 I.R.B. 280 (other than for section 501(c)(3) organizations, eliminating the requirement that tax-exempt nonprofits report the names and addresses of their contributors on their annual information returns). The Senate voted to override the IRS’s decision in late 2018, but the House did not follow suit. See S.J. Res. 64, 115th Cong. (2018).

187 See EXEC. OFF. OF THE PRESIDENT, BIG DATA AND DIFFERENTIAL PRICING 8-13
Fortunately there is little evidence of more nefarious use of such data, whether by criminals or others, although that could be a risk as well in the future. Nevertheless, the greater availability of what many would consider sensitive, personal information may raise concerns with those who consider working for a nonprofit.\textsuperscript{188}

2. Prohibited Discrimination: Can Computers Be Racist?

Another prominent concern with respect to Big Data is that it could lead to discrimination based on race, gender, or other generally prohibited grounds, including inadvertently through, for example, the reproduction of existing patterns of discrimination.\textsuperscript{189} Countering such discrimination can be tricky, including because adjusting the results of Big Data analysis to compensate for such impacts may be technically difficult or may itself raise discrimination concerns.\textsuperscript{190} 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 (because women control a higher share of household spending).\textsuperscript{191} But discrimination concerns can also arise in many other situations, such as access to credit and higher education.\textsuperscript{192}

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

\textsuperscript{188} For similar concerns with respect 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 (“the 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 campaign advertisements”).


\textsuperscript{190} See Barocas & Selbst, supra note 189, at 714-15.


\textsuperscript{192} See EXEC. OFF. OF THE PRESIDENT, supra note 189, at 4.
Be On the Lookout (BOLO) list that unintentionally disparately impacted politically conservative applicants by subjecting them to greater scrutiny than some other applicants (all other things being equal). The same risk exists with respect to 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 such areas. Another example might be if compensation information for nonprofit executives was used to get around laws prohibiting inquiring about compensation history, which laws are usually justified as seeking to reduce gender disparities in compensation.

Computer scientists are well aware of these possible issues and have engaged in extensive research on possible technical fixes to address them. For example, 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. While this may be less possible in the nonprofit context given the enforcement-focused 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, then the necessary resources have to 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.\textsuperscript{197} Big Data, with its mass collection and analysis of information relating to the governed and their organizations, certainly raises such concerns.\textsuperscript{198} This concern is particularly pronounced when government makes what 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.\textsuperscript{199}

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 with respect to any nonprofit in their jurisdiction for any rational reason other than reasons that are constitutionally or otherwise legally prohibited (such as making selections based on expressing certain viewpoints). The one exception for the IRS is with respect to churches, for which Congress requires specific findings to commence an inquiry or audit, but since the IRS lacks annual return data for such entities the application by the IRS of a Big Data approach to them is not feasible at this point under any conditions.\textsuperscript{200} Second, the data collected and analyzed with respect to nonprofits are generally 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.\textsuperscript{201}

There is, however, one issue in this regard. Merely 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, many commentators, including the TE/GE’s own Advisory Committee, raised this concern when the IRS added questions relating to certain governance practices to the returns even though federal tax law does not require any of those practices.\textsuperscript{202} The increasing ability of both the IRS


\textsuperscript{198} See EXEC. OFF. OF THE PRESIDENT, \textit{supra} note 20, at 22.


\textsuperscript{200} See 26 U.S.C. § 7611 (2018); \textit{supra} note 45 and accompanying text.


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. This underlines the dynamic nature of correlations; merely asking certain questions might cause a change in behavior, which might in turn be correlated with greater (or less) legal compliance.

If merely 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 such restraint is lacking at the agency level, for the relevant legislature to forbid them from doing so.

4. Other Legal Limits: Building a Plane While You’re 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. The protections provided by these provisions are related in part, but not entirely, to the privacy concerns discussed above. There are few if any statutory or other legal limitations specifically targeting government use of Big Data methods, supra note 201, at 480 n.3 (other critics).


205 See KUHN, supra note 204, at 11-12 (different conceptualizations of privacy).
however, although the Federal Trade Commission issued non-binding “Fair Information Practices in the Electronic Marketplace” set of guidelines relating to consumer-oriented commercial Internet sites almost 20 years ago that also had implications for government collection and use of data.206 In the federal tax context, the IRS is limited by generally applicable taxpayer privacy requirements with respect to information Congress has not explicitly made subject to public disclosure and also by third party contact rules.207

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.208 Such use implicates the Fourth Amendment to the extent the collection and use of personal data without the knowledge and consent of the affected individuals is a search or seizure in the Fourth Amendment sense.209 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.210

In the nonprofit context these constitutional concerns are relatively 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 (and improper disclosure of the limited information that is not subject to such a requirement appears to be relatively rare).211 But the possible use of Big Data methods to reveal connections involving individuals still has privacy implications, as discussed previously, and also Big Data methods could lead to disparate impacts based on the views expressed by nonprofits and so 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


210 See Kuhn, supra note 204, at 200-01 (due process and privacy as knowledge control).

211 See supra notes 81, 106 and accompanying text (public disclosure of nonprofit government filings); see supra note 183 (instances of improper disclosure).
generally 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 with respect to 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. Current law is that individuals generally do not have an expectation of privacy when such information is held by a third party and so the Fourth Amendment does not apply, although the Supreme Court recently held that an individual does have “a legitimate expectation of privacy in the record of his physical movements as captured by [cell-site location data]” and so such information is protected by the Fourth Amendment.212 Finally, and as Danielle Keats Citron has highlighted, any automated decision making process – such as with respect to selecting nonprofits to audit or otherwise investigate – raises significant due process concerns with respect to notice and the opportunity to heard, and such concerns are heightened if such a process results in the automatic imposition of some type of penalty – such as revocation of tax-exempt status or a cease and desist order with respect to charitable solicitation.213 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 generally 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 with respect to the collection and analysis of data.214 For example, with respect to maintaining taxpayer privacy the IRS is not at liberty to publicly disclose which specific tax-exempt nonprofits it is auditing or even the results of such audits unless the audit leads to revocation of tax exemption.215 This limitation hinders the ability of outside parties to evaluate the effectiveness of the IRS’ 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 generally seeks to obtain the desired information from the audit target, such as a tax-exempt nonprofit, before

212 Carpenter v. United States, 138 S.Ct. 2206, 2216-17 (2018); see also Hatfield, supra note 57, at 331 & n.75.
214 See supra note 207 and accompanying text.
contacting a third party. It is not clear, however, that this limitation would apply to general request for information relating to multiple tax-exempt nonprofits, such as 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 personal involvement on the other end,” which exception might apply to at least some broad inquiries.

The uncertainty regarding whether and how both the constitutional limitations and the applicable statutory limitations apply with respect to Big Data collection and analysis mean that nonprofits subject to such 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. For example, it took congressionally attention to end the IRS tactic of reading emails of taxpayers without consent, notice, or court permission. 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, leads to several recommendations that regulators, researchers, and nonprofits should consider. For both agencies and legislatures, continued pursuit of these methods is clearly desirable, but specific steps must be taken both 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 be aware that other, publicly accessible information such as their websites and social media presences may be compared with those filings for consistency and completeness.

216 Internal Revenue Services (IRS), Internal Revenue Manual § 25.27.1.3(1) (2017).
217 Id. § 25.27.1.2.
A. For Regulators

Given the acknowledged resource limitations faced by both 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.\(^{219}\) At the same time, and as the IRS has acknowledged, not every algorithm will prove effective and efficient in identifying nonprofits that are likely out of compliance with the applicable law.\(^{220}\) This is particularly true given the possible accuracy issues with both the underlying data and the tools used to analysis them.\(^{221}\) It is therefore important both 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 do so as well.\(^{222}\) The IRS has in fact stated that it continues to rely heavily on its “robust referral process” to identify noncompliance, as well as more data-driven methods.\(^{223}\) 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.\(^{224}\)

Regulators may also have to consider making more efforts to ensure that government filings are accurate. While in theory inaccuracies can result in an IRS return being treated as incomplete (and so late filing penalties applying) or even in a perjury conviction for the nonprofit leader who signs the return, it is at best unclear how often the IRS imposes penalties for inaccurate information (as opposed to failing to file a return by the required due date).\(^{225}\)

\(^{219}\) See supra notes 121-123 and accompanying text (resource limitations).

\(^{220}\) See supra note 168 and accompanying text (ineffectiveness of approach used to try to identify noncompliant private foundations).

\(^{221}\) See supra Sections III.A.1, III.A.2.


\(^{223}\) Von Lienen, supra note 43.

\(^{224}\) See supra note 129 and accompanying text.

\(^{225}\) See 26 U.S.C. § 6652(c)(1) (2018) (excise tax can be imposed on a tax-exempt nonprofit for “a failure to include any information required to be shown on [an annual information return] . . . or to show the correct information”); IRS, supra note 46, at 42-42 & 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).
While in early 2018 the IRS implemented a new process for rejecting incomplete or incorrect paper-filed Form 990s, leading to an over 10 percent rejection rate, it is unclear what other steps the IRS has taken to improve return accuracy. 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 with respect to their required filings.

These points are of course relatively 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. More specifically, with respect to 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, although it acknowledged that it does not identify or redact other personally identifiable information such as home addresses and the ages of students receiving scholarships.

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 with respect to 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. Similar considerations need to reflected in evaluations of Big Data methods adopted with respect to overseeing nonprofits. The last point is especially important, because even a highly effective algorithm (in that it is highly successful in detecting legal noncompliance) may have to be set aside if its results also are fundamentally unfair in that they 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

---

226 See Von Lienen, supra note 43.
227 See supra Section III.B.
229 FTC, supra note 143, at iv-v.
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. 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 both 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 repeatedly considered expanding the e-filing requirements for tax-exempt nonprofits, but the requirement has yet to achieve legislative success.

Congress and state legislatures also could take steps to address some of the larger concerns identified above. For example, Congress should 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. This is not unprecedented, as in the election law context Congress has prohibited the sale or use of information regarding individual contributors (which is publicly disclosed) for soliciting contributions or for any commercial purpose. This rule has survived constitutional challenge. Congress 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

---

230 See supra note 218 and accompanying text.
231 See supra note 60 and accompanying text.
232 See supra notes 187-188 and accompanying text.
234 See Federal 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 the statute under the First Amendment).
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. As the amount and visibility of these data grows, presumably this group will also grow. This is a welcome development, as both the resources and interests of government agencies are relatively limited, while private researchers should be able to tap additional funding sources and analyze important issues (for example, gender differences in compensation) that may not attract agency attention. In fact, at least one IRS official has encouraged feedback from private researchers.

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. 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.” The currently available machine-readable data is limited in at 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 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. 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

---

235 See supra notes 133-135 and accompanying text.
236 The example is based on an announced but not yet available article titled Gender and the Compensation of Nonprofit Executive Directors and Chief Financial Officers by Nathan Grasse et al.
237 See Lough, supra note 44 (inviting the public to look at the IRS-related e-filing data and “tell us what they’re seeing”).
238 See supra notes 148-154 and accompanying text.
240 See supra notes 52, 86 and accompanying text.
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 examples, researchers interested in exploring the relative impacts of various nonprofits likely will 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.²⁴¹

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 generally limited to nonprofits that are smaller financially, which may make them less important for some but not all research purposes.²⁴² The latter include most notably churches and church-related entities, but also include, for example, public colleges and universities.²⁴³ So, for example, 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. Researchers that want to analyze issues not fully supported by the existing data because of these limitations therefore need to consider to gather data from other sources to address these representational and 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 both 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 a government filings, the increased access those returns and the data they contain only further emphasizes their public relations role.

²⁴¹ See CHARITY WATCH, OVERHEAD RATIOS ARE ESSENTIAL FOR INFORMED GIVING, https://www.charitywatch.org/charitywatch-articles/overhead-ratios-are-essential-for-informed-giving/133 (last visited Dec. 14, 2018) (summarizing the debate relating to using such ratios to evaluate charities while defending their use).
²⁴² See supra notes 41–48 and accompanying text.
²⁴³ See supra note 45 and accompanying text.
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, for example, 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). While that revelation did not prevent Donald Trump’s election, it did create some negative publicity and force him to both correct the expenditure (presumably by reimbursing the Foundation) and pay a penalty tax to the IRS on behalf of the Foundation. 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, both 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 presents.

---

244 See Allison Graves, Donald Trump, Pam Bondi and $25K: Was it pay to play?, POLITICO FLORIDA (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/ (inaccurate reporting of donation on Foundation IRS annual return); Letter from Noah Bookbinder, Executive Director, Citizens for Responsibility and Ethics in Washington to IRS Commissioner John A. Koskinen 2 (Mar. 21, 2016) (Florida campaign finance filing by recipient political organization revealed donation from Foundation).

245 See Graves, supra note 244.
may present.

Big Data may allow government regulators to deploy their limited resources in a much more efficient manner and so 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 both internal and external evaluators so as to enhance this review and at the same time minimize the perils posed by using Big Data techniques.

With respect to those perils, regulators and private parties must be conscious of the limitations on the data available to them, both with respect to accuracy and comprehensiveness, and on the analytical tools they employ, and not forego the continued use of other methods to help compensate for these limitations. 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, and establish procedures to detect and counter such possibilities if they arise. Fortunately there is time to put such procedures in place at both the federal and state level, but this task must be considered an integral part of establishing a Big Data approach and not merely an afterthought. If it is, then the perils of Big Data are likely to be avoided while its promises are fulfilled with respect to the regulation of nonprofits.